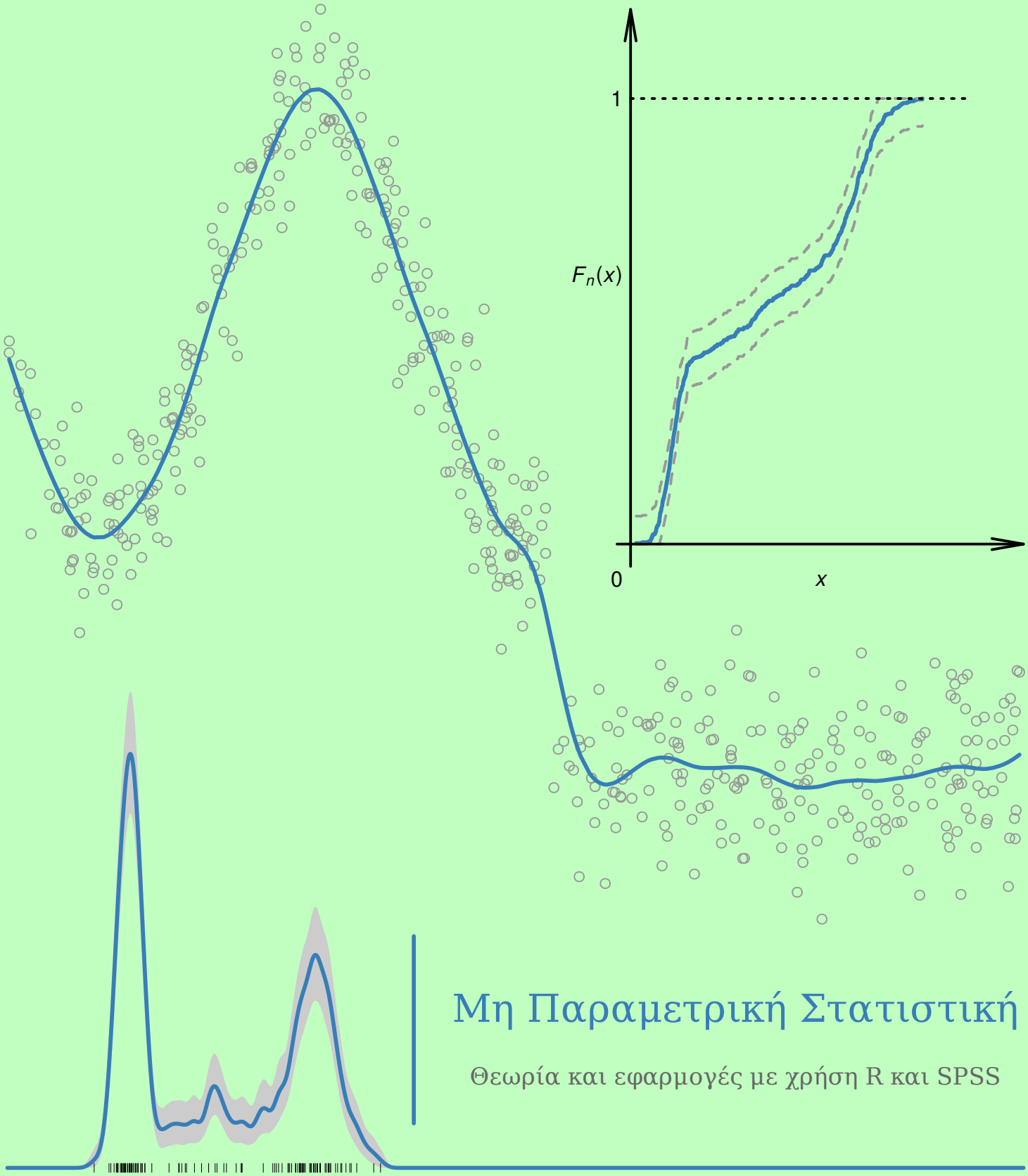


Μη Παραμετρική Στατιστική

Θεωρία και εφαρμογές με χρήση R και SPSS

Α. Μπατσίδης, Π. Παπασταμούλης, Κ. Πετρόπουλος, Α. Ρακιτζής

ΜΗ ΠΑΡΑΜΕΤΡΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

ΘΕΩΡΙΑ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΜΕ ΧΡΗΣΗ R ΚΑΙ SPSS

Μη Παραμετρική Στατιστική

Θεωρία και εφαρμογές με χρήση R και SPSS

Απόστολος Μπατσίδης
Αναπληρωτής Καθηγητής
Πανεπιστήμιο Ιωαννίνων

Παναγιώτης Παπασταμούλης
Επίκουρος Καθηγητής
Οικονομικό Πανεπιστήμιο Αθηνών

Κωνσταντίνος Πετρόπουλος
Αναπληρωτής Καθηγητής
Πανεπιστήμιο Πατρών

Αθανάσιος Ρακιτζής
Επίκουρος Καθηγητής
Πανεπιστήμιο Πειραιώς

ΚΑΛΛΙΠΟΣ
ΑΝΟΙΚΤΕΣ ΑΚΑΔΗΜΑΪΚΕΣ ΕΚΔΟΣΕΙΣ
www.kallipos.gr

Τίτλος πρωτοτύπου: <<*Μη Παραμετρική Στατιστική*>>

Συντελεστές έκδοσης
Γλωσσική επιμέλεια: Βασιλική Τυραϊδή

Κεντρική Ομάδα Υποστήριξης
Γλωσσικός Έλεγχος: Γεωργία Τριανταφυλλίδου

Γραφιστικός Έλεγχος: Χρήστος Κεντρωτής

Βιβλιοθηκονομική Επεξεργασία: Έλενα Αδαμοπούλου

ΚΑΛΛΙΠΟΣ

Εθνικό Μετσόβιο Πολυτεχνείο

Ηρώων Πολυτεχνείου 9

15780 Ζωγράφου

www.kallipos.gr

Βιβλιογραφική αναφορά: Μπατσίδης, Α., Παπασταμούλης, Π., Πετρόπουλος, Κ., & Ρακιτζής, Α. (2022). *Μη Παραμετρική Στατιστική* [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις.

Διαθέσιμο στο: http://dx.doi.org/10.57713/kallipos-102

ISBN: **978-618-5667-92-4**

Στην οικογένειά μου ---Α.Μ.---

Στην κόρη μου Μαρίνα ---Π.Π.---

Στη σύζυγό μου Χριστίνα και στον γιο μου Παναγιώτη ---Κ.Π.---

Στη σύζυγό μου Ερασμία και στους γιους μου Χρήστο και Δημήτρη ---Α.Ρ.---

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΡΟΛΟΓΟΣ

Οι παραδοσιακές μεθοδολογίες της Στατιστικής, όπως οι έλεγχοι υποθέσεων, η ανάλυση διακύμανσης και άλλες, έχουν θεμελιωθεί υπό την υπόθεση ότι τα διαθέσιμα δεδομένα μπορούν να θεωρηθούν ότι προέρχονται από έναν πληθυσμό με γνωστή κατανομή, με τις τιμές των πληθυσμιακών παραμέτρων να είναι συνήθως άγνωστες. Δηλαδή υιοθετείται γνώση της συναρτησιακής μορφής της συνάρτησης πιθανότητας ή της συνάρτησης πυκνότητας πιθανότητας. Στο πλαίσιο αυτό, λόγω του Κεντρικού Οριακού Θεωρήματος και των ιδιοτήτων του, το μοντέλο της κανονικής κατανομής είναι αυτό που κυρίως χρησιμοποιείται. Οι στατιστικές μεθοδολογίες που στηρίζονται στη γνώση της συναρτησιακής μορφής της κατανομής του πληθυσμού αποτελούν τη λεγόμενη Παραμετρική Στατιστική.

Αν και σε πολλά πρακτικά προβλήματα οι υποθέσεις της Παραμετρικής Στατιστικής είναι λογικές, δεν είναι λίγες οι περιπτώσεις όπου οι υποθέσεις αυτές δεν ικανοποιούνται. Επιπλέον, ακόμα και στην περίπτωση όπου δικαιολογείται η χρήση ενός παραμετρικού μοντέλου, ενδέχεται τα συμπεράσματά μας να είναι αρκετά ευαίσθητα στις υποθέσεις αυτού του μοντέλου. Συνεπώς, έχει νόημα η αναζήτηση και η εφαρμογή εναλλακτικών στατιστικών μεθοδολογιών, που δεν περιορίζουν το μοντέλο που χρησιμοποιείται για την περιγραφή των δεδομένων σε μία συγκεκριμένη παραμετρική οικογένεια κατανομών. Η συλλογή αυτών των εναλλακτικών μεθοδολογιών αποτελεί τη λεγόμενη Μη Παραμετρική ή Απαραμετρική Στατιστική. Συνεπώς, μπορούμε να πούμε πως η βασική ιδέα της Μη Παραμετρικής Στατιστικής είναι η συμπερασματολογία με όσο το δυνατόν λιγότερες υποθέσεις.

Βασικός στόχος της συγγραφικής ομάδας ήταν η δημιουργία ενός εύληπτου συγγράμματος για όλους τους/τις φοιτητές/φοιτήτριες τριτοβάθμιων ιδρυμάτων, οι οποίοι/οποίες διδάσκονται για πρώτη φορά το προαναφερθέν αντικείμενο, εφοδιάζοντας το αναγνωστικό κοινό με όλο το απαραίτητο θεωρητικό υπόβαθρο που θα του επιτρέπει να επιλέγει και να εφαρμόζει την κατάλληλη μεθοδολογία για την επίλυση προβλημάτων της επιστημονικής περιοχής του. Απώτερος στόχος είναι να παρουσιαστούν τόσο παραδοσιακές μεθοδολογίες της Μη Παραμετρικής Στατιστικής (π.χ. έλεγχοι τάξεων, ροών, καλής προσαρμογής) όσο και πιο σύγχρονες (π.χ. εκτίμηση της συνάρτησης πυκνότητας πιθανότητας, bootstrap, μη παραμετρική παλινδρόμηση). Ωστόσο επισημαίνεται ότι απαραίτητη προϋπόθεση για τη μελέτη του παρόντος συγγράμματος αποτελεί η πρότερη γνώση βασικών εννοιών Πιθανοτήτων και Στατιστικής.

Στην κατεύθυνση αυτή, η διάρθωση των κεφαλαίων του παρόντος συγγράμματος αναλύεται παρακάτω. Στο Κεφάλαιο 1 γίνεται υπενθύμιση βασικών ορισμών και εννοιών από τη Θεωρία Πιθανοτήτων και τη Στατιστική.

Μπατσίδης, Α., Παπασταμούλης, Π., Πετρόπουλος, Κ., & Ρακιτζής, Α. (2022). *Μη Παραμετρική Στατιστική.* [Προπτυχιακό εγχειρίδιο].

Επίσης, γίνεται μια εισαγωγή στην περιοχή της μη παραμετρικής στατιστικής, στην οποία παρουσιάζονται (α) η αναγκαιότητα των μη παραμετρικών μεθοδολογιών, (β) οι διαφορές τους με τις παραμετρικές μεθόδους και (γ) τα πεδία και οι περιοχές εφαρμογής των μη παραμετρικών μεθόδων.

Στο Κεφάλαιο 2 δίνονται μέθοδοι και τεχνικές για τη μη παραμετρική εκτίμηση της (άγνωστης) αθροιστικής συνάρτησης κατανομής και συναρτησιακών αυτής, ενώ στο Κεφάλαιο 3 οι αντίστοιχες μέθοδοι αφορούν τη μη παραμετρική εκτίμηση της συνάρτησης πυκνότητας πιθανότητας. Στο Κεφάλαιο 4 παρουσιάζονται οι κυριότεροι έλεγχοι καλής προσαρμογής, ενώ στο Κεφάλαιο 5 παρουσιάζονται οι πλέον απλές (αλλά συχνά χρησιμοποιούμενες) τεχνικές για τον έλεγχο υποθέσεων, οι οποίες βασίζονται στη Διωνυμική κατανομή. Στο Κεφάλαιο 6 μελετώνται τεχνικές οι οποίες βασίζονται στη χρήση τάξεων και χρησιμοποιούνται για τον έλεγχο στατιστικών υποθέσεων, ενώ οι έλεγχοι τυχαιότητας παρουσιάζονται στο Κεφάλαιο 7. Στο Κεφάλαιο 8 παρουσιάζονται μη παραμετρικά στατιστικά μέτρα και οι αντίστοιχοι έλεγχοι για τη συσχέτιση δύο μεταβλητών. Τεχνικές μη παραμετρικής παλινδρόμησης αναπτύσσονται στο Κεφάλαιο 9. Στα Κεφάλαια 10 και 11 παρουσιάζονται η μέθοδος Jackknife και η μέθοδος Bootstrap, αντίστοιχα, οι οποίες είναι ευρέως χρησιμοποιούμενες υπολογιστικές στατιστικές τεχνικές. Στο Κεφάλαιο 12 παρουσιάζονται βασικές μη παραμετρικές τεχνικές που χρησιμοποιούνται στον στατιστικό έλεγχο διεργασιών, ενώ στο Κεφάλαιο 13 παρουσιάζεται η εφαρμογή των πιο συχνά χρησιμοποιούμενων μη παραμετρικών μεθοδολογιών με χρήση του στατιστικού πακέτου IBM SPSS, αλλά και της γλώσσας προγραμματισμού `R`. Επιπλέον, στο Παράρτημα του συγγράμματος δίνονται πίνακες τιμών αθροιστικών συναρτήσεων κατανομών, πίνακες κρίσιμων τιμών και ποσοστιαίων σημείων, ώστε να είναι δυνατή η εφαρμογή των στατιστικών τεχνικών που παρουσιάστηκαν στο κύριο μέρος του συγγράμματος, παρότι πολλά από τα αποτελέσματα που εκεί παρατίθενται μπορούν να βρεθούν με τη βοήθεια της γλώσσας προγραμματισμού `R`. Τέλος, από τον ιστότοπο `https://github.com/abatsidis/NPDataSets` μπορεί ο/η αναγνώστης/αναγνώστρια να αποκτήσει πρόσβαση σε σύνολα δεδομένων και σε κώδικες της `R` που έχουν χρησιμοποιηθεί για την υλοποίηση μη παραμετρικών μεθοδολογιών στο πλαίσιο κεφαλαίων αυτού του συγγράμματος.

Ευχόμαστε το παρόν σύγγραμμα να καλύψει πλήρως τις ανάγκες όλων όσων διδάσκουν ή διδάσκονται ή απλά επιθυμούν να πρωτογνωρίσουν τον κόσμο της Μη Παραμετρικής Στατιστικής. Αν και έγινε προσπάθεια ελαχιστοποίησης των λαθών, αστοχιών και παραλείψεων, με υψηλή πιθανότητα τέτοια έχουν παραμείνει στο παρόν κείμενο. Για κάθε παράλειψη ή λάθος υπεύθυνοι είναι οι συγγραφείς και κάθε παρατήρηση και σχόλιο για βελτίωση του παρόντος συγγράμματος είναι ευπρόσδεκτα.

Τέλος, οι συγγραφείς επιθυμούν να ευχαριστήσουν, από τη θέση αυτή, όλους τους/τις φοιτητές/φοιτήτριες που συνέβαλαν με την επισήμανση αστοχιών και αβλεψιών σε αρχικές πανεπιστημιακές διδακτικές σημειώσεις, οι οποίες και αποτέλεσαν τον αρχικό πυρήνα για τη συγγραφή αυτού του βιβλίου.

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ - ΥΠΕΝΘΥΜΙΣΗ ΒΑΣΙΚΩΝ ΕΝΝΟΙΩΝ

Σύνοψη

Στο κεφάλαιο αυτό γίνεται υπενθύμιση βασικών ορισμών και εννοιών, όπως είναι η συνάρτηση πυκνότητας πιθανότητας (σ.π.π), η συνάρτηση πιθανότητας (σ.π.), η αθροιστική συνάρτηση κατανομής (α.σ.κ.), τα ποσοστημόρια (ή ποσοστιαία σημεία), η μέση τιμή και η διακύμανση τυχαίων μεταβλητών. Επιπρόσθετα, δίνονται η σ.π.π., η μέση τιμή και η διακύμανση διατεταγμένων τυχαίων μεταβλητών, καθώς και οι απαραίτητες έννοιες σχετικά με τη σύγκλιση ακολουθιών τυχαίων μεταβλητών. Ακόμη, γίνεται υπενθύμιση βασικών εννοιών σχετιζόμενων με τον έλεγχο υποθέσεων, όπως είναι η στατιστική συνάρτηση ελέγχου (ή ελεγχοσυνάρτηση), το επίπεδο σημαντικότητας και η κρίσιμη περιοχή. Επίσης, γίνεται μια εισαγωγή στην περιοχή της μη παραμετρικής στατιστικής, όπου παρουσιάζονται (α) η αναγκαιότητα των μη παραμετρικών μεθοδολογιών, (β) οι διαφορές τους με τις παραμετρικές μεθόδους, και (γ) τα πεδία και οι περιοχές εφαρμογής των μη παραμετρικών μεθόδων. Τέλος, δίνονται οι τύποι των συναρτήσεων πιθανότητας (περίπτωση διακριτών τυχαίων μεταβλητών) ή των συναρτήσεων πυκνότητας πιθανότητας (περίπτωση συνεχών τυχαίων μεταβλητών) των πιο συχνά χρησιμοποιούμενων πιθανοτικών μοντέλων.

Προαπαιτούμενη γνώση:

Βασικές γνώσεις μαθηματικών, όπως απειροστικού λογισμού και άλγεβρας.

Προσδοκώμενα μαθησιακά αποτελέσματα:

Ο/η φοιτητής/τρια ολοκληρώνοντας την ενότητα αυτή θα μπορεί να προχωρήσει στα επόμενα κεφάλαια του βιβλίου, τα οποία πραγματεύονται θέματα και τεχνικές μη παραμετρικής στατιστικής. Επίσης, θα μπορεί να ανατρέχει στο παρόν κεφάλαιο προκειμένου να ανανεώσει τις γνώσεις του/της σε βασικές έννοιες από τις Πιθανότητες και τη Στατιστική. Επιπλέον, γίνεται μια σύντομη εισαγωγή στην περιοχή της μη παραμετρικής στατιστικής, στους τρόπους εφαρμογής της και παρουσιάζονται τα πλεονεκτήματα/μειονεκτήματα της χρήσης των μη παραμετρικών τεχνικών έναντι της χρήσης των αντίστοιχων παραμετρικών.

Μπατσίδης, Α., Παπασταμούλης, Π., Πετρόπουλος, Κ., & Ρακιτζής, Α. (2022). *Μη Παραμετρική Στατιστική.* [Προπτυχιακό εγχειρίδιο].

Γλωσσάριο επιστημονικών όρων

- Αθροιστική συνάρτηση κατανομής
- Αναμενόμενη τιμή
- Από κοινού κατανομή τυχαίων μεταβλητών
- Διατεταγμένη τυχαία μεταβλητή
- Δείγμα
- Δειγματοχώρος ή δειγματικός χώρος
- Διασπορά
- Διάστημα εμπιστοσύνης
- Έλεγχος υποθέσεων
- Ενδεχόμενα
- Μη παραμετρικός έλεγχος
- Πληθυσμός
- Στατιστική διαδικασία απαλλαγμένη παραμέτρων
- Στατιστική διαδικασία ελεύθερη κατανομής
- Στατιστική συνάρτηση
- Συνάρτηση πιθανότητας
- Συνάρτηση πυκνότητας πιθανότητας

1.1 Εισαγωγή

Υπάρχουν πολλά φαινόμενα -ή συστήματα- των οποίων η επαναλαμβανόμενη παρατήρηση υπό τις ίδιες συνθήκες οδηγεί σε διαφορετικά αποτελέσματα. Χαρακτηριστικό παράδειγμα είναι η ρίψη ενός νομίσματος ή η ρίψη ενός ζαριού. Η μεταβλητότητα στα αποτελέσματα μπορεί να οφείλεται στην τύχη, αλλά κάποιες φορές μπορεί να οφείλεται και στην επίδραση ορισμένων παραγόντων που δεν μπορούν να ελεγχθούν ή είναι άγνωστοι και επηρεάζουν το αποτέλεσμα. Για παράδειγμα, ας υποθέσουμε ότι δύο φάρμακα που χρησιμοποιούνται για την καταπολέμηση της ίδιας ασθένειας έχουν παραχθεί με την ίδια διαδικασία, έχουν δοκιμαστεί υπό τις ίδιες συνθήκες, χρησιμοποιούνται υπό τις ίδιες, γενικά, συνθήκες, αλλά παρατηρείται ότι ο χρόνος αποθεραπείας μεταξύ ομάδων ασθενών, που δοκιμάζουν το συγκεκριμένο φάρμακο, διαφέρει. Η διαφορά αυτή μπορεί να οφείλεται στον τρόπο αντίδρασης του ανθρώπινου οργανισμού στην αγωγή (και η αντίδραση αυτή να είναι διαφορετική μεταξύ των ατόμων) ή στην ύπαρξη κάποιου άλλου παράγοντα (π.χ. βεβαρημένου ιατρικού ιστορικού), ο οποίος παράγοντας διαφοροποιεί τον χρόνο αποθεραπείας στις δύο ομάδες.

Για την περιγραφή φαινομένων, όπως τα παραπάνω, των οποίων η ακριβής συμπεριφορά δεν είναι προβλέψιμη, χρησιμοποιούνται πιθανοτικά πρότυπα ή μοντέλα πιθανότητας ή κατανομές πιθανότητας. Συγκεκριμένα, το πιθανοτικό πρότυπο καταγράφει τα δυνατά αποτελέσματα του φαινομένου που περιγράφει/μοντελοποιεί και αντιστοιχεί πιθανότητες σε αυτά τα δυνατά αποτελέσματα. Οι πιθανότητες αυτές είναι είτε μεμονωμένοι αριθμοί είτε προκύπτουν από κάποιο δεδομένο πιθανοτικό νόμο, δηλαδή μια κατανομή πιθανότητας ή απλά κατανομή.

Πιο συγκεκριμένα, όσον αφορά τα προαναφερθέντα, η **Θεωρία Πιθανοτήτων**, της οποίας κύριο αντικείμενο είναι η μελέτη των κατανομών πιθανότητας και των συστημάτων που περιγράφονται από αυτά, αποτελεί τη βάση για την επιστήμη της **Στατιστικής**. Η Στατιστική ασχολείται (i) με τη συλλογή, (ii) την περιγραφή και (iii) την περιληπτική παρουσίαση δεδομένων, αλλά, κυρίως, (iv) με την εξαγωγή γενικότερων συμπερασμάτων με βάση έναν περιορισμένο αριθμό μετρήσεων ή παρατηρήσεων από τον υπό μελέτη πληθυσμό. Γενικά, όταν υπάρχει ανάγκη για τουλάχιστον ένα εκ των (i)-(iv), τότε είναι απαραίτητη η χρήση των τεχνικών και των μεθόδων της Στατιστικής. Ο κλάδος της Στατιστικής, που πραγματεύεται την περιγραφή και παρουσίαση δεδομένων με χρήση πινάκων, γραφικών παραστάσεων και αριθμητικών μέτρων, ονομάζεται **Περιγραφική Στατιστική**. Εκτός της Περιγραφικής Στατιστικής, ένας άλλος κλάδος της Στατιστικής είναι η **Στατιστική Συμπερασματολογία**, η οποία αποτελεί αντικειμενική μεθοδολογία επαγωγικής γενίκευσης για μια ολότητα τιμών (**πληθυσμός**), που χαρακτηρίζει το υπό μελέτη φαινόμενο, με βάση τις πληροφορίες που περιέχονται σε ένα μέρος αυτού, το οποίο ονομάζεται **δείγμα**.

Οι τιμές του πληθυσμού στο δείγμα καλούνται παρατηρήσεις. Το δείγμα πρέπει να είναι αντιπροσωπευτικό του πληθυσμού. Για να συμβαίνει αυτό, θα πρέπει κάθε στοιχείο (ή μέλος) του πληθυσμού, να έχει συγκεκριμένη πιθανότητα να συμπεριληφθεί στο δείγμα.

Οι τιμές που συνιστούν τον πληθυσμό μπορεί να αντιστοιχούν σε αριθμητικές/ποσοτικές μετρήσεις ενός ποσοτικού χαρακτηριστικού ή να εκφράζουν την παρουσία/απουσία ενός ποιοτικού χαρακτηριστικού, για ένα πλήρως καθορισμένο σύνολο μονάδων (αντικειμένων, ατόμων, οντοτήτων κ.λπ.). Με τον όρο **πληθυσμό** ορίζουμε κάθε καλά ορισμένη συλλογή οντοτήτων ή καταστάσεων ή αντικειμένων ή ατόμων που μοιράζονται κοινά χαρακτηριστικά, τα οποία διέπονται από κάποιο βαθμό τυχαιότητας. Αυτά τα χαρακτηριστικά (ή μεταβλητές), τα οποία μπορεί να είναι ποσοτικά ή ποιοτικά, θέλουμε να τα μελετήσουμε στο πλαίσιο ενός φαινομένου ή ενός συστήματος. Το πλήθος των στοιχείων ενός πληθυσμού καλείται μέγεθος του πληθυσμού.

Ως **δείγμα** ορίζεται ένα μέρος του πληθυσμού, το οποίο επιλέγεται μέσω ενός πιθανοκρατικού μηχανισμού παρατήρησης ή, αλλιώς, πείραμα. Το πλήθος των στοιχείων ενός δείγματος καλείται **μέγεθος δείγματος**.

Στη συνέχεια, στις Ενότητες 1.2-1.8 αυτού του κεφαλαίου, γίνεται υπενθύμιση βασικών ορισμών και εννοιών

της Θεωρίας Πιθανοτήτων και της Στατιστικής. Ο/η ενδιαφερόμενος/η αναγνώστης/τρια, προτού προχωρήσει με τη μελέτη των υπόλοιπων κεφαλαίων, μπορεί ενδεικτικά να ανατρέξει, αν επιθυμεί, σε βασικά βιβλία αναφοράς για τέτοια θέματα, όπως είναι, μεταξύ άλλων, τα ακόλουθα: Δαμιανού και Κούτρας (1998), Ηλιόπουλος (2012), Κούτρας (2018), Κουτρουβέλης (1999a), Κουτρουβέλης (1999b), Παπαϊωάννου και Φερεντίνος (2000), Χαραλαμπίδης (2000a), Χαραλαμπίδης (2000b) και Ζωγράφος (2008), τα οποία και χρησιμοποιήθηκαν για τη συγγραφή αυτού του κεφαλαίου.

1.2 Στοιχεία θεωρίας πιθανοτήτων

Στην ενότητα αυτή παρουσιάζονται η έννοια της πιθανότητας και οι διάφοροι ορισμοί της, τα βασικά εργαλεία υπολογισμού της πιθανότητας ενός ενδεχομένου σε ένα πείραμα τύχης, ενώ αναπτύσσονται οι βασικές μέθοδοι για την επίλυση προβλημάτων υπολογισμού πιθανοτήτων. Με τον όρο **τυχαίο πείραμα** εννοείται οποιαδήποτε καλά καθορισμένη διαδικασία που παράγει ένα παρατηρήσιμο αποτέλεσμα που δεν θα μπορούσε να προβλεφθεί επακριβώς εκ των προτέρων (βλ. Siegel, 2011). Παραδείγματα τυχαίων πειραμάτων είναι η ρίψη ενός νομίσματος, η ρίψη ενός ζαριού, ο αριθμός των εύστοχων τρίποντων σε 10 προσπάθειες ενός καλαθοσφαιριστή, ο χρόνος ζωής ενός ιού κ.λπ.

1.2.1 Δειγματικός χώρος, ενδεχόμενα και πράξεις ενδεχομένων

Ορισμός 1.1: Δειγματικός χώρος

Ως δειγματικός χώρος Ω ορίζεται το σύνολο των δυνατών αποτελεσμάτων του υπό πραγματοποίηση τυχαίου πειράματος.

Παράδειγμα 1.1. Έστω το τυχαίο πείραμα της παρατήρησης του αποτελέσματος που θα φέρει η ρίψη ενός ζαριού. Ποιος είναι ο δειγματικός χώρος Ω; Αν θεωρηθεί ως πείραμα τύχης η ρίψη δύο ζαριών ή τριών ζαριών, ποιος είναι ο δειγματικός χώρος Ω; Αν κατά τη ρίψη των δύο ζαριών μας ενδιαφέρει το άθροισμα των αποτελεσμάτων που θα φέρουμε κατά τη ρίψη τους, ποιος είναι ο δειγματικός χώρος Ω;

Λύση Παραδείγματος 1.1. Για το πείραμα τύχης της ρίψης ενός ζαριού ο δειγματικός χώρος Ω είναι το σύνολο $\{1,2,3,4,5,6\}$. Επίσης, αν θεωρήσουμε το πείραμα της ρίψης δύο ζαριών και της παρατήρησης των αποτελεσμάτων που θα φέρουμε κατά τη ρίψη αυτών, τότε ο δειγματικός χώρος είναι το σύνολο

$$\Omega = \{(x,y) : x = 1,2,\dots,6, y = 1,2,\dots,6\}.$$

Στο δεύτερο πείραμα, ο Ω περιέχει 36 στοιχεία (τα ζεύγη (1,1), (1,2), (1,3), κ.λπ.). Επίσης, ας θεωρήσουμε πάλι ότι εκτελείται το πείραμα της ρίψης των δύο ζαριών και έστω τώρα ότι μας ενδιαφέρει το άθροισμα των αποτελεσμάτων που θα φέρουμε κατά τη ρίψη τους. Τότε, ο δειγματικός χώρος είναι το σύνολο $\Omega = \{2,3,\dots,12\}$, το οποίο περιέχει 11 στοιχεία. Τέλος, αν θεωρήσουμε ότι εκτελείται το πείραμα της ρίψης ενός νομίσματος τρεις φορές, τότε, καθώς σε κάθε ρίψη έχουμε δύο δυνατά αποτελέσματα (Κ=κορώνα ή Γ=Γράμματα), ο δειγματικός χώρος του πειράματος είναι

$$\Omega = \{\text{ΚΚΚ}, \text{ΓΚΚ}, \text{ΚΓΚ}, \text{ΚΚΓ}, \text{ΓΓΚ}, \text{ΓΚΓ}, \text{ΚΓΓ}, \text{ΓΓΓ}\}.$$

□

Παράδειγμα 1.2. Σε ένα εργοστάσιο ελέγχονται τα προϊόντα που βγαίνουν από κάποια γραμμή παραγωγής. Μας ενδιαφέρει ο συνολικός αριθμός των επιθεωρημένων προϊόντων μέχρι το πρώτο ελαττωματικό προϊόν. Ποιος είναι ο δειγματικός χώρος Ω;

Λύση Παραδείγματος 1.2. Σε αυτό το πείραμα τύχης, ο δειγματικός χώρος είναι το σύνολο των φυσικών αριθμών $\Omega = \mathbb{N} \equiv \{1, 2, 3 \ldots\}$. □

Παράδειγμα 1.3. Ας υποθέσουμε ότι σε ένα συγκεκριμένο σημείο ενός αυτοκινητόδρομου καταγράφεται ο χρόνος που μεσολαβεί μεταξύ των διαδοχικών διελεύσεων των αυτοκινήτων από το σημείο αυτό. Όσο μικρότερος είναι ο χρόνος αυτός, τόσο μεγαλύτερος είναι ο αριθμός των αυτοκινήτων που κινούνται στον αυτοκινητόδρομο. Άρα, μπορούμε να θεωρήσουμε ότι αποτελεί και μέτρο αξιολόγησης του κυκλοφοριακού φόρτου. Ποιος είναι ο δειγματικός χώρος στην περίπτωση που μας ενδιαφέρει η μελέτη του ενδιάμεσου χρόνου μεταξύ διαδοχικών διελεύσεων;

Λύση Παραδείγματος 1.3. Ο κατάλληλος δειγματικός χώρος, στην περίπτωση που μας ενδιαφέρει η μελέτη του ενδιάμεσου χρόνου μεταξύ διαδοχικών διελεύσεων, είναι το (άπειρο) σύνολο $\Omega = \{t \in \mathbb{R} : 0 < t < \tau_1\}$, όπου t είναι ο ενδιάμεσος χρόνος (σε sec) μεταξύ των διαδοχικών διελεύσεων και $\tau_1 > 0$. □

Ο δειγματικός χώρος ενός πειράματος μπορεί να είναι είτε ένα σύνολο με πεπερασμένο πλήθος στοιχείων (όπως στο Παράδειγμα 1.1) είτε ένα άπειρο, αλλά αριθμήσιμο σύνολο (όπως στο Παράδειγμα 1.2) είτε, τέλος, ένα άπειρο μη αριθμήσιμο σύνολο (όπως στο Παράδειγμα 1.3). Οι δειγματικοί χώροι των δύο πρώτων περιπτώσεων λέγονται *διακριτοί*, ενώ οι δειγματικοί χώροι που ανήκουν στην τρίτη κατηγορία λέγονται *συνεχείς*.

Ενδεχόμενα (ή γεγονότα) ονομάζονται όλα τα υποσύνολα ενός δειγματικού χώρου (διακριτού ή συνεχούς). Το Ω είναι το βέβαιο ενδεχόμενο, δηλαδή το ενδεχόμενο που πραγματοποιείται πάντοτε, ενώ το $\varnothing$ είναι το αδύνατο ενδεχόμενο, δηλαδή το ενδεχόμενο που δεν μπορεί να πραγματοποιηθεί ποτέ. Επίσης, **απλό** ή **στοιχειώδες ενδεχόμενο** ονομάζεται κάθε μονοσύνολο ενδεχόμενο, δηλαδή κάθε υποσύνολο ενός δειγματοχώρου Ω που αποτελείται από ένα και μόνο στοιχείο αυτού. Θα λέμε ότι ένα ενδεχόμενο E εμφανίζεται σε μία πραγματοποίηση ενός πειράματος (δοκιμής), όταν το αποτέλεσμα ω της δοκιμής, δηλαδή το στοιχειώδες ενδεχόμενο που «έφερε» (εμφάνισε) η δοκιμή, ανήκει στο σύνολο E (δηλαδή $\omega \in E$).

Έστω A και B δύο ενδεχόμενα σε ένα πείραμα τύχης. Τότε μπορούμε να ορίσουμε τις παρακάτω πράξεις μεταξύ ενδεχομένων:

- Η **ένωση** των A και B, $A \cup B$, είναι το ενδεχόμενο που εμφανίζεται σε μία δοκιμή, όταν στην ίδια δοκιμή εμφανιστεί μόνο το A ή μόνο το B ή εμφανιστούν και τα δύο ενδεχόμενα μαζί.
- Η **τομή** των ενδεχομένων A και B, $A \cap B$, είναι το ενδεχόμενο που εμφανίζεται, όταν εμφανιστούν και τα δύο ενδεχόμενα A, B στην ίδια δοκιμή.
- Το **συμπλήρωμα** A' του ενδεχομένου A (ή το συμπληρωματικό ενεδεχόμενο του A) εμφανίζεται, όταν δεν εμφανιστεί το A.
- Η **διαφορά** $A - B$ ορίζεται ως το ενδεχόμενο $A \cap B'$ και εμφανίζεται, όταν στην ίδια δοκιμή εμφανιστεί το A και δεν εμφανιστεί το B. Αν $A \subseteq B$, τότε $A - B = \varnothing$.

Έστω τώρα n το πλήθος ενδεχόμενα $A_1, A_2, \ldots, A_n$ ενός πειράματος τύχης. Τότε, η ένωση $A_1 \cup \ldots \cup A_n = \bigcup_i^n A_i$ εμφανίζεται, όταν εμφανιστεί τουλάχιστον ένα από αυτά. Επίσης, η τομή $A_1 \cap \ldots \cap A_n = \bigcap_i^n A_i$ εμφανίζεται, όταν εμφανιστούν όλα μαζί.

Όταν τα $A_1, A_2, \ldots, A_n$ έχουν την ιδιότητα η εμφάνιση του ενός να αποκλείει την εμφάνιση οποιουδήποτε άλλου στην ίδια δοκιμή, τότε αυτά λέγονται **ασυμβίβαστα ή ξένα ανά δύο**, δηλαδή $A_i \cap A_j = \varnothing$, για κάθε $i \neq j$, $i, j = 1, 2, \ldots, n$. Τέλος, έστω n το πλήθος ενδεχόμενα $A_1, A_2, \ldots, A_n$ ενός πειράματος τύχης. Όταν $A_i \cap A_j = \varnothing$, για κάθε $i \neq j$, $i, j = 1, 2, \ldots, n$, και $\bigcup_{i=1}^n A_i = \Omega$, τότε λέμε ότι τα $A_1, A_2, \ldots, A_n$ εξαντλούν από

κοινού τον δειγματικό χώρο ή, ισοδύναμα, ότι αποτελούν μια **διαμέριση** του Ω.

1.2.2 Συνδυαστική ανάλυση

Για τον εύκολο, γρήγορο και κυρίως σωστό υπολογισμό των πιθανοτήτων πολλές φορές απαιτείται η γνώση τεχνικών Συνδυαστικής Ανάλυσης. Στη συνέχεια, θα περιοριστούμε στην παράθεση της αποκαλούμενης βασικής αρχής απαρίθμησης και κάποιων βασικών στοιχείων συνδυαστικής.

Βασική Αρχή Απαρίθμησης: Αν ένα πείραμα μπορεί να αναλυθεί σε r υποπειράματα, όπου το πρώτο υποπείραμα έχει ν_1 το πλήθος δυνατά αποτελέσματα, το δεύτερο υποπείραμα έχει ν_2 το πλήθος δυνατά αποτελέσματα, ανεξάρτητα από το αποτέλεσμα του πρώτου πειράματος, ..., το r-οστό υποπείραμα έχει ν_r το πλήθος δυνατά αποτελέσματα, ανεξάρτητα από τα αποτελέσματα των προηγούμενων $r-1$ το πλήθος πειραμάτων, τότε το συνολικό πείραμα έχει $\nu = \nu_1 \cdot \nu_2 \cdot \ldots \cdot \nu_r$ το πλήθος δυνατά αποτελέσματα.

Διατάξεις: Έστω ένα σύνολο με ν το πλήθος στοιχεία. Αν πάρουμε r το πλήθος από αυτά ($1 \le r \le \nu$) και τα κατατάξουμε σε μια σειρά, τότε λέμε ότι παίρνουμε μια *διάταξη* των ν στοιχείων ανά r. Το σύνολο των διατάξεων είναι ίσο με

$$P(\nu, r) = \nu \cdot (\nu-1) \cdot \ldots \cdot (\nu-r+1) = \frac{r!}{(\nu-r)!}, \quad r = 1, 2, \ldots, \nu$$

και $P(\nu, \nu) = \nu!$

Διατάξεις με επανάληψη: Το πλήθος των διατάξεων ν το πλήθος στοιχείων ανά r με επανάληψη είναι ίσο με ν^r.

Συνδυασμοί: Έστω ένα σύνολο με ν το πλήθος στοιχεία. Ένα οποιοδήποτε υποσύνολο από r το πλήθος στοιχεία από αυτά ($0 \le r \le \nu$) λέγεται συνδυασμός των ν ανά r. Το πλήθος των συνδυασμών είναι ίσο με

$$C(\nu, r) = \binom{\nu}{r} = \frac{\nu!}{r!(\nu-r)!} = \binom{\nu}{\nu-r}.$$

Συνδυασμοί με επανάληψη: Έστω ένα σύνολο με ν το πλήθος στοιχεία. Το πλήθος των συνδυασμών των ν στοιχείων ανά r με επανάληψη είναι ίσο με

$$\left[\begin{matrix} \nu \\ r \end{matrix} \right] = \frac{\nu(\nu+1)\cdots(\nu+r-1)}{r!} = \binom{\nu+r-1}{r}.$$

1.2.3 Κλασικός ορισμός πιθανότητας

Ο πρώτος ορισμός που θα αναφέρουμε είναι ο αποκαλούμενος κλασικός ορισμός, ο οποίος αναφέρεται και ως «ορισμός πιθανότητας κατά Laplace».

Ορισμός 1.2: Κλασικός ορισμός πιθανότητας

Αν ένα τυχαίο πείραμα έχει ν το πλήθος ισοδύναμα δυνατά αποτελέσματα, η πιθανότητα ενός ενδεχομένου A είναι ο λόγος $\nu(A)/\nu$, όπου $\nu(A)$ είναι ο αριθμός των ευνοϊκών αποτελεσμάτων του

τυχαίου πειράματος για το ενδεχόμενο A.

Ο υπολογισμός της πιθανότητας ενός ενδεχομένου A στην περίπτωση πεπερασμένου δειγματικού χώρου Ω, του οποίου τα στοιχεία είναι ισοπίθανα, μπορεί να γίνει με τη χρήση του κλασικού ορισμού της πιθανότητας. Συγκεκριμένα, πρέπει να υπολογιστούν ο αριθμός $\nu(A)$, δηλαδή το πλήθος των στοιχείων στο ενδεχόμενο A (ευνοϊκές περιπτώσεις), και ο αριθμός ν (πλήθος δυνατών αποτελεσμάτων, δηλαδή το πλήθος των δυνατών περιπτώσεων). Για τον υπολογισμό των $\nu(A)$, ν, χρησιμοποιούνται τεχνικές συνδυαστικής ανάλυσης, κάποιες εκ των οποίων παρουσιάστηκαν στην προηγούμενη υποενότητα.

1.2.4 Εμπειρική πιθανότητα

Παρά την ευκολία του, ο κλασικός ορισμός παρουσιάζει το σημαντικό μειονέκτημα ότι μπορεί να εφαρμοστεί μόνο σε πειράματα τύχης που έχουν πεπερασμένο πλήθος δυνατών αποτελεσμάτων, με όλα τα στοιχειώδη (απλά) ενδεχόμενα να είναι ισοπίθανα. Για την αντιμετώπιση των προβλημάτων του κλασικού ορισμού προτάθηκε ο ορισμός της πιθανότητας ως όριο της σχετικής συχνότητας. Ο ορισμός αυτός αναφέρεται στη βιβλιογραφία ως «ορισμός πιθανότητας κατά Von Mises» ή, πιο απλά, ως «στατιστικός ορισμός» ή ως «εμπειρική πιθανότητα».

Ορισμός 1.3: Στατιστικός ορισμός πιθανότητας

Έστω A ένα ενδεχόμενο του δειγματικού χώρου Ω ενός πειράματος τύχης, $A \subseteq \Omega$. Υποθέτουμε ότι το πείραμα τύχης επαναλαμβάνεται υπό τις ίδιες συνθήκες n το πλήθος φορές. Τότε η πιθανότητα $P(A)$ του ενδεχομένου A ορίζεται ως η σχετική συχνότητα $n(A)/n$ εμφάνισης του A, σε έναν μεγάλο αριθμό n επαναλήψεων του πειράματος.

Ο παραπάνω ορισμός, ως πλαίσιο υπολογισμού της πιθανότητας ενός ενδεχομένου, είναι χρήσιμος στις περιπτώσεις που η βασική διαδικασία του πειράματος μπορεί να επαναληφθεί πολλές φορές υπό τις ίδιες συνθήκες. Καθώς κάτι τέτοιο δεν είναι πάντοτε εφικτό, εισήχθη στη βιβλιογραφία ο αξιωματικός ορισμός της πιθανότητας.

1.2.5 Αξιωματικός ορισμός πιθανότητας

Το 1933 ο Andrey Kolmogorov διατύπωσε έναν αυστηρό ορισμό της πιθανότητας βασιζόμενος σε τρία αξιώματα. Για τον λόγο αυτόν, ο συγκεκριμένος ορισμός αναφέρεται ως «αξιωματικός ορισμός της πιθανότητας». Ο συγκεκριμένος ορισμός ξεπερνάει τις αδυναμίες των προηγούμενων ορισμών και δεν έρχεται σε αντίθεση με τους αρχικούς ορισμούς της πιθανότητας, σε περιπτώσεις που αυτοί μπορούν να εφαρμοστούν.

Ορισμός 1.4: Αξιωματικός Ορισμός Πιθανότητας

Έστω Ω ο δειγματικός χώρος ενός πειράματος, $P(A)$ η πιθανότητα ενός ενδεχομένου A, $A \subseteq \Omega$. Ως πιθανότητα ορίζουμε μια συνάρτηση με πεδίο ορισμού την οικογένεια όλων των υποσυνόλων του Ω που είναι ενδεχόμενα (το πεδίο των ενδεχομένων), η οποία ικανοποιεί τα εξής τρία αξιώματα:

A1. Για κάθε ενδεχόμενο A, $P(A) \geq 0$.

A2. $P(\Omega) = 1$.

A3. Για κάθε αριθμήσιμη ακολουθία ασυμβίβαστων ανά δύο ενδεχομένων $A_1, A_2, \ldots$ ισχύει ότι:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Άμεση απόρροια του παραπάνω ορισμού είναι οι παρακάτω βασικές προτάσεις (Θεωρήματα) του αξιωματικού ορισμού της πιθανότητας.

Βασικές προτάσεις (Θεωρήματα) της αξιωματικής θεωρίας

Π1. $P(\varnothing) = 0$, δηλαδή η πιθανότητα να συμβεί το αδύνατο ενδεχόμενο είναι μηδέν.

Π2. Για κάθε πεπερασμένη ακολουθία ασυμβίβαστων ανά δύο ενδεχομένων $A_1, A_2, \dots, A_n$ του Ω ισχύει

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

Π3. Για κάθε δύο ενδεχόμενα A, B του Ω με $B \subseteq A$ ισχύουν οι ακόλουθες σχέσεις:
- $P(B) \leq P(A)$.
- $P(A - B) = P(A) - P(B)$.

Π4. Για κάθε ενδεχόμενο $A \in \Omega$, $P(A) \leq 1$ (και άρα, $P(A) \in [0,1]$).

Π5. Για κάθε ενδεχόμενο $A \in \Omega$, $P(A') = 1 - P(A)$.

Π6. (**Κανόνας της πρόσθεσης**) Για οποιαδήποτε ενδεχόμενα A, B του Ω ισχύει ότι:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Επίσης, για δύο ενδεχόμενα A, B του δειγματικού χώρου Ω σε ένα πείραμα τύχης, ισχύουν και οι παρακάτω σχέσεις, οι οποίες είναι γνωστές ως «τύποι του De Morgan»:

- $P(A' \cap B') = P((A \cup B)')$
- $P(A' \cup B') = P((A \cap B)')$

1.2.6 Δεσμευμένη πιθανότητα

Η έννοια της δεσμευμένης πιθανότητας μας επιτρέπει τον υπολογισμό της πιθανότητας πραγματοποίησης ενός ενδεχομένου, αν γνωρίζουμε ότι έχει πραγματοποιηθεί ένα άλλο ενδεχόμενο του ίδιου δειγματικού χώρου.

Ορισμός 1.5: Ορισμός Δεσμευμένης Πιθανότητας

Έστω A, B δύο ενδεχόμενα ενός πειράματος τύχης με δειγματικό χώρο Ω και έστω ότι $P(B) > 0$. Η δεσμευμένη πιθανότητα του A όταν γνωρίζουμε ότι έχει συμβεί το B συμβολίζεται με $P(A|B)$ και ορίζεται από τη σχέση:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Επίσης, για δύο ενδεχόμενα A και B ενός πειράματος τύχης, υποθέτοντας ότι οι πιθανότητες $P(B|A), P(A|B)$ μπορούν να οριστούν, ισχύει η παρακάτω σχέση (**Κανόνας του Γινομένου ή Πολλαπλασιαστικός κανόνας**):

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B).$$

1.2.7 Ανεξαρτησία ενδεχομένων

Σε πολλές περιπτώσεις η γνώση της πραγματοποίησης ενός ενδεχομένου αλλάζει την πιθανότητα πραγματοποίησης ενός άλλου ενδεχομένου. Σε κάποιες, όμως, περιπτώσεις, η πληροφορία ότι έχει πραγματοποιηθεί το ένα ενδεχόμενο δεν αλλάζει την πιθανότητα πραγματοποίησης του άλλου. Δηλαδή, σε αυτήν την περίπτωση, η πραγματοποίηση ενός από τα δύο ενδεχόμενα δεν επηρεάζει την πραγματοποίηση ή όχι του άλλου ενδεχομένου. Για τις περιπτώσεις αυτές έχουμε τον ακόλουθο ορισμό.

Ορισμός 1.6: Ανεξάρτητα Ενδεχόμενα

Έστω δύο ενδεχόμενα A, B ενός πειράματος τύχης με δειγματικό χώρο Ω. Τα A, B λέγονται ανεξάρτητα, όταν ισχύει η σχέση $P(A \cap B) = P(A)P(B)$.

Γενικεύοντας τον παραπάνω ορισμό, αν θεωρήσουμε τα n το πλήθος ενδεχόμενα $A_1, A_2, \dots, A_n$ ενός πειράματος τύχης, τότε αυτά θα λέγονται **αμοιβαία ανεξάρτητα**, αν για οποιοδήποτε φυσικό αριθμό k, με $2 \leq k \leq n$ και για οποιοδήποτε υποσύνολο k το πλήθος ενδεχομένων από τα n, έστω $A_{i_1}, A_{i_2}, \dots, A_{i_k}$, ισχύει ότι:

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}).$$

Επίσης, τα n το πλήθος πειράματα $\Pi_1, \dots, \Pi_n$ λέγονται ανεξάρτητα, όταν οποιαδήποτε n ενδεχόμενα (ένα ενδεχόμενο από κάθε πείραμα) είναι ανεξάρτητα ενδεχόμενα.
Άμεσα από τα παραπάνω έχουμε πως, αν A, B είναι ανεξάρτητα ενδεχόμενα ενός πειράματος τύχης, με $P(A), P(B) > 0$, τότε $P(A|B) = P(A)$ και $P(B|A) = P(B)$.

Κλείνουμε την παρούσα υποενότητα, διατυπώνοντας δύο πολύ βασικά θεωρήματα στον λογισμό πιθανοτήτων.

Θεώρημα 1.1: Θεώρημα Ολικής Πιθανότητας

Έστω E οποιοδήποτε ενδεχόμενο του δειγματικού χώρου Ω και $A_1, A_2, \dots, A_n$, ασυμβίβαστα ανά δύο ενδεχόμενα, που εξαντλούν από κοινού τον δειγματικό χώρο Ω και είναι τέτοια ώστε $P(A_i) > 0$, για $i = 1, 2, \dots, n$. Τότε ισχύει ότι:

$$P(E) = P(E|A_1)P(A_1) + \dots + P(E|A_n)P(A_n).$$

Θεώρημα 1.2: Κανόνας του Bayes

Έστω E οποιοδήποτε ενδεχόμενο του δειγματικού χώρου Ω και $A_1, A_2, \dots, A_n$, ασυμβίβαστα ανά δύο ενδεχόμενα, που εξαντλούν από κοινού τον δειγματικό χώρο Ω και είναι τέτοια ώστε $P(A_i) > 0$, για $i = 1, 2, \dots, n$. Τότε, για $i = 1, 2, \dots, n$, ισχύει ότι:

$$P(A_i|E) = \frac{P(E|A_i)P(A_i)}{P(E|A_1)P(A_1) + \dots + P(E|A_n)P(A_n)}.$$

1.3 Τυχαίες μεταβλητές

Έστω $(\Omega, \mathscr{A}, P)$ ένας χώρος πιθανότητας, δηλαδή Ω είναι ο δειγματικός χώρος του πειράματος τύχης, $\mathscr{A}$ η σ-άλγεβρα[1] υποσυνόλων του Ω και P ένα μέτρο πιθανότητας. Τότε, τα στοιχειώδη ενδεχόμενα $\omega \in \Omega$ μπορούν να αναπαρασταθούν ως αριθμοί (για παράδειγμα, αν εκφράζουν ένα ποσοτικό χαρακτηριστικό του τυχαίου πειράματος) ή με σύμβολα (για παράδειγμα, αν εκφράζουν ένα ποιοτικό χαρακτηριστικό). Για να μπορέσουμε να αντιστοιχίσουμε κάθε στοιχειώδες ενδεχόμενο $\omega \in \Omega$ σε έναν αριθμό $x \in \mathbb{R}$, χρειαζόμαστε

[1] Ως σ-άλγεβρα $\mathscr{A}$ ορίζεται κάθε συλλογή από υποσύνολα του Ω που ικανοποιεί τις ακόλουθες ιδιότητες: α) $\Omega \in \mathscr{A}$, β) αν $A \in \mathscr{A}$ τότε $A' \in \mathscr{A}$, και γ) αν έχουμε μια ακολουθία συνόλων A_n, $n = 1, 2, \dots$ στο $\mathscr{A}$, τότε η ένωση των A_n ανήκει επίσης στο $\mathscr{A}$.

μία κατάλληλη συνάρτηση η οποία θα κάνει αυτήν την αντιστοίχιση. Η συνάρτηση αυτή είναι γνωστή και ως **τυχαία μεταβλητή**. Ως **Τυχαία Μεταβλητή** (τ.μ.), έστω X, ορίζεται να είναι μία μονοσήμαντη συνάρτηση με πεδίο ορισμού έναν δειγματικό χώρο Ω και τιμές ένα υποσύνολο των πραγματικών αριθμών, δηλαδή

$$X : \Omega \to \mathbb{R}_X \equiv X(\Omega), \text{ όπου } \mathbb{R}_X \equiv X(\Omega) = \{x \in \mathbb{R} : X(\omega) = x,\ \omega \in \Omega\}.$$

Αν $X(\Omega)$ είναι ένα πεπερασμένο ή αριθμήσιμο σύνολο, τότε η τυχαία μεταβλητή είναι μια **διακριτή τυχαία μεταβλητή**. Αν ο $X(\Omega)$ είναι ένα υπερ-αριθμήσιμο σύνολο, τότε η τυχαία μεταβλητή X είναι μια **συνεχής τυχαία μεταβλητή**.

Έχοντας ορίσει την έννοια της τυχαίας μεταβλητής, μπορούμε να προχωρήσουμε στον ορισμό της αθροιστικής συνάρτησης κατανομής (α.σ.κ.), στον ορισμό της συνάρτησης πιθανότητας (σ.π.) και στον ορισμό της συνάρτησης πυκνότητας πιθανότητας (σ.π.π.).

Ορισμός 1.7: Αθροιστική Συνάρτηση Κατανομής

Έστω $(\Omega, \mathscr{A}, P)$ είναι χώρος πιθανότητας και $X : \Omega \to S_X \subseteq \mathbb{R}$ μία τυχαία μεταβλητή. Το S_X είναι ένα υποσύνολο των πραγματικών αριθμών. Η **συνάρτηση κατανομής** ή **αθροιστική συνάρτηση κατανομής** (α.σ.κ.) της τυχαίας μεταβλητής X συμβολίζεται με $F_X(\cdot)$ και είναι μία πραγματική συνάρτηση με $F_X : \mathbb{R} \to [0,1]$, που ορίζεται από τη σχέση:

$$F_X(x) = P(X \le x) = P(\{\omega \in \Omega : X(\omega) \le x\}), \quad x \in \mathbb{R}.$$

Ορισμός 1.8: Συνάρτηση Πιθανότητας

Έστω X μία διακριτή τυχαία μεταβλητή με σύνολο τιμών ή στήριγμα $S_X = \{x_1, x_2, ..., x_n, ...\}$. Η συνάρτηση $p_X : S_X \to [0,1]$, που ορίζεται από τη σχέση:

$$p_X(x_i) = P(X = x_i), \quad i = 1, 2, ...,$$

ονομάζεται **συνάρτηση πιθανότητας** (σ.π.) της τυχαίας μεταβλητής X.

Ορισμός 1.9: Συνάρτηση Πυκνότητας Πιθανότητας

Έστω X μια τυχαία μεταβλητή με τιμές στο σύνολο των πραγματικών αριθμών $\mathbb{R}$. Η X λέγεται **συνεχής**, αν υπάρχει μία μη αρνητική ολοκληρώσιμη πραγματική συνάρτηση $f_X(\cdot)$ ορισμένη στο σύνολο των πραγματικών αριθμών $\mathbb{R}$, τέτοια ώστε

$$P(X \in B) = \int_B f_X(x)dx, \quad B \subseteq \mathbb{R}.$$

Η συνάρτηση $f_X(\cdot)$ ονομάζεται **συνάρτηση πυκνότητας πιθανότητας** (σ.π.π.) της τυχαίας μεταβλητής X.

Άμεσες συνέπειες του ορισμού είναι ότι η σ.π.π. της τ.μ. X με σύνολο δυνατών τιμών $S_X \subseteq \mathbb{R}$ ικανοποιεί τις ιδιότητες:

$$f_X(x) \ge 0, \text{ για όλα τα } x \in \mathbb{R}$$

και

$$\int_{x \in S_X} f_X(x)dx = 1,$$

που είναι και οι ικανές συνθήκες που πρέπει να πληροί μια πραγματική συνάρτηση για να είναι συνάρτηση πυκνότητας πιθανότητας μιας συνεχούς τυχαίας μεταβλητής.

1.3.1 Χαρακτηριστικά κατανομής τυχαίας μεταβλητής

Η κατανομή πιθανότητας μίας τυχαίας μεταβλητής περιγράφεται από την α.σ.κ. ή από τη σ.π., αν είναι διακριτή, ή τη σ.π.π., αν είναι συνεχής. Σημαντική πληροφορία σχετικά με την κατανομή μίας τυχαίας μεταβλητής μπορούμε να λάβουμε από τις βασικές παραμέτρους της. Τα βασικά χαρακτηριστικά της κατανομής μίας τυχαίας μεταβλητής αποτελούνται από τα (i) μέτρα θέσης ή κεντρικής τάσης (μέση τιμή, διάμεσος, ποσοστιαία σημεία, κορυφή ή επικρατούσα τιμή), (ii) μέτρα διασποράς (διασπορά, τυπική απόκλιση, ενδοτεταρτημοριακό εύρος), (iii) μέτρα ασυμμετρίας και (iv) μέτρα κύρτωσης. Στη συνέχεια, παρατίθενται οι απαραίτητοι ορισμοί.

Ορισμός 1.10: Μέση Τιμή μίας Τυχαίας Μεταβλητής

Η **μαθηματική ελπίδα ή αναμενόμενη τιμή ή μέση τιμή** της τυχαίας μεταβλητής X, συμβολίζεται με $\mathrm{E}(X)$ ή μ. Αν η τ.μ. X είναι διακριτή με σύνολο τιμών $S_X = \{x_1, \ldots x_n, \ldots.\}$ και συνάρτηση πιθανότητας $p_X(\cdot)$ η μέση τιμή της τυχαίας μεταβλητής X ορίζεται από τη σχέση:

$$\mu = \mathrm{E}(X) = \sum_{x \in S_X} x\, p_X(x), \tag{1.1}$$

με την προϋπόθεση ότι η σειρά που εμφανίζεται συγκλίνει απόλυτα, δηλαδή ότι

$$\sum_{x \in S_X} |x| p_X(x) < \infty.$$

Αν η τ.μ. X είναι συνεχής με σύνολο τιμών S_X και συνάρτηση πυκνότητας πιθανότητας $f_X(\cdot)$, τότε η μέση τιμή της τυχαίας μεταβλητής X ορίζεται από τη σχέση:

$$\mu = \mathrm{E}(X) = \int_{x \in S_X} x f_X(x) dx, \tag{1.2}$$

με την προϋπόθεση ότι το ολοκλήρωμα συγκλίνει απόλυτα, δηλαδή ότι

$$\int_{x \in S_X} |x| f_X(x) dx < \infty.$$

Ορισμός 1.11: Διάμεσος

Έστω μία τ.μ. X με α.σ.κ. $F(x)$. **Διάμεσος** της X -ή της κατανομής της X- λέγεται κάθε αριθμός δ που είναι τέτοιος ώστε

$$F(\delta) = P(X \leq \delta) \geqslant \frac{1}{2}, \;\; F(\delta-) = P(X < \delta) \leqslant \frac{1}{2},$$

όπου $F(\delta-) = \lim_{x \to \delta-} F(x)$. Στην περίπτωση που η X είναι συνεχής τ.μ., τότε $F(\delta-) = F(\delta)$ και, άρα, ως διάμεσος της κατανομής της X ορίζεται η (μοναδική) λύση, έστω αυτή δ, της εξίσωσης $F(\delta) = 0.5$.

Ορισμός 1.12: Κορυφή

Έστω X τυχαία μεταβλητή με σ.π. $p_X(x)$ και δυνατές τιμές $x = x_1, x_2, \ldots$ ή με σ.π.π. $f_X(x) > 0, x \in A \subseteq \mathbb{R}$. Τότε:

- αν η X είναι διακριτή τ.μ., το σημείο M καλείται **κορυφή** (ή επικρατούσα τιμή) της κατανομής της X αν έχει θετική πιθανότητα (δηλ. $p_X(M) = P(X = M) > 0$) και ισχύει $p_X(M) \geqslant p_X(x)$ για $x = x_1, x_2, \ldots,$
- αν η X είναι συνεχής τ.μ. και η σ.π.π. αυτής έχει παράγωγο δεύτερης τάξης, το σημείο M καλείται

κορυφή (ή επικρατούσα τιμή) της κατανομής της X, αν $f'_X(M) = 0$ και $f''_X(M) < 0$, δηλαδή αν το M είναι σημείο μεγίστου της $f_X(x)$.

Ορισμός 1.13: Ποσοστιαίο Σημείο

Έστω μία τ.μ. X με α.σ.κ. $F(x)$ και έστω $p \in (0,1)$. Οποιοδήποτε σημείο x_p τέτοιο ώστε

$$P(X < x_p) \leqslant 1 - p \leqslant P(X \leq x_p),$$

ή

$$F(x_p^-) \leqslant 1 - p \leqslant F(x_p),$$

λέγεται p-**ποσοστιαίο σημείο** της X (ή της κατανομής της X). Όταν η τ.μ. X είναι συνεχής, τότε το p-ποσοστιαίο σημείο ($0 < p < 1$) είναι κάθε σημείο x_p που ικανοποιεί την εξίσωση $F(x_p) = 1 - p$. Όταν επιπλέον η $F(\cdot)$ είναι γνησίως αύξουσα στο στήριγμα της X, τότε $F^{-1}(1-p) = x_p$.

Από τον ορισμό του ποσοστιαίου σημείου, έχουμε ότι για $p = 0.5$, το σημείο $x_{0.5}$ είναι η διάμεσος δ της κατανομής της X, ενώ τα σημεία $x_{0.75}$ και $x_{0.25}$ καλούνται πρώτο τεταρτημόριο και τρίτο τεταρτημόριο της κατανομής της X, αντίστοιχα.

Ορισμός 1.14: Διακύμανση (Διασπορά) μίας Τυχαίας Μεταβλητής

Η **διακύμανση** της τυχαίας μεταβλητής X με πεπερασμένη μέση τιμή $\mu = \mathrm{E}(X) < \infty$, συμβολίζεται με $\mathrm{Var}(X)$ ή σ^2. Αν η τ.μ. X είναι διακριτή με σύνολο τιμών $S_X = \{x_1, \ldots x_n, \ldots.\}$ και συνάρτηση πιθανότητας $p_X(\cdot)$, η διακύμανση της τυχαίας μεταβλητής X ορίζεται από τη σχέση:

$$\sigma^2 = \mathrm{E}[(X-\mu)^2] = \sum_{x \in S_X} (x-\mu)^2 p_X(x), \tag{1.3}$$

με την προϋπόθεση ότι η σειρά που εμφανίζεται συγκλίνει απόλυτα. Αν η τ.μ. X είναι συνεχής με σύνολο τιμών $S_X \subseteq \mathbb{R}$ και συνάρτηση πυκνότητας πιθανότητας $f_X(\cdot)$, τότε η διακύμανση της τυχαίας μεταβλητής X ορίζεται από τη σχέση:

$$\sigma^2 = \mathrm{E}[(X-\mu)^2] = \int_{x \in S_X} (x-\mu)^2 f_X(x) dx, \tag{1.4}$$

με την προϋπόθεση ότι το ολοκλήρωμα συγκλίνει απόλυτα.

Ορισμός 1.15: Συντελεστής Λοξότητας

Έστω X μία τυχαία μεταβλητή με $\mu = \mathrm{E}(X)$ και $\sigma^2 = \mathrm{Var}(X)$. Τότε η ποσότητα

$$a_3 = \frac{\mathrm{E}[(X-\mu)^3]}{\sigma^3}$$

καλείται συντελεστής λοξότητας ή απλώς λοξότητα της κατανομής της τ.μ. X.

Ορισμός 1.16: Συντελεστής Κύρτωσης

Έστω X μία τυχαία μεταβλητή με $\mu = \mathrm{E}(X)$ και $\sigma^2 = \mathrm{Var}(X)$. Τότε η ποσότητα

$$a_4 = \frac{\mathrm{E}[(X-\mu)^4]}{\sigma^4}$$

καλείται συντελεστής κύρτωσης ή απλώς κύρτωση της κατανομής της τ.μ. X.

1.4 Βασικά πρότυπα κατανομών

Αν και στην προηγούμενη ενότητα δόθηκαν οι απαραίτητοι ορισμοί για τη γενική μελέτη της κατανομής μιας οποιασδήποτε τυχαίας μεταβλητής, υπάρχουν σημαντικές ειδικές διακριτές και συνεχείς κατανομές, οι οποίες μπορούν να χρησιμοποιηθούν στην πράξη για την περιγραφή (μοντελοποίηση) μιας ευρείας κλάσης στοχαστικών συστημάτων/πειραμάτων/φαινομένων. Αρχικά, παρουσιάζονται οι βασικές διακριτές κατανομές (διακριτά πιθανοτικά πρότυπα) και, στη συνέχεια, οι βασικές συνεχείς κατανομές (συνεχή πιθανοτικά πρότυπα).

1.4.1 Βασικές διακριτές κατανομές

Διωνυμική Κατανομή

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί τη **Διωνυμική κατανομή** με παραμέτρους n και p, $p \in (0,1)$, αν οι δυνατές της τιμές, x, είναι $x \in \{0,1,2,3,\ldots,n\}$, $n \in \mathbb{N}$ και η συνάρτηση πιθανότητας της X δίνεται από τη σχέση:

$$f_{n,p}(x) = \binom{n}{x} p^x (1-p)^{n-x}, \;\; x = 0,1,2,\ldots,n. \tag{1.5}$$

Στην περίπτωση αυτή, θα συμβολίζεται $X \sim B(n,p)$ με $\mathrm{E}(X) = np$, $\mathrm{Var}(X) = np(1-p)$. Επίσης, η αθροιστική συνάρτηση κατανομής της $B(n,p)$ θα συμβολίζεται $F_{n,p}(x)$.
Για $n = 1$, η κατανομή που προκύπτει είναι γνωστή ως **κατανομή Bernoulli** με συνάρτηση πιθανότητας

$$p_X(x) = p^x (1-p)^{1-x}, \;\; x = 0,1, \tag{1.6}$$

με $\mathrm{E}(X) = p$ και $\mathrm{Var}(X) = p(1-p)$.

Γεωμετρική Κατανομή

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί τη **Γεωμετρική κατανομή** με παράμετρο $p \in (0,1)$, αν οι δυνατές της τιμές, x, είναι $x \in \{1,2,3,\ldots\}$ και η συνάρτηση πιθανότητας της X δίνεται από τη σχέση:

$$f_{Geo}(x;p) = p(1-p)^{x-1}, \;\; x = 1,2,\ldots,. \tag{1.7}$$

Στην περίπτωση αυτή, θα συμβολίζεται $X \sim Geo(p)$ με

$$\mathrm{E}(X) = \frac{1}{p}, \; \mathrm{Var}(X) = \frac{1-p}{p^2}.$$

Επίσης, η αθροιστική συνάρτηση κατανομής της $Geo(p)$ θα συμβολίζεται $F_{Geo}(x;p)$.

Αρνητική Διωνυμική Κατανομή

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την **Αρνητική Διωνυμική κατανομή** με παραμέτρους $r \in \{1,2,3,\ldots.\}$ και $p \in (0,1)$ αν οι δυνατές της τιμές, x, είναι $\{r, r+1, r+2, \ldots\}$ και η συνάρτηση πιθανότητας της X δίνεται από τη σχέση:

$$f_{NB}(x;r,p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \;\; x = r, r+1, r+2, \ldots . \tag{1.8}$$

Στην περίπτωση αυτή, θα συμβολίζεται $X \sim NB(r,p)$ με

$$E(X) = \frac{r}{p}, \; Var(X) = \frac{r(1-p)}{p^2}.$$

Επίσης, η αθροιστική συνάρτηση κατανομής της $NB(r,p)$ θα συμβολίζεται $F_{NB}(x;r,p)$.

Υπεργεωμετρική Κατανομή

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την **Υπεργεωμετρική κατανομή** με παραμέτρους N_1, N_2 και n θετικούς ακεραίους, με $n \le N = N_1 + N_2$, αν οι δυνατές της τιμές, x, είναι θετικοί ακέραιοι αριθμοί τέτοιοι ώστε $\max\{0, n - N_2\} \le x \le \min\{N_1, n\}$ και η συνάρτηση πιθανότητας της X δίνεται από τη σχέση:

$$f_{Hg}(x;N_1,N_2,n) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N_1+N_2}{n}}, \ \max\{0,n-N_2\} \le x \le \min\{N_1,n\}. \tag{1.9}$$

Στην περίπτωση αυτή, θα συμβολίζεται $X \sim Hg(N_1,N_2,n)$ με

$$\mathrm{E}(X) = n\frac{N_1}{N_1+N_2}, \ \mathrm{Var}(X) = n\frac{N_1}{N_1+N_2}\cdot\frac{N_2}{N_1+N_2}\cdot\frac{N_1+N_2-n}{N_1+N_2-1}.$$

Επίσης, η αθροιστική συνάρτηση κατανομής της $Hg(N_1,N_2,n)$ θα συμβολίζεται $F_{Hg}(x;N_1,N_2,n)$.

Διακριτή Ομοιόμορφη Κατανομή

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί τη **Διακριτή Ομοιόμορφη κατανομή** με παραμέτρους α, β αν οι δυνατές της τιμές, x, είναι $\{\alpha, \alpha+1, \dots, \beta-1, \beta\}$ και η συνάρτηση πιθανότητας της X δίνεται από τη σχέση:

$$f_{DU}(x;\alpha,\beta) = \frac{1}{\beta-\alpha+1}, \ x = \alpha, \alpha+1, \dots, \beta-1, \beta. \tag{1.10}$$

Στην περίπτωση αυτή, θα συμβολίζεται $X \sim DU(\alpha,\beta)$ με

$$\mathrm{E}(X) = \frac{\alpha+\beta}{2}, \ \mathrm{Var}(X) = \frac{(\beta-\alpha+1)^2-1}{12}.$$

Επίσης, η αθροιστική συνάρτηση κατανομής της $DU(\alpha,\beta)$ θα συμβολίζεται $F_{DU}(x;\alpha,\beta)$.

Κατανομή Poisson

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την **κατανομή Poisson** με παράμετρο λ, $\lambda > 0$, αν οι δυνατές της τιμές, x, είναι $x \in \{0,1,2,3,\dots,\}$ και η συνάρτηση πιθανότητας της X δίνεται από τη σχέση:

$$f_{\mathscr{P}}(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \ x = 0,1,2\dots. \tag{1.11}$$

Στην περίπτωση αυτή, θα συμβολίζεται $X \sim \mathscr{P}(\lambda)$ με $\mathrm{E}(X) = \lambda$, $\mathrm{Var}(X) = \lambda$. Επίσης, η αθροιστική συνάρτηση κατανομής της $\mathscr{P}(\lambda)$ θα συμβολίζεται $F_{\mathscr{P}}(x;\lambda)$.

1.4.2 Βασικές συνεχείς κατανομές

Συνεχής Ομοιόμορφη Κατανομή

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί τη **συνεχή ομοιόμορφη κατανομή** στο διάστημα $[a,b]$ με a, $b \in \mathbb{R}$ και $a < b$, αν οι δυνατές της τιμές, x, ανήκουν στο διάστημα $[a,b]$ και η συνάρτηση πυκνότητας πιθανότητας της X δίνεται από τη σχέση:

$$f_{\mathscr{U}}(x;a,b) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & \text{αλλού.} \end{cases} \tag{1.12}$$

Στην περίπτωση αυτή γράφουμε ότι $X \sim \mathscr{U}(a,b)$ με

$$\mathsf{E}(X) = \frac{b-a}{2},\ \mathsf{Var}(X) = \frac{(b-a)^2}{12}.$$

Επίσης, η αθροιστική συνάρτηση κατανομής της $\mathscr{U}(a,b)$ θα συμβολίζεται $F_{\mathscr{U}}(x;a,b)$.

Κατανομή Βήτα

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί τη **Βήτα κατανομή** με παραμέτρους $a > 0$ και $b > 0$ αν οι δυνατές της τιμές, x, είναι $x \in (0,1)$ και η συνάρτηση πυκνότητας πιθανότητας της X δίνεται από τη σχέση:

$$f_{\mathsf{Beta}}(x;a,b) = \begin{cases} \frac{x^{a-1}(1-x)^{b-1}}{\mathsf{B}(a,b)}, & x \in (0,1), \\ 0, & \text{αλλού}, \end{cases} \tag{1.13}$$

όπου $\mathsf{B}(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$ είναι η Βήτα συνάρτηση.
Στην περίπτωση αυτή γράφουμε ότι $X \sim \mathscr{B}(a,b)$ με

$$\mathsf{E}(X) = \frac{a}{a+b},\ \mathsf{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Επίσης, η αθροιστική συνάρτηση κατανομής της $\mathscr{B}(a,b)$ θα συμβολίζεται $F_{\mathsf{Beta}}(x;a,b)$.

Εκθετική Κατανομή

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την **Εκθετική κατανομή** με παράμετρο $\lambda > 0$, αν οι δυνατές της τιμές, x, είναι $x \geq 0$ και η συνάρτηση πυκνότητας πιθανότητας της X δίνεται από τη σχέση:

$$f_{\mathscr{E}}(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & \text{αλλού}. \end{cases} \tag{1.14}$$

Στην περίπτωση αυτή γράφουμε ότι $X \sim \mathscr{E}(\lambda)$ με

$$\mathsf{E}(X) = \frac{1}{\lambda},\ \mathsf{Var}(X) = \frac{1}{\lambda^2}.$$

Επίσης, η αθροιστική συνάρτηση κατανομής της $\mathscr{E}(\lambda)$ θα συμβολίζεται $F_{\mathscr{E}}(x;\lambda)$.

Γάμμα Κατανομή

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την **κατανομή Γάμμα** με παραμέτρους $a > 0$ και $\lambda > 0$, αν οι δυνατές της τιμές, x, είναι $x \geq 0$ και η συνάρτηση πυκνότητας πιθανότητας της X δίνεται από τη σχέση:

$$f_{\mathscr{G}}(x;a,\lambda) = \begin{cases} \frac{\lambda^a x^{a-1} e^{-\lambda x}}{\Gamma(a)}, & x \geq 0, \\ 0, & \text{αλλού}, \end{cases} \tag{1.15}$$

όπου $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$ είναι η συνάρτηση Γάμμα. Στην περίπτωση αυτή, γράφουμε ότι $X \sim \mathscr{G}(a,\lambda)$ με

$$\mathsf{E}(X) = \frac{a}{\lambda},\ \mathsf{Var}(X) = \frac{a}{\lambda^2}.$$

Επίσης, η αθροιστική συνάρτηση κατανομής της $\mathscr{G}(a,\lambda)$ θα συμβολίζεται $F_{\mathscr{G}}(x;a,\lambda)$.

Παρατήρηση 1.1. Ένας εναλλακτικός ορισμός της κατανομής Γάμμα είναι ο παρακάτω

$$f_{\mathscr{G}}(x;a,b) = \begin{cases} \frac{x^{a-1}e^{-x/b}}{\Gamma(a)b^a}, & x \geq 0, \\ 0, & \text{αλλού}, \end{cases} \tag{1.16}$$

όπου η παράμετρος a είναι η παράμετρος σχήματος (shape parameter) και η παράμετρος b είναι η παράμετρος κλίμακας (scale parameter). Σε αυτή την περίπτωση, γράφουμε $X \sim \mathscr{G}(a,b)$ με

$$\mathsf{E}(X) = ab,\ \mathsf{Var}(X) = ab^2.$$

Κατανομή Weibull

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την **κατανομή Weibull** με παραμέτρους $a > 0$ και $\lambda > 0$, αν οι δυνατές της τιμές, x, είναι $x \geq 0$ και η συνάρτηση πυκνότητας πιθανότητας της X δίνεται από τη σχέση:

$$f_{\mathsf{WEI}}(x;a,\lambda) = \begin{cases} \lambda a x^{a-1}e^{-\lambda x^a}, & x \geq 0, \\ 0, & \text{αλλού}. \end{cases} \tag{1.17}$$

Στην περίπτωση αυτή γράφουμε ότι $X \sim WEI(a,\lambda)$, με

$$\mathsf{E}(X) = \lambda^{-1/a}\Gamma(1+1/a),\ \mathsf{Var}(X) = \lambda^{-2/a}\left[\Gamma(1+2/a) - (\Gamma(1+1/a))^2\right].$$

Επίσης, η αθροιστική συνάρτηση κατανομής της $WEI(a,\lambda)$ θα συμβολίζεται $F_{\mathsf{WEI}}(x;a,\lambda)$.

Κανονική Κατανομή

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την **κανονική κατανομή** με παραμέτρους μ και σ^2, με $\mu \in \mathbb{R}$ και $\sigma > 0$, αν η συνάρτηση πυκνότητας πιθανότητας της X δίνεται από τη σχέση:

$$f_{\mathscr{N}}(x;\mu;\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, x \in \mathbb{R}. \tag{1.18}$$

Στην περίπτωση αυτή, γράφουμε ότι $X \sim \mathscr{N}(\mu,\sigma^2)$ με

$$\mathsf{E}(X) = \mu,\ \mathsf{Var}(X) = \sigma^2.$$

Επίσης, η αθροιστική συνάρτηση κατανομής της $\mathscr{N}(\mu,\sigma^2)$ θα συμβολίζεται $F_{\mathscr{N}}(x;\mu,\sigma^2)$. Στην ειδική περίπτωση όπου $\mu = 0$ και $\sigma = 1$, η κατανομή που προκύπτει λέγεται **τυπική (ή τυποποιημένη) κανονική κατανομή**. Η αθροιστική συνάρτηση κατανομής και η συνάρτηση πυκνότητας πιθανότητας της τυπικής κανονικής κατανομής συμβολίζονται ως $\Phi(\cdot)$ και $\varphi(\cdot)$, αντίστοιχα.

Κατανομή Cauchy

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την **κατανομή Cauchy**, με παραμέτρους μ και σ^2, με $\mu \in \mathbb{R}$ και $\sigma > 0$, αν η συνάρτηση πυκνότητας πιθανότητας της X δίνεται από τη σχέση:

$$f_{\mathscr{C}}(x;\mu,\sigma^2) = \frac{1}{\pi\sigma}\cdot\frac{\sigma^2}{(x-\mu)^2+\sigma^2}, x \in \mathbb{R}. \tag{1.19}$$

Στην περίπτωση αυτή, γράφουμε ότι $X \sim \mathscr{C}(\mu,\sigma^2)$. Η μέση τιμή και η διακύμανση αυτής της κατανομής δεν ορίζονται. Θα συμβολίζουμε την αθροιστική συνάρτηση κατανομής της $\mathscr{C}(\mu,\sigma^2)$ ως $F_{\mathscr{C}}(x;\mu,\sigma^2)$.

Κατανομή χι-τετράγωνο

Έστω $Z_1, Z_2, \ldots, Z_n$, $n \geqslant 1$, ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με κατανομή $\mathscr{N}(0,1)$. Η κατανομή

της τυχαίας μεταβλητής $X = \sum_{i=1}^{n} Z_i^2$ ονομάζεται **κατανομή χι-τετράγωνο** με n βαθμούς ελευθερίας. Η συνάρτηση πυκνότητας πιθανότητας της X δίνεται από τη σχέση:

$$f_{\chi_n^2}(x) = \frac{(1/2)^{n/2}}{\Gamma(n/2)} x^{(n/2)-1} e^{-x/2}, x \in (0, \infty). \tag{1.20}$$

Στην περίπτωση αυτή, γράφουμε ότι $X \sim \chi_n^2$. Η κατανομή αυτή συμπίπτει με την κατανομή $\mathcal{G}(n/2, 1/2)$, ενώ $\mathrm{E}(X) = n$, $\mathrm{Var}(X) = 2n$. Τέλος, θα συμβολίζουμε την αθροιστική συνάρτηση κατανομής της χ_n^2 ως $F_{\chi_n^2}(x)$.

Κατανομή t

Έστω Z και Y ανεξάρτητες τυχαίες μεταβλητές, με $Z \sim \mathcal{N}(0,1)$ και $Y \sim \chi_n^2$. Η κατανομή της τυχαίας μεταβλητής $X = \frac{Z}{\sqrt{Y/n}}$ ονομάζεται **κατανομή** t με n βαθμούς ελευθερίας. Η συνάρτηση πυκνότητας πιθανότητας της X δίνεται από τη σχέση:

$$f_{t_n}(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{(1 + x^2/n)^{(n+1)/2}}, \; x \in \mathbb{R}. \tag{1.21}$$

Στην περίπτωση αυτή γράφουμε ότι $X \sim t_n$, ενώ $\mathrm{E}(X) = 0$, όταν $n > 1$, $\mathrm{Var}(X) = n/(n-2)$, όταν $n > 2$. Αν $n = 1$, η κατανομή t_1 συμπίπτει με την κατανομή $\mathcal{C}(0,1)$. Τέλος, θα συμβολίζουμε την αθροιστική συνάρτηση κατανομής της t_n ως $F_{t_n}(x)$.

Κατανομή $\mathcal{F}$

Έστω X_1 και X_2 ανεξάρτητες τυχαίες μεταβλητές, με $X_1 \sim \chi_{n_1}^2$ και $X_2 \sim \chi_{n_2}^2$. Η κατανομή της τυχαίας μεταβλητής $X = \frac{X_1/n_1}{X_2/n_2}$ ονομάζεται **κατανομή** F (ή κατανομή Snedecor-F προς τιμήν των George Wadel Snedecor (1881-1974) και Sir Ronald Aylmer Fisher (1890-1962)) με n_1 και n_2 βαθμούς ελευθερίας. Η συνάρτηση πυκνότητας πιθανότητας της X δίνεται από τη σχέση:

$$f_{F_{n_1,n_2}}(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)(n_1/n_2)^{(n_1/2)}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \cdot \frac{x^{(n_1/2)-1}}{(1 + n_1 x/n_2)^{(n_1+n_2)/2}}, \; x \in (0, \infty). \tag{1.22}$$

Στην περίπτωση αυτή, γράφουμε ότι $X \sim F_{n_1,n_2}$, ενώ

$$\mathrm{E}(X) = n_2/(n_2-2), \text{ όταν } n_2 > 2, \; \mathrm{Var}(X) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-1)^2(n_2-4)}, \text{ όταν } n_2 > 4.$$

Τέλος, θα συμβολίζουμε την αθροιστική συνάρτηση κατανομής της F_{n_1,n_2} ως $F_{F_{n_1,n_2}}(x)$.

1.4.3 Αναπαραγωγικές ιδιότητες

Κλείνοντας την παρούσα ενότητα, παραθέτουμε κάποιες ιδιότητες αθροισμάτων για κ το πλήθος ανεξάρτητες τυχαίες μεταβλητές. Οι ιδιότητες αυτές είναι γνωστές και ως **αναπαραγωγικές ιδιότητες**.

Έστω $X_1, X_2, \ldots, X_\kappa$ ανεξάρτητες τυχαίες μεταβλητές (το $\kappa \in \{1, 2, \ldots\}$).

- Αν $X_i \sim B(n_i, p)$, $i = 1, 2, \ldots, \kappa$, τότε

$$\sum_{i=1}^{\kappa} X_i \sim B\left(\sum_{i=1}^{\kappa} n_i, p\right).$$

  Στην ειδική περίπτωση που $X_i \sim B(n, p)$, τότε $\sum_{i=1}^{\kappa} X_i \sim B\left(\kappa \cdot n, p\right)$.

- Αν $X_i \sim \mathscr{P}(\lambda_i)$, $i = 1,2,\dots,\kappa$, τότε
$$\sum_{i=1}^{\kappa} X_i \sim \mathscr{P}\left(\sum_{i=1}^{\kappa} \lambda_i\right).$$
Στην ειδική περίπτωση που $X_i \sim \mathscr{P}(\lambda)$, τότε $\sum_{i=1}^{\kappa} X_i \sim \mathscr{P}(\kappa \cdot \lambda)$.
- Αν $X_i \sim Geo(p)$, $i = 1,2,\dots,\kappa$, τότε
$$\sum_{i=1}^{\kappa} X_i \sim NB\left(\kappa, p\right).$$
- Αν $X_i \sim NB(r_i, p)$, $i = 1,2,\dots,\kappa$, τότε
$$\sum_{i=1}^{\kappa} X_i \sim NB\left(\sum_{i=1}^{\kappa} r_i, p\right).$$
Στην ειδική περίπτωση που $X_i \sim NB(r,p)$, τότε $\sum_{i=1}^{\kappa} X_i \sim NB\left(\kappa \cdot r, p\right)$.
- Αν $X_i \sim \mathscr{N}(\mu_i, \sigma_i^2)$, $i = 1,2,\dots,\kappa$, τότε
$$\sum_{i=1}^{\kappa} X_i \sim \mathscr{N}\left(\sum_{i=1}^{\kappa} \mu_i, \sum_{i=1}^{\kappa} \sigma_i^2\right).$$
Στην ειδική περίπτωση που $X_i \sim \mathscr{N}(\mu, \sigma^2)$, τότε $\sum_{i=1}^{\kappa} X_i \sim \mathscr{N}\left(\kappa \cdot \mu, \kappa \cdot \sigma^2\right)$.
- Αν $X_i \sim \mathscr{G}(a_i, \lambda)$, $i = 1,2,\dots,\kappa$, τότε
$$\sum_{i=1}^{\kappa} X_i \sim \mathscr{G}\left(\sum_{i=1}^{\kappa} a_i, \lambda\right).$$
Στην ειδική περίπτωση που $X_i \sim \mathscr{G}(a, \lambda)$, τότε $\sum_{i=1}^{\kappa} X_i \sim \mathscr{G}(\kappa \cdot a, \lambda)$.
- Αν $X_i \sim \mathscr{E}(\lambda)$, $i = 1,2,\dots,\kappa$, τότε
$$\sum_{i=1}^{\kappa} X_i \sim \mathscr{G}(\kappa, \lambda).$$
- Αν $X_i \sim \chi^2_{\nu_i}$, $i = 1,2,\dots,\kappa$, τότε
$$\sum_{i=1}^{\kappa} X_i \sim \chi^2_{\nu},$$
όπου $\nu = \sum_{i=1}^{\kappa} \nu_i$.

1.5 Πολυδιάστατες τυχαίες μεταβλητές

Στην ενότητα αυτή παρουσιάζονται οι βασικές έννοιες για την -ταυτόχρονη- μελέτη δύο ή περισσότερων τυχαίων μεταβλητών. Στην περίπτωση αυτή, τα στοιχειώδη ενδεχόμενα $\omega \in \Omega$ αντιστοιχίζονται σε ένα (γενικά) n-διάστατο διάνυσμα $\boldsymbol{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$, $n \geqslant 2$, γενικεύοντας έτσι την περίπτωση των μονοδιάστατων τυχαίων μεταβλητών και των αντίστοιχων κατανομών. Έτσι, ο ορισμός της αντίστοιχης n-διάστατης τυχαίας μεταβλητής είναι

$$\boldsymbol{X} : \Omega \to \mathbb{R}^n_{\boldsymbol{X}} \equiv \boldsymbol{X}(\Omega), \text{ όπου } \mathbb{R}^n_{\boldsymbol{X}} = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{X}(\omega) = \boldsymbol{x},\ \omega \in \Omega\}.$$

Έχοντας ορίσει την έννοια της n-διάστατης τυχαίας μεταβλητής, μπορούμε να προχωρήσουμε με τον ορισμό της από κοινού (ή κοινής) αθροιστικής συνάρτησης κατανομής, τον ορισμό της από κοινού

συνάρτησης πιθανότητας και τον ορισμό της από κοινού συνάρτησης πυκνότητας πιθανότητας. Σημειώνεται, επίσης, ότι, αν οι $X_1, X_2, \dots, X_n$ είναι όλες διακριτές τυχαίες μεταβλητές, τότε και η αντίστοιχη n-διάστατη τυχαία μεταβλητή $\boldsymbol{X}$ θα αναφέρεται ως διακριτή. Αντίστοιχα, αν οι $X_1, X_2, \dots, X_n$ είναι όλες συνεχείς τυχαίες μεταβλητές, τότε και η αντίστοιχη n-διάστατη τυχαία μεταβλητή $\boldsymbol{X}$ θα αναφέρεται ως συνεχής.

Ορισμός 1.17: Από Κοινού Αθροιστική Συνάρτηση Κατανομής

Έστω $(\Omega, \mathscr{A}, P)$ ένας χώρος πιθανότητας και έστω n το πλήθος τυχαίες μεταβλητές $X_1, X_2, \dots, X_n$ που ορίζονται στον Ω. Θα λέμε ότι οι n τ.μ. είναι από κοινού κατανεμημένες και το διάνυσμα $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$ είναι ένα τυχαίο διάνυσμα. Η συνάρτηση F, η οποία ορίζεται από τη σχέση:

$$F(x_1, x_2, \dots, x_n) = P(X_1 \leqslant x_1, X_2 \leqslant x_2, \dots, X_n \leqslant x_n)$$

για κάθε $\boldsymbol{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ ονομάζεται αθροιστική συνάρτηση κατανομής της n-διάστατης τ.μ. X ή από κοινού αθροιστική συνάρτηση κατανομής των τ.μ. $X_1, X_2, \dots, X_n$. Στην ειδική περίπτωση $n = 2$, για τις τυχαίες μεταβλητές (X, Y), η από κοινού α.σ.κ. ορίζεται από τη σχέση:

$$F(x, y) = P(X \leqslant x, Y \leqslant y), \text{ για κάθε } (x, y) \in \mathbb{R}^2.$$

Ορισμός 1.18: Από Κοινού Συνάρτηση Πιθανότητας

Έστω $(\Omega, \mathscr{A}, P)$ ένας χώρος πιθανότητας και έστω n διακριτές τυχαίες μεταβλητές $X_1, X_2, \dots, X_n$ που ορίζονται στον Ω. Η πολυδιάστατη κατανομή των από κοινού κατανεμημένων διακριτών τ.μ. $X_1, X_2, \dots, X_n$ χαρακτηρίζεται από την κοινή (ή από κοινού) συνάρτηση πιθανότητας (κ.σ.π.). Αυτή είναι μια (πολυμεταβλητή) συνάρτηση $f(x_1, x_2, \dots, x_n)$ με πεδίο ορισμού τον διανυσματικό χώρο $\mathbb{R}^n$, πεδίο τιμών ένα υποσύνολο του διαστήματος $[0,1]$ και ορίζεται από τη σχέση:

$$f(x_1, x_2, \dots, x_n) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n), \quad (x_1, x_2, \dots, x_n) \in \mathbb{R}^n.$$

Στην ειδική περίπτωση $n = 2$, η από κοινού συνάρτηση πιθανότητας των διακριτών τ.μ. X και Y είναι μια συνάρτηση $f(x, y)$ με πεδίο ορισμού το σύνολο $\mathbb{R}^2$ και πεδίο τιμών ένα υποσύνολο του διαστήματος $[0,1]$, τέτοια ώστε:

$$f(x, y) = P(X = x, Y = y), \quad (x, y) \in \mathbb{R}^2.$$

Επιπλέον, η $f(x, y)$ ικανοποιεί τις παρακάτω δύο ιδιότητες:

- $f(x, y) \geq 0, \quad (x, y) \in \mathbb{R}^2$,
- $\sum_x \sum_y f(x, y) = 1$, για τα ζεύγη (x, y) με $f(x, y) > 0$.

Δεν είναι δύσκολο να διαπιστώσουμε ότι ανάλογες ιδιότητες πρέπει να ικανοποιεί και η από κοινού συνάρτηση πιθανότητας μιας (γενικά) n-διάστατης διακριτής τυχαίας μεταβλητής.

Έστω τώρα το τυχαίο διάνυσμα (X, Y). Τότε, αν οι τ.μ. X και Y είναι διακριτές, οι **Περιθώριες Συναρτήσεις Πιθανότητας** του τυχαίου διανύσματος (X, Y) είναι οι σ.π. των X και Y, οι οποίες δίνονται από τις σχέσεις:

$$f_X(x) = \sum_y f(x, y), \; x \in \mathbb{R}, \qquad f_Y(y) = \sum_x f(x, y), \; y \in \mathbb{R}.$$

Επίσης, η **Δεσμευμένη Συνάρτηση Πιθανότητας** της διακριτής τ.μ. X, όταν $Y = y$ συμβολίζεται με $f_{X|Y=y}(x)$ και υπολογίζεται από τη σχέση:

$$f_{X|Y=y}(x) = \frac{f(x, y)}{f_Y(y)}, \quad x \in \mathbb{R},$$

για κάθε $y \in \mathbb{R}$ τέτοιο, ώστε $f_Y(y) > 0$. Ομοίως, η δεσμευμένη σ.π. της διακριτής τ.μ. Y, όταν $X = x$, συμβολίζεται ως $f_{Y|X=x}(y)$ και υπολογίζεται από τη σχέση:

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)}, \quad y \in \mathbb{R},$$

για κάθε $x \in \mathbb{R}$ τέτοιο, ώστε $f_X(x) > 0$.

Ορισμός 1.19: Από Κοινού Συνάρτηση Πυκνότητας Πιθανότητας

Η πολυδιάστατη κατανομή των από κοινού κατανεμημένων συνεχών τυχαίων μεταβλητών $X_1, X_2, \ldots, X_n$ χαρακτηρίζεται από την κοινή (ή από κοινού) συνάρτηση πυκνότητας πιθανότητας. Αυτή είναι μια (πολυμεταβλητή) συνάρτηση $f(\boldsymbol{x}) = f(x_1, x_2, \ldots, x_n)$ με πεδίο ορισμού τον διανυσματικό χώρο $\mathbb{R}^n$ και πεδίο τιμών ένα υποσύνολο του $\mathbb{R}$, για την οποία ισχύει ότι:

$$f(\boldsymbol{x}) \geq 0, \quad \boldsymbol{x} \in \mathbb{R}^n,$$

με

$$\int_{x:x\in\mathbb{R}^n} f(\boldsymbol{x})d\boldsymbol{x} = 1,$$

ενώ για κάθε υποσύνολο $\boldsymbol{E}$ του χώρου $\mathbb{R}^n$ ισχύει ότι:

$$P(\boldsymbol{X} \in \boldsymbol{E}) = \int_{x:x\in\boldsymbol{E}} f(\boldsymbol{x})d\boldsymbol{x}.$$

Παρακάτω δίνονται οι ιδιότητες της από κοινού συνάρτησης πυκνότητας πιθανότητας $f(x,y)$ για ένα τυχαίο διάνυσμα (X,Y) δύο συνεχών τυχαίων μεταβλητών. Συγκεκριμένα, η από κοινού συνάρτηση πυκνότητας πιθανότητας $f(x,y)$ είναι $f(x,y) \geq 0, \quad (x,y) \in \mathbb{R}^2$, με

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1,$$

ενώ για κάθε υποσύνολο $\boldsymbol{E}$ του χώρου $\mathbb{R}^2$ ισχύει ότι:

$$P((X,Y) \in \boldsymbol{E}) = \int\int_{(x,y):(x,y)\in\boldsymbol{E}} f(x,y)dxdy.$$

Ειδικότερα, για a, b, c, d στο $\mathbb{R}$, ισχύει ότι:

$$P(a < X < b, c < Y < d) = \int_c^d \int_a^b f(x,y)dxdy.$$

Έστω τώρα το τυχαίο διάνυσμα (X,Y). Τότε, αν οι τ.μ. X και Y είναι συνεχείς, οι **Περιθώριες Συναρτήσεις Πυκνότητας Πιθανότητας** του τυχαίου διανύσματος (X,Y) είναι οι σ.π.π. των X και Y, οι οποίες δίνονται από τις σχέσεις:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy, \; x \in \mathbb{R}, \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx, \; y \in \mathbb{R}.$$

Με παρόμοιο τρόπο με τη διακριτή περίπτωση, η **Δεσμευμένη Συνάρτηση Πυκνότητας Πιθανότητας** της συνεχούς τ.μ. X, όταν $Y = y$, συμβολίζεται με $f_{X|Y=y}(x)$ και υπολογίζεται από τη σχέση:

$$f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)}, \quad x \in \mathbb{R},$$

για κάθε $y \in \mathbb{R}$ τέτοιο, ώστε $f_Y(y) > 0$.
Ομοίως, η δεσμευμένη σ.π.π. της συνεχούς τ.μ. Y, όταν $X = x$, συμβολίζεται με $f_{Y|X=x}(y)$ και υπολογίζεται από τη σχέση:

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)}, \quad y \in \mathbb{R},$$

για κάθε $x \in \mathbb{R}$ τέτοιο, ώστε $f_X(x) > 0$.

1.5.1 Ανεξαρτησία τυχαίων μεταβλητών

Έστω δύο από κοινού κατανεμημένες τ.μ. (X, Y), διακριτές ή συνεχείς. Τότε αυτές είναι **ανεξάρτητες**, αν αληθεύει μία από τις παρακάτω ισοδύναμες προτάσεις:

- Για κάθε δύο, μετρήσιμα, υποσύνολα E_1 και E_2 των πραγματικών αριθμών ισχύει:

$$P(X \in E_1, Y \in E_2) = P(X \in E_1) \cdot P(Y \in E_2).$$

- Για κάθε $(x,y) \in \mathbb{R}^2$ ισχύει:

$$f(x,y) = f_X(x) \cdot f_Y(y),$$

όπου οι συναρτήσεις $f(x,y)$ και $f_X(x)$, $f_Y(y)$ συμβολίζουν την από κοινού συνάρτηση πιθανότητας και τις περιθώριες συναρτήσεις πιθανότητας, αν πρόκειται για διακριτές τ.μ., ή την από κοινού συνάρτηση πυκνότητας πιθανότητας και τις περιθώριες συναρτήσεις πυκνότητας πιθανότητας, αν πρόκειται για συνεχείς, αντίστοιχα.

- Για κάθε $(x,y) \in \mathbb{R}^2$ τέτοιο, ώστε $f_Y(y) > 0$, ισχύει:

$$f_{X|Y=y}(x) = f_X(x),$$

όπου οι συναρτήσεις $f_{X|Y=y}(x)$ και $f_X(x)$ συμβολίζουν τη δεσμευμένη σ.π. και την περιθώρια σ.π., αν πρόκειται για διακριτές τ.μ. ή τη δεσμευμένη σ.π.π. και την περιθώρια σ.π.π., αν πρόκειται για συνεχείς, αντίστοιχα.

Στη συνέχεια, δίνεται ένα μέτρο για την από κοινού διακύμανση των τυχαίων μεταβλητών X, Y γύρω από τις αντίστοιχες μέσες τιμές τους, το οποίο εκφράζει τη συμμεταβλητότητα των τ.μ. X, Y.

Ορισμός 1.20: Συνδιακύμανση (Συνδιασπορά)

Η συνδιακύμανση ή συνδιασπορά δύο από κοινού κατανεμημένων τ.μ. (X, Y) συμβολίζεται με $\text{Cov}(X, Y)$ και ορίζεται από τη σχέση:

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

Οι βασικές ιδιότητες της συνδιακύμανσης είναι οι ακόλουθες:

- Αν X, Y είναι τυχαίες μεταβλητές με συνδιακύμανση $\text{Cov}(X, Y)$ και $\alpha_i \in \mathbb{R}$, $i = 1,2,3,4$, τότε

$$\text{Cov}(\alpha_1 X + \alpha_2, \alpha_3 Y + \alpha_4) = \alpha_1 \alpha_3 \text{Cov}(X, Y).$$

- Αν X, Y είναι τυχαίες μεταβλητές με συνδιακύμανση $\text{Cov}(X, Y)$, τότε

$$\text{Cov}(X, Y) = \text{E}(XY) - \text{E}(X)\text{E}(Y),$$

υπό την προϋπόθεση ότι υπάρχουν οι παραπάνω μέσες τιμές.

- Αν X, Y και Z είναι τυχαίες μεταβλητές και υπάρχουν οι συνδιακυμάνσεις $\text{Cov}(X, Y)$, $\text{Cov}(X, Z)$, τότε

$$\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z).$$

Ένα μειονέκτημα του μέτρου της συνδιακύμανσης δύο τ.μ. X, Y, το οποίο, εν γένει, μπορεί να μας υποδείξει κατά πόσο δύο τυχαίες μεταβλητές συσχετίζονται μεταξύ τους, είναι το γεγονός ότι εξαρτάται από τις μονάδες μέτρησης των μεταβλητών. Ένα μέτρο του βαθμού συσχέτισης δύο τ.μ. X, Y, το οποίο είναι απαλλαγμένο από μονάδες μέτρησης, είναι ο συντελεστής συσχέτισης ρ.

Ορισμός 1.21: Συντελεστής Συσχέτισης

Ο συντελεστής συσχέτισης ρ δύο από κοινού κατανεμημένων τ.μ. (X, Y) ορίζεται από τη σχέση:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

όπου σ_X, σ_Y είναι οι τυπικές αποκλίσεις των X, Y, αντίστοιχα.

Παρακάτω δίνονται κάποιες βασικές ιδιότητες.

- Αν οι τ.μ. X, Y είναι ανεξάρτητες, τότε $\text{Cov}(X, Y) = 0$.
- Αν a, b είναι δύο πραγματικοί αριθμοί, τότε $\text{Var}(aX + b) = a^2\text{Var}(X)$.
- Αν a, b είναι δύο πραγματικοί αριθμοί και X, Y είναι δύο από κοινού κατανεμημένες τ.μ., τότε

$$\text{Var}(aX \pm bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) \pm 2ab\text{Cov}(X, Y).$$

- Αν οι τ.μ. X, Y είναι ανεξάρτητες, τότε $\rho = 0$.
- Η απόλυτη τιμή του συντελεστή συσχέτισης ισούται με 1, δηλαδή $|\rho| = 1$, αν και μόνο αν οι τ.μ. X, Y συνδέονται με τέλεια γραμμική σχέση, δηλαδή αν και μόνο αν $Y = a \pm bX$.
- Αν $\rho = 0$, τότε οι τ.μ. X, Y λέγονται (γραμμικά) ασυσχέτιστες. Θα πρέπει να σημειωθεί πως δύο ασυσχέτιστες τυχαίες μεταβλητές δεν είναι απαραίτητα ανεξάρτητες, απλά ο βαθμός της γραμμικής σχέσης τους είναι 0. Όμως, αν δύο τυχαίες μεταβλητές είναι ανεξάρτητες, τότε είναι και ασυσχέτιστες.

1.6 Κατανομές διατεταγμένων τυχαίων μεταβλητών

Έστω μια n-διάστατη τυχαία μεταβλητή $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$, ορισμένη σε έναν δειγματικό χώρο Ω. Έστω, επίσης, ότι για κάθε στοιχειώδες ενδεχόμενο $\omega \in \Omega$, οι τιμές $x_i = X_i(\omega) \in \mathbb{R}$, $i = 1, 2, \dots, n$, μπορούν να διαταχθούν κατά αύξουσα τάξη μεγέθους, δηλαδή για κάθε $\omega \in \Omega$ υπάρχει μετάθεση $(i_1, i_2, \dots, i_n)$ των n δεικτών $\{1, 2, \dots, n\}$, τέτοια ώστε $x_{i_1} \leqslant x_{i_2} \leqslant \dots \leqslant x_{i_n}$.

Η **k-τάξης διατεταγμένη τυχαία μεταβλητή** συμβολίζεται με $X_{(k)}$ και ορίζεται ως $X_{(k)}(\omega) = x_{(k)}$, για $\omega \in \Omega$ και $x_{(k)} = x_{i_k}$, $k = 1, 2, \dots, n$. Η τυχαία μεταβλητή $X_{(k)}$ είναι συνάρτηση των τυχαίων μεταβλητών $X_1, X_2, \dots, X_n$. Για $k = 1$, η $X_{(1)} = \min\{X_1, X_2, \dots, X_n\}$, για $k = n$, $X_{(n)} = \max\{X_1, X_2, \dots, X_n\}$, ενώ γενικά $X_{(1)} \leqslant X_{(2)} \leqslant \dots \leqslant X_{(n)}$.

Ας θεωρήσουμε την ειδική περίπτωση όπου οι $X_1, X_2, \dots, X_n$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με α.σ.κ. $F(x) = P(X_i \leqslant x)$, $x \in \mathbb{R}$, $i = 1, 2, \dots, n$. Στη συνέχεια, υπό αυτήν την υπόθεση, δίνουμε βασικά αποτελέσματα για τις κατανομές των παραπάνω διατεταγμένων τυχαίων μεταβλητών.

Αποδεικνύεται ότι η αθροιστική συνάρτηση κατανομής της $X_{(k)}$, $1 \leqslant k \leqslant n$, δίνεται από τη σχέση:

$$F_{X_{(k)}}(x) = P(X_{(k)} \leqslant x) = \sum_{l=k}^{n} \binom{n}{l} [F(x)]^l [1 - F(x)]^{(n-l)}, \quad x \in \mathbb{R}. \tag{1.23}$$

Στην ειδική περίπτωση της $X_{(1)} = \min\{X_1, X_2, \dots, X_n\}$ η αθροιστική συνάρτηση κατανομής της δίνεται από τη σχέση:

$$F_{X_{(1)}}(x) = P(X_{(1)} \leqslant x) = 1 - [1 - F(x)]^n, \quad x \in \mathbb{R}, \tag{1.24}$$

ενώ η αθροιστική συνάρτηση κατανομής της $X_{(n)} = \max\{X_1, X_2, \dots, X_n\}$ δίνεται από τη σχέση:

$$F_{X_{(n)}}(x) = P(X_{(n)} \leqslant x) = [F(x)]^n, \quad x \in \mathbb{R}. \tag{1.25}$$

Αφού είναι γνωστή η αθροιστική συνάρτηση κατανομή της $X_{(k)}$, μπορούμε να προχωρήσουμε με τη συνάρτηση πυκνότητας πιθανότητας αυτής -στη συνεχή περίπτωση. Ειδικότερα, στην ειδική περίπτωση όπου οι $X_1, X_2, \dots, X_n$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με α.σ.κ. $F(x)$, $x \in \mathbb{R}$ και σ.π.π. $f_X(x) = F'(x)$, $x \in \mathbb{R}$, η συνάρτηση πυκνότητας της $X_{(k)}$, $1 \leqslant k \leqslant n$, δίνεται από τη σχέση:

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{(k-1)}[1-F(x)]^{(n-k)} f_X(x), \quad x \in \mathbb{R}, \tag{1.26}$$

Στην ειδική περίπτωση της $X_{(1)} = \min\{X_1, X_2, \dots, X_n\}$ η συνάρτηση πυκνότητας πιθανότητας δίνεται από τη σχέση:

$$f_{X_{(1)}}(x) = n[1-F(x)]^{n-1} f_X(x), \quad x \in \mathbb{R}, \tag{1.27}$$

ενώ η συνάρτηση πυκνότητας πιθανότητας της $X_{(n)} = \max\{X_1, X_2, \dots, X_n\}$ δίνεται από τη σχέση:

$$f_{X_{(n)}}(x) = n[F(x)]^{n-1} f_X(x), \quad x \in \mathbb{R}. \tag{1.28}$$

Με ανάλογο τρόπο, διατυπώνουμε τα αντίστοιχα αποτελέσματα στη διακριτή περίπτωση και συγκεκριμένα για τη συνάρτηση πιθανότητας της $X_{(k)}$. Έστω $X_1, X_2, \dots, X_n$ ανεξάρτητες και ισόνομες διακριτές τυχαίες μεταβλητές με α.σ.κ. $F(x) = P(X_i \leqslant x)$, $x \in \mathbb{R}$ και συνάρτηση πιθανότητας $p_X(x_j) = P(X_i = x_j)$, $j = 0,1,2,\dots$, $i = 1,2,\dots,n$. Έστω, επίσης, ότι μπορούμε να διατάξουμε τις δυνατές τιμές $x_0, x_1, x_2, \dots$ ως $x_0 < x_1 < x_2 < \dots$ Τότε, οι συναρτήσεις πιθανότητας της $X_{(1)}$ και της $X_{(n)}$ δίνονται από τις σχέσεις:

$$\begin{aligned} f_{X_{(1)}}(x_j) = P(X_{(1)} = x_j) &= 1 - [1 - F(x_0)]^n \\ &= [1 - F(x_{j-1})]^n - [1 - F(x_j)]^n, \quad j = 1,2,\dots \end{aligned} \tag{1.29}$$

και

$$\begin{aligned} f_{X_{(n)}}(x_j) = P(X_{(n)} = x_j) &= [F(x_0)]^n \\ &= [F(x_j)]^n - [F(x_{j-1})]^n, \quad j = 1,2,\dots \end{aligned} \tag{1.30}$$

αντίστοιχα.

Κλείνουμε την παρούσα υποενότητα δίνοντας αποτελέσματα σχετικά με την από κοινού συνάρτηση κατανομής και την από κοινού συνάρτηση πυκνότητας πιθανότητας ενός ζεύγους διατεταγμένων τυχαίων μεταβλητών $(X_{(k)}, X_{(r)})$, $1 \leqslant k < r \leqslant n$. Έστω $X_1, X_2, \dots, X_n$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με α.σ.κ. $F(x) = P(X_i \leqslant x)$, $x \in \mathbb{R}$, $i = 1, 2, \dots, n$. Τότε, η από κοινού συνάρτηση κατανομής $F_{X_{(k)},X_{(r)}}(x,y) = P(X_{(k)} \leqslant x, X_{(r)} \leqslant y)$ της διδιάστατης τυχαίας μεταβλητής $(X_{(k)}, X_{(r)})$, $1 \leqslant k < r \leqslant n$, δίνεται από τη σχέση:

$$\begin{aligned} F_{X_{(k)},X_{(r)}}(x,y) &= \sum_{l=r}^{n}\sum_{j=k}^{l} \frac{n!}{j!(l-j)!(n-l)!}[F(x)]^j[F(y)-F(x)]^{l-j}[1-F(y)]^{n-l}, \; x < y, \\ &= \sum_{l=r}^{n}\binom{n}{l}[F(y)]^l[1-F(y)]^{n-l}, \; x \geqslant y. \end{aligned} \tag{1.31}$$

Ειδικότερα, η από κοινού κατανομή των $(X_{(1)}, X_{(n)})$ δίνεται από τη σχέση:

$$\begin{aligned} F_{X_{(1)},X_{(n)}}(x,y) &= [F(y)]^n - [F(y)-F(x)]^n, \; x < y, \\ &= [F(y)]^n, \; x \geqslant y. \end{aligned} \tag{1.32}$$

Αν θέλουμε τώρα να υπολογίσουμε την από κοινού συνάρτηση πιθανότητας της διδιάστατης τυχαίας μεταβλητής $(X_{(k)}, X_{(r)})$ (διακριτή περίπτωση), δεν έχουμε παρά να χρησιμοποιήσουμε την παρακάτω σχέση:

$$\begin{aligned} f_{(X_{(k)},X_{(r)})}(x_{l_1}, x_{l_2}) &= P(X_{(k)} = x_{l_1}, X_{(r)} = x_{l_2}) \\ &= F(x_{l_1}, x_{l_2}) - F(x_{l_1-1}, x_{l_2}) - F(x_{l_1}, x_{l_2-1}) + F(x_{l_1-1}, x_{l_2-1}) \end{aligned}$$

με $l_1, l_2 = 0, 1, 2, \ldots$ και $x_{-1} = y_{-1} = \infty$.

Στη συνεχή περίπτωση, έστω $X_1, X_2, \ldots, X_n$ ανεξάρτητες και ισόνομες συνεχείς τυχαίες μεταβλητές με συνάρτηση κατανομής $F(x) = P(X_i \leqslant x)$, $x \in \mathbb{R}$, $i = 1, 2, \ldots, n$ και συνάρτηση πυκνότητας $f_X(x) = F'(x)$, $x \in \mathbb{R}$. Τότε η από κοινού συνάρτηση πυκνότητας πιθανότητας των $(X_{(k)}, X_{(r)})$, $1 \leqslant k < r \leqslant n$, δίνεται από τη σχέση:

$$\begin{aligned} f_{(X_{(k)},X_{(r)})}(x, y) &= \frac{n![F(x)]^{k-1}[F(y) - F(x)]^{r-k-1}[1 - F(y)]^{n-r}}{(k-1)!(r-k-1)!(n-r)!} f_X(x) f_X(y),\ x < y, \\ &= 0,\ x \geqslant y. \end{aligned}$$

Τέλος, η από κοινού συνάρτηση πυκνότητας πιθανότητας των $(X_{(1)}, X_{(n)})$ δίνεται από τη σχέση:

$$\begin{aligned} f_{(X_{(1)},X_{(n)})}(x, y) &= n(n-1)[F(y) - F(x)]^{n-2} f_X(x) f_X(y),\ x < y, \\ &= 0,\ x \geqslant y. \end{aligned}$$

1.7 Οριακά θεωρήματα

Στην ενότητα αυτή παρουσιάζονται βασικά οριακά θεωρήματα από τη θεωρία πιθανοτήτων. Η μελέτη των οριακών κατανομών τυχαίων μεταβλητών παρουσιάζει ιδιαίτερο ενδιαφέρον, αφού μας δίνει τη δυνατότητα ανάπτυξης ασυμπτωτικών αποτελεσμάτων, δηλαδή αποτελεσμάτων για μεγάλα δείγματα, τα οποία οδηγούν στην ανάπτυξη και χρήση ασυμπτωτικών μεθόδων στη Στατιστική.

Ορισμός 1.22: Σχεδόν βέβαιη Σύγκλιση

Η ακολουθία των τυχαίων μεταβλητών X_n, $n = 1, 2, \ldots$, λέμε ότι συγκλίνει σχεδόν βέβαια (σ.β.) στην τυχαία μεταβλητή X και γράφουμε ότι $X_n \overset{\text{σ.β.}}{\longrightarrow} X$, αν

$$P(\{\omega \in \Omega : X_n(\omega) \to X(\omega)\}) = 1.$$

Ορισμός 1.23: Σύγκλιση κατά Μέσο Τετράγωνο

Η ακολουθία των τυχαίων μεταβλητών X_n, $n = 1, 2, \ldots$, λέμε ότι συγκλίνει κατά μέσο τετράγωνο σε μία σταθερά $c \in \mathbb{R}$, καθώς το $n \to \infty$, αν ισχύει

$$\mathrm{E}(X_n - c)^2 \to 0$$

Ορισμός 1.24: Σύγκλιση κατά Πιθανότητα

Η ακολουθία των τυχαίων μεταβλητών X_n, $n = 1, 2, \ldots$, λέμε ότι συγκλίνει κατά πιθανότητα στην τυχαία μεταβλητή X, καθώς το $n \to \infty$, και γράφουμε ότι $X_n \overset{P}{\longrightarrow} X$, αν για κάθε $\varepsilon > 0$ ισχύει ότι:

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1,$$

ή, ισοδύναμα

$$\lim_{n\to\infty} P(|X_n - X| \geqslant \varepsilon) = 0.$$

Ορισμός 1.25: Σύγκλιση κατά Κατανομή

Έστω η ακολουθία των τυχαίων μεταβλητών X_n, $n = 1,2,\ldots$ και έστω $F_{X_1}, F_{X_2}, \ldots$ η αντίστοιχη ακολουθία των συναρτήσεων κατανομής τους. Λέμε ότι η ακολουθία συγκλίνει κατά κατανομή σε μία τυχαία μεταβλητή X με α.σ.κ. F_X, καθώς το $n \to \infty$, και γράφουμε $X_n \xrightarrow{d} X$, αν

$$F_{X_n}(x) \to F_X(x)$$

για κάθε $x \in \mathbb{R}$ όπου η $F_X(\cdot)$ είναι συνεχής.

Θεώρημα 1.3: (Ασθενής Νόμος Μεγάλων Αριθμών)

Έστω $X_1, X_2, \ldots, X_n$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με $\mathrm{E}(X_i) = \mu < \infty$, για κάθε $i = 1,2,\ldots,n$. Τότε για $n \to \infty$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu. \tag{1.33}$$

Δηλαδή $\lim_{n\to\infty} P(|\bar{X} - \mu| > \varepsilon) = 0$, για κάθε $\varepsilon > 0$.

Θεώρημα 1.4: (Ισχυρός Νόμος Μεγάλων Αριθμών)

Έστω $X_1, X_2, \ldots, X_n$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με $\mathrm{E}(X_i) = \mu < \infty$, για κάθε $i = 1,2,\ldots,n$. Τότε για $n \to \infty$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{σ.β.}} \mu. \tag{1.34}$$

Δηλαδή $P\left(\lim_{n\to\infty} \bar{X} = \mu\right) = 1$.

Θεώρημα 1.5: (Κεντρικό Οριακό Θεώρημα)

Έστω $X_1, X_2, \ldots, X_n$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με $\mathrm{E}(X_i) = \mu < \infty$ και $\mathrm{Var}(X_i) = \sigma^2 < \infty$, για κάθε $i = 1,2,\ldots,n$. Τότε για $n \to \infty$

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1). \tag{1.35}$$

Δηλαδή $\lim_{n\to\infty} P\left(\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leqslant z\right) = \Phi(z)$, όπου $\Phi(\cdot)$ είναι η αθροιστική συνάρτηση κατανομής της τυπικής κανονικής κατανομής $\mathcal{N}(0,1)$.

Θεώρημα 1.6: (Μέθοδος Δέλτα)

Έστω $X_1, X_2, \ldots$ μια ακολουθία τυχαίων μεταβλητών με

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} \mathcal{N}(0,\sigma^2), \quad \text{καθώς } n \to \infty,$$

και h μια παραγωγίσιμη συνάρτηση με $h'(\mu) \neq 0$. Τότε,

$$\sqrt{n}(h(X_n) - h(\mu)) \xrightarrow{d} \mathcal{N}(0, \sigma^2 h'(\mu)^2), \quad \text{καθώς } n \to \infty.$$

Θεώρημα 1.7: (Θεώρημα του Slutsky)

Έστω X_n και Y_n είναι ακολουθίες τυχαίων μεταβλητών ή τυχαίων διανυσμάτων ή τυχαίων πινάκων. Αν η X_n συγκλίνει κατά κατανομή στο τυχαίο X και η Y_n συγκλίνει κατά πιθανότητα στη σταθερά c, τότε

- $X_n + Y_n \xrightarrow{d} X + c$.
- $X_n Y_n \xrightarrow{d} Xc$.
- $X_n Y_n^{-1} \xrightarrow{d} Xc^{-1}$, υπό την προϋπόθεση ότι η σταθερά c αντιστρέφεται.

1.8 Βασικές έννοιες από τη στατιστική

Ο βασικός λόγος για τον οποίο χρησιμοποιούμε στατιστικές τεχνικές και μεθόδους είναι προκειμένου να διερευνήσουμε ένα στοχαστικό φαινόμενο ή πείραμα, με βάση τα διαθέσιμα δεδομένα. Τα δεδομένα έρχονται στη μορφή δείγματος από τον πληθυσμό και με βάση την πληροφορία από το δείγμα προσπαθούμε να εξάγουμε συμπεράσματα για τον πληθυσμό. Όλη αυτή η διαδικασία είναι γνωστή και ως **Στατιστική Συμπερασματολογία**. Η Στατιστική Συμπερασματολογία εστιάζει στην εκτίμηση μιας άγνωστης παραμέτρου από την κατανομή που περιγράφει τον πληθυσμό. Για να εκτιμήσουμε την παράμετρο χρειαζόμαστε έναν εκτιμητή, ο οποίος δεν είναι τίποτα περισσότερο από μια συνάρτηση του δείγματος, δηλαδή μια στατιστική συνάρτηση. Ο ορισμός της στατιστικής συνάρτησης θα δοθεί παρακάτω.

Επίσης, στη Στατιστική, εκτός από την εκτίμηση παραμέτρων, μας ενδιαφέρει και η πρόβλεψη (π.χ. εκτίμηση μιας μελλοντικής τιμής ενός χαρακτηριστικού), όταν είναι γνωστές οι τιμές άλλων μεταβλητών. Η μεταβλητή για την οποία θέλουμε να κάνουμε την πρόβλεψη είναι γνωστή ως αποκριτική (ή εξαρτημένη) μεταβλητή, ενώ οι μεταβλητές οι οποίες μας βοηθούν στην πρόβλεψη της τιμής της είναι γνωστές ως επεξηγηματικές (ή ανεξάρτητες) μεταβλητές. Στη συνέχεια θα παρουσιάσουμε συνοπτικά τις βασικές έννοιες που σχετίζονται με τα προαναφερθέντα πλαίσια εφαρμογής των στατιστικών μεθόδων.

Αρχικά, για να εφαρμόσουμε μεθόδους στατιστικής συμπερασματολογίας είναι απαραίτητη η λήψη δείγματος από τον υπό μελέτη πληθυσμό. Ανάλογα με το είδος ενός συνόλου δεδομένων, επιλέγονται τα κατάλληλα μέτρα και οι κατάλληλες τεχνικές για την περιληπτική παρουσίαση και περιγραφή του. Τα δεδομένα διακρίνονται σε δύο βασικές κατηγορίες: (α) σε ποιοτικά ή κατηγορικά δεδομένα και (β) σε ποσοτικά (ή αριθμητικά) δεδομένα.

Όταν τα ποιοτικά δεδομένα δεν έχουν μια φυσική σειρά ή ιεράρχηση (ή διάταξη), τότε λέγονται **ονομαστικά** (nominal) ή **δεδομένα ονομαστικής κλιμάκωσης**. Αν όμως υπάρχει η δυνατότητα διάταξης, τότε τα ποιοτικά δεδομένα λέγονται **διατάξιμα** (ordinal) ή **δεδομένα ιεραρχικής κλιμάκωσης**.

Επίσης, τα ποσοτικά δεδομένα χωρίζονται σε **διακριτά** (discrete data) και σε **συνεχή** (continuous data). Στην πρώτη περίπτωση, το πλήθος των αριθμητικών τιμών είναι ένα αριθμήσιμο σύνολο, ενώ στη δεύτερη περίπτωση το πλήθος των δυνατών τιμών είναι υπερ-αριθμήσιμο.

Στο σημείο αυτό, αξίζει να γίνει μια αναφορά στους διαφορετικούς τύπους μετρήσεων που μπορούν να ληφθούν. Στην περίπτωση ονομαστικών μετρήσεων (nominal scale), έχουμε τον πλέον απλό τύπο μετρήσεων όπου τα διαφορετικά άτομα/μέλη στο δείγμα τοποθετούνται σε διαφορετικές κλάσεις ή

κατηγορίες απλώς και μόνο από το αποτέλεσμα. Παραδείγματα δεδομένων ονομαστικής κλίμακας είναι τα αποτελέσματα της ρίψης ενός νομίσματος, οι απαντήσεις στο ερώτημα σχετικά με την πρόθεση ψήφου ή την ομάδα ποδοσφαίρου που υποστηρίζουμε.

Ο επόμενος τύπος μετρήσεων είναι αυτός της διατάξιμης κλίμακας, όπου υπάρχει η δυνατότητα (φυσικής) τοποθέτησης των δυνατών αποτελεσμάτων από το μικρότερο στο μεγαλύτερο ή από το καλύτερο στο χειρότερο. Παραδείγματα μετρήσεων διατάξιμης κλίμακας είναι η ηλικιακή ομάδα, όταν δεν ζητείται η ακριβής ηλικία, τα στάδια μιας ασθένειας, η γνώμη που έχει ο/η ερωτώμενος/η ως προς τα θέματα οικονομίας/εξωτερικής πολιτικής π.χ. «συμφωνώ λίγο», «συμφωνώ πολύ», «μάλλον συμφωνώ», «ούτε συμφωνώ/ούτε διαφωνώ» κ.λπ.

Η επόμενη κατηγορία μετρήσεων είναι αυτή των ποσοτικών μετρήσεων με αυθαίρετο μηδέν, γνωστή και ως δεδομένα διαστήματος (interval data). Σε αυτήν την περίπτωση, οι διαθέσιμες μετρήσεις μπορούν να διαταχθούν με φυσικό τρόπο, ενώ έχει φυσικό νόημα και η διαφορά (απόσταση) μεταξύ των μετρήσεων. Συγκεκριμένα, η απόσταση μεταξύ δύο οποιωνδήποτε διαθέσιμων μετρήσεων μπορεί να εκφραστεί σε συγκεκριμένες μονάδες. Χαρακτηριστικό παράδειγμα τέτοιου είδους μετρήσεων είναι οι μετρήσεις θερμοκρασίας, στις οποίες η διαφορά μεταξύ δύο θερμοκρασιών μετριέται σε βαθμούς (μονάδες) της κλίμακας Κελσίου. Οι πραγματικές τιμές της θερμοκρασίας συγκρίνονται με ένα σημείο αυθαίρετου μηδέν (zero degrees). Έτσι, στην περίπτωση των δεδομένων με αυθαίρετο μηδέν χρειάζεται και το σημείο μηδέν, αλλά και ο ορισμός μιας μοναδιαίας απόστασης (unit distance) μεταξύ των μετρήσεων, χωρίς όμως να είναι αυστηρός ο τρόπος επιλογής του σημείου μηδέν ή/και της μοναδιαίας απόστασης.

Τέλος, υπάρχει και η κατηγορία ποσοτικών μετρήσεων με απόλυτο μηδέν, όπου, εκτός από τη δυνατότητα φυσικής διάταξης και απόστασης μεταξύ των μετρήσεων, έχει νόημα και ο λόγος μεταξύ δύο μετρήσεων. Για παράδειγμα, αν έχει νόημα να αναφερθούμε στο ότι μια ποσότητα είναι διπλάσια μιας άλλης, τότε έχει νόημα να αναφερθούμε σε ποσοτικά δεδομένα με απόλυτο μηδέν ή σε δεδομένα λόγου (ratio data). Χαρακτηριστικά παραδείγματα τέτοιων μετρήσεων είναι οι χιλιομετρικές αποστάσεις, τα βάρη, τα ύψη, τα εισοδήματα κ.λπ. Στην πραγματικότητα, η βασική διαφορά μεταξύ μετρήσεων διαστήματος και μετρήσεων λόγου είναι ότι στη δεύτερη περίπτωση η τιμή μηδέν έχει φυσική ερμηνεία. Αντίθετα, στην πρώτη περίπτωση, το μηδέν ορίζεται αυθαίρετα. Αξίζει, επίσης, να αναφέρουμε ότι και στην περίπτωση δεδομένων λόγου, η μοναδιαία απόσταση ορίζεται με αυθαίρετο τρόπο.

Η παραπάνω παρουσίαση των δυνατών τύπων μετρήσεων έγινε προκειμένου ο/η αναγνώστης/τρια, να έχει μια εικόνα του τι δεδομένα μπορούμε να συλλέξουμε από έναν πληθυσμό, δηλαδή τι είδους δεδομένα (μετρήσεις) μπορούμε να έχουμε σε ένα δείγμα. Στη συνέχεια, δίνουμε τον ορισμό του απλού τυχαίου δείγματος.

Ορισμός 1.26: Απλό Τυχαίο Δείγμα

Ένα δείγμα μεγέθους n από έναν πληθυσμό μεγέθους N λέγεται απλό τυχαίο δείγμα, όταν οι $\binom{N}{n}$ το πλήθος δυνατές n-άδες του πληθυσμού έχουν την ίδια πιθανότητα να αποτελούν το δείγμα επιλογής.

Το στάδιο της συλλογής δεδομένων είναι ιδιαίτερα κρίσιμο στη Στατιστική αφού προσδιορίζει το είδος της στατιστικής ανάλυσης που μπορεί να γίνει και άρα επιδρά στην εξαγωγή συμπερασμάτων. Διαφορετικοί τύποι μετρήσεων (δεδομένων) απαιτούν την εφαρμογή διαφορετικών στατιστικών μεθόδων και τεχνικών. Επιπλέον, η λήψη ενός όσο το δυνατόν πιο αντιπροσωπευτικού δείγματος από τον πληθυσμό οδηγεί σε πιο αξιόπιστα συμπεράσματα για το σύνολο του υπό μελέτη πληθυσμού. Έτσι, για να διασφαλίσουμε την αξιοπιστία των συμπερασμάτων που προκύπτει μετά από κάθε στατιστική ανάλυση, προτείνεται η επιλογή ενός απλού τυχαίου δείγματος από τον πληθυσμό. Αξίζει να αναφέρουμε ότι, εκτός από τη λήψη ενός απλού τυχαίου δείγματος, υπάρχει η δυνατότητα εφαρμογής διαφορετικών μεθόδων δειγματοληψίας, ώστε να διασφαλίσουμε τη μεγαλύτερη δυνατή αντιπροσωπευτικότητα του δείγματος, σε σχέση με τον υπό μελέτη πληθυσμό. Τέτοιες μέθοδοι είναι η στρωματοποιημένη δειγματοληψία, η συστηματική δειγματοληψία

και η δειγματοληψία κατά συστάδες. Στη συνέχεια, και εκτός αν αναφέρεται διαφορετικά, η μέθοδος επιλογής του δείγματος είναι η απλή τυχαία δειγματοληψία, η οποία οδηγεί στη λήψη ενός απλού τυχαίου δείγματος.

1.8.1 Στατιστική συμπερασματολογία

Αρχικά, δίνουμε τον ορισμό του τυχαίου δείγματος.

Ορισμός 1.27: Τυχαίο Δείγμα

Αν οι τυχαίες μεταβλητές $X_1, X_2, \dots, X_n$ είναι ανεξάρτητες και έχουν όλες την ίδια κατανομή (δηλαδή είναι και ισόνομες), τότε λέμε ότι αποτελούν ένα τυχαίο δείγμα από αυτήν την κατανομή.

Έστω $\mathbf{X} = (X_1, X_2, \dots, X_n)$ τυχαίο δείγμα από έναν πληθυσμό με $X_i \sim F(x;\theta)$, όπου θ είναι η παράμετρος της κατανομής που περιγράφει τη συμπεριφορά των τιμών του πληθυσμού. Συνήθως, το θ είναι άγνωστο και θέλουμε να το εκτιμήσουμε. Για να το κάνουμε αυτό, θα πρέπει να βρούμε έναν εκτιμητή του θ, ο οποίος είναι μια συνάρτηση του τυχαίου δείγματος $\mathbf{X}$, δηλαδή είναι μια στατιστική συνάρτηση.

Ορισμός 1.28: Στατιστική Συνάρτηση

Ως στατιστική συνάρτηση (σ.σ.) ορίζουμε μία συνάρτηση του τυχαίου δείγματος $\mathbf{X} = (X_1, X_2, \dots, X_n)$. Συνήθως, συμβολίζεται ως $\delta(\mathbf{X})$ ή $T(\mathbf{X})$ και είναι τυχαία μεταβλητή.

Ο εκτιμητής της παραμέτρου θ συμβολίζεται ως $\hat{\theta} \equiv \hat{\theta}(\mathbf{X})$ και αποκαλείται σημειακός εκτιμητής. Για παράδειγμα, αν μ είναι η (άγνωστη) μέση τιμή ενός πληθυσμού και $\mathbf{X} = (X_1, X_2, \dots, X_n)$ τυχαίο δείγμα από τον πληθυσμό, τότε ο δειγματικός μέσος $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ είναι σ.σ. και εκτιμητής του μ. Αν αντικαταστήσουμε τα $X_1, X_2, \dots, X_n$ με τις τιμές $x_1, x_2, \dots, x_n$ από το διαθέσιμο δείγμα, τότε έχουμε μια σημειακή εκτίμηση του μ, τη $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Όμοια, για την (άγνωστη) διασπορά σ^2 ενός πληθυσμού, ένας εκτιμητής είναι ο $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ (δειγματική διασπορά). Η αντίστοιχη σημειακή εκτίμηση του σ^2 με χρήση των τιμών του δείγματος είναι $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Επίσης, αν $\mathbf{X} = (X_1, X_2, \dots, X_n)$ είναι τυχαίο δείγμα από έναν πληθυσμό με μέση τιμή και διακύμανση $\mu = \mathrm{E}(X)$ και $\sigma^2 = \mathrm{Var}(X)$, αντίστοιχα, τότε, για τον δειγματικό μέσο $\bar{X}$ και για τη δειγματική διασπορά S^2 ισχύει ότι $\mathrm{E}(\bar{X}) = \mu$, $\mathrm{E}(S^2) = \sigma^2$ και $\mathrm{Var}(\bar{X}) = \sigma^2/n$.

Για την εύρεση των εκτιμητών υπάρχουν τεχνικές και μεθοδολογίες, όπως η μέθοδος της μέγιστης πιθανοφάνειας και η μέθοδος των ροπών. Επίσης, ένας εκτιμητής επιθυμούμε να πληροί συγκεκριμένα κριτήρια/ιδιότητες, όπως την ιδιότητα της αμεροληψίας, άρα μιλάμε για αμερόληπτους εκτιμητές, του ελαχίστου μέσου τετραγωνικού σφάλματος (ΜΤΣ) ή της συνέπειας, άρα αναφερόμαστε σε συνεπείς εκτιμητές. Παρακάτω δίνονται συνοπτικά κάποια από τα χαρακτηριστικά που έχουν οι «καλοί εκτιμητές».

Ορισμός 1.29: Μέσο Τετραγωνικό Σφάλμα

Έστω $T = T(\mathbf{X})$ εκτιμητής της παραμέτρου θ. Τότε, ως μέσο τετραγωνικό σφάλμα (Mean Square Error, MSE) ορίζεται η ποσότητα

$$\mathrm{MSE}(T) = \mathrm{E}(T-\theta)^2 = \mathrm{Var}(T) + (\mathrm{E}(T) - \theta)^2$$

Γενικά, είναι επιθυμητό ένας εκτιμητής να έχει το ελάχιστο Μέσο Τετραγωνικό Σφάλμα για κάθε θ, αλλά για να συμβεί κάτι τέτοιο θα πρέπει να περιορίσουμε το πρόβλημα σε κατάλληλα υποσύνολα εκτιμητών, όπως είναι για παράδειγμα το σύνολο των *αμερόληπτων εκτιμητών*.

Ορισμός 1.30: Αμερόληπτος Εκτιμητής

Έστω $T = T(\mathbf{X})$ εκτιμητής της παραμέτρου θ. Τότε, λέμε ότι ο T είναι αμερόληπτος εκτιμητής (unbiased estimator) του θ, αν

$$\text{MSE}(T) = \text{Var}(T) \text{ ή } \text{E}(T) = \theta.$$

Ορισμός 1.31: Σχετική Αποτελεσματικότητα

Έστω T_1, T_2 εκτιμητές του θ. Τότε, η σχετική αποτελεσματικότητα του T_1 ως προς τον T_2 είναι

$$\text{eff}(T_1, T_2) = \frac{\text{MSE}(T_2)}{\text{MSE}(T_1)}.$$

Αν οι εκτιμητές T_1, T_2 είναι αμερόληπτοι, τότε $\text{eff}(T_1, T_2) = \text{Var}(T_2)/\text{Var}(T_1)$. Αν για οποιοδήποτε μέγεθος δείγματος n, είναι $\text{eff}(T_1, T_2) \geqslant 1$, τότε ο T_1 είναι πιο αποδοτικός από τον T_2 για την εκτίμηση της παραμέτρου θ.

Ορισμός 1.32: Συνεπείς Εκτιμητές

Έστω T εκτιμητής του θ και για μια ακολουθία τυχαίων μεταβλητών $X_1, X_2, \ldots$, θεωρούμε την αντίστοιχη ακολουθία εκτιμητών $T_1, T_2, \ldots, T_n$ (για τα αντίστοιχα δείγματα μεγέθους $1, 2, \ldots, n$), δηλ. $T_1 = T(X_1)$, $T_2 = T(X_1, X_2)$, ..., $T_n = T(X_1, \ldots, X_n)$. Αν για κάθε $\varepsilon > 0$, ισχύει ότι:

$$\lim_{n\to\infty} P(|T_n - \theta| \geqslant \varepsilon) = 0$$

ή, ισοδύναμα,

$$\lim_{n\to\infty} P(|T_n - \theta| < \varepsilon) = 1,$$

τότε ο T είναι συνεπής εκτιμητής.

Από τα παραπάνω οριακά αποτελέσματα συμπεραίνουμε ότι μια συνεπής ακολουθία εκτιμητών συγκλίνει κατά πιθανότητα στο θ. Επίσης, καθώς το n αυξάνεται, $\text{E}(T_n) \to \theta$ και $\text{Var}(T_n) \to 0$. Σημειώνεται, επίσης, ότι οι παραπάνω συνθήκες είναι ικανές, αλλά όχι αναγκαίες για τη συνέπεια ενός εκτιμητή.

1.8.2 Διαστήματα εμπιστοσύνης

Εκτός από τις μεθόδους εκτίμησης μιας παραμέτρου με σημείο (εύρεση σημειακών εκτιμητών), υπάρχει και η μέθοδος εκτίμησης με διάστημα. Στη μέθοδο αυτή στόχος είναι η εύρεση ενός (τυχαίου) διαστήματος στο οποίο να περιέχεται η άγνωστη τιμή της παραμέτρου θ με συγκεκριμένη πιθανότητα.

Έστω $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ τυχαίο δείγμα από έναν πληθυσμό με σ.π.π. ή σ.π. $f(x;\theta)$. Έστω $L(\mathbf{X}) = L(X_1, X_2, \ldots, X_n)$ και $U(\mathbf{X}) = U(X_1, X_2, \ldots, X_n)$ το κάτω και το άνω όριο ενός διαστήματος. Τα $L(\mathbf{X})$, $U(\mathbf{X})$ είναι στατιστικές συναρτήσεις με $L(\mathbf{X}) < U(\mathbf{X})$. Τότε, αν το (τυχαίο) διάστημα $[L(\mathbf{X}), U(\mathbf{X})]$ χρησιμοποιηθεί για την εκτίμηση του θ, λέμε ότι είναι ένα **διάστημα εμπιστοσύνης** για το θ .

Επίσης, η πιθανότητα το θ να περιέχεται στο διάστημα $[L, U]$, δηλαδή η πιθανότητα

$$P(\theta \in [L(\mathbf{X}), U(\mathbf{X})]) \equiv P(L(\mathbf{X}) \leqslant \theta \leqslant U(\mathbf{X}))$$

λέγεται **πιθανότητα κάλυψης** του διαστήματος, ενώ η ποσότητα

$$\inf_{\theta\in\Theta} P(L(\mathbf{X}) \leqslant \theta \leqslant U(\mathbf{X}))$$

λέγεται **συντελεστής εμπιστοσύνης** του διαστήματος. Αν το διάστημα $[L(\mathbf{X}), U(\mathbf{X})]$ έχει συντελεστή

εμπιστοσύνης τουλάχιστον $1-a$, δηλαδή ισχύει ότι

$$P(L(\mathbf{X}) \leqslant \theta \leqslant U(\mathbf{X})) \geqslant 1-a, \quad \forall\, \theta \in \Theta,$$

τότε λέμε ότι είναι ένα $100(1-a)\%$ διάστημα εμπιστοσύνης για το θ.

Θέλουμε να χρησιμοποιήσουμε το διάστημα $[L(\mathbf{X}), U(\mathbf{X})]$ με σκοπό την εκτίμηση του θ. Για να το επιτύχουμε, θέλουμε:

- το (τυχαίο) διάστημα $[L(\mathbf{X}), U(\mathbf{X})]$ να περιέχει την πραγματική τιμή του θ με μεγάλη πιθανότητα, και
- το (τυχαίο) διάστημα $[L(\mathbf{X}), U(\mathbf{X})]$ να έχει όσον το δυνατόν μικρότερο μήκος.

Σε πρακτικά προβλήματα καθορίζουμε αρχικά την τιμή $1-a$, μέσω της οποίας εκφάζουμε τη βεβαιότητα στο να περιέχει το διάστημα $[L(\mathbf{X}), U(\mathbf{X})]$ την πραγματική τιμή του θ και στη συνέχεια, βρίσκουμε το διάστημα που θα έχει συντελεστή εμπιστοσύνης ίσο με $1-a$. Στις περιπτώσεις που π.χ. η κατανομή των δεδομένων είναι διακριτή, δεν είναι πάντοτε δυνατό να βρούμε ένα διάστημα εμπιστοσύνης με συντελεστή εμπιστοσύνης ακριβώς $1-a$. Τότε αρκεί να βρούμε διαστήματα για τα οποία ο συντελεστής εμπιστοσύνης είναι $\geqslant 1-a$. Τέτοια διαστήματα ονομάζονται συντηρητικά (conservative).

Τα γενικά βήματα για την κατασκευή ενός διαστήματος εμπιστοσύνης είναι τα εξής:

1. Αρχικά, βρίσκουμε έναν «καλό» σημειακό εκτιμητή της παραμέτρου θ (π.χ. έναν εκτιμητή μέγιστης πιθανοφάνειας), έστω αυτός $\hat{\theta}$.
2. Στη συνέχεια, βρίσκουμε την κατανομή του $\hat{\theta}$ και με βάση αυτή, προσδιορίζουμε μια *ποσότητα οδηγό* (ή *αντιστρεπτή ποσότητα*), έστω αυτή $T(\mathbf{X}, \theta)$. Η ποσότητα αυτή εξαρτάται από το θ, αλλά η κατανομή της δεν εξαρτάται από αυτό.
3. Αφού είναι γνωστή η κατανομή της $T(\mathbf{X}, \theta)$, προσδιορίζουμε σταθερές c_1, c_2 τέτοιες ώστε

$$P(c_1 \leqslant T(\mathbf{X}, \theta) \leqslant c_2) = 1-a, \quad \forall\, \theta \in \Theta.$$

4. Επιλύουμε τις ανισότητες $T(\mathbf{X}, \theta) \geqslant c_1$ και $T(\mathbf{X}, \theta) \leqslant c_2$ ως προς θ και έστω ότι οι λύσεις αυτών είναι $L(\mathbf{X}) \leqslant \theta$ και $U(\mathbf{X}) \geqslant \theta$. Πλέον, το τυχαίο διάστημα $[L(\mathbf{X}), U(\mathbf{X})]$ είναι ένα $100(1-a)\%$ διάστημα εμπιστοσύνης για το θ.

Συνήθως, τα c_1, c_2 επιλέγονται έτσι ώστε είτε να ελαχιστοποιείται το μήκος του διαστήματος, οπότε και αναφερόμαστε σε *διαστήματα εμπιστοσύνης ελαχίστου μήκους*, είτε να ισχύει:

$$P(T(\mathbf{X}, \theta) < c_1) = P(T(\mathbf{X}, \theta) > c_2) = a/2, \quad \forall\, \theta \in \Theta.$$

Σε αυτήν την περίπτωση, το διάστημα εμπιστοσύνης είναι γνωστό και ως $100(1-a)\%$ διάστημα εμπιστοσύνης *ίσων ουρών* για το θ.

1.8.3 Έλεγχος υποθέσεων

Μία από τις σημαντικότερες πρακτικές εφαρμογές της Στατιστικής είναι ο έλεγχος στατιστικών υποθέσεων. Ουσιαστικά, ασχολούμαστε με ελέγξιμες υποθέσεις, οι οποίες αναφέρονται σε ιδιότητες πιθανοτικών προτύπων ή στατιστικών πληθυσμών. Οι υποθέσεις αυτές λέγονται στατιστικές υποθέσεις. Ως Έλεγχο Υποθέσεων ορίζουμε τη διαδικασία της Στατιστικής Συμπερασματολογίας με την οποία καταλήγουμε στο να αποφασίσουμε υπέρ μίας από δύο αντικρουόμενες στατιστικές υποθέσεις. Η μία υπόθεση λέγεται μηδενική υπόθεση και συμβολίζεται με H_0, ενώ η άλλη λέγεται εναλλακτική υπόθεση και συμβολίζεται με H_1.

Συνήθως, η μορφή της H_0 είναι $H_0 : \theta = \theta_0$, όπου θ είναι παράμετρος του πληθυσμού, π.χ. $\theta = \mu$ ή $\theta = \mu_1 - \mu_2$. Ως θ_0 συμβολίζουμε μία συγκεκριμένη τιμή της παραμέτρου. Υποθέσεις αυτής της μορφής ονομάζονται απλές υποθέσεις (simple hypothesis). Αν η υπόθεση είναι της μορφής $\theta \geqslant \theta_0$ ή $\theta \leqslant \theta_0$ ή

$\theta \neq \theta_0$, τότε αυτή ονομάζεται σύνθετη (composite hypothesis). Συνήθως, με τη μορφή σύνθετης υπόθεσης διατυπώνεται η H_1, όπως π.χ. $H_1 : \theta > \theta_0$, $H_1 : \theta < \theta_0$ ή $H_1 : \theta \neq \theta_0$. Οι δύο πρώτες υποθέσεις λέμε ότι είναι μονόπλευρες εναλλακτικές, ενώ η τρίτη λέμε ότι είναι δίπλευρη εναλλακτική υπόθεση.

Για να διεξάγουμε έναν στατιστικό έλεγχο υπόθεσης, αρχικά καθορίζουμε τις H_0, H_1. Συνήθως, θεωρούμε ως H_1 τον ισχυρισμό που θέλουμε να υποστηρίξουμε και ως H_0 την άρνηση της H_1, δηλαδή τον αντικρουόμενο ισχυρισμό. Στη συνέχεια, είναι απαραίτητη η χρήση μιας στατιστικής συνάρτησης ελέγχου ή ελεγχοσυνάρτησης (test statistic). Έτσι, ο έλεγχος της H_0 έναντι της H_1 χρησιμοποιεί έναν *κανόνα απόφασης* που βασίζεται στην παρατηρούμενη τιμή της στατιστικής συνάρτησης ελέγχου (σ.σ.ε.) ή ελεγχοσυνάρτησης. Για να μπορεί μια στατιστική συνάρτηση να χρησιμοποιηθεί ως σ.σ.ε., θα πρέπει η κατανομή της υπό την H_0 να είναι πλήρως προσδιορισμένη.

Ο κανόνας απόφασης καθορίζει δύο περιοχές τιμών της κατανομής της σ.σ.ε., οι οποίες είναι ξένες μεταξύ τους. Η μία περιοχή ονομάζεται **περιοχή αποδοχής** της H_0 (acceptance region). Αν βρεθεί η τιμή της σ.σ.ε. σε αυτήν την περιοχή, τότε δεν απορρίπτουμε την H_0. Η δεύτερη περιοχή ονομάζεται **περιοχή απόρριψης** (rejection region) ή *κρίσιμη περιοχή* της H_0. Αν βρεθεί η τιμή της σ.σ.ε. σε αυτήν την περιοχή, τότε απορρίπτουμε την H_0 και αποδεχόμαστε την H_1.

Όμως, επειδή σε κάθε στατιστικό έλεγχο αποφασίζουμε με βάση την τιμή μιας κατάλληλης σ.σ.ε., δηλαδή με βάση τη διαθέσιμη πληροφορία από το δείγμα, για την αποδοχή ή την απόρριψη μίας υπόθεσης υπάρχουν οι παρακάτω δύο «κίνδυνοι» ή σφάλματα. Συγκεκριμένα, τα σφάλματα αυτά είναι τα ακόλουθα:

- **Σφάλμα τύπου I** (Type I Error) το οποίο αντιστοιχεί σε απόρριψη της H_0, ενώ στην πραγματικότητα αυτή είναι αληθής (εσφαλμένη απόρριψη της H_0).
- **Σφάλμα τύπου II** (Type II Error) το οποίο αντιστοιχεί σε αποδοχή της H_0, ενώ στην πραγματικότητα αυτή δεν είναι αληθής (εσφαλμένη αποδοχή της H_0) και άρα είναι αληθής η H_1.

Οι πιθανότητες των σφαλμάτων τύπου I και II συμβολίζονται αντίστοιχα ως a, β και μπορούν να γραφτούν ως

$$a = P(\text{σφάλμα τύπου I}) = P(\text{απορρίπτεται η } H_0 \,|H_0 \text{ αληθής})$$

και

$$\beta = P(\text{σφάλμα τύπου II}) = P(\text{δεν απορρίπτεται η } H_0 \,|H_1 \text{ αληθής})$$

Δείτε και τον παρακάτω Πίνακα 1.1.

	Τι πραγματικά ισχύει	
Τι αποφασίζω	H_0	H_1
Υπέρ H_0	Ορθή Απόφαση με πιθανότητα $1-a$	Εσφαλμένη Απόφαση Σφάλμα Τύπου II
Υπέρ H_1	Εσφαλμένη Απόφαση Σφάλμα Τύπου I	Ορθή Απόφαση με πιθανότητα $1-\beta$

Πίνακας 1.1: Ορθές/Λανθασμένες Αποφάσεις κατά τη διαδικασία ενός Ελέγχου Υποθέσεων.

Το επιθυμητό θα ήταν να επιτυγχάνεται η ταυτόχρονη ελαχιστοποίηση των a και β. Όμως κάτι τέτοιο δεν είναι δυνατόν. Δεν υπάρχει μαθηματική σχέση που να συνδέει άμεσα τα a, β, όμως, για δεδομένο μέγεθος δείγματος n η αύξηση του ενός οδηγεί σε μείωση του άλλου. Αυτό που, συνήθως, κάνουμε κατά τη διεξαγωγή ενός ελέγχου υποθέσεων είναι να προκαθορίζουμε το a και, για το δεδομένο a, ο κανόνας απόφασης να είναι τέτοιος ώστε να ελαχιστοποιείται το β. Ισοδύναμα, αντί να ελαχιστοποιείται το β, επιδιώκουμε να μεγιστοποιήσουμε την **ισχύ του ελέγχου** (power of the test), η οποία ορίζεται ως

$$1-\beta = P(\text{απορρίπτεται η } H_0 \,|H_1 \text{ αληθής}).$$

Όταν η μηδενική υπόθεση είναι της μορφής $H_0 : \theta = \theta_0$, η πιθανότητα σφάλματος τύπου I λέγεται **επίπεδο σημαντικότητας** (ε.σ.) ή μέγεθος του ελέγχου (size of the test). Αξίζει να αναφέρουμε ότι υπάρχουν και περιπτώσεις όπου η πιθανότητα σφάλματος τύπου I δεν είναι μία αλλά πολλές, καθώς καθεμία από αυτές αντιστοιχεί σε μια δεδομένη τιμή του θ μεταξύ αυτών που καθορίζονται υπό την H_0. Για παράδειγμα, αν $H_0 : \theta \leqslant \theta_0$ ή $H_0 : \theta \geqslant \theta_0$, τότε οι τιμές για το θ είναι στο $(-\infty, \theta_0]$ ή στο $[\theta_0, \infty)$, αντίστοιχα. Σε αυτές τις περιπτώσεις, το ε.σ. ορίζεται ως η μέγιστη πιθανότητα σφάλματος τύπου I. Επίσης, επειδή η H_1 είναι συνήθως σύνθετη υπόθεση, η πιθανότητα σφάλματος τύπου II εξαρτάται από την πραγματική τιμή του θ, την οποία, φυσικά, δεν γνωρίζουμε. Η πιθανότητα αυτή ορίζει στην ουσία μια συνάρτηση $\beta(\theta)$, με το θ να ανήκει στην περιοχή τιμών η οποία καθορίζεται υπό την H_1. Η $\beta(\theta)$ είναι γνωστή και ως **λειτουργική χαρακτηριστική καμπύλη** (operating characteristic curve, OC curve) ή απλά, χαρακτηριστική καμπύλη.

Όταν για τη διεξαγωγή ελέγχων χρησιμοποιείται ηλεκτρονικός υπολογιστής (Η/Υ) και στατιστικά πακέτα, όπως το IBM SPSS ή η `R`, τότε έχει επικρατήσει η απόφαση για την αποδοχή ή απόρριψη της H_0 να μην γίνεται εξετάζοντας αν η τιμή της στατιστικής συνάρτησης ελέγχου, η οποία, συνήθως, δίνεται από το πακέτο, ανήκει στην περιοχή απόρριψης, αλλά με χρήση της p-τιμής (p-value), η οποία συχνά εμφανίζεται και ως Sig- από το significance, που σημαίνει σημαντικότητα. Αυτή η μορφή του στατιστικού ελέγχου υποθέσεων είναι γνωστή και ως έλεγχος σημαντικότητας. Στους ελέγχους σημαντικότητας καθορίζονται πρώτα οι υποθέσεις H_0 και H_1, καθώς και η σ.σ.ε., όπως στην περίπτωση των ελέγχων υποθέσεων. Αυτό που δεν καθορίζεται από πριν είναι το ε.σ. a, κάτι το οποίο θα οδηγούσε σε συγκεκριμένη περιοχή απόρριψης της H_0. Εναλλακτικά, προσδιορίζεται το μέγεθος των ενδείξεων έναντι της H_0. Δηλαδή, με βάση τη διαθέσιμη πληροφορία από το δείγμα, προσδιορίζεται αν υπάρχει ή δεν υπάρχει ισχυρή ένδειξη έναντι της H_0. Ο προσδιορισμός αυτός γίνεται με χρήση της p-τιμής, η οποία ονομάζεται **παρατηρούμενο επίπεδο σημαντικότητας** και εκφράζει την πιθανότητα να παρατηρήσουμε στην τύχη, όταν είναι αληθής η H_0, δηλαδή για $\theta = \theta_0$, μια τιμή της σ.σ.ε. που να είναι ίση ή πιο ακραία από αυτήν που ήδη παρατηρήσαμε. Παρακάτω δίνουμε τον σχετικό ορισμό.

Ορισμός 1.33

Το παρατηρούμενο επίπεδο σημαντικότητας (observed level of significance) ή p-τιμή (p-value) είναι το ελάχιστο επίπεδο σημαντικότητας για το οποίο απορρίπτεται η μηδενική υπόθεση (H_0), για τις δοθείσες παρατηρήσεις (διαθέσιμα δεδομένα).

Ο παραπάνω ορισμός της p-τιμής είναι αρκετά γενικός και εφαρμόζεται σε όλες τις περιπτώσεις. Αξίζει, όμως, να αναφέρουμε πως, όταν η H_1 είναι δίπλευρη και η κατανομή της σ.σ.ε. για $\theta = \theta_0$ είναι συμμετρική, τότε το p-value μπορεί να βρεθεί με διπλασιασμό του p-value που αντιστοιχεί σε μονόπλευρη εναλλακτική.

Στην πράξη, όταν είναι διαθέσιμη η p-τιμή, διεξάγουμε τον έλεγχο υπόθεσης με βάση τον παρακάτω κανόνα: Απορρίπτουμε την H_0 αν, p-value $< a$, όπου a είναι το ε.σ. a. Συνήθεις τιμές για το a είναι οι 0.10, 0.05 και 0.01. Αντίθετα, «μεγάλες» τιμές σημαίνει ότι δεν μπορούμε να απορρίψουμε την H_0 στο αντίστοιχο ε.σ., οπότε η απόφαση είναι η αποδοχή (μη απόρριψη) της H_0. Ένας κανόνας απόφασης μέσω της p-τιμής, ο οποίος έχει επικρατήσει, είναι ο εξής:

- αν p-value < 0.05, τότε απορρίπτεται η H_0, ενώ
- αν p-value $\geqslant 0.05$, τότε δεν απορρίπτεται η H_0.

Παρατήρηση 1.2. Στο σημείο αυτό, αξίζει να αναφέρουμε ότι από το 2010 και μετά αυτός ο κανόνας απόφασης δέχεται μία αυστηρή κριτική. Μάλιστα, το 2014, ο Ομότιμος Καθηγητής George Cobb του Mount Holyoke College σε ένα φόρουμ της Αμερικάνικης Στατιστικής Ένωσης (*American Statistical Association*) έθεσε τα εξής ερωτήματα (για λεπτομέρειες βλ. Wasserstein and Lazar, 2016):

> Ερ.: Γιατί τα κολέγια και τα σχολεία διδάσκουν ότι το επίπεδο σημαντικότητας είναι $a = 0.05$;
> Απ.: Επειδή αυτό χρησιμοποιούν η επιστημονική κοινότητα και οι εκδότες των επιστημονικών περιοδικών.

Ερ.: Γιατί τόσοι πολλοί άνθρωποι ακόμα χρησιμοποιούν ως επίπεδο σημαντικότητας το $a = 0.05$;
Απ: Επειδή αυτό μαθαίνουν στα κολέγια και στα σχολεία.

Ένα ενδιαφέρον άρθρο για τα προβλήματα που παρουσιάζονται, όταν παίρνουμε απόφαση μέσω της p-τιμής, είναι αυτό του Harrell (2020)[2]. Στο σημείο αυτό, πρέπει να σημειώσουμε ότι με την παράθεση των παραπάνω, σκοπός μας δεν είναι να επιτείνουμε την αμφισβήτηση για τη χρήση της p-τιμής, αλλά να ενημερώσουμε τον/την αναγνώστη/τρια για τους προβληματισμούς που εγείρει η επιστημονική κοινότητα όσον αφορά την απόφαση που καλούμαστε να πάρουμε μέσω της p-τιμής και, κυρίως, ως προς την ορθή ερμηνεία της συγκεκριμένης τιμής.

1.8.4 Ιδιότητες ελέγχων υποθέσεων

Για την ορθή εφαρμογή ενός ελέγχου υποθέσεων, θα πρέπει να επιλεχθεί κατάλληλη στατιστική συνάρτηση ελέγχου (τεστ). Στις περιπτώσεις που υπάρχουν περισσότερες από μία στατιστικές συναρτήσεις ελέγχου, θα πρέπει η επιλογή να γίνει με βάση το αν ικανοποιούνται οι υποθέσεις για την ορθή εφαρμογή του ελέγχου που βασίζεται στη συγκεκριμένη ελεγχοσυνάρτηση. Για παράδειγμα, αν θέλουμε να ελέγξουμε κατά πόσο η μέση τιμή της κατανομής των τιμών ενός χαρακτηριστικού X σε έναν πληθυσμό είναι ίση με μια δεδομένη τιμή (π.χ. έλεγχος της μορφής $H_0 : \mu = \mu_0$ έναντι $H_0 : \mu = \mu_1$), μια «λογική» επιλογή για τη στατιστική συνάρτηση ελέγχου είναι ο δειγματικός μέσος $\bar{X}$. Από εκεί και έπειτα, ο έλεγχος θα βασιστεί στην κατανομή του $\bar{X}$, η οποία εξαρτάται και από την κατανομή της τυχαίας μεταβλητής X. Για παράδειγμα, υπάρχουν έλεγχοι οι οποίοι βασίζονται στο ότι η κατανομή της X είναι η κανονική κατανομή. Δηλαδή, η υπόθεση σε αυτήν την περίπτωση είναι ότι το τυχαίο δείγμα $X_1, X_2, \dots, X_n$ προέρχεται από πληθυσμό που μοντελοποιείται σύμφωνα με το πρότυπο της κανονικής κατανομής. Αν κάτι τέτοιο δεν ισχύει, τουτέστιν τα διαθέσιμα δεδομένα δείχνουν απόκλιση από την υπόθεση της κανονικότητας, δεν θα πρέπει να επιλεχθεί αυτός ο έλεγχος. Από την άλλη, αν υπάρχουν ενδείξεις ότι, έστω και προσεγγιστικά, η κατανομή της X είναι κανονική, τότε μπορεί να χρησιμοποιηθεί ο συγκεκριμένος έλεγχος, έχοντας όμως κατά νου πως είναι μια προσεγγιστική μέθοδος, με την προσέγγιση να είναι ικανοποιητική κάτω από συγκεκριμένες συνθήκες.

Γενικά, υπάρχουν τρεις βασικές ιδιότητες που πρέπει να ικανοποιεί ένας «καλός» έλεγχος. Συγκεκριμένα,

- ο έλεγχος πρέπει να είναι αμερόληπτος (unbiased test),
- ο έλεγχος πρέπει να είναι συνεπής (consistent test) και
- ο έλεγχος πρέπει να είναι πιο αποτελεσματικός έναντι των υπόλοιπων ελέγχων (efficient test).

Παρακάτω δίνουμε τους σχετικούς ορισμούς.

Ορισμός 1.34: Αμερόληπτος Έλεγχος

Ένας έλεγχος λέμε ότι είναι αμερόληπτος, αν η πιθανότητα απόρριψης της H_0, όταν η H_1 είναι αληθής, είναι πάντοτε μεγαλύτερη ή ίση της πιθανότητας απόρριψης της H_0, όταν η H_0 είναι αληθής.

Άρα, η ισχύς ενός αμερόληπτου ελέγχου είναι πάντοτε τουλάχιστον ίση με το επίπεδο σημαντικότητας. Επίσης, ένας έλεγχος ο οποίος δεν είναι αμερόληπτος καλείται μεροληπτικός (biased test).

Ορισμός 1.35: Συνεπής Έλεγχος

Μια ακολουθία ελέγχων υποθέσεων είναι συνεπής έναντι όλων των εναλλακτικών υποθέσεων H_1, αν η ισχύς του ελέγχου προσεγγίζει το 1, καθώς το $n \to \infty$, για κάθε δυνατή προκαθορισμένη εναλλακτική υπόθεση H_1. Το επίπεδο σημαντικότητας κάθε ελέγχου στην ακολουθία είναι πάντοτε μικρότερο ή ίσο μιας προκαθορισμένης τιμής $a \in (0,1)$.

[2] A Litany of Problems With p-values, https://www.fharrell.com/post/pval-litany/

Θα πρέπει να αναφέρουμε πως ο όρος *συνεπής έλεγχος* αφορά μια ακολουθία ελέγχων υποθέσεων, όπου κάθε έλεγχος αφορά δείγμα μεγέθους n, το οποίο καθώς αυξάνεται προσεγγίζει το μέγεθος του πληθυσμού. Για λόγους ευκολίας έχουμε υποθέσει ότι ο πληθυσμός είναι «απείρου» πλήθους στοιχείων και άρα $n \to \infty$.

Η τρίτη ιδιότητα ενός καλού ελέγχου υποθέσεων είναι αυτή της αποτελεσματικότητας και χρησιμοποιείται για να συγκρίνει δύο ελέγχους υπό παρόμοιες συνθήκες. Στη συνέχεια, δίνουμε τον σχετικό ορισμό, όπου για λόγους ευκολίας έχουμε θεωρήσει την περίπτωση που η H_1 είναι απλή υπόθεση και όχι σύνθετη.

Ορισμός 1.36: Αποτελεσματικός Έλεγχος

Έστω T_1, T_2 δύο στατιστικές συναρτήσεις ελέγχου (σ.σ.ε.) οι οποίες χρησιμοποιούνται για τον έλεγχο της $H_0 : \theta = \theta_0$ έναντι της $H_1 : \theta = \theta_1$, με $\theta_1 \neq \theta_0$. Έστω επίσης ότι για κάθε έλεγχο, το μέγεθός του είναι ίσο με a και το σφάλμα τύπου II είναι ίσο με β. Τότε η σχετική αποδοτικότητα (relative efficiency) της σ.σ.ε. T_1 ως προς τη σ.σ.ε. T_2 είναι ο λόγος n_2/n_1, όπου n_1, n_2 είναι τα μεγέθη δείγματος για τους αντίστοιχους ελέγχους με χρήση των T_1, T_2.

Αν $n_1 < n_2$, τότε η αποδοτικότητα της T_1 ως προς την T_2 είναι μεγαλύτερη του 1 και άρα ο έλεγχος που βασίζεται στην T_1 είναι αποτελεσματικότερος του ελέγχου που βασίζεται στην T_2, αφού για το ίδιο επίπεδο σημαντικότητας, απαιτεί μικρότερο μέγεθος δείγματος ώστε να έχει την ίδια ισχύ με τον έλεγχο που βασίζεται στην T_2.

Ένα μειονέκτημα στη χρήση της σχετικής αποδοτικότητας είναι ότι εξαρτάται από τις τιμές a, β, αλλά και από το αν η εναλλακτική υπόθεση H_1 είναι απλή ή σύνθετη. Στην περίπτωση που είναι σύνθετη, θα πρέπει η σχετική αποτελεσματικότητα να υπολογιστεί για κάθε δυνατή τιμή της παραμέτρου θ στον παραμετρικό χώρο που ορίζεται υπό την H_1. Για τον λόγο αυτό, προκειμένου να συγκρίνουμε δύο ελέγχους υποθέσεων, προτείνεται ο υπολογισμός της ασυμπτωτικής σχετικής αποτελεσματικότητας (asymptotic relative efficiency, ARE).

Ορισμός 1.37: Ασυμπτωτική Σχετική Αποτελεσματικότητα

Έστω T_1, T_2 δύο στατιστικές συναρτήσεις ελέγχου οι οποίες χρησιμοποιούνται για τον ίδιο έλεγχο με μηδενική υπόθεση H_0 και εναλλακτική υπόθεση H_1. Για την ακολουθία των ελέγχων (με καθένα από τα δύο τεστ) υποθέτουμε ότι αυτή είναι συνεπής και το επίπεδο σημαντικότητας είναι a. Έστω επίσης n_1, n_2 τα μεγέθη δείγματος τα οποία απαιτούνται ώστε οι στατιστικές συναρτήσεις ελέγχου T_1, T_2 να έχουν την ίδια ισχύ. Τότε, η ασυμπτωτική σχετική αποτελεσματικότητα (asymptotic relative efficiency, A.R.E.) της σ.σ.ε. T_1 ως προς τη σ.σ.ε. T_2 είναι το όριο του n_2/n_1, καθώς το $n_1 \to \infty$ και η $H_0 \to H_1$.

Συνήθως, η ασυμπτωτική σχετική αποτελεσματικότητα (ARE) αποτελεί ικανοποιητική προσέγγιση της σχετικής αποτελεσματικότητας/αποδοτικότητας σε πολλά πρακτικά προβλήματα.

Κλείνουμε την παρούσα ενότητα με τον ορισμό του συντηρητικού ελέγχου σε αντιστοιχία με τον ορισμό του συντηρητικού διαστήματος εμπιστοσύνης.

Ορισμός 1.38: Συντηρητικός Έλεγχος

Ένας έλεγχος ονομάζεται *συντηρητικός* (conservative) αν το πραγματικό επίπεδο σημαντικότητας είναι μικρότερο από το επιθυμητό.

1.9 Εισαγωγή στη μη παραμετρική στατιστική

Όπως είδαμε και στις προηγούμενες ενότητες, το πλαίσιο παρουσίασης και εφαρμογής των βασικών διαδικασιών της στατιστικής συμπερασματολογίας, όπως π.χ. εύρεση σημειακών εκτιμητών, εύρεση διαστημάτων εμπιστοσύνης, έλεγχοι υποθέσεων κ.λπ., είναι παραμετρικό. Δηλαδή, για την εφαρμογή τους,

βασιζόμαστε σε συγκεκριμένες -ίσως και περιοριστικές- υποθέσεις για τη μορφή της κατανομής του πληθυσμού από τον οποίο λαμβάνεται το δείγμα. Ενδεικτικά, στην πολύ γενική περίπτωση, υποθέτουμε ότι το τυχαίο δείγμα $X_1, X_2, \dots, X_n$ είναι από πληθυσμό ο οποίος μοντελοποιείται σύμφωνα με μια κατανομή πιθανότητας με α.σ.κ. $F(x;\theta)$. Για την $F(x;\theta)$ γνωρίζουμε τη μορφή της και είναι άγνωστες οι παράμετροί της, όπως π.χ. η μέση τιμή και η διασπορά.

Από τις περιοριστικές υποθέσεις που τίθενται για τη μορφή της κατανομής του υπό μελέτη χαρακτηριστικού, η πιο συνηθισμένη είναι η υπόθεση της κανονικότητας. Δηλαδή υποθέτουμε ότι ο πληθυσμός, από τον οποίο επιλέξαμε το δείγμα, μοντελοποιείται σύμφωνα με το πρότυπο της κανονικής κατανομής. Μια άλλη υπόθεση είναι αυτή της ομοσκεδαστικότητας, δηλαδή η υπόθεση ότι η μεταβλητότητα μεταξύ δύο ή περισσότερων υποπληθυσμών μπορεί να θεωρηθεί ίδια. Επίσης, όπως είδαμε, υπάρχουν και άλλα παραμετρικά μοντέλα πιθανότητας, όπως π.χ. η Εκθετική κατανομή, η κατανομή Weibull, η κατανομή Poisson, η Ομοιόμορφη κατανομή. Άρα, κατά την εφαρμογή παραμετρικών μεθόδων στατιστικής ανάλυσης, θα πρέπει να επιλεχθεί το κατάλληλο μοντέλο το οποίο περιγράφει ικανοποιητικά τα διαθέσιμα δεδομένα. Η επιλογή αυτή δεν είναι πάντοτε εύκολη, ενώ σε περίπτωση επιλογής ενός μοντέλου που δεν περιγράφει ικανοποιητικά τα διαθέσιμα δεδομένα και, επομένως, αποκλίνει από το άγνωστο πραγματικό μοντέλο, υπεισέρχονται επιπλέον σφάλματα (misspecification errors).

Έτσι, υποθέσεις οι οποίες προσδιορίζουν τη μορφή της κατανομής των παρατηρήσεων ονομάζονται παραμετρικές, τα αντίστοιχα κριτήρια που χρησιμοποιούνται για τον έλεγχο αυτών των υποθέσεων, οι οποίες σχετίζονται με τις παραμέτρους της κατανομής που έχουμε υποθέσει για τον πληθυσμό, ονομάζονται παραμετρικά κριτήρια, ενώ και οι αντίστοιχες μέθοδοι και τεχνικές στατιστικής συμπερασματολογίας που εφαρμόζονται σε αυτές τις περιπτώσεις αποτελούν μέρος αυτού που ονομάζουμε παραμετρική στατιστική συμπερασματολογία. Όλες οι τεχνικές, που έχουμε αναφέρει έως τώρα, ανήκουν σε αυτήν την κατηγορία.

Όμως, σε πολλά πραγματικά προβλήματα οι υποθέσεις για τη μορφή της κατανομής του πληθυσμού από τον οποίο συλλέξαμε το δείγμα, δεν είναι εύκολο να αιτιολογηθούν. Για παράδειγμα, για τη στοχαστική μοντελοποίηση των τιμών ενός χαρακτηριστικού, οι οποίες τιμές εμφανίζουν συμμετρία, ένα εύλογο ερώτημα που προκύπτει είναι γιατί πρέπει να επιλεχθεί η κανονική κατανομή και όχι κάποια άλλη συμμετρική κατανομή.

Επίσης, ακόμη και αν η μορφή της κατανομής είναι γνωστή, αυτό δεν συνεπάγεται απαραίτητα πως μπορούν άμεσα να χρησιμοποιηθούν οι συνήθεις μεθοδολογίες της παραμετρικής στατιστικής συμπερασματολογίας. Για παράδειγμα, ενδέχεται να μην είναι δυνατή η εξαγωγή αναλυτικών αποτελεσμάτων, η διεξαγωγή ελέγχων στο επιθυμητό ε.σ. ή ακόμη και η αναλυτική εύρεση διαστημάτων εμπιστοσύνης.

Η χρήση προσεγγιστικών μεθόδων αντιμετωπίζει πολλά από τα προβλήματα που αναφέρθηκαν, όμως η ορθή εφαρμογή των μεθόδων αυτών δεν είναι πάντοτε δυνατή. Για παράδειγμα, με εφαρμογή του Κεντρικού Οριακού Θεωρήματος μπορούμε να βρούμε ένα διάστημα εμπιστοσύνης για τη μέση τιμή ενός πληθυσμού, βασιζόμενοι στις ιδιότητες της κανονικής κατανομής. Αν και ο εμπειρικός κανόνας αναφέρει ότι το μέγεθος του δείγματος αρκεί να είναι τουλάχιστον 30, στην πράξη απαιτείται πολύ μεγαλύτερο μέγεθος δείγματος για να επιτύχουμε την επιθυμητή τιμή για τον συντελεστή εμπιστοσύνης. Θα πρέπει, επίσης, να σημειωθεί πως σε πρακτικά προβλήματα δεν είναι καθόλου βέβαιο ότι μπορούμε να λάβουμε αρκετά μεγάλο μέγεθος δείγματος. Πάντως, αν και για το διάστημα εμπιστοσύνης για τη μέση τιμή υπάρχουν λύσεις με τη βοήθεια της ασυμπτωτικής θεωρίας, για τον έλεγχο της ισότητας των διακυμάνσεων δύο ανεξάρτητων πληθυσμών, η απόκλιση από την υπόθεση της κανονικότητας οδηγεί, συνήθως, σε μη αξιόπιστα αποτελέσματα.

Θα θέλαμε, λοιπόν, να μπορούμε να χρησιμοποιήσουμε μια στατιστική μεθοδολογία η οποία να είναι ανθεκτική σε αποκλίσεις από τις βασικές θεωρητικές, και συνήθως, παραμετρικές υποθέσεις. Στο σημείο αυτό, δίνουμε τον ορισμό της ανθεκτικότητας (robustness) ή ευρωστίας μιας στατιστικής διαδικασίας.

Ορισμός 1.39

Μια στατιστική διαδικασία ονομάζεται ανθεκτική ή εύρωστη όταν η ισχύς της δεν επηρεάζεται σημαντικά από αποκλίσεις από τις βασικές (θεωρητικές) προϋποθέσεις εφαρμογής της.

Στην πράξη, είναι αρκετά συχνές οι περιπτώσεις όπου δεν μπορούμε να επιλέξουμε την κανονική κατανομή ως το πλέον κατάλληλο μοντέλο για τη μοντελοποίηση των διαθέσιμων δεδομένων. Για παράδειγμα, υπάρχουν σαφείς ενδείξεις απόκλισης από το μοντέλο της συμμετρίας ή υπάρχει ένα φυσικό σύνορο για τις πιθανές τιμές του χαρακτηριστικού (τ.μ.) X (π.χ. λαμβάνει τιμές στο $[0, \infty)$). Επίσης, όταν έχουμε στη διάθεσή μας κατηγορικά δεδομένα, είτε αυτά είναι ονομαστικά είτε διατάξιμα, τότε, προφανώς, δεν μπορούμε να υποθέσουμε κανονικότητα. Γίνεται, λοιπόν, αντιληπτό ότι απαιτούνται άλλες μέθοδοι στατιστικής συμπερασματολογίας, οι οποίες να μην βασίζονται σε ιδιαίτερα δεσμευτικές υποθέσεις για την κατανομή του πληθυσμού από τον οποίο λαμβάνονται οι παρατηρήσεις.

Οι μέθοδοι αυτές καλούνται *ελεύθερες κατανομής* ή *απαλλαγμένες παραμέτρων* (distribution free methods), αφού η ισχύς τους δεν εξαρτάται από τη μορφή του παραμετρικού μοντέλου το οποίο περιγράφει τη συμπεριφορά των τιμών του χαρακτηριστικού X (βλ., μεταξύ άλλων, Stuart κ.ά., 2010). Ουσιαστικά, οι μέθοδοι, που θεωρούνται απαλλαγμένες παραμέτρων, εφαρμόζονται για μια ευρεία κατηγορία κατανομών, όπως π.χ. σε όλες τις συνεχείς κατανομές. Η περιοχή της Στατιστικής που ασχολείται με τέτοια προβλήματα ονομάζεται **Μη Παραμετρική Στατιστική** ή **Απαραμετρική Στατιστική** (Nonparametric Statistics) και τα αντίστοιχα στατιστικά κριτήρια ονομάζονται μη παραμετρικά. Στο σημείο αυτό πρέπει να σημειώσουμε ότι οι έννοιες «μη παραμετρική» (nonparametric) και «ελεύθερες κατανομής»/«απαλλαγμένες παραμέτρων» (distribution-free) δεν είναι ακριβώς ταυτόσημες. Ο όρος «μη παραμετρική» αναφέρεται στην περιγραφή του προβλήματος, ενώ το «ελεύθερες κατανομής»/«απαλλαγμένες παραμέτρων» στη μέθοδο που χρησιμοποιείται για τη λύση του προβλήματος. Στη συνέχεια, για λόγους απλούστευσης, θα χρησιμοποιούμε τις δύο αυτές έννοιες ισοδύναμα. Τέλος, θα ήταν παράλειψη να μην αναφερθεί αυτό που επισημαίνεται από τον Wasserman (2006), ότι ένα καλύτερο όνομα για τη μη παραμετρική συμπερασματολογία μπορεί να είναι η «μη πεπερασμένων διαστάσεων συμπερασματολογία», υπονοώντας ότι το πλήθος των «παραμέτρων» ενός μη παραμετρικού μοντέλου είναι μη πεπερασμένο.

Ως παραδείγματα εφαρμογής των μη παραμετρικών τεχνικών μπορούμε να θεωρήσουμε (α) τον έλεγχο για τη διάμεση επίδοση σε έναν έλεγχο δεξιοτήτων με χρήση σκορ σε διάφορες επιμέρους δοκιμασίες (μη παραμετρικός έλεγχος υπόθεσης), (β) την εκτίμηση με διάστημα της διαμέσου ενός πληθυσμού με άγνωστη κατανομή (εύρεση μη παραμετρικού διαστήματος εμπιστοσύνης), (γ) τον έλεγχο για την ισότητα της κατανομής του χρόνου ζωής εξαρτημάτων τα οποία παράγονται από δύο διαφορετικές γραμμές παραγωγής (μη παραμετρικός έλεγχος ισότητας κατανομών) και (δ) τον έλεγχο της υπόθεσης ότι τα διαθέσιμα δεδομένα προέρχονται από μία συγκεκριμένη κατανομή με γνωστές (ή άγνωστες) παραμέτρους (έλεγχος καλής προσαρμογής).

Οι περιπτώσεις (α)-(δ) που αναφέρθηκαν παραπάνω είναι ενδεικτικές και δεν εξαντλούν το πού και πώς μπορούν να εφαρμοστούν μη παραμετρικές στατιστικές τεχνικές. Επίσης, για κάποιες από αυτές υπάρχουν και παραμετρικές μεθοδολογίες με τους περιορισμούς που έχουμε ήδη αναφέρει ως προς την εφαρμογή τους. Γεννάται, λοιπόν, το ερώτημα γιατί να προτιμηθούν οι μη παραμετρικές έναντι των παραμετρικών τεχνικών. Καταρχάς, οι μη παραμετρικές τεχνικές χαρακτηρίζονται από την απλότητα κατασκευής τους. Σε πολλές από αυτές, κυρίως στην ανάπτυξη και εφαρμογή ελέγχων υποθέσεων, απαιτούνται γνώσεις συνδυαστικής σε συνδυασμό με τη χρήση απλών αριθμητικών πράξεων, όπως απαρίθμησης, πρόσθεσης, ταξινόμησης. Επίσης, είναι άμεση η εφαρμογή τους για μικρά δείγματα ($n \leqslant 30$), ωστόσο, καθώς το μέγεθος δείγματος n αυξάνεται, απαιτείται περισσότερος χρόνος για την εφαρμογή τους. Πλέον όμως, με τη χρήση των σύγχρονων υπολογιστών, ακόμη και αυτό το πρόβλημα είναι άμεσα αντιμετωπίσιμο.

Σε ότι αφορά την αποτελεσματικότητα των μη παραμετρικών τεχνικών, οι τεχνικές οι οποίες είναι απαλλαγμένες παραμέτρων είναι πιο αποτελεσματικές όταν εφαρμόζονται υπό ένα μη παραμετρικό

πλαίσιο. Όμως, όταν εφαρμόζονται υπό ένα παραμετρικό πλαίσιο, τότε οι μη παραμετρικές τεχνικές είναι λιγότερο αποτελεσματικές, ειδικά για μικρά δείγματα. Η αποτελεσματικότητά τους σε αυτές τις περιπτώσεις αυξάνεται, καθώς αυξάνεται το μέγεθος δείγματος n. Ένας τρόπος για να συγκρίνουμε την αποτελεσματικότητα παραμετρικών και μη παραμετρικών τεχνικών, όπως π.χ. στην περίπτωση ελέγχων υποθέσεων, είναι να χρησιμοποιήσουμε το μέτρο της ασυμπτωτικής σχετικής αποτελεσματικότητας.

Ως προς τη χρησιμότητα των μη παραμετρικών μεθόδων σε πρακτικά προβλήματα, όταν το πιθανοτικό πρότυπο απέχει σημαντικά από τη συμμετρία και τα μεγέθη δείγματος που λαμβάνονται από κάθε πληθυσμό δεν είναι ιδιαίτερα μεγάλα, τότε δεν μπορεί να εφαρμοστεί το Κεντρικό Οριακό Θεώρημα και συνήθεις έλεγχοι, όπως το t-test της ισότητας δύο πληθυσμιακών μέσων τιμών ή το F-test της ισότητας δύο πληθυσμιακών διακυμάνσεων, δεν είναι σωστό να εφαρμοστούν, καθώς είναι πάρα πολύ πιθανό να οδηγήσουν σε εσφαλμένα συμπεράσματα. Αξίζει, επίσης, να σημειωθεί πως σε πολλές περιπτώσεις, λόγω της διακριτής φύσης των περισσότερων μη παραμετρικών ελέγχων, δεν επιτυγχάνεται σχεδόν ποτέ το επιθυμητό επίπεδο σημαντικότητας a. Συνήθως, συμβιβαζόμαστε με το να είναι μικρότερο ή ίσο του a και άρα καταλήγουμε σε ένα *συντηρητικό* κριτήριο.

Για την εφαρμογή μη παραμετρικών στατιστικών τεχνικών, αρκεί να υποθέσουμε ότι ο πληθυσμός είναι συνεχής και σε κάποιες περιπτώσεις και συμμετρικός, χωρίς όμως να υποθέσουμε κάποια συγκεκριμένη μορφή για την κατανομή των δεδομένων. Επιπλέον, και αυτό είναι ένα σημαντικό πλεονέκτημα των μη παραμετρικών μεθόδων έναντι των παραμετρικών, οι μη παραμετρικές στατιστικές τεχνικές μπορούν να εφαρμοστούν σε προβλήματα όπου τα δεδομένα είναι ποιοτικά, τουτέστιν διατάξιμα ή ονομαστικά, ή όταν είναι ομαδοποιημένα σε πίνακα συχνοτήτων. Για την ορθή εφαρμογή των παραμετρικών μεθόδων χρειάζονται οι αρχικές μετρήσεις, δηλαδή τα πρωτογενή δεδομένα (raw data). Ανάλογα με το είδος αυτών, για κάθε παραμετρική μέθοδο υπάρχει και η αντίστοιχη μη παραμετρική, τουλάχιστον για τις πλέον συχνά χρησιμοποιούμενες σε πρακτικά προβλήματα. Επιπροσθέτως, και σε αντιστοιχία με τις παραμετρικές στατιστικές τεχνικές, υπάρχουν μη παραμετρικές στατιστικές τεχνικές για δεδομένα (δείγμα) από έναν πληθυσμό, για δύο ή περισσότερα ανεξάρτητα ή εξαρτημένα δείγματα από τους αντίστοιχους πληθυσμούς.

Τέλος, κατάλληλες μη παραμετρικές μέθοδοι μπορούν να χρησιμοποιηθούν πριν την (ορθή) εφαρμογή μιας παραμετρικής μεθόδου. Για παράδειγμα, προτού εφαρμόσουμε έναν έλεγχο t-test για τη μέση τιμή ενός πληθυσμού, μπορούμε να κάνουμε έναν έλεγχο, ώστε να διαπιστώσουμε, αν τα δεδομένα προέρχονται από κανονική κατανομή. Στη συνέχεια, και ανάλογα με το αποτέλεσμα του ελέγχου, μπορούμε να εφαρμόσουμε το t-test -αν μπορούμε να υποθέσουμε κανονικότητα- ή να προχωρήσουμε με την αντίστοιχη μη παραμετρική μέθοδο.

Δεν είναι, λοιπόν, δύσκολο να διαπιστώσουμε ότι η μη παραμετρική στατιστική έχει, εκτός από θεωρητικό ενδιαφέρον για την περαιτέρω ανάπτυξη και βελτίωση των ήδη υπαρχουσών τεχνικών, πληθώρα εφαρμογών σε πολλά πρακτικά προβλήματα. Στα επόμενα κεφάλαια, θα γίνει αναλυτική παρουσίαση ενός μεγάλου μέρους των πιο γνωστών μεθόδων μη παραμετρικής στατιστικής. Στο σημείο αυτό, θα ήταν παράλειψη να μην αναφέρουμε ότι στη βιβλιογραφία υπάρχει πληθώρα συγγραμμάτων με αντικείμενο τις μη παραμετρικές μεθοδολογίες, εκ των οποίων κάποια έχουν εμφανώς επηρεάσει τον τρόπο παρουσίασης και ανάλυσης της ύλης αυτού του βιβλίου. Ενδεικτικά αναφέρουμε τα ξενόγλωσσα συγγράμματα των Conover (1998), Kvam and Vidakovic (2007), Gibbons and Chakraborti (2020), Sprent and Smeeton (2007), Sprent (1999), Wasserman (2006), Sheskin (2011), Härdle (1990), Chakraborti and Graham (2019), ενώ από την ελληνική βιβλιογραφία ιδιαίτερη μνεία οφείλουμε να κάνουμε στο σύγγραμμα της Ξεκαλάκη (2001), στις διδακτικές σημειώσεις των Ιωαννίδης (2018) και Καρλής (2004), καθώς και στα σχετικά κεφάλαια των βιβλίων των Δαμιανού και Κούτρας (1998) και Παπαϊωάννου και Λουκάς (2002).

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Δαμιανού, Χ. και Κούτρας, Μ. (1998). *Εισαγωγή στη Στατιστική, Μέρος ΙΙ*. Αθήνα: Εκδόσεις Συμμετρία.
Ζωγράφος, Κ. (2008). *Πιθανότητες*. Ιωάννινα: Πανεπιστήμιο Ιωαννίνων.
Ηλιόπουλος, Γ. (2012). *Βασικές Μέθοδοι Εκτίμησης Παραμέτρων*. Αθήνα: Εκδόσεις Σταμούλη.
Ιωαννίδης, Ε. (2018). *Σημειώσεις Μη Παραμετρικής Στατιστικής*. Αθήνα: Ο.Π.Α.
Καρλής, Δ. (2004). *Υπολογιστική Στατιστική*. Οικονομικό Πανεπιστήμιο Αθηνών.
Κούτρας, Μ. (2018). *Εισαγωγή στη Θεωρία Πιθανοτήτων και Εφαρμογές*. Αθήνα: Εκδόσεις Τσότρας.
Κουτρουβέλης, Ι. (1999a). *Πιθανότητες και Στατιστική Ι*. Πάτρα: Ελληνικό Ανοικτό Πανεπιστήμιο.
Κουτρουβέλης, Ι. (1999b). *Πιθανότητες και Στατιστική ΙΙ*. Πάτρα: Ελληνικό Ανοικτό Πανεπιστήμιο.
Ξεκαλάκη, Ε. (2001). *Μη Παραμετρική Στατιστική*. Αθήνα: Εκδόσεις Μπένου.
Παπαϊωάννου, Τ. και Λουκάς, Σ. (2002). *Εισαγωγή στη Στατιστική*. Αθήνα: Εκδόσεις Σταμούλη.
Παπαϊωάννου, Τ. και Φερεντίνος, Κ. (2000). *Μαθηματική Στατιστική*. Αθήνα: Εκδόσεις Σταμούλης.
Χαραλαμπίδης, Χ. (2000a). *Θεωρία Πιθανοτήτων και Εφαρμογές, Τεύχος Ι*. Αθήνα: Εκδόσεις Συμμετρία.
Χαραλαμπίδης, Χ. (2000b). *Θεωρία Πιθανοτήτων και Εφαρμογές, Τεύχος ΙΙ*. Αθήνα: Εκδόσεις Συμμετρία.

Ξενόγλωσση

Chakraborti, S. and Graham, M. (2019). *Nonparametric Statistical Process Control*. John Wiley and Sons.
Conover, W. J. (1998). *Practical Nonparametric Statistics* (3rd ed.). Wiley and Sons, Inc.
Gibbons, J. D. and Chakraborti, S. (2020). *Nonparametric Statistical Inference, Fourth Edition Revised and Expanded*. Chapman and Hall/CRC.
Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.
Kvam, P. and Vidakovic, B. (2007). *Nonparametric Statistics with applications to science and engineering*. Wiley Series in Probability and Statistics.
Sheskin, D. (2011). *Handbook of Parametric and Non-parametric Procedures* (5th ed.). Chapman and Hall/CRC.
Siegel, A. (2011). *Practical Business Statistics*. Elsevier Science.
Sprent, P. (1999). *Applied Nonparametric Statistical Methods*. Chapmann and Hall.
Sprent, P. and Smeeton, N. (2007). *Applied Nonparametric Statistical Methods* (4th ed.). Chapmann and Hall.
Stuart, A., Ord, K. and Arnold, S. (2010). *Kendall's Advanced Theory of Statistics, Volume 2A, Classical Inference and the Linear Model*. Wiley.
Wasserman, L. (2006). *All of Nonparametric Statistics*. New York, NY: Springer Texts in Statistics.
Wasserstein, R. L. and Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), pp. 129–133.

ΚΕΦΑΛΑΙΟ 2

ΜΗ ΠΑΡΑΜΕΤΡΙΚΗ ΕΚΤΙΜΗΣΗ ΤΗΣ ΑΘΡΟΙΣΤΙΚΗΣ ΣΥΝΑΡΤΗΣΗΣ ΚΑΤΑΝΟΜΗΣ ΚΑΙ ΣΥΝΑΡΤΗΣΙΑΚΩΝ ΤΗΣ

Σύνοψη

Στο κεφάλαιο αυτό παρουσιάζονται μέθοδοι μη παραμετρικής εκτίμησης της αθροιστικής συνάρτησης κατανομής (α.σ.κ.) F, καθώς και συναρτησιακών της (statistical functionals), δηλαδή στατιστικών συναρτήσεων της α.σ.κ., όπως είναι η μέση τιμή, η διακύμανση και η συσχέτιση. Στο πλαίσιο αυτό, αρχικά δίνονται ο ορισμός της εμπειρικής αθροιστικής συνάρτησης κατανομής (ε.α.σ.κ) και κάποιες χρήσιμες ιδιότητές της, όπως η μεροληψία, η ιδιότητα της συνέπειας, η ασυμπτωτική κανονικότητα και το Θεώρημα των Glivenko-Cantelli. Τέλος, δίνονται θεωρητικά αποτελέσματα που αφορούν τη μη παραμετρική εκτίμηση συναρτησιακών της α.σ.κ., καθώς και εφαρμογές αυτών.

Προαπαιτούμενη γνώση:
Βασικές γνώσεις Πιθανοτήτων και Στατιστικής.

Προσδοκώμενα μαθησιακά αποτελέσματα:
Ο/η φοιτητής/τρια ολοκληρώνοντας την ενότητα αυτή θα μπορεί να εκτιμά μη παραμετρικά τη συνάρτηση κατανομής και να υπολογίζει ζώνες εμπιστοσύνης, αλλά και ασυμπτωτικά διαστήματα εμπιστοσύνης για αυτήν.

Γλωσσάριο επιστημονικών όρων

- Αθροιστική συνάρτηση κατανομής
- Ανισότητα Dvoretzky–Kiefer–Wolfowitz
- Γραμμικό συναρτησιακό
- Δειγματικό ποσοστιαίο σημείο
- Εκτιμητής αντικατάστασης
- Εμπειρική αθροιστική συνάρτηση κατανομής
- Ζώνη εμπιστοσύνης
- Μη παραμετρική μέθοδος Δέλτα
- Στατιστικό συναρτησιακό
- Συνάρτηση επιρροής

Μπατσίδης, Α., Παπασταμούλης, Π., Πετρόπουλος, Κ., & Ρακιτζής, Α. (2022). *Μη Παραμετρική Στατιστική.* [Προπτυχιακό εγχειρίδιο].

2.1 Εκτίμηση της αθροιστικής συνάρτησης κατανομής

Σε αυτήν την ενότητα θα ασχοληθούμε με την εκτίμηση της αθροιστικής συνάρτησης κατανομής. Έστω $X_1, X_2, \dots, X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής (α.σ.κ.) $F(\cdot)$. Εκτιμούμε την F με την εμπειρική αθροιστική συνάρτηση κατανομής, ο ορισμός της οποίας ακολουθεί.

Ορισμός 2.1

Έστω $X_1, X_2, \dots, X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής (α.σ.κ.) $F(\cdot)$. Η **εμπειρική αθροιστική συνάρτηση κατανομής** συμβολίζεται με $F_n(x)$ και ορίζεται ως εξής:

$$F_n(x) = \frac{\text{πλήθος των } X_i \leqslant x}{n}, \quad x \in \mathbb{R},$$

ή ισοδύναμα

$$F_n(x) = \frac{\sum_{i=1}^{n} I_{(-\infty, x]}(X_i)}{n}, x \in \mathbb{R},$$

όπου

$$I_C(X_i) = \begin{cases} 1, & X_i \in C \\ 0, & X_i \notin C \end{cases}$$

αναφέρεται ως δείκτρια συνάρτηση.

Από τον ορισμό της εμπειρικής αθροιστικής συνάρτησης κατανομής (ή, αλλιώς, εμπειρική συνάρτηση κατανομής) γίνεται άμεσα αντιληπτό ότι πρόκειται για μία στατιστική συνάρτηση, καθώς είναι συνάρτηση των τιμών του δείγματος $X_1, \dots, X_n$. Ο τρόπος υπολογισμού της αποσαφηνίζεται στο επόμενο παράδειγμα.

Παράδειγμα 2.1. Έστω ένα τυχαίο δείγμα $n = 10$ παρατηρήσεων

$$\boldsymbol{x} = (-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14).$$

Να προσδιοριστεί η εμπειρική αθροιστική συνάρτηση κατανομής.

Λύση Παραδείγματος 2.1. Σύμφωνα με τον Ορισμό 2.1, για δοθέν $x \in \mathbb{R}$, η $F_n(x)$ υπολογίζεται μετρώντας το πλήθος των παρατηρήσεων που δεν ξεπερνούν το x και, ακολούθως, διαιρώντας με το μέγεθος δείγματος (εδώ $n = 10$). Για διευκόλυνση στους υπολογισμούς αρχικά διατάσσουμε τις τιμές του δείγματος κατά αύξουσα σειρά μεγέθους, οπότε είναι:

$$(-1.13, -0.90, -0.24, -0.14, -0.08, 0.13, 0.18, 0.71, 1.59, 1.98).$$

Στη συνέχεια, για τον υπολογισμό, για παράδειγμα του $F_n(-0.5)$, σημειώνουμε ότι υπάρχουν δύο παρατηρήσεις, οι δύο πρώτες παρατηρήσεις του διατεταγμένου δείγματος, δηλαδή οι -1.13 και -0.90, οι οποίες δεν ξεπερνούν το $x = -0.5$. Συνεπώς,

$$F_n(-0.5) = \frac{\sum_{i=1}^{10} I_{(-\infty, -0.5]}(x_i)}{10} = \frac{2}{10} = 0.2.$$

Με παρόμοιο τρόπο, αν θεωρήσουμε την τιμή $x = -0.4$, προκύπτει ότι $F_n(-0.4) = 0.2$. Αν, από την άλλη, θεωρήσουμε την τιμή $x = -0.2$, τότε $F_n(-0.2) = 0.3$. Λαμβάνοντας υπόψη όλες τις δυνατές τιμές του x, καταλήγουμε στα αποτελέσματα που δίνονται στον Πίνακα 2.1. □

Παρατήρηση 2.1. Υποθέτοντας ότι δεν υπάρχουν ισοβαθμίες (ties) στο παρατηρηθέν δείγμα, από τον ορισμό της εμπειρικής αθροιστικής συνάρτησης κατανομής και το προηγούμενο παράδειγμα γίνεται σαφές ότι η εμπειρική αθροιστική συνάρτηση κατανομής είναι σταθερή μεταξύ διαδοχικών διατεταγμένων παρατηρήσεων, ενώ αυξάνεται με βήμα $1/n$. Με τον όρο ισοβαθμίες εννοούμε παρατηρήσεις με την ίδια τιμή. Η γραφική παράσταση της εμπειρικής αθροιστικής συνάρτησης κατανομής παρατίθεται στο Σχήμα 2.1.

x	$F_n(x)$
$(-\infty, -1.13)$	0
$[-1.13, -0.90)$	0.1
$[-0.90, -0.24)$	0.2
$[-0.24, -0.14)$	0.3
$[-0.14, -0.08)$	0.4
$[-0.08, 0.13)$	0.5
$[0.13, 0.18)$	0.6
$[0.18, 0.71)$	0.7
$[0.71, 1.59)$	0.8
$[1.59, 1.98)$	0.9
$[1.98, \infty)$	1

Πίνακας 2.1: Υπολογισμός της εμπειρικής συνάρτησης κατανομής για τα δεδομένα του Παραδείγματος 2.1.

Παρατήρηση 2.2. Στην πραγματικότητα, τα δεδομένα του Παραδείγματος 2.1 δημιουργήθηκαν προσομοιώνοντας με χρήση του ελεύθερου λογισμικού `R` από την κατανομή $\mathcal{N}(0,1)$ χρησιμοποιώντας την εντολή `rnorm()`. Αυτό σημαίνει ότι η «πραγματική» συνάρτηση κατανομής είναι η

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \mathrm{d}t, \quad x \in \mathbb{R}.$$

Κατόπιν, οι προσομοιωμένες τιμές στρογγυλοποιήθηκαν σε 2 δεκαδικά ψηφία μέσω της εντολής round(). Ειδικότερα, χρησιμοποιήθηκαν οι παρακάτω εντολές:

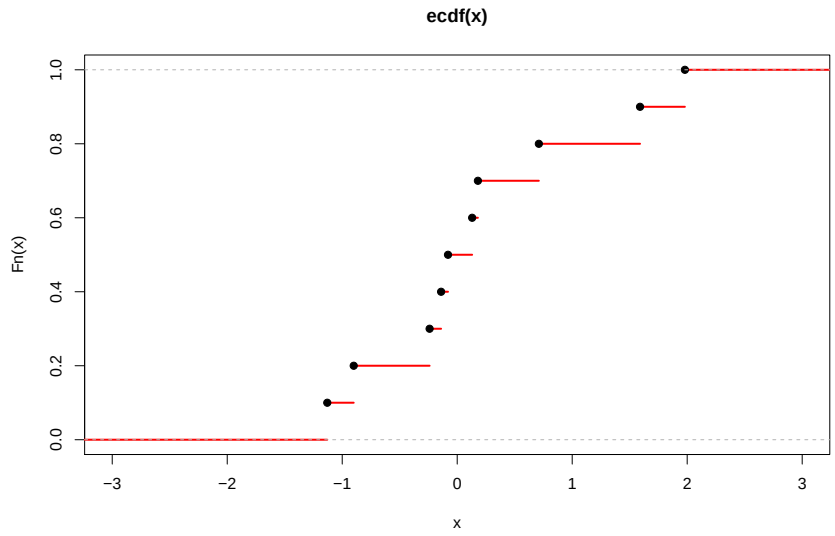
```
> n <- 10
> set.seed(2)
> x <- round(rnorm(n, 0, 1), 2)
> x
 [1] -0.90  0.18  1.59 -1.13 -0.08  0.13  0.71 -0.24  1.98 -0.14
```

Η εντολή `set.seed()` δέχεται ως όρισμα έναν ακέραιο αριθμό (εδώ επιλέξαμε το 2) και αρχικοποιεί τη γεννήτρια τυχαίων αριθμών από την οποία παράγονται τα δεδομένα στη συνέχεια. Αυτό γίνεται ώστε να είναι δυνατή η αναπαραγωγή των αποτελεσμάτων. Στην `R`, για να υπολογίσουμε την εμπειρική αθροιστική συνάρτηση κατανομής, χρησιμοποιούμε την εντολή `ecdf()`. Κατόπιν, μπορούμε να υπολογίσουμε την τιμή της για οποιαδήποτε τιμή. Ακολουθεί η παράθεση των εντολών που επιτρέπουν τον υπολογισμό, για παράδειγμα, του $F_n(x)$ για $x = -0.5$.

```
> Fn <- ecdf(x)
> Fn(-0.5)
[1] 0.2
```

Τέλος, με τη βοήθεια της `R` και της εντολής `plot()`, μπορούμε να κατασκευάσουμε τη γραφική παράσταση της εμπειρικής αθροιστικής συνάρτησης κατανομής. Αυτό γίνεται ως εξής:

```
> plot(Fn, xlim = c(-3, 3), col.hor = "red")
```



Σχήμα 2.1: Γραφική παράσταση της εμπειρικής αθροιστικής συνάρτησης κατανομής για τα δεδομένα του Παραδείγματος 2.1.

Το αποτέλεσμα των παραπάνω εντολών παρατίθεται στο 2.1. Επισημαίνεται ότι τα ορίσματα `xlim = c(-3, 3)` και `col.hor = 'red'` καθορίζουν το εύρος του οριζόντιου άξονα και το χρώμα της γραφικής παράστασης, αντίστοιχα.

Ακολούθως, παρουσιάζονται κάποιες χρήσιμες ιδιότητες της εμπειρικής αθροιστικής συνάρτησης κατανομής (βλ., μεταξύ άλλων, Gibbons and Chakraborti, 2020).

Θεώρημα 2.1

Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(x)$. Τότε, για σταθερό x, η κατανομή της στατιστικής συνάρτησης $nF_n(x)$, όπου $F_n(x)$ η εμπειρική αθροιστική συνάρτηση κατανομής, είναι η διωνυμική με παραμέτρους n και $F(x)$.

Απόδειξη Θεωρήματος 2.1. Από τον ορισμό της εμπειρικής αθροιστικής συνάρτησης κατανομής εύκολα προκύπτει ότι οι δυνατές της τιμές ανήκουν στο ακόλουθο σύνολο: $\{0, 1/n, 2/n, \ldots, (n-1)/n, 1\}$. Για σταθερό x, η τιμή της εμπειρικής αθροιστικής συνάρτησης κατανομής είναι ίση με k/n, $k = 0, \ldots, n$, αν και μόνο αν k το πλήθος δειγματικές τιμές είναι μικρότερες ή ίσες του x.

Επομένως, ο υπολογισμός της πιθανότητας $P\left(F_n(x) = \frac{k}{n}\right)$ ανάγεται στην εύρεση της πιθανότητας να έχουμε k το πλήθος επιτυχίες σε n δοκιμές ενός διωνυμικού πειράματος, όπου επιτυχία θεωρείται όταν μία δειγματική τιμή X_i είναι ίση ή μικρότερη από τη σταθεροποιημένη τιμή x, με πιθανότητα επιτυχίας $p = P(X_i \leq x) = F(x)$. Άρα,

$$P\left(F_n(x) = \frac{k}{n}\right) = \binom{n}{k} (F(x))^k (1 - F(x))^{n-k}, k = 0, \ldots, n,$$

ή ισοδύναμα

$$P(nF_n(x) = k) = \binom{n}{k} (F(x))^k (1 - F(x))^{n-k}, k = 0, \ldots, n,$$

που ολοκληρώνει την απόδειξη. □

Στο προηγούμενο θεώρημα προσδιορίστηκε η ακριβής κατανομή της στατιστικής συνάρτησης $nF_n(x)$, Στο θεώρημα που ακολουθεί το ενδιαφέρον επικεντρώνεται στη συμπεριφορά της εμπειρικής αθροιστικής συνάρτησης κατανομής, για μεγάλο μέγεθος δείγματος (βλ., μεταξύ άλλων, Gibbons and Chakraborti, 2020).

Θεώρημα 2.2

Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από έναν πληθυσμό με συνεχή αθροιστική συνάρτηση κατανομής $F(x)$ και $F_n(x)$ είναι η εμπειρική αθροιστική συνάρτηση κατανομής. Τότε, η εμπειρική αθροιστική συνάρτηση κατανομής $F_n(x)$

α) είναι συνεπής εκτιμητής της αθροιστικής συνάρτησης κατανομής $F(x)$, και

β) ακολουθεί προσεγγιστικά κανονική κατανομή με μέση τιμή $F(x)$ και διακύμανση $\{F(x)(1-F(x))\}/n$.

Απόδειξη Θεωρήματος 2.2. α) Για να αποδείξουμε ότι ένας εκτιμητής, έστω T_n, μίας παραμέτρου θ είναι συνεπής εκτιμητής αυτής της παραμέτρου, δηλαδή ότι $\lim_{n\to\infty} P(\theta-\varepsilon < T_n < \theta+\varepsilon) = 1$, $\forall\, \theta \in \Theta$, αρκεί να δείξουμε ότι T_n είναι αμερόληπτος (ή ασυμπτωτικά αμερόληπτος) εκτιμητής της παραμέτρου θ, δηλαδή ότι $\mathsf{E}(T_n) = \theta$, $\forall\, \theta \in \Theta$ (ή $\mathsf{E}(T_n) \to \theta$, $n \to \infty$), και, επιπλέον, ότι $\mathrm{Var}(T_n) \to 0$, καθώς το $n \to \infty$.

Οπότε, αρχικά θα δείξουμε ότι η εμπειρική αθροιστική συνάρτηση κατανομής είναι αμερόληπτος εκτιμητής της αθροιστικής συνάρτησης κατανομής. Στο Θεώρημα 2.1 αποδείχτηκε ότι $nF_n(x) \sim B(n, F(x))$. Συνεπώς, άμεσα προκύπτει, από τις ιδιότητες της Διωνυμικής κατανομής και της μέσης τιμής, ότι

$$\mathsf{E}(nF_n(x)) = nF(x) \Leftrightarrow \mathsf{E}(F_n(x)) = F(x).$$

Επιπλέον, έχουμε ότι:

$$\mathrm{Var}(nF_n(x)) = nF(x)(1-F(x)) \Leftrightarrow \mathrm{Var}(F_n(x)) = \frac{F(x)(1-F(x))}{n}$$

και, άρα, $\mathrm{Var}(F_n(x)) \to 0$, καθώς το $n \to \infty$, και η απόδειξη του α) ολοκληρώθηκε.

β) Είναι γνωστό από τον ορισμό της εμπειρικής αθροιστικής συνάρτησης κατανομής ότι:

$$F_n(x) = \frac{\sum_{i=1}^n I_{(-\infty,x]}(X_i)}{n}, x \in R.$$

Τουτέστιν, η στατιστική συνάρτηση $F_n(x)$ είναι ο δειγματικός μέσος των τυχαίων μεταβλητών $I_{(-\infty,x]}(X_i)$, η κατανομή των οποίων είναι Bernoulli με πιθανότητα επιτυχίας $F(x)$, μέση τιμή $F(x)$ και διακύμανση $F(x)(1-F(x))$. Εφαρμόζοντας το Κεντρικό Οριακό Θεώρημα προκύπτει άμεσα ότι:

$$F_n(x) = \frac{\sum_{i=1}^n I_{(-\infty,x]}(X_i)}{n} \xrightarrow{d} \mathcal{N}\left(\frac{\mathsf{E}\left(\sum_{i=1}^n I_{(-\infty,x]}(X_i)\right)}{n}, \frac{\mathrm{Var}\left(\sum_{i=1}^n I_{(-\infty,x]}(X_i)\right)}{n^2}\right),$$

ή ισοδύναμα,

$$F_n(x) = \frac{\sum_{i=1}^n I_{(-\infty,x]}(X_i)}{n} \xrightarrow{d} \mathcal{N}\left(F(x), \frac{F(x)(1-F(x))}{n}\right).$$

□

Από τα παραπάνω προκύπτει ότι, αν κάποιος ενδιαφέρεται να εκτιμήσει την αθροιστική συνάρτηση κατανομής $F(x)$ για κάθε x, τότε θα ήθελε να διαπιστώσει πόσο κοντά είναι η εμπειρική αθροιστική συνάρτηση κατανομής στην αθροιστική συνάρτηση κατανομής. Στο επόμενο θεώρημα θεμελιώνεται ότι με

πιθανότητα 1 η σύγκλιση της $F_n(x)$ στην $F(x)$ είναι ομοιόμορφη στο x ή, διαφορετικά, για μεγάλο μέγεθος δείγματος η προσέγγιση της $F(x)$ από την $F_n(x)$ είναι αρκετά ακριβής.

Θεώρημα 2.3: (Glivenko-Cantelli)

Αν $F_n(x)$ είναι η εμπειρική αθροιστική συνάρτηση κατανομής ενός τυχαίου δείγματος $X_1, \dots, X_n$, από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(x)$, τότε:

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{\text{σ.β.}} 0,$$

δηλαδή $P\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0\right) = 1.$

Το παραπάνω θεώρημα αποδείχτηκε από τον Glivenko (1933) για συνεχή συνάρτηση κατανομής F και από τον Cantelli (1933) για γενική συνάρτηση κατανομής F. Αναλυτική απόδειξη μπορεί να βρεθεί στο σύγγραμμα του Loève (1977), στο οποίο το Θεώρημα των Glivenko-Cantelli αναφέρεται ως το «κεντρικό στατιστικό θεώρημα», ενώ στο βιβλίο του Rènyi (1962) αναφέρεται ως «θεμελιώδες στατιστικό θεώρημα». Παραθέτουμε την απόδειξη του Θεώρηματος των Glivenko-Cantelli, όπως αυτή δίνεται από τον Shaikh (2022)[1]. Καθώς στην απόδειξη χρησιμοποιούνται διάφορα λήμματα, η διατύπωση και η απόδειξη αυτών θα προηγηθούν.

Λήμμα 2.1. Έστω F είναι μία αθροιστική συνάρτηση κατανομής στο $\mathbb{R}$. Για κάθε $\varepsilon > 0$ υπάρχει μια πεπερασμένη διαμέριση του $\mathbb{R}$ της μορφής $-\infty = t_0 < t_1 < \dots < t_{k-1} < t_k = \infty$, τέτοια ώστε για $0 \leqslant j \leqslant k-1$,

$$F(t_{j+1}^-) - F(t_j) \leqslant \varepsilon.$$

Απόδειξη Λήμματος 2.1. Έστω δοθέν $\varepsilon > 0$. Θεωρούμε $t_0 = -\infty$ και για $j \geqslant 0$ ορίζουμε

$$t_{j+1} = \sup\{z : F(z) \leqslant F(t_j) + \varepsilon\}.$$

Αρχικά θα δείξουμε ότι $F(t_{j+1}) \geqslant F(t_j) + \varepsilon$. Υποθέτουμε ότι $F(t_{j+1}) < F(t_j) + \varepsilon$ και θα καταλήξουμε σε άτοπο. Επειδή η F, ως αθροιστική συνάρτηση κατανομής, είναι συνεχής από τα δεξιά, υπάρχει $\delta > 0$ τέτοιο ώστε $F(t_{j+1} + \delta) < F(t_j) + \varepsilon$, το οποίο αντίκειται στον ορισμό του t_{j+1}. Έτσι, μεταξύ t_j και t_{j+1}, η F έχει άλματα μήκους τουλάχιστον ε. Αφού αυτό μπορεί να συμβεί το πολύ για ένα πεπερασμένο πλήθος φορών, η διαμέριση είναι της επιθυμητής μορφής, δηλαδή $-\infty = t_0 < t_1 < \dots < t_{k-1} < t_k = \infty$, με k πεπερασμένο. Επίσης, $F(t_{j+1}^-) \leqslant F(t_j) + \varepsilon$. Για να αποδείξουμε το τελευταίο, σημειώνουμε ότι από τον ορισμό του t_{j+1}, έχουμε ότι $F(t_{j+1} - \delta) \leqslant F(t_j) + \varepsilon$, για κάθε $\delta > 0$. Η απόδειξη ολοκληρώνεται λόγω του ορισμού του $F(t_{j+1}^-)$. □

Λήμμα 2.2. Έστω F_n και F είναι αθροιστικές συναρτήσεις κατανομών στο $\mathbb{R}$, τέτοιες ώστε $F_n(x) \to F(x)$ και $F_n(x^-) \to F(x^-)$, για κάθε $x \in \mathbb{R}$. Τότε:

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0.$$

Απόδειξη Λήμματος 2.2. Έστω δοθέν $\varepsilon > 0$. Πρέπει να δείξουμε ότι υπάρχει $N = N(\varepsilon)$ τέτοιο ώστε για $n > N$ και κάθε $x \in \mathbb{R}$

$$|F_n(x) - F(x)| < \varepsilon.$$

Θεωρούμε μια πεπερασμένη διαμέριση του $\mathbb{R}$ της μορφής $-\infty = t_0 < t_1 < \dots < t_{k-1} < t_k = \infty$, οπότε, λόγω του Λήμματος 2.1, για $0 \leqslant j \leqslant k-1$,

$$F(t_{j+1}^-) - F(t_j) \leqslant \frac{\varepsilon}{2}. \tag{2.1}$$

[1] http://home.uchicago.edu/~amshaikh/webfiles/glivenko-cantelli.pdf

Για οποιοδήποτε $x \in \mathbb{R}$, υπάρχει j τέτοιο ώστε $t_j \leqslant x < t_{j+1}$ και

$$F_n(t_j) \leqslant F_n(x) \leqslant F_n(t_{j+1}^-)$$

$$F(t_j) \leqslant F(x) \leqslant F(t_{j+1}^-)$$

το οποίο συνεπάγεται ότι

$$F_n(t_j) - F(t_{j+1}^-) \leqslant F_n(x) - F(x) \leqslant F_n(t_{j+1}^-) - F(t_j).$$

Από την παραπάνω σχέση προκύπτουν τα εξής:

$$F_n(t_j) - F(t_j) + F(t_j) - F(t_{j+1}^-) \leqslant F_n(x) - F(x)$$

και

$$F_n(t_{j+1}^-) - F(t_{j+1}^-) + F(t_{j+1}^-) - F(t_j) \geqslant F_n(x) - F(x).$$

Οπότε, λόγω της σχέσης (2.1), έχουμε ότι:

$$F_n(t_j) - F(t_j) - \frac{\varepsilon}{2} \leqslant F_n(x) - F(x),$$

και

$$F_n(t_{j+1}^-) - F(t_{j+1}^-) + \frac{\varepsilon}{2} \geqslant F_n(x) - F(x).$$

Για κάθε j, έστω $N_j = N_j(\varepsilon)$ τέτοιο ώστε για $n > N_j$

$$F_n(t_j) - F(t_j) > -\frac{\varepsilon}{2}$$

και έστω $M_j = M_j(\varepsilon)$ τέτοιο ώστε για $n > M_j$

$$F_n(t_j^-) - F(t_j^-) < \frac{\varepsilon}{2}.$$

Έστω $N = \max\limits_{1 \leqslant j \leqslant k} \max\{N_j, M_j\}$, για $n > N$ και κάθε $x \in \mathbb{R}$, έχουμε ότι

$$|F_n(x) - F(x)| < \varepsilon.$$

Η απόδειξη του λήμματος ολοκληρώθηκε. □

Λήμμα 2.3. Υποθέτουμε ότι F_n και F είναι αθροιστικές συναρτήσεις κατανομών στο $\mathbb{R}$ τέτοιες ώστε $F_n(x) \to F(x)$, για κάθε $x \in \mathbb{Q}$. Επίσης, υποθέτουμε ότι $F_n(x) - F_n(x^-) \to F(x) - F(x^-)$ για όλα τα σημεία ασυνέχειας της F. Τότε, για κάθε $x \in \mathbb{R}$, $F_n(x) \to F(x)$ και $F_n(x^-) \to F(x^-)$.

Απόδειξη Λήμματος 2.3. Έστω $x \in \mathbb{R}$. Αρχικά θα αποδείξουμε ότι $F_n(x) \to F(x)$. Έστω $s, t \in \mathbb{Q}$ τέτοια ώστε $s < x < t$. Πρώτα υποθέτουμε ότι x είναι σημείο συνέχειας της F. Αφού $F_n(s) \leqslant F_n(x) \leqslant F_n(t)$ και $s, t \in \mathbb{Q}$, έπεται ότι

$$F(s) \leqslant \liminf_{n\to\infty} F_n(x) \leqslant \limsup_{n\to\infty} F_n(x) \leqslant F(t).$$

Αφού x είναι σημείο συνέχειας της F,

$$\lim_{s\to x^-} F(s) = \lim_{t\to x^+} F(t) = F(x),$$

δηλαδή ισχύει το επιθυμητό αποτέλεσμα. Στη συνέχεια, υποθέτουμε ότι x είναι σημείο ασυνέχειας της F. Σημειώνουμε ότι

$$F_n(s) + F_n(x) - F_n(x^-) \leqslant F_n(x) \leqslant F_n(t).$$

Αφού $s,t \in \mathbb{Q}$ και x είναι σημείο ασυνέχειας της F,

$$F(s) + F(x) - F(x^-) \leqslant \liminf_{n\to\infty} F_n(x) \leqslant \limsup_{n\to\infty} F_n(x) \leqslant F(t).$$

Όμως,

$$\lim_{s\to x^-} F(s) = F(x^-) \quad \text{και} \quad \lim_{t\to x^+} F(t) = F(x),$$

άρα ισχύει το επιθυμητό αποτέλεσμα.

Έπειτα, δείχνουμε ότι $F_n(x^-) \to F(x^-)$. Αρχικά υποθέτουμε ότι x είναι σημείο συνέχειας της F. Αφού $F_n(x^-) \leqslant F_n(x)$, έχουμε ότι:

$$\limsup_{n\to\infty} F_n(x^-) \leqslant \limsup_{n\to\infty} F_n(x) = F(x) = F(x^-).$$

Για οποιοδήποτε $s \in \mathbb{Q}$ τέτοιο ώστε $s < x$, έχουμε $F_n(s) \leqslant F_n(x^-)$, από το οποίο προκύπτει ότι

$$F(s) \leqslant \liminf_{n\to\infty} F_n(x^-).$$

Καθώς

$$\lim_{s\to x^-} F(s) = F(x^-),$$

το επιθυμητό αποτέλεσμα ισχύει. Τώρα υποθέτουμε ότι x είναι σημείο ασυνέχειας της F. Από την υπόθεση, $F_n(x) - F_n(x^-) \to F(x) - F(x^-)$ και από το γεγονός ότι, $F_n(x) \to F(x)$, το επιθυμητό αποτέλεσμα ισχύει. □

Απόδειξη Θεωρήματος 2.3. Αν καταφέρουμε να δείξουμε ότι υπάρχει ένα σύνολο N τέτοιο ώστε $P\{N\} = 0$ και για κάθε $\omega \notin N$

(i) $F_n(x,\omega) \to F(x)$, για κάθε $x \in \mathbb{Q}$, και
(ii) $F_n(x,\omega) - F_n(x^-,\omega) \to F(x) - F(x^-)$ για κάθε σημείο συνέχειας της F,

τότε το αποτέλεσμα θα ισχύει χρησιμοποιώντας τα Λήμματα 2.2 και 2.3.

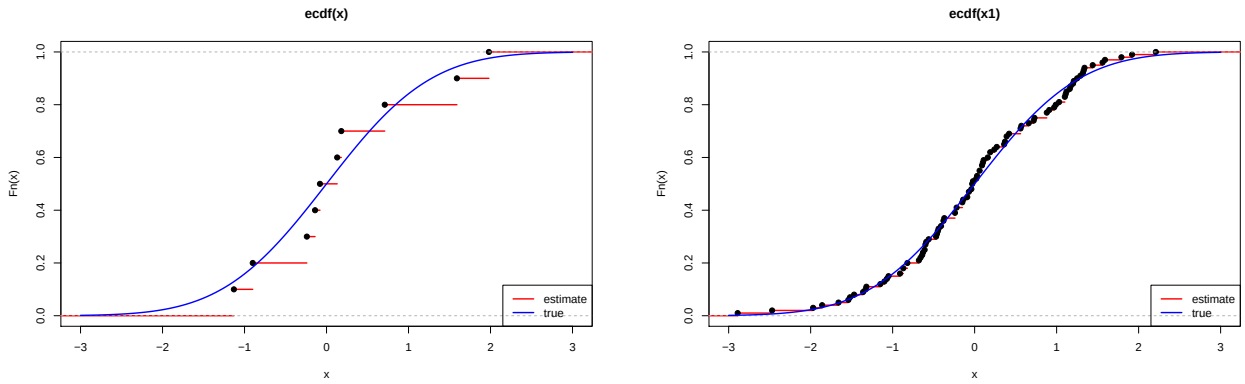
Για κάθε $x \in \mathbb{Q}$, έστω N_x είναι ένα σύνολο τέτοιο ώστε $P\{N_x\} = 0$ και για κάθε $\omega \notin N_x, F_n(x,\omega) \to F(x)$. Έστω $N_1 = \cup_{x\in\mathbb{Q}}$. Τότε, για κάθε $\omega \notin N_1$, $F_n(x,\omega) \to F(x)$ λόγω κατασκευής. Επίσης, αφού $\mathbb{Q}$ είναι μετρήσιμο σύνολο, $P\{N_1\} = 0$.

Για ακέραιο $i \geqslant 1$, ορίζουμε ως J_i το σύνολο των σημείων συνέχειας της F μεγέθους τουλάχιστον $1/i$. Σημειώνουμε ότι για κάθε i, το σύνολο J_i είναι πεπερασμένο. Οπότε, το σύνολο όλων των σημείων συνέχειας της F μπορεί να γραφεί ως $J = \cup_{1\leqslant j<\infty} J_i$. Για κάθε $x \in J$, έστω M_x είναι το σύνολο εκείνο για το οποίο $P\{M_x\} = 0$ και για κάθε $\omega \notin M_x$, $F_n(x,\omega) - F_n(x^-,\omega) \to F(x) - F(x^-)$. Έστω $N_2 = \cup_{x\in J} M_x$. Αφού J είναι ένα μετρήσιμο σύνολο, $P\{N_2\} = 0$.

Για να ολοκληρώσουμε την απόδειξη, έστω $N = N_1 \cup N_2$. Λόγω κατασκευής, για $\omega \notin N_1$, οι συνθήκες (i) και (ii) ισχύουν. Επίσης, $P\{N\} = 0$. Το αποτέλεσμα, πλέον, ισχύει. □

Θα αντιληφθούμε αυτό που περιγράφεται στο Θεώρημα των Glivenko-Cantelli μέσω του παραδείγματος που ακολουθεί, το οποίο αποτελεί συνέχεια του Παραδείγματος 2.1.

Παράδειγμα 2.2 (συνέχεια Παραδείγματος 2.1)**.** Για το τυχαίο δείγμα των 10 τιμών του Παραδείγματος 2.1 σχεδιάστε, με τη βοήθεια της R, την εμπειρική αθροιστική συνάρτηση κατανομής $F_n(x)$ μαζί με τη συνάρτηση κατανομής $\Phi(\cdot)$ της τυπικής κανονικής κατανομής (από την οποία έχουν προσομοιωθεί τα δεδομένα). Στη συνέχεια, με παρόμοιο τρόπο, προσομοιώστε ένα δείγμα, έστω $\boldsymbol{x_1}$ μεγέθους $n = 100$ παρατηρήσεων και επαναλάβετε το ίδιο σχήμα. Τι παρατηρείτε;



(α) $n = 10$ (β) $n = 100$

Σχήμα 2.2: Σύγκριση εμπειρικής αθροιστικής συνάρτησης κατανομής με την πραγματική αθροιστική συνάρτηση κατανομής (η οποία είναι γνωστή λόγω προσομοίωσης των δεδομένων) για διαφορετικές τιμές του μεγέθους δείγματος n.

Λύση Παραδείγματος 2.2. Η εμπειρική αθροιστική συνάρτηση κατανομής $F_n(x)$ του πρώτου δείγματος των 10 παρατηρήσεων μαζί με την αθροιστική συνάρτηση κατανομής της τυπικής κανονικής, $\Phi(x)$, παρατίθεται στο Σχήμα 2.2.(α). Το Σχήμα αυτό προέκυψε εκτελώντας επιπρόσθετα -και χωρίς να κλείσουμε το παράθυρο της R που εμφάνισε το Σχήμα 2.1)- τις παρακάτω εντολές:

```
> xSeq <- seq(-3, 3, length = 1000)
> points(xSeq, pnorm(xSeq),type="l", col = "blue")
> legend("bottomright", c("estimate", "true"), col = c("red", "blue"),
    lty = 1)
```

Αρχικά, σύμφωνα με αυτές τις εντολές, υπολογίσαμε τη συνάρτηση κατανομής $\Phi(x)$ της $\mathcal{N}(0,1)$ για μία ακολουθία 1000 τιμών $x \in [-3,3]$ και, κατόπιν, ενώσαμε τα σημεία $(x,\Phi(x))$ με μία συνεχόμενη γραμμή. Με παρόμοιες εντολές[2] προκύπτει το Σχήμα 2.2.(β), στο οποίο παρατίθεται το γράφημα της εμπειρικής αθροιστικής συνάρτησης κατανομής $F_n(x)$ του δεύτερου δείγματος μεγέθους $n = 100$ μαζί με αυτό της $\Phi(x)$. Παρατηρούμε ότι, καθώς αυξάνεται το μέγεθος δείγματος, η διαφορά της $F_n(x)$ με τη $\Phi(x)$, δηλαδή με την πραγματική συνάρτηση κατανομής, είναι μικρότερη από ότι με την περίπτωση όπου έχουμε μικρό μέγεθος δείγματος. Σύμφωνα με το Θεώρημα Glivenko-Cantelli, η μέγιστη (απόλυτη) διαφορά της εμπειρικής αθροιστικής συνάρτησης κατανομής (κόκκινη) από την πραγματική (μπλε) τείνει στο 0 (με πιθανότητα 1), καθώς $n \to \infty$. Αυτό, βέβαια, δεν σημαίνει ότι οι δύο συναρτήσεις θα ταυτιστούν, καθώς δεν ξεχνάμε ότι για κάθε n η εμπειρική αθροιστική συνάρτηση κατανομής είναι δεξιά συνεχής συνάρτηση, σε αντίθεση με τη $\Phi(x)$ που είναι απόλυτα συνεχής. □

Σύμφωνα, λοιπόν, με το Θεώρημα Glivenko-Cantelli όσο αυξάνεται το μέγεθος του δείγματος, η διαφορά της εμπειρικής αθροιστικής συνάρτησης κατανομής από την πραγματική αθροιστική συνάρτηση κατανομής είναι μικρότερη. Το εύλογο ερώτημα που μπορεί να προκύπτει είναι αν έχει προσδιοριστεί ένα άνω φράγμα για την πιθανότητα η μεγαλύτερη απόλυτη διαφορά τους να είναι μεγαλύτερη από κάποια δοθείσα τιμή. Στη θεωρία πιθανοτήτων και στατιστικής, η ανισότητα Dvoretzky–Kiefer–Wolfowitz (Ανισότητα DKW) οριοθετεί πόσο κοντά είναι η εμπειρική αθροιστική συνάρτηση κατανομής στην αληθινή αθροιστική συνάρτηση κατανομής. Το όνομά της προέρχεται από τους Aryeh Dvoretzky (1916-2008), Jack Kiefer (1924-1981) και Jacob Wolfowitz (1910-1981), οι οποίοι απέδειξαν στην εργασία τους (βλ. Dvoretzky *et al.*, 1956) ότι για κάθε $\varepsilon > 0$ υπάρχει μία πεπερασμένη σταθερά D τέτοια, ώστε για οποιαδήποτε αθροιστική συνάρτηση

[2] Για αναπαραγωγή: `n<-100; set.seed(20); x<-round(rnorm(n,0,1),2);`

κατανομής F στο $\mathbb{R}$ και εμπειρική αθροιστική συνάρτηση κατανομής F_n, να ισχύει:

$$P\left(\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|>\varepsilon\right)\leqslant De^{-2n\varepsilon^2}, \tag{2.2}$$

όπου D μια απροσδιόριστη σταθερά. Ο Massart (1990) απέδειξε ότι η ανισότητα της σχέσης (2.2) ισχύει όταν $D=2$. Συνοψίζοντας, επί του παρόντος, η ανισότητα αυτή είναι γνωστή στη μορφή που δίνεται στο θεώρημα που ακολουθεί. Τέλος, αξίζει να αναφερθεί ότι το αποτέλεσμα έχει πρόσφατα γενικευτεί και στην πολυδιάστατη περίπτωση από τον Naaman (2021). Ειδικότερα, προκύπτει ότι $D=2k$, με k να συμβολίζει τη διάσταση του τυχαίου διανύσματος.

Θεώρημα 2.4: (Ανισότητα Dvoretzky–Kiefer–Wolfowitz (DKW))

Έστω $X_1, X_2, \dots, X_n$ τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(x)$ και $F_n(x)$ είναι η εμπειρική αθροιστική συνάρτηση κατανομής. Τότε, για κάθε $\varepsilon>0$

$$P\left(\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|>\varepsilon\right)\leqslant 2e^{-2n\varepsilon^2}. \tag{2.3}$$

Απόδειξη Θεωρήματος 2.4. Η απόδειξη αυτού του θεωρήματος ξεφεύγει από τους σκοπούς του παρόντος συγγράμματος και παραπέμπουμε τον/την ενδιαφερόμενο/μενη αναγνώστη/στρια στις εργασίες που προαναφέρθηκαν. □

Η ανισότητα DKW δίνει ένα φράγμα για την πιθανότητα ότι η εμπειρική αθροιστική συνάρτηση κατανομής διαφέρει από τη συνάρτηση κατανομής περισσότερο από μια σταθερά ϵ οπουδήποτε στην ευθεία των πραγματικών αριθμών. Επιπλέον, ενισχύει το Θεώρημα Glivenko – Cantelli ποσοτικοποιώντας τον ρυθμό σύγκλισης, καθώς το μέγεθος δείγματος n τείνει στο άπειρο. Μια επιπρόσθετη χρησιμότητα της ανισότητας δίνεται στην επόμενη υποενότητα.

2.1.1 Ζώνη εμπιστοσύνης για την αθροιστική συνάρτηση κατανομής

Μέσω της ανισότητας DKW μπορούμε να κατασκευάσουμε μια **ζώνη εμπιστοσύνης** ή λωρίδα εμπιστοσύνης $100(1-a)\%$ για την αθροιστική συνάρτηση κατανομής F. Συγκεκριμένα, αν θεωρήσουμε $a\in(0,1)$, τότε ισχύουν τα εξής:

Λήμμα 2.4. ($100(1-a)\%$ DKW ζώνη εμπιστοσύνης για την F) Έστω $X_1, X_2, \dots, X_n$ τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(x)$ και $F_n(x)$ είναι η εμπειρική αθροιστική συνάρτηση κατανομής. Έστω

$$L_n(x)=\max\left\{F_n(x)-\sqrt{\frac{1}{2n}\log\frac{2}{a}},0\right\}, \tag{2.4}$$

και

$$U_n(x)=\min\left\{F_n(x)+\sqrt{\frac{1}{2n}\log\frac{2}{a}},1\right\}. \tag{2.5}$$

Τότε για οποιαδήποτε F και για κάθε n,

$$P(L_n(x)\leqslant F(x)\leqslant U_n(x))\geqslant 1-a, \text{ για κάθε } x$$

με $a\in(0,1)$.

Απόδειξη Λήμματος 2.4. Από το Θεώρημα 2.4 και για $\varepsilon_n^2 = \frac{1}{2n}\log\frac{2}{a}$ προκύπτει ότι:

$$\begin{aligned}
P\left(\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|>\varepsilon_n\right) &\leqslant 2e^{-2n\varepsilon_n^2} \Leftrightarrow \\
P(|F_n(x)-F(x)|>\varepsilon_n) &\leqslant 2e^{-2n\frac{1}{2n}\log\frac{2}{a}}, \quad \forall\, x\in\mathbb{R} \Leftrightarrow \\
P(|F_n(x)-F(x)|>\varepsilon_n) &\leqslant a, \quad \forall\, x\in\mathbb{R} \Leftrightarrow \\
P(|F_n(x)-F(x)|\leqslant\varepsilon_n) &\geqslant 1-a, \quad \forall\, x\in\mathbb{R} \Leftrightarrow \\
P(-\varepsilon_n\leqslant F_n(x)-F(x)\leqslant\varepsilon_n) &\geqslant 1-a, \quad \forall\, x\in\mathbb{R} \Leftrightarrow \\
P(-F_n(x)-\varepsilon_n\leqslant -F(x)\leqslant -F_n(x)+\varepsilon_n) &\geqslant 1-a, \quad \forall\, x\in\mathbb{R} \Leftrightarrow \\
P(F_n(x)-\varepsilon_n\leqslant F(x)\leqslant F_n(x)+\varepsilon_n) &\geqslant 1-a, \quad \forall\, x\in\mathbb{R}
\end{aligned}$$

και η απόδειξη ολοκληρώνεται θέτοντας

$$L(x) = \max\{F_n(x)-\varepsilon_n, 0\}$$

και

$$U(x) = \min\{F_n(x)+\varepsilon_n, 1\}.$$

□

Παρατήρηση 2.3. Εναλλακτικά, στην προαναφερθείσα απόδειξη, θα μπορούσαμε να πούμε ότι εξισώνοντας το άνω φράγμα της ανισότητας DKW με a, τότε η λύση ως προς ε δίνεται από $\varepsilon_n^2 = \frac{1}{2n}\log\frac{2}{a}$ και να έχουμε άμεσα ότι

$$P(|F_n(x)-F(x)|>\varepsilon_n)\leqslant a, \quad \forall\, x\in\mathbb{R}.$$

Παράδειγμα 2.3 (συνέχεια Παραδείγματος 2.2)**.** Να υπολογιστούν οι 90% ζώνες εμπιστοσύνης DKW για τα δύο προσομοιωμένα δείγματα μεγέθους $n=10$ και $n=100$ από την τυπική κανονική κατανομή. Να γίνει, με τη βοήθεια της `R`, ένα γράφημα που θα δίνεται η εμπειρική αθροιστική συνάρτηση κατανομής, η συνάρτηση κατανομής της τυπικής κανονικής (από όπου προσομοιώθηκαν τα δεδομένα) και οι 90% ζώνες εμπιστοσύνης για τα δύο προσομοιωμένα δείγματα. Τι παρατηρείτε;

Λύση Παραδείγματος 2.3. Αφού θέλουμε να βρούμε 90% ζώνες εμπιστοσύνης, έχουμε ότι $a=0.1$. Επιπλέον, για το πρώτο δείγμα, καθώς $n=10$, προκύπτει ότι $\sqrt{\frac{1}{2n}\log\frac{2}{a}}\approx 0.387$. Συνεπώς, για κάθε x, υπολογίζουμε τις ποσότητες:

$$\begin{aligned}
L_n(x) &= \max\{F_n(x)-0.387, 0\} \\
U_n(x) &= \min\{F_n(x)+0.387, 1\}.
\end{aligned}$$

Για παράδειγμα, αν $x\in[-0.90,-0.24)$ τότε $F_n(x)=0.2$, οπότε

$$\begin{aligned}
L_n(x) &= \max\{0.2-0.387, 0\} = \max\{-0.187, 0\} = 0 \\
U_n(x) &= \min\{0.2+0.387, 1\} = \min\{0.587, 1\} = 0.587.
\end{aligned}$$

Με παρόμοιο τρόπο γίνονται οι υπολογισμοί για όλες τις δυνατές τιμές του x, τα αποτελέσματα των οποίων δίνονται στον Πίνακα 2.2.

Προφανώς, οι παραπάνω υπολογισμοί είναι ιδιαίτερα χρονοβόροι στην περίπτωση που έχουμε μεγαλύτερο μέγεθος δείγματος, όπως συμβαίνει στη δεύτερη περίπτωση. Το πρόβλημα ξεπερνιέται χρησιμοποιώντας την `R`. Για τον σκοπό αυτόν ορίζουμε τη συνάρτηση `dkw()`, η οποία υπολογίζει τη ζώνη εμπιστοσύνης DKW.

x	$F_n(x)$	$L_n(x)$	$U_n(x)$
$(-\infty, -1.13)$	0	0	0.387
$[-1.13, -0.90)$	0.1	0	0.487
$[-0.90, -0.24)$	0.2	0	0.587
$[-0.24, -0.14)$	0.3	0	0.687
$[-0.14, -0.08)$	0.4	0.013	0.787
$[-0.08, 0.13)$	0.5	0.113	0.887
$[0.13, 0.18)$	0.6	0.213	0.987
$[0.18, 0.71)$	0.7	0.313	1
$[0.71, 1.59)$	0.8	0.413	1
$[1.59, 1.98)$	0.9	0.513	1
$[1.98, \infty)$	1	0.613	1

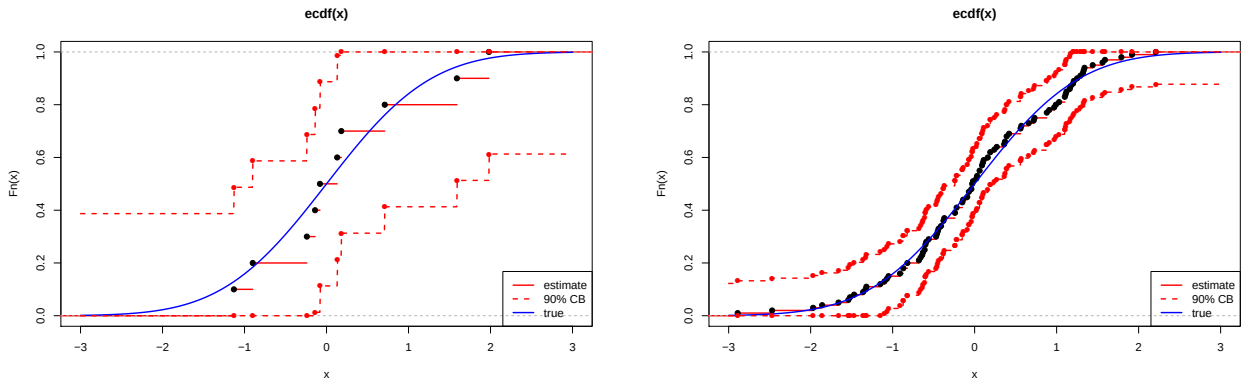
Πίνακας 2.2: Υπολογισμός της εμπειρικής αθροιστικής συνάρτησης κατανομής, του κάτω (L_n) και άνω (U_n) ορίου της 90% ζώνης εμπιστοσύνης DKW στα δεδομένα του Παραδείγματος 2.1.

```
> dkw <- function(x, a, x_evaluate = x){
+          x <- sort(x)
+          Fn <- ecdf(x)
+          n <- length(x)
+          ub <- apply(cbind(1,
+                  Fn(x_evaluate) + sqrt(log(2/a)/(2*n))), 1, min)
+          lb <- apply(cbind(0,
+                  Fn(x_evaluate) - sqrt(log(2/a)/(2*n))), 1, max)
+          cb <- cbind(lb, ub)
+          return(cb)
+ }
```

Η συνάρτηση `dkw()` δέχεται ως υποχρεωτικά ορίσματα τα παρατηρηθέντα δεδομένα (`x`) και έναν αριθμό `a` μεταξύ 0 και 1, ο οποίος χρησιμοποιείται για να καθορίσει τον συντελεστή εμπιστοσύνης, που είναι $1 - a$. Το τρίτο όρισμα είναι προαιρετικό και, αν δεν δοθεί κάποια συγκεκριμένη επιλογή, τότε η συνάρτηση απλώς θα επιστρέψει τα όρια της ζώνης εμπιστοσύνης στα παρατηρηθέντα δεδομένα. Αν δοθεί μία διαφορετική επιλογή, τότε η ζώνη εμπιστοσύνης θα υπολογιστεί στις τιμές του διανύσματος (`x_evaluate`). Έτσι, αν θέλουμε να υπολογίσουμε τις τιμές της ζώνης εμπιστοσύνης με συντελεστή 90% στα δέκα παρατηρηθέντα δεδομένα $\boldsymbol{x}$, γράφουμε στην `R`:

```
> dkw(x, 0.1)
                lb        ub
 [1,] 0.00000000 0.4870228
 [2,] 0.00000000 0.5870228
 [3,] 0.00000000 0.6870228
 [4,] 0.01297724 0.7870228
 [5,] 0.11297724 0.8870228
 [6,] 0.21297724 0.9870228
 [7,] 0.31297724 1.0000000
 [8,] 0.41297724 1.0000000
 [9,] 0.51297724 1.0000000
[10,] 0.61297724 1.0000000
```

Παρατηρήστε ότι οι παραπάνω 10 γραμμές αντιστοιχούν στις τιμές που περιέχονται στις γραμμές 2, 3,..., 11



(α) $n = 10$ (β) $n = 100$

Σχήμα 2.3: 90% ζώνη εμπιστοσύνης DKW μαζί με την εμπειρική αθροιστική συνάρτηση κατανομής και την πραγματική συνάρτηση κατανομής (η οποία είναι γνωστή λόγω προσομοίωσης των δεδομένων) για διαφορετικές τιμές του μεγέθους δείγματος n.

του Πίνακα 2.2. Αν απλά επιθυμούμε να μας επιστραφούν τα όρια της ζώνης εμπιστοσύνης σε μια συγκεκριμένη τιμή, έστω για παράδειγμα την $F(-0.5)$, τότε η εντολή συντάσσεται ως ακολούθως:

```
> dkw(x, 0.1, -0.5)
     lb        ub
[1,]  0 0.5870228
```

Στη συνέχεια, θα κατασκευάσουμε τη γραφική παράσταση που περιέχει την εμπειρική αθροιστική συνάρτηση κατανομής $F_n(x)$ μαζί με τα όρια της 90% ζώνης εμπιστοσύνης. Στο ίδιο διάγραμμα θα παραθέσουμε και την πραγματική συνάρτηση κατανομής $\Phi(x)$.

```
> plot(Fn,xlim=c(-3,3),col.hor = "red", lwd = 2)
> points(xSeq, pnorm(xSeq), type = "l", col = "blue", lwd = 2)
> cb <- dkw(x, 0.1)
> points(sort(x), cb[,1], col = 2, lty = 2, pch = 16, lwd = 2)
> points(sort(x), cb[,2], col = 2, lty = 2, pch = 16, lwd = 2)
> cb <- dkw(x, 0.1, xSeq)
> points(xSeq, cb[,1], col = 2, type = "l", lty = 2, lwd = 2)
> points(xSeq, cb[,2], col = 2, type = "l", lty = 2, lwd = 2)
> legend("bottomright", c("estimate", "90% CB", "true"),
+ col=c("red", "red", "blue"), lty = c(1,2,1), lwd = 2)
```

Το αποτέλεσμα του παραπάνω κώδικα είναι το Σχήμα 2.3.(α). Με παρόμοιο τρόπο και κατάλληλες τροποποιήσεις κατασκευάζεται το 2.3.(β). Παρατηρήστε ότι και οι δύο ζώνες εμπιστοσύνης (για $n = 10$ και $n = 100$) περιέχουν την πραγματική (μπλε) συνάρτηση κατανομής. Σημειώστε ότι εδώ γνωρίζουμε ότι η αθροιστική συνάρτηση κατανομής του πληθυσμού είναι η $\Phi(\cdot)$, αφού έχουμε επιλέξει να προσομοιώσουμε από αυτήν το τυχαίο δείγμα. Ωστόσο, σε πραγματικά προβλήματα, η κατανομή του πληθυσμού δεν είναι γνωστή και για αυτό άλλωστε και θέλουμε να την εκτιμήσουμε. Επίσης, εδώ πρέπει να τονιστεί ότι οι ζώνες εμπιστοσύνης περιέχουν την πραγματική συνάρτηση κατανομής για όλα τα $x \in \mathbb{R}$ και όχι απλώς για κάποιο ή κάποια x. Ακόμη, παρατηρήστε ότι το εύρος της ζώνης εμπιστοσύνης μειώνεται στην περίπτωση που αυξάνεται το μέγεθος δείγματος. □

Εκτός από τη ζώνη εμπιστοσύνης είναι εφικτή και η κατασκευή διαστήματος εμπιστοσύνης για την αθροιστική συνάρτηση κατανομής. Η κατασκευή αυτή αποτελεί αντικείμενο της επόμενης υποενότητας.

2.1.2 Διάστημα εμπιστοσύνης για την αθροιστική συνάρτηση κατανομής

Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(x)$. Θέλουμε να καθορίσουμε στατιστικές συναρτήσεις ℓ_n, u_n τέτοιες, ώστε

$$P(\ell_n \leqslant F(x) \leqslant u_n) \geqslant 1 - a.$$

Στο Θεώρημα 2.1 αποδείχθηκε ότι η τ.μ. $Y = nF_n(x) \sim B(n, F(x))$, για οποιαδήποτε τιμή x. Δηλαδή, ουσιαστικά, η ποσότητα $p = F(x)$ είναι η πιθανότητα επιτυχίας ενός διωνυμικού πειράματος. Επομένως, μπορεί να κατασκευαστεί ένα διάστημα εμπιστοσύνης, με συντελεστή (ή βαθμό) εμπιστοσύνης $1 - a$, για την αθροιστική συνάρτηση κατανομής, υιοθετώντας τις ήδη γνωστές τεχνικές που έχουν αναπτυχθεί στη βιβλιογραφία για την κατασκευή διαστήματος εμπιστοσύνης για την πιθανότητα επιτυχίας της διωνυμικής κατανομής.

Στο πλαίσιο αυτό, ένα πρώτο διάστημα εμπιστοσύνης μπορεί να προκύψει χρησιμοποιώντας τη μέθοδο των Clopper and Pearson (1934), οι οποίοι κατασκεύασαν ένα ακριβές διάστημα εμπιστοσύνης για το p, χρησιμοποιώντας τις αθροιστικές πιθανότητες της διωνυμικής κατανομής. Υιοθετώντας αυτήν τη μέθοδο, το ακριβές διάστημα εμπιστοσύνης για την $F(x)$ είναι το:

$$(\inf S_1, \sup S_2), \text{ όπου } S_1 = \{p | P(Y \leqslant y) > a/2\} \text{ και } S_2 = \{p | P(Y \geqslant y) > a/2\},$$

με $Y \sim B(n, p = F(x))$ και $y = F_n(x)$, η τιμή της εμπειρικής αθροιστικής συνάρτησης κατανομής στο σημείο x.

Εναλλακτικά, χρησιμοποιώντας ασυμπτωτική θεωρία (Κεντρικό Οριακό Θεώρημα, Νόμοι Μεγάλων Αριθμών), μπορεί να κατασκευαστεί ένα ασυμπτωτικό διάστημα εμπιστοσύνης για την $F(x)$, χρησιμοποιώντας την κανονική προσέγγιση της διωνυμικής κατανομής. Το διάστημα αυτό είναι γνωστό στη βιβλιογραφία ως διάστημα εμπιστοσύνης Wald. Υιοθετώντας αυτήν τη μέθοδο, το ασυμπτωτικό διάστημα εμπιστοσύνης για την $F(x)$ είναι το:

$$\left(\widehat{p} - z_{a/2} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + z_{a/2} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \right), \tag{2.6}$$

όπου $\widehat{p} = F_n(x)$ και $z_{a/2}$ είναι το σημείο της τυπικής κανονικής κατανομής, το οποίο που είναι τέτοιο, ώστε, όταν η τυχαία μεταβλητή $Z \sim \mathcal{N}(0,1)$, τότε $P(Z > z_{a/2}) = a/2$.

Οι Newcombe (1998) και Agresti and Coull (1998) έδειξαν ότι το διάστημα εμπιστοσύνης που δίνεται στη σχέση (2.6) μπορεί να έχει πιθανότητα κάλυψης πολύ μικρότερη από $1 - a$ και για αυτό, συνήθως, δεν προτιμάται σε πρακτικές εφαρμογές. Μία βελτίωση, ως προς την πιθανότητα κάλυψης, του διαστήματος εμπιστοσύνης Wald είναι το διάστημα εμπιστοσύνης του Willson (1927), το οποίο κατασκευάστηκε χρησιμοποιώντας έναν μετασχηματισμό της κανονικής προσέγγισης. Το διάστημα εμπιστοσύνης, με συντελεστή εμπιστοσύνης $1 - a$, συμβολίζεται ως (w^-, w^+), όπου

$$w^- = \frac{1}{1 + \frac{z_{a/2}^2}{n}} \left(\widehat{p} + \frac{z_{a/2}^2}{2n} - z_{a/2} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{z_{a/2}^2}{4n^2}} \right)$$

και

$$w^+ = \frac{1}{1 + \frac{z_{a/2}^2}{n}} \left(\widehat{p} + \frac{z_{a/2}^2}{2n} + z_{a/2} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{z_{a/2}^2}{4n^2}} \right).$$

Στις περιπτώσεις όπου είτε το μέγεθος του δείγματος είναι μικρό είτε οι τιμές του p βρίσκονται κοντά στο 0 ή το 1, κάνουμε διόρθωση συνέχειας (Yates' correction) και το διάστημα εμπιστοσύνης που προκύπτει (βλ.

Newcombe, 1998) συμβολίζεται ως (w_{cc}^{-}, w_{cc}^{+}), όπου

$$w_{cc}^{-} = \max\left\{0, \frac{2n\widehat{p} + z_{a/2}^2 - \left[z_{a/2}\sqrt{z_{a/2}^2 - \frac{1}{n} + 4n\widehat{p}(1-\widehat{p}) + (4\widehat{p}-2)} + 1\right]}{2(n + z_{a/2}^2)}\right\}$$

και

$$w_{cc}^{+} = \min\left\{1, \frac{2n\widehat{p} + z_{a/2}^2 + \left[z_{a/2}\sqrt{z_{a/2}^2 - \frac{1}{n} + 4n\widehat{p}(1-\widehat{p}) + (4\widehat{p}-2)} + 1\right]}{2(n + z_{a/2}^2)}\right\}.$$

Προφανώς, στην ειδική περίπτωση του διαστήματος εμπιστοσύνης για την αθροιστική συνάρτηση κατανομής όσα προηγήθηκαν ισχύουν για $\widehat{p} = F_n(x)$.

Μια πιο απλή λύση για τη βελτίωση της πιθανότητας κάλυψης του διαστήματος εμπιστοσύνης Wald έχει δοθεί από τους Agresti and Coull (1998). Έστω x είναι το πλήθος των επιτυχιών που παρατηρήθηκαν σε δείγμα μεγέθους n, $\tilde{n} = n + z_{a/2}^2$ και

$$\tilde{p} = \frac{1}{\tilde{n}}\left(x + \frac{z_{a/2}^2}{2}\right) = \frac{\widehat{p} + \frac{z_{a/2}^2}{2n}}{1 + \frac{z_{a/2}^2}{n}}.$$

Τότε το ασυμπτωτικό διάστημα εμπιστοσύνης για το p που προτάθηκε από τους Agresti and Coull (1998) είναι το:

$$\left(\tilde{p} - z_{a/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}, \tilde{p} + z_{a/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}\right).$$

Σε πρακτικές εφαρμογές, όπου συνήθως $a = 0.05$, αντικαθίσταται το $z_{0.025}$ από το 2 και όχι από την πραγματική τιμή του, η οποία είναι 1.96, οπότε η λογική του «μετασχηματισμού» του n έχει να κάνει με το γεγονός ότι προσθέτουμε στο διωνυμικό πείραμα «2 επιτυχίες και 2 αποτυχίες» (για λεπτομέρειες βλ. Brown *et al.*, 2001).

Τέλος, αναφέρουμε και το ασυμπτωτικό διάστημα εμπιστοσύνης arcsin, που προκύπτει εφαρμόζοντας τον μετασχηματισμό αντίστροφου ημιτόνου. Ο συγκεκριμένος μετασχηματισμός είναι μετασχηματισμός σταθεροποίησης της (ασυμπτωτικής) διασποράς στην κανονική προσέγγιση της διωνυμικής κατανομής και με την εφαρμογή του προκύπτει ότι:

$$\arcsin\left(\sqrt{\widehat{p}}\right) \to \mathcal{N}\left(\arcsin(\sqrt{p}), \frac{1}{4n}\right).$$

Επομένως, σε αυτήν την περίπτωση το ασυμπτωτικό διάστημα εμπιστοσύνης, με συντελεστή εμπιστοσύνης $1 - a$, για το p είναι το:

$$\left(\sin^2\left[\arcsin(\sqrt{\widehat{p}}) - \frac{z_{a/2}}{2\sqrt{n}}\right], \sin^2\left[\arcsin(\sqrt{\widehat{p}}) + \frac{z_{a/2}}{2\sqrt{n}}\right]\right).$$

Παράδειγμα 2.4 (συνέχεια Παραδείγματος 2.3)**.** Να υπολογιστεί το 90% ασυμπτωτικό διάστημα εμπιστοσύνης Wald για την $F(x)$ με $x = 1.64$, χρησιμοποιώντας τα δύο δείγματα μεγέθους $n = 10$ και $n = 100$, αντίστοιχα.

Λύση Παραδείγματος 2.4. Για την περίπτωση του δείγματος με $n = 10$ παρατηρήσεις, η σημειακή εκτίμηση για την $F(1.64)$ είναι η εμπειρική αθροιστική συνάρτηση κατανομής υπολογισμένη στο $x = 1.64$. Επομένως, ανατρέχοντας στον Πίνακα 2.1, έχουμε ότι

$$\widehat{p} = F_n(1.64) = 0.9.$$

Επειδή ο συντελεστής εμπιστοσύνης ισούται με 90%, έχουμε ότι $a = 0.1$ και $z_{0.05} \approx 1.64$ (προσοχή: μην μπερδέψετε αυτήν την τιμή με το $x = 1.64$, που εδώ τυγχάνει να ταυτίζονται!). Αντικαθιστώντας στη σχέση (2.6), λαμβάνουμε το ακόλουθο ασυμπτωτικό διάστημα εμπιστοσύνης:

$$\left(0.9 - 1.64\sqrt{\frac{0.9(1-0.9)}{10}}, 0.9 + 1.64\sqrt{\frac{0.9(1-0.9)}{10}}\right) = (0.744, 1.056).$$

Παρατηρούμε ότι τα άκρα του ασυμπτωτικού διαστήματος εμπιστοσύνης Wald υπερβαίνουν τα όρια του παραμετρικού χώρου (εδώ είναι το $[0,1]$). Επιπλέον, ενθυμούμενοι ότι σε αυτό το παράδειγμα γνωρίζουμε ότι η υποτιθέμενη άγνωστη συνάρτηση κατανομής από την οποία έχουμε προσομοιώσει τα δεδομένα είναι η $\mathcal{N}(0,1)$, έχουμε ότι η πραγματική τιμή της $F(1.64) = \Phi(1.64) \approx 0.95$. Συμπεραίνουμε, λοιπόν, ότι η πραγματική τιμή της αθροιστικής συνάρτησης κατανομής στο σημείο $x = 1.64$ περιέχεται σε αυτό.

Στην περίπτωση του δεύτερου δείγματος (μεγέθους $n = 100$), με αντίστοιχους υπολογισμούς, λαμβάνουμε ότι η πραγματοποίηση του ασυμπτωτικού 90% διαστήματος εμπιστοσύνης είναι το $(0.942, 0.988)$. Παρατηρήστε ότι στην περίπτωση αυτή τα άκρα του διαστήματος βρίσκονται εντός του παραμετρικού χώρου και, επίσης, περιέχουν την (άγνωστη) πραγματική τιμή της $F(1.64) \approx 0.95$. □

2.2 Εκτίμηση συναρτησιακών της αθροιστικής συνάρτησης κατανομής

Έστω $X_1, X_2, \ldots, X_n$, είναι ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(\cdot)$. Στην προηγούμενη ενότητα ασχοληθήκαμε με την εκτίμηση της αθροιστικής συνάρτησης κατανομής. Σε αυτήν την ενότητα το ενδιαφέρον θα επικεντρωθεί στην εκτίμηση παραμέτρων που γράφονται ως συναρτήσεις της συνάρτησης κατανομής F, δηλαδή στην εκτίμηση της $T(F)$, όπου T είναι μια οποιαδήποτε συνάρτηση της F. Έχουμε τότε τον ακόλουθο ορισμό.

Ορισμός 2.2

Έστω X μια τυχαία μεταβλητή από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(\cdot)$. Ονομάζεται **στατιστικό συναρτησιακό** (statistical functional) μια απεικόνιση T, η οποία απεικονίζει την αθροιστική συνάρτηση F σε έναν πραγματικό αριθμό ή διάνυσμα. Στην ειδική περίπτωση που το στατιστικό συναρτησιακό είναι της μορφής $T(F) = \int a(x)dF(x)$, τότε καλείται **γραμμικό συναρτησιακό**.

Από τον προηγούμενο ορισμό προκύπτει ότι παραδείγματα στατιστικών συναρτησιακών είναι, μεταξύ άλλων, η μέση τιμή $\mu = \int x\mathrm{d}F(x)$, η διασπορά $\sigma^2 = \int (x-\mu)^2\mathrm{d}F(x)$ και η διάμεσος $m = F^{-1}(1/2)$.

Ένας εκτιμητής των στατιστικών συναρτησιακών προκύπτει άμεσα αντικαθιστώντας την F με την F_n στη συνάρτηση T. Τότε προκύπτει ο λεγόμενος εκτιμητής αντικατάστασης που ο ορισμός του, για λόγους πληρότητας, ακολουθεί.

Ορισμός 2.3

Έστω $X_1, X_2, \ldots, X_n$, είναι ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(\cdot)$ και $T(F)$ ένα στατιστικό συναρτησιακό. Ο εκτιμητής αντικατάστασης (plug-in estimator) της άγνωστης παραμέτρου $\theta = T(F)$ ορίζεται ως

$$\widehat{\theta}_n = T(F_n),$$

όπου $F_n(\cdot)$ η εμπειρική αθροιστική συνάρτηση κατανομής. Στην περίπτωση που το συναρτησιακό $T(F)$ είναι γραμμικό της μορφής $T(F) = \int a(x)dF(x)$, ο εκτιμητής αντικατάστασης προσδιορίζεται από τη

σχέση

$$T(F_n) = \int a(x)\mathrm{d}F_n(x) = \frac{1}{n}\sum_{i=1}^{n} a(X_i). \tag{2.7}$$

Παρατήρηση 2.4. Σύμφωνα με τον Ορισμό 2.3, ο εκτιμητής αντικατάστασης προκύπτει απλώς αντικαθιστώντας την F με την F_n στη συνάρτηση T. Ωστόσο, σε πολλές περιπτώσεις, παίρνουμε ως εκτιμητή το δειγματικό «εμπειρικό» ανάλογο της ποσότητας που μας ενδιαφέρει. Επιπλέον, οφείλουμε να παρατηρήσουμε ότι, ενώ η αθροιστική συνάρτηση κατανομής F αντιστοιχεί είτε σε συνεχή είτε σε διακριτή τυχαία μεταβλητή, η εμπειρική αθροιστική συνάρτηση κατανομής F_n είναι διακριτή, καθώς θέτει μάζα $1/n$ σε κάθε τυχαία παρατήρηση X_i.

Στη συνέχεια, αναφέρουμε παραδείγματα υπολογισμού του εκτιμητή αντικατάστασης για διάφορα μέτρα του πληθυσμού (βλ. επίσης Wasserman, 2006).

Παράδειγμα 2.5. (Μέση Τιμή) Έστω $X_1, X_2, \ldots, X_n$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(\cdot)$ και μέση τιμή μ. Να δείξετε ότι η πληθυσμιακή μέση τιμή μ είναι ένα γραμμικό συναρτησιακό. Ποιος είναι ο εκτιμητής αντικατάστασης και ποιο το τυπικό σφάλμα του; Προσδιορίστε ένα ασυμπτωτικό διάστημα εμπιστοσύνης για την πληθυσμιακή μέση τιμή.

Λύση Παραδείγματος 2.5. Η πληθυσμιακή μέση τιμή ορίζεται από τη σχέση $\mu = T(F) = \int x dF(x)$ και, επομένως, είναι ένα γραμμικό συναρτησιακό με $a(x) = x$. Ο εκτιμητής αντικατάστασης υπολογίζεται ως:

$$\widehat{\mu} = \int x dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} \alpha(X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}.$$

Το τυπικό σφάλμα του εκτιμητή αντικατάστασης είναι $\mathrm{se} = \sqrt{\mathrm{Var}(\bar{X})} = \dfrac{\sigma}{\sqrt{n}}$. Συνεπώς, αν $\widehat{\sigma}$ είναι ένας εκτιμητής του σ, τότε το εκτιμώμενο τυπικό σφάλμα θα είναι $\widehat{\mathrm{se}} = \dfrac{\widehat{\sigma}}{\sqrt{n}}$. Το ασυμπτωτικό διάστημα εμπιστοσύνης για τη μέση τιμή μ, συντελεστή εμπιστοσύνης $1 - a$, είναι το:

$$\left(\bar{X} - z_{a/2}\frac{\widehat{\sigma}}{\sqrt{n}}, \bar{X} + z_{a/2}\frac{\widehat{\sigma}}{\sqrt{n}}\right).$$

□

Παρατήρηση 2.5. Στο Παράδειγμα 2.5 προσδιορίστηκε το τυπικό σφάλμα se του εκτιμητή $T(F_n)$ και εκτιμήθηκε από τη $\widehat{\mathrm{se}}$. Σε αυτήν την περίπτωση, αλλά και σε πολλές άλλες, ο υπολογισμός αυτός είναι απλός. Ωστόσο, υπάρχουν περιπτώσεις όπου δεν είναι και τόσο προφανής η εκτίμηση του τυπικού σφάλματος. Στις περιπτώσεις όπου υφίσταται αυτή η δυσκολία στον υπολογισμό, μπορεί να δειχθεί ότι:

$$T(F_n) \to \mathcal{N}(T(F), \widehat{\mathrm{se}}^2). \tag{2.8}$$

Επιπροσθέτως, χρησιμοποιώντας το παραπάνω αποτέλεσμα, ένα ασυμπτωτικό διάστημα εμπιστοσύνης για το $T(F)$, με συντελεστή εμπιστοσύνης $1 - a$, είναι το:

$$(T(F_n) - z_{a/2}\widehat{\mathrm{se}}, T(F_n) + z_{a/2}\widehat{\mathrm{se}}). \tag{2.9}$$

Παράδειγμα 2.6. (Διασπορά) Έστω $X_1, X_2, \ldots, X_n$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(\cdot)$ και διασπορά σ^2. Να δείξετε ότι η πληθυσμιακή διασπορά σ^2 είναι ένα γραμμικό συναρτησιακό. Ποιος είναι ο εκτιμητής αντικατάστασης; Ταυτίζεται με τον αμερόληπτο εκτιμητή της διασποράς;

Λύση Παραδείγματος 2.6. Η πληθυσμιακή διασπορά ορίζεται από τη σχέση $\sigma^2 = \int (x-\mu)^2 dF(x)$. Επειδή η μέση τιμή $\mu = \int x \mathrm{d}F(x)$, ισχύει ότι $\sigma^2 = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$. Από τα παραπάνω συμπεραίνουμε ότι $\sigma^2 = T(F)$ και, επιπλέον, είναι ένα γραμμικό συναρτησιακό. Ο εκτιμητής αντικατάστασης της διασποράς υπολογίζεται ως:

$$\begin{aligned}\widehat{\sigma}^2 &= \int x^2 dF_n(x) - \left(\int x dF_n(x)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2.\end{aligned}$$

Παρατηρούμε ότι ο εκτιμητής αυτός είναι διαφορετικός από τη δειγματική διασπορά $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ που είναι αμερόληπτος εκτιμητής της πληθυσμιακής διασποράς σ^2. □

Παράδειγμα 2.7. (Λοξότητα) Έστω $X_1, X_2, \ldots, X_n$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(\cdot)$. Η λοξότητα του πληθυσμού, η οποία μετράει την ασυμμετρία της κατανομής, ορίζεται από τη σχέση:

$$a_3 = \frac{\mathrm{E}[(X-\mu)^3]}{\sigma^3} = \frac{\int (x-\mu)^3 dF(x)}{\left\{\int (x-\mu)^2 dF(x)\right\}^{3/2}},$$

όπου μ και σ^2 υποδηλώνουν την πληθυσμιακή μέση τιμή και διασπορά, αντίστοιχα. Ποιος είναι ο εκτιμητής αντικατάστασης;

Λύση Παραδείγματος 2.7. Από τον ορισμό της λοξότητας προκύπτει:

$$a_3 = \frac{\int (x-\mu)^3 dF(x)}{\left\{\int (x-\mu)^2 dF(x)\right\}^{3/2}}.$$

Από τα Παραδείγματα 2.5 και 2.6 γνωρίζουμε ότι $\widehat{\mu} = \bar{X}$ και $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2$, αντίστοιχα. Επομένως, προκύπτει ότι ο εκτιμητής αντικατάστασης του a_3 είναι ο:

$$\widehat{a}_3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^3}{\widehat{\sigma}^3}.$$

□

Παράδειγμα 2.8. (Συντελεστής Συσχέτισης) Έστω $Z_1 = (X_1, Y_1)$, $Z_2 = (X_2, Y_2)$,..., $Z_n = (X_n, Y_n)$ ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(x,y)$. Εξετάστε αν ο πληθυσμιακός συντελεστής συσχέτισης μεταξύ των μεταβλητών X και Y που ορίζεται από τη σχέση:

$$\rho = T(F) = \frac{\mathrm{E}\left\{(X-\mu_X)(Y-\mu_Y)\right\}}{\sigma_X \sigma_Y} = \frac{\mathrm{E}(XY) - \mu_X \mu_Y}{\sigma_X \sigma_Y}$$

είναι στατιστικό συναρτησιακό. Αν ναι, ποιος είναι ο εκτιμητής αντικατάστασης;

Λύση Παραδείγματος 2.8. Παρατηρούμε ότι ο πληθυσμιακός συντελεστής συσχέτισης μπορεί να γραφτεί στη μορφή

$$T(F) = g(T_1(F), T_2(F), T_3(F), T_4(F), T_5(F)),$$

με

$$g(t_1,t_2,t_3,t_4,t_5)=\frac{t_3-t_1t_2}{\sqrt{\left(t_4-t_1^2\right)\left(t_5-t_2^2\right)}},$$

όπου

$$T_1(F)=\int\int xdF(x,y),\ T_2(F)=\int\int ydF(x,y),$$

$$T_3(F)=\int\int xydF(x,y),\ T_4(F)=\int x^2dF(x,y)$$

και

$$T_5(F)=\int\int y^2dF(x,y).$$

Αντικαθιστώντας τη συνάρτηση κατανομής F με τον εκτιμητή της F_n στις $T_1(F)$, $T_2(F)$, $T_3(F)$, $T_4(F)$, $T_5(F)$, προκύπτει ότι ο εκτιμητής αντικατάστασης είναι

$$\begin{aligned}\widehat{\rho}&=g(T_1(F_n),T_2(F_n),T_3(F_n),T_4(F_n),T_5(F_n))\\&=\frac{\sum_{i=1}^n(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^n\left(X_i-\bar{X}\right)^2}\sqrt{\sum_{i=1}^n\left(Y_i-\overline{Y}\right)^2}}.\end{aligned}$$

Ο παραπάνω εκτιμητής αντικατάστασης ονομάζεται **δειγματικός συντελεστής συσχέτισης**. □

Παράδειγμα 2.9. (Ποσοστιαία Σημεία) Έστω $X_1,X_2,\ldots,X_n$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό με γνησίως αύξουσα αθροιστική συνάρτηση κατανομής $F(\cdot)$. Είναι το $p-$ποσοστιαίο σημείο, $p\in(0,1)$, της κατανομής F στατιστικό συναρτησιακό; Ποιος είναι ο εκτιμητής αντικατάστασης;

Λύση Παραδείγματος 2.9. Το $p-$ποσοστιαίο σημείο της κατανομής F ορίζεται ως

$$x_p=F^{-1}(1-p)=\inf\{x:F(x)\geqslant 1-p\},\ p\in(0,1).$$

Επομένως, είναι στατιστικό συναρτησιακό, καθώς $x_p=T(F)$. Σε αυτήν την περίπτωση, ο εκτιμητής αντικατάστασης είναι,

$$\widehat{x}_p=F_n^{-1}(1-p)=\inf\{x:F_n(x)\geqslant 1-p\}=X_{(j)},$$

όπου $X_{(1)},\ldots,X_{(n)}$ είναι το δείγμα των διατεταγμένων παρατηρήσεων και j είναι εκείνος ο δείκτης για τον οποίο $\frac{j-1}{n}<1-p\leqslant\frac{j}{n}$. Το $\widehat{x}_p$ ονομάζεται **δειγματικό $p-$ποσοστιαίο σημείο**. □

Παρατήρηση 2.6. Μια ενδιαφέρουσα περίπτωση p-ποσοστιαίου σημείου προκύπτει στην ειδική περίπτωση που $p=0.5$. Σε αυτήν την περίπτωση, το ποσοστιαίο σημείο $x_{0.5}$ αναφέρεται ως διάμεσος και συμβολίζεται με m. Από το Παράδειγμα 2.9 προκύπτει ότι ο εκτιμητής αντικατάστασης της διαμέσου δίνεται από τη σχέση:

$$\widehat{m}=\widehat{x}_{0.5}=\widehat{F}_n^{-1}(0.5)=\inf\{x:\widehat{F}_n(x)\geqslant 0.5\}=\begin{cases}X_{(n/2)} & ,\ \text{αν } n \text{ άρτιος}\\ X_{((n+1)/2)} & ,\ \text{αν } n \text{ περιττός.}\end{cases}$$

Στο Παράδειγμα 2.9 το ενδιαφέρον επικεντρώθηκε στη σημειακή εκτίμηση του ποσοστιαίου σημείου. Χρησιμοποιώντας τη διωνυμική κατανομή μπορεί να κατασκευαστεί (ακριβές και όχι ασυμπτωτικό) διάστημα εμπιστοσύνης για το ποσοστιαίο σημείο, όπως αποδεικνύεται στην πρόταση που ακολουθεί (βλ. David and Nagaraja, 2004).

Πρόταση 2.1. Έστω X_1, X_2,...,X_n είναι ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(\cdot)$ και $X_{(1)}$,...,$X_{(n)}$ είναι το δείγμα των διατεταγμένων παρατηρήσεων, ενώ ορίζουμε να είναι $X_{(0)} = -\infty$ και $X_{(n+1)} = \infty$. Επιπρόσθετα, έστω $F_{n,p}(\cdot)$ η αθροιστική συνάρτηση κατανομής της διωνυμικής κατανομής $B(n,p)$. Τότε το τυχαίο διάστημα $\left(X_{(r)}, X_{(s)}\right]$, όπου

$$r = \max\{k = 0,1,\dots,n : F_{n,1-p}(k-1) \leqslant a_1\},$$

και

$$s = \min\{k = 1,\dots,n+1 : F_{n,1-p}(k-1) \geqslant 1 - a_2\}.$$

είναι ένα $100(1-a)\%$ διάστημα εμπιστοσύνης για το ποσοστιαίο σημείο x_p, $p \in (0,1)$, με $a = a_1 + a_2$.

Απόδειξη Πρότασης 2.1. Για την απόδειξη της πρότασης παραπέμπουμε, μεταξύ άλλων, στο σύγγραμμα του Conover (1998). □

Παράδειγμα 2.10 (συνέχεια Παραδείγματος 2.1)**.** Για το τυχαίο δείγμα των 10 τιμών του Παραδείγματος 2.1 να εκτιμηθεί σημειακά η διάμεσος $x_{0.5}$ και να υπολογιστεί το αντίστοιχο 95% Διωνυμικό διάστημα εμπιστοσύνης ίσων ουρών.

Λύση Παραδείγματος 2.10. Έχουμε το τυχαίο δείγμα των ακόλουθων $n = 10$ το πλήθος παρατηρήσεων

$$\boldsymbol{x} = (-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14).$$

Για διευκόλυνση στους υπολογισμούς αρχικά διατάσσουμε τις τιμές του δείγματος κατά αύξουσα σειρά μεγέθους

$$(-1.13, -0.90, -0.24, -0.14, -0.08, 0.13, 0.18, 0.71, 1.59, 1.98).$$

Εφόσον $n = 10$, από το Παράδειγμα 2.9 (βλ. και Παρατήρηση 2.6) έχουμε ότι η σημειακή εκτίμηση της διαμέσου είναι η 5η διατεταγμένη παρατήρηση, δηλαδή:

$$\widehat{m} = \widehat{x}_{0.5} = x_{(5)} = -0.08.$$

Στη συνέχεια, θα εφαρμόσουμε την Πρόταση 2.1 με $a = 0.05$ και $p = 0.5$ για τον υπολογισμό του Διωνυμικού διαστήματος εμπιστοσύνης για τη διάμεσο του πληθυσμού. Επομένως, το 95% διάστημα εμπιστοσύνης (Δ.Ε) για τη διάμεσο είναι το $\left(X_{(r)}, X_{(s)}\right]$, όπου

$$r = \max\{k = 0,1,\dots,10 : F_{10,0.5}(k-1) \leqslant 0.025\},$$

και

$$s = \min\{k = 1,\dots,11 : F_{10,0.05}(k-1) \geqslant 0.975\},$$

με $F_{10,0.5}$ να είναι η αθροιστική συνάρτηση κατανομής της $B(10, 0.5)$. Οι απαραίτητοι υπολογισμοί για τον καθορισμό διατεταγμένων παρατηρήσεων που αντιστοιχούν στους δείκτες r και s συνοψίζονται στον Πίνακα 2.3. Από τον πίνακα αυτό, όπου για διευκόλυνση έχουν επισημανθεί οι αντίστοιχες γραμμές του πίνακα με κίτρινη και πράσινη γραμμή, εύκολα προκύπτει ότι $r = 2$ και $s = 9$, με $x_{(r)} = x_{(2)} = -0.90$ και $x_{(s)} = x_{(9)} = 1.59$. Με βάση το παραπάνω, προκύπτει ότι η πραγματοποίηση του 95% Δ.Ε για τη διάμεσο είναι $\left(x_{(2)}, x_{(9)}\right] = (-0.90, 1.59]$.

Ο παρακάτω κώδικας της R επιστρέφει ένα διάνυσμα που περιέχει τη σημειακή εκτίμηση της διαμέσου, καθώς και τα όρια του 95% Διωνυμικού διαστήματος εμπιστοσύνης, που προσδιορίστηκε στην Πρόταση 2.1. Ο ίδιος κώδικας μπορεί να χρησιμοποιηθεί και για οποιοδήποτε άλλο ποσοστιαίο σημείο x_p, απλά δίνοντας την επιθυμητή τιμή στη μεταβλητή p.

k	$F_{10,0.5}(k-1)$	$x_{(k)}$
0	0.000	$-\infty$
1	0.001	-1.13
2	0.011	-0.90
3	0.055	-0.24
4	0.172	-0.14
5	0.377	-0.08
6	0.623	0.13
7	0.828	0.18
8	0.945	0.71
9	0.989	1.59
10	0.999	1.98
11	1.000	∞

Πίνακας 2.3: Καθορισμός των r και s για τον υπολογισμό του Διωνυμικού διαστήματος εμπιστοσύνης για τη διάμεσο στα δεδομένα του Παραδείγματος 2.10.

```
1  > alpha <- 0.05
2  > p <- 0.5
3  > probTable  <- cbind (0:(n+1), pbinom(q = (-1):n, size = n, prob = 1
     - p), c(-Inf , sort(x), Inf))
4  > bin_lo <- probTable[tail(which(probTable [,2] - alpha/2 <= 0), 1),
     3]
5  > bin_up <- probTable[which(probTable [,2] - 1 + alpha/2 >= 0)[1], 3]
6  > result  <- c(quantile(x, probs = 1 - p, type = 1), bin_lo , bin_up)
7  > names(result) <- c("estimate", "95_low", "95_up")
8  > result
9  estimate   95_low    95_up
10    -0.08    -0.90     1.59
```

Το αντικείμενο `probTable` που ορίζεται στη γραμμή 3 είναι ο Πίνακας 2.3 με τις διωνυμικές πιθανότητες. Τέλος, τονίζουμε ότι η εντολή `quantile(..., probs = 1 - p, type = 1)` υπολογίζει το δειγματικό p-ποσοστιαίο σημείο $\hat{x}_p$ ως την τιμή της αντίστροφης συνάρτησης της εμπειρικής αθροιστικής συνάρτησης κατανομής, όπως περιγράφθηκε στο Παράδειγμα 2.9.

Αξίζει, επίσης, να αναφέρουμε ότι για την επίλυση του συγκεκριμένου παραδείγματος θα μπορούσε να χρησιμοποιηθεί ο Πίνακας Π.7 με τη βοήθεια του οποίου διαπιστώνουμε ότι $F_{10,0.5}(1) = 0.0107 \leqslant 0.025$ και άρα, αφού $k-1=1$, έπεται ότι $r=2$. Με παρόμοιο τρόπο, διαπιστώνουμε ότι $F_{10,0.5}(8) = 0.9893 \geqslant 0.975$ και άρα, αφού $k-1=8$, έπεται ότι $s=9$. □

Παρατήρηση 2.7. Παρατηρούμε ότι το διάστημα εμπιστοσύνης για το p-ποσοστιαίο σημείο που προσδιορίστηκε στην Πρόταση 2.1 δεν εξαρτάται από την κατανομή του πληθυσμού από τον οποίο προέρχεται το αρχικό δείγμα, αλλά εμπλέκεται σε αυτό η διωνυμική κατανομή, η οποία ως διακριτή δημιουργεί το πρόβλημα ότι η πιθανότητα κάλυψης θα είναι τουλάχιστον (και όχι ακριβώς ίση με) $1-a$. Εναλλακτικά, μπορεί να κατασκευαστεί ένα ασυμπτωτικό διάστημα εμπιστοσύνης χρησιμοποιώντας τη (μη παραμετρική) μέθοδο Δέλτα, η οποία θα αναφερθεί στην επόμενη ενότητα.

Το Θεώρημα των Glivenko-Cantelli (Θεώρημα 2.3) μας εξασφαλίζει ότι η εμπειρική αθροιστική συνάρτηση κατανομής F_n συγκλίνει στην F. Είναι, λοιπόν, εύλογο να υποθέσουμε ότι ο εκτιμητής αντικατάστασης $\widehat{\theta} = T(F_n)$ συγκλίνει στο $\theta = T(F)$; Η απάντηση σε αυτό το ερώτημα είναι όχι απαραίτητα και διευκρινίζεται στη συνέχεια μέσω ενός αντιπαραδείγματος. Θεωρούμε ότι η αθροιστική συνάρτηση κατανομής F είναι παραγωγίσιμη με συνάρτηση πυκνότητας πιθανότητας $f(x)$ και έστω $T(F) = F'(x)|_{x=x_0} = f(x_0)$. Τότε η F_n δεν είναι παραγωγίσιμη στα σημεία $x_0 = x_1, \ldots, x_n$, ενώ για $x_0 \neq x_1, \ldots, x_n$ έχουμε $F_n'(x_0) = 0$.

Επομένως, η συνέπεια και η ασυμπτωτική κανονικότητα των εκτιμητών των συναρτησιακών της F απαιτούν επιπλέον ιδιότητες για την $T(F)$, όπως είναι η συνέχεια και η διαφορισιμότητα.

2.3 Συναρτήσεις επιρροής και μη παραμετρική μέθοδος Δέλτα

Η συνάρτηση επιρροής χρησιμοποιείται για την προσέγγιση του τυπικού σφάλματος ενός εκτιμητή αντικατάστασης και ορίζεται ως εξής (Hampel, 1974).

Ορισμός 2.4

Έστω T ένα συναρτησιακό, F είναι συνάρτηση κατανομής στο $\mathbb{R}$ και δ_x είναι η εκφυλισμένη κατανομή στο x. Τότε η **συνάρτηση επιρροής** (Influence Function) ορίζεται ως,

$$\text{IF}_F(x) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)}{\varepsilon},$$

αν το όριο υπάρχει για κάθε $x \in \mathbb{R}$.

Υπενθυμίζεται ότι $\delta_x(y)$ είναι ένα σημείο μάζας στο $x \in \mathbb{R}$, αν

$$\delta_x(y) = \begin{cases} 1 & , \quad y \geqslant x, \\ 0 & , \quad y < x. \end{cases}$$

Ουσιαστικά, η συνάρτηση επιρροής είναι η παράγωγος του συναρτησιακού $T(F)$ στην κατεύθυνση του δ_x. Δηλαδή, αν θεωρήσουμε $F_\varepsilon = (1-\varepsilon)F + \varepsilon\delta_x$, τότε

$$\text{IF}_F(x) = \frac{d}{d\varepsilon} \text{T}(F_\varepsilon)\Big|_{\varepsilon=0}, \tag{2.10}$$

υπό την προϋπόθεση ότι υπάρχει η παράγωγος.

Ορισμός 2.5

Έστω $X_1, X_2, \ldots, X_n$, είναι ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής (α.σ.κ.) $F(\cdot)$. Η **εμπειρική συνάρτηση επιρροής** συμβολίζεται με $\widehat{\text{IF}}(x)$ και ορίζεται ως εξής: $\widehat{\text{IF}}(x) = \text{IF}_{F_n}(x)$, δηλαδή,

$$\widehat{\text{IF}}(x) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F_n + \varepsilon\delta_x) - T(F_n)}{\varepsilon}.$$

Αν $T(F)$ είναι ένα γραμμικό συναρτησιακό, τότε μπορεί να υπολογιστεί η ασυμπτωτική του κατανομή, χρησιμοποιώντας τον Ορισμό 2.4, σύμφωνα με το παρακάτω θεώρημα (βλ. Wasserman, 2006).

Θεώρημα 2.5

Έστω $T(F) = \int a(x)dF(x)$ είναι ένα γραμμικό συναρτησιακό, τότε

1. $\text{IF}_F(x) = a(x) - T(F)$ και $\widehat{\text{IF}}(x) = a(x) - T(F_n)$.
2. Για οποιαδήποτε συνάρτηση κατανομής G,

$$T(G) = T(F) + \int \text{IF}_F(x)dG(x). \tag{2.11}$$

3. $\int \text{IF}_F(x)dF(x) = 0.$

4. Έστω $\tau^2 = \int \mathrm{IF}_F^2(x)dF(x)$. Τότε,

$$\tau^2 = \int (a(x) - T(F))^2 dF(x)$$

και, αν $\tau^2 < \infty$,

$$\sqrt{n}(T(F) - T(F_n)) \to N(0, \tau^2). \tag{2.12}$$

5. Έστω,

$$\widehat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathrm{IF}}^2(X_i) = \frac{1}{n}\sum_{i=1}^{n}(a(X_i) - T(F_n))^2.$$

Τότε, $\widehat{\tau}^2 \xrightarrow{P} \tau^2$ και $\widehat{\mathrm{se}}/\mathrm{se} \xrightarrow{P} 1$, όπου $\widehat{\mathrm{se}} = \widehat{\tau}/\sqrt{n}$ και $\mathrm{se} = \sqrt{\mathrm{Var}(T(F_n))}$.

6. Ισχύει ότι,

$$\frac{\sqrt{n}(T(F) - T(F_n))}{\widehat{\tau}} \to \mathcal{N}(0,1). \tag{2.13}$$

Απόδειξη Θεωρήματος 2.5. 1. Από τον Ορισμό 2.4 και, καθώς $T(F) = \int a(x)dF(x)$,

$$\begin{aligned}
\mathrm{IF}_F(x) &= \lim_{\varepsilon \to 0} \frac{\int a(x)(1-\varepsilon)dF(x) + \varepsilon a(x) - \int a(x)dF(x)}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \frac{-\varepsilon \int a(x)dF(x) + \varepsilon a(x)}{\varepsilon} \\
&= a(x) - T(F).
\end{aligned}$$

Ομοίως, $\widehat{\mathrm{IF}}(x) = a(x) - T(F_n)$.

2. Από το πρώτο μέρος του θεωρήματος έχουμε ότι:

$$\int \mathrm{IF}_F(x)dG(x) = \int (a(x) - T(F))dG(x) = \int a(x)dG(x) - T(F),$$

οπότε η σχέση (2.11) ισχύει.

3. Καθώς το δεύτερο μέρος του θεωρήματος ισχύει για οποιαδήποτε G, προφανώς ισχύει και για την F. Επομένως, έχουμε ότι:

$$\int \mathrm{IF}_F(x)dF(x) = \mathrm{T}(F) - \mathrm{T}(F) = 0.$$

4. Επίσης, λόγω του δεύτερου μέρους του θεωρήματος, προκύπτει ότι:

$$\begin{aligned}
\mathrm{T}(F_n) &= T(F) + \int \mathrm{IF}_F(x)dF_n(x) \\
&= T(F) + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IF}_F(X_i).
\end{aligned}$$

Από το Κεντρικό Οριακό Θεώρημα έχουμε,

$$S = -\sum_{i=1}^{n} \mathrm{IF}_F(X_i) \underset{n\to\infty}{\sim} \mathcal{N}(E(S), \mathrm{Var}(S)).$$

Όμως, από το τρίτο μέρος του θεωρήματος

$$E\left(\mathrm{IF}_F(X_i)\right) = \int \mathrm{IF}_F(x)dF(x) = 0, \forall\ i = 1, \ldots, n.$$

Δηλαδή, $ES = 0$.
Επίσης,

$$\text{Var}\left(\text{IF}_F(X_i)\right) = E\left(\text{IF}_F^2(X_i)\right) - \left[E\left(\text{IF}_F(X_i)\right)\right]^2 = \int \text{IF}_F^2(x)dF(x) = \tau^2, \forall\ i = 1, \dots, n.$$

Επομένως, $\text{Var}(S) = \tau^2/n$ και έπεται ότι η σχέση (2.12) ισχύει.

5. Μπορούμε να γράψουμε τον εκτιμητή του τ^2, ως

$$\widehat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(a(X_i) - T(F)) + (T(F) - T(F_n))\right]^2$$

και κάνοντας τις πράξεις προκύπτει ότι,

$$\widehat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n}(a(X_i) - T(F))^2 + 2(T(F) - T(F_n))\frac{1}{n}\sum_{i=1}^{n}(a(X_i) - T(F)) + (T(F) - T(F_n))^2. \qquad (2.14)$$

Στην απόδειξη του τέταρτου μέρους δείξαμε ότι

$$T(F_n) - T(F) = \frac{1}{n}\sum_{i=1}^{n}\text{IF}_F(X_i),$$

οπότε, λόγω του αποτελέσματος που αποδείχτηκε στο πρώτο μέρος του θεωρήματος

$$T(F_n) - T(F) = \frac{1}{n}\sum_{i=1}^{n}(a(X_i) - T(F)).$$

Από τον Ισχυρό Νόμο των Μεγάλων Αριθμών, καθώς το $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n}\text{IF}_F(X_i) \overset{\text{σ.β.}}{\to} E\left(\text{IF}_F(X_1)\right) = \int \text{IF}_F(x)dF(x) = 0.$$

Δηλαδή ο μεσαίος όρος της σχέσης (2.14) τείνει στο μηδέν σχεδόν βέβαια, όπως και ο τρίτος όρος καθώς, με το ίδιο σκεπτικό,

$$(T(F) - T(F_n))^2 \overset{\text{σ.β.}}{\to} 0.$$

Απομένει να σημειώσουμε ότι ο πρώτος όρος της σχέσης (2.14) είναι

$$\frac{1}{n}\sum_{i=1}^{n}(a(X_i) - T(F))^2 = \frac{1}{n}\sum_{i=1}^{n}\text{IF}_F^2(X_i).$$

Επομένως, από τον Ασθενή Νόμο των Μεγάλων Αριθμών, αφού οι τυχαίες μεταβλητές $a(X_i) - T(F)$ είναι ανεξάρτητες και ισόνομες, έπεται ότι:

$$\frac{1}{n}\sum_{i=1}^{n}\text{IF}_F^2(X_i) \overset{P}{\to} E\left(\text{IF}_F^2(X_1)\right) = \tau^2$$

και το αποτέλεσμα ισχύει.

6. Συνδυάζοντας τα αποτελέσματα του 4ου και 5ου μέρους του θεωρήματος και λαμβάνοντας επιπρόσθετα υπόψη το Θεώρημα του Slutsky, προκύπτει ότι:

$$\frac{\sqrt{n}(T(F) - T(F_n))}{\widehat{\tau}} = \frac{\frac{\sqrt{n}(T(F) - T(F_n))}{\tau}}{\frac{\widehat{\tau}}{\tau}} \to \mathcal{N}(0,1).$$

□

Παρατήρηση 2.8. Όπως επισημαίνεται (βλ., μεταξύ άλλων, Wasserman, 2006) από το παραπάνω θεώρημα προκύπτει ότι η συνάρτηση επιρροής συμπεριφέρεται όπως η συνάρτηση σκορ στην παραμετρική στατιστική συμπερασματολογία. Ειδικότερα, από την παραμετρική στατιστική συμπερασματολογία έχουμε ότι, αν $f(x;\theta)$ είναι η συνάρτηση πυκνότητας πιθανότητας των τ.μ. X_i, $i = 1, \dots, n$, τότε $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ είναι η συνάρτηση πιθανοφάνειας και ο εκτιμητής μέγιστης πιθανοφάνειας $\widehat{\theta}$ της άγνωστης παραμέτρου, είναι εκείνη η τιμή του θ που μεγιστοποιεί την $L(\theta)$. Η συνάρτηση σκορ είναι η συνάρτηση $s_\theta(x) = \partial f(x;\theta)/\partial\theta$, η οποία, υπό συγκεκριμένες συνθήκες ομαλότητας, ικανοποιεί τις συνθήκες:

$$\int s_\theta(x) f(x;\theta) dx = 0 \text{ και } \mathrm{Var}(\widehat{\theta}) \approx \int (s_\theta(x))^2 f(x;\theta) dx/n.$$

Κατά παρόμοιο τρόπο έχουμε ότι για τη συνάρτηση επιρροής ισχύει ότι:

$$\int \mathrm{IF}_F(x) dF(x) = 0 \text{ και } \mathrm{Var}(\mathrm{T}(F_n)) = \int \mathrm{IF}_F^2(x) dF(x)/n.$$

Από τη σχέση (2.13) του Θεωρήματος 2.5 προκύπτει ότι η στατιστική συνάρτηση $\dfrac{T(F) - T(F_n)}{\widehat{\mathrm{se}}}$ ακολουθεί (προσεγγιστικά) την κατανομή $\mathcal{N}(0,1)$. Το αποτέλεσμα αυτό είναι γνωστό στη βιβλιογραφία ως **μη παραμετρική μέθοδος Δέλτα**. Από την κανονική προσέγγιση προκύπτει το ασυμπτωτικό διάστημα εμπιστοσύνης για το γραμμικό συναρτησιακό $T(F)$, που δίνεται στην πρόταση που ακολουθεί.

Θεώρημα 2.6: (Διάστημα Εμπιστοσύνης με τη Μη Παραμετρική Μέθοδο Δέλτα)

Έστω $X_1, X_2, \dots, X_n$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(\cdot)$ και $T(F)$ ένα γραμμικό συναρτησιακό. Υποθέτουμε ότι:

$$\tau^2 = \int (a(x) - T(F))^2 dF(x) < \infty.$$

Τότε ένα $100(1-a)\%$ ασυμπτωτικό διάστημα εμπιστοσύνης για το γραμμικό συναρτησιακό $T(F)$ είναι το:

$$\left(T(F_n) - z_{a/2}\frac{\widehat{\tau}}{\sqrt{n}}, T(F_n) + z_{a/2}\frac{\widehat{\tau}}{\sqrt{n}} \right) \tag{2.15}$$

ή ισοδύναμα το

$$\left(T(F_n) - z_{a/2}\widehat{\mathrm{se}}, T(F_n) + z_{a/2}\widehat{\mathrm{se}} \right)$$

όπου $\widehat{\mathrm{se}} = \dfrac{\widehat{\tau}}{\sqrt{n}}$, με

$$\widehat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n} (a(X_i) - T(F_n))^2.$$

Απόδειξη Θεωρήματος 2.6. Η απόδειξη προκύπτει άμεσα από το 5ο μέρος του Θεωρήματος 2.5. □

Παράδειγμα 2.11. (Ασυμπτωτικό ΔΕ για τη Μέση Τιμή με τη Μη Παραμετρική Μέθοδο Δέλτα) Έστω $X_1, X_2, \dots, X_n$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(\cdot)$. Μπορείτε να προσδιορίσετε ένα $100(1-a)\%$, κατά σημείο, ασυμπτωτικό διάστημα εμπιστοσύνης για την πληθυσμιακή μέση τιμή θ; Ποια υπόθεση χρησιμοποιήσατε;

Λύση Παραδείγματος 2.11. Στο Παράδειγμα 2.5 δείξαμε ότι η πληθυσμιακή μέση τιμή είναι ένα γραμμικό συναρτησιακό, δηλαδή ένα συναρτησιακό της μορφής $T(F) = \int a(x) dF(x)$, με $a(x) = x$, ενώ προσδιορίστηκε

και ο εκτιμητής αντικατάστασης $T(F_n) = \widehat{\theta} = \bar{X}$. Σύμφωνα με το Θεώρημα 2.6, αν

$$\tau^2 = \int (a(x) - T(F))^2 dF(x) = \int (x - \theta)^2 dF(x) < \infty,$$

τότε ένα $100(1-a)\%$ ασυμπτωτικό διάστημα εμπιστοσύνης για τη μέση τιμή είναι το

$$\left(T(F_n) - z_{a/2}\frac{\widehat{\tau}}{\sqrt{n}}, T(F_n) + z_{a/2}\frac{\widehat{\tau}}{\sqrt{n}} \right),$$

όπου $T(F_n) = \bar{X}$ και

$$\widehat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n}(a(X_i) - T(F_n))^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Συνοψίζοντας, το $100(1-a)\%$ ασυμπτωτικό διάστημα εμπιστοσύνης για τη μέση πληθυσμιακή μέση τιμή θ είναι το:

$$(\bar{X} - z_{a/2}\widehat{\text{se}}, \bar{X} + z_{a/2}\widehat{\text{se}}),$$

υπό την υπόθεση ότι η πληθυσμιακή διακύμανση είναι πεπερασμένη. □

Στην προηγούμενη ενότητα δείξαμε ότι μπορούμε να κατασκευάσουμε ένα ακριβές διάστημα εμπιστοσύνης για το ποσοστιαίο σημείο, μέσω της Διωνυμικής κατανομής. Στο παράδειγμα που ακολουθεί προσπαθούμε να κατασκευάσουμε ένα ασυμπτωτικό διάστημα εμπιστοσύνης, χρησιμοποιώντας τη μη παραμετρική μέθοδο Δέλτα.

Παράδειγμα 2.12. (Ασυμπτωτικό ΔΕ για τα Ποσοστιαία σημεία με τη Μη Παραμετρική Μέθοδο Δέλτα) Έστω $X_1, X_2, \ldots, X_n$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(\cdot)$ και συνάρτηση πυκνότητας πιθανότητας $f(\cdot)$. Μπορείτε να προσδιορίσετε ένα $100(1-a)\%$ ασυμπτωτικό διάστημα εμπιστοσύνης για το p-ποσοστιαίο σημείο της κατανομής F; Ποια υπόθεση χρησιμοποιήσατε;

Λύση Παραδείγματος 2.12. Το p-ποσοστιαίο σημείο της κατανομής F προσδιορίζεται από τη σχέση $\theta = T(F) = F^{-1}(p)$, $0 < p < 1$. Για να υπολογίσουμε τη συνάρτηση επιρροής εργαζόμαστε ως εξής: θεωρούμε $F_\varepsilon(y) = (1-\varepsilon)F(y) + \varepsilon\delta_x(y)$, όπου $\delta_x(y)$ είναι η εκφυλισμένη κατανομή στο x, δηλαδή

$$\delta_x(y) = \begin{cases} 1 & , \quad y \geqslant x, \\ 0 & , \quad y < x. \end{cases}$$

Επομένως, από τη σχέση (2.10) αρκεί να υπολογίσουμε την ποσότητα,

$$\frac{d}{d\varepsilon}\mathrm{T}(F_\varepsilon)\Big|_{\varepsilon=0} = \frac{d}{d\varepsilon}F_\varepsilon^{-1}(p).$$

Επειδή $p = F_\varepsilon(\mathrm{T}(F_\varepsilon))$ και $0 = \dfrac{d}{d\varepsilon}F_\varepsilon(\mathrm{T}(F_\varepsilon))$, προκύπτει ότι η συνάρτηση επιρροής είναι (για λεπτομέρειες βλ. Huber, 1981)),

$$\mathrm{IF}_F(x) = \begin{cases} \frac{p-1}{f(\theta)} & , \quad x \leqslant \theta \\ \frac{p}{f(\theta)} & , \quad x > \theta. \end{cases}$$

Οπότε, η ασυμπτωτική διασπορά του $T(F_n)$ είναι:

$$\frac{\tau^2}{n} = \frac{1}{n}\int \mathrm{IF}_F(x) dF(x) = \frac{p(1-p)}{nf^2(\theta)}. \tag{2.16}$$

Δηλαδή, για να εκτιμήσουμε τη διασπορά και να μπορέσουμε να κατασκευάσουμε ένα ασυμπτωτικό διάστημα εμπιστοσύνης, χρησιμοποιώντας τη μέθοδο Δέλτα, χρειάζεται να εκτιμήσουμε τη συνάρτηση πυκνότητας πιθανότητας f. Στο επόμενο κεφάλαιο θα δούμε πώς μπορεί να επιτευχθεί αυτή η εκτίμηση.

□

2.4 Άλλοι μη γραμμικοί εκτιμητές

Σε αυτήν την ενότητα αναφέρουμε δύο προβλήματα (παραδείγματα) εκτίμησης συναρτησιακών, τα οποία δεν είναι γραμμικά, και στα οποία θα δοθεί ιδιαίτερη βαρύτητα στα επόμενα κεφάλαια.

Παράδειγμα 2.13. (χι-τετράγωνο συναρτησιακό) Έστω $A_i \subset \mathbb{R}$, $i = 1,..,k$, σύνολα που αποτελούν μια πεπερασμένη διαμέριση του $\mathbb{R}$, δηλαδή είναι τέτοια ώστε

$$A_i \cap A_j = \varnothing \ , \quad \forall\, i,j = 1,...,k \text{με } i \neq j \ \text{ και } \ \bigcup_{i=1}^{k} A_i = \mathbb{R}.$$

Έστω $X_1, X_2, \dots, X_n$ τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(x)$ και $F_n(x)$ είναι η εμπειρική αθροιστική συνάρτηση κατανομής. Τότε το χι-τετράγωνο συναρτησιακό ορίζεται από τη σχέση

$$T(F) = \sum_{i=1}^{k} p_{i0}^{-1} \left(\int_{A_i} dF(x) - p_{i0} \right)^2, \tag{2.17}$$

όπου

$$p_i = P(A_i) = \int_{A_i} dF(x),$$

και p_{i0}, $i = 1, \dots, k$, γνωστές τιμές, $0 < p_{i0} < 1$. Είναι το χι-τετράγωνο συναρτησιακό ένα γραμμικό συναρτησιακό; Προτείνετε έναν εκτιμητή αντικατάστασης για το $T(F)$.

Λύση Παραδείγματος 2.13. Προφανώς, το $T(F)$ είναι ένα μη γραμμικό συναρτησιακό, καθώς δεν μπορεί να γραφτεί στη μορφή $T(F) = \int a(x)dF(x)$. Ωστόσο, παρατηρούμε ότι εντός της σχέσης (2.17), υπάρχει το συναρτησιακό

$$T_1(F) = p_i = \int_{A_i} dF(x) = \int I_{A_i}(x) dF(x),$$

όπου $I_{A_i}(x)$ είναι η δείκτρια συνάρτηση. Επομένως, $T_1(F)$ είναι ένα γραμμικό συναρτησιακό με $a(x) = I_{A_i}(x)$. Ένας προφανής εκτιμητής αντικατάστασης της $T_1(F)$ είναι ο:

$$T_1(F_n) = \frac{1}{n} \sum_{i=1}^{k} I_{A_i}(x_i) = \frac{\text{πλήθος των } x_i \in A_i}{n} = \frac{n_i}{n}.$$

Τότε ένας εκτιμητής αντικατάστασης του συναρτησιακού της σχέσης (2.17) είναι ο:

$$T(F_n) = \sum_{i=1}^{k} p_{i0}^{-1} \left(\int_{A_i} dF_n(x) - p_{i0} \right)^2 = \sum_{i=1}^{k} p_{i0}^{-1} \left(\frac{n_i}{n} - p_{i0} \right)^2.$$

□

Παρατήρηση 2.9. Από το προηγούμενο παράδειγμα έχουμε ότι:

$$nT(F_n) = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}}$$

η οποία είναι η στατιστική συνάρτηση που χρησιμοποιείται, στο λεγόμενο χι-τετράγωνο έλεγχο καλής προσαρμογής, για το πρόβλημα ελέγχου της μηδενικής υπόθεσης

$$H_0 : p_i = p_{i0} \ , \quad \forall\, i = 1, \dots, k$$

έναντι της εναλλακτικής υπόθεσης $H_1 : p_i \neq p_{i0}$ για κάποιο i, με $1 \leqslant i \leqslant k$. Περισσότερες λεπτομέρειες θα αναφερθούν στο Κεφάλαιο 4.

Παράδειγμα 2.14. (Συναρτησιακό των Mann-Whitney) Έστω X και Y δύο ανεξάρτητες τυχαίες μεταβλητές, με αθροιστική συνάρτηση κατανομής $F(\cdot)$ και $G(\cdot)$, αντίστοιχα. Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από τον πρώτο πληθυσμό και $Y_1, Y_2, \ldots, Y_m$ τυχαίο δείγμα από τον δεύτερο πληθυσμό. Υποθέτουμε ότι τα δύο δείγματα είναι ανεξάρτητα. Να εξετάσετε αν η

$$\theta = P(X \leqslant Y)$$

είναι συναρτησιακό, δηλαδή αν μπορεί να γραφτεί ως $T(F, G)$. Αν ναι, να προσδιορίσετε έναν εκτιμητή αντικατάστασης του θ.

Λύση Παραδείγματος 2.14. Χρησιμοποιώντας την ανεξαρτησία των τυχαίων μεταβλητών X και Y, εύκολα προκύπτει (αφήνεται ως άσκηση για τον/την αναγνώστη/στρια) ότι:

$$\theta = P(X \leqslant Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{y} dF(x) dG(y),$$

ή, ισοδύναμα, ότι:

$$\theta = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I_{(-\infty, y]}(x) dF(x) dG(y),$$

όπου

$$I_{(-\infty, y]}(x) = \begin{cases} 1 & , \quad -\infty < x \leqslant y \\ 0 & , \quad \text{διαφορετικά.} \end{cases} \tag{2.18}$$

Δηλαδή $\theta = T(F, G)$ και, επομένως, είναι ένα συναρτησιακό, το οποίο προφανώς είναι μη γραμμικό. Ωστόσο, μπορούμε για την εκτίμησή του να χρησιμοποιήσουμε τις εμπειρικές συναρτήσεις κατανομών F_n, G_m, οι οποίες βασίζονται στα ανεξάρτητα τυχαία δείγματα $X_1, \ldots, X_n$ και $Y_1, \ldots, Y_m$, αντιστοίχα. Επομένως, ο εκτιμητής αντικατάστασης του θ είναι

$$\widehat{\theta} = T(F_n, G_m) = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, y_j]}(x_i).$$

□

Το συναρτησιακό του προηγούμενου παραδείγματος είναι γνωστό ως συναρτησιακό των Mann-Whitney. Περισσότερες λεπτομέρειες θα αναφερθούν στο Κεφάλαιο 5.

2.5 Εφαρμογή με R

Στην ενότητα αυτή θα εφαρμόσουμε όσα περιγράφθηκαν στις προηγούμενες ενότητες του κεφαλαίου με τη βοήθεια της R. Θα θεωρήσουμε το σύνολο δεδομένων με το οποίο ασχολήθηκαν οι Cox and Lewis (1966) και το οποίο αποτελείται από $n = 799$ το πλήθος τιμές, οι οποίες αναφέρονται σε χρόνους αναμονής μεταξύ διαδοχικών συσπάσεων νευρικής ίνας. Το αρχείο δεδομένων `nerve_data.dat` είναι διαθέσιμο στον ιστότοπο http://www.stat.cmu.edu/~larry/all-of-nonpar/data.html. Αρχικά, αποθηκεύουμε το συγκεκριμένο αρχείο στον υπολογιστή μας με το όνομα `"nerve.dat"`. Κατόπιν, φορτώνουμε στην R τα δεδομένα που είναι στο αρχείο αυτό με τις ακόλουθες εντολές.

```
con <- file("nerve.dat", "r")
x <- as.numeric(
  unlist(
    strsplit(
      readLines(con), "\\t")
    )
```

```
  )
  close(con)
x <- x[!is.na(x)]
hist(x,  main = "", xlab = bquote(italic(x)))
```

Το ιστόγραμμα των δεδομένων που δημιουργείται από την τελευταία εντολή παρατίθεται στο Σχήμα 2.4 (α). Στη συνέχεια, υπολογίζουμε την εμπειρική αθροιστική συνάρτηση κατανομής F_n μαζί με την 95% ζώνη εμπιστοσύνης DKW, όπως παρατίθεται στο Σχήμα 2.4 (β).

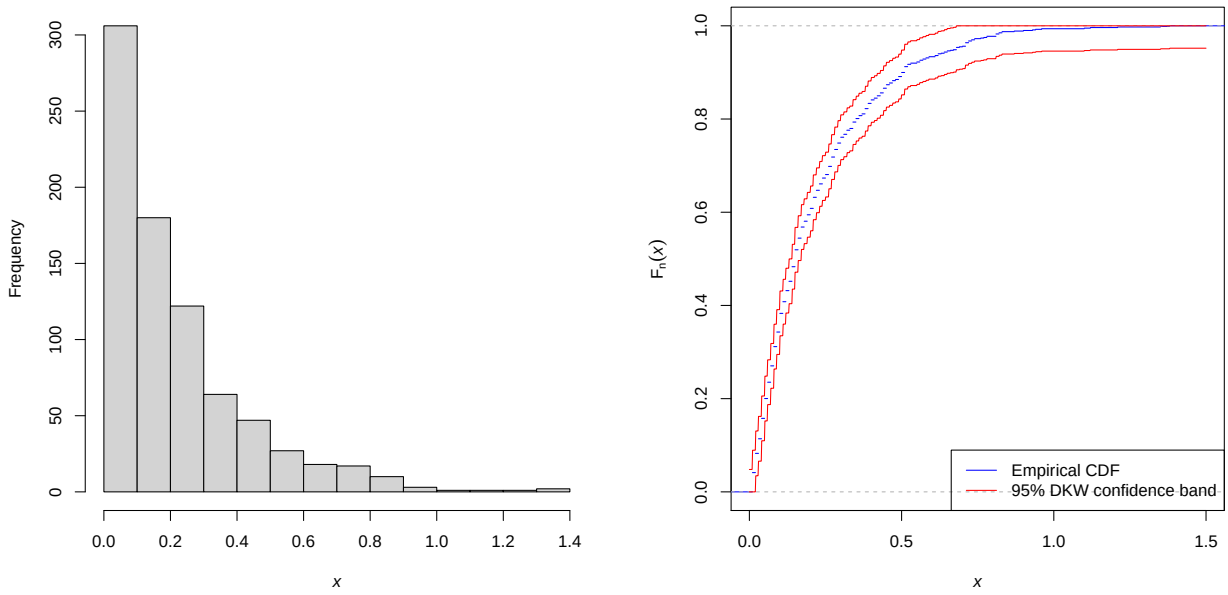
```
Fn <- ecdf(x)
xSeq <- seq(0,1.5,length=1000)
par(mar = c(4,4.5,1,1))
plot(Fn, xlim = c(0,1.5), ylab = bquote(F[n](italic(x))), xlab =
   bquote(italic(x)),
  col.hor = "blue", do.points = FALSE, main = "")
cb <- dkw(x, 0.05, xSeq)
points(xSeq, cb[,1],col = "red", type = "l", lty = 1)
points(xSeq, cb[,2],col = "red", type = "l", lty = 1)
legend('bottomright', c('Empirical CDF', '95% DKW confidence band'),
   col=c('blue','red'),lty=1)
```

Να σημειωθεί στο σημείο αυτό ότι στον παραπάνω κώδικα έχει χρησιμοποιηθεί η συνάρτηση `dkw()`, η οποία έχει οριστεί στο Παράδειγμα 2.2. Ακολούθως, θα υπολογίσουμε την τιμή της $F_n(x)$, τα άκρα της 95% ζώνης εμπιστοσύνης DKW και το αντίστοιχο σημειακό διάστημα εμπιστοσύνης Wald, χρησιμοποιώντας τη σχέση (2.6), για $x \in \{0.1, 0.2, \ldots, 1.2\}$.

```
> xValues <- (1:12)/10
> pn <- Fn(xValues)
> alpha <- 0.05
> dkwV <- dkw(x, alpha, xValues)
> n <- length(x)
> aciL <- pn - qnorm(alpha/2, lower.tail = FALSE)*sqrt(pn*(1-pn)/n)
> aciU <- pn + qnorm(alpha/2, lower.tail = FALSE)*sqrt(pn*(1-pn)/n)
> results <- round(cbind(xValues, pn, dkwV, aciL, aciU), 3)
> colnames(results) <- c("x", "Fn", "DKW_low", "DKW_up", "Wald_low", "
   Wald_up")
> results
         x    Fn DKW_low DKW_up Wald_low Wald_up
 [1,] 0.1 0.383   0.335  0.431    0.349   0.417
 [2,] 0.2 0.608   0.560  0.656    0.574   0.642
 [3,] 0.3 0.761   0.713  0.809    0.731   0.791
 [4,] 0.4 0.841   0.793  0.889    0.816   0.866
 [5,] 0.5 0.900   0.852  0.948    0.879   0.921
 [6,] 0.6 0.934   0.886  0.982    0.916   0.951
 [7,] 0.7 0.956   0.908  1.000    0.942   0.970
 [8,] 0.8 0.977   0.929  1.000    0.967   0.988
 [9,] 0.9 0.990   0.942  1.000    0.983   0.997
[10,] 1.0 0.994   0.946  1.000    0.988   0.999
[11,] 1.1 0.995   0.947  1.000    0.990   1.000
[12,] 1.2 0.996   0.948  1.000    0.992   1.000
```

Στη συνέχεια, υπολογίζουμε τους εκτιμητές αντικατάστασης για τη μέση τιμή (βλ. το Παράδειγμα 2.5), για τη διασπορά (βλ. το Παράδειγμα 2.6) και για τον συντελεστή λοξότητας της κατανομής (δείτε το Παράδειγμα 2.7).

```
> mu = mean(x)
> mu
[1] 0.2185732
> s2 <- var(x) * (n - 1) / n
> s2
[1] 0.0437051
> library("e1071")
> kappa = skewness(x, type = 1)
> kappa
[1] 1.761249
```



(α) (β)

Σχήμα 2.4: Nerve data. (α) Ιστόγραμμα δεδομένων. (β) Εμπειρική Αθροιστική Συνάρτηση Κατανομής (---) και 95% ζώνη εμπιστοσύνης DKW(---).

Τέλος, εκτιμούμε σημειακά και με ένα 95% διάστημα εμπιστοσύνης τα τεταρτημόρια της κατανομής των παρατηρήσεων (x_p, $p = 0.75, 0.5, 0.25$).

```
> alpha <- 0.05; result <- matrix(NA, 3, 4)
> colnames(result) <- c("p", "x_p", "95_low", "95_up")
> j <- 0
> for(p in 3:1/4){
+ probTable <- cbind(0:(n+1), pbinom(q = (-1):n,
+          size = n, prob = 1 - p), c(-Inf, sort(x), Inf))
+ bin_lo <- probTable[tail(which(probTable[,2] - alpha/2<=0),1),3]
+ bin_up <- probTable[which(probTable[,2] - 1 + alpha/2 >= 0)[1],3]
+ j <- j + 1
+ result[j, ] <- c(p, quantile(x, probs = 1 - p, type = 1), bin_lo,
   bin_up)}
> result
         p  x_p 95_low 95_up
[1,] 0.75 0.07   0.06  0.08
[2,] 0.50 0.15   0.14  0.16
[3,] 0.25 0.30   0.28  0.34
```

Για παράδειγμα, η πρώτη γραμμή του αποτελέσματος αντιστοιχεί στο 1ο τεταρτημόριο $x_{0.75}$. Παρατηρούμε ότι ο εκτιμητής αντικατάστασης ισούται με $\hat{x}_{0.75} = 0.07$, ενώ η πραγματοποίηση του Διωνυμικού διαστήματος εμπιστοσύνης με συντελεστή 95% ισούται με $(0.06, 0.08)$.

2.6 Ασκήσεις

Άσκηση 2.1. Έστω F η συνάρτηση κατανομής μιας τυχαίας μεταβλητής X. Υπολογίστε τον εκτιμητή αντικατάστασης, έστω $\widehat{\theta}$, της πιθανότητας $\theta = P(a < X \leqslant \beta)$, για οποιουσδήποτε πραγματικούς αριθμούς $a < \beta$. Ποια θα είναι η συνάρτηση επιρροής αυτής της πιθανότητας;

Άσκηση 2.2. Έστω F η συνάρτηση κατανομής μιας τυχαίας μεταβλητής X. Υπολογίστε τον εκτιμητή αντικατάστασης της διασποράς της κατανομής σ^2, όταν η μέση τιμή μ είναι γνωστή. Ποια θα είναι η συνάρτηση επιρροής του σ^2;

Άσκηση 2.3. Έστω $X_1, \dots, X_n$ τυχαίο δείγμα με $X_i \sim F$, $i = 1, 2, \dots, n$, όπου F είναι συνάρτηση κατανομής και $F_n(x)$ είναι η εμπειρική αθροιστική συνάρτηση κατανομής. Για δοθέν x, προσδιορίστε την ασυμπτωτική κατανομή του $\sqrt{F_n(x)}$.

Άσκηση 2.4. Έστω $X_1, \dots, X_n$ τυχαίο δείγμα με $X_i \sim \mathrm{B}(1, p)$, $i = 1, 2, \dots, n$. Υπολογίστε τον εκτιμητή αντικατάστασης και το εκτιμώμενο τυπικό σφάλμα για το p. Βρείτε το 95% ασυμπτωτικό διάστημα εμπιστοσύνης για το p.

Άσκηση 2.5. Έστω τα ανεξάρτητα τυχαία δείγματα $X_1, \dots, X_n$ με $X_i \sim B(1, p_1)$, $i = 1, 2, \dots, n$ και $Y_1, \dots, Y_m$, με $Y_j \sim B(1, p_2)$, $j = 1, 2, \dots, m$. Υπολογίστε τον εκτιμητή αντικατάστασης και το εκτιμώμενο τυπικό σφάλμα για το $p_1 - p_2$. Βρείτε το 90% ασυμπτωτικό διάστημα εμπιστοσύνης για το $p_1 - p_2$.

Άσκηση 2.6. Στη διάθεσή μας έχουμε το παρακάτω δείγμα $n = 15$ το πλήθος τιμών.

$$0.16,\ 0.36,\ 0.11,\ 0.27,\ 0.17,\ 0.25,\ 0.20,\ 0.13,\ 0.27,\ 0.09,\ 0.28,\ 0.27,\ 0.11,\ 0.39,\ 0.04$$

1. Υπολογίστε την εμπειρική αθροιστική συνάρτηση κατανομής για τα δεδομένα του δείγματος των $n = 15$ τιμών.
2. Υπολογίστε την 95% ζώνη εμπιστοσύνης DKW για τη συνάρτηση κατανομής των δεδομένων του δείγματος.
3. Χρησιμοποιήστε την `R` και επιβεβαιώστε τα αποτελέσματα των προηγούμενων δύο ερωτημάτων. Στη συνέχεια, κατασκευάστε ένα διάγραμμα στο οποίο να απεικονίζεται η εμπειρική αθροιστική συνάρτηση κατανομής, καθώς και η 95% ζώνης εμπιστοσύνης DKW.

Άσκηση 2.7. Οι τιμές του Πίνακα 2.4 καταγράφουν την ταχύτητα (σε km^3/sec) 82 γαλαξιών από 6 περιοχές στον Στέφανο Βόρειο (Corona Borealis) αστερισμό

9.172	9.350	9.483	9.558	9.775	10.227	10.406	16.084
16.170	18.419	18.552	18.600	18.927	19.052	19.070	19.330
19.343	19.349	19.440	19.473	19.529	19.541	19.547	19.663
19.846	19.856	19.863	19.914	19.918	19.973	19.989	20.166
20.175	20.179	20.196	20.215	20.221	20.415	20.629	20.795
20.821	20.846	20.875	20.986	21.137	21.492	21.701	21.814
21.921	21.960	22.185	22.209	22.242	22.249	22.314	22.374
22.495	22.746	22.747	22.888	22.914	23.206	23.241	23.263
23.484	23.538	23.542	23.666	23.706	23.711	24.129	24.285
24.289	24.366	24.717	24.990	25.633	26.960	26.995	32.065
32.789	34.279						

Πίνακας 2.4: Γαλαξιακό σύνολο δεδομένων. Πηγή: Roeder (1990).

1. Υπολογίστε την εμπειρική αθροιστική συνάρτηση κατανομής.

2. Απεικονίστε σε ένα διάγραμμα την εμπειρική αθροιστική συνάρτηση κατανομής. Τι παρατηρείτε;
3. Υπολογίστε την 95% ζώνη εμπιστοσύνης DKW για τη συνάρτηση κατανομής των δεδομένων και προσθέστε την στο διάγραμμα του προηγούμενου ερωτήματος.
4. Για $x = 10, 15, 20, 25, 30, 35$, υπολογίστε τις σημειακές εκτιμήσεις της $F(x)$, τα όρια της 95% ζώνης εμπιστοσύνης DKW και το 95% ασυμπτωτικό διαστημα εμπιστοσύνης Wald ίσων ουρών για την $F(x)$. Υπόδειξη: παρουσιάστε τα αποτέλεσματα σε έναν πίνακα..
5. Εκτιμήστε σημειακά και με ένα 95% διάστημα εμπιστοσύνης τα ποσοστιαία σημεία x_p της κατανομής των παρατηρήσεων, για $p = 0.05, 0.5, 0.95$.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ξενόγλωσση

Agresti, A. and Coull, B. (1998). Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician*, *52*, pp. 119–126.

Brown, L., Cai, T. and DasGupta, A. (2001). Interval estimation for a binomial proportion (with discussion). *Statistical Science*, *16*, pp. 101–133.

Cantelli, F. P. (1933). Sulla determinazione empirica delle leggi di probabilità. *Giorn. Ist. Ital. Attuari*, *4*, pp. 421–424.

Clopper, C. and Pearson, E. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*, pp. 404–416.

Conover, W. J. (1998). *Practical Nonparametric Statistics* (3rd ed.). Wiley and Sons, Inc.

Cox, D. and Lewis, P. (1966). *The Statistical Analysis of Series of Events*. New York, NY: Chapman and Hall.

David, H. and Nagaraja, H. (2004). *Order Statistics*. Wiley Series in Probability and Statistics. Wiley.

Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, *4(3)*, pp. 642–669.

Gibbons, J. D. and Chakraborti, S. (2020). *Nonparametric Statistical Inference, Fourth Edition Revised and Expanded*. Chapman and Hall/CRC.

Glivenko, V. (1933). Sulla determinazione empirica delle leggi di probabilità. *Giorn. Ist. Ital. Attuari*, *4*, pp. 92–99.

Hampel, F. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, *69*, pp. 383–393.

Huber, P. (1981). *Robust Statistics*. John Wiley & Sons, Ltd.

Loève, M. (1977). *Probability Theory I*. New York, NY: Springer.

Massart, P. (1990). The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *Ann. Probab.*, *18(3)*, pp. 1269–1283.

Naaman, M. (2021). On the tight constant in the multivariate Dvoretzky–Kiefer–Wolfowitz inequality. *Statistics & Probability Letters*, *173*, p. 109088.

Newcombe, R. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, *17*, pp. 857–872.

Rènyi, A. (1962). Thèorie des Èlèments Saillants d'Une Suite d'Observations. *in Colloquium on Combinatorial Methods in Probability Theory*, pp. 104–117.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, *85*(411), pp. 617–624.

Shaikh, A. M. (2022). *The Glivenko-Cantelli Theorem*. [Online; accessed 21-February-2023]. URL: http://home.uchicago.edu/~amshaikh/webfiles/glivenko-cantelli.pdf.

Wasserman, L. (2006). *All of Nonparametric Statistics*. New York, NY: Springer Texts in Statistics.

Willson, E. (1927). Probable inference, the law of succession, and statistical inference. *J. Amer. Statist. Assoc.*, *22(158)*, pp. 209–212.

ΚΕΦΑΛΑΙΟ 3

ΜΗ ΠΑΡΑΜΕΤΡΙΚΗ ΕΚΤΙΜΗΣΗ ΤΗΣ ΣΥΝΑΡΤΗΣΗΣ ΠΥΚΝΟΤΗΤΑΣ ΠΙΘΑΝΟΤΗΤΑΣ

Σύνοψη

Όταν έχουμε ένα διαθέσιμο δείγμα από έναν πληθυσμό, ένα από τα κεντρικότερα προβλήματα αποτελεί η εκτίμηση της άγνωστης συνάρτησης πυκνότητας πιθανότητας (σ.π.π.). Η παραμετρική προσέγγιση βασίζεται στην υπόθεση ότι η συναρτησιακή μορφή της σ.π.π. είναι γνωστή, ώστε στη συνέχεια να προχωρήσουμε στην εκτίμηση των άγνωστων παραμέτρων της. Το κεφάλαιο αυτό πραγματεύεται την εκτίμηση της σ.π.π. για την οποία δεν απαιτείται κάποια επιπλέον υπόθεση για τη σ.π.π. από τη συνέχεια ή τη διαφορισιμότητά της. Στο πλαίσιο αυτό, το Κεφάλαιο 3 πραγματεύεται με λεπτομέρεια, παρουσιάζοντας τα πλεονεκτήματα και τα μειονεκτήματα της μεθόδου του ιστογράμματος, η οποία είναι η πιο συνήθης μέθοδος εκτίμησης της συνάρτησης πυκνότητας πιθανότητας. Επίσης, παρουσιάζονται εκτιμητές πυκνοτήτων με χρήση πυρήνα, η οποία είναι μια πιο αποτελεσματική τεχνική από την εκτίμηση μέσω ιστογράμματος.

Προαπαιτούμενη γνώση:

Βασικές γνώσεις Απειροστικού Λογισμού, Πιθανοτήτων και Στατιστικής.

Προσδοκώμενα μαθησιακά αποτελέσματα:

Ο/η φοιτητής/τρια ολοκληρώνοντας την ενότητα αυτή θα μπορεί να εκτιμά μη παραμετρικά τη συνάρτηση πυκνότητας πιθανότητας μέσω ιστογράμματος και μέσω του εκτιμητή που βασίζεται στη χρήση πυρήνα. Επίσης, θα έχει κατανοήσει τις βασικές έννοιες των μεθόδων εξομάλυνσης.

Γλωσσάριο επιστημονικών όρων

- Cross-validated πιθανοφάνεια
- Cross-validation (διεπικύρωση)
- Ανάπτυγμα Taylor
- Ασυμπτωτική ανάλυση
- Εκτιμητής πυκνότητας με χρήση πυρήνα
- Ιστόγραμμα
- Μέθοδοι εξομάλυνσης
- Μεροληψία και διασπορά εκτιμητή
- Ολοκληρωμένο μέσο τετραγωνικό σφάλμα
- Πυρήνας

Μπατσίδης, Α., Παπασταμούλης, Π., Πετρόπουλος, Κ., & Ρακιτζής, Α. (2022). *Μη Παραμετρική Στατιστική.* [Προπτυχιακό εγχειρίδιο].

3.1 Εισαγωγή

Μια συνάρτηση πυκνότητας πιθανότητας μπορεί να εκτιμηθεί κάνοντας χρήση παραμετρικών, ημιπαραμετρικών ή μη παραμετρικών μεθόδων.

Στην πρώτη περίπτωση, υποτίθεται ότι η συνάρτηση πυκνότητας πιθανότητας $f(x)$ είναι μέλος μιας δεδομένης παραμετρικής οικογένειας κατανομών με άγνωστη παράμετρο $\theta \in \Theta$, δηλαδή ότι:

$$f(x) \in \mathscr{F} = \{f(\cdot;\theta); \theta \in \Theta\}.$$

Έτσι, αρκεί να εκτιμηθεί η παράμετρος θ.

Στην περίπτωση που μια συνηθισμένη παραμετρική οικογένεια κατανομών δεν επαρκεί για να περιγράψει ικανοποιητικά τα δεδομένα, τότε μπορούμε να υποθέσουμε ότι η άγνωστη κατανομή των δεδομένων ανήκει σε μία ευρύτερη οικογένεια κατανομών, όπως οι πεπερασμένες μείξεις. Σε αυτήν την περίπτωση, η άγνωστη $f(x)$ εκφράζεται ως σταθμισμένος μέσος όρος διάφορων παραμετρικών μοντέλων. Τα μοντέλα αυτά, που ανήκουν στη δεύτερη κατηγορία των ημιπαραμετρικών μεθόδων, είναι αρκετά ευέλικτα και μπορούν να περιγράψουν ικανοποιητικά πληθώρα άγνωστων κατανομών (Titterington *et al.*, 1985; Frühwirth-Schnatter, 2006), ωστόσο, η εκτίμησή τους είναι πιο απαιτητική.

Στο παρόν κεφάλαιο θα μελετηθεί διεξοδικά η τρίτη περίπτωση και, συγκεκριμένα, θα περιγραφεί η μεθοδολογία μη παραμετρικής εκτίμησης της συνάρτησης πυκνότητας πιθανότητας με χρήση ιστογράμματος και πυρήνα (kernel). Άλλες μη παραμετρικές μέθοδοι εκτίμησης της συνάρτησης πυκνότητας πιθανότητας βασίζονται σε splines, σε ορθογώνιες σειρές και στην ποινικοποιημένη πιθανοφάνεια (Silverman, 1986).

Σε όλες αυτές τις μεθόδους γίνεται χρήση της υπόθεσης ότι η $f(x)$ είναι «ομαλή» συνάρτηση. Αυτό σημαίνει ότι η f είναι μέλος κατάλληλων κλάσεων εκτιμητών που εξαρτώνται από κάποια «παράμετρο εξομάλυνσης» h. Έτσι αρκεί να επιλεχθεί κατάλληλα η παράμετρος εξομάλυνσης. Το κριτήριο που θα χρησιμοποιηθεί για αυτόν τον σκοπό είναι, όπως αναλυτικά θα δούμε, το ολοκληρωμένο μέσο τετραγωνικό σφάλμα, αλλά και η cross validated πιθανοφάνεια.

Η θεωρητική ανάπτυξη των παραπάνω τεχνικών είναι αρκετά παλιά, καθώς, για παράδειγμα, η εκτίμηση πυκνότητας με χρήση πυρήνα είχε προταθεί ήδη από τη δεκαετία του '50 (Rosenblatt, 1956; Parzen, 1962; Epanechnikov, 1969). Ωστόσο, ο μεγάλος όγκος των υπολογισμών που απαιτούνται αποτελούσε εμπόδιο για την πρακτική εφαρμογή της, μέχρι και τη δεκαετία του '80, όπου η ανάπτυξη και η χρήση υπολογιστών ήταν πολύ πιο διαδεδομένες σε σχέση με πριν.

Στην Ενότητα 3.2 γίνεται μια εισαγωγή στις μεθόδους εξομάλυνσης (**smoothing**), οι οποίες μας παρέχουν τα κατάλληλα εργαλεία για την αξιολόγηση εκτιμητών συναρτήσεων, όπως είναι η συνάρτηση πυκνότητας πιθανότητας, ενώ εφαρμόζονται και σε πιο γενικά πλαίσια, όπως είναι η μη παραμετρική παλινδρόμηση. Οι Ενότητες 3.3 και 3.4 αποτελούν την «καρδιά» αυτού του κεφαλαίου, όπου παρουσιάζονται οι εκτιμητές της συνάρτησης πυκνότητας πιθανότητας με χρήση ιστογράμματος και πυρήνα, αντίστοιχα. Ταυτόχρονα, στις Ενότητες 3.3.1 και 3.4.3 μελετώνται θεωρητικές ιδιότητες των εκτιμητών αυτών, ενώ δίνονται και παραδείγματα εφαρμογής των τεχνικών στην `R`. Στην Ενότητα 3.5 γίνεται μια εισαγωγή στην εκτίμηση της συνάρτησης πυκνότητας πιθανότητας σε προβλήματα πολυμεταβλητής φύσης, καθώς και μια σύντομη περιγραφή των δυσκολιών που παρατηρούνται.

3.2 Εισαγωγή στις μεθόδους εξομάλυνσης

Έστω ένα τυχαίο δείγμα $X_1, \ldots, X_n$ από μια κατανομή με (άγνωστη) συνάρτηση πυκνότητας πιθανότητας $f(x)$. Στο πρόβλημα της μη παραμετρικής εκτίμησης της συνάρτησης πυκνότητας πιθανότητας, μας

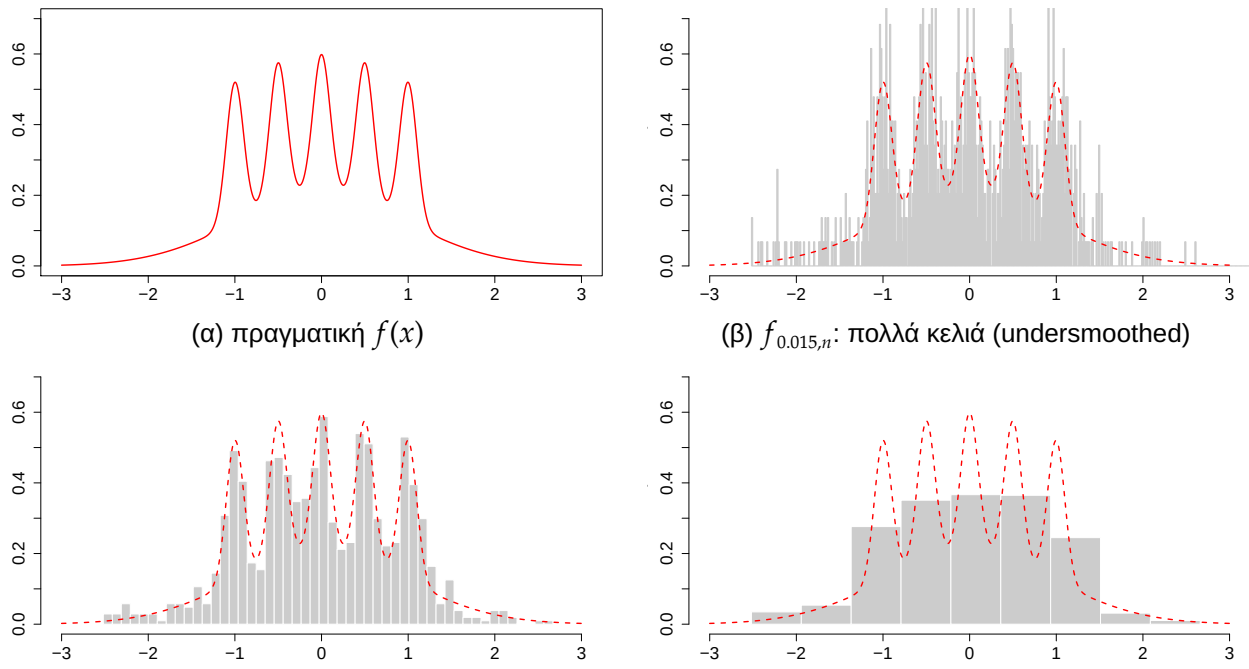
ενδιαφέρει η εκτίμηση της $f(x)$ με βάση το τυχαίο δείγμα $X_1, \dots, X_n$, χωρίς να υποθέσουμε κάποια συγκεκριμένη παραμετρική οικογένεια κατανομών. Καθώς οι εκτιμητές της συνάρτησης πυκνότητας πιθανότητας εξαρτώνται από κάποια **παράμετρο εξομάλυνσης** h θα συμβολίζονται ως $f_{h,n}(x)$. Γενικά, η παράμετρος εξομάλυνσης μπορεί να εξαρτάται από το μέγεθος δείγματος n, δηλαδή είναι $h = h_n$ (για παράδειγμα, $h = 1/\sqrt{n}$), αλλά για λόγους ευκολίας στον συμβολισμό θα συνεχίσουμε να γράφουμε h αντί για h_n.

Έστω η συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \sum_{j=1}^{6} p_j \varphi(x; \mu_j, \sigma_j^2) \tag{3.1}$$

όπου με $\varphi(\cdot; \mu, \sigma^2)$ συμβολίζεται η συνάρτηση πυκνότητας πιθανότητας της κατανομής $\mathcal{N}(\mu, \sigma^2)$, $p_1 = 0.5$, $\mu_1 = 0$, $\sigma_1^2 = 1$, ενώ $p_j = 0.1$, $\mu_j = 0.5j - 2$, $\sigma_j^2 = 0.1^2$, για $j = 2, \dots, 6$. Η (3.1) είναι πράγματι συνάρτηση πυκνότητας πιθανότητας, διότι είναι μία μείξη (έξι) κανονικών κατανομών. Στο Σχήμα 3.1.(α) απεικονίζεται η συνάρτηση πυκνότητας πιθανότητας της σχέσης (3.1), η οποία, για προφανείς λόγους, αναφέρεται ως "Bart Simpson" στο σύγγραμμα του Wasserman (2006).

Υποθέτουμε ότι έχουμε στη διάθεσή μας ένα τυχαίο δείγμα μεγέθους $n = 1000$ παρατηρήσεων από την $f(x)$. Ένας απλός τρόπος εκτίμησης της $f(x)$ είναι μέσω του ιστογράμματος. Οι λεπτομέρειες της μεθόδου θα παρουσιαστούν στην Ενότητα 3.3, ενώ στην ενότητα αυτή θα αρκεστούμε σε μια σύντομη περιγραφή. Αρχικά, γίνεται μια διαμέριση της πραγματικής ευθείας σε διαδοχικά διαστήματα (κελιά) και ακολούθως μετράμε τον αριθμό των παρατηρήσεων εντός κάθε κελιού. Το ύψος της ράβδου που αντιστοιχεί σε καθένα κελί είναι ανάλογο του αριθμού των παρατηρήσεων εντός αυτού. Πιο συγκεκριμένα, η εκτίμηση της συνάρτησης πυκνότητας πιθανότητας μέσω ιστογράμματος για ένα κελί είναι η σχετική συχνότητα των παρατηρήσεων εντός αυτού διαιρεμένη με το αντίστοιχο πλάτος του.



(α) πραγματική $f(x)$ (β) $f_{0.015,n}$: πολλά κελιά (undersmoothed)

(γ) $f_{0.104,n}$: όσα πρέπει (δ) $f_{0.575,n}$: λίγα κελιά (oversmoothed)

Σχήμα 3.1: (α) Η συνάρτηση πυκνότητας πιθανότητας της εξίσωσης (3.1). (β), (γ), (δ) Διάφοροι εκτιμητές $f_{h,n}$ πυκνότητας με χρήση ιστογράμματος και διαφορετικό πλάτος κελιών (h), δοθέντος ενός προσομοιωμένου συνόλου δεδομένων μεγέθους $n = 1000$ από την f.

Στα Σχήματα 3.1.(β), 3.1.(γ) και 3.1.(δ) παρατίθεται η εκτιμηθείσα συνάρτηση πυκνότητας $f(x)$ μέσω ιστογράμματος $f_{h,n}(x)$, με διαφορετικές τιμές του πλάτους κελιών $h = 0.015$, $h = 0.104$ και $h = 0.575$, αντίστοιχα. Γίνεται σαφές ότι στο ιστόγραμμα το πλάτος των κελιών (ισοδύναμα, το πλήθος των κελιών) αποτελεί παράμετρο εξομάλυνσης (smoothing parameter). Στην περίπτωση του Σχήματος 3.1.(β) το h είναι «μικρό» (ισοδύναμα, το πλήθος κελιών είναι «μεγάλο») και αυτό οδηγεί σε μια εκτίμηση της συνάρτησης πυκνότητας πιθανότητας, η οποία προσπαθεί να προσαρμοστεί τέλεια στα παρατηρηθέντα δεδομένα. Αυτό, όμως, έχει ως αποτέλεσμα, όπως θα δούμε, ο εκτιμητής να έχει μικρή μεροληψία, αλλά μεγάλη διασπορά (undersmoothing). Αντίθετα, στην περίπτωση του Σχήματος 3.1.(δ) το h είναι «μεγάλο» (ισοδύναμα, το πλήθος κελιών είναι «μικρό») και αυτό οδηγεί σε μια εκτίμηση της συνάρτησης πυκνότητας πιθανότητας η οποία αγνοεί βασικά χαρακτηριστικά της f, όπως το γεγονός ότι έχει πολλές κορυφές. Αυτό, όμως, έχει ως αποτέλεσμα, όπως θα δούμε, ο εκτιμητής να έχει μεγάλη μεροληψία και μικρή διασπορά (oversmoothing). Μεταξύ αυτών των δύο ακραίων περιπτώσεων βρίσκονται τιμές του h που αντιστοιχούν σε ισορροπία των δύο αυτών αντιμαχόμενων ποσοτήτων (διασποράς και μεροληψίας), όπως στο Σχήμα 3.1.(γ). Συγκρίνοντας το ιστόγραμμα με την πραγματική $f(x)$ (κόκκινη διακεκομμένη καμπύλη) στο Σχήμα 3.1.(γ) παρατηρούμε ότι αναδεικνύονται τα βασικά χαρακτηριστικά της f.

Προτού προχωρήσουμε, θα τονιστούν κάποια πράγματα σχετικά με τον συμβολισμό που θα χρησιμοποιηθεί. Αρχικά, να παρατηρήσουμε ότι στον εκτιμητή $f_{h,n}(x)$ παραλείπουμε να τονίσουμε την εξάρτηση από τα δεδομένα $(X_1, \dots, X_n)$, ορίζοντας:

$$f_{h,n}(x) := f_{h,n}(X_1, \dots, X_n; x).$$

Παρατηρήστε, επίσης, ότι αυτό που ενδιαφέρει είναι η εκτίμηση μιας συνάρτησης (του x) και όχι απλώς μιας άγνωστης παραμέτρου. Είναι σαφές ότι ο εκτιμητής είναι τυχαία μεταβλητή για κάθε x. Έτσι, οι συμβολισμοί $\mathrm{E}(f_{h,n}(x))$ και $\mathrm{Var}\left(f_{h,n}(x)\right)$ αναφέρονται στη μέση τιμή και διασπορά της $f_{h,n}$, αντίστοιχα, θεωρώντας ότι το x είναι σταθερό. Για παράδειγμα, για σταθερό x,

$$\begin{aligned}\mathrm{E}\left(f_{h,n}(x)\right) &= \mathrm{E}\left(f_{h,n}(X_1, \dots, X_n; x)\right) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{h,n}(x_1, \dots, x_n; x) \prod_{i=1}^{n} f(x_i) \mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &=: \bar{f}_h(x). \end{aligned} \tag{3.2}$$

Η μεροληψία ενός εκτιμητή πυκνότητας $f_{h,n}(x)$ που χρησιμοποιείται για την εκτίμηση της $f(x)$ ορίζεται ως:

$$\mathrm{bias}\{f_{h,n}(x), f(x)\} = \mathrm{E}\left(f_{h,n}(x) - f(x)\right). \tag{3.3}$$

Τέλος, η διασπορά του εκτιμητή πυκνότητας $f_{h,n}(x)$ ορίζεται ως:

$$\mathrm{Var}\left(f_{h,n}(x)\right) = \mathrm{E}\left[f_{h,n}(x) - \mathrm{E}\left(f_{h,n}(x)\right)\right]^2. \tag{3.4}$$

3.2.1 Ολοκληρωμένο μέσο τετραγωνικό σφάλμα

Από την προηγούμενη ενότητα είναι σαφές ότι είναι κρίσιμη η επιλογή της παραμέτρου εξομάλυνσης h για έναν εκτιμητή πυκνότητας $f_{h,n}$. Το κεντρικό πρόβλημα στις τεχνικές εξομάλυνσης είναι, φυσικά, η κατάλληλη επιλογή του h, το οποίο δεν πρέπει να είναι ούτε πολύ «μικρό» ούτε πολύ «μεγάλο», ώστε να ισοσταθμίζονται η διασπορά και η μεροληψία του εκτιμητή. Αυτές οι δύο έννοιες, της μεροληψίας και της διασποράς, συνδυάζονται στο **μέσο τετραγωνικό σφάλμα**, το οποίο μας οδηγεί στον ορισμό του **ολοκληρωμένου μέσου τετραγωνικού σφάλματος**, που είναι ένα από τα βασικά κριτήρια που χρησιμοποιούνται για την επιλογή της παραμέτρου εξομάλυνσης h.

Από την Εκτιμητική γνωρίζουμε ότι το τετραγωνικό σφάλμα είναι μία συνάρτηση ζημίας η οποία χρησιμοποιείται αρκετά συχνά στη Στατιστική. Στην περίπτωση της εκτίμησης της συνάρτησης πυκνότητας πιθανότητας f μέσω της $f_{h,n}$, το τετραγωνικό σφάλμα ισούται με:

$$L(f_{h,n}(x), f(x)) = (f_{h,n}(x) - f(x))^2. \tag{3.5}$$

Επειδή το $f_{h,n}(x)$ είναι τυχαία μεταβλητή, θεωρούμε την αναμενόμενη τιμή της (3.5), δηλαδή το μέσο τετραγωνικό σφάλμα το οποίο ορίζεται ως ακολούθως.

Ορισμός 3.1

Το μέσο τετραγωνικό σφάλμα (Mean Square Error) του $f_{h,n}(x)$ για την εκτίμηση της $f(x)$ ορίζεται ως:

$$\text{MSE}\{f_{h,n}(x), f(x)\} = \text{E}\left\{L(f_{h,n}(x), f(x))\right\} = \text{E}\left\{f_{h,n}(x) - f(x)\right\}^2. \tag{3.6}$$

Από τον Ορισμό 3.1 έπεται ότι το μέσο τετραγωνικό σφάλμα μπορεί να εκφραστεί και ως

$$\text{MSE}\{f_{h,n}(x), f(x)\} = \left(\text{bias}\{f_{h,n}(x), f(x)\}\right)^2 + \text{Var}\left(f_{h,n}(x)\right). \tag{3.7}$$

Ξεκάθαρα, το MSE είναι συνάρτηση του x (και φυσικά της άγνωστης $f(x)$). Για να λάβουμε υπόψη όλες τις δυνατές τιμές του x χρησιμοποιούμε το **ολοκληρωμένο μέσο τετραγωνικό σφάλμα** (Integrated Mean Square Error). Στη συνέχεια δίνονται ο ορισμός του ολοκληρωμένου μέσου τετραγωνικού σφάλματος και η τεχνική cross-validation για την εκτίμηση αυτού.

Ορισμός 3.2

Το ολοκληρωμένο μέσο τετραγωνικό σφάλμα της $f_{h,n}(x)$ που χρησιμοποιείται για την εκτίμηση της $f(x)$ ορίζεται από τη σχέση:

$$\begin{aligned}\text{IMSE}\{f_{h,n}, f\} &= \int_{-\infty}^{\infty} \text{MSE}\{f_{h,n}(x), f(x)\}\text{d}x \\ &= \int_{-\infty}^{\infty} \left(\text{bias}\{f_{h,n}(x), f(x)\}\right)^2 \text{d}x + \int_{-\infty}^{\infty} \text{Var}\left(f_{h,n}(x)\right)\text{d}x.\end{aligned} \tag{3.8}$$

Έτσι, το συνολικό σφάλμα υπολογίζεται ολοκληρώνοντας ως προς όλες τις δυνατές τιμές του x. Αν για την εκτίμηση της f είχαμε στη διάθεσή μας δύο εκτιμητές $f_{h_1,n}$ και $f_{h_2,n}$ και ισχύει ότι

$$\text{IMSE}\{f_{h_1,n}, f\} < \text{IMSE}\{f_{h_2,n}, f\}$$

οποιαδήποτε και αν είναι η άγνωστη f, θα πρέπει να προτιμήσουμε τον $f_{h_1,n}$ για την εκτίμησή της.

Αυτό που μας απασχολεί στη συνέχεια είναι η ελαχιστοποίηση του IMSE ως προς h, ώστε να επιλέξουμε την παράμετρο εξομάλυνσης.

Από τη σχέση (3.8) έχουμε ότι:

$$\begin{aligned}\text{IMSE}\{f_{h,n}, f\} &= \int_{-\infty}^{\infty} \text{E}\left\{f_{h,n}(x) - f(x)\right\}^2 \text{d}x \\ &= \text{E}\left[\int_{-\infty}^{\infty} \left\{f_{h,n}(x) - f(x)\right\}^2 \text{d}x\right] \\ &= \text{E}\left[\int_{-\infty}^{\infty} f_{h,n}^2(x)\text{d}x - 2\int_{-\infty}^{\infty} f_{h,n}(x)f(x)\text{d}x + \int_{-\infty}^{\infty} f^2(x)\text{d}x\right].\end{aligned}$$

Ο τελευταίος όρος δεν εξαρτάται από το h, οπότε αρκεί να ελαχιστοποιήσουμε την

$$J(h) := \mathrm{E}\left[\int_{-\infty}^{\infty} f_{h,n}^2(x)\mathrm{d}x - 2\int_{-\infty}^{\infty} f_{h,n}(x)f(x)\mathrm{d}x\right]. \tag{3.9}$$

Είναι σαφές ότι η ποσότητα $J(h)$ είναι άγνωστη, επειδή εξαρτάται από την άγνωστη συνάρτηση πυκνότητας πιθανότητας $f(x)$. Επομένως, θα πρέπει πρώτα να εκτιμηθεί η ποσότητα $J(h)$ και μετά να ελαχιστοποιηθεί ως προς h. Μία πρώτη ιδέα είναι να χρησιμοποιήσουμε τον εκτιμητή αντικατάστασης (θυμηθείτε τον Ορισμό 2.3)

$$\tilde{J}(h) = \int_{-\infty}^{\infty} f_{h,n}^2(x)\mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} f_{h,n}(x_i). \tag{3.10}$$

Δυστυχώς, ο $\tilde{J}(h)$ δεν είναι καλός εκτιμητής λόγω διπλής χρήσης των δεδομένων: Αρχικά τα δεδομένα χρησιμοποιούνται για την εκτίμηση της f και, στη συνέχεια, επιστρατεύονται ξανά για την εκτίμηση του μέσου σφάλματος. Αυτό έχει ως συνέπεια ότι ο $\tilde{J}(h)$ είναι μεροληπτικός και πάντα θα υποδεικνύει ότι το μέσο σφάλμα ελαχιστοποιείται, όταν η παράμετρος εξομάλυνσης $h \approx 0$ (οι ανυπόμονοι/νες αναγνώστες/στριες μπορούν να ανατρέξουν στο Παράδειγμα 3.5).

Αυτή η συμπεριφορά είναι αρκετά συνηθισμένη στη Στατιστική και δεν θα πρέπει να μας εκπλήσσει. Κλασικό παράδειγμα είναι ο συντελεστής προσδιορισμού στη γραμμική παλινδρόμηση: χρησιμοποιούμε τα ίδια δεδομένα για να εκτιμήσουμε ένα μοντέλο αλλά και για να αξιολογήσουμε την προσαρμογή του. Τέτοιου είδους προσεγγίσεις τείνουν να μεροληπτούν υπέρ των πολύπλοκων έναντι των απλούστερων μοντέλων. Το φαινόμενο αυτό είναι γνωστό ως «overfitting» (βλ., μεταξύ άλλων, Hawkins, 2004; Cawley and Talbot, 2010) .

3.2.2 Εκτίμηση ολοκληρωμένου μέσου τετραγωνικού σφάλματος με χρήση cross-validation

Η μέθοδος cross-validation[1] (David, 1974; Stone, 1974, 1977) είναι μία τεχνική για να εξετάσουμε την καλή προσαρμογή ενός μοντέλου στα δεδομένα μας. Βασίζεται στην ιδέα χωρισμού του δείγματος σε δύο τμήματα:

- ένα τμήμα των δεδομένων χρησιμοποιείται για την εκτίμηση του μοντέλου (training dataset), ενώ
- τα υπόλοιπα δεδομένα χρησιμοποιούνται για την αξιολόγηση του μοντέλου (test dataset).

Το πρόβλημα με την παραπάνω διαδικασία είναι ότι, συνήθως, δεν έχουμε επαρκή δεδομένα για να την υλοποιήσουμε χωρίς σημαντική απώλεια πληροφορίας. Για τον λόγο αυτό, η μέθοδος cross-validation εφαρμόζει την ιδέα αυτή χωρίζοντας το δείγμα σε K ομάδες. Χρησιμοποιεί $K-1$ ομάδες για την εκτίμηση και το μοντέλο αξιολογείται μέσω της ομάδας που έμεινε εκτός. Κατόπιν, η διαδικασία επαναλαμβάνεται για καθεμία ομάδα που μένει εκτός (άρα υλοποιείται K φορές). Το αποτέλεσμα προκύπτει ως μέσος όρος όλων αυτών των επαναλήψεων.

Τυπικές επιλογές για τον αριθμό των ομάδων είναι $K = 5,10$ και n. Όταν $K = n$ η τεχνική ονομάζεται Leave-One-Out cross-validation (LOOCV). Σε αυτήν την περίπτωση αφήνουμε 1 παρατήρηση εκτός και το μοντέλο εκτιμάται βάσει $n-1$ παρατηρήσεων. Η παρατήρηση που έμεινε εκτός χρησιμοποιείται για την αξιολόγηση του μοντέλου. Επαναλαμβάνουμε αυτήν τη διαδικασία n το πλήθος φορές, κάθε φορά αφήνοντας και μία διαφορετική παρατήρηση εκτός. Τελικά, υπολογίζουμε τον μέσο όρο όλων των n εκτιμήσεων.

[1] Μία ελληνική μετάφραση του cross-validation είναι διεπικύρωση.

Ορισμός 3.3

Ο εκτιμητής Leave-One-Out cross-validation της ποσότητας $J(h)$ που ορίστηκε στη σχέση (3.9) ορίζεται ως

$$\hat{J}(h) = \int_{-\infty}^{\infty} f_{h,n}^2(x)\mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} f_{h,n-1}^{(-i)}(x_i), \tag{3.11}$$

όπου η $f_{h,n-1}^{(-i)}(x_i)$ είναι η εκτίμηση της πυκνότητας στο x_i, που προκύπτει αφαιρώντας την i-οστή παρατήρηση του δείγματος, $i = 1, \dots, n$.

Στο Σχήμα 3.2 αναπαρίσταται γραφικά η διαδικασία LOOCV για τον υπολογισμό του δεύτερου όρου της σχέσης (3.11). Για κάθε γραμμή, οι $(n-1)$ παρατηρήσεις που αντιστοιχούν στα γκρι κουτάκια χρησιμοποιούνται για την κατασκευή του εκτιμητή $f_{h,n-1}^{(-i)}(\cdot)$ (train data), ενώ η παρατήρηση που αφήνεται εκτός (μπλε κουτάκι) χρησιμοποιείται για τον υπολογισμό της πυκνότητας (test data) σε αυτό το σημείο. Η τελική εκτίμηση προκύπτει ως μέσος όρος.

Πρόταση 3.1. Ο εκτιμητής $\hat{J}(h)$, που δόθηκε στη σχέση (3.11), είναι αμερόληπτος εκτιμητής του $J(h)$, δηλαδή ισχύει ότι:

$$\mathrm{E}(\hat{J}(h)) = J(h).$$

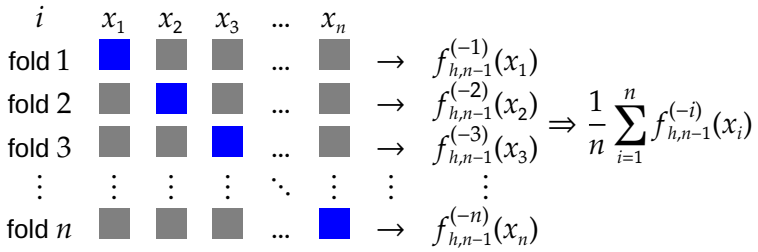
Απόδειξη Πρότασης 3.1. Από τις σχέσεις (3.9) και (3.11), αρκεί να δείξουμε ότι

$$\mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n} f_{h,n-1}^{(-i)}(X_i)\right\} = \mathrm{E}\left\{\int_{-\infty}^{\infty} f_{h,n}(x)f(x)\mathrm{d}x\right\}.$$

Συμβολίζουμε με $\boldsymbol{X}_{[-i]}$ το $(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$, ενώ $\mathrm{d}\boldsymbol{x}_{[-i]} = \prod_{j=1, j\neq i}^{n} \mathrm{d}x_i$. Εξ ορισμού, η $f_{h,n-1}^{(-i)}(\cdot)$ είναι ανεξάρτητη από την X_i, $i = 1, \dots, n$. Άρα

$$\begin{aligned}
\mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n} f_{h,n-1}^{(-i)}(X_i)\right\} &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E} f_{h,n-1}^{(-i)}(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E} f_{h,n-1}(\boldsymbol{X}_{[-i]}; X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{h,n-1}(\boldsymbol{x}_{[-i]}; x_i) \prod_{i=1}^{n} f(x_i)\mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \left\{\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{h,n-1}(\boldsymbol{x}_{[-i]}; x_i) \prod_{j\neq i} f(x_j)\mathrm{d}\boldsymbol{x}_{[-i]}\right\} f(x_i)\mathrm{d}x_i \\
&\overset{(3.2)}{=} \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \bar{f}_h(x_i) f(x_i)\mathrm{d}x_i \\
&= \int_{-\infty}^{\infty} \bar{f}_h(x) f(x)\mathrm{d}x = \mathrm{E}\left\{\int_{-\infty}^{\infty} f_{h,n}(x) f(x)\mathrm{d}x\right\}
\end{aligned}$$

και η απόδειξη ολοκληρώθηκε. □



Σχήμα 3.2: Η διαδικασία Leave-One-Out cross-validation για την εκτίμηση της $\mathrm{E}\left(\int_{-\infty}^{\infty} f_{h,n}(x)f(x)\mathrm{d}x\right)$ στη σχέση (3.9). Για κάθε γραμμή (fold), τα train και test τμήματα του συνόλου δεδομένων αποτελούνται από τα γκρι και μπλε κουτάκια, αντίστοιχα.

3.3 Εκτίμηση της συνάρτησης πυκνότητας πιθανότητας με χρήση ιστογράμματος

Σε αυτήν την ενότητα θεωρούμε το πρόβλημα εκτίμησης της συνάρτησης πυκνότητας πιθανότητας $f(x)$ μέσω ιστογράμματος. Θεωρούμε ένα τυχαίο δείγμα $X_1, \dots, X_n$ από κατανομή με συνάρτηση πυκνότητας πιθανότητας $f(x)$.

Ας υποθέσουμε ότι τα παρατηρηθέντα δεδομένα $x_1, \dots, x_n$ ανήκουν εντός του διαστήματος $[a, b]$. Έστω $m \in \mathbb{Z}_+$ και η διαμέριση του $[a, b]$ σε m διαστήματα (κελιά) ίδιου πλάτους $B_1, \dots, B_m$. Προφανώς, το μήκος κάθε κελιού ισούται με

$$h = \frac{b-a}{m}.$$

Η πιθανότητα να παρατηρήσουμε οποιαδήποτε τιμή εντός του B_j ισούται με

$$p_j := P(X_1 \in B_j) = \int_{B_j} f(x)\mathrm{d}x, \quad j = 1, \dots, m. \tag{3.12}$$

Ένας αμερόληπτος εκτιμητής της p_j είναι, προφανώς, ο αντίστοιχος εκτιμητής αντικατάστασης

$$\hat{p}_j = \frac{n_j}{n}, \tag{3.13}$$

όπου με n_j συμβολίζεται το πλήθος των παρατηρήσεων που ανήκουν εντός του B_j, δηλαδή

$$n_j = \sum_{i=1}^{n} I\{X_i \in B_j\} = \left\{\#X_i : X_i \in B_j\right\}, \quad j = 1, \dots, m. \tag{3.14}$$

Ορισμός 3.4

Η συνάρτηση

$$f_{h,n}(x) = \sum_{j=1}^{m} \frac{\hat{p}_j}{h} I\{x \in B_j\} \tag{3.15}$$

λέγεται εκτιμητής πυκνότητας με χρήση ιστογράμματος με πλάτος κελιών (παράμετρος εξομάλυνσης) $h > 0$.

j	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}	B_{11}	B_{12}
n_j	1	1	1	0	2	1	5	7	5	3	3	1
$f_{h,n}(x)$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	0	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{5}{15}$	$\frac{7}{15}$	$\frac{5}{15}$	$\frac{3}{15}$	$\frac{3}{15}$	$\frac{1}{15}$

Πίνακας 3.1: Ο εκτιμητής $f_{h,n}(x)$ της συνάρτησης πυκνότητας πιθανότητας μέσω ιστογράμματος θεωρώντας τη διαμέριση του διαστήματος $[-3.5, 2.5]$ σε 12 διαστήματα ίσου πλάτους $h = 0.5$ για τα δεδομένα του Παραδείγματος 3.1.

Εφόσον κάθε x ανήκει σε ένα (και μόνο ένα) κελί της διαμέρισης, από τη σχέση (3.15) έπεται ότι

$$f_{h,n}(x) = \frac{\hat{p}_{j_x}}{h},$$

όπου με j_x συμβολίζουμε το κελί που ανήκει το δοθέν x.

Παράδειγμα 3.1. Έστω ένα σύνολο δεδομένων $n = 30$ το πλήθος προσομοιωμένων παρατηρήσεων από την $\mathcal{N}(0,1)$, όπως προκύπτει από τις ακόλουθες εντολές.

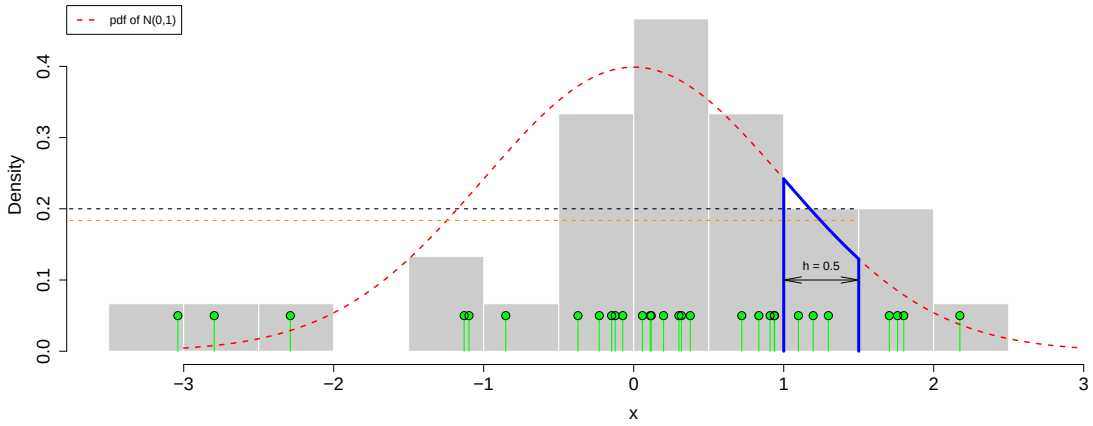
```
1  > n <- 30
2  > set.seed(2020)
3  > x = rnorm(n = n)
4  > x
5   [1]  0.37697212  0.30154837 -1.09802317 -1.13040590 -2.79653432
       0.72057350
6   [7]  0.93912102 -0.22937775  1.75913135  0.11736679 -0.85312282
       0.90925918
7  [13]  1.19637296 -0.37158390 -0.12326023  1.80004312  1.70399588
      -3.03876461
8  [19] -2.28897495  0.05830349  2.17436525  1.09818265  0.31822032
      -0.07314756
9  [25]  0.83426874  0.19875064  1.29784138  0.93671831 -0.14743319
       0.11043199
```

Να υπολογίσετε τον εκτιμητή πυκνότητας μέσω ιστογράμματος και κατόπιν να αναπαρασταθεί γραφικά.

Λύση Παραδείγματος 3.1. Παρατηρούμε ότι τα δεδομένα ανήκουν εντός του διαστήματος $[-3.5, 2.5]$. Έστω η διαμέριση $B_1 = [-3.5, -3), \ldots, B_{12} = [2, 2.5)$ με πλάτος κάθε διαστήματος: $h = 0.5$, το οποίο επιλέγεται αυθαίρετα σε αυτό το παράδειγμα. Θα υπολογίσουμε την εκτίμηση της συνάρτησης πυκνότητας πιθανότητας των δεδομένων με χρήση ιστογράμματος στον Ορισμό 3.4. Για παράδειγμα, ας θεωρήσουμε το διάστημα $B_{10} = [1, 1.5)$. Ο αριθμός των παρατηρήσεων που ανήκουν σε αυτό ισούται με $n_{10} = 3$, άρα από τη σχέση (3.14) έχουμε ότι $\hat{p}_j = \frac{n_j}{n} = \frac{3}{30}$. Επειδή $h = 0.5$, από τη σχέση (3.15) έχουμε ότι η τιμή της $f_{n,h}(x)$ για $x \in B_{10}$ ισούται με $f_{h,n}(x) = \frac{3}{15} = 0.2$. Με αντίστοιχο τρόπο υπολογίζεται η τιμή της $f_{h,n}$ για όλα τα κελιά, όπως παρατίθεται στον Πίνακα 3.1. Τέλος, το Σχήμα 3.3 περιέχει την αντίστοιχη γραφική παράσταση, με τις 30 παρατηρήσεις, οι οποίες αποτελούν το δείγμα, να αντιστοιχούν στις πράσινες κατακόρυφες γραμμές. □

3.3.1 Ιδιότητες εκτιμητή ιστογράμματος

Στην υποενότητα αυτή θα δοθούν κάποιες ιδιότητες του εκτιμητή ιστογράμματος.



Σχήμα 3.3: Οι γκρι ράβδοι αναπαριστούν γραφικά τον εκτιμητή της συνάρτησης πυκνότητας πιθανότητας μέσω ιστογράμματος στον Πίνακα 3.1 για τα δεδομένα (πράσινες κατακόρυφες ράβδοι) του Παραδείγματος 3.1. Η κόκκινη διακεκομμένη γραμμή είναι η συνάρτηση πυκνότητας πιθανότητας της τυπικής κανονικής κατανομής που είναι η πραγματική συνάρτηση πυκνότητας πιθανότητας των δεδομένων.

Πρόταση 3.2. Για σταθερά h και x ισχύει ότι:

$$\mathrm{E}(f_{h,n}(x)) = \frac{p_j}{h} \tag{3.16}$$

και

$$\mathrm{Var}(f_{h,n}(x)) = \frac{p_j(1-p_j)}{nh^2}, \tag{3.17}$$

όπου $p_j = \int_{B_j} f(x)\mathrm{d}x$ και $j = j_x$ το κελί στο οποίο ανήκει το x.

Απόδειξη Πρότασης 3.2. Λόγω της σχέσης (3.14), έχουμε ότι:

$$n_j \sim B(n, p_j).$$

Το αποτέλεσμα προκύπτει άμεσα συνδυάζοντας τη σχέση (1.15) και τις ιδιότητες της διωνυμικής κατανομής. □

Σύμφωνα με τη σχέση (3.16), το ιστόγραμμα είναι **αμερόληπτος εκτιμητής της μέσης πυκνότητας** σε κάθε κελί. Αυτό δεν σημαίνει ότι είναι ένας αμερόληπτος εκτιμητής της f (εκτός αν η f είναι σταθερή στα B_j). Αυτές οι έννοιες εξηγούνται περαιτέρω στο παράδειγμα που ακολουθεί.

Παράδειγμα 3.2. (συνέχεια Παραδείγματος 3.1) Υπολογίστε την εκτίμηση της μέσης πυκνότητας στο κελί B_{10} για τα δεδομένα του Παραδείγματος 3.1 και συγκρίνετέ την με την πραγματική μέση πυκνότητα στο ίδιο διάστημα.

Λύση Παραδείγματος 3.2. Όπως φαίνεται στον Πίνακα 3.1

$$f_{h,n}(x) = \frac{3}{15} = 0.2, \quad x \in B_{10},$$

και αντιστοιχεί στη μαύρη διακεκομμένη γραμμή του Σχήματος 3.3. Θυμηθείτε, στο σημείο αυτό, ότι η πραγματική συνάρτηση κατανομής στο συγκεκριμένο παράδειγμα είναι η $\mathcal{N}(0,1)$, συνεπώς είναι:

$$f(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R},$$

η οποία αντιστοιχεί στην κόκκινη διακεκομμένη καμπύλη του Σχήματος 3.3. Από τη σχέση (3.12), η πιθανότητα να παρατηρήσουμε μία τιμή στο διάστημα B_{10} ισούται με

$$p_{10} = P(X_1 \in B_{10}) = \int_1^{1.5} f(x)\mathrm{d}x \approx 0.092,$$

δηλαδή το εμβαδόν του μπλε χωρίου κάτω από την πυκνότητα της f στο Σχήμα 3.3. Συνεπώς, η αντίστοιχη μέση πυκνότητα της $f(x)$ στο συγκεκριμένο κελί είναι ίση με

$$p_{10}/h = 0.184,$$

και η παραπάνω τιμή αντιστοιχεί στην πορτοκαλί διακεκομμένη γραμμή του Σχήματος 3.3.

Από την Πρόταση 3.2, η τιμή 0.2 (μαύρη γραμμή) είναι η πραγματοποίηση ενός αμερόληπτου εκτιμητή $f_{h,n}(x)$ για $x \in B_{10}$ της τιμής $p_{10}/h = 0.184$ (πορτοκαλί γραμμή) που είναι η (υποτιθέμενη) άγνωστη μέση πυκνότητα στο συγκεκριμένο κελί. Τώρα, πρέπει να έχει καταστεί σαφές ότι το ιστόγραμμα δεν εκτιμά την πυκνότητα σε κάθε κελί, αλλά τη μέση πυκνότητα. □

Το θεώρημα που ακολουθεί (βλ. Scott, 1979) περιγράφει τη μεροληψία, τη διασπορά και το ολοκληρωμένο μέσο τετραγωνικό σφάλμα για τον εκτιμητή ιστογράμματος με πλάτος κελιών h, όταν η πραγματική συνάρτηση πυκνότητας πιθανότητας είναι η $f(x)$.

Παρατήρηση 3.1. Προτού προχωρήσουμε με την παράθεση του Θεωρήματος, κρίνουμε σκόπιμο στο σημείο αυτό να υπενθυμίσουμε κάποιους συμβολισμούς σχετικά με την ασυμπτωτική συμπεριφορά συναρτήσεων. Θα χρησιμοποιήσουμε τον συμβολισμό του Landau, για να περιγράψουμε την ασυμπτωτική συμπεριφορά συναρτήσεων, καθώς το μέγεθος του δείγματος λαμβάνει «μεγάλες» τιμές. Γράφουμε $\alpha_n = O(\beta_n)$, αν υπάρχει n_0 τέτοιο ώστε η $|\alpha_n/\beta_n|$ να είναι φραγμένη για κάθε $n > n_0$. Όπως τονίσαμε στην αρχή της Ενότητας 3.2, η παράμετρος εξομάλυνσης $h > 0$ μπορεί να είναι συνάρτηση του δείγματος, δηλαδή $h = h_n$. Για παράδειγμα, ο συμβολισμός $O(h^2)$ θα πρέπει να ερμηνεύεται ως μια συνάρτηση (του n) της οποίας η απόλυτη τιμή είναι φραγμένη από το Mh_n^2 για «μεγάλα» n, όπου M είναι μια θετική σταθερά.

Θεώρημα 3.1

Έστω ότι $f(x)$ είναι συνεχής συνάρτηση πυκνότητας πιθανότητας με $f'(x)$ και $f''(x)$ συνεχείς και φραγμένες. Τότε

$$\mathrm{bias}\{f_{h,n}(x), f(x)\} = \frac{1}{2}f'(x)[h - 2(x - b_j)] + O(h^2), \tag{3.18}$$

$$\mathrm{Var}\left(f_{h,n}(x)\right) = \frac{f(x)}{nh} + O\left(\frac{1}{n}\right), \tag{3.19}$$

$$\mathrm{IMSE}\{f_{h,n}, f\} = \frac{h^2}{12}\int_{-\infty}^{\infty} f'(u)^2\mathrm{d}u + \frac{1}{nh} + O(h^2) + O\left(\frac{1}{n}\right), \tag{3.20}$$

όπου με b_j συμβολίζεται το αριστερό άκρο του διαστήματος $B_j = [b_j, b_j + h)$.

Απόδειξη Θεωρήματος 3.1. Θεωρώντας το ανάπτυγμα Taylor της f για $u \in B_j$, έχουμε ότι:

$$f(u) = f(x) + (u - x)f'(x) + O(h^2).$$

Αντικαθιστώντας την παραπάνω έκφραση στη σχέση (3.12) προκύπτει ότι:

$$\begin{aligned} p_j = \int_{B_j} f(u)\mathrm{d}u &= \int_{b_j}^{b_j+h} \left(f(x) + (u - x)f'(x) + O(h^2)\right)\mathrm{d}u \\ &= f(x)h + \frac{1}{2}f'(x)[h^2 - 2h(x - b_j)] + O(h^3). \end{aligned}$$

Άρα, λόγω της σχέσης (3.16), η μεροληψία της $f_{h,n}(x)$ ισούται με

$$\begin{aligned}\text{bias}\{f_{h,n}(x), f(x)\} &= \mathsf{E}\left(f_{h,n}(x) - f(x)\right) = \frac{p_j}{h} - f(x) \\ &= \frac{f(x)h + \frac{1}{2}f'(x)[h^2 - 2h(x - b_j)] + O(h^3)}{h} - f(x) \\ &= \frac{1}{2}f'(x)[h - 2(x - b_j)] + O(h^2)\end{aligned}$$

και η απόδειξη της (3.18) ολοκληρώθηκε. Επιπρόσθετα, παρατηρήστε ότι:

$$\left(\text{bias}\{f_{h,n}(x), f(x)\}\right)^2 = \frac{1}{4}f'(x)^2[h - 2(x - b_j)]^2 + O(h^3). \tag{3.21}$$

Για τη διασπορά, λόγω της σχέσης (3.17), έχουμε ότι:

$$\begin{aligned}\text{Var}\left(f_{h,n}(x)\right) &= \frac{p_j(1 - p_j)}{nh^2} \\ &= \frac{\{f(x)h + O(h^2)\}\{1 - O(h)\}}{nh^2} \\ &= \frac{f(x)}{nh} + O\left(\frac{1}{n}\right)\end{aligned}$$

και η απόδειξη της σχέσης (3.19) ολοκληρώθηκε.

Λόγω της σχέσης (3.21) και κάνοντας χρήση του Θεωρήματος Μέσης Τιμής του ολοκληρωτικού λογισμού[2], έχουμε ότι για $\tilde{x}_j \in B_j$

$$\begin{aligned}\int_{B_j}\left(\text{bias}\{f_{h,n}(x), f(x)\}\right)^2 \mathrm{d}x &= \int_{B_j}\frac{1}{4}f'(x)^2[h - 2(x - b_j)]^2\mathrm{d}x + O(h^3) \\ &= \frac{1}{4}f'(\tilde{x}_j)^2\int_{B_j}[h - 2(x - b_j)]^2\mathrm{d}x + O(h^3) \\ &= f'(\tilde{x}_j)^2\frac{h^3}{12} + O(h^3).\end{aligned}$$

Άρα

$$\begin{aligned}\int_{-\infty}^{\infty}\left(\text{bias}\{f_{h,n}(x), f(x)\}\right)^2 \mathrm{d}x &= \sum_{j=1}^{m}\int_{B_j}\text{bias}^2\{f_{h,n}(x), f(x)\}\mathrm{d}x + O(h^3) \\ &= \sum_{j=1}^{m}f'(\tilde{x}_j)^2\frac{h^3}{12} + O(h^3) \\ &= \frac{h^2}{12}\sum_{j=1}^{m}hf'(\tilde{x}_j)^2 + O(h^3) \\ &= \frac{h^2}{12}\int_{-\infty}^{\infty}f'(x)^2\mathrm{d}x + O(h^2).\end{aligned}$$

[2] Η γενικευμένη μορφή του Θεωρήματος Μέσης Τιμής διατυπώνεται ως εξής: Έστω μία συνεχής συνάρτηση $f : [a, b] \to \mathbb{R}$ και g ολοκληρώσιμη συνάρτηση, η οποία διατηρεί σταθερό πρόσημο στο $[a, b]$. Τότε, για κάποιο $c \in (a, b)$ ισχύει ότι $\int_a^b f(x)g(x)\mathrm{d}x = f(c)\int_a^b g(x)\mathrm{d}x$.

Σημειώνεται ότι για τη μετάβαση στην τελευταία γραμμή εκμεταλλευτήκαμε το γεγονός ότι για κάποιο $c_j \in B_j$ θα ισχύει, λόγω του Θεωρήματος Μέσης Τιμής, ότι:

$$\int_{B_j} f'(x)^2 \mathrm{d}x = h f'(c_j)^2, j = 1, \dots, m.$$

Θεωρώντας το ανάπτυγμα Taylor γύρω από το c_j έχουμε ότι $f'(x) = f'(c_j) + O(h)$, από το οποίο προκύπτει ότι:

$$h f'(\tilde{x}_j)^2 = h f'(c_j)^2 + O(h^2) = \int_{B_j} f'(x)^2 \mathrm{d}x + O(h^2).$$

Στη συνέχεια, θεωρούμε το ολοκλήρωμα της διασποράς της $f_{h,n}(x)$. Έχουμε διαδοχικά ότι:

$$\begin{aligned}
\int_{-\infty}^{\infty} \mathrm{Var}\left(f_{h,n}(x)\right) \mathrm{d}x &= \sum_{j=1}^{m} \int_{B_j} \mathrm{Var}\left(f_{h,n}(x)\right) = \sum_{j=1}^{m} \int_{B_j} \frac{p_j(1-p_j)}{nh^2} \\
&= \frac{1}{nh^2} \sum_{j=1}^{m} \int_{B_j} p_j - \frac{1}{nh^2} \sum_{j=1}^{m} \int_{B_j} p_j^2 = \frac{1}{nh} - \frac{1}{nh} \sum_{j=1}^{m} p_j^2 \\
&= \frac{1}{nh} - \frac{1}{nh} \sum_{j=1}^{m} h^2 f^2(\tilde{x}_j) = \frac{1}{nh} - \frac{1}{n} \sum_{j=1}^{m} h f^2(\tilde{x}_j) \\
&= \frac{1}{nh} - \frac{1}{n} \left(\int_{-\infty}^{\infty} f^2(x) \mathrm{d}x + O(1) \right) = \frac{1}{nh} + O\left(\frac{1}{n}\right).
\end{aligned}$$

Η (3.20) προκύπτει μέσω της σχέσης (3.8). □

Τονίζουμε ότι, σύμφωνα με το Θεώρημα 3.1, ο υπολογισμός της μεροληψίας, της διασποράς και του ολοκληρωμένου μέσου τετραγωνικού σφάλματος δεν είναι δυνατός λόγω του ότι η συνάρτηση πυκνότητας πιθανότητας f είναι άγνωστη. Ωστόσο, οι εκφράσεις (3.18) και (3.19) έχουν μία ενδιαφέρουσα ερμηνεία. Ειδικότερα, προκύπτει ότι, καθώς αυξάνεται το πλάτος των κελιών του ιστογράμματος h, μεγαλώνει η μεροληψία, ενώ η μεροληψία της $f_{h,n}$ δεν εξαρτάται από το n. Αντίθετα, καθώς μικραίνει το h, μεγαλώνει η διασπορά. Τέλος, η τιμή που ελαχιστοποιεί ως προς h τη σχέση (3.20) ισούται με

$$h^\star = \frac{1}{n^{1/3}} \left(\frac{6}{\int_{-\infty}^{\infty} f'(u)^2 \mathrm{d}u} \right)^{1/3}. \tag{3.22}$$

Με την παραπάνω επιλογή της παραμέτρου εξομάλυνσης έχουμε ότι:

$$\mathrm{IMSE}(f_{h^\star,n}, f) = O\left(\frac{C}{n^{2/3}}\right), \tag{3.23}$$

όπου $C = (3/4)^{2/3} \left(\int_{-\infty}^{\infty} f'(u)^2 \mathrm{d}u \right)^{1/3}$. Αυτό σημαίνει ότι το IMSE του ιστογράμματος τείνει στο μηδέν με ρυθμό της τάξεως του $n^{(-2/3)}$. Η απόδειξη των παραπάνω σχέσεων είναι άμεση και αφήνεται ως άσκηση στον/στην αναγνώστη/αναγνώστρια (βλ. την Άσκηση 3.1).

Στο επόμενο Θεώρημα διατυπώνεται ότι το ιστόγραμμα είναι ασυμπτωτικά συνεπής εκτιμητής της $f(x)$.

Θεώρημα 3.2

Έστω ότι $f(x)$ συνεχής στο x και $|f'(x)| < M$. Τότε για $h \to 0$ και $nh \to \infty$, καθώς $n \to \infty$ ισχύει ότι:

$$f_{h,n}(x) \xrightarrow{P} f(x).$$

Απόδειξη Θεωρήματος 3.2. Λόγω της σχέσης (3.18), η συνθήκη $h \to 0$ εξασφαλίζει ότι η μεροληψία της $f_{h,n}$ τείνει στο 0. Λόγω της (3.19), η συνθήκη $nh \to \infty$ εξασφαλίζει ότι η διασπορά της $f_{h,n}$ τείνει 0. Άρα, αν ισχύουν οι δύο παραπάνω συνθήκες, τότε η $f_{h,n}(x)$ συγκλίνει κατά μέσο τετράγωνο στην $f(x)$. Η απόδειξη ολοκληρώνεται κάνοντας χρήση του γεγονότος ότι η σύγκλιση κατά μέσο τετράγωνο συνεπάγεται τη σύγκλιση κατά πιθανότητα (για μια απόδειξη αυτού βλ., π.χ. Ρούσσας, 1998, Θεώρημα 10.1). □

3.3.2 Πρακτική επιλογή h μέσω cross-validation

Στην Ενότητα 3.2.2 περιγράψαμε την τεχνική cross-validation για την εκτίμηση της συνάρτησης $J(h)$, η οποία συνάρτηση αντιστοιχεί στο ολοκληρωμένο μέσο τετραγωνικό σφάλμα με τη διαφορά μίας προσθετικής σταθεράς. Έτσι, μπορεί να εκτιμηθεί η παράμετρος εξομάλυνσης h για έναν εκτιμητή πυκνότητας $f_{h,n}(\cdot)$. Σε αυτήν την ενότητα, θα εφαρμόσουμε τη συγκεκριμένη τεχνική για την επιλογή του πλάτους h των κελιών ενός ιστογράμματος. Το θετικό είναι ότι στην περίπτωση του ιστογράμματος, η εκτίμηση (3.11) της ποσότητας $J(h)$ είναι διαθέσιμη αναλυτικά, όπως περιγράφεται στην επόμενη πρόταση.

Πρόταση 3.3. Η cross-validation εκτίμηση (3.11) της ποσότητας $J(h)$ (3.9), για ιστόγραμμα με πλάτος κελιών h είναι ίση με

$$\hat{J}(h) = \frac{2}{h(n-1)} - \frac{n+1}{h(n-1)} \sum_{j=1}^{m} \hat{p}_j^2. \tag{3.24}$$

Απόδειξη Πρότασης 3.3. Αρχικά, να παρατηρήσουμε ότι για το ιστόγραμμα ισχύει ότι:

$$f_{h,n-1}^{(-i)}(x_i) = \frac{n_{j_i} - 1}{h(n-1)} = \frac{n}{n-1} \frac{n_{j_i} - 1}{hn} = \frac{n}{n-1} \left(\frac{n_{j_i}}{hn} - \frac{1}{hn} \right) = \frac{n}{n-1} \left(\frac{\hat{p}_{j_i}}{h} - \frac{1}{hn} \right) \tag{3.25}$$

όπου $j_i := \{j = 1, \dots, m : x_i \in B_j\}$. Από τη σχέση (3.11) έχουμε ότι:

$$\begin{aligned}
\hat{J}(h) &= \int f_{h,n}^2(x) \mathrm{d}x - 2 \sum_{i=1}^{n} \frac{1}{n} f_{h,n-1}^{(-i)}(x_i) \\
&= \sum_{j=1}^{m} \int_{B_j} f_{h,n}^2(x) \mathrm{d}x - \frac{2}{n} \sum_{i=1}^{n} f_{h,n-1}^{(-i)}(x_i) \\
&= \sum_{j=1}^{m} \int_{B_j} \frac{\hat{p}_j^2}{h^2} \mathrm{d}x - \frac{2}{n} \sum_{i=1}^{n} \frac{n}{n-1} \left(\frac{\hat{p}_{j_i}}{h} - \frac{1}{hn} \right)
\end{aligned}$$

όπου στην τελευταία ισότητα χρησιμοποιήθηκε η σχέση (3.25). Είναι τότε:

$$\begin{aligned}\hat{J}(h) &= \frac{2}{h(n-1)} + \frac{1}{h}\sum_{j=1}^{m}\hat{p}_j^2 - \frac{2}{h(n-1)}\sum_{i=1}^{n}\hat{p}_{j_i} \\ &= \frac{2}{h(n-1)} + \frac{1}{h}\sum_{j=1}^{m}\hat{p}_j^2 - \frac{2}{h(n-1)}\sum_{j=1}^{m}\sum_{i:x_i\in B_j}\hat{p}_j \\ &= \frac{2}{h(n-1)} + \frac{1}{h}\sum_{j=1}^{m}\hat{p}_j^2 - \frac{2}{h(n-1)}\sum_{j=1}^{m}n\hat{p}_j^2 \\ &= \frac{2}{h(n-1)} - \frac{n+1}{h(n-1)}\sum_{j=1}^{m}\hat{p}_j^2\end{aligned}$$

και αυτό ολοκληρώνει την απόδειξη. □

Με βάση την Πρόταση 3.3, η διαδικασία για την πρακτική επιλογή του πλάτους κελιών h σε ένα ιστόγραμμα έχει ως εξής:

1. Θεώρησε ένα διακριτό σύνολο ν τιμών $h \in \{h_1, \dots, h_\nu\}$.
2. Υπολόγισε την $\hat{J}(h_i)$, για κάθε $i = 1, \dots, \nu$.
3. Επίλεξε την τιμή $h = h_{i^\star}$, όπου
$$i^\star = \text{argmin}_{i=1,\dots,\nu}\{\hat{J}(h_i)\},$$
δηλαδή επίλεξε εκείνη την τιμή του πλάτους κελιών που (προσεγγιστικά) ελαχιστοποιεί την $\hat{J}(h)$.

Προφανώς, η επιλογή του συνόλου των τιμών $\{h_1, \dots, h_\nu\}$ εξαρτάται από την εκάστοτε περίπτωση. Γενικά, θα πρέπει πάντα να ελέγχουμε αν το ελάχιστο επιτυγχάνεται σε κάποια ενδιάμεση τιμή του συνόλου που θεωρήσαμε και όχι σε κάποιο από τα δύο άκρα. Σε αυτές τις περιπτώσεις, θα πρέπει να αυξήσουμε το εύρος ή/και το πλήθος των δυνατών τιμών.

Εναλλακτικά, μπορούμε να θεωρήσουμε ένα σύνολο δυνατών τιμών του **πλήθους κελιών** αντί του πλάτους κελιών, δοθέντος ενός διαστήματος $[a,b]$ που περιέχει τα δεδομένα. Κατόπιν, εφαρμόζουμε την προηγούμενη διαδικασία, θεωρώντας το αντίστοιχο διάστημα τιμών του πλάτους κελιών. Αυτή η πρακτική είναι, συνήθως, προτιμότερη, καθώς έτσι επιλέγεται ένας φυσικός αριθμός ν, ο οποίος αντιστοιχεί στο μέγιστο πλήθος κελιών, και μετά αρκεί να θεωρήσουμε όλες τις τιμές στο (διακριτό) σύνολο $\{1, 2 \dots, \nu\}$.

Παράδειγμα 3.3 (συνέχεια Παραδείγματος 3.1)**.** Να υπολογίσετε το βέλτιστο πλάτος κελιών για το ιστόγραμμα στα δεδομένα του Παραδείγματος 3.1, θεωρώντας διαμερίσεις του διαστήματος $[-3.5, 2.5]$, με πλήθος κελιών $m = 1, \dots, 14$.

Λύση Παραδείγματος 3.3. Αρχικά, θα θεωρήσουμε ότι το διάστημα $[-3.5, 2.5]$ διαμερίζεται σε m ίσα διαστήματα, όπου $m = 1, 2, 3, \dots, 14$. Για καθεμία από τις τιμές αυτές το πλάτος των κελιών που προκύπτει είναι ίσο με $h_m = 6/m$. Υπολογίζοντας την (3.24) προκύπτουν οι τιμές των δύο πρώτων στηλών στον Πίνακα 3.2. Η ελάχιστη τιμή του εκτιμηθέντος ρίσκου (ως προς μία προσθετική σταθερά) ισούται με -0.236 και επιτυγχάνεται για $h = 1.5$. Το αντίστοιχο διάγραμμα των τιμών της $\hat{J}(h)$) μαζί με το ιστόγραμμα που προκύπτει για $h = 1.5$ παρατίθενται στα Σχήματα 3.4 (α) και 3.4 (β), αντίστοιχα. □

Ένα μειονέκτημα του ιστογράμματος, το οποίο δεν έχει τονιστεί μέχρι τώρα, είναι ότι, εκτός από το πλάτος των κελιών, εξαρτάται και από το αρχικό σημείο της διαμέρισης. Η ιδιότητα αυτή θα αναδειχθεί μέσω του παραδείγματος που ακολουθεί.

	$x_0 = -3.5$		$x_0 = -3.65$	
	h	$\hat{J}(h)$	h	$\hat{J}(h)$
1	6.000	-0.167	6.150	-0.163
2	3.000	-0.219	3.075	-0.214
3	2.000	-0.190	2.050	-0.185
4	1.500	-0.236	1.538	-0.230
5	1.200	-0.224	1.230	-0.224
6	1.000	-0.214	1.025	-0.209
7	0.857	-0.161	0.879	-0.178
8	0.750	-0.199	0.769	-0.195
9	0.667	-0.217	0.683	-0.240
10	0.600	-0.190	0.615	-0.181
11	0.545	-0.183	0.559	-0.200
12	0.500	-0.161	0.513	-0.162
13	0.462	-0.180	0.473	-0.166
14	0.429	-0.177	0.439	-0.157

Πίνακας 3.2: Εκτίμηση μέσου σφάλματος ιστογράμματος $f_{h,n}$ μέσω cross-validation (3.24) για τα δεδομένα του Παραδείγματος 3.1 για δύο διαφορετικά αρχικά σημεία (x_0) της διαμέρισης.

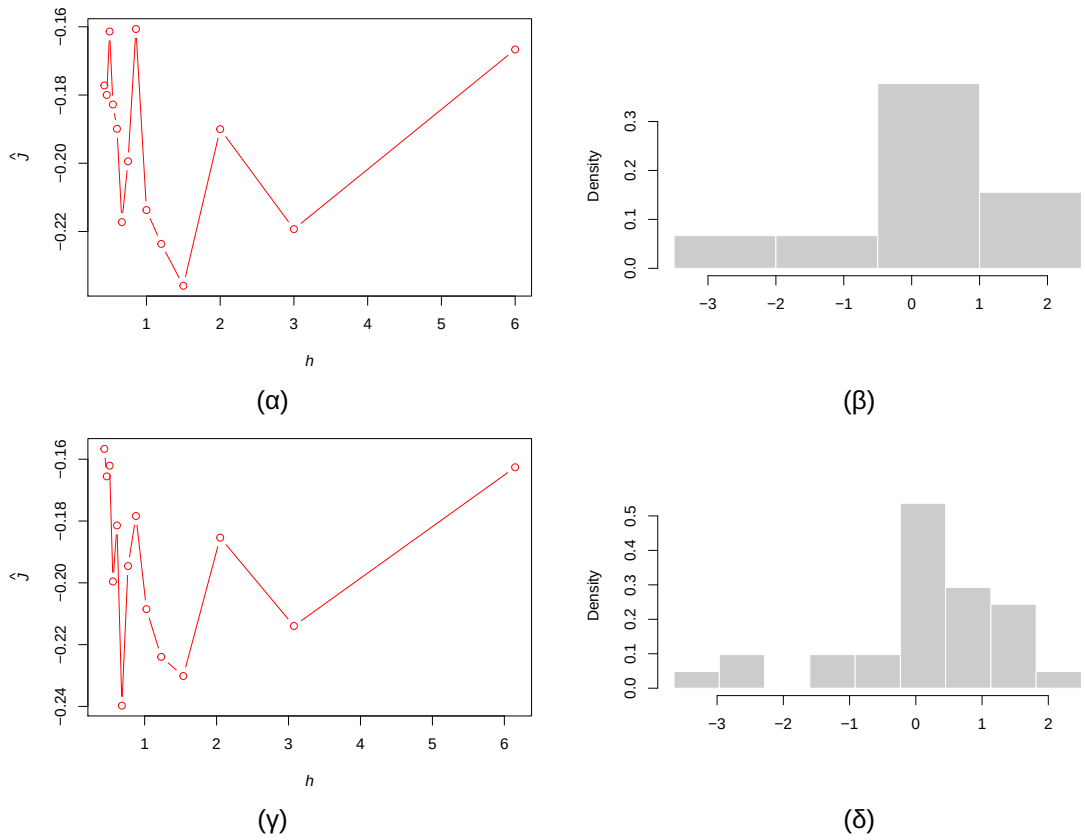
Παράδειγμα 3.4 (συνέχεια Παραδείγματος 3.3)**.** Να υπολογίσετε το βέλτιστο πλάτος κελιών για το ιστόγραμμα στα δεδομένα του Παραδείγματος 3.1, θεωρώντας διαμερίσεις του διαστήματος $[-3.65, 2.5]$, με πλήθος κελιών $m = 1, ..., 14$.

Λύση Παραδείγματος 3.4. Επαναλαμβάνοντας την ίδια διαδικασία με το προηγούμενο παράδειγμα, προκύπτουν οι τιμές των δύο τελευταίων στηλών στον Πίνακα 3.2. Σε αυτήν την περίπτωση, η ελάχιστη τιμή του εκτιμηθέντος ρίσκου (ως προς μία προσθετική σταθερά) ισούται με -0.240 και επιτυγχάνεται για $h = 0.683$. Το αντίστοιχο διάγραμμα των τιμών της $\hat{J}(h)$ μαζί με το ιστόγραμμα που προκύπτει για $h = 0.683$ παρατίθεται στα Σχήματα 3.4 (γ) και 3.4 (δ), αντίστοιχα. Συγκρίνοντας με το βέλτιστο ιστόγραμμα στο Σχήμα 3.4 (γ) παρατηρούμε ότι, τώρα, το πλάτος των κελιών είναι πάνω από δύο φορές μεγαλύτερο σε σχέση με πριν και, συνεπώς, υπάρχουν εμφανείς διαφορές στα δύο ιστογράμματα. □

Ο βασικός σκοπός μιας μη παραμετρικής στατιστικής ανάλυσης είναι το να «αφήσουμε τα δεδομένα να μιλήσουν για τον εαυτό τους». Ωστόσο, από τα προηγούμενα παραδείγματα φαίνεται ότι τα ίδια δεδομένα μπορεί να μην λένε το ίδιο. Για τον σκοπό αυτό έχουν προταθεί πιο σύνθετες τεχνικές που βασίζονται στην ιδέα κατασκευής ενός «μέσου ιστογράμματος» θεωρώντας διαφορετικές αρχικές τιμές για τη διαμέριση. Ο/Η ενδιαφερόμενος/μενη αναγνώστης/στρια παραπέμπεται στο πρώτο κεφάλαιο του συγγράμματος του Härdle (1991).

3.3.3 Εφαρμογές στην `R`

Σε αυτήν την ενότητα θα περιγράψουμε αρχικά την εντολή `hist()`, η οποία χρησιμοποιείται τόσο για τον υπολογισμό της εκτιμηθείσας πυκνότητας $f_{h,n}(x)$ όσο και για τη γραφική αναπαράσταση αυτής. Για να δούμε τα βασικά χαρακτηριστικά αυτής της εντολής, αρχικά προσομοιώνουμε ένα σύνολο δεδομένων 20 το πλήθος τιμών από την τυπική κανονική κατανομή και έπειτα ακολουθούμε τις εντολές που παρατίθενται παρακάτω.



Σχήμα 3.4: Πρώτη γραμμή: εκτιμηθέν μέσο ολοκληρωμένο σφάλμα (α) και βέλτιστο ιστόγραμμα (β) θεωρώντας ως αρχή της διαμέρισης το σημείο $x_0 = -3.5$. Δεύτερη γραμμή: αντίστοιχα διαγράμματα για $x_0 = -3.65$.

```
> x <- rnorm(20)
> f <- hist(x, freq = F)
> f$density
[1] 0.1 0.4 0.7 0.3 0.4 0.1
> f <- hist(x, breaks = c(-2,-1,0,1,2), freq = F)
> f$density
[1] 0.05 0.55 0.35 0.05
```

Οι τιμές `f$density` επιστρέφουν τις τιμές $\frac{\hat{p}_j}{h}$ για καθένα κελί της διαμέρισης. Το πλήθος των κελιών δεν επιλέγεται μέσω του κριτηρίου που περιγράψαμε προηγουμένως. Όμως, μέσω του ορίσματος `breaks = ...` μπορούμε να ορίσουμε ρητά τα κελιά του ιστογράμματος, οπότε να χρησιμοποιήσουμε τα αποτελέσματα της εντολής για περαιτέρω ανάλυση.

Παράδειγμα 3.5. Προσομοιώστε ένα σύνολο δεδομένων 1000 το πλήθος παρατηρήσεων από την κατανομή με συνάρτηση πυκνότητας πιθανότητας που δίνεται στη σχέση (3.1) και την οποία χρησιμοποιήσαμε στην Ενότητα 3.2[3]. Να υπολογίσετε το βέλτιστο πλάτος κελιών για τον εκτιμητή του ιστογράμματος μέσω του εκτιμητή (3.11), ενώ ταυτόχρονα να υπολογίσετε και τον εκτιμητή αντικατάστασης του ολοκληρωμένου μέσου τετραγωνικού σφάλματος που δόθηκε στη σχέση (3.10). **Υπόδειξη:** Θεωρήστε τη διαμέριση του συνόλου των δεδομένων που ορίζεται από την ελάχιστη και τη μέγιστη παρατήρηση

[3]Η διαδικασία προσομοίωσης παρατίθεται για λόγους αναπαραγωγής του συνόλου δεδομένων, χωρίς αναλυτική περιγραφή του μηχανισμού προσομοίωσης.

αυτού, χρησιμοποιώντας m διαστήματα ίσου πλάτους για $m = 7, 8, \ldots, 407$.

Λύση Παραδείγματος 3.5. Ο κώδικας για την προσομοίωση των δεδομένων είναι ο ακόλουθος.

```
> p <- c(0.5,rep(0.1,5))
> mu <- s <- numeric(6)
> mu[1] <- 0
> s[1] <- 1
> for(j in 2:6){
+         mu[j] <- (j-2)/2 - 1
+         s[j] <- 0.1
+ }
> n <- 1000
> set.seed(10)
> x <- numeric(n)
> for(i in 1:n){
+         j <- sample(1:6, 1, prob = p)
+         x[i] <- rnorm(1, mu[j], s[j])
+ }
```

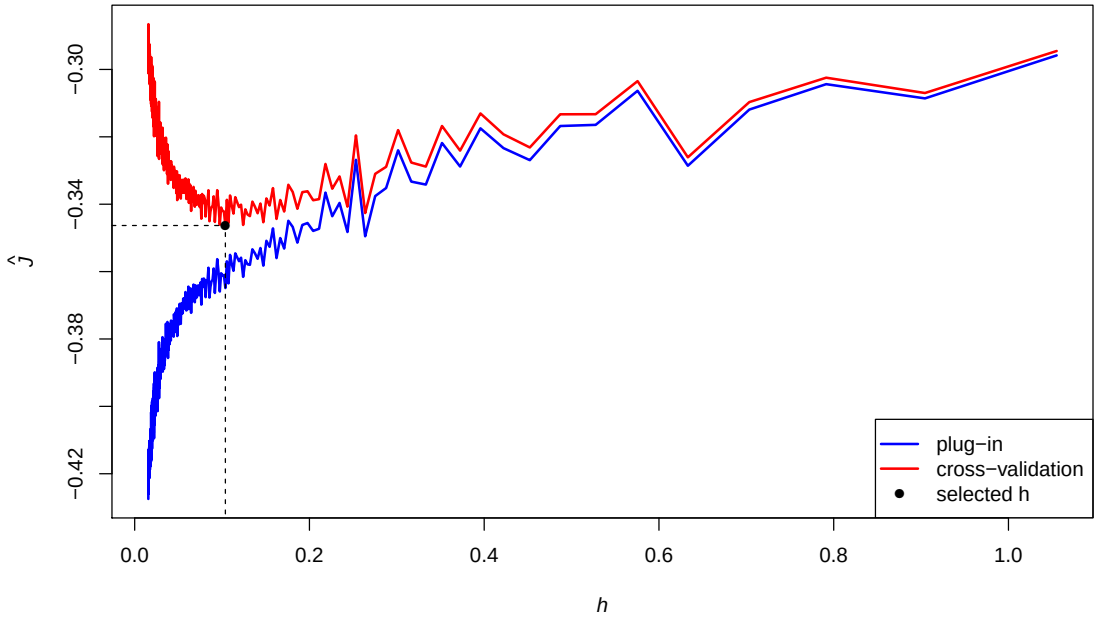
Τα προσομοιωμένα δεδομένα έχουν αποθηκευτεί στη μεταβλητή `x`. Θα θεωρήσουμε τη διαμέριση του συνόλου που ορίζεται από την ελάχιστη και τη μέγιστη παρατήρηση του συνόλου δεδομένων, δηλαδή τη διαμέριση του συνόλου $[-2.51843, 3.81258]$, θεωρώντας m διαστήματα ίσου πλάτους, με $m = 7, 8, \ldots, 407$. Η επιλογή των τιμών m γίνεται για λόγους καλύτερης παρουσίασης της συμπεριφοράς των τιμών h. Θα μπορούσαμε να επιλέξουμε π.χ. το $m = 2, 3, \ldots, 412$. Σε μια τέτοια περίπτωση, όμως, ειδικά για τις μικρές τιμές του m, τα όρια στον άξονα $y'y$ είναι δυσανάλογα μεγάλα και αυτό δυσχεραίνει την ερμηνεία του διαγράμματος.

```
> m_min <- 7
> m_max <- m_min+400
> J_cv <- J_plugin <- hValues <- numeric(m_max-m_min)
> j <- 0
> for(m in m_min:m_max){
+         j <- j + 1
+         breaks = seq(min(x), max(x), length = m)
+         my_hist <- hist(x, breaks = breaks, plot = FALSE)
+         hValues[j] <- h <- (max(x) - min(x))/(m-1)
+         J_cv[j] <- 2/(h*(n-1)) - (n+1)/(h*(n-1))*sum((h*my_hist$
    density)^2)
+         J_plugin[j] <- -1/h*sum((h*my_hist$density)^2)
+ }
> min(J_cv)
[1] -0.3463227
> h <- hValues[which.min(J_cv)]
> h
[1] 0.103787
> m <- (max(x) - min(x))/h + 1
> m
[1] 62
```

Παρατηρούμε ότι η εκτίμηση της ελάχιστης τιμής του εκτιμηθέντος μέσου σφάλματος (3.24) μέσω cross-validation ισούται με

$$\hat{J}(h) \approx -0.346,$$

η οποία επιτυγχάνεται για πλάτος κελιών (παράμετρος εξομάλυνσης) ίσο με $h \approx 0.104$. Αυτή η τιμή αντιστοιχεί σε $m = 62$ κελιά. Οι παρακάτω εντολές κατασκευάζουν το Σχήμα 3.5.



Σχήμα 3.5: Κόκκινη γραμμή: Εκτίμηση μέσου σφάλματος ιστογράμματος μέσω cross-validation (σχέση (3.24)) για τα δεδομένα του Παραδείγματος 3.5. Η μπλε γραμμή αντιστοιχεί στον εκτιμητή αντικατάστασης (σχέση (3.10)).

```
> yli <- range(cbind(J_plugin,J_cv))
> plot(hValues, J_plugin, ylim = yli, type = "l",
+ ylab = bquote(hat(italic(J))),
+ xlab = bquote(italic(h)), col = 'blue', lwd = 2)
> points(hValues,J_cv,type = "l",col = 'red', lwd = 2)
> points(c(-1, h), c(min(J_cv), min(J_cv)), type = "l", lty = 2)
> points(c(h,h), c(-1, min(J_cv)), type = "l", lty = 2)
> points(h,min(J_cv), pch = 16, col = 1)
> legend('bottomright',c('plug-in','cross-validation','selected h'),
+ lty=c(1,1,NA), pch = c(NA, NA, 16),col=c('blue','red','black'), lwd
   = 2)
```

Παρατηρήστε ότι για τον εκτιμητή αντικατάστασης (3.10) το βέλτιστο $h \to 0$ (αναμενόμενο). Τέλος, οι παρακάτω εντολές απεικονίζουν το ιστόγραμμα των δεδομένων που αντιστοιχεί στη βέλτιση επιλογή του πλάτους κελιών βάσει του ολοκληρωμένου μέσου τετραγωνικού σφάλματος.

```
> breaks = seq(min(x), max(x), length = m)
> hist(x, freq = F, breaks = breaks,col = "gray80", border = "white",
+ main = "", xlab = "", cex.axis = 1.5)
```

Το αποτέλεσμα είναι παρόμοιο (η μόνη διαφορά είναι στο εύρος του οριζόντιου άξονα) με αυτό που παρατίθεται στο Σχήμα 3.1 (γ). □

3.4 Εκτίμηση της συνάρτησης πυκνότητας πιθανότητας με χρήση πυρήνα

Σε αυτήν την ενότητα, αρχικά, περιγράφεται ένας απλοϊκός εκτιμητής της συνάρτησης πυκνότητας πιθανότητας, ο οποίος δίνει τη διαίσθηση για την εισαγωγή ενός γενικότερου εκτιμητή με χρήση αυτού που λέμε πυρήνα.

3.4.1 Εισαγωγή στην έννοια του πυρήνα

Έστω συνεχής τυχαία μεταβλητή X, με συνάρτηση πυκνότητας πιθανότητας $f(x)$ και αθροιστική συνάρτηση κατανομής $F(x)$, $x \in \mathbb{R}$. Η συνάρτηση πυκνότητας πιθανότητας μπορεί να εκφραστεί ως:

$$f(x_0) = \lim_{h\to 0} \frac{P(x_0 - h < X < x_0 + h)}{2h} \tag{3.26}$$

διότι

$$\begin{aligned}
\lim_{h\to 0} \frac{P(x_0 - h < X < x_0 + h)}{2h} &= \lim_{h\to 0} \frac{F(x_0 + h) - F(x_0 - h)}{2h} \\
&= \frac{1}{2}\left(\lim_{h\to 0} \frac{F(x_0 + h)}{h} - \lim_{h\to 0} \frac{F(x_0 - h)}{h}\right) \\
&= \frac{1}{2}\left(\lim_{h\to 0} \frac{F(x_0 + h) - F(x_0)}{h} - \lim_{h\to 0} \frac{F(x_0 - h) - F(x_0)}{h}\right) \\
&\overset{\text{θέτω } \ell=-h}{=} \frac{1}{2}\left(\lim_{h\to 0} \frac{F(x_0 + h) - F(x_0)}{h} + \lim_{\ell\to 0} \frac{F(x_0 + \ell) - F(x_0)}{\ell}\right) \\
&= \frac{1}{2}\left(F'(x_0) + F'(x_0)\right) = f(x_0).
\end{aligned}$$

Έστω, τώρα, ότι έχουμε στη διάθεσή μας ένα τυχαίο δείγμα μεγέθους n από την f. Το ερώτημα που τίθεται είναι ποιο είναι το δειγματικό «ανάλογο» της (3.26). Διαισθητικά, για μία δοθείσα παράμετρο $h > 0$, θα πρέπει να χρησιμοποιηθεί η

$$f_{h,n}(x_0) = \frac{1}{n} \frac{\text{πλήθος παρατηρήσεων στο διάστημα } (x_0 - h, x_0 + h)}{2h}. \tag{3.27}$$

Η τελευταία ποσότητα ονομάζεται απλοϊκός (naive) εκτιμητής της συνάρτησης πυκνότητας πιθανότητας f, με παράμετρο εξομάλυνσης h. Η γραφική παράσταση αυτής για δοθείσες τιμές των x_0 και h μοιάζει με ένα ορθογώνιο, το οποίο, καθώς μεταβάλλεται το x_0, θα διανύει τον οριζόντιο άξονα. Για αυτόν τον λόγο, ο απλοϊκός εκτιμητής είναι και γνωστός ως boxcar. Σε αυτό το σημείο δίνουμε τον ακόλουθο ορισμό.

Ορισμός 3.5

Η συνάρτηση

$$f_{h,n}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{I\{|X_i - x| \leqslant h\}}{2h}, \quad x \in \mathbb{R} \tag{3.28}$$

ονομάζεται απλοϊκός εκτιμητής της συνάρτησης πυκνότητας πιθανότητας $f(x)$, με παράμετρο εξομάλυνσης $h > 0$.

Παράδειγμα 3.6. Έστω τυχαίο δείγμα

$$\boldsymbol{x} = (-0.07, -1.36, -0.75, -0.12, 0.56, -0.94, 0.08, 1.09, 2.16, -0.82)$$

Να εκτιμηθεί η $f(1)$ με τον απλοϊκό εκτιμητή και παράμετρο εξομάλυνσης $h = 0.5$ και $h = 1$.

Λύση Παραδείγματος 3.6. Έχουμε $n = 10$ και $x_0 = 1$, άρα από την (3.28)

$$f_{h,10}(1) = \frac{1}{10}\sum_{i=1}^{10}\frac{I\{|x_i - 1| \leqslant h\}}{2h}.$$

Για $h = 0.5$ παρατηρούμε ότι στο διάστημα $[x_0 - h, x_0 + h] = [0.5, 1.5]$ ανήκουν 2 παρατηρήσεις, ήτοι: $0.56, 1.09$. Άρα

$$\sum_{i=1}^{10} I\{|x_i - 1| \leqslant 0.5\} = 2 \Rightarrow f_{0.5,10}(1) = \frac{1}{10}\frac{2}{2\times 0.5} = 0.2.$$

Για $h = 1$ παρατηρούμε ότι στο διάστημα $[x_0 - h, x_0 + h] = [0, 2]$ ανήκουν 3 παρατηρήσεις, ήτοι οι: $0.56, 0.08, 1.09$. Άρα

$$\sum_{i=1}^{10} I\{|x_i - 1| \leqslant 1\} = 3 \Rightarrow f_{1,10}(1) = \frac{1}{10}\frac{3}{2\times 1} = 0.15.$$

□

Μία εναλλακτική έκφραση για τον απλοϊκό εκτιμητή της σχέσης (3.28) είναι η εξής:

$$f_{h,n}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}, \tag{3.29}$$

όπου

$$K(u) := \frac{1}{2} I\{|u| \leqslant 1\}.$$

Παρατηρήστε ότι η $K(u)$ είναι η συνάρτηση πυκνότητας πιθανότητας $U(-1,1)$ και είναι ένα παράδειγμα αυτού που θα ορίσουμε στη συνέχεια ως πυρήνα (kernel). Η $K(u)$ ονομάζεται **απλοϊκός πυρήνας** (boxcar kernel). Η $f_{h,n}(x)$ που δόθηκε στη σχέση (3.29) καλείται **εκτιμητής πυκνότητας απλοϊκού πυρήνα με παράμετρο εξομάλυνσης** h. Στην ενότητα που ακολουθεί θα δούμε ότι υπάρχουν και άλλες επιλογές για τη μορφή του πυρήνα.

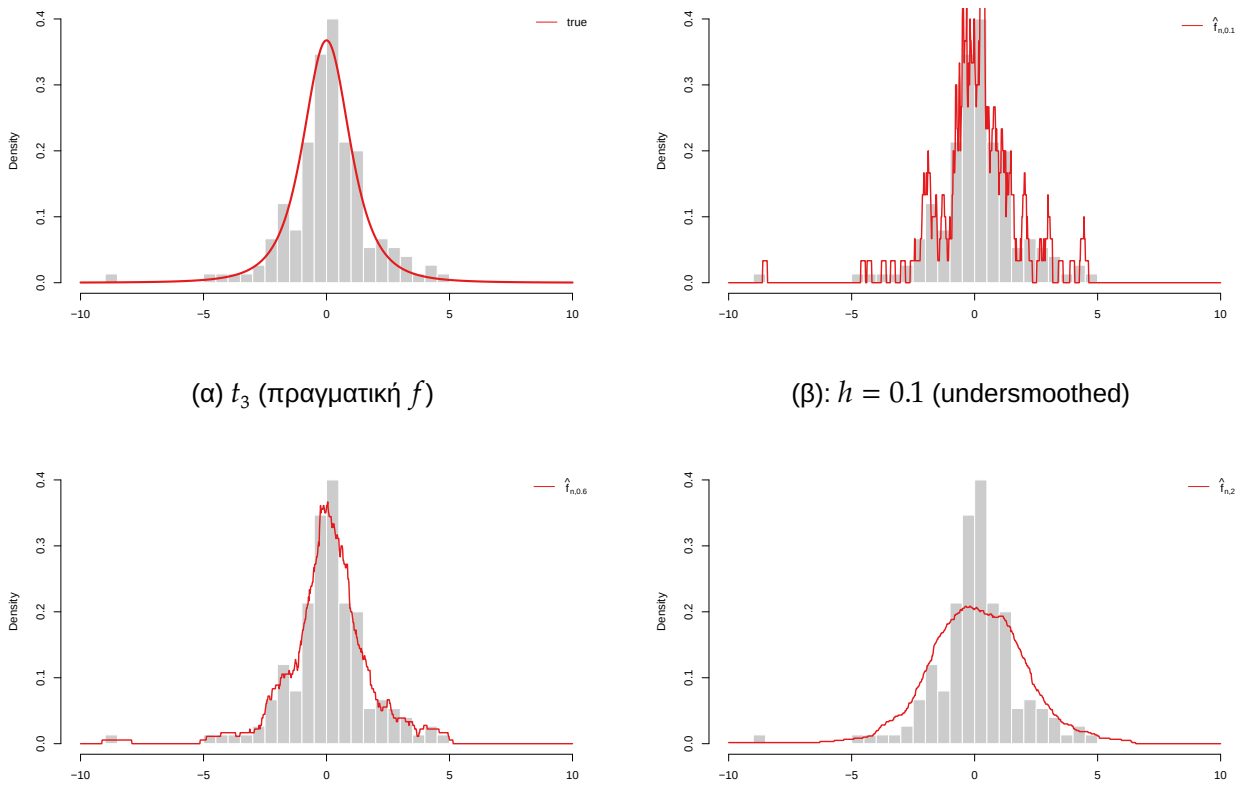
Παράδειγμα 3.7. Προσομοιώστε $n = 150$ το πλήθος παρατηρήσεις από την κατανομή t_3. Στη συνέχεια, να ορίσετε μία συνάρτηση στην `R`, η οποία θα υπολογίζει τον απλοϊκό εκτιμητή πυκνότητας της σχέσης (3.28) και να κατασκευάσετε τη γραφική παράσταση για διαφορετικές τιμές της παραμέτρου εξομάλυνσης μαζί με το ιστόγραμμα των δεδομένων.

Λύση Παραδείγματος 3.7. Αρχικά, προσομοιώνουμε τα δεδομένα μέσω των ακόλουθων εντολών.

```
n <- 150
set.seed(20)
x = rt(n = n, df = 3)
```

Στη συνέχεια, ορίζουμε μία συνάρτηση που υπολογίζει τον απλοϊκό εκτιμητή της σχέσης (3.28).

```
w <- function(y, h, x){
  n <- length(x)
  apply(
    matrix(rep(y, n), ncol = n),
    1,
    function(z){sum( dunif((z-x)/h, min = -1, max = 1) )/(n*h)}
    )
}
```



(α) t_3 (πραγματική f)

(β): $h = 0.1$ (undersmoothed)

(γ): $h = 0.6$ (όχι και άσχημα)

(δ): $h = 2$ (oversmoothed)

Σχήμα 3.6: (α): Ιστόγραμμα των δεδομένων και πραγματική συνάρτηση πυκνότητας πιθανότητας. (β), (γ), (δ): Απλοϊκός εκτιμητής $f_{h,n}$ πυκνότητας για τα δεδομένα του Παραδείγματος 3.7 για διαφορετικές τιμές της παραμέτρου εξομάλυνσης h.

Η συνάρτηση `w()` δέχεται ως ορίσματα μία ακολουθία τιμών `y` στις οποίες θα υπολογιστεί η (3.28), την παράμετρο εξομάλυνσης h και τα παρατηρηθέντα δεδομένα `x`. Η συνάρτηση επιστρέφει ένα διάνυσμα ίδιου μήκους με το `y`. Για παράδειγμα, αν επιθυμούμε να υπολογίσουμε τις τιμές του απλοϊκού εκτιμητή $f_{h,n}(x)$ με παράμετρο εξομάλυνσης $h = 0.6$ για $x = -5, -3, -1, 1, 3, 5$ τότε μπορούμε να εκτελέσουμε την ακόλουθη εντολή. Σημειωτέον ότι σε σχέση με την εξίσωση 3.29 το διάνυσμα `x` αντιστοιχεί στα X_i.

```
1 > w(seq(-5,5,by = 2), 0.6, x)
2 [1] 0.005555556 0.016666667 0.161111111 0.200000000 0.038888889
    0.005555556
```

Ένα ιστόγραμμα των δεδομένων παρατίθεται στο Σχήμα 3.6.(α) μαζί με τη συνάρτηση πυκνότητας πιθανότητας της κατανομής t_3, που είναι η «πραγματική» συνάρτηση πυκνότητας πιθανότητας $f(x)$. Στα Σχήματα 3.6.(β), 3.6.(γ) και 3.6.(δ) παρατίθεται η γραφική παράσταση του απλοϊκού εκτιμητή $f_{h,n}$ (3.28) με παραμέτρους εξομάλυνσης (εύρος παραθύρου) $h = 0.1, 0.6$ και 2, αντίστοιχα.

Οι συγκεκριμένες τιμές του h επιλέχθηκαν αυθαίρετα σε αυτό το παράδειγμα και αναδεικνύουν ότι, όταν η παράμετρος εξομάλυνσης είναι «μικρή» ($h = 0.1$), η εκτίμηση της πυκνότητας προσπαθεί να προσαρμοστεί πολύ «κοντά» στα παρατηρηθέντα δεδομένα και τονίζει τοπικά χαρακτηριστικά αυτών, ίσως παραπάνω από όσο πρέπει. Αυτή η συμπεριφορά είναι τυπική σε περιπτώσεις εκτιμητών με μικρή μεροληψία, αλλά μεγάλη διασπορά (undersmoothing). Από την άλλη μεριά, όταν η παράμετρος εξομάλυνσης είναι «μεγάλη» ($h = 2$), προκύπτει ένας εκτιμητής ο οποίος έχει μικρή διασπορά, αλλά μεγάλη μεροληψία (oversmoothing). Όπως

και στο ιστόγραμμα, το πρόβλημα είναι η κατάλληλη επιλογή του h, έτσι ώστε να επιλέξουμε μία ενδιάμεση τιμή (ούτε πολύ «μικρή» ούτε πολύ «μεγάλη»), ώστε να ισοσταθμίζονται η διασπορά και η μεροληψία του εκτιμητή, όπως για παράδειγμα καταφέρνει η τιμή $h = 0.6$, που αντιστοιχεί στο Σχήμα 3.6.(γ).

Οι παρακάτω εντολές δημιουργούν τα Σχήματα 3.6 (α) και 3.6 (γ).

```
1  densEvaluations <- 1024
2  xSeq <- seq(-10, 10,length = densEvaluations)
3  trueDens <- dt(xSeq, df = 3)
4  library(RColorBrewer)
5  myCol <- brewer.pal(6, "Set1")
6  par(mfrow = c(1, 2))
7  hist(x, 19, freq = F, xlim = c(-10,10), col = "gray80", border = "
       white", main = "", xlab = "")
8  points(xSeq, trueDens, type = "l", col = myCol[1], lwd = 3, lty = 1)
9  legend("topright", "true", lty = 1, col = myCol[1], bty = "n")
10 hist(x, 19, freq = F, xlim = c(-10,10), col = "gray80", border = "
       white", main = "", xlab = "")
11 points(xSeq, w(xSeq, h = 0.6, x), type = "l", lwd = 2, col = myCol[1])
12 legend("topright", legend = bquote(hat(f)["n,0.6"]), lty = 1, col =
       myCol[1], bty = "n")
```

□

3.4.2 Πυρήνες

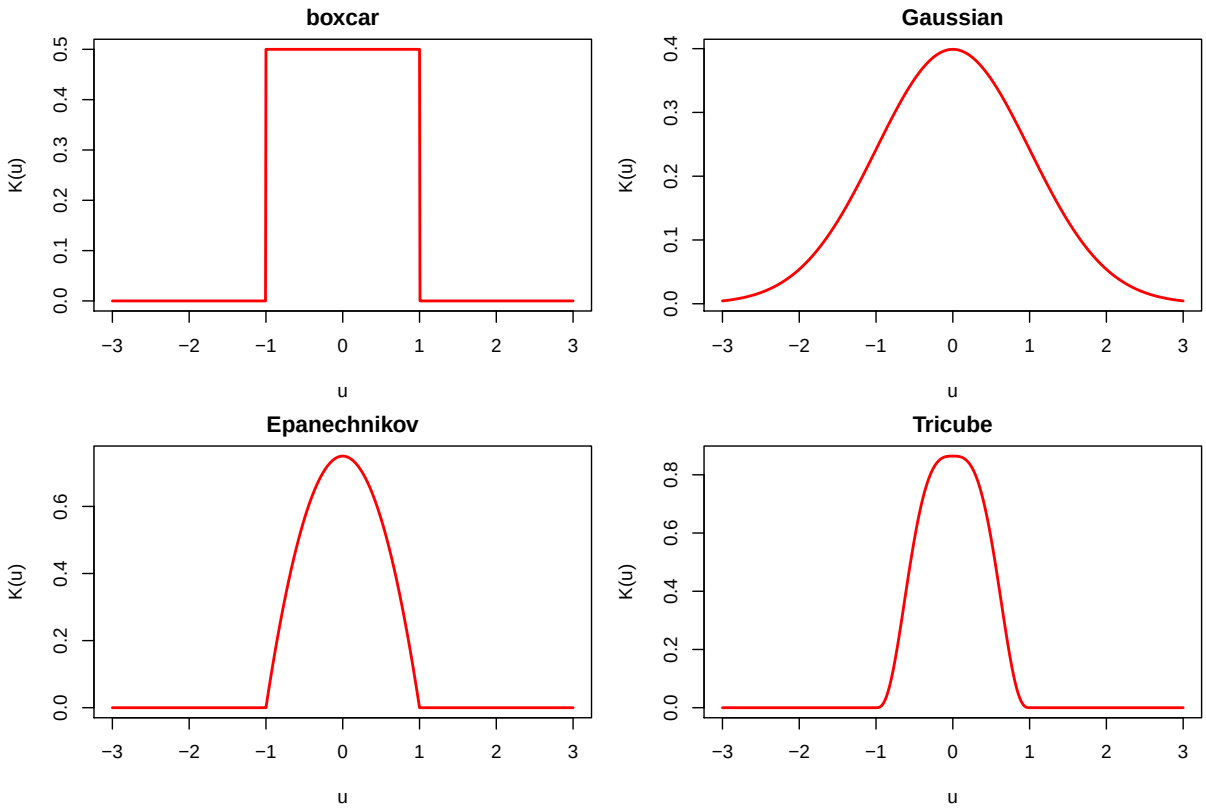
Στην προηγούμενη ενότητα ορίσαμε τον απλοϊκό εκτιμητή της συνάρτησης πυκνότητας πιθανότητας μέσω της σχέσης (3.29), όπου το $K(u)$ είναι η συνάρτηση πυκνότητας πιθανότητας της ομοιόμορφης κατανομής. Από το Σχήμα 3.6 είναι σαφές ότι ο απλοϊκός εκτιμητής έχει το μειονέκτημα ότι παρουσιάζει απότομες μεταβολές (πηδηματάκια) λόγω του ότι βασίζεται στον ομοιόμορφο πυρήνα, δηλαδή στη συνάρτηση πυκνότητας πιθανότητας της ομοιόμορφης κατανομής, ο οποίος μεταβαίνει απότομα από το μηδέν σε θετικές τιμές. Αν, λοιπόν, διαλέξουμε μια άλλη μορφή για τον πυρήνα $K(u)$, μπορούμε να πάρουμε πιο ομαλές εκτιμήσεις.

Γενικά, ως πυρήνας μπορεί να οριστεί οποιαδήποτε συνάρτηση $K(u) : \mathbb{R} \to \mathbb{R}$ που ικανοποιεί τη συνθήκη $\int_{-\infty}^{\infty} K(u)\mathrm{d}u = 1$ (Silverman, 1986; Hansen, 2009). Στο παρόν βιβλίο θα χρησιμοποιηθεί ο εξής ορισμός (βλ., π.χ. Καρλής, 2004; Wasserman, 2006).

Ορισμός 3.6

Μία πραγματική συνάρτηση $K(u)$ ονομάζεται **πυρήνας** (kernel), αν πληροί τις ιδιότητες

1. $K(u) \geqslant 0$, για κάθε $u \in \mathbb{R}$.
2. $\int_{-\infty}^{\infty} K(u)\mathrm{d}u = 1.$
3. $\int_{-\infty}^{\infty} uK(u)\mathrm{d}u = 0.$
4. $0 < \sigma_K^2 = \int_{-\infty}^{\infty} u^2 K(u)\mathrm{d}u < \infty.$



Σχήμα 3.7: Γραφικές παραστάσεις διάφορων πυρήνων.

Παρατήρηση 3.2. Από τον Ορισμό 3.6 προκύπτει ότι ένας πυρήνας είναι μία συνάρτηση πυκνότητας πιθανότητας η οποία έχει μέση τιμή 0 (ιδιότητα 3) και πεπερασμένη διασπορά σ_K^2 (ιδιότητα 4).

Παρακάτω, δίνονται κάποιες από τις πιο συχνά χρησιμοποιούμενες επιλογές πυρήνων $K(u)$, ενώ στο Σχήμα 3.7 απεικονίζονται οι γραφικές παραστάσεις για κάποιες από αυτές τις επιλογές.

$$\textbf{boxcar kernel}: \quad K(u) = \frac{1}{2} I\{|u| \leqslant 1\}$$

$$\textbf{Gaussian kernel}: \quad K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} I\{u \in \mathbb{R}\}$$

$$\textbf{Epanechnikov kernel}: \quad K(u) = \frac{3}{4}\left(1 - u^2\right) I\{|u| \leqslant 1\}$$

$$\textbf{tricube kernel}: \quad K(u) = \frac{70}{81}\left(1 - |u|^3\right)^3 I\{|u| \leqslant 1\}.$$

$$\textbf{biweight kernel}: \quad K(u) = \frac{15}{16}\left(1 - |u|^2\right)^2 I\{|u| \leqslant 1\}.$$

Ένα κοινό χαρακτηρηστικό όλων των παραπάνω επιλογών είναι ότι πρόκειται για συμμετρικούς πυρήνες, δηλαδή ισχύει ότι $K(u) = K(-u)$, για κάθε $u \in \mathbb{R}$. Από την άλλη, υπάρχουν αρκετές διαφορές μεταξύ αυτών, οι οποίες κυρίως έχουν να κάνουν με την κύρτωση. Ο κανονικός πυρήνας είναι ο μόνος που δίνει θετικό βάρος σε τιμές εκτός του διαστήματος $(-1,1)$. Παρατηρήστε ότι πρόκειται για τη συνάρτηση πυκνότητας πιθανότητας της τυπικής κανονικής κατανομής.

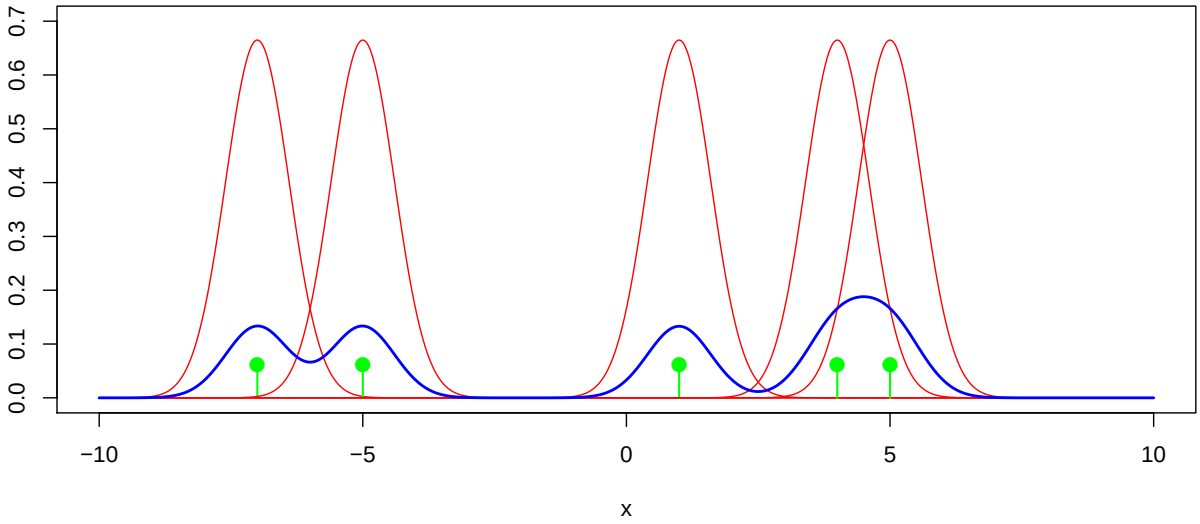
Ο επόμενος ορισμός γενικεύει την έννοια του εκτιμητή της συνάρτησης πυκνότητας πιθανότητας με βάση οποιονδήποτε πυρήνα $K(u)$ που ικανοποιεί τις συνθήκες που αναφέρθηκαν στον Ορισμό 3.6.

Ορισμός 3.7

Η εκτίμηση μιας συνάρτησης πυκνότητας πιθανότητας με τη χρήση ενός πυρήνα $K(u)$ δίνεται από τη σχέση:

$$f_{h,n}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}. \tag{3.30}$$

Η παράμετρος εξομάλυνσης $h > 0$ λέγεται εύρος παραθύρου (ή ζώνης) (window width ή bandwidth) του πυρήνα.



Σχήμα 3.8: Η μπλε γραμμή αντιστοιχεί στον εκτιμητή πυκνότητας με κανονικό πυρήνα για τα δεδομένα του Παραδείγματος 3.8. Για κάθε τιμή x, η εκτίμηση της $f(x)$ προκύπτει ως ο μέσος όρος των 5 κανονικών πυκνοτήτων, οι οποίες απεικονίζονται με κόκκινο χρώμα.

Παράδειγμα 3.8. Να υπολογιστεί ο εκτιμητής με χρήση κανονικού πυρήνα για δείγμα μεγέθους $n = 5$ με τιμές

$$-7, -5, 1, 4, 5$$

θεωρώντας ότι το παράθυρο ισούται με $h = 0.6$.

Λύση Παραδείγματος 3.8. Από τον Ορισμό 3.7 έχουμε για $n = 5$, $h = 0.6$ ότι

$$\begin{aligned} f_{h,n}(x) &= \frac{1}{5\times 0.6}\sum_{i=1}^{5}\frac{1}{\sqrt{2\pi}}e^{-\{(x-x_i)/0.6\}^2/2} \\ &= \frac{1}{3\sqrt{2\pi}}\sum_{i=1}^{5} e^{-25(x-x_i)^2/3}, \quad x \in \mathbb{R}, \end{aligned}$$

με $x_1, \ldots, x_5$, να είναι τα παρατηρηθέντα δεδομένα. Ο παραπάνω εκτιμητής της συνάρτησης πυκνότητας πιθανότητας με χρήση πυρήνα απεικονίζεται στο Σχήμα 3.8, μαζί με τα παρατηρηθέντα δεδομένα (πράσινες κατακόρυφες ράβδοι). Παρατηρήστε ότι για κάθε σημείο x, η εκτίμηση $f_{h,n}(x)$ είναι ο μέσος όρος των πυρήνων με κεντρικό σημείο τα παρατηρηθέντα x_i και αυτό, τελικά, είναι μία ισοβαρής μείξη των κατανομών $\mathcal{N}(x_i, 0.6)$, $i = 1, \ldots, 5$. □

3.4.3 Ιδιότητες εκτιμητή με χρήση πυρήνα

Από τη σχέση (3.30) προκύπτει ότι ο εκτιμητής με χρήση πυρήνα είναι όντως συνάρτηση πυκνότητας πιθανότητας. Προφανώς, $f_{h,n}(x) \geqslant 0$, για κάθε $x \in \mathbb{R}$. Επιπρόσθετα,

$$\begin{aligned}\int_{-\infty}^{\infty} f_{h,n}(x)\mathrm{d}x &= \int_{-\infty}^{\infty} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\mathrm{d}x \\ &= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\left(\frac{x-X_i}{h}\right)\mathrm{d}x \quad \left(\text{θέτουμε} \quad u=\frac{x-X_i}{h}\right) \\ &= \frac{1}{nh}\sum_{i=1}^{n} h\int_{-\infty}^{\infty} K(u)\mathrm{d}u \quad \text{(από ιδιότητα 2 Ορισμού 3.6)} \\ &= \frac{n}{n} = 1.\end{aligned}$$

Είναι εύκολο να δειχτεί ότι η μέση τιμή και η διασπορά αυτής της κατανομής είναι ίσες με $\bar{X}$ και $S^2 + h^2\sigma_K^2$, αντίστοιχα (Άσκηση 3.22). Εφιστούμε την προσοχή στη διάκριση της μέσης τιμής και διασποράς της **κατανομής** $f_{h,n}(\cdot)$ με τη μέση τιμή και διασπορά του **εκτιμητή** $f_{h,n}(x)$! Οι τελευταίες μας απασχολούν στη συνέχεια.

Λήμμα 3.1. Η μέση τιμή του εκτιμητή (3.30) με χρήση πυρήνα $K(u)$ ισούται με

$$\mathsf{E}\left(\{f_{h,n}(x)\}\right) = f(x) + \frac{1}{2}f''(x)h^2\sigma_K^2 + O\left(h^3\right). \tag{3.31}$$

Απόδειξη Λήμματος 3.1. Έχουμε ότι

$$\begin{aligned}\mathsf{E}\left(\{f_{h,n}(x)\}\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}\left(\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right) = \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right)f(y)\mathrm{d}y \\ &= \int_{-\infty}^{\infty} K(u)f(x-hu)\mathrm{d}u \end{aligned} \tag{3.32}$$

όπου για τη μετάβαση στην τελευταία ισότητα θέσαμε $u = \frac{x-y}{h}$. Επειδή η f δεν είναι γνωστή, προσεγγίζουμε το παραπάνω ολοκλήρωμα με το ανάπτυγμα Taylor γύρω από το x για «μικρές» τιμές του h (δηλαδή θεωρώντας ότι $h \to 0$). Είναι

$$\int_{-\infty}^{\infty} K(u)f(x+hu)\mathrm{d}u = \int_{-\infty}^{\infty} K(u)\left\{f(x) - f'(x)hu + \frac{1}{2}f''(x)h^2u^2 + O(h^3)\right\}\mathrm{d}u.$$

Αντικαθιστώντας την τελευταία έκφραση στη σχέση (3.32) προκύπτει ότι:

$$\begin{aligned}\mathsf{E}\left(f_{h,n}(x)\right) &= f(x)\int_{-\infty}^{\infty} K(u)\mathrm{d}u - f'(x)h\int_{-\infty}^{\infty} uK(u)\mathrm{d}u + \frac{1}{2}f''(x)h^2\int_{-\infty}^{\infty} u^2K(u)\mathrm{d}u + O(h^3) \\ &= f(x) + \frac{1}{2}f''(x)h^2\sigma_K^2 + O(h^3),\end{aligned}$$

όπου για τη μετάβαση στην τελευταία ισότητα λάβαμε υπόψη τις ιδιότητες 2, 3 και 4 του Ορισμού 3.6 για καθένα από τα τρία ολοκληρώματα που εμφανίστηκαν στην προηγούμενη γραμμή. Τούτο ολοκληρώνει την απόδειξη της σχέσης (3.31). □

Το επόμενο θεώρημα περιγράφει τη μεροληψία, τη διασπορά και το ολοκληρωμένο μέσο τετραγωνικό σφάλμα ενός εκτιμητή πυκνότητας με χρήση πυρήνα.

Θεώρημα 3.3

Έστω ότι η f είναι τρεις φορές παραγωγίσιμη, με $f^{(3)}(x) < M < \infty$ και ότι K είναι συμμετρικός πυρήνας με διασπορά $\sigma_K^2 = \int_{-\infty}^{\infty} u^2 K(u)\mathrm{d}u < \infty$. Τότε για τον εκτιμητή $f_{h,n}(x)$ με χρήση πυρήνα, που δόθηκε στη σχέση (3.30), ισχύει ότι:

$$\mathrm{bias}\{f_{h,n}(x), f(x)\} = \frac{1}{2}h^2 f''(x)\sigma_K^2 + O(h^4), \tag{3.33}$$

$$\mathrm{Var}\left(f_{h,n}(x)\right) = \frac{f(x)\int_{-\infty}^{\infty} K^2(u)\mathrm{d}u}{nh} + O\left(\frac{1}{n}\right), \tag{3.34}$$

$$\mathrm{IMSE}(f_{h,n}, f) = \frac{1}{4}h^4 \int_{-\infty}^{\infty} \left\{f''(x)\right\}^2 \mathrm{d}x\sigma_K^4 + O(h^6) + \frac{\int_{-\infty}^{\infty} K^2(x)\mathrm{d}x}{nh} + O\left(\frac{1}{n}\right). \tag{3.35}$$

Απόδειξη Θεωρήματος 3.3. Αντικαθιστώντας τη σχέση (3.31) στον ορισμό της μεροληψίας $\mathrm{bias}\{f_{h,n}(x) - f(x)\}$ προκύπτει άμεσα η σχέση (3.33). Για τον υπολογισμό της διασποράς σημειώστε ότι τα $K(\frac{x-X_i}{h})$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές για $i = 1, \ldots, n$. Άρα, έχουμε ότι:

$$\begin{aligned}
\mathrm{Var}\left(f_{h,n}(x)\right) &= \frac{1}{nh^2}\mathrm{Var}\left\{K\left(\frac{x-X_1}{h}\right)\right\} \\
&= \frac{1}{nh^2}\mathrm{E}\left\{K\left(\frac{x-X_1}{h}\right)^2\right\} - \frac{1}{n}\left\{\frac{1}{h}\mathrm{E}\left\{K\left(\frac{x-X_1}{h}\right)\right\}\right\}^2 \\
&= \frac{1}{nh^2}\mathrm{E}\left\{K\left(\frac{x-X_1}{h}\right)^2\right\} - \frac{1}{n}\left\{\mathrm{E}\left(f_{h,n}(x)\right)\right\}^2 \\
&\overset{(3.31)}{=} \frac{1}{nh^2}\mathrm{E}\left\{K\left(\frac{x-X_1}{h}\right)^2\right\} - \frac{1}{n}\left\{f(x) + \frac{1}{2}f''(x)h^2\sigma_K^2 + O(h^3)\right\}^2 \\
&= \frac{1}{nh^2}\mathrm{E}\left\{K\left(\frac{x-X_1}{h}\right)^2\right\} + O\left(\frac{1}{n}\right),
\end{aligned} \tag{3.36}$$

επειδή

$$\lim_{h\to 0}\left(\frac{1}{2}f''(x)h^2\sigma_K^2 + O(h^3)\right) = 0.$$

Για τον πρώτο όρο της σχέσης (3.36) θεωρούμε το ανάπτυγμα Taylor και έχουμε:

$$\begin{aligned}
\frac{1}{nh^2}\mathrm{E}\left\{K\left(\frac{x-X_1}{h}\right)^2\right\} &= \frac{1}{nh^2}\int_{-\infty}^{\infty} K^2\left(\frac{x-v}{h}\right) f(v)\mathrm{d}v \\
&= \frac{1}{nh}\int_{-\infty}^{\infty} K^2(u) f(x-hu)\mathrm{d}u \\
&= \frac{1}{nh}\int_{-\infty}^{\infty} K^2(u)\{f(x) + O(h)\}\mathrm{d}u \\
&= \frac{1}{nh}\int_{-\infty}^{\infty} K^2(u) f(x)\mathrm{d}u + \frac{1}{nh}O(h) \\
&= \frac{1}{nh}\int_{-\infty}^{\infty} K^2(u) f(x)\mathrm{d}u + O\left(\frac{1}{n}\right).
\end{aligned}$$

Αντικαθιστώντας την τελευταία έκφραση στη σχέση (3.36), προκύπτει η σχέση (3.34).

Αντικαθιστώντας τις (3.33) και (3.34) στην (3.7), προκύπτει ότι το μέσο τετραγωνικό σφάλμα ισούται με

$$\begin{aligned}\text{MSE}\{f_{h,n}(x), f(x)\} &= \left\{\frac{1}{2}h^2 f''(x)\sigma_K^2 + O(h^4)\right\}^2 + \frac{f(x)\int_{-\infty}^{\infty} K^2(u)\mathrm{d}u}{nh} + O\left(\frac{1}{n}\right) \\ &= \frac{1}{4}h^4\{f''(x)\}^2\sigma_K^4 + O(h^8) + O(h^4)h^2 f''(x)\sigma_K^2 + \frac{f(x)\int_{-\infty}^{\infty} K^2(u)\mathrm{d}u}{nh} + O\left(\frac{1}{n}\right) \\ &= \frac{1}{4}h^4\left\{f''(x)\right\}^2\sigma_K^4 + \frac{f(x)\int_{-\infty}^{\infty} K^2(u)\mathrm{d}u}{nh} + O(h^6) + O\left(\frac{1}{n}\right).\end{aligned}$$

Ολοκληρώνοντας την τελευταία έκφραση του μέσου τετραγωνικού σφάλματος ως προς x, προκύπτει η σχέση (3.35). □

Από το Θεώρημα 3.3 έπεται ότι, προσεγγιστικά, η μεροληψία της $f_{h,n}$ δεν εξαρτάται από το n (παρατηρήστε ότι το μέγεθος δείγματος δεν υπάρχει πουθενά στο ανάπτυγμα της μεροληψίας). Αυτό σημαίνει ότι ο εκτιμητής με χρήση πυρήνα μπορεί να είναι μεροληπτικός ακόμα και για πολύ μεγάλα μεγέθη δείγματος. Επιπρόσθετα, καθώς αυξάνεται το h, μεγαλώνει η μεροληψία, και, καθώς αυξάνεται το nh, μειώνεται η διασπορά του εκτιμητή. Άρα, ο εκτιμητής πυκνότητας με χρήση πυρήνα είναι ασυμπτωτικά συνεπής εκτιμητής της $f(x)$, αν ικανοποιούνται οι συνθήκες του θεωρήματος που ακολουθεί.

Θεώρημα 3.4

Έστω ότι $f(x)$ συνεχής στο x και $|f'(x)| < M$. Τότε για $h \to 0$ και $nh \to \infty$, καθώς $n \to \infty$, ισχύει ότι:

$$f_{h,n}(x) \xrightarrow{P} f(x).$$

Απόδειξη Θεωρήματος 3.4. Η απόδειξη ακολουθεί παρόμοια επιχειρηματολογία με αυτήν του Θεωρήματος 3.2 και για αυτό παραλείπεται. □

Παραγωγίζοντας τη σχέση (3.35) βρίσκουμε ότι η τιμή της παραμέτρου εξομάλυνσης που ασυμπτωτικά ελαχιστοποιεί το ολοκληρωμένο μέσο τετραγωνικό σφάλμα ισούται με

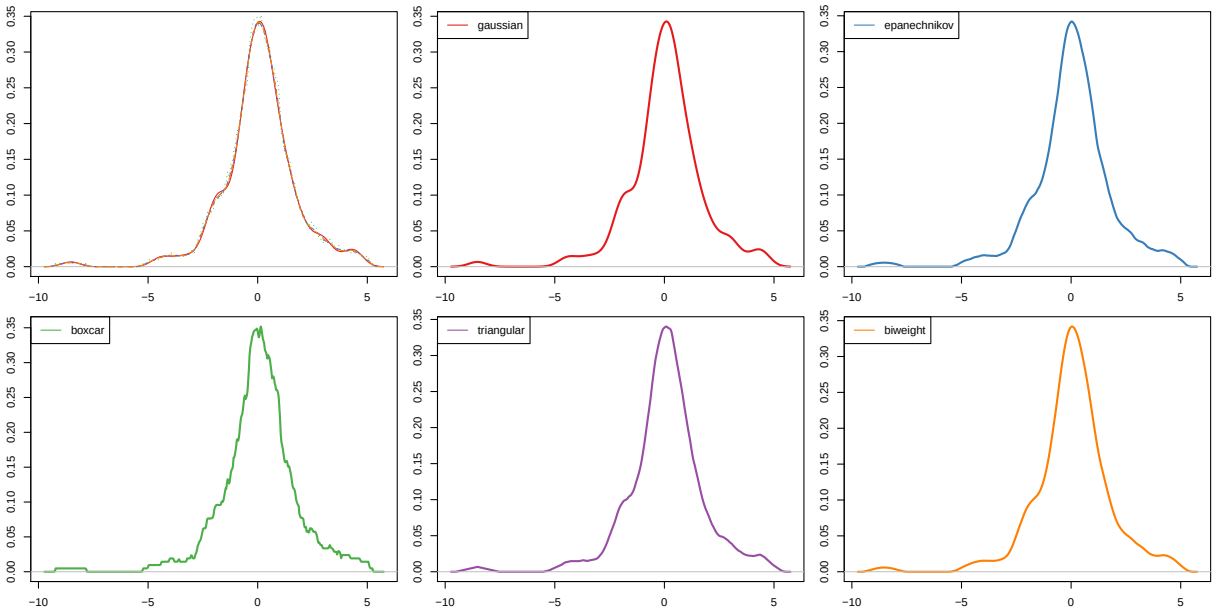
$$h^\star = \left(\frac{c_2}{c_1^2 A(f) n}\right)^{1/5}, \tag{3.37}$$

όπου $c_1 = \int_{-\infty}^{\infty} x^2 K(x)\mathrm{d}x$, $c_2 = \int_{-\infty}^{\infty} K(x)^2\mathrm{d}x$ και $A(f) = \int_{-\infty}^{\infty}\left\{f''(x)\right\}^2\mathrm{d}x$. Με αυτήν την επιλογή της παραμέτρου εξομάλυνσης έχουμε ότι:

$$\text{IMSE}\{f_{h^\star,n}, f\} = O(n^{-4/5}). \tag{3.38}$$

Συγκρίνοντας τα ασυμπτωτικά βέλτιστα ολοκληρωμένα μέσα τετραγωνικά σφάλματα των σχέσεων (3.23) και (3.38) συμπεραίνουμε ότι οι **εκτιμητές πυκνότητας με χρήση πυρήνα συγκλίνουν ταχύτερα από τα ιστογράμματα**, συνεπώς θα πρέπει να προτιμώνται. Περαιτέρω, μπορεί να αποδειχθεί ότι δεν υπάρχει μη παραμετρικός εκτιμητής πυκνότητας που να συγκλίνει με γρηγορότερο ρυθμό από $O(n^{-4/5})$ (βλ., μεταξύ άλλων, Van der Vaart, 2000, Κεφάλαιο 24).

Εκτός από την επιλογή του ασυμπτωτικά βέλτιστου h, μπορούμε να επιλέξουμε και τον βέλτιστο πυρήνα $K(u)$. Σε αυτήν την περίπτωση μπορεί να δειχτεί ότι ο πυρήνας που ασυμπτωτικά ελαχιστοποιεί το ολοκληρωμένο μέσο τετραγωνικό σφάλμα είναι ο Epanechnikov (Muller, 1984). Ωστόσο, όπως φαίνεται και στο Σχήμα 3.9, στην πράξη η επιλογή του πυρήνα δεν παίζει τόσο μεγάλο ρόλο όσο η κατάλληλη επιλογή του h.



Σχήμα 3.9: Εκτιμητές πυκνότητας με χρήση διαφορετικών πυρήνων για τα δεδομένα του Παραδείγματος 3.7. Το πρώτο διάγραμμα περιέχει και τους πέντε εκτιμητές μαζί.

3.4.4 Επιλογή του h για τον εκτιμητή με χρήση πυρήνα

Στην ενότητα αυτή θα παρουσιαστούν τρεις δημοφιλείς τεχνικές για την επιλογή του εύρους παραθύρου h για τον εκτιμητή με χρήση πυρήνα. Συγκεκριμένα, θα αναφερθούμε στον κανονικό κανόνα, στην τεχνική cross-validation και στην τεχνική cross-validated πιθανοφάνειας.

3.4.4.1 Κανονικός κανόνας

Όπως, ίσως, γίνεται αντιληπτό από το όνομα της τεχνικής αυτής, η ιδέα είναι να υποθέσουμε ότι η (άγνωστη) συνάρτηση πυκνότητας πιθανότητας f είναι κάποιο μέλος της κανονικής οικογένειας κατανομών. Στην πραγματικότητα, όμως, αυτό δεν το γνωρίζουμε, και, γενικά, ο κανονικός κανόνας θα πρέπει να χρησιμοποιείται μόνο όταν έχουμε σοβαρούς λόγους να πιστεύουμε ότι η πραγματική f είναι αρκετά λεία και μοιάζει με την κανονική κατανομή.

Έστω ότι χρησιμοποιείται ένας κανονικός πυρήνας και ότι η f είναι η πυκνότητα της $\mathcal{N}(\mu, \sigma^2)$, για κάποια $\mu \in \mathbb{R}$ και $\sigma > 0$. Εδώ, να παρατηρήσουμε ότι αυτό που μας ενδιαφέρει στη σχέση (3.37) δεν είναι η ίδια η $f(x)$, αλλά η $f''(x)$. Θυμηθείτε ότι η δεύτερη παράγωγος μιας συνάρτησης καθορίζει αν αυτή έχει μία ή περισσότερες κορυφές, πόσο ομαλή είναι κ.λπ. Για την περίπτωση της πυκνότητας της $\mathcal{N}(\mu, \sigma^2)$ είναι εύκολο να δούμε ότι:

$$f''(x) = f(x)\left\{\frac{(x-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2}\right\}.$$

Από τη σχέση (3.37) υπολογίζουμε ότι το βέλτιστο h είναι ίσο με (βλ., π.χ. Scott, 2012)

$$h^\star \approx 1.06 \sigma n^{-1/5}.$$

Στην περίπτωση που δεν χρησιμοποιείται ο κανονικός πυρήνας, θα πρέπει να χρησιμοποιηθούν κατάλληλες πολλαπλασιαστικές σταθερές (βλ., π.χ. Καρλής, 2004, Πίνακας 1.3). Συνήθως, η πληθυσμιακή τυπική απόκλιση εκτιμάται από την ποσότητα

$$\min\{s, \text{IQR}/1.34\},$$

όπου s είναι η δειγματική τυπική απόκλιση και $\text{IQR} = \hat{x}_{0.25} - \hat{x}_{0.75}$ το δειγματικό ενδοτεταρτημοριακό εύρος (sample inter-quartile range), όπου τα $\hat{x}_{0.25}$ και $\hat{x}_{0.75}$ είναι αντίστοιχα τα δειγματικά 3ο και 1ο τεταρτημόριο.

3.4.4.2 Cross-validation εκτίμηση του IMSE

Πιο γενικά, για την επιλογή του εύρους παραθύρου h χρησιμοποιούμε το ολοκληρωμένο μέσο τετραγωνικό σφάλμα της σχέσης (3.8). Είδαμε ότι στην περίπτωση του εκτιμητή με χρήση πυρήνα το IMSE δίνεται από τη σχέση (3.35). Παραλείποντας ό,τι δεν εξαρτάται από h ελαχιστοποιούμε τη σχέση (3.9), ήτοι την ποσότητα:

$$J(h) := \mathsf{E}\left[\int_{-\infty}^{\infty} f_{h,n}^2(x)\mathrm{d}x - 2\int_{-\infty}^{\infty} f_{h,n}(x)f(x)\mathrm{d}x\right].$$

Δυστυχώς, όπως και στο ιστόγραμμα, το εμπόδιο είναι ότι η f δεν είναι γνωστή, οπότε το ίδιο ισχύει και για την $J(h)$. Ένας αμερόληπτος εκτιμητής της $J(h)$ μέσω cross-validation είναι ο (βλ. τη σχέση (3.11))

$$\begin{aligned}\hat{J}(h) &= \int_{-\infty}^{\infty} f_{h,n}^2(x)\mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} f_{h,n-1}^{(-i)}(x_i) \\ &\approx \frac{1}{hn^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{\star}\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh}K(0),\end{aligned}$$

όπου $K^{\star}(x) = K^{(2)}(x) - 2K(x)$ και $K^{(2)}(z) = \int_{-\infty}^{\infty} K(z-y)K(y)\mathrm{d}y$. Όταν K είναι ο κανονικός πυρήνας, τότε η $K^{(2)}(z)$ είναι η συνάρτηση πυκνότητας πιθανότητας της $\mathcal{N}(0, 2)$. Η ελαχιστοποίηση του IMSE μέσω της $\hat{J}(h)$ στην παραπάνω εξίσωση είναι απαιτητική, αλλά υπάρχουν τεχνικές, όπως η μέθοδος μέσω του μετασχηματισμού Fourier, για γρήγορο υπολογισμό της. Για σχετικές πληροφορίες δείτε, για παράδειγμα, Silverman (1986), σελ. 61-66.

3.4.4.3 Cross-validated πιθανοφάνεια

Μία εναλλακτική τεχνική για την επιλογή του h είναι να χρησιμοποιήσουμε έναν συνδυασμό της τεχνικής cross-validation με τη μέθοδο μέγιστης πιθανοφάνειας, η οποία είναι γνωστή με το όνομα cross-validated πιθανοφάνεια (Habbema, 1974; Duin, 1976; Van Es, 1991). Η ιδέα είναι να θεωρήσουμε πως κάθε τιμή του h αντιστοιχεί σε ένα μοντέλο που περιγράφει τα δεδομένα. Για δοθέν h, η εκτιμηθείσα πιθανοφάνεια είναι

$$L(h) = \prod_{i=1}^{n} f_{h,n}(x_i).$$

Φυσικά, η παραπάνω πιθανοφάνεια μεγιστοποιείται για $h = 0$, οπότε με βάση αυτήν δεν μπορούμε να επιλέξουμε το h.

Στην Ενότητα 3.2.2 αναφερθήκαμε στην τεχνική του cross-validation για να αξιολογήσουμε πόσο καλό είναι ένα μοντέλο που περιγράφει τα δεδομένα. Πιο συγκεκριμένα στην τεχνική του leave-one-out cross-validation είδαμε ότι μπορούμε να αφήσουμε μία παρατήρηση έξω, κατόπιν να προσαρμόσουμε το μοντέλο στις υπόλοιπες παρατηρήσεις και, στη συνέχεια, να αξιολογήσουμε πόσο καλά δουλεύει το μοντέλο για την παρατήρηση που έμεινε εκτός. Αν επαναλάβουμε την ίδια διαδικασία αφήνοντας έξω μια διαφορετική παρατήρηση, μπορούμε να λάβουμε μία συνολική εκτίμηση για το πόσο καλό είναι το μοντέλο, και, επομένως, να διαλέξουμε ανάμεσα σε διαφορετικά μοντέλα.

Στην περίπτωση της εκτίμησης της συνάρτησης πυκνότητας πιθανότητας με χρήση πυρήνα, τα διαφορετικά μοντέλα είναι οι εκτιμήσεις $f_{h,n}$ που προκύπτουν για διαφορετικές τιμές του εύρους παραθύρου h. Σε αυτό το σημείο δίνουμε τον ακόλουθο ορισμό.

Ορισμός 3.8

Η cross-validated πιθανοφάνεια ορίζεται ως

$$L^{\text{CV}}(h) := \prod_{i=1}^{n} f_{h,n-1}^{(-i)}(x_i), \tag{3.39}$$

όπου $f_{h,n-1}^{(-i)}(x)$ είναι η εκτίμηση της συνάρτησης πυκνότητας πιθανότητας στο x, αν αφαιρέσουμε την i-οστή παρατήρηση, $i = 1, \dots, n$.

Η τιμή του h που μεγιστοποιεί την (3.39) είναι η επιλογή μας για το bandwidth βάσει της cross-validated πιθανοφάνειας. Στην πράξη η μεγιστοποίηση γίνεται δίνοντας ένα σύνολο εύλογων (δυνατών) τιμών για το h και υπολογίζοντας την (3.39) (ή τον λογάριθμο αυτής) για καθεμία τιμή. Το βέλτιστο h αντιστοιχεί στο μέγιστο μεταξύ όλων των τιμών.

3.4.5 Η συνάρτηση `density()` στην `R`

Στην παρούσα ενότητα θα δούμε πώς μπορούμε να εκτιμήσουμε τη συνάρτηση πυκνότητας πιθανότητας με τη χρήση της `R`. Συγκεκριμένα, αυτό μπορεί να γίνει μέσω της εντολής

```
density(x, bw = ..., kernel = ...)
```

όπου

- `x`: τα δεδομένα,
- `bw`: παράμετρος εξομάλυνσης, η οποία μπορεί να είναι
  - ★ είτε κάποια θετική τιμή
  - ★ είτε κάποιος χαρακτήρας (`"nrd"`, `"ucv"`, `"bcv"`, ...) που καθορίζει τον τρόπο επιλογής του h βάσει διαθέσιμων κριτηρίων,
- `kernel`: ο τύπος του πυρήνα με δυνατές επιλογές (μεταξύ άλλων) `"gaussian"`, `"epanechnikov"`, `"rectangular"`, `"triangular"`,`"biweight"`. Σημειώνεται ότι η επιλογή `"rectangular"` αντιστοιχεί στον πυρήνα boxcar.

Όλοι οι πυρήνες στην εντολή `density` είναι παραμετροποιημένοι, έτσι ώστε το `bw` να αντιστοιχεί στην τυπική απόκλιση του πυρήνα. Έτσι, η παράμετρος αυτή δεν παριστάνει το εύρος παραθύρου h του εκτιμητή, όπως περιγράψαμε παραπάνω. Σε κάθε περίπτωση όμως υπάρχει μία 1-1 αντιστοιχία.

Για την επιλογή του h στην `R` με τη χρήση του κανονικού κανόνα (βλ. Ενότητα 3.4.4.1) χρησιμοποιούμε την εντολή

```
density(x, bw = "bw.nrd", kernel = ...)
```

Σημειώστε εδώ ότι η προεπιλεγμένη τεχνική στο όρισμα `bw` είναι η `"bw.nrd0"`, η οποία είναι παρόμοια με αυτήν του κανονικού κανόνα, αλλά με τη διαφορά ότι η πολλαπλασιαστική σταθερά είναι ίση με 0.9. Σε αυτήν την περίπτωση, λέμε ότι προκύπτει ο κανόνας του Silverman (βλ. τη σχέση (3.31) του Silverman, 1986), αντί του πιο διαδεδομένου κανόνα με τη σταθερά 1.06. Ωστόσο, δεν θα πρέπει να θεωρείται ως η πιο κατάλληλη

επιλογή, όπως σημειώνεται και στην περιγραφή της εντολής `density`. Η προεπιλεγμένη τεχνική για το όρισμα `kernel` είναι ο κανονικός πυρήνας.

Για την επιλογή του h στην `R` με ελαχιστοποίηση του IMSE μέσω cross-validation (βλ. Ενότητα 3.4.4.2) χρησιμοποιούμε την εντολή

```
density(x, bw = "ucv", kernel = ...)
```

Η τιμή `'ucv'` στο όρισμα `bw` αντιστοιχεί στη χρήση της μεθόδου cross-validation[4]. Η επιλογή της παραμέτρου εξομάλυνσης μέσω της μεθόδου cross-validated πιθανοφάνειας (που συζητήσαμε στην 3.4.4.3) δεν είναι διαθέσιμη μέσω της εντολής `density`.

Εξ ορισμού, η εκτιμηθείσα πυκνότητα θα υπολογιστεί σε $M = 512$ σημεία τα οποία ισαπέχουν και καλύπτουν το εύρος του παρατηρηθέντος δείγματος. Οι τιμές αυτές επιστρέφονται μέσω της εντολής `f$x`. Οι αντίστοιχες εκτιμήσεις της συνάρτησης πυκνότητας πιθανότητας επιστρέφονται μέσω της εντολής `f$y`. Το πλήθος αυτών των σημείων μπορεί να αλλάξει μέσω του ορίσματος `n`.

Για παράδειγμα, αν επιθυμούμε να υπολογίσουμε την εκτίμηση με τον απλοϊκό πυρήνα σε 1024 σημεία, με χρήση cross-validation για την ελαχιστοποίηση του IMSE, τότε η εντολή συντάσσεται ως

```
f <- density(x, bw = "ucv", kernel = "rectangular", n = 1024)
```

Η συνάρτηση έχει μέθοδο `plot()` για τη γραφική αναπαράστασή της: `plot(f)`.

Παράδειγμα 3.9. (συνέχεια Παραδείγματος 3.5) Θεωρήστε τα δεδομένα του Παραδείγματος 3.5. Να αναπαρασταθεί γραφικά η εκτίμηση της συνάρτησης πυκνότητας πιθανότητας με χρήση διαφορετικών πυρήνων. Η παράμετρος εξομάλυνσης να επιλεχθεί ελαχιστοποιώντας το ολοκληρωμένο μέσο τετραγωνικό σφάλμα, χρησιμοποιώντας την τεχνική cross-validation για την εκτίμησή του.

Λύση Παραδείγματος 3.9. Στον κώδικα που ακολουθεί θα χρησιμοποιήσουμε τα δεδομένα που προσομοιώθηκαν στο Παράδειγμα 3.5 και έχουν αποθηκευτεί στο διάνυσμα `x` το οποίο περιέχει τις 1000 τιμές του δείγματος. Εφόσον ζητείται να εκτιμηθεί το ολοκληρωμένο μέσο τετραγωνικό σφάλμα μέσω cross-validation, επιλέγουμε `bw = "ucv"`. Στο Σχήμα 3.10 παρατίθενται οι γραφικές παραστάσεις 5 διαφορετικών εκτιμήσεων της συνάρτησης πυκνότητας πιθανότητας των δεδομένων. Παρατηρούμε ότι οι διαφορές μεταξύ διαφορετικών πυρήνων είναι αμελητέες.

```
1 library(RColorBrewer)
2 myCol <- brewer.pal(6, "Set1")
3 par(mfrow=c(2,3),mar=c(2,2,1,1))
4 plot(density(x, bw="ucv",kernel="gaussian"),col=myCol[1],lwd=2,main=""
    ,xlab = "",ylab = "")
5 points(density(x, bw="ucv",kernel="epanechnikov"),col=myCol[2],type="l
    ",lwd=2,lty=2)
6 points(density(x, bw="ucv",kernel="rectangular"),col=myCol[3],type="l"
    ,lwd=2,lty=3)
7 points(density(x, bw="ucv",kernel="triangular"),col=myCol[4],type="l",
    lwd=2,lty=4)
8 points(density(x, bw="ucv",kernel="biweight"),col=myCol[5],type="l",
    lwd=2,lty=5)
9 legend('topright', c("gaussian", "epanechnikov", "boxcar", "triangular
    ", "biweight"), col =myCol[1:5], lty=1:5, lwd=2)
```

[4] Συγκεκριμένα, αντιστοιχεί στη μέθοδο unbiased cross-validation, ενώ η τιμή `'bcv'` αντιστοιχεί στη μέθοδο biased cross-validation.

```
plot(density(x, bw="ucv",kernel="gaussian"),col=myCol[1],lwd=2,main=""
   ,xlab = "",ylab = "")
legend('topright',"gaussian",lty=1,col=myCol[1])
plot(density(x, bw="ucv",kernel="epanechnikov"),col=myCol[2],lwd=2,
   main="",xlab = "",ylab = "")
legend('topright',"epanechnikov",lty=1,col=myCol[2])
plot(density(x, bw="ucv",kernel="rectangular"),col=myCol[3],lwd=2,main
   ="",xlab = "",ylab = "")
legend('topright','boxcar',lty=1,col=myCol[3])
plot(density(x, bw="ucv",kernel="triangular"),col=myCol[4],lwd=2,main=
   "",xlab = "",ylab = "")
legend('topright',"triangular",lty=1,col=myCol[4])
plot(density(x, bw="ucv",kernel="biweight"),col=myCol[5],lwd=2,main=""
   ,xlab = "",ylab = "")
legend('topright',"biweight",lty=1,col=myCol[5])
```

□

Παράδειγμα 3.10. (συνέχεια Παραδείγματος 3.5) Θεωρήστε τα δεδομένα του Παραδείγματος 3.5. Να βρείτε την τιμή της παραμέτρου εξομάλυνσης η οποία ελαχιστοποιεί την cross-validated πιθανοφάνεια της σχέσης (3.39) και, κατόπιν, να αναπαραστήσετε σε ένα διάγραμμα τους τρεις εκτιμητές πυκνότητας με χρήση:

1. ιστογράμματος με παράμετρο εξομάλυνσης που ελαχιστοποιεί το IMSE,
2. κανονικού πυρήνα με παράμετρο εξομάλυνσης που ελαχιστοποιεί το IMSE, και
3. κανονικού πυρήνα με παράμετρο εξομάλυνσης που ελαχιστοποιεί την cross-validated πιθανοφάνεια.

Λύση Παραδείγματος 3.10. Θα σχολιάσουμε μόνο τα αποτελέσματα που προκύπτουν, χωρίς να περιγράψουμε τον κώδικα (βλ. Άσκηση 3.10). Αρχικά, εφαρμόζουμε την τεχνική cross-validated πιθανοφάνειας με χρήση κανονικού πυρήνα, θεωρώντας ένα σύνολο 100 τιμών της παραμέτρου εξομάλυνσης h στο διάστημα $(0, 0.5)$. Για καθεμία τιμή υπολογίζουμε την (3.39) και η γραφική παράσταση των τιμών αυτής παρατίθεται στο Σχήμα 3.11 (αριστερά). Παρατηρούμε ότι το μέγιστο της cross-validated πιθανοφάνειας επιτυγχάνεται για $h = 0.085$. Η γραφική παράσταση της εκτιμηθείσας πυκνότητας απεικονίζεται στο Σχήμα 3.11 (δεξιά).

Συγκεντρωτικά, στο Σχήμα 3.12 παρατίθεται (με κόκκινη διακεκομμένη γραμμή) η πραγματική συνάρτηση πυκνότητας πιθανότητας των δεδομένων που δόθηκε στη σχέση (3.1),

- μαζί με το ιστόγραμμα με εύρος κελιών $h = 0.104$, το οποίο είναι το βέλτιστο βάσει ελαχιστοποίησης του IMSE μέσω cross-validation (βλ. το Παράδειγμα 3.5),
- μαζί με τον εκτιμητή της συνάρτησης πυκνότητας πιθανότητας με κανονικό πυρήνα, όπου η παράμετρος εξομάλυνσης ελαχιστοποιεί το IMSE (μπλε γραμμή) το οποίο έχει εκτιμηθεί μέσω cross-validation (βλ. το Παράδειγμα 3.9), και
- μαζί με τον εκτιμητή της συνάρτησης πυκνότητας πιθανότητας με κανονικό πυρήνα και παράμετρο εξομάλυνσης που μεγιστοποιεί την cross-validated πιθανοφάνεια (πράσινη γραμμή).

Είναι προφανές ότι όλες οι τεχνικές αναδεικνύουν τις πολλαπλές κορυφές που έχει η πραγματική συνάρτηση πυκνότητας πιθανότητας που δόθηκε στη σχέση (3.1). Από την άλλη, αν το ζήτημα εξεταστεί ενδελεχώς, θα παρατηρήσετε ότι όλες οι εκτιμήσεις της συνάρτησης πυκνότητας πιθανότητας υποεκτιμούν το ύψος των κορυφών αυτών, όταν συγκρίνονται με την κόκκινη διακεκομμένη γραμμή που αντιστοιχεί στην πραγματική συνάρτηση πυκνότητας πιθανότητας. Αυτό δεν είναι τυχαίο και αποτελεί ένα μειονέκτημα των μη παραμετρικών τεχνικών εκτίμησης της συνάρτησης πυκνότητας πιθανότητας, το οποίο μειονέκτημα σχετίζεται με το γεγονός ότι ο εκτιμητής της συνάρτησης πυκνότητας πιθανότητας (είτε ιστόγραμμα είτε μέσω πυρήνα) είναι μεροληπτικός όσο μεγάλο μέγεθος δείγματος και να διαθέτουμε (θυμηθείτε τις σχέσεις

(3.18) και (3.33)). Όπως, μάλιστα, θα δούμε στο Κεφάλαιο 11, μέσω κατασκευής bootstrap διαστημάτων εμπιστοσύνης της συνάρτησης πυκνότητας πιθανότητας, η αβεβαιότητα είναι αυξημένη στις κορυφές αυτής. □

3.5 Δυσκολίες πολυμεταβλητών προβλημάτων

Τα προβλήματα μη παραμετρικής εκτίμησης της αθροιστικής συνάρτησης κατανομής και της συνάρτησης πυκνότητας πιθανότητας σε περιπτώσεις πολυμεταβλητών παρατηρήσεων είναι πολύ πιο δύσκολα από τα αντίστοιχα στη μονοδιάστατη περίπτωση. Η *κατάρα των μεγάλων διαστάσεων* (curse of dimensionality) έχει ως αποτέλεσμα το απαιτούμενο μέγεθος δείγματος να αυξάνεται εκθετικά με τη διάσταση των δεδομένων. Υπό μία έννοια, σε μεγάλες διαστάσεις, τα δεδομένα είναι σχεδόν πάντα «αραιά» ακόμα και σε περιοχές υψηλής πυκνότητας. Οι Scott and Thompson (1983) χαρακτηρίζουν αυτήν τη συμπεριφορά ως «φαινόμενο κενού χώρου».

Για να γίνει πιο κατανοητό αυτό, ας υποθέσουμε ότι παρατηρούμε n ομοιόμορφα κατανεμημένα σημεία στο διάστημα $[-1,1]$. Στο ερώτημα, πόσα από αυτά αναμένεται να βρεθούν εντός του διαστήματος $[-0.1,0.1]$, η απάντηση είναι $n/10$. Έστω τώρα ότι παρατηρούμε n σημεία στον υπερ-κύβο 10 διαστάσεων

$$[-1,1]^{10} = [-1,1] \times \cdots \times [-1,1].$$

Σε αυτήν την περίπτωση, το αναμενόμενο πλήθος σημείων εντός του συνόλου $[-0.1,0.1]^{10}$ ισούται με:

$$n\left(\frac{0.2}{2}\right)^{10} = \frac{n}{10000000000}.$$

Ξεκάθαρα, σε αυτήν την περίπτωση, το n πρέπει να είναι ένας αρκετά μεγάλος αριθμός, για να εξασφαλιστεί ότι το αναμενόμενο πλήθος παρατηρήσεων σε «μικρά» υποσύνολα του στηρίγματος της κατανομής ξεπερνά ένα προκαθορισμένο μέγεθος.

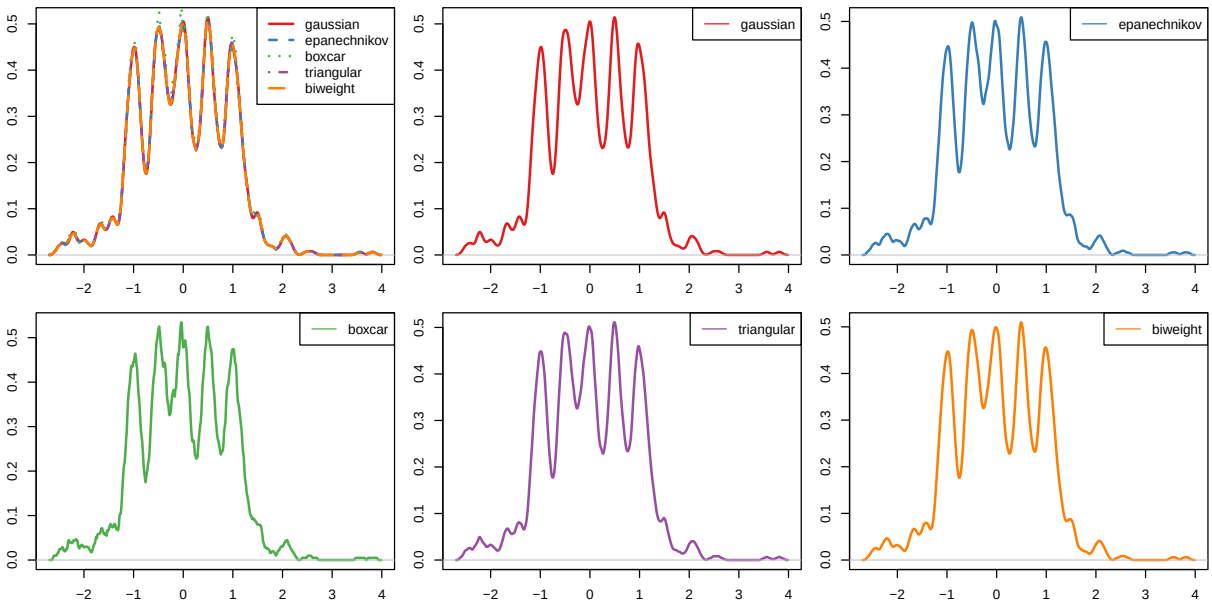
Έστω ότι παρατηρούμε p-διάστατα δεδομένα $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$. Τότε μπορούμε να γενικεύσουμε τον εκτιμητή της συνάρτησης πυκνότητας πιθανότητας με χρήση πυρήνα ως εξής:

$$f_{h,n}(\boldsymbol{x}) = \frac{1}{nh_1 \ldots, h_p} \sum_{i=1}^{n} \prod_{j=1}^{p} K\left(\frac{x_j - X_{ij}}{h_j}\right), \tag{3.40}$$

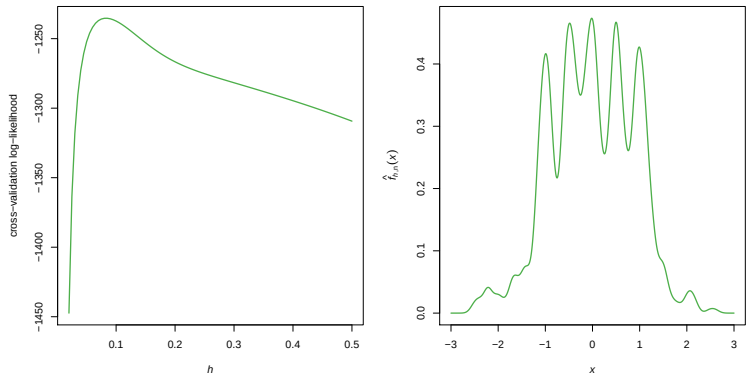
όπου $\boldsymbol{x} = (x_1, \ldots, x_p)$. Σε αυτήν την περίπτωση, μπορεί να δειχτεί (βλ. Wasserman, 2006, Ενότητα 6.5) ότι ο εκτιμητής της συνάρτησης πυκνότητας πιθανότητας με χρήση πυρήνα συγκλίνει με ρυθμό της τάξης $O(n^{-4/(4+p)})$. Παρατηρήστε ότι η σχέση (3.38) αποτελεί ειδική περίπτωση αυτού του αποτελέσματος για $p = 1$. Επομένως, παρατηρούμε πάλι ότι το σφάλμα αυξάνεται πολύ γρήγορα σε σχέση με τον αριθμό των διαστάσεων p. Για να αντιληφθούμε πιο άμεσα πώς η διάσταση επιδρά στην ακρίβεια των αποτελεσμάτων, ας θεωρήσουμε τον Πίνακα 3.3, ο οποίος δείχνει το απαιτούμενο μέγεθος δείγματος για να εξασφαλιστεί ότι

$$\mathrm{E}\left(f_{h,n}(\mathbf{0}) - f(\mathbf{0})\right)^2 / f(\mathbf{0}) \leqslant 0.1$$

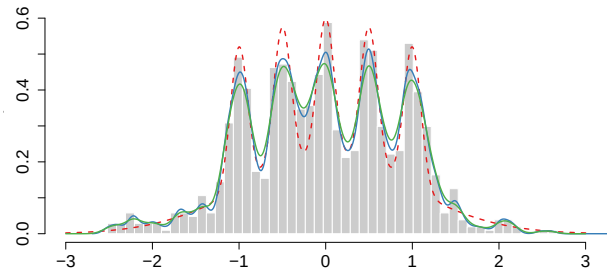
όταν η πραγματική f είναι η συνάρτηση πυκνότητας πιθανότητας της πολυδιάστατης κανονικής κατανομής, χρησιμοποιώντας έναν εκτιμητή με πυρήνα με βέλτιστη παράμετρο εξομάλυνσης. Παρατηρήστε ότι σε 10 διαστάσεις, το απαιτούμενο μέγεθος δείγματος είναι της τάξης του 10^6. Σημειώστε, επίσης, ότι εδώ το $f(\mathbf{0})$ αντιστοιχεί στην κορυφή της συνάρτησης πυκνότητας πιθανότητας και ότι τα πράγματα γίνονται ακόμα χειρότερα στις ουρές της κατανομής. Τα παραπάνω προβλήματα μπορούν να αντιμετωπιστούν μέσω κατάλληλων παραμετρικών ή ημι-παραμετρικών μοντέλων. Ο/Η ενδιαφερόμενος/μενη αναγνώστης/στρια ενδεικτικά παραπέμπεται στο σύγγραμμα της Frühwirth-Schnatter (2006).



Σχήμα 3.10: Διάφοροι εκτιμητές πυκνότητας με χρήση πυρήνα για τα δεδομένα του Παραδείγματος 3.5. Το πάνω αριστερά διάγραμμα περιέχει όλους τους εκτιμητές πυκνότητας μαζί.



Σχήμα 3.11: Αριστερά: η CVL λογαριθμική πιθανοφάνεια. Δεξιά: η εκτίμηση της $f_{h,n}(x)$ χρησιμοποιώντας την τιμή του bandwidth που μεγιστοποιεί την CVL ($h \approx 0.85$) και με χρήση gaussian kernel.



Σχήμα 3.12: Πραγματική συνάρτηση πυκνότητας πιθανότητας (κόκκινη διακεκομμένη γραμμή) και διάφορες μη παραμετρικές εκτιμήσεις αυτής.

p	Απαιτούμενο μέγεθος δείγματος
1	4
2	19
3	67
4	223
5	768
6	2790
7	10700
8	43700
9	187000
10	842000

Πίνακας 3.3: Απαιτούμενο μέγεθος δείγματος, έτσι ώστε το σχετικό μέσο τετραγωνικό σφάλμα να είναι μικρότερο από 0.1, όταν η f είναι η συνάρτηση πυκνότητας πιθανότητας της p-διάστατης κανονικής κατανομής. Πηγή: Κεφάλαιο 4 του Silverman (1986).

3.6 Ασκήσεις

Άσκηση 3.1. Να αποδείξετε τις σχέσεις (3.22) και (3.23).

Άσκηση 3.2. Έστω $Z \sim f_{h,n}(\cdot)$, όπου $f_{h,n}(\cdot)$ η συνάρτηση πυκνότητας πιθανότητας που προκύπτει από τον εκτιμητή πυκνότητας με χρήση πυρήνα που δόθηκε στη σχέση (3.30), με βάση ένα τυχαίο δείγμα $\boldsymbol{X} = (X_1, \dots, X_n)$. Να δειχτεί ότι:

1. $\mathrm{E}(Z) = \bar{X}$
2. $\mathrm{Var}(Z) = \hat{\sigma}^2 + h^2\sigma_K^2$

όπου $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$, με σ_K^2, όπως δόθηκε στην ιδιότητα 4 του Ορισμού 3.6.

Άσκηση 3.3. Έστω η συνάρτηση πυκνότητας πιθανότητας $f(x) = 2x,\ x \in [0,1]$. Θεωρήστε τον εκτιμητή ιστογράμματος με εύρος κελιών $h = 1/m$, $m = 1, 2, \dots$ ξεκινώντας από το $x_0 = 0$. Να δείξετε ότι

$$\mathrm{IMSE}\{f_{h,n}, f\} = \frac{1}{nh} + \frac{h^2}{3} - \frac{4}{3n} + \frac{h^2}{3n}.$$

Άσκηση 3.4. Έστω ότι $\boldsymbol{X} = (X_1, \dots, X_n)$ είναι ένα τυχαίο δείγμα από κατανομή με συνάρτηση πυκνότητας πιθανότητας f και έστω $f_{h,n}$ ο εκτιμητής αυτής με χρήση του πυρήνα

$$K(x) = \begin{cases} 1, & |x| < 1/2 \\ 0, & \text{διαφορετικά.} \end{cases}$$

Να δείξετε ότι

$$\mathrm{E}\left(f_{h,n}(x)\right) = \frac{1}{h}\int_{x-h/2}^{x+h/2} f(y)\mathrm{d}y$$

$$\mathrm{Var}\left(f_{h,n}(x)\right) = \frac{1}{nh^2}\left[\int_{x-h/2}^{x+h/2} f(y)\mathrm{d}y - \left(\int_{x-h/2}^{x+h/2} f(y)\mathrm{d}y\right)^2\right].$$

Για την επίλυση των παρακάτω ασκήσεων να χρησιμοποιηθεί η `R`.

Άσκηση 3.5. Θεωρήστε το σύνολο γαλαξιακών δεδομένων του Κεφαλαίου 2, το οποίο δίνεται στον Πίνακα 2.4.

1. Εκτιμήστε τη συνάρτηση πυκνότητας πιθανότητας των δεδομένων μέσω Ιστογράμματος. Ειδικότερα:
   (α) Θεωρήστε m ισομήκη διαστήματα στο σύνολο $[9, 35]$, όπου $m = 5, \dots, 50$ και βρείτε το βέλτιστο πλήθος κελιών (ισοδύναμα: εύρος κελιών) σύμφωνα με το ολοκληρωμένο μέσο τετραγωνικό σφάλμα.
   (β) Παραθέστε ένα διάγραμμα του εκτιμηθέντος «ρίσκου» $\hat{J}(h)$ για καθεμία τιμή του εκάστοτε εύρους κελιών h.
   (γ) Απεικονίστε το ιστόγραμμα που αντιστοιχεί στο βέλτιστο εύρος κελιών που επιλέξατε.
   (δ) Συγκρίνετε την εκτίμησή σας με το ιστόγραμμα που προκύπτει από την default εντολή της `R`: `hist(x, freq = F)`. Τι διαφορές υπάρχουν; Κατά τη γνώμη σας, ποιο ιστόγραμμα πρέπει να προτιμηθεί, δοθέντος ότι τα δεδομένα προέρχονται από διαφορετικές (μη ομογενείς) περιοχές του αστερισμού;
2. Εκτιμήστε τη συνάρτηση πυκνότητας πιθανότητας των δεδομένων με χρήση πυρήνα, αφού επιλέξετε το βέλτιστο εύρος παραθύρου:

(α) μέσω του ολοκληρωμένου μέσου τετραγωνικού σφάλματος, χρησιμοποιώντας την εντολή `density()` της `R`. Παραθέστε το γράφημα του εκτιμητή με χρήση κανονικού (gaussian), απλοϊκού (boxcar) ή (rectangular) και Epanechnikov πυρήνα, και

(β) μέσω της cross-validated πιθανοφάνειας, μόνο για τον κανονικό πυρήνα. Ποια είναι η τιμή του bandwidth που μεγιστοποιεί την cross-validated πιθανοφάνεια; Παραθέστε το σχετικό διάγραμμα.

3. Απεικονίστε τις 5 εκτιμήσεις της συνάρτησης πυκνότητας πιθανότητας σε ένα κοινό διάγραμμα.

Άσκηση 3.6. Οι παρακάτω τιμές καταγράφουν τους μισθούς των διευθυντών 59 εταιρειών (Πηγή:https://www.statcrunch.com/app/index.php?dataid=285919)

145	621	262	208	362	424	339	736	291	58	498	643	390	332	750
368	659	234	396	300	343	536	543	217	298	1103	406	254	862	204
206	250	21	298	350	800	726	370	536	291	808	543	149	350	242
198	213	296	317	482	155	802	200	282	573	388	250	396	572	

Εκτιμήστε τη συνάρτηση πυκνότητας πιθανότητας των μισθών χρησιμοποιώντας τη μέθοδο του ιστογράμματος και τη μέθοδο εκτίμησης με χρήση πυρήνα. Χρησιμοποιήστε την τεχνική ελαχιστοποίησης του ολοκληρωμένου μέσου τετραγωνικού σφάλματος για την επιλογή της παραμέτρου εξομάλυνσης. Να θεωρήσετε και τον κανονικό κανόνα για την επιλογή αυτής της παραμέτρου στον πυρήνα. Εφαρμόστε διαφορετικούς πυρήνες και σχολιάστε τις αντίστοιχες εκτιμήσεις.

Άσκηση 3.7. Να προτείνετε ένα ασυμπτωτικό διάστημα εμπιστοσύνης για τη διάμεσο θ μιας κατανομής με σ.π.π. f με βάση τη σχέση (2.16) του Κεφαλαίου 2. Εφαρμόστε το αποτέλεσμα για την εύρεση ενός 95% ασυμπτωτικού διαστήματος για τη διάμεσο των δεδομένων του Παραδείγματος 3.7.

Άσκηση 3.8. Μία εκτίμηση της αθροιστικής συνάρτησης κατανομής $F(x)$ μπορεί να προκύψει ολοκληρώνοντας την εκτίμηση της συνάρτησης πυκνότητας πιθανότητας με χρήση πυρήνα, δηλαδή: $F_{h,n}(x) := \int_{-\infty}^{x} f_{h,n}(u)\mathrm{d}u$. Να δείξετε ότι:

$$F_{h,n}(x) = \frac{1}{n}\sum_{i=1}^{n} G\left(\frac{x - X_i}{h}\right),$$

όπου $G(u) = \int_{-\infty}^{x} K(u)\mathrm{d}u$ και $K(u)$ είναι ο πυρήνας που χρησιμοποιήθηκε στον εκτιμητή πυκνότητας $f_{h,n}$. Θεωρήστε τις περιπτώσεις του (i) απλοϊκού (boxcar) και (ii) κανονικού πυρήνα και δώστε τις αντίστοιχες εκφράσεις για την $F_{h,n}(x)$. Ειδικά για την περίπτωση (ii) σχολιάστε τις διαφορές που έχει ένας τέτοιος εκτιμητής από την εμπειρική συνάρτηση κατανομής.

Άσκηση 3.9. Γράψτε τον δικό σας κώδικα στην `R` για τον υπολογισμό της εκτίμησης της συνάρτησης πυκνότητας πιθανότητας με χρήση κανονικού, Epanechnikov και tricube πυρήνα και εφαρμόστε τον στα δεδομένα του Παραδείγματος 3.7. Συγκρίνετε τα αποτελέσματα με αυτά που προκύπτουν από τον απλοϊκό πυρήνα στο Παράδειγμα 3.7.

Άσκηση 3.10. Γράψτε τον δικό σας κώδικα στην `R` για τον υπολογισμό της cross-validated πιθανοφάνειας της σχέσης (3.39). Κατόπιν, να αναπαράγετε τα αποτελέσματα του Παραδείγματος 3.10 στα Σχήματα 3.11 και 3.12.

Οι ασκήσεις που είναι σημειωμένες με ★ είναι πιο δύσκολες.

Άσκηση 3.11. (★) Η τεχνική *naive Bayes classifier* είναι μία δημοφιλής, αν και απλοϊκή, μέθοδος για την ταξινόμηση p-διάστατων μετρήσεων σε μία από K προκαθορισμένες ομάδες ενός πληθυσμού. Περιληπτικά, η τεχνική λειτουργεί ως ακολούθως. Η δεσμευμένη από κοινού συνάρτηση πυκνότητας πιθανότητας του $(X_1, \ldots, X_p)$, δοθέντος ότι αυτό ανήκει στην ομάδα $C = k$, είναι

$$f(\boldsymbol{x}|C = k) = \prod_{j=1}^{p} f_{X_j}(x_j|C = k), \quad k = 1, \ldots, K,$$

όπου $\boldsymbol{x} = (x_1, \ldots, x_p)$. Με άλλα λόγια, υποθέτουμε ότι οι p τυχαίες μεταβλητές $X_1, \ldots, X_p$ είναι δεσμευμένα ανεξάρτητες, δοθείσης της C. Έστω, επίσης, ότι είναι δεδομένες κάποιες εκ των προτέρων πιθανότητες κατάταξης $P(C = k) := \pi_k \geqslant 0$, $k = 1, \ldots, K$, με $\pi_1 + \ldots + \pi_K = 1$. Σύμφωνα με το Θεώρημα Bayes, η εκ των υστέρων πιθανότητα κατάταξης στην ομάδα k ισούται με:

$$P(C = k|\boldsymbol{x}) = \frac{\pi_k f(\boldsymbol{x}|C = k)}{\sum_{j=1}^{K} \pi_j f(\boldsymbol{x}|C = j)}, \quad k = 1, \ldots, K.$$

Επομένως, αν γνωρίζαμε όλες τις ποσότητες στην παραπάνω εξίσωση, μία νέα παρατήρηση $\boldsymbol{x}_0 = (x_{01}, \ldots, x_{op})^\top$, για την οποία δεν γνωρίζουμε την ομάδα του πληθυσμού από την οποία προέρχεται, ταξινομείται στην ομάδα εκείνη για την οποία μεγιστοποιείται η πιθανότητα κατάταξης $P(C = k|\boldsymbol{x}_0)$, $k = 1, \ldots, K$.

Δυστυχώς, οι δεσμευμένες (μονοδιάστατες) πυκνότητες πιθανότητας $f_{X_j}(\cdot|C = k)$, όπου $j = 1, \ldots, p$ και $k = 1, \ldots, K$, θεωρούνται άγνωστες. Ωστόσο, στη διάθεσή μας έχουμε ένα σύνολο p-διάστατων δεδομένων $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ για τα οποία έχει επίσης καταγραφεί και η ομάδα από την οποία προέρχονται, έστω $C_i \in \{1, 2, \ldots, K\}$, όπου $i = 1, \ldots, n$. Περιγράψτε έναν τρόπο εκτίμησης των εκ των υστέρων πιθανοτήτων μέσω εκτιμητών πυκνότητας με χρήση πυρήνα.

Άσκηση 3.12. (★) Το σύνολο δεδομένων του λουλουδιού της Ίριδας (Fisher, 1936) είναι διαθέσιμο στην `R` και σε αυτό καταγράφονται οι τιμές $p = 4$ χαρακτηριστικών για $n = 150$ το πλήθος λουλούδια τα οποία προέρχονται από $K = 3$ διαφορετικά είδη (Iris setosa, versicolor, virginica). Το σύνολο δεδομένων μπορεί να ανακτηθεί μέσω της `R` με την εντολή

```
1 iris
```

Ειδικότερα, στο σύνολο αυτό καταγράφονται δεδομένα για τις ακόλουθες μεταβλητές: Sepal.Length (μήκος σεπάλου), Sepal.Width (πλάτος σεπάλου), Petal.Length (μήκος πετάλου), Petal.Width (πλάτος πετάλου) και Species (είδος).

Το ζητούμενο είναι να βρεθεί ένας κανόνας κατάταξης των λουλουδιών στα τρία είδη με βάση το πλάτος και μήκος σεπάλου και πετάλου. Εφαρμόστε την τεχνική naive Bayes classifier με εκτιμητή πυκνοτήτας με χρήση πυρήνα, όπως περιγράφθηκε στην Άσκηση 3.11. Μπορείτε να χωρίσετε τα δεδομένα σας σε train και test τμήματα για να αξιολογήσετε την προβλεπτική ικανότητα του μοντέλου.

Άσκηση 3.13. (★) Το σύνολο δεδομένων ιταλικών κρασιών (Forina *et al.*, 1986) είναι διαθέσιμο στη βιβλιοθήκη `pgmm` (McNicholas *et al.*, 2019) της `R` και σε αυτό καταγράφονται οι τιμές $p = 27$ χημικών και φυσικών χαρακτηριστικών για $n = 178$ ιταλικά κρασιά, προερχόμενα από $K = 3$ διαφορετικές ποικιλίες (Barolo, Grignolino, Barbera). Το σύνολο δεδομένων μπορεί να ανακτηθεί μέσω της `R` με τις εντολές

```
library("pgmm")
data(wine)
wine
```

Το ζητούμενο είναι να βρεθεί ένας κανόνας κατάταξης των κρασιών με βάση τις φυσικές και χημικές τους ιδιότητες. Εφαρμόστε την τεχνική naive Bayes classifier με εκτιμητή πυκνοτήτας με χρήση πυρήνα, όπως περιγράφθηκε στην Άσκηση 3.11. Μπορείτε να χωρίσετε τα δεδομένα σας σε train και test τμήματα για να αξιολογήσετε την προβλεπτική ικανότητα του μοντέλου.

Άσκηση 3.14. $^{(\star)}$ Προσομοιώστε $n = 100$ το πλήθος παρατηρήσεις από μία διδιάσταστη κανονική κατανομή, αφού επιλέξετε τις παραμέτρους αυτής. Κατόπιν:

1. Γράψτε δικό σας κώδικα στην `R` που να υπολογίζει τον εκτιμητή πυκνότητας με χρήση γινομένου κανονικών πυρήνων, που δόθηκε στη σχέση (3.40).
2. Θεωρήστε την ειδική περίπτωση όπου $h_1 = h_2 = h$ και επιλέξτε την παράμετρο εξομάλυνσης μέσω cross-validated πιθανοφάνειας.
3. Επαναλάβατε το προηγούμενο ερώτημα για τη γενική περίπτωση όπου $h_1 \neq h_2$.

Άσκηση 3.15. $^{(\star)}$ Τα δεδομένα εκρήξεων (Azzalini and Bowman, 1990) καταγράφουν τις τιμές των μεταβλητών eruption: χρόνος έκρηξης σε λεπτά, και waiting: χρόνος αναμονής μέχρι την επόμενη έκρηξη, για $n = 272$ το πλήθος εκρήξεις του θερμοπίδακα Old Faithful στο εθνικό πάρκο Yellowstone της πολιτείας Wyoming των Ηνωμένων Πολιτειών της Αμερικής. Αρχικά, φορτώστε τα δεδομένα αυτά στην `R` μέσω της εντολής:

```
faithful
```

Έπειτα εκτιμήστε τη διδιάστατη συνάρτηση πυκνότητας πιθανότητας κάνοντας χρήση των αποτελεσμάτων της Άσκησης 3.14.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Καρλής, Δ. (2004). *Υπολογιστική Στατιστική*. Οικονομικό Πανεπιστήμιο Αθηνών.

Ρούσσας, Γ. (1998). *Θεωρία Πιθανοτήτων* (2η εκδ.), (Δ. Ιωαννίδης, Μετ.). 2η Έκδοση. Θεσσαλονίκη: Ζήτη.

Ξενόγλωσση

Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *39*(3), pp. 357–365.

Cawley, G. C. and Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, pp. 2079–2107.

David, M. A. (1974). The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction. *Technometrics*, *16*(1), pp. 125–127.

Duin, R. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Transactions on Computers*, *25*(11), pp. 1175–1179.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, *14*(1), pp. 153–158.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), pp. 179–188.

Forina, M., Armanino, C., Castino, M. and Ubigli, M. (1986). Multivariate data analysis as a discriminating method of the origin of wines. *Vitis*, *25*, pp. 189–201.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Science & Business Media.

Habbema, J. D. F. (1974). A stepwise discriminant analysis program using density estimation. In: *Compstat 1974, Proceedings in Computational Statistics,Physica, Vienna.* Pp. 101–110.

Hansen, B. E. (2009). Lecture notes on nonparametrics. *Lecture notes*.

Härdle, W. K. (1991). *Smoothing techniques: with implementation in S*. Springer Science & Business Media.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), pp. 1–12.

McNicholas, P. D., ElSherbiny, A., McDaid, A. F. and Murphy, T. B. (2019). *pgmm: Parsimonious Gaussian Mixture Models*. R package version 1.2.4. URL: https://CRAN.R-project.org/package=pgmm.

Muller, H.-G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *The Annals of Statistics*, *12*, pp. 766–774.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, *33*(3), pp. 1065–1076.

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, *27*(3), pp. 832–837.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, *66*(3), pp. 605–610.

Scott, D. W. (2012). Multivariate density estimation and visualization. In: *Handbook of Computational Statistics*. Springer, pp. 549–569.

Scott, D. W. and Thompson, J. R. (1983). Probability density estimation in higher dimensions. In: *Computer Science and Statistics: Proceedings of the fifteenth symposium on the interface*. Vol. 528. North-Holland, Amsterdam, pp. 173–179.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, New York.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), pp. 111–133.

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), pp. 44–47.

Titterington, D. M., Afm, S., Smith, A. F. and Makov, U. (1985). *Statistical analysis of finite mixture distributions*. Vol. 198. John Wiley & Sons Incorporated.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Vol. 3. Cambridge University Press.

Van Es, B. (1991). Likelihood cross-validation bandwidth selection for nonparametric kernel density estimators. *Journal of Nonparametric Statistics*, *1*(1-2), pp. 83–110.

Wasserman, L. (2006). *All of Nonparametric Statistics*. New York, NY: Springer Texts in Statistics.

ΚΕΦΑΛΑΙΟ 4

ΕΛΕΓΧΟΙ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ

Σύνοψη

Συχνά ο στατιστικός, προτού προχωρήσει στην εφαρμογή συγκεκριμένων μεθοδολογιών, επιθυμεί να ελέγξει κατά πόσο οι διαθέσιμες δειγματικές παρατηρήσεις προσαρμόζονται σε κάποιο συγκεκριμένο (θεωρητικό) μοντέλο. Ο έλεγχος αυτός αναφέρεται στη βιβλιογραφία ως έλεγχος καλής προσαρμογής (goodness-of-fit test) και υποδεικνύει πόσο καλά ένα σύνολο δεδομένων περιγράφεται/προσαρμόζεται από ή σε ένα συγκεκριμένο μοντέλο. Ουσιαστικά, οι έλεγχοι καλής προσαρμογής μας δίνουν τον βαθμό ασυμφωνίας ή τον βαθμό εγγύτητας των παρατηρούμενων τιμών (observed values) στις τιμές που αναμένονται (expected values), αν υιοθετήσουμε το υπό εξέταση μοντέλο. Σκοπός αυτού του κεφαλαίου είναι η παρουσίαση του ελέγχου χι-τετράγωνο καλής προσαρμογής, που αποτελεί τον αρχαιότερο τέτοιο έλεγχο, καθώς και παραδοσιακών ελέγχων που αξιοποιούν τις ιδιότητες της εμπειρικής αθροιστικής συνάρτησης που παρουσιάστηκαν στο Κεφάλαιο 2. Τέλος, λόγω της σπουδαιότητας της κανονικής κατανομής, ειδική αναφορά γίνεται στους τρόπους ελέγχου της καλής προσαρμογής των δεδομένων στο μοντέλο της κανονικής κατανομής (έλεγχοι κανονικότητας).

Προαπαιτούμενη γνώση:
Βασικές γνώσεις Πιθανοτήτων και Στατιστικής.

Προσδοκώμενα μαθησιακά αποτελέσματα:
Ο/η φοιτητής/τρια ολοκληρώνοντας την ενότητα αυτή θα μπορεί να εφαρμόζει κάποιους από τους πλέον γνωστούς ελέγχους καλής προσαρμογής. Ειδικότερα, θα είναι σε θέση να μπορεί να ελέγχει αν ένα σύνολο δεδομένων προέρχεται από έναν πληθυσμό που περιγράφεται ικανοποιητικά από την κανονική κατανομή (έλεγχος κανονικότητας).

Γλωσσάριο επιστημονικών όρων

- Έλεγχος καλής προσαρμογής
- Έλεγχος χι-τετράγωνο καλής προσαρμογής
- Έλεγχοι κανονικότητας
- Έλεγχος Anderson-Darling
- Έλεγχος Cramér - von Mises
- Έλεγχος Kolmogorov-Smirnov
- Έλεγχος Kuiper
- Έλεγχος Shapiro-Wilk
- Έλεγχος Watson

Μπατσίδης, Α., Παπασταμούλης, Π., Πετρόπουλος, Κ., & Ρακιτζής, Α. (2022). *Μη Παραμετρική Στατιστική.* [Προπτυχιακό εγχειρίδιο].

4.1 Εισαγωγή

Έστω ότι $X_1, X_2, \dots, X_n$, είναι τυχαίες παρατηρήσεις μίας τυχαίας μεταβλητής, η οποία μπορεί να είναι συνεχής ή διακριτή, με αθροιστική συνάρτηση κατανομής (α.σ.κ.) $F(x)$, η οποία είναι άγνωστη. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση $H_0 : F(x) = F_0(x)$, για κάθε $x \in \mathbb{R}$, όπου $F_0(x)$ είναι μία ειδική, συγκεκριμένη, αθροιστική συνάρτηση κατανομής. Το πρόβλημα ελέγχου της παραπάνω υπόθεσης καλείται **έλεγχος καλής προσαρμογής**. Οι έλεγχοι καλής προσαρμογής διακρίνονται σε απλούς και σύνθετους ελέγχους. Λέμε ότι έχουμε έναν απλό έλεγχο καλής προσαρμογής, αν η αθροιστική συνάρτηση κατανομής υπό τη μηδενική υπόθεση είναι πλήρως ορισμένη, ενώ, διαφορετικά, λέμε ότι έχουμε έναν σύνθετο έλεγχο καλής προσαρμογής. Για παράδειγμα, πρόκειται για έναν απλό έλεγχο καλής προσαρμογής αν θέλουμε να ελέγξουμε αν το τυχαίο δείγμα προέρχεται από έναν πληθυσμό που περιγράφεται ικανοποιητικά από την κανονική κατανομή με μέση τιμή 5 και τυπική απόκλιση 2. Από την άλλη μεριά, αν θέλουμε να ελέγξουμε αν το τυχαίο δείγμα προέρχεται από έναν πληθυσμό που περιγράφεται ικανοποιητικά από την κανονική κατανομή, χωρίς να προσδιορίζονται οι παράμετροί της, τότε έχουμε έναν σύνθετο έλεγχο καλής προσαρμογής. Επίσης, οι έλεγχοι καλής προσαρμογής διακρίνονται σε δίπλευρους και μονόπλευρους, ανάλογα με τη μορφή της εναλλακτικής υπόθεσης. Μιλάμε για δίπλευρο έλεγχο, αν η εναλλακτική είναι η $H_1 : F(x) \neq F_0(x)$, για κάποιο $x \in \mathbb{R}$, και για μονόπλευρους, αν η εναλλακτική είναι είτε η $H_{1+} : F(x) \geqslant F_0(x)$, για κάποιο $x \in \mathbb{R}$, είτε η $H_{1-} : F(x) \leqslant F_0(x)$, για κάποιο $x \in \mathbb{R}$.

Στο παραπάνω πλαίσιο, έχουν αναπτυχθεί διάφορες στατιστικές μεθοδολογίες, τόσο για τον έλεγχο ότι ένα σύνολο δεδομένων προέρχεται από έναν συγκεκριμένο πληθυσμό, π.χ. κανονικό, εκθετικό κ.ο.κ., όσο και αν δύο ή περισσότερα σύνολα δεδομένων μπορούν να θεωρηθούν ότι προέρχονται από τον ίδιο πληθυσμό. Στο κεφάλαιο αυτό, αρχικά, παρουσιάζεται ο έλεγχος χι-τετράγωνο καλής προσαρμογής, που αποτελεί τον αρχαιότερο τέτοιο έλεγχο. Στη συνέχεια, παρουσιάζονται έλεγχοι καλής προσαρμογής που βασίζονται στην αθροιστική εμπειρική συνάρτηση κατανομής (ή εμπειρική συνάρτηση κατανομής, βλ. Κεφάλαιο 2) και στην εγγύτητά της από την αθροιστική συνάρτηση κατανομής της υπό έλεγχο κατανομής. Τέλος, λόγω της σπουδαιότητας της κανονικής κατανομής, το κεφάλαιο ολοκληρώνεται με μια σύντομη αναφορά σε τρόπους ελέγχου της καλής προσαρμογής των δεδομένων στο μοντέλο της κανονικής κατανομής (έλεγχοι κανονικότητας).

Προτού προχωρήσουμε στην παρουσίαση των παραπάνω, δίνουμε σε αυτήν την εισαγωγική ενότητα του κεφαλαίου, εν συντομία, τις βασικές ιδιότητες της Πολυωνυμικής κατανομής (βλ., επίσης, Κούτρας (2018)), καθώς η κατανομή αυτή εμφανίζεται κατά την παρουσίαση του ελέγχου χι-τετράγωνο καλής προσαρμογής.

Πολυωνυμική κατανομή

Έστω $X_0, X_1, X_2, \dots, X_M$ ο αριθμός εμφανίσεων του αποτελέσματος $F, S_1, S_2, \dots, S_M$ αντίστοιχα σε $n \in \mathbb{N}$ το πλήθος ανεξάρτητες δοκιμές ενός πειράματος τύχης, με $M + 1$ το πλήθος δυνατά αποτελέσματα $F, S_1, S_2, \dots, S_M$. Το αποτέλεσμα F αναφέρεται ως «αποτυχία», ενώ το S_j, $j = 1, 2, \dots, M$ ως «επιτυχία j είδους». Επίσης, οι πιθανότητες εμφάνισης των $M + 1$ αποτελεσμάτων είναι $p_0, p_1, p_2, \dots, p_M$ (ίδιες σε κάθε δοκιμή), με $p_0 = 1 - (p_1 + p_2 + \dots + p_M)$. Η πολυδιάστατη τυχαία μεταβλητή $\mathbf{X} = (X_0, X_1, X_2, \dots, X_M)$ λέγεται ότι ακολουθεί την **Πολυωνυμική κατανομή** με παραμέτρους n και $p_0, p_1, p_2, \dots, p_M$, $p_i \in (0,1)$, $i = 0, 1, \dots, M$, αν για τις δυνατές τιμές της $x_0, x_1, x_2, \dots, x_M$ ισχύει ότι $0 \leqslant x_0, x_1, \dots, x_M \leqslant n$, με $\sum_{i=0}^{M} x_i \leqslant n$ και η από κοινού συνάρτηση πιθανότητάς της δίνεται από τη σχέση:

$$f_{\mathcal{M}}(x_0, x_1, \dots, x_M; n, p_0, p_1, \dots, p_M) = \frac{n!}{x_0! x_1! \dots x_M!} p_0^{x_0} p_1^{x_1} \dots p_M^{x_M}. \tag{4.1}$$

Στην περίπτωση αυτή, γράφουμε ότι

$$\mathbf{X} \sim \mathcal{M}(n; p_0, p_1, \dots, p_M).$$

Επιπλέον, $\mathrm{E}(\mathbf{X}) = (\mathrm{E}(X_0), \mathrm{E}(X_1), \dots, \mathrm{E}(X_M))^T$, όπου $\mathrm{E}(X_i) = np_i$, $i = 0, 1, \dots, M$. Επίσης, για τον πίνακα διακυμάνσεων-συνδιακυμάνσεων $\mathbf{\Sigma}$ της πολυδιάστατης τυχαίας μεταβλητής $\mathbf{X} = (X_0, X_1, X_2, \dots, X_M)^T$ ισχύει

ότι:

$$\mathbf{\Sigma} = \begin{pmatrix} \text{Var}(X_0) & \text{Cov}(X_0, X_1) & \text{Cov}(X_0, X_2) & \dots & \text{Cov}(X_0, X_M) \\ \text{Cov}(X_1, X_0) & \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_M) \\ \text{Cov}(X_2, X_0) & \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_M) \\ \dots & \dots & \dots & \dots & \dots \\ \text{Cov}(X_M, X_0) & \text{Cov}(X_M, X_1) & \text{Cov}(X_M, X_2) & \dots & \text{Var}(X_M) \end{pmatrix}$$

όπου $\text{Var}(X_i) = np_i(1-p_i)$, $i = 0,1,2,\dots,M$ και $\text{Cov}(X_i, X_j) = -np_ip_j$, για $i,j = 0,1,\dots,M$ με $i \neq j$. Θα συμβολίζουμε την αθροιστική συνάρτηση κατανομής της $\mathcal{M}(n;\ p_0,\ p_1,\ \dots\ ,\ p_M)$ ως $F_{\mathcal{M}}(x_0, x_1, \dots, x_M; n, p_0, p_1, \dots, p_M)$.

Για $M = 1$ προκύπτει η Διωνυμική κατανομή, ενώ οι περιθώριες κατανομές των τ.μ. $X_0, X_1, X_2, \dots, X_M$ είναι οι αντίστοιχες Διωνυμικές με παραμέτρους n και p_i.

4.2 Έλεγχος χι-τετράγωνο καλής προσαρμογής

Έστω ότι $X_1, X_2, \dots, X_n$, είναι τυχαίες παρατηρήσεις μίας τυχαίας μεταβλητής, που μπορεί να είναι συνεχής ή διακριτή, με αθροιστική συνάρτηση κατανομής $F(x)$, η οποία είναι άγνωστη. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση:

$$H_0 : F(x) = F_0(x), \text{ για κάθε } x \in \mathbb{R},$$

έναντι της εναλλακτικής υπόθεσης ότι:

$$H_1 : F(x) \neq F_0(x), \text{ για κάποιο } x \in \mathbb{R},$$

όπου $F_0(x)$ είναι μία ειδική (συγκεκριμένη) αθροιστική συνάρτηση κατανομής. Αρχικά υποθέτουμε ότι η $F_0(x)$ μας είναι πλήρως γνωστή (απλός έλεγχος καλής προσαρμογής).

Απλός έλεγχος χι-τετράγωνο καλής προσαρμογής

Ο αρχαιότερος και περισσότερο γνωστός έλεγχος καλής προσαρμογής είναι ο X^2 (χι-τετράγωνο), ο οποίος προτάθηκε από τον Άγγλο μαθηματικό και βιοστατιστικό Karl Pearson (1857-1936) στην εργασία του Pearson (1900)[1]. Ο έλεγχος αυτός απαιτεί τα δεδομένα να είναι σε ομάδες ή να δύναται να χωριστούν σε ομάδες και αξιοποιεί, στη συνέχεια, τις συχνότητες κάθε ομάδας. Επομένως, αν η τυχαία μεταβλητή X είναι συνεχής με σύνολο δυνατών τιμών το διάστημα (α,β), για να εφαρμοστεί ο έλεγχος χι-τετράγωνο το διάστημα (α,β) χωρίζεται σε k το πλήθος μη επικαλυπτόμενα υποδιαστήματα, $I_1 = (a, x_1]$, $I_2 = (x_1, x_2]$,..., $I_k = (x_{k-1}, \beta)$. Επομένως, με την παραπάνω διαδικασία, στην ουσία διακριτοποιούμε ή κατηγοριοποιούμε την κατανομή σε k το πλήθος κατηγορίες, που ουσιαστικά αποτελούν μία διαμέριση του διαστήματος (α,β).

Επιπρόσθετα, συμβολίζουμε με p_{0i} την πιθανότητα, υπό τη μηδενική υπόθεση, μία τυχαία παρατήρηση από την τυχαία μεταβλητή X να ανήκει στην i-οστή κατηγορία, $i = 1,\dots,k$. Επομένως,

$$p_{01} = F_0(x_1) - F_0(\alpha), p_{02} = F_0(x_2) - F_0(x_1), \dots, p_{0k} = F_0(\beta) - F_0(x_{k-1}),$$

με $\sum_{i=1}^{k} p_{0i} = 1$. Τότε, αν $N_1, N_2, \dots, N_k$ είναι οι τυχαίες μεταβλητές που παριστάνουν το πλήθος των παρατηρήσεων που ανήκουν στο υποδιάστημα $I_1,\dots,I_k$, αντίστοιχα, στις n που έχουν επιλεχθεί, είναι προφανές ότι το τυχαίο διάνυσμα $(N_1, \dots, N_k)$ ακολουθεί, υπό τη μηδενική υπόθεση, πολυωνυμική

[1] Η μεθοδολογία πρωτοθεμελιώθηκε από τον Fisher (1924) και μία πλήρης απόδειξη παρατίθεται από τον Cramér (1946).

κατανομή με παραμέτρους n, $p_{01},...,\ p_{0k}$. Επιπλέον, η περιθώρια κατανομή της τυχαίας μεταβλητής N_i ακολουθεί, υπό τη μηδενική υπόθεση, διωνυμική κατανομή με παραμέτρους n και p_{0i}.

Στο παραπάνω πλαίσιο, έστω n_i (e_i) ο παρατηρούμενος (αναμενόμενος υπό τη μηδενική υπόθεση) αριθμός των παρατηρήσεων που ανήκουν στο υποδιάστημα $I_i, i = 1,...,k$. Δηλαδή, n_i είναι το πλήθος των δειγματικών τιμών που ανήκουν στο διάστημα I_i, με $n_1 + n_2 + ... + n_k = n$, ενώ, από το γεγονός ότι υπό τη μηδενική υπόθεση $N_i \sim B\left(n, p_{0i}\right)$, έχουμε ότι:

$$e_i = \mathrm{E}(N_i) = np_{0i}, i = 1,...,k,$$

με $\sum_{i=1}^{k} e_i = n$. Ο X^2 έλεγχος καλής προσαρμογής αντιπαραβάλλει τον αριθμό των παρατηρούμενων παρατηρήσεων που ανήκουν στο i-οστό υποδιάστημα με τον αναμενόμενο αριθμό αυτών υπό την H_0 και βασίζεται στη στατιστική συνάρτηση:

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}. \tag{4.2}$$

Παρατήρηση 4.1. Μια εναλλακτική, ισοδύναμη έκφραση του X^2 ελέγχου καλής προσαρμογής, η οποία προκύπτει με λίγη άλγεβρα, είναι η:

$$X^2 = \sum_{i=1}^{k} \frac{n_i^2}{e_i} - n. \tag{4.3}$$

Είναι προφανές ότι, όταν δεν ισχύει η μηδενική υπόθεση, τότε κάθε παρατηρούμενος αριθμός n_i θα διαφέρει αρκετά από τον αντίστοιχο αναμενόμενο. Άρα, καθώς, η στατιστική συνάρτηση είναι το άθροισμα των τετραγώνων αυτών των διαφορών διαιρεμένων με τον αντίστοιχο αναμενόμενο αριθμό αυτών, συνεπάγεται ότι η μηδενική υπόθεση θα απορρίπτεται για μεγάλες τιμές της στατιστικής συνάρτησης. Το εύλογο ερώτημα που προκύπτει είναι ποιες τιμές της στατιστικής συνάρτησης μπορούν να θεωρηθούν μεγάλες. Η απάντηση αυτή μπορεί να δοθεί, αν προσδιοριστεί η κατανομή της στατιστικής συνάρτησης υπό τη μηδενική υπόθεση. Ο προσδιορισμός αυτός επιτυγχάνεται στην πρόταση που ακολουθεί.

Πρόταση 4.1. Έστω $I_1,...,I_k$ μια διαμέριση του διαστήματος (α, β), που αποτελεί το σύνολο των δυνατών τιμών μιας τυχαίας μεταβλητής X. Έστω $N_1, N_2, \ldots, N_k$ είναι οι τυχαίες μεταβλητές που παριστάνουν το πλήθος των παρατηρήσεων που ανήκουν στα υποδιαστήματα $I_1,...,I_k$, αντίστοιχα, στις n τυχαία επιλεγμένες παρατηρήσεις από την X. Τότε, υπό την υπόθεση ότι η αθροιστική συνάρτηση κατανομής της τυχαίας μεταβλητής X είναι η $F_0(\cdot)$, η τυχαία μεταβλητή

$$X^2 = \sum_{i=1}^{k} \frac{(N_i - e_i)^2}{e_i}, \tag{4.4}$$

όπου $e_i = np_{0i} = nP(X \in I_i|H_0)$, ακολουθεί ασυμπτωτικά ($n \to \infty$) χι-τετράγωνο κατανομή με $k-1$ βαθμούς ελευθερίας (χ^2_{k-1}).

Απόδειξη Πρότασης 4.1. Για την απόδειξη της πρότασης παραπέμπουμε τον/την ενδιαφερόμενο/μενη αναγνώστη/στρια, μεταξύ άλλων, στους Fisher (1924), Cramér (1946), Gibbons and Chakraborti (2020). □

Από την παραπάνω πρόταση προκύπτει ότι απορρίπτουμε σε επίπεδο σημαντικότητας a την H_0, αν για την παρατηρούμενη τιμή της στατιστικής συνάρτησης ελέγχου, έστω X^2_{obs}, ισχύει ότι:

$$X^2_{obs} = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \geqslant \chi^2_{k-1,a}, \tag{4.5}$$

όπου $\chi^2_{k-1,a}$ είναι τέτοιο ώστε

$$P\left(\chi^2_{k-1} \geqslant \chi^2_{k-1,a}\right) = a.$$

Τιμές των $\chi^2_{k-1,a}$ παρατίθενται στους Πίνακες Π.3 και Π.4 του Παραρτήματος για διάφορους βαθμούς ελευθερίας και επίπεδα σημαντικότητας. Άμεσα προκύπτει ότι η $p-$τιμή του ελέγχου προσδιορίζεται από τη σχέση:

$$P\left(\chi^2_{k-1} \geqslant X^2_{obs}\right) = 1 - F_{\chi^2_{k-1}}\left(X^2_{obs}\right),$$

όπου $F_{\chi^2_{k-1}}$ η αθροιστική συνάρτηση κατανομής της χ^2_{k-1} κατανομής.

Παρατήρηση 4.2. Ο έλεγχος χι-τετράγωνο καλής προσαρμογής μπορεί να χρησιμοποιηθεί για οποιαδήποτε μονοδιάστατη κατανομή αρκεί να μπορούμε να υπολογίσουμε την αθροιστική συνάρτηση κατανομής. Ένα πρώτο μειονέκτημά του είναι ότι πρόκειται για ασυμπτωτικό έλεγχο και, επομένως, πρέπει να έχουμε διαθέσιμο μεγάλο σε μέγεθος δείγμα. Επιπλέον, παρατηρούμε ότι απαιτεί τα δεδομένα να διαχωριστούν σε ομάδες-κατηγορίες. Συνεπώς, το ερώτημα που προκύπτει είναι αν ο αριθμός των ομάδων και ο τρόπος καθορισμού τους επηρεάζουν την ισχύ του ελέγχου. Δυστυχώς, η απάντηση είναι καταφατική και έχει οδηγήσει να έχουν προταθεί διάφοροι κανόνες για το πλήθος των ομάδων που πρέπει να χρησιμοποιηθούν και τον τρόπο που αυτές οι ομάδες θα πρέπει να καθοριστούν, έτσι ώστε να έχουμε έναν ισχυρότερο έλεγχο. Για παράδειγμα, έχει προταθεί ο αριθμός των ομάδων k:

- να είναι $k = 2n^{\frac{2}{5}}$ (βλ. Moore, 1986),
- να είναι τέτοιος ώστε: $k \geqslant 3$, $n^2/k \geqslant 10$ και $e_i \geqslant 0.25$ (Kvam and Vidakovic, 2007),
- να είναι τέτοιος ώστε κάθε ομάδα να έχει, υπό τη μηδενική υπόθεση, αναμενόμενο αριθμό παρατηρήσεων μεγαλύτερο από πέντε (Gibbons and Chakraborti, 2020).

Ο τελευταίος κανόνας είναι και ο πιο γνωστός. Σε περίπτωση που ο αναμενόμενος αριθμός παρατηρήσεων σε μια ομάδα δεν ικανοποιεί αυτήν την απαίτηση, τότε ενοποιούμε τις μικρότερες κατηγορίες έτσι ώστε να ισχύει (βλ. Gibbons and Chakraborti, 2020) . Τέλος, όσον αφορά τον τρόπο καθορισμού των ορίων των ομάδων έχει προταθεί να επιλέγονται έτσι ώστε η πιθανότητα η τυχαία μεταβλητή να ανήκει σε καθεμία από αυτές να είναι η ίδια.

Ο απλός έλεγχος καλής προσαρμογής με το χι-τετράγωνο τεστ αποσαφηνίζεται μέσω των παραδειγμάτων που ακολουθούν.

Παράδειγμα 4.1. (X^2 τεστ προσαρμογής Πολυωνυμικής κατανομής). Σύμφωνα με μία θεωρία υπάρχουν 4 ποικιλίες ενός φυτού με αναλογίες $9/16, 3/16, 3/16$ και $1/16$, αντίστοιχα. Σε ένα τυχαίο δείγμα φυτών μεγέθους $n = 278$ βρέθηκε ότι στις 4 ποικιλίες ανήκουν 157, 54, 51 και 16 φυτά, αντίστοιχα. Να ελεγχθεί σε επίπεδο σημαντικότητας $a = 5\%$ η θεωρία σχετικά με τις ποικιλίες.

Λύση Παραδείγματος 4.1. Στην ουσία, έχουμε ένα τυχαίο πείραμα από $n = 278$ ανεξάρτητες δοκιμές στο οποίο μπορούμε να έχουμε $k = 4$ δυνατά αποτελέσματα, ξένα μεταξύ τους. Έστω E_i, $i = 1, ..., 4$, το ενδεχόμενο ένα φυτό να ανήκει στην i-οστή ποικιλία, $i = 1, ..., 4$, με $p_1 = P(E_1)$, $p_2 = P(E_2)$, $p_3 = P(E_3)$ και $p_4 = P(E_4)$ οι αντίστοιχες πιθανότητες πραγματοποίησης καθενός από τα ενδεχόμενα E_i, $i = 1, ..., 4$. Επιπλέον, έστω N_1, N_2, N_3, N_4 οι τυχαίες μεταβλητές που παριστάνουν το πλήθος των φυτών που ανήκουν σε καθεμία από τις τέσσερις ποικιλίες, στα n τυχαία επιλεγμένα φυτά. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση ότι:

$$H_0 : (N_1, ..., N_4) \sim \mathcal{M}(n, 9/16, 3/16, 3/16, 1/16)$$

έναντι της εναλλακτικής ότι τουλάχιστον μια αναλογία διαφέρει από τη δοθείσα αναλογία στην H_0.

Ο έλεγχος γίνεται με τη στατιστική συνάρτηση

$$X^2 = \sum_{i=1}^{4} \frac{(n_i - e_i)^2}{e_i},$$

όπου n_i και e_i είναι ο παρατηρούμενος και αναμενόμενος (υπό τη μηδενική υπόθεση) αριθμός εμφανίσεων του E_i, $i = 1,...,4$. Είναι $e_i = np_{i0}$, $i = 1,...,4$, με $p_{i0} = P(E_i|H_0)$, δηλαδή $(p_{10},...,p_{40}) = (9/16,3/16,3/16,1/16)$. Επομένως,

$$e_1 = 278 \cdot \frac{9}{16} = 156.375,\ e_2 = 278 \cdot \frac{3}{16} = 52.125\ e_3 = 278 \cdot \frac{3}{16} = 52.125 \text{ και } e_4 = 278 \cdot \frac{1}{16} = 17.375.$$

Άρα, καθώς το μέγεθος του δείγματος n είναι μεγάλο και $e_i \geq 5$, $i = 1,...,4$, δηλαδή καθώς πληρούνται οι προϋποθέσεις εφαρμογής του X^2 ελέγχου καλής προσαρμογής, προβαίνουμε στον υπολογισμό του. Είναι

$$X^2 = \sum_{i=1}^{4} \frac{(n_i - e_i)^2}{e_i} = \frac{(157 - 156.375)^2}{156.375} + \frac{(54 - 52.125)^2}{52.125} + \frac{(51 - 52.125)^2}{52.125} + \frac{(16 - 17.375)^2}{17.375}$$

ή, μετά από λίγη άλγεβρα,

$$X^2 = 0.0025 + 0.0674 + 0.0242 + 0.1088 = 0.2029.$$

Απορρίπτουμε την H_0, αν:

$$X^2 \geq \chi^2_{k-1,a} = \chi^2_{3,0.05} = 7.815,$$

όπου η κρίσιμη τιμή $\chi^2_{3,0,05}$ βρέθηκε χρησιμοποιώντας τον Πίνακα Π.4 του Παραρτήματος.

Επομένως, καθώς 0.2029<7.815, δεν μπορεί να απορριφθεί η μηδενική υπόθεση. □

Παράδειγμα 4.2. Κατά το παρελθόν είχε γίνει γνωστό ότι το 75% των Ελλήνων πάει διακοπές 10 μέρες, 15% των Ελλήνων περισσότερο από 10 μέρες, 5% από 5-9 μέρες, 4% από 1-4 μέρες και 1% δεν πάει διακοπές. Όμως, τα τελευταία χρόνια, λόγω των οικονομικών δυσκολιών, ίσως αυτό να μην ισχύει. Σε επίπεδο σημαντικότητας 5% να ελέγξετε την παραπάνω υπόθεση, αν σε ένα τυχαίο δείγμα 200 ατόμων, 165 άτομα πήγαν διακοπές 10 μέρες, 25 άτομα πήγαν διακοπές περισσότερες από 10 μέρες, 6 άτομα πήγαν διακοπές από 5-9 μέρες, 3 άτομα πήγαν διακοπές από 1-4 μέρες και 1 άτομο δεν πήγε διακοπές.

Λύση Παραδείγματος 4.2. Στην ουσία, έχουμε ένα τυχαίο πείραμα από $n = 200$ ανεξάρτητες δοκιμές στο οποίο μπορούμε να έχουμε $k = 5$ το πλήθος δυνατά αποτελέσματα, ξένα μεταξύ τους. Έστω E_i, $i = 1,...,5$, το ενδεχόμενο κάποιος Έλληνας να πάει διακοπές 10 μέρες, περισσότερες από 10 μέρες, 5-9 μέρες, 0-4 μέρες και να μην πάει διακοπές, αντίστοιχα, με $p_1 = P(E_1)$, $p_2 = P(E_2)$, $p_3 = P(E_3)$, $p_4 = P(E_4)$ και $p_5 = P(E_5)$ οι αντίστοιχες πιθανότητες πραγματοποίησης καθενός εκ των ενδεχομένων E_i, $i = 1,...,5$. Επιπλέον, έστω N_1,N_2,N_3,N_4,N_5 οι τυχαίες μεταβλητές που παριστάνουν το πλήθος μεταξύ των n μελών του δείγματος που ανήκουν σε καθεμία από τις πέντε ομάδες. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση ότι:

$$H_0 : (N_1,...,N_5) \sim \mathcal{M}(n, 0.75, 0.15, 0.05, 0.04, 0.01)$$

έναντι της εναλλακτικής ότι τουλάχιστον μία αναλογία διαφέρει από τη δοθείσα αναλογία στην H_0.

Ο έλεγχος γίνεται με τη στατιστική συνάρτηση $X^2 = \sum_{i=1}^{5} \frac{(n_i - e_i)^2}{e_i}$, όπου n_i και e_i είναι ο παρατηρούμενος και αναμενόμενος (υπό τη μηδενική υπόθεση) αριθμός εμφανίσεων του E_i, $i = 1,...,5$. Είναι $e_i = np_{i0}$, $i = 1,...,5$, με $(p_{10},...,p_{50}) = (0.75, 0.15, 0.05, 0.04, 0.01)$, δηλαδή

$$e_1 = 200 \cdot 0.75 = 150, e_2 = 200 \cdot 0.15 = 30, e_3 = 200 \cdot 0.05 = 10,$$

και

$$e_4 = 200 \cdot 0.04 = 8 \text{ και } e_5 = 200 \cdot 0.01 = 2 < 5.$$

Οι προϋποθέσεις εφαρμογής του ελέγχου χι-τετράγωνο καλής προσαρμογής δεν πληρούνται, καθώς $e_5 = 2 < 5$. Για τον λόγο αυτόν και θέλοντας να επιλύσουμε το παραπάνω πρόβλημα, θα προχωρήσουμε σε

συγχώνευση δύο γειτονικών κατηγοριών και, συγκεκριμένα, των δύο τελευταίων, δηλαδή θα ασχοληθούμε με τον έλεγχο της υπόθεσης:

$$H_0' : (N_1, ..., N_4) \sim M(n, 0.75, 0.15, 0.05, 0.05)$$

έναντι της εναλλακτικής ότι τουλάχιστον μια αναλογία διαφέρει από τη δοθείσα αναλογία στην H_0'. Πλέον, η τελευταία κατηγορία αφορά τα άτομα που έκαναν διακοπές το πολύ 4 ημέρες, όπου η απάντηση *0 ημέρες* σημαίνει ότι το άτομο δεν πήγε διακοπές.

Ο έλεγχος γίνεται με τη στατιστική συνάρτηση $X^2 = \sum_{i=1}^{4} \frac{(n_i' - e_i')^2}{e_i}$, όπου n_i' και e_i' ο παρατηρούμενος και αναμενόμενος (υπό την H_0') αριθμός εμφανίσεων του E_i', $i = 1, ..., 4$, με $E_i' = E_i$, για $i = 1, 2, 3$ και $E_4' = E_4 \cup E_5$. Υπό τη μηδενική υπόθεση H_0' ισχύει ότι $e_i' = np_{i0}'$, $i = 1, ..., 4$, με $p_{10}' = 0.75$, $p_{20}' = 0.15$, $p_{30}' = 0.05$ και $p_{40}' = 0.05$. Επομένως, είναι $e_1' = 200 \cdot 0.75 = 150$, $e_2' = 200 \cdot 0.15 = 30$, $e_3' = 200 \cdot 0.05 = 10$ και $e_4' = 200 \cdot 0.05 = 10$. Οπότε, καθώς το μέγεθος του δείγματος n είναι μεγάλο και $e_i' \geqslant 5$, $i = 1, ..., 4$, συνεχίζουμε με τον υπολογισμό της τιμής του X^2. Είναι

$$X^2 = \sum_{i=1}^{4} \frac{(n_i' - e_i')^2}{e_i'} = \frac{(165 - 150)^2}{150} + \frac{(30 - 25)^2}{30} + \frac{(6 - 10)^2}{10} + \frac{(4 - 10)^2}{10}$$

άρα

$$X^2 = 1.5 + 0.83 + 1.6 + 3.6 = 7.53.$$

Απορρίπτουμε την H_0', αν $X^2 \geq \chi^2_{k-1,a} = \chi^2_{3,0.05} = 7.815$. Επομένως, δεν απορρίπτεται η μηδενική υπόθεση, δηλαδή δεν μπορούμε να απορρίψουμε την υπόθεση ότι οι πιθανότητες πραγματοποίησης κάθε ενδεχομένου δεν διαφέρουν στατιστικά σημαντικά από αυτές που δίνονται στην H_0'. □

Παράδειγμα 4.3. Εκλέγεται ένα τυχαίο δείγμα τιμών της τ.μ. X με τα παρακάτω αποτελέσματα:

Διάστημα	(0,1/4]	(1/4,1/2]	(1/2,3/4]	(3/4,1]
Συχνότητα	30	30	10	10

Να ελέγξετε, με επίπεδο σημαντικότητας 5%, αν προέρχεται από την $f_0(x) = 2(1 - x), x \in [0,1]$.

Λύση Παραδείγματος 4.3. Θέλουμε να ελέγξουμε αν η τυχαία μεταβλητή X ακολουθεί κατανομή με συνάρτηση πυκνότητας πιθανότητας:

$$f_0(x) = 2(1 - x), x \in [0,1]$$

ή, ισοδύναμα, με αθροιστική συνάρτηση κατανομής:

$$F_0(x) = \begin{cases} 0 & x < 0 \\ x(2 - x) & 0 \leqslant x < 1 \\ 1 & \text{αλλού} \end{cases}$$

Επιπλέον, το σύνολο των τιμών της τυχαίας μεταβλητής χωρίζεται σε 4 κατηγορίες, ξένες μεταξύ τους. Έστω E_1 το ενδεχόμενο να ανήκουν στο διάστημα (0,1/4], E_2 το ενδεχόμενο να ανήκουν στο διάστημα (1/4,1/2], E_3 το ενδεχόμενο να ανήκουν στο διάστημα (1/2,3/4] και E_4 να ανήκουν στο διάστημα (3/4,1]. Ακολουθεί η εύρεση των πιθανοτήτων $p_{i0} = P(X \in E_i | X \sim F_0)$, $i = 1, ..., 4$. Προφανώς, ο υπολογισμός του γίνεται υπό την υπόθεση ότι η H_0 είναι αληθής.

Για την εύρεση των πιθανοτήτων $p_{i0} = P(X \in E_i | X \sim F_0), i = 1,...,4$, δεν έχουμε παρά να χρησιμοποιήσουμε τον τύπο για την α.σ.κ. $F_0(x)$, καθώς και τη σχέση:

$$P(\alpha < X \leqslant \beta) = F_0(\beta) - F_0(\alpha).$$

Άρα, με αντικατάσταση, έχουμε ότι:

$$p_{10} = P\left(0 < X \leqslant \frac{1}{4}\right) = F_0(1/4) - F_0(0) = \frac{1}{4}\left(2 - \frac{1}{4}\right) = \frac{7}{16}$$

$$p_{20} = P\left(\frac{1}{4} < X \leqslant \frac{1}{2}\right) = F_0(1/2) - F_0(1/4) = \frac{1}{2}\left(2 - \frac{1}{2}\right) - \frac{1}{4}\left(2 - \frac{1}{4}\right) = \frac{5}{16}$$

$$p_{30} = P\left(\frac{1}{2} < X \leqslant \frac{3}{4}\right) = F_0(3/4) - F_0(1/2) = \frac{3}{4}\left(2 - \frac{3}{4}\right) - \frac{1}{2}\left(2 - \frac{1}{2}\right) = \frac{3}{16}$$

$$p_{40} = P\left(\frac{3}{4} < X \leqslant 1\right) = F_0(1) - F_0(3/4) = 1 - \frac{3}{4}\left(2 - \frac{3}{4}\right) = \frac{1}{16}.$$

Ο έλεγχος της μηδενικής υπόθεσης γίνεται με τη στατιστική συνάρτηση $X^2 = \sum_{i=1}^{4} \frac{(n_i - e_i)^2}{e_i}$, όπου e_i ο αναμενόμενος αριθμός εμφανίσεων του E_i, $i = 1,...,4$, όταν ισχύει η μηδενική υπόθεση και n_i ο αντίστοιχος παρατηρούμενος αριθμός. Είναι $n_1 = 30$, $n_2 = 30$, $n_3 = 10$ και $n_4 = 10$, ενώ $e_1 = np_{10} = 80 \cdot \frac{7}{16} = 35$, $e_2 = np_{20} = 80 \cdot \frac{5}{16} = 25$, $e_3 = np_{30} = 80 \cdot \frac{3}{16} = 15$ και $e_4 = np_{40} = 80 \cdot \frac{1}{16} = 5$, αντίστοιχα.

Επομένως,

$$X^2 = \frac{(30-35)^2}{35} + \frac{(30-25)^2}{25} + \frac{(10-15)^2}{15} + \frac{(10-5)^2}{5} = 8.381.$$

Η μηδενική υπόθεση απορρίπτεται αν $X^2 \geq \chi^2_{k-1,a} = \chi^2_{3,0.05} = 7.815$. Αφού $8.381 > 7.815$, η μηδενική υπόθεση απορρίπτεται και εξάγεται το συμπέρασμα ότι το τ.δ. δεν προέρχεται από τη δοθείσα κατανομή. □

Σύνθετος έλεγχος χι-τετράγωνο καλής προσαρμογής

Η περίπτωση μίας πλήρους ορισμένης υποθετικής κατανομής είναι σχεδόν σπάνια στις πρακτικές εφαρμογές. Πολύ συχνά αντιμετωπίζουμε περιπτώσεις, όπου η υποθετική κατανομή περιέχει έναν αριθμό άγνωστων παραμέτρων. Στη γενικότερη περίπτωση ενός τέτοιου ελέγχου έχουμε s (με $s < k - 1$) απροσδιόριστες παραμέτρους $\theta_1, \theta_2, ..., \theta_s$, τις οποίες μπορούμε να συμβολίσουμε συνολικά με το διάνυσμα θ. Δηλαδή σε αυτήν την περίπτωση έχουμε ένα τυχαίο δείγμα $X_1, X_2, ..., X_n$ από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(x)$, η οποία είναι άγνωστη, και θέλουμε να ελέγξουμε τη μηδενική υπόθεση $H_0 : F(x) = F_0(x; \theta)$, για κάθε $x \in \mathbb{R}$ και για κάποιο $\theta \in \Theta$, όπου $F_0(x; \theta)$ είναι μία ειδική αθροιστική συνάρτηση κατανομής, η οποία εξαρτάται από ένα διάνυσμα άγνωστων παραμέτρων θ, το οποίο λαμβάνει τιμές σε έναν παραμετρικό χώρο Θ. Όπως και στην περίπτωση του απλού ελέγχου καλής προσαρμογής, χωρίζουμε πάλι το σύνολο τιμών της τυχαίας μεταβλητής σε k το πλήθος κατηγορίες, όμως τώρα οι πιθανότητες $P(X \in I_i | H_0) = p_{i0}(\theta)$ δεν είναι άμεσα υπολογίσιμες, καθώς είναι συναρτήσεις του διανύσματος των άγνωστων παραμέτρων θ. Σε αυτήν τη συνήθη περίπτωση, θα πρέπει να εκτιμήσουμε τις άγνωστες παραμέτρους, θα πρέπει δηλαδή να εκτιμήσουμε το θ με κάποιο διάνυσμα εκτιμητών, έστω $\hat{\theta}$. Επομένως, για τον έλεγχο της σύνθετης υπόθεσης θα χρησιμοποιήσουμε τη στατιστική συνάρτηση:

$$X^2(\hat{\theta}) = \sum_{i=1}^{k} \frac{(n_i - np_{i0}(\hat{\theta}))^2}{np_{i0}(\hat{\theta})}. \tag{4.6}$$

Ο Fisher (1924) ήταν ο πρώτος που παρατήρησε ότι η ασυμπτωτική κατανομή της στατιστικής συνάρτησης $X^2(\hat{\theta})$, υπό τη μηδενική υπόθεση, δεν είναι απαραίτητο να ακολουθεί χι-τετράγωνο κατανομή με $k-1$ βαθμούς ελευθερίας και ότι διαφορετικές μέθοδοι εκτίμησης των άγνωστων παραμέτρων αντανακλούν στις ιδιότητες της δειγματικής κατανομής του $X^2(\hat{\theta})$. Επίσης, επιχειρηματολόγησε ότι η κατάλληλη μέθοδος εκτίμησης είναι η εκτίμηση με τη μέθοδο της μέγιστης πιθανοφάνειας που στηρίζεται στον παρατηρούμενο αριθμό n_i κάθε ομάδας. Τότε προκύπτει ότι έχουμε να επιλύσουμε το σύστημα των εξισώσεων (βλ. Moore, 1986):

$$\sum_{i=1}^{k} \frac{n_i}{p_{i0}(\theta)} \frac{\partial p_{i0}(\theta)}{\partial \theta_j} = 0, \text{ όπου } j = 1,2,...,s. \tag{4.7}$$

Για τις ειδικές περιπτώσεις της πολυωνυμικής κατανομής και της κατανομής Poisson ισχύει ότι ο εκτιμητής μέγιστης πιθανοφάνειας, χρησιμοποιώντας τον παρατηρούμενο αριθμό των ομάδων, είναι το ίδιο αποδοτικός (efficient) με τον κλασικό εκτιμητή μέγιστης πιθανοφάνειας. Επιπρόσθετα, παρατήρησε ότι ένας ασυμπτωτικά ισοδύναμος εκτιμητής του παραπάνω μπορεί να προκύψει προσδιορίζοντας τις τιμές των παραμέτρων έτσι ώστε το X^2 της σχέσης (4.6) να γίνεται όσο το δυνατό μικρότερο. Αυτή είναι η γνωστή στη στατιστική βιβλιογραφία ως minimum chi-square method of estimation (μέθοδος εκτίμησης του ελαχίστου X^2, βλ. Vuong and Wang, 1993). Τότε, αξιοποιώντας την πρώτη από τις δύο ισοδύναμες εκφράσεις που δόθηκαν στη σχέση (4.3), προκύπτει ότι έχουμε να επιλύσουμε το σύστημα των εξισώσεων:

$$\sum_{i=1}^{k} \left(\frac{n_i}{p_{i0}(\theta)} \right)^2 \frac{\partial p_{i0}(\theta)}{\partial \theta_j} = 0, \text{ όπου } j = 1,2,...,s. \tag{4.8}$$

Προκύπτει, λοιπόν, ότι οι εκτιμητές των άγνωστων παραμέτρων $\theta_1, \theta_2, ..., \theta_s$ αποκτιούνται ως συναρτήσεις των n_i, $i = 1,2,...,s$. Ακόμα και σε απλές περιπτώσεις το σύστημα (4.8) είναι συνήθως πολύ δύσκολο να επιλυθεί, έτσι η εύρεση των εκτιμητών είναι δύσκολη. Η οριακή κατανομή του $X^2(\hat{\theta})$ για αυτήν τη μέθοδο εκτίμησης προσδιορίστηκε από τους Fisher (1924), και τους Neyman and Pearson (1928). Η παραπάνω στατιστική συνάρτηση, αν το μέγεθος του δείγματος n είναι μεγάλο και η μηδενική υπόθεση είναι αληθής, ασυμπτωτικά κατανέμεται ως χ^2_{k-1-s}, δηλαδή ακολουθεί μία χ^2 κατανομή με $k-1-s$ βαθμούς ελευθερίας. Έτσι, απορρίπτουμε τη μηδενική υπόθεση, σε ε.σ. a, αν και μόνο αν $X^2 \geqslant \chi^2_{k-1-s,a}$.

Συνοψίζοντας, προκύπτει ότι, αν έχουμε τον σύνθετο έλεγχο καλής προσαρμογής και χρησιμοποιηθεί η μέθοδος εκτίμησης του ελάχιστου X^2, τότε η κατανομή, υπό τη μηδενική υπόθεση, της στατιστικής συνάρτησης $X^2(\hat{\theta})$ είναι χι-τετράγωνο με $k-1-s$ βαθμούς ελευθερίας. Το ίδιο ισχύει και για την περίπτωση του σύνθετου ελέγχου καλής προσαρμογής στις ειδικές περιπτώσεις της πολυωνυμικής και της Poisson κατανομής, ακόμα και αν χρησιμοποιηθεί ο εκτιμητής μέγιστης πιθανοφάνειας. Απάντηση στο εύλογο ερώτημα για το τι ισχύει για το γενικό πρόβλημα του σύνθετου ελέγχου καλής προσαρμογής στην περίπτωση που χρησιμοποιούνται οι εκτιμητές μέγιστης πιθανοφάνειας δίνεται στην παρατήρηση που ακολουθεί. Για περισσότερες λεπτομέρειες παραπέμπουμε τον/την ενδιαφερόμενο/μενη αναγνώστη/στρια στον Moore (1986) και στις εκεί αναφορές.

Παρατήρηση 4.3. Στην παραπάνω στατιστική συνάρτηση και στη γενική περίπτωση οι άγνωστες πληθυσμιακές παράμετροι θ_i, $i = 1,2,...,s$, εκτιμήθηκαν χρησιμοποιώντας τον παρατηρούμενο αριθμό των παρατηρήσεων που ανήκουν στο i-οστό υποδιάστημα, n_i, αντί των $X_1, \ldots, X_n$. Η μέθοδος αυτή μπορεί να χρησιμοποιηθεί όταν τα δεδομένα καταγράφονται σε ομάδες. Στην περίπτωση, όμως, που είναι διαθέσιμες οι παρατηρήσεις, τότε κάποιος μπορεί να εκτιμήσει τις παραμέτρους θ_i, $i = 1,2,...,s$, πιο αποτελεσματικά, χρησιμοποιώντας π.χ. τους εκτιμητές μέγιστης πιθανοφάνειας που βασίζονται στις παρατηρήσεις $X_1, X_2 \ldots, X_n$. Υπενθυμίζεται ότι, αν $X_1, \ldots, X_n$ αποτελούν ένα τυχαίο δείγμα από έναν πληθυσμό με σ.π. ή σ.π.π. $f(x;\theta)$, όπου θ η άγνωστη παράμετρος ή οι άγνωστες παράμετροι, ο Ε.Μ.Π. της παραμέτρου θ είναι η ποσότητα εκείνη που μεγιστοποιεί ως προς τις τιμές των άγνωστων παραμέτρων τη συνάρτηση πιθανοφάνειας L που δίνεται από τη σχέση $L = \prod_{i=1}^{n} f(x_i;\theta)$.

Όταν χρησιμοποιούνται οι εκτιμητές μέγιστης πιθανοφάνειας, η οριακή κατανομή του X^2 στη σχέση (4.6) παύει να είναι χ^2 με $k-1-s$ βαθμούς ελευθερίας. Οι Chernoff and Lehmann (1954), θεωρώντας τα άκρα των διαστημάτων γνωστά, έδειξαν ότι, όταν s το πλήθος παράμετροι εκτιμώνται από το δείγμα, τότε η οριακή κατανομή του $X^2(\hat{\theta})$ της σχέσης (4.6) φράσσεται μεταξύ μίας χ^2 με $k-1-s$ βαθμούς ελευθερίας και μίας χ^2 με $k-1$ βαθμούς ελευθερίας. Για μεγάλες τιμές του k και ταυτόχρονα μικρές τιμές του s υπάρχει μικρή διαφορά μεταξύ των $\chi^2_{k-1-s,a}$ και $\chi^2_{k-1,a}$ και μπορεί να αγνοηθεί, αλλά για μικρό k η επίδραση της χρήσης της χ^2_{k-1-s} κατανομής για έλεγχους μπορεί να οδηγήσει σε σοβαρά σφάλματα.

Παράδειγμα 4.4. Οι 150 απόγονοι μιας ορισμένης διασταύρωσης, ταξινομημένοι σε 4 ομάδες με βάση ένα συγκεκριμένο χαρακτηριστικό είναι 35 στην πρώτη, 40 στη δεύτερη, 40 στην τρίτη και 35 στην τέταρτη. Σύμφωνα με ένα γενετικό πρότυπο, οι πιθανότητες συμμετοχής σε καθεμία από τις παραπάνω 4 ομάδες είναι:

$$p(1-p)\ ,\ \ p^2\ ,\ \ (1-p)^2\ ,\ \ p(1-p)\ ,$$

αντίστοιχα, όπου $p \in (0,1)$. Να ελεγχθεί αν τα πειραματικά δεδομένα συμφωνούν με αυτό το θεωρητικό πρότυπο σε ε.σ. 10%.

Λύση Παραδείγματος 4.4. Στην ουσία έχουμε ένα τυχαίο πείραμα από $n = 150$ ανεξάρτητες δοκιμές, στο οποίο τυχαίο πείραμα μπορούμε να έχουμε $k = 4$ το πλήθος δυνατά αποτελέσματα, ξένα μεταξύ τους. Έστω E_i, $i = 1,...,4$, το ενδεχόμενο να ανήκει κάποιος απόγονος σε συγκεκριμένη ομάδα. Έστω $p_1 = P(E_1)$, $p_2 = P(E_2)$, $p_3 = P(E_3)$ και $p_4 = P(E_4)$ οι αντίστοιχες πιθανότητες πραγματοποίησης καθενός εκ των τεσσάρων ενδεχομένων. Επιπλέον, έστω N_1, N_2, N_3, N_4 είναι οι τυχαίες μεταβλητές που παριστάνουν το πλήθος των απογόνων που ανήκουν σε καθεμία από τις τέσσερις ομάδες, στα n το πλήθος τυχαία επιλεγμένα άτομα. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση ότι:

$$H_0 : (N_1,...,N_4) \sim \mathcal{M}(n, p_{10}, p_{20}, p_{30}, p_{40})$$

όπου $p_{10} = p(1-p)$, $p_{20} = p^2$, $p_{30} = (1-p)^2$, $p_{40} = p(1-p)$, με $p \in (0,1)$, έναντι της εναλλακτικής ότι τουλάχιστον μία αναλογία διαφέρει από τη δοθείσα αναλογία στην H_0. Επομένως, πρόκειται για έναν σύνθετο έλεγχο καλής προσαρμογής καθώς έχουμε μία άγνωστη παράμετρο (την $\theta = p$), η οποία θα πρέπει να εκτιμηθεί. Η παράμετρος αυτή θα εκτιμηθεί με τη μέθοδο της μέγιστης πιθανοφάνειας. Δηλαδή θέλουμε να μεγιστοποιήσουμε ως προς p τη συνάρτηση πιθανοφάνειας

$$\begin{aligned} L &= \frac{n}{n_1!n_2!n_3!n_4!} p_{10}^{n_1} p_{20}^{n_2} p_{30}^{n_3} p_{40}^{n_4} \\ &= \frac{n}{n_1!n_2!n_3!n_4!} (p(1-p))^{n_1} (p^2)^{n_2} ((1-p)^2)^{n_3} (p(1-p))^{n_4} \end{aligned}$$

ή, ισοδύναμα, θέλουμε να μεγιστοποιήσουμε ως προς p τον λογάριθμο της συνάρτησης πιθανοφάνειας

$$\begin{aligned} \log(L) &= \log\left(\frac{n}{n_1!n_2!n_3!n_4!}\right) + n_1 \cdot \log((p(1-p))) + 2n_2 \cdot \log(p) \\ &+ 2n_3 \cdot \log((1-p)) + n_4 \cdot \log((p(1-p)). \end{aligned}$$

Προκύπτει (επιβεβαιώστε το!) ότι ο εκτιμητής μέγιστης πιθανοφάνειας είναι $\widehat{p} = 1/2$. Θα χρησιμοποιήσουμε τη στατιστική συνάρτηση

$$X^2 = \sum_{i=1}^{4} \frac{\left(n_i - np_{i0}\left(\widehat{p}\right)\right)^2}{np_{i0}\left(\widehat{p}\right)},$$

με κρίσιμη περιοχή $X^2 \geqslant \chi^2_{k-1-s,a} = X^2_{4-1-1,0.1} = 4.605$, όπου $p_{i0}(\widehat{p}) = 0.25$, $i = 1,...,4$. Είναι τότε:

$$e_1 = e_2 = e_3 = e_4 = 150 \cdot 0.25 = 37.5.$$

Άρα, καθώς το μέγεθος του δείγματος n είναι μεγάλο και $e_i \geqslant 5,\ i = 1, ..., 4,$ δηλαδή πληρούνται οι προϋποθέσεις του X^2 ελέγχου καλής προσαρμογής, προβαίνουμε στον υπολογισμό του. Επομένως είναι:

$$\begin{aligned} X^2 &= \frac{(35-37.5)^2}{37.5} + \frac{(40-37.5)^2}{37.5} + \frac{(40-37.5)^2}{37.5} + \frac{(35-37.5)^2}{37.5} \\ &= 0.1666667 + 0.15625 + 0.15625 + 0.1666667 = 0.6458334 \end{aligned}$$

και σε επίπεδο σημαντικότητας 10% η μηδενική υπόθεση δεν απορρίπτεται, καθώς 0.6458334<4.605. □

Παράδειγμα 4.5. (X^2 τεστ καλής προσαρμογής της κατανομής Poisson με άγνωστη παράμετρο). Έστω X η τυχαία μεταβλητή που παριστάνει τον αριθμό των αφίξεων σε ένα κεντρικό φαρμακείο της πόλης των Ιωαννίνων κατά τη διάρκεια ενός μισάωρου. Επιλέγεται τυχαία ένα δείγμα 365 τέτοιων χρονικών περιόδων κατά τη διάρκεια των οποίων έχουμε 730 αφίξεις συνολικά. Ο αριθμός των αφίξεων και οι αντίστοιχες συχνότητες κατά τη διάρκεια αυτών των 365 χρονικών περιόδων δίνονται στον πίνακα που ακολουθεί:

Αριθμός Αφίξεων	0	1	2	3	4	5 ή περισσότερες
Συχνότητα	41	75	120	110	10	9

Να ελεγχθεί με επίπεδο σημαντικότητας 5% η υπόθεση ότι η τυχαία μεταβλητή X ακολουθεί κατανομή Poisson.

Λύση Παραδείγματος 4.5. Θέλουμε να ελέγξουμε την υπόθεση ότι η τυχαία μεταβλητή X, που παριστάνει τον αριθμό των αφίξεων στο φαρμακείο σε ένα μισάωρο, ακολουθεί κατανομή Poisson με συνάρτηση πιθανότητας:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2,...$$

όπου η παράμετρος λ είναι άγνωστη και παριστάνει τον ρυθμό των αφίξεων των πελατών σε χρονικό διάστημα ενός μισάωρου. Η παράμετρος αυτή θα εκτιμηθεί με τη μέθοδο της μέγιστης πιθανοφάνειας, δηλαδή θέλουμε να μεγιστοποιήσουμε ως προς λ τη συνάρτηση πιθανοφάνειας

$$L = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-\lambda n}\lambda^{\sum_{i=1}^n X_i}}{\prod_{i=1}^n x_i!}.$$

Εύκολα προκύπτει ότι $\widehat{\lambda} = \frac{\sum_{i=1}^n X_i}{n}$ και με βάση τις δειγματικές τιμές που δίνονται

$$\widehat{\lambda} = \frac{\sum_{i=1}^n X_i}{n} = \frac{0\cdot 41 + 1\cdot 75 + 2\cdot 120 + 3\cdot 110 + 4\cdot 10 + 5\cdot 9}{365} = \frac{730}{365} = 2.$$

Σημειώνουμε ότι στην εκτίμηση της παραμέτρου λ, έχουμε υποθέσει ότι οι 9 διαθέσιμες τιμές της τελευταίας κατηγορίας είναι όλες ίσες με το 5. Επιπλέον, από τα δεδομένα της εκφώνησης προκύπτει ότι το σύνολο των τιμών της τυχαίας μεταβλητής χωρίζεται σε 6 κατηγορίες, ξένες μεταξύ τους. Έστω E_1 το ενδεχόμενο να μην αφιχθούν πελάτες, E_2 το ενδεχόμενο να αφιχθεί 1 πελάτης κ.ο.κ., έστω E_6 το ενδεχόμενο να αφιχθούν 5 ή περισσότεροι πελάτες, κατά τη διάρκεια ενός μισάωρου.

Θα χρησιμοποιήσουμε τη στατιστική συνάρτηση

$$X^2 = \sum_{i=1}^{6} \frac{\left(n_i - np_{i0}\left(\widehat{\lambda}\right)\right)^2}{np_{i0}\left(\widehat{\lambda}\right)},$$

με κρίσιμη περιοχή $X^2 \geqslant \chi^2_{k-1-s,a} = X^2_{6-1-1,0.05} = 9.488$, όπου $p_{i0}(\hat{\lambda}) = P(E_i|X \sim \mathscr{P}(\hat{\lambda}))$, $i = 1,...,6$. Είναι

$$p_{10}(\hat{\lambda}) = P(X = 0|X \sim \mathscr{P}(2)) = 0.1353,\ p_{20}(\hat{\lambda}) = P(X = 1|X \sim \mathscr{P}(2)) = 0.2707,$$
$$p_{30}(\hat{\lambda}) = P(X = 2|X \sim \mathscr{P}(2)) = 0.2707,\ p_{40}(\hat{\lambda}) = P(X = 3|X \sim \mathscr{P}(2)) = 0.1804,$$
$$p_{50}(\hat{\lambda}) = P(X = 4|X \sim \mathscr{P}(2)) = 0.0902,\ p_{60}(\hat{\lambda}) = 1 - P(X \leqslant 4|X \sim \mathscr{P}(2)) = 0.0527,$$

Είναι τότε:

$$e_1 = 365 \cdot 0.1353 = 49.3845, e_2 = 365 \cdot 0.2707 = 98.8055, e_3 = 365 \cdot 0.2707 = 98.8055,$$
$$e_4 = 365 \cdot 0.1804 = 65.846, e_5 = 365 \cdot 0.0902 = 32.923 \text{ και } e_6 = 365 \cdot 0.0527 = 19.2355.$$

Άρα, καθώς το μέγεθος του δείγματος n είναι μεγάλο και $e_i \geqslant 5$, $i = 1,...,6$, δηλαδή πληρούνται οι προϋποθέσεις του X^2 ελέγχου καλής προσαρμογής, προβαίνουμε στον υπολογισμό του. Επομένως, είναι:

$$\begin{aligned} X^2 &= \frac{(41-49.3845)^2}{49.3845} + \frac{(75-98.8055)^2}{98.8055} + \frac{(120-98.8055)^2}{98.8055} \\ &+ \frac{(110-65.846)^2}{65.846} + \frac{(10-32.923)^2}{32.923} + \frac{(9-19.199)^2}{19.199} \\ &= 1.4235 + 5.7355 + 4.5464 + 29.6081 + 15.9603 + 5.4465 = 62.7204, \end{aligned}$$

και σε επίπεδο σημαντικότητας 5% η μηδενική υπόθεση απορρίπτεται, καθώς 67.7749>9.488. Αυτό σημαίνει ότι η τυχαία μεταβλητή που παριστάνει τον αριθμό των αφίξεων στο φαρμακείο σε ένα μισάωρο δεν περιγράφεται από την κατανομή Poisson. □

Παρατήρηση 4.4. Στα προηγούμενα δύο παραδείγματα, που αφορούν τον σύνθετο έλεγχο καλής προσαρμογής της πολυωνυμικής κατανομής και της Poisson αντίστοιχα, η εκτίμηση με τη μέθοδο της μέγιστης πιθανοφάνειας είναι ασυμπτωτικά ισοδύναμη με αυτήν της μεθόδου εκτίμησης X^2 και για αυτόν το λόγο δεν δημιουργείται πρόβλημα με τους βαθμούς ελευθερίας.

Η χρήση του X^2 τεστ για τον (σύνθετο) έλεγχο καλής προσαρμογής γνωστών συνεχών κατανομών, όπως είναι η εκθετική ή η κανονική κατανομή, δεν προτείνεται λόγω της μικρής ισχύος του. Επιπρόσθετα, όπως αναφέρει και ο Moore (1986), δεν προτείνεται η χρήση του X^2 τεστ στις περιπτώσεις που τα δεδομένα είναι διαθέσιμα. Αυτό οφείλεται στο ότι κατά τη διακριτοποίηση των αρχικών διαθέσιμων δεδομένων έχουμε απώλεια πληροφορίας, που οδηγεί με τη σειρά της σε μικρή ισχύ. Για αυτούς τους λόγους έχει προταθεί στη βιβλιογραφία πληθώρα ελεγχών τόσο για το γενικό υπό μελέτη πρόβλημα όσο και για ειδικές περιπτώσεις κατανομών, όπως είναι η κανονική, η εκθετική και άλλες. Αντικείμενο μελέτης της επόμενης ενότητας είναι οι έλεγχοι καλής προσαρμογής για το γενικό πρόβλημα που βασίζονται στην εμπειρική συνάρτηση κατανομής.

Παρατήρηση 4.5. Για μια εφαρμογή του ελέγχου χι-τετράγωνο καλής προσαρμογής ως ελέγχου ανεξαρτησίας, όταν τα δεδομένα ταξινομούνται σε πίνακες συνάφειας παραπέμπουμε στην Ενότητα 8.6 του παρόντος συγγράμματος.

4.3 Έλεγχοι καλής προσαρμογής που στηρίζονται στην εμπειρική συνάρτηση κατανομής

Στο Κεφάλαιο 2 παρουσιάστηκε μεταξύ άλλων το Θεώρημα των Glivenko-Cantelli, σύμφωνα με το οποίο η διαφορά της εμπειρικής συνάρτησης κατανομής από την πραγματική αθροιστική συνάρτηση κατανομής γίνεται μικρότερη, καθώς αυξάνεται το μέγεθος του δείγματος. Το αποτέλεσμα αυτό και η ερμηνεία του έχουν οδηγήσει να προταθούν στη βιβλιογραφία διάφοροι έλεγχοι καλής προσαρμογής που στηρίζονται σε μέτρα εγγύτητας μεταξύ της εμπειρικής αθροιστικής συνάρτησης κατανομής και της αθροιστικής συνάρτησης κατανομής υπό τη μηδενική υπόθεση. Αντικείμενο μελέτης αυτής της ενότητας αποτελούν οι πιο δημοφιλείς τέτοιοι έλεγχοι, ενώ για άλλους που ανήκουν στην ίδια κατηγορία παραπέμπουμε στη μονογραφία των D'Agostino and Stephens (1986).

4.3.1 Ο έλεγχος Kolmogorov-Smirnov, η γενίκευσή του και οι παραλλαγές του

Μεταξύ των ελέγχων καλής προσαρμογής που έχουν παρουσιαστεί στη βιβλιογραφία και βασίζονται στην εμπειρική αθροιστική συνάρτηση κατανομής, ο πιο δημοφιλής είναι ο λεγόμενος έλεγχος των Kolmogorov-Smirnov ή, απλώς, έλεγχος του Kolmogorov. Προτού προχωρήσουμε στην παρουσίασή του, θα δώσουμε μία εξήγηση για το γεγονός ότι μπορεί να τον συναντήσει κάποιος στη βιβλιογραφία με τις δύο διαφορετικές προαναφερθείσες ονομασίες. Ο Ρώσος μαθηματικός Andrey Nikolaevich Kolmogorov (1903-1987), με πάρα πολύ σημαντική συνεισφορά στη θεωρία πιθανοτήτων, ήταν αυτός που αρχικά τον πρότεινε (βλ. Kolmogorov, 1933). Αργότερα, ο επίσης Ρώσος μαθηματικός, με πολύ σημαντική συμβολή στη θεωρία πιθανοτήτων και στη στατιστική, Nikolai Smirnov (1900-1966), μεταξύ άλλων, επέκτεινε τη στατιστική συνάρτηση που προτάθηκε από τον Kolmogorov για τη σύγκριση της ισότητας των αθροιστικών συναρτήσεων κατανομής δύο πληθυσμών, έχοντας δύο ανεξάρτητα δείγματα, ένα δείγμα από καθέναν από αυτούς (βλ. Smirnov, 1939a,b). Στη συνέχεια αυτής της ενότητας, αρχικά, το ενδιαφέρον επικεντρώνεται στον απλό έλεγχο καλής προσαρμογής με ένα δείγμα.

Kolmogorov-Smirnov απλός έλεγχος καλής προσαρμογής

Ειδικότερα, αν $X_1, X_2, \ldots, X_n$, είναι ένα τυχαίο δείγμα από έναν πληθυσμό με συνεχή, αλλά άγνωστη αθροιστική συνάρτηση κατανομής $F(x)$, για τον έλεγχο της $H_0 : F(x) = F_0(x), \forall x \in \mathbb{R}$, έναντι της εναλλακτικής $H_1 : F(x) \neq F_0(x)$, για κάποιο $x \in \mathbb{R}$, όπου $F_0(x)$ είναι μία πλήρως καθορισμένη, χωρίς άγνωστες παραμέτρους, ειδική (συγκεκριμένη) αθροιστική συνάρτηση κατανομής, ο Kolmogorov (1933) πρότεινε να χρησιμοποιείται η στατιστική συνάρτηση

$$D_n = \sup_{x \in \mathbb{R}} \left| F_n(x) - F_0(x) \right|, \tag{4.9}$$

η οποία, στην ουσία, μετρά πόσο αποκλίνει η ε.α.σ.κ. $F_n(\cdot)$ από την α.σ.κ. $F_0(\cdot)$ και αναζητά τη μέγιστη κατακόρυφη απόσταση μεταξύ των γραφημάτων των $F_n(x)$ και $F_0(x)$. Δηλαδή, η στατιστική συνάρτηση D_n αποτελεί στην πραγματικότητα ένα μέτρο της εγγύτητας των $F_n(x)$ και $F_0(x)$. Από τον ορισμό της στατιστικής συνάρτησης ελέγχου D_n γίνεται άμεσα αντιληπτό ότι απορρίπτεται η μηδενική υπόθεση για μεγάλες τιμές αυτής. Για τον υπολογισμό της στατιστικής συνάρτησης D_n πολύ χρήσιμη είναι η πρόταση που ακολουθεί (βλ., μεταξύ άλλων, Gibbons and Chakraborti, 2020).

Πρόταση 4.2. Έστω $X_1, X_2, \ldots, X_n$, ένα τυχαίο δείγμα από έναν πληθυσμό με συνεχή, άλλα άγνωστη αθροιστική συνάρτηση κατανομής $F(x)$. Η στατιστική συνάρτηση

$$D_n = \sup_{x \in \mathbb{R}} \left| F_n(x) - F_0(x) \right|,$$

για τον απλό έλεγχο καλής προσαρμογής των Kolmogorov-Smirnov, μπορεί να γραφτεί ως $D_n = \max\{D_n^+, D_n^-\}$, με

$$D_n^+ = \sup_{x \in \mathbb{R}} (F_n(x) - F_0(x)) = \max\left\{ \max_{1 \leqslant i \leqslant n} (F_n(X_{(i)}) - F_0(X_{(i)})), 0 \right\} \tag{4.10}$$

και

$$D_n^- = \sup_{x \in \mathbb{R}} (F_0(x) - F_n(x)) = \max\left\{ \max_{1 \leqslant i \leqslant n} (F_0(X_{(i)}) - F_n(X_{(i-1)})), 0 \right\} \tag{4.11}$$

με $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ να είναι οι διατεταγμένες τιμές του δείγματος. Επομένως,

$$D_n = \max\left\{ \max_{1 \leqslant i \leqslant n} (F_n(X_{(i)}) - F_0(X_{(i)})), \max_{1 \leqslant i \leqslant n} (F_0(X_{(i)}) - F_n(X_{(i-1)})), 0 \right\}. \tag{4.12}$$

Απόδειξη Πρότασης 4.2. Αρχικά θεωρούμε την ακόλουθη διαμέριση του $\mathbb{R}$, $\mathbb{R} = \bigcup_{i=0}^{n} \Delta_i$, όπου $\Delta_i = [X_{(i)}, X_{(i+1)})$, $i = 0,1,\dots,n$, με $X_{(0)} = -\infty$ και $X_{(n+1)} = +\infty$. Τότε ισχύει ότι:

$$F_n(x) = F_n(X_i), X_i \leqslant x < X_{i+1}, \text{ για } i = 0,1,\dots,n.$$

Δηλαδή σε κάθε διάστημα Δ_i η εμπειρική συνάρτηση κατανομής υπολογίζεται ως

$$F_n(x) = \begin{cases} 0 & , \quad x < X_{(1)} \\ \frac{1}{n} & , \quad X_{(1)} \leqslant x < X_{(2)} \\ \vdots & \quad \vdots \\ \frac{i}{n} & , \quad X_{(i)} \leqslant x < X_{(i+1)} \\ \vdots & \quad \vdots \\ 1 & , \quad x > X_{(n)}. \end{cases} \tag{4.13}$$

Οπότε, για να υπολογίσουμε το supremum της $|F_n(x) - F_0(x)|$, $\forall\, x \in \mathbb{R}$, αρκεί να το υπολογίσουμε σε κάθε διάστημα Δ_i ξεχωριστά και κατόπιν να πάρουμε το μέγιστο αυτών των supremum, δηλαδή,

$$D_n = \sup_{x\in\mathbb{R}} \left|F_n(x) - F_0(x)\right| = \max_{0\leqslant i\leqslant n} \sup_{x\in\Delta_i} \left|F_n(x) - F_0(x)\right|.$$

Αρχικά, υπολογίζουμε το $D_n^+ = \sup_{x\in\mathbb{R}}(F_n(x) - F_0(x)) = \max_{0\leqslant i\leqslant n} \sup_{x\in\Delta_i}(F_n(x) - F_0(x))$.

Σε κάθε διάστημα $\Delta_i = [X_{(i)}, X_{(i+1)})$, είναι προφανές από τη σχέση (4.13), ότι η $F_n(x)$ είναι σταθερά και, επιπλέον, $F_n(x) = F_n(X_{(i)}) = \frac{i}{n}, x \in \Delta_i, i = 0,1,\dots,n$. Επομένως, για να υπολογίσουμε το $\sup_{x\in\Delta_i}(F_n(x) - F_0(x))$, αρκεί να υπολογίσουμε το $\frac{i}{n} - \inf_{x\in\Delta_i} F_0(x)$. Όμως η F_0, ως συνάρτηση κατανομής, είναι αύξουσα, οπότε $F_0(x) \geqslant F_0(X_{(i)})$, για οποιοδήποτε $x \in \Delta_i$, και τελικά

$$\sup_{x\in\Delta_i}(F_n(x) - F_0(x)) = \frac{i}{n} - F_0(X_{(i)}).$$

Επομένως,

$$D_n^+ = \max_{0\leqslant i\leqslant n}\left\{\frac{i}{n} - F_0(X_{(i)})\right\}$$

το οποίο, ισοδύναμα, γράφεται (βλ., μεταξύ άλλων, Gibbons and Chakraborti, 2020)

$$D_n^+ = \max\left\{\max_{1\leqslant i\leqslant n}\left\{\frac{i}{n} - F_0(X_{(i)})\right\}, 0\right\},$$

καθώς το max για $i = 0$ είναι το μηδέν.

Με παρόμοιο τρόπο υπολογίζουμε το $D_n^- = \sup_{x\in\mathbb{R}}(F_0(x) - F_n(x)) = \max_{0\leqslant i\leqslant n} \sup_{x\in\Delta_i}(F_0(x) - F_n(x))$.

Σε κάθε διάστημα $\Delta_i = [X_{(i)}, X_{(i+1)})$, είναι προφανές από τη σχέση (4.13), ότι η $F_n(x)$ είναι σταθερά και, επιπλέον, $F_n(x) = F_n(X_{(i)}) = \frac{i}{n}, x \in \Delta_i, i = 0,1,\dots,n$. Επομένως, για να υπολογίσουμε το $\sup_{x\in\Delta_i}(F_0(x) - F_n(x))$, αρκεί να υπολογίσουμε το $\sup_{x\in\Delta_i}\left(F_0(x) - \frac{i}{n}\right)$. Όμως η F_0, ως συνάρτηση κατανομής, είναι αύξουσα, οπότε $F_0(x) \leqslant F_0(X_{(i+1)})$, για οποιοδήποτε $x \in \Delta_i$, και τελικά

$$\sup_{x\in\Delta_i}(F_0(x) - F_n(x)) = F_0(X_{(i+1)}) - \frac{i}{n}.$$

Επομένως,

$$D_n^- = \max_{0 \leqslant i \leqslant n} \left\{ F_0(X_{(i+1)}) - \frac{i}{n} \right\}$$

το οποίο, ισοδύναμα, γράφεται (βλ., μεταξύ άλλων, Gibbons and Chakraborti, 2020)

$$D_n^- = \max \left\{ \max_{1 \leqslant i \leqslant n} \left\{ F_0(X_{(i)}) - \frac{i-1}{n} \right\}, 0 \right\},$$

καθώς αρχικά θέτουμε $i+1=j$ και έχουμε ότι:

$$D_n^- = \max_{1 \leqslant j \leqslant n+1} \left\{ F_0(X_{(j)}) - \frac{j-1}{n} \right\}$$

και η ισοδύναμη σχέση προκύπτει, καθώς για $j=n$ το max είναι μηδέν. □

Δύο εύλογα ερωτήματα που μπορεί να έχουν προκύψει είναι ποιες τιμές της στατιστικής συνάρτησης D_n μπορούν να θεωρηθούν μεγάλες, έτσι ώστε να απορρίπτεται η μηδενική υπόθεση, ενώ είναι αληθής με επίπεδο σημαντικότητας a, και αν ο προσδιορισμός αυτών των μεγάλων τιμών εξαρτάται από την προς έλεγχο αθροιστική συνάρτηση κατανομής F_0 (γεγονός που θα αποτελούσε προφανώς μειονέκτημα του ελέγχου). Στο ακόλουθο θεώρημα αποδεικνύεται ότι η κατανομή, υπό τη μηδενική υπόθεση, της στατιστικής συνάρτησης D_n είναι ίδια για όλες τις $F_0(\cdot)$, αλλά διαφορετική για διαφορετικά μεγέθη δείγματος n.

Θεώρημα 4.1

Έστω $X_1, X_2, \ldots, X_n$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό με συνεχή αλλά άγνωστη αθροιστική συνάρτηση κατανομής $F(x)$ και έστω $F_n(\cdot)$ η εμπειρική αθροιστική συνάρτηση κατανομής. Υποθέτουμε ότι η μηδενική υπόθεση $H_0 : F(x) = F_0(x)$ για κάθε $x \in \mathbb{R}$ είναι αληθής. Τότε η κατανομή της στατιστικής συνάρτησης D_n, που ορίστηκε στη σχέση (4.9), είναι ανεξάρτητη της F_0.

Απόδειξη Θεωρήματος 4.1. Για λόγους απλότητας η απόδειξη δίνεται για F_0 γνησίως αύξουσα. Ισχύει ότι

$$P\left(\sup_{x \in \mathbb{R}} \left| F_n(x) - F_0(x) \right| \leq d \right) \overset{\text{θέτω } Y = F_0(X)}{=} P\left(\sup_{y \in (0,1)} \left| F_n\left(F_0^{-1}(y)\right) - y \right| \leq d \right).$$

Από τον ορισμό της εμπειρικής αθροιστικής συνάρτησης κατανομής προκύπτει ότι

$$F_n\left(F_0^{-1}(y)\right) = \frac{\sum_{i=1}^n I_{(-\infty, F_0^{-1}(y)]}(X_i)}{n} = \frac{\sum_{i=1}^n I_{(-\infty, y]}(F_0(X_i))}{n}.$$

Επομένως,

$$P\left(\sup_{y \in (0,1)} \left| F_n\left(F_0^{-1}(y)\right) - y \right| \leq d \right) = P\left(\sup_{y \in (0,1)} \left| \frac{\sum_{i=1}^n I_{(-\infty, y]}(F_0(X_i))}{n} - y \right| \leq d \right).$$

Λαμβάνοντας υπόψη ότι η κατανομή της αθροιστικής συνάρτησης κατανομής είναι ομοιόμορφη στο διάστημα (0,1), καθώς

$$P(F_0(X) \leq t) = P\left(X \leq F_0^{-1}(t)\right) = F_0\left(F_0^{-1}(t)\right) = t, t \in (0,1),$$

και θέτοντας $U_i = F_0(X_i) \sim \mathcal{U}(0,1)$, $i = 1, \ldots, n$, έχουμε ότι:

$$P\left(\sup_{y \in (0,1)} \left| \frac{\sum_{i=1}^n I_{(-\infty, y]}(F_0(X_i))}{n} - y \right| \leq d \right) = P\left(\sup_{y \in (0,1)} \left| \frac{\sum_{i=1}^n I_{(-\infty, y]}(U_i)}{n} - y \right| \leq d \right)$$

που δεν εξαρτάται από την F_0, αλλά προφανώς εξαρτάται από το μέγεθος του δείγματος n. □

Παρατήρηση 4.6. Στο ίδιο συμπέρασμα με αυτό που διατυπώθηκε στο Θεώρημα 4.1 καταλήγουμε χρησιμοποιώντας την Πρόταση 4.2 από όπου έχουμε ουσιαστικά ότι:

$$D_n^+ = \max\left\{\max_{1\leqslant i\leqslant n}\left(i/n - U_{(i)}\right), 0\right\},$$

και

$$D_n^- = \max\left\{\max_{1\leqslant i\leqslant n}\left(U_{(i)} - (i-1)/n\right), 0\right\}$$

όπου $U_{(i)} = F_0\left(X_{(i)}\right)$, με $U_{(1)},\ldots,U_{(n)}$ να είναι, όταν η F_0 είναι συνεχής, διατεταγμένο δείγμα από την ομοιόμορφη κατανομή στο (0,1).

Από το Θεώρημα 4.1 εξάγουμε το συμπέρασμα ότι η κατανομή της στατιστικής συνάρτησης D_n υπό την H_0 για οποιαδήποτε F_0 είναι ίδια. Η ακριβής κατανομή της στατιστικής συνάρτησης D_n υπό την H_0 είναι πολύπλοκη, παρότι έχουν δοθεί στη βιβλιογραφία αναδρομικές (recursive) σχέσεις για τον υπολογισμό πιθανοτήτων της μορφής $P(D_n \leqslant k/n)$ για ακέραιες τιμές του k (βλ. Kolmogorov, 1933). Τα αποτελέσματα αυτά έχουν χρησιμοποιηθεί για να δοθούν πίνακες της συνάρτησης κατανομής της στατιστικής συνάρτησης για $n \leq 100$ και $k = 1,\ldots,15$ (βλ., για παράδειγμα, Birnbaum, 1952).

Παρατήρηση 4.7. Όπως αναφέρουν οι Zhang and Wu (2002), καθώς η ακριβής κατανομή της στατιστικής συνάρτησης D_n είναι διαθέσιμη μόνο για συγκεκριμένες και περιορισμένες τιμές του πηλίκου k/n, αυτή δεν μπορεί να χρησιμοποιηθεί στην πράξη. Λύσεις που έχουν προταθεί είναι είτε να χρησιμοποιηθεί η ασυμπτωτική κατανομή της στατιστικής συνάρτησης είτε μια προσέγγιση της κατανομής ή να προσδιοριστούν οι κρίσιμες τιμές μέσω παρεμβολής. Για περισσότερες πληροφορίες παραπέμπουμε τον/την ενδιαφερόμενο/μενη αναγνώστη/στρια στην εργασία των Zhang and Wu (2002) και τις εκεί αναφορές.

Παράδειγμα 4.6. Να προσδιοριστεί η ακριβής κατανομή, υπό τη μηδενική υπόθεση, της στατιστικής συνάρτησης D_n για τον δίπλευρο έλεγχο, όταν $n = 1$.

Λύση Παραδείγματος 4.6. Η στατιστική συνάρτηση όταν $n = 1$ δίνεται από τη σχέση:

$$D_1 = \left|1 - F_0(x)\right| = 1 - F_0(x),$$

με τιμές μεταξύ μηδέν και ένα. Είναι τότε:

$$P(D_1 \leq t) = P(1 - F_0(x) \leq t) = P(F_0(x) \geq 1 - t) = P\left(X \geq F_0^{-1}(1-t)\right),$$

οπότε

$$P(D_1 \leq t) = 1 - P\left(X \leq F_0^{-1}(1-t)\right) = 1 - F_0\left(F_0^{-1}(1-t)\right) = 1 - (1-t) = t, 0 \leqslant t \leqslant 1.$$

Επομένως, η στατιστική συνάρτηση D_1 ακολουθεί ομοιόμορφη κατανομή στο (0,1). □

Ο Kolmogorov (1933) απέδειξε ότι η οριακή κατανομή του $K_n = \sqrt{n}D_n$ υπό τη μηδενική υπόθεση είναι απαλλαγμένη παραμέτρων (distribution-free) και προσδιορίζεται στο θεώρημα που ακολουθεί.

Θεώρημα 4.2

Έστω ότι $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $F_X(\cdot)$ και $F_n(\cdot)$ η αντίστοιχη εμπειρική αθροιστική συνάρτηση κατανομής. Όταν $n \to +\infty$,

τότε, υπό τη μηδενική υπόθεση $H_0 : F_X(x) = F_0(x)$, $\forall\, x \in \mathbb{R}$, ισχύει ότι:

$$P\left(\sqrt{n}D_n \leqslant x\right) \to H(x), \text{ για κάθε } x \geqslant 0,$$

όπου:

$$H(x) = [1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2x^2}]I_{(0,\infty)}(x).$$

Απόδειξη Θεωρήματος 4.2. Η απόδειξη παραλείπεται, αφού ξεφεύγει από τους σκοπούς του παρόντος συγγράμματος (βλ. Kolmogorov, 1933). □

Γνωρίζοντας ότι η $H_0 : F(x) = F_0(x)$, $\forall\, x \in \mathbb{R}$, έναντι της εναλλακτικής $H_1 : F(x) \neq F_0(x)$, για κάποιο $x \in \mathbb{R}$, απορρίπτεται για μεγάλες τιμές της στατιστικής συνάρτησης D_n έχουμε ότι, για να έχουμε ασυμπτωτικά επίπεδο σημαντικότητας a, θα πρέπει να προσδιορίσουμε την τιμή έστω c_a που είναι τέτοια, ώστε $P\left(\sqrt{n}D_n \geqslant c_a \middle| H_0\right) \approx a$. Επομένως, χρησιμοποιώντας το Θεώρημα 4.2 έχουμε ότι η τιμή c_a θα πρέπει να είναι τέτοια ώστε $1 - H(c_a) \approx a$. Ο Πίνακας Π.11 του Παραρτήματος (βλ., μεταξύ άλλων, Miller, 1956) προσδιορίζει τις τιμές $c_a/\sqrt{n}$ για διάφορα μεγέθη δείγματος και επίπεδα σημαντικότητας. Παρατηρήστε ότι, καθώς δίνεται η τιμή $c_a/\sqrt{n}$, απορρίπτεται η μηδενική υπόθεση του απλού δίπλευρου ελέγχου καλής προσαρμογής, αν η τιμή της στατιστικής συνάρτησης D_n είναι μεγαλύτερη από την τιμή αυτή του Πίνακα Π.11.

Παρατήρηση 4.8. Σε αντίθεση με το κριτήριο χι-τετράγωνο το οποίο βασίζεται μόνο στις συχνότητες των κατηγοριών (ομάδων ή κλάσεων) που τοποθετούνται οι παρατηρήσεις του δείγματος, το κριτήριο Kolmogorov-Smirnov χρησιμοποιεί τις πληροφορίες που παρέχονται από την εμπειρική συνάρτηση κατανομής. Το κριτήριο Kolmogorov-Smirnov, όταν έχουμε διαθέσιμα τα δεδομένα από μια συνεχή τυχαία μεταβλητή, είναι προτιμότερο από το κριτήριο χι-τετράγωνο, γιατί δεν χάνουμε πληροφορία που περιέχεται στις τιμές του δείγματος λόγω διακριτοποίησης. Από την άλλη μεριά, το κριτήριο Kolmogorov-Smirnov, όταν χρησιμοποιείται σε διακριτές κατανομές, καθίσταται συντηρητικό. Για αυτούς τους λόγους έχει γενικά επικρατήσει να χρησιμοποιείται το κριτήριο χι-τετράγωνο για διακριτές και το κριτήριο Kolmogorov-Smirnov για συνεχείς κατανομές. Τέλος, ο στατιστικός έλεγχος των Kolmogorov-Smirnov εφαρμόζεται σε κατανομές με συνεχή αθροιστική κατανομή και τείνει να είναι πιο ευαίσθητος στο κέντρο της κατανομής από ότι στις ουρές (βλ. Kvam and Vidakovic, 2007).

Παράδειγμα 4.7. Έστω οι ακόλουθες 5 δειγματικές τιμές: 0.1, 0.3, 0.24, 0.58, 0.68. Να ελέγξετε, με επίπεδο σημαντικότητας 5%, αν τα δεδομένα προέρχονται από την $\mathcal{U}(0,1)$.

Λύση Παραδείγματος 4.7. Έχουμε ένα τυχαίο δείγμα μεγέθους 5 και θέλουμε να ελέγξουμε την υπόθεση $H_0 : F(x) = F_0(x)$, $\forall\, x \in \mathbb{R}$, έναντι της εναλλακτικής $H_1 : F(x) \neq F_0(x)$, για κάποιο $x \in \mathbb{R}$, όπου $F_0(x)$ η αθροιστική συνάρτηση κατανομής της $\mathcal{U}(0,1)$. Επομένως,

$$F_0(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \leq x < 1, \\ 1, & \text{αλλού.} \end{cases}$$

Θέλοντας να υπολογίσουμε τη στατιστική συνάρτηση $\sqrt{n}D_n$ αρχικά διατάσσουμε τις παρατηρήσεις σε αύξουσα τάξη μεγέθους και υπολογίζουμε τις ποσότητες $F_n\left(X_{(i)}\right)$, $F_n\left(X_{(i)}-\right)$ και $F_0\left(X_{(i)}\right)$. Για διευκόλυνση στους υπολογισμούς δημιουργούμε τον πίνακα που ακολουθεί:

$X_{(i)}$	0.1	0.24	0.3	0.58	0.68
$F_n\left(X_{(i)}\right) = \frac{i}{n}$	0.2	0.4	0.6	0.8	1
$F_n\left(X_{(i)}-\right) = \frac{i-1}{n}$	0	0.2	0.4	0.6	0.8
$F_0\left(X_{(i)}\right)$	0.1	0.24	0.3	0.58	0.68
$F_n\left(X_{(i)}\right) - F_0\left(X_{(i)}\right)$	0.1	0.16	0.3	0.22	0.32
$F_0(X_{(i)}) - F_n(X_{(i)}-)$	0.1	0.04	-0.1	-0.02	-0.12

Επομένως, προκύπτει ότι $D_n = 0.32$. Επιπλέον, καθώς ο Πίνακας Π.11 του Παραρτήματος μας δίνει τιμές του $c_a/\sqrt{n}$, συγκρίνουμε την τιμή του $D_n = 0.32$ με την τιμή του πίνακα για $a = 0.05$, δηλαδή με την τιμή 0.563. Συνεπώς, καθώς η τιμή της στατιστικής συνάρτησης δεν είναι μεγαλύτερη ή ίση από την κρίσιμη τιμή, εξάγεται το συμπέρασμα ότι δεν απορρίπτεται η μηδενική υπόθεση. Άρα δεν υπάρχουν σαφείς ενδείξεις για να απορρίψουμε την υπόθεση ότι τα δεδομένα μπορούν να θεωρηθούν ότι προέρχονται από έναν πληθυσμό που περιγράφεται ικανοποιητικά από την $\mathscr{U}(0,1)$. □

Παράδειγμα 4.8. Έστω ένα τυχαίο δείγμα $n = 20$ το πλήθος παρατηρήσεων από κάποιον συνεχή πληθυσμό. Θέλουμε να ελέγξουμε με επίπεδο σημαντικότητας 5% κατά πόσο τα δεδομένα αυτά προέρχονται από κανονική κατανομή με μέση τιμή 0 και διασπορά 1. Οι διατεταγμένες παρατηρήσεις του δείγματος δίνονται στον επόμενο πίνακα. Υπόδειξη: για διευκόλυνση στους υπολογισμούς χρησιμοποιήστε την R.

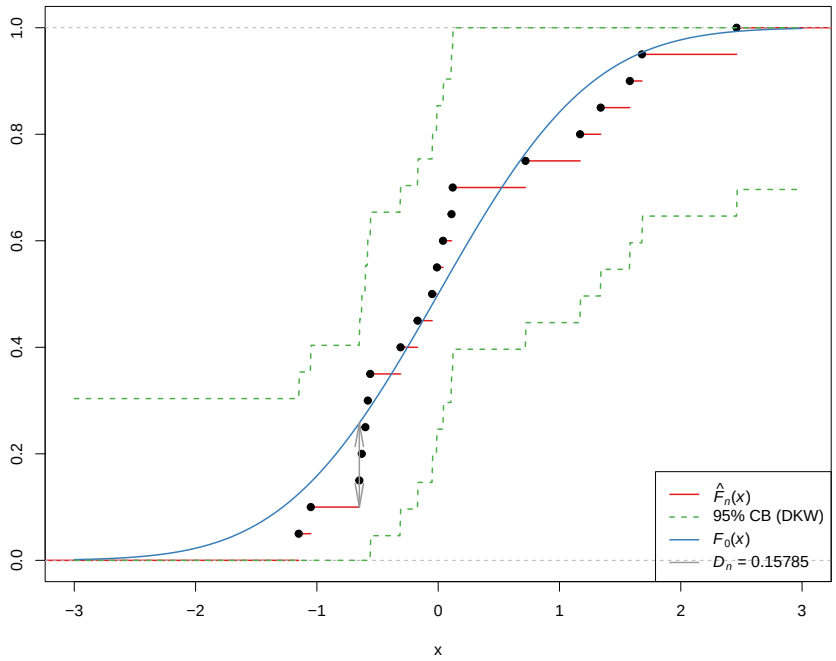
i	1	2	3	4	5	6	7	8	9	10
$X_{(i)}$	-1.15	-1.05	-0.65	-0.63	-0.60	-0.58	-0.56	-0.31	-0.17	-0.05
i	11	12	13	14	15	16	17	18	19	20
$X_{(i)}$	-0.01	0.04	0.11	0.12	0.72	1.17	1.34	1.58	1.68	2.46

Λύση Παραδείγματος 4.8. Για την επίλυση του παραδείγματος και τους απαραίτητους υπολογισμούς θα χρησιμοποιήσουμε τις παρακάτω εντολές της R.

```
1  x1<-c
       (-1.15,-1.05,-0.65,-0.63,-0.60,-0.58,-0.56,-0.31,-0.17,-0.05,-0.01,
        0.04,0.11,0.12,0.72,1.17,1.34,1.58,1.68,2.46)
2  foxi<-pnorm(sort(x1))
3  n<-length(x1)
4  i1<-1:n
5  Dnplus<-(i1/n)-foxi
6  Dnminus<-foxi-(i1-1)/n
7  Dn<-max(max(Dnplus),max(Dnminus))
8  Dn
```

Σημειώνουμε ότι η εντολή pnorm που χρησιμοποιήθηκε εφαρμόζει την αθροιστική συνάρτηση κατανομής της τυπικής κανονικής στις τιμές που δηλώθηκαν (γραμμή 2 στον κώδικα). Από τα παραπάνω προκύπτει ότι $D_{20} = 0.157846$ (γραμμή 7 στον κώδικα) και από τον Πίνακα Π.11 έχουμε ότι η κρίσιμη τιμή του ελέγχου είναι 0.294. Επομένως, η τιμή της στατιστικής συνάρτησης D_{20} δεν είναι μεγαλύτερη από την κρίσιμη τιμή και άρα δεν απορρίπτεται η υπόθεση ότι τα δεδομένα προέρχονται από την τυπική κανονική $\mathscr{N}(0,1)$ κατανομή.

Το Σχήμα 4.1 απεικονίζει την πραγματοποίηση της τιμής D_n μεταξύ της εμπειρικής συνάρτησης κατανομής και της συνάρτησης κατανομής της $\mathscr{N}(0,1)$. Η μέγιστη απόσταση μεταξύ της $F_n(x)$ και $F_0(x)$ επιτυγχάνεται αριστερά της τρίτης διατεταγμένης παρατήρησης -0.65 ($x \uparrow -0.65$) και είναι ίση με $D_n = 0.15785$, η οποία απεικονίζεται με το γκρι βελάκι του Σχήματος 4.1. Σημειώστε ότι ένας εναλλακτικός τρόπος ελέγχου της H_0 σε επίπεδο σημαντικότητας 5% είναι να εξακριβώσουμε αν η 95% ζώνη εμπιστοσύνης DKW περιέχει εξ ολοκλήρου το γράφημα της $F_0(x)$.



Σχήμα 4.1: Παράδειγμα 4.7 Εμπειρική Συνάρτηση Κατανομής (---), 95% ζώνη εμπιστοσύνης DKW(---), αθροιστική συνάρτηση κατανομής της $\mathcal{N}(0,1)$ --- και απόσταση D_n ---.

Αξίζει, επίσης, να αναφέρουμε πως το παράδειγμα θα μπορούσε να λυθεί χρησιμοποιώντας για τους υπολογισμούς είτε μόνο τον Πίνακα Π.1 του Παραρτήματος ή ακόμη μόνο τη συνάρτηση `pnorm` της `R` και δημιουργώντας για διευκόλυνση τον παρακάτω πίνακα. Από τον πίνακα αυτόν προκύπτει ότι $\max_{1\leqslant i\leqslant n}(F_n(X_{(i)}) - \Phi(X_{(i)})) = 0.15224$, ενώ $\max_{1\leqslant i\leqslant n}(\Phi(X_{(i)}) - F_n(X_{(i)}-) = 0.15785$, άρα $D_{20} = 0.15785$ (που συμφωνεί με το προηγούμενο αποτέλεσμα καθώς χρησιμοποιήθηκαν στρογγυλοποιήσεις).

i	1	2	3	4	5	6	7	8
$X_{(i)}$	-1.15	-1.05	-0.65	-0.63	-0.60	-0.58	-0.56	-0.31
$F_n(X_{(i)})$	1/20	2/20	3/20	4/20	5/20	6/20	7/20	8/20
$F_n(X_{(i)}-)$	0	1/20	2/20	3/20	4/20	5/20	6/20	7/20
$\Phi(X_{(i)})$	0.12507	0.14686	0.25785	0.26435	0.27425	0.28096	0.28774	0.37829
$F_n(X_{(i)}) - \Phi(X_{(i)})$	-0.075072	-0.04686	-0.10785	-0.06435	0.02425	0.01904	0.06226	0.02172
$\Phi(X_{(i)}) - F_n(X_{(i)}-)$	0.12507	0.09686	0.15785	0.11435	0.07425	0.03096	-0.01226	0.028280
i	9	10	11	12	13	14	15	16
$X_{(i)}$	-0.17	-0.05	-0.01	0.04	0.11	0.12	0.72	1.17
$F_n(X_{(i)})$	9/20	10/20	11/20	12/20	13/20	14/20	15/20	16/20
$F_n(X_{(i)}-)$	8/20	9/20	10/20	11/20	12/20	13/20	14/20	15/20
$\Phi(X_{(i)})$	0.43251	0.48006	0.49601	0.51595	0.54380	0.54776	0.76115	0.87900
$F_n(X_{(i)}) - \Phi(X_{(i)})$	0.01749	0.01994	0.05399	0.08405	0.10620	0.15224	-0.01424	-0.07900
$\Phi(X_{(i)}) - F_n(X_{(i)}-)$	0.032505	0.03006	-0.00399	-0.034047	-0.05620	-0.10224	0.06424	0.12900
i	17	18	19	20				
$X_{(i)}$	1.34	1.58	1.68	2.46				
$F_n(X_{(i)})$	17/20	18/20	19/20	20/20				
$F_n(X_{(i)}-)$	16/20	17/20	18/20	19/20				
$\Phi(X_{(i)})$	0.90988	0.94295	0.95352	0.99305				
$F_n(X_{(i)}) - \Phi(X_{(i)})$	-0.05988	-0.04295	-0.00352	0.00695				
$\Phi(X_{(i)}) - F_n(X_{(i)}-)$	0.109888	0.09295	0.05352	0.04305				

□

Μέχρι τώρα το ενδιαφέρον επικεντρώθηκε στον δίπλευρο έλεγχο της μηδενικής υπόθεσης $H_0 : F(x) = F_0(x)$, $\forall\, x \in \mathbb{R}$, έναντι της εναλλακτικής $H_1 : F(x) \neq F_0(x)$, για κάποιο $x \in \mathbb{R}$, όπου $F_0(x)$ είναι μία, γνωστή, ειδική αθροιστική συνάρτηση κατανομής, με πλήρως προσδιορισμένες παραμέτρους. Με παρόμοιο τρόπο μπορεί να ελεγχθεί η μηδενική υπόθεση $H_0 : F(x) = F_0(x)$, $\forall\, x \in \mathbb{R}$, έναντι της εναλλακτικής

- $H_{1+} : F(x) \geqslant F_0(x)$, για κάποιο $x \in \mathbb{R}$, ή της
- $H_{1-} : F(x) \leqslant F_0(x)$, για κάποιο $x \in \mathbb{R}$.

Ειδικότερα στην πρώτη περίπτωση χρησιμοποιείται η στατιστική συνάρτηση D_n^+ που παριστάνει τη μέγιστη απόκλιση μεταξύ της εμπειρικής αθροιστικής συνάρτησης κατανομής $F_n(x)$ και της $F_0(x)$ πάνω στις τιμές του x για τις οποίες η $F_n(x)$ είναι πάνω από την $F_0(x)$. Από την άλλη μεριά, στη δεύτερη περίπτωση χρησιμοποιείται η στατιστική συνάρτηση D_n^-, η οποία παριστάνει τη μέγιστη απόκλιση μεταξύ της $F_n(x)$ και της $F_0(x)$ πάνω στις τιμές του x για τις οποίες η $F_n(x)$ είναι κάτω από την $F_0(x)$. Επομένως, είναι προφανές ότι απορρίπτονται οι παραπάνω μηδενικές υποθέσεις έναντι των αντίστοιχων εναλλακτικών για μεγάλες τιμές των στατιστικών συναρτήσεων D_n^+ και D_n^-, αντίστοιχα. Προκύπτει, λοιπόν, ότι απορρίπτονται σε επίπεδο σημαντικότητας a οι μηδενικές υποθέσεις, αν και μόνο αν $\sqrt{n}D_n^+ \geqslant c_{a/2}$ ή $\sqrt{n}D_n^- \geqslant c_{a/2}$, αντίστοιχα, όπου c_a τέτοιο, ώστε $P(\sqrt{n}D_n \geq c_a) \approx a$ και τιμές του $c_a/\sqrt{n}$ δίνονται στον Πίνακα Π.11 του Παραρτήματος.

Παρατήρηση 4.9. Αποδεικνύεται ότι οι κατανομές των D_n^+ και D_n^- υπό τη μηδενική υπόθεση ταυτίζονται (η απόδειξη αφήνεται ως άσκηση).

Παρατήρηση 4.10. Για την αποφυγή εκτεταμένων πινάκων κρίσιμων τιμών των κριτηρίων D_n, D_n^+ και D_n^- για τα διάφορα επίπεδα σημαντικότητας και μεγέθη δείγματος, ο Stephens (1970) τροποποίησε τα κριτήρια αυτά, έτσι ώστε να γίνεται απευθείας σύγκριση για την απόρριψη ή όχι της H_0. Τα τροποποιημένα κριτήρια βασίζονται στις στατιστικές συναρτήσεις:

$$D_n^{*+} = D_n^+ \cdot (\sqrt{n} + 0.12 + 0.11/\sqrt{n}),$$

$$D_n^{*-} = D_n^- \cdot (\sqrt{n} + 0.12 + 0.11/\sqrt{n}),$$

και

$$D_n^* = D_n \cdot (\sqrt{n} + 0.12 + 0.11/\sqrt{n}).$$

Σε αυτό το πλαίσιο, απορρίπτεται, για δοθέν επίπεδο σημαντικότητας a, η μηδενική υπόθεση έναντι της εναλλακτικής, όταν η τιμή του κριτηρίου D_n^{*+}, D_n^{*-} και D_n^* υπερβαίνει την αντίστοιχη τιμή που δίνεται στον Πίνακα 4.1.

Πίνακας 4.1: Κρίσιμες τιμές των τροποποιημένων κριτηρίων των Kolmogorov-Smirnov για τον δίπλευρο έλεγχο.

Κριτήριο	Επίπεδο σημαντικότητας				
	15%	10%	5%	2.5%	1%
D_n^*	1.138	1.224	1.358	1.480	1.628
D_n^{*+}	0.973	1.073	1.224	1.358	1.518
D_n^{*-}	0.973	1.073	1.224	1.358	1.518

Kolmogorov-Smirnov σύνθετος έλεγχος καλής προσαρμογής

Ο Kolmogorov-Smirnov έλεγχος καλής προσαρμογής που παρουσιάστηκε προηγουμένως υποθέτει ότι η μηδενική υπόθεση είναι πλήρως καθορισμένη, δηλαδή ελέγχουμε μία απλή υπόθεση. Με άλλα λόγια, θεωρεί ότι η $F_0(x)$ είναι πλήρως ορισμένη χωρίς την «παρουσία» άγνωστων παραμέτρων. Μια τέτοια υπόθεση τις περισσότερες φορές είναι μη ρεαλιστική και εύλογα κάποιος μπορεί να αναρωτηθεί κατά πόσο ένας τέτοιος έλεγχος μπορεί να επεκταθεί σε περιπτώσεις ελέγχου σύνθετης υπόθεσης, δηλαδή της υπόθεσης $H_0 : F(x) = F_0(x;\theta)$, για κάθε $x \in \mathbb{R}$ και για κάποια τιμή του διανύσματος των άγνωστων παραμέτρων θ που ανήκει στον παραμετρικό χώρο Θ. Για τέτοιες υποθέσεις η $\sup_{x\in\mathbb{R}} \left|F_n(x) - F_0(x;\theta)\right|$ παύει να είναι στατιστική συνάρτηση, αφού εξαρτάται από τις άγνωστες παραμέτρους. Ένας προφανής τρόπος να απαλλαγούμε από τις άγνωστες παραμέτρους είναι να αντικαταστήσουμε τις άγνωστες παραμέτρους με τους εκτιμητές τους, $\hat{\theta}$. Δυστυχώς, όμως, η στατιστική συνάρτηση που προκύπτει, δηλαδή η:

$$D_n\left(\hat{\theta}\right) = \sup_{x\in\mathbb{R}} \left|F_n(x) - F_0(x;\hat{\theta})\right| \tag{4.14}$$

δεν έχει την κατανομή του D_n και κλειστή μορφή για την κατανομή της στατιστικής συνάρτησης $D_n\left(\hat{\theta}\right)$ είναι άγνωστη. Επιπρόσθετα, επισημαίνουμε ότι η μέθοδος εκτίμησης που χρησιμοποιείται (μέγιστης πιθανοφάνειας, ροπών ή κάποια άλλη) επιδρά στην απόδοση του ελέγχου. Για περισσότερες πληροφορίες παραπέμπουμε τον/την ενδιαφερόμενο/μενη αναγνώστη/στρια, μεταξύ άλλων, στην εργασία των Weber *et al.* (2006).

Ένας τρόπος για να ξεπεραστεί το παραπάνω πρόβλημα είναι να χρησιμοποιηθούν τεχνικές Monte Carlo για την εκτίμηση της p-τιμής του ελέγχου. Τα βήματα της μεθόδου για την περίπτωση του δίπλευρου ελέγχου καλής προσαρμογής είναι αυτά που ακολουθούν.

1. Από τις διαθέσιμες παρατηρήσεις εκτιμούμε το διάνυσμα των άγνωστων παραμέτρων θ, το οποίο συμβολίζουμε με $\hat{\theta}$.
2. Υπολογίζουμε την τιμή της στατιστικής συνάρτησης $D_n(\hat{\theta})$.
3. Δημιουργούμε B το πλήθος (B μεγάλος αριθμός) τυχαία δείγματα μεγέθους n από την κατανομή με αθροιστική συνάρτηση κατανομής $F_0(x;\hat{\theta})$ και για καθένα από αυτά υπολογίζουμε την τιμή της στατιστικής συνάρτησης $D_n(\hat{\theta}^{(j)})$ από τη σχέση (4.14), με $\hat{\theta}^{(j)}$, $j = 1, ..., B$ τον εκτιμητή της παραμέτρου θ που προκύπτει από το $j-$οστό δείγμα.
4. Εκτιμούμε την p-τιμή του ελέγχου ως το ποσοστό των φορών που η τιμή της στατιστικής συνάρτησης $D_n(\hat{\theta}^{(j)})$ είναι μεγαλύτερη από την τιμή $D_n(\hat{\theta})$.

Η εφαρμογή των παραπάνω, με τη βοήθεια της `R`, για ένα συγκεκριμένο αριθμητικό παράδειγμα, ακολουθεί, με στόχο να γίνει κατανοητή η παραπάνω διαδικασία.

Παράδειγμα 4.9. Έστω ότι έχουμε ένα τυχαίο δείγμα $n = 12$ το πλήθος παρατηρήσεων από κάποιον συνεχή πληθυσμό. Θέλουμε να ελέγξουμε χρησιμοποιώντας το κριτήριο Kolmogorov-Smirnov, με επίπεδο σημαντικότητας 5%, κατά πόσο τα δεδομένα αυτά προέρχονται από κανονική κατανομή. Οι διατεταγμένες παρατηρήσεις του δείγματος είναι οι: 6900, 7200, 8600, 8700, 9300, 9600, 9800, 10200, 11600, 12200, 15200, 15500. Ποια η p-τιμή που θα προέκυπτε αν αγνοηθεί ότι πρόκειται για σύνθετο έλεγχο καλής προσαρμογής;

Λύση Παραδείγματος 4.9. Καθώς θέλουμε να ελέγξουμε την υπόθεση ότι οι παρατηρήσεις αυτές προέρχονται από κανονική κατανομή, χωρίς να προσδιορίζονται η μέση τιμή και η διακύμανση αυτής, είναι προφανές ότι έχουμε έναν σύνθετο έλεγχο καλής προσαρμογής. Αρχικά, θα πρέπει να εκτιμήσουμε τις άγνωστες παραμέτρους μ και σ^2. Γνωρίζουμε ότι, αν $X_1,...,X_n$ είναι τυχαίο δείγμα από την $\mathcal{N}(\mu,\sigma^2)$, τότε

οι εκτιμητές μέγιστης πιθανοφάνειας αυτών των παραμέτρων είναι:

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \text{ και } \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n}.$$

Στη συνέχεια, υπολογίζουμε τη στατιστική συνάρτηση από τη σχέση (4.14). Αυτό επιτυγχάνεται εύκολα με τις ακόλουθες εντολές στην R.

```
x1<-c(6900, 7200, 8600, 8700, 9300, 9600, 9800, 10200, 11600, 12200,
   15200, 15500)
n<-length(x1)
hatmu<-mean(x1) # estimate mu
hatsigma2<-(n-1)/n*var(x1) # estimate sigma^2
foxi<-pnorm(sort(x1),mean=hatmu,sd=sqrt(hatsigma2))
i1<-1:n
Dnplus<-(i1/n)-foxi
Dnminus<-foxi-(i1-1)/n
Dn<-max(max(Dnplus),max(Dnminus)) # calculate value Dn
Dn
```

Από τα παραπάνω προκύπτει ότι η παρατηρούμενη τιμή της στατιστικής συνάρτησης D_n είναι ίση με 0.1966883. Έπειτα, θέλουμε να υλοποιήσουμε το τρίτο και τέταρτο βήμα που περιγράφθηκε παραπάνω για $B = 10000$ το πλήθος προσομοιωμένα δείγματα.

```
B<-10000
j<-0
Dnsim<-c()
for(i in 1:B) {
x1sim<-sort(rnorm(n, mean =hatmu, sd =sqrt(hatsigma2)))
hatmusim<-mean(x1sim)
hatsigma2sim<-(n-1)/n*var(x1sim)
foxisim<-pnorm(x1sim,mean=hatmusim,sd=sqrt(hatsigma2sim))
i1<-1:n
Dnplussim<-(i1/n)-foxisim
Dnminussim<-foxisim-(i1-1)/n
Dnsim[i]<-max(max(Dnplussim),max(Dnminussim))
if (Dnsim[i]>=Dn) {j=j+1
}
}
pvalue<-j/B
pvalue
```

Από τα παραπάνω προκύπτει ότι η $p-$τιμή του ελέγχου είναι ίση με 0.2342 και, επομένως, καθώς η p-τιμή είναι μεγαλύτερη από 0.05, σε επίπεδο σημαντικότητας 5%, δεν απορρίπτεται η υπόθεση ότι το δείγμα προέρχεται από πληθυσμό που περιγράφεται από την κανονική κατανομή.

Αξίζει να αναφέρουμε πως, αν είχαμε αγνοήσει ότι πρόκειται για σύνθετο έλεγχο και εφαρμόζαμε τον απλό έλεγχο καλής προσαρμογής Kolmogorov-Smirnov, υποθέτοντας δηλαδή ότι οι εκτιμήσεις των παραμέτρων είναι και οι πραγματικές τους τιμές, τότε, από το Θεώρημα 4.2, η p-τιμή του ελέγχου ισούται με:

$$\begin{aligned} P(D_n \geqslant 0.1966883) &= 1 - P\left(\sqrt{12} D_n \leqslant 0.6813484\right) \\ &= 1 - H(0.6813484) \approx 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 (0.6813484)^2}. \end{aligned}$$

Το παραπάνω άθροισμα, περιορίζοντάς το στους 10^8 όρους και υπολογίζοντάς το με τη βοήθεια της R, με τον τρόπο που παρατίθεται παρακάτω, προκύπτει ότι ισούται με 0.3710093. Παρατηρήστε ότι καταλήξαμε στο ίδιο συμπέρασμα με προηγουμένως (μη απόρριψη της H_0). Κάτι τέτοιο, όμως, δεν συνεπάγεται ότι είναι ορθή η εφαρμογή του απλού ελέγχου καλής προσαρμογής Kolmogorov-Smirnov έναντι του σύνθετου, στην περίπτωση που οι παράμετροι της κατανομής είναι άγνωστες.

```
b<-0
for(k in 1:10^8) {
b[k]<- (-1)^{k-1}*exp(-2*(k^2)* (0.6813484)^2)
}
sum(b)
```

□

Εκτός από τον παραπάνω τρόπο αντιμετώπισης του σύνθετου ελέγχου καλής προσαρμογής είναι διαθέσιμοι στη βιβλιογραφία πίνακες κρίσιμων τιμών για κάποιες ειδικές περιπτώσεις κατανομών. Στην πραγματικότητα, χρησιμοποιούνται διαφορετικοί πίνακες κρίσιμων τιμών, ενώ μπορεί να τροποποιείται και η στατιστική συνάρτηση. Δηλαδή, οι πίνακες αυτοί δεν είναι οι ίδιοι για όλες τις κατανομές, αλλά εξαρτώνται από την προς έλεγχο μηδενική υπόθεση. Για παράδειγμα, πίνακες κρίσιμων τιμών για τις ειδικές περιπτώσεις της κανονικής και εκθετικής κατανομής δίνονται στην εργασία του Stephens (1974). Επιπρόσθετα, για συγκεκριμένες κατανομές έχουν προταθεί και μελετηθεί παραλλαγές του ελέγχου Kolmogorov-Smirnov, οι οποίες επιτρέπουν τη χρήση του σε περιπτώσεις όπου οι παράμετροι εκτιμώνται από τα δεδομένα. Στη συνέχεια, θα παραθέσουμε αποτελέσματα που αφορούν τον σύνθετο έλεγχο καλής προσαρμογής για την κανονική και εκθετική κατανομή.

Έλεγχος καλής προσαρμογής της κανονικής κατανομής με άγνωστες παραμέτρους

Μία παραλλαγή του Kolmogorov-Smirnov τεστ για τον έλεγχο της σύνθετης υπόθεσης της κανονικότητας μελετήθηκε από τον Lilliefors (1967). Σύμφωνα με αυτήν την παραλλαγή, αν θέλουμε να ελέγξουμε αν το τ.δ. $X_1, X_2, \ldots, X_n$ προέρχεται από την κανονική κατανομή, με άγνωστη μέση τιμή μ και διασπορά σ^2, υπολογίζουμε αρχικά τις τυποποιημένες τιμές $Z_1, Z_2, \ldots, Z_n$ του τ.δ., που ορίζονται ως εξής:

$$Z_i = \frac{X_i - \bar{X}}{S'}, \tag{4.15}$$

όπου

$$\bar{X} = \sum_{i=1}^{n} X_i/n$$

και

$$S' = \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)}.$$

Τότε η αρχική μηδενική υπόθεση ανάγεται στον έλεγχο της υπόθεσης ότι το τ.δ. $Z_1, Z_2, \ldots, Z_n$ προέρχεται από την τυπική κανονική κατανομή. Η κατάλληλη στατιστική συνάρτηση, στην περίπτωση αυτήν, είναι η μέγιστη κατακόρυφη απόκλιση της εμπειρικής συνάρτησης κατανομής του τυποποιημένου δείγματος από την αθροιστική συνάρτηση κατανομής της τυπικής κανονικής κατανομής, που συμβολίζεται με $\Phi(\cdot)$. Δηλαδή, η ελεγχοσυνάρτηση του Lilliefors ορίζεται από τη σχέση $T_n = \max\{T_n^+, T_n^-\}$, όπου

$$T_n^+ = \max\left\{\max_{1 \leqslant i \leqslant n}\left(i/n - \Phi\left(Z_{(i)}\right)\right), 0\right\}, \tag{4.16}$$

και

$$T_n^- = \max\left\{\max_{1\leqslant i\leqslant n}\left(\Phi\left(Z_{(i)}\right) - (i-1)/n\right), 0\right\}, \tag{4.17}$$

με $Z_{(1)}, Z_{(2)}, \ldots, Z_{(n)}$ να είναι οι διατεταγμένες τυποποιημένες τιμές.

Προφανώς, και στην περίπτωση αυτού του έλεγχου, μεγάλες τιμές της στατιστικής συνάρτησης θα είναι εκείνες που θα συνηγορούν υπέρ της απόρριψης της μηδενικής υπόθεσης, αφού αυτές θα είναι αποτέλεσμα χαμηλού βαθμού εγγύτητας της συνάρτησης κατανομής του τυποποιημένου δείγματος προς την αθροιστική συνάρτηση κατανομής της τυπικής κανονικής κατανομής. Η αναλυτική μορφή της συνάρτησης κατανομής της ελεγχοσυνάρτησης αυτής είναι δύσκολο να προσδιοριστεί. Έτσι, ο Lilliefors μελέτησε και πινακοποίησε την ασυμπτωτική κατανομή της. Τότε, για δοθέν επίπεδο σημαντικότητας a, απορρίπτεται η μηδενική υπόθεση για τον δίπλευρο έλεγχο ότι τα δεδομένα προέρχονται από κανονική κατανομή, όταν η τιμή της στατιστικής συνάρτησης ελέγχου υπερβαίνει την αντίστοιχη τιμή που δίνεται στον Πίνακα Π.12 του Παραρτήματος.

Παράδειγμα 4.10. Χρησιμοποιώντας τον έλεγχο που προτάθηκε από τον Lilliefors ελέγξτε, με επίπεδο σημαντικότητας 5%, αν τα δεδομένα του Παραδείγματος 4.9 προέρχονται από κανονική κατανομή. Υπόδειξη: για διευκόλυνση στους υπολογισμούς χρησιμοποιήστε την R.

Λύση Παραδείγματος 4.10. Αρχικά δημιουργούμε τα τυποποιημένα δεδομένα σύμφωνα με τη σχέση (4.15) και τα διατάσσουμε σε αύξουσα τάξη μεγέθους (εδώ είναι ήδη, αλλά το παράδειγμα επιλύεται ως να μην ήταν).

```
x1<-c(6900, 7200, 8600, 8700, 9300, 9600, 9800, 10200, 11600, 12200,
   15200, 15500)
n<-length(x1)
hatmu<-mean(x1)
hatsigmatonos<-var(x1)
z1<-sort((x1-hatmu)/sqrt(hatsigmatonos)) # standardized sample
```

Έπειτα προσδιορίζουμε το $T_n = \max\{T_n^+, T_n^-\}$ μέσω του υπολογισμού των σχέσεων (4.16) και (4.17).

```
foxi<-pnorm(z1,mean=0,sd=1)
i1<-1:n
Tnplus<-(i1/n)-foxi
Tnminus<-foxi-(i1-1)/n
Tn<-max(max(Tnplus),max(Tnminus)) # calculate Tn value
Tn
```

Από τα παραπάνω προκύπτει ότι η τιμή του ελέγχου είναι ίση με 0.1954125. Καθώς από τον Πίνακα Π.12 του Παραρτήματος έχουμε ότι η κρίσιμη τιμή για $n = 12$, με επίπεδο σημαντικότητας 5%, είναι 0.242, συμπεραίνουμε, καθώς $0.1954125 < 0.242$, ότι δεν απορρίπτεται η υπόθεση ότι το δείγμα προέρχεται από πληθυσμό που περιγράφεται από την κανονική κατανομή.

Προφανώς, το παράδειγμα αυτό θα μπορούσε να επιλυθεί και χωρίς τη βοήθεια των παραπάνω εντολών. Αρχικά, υπολογίζουμε ότι $\bar{X} = \sum_{i=1}^{12} X_i/12 = 10400$ και $S' = \sqrt{\sum_{i=1}^{12}(X_i - \bar{X})^2/11} = 2773.249$. Στη συνέχεια, υπολογίζουμε τις τυποποιημένες τιμές $Z_{(i)}$, $i = 1, \ldots, 12$ (εφαρμόζοντας στρογγυλοποίηση σε δύο δεκαδικά ψηφία) και κάνουμε τις επιπλέον πράξεις. Για τις πράξεις αυτές είναι πολύ βοηθητικός ο πίνακας που ακολουθεί.

i	$Z_{(i)}$	$\Phi(Z_{(i)})$	i/n	$(i-1)/n$	$i/n - \Phi(Z_{(i)})$	$\Phi(Z_{(i)}) - (i-1)/n$
1	-1.26	0.1038347	1/12	0/12	-0.020501348	0.1038346811
2	-1.15	0.1250719	2/12	1/12	0.041594731	0.0417386023
3	-0.65	0.2578461	3/12	2/12	-0.007846111	0.0911794441
4	-0.61	0.2709309	4/12	3/12	0.062402430	0.0209309038
5	-0.40	0.3445783	5/12	4/12	0.072088408	0.0112449251
6	-0.29	0.3859081	6/12	5/12	0.114091881	-0.0307585479
7	-0.22	0.4129356	7/12	6/12	0.170397756	-0.0870644226
8	-0.07	0.4720968	8/12	7/12	0.194569837	-0.1112365035
9	0.43	0.6664022	9/12	8/12	0.083597821	-0.0002644873
10	0.65	0.7421539	10/12	9/12	0.091179444	-0.0078461108
11	1.73	0.9581849	11/12	10/12	-0.041518196	0.1248515291
12	1.84	0.9671159	12/12	11/12	0.032884119	0.0504492147

Επισημαίνεται ότι για τον υπολογισμό των $\Phi(Z_{(i)})$, για $i = 1,...,12$, χρησιμοποιήσαμε την εντολή `pnorm` της `R`. Ωστόσο, είναι προφανές, ότι θα μπορούσε, εναλλακτικά, ο υπολογισμός των $\Phi(Z_{(i)})$, για $i = 1,...,12$, να γίνει χρησιμοποιώντας τον Πίνακα Π.1. Προκύπτει ότι $T_n = \max\{T_n^+, T_n^-\} = \max\{0.194569837, 0.1248515291\}$ και, επομένως, $T_n = 0.194569837$. Οι αποκλίσεις στα αποτελέσματα οφείλονται σε στρογγυλοποιήσεις. □

Ο Stephens (1970) για την ειδική περίπτωση του σύνθετου ελέγχου καλής προσαρμογής για την κανονική κατανομή τροποποίησε το κριτήριο των Kolmogorov-Smirnov θεωρώντας τη στατιστική συνάρτηση:

$$T_n^* = T_n\left(\sqrt{n} - 0.01 + 0.85/\sqrt{n}\right),$$

όπου η $T_n = \max\{T_n^+, T_n^-\}$ με τις στατιστικές συναρτήσεις T_n^+ και T_n^- να υπολογίζονται από τις σχέσεις (4.16) και (4.20), αντίστοιχα. Τότε, για δοθέν επίπεδο σημαντικότητας a, απορρίπτεται η μηδενική υπόθεση για τον δίπλευρο έλεγχο ότι τα δεδομένα προέρχονται από κανονική κατανομή, όταν η τιμή του τροποποιημένου κριτηρίου υπερβαίνει την αντίστοιχη τιμή που δίνεται στον Πίνακα 4.2.

Πίνακας 4.2: Κρίσιμες τιμές του τροποποιημένου κριτηρίου Kolmogorov-Smirnov για τον δίπλευρο έλεγχο κανονικότητας με άγνωστες παραμέτρους.

Επίπεδο σημαντικότητας				
15%	10%	5%	2.5%	1%
0.775	0.819	0.895	0.995	1.035

Έλεγχος καλής προσαρμογής της Εκθετικής κατανομής με άγνωστη παράμετρο

Ο Lilliefors (1969) πρότεινε μία ακόμη παραλλαγή του Kolmogorov-Smirnov τεστ για τον έλεγχο της σύνθετης υπόθεσης της εκθετικής κατανομής. Δηλαδή το ενδιαφέρον τώρα επικεντρώνεται στο να ελέγξουμε αν το τ.δ. $X_1, X_2, \ldots, X_n$ προέρχεται από την εκθετική κατανομή, με άγνωστη μέση τιμή σ, δηλαδή από την κατανομή με αθροιστική συνάρτηση κατανομής $F(x) = 1 - \exp\left(-\frac{x}{\sigma}\right)$, για $x > 0$. Στο πλαίσιο αυτό, υπολογίζουμε αρχικά τις τυποποιημένες τιμές $Z_1, Z_2, ..., Z_n$ του τ.δ. που ορίζονται ως εξής:

$$Z_i = \frac{X_i}{\bar{X}}, i = 1,...,n, \text{ με } \bar{X} = \sum_{i=1}^{n} X_i/n. \tag{4.18}$$

Τότε η αρχική μηδενική υπόθεση ανάγεται στον έλεγχο της υπόθεσης ότι το τ.δ. $Z_1, Z_2, ..., Z_n$ προέρχεται από την εκθετική κατανομή με παράμετρο 1, δηλαδή με αθροιστική συνάρτηση κατανομής

$F_0(x) = 1 - \exp(-x)$, για $x > 0$ και μηδέν αλλού. Η κατάλληλη στατιστική συνάρτηση στην περίπτωση αυτή είναι η μέγιστη κατακόρυφη απόκλιση της εμπειρικής συνάρτησης κατανομής του τυποποιημένου δείγματος από την αθροιστική συνάρτηση κατανομής της εκθετικής με παράμετρο 1. Ομοίως με πριν, η ελεγχοσυνάρτηση του Lilliefors ορίζεται από τη σχέση $K_n = \max\{K_n^+, K_n^-\}$ με

$$K_n^+ = \max\left\{\max_{1\leqslant i\leqslant n}\left(i/n - F_0\left(Z_{(i)}\right)\right), 0\right\}, \tag{4.19}$$

και

$$K_n^- = \max\left\{\max_{1\leqslant i\leqslant n}\left(F_0\left(Z_{(i)}\right) - (i-1)/n\right), 0\right\} \tag{4.20}$$

με $Z_{(1)}, Z_{(2)}, \ldots, Z_{(n)}$, να είναι οι διατεταγμένες τυποποιημένες τιμές.

Προφανώς, και στην περίπτωση αυτού του έλεγχου, μεγάλες τιμές της στατιστικής συνάρτησης θα είναι εκείνες που θα συνηγορούν υπέρ της απόρριψης της μηδενικής υπόθεσης, αφού αυτές θα είναι αποτέλεσμα χαμηλού βαθμού εγγύτητας της συνάρτησης κατανομής του τυποποιημένου δείγματος προς τη συνάρτηση κατανομής της εκθετικής κατανομής με παράμετρο 1. Η κατανομή της K_n έχει μελετηθεί από τον Durbin (1975). Τότε, για δοθέν επίπεδο σημαντικότητας a, απορρίπτεται η μηδενική υπόθεση για τον δίπλευρο έλεγχο ότι τα δεδομένα προέρχονται από εκθετική κατανομή, όταν η τιμή της στατιστικής συνάρτησης ελέγχου υπερβαίνει την αντίστοιχη τιμή που δίνεται στον Πίνακα Π.13 του Παραρτήματος.

Παράδειγμα 4.11. Χρησιμοποιώντας τον έλεγχο που προτάθηκε από τον Lilliefors ελέγξτε, με επίπεδο σημαντικότητας 5%, αν τα ακόλουθα δεδομένα

$$2, 8, 6, 1, 11, 10, 3, 4, 7$$

προέρχονται από μια εκθετική κατανομή. Υπόδειξη: Για διευκόλυνση στους υπολογισμούς μπορείτε να χρησιμοποιήσετε την R.

Λύση Παραδείγματος 4.11. Αρχικά, δημιουργούμε τα τυποποιημένα δεδομένα σύμφωνα με τη σχέση (4.18) και τα διατάσσουμε σε αύξουσα τάξη μεγέθους:

```
x1<-c(2,8,6,1,11,10,3,4,7)
n<-length(x1)
z1<-sort((x1)/(mean(x1)))
```

Έπειτα, προσδιορίζουμε το $K_n = \max\{K_n^+, K_n^-\}$ μέσω του υπολογισμού των σχέσεων (4.19) και (4.20).

```
foxi<-pexp(z1, rate = 1)
i1<-1:n
Knplus<-(i1/n)-foxi
Knminus<-foxi-(i1-1)/n
Kn<-max(max(Knplus),max(Knminus)) # calculation of Kn value
Kn
```

Από τα παραπάνω προκύπτει ότι η τιμή του ελέγχου είναι ίση με 0.2015567. Καθώς από τον Πίνακα Π.13 του Παραρτήματος έχουμε ότι η κρίσιμη τιμή για $n = 9$, με επίπεδο σημαντικότητας 5%, είναι 0.3404, συμπεραίνουμε, καθώς $0.2015567 < 0.3404$, ότι σε ε.σ. 5% δεν απορρίπτεται η υπόθεση ότι το δείγμα προέρχεται από πληθυσμό που περιγράφεται από την εκθετική κατανομή.

Προφανώς, το παράδειγμα αυτό θα μπορούσε να λυθεί και χωρίς τη βοήθεια των παραπάνω εντολών. Αρχικά, υπολογίζουμε ότι $\bar{X} = \sum_{i=1}^{9} X_i/9 = 52/9$. Στη συνέχεια, υπολογίζουμε τις τυποποιημένες τιμές $Z_{(i)}$, $i = 1, \ldots, 9$

(εφαρμόζοντας στρογγυλοποίηση σε δύο δεκαδικά ψηφία) και κάνουμε τις επιπλέον πράξεις. Για τις πράξεις αυτές είναι πολύ βοηθητικός ο πίνακας που ακολουθεί. Για τον υπολογισμό των $F_0(Z_{(i)})$ χρησιμοποιήσαμε την εντολή `pexp` της `R`, αλλά προφανώς θα μπορούσαμε να τις υπολογίσουμε χρησιμοποιώντας άμεσα την α.σ.κ. της $\mathscr{E}(1)$.

i	$Z_{(i)}$	$F_0(Z_{(i)})$	i/n	$(i-1)/n$	$i/n - F_0(Z_{(i)})$	$F_0(Z_{(i)}) - (i-1)/n$
1	0.17	0.1563352	1/9	0/9	-0.04522407	0.15633518
2	0.35	0.2953119	2/9	1/9	-0.07308969	0.18420080
3	0.52	0.4054795	3/9	2/9	-0.07214612	0.18325723
4	0.69	0.4984239	4/9	3/9	-0.05397949	0.16509060
5	1.04	0.6465453	5/9	4/9	-0.09098976	0.20210087
6	1.21	0.7018027	6/9	5/9	-0.03513605	0.14624717
7	1.38	0.7484214	7/9	6/9	0.02935633	0.08175478
8	1.73	0.8227156	8/9	7/9	0.06617330	0.04493781
9	1.90	0.8504314	9/9	8/9	0.14956862	-0.03845751

Προκύπτει ότι $K_n = \max\{K_n^+, K_n^-\} = \max\{0.14956862, 0.20210087\}$ και, επομένως, $K_n = 0.20210087$. Οι αποκλίσεις στα αποτελέσματα οφείλονται σε στρογγυλοποιήσεις. □

Ο Stephens (1970) για την ειδική περίπτωση του σύνθετου ελέγχου καλής προσαρμογής στην εκθετική κατανομή τροποποίησε το κριτήριο των Kolmogorov-Smirnov θεωρώντας τη στατιστική συνάρτηση:

$$\mathrm{K}_n^* = (\mathrm{K}_n - 0.2/n)\left(\sqrt{n} + 0.26 + 0.5/\sqrt{n}\right),$$

όπου η K_n υπολογίζεται από τη σχέση $K_n = \max\{K_n^+, K_n^-\}$, με τις K_n^+ και K_n^- να δίνονται στις σχέσεις (4.19) και (4.20), αντίστοιχα. Τότε απορρίπτεται, για δοθέν επίπεδο σημαντικότητας a ο δίπλευρος έλεγχος ότι τα δεδομένα προέρχονται από εκθετική κατανομή, όταν η τιμή του τροποποιημένου κριτηρίου υπερβαίνει την αντίστοιχη τιμή που δίνεται στον Πίνακα 4.3. Για παράδειγμα, χρησιμοποιώντας τα δεδομένα του προηγούμενου παραδείγματος όπου $K_n = 0.20210087$, $n = 9$, δεν είναι δύσκολο να διαπιστώσουμε ότι $K_n^* = 0.616384$. Από τον Πίνακα 4.3, η κρίσιμη τιμή για έλεγχο, σε επίπεδο σημαντικότητας 5%, είναι 1.094 και άρα δεν απορρίπτουμε την υπόθεση ότι το δείγμα προέρχεται από πληθυσμό που περιγράφεται από την εκθετική κατανομή.

Πίνακας 4.3: Κρίσιμες τιμές του τροποποιημένου κριτηρίου Kolmogorov-Smirnov για τον δίπλευρο έλεγχο της εκθετικής κατανομής με άγνωστη παράμετρο.

Επίπεδο σημαντικότητας				
15%	10%	5%	2.5%	1%
0.926	0.995	1.094	1.184	1.298

Ο έλεγχος Smirnov: η γενίκευση του Kolmogorov-Smirnov για δύο ανεξάρτητα δείγματα

Όπως αναφέρθηκε στην αρχή της ενότητας αυτής, ο έλεγχος του Kolmogorov (1933) για τον έλεγχο της μηδενικής υπόθεσης, ότι ένα τυχαίο δείγμα προέρχεται από συγκεκριμένη κατανομή, επεκτάθηκε από τον Smirnov για τη σύγκριση δύο κατανομών στη βάση δύο ανεξάρτητων δειγμάτων (βλ. Smirnov, 1939a,b). Αναλυτικότερα, έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $F_X(\cdot)$ και $Y_1, Y_2, \ldots, Y_m$, ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση

κατανομής $G_Y(\cdot)$. Υποθέτουμε ότι τα δύο δείγματα είναι ανεξάρτητα. Έστω, επιπρόσθετα, $F_n(\cdot)$ και $G_m(\cdot)$ οι αντίστοιχες εμπειρικές αθροιστικές συναρτήσεις κατανομής αυτών των πληθυσμών. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση $H_0 : F_X(x) = G_Y(x)$, $\forall x \in \mathbb{R}$, έναντι της εναλλακτικής $H_1 : F_X(x) \neq G_Y(x)$, για κάποιο $x \in \mathbb{R}$. Με παρόμοιο σκεπτικό με αυτό που αναπτύχθηκε από τον Kolmogorov για τον έλεγχο καλής προσαρμογής με ένα δείγμα, γίνεται αντιληπτό ότι η στατιστική συνάρτηση για τον έλεγχο της παραπάνω υπόθεσης θα βασίζεται στις εμπειρικές αθροιστικές συναρτήσεις κατανομής αυτών και μάλιστα θα αποτελεί ένα μέτρο της εγγύτητάς τους. Ειδικότερα, χρησιμοποιείται η στατιστική συνάρτηση $D_{n,m}$ που ορίζεται ως:

$$D_{n,m} = \sup_{x\in\mathbb{R}} |F_n(x) - G_m(x)|$$

ή, ισοδύναμα, η στατιστική συνάρτηση

$$D_{n,m} = \max_{1\leqslant i\leqslant n+m} |F_n(Z_i) - G_m(Z_i)|, \tag{4.21}$$

όπου $Z_1, \dots, Z_{n+m}$ είναι το διατεταγμένο δείγμα που προκύπτει από τη σύνθεση των $X_1, X_2, \dots, X_n$ και $Y_1, Y_2, \dots, Y_m$. Στο πλαίσιο αυτό, αν n_i είναι το πλήθος των παρατηρήσεων από το δείγμα $X_1, X_2, \dots, X_n$, που είναι μικρότερες ή ίσες από την τιμή Z_i, για $i = 1, \dots, n + m$, τότε μια ισοδύναμη έκφραση της παραπάνω στατιστικής συνάρτησης είναι η $D_{n,m} = \max\{D^+_{n,m}, D^-_{n.m}\}$, όπου

$$D^+_{n,m} = \sup_{x\in\mathbb{R}} [F_n(x) - G_m(x)], \tag{4.22}$$

και

$$D^-_{n,m} = \sup_{x\in\mathbb{R}} (G_m(x) - F_n(x)). \tag{4.23}$$

Επιπρόσθετα,

$$D^+_{n,m} = \frac{n+m}{m} \max_{1\leqslant i\leqslant n+m} \left(\frac{n_i}{n} - \frac{i}{n+m}\right), \tag{4.24}$$

και

$$D^-_{n,m} = \frac{n+m}{m} \max_{1\leqslant i\leqslant n+m} \left(\frac{i}{n+m} - \frac{n_i}{n}\right), \tag{4.25}$$

όπως αποδεικνύεται στην παρακάτω πρόταση.

Πρόταση 4.3. Έστω $z_1 < z_2 < \dots < z_{n+m}$ οι διατεταγμένες τιμές όλων των $(n + m)$ παρατηρήσεων και n_i είναι το πλήθος των παρατηρήσεων από το δείγμα των $X_1, \dots, X_n$ που είναι μικρότερες ή ίσες από το z_i, $i = 1, 2, \dots, n + m$, τότε

$$D_{n,m} = \max\{D^+_{n,m}, D^-_{n,m}\},$$

όπου $D^+_{n,m} = \dfrac{n+m}{m} \max\limits_{1\leqslant i\leqslant n+m} \left(\dfrac{n_i}{n} - \dfrac{i}{n+m}\right)$ και $D^-_{n,m} = \dfrac{n+m}{m} \max\limits_{1\leqslant i\leqslant n+m} \left(\dfrac{i}{n+m} - \dfrac{n_i}{n}\right)$.

Απόδειξη Πρότασης 4.3. Θεωρούμε τη διαμέριση του $\mathbb{R} = \bigcup\limits_{i=0}^{n} A_i$, όπου $A_i = [z_i, z_{i+1})$, $i = 0, 1, \dots, n + m$, με $z_0 = -\infty$ και $z_{n+m+1} = \infty$. Τότε ισχύει ότι:

$$F_n(z) = \frac{n_i}{n} = F_n(z_i) \text{ και } G_m(z) = \frac{i - n_i}{m} = G_m(z_i), z_i \leqslant z < z_{i+1}, \text{ για } i = 0, 1, \dots, n + m.$$

Δηλαδή, σε κάθε διάστημα A_i, ισχύει ότι $\sup\limits_{z\in A_i} |F_n(z) - G_m(z)| = \left|\frac{n_i}{n} - \frac{i-n_i}{m}\right| = |F_n(z_i) - G_m(z_i)|$, δηλαδή είναι σταθερό. Επομένως,

$$D_{n,m} = \sup_{x\in\mathbb{R}} |F_n(x) - G_m(x)| = \max_{0\leqslant i\leqslant n+m} \sup_{z\in A_i} |F_n(z) - G_m(z)| = \max_{0\leqslant i\leqslant n+m} |F_n(z_i) - G_m(z_i)|.$$

Οπότε, αν βγάλουμε το απόλυτο με θετικό πρόσημο, έχουμε

$$D^{+}_{n,m} = \max_{0 \leqslant i \leqslant n+m} \{F_n(z_i) - G_m(z_i)\} = \max_{0 \leqslant i \leqslant n+m} \left\{ \frac{n_i}{n} - \frac{i - n_i}{m} \right\}.$$

Βγάζοντας κοινό παράγοντα το $\frac{n+m}{m}$ και, μετά από λίγη άλγεβρα, καταλήγουμε ότι

$$D^{+}_{n,m} = \frac{n+m}{m} \max_{0 \leqslant i \leqslant n+m} \left(\frac{n_i}{n} - \frac{i}{n+m} \right).$$

το οποίο ισοδύναμα γράφεται

$$D^{+}_{n,m} = \frac{n+m}{m} \max_{1 \leqslant i \leqslant n+m} \left(\frac{n_i}{n} - \frac{i}{n+m} \right),$$

καθώς το max για $i = 0$ είναι το μηδέν.

Με παρόμοιο τρόπο υπολογίζουμε ότι

$$D^{-}_{n,m} = \max_{0 \leqslant i \leqslant n+m} \{G_m(z_i) - F_n(z_i)\} = \frac{n+m}{m} \max_{1 \leqslant i \leqslant n+m} \left(\frac{i}{n+m} - \frac{n_i}{n} \right),$$

και η απόδειξη ολοκληρώνεται. □

Είναι προφανές ότι η τιμή της στατιστικής συνάρτησης $D_{n,m}$ εξαρτάται από την τάξη των παρατηρήσεων και όχι από την τιμή τους. Επιπρόσθετα, η μηδενική υπόθεση, προφανώς, απορρίπτεται για μεγάλες τιμές της στατιστικής συνάρτησης $D_{n,m}$. Δύο εύλογα ερωτήματα που μπορεί να έχουν προκύψει είναι ποιες τιμές της στατιστικής συνάρτησης μπορούν να θεωρηθούν μεγάλες για δοθέν επίπεδο σημαντικότητας και αν ο προσδιορισμός αυτών των μεγάλων τιμών εξαρτάται από τις προς έλεγχο αθροιστικές συναρτήσεις κατανομής.

Όσον αφορά το δεύτερο ερώτημα, η κατανομή της $D_{n,m}$, υπό τη μηδενική υπόθεση, δεν εξαρτάται από τις κατανομές $F(\cdot)$ και $G(\cdot)$. Αυτό αιτιολογείται καθώς, αν συμβολίσουμε με $F_0(\cdot)$ την κοινή συνεχή αθροιστική συνάρτηση κατανομής, θέτοντας $z = F_0(x)$, είναι $F_n(x) = F_n(z)$ και $G_m(x) = G_m(z)$, όπου η τυχαία μεταβλητή Z, που αντιστοιχεί στο z, ακολουθεί $\mathcal{U}(0,1)$ (βλ. Gibbons and Chakraborti, 2020).

Όσον αφορά το πρώτο ερώτημα, είναι σαφές ότι απαιτείται ο προσδιορισμός της κατανομής της στατιστικής συνάρτησης υπό τη μηδενική υπόθεση, για να είναι εφικτός ο καθορισμός της τιμής της στατιστικής συνάρτησης, που θα θεωρείται μεγάλη και θα οδηγεί σε απόρριψη της μηδενικής υπόθεσης ότι τα δύο ανεξάρτητα δείγματα προέρχονται από έναν κοινό πληθυσμό, με επίπεδο σημαντικότητας a. Στο πλαίσιο αυτό, η ακριβής κατανομή της στατιστικής συνάρτησης $D_{n.m}$, υπό τη μηδενική υπόθεση, έχει αποτελέσει αντικείμενο μελέτης πολλών ερευνητών και ποικίλες μέθοδοι έχουν προταθεί, οι οποίες περιλαμβάνουν επαναληπτικές σχέσεις για τον υπολογισμό των πιθανοτήτων. Για μια σύνοψη αυτών των μεθόδων βλ. Hodges (1958). Στο σημείο αυτό, πρέπει να επισημανθεί ότι μπορεί να βρεθεί σε κλειστή μορφή η ακριβής κατανομή, υπό τη μηδενική υπόθεση, της στατιστικής συνάρτησης $D_{n,m}$. Στο θεώρημα που ακολουθεί αποτελεί αντικείμενο μελέτης η ειδική περίπτωση όπου $m = n$.

Θεώρημα 4.3

Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $F_X(\cdot)$ και $Y_1, Y_2, \ldots, Y_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $G_Y(\cdot)$. Υποθέτουμε ότι τα δύο δείγματα είναι ανεξάρτητα. Έστω, επιπρόσθετα, $F_n(\cdot)$ και $G_n(\cdot)$ οι αντίστοιχες εμπειρικές αθροιστικές συναρτήσεις κατανομής αυτών των πληθυσμών. Η κατανομή, υπό τη μηδενική υπόθεση $H_0 : F_X(x) = G_Y(x)$, $\forall\, x \in \mathbb{R}$, των στατιστικών συναρτήσεων

$D^+_{n,n} = \sup_x [F_n(x) - G_n(x)]$ και $D^-_{n,n} = \sup_{x\in\mathbb{R}} (G_n(x) - F_n(x))$ δίνεται από τη σχέση:

$$P\left(D^+_{n,n} > d\right) = P\left(D^-_{n,n} > d\right) = \frac{\begin{pmatrix} 2n \\ [n(d+1)] \end{pmatrix}}{\begin{pmatrix} 2n \\ n \end{pmatrix}},$$

όπου με $[n(d+1)]$ συμβολίζουμε τον μεγαλύτερο ακέραιο που είναι μικρότερος ή ίσος από $n(d+1)$.

Απόδειξη Θεωρήματος 4.3. Για την απόδειξη στη γενική περίπτωση βλ. Hodges (1958) και τις εκεί αναφορές. Ο τρόπος σκέψης θα δοθεί στο επόμενο παράδειγμα που πραγματεύεται την ειδική περίπτωση που $m = n = 2$. □

Παράδειγμα 4.12. Έστω X_1, X_2, ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $F_X(\cdot)$ και Y_1, Y_2 ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $G_Y(\cdot)$. Υποθέτουμε ότι τα δύο δείγματα είναι ανεξάρτητα. Έστω, επιπρόσθετα, $F_n(\cdot)$ και $G_n(\cdot)$ οι αντίστοιχες εμπειρικές αθροιστικές συναρτήσεις κατανομής αυτών των πληθυσμών. Να προσδιοριστεί η κατανομή της στατιστικής συνάρτησης:

$$D_{2,2} = \sup_{x\in\mathbb{R}} |F_n(x) - G_n(x)|.$$

Λύση Παραδείγματος 4.12. Καθώς X_1, X_2 είναι ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $F_X(\cdot)$ και Y_1, Y_2 είναι ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $G_Y(\cdot)$, συνολικά έχουμε 4 παρατηρήσεις. Αυτό που μας ενδιαφέρει είναι η διάταξή τους, μη λαμβάνοντας υπόψη τη διάταξη εντός των δειγμάτων.

Προκύπτει εύκολα ότι υπάρχουν συνολικά $\begin{pmatrix} 4 \\ 2 \end{pmatrix} = 6$ διατάξεις (ισοπίθανες υπό τη μηδενική υπόθεση), οι οποίες είναι οι:

$$XXYY, XYXY, XYYX, YXXY, YXYX, YYXX.$$

Έτσι, η εμπειρική συνάρτηση κατανομής F_2 λαμβάνει τις ακόλουθες τιμές:

$$1/2\ 1\ 1\ 1, 1/2\ 1/2\ 1\ 1, 1/2\ 1/2\ 1/2\ 1, 0\ 1/2\ 1\ 1, 0\ 1/2\ 1/2\ 1, 0\ 0\ 1/2\ 1,$$

ενώ η εμπειρική συνάρτηση κατανομής G_2 λαμβάνει τις ακόλουθες τιμές:

$$0\ 0\ 1/2\ 1, 0\ 1/2\ 1/2\ 1, 0\ 1/2\ 1\ 1, 1/2\ 1/2\ 1/2\ 1, 1/2\ 1/2\ 1\ 1, 1/2\ 1\ 1\ 1.$$

Οπότε προκύπτει ότι για κάθε περίπτωση οι δυνατές τιμές του $\sup|F_n(x) - G_n(x)|$ είναι: 1, 1/2, 1/2, 1/2, 1/2, 1 αντίστοιχα, με $P\left(D_{2,2} = 1\right) = 2/6 = 1/3$ και $P\left(D_{2,2} = 1/2\right) = 4/6 = 2/3$. □

Από το αποτέλεσμα του Θεωρήματος 4.3 έχει δημιουργηθεί πίνακας κρίσιμων τιμών του ελέγχου Smirnov για την περίπτωση δύο ανεξάρτητων δειγμάτων ίσου μεγέθους (βλ. Πίνακα Π.14 του Παραρτήματος). Από την άλλη μεριά, πίνακας κρίσιμων τιμών του ελέγχου για την περίπτωση δύο ανεξάρτητων δειγμάτων μη ίσου μεγέθους δίνεται για περιορισμένες τιμές των m, n (βλ. Πίνακα Π.15 του Παραρτήματος).

Παράδειγμα 4.13. Στον πίνακα που ακολουθεί καταγράφονται οι τιμές δύο ανεξάρτητων τυχαίων δειγμάτων, μεγέθους 6 και 9, αντίστοιχα από δύο πληθυσμούς με αθροιστικές συναρτήσεις κατανομής $F_X(\cdot)$ και $G_Y(\cdot)$, αντίστοιχα. Να ελέγξετε, με επίπεδο σημαντικότητας 5%, τη μηδενική υπόθεση $H_0 : F_X(x) = G_Y(x)$, $\forall\, x \in \mathbb{R}$, έναντι της εναλλακτικής $H_1 : F_X(x) \neq G_Y(x)$, για κάποιο $x \in \mathbb{R}$.

X_i	5.2	5.4	5.9	6.5	8.2	9.1			
Y_i	7.4	9.8	8.4	9.3	10.1	7.5	6.4	6.9	7.2

Λύση Παραδείγματος 4.13. Από τη θεωρία έχουμε ότι η ελεγχοσυνάρτηση που θα χρησιμοποιηθεί είναι η:

$$D_{6,9} = \sup_{x\in\mathbb{R}} \left|F_6(x) - G_9(x)\right| = \max_{1\leqslant i\leqslant 15} \left|F_{n=6}(Z_i) - G_{m=9}(Z_i)\right|,$$

όπου $Z_1, \dots, Z_{15}$ είναι το διατεταγμένο δείγμα που προκύπτει από τη σύνθεση των $X_1, X_2, \dots, X_6$ και $Y_1, Y_2, \dots, Y_9$, ενώ με $F_n(\cdot)$ και $G_m(\cdot)$ συμβολίζονται οι εμπειρικές αθροιστικές συναρτήσεις κατανομής των X και Y, αντίστοιχα. Είναι τότε (για διευκόλυνση με έντονη γραφή οι τιμές X_i) :

Z_i	$F_n(Z_i)$	$G_m(Z_i)$	$\lvert F_n(Z_i) - G_m(Z_i)\rvert$
5.2	1/6	0	1/6=9/54
5.4	2/6	0	2/6=18/54
5.9	3/6	0	3/6=27/54
6.4	3/6	1/9	3/6-1/9=21/54
6.5	4/6	1/9	4/6-1/9=30/54
6.9	4/6	2/9	4/6-2/9=24/54
7.2	4/6	3/9	4/6-3/9=18/54
7.4	4/6	4/9	4/6-4/9=12/54
7.5	4/6	5/9	4/6-5/9=6/54
8.2	5/6	5/9	5/6-5/9=15/54
8.4	5/6	6/9	5/6-6/9=9/54
9.1	1	6/9	3/9=18/54
9.3	1	7/9	2/9=12/54
9.8	1	8/9	1/9=6/54
10.1	1	1	0

Επομένως, είναι $D_{6,9}$ =30/54=0.556. Από τον Πίνακα Π.15 του Παραρτήματος έχουμε ότι η κρίσιμη τιμή του δίπλευρου ελέγχου, σε επίπεδο σημαντικότητας 5%, είναι ίση με 2/3 (ή 36/54). Επομένως, καθώς $30/54 < 36/54$, δεν μπορεί να απορριφθεί, σε ε.σ. 5%, η μηδενική υπόθεση ότι οι αθροιστικές συναρτήσεις των δύο κατανομών συμπίπτουν. □

Επιπρόσθετα, καθώς η ακριβής κατανομή υπό τη μηδενική υπόθεση της στατιστικής συνάρτησης $D_{n,m}$ στη γενική περίπτωση και ιδιαίτερα για μεγάλες τιμές των m, n είναι πολύπλοκη, στην πράξη χρησιμοποιείται η ασυμπτωτική κατανομή της, η οποία προσδιορίζεται στο θεώρημα που ακολουθεί (βλ. Smirnov, 1939a,b).

Θεώρημα 4.4

Έστω ότι $X_1, X_2, \dots, X_n$, ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $F_X(\cdot)$ και $Y_1, Y_2, \dots, Y_m$, ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $G_Y(\cdot)$. Υποθέτουμε ότι τα δύο δείγματα είναι ανεξάρτητα. Έστω, επιπρόσθετα, $F_n(\cdot)$ και $G_m(\cdot)$ οι αντίστοιχες εμπειρικές αθροιστικές συναρτήσεις κατανομής αυτών. Όταν n και m είναι τέτοια, ώστε $m, n \to \infty$, $\frac{m}{n} \to q$, με q σταθερό θετικό αριθμό, τότε, υπό τη μηδενική υπόθεση $H_0 : F_X(x) = G_Y(x)$, $\forall\, x \in \mathbb{R}$, ισχύει ότι:

$$P\left(\sqrt{\frac{mn}{m+n}} D_{n,m} \leqslant d\right) \approx 1 - 2\sum_{j=1}^{\infty} e^{-2j^2d^2}.$$

Απόδειξη Θεωρήματος 4.4. Η απόδειξη ξεφεύγει από τους σκοπούς του παρόντος συγγράμματος και παραλείπεται. □

Η παραπάνω ασυμπτωτική κατανομή μπορεί να χρησιμοποιηθεί αν $m, n > 30$ και είναι συγκρίσιμα μεταξύ τους (Kvam and Vidakovic, 2007). Επιπλέον, ανατρέχοντας στο Θεώρημα 4.2, παρατηρούμε ότι η ασυμπτωτική κατανομή της στατιστικής συνάρτησης $\sqrt{n}D_n$ για τον έλεγχο καλής προσαρμογής ενός δείγματος είναι ίδια με αυτήν της στατιστικής συνάρτησης $\sqrt{\frac{mn}{m+n}}D_{n,m}$ για τον έλεγχο της ισότητας δύο αθροιστικών συναρτήσεων κατανομής.

Γνωρίζοντας ότι η $H_0 : F_X(x) = G_Y(x), \forall x \in \mathbb{R}$, έναντι της εναλλακτικής $H_1 : F_X(x) \neq G_Y(x)$, για κάποιο $x \in \mathbb{R}$, απορρίπτεται για μεγάλες τιμές της στατιστικής συνάρτησης $D_{n,m}$, έπεται ότι, για να έχουμε, ασυμπτωτικά επίπεδο σημαντικότητας a, θα πρέπει να προσδιορίσουμε την τιμή, έστω c_a', που είναι τέτοια, ώστε $P\left(\sqrt{\frac{mn}{m+n}}D_{n,m} \geqslant c_a' \middle| H_0\right) \approx a$. Επομένως, χρησιμοποιώντας το Θεώρημα 4.4 έχουμε ότι η τιμή c_a' θα πρέπει να είναι τέτοια, ώστε $1 - H(c_a') \approx a$. Συνεπώς, η τιμή c_a' ταυτίζεται με την τιμή c_a του ελέγχου Kolmogorov. Όπως έχουμε δει, ο Πίνακας Π.11 του Παραρτήματος (βλ., μεταξύ άλλων, Miller, 1956) προσδιορίζει τις τιμές $c_a/\sqrt{n}$ για διάφορα μεγέθη δείγματος και επίπεδα σημαντικότητας. Επομένως, ισοδύναμα, θα απορρίπτουμε τη μηδενική υπόθεση, αν

$$\sqrt{\frac{mn}{m+n}}D_{n,m} \geqslant c_a$$

ή αν

$$\sqrt{\frac{m}{m+n}}D_{n,m} \geqslant c_a/\sqrt{n}.$$

Παρατηρήστε ότι, καθώς δίνεται η τιμή $c_a/\sqrt{n}$, απορρίπτεται η μηδενική υπόθεση του απλού δίπλευρου ελέγχου καλής προσαρμογής, αν η τιμή της στατιστικής συνάρτησης $\sqrt{\frac{m}{m+n}}D_{n,m}$ είναι μεγαλύτερη από την κρίσιμη τιμή του Πίνακα Π.11.

Μέχρι τώρα, το ενδιαφέρον επικεντρώθηκε στον δίπλευρο έλεγχο της μηδενικής υπόθεσης $H_0 : F(x) = G(x)$, $\forall x \in \mathbb{R}$, έναντι της εναλλακτικής $H_1 : F(x) \neq G(x)$, για κάποιο $x \in \mathbb{R}$. Με παρόμοιο τρόπο μπορούν να προταθούν τρόποι ελέγχου της ίδιας μηδενικής υπόθεσης έναντι της εναλλακτικής

- $H_{1+} : F(x) > G(x)$, για κάποιο $x \in \mathbb{R}$, ή της
- $H_{1-} : F(x) < G(x)$, για κάποιο $x \in \mathbb{R}$.

Ειδικότερα στην πρώτη περίπτωση, χρησιμοποιείται η στατιστική συνάρτηση D_{mn}^+ που παριστάνει τη μέγιστη απόκλιση μεταξύ της εμπειρικής αθροιστικής συνάρτησης κατανομής $F_n(x)$ και της $G_m(x)$ πάνω στις τιμές του x, για τις οποίες η $F_n(x)$ είναι πάνω από την $G_m(x)$. Από την άλλη μεριά, στη δεύτερη περίπτωση χρησιμοποιείται η στατιστική συνάρτηση $D_{n,m}^-$, η οποία παριστάνει τη μέγιστη απόκλιση μεταξύ της $F_n(x)$ και της $G_m(x)$ πάνω στις τιμές του x, για τις οποίες η $F_n(x)$ είναι κάτω από την $G_m(x)$. Επομένως, είναι προφανές ότι απορρίπτονται οι παραπάνω μηδενικές υποθέσεις έναντι των αντίστοιχων εναλλακτικών για μεγάλες τιμές των στατιστικών συναρτήσεων $D_{n,m}^+$ και $D_{n,m}^-$, αντίστοιχα. Προκύπτει, λοιπόν, ότι απορρίπτονται σε επίπεδο σημαντικότητας a οι μηδενικές υποθέσεις, αν και μόνο αν

$\sqrt{\frac{mn}{m+n}}D_{n,m}^+ \geqslant c_{a/2}'$ ή $\sqrt{\frac{mn}{m+n}}D_{n,m}^- \geqslant c_{a/2}'$, αντίστοιχα, όπου c_a' τέτοιο, ώστε $P\left(\sqrt{\frac{mn}{m+n}}D_{n,m} \geqslant c_a'\right) \approx a$.

Παρατήρηση 4.11. Στη βιβλιογραφία έχουν προταθεί διάφορες παραλλαγές αυτού του στατιστικού τεστ, καθώς και επεκτάσεις του στην περίπτωση k δειγμάτων, $k \geq 3$. Για περισσότερες πληροφορίες παραπέμπουμε στους Kvam and Vidakovic (2007) και στις εκεί αναφορές.

4.3.2 Οι έλεγχοι Cramér-Von Mises, Watson, Kuiper και Anderson-Darling

Εκτός από το Kolmogorov-Smirnov (KS) τεστ υπάρχει ένας μεγάλος αριθμός από ελέγχους καλής προσαρμογής που στηρίζονται στην εμπειρική αθροιστική συνάρτηση κατανομής. Στην ενότητα αυτή, θα παρουσιαστούν πολύ σύντομα οι κυριότεροι τέτοιοι έλεγχοι.

Ο έλεγχος των Cramér-Von Mises

Έστω $X_1, X_2, \dots, X_n$, ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $F_X(\cdot)$ και $F_n(\cdot)$ η αντίστοιχη εμπειρική αθροιστική συνάρτηση κατανομής. Για τον έλεγχο της μηδενικής υπόθεσης $H_0 : F_X(x) = F_0(x), \forall x \in \mathbb{R}$, έναντι της εναλλακτικής υπόθεσης $H_1 : F(x) \neq F_0(x)$, για κάποιο $x \in \mathbb{R}$, ο Σουηδός Harald Cramér (1893 – 1985) και ο Αυστριακός Richard Edler von Mises (1883 – 1953) πρότειναν, ανεξάρτητα, στις εργασίες τους Cramér (1928) και Mises (1928), αντίστοιχα, τη στατιστική συνάρτηση:

$$W_n^2 = n \int_{-\infty}^{\infty} \{F_n(x) - F_0(x)\}^2 dF_0(x),$$

η οποία λαμβάνει την ισοδύναμη υπολογιστική μορφή

$$W_n^2 = \sum_{i=1}^{n} [F_0\left(X_{(i)}\right) - \{(2i-1)/2n\}]^2 + \frac{1}{12n},$$

όπου $X_{(i)}$, $i = 1, ..., n,$ συμβολίζει την $i-$οστή διατεταγμένη παρατήρηση. Από τον τρόπο ορισμού της στατιστικής συνάρτησης W_n^2 είναι προφανές ότι απορρίπτεται η μηδενική υπόθεση για μεγάλες τιμές της. Για την ασυμπτωτική κατανομή αυτής της στατιστικής συνάρτησης, καθώς και για κρίσιμα σημεία αυτής, παραπέμπουμε στην εργασία των Anderson and Darling (1952). Τέλος, επισημαίνεται ότι υπό τη μηδενική υπόθεση η κατανομή της στατιστικής συνάρτησης δεν εξαρτάται από την F_0 και, επίσης, η στατιστική συνάρτηση έχει τροποποιηθεί από τον Αμερικανό μαθηματικό και στατιστικό Theodore Wilbur Anderson (1918 –2016), για τον έλεγχο της ισότητας δύο κατανομών (βλ. Anderson, 1958).

Παρατήρηση 4.12. Για την αποφυγή εκτεταμένων πινάκων για κρίσιμες τιμές του κριτηρίου W_n^2, για τα διάφορα επίπεδα σημαντικότητας και μεγέθη δείγματος, ο Stephens (1970) τροποποίησε το κριτήριο αυτό, έτσι ώστε να γίνεται απευθείας σύγκριση για την απόρριψη ή όχι της H_0 για τον απλό έλεγχο καλής προσαρμογής. Το τροποποιημένο κριτήριο δίνεται από τη σχέση:

$$W_n^{2,*} = (W_n^2 - 0.4/n + 0.6/n^2)(1 + 1/n),$$

και απορρίπτεται, για δοθέν επίπεδο σημαντικότητας a, η μηδενική υπόθεση έναντι της εναλλακτικής, όταν η τιμή του τροποποιημένου κριτηρίου υπερβαίνει την αντίστοιχη τιμή που δίνεται στον Πίνακα 4.4.

Πίνακας 4.4: Κρίσιμες τιμές του τροποποιημένου κριτηρίου Cramér-Von Mises για τον δίπλευρο έλεγχο καλής προσαρμογής.

Επίπεδο σημαντικότητας				
15%	10%	5%	2.5%	1%
0.284	0.347	0.461	0.581	0.743

Ο έλεγχος του Watson

Ο Αυστραλιανός στατιστικός Geoffrey Stuart Watson (1921 –1998) πρότεινε μια τροποποίηση του ελέγχου των Cramér-Von Mises (βλ. Watson, 1961) θεωρώντας τη στατιστική συνάρτηση:

$$U_n^2 = n\int_{-\infty}^{\infty}\left\{F_n(x) - F_0(x) - \int_{-\infty}^{\infty}(F_n(x) - F_0(x))dF_0(x)\right\}^2 dF_0(x),$$

η οποία, ισοδύναμα, δίνεται από τη σχέση:

$$U_n^2 = W_n^2 - n(\bar{F}_0 - 0.5)^2,$$

όπου W_n^2 η στατιστική συνάρτηση των Cramér-Von Mises και $\bar{F}_0 = \dfrac{\sum_{i=1}^{n} F_0(X_i)}{n}$. Το κριτήριο αυτό, όπως και τα προηγούμενα, είναι απαλλαγμένο παραμέτρων (distribution free). Κρίσιμα σημεία για την ασυμπτωτική κατανομή του U_n^2 έχουν δοθεί από τους Watson (1961), Stephens (1963) και Stephens (1964).

Παρατήρηση 4.13. Για την αποφυγή εκτεταμένων πινάκων για κρίσιμες τιμές του κριτηρίου U_n^2, για τα διάφορα επίπεδα σημαντικότητας και μεγέθη δείγματος, ο Stephens (1970) τροποποίησε το κριτήριο αυτό, έτσι ώστε να γίνεται απευθείας σύγκριση για την απόρριψη ή όχι της H_0 για τον απλό έλεγχο καλής προσαρμογής. Το τροποποιημένο κριτήριο δίνεται από τη σχέση:

$$U_n^{2,*} = (U_n^2 - 0.1/n + 0.1/n^2)(1 + 0.8/n),$$

και απορρίπτεται, για δοθέν επίπεδο σημαντικότητας a, η μηδενική υπόθεση έναντι της εναλλακτικής, όταν η τιμή του τροποποιημένου κριτηρίου υπερβαίνει την αντίστοιχη τιμή που δίνεται στον Πίνακα 4.5.

Πίνακας 4.5: Κρίσιμες τιμές του τροποποιημένου κριτηρίου Watson για τον δίπλευρο έλεγχο καλής προσαρμογής.

Επίπεδο σημαντικότητας				
15%	10%	5%	2.5%	1%
0.131	0.152	0.187	0.222	0.268

Ο έλεγχος του Kuiper

Ο έλεγχος που προτάθηκε από τον Ολλανδό μαθηματικό Nicolaas Hendrik "Nico" Kuiper (1920-1994) συνδέεται άμεσα με τον έλεγχο των Kolmogorov–Smirnov. Όπως είδαμε στην προηγούμενη ενότητα, η στατιστική συνάρτηση D_n του ελέγχου των Kolmogorov–Smirnov προσδιορίζεται ως το μέγιστο των D_n^+ και D_n^-. Ο Kuiper (1960) αντί αυτού πρότεινε ως στατιστική συνάρτηση το άθροισμα των D_n^+ και D_n^-, δηλαδή τη $V_n = D_n^+ + D_n^-$. Αυτή η μικρή αλλαγή έχει ως αποτέλεσμα ο έλεγχος που προτάθηκε από τον Kuiper να είναι πιο ευαίσθητος στις ουρές από ό,τι ο αντίστοιχος των Kolmogorov–Smirnov. Πίνακες κρίσιμων τιμών είναι διαθέσιμοι (βλ., για παράδειγμα, Pearson and Hartley, 1972).

Παρατήρηση 4.14. Για την αποφυγή εκτεταμένων πινάκων για κρίσιμες τιμές του κριτηρίου V_n, για τα διάφορα επίπεδα σημαντικότητας και μεγέθη δείγματος, ο Stephens (1970) τροποποίησε το κριτήριο αυτό,

έτσι ώστε να γίνεται απευθείας σύγκριση για την απόρριψη ή όχι της H_0 για τον απλό έλεγχο καλής προσαρμογής. Το τροποποιημένο κριτήριο είναι:

$$V_n^* = V_n(\sqrt{n} + 0.155 + 0.24/\sqrt{n}),$$

και απορρίπτεται, για δοθέν επίπεδο σημαντικότητας a, η μηδενική υπόθεση έναντι της εναλλακτικής, όταν η τιμή του τροποποιημένου κριτηρίου υπερβαίνει την αντίστοιχη τιμή που δίνεται στον Πίνακα 4.6.

Πίνακας 4.6: Κρίσιμες τιμές του τροποποιημένου κριτηρίου Kuiper για τον δίπλευρο έλεγχο καλής προσαρμογής.

Επίπεδο σημαντικότητας				
15%	10%	5%	2.5%	1%
1.537	1.620	1.747	1.862	2.001

Ο έλεγχος των Anderson-Darling

Οι Αμερικανοί Theodore Wilbur Anderson (1918 –2016) και Donald Allan Darling (1915–2014) θέλησαν, μεταξύ άλλων, να τροποποιήσουν τον έλεγχο των Cramér-Von Mises, έτσι ώστε να δίνει μεγαλύτερη προσοχή και βαρύτητα στις ουρές της κατανομής. Σε αυτό το πλαίσιο, οι Anderson and Darling (1952) πρότειναν τη στατιστική συνάρτηση

$$\mathrm{A}_n^2 = n\int_{-\infty}^{\infty} \{F_n(x) - F_0(x)\}^2 \, \psi(F_0(x)) dF_0(x),$$

όπου $\psi(t) \geqslant 0$, $0 \leqslant t \leqslant 1$, είναι μια προκαθορισμένη συνάρτηση βάρους που επιλέγεται, έτσι ώστε να εξετάζονται συγκεκριμένες περιοχές της κατανομής. Για επιπλέον ιδιότητες που πρέπει να πληροί η συνάρτηση βάρους παραπέμπουμε στους Anderson and Darling (1952). Στην ειδική περίπτωση όπου $\psi(t) = 1$, $0 \leqslant t \leqslant 1$, προκύπτει η στατιστική συνάρτηση των Cramér-Von Mises, ενώ οι Anderson and Darling (1952) θεώρησαν την ειδική περίπτωση όπου $\psi(t) = \dfrac{1}{t(1-t)}$, για $0 \leqslant t \leqslant 1$, η οποία οδηγεί σε τεστ που είναι ευαίσθητο στις ουρές. Τότε προκύπτει και η ισοδύναμη έκφραση, η οποία είναι προτιμότερη σε πρακτικές εφαρμογές,

$$A_n^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} (\log t_i + \log(1 - t_{n+1-i}))$$

με $t_i = F_0(X_{(i)})$, $i = 1, 2, \ldots, n$.

Προφανώς, η μηδενική υπόθεση απορρίπτεται για μεγάλες τιμές της στατιστικής συνάρτησης A_n^2. Στην περίπτωση που η αθροιστική συνάρτηση κατανομής είναι πλήρως καθορισμένη (άρα έχουμε έναν απλό έλεγχο καλής προσαρμογής) ανεξάρτητα από το μέγεθος δείγματος για $n \geqslant 5$ ο Stephens (1970) προτείνει για τον απλό έλεγχο καλής προσαρμογής να απορρίπτεται η μηδενική υπόθεση του δίπλευρου ελέγχου, όταν η τιμή της στατιστικής συνάρτησης A_n^2 υπερβαίνει την αντίστοιχη τιμή που δίνεται στον Πίνακα 4.7.

Παρατήρηση 4.15. Η ασυμπτωτική κατανομή της στατιστικής συνάρτησης A_n^2, κρίσιμα σημεία αυτής για διάφορες τιμές του n, καθώς και προσδιορισμός των πιθανοτήτων της μορφής $P(A_n^2 \leqslant d)$ για διάφορες τιμές του d δόθηκαν από τον Lewis (1961). Σημειώνουμε ότι οι πιθανότητες που δίνονται είναι βοηθητικές για την εύρεση της p-τιμής του ελέγχου, ενώ η σύγκλιση του A_n^2 στην ασυμπτωτική του κατανομή είναι πάρα πολύ γρήγορη, με ικανοποιητική προσέγγιση ακόμη και για $n = 5$.

Πίνακας 4.7: Κρίσιμες τιμές του Anderson-Darling κριτηρίου για τον δίπλευρο έλεγχο καλής προσαρμογής.

Επίπεδο σημαντικότητας				
15%	10%	5%	2.5%	1%
1.610	1.933	2.492	3.070	3.880

Παρατήρηση 4.16. Οι D'Agostino and Stephens (1986) στη μονογραφία τους που αφορά τους ελέγχους καλής προσαρμογής παραθέτουν μια σειρά από χρήσιμα συμπεράσματα που έχουν προκύψει από συγκριτικές μελέτες προσομοίωσης ελέγχων προσαρμογής. Ειδικότερα, αναφέρουν ότι οι απλοί έλεγχοι καλής προσαρμογής που βασίζονται στην εμπειρική αθροιστική συνάρτηση κατανομής έχουν συνηθέστερα μεγαλύτερη ισχύ από τον χι-τετράγωνο έλεγχο και αυτό, ίσως, οφείλεται στη διακριτοποίηση των δεδομένων. Παρότι ο έλεγχος των Kolmogorov-Smirnov είναι ο πιο δημοφιλής, είναι συχνά λιγότερο ισχυρός από τον έλεγχο των Anderson-Darling. Τέλος, ο έλεγχος των Anderson-Darling παρότι συμπεριφέρεται παρόμοια με τον έλεγχο των Cramér-Von Mises είναι πιο ισχυρός στον εντοπισμό αποκλίσεων από τη μηδενική υπόθεση στις ουρές και για αυτό προτείνεται μεταξύ των ελέγχων που στηρίζονται στην εμπειρική αθροιστική συνάρτηση σε πρακτικές εφαρμογές.

Οι έλεγχοι Cramér-Von Mises, Watson, Kuiper και Anderson-Darling, όπως και ο έλεγχος των Kolmogorov-Smirnov, είχαν αρχικά αναπτυχθεί για τον απλό έλεγχο καλής προσαρμογής. Όμως, στην πράξη, πολύ συχνά δεν επιθυμούμε να ελέγξουμε αν τα δεδομένα προέρχονται από μία πλήρως καθορισμένη κατανομή, δηλαδή από μία κατανομή με συγκεκριμένες παραμέτρους, αλλά αν προέρχονται από μία κατανομή με κάποια/ες ή και όλες τις παραμέτρους άγνωστες. Σε αυτήν την περίπτωση, με παρόμοιο σκεπτικό με αυτό που αναπτύχθηκε στην προηγούμενη υποενότητα για το κριτήριο των Kolmogorov-Smirnov, μπορεί να εκτιμηθεί η p-τιμή του ελέγχου με μεθόδους προσομοίωσης. Στη βιβλιογραφία, έχουν παρουσιαστεί σύνθετοι έλεγχοι καλής προσαρμογής για συγκεκριμένες περιπτώσεις κατανομών, όπως είναι η κανονική κατανομή, η εκθετική κατανομή, η κατανομή Weibull. Οι έλεγχοι αυτοί βασίζονται σε κατάλληλες τροποποιήσεις τόσο των στατιστικών συναρτήσεων όσο και των κρίσιμων τιμών. Ο/Η ενδιαφερόμενος/μενη αναγνώστης/στρια παραπέμπεται στις εργασίες των Pearson and Hartley (1972), Stephens (1974), στη μονογραφία των D'Agostino and Stephens (1986), καθώς και στις εκεί αναφορές. Στη συνέχεια, θα περιοριστούμε στην παράθεση των αποτελεσμάτων για την κανονική και εκθετική κατανομή.

Έλεγχος καλής προσαρμογής της κανονικής κατανομής με άγνωστες παραμέτρους

Θέλουμε να ελέγξουμε αν το τ.δ. $X_1, X_2, \dots, X_n$ προέρχεται από την κανονική κατανομή, με άγνωστη μέση τιμή μ και άγνωστη διασπορά σ^2. Υπολογίζουμε, αρχικά, τις τυποποιημένες τιμές $Z_1, Z_2, \dots, Z_n$ του τ.δ. χρησιμοποιώντας τη σχέση (4.15). Στη συνέχεια, υπολογίζουμε τις στατιστικές συναρτήσεις

$$W_n^2 = \sum_{i=1}^{n} [\Phi\left(Z_{(i)}\right) - \{(2i-1)/2n\}]^2 + \frac{1}{12n},$$

$$U_n^2 = W_n^2 - n(\bar{F}_0 - 0.5)^2, \text{ με } \bar{F}_0 = \frac{\sum_{i=1}^{n} \Phi(Z_i)}{n}$$

$$V_n = T_n^+ + T_n^-$$

και

$$A_n^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} (\log t_i + \log(1 - t_{n+1-i}))$$

όπου $t_i = \Phi(Z_{(i)})$, $i = 1,2,\dots,n$, και T_n^+, T_n^- οι στατιστικές συναρτήσεις που δόθηκαν στις σχέσεις (4.16) και (4.17), αντίστοιχα. Στον Πίνακα 4.8 δίνονται η προτεινόμενη τροποποίηση και οι αντίστοιχες κρίσιμες τιμές για κάθε έλεγχο (βλ. Pearson and Hartley, 1972; Stephens, 1974). Η μηδενική υπόθεση απορρίπτεται, σε επίπεδο σημαντικότητας a, για τιμές μεγαλύτερες από αυτήν του πίνακα.

Πίνακας 4.8: Τροποποιήσεις και κρίσιμες τιμές των τροποποιημένων Cramér-Von Mises, Watson, Kuiper και Anderson-Darling κριτηρίων για τον έλεγχο της κανονικότητας με άγνωστες παραμέτρους.

	Επίπεδο σημαντικότητας				
Τροποποίηση	15%	10%	5%	2.5%	1%
$V_n(\sqrt{n} + 0.05 + 0.82/\sqrt{n})$	1.320	1.386	1.459	1.585	1.693
$W_n^2(1 + 0.5/n)$	0.091	0.104	0.126	0.148	0.179
$U_n^2(1 + 0.5/n)$	0.085	0.096	0.117	0.136	0.164
$A_n^2(1 + 0.75/n + 2.25/n^2)$	0.561	0.631	0.752	0.873	1.035

Παράδειγμα 4.14. Ένα τυχαίο δείγμα εννέα φοιτητών στο πλαίσιο μερικής απασχόλησής τους κατά τη διάρκεια του καλοκαιριού είχε τις ακόλουθες αποδοχές (σε €): 300 120 210 420 600 240 390 90 270. Μπορεί να θεωρηθεί εύλογη η υπόθεση ότι ο πληθυσμός των «θερινών» εισοδημάτων είναι κανονικός; Χρησιμοποιήστε τον έλεγχο των Anderson-Darling.

Λύση Παραδείγματος 4.14. Σύμφωνα με τη θεωρία που προηγήθηκε, θα πρέπει να υπολογίσουμε αρχικά τις τυποποιημένες τιμές $Z_1,\dots,Z_n$ από τη σχέση (4.15). Για τον λόγο αυτόν υπολογίζουμε από τα δεδομένα που μας δίνονται ότι:

$$\bar{X} = 2640/9 = 293.33 \text{ και } S' = \sqrt{25150} = 158.59.$$

Έπειτα εφαρμόζουμε τον μετασχηματισμό της σχέσης (4.15) και, αφού διατάξουμε τις τιμές κατά αύξουσα τάξη μεγέθους, υπολογίζουμε τις τιμές $t_i = \Phi(Z_(i))$, $i = 1,\dots,9$. Ο υπολογισμός αυτός γίνεται με τη βοήθεια είτε του Πίνακα Π.1 του Παραρτήματος είτε της συνάρτησης pnorm της R. Τότε προκύπτει ο πίνακας που ακολουθεί:

i	$X_{(i)}$	$Z_{(i)}$	$t_i = \Phi(Z_{(i)})$
1	90	-1.28	0.100
2	120	-1.09	0.138
3	210	- 0.53	0.298
4	240	- 0.34	0.367
5	270	- 0.15	0.440
6	300	0.04	0.516
7	390	0.61	0.729
8	420	0.80	0.788
9	600	1.93	0.973

Με τη βοήθεια του παραπάνω πίνακα εύκολα υπολογίζουμε το A_n^2. Πιο συγκεκριμένα:

$$\begin{aligned} A_n^2 &= -n - \sum_{i=1}^{n} \frac{2i-1}{n} (\log t_i + \log(1 - t_{n+1-i})) \\ &= -\frac{1}{9}[\ln(0.1) + \ln(1-0.973)] - \frac{3}{9}[\ln(0.138) + \ln(1-0.788)] \\ &- \frac{5}{9}[\ln(0.298) + \ln(1-0.729)]\frac{7}{9}[\ln(0.367) + \ln(1-0.516)] \\ &- \frac{9}{9}[\ln(0.44) + \ln(1-0.44)] - \frac{11}{9}[\ln(0.516) + \ln(1-0.367)] \\ &- \frac{13}{9}[\ln(0.729) + \ln(1-0.298)] - \frac{15}{9}[\ln(0.788) + \ln(1-0.138)] \\ &- \frac{17}{9}[\ln(0.973) + \ln(1-0.1)] - 9 \\ &= 0.209. \end{aligned}$$

Επομένως, $A^* = A_n^2\left(1 + \frac{0.75}{9} + \frac{2.25}{9^2}\right) = 0.232$. Για $a = 0.05$, από τον Πίνακα 4.8 έχουμε ότι η κρίσιμη τιμή είναι $c = 0.752$, η οποία είναι μεγαλύτερη της τιμής του A^*. Άρα, σε επίπεδο σημαντικότητας 5%, δεν απορρίπτεται η H_0, δηλαδή δεν απορρίπτεται η υπόθεση ότι τα δεδομένα προέρχονται από την κανονική κατανομή. Τα παραπάνω υλοποιούνται στην R με τις εντολές που ακολουθούν.

```
data<-c(300,120,210,420,600,240,390,90,270)
zorderdata<-sort((data-mean(data))/sqrt(var(data)))
tdata<-pnorm(zorderdata,0,1)
voithad<-0
for (i in 1:length(data)){voithad[i]<-(2*i-1)/length(data)*(log(tdata[
   i])+log(1-tdata[length(data)+1-i]))}
ad<- -length(data)-sum(voithad)
ad
adstar<-ad*(1+0.75/length(data)+2.25/((length(data))^2))
adstar
```

□

Έλεγχος καλής προσαρμογής της εκθετικής κατανομής με άγνωστη παράμετρο

Θέλουμε να ελέγξουμε αν το τ.δ. $X_1, X_2, \ldots, X_n$ προέρχεται από την εκθετική κατανομή, με άγνωστη μέση τιμή θ, δηλαδή από την κατανομή με αθροιστική συνάρτηση κατανομής

$$F(x) = 1 - \exp\left(-\frac{x}{\theta}\right),$$

για $x > 0$, και μηδέν αλλού. Υπολογίζουμε, αρχικά, τις τυποποιημένες τιμές $Z_1, Z_2, \ldots, Z_n$, που ορίζονται στη σχέση (4.18) και έπειτα υπολογίζουμε τις στατιστικές συναρτήσεις:

$$W_n^2 = \sum_{i=1}^{n} [F_0\left(Z_{(i)}\right) - \{(2i-1)/2n\}]^2 + \frac{1}{12n},$$

$$U_n^2 = W_n^2 - n(\bar{F}_0 - 0.5)^2, \text{ με } \bar{F}_0 = \frac{\sum_{i=1}^{n} F_0(Z_i)}{n},$$

$$V_n = K_n^+ + K_n^-,$$

και

$$A_n^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} (\log t_i + \log(1 - t_{n+1-i})),$$

όπου $t_i = F_0(Z_{(i)})$, $i = 1, 2, \ldots, n$, K_n^+, K_n^- δίνονται στις σχέσεις (4.19) και (4.20), αντίστοιχα, ενώ $F_0(x) = 1 - \exp(-x)$, για $x > 0$ και μηδέν αλλού. Στον Πίνακα 4.9 δίνονται η προτεινόμενη τροποποίηση και οι αντίστοιχες κρίσιμες τιμές για κάθε έλεγχο (βλ. Pearson and Hartley, 1972; Stephens, 1974). Η μηδενική υπόθεση απορρίπτεται, σε επίπεδο σημαντικότητας a, για τιμές μεγαλύτερες από αυτήν του πίνακα.

Πίνακας 4.9: Τροποποιήσεις και κρίσιμες τιμές των τροποποιημένων Cramér-Von Mises, Watson, Kuiper και Anderson-Darling κριτηρίων για τον έλεγχο της εκθετικής κατανομής με άγνωστη παράμετρο.

	Επίπεδο σημαντικότητας				
Τροποποίηση	15%	10%	5%	2.5%	1%
$(V_n - 0.2/n)(\sqrt{n} + 0.24 + 0.35/\sqrt{n})$	1.445	1.527	1.655	1.774	1.910
$W_n^2(1 + 0.16/n)$	0.148	0.175	0.222	0.271	0.338
$U_n^2(1 + 0.16/n)$	0.112	0.129	0.159	0.189	0.230
$A_n^2(1 + 0.6/n)$	0.916	1.062	1.321	1.591	1.959

4.4 Η ειδική περίπτωση της κανονικής κατανομής

Στις προηγούμενες δύο ενότητες αυτού του κεφαλαίου παρουσιάστηκαν ο έλεγχος χι-τετράγωνο καλής προσαρμογής και δημοφιλείς έλεγχοι που στηρίζονται στην εμπειρική αθροιστική συνάρτηση κατανομής. Η επιλογή αυτή, στο πλαίσιο ενός συγγράμματος που απευθύνεται σε προπτυχιακούς φοιτητές/τριες, είναι πλήρως αιτιολογημένη, καθώς από τη μια πλευρά ο έλεγχος χι-τετράγωνο καλής προσαρμογής είναι ο παλαιότερος έλεγχος καλής προσαρμογής και ιδιαίτερα γνωστός για διακριτά δεδομένα, ενώ από την άλλη πλευρά θέλαμε να αξιοποιηθούν τα αποτελέσματα του Κεφαλαίου 2 που αφορούσαν την εκτίμηση της αθροιστικής συνάρτησης κατανομής, τις ιδιότητές της και, ιδιαίτερα, το Θεώρημα των Glivenko-Cantelli.

Ωστόσο, στη βιβλιογραφία έχουν παρουσιαστεί ποικίλοι άλλοι τρόποι για τον έλεγχο καλής προσαρμογής τόσο στη γενική περίπτωση (είτε απλού είτε σύνθετου ελέγχου) όσο και σε ειδικές περιπτώσεις κατανομών. Η έρευνα σε αυτό το αντικείμενο είναι πάρα πολύ μεγάλη και έχουν εμφανιστεί τόσο γραφικοί όσο και στατιστικοί τρόποι ελέγχου καλής προσαρμογής. Για παράδειγμα, έχουν εμφανιστεί στατιστικοί τρόποι ελέγχου οι οποίοι για τη συνεχή περίπτωση στηρίζονται σε μέτρα αποκλίσεων μεταξύ της εμπειρικής χαρακτηριστικής συνάρτησης και της χαρακτηριστικής συνάρτησης της υπό έλεγχο κατανομής, ενώ για τη διακριτή περίπτωση σε μέτρα αποκλίσεων μεταξύ της εμπειρικής πιθανογεννήτριας συνάρτησης και της πιθανογεννήτριας συνάρτησης της υπό έλεγχο κατανομής. Επιπρόσθετα, έχουν εμφανιστεί στη βιβλιογραφία έλεγχοι που βασίζονται σε χαρακτηρισμούς και ιδιότητες της υπό έλεγχο κατανομής κ.ο.κ. Ειδικότερα, για κάποιες πολύ δημοφιλείς κατανομές, όπως είναι η κανονική, η εκθετική, η Poisson, ο αριθμός των ελέγχων που έχουν προταθεί είναι ιδιαίτερα μεγάλος, καθώς οι τρόποι απόκλισης από μια δοθείσα κατανομή είναι αμέτρητοι και αυτό έχει ως αποτέλεσμα να μην μπορεί να υπάρξει ένας έλεγχος που θα είναι για όλες τις εναλλακτικές κατανομές ο πλέον ισχυρός.

Καθώς ο ρόλος της κανονικής κατανομής είναι πρωταρχικός στη Θεωρία Πιθανοτήτων και στη Στατιστική, αποτελώντας η υπόθεση της κανονικότητας προϋπόθεση για να ισχύει πλήθος στατιστικών μεθοδολογιών, στην ενότητα αυτή θα σχολιαστούν, χωρίς να υπεισέλθουμε σε λεπτομέρειες, οι έλεγχοι κανονικότητας, οι οποίοι θα ταξινομηθούν σε δύο μεγάλες κατηγορίες: τους γραφικούς και τους στατιστικούς.

4.4.1 Γραφικοί τρόποι ελέγχου της κανονικότητας

Υπάρχουν διάφοροι γραφικοί τρόποι ελέγχου της υπόθεσης της κανονικότητας γραφικά. Ειδικότερα, ο πιο απλός και εύκολος τρόπος είναι με το ιστόγραμμα. Από το σχήμα του ιστογράμματος μπορούμε να έχουμε μια πρώτη άποψη για το αν τα δεδομένα που εξετάζουμε αποκλίνουν από την κανονική κατανομή, αν δεν παρατηρείται το σχήμα καμπάνας το οποίο αναμένουμε. Η βασική γραφική μέθοδος για τον έλεγχο της κανονικότητας είναι το Q-Q (quantile-quantile) γράφημα, το οποίο συγκρίνει τα ποσοστιαία σημεία (quantile) του δείγματος έναντι των πληθυσμιακών ποσοστιαίων σημείων της κανονικής κατανομής. Αν τα σημεία είναι κοντά σε ευθεία γραμμή, τότε δεν υπάρχει ένδειξη για απόκλιση από την κανονικότητα. Παρεκκλίσεις από την ευθεία γραμμή δηλώνουν μη κανονικότητα και ο τύπος της μη γραμμικότητας μπορεί να υποδηλώνει τον τρόπο απόκλισης από την κανονικότητα. Ένας επιπλέον τρόπος ελέγχου της κανονικότητας επιτυγχάνεται με τα γραφήματα τύπου P-P (probability-probability). Τα γραφήματα αυτά παριστάνουν σε σύστημα οριζόντιων και κάθετων αξόνων τις τιμές της παρατηρούμενης αθροιστικής συχνότητας (άξονας των x) και της αθροιστικής συχνότητας της υποτιθέμενης κατανομής που ακολουθεί η μεταβλητή που εξετάζεται. Όσο πιο μακριά από τη διχοτόμο της γωνίας των αξόνων είναι συγκεντρωμένα τα σημεία τόσο περισσότερο ενισχύεται η υπόθεση ότι τα δεδομένα αποκλίνουν από την κανονική κατανομή. Τέλος, ένας άλλος γνωστός, εναλλακτικός στους προηγούμενους, τρόπος γραφικού ελέγχου της κανονικότητας επιτυγχάνεται με το λεγόμενο detrended Q-Q γράφημα, όπου απεικονίζονται οι διαφορές μεταξύ των παρατηρούμενων και αναμενόμενων τιμών υπό την υπόθεση της κανονικότητας. Αν τα δεδομένα προέρχονται από την κανονική κατανομή, τότε τα σημεία θα πρέπει να συγκεντρώνονται γύρω από μια οριζόντια γραμμή κοντά στο μηδέν, χωρίς κάποιο ιδιαίτερο μοτίβο.

Οι γραφικοί έλεγχοι έχουν το μειονέκτημα ότι προϋποθέτουν την εμπειρία του/της αναλυτή/αναλύτριας για την κατάλληλη ερμηνεία τους. Για τον λόγο αυτόν πολλές φορές είναι προτιμότερο να χρησιμοποιούνται οι στατιστικοί τρόποι ελέγχου, που γίνονται στη βάση της p- τιμής του ελέγχου.

4.4.2 Στατιστικοί τρόποι ελέγχου της κανονικότητας - Ο έλεγχος Shapiro-Wilk

Στην προηγούμενη ενότητα παρουσιάστηκε ο τρόπος ελέγχου της κανονικότητας με τα κριτήρια των Kolmogorov-Smirnov, Kuiper, Watson, Cramér-Von Mises και Anderson-Darling. Ειδικότερα, ιδιαίτερη μνεία έγινε στις τροποποιήσεις αυτών των στατιστικών συναρτήσεων, για να αντιμετωπιστεί το πρόβλημα που προκύπτει όταν οι παράμετροι της κανονικής κατανομής είναι άγνωστες, που είναι και το πιο σύνηθες. Εκτός, όμως, από τους πέντε αυτούς στατιστικούς ελέγχους, όπως αναφέρει σε μια πρόσφατη εργασία του ο Bayoud (2021), έχουν παρουσιαστεί στη βιβλιογραφία περισσότεροι από σαράντα διαφορετικοί τρόποι ελέγχου της κανονικότητας. Μεταξύ αυτών, οι έλεγχοι που κατέχουν εξέχουσα θέση στη βιβλιογραφία, χρησιμοποιούνται πιο συχνά σε πρακτικές εφαρμογές και είναι διαθέσιμοι στα διάφορα στατιστικά προγράμματα είναι οι: Shapiro-Wilk, Shapiro Francia, Jarque-Bera, D'Agostino skewness, Anscombe-Glynn kurtosis test και D'Agostino-Pearson omnibus test, για να αναφέρουμε απλώς ορισμένους. Με συγκριτικές μελέτες προσομοίωσης έχει προκύψει ότι δεν υπάρχει ένα τεστ το οποίο να είναι βέλτιστο για όλες τις πιθανές αποκλίσεις από την κανονική κατανομή (βλ. Bayoud, 2021; Wijekularathna *et al.*, 2022, και τις εκεί αναφορές). Επομένως, το να προτείνει κάποιος έναν μοναδικό τρόπο ελέγχου της κανονικότητας δεν είναι κάτι που μπορεί να γίνει. Όμως, ένα γενικό συμπέρασμα είναι ότι οι έλεγχοι των Shapiro-Wilk και των Anderson Darling έχουν ικανοποιητική απόδοση, με τον έλεγχο των Shapiro-Wilk να έχει τη μεγαλύτερη ισχύ σε σύγκριση με αυτήν των υπολοίπων για μη συμμετρικές εναλλακτικές κατανομές, ενώ οι έλεγχοι των Kolmogorov Smirnov και χι-τετράγωνο του Pearson δεν έχουν τόσο καλή απόδοση. Για τον λόγο αυτόν, στην υποενότητα αυτή, θα παρουσιαστεί εν συντομία και στο πλαίσιο ενός προπτυχιακού μαθήματος ο έλεγχος των Shapiro-Wilk. Ο έλεγχος αυτός προτάθηκε από τον Αμερικανό στατιστικό Samuel Sanford Shapiro (1930-) και τον Καναδό στατιστικό Martin Bradbury Wilk (1922 – 2013), στην εργασία τους Shapiro and Wilk (1965) και από τότε έχει αποτελέσει αντικείμενο

μελέτης, γενίκευσης, επέκτασης και βελτίωσης τόσο από τους Shapiro και Wilk όσο και από άλλους συγγραφείς (βλ., μεταξύ άλλων, D'Agostino and Stephens, 1986, και τις εκεί αναφορές).

Ας είναι X μία συνεχής τυχαία μεταβλητή και $X_1, X_2, \ldots, X_n$, ένα δείγμα τιμών της. Για τον έλεγχο της υπόθεσης

$$H_0 : X_1, X_2, \ldots, X_n \text{ προέρχονται από κανονικό πληθυσμό } \mathcal{N}(\mu, \sigma^2), \text{ με } \mu, \sigma^2 \text{ άγνωστες,}$$

ο έλεγχος των Shapiro-Wilk βασίζεται στη στατιστική συνάρτηση:

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2},$$

η οποία, ισοδύναμα, εκφράζεται στη μορφή:

$$W = \frac{\left[\sum_{i=1}^{k} a_i (X_{(n-i+1)} - X_{(i)})\right]^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2},$$

όπου a_i, $i = 1, \ldots, n$ είναι σταθεροί συντελεστές που δημιουργούνται από τους μέσους, τις διακυμάνσεις και τις συνδιακυμάνσεις των διατεταγμένων στατιστικών ενός δείγματος μεγέθους n από την κανονική και $k \simeq n/2$. Καθώς ισχύει ότι $a_i = -a_{n+1-i}$, έχουμε επίσης ότι:

$$W = \frac{\left[\sum_{i=1}^{k} a_{n-i+1} (X_{(n-i+1)} - X_{(i)})\right]^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}.$$

Οι τιμές των a_{n-i+1} για $2 \leqslant n \leqslant 50$ δίνονται στον πίνακα 5 της εργασίας των Shapiro and Wilk (1965). Απόσπασμα των τιμών των συντελεστών δίνεται στον Πίνακα 4.10.

Πίνακας 4.10: Συντελεστές του Shapiro-Wilk κριτηρίου.

Μέγεθος δείγματος n	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$	$i=9$	$i=10$
2	0.7071									
3	0.7071	0.0000								
4	0.6872	0.1677								
5	0.6646	0.2413	0.0000							
6	0.6431	0.2806	0.0875							
7	0.6233	0.3031	0.1401	0.0000						
8	0.6052	0.3164	0.1743	0.0561						
9	0.5888	0.3244	0.1976	0.0947	0.0000					
10	0.5739	0.3291	0.2141	0.2141	0.1224	0.0399				
11	0.5601	0.3315	0.2260	0.1429	0.0695	0.0000				
12	0.5475	0.3325	0.2347	0.1586	0.0922	0.0303				
13	0.5359	0.3325	0.2412	0.1707	0.1099	0.0539	0.0000			
14	0.5251	0.3318	0.2460	0.1802	0.1240	0.0727	0.0240			
15	0.5150	0.3306	0.2495	0.1878	0.1353	0.0880	0.0433	0.0000		
16	0.5056	0.3290	0.2521	0.1939	0.1447	0.1005	0.0593	0.0196		
17	0.4968	0.3273	0.2540	0.1988	0.1524	0.1109	0.0725	0.0359	0.0000	
18	0.4886	0.3253	0.2553	0.2027	0.1587	0.1197	0.0837	00.0496	0.0163	
19	0.4808	0.3232	0.2561	0.2059	0.1641	0.1271	0.0932	0.0612	0.0303	0.0000
20	0.4737	0.3211	0.2565	0.2085	0.1686	0.1334	0.1013	0.0711	0.0422	0.0140

Μικρές τιμές του W υποδεικνύουν μη κανονική κατανομή, ενώ τιμές κοντά στη μονάδα δεν μας υποδεικνύουν αποκλίσεις από την κανονικότητα. Επομένως, η κρίσιμη περιοχή είναι της μορφής $W \leqslant w_{1-a}$, όπου w_{1-a} είναι τέτοιο ώστε $P(W \leqslant w_{1-a}|H_0) = a$ ή, ισοδύναμα, $P(W \geqslant w_{1-a}|H_0) = 1 - a$. Ποσοστιαία σημεία για τη μηδενική κατανομή του W συνοψίζονται στην εργασία των Shapiro and Wilk (1965) για p=0.01, 0.02, 0.05, 0.1, 0.5, 0.9, 0.95, 0.98, 0.99 και για δείγματα μεγέθους $n = 3, 4, ..., 50$. Απόσπασμα αυτών των τιμών δίνεται στον Πίνακα Π.16 του Παραρτήματος. Τέλος, ο Royston (1982) επέκτεινε τη στατιστική συνάρτηση W για δείγματα μεγέθους $n \geqslant 50$ και ανέπτυξε έναν μετασχηματισμό για την κατανομή του W υπό την H_0, για να προσεγγίσει την κανονικότητα για μεγέθη δείγματος n, με $7 \leqslant n \leqslant 2000$, που, στη συνέχεια, επεκτάθηκε από τους Rahman and Govidarajulu (1997) μέχρι μέγεθος δείγματος 5000. Ακόμη, ο Royston (1982) πρότεινε μία μέθοδο για τον υπολογισμό του επιπέδου σημαντικότητας για $n < 7$. Με αυτόν τον τρόπο επιτεύχθηκε η ανάπτυξη του W τεστ σε μία μορφή, η οποία μπορεί εύκολα να προγραμματιστεί σε υπολογιστή, για ένα εύρος δειγμάτων μεγέθους $3 < n \leqslant 5000$.

Παράδειγμα 4.15. Χρησιμοποιώντας τον έλεγχο που προτάθηκε από τους Shapiro-Wilk ελέγξτε, με επίπεδο σημαντικότητας 5%, αν τα δεδομένα του Παραδείγματος 4.9 προέρχονται από κανονική κατανομή. Υπόδειξη: Για διευκόλυνση στους υπολογισμούς χρησιμοποιήστε την `R`.

Λύση Παραδείγματος 4.15. Έχουμε ένα τυχαίο δείγμα $n = 12$ το πλήθος παρατηρήσεων και θέλουμε να ελέγξουμε, με επίπεδο σημαντικότητας 5%, χρησιμοποιώντας το κριτήριο Shapiro-Wilk, κατά πόσο τα δεδομένα αυτά προέρχονται από κανονική κατανομή. Οι διατεταγμένες παρατηρήσεις του δείγματος είναι οι: 6900, 7200, 8600, 8700, 9300, 9600, 9800, 10200, 11600, 12200, 15200, 15500.

Ο υπολογισμός της στατιστικής συνάρτησης των Shapiro-Wilk θα γίνει από τη σχέση:

$$W = \frac{\left[\sum_{i=1}^{k} a_i (X_{(n-i+1)} - X_{(i)})\right]^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2},$$

για $n = 12$ και $k = 6$.

Υπολογίζουμε ότι $\bar{X} = 10400$ και $\sum_{i=1}^{n} (X_i - \bar{X})^2 = 84600000$. Για να υπολογίσουμε τον αριθμητή, έχουμε ότι:

$$\begin{aligned} \sum_{i=1}^{6} a_i (X_{(n-i+1)} - X_{(i)}) &= a_1 \cdot (X_{(12)} - X_{(1)}) + a_2 \cdot (X_{(11)} - X_{(2)}) + a_3 \cdot (X_{(10)} - X_{(3)}) \\ &+ a_4 \cdot (X_{(9)} - X_{(4)}) + a_5 \cdot (X_{(8)} - X_{(5)}) + a_6 \cdot (X_{(7)} - X_{(6)}), \end{aligned}$$

όπου οι τιμές των συντελεστών a_i, $i = 1, .., 6$ προσδιορίζονται στον Πίνακα 4.10 (βλ. τη γραμμή $n = 12$). Επομένως, είναι:

$$\begin{aligned} \sum_{i=1}^{6} a_i (X_{(n-i+1)} - X_{(i)}) &= 0.5475 \cdot (15500 - 6900) + 0.3325 \cdot (15200 - 7200) \\ &+ 0.2347 \cdot (12200 - 8600) + 0.1586 \cdot (11600 - 8700) \\ &+ 0.0922 \cdot (10200 - 9300) + 0.0303 \cdot (9800 - 9600) \\ &= 4708.5 + 2660 + 844.92 + 459.94 + 82.98 + 6.06 = 8762.4 \end{aligned}$$

Συνεπώς, είναι $W = (8762.4)^2/(84600000) = 0.9075609$ και η μηδενική υπόθεση απορρίπτεται αν $W \leqslant w_{0.95}$, με την τιμή $w_{0.95}$ να προσδιορίζεται στον Πίνακα Π.16 του Παραρτήματος. Από αυτόν τον πίνακα έχουμε ότι για $n = 12$ είναι $w_{0.95} = 0.859$. Άρα, η μηδενική υπόθεση δεν απορρίπτεται σε επίπεδο σημαντικότητας 5%, δηλαδή δεν απορρίπτεται, σε επίπεδο σημαντικότητας 5%, η υπόθεση ότι τα δεδομένα προέρχονται από πληθυσμό που περιγράφεται από την κανονική κατανομή. □

4.5 Ασκήσεις

Άσκηση 4.1. Οι συχνότητες εμφάνισης των εδρών ενός ζαριού μετά από 310 διαδοχικές ρίψεις δίνονται στον παρακάτω πίνακα:

Έδρα	1	2	3	4	5	6
Συχνότητα Εμφάνισης	38	61	54	65	55	37

Με βάση τα δεδομένα αυτά, να ελέγξετε την υπόθεση ότι το ζάρι είναι δίκαιο. **Υπόδειξη:** Να χρησιμοποιήσετε τον χι-τετράγωνο έλεγχο καλής προσαρμογής και ε.σ. 5%.

Άσκηση 4.2. Στην προηγούμενη εξέταση ενός μαθήματος Στατιστικής εμφανίστηκαν 40 φοιτητές, εκ των οποίων οι 10 ήταν τριτοετείς, 10 τεταρτοετείς και οι υπόλοιποι επί πτυχίω. Ελέγξτε, σε ε.σ. 5%, αν η αναλογία των φοιτητών, ανά έτος, που δίνουν το μάθημα Στατιστικής είναι 0.2, 0.3 και 0.5, αντίστοιχα.

Άσκηση 4.3. Δίνονται τα διαστήματα $\Delta_1 = (0,2]$, $\Delta_2 = (2,4]$, $\Delta_3 = (4,6]$ και $\Delta_4 = (6,\infty)$. Τυχαίο δείγμα $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$ από μια κατανομή στο $(0,\infty)$ έδωσε, αντίστοιχα, τις παρακάτω συχνότητες, 210, 110, 50, 30. Να ελεγχθεί, σε ε.σ. 5%, αν η κατανομή από την οποία προέρχεται το δείγμα είναι η εκθετική με παράμετρο 2.

Άσκηση 4.4. Χρησιμοποιήστε τον χι-τετράγωνο έλεγχο καλής προσαρμογής και ελέγξτε, σε ε.σ. 10%, την υπόθεση ότι τα παρακάτω δεδομένα προέρχονται από την $Hg(10,10,7)$ (Υπεργεωμετρική κατανομή). Τα δεδομένα δίνονται σε μορφή πίνακα συχνοτήτων και αφορούν συνολικά 150 πραγματοποιήσεις επιθεωρήσεων $n = 7$ εξαρτημάτων, τα οποία λαμβάνονται, χωρίς επανατοποθέτηση, από έναν σωρό 20 συνολικά εξαρτημάτων. Σε κάθε επιθεώρηση καταγράφεται το πλήθος των ελαττωματικών εξαρτημάτων, μεταξύ των 7 του δείγματος, και παρατίθεται στον πίνακα που ακολουθεί.

Αριθμός Ελαττωματικών	Πλήθος Επιθεωρήσεων
0	0
1	16
2	38
3	56
4	33
5	7
6	0
7	0

Άσκηση 4.5. Πήραμε ένα δείγμα από 60 δίτεκνες οικογένειες και καταγράψαμε τον αριθμό των αγοριών που υπάρχουν σε αυτές, με τα δεδομένα μας να δίνονται στον παρακάτω πίνακα.

Πλήθος αγοριών (i)	0	1	2
Πλήθος οικογενειών με i αγόρια	15	25	20

Σε μια παλαιότερη έρευνα είχε υποστηριχθεί ότι, αν $1-p$ είναι η πιθανότητα να μην υπάρχει αγόρι στη δίτεκνη οικογένεια, τότε p^2 είναι η πιθανότητα να υπάρχει ένα αγόρι και $p(1-p)$ να υπάρχουν 2 αγόρια. Χρησιμοποιώντας έναν χι-τετράγωνο έλεγχο καλής προσαρμογής και με βάση τα δεδομένα ελέγξτε, σε ε.σ. 1%, αν επαληθεύεται η έρευνα.

Άσκηση 4.6. Με τη χρησιμοποίηση του κριτηρίου Kolmogorov-Smirnov να ελεγχθεί, με επίπεδο σημαντικότητας 1%, η υπόθεση ότι οι παρακάτω μετρήσεις προέρχονται από κανονική κατανομή με μέση τιμή $\mu = 32$ και διακύμανση $\sigma^2 = 3.24$.

31.4, 31.4, 33.3, 33.5, 33.7, 34.4, 34.9, 36.2, 37.0

Άσκηση 4.7. Με τη χρησιμοποίηση του κριτηρίου Kolmogorov-Smirnov να ελεγχθεί, με επίπεδο σημαντικότητας 5%, η υπόθεση ότι οι παρακάτω μετρήσεις προέρχονται από την ομοιόμορφη κατανομή $\mathscr{U}(0,1)$.

0.621, 0.503, 0.203, 0.477, 0.710, 0.581, 0.329, 0.480, 0.554, 0.382.

Άσκηση 4.8. Τυχαίο δείγμα 10 παρατηρήσεων στο $(0,1)$ έδωσε τις εξής τιμές:

$$0.54, 0.72, 0.85, 0.41, 0.22, 0.67, 0.78, 0.90, 0.64, 0.95.$$

Να ελεγχθεί, με επίπεδο σημαντικότητας 10%, η μηδενική υπόθεση $H_0 : F(x) = F_0(x)$, για κάθε $x \in \mathbb{R}$, έναντι της εναλλακτικής $H_1 : F(x) \neq F_0(x)$ για κάποιο $x \in \mathbb{R}$, όπου $F_0(x)$ είναι η συνάρτηση κατανομής της τυχαίας μεταβλητής με συνάρτηση πυκνότητας πιθανότητας $f_0(x) = 2x$, $0 < x < 1$. Υπόδειξη: να χρησιμοποιηθεί το κριτήριο Kolmogorov-Smirnov και Anderson-Darling.

Άσκηση 4.9. Να ελεγχθεί, με επίπεδο σημαντικότητας 5%, αν το παρακάτω τυχαίο δείγμα 12 παρατηρήσεων (διατεταγμένων κατά αύξουσα σειρά) μπορεί να θεωρηθεί ότι αποτελεί δείγμα παρατηρήσεων πάνω σε μία τυχαία μεταβλητή X, της οποίας η κατανομή είναι κανονική με μέση τιμή 30 και διασπορά 100.

$$18.8 \;\; 19.3 \;\; 22.4 \;\; 22.5 \;\; 24.0 \;\; 24.7 \;\; 25.9 \;\; 27.0 \;\; 35.1 \;\; 35.8 \;\; 36.5 \;\; 37.6$$

Για τον έλεγχο να χρησιμοποιηθεί το κριτήριο Kolmogorov-Smirnov, καθώς και το κριτήριο Anderson-Darling.

Άσκηση 4.10. Χορηγούνται δύο διαφορετικά παυσίπονα σε δύο ομάδες ασθενών και μετράμε τις ώρες ανακούφισης, με τα αποτελέσματα να φαίνονται στον παρακάτω πίνακα,

Παυσίπονο 1	6.8	3.1	5.8	4.5	4.2	4.9	3.3	4.7
Παυσίπονο 2	4.4	2.5	2.8	2.1	6.6	0.0	4.8	2.3

Να ελεγχθεί, με επίπεδο σημαντικότητας 5%, αν τα δύο παυσίπονα είναι το ίδιο αποτελεσματικά.

Άσκηση 4.11. Στη διάθεσή μας έχουμε τα παρακάτω τυχαία δείγματα τα οποία αφορούν το βάρος μπαταριών τύπου LR6, οι οποίες έχουν παραχθεί από 2 διαφορετικές εταιρείες (υποθέστε ότι οι παραγωγικές διαδικασίες για την κατασκευή των μπαταριών των 2 εταιρειών, είναι μεταξύ τους ανεξάρτητες).

AA	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X	2.4	2.6	2.9	3.2	3.3	3.4	3.7	3.8	4.0	4.2	4.8	4.9	5.3	6.7	7.6	7.8
Y	1.1	1.8	2.4	2.4	2.6	2.9	3.0	3.2	5.5	6.4	6.7	7.4	8.1	11.0	12.0	13.2

(i) Χρησιμοποιήστε κατάλληλο τεστ και ελέγξετε την υπόθεση (ε.σ. 5%) ότι οι κατανομές του βάρους των μπαταριών, που παράγονται από τις δύο διαφορετικές εταιρείες, είναι ίδιες.

(ii) Να ελέγξετε την υπόθεση ότι η κατανομή του βάρους των μπαταριών της 1ης εταιρείας (τιμές X) είναι η Ομοιόμορφη $\mathscr{U}(2,8)$.

Άσκηση 4.12. Χρησιμοποιήστε το τροποποιημένο κριτήριο των Anderson-Darling και ελέγξτε την υπόθεση ότι οι παρακάτω μετρήσεις προέρχονται από εκθετική κατανομή. Ο έλεγχος να γίνει σε ε.σ. 10%.

$$0.59,\ 0.32,\ 8.11,\ 6.18,\ 2.3,\ 6.09,\ 4.92,\ 4.07,\ 4.27,\ 1.34,\ 1.4.\ 0.47.$$

Άσκηση 4.13. Χρησιμοποιήστε τον έλεγχο των Shapiro-Wilk και ελέγξτε την υπόθεση ότι οι παρακάτω μετρήσεις προέρχονται από κανονική κατανομή. Ο έλεγχος να γίνει σε επίπεδο σημαντικότητας 5%.

$$42,\ 51,\ 58,\ 53,\ 59,\ 56,\ 52,\ 66,\ 62,\ 68,\ 70,\ 76,\ 53,\ 54,\ 55,\ 58.$$

Άσκηση 4.14. Έστω ότι στη διάθεσή μας έχουμε τυχαίο δείγμα χρόνων ζωής (σε ώρες) ενός ηλεκτρονικού εξαρτήματος. Να ελέγξετε την υπόθεση ότι τα δεδομένα προέρχονται από Λογαριθμοκανονική κατανομή με παραμέτρους $\mu = 6, \sigma = 1.5$

$$275.19,\ 14.87,\ 246.73,\ 655.04,\ 233.38,\ 627.58,\ 468.06,\ 3761.7,\ 448.25,\ 545.7,\ 36.8,\ 470.74.$$

Υπόδειξη: Αν η τ.μ. X ακολουθεί Λογαριθμοκανονική κατανομή με παραμέτρους μ και σ^2, τότε η τ.μ. $Y = \ln X \sim \mathcal{N}(\mu, \sigma^2)$.

Άσκηση 4.15. (Conover, 1998). Πενήντα διψήφιοι αριθμοί επιλέχθηκαν τυχαία από έναν τηλεφωνικό κατάλογο. Οι αριθμοί κατά αύξουσα σειρά μεγέθους είναι οι εξής:

23	23	24	27	29	31	32	33	33	35
36	37	40	42	43	43	44	45	48	48
54	54	56	57	57	58	58	58	58	59
61	61	62	63	64	65	66	68	68	70
73	73	74	75	77	81	87	89	93	97

Να ελεγχθεί η υπόθεση ότι οι αριθμοί αυτοί θα μπορούσαν να αποτελούν παρατηρήσεις πάνω σε μια κανονική τυχαία μεταβλητή. **Υπόδειξη:** Χρησιμοποιείστε τον έλεγχο των Shapiro-Wilk, με επίπεδο σημαντικότητας 5%.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Κούτρας, Μ. (2018). *Εισαγωγή στη Θεωρία Πιθανοτήτων και Εφαρμογές*. Αθήνα: Εκδόσεις Σταμούλη.

Ξενόγλωσση

Anderson, T. (1958). On the Distribution of the Two-Sample Cramér–von Mises Criterion. *Annals of Mathematical Statistics*, *33*, pp. 1148–1159.

Anderson, T. and Darling, D. (1952). Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. *Annals of Mathematical Statistics*, *23*, pp. 193–212.

Bayoud, H. A. (2021). Tests of normality: new test and comparative study. *Communications in Statistics - Simulation and Computation*, *50*, pp. 4442–4463.

Birnbaum, Z. (1952). Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. *Journal of the American Statistical Association*, *47*, pp. 425–441.

Chernoff, H. and Lehmann, E. L. (1954). The use of the maximum likelihood estimates on X^2 tests for goodness of fit. *Am. Math. Statist.*, *38*, pp. 52–72.

Conover, W. J. (1998). *Practical Nonparametric Statistics* (3rd ed.). Wiley and Sons, Inc.

Cramér, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, *1*, pp. 13–74.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

D'Agostino, R. and Stephens, M. (1986). *Goodness-of-Fit Techniques*. New York: Marcel-Dekker.

Durbin, J. (1975). Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponential and tests on spacing. *Biometrika*, *62*, pp. 5–22.

Fisher, R. A. (1924). The Conditions Under Which χ2 Measures the Discrepancy Between Observation and Hypothesis. *Journal of the Royal Statistical Society*, *87*, pp. 442–450.

Gibbons, J. D. and Chakraborti, S. (2020). *Nonparametric Statistical Inference, Fourth Edition Revised and Expanded*. Chapman and Hall/CRC.

Hodges, J. L. (1958). The significance probability of the Smirnov two-sample test. *Ark. Mat.*, *3*, pp. 469–486.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari*, *4*, pp. 83–91.

Kuiper, N. H. (1960). Tests concerning random points on a circle. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Series A.*, *63*, pp. 38–47.

Kvam, P. and Vidakovic, B. (2007). *Nonparametric Statistics with applications to science and engineering*. Wiley Series in Probability and Statistics.

Lewis, P. A. (1961). Distribution of the Anderson-Darling statistic. *The Annals of Mathematical Statistics*, *32*, pp. 1118–1124.

Lilliefors, H. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*, pp. 399–402.

Lilliefors, H. (1969). On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown. *Journal of the American Statistical Association*, *64*, pp. 387–389.

Miller, L. (1956). Tables of the percentage points of Kolmogorov statistics. *Journal of the American Statistical Association*, *51*, pp. 111–121.

Mises, R. von (1928). *Wahrscheinlichkeit, Statistik und Wahrheit*. Julius Springer.

Moore, D. S. (1986). Tests of Chi-Squared Type. In: *Goodness-of-fit Techniques*. Ed. by R. D'Agostino and M. Stephens. New York, NY: Marcel Dekker, pp. 63–95.

Neyman, J. and Pearson, E. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, *20*, pp. 175-240 & 263–294.

Pearson, E. and Hartley, H. (1972). *Biometrika Tables for Statisticians, Volume 2.* Cambridge University Press.

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *50*(302), pp. 157–175.

Rahman, M. M. and Govidarajulu, Z. (1997). A modification of the test of Shapiro and Wilk for normality. *Journal of Applied Statistics*, *24*(2), pp. 219–236.

Royston, J. P. (1982). An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *31*(2), pp. 115–124.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, pp. 591–611.

Smirnov, N. (1939a). On the derivations of the empirical distribution curve. *Matematicheskii Sbornik*, *6*, pp. 2–26.

Smirnov, N. (1939b). On the estimation of the discrepancy between empirical curves of distribution for two independent samples (Russian). *Bull. Moscow Univ.*, *1*, pp. 3–16.

Stephens, M. (1963). The distribution of the goodness-of-fit statistic U2N. I. *Biometrika*, *50*, pp. 303–313.

Stephens, M. (1964). The distribution of the goodness-of-fit statistic U2N. II. *Biometrika*, *51*, pp. 393–397.

Stephens, M. (1970). Use of the Kolmogorov-Smirnov, Cramér-Von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *32*, pp. 115–122.

Stephens, M. (1974). EDF statistics for goodness-of-fit and some comparisons. *Journal of the American Statistical Association*, *69*, pp. 730–737.

Vuong, Q. H. and Wang, W. (1993). Minimum chi-square estimation and tests for model selection. *Journal of Econometrics*, *56*(1-2), pp. 141–168.

Watson, G. (1961). Goodness-of-fit tests on a circle. *Biometrika*, *48*, pp. 109–114.

Weber, M. D., Leemis, L. M. and Kincaid, R. K. (2006). Minimum Kolmogorov-Smirnov test statistic parameter estimates. *Journal of Statistical Computation and Simulation*, *76*(3), pp. 195–206.

Wijekularathna, D. K., Manage, A. B. W. and Scariano, S. M. (2022). Power analysis of several normality tests: A Monte Carlo simulation study. *Communications in Statistics - Simulation and Computation*, *51*, pp. 757–773.

Zhang, J. and Wu, Y. (2002). Beta Approximation to the Distribution of Kolmogorov-Smirnov Statistic. *Annals of the Institute of Statistical Mathematics*, *54*, pp. 577–584.

ΚΕΦΑΛΑΙΟ 5

ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ ΒΑΣΙΣΜΕΝΟΙ ΣΤΗ ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

Σύνοψη

Ένα σημαντικό πρόβλημα της Στατιστικής Συμπερασματολογίας είναι η μοντελοποίηση ενός διωνυμικού πειράματος, δηλαδή ενός πειράματος με δύο δυνατά αποτελέσματα τα οποία, συνήθως, αναφέρονται ως «επιτυχία» και «αποτυχία». Παρόλο που προβλήματα τέτοιου τύπου συναντώνται σε προβλήματα Παραμετρικής Στατιστικής, η απλότητα αντιμετώπισης αυτών τα κατατάσσει και στο πεδίο της Μη Παραμετρικής Στατιστικής. Σε αυτό το κεφάλαιο θα ασχοληθούμε με προβλήματα ελέγχου υποθέσεων, των οποίων οι ελεγχοσυναρτήσεις βασίζονται στη Διωνυμική κατανομή. Θα παρουσιαστούν έλεγχοι ποσοστιαίων σημείων (ως παραλλαγή του συνήθους διωνυμικού ελέγχου για τον έλεγχο της διαμέσου ενός πληθυσμού), τόσο για συνεχή όσο και για διακριτά δεδομένα. Έπειτα, το ενδιαφέρον θα επικεντρωθεί στην παρουσίαση των προσημικών ελέγχων, οι οποίοι είναι οι πιο γνωστοί αλλά και συχνά χρησιμοποιούμενοι έλεγχοι που βασίζονται στη διωνυμική κατανομή. Ειδικότερα, θα παρουσιαστούν ο έλεγχος του προσημικού κριτηρίου (sign test), ο προσημικός έλεγχος του McNemar και ο έλεγχος των Cox και Stuart.

Προαπαιτούμενη γνώση:

Βασικές γνώσεις Πιθανοτήτων και Στατιστικής.

Προσδοκώμενα μαθησιακά αποτελέσματα:

Ο/η φοιτητής/τρια ολοκληρώνοντας την ενότητα αυτή θα μπορεί να παίρνει απόφαση για προβλήματα τα οποία αφορούν δύο δυνατά αποτελέσματα. Ειδικότερα, ο/η φοιτητής/τρια θα μπορεί να λαμβάνει αποφάσεις για προβλήματα των οποίων τα δεδομένα προέρχονται από τη Διωνυμική κατανομή, είτε σε αυτά εμπλέκεται ένα δείγμα είτε εμπλέκονται δύο, συνήθως, εξαρτημένα δείγματα.

Γλωσσάριο επιστημονικών όρων

- Διωνυμικός Έλεγχος
- Διωνυμικός Έλεγχος για ποσοστιαία σημεία
- Έλεγχος συσχετίσεων
- Έλεγχος McNemar
- Έλεγχος Cox and Stuart
- Προσημικός Έλεγχος

Μπατσίδης, Α., Παπασταμούλης, Π., Πετρόπουλος, Κ., & Ρακιτζής, Α. (2022). *Μη Παραμετρική Στατιστική.* [Προπτυχιακό εγχειρίδιο].

5.1 Εισαγωγή

Πολλές φορές στην εφαρμοσμένη έρευνα μας ενδιαφέρει να γνωρίζουμε το ποσοστό ενός πληθυσμού που έχει ένα χαρακτηριστικό γνώρισμα. Για παράδειγμα, ένας πολιτικός επιθυμεί να γνωρίζει το ποσοστό του εκλογικού σώματος που πρόκειται να τον ψηφίσει στις επόμενες εκλογές. Ένας αναλυτής αγοράς επιθυμεί να γνωρίζει το ποσοστό των οικογενειών μιας συγκεκριμένης περιοχής που πρόκειται να αλλάξουν τουλάχιστον μία ηλεκτρική συσκευή εντός του επόμενου εξαμήνου. Ενδιαφέρον για τους φροντιστές υγείας παρουσιάζει το ποσοστό των ασθενών οι οποίοι λαμβάνουν μια συγκεκριμένη θεραπεία ή το ποσοστό των ασθενών που έχουν μία συγκεκριμένη ασθένεια και άλλα πολλά. Με λίγα λόγια, τα πειραματικά μας δεδομένα μπορούν να χωριστούν σε δύο κατηγορίες, επιτυχία ή αποτυχία, και μας ενδιαφέρει να γνωρίζουμε το ποσοστό της επιτυχίας. Τότε τα δεδομένα που λαμβάνουμε μπορούν να αναλυθούν με χρήση ελέγχων, οι οποίοι βασίζονται στη διωνυμική κατανομή.

Σε αυτό το κεφάλαιο, θα ασχοληθούμε με την κατασκευή ελέγχων οι οποίοι βασίζονται σε διωνυμικά δεδομένα, χρησιμοποιώντας απλές τεχνικές της Μη Παραμετρικής Στατιστικής Συμπερασματολογίας με σκοπό να παίρνουμε απόφαση, υπέρ ή κατά, μίας συγκεκριμένης υπόθεσης, κάθε φορά. Η σπουδαιότητα των ελέγχων αυτών αιτιολογείται από το γεγονός ότι η διωνυμική κατανομή έχει χρησιμοποιηθεί στη στατιστική συμπερασματολογία για πάνω από 250 χρόνια, με τον Jacques Bernoulli (1655-1705), να είναι ο πρωτοπόρος που ασχολήθηκε με αυτήν. Ειδικότερα, αντικείμενο μελέτης στο βιβλίο του Ars Conjectandi[1] (Bernoulli, 1713) αποτέλεσε η ειδική περίπτωση της μίας επανάληψης του τυχαίου πειράματος.

5.2 Διωνυμικός έλεγχος

Για να μπορέσουμε να κατασκευάσουμε έναν διωνυμικό έλεγχο, θεωρούμε ότι το δείγμα αποτελείται από τα αποτελέσματα n ανεξάρτητων δοκιμών Bernoulli, τα οποία ταξινομούνται σε δύο κατηγορίες. Κάθε αποτέλεσμα ανήκει είτε στην κατηγορία 1, είτε στην κατηγορία 2, ποτέ και στις δύο. Έστω n_1 το πλήθος παρατηρήσεις που ανήκουν στην κατηγορία 1, άρα $n_2 = n - n_1$ το πλήθος παρατηρήσεις που ανήκουν στην κατηγορία 2. Οι υποθέσεις που απαιτούνται, για να συνεχίσουμε τη διαδικασία, είναι οι εξής:

(A1) Το αποτέλεσμα κάθε δοκιμής μπορεί να ταξινομηθεί ως επιτυχία (έστω η κατηγορία 1) ή αποτυχία (κατηγορία 2).
(A2) Η πιθανότητα επιτυχίας, p, είναι η ίδια για όλες τις δοκιμές.
(A3) Οι n το πλήθος δοκιμές είναι αμοιβαία ανεξάρτητες.

5.2.1 Διαδικασία ελέγχου υποθέσεων

Στην ενότητα αυτή θα παρουσιαστεί η διαδικασία ελέγχου υποθέσεων. Η μελέτη θα πραγματοποιηθεί διακρίνοντας δύο περιπτώσεις: τα δίπλευρα και τα μονόπλευρα προβλήματα ελέγχου.

5.2.1.1 Δίπλευρο πρόβλημα ελέγχου

Έστω p_0 μια συγκεκριμένη τιμή στο $[0,1]$. Θέλουμε, αρχικά, να ελέγξουμε το δίπλευρο πρόβλημα

$$\text{(A) } H_0 : p = p_0 \ , \ H_1 : p \neq p_0$$

[1] Το βιβλίο αυτό δημοσιεύτηκε οκτώ χρόνια μετά τον θάνατο του Jacques Bernoulli και περιέχεται σε αυτό μία σημαντική μελέτη πάνω σε τέτοιες δοκιμές, η οποία και θεωρείται ορόσημο στην ιστορία της Θεωρίας Πιθανοτήτων (για λεπτομέρειες βλ. Hollander *et al.*, 2014; Mattmüller, 2014). Τα παραπάνω αιτιολογούν και την ονομασία δοκιμή Bernoulli στη μνήμη του.

σε επίπεδο σημαντικότητας a, $0 < a < 1$.

Θεωρούμε $X_1, X_2, \dots, X_n$ ανεξάρτητες τ.μ. (δοκιμές), όπου για $i = 1, \dots, n$

$$X_i = \begin{cases} 1 & , \text{ αν η } i- \text{ οστή δοκιμή οδηγεί σε επιτυχία,} \\ 0 & , \text{ διαφορετικά.} \end{cases}$$

Αφού θέλουμε να ελέγξουμε ότι η πιθανότητα επιτυχίας παίρνει μια συγκεκριμένη τιμή, έστω αυτή p_0, είναι λογικό να θεωρήσουμε ως ελεγχοσυνάρτηση τη στατιστική συνάρτηση (σ.σ.)

$$T = \sum_{i=1}^{n} X_i,$$

όπου, προφανώς, η $T \sim B(n, p)$. Αν στα αποτελέσματα του δείγματος παρατηρήσουμε n_1 επιτυχίες (αποτελέσματα κατηγορίας 1), τότε η T παίρνει την τιμή n_1. Με βάση την τιμή της T μπορούμε να αποφασίσουμε υπέρ ή κατά της H_0.

Συγκεκριμένα, απορρίπτουμε την H_0, αν η σ.σ. T παίρνει είτε «μικρές», είτε «μεγάλες» τιμές, δηλαδή αν

$$T \leqslant c_1 \text{ ή } T > c_2,$$

όπου οι σταθερές c_1 και c_2 υπολογίζονται από το μέγεθος του ελέγχου. Δηλαδή ο έλεγχος, μεγέθους a, για το πρόβλημα (Α), είναι,

$$\phi(\underline{x}) = \begin{cases} 1 & , \quad T \leqslant c_1 \text{ ή } T > c_2 \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

με $P(T \leqslant c_1 \,|\, p = p_0) = a_1 \simeq a/2$ και $P(T \leqslant c_2 \,|\, p = p_0) = 1 - a_2 \simeq 1 - a/2$, όπου ο υπολογισμός των πιθανοτήτων γίνεται υπό την H_0, δηλαδή για $p = p_0$. Το πραγματικό επίπεδο σημαντικότητας του ελέγχου είναι $a = a_1 + a_2$.

5.2.1.2 Μονόπλευρο πρόβλημα ελέγχου

Πολλές φορές μας ενδιαφέρει να ελέγξουμε αν το ποσοστό ενός πληθυσμού είναι μεγαλύτερο ή μικρότερο από μια συγκεκριμένη τιμή p_0, δηλαδή μας απασχολούν προβλήματα είτε της μορφής,

$$\text{(Β) } H_0 : p \leqslant p_0 \ , \ H_1 : p > p_0$$

είτε της μορφής

$$\text{(Γ) } H_0 : p \geqslant p_0 \ , \ H_1 : p < p_0$$

Απορρίπτουμε την H_0 του προβλήματος (Β), αν η σ.σ. T παίρνει «μεγάλες» τιμές, δηλαδή $T > c$, όπου η σταθερά c υπολογίζεται από το μέγεθος του ελέγχου. Δηλαδή ο έλεγχος, μεγέθους a, για το πρόβλημα (Β), είναι,

$$\phi(\underline{x}) = \begin{cases} 1 & , \quad T > c \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

με $P(T \leqslant c \,|\, p = p_0) \simeq 1 - a$.

Ομοίως, απορρίπτουμε την H_0 του προβλήματος (Γ), αν η σ.σ. T παίρνει «μικρές» τιμές, δηλαδή $T \leqslant c$, όπου η σταθερά c υπολογίζεται από το μέγεθος του ελέγχου. Δηλαδή ο έλεγχος, μεγέθους a, για το πρόβλημα (Γ), είναι,

$$\phi(\underline{x}) = \begin{cases} 1 & , \quad T \leqslant c \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

με $P(T \leqslant c \,|\, p = p_0) \simeq a$.

Παρατήρηση 5.1. Όλες οι σταθερές των παραπάνω ελέγχων υπολογίζονται μέσω της συνάρτησης κατανομής της Διωνυμικής κατανομής (βλ. Πίνακες Π.5-Π.10 στο Παράρτημα) και για συγκεκριμένες τιμές του επιπέδου σημαντικότητας a. Αξίζει, επίσης, να υπενθυμίσουμε ότι λόγω της διακριτής φύσης της κατανομής της σ.σ. T δεν μπορούμε, συνήθως, να επιτύχουμε μέγεθος ελέγχου ακριβώς ίσο με a (*συντηρητικός έλεγχος*).

Παράδειγμα 5.1. Σε ένα δείγμα 15 χαπιών μετρήσαμε την περιεκτικότητα ενός φαρμάκου σε ορισμένη δραστική ουσία και πήραμε τα παρακάτω δεδομένα σε mgr.

$$0.52,\ 0.82,\ 1.25,\ 1.9,\ 2.6,\ 3.86,\ 1.4,\ 1.97,\ 2.85,\ 3.9,\ 0.97,\ 1.5,\ 2.0,\ 2.95,\ 0.99.$$

Ελέγξτε, σε επίπεδο σημαντικότητας $a = 0.05$, αν το 30% των χαπιών έχει περιεκτικότητα τουλάχιστον 1.5 mgr.

Λύση Παραδείγματος 5.1. Θεωρούμε την τ.μ. X_i, η οποία μετράει την περιεκτικότητα της δραστικής ουσίας στο i-οστό χάπι, $i = 1, \dots, 15$. Έπειτα, ορίζουμε τις (Bernoulli) τ.μ. Y_i με

$$Y_i = \begin{cases} 1 & , \ X_i \geqslant 1.5, \\ 0 & , \ X_i < 1.5, \end{cases} \quad i = 1, \dots, 15$$

Προφανώς, κάθε $Y_i \sim B(1, p)$, όπου η πιθανότητα επιτυχίας $p = P(Y_i = 1) = P(X_i \geqslant 1.5)$. Με βάση την παραπάνω μοντελοποίηση, μας ενδιαφέρει το πρόβλημα ελέγχου:

$$H_0 : p = 0.3, \quad H_1 : p \neq 0.3.$$

Ο έλεγχος για το παραπάνω πρόβλημα είναι,

$$\phi(\underline{x}) = \begin{cases} 1 & , \ T \leqslant c_1 \text{ ή } T > c_2, \\ 0 & , \ \text{διαφορετικά}, \end{cases}$$

όπου η στατιστική συνάρτηση ελέγχου είναι η $T = \sum_{i=1}^{n} Y_i \sim B(n, p)$ και οι σταθερές c_1 και c_2 υπολογίζονται από

$$P(T \leqslant c_1 \,|\, p = 0.3) \simeq a/2 = 0.025 \text{ και } P(T \leqslant c_2 \,|\, p = 0.3) \simeq 1 - a/2 = 0.975.$$

Από τον Πίνακα Π.8 του Παραρτήματος και επειδή, υπό τη μηδενική υπόθεση H_0, $\mathrm{T} \sim B(15, 0.3)$, παρατηρούμε ότι $P(T \leqslant 1 \,|\, p = 0.3) = 0.0353$ και $P(T \leqslant 8 \,|\, p = 0.3) = 0.9848$. Οι συγκεκριμένες πιθανότητες 0.0353 και 0.9848 είναι, αντίστοιχα, οι πιο «κοντινές» στις τιμές 0.025 και 0.975. Από τα παραπάνω προκύπτει ότι $c_1 = 1$, $c_2 = 8$ και το πραγματικό επίπεδο σημαντικότητας είναι

$$\begin{aligned} a &= P(T \leqslant 1 \,|\, p = 0.3) + P(T > 8 \,|\, p = 0.3) \\ &= P(T \leqslant 1 \,|\, p = 0.3) + 1 - P(T \leqslant 8 \,|\, p = 0.3) \\ &= 0.0353 + 1 - 0.9848 = 0.0505 \end{aligned}$$

Επομένως, ο έλεγχος για το συγκεκριμένο πρόβλημα γίνεται

$$\phi(\underline{x}) = \begin{cases} 1 & , \ T \leqslant 1 \text{ ή } T > 8 \\ 0 & , \ \text{διαφορετικά}. \end{cases}$$

Όμως, από τα δεδομένα προκύπτει ότι η τιμή της T είναι $\tau = 9 > 8$, δηλαδή απορρίπτουμε την H_0 και δεν μπορούμε να υποθέσουμε, σε επίπεδο σημαντικότητας 5.05%, ότι το 30% των χαπιών έχει περιεκτικότητα τουλάχιστον 1.5 mgr. □

5.2.2 Υπολογισμός p-τιμής Διωνυμικού ελέγχου

Όπως έχουμε ήδη αναφέρει στο Κεφάλαιο 1, εκτός από την εύρεση της κρίσιμης περιοχής για τη διεξαγωγή ενός ελέγχου υποθέσεων, μπορούμε να αποφασίσουμε αν απορρίπτεται ή όχι η H_0 με χρήση του παρατηρούμενου επιπέδου σημαντικότητας, δηλαδή με τη χρήση της γνωστής και ως p-τιμή ή p-value. Ειδικότερα, όταν χρησιμοποιείται κάποιο στατιστικό πακέτο είναι σχεδόν βέβαιο ότι η απόφαση για τα παραπάνω προβλήματα ελέγχου λαμβάνεται με χρήση της p-τιμής, παρά τις αδυναμίες που υπάρχουν στη χρήση και στην ερμηνεία της. Στη συνέχεια, υπολογίζουμε την p-τιμή για καθένα από τα προβλήματα (Α), (Β) ή (Γ) ως ακολούθως.

5.2.2.1 Δίπλευρο πρόβλημα ελέγχου

Για το πρόβλημα (Α), ο έλεγχος, όπως κατασκευάστηκε, απορρίπτει την H_0, όταν η τιμή της ελεγχοσυνάρτησης T, έστω αυτή τ, παίρνει είτε «μικρές», είτε «μεγάλες» τιμές. Είναι προφανές ότι, όταν η τ παίρνει τιμές μικρότερες της αναμενόμενης τιμής της T υπό την H_0, η οποία ισούται με np_0, τότε απορρίπτουμε τη μηδενική υπόθεση, αν $\tau \leqslant c_1$. Όμως, η c_1 προσδιορίζεται από τη σχέση

$$P(T \leqslant c_1 \,|\, p = p_0) = a/2$$

ή διαφορετικά $F_{n,p_0}(c_1) = a/2$, όπου $F_{n,p}(\cdot)$ είναι η συνάρτηση κατανομής της $B(n,p)$. Τελικά, $c_1 = F^{-1}_{n,p_0}(a/2)$ και άρα απορρίπτουμε την H_0, αν $\tau \leqslant F^{-1}_{n,p_0}(a/2)$ ή, ισοδύναμα, αν $a \geqslant 2F_{n,p_0}(\tau)$. Δηλαδή σε αυτήν την περίπτωση η p-τιμή του ελέγχου είναι:

$$p\text{-τιμή} = 2F_{n,p_0}(\tau) = 2P(T \leqslant \tau \,|\, n, p_0).$$

Σημειώνουμε ότι ως $F^{-1}_{n,p}(\cdot; n, p)$ συμβολίζουμε την αντίστροφη συνάρτηση κατανομής της $B(n,p)$.

Με ανάλογο τρόπο, όταν η τ παίρνει τιμές μεγαλύτερες της αναμενόμενης τιμής της T, υπό την H_0, η οποία ισούται με np_0, τότε απορρίπτουμε τη μηδενική υπόθεση, αν $\tau > c_2$. Όμως, η c_2 προσδιορίζεται από τη σχέση

$$P(T \leqslant c_2 \,|\, p = p_0) = 1 - a/2$$

ή, διαφορετικά, $F_{n,p_0}(c_2) = 1 - a/2$. Επομένως, $c_2 = F^{-1}_{n,p_0}(1 - a/2)$, άρα απορρίπτουμε την H_0, αν $\tau > F^{-1}_{n,p_0}(1 - a/2)$. Άμεσα, από την προηγούμενη ανισότητα έπεται ότι $a > 2(1 - F_{n,p_0}(\tau))$ ή $a > 2P(T > \tau \,|\, p = p_0)$. Ωστόσο, καθώς η T είναι μια διακριτή τ.μ. και η τιμή $2P(T > \tau \,|\, p = p_0)$ δεν μπορεί να είναι η μικρότερη τιμή του a για την οποία απορρίπτουμε την H_0 (a είναι αυστηρά μεγαλύτερο του $2P(T > \tau \,|\, p = p_0)$), αναγκαστικά η αμέσως μεγαλύτερη τιμή αυτής της (διπλάσιας) πιθανότητας θα είναι η ζητούμενη p-τιμή, δηλαδή,

$$p\text{-τιμή} = 2P(T \geqslant \tau \,|\, p = p_0).$$

Συνοψίζοντας, για το δίπλευρο πρόβλημα ελέγχου η p-τιμή υπολογίζεται ως

$$p\text{-τιμή} = \begin{cases} 2P(T \leqslant \tau | p = p_0) & , \quad \tau \leqslant np_0 \\ 2P(T \geqslant \tau | p = p_0) & , \quad \tau > np_0. \end{cases} \tag{5.1}$$

5.2.2.2 Μονόπλευρο πρόβλημα ελέγχου

Για το πρόβλημα (Β), ο έλεγχος απορρίπτει την H_0, όταν η τιμή της ελεγχοσυνάρτησης T, έστω αυτή τ, παίρνει «μεγάλες» τιμές. Δηλαδή, απορρίπτουμε τη μηδενική υπόθεση, αν $\tau > c$. Όμως, η c προσδιορίζεται από τη σχέση

$$P(T \leqslant c \,|\, p = p_0) = 1 - a,$$

ή, διαφορετικά, $F_{n,p_0}(c) = 1 - a$. Επομένως, $c = F^{-1}_{n,p_0}(1-a)$, άρα απορρίπτουμε την H_0, αν

$$\tau > F^{-1}_{n,p_0}(1-a) \Leftrightarrow a > (1 - F_{n,p_0}(\tau)).$$

Δηλαδή $a > P(T > \tau \,|p = p_0)$. Επειδή η T είναι μια διακριτή τ.μ. και η τιμή $P(T > \tau \,|p = p_0)$ δεν μπορεί να είναι η μικρότερη τιμή του a για την οποία απορρίπτουμε την H_0 (a είναι αυστηρά μεγαλύτερο του $P(T > \tau \,|p = p_0)$), αναγκαστικά η αμέσως μεγαλύτερη τιμή αυτής της πιθανότητας θα είναι η ζητούμενη p-τιμή. Επομένως,

$$p\text{-τιμή} = P(T \geqslant \tau \,|p = p_0). \tag{5.2}$$

Τέλος, για το πρόβλημα (Γ), ο έλεγχος απορρίπτει την H_0, όταν η τιμή της ελεγχοσυνάρτησης T, έστω αυτή τ, παίρνει «μικρές» τιμές. Οπότε, απορρίπτουμε τη μηδενική υπόθεση, αν $\tau \leqslant c$. Όμως, η c προσδιορίζεται από τη σχέση

$$P(T \leqslant c \,|p = p_0) = a$$

ή, διαφορετικά, $F_{n,p_0}(c) = a$. Επομένως, $c = F^{-1}_{n,p_0}(a)$, άρα απορρίπτουμε την H_0, αν $\tau \leqslant F^{-1}_{n,p_0}(a) \Leftrightarrow a \geqslant F_{n,p_0}(\tau)$. Δηλαδή, σε αυτήν την περίπτωση, η p-τιμή του ελέγχου είναι

$$p\text{-τιμή} = P(T \leqslant \tau \,|p = p_0). \tag{5.3}$$

Παράδειγμα 5.2. Υπολογίστε την p-τιμή του ελέγχου που κατασκευάστηκε στο Παράδειγμα 5.2. Ποιο είναι το συμπέρασμά σας;

Λύση Παραδείγματος 5.2. Υπενθυμίζουμε ότι έχουμε ένα πρόβλημα δίπλευρου ελέγχου, οπότε, για να υπολογίσουμε την p-τιμή, πρέπει να συγκρίνουμε την τιμή της σ.σ. T με την αναμενόμενη τιμή. Επειδή

$$\tau = 9 > 4.5 = 15 \cdot 0.3 = np_0,$$

προκύπτει, από τη σχέση (5.1), ότι η p-τιμή $= 2P(T \geqslant 9)$, όπου, υπό την H_0, η σ.σ. $T \sim B(15,0.3)$. Δηλαδή,

$$p\text{-τιμή} = 2(1 - P(T \leqslant 8)) = 2(1 - 0.9848) = 0.0304$$

(δείτε τον Πίνακα Π.8 στο Παράρτημα). Χρησιμοποιώντας τον κανόνα απόφασης, που αναφέρθηκε στο Κεφάλαιο 1, και επειδή η p-τιμή $= 0.0304 < 0.05$, παίρνουμε την απόφαση ότι η H_0 απορρίπτεται. Μάλιστα, το ελάχιστο ε.σ. για το οποίο απορρίπτεται η H_0, είναι ίσο με 3.04%. □

5.2.3 Κανονική προσέγγιση

Σε περιπτώσεις όπου το μέγεθος του δείγματος n είναι μεγαλύτερο του 20, τότε για την κατασκευή των ελέγχων σε καθένα από τα προβλήματα (Α), (Β) ή (Γ), όπως αυτά αναφέρθηκαν στις προηγούμενες ενότητες, χρησιμοποιούμε την κανονική προσέγγιση, που δίνεται στην πρόταση που ακολουθεί.

Πρόταση 5.1. Για μεγάλες τιμές του μεγέθους δείγματος n η στατιστική συνάρτηση $T = \sum_{i=1}^{n} X_i$ ακολουθεί, υπό τη μηδενική υπόθεση $H_0 : p = p_0$, προσεγγιστικά κανονική κατανομή με μέση τιμή $\mathrm{E}(T) = np_0$ και διακύμανση $\mathrm{Var}(T) = np_0\left(1 - p_0\right)$, δηλαδή ισχύει ότι:

$$\frac{T - np_0}{\sqrt{np_0(1-p_0)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Απόδειξη Πρότασης 5.1. Επειδή η τυχαία μεταβλητή T είναι άθροισμα n ανεξάρτητων και ισόνομων τυχαίων μεταβλητών X_i, οι οποίες τυχαίες μεταβλητές υπό τη μηδενική υπόθεση ακολουθούν κατανομή Bernoulli με πιθανότητα επιτυχίας $p = p_0$, ισχύει, λόγω των Αναπαραγωγικών Ιδιοτήτων (βλ. Κεφάλαιο 1), ότι $T \overset{H_0}{\sim} B\left(n, p_0\right)$. Από το Κεντρικό Οριακό Θεώρημα (Κ.Ο.Θ.) γνωρίζουμε ότι για μεγάλο μέγεθος δείγματος n, το άθροισμα ανεξάρτητων και ισόνομων τυχαίων μεταβλητών προσεγγίζεται από μία κανονική κατανομή. Ειδικότερα, καθώς $n \to \infty$, ισχύει ότι:

$$\frac{T - \mathrm{E}(T)}{\sqrt{\mathrm{Var}(T)}} \xrightarrow{d} \mathcal{N}(0,1),$$

οπότε, υπό τη μηδενική υπόθεση, καθώς $\mathrm{E}(T) = np_0$ και $\mathrm{Var}(T) = np_0\left(1 - p_0\right)$, προκύπτει το ζητούμενο προς απόδειξη αποτέλεσμα. □

Παρατήρηση 5.2. Σύμφωνα με τον Rosner (2015), αν η παράμετρος n είναι σχετικά μεγάλη (π.χ. 100) και η πιθανότητα επιτυχίας είναι είτε κοντά στο μηδέν είτε κοντά στο 1, τότε η διωνυμική κατανομή είναι πολύ θετικά ή αρνητικά λοξή. Το ίδιο ισχύει και όταν το n είναι πολύ μικρό για οποιαδήποτε τιμή της πιθανότητας επιτυχίας p. Στις παραπάνω περιπτώσεις, η προσέγγιση της διωνυμικής από την κανονική κατανομή δεν είναι ικανοποιητική. Από την άλλη πλευρά, η προσέγγιση είναι ικανοποιητική, αν το n είναι σχετικά μεγάλο και η πιθανότητα επιτυχίας p είναι ούτε πολύ μικρή ούτε πολύ μεγάλη. Στην πράξη, πολύ συχνά, τα παραπάνω συνοψίζονται στον λεγόμενο *κανόνα του πέντε* (rule of five), σύμφωνα με τον οποίο (βλ. Rosner, 2015) η κανονική προσέγγιση της διωνυμικής είναι ικανοποιητική, όταν $np \geqslant 5$ και $np(1-p) \geqslant 5$.

Παρατήρηση 5.3. Στη βιβλιογραφία (βλ., μεταξύ άλλων, Κούτρας, 2018) προτείνεται στην ειδική περίπτωση που το Κ.Ο.Θ. χρησιμοποιείται για την προσέγγιση της κατανομής του αθροίσματος διακριτών τυχαίων μεταβλητών να προβαίνουμε στην απαραίτητη διόρθωση συνέχειας (continuity correction). Υπάρχουν δύο κύριοι λόγοι που επιβάλλουν τη διόρθωση συνέχειας. Ο πρώτος είναι ότι μια διακριτή τυχαία μεταβλητή μπορεί να λαμβάνει μόνο συγκεκριμένες τιμές, ενώ από την άλλη πλευρά, μια συνεχής τυχαία μεταβλητή λαμβάνει οποιαδήποτε τιμή σε ένα διάστημα. Επιπλέον, εάν χρησιμοποιήσουμε την κανονική κατανομή ως προσέγγιση μιας διακριτής κατανομής, τότε προκύπτει το πρόβλημα που περιγράφεται μέσω του ακόλουθου παραδείγματος.

Έστω η τυχαία μεταβλητή X που ακολουθεί διωνυμική κατανομή με παραμέτρους $n = 100$ και $p = 0.5$. Η τυχαία αυτή μεταβλητή προσεγγίζεται από την κανονική $\mathcal{N}(50,25)$, αφού $\mu = 100 \cdot 0.5 = 50$, $\sigma^2 = 100 \cdot 0.5 \cdot 0.5 = 25$. Αν, για παράδειγμα, μας ζητούσαν να βρούμε την πιθανότητα $P(X = 3)$, τότε, χρησιμοποιώντας την προσέγγιση μέσω της Κανονικής κατανομής θα είναι μηδέν, ενώ η ακριβής τιμή της είναι ίση με $\binom{100}{3} 0.5^{100}$.

Η διόρθωση συνέχειας έγκειται στην πρόσθεση ή αφαίρεση, ανάλογα, στην τιμή ή στις τιμές της διακριτής μεταβλητής της τιμής 0.5. Έτσι, αν X είναι η διακριτή τυχαία μεταβλητή με μέση τιμή $\mathrm{E}(X) = \mu$ και διασπορά $\mathrm{Var}(X) = \sigma^2$, τότε, με εφαρμογή του Κ.Ο.Θ., ισχύουν οι ακόλουθες σχέσεις:

α) $P(X = a) = P(a - 0.5 \leqslant X \leqslant a + 0.5) \approx \Phi\left(\frac{a+0.5-\mu}{\sigma}\right) - \Phi\left(\frac{a-0.5-\mu}{\sigma}\right)$,

β) $P(X \geq a) = P(X \geqslant a - 0.5) \approx 1 - \Phi\left(\frac{a-0.5-\mu}{\sigma}\right)$,

γ) $P(X \leq a) = P(X \leqslant a + 0.5) \approx \Phi\left(\frac{a+0.5-\mu}{\sigma}\right)$,

δ) $P(X > a) = 1 - P(X \leqslant a) \approx 1 - \Phi\left(\frac{a+0.5-\mu}{\sigma}\right)$,

ε) $P(X < a) = 1 - P(X \geqslant a) \approx \Phi\left(\frac{a-0.5-\mu}{\sigma}\right)$,

στ) $P(a \leq X \leq b) = P(a - 0.5 \leqslant X \leqslant b + 0.5) \approx \Phi\left(\frac{b+0.5-\mu}{\sigma}\right) - \Phi\left(\frac{a-0.5-\mu}{\sigma}\right)$,

όπου $\Phi(\cdot)$ είναι η αθροιστική συνάρτηση κατανομής της $\mathcal{N}(0,1)$.

Από την Πρόταση 5.1 και χρησιμοποιώντας τη διόρθωση συνέχειας συνεπάγεται ότι χρησιμοποιούμε για τον υπό μελέτη έλεγχο τη στατιστική συνάρτηση:

$$Z = \begin{cases} \frac{T-0.5-np_0}{\sqrt{np_0(1-p_0)}} & , \quad \text{αν } T > np_0, \\ \\ \frac{T+0.5-np_0}{\sqrt{np_0(1-p_0)}} & , \quad \text{αν } T \leqslant np_0. \end{cases}$$

Οπότε, για το δίπλευρο πρόβλημα (Α) $H_0 : p = p_0$, $H_1 : p \neq p_0$ και σε επίπεδο σημαντικότητας a, $0 < a < 1$, χρησιμοποιούμε τον έλεγχο

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & , \quad |Z| > z_{a/2} \\ 0 & , \quad \text{διαφορετικά} \end{cases}$$

και η p-τιμή αυτού του ελέγχου δίνεται από τη σχέση:

$$p\text{-τιμή} = \begin{cases} 2\Phi\left(\frac{\tau+0.5-np_0}{\sqrt{np_0(1-p_0)}}\right) & , \quad \tau \leqslant np_0, \\ 2\left(1-\Phi\left(\frac{\tau-0.5-np_0}{\sqrt{np_0(1-p_0)}}\right)\right) & , \quad \tau > np_0, \end{cases} \tag{5.4}$$

όπου $\Phi(\cdot)$ είναι η αθροιστική συνάρτηση κατανομής της $\mathcal{N}(0,1)$.

Αντίστοιχα, για το μονόπλευρο πρόβλημα (Β) $H_0 : p \leqslant p_0$, $H_1 : p > p_0$ και σε επίπεδο σημαντικότητας a, $0 < a < 1$, χρησιμοποιούμε τον έλεγχο

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & , \quad Z > z_a \\ 0 & , \quad \text{διαφορετικά.} \end{cases}$$

Η p-τιμή του παραπάνω ελέγχου είναι p-τιμή $= \Phi\left(\frac{\tau-0.5-np_0}{\sqrt{np_0(1-p_0)}}\right)$. Με ανάλογο τρόπο, για το μονόπλευρο πρόβλημα (Γ) $H_0 : p \geqslant p_0$, $H_1 : p < p_0$ και σε επίπεδο σημαντικότητας a, $0 < a < 1$, χρησιμοποιούμε τον έλεγχο

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & , \quad Z < -z_a \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

με την p-τιμή αυτού του ελέγχου να είναι p-τιμή $= \Phi\left(\frac{\tau+0.5-np_0}{\sqrt{np_0(1-p_0)}}\right)$.

Παράδειγμα 5.3. (Conover, 1998) Σύμφωνα με τον απλό νόμο της κληρονομικότητας του Mendel, η διασταύρωση μεταξύ φυτών δύο συγκεκριμένων γενοτύπων, ενδέχεται να οδηγήσει σε απογόνους, οι οποίοι σε ποσοστό 25% είναι νάνοι και σε ποσοστό 75% είναι κανονικοί. Σε ένα πείραμα, για να προσδιοριστεί κατά πόσο η υπόθεση του απλού νόμου του Mendel είναι εύλογη σε μία συγκεκριμένη περίπτωση, μία διασταύρωση οδήγησε σε απογόνους φυτά από τα οποία τα 243 ήταν νάνοι και τα 682 ήταν κανονικά. Πώς θα ερμηνεύατε τα αποτελέσματα του πειράματος αυτού σε επίπεδο σημαντικότητας ίσο με $a = 0.05$; Υπολογίστε την p-τιμή του παραπάνω ελέγχου.

Λύση Παραδείγματος 5.3. Θεωρούμε τις τ.μ. $X_1, X_2, \ldots, X_n$, $n = 925$, όπου

$$X_i = \begin{cases} 1 & , \quad \text{το } i\text{-οστό φυτό είναι νάνος} \\ 0 & , \quad \text{το } i\text{-οστό φυτό είναι κανονικό} \end{cases} \qquad i = 1, \ldots, 925.$$

Λόγω του μεγέθους δείγματος θα χρησιμοποιήσουμε την κανονική προσέγγιση για να πάρουμε απόφαση για το δίπλευρο πρόβλημα

$$H_0 : p = 0.25 \ , \ H_1 : p \neq 0.25.$$

Επειδή η τιμή της σ.σ. T, $\tau = 243 > 231.25 = 924 \cdot 0.25 = np_0$, θα χρησιμοποιήσουμε τη διόρθωση συνέχειας και την τιμή της σ.σ.

$$Z = \frac{T - 0.5 - np_0}{\sqrt{np_0(1-p_0)}}.$$

Στη συγκεκριμένη περίπτωση $z = \frac{243-0.5-925\cdot 0.25}{\sqrt{925\cdot 0.25\cdot 0.75}} = 0.8542$ και, αφού $|z| = 0.8542 < 1.96 = z_{a/2}$, για $a = 0.05$, έπεται ότι δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση H_0. Δηλαδή δεν υπάρχουν σαφείς ενδείξεις, για να απορρίψουμε τον απλό νόμο της κληρονομικότητας του Mendel.
Η p-τιμή υπολογίζεται από τη σχέση (5.4) ως εξής:

$$p\text{-τιμή} = 2\left(1 - \Phi\left(\frac{243 - 0.5 - 925 \cdot 0.25}{\sqrt{925 \cdot 0.25 \cdot 0.75}}\right)\right) = 2 \cdot (1 - \Phi(0.8542)) \approx 0.392$$

χρησιμοποιώντας τον Πίνακα Π.1 του Παραρτήματος. □

5.3 Διωνυμικός έλεγχος για ποσοστιαία σημεία

Ο διωνυμικός έλεγχος μπορεί να χρησιμοποιηθεί για τον έλεγχο οποιουδήποτε ποσοστιαίου σημείου της κατανομής του πληθυσμού. Συνηθίζεται να αναφέρεται ως quantile test (έλεγχος ποσοστιαίου σημείου) όταν το ποσοστιαίο σημείο είναι άλλο εκτός της διαμέσου. Για την εφαρμογή του κριτηρίου απαιτείται τα δεδομένα να είναι τουλάχιστον διατάξιμα (ordinal). Για τον ορισμό του p-ποσοστιαίου σημείου μιας κατανομής παραπέμπουμε στο Κεφάλαιο 1 (βλ. Ορισμό 1.13). Έστω $X_1, \dots, X_n$ ένα τυχαίο δείγμα από κάποια κατανομή με α.σ.κ. F_X και x^* μία γνωστή τιμή στο πεδίο τιμών της F_X. Τα προβλήματα ελέγχων, που αφορούν ποσοστιαία σημεία και έχουν ένα ιδιαίτερο ενδιαφέρον, είναι της μορφής,

$$\text{(Α)}\ H_0 : x_{p^*} = x^*\ ,\ \ H_1 : x_{p^*} \neq x^*$$

$$\text{(Β)}\ H_0 : x_{p^*} \leqslant x^*\ ,\ \ H_1 : x_{p^*} > x^*.$$

$$\text{(Γ)}\ H_0 : x_{p^*} \geqslant x^*\ ,\ \ H_1 : x_{p^*} < x^*$$

όπου x_{p^*} είναι το p^*-ποσοστιαίο σημείο της κατανομής F. Η κατασκευή ελέγχων για τα παραπάνω προβλήματα γίνεται για δύο διαφορετικές περιπτώσεις. Η μία περίπτωση είναι όταν η F_X είναι συνεχής κατανομή και η άλλη όταν η F_X είναι διακριτή.

5.3.1 Συνεχής περίπτωση

Όταν η τ.μ. $X \sim F_X$ είναι συνεχής, γνωρίζουμε ότι $P(X = x_p) = 0$. Επομένως, ο ορισμός του ποσοστιαίου σημείου απλουστεύεται και προκύπτει ότι το σημείο x_p δίνεται από την εξίσωση

$$P(X > x_p) = p.$$

Ορίζουμε τις τ.μ. $Y_1, \dots, Y_n$, με

$$Y_i = \begin{cases} 1 & , \quad \text{αν } X_i > x^*, \\ 0 & , \quad \text{διαφορετικά.} \end{cases}$$

Προφανώς, κάθε $Y_i \sim B(1, p)$, με $p = P(X_i > x^*)$, $i = 1, \dots, n$. Για το πρόβλημα

$$\text{(Α)}\ H_0 : x_{p^*} = x^*\ ,\ \ H_1 : x_{p^*} \neq x^*$$

παρατηρούμε ότι, υπό την H_0, $p = P(X_i > x^*) = P(X_i > x_{p^*}) = p^*$. Δηλαδή προκύπτει το δίπλευρο πρόβλημα ελέγχου

$$\text{(Α')}\ H_0 : p = p^*\ ,\ \ H_1 : p \neq p^*$$

το οποίο είναι ένα δίπλευρο διωνυμικό πρόβλημα ελέγχου. Οπότε, σύμφωνα με όσα αναφέρθηκαν στην Ενότητα 5.2, ο έλεγχος για το πρόβλημα (Α') και άρα για το πρόβλημα (Α) είναι,

$$\phi(\underline{x}) = \begin{cases} 1 & , \quad T \leqslant c_1 \text{ ή } T > c_2, \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

με στατιστική συνάρτηση ελέγχου την $T = \sum_{i=1}^{n} Y_i$, η οποία εκφράζει το πλήθος των X_i τα οποία είναι μεγαλύτερα της τιμής x^*, δηλαδή $T = \sum_{i=1}^{n} Y_i = \#\{X_i > x^*\}$. Οι σταθερές c_1, c_2 υπολογίζονται έτσι, ώστε $P(T \leqslant c_1 | T \sim B(n,p^*)) \simeq a/2$ και $P(T \leqslant c_2 | T \sim B(n,p^*)) \simeq 1 - a/2$.

Η p-τιμή του ελέγχου θα είναι

$$p\text{-τιμή} = \begin{cases} 2P(T \leqslant \tau | p = p^*) & , \quad \tau \leqslant np^*, \\ 2P(T \geqslant \tau | p = p^*) & , \quad \tau > np^*. \end{cases}$$

Με ανάλογο τρόπο, όπως και προηγουμένως, για το πρόβλημα

$$\text{(Β)} \quad H_0 : x_{p^*} \leqslant x^* \ , \quad H_1 : x_{p^*} > x^*.$$

παρατηρούμε ότι, υπό την H_0, $p = P(X_i > x^*) \leqslant P(X_i > x_{p^*}) = p^*$, οπότε προκύπτει το μονόπλευρο πρόβλημα διωνυμικού ελέγχου

$$\text{(Β')} \quad H_0 : p \leqslant p^* \ , \quad H_1 : p > p^*$$

και ο έλεγχος για το πρόβλημα (Β') και άρα για το πρόβλημα (Β) είναι:

$$\phi(\underline{x}) = \begin{cases} 1 & , \quad T > c, \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

με $T = \sum_{i=1}^{n} Y_i = \#\{X_i > x^*\}$ και η σταθερά c υπολογίζεται από την εξίσωση, $P\left(T \leqslant c \middle| p = p^*\right) = 1 - a$. Επίσης, η p-τιμή του ελέγχου θα είναι p-τιμή $= P(T \geqslant \tau | p = p^*)$.

Τέλος, για το πρόβλημα

$$\text{(Γ)} \quad H_0 : x_{p^*} \geqslant x^* \ , \quad H_1 : x_{p^*} < x^*$$

παρατηρούμε ότι, υπό την H_0, $p = P(X_i > x^*) \geqslant P(X_i > x_{p^*}) = p^*$, οπότε προκύπτει το μονόπλευρο πρόβλημα διωνυμικού ελέγχου

$$\text{(Γ')} \quad H_0 : p \geqslant p^* \ , \quad H_1 : p < p^*$$

και ο έλεγχος για το πρόβλημα (Γ') και άρα για το πρόβλημα (Γ) είναι:

$$\phi(\underline{x}) = \begin{cases} 1 & , \quad T \leqslant c, \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

με $T = \sum_{i=1}^{n} Y_i = \#\{X_i > x^*\}$ και η σταθερά c υπολογίζεται από την εξίσωση, $P\left(T \leqslant c \middle| p = p^*\right) = a$. Η p-τιμή του ελέγχου θα είναι p-τιμή $= P(T \leqslant \tau | p = p^*)$.

Παράδειγμα 5.4. Μια μελέτη, που έγινε πριν από 10 χρόνια, αναφέρει ότι το 40% του ενεργού πληθυσμού εργαζόταν τουλάχιστον 9 ώρες ημερησίως. Σε ένα τωρινό τυχαίο δείγμα ενηλίκων αναφέρονται οι μέσοι χρόνοι εργασίας την ημέρα ως ακολούθως:

$$8.2 \ , \ 9.3 \ , \ 6.6 \ , \ 7.4 \ , \ 8.8 \ , \ 5.2 \ , \ 9.1 \ , \ 6.8, \ 9.5, \ 4.4, \ 9, \ 6.2.$$

Χρησιμοποιήστε τον έλεγχο ποσοστιαίων σημείων για να ελέγξετε, σε ε.σ. μικρότερο του 2%, αν σήμερα το 40% του ενεργού πληθυσμού εργάζεται όπως και 10 χρόνια πριν ή λιγότερο.

Λύση Παραδείγματος 5.4. Ορίζουμε την τ.μ. X, η οποία μετράει τις ημερήσιες ώρες εργασίας ενός ενήλικα την τρέχουσα χρονική περίοδο. Αν $x_{0.40}$ είναι το 40% ποσοστιαίο σημείο της κατανομής της τ.μ. X, τότε, για να προσδιορίσουμε αν οι ενήλικες σήμερα εργάζονται όπως και 10 χρόνια πριν, χρησιμοποιούμε το πρόβλημα

$$H_0 : x_{0.40} \geqslant 9 \ , \ \ H_1 : x_{0.40} < 9.$$

Έχουμε ένα δείγμα το οποίο αποτελείται από $n = 12$ το πλήθος παρατηρήσεις, οπότε ορίζουμε τις τ.μ. $Y_1, \dots, Y_{12}$ ως,

$$Y_i = \begin{cases} 1 & , \quad \text{αν } X_i > 9, \\ 0 & , \quad \text{διαφορετικά.} \end{cases}$$

Προφανώς, κάθε $Y_i \sim B(1,p)$, όπου $p = P(X_i > 9)$, $i = 1, \dots, 12$. Το αντίστοιχο πρόβλημα του διωνυμικού ελέγχου είναι

$$H_0 : p \geqslant 0.40 \ , \ \ H_1 : p < 0.40$$

και ο έλεγχος για αυτό το πρόβλημα είναι,

$$\phi(\underline{x}) = \begin{cases} 1 & , \quad T \leqslant c, \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

με $T = \sum_{i=1}^{n} Y_i = \#\{X_i > 9\}$. Δεν είναι δύσκολο να διαπιστώσουμε από τα διαθέσιμα δεδομένα ότι $\tau = 3$. Η σταθερά c υπολογίζεται από την εξίσωση, $P(T \leqslant c \,|p = 0.4) = a$. Υπό την H_0, $\mathrm{T} \sim B(12, 0.4)$. Από τον Πίνακα Π.8 του Παραρτήματος παρατηρούμε ότι $P(T \leqslant 1) = 0.0196$, η οποία είναι η πλησιέστερη τιμή στο $a = 0.02$, δηλαδή $c = 1$ και $a = 0.0196$. Επειδή $\tau = 3 > 1 = c$, συνάγουμε το συμπέρασμα ότι δεν μπορούμε να απορρίψουμε την H_0, δηλαδή δεν υπάρχουν σαφείς ενδείξεις για να απορριφθεί η υπόθεση ότι το 40% του ενεργού πληθυσμού εργάζεται όπως και 10 χρόνια πριν. □

5.3.2 Διακριτή περίπτωση

Στην περίπτωση όπου η τ.μ. $X \sim F_X$ είναι διακριτή, τότε το ποσοστιαίο σημείο x_p ορίζεται από μία διπλή ανισότητα, επομένως χρειαζόμαστε δύο διωνυμικά πειράματα για να μπορέσουμε να μετασχηματίσουμε τα προβλήματα (Α), (Β) ή (Γ), όπως αυτά έχουν τεθεί στην παρούσα ενότητα. Αρχικά ορίζουμε τις τ.μ. $Y_1, \dots, Y_n$, όπου

$$Y_i = \begin{cases} 1 & , \quad \text{αν } X_i \geqslant x^* \\ 0 & , \quad \text{διαφορετικά.} \end{cases}$$

Προφανώς, κάθε $Y_i \sim B(1,p_1)$, με πιθανότητα επιτυχίας $p_1 = P(X_i \geqslant x^*)$, $i = 1, \dots, n$. Επίσης, θεωρούμε τις τ.μ. $Z_1, \dots, Z_n$, με

$$Z_i = \begin{cases} 1 & , \quad \text{αν } X_i > x^* \\ 0 & , \quad \text{διαφορετικά.} \end{cases}$$

Επομένως, $Z_i \sim B(1,p_2)$, με πιθανότητα επιτυχίας $p_2 = P(X_i > x^*)$, $i = 1, \dots, n$. Για το πρόβλημα

$$\text{(Α)} \ \ H_0 : x_{p^*} = x^* \ , \ \ H_1 : x_{p^*} \neq x^*$$

παρατηρούμε ότι, υπό την H_0, είναι

$$H_0 : p_1 = P(X_i \geqslant x^*) = P(X_i \geqslant x_{p^*}) \geqslant p^* \text{ και } p_2 \leqslant p^*$$

ενώ, υπό την H_1, είναι

$$H_1 : p_1 < p^* \ \text{ ή } \ p_2 > p^*$$

Δηλαδή προκύπτει το δίπλευρο πρόβλημα ελέγχου

$$\text{(Α')} \ \ H_0 : p_1 \geqslant p^* \ \text{ και } \ p_2 \leqslant p^* \ , \ \ H_1 : p_1 < p^* \ \text{ ή } \ p_2 > p^*$$

Θεωρούμε τις στατιστικές συναρτήσεις $T_1 = \sum_{i=1}^{n} Y_i$ και $T_2 = \sum_{i=1}^{n} Z_i$, οι οποίες εκφράζουν το πλήθος των X_i, που είναι τουλάχιστον ίσα με x^* και το πλήθος των X_i, που είναι μεγαλύτερα από την τιμή x^*, αντίστοιχα. Θα γράφουμε $T_1 = \sum_{i=1}^{n} Y_i = \#\{X_i \geqslant x^*\}$, $T_2 = \sum_{i=1}^{n} Z_i = \#\{X_i > x^*\}$ και έστω, επίσης, τ_1, τ_2 οι αντίστοιχες τιμές τους από τα δεδομένα στο δείγμα των $X_1, X_2, \ldots, X_n$. Αξίζει να σημειώσουμε ότι μια μεγάλη τιμή για την T_2 αποτελεί ένδειξη ότι η πιθανότητα $P(X_i > x_{p^*})$ είναι «μεγάλη» και άρα έχουμε ένδειξη ενάντια σε αυτό που πρέπει να ισχύει υπό την H_0 για την p_2. Όμοια, μια μικρή τιμή για την T_1 αποτελεί ένδειξη ότι η πιθανότητα $P(X_i \geqslant x_{p^*})$ είναι «μικρή» και άρα έχουμε ένδειξη ενάντια σε αυτό που πρέπει να ισχύει υπό την H_0 για την p_1. Επομένως, απορρίπτουμε την H_0, αν είτε η σ.σ. T_1 παίρνει «μικρές» τιμές, είτε η σ.σ. T_2 παίρνει «μεγάλες» τιμές. Δηλαδή,

$$\phi(\underline{x}) = \begin{cases} 1 & , \quad T_1 \leqslant c_1 \ \text{ή} \ T_2 > c_2, \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

όπου οι σταθερές c_1, c_2 υπολογίζονται από τις σχέσεις $P(T_1 \leqslant c_1 \,|\, p = p^*) \simeq a/2$ και $P(T_2 \leqslant c_2 \,|\, p = p^*) \simeq 1 - a/2$.

Σε αυτήν την περίπτωση, η p-τιμή υπολογίζεται (βλ. π.χ. Conover, 1998) ότι προσδιορίζεται από τη σχέση:

$$p\text{-τιμή} = 2\min\{P(T_1 \leqslant \tau_1 \,|\, p = p^*), P(T_2 \geqslant \tau_2 \,|\, p = p^*)\}$$

Για τα μονόπλευρα προβλήματα ελέγχου εργαζόμαστε με τον τρόπο που θα εξηγηθεί στη συνέχεια. Αρχικά, για το πρόβλημα

$$\text{(Β)} \quad H_0 : x_{p^*} \leqslant x^* \ , \quad H_1 : x_{p^*} > x^*.$$

παρατηρούμε ότι, υπό την H_0, θα ισχύει $p_2 = P(X_i > x^*) \leqslant P(X_i > x_{p^*}) \leqslant p^*$, οπότε προκύπτει το μονόπλευρο πρόβλημα

$$\text{(Β' ')} \quad H_0 : p_2 \leqslant p^* \ , \quad H_1 : p_2 > p^*.$$

Άρα, ο έλεγχος για το πρόβλημα (Β' ') και κατ' επέκταση για το πρόβλημα (Β) είναι,

$$\phi(\underline{x}) = \begin{cases} 1 & , \quad T_2 > c \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

με στατιστική συνάρτηση ελέγχου $T_2 = \sum_{i=1}^{n} Z_i = \#\{X_i > x^*\}$, ενώ η σταθερά c υπολογίζεται από την εξίσωση, $P(T_2 \leqslant c \,|\, p = p^*) = 1 - a$. Επίσης, η p-τιμή του ελέγχου θα είναι

$$p\text{-τιμή} = P(T_2 \geqslant \tau_2 | p = p^*).$$

Με ανάλογο τρόπο, για το πρόβλημα

$$\text{(Γ)} \quad H_0 : x_{p^*} \geqslant x^* \ , \quad H_1 : x_{p^*} < x^*$$

παρατηρούμε ότι, υπό την H_0, θα ισχύει $p_1 = P(X_i \geqslant x^*) \geqslant P(X_i \geqslant x_{p^*}) \geqslant p^*$, οπότε προκύπτει το μονόπλευρο πρόβλημα διωνυμικού ελέγχου

$$\text{(Γ' ')} \quad H_0 : p_1 \geqslant p^* \ , \quad H_1 : p_1 < p^*.$$

Άρα, ο έλεγχος για το πρόβλημα (Γ' ') και κατ' επέκταση για το πρόβλημα (Γ) είναι,

$$\phi(\underline{x}) = \begin{cases} 1 & , \quad T_1 \leqslant c \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

με στατιστική συνάρτηση ελέγχου $T_1 = \sum_{i=1}^{n} Y_i = \#\{X_i \geqslant x^*\}$, ενώ η σταθερά c υπολογίζεται από την εξίσωση $P(T_1 \leqslant c \,|\, p = p^*) = a$. Η p-τιμή του ελέγχου θα είναι

$$p\text{-τιμή} = P(T_1 \leqslant \tau_1 | p = p^*).$$

Παρατήρηση 5.4. Στην περίπτωση όπου το μέγεθος του δείγματος $n > 20$, τότε για την κατασκευή των ελέγχων σε καθένα από τα προβλήματα (Α), (Β) ή (Γ), όπως αυτά αναφέρθηκαν σε αυτήν την ενότητα, χρησιμοποιούμε την κανονική προσέγγιση (βλ. την υποενότητα 5.2.3).

Ένα από τα σημαντικότερα ποσοστιαία σημεία της κατανομής είναι η διάμεσος, δηλαδή το 50% ποσοστιαίο σημείο της κατανομής. Η διάμεσος είναι μια σημαντική και χρήσιμη παράμετρος σε πολλές περιπτώσεις, ειδικότερα όταν η υποκείμενη κατανομή είναι λοξή και χρειαζόμαστε ένα εναλλακτικό μέτρο κεντρικής τάσης (Gibbons and Chakraborti, 2020).

Παράδειγμα 5.5. Σε δείγμα 20 μαθητών Γ' Γυμνασίου καταγράψαμε τις επιδόσεις τους σε ένα διαγώνισμα Μαθηματικών.

$$14\ \ 13\ \ 10\ \ 12\ \ 13\ \ 10\ \ 15\ \ 12\ \ 9\ \ 14\ \ 9\ \ 12\ \ 16\ \ 14\ \ 16\ \ 17\ \ 8\ \ 12\ \ 13\ \ 10$$

Να ελεγχθεί, σε επίπεδο σημαντικότητας $a \simeq 0.05$, η υπόθεση ότι η διάμεσος της κατανομής των βαθμών είναι η τιμή 12.

Λύση Παραδείγματος 5.5. Ορίζουμε την τ.μ. X, η οποία μετράει τον βαθμό του/της μαθητή/τριας στο διαγώνισμα των Μαθηματικών. Αν $x_{0.5}$ είναι η διάμεσος της κατανομής της τ.μ. X, τότε το πρόβλημα ελέγχου υποθέσεων που χρησιμοποιούμε, με σκοπό να πάρουμε απόφαση, είναι

$$H_0 : x_{0.5} = 12\ ,\ \ H_1 : x_{0.5} \neq 12.$$

Έχουμε ένα δίπλευρο πρόβλημα ελέγχου ποσοστιαίου σημείου σε διακριτή κατανομή (υποθέτουμε ότι οι βαθμοί που μπορούν να πάρουν οι μαθητές είναι στο σύνολο {0,1,…,20}), οπότε σύμφωνα με τη θεωρία που αναπτύξαμε σε αυτήν την υποενότητα ορίζουμε ως

$$Y_i = \begin{cases} 1 & , \ \text{αν } X_i \geqslant 12 \\ 0 & , \ \text{διαφορετικά,} \end{cases}$$

όπου κάθε $Y_i \sim B(1, p_1)$, $p_1 = P(X_i \geqslant 12)$, $i = 1, \ldots, 20$. Επίσης, ορίζουμε

$$Z_i = \begin{cases} 1 & , \ \text{αν } X_i > 12 \\ 0 & , \ \text{διαφορετικά.} \end{cases}$$

Επομένως, $Z_i \sim B(1, p_2)$, $p_2 = P(X_i > 12)$, $i = 1, \ldots, 20$.

Το δίπλευρο πρόβλημα ελέγχου που προκύπτει είναι

$$H_0 : p_1 \geqslant 0.5 \ \text{ και } \ p_2 \leqslant 0.5\ ,\ \ H_1 : p_1 < 0.5 \ \text{ ή } \ p_2 > 0.5$$

με ελεγχοσυνάρτηση

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & , \ T_1 \leqslant c_1 \ \text{ ή } \ T_2 > c_2 \\ 0 & , \ \text{διαφορετικά,} \end{cases}$$

όπου $T_1 = \sum_{i=1}^{20} Y_i = \#\{X_i \geqslant 12\}$, με τιμή $\tau_1 = 14$ και $T_2 = \sum_{i=1}^{20} Z_i = \#\{X_i > 12\}$, με τιμή $\tau_2 = 10$.

Επίσης, οι σταθερές c_1, c_2 υπολογίζονται από τις σχέσεις $P(T_1 \leqslant c_1 \big| p = 0.5) \simeq a/2$ και $P(T_2 \leqslant c_2 \big| p = 0.5) \simeq 1 - a/2$, όπου, υπό την H_0, οι $T_1, T_2 \sim B(20, 0.5)$. Από τον Πίνακα Π.10 του Παραρτήματος παρατηρούμε ότι $P(T_1 \leqslant 5) = 0.0207$ και $P(T_2 \leqslant 14) = 0.9793$. Οι τιμές αυτές είναι οι πλησιέστερες στις τιμές 0.025 και 0.975, αντίστοιχα. Δηλαδή $c_1 = 4$, $c_2 = 14$ και το μέγεθος του ελέγχου είναι

$$a = P(T_1 \leqslant 5) + 1 - P(T_2 \leqslant 14) = 0.0207 + 1 - 0.9793 = 0.0414.$$

Ο τελικός έλεγχος είναι,

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & , \ T_1 \leqslant 5 \ \text{ ή } \ T_2 > 14 \\ 0 & , \ \text{διαφορετικά,} \end{cases}$$

Επειδή $\tau_1 = 14 > 5$ και $\tau_2 = 10 < 14$, καταλήγουμε στο ότι δεν έχουμε ισχυρές ενδείξεις έναντι της H_0 και, άρα, δεν μπορούμε να απορρίψουμε το ότι ο διάμεσος βαθμός για το συγκεκριμένο διαγώνισμα Μαθηματικών είναι η τιμή 12. □

5.4 Προσημικός έλεγχος

Μια παραλλαγή του διωνυμικού ελέγχου είναι ο προσημικός έλεγχος ή έλεγχος προσήμου, ο οποίος είναι ένας από τους πιο παλιούς μη παραμετρικούς ελέγχους που αναφέρονται στη βιβλιογραφία. Ο Arbuthnot (1710) εξέτασε τα στοιχεία των γεννήσεων στο Λονδίνο για μία περίοδο 82 χρόνων μεταξύ $1629 - 1710$. Για καθεμία από αυτές τις χρονιές συνέκρινε τον αριθμό των γεννήσεων των αγοριών με αυτόν των κοριτσιών. Για κάθε χρονιά θεώρησε το γεγονός «περισσότερα αγόρια γεννιούνται παρά κορίτσια» ως « + » και το αντίθετο γεγονός ως « − » (χωρίς δεσμούς - ties). Τότε ένα πρόβλημα της μορφής

$$H_0 : P(+) = P(-) \;\; H_1 : P(+) \neq P(-),$$

αναδύθηκε στην επιφάνεια. Για την ιστορία, ο Arbuthnot (1710) κατέληξε στο συμπέρασμα ότι ο αριθμός γεννήσεων των αγοριών διαφέρει από τον αντίστοιχο των κοριτσιών (βλ. Conover, 1998).

Ο προσημικός έλεγχος, ως μη παραμετρικός έλεγχος, δεν απαιτεί κάποιες ιδιαίτερες προϋποθέσεις εφαρμογής του, παρά μόνο ότι τα δεδομένα πρέπει να είναι τουλάχιστον διατάξιμα . (βλ. π.χ. Kvam and Vidakovic, 2007).

Αρχικά, θα δούμε τον προσημικό έλεγχο, ως έλεγχο διαμέσου, δηλαδή θεωρούμε $X_1, \dots, X_k$ ένα τυχαίο δείγμα και έστω m να είναι η διάμεσος του πληθυσμού, από τον οποίο προέρχεται το δείγμα. Αν m_0 είναι μία γνωστή τιμή, μας ενδιαφέρει να εξετάσουμε αν η διάμεσος ξεπερνάει ή δεν ξεπερνάει ή είναι ίση με την τιμή αυτή. Η μέθοδος του προσημικού ελέγχου στηρίζεται στο πρόσημο της διαφοράς των δειγματικών τιμών $X_1, \dots, X_k$ από την προς έλεγχο τιμή m_0. Επομένως, αρχικά δημιουργούμε τις διαφορές $X_1 - m_0, \dots, X_k - m_0$, και θέτουμε « + », όταν είναι θετικές, δηλαδή όταν $X_i - m_0 > 0$, ενώ, όταν είναι αρνητικές, δηλαδή όταν $X_i - m_0 < 0$, θέτουμε « − ». Οι τιμές του δείγματος, για τις οποίες $X_i - m_0 = 0$ αποτελούν περιπτώσεις ισοβαθμιών ή δεσμών ή συμπτώσεων, δεν μας απασχολούν και εξαιρούνται από το δείγμα.

Παρατήρηση 5.5. Παρότι για συνεχείς κατανομές η περίπτωση $X_i = m_0$ είναι απίθανη, σε κάθε τέτοια περίπτωση λέμε ότι έχουμε δεσμούς (ties). Σε όσα ακολουθούν υποθέτουμε ότι δεν υπάρχουν δεσμοί. Σε αντίθετη περίπτωση, αν υπάρχουν δεσμοί, οι συγκεκριμένες πειραματικές μονάδες αποκλείονται από την περαιτέρω ανάλυση, δεν λαμβάνονται υπόψη και γίνεται κατάλληλη τροποποίηση (μείωση) του αρχικού μεγέθους του δείγματος. Σε όσα έπονται, n είναι το μέγεθος του τροποποιημένου δείγματος ($n \leqslant k$).

Τα προβλήματα που μας ενδιαφέρουν είναι της μορφής,

(Α) $H_0 : m = m_0 \;,\;\; H_1 : m \neq m_0.$

(Β) $H_0 : m \leqslant m_0 \;,\;\; H_1 : m > m_0.$

(Γ) $H_0 : m \geqslant m_0 \;,\;\; H_1 : m < m_0.$

Για την κατασκευή των ελέγχων για αυτά τα προβλήματα θεωρούμε τις τ.μ. $Z_1, \dots, Z_n$, όπου κάθε

$$Z_i = \begin{cases} 1 & , \text{ αν } X_i > m_0 \text{ (δηλαδή +)} \\ 0 & , \text{ αν } X_i < m_0 \text{ (δηλαδή −)} \end{cases} \quad , i = 1, \dots, n.$$

Τότε οι τ.μ. $Z_1, \dots, Z_n$ αποτελούν τυχαίο δείγμα από κατανομή Bernoulli, $B(1, p)$, με πιθανότητα επιτυχίας $p = P(Z_i = 1) = P(X_i > m_0) = P(+)$. Έστω T είναι το πλήθος των θετικών διαφορών $X_i - m_0$, δηλαδή ο αριθμός των προσήμων « + », οπότε $T = \sum_{i=1}^{n} Z_i$. Υπό τη μηδενική υπόθεση $H_0 : m = m_0$, δηλαδή υπό την υπόθεση ότι η διάμεσος m είναι ίση με m_0, προκύπτει από τον ορισμό της διαμέσου ότι:

$$P(X > m_0) = P(X < m_0) = 0.5.$$

Επομένως, υπό την $H_0 : m = m_0$, η στατιστική συνάρτηση T περιγράφει τον αριθμό των θετικών προσήμων (επιτυχία= θετικό πρόσημο) σε n το πλήθος πρόσημα (άρα, σε n δοκιμές ενός πειράματος τύχης με δύο δυνατά αποτελέσματα) με πιθανότητα εμφάνισης θετικού προσήμου ίση με 0.5. Άρα, η T ακολουθεί, υπό τη μηδενική υπόθεση, διωνυμική κατανομή με παραμέτρους n και $p = 0.5$. Δηλαδή

$$T \overset{H_0}{\sim} B(n, 0.5).$$

Επομένως, η στατιστική συνάρτηση T έχει, υπό τη μηδενική υπόθεση H_0, μια πλήρως προσδιορισμένη κατανομή, ανεξάρτητη της άγνωστης παραμέτρου m, και μπορεί να χρησιμοποιηθεί για τον προς μελέτη έλεγχο. Επιπλέον, εύκολα γίνεται αντιληπτό ότι καθένα από τα προβλήματα (Α), (Β) ή (Γ) ανάγονται στα ακόλουθα προβλήματα διωνυμικών ελέγχων:

(Α) $H_0 : p = 0.5\ ,\ \ H_1 : p \neq 0.5.$

(Β) $H_0 : p \leqslant 0.5\ ,\ \ H_1 : p > 0.5.$

(Γ) $H_0 : p \geqslant 0.5\ ,\ \ H_1 : p < 0.5.$

Χρησιμοποιώντας τα αποτελέσματα της Ενότητας 5.2, για το Πρόβλημα (Α) κατασκευάζεται ο έλεγχος μεγέθους a,

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & , \ T \leqslant c_1 \text{ ή } T > c_2, \\ 0 & , \ \text{διαφορετικά}, \end{cases}$$

όπου οι σταθερές c_1 και c_2 υπολογίζονται από τις εξισώσεις,

$$P(T \leqslant c_1 \,|\, p = 0.5) = a_1 \simeq a/2 \ \text{ και } P(T \leqslant c_2 \,|\, p = 0.5) = 1 - a_2 \simeq 1 - a/2.$$

Η p-τιμή αυτού του ελέγχου είναι (βλ. τη σχέση (5.1)),

$$p\text{-τιμή} = \begin{cases} 2P(T \leqslant \tau \,|\, p = 0.5) & , \ \tau \leqslant n/2 \\ 2P(T \geqslant \tau \,|\, p = 0.5) & , \ \tau > n/2 \end{cases}$$

Αντίστοιχα, για το Πρόβλημα (Β) έχουμε τον έλεγχο μεγέθους a

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & , \ T > c \\ 0 & , \ \text{διαφορετικά}, \end{cases}$$

με τη σταθερά c να δίνεται από τη σχέση, $P(T \leqslant c \,|\, p = 0.5) = 1 - a$, ενώ η p-τιμή υπολογίζεται από τη σχέση p-τιμή $= P(T \geqslant \tau \,|\, p = 0.5)$. Τέλος, για το Πρόβλημα (Γ), ο έλεγχος μεγέθους a είναι

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & , \ T \leqslant c \\ 0 & , \ \text{διαφορετικά}, \end{cases}$$

με τη σταθερά c να υπολογίζεται από τη σχέση, $P(T \leqslant c \,|\, p = 0.5) = a$. Σε αυτήν την περίπτωση, η p-τιμή είναι ίση με $P(T \leqslant \tau \,|\, p = 0.5)$.

Όπως προαναφέρθηκε, υπό την H_0, η σ.σ. $T \sim B(n, 0.5)$, η οποία είναι μία συμμετρική κατανομή. Η ιδιότητα αυτή της συμμετρίας αποδεικνύεται, όπως θα δούμε στη συνέχεια, πολύ χρήσιμη για την απλοποίηση του ελέγχου για το Πρόβλημα (Α). Αρχικά, κάνοντας κάποιες απλές πράξεις παρατηρούμε ότι

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n!}{(n-(n-k))!(n-k)!} = \binom{n}{n-k}. \tag{5.5}$$

Από τη σχέση (5.5) έπεται ότι

$$\begin{aligned}\sum_{y=0}^{n-k}\binom{n}{y+k} &= \binom{n}{k}+\binom{n}{k+1}+\ldots+\binom{n}{n}\\ &= \binom{n}{n-k}+\binom{n}{n-(k+1)}+\ldots+\binom{n}{n-n} = \sum_{y=0}^{n-k}\binom{n}{y}.\end{aligned} \tag{5.6}$$

Η επόμενη πρόταση μας δίνει ένα χρήσιμο αποτέλεσμα για την κατασκευή δίπλευρων ελέγχων υπόθεσης με χρήση του Προσημικού Ελέγχου.

Πρόταση 5.2. Έστω ότι η σ.σ. $T \sim B(n,0.5)$, τότε η σ.σ. $n-T \sim B(n,0.5)$.

Απόδειξη Πρότασης 5.2. Αρκεί να δείξουμε ότι οι σ.σ. T και $n-T$ έχουν την ίδια συνάρτηση κατανομής για το ίδιο σημείο t. Γνωρίζουμε ότι

$$F_T(t) = P(T \leqslant t) = \sum_{x=0}^{t}\binom{n}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{n-x} = 1-\sum_{x=t+1}^{n}\binom{n}{x}\left(\frac{1}{2}\right)^n.$$

Θέτουμε $y = x-(t+1)$, επομένως

$$F_T(t) = 1-\left(\frac{1}{2}\right)^n\sum_{y=0}^{n-(t+1)}\binom{n}{y+t+1}.$$

Χρησιμοποιώντας τη σχέση (5.6) προκύπτει ότι:

$$F_T(t) = 1-\left(\frac{1}{2}\right)^n\sum_{y=0}^{n-(t+1)}\binom{n}{y} = 1-\sum_{y=0}^{n-t-1}\binom{n}{y}\left(\frac{1}{2}\right)^y\left(\frac{1}{2}\right)^{n-y} = 1-\sum_{y=0}^{n-t-1}\binom{n}{y}\left(\frac{1}{2}\right)^n.$$

Άρα,

$$F_T(t) = 1-P(T \leqslant n-t-1) = P(T > n-t-1) = P(T \geqslant n-t) = P(n-T \leqslant t) = F_{n-T}(t)$$

□

Από την απόδειξη της Πρότασης 5.2 είναι προφανές ότι οι ουρές της κατανομής της T, υπό την $H_0: p = 0.5$, είναι ίσες (λόγω συμμετρίας της $B(n,0.5)$) και, επομένως, $P(T \leqslant c) = a/2 = P(T \geqslant n-c)$, από το οποίο συνεπάγεται ότι ο έλεγχος του δίπλευρου Προβλήματος (Α) απλοποιείται ως εξής:

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & , \quad T \leqslant c \text{ ή } T \geqslant n-c \\ 0 & , \quad \text{διαφορετικά,}\end{cases}$$

όπου η σταθερά c υπολογίζεται από την εξίσωση,

$$P(T \leqslant c \,|\, p = 0.5) \simeq a/2.$$

Για μεγάλες τιμές του δειγματικού μεγέθους αποδεικνύεται η παρακάτω πρόταση.

Πρόταση 5.3. Για μεγάλες τιμές του μεγέθους δείγματος n η στατιστική συνάρτηση $T = \sum_{i=1}^{n} Z_i$, που παριστάνει το πλήθος των θετικών διαφορών $X_i - m_0$, ακολουθεί, υπό τη μηδενική υπόθεση $H_0: m = m_0$, προσεγγιστικά κανονική κατανομή με μέση τιμή $\mathrm{E}(T) = np = 0.5n$ και διακύμανση $\mathrm{Var}(T) = np\left(1-p\right) = 0.5\cdot 0.5\cdot n = 0.25n$, δηλαδή

$$\frac{T-0.5n}{\sqrt{0.25n}} \xrightarrow{d} \mathcal{N}(0,1).$$

Απόδειξη Πρότασης 5.3. Η απόδειξη είναι παρόμοια με αυτήν της Πρότασης 5.1 και αφήνεται ως άσκηση για τον/την αναγνώστη/στρια. □

Από την Πρόταση 5.3 και χρησιμοποιώντας τη διόρθωση συνέχειας συνεπάγεται ότι χρησιμοποιούμε για τον υπό μελέτη έλεγχο τη στατιστική συνάρτηση:

$$Z = \begin{cases} \frac{T-0.5-0.5n}{\sqrt{0.25n}}, \text{ αν } T > n/2, \\ \frac{T+0.5-0.5n}{\sqrt{0.25n}}, \text{ αν } T < n/2. \end{cases}$$

Οπότε, για το δίπλευρο πρόβλημα (Α) $H_0 : p = 0.5, \quad H_1 : p \neq 0.5$ και σε επίπεδο σημαντικότητας a, $0 < a < 1$, χρησιμοποιούμε τον έλεγχο:

$$\phi(z) = \begin{cases} 1 & , \ |Z| > z_{a/2}, \\ 0 & , \ \text{διαφορετικά} \end{cases}$$

και η p-τιμή αυτού του ελέγχου δίνεται από τη σχέση:

$$p\text{-τιμή} = \begin{cases} 2\Phi\left(\frac{\tau+0.5-0.5n}{\sqrt{0.25n}}\right) & , \ \tau \leqslant n/2, \\ 2\left(1-\Phi\left(\frac{\tau-0.5-0.5n}{\sqrt{0.25n}}\right)\right) & , \ \tau > n/2, \end{cases}$$

όπου $\Phi(\cdot)$ είναι η αθροιστική συνάρτηση κατανομής της $\mathcal{N}(0,1)$.

Αντίστοιχα, για το πρόβλημα (Β) $H_0 : p \leqslant 0.5, \quad H_1 : p > 0.5$ και σε επίπεδο σημαντικότητας a, $0 < a < 1$, χρησιμοποιούμε την ελεγχοσυνάρτηση:

$$\phi(z) = \begin{cases} 1 & , \ Z > z_a, \\ 0 & , \ \text{διαφορετικά.} \end{cases}$$

Η p-τιμή του παραπάνω ελέγχου είναι p-τιμή $= \Phi\left(\frac{\tau-0.5-0.5n}{\sqrt{0.25n}}\right)$. Τέλος, για το πρόβλημα (Γ) $H_0 : p \geqslant 0.5, \quad H_1 : p < 0.5$ και σε επίπεδο σημαντικότητας a, $0 < a < 1$, χρησιμοποιούμε την ελεγχοσυνάρτηση:

$$\phi(z) = \begin{cases} 1 & , \ Z < -z_a, \\ 0 & , \ \text{διαφορετικά,} \end{cases}$$

με την p-τιμή αυτού του ελέγχου να είναι p-τιμή $= \Phi\left(\frac{\tau+0.5-0.5n}{\sqrt{0.25n}}\right)$.

Παράδειγμα 5.6. (Sprent, 1999) Στον πίνακα που ακολουθεί καταγράφεται ο αριθμός των σελίδων 24 τυχαία επιλεγμένων βιβλίων από μία βιβλιοθήκη ενός Τμήματος Μαθηματικών. Αφού κάνετε τις κατάλληλες υποθέσεις, να ελέγξετε, σε ε.σ. 5%, την υπόθεση ότι ο μέσος αριθμός των σελίδων είναι ίσος με 220 σελίδες χρησιμοποιώντας

α) τον ακριβή τρόπο ελέγχου του προσημικού τεστ,
β) τον προσεγγιστικό τρόπο ελέγχου με την κατάλληλη διόρθωση συνέχειας.

153 166 181 192 244 248 258 264 296 305 305 312 330 340 356 361 395 427 433 467 544 551 625 783.

Λύση Παραδείγματος 5.6. Έχουμε ένα τυχαίο δείγμα $X_1, \ldots, X_{24}$, από έναν πληθυσμό με συνεχή αθροιστική συνάρτηση κατανομής F_X. Υποθέτοντας ότι τα δεδομένα προέρχονται από συμμετρικό πληθυσμό, τα αποτελέσματα του ελέγχου της υπόθεσης ότι η διάμεσος m της άγνωστης κατανομής είναι

ίση με $m_0 = 220$, γενικεύονται για την πληθυσμιακή μέση τιμή. Επομένως, θέλουμε να ελέγξουμε τη μηδενική υπόθεση $H_0 : m = 220$ έναντι της εναλλακτικής $H_1 : m \neq 220$.

Το προσημικό τεστ στηρίζεται στη σ.σ. T, που παριστάνει το πλήθος των θετικών διαφορών $X_i - 220$, $i = 1, ..., 24$. Παρατηρούμε ότι υπάρχουν 4 παρατηρήσεις μικρότερες της τιμής 220, δηλαδή οι: 153, 166, 181, 192 και 20 παρατηρήσεις μεγαλύτερες της τιμής 220, δηλαδή οι υπόλοιπες.

α) Υπό την $H_0 : m = 220$, η στατιστική συνάρτηση T περιγράφει τον αριθμό των θετικών προσήμων (επιτυχία= θετικό πρόσημο) σε $n = 24$ το πλήθος πρόσημα (άρα σε $n = 24$ δοκιμές ενός πειράματος τύχης με δύο δυνατά αποτελέσματα) με πιθανότητα επιτυχίας 0.5. Άρα, γίνεται αντιληπτό ότι η σ.σ. T ακολουθεί, υπό τη μηδενική υπόθεση, διωνυμική κατανομή με παραμέτρους $n = 24$ και $p = 0.5$. Δηλαδή

$$T \overset{H_0}{\sim} B(24, 0.5).$$

Η μηδενική υπόθεση απορρίπτεται για $T \leqslant c$ ή $T \geqslant n - c$, όπου c είναι ένας ακέραιος αριθμός που ικανοποιεί τη σχέση,

$$P(T \leqslant c | H_0 \text{ αληθής}) = P(T \leqslant c | T \sim B(24, 0.5)) \approx a/2,$$

Χρησιμοποιώντας την `R` και την εντολή `pbinom(6,24,0.5)` διαπιστώνουμε ότι $P(T \leqslant 6) = 0.01132792 \leqslant 0.025$, η οποία είναι η πλησιέστερη πιθανότητα στο 0.025. Άρα $c = 6$ και $n - c = 24 - 6 = 18$. Επομένως, καθώς $\tau = 20 > n - c = 18$, συμπεραίνουμε ότι, σε επίπεδο σημαντικότητας 5%, έχουμε σαφή ένδειξη ότι η πραγματική τιμή της πληθυσμιακής διαμέσου είναι μεγαλύτερη από 220. Αξίζει να παρατηρήσουμε πως το ακριβές επίπεδο σημαντικότητας (μέγεθος ελέγχου) είναι ίσο με

$$a = P(T \leqslant 6) + P(T \geqslant 18) = 2P(T \leqslant 6) = 2 \cdot 0.011 = 0.022.$$

Τέλος, ο υπολογισμός της p-τιμής γίνεται ως ακολούθως. Αφού $n/2 = 12 < 20 = \tau$, έπεται ότι η p-τιμή ισούται με $2P(T \geqslant 20 \,|\, p = 0.5) = 2(1 - P(T \leqslant 19 \,|\, p = 0.5)) \approx 0.0015$, χρησιμοποιώντας την `R` και την εντολή $2 \cdot$ (1−`pbinom(19,24,0.5)`).

β) Χρησιμοποιώντας τη διόρθωση συνέχειας συνεπάγεται ότι χρησιμοποιούμε για τον υπό μελέτη έλεγχο τη στατιστική συνάρτηση:

$$Z = \frac{T - 0.5 - 0.5n}{\sqrt{0.25n}} \overset{H_0}{\sim} \mathcal{N}(0,1),$$

καθώς $\tau = 20 > n/2 = 12$ και η μηδενική υπόθεση $H_0 : m = m_0$, απορρίπτεται έναντι της $H_1 : m \neq m_0$, αν $|Z| \geqslant z_{a/2} = z_{0.025} = 1.96$. Καθώς

$$z = \frac{20 - 0.5 - 0.5 \cdot 24}{\sqrt{0.25 \cdot 24}} = \frac{7.5}{\sqrt{6}} = \frac{7.5}{2.45} = 3.06,$$

συμπεραίνουμε ότι η μηδενική υπόθεση απορρίπτεται. Δηλαδή, υποθέτοντας ότι τα δεδομένα προέρχονται από συμμετρικό πληθυσμό, προκύπτει ότι υπάρχει σαφής ένδειξη, σε ε.σ. 5%, ότι ο μέσος αριθμός των σελίδων των βιβλίων της βιβλιοθήκης του Τμήματος Μαθηματικών είναι διαφορετικός της τιμής 220. □

5.4.1 Μια εναλλακτική μορφή του προσημικού ελέγχου

Πολλές φορές, ο προσημικός έλεγχος χρησιμοποιείται σε περιπτώσεις που διαθέτουμε ζεύγος παρατηρήσεων (X, Y) και θέλουμε να διαπιστώσουμε αν το πλήθος των περιπτώσεων όπου οι τιμές της X υπερβαίνουν τις τιμές Y (στο ίδιο ζεύγος) είναι το ίδιο ή μεγαλύτερο ή μικρότερο από το πλήθος των περιπτώσεων όπου οι τιμές της Y υπερβαίνουν τις τιμές X. Παρόλο που για το ίδιο πρόβλημα υπάρχουν και άλλοι έλεγχοι στη βιβλιογραφία (παραπέμπουμε στο Κεφάλαιο 6), ο προσημικός έλεγχος είναι συνήθως

απλούστερος στη χρήση, ενώ δεν απαιτούνται ειδικοί στατιστικοί πίνακες για τον καθορισμό της κρίσιμης περιοχής.

Θεωρούμε $(X_1, Y_1), \dots, (X_k, Y_k)$ τυχαίο δείγμα από έναν διδιάστατο πληθυσμό, ο οποίος περιγράφεται από το τυχαίο διάνυσμα (X, Y). Οι X_i, Y_i είναι, συνήθως, εξαρτημένες τ.μ. μιας και στην περίπτωση που αυτές είναι ανεξάρτητες, προτιμάται ο έλεγχος των Mann-Whitney (παραπέμπουμε στο Κεφάλαιο 6), καθώς ο τελευταίος έχει μεγαλύτερη ισχύ από τον προσημικό έλεγχο (Conover, 1998). Η μέθοδος, σε αυτήν την περίπτωση των ζευγαρωτών παρατηρήσεων, βασίζεται στο πρόσημο της διαφοράς των δειγματικών τιμών $X_i - Y_i$. Όταν οι διαφορές είναι θετικές, δηλαδή όταν $X_i - Y_i > 0$, τότε θέτουμε « + », ενώ, όταν είναι αρνητικές, δηλαδή όταν $X_i - Y_i < 0$ ή, ισοδύναμα, όταν $X_i < Y_i$, θέτουμε « − ». Οι τιμές του δείγματος, για τις οποίες $X_i - Y_i = 0$, βγαίνουν έξω από το δείγμα, το οποίο πλέον είναι μεγέθους $n \leqslant k$. Είναι, προφανές, ότι, για να εφαρμοστούν τα παραπάνω, αρκεί οι τιμές να είναι διατάξιμες (ordinal).

Σε αυτό το πλαίσιο, τα προβλήματα ελέγχου που μας ενδιαφέρουν είναι της μορφής:

$$\text{(A)}\ \ H_0 : \mathsf{E}(X) = \mathsf{E}(Y)\ ,\ \ H_1 : \mathsf{E}(X) \neq \mathsf{E}(Y).$$

$$\text{(B)}\ \ H_0 : \mathsf{E}(X) \leqslant \mathsf{E}(Y)\ ,\ \ H_1 : \mathsf{E}(X) > \mathsf{E}(Y).$$

$$\text{(Γ)}\ \ H_0 : \mathsf{E}(X) \geqslant \mathsf{E}(Y)\ ,\ \ H_1 : \mathsf{E}(X) < \mathsf{E}(Y).$$

Παρατήρηση 5.6. Αξίζει να αναφέρουμε ότι για την περίπτωση ενός τυχαίου δείγματος (X_i, Y_i), όπως αυτό ορίστηκε προηγουμένως, η αρχική μορφή των ελέγχων είναι:

$$\text{(A)}\ \ H_0 : P(X_i < Y_i) = P(X_i > Y_i)\ \forall\, i,\ \ H_1 : P(X_i < Y_i) \neq P(X_i > Y_i),\ \text{για τουλάχιστον ένα } i.$$

$$\text{(B)}\ \ H_0 : P(X_i < Y_i) \leqslant P(X_i > Y_i)\ \forall\, i,\ \ H_1 : P(X_i < Y_i) > P(X_i > Y_i),\ \text{για τουλάχιστον ένα } i.$$

$$\text{(Γ)}\ \ H_0 : P(X_i < Y_i) \geqslant P(X_i > Y_i)\ \forall\, i,\ \ H_1 : P(X_i < Y_i) < P(X_i > Y_i),\ \text{για τουλάχιστον ένα } i$$

Για την κατασκευή των ελέγχων για τα παραπάνω προβλήματα θεωρούμε τις τ.μ. $Z_1, \dots, Z_n$, όπου κάθε

$$Z_i = \begin{cases} 1 & , \ \text{αν } X_i > Y_i \ (\text{ή } +) \\ 0 & , \ \text{αν } X_i < Y_i \ (\text{ή } -) \end{cases} \ , \ \text{για } i = 1, \dots, n.$$

Τότε οι τ.μ. $Z_1, \dots, Z_n$ αποτελούν τυχαίο δείγμα από τη $B(1, p)$, με $p = P(Z_i = 1) = P(X_i > Y_i) = P(+)$. Έστω T είναι το πλήθος των θετικών διαφορών $X_i - Y_i$, δηλαδή ο αριθμός των προσήμων « + », οπότε $T = \sum_{i=1}^{n} Z_i$. Υπό τη μηδενική υπόθεση $H_0 : \mathsf{E}(X) = \mathsf{E}(Y)$, η στατιστική συνάρτηση T περιγράφει τον αριθμό των θετικών προσήμων (επιτυχία= θετικό πρόσημο) σε n το πλήθος πρόσημα (άρα σε n δοκιμές ενός πειράματος τύχης με δύο δυνατά αποτελέσματα), με πιθανότητα εμφάνισης θετικού προσήμου ίση με 0.5. Άρα, η T ακολουθεί, υπό τη μηδενική υπόθεση, διωνυμική κατανομή, δηλαδή $T \overset{H_0}{\sim} B(n, 0.5)$ και μπορεί να χρησιμοποιηθεί ως ελεγχοσυνάρτηση. Επιπλέον, εύκολα γίνεται αντιληπτό ότι καθένα από τα προβλήματα (Α), (Β) ή (Γ) ανάγονται στα ακόλουθα προβλήματα διωνυμικών ελέγχων,

$$\text{(A)}\ \ H_0 : p = 0.5\ ,\ \ H_1 : p \neq 0.5.$$

$$\text{(B)}\ \ H_0 : p \leqslant 0.5\ ,\ \ H_1 : p > 0.5.$$

$$\text{(Γ)}\ \ H_0 : p \geqslant 0.5\ ,\ \ H_1 : p < 0.5.$$

Αυτά τα προβλήματα αντιμετωπίστηκαν σε αυτό το κεφάλαιο, οπότε για το Πρόβλημα (Α) κατασκευάζεται ο έλεγχος

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & , \ T \leqslant c \ \text{ή} \ T \geqslant n - c, \\ 0 & , \ \text{διαφορετικά}, \end{cases}$$

όπου η σταθερά c υπολογίζεται από τη σχέση $P(T \leqslant c \,|\, p = 0.5) = a/2$. Η p-τιμή αυτού του ελέγχου είναι (βλ. τη σχέση (5.1))

$$p\text{-τιμή} = \begin{cases} 2P(T \leqslant \tau \,|\, p = 0.5) & , \quad \tau \leqslant n/2, \\ 2P(T \geqslant \tau \,|\, p = 0.5) & , \quad \tau > n/2, \end{cases}$$

όπου τ είναι η τιμή της T στο δείγμα. Αντίστοιχα, για το Πρόβλημα (Β) έχουμε τον έλεγχο:

$$\phi(\underline{x}) = \begin{cases} 1 & , \quad T > c, \\ 0 & , \quad \text{διαφορετικά}, \end{cases}$$

με τη σταθερά c να δίνεται από τη σχέση $P(T \leqslant c \,|\, p = 0.5) = 1 - a$, ενώ η p-τιμή υπολογίζεται από τη σχέση p-τιμή $= P(T \geqslant \tau \,|\, p = 0.5)$. Τέλος, για το Πρόβλημα (Γ):

$$\phi(\underline{x}) = \begin{cases} 1 & , \quad T \leqslant c, \\ 0 & , \quad \text{διαφορετικά}, \end{cases}$$

με τη σταθερά c να υπολογίζεται από τη σχέση $P(T \leqslant c \,|\, p = 0.5) = a$, ενώ η p-τιμή δίνεται από τη σχέση p-τιμή $= P(T \leqslant \tau \,|\, p = 0.5)$.

Παρατήρηση 5.7. Για μεγάλες τιμές του n και χρησιμοποιώντας τη διόρθωση συνέχειας χρησιμοποιούμε για το υπό μελέτη πρόβλημα ελέγχου τη στατιστική συνάρτηση

$$Z = \begin{cases} \frac{T - 0.5 - 0.5n}{\sqrt{0.25n}}, & \text{αν } T > n/2, \\ \frac{T + 0.5 - 0.5n}{\sqrt{0.25n}}, & \text{αν } T \leqslant n/2. \end{cases}$$

οπότε και κάνουμε $z - test$ για καθένα από τα προβλήματα ενδιαφέροντος με τον ίδιο ακριβώς τρόπο, όπως αναφέρθηκε σε αυτήν την ενότητα για το πρόβλημα διαμέσου.

Παράδειγμα 5.7. Μερικοί ερευνητές ισχυρίζονται ότι η ευπάθεια στην ύπνωση μπορεί να βελτιωθεί μέσω της εκπαίδευσης. Για να ερευνήσουμε αυτόν τον ισχυρισμό, μετρήσαμε για έξι άτομα και σε κλίμακα 1-20 την αρχική ευπάθεια στην ύπνωση και, κατόπιν, εκπαιδεύσαμε τα ίδια άτομα για 4 εβδομάδες. Μετά από την περίοδο εκπαίδευσης, κάθε άτομο μετρήθηκε ξεχωριστά και στην ίδια κλίμακα. Στις παρακάτω μετρήσεις, μεγαλύτεροι αριθμοί αναπαριστούν μεγαλύτερη ευπάθεια στην ύπνωση. Διεξάγοντας έναν έλεγχο προσήμου (sign test) πιστεύετε ότι αυτά τα δεδομένα επιβεβαιώνουν τον παραπάνω ισχυρισμό;

Άτομο	1	2	3	4	5	6
Πριν	10	16	7	4	7	2
Μετά	18	19	11	3	5	3

Ο έλεγχος να γίνει με χρήση κατάλληλης κρίσιμης περιοχής, για ε.σ. 10%, αλλά και με χρήση της p-τιμής του παραπάνω ελέγχου.

Λύση Παραδείγματος 5.7. Θεωρούμε την τ.μ. X, η οποία μετράει την ευπάθεια στην ύπνωση πριν την εκπαίδευση, και την τ.μ. Y, η οποία μετράει την ευπάθεια στην ύπνωση μετά την εκπαίδευση. Για να βελτιωθεί η ευπάθεια στην ύπνωση πρέπει οι τιμές μετά την εκπαίδευση να μικρύνουν.

Θεωρούμε τις τ.μ.

$$Z_i = \begin{cases} 1 & , \quad \text{αν } X_i > Y_i \ (+) \\ 0 & , \quad \text{αν } X_i < Y_i \ (-) \end{cases} \quad , i = 1, \ldots, 6.$$

Τότε οι τ.μ. $Z_1, \ldots, Z_6$ αποτελούν τυχαίο δείγμα από την κατανομή $B(1, p)$, με πιθανότητα επιτυχίας $p = P(Z_i = 1) = P(X_i > Y_i) = P(+)$. Για να δώσουμε απάντηση στο ερώτημα που μας τέθηκε, χρησιμοποιούμε το πρόβλημα ελέγχου:

$$H_0 : p \leqslant 0.5 \ , \ \ H_1 : p > 0.5,$$

καθώς, αν απορρίψουμε την H_0, η ευπάθεια στην ύπνωση βελτιώνεται. Σύμφωνα με τα προηγούμενα, απορρίπτουμε την H_0, αν η τιμή τ της σ.σ. $T = \sum_{i=1}^{n} Z_i$ παίρνει «μεγάλες» τιμές, δηλαδή αν $\tau > c$, όπου το c δίνεται από τη σχέση, $P(T \leqslant c \,|p = 0.5) \approx 1 - a$. Άρα, η μορφή της κρίσιμης περιοχής είναι $K : T > c$. Από το δείγμα έχουμε ότι $\tau = 2$, ενώ με χρήση του Πίνακα Π.7 του Παραρτήματος και, επειδή $T \overset{H_0}{\sim} B(6, 0.5)$, έπεται ότι $P(T \leqslant 4) = 0.8906 \approx 0.90$, που σημαίνει ότι $c = 4$. Επίσης, το ακριβές μέγεθος του ελέγχου είναι $a = 0.1094$. Αφού $\tau = 2 < 4$, συμπεραίνουμε ότι δεν μπορούμε να απορρίψουμε την H_0 σε ε.σ. 10.94%, δηλαδή σε ε.σ. 10.94% αποφασίζουμε ότι δεν βελτιώνεται η ευπάθεια στην ύπνωση με την εκπαίδευση που ακολουθήθηκε.

Η p-τιμή του ελέγχου είναι

$$P(T \geqslant 2|p = 0.5) = 1 - P(T < 2|p = 0.5) = 1 - P(T \leqslant 1|p = 0.5) = 1 - 0.1094 = 0.8906,$$

από όπου συμπεραίνουμε ότι δεν έχουμε ισχυρές ενδείξεις έναντι της H_0 σε ε.σ. 10%. □

5.5 Παραλλαγές του Προσημικού ελέγχου

Σε αυτήν την ενότητα θα ασχοληθούμε με κάποιες ειδικές περιπτώσεις (παραλλαγές) του προσημικού ελέγχου, οι οποίες έχουν ιδιαίτερη σημασία στη βιβλιογραφία και χρησιμοποιούνται από πολλούς ερευνητές σε διάφορες εφαρμογές στην ιατρική, την κοινωνιολογία, την ψυχολογία, την οικολογία και αλλού.

5.5.1 Έλεγχος McNemar

Όπως αναφέραμε και στην προηγούμενη ενότητα, ο προσημικός έλεγχος χρησιμοποιείται όταν τα δεδομένα είναι διατάξιμα. Πώς μπορούμε όμως, να εφαρμόσουμε τον προσημικό έλεγχο για τον έλεγχο της διαφοράς ανάμεσα σε συσχετισμένα ποσοστά για τον έλεγχο της σημαντικότητας στην αλλαγή μιας κατάστασης; Δηλαδή το ερώτημα που τίθεται είναι πώς μπορούμε να εφαρμόσουμε τον προσημικό έλεγχο όταν έχουμε δύο εξαρτημένα δείγματα ονομαστικών (nominal) δεδομένων, τα οποία χωρίζονται σε δύο κατηγορίες, έστω «0» και «1». Σε αυτήν την περίπτωση εφαρμόζεται ο έλεγχος McNemar (βλ. π.χ. Sheskin, 2011), ο οποίος προτάθηκε από τον McNemar (1947).

Θεωρούμε $(X_1, Y_1), \ldots, (X_k, Y_k)$ διδιάστατο τυχαίο δείγμα, όπου X_i αναπαριστά την κατάσταση του ατόμου ΠΡΙΝ τη θεραπεία (ή γενικά πριν την εφαρμογή ενός τυχαίου πειράματος) και Y_i αναπαριστά την κατάσταση του ατόμου ΜΕΤΑ τη θεραπεία (ή γενικά μετά την εφαρμογή του τυχαίου πειράματος). Οι X_i και Y_i είναι ονομαστικές τ.μ., εκφράζουν την κατάσταση στην οποία βρίσκεται το άτομο ΠΡΙΝ και ΜΕΤΑ την εφαρμογή του τυχαίου πειράματος, με δύο δυνατές καταστάσεις. Για τον λόγο αυτόν, μπορούμε να δώσουμε τις τιμές 0 και 1 σε αυτές τις δύο καταστάσεις. Άρα, οι δυνατές τιμές του ζεύγους (X_i, Y_i), $i = 1, 2, \ldots, k$, είναι $(0,0)$, $(0,1)$, $(1,0)$ και $(1,1)$. Τα δεδομένα αυτού του ελέγχου, συνήθως, παρουσιάζονται σε έναν 2×2 πίνακα ταξινόμησης (ή πίνακα διπλής εισόδου ή πίνακα συνάφειας), ως ακολούθως:

		ως προς Y_i	
		0	1
ως προς X_i	0	$\alpha = \#(0,0)$	$\beta = \#(0,1)$
	1	$\gamma = \#(1,0)$	$\delta = \#(1,1)$

Πίνακας 5.1: Ταξινόμηση ως προς τις τ.μ. X_i, Y_i.

Σε όσα ακολουθούν

- με α συμβολίζουμε το πλήθος των ατόμων τα οποία από την κατάσταση 0 πριν τη θεραπεία, εξακολουθούν να βρίσκονται στην κατάσταση 0 και μετά τη θεραπεία,
- με β συμβολίζουμε το πλήθος των ατόμων τα οποία από την κατάσταση 0 πριν τη θεραπεία, πάνε στην κατάσταση 1 μετά τη θεραπεία,
- με γ συμβολίζουμε το πλήθος των ατόμων τα οποία από την κατάσταση 1 πριν τη θεραπεία, πάνε στην κατάσταση 0 μετά τη θεραπεία, και
- με δ συμβολίζουμε το πλήθος των ατόμων τα οποία από την κατάσταση 1 πριν τη θεραπεία, βρίσκονται στην κατάσταση 1 και μετά από αυτήν.

Οι υποθέσεις που απαιτούνται για να εφαρμόσουμε τον έλεγχο McNemar είναι οι εξής (Conover, 1998):

(B1) Τα τυχαία διανύσματα (X_i, Y_i) είναι μεταξύ τους ανεξάρτητα.
(B2) Η κλίμακα μέτρησης είναι ονομαστική με δύο κατηγορίες για όλα τα X_i, Y_i.
(B3) Οι διαφορές $P(X_i = 0, Y_i = 1) - P(X_i = 1, Y_i = 0)$ είναι για κάθε $i = 1, 2, \dots, k$, αρνητικές ή μηδέν ή θετικές.

Αυτό που μας ενδιαφέρει είναι να ελέγξουμε αν τα άτομα έχουν υποστεί αλλαγή στη συμπεριφορά τους μετά τη θεραπεία. Για τον σκοπό αυτόν διεξάγουμε το παρακάτω πρόβλημα ελέγχου υποθέσεων.

$$H_0 : P(X_i = 0, Y_i = 1) = P(X_i = 1, Y_i = 0), \text{ για κάθε } i$$

έναντι της

$$H_1 : P(X_i = 0, Y_i = 1) \neq P(X_i = 1, Y_i = 0), \text{ για τουλάχιστον ένα } i.$$

Τα ζεύγη $(0,0)$ και $(1,1)$ μπορούν να αγνοηθούν (περιπτώσεις δεσμών), αφού σε αυτά δεν παρατηρείται κάποια αλλαγή συμπεριφοράς στην κατάσταση των ατόμων. Άρα, το δείγμα μας αποτελείται από $n = \beta + \gamma$ το πλήθος άτομα. Κατασκευάζουμε τον ακριβή έλεγχο McNemar, για $n \leqslant 20$, θεωρώντας τις τ.μ.

$$Z_i = \begin{cases} 1 & , \quad \text{αν } X_i > Y_i \ \ ((1,0)) \\ 0 & , \quad \text{αν } X_i < Y_i \ \ ((0,1)) \end{cases} , \quad i = 1, \dots, n.$$

Τότε $Z_1, \dots, Z_n$ είναι ένα τυχαίο δείγμα από την κατανομή $B(1,p)$, με $p = P(Z_i = 1) = P(X_i > Y_i) = P(1,0)$ και $P(1,0) = P(X_i = 1, Y_i = 0)$. Χρησιμοποιούμε την ελεγχοσυνάρτηση $T = \sum_{i=1}^{n} Z_i$, η οποία εκφράζει το πλήθος των αποτελεσμάτων $(1,0)$ στο δείγμα, δηλαδή η τιμή της είναι $\tau = \#(1,0) = \gamma$. Στη συνέχεια, κατασκευάζουμε έλεγχο για το διωνυμικό πρόβλημα

$$\text{(A)} \quad H_0 : p = 0.5 \ , H_1 : p \neq 0.5$$

σύμφωνα με τη θεωρία που αναπτύχθηκε στην προηγούμενη ενότητα, ως εξής:

$$\phi(z) = \begin{cases} 1 & , \quad T \leqslant c \text{ ή } T \geqslant n - c, \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

όπου η σταθερά c υπολογίζεται από τη σχέση $P(T \leqslant c \,|\, p = 0.5) \approx a/2$ και η p-τιμή είναι

$$p\text{-τιμή} = \begin{cases} 2P(T \leqslant \tau | p = 0.5) & , \quad \tau \leqslant n/2, \\ 2P(T \geqslant \tau | p = 0.5) & , \quad \tau > n/2. \end{cases}$$

Παράδειγμα 5.8. Ένας κουρέας σκέφτεται να αυξήσει την τιμή του κουρέματος κατά ένα ευρώ και να δώσει στους πελάτες ένα κουπόνι για ένα δωρεάν ποτό στο διπλανό μπαράκι. Διεξήγαγε μια έρευνα και επέλεξε τυχαία 50 άτομα που περνούσαν από το κουρείο του (είτε αυτοί ήταν πελάτες του είτε όχι) και τους εξήγησε τι σκέφτεται να κάνει. Προέκυψε ότι 10% των πελατών του είπαν ότι θα πήγαιναν αλλού να κουρευτούν, ενώ 20% των μη πελατών του ισχυρίστηκε ότι θα γίνουν πελάτες του κουρείου του. Αν μόνο 10 άτομα του δείγματος είναι τωρινοί πελάτες του κουρείου, πιστεύετε ότι η προτεινόμενη αλλαγή της πολιτικής του κουρέα θα επηρεάσει τον αριθμό των πελατών του; Να διεξάγετε κατάλληλο έλεγχο σε ε.σ. 5%. Υπολογίστε και ερμηνεύστε την p-τιμή του παραπάνω ελέγχου.

Λύση Παραδείγματος 5.8. Ορίζουμε την τ.μ. X, η οποία καταγράφει την κατάσταση του ατόμου πριν την αλλαγή στην τιμή του κουρέματος, και την τ.μ. Y, η οποία καταγράφει την κατάσταση του ατόμου μετά την αλλαγή στην τιμή του κουρέματος. Θεωρούμε τις καταστάσεις 0 : το άτομο είναι πελάτης και 1 : το άτομο δεν είναι πελάτης. Οπότε προκύπτει ο παρακάτω πίνακας ταξινόμησης των τ.μ. X, Y

		ως προς Y_i	
		0	1
ως προς X_i	0	$\alpha = 9$	$\beta = 1$
	1	$\gamma = 8$	$\delta = 32$

Πίνακας 5.2: Ταξινόμηση ως προς τις τ.μ. X, Y. Δεδομένα Παραδείγματος 5.8.

Θεωρούμε τις τ.μ.

$$Z_i = \begin{cases} 1 & , \quad \text{αν } X_i > Y_i \text{ (δηλαδή } (X_i, Y_i) = (1,0)) \\ 0 & , \quad \text{αν } X_i < Y_i \text{ (δηλαδή } (X_i, Y_i) = (0,1)) \end{cases} , \quad i = 1, \dots, n.$$

Τότε οι τ.μ. $Z_1, \dots, Z_n$ αποτελούν τυχαίο δείγμα από την κατανομή $B(1,p)$, με πιθανότητα επιτυχίας $p = P(Z_i = 1) = P(X_i > Y_i) = P(1,0)$. Το μέγεθος του δείγματος είναι $n = \beta + \gamma = 1 + 8 = 9$ και εφόσον μας ενδιαφέρει αν η πολιτική του κουρέα θα επηρεάσει τον αριθμό των πελατών του, χρησιμοποιούμε το πρόβλημα ελέγχου

$$H_0 : p = 0.5 \, , H_1 : p \neq 0.5,$$

με ελεγχοσυνάρτηση:

$$\phi(z) = \begin{cases} 1 & , \quad T \leqslant c \text{ ή } T \geqslant n - c, \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

όπου η σταθερά c υπολογίζεται από τη σχέση $P(T \leqslant c \,|p = 0.5) \approx a/2$.

Από το δείγμα $\tau = \gamma = 8$, ενώ χρησιμοποιώντας τον Πίνακα Π.7 του Παραρτήματος, αφού $T \overset{H_0}{\sim} B(9, 0.5)$, έπεται ότι $P(T \leqslant 1) = 0.0195$. Η συγκεκριμένη πιθανότητα είναι η πλησιέστερη στην πιθανότητα $a/2 = 0.025$, άρα $c = 6$. Επίσης, το μέγεθος του ελέγχου, λόγω συμμετρίας της κατανομής της σ.σ. T, είναι $2 \cdot 0.0195 = 0.039$. Οπότε, ο έλεγχος γίνεται

$$\phi(z) = \begin{cases} 1 & , \quad T \leqslant 1 \text{ ή } T \geqslant 8, \\ 0 & , \quad \text{διαφορετικά.} \end{cases}$$

Αφού $\tau = 8 \geqslant 8$, συμπεραίνουμε ότι απορρίπτουμε την H_0 σε ε.σ. 3.9%, δηλαδή η πολιτική του κουρέα θα επηρεάσει τον αριθμό των πελατών του.

Επειδή $\tau = 8 > 4.5 = n/2$, η p-τιμή του ελέγχου είναι

$$2P(T \geqslant 8 | p = 0.5) = 2(1 - P(T < 8 | p = 0.5)) = 2(1 - P(T \leqslant 7 | p = 0.5)) = 2(1 - 0.9805) = 0.039.$$

□

Αν το μέγεθος του δείγματος είναι αρκετά μεγάλο (συνήθως $n > 20$), τότε εφαρμόζεται ασυμπτωτικός έλεγχος με χρήση ενός z-test (βλ. Παρατήρηση 5.7). Σε αυτήν την περίπτωση η τ.μ.

$$Z = \frac{T - \mathsf{E}(T)}{\sqrt{\mathsf{Var}(T)}} \xrightarrow{d} \mathcal{N}(0,1)$$

ή διαφορετικά

$$Z = \frac{T - n/2}{\sqrt{n/4}} \xrightarrow{d} \mathcal{N}(0,1).$$

Επειδή $n = \beta + \gamma$ και $T = \gamma$, προκύπτει ότι

$$Z = \frac{\gamma - (\beta+\gamma)/2}{\sqrt{(\beta+\gamma)/4}} = \frac{\gamma - \beta}{\sqrt{\beta+\gamma}} \xrightarrow{d} \mathcal{N}(0,1).$$

Επομένως, η τ.μ. $T_1 = Z^2$ ακολουθεί (ασυμπτωτικά) χι-τετράγωνο κατανομή με έναν βαθμό ελευθερίας, δηλαδή $T_1 \xrightarrow{d} \chi^2_1$.

Στη βιβλιογραφία, ο έλεγχος McNemar παρουσιάστηκε, αρχικά, ως ένας ασυμπτωτικός χ^2 έλεγχος, με ελεγχοσυνάρτηση $T_1 = \dfrac{(\gamma-\beta)^2}{\beta+\gamma}$, οπότε για το πρόβλημα (Α) παίρνουμε απόφαση μέσω του παρακάτω ελέγχου, ο οποίος είναι μεγέθους a,

$$\phi(z) = \begin{cases} 1 & , \quad T_1 > \chi^2_{1,a} \\ 0 & , \quad \text{διαφορετικά.} \end{cases}$$

Ο Edwards (1948) πρότεινε μία διόρθωση συνέχειας θεωρώντας τη στατιστική συνάρτηση

$$T_1 = \frac{(|\gamma - \beta| - 1)^2}{\beta + \gamma},$$

η οποία χρησιμοποιείται ως επί το πλείστον στις εφαρμογές (βλ., μεταξύ άλλων, Riffenburgh, 2006; Kavzoglu, 2017).

Παράδειγμα 5.9. Ένας ερευνητής θέλει να διαπιστώσει αν ένα φάρμακο έχει επίδραση πάνω σε μία συγκεκριμένη ασθένεια. Σε ένα δείγμα 314 ατόμων, διαπιστώνεται ότι 222 άτομα έχουν την εν λόγω ασθένεια. Μετά τη χορήγηση του φαρμάκου, 101 από αυτά εξακολουθούν να ασθενούν, ενώ 59 άτομα από αυτά που δεν είχαν την ασθένεια, τώρα την έχουν. Ελέγξτε αν υπάρχει κάποιου είδους επίδραση του φαρμάκου στην εν λόγω ασθένεια σε ε.σ. 1%.

Λύση Παραδείγματος 5.9. Ορίζουμε την τ.μ. X η οποία περιγράφει την κατάσταση του ατόμου πριν τη χορήγηση του φαρμάκου και την τ.μ. Y η οποία περιγράφει την κατάσταση του ατόμου μετά τη χορήγηση του φαρμάκου. Στο πλαίσιο αυτό, θεωρούμε τις καταστάσεις 0 : το άτομο είναι ασθενής, και 1 : το άτομο δεν είναι ασθενής. Τότε προκύπτει ο πίνακας ταξινόμησης που δίνεται στον Πίνακα 5.3.

		ως προς Y_i	
		0	1
ως προς X_i	0	$\alpha = 101$	$\beta = 121$
	1	$\gamma = 59$	$\delta = 33$

Πίνακας 5.3: Ταξινόμηση ως προς τις τ.μ. X, Y. Δεδομένα Παραδείγματος 5.9.

Θεωρούμε τις τ.μ.

$$Z_i = \begin{cases} 1 & , \quad \text{αν } X_i > Y_i \ \ (\text{δηλαδή αν} (X_i, Y_i) = (1,0)) \\ 0 & , \quad \text{αν } X_i < Y_i \ \ (\text{δηλαδή αν} (X_i, Y_i) = (0,1)) \end{cases} , \quad i = 1, \ldots, n.$$

Τότε, οι $Z_1, \dots, Z_n$ είναι ένα τυχαίο δείγμα από την κατανομή $B(1,p)$, με πιθανότητα επιτυχίας $p = P(Z_i = 1) = P(X_i > Y_i) = P(1,0)$. Εφόσον μας ενδιαφέρει αν υπάρχει κάποιου είδους επίδραση του φαρμάκου στην εν λόγω ασθένεια, χρησιμοποιούμε το πρόβλημα

$$H_0 : p = 0.5 \text{ , έναντι της } H_1 : p \neq 0.5.$$

Το μέγεθος του δείγματος είναι $n = \beta + \gamma = 121 + 59 = 180$, το οποίο, επειδή είναι αρκετά μεγάλο, χρησιμοποιούμε τον ασυμπτωτικό έλεγχο McNemar, ο οποίος βασίζεται στη χρήση κατανομής χι-τετράγωνο, με ελεγχοσυνάρτηση

$$\phi(\underline{z}) = \begin{cases} 1 & , \quad T_1 > \chi^2_{1,a} \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

όπου $T_1 = \dfrac{(|\gamma - \beta| - 1)^2}{\beta + \gamma}$. Από το δείγμα, η τιμή της T_1 είναι $\tau_1 = \dfrac{(|59 - 121| - 1)^2}{121 + 59} = 20.672$, ενώ από τον Πίνακα Π.4 του Παραρτήματος $\chi^2_{1,0.01} = 6.635$. Προφανώς, $\tau_1 = 20.672 > 6.635$, δηλαδή απορρίπτουμε την H_0, επομένως, υπάρχει επίδραση του φαρμάκου στην εν λόγω ασθένεια. □

Παρατήρηση 5.8. Από τη στιγμή που ο έλεγχος McNemar έχει παρουσιαστεί ως ένας έλεγχος προσήμου, είναι εύλογο να χρησιμοποιείται και σε προβλήματα με μονόπλευρες εναλλακτικές υποθέσεις, όπως αυτά παρουσιάστηκαν στην προηγούμενη ενότητα. Αυτό επιτυγχάνεται χρησιμοποιώντας τα ποσοστιαία σημεία της Διωνυμικής κατανομής, όταν $n \leqslant 20$ και κάνοντας $z - test$ για μεγαλύτερα δειγματικά μεγέθη. Για περισσότερες λεπτομέρειες παραπέμπουμε, μεταξύ άλλων, στους Conover (1998), Taillard *et al.* (2008), Agresti (2013), Hollander *et al.* (2014).

Παράδειγμα 5.10. Ρωτήσαμε 100 οδηγούς πόσα χιλιόμετρα διανύουν με το αυτοκίνητό τους κάθε χρόνο. Οι 70 από αυτούς απάντησαν ότι διανύουν περισσότερα από 10000 km, ενώ 30 από αυτούς δήλωσαν ότι διανύουν λιγότερα από 10000 km τον χρόνο. Κατόπιν, ενημερώσαμε τους ίδιους οδηγούς για την πρόθεση της κυβέρνησης να επιβάλει φόρο 0,03 ευρώ/km, οπότε το 30% αυτών που δήλωναν πριν ότι διανύουν περισσότερα από 10000 km τον χρόνο, τώρα δηλώνουν ότι θα διανύσουν λιγότερα από 10000 km τον χρόνο, ενώ το 20% από αυτούς που δήλωσαν ότι διανύουν λιγότερα από 10000 km τον χρόνο, τώρα δηλώνουν ότι θα ξεπερνούν τα 10000 km τον χρόνο. Πιστεύετε ότι η επιβολή του φόρου ανά km, επηρεάζει αρνητικά τη μετακίνηση με αυτοκίνητο; Να διεξάγετε κατάλληλο έλεγχο σε ε.σ. 7%. Ποια είναι η p-τιμή του παραπάνω ελέγχου;

Λύση Παραδείγματος 5.10. Ορίζουμε τις τ.μ. X, Y, όπου η X καταγράφει την κατηγορία στην οποία βρίσκεται ο κάθε οδηγός με βάση τον αριθμό των χιλιομέτρων που διανύει με το αυτοκίνητό του πριν λάβει γνώση της πρόθεσης της κυβέρνησης για επιβολή φόρου 0.03 ευρώ/km, ενώ η τ.μ. Y καταγράφει την κατηγορία στην οποία βρίσκεται ο κάθε οδηγός με βάση τον αριθμό των χιλιομέτρων που σκοπεύει να διανύσει με το αυτοκίνητό του μετά τη λήψη γνώσης για την πρόθεση της κυβέρνησης να επιβάλει φόρο 0.03 ευρώ/km. Οι πιθανές καταστάσεις είναι δύο και, συγκεκριμένα, στην κατάσταση 0 : ο οδηγός δηλώνει ότι διανύει με το αυτοκίνητό του περισσότερα από 10000 km τον χρόνο, ενώ στην κατάσταση 1 : ο οδηγός δηλώνει ότι διανύει με το αυτοκίνητό του λιγότερα από 10000 km τον χρόνο. Οπότε, σύμφωνα με τα δεδομένα της άσκησης, προκύπτει ο πίνακας ταξινόμησης που δίνεται στον Πίνακα 5.4.

Θεωρούμε τις τ.μ.

$$Z_i = \begin{cases} 1 & , \quad \text{αν } X_i > Y_i \ \ (\text{δηλαδή αν}(X_i, Y_i) = (1,0)) \\ 0 & , \quad \text{αν } X_i < Y_i \ \ (\text{δηλαδή αν}(X_i, Y_i) = (0,1)) \end{cases} , \quad i = 1, \dots, n.$$

Τότε, οι τ.μ. $Z_1, \dots, Z_n$ αποτελούν τυχαίο δείγμα από την κατανομή $B(1,p)$, με πιθανότητα επιτυχίας $p = P(Z_i = 1) = P(X_i > Y_i) = P(1,0)$. Εφόσον μας ενδιαφέρει αν η επιβολή του φόρου ανά km, επηρεάζει *αρνητικά* τη μετακίνηση με αυτοκίνητο, χρησιμοποιούμε το μονόπλευρο πρόβλημα ελέγχου

$$H_0 : p \geqslant 0.5 \text{ , κατά } H_1 : p < 0.5.$$

		ως προς Y_i	
		0	1
ως προς X_i	0	$\alpha = 49$	$\beta = 21$
	1	$\gamma = 6$	$\delta = 24$

Πίνακας 5.4: Ταξινόμηση ως προς τις τ.μ. X, Y. Δεδομένα Παραδείγματος 5.10.

Το μέγεθος του δείγματος είναι $n = \beta + \gamma = 21 + 6 = 27 > 20$, οπότε χρησιμοποιούμε ένα $z - test$ με

$$\phi(z) = \begin{cases} 1 & , \ Z < -z_a, \\ 0 & , \ \text{διαφορετικά.} \end{cases}$$

Η στατιστική συνάρτηση ελέγχου, με χρήση της διόρθωσης συνέχειας, είναι η

$$Z = \frac{T + 0.5 - 0.5n}{\sqrt{0.25n}},$$

με $T = \sum_{i=1}^{n} Z_i$. Από τις τιμές στον Πίνακα 5.4 έχουμε πως η τιμή της T είναι $\tau = \#(1,0) = \gamma = 6$. Επομένως, η τιμή της σ.σ. Z είναι ίση με

$$z = \frac{6 + 0.5 - 0.5 \times 27}{\sqrt{0.25 \times 27}} = -2.69.$$

Από τον Πίνακα Π.1 του Παραρτήματος, αφού το ε.σ. είναι 7%, έπεται με χρήση γραμμικής παρεμβολής, ότι η κρίσιμη τιμή για τον έλεγχο είναι, $z_{0.07} = 1.475$. Άρα, επειδή $z = -2.69 < -1.475$, απορρίπτουμε την H_0, δηλαδή η επιβολή του φόρου ανά km, επηρεάζει αρνητικά τη μετακίνηση με αυτοκίνητο σε ε.σ. 7%.

Η p-τιμή αυτού του ελέγχου είναι

$$p\text{-τιμή} = \Phi\left(\frac{\tau + 0.5 - 0.5n}{\sqrt{0.25n}}\right) = \Phi(-2.69) = 1 - \Phi(2.69) = 1 - 0.9964 = 0.0036.$$

□

5.5.2 Έλεγχος Cox-Stuart

Στην παρούσα υποενότητα θα παρουσιάσουμε μια ακόμα παραλλαγή του προσημικού ελέγχου, ο οποίος προτάθηκε από τους Cox and Stuart (1955) και μπορεί να χρησιμοποιηθεί για τον έλεγχο ύπαρξης τάσης σε μια ακολουθία παρατηρήσεων. Για περισσότερους τέτοιους ελέγχους παραπέμπουμε στο Κεφάλαιο 7 του παρόντος συγγράμματος.

Ορισμός 5.1

Μία ακολουθία παρατηρήσεων λέγεται ότι έχει τάση, αν οι τελευταίοι όροι της ακολουθίας έχουν μεγαλύτερες τιμές από τους πρώτους όρους της ακολουθίας (αυξητική τάση) ή μικρότερες τιμές από τους πρώτους όρους της ακολουθίας (πτωτική τάση).

Αν $X_1, X_2, \ldots, X_k$ είναι αμοιβαία ανεξάρτητες τ.μ., σχηματίζουμε ζεύγη της μορφής, $(X_1, X_{1+c}), (X_2, X_{2+c}), \ldots, (X_{k-c}, X_k)$, όπου

$$c = \begin{cases} k/2 & , \ k \text{ άρτιος}, \\ (k-1)/2 & , \ k \text{ περιττός}. \end{cases} \tag{5.7}$$

Μας ενδιαφέρουν τα n ζεύγη, όπου δεν θα υπάρχουν ίδιες τιμές (ισοβαθμίες), δηλαδή προκύπτει ότι $n = \#(X_i < X_{i+c}) + \#(X_i > X_{i+c})$, με $\#(X_i < X_{i+c})$ και $\#(X_i > X_{i+c})$ να συμβολίζουν το πλήθος των X_i που

είναι μικρότερα ή μεγαλύτερα, αντίστοιχα, από X_{i+c}. Αν συμβεί $X_i = X_{i+c}$, τότε έχουμε περίπτωση δεσμού (ισοβαθμίας) και τα αντίστοιχα ζεύγη αφαιρούνται από το αρχικό δείγμα μεγέθους k. Παρατηρήστε, επίσης, πως στην περίπτωση που το k είναι περιττός, τότε η μεσαία παρατήρηση εξαιρείται κατά τη δημιουργία ζευγών. Διαθέτοντας $n \leqslant k$ ζεύγη παρατηρήσεων μπορούμε να εφαρμόσουμε προσημικό έλεγχο θεωρώντας τις τ.μ.

$$Z_i = \begin{cases} 1 & , \quad \text{αν } X_i > X_{i+c} \text{ (ή +)} \\ 0 & , \quad \text{αν } X_i < X_{i+c} \text{ (ή −)} \end{cases} , \; i = 1, \ldots, n.$$

Τότε, οι $Z_1, \ldots, Z_n$ αποτελούν ένα τυχαίο δείγμα από την κατανομή $B(1, p)$, με πιθανότητα επιτυχίας $p = P(Z_i = 1) = P(X_i > X_{i+c}) = P(+)$.

Οι έλεγχοι που μας ενδιαφέρουν δίνονται παρακάτω

(Α) H_0 : Δεν υπάρχει τάση, H_1 : Υπάρχει τάση
(Β) H_0 : Δεν υπάρχει πτωτική τάση, H_1 : Υπάρχει πτωτική τάση
(Γ) H_0 : Δεν υπάρχει αυξητική τάση, H_1 : Υπάρχει αυξητική τάση

Αξίζει να αναφέρουμε πως π.χ. για το Πρόβλημα (Α), είτε οι $X_1, X_2, \ldots, X_k$ θα είναι ισόνομες τ.μ., αν αληθεύει η H_0, διαφορετικά θα υπάρχει τάση στα δεδομένα, αν αληθεύει η H_1. Επίσης, για το Πρόβλημα (Α), δεν είναι δύσκολο να διαπιστώσουμε ότι, όταν είναι αληθής η $H_0 : P(+) = P(-)$, έπεται ότι:

$$P(+) = P(-) \Leftrightarrow p = 1 - p \Leftrightarrow p = 1/2.$$

Επομένως, το διωνυμικό πρόβλημα που καλούμαστε να εξετάσουμε είναι

$$\text{(Α)} \;\; H_0 : p = 0.5 \;, \;\; H_1 : p \neq 0.5$$

Όσον αφορά το Πρόβλημα (Β), η απόρριψη της H_0 σημαίνει ότι $P(+) > P(-)$ και, άρα, $p > 1 - p$ ή $p > 1/2$. Δηλαδή προκύπτει το μονόπλευρο διωνυμικό πρόβλημα ελέγχου

$$\text{(Β)} \;\; H_0 : p \leqslant 0.5 \;, \;\; H_1 : p > 0.5.$$

Τέλος, για το Πρόβλημα (Γ), η απόρριψη της H_0 σημαίνει ότι $P(+) < P(-)$ και, άρα, $p < 1 - p$ ή $p < 1/2$. Οπότε, σε αυτήν την περίπτωση αντιμετωπίζουμε το μονόπλευρο διωνυμικό πρόβλημα ελέγχου

$$\text{(Γ)} \;\; H_0 : p \geqslant 0.5 \;, \;\; H_1 : p < 0.5.$$

Αυτά τα προβλήματα αντιμετωπίστηκαν στην προηγούμενη ενότητα είτε για μεγάλες, είτε για μικρές τιμές του n και δεν χρειάζεται να επαναλάβουμε τις μορφές των στατιστικών συναρτήσεων ελέγχου, κρίσιμων περιοχών και τύπων υπολογισμού των αντιστοίχων p-τιμών. Για να γίνει κατανοητή η εφαρμογή του ελέγχου των Cox-Stuart στην πράξη, δίνουμε ένα παράδειγμα.

Παράδειγμα 5.11. Παρακάτω δίνεται το πλήθος των εργασιών ανά έτος, οι οποίες περιέχουν ιατρικά δεδομένα και είναι δημοσιευμένες σε ένα διεθνές επιστημονικό περιοδικό Στατιστικής:

$$11 \,,\; 6 \,,\; 14 \,,\; 13 \,,\; 18 \,,\; 14 \,,\; 11 \,,\; 22 \,,\; 19 \,,\; 19 \,,\; 25 \,,\; 24 \,,\; 38 \,,\; 19 \,,\; 25 \,,\; 31 \,,\; 19.$$

Ελέγξτε, σε ε.σ. 5%, αν υπάρχει ανοδική (αυξητική) τάση στη δημοσίευση εργασιών με ιατρικά δεδομένα για αυτό το περιοδικό.

Λύση Παραδείγματος 5.11. Ορίζουμε την τ.μ. X η οποία μετράει το πλήθος των εργασιών ανά έτος, με $X_1, X_2, \ldots, X_k$, να είναι το διαθέσιμο δείγμα μεγέθους $k = 17$. Καθώς $k = 17$ περιττός, από τη σχέση (5.7)

είναι $c = \frac{(17+1)}{2} = 9$ και επομένως σχηματίζουμε $k - c = 17 - 9 = 8$ το πλήθος ζεύγη της μορφής $(X_1, X_{1+c}), (X_2, X_{2+c}), \ldots, (X_{k-c}, X_k)$. Δηλαδή, σχηματίζουμε τα παρακάτω ζεύγη παρατηρήσεων:

Ζεύγος παρατηρήσεων	Τιμές	Αποτέλεσμα	Ζεύγος παρατηρήσεων	Τιμές	Αποτέλεσμα
(X_1, X_{10})	(11,19)	−	(X_5, X_{14})	(18,19)	−
(X_2, X_{11})	(6,25)	−	(X_6, X_{15})	(14,25)	−
(X_3, X_{12})	(14,24)	−	(X_7, X_{16})	(11,31)	−
(X_4, X_{13})	(13,38)	−	(X_8, X_{17})	(22,19)	+

Παρατηρήστε ότι η ένατη παρατήρηση δεν θα χρησιμοποιηθεί για τη δημιουργία ζεύγους παρατηρήσεων, καθώς k περιττός και είναι η μεσαία παρατήρηση.

Επιπλέον, θεωρούμε τις τ.μ.

$$Z_i = \begin{cases} 1 & , \text{ αν } X_i > X_{i+c} \text{ (ή +)} \\ 0 & , \text{ αν } X_i < X_{i+c} \text{ (ή −)} \end{cases}, \ i = 1, \ldots, n.$$

Οι τ.μ. $Z_1, \ldots, Z_n$, όπου $n = 8$, αποτελούν ένα τυχαίο δείγμα από την κατανομή $B(1,p)$, με πιθανότητα επιτυχίας $p = P(Z_i = 1) = P(X_i > X_{i+c}) = P(+)$. Θέλουμε να ελέγξουμε αν υπάρχει ανοδική τάση, οπότε ασχολούμαστε με το πρόβλημα:

$$H_0 : p \geqslant 0.5 \ , \ \ H_1 : p < 0.5$$

με ελεγχοσυνάρτηση:

$$\phi(z) = \begin{cases} 1 & , \ T \leqslant c \\ 0 & , \text{ διαφορετικά,} \end{cases}$$

με τη σταθερά c να υπολογίζεται από τη σχέση $P(T \leqslant c \,|\, p = 0.5) = a$. Υπό την H_0, η σ.σ. $T = \sum_{i=1}^{8} Z_i \sim B(8,0.5)$, επομένως, χρησιμοποιώντας τον Πίνακα Π.7 του Παραρτήματος, βρίσκουμε ότι $P(T \leqslant 1) = 0.0352$. Η συγκεκριμένη πιθανότητα είναι η πλησιέστερη στην πιθανότητα $a = 0.05$ (επιθυμητό ε.σ. a). Άρα, $c = 1$, ενώ το ακριβές μέγεθος του ελέγχου είναι $a = 0.0352$. Επειδή η τιμή της T είναι ίση με $\tau = \#(X_i > X_{i+c}) = 1 \leqslant 1 = c$, το συμπέρασμα είναι ότι απορρίπτουμε την H_0 (σε ε.σ. 3.52%), άρα υπάρχει ανοδική τάση στη δημοσίευση εργασιών με ιατρικά δεδομένα για αυτό το περιοδικό Στατιστικής.

Επίσης, η p-τιμή του ελέγχου είναι η $P(T \leqslant 1 \,|\, H_0) = 0.0352$. Άρα, το ελάχιστο επίπεδο σημαντικότητας για το οποίο απορρίπτεται η μηδενική υπόθεση της μη ύπαρξης τάσης έναντι της υπόθεσης ύπαρξης ανοδικής τάσης, είναι ίσο με 3.52%. □

Παρατήρηση 5.9. Πολλές φορές η περιοδικότητα ενός φαινομένου παίζει αρκετά σημαντικό ρόλο και έχει ως αποτέλεσμα να μην μπορούμε να διαπιστώσουμε μέσω του ελέγχου των Cox-Stuart ότι υπάρχει τάση στις τιμές της ακολουθίας. Για παράδειγμα, η άφιξη τουριστών στην Ελλάδα ενδεχομένως να έχει μια αυξητική τάση από τον Μάρτιο έως και τον Αύγουστο και μία πτωτική τάση από τον Σεπτέμβριο έως τον Φεβρουάριο. Για να μην μας διαφύγει αυτή η τάση, ο Conover (1998) προτείνει να εφαρμόζουμε τον προσημικό έλεγχο των Cox and Stuart σε συγκεκριμένες συνθήκες ύπαρξης περιοδικότητας επανατοποθετώντας το δείγμα. Η μέθοδος περιγράφεται μέσω του παραδείγματος που ακολουθεί.

Παράδειγμα 5.12. (Conover, 1998) Ο αριθμός των αυγών που γεννήθηκαν από μια ομάδα εντόμων σε ένα ερευνητικό εργαστήριο μετράται ανά ώρα κατά τη διάρκεια ενός πειράματος 24 ωρών. Θέλουμε να κάνουμε τον έλεγχο

H_0 : Το πλήθος των αυγών που γεννιούνται κατά τη διάρκεια του εικοσιτετραώρου αποτελεί τυχαίο δείγμα

έναντι της

H_1 : ο αριθμός των αυγών τείνει να είναι ελάχιστος περίπου στις 14.15', στη συνέχεια αυξάνει, ώστε να πιάνει τη μέγιστη τιμή του στις 02.15' και εν συνεχεία μειώνεται μέχρι τις 14.15'.

Τα διαθέσιμα δεδομένα δίνονται στον επόμενο πίνακα. Ο έλεγχος να γίνει με χρήση της p-τιμής, σε ε.σ. 2%.

Ώρα	Αριθμός Αυγών	Ώρα	Αριθμός Αυγών	Ώρα	Αριθμός Αυγών
09.00'	151	17.00'	83	01.00'	286
10.00'	119	18.00'	166	02.00'	235
11.00'	146	19.00'	143	03.00'	223
12.00'	111	20.00'	116	04.00'	176
13.00'	63	21.00'	163	05.00'	176
14.00'	84	22.00'	208	06.00'	174
15.00'	60	23.00'	283	07.00'	139
16.00'	109	24.00'	296	08.00'	137

Λύση Παραδείγματος 5.12. Στο συγκεκριμένο πρόβλημα, πρέπει να παρατηρήσουμε αρχικά πως, αν η H_1 είναι αληθής, τότε ο αριθμός των αυγών κοντά στις 14.15' τείνει να είναι ο ελάχιστος, ενώ, καθώς πλησιάζουμε στις 02.15', ο αριθμός των αυγών τείνει να είναι ο μέγιστος. Δηλαδή, στα δεδομένα θα υπάρχει ένδειξη για αυξητική τάση. Άρα, τα δεδομένα αναδιατάσσονται, ξεκινώντας από τις ώρες που είναι πιο κοντά στις 14.15' και καταλήγοντας στις ώρες που είναι πιο κοντά στις 02.15'. Έτσι, προκύπτει ο παρακάτω πίνακας

14.00'	84	10.00'	119	06.00'	174
15.00'	60	19.00'	143	23.00'	283
13.00'	63	09.00'	151	05.00'	176
16.00'	109	20.00'	116	00.00'	296
12.00'	111	08.00'	137	04.00'	176
17.00'	83	21.00'	163	01.00'	286
11.00'	146	07.00'	139	03.00'	223
18.00'	166	22.00'	208	02.00'	235

Στη συνέχεια, θα χρησιμοποιήσουμε το τεστ των Cox-Stuart για να κάνουμε έλεγχο για την ύπαρξη αυξητικής τάσης (μονόπλευρος έλεγχος). Ορίζουμε την τ.μ. X η οποία μετράει τον αριθμό των αυγών ανά ώρα, με $X_1, X_2, \dots, X_k$, να είναι το διαθέσιμο δείγμα μεγέθους $k = 24$. Σχηματίζουμε, επομένως, $c = 24/2 = 12$ το πλήθος ζεύγη της μορφής $(X_1, X_{1+c}), (X_2, X_{2+c}), \dots, (X_{k-c}, X_k)$. Άρα, σχηματίζουμε τα παρακάτω ζεύγη παρατηρήσεων

Ζεύγος παρατηρήσεων	Τιμές	Αποτέλεσμα	Ζεύγος παρατηρήσεων	Τιμές	Αποτέλεσμα
(X_1, X_{13})	$(84,137)$	–	(X_7, X_{19})	$(146,176)$	–
(X_2, X_{14})	$(60,163)$	–	(X_8, X_{20})	$(166,296)$	–
(X_3, X_{15})	$(63,139)$	–	(X_9, X_{21})	$(119,176)$	–
(X_4, X_{16})	$(109,208)$	–	(X_{10}, X_{22})	$(143,286)$	–
(X_5, X_{17})	$(111,174)$	–	(X_{11}, X_{23})	$(151,223)$	–
(X_6, X_{18})	$(83,283)$	–	(X_{12}, X_{24})	$(116,235)$	–

Επιπλέον, θεωρούμε τις τ.μ.

$$Z_i = \begin{cases} 1 & , \text{ αν } X_i > X_{i+c} \text{ (ή } +) \\ 0 & , \text{ αν } X_i < X_{i+c} \text{ (ή } -) \end{cases}, \ i = 1, \dots, n.$$

Οι τ.μ. $Z_1, \dots, Z_n$, όπου $n = 12$, αποτελούν ένα τυχαίο δείγμα από την κατανομή $B(1, p)$, με πιθανότητα επιτυχίας $p = P(Z_i = 1) = P(X_i > X_{i+c}) = P(+)$. Θέλουμε να ελέγξουμε αν υπάρχει ανοδική τάση, οπότε ασχολούμαστε με το πρόβλημα

$$H_0 : p \geqslant 0.5 \ , \ \ H_1 : p < 0.5$$

με ελεγχοσυνάρτηση:

$$\phi(\underline{z}) = \begin{cases} 1 & , \ T \leqslant c, \\ 0 & , \ \text{διαφορετικά}, \end{cases}$$

με τη σταθερά c να υπολογίζεται από τη σχέση, $P(T \leqslant c \big| p = 0.5) = a$. Υπό την H_0, $T = \sum_{i=1}^{12} Z_i \sim B(12, 0.5)$, επομένως, χρησιμοποιώντας τον Πίνακα Π.8 του Παραρτήματος, $P(T \leqslant 2) = 0.0193$, δηλαδή $c = 2$, καθώς για το συγκεκριμένο c η αντίστοιχη πιθανότητα είναι κοντά στην επιθυμητή τιμή για το ε.σ. 5%. Δεν είναι δύσκολο να διαπιστώσουμε ότι το ακριβές μέγεθος του ελέγχου είναι $a = 0.0193$. Επειδή η τιμή της σ.σ. T είναι $\tau = \#(X_i > X_{i+c}) = 0 \leqslant 2 = c$, το συμπέρασμα είναι ότι απορρίπτουμε την H_0, άρα υπάρχει σαφής ένδειξη ότι ο μέγιστος αριθμός αυγών καταμετράται πλησιέστερα στις 02.15' και ο ελάχιστος πλησιέστερα στις 14.15'.

Επίσης, η p-τιμή του ελέγχου είναι η $P(T \leqslant 0 \big| H_0) = 0.00024$ (πάλι με χρήση του Πίνακα Π.8 του Παραρτήματος). Άρα, σε επίπεδο σημαντικότητας 5%, απορρίπτεται η H_0. □

Παρατήρηση 5.10. Ο Stuart (1956) έδειξε ότι η συμπεριφορά του ελέγχου των Cox και Stuart δεν είναι η κατάλληλη, συγκρινόμενη με άλλους παραμετρικούς ελέγχους για την ύπαρξη τάσης, όταν τα δεδομένα προέρχονται από την κανονική κατανομή, χρησιμοποιώντας ως κριτήριο την ασυμπτωτική σχετική αποτελεσματικότητα (για λεπτομέρειες βλ. Conover, 1998). Στο Κεφάλαιο 7 του παρόντος συγγράμματος θα παρουσιαστούν μεθοδολογίες για τον έλεγχο της τυχαιότητας του δείγματος, καθώς και για τον έλεγχο ύπαρξης ανοδικής ή πτωτικής τάσης.

5.5.3 Έλεγχος συσχετίσεων

Στην παρούσα υποενότητα θα παρουσιάσουμε μία άλλη χρήση του Προσημικού Ελέγχου και, συγκεκριμένα, θα δούμε ότι μπορεί να χρησιμοποιηθεί ως έλεγχος ύπαρξης συσχέτισης μεταξύ δύο τυχαίων μεταβλητών X και Y. Ο έλεγχος ύπαρξης συσχέτισης μεταξύ δύο μεταβλητών X και Y είναι αρκετά σημαντικός στη Στατιστική Συμπερασματολογία. Θα λέμε ότι υπάρχει συσχέτιση μεταξύ των τυχαίων μεταβλητών X και Y, αν ισχύουν τα εξής: είτε όσο αυξάνονται οι τιμές της μίας τ.μ., αυξάνονται οι τιμές της άλλης, οπότε έχουμε **θετική συσχέτιση** μεταξύ αυτών, είτε όσο μειώνονται οι τιμές της μίας τ.μ., αυξάνονται οι τιμές της άλλης, σε αυτήν την περίπτωση οι τ.μ. είναι **αρνητικά συσχετισμένες**. Αν οι X, Y δεν είναι αρνητικά ή θετικά συσχετισμένες, τότε λέμε ότι είναι ασυσχέτιστες.

Για να μπορέσουμε να κάνουμε έλεγχο για ύπαρξη συσχέτισης μεταξύ δύο μεταβλητών χρησιμοποιώντας τον Προσημικό Έλεγχο, εργαζόμαστε όπως περιγράφεται στη συνέχεια. Διατάσσουμε, αρχικά, τις τιμές της μίας τυχαίας μεταβλητής και ελέγχουμε αν για τις αντίστοιχες τιμές της άλλης τυχαίας μεταβλητής υπάρχει κάποιου είδους τάση, είτε θετική είτε αρνητική. Δηλαδή εφαρμόζουμε τον έλεγχο των Cox και Stuart, ο οποίος αποτελεί παραλλαγή του Προσημικού Ελέγχου, στο άλλο δείγμα.

Σε αυτό το πλαίσιο, οι έλεγχοι που μας ενδιαφέρουν είναι οι ακόλουθοι:

(Α) H_0 : Δεν υπάρχει συσχέτιση, H_1 : Υπάρχει συσχέτιση.
(Β) H_0 : Δεν υπάρχει αρνητική συσχέτιση, H_1 : Υπάρχει αρνητική συσχέτιση.
(Γ) H_0 : Δεν υπάρχει θετική συσχέτιση, H_1 : Υπάρχει θετική συσχέτιση.

Για το Πρόβλημα (Α), όταν η H_0 είναι αληθής, έχουμε ότι δεν υπάρχει τάση στους όρους του δείγματος που δεν διατάσσουμε, επομένως, το διωνυμικό πρόβλημα που καλούμαστε να εξετάσουμε είναι

$$\text{(Α)} \quad H_0 : p = 0.5 \,, \quad H_1 : p \neq 0.5.$$

Όσον αφορά το Πρόβλημα (Β), απόρριψη της H_0 σημαίνει ότι υπάρχει πτωτική τάση στους όρους του δείγματος που δεν διατάσσουμε, δηλαδή προκύπτει το μονόπλευρο διωνυμικό πρόβλημα ελέγχου:

$$\text{(Β)} \quad H_0 : p \leqslant 0.5 \,, \quad H_1 : p > 0.5.$$

Τέλος, για το Πρόβλημα (Γ), απόρριψη της H_0 σημαίνει ότι υπάρχει αυξητική τάση στους όρους του δείγματος που δεν διατάσσουμε, άρα σε αυτήν την περίπτωση αντιμετωπίζουμε το μονόπλευρο διωνυμικό πρόβλημα

ελέγχου:

$$(\Gamma)\ \ H_0 : p \geqslant 0.5\ ,\ \ H_1 : p < 0.5.$$

Αυτά τα προβλήματα αντιμετωπίστηκαν σε προηγούμενες ενότητες είτε για μεγάλες είτε για μικρές τιμές του n και δεν θα παρουσιάσουμε εκ νέου τις μορφές των στατιστικών συναρτήσεων ελέγχου, κρίσιμων περιοχών και τύπων υπολογισμού των αντιστοίχων p-τιμών. Για να γίνει κατανοητή η εφαρμογή του ελέγχου για την ύπαρξη συσχέτισης με χρήση του ελέγχου των Cox και Stuart, δίνουμε ένα παράδειγμα.

Παράδειγμα 5.13. Ο παρακάτω πίνακας αναφέρεται στις αντιδράσεις ενός τυχαίου δείγματος 10 ασθενών σε δύο φάρμακα A και B. Τα αποτελέσματα αποτελούν τιμές ενός δείκτη αντίδρασης και οι αρνητικές τιμές δείχνουν απόκριση αντίθετη από την επιθυμητή.

Ασθενής	1	2	3	4	5	6	7	8	9	10
A	7	−1.6	−0.2	−1.2	−0.1	3.4	3.7	0.8	0	2
B	1.9	0.8	1.1	0.1	−0.1	4.4	5.5	1.6	4.6	3.4

Να ελεγχθεί αν υπάρχει συσχέτιση μεταξύ των αντιδράσεων ενός ασθενούς στα δύο φάρμακα. Χρησιμοποιήστε ε.σ. 5%.

Λύση Παραδείγματος 5.13. Ορίζουμε την τ.μ. X, η οποία μετράει την αντίδραση του/της ασθενούς στο φάρμακο A, και, αντίστοιχα, την τ.μ. Y, η οποία μετράει την αντίδραση του/της ασθενούς στο φάρμακο B. Διατάσσουμε κατά αύξουσα σειρά τις τιμές της τ.μ. X και αντιστοιχούμε σε αυτές τις τιμές της τ.μ. Y, όπως φαίνεται παρακάτω:

$X_{(i)}$	−1.6	−1.2	−0.2	−0.1	0	0.8	2	3.4	3.7	7
$Y_i^{(*)}$	0.8	0.1	1.1	−0.1	4.6	1.6	3.4	4.4	5.5	1.9

Σημειώνεται πως η $X_{(i)}$ είναι η i-οστή διατεταγμένη παρατήρηση ($i = 1,2,\dots,k$) κατά αύξουσα τάξη μεγέθους, ενώ η $Y_i^{(*)}$ είναι η παρατήρηση από το δείγμα των $Y_1,\dots,Y_k$ η οποία «ζευγαρώνει» με τη $X_{(i)}$.

Στη συνέχεια, εφαρμόζουμε τον έλεγχο των Cox και Stuart στο δείγμα των $Y_1^{(*)}, Y_2^{(*)},\dots,Y_k^{(*)}$, με $k = 10$. Σχηματίζουμε, επομένως, $c = \frac{10}{2} = 5$ το πλήθος ζεύγη της μορφής $(Y_1^{(*)}, Y_{1+c}^{(*)})$, $(Y_2^{(*)}, Y_{2+c}^{(*)})$, ..., $(Y_{k-c}^{(*)}, Y_k^{(*)})$. Άρα, σχηματίζουμε τα παρακάτω ζεύγη παρατηρήσεων:

Ζεύγος παρατηρήσεων	Τιμές	Αποτέλεσμα
$(Y_1^{(*)}, Y_6^{(*)})$	(0.8,1.6)	−
$(Y_2^{(*)}, Y_7^{(*)})$	(0.1,3.4)	−
$(Y_3^{(*)}, Y_8^{(*)})$	(1.1,4.4)	−
$(Y_4^{(*)}, Y_9^{(*)})$	(−0.1,5.5)	−
$(Y_5^{(*)}, Y_{10}^{(*)})$	(4.6,1.9)	+

Επιπλέον, θεωρούμε τις τ.μ.

$$Z_i = \begin{cases} 1 & ,\ \text{αν } Y_i^{(*)} > Y_{i+c}^{(*)}\ (\text{ή } +) \\ 0 & ,\ \text{αν } Y_i^{(*)} < Y_{i+c}^{(*)}\ (\text{ή } -) \end{cases},\ i = 1,\dots,n.$$

Οι τ.μ. $Z_1,\dots,Z_n$, $n = 5$, αποτελούν ένα τυχαίο δείγμα από την κατανομή $B(1,p)$, με πιθανότητα επιτυχίας $p = P(Z_i = 1) = P(Y_i^{(*)} > Y_{i+c}^{(*)}) = P(+)$. Εφόσον μας ενδιαφέρει να εξετάσουμε αν υπάρχει κάποιου είδους συσχέτιση μεταξύ των τ.μ. X και Y, θεωρούμε το διωνυμικό πρόβλημα ελέγχου:

$$H_0 : p = 0.5\ ,\ \ H_1 : p \neq 0.5$$

με ελεγχοσυνάρτηση:

$$\phi(z) = \begin{cases} 1 & , \quad T \leqslant c \text{ ή } T \geqslant n - c \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

με τη σταθερά c να υπολογίζεται από τη σχέση, $P(T \leqslant c \,|p = 0.5) \approx a/2$. Υπό την H_0, η σ.σ. $T = \sum_{i=1}^{5} Z_i \sim B(5, 0.5)$, άρα, με χρήση του Πίνακα Π.7 του Παραρτήματος, έπεται ότι $c = 0$, αφού η $P(T \leqslant 0) = 0.0312$. Η συγκεκριμένη πιθανότητα είναι η πλησιέστερη στο $a/2 = 0.025$. Άρα, αφού ο έλεγχος είναι δίπλευρος και λόγω και της συμμετρίας της κατανομής της T, το μέγεθος του ελέγχου είναι $2 \cdot 0.0315 = 0.0624$. Η μορφή της κρίσιμης περιοχής είναι $K : T \leqslant 0$ ή $T \geqslant 5$, ενώ από τα διαθέσιμα δεδομένα, η τιμή της T είναι $\tau = \#(Y_i > Y_{i+c}) = \#(+) = 1$. Καθώς η τιμή τ της T δεν ανήκει στην κρίσιμη περιοχή του ελέγχου, δεν μπορούμε να απορρίψουμε την H_0 σε ε.σ. 6.24%, δηλαδή δεν μπορούμε να απορρίψουμε την υπόθεση ότι οι τ.μ. X και Y είναι ασυσχέτιστες.

Επίσης, για την p-τιμή του ελέγχου, αφού $\tau = 1 < 2.5 = n/2$, έχουμε ότι ισούται με $2P(T \leqslant 1 \,|H_0) = 0.375$. Άρα, το ελάχιστο επίπεδο σημαντικότητας για το οποίο απορρίπτεται η μηδενική υπόθεση της μη ύπαρξης τάσης έναντι της υπόθεσης ύπαρξης τάσης, είναι ίσο με 37.5% και, άρα, δεν μπορούμε να απορρίψουμε την υπόθεση ότι οι τ.μ. X και Y είναι ασυσχέτιστες. □

Παρατήρηση 5.11. Στις περιπτώσεις ύπαρξης ισοβαθμιών (ties) εντός των δειγμάτων, τότε διατάσσουμε εκείνο το δείγμα με τις λιγότερες ισοβαθμίες και εφαρμόζουμε τον έλεγχο των Cox και Stuart στο άλλο, αφού έχουμε αντιστοιχίσει τις τιμές του. Γενικά, δεν υπάρχει προτίμηση στο ποιο από τα δύο δείγματα πρέπει να διατάξουμε.

Παράδειγμα 5.14. Θεωρούμε τις επιδόσεις 9 φοιτητών/τριών στα μαθήματα Πιθανότητες Ι (X) και Πιθανότητες ΙΙ (Y), όπως φαίνονται παρακάτω,

$$X: \ 5\ 3\ 6\ 3\ 2\ 3\ 2\ 1\ 4$$

$$Y: \ 7\ 2\ 5\ 1\ 1\ 6\ 3\ 4\ 8$$

Ελέγξτε αν υπάρχει θετική συσχέτιση όσον αφορά τις επιδόσεις των φοιτητών/τριών για αυτά τα δύο μαθήματα. Υπολογίστε την p-τιμή του παραπάνω ελέγχου.

Λύση Παραδείγματος 5.14. Ορίζουμε τις τ.μ. X, Y, οι οποίες μετράνε την επίδοση του/της φοιτητή/φοιτήτριας στο μάθημα Πιθανότητες Ι και Πιθανότητες ΙΙ, αντίστοιχα. Επειδή η τ.μ. Y έχει τις λιγότερες περιπτώσεις ισοβαθμιών όσον αφορά τις τιμές της (η τιμή 1 εμφανίζεται δύο φορές), διατάσσουμε κατά αύξουσα σειρά αυτές τις τιμές και αντιστοιχούμε σε αυτές τις τιμές της τ.μ. X, όπως φαίνεται παρακάτω:

$Y_{(i)}$	1	1	2	3	4	5	6	7	8
$X_i^{(*)}$	3	2	3	2	1	6	3	5	4

Εφαρμόζουμε τον έλεγχο των Cox και Stuart στο δείγμα της τ.μ. X, δηλαδή στο δείγμα των $X_1^{(*)}, X_2^{(*)}, \ldots, X_k^{(*)}$, με $k = 9$. Καθώς $k = 9$ περιττός, είναι $c = \frac{(9+1)}{2} = 5$ και, επομένως, σχηματίζουμε $k - c = 9 - 5 = 4$ το πλήθος ζεύγη της μορφής $(X_1^{(*)}, X_{1+c}^{(*)})$, $(X_2^{(*)}, X_{2+c}^{(*)})$, ..., $(X_{k-c}^{(*)}, X_k^{(*)})$. Άρα, σχηματίζουμε τα παρακάτω ζεύγη παρατηρήσεων:

Ζεύγος παρατηρήσεων	Τιμές	Αποτέλεσμα
$(X_1^{(*)}, X_6^{(*)})$	(3,6)	–
$(X_2^{(*)}, X_7^{(*)})$	(2,3)	–
$(X_3^{(*)}, X_8^{(*)})$	(3,5)	–
$(X_4^{(*)}, X_9^{(*)})$	(2,4)	–

Παρατηρήστε ότι η παρατήρηση $X_5^{(*)} = 1$ δεν συμμετέχει στη δημιουργία ζεύγους, καθώς το k είναι περιττός αριθμός και αντιστοιχεί στη μεσαία παρατήρηση.

Έπειτα θεωρούμε τις τ.μ.

$$Z_i = \begin{cases} 1 & , \quad \text{αν } X_i^{(*)} > X_{i+c}^{(*)} \text{ (ή +)} \\ 0 & , \quad \text{αν } X_i^{(*)} < X_{i+c}^{(*)} \text{ (ή −)} \end{cases} , \; i = 1, \dots, n.$$

Οι τ.μ. $Z_1, \dots, Z_n$, με $n = 4$, αποτελούν ένα τυχαίο δείγμα από την κατανομή $B(1, p)$, με πιθανότητα επιτυχίας $p = P(Z_i = 1) = P(X_i > X_{i+c}) = P(+)$. Θέλουμε να εξετάσουμε αν υπάρχει θετική συσχέτιση μεταξύ των τ.μ. X και Y, οπότε θεωρούμε το διωνυμικό πρόβλημα ελέγχου:

$$H_0 : p \geqslant 0.5 \; , \;\; H_1 : p < 0.5,$$

με ελεγχοσυνάρτηση

$$\phi(z) = \begin{cases} 1 & , \quad T \leqslant c, \\ 0 & , \quad \text{διαφορετικά,} \end{cases}$$

με τη σταθερά c να υπολογίζεται από τη σχέση, $P(T \leqslant c \,|p = 0.5) \approx a$. Υπό την H_0, η στατιστική συνάρτηση ελέγχου $T = \sum_{i=1}^{4} Z_i \sim B(4, 0.5)$, επομένως, χρησιμοποιώντας τον Πίνακα Π.7 του Παραρτήματος, παρατηρούμε ότι $P(T \leqslant 0) = 0.0625$. Δεν είναι δύσκολο να διαπιστώσουμε ότι για $c = 0$ προκύπτει η πλησιέστερη πιθανότητα στην επιθυμητή $a = 0.05$ για το ε.σ. Ειδικότερα, επιλέγοντας ως $c = 0$, έπεται ότι το μέγεθος του ελέγχου είναι $a = 0.0625$. Επιπλέον, καθώς η τιμή της T είναι $\tau = \#(+) = 0 \leqslant 0 = c$, απορρίπτουμε την H_0 σε ε.σ. 6.25%, δηλαδή σε αυτό το ε.σ. αποφασίζουμε ότι οι τ.μ. X και Y είναι θετικά συσχετισμένες.

Τέλος, η p-τιμή του ελέγχου είναι $P(T \geqslant \tau | p = 0.5) = P(T \geqslant 0 | p = 0.5) = 0.0625$ και είναι το μικρότερο ε.σ., για το οποίο απορρίπτεται η H_0. □

5.6 Ασκήσεις

Άσκηση 5.1. Οι κάτοικοι μιας περιοχής ανέφεραν στο δημοτικό συμβούλιο ότι τουλάχιστον το 60% των κατοίκων της περιοχής είναι υπέρ της δημιουργίας ενός ψυχαγωγικού κέντρου. Το δημοτικό συμβούλιο επέλεξε ένα τυχαίο δείγμα 100 κατοίκων. Από αυτούς 48 υποστήριξαν τη δημιουργία του κέντρου. Είναι βάσιμος ο ισχυρισμός που διατυπώθηκε στο δημοτικό συμβούλιο; Να ελέγξετε την παραπάνω υπόθεση σε ε.σ. 1%.

Άσκηση 5.2. Από τα 14828 νοικοκυριά μιας πόλης, επελέγη ένα τυχαίο δείγμα 290 νοικοκυριών. Κάθε οικογένεια ερωτήθηκε αν νοίκιαζε ή όχι το σπίτι στο οποίο διέμενε και αν είχε αυτόνομη θέρμανση ή όχι. Τα αποτελέσματα συνοψίζονται στον πίνακα που ακολουθεί.

	Ιδιοκτήτες	Ενοικιαστές
Αυτόνομη θέρμανση	6	34
Κεντρική θέρμανση	141	109

Μπορεί να υποστηριχθεί ο ισχυρισμός ότι τουλάχιστον 15% των ενοικιαζόμενων κατοικιών έχει αυτόνομη θέρμανση; Να ελέγξετε την παραπάνω υπόθεση σε ε.σ. 3%.

Άσκηση 5.3. Για μια περίοδο δέκα εργάσιμων ημερών το πλήθος των πελατών που εξυπηρετούνται σε μία τράπεζα είναι:

$$142 \;\; 134 \;\; 98 \;\; 119 \;\; 131 \;\; 103 \;\; 154 \;\; 122 \;\; 93 \;\; 137.$$

Ελέγξτε, σε ε.σ. 1%, την υπόθεση ότι ο μέσος αριθμός των πελατών που εξυπηρετούνται από τη συγκεκριμένη τράπεζα ξεπερνάει τους 120. Να διατυπώσετε τις υποθέσεις που πρέπει να ισχύουν για την εφαρμογή του συγκεκριμένου ελέγχου.

Άσκηση 5.4. Από τους αποφοίτους του Τμήματος Μαθηματικών που πρόκειται να ορκιστούν, επιλέγουμε τυχαία 12 από αυτούς και καταγράφουμε τον βαθμό του πτυχίου τους:

$$7.05\,,\;\; 6.27\,,\;\; 5.89\,,\;\; 6.44\,,\;\; 6.57\,,\;\; 8.21,\;\; 6.07\,,\;\; 6.12\,,\;\; 6.64\,,\;\; 7.18\,,\;\; 8.63\,,\;\; 6.22$$

Χρησιμοποιώντας έναν έλεγχο για ποσοστιαία σημεία, ελέγξτε αν λιγότερο από το 5% των αποφοίτων που πρόκειται να ορκιστούν έχει πτυχίο Άριστα ($\geqslant 8.5$), σε ε.σ. μικρότερο του 0.1.

Άσκηση 5.5. Τα τελευταία χρόνια έχει παρατηρηθεί το φαινόμενο, να περνάνε τη βάση στο μάθημα «Στατιστική» το πολύ 35% των φοιτητών/τριών που προσέρχονται στις εξετάσεις. Από το σύνολο των φοιτητών/τριών που προσήλθαν στην εξέταση του μαθήματος κατά την τελευταία εξεταστική περίοδο, επιλέξαμε, με τυχαίο τρόπο, ένα δείγμα 10 φοιτητών/τριών και καταγράψαμε τους βαθμούς τους, οι οποίοι είναι οι εξής:

$$3,\;\; 8,\;\; 7,\;\; 4,\;\; 4,\;\; 0,\;\; 2,\;\; 6,\;\; 2,\;\; 5$$

Εφαρμόζοντας έναν έλεγχο για ποσοστιαία σημεία ελέγξτε, σε ε.σ. 5%, κατά πόσο τα σημερινά αποτελέσματα συμπίπτουν με αυτά των παλαιότερων ετών.

Άσκηση 5.6. Οι χρόνοι ζωής, σε ώρες, 16 λυχνιών ραδιοφώνου που επιλέχθηκαν τυχαία από την παραγωγή ενός εργοστασίου είναι οι εξής:

$$46.9,\;\; 56.8,\;\; 63.3,\;\; 97.1,\;\; 47.2,\;\; 59.2,\;\; 63.4,\;\; 67.7,\;\; 49.1,\;\; 59.9,\;\; 63.7,\;\; 73.3,\;\; 56.5,\;\; 63.2,\;\; 64.1,\;\; 78.5$$

Ελέγξτε, σε ε.σ. 5%, αν τουλάχιστον το 5% των λυχνιών έχει χρόνο ζωής που ξεπερνά τις 70 ώρες. Υπολογίστε το μέγεθος του παραπάνω ελέγχου.

Άσκηση 5.7. Ένα τυχαίο δείγμα 6 οδηγών, που συνελήφθησαν να οδηγούν μεθυσμένοι, υποβλήθηκε σε δύο διαφορετικά test για τον καθορισμό του επιπέδου του οινοπνεύματος στο αίμα τους με τα εξής αποτελέσματα:

Οδηγός	1	2	3	4	5	6
test 1	0.170	0.190	0.200	0.183	0.187	0.178
test 2	0.197	0.178	0.150	0.176	0.205	0.153

Θα μπορούσε να ισχυριστεί κάποιος/κάποια ότι τα δύο test δεν δίνουν συνεπή αποτελέσματα και, επομένως, ένας οδηγός που θα καταδικαζόταν με βάση το ένα test, θα μπορούσε να αθωωθεί με βάση το άλλο; Να ελέγξετε την υπόθεση αυτή σε ε.σ. 5%.

Άσκηση 5.8. 100 άτομα ρωτήθηκαν αν στηρίζουν ή όχι την εξωτερική πολιτική μιας κυβέρνησης 2 μήνες και 12 μήνες μετά την ορκωμοσία της, με τα εξής αποτελέσματα.

	Μετά από 2 μήνες	
Μετά από 12 μήνες	Στήριζαν	Δεν στήριζαν
Στηρίζουν	20	40
Δεν στηρίζουν	10	30

Πιστεύετε ότι κατά τη διάρκεια αυτών των μηνών άλλαξε κάτι όσον αφορά την εμπιστοσύνη των πολιτών προς την εξωτερική πολιτική της κυβέρνησης; Να διεξάγετε κατάλληλο έλεγχο σε ε.σ. 1%.

Άσκηση 5.9. Σε 100 άτομα που υποφέρουν από τον πόνο χορηγήσαμε ένα αντίγραφο φαρμάκου (placebo). Στα 35 από αυτά τα άτομα παρατηρήθηκε κάποια ανακούφιση, ενώ στα υπόλοιπα 65 δεν παρατηρήθηκε καμία ανακούφιση. Έπειτα, χορηγήσαμε στα ίδια άτομα ένα καινούριο φάρμακο και 55 από αυτά ανέφεραν ότι υπήρξε μια ανακούφιση του πόνου, κάτι που δεν παρατηρήθηκε για τα υπόλοιπα 45. Από τα 65 άτομα στα οποία δεν παρουσιάστηκε καμιά ανακούφιση από το αντίγραφο, στα 30 δεν παρουσιάστηκε καμία ανακούφιση και από το καινούριο φάρμακο. Υπάρχει ένδειξη ότι το καινούριο φάρμακο είναι πιο αποτελεσματικό από το αντίγραφο; Να διεξάγετε κατάλληλο έλεγχο σε ε.σ. 2%. Να υπολογιστεί και να ερμηνευτεί η p-τιμή του ελέγχου.

Άσκηση 5.10. Στις αρχές του 2009 ρωτήσαμε 60 Έλληνες για το εισόδημά τους. Οι 20 από αυτούς δήλωσαν ότι το εισόδημά τους δεν ξεπερνάει τις 30.000 ευρώ τον χρόνο, ενώ οι υπόλοιποι 40 είχαν εισόδημα μεγαλύτερο από 30.000 ευρώ τον χρόνο. Στις αρχές του 2011 ρωτήσαμε τα ίδια άτομα για το εισόδημά τους. Προέκυψε ότι 10 από τα άτομα που δήλωναν εισοδήματα άνω των 30.000 ευρώ, πλέον δεν ξεπερνούν αυτό το ποσό, ενώ 5 από τα άτομα, που στις αρχές του 2009 δήλωναν εισόδημα κάτω των 30.000 ευρώ, τώρα το ξεπερνούν. Πιστεύετε ότι η οικονομική κρίση επηρέασε το εισόδημα των Ελλήνων; Να διεξάγετε κατάλληλο έλεγχο για την παραπάνω υπόθεση σε ε.σ. 5% και να υπολογίσετε την p-τιμή του παραπάνω ελέγχου.

Άσκηση 5.11. Σε 80 άτομα που πάσχουν από μια συγκεκριμένη ασθένεια, δόθηκε ένα φάρμακο Α και οι 50 από αυτούς παρουσίασαν βελτίωση. Μετά από έναν μήνα στους ίδιους ασθενείς δόθηκε ένα φάρμακο Β. Οι 45 από τους 50, που είχαν παρουσιάσει βελτίωση με το φάρμακο Α, παρουσίασαν βελτίωση και με το φάρμακο Β, ενώ παρουσίασαν βελτίωση με το φάρμακο Β και 10 άτομα από αυτά, που δεν είχαν παρουσιάσει βελτίωση με το φάρμακο Α. Πιστεύετε ότι με το φάρμακο Β οι ασθενείς παρουσίασαν μεγαλύτερη βελτίωση από ό,τι με το φάρμακο Α; Χρησιμοποιήστε ε.σ. $a = 0.05$.

Άσκηση 5.12. Μια επιστημονική ομάδα καταγράφει κάθε μήνα τον μέσο μηνιαίο ρυθμό ροής του νερού ενός μικρού ποταμού, σε κυβικά πόδια ανά δευτερόλεπτο. Στη διάθεσή μας έχουμε τα δεδομένα (μετρήσεις) των τελευταίων 24 μηνών, τα οποία με χρονολογική σειρά από την παλαιότερη προς την πιο πρόσφατη μέτρηση είναι:

$$14.6,\ 12.2,\ 104,\ 220,\ 110,\ 86.0,\ 92.8,\ 74.4,\ 75.4,\ 51.7,\ 29.3,\ 16.0,$$

$$14.2,\ 10.5,\ 123,\ 190,\ 138,\ 98.1,\ 88.1,\ 80.0,\ 75.6,\ 48.8,\ 27.1,\ 15.7$$

Να ελέγξετε την υπόθεση ότι ο ρυθμός ροής του νερού του ποταμού μειώνεται, σε ε.σ. $a \approx 0.1$. **Υπόδειξη:** Τα δεδομένα διαβάζονται κατά γραμμή.

Άσκηση 5.13. Οι τιμές κλεισίματος της μετοχής μιας τράπεζας κατά την περίοδο από 6/8/2013 έως 27/8/2013 είναι οι εξής:

6/8	7/8	8/8	9/8	12/8	13/8	14/8	16/8	19/8	20/8	21/8	22/8	23/8	26/8	27/8
2.51	2.55	2.73	2.79	3.23	3.10	3.03	3.15	2.96	2.84	2.88	2.95	3.05	3.00	2.85

Ελέγξτε αν υπάρχει κάποιου είδους τάση, είτε ανοδική είτε καθοδική, στις τιμές της μετοχής της τράπεζας. Υπολογίστε την p-τιμή του παραπάνω ελέγχου. Να χρησιμοποιηθεί ε.σ. 2%.

Άσκηση 5.14. Οι τιμές που συνοψίζονται στον παρακάτω πίνακα είναι οι μέσες θερμοκρασίες σε μια πόλη της Ελλάδας για 12 συνεχόμενες ημέρες.

Ημέρα	1	2	3	4	5	6	7	8	9	10	11	12
Θερμ. (^0C)	20.2	20.4	20.1	20.3	20.5	20.7	20.5	20.4	20.8	20.8	21.0	20.9

Με βάση τις τιμές αυτές, μπορεί να εξαχθεί το συμπέρασμα ότι η θερμοκρασία παρουσιάζει κάποιου είδους τάση, αυξητική ή πτωτική; Ποια είναι η ελάχιστη τιμή του επιπέδου σημαντικότητας για το οποίο απορρίπτεται η υπόθεση της μη ύπαρξης τάσης;

Άσκηση 5.15. Ρωτήσαμε 10 άτομα που πηγαίνουν γυμναστήριο πόσες ώρες αθλούνται τον μήνα και πόσα φρούτα καταναλώνουν κάθε μήνα. Οι απαντήσεις, που συλλέξαμε, παρουσιάζονται στον παρακάτω πίνακα:

Ώρες άθλησης	12	15	24	18	30	32	17	27	42	8
Κατανάλωση φρούτων	22	28	30	26	26	48	30	32	58	15

Πιστεύετε ότι υπάρχει κάποιου είδους σχέση ανάμεσα στις ώρες που γυμνάζεται ένα άτομο και στην κατανάλωση φρούτων από αυτό; Ποια είναι η p-τιμή του παραπάνω ελέγχου; Ποια η ερμηνεία της;

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Κούτρας, Μ. (2018). *Εισαγωγή στη Θεωρία Πιθανοτήτων και Εφαρμογές*. Αθήνα: Εκδόσεις Σταμούλη.

Ξενόγλωσση

Agresti, A. (2013). *Categorical Data Analysis* (3rd ed.). John Wiley & Sons Inc.

Arbuthnot, J. (1710). An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions of the Royal Society of London*, *27*(328), pp. 186–190.

Bernoulli, J. (1713). *Ars conjectandi: opus posthumum : accedit Tractatus de seriebus infinitis, et Epistola Gallicè scripta de ludo pilae reticularis*. Impensis Thurnisiorum, fratrum.

Conover, W. J. (1998). *Practical Nonparametric Statistics* (3rd ed.). Wiley and Sons, Inc.

Cox, D. and Stuart, A. (1955). Some quick sign tests for trend in location and dispersion. *Biometrika*, *42*(1-2), pp. 80–95.

Edwards, A. L. (1948). Note on the “correction for continuity” in testing the significance of the difference between correlated proportions. *Psychometrika*, *13*, pp. 185–187.

Gibbons, J. D. and Chakraborti, S. (2020). *Nonparametric Statistical Inference, Fourth Edition Revised and Expanded*. Chapman and Hall/CRC.

Hollander, M., Wolfe, D. and Chicken, E. (2014). *Nonparametric Statistical Methods* (3nd ed.). John Wiley and Sons.

Kavzoglu, T. (2017). Chapter 33 - Object-Oriented Random Forest for High Resolution Land Cover Mapping Using Quickbird-2 Imagery. In: *Handbook of Neural Computation*. Ed. by P. Samui, S. Sekhar and V. E. Balas. Academic Press, pp. 607–619.

Kvam, P. and Vidakovic, B. (2007). *Nonparametric Statistics with applications to science and engineering*. Wiley Series in Probability and Statistics.

Mattmüller, M. (2014). The difficult birth of stochastics: Jacob Bernoulli's Ars Conjectandi (1713). *Historia Mathematica*, *41*(3), pp. 277–290.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*, pp. 153–157.

Riffenburgh, R. H. (2006). Chapter 15 - Tests on Categorical Data. In: *Statistics in Medicine (Second Edition)*. Ed. by R. H. Riffenburgh. Second Edition. Burlington: Academic Press, pp. 241–279.

Rosner, B. (2015). *Fundamentals of Biostatistics* (8th ed.). Cengage Learning.

Sheskin, D. (2011). *Handbook of Parametric and Non-parametric Procedures* (5th ed.). Chapman and Hall/CRC.

Sprent, P. (1999). *Applied Nonparametric Statistical Methods*. Chapmann and Hall.

Stuart, A. (1956). The Efficiencies of Tests of Randomness Against Normal Regression. *Journal of the American Statistical Association*, *51*(274), pp. 285–287.

Taillard, E. D., Waelti, P. and Zuber, J. (2008). Few statistical tests for proportions comparison. *European Journal of Operational Research*, *185*(3), pp. 1336–1350.

ΚΕΦΑΛΑΙΟ 6

ΕΛΕΓΧΟΙ ΤΑΞΗΣ

Σύνοψη

Αν και απλοί στη χρήση τους, οι προσημικοί έλεγχοι δεν είναι ιδιαίτερα ισχυροί, αφού λαμβάνουν υπόψη τους μόνο το πρόσημο της διαφοράς μεταξύ των τιμών ενός δείγματος και της τιμής της υποτιθέμενης διαμέσου (όταν εφαρμόζονται ως έλεγχοι διαμέσου) ή το πρόσημο της διαφοράς μεταξύ των τιμών σε ζεύγη παρατηρήσεων. Για τον λόγο αυτόν, έχουν εμφανιστεί στη βιβλιογραφία έλεγχοι που βασίζονται στην τάξη των δεδομένων και όχι απλώς στις τιμές τους και υπό συγκεκριμένες συνθήκες (όπως είναι, για παράδειγμα, η συμμετρία του πληθυσμού) είναι ισχυρότεροι των προσημικών ελέγχων, αφού λαμβάνουν πληροφορία όχι μόνο από το πρόσημο της διαφοράς αλλά και από το μέγεθος της διαφοράς. Σκοπός αυτού του κεφαλαίου είναι η παρουσίαση τέτοιων μεθοδολογιών για τον έλεγχο της ισότητας α) της πληθυσμιακής διαμέσου με δοθείσα τιμή, β) δύο ή περισσότερων πληθυσμιακών διαμέσων (με ανεξάρτητα ή εξαρτημένα δείγματα) και γ) δύο ή περισσότερων πληθυσμιακών διακυμάνσεων.

Προαπαιτούμενη γνώση:
Βασικές γνώσεις Πιθανοτήτων και Στατιστικής.

Προσδοκώμενα μαθησιακά αποτελέσματα:
Ο/η φοιτητής/τρια ολοκληρώνοντας την ενότητα αυτή θα μπορεί να εφαρμόζει κάποιους από τους πλέον γνωστούς ελέγχους τάξης. Ειδικότερα, θα είναι σε θέση να μπορεί να ελέγχει αν ένα σύνολο δεδομένων προέρχεται από έναν πληθυσμό με δοθείσα διάμεσο, την ισότητα δύο ή περισσότερων πληθυσμιακών διαμέσων και τέλος την ισότητα δύο ή περρισότερων πληθυσμιακών διακυμάνσεων.

Γλωσσάριο επιστημονικών όρων

- Γραμμική Στατιστική Συνάρτηση Τάξης
- Δεσμός
- Έλεγχος μίας διαμέσου
- Έλεγχος πληθυσμιακών διακυμάνσεων
- Έλεγχος πληθυσμιακών διαμέσων
- Έλεγχος των τετραγώνων τάξεων μεγέθους (Conover’s squared ranks test)
- Έλεγχος Ansari-Bradley
- Έλεγχος Friedman
- Έλεγχος Kruskal-Wallis
- Έλεγχος Mann-Whitney
- Έλεγχος Siegel-Tukey
- Έλεγχος Wilcoxon
- Τάξη

Μπατσίδης, Α., Παπασταμούλης, Π., Πετρόπουλος, Κ., & Ρακιτζής, Α. (2022). *Μη Παραμετρική Στατιστική.* [Προπτυχιακό εγχειρίδιο].

6.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο, μεταξύ άλλων, παρουσιάστηκαν οι προσημικοί έλεγχοι ως τρόποι ελέγχου της ισότητας μιας πληθυσμιακής διαμέσου με μία δοθείσα τιμή ή της ισότητας δύο πληθυσμιακών διαμέσων με εξαρτημένα δείγματα. Παρότι οι έλεγχοι αυτοί είναι ιδιαίτερα απλοί, δεν είναι πάντοτε ιδιαίτερα ισχυροί. Η ιδιότητά τους αυτή αιτιολογείται πλήρως, καθώς οι προσημικοί έλεγχοι λαμβάνουν υπόψη τους μόνο το πρόσημο της διαφοράς μεταξύ των τιμών ενός δείγματος και της τιμής της υποτιθέμενης διαμέσου (όταν εφαρμόζονται ως έλεγχοι διαμέσου) ή το πρόσημο της διαφοράς μεταξύ των τιμών σε ζεύγη παρατηρήσεων.

Θέλοντας να αξιοποιηθούν όχι μόνο οι τιμές των δεδομένων αλλά και η διάταξη αυτών σε σχέση με τις υπόλοιπες τιμές, έχουν εισαχθεί στη βιβλιογραφία και οι λεγόμενοι έλεγχοι τάξης, δηλαδή οι έλεγχοι που βασίζονται στις τάξεις των δεδομένων. Ο ορισμός του όρου τάξη δίνεται στη συνέχεια.

Ορισμός 6.1

Τάξη (rank) μιας παρατήρησης ενός συνόλου δεδομένων είναι ο αριθμός που δίνει τη θέση που η παρατήρηση έχει στο διατεταγμένο κατά αύξουσα τάξη μεγέθους δείγμα. Αν δύο ή περισσότερες τιμές ταυτίζονται, οπότε και λέμε ότι έχουμε **δεσμούς ή ισοβαθμίες**, τότε ως τάξη καθεμίας θεωρείται ο μέσος όρος των τάξεων που αυτές θα είχαν αν ήταν διαφορετικές (midranks). Για μια παρατήρηση X_i η τάξη της συμβολίζεται με $R(X_i)$.

Ο παραπάνω ορισμός, ο οποίος μπορεί να εφαρμοστεί τόσο σε αριθμητικά όσο και σε μη αριθμητικά δεδομένα, αρκεί αυτά να είναι διατάξιμα, αποσαφηνίζεται στο παράδειγμα που ακολουθεί.

Παράδειγμα 6.1. Υπολογίστε τις τάξεις για κάθε παρατήρηση των ακόλουθων συνόλων δεδομένων: α) 126 142 156 228 245 246 370 419 433 454 478 503, β) 126 142 370 228 245 245 370 370 433 454 433 142, γ) «καθόλου» «λίγο» «πολύ» «λίγο» «μέτρια» «πολύ» «λίγο» «καθόλου» «πολύ» (θεωρήστε, για παράδειγμα, ότι αποτελούν τις απαντήσεις 7 φοιτητών/τριών στην ερώτηση *πόσο ικανοποιημένοι είστε από το πρόγραμμα πρακτικής άσκησης*).

Λύση Παραδείγματος 6.1. α) Σε αυτό το σύνολο δεδομένων οι παρατηρήσεις είναι ήδη διατεταγμένες κατά αύξουσα τάξη μεγέθους. Επιπρόσθετα, παρατηρούμε ότι δεν υπάρχουν δεσμοί (καθώς δεν υπάρχουν παρατηρήσεις με την ίδια τιμή). Έτσι προκύπτει ότι οι τάξεις τους είναι:

$$1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10,\ 11,\ 12,$$

αντίστοιχα.
β) Σε αυτό το σύνολο δεδομένων οι παρατηρήσεις δεν είναι διατεταγμένες κατά αύξουσα τάξη μεγέθους. Επομένως, αρχικά, τις διατάσσουμε κατά αύξουσα τάξη μεγέθους. Είναι τότε:

$$126,\ 142,\ 142,\ 228,\ 245,\ 245,\ 370,\ 370,\ 370,\ 433,\ 433,\ 454.$$

Παρατηρούμε ότι υπάρχουν δεσμοί και θα πρέπει να είμαστε ιδιαίτερα προσεκτικοί στον υπολογισμό των τάξεων. Έτσι, για παράδειγμα, είναι $R(142) = (2 + 3)/2 = 2.5$, καθώς οι τιμές 2 και 3 είναι οι τιμές των τάξεων που θα είχαν οι παρατηρήσεις αυτές αν ήταν διαφορετικές. Με παρόμοιο σκεπτικό είναι $R(245) = (5 + 6)/2 = 5.5$, $R(370) = (7 + 8 + 9)/3 = 8$ και $R(433) = (10 + 11)/2 = 10.5$. Επομένως, οι τάξεις των αρχικών παρατηρήσεων είναι αυτές που δίνονται στον πίνακα που ακολουθεί:

i	1	2	3	4	5	6	7	8	9	10	11	12
X_i	126	142	370	228	245	245	370	370	433	454	433	142
$R(X_i)$	1	2.5	8	4	5.5	5.5	8	8	10.5	12	10.5	2.5

γ) Σε αυτό το σύνολο δεδομένων οι παρατηρήσεις δεν είναι διατεταγμένες κατά αύξουσα τάξη μεγέθους. Επομένως, αρχικά, τις διατάσσουμε κατά αύξουσα τάξη μεγέθους. Είναι τότε: «καθόλου» «καθόλου» «λίγο» «λίγο» «λίγο» «μέτρια» «πολύ» «πολύ» «πολύ». Με παρόμοιο τρόπο, όπως πριν, προκύπτει ότι οι τάξεις αυτών είναι: 1.5, 1.5, 4, 4, 4, 6, 8, 8, 8, αντίστοιχα. Επομένως, οι τάξεις των αρχικών παρατηρήσεων είναι: 1.5, 4, 8, 4, 6, 8, 4, 1.5, 8, αντίστοιχα. Δείτε και τον παρακάτω πίνακα.

i	1	2	3	4	5	6	7	8	9
X_i	«καθόλου»	«λίγο»	«πολύ»	«λίγο»	«μέτρια»	«πολύ»	«λίγο»	«καθόλου»	«πολύ»
$R(X_i)$	1.5	4	8	4	6	8	4	1.5	8

□

Στη συνέχεια θα παραθέσουμε κάποιες ιδιότητες των τάξεων χρήσιμες σε όσα έπονται σε αυτό το κεφάλαιο. Σε όσα ακολουθούν $R_1,...,R_n$ είναι οι τάξεις των δειγματικών τιμών $X_1,\dots,X_n$ στις οποίες υποθέτουμε ότι δεν υπάρχουν δεσμοί, δηλαδή ότι $X_i \neq X_j$, για $i \neq j$, $i,j = 1,...,n$. Τότε, από τον Ορισμό 6.1, προκύπτει ότι:

$$R_i = R(X_i) = \text{αριθμός των } X_j \text{ με } X_j \leqslant X_i,\ i \neq j, i,j = 1,...,n,$$

ή ισοδύναμα

$$R_i = \sum_{j=1, j\neq i}^{n} I\left(X_j \leq X_i\right), i = 1,...,n,$$

όπου $I(\cdot)$ η συνήθης δείκτρια συνάρτηση.

Πρόταση 6.1. (Ιδιότητες τάξεων) Αν $R_1,...,R_n$ είναι οι τάξεις ενός τυχαίου δείγματος $X_1,\dots,X_n$ από μία συνεχή αθροιστική συνάρτηση κατανομής τότε, υποθέτοντας ότι δεν υπάρχουν δεσμοί, ισχύει ότι:

α) $P\left(R_i = j\right) = \frac{1}{n}, j = 1,...,n.$

β) $\mathrm{E}(R_i) = \frac{n+1}{2}$ και $\mathrm{Var}(R_i) = \frac{n^2-1}{12}$.

γ) $\mathrm{Cov}\left(R_i, R_j\right) = -\frac{n+1}{12}$, για $i \neq j$, $i,j = 1,...,n$.

Απόδειξη Πρότασης 6.1. α) Από τον τρόπο ορισμού τους και λαμβάνοντας υπόψη ότι δεν υπάρχουν δεσμοί στο δείγμα των $X_1,\dots,X_n$, προκύπτει άμεσα ότι οι δυνατές τιμές των τάξεων R_i ανήκουν στο σύνολο $\{1,2,...,n\}$ και, επιπλέον, $P\left(R_i = j\right) = \frac{1}{n}, j = 1,...,n$ και η απόδειξη ολοκληρώνεται.

β) Από τον ορισμό της μέσης τιμής μίας διακριτής τυχαίας μεταβλητής προκύπτει ότι:

$$\mathrm{E}(R_i) = \sum_{j=1}^{n} jP\left(R_i = j\right) = \sum_{j=1}^{n} j\frac{1}{n} = \frac{n(n+1)}{2}\frac{1}{n} = \frac{n+1}{2}.$$

Από τον ορισμό της διακύμανσης ισχύει ότι $\mathrm{Var}(R_i) = \mathrm{E}\left(R_i^2\right) - [\mathrm{E}(R_i)]^2$. Στη συνέχεια, θα υπολογιστεί η $\mathrm{E}\left(R_i^2\right)$. Είναι εξ ορισμού:

$$\mathrm{E}\left(R_i^2\right) = \sum_{j=1}^{n} j^2 P\left(R_i = j\right) = \sum_{j=1}^{n} j^2\frac{1}{n} = \frac{n(n+1)(2n+1)}{6}\frac{1}{n} = \frac{(n+1)(2n+1)}{6}.$$

Επομένως,

$$\mathrm{Var}(R_i) = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{2n^2+3n+1}{6} - \frac{n^2+2n+1}{4} = \frac{n^2-1}{12}.$$

γ) Από τον ορισμό της συνδιακύμανσης έχουμε ότι:

$$\mathrm{Cov}\left(R_i, R_j\right) = \sum_{k=1}^{n} \sum_{s=1, k\neq s}^{n} (k - \mathrm{E}(R_i))\left(s - \mathrm{E}\left(R_j\right)\right) P\left(R_i = k, R_j = s\right).$$

Λαμβάνοντας υπόψη το α) και από τον ορισμό της δεσμευμένης πιθανότητας

$$P\left(R_i = k, R_j = s\right) = P(R_i = k) P\left(R_j = s | R_i = k\right) = \frac{1}{n} \cdot \frac{1}{n-1}.$$

Επομένως,

$$\begin{aligned}
\mathrm{Cov}\left(R_i, R_j\right) &= \sum_{k=1}^{n} \sum_{s=1, k\neq s}^{n} \left(k - \frac{n+1}{2}\right)\left(s - \frac{n+1}{2}\right) \frac{1}{n(n-1)} \\
&= \sum_{k=1}^{n} \sum_{s=1}^{n} \left(k - \frac{n+1}{2}\right)\left(s - \frac{n+1}{2}\right) \frac{1}{n(n-1)} - \sum_{k=1}^{n} \left(k - \frac{n+1}{2}\right)^2 \frac{1}{n(n-1)} \\
&= \frac{1}{n(n-1)} \sum_{k=1}^{n} \left(k - \frac{n+1}{2}\right) \sum_{s=1}^{n} \left(s - \frac{n+1}{2}\right) - \frac{1}{n(n-1)} \sum_{k=1}^{n} \left(k - \frac{n+1}{2}\right)^2.
\end{aligned}$$

Όμως,

$$\sum_{k=1}^{n} \left(k - \frac{n+1}{2}\right) = \sum_{s=1}^{n} \left(s - \frac{n+1}{2}\right) = \frac{n(n+1)}{2} - \frac{n(n+1)}{2} = 0,$$

οπότε

$$\begin{aligned}
\mathrm{Cov}\left(R_i, R_j\right) &= -\frac{1}{n(n-1)} \sum_{k=1}^{n} \left(k - \frac{n+1}{2}\right)^2 \\
&= -\frac{1}{n(n-1)} \left\{ \sum_{k=1}^{n} k^2 - 2\frac{n+1}{2} \sum_{k=1}^{n} k + \sum_{k=1}^{n} \left(\frac{n+1}{2}\right)^2 \right\} \\
&= -\frac{1}{n(n-1)} \left\{ \frac{n(n+1)(2n+1)}{6} - \frac{(n+1)^2 n}{2} + \frac{(n+1)^2 n}{4} \right\} \\
&= -\frac{1}{n(n-1)} \left\{ \frac{n(n+1)(2n+1)}{6} - \frac{(n+1)^2 n}{4} \right\} \\
&= -\frac{1}{n(n-1)} \left\{ \frac{4n^3 + 6n^2 + 2n - 3n^3 - 6n^2 - 3n}{12} \right\} \\
&= -\frac{1}{n(n-1)} \frac{n^3 - n}{12} = -\frac{n+1}{12},
\end{aligned}$$

που αποδεικνύει το ζητούμενο. □

Μεγάλο μέρος της μελέτης αυτού του κεφαλαίου θα επικεντρωθεί στο λεγόμενο πρόβλημα δύο δειγμάτων (two-sample problem), δηλαδή σε ελέγχους υποθέσεων που προκύπτουν όταν έχουμε δύο ανεξάρτητα μεταξύ τους τυχαία δείγματα, μεγέθους n_1 και n_2, με $n_1 + n_2 = n$, από τους πληθυσμούς με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1, 2$, αντίστοιχα. Ειδικότερα, έστω $X_{11}, \ldots, X_{1,n_1}$, και $X_{21}, \ldots, X_{2,n_2}$, τα δύο αυτά ανεξάρτητα μεταξύ τους τυχαία δείγματα. Στο πλαίσιο αυτό, διατάσσουμε κατά αύξουσα τάξη μεγέθους τις n συνολικά παρατηρήσεις και συμβολίζουμε με Z_i, $i = 1, \ldots n$, την τυχαία μεταβλητή που παριστάνει αν η i-οστή διατεταγμένη παρατήρηση στο σύνολο των n παρατηρήσεων προέρχεται από το πρώτο δείγμα ή

όχι. Δηλαδή η τυχαία μεταβλητή Z_i λαμβάνει την τιμή 1, αν προέρχεται από το δείγμα του πρώτου πληθυσμού, και την τιμή 0, αν προέρχεται από το δείγμα του δεύτερου πληθυσμού, για $i = 1,...n$. Όπως θα δούμε στη συνέχεια, πολλές από τις στατιστικές συναρτήσεις που θα προταθούν για τους διάφορους ελέγχους που θα εξετάσουμε στο κεφάλαιο αυτό είναι **γραμμικές στατιστικές συναρτήσεις των τάξεων** (linear rank statistics) και ορίζονται από τη σχέση:

$$T_n(Z_1,...,Z_n) = \sum_{i=1}^{n} a_i Z_i, \tag{6.1}$$

όπου a_i είναι δοθείσες σταθερές, που καλούνται βάρη ή συντελεστές στάθμισης (weights). Στη συνέχεια, θα δοθούν κάποια γενικά αποτελέσματα (βλ., μεταξύ άλλων, Gibbons, 2014) που θα διευκολύνουν τη μελέτη ειδικών περιπτώσεων.

Πρόταση 6.2. (Ιδιότητες γραμμικής στατιστικής συνάρτησης τάξεων) Όταν η μηδενική υπόθεση

$$H_0 : F_1(x) = F_2(x) = F(x), \text{ για κάθε } x \in \mathbb{R},$$

είναι αληθής, έχουμε ότι:

$$\mathrm{E}(Z_i) = \frac{n_1}{n}, \mathrm{Var}(Z_i) = \frac{n_1 n_2}{n^2} \text{ και } \mathrm{Cov}(Z_i, Z_j) = -\frac{n_1 n_2}{n^2(n-1)}.$$

Απόδειξη Πρότασης 6.2. Η τυχαία μεταβλητή Z_i λαμβάνει την τιμή 1, αν η i-οστή διατεταγμένη παρατήρηση προέρχεται από το πρώτο δείγμα, και την τιμή 0, αν δεν προέρχεται από αυτό, για $i = 1,...n$. Επομένως, υπό την υπόθεση ότι οι δύο πληθυσμοί ταυτίζονται, είναι $P(Z_i = 1) = \frac{n_1}{n}$ και $P(Z_i = 0) = \frac{n_2}{n}$. Επιπλέον, εύκολα συμπεραίνουμε ότι η Z_i είναι ουσιαστικά μια τυχαία μεταβλητή που ακολουθεί την κατανομή Bernoulli με πιθανότητα επιτυχίας n_1/n. Επομένως, έχουμε ότι:

$$\mathrm{E}(Z_i) = \frac{n_1}{n} \text{ και } \mathrm{Var}(Z_i) = \frac{n_1 n_2}{n^2}.$$

Η συνδιακύμανση των Z_i, Z_j είναι $\mathrm{Cov}(Z_i, Z_j) = \mathrm{E}(Z_i Z_j) - \mathrm{E}(Z_i)\mathrm{E}(Z_j)$. Επομένως, αρκεί να υπολογιστεί η $\mathrm{E}(Z_i Z_j)$, η οποία, λαμβάνοντας υπόψη τις δυνατές τιμές του ζεύγους των (Z_i, Z_j), είναι:

$$\mathrm{E}(Z_i Z_j) = P(Z_i = 1, Z_j = 1) = \frac{\binom{n_1}{2}}{\binom{n}{2}} = \frac{n_1}{n} \cdot \frac{n_1 - 1}{n - 1}.$$

Το ζητούμενο αποτέλεσμα προκύπτει εύκολα συνδυάζοντας τα παραπάνω αποτελέσματα. □

Πρόταση 6.3. Όταν η μηδενική υπόθεση

$$H_0 : F_1(x) = F_2(x) = F(x), \text{ για κάθε } x \in \mathbb{R},$$

είναι αληθής, τότε για τη στατιστική συνάρτηση T_n που ορίστηκε στη σχέση (6.1) έχουμε ότι:

$$\mathrm{E}(T_n) = n_1 \sum_{i=1}^{n} \frac{a_i}{n} \text{ και } \mathrm{Var}(T_n) = \frac{n_1 n_2}{n^2(n-1)} \left\{ n \sum_{i=1}^{n} a_i^2 - \left(\sum_{i=1}^{n} a_i \right)^2 \right\}.$$

Απόδειξη Πρότασης 6.3. Από τον ορισμό της στατιστικής συνάρτησης T_n, τις ιδιότητες της μέσης τιμής και την προηγούμενη πρόταση έχουμε ότι:

$$\mathrm{E}(T_n) = \sum_{i=1}^{n} a_i \mathrm{E}(Z_i) = \sum_{i=1}^{n} a_i \frac{n_1}{n} = n_1 \sum_{i=1}^{n} \frac{a_i}{n}.$$

Επιπλέον, από τον ορισμό της στατιστικής συνάρτησης T_n, τις ιδιότητες της διακύμανσης και την προηγούμενη πρόταση έχουμε ότι:

$$\begin{aligned}
\mathrm{Var}(T_n) &= \sum_{i=1}^{n} a_i^2 \mathrm{Var}(Z_i) + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} a_i a_j \mathrm{Cov}(Z_i, Z_j) \\
&= \frac{n_1 n_2}{n^2} \sum_{i=1}^{n} a_i^2 - \frac{n_1 n_2}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} a_i a_j \\
&= \frac{n_1 n_2}{n^2(n-1)} \left(n \sum_{i=1}^{n} a_i^2 - \sum_{i=1}^{n} a_i^2 - \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} a_i a_j \right) \\
&= \frac{n_1 n_2}{n^2(n-1)} \left\{ n \sum_{i=1}^{n} a_i^2 - \left(\sum_{i=1}^{n} a_i \right)^2 \right\},
\end{aligned}$$

που αποδεικνύει το ζητούμενο. □

Πρόταση 6.4. Όταν η μηδενική υπόθεση

$$H_0 : F_1(x) = F_2(x) = F(x), \text{ για κάθε } x \in \mathbb{R},$$

είναι αληθής, τότε για τη στατιστική συνάρτηση T_n που ορίστηκε στη σχέση (6.1) έχουμε ότι:

$$Z = \frac{T_n - \mathrm{E}(T_n)}{\sqrt{\mathrm{Var}(T_n)}} \xrightarrow{d} \mathcal{N}(0,1)$$

όταν $n_1 + n_2 \to \infty$, με $\frac{n_1}{n_2}$ σταθερά.

Απόδειξη Πρότασης 6.4. Το αποτέλεσμα προκύπτει άμεσα με εφαρμογή του Κεντρικού Οριακού Θεωρήματος για ισόνομες (όχι ανεξάρτητες) τυχαίες μεταβλητές, λαμβάνοντας επιπρόσθετα υπόψη τα αποτελέσματατα της προηγούμενης πρότασης. □

Στη βιβλιογραφία έχει προταθεί πληθώρα ελέγχων οι οποίοι βασίζονται τις τάξεις των δεδομένων και τις ιδιότητες αυτών. Στο πλαίσιο αυτού του συγγράμματος, στη συνέχεια του παρόντος κεφαλαίου, παρουσιάζονται οι σημαντικότεροι και πιο ευρέως χρησιμοποιούμενοι από αυτούς τους ελέγχους.

6.2 Έλεγχοι ισότητας της πληθυσμιακής διαμέσου με δοθείσα τιμή

Έστω $X_1, \ldots, X_n$, ένα τυχαίο δείγμα από έναν πληθυσμό με συνεχή αθροιστική συνάρτηση κατανομής $F(\cdot)$, με διάμεσο m. Επιπλέον, υποθέτουμε ότι η F είναι συμμετρική γύρω από το σημείο m (βλ. Gibbons and Chakraborti, 2020). Ενδιαφερόμαστε να ελέγξουμε μία εκ των τριών ακόλουθων υποθέσεων:

(Α) $H_0 : m = m_0$, έναντι της $H_1 : m \neq m_0$.
(Β) $H_0 : m = m_0$, έναντι της $H_1 : m > m_0$.
(Γ) $H_0 : m = m_0$, έναντι της $H_1 : m < m_0$.

Στη βιβλιογραφία για τα παραπάνω προβλήματα ελέγχου υποθέσεων, εκτός από το προσημικό τεστ, που παρουσιάστηκε στο προηγούμενο κεφάλαιο, έχει παρουσιαστεί και ο λεγόμενος έλεγχος των προσημασμένων τάξεων μεγέθους (Signed Ranks Test) του Wilcoxon. Ο έλεγχος αυτός ονομάζεται έτσι καθώς πρωτοπαρουσιάστηκε από τον χημικό και στατιστικό Frank Wilcoxon (1892-1965) στην εργασία του

Wilcoxon (1945) και στηρίζεται στις τάξεις των διαφορών $D_i = X_i - m_0$, $i = 1, ..., n$, όπου έχουν αποκλειστεί από την περαιτέρω ανάλυση οι δειγματικές τιμές της τυχαίας μεταβλητής X που είναι ίσες με m_0. Ο έλεγχος αυτός παρουσιάζεται αναλυτικά στη συνέχεια.

Αρχικά, δημιουργούμε τις διαφορές $D_i = X_i - m_0$, $i = 1, ..., n$, απαλείφοντας τις μηδενικές διαφορές, οπότε μειώνεται το μέγεθος του δείγματος κατάλληλα. Σε όσα ακολουθούν συνεχίζουμε να συμβολίζουμε με n το μέγεθος του δείγματος που προκύπτει, αποκλείοντας από την περαιτέρω ανάλυση τις προαναφερθείσες τιμές. Επιπρόσθετα, υποθέτουμε ότι οι διαφορές αυτές είναι συμμετρικές γύρω από το μηδέν. Η υπόθεση ότι οι διαφορές προέρχονται από συμμετρικό πληθυσμό με κέντρο συμμετρίας το μηδέν έχει ως συνέπεια οι θετικές και αρνητικές διαφορές να είναι ισοπίθανες. Επίσης, για τις αρνητικές και θετικές διαφορές και για οποιαδήποτε τιμή $c > 0$ ισχύει ότι:

$$F_D(-c) = P(D \leqslant -c) = 1 - F_D(c) = P(D \geqslant c).$$

Στο επόμενο βήμα οι απόλυτες τιμές των διαφορών $|D_1|, \ldots, |D_n|$ διατάσσονται κατά αύξουσα τάξη μεγέθους και υπολογίζονται οι τάξεις τους, έστω $R(|D_i|)$, $i = 1, ..., n$. Στο σημείο αυτό, επισημαίνεται ότι υπό την υπόθεση ότι έχουμε συνεχή πληθυσμό θεωρητικά δεν πρέπει να ασχολούμαστε με την ύπαρξη ούτε μηδενικών διαφορών ούτε δεσμών μεταξύ των παρατηρήσεων.

Σε αυτό το πλαίσιο, έστω T^+ και T^- το άθροισμα των τάξεων που αντιστοιχούν στις θετικές και αρνητικές διαφορές, αντίστοιχα. Δηλαδή

$$T^+ = \sum_{i=1}^{n} I(D_i > 0)\, R(|D_i|), \tag{6.2}$$

και

$$T^- = \sum_{i=1}^{n} I(D_i < 0)\, R(|D_i|), \tag{6.3}$$

όπου $I(\cdot)$ η συνήθης δείκτρια συνάρτηση. Οι παραπάνω σχέσεις ισοδύναμα μπορούν να γραφούν (λόγω του ότι έχουν αποκλειστεί οι μηδενικές διαφορές) ως

$$T^+ = \sum_{i=1}^{n} Z_i R(|D_i|),$$

και

$$T^- = \sum_{i=1}^{n} (1 - Z_i) R(|D_i|),$$

όπου $Z_i = I(D_i > 0)$. Τότε, εύκολα προκύπτει ότι

$$T^+ + T^- = \sum_{i=1}^{n} R(|D_i|) = \frac{n(n+1)}{2}, \tag{6.4}$$

και

$$T^+ - T^- = 2\sum_{i=1}^{n} I(D_i > 0)\, R(|D_i|) - \sum_{i=1}^{n} R(|D_i|) = 2\sum_{i=1}^{n} I(D_i > 0)\, R(|D_i|) - \frac{n(n+1)}{2}.$$

Εναλλακτικά, αν R_i^* είναι οι τυχαίες μεταβλητές που ορίζονται ως:

$$R_i^* = \begin{cases} +R(|D_i|), & \text{αν } D_i > 0, \\ -R(|D_i|), & \text{αν } D_i < 0, \end{cases} \tag{6.5}$$

τότε

$$T^+ - T^- = 2T^+ - \frac{n(n+1)}{2} = \sum_{i=1}^{n} R_i^*. \tag{6.6}$$

Η τ.μ. R_i^* ονομάζεται **προσημασμένος βαθμός** ή **προσημασμένη τάξη μεγέθους**. Σε όσα ακολουθούν συμβολίζουμε με S το άθροισμα των προσημασμένων τάξεων, δηλαδή $S = \sum_{i=1}^{n} R_i^*$.

Υπό τη μηδενική υπόθεση $H_0 : m = m_0$, αναμένεται το άθροισμα των τάξεων που αντιστοιχούν στις θετικές διαφορές (δηλαδή το T^+) να είναι ίσο (ή περίπου ίσο) με το άθροισμα των τάξεων που αντιστοιχούν στις αρνητικές διαφορές (δηλαδή με το T^-). Επομένως, ένα υψηλό άθροισμα θετικών (αρνητικών) τάξεων σε σχέση με το αντίστοιχο των αρνητικών (αντίστοιχα, θετικών) τάξεων θα συνεπάγεται ότι είναι απίθανο να είναι η τιμή m_0 η πληθυσμιακή διάμεσος. Επιπλέον, καθώς το άθροισμα όλων των τάξεων είναι σταθερό και ίσο με $\frac{n(n+1)}{2}$, οι έλεγχοι που στηρίζονται είτε μόνο στη στατιστική συνάρτηση T^+ είτε μόνο στη στατιστική συνάρτηση T^- ή στη στατιστική συνάρτηση $S = T^+ - T^-$ είναι ισοδύναμοι.

Έτσι για τον έλεγχο της $H_0 : m = m_0$, έναντι της εναλλακτικής $H_1 : m > m_0$, η κρίσιμη περιοχή αντιστοιχεί σε εκείνες τις τιμές όπου οι θετικές διαφορές υπερτερούν των αρνητικών διαφορών σε μέγεθος ή/και αριθμό. Επομένως, θα απορρίπτεται η μηδενική υπόθεση για μεγάλες τιμές της T^+ ή μικρές τιμές της T^- ή μεγάλες τιμές της $S = T^+ - T^-$. Επίσης, για τον έλεγχο της $H_0 : m = m_0$, έναντι της εναλλακτικής $H_1 : m < m_0$, η κρίσιμη περιοχή αντιστοιχεί σε εκείνες τις τιμές όπου οι αρνητικές διαφορές υπερτερούν των θετικών διαφορών σε μέγεθος ή/και αριθμό. Επομένως, θα απορρίπτεται η μηδενική υπόθεση για μικρές τιμές της T^+ ή μεγάλες τιμές της T^- ή μικρές τιμές της $S = T^+ - T^-$. Τέλος, για τον έλεγχο της $H_0 : m = m_0$, έναντι της εναλλακτικής $H_1 : m \neq m_0$, η κρίσιμη περιοχή αντιστοιχεί σε εκείνες τις τιμές όπου είτε οι αρνητικές διαφορές υπερτερούν των θετικών διαφορών σε μέγεθος ή/και αριθμό είτε οι θετικές διαφορές υπερτερούν των αρνητικών διαφορών σε μέγεθος ή/και αριθμό. Επομένως, θα απορρίπτεται η μηδενική υπόθεση τόσο για μικρές όσο και για μεγάλες τιμές των T^+, T^- και $S = T^+ - T^-$. Τέλος, αν χρησιμοποιηθεί για τον έλεγχο της υπό μελέτη μηδενικής υπόθεσης η στατιστική συνάρτηση

$$T = \min\{T^+, T^-\} = \min\left\{T^+, \frac{n(n+1)}{2} - T^+\right\},$$

καθώς επιλέγεται η μικρότερη τιμή εκ των T^+ και T^-, η μηδενική υπόθεση απορρίπτεται για μικρές τιμές της στατιστικής συνάρτησης που είναι τέτοιες, ώστε να εξασφαλίζεται ότι έχουμε επίπεδο σημαντικότητας a.

Για τον προσδιορισμό των τιμών των στατιστικών συναρτήσεων που θεωρούνται μικρές (ή μεγάλες ανάλογα) και οδηγούν σε απόρριψη της μηδενικής υπόθεσης απαιτείται ο προσδιορισμός της κατανομής της στατιστικής συνάρτησης υπό τη μηδενική υπόθεση $H_0 : m = m_0$. Για την επίτευξη αυτού του σκοπού θα προσδιοριστεί η ακριβής κατανομή της στατιστικής συνάρτησης T^+ υπό τη μηδενική υπόθεση. Η κατανομή αυτή μπορεί να χρησιμοποιηθεί για μικρές τιμές του μεγέθους δείγματος n και να προσδιοριστούν κατά αυτόν τον τρόπο οι κρίσιμες τιμές του ελέγχου. Επιπλέον, θα δείξουμε ότι το αποτέλεσμα ισχύει και για τη στατιστική συνάρτηση T^- και, επομένως, μπορεί κάποιος να το χρησιμοποιήσει και για τη στατιστική συνάρτηση $T = \min\{T^+, T^-\}$. Όσον αφορά τη στατιστική συνάρτηση S η ακριβής κατανομή της προκύπτει άμεσα ως 1-1 μετασχηματισμός της T^+, ενώ θα προσδιοριστεί στη συνέχεια αυτής της ενότητας η ασυμπτωτική κατανομή της, υπό τη μηδενική υπόθεση.

Θεώρημα 6.1

Έστω $X_1, \ldots, X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με συνεχή αθροιστική συνάρτηση κατανομής $F(\cdot)$. Ενδιαφερόμαστε να ελέγξουμε τη μηδενική υπόθεση ότι η διάμεσος m της άγνωστης κατανομής είναι ίση με m_0, όπου m_0 είναι ένας γνωστός αριθμός. Αν $D_i = X_i - m_0$, $i = 1,\ldots,n$, είναι οι μη μηδενικές διαφορές, τότε, υπό τη μηδενική υπόθεση, η κατανομή της στατιστικής συνάρτησης T^+ που παριστάνει το άθροισμα των τάξεων που αντιστοιχούν στις θετικές διαφορές δίνεται από τη σχέση:

$$P(T^+ = k) = \frac{C_n(k)}{2^n}, \text{ για } k = 0,\ldots,\frac{n(n+1)}{2},$$

όπου $C_n(k)$ ο αριθμός των υποσυνόλων του συνόλου $\{1,2,\ldots,n\}$ των οποίων τα στοιχεία αθροίζουν στο k.

Απόδειξη Θεωρήματος 6.1. Για την απόδειξη του θεωρήματος παραπέμπουμε στο σύγγραμμα των Randles and Wolfe (1979). □

Ένα πρώτο εύλογο ερώτημα είναι αν κάτι αντίστοιχο ισχύει για την κατανομή της στατιστικής συνάρτησης T^-, το οποίο απαντάται στην επόμενη παρατήρηση.

Παρατήρηση 6.1. Είναι $T^+ = \sum_{i=1}^{n} Z_i R(|D_i|)$ και $T^- = \sum_{i=1}^{n} (1 - Z_i) R(|D_i|)$, όπου $Z_i = I(D_i > 0)$. Όμως, υπό τη μηδενική υπόθεση έχουμε ότι η κατανομή της τυχαίας μεταβλητής Z_i είναι Bernoulli με πιθανότητα επιτυχίας 0.5. Επιπρόσθετα, η τυχαία μεταβλητή $1 - Z_i$ ακολουθεί και αυτή υπό τη μηδενική υπόθεση κατανομή Bernoulli με πιθανότητα επιτυχίας 0.5. Επομένως, οι κατανομές των T^+ και T^- υπό τη μηδενική υπόθεση ταυτίζονται.

Από το Θεώρημα 6.1 προκύπτει ότι η ακριβής κατανομή υπό τη μηδενική υπόθεση της στατιστικής συνάρτησης T^+ εξαρτάται από το μέγεθος του δείγματος, αλλά όχι από την κατανομή των X_i. Επομένως, η διαδικασία ελέγχου που βασίζεται στη συγκεκριμένη στατιστική συνάρτηση είναι απαλλαγμένη παραμέτρων (distribution free). Επιπλέον, προκύπτει ότι δεν μπορεί στη γενική περίπτωση να δοθεί σε κλειστή μορφή, καθώς δεν υπάρχει κλειστή μορφή για το $C_n(k)$ παρά μόνο αναδρομική σχέση (βλ. για λεπτομέρειες, μεταξύ άλλων, Randles and Wolfe, 1979). Το παράδειγμα που ακολουθεί έχει διττό στόχο. Από τη μια δίνεται ο τρόπος σκέψης για τον προσδιορισμό της ακριβούς κατανομής, υπό τη μηδενική υπόθεση, των στατιστικών συναρτήσεων T^+ και T^-, όταν $n = 4$ (βλ. επίσης Conover, 1998) και από την άλλη επιβεβαιώνεται, για αυτήν την ειδική περίπτωση, ότι υπό τη μηδενική υπόθεση οι κατανομές των T^+ και T^- ταυτίζονται.

Παράδειγμα 6.2. Αν το μέγεθος του δείγματος είναι ίσο με 4, υποθέτοντας ότι δεν υπάρχουν δεσμοί, να βρεθεί, χωρίς χρήση του Θεωρήματος 6.1, η ακριβής κατανομή υπό τη μηδενική υπόθεση των στατιστικών συναρτήσεων T^+ και T^-. Επιβεβαιώστε το αποτέλεσμα στο οποίο καταλήξατε, χρησιμοποιώντας το Θεώρημα 6.1.

Λύση Παραδείγματος 6.2. Έστω X_1, X_2, X_3, X_4 ένα τυχαίο δείγμα από έναν πληθυσμό με συνεχή αθροιστική συνάρτηση κατανομής $F(\cdot)$. Δημιουργούμε τις διαφορές $D_i = X_i - m_0$, $i = 1, 2, 3, 4$, οι οποίες είναι μη μηδενικές. Υποθέτουμε ότι οι διαφορές αυτές προέρχονται από συμμετρικό περί το μηδέν πληθυσμό. Οι στατιστικές συναρτήσεις T^+ και T^- εκφράζουν το άθροισμα των τάξεων που αντιστοιχούν στις θετικές και αρνητικές διαφορές, αντίστοιχα. Δηλαδή είναι

$$T^+ = \sum_{i=1}^{4} I(D_i > 0) R(|D_i|),$$

και

$$T^- = \sum_{i=1}^{4} I(D_i < 0) R(|D_i|).$$

Αφού το $n = 4$, προκύπτουν οι ακόλουθες 16 το πλήθος περιπτώσεις ως προς το πρόσημο των διαφορών και τον τρόπο διάταξής τους:

- - - -, - + - -, - - + -, - - - +, - + + -, - + - +, - - + +, - + + +, + - - -, + + - -, + - + -, + - - +, + + + -,+ + - +, + - + +, + + + +.

Επομένως, το άθροισμα των τάξεων που αντιστοιχούν σε θετικές διαφορές είναι αντίστοιχα:

$$0,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 9,\ 1,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 10,$$

ενώ το άθροισμα των τάξεων που αντιστοιχούν σε αρνητικές διαφορές είναι αντίστοιχα:

$$10,\ 8,\ 7,\ 6,\ 5,\ 4,\ 3,\ 1,\ 9,\ 7,\ 6,\ 5,\ 4,\ 3,\ 2,\ 0.$$

Για παράδειγμα, για την τετράδα $--\,--$, αφού όλες οι διαφορές είναι αρνητικές, το T^+ ως άθροισμα των τάξεων θετικών διαφορών ισούται με μηδέν, ενώ για την τετράδα $-\,+\,-+$, αφού οι θετικές διαφορές αντιστοιχούν σε διαφορές με τάξεις 2 και 4, το T^+ ισούται με 6. Με ανάλογο τρόπο υπολογίστηκαν και οι υπόλοιπες τιμές των T^+ και T^- για κάθε πιθανή τετράδα αποτελεσμάτων.

Επομένως, η συνάρτηση πιθανότητας της κατανομής των T^+ και T^- υπό τη μηδενική υπόθεση είναι:

$$P(T^+ = t) = P(T^- = t) = \begin{cases} 1/16, & t = 0,1,2,8,9,10, \\ 2/16, & t = 3,4,5,6,7. \end{cases}$$

Στη συνέχεια, θα επιβεβαιώσουμε τα παραπάνω χρησιμοποιώντας το αποτέλεσμα του Θεωρήματος 6.1. Αρχικά, δεν είναι δύσκολο να διαπιστώσουμε πως υπάρχουν $2^n = 2^4 = 16$ διαφορετικά υποσύνολα του συνόλου $\{1,2,3,4\}$ και αυτά είναι τα ακόλουθα:

$$\varnothing, \{1\}, \{2\}, \{3\}, \{4\}, \{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\},$$
$$\{1,2,3\}, \{1,2,4\}, \{1,3,4\}, \{2,3,4\}, \{1,2,3,4\}.$$

Παρατηρήστε ότι τα στοιχεία των υποσυνόλων

$$\varnothing, \{1\}, \{2\}, \{1,3,4\}, \{2,3,4\}, \{1,2,3,4\},$$

αθροίζουν αντίστοιχα στους αριθμούς 0, 1, 2, 8, 9, 10. Από την άλλη πλευρά, τα στοιχεία των $\{3\}$ και $\{1,2\}$ αθροίζουν στο 3, τα στοιχεία των $\{4\}$ και $\{1,3\}$ στο 4, τα στοιχεία των $\{1,4\}$ και $\{2,3\}$ στο 5, τα στοιχεία των $\{2,4\}$ και $\{1,2,3\}$ στο 6 και, τέλος, τα στοιχεία των $\{3,4\}$ και $\{1,2,4\}$ στο 7. Επομένως, είναι:

$$P(T^+ = t) = \begin{cases} 1/16, & t = 0,1,2,8,9,10, \\ 2/16, & t = 3,4,5,6,7. \end{cases}$$

□

Από τη συζήτηση που προηγήθηκε του Θεωρήματος 6.1 και λαμβάνοντας υπόψη την Παρατήρηση 6.1 προκύπτει ότι χρησιμοποιώντας τις στατιστικές συναρτήσεις T^+ και T^- για τη διεξαγωγή των μονόπλευρων ελέγχων σε επίπεδο σημαντικότητας a, θα πρέπει να προσδιοριστούν τα άνω και κάτω a ποσοστιαία σημεία (ανάλογα με το αν έχουμε άνω μονόπλευρο ή κάτω μονόπλευρο έλεγχο), ενώ για τον δίπλευρο έλεγχο απαιτείται και το άνω και το κάτω $a/2$ ποσοστιαίο σημείο. Όμως, όταν χρησιμοποιείται η στατιστική συνάρτηση $T = \min\{T^+, T^-\}$ για τους μεν μονόπλευρους ελέγχους χρειαζόμαστε το κάτω a ποσοστιαίο σημείο, ενώ για τον δίπλευρο έλεγχο το κάτω $a/2$ ποσοστιαίο σημείο της κατανομής της T. Στην πρόταση που ακολουθεί αποδεικνύεται (βλ., μεταξύ άλλων, Randles and Wolfe, 1979) ότι η κατανομή, υπό τη μηδενική υπόθεση, της στατιστικής συνάρτησης T^+, άρα και της στατιστικής συνάρτησης T^-, είναι συμμετρική με κέντρο συμμετρίας το σημείο $\frac{n(n+1)}{4}$. Η ιδιότητα αυτή έχει ως αποτελέσμα να απαιτείται μόνο ο προσδιορισμός των άνω (ή των κάτω) ποσοστιαίων σημείων. Στον Πίνακα Π.18 του Παραρτήματος δίνονται ποσοστιαία σημεία της στατιστικής συνάρτησης T για κάποιες συνήθεις τιμές του επιπέδου σημαντικότητας.

Πρόταση 6.5. Έστω $X_1, \ldots, X_n$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό με συνεχή αθροιστική συνάρτηση κατανομής $F(\cdot)$. Ενδιαφερόμαστε να ελέγξουμε την υπόθεση ότι η διάμεσος m της άγνωστης κατανομής είναι ίση με m_0, όπου m_0 είναι ένας γνωστός αριθμός. Αν $D_i = X_i - m_0$, $i = 1,\ldots,n$, είναι οι μη μηδενικές διαφορές, τότε, υπό τη μηδενική υπόθεση, η κατανομή της στατιστικής συνάρτησης T^+ που παριστάνει το άθροισμα των τάξεων που αντιστοιχούν στις θετικές διαφορές, είναι συμμετρική με κέντρο συμμετρίας την τιμή $\frac{n(n+1)}{4}$.

Απόδειξη Πρότασης 6.5. Από την Παρατήρηση 6.1 έχουμε ότι οι κατανομές των T^+ και T^- ταυτίζονται. Επιπλέον, λαμβάνοντας υπόψη τη σχέση (6.4), έχουμε ότι $T^- = \frac{n(n+1)}{2} - T^+$. Επομένως, ταυτίζονται οι

κατανομές των T^+ και $\frac{n(n+1)}{2} - T^+$ ή, ισοδύναμα, των $T^+ - \frac{n(n+1)}{4}$ και $\frac{n(n+1)}{4} - T^+$, που αποδεικνύει το ζητούμενο.

Εναλλακτικά, μπορεί κάποιος να αποδείξει ότι:

$$P\left(T^+ \leqslant \frac{n(n+1)}{4} - x\right) = P\left(T^+ \geqslant \frac{n(n+1)}{4} + x\right), \text{ για κάθε } x \in \mathbb{R}.$$

Πράγματι,

$$\begin{aligned} P\left(T^+ \leqslant \frac{n(n+1)}{4} - x\right) &= P\left(\frac{n(n+1)}{2} - T^- \leqslant \frac{n(n+1)}{4} - x\right) \\ &= P\left(-T^- \leqslant \frac{n(n+1)}{4} - x - \frac{n(n+1)}{2}\right) \\ &= P\left(-T^- \leqslant -x - \frac{n(n+1)}{4}\right) \\ &= P\left(T^- \geqslant x + \frac{n(n+1)}{4}\right) \\ &= P\left(T^+ \geqslant x + \frac{n(n+1)}{4}\right), \end{aligned}$$

όπου χρησιμοποιήθηκε ότι $T^- = \frac{n(n+1)}{2} - T^+$ και ότι οι κατανομές των T^+ και T^- ταυτίζονται. □

Πρόταση 6.6. Έστω $T_{n,p}$ είναι το p-ποσοστιαίο σημείο της κατανομής της T^+, τότε $T_{n,1-p} = \frac{n(n+1)}{2} - T_{n,p}$.

Απόδειξη Πρότασης 6.6. Από τον ορισμό του ποσοστιαίου σημείου για μια διακριτή κατανομή πιθανότητας, το $T_{n,p}$ είναι τέτοιο, ώστε:

$$P\left(T^+ > T_{n,p}\right) \leqslant p \leqslant P\left(T^+ \geqslant T_{n,p}\right). \tag{6.7}$$

Από τη μία πλευρά της διπλής ανισότητας της σχέσης (6.7), χρησιμοποιώντας ότι $T^- = \frac{n(n+1)}{2} - T^+$, με πιθανότητα 1 και ότι οι κατανομές των T^+ και T^- ταυτίζονται, έχουμε ότι:

$$P\left(T^+ > T_{n,p}\right) \leqslant p \Leftrightarrow P\left(\frac{n(n+1)}{2} - T^+ > T_{n,p}\right) \leqslant p,$$

από όπου προκύπτει ότι:

$$P\left(T^+ < \frac{n(n+1)}{2} - T_{n,p}\right) \leqslant p,$$

ή, ισοδύναμα, ότι:

$$P\left(T^+ \geqslant \frac{n(n+1)}{2} - T_{n,p}\right) \geqslant 1 - p. \tag{6.8}$$

Ομοίως, από την άλλη πλευρά της διπλής ανισότητας της σχέσης (6.7), χρησιμοποιώντας ότι $T^- = \frac{n(n+1)}{2} - T^+$, με πιθανότητα 1 και ότι οι κατανομές των T^+ και T^- ταυτίζονται, έχουμε ότι:

$$P\left(T^+ \geqslant T_{n,p}\right) \geqslant p \Leftrightarrow P\left(\frac{n(n+1)}{2} - T^+ \geqslant T_{n,p}\right) \geqslant p$$

από όπου προκύπτει ότι:

$$P\left(T^+ \leqslant \frac{n(n+1)}{2} - T_{n,p}\right) \geqslant p$$

ή, ισοδύναμα, ότι:

$$P\left(T^+ > \frac{n(n+1)}{2} - T_{n,p}\right) \leqslant 1 - p. \tag{6.9}$$

Από τις σχέσεις (6.8), (6.9) και τον ορισμό του $(1-p)$-ποσοστιαίου σημείου της κατανομής της T^+, είναι προφανές ότι,

$$T_{n,1-p} = \frac{n(n+1)}{2} - T_{n,p},$$

και η απόδειξη ολοκληρώθηκε. □

Χρησιμοποιώντας το Θεώρημα 6.1 και την Πρόταση 6.5 για μικρές τιμές του μεγέθους δείγματος n, προκύπτει, σε επίπεδο σημαντικότητας a, ότι για το πρόβλημα (Α) $H_0 : m = m_0, H_1 : m \neq m_0$ απορρίπτεται η μηδενική υπόθεση, αν $T^+ \leqslant T_{n,1-a/2}$ ή $T^+ > T_{n,a/2}$, για το (Β) $H_0 : m = m_0, H_1 : m > m_0$ απορρίπτεται η μηδενική υπόθεση, αν $T^+ > T_{n,a}$, ενώ για το (Γ) $H_0 : m = m_0, H_1 : m < m_0$ απορρίπτεται η μηδενική υπόθεση, αν $T^+ \leqslant T_{n,1-a}$, όπου $T_{n,p}$ είναι το p-ποσοστιαίο σημείο της κατανομής της T^+ (βλ. σχέση (6.7)). Τα σημεία $T_{n,p}$ για κάποιες συνήθεις τιμές του επιπέδου σημαντικότητας, δίνονται στον Πίνακα Π.17 του Παραρτήματος.

Ο τρόπος υπολογισμού της στατιστικής συνάρτησης και ο προτεινόμενος έλεγχος, χρησιμοποιώντας την ακριβή κατανομή υπό τη μηδενική υπόθεση, διευκρινίζονται στο παράδειγμα που ακολουθεί.

Παράδειγμα 6.3. (Sprent, 1999) Στον πίνακα που ακολουθεί καταγράφεται ο αριθμός των σελίδων 12 τυχαία επιλεγμένων βιβλίων από μία βιβλιοθήκη ενός μαθηματικού τμήματος:

126 142 156 228 245 246 370 419 433 454 478 503.

Χρησιμοποιώντας το τεστ του Wilcoxon και αφού κάνετε τις κατάλληλες υποθέσεις, να ελέγξετε, σε επίπεδο σημαντικότητας 5%, την υπόθεση ότι η πληθυσμιακή διάμεσος του αριθμού των σελίδων των βιβλίων της βιβλιοθήκης είναι ίση με 220 σελίδες.

Λύση Παραδείγματος 6.3. Έχουμε ένα τυχαίο δείγμα $X_1, \ldots, X_{12}$, από έναν πληθυσμό με συνεχή αθροιστική συνάρτηση κατανομής F. Ενδιαφερόμαστε να ελέγξουμε την υπόθεση ότι η διάμεσος m της άγνωστης κατανομής είναι ίση με $m_0 = 220$. Επομένως, θέλουμε να ελέγξουμε τη μηδενική υπόθεση $H_0 : m = 220$ έναντι της εναλλακτικής $H_1 : m \neq 220$.

Αρχικά, δημιουργούμε τις διαφορές $D_i = X_i - m_0$, $i = 1,\ldots,12$, οι οποίες είναι οι:

-94, -78, -64, 8, 25, 26, 150, 199, 213, 234, 258, 283.

Έπειτα οι απόλυτες τιμές των διαφορών $|D_1|, \ldots, |D_{12}|$, διατάσσονται κατά αύξουσα τάξη μεγέθους και υπολογίζονται οι τάξεις τους, έστω $R(|D_i|)$, $i = 1,\ldots,12$, όπως φαίνεται στον πίνακα που ακολουθεί. Για ευκολία στους περαιτέρω υπολογισμούς έχουμε υπογραμμίσει τις αρνητικές διαφορές:

$\lvert D_i\rvert$	8	25	26	<u>64</u>	<u>78</u>	<u>94</u>	150	199	213	234	258	383
$R(\lvert D_i\rvert)$	1	2	3	4	5	6	7	8	9	10	11	12

Καθώς οι αρνητικές διαφορές είναι λιγότερες σε πλήθος, υπολογίζουμε το άθροισμα των τάξεων που αντιστοιχούν σε αυτές. Είναι $T^- = 4 + 5 + 6 = 15$, ενώ το άθροισμα των τάξεων που αντιστοιχούν στις θετικές διαφορές, έστω T^+, είναι $T^+ = \frac{n(n+1)}{2} - T^- = \frac{12\cdot 13}{2} - 15 = 63$. Για τον έλεγχο της υπό μελέτη μηδενικής υπόθεσης, χρησιμοποιείται η στατιστική συνάρτηση $T = \min\{T^+, T^-\} = T^-$. Για ε.σ. 5%, από τον Πίνακα Π.18 του Παραρτήματος έχουμε ότι $T_{n,a/2} = T_{12,0.025} = 13$. Από τη συζήτηση που προηγήθηκε

της Πρότασης 6.5 έχουμε ότι απορρίπτεται η μηδενική υπόθεση, αν $T \leqslant T_{12,0.025} = 13$. Καθώς $T \equiv T^- = 15 \not\leqslant 13 = T_{12,0.025}$ έπεται ότι δεν απορρίπτεται η μηδενική υπόθεση $H_0 : m = 220$ έναντι της $H_1 : m \neq 220$ σε επίπεδο σημαντικότητας 5%.

Εναλλακτικά, θα μπορούσαμε να χρησιμοποιήσουμε τη στατιστική συνάρτηση T^+, με κρίσιμη περιοχή, σε επίπεδο σημαντικότητας 5%, $T^+ \leqslant T_{12,1-0.05/2}$ ή $T^+ > T_{12,0.05/2}$, όπου $T_{12,p}$ είναι το p-ποσοστιαίο σημείο της κατανομής της T^+ και προσδιορίζεται στον Πίνακα Π.17 του Παραρτήματος. Ειδικότερα, είναι $T_{12,1-0.05/2} = 14$ και $T_{12,0.05/2} = \frac{12(12+1)}{2} - T_{12,0.975} = 78 - 14 = 64$. Καθώς $T^+ = 63 \not\leqslant 14$ και $T^+ = 63 \not> 64$, συμπεραίνουμε ότι δεν απορρίπτεται η μηδενική υπόθεση $H_0 : m = 220$ έναντι της $H_1 : m \neq 220$ σε επίπεδο σημαντικότητας 5%.

Δηλαδή, υποθέτοντας ότι τα δεδομένα προέρχονται από συμμετρικό πληθυσμό, προκύπτει ότι ο μέσος αριθμός των σελίδων των βιβλίων της βιβλιοθήκης του μαθηματικού τμήματος δεν είναι στατιστικά σημαντικά διαφορετικός από 220. □

Πίνακες κρίσιμων τιμών για τον έλεγχο του Wilcoxon έχουν δοθεί στη βιβλιογραφία για $n \leqslant 50$. Όταν το μέγεθος του δείγματος είναι μεγάλο (άλλοι συγγραφείς αναφέρουν ως κριτήριο το $n \geqslant 20$ και άλλοι το $n \geqslant 30$) οδηγούμαστε, υπό τη μηδενική υπόθεση, στην εύρεση της προσεγγιστικής κατανομής των στατιστικών συναρτήσεων που παρουσιάστηκαν πρωτύτερα. Προτού προχωρήσουμε στον προσδιορισμό της προσεγγιστικής κατανομής των T^+ και S θα διατυπώσουμε και θα αποδείξουμε ένα πολύ χρήσιμο αποτέλεσμα.

Λήμμα 6.1. Έστω D είναι μια συνεχής τυχαία μεταβλητή από μια κατανομή με κέντρο συμμετρίας το μηδέν. Τότε οι τυχαίες μεταβλητές $Y = |D|$ και $Z = I(D > 0)$, όπου $I(\cdot)$ η συνήθης δείκτρια συνάρτηση, είναι ανεξάρτητες.

Απόδειξη Λήμματος 6.1. Αρκεί να δείξουμε ότι $f_{Y|Z}(y|z) = f_Y(y)$ για τις δυνατές τιμές της διακριτής τυχαίας μεταβλητής Z. Προφανώς, οι δυνατές τιμές της τυχαίας μεταβλητής Z είναι $0,1$. Άρα, δοθέντος ότι $Z = 1$, για $y > 0$ και συμβολίζοντας με dy την οριακή μεταβολή έχουμε ότι:

$$\begin{aligned} P\left(y \leqslant Y \leqslant y + dy | Z = 1\right) &= P\left(y \leqslant Y \leqslant y + dy | D > 0\right) \\ &= \frac{P\left(y \leqslant |D| \leqslant y + dy, D > 0\right)}{P(D > 0)} \\ &= 2P\left(y \leqslant D \leqslant y + dy\right), \end{aligned}$$

όπου χρησιμοποιήθηκε ότι $P(D > 0) = 0.5$, καθώς η D είναι μια συνεχής τυχαία μεταβλητή από μια κατανομή με κέντρο συμμετρίας το μηδέν. Επιπρόσθετα, δοθέντος ότι $Z = 0$, για $y > 0$ είναι:

$$\begin{aligned} P\left(y \leqslant Y \leqslant y + dy | Z = 0\right) &= P\left(y \leqslant Y \leqslant y + dy | D < 0\right) \\ &= \frac{P\left(y \leqslant |D| \leqslant y + dy, D < 0\right)}{P(D < 0)} \\ &= \frac{P\left(y \leqslant -D \leqslant y + dy, D < 0\right)}{P(D < 0)} \\ &= 2P\left(y \leqslant -D \leqslant y + dy\right) \\ &= 2P\left(-(y + dy) \leqslant D \leqslant -y\right) \\ &= 2P\left(y \leqslant D \leqslant y + dy\right), \end{aligned}$$

όπου στην τελευταία σχέση χρησιμοποιήθηκε πάλι ότι η κατανομή της τυχαίας μεταβλητής D είναι συμμετρική γύρω από το μηδέν. Επομένως,

$$f_{Y|Z}(y|z) = 2P\left(y \leqslant D \leqslant y + dy\right), \forall\, y > 0.$$

Επίσης, χρησιμοποιώντας το Θεώρημα Ολικής Πιθανότητας, έπεται ότι:

$$\begin{aligned} f_Y(y) &= P\Big(y \leqslant D \leqslant y+dy\Big) \\ &= P\Big(y \leqslant D \leqslant y+dy\Big|D>0)P(D>0) + P\Big(y \leqslant D \leqslant y+dy\Big|D<0)P(D<0). \end{aligned}$$

Άρα, από τους παραπάνω υπολογισμούς έχουμε,

$$f_Y(y) = 2P\Big(y \leqslant D \leqslant y+dy\Big)\frac{1}{2} + 2P\Big(y \leqslant D \leqslant y+dy\Big)\frac{1}{2} = f_{Y|Z}(y|z)$$

και η απόδειξη ολοκληρώθηκε. □

Από το παραπάνω λήμμα έχουμε ότι, υπό τη μηδενική υπόθεση και την πρόσθετη υπόθεση ότι οι διαφορές προέρχονται από συμμετρική κατανομή, οι $2n$ το πλήθος τυχαίες μεταβλητές

$$I(D_1>0), R(|D_1|), \ldots, I(D_n>0), R(|D_n|)$$

είναι αμοιβαία ανεξάρτητες (mutually independent). Η απόδειξη αυτού του αποτελέσματος αφήνεται ως άσκηση στον/στην αναγνώστη/στρια.

Η παρακάτω πρόταση μας δίνει την ασυμπτωτική κατανομή των στατιστικών συναρτήσεων T^+ και S.

Πρόταση 6.7. Έστω $X_1, \ldots, X_n$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό με συνεχή αθροιστική συνάρτηση κατανομής $F(\cdot)$. Ενδιαφερόμαστε να ελέγξουμε την υπόθεση ότι η διάμεσος m της άγνωστης κατανομής είναι ίση με m_0, όπου m_0 είναι ένας γνωστός αριθμός. Υπό τη μηδενική υπόθεση $H_0 : m = m_0$ και την πρόσθετη υπόθεση ότι δεν υπάρχουν δεσμοί, για μεγάλο μέγεθος δείγματος, ισχύει ότι:

$$W = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{d} \mathcal{N}(0,1),$$

όπου $T^+ = \sum_{i=1}^n I(D_i>0)\,R(|D_i|)$, με $D_i = X_i - m_0$, $i = 1, \ldots, n$, και $R(|D_i|)$, $i = 1, \ldots, n$, είναι οι τάξεις των απόλυτων τιμών των θετικών διαφορών. Δηλαδή T^+ είναι το άθροισμα των τάξεων που αντιστοιχούν στις θετικές διαφορές. Επιπρόσθετα, υπό τη μηδενική υπόθεση,

$$Z = \frac{\sum_{i=1}^n R_i}{\sqrt{\frac{n(n+1)(2n+1)}{6}}} \xrightarrow{d} \mathcal{N}(0,1),$$

όπου οι τυχαίες μεταβλητές R_i είναι οι προσημασμένες τάξεις μεγέθους που προσδιορίζονται από τη σχέση (6.5) και για το άθροισμα των οποίων ισχύει η σχέση (6.6).

Απόδειξη Πρότασης 6.7. Είναι $T^+ = \sum_{i=1}^n I(D_i>0)\,R(|D_i|)$ ή ισοδύναμα, καθώς δεν υπάρχουν δεσμοί, $T^+ = \sum_{i=1}^n I(D_i>0)\,u_i$, με u_i να είναι αριθμοί από $1, \ldots, n$ και να αντιστοιχούν στις τάξεις των $R(|D_i|)$, για $i = 1, \ldots, n$. Επομένως, η στατιστική συνάρτηση T^+ γράφεται ως άθροισμα n τυχαίων μεταβλητών και με χρήση του Κεντρικού Οριακού Θεωρήματος προκύπτει, υπό τη μηδενική υπόθεση, ότι:

$$\frac{T^+ - \mathsf{E}(T^+)}{\sqrt{\mathsf{Var}(T^+)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Επίσης, η $\mathsf{E}(T^+) = \sum_{i=1}^n \mathsf{E}\left(I(D_i>0)\,u_i\right)$, όμως λόγω του ότι υπό τη μηδενική υπόθεση η τυχαία μεταβλητή $I(D_i>0)$ ακολουθεί Bernoulli κατανομή με $p = 1/2$, προκύπτει άμεσα ότι:

$$\mathsf{E}\left(I(D_i>0)\,R(|D_i|)\right) = u_i \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{u_i}{2},$$

και, επομένως,

$$\mathrm{E}(T^+) = \sum_{i=1}^{n} u_i/2. \tag{6.10}$$

Όμως, καθώς έχουμε υποθέσει ότι δεν υπάρχουν δεσμοί[1] η $\mathrm{E}(T^+)$ ισούται με:

$$\mathrm{E}(T^+) = \sum_{i=1}^{n} i/2 = \frac{n(n+1)}{4}. \tag{6.11}$$

Επιπλέον, είναι

$$\mathrm{Var}\left(I\left(D_i > 0\right) R\left(|D_i|\right)\right) = \left(u_i^2 \cdot \frac{1}{2} + 0^2 \cdot \frac{1}{2}\right) - \left(u_i \cdot \frac{1}{2} + 0 \cdot \frac{1}{2}\right)^2 = \frac{u_i^2}{4},$$

και, επομένως,

$$\mathrm{Var}(T^+) = \sum_{i=1}^{n} \frac{u_i^2}{4}. \tag{6.12}$$

Όμως, καθώς έχουμε υποθέσει ότι δεν υπάρχουν δεσμοί, το $\sum_{i=1}^{n} u_i^2/4$ δεν είναι παρά το άθροισμα $1^2 + 2^2 + ... + n^2$ διαιρεμένο με το 4. Άρα,

$$\mathrm{Var}(T^+) = \sum_{i=1}^{n} \frac{i^2}{4} = \frac{n(n+1)(2n+1)}{24}. \tag{6.13}$$

Όσον αφορά την κατανομή της στατιστικής συνάρτησης $S = \sum_{i=1}^{n} R_i$ από τη σχέση (6.6) έχουμε ότι $S = 2T^+ - \frac{n(n+1)}{2}$. Επομένως, είναι $\mathrm{E}(S) = 2\mathrm{E}(T^+) - \frac{n(n+1)}{2}$ και υπό τη μηδενική υπόθεση, χρησιμοποιώντας και τη σχέση (6.10), έχουμε:

$$\mathrm{E}(S) = 2\mathrm{E}(T^+) - \frac{n(n+1)}{2} = \sum_{i=1}^{n} u_i - \frac{n(n+1)}{2},$$

όπου u_i αντιστοιχούν στις τάξεις των $R\left(|D_i|\right)$, για $i = 1, ..., n$. Όπως και πριν το άθροισμα $\sum_{i=1}^{n} u_i$, όταν δεν υπάρχουν δεσμοί (ή όταν υπάρχουν και χρησιμοποιείται η μέθοδος midranks, βλ. προηγούμενη υποσημείωση) είναι ίσο με το $n(n+1)/2$ και, επομένως, $\mathrm{E}(S) = 0$. Επιπλέον, έχουμε ότι $\mathrm{Var}(S) = 4\mathrm{Var}(T^+) = \sum_{i=1}^{n} u_i^2$, όπου u_i αντιστοιχούν στις τάξεις των $R\left(|D_i|\right)$, για $i = 1, ..., n$. Όμως, καθώς έχουμε υποθέσει ότι δεν υπάρχουν δεσμοί το αποτέλεσμα θα προκύπτει ως το άθροισμα $1^2 + 2^2 + ... + n^2$ και, επομένως, είναι

$$\mathrm{Var}(S) = \frac{n(n+1)(2n+1)}{6}.$$

Συνδυάζοντας τα παραπάνω, άμεσα προκύπτει το ζητούμενο αποτέλεσμα. □

Χρησιμοποιώντας την παραπάνω πρόταση, έχουμε τις ακόλουθες κρίσιμες περιοχές για τους αντίστοιχους ελέγχους:

(Α) $|W| \geq z_{a/2}$ για τον έλεγχο της $H_0 : m = m_0$ έναντι της $H_1 : m \neq m_0$.
(Β) $W \geq z_a$ για τον έλεγχο της $H_0 : m = m_0$ έναντι της $H_1 : m > m_0$.
(Γ) $W \leq -z_a$ για τον έλεγχο της $H_0 : m = m_0$ έναντι της $H_1 : m < m_0$.

[1] Στην πραγματικότητα, το ίδιο αποτέλεσμα ισχύει ακόμα και αν υπάρχουν δεσμοί και υπολογίζονται οι τάξεις καθεμίας παρατήρησης που συμμετέχει στον δεσμό ως ο μέσος όρος των τάξεων, καθώς και σε αυτήν την περίπτωση το άθροισμα των τάξεων θα ισούται με το άθροισμα $1 + 2 + ... + n = n(n+1)/2$.

Σε περίπτωση που χρησιμοποιηθεί η στατιστική συνάρτηση Z, οι κρίσιμες περιοχές είναι ανάλογες. Τέλος, λαμβάνοντας υπόψη την Παρατήρηση 6.1, έχουμε ότι οι κατανομές των T^+ και T^- ταυτίζονται. Επομένως, τα παραπάνω συμπεράσματα ισχύουν και για το T^-, άρα μπορεί κάποιος να χρησιμοποιήσει και το $T = \min\{T^+, T^-\}$. Προφανώς, καθώς προσεγγίζουμε διακριτές τυχαίες μεταβλητές από την κανονική κατανομή, μπορεί να γίνει και διόρθωση συνέχειας, με τον τρόπο που αναφέρθηκε στο προηγούμενο κεφάλαιο.

Σε όσα προαναφέρθηκαν είχε υποτεθεί η μη ύπαρξη δεσμών μεταξύ των απόλυτων διαφορών. Σε συνέχεια του Παραδείγματος 6.2 θα εξετάσουμε, αρχικά, αν διαφοροποιείται η ακριβής κατανομή, υπό τη μηδενική υπόθεση, της στατιστικής συνάρτησης T^+, όταν υπάρχουν δεσμοί στις απόλυτες διαφορές.

Παράδειγμα 6.4. (Hollander *et al.*, 2014) Αν το μέγεθος του δείγματος είναι ίσο με 4, να βρεθεί η ακριβής κατανομή του T^+, υπό τη μηδενική υπόθεση και την πρόσθετη υπόθεση ότι υπάρχουν ακριβώς 2 δεσμοί στις τιμές των απόλυτων διαφορών.

Λύση Παραδείγματος 6.4. Είναι εύκολα αντιληπτό ότι, αφού $n = 4$ και υπάρχουν δύο δεσμοί, ο πρώτος εκ των δύο θα είναι μεταξύ της 1ης και 2ης παρατήρησης και ο δεύτερος μεταξύ της 3ης και 4ης παρατήρησης. Επομένως, οι τάξεις των απόλυτων διαφορών σε αυτήν την περίπτωση είναι 1.5, 1.5, 3.5, 3.5, αντίστοιχα. Όταν το μέγεθος δείγματος είναι $n = 4$, έχουμε τις ακόλουθες 16 (2^4) το πλήθος, ισοπίθανες υπό τη μηδενική υπόθεση, περιπτώσεις ως προς το πρόσημο των διαφορών και τον τρόπο διάταξής τους:

$- - --$, $- + --$, $- - +-$, $- - -+$, $- + +-$,$- + -+$, $- - ++$, $- + ++$, $+ - --$,$+ + --$, $+ - +-$, $+ - -+$, $+ + +-$, $+ + -+$, $+ - ++$, $+ + ++$.

Επομένως, μπορούμε να υπολογίσουμε το άθροισμα των τάξεων που αντιστοιχούν στις θετικές διαφορές. Για λόγους ευκολίας στην παρουσίαση των πράξεων, δίνουμε τον Πίνακα 6.1. Στην 1η γραμμή είναι οι τιμές των τάξεων των απόλυτων διαφορών, ενώ, στη συνέχεια, δίνουμε τις 16 το πλήθος τετράδες με τα πρόσημα των διαφορών. Στην τελευταία στήλη, δίνουμε το άθροισμα των τάξεων που αντιστοιχούν στις θετικές διαφορές. Για παράδειγμα, στην 1η κατά σειρά τετράδα $(- - --)$ δεν έχουμε θετικές διαφορές και, άρα, το αντίστοιχο άθροισμα είναι μηδέν. Στην 6η κατά σειρά τετράδα $(- - ++)$ έχουμε θετικές διαφορές στην 3η και 4η θέση, δηλαδή αθροίζουμε τις τάξεις 3.5 και 3.5 και προκύπτει άθροισμα ίσο με 7. Με τον ίδιο τρόπο συμπληρώνονται και οι υπόλοιπες θέσεις της τελευταίας στήλης.

Από τα παραπάνω προκύπτει ότι η συνάρτηση πιθανότητας της κατανομής του T^+ υπό τη μηδενική υπόθεση είναι:

$$P(T^+ = t) = \begin{cases} 1/16, \text{ για } t = 0,3,7,10, \\ 2/16, \text{ για } t = 1.5,3.5,6.5,8.5, \\ 4/16, \text{ για } t = 5. \end{cases}$$

□

Συνδυάζοντας το αποτέλεσμα του προηγούμενου παραδείγματος με το αποτέλεσμα που δόθηκε στο Παράδειγμα 6.2, συμπεραίνουμε ότι η κατανομή του T^+ διαφοροποιείται όταν υπάρχουν δεσμοί. Επομένως, είναι εσφαλμένο για μικρό μέγεθος δείγματος να χρησιμοποιούνται οι κρίσιμες τιμές που δίνονται στο παράρτημα, καθώς αυτές έχουν προκύψει υπό την υπόθεση της μη ύπαρξης δεσμών. Σε περιπτώσεις ύπαρξης δεσμών και μικρού μεγέθους θα πρέπει κάποιος πρώτα να προσδιορίζει την κατανομή της στατιστικής συνάρτησης υπό τη μηδενική υπόθεση με παρόμοιο τρόπο με αυτόν του Παραδείγματος 6.4. Έπειτα, χρησιμοποιώντας αυτήν την κατανομή, θα είναι εφικτό να προσδιορίζει τις κρίσιμες τιμές του ελέγχου.

Το εύλογο ερώτημα που ίσως έχει προκύψει είναι αν τα αποτελέσματα που αφορούν την ασυμπτωτική κατανομή των στατιστικών συναρτήσεων T^+, T^- και S, που προσδιορίστηκαν στην Πρόταση 6.7, ισχύουν όταν υπάρχουν δεσμοί. Για να απαντηθεί το ερώτημα, παρατηρούμε ότι η διακύμανση και των δύο

	1.5	1.5	3.5	3.5	
1	−	−	−	−	0
2	−	+	+	+	8.5
3	+	−	+	+	8.5
4	+	+	−	+	6.5
5	+	+	+	−	6.5
6	−	−	+	+	7
7	−	+	−	+	5
8	−	+	+	−	5
9	+	−	−	+	5
10	+	−	+	−	5
11	+	+	−	−	3
12	−	−	−	+	3.5
13	−	−	+	−	3.5
14	−	+	−	−	1.5
15	+	−	−	−	1.5
16	+	+	+	+	10

Πίνακας 6.1: Υπολογισμοί για τον προσδιορισμό, υπό τη μηδενική υπόθεση, της κατανομής του T^+, για $n = 4$ και την πρόσθετη υπόθεση ότι υπάρχουν ακριβώς 2 δεσμοί στις τιμές των απόλυτων διαφορών.

στατιστικών συναρτήσεων υπολογίστηκε έχοντας υποθέσει ότι δεν υπάρχουν δεσμοί και κατά αυτόν τον τρόπο χρησιμοποιήθηκε ότι $\sum_{i=1}^{n} u_i^2 = 1^2 + 2^2 + ... + n^2 = \frac{n(n+1)(2n+1)}{6}$, όπου u_i είναι οι τιμές που αντιστοιχούν στις τάξεις των $R(|D_i|)$, για $i = 1,...,n$. Ας θεωρήσουμε, για παράδειγμα, την περίπτωση που $n = 4$ και υπάρχουν δύο δεσμοί. Τότε οι τάξεις είναι 1.5, 1.5, 3.5, 3.5 και έχουμε ότι το άθροισμα των τετραγώνων αυτών είναι 29, ενώ, αν δεν υπήρχαν δεσμοί, το άθροισμα των τετραγώνων των τάξεων θα ήταν 30 (αφού $1^2 + ... + 4^2 = 30$). Επομένως, είναι άμεσα αντιληπτό ότι θα πρέπει να υπάρξει κατάλληλη τροποποίηση στη στατιστική συνάρτηση. Για την περίπτωση των στατιστικών συναρτήσεων T^+ ή S η τροποποίηση αυτή (με χρήση των midranks) δίνεται στην πρόταση που ακολουθεί (για την απόδειξη, βλ. π.χ. Lehmann, 2006).

Πρόταση 6.8. Έστω $X_1,\dots,X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(\cdot)$. Ενδιαφερόμαστε να ελέγξουμε την υπόθεση ότι η διάμεσος m της άγνωστης κατανομής είναι ίση με m_0, όπου m_0 είναι ένας γνωστός αριθμός. Έστω, επίσης, $D_i = X_i - m_0$, $i = 1,...,n$, οι μη μηδενικές διαφορές και $d_1,...,d_c$, ο αριθμός των παρατηρήσεων σε καθεμία από τις c διαφορετικές απόλυτες διαφορές (σε αύξουσα τάξη μεγέθους), με $d_i \geq 1$, και $\sum_{i=1}^{c} d_i = n$. Τότε, υπό τη μηδενική υπόθεση, η προσεγγιστική κατανομή της στατιστικής συνάρτησης T^+ που παριστάνει το άθροισμα των τάξεων που αντιστοιχούν στις θετικές διαφορές είναι

$$W = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{0.25\sum_{i=1}^{n} u_i^2}} \xrightarrow{d} \mathcal{N}(0,1)$$

όπου u_i είναι οι τάξεις των $R(|D_i|)$, για $i = 1,...,n$ ή, ισοδύναμα,

$$W = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \sum_{i=1}^{c} \frac{d_i(d_i^2-1)}{48}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Επιπλέον,

$$Z = \frac{S}{\sqrt{\sum_{i=1}^{n} u_i^2}} \xrightarrow{d} \mathcal{N}(0,1),$$

όπου u_i είναι οι τάξεις των $R(|D_i|)$, για $i = 1,...,n$ ή, ισοδύναμα,

$$Z = \frac{S}{\sqrt{\frac{n(n+1)(2n+1)}{6} - \sum_{i=1}^{c} \frac{d_i\left(d_i^2-1\right)}{12}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Απόδειξη Πρότασης 6.8. Συμβολίζουμε με $u_1, u_2, ..., u_n$ τις διαθέσιμες τάξεις. Τότε, από τον ορισμό αυτών και από τις υποθέσεις, έχουμε ότι οι πρώτες d_1 το πλήθος είναι ίσες μεταξύ τους, οι επόμενες d_2 το πλήθος είναι ίσες μεταξύ τους και συνεχίζοντας με παρόμοιο τρόπο οι τελευταίες d_c το πλήθος είναι ίσες μεταξύ τους. Επομένως, χρησιμοποιώντας τα midranks οι πρώτες d_1 το πλήθος (δηλαδή οι τάξεις u_1, u_2,...,u_{d_1}) είναι ίσες με τον μέσο όρο των τάξεων που θα είχαν αν ήταν διαφορετικές, δηλαδή ίσες με τον μέσο όρο των 1, 2, ...d_1. Οπότε,

$$u_1 = ... = u_{d_1} = \frac{1 + ... + d_1}{d_1} = \frac{d_1 + 1}{2}.$$

Με παρόμοιο σκεπτικό οι επόμενες d_2 τάξεις (δηλαδή οι τάξεις u_{d_1+1}, u_{d_1+2},...,$u_{d_1+d_2}$) είναι ίσες με τον μέσο όρο των τάξεων που θα είχαν αν ήταν διαφορετικές, δηλαδή ίσες με τον μέσο όρο των $d_1 + 1$, $d_1 + 2$, ...$d_1 + d_2$. Επομένως,

$$u_{d_1+1} = ... = u_{d_1+d_2} = \frac{(d_1 + 1) + ... + (d_1 + d_2)}{d_2} = d_1 + \frac{d_2 + 1}{2}.$$

Με παρόμοιο τρόπο

$$u_{d_1+d_2+1} = ... = u_{d_1+d_2+d_3} = \frac{(d_1 + d_2 + 1) + ... + (d_1 + d_2 + d_3)}{d_3} = d_1 + d_2 + \frac{d_3 + 1}{2}$$

και συνεχίζουμε παρόμοια.

Επιπλέον, είναι (βλ. και απόδειξη Πρότασης 6.7):

$$T^+ = u_1 I_1 + ... + u_n I_n,$$

με $I_i, i = 1,..,n,$ δείκτρια (τυχαία) μεταβλητή που λαμβάνει την τιμή 1, αν η διαφορά είναι θετική, και την τιμή 0, όταν είναι αρνητική. Προφανώς, η δείκτρια I_i, $i = 1,...,n$ είναι μία Bernoulli διακριτή τυχαία μεταβλήτη με πιθανότητα επιτυχίας, υπό τη μηδενική υπόθεση, $p = 1/2$, οπότε $\mathrm{E}(I_i) = 0.5$ και $\mathrm{Var}(I_i) = 0.25$. Είναι τότε, όπως αναλυτικά αποδείχθηκε στις σχέσεις (6.10) και (6.12), $\mathrm{E}(T^+) = 0.5 \sum_{i=1}^{n} u_i$ και $\mathrm{Var}(T^+) = 0.25 \sum_{i=1}^{n} u_i^2$. Για την εύρεση του πρώτου αθροίσματος αρκεί να παρατηρήσουμε ότι το άθροισμα των τάξεων που συμμετέχουν σε κάθε δεσμό είναι ίσο με το άθροισμα των τάξεων αν δεν υπήρχαν δεσμοί. Επομένως, γίνεται άμεσα αντιληπτό ότι:

$$\mathrm{E}(T^+) = 0.5 \sum_{i=1}^{n} u_i = 0.5 \sum_{i=1}^{n} i = 0.5 \cdot \frac{n(n+1)}{2} = \frac{n(n+1)}{4}.$$

Άρα, υπό τη μηδενική υπόθεση, είναι

$$W = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{0.25 \sum_{i=1}^{n} u_i^2}} \xrightarrow{d} \mathcal{N}(0,1)$$

με u_i να είναι οι τάξεις των $R(|D_i|)$, για $i = 1,...,n$.

Επιπλέον, καθώς $S = 2T^+ - \frac{n(n+1)}{2}$, άμεσα προκύπτει (βλ. και απόδειξη Πρότασης 6.7) ότι:

$$Z = \frac{S}{\sqrt{\sum_{i=1}^{n} u_i^2}} \xrightarrow{d} \mathcal{N}(0,1),$$

με u_i να είναι οι τάξεις των $R(|D_i|)$, για $i = 1,...,n$.

Οι ισοδύναμες μορφές προκύπτουν υπολογίζοντας αναλυτικά το άθροισμα $\sum_{i=1}^{n} u_i^2$. Για τον υπολογισμό αυτού θα χρησιμοποιηθεί c φορές η σχέση:

$$\sum_{i=1}^{k} a_i^2 = \sum_{i=1}^{k} (a_i - \bar{a})^2 + k\bar{a}^2,$$

η οποία ισχύει για οποιαδήποτε επιλογή των a_i και για k θετικό ακέραιο. Αρχικά, θα εφαρμοστεί για την ειδική περίπτωση που $a_1 = 1$, $a_2 = 2$ ...,$a_{d_1} = d_1$, οπότε και είναι $\bar{a} = u_1 = ... = u_{d_1} = \frac{d_1+1}{2}$. Άμεσα έπεται ότι:

$$\sum_{i=1}^{d_1} a_i^2 = 1^2 + ... + d_1^2,$$

$$d_1\bar{a}^2 = \sum_{i=1}^{d_1} u_i^2,$$

και

$$\begin{aligned}
\sum_{i=1}^{d_1} (a_i - \bar{a})^2 &= \sum_{i=1}^{d_1} a_i^2 - 2\bar{a}\sum_{i=1}^{d_1} a_i + d_1\bar{a}^2 \\
&= \frac{d_1(d_1+1)(2d_1+1)}{6} - 2\frac{d_1+1}{2}\frac{d_1(d_1+1)}{2} + d_1\left(\frac{d_1+1}{2}\right)^2 \\
&= \frac{d_1(d_1+1)(2d_1+1)}{6} - \frac{d_1(d_1+1)^2}{4} = \frac{d_1(d_1+1)\{2(2d_1+1) - 3(d_1+1)\}}{12} \\
&= \frac{d_1(d_1+1)(d_1-1)}{12} = \frac{d_1\left(d_1^2-1\right)}{12}.
\end{aligned}$$

Συνδυάζοντας τα προηγούμενα έχουμε ότι:

$$1^2 + ... + d_1^2 = \sum_{i=1}^{d_1} u_i^2 + \frac{d_1\left(d_1^2-1\right)}{12}.$$

Με παρόμοιο τρόπο θεωρώντας $a_1 = d_1 + 1$,...,$a_{d_2} = d_1 + d_2$, οπότε $\bar{a} = u_{d_1+1} = ... = u_{d_1+d_2} = d_1 + \frac{d_2+1}{2}$, προκύπτει ότι:

$$(d_1+1)^2 + ... + (d_1+d_2)^2 = \sum_{i=d_1+1}^{d_1+d_2} u_i^2 + \frac{d_2\left(d_2^2-1\right)}{12}.$$

Συνεχίζοντας, κατά αυτόν τον τρόπο, έπεται ότι:

$$1^2 + ... + n^2 = \sum_{i=1}^{n} u_i^2 + \sum_{i=1}^{c} \frac{d_i\left(d_i^2-1\right)}{12}.$$

Άρα,

$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} i^2 - \sum_{i=1}^{c} \frac{d_i\left(d_i^2-1\right)}{12}$$

ή, ισοδύναμα,

$$\sum_{i=1}^{n} u_i^2 = \frac{n(n+1)(2n+1)}{6} - \sum_{i=1}^{c} \frac{d_i\left(d_i^2-1\right)}{12},$$

που ολοκληρώνει και την απόδειξη. □

Παρατήρηση 6.2. Η Πρόταση 6.7 μπορεί να προκύψει ως ειδική περίπτωση της Πρότασης 6.8 λαμβάνοντας υπόψη ότι στην περίπτωση μη ύπαρξης δεσμών $d_1 = \ldots = d_c = 1$ για $c = n$. Τέλος, σε όσα προηγήθηκαν σε αυτήν την ενότητα οι μηδενικές διαφορές έχουν αποκλειστεί από την ανάλυση και το n αναφέρεται στο μέγεθος του τροποποιημένου δείγματος, ενώ οι τάξεις των παρατηρήσεων σε περιπτώσεις ύπαρξης δεσμών υπολογίστηκαν ως ο μέσος όρος των τάξεων που αυτές οι παρατηρήσεις θα είχαν αν δεν υπήρχαν δεσμοί. Για περισσότερες πληροφορίες σχετικές με τους τρόπους χειρισμού των ισοβαθμιών και των μηδενικών διαφορών παραπέμπουμε στην εργασία του Pratt (1959).

6.3 Έλεγχοι ισότητας δύο πληθυσμιακών διαμέσων

Στην ενότητα αυτή, θα παρουσιαστούν τρόποι ελέγχου της υπόθεσης της ισότητας δύο πληθυσμιακών διαμέσων που βασίζονται στις τάξεις. Η μελέτη θα διεξαχθεί διακρίνοντας δύο περιπτώσεις. Στην πρώτη τα διαθέσιμα δείγματα από τους πληθυσμούς είναι ανεξάρτητα, ενώ στη δεύτερη είναι εξαρτημένα.

6.3.1 Ανεξάρτητα δείγματα: οι έλεγχοι του Wilcoxon και των Mann-Whitney

Έστω ότι έχουμε δύο πληθυσμούς, με αθροιστικές συναρτήσεις κατανομής $F_i, i = 1,2$. Επιπλέον, λαμβάνουμε δύο το πλήθος, ανεξάρτητα μεταξύ τους, τυχαία δείγματα από καθέναν από αυτούς τους δύο πληθυσμούς, μεγέθους n_i, $i = 1,2$, με $n_1 + n_2 = n$. Έστω οι δειγματικές τιμές $X_{i1}, \ldots, X_{i,n_i}$, από τον i-οστό πληθυσμό, $i = 1,2$. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση της ισότητας των κατανομών των πληθυσμών έναντι της εναλλακτικής ότι έχουν την ίδια μορφή, αλλά έχουν διαφορετικό μέτρο κεντρικής τάσης (central tendency). Η εναλλακτική αυτή είναι γνωστή στη βιβλιογραφία (βλ., μεταξύ άλλων, Gibbons and Chakraborti, 2020) ως εναλλακτική θέσης (location alternative). Συμβολικά έχουμε τον έλεγχο της μηδενικής υπόθεσης

$$H_0 : F_1(x) = F_2(x), \text{ για κάθε } x \in \mathbb{R},$$

έναντι μίας εκ των τριών εναλλακτικών

$$\text{(Α): } H_1 : F_1(x) = F_2(x-\theta), \text{ για κάθε } x \in \mathbb{R} \text{ και κάποιο } \theta \neq 0,$$

ή

$$\text{(Β): } H_1 : F_1(x) = F_2(x-\theta), \text{ για κάθε } x \in \mathbb{R} \text{ και κάποιο } \theta < 0,$$

ή

$$\text{(Γ): } H_1 : F_1(x) = F_2(x-\theta), \text{ για κάθε } x \in \mathbb{R} \text{ και κάποιο } \theta > 0.$$

Από τα παραπάνω προκύπτει ότι η α.σ.κ. του πρώτου πληθυσμού είναι μετατοπισμένη προς τα αριστερά, αν $\theta < 0$, ή προς τα δεξιά, αν $\theta > 0$. Η παραπάνω ιδιότητα συνεπάγεται ότι X_1 είναι στοχαστικά μεγαλύτερη

(μικρότερη) από την X_2 όταν $\theta > 0$ ($\theta < 0$, αντίστοιχα). Επομένως, όταν $\theta < 0$, η διάμεσος του δεύτερου πληθυσμού είναι μεγαλύτερη από τη διάμεσο του πρώτου πληθυσμού.

Συνοψίζοντας τα παραπάνω ο αρχικός έλεγχος ανάγεται στον έλεγχο της μηδενικής υπόθεσης $H_0 : m_{X_1} = m_{X_2}$, δηλαδή της υπόθεσης της ισότητας των πληθυσμιακών διαμέσων m_{X_i}, $i = 1,2$ έναντι μίας εκ των τριών εναλλακτικών

$$\text{(Α)}: H_1 : m_{X_1} \neq m_{X_2}$$

ή

$$\text{(Β)}: H_1 : m_{X_1} < m_{X_2}$$

ή

$$\text{(Γ)}: H_1 : m_{X_1} > m_{X_2}.$$

Σε όσα ακολουθούν $R\left(X_{ij}\right)$, $i = 1,2$, $j = 1,\ldots,n_i$, είναι οι τάξεις των διαθέσιμων δειγματικών τιμών των δύο δειγμάτων στο σύνολο των $n = n_1 + n_2$ παρατηρήσεων και R_i το άθροισμα των τάξεων του i-οστού δείγματος, $i = 1,2$. Δηλαδή $R_i = \sum_{j=1}^{n_i} R\left(X_{ij}\right)$, $i = 1,2$. Ειδικότερα, $R_1 = \sum_{j=1}^{n_1} R\left(X_{1j}\right)$, $R_2 = \sum_{j=1}^{n_2} R\left(X_{2j}\right)$ και $\sum_{i=1}^{2} R_i = \frac{n(n+1)}{2}$.

Αν υποθέσουμε ότι η H_0 είναι αληθής, αναμένουμε οι μέσοι όροι των τάξεων σε καθένα από τα δύο δείγματα να είναι περίπου ίσοι μεταξύ τους. Δηλαδή περιμένουμε να ισχύει ότι: $\frac{R_1}{n_1} = \frac{R_2}{n_2}$. Λαμβάνοντας επιπλέον υπόψη ότι $\sum_{i=1}^{2} R_i = \frac{n(n+1)}{2}$ προκύπτει ότι, όταν η H_0 είναι αληθής, τότε:

$$R_1 + \frac{n_2}{n_1} R_1 = \frac{n(n+1)}{2},$$

οπότε,

$$\frac{R_1}{n_1}(n_1 + n_2) = \frac{n(n+1)}{2}.$$

Συνδυάζοντας τα παραπάνω, έχουμε ότι, όταν η H_0 είναι αληθής, τότε:

$$\frac{R_1}{n_1} = \frac{R_2}{n_2} = \frac{n+1}{2}.$$

Από τα παραπάνω εξάγουμε το συμπέρασμα ότι ένας πρακτικός, αλλά όχι στατιστικός, τρόπος για να αποφανθούμε για την αποδοχή ή την απόρριψη της μηδενικής υπόθεσης ότι τα δείγματα προέρχονται από τον ίδιο πληθυσμό, είναι να εξετάζουμε αν οι ποσότητες $\frac{R_i}{n_i}$, $i = 1,2$ είναι περίπου ίσες μεταξύ τους και ίσες με $(n+1)/2$.

Στη βιβλιογραφία, έχουν εμφανιστεί δύο τρόποι ελέγχου οι οποίοι στηρίζονται στην παραπάνω κεντρική ιδέα, ένας από τον Wilcoxon (1945) και ένας από τους Mann and Whitney (1947). Οι έλεγχοι αυτοί, οι οποίοι παρουσιάστηκαν ανεξάρτητα, θα δούμε ότι είναι ισοδύναμοι. Για τον λόγο αυτόν έχει καθιερωθεί και ως έλεγχος Wilcoxon-Mann-Whitney. Στη συνέχεια, θα παρουσιαστεί καθένας εξ αυτών.

Πριν προχωρήσουμε στα όσα προτάθηκαν από τον Wilcoxon (1945), παραθέτουμε μία πρόταση χρήσιμη για όσα ακολουθούν.

Πρόταση 6.9. Υπό τη μηδενική υπόθεση και υποθέτοντας ότι δεν υπάρχουν δεσμοί μεταξύ των δεδομένων αποδεικνύεται ότι:

α) $\mathrm{E}(R_i) = n_i \frac{n+1}{2}$, $i = 1,2$.

β) $\mathrm{Var}(R_i) = n_i \frac{(n+1)(n-n_i)}{12}$, $i = 1,2$.

γ) Για μεγάλα σε μέγεθος δείγματα $R_1 \xrightarrow{d} \mathcal{N}\left(\frac{n_1(n+1)}{2}, n_1 n_2 \frac{(n+1)}{12}\right)$.

Απόδειξη Πρότασης 6.9. α) Χρησιμοποιώντας την Πρόταση 6.1 β) έπεται, άμεσα, ότι:

$$\mathsf{E}(R_i) = \mathsf{E}\left(\sum_{j=1}^{n_i} R\left(X_{ij}\right)\right) = \sum_{j=1}^{n_i} \mathsf{E}\left(R\left(X_{ij}\right)\right) = \sum_{j=1}^{n_i} \frac{n+1}{2} = \frac{n_i(n+1)}{2}.$$

β) Επειδή ισχύει ότι $\sum_{i=1}^{2} R_i = \frac{n(n+1)}{2}$, καταλαβαίνουμε ότι οι τυχαίες μεταβλητές R_1 και R_2 δεν είναι ανεξάρτητες. Άρα, για τον υπολογισμό της διακύμανσης $\mathsf{Var}(R_i)$ θα ισχύει:

$$\mathsf{Var}(R_i) = \mathsf{Var}\left(\sum_{j=1}^{n_i} R\left(X_{ij}\right)\right) = \sum_{j=1}^{n_i} \mathsf{Var}\left(R\left(X_{ij}\right)\right) + \sum_{j=1}^{n_i} \sum_{l=1, l\neq j}^{n_i} \mathsf{Cov}\left(R\left(X_{ij}\right), R\left(X_{il}\right)\right).$$

Χρησιμοποιώντας την Πρόταση 6.1 και συγκεκριμένα τα β) και γ), έπεται ότι:

$$\mathsf{Var}\left(R\left(X_{ij}\right)\right) = \frac{(n+1)(n-1)}{12},$$

και

$$\mathsf{Cov}\left(R\left(X_{ij}\right), R\left(X_{il}\right)\right) = -\frac{n+1}{12}.$$

Επομένως,

$$\begin{aligned} \mathsf{Var}(R_i) &= \sum_{j=1}^{n_i} \frac{(n+1)(n-1)}{12} + \sum_{j=1}^{n_i} \sum_{l=1, l\neq j}^{n_i} -\frac{n+1}{12} \\ &= \frac{n_i(n+1)(n-1)}{12} - \frac{n_i(n_i-1)(n+1)}{12} = \frac{n_i(n+1)(n-n_i)}{12}. \end{aligned}$$

γ) Παρατηρούμε ότι $R_1 = \sum_{j=1}^{n_1} R\left(X_{1j}\right)$, δηλαδή το R_1 είναι άθροισμα n_1 το πλήθος ισόνομων τυχαίων μεταβλητών, με $\mathsf{E}(R_1) = n_1 \frac{n+1}{2}$ και $\mathsf{Var}(R_1) = n_1 n_2 \frac{(n+1)}{12}$. Το ζητούμενο αποτέλεσμα προκύπτει άμεσα ως εφαρμογή του Κεντρικού Οριακού Θεωρήματος. Δηλαδή, υπό τη μηδενική υπόθεση, για μεγάλα σε μέγεθος δείγματα

$$R_1 \xrightarrow{d} \mathcal{N}\left(\frac{n_1(n+1)}{2}, n_1 n_2 \frac{(n+1)}{12}\right),$$

και η απόδειξη ολοκληρώνεται. □

6.3.1.1 Ο έλεγχος του Wilcoxon

Καθώς από την Πρόταση 6.9 έχουμε ότι υπό τη μηδενική υπόθεση $\mathsf{E}(R_i) = n_i \frac{n+1}{2}$, η στατιστική συνάρτηση που προτάθηκε από τον Wilcoxon (1945) για τον έλεγχο της μηδενικής υπόθεσης είναι η

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

ή, εναλλακτικά, η:

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}.$$

Ένα εύλογο ερώτημα είναι ποια από τις δύο στατιστικές συναρτήσεις είναι προτιμότερο να υπολογιστεί και να χρησιμοποιηθεί για τον έλεγχο της μηδενικής υπόθεσης. Για να απαντηθεί αυτό το ερώτημα, αρχικά, παρατηρήστε ότι:

$$U_1 + U_2 = R_1 - \frac{n_1(n_1+1)}{2} + R_2 - \frac{n_2(n_2+1)}{2} = \frac{n(n+1)}{2} - \frac{n_1(n_1+1)}{2} - \frac{n_2(n_2+1)}{2} = n_1 n_2,$$

οπότε:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \text{ και } U_2 = n_1 n_2 - U_1.$$

Επομένως, αν υπολογιστεί η μία εκ των δύο στατιστικών συναρτήσεων άμεσα υπολογίζεται και η άλλη. Για λόγους συντομίας προφανώς προτιμότερο είναι να υπολογιστεί η στατιστική συνάρτηση που αντιστοιχεί στο μικρότερο σε μέγεθος δείγμα. Από τον τρόπο ορισμού των στατιστικών συναρτήσεων έχουμε ότι για τον έλεγχο της $H_0 : m_{X_1} = m_{X_2}$, έναντι της εναλλακτικής $H_1 : m_{X_1} > m_{X_2}$, θα απορρίπτεται η μηδενική υπόθεση για μεγάλες τιμές της U_1 ή μικρές τιμές της U_2. Επίσης, για τον έλεγχο της $H_0 : m_{X_1} = m_{X_2}$ έναντι της εναλλακτικής $H_1 : m_{X_1} < m_{X_2}$, θα απορρίπτεται η μηδενική υπόθεση για μικρές τιμές της U_1 ή μεγάλες τιμές της U_2. Τέλος, για τον έλεγχο της $H_0 : m_{X_1} = m_{X_2}$ έναντι της εναλλακτικής $H_1 : m_{X_1} \neq m_{X_2}$, θα απορρίπτεται η μηδενική υπόθεση τόσο για μικρές όσο και για μεγάλες τιμές των U_1 και U_2. Τέλος, επισημαίνουμε ότι πολλές φορές στην πράξη χρησιμοποιείται η στατιστική συνάρτηση

$$U = \min\{U_1, U_2\} = \min\{U_1, n_1 n_2 - U_1\}.$$

Τότε, καθώς επιλέγεται η μικρότερη τιμή εκ των U_1 και U_2, η μηδενική υπόθεση απορρίπτεται για μικρές τιμές της στατιστικής συνάρτησης που είναι τέτοιες, ώστε να έχουμε επίπεδο σημαντικότητας a.

Για τον προσδιορισμό των τιμών που θεωρούνται μικρές (ή μεγάλες, ανάλογα με τη μορφή της εναλλακτικής) για να απορρίπτεται η μηδενική υπόθεση, απαιτείται ο προσδιορισμός της κατανομής των στατιστικών συναρτήσεων U_1, U_2 υπό τη μηδενική υπόθεση. Για την επίτευξη αυτού του σκοπού αρκεί να προσδιοριστούν οι ακριβείς κατανομές των στατιστικών συναρτήσεων R_1 και R_2 υπό τη μηδενική υπόθεση, καθώς $U_i = R_i - \frac{n_i(n_i+1)}{2}$, για $i = 1,2$, με τις R_i να είναι διακριτές τυχαίες μεταβλητές. Προφανώς, καθώς οι τυχαίες μεταβλητές R_i, U_i, $i = 1,2$, είναι διακριτές τυχαίες μεταβλητές, με $U_1 + U_2 = n_1 n_2$ και $R_1 + R_2 = \frac{n(n+1)}{2}$, αρκεί να προσδιοριστεί η κατανομή μίας εκ των R_i, U_i, για $i = 1,2$. Στην πρόταση που ακολουθεί (βλ., μεταξύ άλλων, Randles and Wolfe, 1979) προσδιορίζεται η ακριβής κατανομή υπό τη μηδενική υπόθεση της R_1. Προφανώς, αντίστοιχο αποτέλεσμα ισχύει για την κατανομή της στατιστικής συνάρτησης R_2 κάνοντας κατάλληλες τροποποιήσεις.

Πρόταση 6.10. Έστω ότι έχουμε δύο πληθυσμούς, με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1,2$. Επιπλέον, έστω $X_{i1},\ldots,X_{i,n_i}$, $i = 1,2$ δύο, ανεξάρτητα μεταξύ τους, τυχαία δείγματα από καθέναν από αυτούς τους δύο πληθυσμούς, μεγέθους n_i, $i = 1,2$, με $n_1 + n_2 = n$. Επίσης, θεωρούμε ότι R_1 είναι το άθροισμα των τάξεων των παρατηρήσεων του πρώτου δείγματος στο σύνολο αυτών των n τιμών. Υπό την υπόθεση ότι τα δύο δείγματα προέρχονται από τον ίδιο πληθυσμό και δεν υπάρχουν δεσμοί μεταξύ αυτών, η συνάρτηση πιθανότητας της στατιστικής συνάρτησης R_1 δίνεται από τη σχέση:

$$P(R_1 = r) = \frac{t_{n_1,n_2}(r)}{\binom{n}{n_1}},\ r = \frac{n_1(n_1+1)}{2},\ldots,\frac{n_1(2n_2+n_1+1)}{2},$$

όπου $t_{n_1,n_2}(r)$ συμβολίζει το πλήθος των μη διατεταγμένων υποσυνόλων n_1 ακεραίων που επιλέγονται χωρίς επανατοποθέτηση από το σύνολο $\{1,\ldots,n\}$ και έχουν άθροισμα ίσο με r και $\binom{n}{n_1}$ το πλήθος των δυνατών διατάξεων των δειγματικών τιμών.

Απόδειξη Πρότασης 6.10. Για την απόδειξη παραπέμπουμε στην εργασία του Wilcoxon (1945). □

$(R(X_{11}), R(X_{12}))$	R_1
(1,2)	3
(1,3)	4
(1,4),(2,3)	5
(1,5),(2,4)	6
(2,5),(3,4)	7
(3,5)	8
(4,5)	9

Πίνακας 6.2: Υπολογισμοί για τον προσδιορισμό, υπό τη μηδενική υπόθεση, της κατανομής του R_1, για $n_1 = 2$ και $n_2 = 3$ (χωρίς δεσμούς).

Από την Πρόταση 6.10 προκύπτει ότι η ακριβής κατανομή, υπό τη μηδενική υπόθεση, της στατιστικής συνάρτησης R_1 εξαρτάται από το μέγεθος κάθε δείγματος, αλλά όχι από την κατανομή του πληθυσμού από όπου προέρχονται οι δειγματικές παρατηρήσεις. Κατά συνέπεια, ο έλεγχος που βασίζεται στη συγκεκριμένη στατιστική συνάρτηση είναι απαλλαγμένος παραμέτρων. Επιπλέον, προκύπτει ότι δεν μπορεί στη γενική περίπτωση να δοθεί η κατανομή σε κλειστή μορφή, καθώς δεν υπάρχει κλειστή μορφή για το $t_{n_1,n_2}(r)$. Στα δύο παραδείγματα που ακολουθούν δίνεται ο τρόπος σκέψης για τον προσδιορισμό της ακριβούς κατανομής, υπό τη μηδενική υπόθεση, των στατιστικών συναρτήσεων R_1 και U_2, αντίστοιχα, στην περίπτωση που έχουμε μικρά σε μέγεθος δείγματα.

Παράδειγμα 6.5. Υποθέτοντας ότι δεν υπάρχουν δεσμοί μεταξύ των δεδομένων να προσδιορίσετε, υπό τη μηδενική υπόθεση, την ακριβή κατανομή του R_1, όταν $n_1 = 2$ και $n_2 = 3$.

Λύση Παραδείγματος 6.5. Έστω ότι X_{11}, X_{12}, και X_{21}, X_{22}, X_{23} είναι οι δειγματικές τιμές από τον πρώτο και δεύτερο πληθυσμό, αντίστοιχα. Επιπλέον, έστω $R\left(X_{ij}\right)$, $i = 1,2$, $j = 1,...,n_i$, με $n_1 = 2$ και $n_2 = 3$, οι τάξεις των διαθέσιμων δειγματικών τιμών των δύο δειγμάτων στο σύνολο των $n = n_1 + n_2 = 5$ το πλήθος παρατηρήσεων. Επίσης, R_1 είναι το άθροισμα των τάξεων του πρώτου δείγματος, δηλαδή $R_1 = \sum_{j=1}^{2} R\left(X_{1j}\right)$. Οι δυνατές τιμές της τυχαίας μεταβλητής R_1, καθώς και οι αντίστοιχες τιμές των $R(X_{11})$ και $R(X_{12})$, παρατίθενται στον Πίνακα 6.2.

Για παράδειγμα, αν $R(X_{11}) = 1$ και $R(X_{12}) = 2$, αυτό σημαίνει ότι στο κοινό δείγμα των πέντε το πλήθος τιμών, δύο από τον πρώτο πληθυσμό και τριών από τον δεύτερο πληθυσμό, οι τιμές από τον πρώτο πληθυσμό είναι οι μικρότερες, αφού έχουν τάξεις 1 και 2, και, άρα, το άθροισμα των τάξεων του πρώτου δείγματος είναι $1 + 2 = 3$. Όμοια, αν $R(X_{11}) = 3$, $R(X_{12}) = 5$, αυτό σημαίνει πως στο κοινό δείγμα των πέντε τιμών, οι τιμές από τον πρώτο πληθυσμό είναι η 3η μεγαλύτερη και η 5η μεγαλύτερη κατά αύξουσα τάξη μεγέθους, αφού έχουν τάξεις 3 και 5, και, άρα, το άθροισμα των τάξεων του πρώτου δείγματος είναι $3 + 5 = 8$.

Επομένως, οι δυνατές τιμές της τυχαίας μεταβλητής R_1 είναι οι $\{3,4,5,6,7,8,9\}$ και η τυχαία μεταβλητή έχει, υπό τη μηδενική υπόθεση, τουτέστιν υπό την υπόθεση ότι κάθε δυνατή διάταξη είναι ισοπίθανη, την ακόλουθη συνάρτηση πιθανότητας:

$$P(R_1 = r) = \begin{cases} 1/10, \text{ για } r = 3,4,8,9, \\ 2/10, \text{ για } r = 5,6,7, \end{cases}$$

αφού καθένα από τα ζεύγη τιμών $(R(X_{11}), R(X_{12}))$ έχει την ίδια πιθανότητα εμφάνισης. □

Παράδειγμα 6.6. Υποθέτοντας ότι δεν υπάρχουν δεσμοί μεταξύ των δεδομένων να προσδιορίσετε, υπό τη μηδενική υπόθεση, την ακριβή κατανομή της στατιστικής συνάρτησης $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$, για $n_1 = 4$ και $n_2 = 2$.

Λύση Παραδείγματος 6.6. Έστω $X_{11}, X_{12}, X_{13}, X_{14}$, και X_{21}, X_{22} είναι οι δειγματικές τιμές από τον πρώτο και δεύτερο πληθυσμό, αντίστοιχα. Επιπλέον, έστω $R\left(X_{ij}\right)$, $i = 1,2$, $j = 1,...,n_i$, με $n_1 = 4$ και $n_2 = 2$, οι

$(R(X_{21}), R(X_{22}))$	R_2	U_2
(1,2)	3	0
(1,3)	4	1
(1,4),(2,3)	5	2
(1,5),(2,4)	6	3
(1,6),(2,5),(3,4)	7	4
(2,6), (3,5)	8	5
(3,6), (4,5)	9	6
(4,6)	10	7
(5,6)	11	8

Πίνακας 6.3: Υπολογισμοί για τον προσδιορισμό, υπό τη μηδενική υπόθεση, της κατανομής της στατιστικής συνάρτησης U_2, για $n_1 = 4$ και $n_2 = 2$ (χωρίς δεσμούς).

τάξεις των διαθέσιμων δειγματικών τιμών των δύο δειγμάτων στο σύνολο των $n = n_1 + n_2 = 6$ το πλήθος παρατηρήσεων. Επίσης, το R_2 είναι το άθροισμα των τάξεων του δεύτερου δείγματος, δηλαδή $R_2 = \sum_{j=1}^{2} R\left(X_{2j}\right)$. Οι δυνατές θέσεις που μπορούν να βρεθούν οι παρατηρήσεις του 2ου πληθυσμού είναι 15 $\left(= \binom{6}{2}\right)$ και είναι οι ακόλουθες: (1,2), (1,3), (1,4), (1,5), (1,6), (2,3), (2,4), (2,5), (2,6), (3,4), (3,5), (3,6), (4,5), (4,6), (5,6). Υπό τη μηδενική υπόθεση καθεμία από αυτές έχει ίδια πιθανότητα εμφάνισης. Επιπλέον, καθώς υποθέτουμε ότι δεν υπάρχουν δεσμοί μεταξύ των δεδομένων, οι παραπάνω τιμές είναι και οι τάξεις κάθε παρατήρησης από τον δεύτερο πληθυσμό. Οι δυνατές τιμές των τυχαίων μεταβλητών R_2 και U_2, καθώς και οι αντίστοιχες τιμές των $R(X_{21})$, $R(X_{22})$ παρατίθενται στον Πίνακα 6.3. Επομένως, η διακριτή τυχαία μεταβλητή U_2 έχει συνάρτηση πιθανότητας:

$$P(U_2 = u) = \begin{cases} 1/15, \text{ για } u = 0,1,7,8, \\ 2/15, \text{ για } u = 2,3,5,6, \\ 3/15, \text{ για } u = 4. \end{cases}$$

□

Οι επόμενες δύο προτάσεις μας δίνουν χρήσιμα αποτελέσματα που βοηθούν στον προσδιορισμό της περιοχής απόρριψης της μηδενικής υπόθεσης. Τα αποτελέσματα δίνονται για τη στατιστική συνάρτηση R_1, ενώ παρόμοια αποτελέσματα ισχύουν για τη στατιστική συνάρτηση R_2.

Πρόταση 6.11. Έστω ότι έχουμε δύο πληθυσμούς, με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1,2$. Επιπλέον, έστω $X_{i1}, \dots, X_{i,n_i}$, $i = 1,2$ δύο, ανεξάρτητα μεταξύ τους, τυχαία δείγματα από καθέναν από αυτούς τους δύο πληθυσμούς, μεγέθους n_i, $i = 1,2$, με $n_1 + n_2 = n$. Επίσης, θεωρούμε ότι R_1 είναι το άθροισμα των τάξεων των παρατηρήσεων του πρώτου δείγματος στο σύνολο αυτών των n τιμών. Υπό την υπόθεση ότι τα δύο δείγματα προέρχονται από τον ίδιο πληθυσμό και δεν υπάρχουν δεσμοί μεταξύ αυτών, η κατανομή της στατιστικής συνάρτησης R_1 είναι συμμετρική γύρω από το σημείο $\frac{n_1(n+1)}{2}$, το οποίο συμπίπτει με τη μέση τιμή της.

Απόδειξη Πρότασης 6.11. Θεωρούμε διάταξη κατά φθίνουσα σειρά (αντίστροφη διάταξη) των X_{1j}, $j = 1, \dots, n_1$ και έστω $R'(X_{11}), \dots, R'(X_{1,n_1})$ οι καινούριες τάξεις μεγέθους αυτών. Οπότε,

$$R'(X_{1j}) = (n+1) - R(X_{1j})$$

και, επομένως,

$$R_1' = \sum_{j=1}^{n_1} R'(X_{1j}) = \sum_{j=1}^{n_1} [(n+1) - R(X_{1j})] = n_1(n+1) - R_1,$$

ή, ισοδύναμα,

$$R_1' - \frac{n_1(n+1)}{2} = \frac{n_1(n+1)}{2} - R_1. \tag{6.14}$$

Υπό τη μηδενική υπόθεση είναι προφανές ότι οι τυχαίες μεταβλητές R_1 και R_1' έχουν την ίδια κατανομή, οπότε λόγω της σχέσης (6.14) προκύπτει ότι οι τυχαίες μεταβλητές $R_1 - \frac{n_1(n+1)}{2}$ και $\frac{n_1(n+1)}{2} - R_1$ έχουν την ίδια κατανομή, δηλαδή για κάθε $k = 0,1,\ldots,\frac{n_1 n_2}{2}$ ισχύει ότι:

$$P\left(R_1 - \frac{n_1(n+1)}{2} = k\right) = P\left(\frac{n_1(n+1)}{2} - R_1 = k\right), \forall\, k \in \{0,1,\ldots,\frac{n_1 n_2}{2}\},$$

ή, διαφορετικά,

$$P\left(R_1 = \frac{n_1(n+1)}{2} + k\right) = P\left(R_1 = \frac{n_1(n+1)}{2} - k\right).$$

Το τελευταίο σημαίνει ότι η R_1 είναι συμμετρική γύρω από το σημείο $\frac{n_1(n+1)}{2}$, το οποίο σημείο συμπίπτει με τη μέση τιμή της (βλ. Πρόταση 6.9). □

Πρόταση 6.12. Έστω w_p είναι το p-ποσοστιαίο σημείο της κατανομής της στατιστικής συνάρτησης R_1, τότε

$$w_{1-p} = n_1(n+1) - w_p, \text{ για κάθε } p \in (0,1).$$

Απόδειξη Πρότασης 6.12. Από τον ορισμό του ποσοστιαίου σημείου μιας διακριτής κατανομής πιθανότητας, το w_p είναι τέτοιο, ώστε

$$P\left(R_1 > w_p\right) \leqslant p \leqslant P\left(R_1 \geqslant w_p\right). \tag{6.15}$$

Από τη μία πλευρά της διπλής ανισότητας της σχέσης (6.15), έχουμε,

$$P\left(R_1 > w_p\right) \leqslant p \Leftrightarrow P\left(R_1 - \frac{n_1(n+1)}{2} > w_p - \frac{n_1(n+1)}{2}\right) \leqslant p$$

ή ισοδύναμα (λόγω του ότι οι τυχαίες μεταβλητές $R_1 - \frac{n_1(n+1)}{2}$ και $\frac{n_1(n+1)}{2} - R_1$ έχουν την ίδια κατανομή)

$$P\left(R_1 < n_1(n+1) - w_p\right) \leqslant p \Leftrightarrow 1 - P\left(R_1 \geqslant n_1(n+1) - w_p\right) \leqslant p,$$

από όπου προκύπτει ότι:

$$P\left(R_1 \geqslant n_1(n+1) - w_p\right) \geqslant 1 - p. \tag{6.16}$$

Ομοίως, χρησιμοποιώντας την άλλη πλευρά της διπλής ανισότητας της σχέσης (6.15), προκύπτει ότι:

$$P\left(R_1 \geqslant w_p\right) \geqslant p \Leftrightarrow P\left(R_1 - \frac{n_1(n+1)}{2} \geqslant w_p - \frac{n_1(n+1)}{2}\right) \geqslant p$$

ή ισοδύναμα

$$P\left(R_1 \leqslant n_1(n+1) - w_p\right) \geqslant p \Leftrightarrow 1 - P\left(R_1 > n_1(n+1) - w_p\right) \geqslant p.$$

Από αυτήν τη σχέση προκύπτει ότι:

$$P\left(R_1 > n_1(n+1) - w_p\right) \leqslant 1 - p. \tag{6.17}$$

Συνδυάζοντας τις σχέσεις (6.16), (6.17) και τον ορισμό του $(1-p)$-ποσοστιαίου σημείου της κατανομής της R_1 είναι προφανές ότι

$$w_{1-p} = n_1(n+1) - w_p, \text{ για κάθε } p \in (0,1),$$

και η απόδειξη ολοκληρώθηκε. □

Από τη συζήτηση που προηγήθηκε της Πρότασης 6.10 προκύπτει ότι, χρησιμοποιώντας τις στατιστικές συναρτήσεις U_1 και U_2 για τη διεξαγωγή των μονόπλευρων ελέγχων σε επίπεδο σημαντικότητας a, θα πρέπει να προσδιοριστούν τα άνω και κάτω a ποσοστιαία σημεία (ανάλογα αν έχουμε άνω μονόπλευρο ή κάτω μονόπλευρο έλεγχο), ενώ για τον δίπλευρο έλεγχο απαιτείται τόσο το άνω όσο και το κάτω $a/2$ ποσοστιαίο σημείο. Από την άλλη μεριά, όταν χρησιμοποιείται η στατιστική συνάρτηση $U = \min\{U_1, U_2\}$, για τους μεν μονόπλευρους ελέγχους χρειαζόμαστε το κάτω a ποσοστιαίο σημείο, ενώ για τον δίπλευρο έλεγχο το κάτω $a/2$ ποσοστιαίο σημείο. Δηλαδή η μηδενική υπόθεση απορρίπτεται για μονόπλευρες εναλλακτικές όταν: $P(U \leq u) < a$, ενώ για τον δίπλευρο έλεγχο όταν $P(U \leq u) < a/2$, όπου u η παρατηρούμενη τιμή της στατιστικής συνάρτησης στο δείγμα. Πίνακες πιθανοτήτων είναι διαθέσιμοι τόσο για την περίπτωση που $n_1 = n_2$, όσο και για την περίπτωση που $n_1 \neq n_2$ (βλ. Πίνακες Π.19-Π.21 του Παραρτήματος).

Στο σημείο αυτό, αξίζει να αναφέρουμε ότι ο έλεγχος μπορεί να πραγματοποιηθεί και με χρήση της στατιστικής συνάρτησης R_1. Συγκεκριμένα, για τη διεξαγωγή των μονόπλευρων ελέγχων σε επίπεδο σημαντικότητας a θα πρέπει να προσδιοριστούν τα άνω και κάτω a ποσοστιαία σημεία (ανάλογα αν έχουμε άνω μονόπλευρο ή κάτω μονόπλευρο έλεγχο), ενώ για τον δίπλευρο έλεγχο απαιτείται τόσο το άνω όσο και το κάτω $a/2$ ποσοστιαίο σημείο. Λόγω της συμμετρίας της κατανομής της τυχαίας μεταβλητής R_1 αρκεί να προσδιορίσουμε είτε το άνω είτε το κάτω ποσοστιαίο σημείο αυτής. Στο Παράρτημα, και συγκεκριμένα στους Πίνακες Π.23-Π.26, δίνονται τα κάτω p-ποσοστιαία σημεία της κατανομής της R_1 για μεγέθη δείγματος $n_1, n_2 \in \{2,3,\ldots,20\}$ και $p \in \{0.001, 0.005, 0.10, 0.025, 0.05, 0.10\}$.

Ο τρόπος υπολογισμού της στατιστικής συνάρτησης U και ο προτεινόμενος έλεγχος, χρησιμοποιώντας την ακριβή κατανομή υπό τη μηδενική υπόθεση, διευκρινίζονται στο παράδειγμα που ακολουθεί.

Παράδειγμα 6.7. Να ελέγξετε αν τα παρακάτω δεδομένα προέρχονται από διαφορετικές κατανομές (δίπλευρος έλεγχος με $a = 5\%$).

X_{1j}	6	8	10	13
X_{2j}	9	12		

Λύση Παραδείγματος 6.7. Έχουμε δύο το πλήθος πληθυσμούς, με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1,2$. Επιπλέον, λαμβάνουμε δύο το πλήθος, ανεξάρτητα μεταξύ τους τυχαία δείγματα από καθέναν από αυτούς τους δύο πληθυσμούς μεγέθους n_i, $i = 1,2$, με $n_1 = 4$, $n_2 = 2$, και $n_1 + n_2 = n = 6$. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση

$$H_0 : F_1(x) = F_2(x), \text{ για κάθε } x \in \mathbb{R},$$

έναντι της εναλλακτικής υπόθεσης

$$H_1 : F_1(x) \neq F_2(x), \text{ για κάποιο } x \in \mathbb{R}.$$

Η στατιστική συνάρτηση που προτάθηκε από τον Wilcoxon (1945) για τον έλεγχο της μηδενικής υπόθεσης είναι η $U = \min\{U_1, U_2\} = \min\{U_1, n_1 n_2 - U_1\}$, όπου $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$, με R_i να συμβολίζει το άθροισμα των τάξεων στο σύνολο των διαθέσιμων δειγματικών τιμών του i-οστού δείγματος, $i = 1,2$. Επομένως, αρχικά αναμειγνύονται τα δύο δείγματα και διατάσσονται κατά αύξουσα τάξη μεγέθους, ενώ στη συνέχεια υπολογίζουμε τις τάξεις των διαθέσιμων δειγματικών τιμών των δύο δειγμάτων στο σύνολο των $n = 6$ το πλήθος παρατηρήσεων. Στον πίνακα που ακολουθεί δίνονται οι 6 παρατηρήσεις κατά αύξουσα τάξη μεγέθους και οι τάξεις αυτών. Για διευκόλυνση στους μετέπειτα υπολογισμούς έχουμε υπογραμμίσει τις τιμές και τις τάξεις των παρατηρήσεων του δεύτερου δείγματος.

Παρατήρηση	6	8	9	10	12	13
Τάξη	1	2	3	4	5	6

Δεν είναι δύσκολο να διαπιστώσουμε ότι:

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2} = (3+5) - \frac{2(2+1)}{2} = 5$$

και, άρα,

$$U_1 = n_1 n_2 - U_2 = 2 \cdot 4 - 5 = 3.$$

Επομένως, χρησιμοποιείται η στατιστική συνάρτηση $U = \min\{U_1, U_2\}$ και η μηδενική υπόθεση απορρίπτεται αν $P(U \leq u) < a/2 = 0.025$, όπου u η παρατηρούμενη τιμή της στατιστικής συνάρτησης ελέγχου στο δείγμα, δηλαδή $u = 3$ σε αυτήν την περίπτωση. Από το Παράδειγμα 6.6 προκύπτει ότι, υπό τη μηδενική υπόθεση, η p-τιμή του ελέγχου είναι $P(U \leq 3) = 6/15 = 0.4$. Καθώς $0.4 > 0.025$, δεν απορρίπτεται η μηδενική υπόθεση και συμπεραίνουμε, σε επίπεδο σημαντικότητας 5%, ότι δεν μπορούμε να απορρίψουμε την υπόθεση ότι τα δύο ανεξάρτητα δείγματα προέρχονται από τον ίδιο πληθυσμό. □

Όταν το πλήθος των n_1, n_2 είναι μεγάλο (ο Lehmann (2006) αναφέρει την τιμή 10) οδηγούμαστε, υπό τη μηδενική υπόθεση, στην εύρεση της προσεγγιστικής κατανομής των στατιστικών συναρτήσεων που παρουσιάστηκαν πρωτύτερα. Η κατανομή αυτή δίνεται στην επόμενη πρόταση.

Πρόταση 6.13. Έστω δύο πληθυσμοί με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1,2$. Επιπλέον, έστω $X_{i1}, \ldots, X_{i,n_i}$, $i = 1,2$ δύο, ανεξάρτητα μεταξύ τους, τυχαία δείγματα από καθένα από αυτούς τους δύο πληθυσμούς, μεγέθους n_i, $i = 1,2$, με $n_1 + n_2 = n$. Έστω R_1 και R_2 το άθροισμα των τάξεων των παρατηρήσεων του πρώτου και δεύτερου δείγματος, αντίστοιχα, στο σύνολο αυτών των n το πλήθος τιμών. Υπό την υπόθεση ότι τα δύο δείγματα προέρχονται από τον ίδιο πληθυσμό και δεν υπάρχουν δεσμοί μεταξύ αυτών, για μεγάλες τιμές των n_1, n_2 ισχύει ότι:

$$Z_i = \frac{R_i - \frac{n_i(n+1)}{2}}{\sqrt{n_1 n_2 \frac{(n+1)}{12}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Απόδειξη Πρότασης 6.13. Η απόδειξη προκύπτει άμεσα από την Πρόταση 6.9 γ) και για τον λόγο αυτό παραλείπεται. □

Χρησιμοποιώντας την παραπάνω πρόταση και τη στατιστική συνάρτηση Z_1 έχουμε τις ακόλουθες κρίσιμες περιοχές για τους αντίστοιχους ελέγχους:

i) $Z_1 \geq z_a$ για τον έλεγχο της $H_0 : m_{X_1} = m_{X_2}$ έναντι της $H_1 : m_{X_1} > m_{X_2}$.
ii) $Z_1 \leq -z_a$ για τον έλεγχο της $H_0 : m_{X_1} = m_{X_2}$ έναντι της $H_1 : m_{X_1} < m_{X_2}$.
iii) $|Z_1| \geq z_{a/2}$ για τον έλεγχο της $H_0 : m_{X_1} = m_{X_2}$ έναντι της $H_1 : m_{X_1} \neq m_{X_2}$.

Σε περίπτωση που χρησιμοποιηθεί η στατιστική συνάρτηση Z_2 οι κρίσιμες περιοχές είναι i) $Z_2 \leq -z_a$, ii) $Z_2 \geq z_a$ και iii) $|Z_2| \geq z_{a/2}$, αντίστοιχα. Προφανώς, καθώς προσεγγίζουμε διακριτές τυχαίες μεταβλητές από την κανονική κατανομή, μπορεί να γίνει και διόρθωση συνέχειας, σύμφωνα με όσα αναφέρθηκαν στο προηγούμενο κεφάλαιο.

Παράδειγμα 6.8. (Sprent, 1999) Στον πίνακα που ακολουθεί καταγράφεται ο αριθμός των σελίδων 16 τυχαία επιλεγμένων βιβλίων από μία βιβλιοθήκη ενός τμήματος Μαθηματικών (Βιβλιοθήκη 1) και 12 επιλεγμένων βιβλίων από μία βιβλιοθήκη ενός τμήματος Φιλολογίας (Βιβλιοθήκη 2). Να ελέγξετε, με επίπεδο σημαντικότητας 5% και κάνοντας τις κατάλληλες υποθέσεις, την υπόθεση ότι ο μέσος αριθμός των σελίδων των βιβλίων ενός τμήματος Φιλολογίας και ενός τμήματος Μαθηματικών δεν διαφέρουν στατιστικά σημαντικά.

Βιβλιοθήκη 1	29	39	60	78	82	112	125	170
	192	224	263	275	276	286	369	756
Βιβλιοθήκη 2	126	142	156	228	245	246	370	419
	433	454	478	503				

Λύση Παραδείγματος 6.8. Έχουμε δύο πληθυσμούς, με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1,2$. Επιπλέον, λαμβάνουμε δύο το πλήθος, ανεξάρτητα μεταξύ τους τυχαία δείγματα, από καθέναν από αυτούς τους δύο πληθυσμούς, μεγέθους n_i, με $n_1 = 16$, $n_2 = 12$ και $n = n_1 + n_2 = 28$. Αν επιπλέον, υποθέσουμε ότι τα σχήματα των κατανομών είναι όμοια (identical) και οι πληθυσμοί είναι συμμετρικοί, τότε τα αποτελέσματα του έλεγχου της μηδενικής υπόθεσης $H_0 : m_1 = m_2$, δηλαδή της ισότητας των πληθυσμιακών διαμέσων m_i, $i = 1,2$, έναντι της $H_1 : m_1 \neq m_2$, γενικεύονται για τις πληθυσμιακές μέσες τιμές. Για να κάνουμε τον παραπάνω έλεγχο, εργαζόμαστε ως εξής: Αρχικά, τα δεδομένα αναμειγνύονται και διατάσσονται κατά αύξουσα τάξη μεγέθους. Στη συνέχεια, υπολογίζονται οι τάξεις αυτών, όπως φαίνονται στον πίνακα που ακολουθεί. Στον πίνακα αυτό, για διευκόλυνση στην κατανόηση της μεθόδου, οι δειγματικές τιμές του 2ου δείγματος και οι αντίστοιχες τάξεις τους έχουν υπογραμμιστεί.

Παρατήρηση	29	39	60	78	82	112	125
Τάξη	1	2	3	4	5	6	7
Παρατήρηση	126	142	156	170	192	224	228
Τάξη	8	9	10	11	12	13	14
Παρατήρηση	245	246	263	275	276	286	369
Τάξη	15	16	17	18	19	20	21
Παρατήρηση	370	419	433	454	478	503	756
Τάξη	22	23	24	25	26	27	28

Υπενθυμίζεται ότι η στατιστική συνάρτηση του ελέγχου είναι η:

$$U = \min\{U_1, U_2\} = \min\{U_1, n_1 n_2 - U_1\},$$

όπου $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$ και $U_2 = n_1 n_2 - U_1$, με το R_i να είναι το άθροισμα των τάξεων του i-οστού δείγματος, $i = 1,2$, στο σύνολο των διαθέσιμων δειγματικών τιμών. Άμεσα, λοιπόν, έπεται ότι

$$R_1 = 1 + 2 + 3 + 4 + 5 + 6 + 7 + 11 + 12 + 13 + 17 + 18 + 19 + 20 + 21 + 28 = 187,$$

οπότε

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} = 187 - \frac{16 \cdot 17}{2} = 51$$

και

$$U_2 = n_1 n_2 - U_1 = 16 \cdot 12 - 51 = 141.$$

Από τον Πίνακα Π.22 του Παραρτήματος έχουμε ότι, για τη συγκεκριμένη περίπτωση, η κρίσιμη τιμή του ελέγχου είναι 53 και, καθώς, η τιμή του στατιστικού $U = \min\{U_1, U_2\}$ είναι μικρότερη από την κρίσιμη τιμή, απορρίπτεται η μηδενική υπόθεση.

Εναλλακτικά, αφού $n_1, n_2 > 10$, μπορούσε να χρησιμοποιηθεί η στατιστική συνάρτηση που προκύπτει από την προσέγγιση της κατανομής της στατιστικής συνάρτησης R_1. Ειδικότερα, κάνοντας και διόρθωση συνέχειας, είναι:

$$Z = \frac{187 - 0.5 - \frac{16 \cdot 17}{2}}{\sqrt{16 \cdot 12 \cdot 28/12}} = \frac{50.5}{\sqrt{448}} = \frac{50.5}{21.17} = 2.39$$

$(R(X_{11}), R(X_{12}))$	R_1
(1,2)	3
(1,3.5)	4.5
(1,3.5)	4.5
(1,5)	6
(2,3.5)	5.5
(2,3.5)	5.5
(2,5)	7
(3.5,3.5)	7
(3.5,5)	8.5
(3.5,5)	8.5

Πίνακας 6.4: Υπολογισμοί για τον προσδιορισμό της κατανομής, υπό τη μηδενική υπόθεση, της στατιστικής συνάρτησης R_1, υπό τη μηδενική υπόθεση για $n_1 = 4$ και $n_2 = 3$ και υποθέτοντας ότι υπάρχει ακριβώς ένας δεσμός μεταξύ της 3ης και 4ης παρατήρησης.

με κρίσιμη περιοχή $|Z| \geq z_{a/2} = z_{0.025} = 1.96$. Επομένως, και πάλι η μηδενική υπόθεση απορρίπτεται. Αυτό σημαίνει ότι υπάρχει στατιστικά σημαντική διαφορά στις διαμέσους των κατανομών των δύο πληθυσμών. Δηλαδή η διάμεσος της κατανομής του αριθμού των σελίδων των βιβλίων στη βιβλιοθήκη του τμήματος Μαθηματικών είναι διαφορετική , σε ε.σ. 5%, από τη διάμεσο της κατανομής του αριθμού των σελίδων στη βιβλιοθήκη του τμήματος Φιλολογίας. Υπό την υπόθεση της συμμετρίας το συμπέρασμα γενικεύεται για τις πληθυσμιακές μέσες τιμές. □

Σε όσα προαναφέρθηκαν είχε υποτεθεί η μη ύπαρξη δεσμών μεταξύ των δεδομένων. Σε συνέχεια του Παραδείγματος 6.5 θα εξετάσουμε αρχικά αν υπάρχει διαφοροποίηση στην ακριβή κατανομή υπό τη μηδενική υπόθεση της στατιστικής συνάρτησης R_1 όταν υπάρχουν δεσμοί. Ως μέθοδος χειρισμού των ισοβαθμιών χρησιμοποιούμε τα midranks, δηλαδή τον μέσο όρο των τάξεων που θα είχαν οι παρατηρήσεις αν δεν υπήρχαν οι δεσμοί.

Παράδειγμα 6.9. (βλ. Hollander *et al.*, 2014) Αν τα μεγέθη των δύο δειγμάτων είναι $n_1 = 2$ και $n_2 = 3$, αντίστοιχα, να βρεθεί η ακριβής κατανομή της στατιστικής συνάρτησης R_1, υπό τη μηδενική υπόθεση και υποθέτοντας ότι υπάρχει ακριβώς ένας δεσμός μεταξύ της 3ης και 4ης παρατήρησης.

Λύση Παραδείγματος 6.9. Έστω οι δειγματικές τιμές X_{11}, X_{12}, και X_{21}, X_{22}, X_{23}, από τον πρώτο και δεύτερο πληθυσμό, αντίστοιχα. Επιπλέον, έστω $R\left(X_{ij}\right)$, $i = 1,2$, $j = 1,...,n_i$, οι τάξεις των διαθέσιμων δειγματικών τιμών των δύο δειγμάτων στο σύνολο των $n = n_1 + n_2 = 5$ το πλήθος παρατηρήσεων, με $n_1 = 2$ και $n_2 = 3$. Είναι R_1 το άθροισμα των τάξεων του πρώτου δείγματος, δηλαδή είναι $R_1 = \sum_{j=1}^{2} R\left(X_{1j}\right)$. Λαμβάνοντας υπόψη ότι υπάρχει ένας δεσμός μεταξύ της 3ης και 4ης παρατήρησης οι τάξεις στο κοινό διατεταγμένο δείγμα είναι

$$1,\ 2,\ 3.5,\ 3.5,\ 5.$$

Οι δυνατές τιμές της τυχαίας μεταβλητής R_1 καθώς και οι αντίστοιχες τιμές των $R(X_{11})$, $R(X_{12})$ παρατίθενται στον Πίνακα 6.4.

Για παράδειγμα, αν είναι $R(X_{11}) = 3.5$ και $R(X_{12}) = 5$, αυτό σημαίνει ότι η μία από τις δύο τιμές του 1ου δείγματος είναι στη θέση 3 (όπου υπάρχει δεσμός) και η άλλη στη θέση 5 ή η μία είναι στη θέση 4 (όπου υπάρχει δεσμός) και η άλλη στη θέση 5. Άρα, σε αυτήν την περίπτωση, το $R_1 = 3.5 + 5 = 8.5$. Με ανάλογο σκεπτικό προκύπτουν και οι υπόλοιπες τιμές. Επομένως, οι δυνατές τιμές της τυχαίας μεταβλητής R_1 είναι οι $\{3,\ 4.5,\ 5.5,\ 6,\ 7,\ 8.5\}$ και η τυχαία μεταβλητή R_1 έχει την ακόλουθη συνάρτηση πιθανότητας:

$$P(R_1 = r) = \begin{cases} 1/10, \text{ για } r = 3,6, \\ 2/10, \text{ για } r = 4.5,5.5,7,8.5. \end{cases}$$

□

Συγκρίνοντας το αποτέλεσμα του προηγούμενου παραδείγματος που αφορά την ακριβή κατανομή υπό τη μηδενική υπόθεση της στατιστικής συνάρτησης R_1, υποθέτοντας ότι υπάρχουν δεσμοί με το αποτέλεσμα που δόθηκε στο Παράδειγμα 6.5 και αφορούσε, υπό το ίδιο δειγματοληπτικό πλαίσιο, την ακριβή κατανομή της ίδιας στατιστικής συνάρτησης υπό την υπόθεση της μη ύπαρξης δεσμών, συμπεραίνουμε ότι αυτές διαφοροποιούνται. Επομένως, είναι εσφαλμένο για μικρό μέγεθος δείγματος να χρησιμοποιούνται οι κρίσιμες τιμές που δίνονται στο παράρτημα, καθώς αυτές έχουν προκύψει υπό την υπόθεση της μη ύπαρξης δεσμών. Σε περιπτώσεις ύπαρξης δεσμών και μικρού μεγέθους θα πρέπει κάποιος πρώτα να προσδιορίζει την κατανομή της στατιστικής συνάρτησης υπό τη μηδενική υπόθεση με παρόμοιο τρόπο με αυτόν του Παραδείγματος 6.9. Έπειτα, χρησιμοποιώντας αυτήν την κατανομή, θα είναι εφικτό να προσδιορίζει τις κρίσιμες τιμές του ελέγχου ή την p-τιμή.

Το εύλογο ερώτημα που ίσως έχει προκύψει είναι αν τα αποτελέσματα της Πρότασης 6.13 ισχύουν. Η απάντηση είναι όχι και οι απαραίτητες τροποποιήσεις, όταν χρησιμοποιούνται τα midranks δίνονται στην πρόταση που έπεται (για την απόδειξη, βλ. μεταξύ άλλων, Lehmann, 2006).

Πρόταση 6.14. Έστω ότι έχουμε δύο το πλήθος πληθυσμούς, με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1,2$. Επιπλέον, έστω $X_{i1}, \ldots, X_{i,n_i}$, $i = 1,2$, δύο ανεξάρτητα μεταξύ τους, τυχαία δείγματα από καθέναν από αυτούς τους δύο πληθυσμούς, μεγέθους n_i, $i = 1,2$, με $n_1 + n_2 = n$. Επιπλέον, έστω $R\left(X_{ij}\right)$, με $i = 1,2$, και $j = 1,\ldots,n_i$ οι τάξεις των διαθέσιμων δειγματικών τιμών στο σύνολο των $n = n_1 + n_2$ το πλήθος παρατηρήσεων. Επιπρόσθετα, υποθέτουμε ότι οι n αυτές δειγματικές παρατηρήσεις λαμβάνουν c το πλήθος διαφορετικές τιμές και έστω, επίσης, ότι d_1 από αυτές είναι ίσες με τη μικρότερη τιμή, d_2 με την αμέσως μεγαλύτερη, ..., d_c από αυτές ίσες με τη μεγαλύτερη, με $d_i \geq 1$ και $\sum_{i=1}^{c} d_i = n$. Υπό την υπόθεση ότι οι δύο πληθυσμοί ταυτίζονται ισχύουν τα ακόλουθα:

α) $\mathrm{E}(R_i) = n_i \frac{n+1}{2}$, $i = 1,2$.

β) $\mathrm{Var}(R_i) = n_1 n_2 \frac{(n+1)}{12} - \frac{n_1 n_2 \sum_{l=1}^{c}\left(d_l^3 - d_l\right)}{12n(n-1)}$, $i = 1,2$.

γ) Για μεγάλα σε μέγεθος δείγματα

$$\frac{R_i - \frac{n_i(n+1)}{2}}{\sqrt{n_1 n_2 \frac{(n+1)}{12} - \frac{n_1 n_2 \sum_{l=1}^{c}\left(d_l^3 - d_l\right)}{12n(n-1)}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Απόδειξη Πρότασης 6.14. Η απόδειξη θα γίνει χωρίς βλάβη της γενικότητας για τη στατιστική συνάρτηση R_1.

Αν $u_1, u_2, \ldots, u_n$ είναι οι διαθέσιμες τάξεις, τότε από τον ορισμό αυτών προκύπτει άμεσα ότι:

$$u_1 = \ldots = u_{d_1} = \frac{1 + \ldots + d_1}{d_1} = \frac{d_1 + 1}{2},$$

$$u_{d_1+1} = \ldots = u_{d_1+d_2} = \frac{(d_1 + 1) + \ldots + (d_1 + d_2)}{d_2} = d_1 + \frac{d_2 + 1}{2},$$

$$u_{d_1+d_2+1} = \ldots = u_{d_1+d_2+d_3} = \frac{(d_1 + d_2 + 1) + \ldots + (d_1 + d_2 + d_3)}{d_3} = d_1 + d_2 + \frac{d_3 + 1}{2},$$

και συνεχίζουμε παρόμοια και για τις υπόλοιπες τάξεις.

Από αυτές τις $u_1, u_2, ..., u_n$, που μπορούμε να τις θεωρήσουμε ως τον πληθυσμό μας, επιλέγονται με απλή τυχαία δειγματοληψία οι n_1 το πλήθος που αντιστοιχούν στις τάξεις του πρώτου δείγματος. Έστω $V_1, ..., V_{n_1}$ οι τυχαίες μεταβλητές που παριστάνουν αυτές τις τάξεις, καθεμία εκ των οποίων έχει την ίδια κατανομή και $R_1 = V_1 + ... + V_{n_1}$.

α) Είναι τότε:

$$\mathsf{E}(R_1) = n_1 \mathsf{E}(V_1) = n_1 \frac{d_1 u_{d_1} + d_2 u_{d_2} + ... + d_c u_{d_c}}{n} = n_1 \frac{n(n+1)}{2n} = n_1 \frac{(n+1)}{n}.$$

Στην παραπάνω σχέση χρησιμοποιήθηκε ότι το άθροισμα των τάξεων με την παρουσία δεσμών είναι ίσο με το άθροισμα των τάξεων αν δεν υπήρχαν δεσμοί, καθώς χρησιμοποιείται η μέθοδος των midranks.

β) Είναι

$$\mathsf{Var}(R_1) = \mathsf{Var}\left(V_1 + ... + V_{n_1}\right) = \sum_{i=1}^{n_1} \mathsf{Var}(V_i) + \sum_{i=1}^{n_1} \sum_{j=1, j\neq i}^{n_1} \mathsf{Cov}(V_i, V_j).$$

Όμως

$$\mathsf{Var}(V_i) = \frac{1}{n} \sum_{i=1}^{n} (u_i - \overline{u})^2 = \frac{1}{n} \sum_{i=1}^{n} u_i^2 - \overline{u}^2,$$

όπου

$$\overline{u} = \frac{d_1 u_{d_1} + d_2 u_{d_2} + ... + d_c u_{d_c}}{n} = \frac{n+1}{2}.$$

Επιπλέον (βλ. και απόδειξη Πρότασης 6.8):

$$1^2 + ... + n^2 = \sum_{i=1}^{n} u_i^2 + \sum_{l=1}^{c} \frac{d_l \left(d_l^2 - 1\right)}{12}.$$

Άρα

$$\mathsf{Var}(V_i) = \frac{1}{n} \frac{n(n+1)(2n+1)}{6} - \frac{1}{n} \sum_{l=1}^{c} \frac{d_l \left(d_l^2 - 1\right)}{12} - \left(\frac{n+1}{2}\right)^2.$$

Επομένως, λαμβάνοντας επιπλέον υπόψη ότι η συνδιακύμανση $\mathsf{Cov}\left(V_i, V_j\right)$ είναι ίδια για καθένα από τα $n_1(n_1 - 1)$ το πλήθος ζεύγη των i, j, με $i, j = 1, ..., n, i \neq j$, έχουμε ότι:

$$\mathsf{Var}(R_1) = \sum_{i=1}^{n_1} \mathsf{Var}(V_i) + \sum_{i=1}^{n_1} \sum_{j=1, j\neq i}^{n_1} \mathsf{Cov}(V_i, V_j) = n_1 \mathsf{Var}(V_i) + n_1 (n_1 - 1) \mathsf{Cov}\left(V_i, V_j\right).$$

Εφαρμόζοντας την παραπάνω σχέση για $n = n_1$ (σε αυτήν την περίπτωση, το άθροισμα των τάξεων του πρώτου δείγματος είναι σταθερό), έχουμε ότι:

$$0 = n \mathsf{Var}(V_i) + n(n-1) \mathsf{Cov}\left(V_i, V_j\right).$$

Άρα $\mathsf{Cov}\left(V_i, V_j\right) = -\frac{\mathsf{Var}(V_i)}{n-1}$. Συνδυάζοντας τα παραπάνω έπεται ότι

$$\mathsf{Var}(R_1) = n_1 \mathsf{Var}(V_i) - n_1 (n_1 - 1) \frac{\mathsf{Var}(V_i)}{n-1} = \frac{n_1 (n - n_1)}{n-1} \mathsf{Var}(V_i) = \frac{n_1 n_2}{n-1} \mathsf{Var}(V_i),$$

και, άρα,

$$\text{Var}(R_i) = n_1 n_2 \frac{(n+1)}{12} - \frac{n_1 n_2 \sum_{l=1}^{c} \left(d_l^3 - d_l\right)}{12n(n-1)}.$$

γ) Προκύπτει με άμεση εφαρμογή του Κ.Ο.Θ. λαμβάνοντας υπόψη τα α) και β). □

Παρατήρηση 6.3. Τα αποτελέσματα της Πρότασης 6.9 μπορούν να προκύψουν από την Πρόταση 6.14 για την ειδική περίπτωση που $d_1 = ... = d_c = 1$, με $c = n$.

6.3.1.2 Ο έλεγχος των Mann-Whitney

Η στατιστική συνάρτηση που προτάθηκε από τους Mann and Whitney (1947) είναι η:

$$\text{MW} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} D_{ij},$$

όπου $D_{ij} = I\left(X_{2j} < X_{1i}\right)$, με $I(\cdot)$ τη συνήθη δείκτρια συνάρτηση, που δίνεται από τη σχέση:

$$I\left(X_{2j} < X_{1i}\right) = \begin{cases} 1, \text{ αν } X_{2j} < X_{1i}, \\ 0, \text{ αλλού.} \end{cases}$$

Για δεδομένο i δεν είναι δύσκολο να διαπιστώσουμε ότι:

$$\sum_{j=1}^{n_2} D_{ij} = D_{i1} + D_{i2} + ... + D_{i,n_2},$$

δηλαδή, είναι ο συνολικός αριθμός των X_{2j}, $j = 1,...,n_2$, που έχουν τιμές μικρότερες του X_{1i}, όπου i είναι δεδομένο και ένα εκ των $1,...,n_1$. Ο αριθμός αυτός, από τον τρόπο ορισμού του, προκύπτει ότι είναι ίσος με την τάξη της δειγματικής τιμής X_{1i} στο δείγμα των $n = n_1 + n_2$ παρατηρήσεων μείον τον αριθμό των X_{1j}, έστω k_i, που είναι μικρότερα από τη συγκεκριμένη παρατήρηση X_{1i}. Δηλαδή k_i είναι η τάξη της X_{1i} στο δείγμα των n_1 το πλήθος παρατηρήσεων από τον πρώτο πληθυσμό. Επομένως, προκύπτει ότι:

$$\text{MW} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} D_{ij} = \sum_{i=1}^{n_1} (R(X_{1i}) - k_i) = \sum_{i=1}^{n_1} R(X_{1i}) - \sum_{i=1}^{n_1} k_i = R_1 - \frac{n_1(n_1+1)}{2},$$

καθώς το άθροισμα των τάξεων $1 + ... + n_1 = \sum_{i=1}^{n_1} k_i = \frac{n_1(n_1+1)}{2}$. Επομένως, $\text{MW} = U_1$, δηλαδή η στατιστική συνάρτηση MW δεν είναι τίποτε άλλο από τη στατιστική συνάρτηση U_1, η οποία προτάθηκε από τον Wilcoxon.

Παρατήρηση 6.4. Για ιστορικούς λόγους αξίζει να αναφερθεί ότι η στατιστική συνάρτηση των Mann-Whitney προτάθηκε αρχικά για τον έλεγχο της υπόθεσης $H_0 : F_{X_1}(x) = F_{X_2}(x), \forall\, x \in \mathbb{R}$, έναντι μίας εκ των εναλλακτικών $H_1 : F_{X_1}(x) \leqslant F_{X_2}(x)$ ή της $H_1 : F_{X_1}(x) \geqslant F_{X_2}(x)$, με γνήσια ανισότητα για κάποιο $x \in R$ ή της $H_1 : F_{X_1}(x) \neq F_{X_2}(x)$, για κάποιο $x \in \mathbb{R}$.

6.3.2 Εξαρτημένα δείγματα

Έστω ένα τυχαίο δείγμα $X_1, \ldots, X_n$, μεγέθους n από έναν πληθυσμό με διάμεσο m_X. Επιπλέον, έστω ένα τυχαίο δείγμα $Y_1, \ldots, Y_n$, μεγέθους n από έναν πληθυσμό με διάμεσο m_Y. Επιπρόσθετα, υποθέτουμε ότι τα δύο δείγματα είναι εξαρτημένα.

Ενδιαφερόμαστε για τους ελέγχους, σε επίπεδο σημαντικότητας a, των παρακάτω υποθέσεων:

i) $H_0 : m_X = m_Y$ έναντι της εναλλακτικής $H_1 : m_X > m_Y$.
ii) $H_0 : m_X = m_Y$ έναντι της εναλλακτικής $H_1 : m_X < m_Y$.
iii) $H_0 : m_X = m_Y$ έναντι της εναλλακτικής $H_1 : m_X \neq m_Y$.

Θεωρώντας το τυχαίο δείγμα των διαφορών $D_i = X_i - Y_i$, $i = 1,...,n$, οι παραπάνω έλεγχοι ανάγονται στους ελέγχους:

i) $H_0 : m_D = 0$ έναντι της εναλλακτικής $H_1 : m_D > 0$.
ii) $H_0 : m_D = 0$ έναντι της εναλλακτικής $H_1 : m_D < 0$.
iii) $H_0 : m_D = 0$ έναντι της εναλλακτικής $H_1 : m_D \neq 0$,

όπου m_D είναι η διάμεσος του πληθυσμού που περιγράφει τη διαφορά $X - Y$. Για τον έλεγχο αυτόν μπορεί να εφαρμοστούν οι μεθοδολογίες που παρουσιάστηκαν στην Ενότητα 6.2 για $m_0 = 0$.

Παρατήρηση 6.5. Τα αποτελέσματα αυτού του ελέγχου γενικεύονται από την πληθυσμιακή διάμεσο στις πληθυσμιακές μέσες τιμές, όταν οι δειγματικές τιμές D_i, $i = 1,...,n$, προέρχονται από έναν συμμετρικό πληθυσμό, δηλαδή όταν $m_D = \mu_D$, όπου μ_D η μέση τιμή της κατανομής της διαφοράς $X - Y$. Παρατηρήστε ότι, αν πράγματι ισχύει η υπόθεση της συμμετρίας, αναμένουμε η δειγματική διάμεσος να είναι περίπου ίση με τη δειγματική μέση τιμή.

Παράδειγμα 6.10. Ένας γιατρός καταγράφει το βάρος 11 ασθενών πριν και μετά από μια τρίμηνη δίαιτα. Τα αποτελέσματα είναι τα ακόλουθα:

Ασθενής	1	2	3	4	5	6	7	8	9	10	11
Πριν	89	84	94	91	90	81	84	89	109	121	78
Μετά	82	79	82	94	95	79	70	71	90	100	79

Με επίπεδο σημαντικότητας 5% ελέγξτε αν η δίαιτα επιφέρει στατιστικά σημαντική διαφοροποίηση στο βάρος.

Λύση Παραδείγματος 6.10. Έστω X και Y οι τυχαίες μεταβλητές που παριστάνουν το βάρος πριν και μετά τη δίαιτα. Έχουμε δύο τυχαία δείγματα $X_1,\dots,X_{11}$ από τον πληθυσμό με διάμεσο m_X και $Y_1,\dots,Y_{11}$, από τον πληθυσμό με διάμεσο m_Y. Επιπρόσθετα, τα δύο δείγματα είναι εξαρτημένα, καθώς οι μετρήσεις γίνονται στις ίδιες πειραματικές μονάδες (ίδιοι ασθενείς). Δημιουργούμε τις διαφορές $D_i = X_i - Y_i$, $i = 1,...,11$, οι οποίες καταγράφονται στον πίνακα που ακολουθεί:

Διαφορά	7	5	12	-3	-5	2	14	18	19	21	-1

Κατά αυτόν τον τρόπο, έχουμε ένα τυχαίο δείγμα $D_1,\dots,D_{11}$, από έναν πληθυσμό με συνεχή αθροιστική συνάρτηση κατανομής F. Επίσης, από τα δεδομένα του δείγματος έχουμε ότι η δειγματική διάμεσος της απώλειας βάρους ισούται με 7, ενώ η δειγματική μέση τιμή της με 8.09. Ενδιαφερόμαστε να ελέγξουμε την υπόθεση ότι η διάμεσος m_D της άγνωστης κατανομής της διαφοράς $X - Y$ είναι ίση με $m_0 = 0$. Επομένως, θέλουμε να ελέγξουμε τη μηδενική υπόθεση $H_0 : m_D = 0$ έναντι της εναλλακτικής $H_1 : m_D \neq 0$.

Οι απόλυτες τιμές των διαφορών $|D_1|,\dots,|D_{11}|$, διατάσσονται κατά αύξουσα τάξη και υπολογίζονται οι τάξεις τους, έστω $R(|D_i|)$, $i = 1,...,11$, όπως φαίνεται στον πίνακα που ακολουθεί, όπου για ευκολία στους περαιτέρω υπολογισμούς έχουμε υπογραμμίσει τις αρνητικές διαφορές και, επιπλέον, έχουμε παρατηρήσει ότι υπάρχει ένας δεσμός δύο παρατηρήσεων.

$\lvert D_i\rvert$:	1	2	3	5	5	7	12	14	18	19	21
$R(\lvert D_i\rvert)$:	1	2	3	4.5	4.5	6	7	8	9	10	11

Το άθροισμα των τάξεων που αντιστοιχούν στις αρνητικές διαφορές είναι $T^- = 1 + 3 + 4.5 = 8.5$, ενώ το άθροισμα των τάξεων που αντιστοιχούν στις θετικές διαφορές, έστω T^+, είναι $T^+ = \frac{n(n+1)}{2} - T^- = \frac{11 \cdot 12}{2} - 8.5 = 57.5$, οπότε $T = \min\{T^+, T^-\} = T^- = 8.5$.

Λαμβάνοντας υπόψη την ύπαρξη ενός δεσμού δύο παρατηρήσεων, οπότε $c = 1$ και $d_1 = 2$, προκύπτει ότι χρησιμοποιείται για τον έλεγχο της υπό μελέτη μηδενικής υπόθεσης η στατιστιστική συνάρτηση:

$$W = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum_{l=1}^{c}\left(d_l^3 - d_l\right)}{48}}}.$$

Λαμβάνοντας υπόψη ότι υπό τη μηδενική υπόθεση η στατιστική συνάρτηση W ακολουθεί (ασυμπτωτικά) την κατανομή $\mathcal{N}(0,1)$, έπεται ότι απορρίπτεται η μηδενική υπόθεση $H_0 : m_D = 0$ έναντι της $H_1 : m_D \neq 0$, αν $|W| \geq z_{a/2} = z_{0.025} = 1.96$. Καθώς:

$$W = \frac{57.5 + 0.5 - \frac{11 \cdot 12}{4}}{\sqrt{\frac{11 \cdot 12 \cdot 23}{24} - \frac{\left(2^3 - 2\right)}{48}}} = \frac{57.5 - 33}{\sqrt{126.5 - 0.125}} = \frac{24.5}{11.24} = 2.18$$

Αν, επιπρόσθετα, χρησιμοποιούσαμε διόρθωση συνέχειας, η τιμή της στατιστικής συνάρτησης W θα ήταν ίση με $W = \frac{24}{11.24} = 2.14$ και, αφού $2.14 > 1.96$, θα απορριπτόταν και πάλι η μηδενική υπόθεση. Άρα, υπάρχει στατιστικά σημαντική διαφοροποίηση στο βάρος πριν και μετά την τρίμηνη δίαιτα σε επίπεδο σημαντικότητας 5%. □

6.4 Έλεγχοι ισότητας περισσότερων των δύο πληθυσμιακών διαμέσων

Στην ενότητα αυτή, αρχικά, θα παρουσιαστεί μια επέκταση του ελέγχου Wilcoxon-Mann-Whitney για τον έλεγχο της υπόθεσης ότι τρία ή περισσότερα τυχαία δείγματα, τα οποία είναι ανεξάρτητα μεταξύ τους, προέρχονται από τον ίδιο πληθυσμό, έναντι της εναλλακτικής υπόθεσης ότι τουλάχιστον δύο από τα δείγματα προέρχονται από πληθυσμούς που διαφέρουν ως προς τις διαμέσους. Σημειώνουμε ότι, καθώς ο έλεγχος που θα παρουσιαστεί, προϋποθέτει στην ουσία ισότητα των πληθυσμιακών διακυμάνσεων, στην επόμενη ενότητα θα ασχοληθούμε με τους ελέγχους της υπόθεσης της ισότητας των πληθυσμιακών διακυμάνσεων. Τέλος, παρουσιάζεται ο μη παραμετρικός έλεγχος όταν τα δείγματα είναι εξαρτημένα.

6.4.1 Ανεξάρτητα δείγματα: ο έλεγχος των Kruskal-Wallis

Το στατιστικό τεστ των Kruskal and Wallis (1952) (βλ. επίσης Kruskal and Wallis, 1953) αποτελεί μία λογική επέκταση του ελέγχου των Wilcoxon-Mann-Whitney. Είναι ένας μη παραμετρικός τρόπος ελέγχου της υπόθεσης ότι τρία ή περισσότερα ανεξάρτητα τυχαία δείγματα προέρχονται από τον ίδιο πληθυσμό, έναντι της εναλλακτικής υπόθεσης ότι τουλάχιστον δύο από τα δείγματα προέρχονται από πληθυσμούς που διαφέρουν ως προς τις διαμέσους. Από τη μορφή της εναλλακτικής υπόθεσης γίνεται άμεσα αντιληπτό ότι το στατιστικό τεστ των Kruskal-Wallis προϋποθέτει στην ουσία ισότητα των πληθυσμιακών διακυμάνσεων. Επιπλέον, υποθέτουμε ότι τα δεδομένα είναι τουλάχιστον διατάξιμα.

Έστω ότι έχουμε k το πλήθος πληθυσμούς, με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1, ..., k$, $k \geqslant 3$. Επιπλέον, λαμβάνουμε k το πλήθος, ανεξάρτητα μεταξύ τους τυχαία δείγματα, ένα από καθέναν από αυτούς τους $i = 1, ..., k$, πληθυσμούς, μεγέθους n_i, $i = 1, ..., k$, με $n_1 + ... + n_k = n$. Έστω $X_{i1}, ..., X_{i,n_i}$ οι δειγματικές τιμές από τον i-οστό πληθυσμό, $i = 1, ..., k$. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση: $H_0 : F_1(x) = ... = F_k(x)$, για κάθε $x \in \mathbb{R}$, όπου $k \geqslant 3$, έναντι της εναλλακτικής υπόθεσης H_1 : ότι υπάρχει τουλάχιστον ένα

ζεύγος i,j με $i \neq j$, $i,j = 1,...,k$, $k \geqslant 3$, τέτοιο ώστε $m_i \neq m_j$, όπου m_l η διάμεσος του l-οστού πληθυσμού, $l = 1,...,k$.

Έστω $R\left(X_{ij}\right)$ $i = 1,...,k, j = 1,...,n_i, k \geqslant 3$, οι τάξεις των διαθέσιμων δειγματικών τιμών των k δειγμάτων στο σύνολο των n παρατηρήσεων. Θα συμβολίζουμε, σε όσα ακολουθούν, με R_i το άθροισμα των τάξεων του i-οστού δείγματος, $i = 1,...,k$, $k \geqslant 3$, δηλαδή $R_i = \sum_{j=1}^{n_i} R\left(X_{ij}\right)$, $i = 1,...,k$, $k \geq 3$. Τότε, εύκολα προκύπτει ότι $\sum_{i=1}^{k} R_i = \frac{n(n+1)}{2}$.

Αν υποθέσουμε ότι η H_0 είναι αληθής, αναμένουμε οι μέσοι όροι των τάξεων σε καθένα από τα i δείγματα, $i = 1,...,k$, $k \geqslant 3$, να είναι περίπου ίσοι μεταξύ τους. Δηλαδή περιμένουμε να ισχύει ότι:

$$\frac{R_1}{n_1} = \frac{R_2}{n_2} = ... = \frac{R_k}{n_k}, k \geqslant 3.$$

Λαμβάνοντας υπόψη ότι $\sum_{i=1}^{k} R_i = \frac{n(n+1)}{2}$, έχουμε ότι:

$$R_1 + \frac{n_2}{n_1}R_1 + ... + \frac{n_k}{n_1}R_1 = \frac{n(n+1)}{2},$$

οπότε:

$$\frac{R_1}{n_1}(n_1 + n_2 + ... + n_k) = \frac{n(n+1)}{2} \Rightarrow \frac{R_1}{n_1} = \frac{R_2}{n_2} = ... = \frac{R_k}{n_k} = \frac{n+1}{2}.$$

Επομένως, ένας πρακτικός, αλλά όχι στατιστικός τρόπος, για να αποφανθούμε για την αποδοχή ή απόρριψη της μηδενικής υπόθεσης, ότι τα τυχαία δείγματα προέρχονται από τον ίδιο πληθυσμό, είναι να εξετάζουμε αν οι ποσότητες $\frac{R_i}{n_i}$, $i = 1,...,k$, $k \geqslant 3$, είναι περίπου ίσες μεταξύ τους και ίσες με $(n+1)/2$ ή, εναλλακτικά, αν η ποσότητα:

$$\sum_{i=1}^{k}\left(\frac{R_i}{n_i} - \frac{n+1}{2}\right)^2,$$

είναι κοντά στο μηδέν. Γίνεται αντιληπτό ότι μεγάλες τιμές της παραπάνω ποσότητας θα υποδεικνύουν απόκλιση από την υπόθεση ότι τα k το πλήθος τυχαία δείγματα προέρχονται από τον ίδιο πληθυσμό.

Στηριζόμενοι σε αυτήν την ιδέα, οι Kruskal and Wallis (1952) πρότειναν στην περίπτωση μη ύπαρξης δεσμών τη στατιστική συνάρτηση:

$$\text{KW} = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{1}{n_i}\left(R_i - \frac{n_i(n+1)}{2}\right)^2. \tag{6.18}$$

Στην πρόταση που ακολουθεί, δίνονται δύο ισοδύναμες μορφές της στατιστικής συνάρτησης.

Πρόταση 6.15. Στην περίπτωση μη ύπαρξης δεσμών, δύο ισοδύναμες εκφράσεις της στατιστικής συνάρτησης KW που δόθηκε στη σχέση (6.18) είναι οι:

$$\text{KW} = \frac{12}{n(n+1)}\sum_{i=1}^{k} n_i\left(\frac{R_i}{n_i} - \frac{(n+1)}{2}\right)^2, \tag{6.19}$$

και

$$\text{KW} = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(n+1). \tag{6.20}$$

Απόδειξη Πρότασης 6.15. Από τη σχέση

$$\text{KW} = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{1}{n_i} \left(R_i - \frac{n_i(n+1)}{2} \right)^2,$$

βάζοντας τον όρο n_i εντός της παρένθεσης που υψώνεται στο τετράγωνο προκύπτει εύκολα η ισοδύναμη έκφραση:

$$\text{KW} = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{n_i^2}{n_i} \left(\frac{R_i}{n_i} - \frac{n_i(n+1)}{2n_i} \right)^2 = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\frac{R_i}{n_i} - \frac{(n+1)}{2} \right)^2.$$

Επιπρόσθετα, είναι:

$$\begin{aligned} \text{KW} &= \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{1}{n_i} \left(R_i - \frac{n_i(n+1)}{2} \right)^2 \\ &= \frac{12}{n(n+1)} \left\{ \sum_{i=1}^{k} \frac{1}{n_i} \left(R_i^2 - 2\frac{n_i(n+1)}{2} R_i + \frac{n_i^2(n+1)^2}{4} \right) \right\}, \end{aligned}$$

οπότε

$$\begin{aligned} \text{KW} &= \frac{12}{n(n+1)} \left\{ \sum_{i=1}^{k} \frac{R_i^2}{n_i} - \frac{(n+1)^2 n}{2} + \frac{(n+1)^2 n}{4} \right\} \\ &= \frac{12}{n(n+1)} \left\{ \sum_{i=1}^{k} \frac{R_i^2}{n_i} - \frac{(n+1)^2 n}{4} \right\} = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \end{aligned}$$

όπου χρησιμοποιήσαμε το γεγονός ότι $\sum_{i=1}^{k} n_i = n$ και, επιπλέον, ότι $\sum_{i=1}^{k} R_i = \frac{n(n+1)}{2}$. □

Προφανώς, καθώς η στατιστική συνάρτηση των Kruskal and Wallis (1952) δίνεται από τη σχέση (6.19) προκύπτει ότι λαμβάνει την τιμή μηδέν, όταν

$$\frac{R_1}{n_1} = \frac{R_2}{n_2} = \ldots = \frac{R_k}{n_k} = \frac{n+1}{2}.$$

Επομένως, όταν η μηδενική υπόθεση είναι αληθής, η στατιστική συνάρτηση KW λαμβάνει τη μικρότερη δυνατή τιμή. Συνεπώς, απορρίπτεται η μηδενική υπόθεση H_0 για μεγάλες τιμές της στατιστικής συνάρτησης KW, δηλαδή αν $\text{KW} \geqslant c$, όπου c είναι ένας αριθμός τέτοιος, ώστε:

$$P\left(\text{KW} \geqslant c \,\middle|\, H_0 \text{ αληθής} \right) = a.$$

Η εύρεση της ακριβούς κατανομής της στατιστικής συνάρτησης KW υπό τη μηδενική υπόθεση είναι πολύ δύσκολη και έχει επιτευχθεί από τους Kruskal and Wallis (1952) στην περίπτωση των τριών πληθυσμών $k = 3$ και για μικρά σε μέγεθος δείγματα τέτοια, ώστε $n_i \leq 5$, για $i = 1,2,3$. Ο τρόπος σκέψης δίνεται στο παράδειγμα που ακολουθεί.

Παράδειγμα 6.11. Να προσδιορίσετε, υποθέτοντας ότι δεν υπάρχουν δεσμοί στις δειγματικές παρατηρήσεις, την ακριβή κατανομή της στατιστικής συνάρτησης των Kruskal-Wallis, όταν $n_1 = 2$, $n_2 = 1$ και $n_3 = 1$, και να δείξετε ότι $P(\text{KW} \geq 2.7) = 0.5$.

	$R(X_{11})$	$R(X_{12})$	$R(X_{21})$	$R(X_{31})$	R_1	R_2	R_3	KW
1	1	2	3	4	3	3	4	2.7
2	1	2	4	3	3	4	3	2.7
3	1	3	2	4	4	2	4	1.8
4	1	3	4	2	4	4	2	1.8
5	1	4	2	3	5	2	3	0.3
6	1	4	3	2	5	3	2	0.3
7	2	3	1	4	5	1	4	2.7
8	2	3	4	1	5	4	1	2.7
9	2	4	1	3	6	1	3	1.8
10	2	4	3	1	6	3	1	1.8
11	3	4	1	2	7	1	2	2.7
12	3	4	2	1	7	2	1	2.7

Πίνακας 6.5: Υπολογισμοί για τον προσδιορισμό, υπό τη μηδενική υπόθεση, της κατανομής της σσ KW, όταν $n_1 = 2$, $n_2 = 1$ και $n_3 = 1$ (μη ύπαρξη δεσμών).

Λύση Παραδείγματος 6.11. Είναι

$$\text{KW} = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{4 \cdot 5} \sum_{i=1}^{3} \frac{R_i^2}{n_i} - 3 \cdot 5,$$

με R_i το άθροισμα των τάξεων στο σύνολο των παρατηρήσεων καθενός εκ των τριών δειγμάτων.

Οι παρατηρήσεις που είναι διαθέσιμες είναι τέσσερις το πλήθος συνολικά και έστω ότι είναι οι $X_{11}, X_{12}, X_{21}, X_{31}$, όπου ο πρώτος δείκτης αναφέρεται στον πληθυσμό από όπου προέρχεται το δείγμα και ο δεύτερος δείκτης στον αριθμό της παρατήρησης εντός αυτού του δείγματος. Θα πρέπει να υπολογιστούν αρχικά οι δυνατές τιμές των R_1, R_2, R_3, όπου $R_i = \sum_{j=1}^{n_i} R\left(X_{ij}\right)$, $i = 1,2,3$ και $R\left(X_{ij}\right)$ $i = 1,2,3$, $j = 1,\ldots,n_i$, οι τάξεις των διαθέσιμων δειγματικών τιμών των τριών δειγμάτων στο σύνολο των $n = 4$ το πλήθος παρατηρήσεων.

Στον Πίνακα 6.5 (στήλες $R(X_{11})$, $R(X_{12})$, $R(X_{21})$, $R(X_{31})$) δίνουμε τις δυνατές περιπτώσεις για τις τάξεις των τεσσάρων παρατηρήσεων (οι δύο πρώτες θέσεις αφορούν τις δειγματικές παρατηρήσεις από τον πρώτο πληθυσμό, οι επόμενες κατά σειρά αυτές από τον δεύτερο και τον τρίτο πληθυσμό, αντίστοιχα). Επίσης, στις στήλες R_1, R_2, R_3 δίνουμε τις τιμές των αθροισμάτων, ενώ στη στήλη KW δίνουμε την τιμή της στατιστικής συνάρτησης KW για καθεμία από τις δυνατές διαφορετικές τετράδες. Υπενθυμίζουμε ότι δεν μας ενδιαφέρει η διάταξη εντός των δειγμάτων (και για αυτόν τον λόγο π.χ. δεν θεωρήθηκε η (2,1,3,4) ως ξεχωριστή διάταξη). Επομένως, οι δυνατές τιμές της στατιστικής συνάρτησης KW είναι $\{0.3, 1.8, 2.7\}$ και η συνάρτηση πιθανότητας, υπό τη μηδενική υπόθεση, είναι:

$$P(\text{KW} = x) = \begin{cases} 3/6, \text{ για } x = 2.7, \\ 2/6, \text{ για } x = 1.8, \\ 1/6, \text{ για } x = 0.3, \\ 0, \text{ αλλού.} \end{cases}$$

Άρα, $P(\text{KW} \geq 2.7) = 0.5$. □

Καθώς η εύρεση της κατανομής της στατιστικής συνάρτησης KW υπό τη μηδενική υπόθεση έχει επιτευχθεί για συγκεκριμένες περιπτώσεις και μόνο για αυτές υπάρχουν διαθέσιμοι πίνακες για τον υπολογισμό πιθανοτήτων ή για τον προσδιορισμό κρίσιμων τιμών, οι ερευνητές οδηγήθηκαν στην εύρεση ενός

προσεγγιστικού στατιστικού τεστ, στην περίπτωση όπου $n_i \geq 10$. Λαμβάνοντας υπόψη ότι $R_i = \sum_{j=1}^{n_i} R\left(X_{ij}\right)$, $i = 1,...,k$, $k \geqslant 3$, δηλαδή ότι είναι στην ουσία άθροισμα n_i το πλήθος τυχαίων (όχι ανεξάρτητων) μεταβλητών, μπορεί να εφαρμοστεί το Κεντρικό Οριακό Θεώρημα (για ισόνομες τυχαίες μεταβλητές), οπότε υπό τη μηδενική υπόθεση και υποθέτοντας μη ύπαρξη δεσμών στις δειγματικές παρατηρήσεις:

$$\frac{R_i - \mathsf{E}(R_i)}{\sqrt{\mathsf{Var}(R_i)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Επομένως, υπό τη μηδενική υπόθεση,

$$\sum_{i=1}^{k}\left(\frac{R_i - \mathsf{E}(R_i)}{\sqrt{\mathsf{Var}(R_i)}}\right)^2 = \sum_{i=1}^{k}\frac{(R_i - \mathsf{E}(R_i))^2}{\mathsf{Var}(R_i)} \xrightarrow{d} \chi^2_{k-1},$$

όπου οι βαθμοί ελευθερίας της χ^2 κατανομής είναι $k-1$ και όχι k, όπως, ίσως, εσφαλμένα, κάποιος θα ανέμενε. Αυτό δικαιολογείται, καθώς, $k-1$ το πλήθος τάξεις R_i ορίζονται ανεξάρτητα, καθώς ισχύει η ισότητα $\sum_{i=1}^{k} R_i = \frac{n(n+1)}{2}$.

Στην επόμενη πρόταση, προσδιορίζονται οι ποσότητες $\mathsf{E}(R_i)$ και $\mathsf{Var}(R_i)$ για τον προσδιορισμό της παραπάνω στατιστικής συνάρτησης.

Πρόταση 6.16. Υπό τη μηδενική υπόθεση και υποθέτοντας ότι δεν υπάρχουν δεσμοί, αποδεικνύεται ότι:

$$\mathsf{E}(R_i) = n_i\frac{n+1}{2} \text{ και } \mathsf{Var}(R_i) = n_i\frac{(n+1)(n-n_i)}{12}.$$

Απόδειξη Πρότασης 6.16. Η απόδειξη είναι παρόμοια με την απόδειξη των α) και β) της Πρότασης 6.9 και για αυτό αφήνεται ως άσκηση για τον/την αναγνώστη/στρια. □

Παρατήρηση 6.6. Τα παραπάνω, ουσιαστικά, αποδεικνύουν ότι, αν X είναι η τ.μ. που παριστάνει το άθροισμα n_i ακέραιων που εκλέγονται στην τύχη χωρίς επανάθεση από τους n το πλήθος πρώτους ακέραιους αριθμούς, δηλαδή από το σύνολο $\{1,...,n\}$, τότε $\mathsf{E}(X) = \frac{n_i(n+1)}{2}$ και $\mathsf{Var}(X) = \frac{n_i(n+1)(n-n_i)}{12}$.

Επομένως, από την Πρόταση 6.16 και τη συζήτηση που προηγήθηκε αυτής προκύπτει ότι υπό τη μηδενική υπόθεση:

$$\sum_{i=1}^{k}\left(\frac{R_i - \mathsf{E}(R_i)}{\sqrt{\mathsf{Var}(R_i)}}\right)^2 = \sum_{i=1}^{k}\frac{\left(R_i - n_i\frac{n+1}{2}\right)^2}{\frac{n_i(n+1)(n-n_i)}{12}} = \frac{12}{(n+1)}\sum_{i=1}^{k}\frac{\left(R_i - n_i\frac{n+1}{2}\right)^2}{n_i(n-n_i)} \xrightarrow{d} \chi k - 1^2.$$

Επιπλέον, στην περίπτωση μη ύπαρξης δεσμών, οι Kruskal and Wallis (1952) απέδειξαν ότι η στατιστική συνάρτηση

$$\mathsf{KW} = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{1}{n_i}\left(R_i - \frac{n_i(n+1)}{2}\right)^2$$

ακολουθεί, υπό τη μηδενική υπόθεση, και αυτή προσεγγιστικά χ^2 κατανομή με $k-1$ βαθμούς ελευθερίας. Επομένως, η υπό μελέτη μηδενική υπόθεση απορρίπτεται, αν $\mathsf{KW} \geqslant \chi^2_{k-1,a}$, όπου $\chi^2_{k-1,a}$ είναι το σημείο εκείνο για το οποίο ισχύει ότι $P\left(\chi^2_{k-1} \geqslant \chi^2_{k-1,a}\right) = a$.

Επισημαίνεται ότι τόσο οι Kruskal and Wallis (1952) όσο και οι Gabriel and Lachebruch (1969) απέδειξαν ότι η παραπάνω προσέγγιση είναι ικανοποιητική ακόμα και για μικρά σε μέγεθος δείγματα.

Παρατήρηση 6.7. Συνήθως η διαπίστωση ότι απορρίπτεται η

$$H_0 : F_1(x) = ... = F_k(x), \text{ για κάθε } x \in \mathbb{R},\ k \geqslant 3,$$

έναντι της εναλλακτικής υπόθεσης

$$H_1 : \text{ υπάρχει τουλάχιστον ένα ζεύγος } (i,j) \text{ με } i \neq j,\ i,j = 1,...,k,\ k \geqslant 3, \text{ τέτοιο, ώστε } m_i \neq m_j,$$

όπου m_l η διάμεσος του l-οστού πληθυσμού, $l = 1,...,k$, δεν αποτελεί τον τελικό σκοπό σε μία στατιστική μελέτη, καθώς το ενδιαφέρον μας επικεντρώνεται στον εντοπισμό των δειγμάτων που προέρχονται από διαφορετικούς πληθυσμούς. Κατά ανάλογο τρόπο, όπως στην κλασική παραμετρική στατιστική, για να δοθεί απάντηση στο παραπάνω ερώτημα, προβαίνουμε στις λεγόμενες πολλαπλές συγκρίσεις (multiple comparisons) . Αξίζει να σημειώσουμε ότι η ονομασία αυτή προέκυψε από το γεγονός ότι στην ουσία έχουμε να κάνουμε $\binom{k}{2}$ το πλήθος στατιστικούς ελέγχους. Σε αυτήν την περίπτωση, οι πολλαπλές συγκρίσεις δεν είναι τίποτε άλλο παρά ο έλεγχος ότι τα δείγματα από τον m-οστό και l-οστό πληθυσμό μπορούμε να θεωρήσουμε ότι προέρχονται από τον ίδιο πληθυσμό έναντι της εναλλακτικής ότι οι πληθυσμιακοί διάμεσοι αυτών διαφέρουν. Για τον σκοπό αυτό χρησιμοποιείται η στατιστική συνάρτηση (βλ. Kvam and Vidakovic, 2007, και τις εκεί αναφορές):

$$t = \frac{\frac{R_m}{n_m} - \frac{R_l}{n_l}}{\sqrt{\frac{S^2(n-1-H^*)}{n-k}\left(\frac{1}{n_l} + \frac{1}{n_m}\right)}},$$

όπου

$$H^* = \frac{1}{S^2}\left\{\sum_{i=1}^{k}\frac{R_i^2}{n_i} - \frac{n(n+1)^2}{4}\right\},$$

και

$$S^2 = (n-1)^{-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i} R\left(X_{ij}\right)^2 - \frac{n(n+1)^2}{4}\right].$$

Η κρίσιμη περιοχή του ελέγχου είναι $|t| \geq t_{n-k,a/2}$, όπου $t_{n-k,a/2}$ τέτοιο, ώστε $P\left(t_{n-k} \geqslant t_{n-k,a/2}\right) = a/2$.

Ένας άλλος τρόπος αντιμετώπισης, σύμφωνα με την εργασία της Dunn (1964), είναι να θεωρήσουμε τη στατιστική συνάρτηση:

$$Z_{ml} = \frac{|\bar{R}_m - \bar{R}_l|}{\sqrt{\frac{n(n+1)}{12}(\frac{1}{n_i} + \frac{1}{n_j})}}\ ,m \neq l\ \ m,\ l = 1,2,\ldots,k.$$

Αν $Z_{ml} > z^* = z_{a/(k(k-1))}$, τότε υπάρχει στατιστικά σημαντική διαφοροποίηση μεταξύ των διαμέσων του m-οστού και l-οστού πληθυσμού. Η ποσότητα a ονομάζεται ολικό επίπεδο σημαντικότητας και εκφράζει την πιθανότητα να κάνουμε τουλάχιστον μία εσφαλμένη απόρριψη ανάμεσα στις $k(k-1)/2$ συγκρίσεις, δηλαδή $1-a$ είναι η πιθανότητα να λάβουμε σε όλους τους ελέγχους ορθή απόφαση. Συνήθως, η τιμή του ολικού επιπέδου που χρησιμοποιείται στις εφαρμογές είναι $a = 0.2$.

Μία επιπρόσθετη εναλλακτική προσέγγιση είναι η διενέργεια των ανά δύο συγκρίσεων με τον έλεγχο των Wilcoxon-Mann-Whitney και με επίπεδο σημαντικότητας για κάθε έλεγχο αυτό που προκύπτει με εφαρμογή της διόρθωσης Bonferonni, ήτοι $2a/(k(k-1))$.

Στο παράδειγμα που ακολουθεί, αποσαφηνίζεται, μέσω ενός αριθμητικού παραδείγματος, ο έλεγχος που προτάθηκε από τους Kruskal-Wallis.

Παράδειγμα 6.12. (Kruskal and Wallis, 1952) Σε ένα εργοστάσιο τρεις μηχανές χρησιμοποιούνται για την παραγωγή δοχείων εμφιαλώσεως. Κατά τη διάρκεια μιας εργάσιμης εβδομάδας καταγράφεται ο αριθμός των δοχείων που κατασκευάστηκαν από κάθε μηχανή και τα αποτελέσματα παρατίθενται στον πίνακα που ακολουθεί, με την επιπλέον επισήμανση ότι κάποιες μέρες δεν παρήχθησαν δοχεία από κάποιες μηχανές λόγω μη λειτουργίας των αντίστοιχων μηχανών.

Μηχανή 1	340	345	330	342	338
Μηχανή 2	339	333	344		
Μηχανή 3	347	343	349	355	

Να ελέγξετε αν υπάρχει στατιστικά σημαντική διαφορά ως προς την παραγωγή δοχείων των τριών μηχανών. Δίνεται ότι $P(\text{KW} \geqslant 5.6564) = 0.049$.

Λύση Παραδείγματος 6.12. Έχουμε διαθέσιμα τρία ανεξάρτητα δείγματα, μεγέθους δείγματος $n_1 = 5$, $n_2 = 3$, $n_3 = 4$, αντίστοιχα, με $n = n_1 + n_2 + n_3 = 12$. Αναμειγνύονται οι δειγματικές παρατηρήσεις X_{ij}, $j = 1,...,n_i$, $i = 1,2,3$, και, στη συνέχεια, διατάσσονται σε αύξουσα τάξη μεγέθους και υπολογίζονται οι τάξεις $R\left(X_{ij}\right)$, $j = 1,...,n_i$, $i = 1,2,3$, κάθε δειγματικής τιμής.

Για ευκολία στους μετέπειτα υπολογισμούς δημιουργούμε τον ακόλουθο πίνακα, όπου στην 3η γραμμή δίνεται το δείγμα από το οποίο προέρχεται η κάθε παρατήρηση:

Τιμή παρατήρησης	330	333	338	339	340	342	343	344	345	347	349	355
Τάξη	1	2	3	4	5	6	7	8	9	10	11	12
Δείγμα	1	2	1	2	1	1	3	2	1	3	3	3

Επομένως, είναι:

$$R_1 = \sum_{j=1}^{5} R\left(X_{1j}\right) = 5 + 9 + 1 + 6 + 3 = 24,$$

$$R_2 = \sum_{j=1}^{3} R\left(X_{2j}\right) = 4 + 2 + 8 = 14,$$

και

$$R_3 = \sum_{j=1}^{4} R\left(X_{3j}\right) = 10 + 7 + 11 + 12 = 40.$$

Άρα

$$\begin{aligned} \text{KW} &= \frac{12}{n(n+1)} \sum_{i=1}^{3} \frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{12(12+1)} \left(\frac{24^2}{5} + \frac{14^2}{3} + \frac{40^2}{4} \right) - 3(12+1) \\ &= \frac{1}{13}(576/5 + 196/3 + 1600/4) - 3 \cdot 13 = \frac{1}{13}(115.2 + 65.333 + 400) - 39 \\ &= 580.533/13 - 39 = 44.6564 - 39 = 5.6564. \end{aligned}$$

Καθώς η ασυμπτωτική κατανομή της σ.σ. KW είναι (υπό την H_0) η $\chi^2_{k-1} \equiv \chi^2_2$, έχουμε ότι $P(\text{KW} \geqslant 5.6564) = 0.049$. Επομένως, έπεται ότι, έστω και οριακά, απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 5%.

Αν χρησιμοποιηθεί ότι η σ.σ. KW υπό τη μηδενική υπόθεση ακολουθεί ασυμπτωτικά $\chi^2_{k-1} \equiv \chi^2_2$, έχουμε ότι δεν απορρίπτεται σε επίπεδο σημαντικότητας 5% η μηδενική υπόθεση, καθώς η τιμή της στατιστικής συνάρτησης

είναι μικρότερη από την τιμή $\chi^2_{2,0.05} = 5.99$. Εναλλακτικά, θα μπορούσε κάποιος να υπολογίσει προσεγγιστικά την $P(\text{KW} \geqslant 5.6564)$, χρησιμοποιώντας την R και την εντολή `1-pchisq(5.6564,2)`, από όπου έχουμε ότι είναι ίση με 0.05911. Επομένως, χρησιμοποιώντας τη χ^2_2 προσέγγιση οδηγούμαστε στο συμπέρασμα ότι, έστω και οριακά, δεν απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 5%, ότι δηλαδή τα τρία δείγματα προέρχονται από τον ίδιο πληθυσμό. □

Στην αρχή αυτής της ενότητας, αναφέρθηκε ότι η στατιστική συνάρτηση των Kruskal-Wallis αποτελεί επέκταση, γενίκευση αυτής των Wilcoxon-Mann-Whitney. Στην πρόταση που ακολουθεί, αποδεικνύεται αυτή η ιδιότητα.

Πρόταση 6.17. Στην περίπτωση των δύο πληθυσμών οι στατιστικοί έλεγχοι που έχουν προταθεί από τους Kruskal-Wallis, Wilcoxon και Mann-Whitney ταυτίζονται.

Απόδειξη Πρότασης 6.17. Έστω ότι έχουμε $k = 2$ το πλήθος πληθυσμούς και δύο ανεξάρτητα τυχαία δείγματα $X_{i1}, \ldots, X_{i,n_i}$ από τον i-οστό πληθυσμό, $i = 1, 2$. Στην ειδική περίπτωση των δύο πληθυσμών η στατιστική συνάρτηση των Kruskal-Wallis λαμβάνει τη μορφή:

$$\text{KW} = \frac{12}{n(n+1)} \sum_{i=1}^{2} \frac{1}{n_i} \left(R_i - \frac{n_i(n+1)}{2} \right)^2 .$$

Λαμβάνοντας υπόψη ότι $R_1 + R_2 = \frac{n(n+1)}{2}$, έχουμε ότι $R_2 = \frac{n(n+1)}{2} - R_1$. Επομένως, είναι

$$\begin{aligned}
\text{KW} &= \frac{12}{n(n+1)} \left[\frac{1}{n_1} \left(R_1 - \frac{n_1(n+1)}{2} \right)^2 + \frac{1}{n_2} \left(R_2 - \frac{n_2(n+1)}{2} \right)^2 \right] \\
&= \frac{12}{n(n+1)} \left[\frac{1}{n_1} \left(R_1 - \frac{n_1(n+1)}{2} \right)^2 + \frac{1}{n_2} \left(\frac{n(n+1)}{2} - R_1 - \frac{n_2(n+1)}{2} \right)^2 \right] \\
&= \frac{12}{n(n+1)} \left[\frac{1}{n_1} \left(R_1 - \frac{n_1(n+1)}{2} \right)^2 + \frac{1}{n_2} \left(\frac{n_1(n+1)}{2} - R_1 \right)^2 \right] \\
&= \frac{12}{n(n+1)} \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \left(R_1 - \frac{n_1(n+1)}{2} \right)^2 \right] \\
&= \frac{12}{n(n+1)} \frac{n_1 + n_2}{n_1 n_2} \left(R_1 - \frac{n_1(n+1)}{2} \right)^2 = \frac{12}{n_1 n_2 (n+1)} \left(R_1 - \frac{n_1(n+1)}{2} \right)^2 .
\end{aligned}$$

Επιπλέον, στην περίπτωση μη ύπαρξης δεσμών η στατιστική συνάρτηση KW ακολουθεί προσεγγιστικά, υπό τη μηδενική υπόθεση, χ^2 κατανομή, με $k - 1 = 2 - 1 = 1$ βαθμούς ελευθερίας, δηλαδή

$$\text{KW} = \frac{12}{n_1 n_2 (n+1)} \left(R_1 - \frac{n_1(n+1)}{2} \right)^2 \xrightarrow{d} \chi^2_1$$

και η μηδενική υπόθεση απορρίπτεται, αν $\text{KW} \geqslant \chi^2_{1,a}$.

Σύμφωνα με τους Wilcoxon και Mann-Whitney η μηδενική υπόθεση απορρίπτεται για πολύ μικρές ή πολύ μεγάλες τιμές της στατιστικής συνάρτησης

$$R_1 - n(n+1)/2 = \sum_{j=1}^{n_1} R\left(X_{ij}\right) - n(n+1)/2.$$

Όμως, για μεγάλα μεγέθη δείγματος, ισχύει ότι:

$$\frac{R_1 - \mathsf{E}(R_1)}{\sqrt{\mathsf{Var}(R_1)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Λαμβάνοντας υπόψη ότι (βλ. Πρόταση 6.9):

$$\mathsf{E}(R_1) = n_1 \frac{n+1}{2} \text{ και } \mathsf{Var}(R_1) = \frac{n_1 n_2 (n+1)}{12}$$

προκύπτει, υπό την H_0, ότι:

$$\frac{R_1 - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}} \xrightarrow{d} \mathcal{N}(0,1),$$

οπότε

$$\left(\frac{R_1 - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}} \right)^2 = \frac{12}{n_1 n_2 (n+1)} \left(R_1 - \frac{n_1(n+1)}{2} \right)^2 \xrightarrow{d} \chi^2_1,$$

που αποδεικνύει το ζητούμενο. □

Σε όσα προαναφέρθηκαν, είχε υποτεθεί η μη ύπαρξη δεσμών μεταξύ των δεδομένων. Σε συνέχεια του Παραδείγματος 6.11 θα εξετάσουμε αρχικά αν υπάρχει διαφοροποίηση στην ακριβή κατανομή υπό τη μηδενική υπόθεση της στατιστικής συνάρτησης KW, όταν υπάρχουν δεσμοί και χρησιμοποιούμε τα midranks, δηλαδή τον μέσο όρο των τάξεων που θα είχαν οι παρατηρήσεις, αν δεν υπήρχαν οι δεσμοί.

Παράδειγμα 6.13. Να προσδιορίσετε την ακριβή κατανομή της στατιστικής συνάρτησης των Kruskal-Wallis, όταν $n_1 = 2$, $n_2 = 1$ και $n_3 = 1$, υποθέτοντας ότι υπάρχει ακριβώς ένας δεσμός μεταξύ της δεύτερης και τρίτης δειγματικής παρατήρησης στο κοινό διατεταγμένο δείγμα.

Λύση Παραδείγματος 6.13. Είναι

$$\mathsf{KW} = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{4 \cdot 5} \sum_{i=1}^{3} \frac{R_i^2}{n_i} - 3 \cdot 5.$$

Οι παρατηρήσεις, που είναι διαθέσιμες, είναι τέσσερις το πλήθος συνολικά και έστω ότι είναι οι $X_{11}, X_{12}, X_{21}, X_{31}$, όπου ο πρώτος δείκτης αναφέρεται στον πληθυσμό από όπου προέρχεται το δείγμα και ο δεύτερος δείκτης στον αριθμό της παρατήρησης εντός αυτού του δείγματος. Θα πρέπει να υπολογιστούν αρχικά οι δυνατές τιμές των R_1, R_2, R_3, όπου $R_i = \sum_{j=1}^{n_i} R\left(X_{ij}\right)$, $i = 1,2,3$, και $R\left(X_{ij}\right)$ $i = 1,2,3$, $j = 1,\ldots,n_i$, οι τάξεις των διαθέσιμων δειγματικών τιμών των τριών δειγμάτων στο σύνολο των $n = 4$ το πλήθος παρατηρήσεων.

Αφού στο κοινό διατεταγμένο δείγμα υπάρχει δεσμός μεταξύ της δεύτερης και της τρίτης παρατήρησης, οι βαθμοί σε αυτές τις δύο παρατηρήσεις θα είναι 2.5 και 2.5. Άρα, οι δυνατές περιπτώσεις για τις τάξεις των τεσσάρων παρατηρήσεων (οι δύο πρώτες θέσεις αφορούν τις δειγματικές παρατηρήσεις από τον πρώτο πληθυσμό, οι επόμενες κατά σειρά αυτές από τον δεύτερο και τον τρίτο πληθυσμό, αντίστοιχα) είναι: (1,2.5,2.5,4), (1,2.5,4,2.5), (1,2.5,2.5,4), (1,2.5,4,2.5), (1,4,2.5,2.5), (1,4,2.5,2.5), (2.5,2.5,1,4), (2.5,2.5,4,1), (2.5,4,1,2.5), (2.5,4,2.5,1), (2.5,4,1,2.5) και (2.5,4,2.5,1).

Στον Πίνακα 6.6 (στήλες $R(X_{11})$, $R(X_{12})$, $R(X_{21})$, $R(X_{31})$) δίνουμε τις δυνατές περιπτώσεις για τις τάξεις των τεσσάρων παρατηρήσεων (οι δύο πρώτες θέσεις αφορούν τις δειγματικές παρατηρήσεις από τον πρώτο πληθυσμό, οι επόμενες κατά σειρά αυτές από τον δεύτερο και τον τρίτο πληθυσμό, αντίστοιχα). Επίσης, στις στήλες R_1, R_2, R_3 δίνουμε τις τιμές των αθροισμάτων, ενώ στη στήλη KW δίνουμε την τιμή της στατιστικής συνάρτησης KW για καθεμία από τις δυνατές διαφορετικές τετράδες. Από τα παραπάνω προκύπτουν οι

	$R(X_{11})$	$R(X_{12})$	$R(X_{21})$	$R(X_{31})$	R_1	R_2	R_3	KW
1	1	2.5	2.5	4	3.5	2.5	4	2.025
2	1	2.5	4	2.5	3.5	4	2.5	2.025
3	1	2.5	2.5	4	3.5	2.5	4	2.025
4	1	2.5	4	2.5	3.5	4	2.5	2.025
5	1	4	2.5	2.5	5	2.5	2.5	0
6	1	4	2.5	2.5	5	2.5	2.5	0
7	2.5	2.5	1	4	5	1	4	2.7
8	2.5	2.5	4	1	5	4	1	2.7
9	2.5	4	1	2.5	6.5	1	2.5	2.025
10	2.5	4	2.5	1	6.5	2.5	1	2.025
11	2.5	4	1	2.5	6.5	1	2.5	2.025
12	2.5	4	2.5	1	6.5	2.5	1	2.025

Πίνακας 6.6: Υπολογισμοί για τον προσδιορισμό, υπό τη μηδενική υπόθεση, της κατανομής της σσ KW, για $n_1 = 2$, $n_2 = 1$ και $n_3 = 1$ και υποθέτοντας ότι υπάρχει ακριβώς ένας δεσμός μεταξύ της δεύτερης και τρίτης δειγματικής παρατήρησης στο κοινό διατεταγμένο δείγμα.

ακόλουθες αντίστοιχες τιμές για τα αθροίσματα των τάξεων:

$$(3.5,2.5,4),(3.5,4,2.5),(3.5,2.5,4),(3.5,4,2.5),(5,2.5,2.5),(5,2.5,2.5),$$

$$(5,1,4),(5,4,1),(6.5,1,2.5),(6.5,2.5,1),(6.5,1,2.5),(6.5,2.5,1).$$

Από αυτές προκύπτουν, ύστερα από αλγεβρικές πράξεις, οι ακόλουθες αντίστοιχες τιμές της στατιστικής συνάρτησης KW:

$$2.025,2.025,2.025,2.025,0,0,2.7,2.7,2.025,2.025,2.025,2.025.$$

Επομένως, οι δυνατές τιμές της στατιστικής συνάρτησης KW είναι $\{0,2.025,2.7\}$ και η συνάρτηση πιθανότητας, υπό τη μηδενική υπόθεση, είναι:

$$P(\text{KW} = x) = \begin{cases} 1/6, \text{ για } x = 2.7, \\ 2/3, \text{ για } x = 2.025, \\ 1/6, \text{ για } x = 0, \\ 0, \text{ αλλού.} \end{cases}$$

Παρατηρήστε ότι είναι τελείως διαφορετική από αυτήν που προσδιορίστηκε στο Παράδειγμα 6.11. □

Για καλύτερη κατανόηση του σκεπτικού εύρεσης της ακριβούς κατανομής για μικρά σε μέγεθος δείγματα υπό την παρουσία δεσμών, δίνεται ακόμη ένα παράδειγμα.

Παράδειγμα 6.14. (Hollander *et al.*, 2014) Να προσδιορίσετε την ακριβή κατανομή του στατιστικού των Kruskal and Wallis (1952), όταν $n_1 = 2$, $n_2 = 2$ και $n_3 = 1$, υποθέτοντας ότι υπάρχουν ακριβώς δύο δεσμοί στις δειγματικές παρατηρήσεις μεταξύ των δύο πρώτων και δύο τελευταίων σε αύξουσα τάξη μεγέθους παρατηρήσεων.

Λύση Παραδείγματος 6.14. Είναι

$$\text{KW} = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{5 \cdot 6} \sum_{i=1}^{3} \frac{R_i^2}{n_i} - 3 \cdot 6,$$

καθώς $n_1 = 2$, $n_2 = 2$ και $n_3 = 1$ ($k = 3$). Επομένως, για την υπό μελέτη περίπτωση θα πρέπει να βρούμε τις δυνατές τιμές της στατιστικής συνάρτησης εξετάζοντας τις δυνατές τιμές των R_i, $i = 1, 2, 3$. Οι παρατηρήσεις, που είναι διαθέσιμες, είναι 5 το πλήθος και, λόγω της ύπαρξης των δεσμών μεταξύ των παρατηρήσεων που αναφέρονται, οι τάξεις είναι $(1.5, 1.5, 3, 4.5, 4.5)$. Θα πρέπει να υπολογιστούν αρχικά οι δυνατές τιμές των R_1, R_2, R_3, όπου $R_i = \sum_{j=1}^{n_i} R\left(X_{ij}\right)$, $i = 1, 2, 3$, και $R\left(X_{ij}\right)$ $i = 1, 2, 3$, $j = 1, ..., n_i$, οι τάξεις των διαθέσιμων δειγματικών τιμών των τριών δειγμάτων στο σύνολο των $n = 5$ το πλήθος παρατηρήσεων. Οι δυνατές περιπτώσεις τότε είναι 30 το πλήθος και παρατίθενται στον Πίνακα 6.7. Επομένως, ύστερα από αλγεβρικές πράξεις, προκύπτει ότι οι δυνατές τιμές της στατιστικής συνάρτησης KW είναι $\{0, 1.35, 3.15, 3.6\}$ και η συνάρτηση πιθανότητας, υπό τη μηδενική υπόθεση, είναι:

$$P(\mathrm{KW} = x) = \begin{cases} 16/30, \text{ για } x = 1.35, \\ 8/30, \text{ για } x = 3.15, \\ 2/30, \text{ για } x = 3.6, \\ 4/30, \text{ για } x = 0. \end{cases}$$

□

Προφανώς, εκτός από την ακριβή κατανομή επηρεάζεται και η ασυμπτωτική κατανομή της στατιστικής συνάρτησης KW όταν υπάρχουν δεσμοί. Σε μία τέτοια περίπτωση ύπαρξης δεσμών ανάμεσα στις δειγματικές παρατηρήσεις, οι Kruskal and Wallis (1952) πρότειναν τη στατιστική συνάρτηση:

$$\mathrm{KW}^* = \frac{\mathrm{KW}}{1 - \frac{\sum_{l=1}^{c}\left(d_l^3 - d_l\right)}{n^3 - n}},$$

όπου KW είναι η συνήθης στατιστική συνάρτηση των Kruskal-Wallis, υπολογισμένη χρησιμοποιώντας τα midranks, και d_l είναι το πλήθος των παρατηρήσεων, που συμμετέχουν στον l-οστό δεσμό ($l = 1, .., c$). Υπό τη μηδενική υπόθεση, η στατιστική συνάρτηση KW^* ακολουθεί χι-τετράγωνο κατανομή με $k - 1$ βαθμούς ελευθερίας και απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας a, αν $\mathrm{KW}^* \geqslant \chi^2_{k-1,a}$. Η παραπάνω τροποποίηση αποσαφηνίζεται στο παράδειγμα που ακολουθεί. Πριν από αυτό απλώς αναφέρουμε ότι εύκολα προκύπτει ότι στην περίπτωση μη ύπαρξης δεσμών, δηλαδή για $c = 0$, είναι $\mathrm{KW}^* = \mathrm{KW}$.

Παράδειγμα 6.15. Στον πίνακα που ακολουθεί καταγράφεται το βάρος νεογνών από 4 διαφορετικές χώρες. Σε επίπεδο σημαντικότητας 5% να εξετάσετε, με το στατιστικό τεστ των Kruskal-Wallis, αν τα 4 δείγματα προέρχονται από τον ίδιο πληθυσμό.

Χώρα Α	2	2.8	3.3	3.2	3.9	3.6	
Χώρα Β	2.8	2	3.5	3,6	3.7	3.8	3.3
Χώρα Γ	2.7	2.1	2.6	2.8	3.7		
Χώρα Δ	2.6	3.3	3.2	2.9	3.8		

Λύση Παραδείγματος 6.15. Έχουμε διαθέσιμα τέσσερα δείγματα, μεγέθους $n_1 = 6$, $n_2 = 7$, $n_3 = 5$, $n_4 = 5$, αντίστοιχα, και, επομένως, το σύνολο των παρατηρήσεων είναι $n = n_1 + n_2 + n_3 = 23$. Αναμειγνύονται οι δειγματικές παρατηρήσεις X_{ij}, $j = 1, ..., n_i$, $i = 1, 2, 3, 4$ και, στη συνέχεια, διατάσσονται σε αύξουσα τάξη μεγέθους και υπολογίζονται οι τάξεις $R\left(X_{ij}\right)$, $j = 1, ..., n_i$, $i = 1, 2, 3, 4$ κάθε δειγματικής τιμής.

Για ευκολία στους μετέπειτα υπολογισμούς δημιουργούμε τον ακόλουθο πίνακα:

Τιμή	2	2	2.1	2.6	2.6	2.7	2.8	2.8	2.8	2.9	3.2	3.2
Τάξη	1.5	1.5	3	4.5	4.5	6	8	8	8	10	11.5	11.5
Δείγμα	1	2	3	3	4	3	1	2	3	4	1	4
Τιμή	3.3	3.3	3.3	3.5	3.6	3.6	3.7	3.7	3.8	3.8	3.9	
Τάξη	14	14	14	16	17.5	17.5	19.5	19.5	21.5	21.5	23	
Δείγμα	1	2	4	2	1	2	2	2	2	4	1	

$R(X_{1j})$	$R(X_{2j})$	$R(X_{3j})$	Τιμή KW
1.5,3	4.5,4.5	1.5	3.15
1.5,3	4.5,4.5	1.5	3.15
1.5,4.5	3,4.5	1.5	1.35
1.5,4.5	3,4.5	1.5	1.35
3,4.5	1.5,4.5	1.5	1.35
3,4.5	15,4.5	1.5	1.35
1.5,4.5	3,4.5	1.5	1.35
1.5,4.5	3,4.5	1.5	1.35
3,4.5	1.5,4.5	1.5	1.35
3,4.5	1.5,4.5	1.5	1.35
4.5,4.5	1.5,3	1.5	3.15
4.5,4.5	1.5,3	1.5	3.15
1.5,1.5	4.5,4.5	3	3.6
1.5,4.5	1.5,4.5	3	0
1.5,4.5	1.5,4.5	3	0
1.5,4.5	1.5,4.5	3	0
1.5,4.5	1.5,4.5	3	0
4.5,4.5	1.5,1.5	3	3.6
1.5,1.5	3,4.5	4.5	3.15
1.5,1.5	3,4.5	4.5	3.15
1.5,3	1.5,4.5	4.5	1.35
1.5,3	1.5,4.5	4.5	1.35
1.5,3	1.5,4.5	4.5	1.35
1.5,3	1.5,4.5	4.5	1.35
1.5,4.5	1.5,3	4.5	1.35
1.5,4.5	1.5,3	4.5	1.35
1.5,4.5	1.5,3	4.5	1.35
1.5,4.5	1.5,3	4.5	1.35
3,4.5	1.5,1.5	4.5	3.15
3,4.5	1.5,1.5	4.5	3.15

Πίνακας 6.7: Υπολογισμοί για τον προσδιορισμό, υπό τη μηδενική υπόθεση, της κατανομής του KW για $n_1 = 2, n_2 = 2$ και $n_3 = 1$ και υποθέτοντας ότι υπάρχουν ακριβώς δύο δεσμοί στις δειγματικές παρατηρήσεις μεταξύ των δύο πρώτων και δύο τελευταίων σε αύξουσα τάξη μεγέθους παρατηρήσεων.

Επομένως, είναι:

$$R_1 = \sum_{j=1}^{6} R\left(X_{1j}\right) = 1.5 + 8 + 11.5 + 14 + 17.5 + 23 = 75.5,$$

$$R_2 = \sum_{j=1}^{7} R\left(X_{2j}\right) = 1.5 + 8 + 14 + 16 + 17.5 + 19.5 + 21.5 = 98,$$

$$R_3 = \sum_{j=1}^{5} R\left(X_{3j}\right) = 3 + 4.5 + 6 + 8 + 19.5 = 41,$$

και

$$R_4 = \sum_{j=1}^{5} R\left(X_{4j}\right) = 4.5 + 10 + 11.5 + 14 + 21.5 = 61.5.$$

Επιπλέον, υπάρχουν συνολικά οκτώ το πλήθος δεσμοί και ο αριθμός των παρατηρήσεων σε κάθε δεσμό είναι: $d_1 = d_2 = 2$, $d_3 = 3$, $d_4 = 2$, $d_5 = 3$ και $d_6 = d_7 = d_8 = 2$, αντίστοιχα.

Επομένως, είναι:

$$\mathrm{KW}^* = \frac{\mathrm{KW}}{1 - \frac{\sum_{l=1}^{8}\left(d_l^3 - d_l\right)}{n^3 - n}},$$

όπου

$$\begin{aligned}\mathrm{KW} &= \frac{12}{n(n+1)}\sum_{i=1}^{4}\frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{23(23+1)}\left(\frac{75.5^2}{6} + \frac{98^2}{7} + \frac{41^2}{5} + \frac{61.5^2}{5}\right) - 3(23+1)\\ &= \frac{1}{46}(950.0417 + 1372 + 336.2 + 756.45) - 72 = \frac{1}{46}3414.6917 - 72 = 2.232\end{aligned}$$

και

$$1 - \frac{\sum_{l=1}^{8}\left(d_l^3 - d_l\right)}{n^3 - n} = 1 - \frac{6\cdot(2^3 - 2) + 2\cdot(3^3 - 3)}{23^3 - 23} = 1 - \frac{84}{44677.5} = 0.99812.$$

Συνδυάζοντας τα παραπάνω έχουμε:

$$\mathrm{KW}^* = \frac{\mathrm{KW}}{1 - \frac{\sum_{l=1}^{8}\left(d_l^3 - d_l\right)}{n^3 - n}} = \frac{2.232}{0.99812} = 2.236.$$

Καθώς, υπό την H_0, η προσεγγιστική κατανομή της στατιστικής συνάρτησης ελέγχου είναι η χ^2_3, προκύπτει, με τη βοήθεια της `R` (και της εντολής `pchisq(2.236,3,lower.tail=F)`), ότι $P\left(\chi^2_3 \geqslant 2.236\right) = 0.5248923$. Επομένως, οδηγούμαστε στο συμπέρασμα ότι σε επίπεδο σημαντικότητας 5%, τα τέσσερα δείγματα προέρχονται από τον ίδιο πληθυσμό. Στο ίδιο συμπέρασμα καταλήγουμε, σε επίπεδο σημαντικότητας 5%, και από το γεγονός ότι η τιμή της στατιστικής συνάρτησης KW^* δεν υπερβαίνει την τιμή $\chi^2_{3,0.05} = 7.81$. □

Παρατήρηση 6.8. Επισημαίνουμε ότι στην περίπτωση που θέλουμε να ελέγξουμε, χρησιμοποιώντας $k \geqslant 3$ το πλήθος ανεξάρτητα τυχαία δείγματα, την ισότητα των πληθυσμιακών διαμέσων έναντι της εναλλακτικής ότι υπάρχει μια δοθείσα διάταξη στις πληθυσμιακές διαμέσους, έχει παρουσιαστεί στη βιβλιογραφία ένας ισχυρότερος έλεγχος από αυτόν των Kruskal–Wallis. Ο έλεγχος αυτός είναι γνωστός ως έλεγχος των Jonckheere–Terpstra και παραπέμπουμε για περισσότερες λεπτομέρειες στις εργασίες Jonckheere (1954) και Terpstra (1952).

6.4.2 Εξαρτημένα δείγματα: ο έλεγχος του Friedman

Έστω ότι έχουμε k το πλήθος πληθυσμούς, $k \geqslant 3$, με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1,...,k$, $k \geqslant 3$. Επιπλέον, λαμβάνουμε k το πλήθος, $k \geqslant 3$, εξαρτημένα τυχαία δείγματα, ένα από καθένα από αυτούς τους $i = 1,...,k$, $k \geqslant 3$, πληθυσμούς. Έστω n το μέγεθος κάθε δείγματος και $X_{i1},...,X_{i,n}$, οι δειγματικές τιμές από τον i-οστό πληθυσμό, $i = 1,...,k$, $k \geqslant 3$. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση της ισότητας των πληθυσμιακών διαμέσων

$$H_0 : m_1 = ... = m_k$$

έναντι της εναλλακτικής

$$H_1 : m_i \neq m_j, \text{ για κάποιο ζεύγος } (i,j), \text{ με } i \neq j,\ i,j = 1,...,k.$$

Στο πλαίσιο αυτό, ο πιο δημοφιλής έλεγχος είναι αυτός που προτάθηκε από τον Αμερικανό οικονομολόγο και στατιστικό Milton Friedman (1912-2006), βραβευμένο το έτος 1976 με Νόμπελ στο γνωστικό αντικείμενο των οικονομικών. Η διαδικασία που προτάθηκε από τον Friedman (1937) για τον έλεγχο περισσότερων των δύο πληθυσμιακών διαμέσων με εξαρτημένα δείγματα αποτελεί αντικείμενο μελέτης αυτής της ενότητας και για την καλύτερη κατανόησή της θα χρησιμοποιήσουμε τον πίνακα διπλής εισόδου που ακολουθεί. Σε αυτόν, σε κάθε γραμμή δίνονται οι παρατηρήσεις σε καθεμία εκ των n το πλήθος διακεκριμένων πειραματικών μονάδων, ενώ σε κάθε στήλη δίνονται οι παρατηρήσεις καθενός εκ των k το πλήθος εξαρτημένων δειγμάτων.

Πειραματική μονάδα	Δείγματα					
	1	2	...	i	...	k
1	X_{11}	X_{21}	...	X_{i1}	...	X_{k1}
2	X_{12}	X_{22}	...	X_{i2}	...	X_{k2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
j	X_{1j}	X_{2j}	...	X_{ij}	...	X_{kj}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
n	X_{1n}	X_{2n}	...	X_{in}	...	X_{kn}

Αρχικά, αποδίδουμε σε κάθε δειγματική παρατήρηση την τιμή της τάξης της, όπως αυτή υπολογίζεται, εντός κάθε γραμμής. Δηλαδή σε κάθε παρατήρηση εντός κάθε γραμμής αποδίδεται μία τάξη από τις $\{1,...,k\}$. Σε όσα ακολουθούν, συμβολίζουμε με $R(X_{ij}) = R_{ij}$, $i = 1,...,k$, $j = 1,...,n$, τις προαναφερθείσες τάξεις. Επομένως, διαμορφώνεται ο ακόλουθος πίνακας:

Πειραματική μονάδα	Δείγματα					
	1	2	...	i	...	k
1	R_{11}	R_{21}	...	R_{i1}	...	R_{k1}
2	R_{12}	R_{22}	...	R_{i2}	...	R_{k2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
j	R_{1j}	R_{2j}	...	R_{ij}	...	R_{kj}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
n	R_{1n}	R_{2n}	...	R_{in}	...	R_{kn}

Ο παραπάνω τρόπος υπολογισμού των τάξεων αποτελεί μια πρώτη σημαντική διαφορά, καθώς οι τάξεις δεν υπολογίζονται ούτε στο σύνολο των nk το πλήθος παρατηρήσεων ούτε εντός κάθε δείγματος. Αφού υπολογιστούν οι τάξεις, με τον προαναφερόμενο τρόπο, υπολογίζουμε το άθροισμα και τον μέσο όρο των τάξεων κάθε δείγματος, έστω R_i και $\bar{R}_i$, αντίστοιχα, για $i = 1,...,k$. Δηλαδή, με άλλα λόγια και αναφερόμενοι στον παραπάνω βοηθητικό πίνακα, υπολογίζουμε το άθροισμα και τον μέσο όρο των τάξεων σε καθεμία από

τις k το πλήθος στήλες. Επομένως, είναι

$$\bar{R}_i = \frac{R_i}{n} = \frac{\sum_{j=1}^{n} R(X_{ij})}{n}, \text{ για } i = 1,...,k.$$

Η ιδέα του Friedman στη συνέχεια είναι η εξής: αν η μηδενική υπόθεση είναι αληθής, τότε το σύνολο των τάξεων σε κάθε στήλη θα παριστάνει ένα τυχαίο δείγμα μεγέθους n από το σύνολο $\{1,...,k\}$. Επιπλέον, κάθε R_{ij} είναι μια τυχαία μεταβλητή με μέση τιμή $\frac{(k+1)}{2}$ και διακύμανση $(k^2-1)/12$ (βλ. Πρόταση 6.1 με τον ρόλο του n να ανήκει στο k). Επομένως, άμεσα προκύπτει ότι υπό τη μηδενική υπόθεση:

$$\mathrm{E}\left(\bar{R}_i\right) = \frac{(k+1)}{2}, \text{ και } \mathrm{Var}\left(\bar{R}_i\right) = \frac{k^2-1}{12n}, \text{ για } i = 1,...,k.$$

Από τα παραπάνω προκύπτει ότι μια στατιστική συνάρτηση για τον έλεγχο της υπό μελέτη μηδενικής υπόθεσης θα μπορούσε να είναι κάποια που θα μετρά την απόκλιση των R_i από τη μέση τιμή τους. Μια λογική, λοιπόν, επιλογή θα ήταν είτε η στατιστική συνάρτηση

$$S^* = \sum_{i=1}^{k} \left(\bar{R}_i - \frac{k+1}{2}\right)^2,$$

είτε η

$$S = \sum_{i=1}^{k} \left(R_i - \frac{n(k+1)}{2}\right)^2 = n^2 S^*,$$

οι οποίες και έχουν χρησιμοποιηθεί για τον εν λόγω έλεγχο. Ωστόσο, ο Friedman θεώρησε τη στατιστική συνάρτηση

$$Q = \frac{12n}{k(k+1)} \sum_{i=1}^{k} \left(\bar{R}_i - \frac{k+1}{2}\right)^2, \tag{6.21}$$

η οποία, λαμβάνοντας υπόψη τις παραπάνω σχέσεις, γράφεται ισοδύναμα ως:

$$Q = \frac{12n}{k(k+1)} S^* = \frac{12n}{k(k+1)} \frac{S}{n^2} = \frac{12S}{nk(k+1)}.$$

Εναλλακτικά, μετά από κάποιες αλγεβρικές πράξεις, η Q γράφεται ισοδύναμα ως:

$$\begin{aligned}
Q &= \frac{12}{nk(k+1)} \left\{ \sum_{i=1}^{k} R_i^2 + \frac{kn^2(k+1)^2}{4} - 2\frac{n(k+1)}{2} \sum_{i=1}^{k} R_i \right\} \\
&= \frac{12}{nk(k+1)} \left\{ \sum_{i=1}^{k} R_i^2 + \frac{kn^2(k+1)^2}{4} - 2\frac{n(k+1)}{2} \sum_{i=1}^{k} \sum_{j=1}^{n} R_{ij} \right\} \\
&= \frac{12}{nk(k+1)} \left\{ \sum_{i=1}^{k} R_i^2 + \frac{kn^2(k+1)^2}{4} - 2\frac{n(k+1)}{2} n\frac{k(k+1)}{2} \right\} \\
&= \frac{12}{nk(k+1)} \left\{ \sum_{i=1}^{k} R_i^2 + \frac{kn^2(k+1)^2}{4} - 2\frac{n^2k(k+1)^2}{4} \right\} \\
&= \frac{12}{nk(k+1)} \left\{ \sum_{i=1}^{k} R_i^2 - \frac{n^2k(k+1)^2}{4} \right\} \\
&= \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1).
\end{aligned}$$

Στο σημείο αυτό, επισημαίνεται ότι η λογική που κρύβεται πίσω από την πρόταση του Friedman θα εξηγηθεί στη συνέχεια, όχι αυστηρά μαθηματικά, αλλά περισσότερο διαισθητικά.

Προφανώς, όποια και από τις παραπάνω στατιστικές συναρτήσεις και αν χρησιμοποιηθεί, θα απορρίπτεται η μηδενική υπόθεση για μεγάλες τιμές της στατιστικής συνάρτησης. Η ακριβής κατανομή της στατιστικής συνάρτησης Q (άρα και των υπόλοιπων, καθώς αποτελούν 1-1 μετασχηματισμό της) έχει προσδιοριστεί από τον Friedman (1937) στην περίπτωση που $k = 3$ και $1 < n \leqslant 9$ ή $k = 4$ και $1 < n \leqslant 4$, ενώ για περισσότερες αναφορές σε προσπάθειες εύρεσης της ακριβούς κατανομής παραπέμπουμε στο σύγγραμμα των Bortz *et al.* (2000) και τις εκεί αναφορές. Επομένως, είναι αναγκαίος ο προσδιορισμός προσεγγιστικής κατανομής για τις παραπάνω στατιστικές συναρτήσεις. Αυτός ήταν και ο λόγος που ο Friedman οδηγήθηκε στη στατιστική συνάρτηση Q. Σύμφωνα με όσα αναφέρθηκαν, για k μεγάλο, κάθε $\bar{R}_i$, $i = 1, ..., k$, μπορεί να προσεγγιστεί από την κανονική κατανομή με μέση τιμή $\frac{k+1}{2}$. Ωστόσο, οι k το πλήθος αυτές τυχαίες μεταβλητές δεν είναι ανεξάρτητες, καθώς το άθροισμά τους είναι ίσο με $k(k+1)/2$. Τότε αποδεικνύεται (δείτε (βλ. Friedman, 1937, και τις εκεί αναφορές στο έργο του Samuel Stanley Wilks (1906-1964)) ότι, καθώς το μέγεθος δείγματος n αυξάνεται, το πηλίκο

$$\frac{(k-1)}{k} \cdot \frac{\sum_{i=1}^{k} \left(\bar{R}_i - \mathrm{E}(\bar{R}_i)\right)^2}{\sqrt{\mathrm{Var}(\bar{R}_i)}}$$

προσεγγίζει τη χ^2_{k-1}, δηλαδή την χι-τετράγωνο κατανομή με $k-1$ βαθμούς ελευθερίας. Επιπρόσθετα, οι ακριβείς τιμές των τριών πρώτων ροπών της στατιστικής συνάρτησης Q υπό τη μηδενική υπόθεση έχουν προσδιοριστεί από τον Friedman (1937), από όπου έχουμε ότι $\mathrm{E}(Q) = k-1$ και $\mathrm{Var}(Q) = \frac{2(k-1)(n-1)}{n}$. Η παραπάνω προσεγγιστική κατανομή είναι, σύμφωνα με τον Friedman (1937), ικανοποιητική όταν $k = 3$ για $n \geqslant 9$, όταν $k \geqslant 4$ για $n \geqslant 6$, ενώ στο σύγγραμμα των Gibbons and Chakraborti (2020) αναφέρεται η τιμή $k > 7$. Συνοψίζοντας τα παραπάνω, η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας a αν $Q \geqslant \chi^2_{k-1,a}$.

Παρατήρηση 6.9. Στην περίπτωση που υπάρχουν δεσμοί, τότε (βλ. Kvam and Vidakovic, 2007) προτείνεται να χρησιμοποιείται η στατιστική συνάρτηση

$$Q' = \frac{k-1}{\sum_{i=1}^{k}\sum_{j=1}^{n} R(X_{ij})^2 - 0.25nk(k+1)^2} \left(\sum_{i=1}^{k} R_i^2 - 0.25n^2k(k+1)^2 \right), \tag{6.22}$$

η οποία ακολουθεί προσεγγιστικά χι-τετράγωνο κατανομή με $k-1$ βαθμούς ελευθερίας, δηλαδή την χ^2_{k-1}.

Αν απορρίψουμε την H_0, τότε (βλ. π.χ. Conover, 1998; Kvam and Vidakovic, 2007) κάνουμε συγκρίσεις (ανά δύο) και απορρίπτουμε την ισότητα των m_i, m_j σε επίπεδο σημαντικότητας a αν

$$|\bar{R}_i - \bar{R}_j| > t_{(k-1)(n-1),a/2} \sqrt{2\frac{n\sum_{i=1}^{k}\sum_{j=1}^{n} R(X_{ij})^2 - \sum_{i=1}^{k} R_i^2}{(k-1)(n-1)}}.$$

Παρατήρηση 6.10. Ο έλεγχος που παρουσιάστηκε σε αυτήν την ενότητα είναι ο πιο γνωστός και πιο ιστορικός έλεγχος για το υπό μελέτη πρόβλημα. Επισημαίνουμε, όμως, ότι δεν είναι ο πλέον ισχυρός (βλ. παρόμοιο σχόλιο από τους Kvam and Vidakovic, 2007). Για παράδειγμα, οι Iman and Davenport (1980) πρότειναν να χρησιμοποιείται η στατιστική συνάρτηση $F = \frac{(n-1)Q}{n(k-1)-Q}$, η οποία προσεγγιστικά έχει F κατανομή με $k-1$ και $(n-1)(k-1)$ βαθμούς ελευθερίας. Ο έλεγχος που προκύπτει τότε είναι εν γένει ισχυρότερος του ελέγχου με τη στατιστική συνάρτηση Q. Για άλλες παραλλαγές του ελέγχου παραπέμπουμε τον/την ενδιαφερόμενο/μενη αναγνώστη/στρια στην εργασία των Eisinga *et al.* (2017) και στις εκεί αναφορές.

Παρατήρηση 6.11. Ένας έλεγχος που έχει παρουσιαστεί στη βιβλιογραφία και μπορεί να θεωρηθεί ειδική περίπτωση του ελέγχου του Friedman είναι ο έλεγχος που είναι γνωστός ως Cochran's Q. Ο μη παραμετρικός αυτός έλεγχος χρησιμοποιείται για τον έλεγχο της μηδενικής υπόθεσης της ισότητας k το πλήθος πληθυσμιακών ποσοστών, με $k > 2$, έναντι της εναλλακτικής υπόθεσης ότι τουλάχιστον δύο εξ αυτών διαφέρουν, όταν έχουμε k το πλήθος εξαρτημένα δείγματα. Για αυτόν τον λόγο θεωρείται και επέκταση του ελέγχου McNemar, ο οποίος παρουσιάστηκε στο Κεφάλαιο 5.

Παράδειγμα 6.16. Σε έναν διαγωνισμό μαγειρικής, 5 κριτές αξιολόγησαν τα πιάτα που έφτιαξαν οι 8 συμμετέχοντες στον τελικό του διαγωνισμού. Τα αποτελέσματα των αξιολογήσεων δίνονται στον επόμενο πίνακα, όπου περιέχονται οι τάξεις των βαθμολογιών που έδωσε κάθε κριτής. Να ελέγξετε την υπόθεση, σε ε.σ. 10%, ότι ο τρόπος βαθμολογίας των κριτών είναι ανεξάρτητος των φιναλίστ.

	Σ1	Σ2	Σ3	Σ4	Σ5	Σ6	Σ7	Σ8
1	8	2	1	3	6	7	4	5
2	1.5	1.5	3	6	6	7	8	5
3	5	6.5	6.5	1	2.5	2.5	4	8
4	7	1	3.5	3.5	5	6	2	8
5	4	3	5.5	5.5	1	2	7.5	7.5

Λύση Παραδείγματος 6.16. Θα χρησιμοποιήσουμε το τεστ του Friedman και, συγκεκριμένα, τη στατιστική συνάρτηση Q', αφού υπάρχουν δεσμοί (βλ. Παρατήρηση 6.9). Είναι $k = 8$, $n = 5$, ενώ οι τάξεις $R(X_{ij})$, $i = 1,2,...,8$, $j = 1,2,3,4,5$, δίνονται στον πίνακα της εκφώνησης. Άρα, άμεσα έπεται ότι

$$R_1 = 8 + 1.5 + 5 + 7 + 4 = 2.5,\ R_2 = 2 + 1.5 + 6.5 + 1 + 3 = 14,\ R_3 = 1 + 3 + 6.5 + 3.5 + 5.5 = 19.5,$$

$$R_4 = 3 + 6 + 1 + 3.5 + 5.5 = 19,\ R_5 = 6 + 6 + 2.5 + 5 + 1 = 20.5,\ R_6 = 7 + 7 + 2.5 + 6 + 2 = 24.5,$$

$$R_7 = 4 + 8 + 4 + 2 + 7.5 = 25.5,\ R_8 = 5 + 5 + 8 + 8 + 7.5 = 33.5.$$

Επίσης, $\sum_{i=1}^{8}\sum_{j=1}^{5} R(X_{ij})^2 = 1037$, $\sum_{i=1}^{8} R_i^2 = 4380.5$, και $0.25nk(k+1)^2 = 810$, ενώ $0.25n^2k(k+1)^2 = 4050$. Άρα, με αντικατάσταση στη σχέση (6.22) έχουμε ότι

$$Q' = \frac{8-1}{1037-810}(4380.5 - 4050) \approx 10.192.$$

Για τον έλεγχο, θα χρησιμοποιηθεί η προσεγγιστική κατανομή της παραπάνω στατιστικής συνάρτησης, η οποία είναι η χ^2_7. Άρα, σε ε.σ. 10%, το άνω 0.10-ποσοστιαίο σημείο της κατανομής χ^2_7 είναι το $\chi^2_{7,0.10} = 12.017$. Αφού $Q' = 10.192 < 12.017 = \chi^2_{7,0.10}$, δεν απορρίπτουμε, σε ε.σ. 10%, την H_0. Δηλαδή, δεν έχουμε σαφείς ενδείξεις ότι ο τρόπος βαθμολογίας των κριτών δεν είναι ανεξάρτητος των συμμετεχόντων στον τελικό του διαγωνισμού. □

6.5 Έλεγχοι ισότητας πληθυσμιακών διακυμάνσεων με ανεξάρτητα δείγματα

Τα στατιστικά τεστ που έχουν προταθεί στη βιβλιογραφία, υπό την υπόθεση της κανονικότητας, για τον έλεγχο της ισότητας δύο ή περισσότερων πληθυσμιακών διακυμάνσεων είναι μη ανθεκτικά σε αποκλίσεις από την κανονική κατανομή. Επομένως, δεν πρέπει να χρησιμοποιούνται σε περιπτώσεις που οι πληθυσμοί δεν περιγράφονται ικανοποιητικά από την κανονική κατανομή. Σε αυτήν την ενότητα, θα αναπτυχθούν μη παραμετρικές μεθοδολογίες που βασίζονται στις τάξεις των δεδομένων, τόσο για τον έλεγχο της ισότητας δύο πληθυσμιακών διακυμάνσεων όσο και για τον έλεγχο της ισότητας περισσότερων των δύο πληθυσμιακών διακυμάνσεων.

6.5.1 Έλεγχοι ισότητας δύο πληθυσμιακών διακυμάνσεων

Πλήθος διαφορετικών μη παραμετρικών μεθοδολογιών που βασίζονται στις τάξεις έχει προταθεί για τον έλεγχο της ισότητας δύο πληθυσμιακών διακυμάνσεων. Στο πλαίσιο αυτού του συγγράμματος θα παρουσιαστεί ο πρώτος χρονολογικά εξ αυτών και έπειτα δύο από τους πιο ευρέως γνωστούς τέτοιους ελέγχους. Για περισσότερες λεπτομέρειες τόσο για τους ελέγχους που θα παρουσιαστούν σε αυτήν την ενότητα, αλλά και για πληροφορίες σχετικά με άλλους παραπέμπουμε τον/την ενδιαφερόμενο/μενη αναγνώστη/στρια στα συγγράμματα των Conover (1998), Gibbons and Chakraborti (2020), καθώς και στην εργασία των Conover *et al.* (1981).

6.5.1.1 Ο έλεγχος του Mood

Ο πρώτος χρονολογικά μη παραμετρικός έλεγχος της υπόθεσης της ισότητας δύο πληθυσμιακών διακυμάνσεων που βασίζεται στις τάξεις παρουσιάστηκε από τον Mood (1954) και αποτελεί αντικείμενο μελέτης αυτής της υποενότητας.

Έστω δύο πληθυσμοί με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1,2$, με πληθυσμιακές διακυμάνσεις σ_1^2 και σ_2^2, αντίστοιχα. Επιπλέον, υποθέτουμε ότι είναι διαθέσιμα δύο ανεξάρτητα μεταξύ τους τυχαία δείγματα από αυτούς τους πληθυσμούς, έστω $X_{11}, \ldots, X_{1,n_1}$ και $X_{21}, \ldots, X_{2,n_2}$, αντίστοιχα, με $n = n_1 + n_2$. Υποθέτουμε ότι οι πληθυσμιακές μέσες τιμές είναι ίσες ή ότι υπάρχει γνωστός μετασχηματισμός, έτσι ώστε να είναι ίσες. Ο Mood (1954) πρότεινε έναν τρόπο ελέγχου της μηδενικής υπόθεσης $H_0 : F_1(u) = F_2(u)$, για κάθε $u \in \mathbb{R}$ έναντι μίας εκ των τριών εναλλακτικών $H_1 : F_1(\theta u) = F_2(u)$, για κάθε $u \in \mathbb{R}$ και κάποιο i) $\theta > 1$, ii) $\theta < 1$, ή iii) $\theta \neq 1$, όπου $\theta = \frac{\sigma_1}{\sigma_2} > 0$. Εναλλακτικά, θεωρώντας ότι $F_1(\theta u) = F_2(u)$, για κάθε $u \in \mathbb{R}$, προβαίνουμε σε έλεγχο της μηδενικής υπόθεσης $\mathrm{H}_0 : \theta = 1$ έναντι της i) $\theta > 1$, ii) $\theta < 1$, ή iii) $\theta \neq 1$, όπου $\theta = \frac{\sigma_1}{\sigma_2} > 0$.

Στο πλαίσιο αυτό, αρχικά, αναμειγνύονται τα δύο δείγματα και διατάσσονται οι δειγματικές παρατηρήσεις σε αύξουσα τάξη μεγέθους. Ο έλεγχος, που προτάθηκε από τον Mood (1954) για την ισότητα των διακυμάνσεων δύο ανεξάρτητων πληθυσμών, βασίζεται (στην περίπτωση μη ύπαρξης δεσμών) στη στατιστική συνάρτηση:

$$M = \sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right)^2 Z_i, \tag{6.23}$$

όπου Z_i είναι η συνήθης δείκτρια μεταβλητή που λαμβάνει την τιμή 1, αν η παρατήρηση που έχει τάξη ίση με i προέρχεται από τον πρώτο πληθυσμό, και την τιμή 0, αν προέρχεται από τον δεύτερο. Από την άλλη μεριά, στην περίπτωση ύπαρξης δεσμών προτείνεται η στατιστική συνάρτηση

$$M^* = \sum_{i=1}^{n} \left(r_i - \frac{n+1}{2} \right)^2 Z_i, \tag{6.24}$$

με r_i την τάξη της i-οστής διατεταγμένης παρατήρησης στο δείγμα όλων των $n = n_1 + n_2$ το πλήθος παρατηρήσεων, με τους δεσμούς να αντιμετωπίζονται με τη μέθοδο των midranks.

Ουσιαστικά (και στις δύο περιπτώσεις) η στατιστική συνάρτηση είναι το άθροισμα των τετραγώνων των διαφορών των τάξεων των παρατηρήσεων του πρώτου δείγματος στο αναμεμειγμένο δείγμα από τον μέσο όρο των τάξεων. Θυμηθείτε ότι ο μέσος όρος των τάξεων, είτε έχουμε δεσμούς είτε όχι, είναι ίσος με $(n+1)/2$, όταν, ως μέθοδος χειρισμού των ισοβαθμιών, επιλέγεται η χρήση των midranks.

Από τον τρόπο ορισμού της στατιστικής συνάρτησης M (αντίστοιχα και της M^*) γίνεται αντιληπτό ότι μεγάλες (μικρές) τιμές της υποδεικνύουν ότι η διακύμανση του πρώτου πληθυσμού είναι μεγαλύτερη (μικρότερη) από την αντίστοιχη του πρώτου. Επομένως, θα απορρίπτεται η $H_0 : \theta = 1$ έναντι της εναλλακτικής i) $H_1 : \theta = \frac{\sigma_1}{\sigma_2} > 1$ για μεγάλες τιμές της στατιστικής συνάρτησης M, έναντι της ii) $H_1 : \theta = \frac{\sigma_1}{\sigma_2} < 1$ για μικρές τιμές της στατιστικής συνάρτησης M ή έναντι της iii) $H_1 : 0 < \theta \neq 1$ για μικρές και μεγάλες τιμές της M.

Για τον προσδιορισμό των τιμών που θεωρούνται μικρές (ή μεγάλες, ανάλογα με τη μορφή της εναλλακτικής υπόθεσης), για να απορρίπτεται η μηδενική υπόθεση, απαιτείται ο προσδιορισμός της κατανομής της στατιστικής συνάρτησης M (ή της M^*) υπό τη μηδενική υπόθεση. Αυτός ο προσδιορισμός είναι εφικτός για μικρά σε μέγεθος δείγματα και για πίνακες κρίσιμων τιμών με βάση την ακριβή κατανομή της στατιστικής συνάρτησης M παραπέμπουμε τον/την αναγνώστη/στρια στους Laubscher *et al.* (1968) και τις εκεί αναφορές. Ο τρόπος σκέψης δίνεται στο παράδειγμα που ακολουθεί.

Παράδειγμα 6.17. Να προσδιορίσετε υπό τη μηδενική υπόθεση και τις υποθέσεις του ελέγχου, που προτάθηκε από τον Mood, την ακριβή κατανομή της στατιστικής συνάρτησης M, στην περίπτωση που $n_1 = 2$ και $n_2 = 3$, όταν είναι γνωστό ότι δεν υπάρχουν δεσμοί.

Λύση Παραδείγματος 6.17. Υπό τη μηδενική υπόθεση και τις υποθέσεις του ελέγχου, που προτάθηκε από τον Mood, όλες οι 10 διατάξεις των X_{11} και X_{12} είναι ισοπίθανες. Αυτές οι διατάξεις (αλλά και οι τάξεις, καθώς δεν υπάρχουν δεσμοί) είναι οι ακόλουθες:

$$(1,2),(1,3),(1,4),(1,5),(2,3),(2,4),(2,5),(3,4),(3,5),(4,5).$$

Χρησιμοποιώντας τη σχέση (6.23) προκύπτει, μετά από λίγη άλγεβρα, ότι οι αντίστοιχες τιμές της στατιστικής συνάρτησης M είναι: 5, 4, 5, 8, 1, 2, 5, 1, 4, 5. Επομένως, σε αυτήν την περίπτωση οι δυνατές τιμές της στατιστικής συνάρτησης M είναι οι: $\{1,2,4,5,8\}$ και έχει συνάρτηση πιθανότητας:

$$P(M=x)=\begin{cases} 2/10, \text{ για } x=1,4, \\ 4/10, \text{ για } x=5, \\ 1/10, \text{ για } x=8,2 \\ 0, \text{ αλλού.} \end{cases}$$

□

Παράδειγμα 6.18. Να προσδιορίσετε υπό τη μηδενική υπόθεση και τις υποθέσεις του ελέγχου, που προτάθηκε από τον Mood, την ακριβή κατανομή της στατιστικής συνάρτησης M^* στην περίπτωση που $n_1 = 2$ και $n_2 = 3$, όταν είναι γνωστό ότι υπάρχει ακριβώς ένας δεσμός μεταξύ της 2ης και 3ης σε διάταξη παρατήρησης.

Λύση Παραδείγματος 6.18. Υπό τη μηδενική υπόθεση και τις υποθέσεις του ελέγχου, που προτάθηκε από τον Mood, όλες οι 10 διατάξεις των X_{11} και X_{12} είναι ισοπίθανες. Αυτές οι διατάξεις είναι οι ακόλουθες:

$$(1,2),(1,3),(1,4),(1,5),(2,3),(2,4),(2,5),(3,4),(3,5),(4,5).$$

Σε αυτήν την περίπτωση, όμως, καθώς υπάρχει δεσμός μεταξύ της 2ης και 3ης σε διάταξη παρατήρησης, οι τάξεις των παρατηρήσεων του πρώτου δείγματος είναι:

$$(1,2.5),(1,2.5),(1,4),(1,5),(2.5,2.5),(2.5,4),(2.5,5),(2.5,4),(2.5,5),(4,5).$$

Χρησιμοποιώντας τη σχέση (6.24) προκύπτει, μετά από λίγη άλγεβρα, ότι οι αντίστοιχες τιμές της στατιστικής συνάρτησης M^* είναι: 4.25, 4.25, 5, 8, 0.5, 1.25, 4.25, 1.25, 4.25, 5. Επομένως, σε αυτήν την περίπτωση, οι δυνατές τιμές της στατιστικής συνάρτησης M είναι οι: $\{0.5,1.25,4.25,5,8\}$ και έχει συνάρτηση πιθανότητας:

$$P(M^*=x)=\begin{cases} 2/10, \text{ για } x=1.25,5, \\ 4/10, \text{ για } x=4.25, \\ 1/10, \text{ για } x=8,0.5, \\ 0, \text{ αλλού.} \end{cases}$$

Παρατηρούμε ότι είναι τελείως διαφορετική από την αντίστοιχη της στατιστικής συνάρτησης M που προσδιορίστηκε στο προηγούμενο παράδειγμα. □

Στην περίπτωση που το μέγεθος δείγματος είναι μεγάλο, χρησιμοποιείται η προσέγγιση της κατανομής των στατιστικών συναρτήσεων M και M^* από την κανονική κατανομή. Η προσέγγιση αυτή αποτελεί αντικείμενο μελέτης της επόμενης πρότασης.

Πρόταση 6.18. Έστω δύο πληθυσμοί με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1,2$, με πληθυσμιακές διακυμάνσεις σ_1^2 και σ_2^2, αντίστοιχα. Επιπλέον, υποθέτουμε ότι είναι διαθέσιμα δύο ανεξάρτητα μεταξύ τους τυχαία δείγματα από αυτούς τους πληθυσμούς, έστω $X_{11}, \dots, X_{1,n_1}$, και $X_{21}, \dots, X_{2,n_2}$, αντίστοιχα, με $n = n_1 + n_2$. Υπό την υπόθεση ότι οι δύο πληθυσμοί ταυτίζονται, ισχύει ότι:

$$\mathrm{Z} = \frac{M^* - \mathsf{E}(M^*)}{\sqrt{\mathsf{Var}(M_*)}} \xrightarrow{d} \mathcal{N}(0,1),$$

καθώς $n \to \infty$, με $\frac{n_1}{n_2}$ σταθερά, όπου

$$\mathsf{E}(M^*) = \frac{n_1}{n} \sum_{i=1}^{n} \left(r_i - \frac{n+1}{2} \right),$$

και

$$\mathsf{Var}(M^*) = \frac{n_1 n_2}{n^2(n-1)} \left\{ n \sum_{i=1}^{n} \left(r_i - \frac{n+1}{2} \right)^4 - \left[\sum_{i=1}^{n} \left(r_i - \frac{n+1}{2} \right)^2 \right]^2 \right\},$$

με r_i την τάξη της i-οστής παρατήρησης στο δείγμα όλων των $n = n_1 + n_2$ το πλήθος παρατηρήσεων.

Στην περίπτωση που δεν υπάρχουν δεσμοί είναι:

$$\mathrm{Z} = \frac{M - \mathsf{E}(M)}{\sqrt{\mathsf{Var}(M)}} \xrightarrow{d} \mathcal{N}(0,1),$$

καθώς $n \to \infty$, με $\frac{n_1}{n_2}$ σταθερά, όπου

$$\mathsf{E}(M) = \frac{n_1(n^2-1)}{12} \text{ και } \mathsf{Var}(M) = \frac{n_1 n_2 (n+1)(n^2-4)}{180}.$$

Απόδειξη Πρότασης 6.18. Παρατηρούμε ότι οι στατιστικές συναρτήσεις M^* και M ανήκουν στην οικογένεια των γραμμικών συναρτήσεων τάξεων, καθώς μπορούν να γραφτούν στη μορφή της σχέσης (6.1) με $a_i = (r_i - \frac{n+1}{2})^2$ και $a_i = (i - \frac{n+1}{2})^2$, αντίστοιχα. Τα προς απόδειξη αποτελέσματα προκύπτουν ως ειδικές περιπτώσεις της Πρότασης 6.3 και της Πρότασης 6.4, λαμβάνοντας, επιπλέον, υπόψη τις ακόλουθες σχέσεις:

$$\sum_{i=1}^{n} i = \frac{(n+1)}{2}, \quad \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6},$$

$$\sum_{i=1}^{n} i^3 = \frac{n^2(n+1)^2}{4}, \quad \sum_{i=1}^{n} i^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{180}.$$

□

Επομένως, συμβολίζοντας με $\theta = \frac{\sigma_1}{\sigma_2} > 0$, σε επίπεδο σημαντικότητας a, έχουμε τις ακόλουθες κρίσιμες περιοχές για τους αντίστοιχους ελέγχους:

i) $Z \geqslant z_a$ για τον έλεγχο της $H_0 : \theta = 1$ έναντι της $H_1 : \theta = \frac{\sigma_1}{\sigma_2} > 1$.

ii) $Z \leqslant -z_a$ για τον έλεγχο της $H_0 : \theta = 1$ έναντι της $H_1 : \theta = \frac{\sigma_1}{\sigma_2} < 1$.

iii) $|Z| \geqslant z_{a/2}$ για τον έλεγχο της $H_0 : \theta = 1$ έναντι της $H_1 : 0 < \theta = \frac{\sigma_1}{\sigma_2} \neq 1$.

6.5.1.2 Ο έλεγχος των Ansari-Bradley

Στην προηγούμενη ενότητα, διαπιστώσαμε ότι η στατιστική συνάρτηση που προτάθηκε από τον Mood για τον υπό μελέτη έλεγχο ανήκει στην οικογένεια των γραμμικών συναρτήσεων τάξεων με τους συντελεστές στάθμισης να είναι οι διαφορές της τάξης της i-οστής παρατήρησης από τον μέσο όρο των τάξεων υψωμένες στο τετράγωνο. Δηλαδή είναι $\mathrm{M}^* = \sum_{i=1}^{n} a_i Z_i$, με $a_i = \left(r_i - \frac{n+1}{2}\right)^2$ και Z_i να είναι η συνήθης δείκτρια μεταβλητή που λαμβάνει την τιμή 1, αν η i-οστή παρατήρηση στο αναμεμειγμένο δείγμα προέρχεται από τον πρώτο πληθυσμό, και την τιμή 0, αν προέρχεται από τον δεύτερο πληθυσμό. Αντίστοιχα, όταν δεν έχουμε δεσμούς, ισχύουν τα παραπάνω για τη στατιστική συνάρτηση M για $r_i = i$. Το τετράγωνο των διαφορών χρησιμοποιήθηκε θέλοντας να αποτελέσει ένα μέτρο της απόκλισης της τάξης κάθε παρατήρησης από τον μέσο όρο των τάξεων υπό την H_0. Προφανώς, μια άλλη επιλογή θα ήταν να χρησιμοποιηθεί η απόλυτη τιμή των διαφορών αντί για το τετράγωνο. Δηλαδή να χρησιμοποιηθεί η στατιστική συνάρτηση:

$$\mathrm{A} = \sum_{i=1}^{n} \left| i - \frac{n+1}{2} \right| Z_i.$$

Όπως επισημαίνουν οι Gibbons and Chakraborti (2020), πολλοί συγγραφείς έχουν προτείνει διάφορες εκδοχές που βασίζονται στην παραπάνω στατιστική συνάρτηση και που έχουν αποδειχθεί να είναι ισοδύναμες, καθώς η μία είναι γραμμική συνάρτηση της άλλης. Για περισσότερες λεπτομέρειες παραπέμπουμε τον/την ενδιαφερόμενο/μενη αναγνώστη/στρια στους Gibbons and Chakraborti (2020) και στις εκεί αναφορές. Στη συνέχεια, στην ενότητα αυτή θα παρουσιαστούν η στατιστική συνάρτηση και ο έλεγχος που προτάθηκε από τους Ansari and Bradley (1960).

Στο παραπάνω πλαίσιο, οι Ansari and Bradley (1960), υποθέτοντας ότι η διαφορά στις παραμέτρους θέσης (διαμέσους) των δύο πληθυσμών είναι γνωστή (θυμηθείτε ότι στον έλεγχο του Mood υποθέτουμε ότι είναι ίσες), πρότειναν στην περίπτωση μη ύπαρξης δεσμών, τη στατιστική συνάρτηση

$$\mathrm{AB} = \frac{n_1(n+1)}{2} - \sum_{i=1}^{n} \left| i - \frac{n+1}{2} \right| Z_i$$

που αποδεικνύεται εύκολα ότι ισοδύναμα γράφεται στη μορφή

$$\mathrm{AB} = \sum_{i=1}^{n} \left(\frac{n+1}{2} - \left| i - \frac{n+1}{2} \right| \right) Z_i. \tag{6.25}$$

Παρατηρούμε ότι η στατιστική συνάρτηση AB ανήκει στην οικογένεια των γραμμικών συναρτήσεων τάξεων, καθώς μπορεί να γραφτεί στη μορφή της σχέσης (6.1) με $a_i = \frac{n+1}{2} - |i - \frac{n+1}{2}|$. Επιπλέον, πρόκειται ουσιαστικά για το άθροισμα των τάξεων που αντιστοιχούν σε παρατηρήσεις που προέρχονται από τον πρώτο πληθυσμό όταν οι τάξεις υπολογίζονται με διαφορετικό τρόπο και, συγκεκριμένα, ως εξής: αποδίδεται η τιμή 1 στη μικρότερη και στη μεγαλύτερη τιμή στο μεικτό δείγμα, η τιμή 2 στην αμέσως μικρότερη και αμέσως μεγαλύτερη και συνεχίζοντας με τον ίδιο τρόπο η τιμή $n/2$ στις δύο μεσαίες παρατηρήσεις, όταν n άρτιος, και η τιμή $(n+1)/2$ στη μία μεσαία, όταν n περιττός.

Παρατήρηση 6.12. Οι Ansari and Bradley (1960) δεν πρότειναν κάποια τροποποίηση της στατιστικής συνάρτησης AB, όταν υπάρχουν δεσμοί μεταξύ των παρατηρήσεων, αναφέροντας ότι θα ήταν αρκετό να αντιμετωπιστεί το πρόβλημα χρησιμοποιώντας τα midranks. Προφανώς, στην περίπτωση αυτή θα πρέπει $a_i = \frac{n+1}{2} - |r_i - \frac{n+1}{2}|$.

Από τον τρόπο ορισμού της στατιστικής συνάρτησης AB γίνεται αντιληπτό ότι μικρές (μεγάλες) τιμές της υποδεικνύουν ότι η διακύμανση του πρώτου πληθυσμού είναι μεγαλύτερη (μικρότερη) από την αντίστοιχη

του πρώτου. Για τον προσδιορισμό των τιμών που θεωρούνται μικρές (ή μεγάλες ανάλογα), για να απορρίπτεται η μηδενική υπόθεση, απαιτείται ο προσδιορισμός της κατανομής της στατιστικής συνάρτησης AB υπό τη μηδενική υπόθεση. Αυτός ο προσδιορισμός είναι εφικτός για μικρά σε μέγεθος δείγματα και πίνακες κρίσιμων τιμών με βάση την ακριβή κατανομή της στατιστικής συνάρτησης AB, οι οποίοι έχουν δοθεί από τους Ansari and Bradley (1960) για τιμές των n_1, n_2 που είναι $n_2 = 2, 3, ..., 10$ και $4 \leqslant n_1 + n_2 \leqslant 20$. Για περισσότερες λεπτομέρειες και αναδρομικές σχέσεις για τον υπολογισμό πιθανοτήτων που συνδέονται με τη στατιστική συνάρτηση AB παραπέμπουμε στην εργασία των Ansari and Bradley (1960). Στο πλαίσιο αυτού του συγγράμματος, στο επόμενο παράδειγμα, αποσαφηνίζεται ο τρόπος σκέψης.

Παράδειγμα 6.19. Να προσδιορίσετε, υπό τη μηδενική υπόθεση και τις υποθέσεις του ελέγχου που προτάθηκε από τους Ansari-Bradley, την ακριβή κατανομή της στατιστικής συνάρτησης AB στην περίπτωση όπου $n_1 = 2$ και $n_2 = 3$, όταν είναι γνωστό ότι δεν υπάρχουν δεσμοί.

Λύση Παραδείγματος 6.19. Υπό τη μηδενική υπόθεση και τις υποθέσεις του ελέγχου που προτάθηκε από τους Ansari-Bradley, όλες οι 10 το πλήθος διατάξεις των X_{11} και X_{12} είναι ισοπίθανες. Αυτές οι διατάξεις (αλλά και οι τάξεις, καθώς δεν υπάρχουν δεσμοί) είναι οι ακόλουθες:

$$(1,2),(1,3),(1,4),(1,5),(2,3),(2,4),(2,5),(3,4),(3,5),(4,5).$$

Χρησιμοποιώντας τη σχέση (6.25) προκύπτει, μετά από λίγη άλγεβρα, ότι οι αντίστοιχες τιμές της στατιστικής συνάρτησης AB είναι: 3, 4, 3, 2, 5, 4, 3, 5, 4, 3. Επομένως, σε αυτήν την περίπτωση οι δυνατές τιμές της στατιστικής συνάρτησης AB είναι οι: $\{2,3,4,5\}$ και έχει συνάρτηση πιθανότητας

$$P(\mathrm{AB} = x) = \begin{cases} 4/10, \text{ για } x = 3, \\ 3/10, \text{ για } x = 4, \\ 2/10, \text{ για } x = 5, \\ 1/10, \text{ για } x = 2, \\ 0 \text{ αλλού.} \end{cases}$$

□

Στην περίπτωση που το μέγεθος δείγματος είναι μεγάλο, χρησιμοποιείται η προσέγγιση της κατανομής της στατιστικής συνάρτησης AB από την κανονική κατανομή. Η προσέγγιση αυτή αποτελεί αντικείμενο μελέτης της επόμενης πρότασης.

Πρόταση 6.19. Έστω δύο πληθυσμοί με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1,2$, με πληθυσμιακές διακυμάνσεις σ_1^2 και σ_2^2, αντίστοιχα. Επιπλέον, υποθέτουμε ότι είναι διαθέσιμα δύο ανεξάρτητα μεταξύ τους τυχαία δείγματα από αυτούς τους πληθυσμούς, έστω $X_{11}, \ldots, X_{1,n_1}$, και $X_{21}, \ldots, X_{2,n_2}$, αντίστοιχα, με $n = n_1 + n_2$. Υπό την υπόθεση ότι οι δύο πληθυσμοί ταυτίζονται, ισχύει ότι:

$$Z = \frac{\mathrm{AB} - \mathrm{E(AB)}}{\sqrt{\mathrm{Var(AB)}}} \to \mathcal{N}(0,1),$$

καθώς $n \to \infty$, με $\frac{n_1}{n_2}$ σταθερά, όπου

$$\mathrm{E(AB)} = \begin{cases} \frac{n_1(n_1+n_2+2)}{4}, \text{ για } n \text{ άρτιο} \\ \frac{n_1(n_1+n_2+1)}{4(n_1+n_2)}, \text{ για } n \text{ περιττό} \end{cases}$$

και

$$\mathrm{Var}(AB) = \begin{cases} \frac{n_1 n_2(n_1+n_2+2)(n_1+n_2-2)}{48(n_1+n_2-1)}, \text{ για } n \text{ άρτιο} \\ \frac{n_1 n_2(n_1+n_2+1)((n_1+n_2)^2+3)}{48(n_1+n_2)^2}, \text{ για } n \text{ περιττό} \end{cases}$$

Απόδειξη Πρότασης 6.19. Παρατηρούμε ότι η στατιστική συνάρτηση AB ανήκει στην οικογένεια των γραμμικών συναρτήσεων τάξεων, καθώς μπορεί να γραφτεί στη μορφή της σχέσης (6.1) με $a_i = \frac{n+1}{2} - |i - \frac{n+1}{2}|$. Το αποτέλεσμα στη συνέχεια προκύπτει ως ειδική περίπτωση της Πρότασης 6.3 και της Πρότασης 6.4, ύστερα από αρκετές αλγεβρικές πράξεις, και αφήνεται ως άσκηση για τον/την αναγνώστη/στρια. □

Επομένως, συμβολίζοντας με $\theta = \frac{\sigma_1}{\sigma_2} > 0$, σε επίπεδο σημαντικότητας a, έχουμε τις ακόλουθες κρίσιμες περιοχές για τους αντίστοιχους ελέγχους:

i) $Z \leqslant -z_a$ για τον έλεγχο της $H_0 : \theta = 1$ έναντι της $H_1 : \theta = \frac{\sigma_1}{\sigma_2} > 1$.

ii) $Z \geqslant z_a$ για τον έλεγχο της $H_0 : \theta = 1$ έναντι της $H_1 : \theta = \frac{\sigma_1}{\sigma_2} < 1$.

iii) $|Z| \geqslant z_{a/2}$ για τον έλεγχο της $H_0 : \theta = 1$ έναντι της $H_1 : 0 < \theta = \frac{\sigma_1}{\sigma_2} \neq 1$.

6.5.1.3 Ο έλεγχος των Siegel-Tukey

Ένας ακόμη έλεγχος για το υπό μελέτη πρόβλημα προτάθηκε από τους Siegel and Tukey (1960). Έστω δύο πληθυσμοί με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1,2$, με πληθυσμιακές διακυμάνσεις σ_1^2 και σ_2^2, αντίστοιχα. Επιπλέον, υποθέτουμε ότι είναι διαθέσιμα δύο ανεξάρτητα μεταξύ τους τυχαία δείγματα από αυτούς τους πληθυσμούς, έστω $X_{11}, \dots, X_{1,n_1}$, και $X_{21}, \dots, X_{2,n_2}$, αντίστοιχα, με $n = n_1 + n_2$. Ο έλεγχος, που προτάθηκε από τους Siegel and Tukey (1960), έχει ως κεντρική ιδέα την ιδιότητα ότι, αν οι δύο πληθυσμοί διαφέρουν μόνο κατά τη διασπορά τους (άρα και αυτός ο έλεγχος θεωρεί ότι οι παράμετροι θέσης είναι ίσες), τότε το δείγμα που προέρχεται από τον πληθυσμό με τη μεγαλύτερη διακύμανση θα είναι περισσότερο απλωμένο και με μεγαλύτερες ακραίες τιμές (μικρές και μεγάλες τιμές). Στη λογική αυτή, προτείνουν αρχικά να αναμειγνύονται τα δύο δείγματα και να διατάσσονται οι δειγματικές παρατηρήσεις σε αύξουσα τάξη μεγέθους. Έστω U_1,...,U_n το διατεταγμένο αυτό δείγμα των $n = n_1 + n_2$ το πλήθος παρατηρήσεων. Οι τάξεις στη συνέχεια για καθεμία από αυτές τις παρατηρήσεις υπολογίζονται με διαφορετικό τρόπο. Αποδίδουμε την τάξη 1 στην παρατήρηση U_1, την τάξη 2 στην παρατήρηση U_n, την τάξη 3 στην παρατήρηση U_{n-1}, την τάξη 4 στην παρατήρηση U_2, την τάξη 5 στην παρατήρηση U_3 και συνεχίζουμε κατά τον ίδιο τρόπο. Σχηματικά, όταν $n = 7$, είναι όπως φαίνεται παρακάτω:

U_1	U_2	U_3	U_4	U_5	U_6	U_7
1	4	5	7	6	3	2

Η στατιστική συνάρτηση, που προτάθηκε από τους Siegel-Tukey, είναι το άθροισμα των τάξεων που αντιστοιχούν στις παρατηρήσεις που προέρχονται από τον πρώτο πληθυσμό.

Ένα πρώτο εύλογο ερώτημα που προκύπτει είναι ποιος είναι ο λόγος της διαφοροποίησης στον τρόπο απόδοσης των τάξεων στις δειγματικές παρατηρήσεις. Όπως εξηγούν οι Siegel and Tukey (1960), αποδίδοντας τις τάξεις κατά αυτόν τον τρόπο, οι ακραίες παρατηρήσεις λαμβάνουν τις μικρότερες τάξεις, ενώ οι υψηλότερες τάξεις αποδίδονται στο μέσο της ακολουθίας των παρατηρήσεων. Αν η μηδενική υπόθεση ότι οι πληθυσμοί ταυτίζονται είναι αληθής, οι παρατηρήσεις των δύο πληθυσμών θα αναμειγνύονται καλά και αυτό θα έχει ως συνέπεια ο μέσος όρος των τάξεων που αποδόθηκαν στο πρώτο δείγμα να είναι περίπου ίσος με τον αντίστοιχο μέσο όρο των τάξεων στο δεύτερο δείγμα. Αν από την άλλη πλευρά, η εναλλακτική υπόθεση ότι οι πληθυσμοί διαφέρουν ως προς τις διασπορές είναι αληθής, αναμένεται περισσότερες παρατηρήσεις από τον πληθυσμό με τη μεγαλύτερη διακύμανση να είναι στα άκρα της διατεταγμένης ακολουθίας και, επομένως, να έχουν μικρότερες τάξεις, ενώ αναμένεται οι περισσότερες εκ των παρατηρήσεων από τον πληθυσμό με τη μικρότερη διακύμανση να είναι κοντά στο μέσο της διατεταγμένης ακολουθίας και, συνεπώς, να αποδίδονται σε αυτές οι υψηλότερες τάξεις. Οπότε, αναμένεται ο μέσος όρος των τάξεων των παρατηρήσεων από τον πληθυσμό με τη μεγαλύτερη διακύμανση να είναι μικρότερος από τον αντίστοιχο μέσο όρο των τάξεων των παρατηρήσεων από τον πληθυσμό με τη

μικρότερη διακύμανση. Όλα τα παραπάνω οδηγούν στη συνέχεια να χρησιμοποιούνται για τον έλεγχο οι πίνακες του ελέγχου των Wilcoxon-Mann-Whitney για τον έλεγχο διαφορών στις διαμέσους.

Συνοψίζοντας, αν R_1 είναι η στατιστική συνάρτηση που παριστάνει το άθροισμα των τάξεων των παρατηρήσεων που προέρχονται από τον πρώτο πληθυσμό, απορρίπτεται, σε επίπεδο σημαντικότητας a, η μηδενική υπόθεση ότι οι πληθυσμοί ταυτίζονται έναντι της εναλλακτικής i) $\sigma_1^2 > \sigma_2^2$, αν $R_1 \leqslant w_{1-a}$, έναντι της εναλλακτικής ii) $\sigma_1^2 < \sigma_2^2$, αν $R_1 \geqslant w_a$, και έναντι της εναλλακτικής iii) $\sigma_1^2 \neq \sigma_2^2$, αν $R_1 \geqslant w_{a/2}$ ή $R_1 \leqslant w_{1-a/2}$, με τις κρίσιμες τιμές να προκύπτουν από τον κλασικό έλεγχο του Wilcoxon (βλ. Ενότητα 6.3). Για την προσέγγιση της κατανομής της στατιστικής συνάρτησης R_1 από την κανονική κατανομή ισχύουν όσα έχουν αναφερθεί για την προσέγγιση της κατανομής της στατιστικής συνάρτησης του ελέγχου του Wilcoxon για δύο ανεξάρτητα δείγματα και παραπέμπουμε τον/την αναγνώστη/στρια στην αντίστοιχη ενότητα.

Τέλος, επισημαίνεται (βλ. Gibbons and Chakraborti, 2020) ότι η στατιστική συνάρτηση των Siegel and Tukey (1960) μπορεί να γραφτεί στη μορφή των γραμμικών συναρτήσεων τάξης της σχέσης (6.1), με τους συντελεστές στάθμισης a_i να προσδιορίζονται από τη σχέση:

$$a_i = \begin{cases} 2i, \text{ για } i \text{ άρτιο } 1 < i \leqslant n/2, \\ 2i-1, \text{ για } i \text{ περιττό } 1 \leqslant i \leqslant n/2, \\ 2(n-i)+2, \text{ για } i \text{ άρτιο } n/2 < i \leqslant n, \\ 2(n-i)+1, \text{ για } i \text{ περιττό } n/2 < i \leqslant n. \end{cases}$$

Παρατήρηση 6.13. Η εφαρμογή του ελέγχου δεν είναι ικανοποιητική, αν, επιπλέον, οι δύο πληθυσμοί διαφέρουν ως προς τη μέση τιμή (ή τη διάμεσο). Σε αυτήν την περίπτωση, όπως και στους ελέγχους που μέχρι τώρα έχουν παρουσιαστεί σε αυτήν την ενότητα, μετασχηματίζουμε τις παρατηρήσεις του δείγματος από τον πληθυσμό με τη μεγαλύτερη μέση τιμή, αφαιρώντας από καθεμία από αυτές τη διαφορά των μέσων. Ωστόσο, στην πράξη ούτε οι πληθυσμιακές μέσες τιμές είναι γνωστές ούτε η διαφορά τους. Για να αντιμετωπιστεί το παραπάνω πρόβλημα, έχει προταθεί στη βιβλιογραφία, μεταξύ άλλων, ο έλεγχος που θα παρουσιαστεί στην επόμενη ενότητα.

6.5.1.4 Ο έλεγχος των Conover-Iman

Έστω δύο πληθυσμοί με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1,2$, με πληθυσμιακές διακυμάνσεις σ_1^2 και σ_2^2, αντίστοιχα. Επιπλέον, υποθέτουμε ότι είναι διαθέσιμα δύο ανεξάρτητα μεταξύ τους τυχαία δείγματα από αυτούς τους πληθυσμούς, έστω $X_{11}, \ldots, X_{1,n_1}$, και $X_{21}, \ldots, X_{2,n_2}$, αντίστοιχα, με $n = n_1 + n_2$. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση

$$H_0 : \sigma_1^2 = \sigma_2^2$$

έναντι μίας εκ των τριών εναλλακτικών υποθέσεων:

(i) $H_1 : \sigma_1^2 \neq \sigma_2^2$,
(ii) $H_1 : \sigma_1^2 > \sigma_2^2$

και

(iii) $H_1 : \sigma_1^2 < \sigma_2^2$.

Δηλαδή, εν αντιθέσει με τους προηγούμενους ελέγχους, δεν έχουμε κάποια πληροφορία ή υπόθεση για τις παραμέτρους θέσης ή τη διαφορά αυτών.

Οι Conover and Iman (1978), αρχικά, επισημαίνουν ότι, χρησιμοποιώντας τον ορισμό της διακύμανσης, οι παραπάνω έλεγχοι ανάγονται στον έλεγχο της

$$H_0 : \mathrm{E}(|X_1 - \mu_1|)^2 = \mathrm{E}(|X_2 - \mu_2|)^2$$

έναντι μίας εκ των τριών εναλλακτικών υποθέσεων:

(i) $H_1 : \mathrm{E}(|X_1 - \mu_1|)^2 \neq \mathrm{E}(|X_2 - \mu_2|)^2$,
(ii) $H_1 : \mathrm{E}(|X_1 - \mu_1|)^2 > \mathrm{E}(|X_2 - \mu_2|)^2$

και

(iii) $H_1 : \mathrm{E}(|X_1 - \mu_1|)^2 < \mathrm{E}(|X_2 - \mu_2|)^2$,

όπου $\mu_i = \mathrm{E}(X_i)$ η μέση τιμή του i–οστού πληθυσμού, $i = 1,2$. Έπειτα, προτείνουν τον σχηματισμό των διαφορών $\left|X_{ij} - \mu_i\right|$, για $i = 1,2$, $j = 1,...,n_i$, στην περίπτωση που οι μέσες τιμές των δύο πληθυσμών είναι γνωστές.

Παρατήρηση 6.14. Αν οι πληθυσμιακές μέσες τιμές είναι άγνωστες, κάτι που είναι σύνηθες στην πράξη, τότε οι Conover and Iman (1978) προτείνουν να σχηματίζουμε τις διαφορές $U_{ij} = \left|X_{ij} - \bar{X}_i\right|$, για $i = 1,2$, $j = 1,...,n_i$, όπου $\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$. Όμως, σε μια τέτοια περίπτωση, επισημαίνουν ότι η ακριβής κατανομή του ελέγχου εξαρτάται από τον αληθινό πληθυσμό. Επομένως, σε αυτήν την περίπτωση, η διαδικασία δεν είναι απαλλαγμένη παραμέτρων (*distribution free*).

Έστω $U_1,...,U_n$ οι διατεταγμένες απόλυτες διαφορές που σχηματίστηκαν. Οι Conover and Iman (1978) προτείνουν τη στατιστική συνάρτηση

$$C = \sum_{i=1}^{n} R(U_i)^2 Z_i, \tag{6.26}$$

όπου Z_i είναι η συνήθης δείκτρια μεταβλητή που λαμβάνει την τιμή 1, αν η i-οστή παρατήρηση στο αναμεμειγμένο δείγμα προέρχεται από τον πρώτο πληθυσμό, και την τιμή 0, αν προέρχεται από τον δεύτερο πληθυσμό.

Χρησιμοποιώντας τη στατιστική συνάρτηση C απορρίπτεται, σε επίπεδο σημαντικότητας a, η μηδενική υπόθεση ότι οι πληθυσμοί έχουν ίδιες διακυμάνσεις έναντι της εναλλακτικής i) $\sigma_1^2 > \sigma_2^2$, αν $C \geqslant c_a$, έναντι της εναλλακτικής ii) $\sigma_1^2 < \sigma_2^2$, αν $C \leqslant c_{1-a}$, και έναντι της εναλλακτικής iii) $\sigma_1^2 \neq \sigma_2^2$, αν $C \geqslant c_{a/2}$ ή $C \leqslant c_{1-a/2}$, με c_a να είναι τέτοιο, ώστε $P(C > c_a|H_0) \leqslant a \leqslant P(C \leqslant c_a|H_0)$. Επειδή ο έλεγχος χρησιμοποιεί το τετράγωνο των τάξεων μεγέθους, στη βιβλιογραφία έχει επικρατήσει, ο έλεγχος αυτός να αναφέρεται ως έλεγχος των τετραγώνων τάξεων μεγέθους (Squared Rank Test).

Για τον προσδιορισμό των τιμών που θεωρούνται μικρές (ή μεγάλες ανάλογα), για να απορρίπτεται η μηδενική υπόθεση, απαιτείται ο προσδιορισμός της κατανομής της στατιστικής συνάρτησης C υπό τη μηδενική υπόθεση. Υποθέτοντας ότι οι πληθυσμιακές μέσες τιμές είναι γνωστές και ότι δεν υπάρχουν δεσμοί μεταξύ των παρατηρήσεων, οι Conover and Iman (1978) δίνουν πίνακες κρίσιμων τιμών για τιμές $n_1 = 3,...,9$, $n_2 = n_1,...,20$ και $n_1 = 10$, $n_2 = 10,...,15$. Ο τρόπος σκέψης για τον υπολογισμό τους δίνεται στο παράδειγμα που ακολουθεί, ενώ τμήμα αυτών των αποτελεσμάτων αποτελεί ο Πίνακας Π.27 του Παραρτήματος.

Παράδειγμα 6.20. Να προσδιορίσετε, υπό τη μηδενική υπόθεση, την ακριβή κατανομή της στατιστικής συνάρτησης C, στην περίπτωση που $n_1 = 3$ και $n_2 = 3$, όταν δεν υπάρχουν δεσμοί μεταξύ των δειγματικών παρατηρήσεων.

Λύση Παραδείγματος 6.20. Υπό τη μηδενική υπόθεση και τις υποθέσεις του ελέγχου, ο οποίος έλεγχος προτάθηκε από τους Conover and Iman (1978), όλες οι 20 πιθανές διατάξεις των παρατηρήσεων από τον πρώτο πληθυσμό στο διατεταγμένο δείγμα των U_i είναι ισοπίθανες. Αυτές οι διατάξεις (και οι τάξεις, καθώς δεν υπάρχουν δεσμοί) είναι οι ακόλουθες:

$$(1,2,3),(1,2,4),(1,2,5),(1,2,6),(1,3,4),(1,3,5),(1,3,6),(1,4,5),(1,4,6),(1,5,6),(2,3,4),$$

$$(2,3,5),(2,3,6),(2,4,5),(2,4,6),(2,5,6),(3,4,5),(3,4,6),(3,5,6),(4,5,6).$$

Χρησιμοποιώντας τη σχέση (6.26) προκύπτει, μετά από λίγη άλγεβρα, ότι οι τιμές της στατιστικής συνάρτησης C είναι:

$$14,\ 21,\ 30,\ 41,\ 26,\ 35,\ 46,\ 42,\ 53,\ 62,\ 29,\ 38,\ 49,\ 45,\ 56,\ 65,\ 50,\ 61,\ 70,\ 77,$$

αντίστοιχα. Επομένως, είναι, για παράδειγμα, $P(C \leqslant 14) = 1/20 = 0.05$, $P(C \leqslant 21) = 2/20 = 0.10$, $P(C \geqslant 70) = 2/20 = 0.1$ και $P(C \geqslant 77) = 1/20 = 0.05$. Οι παραπάνω τιμές συμφωνούν απόλυτα με τις τιμές που δόθηκαν από τους Conover and Iman (1978) για τη συγκεκριμένη ειδική περίπτωση. □

Παράδειγμα 6.21. Να προσδιορίσετε, υπό τη μηδενική υπόθεση, την ακριβή κατανομή της στατιστικής συνάρτησης C, στην περίπτωση που $n_1 = 3$ και $n_2 = 3$, όταν είναι γνωστό ότι υπάρχει δεσμός μεταξύ της 1ης και 2ης διατεταγμένης παρατήρησης του τροποποιημένου δείγματος.

Λύση Παραδείγματος 6.21. Υπό τη μηδενική υπόθεση και τις υποθέσεις του ελέγχου, ο οποίος έλεγχος προτάθηκε από τους Conover and Iman (1978), όλες οι 20 το πλήθος πιθανές διατάξεις των παρατηρήσεων από τον πρώτο πληθυσμό στο διατεταγμένο δείγμα των U_i, $i = 1, ..., n$ είναι ισοπίθανες. Αυτές οι διατάξεις είναι οι ακόλουθες:

$$(1,2,3),(1,2,4),(1,2,5),(1,2,6),(1,3,4),(1,3,5),(1,3,6),(1,4,5),(1,4,6),(1,5,6),(2,3,4),$$

$$(2,3,5),(2,3,6),(2,4,5),(2,4,6),(2,5,6),(3,4,5),(3,4,6),(3,5,6),(4,5,6).$$

Καθώς υπάρχουν δεσμοί μεταξύ της πρώτης και της δεύτερης παρατήρησης θα έχουμε τους ακόλουθους δεσμούς:

$$(1.5,1.5,3),(1.5,1.5,4),(1.5,1.5,5),(1.5,1.5,6),(1.5,3,4),(1.5,3,5),(1.5,3,6),(1.5,4,5),$$

$$(1.5,4,6),(1.5,5,6),(1.5,3,4),$$

$$(1.5,3,5),(1.5,3,6),(1.5,4,5),(1.5,4,6),(1.5,5,6),(3,4,5),(3,4,6),(3,5,6),(4,5,6).$$

Χρησιμοποιώντας τη σχέση (6.26) προκύπτει, μετά από λίγη άλγεβρα, ότι οι αντίστοιχες τιμές της στατιστικής συνάρτησης C είναι:

$$13.5,20.5,29.5,40.5,27.25,36.25,47.25,43.25,54.25,63.25,$$

$$27.25,36.25,47.2543.25,54.25,63.25,50,61,70,77.$$

Επομένως, μπορούμε άμεσα να υπολογίσουμε την αθροιστική συνάρτηση κατανομής, υπό την H_0, για τη στατιστική συνάρτηση ελέγχου C. Για παράδειγμα,

$$P(C \leqslant 13.5) = 1/20 = 0.05,\ P(C \leqslant 20.5) = 2/20 = 0.10,$$

ή

$$P(C \geqslant 70) = 2/20 = 0.1,\ P(C \geqslant 77) = 1/20 = 0.05.$$

Παρατηρούμε ότι η συνάρτηση πιθανότητας της στατιστικής συνάρτησης C είναι σε αυτήν την περίπτωση τελείως διαφορετική από την αντίστοιχη που προσδιορίστηκε στο προηγούμενο παράδειγμα. □

Στην περίπτωση που το μέγεθος δείγματος είναι μεγάλο, χρησιμοποιείται η προσέγγιση της κατανομής της στατιστικής συνάρτησης από την κανονική κατανομή. Η προσέγγιση αυτή, η οποία αποδεικνύεται ότι είναι ικανοποιητική για $n_1, n_2 \geqslant 10$, αποτελεί αντικείμενο μελέτης της επόμενης πρότασης.

Πρόταση 6.20. Έστω δύο πληθυσμοί με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1,2$, με πληθυσμιακές διακυμάνσεις σ_1^2 και σ_2^2, αντίστοιχα. Επιπλέον, υποθέτουμε ότι είναι διαθέσιμα δύο ανεξάρτητα μεταξύ τους τυχαία δείγματα από αυτούς τους πληθυσμούς, έστω $X_{11}, \dots, X_{1n_1}$, και $X_{21}, \dots, X_{2n_2}$, αντίστοιχα, με $n = n_1 + n_2$. Έστω $U_1,\dots,U_n$ είναι το διατεταγμένο δείγμα που προκύπτει υπολογίζοντας τις διαφορές $\left|X_{ij} - \mu_i\right|$, για $i = 1,2$, $j = 1,\dots,n_i$. Υπό την υπόθεση ότι οι δύο πληθυσμοί έχουν ίδιες διακυμάνσεις, ισχύει ότι

$$Z = \frac{C - \mathrm{E}(C)}{\sqrt{\mathrm{Var}(C)}} \xrightarrow{d} \mathcal{N}(0,1),$$

καθώς $n \to \infty$, με $\frac{n_1}{n_2}$ σταθερά, όπου όταν υπάρχουν δεσμοί μεταξύ των παρατηρήσεων είναι

$$\mathrm{E}(C) = \frac{n_1}{n} \sum_{i=1}^{n} R(U_i)^2,$$

και

$$\mathrm{Var}(C) = \frac{n_1 n_2}{n^2(n-1)} \left\{ n \sum_{i=1}^{n} (R(U_i))^4 - \left[\sum_{i=1}^{n} (R(U_i))^2 \right]^2 \right\},$$

με $R(U_i)$ την τάξη της i-οστής παρατήρησης στο δείγμα όλων των $n = n_1 + n_2$ παρατηρήσεων.

Στην περίπτωση που δεν υπάρχουν δεσμοί είναι:

$$\mathrm{E}(C) = \frac{n_1(n+1)(2n+1)}{6} \text{ και } \mathrm{Var}(C) = \frac{n_1 n_2 (n+1)(2n+1)(8n+11)}{180}.$$

Απόδειξη Πρότασης 6.20. Παρατηρούμε ότι η στατιστική συνάρτηση C ανήκει στην οικογένεια των γραμμικών συναρτήσεων τάξεων, καθώς μπορεί να γραφτεί στη μορφή της σχέσης (6.1) με $a_i = R(U_i)^2$, ενώ στην περίπτωση μη ύπαρξης δεσμών είναι $a_i = i^2$. Τα προς απόδειξη αποτελέσματα προκύπτουν ως ειδικές περιπτώσεις της Πρότασης 6.3 και της Πρότασης 6.4, λαμβάνοντας, επιπλέον, υπόψη τις ακόλουθες σχέσεις

$$\sum_{i=1}^{n} i = \frac{(n+1)}{2}, \quad \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum_{i=1}^{n} i^3 = \frac{n^2(n+1)^2}{4}, \quad \sum_{i=1}^{n} i^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1}{180}.$$

□

Επομένως, σε επίπεδο σημαντικότητας a, η μηδενική υπόθεση $H_0 : \sigma_1^2 = \sigma_2^2$ απορρίπτεται έναντι της εναλλακτικής i) $H_1 : \sigma_1^2 > \sigma_2^2$, αν $Z \geqslant z_a$, έναντι της εναλλακτικής ii) $H_1 : \sigma_1^2 < \sigma_2^2$, αν $Z \leqslant -z_a$ ή έναντι της εναλλακτικής iii) $H_1 : \sigma_1^2 \neq \sigma_2^2$, αν $|Z| \geqslant z_{a/2}$.

Παρατήρηση 6.15. Προηγουμένως, μετασχηματίστηκαν τα δύο αρχικά δείγματα και διατάχθηκαν όλες οι παρατηρήσεις κατά αύξουσα τάξη μεγέθους σύμφωνα με τις απόλυτες διαφορές $\left|X_{ij} - \mu_i\right|$, για $i = 1,2$, $j = 1,\dots,n_i$. Η διάταξη αυτή ισοδυναμεί με τη διάταξη κατά αύξουσα τάξη μεγέθους των $\left(X_{ij} - \mu_i\right)^2$, για $i = 1,2$, $j = 1,\dots,n_i$, αλλά προτιμήθηκε γιατί είναι πιο εύκολοι οι υπολογισμοί.

Παράδειγμα 6.22. Για να ελέγξει κατά πόσο δύο διαφορετικές μέθοδοι διδασκαλίας οδηγούν σε απόδοση των μαθητών με διαφορετικού βαθμού διακυμάνσεις, ένας καθηγητής Μαθηματικών εφαρμόζει αυτές τις μεθόδους σε δύο ομάδες μαθητών της ίδιας περίπου ικανότητας στα Μαθηματικά. Η ομάδα 1, αποτελούμενη από 5 μαθητές, διδάσκεται το μάθημα μέσω των παραδόσεων και ενός συμβατικού βιβλίου (μέθοδος Α). Η ομάδα 2,

αποτελούμενη από 6 μαθητές, διδάσκεται το μάθημα μέσω της χρήσης ενός υπολογιστικού πακέτου (μέθοδος Β). Οι εξετάσεις, στο τέλος της χρονιάς, έδωσαν τα εξής αποτελέσματα:

Ομάδα 1 (μέθοδος Α):	14	11	10	8	7	
Ομάδα 2 (μέθοδος Β):	18	17	15	13	12	6

Θα μπορούσε να συμπεράνει ο καθηγητής ότι υπάρχει στατιστικά σημαντική διαφορά μεταξύ των διακυμάνσεων των αποδόσεων των δύο ομάδων μαθητών; Να ελέγξετε την υπόθεση σε ε.σ. $a = 0.05$

(α) χρησιμοποιώντας τον έλεγχο των Siegel-Tukey,
(β) χρησιμοποιώντας τον έλεγχο των Conover-Iman.

Λύση Παραδείγματος 6.22. Θεωρούμε την τυχαία μεταβλητή X_1, η οποία περιγράφει την απόδοση στα μαθηματικά με τη μέθοδο Α διδασκαλίας, και την τυχαία μεταβλητή X_2, η οποία παριστάνει την απόδοση των στα μαθηματικά με τη μέθοδο Β διδασκαλίας. Διαθέτουμε δύο τυχαία δείγματα $X_{11}, \dots, X_{15}$ και $X_{21}, \dots, X_{26}$, ανεξάρτητα μεταξύ τους, και μας ενδιαφέρει να ελέγξουμε το πρόβλημα

$$H_0 : \sigma_1^2 = \sigma_2^2 \,, \;\; H_1 : \sigma_1^2 \neq \sigma_2^2$$

όπου $\sigma_1^2 = \text{Var}(X_1)$ και $\sigma_2^2 = \text{Var}(X_2)$.

(α) Ο έλεγχος των Siegel-Tukey υποθέτει ότι τα δύο δείγματα προέρχονται από πληθυσμούς που έχουν την ίδια μέση τιμή. Στις εφαρμογές, συνήθως, υπολογίζουμε τους δειγματικούς μέσους και χρησιμοποιούμε τη διαφορά αυτών ως εκτίμηση της διαφοράς των πληθυσμιακών μέσων. Στο παράδειγμά μας, $\bar{X}_1 = 10$ και $\bar{X}_2 = 13.5$, επομένως $\bar{X}_2 - \bar{X}_1 = 3.5 > 0$ και αυτή τη διαφορά την αφαιρούμε από το δείγμα με τη μεγαλύτερη μέση τιμή (ελέγξτε ότι η πληθυσμιακή μέση τιμή του δεύτερου πληθυσμού είναι στατιστικά σημαντικά μεγαλύτερη από την αντίστοιχη του πρώτου ή ότι η διαφορά τους είναι ίση με 3.5 μονάδες). Για τον λόγο αυτό, από τις τιμές στην Ομάδα 2, αφαιρούμε το 3.5. Οπότε προκύπτουν τα εξής δείγματα:

X_1	:	14	11	10	8	7	
$X_2^{(*)}$	:	14.5	13.5	11.5	9.5	8.5	2.5.

Διατάσσουμε από κοινού τα δύο δείγματα κατά αύξουσα τάξη μεγέθους και υπολογίζουμε τις τάξεις σε αυτές, σύμφωνα με τη θεωρία που αναφέρθηκε για τον έλεγχο των Siegel-Tukey.

2.5	7	8	8.5	9.5	10	11	11.5	13.5	14	14.5
1	4	5	8	9	11	10	7	6	3	2

Υπολογίζουμε την τιμή της στατιστικής συνάρτησης για τον έλεγχο των Siegel-Tukey, δηλαδή

$$R_1 = \sum_{i=1}^{5} R(X_{1i}) = 4 + 5 + 11 + 10 + 3 = 33.$$

Από τον Πίνακα Π.23 του Παραρτήματος που αφορά τα ποσοστιαία σημεία της στατιστικής συνάρτησης R_1 για τον έλεγχο Wilcoxon-Mann-Whitney τεστ, προκύπτουν ότι τα ποσοστιαία σημεία για $n_1 = 5$ και $n_2 = 6$, όταν $a = 0.05$, για το δίπλευρο πρόβλημα ελέγχου είναι, $w_{0.025} = 19$ και $w_{0.975} = n_1(n_1 + n_2 + 1) - w_{0.025} = 5(5 + 6 + 1) - 19 = 41$. Θα πρέπει να αναφέρουμε ότι τα ποσοστιαία σημεία $w_{0.025}$, $w_{0.0975}$ είναι αντίστοιχα τα κάτω και άνω 0.025 ποσοστιαία σημεία της κατανομής της σ.σ.ε. R_1, όταν η H_0 είναι αληθής. Επειδή $R_1 \not\geq 41$ και $R_1 \not\leq 19$, συμπεραίνουμε ότι δεν μπορούμε να απορρίψουμε την H_0, δηλαδή δεν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των διακυμάνσεων των αποδόσεων των δύο ομάδων μαθητών σε ε.σ. 5%. Υπενθυμίζεται ότι (βλ. Πρόταση 6.12) ότι $w_{1-p} = n_1(n+1) - w_p$, με $n = n_1 + n_2$.

(β) Για τον έλεγχο των Conover-Iman, πρέπει να κάνουμε τους μετασχηματισμούς, $U_{ij} = \left|X_{ij} - \bar{X}_i\right|$, για $i = 1,2,\ j = 1,...,n_i$.

$$\begin{array}{lcccccc} U_1 & : & 10 & 1 & 0 & 2 & 3 \\ U_2 & : & 4.5 & 3.5 & 1.5 & 0.5 & 1.5 & 7.5 \end{array}$$

Διατάσσουμε από κοινού τα δύο δείγματα κατά αύξουσα τάξη μεγέθους και βρίσκουμε την τάξη κάθε (μετασχηματισμένης) παρατήρησης, όπως φαίνεται παρακάτω:

$$\begin{array}{cccccccccccc} 0 & 0.5 & 1 & 1.5 & 1.5 & 2 & 3 & 3.5 & 4.5 & 7.5 & 10 \\ 1 & 2 & 3 & 4.5 & 4.5 & 6 & 7 & 8 & 9 & 10 & 11 \end{array}$$

Υπολογίζουμε την τιμή της στατιστικής συνάρτησης για τον έλεγχο των Conover-Iman, δηλαδή

$$C = \sum_{i=1}^{5} R(U_1)^2 = 1^2 + 3^2 + 6^2 + 7^2 + 10^2 = 195.$$

Από τον Πίνακα Π.27 του Παραρτήματος προκύπτουν ότι τα ποσοστιαία σημεία για $n_1 = 5$ και $n_2 = 6$, όταν $a = 0.05$, για το δίπλευρο πρόβλημα ελέγχου είναι, $w_{0.025} = 363$ και $w_{0.975} = 100$. Επειδή, $w_{0.975} < C < w_{0.025}$, συμπεραίνουμε ότι δεν μπορούμε να απορρίψουμε την H_0, δηλαδή δεν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των διακυμάνσεων των αποδόσεων των δύο ομάδων μαθητών σε ε.σ. 5%.

□

6.5.2 Έλεγχος ισότητας περισσότερων των δύο πληθυσμιακών διακυμάνσεων

Έστω k το πλήθος πληθυσμοί, $k \geq 3$, με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1,2,...,k$, $k \geq 3$, και με πληθυσμιακές διακυμάνσεις σ_i^2, $i = 1,2,...,k$, $k \geqslant 3$. Επιπλέον, υποθέτουμε ότι είναι διαθέσιμα k το πλήθος ανεξάρτητα μεταξύ τους τυχαία δείγματα από αυτούς του πληθυσμούς, έστω $X_{i1},...,X_{1n_i}$, $i = 1,2,...,k$, $k \geqslant 3$.

Στο πλαίσιο αυτό, θέλουμε να ελέγξουμε τη μηδενική υπόθεση

$$H_0 : \text{οι } k \text{ πληθυσμοί ταυτίζονται εκτός ίσως από πιθανές διαφορετικές μέσες τιμές,}$$

έναντι της εναλλακτικής υπόθεσης $H_1 : \sigma_l^2 \neq \sigma_m^2$, όπου $l,m = 1,...,k, l \neq m$, $k \geqslant 3$. Στο παρόν σύγγραμμα θα παρουσιαστεί μόνο η γενίκευση του ελέγχου των Conover and Iman (1978), χωρίς να υπεισέλθουμε σε πολλές λεπτομέρειες, καθώς το σκεπτικό είναι παρόμοιο με αυτό που προηγήθηκε. Επιπρόσθετα, η παρουσίαση θα γίνει για την περίπτωση που οι πληθυσμιακές μέσες τιμές είναι άγνωστες.

Αρχικά, σχηματίζουμε τις διαφορές $\left(X_{ij} - \bar{X}_i\right)^2$ ή τις διαφορές $\left|X_{ij} - \bar{X}_i\right|$ για $i = 1,2,...,k$, $k \geqslant 3$, $j = 1,...,n_i$, όπου $\bar{X}_i$, $i = 1,2,...,k$, $k \geqslant 3$, συμβολίζεται ο συνήθης δειγματικός μέσος για το δείγμα από τον i-οστό πληθυσμό, δηλαδή $\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$, $i = 1,...,k$, $k \geqslant 3$. Στη συνέχεια, οι $n = n_1 + n_2 + ... + n_k$, $k \geqslant 3$, το πλήθος αυτές διαφορές αναμειγνύονται και διατάσσονται κατά αύξουσα τάξη. Αν συμβολίσουμε τις τάξεις αυτών ως $\tilde{R}\left(X_{ij}\right)$, η στατιστική συνάρτηση που χρησιμοποιείται για τον έλεγχο είναι η

$$T = \frac{\sum_{i=1}^{k}\left(T_i^2/n_i\right) - n\overline{T}^2}{V_T},$$

όπου

$$T_i = \sum_{j=1}^{n_i} \tilde{R}\left(X_{ij}\right)^2, \overline{T} = n^{-1}\sum_{i=1}^{k} T_i \text{ και } V_T = (n-1)^{-1}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} \tilde{R}\left(X_{ij}\right)^4 - n\overline{T}^2\right).$$

Υπό τη μηδενική υπόθεση, αποδεικνύεται (βλ., μεταξύ άλλων, Conover, 1998; Kvam and Vidakovic, 2007) ότι:

$$T = \frac{\sum_{i=1}^{k}\left(T_i^2/n_i\right) - n\overline{T}^2}{V_T} \xrightarrow{d} \chi^2_{k-1}.$$

Επομένως, η μηδενική υπόθεση απορρίπτεται, σε επίπεδο σημαντικότητας a, όταν $T \geq \chi^2_{k-1,a}$.

Σε περίπτωση που η μηδενική υπόθεση απορρίπτεται, θέλοντας να εντοπίσουμε σε ποιους πληθυσμούς έχουμε άνισες διακυμάνσεις προχωρούμε σε έλεγχο πολλαπλών συγκρίσεων, δηλαδή στους επιμέρους ελέγχους της $H_0 : \sigma_l^2 = \sigma_m^2$, όπου $l, m = 1, ..., k$, $k \geq 3$, με $l \neq m$, χρησιμοποιώντας τη στατιστική συνάρτηση:

$$t = \frac{\frac{T_l}{n_l} - \frac{T_m}{n_m}}{\sqrt{\left(\frac{1}{n_l} + \frac{1}{n_m}\right) V_T \left(\frac{n-1-T}{n-k}\right)}}$$

και κρίσιμη περιοχή την $|t| \geq t_{n-k,a/2}$.

Παρατήρηση 6.16. Στην περίπτωση της μη ύπαρξης δεσμών η παραπάνω στατιστική συνάρτηση απλοποιείται λαμβάνοντας υπόψη ότι:

$$\overline{T} = (n+1)(2n+1)/6,$$

και

$$V_T = n(n+1)(2n+1)(8n+11)/180.$$

6.6 Έλεγχος ισότητας δύο πληθυσμιακών διακυμάνσεων με εξαρτημένα δείγματα

Στην προηγούμενη ενότητα, το ενδιαφέρον επικεντρώθηκε στην παρουσίαση μεθοδολογιών που βασίζονται στις τάξεις των παρατηρήσεων για τον έλεγχο της ισότητας δύο ή περισσότερων πληθυσμιακών διακυμάνσεων με ανεξάρτητα δείγματα. Στην ενότητα αυτή, θα παρουσιαστεί ένας έλεγχος που έχει προταθεί στη βιβλιογραφία για το πρόβλημα του ελέγχου της ισότητας δύο πληθυσμιακών διακυμάνσεων με εξαρτημένα δείγματα. Ο έλεγχος αυτός παρουσιάστηκε από τον Boehnke (1989) και, παρότι έχει προταθεί εδώ και περισσότερα από τριάντα χρόνια, δεν είναι ευρέως διαδεδομένος, ενώ δεν έχουν επιλυθεί και προβλήματα που έχουν επισημανθεί από τον συγγραφέα του, όπως η ύπαρξη δεσμών.

Έστω δύο πληθυσμοί με αθροιστικές συναρτήσεις κατανομής F_i, $i = 1, 2$, με πληθυσμιακές διακυμάνσεις σ_1^2 και σ_2^2, αντίστοιχα. Επιπλέον, υποθέτουμε ότι είναι διαθέσιμα δύο τυχαία δείγματα από αυτούς τους πληθυσμούς, έστω $X_{11}, \ldots, X_{1,n}$ και $X_{21}, \ldots, X_{2,n}$, τα οποία δείγματα είναι εξαρτημένα. Υποθέτοντας ότι δεν υπάρχουν διαφορές στις διαμέσους, ο έλεγχος που προτάθηκε από τον Boehnke (1989) στηρίζεται στην ιδέα του ελέγχου των Siegel and Tukey (1960), που παρουσιάστηκε στην προηγούμενη ενότητα. Ειδικότερα, αποδίδονται με τον ίδιο τρόπο οι τάξεις στο σύνολο των $2n$ το πλήθος παρατηρήσεων, δηλαδή αποδίδεται η τάξη 1 στη μικρότερη τιμή, οι τάξεις 2 και 3 στις δύο μεγαλύτερες τιμές, οι τάξεις 4 και 5 στις δύο αμέσως μικρότερες τιμές και ούτω καθεξής. Αφού αποδοθούν οι τάξεις, δημιουργούμε τη διαφορά των τάξεων για κάθε πειραματική μονάδα. Η στατιστική συνάρτηση Δ, που προτάθηκε από τον Boehnke (1989), είναι το άθροισμα αυτών των διαφορών. Για να γίνει κατανοητός ο τρόπος υπολογισμού της στατιστικής συνάρτησης Δ, παραθέτουμε το ακόλουθο παράδειγμα από την εργασία του Boehnke (1989).

Παράδειγμα 6.23. Να υπολογιστεί η στατιστική συνάρτηση Δ όταν τα δύο εξαρτημένα δείγματα έχουν τις τιμές που ακολουθούν:

i	X_{1i}	X_{2i}
1	88	92
2	78	28
3	67	96
4	61	55
5	58	81

Λύση Παραδείγματος 6.23. Αρχικά, σύμφωνα με όσα αναφέρθηκαν, πρέπει να υπολογίσουμε τις τάξεις στο αναμεμειγμένο δείγμα και έπειτα να υπολογίσουμε τις διαφορές αυτών των τάξεων. Για τον σκοπό αυτό διατάσσουμε τις παρατηρήσεις κατά αύξουσα τάξη μεγέθους. Προκύπτει το ακόλουθο διατεταγμένο δείγμα: 28, 55, 58, 61, 67, 78, 81, 88, 92, 96. Επομένως, αποδίδονται σε αυτές οι ακόλουθες τάξεις: 1 4 5 8 9 10 7 6 3 2 και σχηματίζεται ο ακόλουθος πίνακας.

i	$R(X_{1i})$	$R(X_{2i})$	$d_i = R(X_{1i}) - R(X_{2i})$
1	6	3	3
2	10	1	9
3	9	2	7
4	8	4	4
5	5	7	-2

Είναι, επομένως, $\Delta = \sum_{i=1}^{n} (R(X_{1i}) - R(X_{2i})) = 21$. □

Η μηδενική υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων απορρίπτεται έναντι της εναλλακτικής υπόθεσης ότι διαφέρουν στατιστικά σημαντικά για απόλυτες τιμές της στατιστικής συνάρτησης Δ μεγαλύτερες από κάποια τιμή, έτσι ώστε να έχουμε επίπεδο σημαντικότητας a. Η ακριβής κατανομή της στατιστικής συνάρτησης Δ έχει βρεθεί αναλυτικά από τον Boehnke (1989) για την περίπτωση που $n = 2$ και αφήνεται ως άσκηση για τον/την αναγνώστη/στρια[2], ενώ έχει επιγραμματικά σχολιαστεί για την περίπτωση που $n = 3$. Για όλες τις άλλες περιπτώσεις ο Boehnke (1989) παραπέμπει τον/την αναγνώστη/στρια στο να δημιουργήσει ένα πρόγραμμα στον υπολογιστή κατάλληλο για τον υπολογισμό της συνάρτησης πιθανότητας της Δ υπό τη μηδενική υπόθεση. Στην περίπτωση που το μέγεθος δείγματος είναι μεγάλο ($n > 30$) αποδεικνύεται ότι η στατιστική συνάρτηση

$$Z = \frac{\Delta - \frac{n(n^2-1)}{6}}{\sqrt{\frac{(n-1)(n+1)^2 n^2}{36}}}$$

ακολουθεί προσεγγιστικά υπό τη μηδενική υπόθεση τυπική κανονική κατανομή $\mathcal{N}(0,1)$ και οι κρίσιμες περιοχές προκύπτουν τότε εύκολα.

[2]Υπόδειξη: θα πρέπει να θεωρήσετε τις 12 διαφορετικές περιπτώσεις διάταξης των τεσσάρων παρατηρήσεων και να υπολογίσετε τις αντίστοιχες 12 τιμές της στατιστικής συνάρτησης Δ.

6.7 Ασκήσεις

Άσκηση 6.1. Ο υπεύθυνος του τμήματος έρευνας αγοράς μιας μεγάλης εταιρείας κατάρτισε ένα ερωτηματολόγιο με στόχο να μελετήσει τις αντιδράσεις του καταναλωτικού κοινού για ένα προϊόν. Υπέθεσε ότι τα ποσοστά των καταναλωτών στους οποίους άρεσε το προϊόν και αυτών στους οποίους δεν άρεσε το προϊόν είναι ίσα. Για να ελέγξει την υπόθεση αυτή, διατύπωσε την εξής ερώτηση: «Πώς αισθάνεστε για το προϊόν αυτό ;». Τα άτομα που επρόκειτο να ερωτηθούν θα έπρεπε να απαντήσουν διαλέγοντας μια απάντηση σε μια κλίμακα από το 1 ως το 7, στην οποία ο αριθμός 1 υποδήλωνε ότι το προϊόν δεν ήταν της αρεσκείας του ερωτώμενου και ο αριθμός 7 δήλωνε ότι το προϊόν ήταν της απόλυτης αρεσκείας του. Η διάμεση απάντηση που αντιστοιχούσε στον αριθμό 4, δήλωνε ότι ο ερωτώμενος ήταν αδιάφορος προς το συγκεκριμένο προϊόν. Σε μια πιλοτική έρευνα που έκανε ενόψει της τελικής διαμόρφωσης του ερωτηματολογίου του, ο ερευνητής πήρε τις εξής απαντήσεις από 12 άτομα:

$$7 \quad 3 \quad 5 \quad 4 \quad 7 \quad 1 \quad 2 \quad 2 \quad 5 \quad 7 \quad 6 \quad 5.$$

Να ελεγχθεί η υπόθεση ότι η ανταπόκριση στο συγκεκριμένο προϊόν είναι αδιάφορη.

Άσκηση 6.2. Πήραμε δείγμα ιζήματος από 8 πλευρές κατά μήκος ενός ποταμού και υπολογίσαμε το μέσο μέγεθος του κόκκου της άμμου, όπως φαίνεται παρακάτω:

$$5.7 \quad 5.6 \quad 4.5 \quad 6.4 \quad 6.9 \quad 7.8 \quad 8.0 \quad 4.9$$

Να ελεγχθεί η υπόθεση ότι το μέσο μέγεθος του κόκκου δεν ξεπερνάει την τιμή 6.2.

Άσκηση 6.3. Η γρίπη και οι συνέπειές της θεωρούνται επικίνδυνες για άτομα ηλικίας 60 ετών και άνω και πολλοί γιατροί συνιστούν στους ασθενείς τους που ανήκουν σε αυτήν την ηλικιακή κατηγορία να εμβολιάζονται με ένα αντιγριπικό εμβόλιο κάθε φθινόπωρο. Ας υποθέσουμε ότι μια φαρμακευτική εταιρεία κατασκευάζει ένα νέο εμβόλιο για μία συγκεκριμένη μορφή γρίπης. Για να ελεγχθούν τα αποτελέσματα αυτού του εμβολίου, απαιτείται η γνώση της μεταβολής της θερμοκρασίας του σώματος που παρατηρείται ακριβώς πριν τον εμβολιασμό και μία ώρα μετά τον εμβολιασμό. Αν το εμβόλιο είναι αποτελεσματικό, ο ασθενής θα πρέπει να παρουσιάσει συμπτώματα ελαφριάς γρίπης (ελαφριά καταρροή, ελαφρύ πονόλαιμο, ρίγη και άνοδο της θερμοκρασίας του σώματος) μέσα σε μία ώρα από τον εμβολιασμό. Τα αποτελέσματα που ακολουθούν αναφέρονται σε ένα δείγμα 10 αρρένων ασθενών (ηλικίας 60 ετών και άνω), οι οποίοι συμμετείχαν στο πείραμα εθελοντικά.

	Εθελοντής									
Θερμοκρασία ^oC	1	2	3	4	5	6	7	8	9	10
Πριν (X_i)	37.0	37.0	36.4	36.7	37.0	36.9	37.0	37.0	36.8	37.0
Μετά (Y_i)	38.0	37.2	37.3	38.6	37.8	36.9	36.9	39.6	37.6	37.5

Αποτελούν τα δεδομένα του πίνακα ένδειξη σημαντικής αύξησης στη θερμοκρασία του σώματος αρρένων ασθενών ηλικίας 60 ετών και άνω μία ώρα μετά τον εμβολιασμό;

Άσκηση 6.4. Στο γυμνάσιο μιας επαρχιακής πόλης, 12 από τους 48 μαθητές της πρώτης τάξης ζούσαν σε αγροικίες. Ο καθηγητής της Φυσικής Αγωγής αυτού του Γυμνασίου, θέλοντας να εξετάσει κατά πόσο τα παιδιά που ζούσαν σε αγροικίες είχαν καλύτερη φυσική κατάσταση από τα παιδιά που ζούσαν στην πόλη, επινόησε ένα τεστ φυσικής κατάστασης στο οποίο και υπέβαλε τους 48 μαθητές. Τα αποτελέσματα του τεστ συνοψίζονται στον πίνακα που ακολουθεί, όπου οι υψηλοί βαθμοί αντανακλούν πολύ καλή φυσική κατάσταση. Ποια είναι τα συμπεράσματά σας; Να διεξάγετε κατάλληλο έλεγχο σε ε.σ. 5%.

Αποτελέσματα του τεστ φυσικής κατάστασης

Παιδιά Αγροικιών		Παιδιά Πόλης					
14.8	10.6	12.7	16.9	7.6	2.4	6.2	9.9
7.3	12.5	14.2	7.9	11.3	6.4	6.1	10.6
5.6	12.9	12.6	16.0	8.3	9.1	15.3	14.8
6.3	16.1	2.1	10.6	6.7	6.7	10.6	5.0
9.0	11.4	17.7	5.6	3.6	18.6	1.8	2.6
4.2	2.7	11.8	5.6	1.0	3.2	5.9	4.0

Άσκηση 6.5. Τα παρακάτω δεδομένα παριστάνουν τον χρόνο ζωής (σε $100h$) δύο τύπων ηλεκτρικών λυχνιών.

X	5.6	4.6	6.8	4.9	6.1	5.3	4.5	5.8	5.4	5.7
Y	7.2	8.1	5.1	7.3	6.9	7.8	5.9	6.7	6.5	7.1

Με επίπεδο σημαντικότητας $a = 0.1$ ελέγξτε την υπόθεση ότι οι δύο τύποι λυχνιών είναι το ίδιο αποτελεσματικοί.

Άσκηση 6.6. Σε ένα μάθημα διοίκησης επιχειρήσεων, οι διαλέξεις δόθηκαν με δύο διαφορετικούς τρόπους. Μία ομάδα φοιτητών παρακολούθησε τις διαλέξεις μέσω τηλεόρασης, ενώ μια δεύτερη ομάδα ζωντανά στο αμφιθέατρο. Σε κάθε περίπτωση όλοι οι φοιτητές εξετάστηκαν πριν από τη διάλεξη και μετά από αυτήν. Οι διαφορές στη βαθμολογία των δύο εξετάσεων για όλους τους φοιτητές καταγράφονται στον παρακάτω πίνακα.

Ζωντανά	20.3	23.5	4.8	21.9	15.5	20.3	26.6	21.9	- 9.4	4.4	-1.6	25.1
Τηλεόραση	6.2	15.6	25.0	4.7	28.1	17.2	14.1	31.2	12.6	9.4	17.2	23.4

Εξετάστε αν υπάρχει κάποια σημαντική διαφορά στη βαθμολογία των δύο ομάδων των φοιτητών σε επίπεδο σημαντικότητας 0.05.

Άσκηση 6.7. Εξετάζουμε την αρτηριακή πίεση 10 ασθενών πριν (X) και μετά (Y) τη χορήγηση κάποιου φαρμάκου κατά της πίεσης. Στη διάθεσή μας έχουμε τα παρακάτω δεδομένα.

i	1	2	3	4	5	6	7	8	9	10
X_i	15	13	17	17	18	15	13	12	13	19
Y_i	11	13	15	14	17	14	14	10	14	13

Να ελέγξετε σε επίπεδο σημαντικότητας 5% ότι η χρήση του φαρμάκου ελαττώνει την αρτηριακή πίεση.

Άσκηση 6.8. Σε 12 ζευγάρια μονοζυγωτικών (identical) διδύμων εφαρμόστηκαν ψυχολογικά κριτήρια με σκοπό να προσδιοριστεί κατά πόσο το πρώτο σε σειρά γέννησης (X) από τα δίδυμα τείνει να γίνει πιο επιθετικό από το δεύτερο (Y). Δόθηκαν τα ακόλουθα αποτελέσματα, όπου ο μεγαλύτερος αριθμός σημαίνει μεγαλύτερη επιθετικότητα. Ελέγξτε σε επίπεδο σημαντικότητας 5% τον παραπάνω ισχυρισμό.

Ζευγάρι	1	2	3	4	5	6	7	8	9	10	11	12
X_i	86	71	77	68	91	72	77	91	70	71	88	87
Y_i	88	77	76	64	96	72	65	90	65	80	81	72

Άσκηση 6.9. Μια έρευνα διεξήχθη για να διαπιστωθούν τα αποτελέσματα δύο «αντιδότων» (Α και Β) για τις συνέπειες του ποτού. Δέκα εθελοντές χωρίστηκαν τυχαία σε δύο ομάδες των πέντε ατόμων. Στη συνέχεια, αφού και τα 10 άτομα ήπιαν την ίδια ποσότητα αλκοόλ, στην πρώτη ομάδα των πέντε ατόμων δόθηκε το «αντίδοτο» Α, ενώ στη δεύτερη ομάδα των υπόλοιπων πέντε ατόμων, δόθηκε το «αντίδοτο» Β. Μετά από μία ώρα, βρέθηκαν από την ανάλυση αίματος οι παρακάτω ποσότητες αλκοόλ στο αίμα (σε mg/ml)

A	76	52	92	80	70
B	110	96	74	105	125

Υπάρχει σαφής ένδειξη για να συμπεράνουμε ότι το ένα αντίδοτο είναι αποτελεσματικότερο του άλλου; Να γίνει ο έλεγχος σε επίπεδο σημαντικότητας 5%.

Άσκηση 6.10. Ένας καθηγητής θέλει να συγκρίνει τις επιδόσεις των φοιτητών στα δύο τμήματα της ίδιας τάξης. Ο καθηγητής διαλέγει 15 φοιτητές τυχαία από κάθε τμήμα και τους τοποθετεί ανάλογα με τους βαθμούς τους. Η σειρά για τη συνδυαζόμενη διάταξη (τάξεις στο κοινό δείγμα) είναι:

Τμήμα 1	1	3	4	6	11	13	14	15	17	19	20	21	27	28	30
Τμήμα 2	2	5	7	8	9	10	12	16	18	22	23	24	25	26	29

Εφαρμόζοντας το κριτήριο των Mann-Whitney, ελέγξτε, σε επίπεδο σημαντικότητας 5%, αν οι επιδόσεις των φοιτητών στα δύο τμήματα διαφέρουν σημαντικά.

Άσκηση 6.11. Τα παρακάτω δεδομένα αφορούν τη συγκέντρωση φωσφόρου σε γεωργικό έδαφος, όπως αυτή μετρήθηκε μετά τη χρήση τεσσάρων διαφορετικών λιπασμάτων. Χρησιμοποιώντας μια κατάλληλη μη παραμετρική μέθοδο, να ελέγξετε την υπόθεση ότι δεν υπάρχει διαφορά στην πραγματική διάμεση συγκέντρωση φωσφόρου για τα τέσσερα διαφορετικά λιπάσματα. Να διατυπωθούν η H_0 και η H_1, να δοθεί η μορφή της κρίσιμης περιοχής για τον έλεγχο, καθώς και το συμπέρασμα του ελέγχου, με επίπεδο σημαντικότητας 1%.

Λίπασμα Α	8.1	4.9	7.0	8.5	8.0
Λίπασμα Β	11.5	11.9	12.1	10.8	10.9
Λίπασμα Γ	15.3	19.4	16.4	16.3	15.0
Λίπασμα Δ	23.0	35.0	28.4	30.1	26.7

Άσκηση 6.12. Ζητήθηκε από πέντε κριτικούς γεύσης να βαθμολογήσουν τέσσερις ετικέτες μπύρας. Η βαθμολογία ήταν από το 1 (η καλύτερη) μέχρι το 4 (η χειρότερη) και τα αποτελέσματα δίνονται στον παρακάτω πίνακα.

	Ετικέτες Μπύρας			
Κριτικός	Α	Β	Γ	Δ
1	2	1	3	4
2	3	1	2	4
3	4	2	1	3
4	3	4	1	2
5	2	1	4	3

Να ελέγξετε με επίπεδο σημαντικότητας 5% την υπόθεση ότι δεν υπάρχει στατιστικά σημαντική διαφοροποίηση στις βαθμολογίες των κριτών για τις τέσσερις διαφορετικές ετικέτες μπύρας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ξενόγλωσση

Ansari, A. R. and Bradley, R. A. (1960). Rank-Sum tests for dispersions. *Ann. Math. Statist.*, *31*(4), pp. 1174–1189.

Boehnke, K. (1989). A Nonparametric test for differences in the dispersion of dependent samples. *Biometrical Journal*, *31*, pp. 421–430.

Bortz, J., Lienert, G. and Boehnke, K. (2000). *Verteilungsfreie Methoden in der Biostatistik*. Springer.

Conover, W. J. (1998). *Practical Nonparametric Statistics* (3rd ed.). Wiley and Sons, Inc.

Conover, W. and Iman, R. (1978). Some exact tables for the squared rank test. *Communication in Statistics*, *5*, pp. 491–513.

Conover, W., Johnson, M. E. and Johnson, M. M. (1981). A Comparative Study of Tests for Homogeneity of Variances with Applications to the Outer Continental Shelf Bidding Data. *Technometrics*, *23*, pp. 351–361.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, *6*, pp. 241–252.

Eisinga, R., Heskes, T., Pelzer, B. and Grotenhuis, M. T. (2017). Exact p-values for pairwise comparison of Friedman rank sums, with application to comparing classifiers. *BMC Bioinformatics*, *18*(68).

Friedman, M. (1937). The use of ranks to avoid assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*, pp. 675–701.

Gabriel, K. and Lachebruch, P. (1969). Non-parametric anova in small samples: A monte carlo study of the adequacy of the asymptotic approximation. *Biometrics*, *25*, pp. 593–596.

Gibbons, J. D. and Chakraborti, S. (2020). *Nonparametric Statistical Inference, Fourth Edition Revised and Expanded*. Chapman and Hall/CRC.

Gibbons, J. D. (2014). Tests of Randomization. In: *Wiley StatsRef: Statistics Reference Online*. American Cancer Society.

Hollander, M., Wolfe, D. and Chicken, E. (2014). *Nonparametric Statistical Methods* (3nd ed.). John Wiley and Sons.

Iman, R. and Davenport, J. (1980). Approximations of the critical region of the Friedman statistic. *Com. Statistics Theory and Methods*, *9*, pp. 571–595.

Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika*, *41*, pp. 133–145.

Kruskal, W. and Wallis, W. (1952). Use of ranks on one-criterion variance analysis. *J. Amer. Statist. Assoc.*, *47*, pp. 583–621.

Kruskal, W. and Wallis, W. (1953). Errata: Use of ranks on one-criterion variance analysis. *J. Amer. Statist. Assoc.*, *48*, pp. 907–911.

Kvam, P. and Vidakovic, B. (2007). *Nonparametric Statistics with applications to science and engineering*. Wiley Series in Probability and Statistics.

Laubscher, N. F., Steffens, F. E. and De Lange, E. M. (1968). Exact Critical Values for Mood's Distribution-Free Test Statistic for Dispersion and Its Normal Approximation. *Technometrics*, *10*(3), pp. 497–507.

Lehmann, E. L. (2006). *Nonparametrics. Statistical methods based on ranks. With the special assistance of H. J. M. D'Abrera. Revised first edition.* Springer, New York.

Mann, H. and Whitney, D. R. (1947). On a test whether one of two random variables is stochastically larger than the other. *Ann. Math. Statistics*, *18*, pp. 50–60.

Mood, A. M. (1954). On the asymptotic efficiency of certain nonparametric two-sample tests. *Ann. Math. Stat.*, *25*, pp. 514–533.

Pratt, J. W. (1959). Remarks on zeros and ties in the Wilcoxon signed rank procedures. *J. Amer. Statist. Assoc.*, *54*, pp. 655–667.

Randles, R. and Wolfe, D. (1979). *Introduction to the Theory of Nonparametric Statistics*. Wiley.

Siegel, S. and Tukey, J. W. (1960). A Nonparametric sum of rans procedure for relative spread in unpaired samples. *Journal of the American Statistical Association.*, *55*, pp. 429–445.

Sprent, P. (1999). *Applied Nonparametric Statistical Methods*. Chapmann and Hall.

Terpstra, T. J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae*, *14*, pp. 327–333.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*, pp. 80–83.

ΚΕΦΑΛΑΙΟ 7

ΕΛΕΓΧΟΙ ΤΥΧΑΙΟΤΗΤΑΣ

Σύνοψη

Οι έλεγχοι τυχαιότητας, δηλαδή οι έλεγχοι της μηδενικής υπόθεσης ότι ένα σύνολο δεδομένων μπορεί να θεωρηθεί ότι είναι τυχαίο, αποτελούν ένα σημαντικό μέρος κάθε στατιστικής ανάλυσης. Η σπουδαιότητά τους εξηγείται από το γεγονός ότι οι περισσότερες από τις στατιστικές μεθοδολογίες, τόσο της Παραμετρικής όσο και της Μη Παραμετρικής Στατιστικής, υποθέτουν ότι το δείγμα μας είναι τυχαίο. Η ύπαρξη πολλών διαφορετικών πιθανών τρόπων απόκλισης από την τυχαιότητα, έχει οδηγήσει στην εμφάνιση στη βιβλιογραφία πληθώρας διαφορετικών ελέγχων τυχαιότητας. Σκοπός του κεφαλαίου αυτού είναι η παρουσίαση ορισμένων από αυτούς τους διαφορετικούς μη παραμετρικούς ελέγχους της τυχαιότητας του δείγματος, κατανοώντας τους διαφορετικούς τύπους αποκλίσεων που καθένας από αυτούς μπορεί να εντοπίσει.

Προαπαιτούμενη γνώση:

Βασικές γνώσεις Πιθανοτήτων και Στατιστικής.

Προσδοκώμενα μαθησιακά αποτελέσματα:

Ο/η φοιτητής/τρια ολοκληρώνοντας την ενότητα αυτή θα μπορεί να εφαρμόζει τους διάφορους ελέγχους τυχαιότητας, να προσδιορίζει την κατανομή, υπό τη μηδενική υπόθεση, των στατιστικών συναρτήσεων που χρησιμοποιούνται σε αυτούς τους ελέγχους και να κατανοεί τους τύπους αποκλίσεων που καθένας εξ αυτών μπορεί να εντοπίσει.

Γλωσσάριο επιστημονικών όρων

- Έλεγχος ροής μέγιστου μήκους
- Έλεγχος σημείων πρώτων διαφορών
- Έλεγχος συνεχόμενων ανοδικών/καθοδικών τιμών
- Έλεγχος τάξης Mann-Kendall
- Έλεγχοι τυχαιότητας με τάξεις
- Ροή
- Τεστ ροών
- Έλεγχος Bartels
- Έλεγχος Wald-Wolfowitz

Μπατσίδης, Α., Παπασταμούλης, Π., Πετρόπουλος, Κ., & Ρακιτζής, Α. (2022). *Μη Παραμετρική Στατιστική.* [Προπτυχιακό εγχειρίδιο].

7.1 Εισαγωγή

Μία από τις βασικότερες παραδοχές για την ανάπτυξη και εφαρμογή των παραδοσιακών μεθόδων στατιστικής συμπερασματολογίας, είτε παραμετρικών, είτε μη παραμετρικών, είναι ότι το δείγμα είναι τυχαίο, δηλαδή ότι οι n το πλήθος δειγματικές παρατηρήσεις $X_1, \dots, X_n$, είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές. Συνήθως αυτό ερμηνεύεται ως απουσία ύπαρξης κάποιου είδους μη τυχαίου μηχανισμού ο οποίος επιδρά στη συμπεριφορά των τιμών που λαμβάνονται από τον πληθυσμό. Ωστόσο, υπάρχουν περιπτώσεις που έχουμε λόγους να αμφιβάλλουμε για την τυχαιότητα των n δειγματικών παρατηρήσεων. Ενδεικτικά, η παρουσία ενός μη τυχαίου μηχανισμού επιλογής δειγματικών παρατηρήσεων μπορεί να είναι υπεύθυνη για την παρουσία αυτοσυσχέτισης στα δεδομένα, για την εμφάνιση μοτίβων και σχηματισμών (π.χ. μείξη κατανομών, τάση, περιοδικότητα). Μια τέτοια περίπτωση μπορεί, για παράδειγμα, να προκύψει όταν οι παρατηρήσεις είναι χρονολογικές, όπως π.χ. αν πρόκειται για το ύψος της βροχής σε μία περιοχή σε n διαφορετικές χρονικές περιόδους. Στο κεφάλαιο αυτό, θα παρουσιαστούν μη παραμετρικές μεθοδολογίες ελέγχου της τυχαιότητας ενός δείγματος. Ειδικότερα, παρουσιάζονται το τεστ των ροών (runs test) ή, όπως αλλιώς είναι γνωστό, το τεστ των Wald and Wolfowitz, ο έλεγχος ροής μέγιστου μήκους, ο έλεγχος τυχαιότητας που βασίζεται στην εμφάνιση συνεχόμενων ανοδικών/καθοδικών τιμών (runs up and down test) και ο έλεγχος σημείων πρώτων διαφορών των Moore and Wallis (1943). Τέλος, θα παρουσιαστούν έλεγχοι τυχαιότητας που βασίζονται στη χρήση τάξεων, όπως το Mann-Kendal Rank test και το Bartels rank test. Σε όλους τους παραπάνω ελέγχους δεν γίνεται καμία υπόθεση για τη μορφή του πληθυσμού από τον οποίο προέρχεται το υπό εξέταση δείγμα και, επομένως, θεωρούνται μη παραμετρικοί έλεγχοι.

Παρατήρηση 7.1. Επισημαίνεται ότι ένας έλεγχος που θα μπορούσε να συμπεριληφθεί σε αυτούς που θα παρουσιαστούν σε αυτό το κεφάλαιο είναι ο έλεγχος των Cox-Stuart, που παρουσιάστηκε στο Κεφάλαιο 5 και ελέγχει την ύπαρξη πτωτικής ή αυξητικής τάσης. Ο έλεγχος αυτός προτιμήθηκε να παρουσιαστεί στο Κεφάλαιο 5, καθώς αποτελεί ειδική περίπτωση του Προσημικού Κριτηρίου (βλ. Ενότητα 5.4) και για τον λόγο αυτό δεν παρουσιάζεται μαζί με τους υπόλοιπους ελέγχους τυχαιότητας.

7.2 Τεστ των ροών

Ένας τρόπος ελέγχου της τυχαιότητας ή μη ενός δείγματος n το πλήθος δειγματικών παρατηρήσεων είναι το λεγόμενο **τεστ των ροών** που είναι γνωστό και ως τεστ των Wald-Wolfowitz, διότι παρουσιάστηκε στη βιβλιογραφία από τους Abraham Wald (1902-1950) και Jacob Wolfowitz (1910-1981) (βλ. Wald and Wolfowitz, 1940). Είναι άξιο αναφοράς ότι σε αυτήν την εργασία χρησιμοποιήθηκε για πρώτη φορά η έννοια της ροής στον έλεγχο στατιστικών υποθέσεων[1]. Από τότε η έννοια της ροής έχει χρησιμοποιηθεί σε διάφορα επιστημονικά πεδία όπως, μεταξύ άλλων, στον στατιστικό έλεγχο ποιότητας, στην αξιοπιστία συστημάτων, στην ψυχολογία και στη βιολογία. Στην ενότητα αυτή, θα παρουσιαστεί διεξοδικά ο έλεγχος των ροών.

Έστω $X_1, \dots, X_n$ οι n το πλήθος διαθέσιμες παρατηρήσεις, διατεταγμένες (συνήθως) σε χρονολογική σειρά. Επιπλέον, υποθέτουμε ότι τα δεδομένα χωρίζονται ή μπορούν να χωριστούν σε δύο ομάδες, έστω στην ομάδα Α και στην ομάδα Β. Για παράδειγμα, αν θέλουμε να ελέγξουμε την τυχαιότητα των υπολοίπων ενός μοντέλου γραμμικής παλινδρόμησης, θεωρούμε ως τιμή διαχωρισμού το μηδέν και, αν η τιμή του i-οστού υπολοίπου είναι αρνητική, δίνουμε στο i-οστό υπόλοιπο το σύμβολο $-$ (ομάδα Α), ενώ αντίστοιχα, στα θετικά υπόλοιπα δίνουμε το σύμβολο $+$ και τα τοποθετούμε στην ομάδα Β. Με αυτόν τον τρόπο το αρχικό δείγμα κωδικοποιείται σε μια ακολουθία συμβόλων με δύο διαφορετικά αποτελέσματα (δύο ομάδες

[1] Οι συγγραφείς ασχολήθηκαν με τον έλεγχο της υπόθεσης ότι δύο δείγματα προέρχονται από τον ίδιο πληθυσμό. Καθώς όμως ο έλεγχος έχει πολύ μικρή ισχύ, δεν παρουσιάστηκε στην αντίστοιχη ενότητα του Κεφαλαίου 6 και προτιμήθηκε ο έλεγχος των Mann-Whitney.

αποτελεσμάτων). Στις υπόλοιπες περιπτώσεις, αν δεν καθορίζεται από τη φύση του προβλήματος κάποιο κριτήριο, χρησιμοποιείται συνηθέστερα ως σημείο διαχωρισμού σε δύο ομάδες η διάμεσος των παρατηρήσεων. Αυτή η θεώρηση έχει προταθεί για παράδειγμα από τον David (1947) και σε αυτές τις περιπτώσεις ο έλεγχος είναι γνωστός ως τεστ των ροών πάνω και κάτω της διαμέσου (runs above and below median).

Παρατήρηση 7.2. Στην περίπτωση που κάποιο υπόλοιπο είναι ίσο με μηδέν ή κάποια από τις δειγματικές τιμές είναι ίση με τη διάμεσο ή με την τιμή του κριτηρίου που έχει καθοριστεί για τον διαχωρισμό των παρατηρήσεων σε δύο ομάδες, τότε αυτές αποκλείονται από την περαιτέρω ανάλυση. Η ανάλυση συνεχίζεται λαμβάνοντας υπόψη την παραπάνω τροποποίηση στο μέγεθος του δείγματος. Σε όσα ακολουθούν συμβολίζεται με n το μέγεθος του τροποποιημένου δείγματος και με $X_1, \dots, X_n$ οι τελικά διαθέσιμες παρατηρήσεις.

Έστω ότι στο δείγμα των n το πλήθος δειγματικών παρατηρήσεων, $X_1, \dots, X_n$, n_1 από αυτές ανήκουν στην ομάδα Α και οι τιμές αυτών αντικαθίστανται με το σύμβολο $+$, ενώ οι $n_2 = n - n_1$ το πλήθος τιμές των υπόλοιπων παρατηρήσεων, που ανήκουν στην ομάδα Β αντικαθίστανται με το σύμβολο $-$. Επομένως, κάθε δειγματική παρατήρηση αντικαθίσταται από το σύμβολο $+$ ή το σύμβολο $-$, τα οποία σύμβολα προσδιορίζουν κατά αυτόν τον τρόπο σε ποια ομάδα από τις δύο ανήκει η δειγματική παρατήρηση. Στο πλαίσιο αυτό ισχύει ο ακόλουθος ορισμός.

Ορισμός 7.1

Σε μια ακολουθία συμβόλων ορίζεται ως **ροή** μια διαδοχή όμοιων συμβόλων, της οποίας προηγούνται και έπονται διαφορετικά σύμβολα ή τίποτα. **Μήκος μιας ροής** είναι ο αριθμός των συμβόλων που περιλαμβάνονται στη ροή.

Τόσο το πλήθος των ροών όσο και το μήκος τους παρέχουν ενδείξεις για αποκλίσεις από την τυχαιότητα. Ειδικότερα, μεγάλος ή μικρός αριθμός ροών, όπως και μια ροή υπερβολικά μεγάλου μήκους, σπάνια εμφανίζονται όταν έχουμε πραγματικά ένα τυχαίο δείγμα. Στη συνέχεια, στην ενότητα αυτή το ενδιαφέρον μας επικεντρώνεται στη μελέτη της στατιστικής συνάρτησης R, που ορίζεται ως ο αριθμός των ακολουθιών όμοιων συμβόλων στην ακολουθία των n το πλήθος συμβόλων $+$ και $-$. Δηλαδή, αρχικά, αυτό που μας ενδιαφέρει είναι η εύρεση του πλήθους των ακολουθιών όμοιων συμβόλων. Έτσι, για παράδειγμα στην περίπτωση της ακολουθίας:

$$++---++--+++$$

είναι $n = 12$, $n_1 = 7$, $n_2 = 5$ και $R = 5$.

Παρατηρήστε ότι, αν $R = 2$, αυτό θα σήμαινε ότι είτε αρχικά όλες οι παρατηρήσεις θα ανήκουν στην ομάδα Α και έπειτα στην ομάδα Β είτε αντίστροφα. Επομένως, μέχρι κάποιο σημείο, κάποια χρονική στιγμή, κάθε παρατήρηση της μιας ομάδας ακολουθείται από παρατήρηση που ανήκει στην ίδια ομάδα. Αν ήταν $R = 12$, θα σήμαινε ότι μια δειγματική παρατήρηση της μιας ομάδας διαδέχεται παρατήρηση της άλλης. Προφανώς, σε αυτές τις δύο περιπτώσεις θα έχουμε ενδείξεις για απόκλιση από την τυχαιότητα.

Από τα παραπάνω γίνεται αντιληπτό ότι η ποσότητα R μπορεί να χρησιμοποιηθεί για τον έλεγχο της τυχαιότητας ή όχι των n το πλήθος δειγματικών παρατηρήσεων και είναι αυτή που προτάθηκε από τους Wald και Wolfowitz. Ειδικότερα, προκύπτει ότι πολύ μικρές τιμές της στατιστικής συνάρτησης R ουσιαστικά υποδηλώνουν μια τάση να δημιουργούνται ομάδες, συστάδες, όμοιων συμβόλων και αποτελεί ένδειξη εξάρτησης των όρων της ακολουθίας από τη (χρονική) σειρά καταγραφής τους ή ένδειξη εξάρτησης άλλης μορφής. Από την άλλη μεριά, πολύ μεγάλες τιμές της στατιστικής συνάρτησης R ουσιαστικά υποδηλώνουν μια εναλλαγή μεταξύ των συμβόλων. Επομένως, πολύ μεγάλες τιμές της R αποτελούν ένδειξη κάποιας μορφής συστηματικής κυκλικής επίδρασης πάνω στις τιμές των όρων της ακολουθίας, άρα και πάλι αποτελεί ένδειξη απόκλισης από την τυχαιότητα.

Επομένως, απορρίπτουμε τη μηδενική υπόθεση

$$H_0 : \text{Η σειρά εμφάνισης των δύο συμβόλων είναι τυχαία,}$$

έναντι της εναλλακτικής

$$H_1 : \text{Η σειρά εμφάνισης των δύο συμβόλων δεν είναι τυχαία,}$$

αν $R \leqslant c_1$ ή $R \geqslant c_2$, όπου οι τιμές c_1 και c_2 θα πρέπει να προσδιοριστούν, έτσι ώστε να έχουμε έλεγχο με επίπεδο σημαντικότητας a. Οι τιμές αυτές επιλέγονται να είναι το $1 - a/2$ και $a/2$ ποσοστιαίο σημείο, αντίστοιχα, της κατανομής της R υπό τη μηδενική υπόθεση. Άρα, για τον προσδιορισμό τους απαιτείται η εύρεση της κατανομής αυτής.

Πριν προχωρήσουμε στην εύρεση της κατανομής της R υπό τη μηδενική υπόθεση θα παραθέσουμε ένα χρήσιμο για την εύρεση αυτής αποτέλεσμα (βλ., επίσης, Gibbons and Chakraborti, 2020).

Πρόταση 7.1. Ο αριθμός των διαφορετικών τρόπων τοποθέτησης n το πλήθος όμοιων στοιχείων σε r το πλήθος μη κενά κελιά είναι ίσος με $\binom{n-1}{r-1}$.

Απόδειξη Πρότασης 7.1. Αρχικά τοποθετούνται τα n το πλήθος στοιχεία, έστω π.χ. άσπρες μπάλες σε μία σειρά. Έπειτα τα r το πλήθος κελιά μπορούν να δημιουργηθούν παρεμβάλλοντας ανάμεσα σε οποιεσδήποτε δύο άσπρες μπάλες $r - 1$ το πλήθος μαύρες μπάλες. Καθώς υπάρχουν $n - 1$ το πλήθος δυνατές θέσεις τέτοιας τοποθέτησης, ο συνολικός αριθμός αυτών των τρόπων είναι ίσος με $\binom{n-1}{r-1}$. □

Μετά και την παράθεση του προηγούμενου αποτελέσματος, στην επόμενη πρόταση προσδιορίζεται η κατανομή του R υπό τη μηδενική υπόθεση.

Πρόταση 7.2. Αν R είναι ο αριθμός των ακολουθιών όμοιων συμβόλων στην ακολουθία n το πλήθος συμβόλων, εκ των οποίων n_1 το πλήθος είναι + και $n_2 = n - n_1$ το πλήθος είναι −, τότε για $r = 2,3,...,n$:

$$f_R(r) = P(R = r) = \begin{cases} \dfrac{2\binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1}}{\binom{n}{n_1}}, \text{ για } r = 2k, \\ \dfrac{\binom{n_1-1}{(r-1)/2}\binom{n_2-1}{(r-3)/2}+\binom{n_1-1}{(r-3)/2}\binom{n_2-1}{(r-1)/2}}{\binom{n}{n_1}}, \text{ για } r = 2k+1 \end{cases}$$

για $k = 1,2,...,n/2$.

Απόδειξη Πρότασης 7.2. Αν η μηδενική υπόθεση είναι αληθής, δηλαδή οι n το πλήθος δειγματικές παρατηρήσεις αποτελούν ένα τυχαίο δείγμα, τότε ο δυνατός αριθμός των διατάξεων των n_1 συμβόλων + είναι $\binom{n}{n_1}$. Αν τώρα το πλήθος των ακολουθιών όμοιων συμβόλων είναι ίσο με r, όπου r άρτιος, δηλαδή $r = 2,4,....$, τότε, υπό τη μηδενική υπόθεση, θα έχουμε $r/2$ ακολουθίες συμβόλων + και $r/2$ ακολουθίες συμβόλων −. Αφού έχουμε $r/2$ ακολουθίες συμβόλων +, αυτό σημαίνει ότι τα n_1 θετικά σύμβολα χωρίζονται σε $r/2$ ομάδες. Αυτό επιτυγχάνεται με $\binom{n_1-1}{r/2-1}$ τρόπους. Όμοια έχουμε $\binom{n_2-1}{r/2-1}$ τρόπους κατασκευής των $r/2$ ροών συμβόλων −. Οπότε από τον πολλαπλασιαστικό κανόνα και καθώς μπορεί να έχουμε είτε πρώτα μια ακολουθία συμβόλων + και μετά μια ακολουθία συμβόλων − και αντίστροφα, προκύπτει ότι:

$$f_R(r) = P(R = r) = \frac{2\binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1}}{\binom{n}{n_1}}, \text{ για } r \text{ άρτιο.}$$

Για την περίπτωση που έχουμε r το πλήθος ακολουθίες όμοιων συμβόλων με r περιττό αριθμό, και υπό τη μηδενική υπόθεση, είτε θα έχουμε $(r-1)/2$ ακολουθίες συμβόλων + και $(r-1)/2+1$ ακολουθίες συμβόλων

$-$, είτε θα έχουμε $(r-1)/2+1$ ακολουθίες συμβόλων $+$ και $(r-1)/2$ ακολουθίες συμβόλων $-$. Με παρόμοιο τρόπο, όπως πριν, προκύπτει ότι:

$$f_R(r) = P(R=r) = \frac{\binom{n_1-1}{(r-1)/2}\binom{n_2-1}{(r-3)/2} + \binom{n_1-1}{(r-3)/2}\binom{n_2-1}{(r-1)/2}}{\binom{n}{n_1}}, \text{ για } r \text{ περιττό,}$$

και η απόδειξη ολοκληρώθηκε. □

Από τα παραπάνω προκύπτει ότι η μηδενική υπόθεση της τυχαιότητας του δείγματος απορρίπτεται έναντι της εναλλακτικής της μη τυχαιότητας του δείγματος (δίπλευρος έλεγχος), είτε όταν $R \leqslant r_{1-a/2}$ είτε $R \geqslant r_{a/2}$, όπου $r_{1-a/2}$ είναι ένας αριθμός τέτοιος, ώστε: $\sum_{r=2}^{r_{1-a/2}} P(R=r) = a/2$. Όμως, ο υπολογισμός αυτός είναι ιδιαίτερα δύσκολος τις περισσότερες φορές και, επιπλέον, όπως συμβαίνει σε όλα τα στατιστικά τεστ που στηρίζονται σε συναρτήσεις πιθανότητας, υπάρχει πεπερασμένο πλήθος επιλογών για το επίπεδο σημαντικότητας a.

Για να γίνουν τα παραπάνω κατανοητά, θεωρήστε την ειδική περίπτωση που $n_1 = 5$ και $n_2 = 4$. Τότε, αν η κρίσιμη περιοχή του δίπλευρου ελέγχου είναι η $K : R \leqslant 2$ ή $R \geqslant 9$, με χρήση της Πρότασης 7.2 προκύπτει ότι $a = 3/126 \approx 0.0238$ κι αυτό διότι $P(R=9) = 1/126$ και $P(R=2) = 2/126$, ενώ, αν η κρίσιμη περιοχή είναι η $K : R \leqslant 3$ ή $R \geqslant 8$, προκύπτει ότι $a = 18/126 \approx 0.1429$, καθώς $P(R=8) = 8/126$ και, άρα, $P(R=8) + P(R=9) = 9/126$, ενώ $P(R=3) = 7/126$ και, επομένως, $P(R=2) + P(R=3) = 9/126$.

Παρατήρηση 7.3. Στην περίπτωση που θέλουμε να ελέγξουμε σε επίπεδο σημαντικότητας a τη μηδενική υπόθεση της τυχαιότητας του δείγματος έναντι της εναλλακτικής, η οποία υποδηλώνει την ύπαρξη τάσης για δημιουργία ομάδων (συστάδων) όμοιων συμβόλων, τότε η μηδενική υπόθεση απορρίπτεται για πολύ μικρές τιμές του R και η μορφή της κρίσιμης περιοχής είναι: $R \leqslant r_{1-a}$. Από την άλλη μεριά, σε επίπεδο σημαντικότητας a, η μηδενική υπόθεση της τυχαιότητας απορρίπτεται έναντι της εναλλακτικής της υπάρξης τάσης εναλλαγής των συμβόλων για πολύ μεγάλες τιμές του R, δηλαδή αν $R \geqslant r_a$, όπου r_a είναι τέτοιο, ώστε υπό τη μηδενική υπόθεση $P(R \geqslant r_a) \geqslant a \geqslant P(R > r_a)$.

Πίνακες κρίσιμων τιμών και πιθανοτήτων για το παραπάνω τεστ είναι διαθέσιμοι στη βιβλιογραφία και έχουν υπολογιστεί χρησιμοποιώντας την Πρόταση 7.2. Ενδεικτικά παραπέμπουμε τον/την ενδιαφερόμενο/μενη αναγνώστη/στρια στους Kokoska and Nevison (1989). Οι πίνακες αυτοί περιορίζονται σε μικρές τιμές των n_1, n_2, συνηθέστερα $n_1, n_2 \leqslant 10$. Για μεγαλύτερες τιμές των n_1, n_2, λαμβάνοντας επιπλέον υπόψη ότι ο παραπάνω έλεγχος οδηγεί σε περιορισμένο αριθμό επιλογών ακριβούς επιπέδου σημαντικότητας οδηγούμαστε στην εύρεση ενός προσεγγιστικού στατιστικού τεστ (βλ., μεταξύ άλλων, Wald and Wolfowitz, 1940), Gibbons and Chakraborti (2020). Ένας πίνακας για την εύρεση κρίσιμων τιμών του τεστ των ροών δίνεται στο Παράρτημα και, συγκεκριμένα, είναι ο Πίνακας Π.28. Αξίζει να αναφέρουμε πως ο συγκεκριμένος πίνακας είναι μόνο για ε.σ. 5%. Για οποιοδήποτε άλλο ε.σ. μπορεί να χρησιμοποιηθεί η `R` και, συγκεκριμένα, οι συναρτήσεις `pruns(x,n1,n2)` και `druns(x,n1,n2)` (βρίσκονται στο πακέτο `randtests`) οι οποίες δίνουν τιμές της συνάρτησης κατανομής και της συνάρτησης πιθανότητας της R, υπολογισμένες στο x, για οποιεσδήποτε τιμές των n_1, n_2. Στη συνέχεια δίνουμε ένα παράδειγμα.

Παράδειγμα 7.1. Ρίχνουμε ένα ζάρι 16 φορές και καταγράφουμε αν το αποτέλεσμα της ρίψης είναι άρτιος (Α) ή περιττός (Π). Τα αποτελέσματα των ρίψεων δίνονται παρακάτω:

ΑΑΑΠΠΑΠΑΑΠΠΠΑΑΑΠ.

Χρησιμοποιήστε το τεστ των ροών και ελέγξτε σε επίπεδο σημαντικότητας $a = 0.05$ (χρησιμοποιήστε τον Πίνακα Π.28 του Παραρτήματος) το αν η σειρά εμφάνισης άρτιας ή περιττής πλευράς στο ζάρι είναι τυχαία. Ποιο το πραγματικό επίπεδο σημαντικότητας του ελέγχου;

Λύση Παραδείγματος 7.1. Αρχικά, καταγράφουμε το πλήθος ροών επιτυχιών τύπου Α και Π. Παρατηρούμε ότι υπάρχουν 4 ροές αποτελεσμάτων Α (με μήκη 3, 1, 2 και 3) καθώς και 4 ροές αποτελεσμάτων Π (με μήκη 2, 1, 3 και 1). Άρα, η τιμή της στατιστικής συνάρτησης ελέγχου είναι $R = 8$. Επίσης, έχουμε ότι $n_1 = 9$ (πλήθος αποτελεσμάτων τύπου Α), $n_2 = 7$ (πλήθος αποτελεσμάτων τύπου Π), με $n = n_1 + n_2 = 16$. Από τον Πίνακα Π.28 του Παραρτήματος σε ε.σ. 5%, η μορφή της κρίσιμης περιοχής είναι $K : R \leqslant 4$ ή $R \geqslant 14$. Αφού η τιμή 8 δεν ανήκει στην κρίσιμη περιοχή, δεν απορρίπτεται, σε επίπεδο σημαντικότητας 5%, η υπόθεση της τυχαιότητας για το διαθέσιμο δείγμα.

Αξίζει να παρατηρήσουμε ότι το πραγματικό ε.σ. του παραπάνω ελέγχου είναι (περίπου) ίσο με 0.01538, αφού προκύπτει με χρήση της συνάρτησης πιθανότητας $P(R = r)$ υπό την H_0 για διάφορες τιμές του r, ότι $P(R \leqslant 4) = 0.00979$ και $P(R \geqslant 14) = 0.005594$. Στο ίδιο αποτέλεσμα καταλήγουμε υπολογίζοντας τις παραπάνω πιθανότητες με την `R` μέσω των εντολών `pruns(4,9,7)` και `1-pruns(13,9,7)`, αντίστοιχα. □

Όπως αναφέρθηκε προηγουμένως, για $n_1, n_2 \leqslant 10$, υπάρχουν πίνακες ποσοστιαίων σημείων της κατανομής της στατιστικής συνάρτησης R με τη βοήθεια των οποίων μπορούμε να βρούμε τη μορφή της κρίσιμης περιοχής. Στη συνέχεια, διατυπώνουμε ένα ασυμπτωτικό αποτέλεσμα με τη βοήθεια του οποίου μπορούμε να διεξάγουμε το τεστ των ροών για μεγάλες τιμές των n_1, n_2.

Πρόταση 7.3. Έστω R ο αριθμός των ακολουθιών όμοιων συμβόλων σε μια ακολουθία n το πλήθος συμβόλων εκ των οποίων n_1 το πλήθος είναι + και $n_2 = n - n_1$ το πλήθος είναι −. Τότε για $n \to \infty$, τέτοιο ώστε $\frac{n_1}{n} \to \lambda$ και $\frac{n_2}{n} \to 1 - \lambda$, για σταθερό λ, με $0 < \lambda < 1$, υπό τη μηδενική υπόθεση της τυχαιότητας του δείγματος, έπεται ότι:

$$Z = \frac{R - \mathsf{E}(R)}{\sqrt{\mathsf{Var}(R)}} \xrightarrow{d} \mathcal{N}(0,1),$$

όπου

$$\mathsf{E}(R) = 1 + \frac{2n_1n_2}{n} \text{ και } \mathsf{Var}(R) = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)},$$

ή ισοδύναμα

$$Z = \frac{R - 2n\lambda(1-\lambda)}{2\sqrt{n\lambda(1-\lambda)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Απόδειξη Πρότασης 7.3. Παρατηρήστε, αρχικά, ότι από τον τρόπο ορισμού του στατιστικού R προκύπτει ότι μπορεί να γραφτεί ως το παρακάτω άθροισμα:

$$R = 1 + \sum_{i=2}^{n} I_i,$$

όπου I_i, $i = 2, \ldots, n$ είναι μία δείκτρια συνάρτηση που λαμβάνει την τιμή 1, αν το i-οστό σύμβολο είναι διαφορετικό από το $(i-1)$-οστό σύμβολο, για $i = 2, \ldots, n$. Τότε, εξ ορισμού, κάθε I_i είναι μία διακριτή τυχαία μεταβλητή που ακολουθεί Bernoulli κατανομή με παράμετρο επιτυχίας

$$p = \frac{n_1n_2}{\binom{n}{2}} = \frac{2n_1n_2}{n(n-1)}.$$

Επιπλέον, λαμβάνοντας υπόψη ότι η μέση τιμή της Bernoulli είναι ίση με την παράμετρο p, προκύπτει ότι:

$$\mathsf{E}(R) = \mathsf{E}\left(1 + \sum_{i=2}^{n} I_i\right) = 1 + \sum_{i=2}^{n} \mathsf{E}(I_i) = 1 + \sum_{i=2}^{n} \frac{2n_1n_2}{n(n-1)} = 1 + \frac{2n_1n_2}{n}.$$

Θα πρέπει, επίσης, να παρατηρήσουμε ότι οι τ.μ. I_i δεν είναι μεταξύ τους ανεξάρτητες, για $i = 2,\ldots,n$. Άρα, για τη διασπορά της τ.μ. R ισχύει ότι:

$$\begin{aligned}
\mathrm{Var}(R) &= \mathrm{Var}\left(1 + \sum_{i=2}^{n} I_i\right) = \mathrm{Var}\left(\sum_{i=2}^{n} I_i\right) = \sum_{i=2}^{n} \mathrm{Var}(I_i) + \sum_{2\le i\ne j\le n} \mathrm{Cov}\left(I_i, I_j\right) \\
&= \sum_{i=2}^{n} \mathrm{Var}(I_i) + \sum_{2\le i\ne j\le n} \left\{\mathrm{E}\left(I_i I_j\right) - \mathrm{E}\left(I_i\right)\mathrm{E}\left(I_j\right)\right\} \\
&= \sum_{i=2}^{n} \mathrm{Var}(I_i) + \sum_{2\le i\ne j\le n} \mathrm{E}\left(I_i I_j\right) - (n-2)(n-1)\left(\mathrm{E}\left(I_i\right)\right)^2 \\
&= \sum_{i=2}^{n} \left\{\mathrm{E}\left(I_i^2\right) - \left(\mathrm{E}\left(I_i\right)\right)^2\right\} + \sum_{2\le i\ne j\le n} \mathrm{E}\left(I_i I_j\right) - (n-2)(n-1)\left(\mathrm{E}\left(I_i\right)\right)^2 \\
&= (n-1)\mathrm{E}(I_i^2) - (n-1)\left(\mathrm{E}\left(I_i\right)\right)^2 - (n-2)(n-1)\left(\mathrm{E}\left(I_i\right)\right)^2 + \sum_{2\le i\ne j\le n} \mathrm{E}\left(I_i I_j\right),
\end{aligned}$$

όπου

$$\mathrm{E}\left(I_i^2\right) = \mathrm{Var}\left(I_i\right) + \left(\mathrm{E}(I_i)\right)^2 = p(1-p) + p^2 = p \equiv \frac{2n_1 n_2}{n(n-1)}.$$

Επομένως, για την εύρεση της διακύμανσης απαιτείται η εύρεση των $(n-1)(n-2)$ το πλήθος ροπών τύπου $\mathrm{E}\left(I_i I_j\right)$. Για τον υπολογισμό αυτών λαμβάνουμε υπόψη τα ακόλουθα: για τις $2(n-2)$ το πλήθος περιπτώσεις που είναι τέτοιες, ώστε είτε $i = j-1$ είτε $i = j+1$, ισχύει ότι:

$$\mathrm{E}\left(I_i I_j\right) = \frac{n_1 n_2 (n_1 - 1) + n_2 n_1 (n_2 - 1)}{n(n-1)(n-2)} = \frac{n_1 n_2}{n(n-1)},$$

ενώ για τις υπόλοιπες $(n-1)(n-2) - 2(n-2) = (n-2)(n-3)$ το πλήθος περιπτώσεις ισχύει ότι

$$\mathrm{E}\left(I_i I_j\right) = \frac{4n_1 n_2 (n_1 - 1)(n_2 - 1)}{n(n-1)(n-2)(n-3)}.$$

Επομένως, είναι:

$$\mathrm{Var}(R) = \frac{2n_1 n_2}{n} + \frac{2(n-2)}{n(n-1)} n_1 n_2 + \frac{4n_1 n_2 (n_1 - 1)(n_2 - 1)}{n(n-1)} - \left(\frac{2n_1 n_2}{n}\right)^2 = \frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}.$$

Το ζητούμενο προκύπτει εφαρμόζοντας το Κ.Ο.Θ., λαμβάνοντας υπόψη τις παραπάνω σχέσεις ή ισοδύναμα ότι:

$$\lim_{n\to\infty} \mathrm{E}(R/n) = 2\lambda(1-\lambda),$$

και

$$\lim_{n\to\infty} \mathrm{Var}(R/\sqrt{n}) = 4\lambda^2(1-\lambda)^2.$$

□

Από την παραπάνω πρόταση έχουμε ότι η μηδενική υπόθεση της τυχαιότητας του δείγματος έναντι της εναλλακτικής υπόθεσης της μη τυχαιότητας του δείγματος απορρίπτεται σε επίπεδο σημαντικότητας a, αν $|Z| \ge z_{a/2}$. Παρόμοια για τους μονόπλευρους ελέγχους, που αναφέρθηκαν στην Παρατήρηση 7.3, έχουμε τις κρίσιμες περιοχές $Z \le -z_a$ και $Z \ge z_a$, αντίστοιχα.

Για μικρές τιμές του n συνιστάται η χρήση της στατιστικής συνάρτησης που προκύπτει με εφαρμογή της διόρθωσης συνεχείας, ώστε να επιτυγχάνεται καλύτερη προσέγγιση στη $\mathcal{N}(0,1)$. Σε αυτήν την περίπτωση, η μηδενική υπόθεση απορρίπτεται, αν

$$Z_{cc}^{(l)} = \frac{R + 0.5 - \mathsf{E}(R)}{\sqrt{\mathsf{Var}(R)}} \leqslant -z_{a/2}$$

ή

$$Z_{cc}^{(u)} = \frac{R - 0.5 - \mathsf{E}(R)}{\sqrt{\mathsf{Var}(R)}} \geqslant z_{a/2}.$$

Οι μονόπλευροι έλεγχοι είναι $Z_{cc}^{(l)} \leqslant -z_a$ και $Z_{cc}^{(u)} \geqslant z_a$, αντίστοιχα.

Παρατήρηση 7.4. Το τεστ των ροών έχει επεκταθεί στην περίπτωση που τα δεδομένα μπορούν να ταξινομηθούν σε μία από k το πλήθος κατηγορίες. Ειδικότερα, αν n_i, $i = 1,2,$ είναι το πλήθος των τιμών που ανήκουν στην i–οστή κατηγορία, με $\sum_{i=1}^{k} n_i = n$, τότε αποδεικνύεται ότι η τ.μ. που παριστάνει τον αριθμό των συνολικών ροών, έστω R, ακολουθεί ασυμπτωτικά κανονική κατανομή με μέση τιμή

$$\mu_R = \mathsf{E}(R) = \frac{n(n+1) - \sum_{i=1}^{k} n_i^2}{n},$$

και διακύμανση

$$\sigma_R^2 = \mathsf{Var}(R) = \frac{\sum_{i=1}^{k} n_i^2 \left(\sum_{i=1}^{k} n_i^2 + n(n+1) \right) - 2n \sum_{i=1}^{k} n_i^3 - n^3}{n^2(n-1)}.$$

Για περισσότερες λεπτομέρειες παραπέμπουμε στους Wallis and Roberts (1956), Sheskin (2011) και στις εκεί αναφορές.

Παράδειγμα 7.2. Μια ομάδα μπάσκετ που συμμετέχει στο πρωτάθλημα του NBA έχει την επόμενη ακολουθία αποτελεσμάτων (N =νίκη, H =ήττα):

$$\mathrm{NHNNHNHHHNHHNNHNNHNHNHNHHNHNH}.$$

Να ελέγξετε, σε επίπεδο σημαντικότητας 5%, αν υπάρχει τυχαιότητα στα αποτελέσματά της και να υπολογίσετε την p-τιμή του ελέγχου.

Λύση Παραδείγματος 7.2. Αν με n_1 συμβολίσουμε το πλήθος των νικών N, τότε $n_1 = 14$, οπότε $n_2 = 15$ και $n = 29$. Επιπλέον, όπως φαίνεται από το πλήθος των όμοιων συμβόλων στην ακολουθία των αποτελεσμάτων, εύκολα υπολογίζουμε ότι $R = 22$, αφού

$$\underline{N}\ \underline{\mathrm{H}}\ \underline{NN}\ \underline{\mathrm{H}}\ \underline{N}\ \underline{\mathrm{HHH}}\ \underline{N}\ \underline{\mathrm{HH}}\ \underline{NN}\ \underline{\mathrm{H}}\ \underline{NN}\ \underline{\mathrm{H}}\ \underline{N}\ \underline{\mathrm{H}}\ \underline{N}\ \underline{\mathrm{H}}\ \underline{N}\ \underline{\mathrm{HH}}\ \underline{N}\ \underline{\mathrm{H}}\ \underline{N}\ \underline{\mathrm{H}}.$$

Είναι,

$$\mathsf{E}(R) = 1 + \frac{2n_1n_2}{n} = 1 + \frac{2 \cdot 14 \cdot 15}{29} = 15.482$$

και

$$\mathsf{Var}(R) = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)} = \frac{2 \cdot 14 \cdot 15(2 \cdot 14 \cdot 15 - 29)}{29^2(29-1)} = 6.97.$$

Επομένως, για τον δίπλευρο έλεγχο της υπόθεσης ότι το δείγμα είναι τυχαίο θα χρησιμοποιήσουμε τη στατιστική συνάρτηση

$$Z = \frac{R - \mathsf{E}(R)}{\sqrt{\mathsf{Var}(R)}} \underset{H_0}{\overset{\text{ασυμπ.}}{\sim}} \mathcal{N}(0,1),$$

με κρίσιμη περιοχή $|Z| \geq z_{a/2} = z_{0.025} = 1.96$. Η τιμή της στατιστικής συνάρτησης υπολογίζεται ως εξής:

$$z = \frac{22 - 15.482}{\sqrt{6.97}} = 2.49$$

και, καθώς 2.49>1.96, συμπεραίνουμε ότι η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας 5%. Αυτό σημαίνει ότι τα αποτελέσματα της ομάδας δεν μπορούν να θεωρηθούν ότι είναι τυχαία.

Η p-τιμή του δίπλευρου ελέγχου υπολογίζεται ως εξής:

$$\begin{aligned} P(|Z| \geqslant 2.49) &= P(Z \leqslant -2.49) + P(Z \geqslant 2.49) = 2 \cdot P(Z \geqslant 2.49) \\ &= 2 \cdot (1 - \Phi(2.49)) = 0.0128. \end{aligned}$$

□

Παράδειγμα 7.3. Για $n = 5, n_1 = 3$ και $n_2 = 2$ να προσδιοριστεί η κατανομή του R, υπό τη μηδενική υπόθεση, χωρίς να χρησιμοποιήσετε την Πρόταση 7.2.

Λύση Παραδείγματος 7.3. Ο αριθμός των δυνατών ακολουθιών, όταν $n = 5$, $n_1 = 3$ και $n_2 = 2$, είναι $\binom{5}{3} = 10$ και είναι οι ακόλουθες:

$$++\,+--, +\,+-+-, +\,+--+, +-++-, +-+-+,$$

και

$$+--++, -+++-, -+-++, --+++, -++-+,$$

με αντίστοιχο αριθμό ροών: 2, 4, 3, 4, 5, 3, 3, 4, 2, 4. Άρα, οι δυνατές τιμές της διακριτής τυχαίας μεταβλητής R, που παριστάνει το πλήθος των ακολουθιών όμοιων συμβόλων, είναι 2, 3, 4 και 5. Τότε, υπό τη μηδενική υπόθεση, δηλαδή υπό την υπόθεση ότι καθεμία εκ των 10 δυνατών ακολουθιών είναι ισοπίθανη, αφού το δείγμα θεωρείται τυχαίο, έχουμε ότι:

r	$P(R = r)$
2	$P(R = 2) = 2/10$
3	$P(R = 3) = 3/10$
4	$P(R = 4) = 4/10$
5	$P(R = 5) = 1/10$

□

Παρατήρηση 7.5. Η ισχύς του τεστ των ροών αυξάνεται όσο αυξάνεται το μέγεθος του δείγματος. Ωστόσο, ο Mogull (1994) απέδειξε ότι ο έλεγχος των ροών δεν θα πρέπει να χρησιμοποιείται στην ακραία περίπτωση όπου ένα δείγμα έχει δύο ροές, καθώς σε τέτοιες περιπτώσεις δεν μπορεί να εντοπίσει απόκλιση από την τυχαιότητα, έχει μειωμένη ισχύ, η οποία μειώνεται όσο αυξάνεται το μέγεθος του δείγματος.

Παρατήρηση 7.6. Όπως αναφέρθηκε στην αρχή αυτής της ενότητας, ο έλεγχος των ροών αρχικά προτάθηκε για τον έλεγχο της υπόθεσης ότι δύο δείγματα προέρχονται από τον ίδιο πληθυσμό. Ειδικότερα, έστω δύο ανεξάρτητα τυχαία δείγματα $X_1,...,X_{n_1}$ και $Y_1,...,Y_{n_2}$ από πληθυσμούς με α.σ.κ. F_X και F_Y, αντίστοιχα. Θέλοντας να ελέγξουμε την $H_0 : F_X(x) = F_Y(x)$, για κάθε $x \in \mathbb{R}$ έναντι της εναλλακτικής ότι $H_1 : F_X(x) \neq F_Y(x)$, για κάποιο $x \in \mathbb{R}$ ακολουθούμε την εξής διαδικασία. Αρχικά, αναμειγνύουμε τα δύο δείγματα και οι δειγματικές παρατηρήσεις διατάσσονται κατά αύξουσα τάξη μεγέθους. Στη συνέχεια, αντικαθίστανται οι τιμές των X και Y με τα σύμβολα + και −, αντίστοιχα, και εφαρμόζεται ο έλεγχος των ροών που παρουσιάστηκε πρωτύτερα. Σημειώνεται, επίσης, ότι στην περίπτωση παρουσίας δεσμών, το κοινό (αναμεμειγμένο) δείγμα, διατάσσεται με τέτοιο τρόπο, ώστε να προκύψει το μέγιστο δυνατό πλήθος ροών συμβόλων + και −.

Ο έλεγχος των ροών που παρουσιάστηκε μπορεί να εντοπίζει ουσιαστικά δύο τύπους αποκλίσεων από την τυχαιότητα, αυτούς που αντιστοιχούν σε πολύ μικρές ή μεγάλες τιμές του συνολικού αριθμού των ροών. Καθώς υπάρχουν διάφοροι πιθανοί τύποι αποκλίσεων από την τυχαιότητα έχει προταθεί πληθώρα τέτοιων ελέγχων στη βιβλιογραφία. Κάποιοι από αυτούς αποτελούν αντικείμενο μελέτης των επόμενων ενοτήτων.

7.3 Έλεγχος ροής μέγιστου μήκους

Στην προηγούμενη ενότητα παρουσιάστηκε το τεστ των ροών για τον έλεγχο της τυχαιότητας του δείγματος, το οποίο βασίζεται στο συνολικό πλήθος ροών επιτυχιών και των δύο αποτελεσμάτων. Ωστόσο, η στατιστική συνάρτηση ελέγχου δεν λαμβάνει υπόψη της το μήκος των ροών. Στην ενότητα αυτή, θα παρουσιαστεί ο έλεγχος που βασίζεται στο μέγιστο μήκος ροής επιτυχιών για ένα από τα δύο δυνατά αποτελέσματα (longest run test) . Λεπτομέρειες για αυτό το τεστ μπορούν να βρεθούν στις εργασίες των Mosteller (1941) και Bateman (1948).

Συγκεκριμένα, έστω $X_1, \ldots, X_n$, οι n το πλήθος διαθέσιμες παρατηρήσεις, διατεταγμένες σε χρονολογική σειρά συνήθως. Επιπλέον, υποθέτουμε ότι τα δεδομένα χωρίζονται ή μπορούν να χωριστούν σε δύο ομάδες, έστω στην ομάδα Α (αποτέλεσμα +) και στην ομάδα Β (αποτέλεσμα −). Αν θεωρήσουμε την ακολουθία

$$+ + + - - - + + - - + + ++$$

δεν είναι δύσκολο να διαπιστώσουμε ότι υπάρχουν συνολικά 5 ροές (3 αφορούν αποτελέσματα + και 2 αφορούν αποτελέσματα −). Όμως, αν θεωρήσουμε ως επιτυχία την εμφάνιση αποτελέσματος +, παρατηρούμε ότι η πρώτη ροή έχει μήκος 3, η δεύτερη έχει μήκος 2, ενώ η τρίτη έχει μήκος 4. Με βάση τα όσα έχουμε πει έως τώρα, είναι $n = 14$, $n_1 = 9$, $n_2 = 5$, $R = 5$.

Αν είναι γνωστός ο αριθμός n_1 των επιτυχιών στις n προσπάθειες, τότε αποδεικνύεται (βλ., μεταξύ άλλων, Balakrishnan and Koutras, 2011) η παρακάτω πρόταση.

Πρόταση 7.4. Έστω L_n η τυχαία μεταβλητή η οποία εκφράζει το μήκος της μέγιστης ροής επιτυχιών σε μια ακολουθία n το πλήθος ανεξάρτητων δοκιμών Bernoulli. Τότε, υπό την υπόθεση της τυχαιότητας του δείγματος,

$$P(L_n < k | S = n_1) = \frac{1}{\binom{n}{n_2}} \sum_{j=0}^{\lfloor \frac{n_1}{k} \rfloor} (-1)^j \binom{n_2 + 1}{j} \binom{n - jk}{n_2}, \tag{7.1}$$

όπου S είναι ο συνολικός αριθμός επιτυχιών στις n δοκιμές και $n_2 = n - n_1$.

Απόδειξη Πρότασης 7.4. Αφού n_1 είναι οι επιτυχίες, οι αποτυχίες θα είναι $n_2 = n - n_1$. Άρα, οι n_2 αποτυχίες μπορούμε να θεωρήσουμε ότι ορίζουν $n_2 + 1$ το πλήθος κελιά (ή διαμερίσεις), όπου οι n_1 επιτυχίες πρέπει να τοποθετηθούν κατάλληλα. Για την πραγματοποίηση του ενδεχομένου $\{L_n < k\}$ θα πρέπει ο μέγιστος αριθμός παρατηρήσεων σε κάθε κελί να είναι $k - 1$. Τότε, το πλήθος των διαφορετικών τρόπων με τους οποίους μπορούμε να τοποθετήσουμε n_1 αντικείμενα σε n_2 διαφορετικά κελιά, με χωρητικότητα το πολύ $k - 1$ είναι:

$$\sum_{j=0}^{\lfloor \frac{n_1}{k} \rfloor} (-1)^j \binom{n_2 + 1}{j} \binom{n - jk}{n_2}.$$

Άρα, διαιρώντας με το πλήθος $\binom{n}{n_1}$ των δυνατών διατάξεων των n_1 συμβόλων +, έπεται το παραπάνω ζητούμενο αποτέλεσμα. □

Παρατήρηση 7.7. Από την Πρόταση 7.4 μπορούμε άμεσα να υπολογίσουμε την κατανομή του μέγιστου μήκους ροής επιτυχιών τύπου + υπό την υπόθεση της τυχαιότητας του δείγματος, δηλαδή υπό την H_0, δοθέντος του αριθμού των επιτυχιών σε n δοκιμές Bernoulli, καθώς

$$P(L_n = k|S = n_1) = P(L_n < k+1|S = n_1) - P(L_n < k|S = n_1),$$

για $k = 0,1,\dots,n_1$.

Παρατήρηση 7.8. Για να βρούμε τη (μη δεσμευμένη) κατανομή $P(L_n = k)$, $k = 0,2,\dots,n$, της μέγιστης ροής επιτυχιών υπό την H_0, αρκεί να παρατηρήσουμε ότι με εφαρμογή του Θεωρήματος Ολικής Πιθανότητας έχουμε:

$$P(L_n < k) = \sum_{n_2=0}^{n} P(L_n < k|S = n_1)P(S = n_1)$$

και με αντικατάσταση στην έκφραση για τη δεσμευμένη πιθανότητα, έπεται, μετά από κάποιες αλγεβρικές πράξεις, ότι:

$$P(L_n < k) = \sum_{n_2=0}^{n} q^{n_2} p^{n-n_2} \sum_{j=0}^{\lfloor \frac{n_1}{k} \rfloor} (-1)^j \binom{n_2+1}{j} \binom{n-jk}{n_2}.$$

Επομένως, $P(L_n = k) = P(L_n < k+1) - P(L_n < k)$.

Παράδειγμα 7.4. Για $n = 5$ να προσδιοριστεί η κατανομή της L_5, υπό τη H_0, χωρίς να χρησιμοποιήσετε το προηγούμενο αποτέλεσμα. Να υπολογιστεί, επίσης, η πιθανότητα $P(L_n < 3|S = 3)$ υπό την υπόθεση της τυχαιότητας του δείγματος.

Λύση Παραδείγματος 7.4. Αρχικά, καταγράφουμε το πλήθος των δυνατών αποτελεσμάτων + και − σε μια ακολουθία $n = 5$ το πλήθος δοκιμών. Λόγω της υπόθεσης της τυχαιότητας του δείγματος, καθεμία από αυτές τις πεντάδες έχει την ίδια πιθανότητα εμφάνισης. Άμεσα διαπιστώνουμε ότι αυτές είναι οι ακόλουθες:

$$\begin{array}{cc}
+++++,++++-,+++-+,++-++, & +-+++,-++++,+++--,+-++-, \\
+-++-,-+++-,++--+,+-+-+, & -++-+,+--++,-+-++,--+++, \\
---++,--+-+,-+--+,+----, & --++-,-+-+-,+--+-,-++--, \\
+-+--,++---,-----+,----+-, & --+--,-+---,+----,-----,
\end{array}$$

με αντίστοιχο μέγιστο μήκος ροής επιτυχιών (συμβόλων +):

$$\begin{array}{c}
5,4,3,2,3,4,3,2,2,3,2,1,2,2,2,3, \\
2,1,1,1,2,1,1,2,1,2,1,1,1,1,1,0.
\end{array}$$

Άρα, οι δυνατές τιμές της διακριτής τυχαίας μεταβλητής L_5, που παριστάνει το μέγιστο μήκος ροής επιτυχιών (συμβόλων +) σε μια ακολουθία 5 το πλήθος δοκιμών Bernoulli είναι 0, 1, 2, 3, 4 και 5. Τότε, υπό τη μηδενική υπόθεση, δηλαδή υπό την υπόθεση ότι καθεμία εκ των 32 δυνατών ακολουθιών είναι ισοπίθανη, αφού το δείγμα θεωρείται τυχαίο, έχουμε ότι:

κ	$P(L_5 = \kappa)$
0	$P(L_5 = 0) = 1/32$
1	$P(L_5 = 1) = 12/32$
2	$P(L_5 = 2) = 11/32$
3	$P(L_5 = 3) = 5/32$
4	$P(L_5 = 4) = 2/32$
5	$P(L_5 = 5) = 1/32$

Τέλος, για τον υπολογισμό της πιθανότητας $P(L_n < 3|S = 3)$, όταν η H_0 είναι αληθής, αρκεί να παρατηρήσουμε ότι το πλήθος των ακολουθιών με 3 ακριβώς αποτελέσματα + είναι 10 και σε αυτές, υπάρχουν 7 συνολικά πεντάδες, όπου η μέγιστη ροή επιτυχιών είναι μικρότερη από 3. Οι ακολουθίες αυτές είναι οι παρακάτω:

$$+-+++-, +-+++-, +++--+, +-++-+, -+++-+, +--++ +, -+--++.$$

Επομένως, $P(L_n < 3|S = 3) = \frac{7}{10}$. □

Από όσα αναφέρθηκαν παραπάνω, μπορούμε να χρησιμοποιήσουμε ως στατιστική συνάρτηση ελέγχου το μήκος της μέγιστης ροής επιτυχιών και να διεξάγουμε τον παρακάτω έλεγχο:

$$H_0 : \text{το δείγμα είναι τυχαίο,}$$

έναντι της εναλλακτικής

$$H_1 : \text{το δείγμα δεν είναι τυχαίο.}$$

Ειδικότερα, η εμφάνιση μιας ασυνήθιστα μεγάλης (σε μήκος) ροής επιτυχιών αποτελεί ένδειξη για παραβίαση της υπόθεσης ότι το δείγμα είναι τυχαίο και, μάλιστα, μπορεί να συνδεθεί με την παρουσία τάσης στα δεδομένα.

Αν υποθέσουμε ότι το επίπεδο σημαντικότητας είναι a, τότε η H_0 απορρίπτεται, αν παρατηρηθεί μία μεγάλη τιμή, έστω l, για την L_n τέτοια, ώστε $P(L_n \geqslant l \,|H_0) = a$. Για την περίπτωση πεπερασμένης ακολουθίας δοκιμών πλήθους n, προτείνεται η χρήση της δεσμευμένης κατανομής της L_n δοθέντος του αριθμού των επιτυχιών, που προσδιορίστηκε στη σχέση (7.1). Για μικρές τιμές των n, n_1, μπορούν να χρησιμοποιηθούν οι πίνακες με τα ποσοστιαία σημεία της κατανομής της L_n υπό την H_0 που παρατίθενται στις εργασίες των Mosteller (1941) και Bateman (1948).

Σύμφωνα με τους Gibbons and Chakraborti (2020), οι έλεγχοι που βασίζονται στο μήκος της μέγιστης ροής επιτυχιών ενδέχεται να είναι ισχυροί υπό προϋποθέσεις. Από τη σύγκριση με τον έλεγχο που βασίζεται στον συνολικό αριθμό ροών, διαπιστώνεται ότι και οι δύο έλεγχοι λαμβάνουν υπόψη τους μέρος της συνολικής διαθέσιμης πληροφορίας. Στον έλεγχο που βασίζεται στον συνολικό αριθμό ροών, αν και το μήκος της κάθε ροής επηρεάζει (έμμεσα) το συνολικό, δεν λαμβάνεται άμεσα υπόψη το μήκος κάθε ροής. Στον έλεγχο που βασίζεται στο μήκος της μέγιστης ροής επιτυχιών, το μήκος της μέγιστης ροής επηρεάζει (επίσης έμμεσα) τόσο το μήκος των υπόλοιπων ροών επιτυχιών όσο και τον συνολικό αριθμό ροών.

Παράδειγμα 7.5. (Gibbons and Chakraborti, 2020) Έστω ότι σε μια ακολουθία $n = 12$ το πλήθος δοκιμών παρατηρήθηκαν $n_1 = 6$ επιτυχίες. Χρησιμοποιήστε τη δεσμευμένη κατανομή της στατιστικής συνάρτησης $L_{12} < k|S = 6$ $(k = 1,2,...,7)$, όταν η H_0 είναι αληθής, και βρείτε την κρίσιμη περιοχή του ελέγχου, ώστε το επίπεδο σημαντικότητας να είναι (περίπου) ίσο με 5%.

Λύση Παραδείγματος 7.5. Από τα δεδομένα του παραδείγματος, έχουμε ότι $n_1 = n_2 = 6$. Θα χρησιμοποιήσουμε τη σχέση (7.1), για $k = 1, 2, ..., 7$ για να υπολογίσουμε τη ζητούμενη δεσμευμένη κατανομή. Άμεσα διαπιστώνουμε ότι $P(L_{12} < 1|S = 6) = 0$, αφού το ελάχιστο μήκος ροής επιτυχιών που μπορούμε να παρατηρήσουμε (όταν έχουν ήδη συμβεί 6 επιτυχίες σε 12 δοκιμές) είναι 1. Επίσης, $\binom{n}{n_2} = \binom{12}{6} = 924$. Με αντικατάσταση στη σχέση (7.1) και, έπειτα από κάποιες αλγεβρικές πράξεις, έχουμε ότι οι υπόλοιπες δεσμευμένες πιθανότητες είναι ίσες με

$$P(L_{12} < 2|S = 6) = \frac{7}{924} \approx 0.00758,\ P(L_{12} < 3|S = 6) = \frac{357}{924} \approx 0.38636,$$

$$P(L_{12} < 4|S = 6) = \frac{728}{924} \approx 0.78788,\ P(L_{12} < 5|S = 6) = \frac{875}{924} \approx 0.94697,$$

$$P(L_{12} < 6|S = 6) = \frac{917}{924} \approx 0.99242, \quad P(L_{12} < 7|S = 6) = 1.$$

Άρα, αν ορίσουμε την περιοχή απόρριψης της H_0 (υπόθεση τυχαιότητας του δείγματος) ως $K : L_{12} \geqslant 5$, δηλαδή, απόρριψη της H_0, αν σε 12 δοκιμές το μήκος της μέγιστης ροής επιτυχιών είναι τουλάχιστον 5, δεδομένου ότι έχουν συμβεί συνολικά 6 επιτυχίες, τότε το επίπεδο σημαντικότητας του ελέγχου είναι

$$P(L_{12} \geqslant 5|S = 6) = 1 - P(L_{12} < 5|S = 6) = 1 - 0.94697 = 0.05303,$$

δηλαδή είναι ίσο με 5.3% και, επομένως, είναι πολύ κοντά στο επιθυμητό. □

7.4 Έλεγχος συνεχόμενων ανοδικών καθοδικών ροών

Στην προηγούμενη ενότητα, παρουσιάστηκε το τεστ των ροών για τον έλεγχο της τυχαιότητας του δείγματος, όταν τα δεδομένα είτε είναι δίτιμα είτε μπορούν να μετατραπούν σε δίτιμα. Ειδικότερα, στην περίπτωση αριθμητικών δεδομένων, αυτό επιτυγχάνεται έχοντας μία τιμή αναφοράς (π.χ. τη διάμεσο, τη μέση τιμή ή οποιαδήποτε άλλη τιμή) για τον διαχωρισμό σε δύο ομάδες. Στην ενότητα αυτή, θα παρουσιαστεί ένας ακόμη έλεγχος τυχαιότητας που εφαρμόζεται σε αριθμητικά δεδομένα χωρίς να είναι απαραίτητο να οριστεί μία τιμή αναφοράς. Στο πλαίσιο αυτό, έστω $X_1, \ldots, X_n$, οι $n > 2$ το πλήθος διαθέσιμες παρατηρήσεις, διατεταγμένες σε χρονολογική σειρά συνήθως. Σχηματίζουμε τις διαδοχικές δειγματικές διαφορές $D_i = X_{i+1} - X_i$, για $i = 1, \ldots, n-1$, δηλαδή τις $D_1 = X_2 - X_1$, $D_2 = X_3 - X_2$, ..., $D_n = X_n - X_{n-1}$. Έπειτα, αγνοώντας τις μηδενικές διαφορές, αντικαθιστούμε τις θετικές διαφορές με το σύμβολο $+$, ενώ τις αρνητικές με το σύμβολο $-$.

Ορισμός 7.2

Ονομάζουμε **ανοδική ροή** μια ακολουθία αριθμών καθένας εκ των οποίων έπεται ενός μικρότερου αριθμού (ή, ισοδύναμα, ακολουθείται από έναν μεγαλύτερο αριθμό). Ονομάζουμε **καθοδική ροή** μια ακολουθία αριθμών καθένας εκ των οποίων έπεται ενός μεγαλύτερου αριθμού (ή, ισοδύναμα, ακολουθείται από έναν μικρότερο αριθμό).

Από τον προηγούμενο ορισμό, γίνεται σαφές ότι μια ανοδική ροή θα προσδιοριστεί από μια ακολουθία θετικών διαφορών, δηλαδή συμβόλων $+$ και θα τερματιστεί όταν εμφανιστεί αρνητική διαφορά. Δηλαδή τερματίζεται όταν διακόπτεται η αύξουσα ακολουθία τιμών από μία μικρότερη τιμή σε σχέση με την προηγούμενη. Από την άλλη, μια καθοδική ροή είναι μια ακολουθία αρνητικών διαφορών, δηλαδή συμβόλων $-$, η οποία θα τερματιστεί όταν εμφανιστεί θετική διαφορά. Δηλαδή τερματίζεται όταν διακόπτεται η φθίνουσα ακολουθία τιμών από μία μεγαλύτερη τιμή σε σχέση με την προηγούμενη. Τα παραπάνω θα διευκρινιστούν μέσω του παραδείγματος που ακολουθεί.

Παράδειγμα 7.6. Υπολογίστε τον αριθμό των ανοδικών και καθοδικών ροών για τα ακόλουθα σύνολα δεδομένων μεγέθους 10:

1ο σύνολο : 43, 49, 52, 55, 56, 59, 55, 57, 49, 50
2ο σύνολο : 49, 45, 38, 37, 36, 43, 31, 29, 27, 35
3ο σύνολο : 43, 45, 48, 49, 46, 43, 41, 39, 37, 35
4ο σύνολο : 43, 49, 42, 45, 43, 46, 45, 48, 47, 51.

Λύση Παραδείγματος 7.6. Υπολογίζουμε, αρχικά, τις διαδοχικές διαφορές (σε πλήθος ίσες με $n - 1 = 9$)

και από αυτές την ακολουθία συμβόλων + και −. Είναι τότε:

$$\begin{aligned}
\text{1ο σύνολο} &: +,+,+,+,+,-,+,-,+ \\
\text{2ο σύνολο} &: -,-,-,-,+,-,-,-,+ \\
\text{3ο σύνολο} &: +,+,+,-,-,-,-,-,-. \\
\text{4ο σύνολο} &: +,-,+,-,+,-,+,-,+.
\end{aligned}$$

Επομένως, στο 1ο σύνολο δεδομένων έχουμε 3 ανοδικές ροές μήκους 5, 1 και 1, αντίστοιχα, και 2 καθοδικές ροές μήκους 1. Στο 2ο σύνολο δεδομένων 2 καθοδικές ροές μήκους 4 και 3, αντίστοιχα, και 2 ανοδικές ροές μήκους 1. Στο 3ο σύνολο δεδομένων έχουμε 1 ανοδική ροή μήκους 3 και 1 καθοδική ροή μήκους 6, ενώ στο 4ο σύνολο έχουμε 5 ανοδικές ροές μήκους 1 και 4 καθοδικές ροές μήκους 1, αντίστοιχα. □

Ο έλεγχος αυτής της ενότητας βασίζεται στον συνολικό αριθμό των ανοδικών και καθοδικών ροών (βλ. Wallis and Moore, 1941). Όταν ο συνολικός αριθμός των ανοδικών και καθοδικών ροών, έστω U, είναι πολύ μικρός (βλ., για παράδειγμα, το 3ο σύνολο δεδομένων) σημαίνει ότι έχουμε είτε αυξανόμενες διαδοχικές τιμές είτε φθίνουσες διαδοχικές τιμές, γεγονός που υποδηλώνει απόκλιση από την τυχαιότητα. Ειδικότερα, σε αυτήν την περίπτωση θα έχουμε ένδειξη παρουσίας τάσης (*trend*) στα δεδομένα, είτε ανοδικής είτε καθοδικής. Από την άλλη, αν ο συνολικός αριθμός των ανοδικών και καθοδικών ροών είναι πολύ μεγάλος (βλ., για παράδειγμα, το 4ο σύνολο δεδομένων), αυτό σημαίνει ότι μια τιμή ακολουθείται από μικρότερη και, στη συνέχεια, από μεγαλύτερη και ούτω καθεξής. Σε αυτήν την περίπτωση, η συμπεριφορά των τιμών παρομοιάζεται με αυτήν της ταλάντωσης (*oscillation*) και αποτελεί ένδειξη παρουσίας κάποιου μη τυχαίου μηχανισμού σε ό,τι αφορά τις τιμές που παρατηρούνται. Προφανώς, και σε αυτήν την περίπτωση, έχουμε απόκλιση από την τυχαιότητα. Επομένως, ο έλεγχος που στηρίζεται στον συνολικό αριθμό των ανοδικών και καθοδικών ροών εντοπίζει δύο τύπους αποκλίσεων από την τυχαιότητα, αυτούς που στη βιβλιογραφία αναφέρονται ως ταλάντωση και ως τάση. Προφανώς, από όσα προηγήθηκαν, γίνεται αντιληπτό ότι ο έλεγχος που θα παρουσιαστεί σε αυτήν την ενότητα εντοπίζει τέτοιες αποκλίσεις, ενώ αποτυγχάνει σε άλλες.

Από τη συζήτηση που προηγήθηκε, συμπεραίνουμε ότι απορρίπτουμε τη μηδενική υπόθεση

$$H_0 : \text{οι διαδοχικές θετικές και αρνητικές μεταβολές είναι τυχαίες}$$

έναντι της εναλλακτικής

$$H_1 : \text{οι διαδοχικές θετικές και αρνητικές μεταβολές είναι μη τυχαίες}$$

αν $U \leqslant c_1$ ή $U \geqslant c_2$, όπου οι τιμές c_1 και c_2 θα πρέπει να προσδιοριστούν, έτσι ώστε να έχουμε έλεγχο σε επίπεδο σημαντικότητας a. Οι τιμές αυτές επιλέγονται να είναι το $1 - a/2$ και $a/2$ ποσοστιαίο σημείο, αντίστοιχα, τα οποία για να προσδιοριστούν απαιτείται η εύρεση της κατανομής του U, υπό τη μηδενική υπόθεση. Ο προσδιορισμός αυτός επιτυγχάνεται υπολογίζοντας όλες τις πιθανές διατάξεις και μεταθέσεις των αποτελεσμάτων + και − σε ένα δεδομένο πλήθος παρατηρήσεων. Για την απόδειξη στη γενική περίπτωση παραπέμπουμε, μεταξύ άλλων, στους Gibbons and Chakraborti (2020) και τις εκεί αναφορές. Στο ακόλουθο παράδειγμα απλώς θα δώσουμε τον τρόπο σκέψης για την εύρεσή της μέσω μιας ειδικής περίπτωσης.

Παράδειγμα 7.7. Για $n = 4$ να προσδιοριστεί η κατανομή του συνολικού αριθμού των ανοδικών και καθοδικών ροών, υπό τη μηδενική υπόθεση της τυχαιότητας του δείγματος και χωρίς την παρουσία δεσμών.

Λύση Παραδείγματος 7.7. Έστω x_1, x_2, x_3, x_4 οι τιμές στο δείγμα και ας υποθέσουμε ότι $x_1 < x_2 < x_3 < x_4$. Καθώς υπάρχουν $n = 4$ το πλήθος παρατηρήσεις, θα έχουμε $n - 1 = 3$ διαφορές με τα ακόλουθα πρόσημα.

Οι πιθανές τριάδες προσήμων δίνονται στον επόμενο πίνακα:

1	x_1	x_2	x_3	x_4	$+$	$+$	$+$
2	x_1	x_2	x_4	x_3	$+$	$+$	$-$
3	x_1	x_3	x_2	x_4	$+$	$-$	$+$
4	x_1	x_3	x_4	x_2	$+$	$+$	$-$
5	x_1	x_4	x_2	x_3	$+$	$-$	$+$
6	x_1	x_4	x_3	x_2	$+$	$-$	$-$
7	x_2	x_1	x_3	x_4	$-$	$+$	$+$
8	x_2	x_1	x_4	x_3	$-$	$+$	$-$
9	x_2	x_3	x_1	x_4	$+$	$-$	$+$
10	x_2	x_3	x_4	x_1	$+$	$+$	$-$
11	x_2	x_4	x_1	x_3	$+$	$-$	$+$
12	x_2	x_4	x_3	x_1	$+$	$-$	$-$
13	x_3	x_1	x_2	x_4	$-$	$+$	$+$
14	x_3	x_1	x_4	x_2	$-$	$+$	$-$
15	x_3	x_2	x_1	x_4	$-$	$-$	$+$
16	x_3	x_2	x_4	x_1	$-$	$+$	$-$
17	x_3	x_4	x_1	x_2	$+$	$-$	$+$
18	x_3	x_4	x_2	x_1	$+$	$-$	$-$
19	x_4	x_1	x_2	x_3	$-$	$+$	$+$
20	x_4	x_1	x_3	x_2	$-$	$+$	$-$
21	x_4	x_2	x_1	x_3	$-$	$-$	$+$
22	x_4	x_2	x_3	x_1	$-$	$+$	$-$
23	x_4	x_3	x_1	x_2	$-$	$-$	$+$
24	x_4	x_3	x_2	x_1	$-$	$-$	$-$

Υπό τη μηδενική υπόθεση της τυχαιότητας καθένα εκ των παραπάνω 24 αποτελεσμάτων (τριάδων) είναι ισοπίθανο. Παρατηρούμε ότι υπάρχουν δύο το πλήθος τριάδες με μια ροή ίδιων συμβόλων μήκους 3 (είναι οι $+++$, $---$), 12 το πλήθος τριάδες με ακριβώς μία ροή ίδιων συμβόλων μήκους 1 και μία ροή ίδιων συμβόλων μήκους δύο (π.χ. οι τριάδες $++-$, $+--$ κ.ά.), ενώ οι υπόλοιπες 10 το πλήθος τριάδες περιέχουν ακριβώς 3 ροές μήκους 1. Άρα, η από κοινού κατανομή του τυχαίου διανύσματος (V_1, V_2, V_3), όπου V_i είναι η τυχαία μεταβλητή που παριστάνει το πλήθος των ροών μήκους i, για $i = 1, 2, 3$, είναι:

$$P(V_1 = 0, V_2 = 0, V_3 = 1) = 2/24 \approx 0.0833,$$

$$P(V_1 = 1, V_2 = 1, V_3 = 0) = 12/24 = 0.5$$

$$P(V_1 = 3, V_2 = 0, V_3 = 0) = 10/24 \approx 0.4167,$$

ενώ για οποιαδήποτε άλλη τριάδα τιμών του τυχαίου διανύσματος (V_1, V_2, V_3), η αντίστοιχη πιθανότητα είναι μηδέν. Πλέον, με τη βοήθεια της παραπάνω κατανομής, μπορούμε να βρούμε τη μορφή της κρίσιμης περιοχής για τον έλεγχο που αφορά την τυχαιότητα του δείγματος. Για παράδειγμα, η πιθανότητα να παρατηρήσουμε το πολύ μία ανοδική (ή μία καθοδική) ροή σε ένα δείγμα μεγέθους $n = 4$, είναι 0.0833. Αντίστοιχα, οι πιθανότητες ο αριθμός αυτός να είναι τουλάχιστον 3 ή τουλάχιστον 2 είναι, αντίστοιχα, 0.4167 και $0.9167 = 0.5 + 0.4167$.

Επισημαίνεται ότι οι παραπάνω κρίσιμες τιμές συμπίπτουν με αυτές του Πίνακα Ε στο σύγγραμμα των Gibbons and Chakraborti (2020). □

Ένας έλεγχος τυχαιότητας μπορεί, επίσης, να αναπτυχθεί με βάση τον συνολικό αριθμό ανοδικών και καθοδικών ροών, λαμβάνοντας υπόψη το μήκος κάθε ροής. Η στατιστική συνάρτηση ελέγχου είναι ο συνολικός αριθμός $V = \sum_{i=1}^{n} V_i$, όπου V_i είναι το πλήθος των ροών μήκους i. Για την ασυμπτωτική μορφή του παραπάνω τεστ, έχουμε την παρακάτω πρόταση (βλ., επίσης, Levene, 1952).

Πρόταση 7.5. Έστω V ο συνολικός αριθμός των ανοδικών και καθοδικών ροών σε μια ακολουθία n το πλήθος παρατηρήσεων. Τότε, για $n \to \infty$, υπό τη μηδενική υπόθεση της τυχαιότητας του δείγματος:

$$W = \frac{V - \mu_V}{\sigma_V} \xrightarrow{d} \mathcal{N}(0,1),$$

όπου

$$\mu_V = \mathrm{E}(V) = \frac{2n-1}{3} \text{ και } \sigma_V^2 = \mathrm{Var}(V) = \frac{16n-29}{90}.$$

Απόδειξη Πρότασης 7.5. Παρατηρήστε, αρχικά, ότι από τον τρόπο ορισμού του στατιστικού V προκύπτει ότι μπορεί να γραφτεί στη μορφή του παρακάτω αθροίσματος:

$$V = 1 + \sum_{i=2}^{n-1} U_i,$$

όπου U_i, $i = 2,\ldots,n-1$, είναι μία δείκτρια συνάρτηση που λαμβάνει την τιμή 1, αν το i-οστό σύμβολο είναι διαφορετικό από το $(i-1)$-οστό σύμβολο, για $i = 2,\ldots,n-1$. Ισοδύναμα η δείκτρια συνάρτηση λαμβάνει την τιμή 1, όταν το πρόσημο της διαφοράς $D_i = X_{i+1} - X_i$ δεν είναι το ίδιο με το πρόσημο της διαφοράς $D_{i-1} = X_i - X_{i-1}$, για $i = 2,\ldots,n-1$. Είναι, λοιπόν, προφανές ότι το πρόσημο είναι διαφορετικό όταν είτε $X_{i+1} < X_i$ και $X_i > X_{i-1}$, αφού τότε $D_{i+1} < 0$ και $D_i > 0$, είτε όταν $X_{i+1} > X_i$ και $X_i < X_{i-1}$, αφού τότε $D_{i+1} > 0$ και $D_i < 0$. Από την άλλη, η δείκτρια λαμβάνει την τιμή 0 όταν $X_{i-1} > X_i > X_{i+1}$ ή $X_{i-1} < X_i < X_{i+1}$. Επομένως, καταλαβαίνουμε ότι η τιμή της δείκτριας μεταβλητής U_i, $i = 2,\ldots,n-1$, καθορίζεται από τη διάταξη της τριάδας των παρατηρήσεων X_{i-1}, X_i, X_{i+1}. Υπό την υπόθεση της τυχαιότητας του δείγματος υπάρχουν 6 διαφορετικές ισοπίθανες διατάξεις των X_{i-1}, X_i, X_{i+1}, οι

$$123,\ 132,\ 213,\ 231,\ 312,\ 321,$$

όπου με 3 συμβολίζουμε τη μεγαλύτερη τιμή και με 1 τη μικρότερη, μεταξύ των X_{i-1}, X_i, X_{i+1}. Από αυτές τις 6 ισοπίθανες περιπτώσεις, εκείνες που οδηγούν σε τιμή της δείκτριας ίση με 0 είναι εύκολα αντιληπτό ότι είναι οι 123 ($X_{i-1} < X_i < X_{i+1}$) και 321 ($X_{i-1} > X_i > X_{i+1}$). Δηλαδή σε 2 από τις 6 διατάξεις, η τιμή της δείκτριας είναι 0, ενώ στις υπόλοιπες 4 από τις 6 διατάξεις, η τιμή της είναι 1. Έτσι, άμεσα διαπιστώνουμε ότι

$$\begin{aligned}
\mathrm{E}(V) &= \mathrm{E}\left(1 + \sum_{i=2}^{n-1} U_i\right) = 1 + \sum_{i=2}^{n-1} \mathrm{E}(U_i) \\
&= 1 + \sum_{i=2}^{n-1} \left(1 \cdot \frac{4}{6} + 0 \cdot \frac{2}{6}\right) \\
&= 1 + \frac{2(n-2)}{3} = \frac{2n-1}{3}.
\end{aligned}$$

Από τον ορισμό της διακύμανσης έχουμε ότι:

$$\mathrm{Var}(V) = \mathrm{Var}\left(\sum_{i=2}^{n-1} U_i\right) = \mathrm{E}\left(\sum_{i=2}^{n-1} U_i\right)^2 - (\mathrm{E}(V))^2,$$

επομένως, αρκεί να υπολογίσουμε την $\mathrm{E}\left(\sum_{i=2}^{n-1} U_i\right)^2$. Είναι τότε (βλ. Kendall and Ord, 1973):

$$\begin{aligned}\mathrm{E}\left(\sum_{i=2}^{n-1} U_i\right)^2 &= (n-2)\mathrm{E}(U_i^2) + 2(n-3)\mathrm{E}(U_i U_{i+1}) \\ &+ 2(n-4)\mathrm{E}(U_i U_{i+2}) + (n-4)(n-5)\mathrm{E}(U_i U_{i+k})\end{aligned}$$

για $k \neq 0,1,2$. Από τον ορισμό της μέσης τιμής έχουμε:

$$\mathrm{E}(U_i^2) = 1^2 \cdot \frac{4}{6} + 0^2 \cdot \frac{2}{6} = \frac{2}{3}.$$

Για $k > 2$ έχουμε ότι οι δείκτριες U_i, U_{i+k}, είναι ανεξάρτητες τυχαίες μεταβλητές, καθώς εμπλέκονται στον ορισμό τους διαφορετικές τριάδες παρατηρήσεων. Επομένως, καθώς $\mathrm{E}(U_i) = 4/6$, έπεται ότι

$$\mathrm{E}(U_i U_{i+k}) = \mathrm{E}(U_i)\mathrm{E}(U_{i+k}) = \frac{4}{9}, \text{ για } k > 2.$$

Μένει να υπολογιστούν οι μέσες τιμές $\mathrm{E}(U_i U_{i+1})$ και $\mathrm{E}(U_i U_{i+2})$. Για τον υπολογισμό της πρώτης μέσης τιμής παρατηρήστε ότι το γινόμενο $U_i U_{i+1}$ λαμβάνει την τιμή 1, αν και οι δύο τ.μ. είναι ίσες με 1, ενώ σε κάθε άλλη περίπτωση λαμβάνει την τιμή 0. Επομένως, για την εύρεση της μέσης τιμής απαιτείται ο υπολογισμός της πιθανότητας και οι δύο τ.μ. να είναι ίσες με 1. Αυτό καθορίζεται από τη διάταξη των 4 διαδοχικών παρατηρήσεων. Υπό την υπόθεση της τυχαιότητας υπάρχουν $4! = 24$ ισοπίθανοι τρόποι, που είναι οι ακόλουθοι:

$$1234, 1243, 1342, 1324, 1432, 1423, 2134, 2143, 2341, 2314, 2431, 2413$$
$$3124, 3142.3241, 3214, 3412, 3421, 4123, 4132, 4213, 4231, 4312, 4321.$$

Από τις παραπάνω διατάξεις, αυτές που οδηγούν σε τιμή 1 για το γινόμενο των δεεκτριών είναι οι:

$$1324, 1423, 2143, 2314, 2413, 3142.3241, 3412, 4132, 4231.$$

Επομένως $\mathrm{E}(U_i U_{i+1}) = 10/24$. Για τον υπολογισμό της μέσης τιμής $\mathrm{E}(U_i U_{i+2})$ παρατηρήστε ότι τώρα εμπλέκονται 5 διαδοχικές παρατηρήσεις σε αυτόν τον υπολογισμό, με δυνατό αριθμό ισοπίθανων μεταθέσεων ίσο με $5! = 120$. Προκύπτει με παρόμοιο τρόπο ότι $\mathrm{E}(U_i U_{i+2}) = 54/120$.

Συνδυάζοντας τις παραπάνω σχέσεις, έπειτα από λίγη άλγεβρα, προκύπτει το ζητούμενο εφαρμόζοντας το Κ.Ο.Θ. □

Από την παραπάνω πρόταση έχουμε ότι στην περίπτωση του δίπλευρου ελέγχου, η H_0 απορρίπτεται σε επίπεδο σημαντικότητας a, αν $|W| \geq z_{a/2}$. Παρόμοια, για τους μονόπλευρους ελέγχους έχουμε τις κρίσιμες περιοχές $W \leqslant -z_a$ και $W \geqslant z_a$, αντίστοιχα.

Για μικρές τιμές του n συνιστάται κατά την εφαρμογή του ασυμπτωτικού ελέγχου, η χρήση της στατιστικής συνάρτησης που προκύπτει με εφαρμογή της διόρθωσης συνεχείας. Συγκεκριμένα, σε επίπεδο σημαντικότητας a, η H_0 απορρίπτεται, αν

$$W_{cc}^{(l)} = \frac{V + 0.5 - \mu_V}{\sigma_V} \leqslant -z_{a/2}$$

ή

$$W_{cc}^{(u)} = \frac{V - 0.5 - \mu_V}{\sigma_V} \geqslant z_{a/2}.$$

Οι μονόπλευροι έλεγχοι είναι $W_{cc}^{(l)} \leqslant -z_a$ και $W_{cc}^{(u)} \geqslant z_a$, αντίστοιχα.

Παράδειγμα 7.8. Ας υποθέσουμε ότι κατά τη διάρκεια ενός συγκεκριμένου μήνα, συνέβησαν 21 τροχαία ατυχήματα κατά μήκος ενός συγκεκριμένου τμήματος της Ολυμπίας Οδού. Οι 20 αποστάσεις μεταξύ των σημείων στα οποία συνέβησαν τα ατυχήματα (σε χιλιόμετρα) δίνονται παρακάτω:

$$0.3, 4.1, 4.2, 3.3, 1.9, 4.8, 0.3, 1.2, 0.8, 10.3, 1.2, 0.1, 10.0, 1.6, 27.6, 12.0, 14.2, 19.7, 15.5, 13.4$$

Αποτελούν τα παραπάνω στοιχεία ένδειξη ότι τα ατυχήματα κατανέμονται τυχαία κατά μήκος του αυτοκινητόδρομου; Να ελέγξετε την παραπάνω υπόθεση χρησιμοποιώντας (i) το τεστ ροών και (ii) τον έλεγχο συνεχόμενων ανοδικών καθοδικών ροών. Για καθένα από τα τεστ χρησιμοποιήστε ε.σ. 5%. Στην περίπτωση του ελέγχου με το τεστ των ροών να προσδιοριστεί το ακριβές επίπεδο σημαντικότητας, καθώς και η p-τιμή του, χρησιμοποιώντας τον Πίνακα Π.28 του Παραρτήματος. Επίσης, να προσδιοριστεί η p-τιμή των ελέγχων στα (i) και (ii), όταν χρησιμοποιείται η ασυμπτωτική κατανομή του τεστ.

Λύση Παραδείγματος 7.8. (i) Αρχικά, για να εφαρμόσουμε το τεστ των ροών, θα πρέπει να επιλέξουμε μία τιμή η οποία θα μετασχηματίσει τα αρχικά δεδομένα σε αποτελέσματα με δύο δυνατές τιμές. Θα χρησιμοποιηθεί η δειγματική διάμεσος, η οποία δεν είναι δύσκολο να διαπιστώσουμε ότι είναι ίση με 4.15. Άρα, η παραπάνω ακολουθία αποτελεσμάτων μετασχηματίζεται στην ακολουθία:

$$B\ B\ A\ B\ B\ A\ B\ B\ B\ A\ B\ B\ A\ B\ A\ A\ A\ A\ A\ A$$

όπου το αποτέλεσμα Β αντιστοιχεί σε τιμή μικρότερη της διαμέσου, ενώ το αποτέλεσμα Α αντιστοιχεί σε τιμή μεγαλύτερη της διαμέσου. Από την ακολουθία των αποτελεσμάτων Α, Β έπεται ότι το συνολικό πλήθος ροών είναι $R = 10$, ενώ το πλήθος τιμών που είναι μικρότερες της διαμέσου είναι $n_1 = 10$ και, άρα, $n_2 = n - n_1 = 20 - 10 = 10$. Σε ε.σ. 5%, η μορφή της κρίσιμης περιοχής από τον Πίνακα Π.28 του Παραρτήματος είναι $K : R \leqslant 6$ ή $R \geqslant 16$. Αφού η τιμή 10 δεν ανήκει στην κρίσιμη περιοχή, δεν απορρίπτουμε την υπόθεση της τυχαιότητας για το διαθέσιμο δείγμα. Άρα, δεν μπορούμε να απορρίψουμε την υπόθεση ότι τα ατυχήματα κατανέμονται τυχαία κατά μήκος του αυτοκινητόδρομου.

Το ακριβές ε.σ. είναι (περίπου) ίσο με 0.037044, καθώς προκύπτει, με χρήση της συνάρτησης πιθανότητας $P(R = r)$ υπό την H_0 για διάφορες τιμές του r, ότι η $P(R \leqslant 6) = 0.018522$, ενώ η $P(R \geqslant 16) = 0.018522$. Οι πιθανότητες αυτές μπορούν να υπολογιστούν και με τη βοήθεια της `R` χρησιμοποιώντας τις εντολές `pruns(6,10,10)` και `1-pruns(15,10,10)`.

Τέλος, η p-τιμή του ελέγχου είναι $2 \cdot \min\{P(R \leqslant 10), P(R \geqslant 10)\}$, όπου με χρήση της συνάρτησης πιθανότητας $P(R = r)$ υπό την H_0 για διάφορες τιμές του r ή με χρήση της `R`, προκύπτει ότι η $P(R \leqslant 10) = 0.4141$, ενώ η $P(R \geqslant 10) = 0.7578$. Άρα, η p-τιμή είναι ίση με $2 \cdot 0.4141 = 0.8282$.

Αξίζει, επίσης, να αναφέρουμε ότι αν χρησιμοποιήσουμε τον ασυμπτωτικό έλεγχο, τότε η μορφή της στατιστικής συνάρτησης ελέγχου είναι

$$Z = \frac{R - \mu_R}{\sigma_R},$$

με

$$\mu_R = 1 + \frac{2n_1n_2}{n} = 1 + \frac{2 \cdot 10 \cdot 10}{20} = 11,$$

και

$$\sigma_R^2 = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)} = \frac{2 \cdot 10 \cdot 10(2 \cdot 10 \cdot 10 - 20)}{20^2(20-1)} \approx 4.7368.$$

Άρα, με αντικατάσταση έχουμε ότι:

$$z = \frac{10 - 11}{\sqrt{4.7368}} \approx -0.4595.$$

Αφού ο έλεγχος είναι δίπλευρος, η p-τιμή του ελέγχου είναι $2(1 - \Phi(|-0.4595|)) \approx 0.6459 \not< 0.05$. Άρα, σε ε.σ. 5%, δεν απορρίπτουμε την H_0 και, άρα, δεν μπορούμε να απορρίψουμε την υπόθεση ότι τα ατυχήματα κατανέμονται τυχαία κατά μήκος του αυτοκινητόδρομου.

(ii) Θα ελέγξουμε την τυχαιότητα του δείγματος χρησιμοποιώντας τον έλεγχο συνεχόμενων ανοδικών καθοδικών ροών. Σχηματίζουμε τις διαφορές $D_i = X_{i+1} - X_i$ δίνοντας το αποτέλεσμα +, αν $D_i > 0$, και −, αν $D_i < 0$. Έτσι, η ακολουθία συμβόλων + και − είναι

$$+ \; + \; - \; - \; + \; - \; + \; - \; + \; - \; - \; + \; - \; + \; - \; + \; + \; - \; -$$

Δεν είναι δύσκολο να διαπιστώσουμε ότι το συνολικό πλήθος ανοδικών και καθοδικών ροών είναι $V = 14$. Συγκεκριμένα, υπάρχουν 7 ροές συμβόλων + με μήκη 2, 1, 1, 1, 1, 1 και 2, αντίστοιχα, καθώς και 7 ροές συμβόλων − με μήκη 2, 1, 1, 2, 1, 1 και 2, αντίστοιχα. Επίσης, είναι

$$\mu_V = \frac{2n-1}{3} = \frac{2 \cdot 20 - 1}{3} = 13,$$

ενώ

$$\sigma_V^2 = \frac{16n - 29}{90} = \frac{16 \cdot 20 - 29}{90} \approx 3.2333.$$

Άρα, για την ασυμπτωτική μορφή του ελέγχου, χρησιμοποιείται η στατιστική συνάρτηση $W = (V - \mu_V)/\sigma_V$ και με αντικατάσταση η τιμή της είναι $w = (14 - 13)/\sqrt{3.2333} \approx 0.5561$.

Αφού ο έλεγχος είναι δίπλευρος, η p-τιμή του ελέγχου είναι $2(1 - \Phi(|0.5561|)) = 0.5781 > 0.05$. Άρα, σε ε.σ. 5%, δεν απορρίπτεται η υπόθεση ότι το δείγμα είναι τυχαίο. □

Παρατήρηση 7.9. Ένας ισοδύναμος έλεγχος με αυτόν που παρουσιάστηκε σε αυτήν την ενότητα είναι ο έλεγχος που βασίζεται στο σύνολο των σημείων αλλαγής (turning points), όπου σημείο αλλαγής σε μια τριάδα διαδοχικών παρατηρήσεων X_{i-1}, X_i και X_{i+1} εμφανίζεται όταν η μεσαία παρατήρηση είναι η μεγαλύτερη ή η μικρότερη. Τότε εύκολα προκύπτει ότι ο συνολικός αριθμός των σημείων αλλαγής, έστω K', είναι ίσος με $V - 1$. Για περισσότερες λεπτομέρειες βλέπε Stuart (1954). Επομένως, όλα τα θεωρητικά αποτελέσματα που παρουσιάστηκαν για τη στατιστική συνάρτηση V μπορούν εύκολα να χρησιμοποιηθούν και να εξαχθούν αντίστοιχα αποτελέσματα για τη στατιστική συνάρτηση K'.

7.5 Έλεγχος σημείων πρώτων διαφορών των Moore και Wallis

Στην ενότητα αυτή θα παρουσιαστεί ένας ακόμη έλεγχος τυχαιότητας που εφαρμόζεται σε αριθμητικά δεδομένα χωρίς να είναι απαραίτητο να οριστεί μια τιμή αναφοράς. Στο πλαίσιο αυτό, έστω $X_1, \ldots, X_n$, οι $n > 2$ το πλήθος διαθέσιμες παρατηρήσεις, διατεταγμένες σε χρονολογική σειρά συνήθως. Σχηματίζουμε τις διαδοχικές δειγματικές διαφορές $D_i = X_{i+1} - X_i$, για $i = 1, \ldots, n-1$, δηλαδή τις $D_1 = X_2 - X_1$, $D_2 = X_3 - X_2$, ..., $D_n = X_n - X_{n-1}$. Έπειτα, αγνοώντας τις μηδενικές διαφορές, αντικαθιστούμε τις θετικές διαφορές με το σύμβολο +, ενώ τις αρνητικές με το σύμβολο −.

Για να γίνει η παραπάνω διαδικασία κατανοητή, θεωρούμε τα παρακάτω σύνολα δεδομένων μεγέθους 10:

1ο σύνολο	:	$43, 49, 52, 55, 56, 59, 55, 57, 49, 50$
2ο σύνολο	:	$49, 45, 38, 37, 36, 43, 31, 29, 27, 35$
3ο σύνολο	:	$43, 45, 48, 49, 46, 43, 41, 39, 37, 35$
4ο σύνολο	:	$43, 49, 42, 45, 43, 46, 45, 48, 47, 51.$

Έπειτα υπολογίζουμε, αρχικά, τις διαδοχικές διαφορές (σε πλήθος ίσες με $n - 1 = 9$) και από αυτές την ακολουθία συμβόλων + και −. Είναι:

$$\begin{aligned}
\text{1ο σύνολο} &: +,+,+,+,+,-,+,-,+\\
\text{2ο σύνολο} &: -,-,-,-,+,-,-,-,+\\
\text{3ο σύνολο} &: +,+,+,-,-,-,-,-,-.\\
\text{4ο σύνολο} &: +,-,+,-,+,-,+,-,+.
\end{aligned}$$

Σε αυτό το πλαίσιο, ο έλεγχος αυτής της ενότητας που προτάθηκε από τους Moore and Wallis (1943) βασίζεται στον αριθμό των συμβόλων +, δηλαδή στο πρόσημο των διαδοχικών διαφορών. Όταν ο συνολικός αριθμός των θετικών διαδοχικών διαφορών, έστω T, είναι πολύ μεγάλος, για παράδειγμα όπως στο 1ο σύνολο δεδομένων, σημαίνει ότι στα δεδομένα μας παρατηρείται η τάση να διαδέχεται μια τιμή μια μεγαλύτερή της. Από την άλλη πλευρά, όταν ο συνολικός αριθμός των θετικών διαδοχικών διαφορών είναι πολύ μικρός, τότε στα δεδομένα έχουμε φθίνουσες διαδοχικές τιμές, όπως για παράδειγμα στο 2ο σύνολο δεδομένων. Επομένως, ο έλεγχος που θα απορρίπτει την υπόθεση της τυχαιότητας για πολύ μεγάλες ή πολύ μικρές τιμές της στατιστικής συνάρτησης T θα εντοπίζει δύο από τους πιθανούς τύπους αποκλίσεων από την τυχαιότητα.

Παρατήρηση 7.10. Όπως είναι προφανές, ο έλεγχος, που στηρίζεται στη στατιστική συνάρτηση T και απορρίπτει την υπόθεση της τυχαιότητας για πολύ μικρές ή μεγάλες τιμές της, δεν μπορεί να εντοπίσει όλες τις δυνατές αποκλίσεις από την τυχαιότητα. Για παράδειγμα, τόσο για το 3ο σύνολο δεδομένων όσο και για το 4ο σύνολο δεδομένων, παρότι υπάρχει ένα μοτίβο που επαναλαμβάνεται, ο αριθμός των συμβόλων + δεν μπορεί να θεωρηθεί ούτε πολύ μεγάλος ούτε πολύ μικρός. Ωστόσο, ανατρέχοντας στην προηγούμενη ενότητα μπορείτε να διαπιστώσετε ότι ο εκεί προτεινόμενος έλεγχος μπορεί να εντοπίσει τέτοιου είδους αποκλίσεις. Αυτή η παρατήρηση δικαιολογεί πλήρως την ύπαρξη διαφορετικών ελέγχων τυχαιότητας.

Από τη συζήτηση που προηγήθηκε, συμπεραίνουμε ότι απορρίπτουμε τη μηδενική υπόθεση της τυχαιότητας του δείγματος έναντι της εναλλακτικής της μη τυχαιότητας (δίπλευρος έλεγχος), αν $T \leqslant c_1$ ή $T \geqslant c_2$, όπου οι τιμές c_1 και c_2 θα πρέπει να προσδιοριστούν έτσι ώστε να έχουμε έλεγχο με επίπεδο σημαντικότητας a. Οι τιμές αυτές επιλέγονται να είναι το $1 - a/2$ και $a/2$ ποσοστιαίο σημείο, αντίστοιχα, τα οποία, για να προσδιοριστούν, απαιτείται η εύρεση της κατανομής της στατιστικής συνάρτησης T, υπό τη μηδενική υπόθεση.

Αναδρομική σχέση για την εύρεση της κατανομής της T υπό την H_0 έχει δοθεί στη γενική περίπτωση από τους Moore and Wallis (1943). Στο ακόλουθο παράδειγμα θα βρούμε τον τρόπο εύρεσής της σε μια ειδική περίπτωση.

Παράδειγμα 7.9. Έστω x_1, x_2, x_3, x_4 οι τιμές στο δείγμα και έστω, επίσης, ότι $x_1 < x_2 < x_3 < x_4$. Για $n = 4$ να προσδιοριστεί η κατανομή του αριθμού θετικών διαδοχικών διαφορών T, υπό την υπόθεση της τυχαιότητας του δείγματος.

Λύση Παραδείγματος 7.9. Καθώς υπάρχουν $n = 4$ το πλήθος παρατηρήσεις, θα έχουμε 24 διαφορετικές διατάξεις των διαθέσιμων παρατηρήσεων:

$$\begin{aligned}
&1234, 1243, 1342, 1324, 1432, 1423, 2134, 2143, 2341, 2314, 2431, 2413\\
&3124, 3142.3241, 3214, 3412, 3421, 4123, 4132, 4213, 4231, 4312, 4321.
\end{aligned}$$

Υπό τη μηδενική υπόθεση της τυχαιότητας κάθε διάταξη είναι ισοπίθανη. Υπολογίζουμε για κάθε δυνατή διάταξη των 4 τιμών το πλήθος των θετικών συμβόλων και είναι:

$$3,2,2,2,1,2,2,1,2,2,1,2,2,1,1,1,2,1,2,1,1,1,1,0,$$

αντίστοιχα. Επομένως, η τυχαία μεταβλητή T, που παριστάνει τον συνολικό αριθμό των θετικών συμβόλων, έχει δυνατές τιμές: 0,1,2,3, με αντίστοιχες πιθανότητες

$$P(T = 0) = P(T = 3) = \frac{1}{24}$$

και

$$P(T=1)=P(T=2)=\frac{11}{24}$$

□

Πίνακες κρίσιμων τιμών για το παραπάνω τεστ είναι διαθέσιμοι στη βιβλιογραφία, για παράδειγμα στην εργασία των Moore and Wallis (1943), όπου έχουμε διαθέσιμες κρίσιμες τιμές για $n \leqslant 12$. Λαμβάνοντας, επιπρόσθετα, υπόψη ότι ο παραπάνω έλεγχος οδηγεί σε περιορισμένο αριθμό επιλογών ακριβούς επιπέδου σημαντικότητας οδηγούμαστε στην εύρεση ενός προσεγγιστικού στατιστικού τεστ.

Πρόταση 7.6. Έστω T ο συνολικός αριθμός των θετικών ανά δύο διαφορών σε μια ακολουθία n το πλήθος τιμών. Τότε, για $n \to \infty$, υπό τη μηδενική υπόθεση της τυχαιότητας του δείγματος:

$$V=\frac{T-\mu_T}{\sigma_T}\xrightarrow{d}\mathcal{N}(0,1),$$

όπου

$$\mu_T=\mathsf{E}(T)=\frac{n-1}{2}\text{ και }\sigma_T^2=\mathsf{Var}(T)=\frac{n+1}{12}.$$

Απόδειξη Πρότασης 7.6. Παρατηρήστε, αρχικά, ότι από τον τρόπο ορισμού του στατιστικού T προκύπτει ότι μπορεί να γραφτεί στη μορφή του παρακάτω αθροίσματος:

$$T=\sum_{i=1}^{n-1}T_i,$$

όπου T_i, $i=1,\ldots,n-1$, μία τυχαία μεταβλητή που λαμβάνει την τιμή 1, όταν το πρόσημο της διαφοράς $D_i=X_{i+1}-X_i$, $i=1,\ldots,n-1$ είναι θετικό, και την τιμή 0, διαφορετικά. Είναι, λοιπόν, προφανές ότι το πρόσημο είναι θετικό όταν $X_{i+1}>X_i$. Επομένως, καταλαβαίνουμε ότι η τιμή της δείκτριας T_i, $i=1,\ldots,n-1$, καθορίζεται από τη διάταξη των παρατηρήσεων X_i, X_{i+1}. Υπό την υπόθεση της τυχαιότητας του δείγματος υπάρχουν 2 ισοπίθανες διαφορετικές διατάξεις και, επομένως,

$$\mathsf{E}(T)=\mathsf{E}\left(\sum_{i=1}^{n-1}T_i\right)=\sum_{i=1}^{n-1}\mathsf{E}(T_i)=\sum_{i=1}^{n-1}\left(1\cdot\frac{1}{2}+0\cdot\frac{1}{2}\right)=\frac{n-1}{2}.$$

Για τη διακύμανση έχουμε:

$$\begin{aligned}\mathsf{Var}(T) &= \mathsf{Var}\left(\sum_{i=1}^{n-1}T_i\right)=\sum_{i=1}^{n-1}\mathsf{Var}(T_i)+\sum_{1\le i\neq j\le(n-1)}\mathsf{Cov}\left(T_i,T_j\right)\\ &= \sum_{i=1}^{n-1}\left(\mathsf{E}(T_i^2)-(\mathsf{E}(T_i))^2\right)+\sum_{1\le i\neq j\le(n-1)}\mathsf{Cov}\left(T_i,T_j\right).\end{aligned}$$

Είναι $\mathsf{E}(T_i^2)=1^2\cdot 0.5+0^2\cdot 0.5=0.5$ και, άρα, $\mathsf{E}(T_i^2)-(\mathsf{E}(T_i))^2=1/4$. Επιπλέον, παρατηρήστε ότι οι δείκτριες T_i, T_{i+k}, για $k>1$ είναι ανεξάρτητες τ.μ., καθώς δεν εμπλέκονται στον υπολογισμό τους κοινές παρατηρήσεις και, επομένως, σε αυτήν την περίπτωση είναι $\mathsf{Cov}(T_i,T_{i+k})=0$. Από την άλλη πλευρά, οι τ.μ. T_i, T_{i+1} είναι συσχετισμένες και ισχύει ότι:

$$\mathsf{Cov}(T_i,T_{i+1})=\mathsf{E}(T_iT_{i+1})-\mathsf{E}(T_i)\cdot\mathsf{E}(T_{i+1})=1\cdot P(T_iT_{i+1}=1)-0.25.$$

Η πιθανότητα $P(T_iT_{i+1}=1)$ καθορίζεται από τη διάταξη των τριών διαδοχικών παρατηρήσεων X_i, X_{i+1}, X_{i+2}. Υπό την υπόθεση της τυχαιότητας υπάρχουν $3!=6$ ισοπίθανοι τρόποι διάταξης αυτών, που είναι οι ακόλουθοι:

$$123, 132, 213, 231, 312, 321.$$

Παρατηρώντας ότι $T_iT_{i+1} = 1$ μόνο στην πρώτη περίπτωση, έχουμε ότι $P(T_iT_{i+1} = 1) = \frac{1}{6}$. Λαμβάνοντας τα προηγούμενα υπόψη, έχουμε ότι:

$$\text{Var}(T) = \sum_{i=1}^{n-1}\left(\frac{1}{4} + 2(n-2)\cdot\left(\frac{1}{6} - \frac{1}{4}\right)\right) = \frac{n-1}{4} - \frac{2(n-2)}{12} = \frac{n+1}{12}.$$

Συνδυάζοντας τις παραπάνω σχέσεις, έπειτα από λίγη άλγεβρα, προκύπτει το ζητούμενο εφαρμόζοντας το Κ.Ο.Θ. □

Από την παραπάνω πρόταση έχουμε ότι ο δίπλευρος έλεγχος απορρίπτεται σε επίπεδο σημαντικότητας a, αν $|V| \geq z_{a/2}$. Παρόμοια, για τους μονόπλευρους ελέγχους έχουμε τις κρίσιμες περιοχές $V \leqslant -z_a$ και $V \geqslant z_a$, αντίστοιχα.

Για μικρές τιμές του n συνιστάται η χρήση της στατιστικής συνάρτησης που προκύπτει με εφαρμογή της διόρθωσης συνεχείας και απορρίπτει τη μηδενική υπόθεση, αν:

$$V_{cc}^{(l)} = \frac{T + 0.5 - \mu_T}{\sigma_T} \leqslant -z_{a/2}$$

ή

$$V_{cc}^{(u)} = \frac{T - 0.5 - \mu_T}{\sigma_T} \geqslant z_{a/2}.$$

Οι μονόπλευροι έλεγχοι είναι $V_{cc}^{(l)} \leqslant -z_a$ και $V_{cc}^{(u)} \geqslant z_a$, αντίστοιχα.

Παράδειγμα 7.10. Στη διάθεσή μας έχουμε τα ημερήσια νέα κρούσματα ενός κορονοϊού για 25 συνεχόμενες ημέρες, όπως αυτά καταγράφηκαν από τον επίσημο φορέα δημόσιας υγείας μιας συγκεκριμένης χώρας.

$$6, 9, 34, 57, 38, 103, 21, 32, 31, 46, 31, 35, 94, 71, 48, 78, 71, 74, 95, 96, 56, 102, 61, 139, 99$$

Χρησιμοποιήστε την ασυμπτωτική μορφή του ελέγχου των Moore και Wallis και ελέγξτε την υπόθεση της τυχαιότητας του δείγματος, σε ε.σ. 5%.

Λύση Παραδείγματος 7.10. Η στατιστική συνάρτηση ελέγχου T για τον έλεγχο των Moore και Wallis ισούται με το πλήθος των θετικών διαφορών. Από την ακολουθία των διαδοχικών διαφορών

$$+ \; + \; + \; - \; + \; - \; + \; - \; + \; - \; + \; + \; - \; - \; + \; - \; + \; + \; + \; - \; + \; - \; + \; -$$

διαπιστώνουμε ότι $T = 14$. Επιπλέον, είναι $n = 25$ και, επομένως

$$\mu_T = \frac{n-1}{2} = \frac{25-1}{2} = 12, \; \sigma_T^2 = \frac{n+1}{12} = \frac{25+1}{12} = 2.1667.$$

Για την ασυμπτωτική μορφή του τεστ, χρησιμοποιείται η στατιστική συνάρτηση $Z = (T - \mu_T)/\sigma_T$, για την οποία γνωρίζουμε ότι υπό την H_0 ακολουθεί (προσεγγιστικά) την τυπική κανονική $\mathcal{N}(0,1)$ κατανομή. Σε ε.σ. 5% η μορφή της κρίσιμης περιοχής είναι $K : |Z| \geqslant 1.96$. Με αντικατάσταση, έπεται ότι η τιμή της Z είναι $(14 - 12)/\sqrt{2.1667} \approx 1.3587$, η οποία δεν ανήκει στην κρίσιμη περιοχή. Άρα, σε ε.σ. 5%, δεν απορρίπτεται η υπόθεση της τυχαιότητας του δείγματος. □

7.6 Mann-Kendal rank test

Στην ενότητα αυτή, θα παρουσιαστεί ένας ακόμη έλεγχος τυχαιότητας, που ίσως είναι και ένας από τους πιο διαδεδομένους σε πρακτικές εφαρμογές από την περιοχή της μετεωρολογίας και της υδρολογίας. Ο έλεγχος αυτός είναι γνωστός ως Mann-Kendall test , καθώς παρότι προτάθηκε από τον Mann (1945), συνδέεται άμεσα με ένα μέτρο συσχέτισης που προτάθηκε από τον Kendall (1938) και για το οποίο θα δοθούν λεπτομέρειες στο επόμενο κεφάλαιο.

Έστω $X_1, \dots, X_n$, οι $n > 2$ το πλήθος διαθέσιμες παρατηρήσεις, διατεταγμένες σε χρονολογική σειρά. Υποθέτουμε ότι οι παρατηρήσεις είναι ποσοτικές (από διακριτή ή από συνεχή κατανομή). Στο πλαίσιο αυτό, ο Mann (1945) πρότεινε έναν τρόπο ελέγχου της υπόθεσης ότι τα δεδομένα αποτελούν ένα τυχαίο δείγμα έναντι της εναλλακτικής ότι αποκλίνουν από την τυχαιότητα και παρουσιάζουν μια μονότονη τάση στην πάροδο του χρόνου. Αυτό ουσιαστικά σημαίνει ότι ο έλεγχος μπορεί να εντοπίσει αν υπάρχει μια πτωτική ή μια αυξητική τάση. Με τον όρο πτωτική (αυξητική) τάση εννοούμε ότι η μεταβλητή παίρνει συνεχώς μικρότερες (μεγαλύτερες) τιμές με την πάροδο του χρόνου, χωρίς απαραίτητα η μεταβολή αυτή να είναι γραμμική. Είναι σαφές, λοιπόν, ότι και αυτός ο έλεγχος προτάθηκε για να εντοπίζει συγκεκριμένους τύπους απόκλισης από την τυχαιότητα. Για παράδειγμα, αν υπάρχει μια κυκλική τάση, δηλαδή μια τάση εμφάνισης αρχικά τιμών που αυξάνονται και στη συνέχεια μειώνονται, ο συγκεκριμένος έλεγχος δεν μπορεί να εντοπίσει αυτήν τη μη τυχαία συμπεριφορά.

Με αυτόν το σκοπό, η σταστιστική συνάρτηση που προτάθηκε από τον Mann (1945) είναι η ακόλουθη:

$$\text{MK} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn\left(X_j - X_i\right) \tag{7.2}$$

όπου

$$sgn(X_j - X_i) = \begin{cases} 1, \text{ αν } X_j - X_i > 0 \\ 0, \text{ αν } X_j - X_i = 0 \\ -1, \text{ αν } X_j - X_i < 0 \end{cases} \tag{7.3}$$

Ο τρόπος υπολογισμού της στατιστικής συνάρτησης MK επεξηγείται στο παράδειγμα που ακολουθεί (Kendall, 1938).

Παράδειγμα 7.11. Υπολογίστε την τιμή της στατιστικής συνάρτησης MK για τα ακόλουθα δεδομένα

$$4,7,2,10,3,6,8,1,5,9.$$

Λύση Παραδείγματος 7.11. Ξεκινάμε από την τιμή 4 και υπολογίζουμε τις διαφορές αυτής της τιμής από τις υπόλοιπες 9 το πλήθος τιμές. Άμεσα μπορούμε να διαπιστώσουμε πως οι διαφορές αυτές είναι:

$$3,-2,6,-1,2,4,-3,1,5$$

και, επομένως, $\sum_{j=2}^{10} sgn\left(X_j - X_1\right) = 3$. Συνεχίζοντας με ανάλογο τρόπο, θεωρούμε τη 2η τιμή στο δείγμα (είναι η τιμή 7) και υπολογίζουμε τις διαφορές αυτής της τιμής από τις υπόλοιπες 8 το πλήθος τιμές. Είναι τότε:

$$-5,3,-4,-1,1,-6,-2,2$$

και, συνεπώς, $\sum_{j=3}^{10} sgn\left(X_j - X_2\right) = -2$. Συνεχίζοντας κατά τον ίδιο τρόπο, έχουμε

$$\sum_{j=4}^{10} sgn\left(X_j - X_3\right) = 5, \quad \sum_{j=5}^{10} sgn\left(X_j - X_4\right) = -6,$$

$$\sum_{j=6}^{10} sgn\left(X_j - X_5\right) = 3, \quad \sum_{j=7}^{10} sgn\left(X_j - X_6\right) = 0,$$

$$\sum_{j=8}^{10} sgn\left(X_j - X_7\right) = -1, \quad \sum_{j=9}^{10} sgn\left(X_j - X_8\right) = 2, \quad \sum_{j=10}^{10} sgn\left(X_j - X_9\right) = 1.$$

Επομένως, προκύπτει ότι ΜΚ $= 3 - 2 + 5 - 6 + 3 + 0 - 1 + 2 + 1 = 5$.

Ένας εναλλακτικός και πιο εύκολος τρόπος υπολογισμού (όταν δεν υπάρχουν δεσμοί) είναι να σκεφτούμε ως εξής. Ο αριθμός 4 (η 1η παρατήρηση στο δείγμα) έχει δεξιά του 6 αριθμούς μεγαλύτερους (τους 7, 10, 6, 8, 5, 9), ο αριθμός 7 (η 2η παρατήρηση στο δείγμα) έχει 3 αριθμούς μεγαλύτερους (τους 10, 8, 9) και συνεχίζοντας με τον ίδιο τρόπο προκύπτει ότι καθένας εκ των $n - 1 = 9$ αριθμών έχει δεξιά του:

$$6,\ 3,\ 6,\ 0,\ 4,\ 2,\ 1,\ 2,\ 1,$$

αντίστοιχα αριθμούς μεγαλύτερους. Οι παραπάνω αριθμοί αθροίζουν στο 25, ενώ οι συνολικοί αριθμοί που συγκρίνονται είναι: $9 + 8 + ... + 1 = 45$ (γενικά $1 + 2 + ... + (n - 1) = \binom{n}{2}$). Επομένως, υπάρχουν και $45 - 25 = 20$, αριθμοί μικρότεροι. Έτσι έχουμε ότι ΜΚ $= 25 - 20$. Εύκολα αποδεικνύεται (αφήνεται ως άσκηση στον/στην αναγνώστη/στρια) ότι, αν k είναι το πλήθος των αριθμών που με τη διαδικασία αυτή είναι μεγαλύτεροι προς τα δεξιά, τότε ΜΚ $= 2k - \frac{n(n-1)}{2}$. □

Παρατήρηση 7.11. Ο συντελεστής συσχέτισης που προτάθηκε από τον Kendall (1938) προκύπτει ως πηλίκο του παραπάνω αθροίσματος (βλ. σχέση (7.2)) ως προς $n(n-1)/2$, όπου ο παρονομαστής είναι η μέγιστη πιθανή τιμή του αθροίσματος ΜΚ.

Παρατηρήστε ότι η λογική της στατιστικής συνάρτησης ΜΚ βασίζεται στις ανά δύο συγκρίσεις κάθε δειγματικού σημείου με όλες τις προηγούμενες δειγματικές τιμές και στον καθορισμό του αριθμού των παρατηρήσεων που έχουμε αύξηση, μείωση ή δεσμό (ισοβαθμία). Μια πολύ μεγάλη θετική τιμή της στατιστικής συνάρτησης θα υποδηλώνει ύπαρξη αυξητικής τάσης στο σύνολο των δεδομένων. Από την άλλη πλευρά, μια μικρή αρνητική τιμή (μεγάλη κατά απόλυτη τιμή) θα υποδηλώνει ύπαρξη πτωτικής (ή αλλιώς μειούμενης) τάσης στο σύνολο δεδομένων. Τέλος, μια τιμή κοντά στο μηδέν θα υποδεικνύει ίσο αριθμό θετικών ή αρνητικών διαφορών, γεγονός που θα αναμενόταν αν οι μετρήσεις ήταν τυχαία κυμαινόμενες γύρω από έναν σταθερό μέσο όρο χωρίς κάποια τάση.

Από τη συζήτηση που προηγήθηκε, συμπεραίνουμε ότι απορρίπτουμε τη μηδενική υπόθεση της τυχαιότητας του δείγματος έναντι της εναλλακτικής της ύπαρξης πτωτικής ή αυξητικής τάσης, αν ΜΚ $\leqslant c_1$ ή ΜΚ $\geqslant c_2$, όπου οι τιμές c_1 και c_2 θα πρέπει να προσδιοριστούν, έτσι ώστε να έχουμε έλεγχο σε επίπεδο σημαντικότητας a. Οι τιμές αυτές επιλέγονται να είναι το $1 - a/2$ και $a/2$ ποσοστιαίο σημείο, αντίστοιχα, τα οποία για να προσδιοριστούν απαιτείται η εύρεση της κατανομής της ΜΚ, υπό τη μηδενική υπόθεση. Οι μονόπλευροι έλεγχοι με εναλλακτικές την ύπαρξη αυξητικής ή πτωτικής τάσης διαμορφώνονται κατά αντιστοιχία με τα προηγούμενα. Για να επιτευχθούν τα παραπάνω απαιτείται η εύρεση της κατανομής της στατιστικής συνάρτησης ΜΚ υπό την υπόθεση της τυχαιότητας του δείγματος. Για τιμές του μεγέθους δείγματος $1 \leqslant n \leqslant 10$ παρατίθεται από τον Kendall (1938) πίνακας υπολογισμού των πιθανοτήτων των δυνατών τιμών της ΜΚ. Ο υπολογισμός για μεγαλύτερες τιμές είναι αρκετά περίπλοκος παρότι έχει δοθεί ένας αναδρομικός τύπος εύρεσης από τον Kendall (1938). Στο επόμενο παράδειγμα θα δοθεί ο τρόπος σκέψης για την ειδική περίπτωση που δεν υπάρχουν ισοβαθμίες και το μέγεθος δείγματος είναι $n = 3$.

Παράδειγμα 7.12. Για $n = 3$ να προσδιοριστεί η κατανομή της στατιστικής συνάρτησης ΜΚ, υπό τη μηδενική υπόθεση της τυχαιότητας του δείγματος και θεωρώντας ότι δεν υπάρχουν ισοβαθμίες μεταξύ των δεδομένων.

Λύση Παραδείγματος 7.12. Υπό την υπόθεση της τυχαιότητας υπάρχουν $3! = 6$ ισοπίθανοι τρόποι διάταξης των παρατηρήσεων, που είναι οι ακόλουθοι:

$$123,\ 132,\ 213,\ 231,\ 321,\ 312.$$

Οι τιμές της στατιστικής συνάρτησης για καθένα σύνολο είναι:

$$1+1+1=3,\ 1+1-1=1,\ -1+1+1=1,\ +1-1-1=-1,\ -1-1-1=-3,\ -1-1+1=-1$$

Επομένως, έχουμε $P(\text{MK}=1) = P(\text{MK}=-1) = 2/6$, $P(\text{MK}=3) = P(\text{MK}=-3) = 1/6$. □

Πίνακες ποσοστιαίων σημείων, όπως προαναφέρθηκε, είναι διαθέσιμοι μέχρι $n \leqslant 10$, κάτι που δεν δημιουργεί πρόβλημα καθώς τόσο ο Kendall (1938) όσο και ο Mann (1945) απέδειξαν ότι η κανονική προσέγγιση είναι ικανοποιητική για $n \geqslant 10$. Ειδικότερα, η στατιστική συνάρτηση

$$Z_{\text{MK}} = \begin{cases} \dfrac{\text{MK}-1}{\sqrt{\text{Var}(\text{MK})}}, & \text{αν MK} > 0, \\ 0, & \text{αν MK} = 0, \\ \dfrac{\text{MK}+1}{\sqrt{\text{Var}(\text{MK})}}, & \text{αν MK} < 0, \end{cases} \tag{7.4}$$

όπου

$$\text{Var}(\text{MK}) = \frac{1}{18}\left(n(n-1)(2n+5) - \sum_{p=1}^{q} t_p(t_p-1)(2t_p+5) \right),$$

με q να είναι το πλήθος των ισοβαθμιών και t_p το πλήθος των τιμών που συμμετέχουν στην p-οστή ισοβαθμία, αποδεικνύεται ότι ακολουθεί ασυμπτωτικά τυπική κανονική κατανομή. Η απόδειξη του συγκεκριμένου αποτελέσματος ξεφεύγει από τους σκοπούς του παρόντος συγγράμματος και για τον λόγο αυτό παραλείπεται. Χρησιμοποιώντας την προσεγγιστική κατανομή, η υπόθεση της τυχαιότητας έναντι της εναλλακτικής της ύπαρξης μονότονης τάσης απορρίπτεται, αν $|Z_{\text{MK}}| \geqslant z_{a/2}$, ενώ απορρίπτεται έναντι της ύπαρξης αυξητικής (πτωτικής) τάσης, αν $Z_{\text{MK}} \geqslant z_a$ (αντίστοιχα αν $Z_{\text{MK}} \leqslant -z_a$).

Παρατήρηση 7.12. Στην περίπτωση μη ύπαρξης ισοβαθμιών, προφανώς, ισχύει ότι $\text{Var}(\text{MK}) = \frac{n(n-1)(2n+5)}{18}$.

Παράδειγμα 7.13. Στη διάθεσή μας έχουμε τις παρακάτω μετρήσεις που αφορούν τη μέση ημερήσια σχετική υγρασία (relative humidity) σε μια συγκεκριμένη αστική περιοχή, για 12 ημέρες. Χρησιμοποιήστε την ασυμπτωτική μορφή του ελέγχου των Mann-Kendall και ελέγξτε σε ε.σ. 1% την υπόθεση της τυχαιότητας για τις τιμές στο δείγμα, οι οποίες είναι:

$$5,\ 6.5,\ 5.75,\ 6.25,\ 7,\ 6.8,\ 7.1,\ 6.7,\ 7.5,\ 7.8,\ 8.3,\ 9.5.$$

Λύση Παραδείγματος 7.13. Η στατιστική συνάρτηση ελέγχου MK για τον έλεγχο των Mann-Kendall δίνεται από τη σχέση (7.2). Για τον υπολογισμό των επιμέρους αθροισμάτων (συνολικά 11 όροι) κατασκευάζεται ο Πίνακας 7.1. Πιο συγκεκριμένα, στον πίνακα δίνονται τα αποτελέσματα των πράξεων $sgn(X_j - X_i)$ για $j = i+1, \ldots, n$, όταν $i = 1, 2, \ldots, n$. Για παράδειγμα, αφού $x_4 = 6.25$ και $x_2 = 6.5$ η διαφορά $x_4 - x_2$ είναι αρνητική και το πρόσημο (τιμή -1) δίνεται στη θέση (4,2) του παραπάνω πίνακα (με έντονη γραμματοσειρά). Στην τελευταία γραμμή δίνονται τα αθροίσματα κατά στήλη και το άθροισμα των τιμών αυτών δίνει την τιμή της στατιστικής συνάρτησης MK. Άμεσα διαπιστώνουμε ότι $\text{MK} = 54$.

Εναλλακτικά, καθώς δεν υπάρχουν δεσμοί στα δεδομένα, ένας τρόπος υπολογισμού της τιμής της στατιστικής συνάρτησης MK είναι ο ακόλουθος. Για καθεμία εκ των 11 πρώτων παρατηρήσεων καταγράφουμε το πλήθος των αριθμών που βρίσκονται δεξιά τους και είναι μεγαλύτεροι από αυτές. Είναι τότε 11, 8, 9, 9, 5, 5, 4, 4, 3, 2, 1, αντίστοιχα με συνολικό άθροισμα 60. Γνωρίζοντας ότι το δυνατό συνολικό άθροισμα είναι $1 + 2 + \ldots + 11 =$

Πίνακας 7.1: Πίνακας υπολογισμών επιμέρους αθροισμάτων για τα δεδομένα του Παραδείγματος 7.13.

		1	2	3	4	5	6	7	8	9	10	11	12
		5	6.5	5.75	6.25	7	6.8	7.1	6.7	7.5	7.8	8.3	9.5
1	5	0											
2	6.5	1	0										
3	5.75	1	-1	0									
4	6.25	1	**-1**	1	0								
5	7	1	1	1	1	0							
6	6.8	1	1	1	1	-1	0						
7	7.1	1	1	1	1	1	1	0					
8	6.7	1	1	1	1	-1	-1	-1	0				
9	7.5	1	1	1	1	1	1	1	1	0			
10	7.8	1	1	1	1	1	1	1	1	1	0		
11	8.3	1	1	1	1	1	1	1	1	1	1	0	
12	9.5	1	1	1	1	1	1	1	1	1	1	1	0
	Άθροισμα	11	6	9	8	3	4	3	4	3	2	1	0

66 έχουμε ότι υπάρχουν και 6 αριθμοί μικρότεροι. Επομένως, $\mathrm{MK} = 60 - 6 = 54$ ή, διαφορετικά, από τη σχέση

$$\mathrm{MK} = 2k - \frac{n(n-1)}{2} = 2 \cdot 60 - \frac{12 \cdot 11}{2} = 120 - 66 = 54.$$

Αφού θα χρησιμοποιηθεί η ασυμπτωτική μορφή του ελέγχου, θα πρέπει να υπολογίσουμε τη διασπορά $\mathrm{Var}(\mathrm{MK})$, η οποία στην περίπτωση μη ύπαρξης ισοβαθμιών ισούται με

$$\mathrm{Var}(\mathrm{MK}) = \frac{n(n-1)(2n+5)}{18} = \frac{12 \cdot 11 \cdot (2 \cdot 12 + 5)}{18} \approx 212.6667.$$

Καθώς $\mathrm{MK} = 54 > 0$, η στατιστική συνάρτηση ελέγχου για το ασυμπτωτικό τεστ είναι $Z_{\mathrm{MK}} = \frac{\mathrm{MK}-1}{\sqrt{\mathrm{Var}(\mathrm{MK})}}$ και με αντικατάσταση έπεται ότι

$$z_{\mathrm{MK}} = \frac{54 - 1}{\sqrt{212.6667}} \approx 3.6343.$$

Σε ε.σ. 1%, το άνω 0.005 ποσοστιαίο σημείο της $\mathcal{N}(0,1)$ είναι 2.58 και η μορφή της κρίσιμης περιοχής είναι $K : |Z| \geqslant 2.58$. Αφού, $3.6343 > 2.58$ απορρίπτεται η H_0 και, άρα, δεν μπορούμε να υποθέσουμε ότι το δείγμα είναι τυχαίο. □

7.7 Bartels rank test

Έστω $X_1, \ldots, X_n$, οι $n > 2$ το πλήθος διαθέσιμες παρατηρήσεις, διατεταγμένες σε χρονολογική σειρά. Αν υποθέσουμε ότι τα δεδομένα αυτά προέρχονται από έναν πληθυσμό που ακολουθεί κανονική κατανομή, τότε ένας τρόπος ελέγχου της υπόθεσης της τυχαιότητας του δείγματος έναντι αποκλίσεων αυτοσυσχέτισης πρώτου βαθμού είναι αυτός που στηρίζεται στο πηλίκο (βλ. Neumann, 1941)

$$\frac{\sum_{i=1}^{n-1}(X_i - X_{i+1})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Όμως, η κατανομή αυτού του πηλίκου υπό την υπόθεση της τυχαιότητας του δείγματος είναι ευαίσθητη όταν τα δεδομένα δεν προέρχονται από κανονικό πληθυσμό (βλ. Bartels, 1982, και τις εκεί αναφορές). Αυτή η

παρατήρηση παρακίνησε τον Bartels (1982) να προτείνει έναν αντίστοιχο μη παραμετρικό έλεγχο, ο οποίος στηρίζεται στη στατιστική συνάρτηση που προκύπτει από το προηγούμενο πηλίκο όταν αντικαθίστανται οι τιμές X_i και X_{i+1} από τις αντίστοιχες τάξεις.

Σε αυτό το πλαίσιο, έστω ότι R_1,R_2,...,R_n είναι οι τάξεις των X_1, X_2,..., X_n αντίστοιχα. Τότε προτείνεται η ακόλουθη στατιστική συνάρτηση για τον έλεγχο της τυχαιότητας του δείγματος

$$\text{RVN} = \frac{\sum_{i=1}^{n-1}(R_i - R_{i+1})^2}{\sum_{i=1}^{n}(R_i - \bar{R})^2}. \tag{7.5}$$

Στην περίπτωση που δεν υπάρχουν δεσμοί στις παρατηρήσεις ισχύει ότι:

$$\bar{R} = \frac{\sum_{i=1}^{n} R_i}{n} = \frac{\sum_{i=1}^{n} i}{n} = \frac{n+1}{2},$$

και

$$\begin{aligned}\sum_{i=1}^{n}(R_i - \bar{R})^2 &= \sum_{i=1}^{n} R_i^2 - n\left(\frac{n+1}{2}\right)^2 \\ &= \sum_{i=1}^{n} i^2 - n\left(\frac{n+1}{2}\right)^2 = \frac{n(n+1)(n-1)}{12},\end{aligned}$$

καθώς $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$. Επομένως, υπό την υπόθεση μη ύπαρξης δεσμών, έχουμε ότι η αρχική στατιστική συνάρτηση RVN είναι ισοδύναμη[2] με τη στατιστική συνάρτηση NM, η οποία ορίζεται από τη σχέση:

$$\text{NM} = \sum_{i=1}^{n-1}(R_i - R_{i+1})^2. \tag{7.6}$$

Η στατιστική συνάρτηση NM έχει εύρος τιμών από $n-1$ έως $\frac{(n-1)(n^2+n-3)}{3}$, όταν n είναι άρτιος αριθμός ή έως $\frac{(n-1)(n^2+n-3)}{3} - 1$, όταν n είναι περιττός αριθμός (βλ. Bartels, 1982).

Καθώς οι δύο στατιστικές συναρτήσεις είναι ισοδύναμες, αρκεί να υπολογιστεί η κατανομή της μίας εκ των δύο υπό τη μηδενική υπόθεση. Ο προσδιορισμός τους βασίζεται σε όλες τις πιθανές μεταθέσεις των πρώτων n ακεραίων. Για μέγεθος δείγματος $4 \leqslant n \leqslant 10$ κρίσιμες τιμές έχουν δοθεί από τον Bartels (1982), ενώ για $10 \leqslant n \leqslant 17$ από τους Mateus and Caeiro (2018). Για μεγαλύτερα μεγέθη δείγματος ο υπολογισμός γίνεται ιδιαίτερα περίπλοκος, αφού αυξάνονται οι δυνατές μεταθέσεις των n ακεραίων. Στο ακόλουθο παράδειγμα απλά θα δώσουμε τον τρόπο σκέψης για την εύρεση της κατανομής της στατιστικής συνάρτησης ελέγχου NM υπό τη μηδενική υπόθεση, μέσω μιας ειδικής περίπτωσης.

Παράδειγμα 7.14. Για $n = 4$, να βρεθεί η κατανομή της στατιστικής συνάρτησης NM υπό τη μηδενική υπόθεση της τυχαιότητας του δείγματος.

Λύση Παραδείγματος 7.14. Έστω x_1, x_2, x_3, x_4 οι τιμές στο δείγμα και έστω, επίσης, ότι $x_1 < x_2 < x_3 < x_4$. Υπό την υπόθεση της τυχαιότητας υπάρχουν $4! = 24$ τρόποι διάταξης των παρατηρήσεων, που οδηγούν στις ακόλουθες τιμές των τάξεων:

$$1234,\ 1243,\ 1342,\ 1324,\ 1432,\ 1423,\ 2134,\ 2143,\ 2341,\ 2314,\ 2431,\ 2413$$
$$3124,\ 3142,\ 3241,\ 3214,\ 3412,\ 3421,\ 4123,\ 4132,\ 4213,\ 4231,\ 4312,\ 4321.$$

[2] Διαφέρουν κατά μία πολλαπλασιαστική σταθερά η οποία εξαρτάται από το μέγεθος δείγματος n. Ειδικότερα, ισχύει ότι: $\text{NM} = \frac{n(n+1)(n-1)}{12}\text{RVN}$.

Καθεμία από αυτές τις διατάξεις οδηγούν, μετά από λίγη άλγεβρα, στις ακόλουθες τιμές για τη στατιστική συνάρτηση NM:

$$3,\ 6,\ 9,\ 9,\ 11,\ 14,\ 6,\ 11,\ 11,\ 14,\ 9,\ 17,\ 9,\ 17,\ 14,\ 11,\ 11,\ 6,\ 11,\ 14,\ 9,\ 9,\ 6,\ 3,$$

αντίστοιχα. Για παράδειγμα, για την περίπτωση 1234 έχουμε ότι

$$\sum_{i=1}^{n-1}(R_i - R_{i+1})^2 = (1-2)^2 + (2-3)^2 + (3-4)^2 = 3.$$

Άμεσα, λοιπόν, διαπιστώνουμε ότι, στην ειδική περίπτωση που δεν έχουμε δεσμούς σε ένα δείγμα μεγέθους $n = 4$ παρατηρήσεων, η κατανομή της στατιστικής συνάρτησης NM υπό την H_0 είναι

k	$P(\mathsf{NM} = k)$
3	$P(\mathsf{NM} = 3) = 2/24$
6	$P(\mathsf{NM} = 6) = 4/24$
9	$P(\mathsf{NM} = 9) = 6/24$
11	$P(\mathsf{NM} = 11) = 6/24$
14	$P(\mathsf{NM} = 14) = 4/24$
17	$P(\mathsf{NM} = 17) = 2/24$

Άμεσα έπεται ότι $P(\mathsf{NM} \leqslant 3) = P(\mathsf{NM} \geqslant 17) = \frac{2}{24}$ και $P(\mathsf{NM} \leqslant 6) = P(\mathsf{NM} \geqslant 14) = \frac{6}{24}$. □

Καθώς το μέγεθος δείγματος n αυξάνεται, αυξάνεται και η πολυπλοκότητα στην εύρεση της ακριβούς κατανομής της στατιστικής συνάρτησης ελέγχου για το τεστ του Bartels. Για τον λόγο αυτό, πολλοί συγγραφείς έχουν ασχοληθεί με την εύρεση της προσεγγιστικής κατανομής για τη στατιστική συνάρτηση NM ή για τη στατιστική συνάρτηση RVN (βλ. Bartels, 1982; Mateus and Caeiro, 2018). Στο πλαίσιο αυτού του συγγράμματος θα παρουσιαστεί η κλασική κανονική προσέγγιση, που προτάθηκε από τον Bartels (1982). Ωστόσο, τόσο η Βήτα προσέγγιση που προτάθηκε από τον ίδιο, όσο και η κανονική προσέγγιση με τη βοήθεια αναπτύγματος Edgeworth, που προτάθηκε από τους Mateus and Caeiro (2018), είναι πιο ικανοποιητικές. Παρά ταύτα, ξεφεύγουν από τους σκοπούς του παρόντος συγγράμματος.

Ειδικότερα, αποδεικνύεται ότι υπό την υπόθεση της τυχαιότητας του δείγματος (και υπό την επιπλέον υπόθεση ότι δεν υπάρχουν δεσμοί) ότι η στατιστική συνάρτηση RVN ακολουθεί ασυμπτωτικά κανονική κατανομή με μέση τιμή 2 και διακύμανση η οποία είναι ίση με:

$$\sigma^2_{\mathsf{RVN}} = \frac{4(n-2)(5n^2 - 2n - 9)}{5n(n+1)(n-1)^2}.$$

Παρατήρηση 7.13. Η ύπαρξη δεσμών επηρεάζει την ακριβή μηδενική κατανομή τόσο της στατιστικής συνάρτησης RVN όσο και της NM και οι πίνακες ποσοστιαίων σημείων ή πιθανοτήτων που έχουν δοθεί στη βιβλιογραφία δεν ισχύουν, παρότι μπορεί να αποτελούν μια καλή προσέγγιση. Επιπλέον, ο Bartels (1982) επισημαίνει ότι αν σε κάθε δεσμό αποδοθεί ο μέσος όρος των τάξεων (midranks), τότε η ασυμπτωτική κατανομή της στατιστικής συνάρτησης RVN δεν επηρεάζεται, σε αντίθεση με την ασυμπτωτική κατανομή της NM. Σε μια τέτοια περίπτωση, λοιπόν, συστήνεται η χρήση της στατιστικής συνάρτησης RVN.

Παρατήρηση 7.14. Μελέτες προσομοίωσης (βλ. Gibbons, 2014, και τις εκεί αναφορές) έχουν δείξει ότι ο έλεγχος RVN είναι ισχυρότερος του ελέγχου του συνολικού αριθμού ανοδικών και καθοδικών ροών για εναλλακτικές που υποδηλώνουν σειριακή αυτοσυσχέτιση.

Παράδειγμα 7.15. Χρησιμοποιήστε τα δεδομένα του Παραδείγματος 7.13 (μετρήσεις σχετικής υγρασίας) και ελέγξτε την υπόθεση της τυχαιότητας του δείγματος, χρησιμοποιώντας αυτήν τη φορά την ασυμπτωτική μορφή του Bartels rank test, σε ε.σ. 1%.

Λύση Παραδείγματος 7.15. Η στατιστική συνάρτηση ελέγχου RVN του Bartels rank test δίνεται από τη σχέση

$$\text{RVN} = \frac{\sum_{i=1}^{n-1}(R_i - R_{i+1})^2}{\sum_{i=1}^{n}(R_i - \bar{R})^2}$$

όπου R_i είναι οι βαθμοί (τάξεις ή ranks) των παρατηρήσεων, ενώ $\bar{R}$ είναι ο μέσος όρος αυτών.

Από τον ορισμό της τάξης μιας παρατήρησης έχουμε τις παρακάτω τάξεις για τις παρατηρήσεις στο δείγμα (γραμμή R_i)

i	1	2	3	4	5	6	7	8	9	10	11	12
X_i	5	6.5	5.75	6.25	7	6.8	7.1	6.7	7.5	7.8	8.3	9.5
R_i	1	4	2	3	7	6	8	5	9	10	11	12

Δεν είναι δύσκολο να διαπιστώσουμε ότι

$$\bar{R} = (1 + 4 + ... + 11 + 12)/12 = 6.5 \text{ και } \sum_{i=1}^{n}(R_i - \bar{R})^2 = 143,$$

καθώς γενικά ισχύει στην περίπτωση μη ύπαρξης δεσμών ότι:

$$\sum_{i=1}^{n} R_i = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

και

$$\sum_{i=1}^{n}(R_i - \bar{R})^2 = \frac{n(n+1)(n-1)}{12}.$$

Επίσης,

$$\sum_{i=1}^{n-1}(R_i - R_{i+1})^2 = (1-4)^2 + (4-2)^2 + ... + (10-11)^2 + (11-12)^2 = 63.$$

Άρα, η τιμή της στατιστικής συνάρτησης $\text{RVN} = 63/143 \approx 0.4406$. Γνωρίζουμε, επίσης, ότι η στατιστική συνάρτηση ελέγχου για το ασυμπτωτικό τεστ είναι $Z = \frac{\text{RVN} - 2}{\sqrt{\sigma^2_{\text{RVM}}}}$, όπου

$$\sigma^2_{\text{RVM}} = \frac{4(n-2)(5n^2 - 2n - 9)}{5n(n+1)(n-1)^2}.$$

Με αντικατάσταση προκύπτει ότι:

$$\sigma^2_{\text{RVM}} = \frac{4 \cdot (12-2) \cdot (5 \cdot 12^2 - 2 \cdot 12 - 9)}{5 \cdot 12 \cdot (12+1) \cdot (12-1)^2} \approx 0.2911.$$

Επομένως, η τιμή της στατιστικής συνάρτησης Z είναι

$$z = \frac{0.4406 - 2}{\sqrt{0.2911}} \approx -2.89.$$

Σε ε.σ. 1%, το άνω 0.005 ποσοστιαίο σημείο της $\mathcal{N}(0,1)$ είναι 2.58 και η μορφή της κρίσιμης περιοχής είναι $K : |Z| \geqslant 2.58$. Αφού $|-2.89| = 2.89 > 2.58$, απορρίπτεται η H_0 και, άρα, δεν μπορούμε να υποθέσουμε ότι το δείγμα είναι τυχαίο. □

Παρατήρηση 7.15. Σε αυτό το κεφάλαιο παρουσιάστηκαν κάποιοι από τους μη παραμετρικούς ελέγχους που έχουν εισαχθεί στη βιβλιογραφία για τον έλεγχο της τυχαιότητας ενός δείγματος. Αν οι δειγματικές παρατηρήσεις είναι απλώς δύο σύμβολα, τότε, σύμφωνα με τους Gibbons and Chakraborti (2020), μεταξύ των ελέγχων, που παρουσιάστηκαν, προτείνεται να χρησιμοποιείται ο έλεγχος των ροών, ενώ, αν είναι αριθμητικά δεδομένα, προτείνεται ο έλεγχος είτε με τη στατιστική συνάρτηση RVN είτε με τον συνολικό αριθμό ανοδικών και καθοδικών ροών. Πληθώρα άλλων ελέγχων έχουν προταθεί στη βιβλιογραφία, με κάποιους από αυτούς να δίνουν ιδιαίτερη έμφαση στην περίπτωση χρονικών σειρών, δηλαδή δεδομένων που δίνονται σε χρονολογική σειρά. Για ελέγχους τυχαιότητας οι οποίοι βασίζονται σε στατιστικές συναρτήσεις του πλήθους των ροών σε μια ακολουθία δίτιμων δοκιμών, παραμέμπουμε τον/την ενδιαφερόμενο/μενη αναγνώστη/στρια στο σύγγραμμα των Balakrishnan and Koutras (2011).

7.8 Ασκήσεις

Άσκηση 7.1. Ρίχνουμε ένα νόμισμα, και τα αποτελέσματα της ρίψης φαίνονται στην παρακάτω ακολουθία, όπου Κ=κεφαλή και Γ=γράμματα.

Κ Κ Κ Γ Γ Γ Γ Κ Κ Γ Κ Γ Γ Γ

Ελέγξτε με το τεστ των ροών και σε επίπεδο σημαντικότητας 5% αν η σειρά εμφάνισης των δύο δυνατών αποτελεσμάτων γίνεται με τυχαίο τρόπο. Ποιο είναι το ακριβές επίπεδο σημαντικότητας του ελέγχου;

Άσκηση 7.2. Σε ένα μικροβιολογικό εργαστήριο, ελέγχονται διαδοχικά δείγματα αίματος με σκοπό την ανίχνευση ενός συγκεκριμένου ιού. Τα αποτελέσματα είναι $+$, αν το τεστ είναι θετικό (παρουσία ιού), ή $-$, αν το τεστ είναι αρνητικό (απουσία ιού). Τα αποτελέσματα στη διάρκεια μίας εβδομάδας είναι τα εξής:

$$+ + - - - - + + - - - - - + + + - - + - - + + + + - + - - - + - + + ++$$

Χρησιμοποιήστε το τεστ των ροών (ασυμπτωτική μορφή ελέγχου) και ελέγξτε την υπόθεση ότι τα αποτελέσματα των εξετάσεων δεν δείχνουν αυξημένο ή μειωμένο ιικό φορτίο στην περιοχή, δηλαδή ότι το δείγμα είναι τυχαίο. Να χρησιμοποιήσετε ε.σ. 1% και ο έλεγχος να γίνει με χρήση κατάλληλων κρίσιμων σημείων. Τέλος, να υπολογιστεί η p-τιμή του ελέγχου.

Άσκηση 7.3. Στη διάθεσή μας έχουμε δεδομένα 100 ημερών, στα οποία έχουμε καταγράψει το ύψος της βροχόπτωσης. Αν η βροχόπτωση ξεπερνά τα 25.4 χιλιοστά, τότε η ημέρα χαρακτηρίζεται ως υγρή (W = wet), διαφορετικά χαρακτηρίζεται ως ξηρή (D = dry). Να ελέγξετε την υπόθεση ότι οι ενδείξεις D,W εμφανίζονται με τυχαίο τρόπο χρησιμοποιώντας την ασυμπτωτική μορφή του τεστ των ροών. Ο έλεγχος να γίνει σε ε.σ. 5%. Να υπολογιστεί η p-τιμή του ελέγχου.

Ημέρα	Αποτέλεσμα	Ημέρα	Αποτέλεσμα	Ημέρα	Αποτέλεσμα	Ημέρα	Αποτέλεσμα	Ημέρα	Αποτέλεσμα
1	D	21	W	41	W	61	D	81	W
2	D	22	W	42	W	62	D	82	D
3	D	23	W	43	W	63	D	83	W
4	D	24	D	44	W	64	D	84	D
5	D	25	D	45	W	65	W	85	D
6	W	26	D	46	D	66	W	86	W
7	W	27	D	47	W	67	D	87	W
8	W	28	D	48	W	68	W	88	D
9	D	29	D	49	D	69	W	89	W
10	D	30	D	50	D	70	D	90	D
11	D	31	D	51	D	71	W	91	W
12	D	32	D	52	D	72	D	92	W
13	D	33	W	53	D	73	D	93	W
14	D	34	D	54	W	74	D	94	W
15	W	35	W	55	W	75	D	95	W
16	W	36	W	56	D	76	D	96	W
17	D	37	D	57	D	77	D	97	W
18	D	38	D	58	D	78	D	98	W
19	D	39	W	59	D	79	W	99	D
20	D	40	W	60	D	80	W	100	W

Άσκηση 7.4. Μια ομάδα μπάσκετ έχει τα παρακάτω αποτελέσματα για μια σειρά 12 αγώνων

Ν Ν Ν Η Η Ν Ν Η Η Η Η Ν

όπου Ν σημαίνει νίκη και Η σημαίνει ήττα. Υπάρχει κάποια ένδειξη τυχαιότητας στη σειρά των αποτελεσμάτων αυτών των αγώνων; Να χρησιμοποιηθεί το τεστ των ροών και ο έλεγχος να γίνει με χρήση κατάλληλης κρίσιμης περιοχής. Να χρησιμοποιηθεί ε.σ. 5%. Ποιο είναι το ακριβές ε.σ. του παραπάνω ελέγχου;

Άσκηση 7.5. Έστω ακολουθία δοκιμών Bernoulli στην οποία παρατηρήσαμε 4 αποτελέσματα $+$ και 3 αποτελέσματα $-$. Να βρεθεί, υπό την υπόθεση της τυχαιότητας, η δεσμευμένη κατανομή της τυχαίας μεταβλητής $L_7 < \ell | S = 4$, όπου L_7 είναι η τυχαία μεταβλητή που παριστάνει το μήκος μέγιστης ροής επιτυχιών στην ακολουθία επτά ανεξάρτητων δοκιμών Bernoulli και S ο συνολικός αριθμός επιτυχιών στις επτά δοκιμές.

Άσκηση 7.6. Χρησιμοποιήστε τα δεδομένα του Παραδείγματος 7.10 (ημερήσια κρούσματα ενός κορονοϊού) και εφαρμόστε τον έλεγχο που βασίζεται στο πλήθος των ανοδικών/καθοδικών ροών για να ελέγξετε την υπόθεση ότι οι διαθέσιμες μετρήσεις αποτελούν τυχαίο δείγμα. Να χρησιμοποιήσετε την ασυμπτωτική μορφή του τεστ και ο έλεγχος να γίνει σε ε.σ. 5%.

Άσκηση 7.7. Οι αφίξεις ξένων τουριστών στην Ελλάδα από το 2000 μέχρι το 2010 δίνονται στον παρακάτω πίνακα (σε χιλιάδες).

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
13095	14057	14179	13969	13313	14765	16039	16153	15939	14915	15006

Πηγή: UNWTO, World Tourism Barometer, Vol. 9, August 2011.

Χρησιμοποιώντας τον έλεγχο των Moore και Wallis και ε.σ. $a = 0.01$, να ελέγξετε την υπόθεση ότι οι αφίξεις τουριστών στην Ελλάδα είναι τυχαίες. Ο έλεγχος να γίνει με χρήση του ασυμπτωτικού ελέγχου, στον οποίο να εφαρμοστεί και η διόρθωση συνεχείας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ξενόγλωσση

Balakrishnan, N. and Koutras, M. V. (2011). *Runs and scans with applications*. Vol. 764. John Wiley & Sons.

Bartels, R. (1982). The Rank version of von Neumann's Ratio Test for Randomness. *Journal of the American Statistical Association*, pp. 40–46.

Bateman, G. (1948). On the power function of the longest run as a test for randomness in a sequence of alternatives. *Biometrika*, *35*, pp. 97–112.

David, F. (1947). A power function for tests of randomness in as sequence of alternatives. *Biometrika*, *34*, pp. 335–339.

Gibbons, J. D. and Chakraborti, S. (2020). *Nonparametric Statistical Inference, Fourth Edition Revised and Expanded*. Chapman and Hall/CRC.

Gibbons, J. D. (2014). Tests of Randomization. In: *Wiley StatsRef: Statistics Reference Online*. American Cancer Society.

Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika*, *30*, pp. 81–93.

Kendall, M. and Ord, J. K. (1973). *Time Series*. Oxford University Press.

Kokoska, S. and Nevison, C. (1989). Critical Values For The Runs Test. In: *In: Statistical Tables and Formulas*. New York, NY: Springer Texts in Statistics, pp. 81–82.

Levene, H. (1952). On the power function of tests of randomness based on runs up and down. *Annals of Mathematical Statistics*, *23*(1), pp. 34–56.

Mann, H. B. (1945). Nonparametric Tests Against Trend. *Econometrica*, *13*, pp. 245–259.

Mateus, A. and Caeiro, F. (2018). Exact and Approximate Probabilities for the Null Distribution of Bartels Randomness Test. In: *Recent Studies on Risk Analysis and Statistical Modeling*. Ed. by T. A. Oliveira, C. P. Kitsos, A. Oliveira and L. Grilo. Springer International Publishing, pp. 227–240.

Mogull, R. G. (1994). The one-sample runs test: A category of exception. *J. Exper. Behav. Statist.*, *19*, pp. 296–303.

Moore, G. H. and Wallis, W. A. (1943). Time Series Significance Tests Based on Signs of Differences. *Journal of the American Statistical Association*, *38*, pp. 153–164.

Mosteller, F. (1941). Note on an application of runs to quality control charts. *Annals of Mathematical Statistics*, *12*, pp. 228–232.

Neumann, J. v. (1941). Distribution of the ratio of the mean square successive difference to the variance. *Ann. Math. Statistics*, *12*, pp. 367–395.

Sheskin, D. (2011). *Handbook of Parametric and Non-parametric Procedures* (5th ed.). Chapman and Hall/CRC.

Stuart, A. (1954). Asymptotic relative efficiencies of distribution-free tests of randomness against normal alternatives. *Journal of the American Statistical Association*, *49*(265), pp. 147–157.

Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Ann. Math. Statistics*, *11*, pp. 147–162.

Wallis, W. A. and Moore, G. H. (1941). A significance test for time series analysis. *Journal of the American Statistical Association*, *36*, pp. 401–409.

Wallis, W. A. and Roberts, H. V. (1956). *Statistics: A new approach*. Glencoe, IL: Free Press.

ΚΕΦΑΛΑΙΟ 8

ΜΕΤΡΑ ΚΑΙ ΕΛΕΓΧΟΙ ΣΥΣΧΕΤΙΣΗΣ ΔΥΟ ΜΕΤΑΒΛΗΤΩΝ

Σύνοψη

Σκοπός του κεφαλαίου αυτού είναι η παράθεση μεθοδολογιών για τη διερεύνηση της ύπαρξης ή μη κάποιου είδους σχέσης (συνήθως γραμμικής) μεταξύ δύο μεταβλητών. Πιο συγκεκριμένα, αρχικά το ενδιαφέρον επικεντρώνεται στη μελέτη της γραμμικής εξάρτησης μεταξύ δύο ποσοτικών τυχαίων μεταβλητών X και Y, μέσω του συντελεστή συσχέτισης του Pearson, που ουσιαστικά αποτελεί μία παραμετρική μέθοδο. Στη συνέχεια, μελετώνται οι συντελεστές συσχέτισης των Spearman και Kendall, για την περίπτωση ύπαρξης μονότονης σχέσης ανάμεσα στις μεταβλητές X και Y. Τέλος, μελετώνται και συναφή μέτρα συνάφειας, όπως ο συντελεστής γ των Goodman and Kruskal και ο συντελεστής Q του Yule.

Προαπαιτούμενη γνώση:

Βασικές γνώσεις Πιθανοτήτων και Στατιστικής.

Προσδοκώμενα μαθησιακά αποτελέσματα:

Ο/η φοιτητής/τρια ολοκληρώνοντας την ενότητα αυτή θα μπορεί να ελέγχει την ύπαρξη ή μη ύπαρξη συσχέτισης μεταξύ δύο μεταβλητών είτε τα δεδομένα είναι ομαδοποιημένα σε πίνακες συνάφειας είτε είναι μη ομαδοποιημένα. Επίσης, ο έλεγχος ανεξαρτησίας σε πίνακες συνάφειας μπορεί να οδηγήσει τον/τη φοιτητή/τρια στο να αποφανθεί αν οι δύο ποιοτικές μεταβλητές είναι ασυσχέτιστες.

Γλωσσάριο επιστημονικών όρων

- Έλεγχος ανεξαρτησίας σε πίνακες συνάφειας
- Διόρθωση Yates
- Μέτρα συνάφειας
- Συντελεστής συσχέτισης
- Συντελεστής συσχέτισης του Pearson
- Συντελεστής συσχέτισης του Spearman
- Συντελεστής συσχέτισης του Kendall
- Συντελεστής συσχέτισης των Goodman-Kruskal
- Συντελεστής συσχέτισης του Yule

Μπατσίδης, Α., Παπασταμούλης, Π., Πετρόπουλος, Κ., & Ρακιτζής, Α. (2022). *Μη Παραμετρική Στατιστική.* [Προπτυχιακό εγχειρίδιο].

8.1 Εισαγωγή

Σκοπός του κεφαλαίου αυτού είναι η παράθεση μεθοδολογιών για τη διερεύνηση της ύπαρξης ή μη γραμμικής σχέσης μεταξύ δύο μεταβλητών. Ειδικότερα, αρχικά το ενδιαφέρον επικεντρώνεται στη μελέτη της γραμμικής εξάρτησης μεταξύ δύο ποσοτικών τυχαίων μεταβλητών X και Y, μέσω του συντελεστή συσχέτισης του Pearson. Στον υπολογισμό του συντελεστή συσχέτισης του Pearson χρησιμοποιείται τόσο η μέση τιμή όσο και η τυπική απόκλιση, ενώ ο έλεγχος της υπόθεσης μη ύπαρξης γραμμικής σχέσης που βασίζεται σε αυτόν υπονοεί κανονικότητα στα δεδομένα. Τα παραπάνω οδηγούν στο συμπέρασμα ότι πρόκειται ουσιαστικά για μια παραμετρική μέθοδο. Για τον λόγο αυτόν στη συνέχεια του κεφαλαίου το ενδιαφέρον εστιάζεται στη μελέτη και παρουσίαση μη παραμετρικών εκδοχών του συντελεστή συσχέτισης, όπως είναι αυτές των Spearman και Kendall (για λεπτομέρειες, βλ. Puth *et al.*, 2015), ενώ ιδιαίτερη αναφορά θα γίνει στον συντελεστή συσχέτισης γ των Goodman και Kruskal, όπως και στην ειδική περίπτωση του συντελεστή Q του Yule.

8.2 Συντελεστής συσχέτισης του Pearson

Έστω οι συνεχείς τυχαίες μεταβλητές X και Y. Το ενδιαφέρον μας επικεντρώνεται στη διερεύνηση της ύπαρξης ή μη γραμμικής σχέσης μεταξύ αυτών. Για παράδειγμα, ενδιαφερόμαστε για την ύπαρξη ή όχι γραμμικής σχέσης μεταξύ του ύψους και του βάρους ενήλικων ατόμων, των εξόδων και των εσόδων μιας εταιρείας κ.ο.κ. Ένα μέτρο συσχέτισης είναι ο πληθυσμιακός ή θεωρητικός συντελεστής συσχέτισης, ο οποίος συμβολίζεται με ρ ή $\rho(X,Y)$ και ο ορισμός του, παρότι δόθηκε στο Κεφάλαιο 1 του παρόντος συγγράμματος, παρατίθεται στη συνέχεια για λόγους πληρότητας.

Ορισμός 8.1: Συντελεστής Συσχέτισης

Ο συντελεστής συσχέτισης ρ δύο από κοινού κατανεμημένων τ.μ. (X,Y) ορίζεται από τη σχέση:

$$\rho(X,Y) := \rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}},$$

όπου $\mathrm{Var}(X)$ και $\mathrm{Var}(Y)$ είναι η πληθυσμιακή διακύμανση των τ.μ. X και Y, αντίστοιχα, ενώ $\mathrm{Cov}(X,Y)$ είναι η συνδιακύμανση των X και Y.

Στον παραπάνω ορισμό $\mathrm{Cov}(X,Y)$ είναι η συνδιακύμανση των X και Y, η οποία δίνεται από τη σχέση:

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))] = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y),$$

και αποτελεί ένα μέτρο της συμμεταβλητότητας των X και Y, μη απαλλαγμένο από τις μονάδες μέτρησης των μεταβλητών. Σε αντίθεση με τη συνδιακύμανση, ο συντελεστής συσχέτισης $\rho(X,Y)$ είναι απαλλαγμένος από μονάδες, δηλαδή πρόκειται για έναν καθαρό αριθμό (βλ., μεταξύ άλλων Κούτρας, 2018). Άλλες βασικές ιδιότητες του συντελεστή συσχέτισης είναι οι ακόλουθες.

- Ο συντελεστής συσχέτισης λαμβάνει τιμές στο διάστημα $[-1,1]$.
- Αν οι τ.μ. X,Y είναι ανεξάρτητες, τότε $\rho = 0$.
- Η απόλυτη τιμή του συντελεστή συσχέτισης ισούται με 1, δηλαδή $|\rho| = 1$, αν και μόνο αν οι τ.μ. X,Y συνδέονται με τέλεια γραμμική σχέση, δηλαδή αν και μόνο αν $Y = a \pm bX$, με $a,b \in \mathbb{R}$.
- Τιμές του συντελεστή συσχέτισης κοντά στο 1 (αντίστοιχα στο -1) υποδεικνύουν την ύπαρξη γραμμικής θετικής (αντίστοιχα αρνητικής) σχέσης. Αυτό σημαίνει ότι μεγάλες τιμές της μιας μεταβλητής αντιστοιχούν σε μεγάλες (αντίστοιχα σε μικρές) τιμές της άλλης.
- Αν $\rho(X,Y) = 0$, τότε οι τ.μ. X,Y λέγονται (γραμμικά) ασυσχέτιστες. Θα πρέπει να σημειωθεί πως δύο

ασυσχέτιστες τυχαίες μεταβλητές δεν είναι απαραίτητα ανεξάρτητες, απλώς ο βαθμός της γραμμικής σχέσης τους είναι 0. Όμως, αν δύο τυχαίες μεταβλητές είναι ανεξάρτητες, τότε είναι και ασυσχέτιστες.

Στην πράξη ο θεωρητικός συντελεστής συσχέτισης είναι δύσκολο να υπολογιστεί. Για τον λόγο αυτόν, εκτιμάται από τον λεγόμενο (δειγματικό) συντελεστή συσχέτισης του Pearson, καθώς οφείλει το όνομά του στον Άγγλο μαθηματικό Karl Pearson (1857-1936), από τον οποίο πρωτοπαρουσιάστηκε στην εργασία του Pearson (1896).

Ορισμός 8.2

Έστω $(X_1, Y_1), \ldots, (X_n, Y_n)$ ένα τυχαίο δείγμα από n ζεύγη παρατηρήσεων για τις τυχαίες μεταβλητές X, Y. Ο συντελεστής συσχέτισης του Pearson συμβολίζεται με $r(X,Y)$ ή r και ορίζεται από τη σχέση:

$$r(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}},$$

ή, ισοδύναμα, από τη σχέση:

$$r(X,Y) = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sqrt{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}\sqrt{\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}}}.$$

Από τον παραπάνω ορισμό άμεσα προκύπτει ότι $r(X,Y) = r(Y,X)$. Επιπρόσθετα, ο δειγματικός συντελεστής συσχέτισης του Pearson λαμβάνει τιμές στο διάστημα $[-1,1]$ με ερμηνεία αντίστοιχη αυτής που δόθηκε για τις τιμές του πληθυσμιακού συντελεστή συσχέτισης.

Παρατήρηση 8.1. Αξίζει να αναφέρουμε ότι ο δειγματικός συντελεστής συσχέτισης του Pearson ορίζεται για ποσοτικές μεταβλητές μόνο και επηρεάζεται από την ύπαρξη ακραίων τιμών στο τυχαίο δείγμα των n το πλήθος παρατηρήσεων $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Έχοντας ορίσει το δειγματικό ανάλογο του πληθυσμιακού συντελεστή συσχέτισης ρ, το ενδιαφέρον στη συνέχεια επικεντρώνεται στο να ελέγξουμε αν δύο ποσοτικές τυχαίες μεταβλητές δεν συνδέονται γραμμικά. Για τον σκοπό αυτόν προβαίνουμε στον στατιστικό έλεγχο μίας εκ των υποθέσεων:

(Α) $H_0 : \rho = 0$ έναντι της $H_1 : \rho \neq 0$, δηλαδή της ύπαρξης είτε γραμμικής θετικής συσχέτισης είτε γραμμικής αρνητικής συσχέτισης.
(Β) $H_0 : \rho = 0$ έναντι της $H_1 : \rho > 0$, δηλαδή της ύπαρξης γραμμικής θετικής συσχέτισης.
(Γ) $H_0 : \rho = 0$ έναντι της $H_1 : \rho < 0$, δηλαδή της ύπαρξης γραμμικής αρνητικής συσχέτισης.

Ο έλεγχος αυτός βασίζεται στο δειγματικό ανάλογο της ποσότητας ρ και στο αποτέλεσμα της πρότασης που ακολουθεί.

Πρόταση 8.1. Έστω $(X_1, Y_1), \ldots, (X_n, Y_n)$ είναι ένα τυχαίο δείγμα από n ζεύγη παρατηρήσεων για τις τυχαίες μεταβλητές X, Y. Επιπρόσθετα, υποθέτουμε ότι το τυχαίο αυτό δείγμα προέρχεται από έναν πληθυσμό που περιγράφεται ικανοποιητικά από τη διδιάστατη κανονική κατανομή. Τότε, υπό τη μηδενική υπόθεση $H_0 : \rho = 0$, ισχύει ότι:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \overset{H_0}{\sim} t_{n-2}.$$

Απόδειξη Πρότασης 8.1. Στο πλαίσιο αυτού του συγγράμματος θα δοθεί μια σκιαγράφηση της απόδειξης, καθώς αυτή γίνεται εύκολα ανακαλώντας αποτελέσματα της Θεωρίας Γραμμικών Μοντέλων και, ειδικότερα,

της απλής γραμμικής παλινδρόμησης. Είναι γνωστό ότι η στατιστική συνάρτηση

$$T = \frac{R^2}{\left(1 - R^2\right)/(n-2)},$$

όπου R^2 ο συντελεστής προσδιορισμού της γραμμικής παλινδρόμησης $Y = a + bX + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, ακολουθεί, υπό την υπόθεση της μη ύπαρξης γραμμικής σχέσης μεταξύ των δύο τυχαίων μεταβλητών X και Y, την κατανομή $F_{1,n-2}$ (βλ. Trosset, 2009). Επιπλέον, ισχύει ότι $R^2 = r^2$ (βλ. Trosset, 2009). Το επιθυμητό αποτέλεσμα προκύπτει συνδυάζοντας τα παραπάνω και λαμβάνοντας επιπλέον υπόψη ότι εξ ορισμού $F_{1,n-2} = t^2_{n-2}$. Για μια λεπτομερή απόδειξη της πρότασης παραπέμπουμε στο σύγγραμμα των Hogg *et al.* (2013). □

Χρησιμοποιώντας το αποτέλεσμα της Πρότασης 8.1 προκύπτει, σε επίπεδο σημαντικότητας a, ότι για το πρόβλημα (Α) $H_0 : \rho = 0, H_1 : \rho \neq 0$ απορρίπτεται η μηδενική υπόθεση, αν $|t| \geq t_{n-2,a/2}$, για το (Β) $H_0 : \rho = 0$, $H_1 : \rho > 0$ απορρίπτεται η μηδενική υπόθεση αν $t \geq t_{n-2,a}$, ενώ για το (Γ) $H_0 : \rho = 0, H_1 : \rho < 0$ απορρίπτεται η μηδενική υπόθεση, αν $t \leq -t_{n-2,a}$.

Παρατήρηση 8.2. Σε κάποιες περιπτώσεις μπορεί κάποιος ερευνητής να ενδιαφέρεται να ελέγξει αν ο πληθυσμιακός συντελεστής συσχέτισης ρ λαμβάνει μια συγκεκριμένη (ορισμένη) τιμή, έστω $\rho_0 \in [-1,1]$, με $\rho_0 \neq 0$. Δηλαδή σε αυτές τις ειδικές περιπτώσεις ενδιαφερόμαστε για τον έλεγχο της μηδενικής υπόθεσης $H_0 : \rho = \rho_0$ έναντι μίας εκ των τριών εναλλακτικών: (Α) $H_1 : \rho \neq \rho_0$, (Β) $H_1 : \rho > \rho_0$ και (Γ) $H_1 : \rho < \rho_0$. Τότε χρησιμοποιείται ένας ασυμπτωτικός έλεγχος, ο οποίος βασίζεται στον μετασχηματισμό του Fisher (1921). Ειδικότερα, υποθέτοντας ότι το τυχαίο διάνυσμα (X, Y) ακολουθεί τη διδιάστατη κανονική κατανομή, για μεγάλες τιμές του δειγματικού μεγέθους n, η r έχει, προσεγγιστικά, κανονική κατανομή $\mathcal{N}(\rho, (1-\rho^2)^2/n)$ (Lehmann, 2004) ή, διαφορετικά,

$$\sqrt{n}(r - \rho) \xrightarrow{d} \mathcal{N}(0, (1-\rho^2)^2). \tag{8.1}$$

Έπειτα, από την (8.1), χρησιμοποιώντας τον μετασχηματισμό του Fisher που δίνεται από τη σχέση:

$$W = \frac{1}{2} \ln \frac{1+\rho}{1-\rho},$$

και τη Μέθοδο Δέλτα (delta method, βλ. Ενότητα 1.7), προκύπτει ότι

$$\sqrt{n}(\hat{W} - W) = \sqrt{n}\left(\frac{1}{2}\ln\frac{1+r}{1-r} - \frac{1}{2}\ln\frac{1+\rho}{1-\rho}\right) \xrightarrow{d} \mathcal{N}(0,1). \tag{8.2}$$

Σύμφωνα με τον Schulze (2004) (βλ., επίσης Welz *et al.*, 2021) για να πετύχουμε καλύτερη προσαρμογή της στατιστικής συνάρτησης της παραπάνω σχέσης στην τυπική κανονική κατανομή, προτείνεται η αντικατάσταση της σταθεράς $\sqrt{n}$, από τη $\sqrt{n-3}$ χωρίς να αλλάζει η ασυμπτωτική κατανομή. Επομένως, προκύπτει ότι, υπό την $H_0 : \rho = \rho_0$,

$$Z = \sqrt{n-3}\left(\frac{1}{2}\ln\frac{1+r}{1-r} - \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}\right) \xrightarrow{d} \mathcal{N}(0,1).$$

Άμεσα, από το προηγούμενο αποτέλεσμα έχουμε, σε επίπεδο σημαντικότητας a, ότι για το πρόβλημα (Α) $H_0 : \rho = \rho_0, H_1 : \rho \neq \rho_0$ απορρίπτεται η μηδενική υπόθεση, αν $|Z| \geq z_{a/2}$, για το (Β) $H_0 : \rho = \rho_0, H_1 : \rho > \rho_0$ απορρίπτεται η μηδενική υπόθεση, αν $Z \geq z_a$, ενώ για το (Γ) $H_0 : \rho = \rho_0, H_1 : \rho_0 < \rho_0$ απορρίπτεται η μηδενική υπόθεση, αν $Z \leq -z_a$.

Επισημαίνουμε ότι η ασυμπτωτική προσέγγιση είναι ικανοποιητική για μέγεθος δείγματος μεγαλύτερο από 50 και μπορεί να χρησιμοποιηθεί και για την κατασκευή διαστήματος εμπιστοσύνης για τον θεωρητικό συντελεστή συσχέτισης (για περισσότερες λεπτομέρειες, βλ. Welz *et al.*, 2021).

Παράδειγμα 8.1. Έστω ότι έχουμε τις ακόλουθες 7 το πλήθος παρατηρήσεις για τις μεταβλητές X και Y που παριστάνουν αντίστοιχα τον αριθμό προϊόντων και το αντίστοιχο κόστος αγοράς (σε ευρώ). Χρησιμοποιώντας τα δεδομένα του πίνακα που ακολουθεί να υπολογιστεί ο συντελεστής συσχέτισης του Pearson και κάνοντας τις κατάλληλες υποθέσεις να ελέγξετε, σε επίπεδο σημαντικότητας 5%, αν συσχετίζονται οι δύο τυχαίες μεταβλητές.

X	6	7	4	8	9	3	5
Y	80	83	75	86	95	72	69

Λύση Παραδείγματος 8.1. Αρχικά υπολογίζουμε τον συντελεστή συσχέτισης του Pearson, χρησιμοποιώντας τη σχέση:

$$r(X,Y) = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}.$$

Από τα διαθέσιμα δεδομένα έπεται άμεσα ότι: $\sum_{i=1}^{7} X_i = 42$ και, επομένως, $\bar{X} = 6$, ενώ $\sum_{i=1}^{7} Y_i = 560$ και $\bar{Y} = 80$. Επιπρόσθετα, ύστερα από αλγεβρικές πράξεις, έχουμε ότι $\sum_{i=1}^{7} X_i Y_i = 3465$, $\sum_{i=1}^{7}(X_i - \bar{X})^2 = 28$ και $\sum_{i=1}^{7}(Y_i - \bar{Y})^2 = 480$. Επομένως, με αντικατάσταση,

$$r(X,Y) = \frac{3465 - \frac{42 \cdot 560}{7}}{\sqrt{28}\sqrt{480}} = \frac{105}{115,93} = 0.906.$$

Άρα ο συντελεστής συσχέτισης του Pearson είναι θετικός (αναμενόμενο από τη φύση των δύο τ.μ.) και κοντά στη μονάδα. Με την προϋπόθεση ότι το διδιάστατο διάνυσμα (X,Y) ακολουθεί διδιάστατη κανονική κατανομή για τον έλεγχο της μηδενικής υπόθεσης $H_0 : \rho = 0$, έναντι της $H_1 : \rho \neq 0$ χρησιμοποιούμε τη στατιστική συνάρτηση:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \overset{H_0}{\sim} t_{n-2},$$

και κρίσιμη περιοχή, με επίπεδο σημαντικότητας $a = 0.05$, $|t| \geq t_{n-2,a/2} = t_{5,0.025} = 2.571$.

Η τιμή της στατιστικής συνάρτησης είναι:

$$t = \frac{0.906\sqrt{7-2}}{\sqrt{1-0.906^2}} = \frac{0.906 \cdot 2.236}{0.4233} = \frac{2.025816}{0.4233} = 4.786,$$

οπότε, καθώς 4.786 >2.571, συμπεραίνουμε ότι απορρίπτεται η μηδενική υπόθεση. Άρα υπάρχει στατιστικά σημαντική γραμμική συσχέτιση μεταξύ του αριθμού των προϊόντων και του κόστους αγοράς και, μάλιστα, θετική. □

Όπως έχει ήδη αναφερθεί, ο συντελεστής συσχέτισης του Pearson προϋποθέτει δύο ποσοτικές μεταβλητές, μη ύπαρξη ακραίων τιμών και, επιπλέον, για τη διενέργεια του στατιστικού ελέγχου υποθέσεων υποθέτουμε ότι το τυχαίο διάνυσμα (X,Y) ακολουθεί διδιάστατη κανονική κατανομή. Για κάθε άλλη περίπτωση, κρίνεται αναγκαία η χρήση των μέτρων συσχέτισης των επόμενων ενοτήτων.

8.3 Συντελεστής συσχέτισης του Spearman

Έστω $(X_1,Y_1),...,(X_n,Y_n)$, ένα τυχαίο δείγμα από n ζεύγη παρατηρήσεων για τις τυχαίες μεταβλητές X και Y. Επιπλέον, έστω $R(X_1) = R_1$, ..., $R(X_n) = R_n$ και $S(X_1) = S_1$, ..., $S(X_n) = S_n$ οι τάξεις των δειγματικών τιμών των X και Y, αντίστοιχα. Επισημαίνουμε ότι σε περίπτωση ύπαρξης δεσμών αντιστοιχούμε σε καθεμία από

τις ίσες αυτές τιμές τον μέσο όρο των τάξεων που θα είχαν αν δεν ταυτίζονταν (midranks). Τότε, ο συντελεστής συσχέτισης του Spearman, που πήρε το όνομά του από τον Άγγλο ψυχολόγο Charles Spearman (1863-1945), καθώς πρωτοπαρουσιάστηκε από αυτόν στην εργασία του Spearman (1904), συμβολίζεται με r_s και δεν είναι τίποτε άλλο παρά ο συντελεστής συσχέτισης του Pearson όταν αυτός εφαρμόζεται στις τάξεις $R_1,...,R_n$, και $S_1,...,S_n$. Ειδικότερα, έχουμε τον ακόλουθο ορισμό.

Ορισμός 8.3

Έστω $(X_1,Y_1),...,(X_n,Y_n)$ ένα τυχαίο δείγμα από n ζεύγη παρατηρήσεων για τις τυχαίες μεταβλητές X, Y. Επιπλέον, έστω $R_1,...,R_n$, και $S_1,...,S_n$, οι τάξεις των δειγματικών τιμών των X και Y, αντίστοιχα. Ο συντελεστής συσχέτισης του Spearman συμβολίζεται με r_s και ορίζεται από τη σχέση:

$$r_s = \frac{\sum_{i=1}^{n}(R_i-\overline{R})(S_i-\overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i-\overline{R})^2}\sqrt{\sum_{i=1}^{n}(S_i-\overline{S})^2}},$$

όπου $\overline{R}=\frac{\sum_{i=1}^{n}R_i}{n}$ και $\overline{S}=\frac{\sum_{i=1}^{n}S_i}{n}$.

Όπως και ο συντελεστής συσχέτισης του Pearson, ο αντίστοιχος του Spearman λαμβάνει τιμές στο διάστημα [-1,1] και ερμηνεύεται κατά τον ίδιο τρόπο.

Στην πρόταση που ακολουθεί, δίνονται ισοδύναμες εκφράσεις για τον συντελεστή συσχέτισης του Spearman στην περίπτωση της μη ύπαρξης δεσμών.

Πρόταση 8.2. Έστω $R_1,...,R_n$ και $S_1,...,S_n$ οι τάξεις των δειγματικών τιμών των X και Y, αντίστοιχα. Υποθέτοντας ότι δεν υπάρχουν δεσμοί, προκύπτουν οι ακόλουθες ισοδύναμες εκφράσεις για τον συντελεστή συσχέτισης του Spearman:

α) $r_s=\frac{12\sum_{i=1}^{n}\left(R_i-\frac{n+1}{2}\right)\left(S_i-\frac{n+1}{2}\right)}{n(n^2-1)}$

β) $r_s=1-\frac{6\sum_{i=1}^{n}(R_i-S_i)^2}{n(n^2-1)}$.

Απόδειξη Πρότασης 8.2. α) Στην περίπτωση μη ύπαρξης δεσμών ισχύουν οι ακόλουθες σχέσεις:

$$\overline{R}=\frac{\sum_{i=1}^{n}R_i}{n}=\frac{1}{n}\sum_{i=1}^{n}i=\frac{1}{n}\frac{n(n+1)}{2}=\frac{n+1}{2},$$

και

$$\overline{S}=\frac{\sum_{i=1}^{n}S_i}{n}=\frac{1}{n}\sum_{i=1}^{n}i=\frac{1}{n}\frac{n(n+1)}{2}=\frac{n+1}{2},$$

ενώ

$$\sum_{i=1}^{n}(R_i-\overline{R})^2=\sum_{i=1}^{n}\left(R_i-\frac{n+1}{2}\right)^2=\sum_{i=1}^{n}R_i^2-(n+1)\sum_{i=1}^{n}R_i+\sum_{i=1}^{n}\left(\frac{n+1}{2}\right)^2.$$

Από την τελευταία σχέση προκύπτει ότι:

$$\begin{aligned}\sum_{i=1}^{n}(R_i-\overline{R})^2 &= \sum_{i=1}^{n}i^2-(n+1)\sum_{i=1}^{n}i+\left(\frac{n+1}{2}\right)^2 n\\ &= \frac{n(n+1)(2n+1)}{6}-\frac{n(n+1)^2}{2}+\frac{n(n+1)^2}{4}\\ &= \frac{n\left(n^2-1\right)}{12}\end{aligned}$$

Με παρόμοιο τρόπο $\sum_{i=1}^{n}(S_i-\overline{S})^2=\frac{n\left(n^2-1\right)}{12}$. Έπειτα, λαμβάνοντας υπόψη τον ορισμό του συντελεστή συσχέτισης του Spearman, προκύπτει ότι:

$$r_s=\frac{\sum_{i=1}^{n}(R_i-\overline{R})(S_i-\overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i-\overline{R})^2}\sqrt{\sum_{i=1}^{n}(S_i-\overline{S})^2}}=\frac{\sum_{i=1}^{n}\left(R_i-\frac{n+1}{2}\right)\left(S_i-\frac{n+1}{2}\right)}{\frac{n(n^2-1)}{12}}.$$

β) Λαμβάνοντας υπόψη το προηγούμενο αποτέλεσμα, αρκεί να δείξουμε ότι:

$$\frac{\sum_{i=1}^{n}\left(R_i-\frac{n+1}{2}\right)\left(S_i-\frac{n+1}{2}\right)}{\frac{n(n^2-1)}{12}}=1-\frac{6\sum_{i=1}^{n}(R_i-S_i)^2}{n(n^2-1)},$$

ή, ισοδύναμα, ότι

$$12\sum_{i=1}^{n}\left(R_i-\frac{n+1}{2}\right)\left(S_i-\frac{n+1}{2}\right)=n(n^2-1)-6\sum_{i=1}^{n}(R_i-S_i)^2.$$

Όμως, καθώς δεν υπάρχουν δεσμοί, ισχύει ότι:

$$\begin{aligned}\sum_{i=1}^{n}\left(R_i-\frac{n+1}{2}\right)\left(S_i-\frac{n+1}{2}\right) &= \sum_{i=1}^{n}R_iS_i-\frac{n+1}{2}\sum_{i=1}^{n}R_i-\frac{n+1}{2}\sum_{i=1}^{n}S_i+\sum_{i=1}^{n}\frac{(n+1)^2}{4}\\ &= \sum_{i=1}^{n}R_iS_i-\frac{n+1}{2}\sum_{i=1}^{n}i-\frac{n+1}{2}\sum_{i=1}^{n}i+\frac{(n+1)^2}{4}n\\ &= \sum_{i=1}^{n}R_iS_i-\frac{n+1}{2}\frac{n(n+1)}{2}-\frac{n+1}{2}\frac{n(n+1)}{2}+\frac{(n+1)^2n}{4}\\ &= \sum_{i=1}^{n}R_iS_i-\frac{(n+1)^2n}{4}.\end{aligned}$$

Επιπλέον, είναι

$$\begin{aligned} n(n^2-1) - 6\sum_{i=1}^{n}(R_i - S_i)^2 &= n(n^2-1) - 6\left\{\sum_{i=1}^{n} R_i^2 - 2\sum_{i=1}^{n} R_i S_i + \sum_{i=1}^{n} S_i^2\right\} \\ &= n(n^2-1) - 6\left\{\sum_{i=1}^{n} i^2 - 2\sum_{i=1}^{n} R_i S_i + \sum_{i=1}^{n} i^2\right\} \\ &= 12\sum_{i=1}^{n} R_i S_i + n(n^2-1) - 12\sum_{i=1}^{n} i^2 \\ &= 12\sum_{i=1}^{n} R_i S_i + n(n^2-1) - 12\frac{n(n+1)(2n+1)}{6} \\ &= 12\sum_{i=1}^{n} R_i S_i + n(n^2-1) - 2n(n+1)(2n+1) \\ &= 12\left\{\sum_{i=1}^{n} R_i S_i - \frac{(n+1)^2 n}{4}\right\} \end{aligned}$$

και η απόδειξη ολοκληρώθηκε. □

Οι σχέσεις που προσδιορίστηκαν στην προηγούμενη πρόταση ισχύουν μόνο όταν δεν υπάρχουν δεσμοί μεταξύ των X_i ή των Y_i, ενώ στην περίπτωση ύπαρξης δεσμών είναι πολύ χρήσιμη η παρατήρηση που ακολουθεί.

Παρατήρηση 8.3. Στην περίπτωση ύπαρξης δεσμών (ties) μεταξύ των X_i ή των Y_i η κλασική αντιμετώπιση, όπως έχει ήδη αναφερθεί, είναι ο υπολογισμός, σε καθεμία από τις ίσες αυτές τιμές, του μέσου όρου των τάξεων που θα είχαν αν δεν ταυτίζονταν. Αν $u_1, u_2, \ldots$ και $v_1, v_2, \ldots$ είναι οι τάξεις των δειγματικών τιμών X_i και Y_i, αντίστοιχα, όπου έχουμε δεσμούς, τότε ο συντελεστής συσχέτισης υπολογίζεται από τη σχέση (για λεπτομέρειες, βλ. Sheskin, 2011):

$$r = \frac{n\left(n^2-1\right) - 6\sum(R_i - S_i)^2 - (U+V)/2}{\left[\left\{n\left(n^2-1\right) - U\right\}\left\{n\left(n^2-1\right) - V\right\}\right]^{1/2}},$$

όπου

$$U = \sum\left(u_i^3 - u_i\right) \text{ και } V = \sum\left(v_i^3 - v_i\right).$$

Όπως και ο συντελεστής συσχέτισης του Pearson, έτσι και ο αντίστοιχος του Spearman χρησιμοποιείται ως στατιστική συνάρτηση για τον έλεγχο της μη ύπαρξης συσχέτισης μεταξύ δύο μεταβλητών. Ειδικότερα, χρησιμοποιείται για τον έλεγχο της μηδενικής υπόθεσης $H_0 : \rho = 0$, δηλαδή της υπόθεσης ότι οι τυχαίες μεταβλητές X και Y είναι αμοιβαία ανεξάρτητες (mutually independent), έναντι μίας εκ των τριών εναλλακτικών:

(Α) $H_1 : \rho > 0$, δηλαδή της τάσης μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μεγάλες τιμές της άλλης και αντίστροφα (περίπτωση θετικής συσχέτισης),

(Β) $H_1 : \rho < 0$, δηλαδή της τάσης μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μικρές τιμές της άλλης και αντίστροφα (περίπτωση αρνητικής συσχέτισης),

(Γ) $H_1 : \rho \neq 0$, δηλαδή της τάσης είτε μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μεγάλες τιμές της άλλης και αντίστροφα είτε μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μικρές τιμές της άλλης και αντίστροφα (περίπτωση θετικής ή αρνητικής συσχέτισης).

Για τον έλεγχο των παραπάνω χρησιμοποιείται ο συντελεστής συσχέτισης του Spearman r_s. Στον Πίνακα Π.29 του Παραρτήματος (βλ., μεταξύ άλλων Χατζηνικολάου, 2002, και αναφορές εκεί) δίνονται οι κρίσιμες τιμές για τον συντελεστή συσχέτισης του Spearman, ενώ στη βιβλιογραφία (Conover, 1998) είναι διαθέσιμα και ποσοστιαία σημεία του r_s υπό τη μηδενική υπόθεση. Τότε οι κρίσιμες περιοχές για τον έλεγχο, σε επίπεδο σημαντικότητας a, καθενός εκ των παραπάνω προβλημάτων είναι:

(Α) Απορρίπτεται η μηδενική υπόθεση για μεγάλες τιμές του r_s, δηλαδή όταν ο συντελεστής συσχέτισης του Spearman ξεπερνά το a ποσοστιαίο σημείο του.

(Β) Απορρίπτεται η μηδενική υπόθεση για μικρές τιμές του r_s, δηλαδή όταν ο συντελεστής συσχέτισης του Spearman δεν ξεπερνά το $1-a$ ποσοστιαίο σημείο του.

(Γ) Απορρίπτεται η μηδενική υπόθεση, αν η απόλυτη τιμή του συντελεστή συσχέτισης του Spearman είναι μεγαλύτερη από το $a/2$ ποσοστιαίο σημείο του.

Αξίζει, επίσης, να αναφέρουμε πως υπάρχει η δυνατότητα να υπολογίσουμε ποσοστιαία σημεία της κατανομής του r_s υπό την H_0 χρησιμοποιώντας την εντολή `qSpearman(p,r)` του πακέτου `SuppDists` της `R`. Για τον υπολογισμό του άνω a ποσοστιαίου σημείου ($0 < a < 1$) πρέπει να δώσουμε `p=1-a`, ενώ στο όρισμα `r` δίνουμε το πλήθος των διαθέσιμων ζευγών παρατηρήσεων.

Παρατήρηση 8.4. Εναλλακτικά, κάποιος μπορεί να χρησιμοποιήσει τη στατιστική συνάρτηση

$$T = \sum_{i=1}^{n} (R_i - S_i)^2,$$

η οποία στη βιβλιογραφία ονομάζεται στατιστική συνάρτηση των Hotelling-Pabst (βλ. Hotelling και Pabst, 1936) και ποσοστιαία σημεία της είναι διαθέσιμα στη βιβλιογραφία (βλ. Conover, 1998). Παρατηρήστε ότι, λόγω της σχέσης β) της Πρότασης 8.2, προκύπτει ότι μεγάλες τιμές του συντελεστή συσχέτισης του Spearman αντιστοιχούν σε μικρές τιμές της στατιστικής συνάρτησης T και αντίστροφα. Επομένως, οι κρίσιμες περιοχές για τον έλεγχο σε επίπεδο σημαντικότητας a, καθενός εκ των παραπάνω προβλημάτων είναι:

(Α) Απορρίπτεται η μηδενική υπόθεση για μικρές τιμές του T, δηλαδή όταν δεν ξεπερνά το $1-a$ ποσοστιαίο σημείο του.

(Β) Απορρίπτεται η μηδενική υπόθεση για μεγάλες τιμές του T, δηλαδή όταν ξεπερνά το a ποσοστιαίο σημείο του.

(Γ) Απορρίπτεται η μηδενική υπόθεση είτε όταν η τιμή της στατιστικής συνάρτησης T ξεπερνά το $a/2$ ποσοστιαίο σημείο του είτε όταν δεν ξεπερνά το 1-$a/2$ ποσοστιαίο σημείο του.

Στην πρόταση που ακολουθεί, προσδιορίζεται η κατανομή του δειγματικού συντελεστή του Spearman υπό τη μηδενική υπόθεση, όταν το το μέγεθος δείγματος n είναι $n \geq 30$, και δεν υπάρχουν δεσμοί μεταξύ των παρατηρήσεων.

Πρόταση 8.3. Υπό την υπόθεση ότι οι τυχαίες μεταβλητές είναι ασυσχέτιστες και με την προϋπόθεση ότι δεν υπάρχουν δεσμοί μεταξύ των παρατηρήσεων ισχύει ότι η στατιστική συνάρτηση $Z = r_s\sqrt{n-1}$ ακολουθεί προσεγγιστικά την τυπική κανονική κατανομή.

Απόδειξη Πρότασης 8.3. Από την Πρόταση 8.2 α) προκύπτει ότι στην περίπτωση μη ύπαρξης δεσμών

$$r_s = \frac{12\sum_{i=1}^{n} R_i S_i}{n(n^2-1)} - \frac{3(n+1)}{n-1}.$$

Τότε, υπό τη μηδενική υπόθεση ότι οι τυχαίες μεταβλητές είναι ασυσχέτιστες, προκύπτει ότι:

$$\begin{aligned}\mathsf{E}(r_s) &= \frac{12\sum_{i=1}^{n}\mathsf{E}(R_iS_i)}{n(n^2-1)} - \frac{3(n+1)}{n-1} \\ &= \frac{12\sum_{i=1}^{n}\mathsf{E}(R_i)\mathsf{E}(S_i)}{n(n^2-1)} - \frac{3(n+1)}{n-1} \\ &= \frac{12n\mathsf{E}(R_i)\mathsf{E}(S_i)}{n(n^2-1)} - \frac{3(n+1)}{n-1}.\end{aligned}$$

Έπειτα, λαμβάνοντας υπόψη την Πρόταση 6.1 σχετικά με τις ιδιότητες των τάξεων, προκύπτει ότι:

$$\mathsf{E}(r_s) = \frac{12n\frac{n+1}{2}\frac{n+1}{2}}{n(n^2-1)} - \frac{3(n+1)}{n-1} = 0.$$

Επιπρόσθετα:

$$\mathsf{Var}(r_s) = \frac{12^2\mathsf{Var}\left(\sum_{i=1}^{n}R_iS_i\right)}{n^2(n^2-1)^2},$$

με

$$\begin{aligned}\mathsf{Var}\left(\sum_{i=1}^{n}R_iS_i\right) &= \sum_{i=1}^{n}\mathsf{Var}(R_iS_i) + \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathsf{Cov}\left(R_iS_i, R_jS_j\right) \\ &= n\mathsf{Var}(R_i)\mathsf{Var}(S_i) + n(n-1)\mathsf{Cov}\left(R_i, R_j\right)\mathsf{Cov}\left(S_i, S_j\right).\end{aligned}$$

Από την Πρόταση 6.1 έχουμε ότι

$$\mathsf{Var}(R_i) = \mathsf{Var}(S_i) = \frac{n^2-1}{12},$$

και

$$\mathsf{Cov}\left(R_i, R_j\right) = \mathsf{Cov}\left(S_i, S_j\right) = -\frac{n+1}{12}.$$

Συνδυάζοντας τα παραπάνω και ύστερα από λίγες αλγεβρικές πράξεις, καταλήγουμε ότι $\mathsf{Var}(r_s) = \frac{1}{n-1}$ και το ζητούμενο προκύπτει με άμεση εφαρμογή του Κεντρικού Οριακού Θεωρήματος. □

Η στατιστική συνάρτηση $Z = r_s\sqrt{n-1}$ χρησιμοποιείται για τον έλεγχο της μηδενικής υπόθεσης $H_0 : \rho = 0$, έναντι μίας εκ των τριών εναλλακτικών:

(Α) $H_1 : \rho > 0$, με κρίσιμη περιοχή $Z = r_s\sqrt{n-1} \geq z_a$,
(Β) $H_1 : \rho < 0$, με κρίσιμη περιοχή $Z = r_s\sqrt{n-1} \leq -z_a$,
(Γ) $H_1 : \rho \neq 0$, με κρίσιμη περιοχή $|Z| = \left|r_s\sqrt{n-1}\right| \geq z_{a/2}$.

Παρατήρηση 8.5. Μία καλύτερη προσέγγιση, η οποία συνήθως χρησιμοποιείται για μέγεθος δείγματος $n \geqslant 10$, προτάθηκε από τους Kendall and Gibbons (1990) και είναι ανάλογη με αυτήν της Πρότασης 8.1. Δηλαδή χρησιμοποιείται η στατιστική συνάρτηση

$$t_s = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

η οποία υπό την $H_0 : \rho = 0$ έχει κατά προσέγγιση κατανομή t με $n-2$ βαθμούς ελευθερίας (βλ. Sprent and Smeeton, 2007).

Επίσης, μπορούμε να χρησιμοποιήσουμε τον μετασχηματισμό του Fisher,

$$W_s = \frac{1}{2}\ln\frac{1+r_s}{1-r_s},$$

όπου, υπό την υπόθεση της (διδιάστατης) κανονικότητας, η σ.σ. W_s έχει, προσεγγιστικά, κανονική κατανομή $N\left(0, \frac{1.06}{n-3}\right)$, υπό την $H_0 : \rho = 0$. Στην περίπτωση που παραβιάζεται η υπόθεση της κανονικότητας (από τα δεδομένα), τότε η παραπάνω προσέγγιση για την κατανομή του W_s είναι βάσιμη, όταν $n > 50$ (βλ. Yu and Hutson, 2022).

Παράδειγμα 8.2. Να υπολογίσετε τον συντελεστή συσχέτισης του Spearman για τα δεδομένα του Παραδείγματος 8.1 (τιμές (X_i, Y_i), $i = 1, 2, \ldots, 7$) και να ελέγξετε, σε επίπεδο σημαντικότητας 5%, την υπόθεση ότι οι δύο μεταβλητές δεν συσχετίζονται.

Λύση Παραδείγματος 8.2. Αν με R_i και S_i συμβολίσουμε τις τάξεις των δειγματικών τιμών των τυχαίων μεταβλητών X και Y, αντίστοιχα, τότε είναι:

X_i	6	7	4	8	9	3	5
Y_i	80	83	75	86	95	72	69
R_i	4	5	2	6	7	1	3
S_i	4	5	3	6	7	2	1
$(R_i - S_i)^2$	0	0	1	0	0	1	4

Επομένως, ο συντελεστής συσχέτισης του Spearman, καθώς δεν υπάρχουν δεσμοί, είναι:

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(R_i - S_i)^2}{n(n^2-1)} = 1 - \frac{6 \cdot (0+0+1+0+0+1+4)}{7(7^2-1)} = 1 - \frac{36}{336} = 0.893.$$

Χρησιμοποιώντας τον Πίνακα Π.29 του Παραρτήματος για τον δίπλευρο έλεγχο απορρίπτεται η μηδενική υπόθεση, όταν ο συντελεστής συσχέτισης του Spearman ξεπερνά κατά απόλυτη τιμή την τιμή 0.786. Συνεπώς, καθώς $|r_s| > 0.786$, η μηδενική υπόθεση απορρίπτεται. Αυτό σημαίνει ότι υπάρχει στατιστικά σημαντική συσχέτιση μεταξύ των δύο μεταβλητών και, μάλιστα, θετική συσχέτιση. □

8.4 Συντελεστής συσχέτισης του Kendall

Ο Kendall (1938) πρότεινε ένα εναλλακτικό μέτρο διδιάστατης εξάρτησης στηριζόμενος στις έννοιες των «σύμφωνων ζευγαριών» (concordant pairs) και των «ασύμφωνων ζευγαριών» (discordant pairs), οι ορισμοί των οποίων ακολουθούν.

Ορισμός 8.4

Σύμφωνα ζευγάρια είναι εκείνα για τα οποία οι τιμές των (X_i, Y_i) και $\left(X_j, Y_j\right)$, $i \neq j$, $i,j = 1,\ldots,n$, έχουν την ίδια διάταξη και για τις δύο μεταβλητές, δηλαδή είτε είναι $X_i < X_j$ και $Y_i < Y_j$ είτε $X_i > X_j$ και $Y_i > Y_j$. Εναλλακτικά, μπορούμε να πούμε ότι σύμφωνα ζευγάρια είναι εκείνα στα οποία το πρόσημο της διαφοράς $X_i - X_j$ είναι σύμφωνο με το πρόσημο της διαφοράς $Y_i - Y_j$.

Ορισμός 8.5

Ασύμφωνα ζευγάρια είναι εκείνα για τα οποία οι τιμές των (X_i, Y_i) και $\left(X_j, Y_j\right)$, $i \neq j$, $i,j = 1,\ldots,n$, δεν έχουν την ίδια διάταξη και για τις δύο μεταβλητές, δηλαδή είτε είναι $X_i < X_j$ και $Y_i > Y_j$ είτε $X_i > X_j$ και $Y_i < Y_j$. Εναλλακτικά, μπορούμε να πούμε ότι ότι ασύμφωνα ζευγάρια είναι εκείνα στα οποία το πρόσημο της διαφοράς $X_i - X_j$ δεν είναι σύμφωνο με το πρόσημο της διαφοράς $Y_i - Y_j$.

Ορισμός 8.6

Ισοβαθμισμένα ζευγάρια είναι εκείνα για τα οποία οι τιμές των (X_i, Y_i) και (X_j, Y_j), $i \neq j$, $i,j = 1,\ldots,n$, είναι τέτοιες, ώστε ή $X_i = X_j$ ή $Y_i = Y_j$ ή $X_i = X_j$ και $Y_i = Y_j$. Εναλλακτικά, μπορούμε να πούμε ότι ότι ισοβαθμισμένα ζευγάρια είναι εκείνα στα οποία ή $X_i - X_j = 0$ ή $Y_i - Y_j = 0$ ή και τα δύο.

Έστω $(X_1, Y_1),\ldots,(X_n, Y_n)$ ένα τυχαίο δείγμα από n ζεύγη παρατηρήσεων για τις τυχαίες μεταβλητές X και Y. Επιπλέον, έστω n_c, n_d και n_0 ο αριθμός των σύμφωνων, ασύμφωνων και ισοβαθμισμένων ζευγαριών, αντίστοιχα. Τότε είναι

$$n_c + n_d + n_0 = \frac{n(n-1)}{2} = \binom{n}{2}$$

καθώς το άθροισμα των n_c, n_d και n_0 ισούται με τον δυνατό αριθμό των ζευγαριών. Στο παραπάνω πλαίσιο, ο Kendall (1938) πρότεινε τη στατιστική συνάρτηση:

$$\tau = \frac{n_c - n_d}{n(n-1)/2}. \tag{8.3}$$

Η παραπάνω στατιστική συνάρτηση δεν είναι τίποτε άλλο παρά το δειγματικό ανάλογο ενός συντελεστή που ορίστηκε από τον Kendall (1938) ως η διαφορά των πιθανοτήτων συμφωνίας και ασυμφωνίας p_c και p_d, αντίστοιχα, των (X_i, Y_i) και (X_j, Y_j). Ειδικότερα:

$$p_c = P\left[\left(X_i - X_j\right)\left(Y_i - Y_j\right) > 0\right]$$

και

$$p_d = P\left[\left(X_i - X_j\right)\left(Y_i - Y_j\right) < 0\right].$$

Εύκολα, μπορεί να διαπιστώσει κανείς ότι αν X και Y είναι ανεξάρτητες συνεχείς τυχαίες μεταβλητές, τότε $p_c - p_d = 0$, ενώ το αντίστροφο δεν ισχύει πάντοτε. Παρόλα αυτά, αποδεικνύεται ότι ο συντελεστής τ είναι ίσος με το μηδέν αν και μόνο αν οι τυχαίες μεταβλητές είναι ασυσχέτιστες (Sheskin, 2011).

Παρατήρηση 8.6. Για τον εύκολο υπολογισμό των n_c και n_d καλό είναι να ακολουθείται η μεθοδολογία που περιγράφεται μεταξύ άλλων από τους Conover *et al.* (1981). Πιο συγκεκριμένα, αρχικά τα ζεύγη των n παρατηρήσεων $(X_1, Y_1),\ldots,(X_n, Y_n)$ διατάσσονται κατά αύξουσα τάξη μεγέθους ως προς τη μεταβλητή X. Τότε η εύρεση του πλήθους των σύμφωνων ζευγαριών κάθε ζεύγους έγκειται στον υπολογισμό του αριθμού των παρατηρήσεων με τιμές στην τυχαία μεταβλητή Y μεγαλύτερη από τη συγκεκριμένη παρατήρηση, ενώ η εύρεση του πλήθους των ασύμφωνων ζευγαριών κάθε ζεύγους έγκειται στον υπολογισμό του αριθμού των παρατηρήσεων με τιμές στην τυχαία μεταβλητή Y μικρότερη από τη συγκεκριμένη παρατήρηση.

Παρατήρηση 8.7. Από τη σχέση (8.3) προκύπτει ότι στην περίπτωση που όλα τα δυνατά ζευγάρια είναι σύμφωνα, τότε $\tau = 1$, ενώ, όταν όλα τα δυνατά ζευγάρια είναι ασύμφωνα έχουμε ότι $\tau = -1$.

Όπως και ο συντελεστής συσχέτισης του Spearman έτσι και ο συντελεστής του Kendall (ή μια συνάρτηση αυτού) χρησιμοποιείται ως στατιστική συνάρτηση για τον έλεγχο της μη ύπαρξης συσχέτισης δύο μεταβλητών. Ειδικότερα, χρησιμοποιείται για τον έλεγχο της μηδενικής υπόθεσης $H_0 : \rho = 0$, δηλαδή της υπόθεσης ότι οι τυχαίες μεταβλητές X και Y είναι αμοιβαία ανεξάρτητες (mutually independent), έναντι μίας εκ των τριών εναλλακτικών:

(A) $H_1 : \rho > 0$, δηλαδή της τάσης μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μεγάλες τιμές της άλλης και αντίστροφα,

(B) $H_1 : \rho < 0$, δηλαδή της τάσης μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μικρές τιμές της άλλης και αντίστροφα,

(Γ) $H_1 : \rho \neq 0$, δηλαδή της τάσης είτε μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μεγάλες τιμές της άλλης και αντίστροφα είτε μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μικρές τιμές της άλλης και αντίστροφα.

Για τον έλεγχο των παραπάνω χρησιμοποιείται είτε ο συντελεστής του Kendall είτε για ευκολία η στατιστική συνάρτηση

$$T = n_c - n_d = \binom{n}{2}\tau.$$

Αποδεικνύεται ότι η κατανομή της στατιστικής συνάρτησης T υπό τη μηδενική υπόθεση είναι συμμετρική γύρω από τη μέση τιμή της, που είναι η τιμή 0. Επομένως, ισχύει, υπό τη μηδενική υπόθεση, ότι $P(T \leqslant -x) = P(T \geqslant x)$. Στον Πίνακα Π.30 του Παραρτήματος (βλ., επίσης, Conover *et al.*, 1981) δίνονται ποσοστιαία σημεία T_p της στατιστικής συνάρτησης T υπό τη μηδενική υπόθεση, που είναι τέτοια ώστε $P(T > T_p) \leqslant p \leqslant P(T \geqslant T_p)$.

Αξίζει, επίσης, να αναφέρουμε πως υπάρχει η δυνατότητα να υπολογίσουμε ποσοστιαία σημεία της κατανομής της T υπό την H_0 χρησιμοποιώντας την εντολή `0.5*N*(N-1)*qKendall(p,N)`. Η εντολή `qKendall(p,N)` (υπάρχει στο πακέτο `SuppDists` της `R`) δίνει ποσοστιαία σημεία της κατανομής του συντελεστή συσχέτισης τ όταν η H_0 είναι αληθής. Για τον υπολογισμό του άνω a ποσοστιαίου σημείου ($0 < a < 1$) πρέπει να δώσουμε `p=1-a`, ενώ στο όρισμα `N` δίνουμε το πλήθος των διαθέσιμων ζευγών παρατηρήσεων (μέγεθος δείγματος n). Για να μετατρέψουμε τα ποσοστιαία σημεία της κατανομής της τ σε ποσοστιαία σημεία της T αρκεί να τα πολλαπλασιάσουμε με $n(n-1)/2$.

Τότε οι κρίσιμες περιοχές για καθέναν από τους παραπάνω ελέγχους σε επίπεδο σημαντικότητας a είναι:

(Α) Απορρίπτεται η μηδενική υπόθεση, όταν η τιμή της στατιστικής συνάρτησης T είναι μεγαλύτερη από την τιμή του Πίνακα Π.30 για $p = a$.

(Β) Απορρίπτεται η μηδενική υπόθεση, όταν η τιμή της στατιστικής συνάρτησης T είναι μικρότερη από την αντίθετη της τιμής του Πίνακα Π.30 για $p = a$.

(Γ) Λαμβάνοντας υπόψη τη συμμετρία της κατανομής της στατιστικής συνάρτησης T, η μηδενική υπόθεση απορρίπτεται, αν η τιμή της στατιστικής συνάρτησης είναι μεγαλύτερη κατά απόλυτη τιμή από την τιμή του Πίνακα Π.30 για $p = a/2$.

Τέλος, για μεγάλο μέγεθος δείγματος αποδεικνύεται ότι (Kendall and Stuart, 1961),

$$Z = \frac{3\tau\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} \underset{H_0}{\overset{\alpha\sigma\upsilon\mu\pi.}{\sim}} \mathcal{N}(0,1). \tag{8.4}$$

Μάλιστα, η παραπάνω προσέγγιση θεωρείται ικανοποιητική για $n > 10$ (Sheskin, 2011).

Παρατήρηση 8.8. Εύλογα μπορεί να αναρωτηθεί κάποιος αν υπάρχουν διαφορές μεταξύ του συντελεστή του Spearman και του Kendall. Ο συντελεστής του Spearman τείνει να λαμβάνει μεγαλύτερες απόλυτες τιμές από τις αντίστοιχες του Kendall, ενώ, από την άλλη, η κανονική προσέγγιση της κατανομής του συντελεστή του Kendall, τ, είναι πολύ πιο γρήγορη από αυτήν της κατανομής του Spearman. Η παραπάνω ιδιότητα έχει ως συνέπεια τα προσεγγιστικά ποσοστιαία σημεία που βασίζονται στην τ να είναι πιο αξιόπιστα (για λεπτομέρειες, βλ. Gibbons and Chakraborti, 2020).

Παράδειγμα 8.3. (Conover *et al.*, 1981). Δέκα ζευγάρια που πηγαίνουν συχνά για μπόουλινγκ καταγράφουν στον πίνακα που ακολουθεί την επίδοση (σκορ) του άνδρα (τ.μ. X) και την αντίστοιχη επίδοση σκορ της γυναίκας (τ.μ. Y). Υπολογίστε τον συντελεστή συσχέτισης του Kendall και αποφανθείτε με επίπεδο σημαντικότητας 5% αν οι επιδόσεις ανδρών και γυναικών είναι αμοιβαία ανεξάρτητες.

Επιδόσεις Άντρα X_i	147	158	131	142	183	151	196	129	155	158
Επιδόσεις Γυναίκας Y_i	122	128	125	123	115	120	108	143	124	123

Λύση Παραδείγματος 8.3. Αρχικά, τα ζεύγη των $n = 10$ παρατηρήσεων (X_i, Y_i), $i = 1,...,10$, διατάσσονται κατά αύξουσα τάξη μεγέθους ως προς τη μεταβλητή X, που περιγράφει τις επιδόσεις του άντρα στο μπόουλινγκ. Τότε η εύρεση του πλήθους των σύμφωνων ζευγαριών κάθε ζεύγους έγκειται στον υπολογισμό του αριθμού των παρατηρήσεων με τιμές στην τυχαία μεταβλητή Y, που περιγράφει την επίδοση της γυναίκας στο μπόουλινγκ, μεγαλύτερες από την τιμή της συγκεκριμένης παρατήρησης, ενώ η εύρεση του πλήθους των ασύμφωνων ζευγαριών κάθε ζεύγους έγκειται στον υπολογισμό του αριθμού των παρατηρήσεων με τιμές στην τυχαία μεταβλητή Y μικρότερες από την τιμή της συγκεκριμένης παρατήρησης.

Για μεγαλύτερη ευκολία μπορούμε να χρησιμοποιήσουμε τον ακόλουθο πίνακα:

$X_{(i)}$	Αντίστοιχες τιμές της τ.μ. Y	Αριθμός σύμφωνων	Αριθμός ασύμφωνων
129	143	0	9
131	125	1	7
142	123	2	4
147	122	3	3
151	120	3	2
155	124	1	3
158	123	0	2
158	128	0	2
183	115	0	1
196	108	0	0
Σύνολο		$n_c = 10$	$n_d = 33$

Ο πίνακας αυτός μας δίνει τις απαραίτητες τιμές για τον υπολογισμό της στατιστικής συνάρτησης για τον έλεγχο της μηδενικής υπόθεσης $H_0 : \rho = 0$, δηλαδή της υπόθεσης ότι οι τυχαίες μεταβλητές X και Y είναι αμοιβαία ανεξάρτητες, έναντι της δίπλευρης εναλλακτικής υπόθεσης $H_1 : \rho \neq 0$, δηλαδή της ύπαρξης τάσης είτε μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μεγάλες τιμές της άλλης, και αντίστροφα, είτε μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μικρές τιμές της άλλης, και αντίστροφα. Η στατιστική συνάρτηση του ελέγχου είναι $T = n_c - n_d$ με τιμή $T = 10 - 33 = -23$. Όπως έχουμε ήδη αναφέρει, η μηδενική υπόθεση απορρίπτεται, έναντι της δίπλευρης εναλλακτικής, όταν η τιμή της στατιστικής συνάρτησης T ξεπερνά κατά απόλυτη τιμή το $a/2$ ποσοστιαίο σημείο της κατανομής της T υπό την H_0, έστω αυτό $w_{a/2}$. Τιμές των ποσοστιαίων σημείων $w_{a/2}$, για διάφορες τιμές του επιπέδου σημαντικότητας a, δίνονται στον Πίνακα Π.30. Άρα, για $a = 0.05$, η μορφή της κρίσιμης περιοχής του ελέγχου είναι $|T| > w_{0.025} = 21$. Επομένως, καθώς $|-23| > 21$, η μηδενική υπόθεση απορρίπτεται. Αυτό σημαίνει ότι υπάρχει στατιστικά σημαντική τάση είτε μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μεγάλες τιμές της άλλης, και αντίστροφα, είτε μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μικρές τιμές της άλλης, και αντίστροφα. □

Παρατήρηση 8.9. Όταν υπάρχει σημαντικός αριθμός ισοβαθμιών (ties) στο δείγμα, δεν υπολογίζεται η τιμή της στατιστικής συνάρτησης της σχέσης (8.4), αλλά υπολογίζεται η τιμή $z = (n_c - n_d)/\xi$, όπου η ποσότητα ξ είναι μια «διορθωμένη» τυπική απόκλιση, έχοντας λάβει υπόψη το πλήθος των ισοβαθμιών. Συγκεκριμένα, αν t_i, u_j εκφράζουν, αντίστοιχα, το πλήθος των στοιχείων στην i-οστή ομάδα ισόβαθμων παρατηρήσεων για το δείγμα των X και το πλήθος των στοιχείων στην j-οστή ομάδα ισόβαθμων παρατηρήσεων για το δείγμα των Y, τότε

$$\xi = \sqrt{\frac{\xi_0 - \xi_x - \xi_y}{18} + \xi_1 + \xi_2},$$

όπου

$$\xi_0 = n(n-1)(2n+5),\ \xi_x = \sum_i t_i(t_i-1)(2t_i+5),\ \xi_y = \sum_j u_j(u_j-1)(2u_j+5),$$

$$\xi_1 = \frac{\sum_i t_i(t_i-1)\cdot\sum_j u_j(u_j-1)}{2n(n-1)},$$
$$\xi_2 = \frac{\sum_i t_i(t_i-1)(t_i-2)\cdot\sum_j u_j(u_j-1)(u_j-2)}{9n(n-1)(n-2)}.$$

Για περισσότερες λεπτομέρειες παραπέμπουμε τον/την ενδιαφερόμενο/μενη αναγνώστη/στρια στο σύγγραμμα των Kendall and Gibbons (1990).

8.5 Συντελεστής συσχέτισης των Goodman and Kruskal

Στις περιπτώσεις όπου τα δεδομένα μας ταξινομούνται σε πίνακες συνάφειας όπου και οι δύο μεταβλητές είναι διατάξιμες (ordinal) και θέλουμε να εξετάσουμε αν υπάρχει συσχέτιση μεταξύ αυτών των μεταβλητών, χρησιμοποιείται ο συντελεστής συσχέτισης των Goodman and Kruskal (1954). Ο συντελεστής αυτός συμβολίζεται με γ και ορίζεται ως

$$\gamma = \frac{n_c - n_d}{n_c + n_d} \tag{8.5}$$

όπου n_c και n_d είναι ο αριθμός των σύμφωνων και ασύμφωνων ζευγαριών αντίστοιχα, όπως αυτά ορίστηκαν στην προηγούμενη ενότητα. Η χρήση της (8.5) είναι δυνατή και στην περίπτωση όπου υπάρχουν δεσμοί στα δεδομένα. Ο ορισμός του συντελεστή συσχέτισης γ είναι παρόμοιος με αυτόν του συντελεστή τ του Kendall (βλέπε τη σχέση (8.3)). Επομένως, οι τιμές του συντελεστή γ ανήκουν στο διάστημα $[-1,1]$ και προκύπτει ότι τιμές του γ κοντά στο -1 (1, αντίστοιχα) αποτελούν ένδειξη ότι οι μεταβλητές X και Y είναι αρνητικά (θετικά, αντίστοιχα) συσχετισμένες. Τιμές του γ κοντά στο 0 δηλώνουν ότι οι X και Y είναι ασυσχέτιστες. Για περισσότερες λεπτομέρειες πάνω σε μέτρα συσχέτισης, παραπέμπουμε τον/την αναγνώστη/στρια στις εργασίες-μελέτες των Goodman and Kruskal (1954; 1959; 1963; 1972).

Βέβαια, αν θέλουμε να ελέγξουμε την ύπαρξη ή μη της συσχέτισης των μεταβλητών X και Y, μπορούμε να διεξάγουμε τον έλεγχο της μηδενικής υπόθεσης $H_0 : \rho = 0$, δηλαδή της υπόθεσης ότι οι τυχαίες μεταβλητές X και Y είναι αμοιβαία ανεξάρτητες (mutually independent), έναντι μίας εκ των τριών εναλλακτικών:

(Α) $H_1 : \rho > 0$, δηλαδή της τάσης μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μεγάλες τιμές της άλλης, και αντίστροφα,
(Β) $H_1 : \rho < 0$, δηλαδή της τάσης μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μικρές τιμές της άλλης, και αντίστροφα,
(Γ) $H_1 : \rho \neq 0$, δηλαδή της τάσης είτε μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μεγάλες τιμές της άλλης και αντίστροφα είτε μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μικρές τιμές της άλλης, και αντίστροφα.

Συνήθως, για τα προβλήματα (Α), (Β) ή (Γ) χρησιμοποιείται ένας ασυμπτωτικός έλεγχος (βλ. Sheskin, 2011) κάνοντας τον μετασχηματισμό

$$\mathrm{Z} = \gamma\sqrt{\frac{n_c + n_d}{n(1-\gamma^2)}}$$

όπου για μεγάλες τιμές του n και υπό τη μηδενική υπόθεση $H_0 : \rho = 0$, αποδεικνύεται (Goodman and Kruskal, 1963) ότι $Z \sim \mathcal{N}(0,1)$. Επομένως, χρησιμοποιώντας τα ποσοστιαία σημεία της τυπικής κανονικής κατανομής και σε επίπεδο σημαντικότητας a, $0 < a < 1$, έχουμε ότι:

(Α) Απορρίπτεται η μηδενική υπόθεση, όταν $Z > z_a$.
(Β) Απορρίπτεται η μηδενική υπόθεση, όταν $Z \leqslant -z_a$.
(Γ) Απορρίπτεται η μηδενική υπόθεση, όταν $|Z| > z_{a/2}$.

Παράδειγμα 8.4. (Sheskin, 2011) Ένας ερευνητής θέλει να διαπιστώσει αν υπάρχει σχέση μεταξύ του βάρους ενός ανθρώπου (το οποίο ορίζουμε ως μεταβλητή X) και της σειράς γέννησης (την οποία ορίζουμε

ως μεταβλητή Y). Αφού καταγράφηκαν το βάρος και η σειρά γέννησης 300 ατόμων, τα δεδομένα ταξινομήθηκαν σε τρεις κατηγορίες βάρους και τέσσερις κατηγορίες που αφορούν τη σειρά γέννησης. Ειδικότερα, εμπλέκονται τρεις κατηγορίες σχετιζόμενες με το βάρος: κάτω από τον μέσο όρο (Μ.Ο.), μέσος όρος και πάνω από τον μέσο όρο. Οι παρακάτω τέσσερις κατηγορίες σχετίζονται με τη σειρά γέννησης: πρωτότοκος, δευτερότοκος, τριτότοκος, όλοι οι υπόλοιποι. Στον Πίνακα 8.1 συνοψίζουμε τα δεδομένα στις παραπάνω κατηγορίες.

		Σειρά Γέννησης			
		Πρωτότοκος	**Δευτερότοκος**	**Τριτότοκος**	**Υπόλοιποι**
Βάρος	**< Μ.Ο.**	70	15	10	5
	Μ.Ο.	10	60	20	10
	> Μ.Ο.	10	15	35	40

Πίνακας 8.1: Δεδομένα Παραδείγματος 8.4.

Σε επίπεδο σημαντικότητας 5% ελέγξτε αν υπάρχει σημαντική σχέση ανάμεσα στο βάρος του ατόμου και στη σειρά γέννησης.

Λύση Παραδείγματος 8.4. Συμβολίζουμε με K_{ij} το κελί της i-οστής γραμμής, $i = 1,2,3$, και j-οστής στήλης, $j = 1,2,3,4$ του Πίνακα 8.1. Αρχικά, για τον υπολογισμό του n_c χρησιμοποιούμε τη βηματική διαδικασία που περιγράφεται στη συνέχεια.

(Β1) Ξεκινάμε από το κελί K_{11} και πολλαπλασιάζουμε τη συχνότητα αυτού του κελιού με το άθροισμα των συχνοτήτων όλων των άλλων κελιών του Πίνακα που βρίσκονται κάτω από την 1η γραμμή και δεξιά της 1ης στήλης. Στον Πίνακα 8.1 αναφερόμαστε στα κελιά K_{22}, K_{23}, K_{24}, K_{32}, K_{33} και K_{34}. Είναι προφανές ότι κάθε ζεύγος τιμών που βρίσκεται σε ένα από τα προαναφερθέντα κελιά είναι σύμφωνο με κάθε ζεύγος του K_{11}. Επομένως, αν n_{ij} είναι το πλήθος των τιμών του κελιού K_{ij}, τότε ο αριθμός των σύμφωνων ζευγαριών του K_{11} με τα υπόλοιπα, θα είναι $n_{11} \cdot (n_{22} + n_{23} + n_{24} + n_{32} + n_{33} + n_{34})$. Δηλαδή, για τον Πίνακα 8.1, αυτός ο αριθμός θα είναι $70 \cdot (60 + 20 + 10 + 15 + 35 + 40) = 12600$.

(Β2) Την ίδια διαδικασία που περιγράψαμε προηγουμένως, ακολουθούμε και για τα υπόλοιπα κελιά. Για παράδειγμα, τα ζευγάρια του κελιού K_{12} είναι σύμφωνα με τα ζευγάρια τιμών που υπάρχουν στα κελιά κάτω από την 1η γραμμή και δεξιά της 2ης στήλης, δηλαδή τα K_{23}, K_{24}, K_{33} και K_{34}. Ο αριθμός των σύμφωνων ζευγαριών του K_{12} με τα υπόλοιπα θα είναι $n_{12} \cdot (n_{23} + n_{24} + n_{33} + n_{34})$. Στην περίπτωση του παραδείγματος, ο αριθμός αυτός είναι $15 \cdot (20 + 10 + 35 + 40) = 1575$.

(Β3) Στον Πίνακα 8.2 υπολογίζουμε τον αριθμό των σύμφωνων ζευγαριών για όλες τις δυνατές περιπτώσεις.

Για να υπολογίσουμε το n_d, δηλαδή τον αριθμό των ασύμφωνων ζευγαριών, χρησιμοποιούμε μία ανάλογη διαδικασία. Τα βήματά της είναι:

(Β1) Σε αυτήν την περίπτωση, ξεκινάμε από το κελί K_{14}, δηλαδή αυτό που βρίσκεται στην πάνω δεξιά γωνία του πίνακα, και πολλαπλασιάζουμε τη συχνότητα αυτού του κελιού με το άθροισμα των συχνοτήτων όλων των άλλων κελιών του Πίνακα που βρίσκονται κάτω από την 1η γραμμή και αριστερά της 4ης στήλης. Στον Πίνακα 8.1 αναφερόμαστε στα κελιά K_{21}, K_{22}, K_{23}, K_{31}, K_{32} και K_{33}. Είναι, πάλι, προφανές ότι κάθε ζεύγος τιμών που βρίσκεται σε ένα από τα προαναφερθέντα κελιά είναι ασύμφωνο με κάθε ζεύγος του K_{14}. Επομένως, ο αριθμός των ασύμφωνων ζευγαριών του K_{14} με τα υπόλοιπα, θα είναι $n_{14} \cdot (n_{21} + n_{22} + n_{23} + n_{31} + n_{32} + n_{33})$. Δηλαδή, για τον Πίνακα 8.1, αυτός ο αριθμός θα είναι $5 \cdot (10 + 60 + 20 + 10 + 15 + 35) = 750$.

(Β2) Την ίδια διαδικασία που περιγράψαμε προηγουμένως, ακολουθούμε και για τα υπόλοιπα κελιά. Για παράδειγμα, τα ζευγάρια του κελιού K_{13} είναι ασύμφωνα με τα ζευγάρια τιμών που υπάρχουν στα κελιά κάτω από την 1η γραμμή και αριστερά της 3ης στήλης, δηλαδή τα K_{21}, K_{22}, K_{31} και K_{32}. Ο αριθμός των ασύμφωνων ζευγαριών του K_{13} με τα υπόλοιπα θα είναι $n_{13} \cdot (n_{21} + n_{22} + n_{31} + n_{32})$. Στην περίπτωση του παραδείγματος, ο αριθμός αυτός είναι $10 \cdot (10 + 60 + 10 + 5) = 950$.

$K_{11} : 70(60 + 20 + 10 + 15 + 35 + 40)$	=	12600
$K_{12} : 15 \cdot (20 + 10 + 35 + 40)$	=	1575
$K_{13} : 10 \cdot (10 + 40)$	=	500
$K_{14} : 5 \cdot (0)$	=	0
$K_{21} : 10 \cdot (15 + 35 + 40)$	=	900
$K_{22} : 60 \cdot (35 + 40)$	=	4500
$K_{23} : 20 \cdot (40)$	=	800
$K_{24} : 10 \cdot (0)$	=	0
$K_{31} : 10 \cdot (0)$	=	0
$K_{32} : 15 \cdot (0)$	=	0
$K_{33} : 35 \cdot (0)$	=	0
$K_{34} : 40 \cdot (0)$	=	0
n_c = ο συνολικός αριθμός των σύμφωνων ζευγαριών	=	20875

Πίνακας 8.2: Υπολογισμός του n_c για το Παράδειγμα 8.4.

(Β3) Στον Πίνακα 8.3 υπολογίζουμε τον αριθμό των σύμφωνων ζευγαριών για όλες τις δυνατές περιπτώσεις.

$K_{14} : 5 \cdot (10 + 60 + 20 + 10 + 15 + 35)$	=	750
$K_{13} : 10 \cdot (10 + 60 + 10 + 5)$	=	950
$K_{12} : 15 \cdot (10 + 10)$	=	300
$K_{11} : 70 \cdot (0)$	=	0
$K_{24} : 10 \cdot (10 + 15 + 35)$	=	600
$K_{23} : 20 \cdot (10 + 15)$	=	500
$K_{22} : 60 \cdot (10)$	=	600
$K_{21} : 10 \cdot (0)$	=	0
$K_{34} : 40 \cdot (0)$	=	0
$K_{33} : 35 \cdot (0)$	=	0
$K_{32} : 15 \cdot (0)$	=	0
$K_{31} : 10 \cdot (0)$	=	0
n_d = ο συνολικός αριθμός των ασύμφωνων ζευγαριών	=	3700

Πίνακας 8.3: Υπολογισμός του n_d για το Παράδειγμα 8.4.

Από τη σχέση 8.5 υπολογίζουμε την τιμή του συντελεστή γ ως

$$\gamma = \frac{20875 - 3700}{20875 + 3700} = 0.7$$

και, επομένως, μπορούμε να κάνουμε ένα $z - test$, χρησιμοποιώντας τη στατιστική συνάρτηση

$$\mathrm{Z} = \gamma \sqrt{\frac{n_c + n_d}{n(1 - \gamma^2)}}$$

η οποία για το παράδειγμά μας έχει τιμή

$$z = 0.7 \sqrt{\frac{20875 + 3700}{300(1 - (0.7)^2)}} = 8.87$$

Σε επίπεδο σημαντικότητας 5%, $|z| = 8.87 > 1.645 = z_{0.025}$, επομένως απορρίπτουμε τη μηδενική υπόθεση του δίπλευρου προβλήματος (Γ) $H_0 : \rho = 0$, έναντι $H_1 : \rho \neq 0$, δηλαδή υπάρχει σημαντική σχέση ανάμεσα στο βάρος του ατόμου και στη σειρά γέννησης. □

8.5.1 Συντελεστής συσχέτισης του Yule

Στην περίπτωση που τα δεδομένα μπορούν να ταξινομηθούν σε έναν 2×2 πίνακα συνάφειας (βλ. Πίνακα 8.4), τότε χρησιμοποιείται ο συντελεστής συσχέτισης Q του Yule, ο οποίος αποτελεί ειδική περίπτωση του συντελεστή συσχέτισης των Goodman and Kruskal.

	Στήλη 1	**Στήλη 2**
Γραμμή 1	a	b
Γραμμή 2	c	d

Πίνακας 8.4: Μοντέλο 2×2 Πίνακα Συνάφειας.

Πράγματι, η τιμή του συντελεστή Q δίνεται από τη σχέση (Yule, 1900)

$$Q = \frac{ad - bc}{ad + bc}. \tag{8.6}$$

Παρατηρήστε ότι, χρησιμοποιώντας τη λογική που αναπτύχθηκε μέσα από το Παράδειγμα 8.4, είναι σαφές ότι $n_c = ad$ και $n_d = bc$, οπότε ο συντελεστής του Yule συμπίπτει με τον συντελεστή γ της σχέσης (8.5). Γενικά, ο συντελεστής του Yule υπερεκτιμά τον βαθμό συσχέτισης μεταξύ των δύο μεταβλητών (βλ. Sheskin, 2011) και, μάλιστα, οι Ott *et al.* (1992) αναφέρουν ότι, αν η απόλυτη τιμή του Q είναι 1, αυτό δεν σημαίνει ότι υπάρχει απόλυτη συσχέτιση ανάμεσα στις δύο μεταβλητές. Για παράδειγμα, αν μία από τις παρατηρηθείσες τιμές του Πίνακα 8.4 είναι 0, τότε ο συντελεστής του Yule θα είναι -1 ή $+1$, για αυτό σε αυτές τις περιπτώσεις θα πρέπει να είμαστε ιδιαιτέρως προσεκτικοί.

Στην περίπτωση όπου θέλουμε να ελέγξουμε την ύπαρξη ή μη της συσχέτισης των μεταβλητών X και Y, μπορούμε να διεξάγουμε τον έλεγχο της μηδενικής υπόθεσης $H_0 : \rho = 0$, έναντι μίας εκ των τριών εναλλακτικών:

(Α) $H_1 : \rho > 0$, δηλαδή της τάσης μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μεγάλες τιμές της άλλης, και αντίστροφα,

(Β) $H_1 : \rho < 0$, δηλαδή της τάσης μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μικρές τιμές της άλλης, και αντίστροφα,

(Γ) $H_1 : \rho \neq 0$, δηλαδή της τάσης είτε μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μεγάλες τιμές της άλλης, και αντίστροφα, είτε μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μικρές τιμές της άλλης, και αντίστροφα.

Για τα προβλήματα (Α), (Β) ή (Γ) εφαρμόζεται ένας ασυμπτωτικός έλεγχος (βλ. Ott *et al.*, 1992) χρησιμοποιώντας τη σ.σ.

$$Z = \frac{Q}{\sqrt{\frac{1}{4}(1 - Q^2)^2 \left[\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right]}}$$

όπου για μεγάλες τιμές του n και υπό τη μηδενική υπόθεση $H_0 : \rho = 0$, αποδεικνύεται ότι $Z \sim \mathcal{N}(0,1)$. Επομένως, χρησιμοποιώντας τα ποσοστιαία σημεία της τυπικής κανονικής κατανομής και σε επίπεδο σημαντικότητας a, $0 < a < 1$,

(Α) Απορρίπτεται η μηδενική υπόθεση, όταν $Z > z_a$.

(Β) Απορρίπτεται η μηδενική υπόθεση, όταν $Z \leqslant -z_a$.

(Γ) Απορρίπτεται η μηδενική υπόθεση, όταν $|Z| > z_{a/2}$.

Εδώ, σημειώνουμε ότι, ενώ ο συντελεστής συσχέτισης των Goodman and Kruskal χρησιμοποιείται στην περίπτωση διατάξιμων πινάκων συνάφειας, αντιθέτως, στην περίπτωση του συντελεστή συσχέτισης του Yule, η μη ύπαρξη διάταξης δεν περιορίζει την εφαρμογή του συντελεστή Q.

8.6 Έλεγχος ανεξαρτησίας ως έλεγχος συσχέτισης σε πίνακες συνάφειας

Ένα από τα πιο γνωστά αποτελέσματα της Πιθανοθεωρίας είναι αυτό σύμφωνα με το οποίο η ανεξαρτησία δύο τυχαίων μεταβλητών X και Y συνεπάγεται ότι οι μεταβλητές αυτές είναι και ασυσχέτιστες, ενώ το αντίθετο, γενικά, δεν συμβαίνει εκτός αν αυτές ακολουθούν κανονική κατανομή (βλ., μεταξύ άλλων, Gubner, 2006). Επομένως, ένας τρόπος για να δείξουμε ότι δύο μεταβλητές είναι ασυσχέτιστες είναι να δείξουμε ότι αυτές είναι ανεξάρτητες. Μία μεθοδολογία, που έχουμε ήδη αναφέρει, η οποία χρησιμοποιείται ως έλεγχος ανεξαρτησίας όταν τα δεδομένα ταξινομούνται σε πίνακες συνάφειας, είναι ο έλεγχος χι-τετράγωνο καλής προσαρμογής (βλ. Κεφάλαιο 4). Πριν προχωρήσουμε στην παρουσίαση της μεθοδολογίας, παραπέμπουμε τον/την αναγνώστη/στρια να θυμηθεί τον ορισμό της Πολυωνυμικής κατανομής, που δόθηκε στην εισαγωγική ενότητα στο Κεφάλαιο 4.

Θεωρούμε δύο κατηγορικές τυχαίες μεταβλητές A και B, με k κατηγορίες $A_1, \dots, A_k$, και s κατηγορίες $B_1, \dots, B_s$, αντίστοιχα. Αν X_{ij} είναι η τυχαία μεταβλητή που παριστάνει το πλήθος των ατόμων που ανήκουν στην κατηγορία A_i και B_j, τότε

$$\mathbf{X} = (X_{ij} : i = 1, \dots, k, j = 1, \dots, s) \sim \mathcal{M}(n; p_{ij}; i = 1, \dots, k, j = 1, \dots, s),$$

όπου $p_{ij} = P(A_i \cap B_j)$, για $i = 1, ..., k$, $j = 1, ..., s$. Επιπλέον, θεωρούμε τις πιθανότητες $p_i = P(A_i)$ και $q_j = P(B_j)$, για $i = 1, ..., k$, $j = 1, ..., s$. Από τον ορισμό της Πολυωνυμικής κατανομής έπεται ότι το πλήθος των δυνατών αποτελεσμάτων είναι $M + 1 = ks$. Τα χαρακτηριστικά A, B είναι ανεξάρτητα, αν και μόνο αν $p_{ij} = p_i q_j$, $\forall\ i, j$. Άρα, το πρόβλημα ελέγχου υποθέσεων που καλούμαστε να αντιμετωπίσουμε είναι της μορφής

$$H_0 : p_{ij} = p_i q_j\ ,\ \ \forall\ i, j\ \ \text{έναντι της}\ \ H_1 : p_{ij} \neq p_i q_j\ ,\ \ \text{για κάποιο}\ \ (i, j),\ \text{για}\ i = 1, ..., k, j = 1, ..., s.$$

Έστω n_{ij} είναι οι παρατηρηθείσες τιμές και e_{ij} είναι οι αναμενόμενες τιμές, υπό την H_0. Στο συγκεκριμένο μοντέλο, $e_{ij} = np_{ij} = np_i q_j$. Επειδή τα p_i και q_j είναι άγνωστα, εκτιμώνται χρησιμοποιώντας τους εκτιμητές μέγιστης πιθανοφάνειας (ΕΜΠ) $\hat{p}_i = \dfrac{n_{i\cdot}}{n}$ και $\hat{q}_j = \dfrac{n_{\cdot j}}{n}$, όπου $n_{i\cdot} = \sum_{j=1}^{s} n_{ij}$ και $n_{\cdot j} = \sum_{i=1}^{k} n_{ij}$. Επομένως, υπό την H_0, οι εκτιμώμενες αναμενόμενες τιμές είναι

$$e_{ij} = n\hat{p}_{ij} = n\hat{p}_i\hat{q}_j = \frac{n_{i\cdot}n_{\cdot j}}{n} \tag{8.7}$$

και η σ.σ. του ελέγχου καλής προσαρμογής (βλ. Κεφάλαιο 4), που χρησιμοποιείται για να ελέγξουμε την υπόθεση ανεξαρτησίας, λαμβάνει τη μορφή

$$X^2 = \sum_{j=1}^{s}\sum_{i=1}^{k} \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}} = \sum_{j=1}^{s}\sum_{i=1}^{k} \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}. \tag{8.8}$$

Η ασυμπτωτική κατανομή της X^2, υπό την H_0, είναι μία χι-τετράγωνο (Conover, 1998) με r βαθμούς ελευθερίας, όπου $r = ks - 1 -$(πλήθος ανεξάρτητων εκτιμώμενων παραμέτρων υπό την H_0) $= ks - 1 - (k - 1 + s - 1) = (k - 1)(s - 1)$. Τελικά, σε επίπεδο σημαντικότητας a, χρησιμοποιούμε τον έλεγχο,

$$\phi(\underline{x}) = \begin{cases} 1 & , \quad X^2 > \chi^2_{(k-1)(s-1),a} \\ 0 & , \quad X^2 \leqslant \chi^2_{(k-1)(s-1),a} \end{cases}$$

Παράδειγμα 8.5. Τυχαίο δείγμα 200 εργαζόμενων σε μεγάλη επιχείρηση ταξινομήθηκε στον παρακάτω πίνακα ανάλογα με τη θέση τους στην ιεραρχική κλίμακα της επιχείρησης και την άποψή τους για το προτεινόμενο σύστημα αμοιβής.

	Άποψη	
Βαθμίδα Ιεραρχίας	Θετική	Αρνητική
Ανώτερη	30	50
Μέση	30	45
Κατώτερη	20	25

Χρησιμοποιώντας χι-τετράγωνο έλεγχο καλής προσαρμογής διαπιστώστε αν η άποψη των εργαζόμενων για το προτεινόμενο σύστημα είναι ανεξάρτητη από τη θέση τους στην ιεραρχική κλίμακα, σε επίπεδο σημαντικότητας 10%.

Λύση Παραδείγματος 8.5. Ορίζουμε ως χαρακτηριστικό Α την ιεραρχική κλίμακα της επιχείρησης η οποία έχει τρεις κατηγορίες A_1 : Ανώτερη, A_2 : Μέση, A_3 : Κατώτερη και ως Β την άποψη των εργαζομένων για το προτεινόμενο σύστημα αμοιβής με δύο κατηγορίες B_1 : Θετική, B_2 : Αρνητική. Οπότε, αν X_{ij} η τυχαία μεταβλητή που παριστάνει το πλήθος των ατόμων που ανήκουν στην κατηγορία A_i, $i = 1,2,3$ και B_j, $j = 1,2$, τότε

$$\mathbf{X} = (X_{11}, X_{12}, X_{21}, X_{22}, X_{31}, X_{32}) \sim \mathcal{M}(n; p_{ij}; i = 1,2,3, j = 1,2),$$

όπου $p_{ij} = P(A_i \cap B_j)$, για $i = 1,...,k$, $j = 1,...,s$. Επίσης, αν $p_i = P(A_i)$ και $q_j = P(B_j)$, τότε το πρόβλημα ελέγχου υποθέσεων που καλούμαστε να αντιμετωπίσουμε είναι

$$H_0 : p_{ij} = p_i q_j \ , \ \forall \ i,j \text{ έναντι της } H_1 : p_{ij} \neq p_i q_j \ , \quad \text{για κάποιο } (i,j) \text{ για } i = 1,2,3, j = 1,2.$$

Οι αναμενόμενες τιμές μέσα σε κάθε κελί υπολογίζονται μέσω της σχέσης (8.7) και δίνονται στον παρακάτω πίνακα:

	Άποψη	
Βαθμίδα Ιεραρχίας	Θετική	Αρνητική
Ανώτερη	32	48
Μέση	30	45
Κατώτερη	18	27

Από τη σχέση (8.8) και με αντικατάσταση των τιμών n_{ij}, e_{ij}, η τιμή της είναι, μετά από αλγεβρικές πράξεις, $X^2 = 0.578$. Η τιμή αυτή συγκρίνεται με το 10% ποσοστιαίο σημείο της κατανομής χ^2 με $(k-1)(s-1) = (3-1)(2-1) = 2$ βαθμούς ελευθερίας. Από τον Πίνακα Π.3 του Παραρτήματος προκύπτει ότι $\chi^2_{2,0.10} = 4.605$, το οποίο είναι μεγαλύτερο του 0.578. Επομένως, δεν απορρίπτουμε την H_0. Άρα, η άποψη των εργαζομένων για το προτεινόμενο σύστημα είναι ανεξάρτητη από τη θέση τους στην ιεραρχική κλίμακα, σε επίπεδο σημαντικότητας 10%.

Από το παραπάνω συμπέρασμα προκύπτει ότι, αν η τ.μ. X μετράει την άποψη του εργαζόμενου για το προτεινόμενο σύστημα αμοιβής και η τ.μ. Y μετράει τη θέση του εργαζόμενου στην ιεραρχική κλίμακα, τότε οι τ.μ. X και Y είναι ανεξάρτητες, επομένως και ασυσχέτιστες. □

Στην περίπτωση που ο πίνακας συνάφειας είναι 2×2 (βλ. Πίνακα 8.4) μπορεί να δοθεί άμεσα η μορφή τόσο της στατιστικής συνάρτησης ελέγχου χ^2 όσο και του ελέγχου για το πρόβλημα ανεξαρτησίας. Πράγματι, χρησιμοποιώντας τη σχέση (8.8), προκύπτει ότι

$$X^2 = \frac{\left(a - \frac{(a+c)(a+b)}{n}\right)^2}{\frac{(a+c)(a+b)}{n}} + \frac{\left(b - \frac{(a+b)(b+d)}{n}\right)^2}{\frac{(a+b)(b+d)}{n}} + \frac{\left(c - \frac{(a+c)(c+d)}{n}\right)^2}{\frac{(a+c)(c+d)}{n}} + \frac{\left(d - \frac{(b+d)(c+d)}{n}\right)^2}{\frac{(b+d)(c+d)}{n}}$$

ή, διαφορετικά,

$$X^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \tag{8.9}$$

και ο έλεγχος για το πρόβλημα ανεξαρτησίας σε πίνακα συνάφειας 2×2 γίνεται

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & , \quad X^2 > \chi^2_{1,a} \\ 0 & , \quad X^2 \leqslant \chi^2_{1,a} \end{cases}$$

Παράδειγμα 8.6. Τυχαίο δείγμα 1000 ατόμων ταξινομήθηκε στον παρακάτω πίνακα ανάλογα με το φύλο του (Α=αγόρι, Κ=κορίτσι) και το αν πάσχει ή όχι από αχρωματοψία.

	Αχρωματοψία	
Φύλο	Όχι	ναι
Αγόρι	442	38
Κορίτσι	480	40

Χρησιμοποιώντας έναν χι-τετράγωνο έλεγχο καλής προσαρμογής διαπιστώστε κατά πόσο τα χαρακτηριστικά «Φύλο» και «Αχρωματοψία» είναι ανεξάρτητα, σε επίπεδο σημαντικότητας 5%.

Λύση Παραδείγματος 8.6. Ορίζουμε ως χαρακτηριστικό Α το Φύλο που έχει δύο κατηγορίες A_1 : Αγόρι, A_2 : Κορίτσι και ως χαρακτηριστικό Β την ύπαρξη Αχρωματοψίας ή όχι, με B_1 : Οχι, B_2 : Ναι. Οπότε, αν X_{ij} η τυχαία μεταβλητή που παριστάνει το πλήθος των ατόμων που ανήκουν στην κατηγορία A_i, $i = 1,2$ και B_j, $j = 1,2$ τότε

$$\boldsymbol{X} = (X_{11}, X_{12}, X_{21}, X_{22}) \sim \mathcal{M}(n; p_{11}, p_{12}, p_{21}, p_{22}),$$

όπου $p_{ij} = P(A_i \cap B_j)$. Επίσης, αν $p_i = P(A_i)$ και $q_j = P(B_j)$, τότε το πρόβλημα ελέγχου υποθέσεων που καλούμαστε να αντιμετωπίσουμε είναι

$$H_0 : p_{ij} = p_i q_j \ , \ \ \forall \ i,j \text{ έναντι της } \ H_1 : p_{ij} \neq p_i q_j \ , \quad \text{για κάποιο } \ (i,j), \text{ με } i = 1,2, j = 1,2.$$

Χρησιμοποιώντας τη σχέση (8.9), υπολογίζουμε ότι $X^2 = 0.017$. Από τον Πίνακα Π.3 του Παραρτήματος προκύπτει ότι $\chi^2_{1,0.05} = 3.841$, το οποίο είναι μεγαλύτερο του 0.017, επομένως αποδεχόμαστε την H_0. Άρα, τα χαρακτηριστικά «Φύλο» και «Αχρωματοψία» είναι ανεξάρτητα, σε επίπεδο σημαντικότητας 5%. □

Επειδή στην πραγματικότητα, αυτός ο έλεγχος καλής προσαρμογής χρησιμοποιεί μία συνεχή κατανομή για να προσεγγίσει μία διακριτή κατανομή (κατά βάση τα δεδομένα μας είναι Πολυωνυμικά), συνήθως, χρησιμοποιείται μία διόρθωση συνεχείας (Kendall and Stuart, 1961), η οποία αποτελεί αντικείμενο μελέτης της επόμενης υποενότητας.

8.6.1 Διόρθωση Yates

Η διόρθωση συνεχείας βασίζεται στο γεγονός ότι, αν μια συνεχής κατανομή χρησιμοποιείται για να προσεγγίσει μια διακριτή κατανομή, μια τέτοια προσέγγιση ενδέχεται να μεγαλώσει την πιθανότητα σφάλματος τύπου Ι. Χρησιμοποιώντας τη διόρθωση, το σφάλμα τύπου Ι θα είναι πιο συμβατό με το επιθυμητό επίπεδο σημαντικότητας (Sheskin, 2011). Συνήθως, η διόρθωση συνεχείας που προτείνεται στην περίπτωση του χ^2 ελέγχου καλής προσαρμογής είναι αυτή του Yates (1934), η οποία προτάθηκε για 2×2 πίνακα συνάφειας και στη γενική της μορφή είναι,

$$X^2 = \sum_{j=1}^{s} \sum_{i=1}^{k} \frac{\left(|n_{ij} - e_{ij}| - 0.5\right)^2}{e_{ij}} \tag{8.10}$$

Σημειώνεται ότι η τιμή της X^2 της σχέσης (8.10) είναι σαφώς μικρότερη από αυτήν της σχέσης (8.8) και, επομένως, όταν αποδεχόμαστε την H_0 (βλ. Παραδείγματα 8.5 και 8.9) η διόρθωση Yates δεν χρειάζεται. Σε γενικές γραμμές, προτείνεται να χρησιμοποιείται η διόρθωση Yates όταν η αναμενόμενη τιμή σε κάποιο κελί είναι το πολύ 1 ή πάνω από το 20% των κελιών έχουν αναμενόμενη τιμή < 5 (Cochran, 1954). Στη βιβλιογραφία, πάντως, υπάρχει διαφωνία αν τελικά αυτή η διόρθωση είναι προτέρημα, καθώς τελικά ο έλεγχος καθίσταται συντηρητικός (βλ., για παράδειγμα την εργασία των Storer and Kim, 1990, και τις εκεί αναφορές).

Παράδειγμα 8.7. 60 πιλότοι, υποψήφιοι για κάποιες θέσεις σε μια αεροπορική εταιρεία, υποβάλλονται σε ένα ψυχολογικό τεστ σύμφωνα με το οποίο ταξινομούνται ως εσωστρεφείς ή εξωστρεφείς, και σε ένα τεστ δεξιότητας στο οποίο βαθμολογούνται ως επιτυχόντες ή αποτυχόντες. Από τους 36 επιτυχόντες, μόνο οι 3 θεωρήθηκαν εσωστρεφείς, ενώ 17 αποτυχόντες θεωρήθηκαν εξωστρεφείς. Θα ήταν εύλογο να υποθέσει κανείς ότι τα δεδομένα υπαινίσσονται ύπαρξη συσχέτισης μεταξύ του τύπου της προσωπικότητας και των δεξιοτήτων των πιλότων; Να γίνει ο σχετικός έλεγχος σε ε.σ. 5%.

Λύση Παραδείγματος 8.7. Ορίζουμε ως χαρακτηριστικό Α το αποτέλεσμα του ψυχολογικού τεστ που έχει δύο κατηγορίες A_1 : Εσωστρεφής, A_2 : Εξωστρεφής και ως χαρακτηριστικό Β το αποτέλεσμα του τεστ δεξιότητας, με κατηγορίες B_1 : Επιτυχών, B_2 : Αποτυχών. Οπότε, αν X_{ij} η τυχαία μεταβλητή που παριστάνει το πλήθος των ατόμων που ανήκουν στην κατηγορία A_i, $i = 1, 2$ και B_j, $j = 1, 2$ τότε $\boldsymbol{X} = (X_{11}, X_{12}, X_{21}, X_{22}) \sim \mathcal{M}(n; p_{11}, p_{12}, p_{21}, p_{22})$, όπου $p_{ij} = P(A_i \cap B_j)$. Επίσης, αν $p_i = P(A_i)$ και $q_j = P(B_j)$, τότε το πρόβλημα ελέγχου υποθέσεων που καλούμαστε να αντιμετωπίσουμε είναι

$$H_0 : p_{ij} = p_i q_j \ , \ \forall\ i,j \ \ vs. \ \ H_1 : p_{ij} \neq p_i q_j \ , \quad \text{για κάποιο } (i,j), \text{ με } i = 1,2, j = 1,2.$$

Από τα δεδομένα του Παραδείγματος προκύπτει ο παρακάτω 2×2 πίνακας συνάφειας,

	Επιτυχόντες	Αποτυχόντες
Εσωστρεφείς	3	7
Εξωστρεφείς	33	17

Οι αναμενόμενες τιμές μέσα σε κάθε κελί υπολογίζονται μέσω της σχέσης (8.7) και δίνονται στον παρακάτω πίνακα,

	Επιτυχόντες	Αποτυχόντες
Εσωστρεφείς	6	4
Εξωστρεφείς	30	20

Υπολογίζουμε το στατιστικό του ελέγχου ως

$$X^2 = \frac{(3-6)^2}{6} + \frac{(7-4)^2}{4} + \frac{(33-30)^2}{30} + \frac{(17-20)^2}{20} = 4.5.$$

Από τον Πίνακα Π.3 του Παραρτήματος προκύπτει ότι $\chi^2_{1,0.05} = 3.841$, το οποίο είναι μικρότερο του 4.5, επομένως απορρίπτουμε την H_0. Επειδή ένα από τα 4 κελιά του πίνακα έχει αναμενόμενη τιμή 4 (< 5) και η μηδενική υπόθεση έχει απορριφθεί, πρέπει να κάνουμε διόρθωση Yates, οπότε υπολογίζεται εκ νέου η σ.σ. X^2 χρησιμοποιώντας τη σχέση (8.10). Έτσι έχουμε ότι

$$X^2 = \frac{(|3-6|-0.5)^2}{6} + \frac{(|7-4|-0.5)^2}{4} + \frac{(|33-30|-0.5)^2}{30} + \frac{(|17-20|-0.5)^2}{20} = 3.125.$$

Αφού $\chi^2_{1,0.05} = 3.841$, το οποίο είναι μεγαλύτερο του 3.125, προκύπτει ότι τώρα αλλάζουμε απόφαση και αποδεχόμαστε την H_0. Αυτό ακριβώς είναι ένα πρόβλημα του X^2 ελέγχου καλής προσαρμογής και σε αυτές τις περιπτώσεις είναι προτιμότερο να ακολουθήσουμε μία άλλη διαδικασία για έλεγχο ανεξαρτησίας, όπως, π.χ. τον ακριβή έλεγχο του Fisher ή τον Ασυμπτωτικό έλεγχο λόγου πιθανοφανειών (ΕΛΠ)). Για περισσότερες λεπτομέρειες παραπέμπουμε, μεταξύ άλλων, στο σύγγραμμα Hoffman (2015). □

Παρατήρηση 8.10. Πολλοί συγγραφείς συμφωνούν ότι όταν το μέγεθος του δείγματος είναι αρκετά μικρό (συνήθως μικρότερο του 20), τότε πρέπει να αποφεύγεται η χρήση του χ^2 ελέγχου καλής προσαρμογής για πίνακα συνάφειας 2×2. Ο Cochran (1952; 1954), επιπλέον, αναφέρει ότι για $20 < n < 40$, ο έλεγχος μπορεί να πραγματοποιείται, όταν η αναμενόμενη τιμή σε κάθε κελί είναι μεγαλύτερη από 5 και για $n > 40$, όταν όλα τα κελιά έχουν αναμενόμενες τιμές μεγαλύτερες ή ίσες της μονάδας.

8.7 Ασκήσεις

Άσκηση 8.1. Ρωτήσαμε 10 άτομα που πηγαίνουν γυμναστήριο, πόσες ώρες αθλούνται τον μήνα και πόσα φρούτα καταναλώνουν (κάθε μήνα). Οι απαντήσεις που συλλέξαμε, παρουσιάζονται στον παρακάτω πίνακα.

Ώρες άθλησης	12	15	24	18	30	32	17	27	42	8
Κατανάλωση φρούτων	22	28	30	26	26	48	30	32	58	15

Χρησιμοποιώντας τον συντέλεστη συσχέτισης του Pearson, ελέγξτε σε ε.σ. 5%, αν υπάρχει συσχέτιση ανάμεσα στις ώρες που γυμνάζεται ένα άτομο και στην κατανάλωση φρούτων από αυτό.

Άσκηση 8.2. Δέκα παιδιά της Ε' Δημοτικού επιλέχθηκαν τυχαία και έδωσαν ένα τεστ αριθμητικής και ένα γλώσσας, με τα παρακάτω αποτελέσματα (σε κλίμακα βαθμολόγησης 0-100).

Παιδί	1	2	3	4	5	6	7	8	9	10
Τεστ Γλώσσας	26	31	42	46	54	54	61	62	66	68
Τεστ Αριθμητικής	21	32	28	29	44	38	37	48	43	46

Ελέγξτε, σε ε.σ. 5%, αν υπάρχει θετική συσχέτιση ανάμεσα στις μεταβλητές, χρησιμοποιώντας:

(α) τον συντέλεστη συσχέτισης του Pearson,
(β) τον συντέλεστη συσχέτισης του Kendall.

Άσκηση 8.3. Το Υπουργείο Περιβάλλοντος του Ηνωμένου Βασιλείου διεξήγαγε μια έρευνα ανάμεσα στις χρονιές 1986 και 1987, για να διαπιστώσει την επίδραση του μολύβδου, που προέρχεται από την εκπομπή καυσαερίων από τους κινητήρες των αυτοκινήτων, στους πολίτες. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα.

Ηλικία	Συγκέντρωση μολύβδου στο αίμα (mg/100ml)
$18-20$	4.9
$21-25$	5.1
$26-30$	5.2
$31-35$	5.3
$36-40$	5.6
$41-45$	5.7
$46-50$	6.1
$51-55$	7.5
$56-60$	7.4
$61-64$	7.5

Χρησιμοποιώντας τον συντέλεστη συσχέτισης του Spearman ελέγξτε, σε ε.σ. 10%, αν υπάρχει θετική συσχέτιση ανάμεσα στις μεταβλητές ηλικία και συγκέντρωση μολύβδου στο αίμα.

Άσκηση 8.4. Μία έρευνα που έγινε από μια εταιρεία είχε σκοπό να διαπιστώσει αν η ανάπτυξη της αγοράς και η τοποθεσία στην οποία βρισκόταν η εταιρεία, ήταν στατιστικά ανεξάρτητες. Χρησιμοποιώντας έναν χι-τετράγωνο έλεγχο καλής προσαρμογής, τι συμπεραίνετε, σε ε.σ. 5%, με βάση τα παρακάτω δεδομένα;

	Τοποθεσία		
Ανάπτυξη	Αστική	Ημιαστική	Αγροτική
Μεγάλη	125	210	65
Μεσαία	100	180	70
Μηδενική	75	110	65

Άσκηση 8.5. Τυχαίο δείγμα 100 ενηλίκων ανδρών ταξινομήθηκαν στον παρακάτω πίνακα ανάλογα με το δικό τους επίπεδο εκπαίδευσης και αυτό του πατέρα τους.

	Επίπεδο Εκπαίδευσης Πατέρα		
Επίπεδο Εκπαίδευσης Παιδιού	Πρωτοβάθμια	Δευτεροβάθμια	Τριτοβάθμια
Πρωτοβάθμια	12	5	3
Δευτεροβάθμια	10	9	11
Τριτοβάθμια	9	24	17

(α) Χρησιμοποιώντας τον συντελεστή συσχέτισης των Goodman and Kruskal ελέγξτε, σε ε.σ. 5%, αν υπάρχει συσχέτιση ανάμεσα στα επίπεδα εκπαίδευσης πατέρα και γιου.

(β) Χρησιμοποιώντας έναν χι-τετράγωνο έλεγχο καλής προσαρμογής, να ελεγχθεί, σε ε.σ. 10%, η υπόθεση ότι τα επίπεδα εκπαίδευσης πατέρα και γιου είναι ανεξάρτητα.

Άσκηση 8.6. Τυχαίο δείγμα 120 ατόμων ταξινομήθηκε στον παρακάτω πίνακα ανάλογα με το φύλο του και αν καπνίζει ή όχι.

	Καπνιστής	**Μη καπνιστής**
Άνδρας	66	4
Γυναίκα	38	12

Χρησιμοποιώντας έναν χι-τετράγωνο έλεγχο καλής προσαρμογής, ελέγξτε, σε ε.σ. 1%, αν υπάρχει σχέση ανάμεσα στο φύλο και στο αν ένα άτομο καπνίζει ή όχι.

Άσκηση 8.7. Μια εβδομάδα πριν τις πρόσφατες ευρωεκλογές διεξήχθη μία δημοσκοπική έρευνα όπου, μεταξύ άλλων, τα άτομα που μετείχαν σε αυτήν ρωτήθηκαν αν πρόκειται να συμμετάσχουν σε αυτές τις εκλογές. Τα αποτελέσματα που προέκυψαν ανά ηλικιακή κατηγορία εμφανίζονται στον πίνακα που ακολουθεί.

	Ηλικιακές Κατηγορίες				
	18 – 24	25 – 34	35 – 44	45 – 54	55+
Ναι	8	38	22	76	36
Όχι	12	52	18	34	4

Χρησιμοποιώντας έναν χι-τετράγωνο έλεγχο καλής προσαρμογής, ελέγξτε, σε ε.σ. 5%, αν υπάρχει κάποια σχέση ανάμεσα στην ηλικία των ατόμων και στο εάν θα πάνε να ψηφίσουν ή όχι.

Άσκηση 8.8. Τυχαίο δείγμα 80 ατόμων ταξινομήθηκε στον παρακάτω πίνακα ανάλογα με το φύλο του και αν συμφωνεί ή όχι με μία κυβερνητική απόφαση.

	Απάντηση	
Φύλο	Ναι	Όχι
Άνδρας	6	32
Γυναίκα	2	40

(α) Χρησιμοποιώντας τον συντελεστή συσχέτισης του Yule ελέγξτε, σε ε.σ. 5%, αν υπάρχει συσχέτιση ανάμεσα στο φύλο και την απάντηση που έδωσαν οι πολίτες.

(β) Χρησιμοποιώντας έναν χι-τετράγωνο έλεγχο καλής προσαρμογής, ελέγξτε την υπόθεση ότι το φύλο είναι ανεξάρτητο της απάντησης που έδωσαν οι πολίτες, σε ε.σ. 1%. Για να δώσετε την τελική σας απάντηση, χρειάζεται να κάνετε τη διόρθωση Yates; Να αιτιολογήσετε την απάντησή σας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Κούτρας, Μ. (2018). *Εισαγωγή στη Θεωρία Πιθανοτήτων και Εφαρμογές*. Αθήνα: Εκδόσεις Σταμούλη.
Χατζηνικολάου, Δ. (2002). *Στατιστική για οικονομολόγους*. Εκδόσεις Κιόρογλου Λαμπρινή.

Ξενόγλωσση

Cochran, W. G. (1952). The χ^2 Test of Goodness of Fit. *The Annals of Mathematical Statistics*, *23*(3), pp. 315–345.
Cochran, W. G. (1954). Some Methods for Strengthening the Common χ2 Tests. *Biometrics*, *10*(4), pp. 417–451.
Conover, W. J. (1998). *Practical Nonparametric Statistics* (3rd ed.). Wiley and Sons, Inc.
Conover, W., Johnson, M. E. and Johnson, M. M. (1981). A Comparative Study of Tests for Homogeneity of Variances with Applications to the Outer Continental Shelf Bidding Data. *Technometrics*, *23*, pp. 351–361.
Fisher, R. (1921). On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample. *Metron.*, *1*, pp. 3–32.
Gibbons, J. D. and Chakraborti, S. (2020). *Nonparametric Statistical Inference, Fourth Edition Revised and Expanded*. Chapman and Hall/CRC.
Goodman, L. and Kruskal, W. (1954). Measures of Association for Cross Classifications*. *Journal of the American Statistical Association*, *49*(268), pp. 732–764.
Goodman, L. and Kruskal, W. (1959). Measures of Association for Cross Classifications. II: Further Discussion and References. *Journal of the American Statistical Association*, *54*(285), pp. 123–163.
Goodman, L. and Kruskal, W. (1963). Measures of Association for Cross Classifications III: Approximate Sampling Theory. *Journal of the American Statistical Association*, *58*(302), pp. 310–364.
Goodman, L. and Kruskal, W. (1972). Measures of Association for Cross Classifications, IV: Simplification of Asymptotic Variances. *Journal of the American Statistical Association*, *67*(338), pp. 415–421.
Gubner, J. A. (2006). *Probability and Random Processes for Electrical and Computer Engineers*. Cambridge University Press.
Hoffman, J. I. (2015). *Biostatistics for Medical and Biomedical Practitioners*. Academic Press.
Hogg, R., McKean, J. and Craig, A. (2013). *Introduction to Mathematical Statistics. 7th ed.* Boston: Pearson.
Hotelling, H. and Pabst, M. (1936). Rank correlation and tests of significance involving the assumption of normality. *Annals of Mathematical Statistics*, *7*, pp. 29–43.
Kendall, M. and Gibbons, J. (1990). *Rank Correlation Methods* (5th ed.). Edward Arnold, London.
Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika*, *30*, pp. 81–93.
Kendall, M. and Stuart, A. (1961). *The Advanced Theory of Statistics: Vol. 2 - Inference and Relationship*. Hafner Publishing Company, London.
Lehmann, E. L. (2004). *Elements of large-sample theory.* Springer, New York, NY.

Ott, R. L., Larson, R., Rexroat, C. and Mendenhall, W. (1992). *Statistics: A tool for the social sciences* (5th ed.). PWS–Kent Publishing Company, Boston.

Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. *Philosophical Transactions of the Royal Society of London*, *187*, pp. 253–318.

Puth, M.-T., Neuhäuser, M. and Ruxton, G. D. (2015). Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. *Animal Behaviour*, *102*, pp. 77–84.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches.* Hogrefe Publishing.

Sheskin, D. (2011). *Handbook of Parametric and Non-parametric Procedures* (5th ed.). Chapman and Hall/CRC.

Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. *The American Journal of Psychology*, *15*, pp. 201–292.

Sprent, P. and Smeeton, N. (2007). *Applied Nonparametric Statistical Methods* (4th ed.). Chapmann and Hall.

Storer, B. E. and Kim, C. (1990). Exact Properties of Some Exact Test Statistics for Comparing Two Binomial Proportions. *Journal of the American Statistical Association*, *85*(409), pp. 146–155.

Trosset, M. (2009). *An Introduction to Statistical Inference and Its Applications with R*. Chapman and Hall/CRC.

Welz, T., Doebler, P. and Pauly, M. (2021). Fisher transformation based confidence intervals of correlations in fixed- and random-effects meta-analysis. *The British Journal of Mathematical and Statistical Psychology*, *75*, pp. 1–22.

Yates, F. (1934). Contingency Tables Involving Small Numbers and the χ2 Test. *Supplement to the Journal of the Royal Statistical Society*, *1*(2), pp. 217–235.

Yu, H. and Hutson, A. D. (2022). A Robust Spearman Correlation Coefficient Permutation Test. *Communications in Statistics - Theory and Methods*, *DOI: 10.1080/03610926.2022.2121144*.

Yule, G. (1900). On the association of attributes in statistics: with illustrations from the material of the childhood society, &c. *Philosophical Transactions of the Royal Society of London. Series A*, *194*, pp. 257–319.

ΚΕΦΑΛΑΙΟ 9

ΜΗ ΠΑΡΑΜΕΤΡΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Σύνοψη

Το παρόν κεφάλαιο αποτελεί μία εισαγωγή στο πεδίο της μη παραμετρικής παλινδρόμησης. Δίνεται έμφαση στην περίπτωση μιας επεξηγηματικής μεταβλητής και παρουσιάζονται οι τεχνικές εκτίμησης της συνάρτησης παλινδρόμησης μέσω του παλινδρογράμματος (regressogram), του εκτιμητή Nadaraya-Watson, των τοπικών πολυωνύμων (local polynomial regression) και των εξομαλυντών spline. Καθώς οι τεχνικές αυτές εξαρτώνται από κάποια παράμετρο εξομάλυνσης, θα φανούν ιδιαίτερα χρήσιμες στον/στην αναγνώστη/στρια οι έννοιες που εισήχθησαν στο Κεφάλαιο 3 του παρόντος συγγράμματος. Γίνεται, επίσης, μια σύντομη αναφορά στα προσθετικά μοντέλα (additive models) και στα δέντρα παλινδρόμησης (regression trees) για την περίπτωση περισσότερων επεξηγηματικών μεταβλητών.

Προαπαιτούμενη γνώση:
Βασικές γνώσεις Παλινδρόμησης και Εκτιμητικής.
Κεφάλαιο 3 του παρόντος συγγράμματος.

Προσδοκώμενα μαθησιακά αποτελέσματα:
Ο/η φοιτητής/τρια ολοκληρώνοντας την ενότητα αυτή θα μπορεί να εφαρμόζει βασικές μεθόδους Μη Παραμετρικής Παλινδρόμησης.

Γλωσσάριο επιστημονικών όρων

- Ανθεκτική γραμμική παλινδρόμηση
- Απλοϊκός εκτιμητής παλινδρόμησης
- Βαθμοί ελευθερίας εξομαλυντή
- Γραμμικός εξομαλυντής
- Δέντρα παλινδρόμησης
- Εκτιμητής Nadaraya-Watson
- Εκτιμητής τοπικών μέσων
- Εξομαλυντής spline
- Παλινδρόγραμμα
- Ποινικοποιημένη παλινδρόμηση
- Προσθετικά μοντέλα
- Τοπικά πολυώνυμα

Μπατσίδης, Α., Παπασταμούλης, Π., Πετρόπουλος, Κ., & Ρακιτζής, Α. (2022). *Μη Παραμετρική Στατιστική.* [Προπτυχιακό εγχειρίδιο].

9.1 Εισαγωγή

Οι μέθοδοι παλινδρόμησης έχουν ως κύριο στόχο τη μοντελοποίηση της σχέσης μεταξύ μίας εξαρτημένης μεταβλητής Y και μίας ή περισσότερων ανεξάρτητων μεταβλητών X_1, X_2,...,X_k, με απώτερο σκοπό την πρόβλεψη της τιμής της Y όταν είναι γνωστή η τιμή των X_1, X_2,...,X_k. Εν γένει, υπάρχουν τρεις τύποι παλινδρόμησης: η γραμμική, η μη γραμμική παραμετρική και η μη παραμετρική. Στους δύο πρώτους τύπους παλινδρόμησης, το μοντέλο προσδιορίζεται από τη σχέση:

$$Y_i = f(x_{i1}, x_{i2}, \ldots, x_{ik}; \beta_0, \beta_1, \beta_2, \ldots, \beta_k) + \epsilon_i,$$

όπου υποθέτουμε ότι τα σφάλματα ϵ_i ακολουθούν κανονική κατανομή με μέση τιμή 0 και σταθερή διακύμανση, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$ είναι το διάνυσμα των άγνωστων παραμέτρων που θα εκτιμηθούν, $\boldsymbol{x} = (x_{i1}, x_{i2}, \ldots, x_{ik})$ είναι το διάνυσμα των ανεξάρτητων (επεξηγηματικών) μεταβλητών για την i-οστή παρατήρηση, ενώ y_i είναι η μέτρηση της εξαρτημένης μεταβλητής στην i-οστή παρατήρηση, $i = 1, \ldots, n$. Στην περίπτωση της γραμμικής (αντίστοιχα μη γραμμικής) παλινδρόμησης, η συνάρτηση f που ορίζει την προκαθορισμένη σχέση μεταξύ της μέσης τιμής της εξαρτημένης μεταβλητής Y και των επεξηγηματικών μεταβλητών $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ είναι γνωστής μορφής και συγκεκριμένα γραμμική (αντίστοιχα μη γραμμική).

Όμως, στις περιπτώσεις που οι παραπάνω υποθέσεις είτε παραβιάζονται είτε είναι μη ρεαλιστικές, πρέπει να υιοθετηθεί ένα άλλο μοντέλο για την περιγραφή της σχέσης μεταξύ εξαρτημένης μεταβλητής και των επεξηγηματικών μεταβλητών και, συγκεκριμένα, το γενικό μη παραμετρικό μοντέλο παλινδρόμησης, το οποίο δίνεται από τη σχέση:

$$Y_i = f(x_{i1}, x_{i2}, \ldots, x_{ip}) + \epsilon_i. \tag{9.1}$$

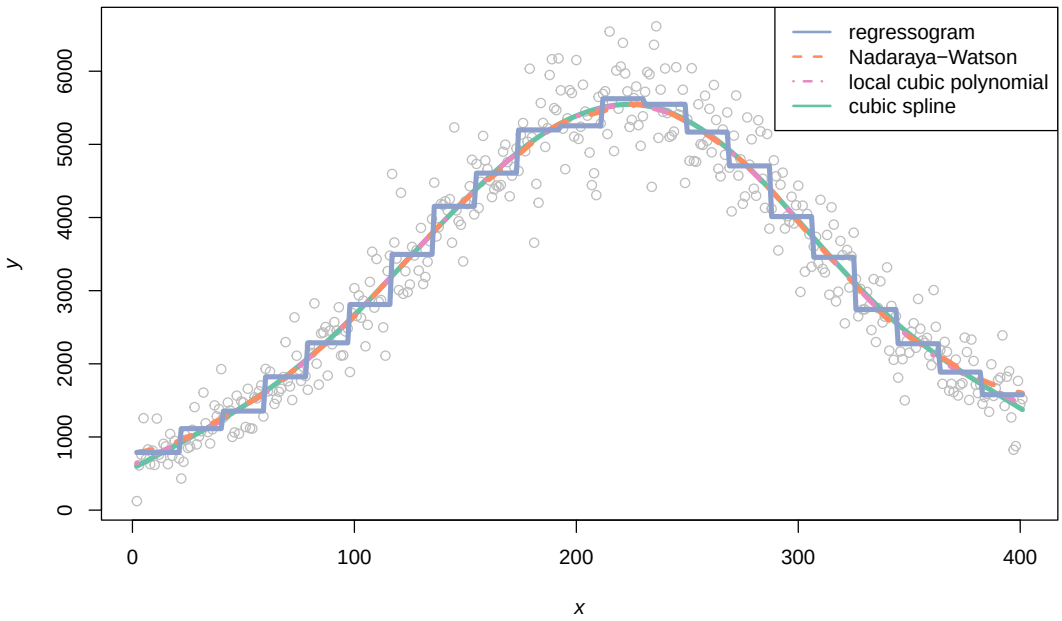
Έτσι, υπό το γενικό μη παραμετρικό μοντέλο παλινδρόμησης, ο στόχος είναι να εκτιμηθεί η άγνωστη συνάρτηση f, χωρίς να υποθέσουμε ότι η f είναι γνωστής μορφής, π.χ. γραμμική. Η περίπτωση που θα επικεντρωθούμε είναι αυτή του απλού μοντέλου παλινδρόμησης:

$$Y_i = f(x_i) + \epsilon_i, \tag{9.2}$$

όπου $f(\cdot)$ είναι μια άγνωστη συνάρτηση, η οποία είναι η βασική προς εκτίμηση ποσότητα. Οι περισσότερες τεχνικές μη παραμετρικής παλινδρόμησης υποθέτουν ότι η άγνωστη f είναι ομαλή και συνεχής καμπύλη. Η περίπτωση του πολλαπλού μοντέλου παλινδρόμησης θα συζητηθεί εν συντομία στην Ενότητα 9.7.

Η περιοχή της μη παραμετρικής παλινδρόμησης είναι πολύ μεγάλη για να μελετηθεί επαρκώς σε ένα κεφάλαιο. Εδώ, θα αρκεστούμε να παρουσιάσουμε τμήμα των διαθέσιμων τεχνικών, ώστε ο/η αναγνώστης/στρια να πάρει μία πρώτη βασική ιδέα. Δεν θα αναφερθούμε καθόλου σε σημαντικά ζητήματα, όπως στην εκτίμηση της διασποράς των σφαλμάτων (βλ. όμως την ΄Ασκηση 9.6) και στην κατασκευή ζωνών εμπιστοσύνης, για τα οποία παραπέμπουμε τον/την αναγνώστη/αναγνώστρια στα συγγράμματα των Wasserman (2006, Κεφάλαιο 5), Hastie and Tibshirani (2017, Κεφάλαιο 3) και Loader (2012). Αναγνώστες/στριες που ενδιαφέρονται να μελετήσουν πιο ολοκληρωμένα τις τεχνικές της μη παραμετρικής παλινδρόμησης παραπέμπονται στα συγγράμματα των Fan and Gijbels (2018), Härdle (1990), Loader (2006), Hastie and Tibshirani (2017) και Green and Silverman (2019) .

Παράδειγμα 9.1. Η κοσμική ακτινοβολία υποβάθρου (Cosmic Microwave Background (CMB) radiation) είναι το ίχνος ή το υπόλειμμα της ακτινοβολίας που εξέπεμπε το σύμπαν κατά την αρχή της δημιουργίας του (Big Bang). Στην επιστήμη της κοσμολογίας παρουσιάζει ενδιαφέρον η μελέτη της έντασης (y) της μεταβλητότητας της θερμοκρασίας ανά συχνότητα (x) (Genovese *et al.*, 2004). Ένα σύνολο δεδομένων που αποτελείται από $n = 400$ το πλήθος παρατηρήσεις[1] αυτού του είδους παρατίθεται στο Σχήμα 9.1. Σκοπός του παραδείγματος είναι να δώσει μια πρώτη γεύση της προσαρμογής μη παραμετρικών μοντέλων παλινδρόμησης.



Σχήμα 9.1: CMB data (Genovese *et al.*, 2004): Παρατηρηθέντα δεδομένα μαζί με τέσσερις μη παραμετρικές εκτιμήσεις παλινδρόμησης: παλινδρόγραμμα (regressogram), εκτιμητής Nadaraya-Watson, τοπικό πολυώνυμο βαθμού 3 και cubic spline.

Λύση Παραδείγματος 9.1. Ξεκάθαρα, τα παρατηρήθεντα δεδομένα $(y_1, x_1), \ldots, (y_n, x_n)$ δεν υπακούουν στις συνήθεις υποθέσεις ενός μοντέλου γραμμικής παλινδρόμησης. Στο Σχήμα 9.1 παρατίθενται τέσσερις εκτιμητές με τις μη παραμετρικές τεχνικές παλινδρόμησης: παλινδρόγραμμα (regressogram), εκτιμητής Nadaraya-Watson, χρήση τοπικού πολυωνύμου και smoothing splines, οι οποίες τεχνικές εκτίμησης θα παρουσιαστούν στη συνέχεια του παρόντος κεφαλαίου. Παρατηρήστε ότι οι εκτιμήσεις της συνάρτησης παλινδρόμησης μοιάζουν αρκετά μεταξύ τους, με την τεχνική του παλινδρογράμματος να διαφοροποιείται πιο έντονα από τις υπόλοιπες λόγω των χαρακτηριστικών σημείων ασυνέχειας της εκτιμηθείσας συνάρτησης. □

9.2 Το παλινδρόγραμμα

Το παλινδρόγραμμα[2] βασίζεται στην υιοθέτηση του ιστογράμματος στο πλαίσιο της παλινδρόμησης. Πρόκειται για μια απλή μέθοδο η οποία γενικά υπολείπεται των τεχνικών που θα παρουσιαστούν στη συνέχεια, αλλά είναι χρήσιμη για την εισαγωγή των βασικών χαρακτηριστικών των μη παραμετρικών εκτιμητών παλινδρόμησης.

Έστω ότι $a \leqslant x_i \leqslant b$, $i = 1, \ldots, n$. Θεωρούμε μια διαμέριση του διαστήματος (a, b) σε m διαστήματα $B_1, \ldots, B_m$ (ίσου) μήκους $h = \frac{b-a}{m}$.

[1] Χρησιμοποιούμε τις 400 πρώτες το πλήθος παρατηρήσεις από το σύνολο των δεδομένων των Genovese *et al.* (2004), η πλήρης μορφή του οποίου αποτελείται από $n = 899$ παρατηρήσεις και το οποίο είναι διαθέσιμο στον σύνδεσμο http://www.stat.cmu.edu/~larry/all-of-nonpar/=data/wmap.dat.

[2] Ως παλινδρόγραμμα αποδίδεται η μετάφραση του όρου regressogram που χρησιμοποίησε ο Tukey (1961) (regressogram = regression + histogram).

Ορισμός 9.1

Το **παλινδρόγραμμα** ορίζεται ως

$$\hat{f}(x) = \frac{1}{k_j} \sum_{i: x_i \in B_j} Y_i, \quad x \in B_j, \tag{9.3}$$

όπου $k_j = \sum_{i=1}^{n} I(x_i \in B_j)$.

Από τον Ορισμό 9.2 έπεται ότι το παλινδρόγραμμα δεν είναι τίποτε άλλο παρά ο δειγματικός μέσος των τιμών της μεταβλητής απόκρισης που αντιστοιχούν σε κάθε διάστημα της διαμέρισης. Από αυτήν την παρατήρηση εξάγουμε το συμπέρασμα ότι η προκύπτουσα εκτίμηση της $f(\cdot)$ είναι μία κλιμακωτή (άρα μη συνεχής) συνάρτηση. Η μπλε γραμμή του Σχήματος 9.1 απεικονίζει το παλινδρόγραμμα για τα δεδομένα CMB του Παραδείγματος 9.1, θεωρώντας $m = 21$ το πλήθος διαστήματα.

Έστω τώρα ότι $x \in B_{j^\star}$ και ας ορίσουμε

$$w_i(x) = \begin{cases} \dfrac{1}{k_{j^\star}}, & x_i \in B_{j^\star} \\ 0, & \text{διαφορετικά.} \end{cases} .$$

Από τη σχέση (9.3) έπεται ότι:

$$\hat{f}(x) = \sum_{i=1}^{n} w_i(x) Y_i. \tag{9.4}$$

Ορισμός 9.2

Αν για κάθε x υπάρχει διάνυσμα $\boldsymbol{w}(x) = (w_1(x), \dots, w_n(x))^\top$, έτσι ώστε η $\hat{f}$ να γράφεται στη μορφή της σχέσης (9.4), τότε ο εκτιμητής $\hat{f}$ καλείται **γραμμικός εξομαλυντής** (linear smoother).

Συνεπώς, το παλινδρόγραμμα είναι ένα παράδειγμα γραμμικού εξομαλυντή. Στις επόμενες ενότητες, θα δούμε και άλλα παραδείγματα γραμμικών εξομαλυντών.

Παρατήρηση 9.1. Προσοχή: Ένας γραμμικός εξομαλυντής δεν υποθέτει ότι η (άγνωστη) συνάρτηση $f(x)$ είναι γραμμική.

Έστω τώρα ο $n \times n$ πίνακας $\boldsymbol{W}$, με το (i,j) στοιχείο του να είναι $W_{ij} = w_j(x_i)$, όπου $i = 1, \dots, n$ και $j = 1, \dots, n$. Για παράδειγμα, ας θεωρήσουμε $n = 9$ παρατηρήσεις και $m = 3$ διαστήματα, όπου καθένα από αυτά περιέχει $k_1 = 3$ (έστω τις x_1, x_2, x_3), $k_2 = 4$ (έστω τις x_4, x_5, x_6, x_7) και $k_3 = 2$ (x_8, x_9) το πλήθος παρατηρήσεις, αντίστοιχα. Ο πίνακας W σε αυτήν την περίπτωση ισούται με

$$\boldsymbol{W} = \begin{pmatrix} 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \end{pmatrix}.$$

Ας ορίσουμε, τώρα, το διάνυσμα των **προσαρμοσμένων** τιμών $\hat{\boldsymbol{Y}} = (\hat{f}(x_1), \dots, \hat{f}(x_n))^\top$. Οι προσαρμοσμένες τιμές γράφονται με τη βοήθεια του πίνακα $\boldsymbol{W}$ ως γραμμικός συνδυασμός των παρατηρήσεων, ήτοι γράφονται στη μορφή:

$$\hat{\boldsymbol{Y}} = \boldsymbol{W}\boldsymbol{Y} \tag{9.5}$$

όπου $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$.

Παρατήρηση 9.2. Το ίχνος του πίνακα $\boldsymbol{W}$ στη σχέση (9.5) για έναν γραμμικό εξομαλυντή, που προσδιορίζεται από τη σχέση (9.4), ονομάζεται **βαθμοί ελευθερίας του εξομαλυντή**:

$$df = \text{tr}(\boldsymbol{W}).$$

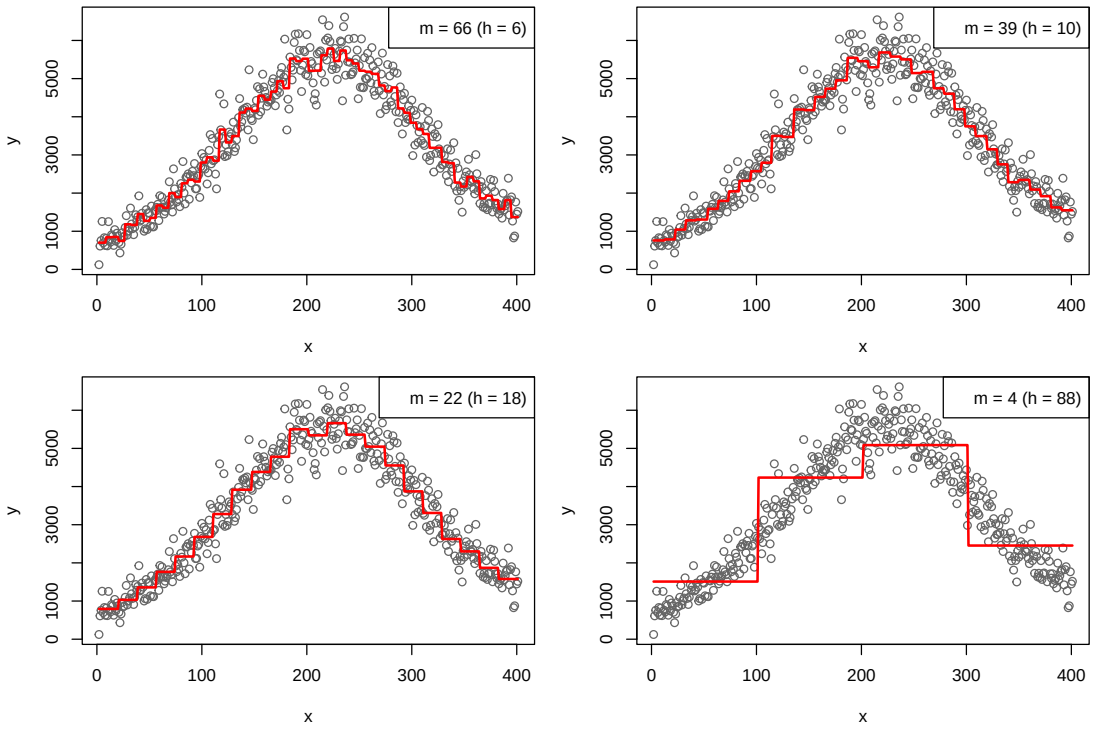
Οι βαθμοί ελευθερίας ενός εξομαλυντή αντιστοιχούν στον «αποτελεσματικό αριθμό παραμέτρων» (βλ. Ενότητα 7.6 των Hastie *et al.*, 2009, για περισσότερες λεπτομέρειες) που χρησιμοποιεί ένας εξομαλυντής. Η ποσότητα αυτή μας επιτρέπει να συγκρίνουμε την πολυπλοκότητα διαφορετικών εξομαλυντών μεταξύ τους. Παρατηρήστε ότι στο προηγούμενο παράδειγμα $df = \text{tr}(\boldsymbol{W}) = 3$, δηλαδή οι βαθμοί ελευθερίας είναι ίσοι και με τον αριθμό των διαστημάτων του παλινδρογράμματος.

Παράδειγμα 9.2 (συνέχεια Παραδείγματος 9.1)**.** Στα δεδομένα του Παραδείγματος 9.1 να υπολογιστεί με χρήση της `R` το παλινδρόγραμμα. Θεωρήστε διαφορετικές τιμές του πλήθους κελιών και συγκρίνετε τα αποτελέσματα γραφικά.

Λύση Παραδείγματος 9.2. Αρχικά, θα ορίσουμε τη συνάρτηση `regressogram()` η οποία θα επιστρέφει τις προσαρμοσμένες τιμές του παλινδρογράμματος (βλ. σχέση (9.5)). Το όρισμα αυτής της συνάρτησης θα είναι το εύρος των κελιών της διαμέρισης h (φυσικά, αυτή η συνάρτηση εξαρτάται και από τα παρατηρηθέντα δεδομένα). Κατόπιν, θα υπολογίσουμε το παλινδρόγραμμα θεωρώντας τέσσερις διαφορετικές τιμές του εύρους των διαστημάτων διαμέρισης ($h = 6, 10, 18, 88$), οι οποίες αντιστοιχούν σε πλήθος διαστημάτων $m = 66, 39, 22,$ και 4, αντίστοιχα. Στον κώδικα που ακολουθεί με `x` συμβολίζουμε το διάνυσμα των 400 παρατηρήσεων του δείγματός μας (βλ. Παράδειγμα 9.1 για την ανάκτηση των δεδομένων).

```
1  regressogram <- function(h,plot=F){
2    b <- max(x)
3    a <- min(x)
4    nBreaks <- (b-a)/h
5    binnedX <- cut(x,breaks=nBreaks)
6    # split y according to binnedX
7    splitY <- split(y,binnedX)
8    splitIndex<-split(1:n,binnedX)
9    # smoothing matrix
10   L <- matrix(data=0, nrow = n, ncol = n)
11   for(j in 1:length(splitIndex)){
12     L[splitIndex[[j]],splitIndex[[j]]] <- 1/length(splitIndex[[j]])
13   }
14   rg <- L % * % matrix(y)
15   if(plot==TRUE){
16     plot(x,y,col='gray40')
17     points(x,rg,type='l',col='red',lwd=2)
18     legend('topright',paste0('m = ',floor((b-a)/h), ' (h = ', round(h,
         0),')'))
19   }
20   return(rg)
21 }
22
23 par(mfrow = c(2,2),mar=c(4,4,1,1))
24 regressogram(6,  plot = T)
25 regressogram(10, plot = T)
26 regressogram(18, plot = T)
27 regressogram(88, plot = T)
```

Το αποτέλεσμα των παραπάνω εντολών κατασκευάζει το Σχήμα 9.2. Παρατηρούμε ότι το πλήθος των κελιών είναι καθοριστικό για το πόσο λεία είναι η τελική εκτίμηση, γεγονός που ήταν εκ των προτέρων αναμενόμενο. Ειδικότερα, καθώς μειώνεται το πλήθος των κελιών (ισοδύναμα, καθώς αυξάνεται το h_x) η εκτίμηση γίνεται όλο και πιο λεία. Τίθεται λοιπόν το ερώτημα: με βάση τα συγκεκριμένα διαγράμματα, ποια τιμή του h_x θα επιλέξουμε; Η απάντηση αυτή θα μας απασχολήσει στη συνέχεια σε ξεχωριστή ενότητα (βλ. Ενότητα 9.6) τόσο για αυτήν την τεχνική αλλά και για όσες τεχνικές παρουσιαστούν στη συνέχεια. □



Σχήμα 9.2: Οι εκτιμήσεις της συνάρτησης παλινδρόμησης με βάση το παλινδρόγραμμα για διαφορετικό πλήθος κελιών για τα δεδομένα CMB του Παραδείγματος 9.1.

9.3 Ο εκτιμητής των Nadaraya-Watson

Στην ενότητα αυτή θα δούμε πώς βρίσκει εφαρμογή η μέθοδος των kernels στη μη παραμετρική παλινδρόμηση, δηλαδή ουσιαστικά στην εκτίμηση της δεσμευμένης αναμενόμενης τιμής της εξαρτημένης μεταβλητής Y δοθείσης της τιμής της ανεξάρτητης μεταβλητής X.

Έστω (x_1, y_1),...,(x_n, y_n) ένα δείγμα ανεξάρτητων παρατηρήσεων από ένα διδιάστατο τυχαίο διάνυσμα (X, Y), με από κοινού αθροιστική συνάρτηση κατανομής $F_{XY}(x, y)$ και από κοινού συνάρτηση πυκνότητας πιθανότητας $f_{XY}(x, y)$. Επιπρόσθετα, έστω $f_X(\cdot)$ και $f_Y(\cdot)$ οι περιθώριες συναρτήσεις πυκνότητας πιθανότητας των τυχαίων μεταβλητών X και Y, αντίστοιχα. Τέλος, έστω $m(x)$ η εξίσωση παλινδρόμησης της Y ως προς X, δηλαδή $m(x) = \mathrm{E}(Y|X = x)$. Προφανώς, αν η $m(x)$ είναι γνωστής αναλυτικής μορφής και περιλαμβάνει έναν συγκεκριμένο αριθμό άγνωστων παραμέτρων, τότε υπάρχουν πολύ γνωστές μέθοδοι για την εκτίμηση αυτών των παραμέτρων από τα εμπειρικά δεδομένα, όπως, π.χ. η μέθοδος ελαχίστων τετραγώνων. Ωστόσο, το πρόβλημα έγκειται στον προσδιορισμό ενός εκτιμητή $\widehat{m}(x)$, ο οποίος θα συγκλίνει στην άγνωστη $m(x)$, ανεξάρτητα από τη μορφή των $f_{XY}(x, y)$ και $m(x)$. Ένας τέτοιος εκτιμητής προτάθηκε ανεξάρτητα και ταυτόχρονα από τους Nadaraya (1964) και Watson (1964). Ειδικότερα, από τις

ιδιότητες της δεσμευμένης τιμής έχουμε ότι:

$$m(x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy = \int_{-\infty}^{\infty} y \frac{f_{X,Y}(x,y)}{f_X(x)} dy.$$

Η ιδέα λοιπόν έγκειται στην αντικατάσταση των $f_{X,Y}(x,y)$ και $f_X(x)$ από τους εκτιμητές πυρήνα (βλ. Κεφάλαιο 3)

$$\widehat{f}_{X,Y}(x,y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right) K_y\left(\frac{y-y_i}{h_y}\right) \tag{9.6}$$

και

$$\widehat{f}_X(x) = \frac{1}{nh_x} \sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right), \tag{9.7}$$

αντίστοιχα. Στο σημείο αυτό, θα πρέπει να σημειώσουμε ότι για λόγους απλούστευσης στην περίπτωση της από κοινού συνάρτησης πυκνότητας πιθανότητας έχει χρησιμοποιηθεί το γινόμενο δύο ανεξάρτητων kernels, ένα για καθεμία εκ των δύο μεταβλητών. Προφανώς, κάποιος θα μπορούσε να έχει χρησιμοποιήσει πολυπλοκότερους πυρήνες. Επίσης, ο δείκτης τόσο στον πυρήνα $K(\cdot)$, όσο και στο παράθυρο $h(\cdot)$, αναφέρεται στη μεταβλητή που χρησιμοποιούμε. Αντικαθιστώντας τις παραπάνω εκτιμήσεις στη δεσμευμένη αναμενόμενη τιμή έχουμε:

$$\begin{aligned}
\widehat{m}(x) &= \int_{-\infty}^{\infty} y \frac{\frac{1}{nh_x h_y} \sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right) K_y\left(\frac{y-y_i}{h_y}\right)}{\frac{1}{nh_x} \sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right)} dy \\
&= \frac{\sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right) \int_{-\infty}^{\infty} \frac{y}{h_y} K_y\left(\frac{y-y_i}{h_y}\right) dy}{\sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right)}.
\end{aligned}$$

Για τον υπολογισμό του τελευταίου ολοκληρώματος θέτουμε $(y - y_i)/h_y = u$ και έχουμε ότι:

$$\begin{aligned}
\widehat{m}(x) &= \frac{\sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right) \int_{-\infty}^{\infty} (uh_y + y_i) K_y(u)\, du}{\sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right)} \\
&= \frac{\sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right) y_i}{\sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right)}
\end{aligned}$$

καθώς (βλ. Ορισμό 3.6, Ενότητα 3.4.2) $\int_{-\infty}^{\infty} K_y(u)du = 1$ και $\int_{-\infty}^{\infty} uK_y(u)du = 0$. Επομένως,

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right) y_i}{\sum_{i=1}^{n} K_x\left(\frac{x-x_i}{h_x}\right)} = \sum_{i=1}^{n} y_i \frac{K_x\left(\frac{x-x_i}{h_x}\right)}{\sum_{j=1}^{n} K_x\left(\frac{x-x_j}{h_x}\right)},$$

και αυτό μας οδηγεί στον ακόλουθο ορισμό.

Ορισμός 9.3

Ο εκτιμητής **Nadaraya-Watson** ορίζεται ως

$$\widehat{m}_{NW}(x) = \sum_{i=1}^{n} w_i(x) Y_i, \tag{9.8}$$

με

$$w_i(x) = \frac{K_x\left(\frac{x-x_i}{h_x}\right)}{\sum_{j=1}^{n} K_x\left(\frac{x-x_j}{h_x}\right)}, \tag{9.9}$$

όπου η συνάρτηση K_x είναι ένας πυρήνας (βλ. Ορισμό 3.6) με εύρος παραθύρου $h_x > 0$.

Από τη σχέση (9.8) είναι φανερό ότι ο εκτιμητής των Nadaraya-Watson είναι ένας γραμμικός εξομαλυντής της μορφής (9.4), διότι μπορεί να γραφτεί ως γραμμικός συνδυασμός ή σταθμισμένο άθροισμα των παρατηρήσεων y_i, με βάρη ή συντελεστές στάθμισης w_i. Ο εκτιμητής αυτός δίνει περισσότερο βάρος σε σημεία που είναι πλησιέστερα στο εκάστοτε x. Ισοδύναμα, λαμβάνοντας υπόψη τη σχέση (9.7), οι συντελεστές στάθμισης στη σχέση (9.9) γράφονται και ως

$$w_i(x) = \frac{1}{nh_x} \cdot \frac{K_x\left(\frac{x-x_i}{h_x}\right)}{\widehat{f}_X(x)}. \tag{9.10}$$

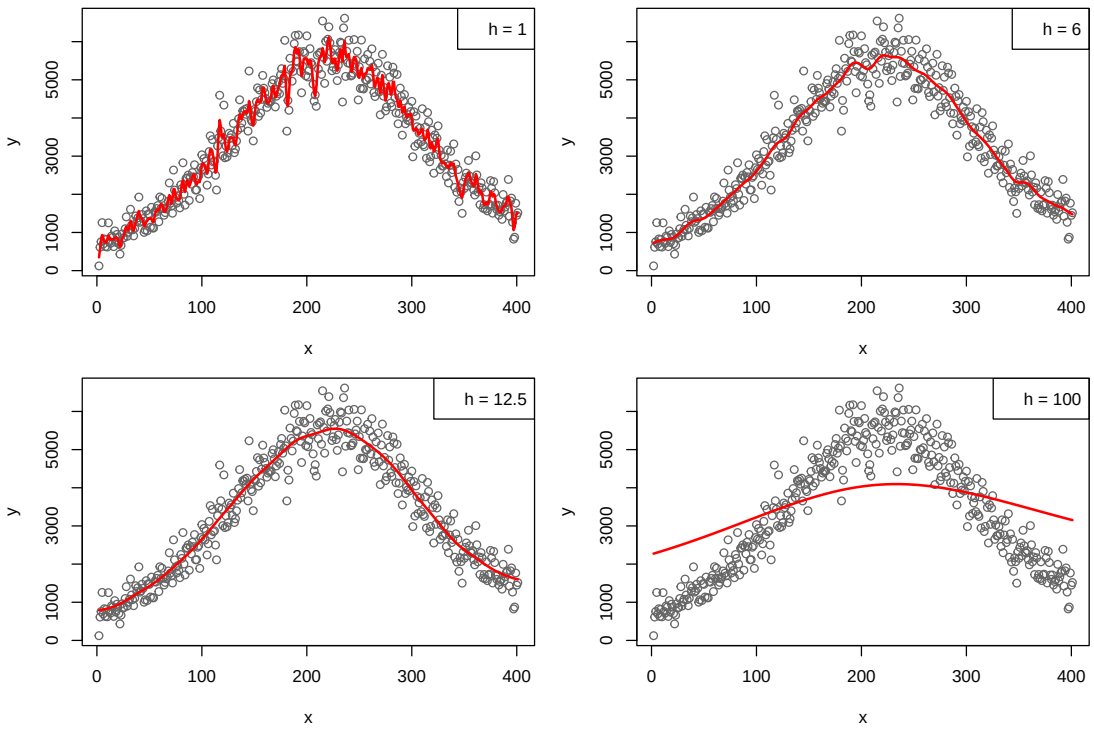
Παράδειγμα 9.3 (συνέχεια Παραδείγματος 9.1). Στα δεδομένα του παραδείγματος 9.1 να υπολογιστεί με χρήση της `R` ο εκτιμητής Nadaraya-Watson κάνοντας χρήση κανονικού πυρήνα για τις εξής τιμές του εύρους παραθύρου $h_x = 1, 6, 12.5, 100$. Κατόπιν, να απεικονιστούν γραφικά οι εκτιμήσεις της συνάρτησης παλινδρόμησης πάνω από το διάγραμμα διασποράς των δεδομένων. Να σχολιαστεί το αποτέλεσμα όσον αφορά τις διαφορετικές επιλογές του εύρους παραθύρου.

Λύση Παραδείγματος 9.3. Ο παρακάτω κώδικας R ορίζει αρχικά τη συνάρτηση `ker()`, η οποία υπολογίζει διάφορους πυρήνες. Εδώ, έχουν χρησιμοποιηθεί οι πυρήνες Gaussian, boxcar, Epanechnikov, tricube (βλ. Κεφάλαιο 3, Ενότητα 3.4.2). Κατόπιν, υπολογίζεται ο εκτιμητής Nadaraya-Watson μέσω της συνάρτησης `nadarayaWatson()` που ορίζεται στη συνέχεια. Η συνάρτηση αυτή επιστρέφει τις τιμές του εκτιμητή (9.8) σε καθεμία από τις παρατηρήσεις του διαθέσιμου δείγματος (`x`).

```
ker <- function(x,type){
  if(type=='gaussian'){
    return(dnorm(x))
  }
  if(type=='boxcar'){
    return(0.5*(abs(x)<=1))
  }
  if(type=='epanechnikov'){
    return(0.75*(1-x^2)*(abs(x)<=1))
  }
  if(type=='tricube'){
    return((70*(1-abs(x)^3)^3/81)*(abs(x)<=1))
  }
}

nadarayaWatson <- function(h,plot=F,type){
  # smoothing matrix
  L <- matrix(data=0, nrow = n, ncol = n)
  for(j in 1:n){
```

```
    L[j,] <- ker((x-x[j])/h,type=type)
  }
  for(i in 1:n){
    L[i,] <- L[i,]/sum(L[i,])
  }
  rg <- L % * % matrix(y)
  if(plot==TRUE){
    plot(x,y,col='gray40')
    points(x,rg,type='l',col='red',lwd=2)
    legend('topright',paste0('h = ', h))
  }
  return(rg)
}

par(mfrow = c(2,2), mar = c(4, 4, 1, 1))
nadarayaWatson(1, plot = T, type = "gaussian")
nadarayaWatson(6, plot = T, type = "gaussian")
nadarayaWatson(12.5, plot = T, type = "gaussian")
nadarayaWatson(100, plot = T, type = "gaussian")
```

Το αποτέλεσμα των παραπάνω εντολών κατασκευάζει το Σχήμα 9.3. Παρατηρούμε, όπως είναι αναμενόμενο, ότι το εύρος παραθύρου είναι καθοριστικό για το πόσο λεία είναι η τελική εκτίμηση. Ειδικότερα, καθώς αυξάνεται το h_x, η εκτίμηση γίνεται όλο και πιο λεία. Το ερώτημα που τίθεται είναι: με βάση τα συγκεκριμένα διαγράμματα, ποια τιμή του h_x θα επιλέγατε;



Σχήμα 9.3: Οι εκτιμήσεις της συνάρτησης παλινδρόμησης με βάση τον εκτιμητή Nadaraya-Watson με χρήση κανονικού πυρήνα και διαφορετικές τιμές του εύρους παραθύρου h_x για τα δεδομένα CMB του Παραδείγματος 9.1.

□

Χρησιμοποιώντας στη σχέση (9.9) τον απλοϊκό (boxcar) πυρήνα (βλ. Κεφάλαιο 3, Ενότητα 3.4.2) προκύπτει

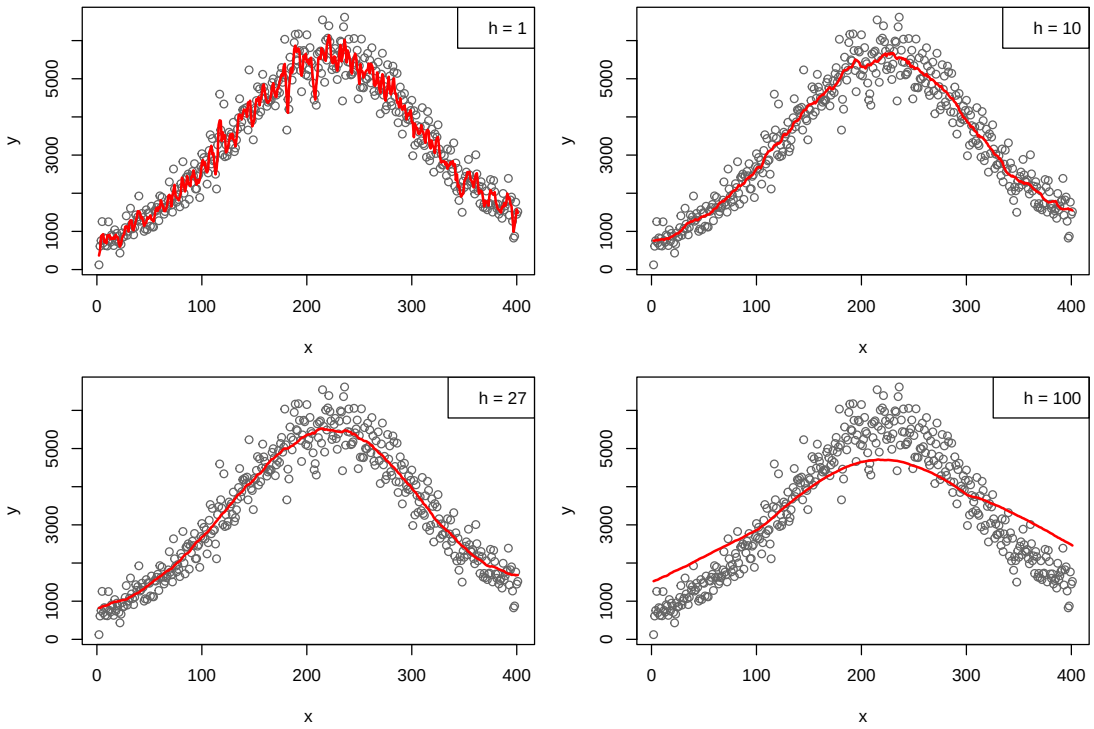
ο εκτιμητής:

$$\widehat{m}_{\mathrm{N}}(x;h) = \frac{\sum_i^n I(x_i - h < x < x_i + h) y_i}{\sum_{i=1}^n I(x_i - h < x < x_i + h)} = \sum_{i=1}^n \frac{1}{|N(x;h)|} \sum_{i \in N(x;h)} y_i, \tag{9.11}$$

όπου $N(x;h) := \{i = 1,...,n : |x_i - x| < h\}$ είναι το σύνολο των δεικτών του δείγματος εντός της γειτονιάς του x και $|N(x;h)|$ είναι το πλήθος των στοιχείων αυτού του συνόλου. Ο εκτιμητής της σχέσης (9.11) καλείται εκτιμητής τοπικών μέσων (local average estimator κατά τον Wasserman, 2006) ή απλοϊκός εκτιμητής παλινδρόμησης (naive regression estimator). Από τη δεύτερη έκφραση του εκτιμητή στη σχέση (9.11) προκύπτει ότι ο εκτιμητής αυτός δεν είναι τίποτε άλλο παρά ο δειγματικός μέσος εκείνων των Y_i που αντιστοιχούν στα x_i, τα οποία ανήκουν στη γειτονιά που ορίζεται από τη σχέση $|x_i - x| \leqslant h$, για το δοθέν x.

Παράδειγμα 9.4 (συνέχεια Παραδείγματος 9.1)**.** Στα δεδομένα του Παραδείγματος 9.1 να υπολογιστεί με χρήση της R ο εκτιμητής τοπικών μέσων για τις εξής τιμές του εύρους παραθύρου $h_x = 1,10,27,100$. Κατόπιν, να απεικονιστούν γραφικά οι εκτιμήσεις της συνάρτησης παλινδρόμησης πάνω από το διάγραμμα διασποράς των δεδομένων. Να σχολιαστεί το αποτέλεσμα όσον αφορά τις διαφορετικές επιλογές του εύρους παραθύρου.

Λύση Παραδείγματος 9.4. Το αποτέλεσμα των παραπάνω εντολών κατασκευάζει το Σχήμα 9.4. Παρατηρούμε ότι το εύρος παραθύρου είναι καθοριστικό για το πόσο λεία είναι η τελική εκτίμηση, όπως είναι αναμενόμενο. Ειδικότερα, καθώς αυξάνει το h_x, η εκτίμηση γίνεται όλο και πιο λεία. Το ερώτημα που τίθεται είναι: με βάση τα συγκεκριμένα διαγράμματα, ποια τιμή του h_x θα επιλέγατε; □



Σχήμα 9.4: Οι εκτιμήσεις της συνάρτησης παλινδρόμησης με βάση τον εκτιμητή τοπικών μέσων όρων (δηλαδή τον εκτιμητή Nadaraya-Watson με χρήση απλοϊκού πυρήνα) και διαφορετικές τιμές του εύρους παραθύρου h για τα δεδομένα CMB του Παραδείγματος 9.1.

9.4 Τοπική πολυωνυμική παλινδρόμηση

Ο εκτιμητής των Nadaraya–Watson, που παρουσιάστηκε στην προηγούμενη ενότητα, μπορεί να θεωρηθεί ειδική περίπτωση μιας ευρύτερης οικογένειας μη παραμετρικών εκτιμητών, των γνωστών ως τοπικών

πολυωνυμικών εκτιμητών, που αποτελούν αντικείμενο μελέτης αυτής της ενότητας. Η κεντρική ιδέα των τοπικών πολυωνύμων προκύπτει από την προσπάθεια εύρεσης ενός εκτιμητή (έστω αυτός $\widehat{m}(\cdot)$) της συνάρτησης $m(\cdot)$, ο οποίος ελαχιστοποιεί το άθροισμα τετραγώνων των σφαλμάτων:

$$\sum_{i=1}^{n}(y_i - m(x_i))^2, \tag{9.12}$$

χωρίς να υποθέτουμε κάποια συγκεκριμένη μορφή για την αληθινή συνάρτηση $m(x)$, πέραν του ότι είναι «ομαλή». Να τονίσουμε εδώ ότι η ελαχιστοποίηση της (9.12) στον χώρο όλων των συναρτήσεων δίνει την (καθόλου ενδιαφέρουσα) λύση μιας συνάρτησης που παρεμβάλλει τα παρατηρηθέντα δεδομένα. Θυμηθείτε ότι στην κλασική γραμμική παλινδρόμηση υποθέτουμε ότι $m(x) = a + \beta x$ και το πρόβλημα ανάγεται στην εύρεση των παραμέτρων για τις οποίες ελαχιστοποιείται το $\sum_{i=1}^{n}(y_i - a - \beta x_i)^2$. Όταν λοιπόν η $m(\cdot)$ δεν έχει κάποια γνωστή παραμετρική μορφή και μπορεί να λάβει οποιαδήποτε μαθηματική έκφραση, θα πρέπει να ακολουθηθεί μια εναλλακτική προσέγγιση. Μια τέτοια περιγράφεται στη συνέχεια.

Αρχικά, δημιουργούμε μια τοπική παραμετροποίηση (local parametrization) για τη συνάρτηση $m(\cdot)$. Κάτι τέτοιο μπορεί να επιτευχθεί χρησιμοποιώντας το ανάπτυγμα Taylor τάξης p της $m(x_i)$ γύρω από το x, για x που είναι κοντά στο x_i. Σε όσα ακολουθούν, υποθέτουμε ότι οι παράγωγοι $p+1$ τάξης της $m(x)$ υπάρχουν στο σημείο x και είναι συνεχείς. Επομένως,

$$m(x_i) \approx m(x) + m^{(1)}(x)(x_i - x) + \frac{m^{(2)}(x)}{2!}(x_i - x)^2 + \ldots + \frac{m^{(p)}(x)}{p!}(x_i - x)^p, \tag{9.13}$$

όπου με $m^{(k)}(x)$ συμβολίζουμε την παράγωγο k τάξης της $m(x)$ ως προς x, ενώ ορίζουμε να είναι $m^{(0)}(x) := m(x)$. Αντικαθιστώντας τη σχέση (9.13) στη σχέση (9.12), προκύπτει ότι πρέπει να ελαχιστοποιήσουμε το ακόλουθο άθροισμα τετραγώνων:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p}\frac{m^{(j)}(x)}{j!}(x_i - x)^j\right)^2. \tag{9.14}$$

Ωστόσο, και η παραπάνω σχέση δεν μπορεί να χρησιμοποιηθεί, καθώς εξαρτάται από τις $m^{(j)}$, $j = 0,\ldots,p$, οι οποίες είναι και αυτές προφανώς άγνωστες, αφού η $m(\cdot)$ είναι άγνωστη. Στη συνέχεια, η επόμενη βασική ιδέα είναι να αντικατασταθούν οι ποσότητες $\frac{m^{(j)}(x)}{j!}$ από την παράμετρο β_j και έτσι, από τη σχέση (9.14), να οδηγηθούμε σε ένα πρόβλημα γραμμικής παλινδρόμησης όπου θα πρέπει να εκτιμηθούν οι παράμετροι $\beta_0,\ldots\beta_p$ ή, ισοδύναμα, το διάνυσμα των παραμέτρων $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top$. Επομένως, θέλουμε να ελαχιστοποιήσουμε ως προς $\beta_0,\ldots\beta_p$ το ακόλουθο άθροισμα τετραγώνων

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p}\beta_j(x_i - x)^j\right)^2. \tag{9.15}$$

Η ελαχιστοποίηση της σχέσης (9.15) ως προς β_j, $j = 0,1,\ldots,p$ μπορεί να επιτευχθεί με την κλασική θεωρία των εκτιμητών ελαχίστων τετραγώνων. Ωστόσο, θέλοντας να σταθμίσουμε τη συνεισφορά κάθε ζεύγους τιμών (x_i, y_i) στην εκτίμηση της $m(x)$ ανάλογα με την απόσταση του x_i από το x, τροποποιούμε τη σχέση (9.15) και οδηγούμαστε στο ότι πρέπει να ελαχιστοποιηθεί ως προς β_j, $j = 0,1,\ldots,p$ η ακόλουθη συνάρτηση:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p}\beta_j(x_i - x)^j\right)^2 K\left(\frac{x - x_i}{h}\right), \tag{9.16}$$

για κάποιον πυρήνα $K(\cdot)$ και κάποια παράμετρο εξομάλυνσης $h > 0$. Συμβολίζοντας με $\mathbf{X}_x$ τον $n \times (p+1)$ πίνακα

$$\boldsymbol{X}_x = \begin{pmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^p \\ 1 & x_2 - x & \cdots & (x_2 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^p \end{pmatrix} \tag{9.17}$$

με $\boldsymbol{W}_x$ τον $n \times n$ διαγώνιο πίνακα

$$\boldsymbol{W}_x = diag\left(K\left(\frac{x - x_1}{h}\right), ..., K\left(\frac{x - x_n}{h}\right)\right), \tag{9.18}$$

και με $\boldsymbol{Y} = (y_i, ..., y_n)^\top$, μετατρέπουμε το πρόβλημα σε πρόβλημα εύρεσης εκτιμητών ελαχίστων τετραγώνων με βάρη/συντελεστές στάθμισης (weighted least squares). Πιο συγκεκριμένα, θέλουμε να ελαχιστοποιήσουμε ως προς $\boldsymbol{\beta}$ την ποσότητα

$$(\boldsymbol{Y} - \boldsymbol{X}_x\boldsymbol{\beta})^\top \boldsymbol{W}_x (\boldsymbol{Y} - \boldsymbol{X}_x\boldsymbol{\beta}),$$

από όπου προκύπτει ότι

$$\widehat{\boldsymbol{\beta}}_h = (\boldsymbol{X}_x^\top \boldsymbol{W}_x \boldsymbol{X}_x)^{-1} \boldsymbol{X}_x^\top \boldsymbol{W}_x \boldsymbol{Y}. \tag{9.19}$$

Λαμβάνοντας υπόψη ότι $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^\top = \left(m(x), m^{(1)}/1!, ..., m^{(p)}/p!\right)^\top$, άμεσα προκύπτει ότι ο εκτιμητής της $m(x)$ είναι ο εκτιμητής του σταθερού όρου β_0. Επομένως, από τη σχέση (9.19) έπεται ο ακόλουθος ορισμός:

Ορισμός 9.4

Ο εκτιμητής τοπικού πολυωνύμου είναι:

$$\widehat{m}(x; p, h) = \sum_{i=1}^{n} w_i(x) Y_i, \tag{9.20}$$

με

$$\boldsymbol{w}(x)^\top = (w_1(x), \dots, w_n(x)) = \boldsymbol{e}_1^\top (\boldsymbol{X}_x^\top \boldsymbol{W}_x \boldsymbol{X}_x)^{-1} \boldsymbol{X}_x^\top \boldsymbol{W}_x,$$

όπου $\boldsymbol{X}_x$ και $\boldsymbol{W}_x$ ορίζονται μέσω των σχέσεων (9.17) και (9.18), αντίστοιχα, και $\boldsymbol{e}_1 = (1, 0, ..., 0)^\top$ είναι το $(p+1) \times 1$ διάνυσμα που έχει την τιμή 1 στην πρώτη θέση και όλες τις άλλες συνιστώσες ίσες με μηδέν.

Παρατήρηση 9.3. Παρατηρήστε ότι χρησιμοποιήθηκε ο συμβολισμός $\widehat{m}(x; p, h)$, για να δείξουμε ότι ο εκτιμητής της $m(x)$ εξαρτάται τόσο από την τάξη του πολυωνύμου p στο ανάπτυγμα Taylor όσο και από την παράμετρο εξομάλυνσης h του πυρήνα K. Τέλος, από τη σχέση (9.20) προκύπτει ότι ο εκτιμητής τοπικού πολυωνύμου είναι γραμμικός εξομαλυντής της μορφής (9.4).

Παρατήρηση 9.4. Όταν $p = 0$, έχουμε ότι $\boldsymbol{e}_1 = 1$ και $\boldsymbol{X}_x = (1, 1, \dots, 1)^\top$, ενώ τα βάρη που εμφανίζονται στη σχέση (9.20) δίνονται από τη σχέση:

$$w_i(x) = \left(\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right)\right)^{-1} K\left(\frac{x - x_i}{h}\right),$$

δηλαδή προέκυψαν κατά αυτόν τον τρόπο τα βάρη του εκτιμητή των Nadaraya-Watson (βλ. σχέση (9.9)), που για τον λόγο αυτόν ονομάζεται τοπικός σταθερός εκτιμητής.

Παράδειγμα 9.5 (συνέχεια Παραδείγματος 9.1)**.** Στα δεδομένα του Παραδείγματος 9.1 να υπολογιστεί με χρήση της `R` ο εκτιμητής τοπικής πολυωνυμικής παλινδρόμησης για $p = 3$, κάνοντας χρήση κανονικού πυρήνα για τις εξής τιμές του εύρους παραθύρου $h = 1, 10, 35, 100$. Κατόπιν, να απεικονιστούν γραφικά οι εκτιμήσεις της συνάρτησης παλινδρόμησης πάνω από το διάγραμμα διασποράς των δεδομένων. Να σχολιαστεί το αποτέλεσμα όσον αφορά τις διαφορετικές επιλογές του εύρους παραθύρου.

Λύση Παραδείγματος 9.5. Ο παρακάτω κώδικας R υπολογίζει και απεικονίζει γραφικά τον εκτιμητή κυβικού τοπικού πολυωνύμου με χρήση κανονικού πυρήνα μέσω της βιβλιοθήκης `NonpModelCheck` (Zambom and Akritas, 2017).

```
library("NonpModelCheck")
fit1 <- localpoly.reg(x, y, degree.pol = 3, kernel.type = "gaussian",
   bandwidth = 1)
fit2 <- localpoly.reg(x, y, degree.pol = 3, kernel.type = "gaussian",
   bandwidth = 15)
fit3 <- localpoly.reg(x, y, degree.pol = 3, kernel.type = "gaussian",
   bandwidth = 35)
fit4 <- localpoly.reg(x, y, degree.pol = 3, kernel.type = "gaussian",
   bandwidth = 100)
par(mfrow = c(2,2), mar=c(4,4,1,1))
plot(x, y, col='gray40')
points(x, fit1$predicted, type = 'l', col = 'red', lwd = 2)
legend('topright', "h = 1")
plot(x, y, col='gray40')
points(x, fit2$predicted, type = 'l', col = 'red', lwd = 2)
legend('topright', "h = 15")
plot(x, y, col='gray40')
points(x, fit3$predicted, type = 'l', col = 'red', lwd = 2)
legend('topright', "h = 35")
plot(x, y, col='gray40')
points(x, fit4$predicted, type = 'l', col = 'red', lwd = 2)
legend('topright', "h = 100")
```

Το αποτέλεσμα των παραπάνω εντολών κατασκευάζει το Σχήμα 9.5. Παρατηρούμε ότι το εύρος παραθύρου είναι καθοριστικό για το πόσο λεία είναι η τελική εκτίμηση, όπως είναι αναμενόμενο. Καθώς αυξάνεται το h, η εκτίμηση γίνεται όλο και πιο λεία. Με βάση τα συγκεκριμένα διαγράμματα, ποια τιμή του h θα επιλέγατε;

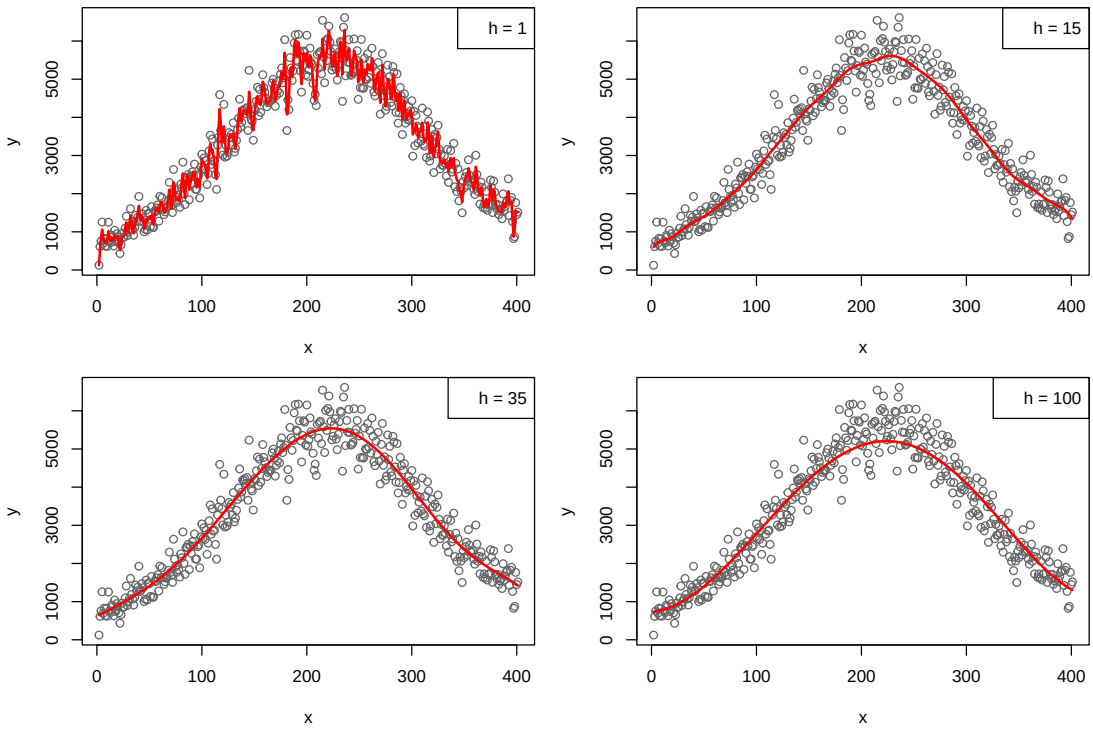
□

Παρατήρηση 9.5. Ο Fan (1992) έδειξε ότι ασυμπτωτικά ο εκτιμητής τοπικού πολυωνύμου υπερτερεί του εκτιμητή Nadaraya-Watson. Ένα χαρακτηριστικό μειονέκτημα του εκτιμητή Nadaraya-Watson είναι η μεροληψία στα άκρα, κάτι που δεν ισχύει για τον εκτιμητή τοπικού πολυωνύμου. Παρατηρήστε, για παράδειγμα, τη συμπεριφορά του εκτιμητή Nadaraya-Watson με χρήση κανονικού πυρήνα στο Σχήμα 9.3, για $h = 12.5$ ή με χρήση απλοϊκού πυρήνα στο Σχήμα 9.4 για $h = 27$. Εύκολα προκύπτει ότι και στις δύο περιπτώσεις υπάρχει μεροληψία στα άκρα, δηλαδή για $x \to 0$ και $x \to 400$. Αντίθετα, ο εκτιμητής τοπικού πολυωνύμου στο Σχήμα 9.6 για $h = 35$ δεν παρουσιάζει (τουλάχιστον οπτικά) κάποιου είδους συστηματική μεροληψία.

9.5 Εξομαλυντές splines

Για μία ολοκληρωμένη παρουσίαση της εξομάλυνσης μέσω splines παραπέμπουμε στο σύγγραμμα της Wahba (1990) (βλ., επίσης Reinsch, 1967; Wahba, 1975). Για τις ανάγκες του παρόντος συγγράμματος, θα περιοριστούμε στην αναζήτηση μιας συνάρτησης $\hat{f}(x)$, η οποία πρέπει να έχει συνεχείς παραγώγους πρώτης και δεύτερης τάξης και να ελαχιστοποιεί την

$$SS(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - g(x_i))^2 + h\int_a^b (g''(x))^2 \mathrm{d}x, \tag{9.21}$$



Σχήμα 9.5: Οι εκτιμήσεις της συνάρτησης παλινδρόμησης με βάση τον εκτιμητή κυβικού τοπικού πολυωνύμου με χρήση κανονικού πυρήνα και διαφορετικές τιμές του εύρους παραθύρου h για τα δεδομένα CMB του Παραδείγματος 9.1.

όπου $[a, b]$ είναι ένα διάστημα το οποίο περιέχει τα παρατηρηθέντα δεδομένα, δηλαδή $a < x_i < b$, για κάθε $i = 1, \ldots, n$. Ουσιαστικά, το πρόβλημα ελαχιστοποίησης της $SS(h)$ ως προς g είναι ένα πρόβλημα **ποινικοποιημένης παλινδρόμησης** (penalized regression). Ο πρώτος όρος της σχέσης (9.21) είναι το (μέσο) άθροισμα τετραγώνων των σφαλμάτων και ενθαρρύνει τη συνάρτηση την οποία επιζητούμε να είναι «κοντά» στα παρατηρηθέντα δεδομένα. Ο δεύτερος όρος στη σχέση (9.21) είναι ένας όρος ποινικοποίησης που αποτελείται από μία παράμετρο εξομάλυνσης $h > 0$, ενώ το $\int_a^b (g''(x))^2 \mathrm{d}x$ ελέγχει την τραχύτητα της συνάρτησης g.

Θυμηθείτε ότι η δεύτερη παράγωγος μιας συνάρτησης αντιστοιχεί στον ρυθμό μεταβολής της πρώτης παραγώγου: μία γραμμική συνάρτηση έχει μηδενική δεύτερη παράγωγο, ενώ μία συνάρτηση με πολύ απότομες μεταβολές θα έχει δεύτερη παράγωγο με μεγάλες τιμές κατά απόλυτη τιμή. Έτσι, ο όρος $\int_a^b (g''(x))^2 \mathrm{d}x$ της σχέσης (9.21) μετράει συνολικά το πόσο απότομα μεταβάλλεται η κλίση της συνάρτησης g. Υπο αυτήν την έννοια, όσο πιο απότομη είναι η μεταβολή της κλίσης της $g(\cdot)$, τόσο πιο «τραχειά» θεωρείται αυτή. Η παράμετρος εξομάλυνσης h καθορίζει την ισορροπία μεταξύ των δύο αυτών όρων (προσαρμογής και ποινικοποίησης). Όταν $h = 0$, η λύση είναι η συνάρτηση παρεμβολής μεταξύ των $(x_1, Y_1), \ldots, (x_n, Y_n)$. Από την άλλη, όταν $h \to \infty$, η λύση είναι η ευθεία ελαχίστων τετραγώνων.

Ένα spline είναι μία συνάρτηση η οποία εκφράζεται ως πολυώνυμο εντός συγκεκριμένων τμημάτων. Τα κυβικά splines είναι αυτά που χρησιμοποιούνται πιο συχνά στην πράξη και για αυτό δίνεται ο επόμενος ορισμός.

Ορισμός 9.5

Έστω το σύνολο σημείων $\Xi = \{\xi_1, \xi_2, \ldots, \xi_k\}$, με $\xi_1 < \xi_2 < \ldots < \xi_k$, που περιέχεται εντός ενός συνόλου (a, b). Κάθε $\xi \in \Xi$ καλείται **δεσμός** ή **κόμβος**. Ένα **κυβικό spline** είναι μια συνεχής συνάρτηση r τέτοια ώστε:

(i) η $r(\cdot)$ είναι κυβικό πολυώνυμο σε κάθε διάστημα (ξ_1, ξ_2), ..., (ξ_{k-1}, ξ_k), και
(ii) η $r(\cdot)$ έχει συνεχή πρώτη και δεύτερη παράγωγο στους δεσμούς.

Ένα spline το οποίο είναι γραμμική συνάρτηση εκτός των οριακών δεσμών, δηλαδή αριστερά του ξ_1 και δεξιά του ξ_k, καλείται **φυσικό spline**.

Θεώρημα 9.1

Η συνάρτηση $\hat{f}(x)$ η οποία ελαχιστοποιεί την (9.21) είναι ένα φυσικό κυβικό spline με δεσμούς στις διακριτές τιμές των $x_1, \ldots, x_n$. Ο αντίστοιχος εκτιμητής καλείται **εξομαλυντής spline**.

Απόδειξη Θεωρήματος 9.1. Η απόδειξη βασίζεται σε έννοιες του λογισμού μεταβολών οι οποίες ξεφεύγουν από τους σκοπούς του παρόντος συγγράμματος. Ο/Η ενδιαφερόμενος/μενη αναγνώστης/στρια παραπέμπεται στα άρθρα Schoenberg (1964a) και Schoenberg (1964b). □

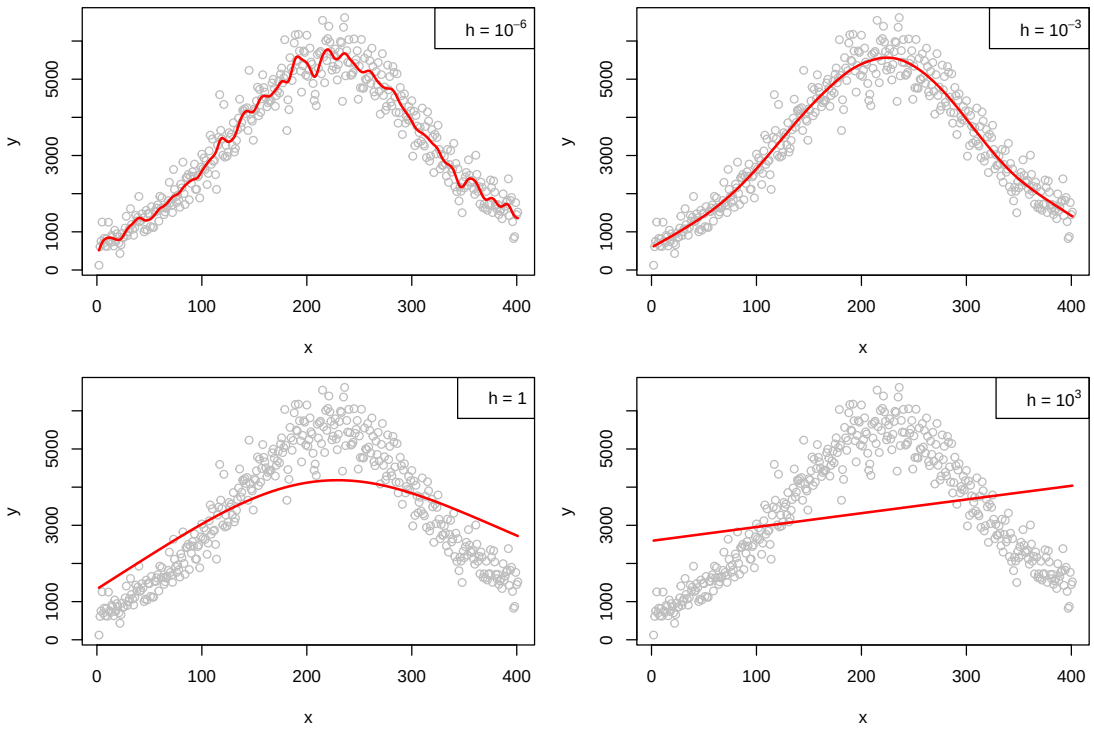
Μπορεί να δειχθεί ότι (βλ. Wasserman, 2006) ο εξομαλυντής spline είναι, επίσης, ένας γραμμικός εξομαλυντής της μορφής (9.4).

Παράδειγμα 9.6 (συνέχεια Παραδείγματος 9.1)**.** Στα δεδομένα του Παραδείγματος 9.1 να υπολογιστεί με χρήση της `R` ο εξομαλυντής spline για $h = 10^{-6}, 10^{-3}, 1, 10^{3}$. Κατόπιν, να απεικονιστούν γραφικά οι εκτιμήσεις της συνάρτησης παλινδρόμησης πάνω από το διάγραμμα διασποράς των δεδομένων. Να σχολιαστεί το αποτέλεσμα όσον αφορά τις διαφορετικές επιλογές της παραμέτρου εξομάλυνσης h.

Λύση Παραδείγματος 9.6. Ο παρακάτω κώδικας R υπολογίζει και απεικονίζει γραφικά τον εξομαλυντή spline μέσω της βιβλιοθήκης `splines`, η οποία είναι μέρος της `R`.

```
require("splines")
fit1 <- smooth.spline(x, y, lambda=0.000001)
fit2 <- smooth.spline(x, y, lambda=0.001)
fit3 <- smooth.spline(x, y, lambda=1)
fit4 <- smooth.spline(x, y, lambda=1000)
par(mfrow = c(2,2), mar=c(4,4,1,1))
plot(x, y, col="grey", xlab="x", ylab="y")
lines(fit1, col = "red", lwd = 2)
legend("topright", as.expression(bquote(lambda*' = 10'^{-6})))
plot(x, y, col="grey", xlab="x", ylab="y")
lines(fit2, col = "red", lwd = 2)
legend("topright", as.expression(bquote(lambda*' = 10'^{-3})))
plot(x, y, col="grey", xlab="x", ylab="y")
lines(fit3, col = "red", lwd = 2)
legend("topright", as.expression(bquote(lambda*' = 1')))
plot(x, y, col="grey", xlab="x", ylab="y")
lines(fit4, col = "red", lwd = 2)
legend("topright", as.expression(bquote(lambda*' = 10'^{3})))
```

Η παράμετρος `lambda` της εντολής `smooth.spline(...)` αντιστοιχεί στην παράμετρο εξομάλυνσης h της σχέσης (9.21). Το αποτέλεσμα των παραπάνω εντολών κατασκευάζει το Σχήμα 9.6. Παρατηρούμε ότι η παράμετρος εξομάλυνσης είναι καθοριστική για το πόσο λεία είναι η τελική εκτίμηση, όπως είναι αναμενόμενο. Ειδικότερα, καθώς αυξάνεται το h, η εκτίμηση γίνεται όλο και πιο λεία. Το ερώτημα που τίθεται είναι: με βάση τα συγκεκριμένα διαγράμματα, ποια τιμή του h θα επιλέγατε; □



Σχήμα 9.6: Οι εκτιμήσεις της συνάρτησης παλινδρόμησης με εξομαλυντή spline και διαφορετικές τιμές της παραμέτρου εξομάλυνσης h για τα δεδομένα CMB του Παραδείγματος 9.1.

9.6 Επιλογή παραμέτρου εξομάλυνσης

Είδαμε στις προηγούμενες ενότητες ότι οι μη παραμετρικοί εκτιμητές παλινδρόμησης εξαρτώνται από κάποια παράμετρο εξομάλυνσης h. Κατ' αναλογία με το Κεφάλαιο 3, χρησιμοποιούμε ως συνάρτηση κινδύνου το μέσο τετραγωνικό σφάλμα:

$$R(h) = \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - f(x_i))^2\right). \tag{9.22}$$

Ιδανικά, θα θέλαμε να ελαχιστοποιήσουμε την (9.22) ως προς h, αλλά αυτό δεν είναι δυνατόν διότι η $R(h)$ εξαρτάται από την (άγνωστη) συνάρτηση $f(x)$. Έτσι, ελαχιστοποποιείται μία εκτίμηση $\hat{R}(h)$ της $R(h)$.

Μία επιλογή που ενδεχομένως αποτελεί την πρώτη σκέψη μας για την εκτίμηση $\hat{R}(h)$ είναι η:

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(x_i))^2.$$

Δυστυχώς, η παραπάνω εκτίμηση δεν είναι καλή, καθώς τείνει να υποεκτιμά την $R(h)$ και, τυπικά, οδηγεί σε εκτιμήσεις της παραμέτρου h οι οποίες είναι μικρότερες από την ιδανική (overfitting). Όπως εξηγήσαμε στο Κεφάλαιο 3, αυτό συμβαίνει λόγω διπλής χρήσης των δεδομένων: μία για την εκτίμηση της συνάρτησης $f(x)$ και μία για την εκτίμηση του $R(h)$. Έτσι, θα χρησιμοποιήσουμε την τεχνική cross-validation για την εκτίμηση της $R(h)$ (θυμηθείτε την Ενότητα 3.2.2).

Ορισμός 9.6

Η εκτίμηση της συνάρτησης κινδύνου (9.22) μέσω της τεχνικής leave-one-out cross validation ορίζεται

ως:

$$\text{CV} = \hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{(-i)}(x_i))^2 \tag{9.23}$$

όπου $\hat{f}_{(-i)}(x_i)$ είναι ο εκτιμητής που προκύπτει αγνοώντας την i-οστή παρατήρηση (x_i, Y_i).

Στις περιπτώσεις όπου η εκτίμηση $\hat{f}(x)$ είναι γραμμική συνάρτηση των $Y_1, \ldots, Y_n$ (όπως όλες οι εκτιμήσεις που γνωρίσαμε παραπάνω), η τιμή της $\hat{R}(h)$ μπορεί να υπολογιστεί εύκολα, όπως περιγράφεται στη συνέχεια, χωρίς να απαιτείται ο επανυπολογισμός μετά την εξαίρεση καθεμιάς από τις n παρατηρήσεις.

Θεώρημα 9.2

Έστω $\hat{f}(x) = \sum_{i=1}^{n} w_i(x) Y_i$ ένας γραμμικός εκτιμητής της συνάρτησης παλινδρόμησης $f(x)$ και ο $n \times n$ πίνακας $\boldsymbol{W}$, με $W_{ij} = w_j(x_i)$ για $i = 1, \ldots, n$ και $j = 1, \ldots, n$. Τότε, η εκτίμηση (9.23) της συνάρτησης κινδύνου $R(h)$ μέσω της τεχνικής leave-one-out cross validation ισούται με

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{f}(x_i)}{1 - W_{ii}}\right)^2, \tag{9.24}$$

όπου το $W_{ii} = w_i(x_i)$ είναι το στοιχείο στη θέση (i,i) του $n \times n$ πίνακα $\boldsymbol{W}$.

Απόδειξη Θεωρήματος 9.2. Παραπέμπουμε ενδεικτικά στο σύγγραμμα των Hastie and Tibshirani (2017), σελ. 47. □

Επομένως, αφού υπολογιστεί η $\hat{R}(h)$ μέσω της (9.24) για διαφορετικές τιμές της παραμέτρου εξομάλυνσης h, επιλέγεται η τιμή που ελαχιστοποιεί την $\hat{R}(h)$. Για περισσότερες λεπτομέρειες, παραπέμπουμε στους Härdle *et al.* (1988), Wasserman (2006).

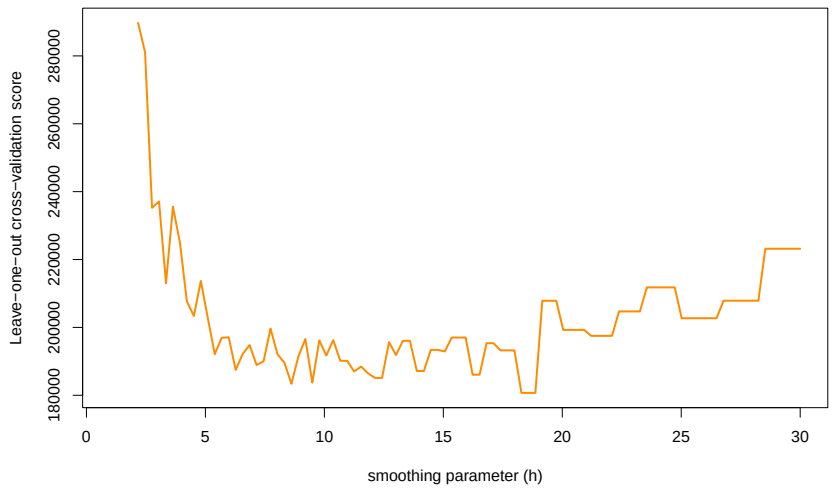
Παράδειγμα 9.7 (συνέχεια Παραδείγματος 9.2)**.** Να επιλεχθεί η παράμετρος εξομάλυνσης h για το παλινδρόγραμμα στα δεδομένα του Παραδείγματος 9.1.

Λύση Παραδείγματος 9.7. Ο παρακάτω κώδικας υπολογίζει την εκτίμηση της συνάρτησης κινδύνου (9.22) μέσω της τεχνικής leave-one-out cross validation. Για τον σκοπό αυτό θα υπολογιστεί η εκτίμηση $\hat{R}(h)$ στην (9.24) για 100 τιμές της παραμέτρου εξομάλυνσης $h \in [0, 30]$.

```
cvScore <- function(h){
  b <- max(x)
  a <- min(x)
  nBreaks <- (b-a)/h
  binnedX <- cut(x,breaks=nBreaks)
  # split y according to binnedX
  splitY <- split(y,binnedX)
  splitIndex<-split(1:n,binnedX)
  # smoothing matrix
  L <- matrix(data=0, nrow = n, ncol = n)
  for(j in 1:length(splitIndex)){
    L[splitIndex[[j]],splitIndex[[j]]] <- 1/length(splitIndex[[j]])
  }
  rg <- L % * % matrix(y)
  return(
    mean(
      ( (y-rg)/(1-diag(L)) )^2
    )
  )
```

```
}

h <- seq(1,30,length=100)
cvValues <- numeric(100)
j<-0
for(i in h){
  j<-j+1
  cvValues[j] <- cvScore(i)
}
>  h[which.min(cvValues)]
[1] 18.28283
```

Το Σχήμα 9.7 απεικονίζει το διάγραμμα των εκτιμήσεων $\hat{R}(h)$ για διαφορετικές τιμές της παραμέτρου εξομάλυνσης. Η τιμή που ελαχιστοποιεί την $\hat{R}$ ισούται με $h \approx 18.2$.



Σχήμα 9.7: Εκτίμηση της $R(h)$ στην (9.22) για διαφορετικές τιμές της παραμέτρου εξομάλυνσης h μέσω της (9.24) για το παλινδρόγραμμα στα δεδομένα CMB του Παραδείγματος 9.1.

□

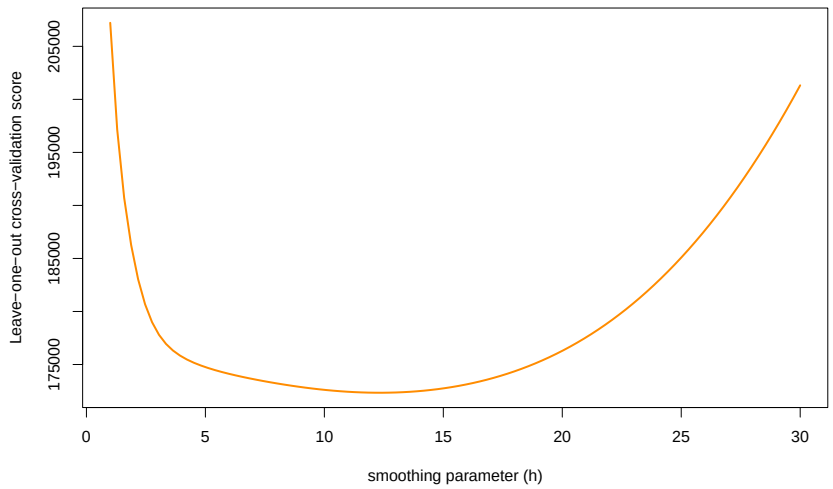
Παράδειγμα 9.8 (συνέχεια Παραδείγματος 9.3)**.** Να επιλεχθεί η παράμετρος εξομάλυνσης h για τον εκτιμητή Nadaraya-Watson με χρήση κανονικού πυρήνα στα δεδομένα του Παραδείγματος 9.1.

Λύση Παραδείγματος 9.8. Ο παρακάτω κώδικας υπολογίζει την εκτίμηση της συνάρτησης κινδύνου μέσω της τεχνικής leave-one-out cross validation. Για τον σκοπό αυτόν, θα υπολογιστεί η εκτίμηση $\hat{R}(h)$ στην (9.24) για 100 τιμές της παραμέτρου εξομάλυνσης $h \in [0,30]$.

```
cvScoreNW <- function(h,type){
  # smoothing matrix
  L <- matrix(data=0, nrow = n, ncol = n)
  for(j in 1:n){
    L[j,] <- ker((x-x[j])/h,type=type)
    L[j,] <- L[j,]/sum(L[j,])
  }
  rg <- L % * % matrix(y)
  return(
    mean(
```

```
      ( (y-rg)/(1-diag(L)) )^2
    )
  )
}


h <- seq(1,30,length=100)
cvValues <- numeric(100)
j<-0
for(i in h){
  j<-j+1
  cvValues[j] <- cvScoreNW(i,type='gaussian')
}
> h[which.min(cvValues)]
[1] 12.42424
```

Το Σχήμα 9.8 απεικονίζει το διάγραμμα των εκτιμήσεων $\hat{R}(h)$ για διαφορετικές τιμές της παραμέτρου εξομάλυνσης. Η τιμή που ελαχιστοποιεί την εκτιμηθείσα συνάρτηση κινδύνου ισούται με $h \approx 12.42$. □



Σχήμα 9.8: Εκτίμηση της $R(h)$ στην (9.22) για διαφορετικές τιμές της παραμέτρου εξομάλυνσης h μέσω της (9.24) για τον εκτιμητή Nadaraya-Watson με χρήση κανονικού πυρήνα στα δεδομένα CMB του Παραδείγματος 9.1.

Παρατήρηση 9.6. Για την επιλογή της παραμέτρου εξομάλυνσης μέσω της (9.24) για τον εκτιμητή τοπικού πολυωνύμου παραπέμπουμε στο όρισμα `bandwidth = "CV"` της εντολής `localpoly.reg(...)` της βιβλιοθήκης `NonpModelCheck`.

Παρατήρηση 9.7. Για την επιλογή της παραμέτρου εξομάλυνσης μέσω της (9.24) για εξομαλυντή spline παραπέμπουμε στο όρισμα `cv = TRUE` της εντολής `smooth.spline(...)` της βιβλιοθήκης `splines`.

9.7 Πολλαπλή μη παραμετρική παλινδρόμηση

Μέχρι τώρα έχουμε αναφερθεί στην περίπτωση μίας επεξηγηματικής μεταβλητής. Ωστόσο, οι έννοιες στις οποίες βασίζονται οι προηγούμενοι μη παραμετρικοί εκτιμητές της συνάρτησης παλινδρόμησης μπορούν να

επεκταθούν στην περίπτωση δύο ή περισσότερων επεξηγηματικών μεταβλητών (βλ. σχέση (9.1)). Όπως αναφέρθηκε και στην Ενότητα 3.5 του Κεφαλαίου 3, καθώς η διάσταση του προβλήματος αυξάνεται οι αντίστοιχοι μη παραμετρικοί εκτιμητές απαιτούν εκθετικά μεγαλύτερα μεγέθη δείγματος για να επιτύχουν παρόμοια επίπεδα ακρίβειας με τη μονοδιάστατη περίπτωση. Το πρόβλημα αυτό είναι γνωστό ως «κατάρα των μεγάλων διαστάσεων» (curse of dimensionality).

Στην πράξη, αντί να θεωρούμε μία γενική συνάρτηση παλινδρόμησης $f(x_1, \dots, x_p)$ περιορίζουμε το πρόβλημα στην ειδική περίπτωση όπου η f είναι μια **προσθετική συνάρτηση** της μορφής:

$$f(x_1, \dots, x_p) = f_1(x_1) + \dots + f_p(x_p). \tag{9.25}$$

Παρότι, λοιπόν, είναι γενικά δύσκολο να εκτιμηθεί μία πολυδιάστατη συνάρτηση παλινδρόμησης, η βέλτιστη προσθετική προσέγγιση αυτής μέσω της σχέσης (9.25) μπορεί να εκτιμηθεί εύκολα και αποδοτικά (βλ. Stone, 1985). Για μία πλήρη παρουσίαση (γενικευμένων) προσθετικών μοντέλων παραπέμπουμε στο σύγγραμμα των Hastie and Tibshirani (2017).

Η εκτίμηση (γενικευμένων) προσθετικών μοντέλων είναι πολύ απλή: αρκεί να επιλέξουμε οποιαδήποτε από τις διαθέσιμες μη παραμετρικές τεχνικές για τη μονοδιάστατη περίπτωση (π.χ. τον εκτιμητή Nadaraya-Watson, τον εξομαλυντή `spline` κ.ά.) και να εκτιμηθεί διαδοχικά καθεμία συνάρτηση f_j, $j = 1, \dots, p$, μέσω μιας διαδικασίας που ονομάζεται οπισθοδρομική προσαρμογή (backfitting Breiman and Friedman, 1985; Buja *et al.*, 1989). Για την πρακτική υλοποίηση των τεχνικών αυτών μέσω της `R` παραπέμπουμε στις βιβλιοθήκες `gam` (βασίζεται στη μεθοδολογία που περιγράφεται στο Κεφάλαιο 7 των Chambers and Hastie, 1991) και `mgcv` (Wood, 2003, 2004, 2011; Wood *et al.*, 2016; Wood, 2017).

Υπάρχουν διάφορες παραλλαγές των προσθετικών μοντέλων. Για παράδειγμα, ας θεωρήσουμε ένα προσθετικό μοντέλο της μορφής:

$$f(x_1, x_2, \dots, x_p) = f_1(x_1, x_2) + f_2(x_3) + \dots + f_{p-1}(x_p).$$

Σε αυτήν την περίπτωση, η συνάρτηση παλινδρόμησης έχει $p - 1$ προσθετικούς όρους, όπου ο πρώτος εξ αυτών είναι μία συνάρτηση που συνδυάζει τις δύο πρώτες επεξηγηματικές μεταβλητές x_1 και x_2. Μία οικογένεια συναρτήσεων που χρησιμοποιείται συχνά για τέτοιους συνδυαστικούς όρους είναι τα **thin plate splines** (Duchon, 1977) και πρόκειται για συναρτήσεις που ελαχιστοποιούν (ως προς g) την έκφραση

$$\sum_{i=1}^{n}(y_i - g(x_1, x_2))^2 + \lambda \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(g^2_{x_1,x_1}(u,v) + g^2_{x_1,x_2}(u,v) + g^2_{x_2,x_2}(u,v))\mathrm{d}u\mathrm{d}v, \tag{9.26}$$

όπου $g_{x_1,x_2}(u,v) = \left.\frac{\partial^2 g}{\partial x_1 \partial x_2}\right|_{(u,v)}$. Μπορεί να δειχθεί ότι ένα thin plate regression spline έχει λιγότερες παραμέτρους από ένα αντίστοιχο πολυδιάστατο spline και ότι μπορεί να εκτιμηθεί εύκολα. Τα thin plate regression splines μπορούν να χρησιμοποιηθούν για τη μοντελοποίηση όρων αλληλεπίδρασης μεταξύ επεξηγηματικών μεταβλητών στη μη παραμετρική παλινδρόμηση. Για περισσότερες λεπτομέρειες, ο/η αναγνώστης/στρια παραπέμπεται στον Wood (2003). Ένα παράδειγμα που αναδεικνύει την ευελιξία αυτών των μοντέλων δίνεται στη συνέχεια.

Παράδειγμα 9.9. Προσομοιώστε $n = 200$ το πλήθος παρατηρήσεις από τη συνάρτηση:

$$z = \frac{\sin(\sqrt{x^2 + y^2})}{\sqrt{x^2 + y^2}} + \epsilon,$$

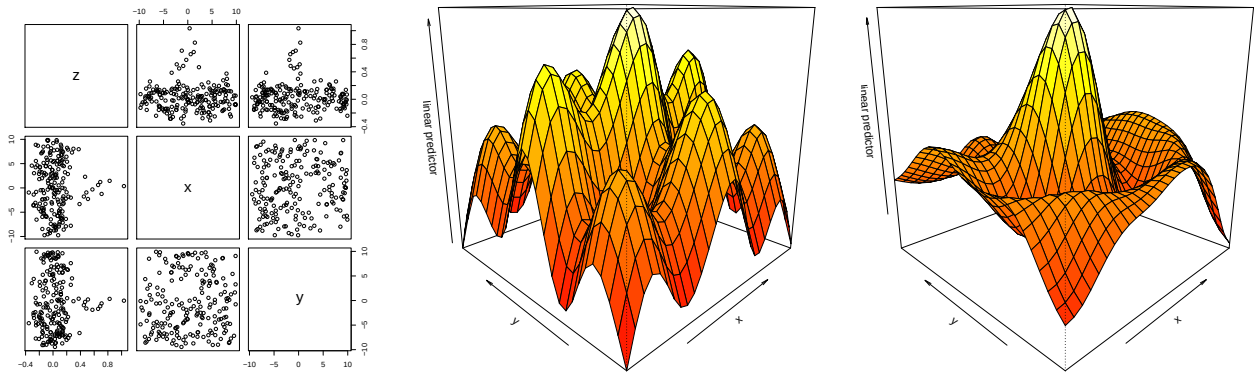
όπου $\epsilon \sim \mathcal{N}(0, 0.1^2)$. Οι τιμές των δύο επεξηγηματικών μεταβλητών (x και y) να προσομοιωθούν από ανεξάρτητες ομοιόμορφες κατανομές στο διάστημα $(-10, 10)$. Να χρησιμοποιήσετε την εντολή `gam(...)` για να εκτιμήσετε:

1. προσθετικό μοντέλο της μορφής $z = f_1(x) + f_2(y) + \epsilon$ και
2. μοντέλο με αλληλεπίδραση $z = f(x,y) + \epsilon$, όπου η $f(x,y)$ είναι μια thin plate regression spline.

Έπειτα, να οπτικοποιήσετε το αποτέλεσμα.

Λύση Παραδείγματος 9.9. Ο παρακάτω κώδικας προσομοιώνει ένα σύνολο δεδομένων μεγέθους $n = 200$ σύμφωνα με το μοντέλο που περιγράφεται στην εκφώνηση. Στο Σχήμα 9.9.(α) απεικονίζονται τα διαγράμματα διασποράς μεταξύ των προσομοιωμένων τιμών της μεταβλητής απόκρισης z και των δύο επεξηγηματικών μεταβλητών (x και y).

```
library(mgcv)
set.seed(1)
n <- 200
x <- -10 + 20*runif(n)
y <- -10 + 20*runif(n)
z <- sin(sqrt(x^2+y^2))/sqrt(x^2+y^2) + rnorm(n, sd = 0.1)
pairs(~z+x+y)
b1 <- gam(z ~ s(x) + s(y))
b2 <- gam(z ~ s(x, y))
vis.gam(b1, theta = -45)
vis.gam(b2, theta = -45)
```



(a) (b) (g)

Σχήμα 9.9: (α) Ανά δύο διαγράμματα διασποράς των δεδομένων του Παραδείγματος 9.9. Εκτίμηση προσθετικού μοντέλου (β) $z = f_1(x) + f_2(y) + \epsilon$ και (γ) μοντέλου με όρο αλληλεπίδρασης $z = f(x,y) + \epsilon$, μέσω της `gam(...)` της βιβλιοθήκης `mgcv`.

□

Κατόπιν, χρησιμοποιείται η εντολή `gam(z ~ s(x) + s(y))`, για να εκτιμήσει ένα προσθετικό μοντέλο της μορφής $z = f_1(x) + f_2(y) + \epsilon$. Στη συνέχεια, χρησιμοποιείται η εντολή `gam(z ~ s(x, y))`, για να εκτιμηθεί ένα προσθετικό μοντέλο της μορφής $z = f(x,y) + \epsilon$. Η παράμετρος εξομάλυνσης επιλέγεται μέσω μιας τροποποίησης του κριτηρίου cross validation, η οποία ονομάζεται generalized cross validation (βλ. Άσκηση 9.7). Η εκτιμηθείσα συνάρτηση παλινδρόμησης οπτικοποιείται μέσω της εντολής `vis.gam(...)` και το αποτέλεσμα παρατίθεται στο Σχήμα 9.9 (β) και 9.9 (γ). Στην περίπτωση του thin plate regression spline στο Σχήμα 9.9 (γ) η εκτιμηθείσα συνάρτηση παλινδρόμησης είναι αρκετά κοντά στην πραγματική (πειστείτε για αυτό!).

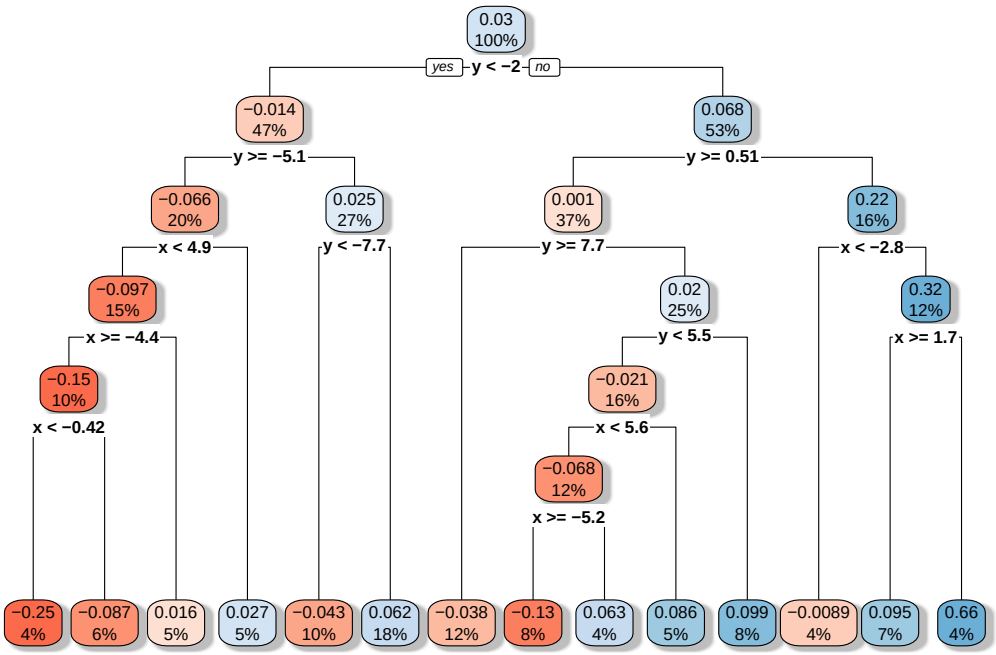
Θα πρέπει να αναφερθεί ότι, εκτός από τα προσθετικά μοντέλα, υπάρχει πληθώρα διαφορετικών μη παραμετρικών μεθοδολογίων για το πρόβλημα της πολλαπλής παλινδρόμησης. Μία αρκετά δημοφιλής

τεχνική είναι τα **δέντρα παλινδρόμησης** (Breiman *et al.*, 2017). Ένα δέντρο παλινδρόμησης είναι ένα μοντέλο της μορφής $y = f(\boldsymbol{x}) + \epsilon$, με

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} c_m I(\boldsymbol{x} \in R_m),$$

όπου $c_1, \ldots, c_M$ σταθερές και $R_1, \ldots, R_M$ μία διαμέριση (που αποτελείται από ορθογώνια) του χώρου των επεξηγηματικών μεταβλητών $\boldsymbol{x} = (x_1, \ldots, x_p)$. Για αυτές τις τεχνικές παραπέμπουμε τον/την αναγνώστη/στρια στο Κεφάλαιο 8 του συγγράμματος των James *et al.* (2013). Τα δέντρα παλινδρόμησης μπορούν να εφαρμοστούν στην `R` μέσω των βιβλιοθηκών `tree` (Ripley, 2021) και `rpart` (Therneau and Atkinson, 2019). Στη συνέχεια, δίνουμε ένα παράδειγμα για την εκτίμηση ενός δέντρου παλινδρόμησης μέσω της `R`.

Παράδειγμα 9.10 (συνέχεια Παραδείγματος 9.9)**.** Να προσαρμόσετε ένα δέντρο παλινδρόμησης στα δεδομένα του Παραδείγματος 9.9.



Σχήμα 9.10: Δέντρο παλινδρόμησης στα προσομοιωμένα δεδομένα του Παραδείγματος 9.9.

Λύση Παραδείγματος 9.10. Ο παρακάτω κώδικας εκτιμά το δέντρο παλινδρόμησης στα δεδομένα του Παραδείγματος 9.9.

```
library("rpart")
library("rpart.plot")
myTree <- rpart(z ~ x+y)
rpart.plot(myTree, box.palette="RdBu", shadow.col="gray")
```

Το δέντρο παλινδρόμησης απεικονίζεται στο Σχήμα 9.10. Οι τιμές σε κάθε κουτάκι της βάσης του Σχήματος 9.10 δίνουν τους εκτιμηθέντες συντελεστές $c_1, \ldots, c_M$, για καθένα από τα $M = 14$ ορθογώνια της διαμέρισης του επιπέδου των επεξηγηματικών μεταβλητών x, y. Για παράδειγμα, αν $-5.1 \leqslant y < -2$ και $-4.4 \leqslant x < -0.42$, τότε $\hat{f}(x, y) = c_1 = -0.25$. Οπότε, το πρώτο ορθογώνιο της διαμέρισης είναι το

$$R_1 = [-4.4, -0.42) \times [-5.1, -2).$$

Στο πρώτο κουτάκι, επίσης, αναγράφεται το ποσοστό 4% και παριστάνει τη σχετική συχνότητα των παρατηρήσεων του συνόλου των δεδομένων μας που ανήκουν στο συγκεκριμένο υποσύνολο του χώρου των επεξηγηματικών μεταβλητών (8 από τις 200 παρατηρήσεις). Η τιμή $c_1 = \hat{f}(x, y) = -0.25$ αντιστοιχεί στον δειγματικό μέσο της μεταβλητής απόκρισης για τις 8 παρατηρήσεις που ανήκουν στο συγκεκριμένο ορθογώνιο, δηλαδή

$$c_1 = \frac{1}{||R_1||} \sum_{z_i:(x_i,y_i)\in R_1} z_i,$$

όπου $||R_1|| = 8$. Με παρόμοιο τρόπο ερμηνεύονται οι τιμές στα επόμενα κουτάκια. □

Ένα από τα μειονεκτήματα των δέντρων παλινδρόμησης είναι ότι δεν μοντελοποιούν εύκολα κύριες επιδράσεις επεξηγηματικών μεταβλητών, ενώ συνήθως έχουν την τάση να υπερ-προσαρμόζονται στα δεδομένα (overfitting). Για αυτούς του λόγους έχουν προταθεί ποικίλες γενικεύσεις/τροποποιήσεις, όπως τα μοντέλα MARS (Multivariate Adaptive Regression Splines) (Friedman, 1991) και τα τυχαία δάση (random forests) (Ho, 1995; Breiman, 2001).

9.8 Ανθεκτική γραμμική παλινδρόμηση

Στην ενότητα αυτή, θα παρουσιαστεί η πιο δημοφιλής μέθοδος για την εκτίμηση γραμμικής τάσης. Η μέθοδος αυτή αποτελεί ουσιαστικά το μη παραμετρικό ανάλογο της απλής γραμμικής παλινδρόμησης, δηλαδή της παλινδρόμησης κατά την οποία υπάρχει μία εξαρτημένη και μία ανεξάρτητη μεταβλητή. Επιπλέον, στη βιβλιογραφία είναι γνωστή είτε ως Sen-Theil είτε ως Kendall-Theil παλινδρόμηση, ενώ αναφέρεται και ως single median μέθοδος. Οι παραπάνω ονομασίες είναι πλήρως αιτιολογημένες, καθώς ο Ολλανδός οικονομέτρης Henri Theil (1924 –2000), αρχικά, πρότεινε έναν τρόπο εκτίμησης της κλίσης μιας ευθείας παλινδρόμησης (βλ. Theil, 1950), ο οποίος, στη συνέχεια, επεκτάθηκε από τον Ινδό στατιστικό Pranab Kumar Sen (1937-), στην εργασία του Sen (1968), αντίστοιχα. Τέλος, η σύνδεση αυτού του εκτιμητή με τον συντελεστή συσχέτισης του Kendall, αιτιολογεί την προσθήκη του ονόματος του Βρετανού στατιστικού Sir Maurice George Kendall (1907–1983).

Έστω το μοντέλο απλής γραμμικής παλινδρόμησης

$$Y = \alpha + \beta X + \epsilon,$$

όπου α και β είναι οι άγνωστες παράμετροι και ϵ η τ.μ. που παριστάνει το σφάλμα, με άγνωστη αθροιστική συνάρτηση κατανομής F η οποία έχει διάμεσο ίση με το 0. Επιπρόσθετα, υποθέτουμε ότι είναι διαθέσιμα n το πλήθος ζεύγη παρατηρήσεων, (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), όπου $x_i, y_i \in \mathbb{R}$, με τα x_i να είναι γνωστές μη ταυτόσημες σταθερές.

Θέλοντας να εκτιμήσει την παράμετρο β, ο Theil (1950) υπέθεσε ότι όλες οι τιμές x_i, $i = 1, ..., n$ είναι διακεκριμένες και πρότεινε την ακόλουθη διαδικασία:

1. Για κάθε ζεύγος (x_i, y_i) και (x_j, y_j), με $1 \leqslant i < j \leqslant n$, υπολογίζουμε την αντίστοιχη κλίση:

$$S_{ij} = \frac{y_j - y_i}{x_j - x_i}. \tag{9.27}$$

   Επομένως, υπολογίζονται $\binom{n}{2}$ το πλήθος τιμές S_{ij}.
2. Ο εκτιμητής $\tilde{\beta}$ της παραμέτρου β ορίζεται να είναι η διάμεσος των τιμών, δηλαδή

$$\tilde{\beta} = median\left\{S_{ij},\ 1 \leqslant i < j \leqslant n\right\}. \tag{9.28}$$

Καθώς η παραπάνω διαδικασία εκτίμησης βασίζεται στη μη ρεαλιστική υπόθεση ότι οι τιμές των x_i είναι διακεκριμένες, επεκτάθηκε από τον Sen (1968), έτσι ώστε να μπορεί να χρησιμοποιηθεί και σε περιπτώσεις όπου παραβιάζεται αυτή η υπόθεση. Ειδικότερα, ο Sen (1968) πρότεινε να προσδιορίζεται ο (τροποποιημένος) εκτιμητής $\tilde{\beta}_{mod}$ της παραμέτρου β από τη σχέση:

$$\tilde{\beta}_{mod} = median(S_0), \tag{9.29}$$

όπου

$$S_0 = \left\{ S_{ij} : S_{ij} = \frac{y_j - y_i}{x_j - x_i}, \text{ αν } x_i \neq x_j, 1 \leqslant i < j \leqslant n \right\}.$$

Επομένως, ο εκτιμητής, σε αυτήν την περίπτωση, ισούται με τη διάμεσο των δειγματικών κλίσεων που μπορούν να οριστούν.

Παρατήρηση 9.8. Ο εκτιμητής $\tilde{\beta}$ της σχέσης (9.28) είναι αυτός που είναι γνωστός ως εκτιμητής των *Theil–Sen*. Είναι ανθεκτικός στην ύπαρξη ακραίων τιμών στις τιμές της εξαρτημένης μεταβλητής, υπολογίζεται εύκολα και η ασυμπτωτική κατανομή του (υπό συγκεκριμένες υποθέσεις για την κατανομή των σφαλμάτων) είναι η κανονική κατανομή (βλ. Sen, 1968). Επιπρόσθετα, η ασυμπτωτική σχετική αποτελεσματικότητά του (asymptotic relative efficiency, ARE) σε σύγκριση με τον εκτιμητή ελαχίστων τετραγώνων έχει μελετηθεί από τους Sen (1968) και Wilcox (1998), από όπου, μεταξύ άλλων, προκύπτει ότι, αν η αθροιστική συνάρτηση κατανομής των σφαλμάτων ϵ είναι κανονική, τότε $ARE = 3/\pi = 0.955$, αν είναι η λογιστική κατανομή, τότε είναι $ARE > 1$, ενώ για οποιαδήποτε συνεχή αθροιστική συνάρτηση κατανομής F είναι $ARE \geqslant 0.864$. Τέλος, υπό συνθήκες ομαλότητας, ο εκτιμητής $\tilde{\beta}_{mod}$ είναι αμερόληπτος εκτιμητής της παραμέτρου β (βλ. Wang and Yu, 2005, και τις εκεί αναφορές).

Για την εκτίμηση του σταθερού όρου α, ο εκτιμητής που προτάθηκε από τους Hettmansperger *et al.* (1997) δίνεται από τη σχέση:

$$\tilde{\alpha} = median\{y_i - \tilde{\beta} \cdot x_i\}. \tag{9.30}$$

Προφανώς, χρησιμοποιώντας τους παραπάνω εκτιμητές, μπορεί άμεσα να βρεθεί η εκτιμώμενη εξίσωση παλινδρόμησης.

Παρατήρηση 9.9. Για την εκτίμηση του σταθερού όρου έχει προταθεί και η παραλλαγή που προτείνει να προσδιορίζεται ο εκτιμητής του σταθερού όρου από τη σχέση:

$$\tilde{\alpha} = median\{y_i\} - \tilde{\beta} \cdot median\{x_i\}. \tag{9.31}$$

Η τελευταία παραλλαγή προτάθηκε, έτσι ώστε η εκτιμώμενη ευθεία παλινδρόμησης να διέρχεται από τη διάμεσο των παρατηρήσεων σε πλήρη αντιστοίχιση με την ευθεία ελαχίστων τετραγώνων που διέρχεται από το μέσο των παρατηρήσεων.

Παράδειγμα 9.11. Χρησιμοποιώντας τα δεδομένα του Πίνακα 9.1 (βλ. Sen, 1968) υπολογίστε τους εκτιμητές των παραμέτρων της απλής παλινδρόμησης με τη μέθοδο των Theil-Sen. Για τις πράξεις χρησιμοποιήστε (όπου είναι απαραίτητο) ακρίβεια 3 δεκαδικών ψηφίων.

x_i	1	2	3	4	10	12	18
y_i	9	15	19	20	45	55	78

Πίνακας 9.1: Δεδομένα εφαρμογής απλής παλινδρόμησης με τη μέθοδο των Theil-Sen.

Λύση Παραδείγματος 9.11. Αρχικά, παρατηρούμε ότι δεν υπάρχουν δεσμοί στις τιμές των x_i, $i = 1,...,7$. Θα υπολογιστούν συνολικά $\binom{7}{2} = 21$ τιμές S_{ij}, $1 \leqslant i < j \leqslant n$, από τη σχέση (9.27). Είναι

$$S_{12} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{15-9}{2-1} = 6, \;\; S_{13} = \frac{y_3 - y_1}{x_3 - x_1} = \frac{19-9}{3-1} = 5,$$

$$S_{14} = \frac{y_4 - y_1}{x_4 - x_1} = \frac{20-9}{4-1} = \frac{11}{3} = 3.667, \;\; S_{15} = \frac{y_5 - y_1}{x_5 - x_1} = \frac{45-9}{10-1} = 4,$$

$$S_{16} = \frac{y_6 - y_1}{x_6 - x_1} = \frac{55-9}{12-1} = \frac{46}{11} = 4.182, \;\; S_{17} = \frac{y_7 - y_1}{x_7 - x_1} = \frac{78-9}{18-1} = \frac{69}{17} = 4.059.$$

Με παρόμοιο τρόπο, έχουμε ότι:

$$S_{23} = 4,\; S_{24} = 2.5,\; S_{25} = 30/8 = 3.75,\; S_{26} = 4,\; S_{27} = 63/16 = 3.937,$$

$$S_{34} = 1,\; S_{35} = 26/7 = 3.714,\; S_{36} = 4,\; S_{37} = 59/15 = 3.933,\; S_{45} = 25/6 = 4.167,$$

$$S_{46} = 35/8 = 4.375,\; S_{47} = 58/14 = 4.143,\; S_{56} = 5,\; S_{57} = 33/8 = 4.125,\; S_{67} = 23/6 = 3.833.$$

Στη συνέχεια, διατάσσουμε τις 21 τιμές αυτές, έτσι ώστε να υπολογίσουμε εύκολα τη διάμεσό τους. Το διατεταγμένο δείγμα είναι το:

$$1,\; 2.5,\; 11/3,\; 30/8,\; 26/7,\; 23/6,\; 59/15,\; 63/16,\; 4,\; 4,\; 4,\; 4,\; 69/17,\; 33/8,\; 58/14,\; 25/6,$$

$$46/11,\; 35/8,\; 5,\; 5,\; 6,\; 6,\; 7.$$

Καθώς η διάμεσος των παραπάνω τιμών είναι η τιμή της 11ης παρατήρησης, προκύπτει ότι είναι ίση με 4 και, άρα, $\tilde{\beta} = 4$. Επιπλέον, αφού η διάμεσος των τιμών x_i είναι 4, και αντίστοιχα των τιμών y_i είναι ίση με 20, με εφαρμογή της σχέσης (9.31) έπεται ότι:

$$\tilde{\alpha}_1 = median\{y_i\} - \tilde{\beta}_1 \cdot median\{x_i\} = 20 - 4 \cdot 4 = 4,$$

ενώ με εφαρμογή της σχέσης (9.30) είναι:

$$\tilde{\alpha} = median\{y_i - \tilde{\beta} \cdot x_i\} = median\{5,7,7,4,5,7,6\} = 6.$$

□

9.8.1 Έλεγχοι υποθέσεων στην ανθεκτική γραμμική παλινδρόμηση

Έπειτα από την εκτίμηση των παραμέτρων, το ενδιαφέρον επικεντρώνεται στον έλεγχο της μηδενικής υπόθεσης $H_0 : \beta = \beta_0$. Σε όσα ακολουθούν υποθέτουμε (βλ. Hollander *et al.*, 2014) ότι οι τιμές στην ανεξάρτητη μεταβλητή είναι διακεκριμένες και χωρίς βλάβη της γενικότητας ότι $x_1 < ... < x_n$, ενώ τα σφάλματα προέρχονται από συνεχή πληθυσμό με διάμεσο ίση με μηδέν. Επίσης, θα συμβολίζουμε $D_i = y_i - \beta_0 x_i$, $i = 1,...,n$.

Στο παραπάνω πλαίσιο, η στατιστική συνάρτηση που προτάθηκε από τον Theil (1950) για τον έλεγχο της $H_0 : \beta = \beta_0$ δίνεται από τη σχέση:

$$C = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c(D_j - D_i), \tag{9.32}$$

όπου

$$c(a) = \begin{cases} -1, & \text{αν } a < 0 \\ 0, & \text{αν } a = 0 \\ 1, & \text{αν } a > 0. \end{cases}$$

Επομένως, για κάθε ζεύγος δεικτών (i,j), με $1 \leqslant i < j \leqslant n$, το σκορ $c(D_j - D_i)$ είναι 1 (αντίστοιχα, –1), αν η διαφορά $D_j - D_i$ είναι θετική (αντίστοιχα, αρνητική), και 0, αν η διαφορά $D_j - D_i$ είναι ίση με 0. Επομένως, η στατιστική συνάρτηση είναι το άθροισμα τιμών 1 και –1 (το πλήθος τους καθορίζεται από το πόσες θετικές και πόσες αρνητικές διαφορές $D_j - D_i$ παρατηρούνται στα δεδομένα).

Παρατηρούμε, τώρα, ότι υπό τη μηδενική υπόθεση $H_0 : \beta = \beta_0$ (βλ. Hollander *et al.*, 2014)

$$\begin{aligned} D_j - D_i &= y_j - y_i + \beta_0(x_i - x_j) \\ &= \alpha + \beta x_j - \alpha - \beta x_i + \epsilon_j - \epsilon_i + \beta_0(x_i - x_j) \\ &= (\beta - \beta_0)(x_j - x_i) + \epsilon_j - \epsilon_i. \end{aligned}$$

Λαμβάνοντας υπόψη ότι έχουμε υποθέσει ότι η διάμεσος της κατανομής των σφαλμάτων είναι ίση με μηδέν και ότι $x_1 < \ldots < x_n$, δηλαδή υποθέτοντας, επιπλέον, ότι δεν υπάρχουν δεσμοί στις τιμές $x_1, \ldots, x_n$, προκύπτει ότι απορρίπτεται η μηδενική υπόθεση $H_0 : \beta = \beta_0$ έναντι της εναλλακτικής $H_1 : \beta > \beta_0$ ($H_1 : \beta < \beta_0$), για μεγάλες (μικρές, αντίστοιχα) τιμές της στατιστικής συνάρτησης C. Η μορφή της κρίσιμης περιοχής αιτιολογείται πλήρως, καθώς όταν $\beta > \beta_0$ ($\beta < \beta_0$, αντίστοιχα), τότε υπάρχει η τάση να παρατηρούμε θετικές (αρνητικές) διαφορές, γεγονός που οδηγεί σε μεγάλες (μικρές, αντίστοιχα) τιμές της στατιστικής συνάρτησης C. Προφανώς, στην περίπτωση δίπλευρου ελέγχου, η μηδενική υπόθεση απορρίπτεται είτε για μεγάλες είτε για μικρές τιμές της στατιστικής συνάρτησης C.

Για τον καθορισμό του ποια τιμή της στατιστικής συνάρτησης C θεωρείται μικρή και ποια μεγάλη θα πρέπει να προσδιοριστεί η κατανομή της υπό τη μηδενική υπόθεση. Μπορούμε να διαπιστώσουμε ότι η στατιστική συνάρτηση C, που δίνεται στη σχέση (9.32), δεν είναι τίποτε άλλο παρά ο συντελεστής συσχέτισης του Kendall υπολογισμένος με τις τιμές x και $y - \beta_0 x$. Αυτό σημαίνει ότι υπό τη μηδενική υπόθεση και υποθέτοντας ότι δεν υπάρχουν ισοβαθμίες, η κατανομή της στατιστικής συνάρτησης C ταυτίζεται με την κατανομή της στατιστικής συνάρτησης K του Kendall, που παρουσιάστηκε στο προηγούμενο κεφάλαιο και για τη διενέργεια ελέγχου σε επίπεδο σημαντικότητας a χρησιμοποιείται ο Πίνακας Π.30 του παραρτήματος, δηλαδή ο πίνακας ποσοστιαίων σημείων του στατιστικού τεστ του Kendall. Επομένως:

(Α) Απορρίπτεται η μηδενική υπόθεση $H_0 : \beta = \beta_0$ έναντι της εναλλακτικής $H_0 : \beta > \beta_0$, όταν η τιμή της στατιστικής συνάρτησης C ξεπερνά το $1 - a$ ποσοστιαίο σημείο του, δηλαδή αν η τιμή της στατιστικής συνάρτησης C είναι μεγαλύτερη από την τιμή του Πίνακα Π.30 για $p = 1 - a$.

(Β) Απορρίπτεται η μηδενική υπόθεση $H_0 : \beta = \beta_0$ έναντι της εναλλακτικής $H_0 : \beta < \beta_0$, όταν η τιμή της στατιστικής συνάρτησης C δεν ξεπερνά το a ποσοστιαίο σημείο του ή, ισοδύναμα, λόγω συμμετρίας, αν τη τιμή της στατιστικής συνάρτησης C είναι μικρότερη από την αντίθετη της τιμής του Πίνακα Π.30 για $p = 1 - a$.

(Γ) Απορρίπτεται η μηδενική υπόθεση $H_0 : \beta = \beta_0$ έναντι της εναλλακτικής $H_0 : \beta \neq \beta_0$, είτε όταν η τιμή της στατιστικής συνάρτησης C ξεπερνά το $1 - a/2$ ποσοστιαίο σημείο του είτε όταν η τιμή της στατιστικής συνάρτησης T δεν ξεπερνά το $a/2$ ποσοστιαίο σημείο του. Επομένως, λαμβάνοντας υπόψη τη συμμετρία της κατανομής της στατιστικής συνάρτησης C, η μηδενική υπόθεση απορρίπτεται, αν η τιμή της στατιστικής συνάρτησης C είναι μεγαλύτερη κατά απόλυτη τιμή από την τιμή του Πίνακα Π.30 για $p = 1 - a/2$.

Τέλος, για μεγάλο μέγεθος δείγματος αποδεικνύεται ότι υπό τη μηδενική υπόθεση:

$$Z = \frac{C}{\sqrt{n(n-1)(2n+5)/18}} \xrightarrow{d} \mathcal{N}(0,1).$$

Στη συνέχεια, δίνεται ένα παράδειγμα για να γίνει πιο κατανοητή η διαδικασία ελέγχου υποθέσεων που μόλις περιγράψαμε.

Παράδειγμα 9.12. Χρησιμοποιώντας τα δεδομένα του Πίνακα 9.2 (βλ. Hollander *et al.*, 2014) ελέγξτε τη μηδενική υπόθεση $H_0 : \beta = 0$ έναντι της εναλλακτικής $H_1 : \beta \neq 0$. Ο έλεγχος να γίνει σε επίπεδο σημαντικότητας 5%.

x_i	1	2	3	4	5
y_i	1.26	1.27	1.12	1.16	1.03

Πίνακας 9.2: Δεδομένα εφαρμογής απλής παλινδρόμησης και ελέγχων υποθέσεων (πηγή: Hollander *et al.*, 2014).

Λύση Παραδείγματος 9.12. Θέλουμε να ελέγξουμε, σε επίπεδο σημαντικότητας a, τη μηδενική υπόθεση $H_0 : \beta = 0$ έναντι της εναλλακτικής $H_1 : \beta \neq 0$. Σύμφωνα με τη θεωρία που προηγήθηκε, η μηδενική υπόθεση απορρίπτεται, σε επίπεδο σημαντικότητας 5%, αν η τιμή της στατιστικής συνάρτησης C είναι κατά απόλυτη τιμή μεγαλύτερη από την τιμή 8, καθώς από τον Πίνακα Π.30 το ποσοστιαίο 0.975 σημείο είναι ίσο με 8 για $n = 5$. Στη συνέχεια, θα υπολογιστεί η τιμή της στατιστικής συνάρτησης C.

Αρχικά, παρατηρούμε ότι είναι $\beta_0 = 0$ και, επομένως, σε αυτό το παράδειγμα οι διαφορές D_i, $i = 1,...,5$, ταυτίζονται με τις τιμές y_i. Στη συνέχεια, υπολογίζουμε τις $\binom{n}{2} = \binom{5}{2} = 10$ το πλήθος διαφορές $D_j - D_i$:

$$D_2 - D_1 = 0.01, D_3 - D_1 = -0.14, D_4 - D_1 = -0.1, D_5 - D_1 = -0.23, D_3 - D_2 = -0.15,$$

$$D_4 - D_2 = -0.11, D_5 - D_2 = -0.24, D_4 - D_3 = 0.04, D_5 - D_3 = -0.09, D_5 - D_4 = -0.13.$$

Επομένως, έχουμε 2 το πλήθος θετικές διαφορές και 8 το πλήθος αρνητικές διαφορές, οπότε $C = 2 - 8 = -6$ και η μηδενική υπόθεση δεν απορρίπτεται σε επίπεδο σημαντικότητας 5%, καθώς $C = -6 \not< -8$ και $C = -6 \not> 8$

Τα παραπάνω θα μπορούσαν να υλοποιηθούν στη γλώσσα προγραμματισμού `R` χρησιμοποιώντας τις ακόλουθες εντολές:

```
> library(NSM3)
> x<-c(1,2,3,4,5)
> y<-c(1.26,1.27,1.12,1.16,1.03)
> theil(x,y,beta.0=0,type='t')
```

Αρχικά, φορτώνουμε το πακέτο `NSM3` και χρησιμοποιούμε την εντολή `theil(...)` εισάγοντας τα διανύσματα `x`, `y` στα οποία έχουμε καταχωρίσει τις τιμές από τον Πίνακα 9.2. Χρησιμοποιούμε το όρισμα `type='t'`, αν θέλουμε να κάνουμε δίπλευρο έλεγχο (όπου `'t'` είναι για two-sided). Αν θέλαμε να κάνουμε τον έλεγχο με εναλλακτική την $H_1 : \beta < \beta_0$, θα έπρεπε να χρησιμοποιήσουμε `type='l'`, ενώ, αν η εναλλακτική ήταν η $H_1 : \beta > \beta_0$, θα έπρεπε να χρησιμοποιήσουμε `type='u'`. Το αποτέλεσμα της ανάλυσης δίνεται παρακάτω:

```
Alternative: beta not equal to 0
C = -6, C.bar = -0.6, P = 0.233
beta.hat = -0.056
alpha.hat = 1.316

1 - alpha = 0.95 two-sided CI for beta:
-0.15, 0.04
```

Παρατηρούμε ότι η τιμή της στατιστικής συνάρτησης ελέγχου C είναι ίση με -6 (όπως την υπολογίσαμε «με το χέρι»), ενώ η τιμή `C.bar` δεν είναι παρά ο συντελεστής συσχέτισης του Kendall για τα δεδομένα του Πίνακα 9.2. Η p-τιμή του ελέγχου είναι 0.233 και σε ε.σ. 5% δεν απορρίπτουμε τη μηδενική υπόθεση $H_0 : \beta = 0$. Από το output της ανάλυσης δίνονται, επίσης, οι εκτιμήσεις $\tilde{\alpha} = 1.316$, $\tilde{\beta} = -0.056$ των παραμέτρων α, β του μοντέλου απλής γραμμικής παλινδρόμησης. □

9.9 Ασκήσεις

Άσκηση 9.1 (Auto-Mpg Data)**.** Θεωρήστε το σύνολο δεδομένων Auto-Mpg Data το οποίο είναι διαθέσιμο στον σύνδεσμο https://archive.ics.uci.edu/ml/datasets/auto+mpg. Ενδιαφέρει η μελέτη του δείκτη κατανάλωσης καυσίμων (mpg: μεταβλητή απόκρισης) σε σχέση με το βάρος (wt: επεξηγηματική μεταβλητή) του αυτοκινήτου. Να υπολογίσετε το παλινδρόγραμμα για την παλινδρόμηση της mpg στη wt. Η παράμετρος εξομάλυνσης να επιλεχθεί με βάση το κριτήριο leave-one-out cross validation. Να αναπαραστήσετε γραφικά την εκτιμηθείσα συνάρτηση παλινδρόμησης.

Άσκηση 9.2 (Auto-Mpg Data)**.** Θεωρήστε τα δεδομένα της Άσκησης 9.1. Να υπολογίσετε τον εκτιμητή Nadaraya-Watson (με χρήση κανονικού πυρήνα) για την παλινδρόμηση της mpg στη wt. Η παράμετρος εξομάλυνσης να επιλεχθεί με βάση το κριτήριο leave-one-out cross validation. Να αναπαραστήσετε γραφικά την εκτιμηθείσα συνάρτηση παλινδρόμησης.

Άσκηση 9.3 (Auto-Mpg Data)**.** Να υπολογίσετε τους βαθμούς ελευθερίας (βλ. Παρατήρηση 9.2) των εξομαλυντών στις Ασκήσεις 9.1 και 9.2.

Άσκηση 9.4 (Auto-Mpg Data)**.** Θεωρήστε τα δεδομένα της Άσκησης 9.1. Να υπολογίσετε τον εκτιμητή τοπικής πολυωνυμικής (κυβικής) παλινδρόμησης για την παλινδρόμηση της mpg στη wt. Η παράμετρος εξομάλυνσης να επιλεχθεί με βάση το κριτήριο leave-one-out cross validation. Να αναπαραστήσετε γραφικά την εκτιμηθείσα συνάρτηση παλινδρόμησης.

Άσκηση 9.5 (Auto-Mpg Data)**.** Θεωρήστε τα δεδομένα της Άσκησης 9.1. Να υπολογίσετε τον εκτιμητή της συνάρτησης παλινδρόμησης μέσω εξομαλυντή spline για την παλινδρόμηση της mpg στη wt. Η παράμετρος εξομάλυνσης να επιλεχθεί με βάση το κριτήριο leave-one-out cross validation. Να αναπαραστήσετε γραφικά την εκτιμηθείσα συνάρτηση παλινδρόμησης.

Άσκηση 9.6 (Εκτίμηση διασποράς σφαλμάτων)**.** Έστω το απλό γραμμικό μοντέλο (9.2) και ας υποθέσουμε ότι τα σφάλματα έχουν σταθερή διασπορά, δηλαδή: $\text{Var}(\epsilon_i) = \sigma^2$, για κάθε $i = 1, \dots, n$. Έστω ότι $\hat{f}(x)$ είναι ένας γραμμικός εξομαλυντής της μορφής (9.4) και $\boldsymbol{W}$ ο $n \times n$ πίνακας εξομάλυνσης, με στοιχεία $W_{ij} = w_j(x_i)$, $i, j = 1, \dots, n$. Υπό κάποιες συνθήκες ομαλότητας, μπορεί να δειχθεί ότι ο

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{f}(x_i))^2}{n - 2\nu + \tilde{\nu}},$$

είναι ένας συνεπής εκτιμητής της διασποράς των σφαλμάτων (βλ. Wasserman, 2006, Ενότητα 5.6), όπου $\nu = \text{tr}(\boldsymbol{W})$ (οι βαθμοί ελευθερίας του εξομαλυντή, βλ. Παρατήρηση 9.2) και $\tilde{\nu} = \text{tr}(\boldsymbol{W}^\top \boldsymbol{W})$. Να υπολογιστεί η εκτίμηση της διασποράς των σφαλμάτων στα δεδομένα του Παραδείγματος 9.1 χρησιμοποιώντας:

1. το παλινδρόγραμμα, και
2. τον εκτιμητή Nadaraya-Watson με χρήση κανονικού πυρήνα.

Υπόδειξη: Θεωρήστε ότι η παράμετρος εξομάλυνσης του γραμμικού εξομαλυντή επιλέγεται βάσει της τεχνικής leave-one-out cross validation.

Άσκηση 9.7 (Generalized cross-validation)**.** Να επιλέξετε την παράμετρο εξομάλυνσης h στα δεδομένα του Παραδείγματος 9.1 ελαχιστοποιώντας το κριτήριο γενικευμένης cross validation το οποίο ισούται με:

$$\text{GCV}(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i - \hat{f}(x_i)}{1 - \nu/n} \right)^2,$$

όπου $\nu = \text{tr}(W)$ (οι βαθμοί ελευθερίας του εξομαλυντή, βλ. Παρατήρηση 9.2).

Άσκηση 9.8 (CMB data)**.** Θεωρήστε το πλήρες ($n = 899$ το πλήθος παρατηρήσεις) σύνολο δεδομένων κοσμικής ακτινοβολίας υποβάθρου, το οποίο είναι διαθέσιμο στον σύνδεσμο http://www.stat.cmu.edu/~larry/all-of-nonpar/=data/wmap.dat. Να εκτιμήσετε τη συνάρτηση παλινδρόμησης κάνοντας χρήση παλινδρογράμματος, εκτιμητή Nadaraya-Watson, τοπικού πολυωνύμου και εξομαλυντή spline. Σε κάθε περίπτωση χρησιμοποιήστε το κριτήριο leave-one-out cross validation για την επιλογή της παραμέτρου εξομάλυνσης.

Άσκηση 9.9. Θεωρήστε το σύνολο δεδομένων `trees` που είναι διαθέσιμο στην `R`, το οποίο περιέχει τις εξής μεταβλητές

- `Girth`: διάμετρος σε ίντσες,
- `Height`: ύψος σε πόδια,
- `Volume`: όγκος σε κυβικά πόδια,

για $n = 31$ το πλήθος δέντρα. Σκοπός είναι η περιγραφή του όγκου σε σχέση με τη διάμετρο και το ύψος.

1. Προσαρμόστε ένα προσθετικό μοντέλο με χρήση της εντολής `gam` της βιβλιοθήκης `mgcv`. Κατόπιν να οπτικοποιήσετε το μοντέλο που εκτιμήσατε, καθώς και τα κατάλοιπα αυτού.
2. Προσαρμόστε ένα δέντρο παλινδρόμησης με χρήση της εντολής `rpart` της βιβλιοθήκης `rpart`. Κατόπιν, να οπτικοποιήσετε το μοντέλο που εκτιμήσατε.
3. Χρησιμοποιήστε την εντολή `predict()` για να προβλέψετε τον όγκο ενός δέντρου με `Girth = 15` και `Height = 70`, με βάση τα δύο προηγούμενα μοντέλα.

Υπόδειξη: Θεωρήστε ως μεταβλητή απόκρισης τον λογάριθμο του όγκου.

Άσκηση 9.10. Τα δεδομένα στον Πίνακα 9.3 καταγράφουν τον αριθμό διεθνών τηλεφωνικών κλήσεων σε δεκάδες εκατομμύρια (y) ανά χρονιά $x = 50, 51, \ldots, 73$ στο Βέλγιο[3]. Χρησιμοποιήστε τα διαθέσιμα δεδομένα και εκτιμήστε ένα γραμμικό μοντέλο για την παλινδρόμηση της y στη x με την τεχνική των Kendall-Theil και συγκρίνετε αυτά τα αποτελέσματα με εκείνα που προκύπτουν με τη μέθοδο ελαχίστων τετραγώνων.

x	y	x	y	x	y	x	y
50	0.44	56	0.81	62	1.61	68	18.2
51	0.47	57	0.88	63	2.12	69	21.2
52	0.47	58	1.06	64	11.9	70	4.30
53	0.59	59	1.20	65	12.4	71	2.40
54	0.66	60	1.35	66	14.2	72	2.70
55	0.73	61	1.49	67	15.9	73	2.90

Πίνακας 9.3: Δεδομένα διεθνών τηλεφωνικών κλήσεων. Πηγή: Rousseeuw and Yohai (1984).

Άσκηση 9.11. Θεωρήστε τα δεδομένα της Άσκησης 9.1 και εκτιμήστε ένα γραμμικό μοντέλο για την παλινδρόμηση της y(=mpg) στη x(=wt) με την τεχνική των Kendall-Theil. Να συγκρίνετε τα αποτελέσματα με εκείνα που προέκυψαν με τη μέθοδο του παλινδρογράμματος.

[3] Είναι γνωστό ότι οι συγκεκριμένες μετρήσεις δεν είναι αξιόπιστες, διότι κατά τις χρονιές 1964-1969 χρησιμοποιήθηκε διαφορετικό σύστημα καταγραφής, το οποίο μετρούσε τον συνολικό αριθμό λεπτών αυτών των κλήσεων αντί του συνολικού αριθμού αυτών. Παρά ταύτα, τα δεδομένα δίνονται για εκπαιδευτικούς σκοπούς και εξάσκηση στην εκτίμηση μοντέλων ανθεκτικής γραμμικής παλινδρόμησης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ξενόγλωσση

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), pp. 5–32.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, *80*(391), pp. 580–598.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (2017). *Classification and regression trees*. Routledge.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, *17*, pp. 453–510.

Chambers, J. M. and Hastie, T. (1991). *Statistical Models in S*. CRC Press, Inc.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In: *Constructive theory of functions of several variables*. Springer, pp. 85–100.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American statistical Association*, *87*(420), pp. 998–1004.

Fan, J. and Gijbels, I. (2018). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Routledge.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*, pp. 1–67.

Genovese, C. R., Miller, C. J., Nichol, R. C., Arjunwadkar, M. and Wasserman, L. (2004). Nonparametric Inference for the Cosmic Microwave Background. *Statistical Science*, *19*(2), pp. 308–321.

Green, P. J. and Silverman, B. W. (2019). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall/CRC.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association*, *83*(401), pp. 86–95.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

Hastie, T. J. and Tibshirani, R. J. (2017). *Generalized additive models*. Routledge.

Hettmansperger, T. P., McKean, J. W. and Sheather, S. J. (1997). 7 Rank-based analyses of linear models. In: *Robust Inference*. Vol. 15. Handbook of Statistics. Elsevier, pp. 145–173.

Ho, T. K. (1995). Random decision forests. In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE, pp. 278–282.

Hollander, M., Wolfe, D. and Chicken, E. (2014). *Nonparametric Statistical Methods* (3nd ed.). John Wiley and Sons.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Vol. 112. Springer.

Loader, C. (2012). Smoothing: local regression techniques. In: *Handbook of Computational Statistics*. Springer, pp. 571–596.

Loader, C. (2006). *Local regression and likelihood*. Springer Science & Business Media.

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability and Its Applications*, *9*, pp. 141–2.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische mathematik*, *10*(3), pp. 177–183.

Ripley, B. (2021). *tree: Classification and Regression Trees*. R package version 1.0-41. URL: https://CRAN.R-project.org/package=tree.

Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of S-estimators. In: *Robust and nonlinear time series analysis*. Springer, pp. 256–272.

Schoenberg, I. J. (1964a). Spline Functions and the problem of graduation. *Proceedings of the National Academy of Sciences*, *52*(4), pp. 947–950.

Schoenberg, I. J. (1964b). On interpolation by spline functions and its minimal properties. In: *On Approximation Theory/Über Approximationstheorie*. Springer, pp. 109–129.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, *63*, pp. 1379–1389.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, pp. 689–705.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. I, II, III. *Nederl. Akad. Wetensch., Proc.*, *53*, pp. 386–392, 521–525, 1397–1412.

Therneau, T. and Atkinson, B. (2019). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15. URL: https://CRAN.R-project.org/package=rpart.

Tukey, J. W. (1961). Curves as parameters, and touch estimation. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Wahba, G. (1975). Smoothing noisy data with spline functions. *Numerische mathematik*, *24*(5), pp. 383–393.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wang, X. and Yu, Q. (2005). Unbiasedness of the Theil–Sen estimator. *Nonparametric Statistics*, *17*, pp. 685–695.

Wasserman, L. (2006). *All of Nonparametric Statistics*. New York, NY: Springer Texts in Statistics.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A.*, *26*, pp. 359–372.

Wilcox, R. (1998). Simulations on the Theil–Sen regression estimator with right-censored data. *Statistics and Probability Letters*, *39*, pp. 43–47.

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, *65*(1), pp. 95–114.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), pp. 673–686.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), pp. 3–36.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman and Hall/CRC.

Wood, S., Pya, N. and Safken, B. (2016). Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association*, *111*, pp. 1548–1575.

Zambom, A. Z. and Akritas, M. G. (2017). NonpModelCheck: An R Package for Nonparametric Lack-of-Fit Testing and Variable Selection. *Journal of Statistical Software*, *77*(10), pp. 1–28.

ΚΕΦΑΛΑΙΟ 10

Η ΜΕΘΟΔΟΣ JACKKNIFE

Σύνοψη

Η μέθοδος jackknife είναι μια τεχνική επαναδειγματοληψίας που επιτρέπει την εκτίμηση της μεροληψίας και του τυπικού σφάλματος ενός εκτιμητή σε ένα μη παραμετρικό πλαίσιο. Σε αρκετές περιπτώσεις προβλημάτων εκτίμησης, η τεχνική αυτή οδηγεί σε εκτιμητές με μικρότερη μεροληψία και σε συνεπή εκτίμηση του τυπικού σφάλματος. Στην Ενότητα 10.1 παρουσιάζεται η βασική ιδέα της μεθόδου και περιγράφεται, μέσω απλών παραδειγμάτων, η διαδικασία για την εφαρμογή της τεχνικής. Θεωρητικές ιδιότητες σχετικά με την εκτίμηση της μεροληψίας και της διασποράς εκτιμητών μέσω του jackknife παρουσιάζονται στις Ενότητες 10.2 και 10.3. Στην Ενότητα 10.4 γίνεται μια σύντομη αναφορά στην τεχνική jackknife ανώτερης τάξης, ενώ το Κεφάλαιο ολοκληρώνεται με μία εφαρμογή μέσω `R` σε πραγματικά δεδομένα στην Ενότητα 10.5.

Προαπαιτούμενη γνώση:
Βασικές γνώσεις Εκτιμητικής.

Προσδοκώμενα μαθησιακά αποτελέσματα:
Ο/η φοιτητής/τρια ολοκληρώνοντας την ενότητα αυτή θα μπορεί να εφαρμόζει την τεχνική jackknife για την εκτίμηση της μεροληψίας και του τυπικού σφάλματος ενός εκτιμητή. Επίσης, θα έχει κατανοήσει τις βασικές ιδιότητες του εκτιμητή jackknife και πότε αυτός μπορεί να χρησιμοποιηθεί.

Γλωσσάριο επιστημονικών όρων

- Επαναδειγματοληψία
- Μεροληψία και διασπορά εκτιμητή
- Τυπικό σφάλμα
- Jackknife

Μπατσίδης, Α., Παπασταμούλης, Π., Πετρόπουλος, Κ., & Ρακιτζής, Α. (2022). *Μη Παραμετρική Στατιστική.* [Προπτυχιακό εγχειρίδιο].

10.1 Ο εκτιμητής jackknife

Η μέθοδος jackknife είναι μια τεχνική επαναδειγματοληψίας που εισήχθη στη βιβλιογραφία ως μία τεχνική μείωσης της μεροληψίας ενός εκτιμητή από τον Maurice Quenouille (1924-1973) στις εργασίες του Quenouille (1949) και Quenouille (1956). Αργότερα, ο Tukey (1958) παρατήρησε ότι το jackknife μπορεί να χρησιμοποιηθεί εξίσου και ως μία μέθοδος εκτίμησης τυπικών σφαλμάτων εκτιμητών, παρατήρηση που έκανε την τεχνική αυτή αρκετά δημοφιλή. Η μέθοδος jackknife μπορεί να δώσει απαντήσεις σε περιπτώσεις όπου αναλυτικοί υπολογισμοί ή κλασικά ασυμπτωτικά αποτελέσματα δεν είναι εύκολο να χρησιμοποιηθούν. Εξαιτίας αυτού, η τεχνική μπορεί να θεωρηθεί ως ένα αρκετά χρήσιμο εργαλείο στα χέρια ενός Στατιστικού, όπως ακριβώς ένας σουγιάς (jackknife) μπορεί να χρησιμεύσει στα χέρια ενός περιηγητή.

Η βασική ιδέα του jackknife στηρίζεται στον διαδοχικό υπολογισμό της εκτίμησης, αφαιρώντας παρατηρήσεις από το αρχικό δείγμα. Σε κάθε βήμα χρησιμοποιείται ένα διαφορετικό «δείγμα» και αυτό παρέχει, όπως θα δούμε, πληροφορία για τη μεταβλητότητα του εκτιμητή. Η πιο τυπική περίπτωση της τεχνικής jackknife είναι να αφαιρούμε κάθε φορά από μία παρατήρηση. Κατά αυτόν τον τρόπο, προκύπτουν και λαμβάνονται υπόψη n το πλήθος διαφορετικές εκτιμήσεις (μία για κάθε δείγμα). Κατόπιν, χρησιμοποιούμε τις n αυτές τιμές για να εκτιμήσουμε τη μεροληψία του εκτιμητή. Τελικά, ο εκτιμητής jackknife προκύπτει αφαιρώντας από τον αρχικό εκτιμητή την εκτιμηθείσα μεροληψία. Σε ό,τι ακολουθεί θα υποθέσουμε ότι έχουμε στη διάθεσή μας ένα τυχαίο δείγμα $X_1, \dots, X_n$ από πληθυσμό με κατανομή F και ότι μας ενδιαφέρει η εκτίμηση μιας παραμέτρου (ή ενός συναρτησιακού) $\theta \in \Theta$. Υποθέτουμε, επίσης, ότι έχουμε στη διάθεσή μας έναν αρχικό εκτιμητή $T_n = T(X_1, \dots, X_n)$ της παραμέτρου θ: $T_n = \widehat{\theta}_n$. Υπενθυμίζουμε ότι η *μεροληψία* του εκτιμητή T_n, που χρησιμοποιείται για την εκτίμηση της παραμέτρου $\theta \in \Theta$, ορίζεται από τη σχέση:

$$\text{bias}(T_n) = \mathsf{E}(T_n) - \theta, \text{ για κάθε } \theta \in \Theta. \tag{10.1}$$

Ορίζουμε ως

$$T(-i) = T(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n), \quad i = 1, \dots, n,$$

τον εκτιμητή που προκύπτει αν **αφαιρέσουμε** την i-οστή δειγματική παρατήρηση.

Ορισμός 10.1

Ο **εκτιμητής jackknife** ορίζεται ως

$$T_{\text{jack}} := T_n - b_{\text{jack}} \tag{10.2}$$

όπου με b_{jack} ορίζουμε την **jackknife εκτίμηση της μεροληψίας**

$$b_{\text{jack}} := (n-1)\left(\overline{T}_n - T_n\right), \tag{10.3}$$

με

$$\overline{T}_n = \frac{1}{n}\sum_{i=1}^{n} T(-i). \tag{10.4}$$

Είναι εύκολο να αποδειχθεί (βλ. Άσκηση 10.1) ότι ο εκτιμητής jackknife μπορεί να εκφραστεί εναλλακτικά ως

$$T_{\text{jack}} = nT_n - (n-1)\overline{T}_n, \tag{10.5}$$

και

$$T_{\text{jack}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{T}_i, \tag{10.6}$$

όπου στη σχέση (10.6) οι **ψευδοτιμές** $\tilde{T}_i$ ορίζονται ως

$$\tilde{T}_i := nT_n - (n-1)T(-i), \quad i = 1, \dots, n. \tag{10.7}$$

Η ιδέα πίσω από την αναπαράσταση μέσω των ψευδοτιμών είναι ότι ο εκτιμητής jackknife γράφεται ως ο δειγματικός μέσος n τιμών. Ωστόσο, οι ψευδοτιμές $\tilde{T}_i$, $i = 1, \dots, n$, δεν είναι ανεξάρτητες, παρά μόνο σε ειδικές περιπτώσεις (βλ. Άσκηση 10.2). Στις επόμενες ενότητες θα συζητήσουμε κάποιες ιδιότητες του εκτιμητή jackknife. Πριν φτάσουμε εκεί όμως, θα πάρουμε μια (ας μας επιτραπεί η έκφραση) πρώτη γεύση για αυτές μέσω του παραδείγματος που ακολουθεί.

Παράδειγμα 10.1. Έστω τυχαίο δείγμα έξι τιμών $(4, 3, 7, 6, 5, 9)$ από πληθυσμό με κατανομή F. Να υπολογιστούν οι εκτιμητές jackknife για τα συναρτησιακά $T_1(F) = \mu$ και $T_2(F) = \sigma^2$, με βάση τους αντίστοιχους εκτιμητές αντικατάστασης.

Λύση Παραδείγματος 10.1. Ξέρουμε ότι οι εκτιμητές αντικατάστασης είναι

$$\hat{\mu} = T_1(\hat{F}_n) = \bar{X}_n$$

και

$$\hat{\sigma}^2 = T_2(\hat{F}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2,$$

αντίστοιχα. Μετά από αλγεβρικές πράξεις, προκύπτει ότι

$$\hat{\mu} = \frac{34}{6} \text{ και } \hat{\sigma}^2 = \frac{23.3333}{6}.$$

Στη συνέχεια, θέλουμε να υπολογίσουμε τον εκτιμητή jackknife των παραμέτρων μ και σ^2. Για τον σκοπό αυτόν στον Πίνακα 10.1 δίνονται οι απαραίτητες ποσότητες για τον υπολογισμό τους.

Παρατηρήσεις		Μέσος		Διασπορά	
i	x_i	$T(-i)$	$\tilde{T}_i$	$T(-i)$	$\tilde{T}_i$
1	4.00	6.00	4.00	4.00	3.3333
2	3.00	6.20	3.00	2.96	8.5333
3	7.00	5.40	7.00	4.24	2.1333
4	6.00	5.60	6.00	4.64	0.1333
5	5.00	5.80	5.00	4.56	0.5333
6	9.00	5.00	9.00	2.00	13.3333

Πίνακας 10.1: Υπολογισμός των εκτιμήσεων $T(-i)$ και των αντίστοιχων ψευδοτιμών για καθεμία από τις έξι παρατηρήσεις του δείγματος στο Παράδειγμα 10.1.

Για την κατανόηση των παραπάνω υπολογισμών ας πάρουμε για παράδειγμα την πρώτη παρατήρηση $x_1 = 4$. Αρχικά, πρέπει να υπολογίσουμε την εκτίμηση $T(-1)$ που προκύπτει αφαιρώντας την πρώτη παρατήρηση από το δείγμα των έξι τιμών, δηλαδή να υποθέσουμε ότι παρατηρήσαμε τις τιμές $(x_2, x_3, x_4, x_5, x_6) = (3, 7, 6, 5, 9)$. Έτσι, για τη μέση τιμή θα έχουμε ότι:

$$T(-1) = (3 + 7 + 6 + 5 + 9)/5 = 6 \text{ και } \tilde{T}_1 = 6 \cdot \hat{\mu} - (6-1)T(-1) = 6 \cdot \frac{34}{6} - 5 \cdot 6 = 4.$$

Αντίστοιχα, για τη διασπορά έχουμε ότι:

$$T(-1) = \frac{(3-6)^2 + (7-6)^2 + (6-6)^2 + (5-6)^2 + (9-6)^2}{5} = 4$$

και

$$\tilde{T}_1 = 6 \cdot \hat{\sigma}^2 - (6-1)T(-1) = 6 \cdot \frac{23.3333}{6} - 5 \cdot 4 = 3.3333.$$

Με παρόμοιο τρόπο υπολογίζουμε τις αντίστοιχες ποσότητες στις επόμενες γραμμές του Πίνακα 10.1.

Έτσι, αντικαθιστώντας στη σχέση (10.6), έχουμε ότι η jackknife εκτίμηση για τον μέσο είναι ίση με

$$T_{\text{jack}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{T}_i = \frac{34}{6}.$$

Παρατηρούμε ότι η εκτίμηση αυτή συμπίπτει με την εκτίμηση που δίνει ο δειγματικός μέσος.

Για τη διασπορά, θυμηθείτε ότι η αρχική εκτίμηση είναι $T_n = \widehat{\sigma}^2 = \frac{23.3333}{6}$, ενώ η τεχνική jackknife δίνει

$$T_{\text{jack}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{T}_i = \frac{27.9998}{6} \approx 4.67.$$

Εδώ, λοιπόν, παρατηρούμε ότι οι δύο εκτιμήσεις διαφέρουν. Ας παρατηρήσουμε σε αυτό το σημείο ότι η δειγματική διασπορά είναι $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \approx 4.67$. Δηλαδή, η jackknife εκτίμηση αφαίρεσε τη μεροληψία του $\widehat{\sigma}^2$ και κατέληξε στον αμερόληπτο εκτιμητή S_n^2! □

Είδαμε στο προηγούμενο παράδειγμα ότι η jackknife εκτίμηση για τη μέση τιμή συμπίπτει με τον δειγματικό μέσο και ότι η jackknife εκτίμηση της διασποράς συμπίπτει με τη δειγματική διασπορά. Στην πρόταση που ακολουθεί αποδεικνύεται ότι το παραπάνω συμβαίνει πάντοτε.

Πρόταση 10.1. Ο jackknife εκτιμητής της μέσης τιμής είναι

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

ενώ ο jackknife εκτιμητής της διασποράς είναι

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2. \tag{10.8}$$

Απόδειξη Πρότασης 10.1. Για τη μέση τιμή έχουμε ότι οι ψευδοτιμές είναι:

$$\tilde{T}_i = nT_n - (n-1)T(-i) = n\frac{1}{n}\sum_{i=1}^{n} X_i - (n-1)\frac{1}{n-1}\sum_{j\neq i}^{n} X_j = \sum_{i=1}^{n} X_i - \left(\sum_{j=1}^{n} X_j - X_i\right) = X_i.$$

Άρα, σε αυτήν την περίπτωση, παρατηρούμε ότι οι ψευδοτιμές συμπίπτουν με τις αρχικές παρατηρήσεις. Επομένως, έχουμε από τη σχέση (10.6) ότι $T_{\text{jack}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{T}_i = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$.

Για τη διασπορά έχουμε ότι:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{\sum_{i=1}^{n} X_i^2}{n} - n\bar{X}^2$$

ή, ισοδύναμα,

$$\hat{\sigma}^2 = \frac{1}{n^2}\left(n\sum_{i=1}^{n} X_i^2 - (n\bar{X})^2\right)$$

και αναπτύσσοντας το τετράγωνο του $\bar{X}$ προκύπτει ότι:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n^2}\left\{(n-1)\sum_{i=1}^{n}X_i^2 - \sum_{i=1}^{n}\sum_{j\neq i}X_iX_j\right\},$$

ή, ισοδύναμα,

$$n\widehat{\sigma}^2 = nT_n = \frac{1}{n}\left\{(n-1)\sum_{i=1}^{n}X_i^2 - \sum_{i=1}^{n}\sum_{j\neq i}X_iX_j\right\}. \tag{10.9}$$

Με παρόμοιο τρόπο, στην περίπτωση που θεωρήσουμε τις παρατηρήσεις $X_1, \ldots, X_i, X_{i+1}, \ldots, X_n$, έχουμε ότι:

$$(n-1)T(-i) = \frac{1}{n-1}\left\{(n-2)\sum_{j\neq i}X_j^2 - \sum_{k\neq i}\sum_{j\neq k}X_kX_j\right\}.$$

Άρα, έχουμε ότι:

$$\begin{aligned}
(n-1)\overline{T}_n &= \frac{n-1}{n}\sum_{i=1}^{n}T(-i)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n-1}\left\{(n-2)\sum_{j\neq i}X_j^2 - \sum_{k\neq i}\sum_{j\neq k}X_kX_j\right\}\right]\\
&= \frac{1}{n(n-1)}\left\{(n-2)\sum_{i=1}^{n}\sum_{j\neq i}X_j^2 - \sum_{i=1}^{n}\sum_{k\neq i}\sum_{j\neq k}X_kX_j\right\}\\
&= \frac{1}{n(n-1)}\left\{(n-2)(n-1)\sum_{i=1}^{n}X_i^2 - (n-2)\sum_{i=1}^{n}\sum_{j\neq i}X_iX_j\right\}\\
&= \frac{n-2}{n-1}\frac{1}{n}\left\{(n-1)\sum_{i=1}^{n}X_i^2 - \sum_{i=1}^{n}\sum_{j\neq i}X_iX_j\right\}\\
&\overset{(10.9)}{=} \frac{n-2}{n-1}n\hat{\sigma}^2.
\end{aligned}$$

Επομένως, ο jackknife εκτιμητής θα είναι

$$T_{\text{jack}} = nT_n - (n-1)\overline{T}_n = n\hat{\sigma}^2 - \frac{n-2}{n-1}n\hat{\sigma}^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

δηλαδή η δειγματική διασπορά S_n^2. □

10.2 Μεροληψία και εκτιμητής jackknife

Στην προηγούμενη ενότητα, είδαμε ότι η ιδέα πάνω στην οποία βασίζεται ο εκτιμητής jackknife είναι η εκτίμηση της μεροληψίας του αρχικού εκτιμητή T_n και, κατόπιν, η αφαίρεση αυτής της ποσότητας από τον αρχικό εκτιμητή. Τι γίνεται, λοιπόν, στην περίπτωση που θα εφαρμόσουμε την τεχνική jackknife σε έναν αμερόληπτο εκτιμητή; Σύμφωνα με την πρόταση που ακολουθεί, ο εκτιμητής που προκύπτει παραμένει αμερόληπτος.

Πρόταση 10.2. Αν T_n είναι αμερόληπτος εκτιμητής της παραμέτρου θ, τότε το ίδιο ισχύει και για τον εκτιμητή jackknife T_{jack}. Δηλαδή

$$\mathsf{E}(T_n) = \theta, \quad \forall\, \theta \in \Theta \quad \Rightarrow \quad \mathsf{E}(T_{\text{jack}}) = \theta, \quad \forall\, \theta \in \Theta. \tag{10.10}$$

Απόδειξη Πρότασης 10.2. Έστω ότι ο T_n είναι αμερόληπτος εκτιμητής της παραμέτρου θ, δηλαδή $\mathsf{E}(T_n) = \theta$, $\forall\, \theta \in \Theta$. Η ιδιότητα της αμεροληψίας θα ισχύει τότε και για τους εκτιμητές $T(-i)$, $i = 1, \ldots, n$, δηλαδή

$$\mathsf{E}(T(-i)) = \theta, \quad \forall\, \theta \in \Theta.$$

Συνεπώς,

$$\mathsf{E}(T_{\text{jack}}) = \mathsf{E}\left\{nT_n - (n-1)\overline{T}_n\right\} = n\mathsf{E}(T_n) - (n-1)\mathsf{E}(\overline{T}_n) = n\theta - (n-1)\theta = \theta, \quad \forall\, \theta \in \Theta.$$

Επομένως, αποδείξαμε ότι και ο T_{jack} είναι αμερόληπτος. □

Σε αυτό το σημείο, είναι εύλογο και λογικό να διερωτηθούμε αν μπορούμε να πούμε κάτι σχετικά με τη μεροληψία του εκτιμητή jackknife T_{jack} στη γενικότερη περίπτωση όπου ο αρχικός εκτιμητής T_n δεν είναι αμερόληπτος. Απάντηση στο παραπάνω ερώτημα δίνεται στην πρόταση που ακολουθεί, υπό την υπόθεση ότι η μεροληψία του εκτιμητή T_n είναι της μορφής

$$\text{bias}(T_n) = \frac{\alpha_1(\theta)}{n} + \frac{\alpha_2(\theta)}{n^2} + \cdots \tag{10.11}$$

όπου $\alpha_1, \alpha_2, \ldots$ είναι συναρτήσεις που δεν εξαρτώνται από το n. Η παραπάνω υπόθεση (10.11) για τη μορφή της μεροληψίας του εκτιμητή δεν είναι στην πράξη δεσμευτική, καθώς μπορεί να εξασφαλιστεί για αρκετούς εκτιμητές, συμπεριλαμβάνοντας τους περισσότερους εκτιμητές μέγιστης πιθανοφάνειας υπό κάποιες συνθήκες ομαλότητας (Quenouille, 1956; Schucany *et al.*, 1971). Για παράδειγμα, αν μας ενδιαφέρει η εκτίμηση του $\theta = \sigma^2$ και χρησιμοποιείται ο εκτιμητής $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, είναι εύκολο να δειχθεί ότι

$$\mathsf{E}(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{1}{n}\sigma^2$$

και, επομένως, έχουμε ότι $\text{bias}(\hat{\sigma}^2) = -\sigma^2/n$, που είναι πράγματι μια συνάρτηση της μορφής (10.11), με $\alpha_1(\theta) = -\theta$ και $\alpha_2(\theta) = 0$.

Πρόταση 10.3. Υποθέτοντας ότι ισχύει η σχέση (10.11) για τη μεροληψία του T_n, τότε

$$\mathsf{E}(b_{\text{jack}}) = \text{bias}(T_n) + O\left(\frac{1}{n^2}\right). \tag{10.12}$$

και

$$\text{bias}(T_{\text{jack}}) = O\left(\frac{1}{n^2}\right). \tag{10.13}$$

Απόδειξη Πρότασης 10.3. Από τη σχέση (10.11) έχουμε ότι:

$$\text{bias}(T_n) = \frac{\alpha_1(\theta)}{n} + \frac{\alpha_2(\theta)}{n^2} + O\left(\frac{1}{n^3}\right),$$

όπου $O\left(\frac{1}{n^3}\right)$ είναι μια ποσότητα η οποία δεν ξεπερνά το $M\frac{1}{n^3}$ κατά απόλυτη τιμή, για κάποια θετική σταθερά M και για μεγάλες τιμές του n. Η μεροληψία των $T(-i)$ γράφεται ως

$$\text{bias}\,(T(-i)) = \frac{\alpha_1(\theta)}{n-1} + \frac{\alpha_2(\theta)}{(n-1)^2} + O\left(\frac{1}{n^3}\right),$$

διότι $O(1/(n-1)^3) = O(1/n^3)$. Άρα,

$$\text{bias}\left(\overline{T}_n\right) = \text{bias}\left(\frac{1}{n}\sum_{i=1}^{n} T(-i)\right) = \frac{\alpha_1(\theta)}{n-1} + \frac{\alpha_2(\theta)}{(n-1)^2} + O\left(\frac{1}{n^3}\right).$$

Συνεπώς, η μέση τιμή του jackknife εκτιμητή της μεροληψίας είναι

$$\begin{aligned}
\text{E}(b_{\text{jack}}) &= (n-1)\left\{\text{E}(\overline{T}_n) - \text{E}(T_n)\right\} \\
&= (n-1)\left\{(\text{E}(\overline{T}_n) - \theta) - (\text{E}(T_n) - \theta)\right\} \\
&= (n-1)\left\{\text{bias}(\overline{T}_n) - \text{bias}(T_n)\right\} \\
&= (n-1)\left\{\left(\frac{1}{n-1} - \frac{1}{n}\right)\alpha_1(\theta) + \left(\frac{1}{(n-1)^2} - \frac{1}{n^2}\right)\alpha_2(\theta) + O\left(\frac{1}{n^3}\right)\right\} \\
&= \frac{\alpha_1(\theta)}{n} + \frac{(2n-1)\alpha_2(\theta)}{n^2(n-1)} + O\left(\frac{1}{n^2}\right) \\
&= \frac{\alpha_1(\theta)}{n} + O\left(\frac{1}{n^2}\right) \\
&= \text{bias}(T_n) + O\left(\frac{1}{n^2}\right).
\end{aligned}$$

Άρα, το b_{jack} εκτιμά τη μεροληψία του T_n με ακρίβεια της τάξης του $1/n^2$. Με παρόμοιο τρόπο μπορούμε να δείξουμε ότι

$$\text{bias}(T_{\text{jack}}) = -\frac{\alpha_2(\theta)}{n(n-1)} + O\left(\frac{1}{n^2}\right) = O\left(\frac{1}{n^2}\right)$$

και η απόδειξη ολοκληρώνεται. □

Παρατήρηση 10.1. Θυμηθείτε ότι λόγω της σχέσης (10.11) έχουμε ότι η μεροληψία του αρχικού εκτιμητή είναι τάξης $O(1/n)$. Σύμφωνα, όμως, με τη σχέση (10.13) η μεροληψία του T_{jack} είναι τάξης $O(1/n^2)$, δηλαδή **μικρότερη** κατά μία τάξη μεγέθους συγκριτικά με τη μεροληψία του αρχικού εκτιμητή T_n.

Στην πρόταση που ακολουθεί, διατυπώνεται το συμπέρασμα ότι, αν ο αρχικός εκτιμητής T_n είναι *τετραγωνική στατιστική συνάρτηση*, δηλαδή αν μπορεί να εκφραστεί στη μορφή:

$$T_n = \text{E}(T_n) + \frac{1}{n}\sum_{i=1}^{n}\alpha^{(n)}(x_i) + \frac{1}{n^2}\sum_{1\leqslant i\leqslant j\leqslant n}\beta^{(n)}(x_i, x_j), \tag{10.14}$$

όπου $\alpha^{(n)}(x)$ και $\beta^{(n)}(x,y)$ είναι συναρτήσεις των x και x,y, αντίστοιχα, οι οποίες δύνανται να εξαρτώνται και από το μέγεθος δείγματος n, τότε ο b_{jack} εκτιμά αμερόληπτα τη μεροληψία του αρχικού εκτιμητή T_n. Για παράδειγμα, ο εκτιμητής αντικατάστασης της διασποράς είναι τετραγωνική στατιστική συνάρτηση (βλ. Άσκηση 10.7).

Πρόταση 10.4. Αν ο αρχικός εκτιμητής T_n είναι τετραγωνική στατιστική συνάρτηση της μορφής της σχέσης (10.14), ο b_{jack}, που ορίστηκε στη σχέση (10.3), εκτιμά αμερόληπτα τη μεροληψία του αρχικού εκτιμητή T_n.

Απόδειξη Πρότασης 10.4. Για την απόδειξη παραπέμπουμε στο σύγγραμμα των Efron and Stein (1981). □

10.3 Τυπικά σφάλματα και διαστήματα εμπιστοσύνης

Ο Tukey (1958), θεωρώντας ότι οι ψευδοτιμές (βλ. σχέση (10.7)) είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές, χρησιμοποίησε ως εκτίμηση της διασποράς του εκτιμητή jackknife την ποσότητα

$$v_{\text{jack}} = \frac{\tilde{s}_n^2}{n} \tag{10.15}$$

όπου

$$\tilde{s}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\tilde{T}_i - \frac{1}{n}\sum_{j=1}^{n}\tilde{T}_j\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\tilde{T}_i - T_{\text{jack}}\right)^2.$$

Σημειώνουμε, εδώ, ότι το v_{jack} μπορεί να θεωρηθεί ως εκτίμηση της διασποράς τόσο του αρχικού εκτιμητή, δηλαδή της $\text{Var}\{T_n\}$, όσο και του εκτιμητή jackknife, δηλαδή της $\text{Var}\{T_{\text{jack}}\}$ (βλ. Hinkley, 1978). Επομένως, η εκτίμηση του τυπικού σφάλματος ισούται με

$$\text{se}_{\text{jack}} = \sqrt{v_{\text{jack}}}.$$

Γενικά, μπορεί να δειχθεί ότι η εκτίμηση (10.15) είναι μεροληπτική και, πιο συγκεκριμένα, ότι οδηγεί σε υπερεκτίμηση της πραγματικής διασποράς $\text{Var}\,(T_n)$ (Efron and Stein, 1981). Σύμφωνα όμως με το παρακάτω θεώρημα, η εκτίμηση είναι ασυμπτωτικά συνεπής στην περίπτωση όπου ο εκτιμητής είναι συνάρτηση του δειγματικού μέσου. Σημειώνουμε ότι ένα πιο γενικό αποτέλεσμα για την περίπτωση πολυμεταβλητών δεδομένων με πεπερασμένη μέση τιμή και πίνακα διασπορών - συνδιασπορών δίνεται στους Shao and Tu (2012).

Θεώρημα 10.1: Miller (1964)

Έστω ότι $\text{E}(X_1) = \mu$, $\text{Var}(X_1) = \sigma^2 < \infty$ και ότι $T_n = g(\bar{X})$, όπου g είναι μια συνάρτηση η οποία έχει συνεχή και μη μηδενική πρώτη παράγωγο στο μ. Τότε

$$\frac{T_n - g(\mu)}{\sigma_n} \xrightarrow{d} \mathcal{N}(0,1),$$

όπου $\sigma_n^2 = n^{-1}[g'(\mu)]^2\sigma^2$ και ο εκτιμητής είναι συνεπής, δηλαδή

$$\frac{v_{\text{jack}}}{\sigma_n^2} \xrightarrow{\text{σ.β.}} 1.$$

Απόδειξη Θεωρήματος 10.1. Για την απόδειξη παραπέμπουμε στην απόδειξη του Θεωρήματος 1 στην εργασία του Miller (1964). □

Ξεκάθαρα, οι περιπτώσεις που καλύπτονται από το προηγούμενο θεώρημα είναι περιορισμένες, καθώς πρόκειται για στατιστικές συναρτήσεις που εκφράζονται ως συνάρτηση του δειγματικού μέσου. Ωστόσο, μπορεί να αποδειχθεί ότι ανάλογο αποτέλεσμα καλύπτει και περιπτώσεις στατιστικών συναρτήσεων που εκφράζονται ως συνάρτηση της δειγματικής διασποράς, δηλαδή όταν $T_n = g(S_n^2)$ (Miller, 1968). Τέλος, όπως διατυπώνεται στην πρόταση που ακολουθεί, αποδεικνύεται ότι η v_{jack} δεν είναι συνεπής για μη «ομαλά» στατιστικά, όπως είναι για παράδειγμα τα δειγματικά ποσοστιαία σημεία $T_n = \tilde{F}_n^{-1}(p)$.

Θεώρημα 10.2: Efron (1982)

Ο jackknife εκτιμητής της διασποράς του εκτιμητή του p ποσοστιαίου σημείου $T(F) = F^{-1}(p)$ μιας κατανομής F δεν είναι συνεπής.

Απόδειξη Θεωρήματος 10.2. Για την απόδειξη παραπέμπουμε στο σύγγραμμα του Efron (1982). □

Παράδειγμα 10.2. Να εκτιμηθεί μέσω jackknife το τυπικό σφάλμα του δειγματικού μέσου στα δεδομένα του Παραδείγματος 10.1.

Λύση Παραδείγματος 10.2. Μέσω του jackknife εκτιμούμε τη διασπορά του δειγματικού μέσου από τη σχέση (10.15). Προκύπτει ότι: $\tilde{s}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\tilde{T}_i - T_{\text{jack}}\right)^2 = 4.67$. Άρα,

$$v_{\text{jack}} = \frac{\tilde{s}_n^2}{n} = 0.78$$

οπότε $\text{se}_{\text{jack}} = \sqrt{v_{\text{jack}}} = \sqrt{0.78}$.

Παρατηρήστε εδώ ότι έχουμε καταλήξει στο ίδιο αποτέλεσμα με αυτό που προβλέπει η θεωρία για τον δειγματικό μέσο! Γνωρίζουμε ότι $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$ που εκτιμάται με $\widehat{\text{Var}}(\bar{X}_n) = \frac{S_n^2}{n}$. Η εκτίμηση του τυπικού σφάλματος του δειγματικού μέσου είναι $\widehat{\text{se}} = \sqrt{\frac{S_n^2}{n}}$, από όπου προκύπτει ότι $\widehat{\text{se}} = \sqrt{0.78}$. □

Ο Tukey (1958) πρότεινε τη χρήση του διαστήματος

$$T_{\text{jack}} \pm t_{n-1;\alpha/2}\,\text{se}_{\text{jack}}, \tag{10.16}$$

ως ένα προσεγγιστικό $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την παράμετρο θ. Το επιχείρημα του Tukey για την κατασκευή αυτού του διαστήματος ήταν ότι σε πολλές περιπτώσεις οι ψευδοτιμές $\tilde{T}_i$ στη σχέση (10.6) μπορούν κατά προσέγγιση να θεωρηθούν ως ανεξάρτητες και ισόνομες τυχαίες μεταβλητές. Σε αυτές τις περιπτώσεις, λοιπόν, η

$$\frac{\sqrt{n}(T_{\text{jack}} - \theta)}{\tilde{s}_n^2}$$

θα ακολουθεί κατά προσέγγιση την κατανομή t_{n-1} και μπορεί να θεωρηθεί ως ποσότητα οδηγός για την παράμετρο θ.

Παρατήρηση 10.2. Τονίζουμε ξανά ότι οι ψευδοτιμές δεν είναι γενικά ανεξάρτητες, οπότε το παραπάνω διάστημα εμπιστοσύνης δεν είναι ακριβές. Σύμφωνα με τον Efron (1982), το διάστημα στην (10.16) θα πρέπει να χρησιμοποιείται μόνο όταν το μέγεθος δείγματος είναι μεγάλο. Παρατηρήστε, επίσης, ότι, εφόσον σε τέτοιες περιπτώσεις η κατανομή t_n συγκλίνει στην τυπική κανονική κατανομή, δεν υπάρχει κάποια πρακτική διαφορά μεταξύ της χρήσης ποσοστιαίων σημείων της κατανομής t_{n-1} ή ποσοστιαίων σημείων της τυπικής κανονικής κατανομής.

10.4 Jackknife ανώτερης τάξης

Έως τώρα, έχουμε δει ότι η βασική ιδέα του jackknife είναι ο διαδοχικός υπολογισμός του εκτιμητή μετά την αφαίρεση μίας παρατήρησης τη φορά. Η ιδέα αυτή μπορεί να γενικευτεί και να αφαιρούμε παραπάνω από μία παρατηρήσεις τη φορά. Τέτοιου είδους τακτικές είναι γνωστές με το όνομα jackknife ανώτερης τάξης. Για παράδειγμα, αν αφαιρούμε ζευγάρια παρατηρήσεων, κάνουμε λόγο για jackknife δεύτερης τάξης.

Στο πλαίσιο αυτό, υποθέτουμε ότι για το μέγεθος του δείγματος n ισχύει ότι $n = gh$ για κάποιους ακεραίους g και h. Τότε μπορούμε να αφαιρέσουμε παρατηρήσεις σε ομάδες μεγέθους h. Ειδικότερα, η πρώτη ομάδα δύναται να αποτελείται από τις παρατηρήσεις $x_1, \ldots, x_h$, η δεύτερη ομάδα να αποτελείται από τις $x_{h+1}, \ldots, x_{2h}$ και ούτω καθεξής. Επεκτείνοντας τον ορισμό του εκτιμητή Jackknife, που δόθηκε στη σχέση (10.2), ορίζουμε

ως $T(-i)$ τον εκτιμητή που προκύπτει όταν αφαιρέσουμε την i-οστή ομάδα, για $i = 1, \dots, g$. Τότε ο εκτιμητής jackknife τάξης h ορίζεται ως

$$T_{\text{jack}} = gT_n - (g-1)\overline{T}_g,$$

όπου $\overline{T}_g = \frac{1}{g}\sum_{i=1}^{g} T(-i)$.

Είναι προφανές ότι η περίπτωση $h = 1$ αντιστοιχεί στον τυπικό jackknife εκτιμητή με $g = n$ ομάδες, όπως έχουμε δει στα προηγούμενα.

Παρατήρηση 10.3. Στην περίπτωση που δεν είναι απαγορευτικό, είναι προτιμότερο να ορίσουμε ως

$$\overline{T}_g = \frac{1}{\binom{n}{h}} \sum_{\tilde{i}} T(-\tilde{i}),$$

όπου το $\tilde{i}$ συμβολίζει ένα υποσύνολο μεγέθους h του $\{1, \dots, n\}$, ενώ με $\sum_{\tilde{i}}$ συμβολίζεται το άθροισμα σε όλα τα $\tilde{i}$. Σε αυτήν την περίπτωση, ο εκτιμητής jackknife έχει ίδια μέση τιμή με πριν, αλλά μικρότερη διασπορά.

Παρατήρηση 10.4. Σημειώνουμε ότι στην πρώτη δημοσιευμένη εργασία σχετικά με το jackknife (βλ. Quenouille, 1949) θεωρήθηκε η περίπτωση όπου $g = 2$. Μπορεί να δειχθεί ότι τέτοιες παραλλαγές είναι ικανές να «εξαφανίσουν» και επιπλέον όρους της μεροληψίας (βλ. Miller, 1974).

10.5 Εφαρμογή με `R`

Θα θεωρήσουμε τα δεδομένα συσπάσεων νευρικής ίνας της Ενότητας 2.5 και έστω ότι μας ενδιαφέρει η συμπερασματολογία για τον συντελεστή λοξότητας $T(F) = \frac{1}{\sigma^3}\int (x-\mu)^3 \mathrm{d}F$ της άγνωστης πληθυσμιακής κατανομής F, όπου με μ και σ συμβολίζουμε τη μέση τιμή και την τυπική απόκλιση του πληθυσμού, αντίστοιχα. Για τον σκοπό αυτό, θα θεωρήσουμε τον εκτιμητή αντικατάστασης $T(\hat{F}_n)$ (βλ. Παράδειγμα 2.7, Κεφάλαιο 2) και θα εκτιμήσουμε τη μεροληψία και το τυπικό σφάλμα του εκτιμητή αντικατάστασης μέσω της μεθόδου jackknife. Στη συνέχεια, θα υπολογιστεί ένα 95% προσεγγιστικό διάστημα εμπιστοσύνης.

Σύμφωνα με το Παράδειγμα 2.7, ο εκτιμητής αντικατάστασης του $T(F)$ είναι ο

$$T(\hat{F}_n) = \hat{a}_3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^3}{\hat{\sigma}^3}.$$

Από την αντίστοιχη εφαρμογή στην Ενότητα 2.5, έχουμε ότι οι εκτιμητές αντικατάστασης της μέσης τιμής και τυπικής απόκλισης είναι ίσοι με $\hat{\mu} \approx 0.22$ και $\hat{\sigma} \approx 0.21$, αντίστοιχα, ενώ η σημειακή εκτίμηση του συντελεστή λοξότητας ισούται με $T(\hat{F}_n) \approx 1.76$.

Ο παρακάτω κώδικας `R` υπολογίζει τις jackknife εκτιμήσεις της μεροληψίας και του τυπικού σφάλματος του εκτιμητή αντικατάστασης $T(\hat{F}_n)$, καθώς και το προσεγγιστικό 95% διάστημα εμπιστοσύνης για τον συντελεστή λοξότητας. Πριν εκτελέσουμε τις εντολές που ακολουθούν, διαβάζουμε τα δεδομένα (βλ. Ενότητα 2.5) και, κατόπιν, τα αποθηκεύουμε στο διάνυσμα `x`, το οποίο περιέχει τις $n = 799$ το πλήθος παρατηρήσεις του δείγματος.

```
> library("e1071")
> kappa = skewness(x, type = 1)
> kappa
[1] 1.761249
> #jackknife the skewness coefficient
> library("e1071")
```

```
> kappa = skewness(x, type = 1)
> jk_kappa <- pseudo_Value <- numeric(n)
> for(i in 1:n){
+         jk_kappa[i] <- skewness(x[-i], type = 1)
+         pseudo_Value[i] <- n * kappa - (n - 1) * jk_kappa[i]
+ }
> kappa_jack <- mean(pseudo_Value)
> bias_jack <- (n - 1) * (mean(jk_kappa) - kappa)
> se_jack <- sqrt(var(pseudo_Value)/n)
> ci_jack <- kappa_jack + c(-1, 1) * qnorm(0.025, lower.tail = F) * se
   _jack
> cat(paste0("jackknife estimate of bias: ", round(bias_jack, 3)),"\n"
   )
jackknife estimate of bias: -0.017
> cat(paste0("jackknife estimate: ", round(kappa_jack, 3)),"\n")
jackknife estimate: 1.778
> cat(paste0("jackknife estimate of standard error: ", round(se_jack,
   3)),"\n")
jackknife estimate of standard error: 0.172
> cat(paste0("approximate 95% CI: [", round(ci_jack[1], 3), ", ",
   round(ci_jack[2], 3),"]"),"\n")
approximate 95% CI: [1.441, 2.115]
```

Αρχικά, υπολογίζουμε τον εκτιμητή αντικατάστασης $T(\hat{F}_n)$ μέσω της εντολής `skewness()` (απαιτείται η βιβλιοθήκη `e1071` για αυτό). Ορίζουμε τα διανύσματα `jk_kappa` και `pseudo_Value` τα οποία θα τα χρησιμοποιήσουμε για να αποθηκεύσουμε τις n το πλήθος εκτιμήσεις $T(-i)$ και $\tilde{T}_i$, αντίστοιχα. Έτσι, το $T(-i)$ είναι η τιμή του εκτιμητή αντικατάστασης, εξαιρώντας την i-οστή δειγματική παρατήρηση, δηλαδή:

$$T(-i) = \frac{1}{n-1} \frac{\sum_{j=1, j\neq i}^{n} (X_j - \hat{\mu}(-i))^3}{\hat{\sigma}^3(-i)}, \quad i = 1, \ldots, n,$$

όπου τα $\hat{\mu}(-i)$ και $\hat{\sigma}(-i)$ συμβολίζουν τη δειγματική μέση τιμή και τη δειγματική τυπική απόκλιση, αντίστοιχα, στο δείγμα που προκύπτει αφαιρώντας την i-οστή παρατήρηση. Αυτό υπολογίζεται μέσω της εντολής `skewness(x[-i], type = 1)` στην 6η γραμμή του κώδικα. Η ψευδοτιμή $\tilde{T}_i$, που αντιστοιχεί σε αυτήν την περίπτωση, υπολογίζεται στην επόμενη γραμμή του κώδικα και αποθηκεύεται στην i-οστή θέση του διανύσματος `pseudo_Value`.

Αυτή η διαδικασία επαναλαμβάνεται n το πλήθος φορές συνολικά μέσω της εντολής `for()`, μέχρις ότου να εξαντλήσουμε όλες τις διαφορετικές παρατηρήσεις $i = 1, \ldots, n$. Τέλος, στις μεταβλητές `bias_jack` και `se_jack` επιστρέφεται η jackknife εκτίμηση της μεροληψίας μέσω της σχέσης (10.3) και η εκτίμηση του τυπικού σφάλματος $\sqrt{v}_{\text{jack}}$ (βλ. την (10.15)) του εκτιμητή, αντίστοιχα.

Από τα παραπάνω έχουμε ότι η jackknife εκτίμηση της μεροληψίας του εκτιμητή αντικατάστασης ισούται με $b_{\text{jack}} = -0.017$. Συνεπώς, η jackknife εκτίμηση της λοξότητας από τη σχέση (10.2) προκύπτει ότι ισούται με:

$$T_{\text{jack}} = T_n - b_{\text{jack}} = 1.761 - (-0.017) = 1.778,$$

το οποίο μπορεί να υπολογιστεί και εναλλακτικά μέσω των ψευδοτιμών στη σχέση (10.6).

Τα παραπάνω μπορούν να υλοποιηθούν αυτόματα μέσω της χρήσης του πακέτου `bootstrap` μέσω των παρακάτω εντολών.

```
> #  using bootstrap package
> library("bootstrap")
> theta <- function(x){skewness(x, type = 1)}
> results <- jackknife(x,theta)
```

```
> # jackknife estimate of standard error
> round(results[[1]], 3)
[1] 0.172
> # jackknife estimate of bias
> round(results[[2]], 3)
[1] -0.017
```

Φυσικά, τα αποτελέσματα ταυτίζονται με αυτά του δικού μας κώδικα.

Η jackknife εκτίμηση του τυπικού σφάλματος είναι ίση με 0.17. Εφόσον σε αυτήν την περίπτωση το μέγεθος του δείγματος ισούται με $n = 799$, μπορούμε να χρησιμοποιήσουμε το διάστημα εμπιστοσύνης που δόθηκε στη σχέση (10.9). Επομένως, ένα (προσεγγιστικό) 95% διάστημα εμπιστοσύνης για το $T(F)$ είναι το

$$1.778 \pm 1.96 \times 0.171 = [1.441, 2.115].$$

Παρατηρούμε ότι το παραπάνω διάστημα δεν περιέχει την τιμή 0, το οποίο υποδεικνύει ότι τα δεδομένα δεν προέρχονται από κανονική κατανομή (η λοξότητα της κανονικής κατανομής είναι ίση με 0).

10.6 Ασκήσεις

Άσκηση 10.1. Να αποδείξετε ότι ισχύουν οι σχέσεις (10.5) και (10.6).

Άσκηση 10.2. Υποθέστε ότι ο εκτιμητής T_n γράφεται στη μορφή $T_n = \frac{1}{n}\sum_{i=1}^{n}\alpha(X_i)$, όπου $\alpha(x)$ είναι μια γραμμική συνάρτηση του x. Να δείξετε ότι οι ψευδοτιμές $\widetilde{T}_i$ που ορίζονται στη σχέση (10.7) είναι ανεξάρτητες τυχαίες μεταβλητές.

Άσκηση 10.3. Έστω δείγμα μεγέθους 3 και οι διατεταγμένες παρατηρήσεις $X_1 < X_2 < X_3$. Δείξτε ότι η στατιστική συνάρτηση

$$\frac{7X_2 - 2(X_1 + X_3)}{3}.$$

είναι ο jackknife εκτιμητής της διαμέσου.

Άσκηση 10.4. Έστω τυχαίο δείγμα $X_1, \ldots, X_n$ από την κατανομή $B(1,p)$ και υποθέτουμε ότι ενδιαφερόμαστε για την εκτίμηση του p^2. Θεωρήστε τον εκτιμητή μέγιστης πιθανοφάνειας $T_n = \frac{1}{n^2}\left(\sum_{i=1}^{n} X_i\right)^2$. Να δείξετε ότι ο εκτιμητής T_n έχει μεροληψία ίση με

$$\text{bias}(T_n) = \frac{p(1-p)}{n}.$$

Στη συνέχεια, να αποδείξετε ότι ο αντίστοιχος jackknife εκτιμητής είναι ο

$$T_{\text{jack}} = \frac{1}{n(n-1)}\sum_{i=1}^{n} X_i\left(\sum_{i=1}^{n} X_i - 1\right),$$

ο οποίος είναι αμερόληπτος.

Άσκηση 10.5. Να εκτιμηθεί μέσω jackknife το τυπικό σφάλμα του εκτιμητή αντικατάστασης της διασποράς στα δεδομένα του Παραδείγματος 10.1.

Άσκηση 10.6. Προσομοιώστε $n = 30$ το πλήθος παρατηρήσεις από τη διδιάστατη κανονική κατανομή $\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}2 & 1\\1 & 2\end{pmatrix}\right)$ και υπολογίστε την jackknife εκτίμηση του συντελεστή συσχέτισης, καθώς και το τυπικό σφάλμα της εκτίμησης. Υπόδειξη: για την προσομοίωση χρησιμοποιήστε την εντολή `rmvnorm(...)` από τη βιβλιοθήκη `mvtnorm` της `R`.

Άσκηση 10.7. Έστω τυχαίο δείγμα $X_1, \ldots, X_n$ από κατανομή με μέση τιμή μ και διασπορά σ^2. Να δείξετε ότι ο εκτιμητής αντικατάστασης της διασποράς

$$T_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

γράφεται ως τετραγωνική στατιστική συνάρτηση της μορφής (10.14), με

$$\alpha^{(n)}(x) = \frac{n-1}{n}[(x-\mu)^2 + \sigma^2] \quad \text{και} \quad \beta^{(n)}(x,y) = -2(x-\mu)(y-\mu).$$

Άσκηση 10.8. Ο συντελεστής Gini (βλ. Gini, 1936) χρησιμοποιείται από τους οικονομολόγους για μέτρηση της ανισοκατανομής του εισοδήματος σε έναν πληθυσμό και δίνεται από τη σχέση

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\bar{x}},$$

με τιμές κοντά στο 0 να εκφράζουν ισοκατανομή του εισοδήματος, ενώ τιμές κοντά στο 1 να εκφράζουν ανισοκατανομή του εισοδήματος.

Θεωρήστε τα δεδομένα του Πίνακα 10.2, στον οποίο καταγράφονται τα εισοδήματα (σε εκατοντάδες χιλιάδες ευρώ) ενός τυχαίου δείγματος 25 ατόμων από τρία διαφορετικά χωριά Α, Β και C. Για κάθε χωριό χρησιμοποιήστε jackknife για

1. να εκτιμήσετε τη μεροληψία του δειγματικού συντελεστή Gini,
2. να εκτιμήσετε το τυπικό σφάλμα του δειγματικού συντελεστή Gini, και
3. να υπολογίσετε ένα 95% προσεγγιστικό διάστημα εμπιστοσύνης ίσων ουρών για τον πληθυσμιακό συντελεστή Gini.

A/A	A	B	C
1	0.38	0.10	1.28
2	0.05	0.88	1.86
3	0.68	0.42	0.83
4	1.06	0.18	0.17
5	0.55	2.64	1.34
6	0.03	0.31	0.44
7	1.23	6.82	0.52
8	1.43	0.70	1.76
9	0.52	0.40	0.26
10	0.77	0.49	0.55
11	3.42	1.86	1.45
12	0.01	0.77	1.15
13	1.56	0.01	0.57
14	1.07	1.49	0.07
15	3.60	0.01	2.80
16	1.30	0.13	1.06
17	1.28	2.58	1.18
18	0.00	0.43	0.59
19	3.52	0.35	1.04
20	0.18	4.83	0.14
21	4.53	0.50	0.00
22	0.78	4.36	0.53
23	0.32	0.78	1.18
24	0.76	0.67	0.22
25	0.51	0.92	0.63

Πίνακας 10.2: Εισόδημα για ένα τυχαίο δείγμα 25 νοικοκυριών σε 3 διαφορετικά χωριά (πηγή: Καρλής, 2004).

Άσκηση 10.9. Έστω X (Y, αντίστοιχα) η τυχαία μεταβλητή που παριστάνει τον πληθυσμό σε χιλιάδες μιας πόλης των Ηνωμένων Πολιτειών Αμερικής το έτος 1920 (1930, αντίστοιχα). Έστω (X_i, Y_i), $i = 1, \dots, n$, τυχαίο δείγμα n το πλήθος πόλεων των Ηνωμένων Πολιτειών Αμερικής τα έτη 1920 και 1930, αντίστοιχα. Ενδιαφέρει η εκτίμηση του συνολικού πληθυσμού, έστω d, των ΗΠΑ το 1930, με βάση το συγκεκριμένο δείγμα και γνωρίζοντας ότι το 1920 ο πληθυσμός των ΗΠΑ ήταν ίσος με a. Αν οι ΗΠΑ έχουν συνολικά k το πλήθος πόλεις, τότε

$$\mathrm{E}(X) = a/k, \quad \mathrm{E}(Y) = d/k$$

και, επομένως, ο συνολικός πληθυσμός d το 1930 θα είναι

$$d = a\theta,$$

όπου $\theta = \frac{\mathrm{E}(Y)}{\mathrm{E}(X)}$. Να θεωρήσετε τα δεδομένα του Πίνακα 10.3 και τον εκτιμητή αντικατάστασης του λόγου μέσων τιμών θ, δηλαδή

$$T_n = \frac{\bar{Y}_n}{\bar{X}_n}$$

όπου $\bar{X}_n$ και $\bar{Y}_n$ οι δειγματικοί μέσοι κατά το 1920 και 1930, αντίστοιχα. Να εκτιμήσετε τη μεροληψία και το τυπικό σφάλμα του T_n μέσω jackknife. Να υπολογίσετε και ένα 95% προσεγγιστικό διάστημα εμπιστοσύνης για το d το 1930.

x_i	y_i	x_i	y_i	x_i	y_i
138	143	76	80	67	67
93	104	381	464	120	115
61	69	387	459	172	183
179	260	78	106	66	86
48	75	60	57	46	65
37	63	507	634	121	113
29	50	50	64	44	58
23	48	77	89	64	63
30	111	64	77	56	142
2	50	40	60	40	64
38	52	136	139	116	130
46	53	243	291	87	105
71	79	256	288	43	61
25	57	94	85	43	50
298	317	36	46	161	232
74	93	45	53	36	54
50	58				

Πίνακας 10.3: Μέγεθος (σε χιλιάδες κατοίκους) 49 πόλεων των ΗΠΑ το 1920 (x_i) και το 1930 (y_i) (πηγή: Cochran, 2007).

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Καρλής, Δ. (2004). *Υπολογιστική Στατιστική*. Οικονομικό Πανεπιστήμιο Αθηνών.

Ξενόγλωσση

Cochran, W. G. (2007). *Sampling techniques*. John Wiley & Sons.

Efron, B. and Stein, C. (1981). The Jackknife Estimate of Variance. *The Annals of Statistics*, *9*(3), pp. 586–596.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. SIAM.

Gini, C. (1936). On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series*, *208*(1), pp. 73–79.

Hinkley, D. V. (1978). Improving the jackknife with special reference to correlation estimation. *Biometrika*, *65*(1), pp. 13–21.

Miller, R. G. (1964). A trustworthy jackknife. *The Annals of Mathematical Statistics*, *35*(4), pp. 1594–1605.

Miller, R. G. (1968). Jackknifing Variances. *The Annals of Mathematical Statistics*, *39*(2), pp. 567–582.

Miller, R. G. (1974). The jackknife-a review. *Biometrika*, *61*(1), pp. 1–15.

Quenouille, M. H. (1949). Problems in Plane Sampling. *The Annals of Mathematical Statistics*, *20*(3), pp. 355–375.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, *43*(3/4), pp. 353–360.

Schucany, W., Gray, H. and Owen, D. (1971). On bias reduction in estimation. *Journal of the American Statistical Association*, *66*(335), pp. 524–533.

Shao, J. and Tu, D. (2012). *The jackknife and bootstrap*. Springer Science & Business Media.

Tukey, J. (1958). Bias and confidence in not quite large samples. *Ann. Math. Statist.*, *29*, p. 614.

ΚΕΦΑΛΑΙΟ 11

Η ΜΕΘΟΔΟΣ BOOTSTRAP

Σύνοψη

Η μέθοδος bootstrap είναι μια άλλη τεχνική επαναδειγματοληψίας, γενικότερη της μεθόδου jackknife, που επιτρέπει την εκτίμηση της μεταβλητότητας εκτιμητών, χωρίς τη χρησιμοποίηση επιπλέον δεδομένων, παρά μόνο με τη βοήθεια του αρχικού δείγματος. Η αύξηση της υπολογιστικής ισχύος που σημειώθηκε κατά τη δεκαετία του '80 σε συνδυασμό με την ευελιξία εφαρμογής του bootstrap σε ευρύ φάσμα προβλημάτων, αλλά και η αμεσότητα που παρέχει η μέθοδος διαισθητικά, συνέβαλαν καθοριστικά στην εδραίωσή της ως ένα από τα πιο δημοφιλή μη παραμετρικά εργαλεία. Στο κεφάλαιο αυτό, αρχικά, παρουσιάζεται, μέσω απλών παραδειγμάτων για την εκτίμηση της μέσης τιμής και της διακύμανσης της κατανομής του πληθυσμού, η βασική ιδέα της μεθόδου bootstrap. Κατόπιν, περιγράφονται τα bootstrap διαστήματα εμπιστοσύνης και έλεγχοι υποθέσεων για έναν ή περισσότερους πληθυσμούς. Τέλος, παρουσιάζονται περαιτέρω εφαρμογές της μεθόδου σε προβλήματα παλινδρόμησης και εκτίμησης της συνάρτησης πυκνότητας.

Προαπαιτούμενη γνώση:

Εκτιμητική (με σημείο και με διάστημα) και έλεγχοι υποθέσεων.

Στοιχειώδεις γνώσεις προσομοίωσης Monte Carlo.

Κεφάλαια 2, 3 και 10 του παρόντος συγγράμματος.

Προσδοκώμενα μαθησιακά αποτελέσματα:

Ο/η φοιτητής/τρια ολοκληρώνοντας την ενότητα αυτή θα μπορεί να εφαρμόζει την τεχνική bootstrap για την εκτίμηση της μεροληψίας και του τυπικού σφάλματος ενός εκτιμητή. Επίσης, θα έχει κατανοήσει τα bootstrap διαστήματα εμπιστοσύνης, καθώς και τον τρόπο διεξαγωγής με τη βοήθεια της μεθόδου bootstrap ελέγχων υποθέσεων για έναν ή περισσότερους πληθυσμούς.

Γλωσσάριο επιστημονικών όρων

- Εμπειρική αθροιστική συνάρτηση κατανομής
- Bootstrap διαστήματα εμπιστοσύνης
- Bootstrap εκτίμηση μεροληψίας
- Bootstrap εκτίμηση συνάρτησης πυκνότητας
- Bootstrap εκτίμηση τυπικού σφάλματος
- Monte Carlo

Μπατσίδης, Α., Παπασταμούλης, Π., Πετρόπουλος, Κ., & Ρακιτζής, Α. (2022). *Μη Παραμετρική Στατιστική.* [Προπτυχιακό εγχειρίδιο].

11.1 Το Bootstrap

Η μεθοδολογία bootstrap (Efron, 1979) εισήχθη λίγο πριν από την αυγή της δεκαετίας του '80. Πρόκειται για μία περίοδο δραστικών αλλαγών στην υπολογιστική επιστήμη (κατ' επέκταση και στην Υπολογιστική Στατιστική), κυρίως λόγω της εξάπλωσης των προσωπικών ηλεκτρονικών υπολογιστών. Η εκτέλεση προγραμμάτων που μέχρι πρότινος απαιτούσε εξειδικευμένο προσωπικό να χειρίζεται διάτρητες κάρτες σε κεντρικές μονάδες υπολογιστών, πλέον μπορούσε να γίνει άμεσα. Χαρακτηριστικό αυτού του κλίματος, και αφού οι πιο παραδοσιακοί Στατιστικοί είχαν ήδη αρχίσει να ξύνουν τα μολύβια τους για τη μελέτη θεωρητικών ιδιοτήτων του bootstrap, είναι η περιγραφή που ακολουθεί:

> Τυπικές υποθέσεις για τα δεδομένα αντικαθίστανται από υπολογισμούς μεγάλης κλίμακας. Το "bootstrap" είναι μία μέθοδος μέσω της οποίας έχει αναθεωρηθεί η αξιοπιστία προηγούμενων επιστημονικών συμπερασμάτων.

η οποία συνόδευε άρθρο των Diaconis and Efron (1983) στο δημοφιλές περιοδικό Scientific American.

Αρχικά, θα χρησιμοποιήσουμε την περίπτωση ενός τυχαίου δείγματος από έναν πληθυσμό για να παρουσιάσουμε τη βασική ιδέα των μεθόδων bootstrap. Έστω, λοιπόν, ότι έχουμε στη διάθεσή μας παρατηρήσεις $\boldsymbol{x} = (x_1, \dots, x_n)$. Υποθέτουμε ότι αυτές οι τιμές είναι πραγματοποιήσεις ανεξάρτητων και ισόνομων τυχαίων μεταβλητών X_i, $i = 1, \dots, n$, με $X_i \sim F$, όπου F κάποια αθροιστική συνάρτηση κατανομής. Σκοπός της ανάλυσης είναι η συμπερασματολογία για κάποιο πληθυσμιακό χαρακτηριστικό (συναρτησιακό) $\theta = T(F)$. Υποθέτουμε, εδώ, ότι θα έχουμε στη διάθεσή μας έναν εκτιμητή του θ, όπως για παράδειγμα τον εκτιμητή αντικατάστασης (θυμηθείτε τον Ορισμό 2.3)

$$T_n := T(F_n),$$

όπου F_n είναι η εμπειρική αθροιστική συνάρτηση κατανομής. Ενδιαφερόμαστε για την εκτίμηση του τυπικού σφάλματος του T_n, την κατασκευή διαστημάτων εμπιστοσύνης για το θ και, πιο γενικά, για τον προσδιορισμό της δειγματικής κατανομής του εκτιμητή.

Στο προηγούμενο κεφάλαιο γνωρίσαμε την τεχνική jackknife, η οποία μας επιτρέπει την εκτίμηση τυπικών σφαλμάτων εκτιμητών. Ωστόσο, είδαμε ότι η τεχνική αυτή μπορεί να χρησιμοποιηθεί σε περιορισμένες περιπτώσεις. Μια πιο ευέλικτη και γενική τεχνική είναι η μέθοδος bootstrap (Efron, 1979). Υπολογιστικά, η μέθοδος jackknife είναι λιγότερο απαιτητική, αλλά η μέθοδος bootstrap έχει περισσότερα πλεονεκτήματα. Ενδεικτικά, η μέθοδος bootstrap μας δίνει τη δυνατότητα εκτίμησης της δειγματικής κατανομής εκτιμητών συναρτησιακών, κάτι που δεν μπορεί να απαντηθεί μέσω της μεθόδου jackknife. Γενικά, η τεχνική μπορεί να εφαρμοστεί τόσο σε παραμετρικά όσο και σε μη παραμετρικά πλαίσια. Σε αυτό το σύγγραμμα, ο όρος bootstrap θα αναφέρεται πάντα στη μη παραμετρική περίπτωση.

Σε αυτό το σημείο, ας θυμηθούμε τι σημαίνει δειγματική κατανομή ενός εκτιμητή βάσει επαναλαμβανόμενης δειγματοληψίας από τον υπό μελέτη πληθυσμό.

1. Πάρε πολλά δείγματα (θεωρητικά: άπειρα) μεγέθους n από τον *πληθυσμό*

$$X_i \sim F, \quad \text{ανεξάρτητα για} \quad i = 1, \dots, n.$$

2. Υπολόγισε την εκτίμηση που προκύπτει σε κάθε δείγμα.
3. Η κατανομή (των εκτιμήσεων) που προκύπτει είναι η *δειγματική κατανομή του εκτιμητή*.

Στην πράξη, όμως, έχουμε μόνο ένα δείγμα από τον πληθυσμό. Η ιδέα του Efron για να εκτιμήσει τη δειγματική κατανομή ενός εκτιμητή είναι να αντικαταστήσει την πληθυσμιακή αθροιστική συνάρτηση κατανομής (F) στο Βήμα 1 με μία εκτίμηση αυτής και, πιο συγκεκριμένα, από την εμπειρική αθροιστική συνάρτηση κατανομής (F_n). Τότε προκύπτουν τα ακόλουθα βήματα:

1. Πάρε πολλά δείγματα μεγέθους n από την εμπειρική αθροιστική συνάρτηση κατανομής των δεδομένων
$$X_i^\star \sim F_n, \quad \text{ανεξάρτητα για} \quad i = 1, \dots, n.$$
2. Υπολόγισε την εκτίμηση που προκύπτει σε κάθε δείγμα.
3. Η κατανομή που προκύπτει είναι η *bootstrap* κατανομή του εκτιμητή και χρησιμοποιείται για να προσεγγίσει την πραγματική του κατανομή.

Αρχικά, να παρατηρήσουμε ότι στην περίπτωση όπου $F_n = F$, η bootstrap κατανομή συμπίπτει με την πραγματική κατανομή του εκτιμητή. Αυτό, ουσιαστικά, σημαίνει ότι το bootstrap είναι συνεπές κατά Fisher (Fisher, 1922).

Το κομβικό σημείο για την εφαρμογή του bootstrap είναι ο υπολογισμός της bootstrap κατανομής. Δυστυχώς, ο αναλυτικός υπολογισμός αυτής είναι δυνατός μόνο σε πολύ απλές περιπτώσεις (μην ξεχνάτε ότι η κατανομή F θεωρείται άγνωστη!). Στην πράξη, το Βήμα 1 προσεγγίζεται μέσω προσομοίωσης Monte Carlo. Στην ενότητα που ακολουθεί, θα εξηγήσουμε ότι αυτό, τελικά, ισοδυναμεί με επαναλαμβανόμενη δειγματοληψία με επανάθεση από το παρατηρηθέν δείγμα.

11.1.1 Ο αλγόριθμος Monte Carlo Bootstrap για ένα τυχαίο δείγμα

Από το Κεφάλαιο 2 γνωρίζουμε ότι η εμπειρική αθροιστική συνάρτηση κατανομής που ορίζεται από τη σχέση:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \leqslant x\}$$

είναι ένας συνεπής εκτιμητής της F. Έστω, λοιπόν, ότι μας ενδιαφέρει η συμπερασματολογία για το συναρτησιακό $\theta = T(F)$ με βάση έναν εκτιμητή $T_n = T_n(\boldsymbol{X})$, όπου $\boldsymbol{X} = (X_1, \dots, X_n)$ είναι το τυχαίο δείγμα. Η εκτίμηση που προκύπτει με βάση αυτές τις n το πλήθος παρατηρήσεις $x_1, \dots, x_n$ του δείγματος είναι, φυσικά, η $T_n(\boldsymbol{x})$. Αυτό που μας ενδιαφέρει πρωτίστως είναι η δειγματική κατανομή του εκτιμητή. Η ιδέα του bootstrap είναι ότι μπορούμε να προσομοιώσουμε από την F_n για να εκτιμήσουμε τη δειγματική κατανομή του εκτιμητή και ποσότητες που απορρέουν από αυτήν, όπως για παράδειγμα τη μεροληψία και το τυπικό σφάλμα.

Στο σημείο αυτό, θα πρέπει να θυμηθούμε ότι F_n είναι συνάρτηση κατανομής διακριτής τυχαίας μεταβλητής, καθώς το στήριγμά της είναι το σύνολο των n το πλήθος παρατηρηθέντων τιμών $\{x_1, \dots, x_n\}$, με καθεμία από αυτές τις τιμές να έχει μάζα πιθανότητας ίση με $1/n$. Η προσομοίωση n το πλήθος τιμών από την F_n ισοδυναμεί με τυχαία δειγματοληψία με επανάθεση από το παρατηρηθέν δείγμα. Επομένως, για τη Monte Carlo εκτίμηση χαρακτηριστικών (τυπικό σφάλμα κ.λπ.) του εκτιμητή $T_n := T(F_n)$ αρκεί να λάβουμε με επανάθεση πολλά τέτοια δείγματα (έστω B το πλήθος) μεγέθους n. Έστω

$$\boldsymbol{x}_b := (x_{b,1}^\star, \dots, x_{b,n}^\star)$$

με $b = 1, \dots, B$, τα δείγματα αυτά που προκύπτουν με επανάθεση. Κατόπιν, για καθένα προσομοιωμένο δείγμα υπολογίζουμε την τιμή του εκτιμητή, την οποία θα τη συμβολίζουμε με $T_{n,b}^\star = T_n(\boldsymbol{x}_b)$. Αυτές οι τιμές χρησιμοποιούνται στη συνέχεια για εκτίμηση μεροληψίας, τυπικών σφαλμάτων και κατασκευή διαστημάτων εμπιστοσύνης. Ας συνοψίσουμε τα βήματα της διαδικασίας στον παρακάτω αλγόριθμο.

1. Για $b = 1, \dots, B$
   1.1. Κάνε δειγματοληψία με επανάθεση n το πλήθος τιμών από τα $(x_1, \dots, x_n)$. Έστω ότι προκύπτουν οι τιμές
   $$(x_{b,1}^\star, \dots, x_{b,n}^\star).$$
   Σημείωση: το βήμα αυτό ισοδυναμεί με προσομοίωση ενός δείγματος μεγέθους n από την F_n.

1.2. Υπολόγισε την εκτίμηση για το συγκεκριμένο bootstrap δείγμα, δηλαδή υπολόγισε την

$$T^{\star}_{n,b} := T_n(x^{\star}_{b,1}, \dots, x^{\star}_{b,n}).$$

2. Μέσω των $\{T^{\star}_{n,b}; b = 1, \dots, B\}$ εκτιμάται η δειγματική κατανομή του $T_n = T_n(X_1, \dots, X_n)$, η οποία μπορεί να χρησιμοποιηθεί για περαιτέρω συμπερασματολογία όπως,
   - εκτίμηση τυπικού σφάλματος του εκτιμητή,
   - κατασκευή διαστημάτων εμπιστοσύνης.

Παράδειγμα 11.1. Έστω τυχαίο δείγμα $n = 20$ το πλήθος τιμών

$$(19.15, 19.43, 18.20, 21.41, 24.18, 23.27, 18.92, 18.79, 19.59, 19.38,$$
$$21.91, 19.22, 19.71, 19.26, 21.22, 16.67, 17.56, 22.55, 20.47, 21.87).$$

Να βρεθούν οι δειγματικές κατανομές των $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ και $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, κάνοντας χρήση της υπόθεσης ότι τα δεδομένα προέρχονται από κανονικό πληθυσμό με μέση τιμή μ και διασπορά σ^2. Στη συνέχεια, να ληφθεί δείγμα $B = 1000$ bootstrap τιμών και να κατασκευαστεί ένα ιστόγραμμα σχετικών συχνοτήτων για τις δειγματικές κατανομές των $\bar{X}$ και $\hat{\sigma}^2$. Να συγκρίνετε τα ιστογράμματα με τις συναρτήσεις πυκνότητας πιθανότητας των πραγματικών δειγματικών κατανομών (αφού εκτιμηθούν οι άγνωστες παράμετροι αυτών).

Λύση Παραδείγματος 11.1. Από τις ιδιότητες της κανονικής κατανομής είναι εύκολο να δειχθεί ότι [1]

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{και} \quad \hat{\sigma}^2 \sim \mathcal{G}\left(\frac{n-1}{2}, \frac{n}{2\sigma^2}\right).$$

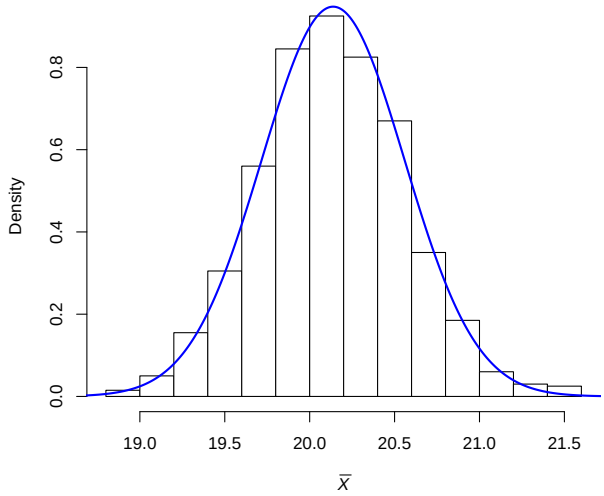
Παρατηρήστε ότι οι παραπάνω κατανομές δεν είναι πλήρως καθορισμένες, αφού εξαρτώνται από τις άγνωστες παραμέτρους μ και σ^2. Από το διαθέσιμο σύνολο δεδομένων προκύπτει ότι οι σημειακές εκτιμήσεις αυτών είναι $\bar{x} = 20.13$ και $\sum_{i=1}^{20}(x_i - \bar{x})^2/20 = 3.544$, οπότε οι εκτιμήσεις των θεωρητικών δειγματικών κατανομών, διατηρώντας δύο δεκαδικά ψηφία, είναι

$$\bar{X} \sim \mathcal{N}(20.13, 0.18) \quad \text{και} \quad \hat{\sigma}^2 \sim \mathcal{G}(9.5, 2.82).$$

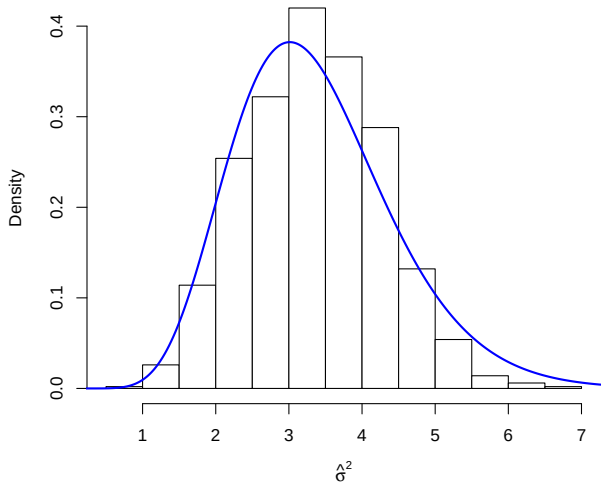
Στη συνέχεια, θα λάβουμε 1000 bootstrap δείγματα των δύο εκτιμητών από το παρατηρηθέν δείγμα, μέσω της `R`. Για τον σκοπό αυτό θα χρησιμοποιηθεί η εντολή `sample(n, n, replace = TRUE)`, η οποία λαμβάνει τυχαίο δείγμα με επανάθεση μεγέθους n από n στοιχεία.

```
x <- c(19.15, 19.43, 18.2, 21.41, 24.18, 23.27, 18.92, 18.79, 19.59,
19.38, 21.91, 19.22, 19.71, 19.26, 21.22, 16.67, 17.56, 22.55, 20.47,
   21.87)
n <- length(x)
# set the seed for reproducibility purposes!
set.seed(1)
theta_hat1 <- mean(x)
theta_hat2 <- (n-1)*var(x)/n
B = 1000
theta_values1 <- theta_values2 <- numeric(B)
for (i in 1:B){
  y <- x[sample(n, n, replace = TRUE)]
  theta_values1[i] <- mean(y)
  theta_values2[i] <- (n-1)*var(y)/n
}
```

[1] Υπενθύμιση: $\frac{\sum_i (X_i - \bar{X})^2}{\sigma^2} \sim X^2_{n-1} \equiv \mathcal{G}((n-1)/2, 1/2)$ (shape - rate).



(α) bootstrap στο $\bar{X}$



(β) bootstrap στο $\hat{\sigma}^2$

Σχήμα 11.1: Ιστόγραμμα 1000 bootstrap δειγμάτων μαζί με τη συνάρτηση πυκνότητας πιθανότητας (μπλε γραμμή) της αντίστοιχης θεωρητικής κατανομής (μετά την εκτίμηση άγνωστων παραμέτρων) στα κανονικά δεδομένα του Παραδείγματος 11.1.

```
15
16 xSeq1 <- seq(18, 22, length = 1000)
17 hist(theta_values1,  freq=F, xlab = bquote(bar(italic(X))),main='',
       col= 'white')
18 points(xSeq1, dnorm(xSeq1, theta_hat1, sqrt(theta_hat2/n)), type = 'l'
       , lwd = 2, col = 'blue')
19 xSeq2 <- seq(0, 10, length = 1000)
20 hist(theta_values2,  freq=F, xlab = bquote(hat(italic(sigma))^2),main=
       '', col= 'white')
21 points(xSeq2, dgamma(xSeq2, shape = (n - 1)/2, rate = n/(2*theta_hat2)
       ), type = 'l', lwd = 2, col = 'blue')
```

Το ιστόγραμμα των 1000 bootstrap τιμών του $\bar{X}$ και του $\hat{\sigma}^2$ παρατίθεται στο Σχήμα 11.1.(α) και 11.1.(β), αντίστοιχα. Παρατηρήστε ότι και στις δύο περιπτώσεις το ιστόγραμμα είναι αρκετά «κοντά» στην αντίστοιχη συνάρτηση πυκνότητας. Άρα, μπορούμε να χρησιμοποιήσουμε τις προσομοιωμένες bootstrap τιμές των στατιστικών συναρτήσεων ως μια καλή προσέγγιση της (υποτιθέμενης) άγνωστης δειγματικής κατανομής των εκτιμητών αυτών. □

Παρατήρηση 11.1. Υπάρχουν δύο πηγές σφαλμάτων στη συμπερασματολογία όταν χρησιμοποιούνται τεχνικές επαναδειγματοληψίας, όπως το Monte Carlo Bootstrap: **στατιστικό σφάλμα** και το **σφάλμα λόγω προσομοίωσης**. Το πρώτο είδος σφάλματος αναφέρεται στο γεγονός ότι χρησιμοποιούμε την F_n (δηλαδή μία προσέγγιση της F) και όχι την ίδια την F για τη συμπερασματολογία. Το δεύτερο είδος σφάλματος αναφέρεται στο γεγονός ότι χρησιμοποιούνται εμπειρικές εκτιμήσεις που προκύπτουν από προσομοίωση πεπερασμένου πλήθους δειγμάτων από την F_n. Το σφάλμα αυτό θα μπορούσε να εξαλειφθεί αν μπορούσαμε να κάνουμε προσομοίωση άπειρων δειγμάτων από την F_n, το οποίο όμως είναι αδύνατο και όχι απαραίτητο στην πράξη. Για μια πολύ ενδιαφέρουσα παρουσίαση των δύο αυτών πηγών σφαλμάτων παραπέμπουμε τον ενδιαφερόμενο αναγνώστη στην Ενότητα 2.5 των Davison and Hinkley (1997).

Παρατήρηση 11.2. Ο Brad Efron, που εισήγαγε την τεχνική bootstrap, συνέστησε ότι μερικές δεκάδες bootstrap δειγμάτων είναι αρκετές για την εκτίμηση τυπικών σφαλμάτων, ενώ μερικές εκατοντάδες δειγμάτων είναι αρκετές για την κατασκευή διαστημάτων εμπιστοσύνης (Efron and Tibshirani, 1986; Efron

and Tibshirani, 1994). Βέβαια, την εποχή εκείνη, η διαθέσιμη υπολογιστική ταχύτητα ήταν περιορισμένη σε σχέση με τα σημερινά δεδομένα, οπότε στις μέρες μας η σύσταση είναι να θεωρούνται μερικές χιλιάδες bootstrap δειγμάτων. Μπορείτε να διαπιστώσετε και μόνοι σας ότι ο χρόνος εκτέλεσης του παραπάνω αλγορίθμου παραμένει πολύ μικρός, ακόμα και αν αυξήσετε τις επαναλήψεις σε $B = 10000$.

11.1.2 Bootstrap εκτίμηση μεροληψίας και τυπικού σφάλματος

Σε αυτήν την ενότητα, θα δούμε πώς μπορεί να αξιοποιηθεί το bootstrap δείγμα για την εκτίμηση διάφορων ποσοτήτων που μας ενδιαφέρουν. Αρχικά, θεωρούμε το πρόβλημα εκτίμησης της μεροληψίας του T_n, δηλαδή της ποσότητας

$$\text{bias}(T_n) = \text{E}(T_n) - \theta = \text{E}(T_n) - T(F).$$

Φυσικά, η F είναι άγνωστη και κατ' επέκταση το ίδιο ισχύει και για το συναρτησιακό $\theta = T(F)$. Όμως, από το παρατηρηθέν δείγμα έχουμε ήδη την εκτίμηση T_n (π.χ. μέσω του εκτιμητή αντικατάστασης $T_n = T(F_n)$). Τι γίνεται, όμως, με την (άγνωστη) μέση τιμή του εκτιμητή, δηλαδή με την $\text{E}(T_n)$; Με βάση το bootstrap μπορούμε να την εκτιμήσουμε με την

$$T_n^\star := \frac{1}{B}\sum_{b=1}^{B} T_{n,b}^\star. \tag{11.1}$$

Παρατηρήστε ότι η $T_n^\star$ που δίνεται στη σχέση (11.1) δεν είναι τίποτε άλλο παρά ο δειγματικός μέσος των B προσομοιωμένων bootstrap τιμών $T_{n,b}^\star$. Επομένως, μια εκτίμηση της $\text{bias}(T_n)$ είναι απλώς η διαφορά μεταξύ του δειγματικού μέσου των bootstrap τιμών ($T_n^\star$) και της σημειακής εκτίμησης T_n, σύμφωνα με τον ορισμό που ακολουθεί.

Ορισμός 11.1

Η bootstrap εκτίμηση της μεροληψίας του εκτιμητή T_n είναι

$$\widehat{\text{bias}}(T_n) = T_n^\star - T_n \tag{11.2}$$

όπου $T_n^\star := \frac{1}{B}\sum_{b=1}^{B} T_{n,b}^\star$ ο δειγματικός μέσος των B προσομοιωμένων bootstrap τιμών $T_{n,b}^\star$.

Τονίζουμε ότι η $\widehat{\text{bias}}(T_n)$ είναι **εκτίμηση** της μεροληψίας του εκτιμητή και, προφανώς, μπορεί να είναι διαφορετική από μηδέν, ακόμα και για αμερόληπτους εκτιμητές. Να παρατηρήσουμε στο σημείο αυτό ότι, κατ' αναλογία με το jackknife, μπορούμε να κατασκευάσουμε έναν βελτιωμένο εκτιμητή αφού διορθώσουμε τη μεροληψία του αρχικού εκτιμητή. Έτσι, μετά τη διόρθωση μεροληψίας, ο bootstrap εκτιμητής ορίζεται ως

$$T_{\text{boot}} = T_n - \widehat{\text{bias}}(T_n) = 2T_n - T_n^\star. \tag{11.3}$$

Μία άλλη ποσότητα που παρουσιάζει ιδιαίτερο ενδιαφέρον για την κατασκευή διαστημάτων εμπιστοσύνης και για την εξαγωγή συμπερασμάτων είναι η διασπορά του εκτιμητή, δηλαδή

$$\text{Var}(T_n) = \text{E}\{T_n - \text{E}(T_n)\}^2.$$

Με βάση το bootstrap μπορούμε να εκτιμήσουμε τη διασπορά και το τυπικό σφάλμα του T_n ως ακολούθως.

Ορισμός 11.2

Η bootstrap εκτίμηση της διασποράς του εκτιμητή είναι:

$$v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^{\star} - T_{n}^{\star}\right)^2, \tag{11.4}$$

ενώ η bootstrap εκτίμηση του τυπικού σφάλματος του εκτιμητή είναι:

$$\text{se}_{\text{boot}} = \sqrt{v_{\text{boot}}}, \tag{11.5}$$

Σημειώνουμε ότι μερικοί συγγραφείς (Efron and Tibshirani, 1986) θεωρούν την ποσότητα $B-1$ στον παρονομαστή της σχέσης (11.4), αντί του B. Συνήθως, ο αριθμός των bootstrap δειγμάτων είναι μεγάλος (θυμηθείτε την Παρατήρηση 11.2), οπότε η διαφορά μεταξύ των δύο εκτιμήσεων είναι στην πραγματικότητα αμελητέα.

Παράδειγμα 11.2 (Συνέχεια Παραδείγματος 11.1)**.** Να εκτιμηθούν μέσω bootstrap το τυπικό σφάλμα και η μεροληψία του $T_n = \bar{X}$ στα δεδομένα του Παραδείγματος 11.1 και να συγκριθούν με τις αντίστοιχες εκτιμήσεις που προκύπτουν μέσω jackknife και αναλυτικών υπολογισμών.

Λύση Παραδείγματος 11.2. Για τον δειγματικό μέσο, από το Παράδειγμα 11.1, γνωρίζουμε ότι $T_n = \bar{x} = 20.138$. Ο δειγματικός μέσος των $B = 1000$ bootstrap τιμών ισούται με

$$T_n^{\star} = \frac{1}{1000}\sum_{b=1}^{1000} T_{n,b} = \frac{1}{1000}\sum_{b=1}^{1000}\left(\frac{1}{20}\sum_{i=1}^{20} x_{b,i}^{\star}\right) = 20.132,$$

όπως προκύπτει χρησιμοποιώντας την `R`, κρατώντας ακρίβεια τριών δεκαδικών ψηφίων. Οπότε, τώρα μπορούμε να εκτιμήσουμε τη μεροληψία του $\bar{X}$ μέσω της σχέσης (11.2). Είναι:

$$\widehat{\text{bias}}(T_n) = T_n^{\star} - T_n = 20.132 - 20.138 = -0.006.$$

Το παραπάνω μπορεί να υπολογιστεί απευθείας μέσω της `R`, υπολογίζοντας τον δειγματικό μέσο των 1000 τιμών των δειγματικών μέσων ανά δείγμα που έχουν αποθηκευτεί στο διάνυσμα `theta_values1` (βλ. τον κώδικα του Παραδείγματος 11.1):

```
> theta_star1 <- mean(theta_values1)
round(theta_star1, 3)
[1] 20.132
# bias estimate for mean
> round(theta_star1 - theta_hat1, 3)
[1] -0.006
```

Να τονίσουμε, εδώ, ότι είναι γνωστό ότι ο δειγματικός μέσος είναι αμερόληπτος εκτιμητής της μέσης τιμής του πληθυσμού, οπότε δεν υπάρχει λόγος να εκτιμηθεί η μεροληψία του $T_n = \bar{X}$. Σε κάθε περίπτωση όμως, είναι καθησυχαστικό το γεγονός ότι η εκτίμηση −0.006 (μοιάζει να) είναι πολύ κοντά στο μηδέν. Στη συνέχεια, θα εκτιμήσουμε μέσω bootstrap το τυπικό σφάλμα του δειγματικού μέσου. Από τη σχέση (11.5) και με τη βοήθεια της `R` προκύπτει ότι: $\text{se}_{\text{boot}} = \sqrt{v_{\text{boot}}} = 0.426$, το οποίο υπολογίζεται μέσω της `R` ως εξής:

```
> standard_error1 <- sqrt(sum((theta_values1 - theta_star1)^2)/(B-1))
> round(standard_error1, 3)
[1] 0.426
```

	$\bar{X}$		$\hat{\sigma}^2$	
	$\widehat{\text{se}}$	$\widehat{\text{bias}}$	$\widehat{\text{se}}$	$\widehat{\text{bias}}$
jackknife	0.432	0	1.053	−0.187
bootstrap	0.426	−0.006	0.925	−0.201
analutik'a	0.421	0	1.092	−0.177

Πίνακας 11.1: Εκτιμήσεις μεροληψίας και τυπικού σφάλματος των εκτιμητών $\bar{X}$ και $\hat{\sigma}^2$ στα δεδομένα του Παραδείγματος 11.1.

Φυσικά, στο συγκεκριμένο παράδειγμα, μπορούμε να προσδιορίσουμε αναλυτικά την έκφραση του τυπικού σφάλματος του δειγματικού μέσου. Ειδικότερα είναι

$$\text{se}(\bar{X}) = \sqrt{\frac{\sigma^2}{n}}.$$

Επομένως, μια εκτίμηση του τυπικού σφάλματος του δειγματικού μέσου είναι η:

$$\widehat{\text{se}}(\bar{X}) = \sqrt{\frac{\hat{\sigma}^2}{n}} = \sqrt{\frac{3.544}{20}} = 0.421.$$

Παρατηρήστε ότι για την εκτίμηση του άγνωστου σ^2, που εμφανίζεται στο τυπικό σφάλμα του $\bar{X}$ ($\text{se}(\bar{X})$) χρησιμοποιήσαμε τον εκτιμητή αντικατάστασης του σ^2 (δηλαδή τον $\hat{\sigma}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$) και προέκυψε μία εκτίμηση του τυπικού σφάλματος ($\widehat{\text{se}}(\bar{X})$). Τα αποτελέσματα συνοψίζονται στις δύο πρώτες στήλες του Πίνακα 11.1, μαζί με τις αντίστοιχες εκτιμήσεις που προκύπτουν από το jackknife (για τις τελευταίες δύο στήλες αυτού του πίνακα παραπέμπουμε στο επόμενο παράδειγμα). □

Παρατήρηση 11.3. Η εφαρμογή jackknife στα δεδομένα του προηγούμενου παραδείγματος και η επαλήθευση των αποτελεσμάτων που δίνονται στην πρώτη γραμμή του Πίνακα 11.1 αφήνονται ως άσκηση στον/στην αναγνώστη/στρια.

Παράδειγμα 11.3 (συνέχεια Παραδείγματος 11.1)**.** Να εκτιμηθούν μέσω bootstrap το τυπικό σφάλμα και η μεροληψία του $T_n = \hat{\sigma}^2$ στα δεδομένα του Παραδείγματος 11.1 και να συγκριθούν με τις αντίστοιχες εκτιμήσεις που προκύπτουν μέσω jackknife και αναλυτικών υπολογισμών.

Λύση Παραδείγματος 11.3. Για τον εκτιμητή αντικατάστασης της διασποράς, από το Παράδειγμα 11.1, γνωρίζουμε ότι $T_n = \hat{\sigma}^2 = 3.544$. Ο δειγματικός μέσος των $B = 1000$ bootstrap τιμών ισούται με

$$T_n^\star = \frac{1}{1000}\sum_{b=1}^{1000} T_{n,b} = \frac{1}{1000}\sum_{b=1}^{1000}\left(\frac{1}{20}\sum_{i=1}^{20}(x_{b,i}^\star - \bar{x}_{(b)}^\star)^2\right) = 3.343,$$

όπου $\bar{x}_{(b)}^\star = \frac{1}{20}\sum_{i=1}^{n} x_{b,i}^\star$. Τα αποτελέσματα προέκυψαν με τη βοήθεια της `R` διατηρώντας ακρίβεια τριών δεκαδικών ψηφίων. Οπότε, τώρα μπορούμε να εκτιμήσουμε τη μεροληψία του $\hat{\sigma}^2$ μέσω της σχέσης (11.2). Ειδικότερα, είναι:

$$\widehat{\text{bias}}(T_n) = T_n^\star - T_n = 3.343 - 3.544 = -0.201.$$

Το παραπάνω μπορεί να υπολογιστεί απευθείας μέσω της `R`, υπολογίζοντας τον δειγματικό μέσο των 1000 τιμών των εκτιμητών διασποράς ανά bootstrap δείγμα, οι οποίοι εκτιμητές έχουν αποθηκευτεί στο διάνυσμα `theta_values2` (βλ. κώδικα του Παραδείγματος 11.1):

```
> theta_star2 <- mean(theta_values2)
round(theta_star2, 3)
[1] 3.343
# bias estimate for variance estimator
> round(theta_star2 - theta_hat2, 3)
[1] -0.201
```

Στη συνέχεια, υπολογίζουμε τις εκτιμήσεις της μεροληψίας και του τυπικού σφάλματος του $\hat{\sigma}^2$. Λαμβάνοντας υπόψη ότι όταν έχουμε δειγματοληψία από κανονικό πληθυσμό

$$\hat{\sigma}^2 \sim \mathcal{G}\left(\frac{n-1}{2}, \frac{n}{2\sigma^2}\right),$$

προκύπτει από τις σχέσεις για τη μέση τιμή και τη διακύμανση της Γάμμα κατανομής ότι:

$$\text{bias}(\hat{\sigma}^2) = \text{E}(\hat{\sigma}^2) - \sigma^2 = \left(\frac{n-1}{n} - 1\right)\sigma^2 = -\frac{1}{n}\sigma^2,$$

και

$$\text{se}(\hat{\sigma}^2) = \sqrt{\frac{2(n-1)}{n^2}\sigma^4}.$$

Επομένως, η εκτίμηση της μεροληψίας και του τυπικού σφάλματος είναι:

$$\widehat{\text{bias}}(\hat{\sigma}^2) = -\frac{1}{n}\hat{\sigma}^2$$

και

$$\widehat{\text{se}}(\hat{\sigma}^2) = \sqrt{\frac{2(n-1)}{n^2}\hat{\sigma}^4},$$

αντίστοιχα. Τα αποτελέσματα συνοψίζονται στις δύο τελευταίες στήλες του Πίνακα 11.1, μαζί με τις αντίστοιχες εκτιμήσεις που προκύπτουν από το jackknife. □

Παρατήρηση 11.4. Τονίζουμε ότι η εκτίμηση της μεροληψίας και του τυπικού σφάλματος μέσω αναλυτικών εκφράσεων δεν είναι πάντοτε δυνατή, και είναι ακριβώς αυτές οι περιπτώσεις όπου αναδεικνύεται η πραγματική συνεισφορά των τεχνικών bootstrap. Στα προηγούμενα παραδείγματα, κάτι τέτοιο ήταν δυνατό γιατί κάναμε χρήση του γεγονότος ότι η κατανομή των δεδομένων ήταν η κανονική κατανομή, και παραθέσαμε τις αναλυτικές εκτιμήσεις για λόγους σύγκρισης με τις μεθόδους επαναδειγματοληψίας.

11.2 Bootstrap διαστήματα εμπιστοσύνης

Έστω ότι ο T_n, π.χ. $T_n = T(F_n)$, εκτιμά κάποια μονοδιάστατη ποσότητα θ και ας συμβολίσουμε με c_α: το (κάτω) α-ποσοστιαίο σημείο της κατανομής (έστω H) της $T_n - \theta$, για κάποιο $\alpha \in (0,1)$. Αν γνωρίζαμε την κατανομή αυτή (και κατ' επέκταση τα ποσοστιαία σημεία αυτής), τότε θα μπορούσαμε να επιλέξουμε σταθερές c_{α_1} και c_{α_2}, έτσι ώστε να περιέχουν την τυχαία μεταβλητή $T_n - \theta$ με κάποια προκαθορισμένη πιθανότητα, έστω α. Δηλαδή, αν γνωρίζαμε την κατανομή H, θα μπορούσαμε να προσδιορίσουμε σταθερές c_{α_1} και c_{α_2}, τέτοιες ώστε

$$P\left(c_{\alpha_1} \leqslant T_n - \theta \leqslant c_{\alpha_2}\right) = \alpha, \quad \forall\, \theta \in \Theta,$$

με $\alpha_1 > 0, \alpha_2 > 0$ και $\alpha_1 + \alpha_2 = \alpha$. Μία συνηθισμένη επιλογή είναι να θεωρήσουμε ίσες πιθανότητες αριστερά και δεξιά, δηλαδή να θεωρήσουμε ότι οι σταθερές c_{α_1} και c_{α_2}, επιπλέον, πληρούν τις σχέσεις:

$$P\left(T_n - \theta \leqslant c_{\alpha/2}\right) = \frac{\alpha}{2} = P\left(T_n - \theta \geqslant c_{1-\alpha/2}\right), \quad \forall\, \theta \in \Theta.$$

Οπότε,

$$T_n - \theta \leqslant c_{\alpha/2} \Leftrightarrow \theta \geqslant T_n - c_{\alpha/2}$$

και

$$T_n - \theta \geqslant c_{1-\alpha/2} \Leftrightarrow \theta \leqslant T_n - c_{1-\alpha/2}.$$

Άρα, το (τυχαίο) διάστημα

$$\left[T_n - c_{1-\alpha/2}, T_n - c_{\alpha/2}\right] = \left[T_n - H^{-1}(1-\alpha/2), T_n - H^{-1}(\alpha/2)\right] \tag{11.6}$$

είναι ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης (ίσων ουρών) για το θ. Δυστυχώς, η κατανομή του $T_n - \theta$ σπανιώς είναι γνωστή (μην ξεχνάτε ότι δεν κάνουμε καμία υπόθεση για την κατανομή των αρχικών δεδομένων) και όλες οι μέθοδοι που θα περιγράψουμε στη συνέχεια έχουν σκοπό να προσεγγίσουν τα ποσοστημόρια της κατανομής του $T_n - \theta$.

Είναι σημαντικό να τονίσουμε ότι όλα τα bootstrap διαστήματα εμπιστοσύνης που θα περιγράψουμε στη συνέχεια είναι προσεγγιστικά, με την έννοια ότι ο συντελεστής εμπιστοσύνης τους είναι περίπου ίσος με το επιθυμητό επίπεδο. Λέμε ότι ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης $[\theta_\ell, \theta_u]$ για το θ έχει ακρίβεια πρώτης τάξης, αν $P([\theta_\ell, \theta_u] \ni \theta) = 1 - \alpha + O(n^{-1/2})$. Αντίστοιχα, το διάστημα έχει ακρίβεια δεύτερης τάξης, αν ισχύει ότι $P([\theta_\ell, \theta_u] \ni \theta) = 1 - \alpha + O(n^{-1})$.

11.2.1 Βασική μέθοδος

Η ιδέα είναι να επιστρατεύσουμε την εμπειρική bootstrap συνάρτηση κατανομής $H_B(u)$ για να εκτιμήσουμε την πραγματική κατανομή $H(u)$ της $T_n - \theta$ στην (11.6), μέσω των προσομοιωμένων τιμών $\{T^\star_{n,b} - T_n; b = 1, \ldots, B\}$. Έτσι λοιπόν, για να εκτιμηθεί η (άγνωστη) πιθανότητα

$$H(u) = P(T_n - \theta \leqslant u), \quad \text{για δοθέν} \quad u \in \mathbb{R},$$

μπορούμε να χρησιμοποιήσουμε τη Monte Carlo εκτίμηση αυτής με βάση το bootstrap δείγμα των B τιμών. Επομένως, μπορούμε να χρησιμοποιήσουμε την

$$H_B(u) = \frac{\#\{T^\star_{n,b} - T_n \leqslant u\}}{B} = \frac{1}{B}\sum_{b=1}^{B} I\{T^\star_{n,b} - T_n \leqslant u\}.$$

Παρατηρήστε ότι αυτό που μας ενδιαφέρει για τον υπολογισμό του διαστήματος (11.6) είναι η εκτίμηση των $1-\alpha/2$ και $\alpha/2$ ποσοστιαίων σημείων της H. Για τον σκοπό αυτόν, θα χρησιμοποιήσουμε τον εκτιμητή αντικατάστασης, όπως είδαμε στο Παράδειγμα 2.9 του Κεφαλαίου 2. Έστω, λοιπόν, $c_a = H^{-1}(a)$ το κάτω a-ποσοστιαίο σημείο της H, για κάποιο $a \in (0,1)$. Τότε

$$\hat{c}_a = H_B^{-1}(a) = \inf\{u : H_B(u) \geqslant a\}.$$

Συνεπώς, το $100(1-\alpha)\%$ βασικό bootstrap διάστημα εμπιστοσύνης για το θ είναι το:

$$\left[T_n - H_B^{-1}(1-\alpha/2), T_n - H_B^{-1}(\alpha/2)\right]. \tag{11.7}$$

Η αντιστοιχία μεταξύ των σχέσεων (11.7) και (11.6) είναι άμεση. Η μόνη διαφορά είναι η αντικατάσταση των (άγνωστων) ποσοστιαίων σημείων της H με τα δειγματικά ανάλογα στο boostrap δείγμα.

Το a-ποσοστιαίο σημείο των προσομοιωμένων τιμών $\{T^\star_{n,b} - T_n; b = 1, \ldots, B\}$, είναι ίσο με

$$\hat{c}_a = T^\star_{[n,ja]} - T_n, \quad a \in (0,1) \tag{11.8}$$

όπου με $T^{\star}_{[n,1]} \leqslant T^{\star}_{[n,2]} \leqslant \ldots \leqslant T^{\star}_{[n,B]}$ συμβολίζουμε το διατεταγμένο bootstrap δείγμα. Ο θετικός ακέραιος j_α είναι τέτοιος, ώστε:

$$j_\alpha \in \{1,2,\ldots,B\} : \frac{j_\alpha - 1}{B} < \alpha \leqslant \frac{j_\alpha}{B} \Leftrightarrow aB \leqslant j_\alpha < \alpha B + 1 \Leftrightarrow j_\alpha = \lfloor \alpha B + 0.5 \rfloor,$$

όπου το $\lfloor y \rfloor$ συμβολίζει το ακέραιο μέρος του y. Αντικαθιστώντας την (11.8) στην (11.7) προκύπτει ότι μία ισοδύναμη έκφραση για το $100(1-\alpha)\%$ βασικό bootstrap διάστημα εμπιστοσύνης είναι η ακόλουθη:

$$\left[2T_n - T^{\star}_{[n,\lfloor B(1-\alpha/2)+0.5 \rfloor]}, 2T_n - T^{\star}_{[n,\lfloor B\alpha/2+0.5 \rfloor]}\right]. \tag{11.9}$$

Παρατήρηση 11.5. Στην R μπορεί να χρησιμοποιηθει η εντολή `quantile(..., type = 1)` για τον υπολογισμό του επιθυμητού δειγματικού ποσοστιαίου σημείου. Επομένως, αρκεί να καλέσουμε την εντολή αυτή στο διάνυσμα των bootstrap τιμών $\{T^{\star}_{n,b} - T_n; b = 1,\ldots,B\}$.

Παράδειγμα 11.4 (συνέχεια Παραδείγματος 11.1)**.** Να υπολογιστεί ένα 95% βασικό bootstrap διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού με βάση τα δεδομένα του Παραδείγματος 11.1.

Λύση Παραδείγματος 11.4. Θα εφαρμόσουμε τη σχέση (11.9) για $\alpha = 0.05$ και $T_n = \bar{X}$. Εφόσον έχουμε $B = 1000$ bootstrap τιμές του δειγματικού μέσου, επιλέγουμε την $j_{0.025} = 25$η και $j_{0.975} = 975$η διατεταγμένη τιμή. Χρησιμοποιώντας την R, προκύπτει ότι οι τιμές αυτές είναι ίσες με $T^{\star}_{[n,25]} = 19.336$ και $T^{\star}_{[n,975]} = 20.967$, αντίστοιχα. Υπενθυμίζουμε ότι με βάση τον κώδικα του Παραδείγματος 11.1 αυτά τα δειγματικά ποσοστιαία σημεία μπορούν να υπολογιστούν μέσω της R ως εξής:

```
> quantile(theta_values1, probs = c(0.025, 0.975), type = 1)
 2.5%  97.5%
19.336 20.967
```

όπου το διάνυσμα `theta_values1` περιέχει τις προσομοιωμένες bootstrap τιμές του δειγματικού μέσου. Αντικαθιστώντας στη σχέση (11.9), όπου $T_n = \bar{x} = 20.138$, λαμβάνουμε, τελικά, ότι το 95% βασικό bootstrap διάστημα εμπιστοσύνης για τη μέση τιμή είναι το:

$$[2 \cdot 20.138 - 20.967, 2 \cdot 20.138 - 19.336] = [19.309, 20.940].$$

□

Παράδειγμα 11.5 (συνέχεια Παραδείγματος 11.1)**.** Να υπολογιστεί ένα 95% βασικό bootstrap διάστημα εμπιστοσύνης για τη διασπορά του πληθυσμού με βάση τα δεδομένα του Παραδείγματος 11.1.

Λύση Παραδείγματος 11.5. Θα εφαρμόσουμε τη σχέση (11.9) για $\alpha = 0.05$ και $T_n = \hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$. Εφόσον έχουμε $B = 1000$ bootstrap τιμές του $\hat{\sigma}^2$, επιλέγουμε $j_{0.025} = 25$η και $j_{0.975} = 975$η διατεταγμένη τιμή. Χρησιμοποιώντας την R, προκύπτει ότι οι τιμές αυτές είναι ίσες με $T^{\star}_{[n,25]} = 1.640$ και $T^{\star}_{[n,975]} = 5.181$, αντίστοιχα. Υπενθυμίζουμε ότι με βάση τον κώδικα του Παραδείγματος 11.1 αυτά τα δειγματικά ποσοστιαία σημεία μπορούν να υπολογιστούν μέσω της R ως εξής:

```
> quantile(theta_values2,probs = c(0.025, 0.975), type = 1)
    2.5%    97.5%
1.640 5.181
```

όπου το διάνυσμα `theta_values2` περιέχει τις προσομοιωμένες bootstrap τιμές του $\hat{\sigma}^2$. Αντικαθιστώντας στη σχέση (11.9), όπου $T_n = \hat{\sigma}^2 = 3.544$, λαμβάνουμε, τελικά, ότι το 95% βασικό bootstrap διάστημα εμπιστοσύνης για τη διασπορά είναι το:

$$[2 \cdot 3.544 - 5.181, 2 \cdot 3.544 - 1.640] = [1.907, 5.448].$$

□

11.2.2 Κανονική μέθοδος

Η επόμενη μέθοδος κατασκευής bootstrap διαστημάτων εμπιστοσύνης βασίζεται στην υπόθεση ότι η κατανομή H προσεγγίζεται ικανοποιητικά από την κανονική, δηλαδή θεωρούμε ότι $T_n - \theta \sim \mathcal{N}(0, v^2), \forall\, \theta$. Να σημειώσουμε, εδώ, ότι μία τέτοια υπόθεση δεν αντιβαίνει με τις υποθέσεις μιας μη παραμετρικής ανάλυσης, καθώς δεν υποθέτουμε κάτι για την πληθυσμιακή κατανομή των δεδομένων, αλλά για τη δειγματική κατανομή του εκτιμητή. Για παράδειγμα, γνωρίζουμε ότι η συγκεκριμένη υπόθεση δικαιολογείται, τουλάχιστον ασυμπτωτικά, για τους εκτιμητές μέγιστης πιθανοφάνειας, υπό κάποιες συνθήκες ομαλότητας.

Αν η διασπορά v^2 ήταν γνωστή, τα άκρα του διαστήματος εμπιστοσύνης στη σχέση (11.6) θα ήταν

$$\left[T_n - z_{\alpha/2} v, T_n + z_{\alpha/2} v\right]$$

όπου $z_{\alpha/2}$ το άνω-$\alpha/2$ ποσοστιαίο σημείο της $\mathcal{N}(0,1)$. Στην πράξη η v είναι άγνωστη και εκτιμάται από την bootstrap εκτίμηση του τυπικού σφάλματος (11.5). Οπότε το

$$\left[T_n - z_{\alpha/2} \mathrm{se}_{\mathrm{boot}}, T_n + z_{\alpha/2} \mathrm{se}_{\mathrm{boot}}\right]$$

είναι ένα $100(1-\alpha)\%$ bootstrap διάστημα εμπιστοσύνης για το θ βασισμένο στην Κανονική προσέγγιση.

Ένα μειονέκτημα του προηγούμενου διαστήματος εμπιστοσύνης είναι ότι βασίζεται στην υπόθεση ότι ο εκτιμητής T_n είναι αμερόληπτος[2] για το θ. Στην Ενότητα 11.1.2 είδαμε ότι η (όποια) μεροληψία του T_n εκτιμάται μέσω της σχέσης (11.2) και ένας βελτιωμένος bootstrap εκτιμητής με διόρθωση μεροληψίας δίνεται από τον T_{boot} στη σχέση (11.3). Ακολουθώντας παρόμοιους συλλογισμούς με τον T_{boot} στη θέση του T_n, καταλήγουμε ότι το τυχαίο διάστημα

$$\left[2T_n - T_n^\star - z_{\alpha/2} \mathrm{se}_{\mathrm{boot}}, 2T_n - T_n^\star + z_{\alpha/2} \mathrm{se}_{\mathrm{boot}}\right] \tag{11.10}$$

είναι ένα $100(1-\alpha)\%$ bootstrap διάστημα εμπιστοσύνης για το θ βασισμένο στην Κανονική προσέγγιση, με διόρθωση μεροληψίας.

Παράδειγμα 11.6 (συνέχεια Παραδείγματος 11.1)**.** Να υπολογιστεί ένα 95% κανονικό bootstrap διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού με βάση τα δεδομένα του Παραδείγματος 11.1.

Λύση Παραδείγματος 11.6. Έχουμε ότι $\theta = \mu$, όπου $\mu \in \mathbb{R}$ η πληθυσμιακή μέση τιμή. Για την εκτίμηση του θ χρησιμοποιούμε τον εκτιμητή $T_n = \bar{X}$ και έχουμε ήδη προσομοιώσει ένα bootstrap δείγμα 1000 τιμών στο Παράδειγμα 11.1. Με βάση αυτό, στο Παράδειγμα 11.2 έχει εκτιμηθεί τόσο η μεροληψία του εκτιμητή όσο και το τυπικό του σφάλμα. Οπότε μπορούμε να αντικαταστήσουμε όλες τις απαραίτητες ποσότητες στην (11.10), για $\alpha = 0.05$, και να λάβουμε ότι το

$$2T_n - T_n^\star \pm z_{0.025} \mathrm{se}_{\mathrm{boot}} = 2 \cdot 20.138 - 20.132 \pm 1.96 \cdot 0.426 = [19.309, 20.979]$$

είναι ένα 95% bootstrap διάστημα εμπιστοσύνης για τη μέση τιμή. □

Παράδειγμα 11.7 (συνέχεια Παραδείγματος 11.1)**.** Να υπολογιστεί ένα 95% κανονικό bootstrap διάστημα εμπιστοσύνης για τη διασπορά του πληθυσμού με βάση τα δεδομένα του Παραδείγματος 11.1.

Λύση Παραδείγματος 11.7. Έχουμε ότι $\theta = \sigma^2$, όπου $\sigma^2 \in (0, \infty)$ η πληθυσμιακή διασπορά. Για την εκτίμηση του θ χρησιμοποιούμε τον εκτιμητή $T_n = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ και έχουμε ήδη προσομοιώσει ένα bootstrap δείγμα 1000 τιμών στο Παράδειγμα 11.1. Με βάση αυτό, στο Παράδειγμα 11.3 έχει εκτιμηθεί τόσο η μεροληψία του εκτιμητή όσο και το τυπικό του σφάλμα. Επομένως, μπορούμε να αντικαταστήσουμε όλες τις απαραίτητες ποσότητες στη σχέση (11.10) για $\alpha = 0.05$ και να λάβουμε ότι το

$$2T_n - T_n^\star \pm z_{0.025} \mathrm{se}_{\mathrm{boot}} = 2 \cdot 3.544 - 3.343 \pm 1.96 \cdot 0.925 = [1.932, 5.558]$$

είναι ένα 95% bootstrap διάστημα εμπιστοσύνης για τη διασπορά. □

[2] Αφήνεται ως άσκηση να αιτιολογήσετε πώς προκύπτει αυτός ο ισχυρισμός.

Παρατήρηση 11.6. Στο Παράδειγμα 11.1 κατασκευάστηκαν τα ιστογράμματα των bootstrap δειγμάτων των δύο εκτιμητών. Τα διαστήματα εμπιστοσύνης που υπολογίστηκαν στα Παραδείγματα 11.6 και 11.7 βασίζονται στην υπόθεση ότι η κατανομή του εκτιμητή T_n προσεγγίζεται ικανοποιητικά από μια κανονική κατανομή. Αφήνεται ως άσκηση στον/στην αναγνώστη/στρια ο σχολιασμός της εγκυρότητας αυτού του ισχυρισμού με βάση τα ιστογράμματα του Σχήματος 11.1.

Παρατήρηση 11.7. Στην περίπτωση που η υπόθεση της κανονικότητας της κατανομής του εκτιμητή φαίνεται να μην ισχύει, είναι συνηθισμένη η αναζήτηση κατάλληλου μετασχηματισμού του εκτιμητή $g(T_n)$, έτσι ώστε η κατανομή της τυχαίας μεταβλητής $g(T_n) - g(\theta)$ να προσεγγίζεται από κάποια κανονική κατανομή. Στη συνέχεια, υπολογίζεται ένα κανονικό διάστημα εμπιστοσύνης για τη μετασχηματισμένη παράμετρο και, τελικά, αναφέρεται ένα διάστημα εμπιστοσύνης στην αρχική κλίμακα εφαρμόζοντας τον αντίστροφο μετασχηματισμό.

11.2.3 Μέθοδος ποσοστιαίων σημείων

Το $100(1-\alpha)\%$ bootstrap διάστημα εμπιστοσύνης ποσοστιαίων σημείων (Bootstrap percentile CI) ορίζεται ως

$$\left[G_n^{\star -1}(\alpha/2), G_n^{\star -1}(1-\alpha/2)\right] = \left[T^{\star}_{[n,\lfloor B\alpha/2+0.5\rfloor]}, T^{\star}_{[n,\lfloor B(1-\alpha/2)+0.5\rfloor]}\right] \tag{11.11}$$

όπου με $G_n^{\star}$ συμβολίζουμε την εμπειρική κατανομή του bootstrap δείγματος. Υπολογιστικά, πρόκειται για την πιο άμεση τεχνική, καθώς αρκεί να υπολογιστούν τα $\alpha/2$ και $1-\alpha/2$ ποσοστιαία σημεία του bootstrap δείγματος $\{T^{\star}_{n,1}, \ldots, T^{\star}_{n,B}\}$.

Η τεχνική αυτή βασίζεται στην υπόθεση ότι υπάρχει ένας μονότονος μετασχηματισμός g τέτοιος, ώστε

$$U := g(T_n) \sim \mathcal{N}\left(g(\theta), c^2\right), \tag{11.12}$$

όπως περιγράφτηκε στην Παρατήρηση 11.7. Ωστόσο, η γνώση του μετασχηματισμού αυτού δεν παίζει ρόλο στην εφαρμογή της τεχνικής - αρκεί μόνο η υπόθεση ύπαρξης αυτού! Ο/η αναγνώστης/στρια παραπέμπεται στην Ενότητα 3.4 του Wasserman (2006) για περισσότερες λεπτομέρειες.

Παράδειγμα 11.8 (συνέχεια Παραδείγματος 11.1)**.** Να υπολογιστεί ένα 95% bootstrap διάστημα εμπιστοσύνης με τη μέθοδο των ποσοστιαίων σημείων για τη μέση τιμή του πληθυσμού με βάση τα δεδομένα του Παραδείγματος 11.1.

Λύση Παραδείγματος 11.8. Εφόσον έχουμε ότι $\theta = \mu$, όπου $\mu \in \mathbb{R}$ η (άγνωστη) πληθυσμιακή μέση τιμή, θα χρησιμοποιηθεί το bootstrap δείγμα των δειγματικών μέσων του Παραδείγματος 11.2. Το 0.025 δειγματικό ποσοστιαίο σημείο του bootstrap δείγματος αντιστοιχεί στην 25η διατεταγμένη παρατήρηση, δηλαδή: $T^{\star}_{[n,25]} = 19.336$. Το 0.975 δειγματικό ποσοστιαίο σημείο του bootstrap δείγματος αντιστοιχεί στην 975η διατεταγμένη παρατήρηση, δηλαδή: $T^{\star}_{[n,975]} = 20.967$. Αντικαθιστώντας αυτές τις τιμές στην (11.11), λαμβάνουμε ότι το

$$[T^{\star}_{[n,25]}, T^{\star}_{[n,975]}] = [19.336, 20.967]$$

είναι ένα 95% bootstrap διάστημα εμπιστοσύνης για τη μέση τιμή, με βάση τη μέθοδο των ποσοστιαίων σημείων. □

Παράδειγμα 11.9 (συνέχεια Παραδείγματος 11.1)**.** Να υπολογιστεί ένα 95% κανονικό bootstrap διάστημα εμπιστοσύνης με τη μέθοδο των ποσοστιαίων σημείων για τη διασπορά του πληθυσμού με βάση τα δεδομένα του Παραδείγματος 11.1.

Λύση Παραδείγματος 11.9. Εφόσον έχουμε ότι $\theta = \sigma^2$, όπου $\sigma^2 > 0$ η (άγνωστη) πληθυσμιακή διασπορά, θα χρησιμοποιηθεί το bootstrap δείγμα των $\hat{\sigma}^2$ του Παραδείγματος 11.2. Το 0.025 δειγματικό ποσοστιαίο σημείο του bootstrap δείγματος αντιστοιχεί στην 25η διατεταγμένη παρατήρηση, δηλαδή: $T^{\star}_{[n,25]} = 1.640$. Το

0.975 δειγματικό ποσοστιαίο σημείο του bootstrap δείγματος αντιστοιχεί στην 975η διατεταγμένη παρατήρηση, δηλαδή: $T_{[}n,975]^{\star} = 5.181$. Αντικαθιστώντας αυτές τις τιμές στην (11.11), λαμβάνουμε ότι το

$$[T^{\star}_{[n,25]}, T^{\star}_{[n,975]}] = [1.640, 5.181]$$

είναι ένα 95% bootstrap διάστημα εμπιστοσύνης για τη διασπορά, με βάση τη μέθοδο των ποσοστιαίων σημείων. □

11.2.4 Βελτιωμένες μέθοδοι

Τα διαστήματα που κατασκευάσαμε στις προηγούμενες ενότητες (βασικά, κανονικά και ποσοστιαίων σημείων) έχουν όλα ακρίβεια πρώτης τάξης. Δύο τεχνικές που οδηγούν σε βελτιωμένη ακρίβεια είναι τα διαστήματα studentized bootstrap και τα adjusted percentile (Bias Corrected – BC_{α}), τα οποία έχουν ακρίβεια δεύτερης τάξης (Davison and Hinkley, 1997; Hall, 2013).

Τα studentized bootstrap διαστήματα εμπιστοσύνης βελτιώνουν τα κανονικά bootstrap διαστήματα εμπιστοσύνης και είναι της μορφής:

$$[T_n + \zeta_{\alpha/2}\text{se}_{\text{boot}}, T_n + \zeta_{1-\alpha/2}\text{se}_{\text{boot}}]$$

όπου με ζ_{α} συμβολίζουμε το κάτω α-ποσοστιαίο σημείο της κατανομής των

$$\frac{T^{\star}_{n,b} - T_n}{\text{se}(T^{\star}_{n,b})}.$$

Το πλεονέκτημα των διαστημάτων αυτών, όπως τονίσαμε ήδη, είναι ότι έχουν ακρίβεια δεύτερης τάξης. Από την άλλη, το studentized bootstrap διάστημα έχει νόημα κυρίως σε περιπτώσεις παραμέτρων θέσης και, συνήθως, δεν προτιμάται για παραμέτρους κλίμακας. Υπολογιστικά, υπάρχει το μειονέκτημα ότι χρειάζεται εκτίμηση του $\text{se}(T^{\star}_{n,b})$, η οποία εκτίμηση μπορεί να γίνει εφαρμόζοντας τεχνικές επαναδειγματοληψίας (jackknife, bootstrap) για κάθε bootstrap επανάληψη $b = 1, \ldots, B$.

Παράδειγμα 11.10 (συνέχεια Παραδείγματος 11.1)**.** Να υπολογιστούν 95% studentized bootstrap διαστήματα εμπιστοσύνης για τη μέση τιμή και τη διασπορά των δεδομένων του Παραδείγματος 11.1.

Λύση Παραδείγματος 11.10. Ο κώδικας `R`, που ακολουθεί, εφαρμόζει bootstrap 50 επαναλήψεων εντός του κυρίως bootstrap των 1000 επαναλήψεων, ώστε να υπολογιστούν τα ζητούμενα studentized bootstrap διαστήματα εμπιστοσύνης.

```
> set.seed(1)
> B2 <- 50
> theta_hat1 <- mean(x)
> theta_hat2 <- (n-1)*var(x)/n
> ksi1 <- theta_values1 <- numeric(B)
> ksi2 <- theta_values2 <- numeric(B)
> t1 <- t2 <- numeric(B2)
> for (i in 1:B){
+ y <- x[sample(n, n, replace = TRUE)]
+ theta_values1[i] <- mean(y)
+ theta_values2[i] <- (n-1)*var(y)/n
+ for(j in 1:B2){
+ bootstar <- sample(y, replace=T)
+ t1[j] <- mean(bootstar)
+ t2[j] <- (n-1)*var(bootstar)/n
+ }
```

```
+ ksi1[i]<- (theta_values1[i] - theta_hat1)/sqrt(var(t1))
+ ksi2[i]<- (theta_values2[i] - theta_hat2)/sqrt(var(t2))
+ }
> studentizedCIMeanB <- theta_hat1 + quantile(ksi1,probs=c(0.025,
   0.975))*sqrt(var(theta_values1))
> studentizedCIVarB <- theta_hat2 + quantile(ksi2,probs=c(0.025,
   0.975))*sqrt(var(theta_values2))
> studentizedCIMeanB
    2.5%    97.5%
19.06641 21.05746
> studentizedCIVarB
      2.5%      97.5%
-0.2874693  4.9498293
```

Το αποτέλεσμα φαίνεται στις δύο τελευταίες γραμμές. Παρατηρήστε ότι το διάστημα εμπιστοσύνης για τη διασπορά παραβιάζει τα όρια του παραμετρικού χώρου. □

Τέλος, αναφέρουμε τα διαστήμα εμπιστοσύνης adjusted percentile (Bias Corrected - BC_α), τα οποία βελτιώνουν τα bootstrap διαστήματα ποσοστιαίων σημείων λαμβάνοντας υπόψη διορθώσεις μεροληψίας και ασυμμετρίας. Για περισσότερες πληροφορίες παραπέμπουμε στον Efron (1987). Τα διαστήματα αυτά (καθώς και όλα τα υπόλοιπα) μπορούν να υπολογιστούν και μέσω του πακέτου boot στην R.

Παρατήρηση 11.8. Τα bootstrap διαστήματα εμπιστοσύνης, που υπολογίστηκαν στα Παραδείγματα 11.2 ως 11.10, μαζί με τα αντίστοιχα jackknife και ακριβή διαστήματα εμπιστοσύνης (που εδώ είναι δυνατόν να υπολογιστούν κάνοντας χρήση της παραμετρικής υπόθεσης ότι η πληθυσμιακή κατανομή είναι γνωστή και, συγκεκριμένα, η κανονική) συνοψίζονται στον Πίνακα 11.2. Υπενθυμίζεται ότι το $100(1-\alpha)\%$ ακριβές διάστημα εμπιστοσύνης ίσων ουρών για τη μέση τιμή είναι το:

$$\left[\bar{X} - t_{n-1;\alpha/2} S_n/\sqrt{n}, \bar{X} + t_{n-1;\alpha/2} S_n/\sqrt{n}\right]$$

ενώ για τη διασπορά είναι το:

$$\left[\frac{(n-1)S_n^2}{\chi^2_{n-1;\alpha/2}}, \frac{(n-1)S_n^2}{\chi^2_{n-1;1-\alpha/2}}\right].$$

Παρατήρηση 11.9. Για να πάρουμε μια εμπειρική ιδέα για την ακρίβεια των διαφορετικών διαστημάτων εμπιστοσύνης που υπολογίσαμε στα προηγούμενα παραδείγματα μπορούμε να κάνουμε μια μελέτη Monte Carlo. Προσομοιώνουμε δείγμα $(x_1, \ldots, x_n)$ από $\mathcal{N}(\mu, \sigma^2)$ και, κατόπιν, υπολογίζουμε τα προηγούμενα διαστήματα εμπιστοσύνης με συντελεστή εμπιστοσύνης 95%. Επαναλαμβάνουμε τη διαδικασία αυτή πολλές φορές (εδώ $M = 10000$), ώστε να λάβουμε τη Monte Carlo εκτίμηση των αριστερών και δεξιών πιθανοτήτων αστοχίας:

$$\widehat{P}(\theta_{\text{low}} > \theta) = \frac{\text{αριθμός φορών που } \theta_{\text{low}} > \theta}{M},$$

και

$$\widehat{P}(\theta_{\text{up}} < \theta) = \frac{\text{αριθμός φορών που } \theta_{\text{up}} < \theta}{M},$$

όπου τα θ_{low} και θ_{up} συμβολίζουν το άνω και κάτω όριο του διαστήματος εμπιστοσύνης. Ως αριστερή (αντίστοιχα δεξιά) πιθανότητα αστοχίας ορίζουμε την πιθανότητα η πραγματική τιμή της παραμέτρου θ να είναι μικρότερη (αντίστοιχα μεγαλύτερη) του κάτω (αντίστοιχα άνω) ορίου του διαστήματος εμπιστοσύνης. Τα αποτελέσματα παρατίθενται στους Πίνακες 11.3 και 11.4 για τρία διαφορετικά μεγέθη δείγματος. Παρατηρήστε ότι δεν υπάρχουν έντονες διαφορές στην περίπτωση της μέσης τιμής. Στην περίπτωση της διασποράς, τα βελτιωμένα διαστήματα bootstrap BCa είναι πιο κοντά στην ονομαστική τιμή (2.5%) σε σχέση με τα υπόλοιπα.

	μ_{low}	μ_{up}	σ^2_{low}	σ^2_{up}
Ακριβές	19.234	21.042	2.158	7.958
Jackknife	19.292	20.984	1.667	5.794
bootstrap (κανονικό)	19.309	20.979	1.932	5.557
bootstrap (βασικό)	19.309	20.940	1.907	5.448
bootstrap (ποσοστιαίων σημείων)	19.336	20.967	1.640	5.181
bootstrap (studentized)	19.066	21.057	-0.287	4.950
bootstrap (BCa)	19.355	21.013	2.116	6.731

Πίνακας 11.2: Συγκεντρωτικά 95% διαστήματα εμπιστοσύνης για τα δεδομένα του Παραδείγματος 11.1.

	$n = 50$		$n = 100$		$n = 200$	
	μ_{low}	μ_{up}	μ_{low}	μ_{up}	μ_{low}	μ_{up}
jackknife	0.032	0.029	0.027	0.028	0.027	0.025
bootstrap (κανονικό)	0.032	0.030	0.028	0.028	0.028	0.025
bootstrap (βασικό)	0.032	0.030	0.029	0.028	0.028	0.025
bootstrap (ποσοστιαίων σημείων)	0.033	0.030	0.028	0.028	0.028	0.025
bootstrap (BCa)	0.033	0.030	0.028	0.028	0.028	0.024
bootstrap (studentized)	0.025	0.022	0.023	0.023	0.024	0.022

Πίνακας 11.3: Monte Carlo εκτίμηση των αριστερών και δεξιών πιθανοτήτων αστοχίας των 95% ΔΕ ίσων ουρών για το $\theta = \mu$.

	$n = 50$		$n = 100$		$n = 200$	
	σ^2_{low}	σ^2_{up}	σ^2_{low}	σ^2_{up}	σ^2_{low}	σ^2_{up}
jackknife	0.009	0.078	0.008	0.057	0.014	0.044
bootstrap (κανονικό)	0.011	0.085	0.009	0.061	0.015	0.046
bootstrap (βασικό)	0.008	0.098	0.007	0.070	0.012	0.054
bootstrap (ποσοστιαίων σημείων)	0.007	0.095	0.007	0.064	0.014	0.048
bootstrap (BCa)	0.026	0.048	0.024	0.038	0.026	0.026
bootstrap (studentized)	0.000	0.127	0.001	0.092	0.003	0.069

Πίνακας 11.4: Monte Carlo εκτίμηση των αριστερών και δεξιών πιθανοτήτων αστοχίας των 95% ΔΕ ίσων ουρών για το $\theta = \sigma^2$.

Παρατήρηση 11.10. Το πακέτο `boot` στην `R` είναι αρκετά δημοφιλές για αυτόματη εφαρμογή bootstrap συμπερασματολογίας. Τα βασικά βήματα για τη χρήση του πακέτου είναι τα εξής. Αρχικά, πρέπει να οριστεί μία συνάρτηση η οποία επιστρέφει τη στατιστική συνάρτηση για την οποία επιθυμούμε να εφαρμόσουμε bootstrap. Η συνάρτηση αυτή δέχεται δύο ορίσματα: τα παρατηρηθέντα δεδομένα (`data`) και ένα σύνολο δεικτών (`indices`) μεγέθους n με επανάθεση από το $\{1,2,\dots,n\}$. Κατόπιν, χρησιμοποιούμε τη συνάρτηση `boot(...)` για να προσομοιωθεί το bootstrap δείγμα. Το αποτέλεσμα αυτής της εντολής μπορεί να χρησιμοποιηθεί μετά στη συνάρτηση `boot.ci(...)` για τον υπολογισμό bootstrap διαστημάτων εμπιστοσύνης. Ας θεωρήσουμε για παράδειγμα ότι το ενδιαφέρον επικεντρώνεται στην περίπτωση της $\hat{\sigma}^2$ στα δεδομένα του Παραδείγματος 11.1. Τότε έχουμε τον ακόλουθο κώδικα στην `R`.

```
> #using the boot package
> #1.Define a function that returns the statistic we want
> get_theta <- function(data, indices){
+ y <- data[indices]
+ theta <- (n-1)*var(y)/n
+ return(theta)
+ }
> #2.Use the boot function to get R bootstrap replicates of the
   statistic
> set.seed(12345)
> boot_out <- boot(x, R = 1000, statistic = get_theta)
> #3.Use the boot.ci function to get the confidence intervals
> boot.ci(boot_out)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_out)

Intervals :
Level      Normal              Basic
95%   ( 1.888,  5.613 )   ( 1.803,  5.542 )

Level     Percentile            BCa
95%   ( 1.546,  5.285 )   ( 2.116,  6.731 )
Calculations and Intervals on Original Scale
Some BCa intervals may be unstable
```

Παρατηρήστε ότι η εντολή `boot.ci(...)` δεν επιστρέφει αυτόματα τα studentized διαστήματα εμπιστοσύνης. Για να γίνει κάτι τέτοιο θα πρέπει επίσης να εισάγουμε τις εκτιμήσεις των διασπορών των bootstrap δειγμάτων. Αυτό γίνεται με χρήση του ορίσματος `var.t0` στο οποίο εισάγονται οι εκτιμήσεις των διασπορών. Για περισσότερες λεπτομέρειες εκτελέστε την εντολή `?boot.ci`.

11.3 Εφαρμογές

Στην ενότητα αυτή, θα παρουσιαστούν περαιτέρω εφαρμογές της μεθόδου bootstrap σε προβλήματα παλινδρόμησης, στον έλεγχο υποθέσεων και στην εκτίμηση της συνάρτησης πυκνότητας πιθανότητας.

11.3.1 Γραμμική παλινδρόμηση

Έστω το απλό γραμμικό μοντέλο

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1,\dots,n,$$

όπου $x_1, \ldots, x_n$ δεδομένες σταθερές, α και β είναι άγνωστες παράμετροι, τα σφάλματα ϵ_i, $i = 1, \ldots, n$, είναι τυχαίο δείγμα από κάποια κατανομή με μέση τιμή 0 και σταθερή (αλλά άγνωστη) διασπορά σ^2. Τυπικά, σε ένα παραμετρικό πλαίσιο υποθέτουμε περαιτέρω ότι $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Η υπόθεση αυτή μας επιτρέπει να κατασκευάζουμε διαστήματα εμπιστοσύνης για τις παραμέτρους του γραμμικού μοντέλου και να ελέγχουμε υποθέσεις που διατυπώνονται για αυτές. Υπάρχουν πολλές περιπτώσεις, όμως, όπου η υπόθεση κανονικότητας δεν ευσταθεί. Σε αυτές τις περιπτώσεις υπάρχουν δύο δυνατά σενάρια: αν το μέγεθος δείγματος είναι «μεγάλο», μπορεί να γίνει συμπερασματολογία, όπως στο κανονικό γραμμικό μοντέλο, διότι οι δειγματικές κατανομές των εκτιμητών θα προσεγγίζονται ικανοποιητικά από κανονικές κατανομές. Αν αυτό δεν ισχύει, θα πρέπει να καταφύγουμε σε μη παραμετρική στατιστική συμπερασματολογία και το bootstrap είναι το κατάλληλο εργαλείο για αυτό.

	dose	surv
1	117.50	44.00
2	117.50	55.00
3	235.00	16.00
4	235.00	13.00
5	470.00	4.00
6	470.00	1.96
7	470.00	6.12
8	705.00	0.50
9	705.00	0.32
10	940.00	0.11
11	940.00	0.01
12	940.00	0.02
13	1410.00	0.70
14	1410.00	0.01

Πίνακας 11.5: Ποσοστά επιβίωσης αρουραίων (στήλη surv) για διαφορετικά επίπεδα ραδιενέργειας (στήλη dose). Πηγή: Efron (1988).

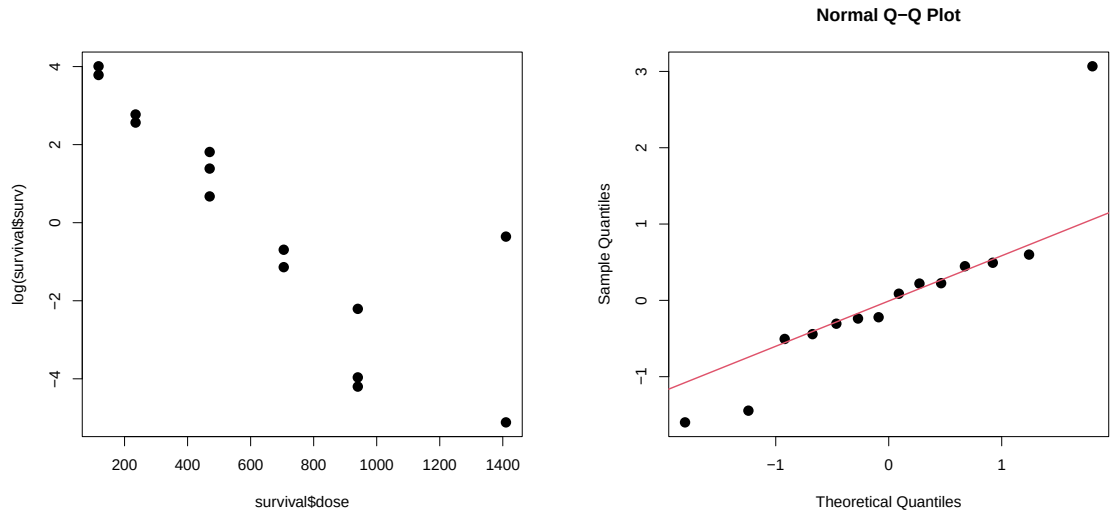
Ας θεωρήσουμε για παράδειγμα τα δεδομένα που παρατίθενται στον Πίνακα 11.5 και τα οποία προέρχονται από την εργασία του Efron (1988). Στα δεδομένα αυτά καταγράφονται τα ποσοστά επιβίωσης αρουραίων (στήλη surv) για διαφορετικά επίπεδα ραδιενέργειας (στήλη dose). Αυτό που ενδιαφέρει εδώ είναι η προσαρμογή ενός γραμμικού μοντέλου της μορφής:

$$\log(\text{surv}) = \alpha + \beta \cdot \text{dose},$$

όπου με $\log(\cdot)$ συμβολίζεται εδώ ο φυσικός λογάριθμος. Τα δεδομένα $(\log(\text{surv}), \text{dose})$, παριστάνονται στο διάγραμμα διασποράς στο Σχήμα 11.2.(α). Είναι προφανές από το qq-plot των studentized υπολοίπων του προσαρμοσμένου γραμμικού μοντέλου ότι η υπόθεση της κανονικότητας των σφαλμάτων δεν ικανοποιείται. Δεδομένου ότι το μέγεθος του δείγματος είναι μικρό ($n = 14$) και της αναντιστοιχίας μεταξύ των ποσοστημορίων στο qq-plot, θα πρέπει να είμαστε διστακτικοί στη χρήση των διαστημάτων εμπιστοσύνης του κανονικού γραμμικού μοντέλου.

Στη συνέχεια, θα περιγράψουμε πώς μπορούμε να κάνουμε συμπερασματολογία για τις παραμέτρους του μοντέλου χωρίς την υπόθεση της κανονικότητας των σφαλμάτων. Δύο προσεγγίσεις bootstrap για ένα μοντέλο παλινδρόμησης είναι οι εξής (Freedman, 1981):

1. δειγματοληψία με επανάθεση από τα ζεύγη $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$, και
2. δειγματοληψία με επανάθεση στα (εκτιμηθέντα) κατάλοιπα $e_i = Y_i - \widehat{\alpha} - \widehat{\beta} x_i$, $i = 1, \ldots, n$, όπου $\widehat{\alpha}$ και $\widehat{\beta}$ είναι οι εκτιμητές ελαχίστων τετραγώνων. Κατόπιν, θεωρούμε το δείγμα $(x_1, Y_1^\star), (x_2, Y_2^\star), \ldots, (x_n, Y_n^\star)$,



Σχήμα 11.2: Δεδομένα επιβίωσης ποντικιών σε ραδιενέργεια: (α) διάγραμμα διασποράς δεδομένων, (β): κανονικό qq-plot των studentized υπολοίπων του γραμμικού μοντέλου.

όπου

$$Y_i^\star = \widehat{\alpha} + \widehat{\beta} x_i + e_i^\star$$

και $e_i^\star$ δείγματα bootstrap από τα e_i, $i = 1, \ldots, n$.

Παρότι σε αρκετές περιπτώσεις (αλλά όχι πάντα) οι δύο μέθοδοι δίνουν παρόμοια αποτελέσματα, η 2η είναι πιο σωστή θεωρητικά[3]. Στη συνέχεια, θα περιγράψουμε τη δεύτερη τεχνική, δηλαδή θα λάβουμε bootstrap δείγματα των εκτιμηθέντων καταλοίπων.

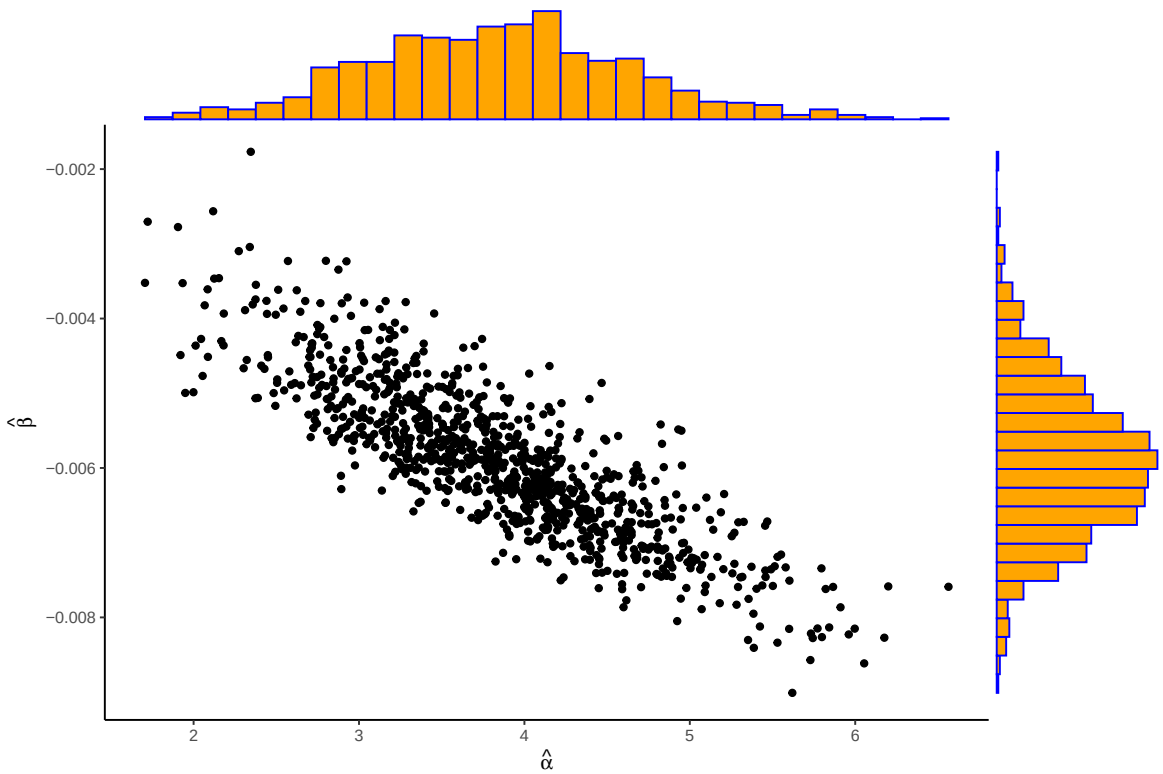
Παράδειγμα 11.11. Θεωρήστε τα δεδομένα του Πίνακα 11.5 και το γραμμικό μοντέλο $\log(\text{surv}) = \alpha + \beta \cdot \text{dose}$. Να υπολογιστούν 95% bootstrap διαστήματα εμπιστοσύνης για τις παραμέτρους α και β.

Λύση Παραδείγματος 11.11. Ο παρακάτω κώδικας `R` εφαρμόζει bootstrap 1000 τιμών στα κατάλοιπα του εκτιμηθέντος γραμμικού μοντέλου και, κατόπιν, υπολογίζει διαστήματα εμπιστοσύνης (κανονικά, βασικά, ποσοστιαίων σημείων και BCa) μέσω του πακέτου boot.

```
 > library(boot)
> data(survival)
> fit <- lm(log(survival$surv) ~ survival$dose)
> e <- fit$residuals
> a <- fit$coefficients[1]
> b <- fit$coefficients[2]
> surv.fun <- function(data, indices){
+ d <- data[indices]
+ y_boot <- a + b*survival$dose + d
+ d.reg <- lm(y_boot ~ survival$dose)
+ c(coef(d.reg))
+ }
> set.seed(1)
> surv.boot <- boot(e,surv.fun,R=1000)

# bootstrap CIs for alpha
> boot.ci(surv.boot, index = 1)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
```

[3] Αφήνεται ως άσκηση η αιτιολόγηση αυτού του ισχυρισμού.

```
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = surv.boot, index = 1)

Intervals :
Level      Normal              Basic
95%   ( 2.261,  5.347 )   ( 2.145,  5.301 )

Level     Percentile            BCa
95%   ( 2.346,  5.502 )   ( 2.379,  5.552 )

# bootstrap CIs for beta
> boot.ci(surv.boot, index = 2)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = surv.boot, index = 2)

Intervals :
Level      Normal              Basic
95%   (-0.0079, -0.0039 )   (-0.0080, -0.0041 )

Level     Percentile            BCa
95%   (-0.0078, -0.0038 )   (-0.0078, -0.0038 )
```



Σχήμα 11.3: Bootstrap τιμές των εκτιμητών ελαχίστων τετραγώνων $(\hat{\alpha}, \hat{\beta})$ και τα αντίστοιχα ιστογράμματα, για τα δεδομένα (dose, log(surv)) του Πίνακα 11.5.

Στις πρώτες δύο γραμμές του κώδικα φορτώνουμε την απαραίτητη βιβλιοθήκη και το σύνολο δεδομένων. Έπειτα (γραμμή 3 του κώδικα), προσαρμόζουμε το γραμμικό μοντέλο με εξαρτημένη μεταβλητή τον

λογάριθμο του χρόνου επιβίωσης και ανεξάρτητη μεταβλητή την ποσότητα που χορηγήθηκε ως δόση. Στη γραμμή 4 του κώδικα ζητούμε την αποθήκευση των εκτιμηθέντων καταλοίπων του γραμμικού μοντέλου, δηλαδή των τιμών $y_i - \hat{y}_i$, $i = 1, \ldots, n$. Οι 14 τιμές των καταλοίπων αποθηκεύονται στη μεταβλητή `e`. Οι εκτιμητές ελαχίστων τετραγώνων του σταθερού όρου του μοντέλου και της παραμέτρου της ανεξάρτητης μεταβλητής αποθηκεύονται στις μεταβλητές `a` και `b`, αντίστοιχα. Μπορούμε να διαπιστώσουμε ότι οι τιμές που αποθηκεύονται είναι $\hat{\alpha} = 3.8236$ και $\hat{\beta} = -0.0059$, αντίστοιχα.

Η συνάρτηση `surv.fun` (που ορίζεται στη γραμμή 7 του κώδικα) εφαρμόζει bootstrap στα κατάλοιπα και επιστρέφει σε ένα διάνυσμα δύο τιμών τις bootstrap τιμές των εκτιμήσεων ελαχίστων τετραγώνων. Είναι σημαντικό να τονίσουμε εδώ, ότι αυτή η συνάρτηση δέχεται ως όρισμα τα εκτιμηθέντα κατάλοιπα του προσαρμοσμένου γραμμικού μοντέλου (και όχι τα ίδια τα δεδομένα). Αυτό είναι σαφές από τον τρόπο που καλείται αυτή η συνάρτηση μέσω της εντολής `boot(e, surv.func, ...)`: το `e` είναι το διάνυσμα με τα εκτιμηθέντα κατάλοιπα. Τελικά, το bootstrap δείγμα (`surv.boot`) που παρήγαγε η εντολή boot εισάγεται στην εντολή `boot.ci` για τον υπολογισμό των διαστημάτων εμπιστοσύνης, τα οποία εκτυπώνονται στο τέλος του παραπάνω κώδικα. □

Να παρατηρήσουμε, εδώ, ότι τα άκρα όλων των διαστημάτων εμπιστοσύνης που υπολογίστηκαν προηγουμένως δεν περιέχουν το μηδέν. Αυτό σημαίνει ότι υπάρχει σημαντική (αρνητική) γραμμική επίδραση της επεξηγηματικής μεταβλητής (dose) στη μεταβλητή απόκρισης (survival). Είναι αναμενόμενο ότι η επιβίωση των αρουραίων μειώνεται καθώς αυξάνεται το επίπεδο ραδιενέργειας. Στο σημείο αυτό προτρέπουμε τον/την αναγνώστη/στρια να συνεχίσει τη μελέτη του/της επιλύοντας την Άσκηση 11.6 και την Άσκηση 11.7.

11.3.2 Έλεγχοι υποθέσεων

Έστω ότι θέλουμε να ελέγξουμε μία υπόθεση της μορφής

$$H_0 : \theta = \theta_0 \quad \text{έναντι της} \quad H_1 : \theta \neq \theta_0.$$

Στη διάθεσή μας έχουμε μία στατιστική συνάρτηση $T_n = T(X_1, \ldots, X_n)$ που μπορεί να χρησιμοποιηθεί για τον έλεγχο. Φυσικά, η κατανομή του πληθυσμού θεωρείται άγνωστη και, συνεπώς, η κατανομή της T_n υπό την ισχύ της μηδενικής υπόθεσης είναι επίσης άγνωστη.

Ας υποθέσουμε ότι η παρατηρηθείσα τιμή της T_n είναι t και ότι η κατανομή της είναι συμμετρική γύρω από το 0 υπό την ισχύ της μηδενικής υπόθεσης. Τότε, η p-value του ελέγχου δίνεται από τη σχέση:

$$\text{p} - \text{value} = P(|T_n| \geqslant |t||H_0), \tag{11.13}$$

όπου η πιθανότητα στο δεξί μέλος αναφέρεται στην κατανομή της T_n υπό την ισχύ της μηδενικής υπόθεσης. Συνεπώς, για να εκτιμήσουμε την p-value του ελέγχου μέσω bootstrap θα πρέπει να προσομοιώσουμε δείγματα με επανάθεση μεγέθους n όχι απευθείας από την F_n, αλλά από την F_n **υπό την ισχύ της** H_0. Έτσι, η γενική διαδικασία εκτίμησης της p-value μέσω bootstrap είναι αυτή που περιγράφεται στη συνέχεια

1. Προσομοιώνουμε bootstrap δείγματα από την F_n υπό την ισχύ της H_0

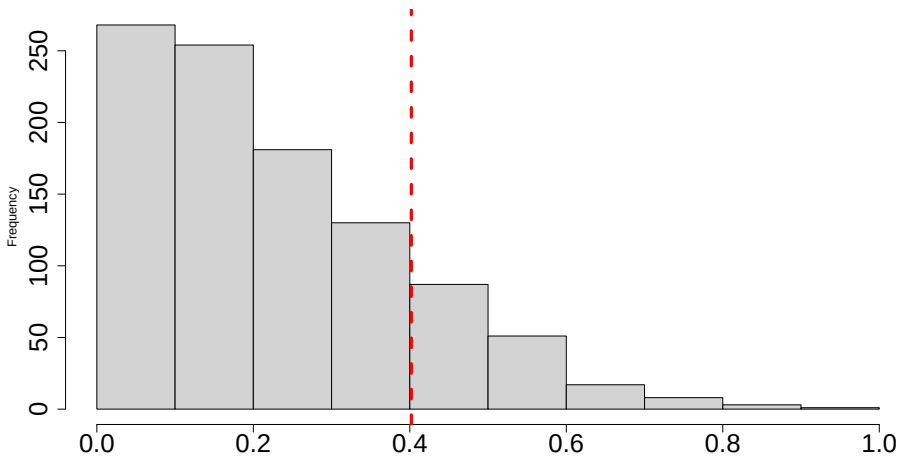
$$(\tilde{x}^{\star}_{b,1}, \tilde{x}^{\star}_{b,2}, \ldots, \tilde{x}^{\star}_{b,n}), \text{ για } b = 1, \ldots, B.$$

2. Κατόπιν, υπολογίζουμε την T_n για κάθε $b = 1, \ldots, B$:

$$T^{\star}_{n,b} := T_n(\tilde{x}^{\star}_{b,1}, \tilde{x}^{\star}_{b,2}, \ldots, \tilde{x}^{\star}_{b,n}).$$

3. Bootstrap εκτίμηση του p-value (για δίπλευρη εναλλακτική υπόθεση)

$$\text{p} - \text{value}_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} I\{|T^{\star}_{n,b}| \geqslant |t|\}.$$



Σχήμα 11.4: Ιστόγραμμα $B = 1000$ bootstrap τιμών της $T_n = \left|\overline{\tilde{X}}_n - 1\right|$ στα δεδομένα του Παραδείγματος 11.12. Η διακεκομμένη γραμμή αντιστοιχεί στην παρατηρηθείσα τιμή της $T_n = 0.402$.

Το κομβικό σημείο, για την εφαρμογή του παραπάνω, είναι η προσομοίωση bootstrap δειγμάτων από την εμπειρική συνάρτηση κατανομής υπό την ισχύ της H_0. Η γενική ιδέα είναι ότι τα αρχικά δεδομένα θα πρέπει να μετασχηματιστούν, έτσι ώστε η εμπειρική συνάρτηση κατανομής των μετασχηματισμένων δεδομένων να ικανοποιεί τη μηδενική υπόθεση. Ο τρόπος που κάτι τέτοιο μπορεί να επιτευχθεί εξαρτάται από την εκάστοτε περίπτωση, όπως θα γίνει σαφές μέσω των παραδειγμάτων που ακολουθούν.

Παράδειγμα 11.12 (bootstrap έλεγχος μέσου)**.** Έστω οι δέκα παρατηρήσεις

$$-0.89, -0.47, 0.05, 0.155, 0.279, 0.775, 1.0016, 1.23, 1.89, 1.96.$$

Να ελεγχθεί, στα συνήθη επίπεδα σημαντικότητας, η υπόθεση ότι ο μέσος του πληθυσμού είναι 1, έναντι της εναλλακτικής ότι διαφέρει.

Λύση Παραδείγματος 11.12. Πρόκειται για τον ελέγχο της $H_0 : \mu = 1$ έναντι της $H_1 : \mu \neq 1$. Ας υποθέσουμε ότι χρησιμοποιείται η στατιστική συνάρτηση $T_n = \bar{X} - 1$ για να μας δώσει στοιχεία κατά της μηδενικής υπόθεσης. Παρατηρήστε ότι «μεγάλες» τιμές της $|T_n|$ συνηγορούν κατά της H_0, οπότε η p-value θα δίνεται πράγματι από την έκφραση (11.13). Η παρατηρηθείσα τιμή της $|T_n|$ στο δείγμα ισούται με:

$$|t| = |\bar{x} - 1| = |0.598 - 1| = 0.402.$$

Παρατηρήστε ότι, ενώ η H_0 υποθέτει μέση τιμή 1, η παρατηρηθείσα τιμή του δειγματικού μέσου είναι $\bar{x} = 0.598$. Οπότε, αν πάρουμε bootstrap δείγματα από την F_n, δεν προσομοιώνουμε από μία κατανομή που «υπακούει» στην H_0. Για να συνεχίσουμε λοιπόν θα πρέπει να μετασχηματίσουμε τα δεδομένα ώστε να ικανοποιείται η υπόθεση ότι η μέση τιμή της εμπειρικής κατανομής είναι 1. Δεδομένου ότι ο δειγματικός μέσος ισούται με 0.598, αν σε κάθε παρατήρηση προσθέσουμε 0.402:

$$\tilde{X}_i = X_i + 0.402, \quad i = 1, \dots, n,$$

θα έχουμε ότι η εμπειρική κατανομή των μετασχηματισμένων δεδομένων $(\tilde{X}_1, \dots, \tilde{X}_n)$ έχει πράγματι μέση τιμή ίση με 1, δηλαδή αυτήν την τιμή που υποθέτει η μηδενική υπόθεση. Συνεπώς, η εμπειρική συνάρτηση κατανομής των $\tilde{X}_i$ μπορεί να χρησιμοποιηθεί για προσομοίωση υπό την H_0. Παρατηρήστε, επίσης, ότι, ενώ διορθώνουμε τη μέση τιμή, άλλα χαρακτηριστικά, όπως η διασπορά και η συμμετρία, παραμένουν ίδια. Ο κώδικας που ακολουθεί εφαρμόζει `bootstrap` $B = 1000$ τιμών για την εκτίμηση της p-value.

```
> x <- c(-0.89,-0.47,0.05,0.155,0.279,0.775,1.0016,1.23,1.89,1.96)
> x_tilde <- x + 1 - mean(x)
> n <- length(x)
> t0 <- abs(mean(x) - 1)
> B <- 1000
> theta_values <- numeric(B)
> set.seed(1)
> for (b in 1:B){
+ bootsample <- sample(x_tilde,replace=TRUE)
+ theta_values[b] <- abs(mean(bootsample) - 1)
+ }
> pvalue_boot <- sum(theta_values > t0)/B
> pvalue_boot
[1] 0.166
```

Στο Σχήμα 11.4 παρατίθεται το ιστόγραμμα των bootstrap τιμών. Παρατηρήστε ότι πρόκειται για ιδιαίτερα ασύμμετρη κατανομή. Η bootstrap εκτίμηση της p-value ισούται με τη σχετική συχνότητα των τιμών που ξεπερνούν την παρατηρηθείσα τιμή της Tn (που αντιστοιχεί στην κόκκινη διακεκομμένη γραμμή), δηλαδή

$$\text{p} - \text{value}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B} I\{|T^{\star}_{n,b}| \geqslant |t|\} = 0.166.$$

Το συμπέρασμα είναι ότι δεν έχουμε στοιχεία για να απορρίψουμε την H_0 στα συνήθη επίπεδα σημαντικότητας. □

Παρατήρηση 11.11. Η επιλογή της ελεγχοσυνάρτησης για τη διεξαγωγή ενός bootstrap ελέγχου παίζει σημαντικό ρόλο. Ιδανικά, θα πρέπει η κατανομή της ελεγχοσυνάρτησης (που μας είναι άγνωστη) να μην εξαρτάται από άγνωστες παραμέτρους. Μία ελεγχοσυνάρτηση που έχει την ιδιότητα ότι η κατανομή της, υπό την H_0, δεν εξαρτάται από άγνωστες παραμέτρους ονομάζεται ποσότητα οδηγός (Barnard, 1974), όπως αναφέρθηκε και στην Ενότητα 1.8.2. Για παράδειγμα, ας θεωρήσουμε την περίπτωση που ενδιαφερόμαστε για τον έλεγχο της μέση τιμής πληθυσμού $H_0 : \mu = \mu_0$. Υπό την H_0, η ασυμπτωτική κατανομή του $T_1 = \bar{X} - \mu_0$ είναι η $\mathcal{N}(0, \sigma^2/n)$, όπου το σ^2 είναι η πληθυσμιακή διασπορά. Αντίθετα, η ασυμπτωτική κατανομή της $T_2 = \sqrt{n}\frac{\bar{X}-\mu_0}{S_n}$ είναι η $\mathcal{N}(0,1)$. Επομένως, η T_2 είναι ποσότητα οδηγός (pivotal quantity), σε αντίθεση με την T_1. Γενικά, μέσω bootstrap δεν είναι εύκολο να εξακριβώσουμε αν η ελεγχοσυνάρτηση που χρησιμοποιούμε είναι πράγματι ποσότητα οδηγός. Στο σημείο αυτό, προτρέπουμε τον/την αναγνώστη/στρια να λύσει την Άσκηση 11.9.

Στη συνέχεια, θεωρούμε το πρόβλημα της σύγκρισης δύο πληθυσμιακών μέσων τιμών με ανεξάρτητα δείγματα. Έστω $\mathbf{X} = (X_1, \ldots, X_n)$ τυχαίο δείγμα από κατανομή με μέση τιμή μ_1 και $\mathbf{Y} = (Y_1, \ldots, Y_m)$ τυχαίο δείγμα από κατανομή με μέση τιμή μ_2. Περαιτέρω, υποθέτουμε ότι τα X_i, Y_j είναι ανεξάρτητα για κάθε i, j. Στο παραπάνω πλαίσιο μας ενδιαφέρει ο έλεγχος της υπόθεσης:

$$H_0 : \mu_1 = \mu_2 \quad \text{έναντι της} \quad H_1 : \mu_1 > \mu_2.$$

Σημειώστε ότι στην ειδική περίπτωση όπου οι δύο πληθυσμοί θεωρούνται κανονικοί με κοινή διασπορά, τότε ο έλεγχος αυτής της υπόθεσης γίνεται μέσω του t-test για δύο πληθυσμούς. Το μη παραμετρικό ανάλογο, όπως είδαμε στο Κεφάλαιο 6, είναι ο έλεγχος Wilcoxon-Mann-Whitney, υπό την υπόθεση ότι οι κατανομές που περιγράφουν τους δύο πληθυσμούς προκύπτουν από μετατόπιση θέσης.

Εν συνεχεία, θα εξετάσουμε το πρόβλημα του ελέγχου της ισότητας των μέσων χωρίς να υποθέσουμε κανονικότητα (όπως στο t test) ή ότι οι συναρτήσεις κατανομών των δύο πληθυσμών ταυτίζονται υπό τη

μηδενική υπόθεση (όπως στον έλεγχο Wilcoxon-Mann-Whitney). Μία στατιστική συνάρτηση που μπορεί να επιστρατευθεί για τον παραπάνω έλεγχο είναι η:

$$T(\boldsymbol{X},\boldsymbol{Y}) = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}, \tag{11.14}$$

όπου $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\bar{Y}_m = \frac{1}{m}\sum_{i=1}^{m} Y_i$, $S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ και $S_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2$. Στην περίπτωση όπου τα δείγματα είναι κανονικά και υπό την υπόθεση της ανεξαρτησίας, η κατανομή της $T(\boldsymbol{X},\boldsymbol{Y})$ υπό την H_0 προσεγγίζεται[4] από μία κατανομή t (Welch, 1938, 1947). Η p-value για την παραπάνω μονόπλευρη εναλλακτική υπόθεση ισούται με $P(T > t|H_0)$, όπου $t = T(\boldsymbol{x},\boldsymbol{y})$ η παρατηρηθείσα τιμή της T.

Για να εφαρμόσουμε bootstrap στην (11.14), θα πρέπει να μετασχηματίσουμε τα δεδομένα, έτσι ώστε να ικανοποιείται η H_0, δηλαδή να ισχύει ότι οι πληθυσμοί έχουν την ίδια μέση τιμή. Για τον λόγο αυτόν, θεωρούμε τα μετασχηματισμένα δεδομένα

$$\tilde{X}_i = X_i - \bar{X}_n + \bar{Z}, \quad i = 1,\dots,n,$$

και

$$\tilde{Y}_j = Y_j - \bar{Y}_m + \bar{Z}, \quad j = 1,\dots,m,$$

όπου

$$\bar{Z} = \frac{1}{n+m}\left(\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j\right).$$

Τότε, εύκολα, προκύπτει ότι:

$$\mathrm{E}(\tilde{X}_i) = \mathrm{E}(\tilde{Y}_j) = \frac{n}{n+m}\mu_1 + \frac{m}{n+m}\mu_2,$$

για κάθε $i = 1,\dots,n$ και $j = 1,\dots,m$, που σημαίνει ότι τα μετασχηματισμένα δεδομένα $\tilde{X}_i$ και $\tilde{Y}_j$ προέρχονται πράγματι από μία κατανομή με κοινή μέση τιμή, όπως υπαγορεύει η μηδενική υπόθεση. Συνεπώς, ο αλγόριθμος bootstrap για τον έλεγχο ισότητας μέσων τιμών δύο πληθυσμών με ανεξάρτητα δείγματα έχει ως εξής:

(1) Θέσε $\tilde{x}_i = x_i - \bar{x} + \bar{z}$ και $\tilde{y}_j = y_j - \bar{y}_m + \bar{z}$, για $i = 1,\dots,n$, $j = 1,\dots,m$.

(2) Για $b = 1,\dots,B$

(2.1) Πάρε δείγμα

$$\tilde{\boldsymbol{x}}_b^{\star} = (\tilde{x}_{b,1}^{\star},\dots,\tilde{x}_{b,n}^{\star})$$

με επανάθεση από τα $(\tilde{x}_1,\dots,\tilde{x}_n)$.

(2.2) Πάρε δείγμα

$$\tilde{\boldsymbol{y}}_b^{\star} = (\tilde{y}_{b,1}^{\star},\dots,\tilde{y}_{b,m}^{\star})$$

με επανάθεση από τα $(\tilde{y}_1,\dots,\tilde{y}_m)$.

(2.3) Υπολόγισε την τιμή της ελεγχοσυνάρτησης Welch, η οποία δίνεται στη σχέση (11.14). Έστω

$$T_{n,b}^{\star} := T(\tilde{\boldsymbol{x}}_b^{\star},\tilde{\boldsymbol{y}}_b^{\star}).$$

(3) Υπολόγισε την bootstrap εκτίμηση της p-value (για μονόπλευρη εναλλακτική $\mu_1 > \mu_2$)

$$\text{p} - \text{value}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B} I\{T_{n,b}^{\star} \geqslant t\},$$

όπου με $t = T(\boldsymbol{x},\boldsymbol{y})$ συμβολίζουμε την πραγματοποίηση στο δείγμα της στατιστικής συνάρτησης της σχέσης (11.14).

[4] Το πρόβλημα εύρεσης της ακριβούς κατανομής της (11.14) είναι το πρόβλημα Behrens-Fisher (Fisher, 1935).

Παράδειγμα 11.13 (Σύγκριση μέσων τιμών δύο ανεξάρτητων πληθυσμών). Δεκαέξι ποντίκια συμμετείχαν σε ένα πείραμα, όπου σε επτά από αυτά δόθηκε ένα καινούριο φάρμακο και στα υπόλοιπα δόθηκε ένα υπάρχον φάρμακο. Κατόπιν, καταγράφηκαν για κάθε ποντίκι οι μέρες επιβίωσης.

Φάρμακο	Μέρες Επιβίωσης
Νέο	94,197,16,38,99,141,23
Παλιό	52,104,146,10,51,30,40,27,46

Εξετάστε αν το νέο φάρμακο επιφέρει αύξηση στον μέσο αριθμό ημερών επιβίωσης.

Λύση Παραδείγματος 11.13. Πρόκειται για έλεγχο ισότητας δύο μέσων τιμών με ανεξάρτητα δείγματα. Ειδικότερα, θέλουμε να ελέγξουμε την

$$H_0 : \mu_1 = \mu_2 \quad \text{έναντι της} \quad H_1 : \mu_1 > \mu_2,$$

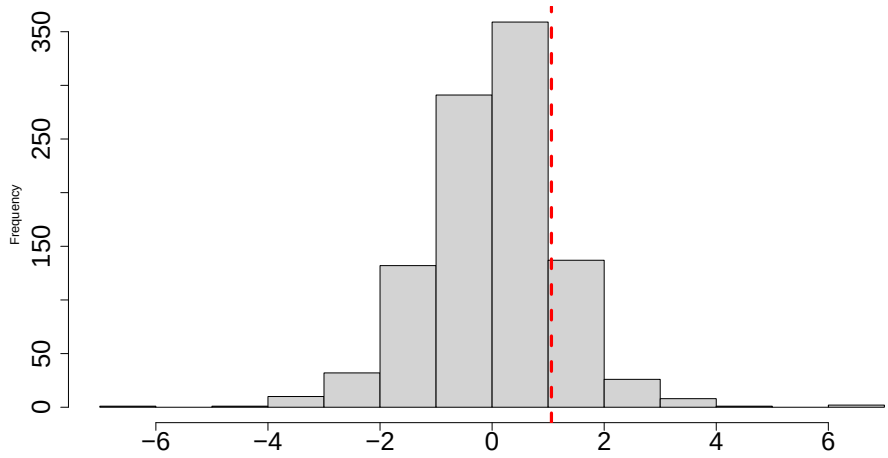
όπου μ_1, μ_2 η μέση τιμή της επιβίωσης για το νέο και παλιό φάρμακο, αντίστοιχα. Έχουμε $n = 7$, $m = 9$ το πλήθος παρατηρήσεις από τους δύο πληθυσμούς, με $\bar{x}_n = 86.86$, $\bar{y}_m = 56.22$, $s_1^2 = 4457.81$ και $s_2^2 = 1804.194$. Συνεπώς, η παρατηρηθείσα τιμή της (11.14) ισούται με

$$t = T(\boldsymbol{x}, \boldsymbol{y}) = \frac{86.86 - 56.22}{\sqrt{4457.81/7 + 1804.194/9}} = 1.058.$$

Ο παρακάτω κώδικας λαμβάνει το bootstrap δείγμα και, κατόπιν, υπολογίζει την εκτίμηση της p-τιμής.

```
> x <- c( 94,  197,   16,   38,  99,  141,  23)
> y <- c(52,  104,  146,  10,  51,   30,   40,  27,  46)
> x_tilde <- x - mean(x) + mean(c(x,y))
> y_tilde <- y - mean(y) + mean(c(x,y))
> t0 <-  t.test(x,y,alternative='greater',paired=F,var.equal=F)$
   statistic
> B <- 1000
> theta_values <- numeric(B)
> n1 <- length(x)
> n2 <- length(y)
> set.seed(1)
> for (i in 1:B){
+ x_b <- x_tilde[sample(n1, n1, replace = TRUE)]
+ y_b <- y_tilde[sample(n2, n2, replace = TRUE)]
+ theta_values[i] <- t.test(x_b,y_b,alternative='greater',paired=F,var
   .equal=F)$statistic
+ }
> pvalue_boot <- length(theta_values[theta_values>t0])/B
> pvalue_boot
[1] 0.162
```

Το ιστόγραμμα των προσομοιωμένων τιμών παρατίθεται στο Σχήμα 11.5. Παρατηρούμε ότι η bootstrap εκτίμηση της p-value με βάση 1000 επαναλήψεις ισούται με 0.162, το οποίο αντιστοιχεί στη σχετική συχνότητα των φορών όπου οι bootstrap τιμές ξεπερνούν την τιμή 1.058 (δηλαδή την κόκκινη γραμμή του Σχήματος 11.5). Συμπεραίνουμε ότι δεν έχουμε στοιχεία για την απόρριψη της μηδενικής υπόθεσης κατά της εναλλακτικής στα συνήθη επίπεδα σημαντικότητας. Το αποτέλεσμα αυτό συμφωνεί με το συμπέρασμα που προκύπτει μέσω του t-test και του ελέγχου Wilcoxon-Mann-Whitney, όπως φαίνεται στον Πίνακα 11.5 (αφήνεται ως άσκηση στον/στην αναγνώστη/στρια η επιβεβαίωση του αποτελέσματος των ελέγχων t-test και Wilcoxon-Mann-Whitney). □



Σχήμα 11.5: Ιστόγραμμα $B = 1000$ bootstrap τιμών της (11.14) για τα δεδομένα του Παραδείγματος 11.13. Η κόκκινη γραμμή αντιστοιχεί στην παρατηρηθείσα τιμή της (11.14).

Μέθοδος	bootstrap	t-test	Mann-Whitney
p-value	0.162	0.158	0.340

Πίνακας 11.6: Οι p-τιμές διαφορετικών τεχνικών για τον έλεγχο της $H_0 : \mu_1 = \mu_2$ έναντι της $H_1 : \mu_1 > \mu_2$ στα δεδομένα του Παραδείγματος 11.13.

11.3.3 Εκτίμηση συνάρτησης πυκνότητας

Στο Κεφάλαιο 3 γνωρίσαμε την τεχνική εκτίμησης μιας άγνωστης συνάρτησης πυκνότητας πιθανότητας $f(x)$ μέσω πυρήνα $f_{h,n}(x)$. Με βάση το bootstrap μπορεί να υπολογιστεί εύκολα ένα διάστημα εμπιστοσύνης. Θα χρησιμοποιήσουμε, ως παράδειγμα, τα δεδομένα Bart Simpson του Κεφαλαίου 3, επιλέγοντας την εκτίμηση της πυκνότητας μέσω κανονικού πυρήνα. Στη συνέχεια, παρατίθεται ο ψευδοκώδικας που υλοποιεί το bootstrap στο πρόβλημα αυτό.

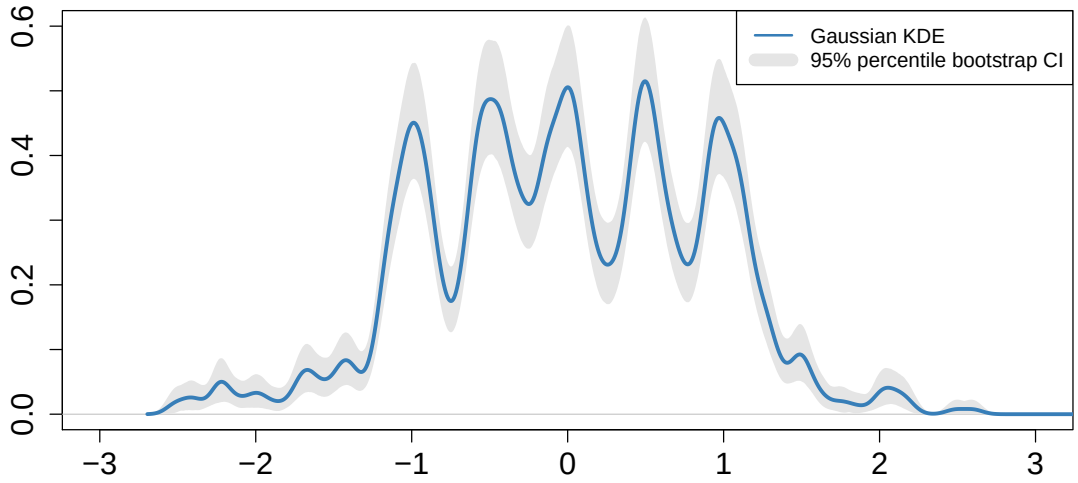
Από το Παράδειγμα 3.9, έχουμε ότι η τιμή της παραμέτρου εξομάλυνσης (εύρος παραθύρου), που ελαχιστοποιεί το IMSE, ισούται με $h = 0.059$, η οποία αντιστοιχεί στην μπλε γραμμή του Σχήματος 11.6. Η γκρι ζώνη απεικονίζει το 95% bootstrap διάστημα εμπιστοσύνης ποσοστιαίων σημείων για κάθε τιμή x του οριζόντιου άξονα. Παρατηρήστε ότι υπάρχει αυξημένη αβεβαιότητα στις κορυφές της πυκνότητας (μεγαλύτερο εύρος διαστήματος εμπιστοσύνης) σε σχέση με τις υπόλοιπες τιμές. Κάποια μειονεκτήματα του συγκεκριμένου διαστήματος είναι ότι πρόκειται για σημειακό διάστημα (για καθένα x) και όχι για όλες τις τιμές της συνάρτησης πυκνότητας πιθανότητας ταυτόχρονα. Επίσης, πρόκειται για διάστημα εμπιστοσύνης για την $\bar{f}_h(x) := \mathrm{E}\left(f_{h,n}(x)\right)$ (θυμηθείτε τη σχέση (3.2)) και όχι για την ίδια την $f(x)$. Σε κάθε περίπτωση, όμως, μέσω της ζώνης εμπιστοσύνης αντανακλώνται πράγματι τα βασικά χαρακτηριστικά της f, όπως η ύπαρξη πέντε κορυφών. Αφού μελετήσετε τον ψευδοκώδικα που υλοποιεί το bootstrap στο πρόβλημα αυτό, ανατρέξτε στην Άσκηση 11.11.

Algorithm 1: Αλγόριθμος bootstrap για την εκτίμηση συνάρτησης πυκνότητας πιθανότητας.

Input : $\alpha \in (0,1)$: $1-$συντελεστής εμπιστοσύνης
$B \in \mathbb{Z}_+$: αριθμός bootstrap δειγμάτων
$\boldsymbol{x} = (x_1, \ldots, x_n)$: παρατηρηθέντα δεδομένα
$M \in \mathbb{Z}_+$: ακέραιος που θα καθορίσει ένα σύνολο τιμών $(t_1, \ldots, t_M)$, στις οποίες θα υπολογιστεί η $f_{h,n}(\cdot)$
$h > 0$: bandwidth.
Output: $C : M \times 2$ πίνακας με το άνω και κάτω όριο του ΔΕ στο t_m, $m = 1, \ldots, M$
Step 1: Bootstrap sampling
  for $b = 1$ **to** B
    Step 1.1 sampling:
      Λάβε τυχαίο δείγμα $\boldsymbol{y}$ με επανάθεση από το $\boldsymbol{x}$, μεγέθους n
    Step 1.2 Για κάθε t_m υπολόγισε την εκτίμηση της πυκνότητας, δοθέντος $\boldsymbol{y}$:
      for all $m \in \{1, \ldots, M\}$: $T^{\star}_{b,m} = f^{\star}_{h,n}(t_m)$;
  endfor

Step 2: $100(1-\alpha)\%$ **percentile boostrap CI**
  for $m = 1$ **to** M
    $C_{m,1} = \alpha/2$ – κάτω ποσοστιαίο σημείο του $T^{\star}_{1:B,m}$
    $C_{m,2} = 1 - \alpha/2$ – άνω ποσοστιαίο σημείο του $T^{\star}_{1:B,m}$
  endfor

END of algorithm



Σχήμα 11.6: Δεδομένα Bart Simpson: εκτίμηση συνάρτησης πυκνότητας πιθανότητας με χρήση κανονικού πυρήνα και (σημειακά) 95% bootstrap διαστήματα εμπιστοσύνης με τη μέθοδο των ποσοστιαίων σημείων.

11.4 Θέματα για περαιτέρω μελέτη

Η χρησιμότητα του bootstrap αναδεικνύεται σε προβλήματα στα οποία δεν είναι εύκολη ή εύλογη η χρήση παραμετρικών υποθέσεων και η ενότητα των ασκήσεων που ακολουθεί παρουσιάζει αρκετά τέτοια προβλήματα. Υπάρχουν, επιπλέον, και οι περιπτώσεις όπου το bootstrap δίνει απαντήσεις όταν άλλες μη

παραμετρικές τεχνικές (π.χ. το jackknife) αποτυγχάνουν: ένα τέτοιο παράδειγμα δίνεται στην Άσκηση 11.1. Θα πρέπει να τονίσουμε ότι η μέθοδος bootstrap δεν είναι δυνατόν να καλυφθεί επαρκώς σε ένα μόνο κεφάλαιο ενός προπτυχιακού συγγράμματος. Για πληρέστερη παρουσίαση της τεχνικής παραπέμπουμε στους Hall (1992), Efron and Tibshirani (1994), Davison and Hinkley (1997), Shao and Tu (2012), Boos (2003), Efron (2003), Janssen and Pauls (2003), Politis and Romano (1994), Wu (1986) και Mammen and Nandi (2012). Ωστόσο, το bootstrap δεν είναι πανάκεια για όλα τα προβλήματα που μπορεί να αντιμετωπίσει ένας στατιστικός. Λέμε ότι το bootstrap είναι συνεπές όταν η διαφορά (μετρούμενη μέσω κατάλληλων αποστάσεων) μεταξύ της bootstrap εκτίμησης και της πραγματικής τιμής του συναρτησιακού τείνει στο μηδέν. Όταν το συναρτησιακό είναι διαφορίσιμο κατά Hadamard (Shapiro, 1990; Van Der Vaart, 1991), μπορεί να δειχθεί ότι το bootstrap είναι συνεπές (βλ. Θεώρημα 3.21 στο βιβλίο του Wasserman, 2006). Η περιγραφή τέτοιων ιδιοτήτων ξεφεύγει από τους σκοπούς του παρόντος συγγράμματος και παραπέμπουμε τους/τις ενδιαφερόμενους/μενες αναγνώστες/στριες στις εργασίες των Bickel and Freedman (1981), Beran and Ducharme (1991) και Inoue and Kilian (2003).

Κάποιες χαρακτηριστικές περιπτώσεις, όπου το τυπικό bootstrap αποτυγχάνει, είναι οι εξής:

- Μικρά σύνολα δεδομένων (η F_n δεν είναι καλή προσέγγιση της F).
- Προβλήματα μεγάλων διαστάσεων: καθώς η διάσταση των μεταβλητών αυξάνεται, η F_n δεν αποτελεί καλή προσέγγιση της F για πεπερασμένο n (ανακαλέστε τη σχετική συζήτηση στην 3.5).
- Κατανομές με μη πεπερασμένες ροπές (Athreya, 1987).
- Εξαρτημένες παρατηρήσεις (π.χ. χρονολογικές σειρές, χωρικά προβλήματα). Σε τέτοιου είδους περιπτώσεις υπάρχουν παραλλαγές του τυπικού bootstrap (Bühlmann, 2002).
- Εκτίμηση ακραίων τιμών, όπως
  - εκτίμηση του 99.99% ποσοστιαίου σημείου,
  - εκτίμηση παραμέτρων που βρίσκονται στο όριο του παραμετρικού χώρου: για παράδειγμα, θεωρήστε την περίπτωση τυχαίου δείγματος από την κατανομή $\mathscr{U}(0, \theta)$ και ότι ενδιαφέρει η εκτίμηση του θ.

Υπάρχουν αρκετές παραλλαγές του τυπικού bootstrap που διορθώνουν αδύνατα σημεία της μεθόδου. Η τεχνική smoothed bootstrap (Silverman and Young, 1987; Hall *et al.*, 1989; De Angelis and Young, 1992) αντικαθιστά την (διακριτή) εμπειρική αθροιστική συνάρτηση κατανομής με μία λεία (συνεχή) εκτίμηση της συνάρτησης πυκνότητας πιθανότητας που προκύπτει μέσω πυρήνα (βλ. Άσκηση 3.8, Κεφάλαιο 2). Η τεχνική του επαναληπτικού bootstrap (Booth and Hall, 1994; Lee and Young, 1995; Booth and Presnell, 1998) μπορεί να βελτιώσει τη συμπερασματολογία σε περιπτώσεις μη «λείων» συναρτησιακών, όπως είναι τα ποσοστιαία σημεία. Η Μπεϋζιανή bootstrap προσέγγιση (Rubin, 1981; Lo, 1987, 1988) αντιστοιχίζει σε κάθε παρατήρηση μία πιθανότητα η οποία μεταβάλλεται από επανάληψη σε επανάληψη, αντί να δίνει βάρος $1/n$ σε κάθε παρατήρηση. Όπως υπαινίσσεται η ονομασία αυτής της παραλλαγής, η κατανομή της παραμέτρου πάνω στην οποία εφαρμόζεται το Μπεϋζιανό bootstrap μπορεί να θεωρηθεί ως εκ των υστέρων κατανομή. Τέλος, σε περιπτώσεις εξαρτημένων παρατηρήσεων χρησιμοποιούνται τεχνικές block bootstrap (Lahiri, 1993; Carlstein *et al.*, 1998; Lahiri, 1999).

11.5 Ασκήσεις

Άσκηση 11.1 (bootstrap στη διάμεσο)**.** Προσομοιώστε $n = 30$ το πλήθος παρατηρήσεις από την κατανομή $\mathscr{E}(\log 2)$ (δηλαδή από Εκθετική κατανομή με μέση τιμή $1/\log 2$) μέσω των

```
R> set.seed(1)
R> x <- rexp(n = 30, rate = log(2))
```

1. Εφαρμόστε jackknife για να εκτιμήσετε το τυπικό σφάλμα εκτιμητή αντικατάστασης για τη διάμεσο. Τι παρατηρείτε;
2. Προσομοιώστε $B = 1000$ bootstrap τιμές της δειγματικής διαμέσου και κατασκευάστε ένα ιστόγραμμα των προσομοιωμένων τιμών.
3. Κατασκευάστε 95% bootstrap διαστήματα εμπιστοσύνης (κανονικά, βασικά και ποσοστιαίων σημείων) για τη διάμεσο.
4. Υπολογίστε το 95% διωνυμικό διάστημα εμπιστοσύνης για τη διάμεσο.

Άσκηση 11.2. Θεωρήστε το αρχείο δεδομένων συσπάσεων νευρικής ίνας (βλ. Ενότητα 2.5). Να υπολογιστούν 95% bootstrap διαστήματα εμπιστοσύνης (κανονικά, βασικά και ποσοστιαίων σημείων) για τα τεταρτημόρια της κατανομής των παρατηρήσεων (q_p, $p = 0.25, 0.5, 0.75$).

Άσκηση 11.3 (bootstrap στον συντελεστή Gini)**.** Να γίνει bootstrap στον συντελεστή Gini, χρησιμοποιώντας τα δεδομένα της Άσκησης 10.8, και να υπολογιστούν 95% bootstrap διαστήματα εμπιστοσύνης.

Άσκηση 11.4 (bootstrap στον λόγο δειγματικών μέσων)**.** Να γίνει bootstrap στον λόγο δειγματικών μέσων, χρησιμοποιώντας τα δεδομένα της Άσκησης 10.9, και να υπολογιστούν 95% bootstrap διαστήματα εμπιστοσύνης.

Άσκηση 11.5 (bootstrap για συντελεστή συσχέτισης)**.** Σε ένα τυχαίο δείγμα $n = 15$ το πλήθος φοιτητών σε σχολές νομικής στην Αμερική, μετρήθηκαν οι μεταβλητές LSAT (average score on a national law test) και GPA (average undergraduate grade-point average). Τα δεδομένα παρατίθενται στον Πίνακα 11.7.

	LSAT	GPA
1	576	3.39
2	635	3.30
3	558	2.81
4	578	3.03
5	666	3.44
6	580	3.07
7	555	3.00
8	661	3.43
9	651	3.36
10	605	3.13
11	653	3.12
12	575	2.74
13	545	2.76
14	572	2.88
15	594	2.96

Πίνακας 11.7: Μετρήσεις 15 φοιτητών σχολών νομικής που αφορούν τις μεταβλητές LSAT (average score on a national law test) και GPA (average undergraduate grade-point average). Πηγή: Efron and Tibshirani (1994).

1. Προσομοιώστε $B = 1000$ bootstrap τιμές του συντελεστή συσχέτισης μεταξύ LSAT και GPA και κατασκευάστε ένα ιστόγραμμα των προσομοιωμένων τιμών.
2. Κατασκευάστε 95% bootstrap διαστήματα εμπιστοσύνης (κανονικά, βασικά και ποσοστιαιών σημείων) για τον συντελεστή συσχέτισης μεταξύ των δύο μεταβλητών.

Υπόδειξη: τα παραπάνω να προγραμματιστούν βάσει δικού σας κώδικα αλλά και του πακέτου `boot`.

Άσκηση 11.6. Θεωρήστε τα δεδομένα του Πίνακα 11.5 και εκτιμήστε το απλό γραμμικό μοντέλο με εξαρτημένη μεταβλητή την log(surv) και ανεξάρτητη την dose. Στη συνέχεια, κάντε bootstrap στα κατάλοιπα του γραμμικού μοντέλου και υπολογίστε 95% διαστήματα εμπιστοσύνης των εκτιμητών ελαχίστων τετραγώνων χρησιμοποιώντας δικό σας κώδικα.

Άσκηση 11.7. Θεωρήστε τα δεδομένα του Πίνακα 11.5 και εκτιμήστε το απλό γραμμικό μοντέλο log(surv) και ανεξάρτητη την dose. Στη συνέχεια, κάντε bootstrap στις παρατηρήσεις (x_i, y_i) και υπολογίστε 95% διαστήματα εμπιστοσύνης των εκτιμητών ελαχίστων τετραγώνων χρησιμοποιώντας δικό σας κώδικα (χωρίς χρήση του πακέτου `boot`). Παρατηρείτε διαφορές σε σχέση με την ανάλυση μέσω `bootstrap` στα κατάλοιπα;

Άσκηση 11.8 (bootstrap σε πίνακα συνάφειας)**.** Στον Πίνακα 11.8 δίνεται δείγμα 419 ατόμων ταξινομημένο ως προς το αν πάσχουν από κατάθλιψη και αν είχαν κάποια τραυματική εμπειρία στο παρελθόν. Ενδιαφερόμαστε να διερευνήσουμε αν η κατάθλιψη επηρεάζεται από την ύπαρξη τραυματικής εμπειρίας μέσω της διαφοράς των δεσμευμένων πιθανοτήτων

$$\theta = P(Y = 1|X = 1) - P(Y = 1|X = 2), \tag{11.15}$$

όπου οι αριθμοί υποδηλώνουν τον δείκτη γραμμών/στηλών του Πίνακα 11.8. Από την ανάλυση κατηγορικών δεδομένων, γνωρίζουμε ότι ο εκτιμητής μέγιστης πιθανοφάνειας του θ είναι ο:

$$T = \widehat{P}(Y = 1|X = 1) - \widehat{P}(Y = 1|X = 2) = \frac{n_{11}}{n_{11} + n_{12}} - \frac{n_{21}}{n_{21} + n_{22}},$$

όπου n_{ij} η παρατηρηθείσα συχνότητα στο κελί (i, j) του πίνακα.

		Κατάθλιψη (Y)	
		ναι	όχι
Τραυματική εμπειρία (X)	ναι	33	131
	όχι	4	251

Πίνακας 11.8: Δεδομένα κατάθλιψης και τραυματικής εμπειρίας.

1. Περιγράψτε τον τρόπο με τον οποίο θα λάβετε bootstrap δείγματα από τα παραπάνω δεδομένα.
2. Προσομοιώστε $B = 10000$ bootstrap τιμές T και κατασκευάστε ένα ιστόγραμμα των προσομοιωμένων τιμών.
3. Κατασκευάστε 95% bootstrap διαστήματα εμπιστοσύνης (κανονικά, βασικά και ποσοστιαιών σημείων) για το θ. Συγκρίνετε τα αποτελέσματα με την πραγματοποίηση του 95% ασυμπτωτικού διαστήματος εμπιστοσύνης για το θ: $[0.130, 0.241]$.

Άσκηση 11.9. Να εκτιμήσετε μέσω bootstrap την υπόθεση $H_0 : \mu = 1$ έναντι της $H_1 : \mu \neq 1$ στα δεδομένα του Παραδείγματος 11.12 χρησιμοποιώντας τη στατιστική συνάρτηση $T_n = \sqrt{n}\frac{\bar{X}-\mu_0}{S_n}$, όπου $S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$. Να συγκρίνετε το αποτέλεσμά σας με αυτό του Παραδείγματος 11.12.

Ναρκωτικό	Καταθλιπτική Επίδραση
MDMA	$28, 35, 35, 24, 39, 32, 27, 29, 36, 35$
Αλκοόλ	$5, 6, 30, 8, 9, 7, 6, 17, 3, 10$

Πίνακας 11.9: Καταθλιπτική επίδραση ναρκωτικών. Πηγή: Field *et al.* (2012).

Άσκηση 11.10. Τα δεδομένα στον Πίνακα 11.9 αφορούν την καταθλιπτική επίδραση δύο ειδών ναρκωτικών σε δείγμα 20 clubbers, όπου σε 10 δόθηκε MDMA (3,4-Methylenedioxymethamphetamine, γνωστό και ως ecstasy) και στους υπόλοιπους αλκοόλ.

Να εκτιμηθεί μέσω bootstrap η p-value του ελέγχου υπόθεσης

$$H_0 : \mu_1 = \mu_2 \quad \text{έναντι της} \quad H_1 : \mu_1 > \mu_2$$

όπου μ_1 και μ_2 η μέση καταθλιπτική επίδραση του MDMA και του αλκοόλ, αντίστοιχα.

Άσκηση 11.11. Θεωρήστε τα δεδομένα Bart Simpson του Κεφαλαίου 3 και τον εκτιμητή πυκνότητας με χρήση κανονικού πυρήνα. Προσομοιώστε bootstrap τιμές της εκτίμησης της συνάρτησης πυκνότητας πιθανότητας και υπολογίστε 95% bootstrap διαστήματα εμπιστοσύνης ποσοστιαίων σημείων με βάση τον ψευδοκώδικα που δόθηκε στην Ενότητα 11.3.3. Απεικονίστε σε ένα κοινό διάγραμμα την εκτιμηθείσα πυκνότητα μαζί με το διάστημα εμπιστοσύνης. Υπόδειξη: το αποτέλεσμα θα πρέπει να μοιάζει με το Σχήμα 11.6.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ξενόγλωσση

Athreya, K. B. (1987). Bootstrap of the Mean in the Infinite Variance Case. *The Annals of Statistics*, *15*(2), pp. 724–731.

Barnard, B. (1974). Conditionality, pivotals, and robust estimation. In: *Proceedings of the Conference on Foundational Questions in Statistical Inference. Memoirs*. 1.

Beran, R. and Ducharme, G. R. (1991). *Asympotic Theory for Bootstrap Methods in Statistics*. Centre De Recherches Mathematiques.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, *9*, pp. 1196–1217.

Boos, D. D. (2003). Introduction to the Bootstrap World. *Statistical Science*, *18*(2), pp. 168–174.

Booth, J. and Presnell, B. (1998). Allocation of Monte Carlo resources for the iterated bootstrap. *Journal of Computational and Graphical Statistics*, *7*(1), pp. 92–112.

Booth, J. G. and Hall, P. (1994). Monte Carlo approximation and the iterated bootstrap. *Biometrika*, *81*(2), pp. 331–340.

Bühlmann, P. (2002). Bootstraps for time series. *Statistical Science*, *17*, pp. 52–72.

Carlstein, E., Do, K.-A., Hall, P., Hesterberg, T. and Künsch, H. R. (1998). Matched-block bootstrap for dependent data. *Bernoulli*, *4*, pp. 305–328.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. 1. Cambridge University Press.

De Angelis, D. and Young, G. A. (1992). Smoothing the bootstrap. *International Statistical Review/Revue Internationale de Statistique*, *60*, pp. 45–56.

Diaconis, P. and Efron, B. (1983). Computer-Intensive Methods in Statistics. *Scientific American*, *248*(5), pp. 116–131.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), pp. 1–26.

Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, *82*(397), pp. 171–185.

Efron, B. (1988). Computer-intensive methods in statistical regression. *Siam Review*, *30*(3), pp. 421–449.

Efron, B. (2003). Second thoughts on the bootstrap. *Statistical Science*, *18*, pp. 135–140.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*, pp. 54–75.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC press.

Field, A., Miles, J. and Field, Z. (2012). *Discovering statistics using R*. Sage publications.

Fisher, R. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*, pp. 309–368.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, *6*(4), pp. 391–398.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, *9*(6), pp. 1218–1228.

Hall, P. (1992). On bootstrap confidence intervals in nonparametric regression. *The Annals of Statistics*, *20*, pp. 695–711.

Hall, P. (2013). *The bootstrap and Edgeworth expansion*. Springer Science & Business Media.

Hall, P., DiCiccio, T. J. and Romano, J. P. (1989). On smoothing and the bootstrap. *The Annals of Statistics*, *17*, pp. 692–704.

Inoue, A. and Kilian, L. (2003). The continuity of the limit distribution in the parameter of interest is not essential for the validity of the bootstrap. *Econometric Theory*, *19*(6), pp. 944–961.

Janssen, A. and Pauls, T. (2003). How do bootstrap and permutation tests work? *The Annals of Statistics*, *31*(3), pp. 768–806.

Lahiri, S. N. (1999). Theoretical comparisons of block bootstrap methods. *Annals of Statistics*, *27*, pp. 386–404.

Lahiri, S. N. (1993). On the moving block bootstrap under long range dependence. *Statistics & Probability Letters*, *18*(5), pp. 405–413.

Lee, S. M. and Young, G. A. (1995). Asymptotic iterated bootstrap confidence intervals. *The Annals of Statistics*, *23*, pp. 1301–1330.

Lo, A. Y. (1987). A large sample study of the Bayesian bootstrap. *The Annals of Statistics*, *15*, pp. 360–375.

Lo, A. Y. (1988). A Bayesian bootstrap for a finite population. *The Annals of Statistics*, *16*, pp. 1684–1695.

Mammen, E. and Nandi, S. (2012). Bootstrap and resampling. In: *Handbook of Computational Statistics*. Springer, pp. 499–527.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, *89*(428), pp. 1303–1313.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, *9*, pp. 130–134.

Shao, J. and Tu, D. (2012). *The jackknife and bootstrap*. Springer Science & Business Media.

Shapiro, A. (1990). On concepts of directional differentiability. *Journal of Optimization Theory and Applications*, *66*(3), pp. 477–487.

Silverman, B. and Young, G. (1987). The bootstrap: To smooth or not to smooth? *Biometrika*, *74*(3), pp. 469–479.

Van Der Vaart, A. (1991). Efficiency and Hadamard differentiability. *Scandinavian Journal of Statistics*, *18*, pp. 63–75.

Wasserman, L. (2006). *All of Nonparametric Statistics*. New York, NY: Springer Texts in Statistics.

Welch, B. L. (1938). The Significance of the Difference Between Two Means when the Population Variances are Unequal. *Biometrika*, *29*(3/4), pp. 350–362.

Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, *34*(1-2), pp. 28–35.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, *14*(4), pp. 1261–1295.

ΚΕΦΑΛΑΙΟ 12

ΜΗ ΠΑΡΑΜΕΤΡΙΚΟΣ ΣΤΑΤΙΣΤΙΚΟΣ ΕΛΕΓΧΟΣ ΔΙΕΡΓΑΣΙΩΝ

Σύνοψη

Σκοπός του κεφαλαίου αυτού είναι η παρουσίαση μίας εκ των πλέον σύγχρονων εφαρμογών της Μη Παραμετρικής Στατιστικής, του στατιστικού ελέγχου διεργασιών. Ο στατιστικός έλεγχος διεργασιών είναι ένα σύνολο (στατιστικών) μεθόδων οι οποίες έχουν ως στόχο την παρακολούθηση διαδικασιών που εξελίσσονται κυρίως ως προς τον χρόνο, με σκοπό να ανιχνευθούν αλλαγές σε αυτές. Βασικό εργαλείο του στατιστικού ελέγχου διεργασιών είναι το διάγραμμα ελέγχου. Μετά από μία σύντομη εισαγωγή στην έννοια του διαγράμματος ελέγχου, παρουσιάζονται η λειτουργία και ο τρόπος εφαρμογής των πιο συχνά χρησιμοποιούμενων μη παραμετρικών διαγραμμάτων ελέγχου, τα οποία βασίζονται σε μη παραμετρικούς ελέγχους, όπως το προσημικό κριτήριο (sign test), το κριτήριο των προσημασμένων τάξεων (Wilcoxon's test), το κριτήριο Προηγήσεων (Precedence test) ή το κριτήριο του αθροίσματος τάξεων (Mann-Whitney test).

Προαπαιτούμενη γνώση:

Βασικές γνώσεις Πιθανοτήτων και Στατιστικής.

Κεφάλαια 5 και 6 του παρόντος συγγράμματος.

Προσδοκώμενα μαθησιακά αποτελέσματα:

Ο/η φοιτητής/τρια ολοκληρώνοντας την ενότητα αυτή θα μπορεί να αναπτύσσει και να εφαρμόζει μη παραμετρικά διαγράμματα ελέγχου για την παρακολούθηση διεργασιών, να υπολογίζει την απόδοσή τους και να τα συγκρίνει μεταξύ τους.

Γλωσσάριο επιστημονικών όρων

- Διάγραμμα Ελέγχου
- Έλεγχος Mann-Whitney
- Κατανομή Μήκους Ροής
- Κριτήριο Προσημασμένων Τάξεων
- Κριτήριο Προηγήσεων
- Μέσο Μήκος Ροής
- Πιθανότητα Εσφαλμένου Συναγερμού
- Προσημικό Κριτήριο
- Στατιστικός έλεγχος διεργασιών

Μπατσίδης, Α., Παπασταμούλης, Π., Πετρόπουλος, Κ., & Ρακιτζής, Α. (2022). *Μη Παραμετρική Στατιστική.* [Προπτυχιακό εγχειρίδιο].

12.1 Εισαγωγή

Ο Στατιστικός Έλεγχος Διεργασιών αποτελείται από ένα σύνολο στατιστικών μεθόδων και τεχνικών, οι οποίες έχουν ως στόχο την παρακολούθηση διεργασιών (στοχαστικών διαδικασιών) προκειμένου να ανιχνευθούν έγκαιρα πιθανές ανεπιθύμητες καταστάσεις. Το πιο σημαντικό εργαλείο όλων αυτών των μεθόδων είναι το διάγραμμα ελέγχου. Τα διαγράμματα ελέγχου προτάθηκαν από τον Walter Andrew Shewhart (1891-1967) στη διάρκεια της δεκαετίας του 1920 και είναι μια γραφική αναπαράσταση σημείων στο επίπεδο. Κάθε σημείο που αναπαρίσταται αποτελεί την τιμή μιας κατάλληλης στατιστικής συνάρτησης, η οποία περιέχει πληροφορία σχετικά με τη συμπεριφορά της διεργασίας. Έτσι, παρατηρώντας τις τιμές των σημείων στο επίπεδο και το πώς αυτές συμπεριφέρονται, μπορούμε να εξάγουμε χρήσιμα συμπεράσματα σχετικά με τη διεργασία. Ουσιαστικά, αυτό που μας ενδιαφέρει είναι να συμπεράνουμε αν η διεργασία συμπεριφέρεται «ομαλά» (δηλαδή είναι εντός ελέγχου) ή όχι (και άρα είναι εκτός ελέγχου).

Εκτός από τα σημεία που απεικονίζονται στο επίπεδο, σε ένα διάγραμμα ελέγχου απεικονίζεται και ένα όριο ελέγχου (όταν αναφερόμαστε σε μονόπλευρα διαγράμματα ελέγχου) ή δύο όρια ελέγχου (στην περίπτωση των δίπλευρων διαγραμμάτων ελέγχου). Στην πραγματικότητα, τα όρια ελέγχου δεν είναι παρά γραμμές στο επίπεδο, οι οποίες, σε αναλογία με τη διαδικασία του ελέγχου υποθέσεων, καθορίζουν περιοχές απόφασης και προσδιορίζονται με στατιστικά κριτήρια. Ανάλογα με το πού απεικονίζονται τα σημεία (εντός του διαστήματος που ορίζουν τα όρια ελέγχου ή εκτός αυτού), μπορούμε να αποφασίσουμε αν η διεργασία λειτουργεί όπως επιθυμούμε ή ότι έχουν συμβεί αλλαγές για τις οποίες πρέπει να ληφθεί ειδική μέριμνα. Για να λάβουμε απόφαση, θα πρέπει να ορίσουμε κατάλληλους κανόνες με τη βοήθεια των οποίων ανακηρύσσουμε μια διεργασία ως εκτός ελέγχου ή όχι.

Παρακάτω δίνουμε κάποιες βασικές έννοιες του στατιστικού ελέγχου διεργασιών, καθώς και των διαγραμμάτων ελέγχου, οι οποίες θα μας χρειαστούν στη συνέχεια της μελέτης αυτού του κεφαλαίου. Δεν είναι δυνατόν να παρουσιαστούν όλες οι τεχνικές και όλα τα είδη των διαγραμμάτων ελέγχου, διότι κάτι τέτοιο ξεφεύγει από τους σκοπούς του παρόντος συγγράμματος. Για όποιον/όποια επιθυμεί να εμβαθύνει στις έννοιες και στις μεθόδους του στατιστικού ελέγχου διεργασιών παραπέμπουμε στο βιβλίο του Montgomery (2020), καθώς και στις πανεπιστημιακές σημειώσεις/συγγράμματα των Αντζουλάκος (2003), Ταγαράς (2001) και Μπερσίμης κ.ά. (2021).

12.1.1 Κατηγορίες αιτιών μεταβλητότητας

Τα διαγράμματα ελέγχου, αρχικά, εφαρμόστηκαν για την παρακολούθηση παραγωγικών διεργασιών (παραγωγή προϊόντων/αντικειμένων στη βιομηχανία), αλλά πλέον μπορούν να εφαρμοστούν και για την παρακολούθηση μη βιομηχανικών διεργασιών. Είναι γνωστό ότι κάθε διεργασία, όσο καλά σχεδιασμένη και αν είναι και όσο προσεκτικά και αν επιβλέπεται, θα λειτουργεί παρουσία μιας φυσικής μεταβλητότητας. Ως φυσική μεταβλητότητα εννοούμε το αθροιστικό αποτέλεσμα πολλών μικρών αιτιών, οι οποίες αναφέρονται ως κοινές ή τυχαίες αιτίες μεταβλητότητας (common or chance causes of variation). Η φυσική μεταβλητότητα είναι συνήθως μικρή σε μέγεθος και δεν μπορεί να αποδοθεί σε ελέγξιμους παράγοντες. Μια διεργασία η οποία λειτουργεί μόνο με την παρουσία φυσικής μεταβλητότητας λέμε ότι είναι μια εντός ελέγχου διεργασία (in-control process ή IC). Μια διεργασία η οποία είναι εντός ελέγχου θεωρείται ότι λειτουργεί σε σταθερή κατάσταση και σύμφωνα με τα πρότυπα (standards) που έχουμε θέσει.

Όμως, σε μια διεργασία μπορεί να εμφανίζονται περιστασιακά και άλλες μορφές μεταβλητότητας οι οποίες δεν οφείλονται σε τυχαίες αιτίες, αλλά αφορούν τη συστηματική αλλαγή στο επίπεδο κάποιου ή κάποιων παραγόντων που καθορίζουν την ποιότητα των προϊόντων που παράγονται. Τέτοιες αιτίες μπορεί να είναι η λανθασμένη ρύθμιση των μηχανών παραγωγής, τα λάθη των χειριστών των μηχανημάτων ή η κακής ποιότητας πρώτη ύλη. Η μεταβλητότητα αυτή είναι σε μέγεθος πολύ μεγαλύτερη της φυσικής μεταβλητότητας και η παρουσία της οδηγεί συνήθως σε μη αποδεκτά επίπεδα λειτουργίας της διεργασίας.

Θα αναφερόμαστε σε αυτή ως ειδική μεταβλητότητα και οι αιτίες που οδηγούν στην εμφάνισή της ονομάζονται ειδικές αιτίες μεταβλητότητας (assignable causes of variation). Μια διεργασία η οποία λειτουργεί με την παρουσία ειδικών αιτιών μεταβλητότητας λέμε ότι είναι μια εκτός ελέγχου διεργασία (out-of-control process ή OOC). Μια διεργασία η οποία είναι εκτός ελέγχου θεωρείται ότι λειτουργεί σε μη σταθερή κατάσταση και έχουν συμβεί σημαντικές αλλαγές. Οι αλλαγές αυτές θα πρέπει να ανιχνευθούν και να απομακρυνθούν ώστε η διεργασία να επανέλθει σε εντός ελέγχου κατάσταση.

Έτσι, τα διαγράμματα ελέγχου μπορούν να χρησιμοποιηθούν για την επίβλεψη (παρακολούθηση μιας διεργασίας), ώστε να διαπιστωθεί αν αυτή λειτουργεί παρουσία μόνο φυσικής μεταβλητότητας (και άρα είναι εντός ελέγχου) ή αν λειτουργεί παρουσία ειδικών αιτιών μεταβλητότητας (και άρα είναι εκτός ελέγχου).

12.1.2 Προοπτική και αναδρομική εφαρμογή διαγραμμάτων ελέγχου

Τα διαγράμματα ελέγχου μπορούν να χρησιμοποιηθούν είτε προοπτικά (για την παρακολούθηση της διεργασίας σε πραγματικό χρόνο, online ή prospective monitoring) είτε αναδρομικά (για εκτίμηση των παραμέτρων της διεργασίας όταν αυτή είναι εντός ελέγχου, offline ή retoprospective monitoring). Οι δύο αυτές χρήσεις των διαγραμμάτων ελέγχου συνδέονται με δύο διακριτές φάσεις εφαρμογής τους, την ανάλυση Φάσης I και την ανάλυση Φάσης II.

Κατά την ανάλυση Φάσης I συλλέγεται ένα σύνολο δεδομένων από τη διεργασία (προκαταρκτικό δείγμα ή δείγμα αναφοράς) και, στη συνέχεια, με την εφαρμογή κατάλληλων στατιστικών μεθόδων, προσπαθούμε να απαντήσουμε στο ερώτημα αν η διεργασία ήταν εντός ή εκτός ελέγχου κατά τη χρονική περίοδο συλλογής των δεδομένων. Σε αυτήν τη φάση τα διαγράμματα ελέγχου βοηθούν τον διαχειριστή της διεργασίας να εκτιμήσει τις τιμές των παραμέτρων της διεργασίας όταν αυτή είναι εντός ελέγχου. Επίσης, προσδιορίζονται οι τιμές των ορίων ελέγχου οι οποίες θα χρησιμοποιηθούν για την παρακολούθηση της διεργασίας σε πραγματικό χρόνο. Άμεσα, λοιπόν, διαπιστώνουμε ότι η εφαρμογή των διαγραμμάτων ελέγχου Φάσης I αποτελεί αναδρομική εφαρμογή αυτών.

Κατά την ανάλυση Φάσης II τα διαγράμματα ελέγχου χρησιμοποιούνται προκειμένου να παρακολουθήσουμε τη διεργασία σε πραγματικό χρόνο ώστε να ανιχνεύσουμε εγκαίρως μια πιθανή αλλαγή στο μέσο επίπεδο ή/και τη μεταβλητότητα της κατανομής του χαρακτηριστικού, το οποίο καθορίζει την ποιότητα των προϊόντων που παράγονται. Κατά την ανάλυση Φάσης II, σε κάθε χρονική περίοδο που ένα δείγμα λαμβάνεται από τη διεργασία, ο διαχειριστής δίνει άμεσα μια απάντηση στο ερώτημα «είναι εντός ελέγχου η διεργασία;». Επίσης, κατά την ανάλυση Φάσης II δεν μας ενδιαφέρει το πώς έχουν εκτιμηθεί οι τιμές των παραμέτρων της διεργασίας ή αν ήταν από πριν γνωστές οι τιμές τους.

12.1.3 Τυπική μορφή διαγράμματος ελέγχου

Όπως αναφέραμε και προηγουμένως, ουσιαστικά, οι αλλαγές που θέλουμε να ανιχνεύσουμε αφορούν αλλαγές στην κατανομή ενός χαρακτηριστικού X (τ.μ.), οι τιμές του οποίου καθορίζουν την ποιότητα των προϊόντων που παράγονται. Γενικά, για το χαρακτηριστικό X υποθέτουμε ότι ακολουθεί κάποια κατανομή με συνάρτηση πιθανότητας ή συνάρτηση πυκνότητας πιθανότητας $f(x;\theta)$. Για να εφαρμόσουμε το διάγραμμα ελέγχου για την παρακολούθηση της διεργασίας, επιλέγουμε, σε διαφορετικές χρονικές στιγμές μεταξύ τους, τυχαία δείγματα μεγέθους $n \geqslant 1$ των τιμών της X, έστω $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{in})$, $i = 1, 2, \ldots$, και υπολογίζουμε την τιμή $W_i = g(\mathbf{X}_i)$ μιας κατάλληλης στατιστικής συνάρτησης η οποία εκτιμά την υπό παρακολούθηση παράμετρο (π.χ. μέση τιμή, διακύμανση). Αν σε κάθε δειγματοληψία έχουμε δείγμα μεγέθους $n \geqslant 2$, τότε λέμε ότι συλλέγουμε ορθολογικά δείγματα (rational subgroups), ενώ, αν $n = 1$, τότε λέμε ότι συλλέγουμε μεμονωμένες παρατηρήσεις (individual observations). Οπότε, τα αντίστοιχα διαγράμματα ελέγχου είναι διαγράμματα ελέγχου για ορθολογικά δείγματα και διαγράμματα ελέγχου για μεμονωμένες παρατηρήσεις. Για παράδειγμα, αν επιθυμούμε να παρακολουθήσουμε το μέσο επίπεδο της

διεργασίας και να ανιχνεύσουμε τυχόν αλλαγές σε αυτό, τότε από κάθε δείγμα μεγέθους n υπολογίζουμε τη δειγματική μέση τιμή $\bar{X}$ και οι διαδοχικές τιμές $W_1, W_2, \ldots$ απεικονίζονται στο διάγραμμα ελέγχου. Δηλαδή, σε αυτή την περίπτωση, $W_i = \bar{X}_i$, $i = 1, 2, \ldots$.

Παρατήρηση 12.1. Στο σημείο αυτό, αξίζει να αναφέρουμε πως στην περίπτωση των ορθολογικών δειγμάτων μεγέθους $n \geqslant 2$ η συλλογή των τιμών σε κάθε δείγμα μπορεί να γίνει με δύο τρόπους: (i) οι τιμές κάθε δείγματος θα πρέπει να προκύπτουν από προϊόντα που έχουν παραχθεί την ίδια χρονική στιγμή (αν κάτι τέτοιο είναι εφικτό) και (ii) οι τιμές κάθε δείγματος θα πρέπει να προκύπτουν από προϊόντα που έχουν επιλεχθεί τυχαία από τη συνολική παραγωγή των προϊόντων που έχουν παραχθεί καθ' όλη τη διάρκεια του διαστήματος δειγματοληψίας. Ουσιαστικά, τα ορθολογικά δείγματα πρέπει να συλλέγονται με τέτοιο τρόπο, ώστε, αν εμφανίζονται ειδικές αιτίες μεταβλητότητας σε μια διεργασία, οι διαφορές (μεταβλητότητα) να μεγιστοποιούνται μεταξύ των δειγμάτων (between samples), ενώ οι διαφορές μέσα στα δείγματα (within samples) να ελαχιστοποιούνται.

Ανάλογα με την κατανομή του X έχουμε, επίσης, δύο βασικές κατηγορίες διαγραμμάτων ελέγχου. Αν η κατανομή είναι συνεχής, τότε έχουμε διαγράμματα ελέγχου μεταβλητών (control charts for variables), ενώ, αν η κατανομή είναι διακριτή, έχουμε διαγράμματα ελέγχου ιδιοτήτων (control charts for attributes).

Παρακάτω δίνεται η συνήθης εικόνα ενός διαγράμματος ελέγχου (Σχήμα 12.1). Παρατηρήστε πως εκτός από τα σημεία που απεικονίζονται σε αυτό (τιμές $W_1, W_2, \ldots$), έχουν σχεδιαστεί και τρεις ακόμη οριζόντιες γραμμές, οι οποίες είναι τα όρια ελέγχου και η κεντρική γραμμή του διαγράμματος. Η κεντρική γραμμή (center line ή CL) του διαγράμματος συνήθως αντιστοιχεί στη μέση τιμή της κατανομής του X ή στη διάμεσο αυτής και υπολογίζεται όταν η διεργασία είναι εντός ελέγχου. Δηλαδή όταν η κατανομή που ακολουθεί η X είναι αυτή που περιγράφει τη συμπεριφορά των τιμών του χαρακτηριστικού, όταν η διεργασία λειτουργεί παρουσία μόνο φυσικών αιτιών μεταβλητότητας. Οι δύο άλλες οριζόντιες γραμμές είναι τα όρια ελέγχου (control limits), όπου το UCL (upper control limit) είναι το άνω όριο ελέγχου, ενώ το LCL (lower control limit) είναι το κάτω όριο ελέγχου. Επίσης, τα LCL, UCL υπολογίζονται όταν η διεργασία είναι εντός ελέγχου.

Για τον προσδιορισμό των τιμών της κεντρικής γραμμής και των ορίων ελέγχου μπορούν να χρησιμοποιηθούν ποσοστιαία σημεία της κατανομής της $W = g(\mathbf{X})$, όταν η διεργασία είναι εντός ελέγχου. Τα όρια αυτά είναι γνωστά και ως όρια ελέγχου πιθανότητας (probability limits). Επίσης, υπάρχει η δυνατότητα κατασκευής ορίων ελέγχου με βάση το λεγόμενο πρότυπο ορίων $A\sigma$ ($A > 0$) που είναι το:

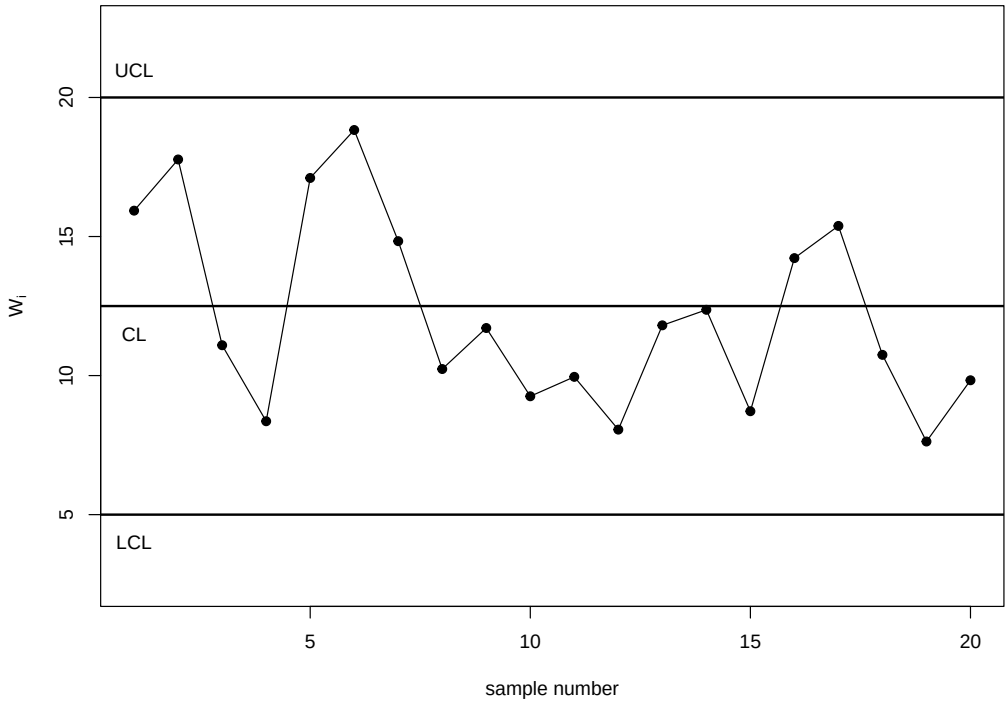
$$LCL = \mu_{0,W} - A\sigma_{0,W}, \quad CL = \mu_{0,W}, \quad UCL = \mu_{0,W} + A\sigma_{0,W},$$

όπου $\mu_{0,W}$ και $\sigma_{0,W}$ είναι η μέση τιμή και η τυπική απόκλιση της εντός ελέγχου κατανομής της W.

Εφόσον τα σημεία του διαγράμματος βρίσκονται μεταξύ των ορίων ελέγχου LCL, UCL η διεργασία θεωρείται ότι βρίσκεται εντός ελέγχου ή, αλλιώς, λέμε ότι «δεν έχουμε ένδειξη για εκτός ελέγχου διεργασία για το δοθέν επίπεδο σημαντικότητας». Αν, όμως, κάποιο σημείο βρεθεί εκτός των ορίων ελέγχου λέμε ότι «υπάρχει ένδειξη ότι η διεργασία είναι εκτός ελέγχου», οπότε αντιμετωπίζουμε κατάσταση συναγερμού και πρέπει να γίνει ειδική διερεύνηση αν πράγματι υπάρχουν ειδικές αιτίες μεταβλητότητας. Αν κριθεί ότι η διεργασία λειτουργεί παρουσία ειδικών αιτιών μεταβλητότητας, θα πρέπει αυτές να απομακρυνθούν και να γίνουν κατάλληλες διορθωτικές ενέργειες ώστε η διεργασία να έρθει εντός ελέγχου.

Πριν προχωρήσουμε, αξίζει να αναφέρουμε πως το διάγραμμα ελέγχου στο Σχήμα 12.1 είναι ένα δίπλευρο διάγραμμα ελέγχου, αφού χρησιμοποιούνται δύο όρια. Αν χρησιμοποιούνταν ένα μόνο όριο, τότε θα αναφερόμασταν σε μονόπλευρο διάγραμμα ελέγχου και ειδικότερα, σε άνω μονόπλευρο διάγραμμα ελέγχου, αν χρησιμοποιούσαμε μόνο το UCL, ή σε κάτω μονόπλευρο διάγραμμα ελέγχου, αν χρησιμοποιούσαμε μόνο το LCL.

Η πιο γνωστή κατηγορία διαγραμμάτων ελέγχου είναι τα διαγράμματα ελέγχου τύπου Shewhart. Σε ένα διαγράμμα ελέγχου Shewhart απεικονίζονται οι τιμές της στατιστικής συνάρτησης W_i, $i = 1, 2, \ldots$, και η



Σχήμα 12.1: Συνήθης μορφή διαγράμματος ελέγχου.

απόφαση λαμβάνεται με βάση την τιμή της W_i. Επειδή η απόφαση για το αν η διεργασία είναι εντός ή εκτός ελέγχου λαμβάνεται μόνο με βάση την τιμή του πιο πρόσφατου δείγματος, ένα διάγραμμα ελέγχου Shewhart είναι ένα διάγραμμα ελέγχου χωρίς μνήμη. Υπάρχουν και διαγράμματα ελέγχου με μνήμη, όπως είναι τα διαγράμματα ελέγχου τύπου EWMA (Exponentially Weighted Moving Average) και τα διαγράμματα ελέγχου τύπου CUSUM (Cumulative Sum). Για περισσότερες λεπτομέρειες παραπέμπουμε τον/την ενδιαφερόμενο/μενη αναγνώστη/στρια στο σύγγραμμα του Montgomery (2020). Στη συνέχεια, θα μας απασχολήσει μόνο η περίπτωση των διαγραμμάτων ελέγχου Shewhart.

12.1.4 Μέτρα απόδοσης ενός διάγραμματος ελέγχου

Η αξιολόγηση της απόδοσης ενός διαγράμματος ελέγχου γίνεται με χρήση της κατανομής του μήκους ροής (run length distribution). Ως μήκος ροής (Run Length) ορίζουμε την τ.μ. N, η οποία εκφράζει το πλήθος των σημείων, τα οποία απεικονίζονται στο διάγραμμα ελέγχου, μέχρις ότου το διάγραμμα ελέγχου να δώσει για πρώτη φορά ένδειξη εκτός ελέγχου διεργασίας. Για τα διαγράμματα ελέγχου τύπου Shewhart Φάσης ΙΙ, η κατανομή του μήκους ροής είναι η Γεωμετρική κατανομή με πιθανότητα επιτυχίας

$$p_{out} = P(W_i > UCL) + P(W_i < LCL).$$

Σε αυτήν την περίπτωση, η συνάρτηση πιθανότητας είναι

$$P(N = x) = p_{out}(1 - p_{out})^{x-1}\ x = 1, 2, \dots,$$

με μέση τιμή $\mathrm{E}(N) = 1/p_{out}$ και διασπορά $\mathrm{Var}(X) = (1 - p_{out})/p_{out}^2$. Η πιθανότητα p_{out} δεν είναι παρά η πιθανότητα να βρεθεί ένα σημείο εκτός των ορίων ελέγχου, δηλαδή είναι η πιθανότητα να δώσει το διάγραμμα ελέγχου ένδειξη εκτός ελέγχου διεργασίας.

Αν η διεργασία στην πραγματικότητα είναι εντός ελέγχου, η πιθανότητα p_{out} δεν είναι παρά πιθανότητα εσφαλμένου συναγερμού (false alarm probability) και, σε αναλογία με τον συμβολισμό που χρησιμοποιείται στους στατιστικούς ελέγχους υποθέσεων, συμβολίζεται με a. Αν η διεργασία είναι εκτός ελέγχου, η πιθανότητα p_{out} δεν είναι παρά η πιθανότητα συναγερμού (alarm probability), ενώ η πιθανότητα $1 - p_{out}$ είναι η πιθανότητα μη ένδειξης εκτός ελέγχου διεργασίας, όταν, όμως, στην πραγματικότητα η διεργασία είναι εκτός ελέγχου. Σε αυτήν την περίπτωση, σε αναλογία με τον συμβολισμό που χρησιμοποιείται στους στατιστικούς ελέγχους υποθέσεων, θα συμβολίζουμε ως $\beta = 1 - p_{out}$.

Η μέση τιμή $\mathrm{E}(N)$ είναι ο αναμενόμενος αριθμός των σημείων που απεικονίζονται στο διάγραμμα ελέγχου, μέχρις ότου το διάγραμμα ελέγχου να δώσει για πρώτη φορά ένδειξη εκτός ελέγχου διεργασίας και είναι γνωστή ως **μέσο μήκος ροής** (averge run length ή ARL). Το ARL αποτελεί το πιο γνωστό και συχνά χρησιμοποιούμενο μέτρο απόδοσης ενός διαγράμματος ελέγχου. Ανάλογα με το αν η διεργασία είναι εντός ή εκτός ελέγχου, υπολογίζεται και η αντίστοιχη τιμή για το ARL. Το εντός ελέγχου ARL (ή ARL_0) είναι $ARL_0 = 1/a$, ενώ το εκτός ελέγχου ARL (ή ARL_1) είναι $ARL_1 = 1/(1 - \beta)$. Για μια διεργασία η οποία βρίσκεται εντός ελέγχου, θέλουμε να έχουμε μεγάλη τιμή για το ARL_0 (ισοδύναμα, μικρή πιθανότητα εσφαλμένου συναγερμού), ώστε να μην έχουμε συχνές και, στην πραγματικότητα, άσκοπες διακοπές στην παρακολούθηση της διεργασίας. Για μια διεργασία, η οποία βρίσκεται εκτός στατιστικού ελέγχου, θέλουμε να έχουμε μικρή τιμή για το ARL_1, ώστε κατά μέσο όρο να μην χρειάζεται πολύς χρόνος (μεγάλος αριθμός σημείων/δειγμάτων) μέχρις ότου να ανιχνεύσουμε την παρουσία ειδικών αιτιών μεταβλητότητας στη διεργασία.

Συνήθως, ο προσδιορισμός των ορίων ελέγχου του διαγράμματος γίνεται για δεδομένη τιμή της πιθανότητας εσφαλμενου συναγερμού a ή για δεδομένη τιμή του εντός ελέγχου ARL. Η επιλογή του a επηρεάζει την πιθανότητα β ή, ισοδύναμα, τις τιμές του ARL_1. Καθώς μειώνεται το a (άρα αυξάνεται η τιμή ARL_0, οπότε και η πιθανότητα εσφαλμένου συναγερμού μειώνεται), αυξάνεται η πιθανότητα β. Αυτό σημαίνει πως αυξάνεται η πιθανότητα το διαγράμμα να μην δώσει ένδειξη εκτός ελέγχου διεργασίας (όταν στην πραγματικότητα η διεργασία είναι εκτός ελέγχου). Δηλαδή, αυξάνεται και ο αναμενόμενος αριθμός σημείων που απεικονίζονται σε αυτό μέχρις ότου να δώσει ένδειξη εκτός ελέγχου διεργασίας (μεγάλη τιμή ARL_1). Συνήθως, επιλέγεται μια τιμή για το a και με βάση αυτή προσδιορίζονται τα όρια ελέγχου και, στη συνέχεια, η απόδοση του διαγράμματος (με κριτήριο το ARL) για διάφορες περιπτώσεις εκτός ελέγχου διεργασίας (π.χ. για αλλαγές στο μέσο επίπεδο της διεργασίας, για αλλαγές στη μεταβλητότητα της διεργασίας και ούτω καθεξής).

Αξίζει, επίσης, να αναφέρουμε πως ο προσδιορισμός του ARL μας βοηθάει στο να υπολογίσουμε τις τιμές των ορίων ελέγχου του διαγράμματος (είτε με βάση το μοντέλο ορίων πιθανότητας είτε με βάση το μοντέλο ορίων $A\sigma$). Συνήθως, τα όρια ελέγχου προσδιορίζονται ώστε το διάγραμμα να έχει δεδομένη τιμή για το εντός ελέγχου ARL.

12.1.5 Μη παραμετρικά διαγράμματα ελέγχου

Βασική υπόθεση για την ανάπτυξη διαγραμμάτων ελέγχου τύπου Shewhart, είναι η υπόθεση της κανονικότητας των δεδομένων. Φυσικά, υπάρχουν και άλλα παραμετρικά μοντέλα τα οποία μπορούν να χρησιμοποιηθούν για τη μοντελοποίηση των διαθέσιμων μετρήσεων. Αν για τα δεδομένα (μετρήσεις/παρατηρήσεις), τα οποία λαμβάνονται από τη διεργασία, μπορούμε να θεωρήσουμε ότι προέρχονται από πληθυσμό που μοντελοποιείται σύμφωνα με το μοντέλο της κανονικής κατανομής (ή οποιοδήποτε άλλο πιθανοτικό πρότυπο), τότε το διάγραμμα ελέγχου που θα αναπτύξουμε είναι ένα παραμετρικό διάγραμμα ελέγχου.

Όμως, στην περίπτωση που παραβιάζεται η υπόθεση της κανονικότητας ή δεν μπορούμε να υποθέσουμε κανονικότητα για τις διαθέσιμες μετρήσεις ή δεν θέλουμε να κάνουμε κάποια υπόθεση για τη μορφή του πληθυσμού από τον οποίο συλλέγονται οι μετρήσεις, τότε θα πρέπει να χρησιμοποιήσουμε μη παραμετρικά

διαγράμματα ελέγχου (nonparametric control charts).

Αν και η χρήση των μη παραμετρικών διαγραμμάτων θα μπορούσε να είναι εκτεταμένη, υπάρχει ένα βασικό μειονέκτημα το οποίο περιορίζει τη δημοφιλία τους μεταξύ των χρηστών. Το μειονέκτημα αυτό είναι η δυσκολία εύρεσης αναλυτικών αποτελεσμάτων για τα χαρακτηριστικά του διαγράμματος (π.χ. ο υπολογισμός σε κλειστή μορφή των ορίων ελέγχου). Για τον λόγο αυτόν, συνήθως, χρησιμοποιείται προσομοίωση ώστε να μπορέσει κανείς να σχεδιάσει το διάγραμμα και να υπολογίσει την απόδοσή του. Όταν αναφερόμαστε στον σχεδιασμό του διαγράμματος θα εννοούμε τον προσδιορισμό των ορίων ελέγχου (αν είναι δίπλευρο διάγραμμα) ή τον προσδιορισμό ενός ορίου ελέγχου (αν είναι μονόπλευρο διάγραμμα), ώστε η απόδοση του διαγράμματος, όταν δεν υπάρχουν αλλαγές στη διεργασία, δηλαδή όταν η διεργασία είναι εντός ελέγχου, να είναι επιθυμητή. Η διαδικασία του προσδιορισμού ενός ορίου ελέγχου ή ενός ζεύγους ορίων ελέγχου είναι παρόμοια με τη διαδικασία εύρεσης της κρίσιμης περιοχής σε έναν στατιστικό έλεγχο υποθέσεων ώστε το σφάλμα τύπου I να είναι το επιθυμητό.

Στο σημείο αυτό, πρέπει να αναφέρουμε πως μια ιδιαίτερα χρήσιμη ιδιότητα ενός μη παραμετρικού διαγράμματος ελέγχου είναι ότι η εντός ελέγχου απόδοσή του δεν εξαρτάται από το ποια είναι η (συνεχής) κατανομή για το X, αλλά είναι η ίδια για όλες τις συνεχείς κατανομές. Για τον λόγο αυτόν, τέτοια μη παραμετρικά διαγράμματα ελέγχου είναι γνωστά και ως διαγράμματα ελέγχου απαλλαγμένα παραμέτρων (distribution-free control charts). Επίσης, ένας άλλος λόγος που τα μη παραμετρικά διαγράμματα ελέγχου δεν είναι ιδιαίτερα διαδεδομένα και σε ευρεία χρήση στις πρακτικές εφαρμογές είναι ότι οι κατανομές των στατιστικών συναρτήσεων που χρησιμοποιούνται για την ανάπτυξη αυτών δεν είναι (σε αρκετές περιπτώσεις) κάποιες από τις γνωστές, με αποτέλεσμα να απαιτείται η χρήση εξειδικευμένου λογισμικού (π.χ. της `R`).

Στη συνέχεια, παρουσιάζονται τα κυριότερα μη παραμετρικά διαγράμματα ελέγχου, τα οποία βασίζονται σε στατιστικές συναρτήσεις, οι οποίες έχουν χρησιμοποιηθεί για την ανάπτυξη μη παραμετρικών ελέγχων. Θα εστιάσουμε στο πρόβλημα της ανίχνευσης αλλαγών σε μια παράμετρο θέσης της (άγνωστης) κατανομής, η οποία περιγράφει τη διεργασία. Επίσης, θα δοθεί και ένα παράδειγμα για την ανίχνευση αλλαγών στη μεταβλητότητα. Στόχος δεν είναι η εκτενής παρουσίαση όλων των δυνατών μη παραμετρικών διαγραμμάτων ελέγχου αλλά η παρουσίαση βασικών τεχνικών με πρακτική σημασία. Πιο συγκεκριμένα, θα μας απασχολήσουν διαγράμματα ελέγχου τα οποία βασίζονται στο Προσημικό Κριτήριο (Sign Test, βλ. Κεφάλαιο 5), το Κριτήριο Προσημασμένων Τάξεων (Signed-Rank Test, βλ. Κεφάλαιο 6) και το Κριτήριο των Mann-Whitney (βλ., επίσης, Κεφάλαιο 6). Θα αναφερθούμε, επιπλέον, και σε ένα διάγραμμα ελέγχου το οποίο βασίζεται στο κριτήριο των Προηγήσεων (Precedence Test). Για περισσότερες λεπτομέρειες σχετικά με μεθόδους και τεχνικές του μη παραμετρικού στατιστικού ελέγχου διεργασιών, παραπέμπουμε στο σύγγραμμα των Chakraborti and Graham (2019).

12.2 Διαγράμματα ελέγχου με χρήση του προσημικού κριτηρίου

Στην παρούσα ενότητα, θα παρουσιάσουμε το διάγραμμα ελέγχου τύπου Shewhart SN-chart, το οποίο βασίζεται στον Προσημικό έλεγχο (sign test). Τα διαγράμματα αυτά έχουν μελετηθεί εκτενώς στη βιβλιογραφία και ο/η ενδιαφερόμενος/η αναγνώστης/στρια μπορεί να βρει περισσότερες λεπτομέρειες στην εργασία των Amin *et al.* (1995). Το συγκεκριμένο διάγραμμα είναι το απλούστερο μη παραμετρικό διάγραμμα ελέγχου, καθώς για την ορθή εφαρμογή του δεν χρειάζεται να υποθέσουμε κάποιο συγκεκριμένο παραμετρικό μοντέλο για την κατανομή των παρατηρήσεων. Αρχικά, θα δείξουμε πώς μπορεί να χρησιμοποιηθεί στην περίπτωση που θέλουμε να ανιχνεύσουμε αλλαγές στη διάμεσο της διεργασίας (ή διάμεσο του πληθυσμού). Εδώ, η διάμεσος θεωρείται ως ένα αντιπροσωπευτικό μέτρο θέσης της κατανομής των παρατηρήσεων οι οποίες συλλέγονται από τη διεργασία.

Συγκεκριμένα, έστω ότι συλλέγουμε σε κάθε χρονική στιγμή $i = 1, 2, \ldots$ τυχαίο δείγμα μεγέθους $n \geqslant 2$ και

συγκρίνουμε κάθε παρατήρηση (εντός του δείγματος) με την τιμή στόχο m_0, η οποία είναι η εντός ελέγχου τιμή της διαμέσου m. Από κάθε δείγμα, υπολογίζουμε την τιμή της σ.σ.:

$$SN_i = \sum_{j=1}^{n} sgn(X_{ij} - m_0),$$

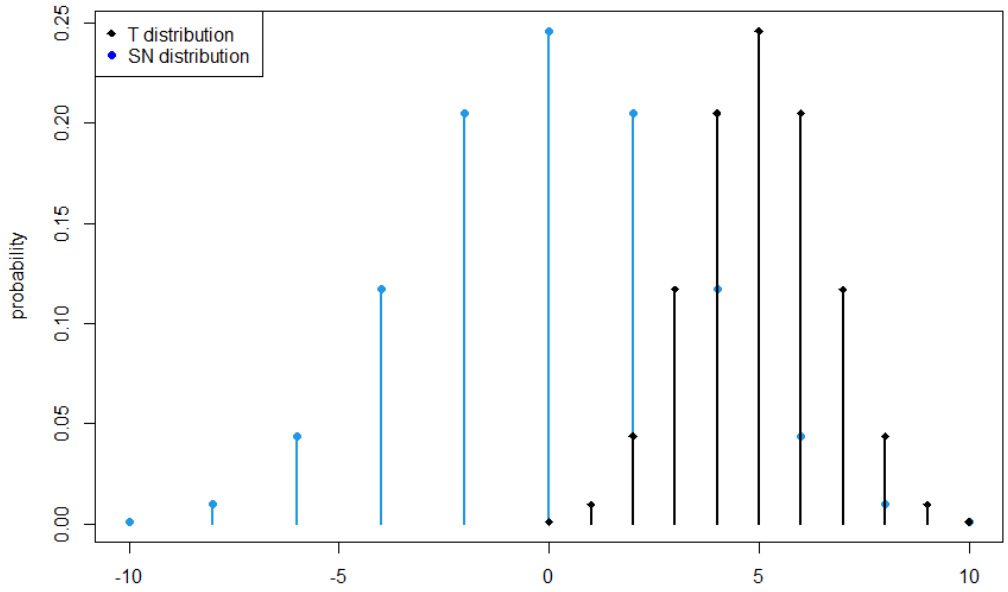
όπου $sgn(x)$ είναι η συνάρτηση προσήμου (βλ. π.χ. Κεφάλαιο 5). Δεν είναι δύσκολο να διαπιστώσουμε ότι η τιμή SN_i είναι ακριβώς η διαφορά μεταξύ του αριθμού των παρατηρήσεων που είναι μεγαλύτερες και μικρότερες από την τιμή m_0 στο i-οστό δείγμα. Αν η κατανομή του πληθυσμού είναι συνεχής, το ενδεχόμενο $\{(X_{ij} - m_0) = 0\}$ έχει μηδενική πιθανότητα να συμβεί, αφού η σύμπτωση της τιμής x_{ij} με την τιμή m_0 οφείλεται σε στρογγυλοποίηση της ακριβούς τιμής για τη X_{ij}. Άρα, στη περίπτωση που η κατανομή της X είναι συνεχής, θα υποθέσουμε ότι δεν είναι δυνατή η εμφάνιση ισοβαθμιών (δεσμών, ties) και η σ.σ. SN_i γράφεται ως

$$T_i = \frac{SN_i + n}{2},$$

ή $SN_i = 2T_i - n$, όπου η τυχαία μεταβλητή T_i ακολουθεί τη διωνυμική κατανομή $B(n,p)$, με $p = P(X_{ij} > m_0)$. Ουσιαστικά, η T_i καταμετρά το πλήθος των παρατηρήσεων στο i-οστό δείγμα οι οποίες είναι μεγαλύτερες του m_0. Κατά συνέπεια, η μέση τιμή και η διασπορά της SN_i δίνονται από τις σχέσεις

$$E(SN_i) = n(2p-1), \quad V(SN_i) = 4np(1-p).$$

Στην περίπτωση που η διεργασία είναι εντός στατιστικού ελέγχου (άρα $p = P(X_{ij} > m_0) = 0.5$), έπεται άμεσα ότι $E(SN_i) = 0$ και $V(SN_i) = n$. Δεν είναι δύσκολο να διαπιστώσουμε (βλ., επίσης, Σχήμα 12.2) ότι οι κατανομές των SN_i και T_i είναι συμμετρικές περί τη μέση τιμή τους. Παρατηρήστε, επίσης, ότι οι δυνατές τιμές της SN_i είναι στο σύνολο $\{-n, -n+2, \ldots, 0, \ldots, n-2, n\}$.



Σχήμα 12.2: Συνάρτηση πιθανότητας των τ.μ. T_i, SN_i για $n = 10$.

Λόγω αυτής της συμμετρίας, τα όρια ελέγχου και η κεντρική γραμμή για το διάγραμμα ελέγχου SN-chart είναι συμμετρικά περί την εντός ελέγχου μέση τιμή $E(SN_i) = 0$. Έτσι, αν $b_{SN} \in \{0,1,\ldots,n\}$, μπορούμε να

χρησιμοποιήσουμε ως όρια ελέγχου τα $UCL = b_{SN}$, $LCL = -b_{SN}$ και η κεντρική γραμμή να είναι ίση με $CL = 0$. Αφού έχουμε προσδιορίσει το πού θα βρίσκονται τα όρια ελέγχου, στη συνέχεια απεικονίζουμε τις διαδοχικές τιμές SN_i, $i = 1, 2, \ldots$. Αν οι τιμές της SN_i είναι στο διάστημα $(-b_{SN}, b_{SN})$, τότε η διεργασία είναι εντός στατιστικού ελέγχου, δηλαδή η διεργασία λειτουργεί χωρίς αλλαγές. Διαφορετικά, όταν για πρώτη φορά συμβεί $SN_i \geqslant b_{SN}$ ή $SN_i \leqslant -b_{SN}$, τότε έχουμε ένδειξη εκτός ελέγχου διεργασίας.

Ο προσδιορισμός της τιμής b_{SN} γίνεται με τέτοιο τρόπο ώστε το διάγραμμα να έχει την επιθυμητή εντός ελέγχου απόδοση. Για τον υπολογισμό της απόδοσης του διαγράμματος θα χρησιμοποιηθεί το μέσο μήκος ροής ARL. Η πιθανότητα να δώσει το διάγραμμα ελέγχου SN-chart ένδειξη εκτός ελέγχου διεργασίας στο i-οστό δείγμα ισούται με:

$$p_{out} = P(SN_i \geqslant b_{SN} \text{ ή } SN_i \leqslant -b_{SN}).$$

Κατά τον σχεδιασμό ενός διαγράμματος ελέγχου, η τιμή του a προεπιλέγεται και, στη συνέχεια, προσδιορίζονται οι τιμές των ορίων ελέγχου. Λόγω συμμετρίας της διωνυμικής κατανομής $B(n, 0.5)$, μπορούμε να επιλέξουμε την τιμή b_{SN} ως τον ελάχιστο ακέραιο για τον οποίο ισχύει $P(SN_i \geqslant b_{SN}) \leqslant a/2$ ή, ισοδύναμα, λαμβάνοντας υπόψη ότι $SN_i = 2T_i - n$ έχουμε

$$P(SN_i \geqslant b_{SN}) = P(2T_i - n \geqslant b_{SN}) = P\left(T_i \geqslant \frac{n + b_{SN}}{2}\right).$$

Για παράδειγμα, αν $n = 9$, τότε οι τιμές της α.σ.κ. της $B(9, 0.5)$ δίνονται (π.χ. με χρήση της εντολής `pbinom(0:9,9,0.5)` της `R`) στον Πίνακα 12.1.

x	$P(X \leqslant x)$
0	0.00195
1	0.01953
2	0.08984
3	0.25391
4	0.50000
5	0.74609
6	0.91016
7	0.98047
8	0.99805
9	1.00000

Πίνακας 12.1: Συνάρτηση πιθανότητας της $B(9, 0.5)$.

Αν επιλέξουμε $a = 0.0027$ (μια τιμή η οποία είναι αρκετά δημοφιλής στην περιοχή του στατιστικού ελέγχου διεργασιών) τότε, από τον Πίνακα 12.1, διαπιστώνουμε ότι η $P(T_i \geqslant 9) = 0.00195$. Η τιμή αυτή δεν είναι μικρότερη από το $a/2 = 0.00135$, όμως είναι η πιο κοντινή σε αυτή. Άρα, προκύπτει ότι $(9 + b_{SN})/2 = 9$, δηλαδή $b_{SN} = 9$. Οπότε, για το συγκεκριμένο διάγραμμα ελέγχου SN-chart, με όρια ελέγχου $UCL = 9$, $LCL = -9$, ενώ το εντός ελέγχου ARL είναι ίσο με 256, καθώς

$$\begin{aligned} ARL_0 &= \frac{1}{a} = \frac{1}{P(SN_i \geqslant b_{SN}) + P(SN_i \leqslant -b_{SN})} \\ &= \frac{1}{2 \cdot P(SN_i \geqslant 9)} = \frac{1}{2 \cdot 0.5^9} \\ &= \frac{1}{0.5^8} \approx 256. \end{aligned}$$

Παρατηρήστε ότι παραπάνω χρησιμοποιήθηκε η συμμετρία περί το μηδέν της SN_i, ενώ η πιθανότητα $P(SN_i \geqslant n) = P(T_i \geqslant n) = 0.5^n$, αφού, όταν η διεργασία είναι εντός στατιστικού ελέγχου, η σ.σ. T_i ακολουθεί τη διωνυμική κατανομή $B(n, 0.5)$.

Η τιμή $ARL_0 = 256$ σημαίνει ότι ο αναμενόμενος αριθμός των σημείων που απεικονίζονται στο διάγραμμα, μέχρι αυτό να δώσει για πρώτη φορά (εσφαλμένα) ένδειξη εκτός ελέγχου διεργασίας, είναι 256. Επίσης, πρέπει να αναφέρουμε πως, λόγω της διακριτής φύσης των δεδομένων, δεν είναι πάντοτε εφικτή η επιθυμητή εντός ελέγχου απόδοση. Δηλαδή δεν μπορούμε σε κάθε περίπτωση να προσδιορίσουμε τις τιμές των ορίων ελέγχου ώστε η τιμή για το εντός ελέγχου ARL να είναι ίση με την επιθυμητή. Κάτι αντίστοιχο συμβαίνει κατά τον υπολογισμό της ακριβούς πιθανότητας Σφάλματος Τύπου I του Προσημικού κριτηρίου.

Ενδεικτικά, παρουσιάζουμε τον Πίνακα 12.2, όπου για διάφορες τιμές του n δίνουμε τις τιμές των ορίων ελέγχου (τιμή b_{SN}), την πραγματική πιθανότητα εσφαλμένου συναγερμού a' και την αντίστοιχη τιμή για το εντός ελέγχου ARL. Ο υπολογισμός των τιμών b_{SN} αλλά και της ακριβούς πιθανότητας εσφαλμένου συναγερμού a', όπως και του ARL, γίνεται με τον τρόπο που περιγράψαμε παραπάνω (βλ., επίσης Chakraborti and Graham, 2014). Για το a έχουμε επιλέξει -ενδεικτικά πάλι- την τιμή 0.0027, δηλαδή $a/2 = 0.00135$. Αξίζει να αναφέρουμε πως η δημοφιλία της συγκεκριμένης τιμής του a οφείλεται στο ότι σχετίζεται με την πιθανότητα εσφαλμένου συναγερμού σε ένα δίπλευρο διάγραμμα ελέγχου Φάσης II $\bar{X}$, με όρια ελέγχου τα οποία υπολογίζονται από μοντέλο ορίων $A\sigma$ με $A = 3$, και για την περίπτωση που η κατανομή του χαρακτηριστικού X είναι η κανονική. Παρατηρήστε πως χρειαζόμαστε αρκετά μεγάλο μέγεθος δείγματος ώστε να έχουμε εντός ελέγχου απόδοση κοντά στην επιθυμητή. Στη συνέχεια, δίνουμε ένα παράδειγμα εφαρμογής του διαγράμματος ελέγχου SN-chart.

n	b_{SN}	a'	ARL
5	5	0.0625	16
6	6	0.03125	32
7	7	0.015625	64
8	8	0.0078125	128
9	9	0.00390625	256
10	10	0.001953125	512
15	13	0.0009765625	1024
20	14	0.002576828	388.10

Πίνακας 12.2: Τιμές πιθανότητας εσφαλμένου συναγερμού και εντός ελέγχου ARL για διάφορα μεγέθη δείγματος n.

Παράδειγμα 12.1. Στον Πίνακα 12.3 δίνονται, με ακρίβεια δύο δεκαδικών ψηφίων, τα δεδομένα από 40 δείγματα μεγέθους $n = 5$ παρατηρήσεων από άγνωστο πληθυσμό με διάμεσο $m = m_0 = 3$. Θέλουμε να παρακολουθήσουμε τη διεργασία και να ανιχνεύσουμε τυχόν αλλαγές στη διάμεσο του πληθυσμού χρησιμοποιώντας ένα διάγραμμα ελέγχου SN-chart.

Λύση Παραδείγματος 12.1. Αφού το $n = 5$, από τον Πίνακα 12.2 προκύπτει ότι $b_{SN} = 5$. Τότε, η πιθανότητα εσφαλμένου συναγερμού είναι ίση με 0.0625, ενώ το εντός ελέγχου ARL ισούται με 16.

Από τα δεδομένα του προβλήματος έχουμε ότι η διάμεσος είναι $m_0 = 3$. Άρα, πρέπει αρχικά να καταγράψουμε το πλήθος των παρατηρήσεων από κάθε δείγμα που είναι μεγαλύτερες του 3, δηλαδή τις τιμές T_i, και στη συνέχεια τις τιμές $SN_i = 2 \cdot T_i - n$. Δεν είναι δύσκολο να διαπιστώσουμε ότι οι τιμές αυτές είναι αυτές οι οποίες παρατίθενται στον Πίνακα 12.4.

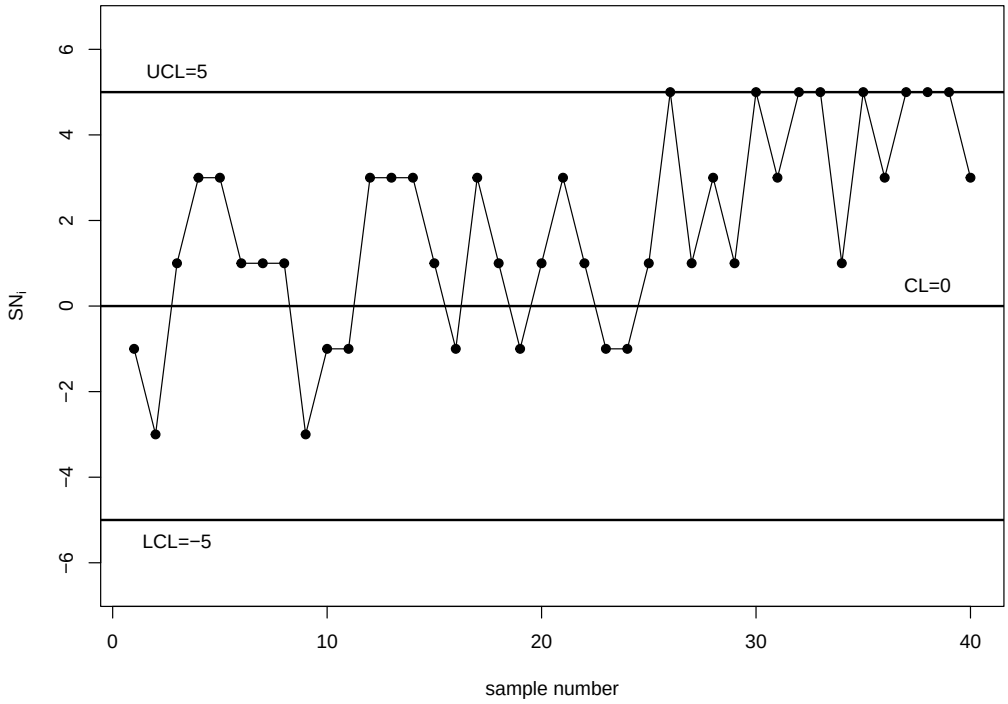
Αφού έχουμε επιλέξει τις τιμές των ορίων ελέγχου (η κεντρική γραμμή θα είναι ίση με 0) και γνωρίζουμε τις τιμές της σ.σ. που πρέπει να απεικονίσουμε στο διάγραμμα, μπορούμε να κατασκευάσουμε το ζητούμενο διάγραμμα ελέγχου SN-chart. Η εικόνα του διαγράμματος δίνεται στο Σχήμα 12.3.

A/A	X_1	X_2	X_3	X_4	X_5	A/A	X_1	X_2	X_3	X_4	X_5
1	1.02	3.86	3.51	1.56	0.85	21	3.01	1.85	4.03	4.25	6.27
2	2.88	3.42	2.42	1.33	1.26	22	1.90	4.13	4.59	0.80	13.62
3	3.95	1.73	5.19	2.38	3.94	23	1.99	9.24	2.61	2.10	4.93
4	7.45	5.74	5.24	2.85	9.51	24	3.82	1.34	2.24	7.20	2.26
5	5.35	3.93	5.51	2.81	4.51	25	3.14	9.61	0.47	7.62	1.11
6	5.23	1.73	2.48	4.08	5.45	26	8.79	7.68	4.91	9.16	5.49
7	2.95	1.62	4.04	3.17	5.20	27	6.27	1.50	10.74	10.36	1.74
8	5.11	8.00	1.18	8.36	2.19	28	12.41	14.19	6.19	1.56	4.62
9	0.51	3.56	2.75	2.82	2.39	29	5.91	13.61	2.67	1.10	7.95
10	0.97	4.31	6.14	1.33	1.75	30	4.54	4.41	12.93	9.22	5.95
11	1.81	3.18	2.19	5.46	1.62	31	8.86	1.76	18.39	6.57	9.32
12	5.98	1.97	5.55	4.13	3.51	32	11.32	7.86	11.22	4.75	11.89
13	5.85	2.53	4.67	3.94	4.13	33	5.37	11.56	7.68	5.19	4.18
14	3.12	3.37	3.24	2.48	3.41	34	3.23	2.43	2.42	6.32	6.56
15	9.05	2.48	9.24	2.03	4.68	35	11.60	4.01	3.66	8.44	7.35
16	7.88	1.55	2.30	2.82	4.25	36	3.69	4.05	1.17	6.01	5.11
17	2.45	3.62	8.26	6.63	3.25	37	4.18	3.20	8.87	5.82	7.80
18	4.53	2.85	1.22	3.37	3.82	38	4.53	4.42	17.29	4.77	7.34
19	2.97	1.44	4.47	1.59	4.50	39	6.04	11.39	8.14	3.30	13.62
20	2.08	1.83	5.23	3.69	4.46	40	3.94	1.93	6.39	6.83	5.29

Πίνακας 12.3: Δεδομένα εφαρμογής διαγράμματος SN-chart.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
T_i	2	1	3	4	4	3	3	3	1	2	2	4	4	4	3	2	4	3	2	3
SN_i	-1	-3	1	3	3	1	1	1	-3	-1	-1	3	3	3	1	-1	3	1	-1	1
i	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
T_i	4	3	2	2	3	5	3	4	3	5	4	5	5	3	5	4	5	5	5	4
SN_i	3	1	-1	-1	1	5	1	3	1	5	3	5	5	1	5	3	5	5	5	3

Πίνακας 12.4: Τιμές των σ.σ. T_i και SN_i για τα δεδομένα του Πίνακα 12.3.



Σχήμα 12.3: Διάγραμμα ελέγχου SN-chart για τα δεδομένα του Πίνακα 12.3.

Από την εικόνα του διαγράμματος συμπεραίνουμε ότι το διάγραμμα ελέγχου δίνει για πρώτη φορά ένδειξη εκτός ελέγχου διεργασίας στο δείγμα 26, αφού η τιμή του είναι ίση με 5 και «πέφτει» πάνω στο άνω όριο ελέγχου. Θα πρέπει να γίνει ειδική διερεύνηση αν πράγματι η διεργασία λειτουργεί υπό την παρουσία ειδικών αιτιών μεταβλητότητας (δηλαδή είναι πράγματι εκτός ελέγχου) ή αν η συγκεκριμένη ένδειξη είναι ψευδής (οπότε και η διεργασία είναι στην πραγματικότητα εντός ελέγχου).

Προτού ολοκληρώσουμε τη λύση του συγκεκριμένου παραδείγματος, κρίνουμε σκόπιμο για λόγους πληρότητας να δώσουμε και τις εντολές στην R οι οποίες υπολογίζουν τις τιμές της σ.σ. που απεικονίζεται στο SN-chart, καθώς και αυτές με τις οποίες κατασκευάζεται το διάγραμμα ελέγχου στο Σχήμα 12.3. □

```
> m0<-3 # population median
> # data entry
> data<-matrix(c(1.02,2.88,3.95,7.45,5.35,5.23,2.95,5.11,0.51,
+ 0.97,1.81,5.98,5.85,3.12,9.05,7.88,2.45,4.53,3.00,2.08,3.01,
+ 1.90,1.99,3.82,3.14,8.79,6.27,12.41,5.91,4.54,8.86,11.32,
+ 5.37,3.23,11.60,3.69,4.18,4.53,6.04,3.94,3.86,3.42,1.73,
+ 5.74,3.93,1.73,1.62,8.00,3.56,4.31,3.18,1.97,2.53,3.37,
+ 2.48,1.55,3.62,2.85,1.44,1.83,1.85,4.13,9.24,1.34,9.61,
+ 7.68,1.50,14.19,13.61,4.41,1.76,7.86,11.56,2.43,4.01,4.05,
+ 3.20,4.42,11.39,1.93,3.51,2.42,5.19,5.24,5.51,2.48,4.04,
+ 1.18,2.75,6.14,2.19,5.55,4.67,3.24,9.24,2.30,8.26,1.22,
+ 4.47,5.23,4.03,4.59,2.61,2.24,0.47,4.91,10.74,6.19,2.67,
+ 12.93,18.39,11.22,7.68,2.42,3.66,1.17,8.87,17.29,8.14,
+ 6.39,1.56,1.33,2.38,2.85,2.81,4.08,3.17,8.36,2.82,1.33,
+ 5.46,4.13,3.94,2.48,2.03,2.82,6.63,3.37,1.59,3.69,4.25,
+ 0.80,2.10,7.20,7.62,9.16,10.36,1.56,1.10,9.22,6.57,4.75,
+ 5.19,6.32,8.44,6.01,5.82,4.77,3.30,6.83,0.85,1.26,3.94,
+ 9.51,4.51,5.45,5.20,2.19,2.39,1.75,1.62,3.51,4.13,3.41,
```

```
+ 4.68,4.25,3.25,3.82,4.50,4.46,6.27,13.62,4.93,2.26,1.11,
+ 5.49,1.74,4.62,7.95,5.95,9.32,11.89,4.18,6.56,7.35,5.11,
+ 7.80,7.34,13.62,5.29),ncol=5,byrow=FALSE)
> vec1<-apply(data>m0,1,sum) # Ti values
> SN1<-2*vec1-5 # SNi values
> # construction of SN-chart
> plot(1:40,SN1,type='n',xlab='sample number',ylab=expression(SN[i]),
+ ylim=c(-6.5,6.5))
> points(1:40,SN1,pch=19)
> lines(1:40,SN1,lty=1,lwd=1)
> abline(h=5,lwd=2)
> abline(h=-5,lwd=2)
> abline(h=0,lwd=2)
> text(3,5.5,'UCL=5')
> text(3,-5.5,'LCL=-5')
> text(38,0.5,'CL=0')
```

12.3 Διαγράμματα ελέγχου με χρήση του κριτηρίου προσημασμένων τάξεων

Στη παρούσα ενότητα, θα παρουσιάσουμε ένα άλλο μη παραμετρικό διάγραμμα ελέγχου, κατάλληλο για την ανίχνευση αλλαγών στη διάμεσο ενός συνεχούς και συμμετρικού πληθυσμού. Είναι γνωστό ότι το κριτήριο του Wilcoxon (Wilcoxon's Signed-Rank test ή SR, βλ. Κεφάλαιο 6) είναι κατάλληλο για τον έλεγχο υπόθεσης σχετικά με τη διάμεσο ενός πληθυσμού, ο οποίος μοντελοποιείται σύμφωνα με μια συνεχή και συμμετρική κατανομή. Αν και το προσημικό κριτήριο μπορεί να εφαρμοστεί για οποιαδήποτε συνεχή κατανομή (δεν είναι απαραίτητη η υπόθεση της συμμετρίας), στην περίπτωση που η κατανομή είναι συμμετρική, τότε το τεστ του Wilcoxon είναι πιο ισχυρό. Από την άλλη, το προσημικό κριτήριο είναι κατάλληλο για τον έλεγχο της υπόθεσης οποιουδήποτε ποσοστιαίου σημείου, ενώ το τεστ του Wilcoxon χρησιμοποιείται για τον έλεγχο της διαμέσου. Ο Bakir (2004) πρότεινε και μελέτησε μη παραμετρικά διαγράμματα ελέγχου τύπου Shewhart, τα οποία βασίζονται στη στατιστική συνάρτηση SR για τον έλεγχο του Wilcoxon. Το διάγραμμα Shewhart, το οποίο βασίζεται σε αυτήν τη στατιστική συνάρτηση, θα αναφέρεται ως SR-chart.

Συγκεκριμένα, έστω ότι σε κάθε χρονική στιγμή $i = 1,2,\ldots$ συλλέγονται δείγματα μεγέθους n από πληθυσμό με άγνωστη κατανομή και διάμεσο m_0. Έστω ότι το i-οστό δείγμα είναι το $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{in})$ και έστω, επίσης, ότι $|X_{i1} - m_0|$, $|X_{i2} - m_0|$, ..., $|X_{in} - m_0|$ είναι οι απόλυτες τιμές των αποκλίσεων των τιμών του i-οστού δείγματος από τη διάμεσο m_0 του πληθυσμού. Έστω, επιπλέον, ότι R_{ij} είναι η τάξη (rank) της παρατήρησης $|X_{ij} - m_0|$ μεταξύ των n απόλυτων αποκλίσεων. Στο διάγραμμα ελέγχου απεικονίζονται οι τιμές της στατιστικής συνάρτησης

$$SR_i = \sum_{j=1}^{n} sgn(X_{ij} - m_0)R_{ij},$$

για $i = 1,2,\ldots$. Άμεσα μπορούμε να διαπιστώσουμε ότι η στατιστική συνάρτηση SR_i αποτελεί τη διαφορά μεταξύ του αθροίσματος των τάξεων (για τις απόλυτες τιμές των αποκλίσεων) που αντιστοιχούν σε θετικές και σε αρνητικές αποκλίσεις, αντίστοιχα. Η στατιστική συνάρτηση SR γράφεται, ισοδύναμα, ως

$$SR_i = 2W_i^+ - \frac{n(n+1)}{2}, \tag{12.1}$$

όπου W_i^+ είναι το άθροισμα των τάξεων, όπως αυτές υπολογίστηκαν, των θετικών διαφορών των τιμών του i-οστού δείγματος από τη διάμεσο m_0.

Όταν η διεργασία είναι εντός στατιστικού ελέγχου, τότε η κατανομή της σ.σ. SR_i είναι συμμετρική περί το μηδέν. Άρα, τα όρια ελέγχου μπορούν να επιλεχθούν, ώστε $LCL = -UCL$. Αν b_{SR} είναι ένας θετικός ακέραιος,

τότε ως όρια ελέγχου επιλέγουμε τα $LCL = -b_{SR}$, $UCL = b_{SR}$, ενώ η κεντρική γραμμή είναι ίση με το μηδέν. Στη συνέχεια, στο διάγραμμα SR-chart απεικονίζονται τιμές της σ.σ. SR_i και, όταν για πρώτη φορά το $SR_i \geqslant b_{SR}$ ή $SR_i \leqslant -b_{SR}$ (οτιδήποτε από τα δύο συμβεί πρώτο), το διάγραμμα δίνει ένδειξη εκτός ελέγχου διεργασίας στο i-οστό δείγμα.

Η τιμή b_{SR} προσδιορίζεται για δεδομένη πιθανότητα εσφαλμένου συναγερμού a και είναι ο μικρότερος ακέραιος για τον οποίο ισχύει $P(SR_i \geqslant b_{SR}) \leqslant a/2$ ή, ισοδύναμα, λαμβάνοντας υπόψη τη σχέση (12.1)

$$P(SR_i \geqslant b_{SR}) = P\left(2W_i^+ - \frac{n(n+1)}{2} \geqslant b_{SR}\right) = P\left(W_i^+ \geqslant \frac{2b_{SR} + n(n+1)}{4}\right) \leqslant a/2.$$

Με χρήση της εντολής `psignrank(x,n,...)` της `R` μπορούμε να υπολογίσουμε τιμές της α.σ.κ. της τ.μ. W_i^+ για διάφορα μεγέθη δείγματος n. Υπενθυμίζεται ότι οι δυνατές τιμές της W_i^+ είναι στο σύνολο $\{0, 1, \ldots, n(n+1)/2\}$. Στον Πίνακα 12.5 δίνεται η αθροιστική συνάρτηση κατανομής της W_i^+ για $n = 7$, η οποία, καθώς $n(n+1)/2 = 7 \cdot 8/2 = 28$, προκύπτει άμεσα με την εντολή `psignrank(0:28,7)`.

x	$P(W^+ \leqslant x)$	x	$P(W^+ \leqslant x)$	x	$P(W^+ \leqslant x)$	x	$P(W^+ \leqslant x)$
0	0.0078125	8	0.1875000	16	0.6562500	24	0.9609375
1	0.0156250	9	0.2343750	17	0.7109375	25	0.9765625
2	0.0234375	10	0.2890625	18	0.7656250	26	0.9843750
3	0.0390625	11	0.3437500	19	0.8125000	27	0.9921875
4	0.0546875	12	0.4062500	20	0.8515625	28	1.0000000
5	0.0781250	13	0.4687500	21	0.8906250		
6	0.1093750	14	0.5312500	22	0.9218750		
7	0.1484375	15	0.5937500	23	0.9453125		

Πίνακας 12.5: Συνάρτηση κατανομής της τ.μ. W_i^+ για $n = 7$.

Αν επιλέξουμε $a = 0.0027$, τότε, από τον Πίνακα 12.5, διαπιστώνουμε ότι

$$P(W_i^+ \geqslant 28) = 0.0078125 \not\leqslant 0.00135.$$

Δηλαδή, δεν μπορούμε να βρούμε κατάλληλη τιμή για το ζητούμενο όριο ελέγχου ώστε να ικανοποιείται η απαίτηση για το δεδομένο επίπεδο σημαντικότητας. Παρά ταύτα, η συγκεκριμένη τιμή είναι η πιο κοντινή στην τιμή $a/2$, οπότε και θα χρησιμοποιήσουμε την αντίστοιχη τιμή της κατανομής της W_i^+, για να προσδιορίσουμε τα όρια ελέγχου. Άρα, έπεται ότι

$$\frac{2b_{SR} + n(n+1)}{4} = 28,$$

δηλαδή $b_{SR} = 28$. Επιπρόσθετα, για το συγκεκριμένο διάγραμμα ελέγχου SR-chart, με όρια ελέγχου $UCL = 28$, $LCL = -28$, το εντός ελέγχου ARL είναι ίσο με 64, αφού

$$ARL = \frac{1}{a} = \frac{1}{P(SR_i \geqslant b_{SR}) + P(SR_i \leqslant -b_{SR})} = \frac{1}{2 \cdot P(SR_i \geqslant b_{SR})} = \frac{1}{2 \cdot 0.0078125} = 64.$$

Αυτό σημαίνει ότι ο αναμενόμενος αριθμός σημείων που απεικονίζονται στο διάγραμμα, μέχρις ότου αυτό να δώσει για πρώτη φορά ένδειξη εκτός ελέγχου διεργασίας, είναι 64. Επίσης, πρέπει να αναφέρουμε πως, όπως και στην περίπτωση του διαγράμματος ελέγχου SN-chart, λόγω του ότι η κατανομή της W_i^+ είναι διακριτή, δεν είναι πάντοτε εφικτός ο προσδιορισμός των τιμών των ορίων ελέγχου ώστε η εντός ελέγχου απόδοση του διαγράμματος (τιμή εντός ελέγχου ARL) να είναι ακριβώς ίση με την επιθυμητή. Στο παραπάνω παράδειγμα, αν και θέλαμε $a = 0.0027$, για $n = 7$ η ακριβής πιθανότητα εσφαλμένου συναγερμού είναι $a = 2 \cdot 0.0078125 = 0.015625$, η οποία φυσικά είναι αρκετά μεγαλύτερη της επιθυμητής. Παρατηρήστε,

επίσης, πως αυτή είναι η μικρότερη δυνατή τιμή που μπορούμε να επιτύχουμε για το συγκεκριμένο μέγεθος δείγματος. Είναι, λοιπόν, εμφανές ότι για μικρά μεγέθη δείγματος, υπάρχουν σημαντικοί περιορισμοί στον προσδιορισμό των ορίων ελέγχου ενός διαγράμματος SR-chart (όπως και ενός SN-chart), ώστε η πιθανότητα εσφαλμένου συναγερμού να είναι η επιθυμητή.

Αξίζει, επίσης, να αναφέρουμε πως η πιθανότητα $P(W^+ \leqslant 0)$ αντιστοιχεί στην πιθανότητα εμφάνισης μιας επτάδας, δηλαδή δείγματος μεγέθους $n = 7$, με όλες τις διαφορές $X_{ij} - m_0$, $j = 1, 2, \dots, 7$, να είναι αρνητικές. Σε αυτήν την περίπτωση, η τιμή της W_i^+ είναι ίση με 0 και η πιθανότητα να συμβεί αυτό είναι $1/2^7 = 0.0078125$, όπου 2^7 είναι το σύνολο των επτάδων με πιθανά αποτελέσματα σε κάθε συντεταγμένη $+$ (για θετικές διαφορές) ή $-$ (για αρνητικές διαφορές). Λόγω συμμετρίας της κατανομής της W_i, ισχύει ότι $P(W_i^+ \leqslant 0) = P(W_i^+ \geqslant 28)$ και με αυτόν τον τρόπο μπορεί να υπολογιστεί η πιθανότητα εσφαλμένου συναγερμού. Φυσικά, ο ίδιος συλλογισμός μπορεί να ακολουθηθεί για τον προσδιορισμό της πιθανότητας εσφαλμένου συναγερμού για οποιαδήποτε n και b_{SR}.

Ενδεικτικά, παρουσιάζουμε τον Πίνακα 12.6 (βλ., επίσης Chakraborti and Graham, 2014), όπου για διάφορες τιμές του n δίνουμε τις τιμές των ορίων ελέγχου, την πραγματική πιθανότητα εσφαλμένου συναγερμού και την αντίστοιχη τιμή για το εντός ελέγχου ARL. Ο υπολογισμός των τιμών b_{SR}, της ακριβούς πιθανότητας εσφαλμένου συναγερμού a', καθώς και του ARL γίνεται με τον τρόπο που περιγράψαμε παραπάνω. Για το a έχουμε επιλέξει ενδεικτικά την τιμή 0.0027. Παρατηρούμε πως για να πετύχουμε μια τιμή κοντά στην επιθυμητή πιθανότητα εσφαλμένου συναγερμού, θα πρέπει να αυξήσουμε σημαντικά το μέγεθος δείγματος (ενδεικτικά για $n = 20$ ή μεγαλύτερο), το οποίο, όμως, σε πρακτικές εφαρμογές δεν είναι πάντοτε εφικτό.

n	b_{SR}	a'	ARL
5	15	0.0625	16
6	21	0.03125	32
7	28	0.015625	64
8	36	0.007813	128
9	45	0.003906	256
10	55	0.001953	512
15	100	0.002625	381.02
20	156	0.002325	430.10

Πίνακας 12.6: Τιμές πιθανότητας εσφαλμένου συναγερμού και εντός ελέγχου ARL για διάφορα μεγέθη δείγματος n.

Ακολούθως, δίνουμε ένα παράδειγμα εφαρμογής του διαγράμματος ελέγχου SR-chart.

Παράδειγμα 12.2. Χρησιμοποιώντας τα δεδομένα του Παραδείγματος 12.1, να κατασκευαστεί το διάγραμμα ελέγχου SR-chart.

Λύση Παραδείγματος 12.2. Αφού το $n = 5$, από τον Πίνακα 12.6 θα χρησιμοποιήσουμε $b_{SR} = 15$. Τότε, η πιθανότητα εσφαλμένου συναγερμού είναι ίση με 0.0625, ενώ το εντός ελέγχου ARL ισούται με 16.

Αφού η διάμεσος είναι $m_0 = 3$, βρίσκουμε τις απόλυτες τιμές των αποκλίσεων $|X_{ij} - m_0|$, $i = 1, 2, \dots, 40$ και, έπειτα, υπολογίζουμε τις τάξεις τους. Κατόπιν, βρίσκουμε το πρόσημο των διαφορών $X_{ij} - m_0$ και υπολογίζουμε για το i-οστό δείγμα την τιμή της στατιστικής συνάρτησης R_i. Ενδεικτικά, για το 1ο δείγμα, οι απαιτούμενες πράξεις παρατίθενται στον Πίνακα 12.7.

Άρα, η τιμή SR_1 είναι ίση με

$$SR_1 = \sum_{j=1}^{n} sgn(X_{1j} - m_0) R_{1j} = 4 \cdot (-1) + 2 \cdot 1 + 1 \cdot 1 + 3 \cdot (-1) + 5 \cdot (-1) = -9,$$

j	1	2	3	4	5
X_{1j}	1.02	3.86	3.51	1.56	0.85
$X_{1j} - m_0$	-1.98	0.86	0.51	-1.44	-2.15
$sgn(X_{1j} - m_0)$	-1	1	1	-1	-1
$\|X_{1j} - m_0\|$	1.98	0.86	0.51	1.44	2.15
R_{1j}	4	2	1	3	5

Πίνακας 12.7: Υπολογισμός τιμής SR_1 για τα δεδομένα του Πίνακα 12.3.

ενώ με τον ίδιο τρόπο υπολογίζονται και οι υπόλοιπες τιμές SR_i, για $i = 1,...,40$, και οι οποίες δίνονται στον Πίνακα 12.8.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
SR_i	-9	-11	5	13	13	7	5	9	-9	-1	-3	11	13	5	9	1	11	3	-1	5
i	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
SR_i	9	5	3	1	5	15	9	13	9	15	13	15	15	5	15	9	15	15	15	11

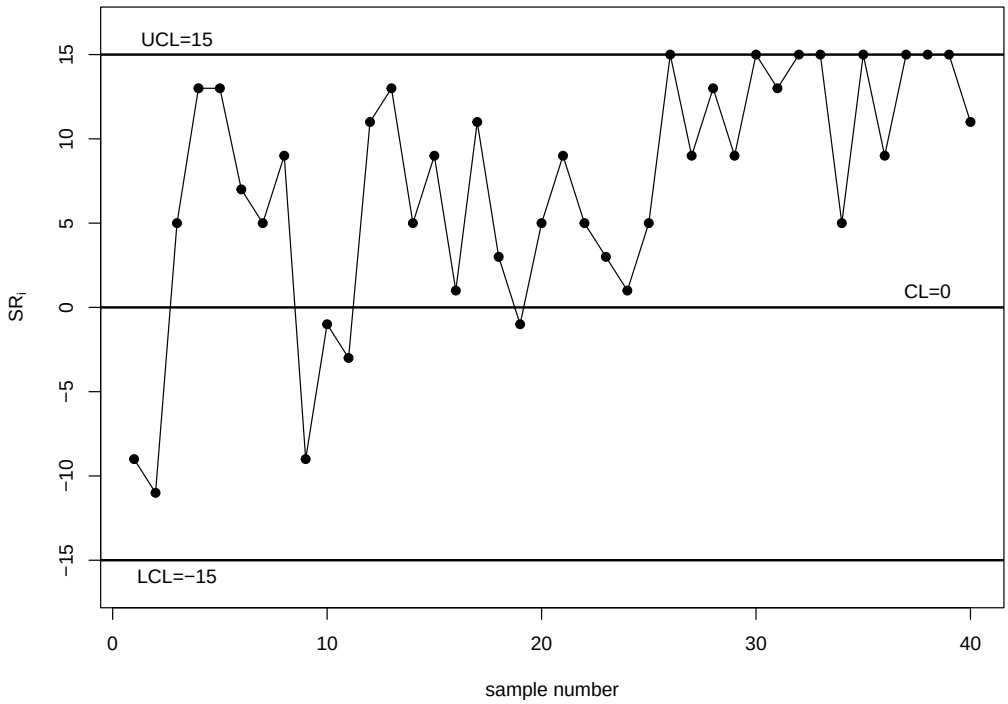
Πίνακας 12.8: Τιμές της σ.σ. SR_i για τα δεδομένα του Πίνακα 12.3.

Αφού έχουμε επιλέξει τις τιμές των ορίων ελέγχου (η κεντρική γραμμή θα είναι ίση με 0) και γνωρίζουμε τις τιμές της σ.σ. που πρέπει να απεικονίσουμε στο διάγραμμα, μπορούμε να κατασκευάσουμε το ζητούμενο διάγραμμα ελέγχου SR-chart. Η εικόνα του διαγράμματος δίνεται στο Σχήμα 12.4.

Προτού ολοκληρώσουμε τη λύση του συγκεκριμένου παραδείγματος, κρίνουμε σκόπιμο για λόγους πληρότητας να δώσουμε και τις εντολές στην R, οι οποίες υπολογίζουν τις τιμές της σ.σ. που απεικονίζεται στο SR-chart, καθώς και αυτές με τις οποίες κατασκευάζεται το διάγραμμα ελέγχου στο Σχήμα 12.4.

```
> m0<-3 # population median
> # data entry
> data<-matrix(c(1.02,2.88,3.95,7.45,5.35,5.23,2.95,5.11,0.51,
+ 0.97,1.81,5.98,5.85,3.12,9.05,7.88,2.45,4.53,3.00,2.08,3.01,
+ 1.90,1.99,3.82,3.14,8.79,6.27,12.41,5.91,4.54,8.86,11.32,
+ 5.37,3.23,11.60,3.69,4.18,4.53,6.04,3.94,3.86,3.42,1.73,
+ 5.74,3.93,1.73,1.62,8.00,3.56,4.31,3.18,1.97,2.53,3.37,
+ 2.48,1.55,3.62,2.85,1.44,1.83,1.85,4.13,9.24,1.34,9.61,
+ 7.68,1.50,14.19,13.61,4.41,1.76,7.86,11.56,2.43,4.01,4.05,
+ 3.20,4.42,11.39,1.93,3.51,2.42,5.19,5.24,5.51,2.48,4.04,
+ 1.18,2.75,6.14,2.19,5.55,4.67,3.24,9.24,2.30,8.26,1.22,
+ 4.47,5.23,4.03,4.59,2.61,2.24,0.47,4.91,10.74,6.19,2.67,
+ 12.93,18.39,11.22,7.68,2.42,3.66,1.17,8.87,17.29,8.14,
+ 6.39,1.56,1.33,2.38,2.85,2.81,4.08,3.17,8.36,2.82,1.33,
+ 5.46,4.13,3.94,2.48,2.03,2.82,6.63,3.37,1.59,3.69,4.25,
+ 0.80,2.10,7.20,7.62,9.16,10.36,1.56,1.10,9.22,6.57,4.75,
+ 5.19,6.32,8.44,6.01,5.82,4.77,3.30,6.83,0.85,1.26,3.94,
+ 9.51,4.51,5.45,5.20,2.19,2.39,1.75,1.62,3.51,4.13,3.41,
+ 4.68,4.25,3.25,3.82,4.50,4.46,6.27,13.62,4.93,2.26,1.11,
+ 5.49,1.74,4.62,7.95,5.95,9.32,11.89,4.18,6.56,7.35,5.11,
+ 7.80,7.34,13.62,5.29),ncol=5,byrow=FALSE)
> dataSRabs<-abs(data-m0) # values |Xij-m0|
> dataSRabsranks<-apply(dataSRabs,1,rank) # ranks of values |Xij-m0|
> dataSRsgn<-sign(data-m0) # signs of differences Xij-m0
> vec1<-c() # empty vector
> # calculation of SRi values
> for(j0 in 1:40){
+   vec1[j0]<-sum(dataSRsgn[j0,]*dataSRabsranks[,j0])
```

```
+ }
> # construction of SR-chart
> plot(1:40,vec1,type='n',xlab='sample number',ylab=expression(SR[i]),
   ylim=c(-16.5,16.5))
> points(1:40,vec1,pch=19)
> lines(1:40,vec1,lty=1,lwd=1)
> abline(h=15,lwd=2)
> abline(h=-15,lwd=2)
> abline(h=0,lwd=2)
> text(3,15.95,'UCL=15')
> text(3,-15.95,'LCL=-15')
> text(38,0.95,'CL=0')
```



Σχήμα 12.4: Διάγραμμα ελέγχου SR-chart για τα δεδομένα του Πίνακα 12.3.

Από την εικόνα του διαγράμματος συμπεραίνουμε ότι το διάγραμμα ελέγχου δίνει για πρώτη φορά ένδειξη εκτός ελέγχου διεργασίας στο δείγμα 27, αφού η τιμή SR_{26} είναι ίση με 15 και «πέφτει» πάνω στο άνω όριο ελέγχου. Θα πρέπει να γίνει ειδική διερεύνηση αν πράγματι η διεργασία λειτουργεί παρουσία ειδικών αιτιών μεταβλητότητας (δηλαδή είναι πράγματι εκτός ελέγχου) ή αν η συγκεκριμένη ένδειξη είναι ψευδής (οπότε και η διεργασία είναι στην πραγματικότητα εντός ελέγχου). □

12.4 Μη παραμετρικά διαγράμματα ελέγχου για τη διασπορά

Στην παρούσα ενότητα, θα περιγράψουμε τον τρόπο κατασκευής ενός διαγράμματος ελέγχου για την ανίχνευση αλλαγών στη μεταβλητότητα της διεργασίας. Το διάγραμμα αυτό προτάθηκε από τους Amin *et al.* (1995) και βασίζεται στο Προσημικό κριτήριο. Όμως, αντί να καταγράφουμε σε κάθε δείγμα το πλήθος των

παρατηρήσεων που είναι μεγαλύτερες από τη διάμεσο m_0, καταγράφουμε το πλήθος των παρατηρήσεων οι οποίες βρίσκονται εκτός του διαστήματος το οποίο ορίζεται από τα τεταρτημόρια της εντός ελέγχου κατανομής του X. Όπως και στην περίπτωση του διαγράμματος ελέγχου SN-chart, όπου θεωρήσαμε ότι η διάμεσος m_0 είναι γνωστή, έτσι και εδώ, θα υποθέσουμε ότι το 1ο και το 3ο τεταρτημόριο της εντός ελέγχου κατανομής του X είναι γνωστά.

Για να κάνουμε έλεγχο διασποράς με χρήση του Προσημικού κριτηρίου, μπορούμε να χρησιμοποιήσουμε το κριτήριο που μελετήθηκε από τον Bradley (1968) και είναι γνωστό και ως *Westenberg's two-sample interquartile range test*. Το κριτήριο αυτό βασίζεται στη συγχώνευση (pooling) δύο δειγμάτων σε ένα κοινό δείγμα και στην καταμέτρηση του πλήθους των παρατηρήσεων από το 1ο δείγμα, που είναι μεγαλύτερες από το 3ο τεταρτημόριο ($x_{0.25}$) ή μικρότερες από το 1ο τεταρτημόριο ($x_{0.75}$), όπου τα $x_{0.75}$, $x_{0.25}$ είναι τα τεταρτημόρια που προκύπτουν από το κοινό δείγμα. Εδώ, όμως, θα δούμε πως μπορεί να τροποποιηθεί το παραπάνω κριτήριο για τον έλεγχο διασποράς, ώστε να χρησιμοποιηθεί ως στατιστική συνάρτηση σε ένα μη παραμετρικό διάγραμμα ελέγχου. Ουσιαστικά, οι Amin *et al.* (1995) χρησιμοποίησαν το τεστ που μελετήθηκε από τον Bradley (1968) και ανέπτυξαν ένα μη παραμετρικό διάγραμμα ελέγχου για την ανίχνευση αλλαγών στη διασπορά της διεργασίας.

Στην πράξη, τα $x_{0.75}$, $x_{0.25}$ τα γνωρίζουμε από την ίδια τη διεργασία (π.χ. μέσω των μηχανικών παραγωγής) ή τα εκτιμούμε από ένα κατάλληλο δείγμα Φάσης Ι, το οποίο -προφανώς- θα πρέπει να έχει ληφθεί όταν η διεργασία είναι εντός ελέγχου. Έστω ότι η τυχαία μεταβλητή U_{ij} ορίζεται ως εξής:

$$U_{ij} = \begin{cases} 1, & \text{αν } X_{ij} < x_{0.75} \text{ ή } X_{ij} > x_{0.25} \\ 0, & \text{αν } X_{ij} = x_{0.75} \text{ ή } X_{ij} = x_{0.25} \\ -1, & \text{αν } x_{0.75} \leqslant X_{ij} \leqslant x_{0.25}. \end{cases}$$

όπου X_{ij} είναι η j-οστή παρατήρηση του i-οστού δείγματος. Τότε, κατ' αντιστοιχία με τη στατιστική συνάρτηση T_i του προσημικού κριτηρίου που παρουσιάστηκε στην Ενότητα 12.2, η στατιστική συνάρτηση του κριτηρίου είναι

$$U_i = \sum_{j=1}^{n} U_{ij}$$

όπου $U_i = 2V_i - n$, με την τυχαία μεταβλητή V_i να ακολουθεί τη διωνυμική κατανομή $B(n,p)$ με $p = P(U_{ij} = 1)$.

Όταν η διεργασία είναι εντός στατιστικού ελέγχου, η πιθανότητα επιτυχίας p ισούται με 0.5. Άρα, μπορούμε να ορίσουμε ένα άνω μονόπλευρο διάγραμμα ελέγχου, το οποίο ονομάζεται V- chart, για την παρακολούθηση της μεταβλητότητας της διεργασίας. Στα προβλήματα του στατιστικού ελέγχου διεργασιών η ανίχνευση αύξησης στη μεταβλητότητα της διεργασίας είναι ιδιαίτερα σημαντική, διότι συνδέεται άμεσα με τη χειροτέρευση των προϊόντων που παράγει η διεργασία. Όμως, θα μπορούσαμε να αναπτύξουμε και ένα δίπλευρο διάγραμμα ελέγχου με βάση τη σ.σ. V_i, ώστε να μπορούμε να ανιχνεύσουμε είτε αυξήσεις είτε μειώσεις στη μεταβλητότητα της διεργασίας. Η ανίχνευση μείωσης στη μεταβλητότητα της διεργασίας συνδέεται με τη βελτίωση της ποιότητας των προϊόντων που παράγει η διεργασία και οι σύγχρονες τεχνικές του στατιστικού ελέγχου διεργασιών δίνουν ιδιαίτερη έμφαση στην έγκαιρη ανίχνευση βελτίωσης, ειδικά μετά από την εφαρμογή διορθωτικών αλλαγών σε αυτή. Στη συνέχεια, δεν θα μας απασχολήσει η περίπτωση του δίπλευρου διαγράμματος ελέγχου V-chart και αφήνεται ως άσκηση.

Έτσι, το άνω μονόπλευρο διάγραμμα ελέγχου V- chart δίνει ένδειξη εκτός ελέγχου διεργασίας, όταν για πρώτη φορά $V_i \geqslant UCL$, όπου το V_i είναι ο αριθμός των X_{ij} που είτε είναι μεγαλύτερα της τιμής $x_{0.25}$ είτε είναι μικρότερα της τιμής $x_{0.75}$. Τα $x_{0.75}$ και $x_{0.25}$ αντιστοιχούν στο 1ο και 3ο τεταρτημόριο της κατανομής του χαρακτηριστικού όταν η διεργασία είναι εντός ελέγχου. Παρακάτω, δίνεται ένα παράδειγμα εφαρμογής της παραπάνω διαδικασίας. Αξίζει να αναφέρουμε πως είναι δυνατή η γενίκευσή της χρησιμοποιώντας αντί των $x_{0.75}$, $x_{0.25}$ οποιοδήποτε άλλο ζεύγος ποσοστιαίων σημείων επιθυμούμε (βλ. Άσκηση 12.5).

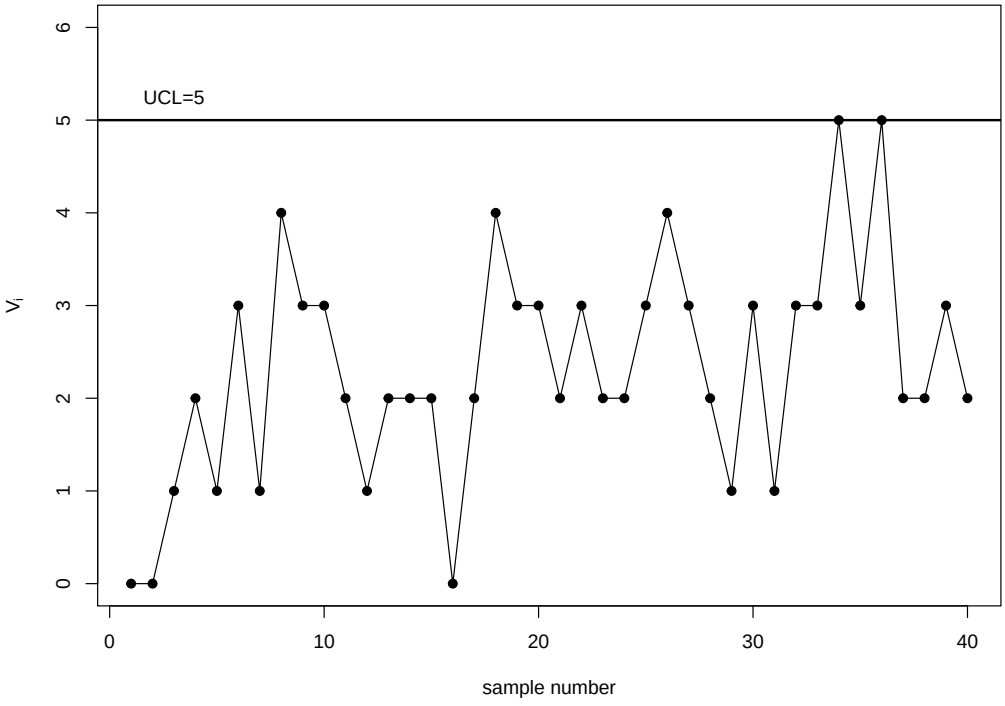
Παράδειγμα 12.3. Τα δεδομένα στον Πίνακα 12.9 προέρχονται από άγνωστο πληθυσμό, για τον οποίο από προηγούμενες μελέτες γνωρίζουμε ότι το 25% των τιμών του είναι μικρότερες του 248, ενώ ένα 25% των τιμών του ξεπερνά την τιμή 251. Να παρακολουθηθεί η διεργασία και, συγκεκριμένα, η μεταβλητότητα αυτής, χρησιμοποιώντας ένα μονόπλευρο διάγραμμα ελέγχου V-chart. Ως άνω όριο ελέγχου να χρησιμοποιηθεί η τιμή $UCL = 5$.

Λύση Παραδείγματος 12.3. Από τα δεδομένα έχουμε ότι $x_{0.75} = 248$ και $x_{0.25} = 251$. Στη συνέχεια, θα πρέπει να υπολογίσουμε τις τιμές της σ.σ. V, οι οποίες απεικονίζονται στο διάγραμμα. Για να γίνει κατανοητή η διαδικασία, ας θεωρήσουμε το 1ο δείγμα τιμών, δηλαδή τις τιμές 248.75, 249.89, 250.8, 250.58, 248.76. Άμεσα διαπιστώνουμε ότι όλες οι τιμές είναι μεταξύ 248 και 251. Άρα, για $j = 1,2,\ldots,5$ είναι $U_{1j} = -1$, οπότε

$$U_1 = \sum_{j=1}^{5} U_{1j} = -5,\ V_1 = (U_1 + n)/2 = (-5 + 5)/2 = 0.$$

Με τον ίδιο τρόπο υπολογίζονται και οι υπόλοιπες τιμές της V_i, για $i = 1,\ldots,40$, οι οποίες δίνονται στον Πίνακα 12.10.

Το ζητούμενο διάγραμμα δίνεται στο Σχήμα 12.5. Αφού η απεικονιζόμενη σ.σ. είναι η V_i, για την οποία γνωρίζουμε πως όταν η διεργασία είναι εντός στατιστικού ελέγχου ακολουθεί διωνυμική κατανομή $B(n,0.5)$, έπεται ότι η πιθανότητα εσφαλμένου συναγερμού είναι $P(V_i \geqslant UCL) = P(V_i \geqslant 5) = 0.03125$. Παρατηρήστε, επίσης, πως, αφού το μέγεθος δείγματος είναι $n = 5$, αυτή είναι η μικρότερη δυνατή τιμή που μπορούμε να πετύχουμε για την πιθανότητα εσφαλμένου συναγερμού, ενώ η αντίστοιχη τιμή για το εντός ελέγχου ARL είναι $1/0.03125 = 32$. Από την εικόνα του διαγράμματος (Σχήμα 12.5) παρατηρούμε ότι έχουμε για πρώτη φορά ένδειξη εκτός ελέγχου διεργασίας στο δείγμα 34, αφού το αντίστοιχο σημείο (τιμή V_{34}) βρίσκεται ακριβώς πάνω στο άνω όριο ελέγχου $UCL = 5$.



Σχήμα 12.5: Διάγραμμα ελέγχου V-chart για τα δεδομένα του Πίνακα 12.9.

A/A	X_1	X_2	X_3	X_4	X_5
1	248.75	249.89	250.8	250.58	248.76
2	250.37	249.69	248.78	249.11	250.08
3	248.33	247.06	250.68	250	248.18
4	253.19	249.04	247.74	250.15	250.32
5	250.66	250.84	252.87	248.82	248.69
6	248.36	252.72	253.96	248.86	253.53
7	250.97	249.79	249.27	249.73	251.43
8	251.48	250.78	247.91	252.36	251.82
9	251.15	249.89	251.14	246.95	250.77
10	249.39	247.25	249.73	251.19	253.36
11	253.02	249.17	254.8	250.67	248.73
12	250.78	249.21	249.92	252.13	249.08
13	248.76	249.88	251.38	249.39	252.86
14	245.57	252.2	250.06	250.74	248.7
15	252.25	251.53	248.51	250.53	249.59
16	249.91	249.67	250.38	248.91	249.21
17	249.97	249.49	246.39	252.42	249.36
18	251.89	251.39	252.93	252.32	249.44
19	251.64	251.11	250.31	251.4	250.99
20	251.19	248.62	254.35	253.17	249.65
21	251.84	248.59	250.95	251.12	248.99
22	251.56	250.73	248.58	247.45	252.69
23	250.15	251.54	251.22	248.85	249.57
24	246.02	249.78	248.13	247.55	249.64
25	251.24	251.76	247.49	249.05	249.8
26	251.78	245.21	247.31	255.77	249.6
27	249.82	252.94	252.5	250.26	253.66
28	249.91	245.84	248.45	251.14	248.08
29	248.3	248.84	246.54	249.81	248.92
30	249.19	247.21	254.67	249.16	247.68
31	250.15	248.12	251.06	249.91	249.56
32	248.53	255.22	249.4	251.97	251.01
33	251.33	250.04	252.65	255.19	248.17
34	246.2	246.78	252.22	252.57	252.08
35	250.77	245.9	248.45	253.02	246.98
36	246.16	251.13	255.52	246.92	247.38
37	249.25	249.95	249.36	252.46	253.6
38	248.68	249.2	246.44	250.55	247.46
39	248.37	247.68	249.64	246.33	251.03
40	249.86	246.28	250.52	251.3	249.05

Πίνακας 12.9: Δεδομένα εφαρμογής διαγράμματος ελέγχου V-chart.

i	1	2	3	4	5	6	7	8	9	10
V_i	0	0	3	2	1	4	1	4	1	2
i	11	12	13	14	15	16	17	18	19	20
V_i	2	1	2	2	3	0	2	4	2	3
i	21	22	23	24	25	26	27	28	29	30
V_i	2	4	1	3	2	4	3	3	2	3
i	31	32	33	34	35	36	37	38	39	40
V_i	1	3	3	5	4	4	2	2	3	1

Πίνακας 12.10: Τιμές της σ.σ. V_i για τα δεδομένα του Πίνακα 12.9.

Ολοκληρώνοντας τη λύση του παραδείγματος, δίνουμε και τις εντολές στην R για τον υπολογισμό των τιμών στον Πίνακα 12.10 καθώς και για την κατασκευή του διαγράμματος ελέγχου στο Σχήμα 12.5.

```
> n<-5 # sample size
> q1<-248 # Q1
> q3<-251 # Q3
> # data entry
> data<-matrix(c(248.75,250.37,248.33,253.19,250.66,248.36,250.97,
+ 251.48,251.15,249.39,253.02,250.78,248.76,245.57,252.25,249.91,
+ 249.97,251.89,251.64,251.19,251.84,251.56,250.15,246.02,251.24,
+ 251.78,249.82,249.91,248.30,249.19,250.15,248.53,251.33,246.20,
+ 250.77,246.16,249.25,248.68,248.37,249.86,249.89,249.69,247.06,
+ 249.04,250.84,252.72,249.79,250.78,249.89,247.25,249.17,249.21,
+ 249.88,252.20,251.53,249.67,249.49,251.39,251.11,248.62,248.59,
+ 250.73,251.54,249.78,251.76,245.21,252.94,245.84,248.84,247.21,
+ 248.12,255.22,250.04,246.78,245.90,251.13,249.95,249.20,247.68,
+ 246.28,250.80,248.78,250.68,247.74,252.87,253.96,249.27,247.91,
+ 251.14,249.73,254.80,249.92,251.38,250.06,248.51,250.38,246.39,
+ 252.93,250.31,254.35,250.95,248.58,251.22,248.13,247.49,247.31,
+ 252.50,248.45,246.54,254.67,251.06,249.40,252.65,252.22,248.45,
+ 255.52,249.36,246.44,249.64,250.52,250.58,249.11,250.00,250.15,
+ 248.82,248.86,249.73,252.36,246.95,251.19,250.67,252.13,249.39,
+ 250.74,250.53,248.91,252.42,252.32,251.40,253.17,251.12,247.45,
+ 248.85,247.55,249.05,255.77,250.26,251.14,249.81,249.16,249.91,
+ 251.97,255.19,252.57,253.02,246.92,252.46,250.55,246.33,251.30,
+ 248.76,250.08,248.18,250.32,248.69,253.53,251.43,251.82,250.77,
+ 253.36,248.73,249.08,252.86,248.70,249.59,249.21,249.36,249.44,
+ 250.99,249.65,248.99,252.69,249.57,249.64,249.80,249.60,253.66,
+ 248.08,248.92,247.68,249.56,251.01,248.17,252.08,246.98,247.38,
+ 253.60,247.46,251.03,249.05),ncol=5,byrow=F)
> vec1<-apply(ifelse(data>q3|data<q1,1,-1),1,sum)
> V1<-(vec1+n)/2 # Vi values
> # construction of V-chart
> plot(1:40,V1,type='n',xlab='sample number',ylab=expression(V[i]),
+ ylim=c(0,6.0))
> points(1:40,V1,pch=19)
> lines(1:40,V1,lty=1,lwd=1)
> abline(h=5,lwd=2)
> text(3,5.25,'UCL=5')
```

□

12.5 Διαγράμματα ελέγχου με χρήση του κριτηρίου προηγήσεων

Γενικά, η χρήση των διαγραμμάτων ελέγχου, και ειδικά αυτών που αναφέρονται ως παραμετρικά, είναι πιο άμεση όταν είναι γνωστές οι παράμετροι του πληθυσμού. Στην πράξη όμως, οι τιμές των παραμέτρων δεν είναι γνωστές και θα πρέπει να προσδιοριστούν. Όπως έχουμε αναφέρει και στην εισαγωγική ενότητα του παρόντος κεφαλαίου, πριν την εφαρμογή ενός διαγράμματος ελέγχου στην πράξη, απαιτείται η διεξαγωγή της ανάλυσης Φάσης I, ώστε να πιστοποιηθεί η διεργασία, δηλαδή πώς λειτουργεί, όταν είναι εντός ελέγχου και, στη συνέχεια, να υπολογιστούν (εκτιμηθούν) οι τιμές των ορίων ελέγχου από τα δεδομένα της Φάσης I. Τα δεδομένα της Φάσης I τα οποία χρησιμοποιούνται για τον σκοπό αυτό είναι γνωστά και ως δεδομένα αναφοράς (reference data) ή ως δείγμα αναφοράς (reference sample). Αφού εκτιμήσουμε τις τιμές των ορίων ελέγχου, στη συνέχεια προχωράμε στην ανάλυση Φάσης II, κατά την οποία παρακολουθούμε τη διεργασία σε πραγματικό χρόνο.

Στη συνέχεια, θα δείξουμε πώς μπορούμε να εκτιμήσουμε τα όρια ελέγχου ενός μη παραμετρικού διαγράμματος ελέγχου με βάση ένα δείγμα αναφοράς. Οι Janacek and Meikle (1997) πρότειναν ένα μη παραμετρικό διάγραμμα ελέγχου στο οποίο τα όρια ελέγχου αποτελούν δύο διατεταγμένες τιμές από ένα δείγμα αναφοράς και, στη συνέχεια, η τιμή που απεικονίζεται στο διάγραμμα (κατά την ανάλυση Φάσης II) είναι η διάμεσος m_i του i-οστού δείγματος, $i = 1, 2, \ldots$. Οι Chakraborti *et al.* (2004) επέκτειναν τη μεθοδολογία των Janacek and Meikle (1997) και, αντί της διαμέσου, μπορεί να χρησιμοποιηθεί οποιαδήποτε διατεταγμένη στατιστική συνάρτηση (κατά την ανάλυση Φάσης II). Το διάγραμμα των Chakraborti *et al.* (2004) είναι γνωστό ως διάγραμμα ελέγχου Προηγήσεων (Precedence control chart), ενώ το διάγραμμα ελέγχου των Janacek and Meikle (1997) αναφέρεται συνήθως ως διάγραμμα ελέγχου διαμέσου (median control chart). Αξίζει να αναφέρουμε πως η στατιστική συνάρτηση του διαγράμματος ελέγχου διαμέσων προκύπτει από τον έλεγχο που προτάθηκε από τον Mathisen (1943) και αφορά τον έλεγχο της υπόθεσης ότι δύο ανεξάρτητα δείγματα προέρχονται από τον ίδιο πληθυσμό, χρησιμοποιώντας ως στατιστική συνάρτηση ελέγχου το πλήθος των στοιχείων του 2ου δείγματος, των οποίων οι τιμές είναι μικρότερες από τη διάμεσο των τιμών στο 1ο δείγμα.

Έστω ότι έχουμε στη διάθεσή μας ένα δείγμα αναφοράς μεγέθους n_1 από μια εντός ελέγχου διεργασία με α.σ.κ. $F_X(x) = F(x - \theta)$, $x \in \mathbb{R}$ και θ είναι η παράμετρος θέσης του πληθυσμού. Εδώ, θεωρούμε ότι το θ είναι η διάμεσος του πληθυσμού από τον οποίο προέρχεται το δείγμα αναφοράς. Τα όρια ελέγχου του διαγράμματος ελέγχου Προηγήσεων δίνονται από τις τιμές δύο διατεταγμένων στατιστικών συναρτήσεων του δείγματος αναφοράς, έστω $LCL = X_{(\nu):n_1}$ και $UCL = X_{(\kappa):n_1}$, με $1 \leqslant \nu < \kappa \leqslant n_1$. Έστω, επίσης, $Y^{(h)}_{(j):n_2}$ η j διατεταγμένη παρατήρηση από το h-οστό δείγμα ($h = 1, 2, \ldots$) μεγέθους n_2, το οποίο συλλέγεται κατά την ανάλυση Φάσης II. Το δείγμα αυτό είναι γνωστό και ως δείγμα ελέγχου (test group) και υποθέτουμε ότι η α.σ.κ. του πληθυσμού, από τον οποίο προέρχονται τα δείγματα ελέγχου, είναι η $G(x) = F(x - \theta_h)$, όπου θ_h είναι η παράμετρος θέσης της κατανομής από την οποία προέρχεται το h-οστό δείγμα ελέγχου. Για λόγους απλότητας θα υποθέσουμε ότι η $Y^{(h)}_{(j):n}$ αντιστοιχεί στη διάμεσο του δείγματος (περίπτωση του διαγράμματος ελέγχου που προτάθηκε από τους Janacek and Meikle, 1997), αλλά μπορεί να χρησιμοποιηθεί οποιαδήποτε άλλη διατεταγμένη στατιστική συνάρτηση.

Όταν η διεργασία βρίσκεται εντός ελέγχου, τότε η κατανομή του πληθυσμού, από την οποία προέρχονται τα δείγματα ελέγχου είναι η ίδια με την κατανομή του πληθυσμού, από την οποία προέρχεται το δείγμα αναφοράς. Έτσι όπως έχουμε διατυπώσει το πρόβλημα, η διεργασία παραμένει εντός στατιστικού ελέγχου όσο ισχύει ότι $\theta = \theta_h$, $h = 1, 2, \ldots$ (δεν υπάρχει διαφορά στη διάμεσο μεταξύ των πληθυσμών από τους οποίους προέρχονται το δείγμα αναφοράς και τα δείγματα ελέγχου). Αν συμβολίζουμε ως W_j το πλήθος των παρατηρήσεων X που *προηγούνται* (δηλαδή δεν είναι μεγαλύτερες) της $Y^{(h)}_{(j):n}$, τότε η στατιστική συνάρτηση W_j είναι γνωστή ως στατιστική συνάρτηση προήγησης (precedence statistic) και ο έλεγχος υπόθεσης που βασίζεται σε αυτήν είναι γνωστός ως κριτήριο Προηγήσεων. Το κριτήριο αυτό έχει χρησιμοποιηθεί ως ένας γρήγορος και απλός, μη παραμετρικός έλεγχος σε δεδομένα χρόνων ζωής (βλ., π.χ. Nelson, 1963, 1993).

Για περισσότερες λεπτομέρειες σχετικά με το κριτήριο Προηγήσεων, αλλά και επεκτάσεις αυτού, παραπέμπουμε στην εργασία των Chakraborti and Van der Laan (1996) και στο σύγγραμμα των Balakrishnan and Ng (2006).

Για τον στατιστικό σχεδιασμό του διαγράμματος ελέγχου Προηγήσεων πρέπει να προσδιοριστούν οι τιμές των ορίων ελέγχου (δηλαδή οι τιμές των ν, κ), ώστε η πιθανότητα εσφαλμένου συναγερμού να είναι η επιθυμητή. Αν το δείγμα αναφοράς έχει μέγεθος n_1, τα δείγματα που συλλέγονται κατά την ανάλυση Φάσης II (δείγματα ελέγχου) έχουν μέγεθος n_2 και στο διάγραμμα απεικονίζεται η j-οστή διατεταγμένη παρατήρηση από τα δείγματα ελέγχου, τότε για τον προσδιορισμό των ν, κ χρησιμοποιείται η παρακάτω σχέση (βλ. Chakraborti *et al.*, 2004)

$$\sum_{u=\nu}^{\kappa-1} \frac{\binom{j+u-1}{u}\binom{n_1+n_2-j-u}{n_1-u}}{\binom{n_1+n_2}{n_1}} \geqslant 1-a, \tag{12.2}$$

όπου a είναι η πιθανότητα εσφαλμένου συναγερμού.

Χρησιμοποιώντας την παραπάνω ανισότητα οι Chakraborti *et al.* (2004) έδωσαν πίνακες με τις τιμές των ν, κ για διάφορες τιμές των n_1, n_2 και a. Επιπλέον, η μελέτη τους έδειξε ότι το διάγραμμα ελέγχου Προηγήσεων είναι πιο ανθεκτικό στην παραβίαση της υπόθεσης της κανονικότητας από το σύνηθες παραμετρικό διάγραμμα ελέγχου Shewhart $\bar{X}$. Ως μη παραμετρικό διάγραμμα ελέγχου έχει την ιδιότητα η πιθανότητα εσφαλμένου συναγερμού να είναι ίδια για οποιαδήποτε συνεχή κατανομή πιθανότητας, όταν η διεργασία είναι εντός ελέγχου. Στα παραμετρικά διαγράμματα ελέγχου, η συγκεκριμένη πιθανότητα υπολογίζεται με βάση το υποτιθέμενο πρότυπο (π.χ. κατανομή από την οποία προέρχεται το δείγμα αναφοράς).

Παρακάτω, δίνουμε τις εντολές στην R για τον υπολογισμό των ν, κ για δεδομένες τιμές των n_1, n_2 και a. Με τη βοήθεια αυτών των εντολών μπορούμε να επιβεβαιώσουμε τον πίνακα Table 1 στην εργασία των Chakraborti *et al.* (2004). Η τελευταία εντολή τυπώνει τις τιμές των ν, κ, καθώς και την πραγματική πιθανότητα εσφαλμένου συναγερμού, η οποία δεν θέλουμε να ξεπερνά την επιθυμητή. Αξίζει να αναφέρουμε πως ο αλγόριθμος ξεκινά με αρχικές τιμές για τα ν και κ, οι οποίες δίνονται στις γραμμές 6 και 7, αντίστοιχα. Στην πραγματικότητα, αφού η απεικονιζόμενη σ.σ. είναι η διάμεσος, αρκεί να βρούμε την τιμή για το ν, αφού, για δεδομένη τιμή του ν, το κ υπολογίζεται ως $n_1 - \nu + 1$. Σε κάθε βήμα ο αλγόριθμος ελέγχει αν για τη δεδομένη τιμή ν ικανοποιείται η ανισότητα της εξίσωσης (12.2). Αν ικανοποιείται, ο αλγόριθμος προχωρά έχοντας αυξήσει κατά 1 μονάδα την τιμή του ν. Όταν για πρώτη φορά παύει να ισχύει η ανισότητα της εξίσωσης (12.2), τότε τερματίζει ο αλγόριθμος και τυπώνονται οι τιμές των ν, κ, καθώς και η πιθανότητα εσφαλμένου συναγερμού του διαγράμματος.

```
1  > a<-0.01 # false alarm rate
2  > p0<-(1-a)
3  > n1<-50 # size of reference sample
4  > n2<-5 # size of test sample
5  > j0<-3 # this is the case of median mi
6  > nu0<-1 # starting value for nu
7  > kappa0<-n1-nu0+1 # kappa in the case of median only
8  > myfun1<-function(w){
9  +   choose(j0+w-1,w)*choose(n1+n2-j0-w,n1-w)/choose(n1+n2,n1)
10 +  }
11 > S1<-sum(myfun1(nu0:(kappa0-1)))
12 > while(S1>=p0){
13 +   nu0<-nu0+1
14 +   kappa0<-n1-nu0+1
15 +   S1<-sum(myfun1(nu0:(kappa0-1)))
16 +  }
17 > nu0<-nu0-1
```

```
> kappa0<-n1-nu0+1
> S1<-sum(myfun1(nu0:(kappa0-1)))
> cat('nu:',nu0,' kappa:',kappa0,' FAR:',1-S1,'\n')
```

Στη συνέχεια, δίνουμε ένα παράδειγμα εφαρμογής του διαγράμματος ελέγχου Προηγήσεων.

Παράδειγμα 12.4. Στον Πίνακα 12.11 δίνεται ένα δείγμα αναφοράς το οποίο συλλέχθηκε από μια εντός ελέγχου διεργασία. Η κατανομή του πληθυσμού, από την οποία προήλθε το δείγμα, είναι άγνωστη. Χρησιμοποιώντας ως δεδομένα για την ανάλυση Φάσης ΙΙ, που δίνονται στον Πίνακα 12.12, να αναπτυχθεί, με $a = 0.01$, ένα διάγραμμα ελέγχου Προηγήσεων, με στατιστική συνάρτηση τη διάμεσο. Για διευκόλυνση στον Πίνακα 12.12, στη στήλη $Y^{(h)}_{(3):5}$, δίνεται η διάμεσος του h-οστού δείγματος, $h = 1, 2, \ldots, 15$.

A/A	X_1	X_2	X_3	X_4	X_5
1	1.332	26.052	17.397	10.861	14.17
2	20.036	5.642	9.354	31.706	28.449
3	11.938	10.596	54.033	21.276	22.937
4	9.469	23.216	10.295	11.938	46.044
5	7.385	28.938	34.768	6.109	5.203
6	20.785	10.344	8.714	26.634	32.735
7	12.186	22.669	8.999	5.934	18.872
8	18.122	10.556	8.308	35.719	7.87
9	16.567	11.471	25.291	9.26	9.649
10	11.451	5.501	12.763	6.553	12.759
11	38.432	24.693	5.543	8.843	27.237
12	13.813	7.666	18.805	23.483	10.442
13	21.732	5.188	6.568	10.561	5.074
14	12.715	4.792	14.468	12.116	5.038
15	33.104	12.402	15.01	2.84	12.839
16	9.685	16.903	24.041	6.356	13.839
17	5.697	15.95	10.931	1.134	2.01
18	14.409	18.601	17.635	30.517	12.814
19	4.041	5.732	13.912	10.493	19.528
20	24.766	21.048	27.813	30.586	33.047

Πίνακας 12.11: Δείγμα αναφοράς μεγέθους $n_1 = 100$.

Αφού σε κάθε δείγμα έχουμε 5 μετρήσεις και συνολικά 20 δείγματα, το μέγεθος του δείγματος αναφοράς είναι $n_1 = 100$, ενώ είναι $n_2 = 5$. Χρησιμοποιώντας τις εντολές στην R, που δόθηκαν προηγουμένως, για τις δεδομένες τιμές των n_1, n_2 και $a = 0.01$ (ή τον πίνακα Table 1 στην εργασία των Chakraborti *et al.*, 2004) έχουμε ότι $\nu = 7$ και $\kappa = 94$. Παρακάτω δίνονται οι σχετικές εντολές.

```
> a<-0.01 # false alarm rate
> p0<-(1-a)
> n1<-100 # size of reference sample
> n2<-5 # size of test sample
> j0<-3 # this is the case of median mi
> nu0<-1 # starting value for nu
> kappa0<-n1-nu0+1 # kappa in the case of median only
> myfun1<-function(w){
+   choose(j0+w-1,w)*choose(n1+n2-j0-w,n1-w)/choose(n1+n2,n1)
+  }
```

```
> S1<-sum(myfun1(nu0:(kappa0-1)))
> while(S1>=p0){
+   nu0<-nu0+1
+   kappa0<-n1-nu0+1
+   S1<-sum(myfun1(nu0:(kappa0-1)))
+  }
> nu0<-nu0-1
> kappa0<-n1-nu0+1
> S1<-sum(myfun1(nu0:(kappa0-1)))
> cat('nu:',nu0,' kappa:',kappa0,' FAR:',1-S1,'\n')
nu: 7  kappa: 94  FAR: 0.008186813
```

Άρα, από τον Πίνακα 12.11 έχουμε ότι $LCL = X_{(7):100} = 5.038$ (7η διατεταγμένη παρατήρηση) και $UCL = X_{(94):100} = 33.047$ (94η διατεταγμένη παρατήρηση). Επιπλέον, η πραγματική πιθανότητα εσφαλμένου συναγερμού είναι 0.0082 (βλ. την τιμή `FAR` στην τελευταία γραμμή των εντολών της `R` που δόθηκαν προηγουμένως), ενώ η τιμή του εντός ελέγχου ARL είναι (περίπου) ίση με 122, αφού $1/0.008186813 = 122.1477 \approx 122$. Στη συνέχεια, κατασκευάζουμε το διάγραμμα ελέγχου Προηγήσεων για τα δεδομένα ανάλυσης Φάσης II (βλ. Πίνακα 12.12). Η εικόνα του διαγράμματος δίνεται στο Σχήμα 12.6. Τα σημεία στο διάγραμμα είναι οι δειγματικές διάμεσοι από κάθε δείγμα ελέγχου (βλ. στήλη $Y^{(h)}_{(3):5}$ στον Πίνακα 12.12).

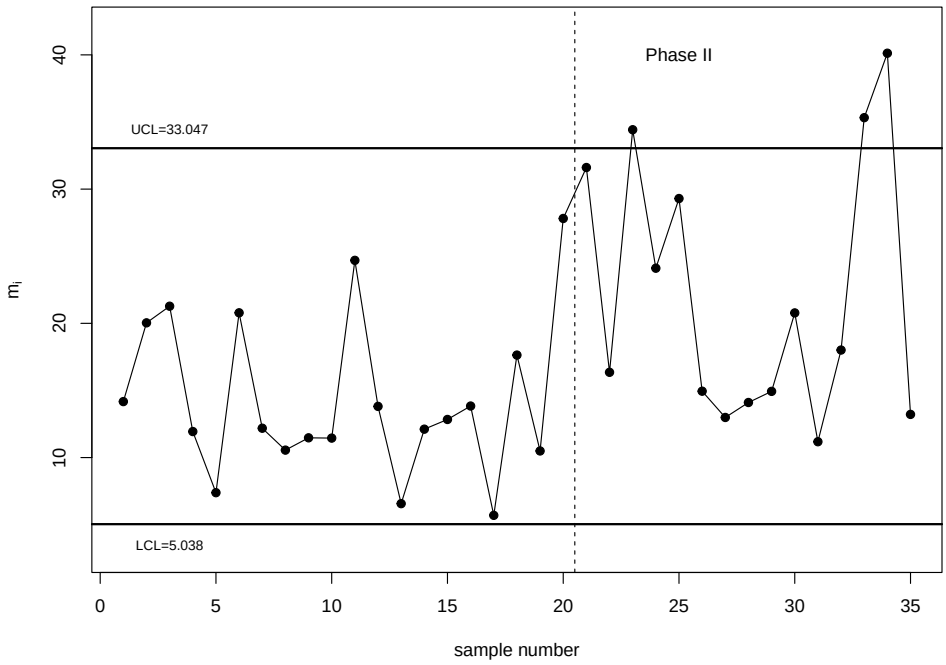
h	X_1	X_2	X_3	X_4	X_5	$Y^{(h)}_{(3):5}$
1	7.064	37.74	5.369	47.485	31.608	31.608
2	5.064	35.562	6.9	16.353	25.903	16.353
3	46.955	34.424	17.243	61.201	8.681	34.424
4	24.108	16.529	26.721	47.875	14.088	24.108
5	39.894	32.364	9.463	29.298	17.714	29.298
6	5.315	10.698	14.945	20.115	64.977	14.945
7	7.487	12.988	9.091	18.254	30.433	12.988
8	23.691	26.753	11.458	14.103	12.812	14.103
9	42.153	34.864	3.741	14.932	11.248	14.932
10	21.924	20.777	18.993	23.948	18.918	20.777
11	6.388	17.387	11.032	26.984	11.18	11.18
12	18.01	3.227	3.968	18.309	26.296	18.01
13	8.404	35.322	7.903	44.885	41.889	35.322
14	17.132	46.782	27.048	42.23	40.124	40.124
15	38.576	13.216	18.958	10.822	12.136	13.216

Πίνακας 12.12: Δεδομένα ανάλυσης Φάσης II για το Παράδειγμα 12.4.

Παρατηρούμε ότι το διάγραμμα δίνει για πρώτη φορά ένδειξη εκτός ελέγχου διεργασίας στο 23ο δείγμα (ή 3ο δείγμα από την έναρξη της Φάσης II). Επιπλέον, από την ανάλυση Φάσης I (δείγματα 1-20) δεν έχουμε κάποιο σημείο εκτός των ορίων ελέγχου. Άρα κατά τη συλλογή του δείγματος αναφοράς, η διεργασία ήταν εντός στατιστικού ελέγχου.

12.6 Διαγράμματα ελέγχου με χρήση του κριτηρίου Mann-Whitney

Δεν είναι δύσκολο να διαπιστώσουμε ότι το κριτήριο των προηγήσεων δεν είναι παρά ένας μη παραμετρικός έλεγχος για δύο ανεξάρτητα δείγματα. Το πλέον γνωστό μη παραμετρικό κριτήριο σε αυτήν την περίπτωση



Σχήμα 12.6: Διάγραμμα ελέγχου με χρήση Κριτηρίου Προηγήσεων για τα δεδομένα των Πινάκων 12.11 και 12.12.

είναι το τεστ των Mann-Whitney (βλ. Κεφάλαιο 6, Ενότητα 6.3). Μπορούμε, λοιπόν, να αναπτύξουμε ένα μη παραμετρικό διάγραμμα ελέγχου με βάση το τεστ των Mann-Whitney το οποίο θα έχει όλα τα πλεονεκτήματα των μη παραμετρικών διαγραμμάτων ελέγχου που αναφέρθηκαν έως τώρα, όπως, π.χ. να είναι απαλλαγμένο παραμέτρων και λόγω αυτού η εντός ελέγχου απόδοσή του να είναι η ίδια για όλες τις συνεχείς κατανομές. Όπως και στην περίπτωση του διαγράμματος ελέγχου Προηγήσεων, τα όρια ελέγχου του διαγράμματος προσδιορίζονται από ένα δείγμα αναφοράς Φάσης I. Οι ιδιότητες του διαγράμματος αυτού μελετήθηκαν από τους Chakraborti and Van de Wiel (2008), οι οποίοι έδωσαν και πίνακες για τον στατιστικό σχεδιασμό (επιλογή των τιμών των ορίων ελέγχου) του διαγράμματος ελέγχου Mann-Whitney. Επιπλέον, έδειξαν ότι έχει καλύτερη απόδοση έναντι του διαγράμματος ελέγχου Προηγήσεων.

Έστω ότι ένα δείγμα αναφοράς $X_1, X_2, \ldots, X_{n_1}$ μεγέθους n_1 είναι διαθέσιμο από μια εντός ελέγχου διεργασία. Έστω, επίσης, ότι κατά την ανάλυση Φάσης II (παρακολούθηση διεργασίας σε πραγματικό χρόνο) συλλέγονται δείγματα $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ μεγέθους n_2 με $\boldsymbol{Y}_h = (Y_{h1}, Y_{h2}, \ldots, Y_{hn_2}), h = 1, 2, \ldots$. Υποθέτουμε ότι τα δείγματα που συλλέγονται κατά τη Φάση II (δείγματα ελέγχου) είναι ανεξάρτητα μεταξύ τους, όπως επίσης είναι και ανεξάρτητα από το δείγμα αναφοράς. Όπως έχει ήδη αναφερθεί στο Κεφάλαιο 6, η στατιστική συνάρτηση για τον έλεγχο των Mann-Whitney βασίζεται στον συνολικό αριθμό των ζευγών (X_i, Y_j) με $Y_j < X_i$. Στο σημείο αυτό πρέπει να επισημάνουμε πως ο τρόπος με τον οποίο ανέπτυξαν οι Chakraborti and Van de Wiel (2008) το διάγραμμα ελέγχου MW βασίζεται στο πλήθος των ζευγών (X_i, Y_j) με $Y_j > X_i$. Η μετάβαση από τον έναν ορισμό στον άλλο είναι άμεση, όπως θα δούμε και στη συνέχεια. Έτσι, με βάση τον συμβολισμό που έχουμε χρησιμοποιήσει έως τώρα, η στατιστική συνάρτηση που χρησιμοποιείται για την ανάπτυξη του διαγράμματος ελέγχου MW είναι η

$$MW = R_2 - \frac{n_2(n_2+1)}{2},$$

όπου R_2 είναι το άθροισμα των τάξεων για τις παρατηρήσεις του δείγματος των Y, με τις τάξεις να έχουν υπολογιστεί στο κοινό δείγμα των $n_1 + n_2$ τιμών. Άμεσα έπεται ότι για κάθε δείγμα $\boldsymbol{Y}_h$ απαιτείται ο

υπολογισμός της αντίστοιχης τιμής MW_h, οπότε και προκύπτει μια ακολουθία τιμών $MW_1, MW_2, \ldots$, οι οποίες απεικονίζονται στο διάγραμμα. Υπενθυμίζεται ότι οι δυνατές τιμές της στατιστικής συνάρτησης MW είναι στο σύνολο $\{0, 1, \ldots, n_1 n_2\}$. Μεγάλες τιμές της MW αποτελούν ένδειξη ότι υπάρχει αύξηση στην παράμετρο θέσης της κατανομής από την οποία προέρχονται τα δεδομένα, ενώ, αντίστοιχα, μικρές τιμές αποτελούν ένδειξη μείωσης.

Το διάγραμμα δίνει ένδειξη εκτός ελέγχου διεργασίας όταν για πρώτη φορά είναι $MW_h < L_{n_1 n_2}$ ή $MW_h > U_{n_1 n_2}$, οτιδήποτε από τα δύο συμβεί πρώτο. Οι τιμές $L_{n_1 n_2}$, $U_{n_1 n_2}$ είναι το κάτω και το άνω όριο του διαγράμματος ελέγχου. Όταν η διεργασία είναι εντός ελέγχου, λόγω της συμμετρίας της κατανομής της στατιστικής συνάρτησης MW γύρω από την τιμή $n_1 n_2/2$, χρησιμοποιούμε ως κάτω όριο ελέγχου την τιμή $L_{n_1 n_2} = n_1 n_2 - U_{n_1 n_2}$. Άρα, ο στατιστικός σχεδιασμός του διαγράμματος ελέγχου MW απαιτεί τον προσδιορισμό μόνο της τιμής $U_{n_1 n_2}$, ώστε να έχει την επιθυμητή εντός ελέγχου απόδοση.

Στην πράξη, τα όρια ελέγχου του διαγράμματος προσδιορίζονται, ώστε το εντός ελέγχου ARL να έχει την επιθυμητή τιμή. Στο σημείο αυτό, θα πρέπει να αναφέρουμε πως, αν οι διαδοχικές στατιστικές συναρτήσεις $MW_1, MW_2, \ldots$, οι τιμές των οποίων απεικονίζονται στο διάγραμμα, ήταν ανεξάρτητες, τότε το εντός ελέγχου ARL θα ήταν απλά το αντίστροφο της πιθανότητας εσφαλμένου συναγερμού, η οποία ισούται με την πιθανότητα $2P(MW > U_{n_1 n_2})$, λόγω συμμετρίας της κατανομής. Άρα, θα μπορούσαμε να χρησιμοποιήσουμε ως άνω όριο ελέγχου $U_{n_1 n_2}$ (και να προκύψει άμεσα και το κάτω όριο ελέγχου $L_{n_1 n_2} = n_1 n_2 - U_{n_1 n_2}$) από τα ποσοστιαία σημεία της κατανομής της στατιστικής συνάρτησης για το τεστ των Mann-Whitney. Για τον υπολογισμό αυτών των ποσοστιαίων σημείων μπορεί να χρησιμοποιηθεί η συνάρτηση `qwilcox(...)` της `R`.

Όμως, οι διαδοχικές στατιστικές συναρτήσεις $MW_1, MW_2, \ldots$ είναι συσχετισμένες, αφού τα δείγματα μεγέθους n_2 που λαμβάνονται κατά την ανάλυση Φάσης II συγκρίνονται όλα με τα ίδια όρια ελέγχου, τα οποία έχουν προκύψει από το ίδιο δείγμα αναφοράς. Ο παραπάνω τρόπος προσδιορισμού των ορίων ελέγχου θα μπορούσε να χρησιμοποιηθεί για αρκετά μεγάλο μέγεθος n_1 του δείγματος αναφοράς και να αποτελέσει μια απλή αλλά και συνάμα προσεγγιστική λύση. Όμως υπάρχουν σημαντικά μειονεκτήματα μέχρι την τελική υιοθέτησή της, αφού θα πρέπει να περιμένουμε αρκετά μέχρι να συλλέξουμε το δείγμα αναφοράς, με αποτέλεσμα να καθυστερεί η έναρξη της παρακολούθησης της διεργασίας. Επίσης, δεν μπορούμε να γνωρίζουμε με μεγάλη βεβαιότητα πόσο πρέπει να είναι το n_1 (μέγεθος του δείγματος αναφοράς), ώστε η μη ανεξαρτησία να μην δημιουργεί προβλήματα στην απόδοση του διαγράμματος. Λόγω της μη ανεξαρτησίας των στατιστικών συναρτήσεων $MW_1, MW_2, \ldots$ θα υπάρχουν σημαντικές αποκλίσεις μεταξύ της ακριβούς τιμής της πιθανότητας εσφαλμένου συναγερμού και της επιθυμητής.

Οι Chakraborti and Van de Wiel (2008) ανέπτυξαν μεθόδους για τον υπολογισμό του ARL και έδωσαν πίνακες με τον στατιστικό σχεδιασμό ενός δίπλευρου διαγράμματος ελέγχου MW. Η αναλυτική επεξήγηση των τεχνικών και μεθόδων ξεφεύγει από τους σκοπούς του παρόντος συγγράμματος. Παρακάτω δίνονται οι εντολές στην `R`, όπου με χρήση προσομοίωσης είναι δυνατός ο υπολογισμός των τιμών των ορίων ελέγχου για δεδομένες τιμές n_1, n_2 και πιθανότητας εσφαλμένου συναγερμού a. Με τις συγκεκριμένες εντολές[1] είναι δυνατή η επιβεβαίωση των αποτελεσμάτων του πίνακα Table 3 στην εργασία των Chakraborti and Van de Wiel (2008).

```
> sims<-100 # number of simulation runs
> n1<-100 # reference sample size
> n2<-5 # test sample size
> FAR0<-0.0027 # false alarm rate for IC ARL=370.4=1/0.0027
> # starting values for U, L (control limits)
> j2<-0
```

[1] Οι εντολές αυτές αποτελούν μια εξαιρετικά απλή υλοποίηση της γενικής αλγοριθμικής διαδικασίας για τον υπολογισμό του εντός ελέγχου ARL του διαγράμματος ελέγχου MW. Για πιο αποδοτικές (και σαφώς ταχύτερες) μεθόδους υπολογισμού παραπέμπουμε στην εργασία των Chakraborti and Van de Wiel (2008).

```
> UMN<-qwilcox(1-FAR0/2,n2,n1)-j2
> LMN<-n1*n2-UMN
> listRL<-c()
> for(j in 1:sims){
+  j1<-1
+  x0<-runif(n1)
+  x1<-matrix(runif(n2),ncol=5)
+  MWtest<-wilcox.test(x1,x0,paired=FALSE)
+  MWs<-MWtest$statistic
+  while(MWs<=UMN&MWs>=LMN){
+    j1<-j1+1
+    x1<-matrix(runif(n2),ncol=5)
+    MWtest<-wilcox.test(x1,x0,paired=FALSE)
+    MWs<-MWtest$statistic
+  }
+ listRL[j]<-j1
+ }
> ARLin<-mean(listRL)
> while(ARLin>1/FAR0){
+  j2<-(j2+1)
+  UMN<-qwilcox(1-FAR0/2,n2,n1)-j2
+  LMN<-n1*n2-UMN
+  listRL<-c()
+  for(j in 1:sims){
+  j1<-1
+  x0<-runif(n1)
+  x1<-matrix(runif(n2),ncol=5)
+  MWtest<-wilcox.test(x1,x0,paired=FALSE)
+  MWs<-MWtest$statistic
+  while(MWs<=UMN&MWs>=LMN){
+      j1<-j1+1
+      x1<-matrix(runif(n2),ncol=5)
+      MWtest<-wilcox.test(x1,x0,paired=FALSE)
+      MWs<-MWtest$statistic
+    }
+    listRL[j]<-j1
+  }
+  ARLin<-mean(listRL)
+ }
> cat('Lmn:',LMN,' Umn:',UMN,' ARL:',mean(listRL),'\n')
```

Αξίζει να σταθούμε σε δύο σημεία των παραπάνω εντολών. Αρχικά, χρησιμοποιούμε ως αρχικές τιμές για το άνω όριο ελέγχου $U_{n_1 n_2}$ ποσοστιαία σημεία της κατανομής της στατιστικής συνάρτησης για τον έλεγχο των Mann-Whitney. Αυτό γίνεται με χρήση της συνάρτησης `qwilcox(1-FAR0,n2,n1)` (γραμμές κώδικα 7 και 8). Παρατηρήστε ότι, για να έχουμε τα ίδια αποτελέσματα με τους Chakraborti and Van de Wiel (2008), ακολουθώντας όμως και τον συμβολισμό που έχουμε χρησιμοποιήσει στο Κεφάλαιο 6, θα πρέπει να δώσουμε πρώτα το n_2 και μετά το n_1. Επίσης, η εφαρμογή του τεστ των Mann-Whitney στην `R` γίνεται με τη συνάρτηση `wilcox.test(x1,x0,paired=FALSE)`. Πάλι, για να έχουμε τον ίδιο συμβολισμό με αυτόν στο Κεφάλαιο 6 αλλά και για να ακολουθήσουμε τη μεθοδολογία των Chakraborti and Van de Wiel (2008), θα πρέπει να δώσουμε πρώτα το δείγμα από την ανάλυση Φάσης II (είναι το `x1`) και μετά το δείγμα αναφοράς (είναι το `x0`). Επίσης, πρέπει να χρησιμοποιήσουμε το όρισμα `paired='FALSE'`. Η τιμή της στατιστικής συνάρτησης προκύπτει με χρήση του `$statistic`. Δείτε π.χ. τις γραμμές 34 και 35 στον παραπάνω κώδικα.

Παράδειγμα 12.5. Χρησιμοποιώντας τα δεδομένα του Παραδείγματος 12.4, να αναπτυχθεί ένα διάγραμμα ελέγχου Mann-Whitney, υποθέτοντας ότι η επιθυμητή τιμή για το εντός ελέγχου ARL είναι 370.4.

Λύση Παραδείγματος 12.4. Για $n_1 = 100$, $n_2 = 5$, είτε χρησιμοποιώντας τις εντολές της R που δόθηκαν παραπάνω είτε χρησιμοποιώντας τον πίνακα Table 3 από την εργασία των Chakraborti and Van de Wiel (2008), έχουμε ότι $U_{n_1 n_2} = 431$ και $L_{n_1 n_2} = n_1 n_2 - U_{n_1 n_2} = 500 - 431 = 69$.

Στη συνέχεια, θα πρέπει να υπολογίσουμε τις τιμές των $MW_1, MW_2, \ldots, MW_{15}$ για καθένα από τα 15 δείγματα της ανάλυσης Φάσης ΙΙ. Αυτό σημαίνει πως θα πρέπει να διεξάγουμε 15 ελέγχους Mann-Whitney με το πρώτο δείγμα να είναι το δείγμα αναφοράς και το δεύτερο δείγμα να είναι καθένα από τα δείγματα που συλλέγονται κατά την ανάλυση Φάσης ΙΙ. Για παράδειγμα, αν το πρώτο δείγμα έχει τις τιμές 7.064, 37.74, 5.369, 47.485, 31.608 (βλ. Πίνακα 12.12), τότε στο κοινό διατεταγμένο δείγμα των $n_1 + n_2 = 105$ παρατηρήσεων το άθροισμα των τάξεων για τις τιμές του δεύτερου δείγματος είναι 332 (επιβεβαιώστε το!). Άρα, η πρώτη τιμή που απεικονίζεται στο διάγραμμα είναι η $MW = R_2 - n(n+1)/2 = 332 - 5 \cdot 6/2 = 317$. Με τον ίδιο τρόπο υπολογίζονται και οι τιμές των υπόλοιπων σημείων στο διάγραμμα.

Παρακάτω, για λόγους ευκολίας, παραθέτουμε τις σχετικές εντολές στην R. Αρχικά, εισάγουμε σε ένα διάνυσμα `x0` τις τιμές του δείγματος αναφοράς.

```
> # reference sample
> x0<-c(1.332,26.052,17.397,10.861,14.170,20.036,5.642,9.354,31.706,
+ 28.449,11.938,10.596,54.033,21.276,22.937,9.469,23.216,10.295,
+ 11.938,46.044,7.385,28.938,34.768,6.109,5.203,20.785,10.344,8.714,
+ 26.634,32.735,12.186,22.669,8.999,5.934,18.872,18.122,10.556,8.308,
+ 35.719,7.870,16.567,11.471,25.291,9.26,9.649,11.451,5.501,12.763,
+ 6.553,12.759,38.432,24.693,5.543,8.843,27.237,13.813,7.666,18.805,
+ 23.483,10.442,21.732,5.188,6.568,10.561,5.074,12.715,4.792,14.468,
+ 12.116,5.038,33.104,12.402,15.010,2.840,12.839,9.685,16.903,24.041,
+ 6.356,13.839,5.697,15.950,10.931,1.134,2.010,14.409,18.601,17.635,
+ 30.517,12.814,4.041,5.732,13.912,10.493,19.528,24.766,21.048,27.813,
+ 30.586,33.047)
```

Οι τιμές των δειγμάτων κατά την ανάλυση Φάσης ΙΙ εισάγονται σε μορφή πίνακα με 5 στήλες και 15 γραμμές (είναι ο `x1`).

```
> # test samples as matrix
> x1<-matrix(c(7.064,37.74,5.369,47.485,31.608,5.064,35.562,6.9,
+ 16.353,25.903,46.955,34.424,17.243,61.201,8.681,24.108,16.529,
+ 26.721,47.875,14.088,39.894,32.364,9.463,29.298,17.714,5.315,
+ 10.698,14.945,20.115,64.977,7.487,12.988,9.091,18.254,30.433,
+ 23.691,26.753,11.458,14.103,12.812,42.153,34.864,3.741,14.932,
+ 11.248,21.924,20.777,18.993,23.948,18.918,6.388,17.387,11.032,
+ 26.984,11.18,18.01,3.227,3.968,18.309,26.296,8.404,35.322,7.903,
+ 44.885,41.889,17.132,46.782,27.048,42.23,40.124,38.576,13.216,
+ 18.958,10.822,12.136),byrow=T,ncol=5)
```

Έπειτα, υπολογίζουμε τις τιμές MW_1, MW_2, ... , MW_{15}. Αυτό γίνεται με τη συνάρτηση `wilcox.test(x1,x0,paired=FALSE)` και η τιμή της στατιστικής συνάρτησης προκύπτει με χρήση του `$statistic`. Σε κάθε επανάληψη, η τιμή της στατιστικής συνάρτησης καταχωρίζεται στο διάνυσμα `listMW` και στο τέλος, αυτό περιέχει τις τιμές που απεικονίζονται στο διάγραμμα.

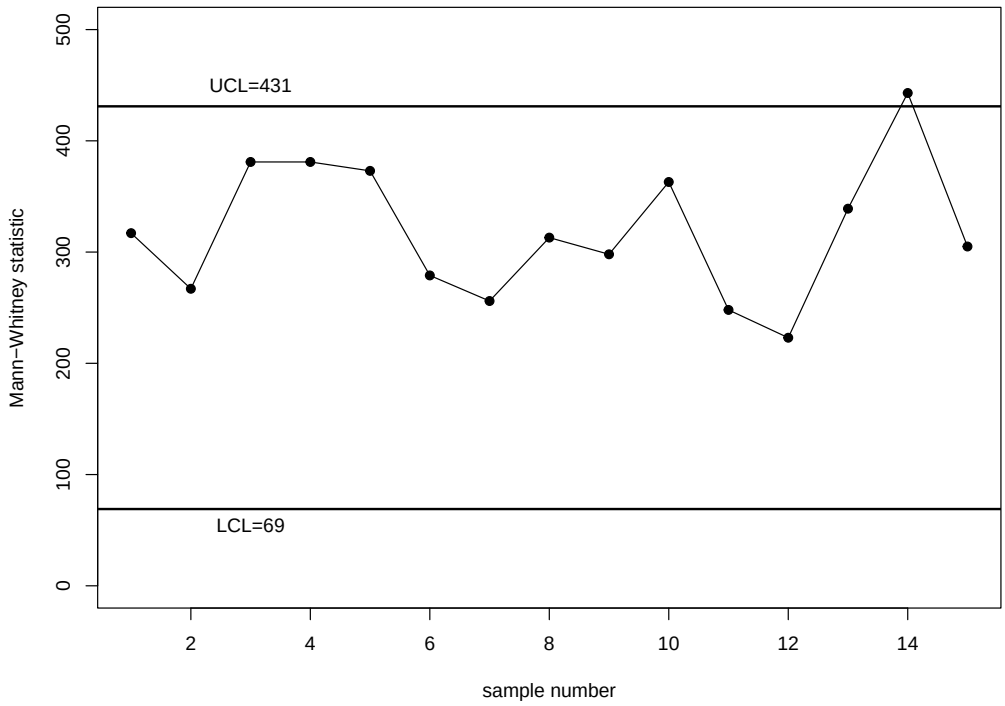
```
> # calculating the values of the test statistic
> listMW<-c()
> for(j0 in 1:dim(x1)[1]){
+   MWtest<-wilcox.test(x1[j0,],x0,paired=FALSE)
+   listMW[j0]<-MWtest$statistic # MW statistic, I(X2j<X1i)
+ }
```

```
> listMW # values of the test statistic
 [1] 317 267 381 381 373 279 256 313 298 363 248 223 339 443 305
```

Παρακάτω δίνονται οι εντολές για την κατασκευή του διαγράμματος με χρήση της R.

```
> UCL=431
> LCL=69
> plot(1:length(listMW),listMW,type='n',xlab='sample number',
+ ylab='Mann-Whitney statistic',ylim=c(0,500))
> points(1:length(listMW),listMW,pch=19)
> lines(1:length(listMW),listMW,lty=1,lwd=1)
> abline(h=UCL,lwd=2)
> abline(h=LCL,lwd=2)
> abline(v=20.5,lty=2)
> text(3,450,'UCL=431')
> text(3,55,'LCL=69')
```

Παρατηρούμε ότι το διάγραμμα δίνει για πρώτη φορά ένδειξη εκτός ελέγχου διεργασίας στο 14ο δείγμα από την έναρξη της Φάσης II.



Σχήμα 12.7: Διάγραμμα ελέγχου με χρήση του τεστ των Mann-Whitney για τα δεδομένα των Πινάκων 12.11 και 12.12.

□

12.7 Ασκήσεις

Για την επίλυση των παρακάτω ασκήσεων να χρησιμοποιηθεί η R.

Άσκηση 12.1. Τα παρακάτω δεδομένα (Πίνακας 12.13) αφορούν τις μετρήσεις σε χιλιοστά της διαμέτρου μεταλλικών στεφάνων και είναι γνωστά ως piston rings data (Montgomery, 2020). Χρησιμοποιήστε τις μετρήσεις από τα 25 πρώτα δείγματα και εκτιμήστε τη διάμεσο του πληθυσμού. Υποθέστε πως τα δείγματα προέρχονται από εντός ελέγχου διεργασία. Στη συνέχεια, για τα υπόλοιπα 15 δείγματα ως δεδομένα για ανάλυση Φάσης ΙΙ, κατασκευάστε το διάγραμμα ελέγχου SN-chart, χρησιμοποιώντας όρια ελέγχου $UCL = 5, LCL = -5$. Να ερμηνεύσετε την εικόνα του διαγράμματος.

A/A	X_1	X_2	X_3	X_4	X_5
1	74.030	74.002	74.019	73.992	74.008
2	73.995	73.992	74.001	74.011	74.004
3	73.988	74.024	74.021	74.005	74.002
4	74.002	73.996	73.993	74.015	74.009
5	73.992	74.007	74.015	73.989	74.014
6	74.009	73.994	73.997	73.985	73.993
7	73.995	74.006	73.994	74.000	74.005
8	73.985	74.003	73.993	74.015	73.988
9	74.008	73.995	74.009	74.005	74.004
10	73.998	74.000	73.990	74.007	73.995
11	73.994	73.998	73.994	73.995	73.990
12	74.004	74.000	74.007	74.000	73.996
13	73.983	74.002	73.998	73.997	74.012
14	74.006	73.967	73.994	74.000	73.984
15	74.012	74.014	73.998	73.999	74.007
16	74.000	73.984	74.005	73.998	73.996
17	73.994	74.012	73.986	74.005	74.007
18	74.006	74.010	74.018	74.003	74000
19	73.984	74.002	74.003	74.005	73.997
20	74.000	74.010	74.013	74.020	74.003
21	73.988	74.001	74.009	74.005	73.996
22	74.004	73.999	73.990	74.006	74.009
23	74.010	73.989	73.990	74.009	74.014
24	74.015	74.008	73.993	74.000	74.010
25	73.982	73.984	73.995	74.017	74.013
26	74.012	74.015	74.030	73.986	74.000
27	73.995	74.010	73.990	74.015	74.001
28	73.987	73.999	73.985	74.000	73.990
29	74.008	74.010	74.003	73.991	74.006
30	74.003	74.000	74.001	73.986	73.997
31	73.994	74.003	74.015	74.020	74.004
32	74.008	74.002	74.018	73.995	74.005
33	74.001	74.004	73.990	73.996	73.998
34	74.015	74.000	74.016	74.025	74.000
35	74.030	74.005	74.000	74.016	74.012
36	74.001	73.990	73.995	74.010	74.024
37	74.015	74.020	74.024	74.005	74.019
38	74.035	74.010	74.012	74.015	74.026
39	74.017	74.013	74.036	74.025	74.026
40	74.010	74.005	74.029	74.000	74.020

Πίνακας 12.13: Δεδομένα μετρήσεων διαμέτρου μεταλλικών στεφάνων (σε mm).

Άσκηση 12.2. Χρησιμοποιήστε τα δεδομένα του Πίνακα 12.13 και κατασκευάστε ένα διάγραμμα ελέγχου SR-chart με όρια ελέγχου $UCL = 15, LCL = -15$. Χρησιμοποιήστε τα 25 πρώτα δείγματα και εκτιμήστε τη διάμεσο του πληθυσμού. Υποθέστε ότι τα δείγματα αυτά προέρχονται από μια εντός ελέγχου διεργασία. Στη

συνέχεια, χρησιμοποιήστε τα υπόλοιπα 15 δείγματα ως δεδομένα ανάλυσης Φάσης ΙΙ. Να ερμηνεύσετε την εικόνα του διαγράμματος.

Άσκηση 12.3. Υποθέστε ότι τα 25 πρώτα δείγματα του Πίνακα 12.13 αποτελούν ένα δείγμα αναφοράς και τα υπόλοιπα 15 είναι δείγματα ελέγχου. Να κατασκευάσετε:

(i) το διάγραμμα ελέγχου Προηγήσεων με όρια ελέγχου τα οποία προκύπτουν από την 7η και την 119η διατεταγμένη παρατήρηση στο δείγμα αναφοράς και

(ii) το διάγραμμα ελέγχου MW με όρια ελέγχου $LCL = 85$, $UCL = 540$.

Να ερμηνεύσετε τις εικόνες των διαγραμμάτων.

Άσκηση 12.4. Να αναπτύξετε ένα άνω μονόπλευρο διάγραμμα ελέγχου V-chart για την παρακολούθηση της διασποράς μιας διεργασίας χρησιμοποιώντας τα δεδομένα του Πίνακα 12.14. Υποθέστε ότι τα 20 πρώτα δείγματα προέρχονται από εντός ελέγχου διεργασία και εκτιμήστε το 1ο και το 3ο τεταρτημόριο της κατανομής των δεδομένων. Στη συνέχεια, υποθέστε πως τα υπόλοιπα 15 δείγματα αποτελούν δεδομένα ανάλυσης Φάσης ΙΙ. Κατασκευάστε το διάγραμμα και ερμηνεύστε την εικόνα του. Ως άνω όριο ελέγχου να χρησιμοποιηθεί η τιμή 9.

A/A	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	0.601	3.513	4.241	2.601	1.360	2.559	2.825	3.210	2.514	2.510
2	0.190	1.227	1.687	0.723	2.342	0.472	3.585	0.962	1.778	1.331
3	5.663	1.074	2.558	0.983	1.112	3.378	3.454	14.019	0.083	3.925
4	3.120	8.324	0.552	2.117	1.738	1.210	0.543	0.588	2.251	2.569
5	0.457	0.928	0.383	1.321	0.368	2.781	0.316	0.687	2.693	0.664
6	0.045	1.125	0.013	0.863	1.031	4.294	0.120	3.585	0.881	0.283
7	0.085	0.345	5.044	1.364	10.326	2.588	2.641	1.683	0.264	0.626
8	3.476	0.458	3.218	0.087	1.533	1.730	0.637	2.075	1.064	0.879
9	0.654	0.198	0.907	5.995	0.249	1.144	0.071	1.028	3.908	0.003
10	3.160	1.012	0.210	6.795	0.463	1.730	1.632	0.057	1.273	1.740
11	1.865	1.894	2.880	0.883	1.061	1.228	2.657	0.633	0.033	1.028
12	0.357	5.650	5.088	0.228	5.028	0.454	1.669	0.111	0.280	1.982
13	0.802	0.894	5.423	1.548	0.569	7.971	1.517	0.030	0.576	0.512
14	0.392	1.298	0.795	1.322	1.252	4.212	0.965	1.020	0.290	3.476
15	3.074	0.697	2.055	0.261	2.439	6.982	6.234	6.395	0.297	0.075
16	8.371	1.987	0.805	1.122	0.814	0.250	12.518	0.881	1.459	0.034
17	4.671	0.229	3.803	0.153	0.575	0.432	2.548	6.286	0.370	0.081
18	3.452	2.787	2.916	0.495	0.343	2.247	0.599	1.393	0.899	4.100
19	9.270	0.639	1.896	2.090	0.534	0.211	1.043	1.992	2.491	0.773
20	0.146	0.412	1.926	2.574	0.881	2.285	2.684	4.132	0.648	1.242
21	5.518	1.152	6.429	0.120	3.000	0.582	0.335	4.138	2.109	2.946
22	0.520	1.842	1.127	0.242	13.794	6.568	6.160	3.085	0.734	0.417
23	1.296	0.337	5.765	2.121	4.230	1.736	2.900	8.495	3.518	2.612
24	2.248	8.213	3.480	2.776	5.165	1.652	0.876	2.002	1.318	5.314
25	0.657	3.601	5.542	0.084	1.317	2.905	0.779	1.163	4.893	1.144
26	3.453	1.083	2.613	4.576	7.363	8.739	14.035	0.588	0.942	2.034
27	8.242	17.510	7.059	4.445	2.959	1.659	10.012	6.635	0.235	3.954
28	1.530	9.623	10.863	0.365	0.107	0.613	0.576	1.212	3.664	2.775
29	2.941	1.417	3.794	7.907	2.787	3.876	0.163	2.417	0.257	3.383
30	0.790	5.695	4.518	3.388	2.796	1.128	1.531	0.012	5.083	4.698
31	12.674	0.995	2.474	3.711	3.872	8.827	0.533	1.239	0.287	0.064
32	2.077	2.446	0.853	0.650	0.757	1.638	2.179	6.194	0.400	7.986
33	4.568	4.351	8.684	1.395	0.109	0.557	0.196	0.227	6.263	0.483
34	4.559	2.557	0.164	4.636	4.197	2.183	1.064	8.941	0.884	5.246
35	0.840	1.092	0.557	3.420	1.968	3.780	3.347	2.869	0.523	0.105

Πίνακας 12.14: Δεδομένα Άσκησης 12.4.

Άσκηση 12.5. Να επαναλάβετε την Άσκηση 12.4, όμως, αυτήν τη φορά, χρησιμοποιήστε τα 20 πρώτα δείγματα και εκτιμήστε τα 0.2 και 0.8 ποσοστιαία σημεία της κατανομής του χαρακτηριστικού X. Στη συνέχεια, εφαρμόστε το Προσημικό κριτήριο της Ενότητας 12.4, όπου αντί για το 1ο και το 3ο τεταρτημόριο, να χρησιμοποιηθούν τα ποσοστιαία σημεία $x_{0.8}$ και $x_{0.2}$ που μόλις υπολογίσατε. Κατασκευάστε το διάγραμμα και ερμηνεύστε την εικόνα του. Ως άνω όριο ελέγχου να χρησιμοποιηθεί η τιμή 9.

Άσκηση 12.6. (**Διαγράμματα Ελέγχου τύπου EWMA**) Στα διαγράμματα ελέγχου εκθετικά σταθμισμένου κινητού μέσου (Exponentially Weighted Moving Average control charts) απεικονίζονται τιμές της στατιστικής συνάρτησης

$$Z_i = \lambda W_i' + (1-\lambda)Z_{i-1},$$

όπου η παράμετρος $\lambda \in (0,1]$ είναι γνωστή ως παράμετρος εξομάλυνσης, ενώ η αρχική τιμή Z_0 ισούται με την εντός ελέγχου μέση τιμή $\mathrm{E}(W_i')$ της σ.σ. W_i'. Οι τιμές $Z_1, Z_2, \ldots$ που απεικονίζονται στο διάγραμμα συγκρίνονται με όρια ελέγχου τα οποία δίνονται από τις σχέσεις:

$$LCL = \mathrm{E}(W_i') - \mathrm{A}\sqrt{\frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2i})\mathrm{Var}(W_i')},$$

$$UCL = \mathrm{E}(W_i') - \mathrm{A}\sqrt{\frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2i})\mathrm{Var}(W_i')},$$

ενώ η κεντρική γραμμή είναι ίση με $CL = \mathrm{E}(W_i')$. Οι τιμές των ορίων ελέγχου υπολογίζονται όταν η διεργασία είναι εντός ελέγχου. Παρατηρήστε, επίσης, ότι εξαρτώνται από το δείγμα i. Τα όρια αυτά είναι γνωστά ως *χρονομεταβλητά όρια ελέγχου* (time-varying control limits).

Να κατασκευάσετε ένα διάγραμμα ελέγχου SN-EWMA για τα δεδομένα του Πίνακα 12.13 χρησιμοποιώντας ως W_i' τη στατιστική συνάρτηση SN_i. Υπενθυμίζεται πως, όταν η διεργασία είναι εντός στατιστικού ελέγχου, η $\mathrm{E}(SN_i) = 0$ και η $\mathrm{Var}(SN_i) = n/2$. Για το ζεύγος τιμών (λ, A) χρησιμοποιήστε τις τιμές $(0.05, 2.5)$ και ερμηνεύστε την εικόνα του διαγράμματος.

Άσκηση 12.7. Να επαναλάβετε την Άσκηση 12.6, όμως, αυτήν τη φορά, να αναπτύξετε ένα διάγραμμα ελέγχου τύπου EWMA χρησιμοποιώντας ως W_i' τη στατιστική συνάρτηση SR_i. Υπενθυμίζεται πως όταν η διεργασία είναι εντός στατιστικού ελέγχου, η $\mathrm{E}(SR_i) = 0$ (αφού η $\mathrm{E}(W_i^+) = n(n+1)/2$ και η $\mathrm{Var}(SR_i) = 4\mathrm{Var}(W_i^+)$, όπου η $\mathrm{Var}(W_i^+) = n(n+1)(2n+1)/6$ (βλ. Κεφάλαιο 6, Ενότητα 6.2). Για το ζεύγος τιμών (λ, A) χρησιμοποιήστε τις τιμές $(0.05, 2.5)$ και ερμηνεύστε την εικόνα του διαγράμματος.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Αντζουλάκος, Δ. (2003). *Στατιστικός Έλεγχος Ποιότητας* (1η εκδ.). Πειραιάς: Πανεπιστημιακές Σημειώσεις.

Μπερσίμης, Σ., Ρακιτζής, Α. και Σαχλάς, Α. (2021). *Στατιστικός Έλεγχος Ποιότητας* (1η εκδ.). Θεσσαλονίκη: Τζιόλας.

Ταγαράς, Γ. (2001). *Στατιστικός Έλεγχος Ποιότητας*(1η εκδ.). Θεσσαλονίκη: ΖΗΤΗ.

Ξενόγλωσση

Amin, R. W., Reynolds Jr, M. R. and Saad, B. (1995). Nonparametric quality control charts based on the sign statistic. *Communications in Statistics-Theory and Methods*, *24*(6), pp. 1597–1623.

Bakir, S. T. (2004). A distribution-free Shewhart quality control chart based on signed-ranks. *Quality Engineering*, *16*(4), pp. 613–623.

Balakrishnan, N. and Ng, H. T. (2006). *Precedence-type tests and applications*. John Wiley & Sons.

Bradley, J. V. (1968). *Distribution-free statistical tests*. Prentice-Hall.

Chakraborti, S. and Van der Laan, P. (1996). Precedence tests and confidence bounds for complete data: an overview and some results. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *45*(3), pp. 351–369.

Chakraborti, S. and Graham, M. (2019). *Nonparametric Statistical Process Control*. John Wiley and Sons.

Chakraborti, S. and Van de Wiel, M. A. (2008). A nonparametric control chart based on the Mann-Whitney statistic. In: *Beyond parametrics in interdisciplinary research: Festschrift in Honor of Professor Pranab K. Sen*. Ed. by N. Balakrishnan, E. A. Pena and M. J. Silvapulle. Institute of Mathematical Statistics, pp. 156–172.

Chakraborti, S., Van der Laan, P. and Van de Wiel, M. A. (2004). A class of distribution-free control charts. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *53*(3), pp. 443–462.

Chakraborti, S. and Graham, M. A. (2014). Control charts, nonparametric. *Wiley StatsRef: Statistics Reference Online*.

Janacek, G. J. and Meikle, S. E. (1997). Control charts based on medians. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *46*(1), pp. 19–31.

Mathisen, H. C. (1943). A method of testing the hypothesis that two samples are from the same population. *The Annals of Mathematical Statistics*, *14*(2), pp. 188–194.

Montgomery, D. (2020). *Introduction to Statistical Quality Control*. John Wiley & Sons.

Nelson, L. S. (1963). Tables for a precedence life test. *Technometrics*, *5*(4), pp. 491–499.

Nelson, L. S. (1993). Tests on early failures—the precedence life test. *Journal of Quality Technology*, *25*(2), pp. 140–143.

ΚΕΦΑΛΑΙΟ 13

ΥΛΟΠΟΙΗΣΗ ΜΗ ΠΑΡΑΜΕΤΡΙΚΩΝ ΣΤΑΤΙΣΤΙΚΩΝ ΤΕΧΝΙΚΩΝ ΣΤΟ SPSS ΚΑΙ ΣΤΗΝ R

Σύνοψη

Σημαντικό ρόλο στην εξέλιξη (αλλά και τη δημοφιλία) των στατιστικών τεχνικών, παραμετρικών ή μη παραμετρικών, έχει διαδραματίσει η χρήση των ηλεκτρονικών υπολογιστών. Στο κεφάλαιο αυτό, παρουσιάζεται η υλοποίηση των κυριότερων μη παραμετρικών στατιστικών μεθοδολογιών με χρήση του στατιστικού προγράμματος (πακέτου), του SPSS, καθώς και της γλώσσας στατιστικού προγραμματισμού R. Στόχος είναι να παρουσιαστούν οι βασικές μη παραμετρικές τεχνικές που αποτέλεσαν αντικείμενο μελέτης σε προηγούμενα κεφάλαια αυτού του συγγράμματος και οι οποίες μπορούν να υλοποιηθούν και με τα δύο προγράμματα. Επομένως, στόχος του κεφαλαίου δεν είναι να γίνει πλήρης παρουσίαση όλων των δυνατοτήτων των δύο αυτών προγραμμάτων.

Προαπαιτούμενη γνώση:

Βασικές γνώσεις SPSS και R.

Κεφάλαια 4, 5, 6, 7 και 8 του παρόντος συγγράμματος.

Προσδοκώμενα μαθησιακά αποτελέσματα:

Ο/η φοιτητής/τρια ολοκληρώνοντας την ενότητα αυτή θα μπορεί να εφαρμόζει σε πρακτικά προβλήματα τις κατάλληλες μη παραμετρικές στατιστικές τεχνικές, είτε με χρήση του SPSS είτε με χρήση της R.

Γλωσσάριο επιστημονικών όρων

- SPSS
- R
- Διωνυμικός έλεγχος
- Έλεγχος Friedman
- Έλεγχος Kruskal-Wallis
- Έλεγχος Mann-Whitney
- Έλεγχος McNemar
- Έλεγχος Wilcoxon
- Έλεγχος Cochran's Q
- Έλεγχος Jonckheere–Terpstra

Μπατσίδης, Α., Παπασταμούλης, Π., Πετρόπουλος, Κ., & Ρακιτζής, Α. (2022). *Μη Παραμετρική Στατιστική.* [Προπτυχιακό εγχειρίδιο].

- Έλεγχος καλής προσαρμογής Kolmogorov-Smirnov
- Έλεγχος καλής προσαρμογής χ^2
- Έλεγχος καλής προσαρμογής Lilliefors για εκθετική κατανομή
- Έλεγχος κανονικότητας Anderson-Darling
- Έλεγχος κανονικότητας Lilliefors
- Έλεγχος κανονικότητας Shapiro-Wilk
- Έλεγχος τυχαιότητας runs test
- Πολλαπλές συγκρίσεις
- Προσημικός έλεγχος
- Συντελεστής συσχέτισης Spearman
- Συντελεστής συσχέτισης Kendall

13.1 Εισαγωγή

Τα στατιστικά προγράμματα (ή στατιστικά πακέτα) αποτελούν, πλέον, αναπόσπαστο κομμάτι της στατιστικής ανάλυσης δεδομένων. Λόγω του όγκου των δεδομένων που συλλέγονται, δεν είναι δυνατή η εφαρμογή μεθόδων στατιστικής συμπερασματολογίας «με το χέρι». Κατά συνέπεια, θα πρέπει να χρησιμοποιηθεί ηλεκτρονικός υπολογιστής και συγκεκριμένο κατάλληλο λογισμικό, το οποίο παρέχει τη δυνατότητα διεξαγωγής εξειδικευμένων στατιστικών αναλύσεων. Ως συνέπεια αυτού, έχουν αναπτυχθεί στατιστικά πακέτα (δηλαδή λογισμικό για Η/Υ) τα οποία μπορούν να χρησιμοποιηθούν για να κάνουμε διαφόρων τύπων στατιστικές αναλύσεις, μεταξύ αυτών και των μη παραμετρικών ελέγχων που παρουσιάστηκαν στα προηγούμενα κεφάλαια του παρόντος συγγράμματος. Μεταξύ όλων των διαθέσιμων στατιστικών πακέτων, στο σύγγραμμα αυτό επιλέχθηκε να υλοποιηθούν οι κυριότερες μη παραμετρικές μεθοδολογίες των προηγούμενων κεφαλαίων με χρήση του SPSS και της R, για τα οποία παρατίθενται κάποιες χρήσιμες πληροφορίες στη συνέχεια. Πρέπει, επίσης, να σημειώσουμε ότι σε αυτό το κεφάλαιο δεν θα παρουσιάσουμε τον τρόπο εκτίμησης της α.σ.κ. ή/και της σ.π.π., όπως και τις μεθόδους jackknife, bootstrap και της μη παραμετρικής παλινδρόμησης, καθώς η υλοποίηση αυτών των μεθόδων με χρήση της R παρουσιάζεται στα αντίστοιχα κεφάλαια.

Το SPSS (Statistical Package for the Social Sciences) είναι ένα από τα πιο γνωστά στατιστικά προγράμματα. Πρόκειται για ένα πρόγραμμα με μεγάλες δυνατότητες στατιστικής επεξεργασίας και ανάλυσης που δουλεύει σε περιβάλλον Windows, MacOS και Linux. Το SPSS αναπτύχθηκε το 1968 από τους Norman H. Nie, Dale H. Bent και C. Hadlai Hull. Από την 17η έκδοσή του και έπειτα μετονομάστηκε σε PASW Statistics (Predictive Analytics SoftWare Statistics), ενώ ταυτόχρονα επανασχεδιάστηκε, προσφέροντας αναβαθμισμένο γραφικό περιβάλλον και εμπλουτίστηκε με νέες δυνατότητες, τόσο κατά τη διαχείριση, όσο και κατά την ανάλυση των δεδομένων. Από τον Αύγουστο του 2010 είναι γνωστό ως IBM SPSS Statistics (με τον τίτλο αυτό κυκλοφόρησε η 19η έκδοση του προγράμματος), αφού είχε ήδη εξαγοραστεί από την IBM. Πλέον, βρίσκεται στην 28η έκδοσή του. Στο παρόν κεφάλαιο θα δούμε τις δυνατότητες της 27ης έκδοσης του προγράμματος, ως προς την εφαρμογή μη παραμετρικών στατιστικών τεχνικών και μεθόδων. Η 27η έκδοση κυκλοφόρησε τον Ιούνιο του 2019. Επισημαίνεται ότι το SPSS απαιτεί την αγορά άδειας χρήσης ώστε να μπορούμε να το χρησιμοποιούμε.

Η γλώσσα προγραμματισμού R ή το στατιστικό πακέτο R είναι ένα ισχυρό υπολογιστικό εργαλείο κατάλληλο για συγγραφή προγραμμάτων με σκοπό την επίλυση στατιστικών προβλημάτων. Αναπτύχθηκε από τους Ross Ihaka και Robert Gentleman από το Πανεπιστήμιο του Auckland στη Νέα Ζηλανδία, στις αρχές της δεκαετίας του 1990. Βασίστηκε στη γλώσσα προγραμματισμού S, η οποία είναι ενσωματωμένη στο εμπορικό πρόγραμμα S-Plus. Για πιο εύκολη χρήση της R υπάρχει η δυνατότητα εγκατάστασης και ενός Ενσωματωμένου Αναπτυξιακού Περιβάλλοντος (Integrated Development Environment, IDE), όπως το R Studio (https://rstudio.com/products/rstudio/download/). Βασικό πλεονέκτημα της R είναι η δωρεάν διάθεσή της μέσα από την ιστοσελίδα http://www.r-project.org. Όποιος/α επιθυμεί μπορεί να τη μεταφορτώσει στον Η/Υ ή στο laptop του/της και να τη χρησιμοποιήσει δωρεάν. Επίσης, χρησιμοποιείται και υποστηρίζεται από έναν πολύ μεγάλο αριθμό ερευνητών παγκοσμίως. Συνεχώς αναπτύσσονται πακέτα (packages) που ενσωματώνουν ποικίλες λειτουργίες, στατιστικές τεχνικές και αλγορίθμους. Πλέον, κάθε καινούρια στατιστική μεθοδολογία, που προτείνεται, συνοδεύεται και από το αντίστοιχο πακέτο στην R. Αυτήν τη στιγμή στην R είναι διαθέσιμα περισσότερα από 10.000 το πλήθος πακέτα. Στο παρόν κεφάλαιο θα δούμε τις δυνατότητες της R (ως προς την εφαρμογή μη παραμετρικών στατιστικών τεχνικών και μεθόδων). Θα χρησιμοποιήσουμε την έκδοση 4.1.2, η οποία κυκλοφόρησε τον Νοέμβριο του 2021.

Από τα παραπάνω, είναι προφανές ότι η επιλογή αυτών των στατιστικών πακέτων έγινε λόγω της δημοφιλίας των δύο αυτών προγραμμάτων, καθώς το SPSS χρησιμοποιείται από πληθώρα επιστημόνων (όχι μόνο μαθηματικών ή/και στατιστικών) για τη διεξαγωγή στατιστικών αναλύσεων, ενώ το ίδιο συμβαίνει

και με την R, η οποία όμως έχει το ιδιαίτερο χαρακτηριστικό ότι είναι διαθέσιμη δωρεάν. Στο πλαίσιο αυτού του συγγράμματος δεν θα δοθούν πληροφορίες για τις βασικές λειτουργίες αυτών των δύο προγραμμάτων. Για εισαγωγικές έννοιες αλλά και για μια πιο πλήρη παρουσίαση του SPSS ο/η ενδιαφερόμενος/μενη αναγνώστης/στρια παραπέμπεται στα συγγράμματα των Σαχλάς και Μπερσίμης (2016) και Χαλικιάς κ.ά. (2015), ενώ για την R αναφέρουμε ενδεικτικά τα συγγράμματα των Ντζούφρας και Καρλής (2015), Φουσκάκης (2021) και τις Πανεπιστημιακές σημειώσεις των Φωκιανός και Χαραλάμπους (2010) και Αντζουλάκος (2013).

Στη συνέχεια αυτού του κεφαλαίου γίνεται παρουσίαση των πιο βασικών μη παραμετρικών στατιστικών τεχνικών, οι οποίες μπορούν να υλοποιηθούν και με τα δύο προγράμματα. Κατά συνέπεια, δεν παρουσιάζονται τεχνικές που είναι διαθέσιμες σε ένα μόνο από τα δύο προγράμματα. Οι τεχνικές παρουσιάζονται μέσα από παραδείγματα σε προβλήματα που μπορούν να θεωρηθούν παρόμοια με αυτά που συναντώνται στην πράξη. Πρέπει, επίσης, να σημειωθεί ότι όλα τα δεδομένα είναι προσομοιωμένα και δεν προέρχονται από πραγματικά προβλήματα.

13.2 Έλεγχοι καλής προσαρμογής

Σκοπός αυτής της ενότητας είναι η υλοποίηση, μέσω παραδειγμάτων, με χρήση του SPSS και της R των κυριότερων ελέγχων καλής προσαρμογής που παρουσιάστηκαν στο Κεφάλαιο 4 του παρόντος συγγράμματος.

13.2.1 Έλεγχοι καλής προσαρμογής: χι-τετράγωνο και Kolmogorov-Smirnov

Στην υποενότητα αυτή θα δούμε πώς υλοποιούνται ο χι-τετράγωνο ελέγχος καλής προσαρμογής, που αποτελεί τον αρχαιότερο τέτοιο έλεγχο, καθώς και ο έλεγχος Kolmogorov-Smirnov που αξιοποιεί τις ιδιότητες της εμπειρικής αθροιστικής συνάρτησης (βλ. Ενότητα 4.2 και Ενότητα 4.3, αντίστοιχα).

Παράδειγμα 13.1. (**One-sample Kolmogorov-Smirnov test**): Στον Πίνακα 13.1 δίνονται οι μετρήσεις ενός τυχαίου δείγματος 40 τιμών από πληθυσμό με άγνωστη κατανομή. Χρησιμοποιώντας τον έλεγχο των Kolmogorov-Smirnov να ελέγξετε σε ε.σ. 10% την υπόθεση ότι τα δεδομένα προέρχονται από την κανονική κατανομή $\mathcal{N}(12, 4.9^2)$.

A/A	X	A/A	X	A/A	X	A/A	X	A/A	X
1	12.81	9	22.95	17	8.74	25	15.78	33	13.87
2	14.56	10	19.44	18	14.18	26	16.83	34	8.73
3	18.76	11	10.19	19	9.09	27	8.98	35	17.9
4	14.51	12	14.91	20	10.43	28	14.14	36	8.82
5	19.37	13	8.51	21	12.86	29	13.78	37	23.77
6	20.13	14	3.47	22	11.63	30	16.42	38	11.16
7	6.51	15	9.34	23	16.61	31	9.74	39	16.18
8	10.82	16	14.26	24	5.9	32	9.28	40	14.69

Πίνακας 13.1: Τυχαίο δείγμα 40 τιμών από πληθυσμό με άγνωστη κατανομή.

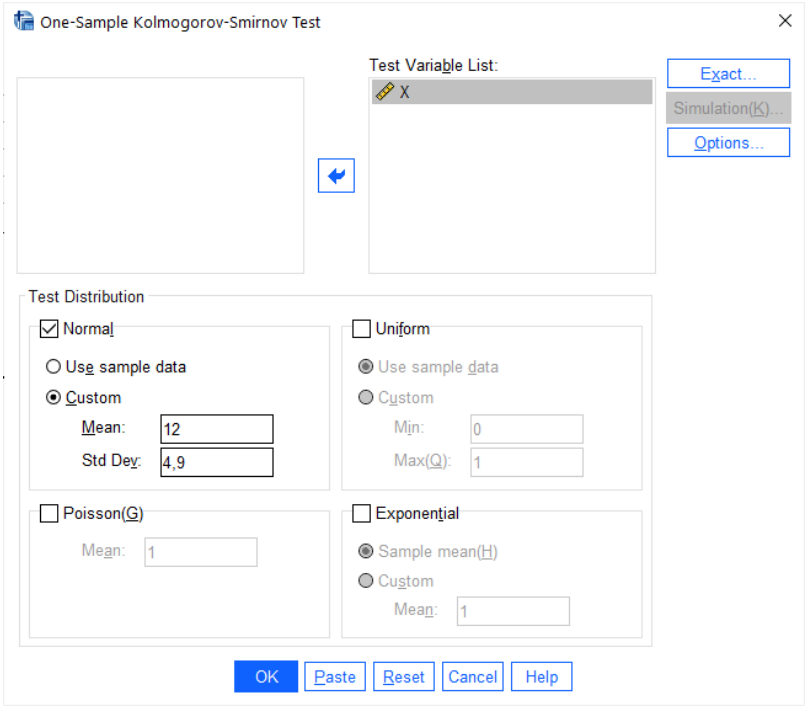
Λύση Παραδείγματος 13.1. Από τα δεδομένα της εκφώνησης άμεσα προκύπτει ότι έχουμε έναν απλό έλεγχο καλής προσαρμογής, καθώς υπό τη μηδενική υπόθεση η κατανομή είναι πλήρως καθορισμένη, χωρίς άγνωστες παραμέτρους. Θα χρησιμοποιηθεί ο έλεγχος των Kolmogorov-Smirnov που εκτενώς παρουσιάστηκε στην Ενότητα 4.3.1.

(με χρήση SPSS): Θα χρησιμοποιήσουμε το SPSS και θα εφαρμόσουμε το τεστ των Kolmogorov-Smirnov. Αρχικά, εισάγουμε σε μια στήλη ενός κενού φύλλου εργασίας του SPSS τα παραπάνω δεδομένα και μετονομάζουμε τη στήλη σε X.

Στη συνέχεια, επιλέγουμε από το κεντρικό παράθυρο διαλόγου:

Analyze / Nonparametric Tests / Legacy Dialogs / 1-Sample K-S

Στο νέο παράθυρο διαλόγου που προκύπτει (βλ. Εικόνα 13.1) εισάγουμε στο πεδίο Test Variable List τη X, επιλέγουμε Test Distribution: Normal και, πατώντας το Custom, δίνουμε στο Mean την τιμή 12 και στο Std Dev την τιμή 4.9, καθώς αυτή είναι η κανονική την οποία θέλουμε να ελέγξουμε υπό τη μηδενική υπόθεση. Πατάμε OK και προκύπτει το output της ανάλυσης, το οποίο δίνεται στην Εικόνα 13.2.



Εικόνα 13.1: Παράθυρο διαλόγου για έλεγχο One-Sample Kolmogorov-Smirnov Test.

Στο output της ανάλυσης δίνεται η τιμή της στατιστικής συνάρτησης ελέγχου Z (ασυμπτωτική μορφή του τεστ), ενώ η p-τιμή είναι ίση με 0.216. Συμπεραίνουμε ότι δεν μπορούμε να απορρίψουμε την H_0 και άρα δεν μπορούμε να απορρίψουμε την υπόθεση ότι τα δεδομένα προέρχονται από πληθυσμό με κατανομή $\mathcal{N}(12, 4.9^2)$.

(με χρήση R): Θα χρησιμοποιήσουμε την `R` και θα εφαρμόσουμε το τεστ των Kolmogorov-Smirnov. Αρχικά, εισάγουμε σε ένα διάνυσμα, έστω αυτό `x`, τις διαθέσιμες μετρήσεις. Στη συνέχεια, χρησιμοποιούμε την εντολή `ks.test(...)`, όπως φαίνεται παρακάτω.

One-Sample Kolmogorov-Smirnov Test

		X
N		40
Normal Parameters[a,b]	Mean	12
	Std. Deviation	4,9
Most Extreme Differences	Absolute	,167
	Positive	,008
	Negative	-,167
Kolmogorov-Smirnov Z		1,055
Asymp. Sig. (2-tailed)		,216

a. Test distribution is Normal.

b. User-Specified

Εικόνα 13.2: Αποτελέσματα ελέγχου One-sample Kolmogorov-Smirnov για τα δεδομένα του Πίνακα 13.1.

```
> x<-c(12.81,22.95,8.74,15.78,13.87,14.56,19.44,14.18,16.83,8.73,
+ 18.76,10.19,9.09,8.98,17.9,14.51,14.91,10.43,14.14,8.82,19.37,
+ 8.51,12.86,13.78,23.77,20.13,3.47,11.63,16.42,11.16,6.51,9.34,
+ 16.61,9.74,16.18,10.82,14.26,5.9,9.28,14.69)
> ks.test(x, 'pnorm', 12, 4.9, exact = FALSE)
```

Για τη χρήση της `ks.test` στο συγκεκριμένο πρόβλημα, εισάγουμε το διάνυσμα `x` με τις διαθέσιμες μετρήσεις, ενώ χρησιμοποιώντας το όρισμα `'pnorm'` δηλώνουμε ότι θέλουμε να κάνουμε έλεγχο καλής προσαρμογής της κανονικής κατανομής. Επίσης, καθώς θέλουμε να ελέγξουμε την προσαρμογή των δεδομένων σε συγκεκριμένη κανονική κατανομή, δίνουμε τη μέση τιμή μ και την τυπική απόκλιση σ (οι τιμές 12 και 4.9, αντίστοιχα). Επίσης, χρησιμοποιώντας το όρισμα `exact=FALSE` ζητάμε από την `R` να κάνει τον έλεγχο με χρήση της ασυμπτωτικής κατανομής της στατιστικής συνάρτησης ελέγχου (σ.σ.ε.). Παρακάτω δίνουμε το output της ανάλυσης.

```
        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.1668, p-value = 0.2157
alternative hypothesis: two-sided
```

Από το output της ανάλυσης στην `R` έχουμε ότι η τιμή της σ.σ.ε. είναι 0.1668, ίδια με αυτήν που δίνει το SPSS, στη γραμμή Most Extreme Differences Absolute. Επίσης, η p-τιμή ισούται με 0.2157 (ίδια με αυτήν που δίνει το SPSS). Αφού η p-τιμή είναι $0.2157 > 0.10$ δεν μπορούμε να απορρίψουμε την H_0 και άρα σε ε.σ. 10% δεν μπορούμε να απορρίψουμε την υπόθεση ότι τα δεδομένα προέρχονται από $\mathcal{N}(12, 4.9^2)$.

Αξίζει, επίσης, να αναφέρουμε ότι, αν χρησιμοποιήσουμε το όρισμα `exact=TRUE`, τότε η `R` κάνει τον έλεγχο με χρήση της ακριβούς κατανομής της σ.σ.ε. D_n. Παρακάτω δίνουμε την εντολή μαζί με το output της ανάλυσης. Σε αυτήν την περίπτωση η p-τιμή είναι $0.1927 > 0.10$ και άρα, δεν έχουμε σαφείς ενδείξεις για να απορρίψουμε την υπόθεση ότι τα δεδομένα προέρχονται από $\mathcal{N}(12, 4.9^2)$.

```
> x<-c(12.81,22.95,8.74,15.78,13.87,14.56,19.44,14.18,16.83,8.73,
+ 18.76,10.19,9.09,8.98,17.9,14.51,14.91,10.43,14.14,8.82,19.37,
+ 8.51,12.86,13.78,23.77,20.13,3.47,11.63,16.42,11.16,6.51,9.34,
+ 16.61,9.74,16.18,10.82,14.26,5.9,9.28,14.69)
> ks.test(x, 'pnorm', 12, 4.9, exact = TRUE)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.1668, p-value = 0.1927
alternative hypothesis: two-sided
```

□

Παράδειγμα 13.2. (**One-sample Kolmogorov-Smirnov test, Exponential distribution**): Στον Πίνακα 13.2 δίνονται οι μετρήσεις ενός τυχαίου δείγματος 35 τιμών από πληθυσμό με άγνωστη κατανομή. Χρησιμοποιώντας τον έλεγχο των Kolmogorov-Smirnov να ελέγξετε σε ε.σ. 10%. την υπόθεση ότι τα δεδομένα προέρχονται από πληθυσμό που μοντελοποιείται σύμφωνα με το πιθανοτικό πρότυπο της εκθετικής κατανομής.

1.98	5.08	0.26	8.11	4.04
0.73	2.05	2.92	0.27	3.81
4.03	3.31	0.61	3.25	5.35
2.40	1.97	0.77	0.22	2.99
0.44	3.74	0.54	2.76	2.35
1.05	6.12	9.00	10.76	2.32
6.98	0.19	1.33	0.87	0.34

Πίνακας 13.2: Τυχαίο δείγμα 35 τιμών.

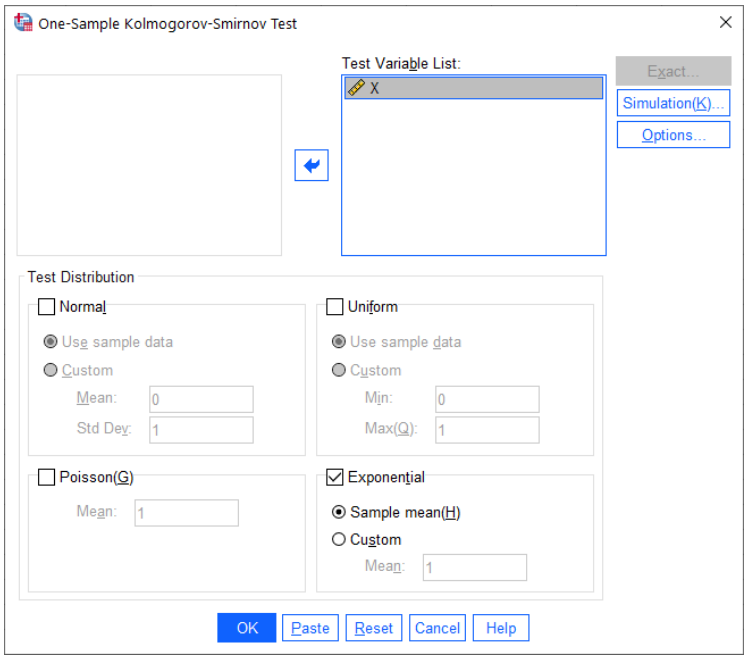
Λύση Παραδείγματος 13.2. Από τα δεδομένα της εκφώνησης άμεσα προκύπτει ότι έχουμε έναν σύνθετο έλεγχο καλής προσαρμογής καθώς υπό τη μηδενική υπόθεση η κατανομή δεν είναι πλήρως καθορισμένη, αφού υπάρχουν άγνωστες παράμετροι. Θα χρησιμοποιηθεί ο έλεγχος των Kolmogorov-Smirnov, που εκτενώς παρουσιάστηκε στην Ενότητα 4.3.1, με την παράμετρο της εκθετικής κατανομής να εκτιμάται από τα δεδομένα.

(με χρήση SPSS): Αρχικά, εισάγουμε σε μια στήλη ενός κενού φύλλου εργασίας του SPSS τα δεδομένα και μετονομάζουμε τη στήλη σε X. Στη συνέχεια, από το κεντρικό παράθυρο διαλόγου επιλέγουμε:

Analyze / Nonparametric Tests / Legacy Dialogs / 1-Sample K-S

Στο νέο παράθυρο διαλόγου που προκύπτει (βλ. Εικόνα 13.3) εισάγουμε στο πεδίο Test Variable List τη X. Επιλέγουμε Test Distribution: Exponential, αποεπιλέγουμε το Normal (εκτός αν επιθυμούμε να ελέγξουμε και την υπόθεση της κανονικότητας) και αφήνουμε την προεπιλογή Use sample data. Πατάμε OK και προκύπτει το output της ανάλυσης, το οποίο δίνεται στην Εικόνα 13.4.

Στο output της ανάλυσης δίνεται η τιμή της στατιστικής συνάρτησης ελέγχου (γραμμή Test Statistic) η οποία αντιστοιχεί στην απόλυτη τιμή της μεγαλύτερης διαφοράς μεταξύ της υποτιθέμενης κατανομής, υπό την H_0,



Εικόνα 13.3: Παράθυρο διαλόγου για έλεγχο One-Sample Kolmogorov-Smirnov Test - Περίπτωση Εκθετικής κατανομής.

και της εμπειρικής συνάρτησης κατανομής. Η τιμή αυτή δίνεται και στη γραμμή Absolute, στο Most Extreme Differences. Παρατηρήστε, επίσης, ότι η παράμετρος της κατανομής έχει εκτιμηθεί από τα δεδομένα και είναι ίση με $\hat{\theta} = 2.9411$. Αξίζει να αναφέρουμε ότι η τιμή αυτή είναι η εκτίμηση της μέσης τιμής της εκθετικής κατανομής. Η p-τιμή του ελέγχου είναι ίση με 0.473 και βασίζεται στο τεστ του Lilliefors λόγω του ότι η παράμετρος της κατανομής υπό την H_0 δεν είναι πλήρως καθορισμένη και έχει εκτιμηθεί από τα δεδομένα. Η εκτίμηση βασίζεται σε 10.000 το πλήθος Monte Carlo επαναλήψεις, ενώ δίνεται και ένα 99% διάστημα εμπιστοσύνης για αυτήν. Η p-τιμή του ελέγχου εκτιμάται (με βεβαιότητα 99%) ότι είναι μεταξύ 0.460 και 0.485 και άρα, δεν μπορούμε να απορρίψουμε την H_0. Δηλαδή, δεν μπορούμε να απορρίψουμε την υπόθεση ότι τα δεδομένα προέρχονται από εκθετική κατανομή.

One-Sample Kolmogorov-Smirnov Test

			X
N			35
Exponential parameter.[a,b]	Mean		2,9411
Most Extreme Differences	Absolute		,117
	Positive		,058
	Negative		-,117
Test Statistic			,117
Monte Carlo Sig. (2-tailed)[c]	Sig.		,473
	99% Confidence Interval	Lower Bound	,460
		Upper Bound	,485

a. Test Distribution is Exponential.

b. Calculated from data.

c. Lilliefors' method based on 10000 Monte Carlo samples with starting seed 624387341.

Εικόνα 13.4: Αποτελέσματα ελέγχου One-Sample Kolmogorov-Smirnov για τα δεδομένα του Πίνακα 13.2.

(με χρήση R): Για να μπορέσουμε να εκτελέσουμε στην `R` τον έλεγχο Lilliefors για την υπόθεση ότι τα δεδομένα προέρχονται από εκθετική κατανομή, πρέπει να φορτώσουμε το πακέτο `KScorrect`. Στο πακέτο αυτό περιέχεται η συνάρτηση `LcKS(...)`, η οποία εκτελεί τον έλεγχο One-Sample Kolmogorov-Smirnov με τη διόρθωση Lilliefors. Υπενθυμίζεται ότι ο έλεγχος αυτός πρέπει να προτιμάται

όταν οι παράμετροι του υποτιθέμενου μοντέλου είναι άγνωστες και πρέπει να εκτιμηθούν, τουτέστιν όταν έχουμε να κάνουμε έναν σύνθετο έλεγχο καλής προσαρμογής. Ο υπολογισμός της p-τιμής γίνεται μέσω προσομοίωσης, ενώ, εκτός από την εκθετική κατανομή, μπορεί να χρησιμοποιηθεί και για ελέγχους άλλων μοντέλων (πιθανοτικών προτύπων), όπως της κανονικής κατανομής, της ομοιόμορφης κατανομής, της κατανομής Γάμμα και της κατανομής Weibull.

Αφού φορτώσουμε το πακέτο, εισάγουμε τα δεδομένα (35 μετρήσεις) σε ένα διάνυσμα, έστω αυτό `x`. Στη συνέχεια, χρησιμοποιούμε τη συνάρτηση `LcKS(x,'pexp')`, όπου στο πρώτο όρισμα είναι το διάνυσμα με τις διαθέσιμες τιμές και στο δεύτερο είναι το μοντέλο του οποίου θέλουμε να ελέγξουμε την καλή προσαρμογή στα δεδομένα. Για την περίπτωση της εκθετικής εισάγουμε `'pexep'`. Παρακάτω δίνονται οι σχετικές εντολές και το αποτέλεσμα της ανάλυσης.

```
> library(KScorrect)
> x<-c(1.98,5.08,0.26,8.11,4.04,0.73,2.05,2.92,0.27,3.81,4.03,
+ 3.31,0.61,3.25,5.35,2.40,1.97,0.77,0.22,2.99,0.44,3.74,0.54,
+ 2.76,2.35,1.05,6.12,9.00,10.76,2.32,6.98,0.19,1.33,
+ 0.87,0.34)
> set.seed(9)
> LcKS.results<-LcKS(x, 'pexp')
> LcKS.results$p.value # print only the simulated pvalue
[1] 0.492
```

Η p-τιμή εκτιμάται ότι είναι ίση με 0.492 (το SPSS έδωσε αντίστοιχη p-τιμή ίση με 0.473) και άρα, σε ε.σ. 5%, δεν μπορούμε να απορρίψουμε την υπόθεση ότι τα δεδομένα προέρχονται από εκθετική κατανομή. Παρατηρήστε ότι χρησιμοποιήσαμε set.seed(9), ενώ θα μπορούσε να χρησιμοποιηθεί οποιαδήποτε άλλη τιμή (και όχι οπωσδήποτε το 9). Αυτό έγινε ώστε να είναι δυνατή η επιβεβαίωση των αποτελεσμάτων. Επειδή ο υπολογισμός της p-τιμής γίνεται μέσω προσομοίωσης, αν δεν χρησιμοποιηθεί set.seed θα υπάρχουν (μικρές) διαφοροποιήσεις στην εκτίμηση της p-τιμής από υπολογιστή σε υπολογιστή. □

Παράδειγμα 13.3. (**GOF-Chi square test**): Στον Πίνακα 13.3 δίνονται 100 ψηφία από το σύνολο $\{0,1,...,9\}$ τα οποία έχουν παραχθεί από ένα πρόγραμμα Η/Υ. Αν ο Η/Υ παράγει πράγματι τυχαίους αριθμούς, θα πρέπει καθένα από τα ψηφία να έχει την ίδια πιθανότητα να εμφανιστεί. Σε ε.σ. 1% να ελέγξετε την υπόθεση της τυχαιότητας εμφάνισης των ψηφίων με χρήση του ελέγχου χι-τετράγωνο καλής προσαρμογής.

Λύση Παραδείγματος 13.3. Στην ουσία έχουμε ένα τυχαίο πείραμα από $n = 100$ ανεξάρτητες δοκιμές στο οποίο μπορούμε να έχουμε $k = 10$ δυνατά αποτελέσματα, ξένα μεταξύ τους. Έστω E_i, $i = 0,...,9$, το ενδεχόμενο εμφάνισης του i-οστού ψηφίου, $i = 0,...,9$, με $p_i = P(E_i)$ τις αντίστοιχες πιθανότητες πραγματοποίησης καθενός ενδεχομένου. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση ότι $H_0 : p_i = \frac{1}{10}$, $i = 0,2,...,9$ έναντι της H_1 : όχι η H_0. Πρόκειται για απλό έλεγχο καλής προσαρμογής και ο χι-τετράγωνο έλεγχος γίνεται με τη στατιστική συνάρτηση:

$$X^2 = \sum_{i=0}^{9} \frac{(n_i - e_i)^2}{e_i},$$

όπου n_i και e_i είναι ο παρατηρούμενος και αναμενόμενος, υπό τη μηδενική υπόθεση, αριθμός εμφανίσεων του E_i, $i = 0,...,9$. Δηλαδή είναι $e_i = np_{i0}$, $i = 0,,...,9$, με $p_{i0} = 1/10$. Για περισσότερες λεπτομέρειες παραπέμπουμε στην Ενότητα 4.2.

(με χρήση SPSS): Αρχικά, εισάγουμε τις 100 το πλήθος τιμές σε μια κενή στήλη ενός φύλλου εργασίας του SPSS, την οποία μετονομάζουμε σε X. Έπειτα από το κεντρικό παράθυρο διαλόγου επιλέγουμε:

Analyze / Nonparametric Tests / Legacy Dialogs / Chi-square Test

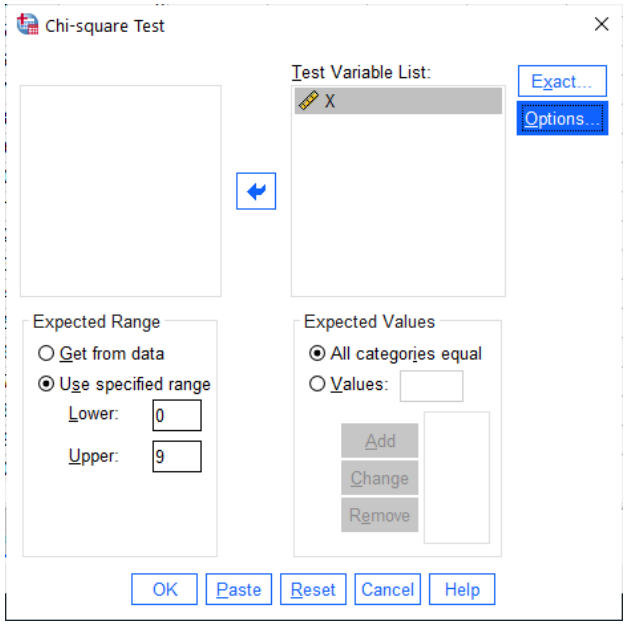
A/A	X	A/A	X	A/A	X	A/A	X	A/A	X
1	2	21	5	41	2	61	3	81	1
2	9	22	4	42	8	62	9	82	4
3	1	23	7	43	3	63	7	83	7
4	6	24	1	44	7	64	5	84	7
5	9	25	2	45	3	65	5	85	7
6	2	26	2	46	1	66	4	86	1
7	8	27	5	47	0	67	5	87	2
8	6	28	2	48	2	68	9	88	6
9	5	29	5	49	1	69	9	89	5
10	4	30	6	50	0	70	2	90	2
11	0	31	3	51	0	71	9	91	8
12	4	32	3	52	1	72	6	92	4
13	6	33	2	53	9	73	1	93	2
14	9	34	9	54	5	74	6	94	3
15	0	35	0	55	1	75	4	95	1
16	6	36	8	56	6	76	2	96	5
17	7	37	1	57	4	77	8	97	3
18	5	38	7	58	3	78	9	98	6
19	1	39	6	59	8	79	5	99	8
20	5	40	8	60	7	80	1	100	0

Πίνακας 13.3: Δεδομένα τυχαίων αριθμών.

Στο νέο παράθυρο που εμφανίζεται (βλ. Εικόνα 13.5) εισάγουμε τη στήλη X στο πεδίο Test Variable List. Στο πεδίο Expected Range επιλέγουμε Use specified range, όπου στο Lower δίνουμε την τιμή 0 (καθώς αυτή είναι η μικρότερη διαθέσιμη τιμή) και στο Upper την τιμή 9 (καθώς αυτή είναι η μεγαλύτερη διαθέσιμη τιμή). Επίσης, στο Expected Values επιλέγουμε All categories equal (καθώς υπό τη μηδενική υπόθεση θέλουμε ίσες πιθανότητες εμφάνισης των 10 το πλήθος ψηφίων). Πατάμε OK και προκύπτει το output της ανάλυσης, που δίνεται στις Εικόνες 13.6 και 13.7.

Στην Εικόνα 13.6 δίνεται ο πίνακας με τις παρατηρούμενες και αναμενόμενες συχνότητες για καθεμία από τις κατηγορίες αποτελεσμάτων. Σε αυτήν την περίπτωση, οι κατηγορίες αυτές είναι τα ψηφία $0,1,...,9$. Επίσης, στην Εικόνα 13.7 δίνεται η τιμή της στατιστικής συνάρτησης ελέγχου (γραμμή Chi-Square) καθώς και η p-τιμή του ελέγχου. Αφού η p-τιμή είναι ίση με 0.834, δεν απορρίπτουμε την H_0. Άρα δεν μπορούμε να απορρίψουμε την υπόθεση ότι ο Η/Υ παράγει πράγματι τυχαία ψηφία από το σύνολο $\{0,1,...,9\}$, δηλαδή δεν μπορούμε να απορρίψουμε την υπόθεση ότι καθένα από τα 10 ψηφία έχει την ίδια πιθανότητα εμφάνισης.

(με χρήση R): Εισάγουμε τα δεδομένα σε ένα διάνυσμα (έστω `x`). Αφού εισάγουμε τα δεδομένα, χρησιμοποιούμε την εντολή `table(x)`, ώστε να φτιάξουμε τον πίνακα συχνοτήτων των τιμών 0, 1, ..., 9 στο `x`. Στη συνέχεια, χρησιμοποιούμε την εντολή `chisq.test(tab)`, όπως παρακάτω:



Εικόνα 13.5: Παράθυρο διαλόγου Chi-square Test του SPSS.

Frequencies

X

	Category	Observed N	Expected N	Residual
1	0	7	10,0	-3,0
2	1	13	10,0	3,0
3	2	13	10,0	3,0
4	3	8	10,0	-2,0
5	4	8	10,0	-2,0
6	5	13	10,0	3,0
7	6	11	10,0	1,0
8	7	9	10,0	-1,0
9	8	8	10,0	-2,0
10	9	10	10,0	,0
Total		100		

Εικόνα 13.6: Πίνακας παρατηρούμενων και αναμενόμενων συχνοτήτων.

Test Statistics

	X
Chi-Square	5,000[a]
df	9
Asymp. Sig.	,834

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 10,0.

Εικόνα 13.7: Αποτελέσματα ελέγχου χι-τετράγωνο καλής προσαρμογής.

```
> x<-c(2,5,2,3,1,9,4,8,9,4,1,7,3,7,7,6,1,7,5,7,
+9,2,3,5,7,2,2,1,4,1,8,5,0,5,2,6,2,2,9,6,
+5,5,1,9,5,4,6,0,2,2,0,3,0,9,8,4,3,1,6,4,
+6,2,9,1,2,9,9,5,6,3,0,0,1,4,1,6,8,6,2,5,
+7,1,4,8,3,5,7,3,9,6,1,6,8,5,8,5,8,7,1,0)
> tab<-table(x) # create the frequency table
> tab # print the frequency table
x
 0  1  2  3  4  5  6  7  8  9
 7 13 13  8  8 13 11  9  8 10
> chisq.test(tab) # apply GOF chi-square test
```

Πριν προχωρήσουμε στον σχολιασμό των αποτελεσμάτων της ανάλυσης, θα πρέπει να αναφέρουμε ότι ο τρόπος με τον οποίο χρησιμοποιούμε την εντολή `chisq.test(tab)` υποθέτει ότι οι διαφορετικές τιμές (κατηγορίες) είναι ισοπίθανες. Αν θέλουμε να το αλλάξουμε, θα πρέπει να χρησιμοποιήσουμε το όρισμα `p = ...` και να εισάγουμε το διάνυσμα με τις πιθανότητες για κάθε κατηγορία. Το αποτέλεσμα της ανάλυσης είναι το ακόλουθο:

```
        Chi-squared test for given probabilities

data:  tab
X-squared = 5, df = 9, p-value = 0.8343
```

Η p-τιμή εκτιμάται ότι είναι ίση με 0.8343 (είναι η ίδια με αυτήν που έδωσε το SPSS). Άρα, σε ε.σ. 5%, δεν μπορούμε να απορρίψουμε την υπόθεση ότι ο Η/Υ παράγει πράγματι τυχαία ψηφία από το σύνολο $\{0,1,...,9\}$. □

Παράδειγμα 13.4. (**GOF-Chi square test - Σύνθετος έλεγχος καλής προσαρμογής**): Στον Πίνακα 13.4 δίνονται οι μετρήσεις ενός τυχαίου δείγματος 50 τιμών από πληθυσμό με άγνωστη διακριτή κατανομή. Χρησιμοποιώντας τον χι-τετράγωνο έλεγχο καλής προσαρμογής να ελέγξετε, σε ε.σ. 5%, την υπόθεση ότι τα δεδομένα προέρχονται από πληθυσμό που μοντελοποιείται σύμφωνα με το πιθανοτικό πρότυπο της κατανομής Poisson (με άγνωστη μέση τιμή θ).

5	2	5	3	4
3	3	3	4	3
7	5	3	4	4
2	4	8	4	2
2	5	5	2	3
4	6	8	3	2
4	5	5	3	4
5	7	2	5	4
5	4	3	3	3
2	3	8	5	3

Πίνακας 13.4: Τυχαίο δείγμα 50 τιμών από άγνωστη διακριτή κατανομή.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
X	50	2	8	4,02	1,610
Valid N (listwise)	50				

Εικόνα 13.8: Αποτέλεσμα της διαδικασίας Descriptives του SPSS.

Λύση Παραδείγματος 13.4. Από την εκφώνηση προκύπτει ότι έχουμε έναν σύνθετο έλεγχο καλής προσαρμογής, καθώς η υποθετική υπό τη μηδενική υπόθεση κατανομή περιέχει μία άγνωστη παράμετρο. Δηλαδή σε αυτήν την περίπτωση έχουμε ένα τυχαίο δείγμα $X_1, X_2, \dots, X_{50}$ από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(x)$, η οποία είναι άγνωστη, και θέλουμε να ελέγξουμε τη μηδενική υπόθεση $H_0 : F(x) = F_0(x;\theta)$, για κάθε $x \in \mathbb{R}$ και για κάποιο $\theta \in \Theta$, όπου $F_0(x;\theta)$ είναι η αθροιστική συνάρτηση κατανομής της κατανομής Poisson. Όπως είδαμε αναλυτικά στην Ενότητα 4.2, για να μπορέσουμε να εφαρμόσουμε τον χι-τετράγωνο έλεγχο καλής προσαρμογής, θα πρέπει αρχικά να εκτιμήσουμε την άγνωστη παράμετρο και, στη συνέχεια, να ομαδοποιήσουμε τα δεδομένα σε κατηγορίες κατά τέτοιο τρόπο ώστε οι αναμενόμενες συχνότητες σε κάθε κατηγορία να είναι μεγαλύτερες ή ίσες από 5. Η υλοποίηση των παραπάνω με χρήση SPSS και `R` δίνεται παρακάτω.

(με χρήση SPSS): Αρχικά, σε μια στήλη ενός κενού φύλλου εργασίας του SPSS εισάγουμε τις τιμές του Πίνακα 13.4 και μετονομάζουμε τη στήλη σε X. Στη συνέχεια, θα πρέπει να εκτιμήσουμε την παράμετρο της κατανομής Poisson καθώς από τα δεδομένα του προβλήματος δεν γνωρίζουμε την τιμή της. Επιλέγουμε από το κεντρικό παράθυρο διαλόγου

Analyze / Descriptive Statistics / Descriptives

Στο νέο παράθυρο διαλόγου που προκύπτει στο πεδίο Variables(s) εισάγουμε τη μεταβλητή X και πατάμε OK. Από το output που προκύπτει (βλ. Εικόνα 13.8) έχουμε ότι η εκτιμώμενη τιμή της μέσης τιμής της Poisson είναι $\hat{\theta} = 4.02$. Παρατηρούμε, επίσης, ότι η ελάχιστη τιμή στο δείγμα είναι το 2, ενώ η μέγιστη είναι το 8. Είναι γνωστό ότι το στήριγμα της κατανομής Poisson είναι το σύνολο των θετικών ακεραίων $\{0,1,2,...\}$.

Για να υπολογίσουμε τις αναμενόμενες συχνότητες υπό την υπόθεση της κατανομής Poisson με παράμετρο $\hat{\theta} = 4.02$, θα χρησιμοποιήσουμε το SPSS με τον τρόπο που περιγράφεται στη συνέχεια.

Από τα δεδομένα της εκφώνησης προκύπτει ότι το σύνολο των τιμών της τυχαίας μεταβλητής χωρίζεται σε 7 κατηγορίες, ξένες μεταξύ τους. Έστω E_1 το ενδεχόμενο να έχουμε τιμή μικρότερη ή ίση από 2, E_2 το ενδεχόμενο να έχουμε τιμή ίση με 3, ..., E_6 το ενδεχόμενο να έχουμε τιμή ίση με 6 και E_7 το ενδεχόμενο να έχουμε τιμή μεγαλύτερη ή ίση από 8. Αρχικά, θα υπολογίσουμε τις πιθανότητες πραγματοποίησης κάθε

ενδεχομένου, όταν έχουμε κατανομή Poisson με παράμετρο 4.02. Για να επιτευχθεί αυτό, εισάγουμε σε μια νέα στήλη του SPSS τις τιμές 0, 1, 2,...,8. Μετονομάζουμε τη στήλη σε Y. Στη συνέχεια, επιλέγουμε από το κεντρικό παράθυρο διαλόγου:

Transform / Compute Variable

και στο παράθυρο διαλόγου που ανοίγει δίνουμε την ονομασία pdfPoisson στη νέα στήλη μέσω του πεδίου Target Variable, ενώ στο Numeric Expression δίνουμε PDF.POISSON(Y,4.02) και πατάμε OK. Με τον τρόπο αυτό υπολογίσαμε τις πιθανότητες $P(X = x) = e^{-4.02}\frac{4.02^x}{x!}$, $x = 0,1,2,...,8$. Από τις τιμές αυτές προκύπτουν τα ακόλουθα:

$$p_{10}(\hat{\lambda}) = P(X \leqslant 2|X \sim \mathscr{P}(4.02)) = 0.23518743$$

$$p_{20}(\hat{\lambda}) = P(X = 3|X \sim \mathscr{P}(4.02)) = 0.19438515,$$

$$p_{30}(\hat{\lambda}) = P(X = 4|X \sim \mathscr{P}(4.02)) = 0.19535708,$$

$$p_{40}(\hat{\lambda}) = P(X = 5|X \sim \mathscr{P}(4.02)) = 0.15706709,$$

$$p_{50}(\hat{\lambda}) = P(X = 6|X \sim \mathscr{P}(4.02)) = 0.10523495,$$

$$p_{60}(\hat{\lambda}) = P(X = 7|X \sim \mathscr{P}(4.02)) = 0.06043493,$$

$$p_{70}(\hat{\lambda}) = P(X \geqslant 8|X \sim \mathscr{P}(4.02)) = 1 - P(X \leqslant 7|X \sim \mathscr{P}(4.02)) = 0.05233336.$$

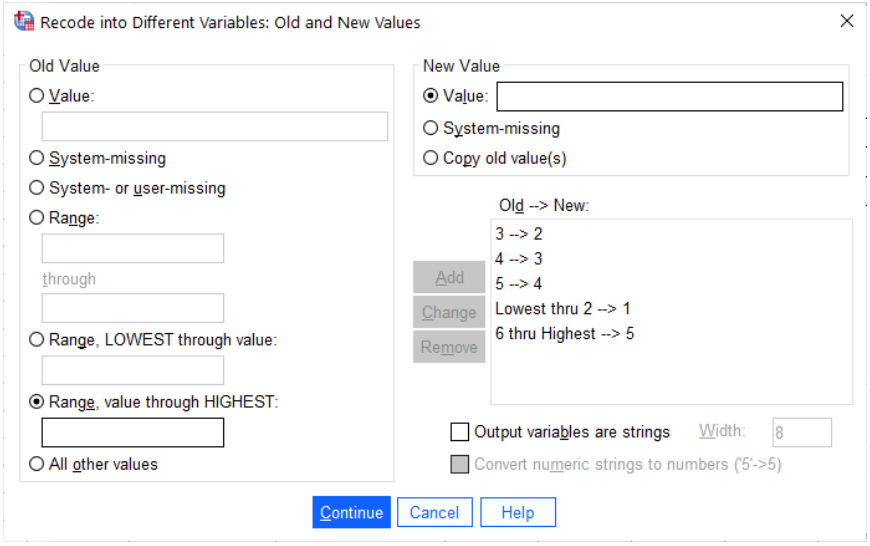
Εισάγουμε τις παραπάνω τιμές σε μια νέα στήλη του SPSS και τη μετονομάζουμε σε z. Στη συνέχεια, επιλέγουμε πάλι **Transform / Compute Variable** και στο παράθυρο διαλόγου που ανοίγει δίνουμε nz στο Target Variable, ενώ στο Numeric Expression δίνουμε 50*z. Πατάμε OK και έχουμε υπολογίσει τις αναμενόμενες συχνότητες εμφάνισης των 7 ενδεχομένων. Προκύπτουν τότε τα αποτελέσματα που δίνονται στη συνέχεια:

E1	0.23518743	11.759372
E2	0.19438515	9.719258
E3	0.19535708	9.767854
E4	0.15706709	7.853355
E5	0.10523495	5.261748
E6	0.06043493	3.021746
E7	0.05233336	2.616668

Πίνακας 13.5: Πίνακας πιθανοτήτων και αναμενόμενων συχνοτήτων υπό την υπόθεση της κατανομής Poisson για τα δεδομένα του Παραδείγματος 13.4.

Από τον Πίνακα 13.5 παρατηρούμε ότι θα πρέπει να συγχωνεύσουμε κάποιες από τις κατηγορίες, ώστε να έχουμε αναμενόμενες συχνότητες $\geqslant 5$. Δεν είναι δύσκολο να διαπιστώσουμε ότι δύναται να συνενωθούν οι κατηγορίες E5, E6 και E7 του Πίνακα 13.5. Κατά αυτόν τον τρόπο θα προκύψουν οι ακόλουθες νέες κατηγορίες:

- E1, για τις τιμές που είναι μικρότερες ή ίσες από 2 (δηλ. $\{\leqslant 2\}$),
- E2, για τις τιμές που είναι ίσες με 3 (δηλ. $\{= 3\}$),
- E3, για τις τιμές που είναι ίσες με 4 (δηλ. $\{= 4\}$)
- E4, για τις τιμές που είναι ίσες με 5 (δηλ. $\{= 5\}$)
- E5, για τις τιμές που είναι ίσες ή μεγαλύτερες από 6 (δηλ. $\{\geqslant 6\}$).



Εικόνα 13.9: Κωδικοποίηση των τιμών του Πίνακα 13.4.

Οι αντίστοιχες πιθανότητες είναι 0.23518743, 0.19438515, 0.19535708, 0.15706709, 0.2180032. Αντίστοιχα, από τη συνένωση των κατηγοριών, οι αντίστοιχες αναμενόμενες συχνότητες είναι 11.759372, 9.719258, 9.767854, 7.853355 και 10.90016.

Στη συνέχεια, θα πρέπει να τοποθετήσουμε στις 5 κατηγορίες τις τιμές που έχουν δοθεί στη στήλη X. Για να το κάνουμε αυτό, εργαζόμαστε με τον τρόπο που ακολουθεί. Επιλέγουμε

Transform / Recode into Different Variables.

και στο νέο παράθυρο διαλόγου, που προκύπτει, εισάγουμε στο Numeric Variable -> Output Variable τη στήλη X και στο Output Variable δίνουμε Xcat και πατάμε Change.

Στη συνέχεια, πατάμε Old and New Values και στο παράθυρο που ανοίγει (βλ. Εικόνα 13.9), μας δίνεται η δυνατότητα να κάνουμε επανακωδικοποίηση των τιμών της X. Αρχικά, επιλέγουμε Range, LOWEST through value και δίνουμε την τιμή 2, ενώ στο New Value: Value δίνουμε την τιμή 1 (η 1η κατηγορία) και πατάμε Add. Στη συνέχεια, στο Old value / Value δίνουμε την τιμή 3 και στο New Value την τιμή 2 (η 2η κατηγορία) και πατάμε Add. Με τον ίδιο τρόπο ορίζουμε τις κατηγορίες 3 και 4, οι οποίες αντιστοιχούν στις τιμές 4 και 5 της στήλης X. Τέλος, επιλέγουμε Range, value through HIGHEST και δίνουμε την τιμή 6, ενώ στο New Value: Value δίνουμε την τιμή 5 (η 5η κατηγορία) και πατάμε Add. Πατάμε Continue και μετά OK.

Πλέον, είμαστε έτοιμοι να διεξάγουμε τον σύνθετο χι-τετράγωνο έλεγχο καλής προσαρμογής. Από το κεντρικό παράθυρο διαλόγου επιλέγουμε:

Analyze / Nonparametric Tests / Legacy Dialogs / Chi-square

Στο νέο παράθυρο διαλόγου που προκύπτει στο πεδίο Test Variable List εισάγουμε τη μεταβλητή Xcat, ενώ στο Expected Values επιλέγουμε Values και εισάγουμε τις αναμενόμενες συχνότητες, που αντιστοιχούν στις κατηγορίες 1 έως 5, με την αντίστοιχη σειρά. Έτσι, για την 1η κατηγορία εισάγουμε στο Values την τιμή 11.759372 και πατάμε Add. Για τη 2η κατηγορία εισάγουμε στο Values την τιμή 9.719258 και πατάμε Add. Με τον ίδιο τρόπο εισάγουμε και τις αναμενόμενες συχνότητες για τις υπόλοιπες κατηγορίες. Πατάμε OK και προκύπτει το output της ανάλυσης.

Θα πρέπει να αναφέρουμε ότι από το output παραθέτουμε μόνο ένα τμήμα του στην Εικόνα 13.10. Αυτό γίνεται καθώς το τμήμα των αποτελεσμάτων που αφορά τον υπολογισμό της p-τιμής είναι λανθασμένο,

αφού το SPSS την υπολογίζει χωρίς να αφαιρεί έναν βαθμό ελευθερίας από τη χι-τετράγωνο κατανομή (λόγω της εκτίμησης μιας παραμέτρου). Στην Εικόνα 13.10 δίνεται ο πίνακας με τις παρατηρούμενες και με τις αναμενόμενες συχνότητες για τις 5 κατηγορίες, όπως αυτές προέκυψαν μετά από τη συνένωση που περιγράψαμε παραπάνω. Για τον σωστό υπολογισμό της p-τιμής λαμβάνουμε υπόψη ότι η τιμή της σ.σ. χ^2 είναι ίση με 6.706.

Xcat

	Observed N	Expected N	Residual
1,00	8	11,8	-3,8
2,00	14	9,7	4,3
3,00	11	9,8	1,2
4,00	11	7,9	3,1
5,00	6	10,9	-4,9
Total	50		

Εικόνα 13.10: Πίνακας παρατηρούμενων και αναμενόμενων συχνοτήτων μετά την ένωση κατηγοριών.

Για να υπολογίσουμε τη σωστή p-τιμή με το SPSS δουλεύουμε ως εξής: επιλέγουμε Transform / Compute Variable, στο Target Variable δίνουμε PVALUE, ενώ στο Numeric Expression εισάγουμε την έκφραση 1-CDF.CHISQ(6.706,3). Πατάμε OK και η p-τιμή καταχωρίζεται ως νέα μεταβλητή/στήλη στο φύλλο εργασίας του SPSS και είναι ίση με 0.08188. Συμπεραίνουμε ότι το ελάχιστο ε.σ. για το οποίο δεν απορρίπτεται η H_0 είναι 8.188%.

(με χρήση R): Εισάγουμε τα δεδομένα σε ένα διάνυσμα, έστω αυτό `x`, και εκτελούμε τις παρακάτω εντολές:

```
> x<-c(5,2,5,3,4,3,3,3,4,3,7,5,3,4,4,2,4,8,4,2,2,5,5,2,3,4,6,8,
+ 3,2,4,5,5,3,4,5,7,2,5,4,5,4,3,3,3,2,3,8,5,3)
> lambdaest<-mean(x) # ML estimator of Poisson parameter
> lambdaest # print the estimate
[1] 4.02
> tab<-table(x) # frequency table from x
> tab # print the frequency table
x
 2  3  4  5  6  7  8
 8 14 11 11  1  2  3
# vector of probabilities
> probs<-c(ppois(2,lambdaest),dpois(c(3,4,5,6,7),lambdaest),
+ 1-ppois(7,lambdaest))
> OBServed<-c(8,14,11,11,1,2,3) # observed frequencies
> chisq.test(OBServed,correct=F,p=probs,simulate.p.value=T,B=2000)
```

Πιο συγκεκριμένα, αρχικά, υπολογίζουμε την εκτιμώμενη τιμή της άγνωστης πληθυσμιακής παραμέτρου της κατανομής Poisson. Στη συνέχεια, χρησιμοποιούμε την εντολή `table(x)`, ώστε να φτιάξουμε τον πίνακα συχνοτήτων των τιμών διαφορετικών τιμών στο `x`. Έπειτα χρησιμοποιούμε την εντολή `chisq.test(tab)`. Το διάνυσμα `probs` περιέχει τις πιθανότητες για καθεμία από τις κλάσεις, με βάση την κατανομή Poisson με παράμετρο 4.02. Επίσης, έχουμε επιλέξει και την τιμή `TRUE` για το όρισμα `simulate.p.value`, ώστε, αν κάποια από τις αναμενόμενες συχνότητες είναι < 5 (και άρα η προσέγγιση προς την κατανομή χ^2 δεν θα είναι ικανοποιητική), τότε η p-τιμή του ελέγχου να υπολογιστεί μέσω προσομοίωσης (για 2000 το πλήθος επαναλήψεις). Το αποτέλεσμα των παραπάνω εντολών είναι το ακόλουθο:

```
        Chi-squared test for given probabilities with simulated p-value (based
           on 2000 replicates)

data:  OBServed
X-squared = 8.3569, df = NA, p-value = 0.2119
```

Η p-τιμή εκτιμάται ότι είναι ίση με 0.2119, διαφορετική από αυτήν την οποία έδωσε το SPSS, καθώς βασίζεται σε προσομοίωση. Άρα, σε ε.σ. 5%, δεν απορρίπτουμε την H_0 και άρα δεν μπορούμε να απορρίψουμε την υπόθεση ότι τα δεδομένα προέρχονται από κατανομή Poisson. □

Παράδειγμα 13.5. (**GOF-Chi square test & Kolmogorov-Smirnov test**): Να ελέγξετε την υπόθεση ότι τα δεδομένα του Πίνακα 13.6 (50 τιμές) προέρχονται από την κατανομή Γάμμα $\mathscr{G}(2,1/4)$. Ο έλεγχος να γίνει σε ε.σ. 5% με χρήση των ελέγχων καλής προσαρμογής χ^2 και one-sample Kolmogorov-Smirnov. Θα πρέπει να σημειωθεί ότι κατά τη συλλογή (καταγραφή) των δεδομένων έγινε στρογγυλοποίηση στις αρχικές μετρήσεις ώστε όλες να δίνονται με ακρίβεια 2 δεκαδικών ψηφίων.

14.60	0.58	7.73	2.52	3.47
1.91	20.00	10.29	2.34	3.83
21.93	1.72	15.59	8.46	5.66
1.14	6.67	5.10	5.47	10.58
2.96	5.31	4.19	3.10	7.36
11.12	5.42	4.77	1.91	4.32
5.18	2.50	13.54	7.00	2.88
4.28	3.16	10.18	8.12	15.58
12.33	9.70	20.25	2.19	8.42
7.84	26.18	19.25	3.65	10.65

Πίνακας 13.6: Δεδομένα από άγνωστη συνεχή κατανομή.

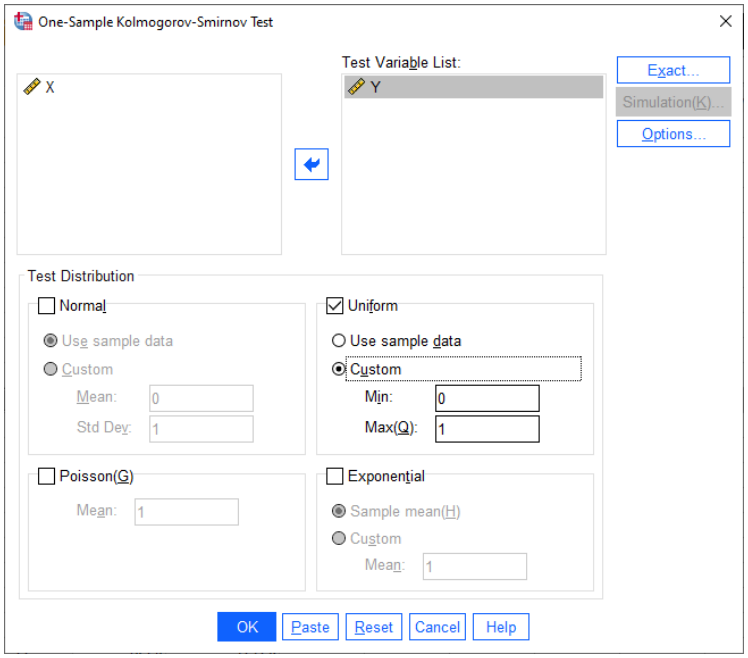
Λύση Παραδείγματος 13.5. Πρόκειται για απλό έλεγχο καλής προσαρμογής, καθώς γνωρίζουμε πλήρως την κατανομή υπό τη μηδενική υπόθεση, και θα υλοποιήσουμε τον χι-τετράγωνο έλεγχο καλής προσαρμογής και τον Kolmogorov-Smirnov έλεγχο.

(**με χρήση SPSS**): Αρχικά, σε μια στήλη εισάγουμε τις τιμές του Πίνακα 13.6 και μετονομάζουμε τη στήλη σε X. Το SPSS δεν μας δίνει άμεσα τρόπο να ελέγξουμε την υπόθεση ότι τα δεδομένα προσαρμόζονται στην κατανομή Γάμμα. Μπορούμε, όμως, να μετασχηματίσουμε τα παραπάνω δεδομένα σε τιμές από την $\mathscr{U}(0,1)$. Γνωρίζουμε ότι, αν X τ.μ. με α.σ.κ. $F_X(x)$, τότε η μετασχηματισμένη τ.μ. $Y = F_X(X)$ ακολουθεί την $\mathscr{U}(0,1)$. Άρα, χρησιμοποιώντας την α.σ.κ. της $\mathscr{G}(2,1/4)$, θα μετασχηματίσουμε τις τιμές στη στήλη X. Επιλέγουμε **Transform / Compute Variable** και στο παράθυρο διαλόγου που ανοίγει, δίνουμε Y στο **Target Variable** και στο **Numeric Expression** εισάγουμε CDF.GAMMA(X,2,1/4), το οποίο θα εφαρμόσει τον παραπάνω μετασχηματισμό στα δεδομένα. Πατάμε OK. Σημειώνεται ότι ο τρόπος με τον οποίο ορίζει το SPSS την κατανομή Γάμμα είναι τέτοιος ώστε η μέση τιμή της είναι α/β, όπου α, β είναι οι παράμετροι σχήματος και κλίμακας, αντίστοιχα.

Στη συνέχεια, από το κεντρικό παράθυρο διαλόγου επιλέγουμε:

Analyze / Nonparametric Tests / Legacy Dialogs / 1-Samples K-S.

Στο νέο παράθυρο διαλόγου που προκύπτει (βλ. την Εικόνα 13.11) στο πλαίσιο **Test Variable List** εισάγουμε τη στήλη Y, ενώ στο πλαίσιο Test Distribution επιλέγουμε μόνο Uniform και πατάμε Custom. Αφήνουμε τις τιμές 0 και 1 στα Min και Max(Q) και πατάμε OK.



Εικόνα 13.11: Παράθυρο διαλόγου One-Sample Kolmogorov-Smirnov Test για έλεγχο δεδομένων από Ομοιόμορφη κατανομή.

Το output της ανάλυσης δίνεται στην Εικόνα 13.12. Από την p-τιμή του ασυμπτωτικού ελέγχου, προκύπτει ότι δεν απορρίπτουμε τη μηδενική υπόθεση, ότι οι (μετασχηματισμένες) τιμές $y_1, y_2, \ldots, y_{50}$ προέρχονται από κατανομή $\mathcal{U}(0,1)$. Άρα, δεν μπορούμε να απορρίψουμε την υπόθεση ότι οι αρχικές παρατηρήσεις $x_1, x_2, \ldots, x_{50}$ προέρχονται από κατανομή $\mathcal{G}(2,1/4)$.

One-Sample Kolmogorov-Smirnov Test

		Y
N		50
Uniform Parameters[a,b]	Minimum	0
	Maximum	1
Most Extreme Differences	Absolute	,106
	Positive	,106
	Negative	-,053
Kolmogorov-Smirnov Z		,751
Asymp. Sig. (2-tailed)		,626

a. Test distribution is Uniform.

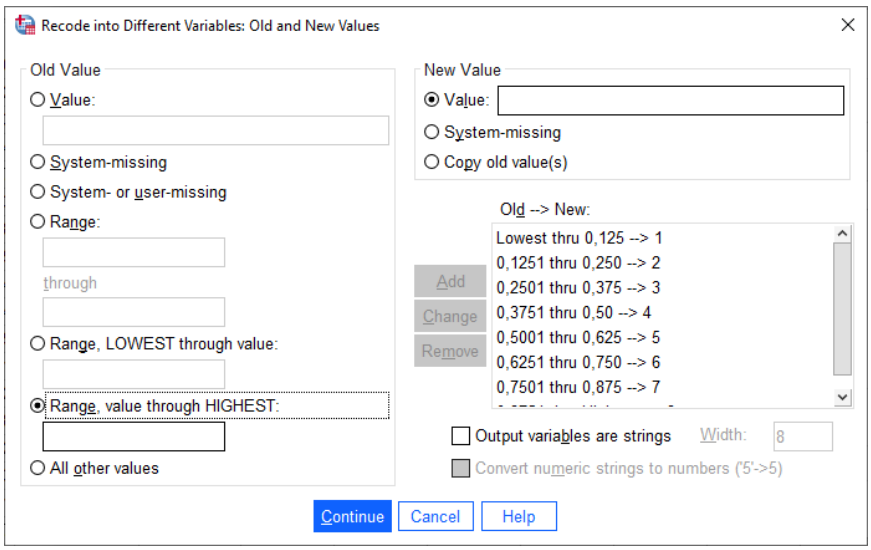
b. User-Specified

Εικόνα 13.12: Αποτέλεσμα ελέγχου One-Sample Kolmogorov-Smirnov test για τα δεδομένα του Πίνακα 13.6.

Στη συνέχεια, θα δούμε ότι με τη βοήθεια του SPSS μπορούμε να εφαρμόσουμε τον χι-τετράγωνο έλεγχο καλής προσαρμογής για να ελέγξουμε την υπόθεση ότι οι παρατηρήσεις προέρχονται από την κατανομή $\mathcal{G}(2,1/4)$. Θα βασιστούμε πάλι στον μετασχηματισμό των αρχικών παρατηρήσεων μέσω της α.σ.κ. της κατανομής $\mathcal{G}(2,1/4)$. Συγκεκριμένα, θα χρησιμοποιήσουμε τις μετασχηματισμένες τιμές $y_1, y_2, \ldots, y_{50}$, για να ελέγξουμε την υπόθεση ότι αυτές προέρχονται από την $\mathcal{U}(0,1)$ κατανομή.

Για την εφαρμογή του χι-τετράγωνο ελέγχου καλής προσαρμογής απαιτείται ο καθορισμός τάξεων (κατηγοριών) στις οποίες θα ομαδοποιήσουμε τις διαθέσιμες παρατηρήσεις. Επιπλέον, τα διαστήματα που θα χρησιμοποιήσουμε πρέπει να είναι τέτοια ώστε οι αναμενόμενες συχνότητες σε κάθε διάστημα (κατηγορία) να είναι $\geqslant 5$. Άρα, μπορούμε να θεωρήσουμε τα διαστήματα $[0,1/8]$, $(1/8,2/8]$, $(2/8,3/8]$, ..., $(7/8,1]$, όπου η πιθανότητα εμφάνισης μιας τιμής σε καθένα από αυτά τα διαστήματα είναι ίση με 1/8 (λόγω της υπόθεσης της κατανομής $\mathscr{U}(0,1)$ για τα μετασχηματισμένα δεδομένα).

Για να κωδικοποιήσουμε τις τιμές της στήλης Y επιλέγουμε **Transform / Recode into Different Variable** και, στο παράθυρο διαλόγου που ανοίγει, εισάγουμε στο πλαίσιο **Numeric Variable -> Output Variable** τη στήλη Y και στο πλαίσιο **Output Variable** δίνουμε Ycat. Πατάμε Change. Στη συνέχεια, επιλέγουμε Old and New Values (βλ. Εικόνα 13.13) και στο Range, LOWEST through value δίνουμε την τιμή 0.125, ενώ στο New Value / Value δίνουμε την τιμή 1 (1η κατηγορία). Για τη 2η κατηγορία, επιλέγουμε Old and New Values και στο Range δίνουμε 0.1251 through 0.250, ενώ στο New Value / Value δίνουμε την τιμή 2. Για την 3η κατηγορία, επιλέγουμε Old and New Values και στο Range δίνουμε 0.2501 through 0.375, ενώ στο New Value / Value δίνουμε την τιμή 3. Συνεχίζουμε με τον ίδιο τρόπο και για τις υπόλοιπες κατηγορίες. Για την 8η κατηγορία, επιλέγουμε Old and New Values και στο Range, value through HIGHEST δίνουμε την τιμή 0.8751, ενώ στο New Value / Value δίνουμε την τιμή 8. Πατάμε Continue και μετά OK και δημιουργείται μια νέα στήλη με τιμές 1, 2, ...8, με την αντιστοίχιση καθεμίας από τις 50 τιμές της στήλης Y στις 8 κατηγορίες.



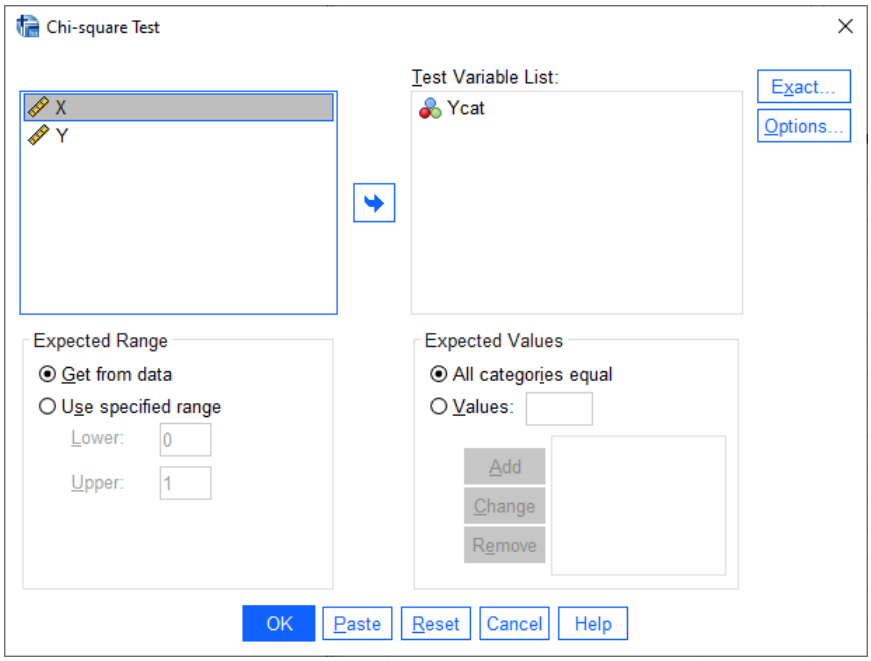
Εικόνα 13.13: Κωδικοποίηση μετασχηματισμένων τιμών σε 8 κλάσεις.

Στη συνέχεια, από το κεντρικό παράθυρο διαλόγου επιλέγουμε

Analyze / Nonparametric Tests / Legacy Dialogs / Chi-Square.

Στο παράθυρο διαλόγου που ανοίγει (βλ. Εικόνα 13.14) στο πλαίσιο Test Variable List εισάγουμε Ycat, στο πλαίσιο Expected Range επιλέγουμε Get from data, ενώ στο Expected Values, επιλέγουμε All categories equal. Πατάμε OK και προκύπτει το output της ανάλυσης.

Από την Εικόνα 13.15 έχουμε τον πίνακα με τις παρατηρούμενες και τις αναμενόμενες συχνότητες (υπό το μοντέλο της ομοιόμορφης κατανομής για τις τιμές στη στήλη Y). Επίσης, από την Εικόνα 13.16 η τιμή της στατιστικής συνάρτησης είναι ίση με 4.080, ενώ η p-τιμή του ελέγχου είναι 0.771. Άρα, δεν μπορούμε να απορρίψουμε την υπόθεση ότι οι (μετασχηματισμένες) τιμές $y_1, y_2, \ldots, y_{50}$ προέρχονται από την $\mathscr{U}(0,1)$



Εικόνα 13.14: Παράθυρο διαλόγου Chi-square test για τον χι-τετράγωνο έλεγχο καλής προσαρμογής των μετασχηματισμένων δεδομένων.

κατανομή. Δηλαδή, δεν μπορούμε να απορρίψουμε την υπόθεση ότι οι αρχικές παρατηρήσεις $x_1, x_2, \ldots, x_{50}$ προέρχονται από την κατανομή $\mathcal{G}(2,1/4)$.

Ycat

	Observed N	Expected N	Residual
1,00	7	6,3	,8
2,00	9	6,3	2,8
3,00	6	6,3	-,2
4,00	5	6,3	-1,2
5,00	7	6,3	,8
6,00	5	6,3	-1,2
7,00	3	6,3	-3,2
8,00	8	6,3	1,8
Total	50		

Εικόνα 13.15: Παρατηρούμενες και αναμενόμενες συχνότητες στις 8 κατηγορίες για τα δεδομένα της Ycat.

(με χρήση R): Εισάγουμε τα δεδομένα του Πίνακα 13.6 σε ένα διάνυσμα, έστω αυτό `x`. Αρχικά, θα χρησιμοποιήσουμε τον έλεγχο καλής προσαρμογής One-Sample Kolmogorov-Smirnov. Αυτό γίνεται με χρήση της εντολής `ks.test(...)`. Οι εντολές δίνονται παρακάτω:

Test Statistics

	Ycat
Chi-Square	4,080[a]
df	7
Asymp. Sig.	,771

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 6,3.

Εικόνα 13.16: Αποτελέσματα χι-τετράγωνο ελέγχου καλής προσαρμογής για τις τιμές της Ycat.

```
> x<-c(14.60,0.58,7.73,2.52,3.47,1.91,20.00,10.29,2.34,3.83,
+ 21.93,1.72,15.59,8.46,5.66,1.14,6.67,5.10,5.47,10.58,2.96,
+ 5.31,4.19,3.10,7.36,11.12,5.42,4.77,1.91,4.32,5.18,2.50,
+ 13.54,7.00,2.88,4.28,3.16,10.18,8.12,15.58,12.33,9.70,
+ 20.25,2.19,8.42,7.84,26.18,19.25,3.65,10.65)
> ks.test(x,'pgamma',2,1/4)
```

Κατά την εφαρμογή της εντολής `ks.test(...)` εισάγουμε, εκτός από το διάνυσμα `x`, το όνομα της κατανομής που έχουμε υποθέσει ως το θεωρητικό μοντέλο για την κατανομή του πληθυσμού (δηλαδή σε αυτήν την περίπτωση την `pgamma`), ενώ, καθώς ο έλεγχος είναι απλός, δίνουμε και τις τιμές των παραμέτρων της κατανομής. Αφού η κατανομή είναι η $\mathcal{G}(2,1/4)$, πρέπει να δώσουμε ως παράμετρο σχήματος (shape parameter) την τιμή 2 και ως παράμετρο ρυθμού (rate parameter) την τιμή 1/4. Το αποτέλεσμα της ανάλυσης δίνεται παρακάτω:

```
        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.10667, p-value = 0.62
alternative hypothesis: two-sided

Warning message:
In ks.test(x, 'pgamma', 2, 1/4) :
  ties should not be present for the Kolmogorov-Smirnov test
```

Η τιμή D της στατιστικής συνάρτησης ελέγχου, καθώς και η p-τιμή είναι ίδιες με τις αντίστοιχες τιμές που έδωσε το SPSS, στις γραμμές Most Extreme Differences: Absolute και Asymp. Sig. (2-tailed), αντίστοιχα, στην Εικόνα 13.12. Άρα, αφού η p-τιμή είναι 0.62, δεν μπορούμε να απορρίψουμε την υπόθεση ότι τα δεδομένα προέρχονται από κατανομή $\mathcal{G}(2,1/4)$.

Στη συνέχεια, θα εφαρμόσουμε τον χι-τετράγωνο έλεγχο καλής προσαρμογής. Αρχικά, μετασχηματίζουμε τις διαθέσιμες τιμές εφαρμόζοντας σε αυτές την α.σ.κ. της κατανομής $\mathcal{G}(2,1/4)$. Ακολούθως, χρησιμοποιούμε τις μετασχηματισμένες τιμές $y_1, y_2, ..., y_{50}$, για να ελέγξουμε την υπόθεση ότι αυτές προέρχονται από την κατανομή $\mathcal{U}(0,1)$. Όπως και στη λύση που δόθηκε με χρήση του SPSS, θα ομαδοποιήσουμε τις μετασχηματισμένες τιμές στα διαστήματα $[0,1/8]$, $(1/8,2/8]$, $(2/8,3/8]$, ..., $(7/8,1]$,

όπου η πιθανότητα εμφάνισης μιας τιμής σε καθένα από αυτά τα διαστήματα είναι ίση με 1/8 (αφού η κατανομή που έχουμε υποθέσει για τα μετασχηματισμένα δεδομένα είναι η $\mathcal{U}(0,1)$. Στη συνέχεια, αφού έχουμε κάνει την ομαδοποίηση, χρησιμοποιούμε την εντολή `chisq.test(...)`. Οι σχετικές εντολές δίνονται παρακάτω:

```
> y<-pgamma(x,2,1/4)# transformed values
> breaks1<-seq(0,1,by=1/8) # creating bins
> cutY<-cut(y,breaks1)# grouping the y values in the bins
> table(cutY) # frequency table
> OBS1<-c()# empty vector
> # put the observed frequencies
> # in the OBS1 vector
> for(j in 1:(length(breaks1)-1)){
+ OBS1[j]<-table(cutY)[[j]][1]
+ }
> # apply the GOF chi-square test
> chisq.test(OBS1,p=rep(1/length(OBS1),length(OBS1)),correct=TRUE,
   rescale.p=FALSE,simulate.p.value=FALSE)
```

Οι μετασχηματισμένες τιμές δίνονται στο διάνυσμα `y`, ενώ στο διάνυσμα `breaks1` έχουμε τα όρια των διαστημάτων. Με την εντολή `cut(y,breaks1)` ομαδοποιούμε τις τιμές που δόθηκαν στο `y` στα 8 διαστήματα, τα άκρα των οποίων δίνονται στο διάνυσμα `breaks`, το οποίο δημιουργείται στη δεύτερη γραμμή του κώδικα. Μπορούμε να έχουμε πίνακα συχνοτήτων του διανύσματος `cutY`, ο οποίος δεν είναι παρά ο πίνακας στην Εικόνα 13.15. Οι τιμές αυτές είναι οι παρατηρούμενες συχνότητες. Στη συνέχεια, καταχωρίζουμε σε ένα διάνυσμα (στο `OBS1`) τις παρατηρούμενες συχνότητες και χρησιμοποιούμε την εντολή `chisq.test`. Η χρήση αυτής της εντολής έχει επεξηγηθεί στο προηγούμενο παράδειγμα και για τον λόγο αυτόν, δεν δίνουμε στο σημείο αυτό περισσότερες πληροφορίες. Το αποτέλεσμα της ανάλυσης δίνεται παρακάτω:

```
cutY
(0,0.125] (0.125,0.25] (0.25,0.375]  (0.375,0.5]  (0.5,0.625] (0.625,0.75]
        7             9            6            5            7            5
(0.75,0.875]     (0.875,1]
           3             8

        Chi-squared test for given probabilities

data:  OBS1
X-squared = 4.08, df = 7, p-value = 0.7705
```

Παρατηρούμε ότι η τιμή X^2 της στατιστικής συνάρτησης ελέγχου, καθώς και η p-τιμή, είναι ίδιες με τις αντίστοιχες τιμές που έδωσε το SPSS (βλ. Εικόνα 13.16). Άρα, αφού η p-τιμή είναι 0.7705, δεν μπορούμε να απορρίψουμε την υπόθεση ότι τα μετασχηματισμένα δεδομένα προέρχονται από κατανομή $\mathcal{U}(0,1)$, δηλαδή δεν μπορούμε να απορρίψουμε την υπόθεση ότι οι αρχικές παρατηρήσεις προέρχονται από την κατανομή $\mathcal{G}(2,1/4)$. □

13.2.2 Έλεγχοι κανονικότητας

Λόγω της σπουδαιότητας της κανονικής κατανομής, σε αυτήν την υποενότητα, θα γίνει ειδική αναφορά στην υλοποίηση των τρόπων ελέγχου της καλής προσαρμογής των δεδομένων στο μοντέλο της κανονικής

κατανομής (βλ. Ενότητα 4.4).

Παράδειγμα 13.6. (**Έλεγχος κανονικότητας για ένα δείγμα**): Στον Πίνακα 13.7 δίνονται οι μετρήσεις ενός τυχαίου δείγματος μεγέθους 22 από έναν πληθυσμό με άγνωστη συνεχή κατανομή. Να ελέγξετε την υπόθεση της κανονικότητας για το διαθέσιμο δείγμα παρατηρήσεων. Να χρησιμοποιηθούν οι έλεγχοι κανονικότητας Kolmogorov-Smirnov, με χρήση της διόρθωσης Anderson-Darling, ο έλεγχος Lilliefors, καθώς και ο έλεγχος των Shapiro-Wilk.

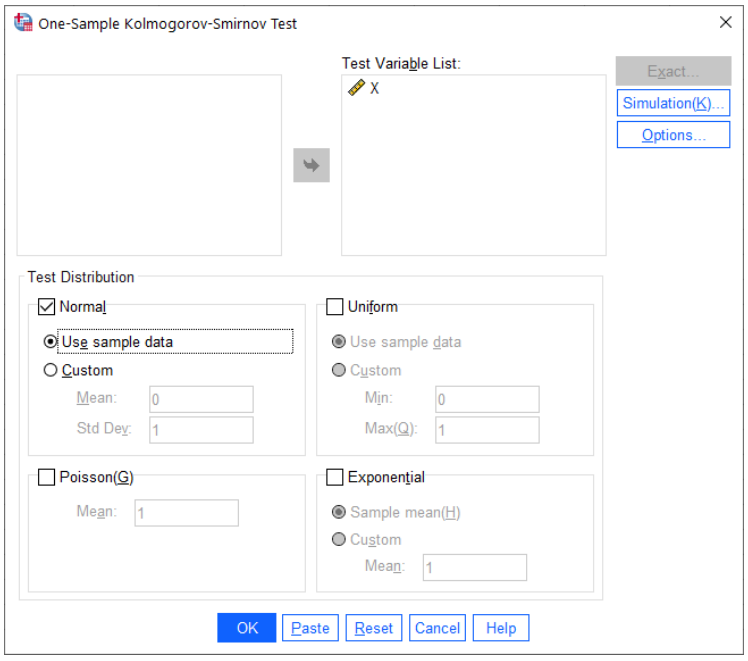
247.61,	249.14,	247.57,	250.10,	251.26,	246.32,	252.96,	251.51,	252.48,	242.43,	246.36,
252.37,	249.80,	250.64,	245.11,	247.08,	253.47,	246.53,	251.69,	252.52,	249.57,	251.62

Πίνακας 13.7: Δεδομένα από άγνωστη συνεχή κατανομή.

Λύση Παραδείγματος 13.6. (με χρήση SPSS): Θα χρησιμοποιήσουμε το SPSS και θα εφαρμόσουμε το τεστ των Kolmogorov-Smirnov. Πριν την εφαρμογή του, αξίζει να παρατηρήσουμε ότι οι παράμετροι της υποτιθέμενης κατανομής (κανονική) δεν είναι γνωστές και άρα θα εκτιμηθούν από τα δεδομένα. Αρχικά, εισάγουμε σε μια στήλη ενός κενού φύλλου εργασίας του SPSS τα παραπάνω δεδομένα και τη μετονομάζουμε σε X. Στη συνέχεια, επιλέγουμε από το κεντρικό παράθυρο διαλόγου:

Analyze / Nonparametric Tests / Legacy Dialogs / 1-Sample K-S.

Στο νέο παράθυρο διαλόγου που ανοίγει (βλ. Εικόνα 13.17) εισάγουμε στο πλαίσιο Test Variable List τη X, επιλέγουμε Test Distribution: Normal και αφήνουμε την προεπιλογή Use sample data. Πατάμε OK και προκύπτει το output της ανάλυσης που δίνεται στην Εικόνα 13.18.



Εικόνα 13.17: Παράθυρο διαλόγου One-Sample Kolmogorov-Smirnov Test για έλεγχο κανονικότητας.

Στο output της ανάλυσης δίνεται η τιμή της στατιστικής συνάρτησης ελέγχου Z (γραμμή Test Statistic), η οποία αντιστοιχεί στην απόλυτη τιμή της μεγαλύτερης διαφοράς μεταξύ της υποτιθέμενης κατανομής (υπό την H_0) και της εμπειρικής συνάρτησης κατανομής. Η τιμή αυτή δίνεται και στη γραμμή Absolute, στο Most Extreme Differences. Παρατηρήστε, ακόμα, ότι οι παράμετροι της κατανομής έχουν εκτιμηθεί από τα δεδομένα και είναι ίσες με $\hat{\mu} = 249.4155$ και $\hat{\sigma} = 2.91375$. Η p-τιμή του ασυμπτωτικού ελέγχου δίνεται ότι είναι ίση με

0.200. Το SPSS μας ενημερώνει ότι έχει εφαρμοστεί η διόρθωση κατά Lilliefors λόγω του ότι οι παράμετροι της κατανομής υπό την H_0 δεν είναι γνωστές και έχουν εκτιμηθεί. Επίσης, η p-τιμή έχει εκτιμηθεί και μέσω προσομοίωσης (δίνεται και ένα 99% διάστημα εμπιστοσύνης για αυτήν). Η p-τιμή του ελέγχου εκτιμάται (με βεβαιότητα 99%) ότι είναι μεταξύ 0.240 και 0.262 και άρα, δεν μπορούμε να απορρίψουμε την υπόθεση ότι τα δεδομένα προέρχονται από κανονική κατανομή.

One-Sample Kolmogorov-Smirnov Test

			X
N			22
Normal Parameters[a,b]	Mean		249,4609
	Std. Deviation		2,95564
Most Extreme Differences	Absolute		,138
	Positive		,098
	Negative		-,138
Test Statistic			,138
Asymp. Sig. (2-tailed)[c]			,200[d]
Monte Carlo Sig. (2-tailed)[e]	Sig.		,331
	99% Confidence Interval	Lower Bound	,318
		Upper Bound	,343

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

d. This is a lower bound of the true significance.

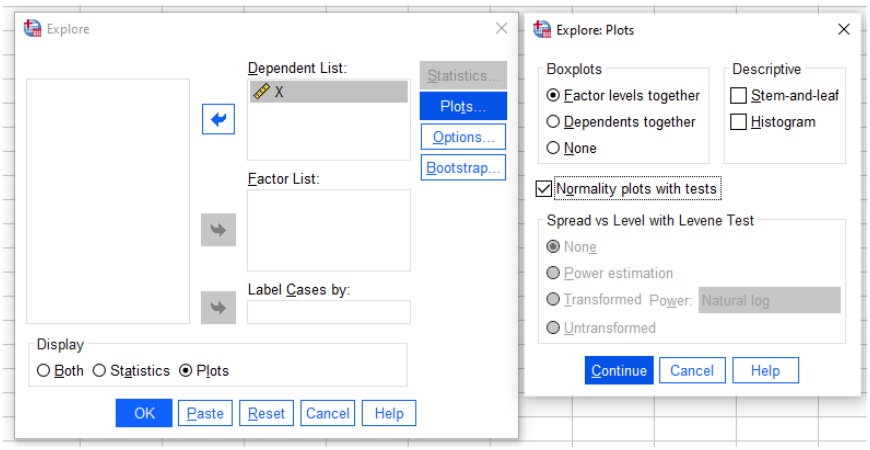
e. Lilliefors' method based on 10000 Monte Carlo samples with starting seed 2000000.

Εικόνα 13.18: Αποτέλεσμα ελέγχου One-Sample KS test για τα δεδομένα του Πίνακα 13.7.

Για την εφαρμογή του ελέγχου κανονικότητας Shapiro-Wilk με χρήση του SPSS, επιλέγουμε από το κεντρικό παράθυρο διαλόγου:

Analyze / Descriptive Statistics / Explore.

Στο παράθυρο διαλόγου που ανοίγει (βλ. Εικόνα 13.19) εισάγουμε στο πλαίσιο Dependent List τη X και στο πλαίσιο Display επιλέγουμε Plots. Επιλέγουμε Plots, στη συνέχεια επιλέγουμε Normality plots with tests και πατάμε Continue και μετά OK και προκύπτει το output της ανάλυσης, το οποίο δίνεται στην Εικόνα 13.20.



Εικόνα 13.19: Παράθυρο διαλόγου Explore και επιλογή ελέγχων κανονικότητας.

Από το output της ανάλυσης έχουμε ότι η τιμή της στατιστικής συνάρτησης ελέγχου για τον έλεγχο των Shapiro-Wilk είναι ίση με 0.938 (στήλη Statistic), ενώ η p-τιμή του ελέγχου είναι ίση με 0.178. Άρα, σε ε.σ. 5%, δεν μπορούμε να απορρίψουμε την υπόθεση της κανονικότητας. Αξίζει να σημειωθεί ότι μέσα από τη διαδικασία Explore, εκτός του ελέγχου για κανονικότητα των Shapiro-Wilk, διεξάγεται και ο έλεγχος για κανονικότητα των Kolmogorov-Smirnov με χρήση της διόρθωσης Lilliefors. Για τον συγκεκριμένο έλεγχο

δίνεται ένα κάτω όριο για την p-τιμή και, αφού αυτή είναι μεγαλύτερη από 0.200, δεν μπορούμε να απορρίψουμε την υπόθεση της κανονικότητας.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
X	,138	22	,200*	,938	22	,178

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Εικόνα 13.20: Αποτελέσματα ελέγχων κανονικότητας Shapiro-Wilk και One-Sample KS test με χρήση της διόρθωσης Lilliefors.

(με χρήση R): Θα χρησιμοποιήσουμε την `R` και θα εφαρμόσουμε το τεστ των Kolmogorov-Smirnov με χρήση της διόρθωσης των Anderson-Darling, καθώς και τον έλεγχο των Shapiro-Wilk. Για να το κάνουμε αυτό, θα φορτώσουμε αρχικά το πακέτο `nortest`. Στη συνέχεια, καταχωρίζουμε τα δεδομένα του Πίνακα 13.7 σε ένα διάνυσμα, έστω αυτό `x`. Για τον έλεγχο των Shapiro-Wilk, χρησιμοποιείται η εντολή `shapiro.test(...)`, ενώ για τον έλεγχο των Kolmogorov-Smirnov, σύμφωνα με την τροποποίηση που προτάθηκε από τους Anderson-Darling, πρέπει να χρησιμοποιήσουμε την εντολή `ad.test(...)` από το πακέτο `nortest`. Επίσης, στο ίδιο πακέτο, υπάρχει η εντολή `lillie.test(...)`, η οποία διεξάγει τον έλεγχο των Kolmogorov-Smirnov με τη διόρθωση κατά Lilliefors. Παρακάτω δίνουμε τις σχετικές εντολές μαζί με τα αποτελέσματα των ελέγχων.

```
> library(nortest)
> x<-c(247.61,249.14,247.57,250.10,251.26,246.32,252.96,251.51,
+ 252.48,242.43,246.36,252.37,249.80,250.64,245.11,247.08,253.47,
+ 246.53,251.69,252.52,249.57,251.62)
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.93781, p-value = 0.1783

> ad.test(x) # anderson-darling

        Anderson-Darling normality test

data:  x
A = 0.49474, p-value = 0.1929

> lillie.test(x) # lilliefors test

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  x
D = 0.13773, p-value = 0.3411
```

Από τα αποτελέσματα των τεστ βλέπουμε ότι δεν απορρίπτεται η υπόθεση της κανονικότητας. Το τεστ των Shapiro-Wilk έδωσε τιμή 0.93947 για τη στατιστική συνάρτηση ελέγχου και p-τιμή ίση με $0.193 > 0.05$. Σημειώνουμε ότι τις ίδιες τιμές έδωσε και ο έλεγχος με χρήση του SPSS. Επίσης, η τιμή της στατιστικής συνάρτησης για τον έλεγχο των Anderson-Darling είναι 0.49474, με p-τιμή ίση με $0.1929 > 0.05$. Τέλος, ο έλεγχος για κανονικότητα με το τεστ του Lilliefors έδωσε τιμή στατιστικής συνάρτησης ίση με 0.13773 και p-τιμή ίση με $0.3411 > 0.05$. □

13.3 Έλεγχοι υποθέσεων βασισμένοι στη διωνυμική κατανομή

Στην ενότητα αυτή θα παρουσιαστεί, μέσω παραδειγμάτων, η υλοποίηση, με χρήση του SPSS αλλά και της R, του ελέγχου του προσημικού κριτηρίου (sign test), του προσημικού ελέγχου του McNemar, του διωνυμικού ελέγχου (Binomial test) και του διωνυμικού ελέγχου για ποσοστιαία σημεία. Οι έλεγχοι αυτοί είναι οι πιο γνωστοί και συχνά χρησιμοποιούμενοι έλεγχοι που βασίζονται στη διωνυμική κατανομή και αποτέλεσαν αντικείμενο μελέτης στο Κεφάλαιο 5 του παρόντος συγγράμματος.

Παράδειγμα 13.7. (**Έλεγχος Sign test**): Στον Πίνακα 13.7 δίνονται μετρήσεις που αφορούν την αρτηριακή πίεση 15 ασθενών, πριν και μετά τη λήψη ενός χαπιού. Χρησιμοποιώντας ε.σ. $a = 5\%$, να ελέγξετε την υπόθεση ότι η διάμεση πίεση, πριν και μετά τη λήψη χαπιού, είναι διαφορετική.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X_i	125	115	130	140	142	117	143	123	139	134	136	118	121	122	133
Y_i	110	122	125	120	140	124	123	137	135	145	144	112	119	121	118

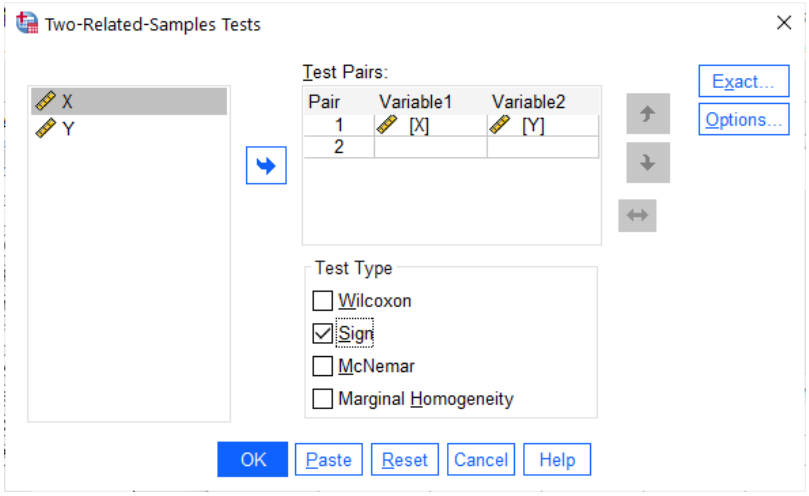
Πίνακας 13.8: Δεδομένα αρτηριακής πίεσης για $n = 15$ ασθενείς.

Λύση Παραδείγματος 13.7. Έστω X (Y, αντίστοιχα) η τ.μ. που παριστάνει την τιμή της αρτηριακής πίεσης πριν (μετά, αντίστοιχα) τη λήψη του χαπιού. Η υπόθεση που θέλουμε να ελέγξουμε είναι η $H_0 : P(X < Y) = P(X > Y)$ έναντι $H_1 :$ όχι η H_0. Για τον συγκεκριμένο έλεγχο μπορούμε να χρησιμοποιήσουμε τον Προσημικό έλεγχο (sign test) που παρουσιάστηκε στην Ενότητα 5.4. Χρησιμοποιώντας τα αριθμητικά δεδομένα που δίνονται, θα υλοποιήσουμε τον έλεγχο με χρήση του SPSS αλλά και της R.

(με χρήση SPSS) Εισάγουμε αρχικά τα παραπάνω δεδομένα σε δύο στήλες ενός φύλλου εργασίας του SPSS και μετονομάζουμε τις στήλες σε X και Y, αντίστοιχα. Έπειτα, επιλέγουμε από το κεντρικό παράθυρο διαλόγου:

Analyze / Nonparametric Tests / 2 Related Samples

και στο νέο παράθυρο διαλόγου που προκύπτει (βλ. Εικόνα 13.21) εισάγουμε τις μεταβλητές X, Y στα πεδία Variable 1 και Variable 2, αντίστοιχα. Στο Test Type επιλέγουμε Sign και πατάμε OK.



Εικόνα 13.21: Επιλογές του Nonparametric Tests / 2 Related Samples του SPSS.

Τα αποτελέσματα της παραπάνω ανάλυσης δίνονται στην Εικόνα 13.22 και στην Εικόνα 13.23. Ειδικότερα, στην Εικόνα 13.22 δίνεται το πλήθος των αρνητικών και θετικών διαφορών $Y - X$ (difference). Το πλήθος

αυτών των διαφορών αποτελεί ένδειξη για απόρριψη ή όχι της H_0. Ένας πολύ μεγάλος ή ένας πολύ μικρός αριθμός διαφορών, ενδέχεται να οδηγήσει σε απόρριψη της H_0.

Frequencies

		N
Y - X	Negative Differences[a]	10
	Positive Differences[b]	5
	Ties[c]	0
	Total	15

a. Y < X

b. Y > X

c. Y = X

Εικόνα 13.22: Πλήθος θετικών και αρνητικών διαφορών.

Τέλος, από την Εικόνα 13.23 παρατηρούμε ότι το SPSS εκτελεί το ακριβές sign test, με χρήση της διωνυμικής κατανομής για τον υπολογισμό της p-τιμής. Η p-τιμή είναι ίση με $0.302 > 0.05$, από την οποία συμπεραίνουμε πως, σε ε.σ. 5%, δεν απορρίπτεται η H_0. Επομένως, σε ε.σ. 5% δεν μπορούμε να απορρίψουμε την υπόθεση ότι οι τιμές αρτηριακής πίεσης, πριν και μετά τη λήψη του χαπιού, προέρχονται από τον ίδιο πληθυσμό. Δηλαδή, δεν μπορούμε να απορρίψουμε, σε ε.σ. 5%, την υπόθεση ότι η λήψη του χαπιού δεν επιφέρει στατιστικά σημαντική διαφοροποίηση στη διάμεση τιμή της αρτηριακής πίεσης.

Test Statistics[a]

	Y - X
Exact Sig. (2-tailed)	,302[b]

a. Sign Test

b. Binomial distribution used.

Εικόνα 13.23: p-τιμή για το Προσημικό κριτήριο.

(με χρήση R) Αρχικά, εισάγουμε τα δεδομένα σε δύο διανύσματα, έστω αυτά `x` και `y`. Με χρήση των εντολών `length(which(x>y))` βρίσκουμε το πλήθος των περιπτώσεων $X > Y$ και αυτό μας δίνει τη δυνατότητα να μπορούμε να εφαρμόσουμε τον διωνυμικό έλεγχο (binomial test) χρησιμοποιώντας την εντολή `binom.test`. Συγκεκριμένα, εισάγουμε ως `x` το πλήθος των περιπτώσεων $X > Y$ (εδώ είναι 10), ενώ εισάγουμε ως `n` το πλήθος των δοκιμών (εδώ ταυτίζονται με το πλήθος των τιμών στα `x` και `y`). Ως `p` δίνουμε την τιμή 0.5, ενώ με το όρισμα `alternative = 'two.sided'` επιλέγουμε να κάνουμε τον έλεγχο με δίπλευρη εναλλακτική. Αν θέλουμε να κάνουμε κάποιον από τους μονόπλευρους ελέγχους, δίνουμε `alternative = 'less'` ή `alternative = 'greater'`, οπότε οι αντίστοιχες εναλλακτικές είναι $H_1 : p < 0.5$ και $H_1 : p > 0.5$. Επίσης, με χρήση του ορίσματος `conf.level = 0.95`, επιλέγουμε συντελεστή εμπιστοσύνης ίσο με 95% για το διάστημα εμπιστοσύνης για την πραγματική πιθανότητα επιτυχίας $p = P(X > Y)$. Παρακάτω δίνονται οι εντολές, καθώς και το output της ανάλυσης.

```
1 > x<-c(125,115,130,140,142,117,143,123,139,134,136,118,121,122,133)
2 > y<-c(110,122,125,120,140,124,123,137,135,145,144,112,119,121,118)
3 > length(which(x>y))
```

```
> binom.test(x=length(which(x>y)), n=length(x), p = 0.5,
+ alternative = 'two.sided',conf.level = 0.95)
```

```
Exact binomial test
data:  length(which(x > y)) and length(x)
number of successes = 10, number of trials = 15, p-value = 0.3018
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3838037 0.8817589
sample estimates:
probability of success
             0.6666667
```

Από τα αποτελέσματα της ανάλυσης διαπιστώνουμε ότι η τιμή της σ.σ.ε. είναι ίση με 10, δηλαδή είναι ίση με το πλήθος των περιπτώσεων $\{X_i > Y_i\}$. Επίσης, η p-τιμή είναι ίση με $0.3018 > 0.05$ (ακριβώς ίδια με αυτήν που δίνει το SPSS) και άρα συμπεραίνουμε ότι δεν απορρίπτεται η H_0 σε ε.σ. 5%. Επίσης, παρατηρήστε ότι δίνεται και ένα 95% διάστημα εμπιστοσύνης για την πραγματική πιθανότητα $p = P(X > Y)$. Επισημαίνεται ότι ο υπολογισμός του συγκεκριμένου διαστήματος γίνεται με βάση τη σχέση των Clopper and Pearson (1934) (βλ. σχετικά Κεφάλαιο 5). Από το διάστημα εμπιστοσύνης προκύπτει ότι εκτιμούμε, λοιπόν, ότι, με βεβαιότητα 95%, η πραγματική πιθανότητα $p = P(X > Y)$ είναι μεταξύ 0.3838 και 0.8818. □

Παράδειγμα 13.8. (**Έλεγχος McNemar test - Περίπτωση μη ομαδοποιημένων δεδομένων**): Ένας κλινικός διατροφολόγος θέλοντας να ελέγξει την επίδραση μιας δίαιτας σε ενήλικες άνδρες καταγράφει σε ένα δείγμα 20 τυχαία επιλεγμένων ανδρών την κατάστασή τους πριν και μετά την εφαρμογή της δίαιτας. Στον Πίνακα 13.8 δίνονται τα δεδομένα αυτά, όπου X είναι η καταγραφή πριν την εφαρμογή της δίαιτας (1 για υπέρβαρους, 0 για μη υπέρβαρους) και Y είναι η καταγραφή μετά την εφαρμογή της δίαιτας. Χρησιμοποιήστε το τεστ του McNemar για να ελέγξετε, σε ε.σ. 5%, το κατά πόσο η δίαιτα έχει στατιστικά σημαντική επίδραση στο βάρος των ασθενών.

Λύση Παραδείγματος 13.8. Έστω X (Y, αντίστοιχα) η τ.μ. που παριστάνει αν ένας άνδρας είναι υπέρβαρος ή όχι πριν (μετά, αντίστοιχα) την εφαρμογή της δίαιτας. Δηλαδή οι τ.μ. λαμβάνουν δύο τιμές, έστω «0» και «1», και έχοντας εξαρτημένα δείγματα, αφού οι μετρήσεις γίνονται στο ίδιο άτομο πριν και μετά την εφαρμογή της δίαιτας, αυτό που μας ενδιαφέρει είναι να ελέγξουμε αν τα άτομα έχουν υποστεί αλλαγή στη συμπεριφορά τους μετά τη θεραπεία. Για αυτόν τον σκοπό ελέγχουμε το παρακάτω πρόβλημα ελέγχου υποθέσεων H_0 : $P(X_i = 0, Y_i = 1) = P(X_i = 1, Y_i = 0)$, για κάθε i, έναντι της H_1 : $P(X_i = 0, Y_i = 1) \neq P(X_i = 1, Y_i = 0)$, για τουλάχιστον ένα i.

Για τον συγκεκριμένο έλεγχο μπορούμε να χρησιμοποιήσουμε τον έλεγχο McNemar (βλ. σχετικά Κεφάλαιο 5, Ενότητα 5.5.1). Χρησιμοποιώντας τα αριθμητικά δεδομένα που δίνονται θα υλοποιήσουμε τον έλεγχο με χρήση του SPSS αλλά και της R.

(**με χρήση SPSS**) Εισάγουμε, αρχικά, τα παραπάνω δεδομένα σε δύο στήλες ενός φύλλου εργασίας του SPSS και τις μετονομάζουμε σε X και Y, αντίστοιχα. Στη συνέχεια, από το κεντρικό παράθυρο διαλόγουμε επιλέγουμε:

Analyze / Nonparametric Tests / Legacy Dialogs / 2 Related Samples

και στο νέο παράθυρο διαλόγου που ανοίγει (βλ. Εικόνα 13.24) εισάγουμε τη X στο πεδίο Variable 1 και την Y στο πεδίο Variable 2, ενώ στο πεδίο Test Type επιλέγουμε McNemar και πατάμε OK.

A/A	X	Y	A/A	X	Y
1	1	0	21	0	0
2	0	0	22	0	0
3	1	0	23	1	1
4	1	1	24	0	0
5	0	0	25	1	0
6	0	0	26	0	0
7	1	0	27	1	0
8	1	1	28	1	1
9	1	0	29	0	0
10	0	0	30	0	0
11	0	0	31	1	1
12	0	0	32	0	0
13	0	1	33	0	1
14	0	0	34	0	0
15	1	0	35	0	0
16	1	0	36	1	0
17	0	0	37	0	0
18	1	0	38	1	1
19	1	1	39	0	0
20	0	0	40	0	1

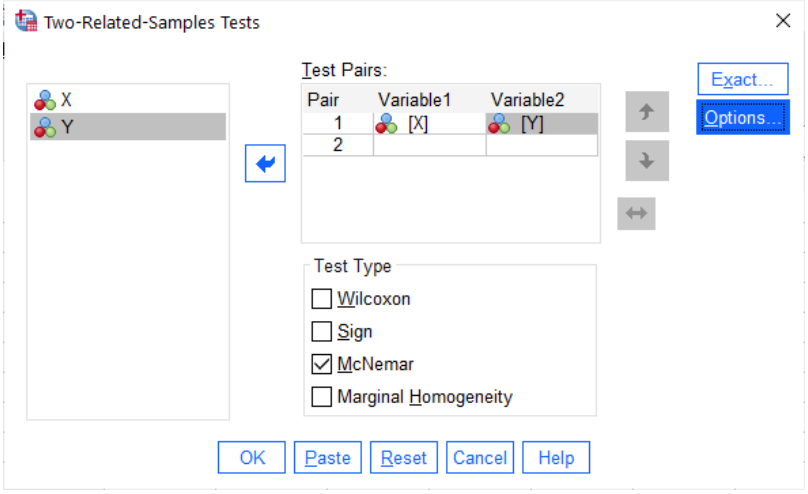
Πίνακας 13.9: Καταγραφή Υπέρβαρων - Μη υπέρβαρων ανδρών, πριν (X) και μετά (Y) την εφαρμογή της δίαιτας.

Το output της ανάλυσης δίνεται στις Εικόνες 13.25 και 13.26. Συγκεκριμένα, στην Εικόνα 13.25 βλέπουμε ότι πριν την εφαρμογή της δίαιτας, στο δείγμα των 40 ατόμων, υπάρχουν 23 μη υπέρβαροι και 17 υπέρβαροι. Επίσης, παρατηρούμε ότι 3 άτομα από τα αρχικά μη υπέρβαρα βρέθηκαν, μετά την εφαρμογή της δίαιτας, να είναι υπέρβαρα, ενώ 10 άτομα τα οποία ήταν υπέρβαρα πριν την εφαρμογή της δίαιτας, ταξινομήθηκαν ως μη υπέρβαρα μετά την εφαρμογή αυτής.

Το αποτέλεσμα του ελέγχου δίνεται στην Εικόνα 13.26 όπου έχει χρησιμοποιηθεί η ακριβής κατανομή (διωνυμική) της στατιστικής συνάρτησης ελέγχου. Η p-τιμή είναι 0.092 και σε ε.σ. 5% δεν απορρίπτουμε την H_0. Επομένως, σε ε.σ. 5% δεν μπορούμε να ισχυριστούμε ότι η δίαιτα είχε στατιστικά σημαντική επίδραση στην απώλεια βάρους.

(**με χρήση R**) Αρχικά, εισάγουμε τα δεδομένα σε δύο διανύσματα, έστω αυτά `x` (αποτέλεσμα πριν την εφαρμογή της δίαιτας) και `y` (αποτέλεσμα μετά την εφαρμογή της δίαιτας). Στη συνέχεια, χρησιμοποιούμε την εντολή `mcnemar.test(...)`. Συγκεκριμένα, εισάγουμε ως ορίσματα τα διανύσματα `x` και `y`. Επίσης, με χρήση του ορίσματος `correct = TRUE` εφαρμόζεται η ασυμπτωτική μορφή του ελέγχου του McNemar με χρήση της διόρθωσης συνεχείας (βλ. Κεφάλαιο 5). Η προεπιλεγμένη τιμή για τον συντελεστή εμπιστοσύνης είναι 0.95. Παρακάτω δίνονται οι εντολές στην `R`, καθώς και το αποτέλεσμα της ανάλυσης.

```
> x<-c(1,0,1,1,0,0,1,1,1,0,0,0,0,0,1,1,0,1,1,0,0,0,1,0,1,0,
+ 1,1,0,0,1,0,0,0,0,1,0,1,0,0)
> y<-c(0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,0,0,1,0,0,0,1,0,0,0,
+ 0,1,0,0,1,0,1,0,0,0,0,1,0,1)
> mcnemar.test(x, y, correct = TRUE)
```



Εικόνα 13.24: Παράθυρο διαλόγου Nonparametric Tests / Two Related Samples του SPSS.

X & Y

X	Y: no overweight	Y: overweight
no overweight	20	3
overweight	10	7

Εικόνα 13.25: Πίνακας Ταξινόμησης δεδομένων του Παραδείγματος 13.8.

```
McNemar's Chi-squared test with continuity correction
data:  x and y
McNemar's chi-squared = 2.7692, df = 1, p-value = 0.09609
```

Από τα αποτελέσματα της ανάλυσης διαπιστώνουμε ότι η p-τιμή είναι ίση με $0.09609 > 0.05$ και, επομένως, συμπεραίνουμε ότι δεν απορρίπτεται η H_0 σε ε.σ. 5%. Τέλος, σημειώνουμε ότι η p-τιμή που δίνει η R είναι ελαφρώς διαφορετική από αυτήν που δίνει το SPSS, καθώς στην R έχει χρησιμοποιηθεί η ασυμπτωτική κατανομή του ελέγχου με χρήση της διόρθωσης συνεχείας, ενώ το SPSS χρησιμοποιεί τη διωνυμική κατανομή (ακριβή μορφή του τεστ) για τον υπολογισμό της p-τιμής. Για τον υπολογισμό της p-τιμής του ελέγχου McNemar με χρήση της διωνυμικής κατανομής δεν υπάρχει έτοιμη εντολή στην R. Για τον λόγο αυτό προτείνεται η χρήση των παρακάτω εντολών (βλ., επίσης, Ενότητα 5.5.1). Άμεσα διαπιστώνουμε ότι επιβεβαιώνεται η p-τιμή που δόθηκε από το SPSS. □

```
> n<-sum(x!=y) # calculate the sample size n = beta + gamma
> tau<-sum(x<y) # calculate the value of T statistic
> # pvalue calculation
> pvalue<-ifelse(tau>n/2,2*pbinom(tau,n,0.5,lower.tail=F),
+ 2*pbinom(tau,n,0.5))
> pvalue # print pvalue
[1] 0.09228516
```

Παράδειγμα 13.9. (**Έλεγχος McNemar test - Περίπτωση ομαδοποιημένων δεδομένων**): Εταιρεία δημοσκοπήσεων θέλει να διαπιστώσει αν η επερχόμενη τηλεμαχία (debate) μεταξύ δύο υποψηφίων (έστω αυτοί Α και Β) για τη δημαρχία μιας πόλης θα έχει στατιστικά σημαντική επίδραση στο εκλογικό σώμα. Για τον λόγο αυτό, μία ημέρα πριν την τηλεμαχία επιλέγει από το εκλογικό σώμα, με τυχαίο τρόπο, δείγμα 785

Test Statistics[a]

	X & Y
N	40
Exact Sig. (2-tailed)	,092[b]

a. McNemar Test

b. Binomial distribution used.

Εικόνα 13.26: Αποτέλεσμα ελέγχου McNemar με χρήση SPSS.

ατόμων και καταγράφει την πρόθεση ψήφου τους (υπέρ του Α ή υπέρ του Β). Στη συνέχεια, μία ημέρα μετά την τηλεμαχία, ρωτάει το ίδιο ακριβώς δείγμα ατόμων ως προς την πρόθεση ψήφου. Χρησιμοποιώντας τα δεδομένα που δίνονται στον Πίνακα 13.10 να ελέγξετε σε ε.σ. 1% την υπόθεση ότι η εμφάνιση των δύο υποψηφίων στην τηλεμαχία, δεν επηρέασε την πρόθεση ψήφου.

		Μετά την τηλεμαχία	
		A	B
Πριν την	A	396	72
τηλεμαχία	B	6	311

Πίνακας 13.10: Δεδομένα πρόθεσης ψήφου πριν και μετά την τηλεμαχία.

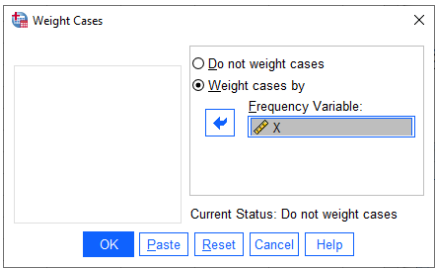
Λύση Παραδείγματος 13.9. Έστω X (Y, αντίστοιχα) η τ.μ. που παριστάνει αν ο/η ερωτώμενος/ερωτώμενη είναι υπέρ του Α ή του Β υποψήφιου, δηλαδή λαμβάνει δύο τιμές, έστω «0» και «1», πριν (μετά, αντίστοιχα) την τηλεμαχία. Αυτό που μας ενδιαφέρει είναι να ελέγξουμε αν τα άτομα έχουν υποστεί αλλαγή στη συμπεριφορά τους μετά την τηλεμαχία, για αυτόν τον σκοπό εξετάζουμε το παρακάτω πρόβλημα ελέγχου υποθέσεων H_0 : $P(X_i = 0, Y_i = 1) = P(X_i = 1, Y_i = 0)$, για κάθε i, έναντι της εναλλακτικής H_1 : $P(X_i = 0, Y_i = 1) \neq P(X_i = 1, Y_i = 0)$, για τουλάχιστον ένα i. Για τον συγκεκριμένο έλεγχο μπορούμε να χρησιμοποιήσουμε τον έλεγχο McNemar (βλ. σχετικά Κεφάλαιο 5, Ενότητα 5.5.1). Χρησιμοποιώντας τα αριθμητικά δεδομένα που δίνονται θα υλοποιήσουμε τον έλεγχο με χρήση του SPSS αλλά και της R.

(**με χρήση SPSS**): Για να εφαρμόσουμε το τεστ του McNemar όταν τα δεδομένα δίνονται ομαδοποιημένα και σε έναν πίνακα συνάφειας 2×2, θα δουλέψουμε με έναν διαφορετικό τρόπο σε σχέση με αυτόν που δουλέψαμε στο Παράδειγμα 13.8. Αρχικά, σε ένα κενό φύλλο εργασίας του SPSS εισάγουμε σε μια στήλη τις τιμές 396, 6, 72, 311 και μετονομάζουμε τη στήλη σε X. Στη συνέχεια, εισάγουμε στην αμέσως επόμενη στήλη τις τιμές A, B, A, B και τη μετονομάζουμε σε ROW. Τέλος, εισάγουμε σε μια 3η στήλη τις τιμές A, A, B, B και τη μετονομάζουμε σε COL. Για να δώσουμε στο SPSS να «καταλάβει» ότι έχουμε δεδομένα της μορφής του Πίνακα 13.10, επιλέγουμε **Data / Weight Cases** και στο παράθυρο διαλόγου που ανοίγει (βλ. Εικόνα 13.27) επιλέγουμε Weight cases by και στο Frequency Variable εισάγουμε τη στήλη X και πατάμε OK. Μετά από την εισαγωγή των δεδομένων, κατά αυτόν τον τρόπο, από το κεντρικό παράθυρο διαλόγου επιλέγουμε:

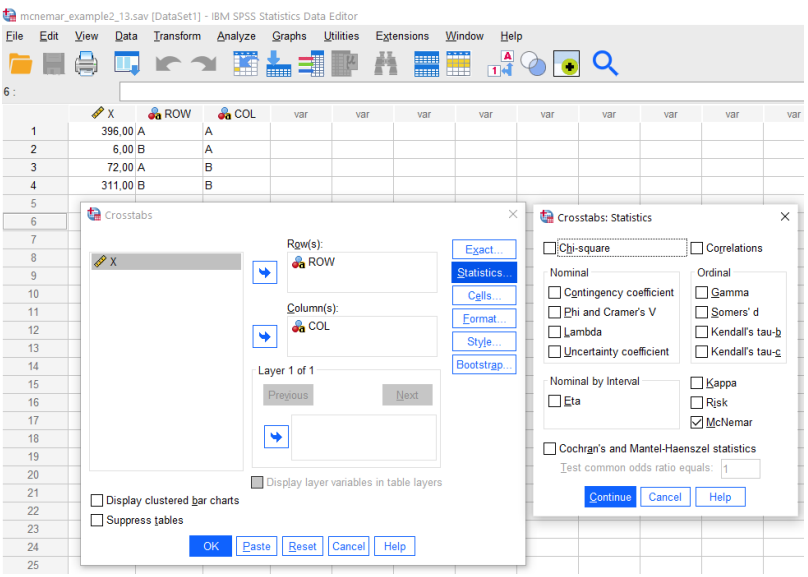
Analyze / Descriptive Statistics / Crosstabs

Στο νέο παράθυρο διαλόγου που προκύπτει (βλ. Εικόνα 13.28) τοποθετούμε στο Row(s) τη στήλη με όνομα ROW και στο Column(s) τη στήλη με όνομα COL. Στη συνέχεια, πατάμε το πεδίο Statistics και επιλέγουμε McNemar. Τέλος, πατάμε Continue και μετά OK.

Το output της ανάλυσης δίνεται στις Εικόνες 13.29 και 13.30. Συγκεκριμένα, στην Εικόνα 13.29 βλέπουμε ότι πριν την τηλεμαχία, στο δείγμα των 785 ατόμων, οι 468 προτίθενται να ψηφίσουν τον υποψήφιο Α και οι



Εικόνα 13.27: Το παράθυρο της διαδικασίας Weight Cases.



Εικόνα 13.28: Το παράθυρο της ανάλυσης Crosstabs.

υπόλοιποι 317 τον υποψήφιο Β. Ωστόσο, μετά την τηλεμαχία, υπάρχουν 6 μέλη του δείγματος που μετακινήθηκαν από τον υποψήφιο Β στον υποψήφιο Α, αλλά και 72 μέλη του δείγματος που μετακινήθηκαν από τον υποψήφιο Α στον υποψήφιο Β.

ROW * COL Crosstabulation

Count

		COL		
		A	B	Total
ROW	A	396	72	468
	B	6	311	317
Total		402	383	785

Εικόνα 13.29: Πίνακας Ταξινόμησης ως προς X και Y.

Το αποτέλεσμα του τεστ δίνεται στην Εικόνα 13.30, από όπου έχουμε ότι η p-τιμή του ελέγχου με χρήση της ακριβούς κατανομής (διωνυμικής) της στατιστικής συνάρτησης είναι μικρότερη από 0.001. Άρα, σε ε.σ. 1%, απορρίπτουμε την H_0, δηλαδή σε ε.σ. 1% μπορούμε να ισχυριστούμε ότι η τηλεμαχία είχε στατιστικά σημαντική επίδραση στην πρόθεση ψήφου του εκλογικού σώματος.

Chi-Square Tests

	Value	Exact Sig. (2-sided)
McNemar Test		<,001[a]
N of Valid Cases	785	

a. Binomial distribution used.

Εικόνα 13.30: Αποτελέσματα ελέγχου McNemar με χρήση της διαδικασίας Crosstabs του SPSS.

Αξίζει να αναφέρουμε ότι στο SPSS υπάρχει και ένας εναλλακτικός τρόπος για να διεξαχθεί το τεστ του McNemar. Η διαφοροποίηση αρχικά έγκειται στο ότι κατά τη διαδικασία εισαγωγής των δεδομένων δίνονται αριθμητικές τιμές, δηλαδή αντί για Α και Β, δίνονται π.χ. οι τιμές 1 (αντί Α) και 2 (αντί Β). Στη συνέχεια, μέσω της καρτέλας Variable View, μέσω του πλαισίου Measure καθορίζουμε τη φύση των μεταβλητών, δίνοντας την τιμή Nominal (ονομαστική μεταβλητή) για τις στήλες COL και ROW. Τέλος, στο Values, δίνουμε όπου 1 την ετικέτα (Label) «Α» και όπου 2 την ετικέτα «Β». Στη συνέχεια, προχωράμε με τη στάθμιση των συχνοτήτων με χρήση της Weight Cases, όπως είδαμε προηγούμενα.

Έπειτα, από το κεντρικό παράθυρο διαλόγου επιλέγουμε

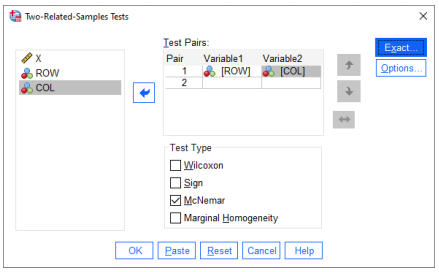
Analyze / Nonparametric Tests / Legacy Dialogs / 2 Related Samples

και στο νέο παράθυρο διαλόγου που προκύπτει (βλ. Εικόνα 13.31) εισάγουμε στο πλαίσιο Variable 1 τη μεταβλητή ROW και στο πλαίσιο Variable 2 τη μεταβλητή COL, ενώ στο πλαίσιο Test Type επιλέγουμε McNemar και πατάμε OK.

Το αποτέλεσμα της ανάλυσης δίνεται στην Εικόνα 13.32. Ο έλεγχος γίνεται με χρήση της στατιστικής συνάρτησης (βλ. Κεφάλαιο 5) $\chi^2 = (|\beta - \gamma| - 1)^2/(\beta + \gamma)$, όπου με βάση τα δεδομένα που δίνονται είναι $\gamma = 6$ και $\beta = 72$. Άρα,

$$\chi^2 = \frac{(|6 - 72| - 1)^2}{6 + 72} \approx 54.1667$$

και η p-τιμή είναι $1 - F_{\chi_1^2}(54.1667)$, με $F_{\chi_1^2}(\cdot)$ να είναι η α.σ.κ. της κατανομής χ_1^2. Προκύπτει, λοιπόν, ότι η p-τιμή είναι πολύ μικρότερη από το 0.001 και άρα, σε ε.σ. 1% απορρίπτουμε την H_0. Δηλαδή, σε ε.σ. 1%



Εικόνα 13.31: Παράθυρο διαλόγου Nonparametric Tests / Legacy Dialogs / 2 Related Samples του SPSS για τον έλεγχο McNemar.

Test Statistics[a]

	ROW & COL
N	785
Chi-Square[b]	54,167
Asymp. Sig.	<,001

a. McNemar Test

b. Continuity Corrected

Εικόνα 13.32: Αποτέλεσμα ελέγχου McNemar - Ασυμπτωτικός έλεγχος.

μπορούμε να ισχυριστούμε ότι η τηλεμαχία είχε σημαντική επίδραση στην πρόθεση ψήφου του εκλογικού σώματος.

(**με χρήση R**) Αρχικά, εισάγουμε τα δεδομένα σε έναν πίνακα, έστω αυτός `x`, όμοιο με αυτόν που δίνεται στην εκφώνηση της άσκησης. Αυτό επιτυγχάνεται με χρήση της εντολής `matrix`, ενώ η κύρια ανάλυση διεξάγεται μέσω της εντολής `mcnemar.test(...)`, όπως φαίνεται παρακάτω:

```
> x<-matrix(c(396,6,72,311),nrow=2)
> mcnemar.test(x,correct=TRUE)
> x
```

Όπως είδαμε, η εντολή `mcnemar.test(...)` υλοποιεί τον έλεγχο McNemar. Στην περίπτωση που έχουμε τον Πίνακα Ταξινόμησης (ή Πίνακα Συνάφειας 2×2) των αποτελεσμάτων πριν και μετά, αρκεί να εισάγουμε ως όρισμα μόνο τον πίνακα `x`. Επίσης, με χρήση του ορίσματος `correct = TRUE` εφαρμόζεται η ασυμπτωτική μορφή του ελέγχου του McNemar με χρήση της διόρθωσης συνεχείας.

```
     [,1] [,2]
[1,]  396   72
[2,]    6  311
McNemar's Chi-squared test with continuity correction
data:  x
McNemar's chi-squared = 54.167, df = 1, p-value = 1.842e-13
```

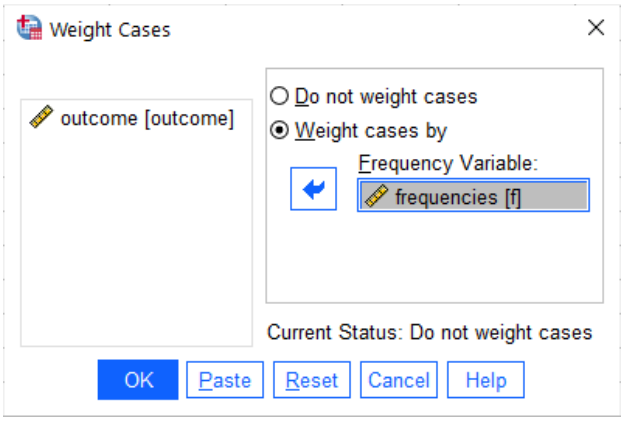
Από τα αποτελέσματα της ανάλυσης διαπιστώνουμε πως η p-τιμή είναι ίση με $1.842 \cdot 10^{-13}$, ενώ η τιμή της στατιστικής συνάρτησης ελέγχου είναι ίση με την αντίστοιχη τιμή που έδωσε το SPSS. Άρα απορρίπτεται η H_0 σε όλα τα συνήθη ε.σ. (και προφανώς και στο 1%) και, επομένως, σε ε.σ. 1% προκύπτει ότι η εμφάνιση των δύο υποψήφιων στην τηλεμαχία είχε στατιστικά σημαντική επίδραση στην πρόθεση ψήφου.

□

Παράδειγμα 13.10. (**Διωνυμικός Έλεγχος-Binomial test**): Έστω ότι ρίχνουμε ένα νόμισμα 40 φορές και καταγράφουμε τα αποτελέσματα κάθε ρίψης («Κ» ή «Γ»). Σε αυτές τις 40 ρίψεις, 24 φορές εμφανίστηκε η όψη «Κ» και 16 φορές εμφανίστηκε η όψη «Γ». Να ελέγξετε σε ε.σ. 5% την υπόθεση ότι το νόμισμα είναι δίκαιο. Να υποθέσετε ότι όταν ρίχνουμε ένα δίκαιο νόμισμα, η πιθανότητα εμφάνισης «Κ» είναι ίση με 1/2.

Λύση Παραδείγματος 13.10. Έχουμε $n = 40$ το πλήθος ανεξάρτητες δοκιμές Bernoulli, τις ρίψεις του νομίσματος, με επιτυχία να θεωρείται η εμφάνιση Κ. Έστω p η πιθανότητα εμφάνισης Κ, η οποία παραμένει σταθερή σε κάθε ρίψη, ενώ το αποτέλεσμα κάθε ρίψης είναι ανεξάρτητο από κάθε άλλο. Θέλουμε να ελέγξουμε σε επίπεδο σημαντικότητας τη μηδενική υπόθεση $H_0 : p = 0.5$ έναντι της εναλλακτικής $H_1 : p \neq 0.5$. Για τον σκοπό αυτό, θα υλοποιηθεί με χρήση του SPSS και της `R` ο διωνυμικός έλεγχος που αναλυτικά παρουσιάστηκε στο Κεφάλαιο 5 του παρόντος συγγράμματος (βλ. Ενότητα 5.2).

(**με χρήση SPSS**): Εισάγουμε αρχικά τα δεδομένα στο SPSS σύμφωνα με τη διαδικασία που περιγράφεται παρακάτω. Σε μια στήλη εισάγουμε τις τιμές 24 και 16 και ονομάζουμε τη στήλη αυτή frequencies. Στη συνέχεια, στη διπλανή στήλη, εισάγουμε τις τιμές 1 (για «Κορώνα» ή «Κ») και 0 (για «Γράμματα» ή «Γ») και ονομάζουμε αυτήν τη στήλη outcome. Τέλος, χρησιμοποιούμε το Weight Cases για να «σταθμίσουμε» ως προς τη στήλη των συχνοτήτων εμφάνισης των δυνατών αποτελεσμάτων, όπως φαίνεται στην Εικόνα 13.33.



Εικόνα 13.33: Διαδικασία Weight Cases.

Στη συνέχεια, διεξάγεται η κύρια ανάλυση επιλέγοντας από το κεντρικό παράθυρο διαλόγου:

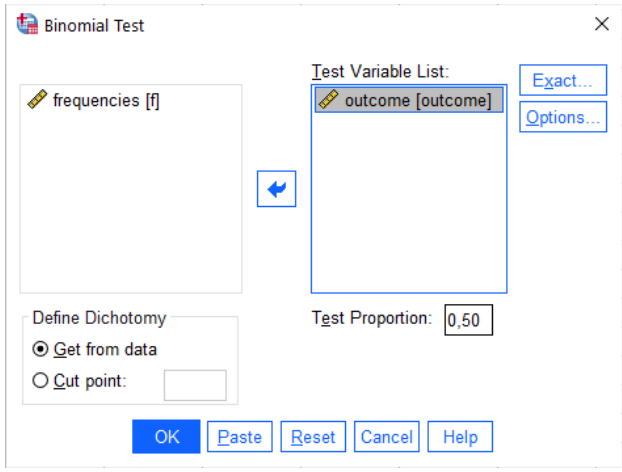
Analyze / Nonparametric Tests / Binomial

Στο νέο παράθυρο διαλόγου που προκύπτει (βλ. Εικόνα 13.34) εισάγουμε τη στήλη των αποτελεσμάτων (δηλαδή τη μεταβλητή outcome) στο πεδίο Test Variables και πατάμε OK.

Τα αποτελέσματα του ελέγχου δίνονται στην Εικόνα 13.35. Η p-τιμή είναι ίση με $0.268 > 0.05$, οπότε σε ε.σ. $a = 0.05$, δεν απορρίπτουμε την H_0. Επομένως, σε ε.σ. 5%, μπορούμε να θεωρήσουμε ότι το νόμισμα είναι δίκαιο.

(με χρήση R) Καθώς από τα δεδομένα της άσκησης έχουμε ότι στο σύνολο $n = 40$ προσπαθειών, με δύο δυνατά αποτελέσματα, το σύνολο των επιτυχιών είναι 24 αρκεί να χρησιμοποιήσουμε την εντολή `binom.test(...)` με τον τρόπο που ακολουθεί:

```
> binom.test(24,40,p=0.5,alternative='two.sided')
```



Εικόνα 13.34: Το παράθυρο διαλόγου Nonparametric Tests / Binomial Test.

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
outcome	Group 1	Head	24	,60	,50	,268
	Group 2	Tail	16	,40		
	Total		40	1,00		

Εικόνα 13.35: Αποτελέσματα εφαρμογής του Binomial Test.

Σημειώνουμε ότι, καθώς η εντολή `binom.test(...)` έχει επεξηγηθεί στο Παράδειγμα 13.7, δεν δίνονται εκ νέου λεπτομέρειες σχετικά με τη χρήση της. Παρακάτω δίνεται το output της ανάλυσης.

```
Exact binomial test
data:  24 and 40
number of successes = 24, number of trials = 40, p-value = 0.2682
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4332671 0.7513500
sample estimates:
probability of success
                   0.6
```

Από τα αποτελέσματα της ανάλυσης διαπιστώνουμε ότι η p-τιμή είναι ίση με $0.2682 > 0.05$ (ακριβώς ίδια με αυτήν που δίνει το SPSS). Συνεπώς, συμπεραίνουμε ότι δεν απορρίπτεται η H_0 σε ε.σ. 5%, δηλαδή σε ε.σ. 5% δεν μπορούμε να απορρίψουμε την υπόθεση ότι το νόμισμα είναι δίκαιο. □

Παράδειγμα 13.11. (**Έλεγχος διαμέσου ενός πληθυσμού - Sign test**): Χρησιμοποιώντας τα δεδομένα που ακολουθούν (Πίνακας 13.11), να ελέγξετε την υπόθεση ότι η διάμεσος της κατανομής από την οποία προέρχονται τα δεδομένα είναι ίση με 1.6, δηλαδή να γίνει ο έλεγχος της $H_0 : x_{0.5} = 1.6$ έναντι της $H_1 : x_{0.5} \neq 1.6$, όπου $x_{0.5}$ είναι η διάμεσος της κατανομής. Ο έλεγχος να γίνει με χρήση της p-τιμής σε ε.σ. 5%.

Λύση Παραδείγματος 13.11. Στο παράδειγμα αυτό θα υλοποιηθεί ο έλεγχος διαμέσου ενός πληθυσμού που αναλυτικά παρουσιάστηκε στο Κεφάλαιο 5 με χρήση του προσημικού κριτηρίου (Sign Test, βλ. Ενότητα

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	0.1	0.2	0.2	0.25	0.25	0.6	0.75	0.8	0.85	1.2
i	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	1.1	1.5	1.7	1.9	2.2	2.3	2.4	2.5	2.9	3.2
i	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	5.1	7.2	7.6	7.9	9.4	9.9	10.5	13.4	17.6	22.1

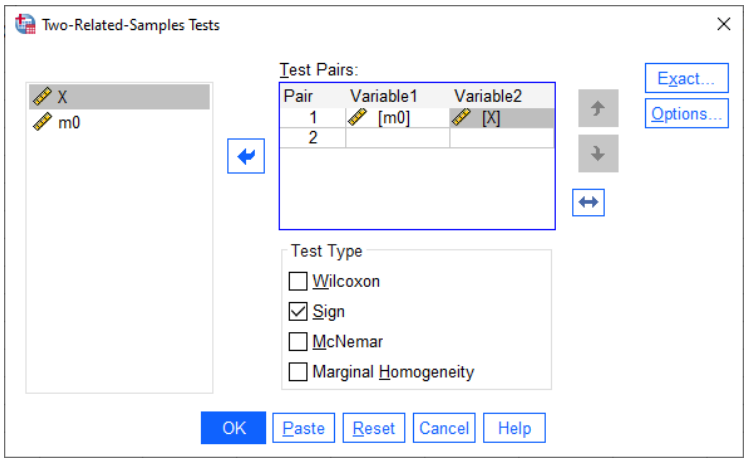
Πίνακας 13.11: Δεδομένα για έλεγχο της διαμέσου ενός πληθυσμού.

5.4). Θεωρούμε ότι έχουμε $n = 30$ το πλήθος ανεξάρτητες δοκιμές Bernoulli (τις μετρήσεις της τυχαίας μεταβλητής). Η μέθοδος του προσημικού ελέγχου στηρίζεται στο πρόσημο της διαφοράς των δειγματικών τιμών $X_1, \dots, X_n$ από την προς έλεγχο τιμή m_0 (εδώ $m_0 = 1.6$). Επομένως, αρχικά δημιουργούμε τις διαφορές $X_1 - m_0, \dots, X_n - m_0$ και θέτουμε «+», όταν είναι θετικές, δηλαδή όταν $X_i - m_0 > 0$, ενώ, όταν είναι αρνητικές, δηλαδή όταν $X_i - m_0 < 0$, θέτουμε «−». Η σ.σ.ε. T είναι το πλήθος των θετικών διαφορών $X_i - m_0$, δηλαδή ο αριθμός των προσήμων «+». Ο έλεγχος που θέλουμε να κάνουμε έχει τη μορφή $H_0 : m = 1.6$ έναντι της εναλλακτικής $H_1 : m \neq 1.6$.

(**με χρήση SPSS**) Αρχικά, σε ένα φύλλο εργασίας στο SPSS, εισάγουμε τα δεδομένα του Πίνακα 13.11 σε μια στήλη, την οποία και ονομάζουμε X. Σε διπλανή στήλη εισάγουμε την τιμή 1.6 σε πλήθος γραμμών ίσο με το πλήθος των τιμών στη στήλη X. Μετονομάζουμε τη στήλη σε $m0$. Στη συνέχεια, η κύρια ανάλυση διεξάγεται επιλέγοντας από το κεντρικό παράθυρο διαλόγου:

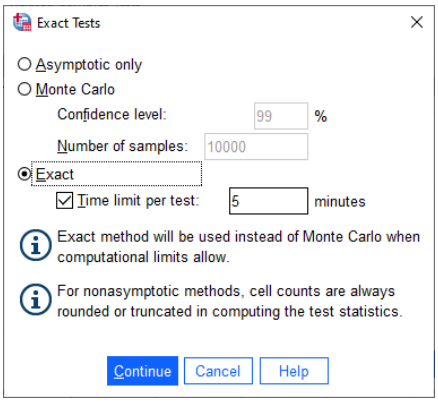
Analyze / Nonparametric Tests / Legacy Dialogs/ 2 Related Samples

Στο νέο παράθυρο διαλόγου (βλ. Εικόνα 13.36) εισάγουμε τη στήλη $m0$ στο πεδίο Test Pairs: Variable 1, τη στήλη X στο πεδίο Test Pairs: Variable 2, ενώ στο Test Type επιλέγουμε Sign. Επιλέγουμε Exact (βλ. Εικόνα 13.37), όπου μας δίνεται η δυνατότητα είτε να εφαρμόσουμε τον έλεγχο με χρήση της ακριβούς κατανομής (διωνυμική) επιλέγοντας Exact είτε να εφαρμόσουμε ασυμπτωτικό έλεγχο (με χρήση z-test) επιλέγοντας Asymptotic Only. Επισημαίνεται ότι εφαρμόζουμε τον ασυμπτωτικό έλεγχο αν το μέγεθος δείγματος είναι αρκετά μεγάλο. Στο συγκεκριμένο παράδειγμα, για εκπαιδευτικούς λόγους, υλοποιούμε τον έλεγχο με χρήση της ακριβούς κατανομής (διωνυμικής) επιλέγοντας Exact.



Εικόνα 13.36: Το παράθυρο διαλόγου Two-Related-Samples-Tests - Προσημικός Έλεγχος.

Από τον πίνακα στην Εικόνα 13.38 έχουμε ότι η τιμή της σ.σ.ε. T είναι ίση με 18 (αφού το πλήθος των θετικών διαφορών $X_i - 1.6$ δίνεται στη γραμμή Positive Differences). Τα αποτελέσματα του ελέγχου δίνονται στην Εικόνα 13.39. Παρατηρήστε ότι στη γραμμή Exact Sig. (2-tailed) δίνεται η ζητούμενη p-τιμή, η



Εικόνα 13.37: Το παράθυρο διαλόγου Exact στο Two-Related-Samples-Tests.

οποία δεν είναι παρά η πιθανότητα $2P(T \geqslant 18) \approx 0.362$, όταν η $T \sim B(30, 0.5)$. Άρα, σε ε.σ. $a = 0.05$, δεν απορρίπτουμε την H_0 και άρα δεν μπορούμε να απορρίψουμε την υπόθεση ότι η διάμεσος της κατανομής του πληθυσμού είναι ίση με 1.6.

Frequencies

		N
X - m0	Negative Differences[a]	12
	Positive Differences[b]	18
	Ties[c]	0
	Total	30

a. X < m0
b. X > m0
c. X = m0

Εικόνα 13.38: Υπολογισμός τιμής σ.σ.ε. T του Προσημικού κριτηρίου.

Test Statistics[a]

	X - m0
Z	-,913
Asymp. Sig. (2-tailed)	,361
Exact Sig. (2-tailed)	,362
Exact Sig. (1-tailed)	,181
Point Probability	,081

a. Sign Test

Εικόνα 13.39: Αποτελέσματα εφαρμογής Προσημικού κριτηρίου.

(**με χρήση R**) Αρχικά, εισάγουμε σε ένα διάνυσμα, έστω αυτό `x`, τα διαθέσιμα δεδομένα. Στη συνέχεια, χρησιμοποιούμε την εντολή `binom.test(...)`, όπως παρακάτω:

```
> x<-c(0.1,0.2,0.2,0.25,0.25,0.6,0.75,0.8,0.85,1.2,1.1,
+1.5,1.7,1.9,2.2,2.3,2.4,2.5,2.9,3.2,5.1,7.2,7.6,7.9,
+9.4,9.9,10.5,13.4,17.6,22.1)
> tstar<-sum(x-1.6>0) # value of T
> binom.test(tstar,length(x))
```

Η εντολή `binom.test(…)` έχει επεξηγηθεί στο Παράδειγμα 13.7 και για αυτόν τον λόγο δεν δίνουμε εκ νέου λεπτομέρειες σχετικά με τη χρήση της. Παρακάτω δίνεται το output της ανάλυσης.

```
        Exact binomial test

data:  tstar and length(x)
number of successes = 18, number of trials = 30, p-value = 0.3616
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4060349 0.7734424
sample estimates:
probability of success
                   0.6
```

Από τα αποτελέσματα της ανάλυσης διαπιστώνουμε ότι η p-τιμή είναι ίση με $0.3616 > 0.05$ (ακριβώς ίδια με αυτήν που δίνει το SPSS) και άρα συμπεραίνουμε ότι δεν απορρίπτεται η H_0 σε ε.σ. 5%. Άρα, σε ε.σ. 5%, δεν μπορούμε να απορρίψουμε την υπόθεση ότι η διάμεσος της κατανομής του πληθυσμού είναι ίση με 1.6. □

Παράδειγμα 13.12. (**Διωνυμικός Έλεγχος για ποσοστιαία σημεία-Quantile test**): Χρησιμοποιώντας τα δεδομένα που ακολουθούν, να ελέγξετε την υπόθεση ότι το 3ο τεταρτημόριο της κατανομής από την οποία προέρχονται τα δεδομένα είναι μεγαλύτερο από το 6, δηλαδή να γίνει ο έλεγχος της $H_0 : x_{0.25} \leqslant 6$ έναντι της $H_1 : x_{0.25} > 6$, όπου $x_{0.25}$ είναι το 3ο τεταρτημόριο της κατανομής. Ο έλεγχος να γίνει με χρήση της p-τιμής, την οποία και να ερμηνεύσετε. Να υποθέσετε ότι η κατανομή του πληθυσμού, από τον οποίο προέρχεται το παρακάτω δείγμα, είναι συνεχής.

$$8.4,\ 7.6,\ 4.9,\ 8.8,\ 6.3,\ 3.3,\ 3.1,\ 4.1,\ 7.7,\ 1.8,\ 4.3,\ 6.5,\ 9.8$$

$$5.7,\ 6.5,\ 5.7,\ 7.1,\ 6.7,\ 1.7,\ 5.6,\ 3.6,\ 5.5,\ 4.2,\ 16.3,\ 3.2$$

Λύση Παραδείγματος 13.12. Στο παράδειγμα αυτό θα υλοποιηθεί ο έλεγχος ποσοστιαίου σημείου που αναλυτικά παρουσιάστηκε στο Κεφάλαιο 5 και διεξάγεται ως διωνυμικός έλεγχος (βλ. Ενότητα 5.3). Ειδικότερα, θεωρούμε ότι έχουμε $n = 25$ το πλήθος ανεξάρτητες δοκιμές Bernoulli (τις μετρήσεις της τυχαίας μεταβλητής) με επιτυχία να θεωρείται η εμφάνιση τιμής μεγαλύτερης από 6. Έστω p η πιθανότητα εμφάνισης τιμής μεγαλύτερης από 6, η οποία παραμένει σταθερή σε κάθε δοκιμή του τυχαίου πειράματος, ενώ το αποτέλεσμα κάθε δοκιμής είναι ανεξάρτητο από κάθε άλλο. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση $H_0 : p \leqslant 0.25$ έναντι της $H_1 : p > 0.25$, δηλαδή θέλουμε να ελέγξουμε την υπόθεση ότι η πιθανότητα επιτυχίας (δηλαδή η $p = P(X > 6)$) είναι ίση ή μικρότερη από 0.25. Επομένως, ουσιαστικά θα ελέγξουμε αν το 3ο τεταρτημόριο του πληθυσμού είναι μεγαλύτερο του 6. Αξίζει να αναφέρουμε πως με βάση τα όσα έχουμε αναφέρει στην Ενότητα 5.3, θέλουμε να διεξάγουμε τον έλεγχο

$$H_0 : x_{0.25} \leqslant 6,\ \ H_1 : x_{0.25} > 6$$

ή, ισοδύναμα,

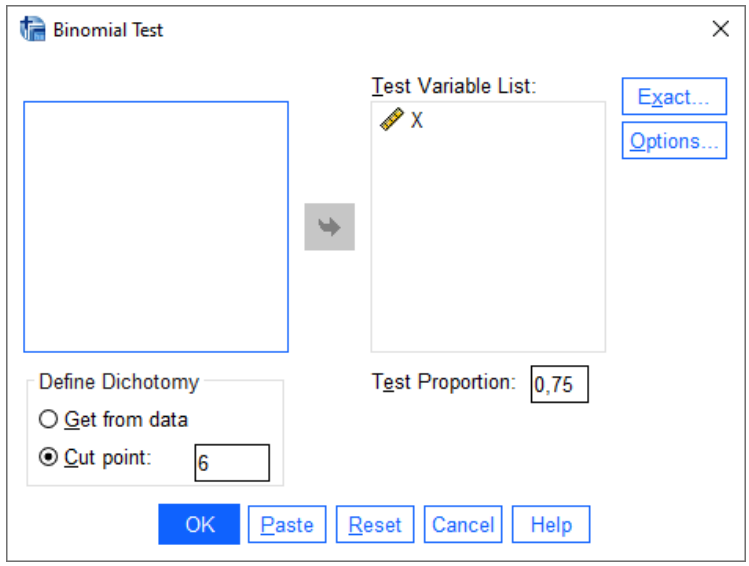
$$H_0 : p \leqslant 0.25, \;\; H_1 : p > 0.25$$

Υπενθυμίζεται, επίσης, ότι η στατιστική συνάρτηση του ελέγχου είναι η $T = \#\{X_i > 6\}$, δηλαδή το πλήθος των τιμών του δείγματος που είναι μεγαλύτερες της τιμής 6.

(**με χρήση SPSS**) Αρχικά, σε ένα φύλλο εργασίας στο SPSS, εισάγουμε τα δεδομένα σε μια στήλη, την οποία και ονομάζουμε X. Στη συνέχεια, η κύρια ανάλυση διεξάγεται επιλέγοντας από το κεντρικό παράθυρο διαλόγου:

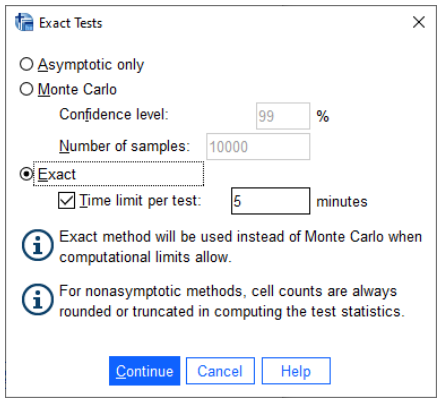
Analyze / Nonparametric Tests / Legacy Dialogs/ Binomial

Στο νέο παράθυρο διαλόγου (βλ. Εικόνα 13.40) εισάγουμε τη στήλη X στο πεδίο Test Variables, στο πεδίο Define Dichotomy επιλέγουμε Cut Point και δίνουμε την τιμή 6, ενώ στο Test Proportion δίνουμε την τιμή 0.75. Οι παραπάνω επιλογές είναι πλήρως αιτιολογημένες καθώς, αν πράγματι το 6 είναι το 3ο τεταρτημόριο της κατανομής του πληθυσμού από τον οποίο λάβαμε το δείγμα, τότε το 75% των τιμών αυτού θα είναι μικρότερο ή ίσο από το 6 και το υπόλοιπο 25% των τιμών θα είναι μεγαλύτερο από το 6. Επίσης, αν το 3ο τεταρτημόριο είναι στην πραγματικότητα μεγαλύτερο του 6, τότε το ποσοστό των τιμών του πληθυσμού, οι οποίες είναι μεγαλύτερες του 6, θα είναι μεγαλύτερο από 25%. Θα πρέπει, όμως, να σημειώσουμε πως στο SPSS πρέπει να δοθεί η πιθανότητα που αφορά το ενδεχόμενο $\{X \leqslant 0.75\}$. Τέλος, πατώντας το πλαίσιο Exact (βλ. Εικόνα 13.41) μας δίνεται η δυνατότητα είτε να εφαρμόσουμε τον έλεγχο με χρήση της ακριβούς κατανομής (διωνυμική) επιλέγοντας Exact είτε να εφαρμόσουμε ασυμπτωτικό έλεγχο (με χρήση z-test) επιλέγοντας Asymptotic Only. Επισημαίνεται ότι εφαρμόζουμε τον ασυμπτωτικό έλεγχο αν το μέγεθος δείγματος είναι αρκετά μεγάλο. Στο συγκεκριμένο παράδειγμα, κυρίως για λόγους επίδειξης, υλοποιούμε τον έλεγχο με χρήση της ακριβούς κατανομής (διωνυμικής) επιλέγοντας Exact.



Εικόνα 13.40: Το παράθυρο διαλόγου Binomial Test - Έλεγχος Quantile Test.

Τα αποτελέσματα του ελέγχου δίνονται στην Εικόνα 13.42. Παρατηρήστε ότι στη στήλη Exact Sig. (1-tailed) δίνεται η ζητούμενη p-τιμή, η οποία δεν είναι παρά η πιθανότητα $P(T \geqslant 11) \approx 0.030$, όταν η $T \sim B(25, 0.25)$. Σημειώστε ότι με βάση τα όσα έχουμε αναφέρει στην Ενότητα 5.3, η τιμή της στατιστικής συνάρτησης είναι 11 (δίνεται στη γραμμή Group 2), ενώ η p-τιμή είναι $P(T \geqslant 11 | T \sim B(25, 0.25))$. Άρα, σε ε.σ. $a = 0.05$, απορρίπτουμε την H_0 και μπορούμε να υποθέσουμε ότι το 3ο τεταρτημόριο της κατανομής των δεδομένων είναι μεγαλύτερο του 6. Επίσης, μπορούμε να πούμε ότι το ελάχιστο επίπεδο σημαντικότητας για το οποίο απορρίπτεται η H_0 είναι ίσο με 3%.



Εικόνα 13.41: Το παράθυρο διαλόγου Exact στο Binomial Test.

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)	Point Probability
X	Group 1	<= 6	14	,56	,75	,030[a]	,019
	Group 2	> 6	11	,44			
	Total		25	1,00			

a. Alternative hypothesis states that the proportion of cases in the first group < ,75.

Εικόνα 13.42: Αποτελέσματα εφαρμογής Binomial Test - Έλεγχος Quantile Test.

(**με χρήση R**) Αρκεί να δώσουμε τις παρακάτω εντολές:

```
1  > x<-c(8.4,7.6,4.9,8.8,6.3,3.3,3.1,4.1,7.7,1.8,4.3,6.5,9.8,
2  + 5.7,6.5,5.7,7.1,6.7,1.7,5.6,3.6,5.5,4.2,16.3,3.2)
3  > theta0<-6
4  > Tau<-length(which(x>theta0)) # number of Xi>6
5  > pstar<-0.25
6  > n<-length(x) # sample size n
7  > # apply the quantile test as a binomial test
8  > binom.test(Tau,n,p=pstar,alternative='greater')
```

Η εντολή `binom.test(...)` έχει επεξηγηθεί στο Παράδειγμα 13.7 και για αυτόν τον λόγο δεν δίνουμε εκ νέου λεπτομέρειες σχετικά με τη χρήση της. Στη γραμμή 4 των παραπάνω εντολών υπολογίζεται η τιμή της στατιστικής συνάρτησης T. Παρακάτω δίνεται το output της ανάλυσης.

```
        Exact binomial test

data:  11 and 25
number of successes = 11, number of trials = 25, p-value = 0.02967
alternative hypothesis: true probability of success is greater than 0.25
95 percent confidence interval:
 0.2698531 1.0000000
sample estimates:
probability of success
                  0.44
```

Από τα αποτελέσματα της ανάλυσης διαπιστώνουμε πως η p-τιμή είναι ίση με $0.02967 < 0.05$ (ακριβώς ίδια με αυτήν που δίνει το SPSS με τη διαφορά να είναι λόγω στρογγυλοποίησης στο τελικό αποτέλεσμα) και άρα

συμπεραίνουμε ότι απορρίπτεται η H_0 σε ε.σ. 5%. Άρα, μπορούμε να θεωρήσουμε ότι το 3ο τεταρτημόριο της κατανομής του πληθυσμού είναι μεγαλύτερο του 6.

□

13.4 Έλεγχοι τάξης

Σκοπός αυτής της ενότητας είναι η υλοποίηση, μέσω παραδειγμάτων και με χρήση του SPPS και της R, των μεθοδολογιών για τον έλεγχο της ισότητας δύο ή περισσότερων πληθυσμιακών διαμέσων (με ανεξάρτητα ή εξαρτημένα δείγματα), που βασίζονται στην τάξη των δεδομένων και όχι απλώς στις τιμές τους. Οι έλεγχοι αυτοί αποτέλεσαν αντικείμενο μελέτης στο Κεφάλαιο 6 και θα ταξινομηθούν σε εκείνους που αφορούν δύο το πλήθος πληθυσμούς και σε εκείνους που αφορούν περισσότερους από δύο το πλήθος πληθυσμούς.

13.4.1 Δύο πληθυσμοί

Στην ενότητα αυτή θα υλοποιηθούν, μέσω παραδειγμάτων και με χρήση του SPPS και της R, οι έλεγχοι που εκτενώς μελετήθηκαν στην Ενότητα 6.3 και αφορούν τον έλεγχο διαμέσων δύο πληθυσμών.

Παράδειγμα 13.13. (**Έλεγχος Wilcoxon για 2 εξαρτημένα δείγματα**): Χρησιμοποιώντας τα δεδομένα του Πίνακα 13.7 θέλουμε να ελέγξουμε σε ε.σ. $a = 5\%$ αν υπάρχει διαφορά στη διάμεση αρτηριακή πίεση πριν και μετά τη λήψη χαπιού. Υπόδειξη: να γίνει ο έλεγχος χρησιμοποιώντας τον έλεγχο του Wilcoxon για 2 εξαρτημένα δείγματα.

Λύση Παραδείγματος 13.13. Έχουμε ένα τυχαίο δείγμα $X_1, \ldots, X_n$, μεγέθους $n = 15$ από τον πληθυσμό, ο οποίος περιγράφει την αρτηριακή πίεση πριν τη λήψη του χαπιού. Επιπλέον, έχουμε ένα τυχαίο δείγμα $Y_1, \ldots, Y_n$, μεγέθους $n = 15$, από τον πληθυσμό ο οποίος περιγράφει την αρτηριακή πίεση μετά τη λήψη του χαπιού. Καθώς τα δύο δείγματα είναι εξαρτημένα, θεωρώντας το τυχαίο δείγμα των διαφορών $D_i = X_i - Y_i$, $i = 1, \ldots, n$ ο έλεγχος της ισότητας των πληθυσμιακών διαμέσων ανάγεται (βλ. για περισσότερες λεπτομέρειες Ενότητα 6.3.2) στον έλεγχο της $H_0 : m_D = 0$ έναντι της εναλλακτικής $H_1 : m_D \neq 0$, όπου m_D είναι η διάμεσος του πληθυσμού που περιγράφει τη διαφορά $X - Y$. Επομένως, μπορούμε πλέον να εφαρμόσουμε τον έλεγχο Wilcoxon που παρουσιάστηκε στην Ενότητα 6.2.

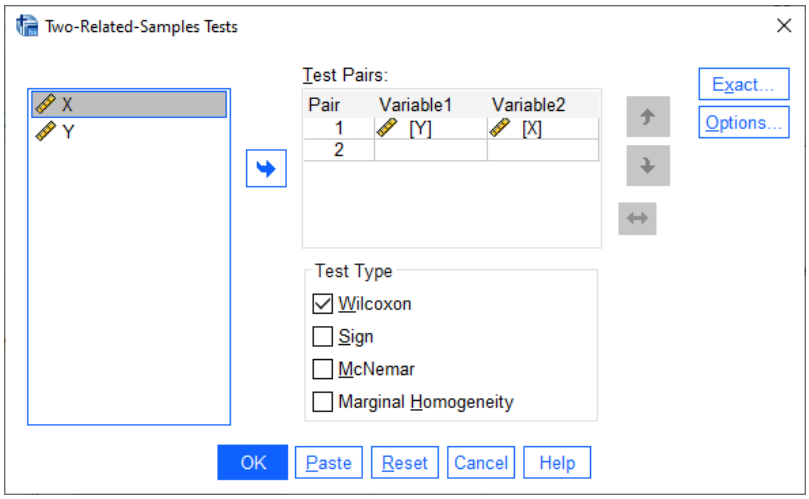
(**με χρήση SPSS**): Αρχικά, εισάγουμε τις τιμές της αρτηριακής πίεσης, πριν και μετά τη λήψη του χαπιού, σε δύο στήλες του SPSS και τις ονομάζουμε X (για τις μετρήσεις πριν τη λήψη) και Y (για τις μετρήσεις μετά τη λήψη), αντίστοιχα. Έπειτα για τη διεξαγωγή της κυρίως ανάλυσης επιλέγουμε από το κεντρικό παράθυρο διαλόγου:

Analyze / Nonparametric Tests / Legacy Dialogs/ 2 Related Samples

Στο νέο παράθυρο διαλόγου που προκύπτει (βλ. Εικόνα 13.43) εισάγουμε τις μεταβλητές X, Y στα πεδία Variable 2 και Variable 1, αντίστοιχα, ενώ στο πλαίσιο Test Type επιλέγουμε Wilcoxon και πατάμε OK. Αξίζει να αναφέρουμε ότι το SPSS θα εκτελέσει τον ασυμπτωτικό έλεγχο.

Στην Εικόνα 13.44 δίνεται το πλήθος των τάξεων (ranks) που βασίζονται σε θετικές και σε αρνητικές διαφορές $Z_i = X_i - Y_i$, όπως και ο μέσος και το άθροισμα αυτών των τάξεων. Ως θετική (αντίστοιχα, αρνητική) διαφορά, ορίζεται η $X_i - Y_i > 0$ (αντίστοιχα, $X_i - Y_i < 0$).

Στην Εικόνα 13.45 δίνονται τα αποτελέσματα του ελέγχου. Η τιμή της σ.σ.ε. T είναι 45 (βλ. στήλη Sum of Ranks, Εικόνα 13.44), η οποία ισούται με το άθροισμα των τάξεων που αντιστοιχούν σε αρνητικές διαφορές. Παρατηρούμε ότι το SPSS εκτελεί ως προεπιλογή τον ασυμπτωτικό έλεγχο Wilcoxon, ωστόσο



Εικόνα 13.43: Το παράθυρο διαλόγου Two-Related Samples Tests για τον έλεγχο Wilcoxon.

Ranks

		N	Mean Rank	Sum of Ranks
X - Y	Negative Ranks	5[a]	9,00	45,00
	Positive Ranks	10[b]	7,50	75,00
	Ties	0[c]		
	Total	15		

a. X < Y
b. X > Y
c. X = Y

Εικόνα 13.44: Πλήθος τάξεων που αντιστοιχούν σε θετικές και αρνητικές διαφορές.

από την επιλογή Exact (βλ. Εικόνα 13.43) μπορούμε να επιλέξουμε να υπολογιστεί η p-τιμή, χρησιμοποιώντας την ακριβή κατανομή του ελέγχου Wilcoxon. Από τα αποτελέσματα της ανάλυσης στην Εικόνα 13.45 έπεται ότι η p-τιμή είναι ίση με $0.394 > 0.05$. Άρα συμπεραίνουμε ότι σε ε.σ. 5% δεν απορρίπτεται η H_0, δηλαδή δεν υπάρχουν σαφείς ενδείξεις ότι η λήψη χαπιού επιφέρει στατιστικά σημαντική διαφοροποίηση στη διάμεση αρτηριακή πίεση.

Test Statistics[a]

	X - Y
Z	-,853[b]
Asymp. Sig. (2-tailed)	,394

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

Εικόνα 13.45: Τιμή στατιστικής συνάρτησης και p-τιμή για το τεστ του Wilcoxon.

(**με χρήση R**): Αρχικά, εισάγουμε τα δεδομένα σε δύο διανύσματα, έστω αυτά `x` και `y`. Στη συνέχεια, χρησιμοποιούμε την εντολή `wilcox.test(...)`, όπως φαίνεται παρακάτω:

```
> x<-c(125,115,130,140,142,117,143,123,139,134,136,118,121,122,133)
> y<-c(110,122,125,120,140,124,123,137,135,145,144,112,119,121,118)
> wilcox.test(y, x,alternative = 'two.sided',mu = 0, paired = TRUE,
+ correct = TRUE,conf.level = 0.95)
```

Ειδικότερα, στην εντολή `wilcox.test(…)` εισάγουμε αρχικά ως όρισμα τα διανύσματα με τις διαθέσιμες μετρήσεις και, στη συνέχεια, δηλώνουμε τη μορφή της εναλλακτικής υπόθεσης. Στο όρισμα `mu` δηλώνεται η προς έλεγχο τιμή για την πληθυσμιακή διάμεσο των διαφορών (εδώ η τιμή 0), ενώ στο όρισμα `alternative = 'two.sided'` επιλέγουμε να κάνουμε τον έλεγχο με δίπλευρη εναλλακτική, δηλαδή τον έλεγχο $\mathrm{H}_0 : m_D = 0$ έναντι της $\mathrm{H}_1 : m_D \neq 0$. Επισημαίνεται ότι, αν θέλουμε να κάνουμε κάποιον από τους μονόπλευρους ελέγχους, δηλώνουμε `alternative = 'less'` ή `alternative = 'greater'`, οπότε οι αντίστοιχες εναλλακτικές είναι $H_1 : m_D < 0$ και $H_1 : m_D > 0$, αντίστοιχα. Καθώς πρόκειται για εξαρτημένα δείγματα δηλώνουμε στο όρισμα `paired` την τιμή `TRUE`. Επίσης, με χρήση του ορίσματος `correct = TRUE` χρησιμοποιείται η διόρθωση συνεχείας κατά την εφαρμογή του ασυμπτωτικού ελέγχου Wilcoxon, ενώ στο όρισμα `conf.level` δηλώνεται ο συντελεστής εμπιστοσύνης (εδώ δηλώθηκε να είναι ίσος με 95%). Το output της παραπάνω ανάλυσης είναι το ακόλουθο:

```
        Wilcoxon signed rank test with continuity correction

data:  y and x
V = 45, p-value = 0.4098
alternative hypothesis: true location shift is not equal to 0
```

Από τα αποτελέσματα της ανάλυσης διαπιστώνουμε ότι η τιμή της σ.σ.ε. είναι ίση με 45, ίση με αυτή που προέκυψε από την ανάλυση με το SPSS. Η συγκεκριμένη τιμή είναι το άθροισμα των τάξεων που αντιστοιχούν σε αρνητικές διαφορές $X < Y$. Επίσης, η p-τιμή είναι ίση με 0.4098, ελαφρώς διαφορετική από την αντίστοιχη που πήραμε από το SPSS, το οποίο οφείλεται στο ότι εδώ επιλέξαμε να εφαρμόσουμε τη διόρθωση συνεχείας. Άρα, αφού η p-τιμή είναι ίση με $0.4098 > 0.05$, συμπεραίνουμε ότι δεν απορρίπτεται σε ε.σ. 5% η H_0. Επομένως, δεν μπορούμε να απορρίψουμε, σε ε.σ. 5%, την υπόθεση ότι η διάμεσος της κατανομής των διαφορών είναι ίση με μηδέν. □

Παράδειγμα 13.14. (**Έλεγχος διαμέσου ενός πληθυσμού - Wilcoxon test**): Χρησιμοποιώντας τα δεδομένα του Πίνακα 13.12 να ελέγξετε την υπόθεση ότι η διάμεσος της κατανομής, από την οποία προέρχονται τα δεδομένα, είναι ίση με 37.5, δηλαδή να γίνει ο έλεγχος της $H_0 : x_{0.5} = 37.5$ έναντι της $H_1 : x_{0.5} \neq 37.5$, όπου $x_{0.5}$ είναι η διάμεσος της κατανομής. Ο έλεγχος να γίνει με χρήση του ελέγχου Wilcoxon για 2 εξαρτημένα δείγματα σε ε.σ. 5%.

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	53.3	38.2	44.6	34.15	27.4	45.3	35.9	48.3	40.5	44.4
i	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	38.2	34.1	39.3	37.2	45.6	47.8	44.9	45.1	39.2	48.2
i	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	35.3	45.1	42.6	50.1	35.7	47.3	44.9	42.1	39.5	36.7

Πίνακας 13.12: Δεδομένα για έλεγχο διαμέσου ενός πληθυσμού.

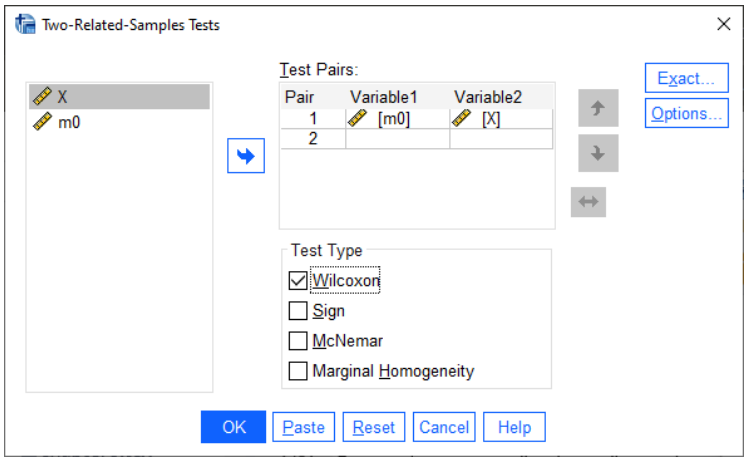
Λύση Παραδείγματος 13.14. Στο παράδειγμα αυτό, θα υλοποιηθεί ο έλεγχος διαμέσου ενός πληθυσμού που αναλυτικά παρουσιάστηκε στο Κεφάλαιο 6 με χρήση του ελέγχου Wilcoxon (βλ. Ενότητα 6.2). Έχουμε ένα τυχαίο δείγμα $X_1, \ldots, X_n$, μεγέθους $n = 30$ από τον πληθυσμό με άγνωστη κατανομή. Αφού θέλουμε να κάνουμε τον έλεγχο της $H_0 : m = 37.5$ έναντι της εναλλακτικής $H_1 : m \neq 37.5$, όπου m είναι η διάμεσος του πληθυσμού, μπορούμε να θεωρήσουμε ένα δεύτερο δείγμα 30 τιμών, οι οποίες είναι όλες ίσες με 37.5 και να εφαρμόσουμε τη διαδικασία που περιγράφθηκε στο Παράδειγμα 13.13. Δηλαδή να θεωρήσουμε το τυχαίο

δείγμα των διαφορών $D_i = X_i - 37.5$, $i = 1, ..., 30$ και να διεξάγουμε τον έλεγχο της $H_0 : m - 37.5 = 0$ έναντι της εναλλακτικής $H_1 : m - 37.5 \neq 0$.

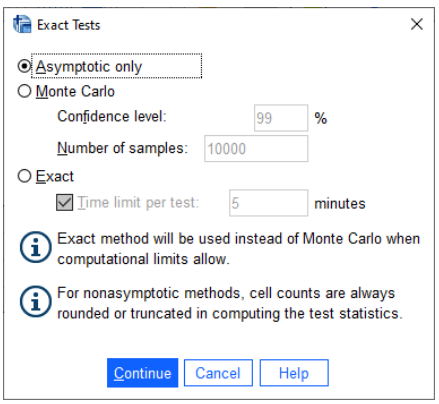
(**με χρήση SPSS**) Αρχικά, σε ένα φύλλο εργασίας στο SPSS, εισάγουμε τα δεδομένα του Πίνακα 13.11 σε μια στήλη, την οποία και ονομάζουμε X. Σε διπλανή στήλη εισάγουμε την τιμή 37.5 σε πλήθος γραμμών ίσο με το πλήθος των τιμών στη στήλη X. Μετονομάζουμε τη στήλη σε $m0$. Στη συνέχεια, η κύρια ανάλυση διεξάγεται επιλέγοντας από το κεντρικό παράθυρο διαλόγου:

Analyze / Nonparametric Tests / Legacy Dialogs/ 2 Related Samples

Στο νέο παράθυρο διαλόγου (βλ. Εικόνα 13.46) εισάγουμε τη στήλη $m0$ στο πεδίο Test Pairs: Variable 1, τη στήλη X στο πεδίο Test Pairs: Variable 2, ενώ στο Test Type επιλέγουμε Wilcoxon. Επιλέγουμε το πλαίσιο Exact (βλ. Εικόνα 13.47), από όπου μας δίνεται η δυνατότητα είτε να εφαρμόσουμε τον έλεγχο με χρήση της ακριβούς κατανομής επιλέγοντας Exact είτε να εφαρμόσουμε ασυμπτωτικό έλεγχο (με χρήση z-test) επιλέγοντας Asymptotic Only. Επισημαίνεται ότι εφαρμόζουμε τον ασυμπτωτικό έλεγχο αν το μέγεθος δείγματος είναι αρκετά μεγάλο. Στο συγκεκριμένο παράδειγμα, για εκπαιδευτικούς λόγους, υλοποιούμε τον ασυμπτωτικό έλεγχο επιλέγοντας Asymptotic only.



Εικόνα 13.46: Το παράθυρο διαλόγου Two-Related-Samples-Tests - Έλεγχος Wilcoxon.



Εικόνα 13.47: Το παράθυρο διαλόγου Exact στο Wilcoxon.

Από τον πίνακα στην Εικόνα 13.48 έχουμε ότι η τιμή της σ.σ.ε. $T = \min\{T^+, T^-\}$ είναι ίση με 77.50 (αντιστοιχεί στην τιμή της T^-, η οποία δίνει το άθροισμα των τάξεων για τις αρνητικές διαφορές και δίνεται στη στήλη Negative Differences). Τα αποτελέσματα του ελέγχου δίνονται στην Εικόνα 13.49. Παρατηρήστε ότι στη γραμμή Asymp. Sig. (2-tailed) δίνεται η ζητούμενη p-τιμή, η οποία δεν είναι τίποτα άλλο παρά η

πιθανότητα $2\Phi(-3.189) \approx 0.001$. Εδώ, χρησιμοποιείται η ασυμπτωτική κατανομή της T υπό την H_0 (η οποία είναι η $\mathcal{N}(0,1)$) με την τιμή της να υπολογίζεται από τη σχέση

$$(T - n(n+1)/4)/\sqrt{n(n+1)(2n+1)/24}.$$

Άρα, σε ε.σ. $a = 0.05$ απορρίπτουμε την H_0, δηλαδή απορρίπτεται ο ισχυρισμός ότι η διάμεσος της κατανομής του πληθυσμού είναι ίση με 37.5.

Ranks

		N	Mean Rank	Sum of Ranks
X - m0	Negative Ranks	8[a]	9,69	77,50
	Positive Ranks	22[b]	17,61	387,50
	Ties	0[c]		
	Total	30		

a. X < m0
b. X > m0
c. X = m0

Εικόνα 13.48: Υπολογισμός τιμής σ.σ.ε. T του ελέγχου Wilcoxon.

Test Statistics[a]

	X - m0
Z	-3,189[b]
Asymp. Sig. (2-tailed)	,001

a. Wilcoxon Signed Ranks Test
b. Based on negative ranks.

Εικόνα 13.49: Αποτελέσματα εφαρμογής ελέγχου Wilcoxon.

(**με χρήση R**) Αρχικά, εισάγουμε σε ένα διάνυσμα, έστω αυτό `x`, τα διαθέσιμα δεδομένα. Επίσης, φτιάχνουμε και ένα διάνυσμα `y` ίσης διάστασης με το διάνυσμα `x`, στο οποίο περιέχεται μόνο η τιμή 37.5. Στη συνέχεια, χρησιμοποιούμε την εντολή `wilcox.test(...)`, όπως παρακάτω:

```
1 > x<-c(53.30,38.20,44.60,34.10,27.40,45.30,35.90,48.30,40.50,
2 + 44.40,38.20,34.10,39.30,37.20,45.60,47.80,44.90,45.10,
3 + 39.20,48.20,35.30,45.10,42.60,50.10,35.70,47.30,44.90,
4 + 42.10,39.50,36.70)
5 > y<-rep(37.5,length(x))
6 > wilcox.test(y,x,paired=T,correct=F)
```

Η εντολή `wilcox.test(…)` έχει επεξηγηθεί στο Παράδειγμα 13.13 και δεν δίνουμε εκ νέου λεπτομέρειες σχετικά με τη χρήση της. Παρακάτω δίνεται το output της ανάλυσης.

```
        Wilcoxon signed rank test

data:  y and x
V = 77.5, p-value = 0.00143
alternative hypothesis: true location shift is not equal to 0
```

Από τα αποτελέσματα της ανάλυσης διαπιστώνουμε ότι η p-τιμή είναι ίση με $0.00143 < 0.05$ (ακριβώς ίδια με αυτή που δίνει το SPSS, αφού δεν έχουμε χρησιμοποιήσει τη διόρθωση συνεχείας) και άρα συμπεραίνουμε ότι απορρίπτεται η H_0 σε ε.σ. 5%. Άρα, σε ε.σ. 5%, απορρίπτεται ο ισχυρισμός ότι η διάμεσος της κατανομής του πληθυσμού είναι ίση με 37.5. □

Παράδειγμα 13.15. (**Έλεγχος Mann-Whitney - Ακριβές Τεστ**): Ένας ψυχολόγος θέλει να ελέγξει αν ο διάμεσος χρόνος επίλυσης ενός προβλήματος από ένα παιδί με μαθησιακές δυσκολίες διαφέρει από τον αντίστοιχο χρόνο επίλυσης του προβλήματος από ένα παιδί χωρίς μαθησιακές δυσκολίες. Για τον λόγο αυτό κατέγραψε στον Πίνακα 13.15 (βλ., επίσης, Sprent, 1999) τον χρόνο (σε δευτερόλεπτα) που χρειάστηκε καθένα από τα 7 παιδιά χωρίς μαθησιακές δυσκολίες (ομάδα 1) και καθένα από τα 8 παιδιά με μαθησιακές δυσκολίες (ομάδα 2). Να διεξάγετε κατάλληλο έλεγχο, σε ε.σ. 2%, και να διατυπώσετε το συμπέρασμα του ελέγχου.

i	1	2	3	4	5	6	7	8
X	204	218	197	183	227	233	191	
Y	243	228	261	202	343	242	220	239

Πίνακας 13.13: Δεδομένα χρόνων επίλυσης προβλήματος από δύο ανεξάρτητους πληθυσμούς.

Λύση Παραδείγματος 13.15. Έστω X και Y οι τ.μ. που περιγράφουν τον χρόνο επίλυσης σε δευτερόλεπτα ενός θέματος από άτομα χωρίς μαθησιακές και με μαθησιακές δυσκολίες, αντίστοιχα. Έχουμε δύο το πλήθος, ανεξάρτητα μεταξύ τους, τυχαία δείγματα από καθέναν από αυτούς τους δύο πληθυσμούς, μεγέθους $n_1 = 7$ και $n_2 = 8$. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση $H_0 : m_X = m_Y$, δηλαδή την υπόθεση της ισότητας των πληθυσμιακών διαμέσων έναντι της εναλλακτικής $H_1 : m_X \neq m_Y$, όπου με m_X και m_Y συμβολίζουμε τη διάμεσο στον 1ο και στον 2ο πληθυσμό, αντίστοιχα. Θα εφαρμόσουμε τον έλεγχο των Mann-Whitney (βλ. Ενότητα 6.3).

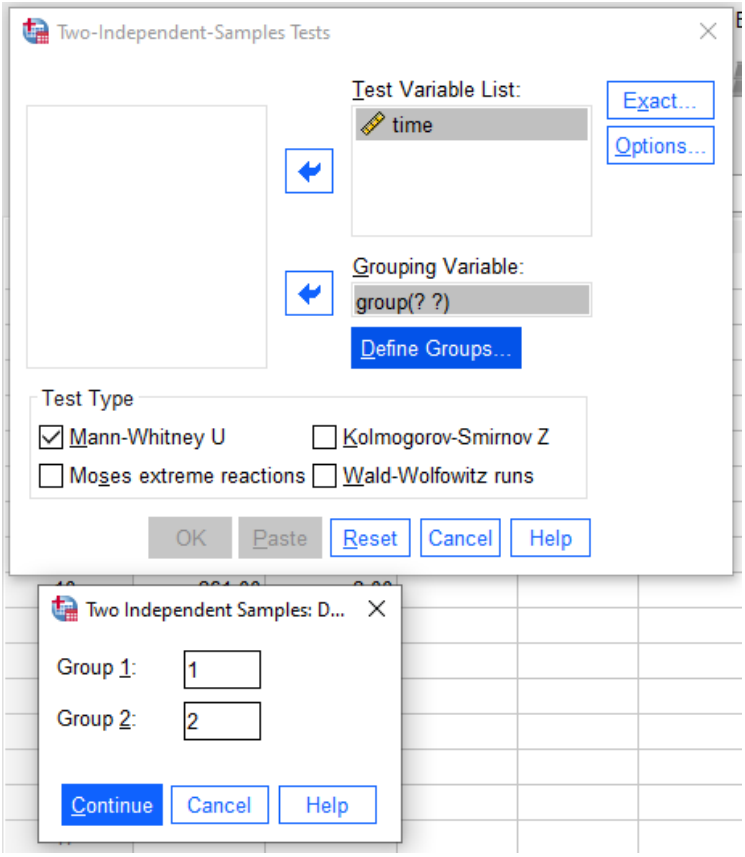
(με χρήση SPSS) Αρχικά, εισάγουμε τα δεδομένα σε ένα φύλλο εργασίας του SPSS, εισάγοντας πρώτα σε μια στήλη τις τιμές X (γραμμές 1-7), ενώ, στη συνέχεια, ακριβώς από κάτω, εισάγουμε τις τιμές Y (γραμμές 8-15). Μετονομάζουμε τη στήλη σε time. Στη διπλανή στήλη, εισάγουμε την τιμή 1 στις γραμμές 1-7 και την τιμή 2 στις γραμμές 8-15, για να δηλώσουμε κατά αυτόν τον τρόπο από ποιον πληθυσμό (ομάδα) έχει προέλθει η κάθε τιμή. Μετονομάζουμε τη στήλη σε group.

Έπειτα διεξάγουμε την κύρια ανάλυση επιλέγοντας από το κεντρικό παράθυρο διαλόγου:

Analyze / Nonparametric Tests / Legacy Dialogs / 2 Independent Samples

Στο νέο παράθυρο διαλόγου που προκύπτει (βλ. Εικόνα 13.50) εισάγουμε τη στήλη time στο πεδίο Test Variable List και τη στήλη group στο πεδίο Grouping Variable. Στη συνέχεια, πατάμε το πλαίσιο Define Groups και στο παράθυρο διαλόγου, που ανοίγει, δίνουμε στο Group 1 την τιμή 1, στο Group 2 την τιμή 2 και πατάμε Continue. Στη συνέχεια πατάμε OK και προκύπτει το output της ανάλυσης.

Από το output της ανάλυσης που δίνεται στην Εικόνα 13.51 έχουμε τη μέση τιμή των τάξεων, καθώς και το άθροισμα αυτών, για τις συνολικά 15 παρατηρήσεις (κοινό δείγμα). Υπενθυμίζεται ότι οι τάξεις υπολογίζονται ως προς το κοινό διατεταγμένο δείγμα (συνολικά 15 παρατηρήσεις). Από τις τιμές στη στήλη MeanRank έχουμε ενδείξεις ότι ο διάμεσος χρόνος που χρειάζεται η 2η ομάδα για την επίλυση ενός προβλήματος είναι μεγαλύτερος από τον χρόνο που χρειάζεται η 1η ομάδα. Όμως, αυτό δεν είναι παρά μια ένδειξη και δεν αποτελεί στατιστική συμπερασματολογία. Θα πρέπει να προχωρήσουμε με τη διενέργεια κατάλληλου ελέγχου



Εικόνα 13.50: Το παράθυρο διαλόγου Two-Independent Samples Tests για τον έλεγχο Mann-Whitney.

υπόθεσης ώστε να διαπιστώσουμε αν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των δύο ομάδων ως προς τον διάμεσο χρόνο επίλυσης ενός προβλήματος.

Στη συνέχεια, από τον πίνακα στην Εικόνα 13.52, έχουμε την τιμή της στατιστικής συνάρτησης (γραμμή Mann-Whitney U), καθώς και την τιμή της στατιστικής συνάρτησης Z για την ασυμπτωτική μορφή του τεστ. Όμως, για την ορθή εφαρμογή του ασυμπτωτικού ελέγχου των Mann-Whitney U, θα πρέπει τα μεγέθη δείγματος από κάθε πληθυσμό να είναι τουλάχιστον 10. Καθώς σε αυτήν την περίπτωση έχουμε $n_1 = 7$ και $n_2 = 8$, θα προχωρήσουμε με χρήση του ακριβούς ελέγχου.

Για να κάνουμε τον έλεγχο Mann-Whitney χρησιμοποιώντας την ακριβή κατανομή του ελέγχου, πρέπει να το επιλέξουμε από το Exact (βλ. Εικόνα 13.50). Σε αυτήν την περίπτωση το output της ανάλυσης δίνεται στην Εικόνα 13.53. Η ακριβής τιμή για την p-τιμή του δίπλευρου ελέγχου είναι $0.014 < 0.02$ (στο Exact Sig. (2-tailed)) και άρα απορρίπτεται η H_0. Από τις δειγματικές διαμέσους έπεται ότι ο διάμεσος χρόνος που απαιτεί η 1η ομάδα εκτιμάται ότι είναι ίσος με 204, ενώ για τη 2η ομάδα, εκτιμάται ότι είναι ίσος με 240.5. Άρα έχουμε ενδείξεις ότι η διάμεσος της κατανομής του χρόνου επίλυσης προβλήματος στη 2η ομάδα είναι μεγαλύτερη έναντι της αντίστοιχης διαμέσου της 1ης ομάδας.

Ολοκληρώνοντας τη λύση του προβλήματος με χρήση SPSS, αξίζει να αναφέρουμε ότι η τιμή 7 στη γραμμή Mann-Whitney U αντιστοιχεί στην τιμή της στατιστικής συνάρτησης που προτάθηκε από τους Mann and Whitney (1947), ενώ στη γραμμή Wilcoxon W δίνεται το άθροισμα των τάξεων για τις παρατηρήσεις από την 1η ομάδα (δηλαδή είναι η τιμή για το group = 1, στη στήλη Sum of Ranks στον Πίνακα 13.51).

(**με χρήση R**) Εισάγουμε τα δεδομένα σε δύο διανύσματα, έστω τα `x` και `y`, και χρησιμοποιούμε την εντολή `wilcox.test(...)`, όπως φαίνεται στις εντολές που παρατίθενται παρακάτω. Ειδικότερα, στην εντολή `wilcox.test(...)` εισάγουμε αρχικά ως όρισμα τα διανύσματα με τις διαθέσιμες μετρήσεις και, στη συνέχεια, δηλώνουμε τον τύπο της εναλλακτικής υπόθεσης. Πιο συγκεκριμένα, στο όρισμα `mu` δηλώνεται η

Ranks

	group	N	Mean Rank	Sum of Ranks
time	1,00	7	5,00	35,00
	2,00	8	10,63	85,00
	Total	15		

Εικόνα 13.51: Άθροισμα τάξεων (Sum of Ranks) και μέσος όρος τάξεων (Mean Rank) στις δύο ομάδες.

Test Statistics[a]

	time
Mann-Whitney U	7,000
Wilcoxon W	35,000
Z	-2,430
Asymp. Sig. (2-tailed)	,015
Exact Sig. [2*(1-tailed Sig.)]	,014[b]

a. Grouping Variable: group

b. Not corrected for ties.

Εικόνα 13.52: Αποτελέσματα ελέγχου Mann-Whitney.

προς έλεγχο τιμή για τη διαφορά των πληθυσμιακών διαμέσων $m_X - m_Y$ (εδώ η τιμή 0), ενώ στο όρισμα `alternative = 'two.sided'` επιλέγουμε να κάνουμε τον έλεγχο με δίπλευρη εναλλακτική, δηλαδή τον έλεγχο $\mathrm{H}_0 : m_X - m_Y = 0$ έναντι της $\mathrm{H}_1 : m_X - m_Y \neq 0$. Επισημαίνεται ότι, αν θέλουμε να κάνουμε κάποιον από τους μονόπλευρους ελέγχους, δηλώνουμε `alternative = 'less'` ή `alternative = 'greater'`, οπότε οι αντίστοιχες εναλλακτικές είναι $H_1 : m_X - m_Y < 0$ και $H_1 : m_X - m_Y > 0$, αντίστοιχα. Καθώς πρόκειται για ανεξάρτητα δείγματα δηλώνουμε στο όρισμα `paired` την τιμή `FALSE`. Επίσης, με χρήση του ορίσματος `exact = TRUE` δηλώνουμε ότι θέλουμε να διεξάγουμε τον έλεγχο με χρήση της ακριβούς (και όχι της ασυμπτωτικής) κατανομής της σ.σ.ε.

```
1  > x<-c(204,218,197,183,227,233,191)
2  > y<-c(243,228,261,202,343,242,220,239)
3  > wilcox.test(x, y,alternative = 'two.sided',mu = 0,
4  + paired = FALSE ,exact=TRUE)
```

Το output της παραπάνω ανάλυσης είναι το ακόλουθο:

```
        Wilcoxon rank sum exact test

data:  x and y
W = 7, p-value = 0.01399
alternative hypothesis: true location shift is not equal to 0
```

Από τα αποτελέσματα της ανάλυσης διαπιστώνουμε ότι η τιμή της σ.σ.ε. είναι ίση με 7. Η τιμή αυτή υπολογίζεται από τη σχέση (βλ. Κεφάλαιο 6) $U = \min\{U_1, U_2\}$, όπου $U_i = R_i - n_i(n_i + 1)/2$, με R_i να είναι

Test Statistics[a]

	time
Mann-Whitney U	7,000
Wilcoxon W	35,000
Z	-2,430
Asymp. Sig. (2-tailed)	,015
Exact Sig. [2*(1-tailed Sig.)]	,014[b]
Exact Sig. (2-tailed)	,014
Exact Sig. (1-tailed)	,007
Point Probability	,002

a. Grouping Variable: group

b. Not corrected for ties.

Εικόνα 13.53: Αποτελέσματα ελέγχου Mann-Whitney με χρήση της επιλογής Exact.

το άθροισμα των τάξεων των παρατηρήσεων του $i-$οστού δείγματος, ως προς το σύνολο των διαθέσιμων μετρήσεων (κοινό δείγμα) και n_1, n_2 να είναι το μέγεθος του 1ου και του 2ου δείγματος, αντίστοιχα. Επιπλέον, έχουμε ότι η p-τιμή είναι ίση με 0.01399. Η τιμή αυτή ταυτίζεται με την p-τιμή που προκύπτει με χρήση του SPSS στη γραμμή Exact Sig. (2-tailed) Άρα, αφού η p-τιμή είναι ίση με $0.01344 < 0.02$, συμπεραίνουμε ότι σε ε.σ. 2% απορρίπτεται η H_0. Επομένως, σε ε.σ. 2%, συμπεραίνουμε ότι ο διάμεσος χρόνος επίλυσης του προβλήματος διαφέρει μεταξύ παιδιών με μαθησιακές και χωρίς μαθησιακές δυσκολίες. Από τις δειγματικές διαμέσους έπεται ότι ο διάμεσος χρόνος που απαιτεί η 1η ομάδα εκτιμάται ότι είναι ίσος με 204, ενώ για τη 2η ομάδα, εκτιμάται ότι είναι ίσος με 240.5. Άρα έχουμε ενδείξεις ότι η διάμεσος της κατανομής του χρόνου επίλυσης προβλήματος στη 2η ομάδα είναι μεγαλύτερη έναντι της διαμέσου της κατανομής στην 1η ομάδα. □

Παράδειγμα 13.16. (**Έλεγχος Mann-Whitney - Ασυμπτωτικό Τεστ**): Στον Πίνακα 13.16 δίνονται τιμές δύο τυχαίων και ανεξάρτητων δειγμάτων από δύο πληθυσμούς. Να ελέγξετε σε ε.σ. 10% την υπόθεση ότι οι δύο πληθυσμοί μπορούν να θεωρηθούν ταυτόσημοι, εφαρμόζοντας το τεστ των Mann-Whitney.

Λύση Παραδείγματος 13.16. Θέλουμε να ελέγξουμε την υπόθεση ότι τα δύο δείγματα προέρχονται από τον ίδιο πληθυσμό. Ο έλεγχος διατυπώνεται ως:

$$H_0 : P(X < Y) = \frac{1}{2} \text{ έναντι της } H_1 : P(X < Y) \neq \frac{1}{2}.$$

Θα εφαρμόσουμε τον έλεγχο των Mann-Whitney που παρουσιάστηκε στο Κεφάλαιο 6 (βλ. Ενότητα 6.3).

(με χρήση SPSS) Αρχικά, εισάγουμε τα δεδομένα σε ένα κενό φύλλο εργασίας του SPSS με τον ακόλουθο τρόπο. Εισάγουμε σε μια στήλη πρώτα τις τιμές X (γραμμές 1-40) και, στη συνέχεια, ακριβώς από κάτω, εισάγουμε τις τιμές Y (γραμμές 41-80). Μετονομάζουμε τη στήλη σε measurement. Στη διπλανή στήλη, εισάγουμε την τιμή 1 στις γραμμές 1-40 και την τιμή 2 στις γραμμές 41-80, για να δηλώσουμε κατά αυτόν τον τρόπο από ποιον πληθυσμό έχει προέλθει η κάθε τιμή. Μετονομάζουμε τη στήλη σε group.

Στη συνέχεια, από το κεντρικό παράθυρο διαλόγου επιλέγουμε:

Analyze / Nonparametric Tests / 2 Independent Samples

Α/Α	X	Y	Α/Α	X	Y
1	20.75	22.16	21	18.78	13.18
2	23.20	38.08	22	19.01	14.08
3	18.64	12.64	23	23.03	36.83
4	20.22	19.43	24	16.20	5.70
5	17.77	9.70	25	19.94	18.10
6	19.59	16.48	26	18.56	12.37
7	22.02	29.78	27	23.09	37.29
8	20.38	20.27	28	19.89	17.87
9	20.97	23.39	29	21.49	26.41
10	20.83	22.63	30	20.23	19.50
11	20.08	18.75	31	20.98	23.46
12	19.07	14.31	32	18.02	10.49
13	22.31	31.71	33	17.85	9.98
14	19.09	14.38	34	19.85	17.67
15	20.11	18.90	35	19.94	18.08
16	21.14	24.33	36	18.63	12.60
17	15.71	4.74	37	24.33	47.42
18	22.70	34.40	38	18.04	10.57
19	19.85	17.68	39	20.77	22.32
20	20.91	23.05	40	21.30	25.28

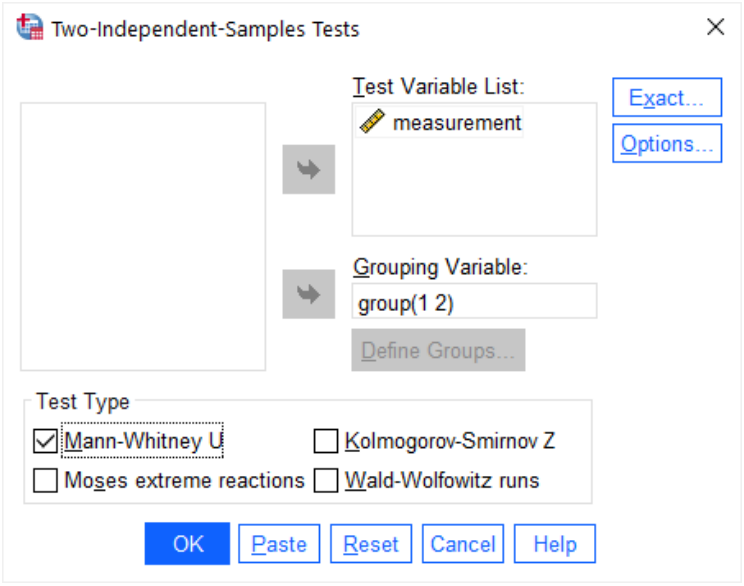
Πίνακας 13.14: Τιμές δύο τυχαίων και ανεξάρτητων δειγμάτων από δύο πληθυσμούς.

Στο νέο παράθυρο διαλόγου που προκύπτει (βλ. Εικόνα 13.54) εισάγουμε τη στήλη measurement στο πεδίο Test Variable List και τη στήλη group στο πεδίο Grouping Variable. Πατάμε το Define Groups και στο παράθυρο διαλόγου που ανοίγει δίνουμε στο Group 1 την τιμή 1, στο Group 2 την τιμή 2 και πατάμε Continue. Στη συνέχεια πατάμε OK.

Αρχικά, από το output της ανάλυσης (βλ. Εικόνα 13.55) έχουμε τη μέση τιμή των τάξεων, καθώς και το άθροισμα αυτών, για τις 40 παρατηρήσεις κάθε δείγματος. Υπενθυμίζεται ότι οι τάξεις υπολογίζονται ως προς το κοινό διατεταγμένο δείγμα (συνολικά 80 παρατηρήσεις).

Στη συνέχεια, από τον πίνακα αποτελεσμάτων που δίνεται στην Εικόνα 13.56, έχουμε την τιμή της στατιστικής συνάρτησης (γραμμή Mann-Whitney U) καθώς και την τιμή της στατιστικής συνάρτησης Z για την ασυμπτωτική μορφή του τεστ. Σε αντίθεση με το προηγούμενο παράδειγμα, εδώ χρησιμοποιήσαμε τον ασυμπτωτικό έλεγχο καθώς τα μεγέθη δείγματος από κάθε πληθυμό είναι μεγαλύτερα του 10. Η p-τιμή του ελέγχου είναι $0.379 > 0.10$ και άρα, ακόμη και σε ε.σ. 10%, δεν απορρίπτεται η H_0. Άρα δεν μπορούμε να απορρίψουμε την υπόθεση ότι τα δύο δείγματα προέρχονται από τον ίδιο πληθυσμό.

(με χρήση R) Αρχικά, εισάγουμε τα δεδομένα σε δύο διανύσματα, έστω αυτά `x` και `y`, ενώ υλοποιούμε τον έλεγχο με την εντολή `wilcox.test(...)`, όπως φαίνεται παρακάτω. Ειδικότερα, στην εντολή `wilcox.test(...)` εισάγουμε, αρχικά, ως όρισμα τα διανύσματα με τις διαθέσιμες μετρήσεις και, στη συνέχεια, δηλώνουμε τον τύπο της εναλλακτικής υπόθεσης. Πιο συγκεκριμένα, στο όρισμα `mu` δηλώνεται η προς έλεγχο τιμή για τη διαφορά των πληθυσμιακών διαμέσων $m_X - m_Y$ (εδώ η τιμή 0), ενώ στο όρισμα `alternative = 'two.sided'` επιλέγουμε να κάνουμε τον έλεγχο με δίπλευρη εναλλακτική. Επισημαίνεται ότι, αν θέλουμε να κάνουμε κάποιον από τους μονόπλευρους ελέγχους, δηλώνουμε `alternative = 'less'` ή `alternative = 'greater'`, οπότε οι αντίστοιχες εναλλακτικές είναι $H_1 : P(X < Y) < 1/2$ και $H_1 : P(X < Y) > 1/2$, αντίστοιχα. Καθώς πρόκειται για ανεξάρτητα δείγματα,



Εικόνα 13.54: Παράθυρο διαλόγου Two-Independent Samples Tests.

Ranks

	group	N	Mean Rank	Sum of Ranks
measurement	1,00	40	42,79	1711,50
	2,00	40	38,21	1528,50
	Total	80		

Εικόνα 13.55: Άθροισμα τάξεων και μέσος όρος τάξεων.

δηλώνουμε στο όρισμα `paired` την τιμή `FALSE`. Επίσης, με χρήση του ορίσματος `correct = TRUE` ή `correct = FALSE` χρησιμοποιείται ή όχι η διόρθωση συνεχείας κατά την εφαρμογή του ασυμπτωτικού ελέγχου. Στο συγκεκριμένο παράδειγμα επιλέξαμε να μην υλοποιηθεί η διόρθωση συνεχείας (ώστε να έχουμε ταύτιση των αποτελεσμάτων με αυτά του SPSS).

```
> x<-c(20.75,23.20,18.64,20.22,17.77,19.59,22.02,20.38,20.97,20.83,
+ 20.08,19.07,22.31,19.09,20.11,21.14,15.71,22.70,19.85,20.91,
+ 18.78,19.01,23.03,16.20,19.94,18.56,23.09,19.89,21.49,20.23,
+ 20.98,18.02,17.85,19.85,19.94,18.63,24.33,18.04,20.77,21.30)
> y<-c(22.16,38.08,12.64,19.43,9.70,16.48,29.78,20.27,23.39,22.63,
+ 18.75,14.31,31.71,14.38,18.90,24.33,4.74,34.40,17.68,23.05,
+ 13.18,14.08,36.83,5.70,18.10,12.37,37.29,17.87,26.41,19.50,
+ 23.46,10.49,9.98,17.67,18.08,12.60,47.42,10.57,22.32,25.28)
> wilcox.test(x, y,alternative = 'two.sided',mu = 0,
+ paired = FALSE,correct = FALSE)
```

Το output της παραπάνω ανάλυσης είναι το ακόλουθο:

```
        Wilcoxon rank sum test
data:  x and y
W = 891.5, p-value = 0.3786
alternative hypothesis: true location shift is not equal to 0
```

Από τα αποτελέσματα της ανάλυσης διαπιστώνουμε ότι η τιμή της σ.σ.ε. είναι ίση με 891.5. Η τιμή αυτή υπολογίζεται από τη σχέση (βλ. Κεφάλαιο 6) $U = \min\{U_1, U_2\}$, όπου $U_i = R_i - n_i(n_i + 1)/2$ και R_i είναι το

Test Statistics[a]

	measurement
Mann-Whitney U	708,500
Wilcoxon W	1528,500
Z	-,880
Asymp. Sig. (2-tailed)	,379

a. Grouping Variable: group

Εικόνα 13.56: Αποτελέσματα ελέγχου Mann-Whitney.

άθροισμα των τάξεων των παρατηρήσεων του i-οστού δείγματος, ως προς το σύνολο των διαθέσιμων μετρήσεων (κοινό δείγμα), με n_1, n_2 να είναι το μέγεθος του 1ου και του 2ου δείγματος, αντίστοιχα. Η p-τιμή είναι ίση με 0.3786, ίδια με αυτήν που προέκυψε με χρήση του SPSS, καθώς και τα δύο πακέτα, χρησιμοποιούν την ασυμπτωτική κατανομή του ελέγχου (το SPSS χωρίς τη διόρθωση συνεχείας). Άρα, αφού η p-τιμή είναι ίση με $0.3786 > 0.05$, συμπεραίνουμε πως δεν απορρίπτεται η H_0. Επομένως, δεν μπορούμε να απορρίψουμε την υπόθεση ότι οι τιμές X_i, Y_i προέρχονται από πληθυσμούς με κοινή διάμεσο. □

13.4.2 Περισσότεροι από δύο πληθυσμοί

Στην ενότητα αυτή θα υλοποιηθούν, μέσω παραδειγμάτων και με χρήση του SPSS και της R, οι έλεγχοι που εκτενώς μελετήθηκαν στην Ενότητα 6.4 και αφορούν τον έλεγχο διαμέσων περισσότερων από δύο πληθυσμών.

Παράδειγμα 13.17. (**Έλεγχος Kruskal-Wallis**): Τα δεδομένα του Πίνακα 13.17 αφορούν τη συγκέντρωση φωσφόρου σε γεωργικό έδαφος, όπως αυτή μετρήθηκε μετά τη χρήση τεσσάρων διαφορετικών λιπασμάτων. Χρησιμοποιώντας τα δεδομένα αυτά να ελέγξετε (ε.σ. 1%) την υπόθεση της ισότητας των διαμέσων των κατανομών της συγκέντρωσης φωσφόρου στα τέσσερα είδη λιπασμάτων και να εντοπίσετε τις όποιες διαφορές, σε περίπτωση που υπάρχουν.

Λίπασμα 1	8.1	4.9	7.0	8.5	8.0	6.8	5.3		
Λίπασμα 2	11.5	11.9	12.1	10.8	10.9	12.3			
Λίπασμα 3	15.3	19.4	16.4	16.3	15.0	16.8	17.2	17.6	15.9
Λίπασμα 4	23.0	35.0	28.4	30.1	26.7	33.4	32.9	27.6	29.6

Πίνακας 13.15: Δεδομένα συγκέντρωσης φωσφόρου.

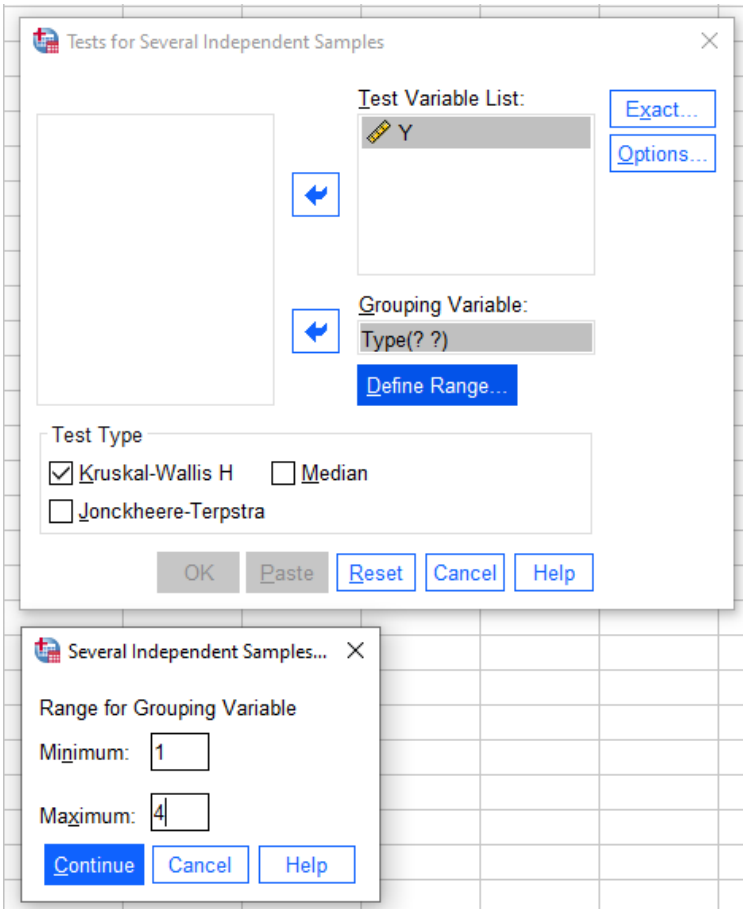
Λύση Παραδείγματος 13.17. Έστω F_i, $i = 1,...,4$, η αθροιστική συνάρτηση κατανομής που περιγράφει τη συγκέντρωση φωσφόρου σε γεωργικό έδαφος χρησιμοποιώντας το Λίπασμα i, $i = 1,...,4$. Έχουμε 4 το πλήθος ανεξάρτητα μεταξύ τους τυχαία δείγματα, ένα δείγμα από καθέναν από αυτούς τους τέσσσερις πληθυσμούς, μεγέθους $n_1 = 7$, $n_2 = 6$, $n_3 = 9$ και $n_4 = 9$, αντίστοιχα. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση: $H_0 : F_1(x) = ... = F_4(x)$, για κάθε $x \in \mathbb{R}$ έναντι της εναλλακτικής υπόθεσης H_1 : όχι η H_0. Αν υποθέσουμε ότι οι κατανομές F_i διαφέρουν μόνο ως προς τη θέση τους, τότε ο συγκεκριμένος έλεγχος διατυπώνεται ως $H_0 : m_1 = ... = m_4$ έναντι της εναλλακτικής υπόθεσης H_1 : υπάρχει τουλάχιστον ένα ζεύγος i,j με $i \neq j$, $i,j = 1,...,4$, τέτοιο ώστε $m_i \neq m_j$, όπου m_l η διάμεσος του l-οστού πληθυσμού, $l = 1,...,4$. Σε αυτήν την περίπτωση, ο έλεγχος που θα διεξαχθεί είναι αυτός των Kruskal-Wallis (βλ. Ενότητα 6.4.1).

(με χρήση SPSS) Για να χρησιμοποιήσουμε το SPSS, εισάγουμε τα δεδομένα του Πίνακα 13.17 σε ένα φύλλο εργασίας με τον τρόπο που περιγράφεται στη συνέχεια. Στο πλαίσιο αυτό, αρχικά, εισάγουμε σε μία στήλη τις συγκεντρώσεις φωσφόρου από τα 31 το πλήθος γεωγραφικά εδάφη, δηλαδή εισάγουμε συνολικά $n = 31$ το πλήθος τιμές ($n = n_1 + n_2 + n_3 + n_4 = 31$). Μετονομάζουμε τη στήλη σε Y. Στη διπλανή στήλη εισάγουμε αρχικά την τιμή 1 (=Λίπασμα 1) στις 7 πρώτες γραμμές, στις γραμμές 8-13 εισάγουμε την τιμή 2, στις γραμμές 14-22 εισάγουμε την τιμή 3 και στις γραμμές 23-31 εισάγουμε την τιμή 4. Μετονομάζουμε τη στήλη σε Type.

Από το κεντρικό παράθυρο διαλόγου επιλέγουμε

Analyze / Nonparametric Tests / Legacy Dialogs / k-Independent Samples

Στο νέο παράθυρο διαλόγου που προκύπτει (βλ. Εικόνα 13.57) εισάγουμε στο πεδίο Test Variable List τη στήλη Y, ενώ στο πεδίο Grouping Variable εισάγουμε τη στήλη Type. Πατάμε το πλαίσιο Define Range και δίνουμε την τιμή 1 στο Minimum και την τιμή 4 στο Maximum. Πατάμε Continue, επιλέγουμε στο Test Type το Kruskal-Wallis H και πατάμε OK.



Εικόνα 13.57: Παράθυρο διαλόγου Tests for Several Independent Samples του SPSS.

Το output της ανάλυσης δίνεται στις Εικόνες 13.58 και 13.59. Αρχικά, στην Εικόνα 13.58 έχουμε τις μέσες τιμές των τάξεων (στήλη Mean Rank) των παρατηρήσεων από καθεμία υποομάδα (κατηγορία λιπάσματος), ενώ στην Εικόνα 13.59 έχουμε το αποτέλεσμα του ελέγχου.

Παρατηρούμε ότι ο έλεγχος γίνεται με χρήση της στατιστικής συνάρτησης ελέγχου H (ασυμπτωτική μορφή ελέγχου). Η τιμή της στατιστικής συνάρτησης H είναι ίση με 27.998, ενώ η p-τιμή είναι < 0.001. Άρα έχουμε

Ranks

	Type	N	Mean Rank
Y	1	7	4,00
	2	6	10,50
	3	9	18,00
	4	9	27,00
	Total	31	

Εικόνα 13.58: Πίνακας με τη μέση τιμή των τάξεων (Mean Rank) στις 4 ομάδες.

Test Statistics[a,b]

	Y
Kruskal-Wallis H	27,998
df	3
Asymp. Sig.	<,001

a. Kruskal Wallis Test

b. Grouping Variable: Type

Εικόνα 13.59: Αποτελέσματα του ελέγχου Kruskal-Wallis με χρήση SPSS.

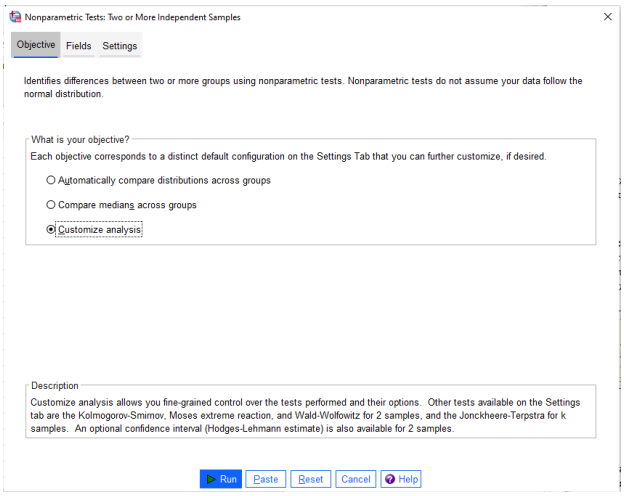
ισχυρή ένδειξη έναντι της H_0 και δεν μπορούμε σε κάποιο σύνηθες ε.σ. να υποθέσουμε ότι η διάμεση συγκέντρωση φωσφόρου είναι η ίδια μεταξύ των 4 λιπασμάτων. Αυτό σημαίνει ότι τουλάχιστον δύο από τις διαμέσους θα διαφέρουν μεταξύ τους. Μια πρώτη ένδειξη μπορούμε να λάβουμε από τον πίνακα που δίνεται στην Εικόνα 13.58. Παρατηρούμε ότι η μέση τιμή των τάξεων για τις τιμές της συγκέντρωσης υπό το Λίπασμα 1 είναι αρκετά μικρότερη έναντι της αντίστοιχης μέσης τιμής των τάξεων για τις τιμές της συγκέντρωσης υπό το Λίπασμα 3 ή το Λίπασμα 4. Άρα μικρότερες τιμές συγκέντρωσης φωσφόρου παρατηρούνται κατά τη χρήση του Λιπάσματος 1, ενώ μεγαλύτερες κατά τη χρήση των Λιπασμάτων 3 ή 4. Φυσικά, κάτι τέτοιο δεν αποτελεί στατιστική συμπερασματολογία, απλώς μας παρέχει κάποιες ενδείξεις.

Αφού απορρίψαμε την υπόθεση της ισότητας των διαμέσων, θα προχωρήσουμε με τη διεξαγωγή πολλαπλών συγκρίσεων, ώστε να διαπιστώσουμε ποιες από τις διαμέσους διαφέρουν στατιστικά σημαντικά έναντι των υπολοίπων. Για τον σκοπό αυτόν επιλέγουμε από το κεντρικό παράθυρο διαλόγου:

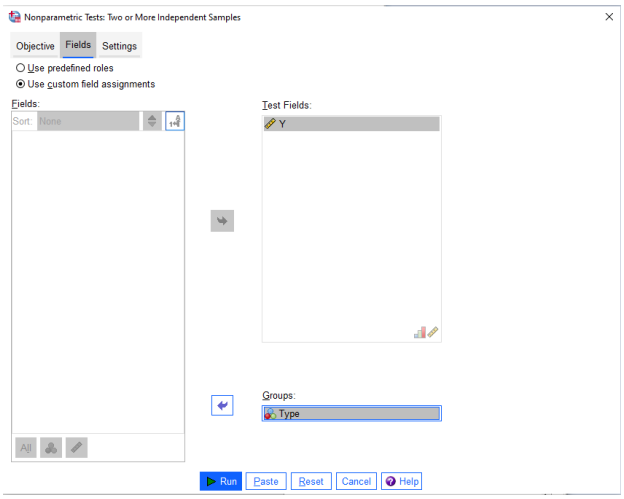
Analyze / Nonparametric Tests / Independent Samples

Στο νέο παράθυρο διαλόγου που εμφανίζεται (Εικόνα 13.60) υπάρχουν τρεις καρτέλες (Objective, Fields, Settings). Στην καρτέλα Objective, επιλέγουμε Customize Analysis ώστε να παραμετροποιήσουμε, όπως επιθυμούμε, την ανάλυση. Έπειτα, στην καρτέλα Fields (Εικόνα 13.61) επιλέγουμε Use custom field assignments και εισάγουμε στο Test Fields τη στήλη Y, ενώ στο Groups εισάγουμε τη στήλη Type. Στη συνέχεια, στην καρτέλα Settings, στο Select an item: Choose Tests (Εικόνα 13.62), επιλέγουμε Customize tests και μετά Kruskal-Wallis 1-way ANOVA (k samples). Για να διεξάγουμε πολλαπλές συγκρίσεις, επιλέγουμε στο Multiple comparisons: All pairwise. Με τη συγκεκριμένη επιλογή διεξάγουμε ελέγχους για την ισότητα δύο διαμέσων, για όλα τα δυνατά ζεύγη των $k = 4$ υποπληθυσμών, δηλαδή θα διεξαχθούν συνολικά 6 ζεύγη ελέγχων. Τέλος, αν επιλέξουμε Select an item: Test Options (Εικόνα 13.63), μπορούμε αλλάξουμε το επίπεδο σημαντικότητας για τους ελέγχους, καθώς και τον συντελεστή εμπιστοσύνης των

αντίστοιχων διαστημάτων εμπιστοσύνης. Στο πλαίσιο της υλοποίησης σε αυτό το παράδειγμα επιλέγουμε ε.σ. 1% (Significance level: 0.01) και συντελεστή εμπιστοσύνης 99% (Confidence interval(%): 99.0). Πατάμε Run και προκύπτει το output της ανάλυσης.



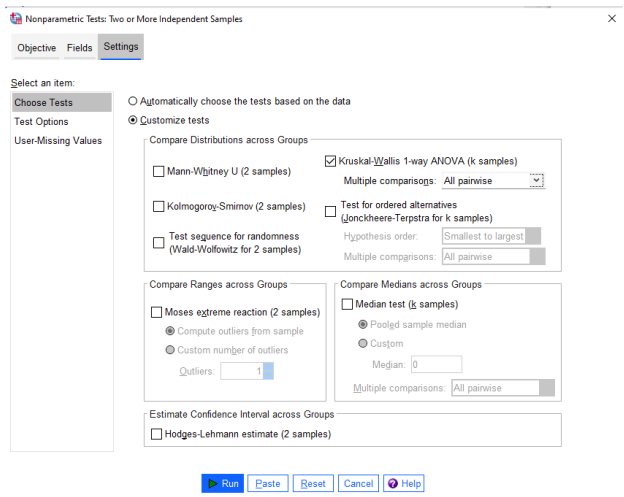
Εικόνα 13.60: Καρτέλα Objective στο Independent Samples.



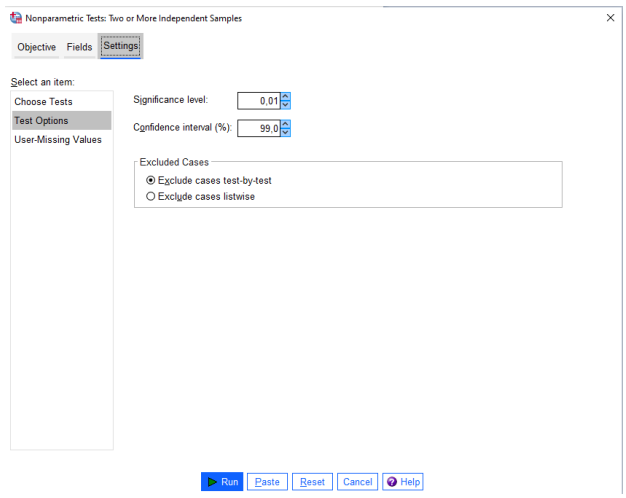
Εικόνα 13.61: Καρτέλα Fields στο Independent Samples.

Τα αποτελέσματα των πολλαπλών συγκρίσεων δίνονται στον πίνακα με τίτλο Pairwise Comparisons of Type (Εικόνα 13.64). Για τον υπολογισμό των p-τιμών των επιμέρους συγκρίσεων έχει χρησιμοποιηθεί η διόρθωση Bonferroni έτσι ώστε το ολικό επίπεδο σημαντικότητας να διατηρηθεί στα επιθυμητά επίπεδα του 1%. Οι τιμές στη στήλη Std. Test Statistic έχουν υπολογιστεί με χρήση της μεθόδου του Dunn (1964) (βλ., επίσης, Κεφάλαιο 6, Ενότητα 6.4.1). Επίσης, οι τιμές στη στήλη Sig. αντιστοιχούν στην p-τιμή, η οποία είναι ίση με $2 \cdot [1 - \Phi(|z^*|)]$, όπου z^* είναι η τιμή στη στήλη Std. Test Statistic. Στη συνέχεια, εφαρμόζεται η διόρθωση Bonferroni, όπου κάθε τιμή στη στήλη Sig. πολλαπλασιάζεται με το πλήθος των ανά δύο συγκρίσεων (εδώ είναι 6) και έτσι προκύπτουν οι τιμές που δίνονται υπό τη στήλη Adj. Sig. στην Εικόνα 13.64.

Συγκρίνοντας τις τιμές στη στήλη Adj. Sig. με την τιμή 0.01, παρατηρούμε ότι η συγκέντρωση φωσφόρου είναι στατιστικά σημαντική μεταξύ των λιπασμάτων 1 και 4, όπως και μεταξύ των λιπασμάτων 2 και 4. Ακόμα, δεν μπορούμε να υποθέσουμε ότι η διάμεση συγκέντρωση φωσφόρου διαφέρει μεταξύ των λιπασμάτων 1 και 2, των λιπασμάτων 2 και 3, όπως και για τα λιπάσματα 3 και 4.



Εικόνα 13.62: Καρτέλα Settings: Choose Tests στο Independent Samples.



Εικόνα 13.63: Καρτέλα Settings: Test Options στο Independent Samples.

Pairwise Comparisons of Type

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
1-2	-6,500	5,058	-1,285	,199	1,000
1-3	-14,000	4,582	-3,055	,002	,013
1-4	-23,000	4,582	-5,020	<,001	,000
2-3	-7,500	4,792	-1,565	,118	,705
2-4	-16,500	4,792	-3,443	<,001	,003
3-4	-9,000	4,286	-2,100	,036	,214

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is ,010.

a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

Εικόνα 13.64: Αποτελέσματα πολλαπλών συγκρίσεων με χρήση SPSS.

(με χρήση R) Αρχικά, εισάγουμε σε ένα διάνυσμα, έστω `x`, τις μετρήσεις συγκέντρωσης φωσφόρου και σε ένα διάνυσμα-παράγοντα (factor), έστω `y`, εισάγουμε τις τιμές 1-4, ανάλογα με το λίπασμα που έχει χρησιμοποιηθεί στο γεωργικό έδαφος. Στη συνέχεια, χρησιμοποιούμε την εντολή `kruskal.test(...)`, όπως παρακάτω:

```
> x<-c(8.10,4.90,7.00,8.50,8.00,6.80,5.30,11.50,11.90,12.10,
+ 10.80,10.90,12.30,15.30,19.40,16.40,16.30,15.00,16.80,
+ 17.20,17.60,15.90,23.00,35.00,28.40,30.10,26.70,33.40,
+ 32.90,27.60,29.60)
> y<-c(1,1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,4,4,4,4,4,4,4,4,4)
> kruskal.test(x,as.factor(y))
```

```
        Kruskal-Wallis rank sum test

data:  x and as.factor(y)
Kruskal-Wallis chi-squared = 27.998, df = 3, p-value = 3.636e-06
```

Από τα αποτελέσματα της ανάλυσης διαπιστώνουμε ότι η τιμή της σ.σ.ε. είναι ίση με 27.998, ίση με την τιμή που έδωσε το SPSS, καθώς και τα δύο πακέτα χρησιμοποιούν την ασυμπτωτική κατανομή του ελέγχου, η οποία (προσεγγιστικά και υπό την H_0) είναι η κατανομή χ^2_3. Αφού η p-τιμή είναι ίση με $3.636 \cdot 10^{-6}$, συμπεραίνουμε ότι απορρίπτεται η H_0, δηλαδή δεν μπορούμε να θεωρήσουμε ότι η διάμεση συγκέντρωση φωσφόρου είναι η ίδια μεταξύ των 4 λιπασμάτων.

Αφού απορρίψαμε την H_0, θα προχωρήσουμε στη διεξαγωγή πολλαπλών συγκρίσεων, αφού πρώτα εγκαταστήσουμε τη βιβλιοθήκη `FSA` της `R`. Έπειτα η υλοποίηση των πολλαπλών συγκρίσεων επιτυγχάνεται μέσω της εντολής `dunnTest` ως εξής:

```
library(FSA)
> x<-c(8.10,4.90,7.00,8.50,8.00,6.80,5.30,11.50,11.90,
+ 12.10,10.80,10.90,12.30,15.30,19.40,16.40,16.30,
+ 15.00,16.80,17.20,17.60,15.90,23.00,35.00,28.40,
+ 30.10,26.70,33.40,32.90,27.60,29.60)
> y<-c(1,1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,4,4,4,4,4,4,4,4,4)
> DUNN1<-dunnTest(x,as.factor(y),method='bonferroni')
> print(DUNN1)
```

Αρχικά, όπου `x` είναι το διάνυσμα με τις τιμές συγκέντρωσης φωσφόρου και `y` είναι το διάνυσμα (ως παράγοντας) με τις διαφορετικές κατηγορίες λιπάσματος. Επίσης, χρησιμοποιούμε τη μέθοδο Bonferroni (με χρήση του ορίσματος `method='bonferroni'`) ώστε το ολικό επίπεδο σημαντικότητας να διατηρηθεί στα επιθυμητά επίπεδα. Παρακάτω δίνονται τα αποτελέσματα της ανάλυσης.

```
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Bonferroni method.

  Comparison         Z      P.unadj        P.adj
1      1 - 2 -1.284994 1.987944e-01 1.000000e+00
2      1 - 3 -3.055435 2.247340e-03 1.348404e-02
```

```
3      2 - 3 -1.565119 1.175551e-01 7.053304e-01
4      1 - 4 -5.019644 5.176736e-07 3.106042e-06
5      2 - 4 -3.443261 5.747437e-04 3.448462e-03
6      3 - 4 -2.099827 3.574405e-02 2.144643e-01
```

Από το output της ανάλυσης βλέπουμε ότι έχουμε ακριβώς τις ίδιες τιμές στις στήλες Z, P.unadj και P.adj με τις στήλες Std.Test Statistic, Sig. και Adj. Sig. που προέκυψαν στο output του SPSS (βλ. Εικόνα 13.64). Στη συνέχεια, συγκρίνουμε τις τιμές στη στήλη P.adj με την τιμή 0.01. Παρατηρούμε ότι η συγκέντρωση φωσφόρου είναι στατιστικά σημαντική μεταξύ των λιπασμάτων 1 και 4, όπως και μεταξύ των λιπασμάτων 2 και 4. Επίσης, δεν μπορούμε να υποθέσουμε ότι η διάμεση συγκέντρωση φωσφόρου διαφέρει στατιστικά σημαντικά μεταξύ των λιπασμάτων 1 και 2, των λιπασμάτων 2 και 3, όπως και μεταξύ των λιπασμάτων 3 και 4.

Κλείνοντας την παρουσίαση της λύσης του συγκεκριμένου παραδείγματος με χρήση της `R`, αξίζει να αναφέρουμε πως τα πακέτα `agricolae` και `PMCMRplus` διαθέτουν περισσότερες επιλογές ως προς τη μέθοδο διατήρησης του ολικού επιπέδου σημαντικότητας στα επιθυμητά επίπεδα, ενώ δίνουν και μια ομαδοποίηση των διαφορετικών διαμέσων. Όμως, η αναλυτική παρουσίαση των συγκεκριμένων μεθόδων και τεχνικών από το κάθε πακέτο, ξεφεύγει από τους σκοπούς αυτού του κεφαλαίου. □

Παράδειγμα 13.18. (Έλεγχος Friedman) Στον Πίνακα 13.18 δίνονται μετρήσεις της συστολικής πίεσης 14 ασθενών σε τέσσερις διαφορετικές χρονικές στιγμές μιας συγκεκριμένης αγωγής. Ειδικότερα, η 1η μέτρηση λαμβάνεται πριν την έναρξη του προγράμματος, η 2η μέτρηση μετά τη χορήγηση χαπιού, η 3η μέτρηση μετά την τροποποίηση διαιτολογίου και η 4η μέτρηση μετά τη συμμετοχή σε πρόγραμμα καθημερινής άσκησης. Χρησιμοποιώντας τα δεδομένα αυτά να ελέγξετε, σε ε.σ. 5%, την υπόθεση της ισότητας των διαμέσων της συστολικής πίεσης στα τέσσερα στάδια της αγωγής και να εντοπίσετε τις όποιες διαφορές, σε περίπτωση που υπάρχουν.

Ασθενής	1η μέτρηση	2η μέτρηση	3η μέτρηση	4η μέτρηση
1	12.20	12.00	11.70	11.50
2	13.40	13.10	12.90	12.40
3	14.10	13.60	13.70	13.40
4	12.90	13.10	12.70	12.40
5	13.50	12.90	13.10	12.50
6	13.60	13.50	13.20	12.90
7	14.20	13.90	13.80	13.50
8	12.80	12.40	12.50	12.40
9	14.40	14.10	13.20	13.30
10	13.90	13.90	13.10	12.40
11	14.30	13.20	13.40	12.90
12	12.70	12.40	12.50	12.40
13	12.60	12.80	12.30	11.80
14	13.10	13.20	12.70	12.10

Πίνακας 13.16: Δεδομένα συστολικής πίεσης (σε mmHg) υπερτασικών ασθενών.

Λύση Παραδείγματος 13.18. Έστω F_i, $i = 1,...,4$, η αθροιστική συνάρτηση κατανομής που περιγράφει τη συστολική πίεση στο i-οστό στάδιο της αγωγής, $i = 1,...,4$,. Έχουμε 4 το πλήθος εξαρτημένα μεταξύ τους τυχαία δείγματα, ένα από καθέναν από αυτούς τους τέσσσερις πληθυσμούς, μεγέθους $n = 14$. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση της ισότητας των πληθυσμιακών διαμέσων

$$H_0 : m_1 = ... = m_4$$

έναντι της εναλλακτικής

$$H_1 : m_i \neq m_j, \text{ για κάποιο ζεύγος } (i,j), \text{ με } i \neq j,\ i,j = 1,...,4.$$

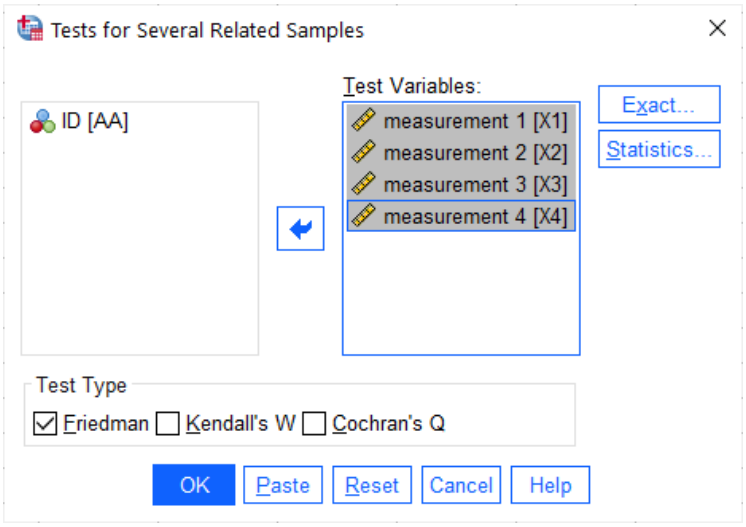
Τότε ο έλεγχος που θα διεξαχθεί είναι αυτός του Friedmann (βλ. Ενότητα 6.4.2).

(με χρήση SPSS): Για να χρησιμοποιήσουμε το SPSS, εισάγουμε τα δεδομένα του Πίνακα 13.18 (1η - 4η μέτρηση) σε ένα φύλλο εργασίας. Η εισαγωγή πρέπει να γίνει σε 4 στήλες, όπως ακριβώς παρουσιάζονται στον Πίνακα 13.18. Μετονομάζουμε τις 4 στήλες σε X1, X2, X3, X4.

Έπειτα, από το κεντρικό παράθυρο διαλόγου επιλέγουμε

Analyze / Nonparametric Tests / Legacy Dialogs / K Related Samples

Στο νέο παράθυρο διαλόγου που προκύπτει (βλ. Εικόνα 13.65) εισάγουμε στο πεδίο Test Variables τις στήλες X1-X4, ενώ στο πεδίο Test Type επιλέγουμε Friedman και πατάμε OK.



Εικόνα 13.65: Παράθυρο διαλόγου Tests for Several Related Samples του SPSS.

Το output της ανάλυσης δίνεται στις Εικόνες 13.66 και 13.67. Αρχικά, στην Εικόνα 13.66 έχουμε τις μέσες τιμές των τάξεων (στήλη Mean Rank) των παρατηρήσεων που αντιστοιχούν σε καθεμία από τις 4 μετρήσεις. Στην Εικόνα 13.67 έχουμε το αποτέλεσμα του ελέγχου. Ο έλεγχος γίνεται με χρήση του ασυμπτωτικού ελέγχου, όπου η ασυμπτωτική κατανομή της στατιστικής συνάρτησης ελέγχου ακολουθεί (υπό την H_0) την χ^2_3 κατανομή (για περισσότερες λεπτομέρειες βλ. Ενότητα 6.4.2). Η τιμή της στατιστικής συνάρτησης είναι ίση με 30.504, ενώ η p-τιμή είναι < 0.001. Άρα έχουμε ισχυρή ένδειξη έναντι της H_0 και δεν μπορούμε να υποθέσουμε ότι η διάμεση συστολική πίεση παραμένει η ίδια στα τέσσερα στάδια της αγωγής.

Αφού απορρίψαμε την H_0, θα προχωρήσουμε στη διεξαγωγή πολλαπλών συγκρίσεων. Αρχικά, από την Εικόνα 13.66 έχουμε ένδειξη ότι ο μέσος όρος των τάξεων των παρατηρήσεων που αντιστοιχούν στην 1η μέτρηση είναι μεγαλύτερος έναντι αυτών που αντιστοιχούν στην 4η μέτρηση, ενώ οι μέσοι όροι των τάξεων της 2ης και 3ης μέτρησης δεν φαίνεται να διαφέρουν σημαντικά. Φαίνεται πως στο τέλος του προγράμματος έχουμε μείωση στην τιμή της συστολικής πίεσης, ενώ δεν φαίνεται να υπάρχει μεγάλη διαφορά μετά τη λήψη χαπιού (2η μέτρηση) και μετά την αλλαγή του διαιτολογίου (3η μέτρηση). Πάντως, σε σχέση με την 1η μέτρηση, η οποία ελήφθη πριν την έναρξη του προγράμματος, έχουμε ενδείξεις ότι το πρόγραμμα επιδρά θετικά στη μείωση της συστολικής πίεσης. Όλα αυτά, όμως, δεν είναι παρά ενδείξεις και θα πρέπει να διεξάγουμε κατάλληλους στατιστικούς ελέγχους ώστε να διαπιστώσουμε ποιες πληθυσμιακές διάμεσοι διαφέρουν στατιστικά σημαντικά από τις υπόλοιπες. Για τον σκοπό αυτόν επιλέγουμε από το κεντρικό παράθυρο διαλόγου:

Ranks

	Mean Rank
measurement 1	3,75
measurement 2	2,82
measurement 3	2,29
measurement 4	1,14

Εικόνα 13.66: Πίνακας μέσων τιμών των τάξεων (Mean Rank) σε καθένα από τα 4 σύνολα μετρήσεων.

Test Statistics[a]

N	14
Chi-Square	30,504
df	3
Asymp. Sig.	<,001

a. Friedman Test

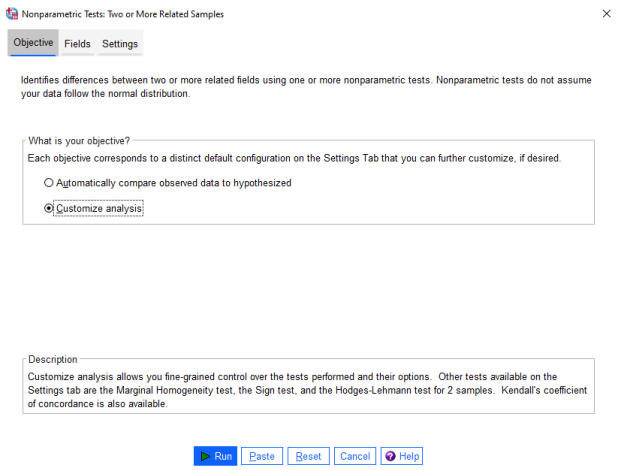
Εικόνα 13.67: Αποτελέσματα του ελέγχου Friedman για τα δεδομένα του Πίνακα 13.18.

Analyze / Nonparametric Tests / Related Samples

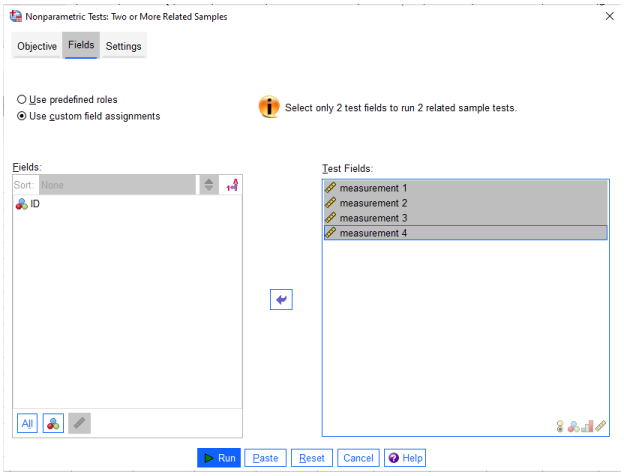
Στο νέο παράθυρο διαλόγου που εμφανίζεται (βλ. Εικόνα 13.68) υπάρχουν τρεις καρτέλες (Objective, Fields, Settings). Στην καρτέλα Objective επιλέγουμε Customize Analysis ώστε να παραμετροποιήσουμε, όπως επιθυμούμε, την ανάλυση. Στην καρτέλα Fields (Εικόνα 13.69) επιλέγουμε Use custom field assignments και εισάγουμε στο Test Fields τις στήλες X1-X4. Στη συνέχεια, στην καρτέλα Settings, στο Select an item: Choose Tests (Εικόνα 13.70), επιλέγουμε Customize tests και μετά Friedman (k samples). Για να διεξάγουμε πολλαπλές συγκρίσεις επιλέγουμε στο Multiple comparisons: All pairwise. Με τη συγκεκριμένη επιλογή διεξάγουμε ελέγχους για την ισότητα δύο πληθυσμιακών διαμέσων, για όλα τα δυνατά ζεύγη των $k = 4$ υποπληθυσμών. Τέλος, αν επιλέξουμε Select an item: Test Options (βλ. Εικόνα 13.71), μπορούμε να αλλάξουμε το επίπεδο σημαντικότητας για τους ελέγχους καθώς και τον συντελεστή εμπιστοσύνης των αντίστοιχων διαστημάτων εμπιστοσύνης. Στο πλαίσιο της υλοποίησης σε αυτό το παράδειγμα επιλέγουμε ε.σ. 5% (Significance level: 0.05) και συντελεστή εμπιστοσύνης 95% (Confidence interval(%): 95.0). Πατάμε Run και προκύπτει το output της ανάλυσης.

Τα αποτελέσματα των πολλαπλών συγκρίσεων δίνονται στον πίνακα με τίτλο Pairwise Comparisons (Εικόνα 13.72). Οι τιμές στη στήλη Test Statistic είναι οι διαφορές $|\bar{R}_i - \bar{R}_j|$, όπου $\bar{R}_i$ είναι ο μέσος όρος των τάξεων για την i-οστή μέτρηση. Οι τιμές στη στήλη Std. Test Statistic προκύπτουν ως το πηλίκο των αντίστοιχων τιμών στις στήλες Test Statistic και Std. Error. Η στήλη Sig. δίνει τις p-τιμές οι οποίες υπολογίζονται με χρήση του τύπου $2 \cdot [1 - \Phi(|z^*|)]$, όπου z^* είναι η τιμή στη στήλη Std. Test Statistic. Στη συνέχεια, εφαρμόζεται η διόρθωση Bonferroni, έτσι ώστε το ολικό επίπεδο σημαντικότητας στο οποίο διεξάγονται οι έλεγχοι να διατηρηθεί στα επιθυμητά επίπεδα, δηλαδή να είναι 5%. Κάθε τιμή στη στήλη Sig. πολλαπλασιάζεται με το πλήθος των ανά δύο συγκρίσεων (εδώ είναι 6) και έτσι προκύπτουν οι τιμές στη στήλη Adj. Sig.

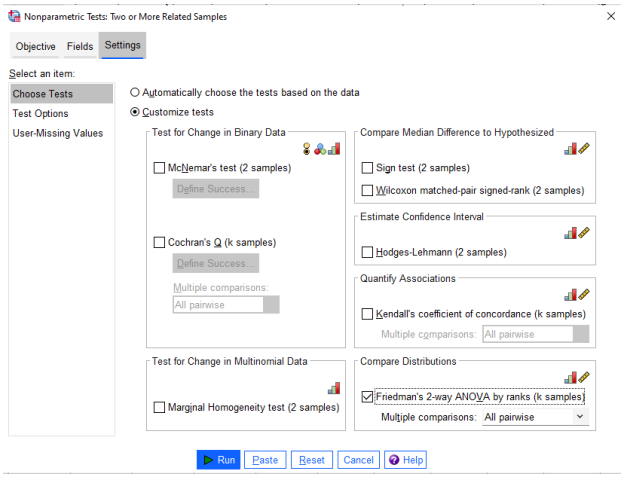
Συγκρίνοντας τις τιμές στη στήλη Adj. Sig. με την τιμή 0.05, παρατηρούμε ότι η διάμεση συστολική πίεση διαφέρει σημαντικά μεταξύ 2ης και 4ης μέτρησης, μεταξύ 1ης και 4ης, καθώς και μεταξύ 1ης και 3ης. Επίσης,



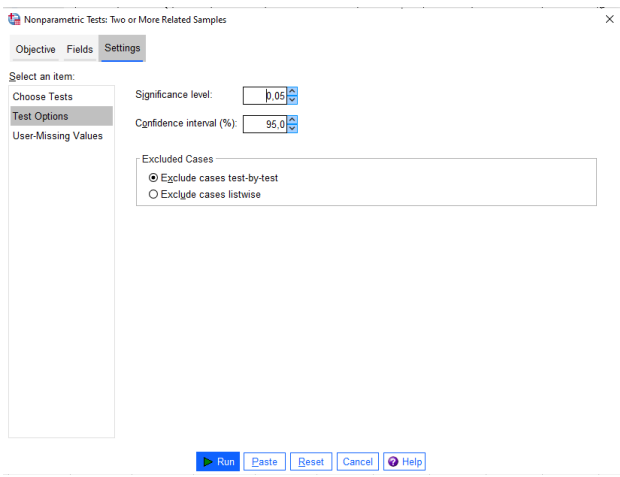
Εικόνα 13.68: Καρτέλα Objective στο Related Samples.



Εικόνα 13.69: Καρτέλα Fields στο Related Samples.



Εικόνα 13.70: Καρτέλα Settings: Choose Tests στο Related Samples.



Εικόνα 13.71: Καρτέλα Settings: Test Options στο Related Samples.

προκύπτει ότι δεν υπάρχουν σαφείς ενδείξεις ότι η διαφορά είναι (στατιστικά) σημαντική μεταξύ 1ης και 2ης μέτρησης, όπως και μεταξύ 2ης και 3ης, αλλά και μεταξύ 3ης και 4ης μέτρησης.

Pairwise Comparisons

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
measurement 4-measurement 3	1,143	,488	2,342	,019	,115
measurement 4-measurement 2	1,679	,488	3,440	<,001	,003
measurement 4-measurement 1	2,607	,488	5,343	<,001	,000
measurement 3-measurement 2	,536	,488	1,098	,272	1,000
measurement 3-measurement 1	1,464	,488	3,001	,003	,016
measurement 2-measurement 1	,929	,488	1,903	,057	,342

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is ,050.

a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

Εικόνα 13.72: Αποτελέσματα πολλαπλών συγκρίσεων για $k = 4$ εξαρτημένα δείγματα με χρήση SPSS.

(με χρήση R) Αρχικά, εισάγουμε σε έναν πίνακα (matrix), έστω αυτός `x`, όλες τις διαθέσιμες μετρήσεις που δίνονται στον Πίνακα 13.18. Ειδικότερα, η εισαγωγή γίνεται δίνοντας αρχικά το $c(12.20, ...12.10)$, δηλαδή όλες τις μετρήσεις κατά τέτοιο τρόπο ώστε να προηγούνται οι μετρήσεις συστολικής πίεσης του 1ου ατόμου στις 4 χρονικές στιγμές, να έπονται οι μετρήσεις συστολικής πίεσης του 2ου ατόμου στις 4 χρονικές στιγμές και τέλος να δίνονται οι μετρήσεις συστολικής πίεσης του 14ου ατόμου στις 4 χρονικές στιγμές. Τα δεδομένα αυτά συνθέτουν έναν πίνακα 14×4 μέσω της δήλωσης ότι `ncol=4`. Στη συνέχεια, χρησιμοποιούμε την εντολή `friedman.test(x)`, όπως παρακάτω:

```
> x<-matrix(c(12.20,12.00,11.70,11.50,13.40,13.10,12.90,
+ 12.40,14.10,13.60,13.70,13.40,12.90,13.10,12.70,12.40,
+ 13.50,12.90,13.10,12.50,13.60,13.50,13.20,12.90,14.20,
+ 13.90,13.80,13.50,12.80,12.40,12.50,12.40,14.40,14.10,
+ 13.20,13.30,13.90,13.90,13.10,12.40,14.30,13.20,13.40,
+ 12.90,12.70,12.40,12.50,12.40,12.60,12.80,12.30,11.80,
+ 13.10,13.20,12.70,12.10),
+ ncol=4,byrow=T)
> friedman.test(x)
```

Στη συνέχεια, δίνονται τα αποτελέσματα της ανάλυσης από όπου και διαπιστώνουμε ότι η τιμή της σ.σ.ε. είναι ίση με 30.504, ίση με την τιμή που έδωσε το SPSS, καθώς και τα δύο πακέτα χρησιμοποιούν την

ασυμπτωτική κατανομή του ελέγχου, η οποία (προσεγγιστικά και υπό την H_0) είναι η κατανομή χ^2_3. Αφού η p-τιμή είναι ίση με $1.081 \cdot 10^{-6}$, συμπεραίνουμε ότι απορρίπτεται η H_0 και άρα δεν μπορούμε να θεωρήσουμε ότι η διάμεσος της συστολικής πίεσης παραμένει η ίδια στα τέσσερα στάδια της αγωγής.

```
        Friedman rank sum test

data:  x
Friedman chi-squared = 30.504, df = 3, p-value = 1.081e-06
```

Αφού απορρίψαμε την H_0, θα προχωρήσουμε στη διεξαγωγή πολλαπλών συγκρίσεων. Αυτό θα γίνει με χρήση της βιβλιοθήκης `PMCMRplus`. Οι σχετικές εντολές δίνονται παρακάτω:

```
> library(PMCMRplus)
> x<-matrix(c(12.20,12.00,11.70,11.50,13.40,13.10,12.90,12.40,
+ 14.10,13.60,13.70,13.40,12.90,13.10,12.70,12.40,13.50,12.90,
+ 13.10,12.50,13.60,13.50,13.20,12.90,14.20,13.90,13.80,13.50,
+ 12.80,12.40,12.50,12.40,14.40,14.10,13.20,13.30,13.90,
+ 13.90,13.10,12.40,14.30,13.20,13.40,12.90,12.70,12.40,
+ 12.50,12.40,12.60,12.80,12.30,11.80,13.10,13.20,12.70,12.10),
+ ncol =4, byrow =T,dimnames=list(1:14,c('X1','X2','X3','X4')))
> frdAllPairsSiegelTest(y=x, p.adjust = 'bonferroni')
```

Αρχικά, χρησιμοποιούμε τον πίνακα `x` στον οποίο βρίσκονται τα διαθέσιμα δεδομένα. Με το όρισμα dimnames δίνουμε ως λίστα τα blocks (διάνυσμα `1:14`), καθώς και τις μετρήσεις (το διάνυσμα `c('X1','X2','X3','X4')`, δηλώνει το ότι έχουμε 4 διαφορετικές μετρήσεις για κάθε άτομο). Στη συνέχεια, χρησιμοποιούμε την εντολή `frdAllPairsSiegelTest(...)`, όπου δίνουμε στο όρισμα `y` τον πίνακα `x` και ως μέθοδο διόρθωσης, ώστε το ολικό επίπεδο σημαντικότητας να παραμένει στα επιθυμητά επίπεδα, δίνουμε τη μέθοδο `'bonferroni'` (όρισμα `p.adjust ='bonferroni'`). Το αποτέλεσμα της ανάλυσης δίνεται παρακάτω:

```
        Pairwise comparisons using Siegel-Castellan all-pairs test for
        a two-way balanced complete block design

data: y

   X1      X2     X3
X2 0.3422  -      -
X3 0.0162  1.0000 -
X4 5.5e-07 0.0035 0.1150

P value adjustment method: bonferroni
```

Από τα αποτελέσματα της ανάλυσης βλέπουμε ότι οι τιμές στον πίνακα είναι ακριβώς οι τιμές της στήλης Adj.Sig. στον πίνακα Pairwise Comparisons που έδωσε το SPSS (βλ. Εικόνα 13.72). Συγκρίνοντας τις τιμές αυτές με το ολικό επίπεδο σημαντικότητας $\alpha = 0.05$, έχουμε ότι η διάμεση συστολική πίεση διαφέρει στατιστικά σημαντικά μεταξύ 2ης και 4ης μέτρησης, μεταξύ 1ης και 4ης, καθώς και μεταξύ 1ης και 3ης. Επίσης, προκύπτει ότι δεν υπάρχουν σαφείς ενδείξεις ότι η διαφορά είναι (στατιστικά) σημαντική μεταξύ 1ης και 2ης μέτρησης, όπως και μεταξύ 2ης και 3ης, αλλά και μεταξύ 3ης και 4ης μέτρησης.

Κλείνοντας την παρουσίαση της λύσης του συγκεκριμένου παραδείγματος με χρήση της `R`, πρέπει να αναφέρουμε ότι ο τρόπος διεξαγωγής των πολλαπλών συγκρίσεων βασίζεται στη μέθοδο που προτάθηκε από τους Siegel and Castellan (1988). Χρησιμοποιώντας τις εντολές `frdAllPairsNemenyiTest(...)` και `frdAllPairsConoverTest(...)` μπορούμε να διεξάγουμε πολλαπλές συγκρίσεις για τον έλεγχο του Friedman εφαρμόζοντας τον έλεγχο των Wilcoxon–Nemenyi–McDonald–Thompson. Ενδεικτικά, παραπέμπουμε στην Ενότητα 7.3 του συγγράμματος των Hollander *et al.* (2014) και στον έλεγχο του Conover (1998), αντίστοιχα. Ο έλεγχος του Conover (1998) είναι αυτός που παρουσιάστηκε στο Κεφάλαιο 6. Επίσης, εκτός από το πακέτο `PMCMRplus`, υπάρχει και το πακέτο `agricolae`, στο οποίο με χρήση της εντολής `friedman(...)` μπορούμε να διεξάγουμε πολλαπλές συγκρίσεις για τον έλεγχο του Friedman. Η αναλυτική παρουσίαση όλων αυτών των τεχνικών και μεθόδων, ξεφεύγει από τους σκοπούς αυτού του κεφαλαίου. □

Παράδειγμα 13.19. (**Έλεγχος Cochran's Q**): Μια εταιρεία πρόκειται να αξιολογήσει τους υπαλλήλους της υποβάλλοντάς τους σε 5 διαφορετικά τεστ δεξιοτήτων. Θέλοντας να ελέγξει αν αυτά τα τεστ έχουν την ίδια δυσκολία, 9 υπάλληλοι του τμήματος πωλήσεων μιας εταιρείας καλλυντικών αξιολογήθηκαν από το τμήμα προσωπικού με καθένα από τα 5 τεστ. Τα δεδομένα που προέκυψαν δίνονται στον Πίνακα 13.19, όπου με 1 έχουμε συμβολίσει την επιτυχία στο τεστ και με 0 την αποτυχία. Ποιο είναι το συμπέρασμά σας σε ε.σ. 1%;

Υπάλληλος	Τεστ 1	Τεστ 2	Τεστ 3	Τεστ 4	Τεστ 5
1	1	1	0	1	1
2	1	0	0	0	1
3	0	1	1	0	1
4	0	1	0	1	1
5	1	1	1	0	1
6	1	1	1	1	1
7	1	0	0	0	0
8	0	1	0	1	1
9	1	1	0	0	1

Πίνακας 13.17: Επιδόσεις υπαλλήλων στα 5 τεστ.

Λύση Παραδείγματος 13.19. Θέλουμε να ελέγξουμε αν το ποσοστό επιτυχίας στα 5 τεστ είναι το ίδιο ή όχι. Καθώς έχουμε 5 το πλήθος εξαρτημένα δείγματα, θα χρησιμοποιήσουμε τον έλεγχο που είναι γνωστός ως Cochran's Q και που, ουσιαστικά, αποτελεί ειδική περίπτωση του ελέγχου του Friedman για δίτιμες μεταβλητές (βλ. Παρατήρηση 6.11).

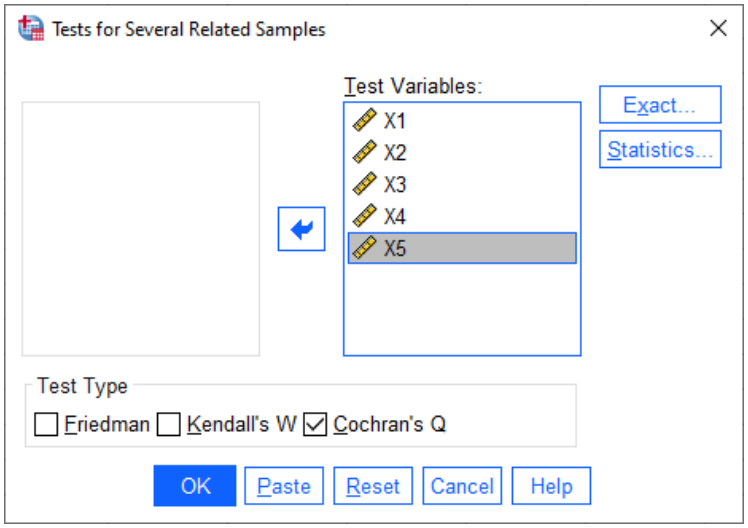
(με χρήση SPSS): Αρχικά, εισάγουμε στο SPSS τα αποτελέσματα (τιμές 0 ή 1) των 5 τεστ σε 5 στήλες, τις οποίες μετονομάζουμε σε X1, X2, X3, X4, X5.

Έπειτα, από το κεντρικό παράθυρο διαλόγου επιλέγουμε

Analyze / Nonparametric Tests / Legacy Dialogs / k Related Samples

Στο νέο παράθυρο διαλόγου, που ανοίγει, (βλ. Εικόνα 13.73) εισάγουμε τις στήλες X1-X5 στο πεδίο Test Variables και επιλέγουμε Cochran's Q στο πλαίσιο Test Type. Πατάμε OK και προκύπτει το output της ανάλυσης.

Από τα αποτελέσματα στην Εικόνα 13.74 έχουμε ότι η p-τιμή του ελέγχου είναι 0.085, δηλαδή δεν είναι μικρότερη από 0.01, και άρα, σε ε.σ. 1%, δεν απορρίπτουμε την υπόθεση ότι όλα τα τεστ έχουν την ίδια πιθανότητα επιτυχίας, δηλαδή την υπόθεση ότι είναι της ίδιας δυσκολίας.



Εικόνα 13.73: Παράθυρο διαλόγου του Tests for Several Related Samples του SPSS για τον έλεγχο Cochran's Q.

Test Statistics

N	9
Cochran's Q	8,190[a]
df	4
Asymp. Sig.	,085

a. 1 is treated as a success.

Εικόνα 13.74: Αποτελέσματα του ελέγχου Cochran's Q.

Επισημαίνεται ότι σε περίπτωση απόρριψης της μηδενικής υπόθεσης, θα πρέπει να συνεχίσουμε με τη διεξαγωγή πολλαπλών συγκρίσεων. Ένας τρόπος θα ήταν να χρησιμοποιήσουμε το τεστ του McNemar για να κάνουμε ελέγχους ποσοστών ανά ζεύγη, σε όλα τα δυνατά ζεύγη. Όμως, για να μπορέσουμε να διατηρήσουμε το ολικό επίπεδο σημαντικότητας στα επιθυμητά επίπεδα, θα πρέπει να χρησιμοποιήσουμε τη μέθοδο Bonferroni και να διορθώσουμε τις p-τιμές που προκύπτουν από την εφαρμογή του ελέγχου McNemar. Στο συγκεκριμένο παράδειγμα, απαιτείται η δειξαγωγή συνολικά 10 τέτοιων ελέγχων, οπότε, αν επιθυμούμε το ολικό επίπεδο σημαντικότητας να είναι π.χ. 1%, πρέπει αρχικά να πολλαπλασιάσουμε την p-τιμή κάθε ελέγχου με 10 και, στη συνέχεια, να συγκρίνουμε τη «διορθωμένη» p-τιμή με την τιμή 0.01. Επίσης, μπορούμε να χρησιμοποιήσουμε και τη διαδικασία του SPSS

Analyze / Nonparametric Tests / Related Samples

Θα πρέπει στην καρτέλα Settings, στο Select an item:Choose Tests να επιλέξουμε αρχικά Customize tests και μετά Cochran's Q (k samples) μαζί με το Multiple comparisons: All pairwise (βλ. Εικόνα 13.70).

(με χρήση R) Για να κάνουμε το τεστ Cochran's Q με χρήση της `R`, θα χρειαστεί, αρχικά, να φορτώσουμε τη βιβλιοθήκη `RVAideMemoire`. Στη συνέχεια, εισάγουμε τα δεδομένα ως εξής: αρχικά, εισάγουμε τις 45 το πλήθος διαθέσιμες τιμές σε ένα διάνυσμα (έστω αυτό `x`), εισάγοντας τις μετρήσεις ανά πειραματική μονάδα. Δηλαδή οι πρώτες 5 τιμές αφορούν τις επιδόσεις της πρώτης πειραματικής μονάδας στις 5 διαφορετικές αξιολογήσεις, οι επόμενες 5 τιμές αφορούν τις αντίστοιχες μετρήσεις για τη δεύτερη πειραματική μονάδα και ούτω καθεξής. Στο διάνυσμα `fact1` δηλώνουμε τους παράγοντες, δηλαδή ότι έχουμε 5 παράγοντες με μία μέτρηση σε κάθε πειραματική μονάδα σε σύνολο 45 πειραματικών μονάδων, ενώ στο `block1` δηλώνουμε τα blocks, δηλαδή ότι έχουμε 9 το πλήθος πειραματικές μονάδες σε καθέναν από τους 5 παράγοντες. Έπειτα, χρησιμοποιούμε την εντολή `cochran.qtest(...)` με τον τρόπο που φαίνεται παρακάτω.

```
> library(RVAideMemoire)
> response1 <-c(1,1,0,1,1,1,1,0,0,0,1,0,1,1,0,1,0,1,0,1,1,1,1,1,0,
+ 1,1,1,1,1,1,1,1,0,0,0,0,0,1,0,1,1,1,1,1,0,0,1)
> fact1 <- gl(5,1,45,labels=LETTERS[1:5])
> block1 <- gl(9,5,labels=letters[1:9])
> cochran.qtest(response1 ~ fact1|block1)
```

Από τα αποτελέσματα της ανάλυσης έχουμε ότι η τιμή της στατιστικής συνάρτησης Q, όπως και η p-τιμή, είναι ίσες με τις αντίστοιχες που προέκυψαν από το SPSS. Η H_0 απορρίπτεται σε ε.σ. 10%, αλλά όχι σε ε.σ. 5%. Το μικρότερο επίπεδο σημαντικότητας για το οποίο απορρίπτεται η H_0 είναι 8.48%. Τέλος, δίνονται και οι αναλογίες επιτυχημένων τεστ, σε καθεμία από τις 5 κατηγορίες τεστ.

```
        Cochran's Q test
data:  response1 by fact1, block = block1
Q = 8.1905, df = 4, p-value = 0.08484
alternative hypothesis: true difference in probabilities is not equal to 0
sample estimates:
proba in group A proba in group B proba in group C proba in group D
       0.6666667        0.7777778        0.3333333        0.4444444
proba in group E
       0.8888889
```

Πριν κλείσουμε την παρουσίαση της λύσης του συγκεκριμένου παραδείγματος με χρήση της `R`, αξίζει να αναφέρουμε ότι, αν επιθυμούμε να διεξάγουμε πολλαπλές συγκρίσεις, σε περίπτωση που η H_0 απορριφθεί, μπορούμε να το κάνουμε με χρήση της εντολής

```
cochran.qtest(...,alpha=...,p.method=...).
```

Η συγκεκριμένη εντολή εκτελεί συγκρίσεις ποσοστών ανά δύο με χρήση του προσημικού κριτηρίου (δηλαδή με το τεστ του McNemar) σε επίπεδο σημαντικότητας το οποίο καθορίζεται από την τιμή του ορίσματος `alpha`. Δίνοντας `p.method='bonferroni'` χρησιμοποιούμε τη μέθοδο Bonferroni για να διορθώσουμε τις p-τιμές των ελέγχων για τα ζεύγη συσχετισμένων ποσοστών. □

Παράδειγμα 13.20. (Έλεγχος Jonckheere-Terpstra): Στον Πίνακα 13.18 δίνονται δεδομένα τα οποία αφορούν τη θερμοκρασία σώματος 24 ενηλίκων, σε διάστημα 24 ωρών μετά τον εμβολιασμό τους έναντι μιας μεταδοτικής ασθένειας. Καθένας από τους 22 ενήλικες ταξινομήθηκε σε μια ηλικιακή ομάδα μεταξύ των 20-29, 30-39, 40-49 και 50-59. Να ελέγξετε σε ε.σ. 5% την υπόθεση ότι δεν υπάρχει διαφοροποίηση στη διάμεση θερμοκρασία σώματος μετά τον εμβολιασμό έναντι της εναλλακτικής ότι καθώς αυξάνεται η ηλικία αυξάνεται και η θερμοκρασία σώματος.

20-29	37.2	37.1	38.0	37.2	37.1		
30-39	37.5	37.4	37.9	38.3	38.6	38.4	
40-49	39.0	38.9	37.5	39.3	39.5	39.5	
50-59	38.2	37.8	37.9	38.1	37.7	37.9	37.2

Πίνακας 13.18: Δεδομένα θερμοκρασίας σώματος 24 ενηλίκων.

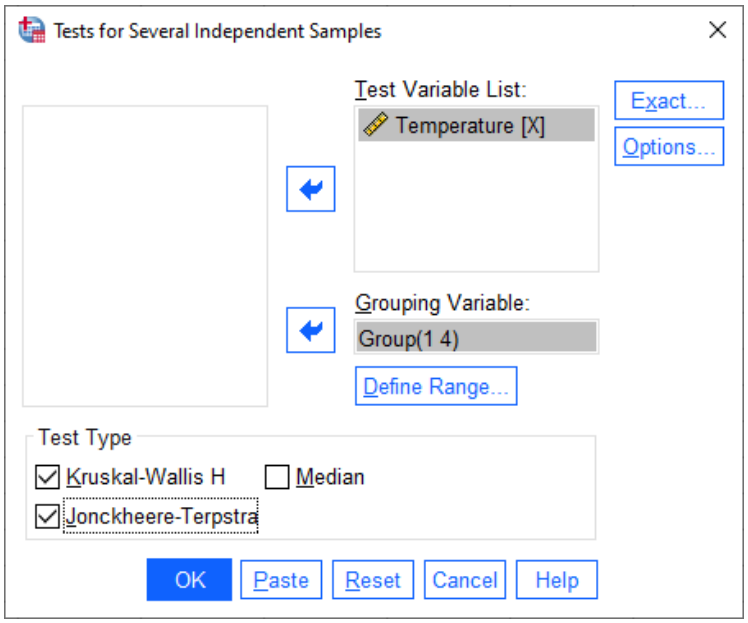
Λύση Παραδείγματος 13.20. Όπως έχει αναφερθεί στην Παρατήρηση 6.8, όταν θέλουμε να ελέγξουμε, χρησιμοποιώντας $k \geqslant 3$ (εδώ $k = 4$) το πλήθος ανεξάρτητα τυχαία δείγματα, την ισότητα των πληθυσμιακών διαμέσων έναντι της εναλλακτικής ότι υπάρχει μια δοθείσα διάταξη στις πληθυσμιακές διαμέσους είναι προτιμότερο αντί του ελέγχου των Kruskal–Wallis να εφαρμόζουμε τον έλεγχο των Jonckheere–Terpstra. Στη συνέχεια, θα δούμε την υλοποίηση του ελέγχου με χρήση SPSS και με χρήση `R`.

(με χρήση SPSS): Αρχικά εισάγουμε το αποτέλεσμα της θερμομέτρησης (μετονομάζουμε τη στήλη σε X) και σε μια διπλανή στήλη (την ονομάζουμε Group) εισάγουμε τις τιμές 1, 2, 3, 4 ανάλογα με την ηλικιακή ομάδα στην οποία ανήκει ο κάθε ενήλικας (1 για 20-29, 2 για 30-39, 3 για 40-49 και 4 για 50-59). Από το κεντρικό παράθυρο διαλόγου επιλέγουμε:

Analyze / Nonparametric Tests / Legacy Dialogs / k Independent Samples

Στο νέο παράθυρο διαλόγου, που ανοίγει, (βλ. Εικόνα 13.75) εισάγουμε στο Test Variable List τη στήλη X και στο Grouping Variable εισάγουμε τη στήλη Group. Πατάμε Define Range και δίνουμε τις τιμές 1 και 4 στα πεδία Minimum και Maximum, αντίστοιχα. Πατάμε Continue και στο Test Type επιλέγουμε Kruskal-Waliis και Jonckheere-Terpstra. Πατάμε OK και προκύπτει το output της ανάλυσης.

Αρχικά, στην Εικόνα 13.76 δίνονται οι μέσες τιμές των τάξεων για τις παρατηρήσεις που ανήκουν σε καθεμία από τις 4 ηλικιακές ομάδες. Στη συνέχεια, στην Εικόνα 13.77 έχουμε τα αποτελέσματα της εφαρμογής του τεστ των Kruskal-Wallis. Βλέπουμε ότι η μηδενική υπόθεση απορρίπτεται. Συμπεραίνουμε, επομένως, ότι η διάμεση θερμοκρασία σώματος μιας τουλάχιστον ηλικιακής ομάδας διαφέρει από την αντίστοιχη των υπόλοιπων ηλικιακών ομάδων. Από τον πίνακα με τις μέσες τιμές των τάξεων των παρατηρήσεων στις 4 ομάδες έχουμε ενδείξεις για διαφοροποίηση στη διάμεση θερμοκρασία. Αν θέλαμε να



Εικόνα 13.75: Παράθυρο διαλόγου του Tests for Several Independent Samples του SPSS για τους ελέγχους Kruskal-Wallis και Jonckheere-Terpstra.

εντοπίσουμε την ύπαρξη στατιστικά σημαντικά διαφορών (διεξάγοντας κατάλληλο στατιστικό έλεγχο), θα έπρεπε να προχωρήσουμε διεξάγοντας πολλαπλές συγκρίσεις. Οι διαδικασίες πολλαπλών συγκρίσεων σε μη παραμετρικούς ελέγχους με χρήση του SPSS έχουν παρουσιαστεί σε προηγούμενο παράδειγμα. Ως εκ τούτου, δεν θα τις περιγράψουμε εκ νέου και αφήνεται ως άσκηση για τον/την αναγνώστη/στρια η υλοποίησή τους και η ερμηνεία τους.

Ranks

	Group	N	Mean Rank
Temperature	1,00	5	5,00
	2,00	6	13,25
	3,00	6	19,58
	4,00	7	11,14
	Total	24	

Εικόνα 13.76: Μέσες τιμές των τάξεων (Mean Rank) των παρατηρήσεων στις 4 ομάδες.

Στην Εικόνα 13.79 έχουμε τα αποτελέσματα του τεστ των Jonckheere-Terpstra. Η p-τιμή του ελέγχου είναι 0.121 και άρα, σε ε.σ. 5%, δεν απορρίπτουμε την υπόθεση της ισότητας των πληθυσμιακών διαμέσων έναντι της εναλλακτικής ότι καθώς αυξάνεται η ηλικιακή ομάδα, ο πυρετός που εκδηλώνεται μετά τον εμβολιασμό είναι υψηλότερος.

(με χρήση R) Για να κάνουμε τον έλεγχο των Jonckheere-Terpstra με χρήση της `R`, θα πρέπει αρχικά να εγκαταστήσουμε το πακέτο `DescTools` και υλοποιούμε τον έλεγχο με τον τρόπο που φαίνεται παρακάτω.

```
> library(DescTools)
> x<-c(37.2,37.1,38.0,37.2,37.1,37.5,37.4,37.9,38.3,38.6,38.4,39.0,
+ 38.9,37.5,39.3,39.5,39.5,38.2,37.8,37.9,38.1,37.7,37.9,37.2)
> g2<-factor(c(rep('20-29',5),rep('30-39',6),rep('40-49',6),
+ rep('50-59',7)),ordered=TRUE)
> JonckheereTerpstraTest(x, g2, alternative ='increasing')
```

Test Statistics[a,b]

	Temperature
Kruskal-Wallis H	12,029
df	3
Asymp. Sig.	,007

a. Kruskal Wallis Test

b. Grouping Variable: Group

Εικόνα 13.77: Αποτέλεσμα του ελέγχου των Kruskal-Wallis στα δεδομένα του Παραδείγματος 13.20.

Pairwise Comparisons of Group

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
1,00-4,00	-6,143	4,130	-1,487	,137	,822
1,00-2,00	-8,250	4,271	-1,931	,053	,321
1,00-3,00	-14,583	4,271	-3,414	<,001	,004
4,00-2,00	2,107	3,925	,537	,591	1,000
4,00-3,00	8,440	3,925	2,151	,032	,189
2,00-3,00	-6,333	4,073	-1,555	,120	,720

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.

Asymptotic significances (2-sided tests) are displayed. The significance level is ,050.

a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

Εικόνα 13.78: Αποτέλεσμα πολλαπλών συγκρίσεων με χρήση του SPSS για το τεστ των Kruskal-Wallis στα δεδομένα του Παραδείγματος 13.20.

Jonckheere-Terpstra Test[a]

	Temperature
Number of Levels in Group	4
N	24
Observed J-T Statistic	137,500
Mean J-T Statistic	107,500
Std. Deviation of J-T Statistic	19,361
Std. J-T Statistic	1,550
Asymp. Sig. (2-tailed)	,121

a. Grouping Variable: Group

Εικόνα 13.79: Αποτελέσματα του ελέγχου των Jonckheere-Terpstra με το SPSS στα δεδομένα του Παραδείγματος 13.20.

Ειδικότερα, αρχικά εισάγουμε τις μετρήσεις σε ένα διάνυσμα, έστω αυτό `x`, δίνοντας πρώτα τις τιμές του 1ου τυχαίου δείγματος και, στο τέλος, τις τιμές του 5ου τυχαίου δείγματος. Στο διάνυσμα `g2` δηλώνουμε το όνομα κάθε επιπέδου του παράγοντα (εδώ κάθε ηλικιακής κατηγορίας) και το πλήθος των παρατηρήσεων που έχουμε σε καθένα τυχαίο δείγμα (εδώ 5, 6, 6, και 7), αντίστοιχα. Χρησιμοποιούμε την εντολή `factor(...)` με το όρισμα `ordered=TRUE`, ώστε να δημιουργήσουμε ένα διάνυσμα διατάξιμων κατηγοριών. Ο τρόπος που εισάγονται τα επίπεδα του παράγοντα είναι από τη μικρότερη ηλικιακή κατηγορία προς τη μεγαλύτερη. Τέλος, στην εντολή `JonckheereTerpstraTest` δηλώνουμε ότι η εναλλακτική υπόθεση είναι η ύπαρξη αύξουσας διάταξης στη θερμοκρασία σώματος ως προς τη διάταξη της ηλικιακής κατηγορίας (`alternative ='increasing'`).

Τα αποτελέσματα του τεστ των Jonckheere-Terpstra είναι τα ακόλουθα:

```
        Jonckheere-Terpstra test

data:  x and g2
JT = 137.5, p-value = 0.06114
alternative hypothesis: increasing

Warning message:
In JonckheereTerpstraTest.default(x, g2, alternative = 'increasing') :
  Sample size > 100 or data with ties
 p-value based on normal approximation. Specify nperm for permutation p-value
```

Άμεσα συμπεραίνουμε ότι δεν απορρίπτεται η H_0. Η τιμή της στατιστικής συνάρτησης του ελέγχου, καθώς και η p-τιμή είναι ακριβώς ίδιες με αυτές που δίνει το SPSS. Αξίζει να αναφέρουμε ότι στο SPSS δίνεται η p-τιμή για δίπλευρο έλεγχο με χρήση ασυμπτωτικού ελέγχου (z-test). Όμως, η `R` δίνει ως p-τιμή την τιμή 0.06114, η οποία αντιστοιχεί στο μισό της p-τιμής που δίνει το SPSS. Αυτό αιτιολογείται καθώς ο έλεγχος που ζητήσαμε να διεξάγει η `R` είναι μονόπλευρος. □

13.5 Έλεγχοι τυχαιότητας

Οι έλεγχοι τυχαιότητας, δηλαδή οι έλεγχοι της μηδενικής υπόθεσης ότι ένα σύνολο δεδομένων μπορεί να θεωρηθεί ότι είναι τυχαίο, αποτελούν ένα σημαντικό μέρος κάθε στατιστικής ανάλυσης και εκτενώς παρουσιάστηκαν στο Κεφάλαιο 7 του παρόντος συγγράμματος. Στην ενότητα αυτή, θα υλοποιηθεί με τη χρήση τόσο του SPSS όσο και της `R` ο έλεγχος των ροών που παρουσιάστηκε στην Ενότητα 7.2.

Παράδειγμα 13.21. (**Έλεγχος Τυχαιότητας με Χρήση Ροών**): Στη διάθεσή μας έχουμε τον ημερήσιο όγκο συναλλαγών (σε εκατομμύρια ευρώ) που διακινήθηκαν σε μεγάλο χρηματιστήριο της Ευρώπης. Θέλουμε να ελέγξουμε σε ε.σ. 5% την υπόθεση ότι οι παρακάτω τιμές αποτελούν τυχαίο δείγμα.

$$98,\ 93,\ 82,\ 103,\ 113,\ 111,\ 104,\ 103,\ 114,\ 107,\ 111,\ 109,\ 109,\ 108,\ 128,\ 92.$$

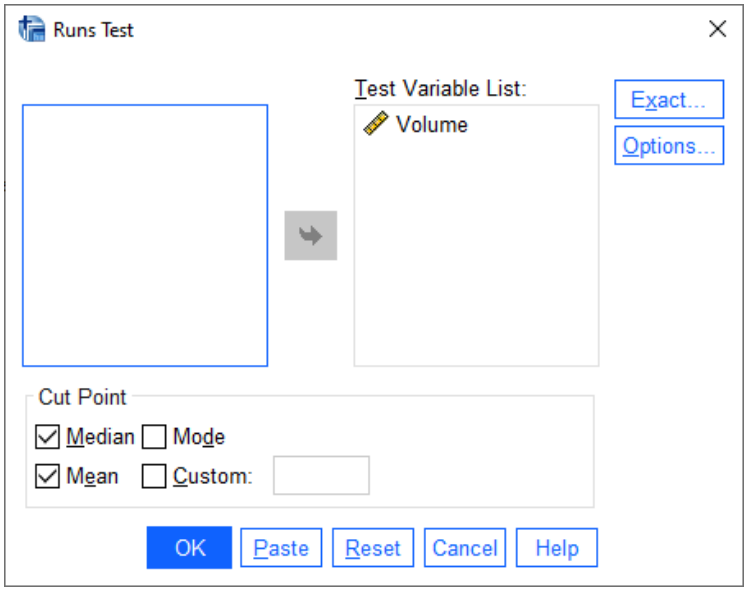
Λύση Παραδείγματος 13.21. Έχουμε $n = 16$ το πλήθος διαθέσιμες παρατηρήσεις, διατεταγμένες σε χρονολογική σειρά και θέλουμε να ελέγξουμε την υπόθεση ότι αποτελούν ένα τυχαίο δείγμα. Θα εφαρμόσουμε τον έλεγχο τυχαιότητας με χρήση ροών (Randomness test based on runs) (βλ. Ενότητα 7.2). Στο σημείο αυτό, απλώς υπενθυμίζουμε ότι ο έλεγχος αυτός βασίζεται στην κωδικοποίηση του αρχικού δείγματος σε μια ακολουθία συμβόλων με δύο διαφορετικά αποτελέσματα (δύο ομάδες αποτελεσμάτων) και, επομένως, απαιτείται ο χωρισμός των δεδομένων σε δύο ομάδες. Καθώς από τη φύση του

προβλήματος δεν καθορίζεται κάποιο κριτήριο, θα χρησιμοποιήσουμε τα συνηθέστερα σημεία διαχωρισμού που είναι η δειγματική διάμεσος και η δειγματική μέση τιμή των παρατηρήσεων. Μεταξύ των δύο προτείνεται η χρήση της μέσης τιμής, αν τα διαθέσιμα δεδομένα εμφανίζουν συμμετρία, ενώ, αν στα δεδομένα υπάρχουν ακραίες τιμές, τότε προτείνεται η χρήση της διαμέσου, διότι επηρεάζεται λιγότερο από την παρουσία ακραίων τιμών.

(**με χρήση SPSS**) Εισάγουμε τις ημερήσιες τιμές του όγκου συναλλαγών σε μια στήλη ενός κενού φύλλου εργασίας του SPSS, την οποία την μετονομάζουμε σε Volume. Έπειτα από το κεντρικό παράθυρο διαλόγου επιλέγουμε:

Analyze / Nonparametric Tests / Legacy Dialogs / Runs

Στο νέο παράθυρο διαλόγου που προκύπτει (βλ. Εικόνα 13.80) στο πλαίσιο Test Variable δίνουμε Volume, ενώ στο πλαίσιο Cut Point επιλέγουμε Median και Mean, ώστε να δούμε μήπως πάρουμε διαφορετικά αποτελέσματα, χρησιμοποιώντας τη διάμεσο και το μέσο του δείγματος ως σημεία καθορισμού των δύο διαφορετικών αποτελεσμάτων. Πατάμε OK και προκύπτει το output της ανάλυσης.



Εικόνα 13.80: Παράθυρο διαλόγου Runs Test.

Ειδικότερα, στην περίπτωση που χρησιμοποιηθεί ως σημείο διαχωρισμού η διάμεσος, το αποτέλεσμα του τεστ δίνεται στην Εικόνα 13.81. Η διάμεσος του δείγματος είναι ίση με 107.50 και υπάρχουν 8 τιμές κάτω από τη διάμεσο και άλλες 8 πάνω από αυτήν. Το πλήθος των ροών στο δείγμα είναι 7. Αν συμβολίσουμε με $+$ τις δειγματικές τιμές που είναι μεγαλύτερες της διαμέσου και με $-$ τις τιμές που είναι μικρότερες της διαμέσου, τότε οι αρχικές τιμές στο δείγμα κωδικοποιούνται στην παρακάτω ακολουθία συμβόλων $+$ και $-$:

$$- \; - \; - \; - \; + \; + \; - \; - \; + \; - \; + \; + \; + \; + \; + \; -$$

Η τιμή της στατιστικής συνάρτησης ελέγχου (ασυμπτωτικό τεστ) είναι ίση με $Z = -0.776$ και, αφού η p-τιμή είναι $0.438 > 0.05$, δεν απορρίπτουμε την υπόθεση της τυχαιότητας του δείγματος σε ε.σ. 5%.

Στην περίπτωση που χρησιμοποιηθεί ως σημείο διαχωρισμού η δειγματική μέση τιμή, το αποτέλεσμα του τεστ δίνεται στην Εικόνα 13.82. Η δειγματική μέση τιμή είναι ίση με 105.3125 και στο δείγμα υπάρχουν 7 τιμές μικρότερες από αυτήν και 9 τιμές μεγαλύτερες από αυτήν. Το πλήθος των ροών στο δείγμα είναι 5. Αν

Runs Test

	Volume
Test Value[a]	107,50
Cases < Test Value	8
Cases >= Test Value	8
Total Cases	16
Number of Runs	7
Z	-,776
Asymp. Sig. (2-tailed)	,438

a. Median

Εικόνα 13.81: Αποτελέσματα ελέγχου τυχαιότητας Runs Test - Σημείο διαχωρισμού η διάμεσος.

συμβολίσουμε με $+$ την περίπτωση μιας τιμής στο δείγμα να είναι μεγαλύτερη της δειγματικής μέσης τιμής και με $-$ την περίπτωση που η τιμή είναι μικρότερη της δειγματικής μέσης τιμής, οι αρχικές τιμές στο δείγμα κωδικοποιούνται στην παρακάτω ακολουθία συμβόλων:

$$- \; - \; - \; - \; + \; + \; - \; - \; + \; + \; + \; + \; + \; + \; + \; -$$

Η τιμή της στατιστικής συνάρτησης ελέγχου (ασυμπτωτικό τεστ) είναι ίση με $Z = -1.776$ και, αφού η p-τιμή είναι $0.076 > 0.05$, δεν απορρίπτουμε την υπόθεση της τυχαιότητας του δείγματος σε ε.σ. 5%.

Runs Test 2

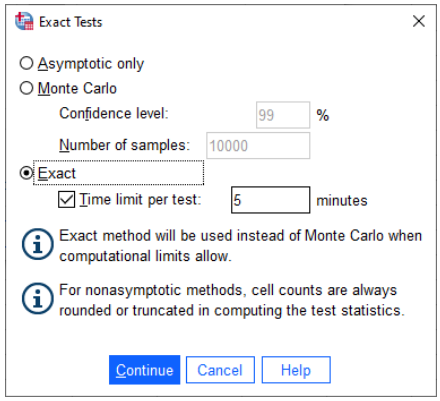
	Volume
Test Value[a]	105,3125
Cases < Test Value	7
Cases >= Test Value	9
Total Cases	16
Number of Runs	5
Z	-1,776
Asymp. Sig. (2-tailed)	,076

a. Mean

Εικόνα 13.82: Αποτελέσματα ελέγχου τυχαιότητας Runs Test - Σημείο διαχωρισμού η μέση τιμή.

Επειδή το πλήθος των αποτελεσμάτων από κάθε κατηγορία (δηλαδή το πλήθος των $+$ και $-$) δεν είναι μεγαλύτερο του 20, δεν συστήνεται η χρήση του ασυμπτωτικού ελέγχου. Για να κάνουμε τον ακριβή έλεγχο στο SPSS, επιλέγουμε Exact στο παράθυρο διαλόγου Runs Test (βλ. Εικόνα 13.80) και, στη συνέχεια, επιλέγουμε Exact (βλ. Εικόνα 13.83). Πατάμε Continue και μετά OK, οπότε προκύπτει το output της ανάλυσης (βλ. Εικόνες 13.84 και 13.85).

Από τα αποτελέσματα της ανάλυσης έχουμε ότι για την περίπτωση που το σημείο διαχωρισμού είναι η διάμεσος, η p-τιμή του ακριβούς ελέγχου είναι $0.429 > 0.05$ (γραμμή Asymp. Sig. (2-tailed)) και άρα, σε ε.σ. 5%, δεν απορρίπτουμε την υπόθεση της τυχαιότητας του δείγματος. Σε ανάλογα αποτελέσματα



Εικόνα 13.83: Επιλογή ακριβούς ελέγχου τυχαιότητας Runs Test.

Runs Test

	Volume
Test Value[a]	107,50
Cases < Test Value	8
Cases >= Test Value	8
Total Cases	16
Number of Runs	7
Z	-,776
Asymp. Sig. (2-tailed)	,438
Exact Sig. (2-tailed)	,429
Point Probability	,114

a. Median

Εικόνα 13.84: Αποτελέσματα ελέγχου τυχαιότητας Runs Test - Σημείο διαχωρισμού η διάμεσος. Ακριβής έλεγχος.

Runs Test 2

	Volume
Test Value[a]	105,3125
Cases < Test Value	7
Cases >= Test Value	9
Total Cases	16
Number of Runs	5
Z	-1,776
Asymp. Sig. (2-tailed)	,076
Exact Sig. (2-tailed)	,060
Point Probability	,025

a. Mean

Εικόνα 13.85: Αποτελέσματα ελέγχου τυχαιότητας Runs Test - Σημείο διαχωρισμού η μέση τιμή. Ακριβής έλεγχος.

καταλήγουμε και στην περίπτωση που το σημείο διαχωρισμού είναι η μέση τιμή. Σε εκείνη την περίπτωση, η p-τιμή είναι $0.06 > 0.05$.

(με χρήση R) Αρχικά, εισάγουμε τα δεδομένα σε ένα διάνυσμα, έστω αυτό `x`. Θα χρειαστούμε το πακέτο `randtests` το οποίο, όπως αναφέρθηκε στην εισαγωγή αυτής της ενότητας, μας δίνει τη δυνατότητα υλοποίησης των περισσότερων από τους ελέγχους που παρουσιάστηκαν στο Κεφάλαιο 7. Αφού φορτώσουμε το πακέτο, χρησιμοποιούμε την εντολή `runs.test(...)`, όπως φαίνεται παρακάτω:

```
> library(randtests)
> x<-c(98,93,82,103,113,111,104,103,114,107,111,109,109,108,128,92)
> # asymptotic test using median as test value
> runs.test(x,alternative='two.sided',threshold=median(x),
+ pvalue='normal')
> # asymptotic test using mean as test value
> runs.test(x,alternative='two.sided',threshold=mean(x),
+ pvalue='normal')
```

Ειδικότερα, στο όρισμα `alternative` μας δίνεται η δυνατότητα να δηλώσουμε αν θέλουμε αριστερόπλευρο (`'left.sided'`, έλεγχος τυχαιότητας έναντι της ύπαρξης τάσης), δεξιόπλευρο (`'right.sided'`, έλεγχος τυχαιότητας έναντι της ύπαρξης πρώτης τάξης αρνητικής συσχέτισης) ή δίπλευρο έλεγχο ('two.sided', έλεγχος τυχαιότητας έναντι της μη τυχαιότητας), με τον τελευταίο να αποτελεί και την προεπιλογή. Επίσης, στο όρισμα `threshold` δηλώνεται το σημείο διαχωρισμού που αναφέρθηκε παραπάνω, με την προεπιλογή να είναι η δειγματική διάμεσος. Εδώ, υλοποιούμε τον έλεγχο με σημείο διαχωρισμού τόσο τη δειγματική διάμεσο όσο και τη δειγματική μέση τιμή. Τέλος, στο όρισμα `pvalue` δηλώνεται αν θα χρησιμοποιηθεί η ακριβής κατανομή (`'exact'`) της στατιστικής συνάρτησης R ή η κανονική προσέγγισή της (`'normal'`). Η ακριβής κατανομή προτείνεται να χρησιμοποιείται για μικρές τιμές των n_1, n_2, συνηθέστερα $n_1, n_2 \leqslant 20$, με την προεπιλογή να είναι η κανονική προσέγγιση. Στη συνέχεια, για εκπαιδευτικούς λόγους, θα διεξάγουμε τον έλεγχο και με τους δύο τρόπους. Αρχικά, για τον ασυμπτωτικό έλεγχο, οι παραπάνω εντολές δίνουν τα ακόλουθα αποτελέσματα:

```
 Runs Test

data:  x
statistic = -1.0351, runs = 7, n1 = 8, n2 = 8, n = 16, p-value = 0.3006
alternative hypothesis: nonrandomness

        Runs Test

data:  x
statistic = -2.0397, runs = 5, n1 = 9, n2 = 7, n = 16, p-value =
0.04139
alternative hypothesis: nonrandomness
```

Από τα αποτελέσματα της ανάλυσης παρατηρούμε ότι οι p-τιμές των δύο ελέγχων (0.3006 και 0.04139) διαφέρουν από τις αντίστοιχες τιμές που δίνει το SPSS (0.438 και 0.076, αντίστοιχα). Αυτό οφείλεται στο γεγονός ότι το SPSS διεξάγει τον ασυμπτωτικό έλεγχο με χρήση διόρθωσης συνεχείας. Συμπεραίνουμε ότι, σε ε.σ. 5%, δεν απορρίπτεται η H_0 αν χρησιμοποιήσουμε τον έλεγχο με σημείο διαχωρισμού τη διάμεσο. Αντίθετα, στην περίπτωση που το σημείο διαχωρισμού είναι η μέση τιμή, η H_0 απορρίπτεται σε ε.σ. 5% (έστω και οριακά).

Παρακάτω δίνουμε τα αποτελέσματα της ανάλυσης χρησιμοποιώντας την ακριβή κατανομή του ελέγχου. Δεν θα επαναλάβουμε τις εντολές, αφού το μόνο που αλλάζει είναι η ανάθεση στο όρισμα `pvalue`, το οποίο όρισμα πλέον έχει την τιμή `'exact'`.

```
        Runs Test

data:  x
statistic = -1.0351, runs = 7, n1 = 8, n2 = 8, n = 16, p-value = 0.4289
alternative hypothesis: nonrandomness

        Runs Test

data:  x
statistic = -2.0397, runs = 5, n1 = 9, n2 = 7, n = 16, p-value = 0.06993
alternative hypothesis: nonrandomness
```

Από τα αποτελέσματα της ανάλυσης έχουμε ότι για την περίπτωση που το σημείο διαχωρισμού είναι η διάμεσος, η p-τιμή του ακριβούς ελέγχου είναι $0.4289 > 0.05$, ίδια με αυτήν που έδωσε και το SPSS. Άρα σε ε.σ. 5% δεν απορρίπτουμε την υπόθεση της τυχαιότητας του δείγματος. Σε ανάλογα αποτελέσματα καταλήγουμε και στην περίπτωση που το σημείο διαχωρισμού είναι η μέση τιμή. Σε εκείνη την περίπτωση, η p-τιμή είναι $0.06993 > 0.05$. Αξίζει να αναφέρουμε ότι η συγκεκριμένη τιμή διαφέρει ελαφρώς από την τιμή που έδωσε το SPSS, η οποία είναι 0.06014 (με ακρίβεια 5 δεκαδικών ψηφίων). Λαμβάνοντας υπόψη ότι τα $n_1, n_2 < 20$, συμπεραίνουμε, από τη χρήση του ακριβούς ελέγχου ότι, σε ε.σ. 5%, δεν μπορούμε να απορρίψουμε την υπόθεση της τυχαιότητας του συγκεκριμένου δείγματος. □

13.6 Έλεγχοι συσχέτισης δύο μεταβλητών

Στο Κεφάλαιο 8, μεταξύ άλλων, το ενδιαφέρον εστιάστηκε στη μελέτη μη παραμετρικών εκδοχών του συντελεστή συσχέτισης, όπως είναι οι εκδοχές που προτάθηκαν από τους Spearman και Kendall. Στην ενότητα αυτή το ενδιαφέρον επικεντρώνεται στην υλοποίηση των τεστ, που βασίζονται σε αυτούς τους συντελεστές και οι οποίοι έλεγχοι εκτενώς παρουσιάστηκαν στην Ενότητα 8.3 και στην Ενότητα 8.4. Τα τεστ αυτά χρησιμοποιούνται για τον έλεγχο της μηδενικής υπόθεσης $H_0 : \rho = 0$, δηλαδή της υπόθεσης ότι οι τυχαίες μεταβλητές X και Y είναι αμοιβαία ανεξάρτητες (mutually independent) έναντι μίας εκ των τριών εναλλακτικών:

(Α) $H_1 : \rho > 0$, δηλαδή της τάσης μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μεγάλες τιμές της άλλης και αντίστροφα (ύπαρξη θετικής τάσης/συσχέτισης),

(Β) $H_1 : \rho < 0$, δηλαδή της τάσης μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μικρές τιμές της άλλης και αντίστροφα (ύπαρξη αρνητικής τάσης/συσχέτισης),

(Γ) $H_1 : \rho \neq 0$, δηλαδή της τάσης είτε μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μεγάλες τιμές της άλλης και αντίστροφα είτε μεγάλες τιμές της μιας μεταβλητής να αντιστοιχούν σε μικρές τιμές της άλλης και αντίστροφα (ύπαρξη τάσης/συσχέτισης, είτε θετικής είτε αρνητικής).

Παράδειγμα 13.22. (**Συντελεστές Συσχέτισης Spearman και Kendall**): Σε ένα τυχαίο δείγμα 45 φοιτητών/τριών ενός Τμήματος Μαθηματικών, καταγράψαμε τις επιδόσεις τους στην πρόοδο και στην τελική εξέταση ενός μαθήματος του προπτυχιακού προγράμματος σπουδών. Τα δεδομένα δίνονται στον Πίνακα 13.19.

Α/Α	Πρόοδος	Τελική	Α/Α	Πρόοδος	Τελική	Α/Α	Πρόοδος	Τελική
1	5.5	6	16	2.5	5	31	7	8.5
2	4	3.5	17	1.5	3	32	9	8.5
3	2	3.5	18	1	6	33	8	8.5
4	2	1	19	8	5	34	5.5	6.5
5	1.5	2.5	20	9	9.5	35	5	3
6	6	7.5	21	0.5	3	36	3.5	3
7	5	5	22	4.5	6	37	2	4
8	5.5	7	23	4	5.5	38	1.5	1
9	7	5	24	5	6	39	3.5	5
10	6	5.5	25	3	3.5	40	6	5
11	4	5	26	6	5.5	41	7.5	7
12	1	3	27	2	4	42	5.5	9
13	2	2.5	28	2.5	4	43	3	5
14	0.5	0	29	3	5	44	2	2.5
15	6	3.5	30	1.5	2	45	2.5	4

Πίνακας 13.19: Αποτελέσματα επίδοσης 45 φοιτητών/τριών σε ενδιάμεση πρόοδο και τελική εξέταση.

(i) Να υπολογιστεί ο συντελεστής συσχέτισης του Spearman και να ερμηνεύσετε την τιμή του.
(ii) Να υπολογιστεί ο συντελεστής συσχέτισης του Kendall και να ερμηνεύσετε την τιμή του.
(iii) Χρησιμοποιώντας τον συντελεστή συσχέτισης του Spearman, να ελέγξετε την υπόθεση (ε.σ. $a = 0.01$) ότι δεν υπάρχει στατιστικά σημαντική συσχέτιση μεταξύ της επίδοσης στην πρόοδο και της επίδοσης στην τελική εξέταση του μαθήματος.
(iv) Να επαναλάβετε το προηγούμενο ερώτημα, χρησιμοποιώντας αυτήν τη φορά τον συντελεστή συσχέτισης του Kendall.

Λύση Παραδείγματος 13.22. Έχουμε ένα τυχαίο δείγμα $(X_1, Y_1), ..., (X_n, Y_n)$, από $n = 45$ ζεύγη παρατηρήσεων για τις τυχαίες μεταβλητές X: επίδοση στην ενδιάμεση πρόοδο και Y: επίδοση στην τελική εξέταση. Θα υπολογίσουμε τους μη παραμετρικούς συντελεστές συσχέτισης των Spearman και Kendall, και θα διεξάγουμε έλεγχο σε ε.σ. 1% της υπόθεσης ότι οι τυχαίες μεταβλητές X και Y είναι αμοιβαία ανεξάρτητες (mutually independent), έναντι της εναλλακτικής της ύπαρξης τάσης, είτε θετικής είτε αρνητικής.

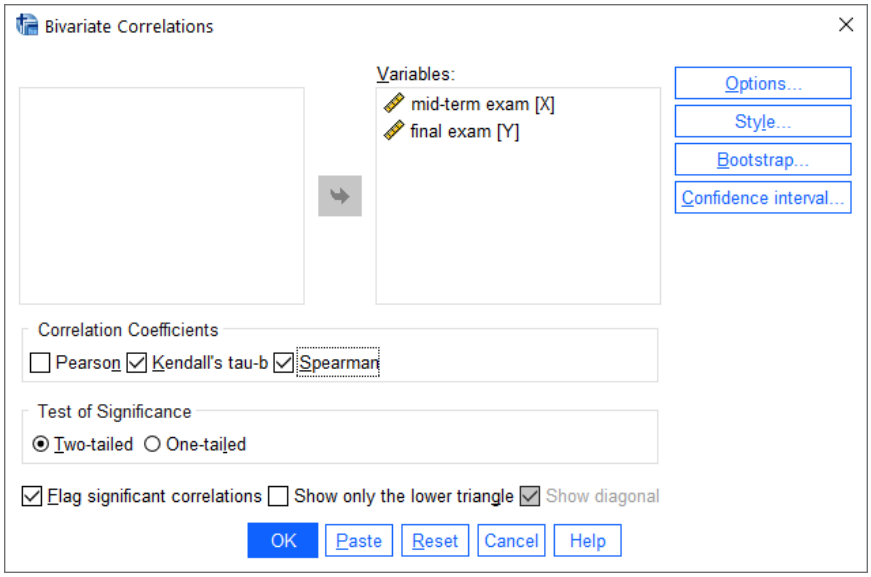
(**με χρήση SPSS**): Αρχικά, εισάγουμε τα δεδομένα από τις στήλες Πρόοδος και Τελική σε δύο στήλες ενός κενού φύλλου εργασίας του SPSS και τις μετονομάζουμε σε X και Y, αντίστοιχα.

Στη συνέχεια, από το κεντρικό παράθυρο διαλόγου επιλέγουμε:

Analyze / Correlate / Bivariate

Στο νέο παράθυρο διαλόγου που ανοίγει (βλ. Εικόνα 13.86), εισάγουμε στο πλαίσιο Variables τις στήλες X και Y. Στο πλαίσιο Correlation Coefficients επιλέγουμε Spearman (για υπολογισμό του συντελεστή συσχέτισης του Spearman) και Kendall's tau-b (για υπολογισμό του συντελεστή συσχέτισης του Kendall). Επιπρόσθετα, στο πεδίο Test of Significance, επιλέγουμε Two-tailed, αν θέλουμε να κάνουμε δίπλευρο έλεγχο. Σημειώνουμε ότι αν επιθυμούμε να ελέγξουμε την ύπαρξη συσχέτισης μεταξύ των X, Y μιας συγκεκριμένης κατεύθυνσης, δηλαδή είτε θετικής είτε αρνητικής, πρέπει να επιλέξουμε One-tailed. Στο συγκεκριμένο παράδειγμα θα επιλέξουμε Two-tailed. Πατάμε OK και προκύπτει το output της ανάλυσης, που δίνεται στην Εικόνα 13.87.

Από το output της ανάλυσης έχουμε ότι ο συντελεστής συσχέτισης του Spearman ισούται με $r_S = 0.754$, το οποίο αποτελεί ένδειξη ισχυρής θετικής συσχέτισης μεταξύ του βαθμού στην ενδιάμεση πρόοδο και στην



Εικόνα 13.86: Παράθυρο διαλόγου Bivariate Correlations.

Correlations

			mid-term exam	final exam
Kendall's tau_b	mid-term exam	Correlation Coefficient	1,000	,613**
		Sig. (2-tailed)	.	<,001
		N	45	45
	final exam	Correlation Coefficient	,613**	1,000
		Sig. (2-tailed)	<,001	.
		N	45	45
Spearman's rho	mid-term exam	Correlation Coefficient	1,000	,754**
		Sig. (2-tailed)	.	<,001
		N	45	45
	final exam	Correlation Coefficient	,754**	1,000
		Sig. (2-tailed)	<,001	.
		N	45	45

**. Correlation is significant at the 0.01 level (2-tailed).

Εικόνα 13.87: Τιμές συντελεστών συσχέτισης Spearman και Kendall για τα δεδομένα του Πίνακα 13.19.

τελική εξέταση. Μπορούμε να συμπεράνουμε ότι όσοι έχουν υψηλούς βαθμούς στην πρόοδο έχουν και υψηλό βαθμό στην τελική εξέταση. Η τιμή του συντελεστή συσχέτισης του Kendall είναι $\tau = 0.613$ και η ερμηνεία του είναι αντίστοιχη με αυτήν του συντελεστή του Spearman.

Επίσης, από τη γραμμή Sig. (2-tailed) συμπεραίνουμε ότι η p-τιμή για τον έλεγχο της μηδενικής υπόθεσης είναι μικρότερη από 0.001 (είτε με χρήση του συντελεστή Spearman είτε με χρήση του συντελεστή Kendall). Άρα συμπεραίνουμε, σε ε.σ. 1% ότι οι επιδόσεις στην ενδιάμεση πρόοδο και στην τελική εξέταση είναι συσχετισμένες. Μάλιστα, αφού η τιμή του συντελεστή συσχέτισης είναι θετική, μπορούμε να ισχυριστούμε ότι, σε ε.σ. 5%, υπάρχει στατιστικά σημαντική θετική συσχέτιση μεταξύ του βαθμού στην ενδιάμεση πρόοδο και του βαθμού στην τελική εξέταση του μαθήματος.

(με χρήση R) Αρχικά, εισάγουμε τα δεδομένα σε δύο διανύσματα, έστω αυτά `x` (για τα δεδομένα της Προόδου) και `y` (για τα δεδομένα της Τελικής εξέτασης). Στη συνέχεια, χρησιμοποιούμε την εντολή `cor.test(...)` εισάγοντας τα διανύσματα `x` και `y`. Με το όρισμα `method` μπορούμε να ζητήσουμε τον υπολογισμό του συντελεστή συσχέτισης που επιθυμούμε. Έτσι, δίνοντας `method='spearman'` η `R` επιστρέφει την τιμή του συντελεστή συσχέτισης Spearman, ενώ με `method='kendall'` η `R` επιστρέφει την τιμή του συντελεστή συσχέτισης Kendall. Επίσης, με το όρισμα `alternative` μπορούμε να επιλέξουμε να κάνουμε είτε δίπλευρο είτε μονόπλευρο έλεγχο. Δίνοντας `alternative='two.sided'` διεξάγουμε τον δίπλευρο έλεγχο, ενώ αν δηλώσουμε `alternative='greater'` (`alternative='less'`) μπορούμε να κάνουμε τον έλεγχο για θετική συσχέτιση (αντίστοιχα αρνητική συσχέτιση) μεταξύ των χαρακτηριστικών. Τέλος, με το όρισμα `exact` μπορούμε να επιλέξουμε αν θα κάνουμε ασυμπτωτικό έλεγχο (για `exact='FALSE'`) ή τον ακριβή έλεγχο (για `exact='TRUE'`). Σε αυτό το παράδειγμα, επιλέγουμε να κάνουμε τον ασυμπτωτικό έλεγχο, αφού το διαθέσιμο δείγμα μπορεί να θεωρηθεί μεγάλο. Παρακάτω δίνονται οι σχετικές εντολές και το αποτέλεσμα της ανάλυσης.

```
> x<-c(5.5,4,2,2,1.5,6,5,5.5,7,6,4,1,2,0.5,6,
+ 2.5,1.5,1,8,9,0.5,4.5,4,5,3,6,2,2.5,3,1.5,
+ 7,9,8,5.5,5,3.5,2,1.5,3.5,6,7.5,5.5,3,2,2.5)
> y<-c(6,3.5,3.5,1,2.5,7.5,5,7,5,5.5,5,3,2.5,0,3.5,
+ 5,3,6,5,9.5,3,6,5.5,6,3.5,5.5,4,4,5,2,
+ 8.5,8.5,8.5,6.5,3,3,4,1,5,5,7,9,5,2.5,4)
> cor.test(x,y,method='spearman',alternative = 'two.sided',
+ exact = FALSE)
> cor.test(x,y,method='kendall',alternative = 'two.sided',
+ exact = FALSE)
```

```
        Spearman's rank correlation rho

data:  x and y
S = 3741.1, p-value = 2.317e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7535521

        Kendall's rank correlation tau

data:  x and y
z = 5.6076, p-value = 2.051e-08
alternative hypothesis: true tau is not equal to 0
sample estimates:
```

```
      tau
0.6130088
```

Άμεσα διαπιστώνουμε ότι οι τιμές των συντελεστών συσχέτισης είναι οι ίδιες με αυτές που έδωσε το SPSS (τιμή για τον συντελεστή συσχέτισης Spearman ίση με $r_s = 0.7535521$ και η τιμή για τον συντελεστή συσχέτισης Kendall ίση με $\tau = 0.6130088$). Αξίζει να αναφέρουμε ότι η τιμή S στο αποτέλεσμα του τεστ με χρήση του συντελεστή συσχέτισης Spearman είναι η τιμή της στατιστικής συνάρτησης

$$T = \sum_{i=1}^{n} (R_i - S_i)^2,$$

η οποία στη βιβλιογραφία ονομάζεται στατιστική συνάρτηση των Hotelling-Pabst (βλ. Hotelling and Pabst, 1936) και ποσοστιαία σημεία της είναι διαθέσιμα στη βιβλιογραφία (βλ. Conover, 1998)). Για περισσότερες λεπτομέρειες παραπέμπουμε στην Παρατήρηση 8.4.

Επίσης, για τον έλεγχο με χρήση του συντελεστή συσχέτισης Kendall, η τιμή z δεν έχει υπολογιστεί μέσω της σχέσης (8.4), αλλά καθώς υπάρχει σημαντικός αριθμός ισοβαθμιών (ties) στο δείγμα έχουν εφαρμοστεί όσα έχουν αναφερθεί στην Παρατήρηση 8.9.

Από τις p-τιμές των ελέγχων έπεται ότι η μηδενική υπόθεση περί αμοιβαία ανεξάρτητων μεταβλητών X, Y απορρίπτεται σε ε.σ. 1%. Άρα οι επιδόσεις στην ενδιάμεση πρόοδο και στην τελική εξέταση είναι συσχετισμένες. Μάλιστα, αφού η τιμή του συντελεστή συσχέτισης είναι θετική, μπορούμε να ισχυριστούμε ότι υπάρχει στατιστικά σημαντική θετική συσχέτιση μεταξύ του βαθμού στην ενδιάμεση πρόοδο και του βαθμού στην τελική εξέταση του μαθήματος. □

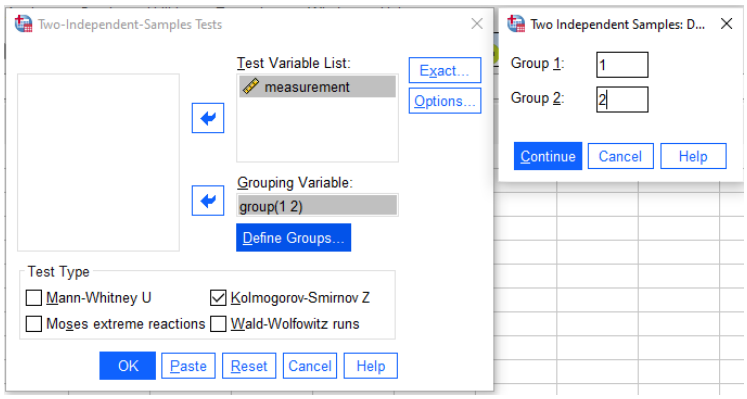
13.7 Ο έλεγχος Smirnov και ο έλεγχος ροών για σύγκριση δύο κατανομών

Έστω $X_1, X_2, \ldots, X_n$, ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $F_X(\cdot)$ και $Y_1, Y_2, \ldots, Y_m$, ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $G_Y(\cdot)$. Υποθέτουμε ότι τα δύο δείγματα είναι ανεξάρτητα. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση $H_0 : F_X(x) = G_Y(x)$, $\forall\, x \in \mathbb{R}$, έναντι της εναλλακτικής $H_1 : F_X(x) \neq G_Y(x)$, για κάποιο $x \in \mathbb{R}$. Πλην των ελέγχων που βασίζονται στις τάξεις και παρουσιάστηκαν στο Κεφάλαιο 6 (βλ. υλοποίηση στην Ενότητα 13.4.1) στο Κεφάλαιο 4 είδαμε ότι ένας τρόπος διενέργειας του παραπάνω ελέγχου είναι η γενίκευση του ελέγχου του Kolmogorov που προτάθηκε από τον Smirnov (βλ. Ενότητα 4.3.1). Επιπρόσθετα, στο Κεφάλαιο 7, και ειδικότερα στην Ενότητα 7.2 αναφέρθηκε ότι το τεστ των Wald-Wolfowitz (βλ. Wald and Wolfowitz, 1940) μπορεί, επίσης, να χρησιμοποιηθεί για τον έλεγχο της υπόθεσης ότι δύο ανεξάρτητα τυχαία δείγματα προέρχονται από τον ίδιο πληθυσμό. Καθώς όμως ο έλεγχος αυτός έχει πολύ μικρή ισχύ, δεν παρουσιάστηκε εκτενώς (βλ. και Παρατήρηση 7.6). Ωστόσο, στην ενότητα αυτή θα δούμε πώς υλοποιούνται οι παραπάνω έλεγχοι με χρήση του SPSS και της R.

Παράδειγμα 13.23. (**Έλεγχος 2-sample Kolmogorov-Smirnov**): Χρησιμοποιήστε τα δεδομένα του Πίνακα 13.16 και με τη χρήση του 2-sample Kolmogorov-Smirnov test ελέγξτε σε ε.σ. 5% την υπόθεση ότι οι πληθυσμοί από τους οποίους προέρχονται τα δύο τυχαία δείγματα $X_1, X_2, ..., X_{40}$, $Y_1, Y_2, ..., Y_{40}$ μπορούν να θεωρηθούν ταυτόσημοι.

Λύση Παραδείγματος 13.23. (με χρήση SPSS) Εισάγουμε σε μια στήλη όλες τις μετρήσεις (στήλες X και Y) (την ονομάζουμε measurement) και σε διπλανή στήλη, την οποία ονομάζουμε group, εισάγουμε τις τιμές 1 (για τα δεδομένα από τη X, γραμμές 1-40) και 2 (για τα δεδομένα από την Y, γραμμές 41-80).

Στη συνέχεια, από το κεντρικό παράθυρο διαλόγου επιλέγουμε



Εικόνα 13.88: Παράθυρο Διαλόγου για την Εφαρμογή του ελέγχου 2 sample Kolmogorov-Smirnov.

Analyze / Nonparametric Tests / Legacy Dialogs / 2 Independent Samples.

Στο νέο παράθυρο διαλόγου που προκύπτει (βλ. Εικόνα 13.88) εισάγουμε τη στήλη measurement στο πεδίο Test Variable List και τη στήλη group στο πεδίο Grouping Variable. Πατάμε το Define Groups και στο παράθυρο διαλόγου, που ανοίγει, δίνουμε στο Group 1 την τιμή 1, στο Group 2 την τιμή 2 και πατάμε Continue. Τέλος, επιλέγουμε Kolmogorov-Smirnov Z, αντί Mann-Whitney U (που είναι η προεπιλογή) και πατάμε OK.

Το output της ανάλυσης δίνεται στην Εικόνα 13.89. Στις γραμμές Most Extreme Differences έχουμε τις τιμές των διαφορών D_n^+, D_n^- και D_n, ενώ η τιμή της στατιστικής συνάρτησης ελέγχου (για το ασυμπτωτικό τεστ) βρίσκεται στη γραμμή Kolmogorov-Smirnov Z. Η p-τιμή είναι ίση με $0.015 < 0.05$ και σε ε.σ. 5% απορρίπτεται η H_0, δηλαδή σε ε.σ. 5% δεν μπορούμε να θεωρήσουμε ότι τα δύο τυχαία δείγματα προέρχονται από τον ίδιο πληθυσμό.

Test Statistics[a]

		measurement
Most Extreme Differences	Absolute	,350
	Positive	,275
	Negative	-,350
Kolmogorov-Smirnov Z		1,565
Asymp. Sig. (2-tailed)		,015

a. Grouping Variable: group

Εικόνα 13.89: Αποτελέσματα ελέγχου Two Sample Kolmogorov-Smirnov.

(με χρήση R) Αρχικά, εισάγουμε τα δεδομένα σε δύο διανύσματα, έστω αυτά `x` και `y`. Χρησιμοποιούμε την εντολή `ks.test(...)` και εισάγουμε ως όρισμα τα διανύσματα με τις διαθέσιμες μετρήσεις. Με το όρισμα `alternative = 'two.sided'` επιλέγουμε να κάνουμε τον έλεγχο με δίπλευρη εναλλακτική. Αν θέλουμε να κάνουμε κάποιον από τους μονόπλευρους ελέγχους, δίνουμε `alternative = 'less'` ή `alternative = 'greater'`, οπότε οι αντίστοιχες εναλλακτικές είναι $H_1 : F(x) < G(x)$ ή $H_1 : F(x) > G(x)$, αντίστοιχα. Οι σχετικές εντολές και τα αποτελέσματα του ελέγχου δίνονται παρακάτω:

```
> x<-c(20.75,23.20,18.64,20.22,17.77,19.59,22.02,20.38,20.97,
+ 20.83,20.08,19.07,22.31,19.09,20.11,21.14,15.71,22.70,19.85,
+ 20.91,18.78,19.01,23.03,16.20,19.94,18.56,23.09,19.89,21.49,
+ 20.23,20.98,18.02,17.85,19.85,19.94,18.63,24.33,18.04,20.77,21.30)
```

```
> y<-c(22.16,38.08,12.64,19.43,9.70,16.48,29.78,20.27,23.39,
+ 22.63,18.75,14.31,31.71,14.38,18.90,24.33,4.74,34.40,17.68,
+ 23.05,13.18,14.08,36.83,5.70,18.10,12.37,37.29,17.87,26.41,
+ 19.50,23.46,10.49,9.98,17.67,18.08,12.60,47.42,10.57,22.32,25.28)
> ks.test(x,y,alternative='two.sided')

        Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.35, p-value = 0.01489
alternative hypothesis: two-sided

Warning message:
In ks.test(x, y, alternative = 'two.sided') :
  cannot compute exact p-value with ties
```

Από τα αποτελέσματα της ανάλυσης διαπιστώνουμε ότι η τιμή της σ.σ.ε. είναι ίση με $D_n = 0.35$, δηλαδή ίση με την τιμή που έδωσε το SPSS στη γραμμή Most Extreme Differences: Absolute. Αυτό συμβαίνει καθώς και τα δύο πακέτα χρησιμοποιούν την ασυμπτωτική κατανομή του ελέγχου. Αφού η p-τιμή είναι ίση με $0.01489 < 0.05$ (ίση με αυτήν που δίνει το SPSS), συμπεραίνουμε ότι απορρίπτεται η H_0 και άρα δεν μπορούμε να θεωρήσουμε ότι οι τιμές X_i, Y_i προέρχονται από τον ίδιο πληθυσμό. □

Παράδειγμα 13.24. (**Έλεγχος Wald-Wolfowitz**): Χρησιμοποιήστε τα δεδομένα του Πίνακα 13.16 και με τη χρήση του Wald-Wolfowitz Runs test ελέγξτε σε ε.σ. 5% την υπόθεση ότι οι πληθυσμοί από τους οποίους προέρχονται τα δύο τυχαία δείγματα $X_1, X_2, ..., X_{40}$, $Y_1, Y_2, ..., Y_{40}$ μπορούν να θεωρηθούν ταυτόσημοι.

Λύση Παραδείγματος 13.24. Θα εφαρμόσουμε τον έλεγχο των Wald-Wolfowitz για δύο ανεξάρτητα δείγματα, για να ελέγξουμε την υπόθεση ότι αυτά προέρχονται από τον ίδιο πληθυσμό. Η μηδενική υπόθεση του ελέγχου είναι $H_0 : F_X(x) = G_Y(x)$ για κάθε $x \in \mathbb{R}$ έναντι της $H_1 : F_X(x) \neq G_Y(x)$ για τουλάχιστον ένα $x \in \mathbb{R}$. Οι $F_X(\cdot)$, $G_Y(\cdot)$ είναι οι α.σ.κ. των δύο ανεξάρτητων πληθυσμών.

(με χρήση SPSS) Εισάγουμε τα δεδομένα, όπως και στην περίπτωση εφαρμογής των τεστ των Mann-Whitney και Two sample Kolmogorov-Smirnov. Στη συνέχεια, από το κεντρικό παράθυρο διαλόγου επιλέγουμε

Analyze / Nonparametric Tests / Legacy Dialogs / 2 Independent Samples

Στο νέο παράθυρο διαλόγου, που προκύπτει, εισάγουμε τη στήλη measurement) στο πεδίο Test Variable List και τη στήλη group στο πεδίο Grouping Variable. Πατάμε το Define Groups και στο παράθυρο διαλόγου, που ανοίγει, δίνουμε στο Group 1 την τιμή 1, στο Group 2 την τιμή 2 και πατάμε Continue. Επιλέγουμε Wald-Wolfowitz, αντί Kolmogorov-Smirnov Z που είχαμε επιλέξει στο προηγούμενο παράδειγμα και πατάμε OK.

Το output της ανάλυσης δίνεται στην Εικόνα 13.90. Στις γραμμές Minimum Possible και Maximum Possible έχουμε το ελάχιστο και το μέγιστο πλήθος ροών που μπορούν να σχηματιστούν κατά την ένωση των τιμών από τα δύο δείγματα σε ένα κοινό δείγμα, ενώ προκύπτει ότι, και στις δύο περιπτώσεις, η p-τιμή (στήλη Asymp. Sig) είναι μικρότερη από 0.001. Ο έλεγχος γίνεται με χρήση του ασυμπτωτικού τεστ και η τιμή της στατιστικής συνάρτησης ελέγχου Z δίνεται στην αντίστοιχη στήλη. Άρα, σε ε.σ. 5%, απορρίπτεται η H_0 και δεν μπορούμε να θεωρήσουμε ότι τα δύο τυχαία δείγματα προέρχονται από τον ίδιο πληθυσμό.

Test Statistics[a,b]

		Number of Runs	Z	Asymp. Sig. (1-tailed)
measurement	Minimum Possible	25[c]	-3,601	<,001
	Maximum Possible	25[c]	-3,601	<,001

a. Wald-Wolfowitz Test

b. Grouping Variable: group

c. There are 1 inter-group ties involving 2 cases.

Εικόνα 13.90: Αποτελέσματα ελέγχου Wald-Wolfowitz.

(με χρήση R) Αρχικά, εισάγουμε τα δεδομένα σε δύο διανύσματα, έστω αυτά `x` και `y`. Η `R` δεν δίνει άμεσα το τεστ των Wald-Wolfowitz για τον έλεγχο της ισότητας των δύο κατανομών. Για τον λόγο αυτό, δίνουμε και τις σχετικές εντολές οι οποίες θα διεξάγουν το συγκεκριμένο τεστ, όπως φαίνεται παρακάτω. Επίσης, για μεγαλύτερη ευκολία, θα φορτώσουμε και το πακέτο `randtests`.

```
> library(randtests)
> x<-c(20.75,23.20,18.64,20.22,17.77,19.59,22.02,20.38,20.97,20.83,
+ 20.08,19.07,22.31,19.09,20.11,21.14,15.71,22.70,19.85,20.91,18.78,
+ 19.01,23.03,16.20,19.94,18.56,23.09,19.89,21.49,20.23,20.98,18.02,
+ 17.85,19.85,19.94,18.63,24.33,18.04,20.77,21.30)
> y<-c(22.16,38.08,12.64,19.43,9.70,16.48,29.78,20.27,23.39,22.63,
+ 18.75,14.31,31.71,14.38,18.90,24.33,4.74,34.40,17.68,23.05,13.18,
+ 14.08,36.83,5.70,18.10,12.37,37.29,17.87,26.41,19.50,23.46,10.49,
+ 9.98,17.67,18.08,12.60,47.42,10.57,22.32,25.28)
> ### R code for the wald-wolfowitz test
> vec1<-c(x,y) # create the joint sample
> vec2<-c(rep(1,length(x)),rep(2,length(y))) # grouping variable
> # order of every observation in the joint sample
> id1<-order(vec1)
> runs.test(vec2[id1])
```

Ειδικότερα, αφού έχουμε εισάγει τα δεδομένα στα διανύσματα `x` και `y`, φτιάχνουμε το κοινό δείγμα (διάνυσμα `vec1`), ενώ φτιάχνουμε και τη μεταβλητή ομαδοποίησης (grouping variable), η οποία δείχνει σε ποιο δείγμα ανήκει η κάθε παρατήρηση. Αυτό επιτυγχάνεται με τη δημιουργία ενός διανύσματος (το `vec2`) στο οποίο το πλήθος των 1 ισούται με το πλήθος των τιμών που δηλώθηκαν στο διάνυσμα `x` ενώ, τα υπόλοιπα στοιχεία του `vec2` είναι ίσα με 2. Το πλήθος των 2 ισούται με το πλήθος των τιμών που δηλώθηκαν στο διάνυσμα `y`. Στη συνέχεια, με την εντολή `order` καταχωρίζουμε στο διάνυσμα `id1` τη διάταξη της κάθε παρατήρησης στο κοινό δείγμα και χρησιμοποιούμε τη συνάρτηση `runs.test` η οποία εκτελεί τον έλεγχο τυχαιότητας των Wald-Wolfowitz στο διάνυσμα `vec2[id1]`. Το διάνυσμα αυτό περιέχει δύο δυνατά αποτελέσματα (τιμές 1 ή 2) και ουσιαστικά το αρχικό πρόβλημα του ελέγχου της ισότητας των δύο κατανομών μετατρέπεται σε έλεγχο τυχαιότητας (βλ. υλοποίηση ελέγχων τυχαιότητας με χρήση της `R` στην Ενότητα 13.5).

Παρακάτω, δίνονται τα αποτελέσματα της ανάλυσης, από τα οποία διαπιστώνουμε ότι η τιμή της σ.σ.ε. είναι ίση με -3.6006, ίση με αυτήν που έδωσε ο έλεγχος με χρήση του SPSS. Η p-τιμή είναι $0.0003175 < 0.05$ και άρα, σε ε.σ. 5%, απορρίπτουμε την υπόθεση της ισότητας των δύο κατανομών. □

```
        Runs Test

data:  vec2[id1]
statistic = -3.6006, runs = 25, n1 = 40, n2 = 40, n = 80, p-value = 0.0003175
alternative hypothesis: nonrandomness
```

13.8 Ασκήσεις

Οι παρακάτω ασκήσεις να λυθούν με χρήση SPSS αλλά και με χρήση R. Σε όλες τις ασκήσεις να χρησιμοποιήσετε ε.σ. 5%, εκτός αν αναφέρεται διαφορετικά.

Άσκηση 13.1. Χρησιμοποιήστε τον χι-τετράγωνο έλεγχο καλής προσαρμογής και ελέγξτε την υπόθεση ότι τα παρακάτω δεδομένα προέρχονται από πληθυσμό που μοντελοποιείται σύμφωνα με την Υπεργεωμετρική κατανομή $Hg(10,10,7)$. Τα δεδομένα δίνονται σε μορφή πίνακα συχνοτήτων και αφορούν συνολικά 150 πραγματοποιήσεις επιθεωρήσεων $n = 7$ το πλήθος εξαρτήματα, τα οποία λαμβάνονται, χωρίς επανατοποθέτηση, από έναν σωρό 20 συνολικά εξαρτημάτων. Σε κάθε επιθεώρηση καταγράφεται το πλήθος των ελαττωματικών εξαρτημάτων, μεταξύ των 7 του δείγματος, και παρατίθεται στον πίνακα που ακολουθεί.

Αριθμός Ελαττωματικών	Πλήθος Επιθεωρήσεων
0	0
1	16
2	38
3	56
4	33
5	7
6	0
7	0

Άσκηση 13.2. Χρησιμοποιήστε το κριτήριο Kolmogorov-Smirnov και ελέγξετε την υπόθεση ότι οι παρακάτω μετρήσεις προέρχονται από την κατανομή $\mathscr{B}(2,2)$ (κατανομή Βήτα).

0.621, 0.503, 0.203, 0.477, 0.710, 0.581, 0.329, 0.480, 0.554, 0.382.

Άσκηση 13.3. Να ελέγξετε την υπόθεση ότι το παρακάτω τυχαίο δείγμα 12 παρατηρήσεων (διατεταγμένων κατά αύξουσα σειρά) μπορεί να θεωρηθεί ότι αποτελεί δείγμα παρατηρήσεων πάνω σε μία τυχαία μεταβλητή X, της οποίας η κατανομή είναι κανονική με μέση τιμή 30 και διασπορά 100.

18.8 19.3 22.4 22.5 24.0 24.7 25.9 27.0 35.1 35.8 36.5 37.6

Για τον έλεγχο να χρησιμοποιηθούν τα κριτήρια Kolmogorov-Smirnov και Anderson-Darling.

Άσκηση 13.4. (Conover, 1998) Πενήντα διψήφιοι αριθμοί επελέγησαν τυχαία από έναν τηλεφωνικό κατάλογο. Οι αριθμοί κατά αύξουσα σειρά μεγέθους είναι οι εξής:

23 23 24 27 29 31 32 33 33 35 36 37 40 42 43 43 44 45 48 48 54 54 56 57 57

58 58 58 58 59 61 61 62 63 64 65 66 68 68 70 73 73 74 75 77 81 87 89 93 97

Να ελεγχθεί η υπόθεση ότι οι αριθμοί αυτοί θα μπορούσαν να αποτελούν παρατηρήσεις πάνω σε μια κανονική τυχαία μεταβλητή. Ο έλεγχος να γίνει με χρήση του τεστ των Shapiro-Wilk.

Άσκηση 13.5. Στη διάθεσή μας έχουμε τα παρακάτω τυχαία δείγματα τα οποία αφορούν το βάρος μπαταριών τύπου LR6, οι οποίες έχουν παραχθεί από 2 διαφορετικές εταιρείες (υποθέστε ότι οι παραγωγικές διαδικασίες για την κατασκευή των μπαταριών των 2 εταιρειών, είναι μεταξύ τους ανεξάρτητες).

ΑΑ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X	2.4	2.6	2.9	3.2	3.3	3.4	3.7	3.8	4.0	4.2	4.8	4.9	5.3	6.7	7.6	7.8
Y	1.1	1.8	2.4	2.4	2.6	2.9	3.0	3.2	5.5	6.4	6.7	7.4	8.1	11.0	12.0	13.2

(i) Χρησιμοποιήστε κατάλληλο τεστ και ελέγξτε την υπόθεση ότι οι κατανομές του βάρους των μπαταριών, που παράγονται από τις δύο διαφορετικές εταιρείες, είναι ίδιες.

(ii) Να ελέγξετε την υπόθεση ότι η κατανομή του βάρους των μπαταριών της 1ης εταιρείας (τιμές X) είναι η Ομοιόμορφη $\mathscr{U}(2,8)$.

Άσκηση 13.6. Από τους αποφοίτους του Τμήματος Μαθηματικών, που πρόκειται να ορκιστούν, επιλέγουμε τυχαία 12 από αυτούς και καταγράφουμε τον βαθμό του πτυχίου τους:

$$7.05\,,\ 6.27\,,\ 5.89\,,\ 6.44\,,\ 6.57\,,\ 8.21,\ 6.07\,,\ 6.12\,,\ 6.64\,,\ 7.18\,,\ 8.63\,,\ 6.22.$$

Χρησιμοποιώντας έναν έλεγχο για ποσοστιαία σημεία, ελέγξτε αν λιγότερο από το 5% των αποφοίτων που πρόκειται να ορκιστούν έχει πτυχίο Άριστα ($\geqslant 8.5$), σε ε.σ. μικρότερο του 0.1.

Άσκηση 13.7. Στο πλαίσιο μιας κλινικής μελέτης συμμετείχαν 100 άτομα που υποφέρουν από κεφαλαλγία. Αρχικά τους χορηγήθηκε ένα αντίγραφο φαρμάκου (placebo) και σε 35 από αυτά τα άτομα παρατηρήθηκε κάποια ανακούφιση, ενώ στα υπόλοιπα 65 δεν παρατηρήθηκε καμία ανακούφιση. Έπειτα, στα ίδια άτομα, χορηγήθηκε ένα καινούριο φάρμακο. Σε 55 από αυτά αναφέρθηκε ότι υπήρξε μια ανακούφίση του πόνου, κάτι που δεν παρατηρήθηκε στα υπόλοιπα 45. Από τα 65 άτομα στα οποία δεν παρουσιάστηκε καμιά ανακούφιση από το αντίγραφο, στα 30 δεν παρουσιάστηκε καμία ανακούφιση και από το καινούριο φάρμακο. Υπάρχει ένδειξη ότι το καινούριο φάρμακο είναι πιο αποτελεσματικό από το αντίγραφο;

Άσκηση 13.8. Οι χρόνοι ζωής (σε ώρες) 16 λυχνιών ραδιοφώνου που επιλέχθηκαν τυχαία από την παραγωγή ενός εργοστασίου είναι οι εξής:

$$46.9\ \ 56.8\ \ 63.3\ \ 97.1\ \ 47.2\ \ 59.2\ \ 63.4\ \ 67.7\ \ 49.1\ \ 59.9\ \ 63.7\ \ 73.3\ \ 56.5\ \ 63.2\ \ 64.1\ \ 78.5.$$

Ελέγξτε σε ε.σ. 5%, αν τουλάχιστον το 5% των λυχνιών έχει χρόνο ζωής που ξεπερνά τις 70 ώρες.

Άσκηση 13.9. Δύο κριτικοί γεύσης βαθμολόγησαν 15 εστιατόρια σε μια συνεχή κλίμακα από το 0 έως το 10. Τα δεδομένα δίνονται στον παρακάτω πίνακα. Να ελέγξετε την υπόθεση ότι ο κριτικός Β έμεινε περισσότερο ικανοποιημένος από τον Α. Για τον έλεγχο να χρησιμοποιηθεί το προσημικό κριτήριο (sign test).

Εστιατόριο	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Κριτικός Α	6.1	5.2	8.9	7.4	4.3	9.7	8.5	6.5	9.2	7.1	5.2	5.9	7.5	8.1	8.8
Κριτικός Β	7.3	5.5	9.1	7.0	5.1	9.8	7.7	6.1	9.3	8.0	4.4	5.5	6.4	6.9	9.2

Άσκηση 13.10. Πήραμε δείγμα ιζήματος από 8 πλευρές κατά μήκος ενός ποταμού και υπολογίσαμε το μέσο μέγεθος του κόκκου της άμμου (σε mm), όπως φαίνεται παρακάτω. Να ελεγχθεί η υπόθεση ότι το μέσο μέγεθος του κόκκου δεν ξεπερνάει την τιμή 6.2.

$$5.7\ \ 5.6\ \ 4.5\ \ 6.4\ \ 6.9\ \ 7.8\ \ 8.0\ \ 4.9.$$

Άσκηση 13.11. Στο γυμνάσιο μιας επαρχιακής πόλης 12 από τους 48 μαθητές της πρώτης τάξης ζούσαν σε αγροικίες. Ο καθηγητής της Φυσικής Αγωγής αυτού του Γυμνασίου, θέλοντας να εξετάσει κατά πόσο τα παιδιά που ζούσαν σε αγροικίες είχαν καλύτερη φυσική κατάσταση από τα παιδιά που ζούσαν στην πόλη, επινόησε ένα τεστ φυσικής κατάστασης στο οποίο και υπέβαλε τους 48 μαθητές. Τα αποτελέσματα του τεστ συνοψίζονται στον πίνακα που ακολουθεί (οι υψηλοί βαθμοί αντανακλούν πολύ καλή φυσική κατάσταση). Ποια είναι τα συμπεράσματά σας;

Αποτελέσματα του τεστ φυσικής κατάστασης

Παιδιά Αγροικιών		Παιδιά Πόλης					
14.8	10.6	12.7	16.9	7.6	2.4	6.2	9.9
7.3	12.5	14.2	7.9	11.3	6.4	6.1	10.6
5.6	12.9	12.6	16.0	8.3	9.1	15.3	14.8
6.3	16.1	2.1	10.6	6.7	6.7	10.6	5.0
9.0	11.4	17.7	5.6	3.6	18.6	1.8	2.6
4.2	2.7	11.8	5.6	1.0	3.2	5.9	4.0

Άσκηση 13.12. Σε ένα μάθημα διοίκησης επιχειρήσεων, οι διαλέξεις δόθηκαν με δύο διαφορετικούς τρόπους. Μία ομάδα φοιτητών παρακολούθησε τις διαλέξεις μέσω τηλεδιάσκεψης, ενώ μια δεύτερη ομάδα διά ζώσης στο αμφιθέατρο. Σε κάθε περίπτωση, όλοι οι φοιτητές εξετάστηκαν πριν από τη διάλεξη και μετά από αυτήν. Θετική βαθμολογία σημαίνει πως η επίδοση στην εξέταση πριν τη διάλεξη είναι μεγαλύτερη της επίδοσης στην εξέταση μετά τη διάλεξη. Οι διαφορές στη βαθμολογία των δύο εξετάσεων για όλους τους φοιτητές καταγράφονται στον παρακάτω πίνακα. Εξετάστε αν υπάρχει κάποια σημαντική διαφορά στη βαθμολογία των δύο ομάδων των φοιτητών.

Δια ζώσης	20.3	23.5	4.8	21.9	15.5	20.3	26.6	21.9	- 9.4	4.4	-1.6	25.1
Τηλεδιάσκεψη	6.2	15.6	25.0	4.7	28.1	17.2	14.1	31.2	12.6	9.4	17.2	23.4

Άσκηση 13.13. Σε 12 ζευγάρια μονοζυγωτικών (identical) διδύμων εφαρμόστηκαν ψυχολογικά κριτήρια με σκοπό να προσδιοριστεί το γεγονός κατά πόσο το πρώτο σε σειρά γέννησης (X) από τα δίδυμα τείνει να γίνει πιο επιθετικό από το δεύτερο (Y). Δόθηκαν τα ακόλουθα αποτελέσματα, όπου ο μεγαλύτερος αριθμός σημαίνει μεγαλύτερη επιθετικότητα. Χρησιμοποιήστε κατάλληλο τεστ και ελέγξτε τον παραπάνω ισχυρισμό.

Ζευγάρι	1	2	3	4	5	6	7	8	9	10	11	12
X_i	86	71	77	68	91	72	77	91	70	71	88	87
Y_i	88	77	76	64	96	72	65	90	65	80	81	72

Άσκηση 13.14. Σε τυχαίο δείγμα 20 ενηλίκων ζητήθηκε να αξιολογήσουν, μέσω «τυφλού» (blind) τεστ, την ποιότητα δύο ποικιλιών ντομάτας, μιας εγχώριας και μιας εισαγόμενης. Οι βαθμολογίες ήταν από 1 (χαμηλότερη ποιότητα) έως 10 (υψηλότερη ποιότητα). Τα αποτελέσματα δίνονται στον παρακάτω πίνακα. Χρησιμοποιώντας τον έλεγχο του Wilcoxon, να ελέγξετε την υπόθεση ότι οι ενήλικες προτιμούν την εγχώρια ντομάτα.

Ενήλικας	Εισαγωγής	Εγχώρια	Ενήλικας	Εισαγωγής	Εγχώρια
1	2	6	11	6	8
2	3	5	12	4	5
3	7	6	13	2	3
4	8	8	14	6	4
5	7	5	15	5	4
6	4	8	16	8	9
7	3	9	17	7	5
8	4	6	18	9	8
9	5	4	19	6	7
10	6	9	20	3	2

Άσκηση 13.15. Οκτώ παραγωγοί βρώσιμης ελιάς επιλέχθηκαν με τυχαίο τρόπο και τους δόθηκε η δυνατότητα να καλλιεργήσουν 4 διαφορετικές ποικιλίες (Π1 – Π4), σε ακριβώς 4 ίδια τεμάχια των χωραφιών τους. Μετά από καθορισμένο χρόνο, καταγράφηκε η παραγωγή του κάθε καλλιεργητή (σε κιλά), για καθεμία από τις 4 διαφορετικές ποικιλίες. Χρησιμοποιώντας τα δεδομένα που δίνονται στον παρακάτω πίνακα και το τεστ του Friedman, ελέγξτε την υπόθεση ότι δεν υπάρχει διαφοροποίηση στην απόδοση των διαφορετικών καλλιεργειών. Σε περίπτωση που απορριφθεί η H_0, να διεξάγετε πολλαπλές συγκρίσεις με ολικό επίπεδο σημαντικότητας 5%.

Α/Α	Π1	Π2	Π3	Π4
1	80	85	110	64
2	83	90	108	55
3	58	66	89	48
4	55	71	90	49
5	82	48	105	47
6	89	70	112	59
7	60	89	85	66
8	89	58	83	62

Άσκηση 13.16. Οι υπάλληλοι μιας εταιρείας συμμετέχουν σε εκπαίδευση σε ένα νέο λογισμικό. Οι συμμετέχοντες χωρίστηκαν σε 4 ομάδες των 5 ατόμων και κάθε ομάδα ανέλαβε διαφορετικός εκπαιδευτής (συνολικά 4 εκπαιδευτές). Στο τέλος της περιόδου εκπαίδευσης οι συμμετέχοντες αξιολογήθηκαν από τους εκπαιδευτές τους σε κλίμακα 0-100. Παρακάτω δίνονται οι βαθμολογίες όλων των υπαλλήλων. Να ελέγξετε την υπόθεση ότι δεν υπάρχει διαφορά στη διάμεση επίδοση μεταξύ των 4 ομάδων. Σε περίπτωση που απορριφθεί η προς έλεγχο υπόθεση, να διεξάγετε πολλαπλές συγκρίσεις με ολικό επίπεδο σημαντικότητας 5%.

Ομάδα	1	2	3	4	5
1	45	55	62	64	49
2	65	72	71	79	76
3	85	80	78	89	91
4	88	76	90	91	95

Άσκηση 13.17. Τα παρακάτω δεδομένα αφορούν την ετήσια σοδειά καλαμποκιού (σε τόνους). Έχουν ληφθεί 4 διαφορετικά δείγματα για καθεμία από τις 4 διαφορετικές μεθόδους καλλιέργειας. Να ελέγξετε την υπόθεση ότι δεν υπάρχει διαφορά στη διάμεση ετήσια σοδειά μεταξύ των 4 διαφορετικών μεθόδων καλλιέργειας. Σε περίπτωση που απορριφθεί η H_0, να διεξάγετε πολλαπλές συγκρίσεις με ολικό επίπεδο σημαντικότητας 5%.

Ομάδα	1	2	3	4	5	6	7	8	9	10
1	82	92	93	91	89	96	78	92	90	
2	95	90	80	82	85	83	87	94	89	88
3	100	102	92	96	91	93	92			
4	79	83	82	84	75	85	80	81	89	86

Άσκηση 13.18. Μια νευροψυχολόγος θέλει να αξιολογήσει την ικανότητα μνήμης νέων ενηλίκων. Στο πλαίσιο της μελέτης επιλέγονται με τυχαίο τρόπο 12 φοιτητές/τριες για να συμμετάσχουν στο παρακάτω

πείραμα. Σε καθέναν από αυτούς δίνονται 4 λίστες, με 20 λέξεις σε καθεμία. Αφού τις διαβάσουν, στη συνέχεια προσπαθούν να τις θυμηθούν. Τα δεδομένα αφορούν το πλήθος των λέξεων από κάθε λίστα τις οποίες θυμήθηκαν σωστά οι συμμετέχοντες και δίνονται στον παρακάτω πίνακα. Να ελέγξετε την υπόθεση ότι όλες οι λίστες έχουν την ίδια δυσκολία στην απομνημόνευση. Αν απορρίψετε την H_0, να διεξάγετε πολλαπλές συγκρίσεις (με ολικό επίπεδο σημαντικότητας 5%), ώστε να διαπιστώσετε ποιες λίστες διαφέρουν ως προς τη δυσκολία απομνημόνευσης των λέξεων που περιέχουν.

	Πλήθος λέξεων			
Φοιτητής	Λίστα 1	Λίστα 2	Λίστα 3	Λίστα 4
1	18	14	16	20
2	7	6	5	10
3	13	14	16	17
4	15	10	12	14
5	12	11	12	18
6	11	9	9	16
7	15	16	10	14
8	10	8	11	16
9	14	12	13	15
10	9	9	9	10
11	8	6	9	14
12	10	11	13	16

Άσκηση 13.19. Σε ένα μικροβιολογικό εργαστήριο, ελέγχονται διαδοχικά δείγματα αίματος με σκοπό την ανίχνευση ενός συγκεκριμένου ιού. Τα αποτελέσματα είναι + αν το τεστ είναι θετικό (παρουσία ιού) ή − αν το τεστ είναι αρνητικό (απουσία ιού). Τα αποτελέσματα στη διάρκεια μιας εβδομάδας είναι τα εξής:

$$++---+--+-----++---+--++++-+---+-++$$

Χρησιμοποιήστε το τεστ των ροών (ασυμπτωτική μορφή ελέγχου) και ελέγξτε την υπόθεση ότι τα αποτελέσματα των εξετάσεων δεν δείχνουν αυξημένο ή μειωμένο ιικό φορτίο στην περιοχή, δηλαδή ότι το δείγμα είναι τυχαίο.

Άσκηση 13.20. Στη διάθεσή μας έχουμε δεδομένα 100 ημερών, στα οποία έχουμε καταγράψει το ύψος της βροχόπτωσης. Αν η βροχόπτωση ξεπερνά τα 25,4 χιλιοστά, τότε η ημέρα χαρακτηρίζεται ως υγρή (W = wet), διαφορετικά χαρακτηρίζεται ως ξηρή (D = dry). Να ελέγξετε την υπόθεση ότι οι ενδείξεις D,W εμφανίζονται με τυχαίο τρόπο χρησιμοποιώντας την ασυμπτωτική μορφή του τεστ των ροών.

Ημέρα	Αποτέλεσμα	Ημέρα	Αποτέλεσμα	Ημέρα	Αποτέλεσμα	Ημέρα	Αποτέλεσμα	Ημέρα	Αποτέλεσμα
1	D	21	W	41	W	61	D	81	W
2	D	22	W	42	W	62	D	82	D
3	D	23	W	43	W	63	D	83	W
4	D	24	D	44	W	64	D	84	D
5	D	25	D	45	W	65	W	85	D
6	W	26	D	46	D	66	W	86	W
7	W	27	D	47	W	67	D	87	W
8	W	28	D	48	W	68	W	88	D
9	D	29	D	49	D	69	W	89	W
10	D	30	D	50	D	70	D	90	D
11	D	31	D	51	D	71	W	91	W
12	D	32	D	52	D	72	D	92	W
13	D	33	W	53	D	73	D	93	W
14	D	34	D	54	W	74	D	94	W
15	W	35	W	55	W	75	D	95	W
16	W	36	W	56	D	76	D	96	W
17	D	37	D	57	D	77	D	97	W
18	D	38	D	58	D	78	D	98	W
19	D	39	W	59	D	79	W	99	D
20	D	40	W	60	D	80	W	100	W

Άσκηση 13.21. Ρωτήσαμε 10 άτομα που πηγαίνουν γυμναστήριο πόσες ώρες αθλούνται τον μήνα και πόσα φρούτα καταναλώνουν (κάθε μήνα). Οι απαντήσεις που συλλέξαμε, παρουσιάζονται στον παρακάτω πίνακα. Χρησιμοποιώντας τον συντέλεστη συσχέτισης του Spearman, να ελέγξετε την υπόθεση ότι δεν υπάρχει συσχέτιση ανάμεσα στις ώρες που γυμνάζεται ένα άτομο και στην κατανάλωση φρούτων από αυτό.

Ώρες άθλησης	12	15	24	18	30	32	17	27	42	8
Κατανάλωση φρούτων	22	28	30	26	26	48	30	32	58	15

Άσκηση 13.22. Μία έρευνα που έγινε από μια εταιρεία είχε σκοπό να διαπιστώσει την ύπαρξη στατιστικά σημαντικής εξάρτησης μεταξύ της ανάπτυξης της αγοράς και της τοποθεσίας στην οποία βρισκόταν. Να ελέγξετε την υπόθεση αυτή, χρησιμοποιώντας έναν χι-τετράγωνο έλεγχο καλής προσαρμογής με βάση τα παρακάτω δεδομένα. Να διατυπώσετε το συμπέρασμά σας.

	Τοποθεσία		
Ανάπτυξη	Αστική	Ημιαστική	Αγροτική
Μεγάλη	125	210	65
Μεσαία	100	180	70
Μηδενική	75	110	65

Άσκηση 13.23. Στη διάθεσή μας έχουμε τα παρακάτω δεδομένα τα οποία αφορούν το πλήθος των λέξεων της 1ης πλήρους πρότασης από 10 τυχαίες επιλεγμένες σελίδες δύο βιβλίων. Το βιβλίο 1 (Β1) είναι από τον συγγραφέα Σ1, ενώ το βιβλίο 2 (Β2) είναι από τον συγγραφέα Σ2. Να ελέγξετε την υπόθεση ότι η κατανομή του μήκους των προτάσεων στα δύο βιβλία είναι διαφορετική.

Σελ	1	2	3	4	5	6	7	8	9	10
Β1	21	20	17	25	29	24	38	18	32	31
Β2	45	14	13	33	35	19	58	41	64	26

Άσκηση 13.24. Στο πλαίσιο ενός φαρμακολογικού πειράματος ελήφθη δείγμα 7 υγιών σκυλιών και μετρήθηκε η κατανάλωση καρδιακού οξυγόνου (MVO), καθώς και η πίεση αριστερής κοιλίας (LVP). Τα δεδομένα δίνονται στον πίνακα που ακολουθεί.

Σκυλί	1	2	3	4	5	6	7
MVO	78	92	116	90	106	78	99
LVP	32	33	45	30	38	24	44

(i) Να υπολογίσετε και να ερμηνεύσετε τους συντελεστές συσχέτισης Spearman και Kendall χρησιμοποιώντας τα διαθέσιμα δεδομένα.

(ii) Να ελέγξετε την υπόθεση ότι η κατανάλωση καρδιακού οξυγόνου και η πίεση αριστερής κοιλίας είναι ασυσχέτιστες σε υγιή σκυλιά. Ο έλεγχος να γίνει και με τους δύο συντελεστές.

Άσκηση 13.25. Οι 90 μαθητές της 1ης τάξης ενός Δημοτικού Σχολείου χωρίστηκαν σε δύο τμήματα. Στο ένα τμήμα (40 μαθητές) διδάχθηκε ανάγνωση με τη μέθοδο Μ1 και στο 2ο τμήμα (50 μαθητές) χρησιμοποιήθηκε η μέθοδος Μ2. Κατόπιν, οι μαθητές και των δυο ομάδων αξιολογήθηκαν (στην κλίμακα 1-10) και τα τελικά αποτελέσματα της επίδοσής τους στην ανάγνωση δίνονται στον παρακάτω πίνακα.

Βαθμός (X)	1	2	3	4	5	6	7	8	9	10
Αριθμός Μαθητών στη Μ1	0	0	4	10	13	8	4	0	1	0
Αριθμός Μαθητών στη Μ2	1	2	3	10	16	7	3	3	4	1

Χρησιμοποιώντας το τεστ των Kolmogorov-Smirnov για δύο δείγματα, να ελέγξετε την υπόθεση ότι και οι δύο μεθοδολογίες είναι το ίδιο αποτελεσματικές.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Αντζουλάκος, Δ. (2013). *Ανάλυση Δεδομένων με τη Χρήση Στατιστικών Πακέτων. Εισαγωγή στο R* (2η εκδ.). Πειραιάς: Πανεπιστημιακές Σημειώσεις.

Ντζούφρας, Ι. και Καρλής, Δ. (2015). *Εισαγωγή στον προγραμματισμό και στη στατιστική ανάλυση με R* [Προπτυχιακό εγχειρίδιο]. Αθήνα: Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. URL: http://hdl.handle.net/11419/2601.

Σαχλάς, Α. και Μπερσίμης, Σ. (2016). *Εφαρμοσμένη Στατιστική με Χρήση του IBM SPSS Statistics 23 : Με έμφαση στις Επιστήμες Υγείας*. Θεσσαλονίκη: Τζιόλας.

Φουσκάκης, Δ. (2021). *Ανάλυση δεδομένων με χρήση της R* (2η εκδ.). Αθήνα: Τσότρας.

Φωκιανός, Κ. και Χαραλάμπους, Χ. (2010). *Εισαγωγή στην R: Πρόχειρες Σημειώσεις* (2η εκδ.). Κύπρος: Πανεπιστημιακές Σημειώσεις.

Χαλικιάς, Μ., Λάλου, Π. και Μανωλέσου, Α. (2015). *Μεθοδολογία έρευνας και εισαγωγή στη Στατιστική Ανάλυση Δεδομένων με το IBM SPSS STATISTICS*[Εργαστηριακός Οδηγός]. Αθήνα: Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. URL: http://hdl.handle.net/11419/5075.

Ξενόγλωσση

Clopper, C. and Pearson, E. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*, pp. 404–416.

Conover, W. J. (1998). *Practical Nonparametric Statistics* (3rd ed.). Wiley and Sons, Inc.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, *6*, pp. 241–252.

Hollander, M., Wolfe, D. and Chicken, E. (2014). *Nonparametric Statistical Methods* (3nd ed.). John Wiley and Sons.

Hotelling, H. and Pabst, M. (1936). Rank correlation and tests of significance involving the assumption of normality. *Annals of Mathematical Statistics*, *7*, pp. 29–43.

Mann, H. and Whitney, D. R. (1947). On a test whether one of two random variables is stochastically larger than the other. *Ann. Math. Statistics*, *18*, pp. 50–60.

Siegel, S. and Castellan, N. (1988). *Nonparametric Statistics for the Behavioral Sciences*. Springer.

Sprent, P. (1999). *Applied Nonparametric Statistical Methods*. Chapmann and Hall.

Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Ann. Math. Statistics*, *11*, pp. 147–162.

ΠΑΡΑΡΤΗΜΑ

ΠΙΝΑΚΕΣ

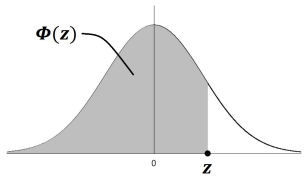
Στο παράρτημα αυτό θα δοθούν οι πίνακες κρίσιμων τιμών/ποσοστιαίων σημείων των κατανομών των ελεγχοσυναρτήσεων των παρακάτω κριτηρίων:

- Έλεγχος καλής προσαρμογής Kolmogorov-Smirnov.
- Έλεγχος Smirnov για δύο ανεξάρτητα δείγματα (για ισομεγέθη αλλά και για ανισομεγέθη δείγματα).
- Έλεγχος Wilcoxon Signed Rank.
- Έλεγχος Mann-Whitney (με χρήση της στατιστικής συνάρτησης U αλλά και της στατιστικής συνάρτησης R_1).
- Έλεγχος συσχέτισης με χρήση συντελεστή Spearman's r_s.
- Έλεγχος συσχέτισης με χρήση συντελεστή Kendall's τ.
- Έλεγχος τυχαιότητας με χρήση ροών (Runs Test).

Επιπλέον, δίνονται πίνακες ποσοστιαίων σημείων των κατανομών χι-τετράγωνο, Student's t και τυπικής Κανονικής κατανομής, καθώς και τιμές της αθροιστικής συνάρτησης κατανομής της Διωνυμικής κατανομής $\mathrm{B}(n,p)$ για διάφορες τιμές των παραμέτρων n,p.

Τυπική Κανονική κατανομή - $\mathcal{N}(0,1)$

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{(2\pi)^{1/2}} e^{-t^2/2} dt$$



Πίνακας Π.1: Πίνακας Τυπικής Κανονικής κατανομής - $Z \sim \mathcal{N}(0,1)$. Ο πίνακας δίνει τις τιμές $\Phi(z) = P(Z \leqslant z)$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997

Κατανομή t (Student)

Πίνακας Π.2: Πίνακας t κατανομής. Ο πίνακας δίνει τις τιμές $t_{\nu,a}$: $P(t_\nu \geqslant t_{\nu,a}) = a$.

	a												
ν	.4	.3	.25	.2	.15	.1	.075	.05	.025	.01	.005	.001	.0005
1	0.325	0.727	1.000	1.376	1.963	3.078	4.165	6.314	12.706	31.821	63.657	318.310	636.620
2	0.289	0.617	0.816	1.061	1.386	1.886	2.282	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.584	0.765	0.978	1.250	1.638	1.924	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.569	0.741	0.941	1.190	1.533	1.778	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.559	0.727	0.920	1.156	1.476	1.699	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.553	0.718	0.906	1.134	1.440	1.650	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.549	0.711	0.896	1.119	1.415	1.617	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.546	0.706	0.889	1.108	1.397	1.592	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.543	0.703	0.883	1.100	1.383	1.574	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.542	0.700	0.879	1.093	1.372	1.559	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.540	0.697	0.876	1.088	1.363	1.548	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.539	0.695	0.873	1.083	1.356	1.538	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.538	0.694	0.870	1.079	1.350	1.530	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.537	0.692	0.868	1.076	1.345	1.523	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.536	0.691	0.866	1.074	1.341	1.517	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.535	0.690	0.865	1.071	1.337	1.512	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.534	0.689	0.863	1.069	1.333	1.508	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.534	0.688	0.862	1.067	1.330	1.504	1.734	2.101	2.552	2.878	3.610	3.922
19	0.257	0.533	0.688	0.861	1.066	1.328	1.500	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.533	0.687	0.860	1.064	1.325	1.497	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.532	0.686	0.859	1.063	1.323	1.494	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.532	0.686	0.858	1.061	1.321	1.492	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.532	0.685	0.858	1.060	1.319	1.489	1.714	2.069	2.500	2.807	3.485	3.768
24	0.256	0.531	0.685	0.857	1.059	1.318	1.487	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.531	0.684	0.856	1.058	1.316	1.485	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.531	0.684	0.856	1.058	1.315	1.483	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.531	0.684	0.855	1.057	1.314	1.482	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.530	0.683	0.855	1.056	1.313	1.480	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.530	0.683	0.854	1.055	1.311	1.479	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.530	0.683	0.854	1.055	1.310	1.477	1.697	2.042	2.457	2.750	3.385	3.646
31	0.256	0.530	0.682	0.853	1.054	1.309	1.476	1.696	2.040	2.453	2.744	3.375	3.633
32	0.255	0.530	0.682	0.853	1.054	1.309	1.475	1.694	2.037	2.449	2.738	3.365	3.622
33	0.255	0.530	0.682	0.853	1.053	1.308	1.474	1.692	2.035	2.445	2.733	3.356	3.611
34	0.255	0.529	0.682	0.852	1.052	1.307	1.473	1.691	2.032	2.441	2.728	3.348	3.601
35	0.255	0.529	0.682	0.852	1.052	1.306	1.472	1.690	2.030	2.438	2.724	3.340	3.591
36	0.255	0.529	0.681	0.852	1.052	1.306	1.471	1.688	2.028	2.434	2.719	3.333	3.582
37	0.255	0.529	0.681	0.851	1.051	1.305	1.470	1.687	2.026	2.431	2.715	3.326	3.574
38	0.255	0.529	0.681	0.851	1.051	1.304	1.469	1.686	2.024	2.429	2.712	3.319	3.566
39	0.255	0.529	0.681	0.851	1.050	1.304	1.468	1.685	2.023	2.426	2.708	3.313	3.558
40	0.255	0.529	0.681	0.851	1.050	1.303	1.468	1.684	2.021	2.423	2.704	3.307	3.551
45	0.255	0.528	0.680	0.850	1.049	1.301	1.465	1.679	2.014	2.412	2.690	3.281	3.520
50	0.255	0.528	0.679	0.849	1.047	1.299	1.462	1.676	2.009	2.403	2.678	3.261	3.496
60	0.254	0.527	0.679	0.848	1.045	1.296	1.458	1.671	2.000	2.390	2.660	3.232	3.460
70	0.254	0.527	0.678	0.847	1.044	1.294	1.456	1.667	1.994	2.381	2.648	3.211	3.435
80	0.254	0.526	0.678	0.846	1.043	1.292	1.453	1.664	1.990	2.374	2.639	3.195	3.416
90	0.254	0.526	0.677	0.846	1.042	1.291	1.452	1.662	1.987	2.368	2.632	3.183	3.402
100	0.254	0.526	0.677	0.845	1.042	1.290	1.451	1.660	1.984	2.364	2.626	3.174	3.390
120	0.254	0.526	0.677	0.845	1.041	1.289	1.449	1.658	1.980	2.358	2.617	3.160	3.373
∞	0.253	0.524	0.674	0.842	1.036	1.282	1.440	1.645	1.960	2.326	2.576	3.090	3.291

Κατανομή χ^2

Πίνακας Π.3: Πίνακας χ^2 κατανομής. Ο πίνακας δίνει τις τιμές $\chi^2_{\nu,a}$: $P(\chi^2_\nu \geqslant \chi^2_{\nu,a}) = a$.

	a											
ν	.9995	.999	.995	.91	.975	.95	.925	.9	.85	.8	.7	0.6
1	$.0^6393$	$.0^5157$	$.0^4393$	$.0^3157$	0.001	0.004	0.009	0.016	0.036	0.064	0.148	0.275
2	0.001	0.002	0.010	0.020	0.051	0.103	0.156	0.211	0.325	0.446	0.713	1.022
3	0.015	0.024	0.072	0.115	0.216	0.352	0.472	0.584	0.798	1.005	1.424	1.869
4	0.064	0.091	0.207	0.297	0.484	0.711	0.897	1.064	1.366	1.649	2.195	2.753
5	0.158	0.210	0.412	0.554	0.831	1.145	1.394	1.610	1.994	2.343	3.000	3.655
6	0.299	0.381	0.676	0.872	1.237	1.635	1.941	2.204	2.661	3.070	3.828	4.570
7	0.485	0.598	0.989	1.239	1.690	2.167	2.528	2.833	3.358	3.822	4.671	5.493
8	0.710	0.857	1.344	1.646	2.180	2.733	3.144	3.490	4.078	4.594	5.527	6.423
9	0.972	1.152	1.735	2.088	2.700	3.325	3.785	4.168	4.817	5.380	6.393	7.357
10	1.265	1.479	2.156	2.558	3.247	3.940	4.446	4.865	5.570	6.179	7.267	8.295
11	1.587	1.834	2.603	3.053	3.816	4.575	5.124	5.578	6.336	6.989	8.148	9.237
12	1.934	2.214	3.074	3.571	4.404	5.226	5.818	6.304	7.114	7.807	9.034	10.182
13	2.305	2.617	3.565	4.107	5.009	5.892	6.524	7.042	7.901	8.634	9.926	11.129
14	2.697	3.041	4.075	4.660	5.629	6.571	7.242	7.790	8.696	9.467	10.821	12.078
15	3.108	3.483	4.601	5.229	6.262	7.261	7.969	8.547	9.499	10.307	11.721	13.030
16	3.536	3.942	5.142	5.812	6.908	7.962	8.707	9.312	10.309	11.152	12.624	13.983
17	3.980	4.416	5.697	6.408	7.564	8.672	9.452	10.085	11.125	12.002	13.531	14.937
18	4.439	4.905	6.265	7.015	8.231	9.390	10.205	10.865	11.946	12.857	14.440	15.893
19	4.912	5.407	6.844	7.633	8.907	10.117	10.965	11.651	12.773	13.716	15.352	16.850
20	5.398	5.921	7.434	8.260	9.591	10.851	11.732	12.443	13.604	14.578	16.266	17.809
21	5.896	6.447	8.034	8.897	10.283	11.591	12.504	13.240	14.439	15.445	17.182	18.768
22	6.404	6.983	8.643	9.542	10.982	12.338	13.282	14.041	15.279	16.314	18.101	19.729
23	6.924	7.529	9.260	10.196	11.689	13.091	14.065	14.848	16.122	17.187	19.021	20.690
24	7.453	8.085	9.886	10.856	12.401	13.848	14.853	15.659	16.969	18.062	19.943	21.652
25	7.991	8.649	10.520	11.524	13.120	14.611	15.645	16.473	17.818	18.940	20.867	22.616
26	8.538	9.222	11.160	12.198	13.844	15.379	16.441	17.292	18.671	19.820	21.792	23.579
27	9.093	9.803	11.808	12.879	14.573	16.151	17.241	18.114	19.527	20.703	22.719	24.544
28	9.656	10.391	12.461	13.565	15.308	16.928	18.045	18.939	20.386	21.588	23.647	25.509
29	10.227	10.986	13.121	14.256	16.047	17.708	18.853	19.768	21.247	22.475	24.577	26.475
30	10.804	11.588	13.787	14.953	16.791	18.493	19.664	20.599	22.110	23.364	25.508	27.442
31	11.389	12.196	14.458	15.655	17.539	19.281	20.478	21.434	22.976	24.255	26.440	28.409
32	11.979	12.811	15.134	16.362	18.291	20.072	21.295	22.271	23.844	25.148	27.373	29.376
33	12.576	13.431	15.815	17.074	19.047	20.867	22.115	23.110	24.714	26.042	28.307	30.344
34	13.179	14.057	16.501	17.789	19.806	21.664	22.938	23.952	25.586	26.938	29.242	31.313
35	13.787	14.688	17.192	18.509	20.569	22.465	23.763	24.797	26.460	27.836	30.178	32.282
36	14.401	15.324	17.887	19.233	21.336	23.269	24.591	25.643	27.336	28.735	31.115	33.252
37	15.020	15.965	18.586	19.960	22.106	24.075	25.421	26.492	28.214	29.635	32.053	34.222
38	15.644	16.611	19.289	20.691	22.878	24.884	26.254	27.343	29.093	30.537	32.992	35.192
39	16.273	17.262	19.996	21.426	23.654	25.695	27.089	28.196	29.974	31.441	33.932	36.163
40	16.906	17.916	20.707	22.164	24.433	26.509	27.926	29.051	30.856	32.345	34.872	37.134
45	20.137	21.251	24.311	25.901	28.366	30.612	32.140	33.350	35.290	36.884	39.585	41.995
50	23.461	24.674	27.991	29.707	32.357	34.764	36.397	37.689	39.754	41.449	44.313	46.864
60	30.340	31.738	35.534	37.485	40.482	43.188	45.016	46.459	48.759	50.641	53.809	56.620
70	37.467	39.036	43.275	45.442	48.758	51.739	53.748	55.329	57.844	59.898	63.346	66.396
80	44.791	46.520	51.172	53.540	57.153	60.391	62.568	64.278	66.994	69.207	72.915	76.188
90	52.276	54.155	59.196	61.754	65.647	69.126	71.460	73.291	76.195	78.558	82.511	85.993
100	59.896	61.918	67.328	70.065	74.222	77.929	80.412	82.358	85.441	87.945	92.129	95.808
120	75.467	77.755	83.852	86.923	91.573	95.705	98.464	100.624	104.037	106.806	111.419	115.465
150	99.463	102.113	109.142	112.668	117.985	122.692	125.827	128.275	132.137	135.263	140.457	145.000

Κατανομή χ^2

Πίνακας Π.4: Πίνακας χ^2 κατανομής. Ο πίνακας δίνει τις τιμές $\chi^2_{\nu,a}$: $P(\chi^2_\nu \geqslant \chi^2_{\nu,a}) = a$.

	a												
ν	.5	.4	.3	.2	.15	.1	.0755	.05	.025	.01	.005	.0001	.0005
1	0.455	0.708	1.074	1.642	2.072	2.706	3.170	3.841	5.024	6.635	7.879	10.828	12.116
2	1.386	1.833	2.408	3.219	3.794	4.605	5.181	5.991	7.378	9.210	10.597	13.816	15.202
3	2.366	2.946	3.665	4.642	5.317	6.251	6.905	7.815	9.348	11.345	12.838	16.266	17.730
4	3.357	4.045	4.878	5.989	6.745	7.779	8.496	9.488	11.143	13.277	14.860	18.467	19.997
5	4.351	5.132	6.064	7.289	8.115	9.236	10.008	11.070	12.833	15.086	16.750	20.515	22.105
6	5.348	6.211	7.231	8.558	9.446	10.645	11.466	12.592	14.449	16.812	18.548	22.458	24.103
7	6.346	7.283	8.383	9.803	10.748	12.017	12.883	14.067	16.013	18.475	20.278	24.322	26.018
8	7.344	8.351	9.524	11.030	12.027	13.362	14.270	15.507	17.535	20.090	21.955	26.124	27.868
9	8.343	9.414	10.656	12.242	13.288	14.684	15.631	16.919	19.023	21.666	23.589	27.877	29.666
10	9.342	10.473	11.781	13.442	14.534	15.987	16.971	18.307	20.483	23.209	25.188	29.588	31.420
11	10.341	11.530	12.899	14.631	15.767	17.275	18.294	19.675	21.920	24.725	26.757	31.264	33.137
12	11.340	12.584	14.011	15.812	16.989	18.549	19.602	21.026	23.337	26.217	28.300	32.909	34.821
13	12.340	13.636	15.119	16.985	18.202	19.812	20.897	22.362	24.736	27.688	29.819	34.528	36.478
14	13.339	14.685	16.222	18.151	19.406	21.064	22.180	23.685	26.119	29.141	31.319	36.123	38.109
15	14.339	15.733	17.322	19.311	20.603	22.307	23.452	24.996	27.488	30.578	32.801	37.697	39.719
16	15.338	16.780	18.418	20.465	21.793	23.542	24.716	26.296	28.845	32.000	34.267	39.252	41.308
17	16.338	17.824	19.511	21.615	22.977	24.769	25.970	27.587	30.191	33.409	35.718	40.790	42.879
18	17.338	18.868	20.601	22.760	24.155	25.989	27.218	28.869	31.526	34.805	37.156	42.312	44.434
19	18.338	19.910	21.689	23.900	25.329	27.204	28.458	30.144	32.852	36.191	38.582	43.820	45.973
20	19.337	20.951	22.775	25.038	26.498	28.412	29.692	31.410	34.170	37.566	39.997	45.315	47.498
21	20.337	21.991	23.858	26.171	27.662	29.615	30.920	32.671	35.479	38.932	41.401	46.797	49.011
22	21.337	23.031	24.939	27.301	28.822	30.813	32.142	33.924	36.781	40.289	42.796	48.268	50.511
23	22.337	24.069	26.018	28.429	29.979	32.007	33.360	35.172	38.076	41.638	44.181	49.728	52.000
24	23.337	25.106	27.096	29.553	31.132	33.196	34.572	36.415	39.364	42.980	45.559	51.179	53.479
25	24.337	26.143	28.172	30.675	32.282	34.382	35.780	37.652	40.646	44.314	46.928	52.620	54.947
26	25.336	27.179	29.246	31.795	33.429	35.563	36.984	38.885	41.923	45.642	48.290	54.052	56.407
27	26.336	28.214	30.319	32.912	34.574	36.741	38.184	40.113	43.195	46.963	49.645	55.476	57.858
28	27.336	29.249	31.391	34.027	35.715	37.916	39.380	41.337	44.461	48.278	50.993	56.892	59.300
29	28.336	30.283	32.461	35.139	36.854	39.087	40.573	42.557	45.722	49.588	52.336	58.301	60.735
30	29.336	31.316	33.530	36.250	37.990	40.256	41.762	43.773	46.979	50.892	53.672	59.703	62.162
31	30.336	32.349	34.598	37.359	39.124	41.422	42.948	44.985	48.232	52.191	55.003	61.098	63.582
32	31.336	33.381	35.665	38.466	40.256	42.585	44.131	46.194	49.480	53.486	56.328	62.487	64.995
33	32.336	34.413	36.731	39.572	41.386	43.745	45.311	47.400	50.725	54.776	57.648	63.870	66.403
34	33.336	35.444	37.795	40.676	42.514	44.903	46.488	48.602	51.966	56.061	58.964	65.247	67.803
35	34.336	36.475	38.859	41.778	43.640	46.059	47.663	49.802	53.203	57.342	60.275	66.619	69.199
36	35.336	37.505	39.922	42.879	44.764	47.212	48.835	50.998	54.437	58.619	61.581	67.985	70.588
37	36.336	38.535	40.984	43.978	45.886	48.363	50.005	52.192	55.668	59.893	62.883	69.346	71.972
38	37.335	39.564	42.045	45.076	47.007	49.513	51.173	53.384	56.896	61.162	64.181	70.703	73.351
39	38.335	40.593	43.105	46.173	48.126	50.660	52.338	54.572	58.120	62.428	65.476	72.055	74.725
40	39.335	41.622	44.165	47.269	49.244	51.805	53.501	55.758	59.342	63.691	66.766	73.402	76.095
45	44.335	46.761	49.452	52.729	54.810	57.505	59.287	61.656	65.410	69.957	73.166	80.077	82.876
50	49.335	51.892	54.723	58.164	60.346	63.167	65.030	67.505	71.420	76.154	79.490	86.661	89.561
60	59.335	62.135	65.227	68.972	71.341	74.397	76.411	79.082	83.298	88.379	91.952	99.607	102.695
70	69.334	72.358	75.689	79.715	82.255	85.527	87.680	90.531	95.023	100.425	104.215	112.317	115.578
80	79.334	82.566	86.120	90.405	93.106	96.578	98.861	101.879	106.629	112.329	116.321	124.839	128.261
90	89.334	92.761	96.524	101.054	103.904	107.565	109.969	113.145	118.136	124.116	128.299	137.208	140.782
100	99.334	102.946	106.906	111.667	114.659	118.498	121.017	124.342	129.561	135.807	140.169	149.449	153.167
120	119.334	123.289	127.616	132.806	136.062	140.233	142.965	146.567	152.211	158.950	163.648	173.617	177.603
150	149.334	153.753	158.577	164.349	167.962	172.581	175.602	179.581	185.800	193.208	198.360	209.265	213.613

Πίνακας Π.5: Πίνακας Διωνυμικής κατανομής. Αν $X \sim B(n,p)$, ο πίνακας δίνει την πιθανότητα $P(X \leqslant x)$.

n	x	p = **0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
1	0	0.9900	0.9800	0.9700	0.9600	0.9500	0.9400	0.9300	0.9200	0.9100
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	0	0.9801	0.9604	0.9409	0.9216	0.9025	0.8836	0.8649	0.8464	0.8281
	1	0.9999	0.9996	0.9991	0.9984	0.9975	0.9964	0.9951	0.9936	0.9919
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	0	0.9703	0.9412	0.9127	0.8847	0.8574	0.8306	0.8044	0.7787	0.7536
	1	0.9997	0.9988	0.9974	0.9953	0.9928	0.9896	0.9860	0.9818	0.9772
	2	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9997	0.9995	0.9993
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	0	0.9606	0.9224	0.8853	0.8493	0.8145	0.7807	0.7481	0.7164	0.6857
	1	0.9994	0.9977	0.9948	0.9909	0.9860	0.9801	0.9733	0.9656	0.9570
	2	1.0000	1.0000	0.9999	0.9998	0.9995	0.9992	0.9987	0.9981	0.9973
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	0	0.9510	0.9039	0.8587	0.8154	0.7738	0.7339	0.6957	0.6591	0.6240
	1	0.9990	0.9962	0.9915	0.9852	0.9774	0.9681	0.9575	0.9456	0.9326
	2	1.0000	0.9999	0.9997	0.9994	0.9988	0.9980	0.9969	0.9955	0.9937
	3	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9997
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	0	0.9415	0.8858	0.8330	0.7828	0.7351	0.6899	0.6470	0.6064	0.5679
	1	0.9985	0.9943	0.9875	0.9784	0.9672	0.9541	0.9392	0.9227	0.9048
	2	1.0000	0.9998	0.9995	0.9988	0.9978	0.9962	0.9942	0.9915	0.9882
	3	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9995	0.9992
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	0	0.9321	0.8681	0.8080	0.7514	0.6983	0.6485	0.6017	0.5578	0.5168
	1	0.9980	0.9921	0.9829	0.9706	0.9556	0.9382	0.9187	0.8974	0.8745
	2	1.0000	0.9997	0.9991	0.9980	0.9962	0.9937	0.9903	0.9860	0.9807
	3	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9993	0.9988	0.9982
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
8	0	0.9227	0.8508	0.7837	0.7214	0.6634	0.6096	0.5596	0.5132	0.4703
	1	0.9973	0.9897	0.9777	0.9619	0.9428	0.9208	0.8965	0.8702	0.8423
	2	0.9999	0.9996	0.9987	0.9969	0.9942	0.9904	0.9853	0.9789	0.9711
	3	1.0000	1.0000	0.9999	0.9998	0.9996	0.9993	0.9987	0.9978	0.9966
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	0	0.9135	0.8337	0.7602	0.6925	0.6302	0.5730	0.5204	0.4722	0.4279
	1	0.9966	0.9869	0.9718	0.9522	0.9288	0.9022	0.8729	0.8417	0.8088
	2	0.9999	0.9994	0.9980	0.9955	0.9916	0.9862	0.9791	0.9702	0.9595
	3	1.0000	1.0000	0.9999	0.9997	0.9994	0.9987	0.9977	0.9963	0.9943
	4	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9995
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	0	0.9044	0.8171	0.7374	0.6648	0.5987	0.5386	0.4840	0.4344	0.3894
	1	0.9957	0.9838	0.9655	0.9418	0.9139	0.8824	0.8483	0.8121	0.7746
	2	0.9999	0.9991	0.9972	0.9938	0.9885	0.9812	0.9717	0.9599	0.9460
	3	1.0000	1.0000	0.9999	0.9996	0.9990	0.9980	0.9964	0.9942	0.9912
	4	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9994	0.9990
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
11	0	0.8953	0.8007	0.7153	0.6382	0.5688	0.5063	0.4501	0.3996	0.3544
	1	0.9948	0.9805	0.9587	0.9308	0.8981	0.8618	0.8228	0.7819	0.7399
	2	0.9998	0.9988	0.9963	0.9917	0.9848	0.9752	0.9630	0.9481	0.9305
	3	1.0000	1.0000	0.9998	0.9993	0.9984	0.9970	0.9947	0.9915	0.9871
	4	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9995	0.9990	0.9983
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
12	0	0.8864	0.7847	0.6938	0.6127	0.5404	0.4759	0.4186	0.3677	0.3225
	1	0.9938	0.9769	0.9514	0.9191	0.8816	0.8405	0.7967	0.7513	0.7052
	2	0.9998	0.9985	0.9952	0.9893	0.9804	0.9684	0.9532	0.9348	0.9134
	3	1.0000	0.9999	0.9997	0.9990	0.9978	0.9957	0.9925	0.9880	0.9820
	4	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9991	0.9984	0.9973
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Πίνακας Π.6: Πίνακας Διωνυμικής κατανομής. Αν $X \sim B(n,p)$, ο πίνακας δίνει την πιθανότητα $P(X \leqslant x)$.

n	x	p = **0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
13	0	0.8775	0.7690	0.6730	0.5882	0.5133	0.4474	0.3893	0.3383	0.2935
	1	0.9928	0.9730	0.9436	0.9068	0.8646	0.8186	0.7702	0.7206	0.6707
	2	0.9997	0.9980	0.9938	0.9865	0.9755	0.9608	0.9422	0.9201	0.8946
	3	1.0000	0.9999	0.9995	0.9986	0.9969	0.9940	0.9897	0.9837	0.9758
	4	1.0000	1.0000	1.0000	0.9999	0.9997	0.9993	0.9987	0.9976	0.9959
	5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997	0.9995
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
14	0	0.8687	0.7536	0.6528	0.5647	0.4877	0.4205	0.3620	0.3112	0.2670
	1	0.9916	0.9690	0.9355	0.8941	0.8470	0.7963	0.7436	0.6900	0.6368
	2	0.9997	0.9975	0.9923	0.9833	0.9699	0.9522	0.9302	0.9042	0.8745
	3	1.0000	0.9999	0.9994	0.9981	0.9958	0.9920	0.9864	0.9786	0.9685
	4	1.0000	1.0000	1.0000	0.9998	0.9996	0.9990	0.9980	0.9965	0.9941
	5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9992
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
15	0	0.8601	0.7386	0.6333	0.5421	0.4633	0.3953	0.3367	0.2863	0.2430
	1	0.9904	0.9647	0.9270	0.8809	0.8290	0.7738	0.7168	0.6597	0.6035
	2	0.9996	0.9970	0.9906	0.9797	0.9638	0.9429	0.9171	0.8870	0.8531
	3	1.0000	0.9998	0.9992	0.9976	0.9945	0.9896	0.9825	0.9727	0.9601
	4	1.0000	1.0000	0.9999	0.9998	0.9994	0.9986	0.9972	0.9950	0.9918
	5	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997	0.9993	0.9987
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
16	0	0.8515	0.7238	0.6143	0.5204	0.4401	0.3716	0.3131	0.2634	0.2211
	1	0.9891	0.9601	0.9182	0.8673	0.8108	0.7511	0.6902	0.6299	0.5711
	2	0.9995	0.9963	0.9887	0.9758	0.9571	0.9327	0.9031	0.8689	0.8306
	3	1.0000	0.9998	0.9989	0.9968	0.9930	0.9868	0.9779	0.9658	0.9504
	4	1.0000	1.0000	0.9999	0.9997	0.9991	0.9981	0.9962	0.9932	0.9889
	5	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9995	0.9990	0.9981
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
17	0	0.8429	0.7093	0.5958	0.4996	0.4181	0.3493	0.2912	0.2423	0.2012
	1	0.9877	0.9554	0.9091	0.8535	0.7922	0.7283	0.6638	0.6005	0.5396
	2	0.9994	0.9956	0.9866	0.9714	0.9497	0.9218	0.8882	0.8497	0.8073
	3	1.0000	0.9997	0.9986	0.9960	0.9912	0.9836	0.9727	0.9581	0.9397
	4	1.0000	1.0000	0.9999	0.9996	0.9988	0.9974	0.9949	0.9911	0.9855
	5	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9993	0.9985	0.9973
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
18	0	0.8345	0.6951	0.5780	0.4796	0.3972	0.3283	0.2708	0.2229	0.1831
	1	0.9862	0.9505	0.8997	0.8393	0.7735	0.7055	0.6378	0.5719	0.5091
	2	0.9993	0.9948	0.9843	0.9667	0.9419	0.9102	0.8725	0.8298	0.7832
	3	1.0000	0.9996	0.9982	0.9950	0.9891	0.9799	0.9667	0.9494	0.9277
	4	1.0000	1.0000	0.9998	0.9994	0.9985	0.9966	0.9933	0.9884	0.9814
	5	1.0000	1.0000	1.0000	0.9999	0.9998	0.9995	0.9990	0.9979	0.9962
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9994
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
19	0	0.8262	0.6812	0.5606	0.4604	0.3774	0.3086	0.2519	0.2051	0.1666
	1	0.9847	0.9454	0.8900	0.8249	0.7547	0.6829	0.6121	0.5440	0.4798
	2	0.9991	0.9939	0.9817	0.9616	0.9335	0.8979	0.8561	0.8092	0.7585
	3	1.0000	0.9995	0.9978	0.9939	0.9868	0.9757	0.9602	0.9398	0.9147
	4	1.0000	1.0000	0.9998	0.9993	0.9980	0.9956	0.9915	0.9853	0.9765
	5	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9986	0.9971	0.9949
	6	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9991
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
20	0	0.8179	0.6676	0.5438	0.4420	0.3585	0.2901	0.2342	0.1887	0.1516
	1	0.9831	0.9401	0.8802	0.8103	0.7358	0.6605	0.5869	0.5169	0.4516
	2	0.9990	0.9929	0.9790	0.9561	0.9245	0.8850	0.8390	0.7879	0.7334
	3	1.0000	0.9994	0.9973	0.9926	0.9841	0.9710	0.9529	0.9294	0.9007
	4	1.0000	1.0000	0.9997	0.9990	0.9974	0.9944	0.9893	0.9817	0.9710
	5	1.0000	1.0000	1.0000	0.9999	0.9997	0.9991	0.9981	0.9962	0.9932
	6	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9994	0.9987
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Πίνακας Π.7: Πίνακας Διωνυμικής κατανομής. Αν $X \sim B(n,p)$, ο πίνακας δίνει την πιθανότητα $P(X \leqslant x)$.

n	x	$p = 0.10$	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
1	0	0.9000	0.8500	0.8000	0.7500	0.7000	0.6500	0.6000	0.5500	0.5000
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	0	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
	1	0.9900	0.9775	0.9600	0.9375	0.9100	0.8775	0.8400	0.7975	0.7500
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	0	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
	1	0.9720	0.9393	0.8960	0.8438	0.7840	0.7183	0.6480	0.5748	0.5000
	2	0.9990	0.9966	0.9920	0.9844	0.9730	0.9571	0.9360	0.9089	0.8750
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	0	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
	1	0.9477	0.8905	0.8192	0.7383	0.6517	0.5630	0.4752	0.3910	0.3125
	2	0.9963	0.9880	0.9728	0.9492	0.9163	0.8735	0.8208	0.7585	0.6875
	3	0.9999	0.9995	0.9984	0.9961	0.9919	0.9850	0.9744	0.9590	0.9375
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	0	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0312
	1	0.9185	0.8352	0.7373	0.6328	0.5282	0.4284	0.3370	0.2562	0.1875
	2	0.9914	0.9734	0.9421	0.8965	0.8369	0.7648	0.6826	0.5931	0.5000
	3	0.9995	0.9978	0.9933	0.9844	0.9692	0.9460	0.9130	0.8688	0.8125
	4	1.0000	0.9999	0.9997	0.9990	0.9976	0.9947	0.9898	0.9815	0.9688
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	0	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
	1	0.8857	0.7765	0.6554	0.5339	0.4202	0.3191	0.2333	0.1636	0.1094
	2	0.9842	0.9527	0.9011	0.8306	0.7443	0.6471	0.5443	0.4415	0.3437
	3	0.9987	0.9941	0.9830	0.9624	0.9295	0.8826	0.8208	0.7447	0.6563
	4	0.9999	0.9996	0.9984	0.9954	0.9891	0.9777	0.9590	0.9308	0.8906
	5	1.0000	1.0000	0.9999	0.9998	0.9993	0.9982	0.9959	0.9917	0.9844
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	0	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
	1	0.8503	0.7166	0.5767	0.4449	0.3294	0.2338	0.1586	0.1024	0.0625
	2	0.9743	0.9262	0.8520	0.7564	0.6471	0.5323	0.4199	0.3164	0.2266
	3	0.9973	0.9879	0.9667	0.9294	0.8740	0.8002	0.7102	0.6083	0.5000
	4	0.9998	0.9988	0.9953	0.9871	0.9712	0.9444	0.9037	0.8471	0.7734
	5	1.0000	0.9999	0.9996	0.9987	0.9962	0.9910	0.9812	0.9643	0.9375
	6	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9984	0.9963	0.9922
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
8	0	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
	1	0.8131	0.6572	0.5033	0.3671	0.2553	0.1691	0.1064	0.0632	0.0352
	2	0.9619	0.8948	0.7969	0.6785	0.5518	0.4278	0.3154	0.2201	0.1445
	3	0.9950	0.9786	0.9437	0.8862	0.8059	0.7064	0.5941	0.4770	0.3633
	4	0.9996	0.9971	0.9896	0.9727	0.9420	0.8939	0.8263	0.7396	0.6367
	5	1.0000	0.9998	0.9988	0.9958	0.9887	0.9747	0.9502	0.9115	0.8555
	6	1.0000	1.0000	0.9999	0.9996	0.9987	0.9964	0.9915	0.9819	0.9648
	7	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9983	0.9961
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	0	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
	1	0.7748	0.5995	0.4362	0.3003	0.1960	0.1211	0.0705	0.0385	0.0195
	2	0.947	0.8591	0.7382	0.6007	0.4628	0.3373	0.2318	0.1495	0.08984
	3	0.9917	0.9661	0.9144	0.8343	0.7297	0.6089	0.4826	0.3614	0.2539
	4	0.9991	0.9944	0.9804	0.9511	0.9012	0.8283	0.7334	0.6214	0.5000
	5	0.9999	0.9994	0.9969	0.9900	0.9747	0.9464	0.9006	0.8342	0.7461
	6	1.0000	1.0000	0.9997	0.9987	0.9957	0.9888	0.9750	0.9502	0.9102
	7	1.0000	1.0000	1.0000	0.9999	0.9996	0.9986	0.9962	0.9909	0.9805
	8	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9980
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	0	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
	1	0.7361	0.5443	0.3758	0.2440	0.1493	0.0860	0.0464	0.0233	0.0107
	2	0.9298	0.8202	0.6778	0.5256	0.3828	0.2616	0.1673	0.0996	0.0547
	3	0.9872	0.9500	0.8791	0.7759	0.6496	0.5138	0.3823	0.2660	0.1719
	4	0.9984	0.9901	0.9672	0.9219	0.8497	0.7515	0.6331	0.5044	0.3770
	5	0.9999	0.9986	0.9936	0.9803	0.9527	0.9051	0.8338	0.7384	0.6230
	6	1.0000	0.9999	0.9991	0.9965	0.9894	0.9740	0.9452	0.8980	0.8281
	7	1.0000	1.0000	0.9999	0.9996	0.9984	0.9952	0.9877	0.9726	0.9453
	8	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983	0.9955	0.9893
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Πίνακας Π.8: Πίνακας Διωνυμικής κατανομής. Αν $X \sim B(n,p)$, ο πίνακας δίνει την πιθανότητα $P(X \leqslant x)$.

n	x	p = **0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
11	0	0.3138	0.1673	0.0859	0.0422	0.0198	0.0088	0.0036	0.0014	0.0004
	1	0.6974	0.4922	0.3221	0.1971	0.1130	0.0606	0.0302	0.0139	0.0059
	2	0.9104	0.7788	0.6174	0.4552	0.3127	0.2001	0.1189	0.0652	0.0327
	3	0.9815	0.9306	0.8389	0.7133	0.5696	0.4256	0.2963	0.1911	0.1133
	4	0.9972	0.9841	0.9496	0.8854	0.7897	0.6683	0.5328	0.3971	0.2744
	5	0.9997	0.9973	0.9883	0.9657	0.9218	0.8513	0.7535	0.6331	0.5000
	6	1.0000	0.9997	0.9980	0.9924	0.9784	0.9499	0.9006	0.8262	0.7256
	7	1.0000	1.0000	0.9998	0.9988	0.9957	0.9878	0.9707	0.9390	0.8867
	8	1.0000	1.0000	1.0000	0.9999	0.9994	0.9980	0.9941	0.9852	0.9673
	9	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9993	0.9978	0.9941
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9995
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
12	0	0.2824	0.1422	0.0687	0.0317	0.0138	0.0057	0.0022	0.0008	0.0002
	1	0.6590	0.4435	0.2749	0.1584	0.0850	0.0424	0.0196	0.0083	0.0032
	2	0.8891	0.7358	0.5583	0.3907	0.2528	0.1513	0.0834	0.0421	0.0193
	3	0.9744	0.9078	0.7946	0.6488	0.4925	0.3467	0.2253	0.1345	0.0730
	4	0.9957	0.9761	0.9274	0.8424	0.7237	0.5833	0.4382	0.3044	0.1938
	5	0.9995	0.9954	0.9806	0.9456	0.8822	0.7873	0.6652	0.5269	0.3872
	6	0.9999	0.9993	0.9961	0.9857	0.9614	0.9154	0.8418	0.7393	0.6128
	7	1.0000	0.9999	0.9994	0.9972	0.9905	0.9745	0.9427	0.8883	0.8062
	8	1.0000	1.0000	0.9999	0.9996	0.9983	0.9944	0.9847	0.9644	0.9270
	9	1.0000	1.0000	1.0000	1.0000	0.9998	0.9992	0.9972	0.9921	0.9807
	10	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9968
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
13	0	0.2542	0.1209	0.0550	0.0238	0.0097	0.0037	0.0013	0.0004	0.0001
	1	0.6213	0.3983	0.2336	0.1267	0.0637	0.0296	0.0126	0.0050	0.0017
	2	0.8661	0.6920	0.5017	0.3326	0.2025	0.1132	0.0579	0.0269	0.0112
	3	0.9658	0.8820	0.7473	0.5843	0.4206	0.2783	0.1686	0.0929	0.0461
	4	0.9935	0.9658	0.9009	0.7940	0.6543	0.5005	0.3530	0.2279	0.1334
	5	0.9991	0.9925	0.9700	0.9198	0.8346	0.7159	0.5744	0.4268	0.2905
	6	0.9999	0.9987	0.9930	0.9757	0.9376	0.8705	0.7712	0.6437	0.5000
	7	1.0000	0.9998	0.9988	0.9944	0.9818	0.9538	0.9023	0.8212	0.7095
	8	1.0000	1.0000	0.9998	0.9990	0.9960	0.9874	0.9679	0.9302	0.8666
	9	1.0000	1.0000	1.0000	0.9999	0.9993	0.9975	0.9922	0.9797	0.9539
	10	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9987	0.9959	0.9888
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
14	0	0.2288	0.1028	0.0440	0.0178	0.0068	0.0024	0.0008	0.0002	0.0000
	1	0.5846	0.3567	0.1979	0.1010	0.0475	0.0205	0.0081	0.0029	0.0009
	2	0.8416	0.6479	0.4481	0.2811	0.1608	0.0839	0.0398	0.0170	0.0065
	3	0.9559	0.8535	0.6982	0.5213	0.3552	0.2205	0.1243	0.0632	0.0287
	4	0.9908	0.9533	0.8702	0.7415	0.5842	0.4227	0.2793	0.1672	0.0898
	5	0.9985	0.9885	0.9561	0.8883	0.7805	0.6405	0.4859	0.3373	0.2120
	6	0.9998	0.9978	0.9884	0.9617	0.9067	0.8164	0.6925	0.5461	0.3953
	7	1.0000	0.9997	0.9976	0.9897	0.9685	0.9247	0.8499	0.7414	0.6047
	8	1.0000	1.0000	0.9996	0.9978	0.9917	0.9757	0.9417	0.8811	0.7880
	9	1.0000	1.0000	1.0000	0.9997	0.9983	0.9940	0.9825	0.9574	0.9102
	10	1.0000	1.0000	1.0000	1.0000	0.9998	0.9989	0.9961	0.9886	0.9713
	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9978	0.9935
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9991
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
15	0	0.2059	0.0874	0.0352	0.0134	0.0047	0.0016	0.0005	0.0001	0.0000
	1	0.5490	0.3186	0.1671	0.0802	0.0353	0.0142	0.0052	0.0017	0.0005
	2	0.8159	0.6042	0.3980	0.2361	0.1268	0.0617	0.0271	0.0107	0.0037
	3	0.9444	0.8227	0.6482	0.4613	0.2969	0.1727	0.0905	0.0424	0.0176
	4	0.9873	0.9383	0.8358	0.6865	0.5155	0.3519	0.2173	0.1204	0.0592
	5	0.9978	0.9832	0.9389	0.8516	0.7216	0.5643	0.4032	0.2608	0.1509
	6	0.9997	0.9964	0.9819	0.9434	0.8689	0.7548	0.6098	0.4522	0.3036
	7	1.0000	0.9994	0.9958	0.9827	0.9500	0.8868	0.7869	0.6535	0.5000
	8	1.0000	0.9999	0.9992	0.9958	0.9848	0.9578	0.9050	0.8182	0.6964
	9	1.0000	1.0000	0.9999	0.9992	0.9963	0.9876	0.9662	0.9231	0.8491
	10	1.0000	1.0000	1.0000	0.9999	0.9993	0.9972	0.9907	0.9745	0.9408
	11	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9981	0.9937	0.9824
	12	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9963
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Πίνακας Π.9: Πίνακας Διωνυμικής κατανομής. Αν $X \sim B(n,p)$, ο πίνακας δίνει την πιθανότητα $P(X \leqslant x)$.

n	x	p = 0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
16	0	0.1853	0.0743	0.0281	0.0100	0.0033	0.0010	0.0003	0.0000	0.0000
	1	0.5147	0.2839	0.1407	0.0635	0.0261	0.0098	0.0033	0.0010	0.0003
	2	0.7892	0.5614	0.3518	0.1971	0.0994	0.0451	0.0183	0.0066	0.0021
	3	0.9316	0.7899	0.5981	0.4050	0.2459	0.1339	0.0651	0.0281	0.0106
	4	0.9830	0.9209	0.7982	0.6302	0.4499	0.2892	0.1666	0.0853	0.0384
	5	0.9967	0.9765	0.9183	0.8103	0.6598	0.4900	0.3288	0.1976	0.1051
	6	0.9995	0.9944	0.9733	0.9204	0.8247	0.6881	0.5272	0.3660	0.2272
	7	0.9999	0.9989	0.9930	0.9729	0.9256	0.8406	0.7161	0.5629	0.4018
	8	1.0000	0.9998	0.9985	0.9925	0.9743	0.9329	0.8577	0.7441	0.5982
	9	1.0000	1.0000	0.9998	0.9984	0.9929	0.9771	0.9417	0.8759	0.7728
	10	1.0000	1.0000	1.0000	0.9997	0.9984	0.9938	0.9809	0.9514	0.8949
	11	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9951	0.9851	0.9616
	12	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9991	0.9965	0.9894
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9979
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
17	0	0.1668	0.0631	0.0230	0.0075	0.0023	0.0007	0.0002	0.0000	0.0000
	1	0.4818	0.2525	0.1182	0.0501	0.0193	0.0067	0.0021	0.0006	0.0001
	2	0.7618	0.5198	0.3096	0.1637	0.0774	0.0327	0.0123	0.0041	0.0012
	3	0.9174	0.7556	0.5489	0.3530	0.2019	0.1028	0.0464	0.0184	0.0064
	4	0.9779	0.9013	0.7582	0.5739	0.3887	0.2348	0.1260	0.0596	0.0245
	5	0.9953	0.9681	0.8943	0.7653	0.5968	0.4197	0.2639	0.1471	0.0717
	6	0.9992	0.9917	0.9623	0.8929	0.7752	0.6188	0.4478	0.2902	0.1662
	7	0.9999	0.9983	0.9891	0.9598	0.8954	0.7872	0.6405	0.4743	0.3145
	8	1.0000	0.9997	0.9974	0.9876	0.9597	0.9006	0.8011	0.6626	0.5000
	9	1.0000	1.0000	0.9995	0.9969	0.9873	0.9617	0.9081	0.8166	0.6855
	10	1.0000	1.0000	0.9999	0.9994	0.9968	0.988	0.9652	0.9174	0.8338
	11	1.0000	1.0000	1.0000	0.9999	0.9993	0.997	0.9894	0.9699	0.9283
	12	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9975	0.9914	0.9755
	13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9981	0.9936
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9988
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
18	0	0.1501	0.0536	0.0180	0.0056	0.0016	0.0004	0.0001	0.0000	0.0000
	1	0.4503	0.2241	0.0991	0.0395	0.0142	0.0046	0.0013	0.0003	0.0000
	2	0.7338	0.4797	0.2713	0.1353	0.0600	0.0236	0.0082	0.0025	0.0007
	3	0.9018	0.7202	0.5010	0.3057	0.1646	0.0783	0.0328	0.0120	0.0038
	4	0.9718	0.8794	0.7164	0.5187	0.3327	0.1886	0.0942	0.0411	0.0154
	5	0.9936	0.9581	0.8671	0.7175	0.5344	0.355	0.2088	0.1077	0.0481
	6	0.9988	0.9882	0.9487	0.8610	0.7217	0.5491	0.3743	0.2258	0.1189
	7	0.9998	0.9973	0.9837	0.9431	0.8593	0.7283	0.5634	0.3915	0.2403
	8	1.0000	0.9995	0.9957	0.9807	0.9404	0.8609	0.7368	0.5778	0.4073
	9	1.0000	0.9999	0.9991	0.9946	0.9790	0.9403	0.8653	0.7473	0.5927
	10	1.0000	1.0000	0.9998	0.9988	0.9939	0.9788	0.9424	0.872	0.7597
	11	1.0000	1.0000	1.0000	0.9998	0.9986	0.9938	0.9797	0.9463	0.8811
	12	1.0000	1.0000	1.0000	1.0000	0.9997	0.9986	0.9942	0.9817	0.9519
	13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9951	0.9846
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9990	0.9962
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9990	0.9962
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Πίνακας Π.10: Πίνακας Διωνυμικής κατανομής. Αν $X \sim B(n,p)$, ο πίνακας δίνει την πιθανότητα $P(X \leqslant x)$.

n	x	p 0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
19	0	0.1351	0.0456	0.0144	0.0042	0.0011	0.0003	0.0000	0.0000	0.0000
	1	0.4203	0.1985	0.0829	0.0310	0.0104	0.0031	0.0008	0.0002	0.0000
	2	0.7054	0.4413	0.2369	0.1113	0.0462	0.0170	0.0055	0.0015	0.0004
	3	0.8850	0.6841	0.4551	0.2631	0.1332	0.0591	0.0230	0.0077	0.0022
	4	0.9648	0.8556	0.6733	0.4654	0.2822	0.1500	0.0696	0.0280	0.0096
	5	0.9914	0.9463	0.8369	0.6678	0.4739	0.2968	0.1629	0.0777	0.0318
	6	0.9983	0.9837	0.9324	0.8251	0.6655	0.4812	0.3081	0.1727	0.0835
	7	0.9997	0.9959	0.9767	0.9225	0.8180	0.6656	0.4878	0.3169	0.1796
	8	1.0000	0.9992	0.9933	0.9713	0.9161	0.8145	0.6675	0.4940	0.3238
	9	1.0000	0.9999	0.9984	0.9911	0.9674	0.9125	0.8139	0.6710	0.5000
	10	1.0000	1.0000	0.9997	0.9977	0.9895	0.9653	0.9115	0.8159	0.6762
	11	1.0000	1.0000	1.0000	0.9995	0.9972	0.9886	0.9648	0.9129	0.8204
	12	1.0000	1.0000	1.0000	0.9999	0.9994	0.9969	0.9884	0.9658	0.9165
	13	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9969	0.9891	0.9682
	14	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9972	0.9904
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9978
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
20	0	0.1216	0.0388	0.0115	0.0032	0.0008	0.0001	0.0000	0.0000	0.0000
	1	0.3917	0.1756	0.0692	0.0243	0.0076	0.0021	0.0005	0.0001	0.0000
	2	0.6769	0.4049	0.2061	0.0913	0.0355	0.0121	0.0036	0.0009	0.0002
	3	0.8670	0.6477	0.4114	0.2252	0.1071	0.0444	0.0160	0.0049	0.0013
	4	0.9568	0.8298	0.6296	0.4148	0.2375	0.1182	0.0510	0.0189	0.0059
	5	0.9887	0.9327	0.8042	0.6172	0.4164	0.2454	0.1256	0.0553	0.0207
	6	0.9976	0.9781	0.9133	0.7858	0.6080	0.4166	0.2500	0.1299	0.0577
	7	0.9996	0.9941	0.9679	0.8982	0.7723	0.6010	0.4159	0.2520	0.1316
	8	0.9999	0.9987	0.9900	0.9591	0.8867	0.7624	0.5956	0.4143	0.2517
	9	1.0000	0.9998	0.9974	0.9861	0.9520	0.8782	0.7553	0.5914	0.4119
	10	1.0000	1.0000	0.9994	0.9961	0.9829	0.9468	0.8725	0.7507	0.5881
	11	1.0000	1.0000	0.9999	0.9991	0.9949	0.9804	0.9435	0.8692	0.7483
	12	1.0000	1.0000	1.0000	0.9998	0.9987	0.9940	0.9790	0.9420	0.8684
	13	1.0000	1.0000	1.0000	1.0000	0.9997	0.9985	0.9935	0.9786	0.9423
	14	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9984	0.9936	0.9793
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9985	0.9941
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Πίνακας Π.11: Πίνακας κρίσιμων τιμών για τον έλεγχο των Kolmogorov-Smirnov. Πηγή: Conover (1998) Table A 13.

n	$a = .20$	$a = .10$	$a = .05$	$a = .01$
1	.900	.950	.975	.995
2	.684	.776	.842	.929
3	.565	.636	.708	.829
4	.493	.565	.624	.734
5	.447	.509	.563	.669
6	.410	.468	.519	.617
7	.381	.436	.483	.576
8	.358	.410	.454	.542
9	.339	.387	.430	.513
10	.323	.369	.409	.489
11	.308	.352	.391	.468
12	.296	.338	.375	.449
13	.285	.325	.361	.432
14	.275	.314	.349	.418
15	.266	.304	.338	.404
16	.258	.295	.327	.392
17	.250	.286	.318	.381
18	.244	.279	.309	.371
19	.237	.271	.301	.361
20	.232	.265	.294	.352
21	.226	.259	.287	.344
22	.221	.253	.281	.337
23	.216	.247	.275	.330
24	.212	.242	.269	.323
25	.208	.238	.264	.317
26	.204	.233	.259	.311
27	.200	.229	.254	.305
28	.197	.225	.250	.300
29	.193	.221	.246	.295
30	.190	.218	.242	.290
31	.187	.214	.238	.285
32	.184	.211	.234	.281
33	.182	.208	.231	.277
34	.179	.205	.227	.273
35	.177	.202	.224	.269
36	.174	.199	.221	.265
37	.172	.196	.218	.262
38	.170	.194	.215	.258
39	.168	.191	.213	.255
40	.165	.189	.210	.252
$n > 40$	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Πίνακας Π.12: Πίνακας κρίσιμων τιμών της ελεγχοσυνάρτησης του Lilliefors για τον έλεγχο της κανονικότητας. Πηγή: Sheskin (2011) Table A 22.

n	$a = 0.20$	$a = 0.15$	$a = 0.10$	$a = 0.05$	$a = 0.01$
4	.300	.319	.352	.381	.417
5	.285	.299	.315	.337	.405
6	.265	.277	.294	.319	.364
7	.247	.258	.276	.300	.348
8	.233	.244	.261	.285	.331
9	.223	.233	.249	.271	.311
10	.215	.224	.239	.258	.294
11	.206	.217	.230	.249	.284
12	.199	.212	.223	.242	.275
13	.190	.202	.214	.234	.268
14	.183	.194	.207	.227	.261
15	.177	.187	.201	.220	.257
16	.173	.182	.195	.213	.250
17	.169	.177	.189	.206	.245
18	.166	.173	.184	.200	.239
19	.163	.169	.179	.195	.235
20	.160	.166	.174	.190	.231
25	.142	.147	.158	.173	.200
30	.131	.136	.144	.161	.187
$n > 30$	$\frac{.736}{\sqrt{n}}$	$\frac{.768}{\sqrt{n}}$	$\frac{.805}{\sqrt{n}}$	$\frac{.886}{\sqrt{n}}$	$\frac{1.031}{\sqrt{n}}$

Πίνακας Π.13: Πίνακας κρίσιμων τιμών της ελεγχοσυνάρτησης του Lilliefors για τον έλεγχο της εκθετικής κατανομής. Πηγή: Conover (1998) Table A 15.

Πηγή: Conover (1998) Table A 13

n	.95	.90	.80	.70	.50	.30	.20	.10	.05	.01	.001
2	.3127	.3200	.3337	.3617	.4337	.5034	.5507	.5934	.6133	.6284	.6317
3	.2299	.2544	.2899	.3166	.3645	.4122	.4508	.5111	.5508	.6003	.6296
4	.2072	.2281	.2545	.2766	.3163	.3685	.4007	.4442	.4844	.5574	.6215
5	.1884	.2052	.2290	.2483	.2877	.3317	.3603	.4045	.4420	.5127	.5814
6	.1726	.1882	.2102	.2290	.2645	.3045	.3320	.3732	.4085	.4748	.5497
7	.1604	.1750	.1961	.2136	.2458	.2838	.3098	.3481	.3811	.4459	.5181
8	.1506	.1646	.1845	.2006	.2309	.2671	.2914	.3274	.3590	.4208	.4913
9	.1426	.1561	.1746	.1897	.2186	.2529	.2758	.3101	.3404	.3995	.4679
10	.1359	.1486	.1661	.1805	.2082	.2407	.2626	.2955	.3244	.3813	.4473
12	.1249	.1364	.1524	.1657	.1912	.2209	.2411	.2714	.2981	.3511	.4132
14	.1162	.1268	.1418	.1542	.1778	.2054	.2242	.2525	.2774	.3272	.3858
16	.1091	.1191	.1332	.1448	.1669	.1929	.2105	.2371	.2606	.3076	.3632
18	.1032	.1127	.1260	.1369	.1578	.1824	.1990	.2242	.2465	.2911	.3441
20	.0982	.1073	.1199	.1303	.1501	.1735	.1893	.2132	.2345	.2771	.3277
22	.0939	.1025	.1146	.1245	.1434	.1657	.1809	.2038	.2241	.2649	.3135
24	.0901	.0984	.1099	.1195	.1376	.1590	.1735	.1954	.2150	.2542	.3010
26	.0868	.0947	.1058	.1150	.1324	.1530	.1670	.1881	.2069	.2447	.2899
28	.1838	.0914	.1021	.1110	.1278	.1477	.1611	.1815	.1997	.2362	.2799
30	.0811	.0885	.0988	.1074	.1236	.1428	.1559	.1756	.1932	.2286	.2709
35	.0754	.0822	.0918	.0997	.1148	.1326	.1447	.1630	.1793	.2123	.2517
40	.0707	.0771	.0861	.0935	.1077	.1243	.1356	.1528	.1681	.1990	.2361
45	.0668	.0729	.0814	.0884	.1017	.1174	.1281	.1443	.1588	.1880	.2231
50	.0636	.0693	.0774	.0840	.0966	.1116	.1217	.1371	.1509	.1787	.2121
60	.0582	.0635	.0708	.0769	.0885	.1021	.1114	.1255	.1381	.1635	.1943
70	.0541	.0589	.0658	.0714	.0821	.0946	.1033	.1164	.1281	.1517	-
80	.0507	.0553	.0616	.0669	.0769	.0887	.0968	.1090	.1200	.1421	-
90	.0479	.0522	.0582	.0632	.0726	.0838	.0914	.1029	.1132	.1341	-
100	.0455	.0496	.0553	.0600	.0690	.0796	.0868	.0977	.1075	.1274	-
$n > 100$	$\frac{.4550}{\sqrt{n}}$	$\frac{.4959}{\sqrt{n}}$	$\frac{.5530}{\sqrt{n}}$	$\frac{.6000}{\sqrt{n}}$	$\frac{.6898}{\sqrt{n}}$	$\frac{.7957}{\sqrt{n}}$	$\frac{.8678}{\sqrt{n}}$	$\frac{.9773}{\sqrt{n}}$	$\frac{.1.0753}{\sqrt{n}}$	$\frac{1.2743}{\sqrt{n}}$	$\frac{1.5010}{\sqrt{n}}$

Πίνακας Π.14: Πίνακας κρίσιμων τιμών της ελεγχοσυνάρτησης του Smirnov για δύο ισομεγέθη δείγματα. Πηγή: Conover (1998) Table A 19.

n	$a = .20$	$a = .10$	$a = .05$	$a = .02$
3	2/3	2/3		
4	3/4	3/4	3/4	
5	3/5	3/5	4/5	4/5
6	3/6	4/6	4/6	5/6
7	4/7	4/7	5/7	5/7
8	4/8	4/8	5/8	5/8
9	4/9	5/9	5/9	6/9
10	4/10	5/10	6/10	6/10
11	5/11	5/11	6/11	7/11
12	5/12	5/12	6/12	7/12
13	5/13	6/13	6/13	7/13
14	5/14	6/14	7/14	7/14
15	5/15	6/15	7/15	8/15
16	6/16	6/16	7/16	8/16
17	6/17	7/17	7/17	8/17
18	6/18	7/18	8/17	9/17
19	6/19	7/19	8/19	9/19
20	6/20	7/20	8/20	9/20
21	6/21	7/21	8/21	9/21
22	7/22	8/22	8/22	10/22
23	7/23	8/23	9/23	10/23
24	7/24	8/24	9/24	10/24
25	7/25	8/25	9/25	10/25
26	7/26	8/26	9/26	10/26
27	7/27	8/27	9/27	11/27
28	8/28	9/28	10/28	11/28
29	8/29	9/29	10/29	11/29
30	8/30	9/30	10/30	11/30
31	8/31	9/31	10/31	11/31
32	8/32	9/32	10/32	12/32
34	8/34	10/34	11/34	12/34
36	9/36	10/36	11/36	12/36
38	9/38	10/38	11/38	13/38
40	9/40	10/40	12/40	13/40
$n > 40$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.73}{\sqrt{n}}$	$\frac{1.92}{\sqrt{n}}$	$\frac{2.15}{\sqrt{n}}$

Πίνακας Π.15: Πίνακας κρίσιμων τιμών της ελεγχοσυνάρτησης του Smirnov για δύο ανισομεγέθη δείγματα. Πηγή: Conover (1998) Table A 20.

m	n	$a = .20$	$a = .10$	$a = .05$	$a = .02$
1	9	17/18			
	10	9/10			
2	3	5/6			
	4	3/4			
	5	4/5	4/5		
	6	5/6	5/6		
	7	5/7	6/7		
	8	3/4	7/8	7/8	
	9	7/9	8/9	8/9	
	10	7/10	4/5	9/10	
3	4	3/4	3/4		
	5	2/3	4/5	4/5	
	6	2/3	2/3	5/6	
	7	2/3	5/7	6/7	6/7
	8	5/8	3/4	3/4	7/8
	9	2/3	2/3	7/9	8/9
	10	3/5	7/10	4/5	9/10
	12	7/12	2/3	3/4	5/6
4	5	3/5	3/4	4/5	4/5
	6	7/12	2/3	3/4	5/6
	7	17/28	5/7	3/4	6/7
	8	5/8	5/8	3/4	7/8
	9	5/9	2/3	3/4	7/9
	10	11/20	13/20	7/10	4/5
	12	7/12	2/3	2/3	3/4
	16	9/16	5/8	11/16	3/4
5	6	3/5	2/3	2/3	5/6
	7	4/7	23/35	5/7	29/35
	8	11/20	5/8	27/40	4/5
	9	5/9	3/5	31/45	7/9
	10	1/2	3/5	7/10	7/10
	15	8/15	3/5	2/3	11/15
	20	1/2	11/20	3/5	7/10
6	7	23/42	4/7	29/42	5/7
	8	1/2	7/12	2/3	3/4
	9	1/2	5/9	2/3	13/18
	10	1/2	17/30	19/30	7/10
	12	1/2	7/12	7/12	2/3
	18	4/9	5/9	11/18	2/3
	24	11/24	1/2	7/12	5/8
7	8	27/56	33/56	5/8	41/56
	9	31/63	5/9	40/63	5/7
	10	33/70	39/70	43/70	7/10
	14	3/7	1/2	4/7	9/14
	28	3/7	13/28	15/28	17/28
8	9	4/9	13/24	5/8	2/3
	10	19/40	21/40	23/40	27/40
	12	11/24	1/2	7/12	5/8
	16	7/16	1/2	9/16	5/8
	32	13/32	7/16	1/2	9/16
9	10	7/15	1/2	26/45	2/3
	12	4/9	1/2	5/9	11/18
	15	19/45	22/45	8/15	3/5
	18	7/18	4/9	1/2	5/9
	36	13/36	5/12	17/36	19/36
10	15	2/5	7/15	1/2	17/30
	20	2/5	9/20	1/2	11/20
	40	7/20	2/5	9/20	1/2
12	15	23/60	9/20	1/2	11/20
	16	3/8	7/16	23/48	13/24
	18	13/36	5/12	17/36	19/36
	20	11/30	5/12	7/15	31/60
15	7	7/20	2/5	13/30	29/60
16	20	27/80	31/80	17/40	19/40
Μεγαλύτερα n,m		$1.07\sqrt{\frac{m+n}{mn}}$	$1.22\sqrt{\frac{m+n}{mn}}$	$1.36\sqrt{\frac{m+n}{mn}}$	$1.52\sqrt{\frac{m+n}{mn}}$

Πίνακας Π.16: Κρίσιμες τιμές w_{1-a} του Shapiro-Wilk κριτηρίου. Πηγή: Conover (1998) Table A 17.

Μέγεθος δείγματος	Επίπεδο σημαντικότητας a			
n	1%	2%	5%	10%
3	0.753	0.756	0.767	0.789
4	0.687	0.707	0.748	0.792
5	0.686	0.715	0.762	0.806
6	0.713	0.743	0.788	0.826
7	0.730	0.760	0.803	0.838
8	0.749	0.778	0.818	0.851
9	0.764	0.791	0.829	0.859
10	0.781	0.806	0.842	0.869
11	0.792	0.817	0.850	0.876
12	0.805	0.828	0.859	0.883
13	0.814	0.837	0.866	0.889
14	0.825	0.846	0.874	0.895
15	0.835	0.855	0.881	0.901
16	0.844	0.863	0.887	0.906
17	0.851	0.869	0.892	0.910
18	0.858	0.874	0.897	0.914
19	0.863	0.879	0.901	0.917
20	0.868	0.884	0.905	0.920
21	0.873	0.888	0.908	0.923
22	0.878	0.892	0.911	0.926
23	0.881	0.895	0.914	0.928
24	0.884	0.898	0.916	0.930
25	0.888	0.901	0.918	0.931
26	0.891	0.904	0.921	0.933
27	0.894	0.906	0.923	0.935
28	0.896	0.908	0.924	0.936
29	0.898	0.910	0.926	0.937
30	0.900	0.912	0.927	0.939
31	0.902	0.914	0.929	0.940
32	0.904	0.915	0.930	0.941
33	0.906	0.917	0.931	0.942
34	0.908	0.919	0.933	0.943
35	0.910	0.920	0.934	0.944
36	0.912	0.922	0.935	0.945
37	0.914	0.924	0.936	0.946
38	0.916	0.925	0.938	0.947
39	0.917	0.927	0.939	0.948
40	0.919	0.928	0.940	0.949
41	0.920	0.929	0.941	0.950
42	0.922	0.930	0.942	0.951
43	0.923	0.932	0.943	0.951
44	0.924	0.933	0.944	0.952
45	0.926	0.934	0.945	0.953
46	0.927	0.935	0.945	0.953
47	0.928	0.936	0.946	0.954
48	0.929	0.937	0.947	0.954
49	0.929	0.937	0.947	0.955
50	0.930	0.938	0.947	0.955

Πίνακας Π.17: Ποσοστιαία σημεία για τον προσημικό έλεγχο τάξης Wilcoxon. Πηγή: Conover (1998) Table A 12.

n	$T_{n,0.995}$	$T_{n,0.99}$	$T_{n,0.975}$	$T_{n,0.95}$	$T_{n,0.90}$	$T_{n,0.80}$	$T_{n,0.70}$	$T_{n,0.60}$	$T_{n,0.50}$	$\frac{n(n+1)}{2}$
4	0	0	0	0	1	3	3	4	5	10
5	0	0	0	1	3	4	5	6	7.5	15
6	0	0	1	3	4	6	8	9	10.5	21
7	0	1	3	4	6	9	11	12	14	28
8	1	2	4	6	9	12	14	16	18	36
9	2	4	6	9	11	15	18	20	22.5	45
10	4	6	9	11	15	19	22	25	27.5	55
11	6	8	11	14	18	23	27	30	33	66
12	8	10	14	18	22	28	32	36	39	78
13	10	13	18	22	27	33	38	42	45.5	91
14	13	16	22	26	32	39	44	48	52.5	105
15	16	20	26	31	37	45	51	55	60	120
16	20	24	30	36	43	51	58	63	68	136
17	24	28	35	42	49	58	65	71	76.5	153
18	28	33	41	48	56	66	73	80	85.5	171
19	33	38	47	54	63	74	82	89	95	190
20	38	44	53	61	70	83	91	98	105	210
21	44	50	59	68	78	91	100	108	115.5	231
22	49	56	67	76	87	100	110	119	126.5	253
23	55	63	74	84	95	110	120	130	138	276
24	62	70	82	92	105	120	131	141	150	300
25	69	77	90	101	114	131	143	153	162.55	325
26	76	85	99	111	125	142	155	165	175.5	351
27	84	94	108	120	135	154	167	178	189	378
28	92	102	117	131	146	166	180	192	203	406
29	101	111	127	141	158	178	193	206	217.5	435
30	110	121	138	152	170	191	207	220	232.5	465

Πίνακας Π.18: Κρίσιμες τιμές για τον δίπλευρο έλεγχο του Wilcoxon Signed Rank Test σε επίπεδο σημαντικότητας a. Πηγή: Sheskin (2011) Table A 5.

	ε.σ. για μονόπλευρο έλεγχο			
	0.05	0.025	0.01	0.005
	ε.σ. για δίπλευρο έλεγχο			
n	0.10	0.05	0.02	0.01
5	0			
6	2	0		
7	3	2	0	
8	5	3	1	0
9	8	5	3	1
10	10	8	5	3
11	13	10	7	5
12	17	13	9	7
13	21	17	12	9
14	25	21	15	12
15	30	25	19	15
16	35	29	23	19
17	41	34	27	23
18	47	40	32	27
19	53	46	37	32
20	60	52	43	37
21	67	58	49	42
22	75	65	55	48
23	89	73	62	54
24	91	81	69	61
25	100	89	76	68
26	110	98	84	75
27	119	107	92	83
28	130	116	101	91
29	140	126	110	100
30	151	137	120	109

Πίνακας Π.19: Υπολογισμός πιθανοτήτων της μορφής $P(U \leq u)$, όπου u η παρατηρούμενη τιμή του τεστ των Wilcoxon-Mann-Whitney για $n_1 = 3$. Πηγή: Παπαϊωάννου και Λουκάς (2002).

$U \backslash n_2$	1	2	3
0	0.250	0.100	0.050
1	0.500	0.200	0.100
2	0.750	0.400	0.200
3		0.600	0.350
4			0.500
5			0.650

Πίνακας Π.20: Υπολογισμός πιθανοτήτων της μορφής $P(U \leq u)$, όπου u η παρατηρούμενη τιμή του τεστ των Wilcoxon-Mann-Whitney για $n_1 = 4$. Πηγή: Παπαϊωάννου και Λουκάς (2002).

$U \backslash n_2$	1	2	3	4
0	0.200	0.067	0.028	0.014
1	0.400	0.133	0.057	0.029
2	0.600	0.267	0.114	0.057
3		0.400	0.200	0.100
4		0.600	0.314	0.171
5			0.429	0.343
6			0.571	0.343
7				0.443
8				0.557

Πίνακας Π.21: Υπολογισμός πιθανοτήτων της μορφής $P(U \leq u)$, όπου u η παρατηρούμενη τιμή του τεστ των Wilcoxon-Mann-Whitney για $n_1 = 5$. Πηγή: Παπαϊωάννου και Λουκάς (2002).

$U \backslash n_2$	1	2	3	4	5
0	0.167	0.047	0.018	0.008	0.004
1	0.333	0.095	0.036	0.016	0.008
2	0.500	0.190	0.071	0.032	0.016
3	0.667	0.286	0.125	0.056	0.028
4		0.429	0.196	0.095	0.048
5		0.571	0.286	0.143	0.075
6			0.393	0.206	0.111
7			0.500	0.278	0.155
8			0.607	0.365	0.210
9				0.452	0.274
10				0.548	0.345
11					0.421
12					0.500
13					0.579

Πίνακας Π.22: Κρίσιμες τιμές του Mann-Whitney με επίπεδο σημαντικότητας 5% για δίπλευρο έλεγχο ή 0.025 για μονόπλευρο, με n_1 και n_2 το μέγεθος του μικρότερου και μεγαλύτερου δείγματος. Πηγή: Sheskin (2011) Table A 11.

	n_2															
n_1	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6		5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7			8	10	12	14	16	18	20	22	24	26	28	30	32	34
8				13	15	17	19	22	24	26	29	31	34	36	38	41
9					17	20	23	26	28	31	34	37	39	42	45	48
10						23	26	29	33	36	39	42	45	48	52	55
11							30	33	37	40	44	47	51	55	58	62
12								37	41	45	49	53	57	61	65	69
13									45	50	54	59	63	67	72	76
14										55	59	64	67	74	78	83
15											64	70	75	80	85	90
16												75	81	86	92	98
17													87	93	99	105
18														99	106	112
19															113	119
20																127

Πίνακας Π.23: Κάτω p ποσοστιαία σημεία της στατιστικής συνάρτησης R_1 για τον έλεγχο των Mann-Whitney, $n_1 = 2,3,...,11$, $n_2 = 2,3,...,11$. Πηγή: Ξεκαλάκη (2001), Πίνακας 9, σελ. 776-780 .

	p	$n_2=2$	3	4	5	6	7	8	9	10	11
$n_1=2$	0.001	3	3	3	3	3	3	3	3	3	3
	0.005	3	3	3	3	3	3	3	3	3	3
	0.010	3	3	3	3	3	3	3	3	3	3
	0.025	3	3	3	3	3	3	4	4	4	5
	0.050	3	3	3	4	4	4	5	5	5	5
	0.100	3	4	4	5	5	5	6	6	7	7
$n_1=3$	0.001	6	6	6	6	6	6	6	6	6	6
	0.005	6	6	6	6	6	6	6	7	7	7
	0.010	6	6	6	6	6	7	7	8	8	8
	0.025	6	6	6	7	8	8	9	9	10	10
	0.050	6	7	7	8	9	9	10	11	11	12
	0.100	7	8	8	9	10	11	12	12	13	14
$n_1=4$	0.001	10	10	10	10	10	10	10	10	11	11
	0.005	10	10	10	10	11	11	12	12	13	13
	0.010	10	10	10	11	12	12	13	14	14	15
	0.025	10	10	11	12	13	14	15	15	16	17
	0.050	10	11	12	13	14	15	16	17	18	19
	0.100	11	12	14	15	16	17	18	20	21	22
$n_1=5$	0.001	15	15	15	15	15	15	16	17	17	18
	0.005	15	15	15	16	17	17	18	19	20	21
	0.010	15	15	16	17	18	19	20	21	22	23
	0.025	15	16	17	18	19	21	22	23	24	25
	0.050	16	17	18	20	21	22	24	25	27	28
	0.100	17	18	20	21	23	24	26	28	29	31
$n_1=6$	0.001	21	21	21	21	21	21	23	24	25	26
	0.005	21	21	22	23	24	25	26	27	28	29
	0.010	21	21	23	24	25	26	28	29	30	31
	0.025	21	23	24	25	27	28	30	32	33	35
	0.050	22	24	25	27	29	30	32	34	36	38
	0.100	23	25	27	29	31	33	35	37	39	41
$n_1=7$	0.001	28	28	28	28	29	30	31	32	34	35
	0.005	28	28	29	30	32	33	35	36	38	39
	0.010	28	29	30	32	33	35	36	38	40	41
	0.025	28	30	32	34	35	37	39	41	43	45
	0.050	29	31	33	35	37	40	42	44	46	48
	0.100	30	33	35	37	40	42	45	47	50	52
$n_1=8$	0.001	36	36	36	37	38	39	41	42	43	45
	0.005	36	36	38	39	41	43	44	46	48	50
	0.010	36	37	39	41	43	44	46	48	50	52
	0.025	37	39	41	43	45	47	50	52	54	56
	0.050	38	40	42	45	47	50	52	55	57	60
	0.100	39	42	44	47	50	53	56	59	61	64
$n_1=9$	0.001	45	45	45	47	48	49	51	53	54	56
	0.005	45	46	47	49	51	53	55	57	59	62
	0.010	45	47	49	51	53	55	57	60	62	64
	0.025	46	48	50	53	56	58	61	63	66	69
	0.050	47	50	52	55	58	61	64	67	70	73
	0.100	48	51	55	58	61	64	68	71	74	77
$n_1=10$	0.001	55	55	56	57	59	61	62	64	66	68
	0.005	55	56	58	60	62	65	67	69	72	74
	0.010	55	57	59	62	64	67	69	72	75	78
	0.025	56	59	61	64	67	70	73	76	79	82
	0.050	57	60	63	67	70	73	76	80	83	87
	0.100	59	62	66	69	73	77	80	84	88	92
$n_1=11$	0.001	66	66	67	69	71	73	75	77	79	82
	0.005	66	67	69	72	74	77	80	83	85	88
	0.010	66	68	71	74	76	79	82	85	89	92
	0.025	67	70	73	76	80	83	86	90	93	97
	0.050	68	72	75	79	83	86	90	94	98	101
	0.100	70	74	78	82	86	90	94	98	103	107

Πίνακας Π.24: Κάτω p ποσοστιαία σημεία της στατιστικής συνάρτησης R_1 για τον έλεγχο των Mann-Whitney, $n_1 = 12,13,...,20$, $n_2 = 2,3,...,11$. Πηγή: Ξεκαλάκη (2001), Πίνακας 9, σελ. 776-780 .

	p	$n_2=2$	3	4	5	6	7	8	9	10	11
$n_1=12$	0.001	78	78	79	81	83	86	88	91	93	96
	0.005	78	80	82	85	88	91	94	97	100	103
	0.010	78	81	84	87	90	93	96	100	103	107
	0.025	80	83	86	90	93	97	101	105	108	112
	0.050	81	84	88	92	96	100	105	109	111	117
	0.100	83	87	91	96	100	105	109	114	118	123
$n_1=13$	0.001	91	91	93	95	97	100	103	106	109	112
	0.005	91	93	95	99	102	105	109	112	116	119
	0.010	92	94	97	101	104	108	112	115	119	123
	0.025	93	96	100	104	108	112	116	120	125	129
	0.050	94	98	102	107	111	116	120	125	129	134
	0.100	96	101	105	110	115	120	125	130	135	140
$n_1=14$	0.001	105	105	107	109	112	115	118	121	125	128
	0.005	105	107	110	113	117	121	124	128	132	136
	0.010	106	108	112	116	119	123	128	132	136	140
	0.025	107	111	115	119	123	128	132	137	142	146
	0.050	109	113	117	122	127	132	137	142	147	152
	0.100	110	116	121	126	131	137	142	147	153	158
$n_1=15$	0.001	120	120	122	125	128	133	135	138	142	145
	0.005	120	123	126	129	133	137	141	145	150	154
	0.010	121	124	128	132	136	140	145	149	154	158
	0.025	122	126	131	135	140	145	150	155	160	165
	0.050	124	128	133	139	144	149	154	160	165	171
	0.100	126	131	137	143	148	154	160	166	172	178
$n_1=16$	0.001	136	136	139	142	145	148	152	156	160	164
	0.005	136	139	142	146	150	155	159	164	168	173
	0.010	137	140	144	149	153	158	163	168	173	178
	0.025	138	143	148	152	158	163	168	174	179	184
	0.050	140	145	151	156	162	167	173	179	185	191
	0.100	142	148	154	160	166	173	179	185	191	198
$n_1=17$	0.001	153	154	156	159	163	167	171	175	179	183
	0.005	153	156	160	164	169	173	178	183	188	193
	0.010	154	158	162	167	172	177	182	187	192	198
	0.025	156	160	165	171	176	182	188	193	199	205
	0.050	157	163	169	174	180	187	193	199	205	211
	0.100	160	166	172	179	185	192	199	206	212	219
$n_1=18$	0.001	171	172	175	178	182	186	190	195	199	204
	0.005	171	174	178	183	188	193	198	203	209	214
	0.010	172	176	181	186	191	196	202	208	213	219
	0.025	174	179	184	190	196	202	208	214	220	227
	0.050	176	181	188	194	200	207	213	220	227	233
	0.100	178	185	192	199	206	213	220	227	234	241
$n_1=19$	0.001	190	191	194	198	202	206	211	216	220	225
	0.005	191	194	198	203	208	213	219	224	230	236
	0.010	192	195	200	206	211	217	223	229	235	241
	0.025	193	198	204	210	216	223	229	236	243	249
	0.050	195	201	208	214	221	228	235	242	249	256
	0.100	198	205	212	219	227	234	242	249	257	264
$n_1=20$	0.001	210	211	214	218	223	227	232	237	243	248
	0.005	211	214	219	224	229	235	241	247	253	259
	0.010	212	216	221	227	233	239	245	251	258	264
	0.025	213	219	225	231	238	245	251	259	266	273
	0.050	215	222	229	236	243	250	258	265	273	280
	0.100	218	226	233	241	249	257	265	273	281	289

Πίνακας Π.25: Κάτω p ποσοστιαία σημεία της στατιστικής συνάρτησης R_1 για τον έλεγχο των Mann-Whitney, $n_1 = 2,3,...,11$, $n_2 = 12,13,...,20$. Πηγή: Ξεκαλάκη (2001), Πίνακας 9, σελ. 776-780 .

	p	$n_2 = 12$	13	14	15	16	17	18	19	20
$n_1 = 2$	0.001	3	3	3	3	3	3	3	3	3
	0.005	3	3	3	3	3	3	3	4	4
	0.010	3	4	4	4	4	4	4	5	5
	0.025	5	5	5	5	5	6	6	6	6
	0.050	6	6	7	7	7	7	8	8	8
	0.100	8	8	8	9	9	10	10	11	11
$n_1 = 3$	0.001	6	6	6	6	6	7	7	7	7
	0.005	8	8	8	9	9	9	9	10	10
	0.010	9	9	9	10	10	11	11	11	12
	0.025	11	11	12	12	13	13	14	14	15
	0.050	12	13	14	14	15	16	16	17	18
	0.100	15	16	17	17	18	19	20	21	22
$n_1 = 4$	0.001	11	12	12	12	13	13	14	14	14
	0.005	14	14	15	16	16	17	17	18	19
	0.010	16	16	17	18	18	19	20	20	21
	0.025	18	19	20	21	22	22	23	24	25
	0.050	20	21	22	23	25	26	27	28	29
	0.100	23	24	26	27	28	29	31	32	33
$n_1 = 5$	0.001	18	19	19	20	21	21	22	23	23
	0.005	22	23	23	24	25	26	27	28	29
	0.010	24	25	26	27	28	29	30	31	32
	0.025	27	28	29	30	31	33	34	35	36
	0.050	29	31	32	34	35	36	38	39	41
	0.100	33	34	36	38	39	41	43	44	46
$n_1 = 6$	0.001	26	27	28	29	30	31	32	33	34
	0.005	31	32	33	34	35	37	38	39	40
	0.010	33	34	35	37	38	40	41	42	44
	0.025	36	38	39	41	43	44	46	47	49
	0.050	39	41	43	45	47	48	50	52	54
	0.100	43	45	47	49	51	53	56	58	60
$n_1 = 7$	0.001	36	37	38	39	40	42	43	44	45
	0.005	41	42	44	45	47	48	50	51	53
	0.010	43	45	46	48	50	52	53	55	57
	0.025	47	49	51	53	55	57	59	61	63
	0.050	50	53	55	57	59	62	64	66	68
	0.100	55	57	60	62	65	67	70	72	75
$n_1 = 8$	0.001	46	48	49	51	52	54	55	57	58
	0.005	52	54	55	57	59	61	63	65	67
	0.010	54	56	59	61	63	65	67	69	71
	0.025	59	61	63	66	68	71	73	75	78
	0.050	63	65	68	70	73	76	78	81	84
	0.100	67	70	73	76	79	82	85	88	91
$n_1 = 9$	0.001	58	60	61	63	65	67	69	71	72
	0.005	64	66	68	70	73	75	77	79	82
	0.010	67	69	72	74	77	79	82	84	86
	0.025	72	74	77	80	83	85	88	91	94
	0.050	76	79	82	85	88	91	94	97	100
	0.100	81	84	87	91	94	98	101	104	108
$n_1 = 10$	0.001	70	73	75	77	79	81	83	85	88
	0.005	77	80	82	85	87	90	93	95	98
	0.010	80	83	86	89	92	94	97	100	103
	0.025	85	89	92	95	98	101	104	108	111
	0.050	90	93	97	100	104	107	111	114	118
	0.100	95	99	103	107	110	114	118	122	126
$n_1 = 11$	0.001	84	87	89	91	94	96	99	101	104
	0.005	91	94	97	100	103	106	109	112	115
	0.010	95	98	101	104	108	111	114	117	120
	0.025	100	104	107	111	114	118	122	125	129
	0.050	105	109	113	117	121	124	128	132	136
	0.100	111	115	119	124	128	132	136	140	145

Πίνακας Π.26: Κάτω p ποσοστιαία σημεία της στατιστικής συνάρτησης R_1 για τον έλεγχο των Mann-Whitney, $n_1 = 12,13,...,20$, $n_2 = 12,13,...,20$. Πηγή: Ξεκαλάκη (2001), Πίνακας 9, σελ. 776-780 .

	p	$n_2 = 12$	13	14	15	16	17	18	19	20
$n_1 = 12$	0.001	98	102	104	106	110	113	116	118	121
	0.005	106	110	113	116	120	123	126	130	133
	0.010	110	114	117	121	125	128	132	135	139
	0.025	116	120	124	128	132	136	140	144	148
	0.050	121	126	130	134	139	143	147	151	156
	0.100	128	132	137	142	146	151	156	160	165
$n_1 = 13$	0.001	115	118	121	124	127	130	134	137	140
	0.005	123	126	130	134	137	141	145	149	152
	0.010	127	131	135	139	143	147	151	155	159
	0.025	133	137	142	146	151	155	159	164	168
	0.050	139	143	148	153	157	162	167	172	176
	0.100	145	150	155	160	166	171	176	181	186
$n_1 = 14$	0.001	131	135	138	142	145	149	152	156	160
	0.005	140	144	148	152	156	160	164	169	173
	0.010	144	149	153	157	162	166	171	175	179
	0.025	151	156	161	165	170	175	180	184	189
	0.050	157	162	167	172	177	183	188	193	198
	0.100	164	169	175	180	186	191	197	203	208
$n_1 = 15$	0.001	149	153	157	161	164	168	172	176	180
	0.005	158	163	167	172	176	181	185	190	194
	0.010	163	168	172	177	182	187	191	196	201
	0.025	170	175	180	185	191	196	201	206	211
	0.050	176	182	187	193	198	204	209	215	221
	0.100	184	189	195	201	207	213	219	225	231
$n_1 = 16$	0.001	168	172	176	180	185	189	193	197	202
	0.005	178	182	187	192	197	202	207	211	216
	0.010	183	188	193	198	203	208	213	219	224
	0.025	190	196	201	207	212	218	223	229	235
	0.050	197	202	208	214	220	226	232	238	244
	0.100	204	211	217	223	230	236	243	249	256
$n_1 = 17$	0.001	188	192	197	201	206	211	215	220	224
	0.005	198	203	208	214	219	224	229	235	240
	0.010	203	209	214	220	225	231	236	242	247
	0.025	211	217	223	229	235	241	247	253	259
	0.050	218	224	231	237	243	250	256	263	269
	0.100	226	233	239	246	253	260	267	274	281
$n_1 = 18$	0.001	209	214	218	223	228	233	238	243	248
	0.005	219	225	230	236	242	247	253	259	264
	0.010	225	231	237	242	248	254	260	266	272
	0.025	233	239	246	252	258	265	271	278	284
	0.050	240	247	254	260	267	274	281	288	295
	0.100	249	256	263	270	278	285	292	300	307
$n_1 = 19$	0.001	231	236	241	246	251	257	262	268	273
	0.005	242	248	254	260	265	272	278	284	290
	0.010	247	254	260	266	273	279	285	292	298
	0.025	256	263	269	276	283	290	297	304	310
	0.050	263	271	278	285	292	300	307	314	321
	0.100	272	280	288	295	303	311	319	326	334
$n_1 = 20$	0.001	253	259	265	270	276	281	287	293	299
	0.005	265	271	278	284	290	297	303	310	316
	0.010	271	278	284	291	298	304	311	318	325
	0.025	280	287	294	301	309	316	323	330	338
	0.050	288	295	303	311	318	326	334	341	349
	0.100	297	305	313	321	330	338	346	354	362

Πίνακας Π.27: Ποσοστιαία σημεία της στατιστικής συνάρτησης του ελέγχου των τετραγώνων τάξεως μεγέθους (Squared Rank Test). Πηγή: Conover (1998) Table A 9.

n_1		$n_2=3$	$n_2=4$	$n_2=5$	$n_2=6$	$n_2=7$	$n_2=8$	$n_2=9$	$n_2=10$
3	a=0.995	14	14	14	14	14	14	21	21
	a=0.99	14	14	14	14	21	21	26	26
	a=0.975	14	14	21	26	29	30	35	41
	a=0.95	21	21	26	30	38	42	49	54
	a=0.90	26	29	35	42	50	59	69	77
	a=0.10	65	90	1 17	149	182	221	260	305
	a=0.05	70	101	129	161	197	238	285	333
	a=0.025	77	110	138	170	213	257	308	362
	a=0.01	77	110	149	194	230	285	329	394
	a=0.005	77	110	149	194	245	302	346	413
4	a=0.995	30	30	30	39	39	46	50	54
	a=0.99	30	30	39	46	50	51	62	66
	a=0.975	30	39	50	54	63	71	78	90
	a=0.95	39	50	57	66	78	90	102	114
	a=0.90	50	62	71	85	99	114	130	149
	a=0.10	111	142	182	222	270	321	375	435
	a=0.05	119	154	197	246	294	350	413	476
	a=0.025	126	165	206	255	311	374	439	510
	a=0.01	126	174	219	270	334	401	470	545
	a=0.005	126	174	230	281	351	414	494	567
5	a=0.995	55	55	66	75	79	88	99	110
	a=0.99	55	66	75	82	90	103	115	127
	a=0.975	66	79	88	100	114	131	145	162
	a=0.95	75	88	103	120	135	155	175	195
	a=0.90	87	103	121	142	163	187	212	239
	a=0.10	169	214	264	319	379	445	514	591
	a=0.05	178	228	282	342	410	479	558	639
	a=0.025	183	235	297	363	433	508	592	680
	a=0.01	190	246	310	382	459	543	631	727
	a=0.005	190	255	319	391	478	559	654	754
6	a=0.995	91	104	115	124	136	152	167	182
	a=0.99	91	115	124	139	155	175	191	210
	a=0.975	115	130	143	164	184	208	231	255
	a=0.95	124	139	164	187	211	239	268	299
	a=0.90	136	163	187	215	247	280	315	352
	a=0.10	243	300	364	435	511	592	679	772
	a=0.05	255	319	386	463	545	634	730	831
	a=0.025	259	331	406	486	574	670	771	880
	a=0.01	271	339	424	511	607	706	817	935
	a=0.005	271	346	431	526	624	731	847	970

Πίνακας Π.28: Ποσοστιαία σημεία για δίπλευρο έλεγχο μεγέθους 0.05 της ελεγχοσυνάρτησης R του ελέγχου των ροών. Οι κρίσιμες περιοχές αντιστοιχούν σε τιμές της R μικρότερες ή ίσες από την τιμή $r_{0.975}$ (πάνω γραμμή) είτε μεγαλύτερες ή ίσες από την τιμή $r_{0.025}$ (κάτω γραμμή). Πηγή: Sheskin (2011) Table A 8.

	$n_1 = 2$	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
$n_2 = 2$											2 -	2 -	2 -	2 -	2 -	2 -	2 -	2 -	2 -
3					2 -	2 -	2 -	2 -	2 -	2 -	2 -	2 -	2 -	3 -	3 -	3 -	3 -	3 -	3 -
4				2 9	2 9	2 -	3 -	3 -	3 -	3 -	3 -	3 -	3 -	3 -	4 -	4 -	4 -	4 -	4 -
5			2 9	2 10	3 10	3 11	3 11	3 -	3 -	4 -	4 -	4 -	4 -	4 -	4 -	4 -	4 -	5 -	5 -	5 -
6		2 -	2 9	3 10	3 11	3 12	3 12	4 13	4 13	4 13	4 13	5 -	5 -	5 -	5 -	5 -	5 -	6 -	6 -
7		2 -	2 -	3 11	3 12	3 13	4 13	4 14	5 14	5 14	5 14	5 15	5 15	6 15	6 -	6 -	6 -	6 -	6 -
8		2 -	3 -	3 11	3 12	4 13	4 14	5 14	5 15	5 15	6 16	6 16	6 16	6 16	6 17	7 17	7 17	7 17	7 17
9		2 -	3 -	3 -	4 13	4 14	5 14	5 15	5 16	6 16	6 16	6 17	7 17	7 18	7 18	7 18	8 18	8 18	8 18
10		2 -	3 -	3 -	4 13	5 14	5 15	5 16	6 16	6 17	7 17	7 18	7 18	7 18	8 19	8 19	8 19	8 20	9 20
11		2 -	3 -	4 -	4 13	5 14	5 15	6 16	6 17	7 17	7 18	7 19	8 19	8 19	8 20	9 20	9 20	9 21	9 21
12	2 -	2 -	3 -	4 -	4 13	5 14	6 16	6 16	7 17	7 18	7 19	8 19	8 20	8 20	9 21	9 21	9 21	10 22	10 22
13	2 -	2 -	3 -	4 -	5 -	5 15	6 16	6 17	7 18	7 19	8 19	8 20	9 20	9 21	9 21	10 22	10 22	10 23	10 23
14	2 -	2 -	3 -	4 -	5 -	5 15	6 16	7 17	7 18	8 19	8 20	9 20	9 21	9 22	10 22	10 23	10 23	11 23	11 24
15	2 -	3 -	3 -	4 -	5 -	6 15	6 16	7 18	7 18	8 19	8 20	9 21	9 22	10 22	10 23	11 23	11 24	11 24	12 25
16	2 -	3 -	4 -	4 -	5 -	6 -	6 17	7 18	8 19	8 20	9 21	9 21	10 22	10 23	11 23	11 24	11 25	12 25	12 25
17	2 -	3 -	4 -	4 -	5 -	6 -	7 17	7 18	8 19	9 20	9 21	10 22	10 23	11 23	11 24	11 25	12 25	12 26	13 26
18	2 -	3 -	4 -	5 -	5 -	6 -	7 17	8 18	8 19	9 20	9 21	10 22	10 23	11 24	11 25	12 25	12 26	13 26	13 27
19	2 -	3 -	4 -	5 -	6 -	6 -	7 17	8 18	8 20	9 21	10 22	10 23	11 23	11 24	12 25	12 26	13 26	13 27	14 27
20	2 -	3 -	4 -	5 -	6 -	6 -	7 17	8 18	9 20	9 21	10 22	10 23	11 24	12 25	12 25	13 26	13 27	13 27	14 28

Πίνακας Π.29: Κρίσιμες τιμές για τον συντελεστή συσχέτισης r_s του Spearman. Πηγή: Sheskin (2011) Table A 18.

	ε.σ. για μονόπλευρο έλεγχο			
	0.05	0.025	0.01	0.005
	ε.σ. για δίπλευρο έλεγχο			
n	0.10	0.05	0.02	0.01
5	0.900	1.000	1.000	
6	0.829	0.886	0.943	1.000
7	0.714	0.786	0.893	0.929
8	0.643	0.738	0.833	0.881
9	0.600	0.700	0.783	0.833
10	0.564	0.648	0.745	0.794
11	0.536	0.618	0.709	0.755
12	0.503	0.587	0.671	0.727
13	0.484	0.560	0.648	0.703
14	0.464	0.538	0.622	0.675
15	0.443	0.521	0.604	0.654
16	0.429	0.503	0.582	0.635
17	0.414	0.485	0.566	0.615
18	0.401	0.472	0.550	0.600
19	0.391	0.460	0.535	0.584
20	0.380	0.447	0.520	0.570
21	0.370	0.435	0.508	0.556
22	0.361	0.425	0.496	0.544
23	0.353	0.415	0.486	0.532
24	0.344	0.406	0.476	0.521
25	0.337	0.398	0.466	0.511
26	0.331	0.390	0.457	0.501
27	0.324	0.382	0.448	0.491
28	0.317	0.375	0.440	0.483
29	0.312	0.368	0.433	0.475
30	0.306	0.362	0.425	0.467

Πίνακας Π.30: Άνω ποσοστιαία σημεία του Kendall στατιστικού τεστ. Πηγή: Conover (1998) Table A 11.

n	$p = .100$	$p = 0.05$	$p = .025$	$p = .010$	$p = 0.005$
4	4	4	6	6	6
5	6	6	8	8	10
6	7	9	11	11	13
7	9	11	13	15	17
8	10	14	16	18	20
9	12	16	18	22	24
10	15	19	21	25	27
11	17	21	25	29	31
12	18	24	28	34	36
13	22	28	32	38	42
14	23	31	35	41	45
15	27	33	39	47	51
16	28	36	44	50	56
17	32	40	48	56	62
18	35	43	51	61	67
19	37	47	55	65	73
20	40	50	60	70	78
21	42	54	64	76	84
22	45	59	69	81	89
23	49	63	73	87	97
24	52	66	78	92	102
25	56	70	84	98	108
26	59	75	89	105	115
27	61	79	93	111	123
28	66	84	98	116	128
29	68	88	104	124	136
30	73	93	109	129	143

ΕΥΡΕΤΗΡΙΟ

Το παρόν σύγγραμμα χρηματοδοτήθηκε από το Πρόγραμμα Δημοσίων Επενδύσεων του Υπουργείου Παιδείας.