# ΜΗ ΠΑΡΑΜΕΤΡΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

**Ιωάννης Βρόντος**

**ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ**
**ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ**

**Δεκέμβριος 2006**

## ΣΚΟΠΟΣ ΤΟΥ ΜΑΘΗΜΑΤΟΣ

Σκοπός του μαθήματος είναι η παρουσίαση, μελέτη και ανάπτυξη τεχνικών και μεθοδολογιών για την αντιμετώπιση στατιστικών προβλημάτων. Το κύριο πλεονέκτημα των τεχνικών αυτών είναι ότι δεν προϋποθέτουν γνώση της κατανομής του πληθυσμού από τον οποίο έχουν προέλθει τα υπό μελέτη στοιχεία. Για τον λόγο αυτό, οι μη παραμετρικές τεχνικές αποβλέπουν σε ευρύτερα πεδία εφαρμογής από τις αντίστοιχες παραμετρικές τεχνικές.

Η μη παραμετρική ανάλυση δεδομένων ασχολείται, γενικά, με τον έλεγχο υποθέσεων που αφορά παραμέτρους του υπό μελέτη πληθυσμού, τον έλεγχο κατανομών, την (μη παραμετρική) παλινδρόμηση, την ανάλυση δεδομένων ταξινομημένων σε πίνακες συνάφειας. Θα δοθεί έμφαση τόσο στη θεωρία, όσο και στην αντιμετώπιση και εφαρμογή των τεχνικών αυτών σε δεδομένα.

Η αξιολόγηση θα γίνει με βάση την βαθμολογία στην τελική εξέταση.

### *Διδακτικά Βοηθήματα*

Ξεκαλάκη Ε. (2001). Μη Παραμετρική Στατιστική

### *Προτεινόμενη Βιβλιογραφία*

Conover, W.J. (1999). Practical Nonparametric Statistics, Third Edition, Wiley.

# ΚΕΦΑΛΑΙΟ 1:ΕΙΣΑΓΩΓΗ

## Παραμετρική Στατιστική

- Πολλές παραμετρικές μέθοδοι ελέγχου υποθέσεων στηρίζονται στην υπόθεση ότι η μεταβλητότητα των δεδομένων περιγράφεται από κάποια κατανομή συγκεκριμένης μορφής.
- Σε περιπτώσεις που δεν είναι επιτρεπτή οποιαδήποτε υπόθεση για την μορφή του πληθυσμού, ο ερευνητής χρησιμοποιεί το κεντρικό οριακό θεώρημα, και στηρίζεται συνήθως στην υπόθεση της κανονικότητας για να διεξάγει τους ελέγχους που τον ενδιαφέρουν.
- Πρόβλημα υπάρχει όταν τα δείγματα που έχουμε συλλέξει από τον υπό μελέτη πληθυσμό είναι μικρά, και τα δεδομένα δεν κατανέμονται από κανονικό πληθυσμό.

## Μη Παραμετρική Στατιστική

- Το κυριότερο πλεονέκτημα των μη παραμετρικών τεχνικών είναι ότι οι έλεγχοι δεν προϋποθέτουν γνώση της κατανομής του πληθυσμού από τον οποίο έχουν προέλθει τα υπό μελέτη στοιχεία.
- Εφαρμόζονται (συνήθως) στις τάξεις μεγέθους και όχι στα ιδία τα στοιχεία.
- Οι μη παραμετρικές τεχνικές χρησιμοποιούνται σε περίπτωση μικρών δειγμάτων (λόγω της φύσης του προβλήματος, ή όταν κάνουμε πιλοτική έρευνα).
- Οι μη παραμετρικοί έλεγχοι χρησιμοποιούνται επίσης για δεδομένα τα οποία είναι ταξινομημένα σε κατηγορίες (κατηγορικά δεδομένα, π.χ. κλίμακα διάταξης, ονομαστική κλίμακα).

## Γενικά Σχόλια

- Σε περίπτωση που είναι γνωστό ότι τα δεδομένα προέρχονται από κανονική κατανομή τότε οι παραμετρικοί έλεγχοι είναι πιο ισχυροί.
- Αν η κατανομή των δεδομένων δεν είναι κανονική τότε οι μη παραμετρικοί έλεγχοι έχουν πλεονέκτημα έναντι των παραμετρικών ελέγχων.

# ΚΕΦΑΛΑΙΟ 2: ΜΕΡΙΚΟΙ ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ ΒΑΣΙΣΜΕΝΟΙ ΣΤΗΝ ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

## 2.1 Διωνυμικός έλεγχος

- Έστω $X_1, X_2, ..., X_n$ είναι τα αποτελέσματα n ανεξάρτητων δοκιμών, όπου

$$X_i = \begin{cases} 1\ (\text{επιτυχία}),\ \text{με πιθανότητα p} \\ 0\ (\text{αποτυχία}),\ \text{με πιθανόητα 1-p} = \text{q} \end{cases},\ i = 1, 2, ..., n$$

π.χ. τα n αποτελέσματα μπορεί να είναι μια ακολουθία της μορφής: 0, 1, 1, 0, 0, 1, 0, 1, ...

Συμβολίζουμε: $O_1 = \#\ \text{παρατηρήσεων}\ 1\ (\#\ \text{επιτυχιών})$

$$O_2 = \#\ \text{των αποτυχιών}\ = \text{n} - \text{O}_1$$

- Ο διωνυμικός έλεγχος αφορά το ποσοστό του ενδεχομένου ‘επιτυχία’ στον πληθυσμό, και μπορεί να έχει μία από τις ακόλουθες μορφές:

**Αμφίπλευρος έλεγχος**: $H_o : p = p_o$

$H_1 : p \neq p_o$

**Μονόπλευρο έλεγχος**: $H_o : p \leq p_o$ ή $H_o : p \geq p_o$

$H_1 : p > p_o$ $H_1 : p < p_o$

- Για να πραγματοποιήσουμε τον έλεγχο χρειαζόμαστε μια στατιστική συνάρτηση ελέγχου, και την κατανομή της. Η Στατιστική συνάρτηση ελέγχου είναι ο αριθμός των επιτυχιών στις n ανεξάρτητες δοκιμές :

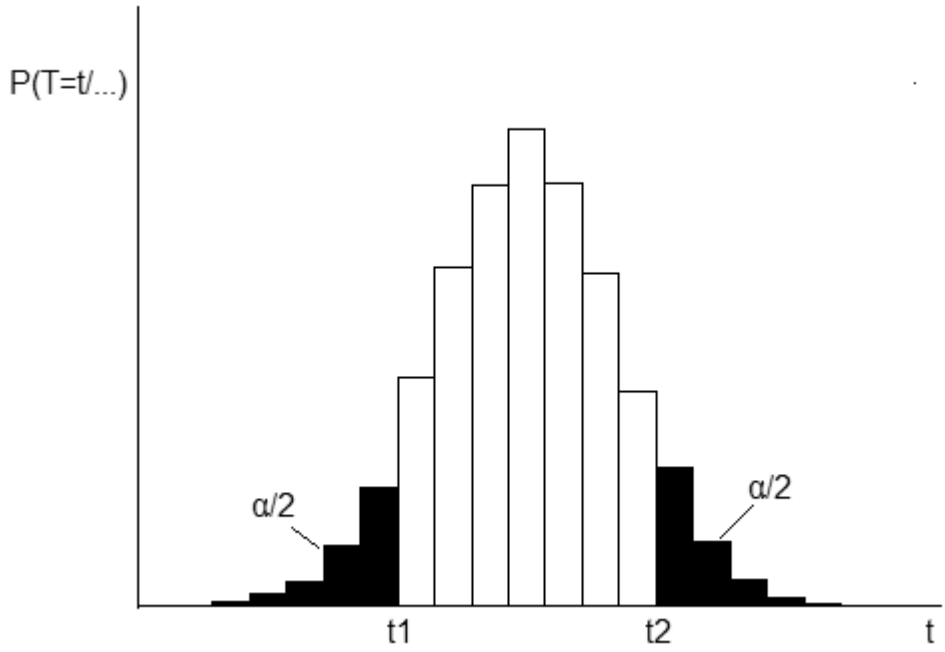
$$T = \sum_{i=1}^{n} X_i \Rightarrow T = O_1 = \ \#\ \text{παρατηρήσεων με επιτυχία 1}$$

$\text{T} \sim \text{Bin}\,(\text{n}, \text{p})$: Η κατανομή της στατιστικής συνάρτησης ακολουθεί την διωνυμική κατανομή με παραμέτρους n, p.

Η στατιστική συνάρτηση T είναι διακριτή τυχαία μεταβλητή, δηλαδή ο έλεγχος υποθέσεων σπάνια μπορεί να γίνει σε επίπεδο στατιστικής σημαντικότητας ακριβώς ίσο με το επιθυμητό α.

- **Πότε απορρίπτουμε την Ηο: (Χρήση Συνάρτησης ελέγχου και κριτικών τιμών)**

**1)** $H_o : p = p_o$ , $H_1 : p \neq p_o$




Απορρίπτουμε την $H_o$ αν $T \le t_1$ ή $T > t_2$ δηλαδή αν η συνάρτηση ελέγχου παίρνει ή πολύ μικρές ή πολύ μεγάλες τιμές. Η κρίσιμη περιοχή είναι όπως φαίνεται και από το γράφημα: $(-\infty,\ t_1] \cup (t_2, \infty)$.

Καθορισμός $t_1$, $t_2$ :

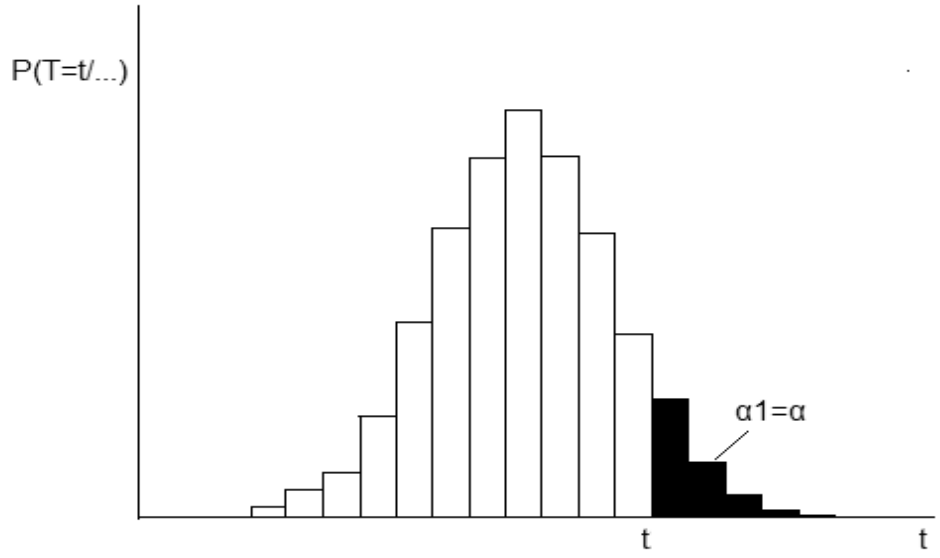
$$a = P\left[\left(\{T \le t_1\} \cup \{T > t_2\}\right) \ / \mathrm{n}, \mathrm{p} = \mathrm{p_o}\right] =$$
$$= P\left(T \le t_1 / p = p_o, n\right) + P\left(T > t_2 / p = p_o, n\right)$$

Το $t_1$ προσδιορίζεται από: $P\left(T \le t_1 / n,\ \mathrm{p} = \mathrm{p_o}\right) = a_1 \cong a/2$

Το $t_2$ προσδιορίζεται από: $P\left(T > t_2 / n,\ \mathrm{p} = \mathrm{p_o}\right) = a_2 \cong a/2$ ή

$$P\left(T \le t_2 / n,\ \mathrm{p} = \mathrm{p_o}\right) = 1 - a/2$$

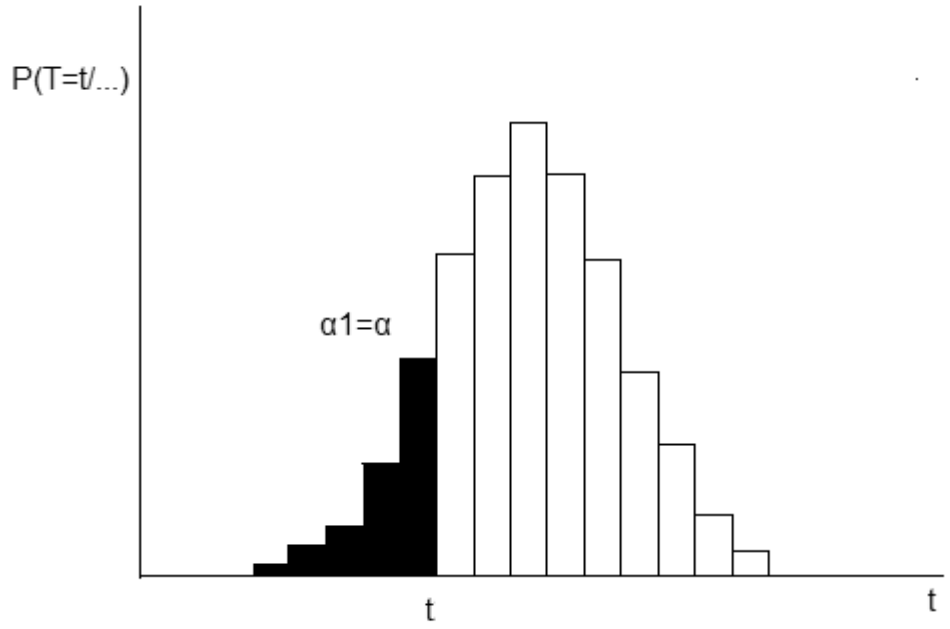
**2)** $H_o : p \le p_o$ , $H_1 : p > p_o$




Απορρίπτουμε την $H_o$ αν $T > t$ δηλαδή για μεγάλες τιμές της συνάρτησης ελέγχου. Η κρίσιμη περιοχή είναι όπως φαίνεται και από το γράφημα: $(t, \infty)$.

Καθορισμός $t$ : $P(T > t/n,\ \mathrm{p} = \mathrm{p}_o) = a_1 \cong a$ ή $P(T \le t/n,\ \mathrm{p} = \mathrm{p}_o) = 1 - a_1 = 1 - a$

**3)** $H_o : p \ge p_o$ , $H_1 : p < p_o$




Απορρίπτουμε την $H_o$ αν $T \le t$ δηλαδή για μικρές τιμές της συνάρτησης ελέγχου. Η κρίσιμη περιοχή είναι όπως φαίνεται και από το γράφημα: $(-\infty,\ t]$.

Καθορισμός $t$ : $P(T \le t/n,\ \mathrm{p} = \mathrm{p}_o) = a_1 \cong a$ .

- **Πότε απορρίπτουμε την Ho: (Χρήση του κρίσιμου επιπέδου, $\hat{a}$, p-value)**

Ορίζουμε ως $\hat{a}$ (p-value) την μέγιστη πιθανότητα να παρατηρηθεί μια τιμή της στατιστικής συνάρτησης T, το ίδιο ακραία ή περισσότερο ακραία από την τιμή που παρατηρήθηκε, δοθέντος ότι η μηδενική υπόθεση $H_0$ είναι αληθής.

**1)** $H_o : p = p_o$ , $H_1 : p \neq p_o$

$T \overset{H_0}{\sim} Bin(n, p = p_0)$. Στην συγκεκριμένη περίπτωση του διωνυμικού ελέγχου διακρίνουμε τις εξής περιπτώσεις:

Αν $\tau \leq np_0$ τότε ισχύει ότι $\frac{\hat{a}}{2} = P(T \leq \tau / n, p = p_0) \Rightarrow \hat{a} = 2P(T \leq \tau / n, p = p_0)$

Αν $\tau > np_0$ τότε ισχύει ότι $\frac{\hat{a}}{2} = P(T \geq \tau / n, p = p_0) \Rightarrow \hat{a} = 2P(T \geq \tau / n, p = p_0)$

Απορρίπτουμε την μηδενική υπόθεση $H_0$ αν $a > \hat{a}$.

**2)** $H_o : p \leq p_o$ , $H_1 : p > p_o$

Στην περίπτωση αυτή το παρατηρούμενο επίπεδο σημαντικότητας $\hat{a}$ υπολογίζεται ως εξής:

$$\hat{a} = P(T \geq \tau / n, p = p_0).$$

Απορρίπτουμε την μηδενική υπόθεση $H_0$ αν $a > \hat{a}$.

**3)** $H_o : p \geq p_o$ , $H_1 : p < p_o$

Στην περίπτωση αυτή το παρατηρούμενο επίπεδο σημαντικότητας $\hat{a}$ υπολογίζεται ως εξής:

$$\hat{a} = P(T \leq \tau / n, p_0).$$

Απορρίπτουμε την μηδενική υπόθεση $H_0$ αν $a > \hat{a}$.

**Παράδειγμα 2.1.1 (σελ. 48 βιβλίου)**

5 από τους 13 θανάτους εργατών οφείλονται σε καρκίνο. Μας ενδιαφέρει να ελεγχθεί το ζεύγος των υποθέσεων

$$H_0 : p = 0.25$$
$$H_1 : p \neq 0.25$$

σε επίπεδο στατιστικής σημαντικότητας α=0.05, όπου $p$ είναι το ποσοστό των θανάτων στον πληθυσμό που οφείλονται σε καρκίνο.

**Λύση**

**Χρήση Συνάρτησης ελέγχου και κριτικών τιμών**

n=13 και $\tau$ =5 (η τιμή της στατιστικής συνάρτησης ελέγχου). Στο συγκεκριμένο έλεγχο απορρίπτουμε την μηδενική υπόθεση αν ισχύει $T > t_2$ ή $T \le t_1$ όπου οι κρίσιμες περιοχές ορίζονται από τις σχέσεις:

$$P(T \le t_1 / n = 13, p = 0.25) = a_1 \cong 2.5\% \qquad (1)$$

$$P(T > t_2 / n = 13, p = 0.25) = a_2 \cong 2.5\% \text{ ή}$$
$$P(T \le t_2 / n = 13, p = 0.25) = 1 - a_2 \cong 97.5\% \quad (2)$$

$$(1) \Rightarrow P(T \le 0 / n = 13, p = 0.25) = 0.0238 \cong 0.025$$

$$(2) \Rightarrow P(T \le 6 / n = 13, p = 0{,}25) = 0.9757 \cong 0.975$$

Από τα παραπάνω αποτελέσματα έχουμε ότι $a_1 = 0.0238$ και $1 - a_2 = 0.9757$ και συνεπώς συμπεραίνουμε ότι $a_1 + a_2 = 0.0238 + (1 - 0.9757) = 0.0238 + 0.0243 = 0.0481$.

Άρα $t_1 = 0$ και $t_2 = 6$. Επειδή όμως T=$\tau$ =5 (και όχι $T \le 0$ ή $T \ge 6$) συμπεραίνουμε ότι δεν απορρίπτουμε την $H_0$ σε α=0.0481.

**Χρήση του κρίσιμου επιπέδου, $\hat{a}$ , p-value**

Παρατηρούμε ότι $\tau$ =5 και ότι $np_0 = 13 \cdot 0.25 = 3.25$. Επειδή $\tau = 5 > np_0$ τότε

$$\frac{\hat{a}}{2} = P(T \ge 5 / n = 13, p = 0.25) = 1 - P(T \le 4 / n = 13, p = 0.25) = 1 - 0.7940 = 0.206$$

Επομένως $\hat{a} = 2 \cdot 0.206 = 0.412$. Άρα αφού α=0.0481< $\hat{a}$ =0.412 συμπεραίνουμε ότι δεν απορρίπτουμε την μηδενική υπόθεση $H_0$.

**Προσέγγιση της διωνυμικής από την κανονική κατανομή**

Η προσέγγιση της διωνυμικής από την κανονική κατανομή μπορεί να γίνει όταν το μέγεθος του δείγματος n των ανεξάρτητων δοκιμών είναι μεγάλο (συνήθως n>20) και η κατανομή πιθανότητας είναι περίπου συμμετρική ($np \geq 5$, $n(1-p) \geq 5$). Στην περίπτωση αυτή η διωνυμική προσεγγίζεται από την κανονική κατανομή με μέσο $np$ και διακύμανση $npq$. Το κρίσιμο σημείο δίνεται από τον τύπο $t = np_0 + Z_r\sqrt{np_0(1-p_0)}$, όπου $Z_r$ είναι το r ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής.

**<u>Παράδειγμα 2.1.3 (σελ. 64 βιβλίου)</u>**

Σύμφωνα με τα δεδομένα του παραδείγματος έχουμε n=925 απογόνους φυτών, από τα οποία οι 243 είναι νάνοι και ότι οι 682 είναι κανονικοί. Μας ενδιαφέρει να εξετάσουμε το ζεύγος των υποθέσεων της μορφής: $\begin{matrix} H_0 : p = 3/4 \\ H_1 : p \neq 3/4 \end{matrix}$, σε επίπεδο στατιστικής σημαντικότητας α=0.05, όπου $p$ είναι το ποσοστό των φυτών – απογόνων που είναι κανονικά.

**<u>Λύση</u>**

**Χρήση Συνάρτησης ελέγχου και κριτικών τιμών**

T= $\tau$ =682, n=243+682=925.

$np_0 = 925 \cdot 0.75 = 693.75 \geq 5$

$n(1-p_0) = 925 \cdot 0.25 = 231.25 \geq 5$

$np_0(1-p_0) = 925 \cdot 0.25 \cdot 0.75 = 173.4375$.

Επομένως:

$$\frac{a}{2} = P(T \leq t_1 / n = 925, p = p_0 = 0.75)$$

$$\cong P(\frac{T-\mu}{\sigma} \leq \frac{t_1-\mu}{\sigma} / \mu = np_0 = 925 \cdot 0.75 = 693.75,\ \sigma^2 = np_0q_0 = 173.4375)$$

$$= P(Z \leq \frac{t_1-\mu}{\sigma} / \mu = 693.75,\ \sigma^2 = 173.4375) \quad (1)$$

$$\frac{a}{2} = P(T > t_2 / n = 925,\ p = p_0 = 0.75) \Rightarrow$$

$$\frac{\alpha}{2} = 1 - P(T \le t_2 / n = 925,\ p = p_0 = 0.75) \Rightarrow$$

$$1 - \frac{a}{2} = P(T \le t_2 / n = 925,\ p = p_0 = 0.75)$$

$$\cong P(\frac{T-\mu}{\sigma} \le \frac{t_2-\mu}{\sigma} / \mu = np_0 = 925 \cdot 0.75 = 693.75,\ \sigma^2 = np_0q_0 = 173.4375)$$

$$= P(Z \le \frac{t_2-\mu}{\sigma} / \mu = 693.75,\ \sigma^2 = 173.4375) \quad (2)$$

Από τις παραπάνω σχέσεις μπορούμε να δούμε ότι ισχύουν τα ακόλουθα:

$$\frac{t_1-\mu}{\sigma} = Z_{\alpha/2} \Rightarrow t_1 = \mu + Z_{\alpha/2}\sigma = np_0 + Z_{a/2}\sqrt{np_0(1-p_0)} \Rightarrow t_1 = 693.75 - 1.96\sqrt{173.4}$$

$$\Rightarrow t_1 = 693.75 - 1.96 \cdot 13.16 = 693.75 - 25.79 = 667.96$$

$$\frac{t_2-\mu}{\sigma} = Z_{1-\alpha/2} \Rightarrow t_2 = \mu + Z_{1-\alpha/2}\sigma = np_0 + Z_{1-a/2}\sqrt{np_0(1-p_0)} \Rightarrow t_2 = 693.75 + 1.96\sqrt{173.4}$$

$$\Rightarrow t_2 = 693.75 + 1.96 \cdot 13.16 = 693.75 + 25.79 = 719.54$$

Άρα από τα παραπάνω αποτελέσματα ($t_1 = 667.96$ <T=682< $t_2 = 719.54$) βγάζουμε το συμπέρασμα ότι δεν απορρίπτουμε την μηδενική υπόθεση $H_0$.

**Χρήση του κρίσιμου επιπέδου, $\hat{a}$, p-value**

Αφού $\tau = 682 < np_0 = 693.75$, το p-value υπολογίζεται ως εξής:

$$\frac{\hat{a}}{2} = P(T \le \tau / n,\ p = p_0) = P(Z \le \frac{682-\mu}{\sigma} / \mu = np_0, \sigma^2 = np_0(1-p_0)) =$$

$$P(Z \le \frac{682-693.75}{\sqrt{173.4375}}) = \Phi(\frac{-11}{13.16}) = \Phi(-0.83) = 0.2033$$

Συνεπώς έχουμε ότι $\hat{a} = p-value = 2 \cdot 0.2033 = 0.4066$. Δηλαδή σε επίπεδο στατιστικής σημαντικότητας α=0.05 συμπεραίνουμε ότι δεν απορρίπτουμε την μηδενική υπόθεση $H_0$ επειδή $\hat{a} > a$.

**Άσκηση 1 (Θέμα 1ο, Εξετάσεις Φεβρουαρίου 2005)**

Μία εταιρεία απορρυπαντικών ενδιαφέρεται να εκτιμήσει το ποσοστό των νοικοκυρών που χρησιμοποιούν κάποιο από τα απορρυπαντικά που παράγει. Σε μία έρευνα ρωτήθηκαν 14 νοικοκυρές και βρέθηκε ότι 4 χρησιμοποιούν προϊόντα ή απορρυπαντικά που παράγει η συγκεκριμένη εταιρεία. Να εξεταστεί σε επίπεδο σημαντικότητας 0.05, το ακόλουθο ζεύγος των υποθέσεων:

$$H_0 : p = 0.35$$
$$H_1 : p \neq 0.35$$

όπου p είναι το ποσοστό των νοικοκυρών που χρησιμοποιούν κάποιο από τα απορρυπαντικά που παράγει η εταιρεία (επί του συνόλου των νοικοκυρών). Ο έλεγχος να πραγματοποιηθεί με 2 τρόπους α)χρησιμοποιώντας τις κριτικές τιμές β)χρησιμοποιώντας το κρίσιμο επίπεδο του ελέγχου. Ποιο είναι το συμπέρασμα στο οποίο καταλήγετε;

**Λύση**

**α) Χρήση Συνάρτησης ελέγχου και κριτικών τιμών**

$t = 4$ , $n = 14$

Απορρίπτουμε την $H_o$ αν $T > t_2$ ή $T \le t_1$. Οι κρίσιμες τιμές ορίζονται ως εξής

$$P(T \le t_1 / n = 14, p = 0.35) = a_1 = 0.025 \qquad (1)$$

$$P(T > t_2 / n = 14, p = 0.35) = a_2 = 0.025 \qquad \text{ή} \qquad P(T \le t_2 / n, p) = 1 - a_2 = 0.975 \qquad (2)$$

$$(1) \Rightarrow P(T \le 1 / n = 14, p = 0.35) = 0.0205 \qquad t_1 = 1$$

$$(2) \Rightarrow P(T \le 8 / n = 14, p = 0.35) = 0.9757 \qquad t_2 = 8$$

Επειδή T= $\tau$ =4 (μεγαλύτερο από $t_1 = 1$, και μικρότερο από $t_2 = 8$) δεν απορρίπτω την $H_o$ σε α=0.0448.

$$\left.\begin{matrix} a_1 = 0.0205 \\ 1 - a_2 = 0.9757 \end{matrix}\right\} a_1 + a_2 = 0.0205 + (1 - 0.9757) = 0.0205 + 0.0243 = 0.0448$$

**β) Χρήση του κρίσιμου επιπέδου, $\hat{a}$ , p-value**

$n \cdot p_o = 14 \cdot 0.35 = 4.9$ . Επειδή $\tau = 4 < n \cdot p_o$ τότε

$$\frac{\hat{a}}{2} = P(T \le \tau / n = 14, p = 0.35) \Rightarrow \hat{a} = 2 \cdot P\left(T \le 4 / n = 14, p = 0.35\right) = 2 \cdot 0.4227 = 0.8454$$

Άρα σε α=5% επειδή $\hat{a} < 0.8454$ δεν απορρίπτουμε την Ho.

**<u>Άσκηση 2 (Θέμα 1ο, Εξετάσεις Σεπτεμβρίου 2005)</u>**

Ο διευθυντής ενός αεροδρομίου ισχυρίζεται ότι περισσότερες από το 45% των πτήσεων εξωτερικού δεν έχουν χρόνο καθυστέρησης στις προγραμματισμένες αναχωρήσεις τους. Σε ένα τυχαίο δείγμα 14 πτήσεων εξωτερικού, διαπιστώνεται ότι υπάρχει καθυστέρηση στις αναχωρήσεις 4 πτήσεων. Να εξεταστεί σε επίπεδο σημαντικότητας 0.05, το ακόλουθο ζεύγος των υποθέσεων:

$$H_0 : p \le 0.45$$
$$H_1 : p > 0.45$$

όπου p είναι το ποσοστό των πτήσεων εξωτερικού που δεν έχουν χρόνο καθυστέρησης στις προγραμματισμένες αναχωρήσεις τους. Ποιο είναι το συμπέρασμα στο οποίο καταλήγετε;

**<u>Λύση</u>**

**α) Χρήση Συνάρτησης ελέγχου και κριτικών τιμών**

n=14 $\tau = 10$

$P\left(T > t / n = 14, p = 0.45\right) = 0.05$ ή $P\left(T \le t / n = 14, p = 0.45\right) = 0.95$ ή

$P\left(T \le 9 / n = 14, p = 0.45\right) = 0.9574$

Άρα t=9 αφού T=10>t=9 απορρίπτουμε την $H_o$. Επομένως ισχύει ο ισχυρισμός του διευθυντή.

**β) Χρήση του κρίσιμου επιπέδου, $\hat{a}$ , p-value**

$$\hat{a} = p - value = P\left(T \ge \tau / n = 14, p = 0.45\right) = P\left(T \ge 10 / n = 14, p = 0.45\right) =$$
$$= 1 - P\left(T < 10 / n = 14, p = 0.45\right) = 1 - P(T \le 9 / n = 14, p = 0.45) = 1 - 0.9574 = 0.0426$$

Επειδή ο έλεγχος γίνεται σε επίπεδο σημαντικότητας α=0.05 και επειδή έχουμε p-value=0.0426 συμπεραίνουμε ότι η $H_o$ απορρίπτεται. Επομένως ισχύει ο ισχυρισμός του διευθυντή.

## 2.2 Προσημικός έλεγχος ή έλεγχος προσήμων

- Ουσιαστικά ο προσημικός έλεγχος είναι ο διωνυμικός έλεγχος για την περίπτωση όπου $p_0 = \frac{1}{2}$. Ο προσημικός έλεγχος χρησιμοποιείται για:

  Έλεγχο ότι οι τιμές μιας από τις μεταβλητές του ζεύγους (X, Y) τείνουν να είναι μεγαλύτερες από τις τιμές της άλλης.
  Έλεγχο ύπαρξης τάσης σε μια ακολουθία μετρήσεων σε κλίμακα διάταξης
  Έλεγχο ύπαρξης συσχέτισης

- Τα δεδομένα για τον προσημικό έλεγχο είναι παρατηρήσεις σε ένα διδιάστατο τυχαίο δείγμα $(X_i, Y_i)$, $i = 1, 2, ..., n'$, όπου η $X_i$ δεν είναι ανεξάρτητη από την $Y_i$ (εξαρτημένες τιμές).

  Για κάθε ζεύγος παρατηρήσεων κάνουμε τις συγκρίσεις: Αν $X_i < Y_i$, η παρατήρηση ταξινομείται ως «+», αν $X_i > Y_i$ ταξινομείται ως «-» και τέλος αν $X_i = Y_i$ ταξινομείται ως 0.

- Αν οι τιμές της X τείνουν να είναι μεγαλύτερες από τις τιμές Y τότε θα ισχύει ότι $P(-) > P(+)$. Οι δυνατές μορφές που μπορεί να πάρει ο προσημικός έλεγχος δίνονται παρακάτω:

$$\begin{matrix} H_0 : P(+) = P(-) \\ H_1 : P(+) \neq P(-) \end{matrix} \text{ ή } \begin{matrix} H_0 : E(X) = E(Y) \\ H_1 : E(X) \neq E(Y) \end{matrix} \quad (1)$$

$$\begin{matrix} H_0 : P(+) \leq P(-) \\ H_1 : P(+) > P(-) \end{matrix} \text{ ή } \begin{matrix} H_0 : E(X) \geq E(Y) \\ H_1 : E(X) < E(Y) \end{matrix} \quad (2)$$

$$\begin{matrix} H_0 : P(+) \geq P(-) \\ H_1 : P(+) < P(-) \end{matrix} \text{ ή } \begin{matrix} H_0 : E(X) \leq E(Y) \\ H_1 : E(X) > E(Y) \end{matrix} \quad (3)$$

Από την μελέτη των παραπάνω μορφών που μπορεί να πάρει ο προσημικός έλεγχος γίνεται εμφανές ότι ο συγκεκριμένος έλεγχος μπορεί να χρησιμοποιηθεί και για έλεγχο διαμέσων.

- Στον προσημικό έλεγχο η στατιστική συνάρτηση ελέγχου T είναι ο αριθμός των «+» ζευγών και το συνολικό μέγεθος του δείγματος είναι ίσο με n, όπου n είναι το άθροισμα του αριθμού των «+» και του αριθμού των «-» ζευγών. Άρα βάσει αυτών μπορούμε να πούμε ότι η στατιστική συνάρτηση ελέγχου T ακολουθεί την διωνυμική κατανομή $T \sim Bin(n, p = p(+))$. Έτσι λοιπόν οδηγούμαστε με τον τρόπο αυτό σε μια άλλη δυνατή μορφή που μπορεί να πάρει ο προσημικός έλεγχος η οποία είναι η ακόλουθη:

$$\begin{matrix} H_0 : P(+) = \frac{1}{2} \\ H_1 : P(+) \neq \frac{1}{2} \end{matrix} \quad (1) \qquad \begin{matrix} H_0 : P(+) \leq \frac{1}{2} \\ H_1 : P(+) > \frac{1}{2} \end{matrix} \quad (2) \qquad \begin{matrix} H_0 : P(+) \geq \frac{1}{2} \\ H_1 : P(+) < \frac{1}{2} \end{matrix} \quad (3)$$

Στις περιπτώσεις αυτές μπορούμε να πούμε ότι $T \sim Bin(n, p = p(+) = \frac{1}{2})$.

- Για την περίπτωση (1) απορρίπτουμε την $H_0$ αν $T \leq t_1$ ή αν $T > t_2$ όπου τα $t_1, t_2$ καθορίζονται από τις παρακάτω σχέσεις:

$$P(T \leq t_1 / n, p = 0{,}5) \cong a/2 \quad \text{και} \quad P(T > t_2 / n, p = 0{,}5) \cong a/2$$

Λόγω της συμμετρίας ισχύει ότι

$$\begin{aligned} a/2 &\cong P(T \leq t_1 / n, p = 0.5) \\ &= P(T > t_2 / n, p = 0.5) \\ &= P(T \geq t_2 + 1 / n, p = 0.5) \\ &= P(T \geq n - t_1 / n, p = 0.5) \end{aligned}$$

Οπότε στην συγκεκριμένη περίπτωση (1) απορρίπτουμε την μηδενική υπόθεση αν $T \leq t$ ή αν $T \geq n - t$ όπου το t καθορίζεται από την σχέση $P(T \leq t / n, p = 0{,}5) = a/2$.

Για την περίπτωση (2) απορρίπτουμε την $H_0$ αν ισχύει T>t όπου το t καθορίζεται από την σχέση $P(T > t / n, p = 0.5) \cong \alpha$.

Τέλος για την περίπτωση (3) απορρίπτουμε την $H_0$ αν $T \leq t$ όπου το t καθορίζεται από την σχέση $P(T \leq t / n, p = 0.5) \cong \alpha$

**Παράδειγμα 2.2.1 (σελ. 76 βιβλίου)**

| $X_i$ (βάρη πριν τη δίαιτα) | $Y_i$ (βάρη μετά την δίαιτα) | Πρόσημο ζεύγους $(X_i, Y_i)$ |
|---|---|---|
| 174 | 165 | - |
| 191 | 186 | - |
| 188 | 183 | - |
| 182 | 178 | - |
| 201 | 203 | + |
| 188 | 181 | - |

Θέλουμε να εξετάσουμε αν παρέχουν τα δεδομένα ενδείξεις ότι η συγκεκριμένη δίαιτα είναι αποτελεσματική. Οπότε μπορούμε να πούμε ότι η δίαιτα θα είναι αποτελεσματική αν P(+)<P(-) όπου «+» είναι η περίπτωση όπου X<Y. Άρα ο έλεγχος που θα εφαρμόσουμε θα είναι της μορφής $\begin{matrix} H_0 : P(+) \geq P(-) \ \ (\text{μη αποτελεσματική δίαιτα}) \\ H_1 : P(+) < P(-) \ \ (\text{αποτελεσματική δίαιτα}) \end{matrix}$ ή αλλιώς $\begin{matrix} H_0 : P(+) \geq 0.5 \\ H_1 : P(+) < 0.5 \end{matrix}$.

Η συνάρτηση T θα είναι ίση με τον αριθμό των «+», δηλαδή εδώ έχουμε ότι T=1. Επίσης έχουμε ότι n=6 και με βάση τα στοιχεία αυτά θα υπολογίσουμε το p-value.

Έχουμε ότι $p-value = \hat{a} = P(T \leq \tau_1) = P(T \leq 1/n = 6, p = 0.5) = 0.1094$. Άρα λοιπόν σε επίπεδο στατιστικής σημαντικότητας α=0.05 συμπεραίνουμε ότι δεν απορρίπτουμε την $H_0$ επειδή α<p-value=$\hat{a}$.

Ας αντιμετωπίσουμε τώρα τον παραπάνω έλεγχο με έναν εναλλακτικό τρόπο: θα θεωρούμε ως επιτυχία τις περιπτώσεις όπου έχουμε «-». Άρα λοιπόν η συνάρτηση ελέγχου T θα είναι ίση με τον αριθμό των «-». Στην περίπτωση μας έχουμε ότι T=5 και το p-value θα δίνεται από την σχέση $P(T \geq 5) = 1 - P(T \leq 4) = 1 - 0.8906 = 0.1094$. Επομένως, δεν απορρίπτουμε την μηδενική υπόθεση $H_0$.

## 2.3 Παραλλαγές προσημικού ελέγχου

### 2.3.1 Έλεγχος Mc Nemar για την σημαντικότητα αλλαγής μιας κατάστασης

- Τα δεδομένα αφορούν n' ανεξάρτητες παρατηρήσεις δύο τυχαίων μεταβλητών $(X_i,\ Y_i),\ \mathrm{i}=1,...,\mathrm{n}'$
- Η κλίμακα μέτρησης των $X_i, Y_i$ είναι ονομαστική με δύο κατηγορίες "0" και "1". Δυνατές τιμές ζευγών είναι (0, 0) , (0, 1) , (1, 0) , (1, 1).
- Σκοπός είναι να ανιχνεύσουμε αν υπάρχει διαφορά μεταξύ της πραγματοποίησης του ενδεχομένου (0, 1) και του (1, 0).
- Ταξινόμηση δεδομένων

| | | Y | |
|---|---|---|---|
| | | 0 | 1 |
| X | 0 | (0, 0) | (0, 1) |
| | 1 | (1, 0) | (1, 1) |

Μας ενδιαφέρει να ελέγξουμε το ζεύγος των υποθέσεων

$$H_O : P(X=0, Y=1) = P(X=1, Y=0)$$

$$H_1 : P(X=0, Y=1) \neq P(X=1, Y=0)$$

όπου n = ο αριθμός ζευγών (0, 1) + αριθμός ζευγών (1,0)

Έστω ότι το ζεύγος (0, 1) θεωρείται "επιτυχία" (+), και το ζεύγος (1,0) θεωρείται "αποτυχία" (-). Τότε η συνάρτηση ελέγχου T είναι ο αριθμός των επιτυχιών και ακολουθεί την Διωνυμική κατανομή με παραμέτρους n, p. Δηλαδή, $T \sim Bin(n, p=1/2)$. Το ζεύγος των υποθέσεων που θέλουμε να ελέγξουμε είναι το ακόλουθο:

$$H_o : P(+) = P(-), \qquad \text{ή} \qquad H_o : P(+) = 1/2,$$

$$H_1 : P(+) \neq P(-), \qquad H_1 : P(+) \neq 1/2.$$

**<u>Παράδειγμα 2.3.1 (σελ. 88 βιβλίου)</u>**

| | | **Y** | | |
|---|---|---|---|---|
| | | **0** | **1** | **Σύνολο** |
| **X** | **0** | 7 | 36(+) | 43 |
| | **1** | 30(-) | 62 | 92 |
| | | 37 | 98 | 135 |

X: η γνώμη του ατόμου αρχικά (την πρώτη φορά)

Y: η γνώμη του ατόμου μετά από κάποιο διάστημα (την δεύτερη φορά)

0: αρνητική γνώμη

1: θετική γνώμη

**<u>Λύση</u>**

Από 135 άτομα 43 ήταν (αρχικά) εναντίον της ακολουθούμενης εξωτερικής πολιτικής. Όταν ξαναρωτήθηκαν τα 135 άτομα, 37 ήταν εναντίον της ακολουθούμενης εξωτερικής πολιτικής, από τα οποία τα 30 ήταν αρχικά υπέρ.

Θέλουμε να ελέγξουμε εάν η μεταβολή στον αριθμό των ατόμων που ήταν εναντίον της εξωτερικής πολιτικής είναι σημαντική. $H_o : P(0,1) = 1/2, \qquad H_1 : P(0,1) \neq 1/2$

n=36+30=66.

Η συνάρτηση ελέγχου είναι ο αριθμός των «+», δηλαδή T=36. $T \sim Bin(n = 66, p = 0.5)$.

Επειδή n>20 μπορούμε να πούμε ότι $T \sim N(\mu = np = \frac{n}{2}, \sigma^2 = npq = \frac{n}{4})$. Η κρίσιμη περιοχή μπορεί να βρεθεί από τις σχέσεις $T \le t$ και $T \ge n - t$ όπου το t ορίζεται από την σχέση $P(T \le t / n, p = 0,5) \cong a/2$.

Έχουμε όμως ότι $\frac{a}{2} = P(T \le t / n = 66, p = 0.5) = P(\frac{T - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \le \frac{t - \frac{n}{2}}{\sqrt{\frac{n}{4}}}) = P(Z \le \frac{t - \frac{n}{2}}{\sqrt{\frac{n}{4}}})$.

Επίσης έχουμε ότι $\frac{t - \frac{n}{2}}{\sqrt{\frac{n}{4}}} = Z_{a/2} = -Z_{1-a/2} \Rightarrow t = -Z_{1-a/2} \cdot \sqrt{n/4} + \frac{n}{2}$.

Χρησιμοποιώντας επίπεδο σημαντικότητας α=0.10 $\Rightarrow Z_{1-a/2} = Z_{0.95} = 1.645$

Η κρίσιμη περιοχή δίνεται από την σχέση $\{T \le t\} \bigcup \{T \ge n - t\}$ οπότε

$$t = \frac{66}{2} - 1.645 \cdot \sqrt{\frac{66}{4}} = 33 - 1.645 \cdot \sqrt{16.5} = 27$$

Άρα λοιπόν θα απορρίπτουμε την $H_0$ αν $T \le 27$ ή αν $T \ge 66 - 27 = 39$

Επειδή όμως $\tau$ =36 συμπεραίνουμε ότι δεν απορρίπτουμε την $H_0$.

### 2.3.1 Έλεγχος Cox Stuart για την Ύπαρξη Τάσης σε μια Ακολουθία Παρατηρήσεων.

Στον Έλεγχο Cox Stuart τα δεδομένα αποτελούνται από n παρατηρήσεις πάνω σε μια ακολουθία τυχαίων μεταβλητών $X_1, X_2, \ldots, X_n$, των οποίων οι δείκτες υποδεικνύουν την σειρά με την οποία οι τυχαίες μεταβλητές παρατηρήθηκαν. Πραγματοποιώντας τον έλεγχο Cox Stuart θέλουμε να δούμε αν η εν λόγω σειρά έχει αυξητική τάση.

Έστω $c = \begin{cases} n'/2 & \text{αν n' άρτιος} \\ (n'+1)/2 & \text{αν n' περιττός} \end{cases}$

Φτιάχνουμε ζεύγη $(X_1, X_{1+c}), (X_2, X_{2+c}), \ldots, (X_{n'-c}, X_{n'})$ και το κάθε ζεύγος το "βαθμολογούμε" με:

- **"+"**, αν $X_i < X_{i+c}$
- **"-"**, αν $X_i > X_{i+c}$
- 0, αν $X_i < X_{i+c}$

Επίσης έστω n ο αριθμός των "+" και "-" ζευγών. Η ελεγχοσυνάρτηση εδώ είναι T: ο αριθμός των "+", T~Διωνυμική(n,p).

**(Α) Αμφίπλευρος Έλεγχος**

$$\left.\begin{array}{l} H_o: \text{δεν υπάρχει τάση (συχνά στον χρόνο δεν υπάρχει σχέση)} \\ \text{H}_1: \text{υπάρχει τάση (υπάρχει σχέση στον χρόνο)} \end{array}\right\} \Leftrightarrow \begin{array}{l} \text{H}_0: P(+) = P(-) \\ H_1: P(+) \neq P(-) \end{array}$$

Μπορεί να χρησιμοποιηθεί για ανίχνευση μη τυχαίου σήματος στη σειρά

**(Β) Μονόπλευρος Έλεγχος**

$$\left.\begin{array}{l} H_o: \text{δεν υπάρχει αυξητική τάση )} \\ \text{H}_1: \text{υπάρχει αυξητική τάση} \end{array}\right\} \Leftrightarrow \begin{array}{l} \text{H}_0: P(+) \leq P(-) \\ H_1: P(+) > P(-) \end{array}$$

**<u>Παράδειγμα 2.3.2 (σελ. 100 βιβλίου)</u>**

Ρυθμός θανάτων από τροχαία από 100.000 κατοίκους για 15 χρόνια.

(η παρατήρηση 19.2 αγνοείται για δεν υπάρχει παρατήρηση που απέχει από αυτή 8 χρόνια)

$x_i$ : 17.3 17.9 18.4 18.1 18.3 19.6 18.6 19.2

$x_{i+8}$ : 17.7 20.0 19.0 18.8 19.3 20.2 19.9

Πρόσημο:+ + + + + + +

Τείνει ο ρυθμός θανάτων να αυξάνει;

n'=15 και c=16/2=8 α=0,05 T= ο # των "+"=7 και n=7

$$\left.\begin{array}{l} H_o: \text{δεν υπάρχει αυξητική τάση )} \\ \text{H}_1: \text{υπάρχει αυξητική τάση} \end{array}\right\} \Leftrightarrow \begin{array}{l} \text{H}_0: P(+) \leq 1/2 \\ H_1: P(+) > 1/2 \end{array}$$

Κρίσιμη περιοχή $T \geq n - t$ όπου $\Pr\left(T \leq t/T \sim \text{Bin}\left(7, \frac{1}{2}\right)\right) = 0{,}05$

για $t = 1$ πραγματικό σημείο $\alpha = 0{,}0625$ $\quad T \geq 7 - 1 = 6$

στατιστικής σημαντικότητας $= 1 - P(T \leq 6/n = 7, \rho = 1/2) = 1 - 0.9922$

Κρίσιμο επίπεδο : $\hat{\alpha} = \Pr(T \geq 7/\text{Bin}(7,1/2)) = 0{,}0078$

Το 0,0078 δείχνει πόσο πιθανό είναι αυτό που είδα κάτω από την $H_o$ επομένως απορρίπτουμε $H_o$ , αλλά και για α=0,01 πάλι θα απέρριπτα την $H_o$.

### 2.3.1 Έλεγχος Ύπαρξης Συσχέτισης Cox Stuart

Σε αυτό τον έλεγχο η μηδενική και η εναλλακτική υπόθεση είναι:

$H_o$ : δεν υπάρχει συσχέτιση

$H_1$ : υπάρχει συσχέτιση

Διατάσσουμε τις παρατηρήσεις έτσι ώστε η επιλεγμένη μεταβλητή να είναι σε αύξουσα σειρά, κατόπιν εξετάζουμε για ύπαρξη τάσης στην άλλη μεταβλητή

Παράδειγμα

n'=10 ασθενείς (α=0,05). Υπάρχει θετική συσχέτιση των αντιδράσεων στο Α,Β;

n'=10 c=5

| **Ασθενής** | **Φάρμ.Α** | **Φάρμ.Β** | **Ασθενείς** | **X** | **Y** | $Y_i$ | $Y_{i+5}$ | **Πρόσημο** |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.7 | 1.9 | 2 | -1.6 | 0.8 | 0.8 | 1.9 | + |
| 2 | -1.6 | 0.8 | 4 | -1.2 | 0.1 | 0.1 | 1.6 | + |
| 3 | -0.2 | 1.1 | 3 | -0.2 | 1.1 | 0.1 | 3.4 | + |
| 4 | -1.2 | 0.1 | 5 | -0.1 | -0.1 | -0.1 | 4.4 | + |
| 5 | -0.1 | -0.1 | 9 | 0 | 4.6 | 4.6 | 5.5 | + |
| 6 | 3.4 | 4.4 | 1 | 0.7 | 1.9 | | | |
| 7 | 3.7 | 5.5 | 8 | 0.8 | 1.6 | | | |
| 8 | 0.8 | 1.6 | 10 | 2 | 3.4 | | | |
| 9 | 0 | 4.6 | 6 | 3.4 | 4.4 | | | |
| 10 | 2 | 3.4 | 7 | 3.7 | 5.5 | | | |

$$\left.\begin{array}{l} H_o : \text{δεν υπάρχει συσχέτιση} \\ H_1 : \text{υπάρχει συσχέτιση} \end{array}\right\} \Leftrightarrow \begin{array}{l} H_O : \text{δεν υπάρχει αυξητική τάση} \\ H_1 : \text{υπάρχει αυξητική τάση} \end{array}$$

Τ= # θετικών προσήμων

$T \geq n - t$ όπου $\Pr(T \leq t / T \sim Bin(5, 0.5)) = a$ t=0 α=0,0312

Απορρίπτω $H_o$ αν $T \geq n - t \Leftrightarrow T \geq n - 0 \Leftrightarrow T \geq 5$. Εδώ απορρίπτεται η $H_o$ σε α=0,0312, άρα υπάρχει θετική συσχέτιση.

$\hat{a} = \Pr(T \geq 5 / T \sim Bin(5, 0.5)) = 1 - P(T \leq 4) = 1 - 0.9688 = 0.0312$

## ΚΕΦΑΛΑΙΟ 3: ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ ΒΑΣΙΣΜΕΝΕΣ ΣΤΙΣ ΤΑΞΕΙΣ ΜΕΓΕΘΟΥΣ ΤΩΝ ΠΑΡΑΤΗΡΗΣΕΩΝ ΕΝΟΣ Η ΔΥΟ ΔΕΙΓΜΑΤΩΝ

### 3.1 Έλεγχος Wilcoxon για ένα δείγμα παρατηρήσεων ή ζεύγος παρατηρήσεων

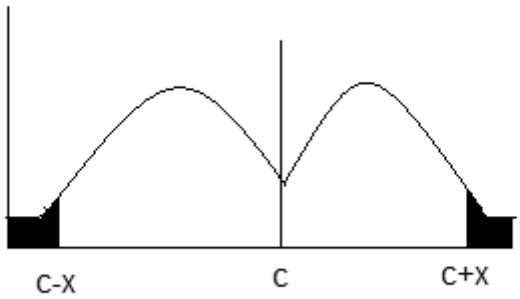
- Έλεγχος παραμέτρων κεντρικής τάσης . (μέσο, διάμεσο)
- Ένα δείγμα παρατηρήσεων ή ζεύγη παρατηρήσεων.
- Παραμετρικός ανάλογος έλεγχος t.

**Προσημικός:** Κοιτά μόνο τις διαφορές (θετικές, αρνητικές). + ή -

**Wilcoxon:**

- Κοιτά και το μέγεθος της διαφοράς.
- Υπόθεση συμμετρίας.

Συμμετρία: Συνεχείς μεταβλητές: $P(X \geq c + x) = P(X \leq c - x)$ συμμετρία γύρω από την ευθεία x=c.




- Συμμετρία: μέσος = διάμεσος
- Υπόθεση συμμετρίας λιγότερο αυστηρή από της κανονικότητας
- Ο έλεγχος χρησιμοποιείται και για διατεταγμένες παρατηρήσεις.

### 3.1.1 Έλεγχος Προσημασμένων Τάξεων Μεγέθους του Wilcoxon για την διάμεσο ενός πληθυσμού.

Έστω τυχαίο δείγμα $X_1,...,X_{n'}$ από πληθυσμό με συμμετρική κατανομή.

Έστω ότι συμβολίζουμε την διάμεσο με $X_{0.5}$. Υπάρχουν τρεις έλεγχοι:

| Α. Μονόπλευρος Έλεγχος | Β. Μονόπλευρος Έλεγχος | Γ. Αμφίπλευρος Έλεγχος |
|---|---|---|
| $H_O : X_{0.5} \geq m$ | $H_O : X_{0.5} \leq m$ | $H_O : X_{0.5} = m$ |
| $H_1 : X_{0.5} < m$ | $H_1 : X_{0.5} > m$ | $H_O : X_{0.5} \neq m$ |

Ισχύει και για έλεγχο της μέσης τιμής (αντί για $X_{0.5} \to E(x)$ )

Για την διάμεσο ισχύει:

- Για συνεχή πληθυσμό $P(X < X_{0.5}) = P(X > X_{0.5}) = 0.5$
- Για διακριτή τυχαία μεταβλητή

$P(X < X_{0.5}) \leq 0.5 \qquad P(X \leq X_{0.5}) \geq 1/2$

$P(X > X_{0.5}) \leq 0.5 \qquad P(X \geq X_{0.5}) \geq 1/2$

**Βήματα του Ελέγχου:**

1. Δημιουργούμε τις διαφορές $D_i = X_i - m$. (εξαιρούμε τις τιμές $D_i = 0$. Άρα n μέγεθος διαφορών $n \leq n'$)
2. Κατασκευάζουμε τις απόλυτες διαφορές $|D_i| = |X_i - m|$
3. Διατάσσουμε τις απόλυτες διαφορές $|D_i|$ και αντιστοιχούμε βαθμούς $R(|D_i|)$ από το 1(μικρότερη) έως n (μεγαλύτερη). Αν έχουμε απόλυτες διαφορές που είναι ίσες βάζουμε τον μέσο όρο.
4. Δημιουργούμε την προσημασμένη τάξη μεγέθους

$R_i = \begin{cases} + R(|D_i|), & \text{αν } D_i = X_i - m < 0 \\ - R(|D_i|), & \text{αν } D_i = X_i - m > 0 \end{cases}$ Η μεταβλητή R ονομάζεται προσημασμένος βαθμός ή προσημασμένη τάξη μεγέθους.

Έστω $S = \sum R_i$ Αποδεικνύεται ότι $E(S) = E(\sum R_i) = 0$ και $V(S) = \sum_{i=1}^{n} R_i^2$

Η ελεγχοσυνάρτηση $T = \frac{\sum_{i=1}^{n} R_i}{\sqrt{\sum_{i=1}^{n} R_i^2}}$ είναι η τυποποιημένη μορφή του αθροίσματος των προσημασμένων τάξεων μεγέθους των διαφορών $|D_i| = |X_i - m|$

- Λαμβάνει υπόψη το μέγεθος των διαφορών
- Λαμβάνει υπόψη το πρόσημο των διαφορών

Εάν δεν υπάρχουν τιμές ίδιες στα $R_i$ τότε

$$T = \frac{\sum R_i}{\sqrt{n(n+1)(2n+1)/\sigma}}$$

Η κατανομή της Τ προσεγγίζεται ικανοποιητικά από την τυποποιημένη κανονική κατανομή.

Πρέπει να σημειωθεί εδώ ότι στην περίπτωση που οι τιμές των παρατηρήσεων του δείγματος $R_1, R_2, ..., R_n$ είναι διακεκριμένες συχνά χρησιμοποιείται η στατιστική συνάρτηση

$$T^+ = \text{άθροισμα των θετικών } \mathrm{R_i}$$

για τον έλεγχο των υποθέσεων Α, Β και Γ. Τα ποσοστιαία σημεία $w_p$ της κατανομής της στατιστικής αυτής συνάρτησης περιέχονται στον πίνακα 8 του παραρτήματος του βιβλίου.

| Α. Μονόπλευρος Έλεγχος | Β. Μονόπλευρος Έλεγχος | Γ. Αμφίπλευρος Έλεγχος |
|---|---|---|
| Απορ. $H_O$ αν $T^+ < w_a$ | Απορ. $H_O$ αν $T^+ > w_{1-a}$ | Απορ. $H_O$ αν $T^+ > W_{1-a/2}$ ή $T^+ < W_{\alpha/2}$ |
| ή $T < Z_a (= -Z_{1-a})$ | ή $T > Z_{1-a}$ | Απορ. $H_O$ αν $T > Z_{1-a/2}$ ή $T < Z_{\alpha/2}$ |

**Έλεγχος Wilcoxon για ένα δείγμα παρατηρήσεων.**

**Αμφίπλευρος έλεγχος**: $H_o : X_{0.5} = m$, $H_1 : X_{0.5} \neq m$

(1) $T^+ = \sum R_i$ (για θετικές διαφορές) όταν δεν υπάρχουν ισοβαθμίες (είναι όλες διακεκριμένες.

Απορρίπτουμε την $H_o$ αν $T^+ > w_{1-a/2}$ ή $T^+ < w_{a/2}$. Στους πίνακες δίνονται οι τιμές του $w_{a/2}$, για $w_p$ όπου $p \leq 0.5$.

Για $p > 0.5$ χρησιμοποιούμε $w_p = \frac{n(n+1)}{2} - w_{1-p}$ δηλαδή $w_{1-\alpha/2} = \frac{n(n+1)}{2} - w_{\alpha/2}$. π.χ $w_{0.95} = \frac{n(n+1)}{2} - w_{0.05}$.

(2) Όταν έχουμε ισοβαθμίες στις τάξεις μεγέθους και αν n>50 χρησιμοποιούμε την ελεγχοσυνάρτηση $T = \frac{\sum R_i}{\sqrt{\sum R_i^2}}$. Απορρίπτουμε την $H_o$ αν $T > Z_{1-\alpha/2}$ ή $T < Z_{\alpha/2}$

**Μονόπλευρος έλεγχος**

- $H_o : X_{0.5} \geq m$, $H_1 : X_{0.5} < m$ Απορρίπτουμε την $H_o$ αν $T^+ < w_a$ (1) ή $T < Z_a = -Z_{1-a}$ (2).
- $H_o : X_{0.5} \leq m$, $H_1 : X_{0.5} > m$ Απορρίπτουμε την $H_o$ αν $T^+ > w_{1-a}$ (1) ή $T > Z_{1-a}$ (2)

**3.1.2 Έλεγχος Προσημασμένων Τάξεων Μεγέθους του Wilcoxon για δείγμα ζεύγους παρατηρήσεων.**

- Δείγμα ζεύγους παρατηρήσεων $(x_1, y_1), (x_2, y_2), ..., (x_{n'}, y_{n'})$
- Θεωρούμε διαφορές $D_i = Y_i - X_i$ για i=1,2,3,…,n΄.
- Υπολογίζουμε $|D_i| = |Y_i - X_i|$, i=1,2,3,…,n΄,αγνοούμε τα ζεύγη για τα οποία $Y_i = X_i$ άρα $D_i = 0$. ($n \leq n'$)
- Βρίσκουμε τις τάξεις μεγέθους $R(|D_i|)$ με τιμές 1 έως n αν δεν υπάρχουν ίδιες τιμές για τα $|D_i|$.
- Υπολογίζουμε $R_i = \begin{cases} +R(|D_i|), \text{αν } D=Y_i - X_i > 0 \\ -R(|D_i|), \text{αν } D=Y_i - X_i < 0 \end{cases}$

Με $d_{0.5}$ συμβολίζουμε την διάμεσο του πληθυσμού των διαφορών D=Y-X

αν Y>X τότε d>0 (αν οι τιμές της Y τείνουν να είναι μεγαλύτερες από της X η διάμεσος διαφορά είναι θετική).

αν Y<X τότε d<0

| Α. Μονόπλευρος Έλεγχος | Β. Μονόπλευρος Έλεγχος | Γ. Αμφίπλευρος Έλεγχος |
|---|---|---|
| $H_O : d_{0.5} \geq m$ | $H_O : d_{0.5} \leq m$ | $H_O : d_{0.5} = m$ |
| $H_1 : d_{0.5} < m$ | $H_1 : d_{0.5} > m$ | $H_O : d_{0.5} \neq m$ |

Για τον έλεγχο των υποθέσεων των περιπτώσεων Α, Β και Γ χρησιμοποιείται η στατιστική συνάρτηση $T = \frac{\sum R_i}{\sqrt{\sum R_i^2}}$, ή όταν όλες οι διαφορές $D_i = Y_i - X_i$ είναι διακεκριμένες χρησιμοποιούμε την $T = \frac{\sum R_i}{\sqrt{n(n+1)(2n+1)/6}}$ ή $T^+ = $ άθροισμα των θετικών $R_i$

| Α. Μονόπλευρος Έλεγχος | Β. Μονόπλευρος Έλεγχος | Γ. Αμφίπλευρος Έλεγχος |
|---|---|---|
| Απορ. $H_O$ αν $T^+ < w_a$ | Απορ. $H_O$ αν $T^+ > w_{1-a}$ | Απορ. $H_O$ αν $T^+ > W_{1-\alpha/2}$ ή $T^+ < W_{\alpha/2}$ |
| ή $T < Z_a (= -Z_{1-a})$ | ή $T > Z_{1-a}$ | Απορ. $H_O$ αν $T > Z_{1-\alpha/2}$ ή $T < Z_{\alpha/2}$ |

**3.2 Περίπτωση Ανεξάρτητων Δειγμάτων. Έλεγχος Mann-Whitney, Wilcoxon.**

- Δύο ανεξάρτητα δείγματα.
- Έλεγχο για διαφορά στην θέση των δύο πληθυσμών. → μέση τιμή, διάμεσο
- Αντίστοιχος παραμετρικός έλεγχος t
- Δεδομένα σε κλίμακα μέτρησης η οποία είναι τουλάχιστον διάταξης.

Έχουμε $X_1, X_2, ..., X_n$ δείγμα n και $Y_1, Y_2, ..., Y_m$ δείγμα m.

1. Ενώνουμε το δείγμα (n+m)=N
2. Βρίσκουμε τις τάξεις μεγέθους R.

Έστω $R(x_i)$, και $R(y_i)$, οι βαθμοί που αντιστοιχούν στα $X_i$ και στα $Y_i$. Έστω $F(.)$ και $G(.)$ οι συναρτήσεις των κατανομών των τυχαίων μεταβλητών X και Y, αντίστοιχα.

| Α. Μονόπλευρος Έλεγχος | Β. Μονόπλευρος Έλεγχος | Γ. Αμφίπλευρος Έλεγχος |
|---|---|---|
| $H_O : F(x) \leq G(x)$ | $H_O : F(x) \geq G(x)$ | $H_O : F(x) \geq G(x)$ |
| $H_O : F(x) > G(x)$ | $H_O : F(x) < G(x)$ | $H_O : d_{0.5} \neq m$ |

Για τον αμφίπλευρο έλεγχο θα πρέπει $P(X \le x) \ne P(Y \le x)$ για κάποιο x ή $P(X > x) \ne P(Y > x)$.

Έστω ότι οι τιμές της X τείνουν να είναι μεγαλύτερες της Y, $P(X > Y) \ne P(Y > X)$ δηλαδή $P(X < Y) \ne 1/2$.

| Α. Μονόπλευρος Έλεγχος | Β. Μονόπλευρος Έλεγχος | Γ. Αμφίπλευρος Έλεγχος |
|---|---|---|
| $H_O : P(X < Y) \le 1/2$ | $H_O : P(X < Y) \ge 1/2$ | $H_O : P(X < Y) = 1/2$ |
| $H_O : P(X < Y) > 1/2$ | $H_O : P(X < Y) < 1/2$ | $H_O : P(X < Y) \ne 1/2$ |

Στατιστική Συνάρτηση Ελέγχου

- Όταν δεν υπάρχουν περιπτώσεις ταύτισης τιμών στο δείγμα ή όταν υπάρχουν λίγες ταυτίσεις χρησιμοποιούμε την ελεγχοσυνάρτηση $T = \sum_{i=1}^{n} R(X_i)$. Τα ποσοστιαία σημεία της κατανομής της στατιστικής συνάρτησης T δίνονται στον πίνακα 9 του παραρτήματος. Ο πίνακας δίνει μόνο τα κάτω ποσοστιαία σημεία, $w_p$, p= 0.001, 0.005, 0.01, 0.025, 0.05, 0.1. Τα πάνω ποσοστιαία σημεία μπορούν να υπολογισθούν από την σχέση $w_{1-p} = n(N+1) - w_p$.
- Όταν υπάρχουν πολλές ταυτίσεις χρησιμοποιείται η:

$$T_1 = \frac{T - E(T)}{\sqrt{V(T)}} = \frac{T - \frac{n(N+1)}{2}}{\sqrt{\frac{nm}{N(N-1)} \cdot \sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}}$$

Τα ποσοστιαία σημεία της κατανομής της $T_1$ προσεγγίζονται από τα αντίστοιχα ποσοστιαία σημεία της τυποποιημένης κανονικής κατανομής.

| Α. Μονόπλευρος Έλεγχος | Β. Μονόπλευρος Έλεγχος | Γ. Αμφίπλευρος Έλεγχος |
|---|---|---|
| Απορ. $H_O$ αν $T < w_a$ | Απορ. $H_O$ αν $T > w_{1-a}$ | Απορ. $H_O$ αν $T > W_{1-a/2}$ ή $T < W_{\alpha/2}$ |
| ή $Z < Z_a (= -Z_{1-a})$ | ή $T_1 > Z_{1-a}$ | Απορ. $H_O$ αν $T_1 > Z_{1-a/2}$ ή $T_1 < Z_{\alpha/2}$ |

Σημείωση: Όταν η $F(x)$ δεν ταυτίζεται με την $G(x)$, τότε μπορεί να ταυτίζεται με την $G(x+c)$ και κάνουμε υποθέσεις που αναφέρονται στην μέση τιμή ή στην διάμεσο.

| Α. Μονόπλευρος Έλεγχος | Β. Μονόπλευρος Έλεγχος | Γ. Αμφίπλευρος Έλεγχος |
|---|---|---|
| $H_O : E(X) \geq E(Y)$ | $H_O : E(X) \leq E(Y)$ | $H_O : E(X) = E(Y)$ |
| $H_O : E(X) < E(Y)$ | $H_O : E(X) > E(Y)$ | $H_O : E(X) \neq E(Y)$ |

**Ασκήσεις**

**Άσκηση 1** (Wilcoxon για ένα δειγμα)

Λαμβάνεται απλό τυχαίο δείγμα 15 παρατηρήσεων από έναν πληθυσμό, για τον οποίο θέλουμε να ελέγξουμε αν η διάμεσος του είναι μεγαλύτερη του 30 ή όχι, σε επίπεδο σημαντικότητας α=0.05. Οι παρατηρήσεις του δείγματος είναι: 26.0 , 27.4 , 30.3 , 31.2 , 33.2 33.9 , 23.8 , 26.9 , 35.9 , 34.9 , 28.0 , 30.7 , 32.8 , 34.3 , 35.0.

Λύση

$H_o : X_{0.5} \geq 30$

$H_1 : X_{0.5} < 30$

| $X_i$ | $D_i = X_i - 30$ | $\lvert D_i \rvert = \lvert X_i - 30 \rvert$ | $R_i(\lvert D_i \rvert)$ | $R_i$ |
|---|---|---|---|---|
| 23.8 | -6.2 | 6.2 | 15 | -15 |
| 26.0 | -4 | 4 | 10 | -10 |
| 26.9 | -3.1 | 3.1 | 7 | -7 |
| 27.4 | -2.6 | 2.6 | 5 | -5 |
| 28.0 | -2.0 | 2 | 4 | -4 |
| 30.3 | 0.3 | 0.3 | 1 | 1 |
| 30.7 | 0.7 | 0.7 | 2 | 2 |
| 31.2 | 1.2 | 1.2 | 3 | 3 |
| 32.8 | 2.8 | 2.8 | 6 | 6 |
| 33.2 | 3.2 | 3.2 | 8 | 8 |
| 33.9 | 3.9 | 3.9 | 9 | 9 |
| 34.3 | 4.3 | 4.3 | 11 | 11 |
| 34.9 | 4.9 | 4.9 | 12 | 12 |
| 35.0 | 5.0 | 5.0 | 13 | 13 |
| 35.9 | 5.9 | 5.9 | 14 | 14 |

Δεν υπάρχουν περιπτώσεις ταύτισης: Η ελεγχοσυνάρτηση είναι $T^+ = \sum R_i$ (θετικών $R_i$)=79. Η περιοχή απόρριψης ορίζεται από την $T^+ < w_a = w_{0.05} = 31$.

Απορρίπτω την $H_o$ αν $T^+ < w_a$ ή αν $T^+ = 79 < w_a = 31$ είναι το 79<31 <u>όχι</u> άρα δεν απορρίπτω την $H_o$, οπότε η πραγματική διάμεσος είναι μεγαλύτερη από 30.

$$\hat{a} = \max P\left(T^+ \le 79 / H_o\right) = \max P\left(T^+ \le 79 / X_{0.5} \ge 30\right) = P\left(T^+ \le 79 / X_{0.5} = 30\right) \cong 0.85$$

Αν $a > \hat{a} = p - value \Rightarrow$ απορρίπτουμε την $\mathrm{H_o}$ εδώ έχουμε p-value 0.85 άρα δεν απορρίπτουμε την $\mathrm{H}_o$ σε α=0.05.

$$w_p = \frac{n(n+1)}{2} - w_{1-p}$$

$$w_{0.60} = \frac{n(n+1)}{2} - w_{0.40} = 120 - 55 = 65, \quad w_{0.70} = \frac{n(n+1)}{2} - w_{0.30} = 120 - 51 = 69$$

$$w_{0.80} = \frac{n(n+1)}{2} - w_{0.20} = 120 - 45 = 75, \quad w_{0.90} = \frac{n(n+1)}{2} - w_{0.10} = 120 - 37 = 83$$

**<u>Άσκηση 3.1.5 σελίδα 247 φωτοτυπία</u>**

**<u>Άσκηση 2</u>** <u>(Δείγμα Ζευγών Παρατηρήσεων)</u>

Τα ποσοστά τηλεθέασης των προγραμμάτων δύο καναλιών που προβάλλονται τις ίδιες ώρες παρουσιάζονται στον παρακάτω πίνακα:

| Χρόνος Εκπομπής | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|---|
| $X_i$ Καν.Α | 14.5 | 21.4 | 9.9 | 12.8 | 19.2 | 28.1 | 14.2 | 23.6 | 13.2 |
| $Y_i$ Καν.Α | 12.3 | 21.9 | 8.8 | 11.1 | 18.1 | 26.4 | 11.0 | 20.1 | 11.5 |

Διαφέρουν σημαντικά (α=0.05) τα ποσοστά τηλεθέασης του κοινού για τα 2 κανάλια;

<u>Λύση:</u>

- Υπόθεση συμμετρίας
- Γίνεται και με προσημικό έλεγχο

$H_o : E(x) = E(y)$ ή $\mathrm{H_o} : d_{0.5} = 0$

$H_1 : E(x) \neq E(y)$ ή $\mathrm{H_1} : d_{0.5} \neq 0$

| $X_i$ | $Y_i$ | $D_i = X_i - Y_i$ | $\lvert D_i\rvert = \lvert X_i - Y_i\rvert$ | $R_i(\lvert D_i\rvert)$ | $R_i$ | $R_i^2$ |
|---|---|---|---|---|---|---|
| 14.5 | 12.3 | 2.2 | 2.2 | 7 | 7 | 49 |
| 21.4 | 21.9 | -0.5 | 0.5 | 1 | -1 | 1 |
| 9.9 | 8.8 | 1.1 | 1.1 | 2.5 | 2.5 | 6.25 |
| 12.8 | 11.1 | 1.7 | 1.7 | 5 | 5 | 25 |
| 19.2 | 18.1 | 1.1 | 1.1 | 2.5 | 2.5 | 6.25 |
| 28.1 | 26.4 | 1.7 | 1.7 | 5 | 5 | 25 |
| 14.2 | 11.0 | 3.2 | 3.2 | 8 | 8 | 64 |
| 23.6 | 20.1 | 3.5 | 3.5 | 9 | 9 | 81 |
| 13.2 | 11.5 | 1.7 | 1.7 | 5 | 5 | 25 |
| **Σύνολο** | | | | | **43** | **282.5** |

Έχουμε ισοβαθμίες άρα: $T = \dfrac{\sum_{i=1}^{n} R_i}{\sqrt{\sum R_i^2}} = \dfrac{43}{\sqrt{282.5}} = 2.56$

Απορρίπτουμε την $H_o$ αν $T > Z_{1-a/2}$ ή αν $T < Z_{a/2}$

$Z_{1-a/2} = Z_{1-\frac{0.05}{2}} = Z_{0.975} = 1.96.$ $\quad Z_{a/2} = -Z_{1-a/2} = -1.96$

Αφού $T > Z_{1-a/2}$ ή $2.56>1.96 \Rightarrow H_o$ απορρίπτεται $\Rightarrow$ τα ποσοστά τηλεθέασης (κατά μέσο όρο ή η διάμεσος τους) των δύο καναλιών διαφέρουν.

$$\frac{\hat{a}}{2} = P(T \geq 2.56/H_o) = 1 - P(T < 2.56/H_o) = 1 - 0.9948 = 0.0052 \Rightarrow \hat{a} = 2 \cdot 0.0052$$

$$\Rightarrow \hat{a} = 0.0104$$

Αφού $a = 0.05 > \hat{a} = 0.0104 \Rightarrow$ Απορρίπτεται $H_o$

**<u>Παράδειγμα 3.2.2 (σελίδα 282)</u>**(2 ανεξάρτητα δείγματα)

| Προέλευση Τεμαχίου | A | A | A | B | A | B | B | B | B |
|---|---|---|---|---|---|---|---|---|---|
| Βαθμός Σκληρότητας | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

α=0.05

$H_o$ : τα 2 είδη πυρόλιθου δεν διαφέρουν ως προς τον βαθμό σκληρότητας

$H_1$ : τα 2 είδη πυρόλιθου διαφέρουν ως προς τον βαθμό σκληρότητας

Δεν υπάρχουν περιπτώσεις ταύτισης

$$T = \sum_{i=1}^{4} R(X_i)(\text{απο περιοχή A}) = 1 + 2 + 3 + 5 = 11$$

Απορρίπτουμε $H_o$ αν $T > w_{1-a/2}$ ή αν $T < w_{a/2}$ α=0.05 $w_{a/2} = w_{0.05/2} = w_{0.025} = 12$

$w_{1-a/2} = n\cdot(N+1) - w_{a/2} = 4\cdot 10 - 12 = 40 - 12 = 28$

Απορρίπτουμε $H_o$ αν 11>28 ή αν 11<12

Άρα απορρίπτω την $H_o$, άρα σε επίπεδο σημαντικότητας α=0.05 έχουμε ενδείξεις ότι τα 2 είδη πυρόλιθου διαφέρουν ως προς τον βαθμό σκληρότητας.

$$\frac{\hat{a}}{2} = MaxP(T \le 11/H_o) = P(T \le 11/H_o) = 0.01 \Rightarrow \hat{a} = 0.01\cdot 2 \Rightarrow \hat{a} = 0.02$$

Αφού $a = 0.05 > \hat{a} = 0.02 \Rightarrow$ απορρίπτουμε την $H_o$

Πίνακα 9. (φωτοτυπίας)

**<u>Άσκηση</u>**

Οι βαθμοί στο μάθημα της στατιστικής για ένα δείγμα n=15 φοιτητών από την asoee δίνονται στην στήλη X. Οι βαθμοί στο μάθημα της στατιστικής για ένα δείγμα 17 φοιτητών στο ίδιο μάθημα από κάποιο επαρχιακό φορέα δίνονται στην στήλη Y. α=0.05

Να ελεγχθεί αν οι βαθμοί των φοιτητών της επαρχίας τείνουν να είναι μεγαλύτεροι από αυτούς των φοιτητών της πρωτεύουσας,

Λύση:

(Βλέπε πίνακα που ακολουθεί)

$$H_o: P(X<Y) \le 1/2 \qquad \text{ή} \qquad H_o: E(X) \ge E(Y)$$
$$H_1: P(X<Y) > 1/2 \qquad\qquad H_1: E(X) < E(Y)$$

n=15, m=17, n+m=32 $\sum_{i=1}^{n} R(X_i) = 243.5$ $\sum_{i=1}^{N} R_i^2 = 11398.5$ $T = \sum_{i=1}^{n} R(X_i) = 243.5$

$$T_1 = \frac{T - \frac{n\cdot(N+1)}{2}}{\sqrt{\frac{nm}{N(N-1)}\cdot\sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}} = \frac{243.5 - \frac{15\cdot 33}{2}}{\sqrt{\frac{15\cdot 17}{32\cdot 31}\cdot 11398.5 - \frac{15\cdot 17\cdot 33^2}{4\cdot 31}}} =$$

$$= \frac{243.5 - 247.5}{\sqrt{2930.05 - 2239.47}} = \frac{-4}{\sqrt{690.58}} = \frac{-4}{26.27} = -0.15$$

Απορρίπτω $H_o$ αν $T_1 < Z_a = -Z_{1-a} = -Z_{0.95} = -1.645$.

Αφού $T_1 = -0.15$ δεν είναι < -1.645 $\Rightarrow$ δεν απρρίπτεται η $H_o$

$\hat{\alpha} = P(T_1 \le -0.15/H_o) = P(Z \le -0.15) = 0.4404$ σε α=0.05 δεν απορρίπτεται η $H_o$ αφού α=0.05 δεν είναι μεγαλύτερο από το $\hat{a} = 0.4404$.

| | C1 | C2 | C3 | C4 | C5 | C6 | C8 |
|---|---|---|---|---|---|---|---|
| | asoee | eparxia | enop_sam | categ | ranks(Ri) | Ri^2 | |
| **1** | 7 | 6 | 7 | 1 | 19.0 | 361.0 | 19.0 |
| **2** | 8 | 5 | 8 | 1 | 23.5 | 552.3 | 23.5 |
| **3** | 6 | 3 | 6 | 1 | 14.5 | 210.3 | 14.5 |
| **4** | 9 | 7 | 9 | 1 | 28.0 | 784.0 | 28.0 |
| **5** | 4 | 9 | 4 | 1 | 6.5 | 42.3 | 6.5 |
| **6** | 2 | 8 | 2 | 1 | 1.5 | 2.3 | 1.5 |
| **7** | 6 | 6 | 6 | 1 | 14.5 | 210.3 | 14.5 |
| **8** | 5 | 9 | 5 | 1 | 10.5 | 110.3 | 10.5 |
| **9** | 8 | 10 | 8 | 1 | 23.5 | 552.3 | 23.5 |
| **10** | 9 | 5 | 9 | 1 | 28.0 | 784.0 | 28.0 |
| **11** | 10 | 2 | 10 | 1 | 31.5 | 992.3 | 31.5 |
| **12** | 4 | 3 | 4 | 1 | 6.5 | 42.3 | 6.5 |
| **13** | 5 | 7 | 5 | 1 | 10.5 | 110.3 | 10.5 |
| **14** | 7 | 8 | 7 | 1 | 19.0 | 361.0 | 19.0 |
| **15** | 4 | 9 | 4 | 1 | 6.5 | 42.3 | 6.5 |
| **16** | | 7 | 6 | 2 | 14.5 | 210.3 | * |
| **17** | | 4 | 5 | 2 | 10.5 | 110.3 | 243.5 |
| **18** | | | 3 | 2 | 3.5 | 12.3 | |
| **19** | | | 7 | 2 | 19.0 | 361.0 | |
| **20** | | | 9 | 2 | 28.0 | 784.0 | |
| **21** | | | 8 | 2 | 23.5 | 552.3 | |
| **22** | | | 6 | 2 | 14.5 | 210.3 | |
| **23** | | | 9 | 2 | 28.0 | 784.0 | |
| **24** | | | 10 | 2 | 31.5 | 992.3 | |
| **25** | | | 5 | 2 | 10.5 | 110.3 | |
| **26** | | | 2 | 2 | 1.5 | 2.3 | |
| **27** | | | 3 | 2 | 3.5 | 12.3 | |
| **28** | | | 7 | 2 | 19.0 | 361.0 | |
| **29** | | | 8 | 2 | 23.5 | 552.3 | |
| **30** | | | 9 | 2 | 28.0 | 784.0 | |
| **31** | | | 7 | 2 | 19.0 | 361.0 | |
| **32** | | | 4 | 2 | 6.5 | 42.3 | |
| **33** | | | | | | 11398.5 | |

### 3.2.2 Έλεγχος Kruskal-Walls

κ>2 ανεξάρτητα τυχαία δείγματα

- $H_o$ : οι συναρτήσεις κατανομής των κ πληθυσμών είναι ίσες

  $H_1$ : τουλάχιστον 2 από τους πληθυσμούς έχουν διαφορετικές μέσες τιμές

  $H_o$ : οι πληθυσμοί έχουν ίσες μέσες τιμές

  $H_1$ : τουλάχιστον 2 από τους πληθυσμούς έχουν διαφορετικές μέσες τιμές.

- Παραμετρικό ανάλογο είναι η ανάλυση διασποράς κατά ένα κριτήριο.
- Δεδομένα: κ ανεξάρτητα τυχαία δείγματα.(ίσως διαφορετικού μεγέθους)

| Δείγμα 1 | Δείγμα 2 | …. | Δείγμα κ |
|---|---|---|---|
| $X_{1,1}$ | $X_{2,1}$ | .... | $X_{k,1}$ |
| $X_{1,2}$ | $X_{2,2}$ | …. | $X_{k,2}$ |
| ⋮ | ⋮ | …. | ⋮ |
| $X_{1,n_1}$ | $X_{2,n_2}$ | …. | $X_{k,n_k}$ |

$$N = \sum_{i=1}^{k} n_i$$

- Κλίμακα μέτρησης: τουλάχιστον κλίμακα διάταξης

Μεθοδολογία:

- Ενοποιούμε τα δείγματα
- Βρίσκουμε τις τάξεις μεγέθους.
- Υπολογίζουμε $R_i = \sum_{j=1}^{n_i} R(X_{ij}), \mathrm{i} = 1,2,...,\mathrm{k}$ (άθροισμα τάξεων μεγέθους του i δείγματος).

Ελεγχοσυνάρτηση:

1) Για ισοβαθμίες

- $T = \frac{1}{S^2}\left(\sum_{i=1}^{k} \frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\right)$
- $S^2 = \frac{1}{N-1}\left(\sum_{i,j} R(X_{i,j})^2 - \frac{N(N+1)^2}{4}\right)$

2) Χωρίς ισοβαθμίες

- $T = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$
- $S^2 = \frac{N(N+1)}{12}$

Τα κρίσιμα σημεία της κατανομής της στατιστικής συνάρτησης Τα περιέχονται στον πίνακα 26 του παραρτήματος για κ=3, και $n_i \leq 5$ (i=1,2,3) για την περίπτωση που δεν υπάρχουν ισοβαθμούσες παρατηρήσεις. Η κατανομή της Τα προσεγγίζεται ικανοποιητικά από την κατανομή $x^2$ με k -1 βαθμούς ελευθερίας.

- Σε περίπτωση που απορρίψουμε την μηδενική υπόθεση $H_0$, ακολουθούμε την διαδικασία των πολλαπλών συγκρίσεων για να προσδιορίσουμε τα ζεύγη εκείνα του πληθυσμού τα οποία διαφέρουν.
- Στην περίπτωση αυτή έχουμε ότι οι πληθυσμοί i και j διαφέρουν αν ισχύει $\left| \frac{R_i}{n_i} - \frac{R_j}{n_j} \right| > t_{N-K,1-\frac{a}{2}} \cdot \left( s^2 \cdot \frac{N-1-T}{N-K} \right)^{\frac{1}{2}} \cdot \left( \frac{1}{n_i} + \frac{1}{n_j} \right)^{\frac{1}{2}}$.

**3.4 Μέτρα συσχέτισης τάξης μεγέθους**

Σε αυτή την περίπτωση έχουμε τις ακόλουθες ιδιότητες ώστε να είναι κάποιο μέτρο συσχέτισης μεταξύ των μεταβλητών Χ και Υ στατιστικά αποδεκτό:

- Η τιμή του μέτρου συσχέτισης θα πρέπει πάντα να είναι μεταξύ -1 και 1
- Όταν έχουμε θετική συσχέτιση τότε το μέτρο συσχέτισης θα πρέπει να είναι θετικό και η τιμή του να τείνει προς την τιμή +1.
- Όταν έχουμε αρνητική συσχέτιση τότε το μέτρο συσχέτισης θα πρέπει να είναι πάντα αρνητικό και η τιμή του να τείνει προς την τιμή -1.
- Όταν δεν υπάρχει σχέση μεταξύ των μεταβλητών Χ και Υ τότε θα πρέπει το μέτρο συσχέτισης θα παίρνει την τιμή 0.

**Συντελεστής συσχέτισης του Pearson**

Ο συντελεστής συσχέτισης του Pearson είναι ο $r=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\cdot\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}}$ ο οποίος παίρνει την μορφή $r=\frac{\sigma_{12}}{\sqrt{\sigma_1^2\cdot\sqrt{\sigma_2^2}}}=\frac{\frac{1}{n}\cdot\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)}{\sqrt{\frac{1}{n}\cdot\Sigma\left(x_i-\bar{x}\right)^2}\cdot\sqrt{\frac{1}{n}\cdot\Sigma\left(y_i-\bar{y}\right)^2}}$.

Στο σημείο αυτό θα αναφέρουμε ότι η κατανομή του συντελεστή r εξαρτάται από την διμεταβλητή κατανομή του διανύσματος (X,Y).

Στην συνέχεια θα εισάγουμε κάποια μέτρα συσχέτισης τα οποία εξαρτώνται από τις τάξεις μεγέθους και έχουν κατανομές που είναι ανεξάρτητες από την κατανομή του διανύσματος (X,Y).

**Συντελεστής Spearman**

Έστω ότι έχουμε $(X_n,Y_n)$ ένα δείγμα n παρατηρήσεων. Τα δεδομένα μας μπορούν να είναι συνεχείς τυχαίες μεταβλητές ή να έχουν μετρήσεις σε κλίμακα διάταξης (βαθμός προτίμησης προϊόντος κτλ). Τότε ορίζουμε ως συντελεστή συσχέτισης Spearman την αριθμητική ποσότητα $\rho=\frac{\sum_{i=1}^{n}\left[R\left(X_i\right)-\overline{R\left(X\right)}\right]\cdot\left[R\left(Y_i\right)-\overline{R(Y)}\right]}{\left(\sum_{i=1}^{n}\left(\left(R\left(X_i\right)-\overline{R(X)}\right)^2\right)\right)^{\frac{1}{2}}\cdot\left(\sum_{i=1}^{n}\left(R\left(Y_i\right)-\overline{R\left(Y\right)}\right)^2\right)^{\frac{1}{2}}}$ όπου $\overline{R\left(X\right)}=\frac{\sum_{i=1}^{n}R\left(X_i\right)}{n}$ και $\overline{R\left(Y\right)}=\frac{\sum_{i=1}^{n}R(Y_i)}{n}$. Ουσιαστικά ο συντελεστής Spearman είναι ο συντελεστής Pearson αλλά με τις τάξεις μεγέθους αντί των X και Y.

- Περίπτωση όπου δεν υπάρχουν ισοπαλίες

$$\overline{R(X)}=\frac{1}{n}\cdot\sum_{i=1}^{n}R\left(X_i\right)=\frac{1}{n}\sum_{i=1}^{n}i=\frac{1}{n}\cdot\frac{n\left(n+1\right)}{2}=\frac{n+1}{2}$$

$$\overline{R\left(x\right)}=\frac{n+1}{2}$$

$$\sum_{i=1}^{n}\left[R(X_i)-\overline{R(X)}\right]^2=\frac{n\cdot\left(n^2-1\right)}{12}$$

Άρα όταν έχουμε διακριτές τιμές ο συντελεστής μπορεί να πάρει την μορφή

$$\rho=\frac{\sum_{i=1}^{n}\left(R(X_i)-\frac{n+1}{2}\right)\left(R(Y_i)-\frac{n+1}{2}\right)}{\frac{n\left(n^2-1\right)}{12}}\Rightarrow\rho=1-\frac{6T}{n\left(n^2-1\right)}$$ όπου

$$\mathrm{T}=\sum_{i=1}^{n}\left(R(X_i)-R(Y_i)\right)^2.$$

- Περίπτωση όπου έχουμε αρκετές ισοπαλίες

Σε αυτή την περίπτωση χρησιμοποιείται άλλος τύπος από τον αρχικό τύπο του Spearman. Ο τύπος αυτός έχει την ακόλουθη μορφή

$$\rho=\frac{\sum_{i=1}^{n}R(X_i)R(Y_i)-n\left[\frac{n+1}{2}\right]^2}{\sqrt{\sum_{i=1}^{n}R(X_i)^2-n\left(\frac{n+1}{2}\right)^2}\sqrt{\sum_{i=1}^{n}R(Y_i)^2-n\left(\frac{n+1}{2}\right)^2}}$$

Επίσης στο σημείο αυτό να αναφέρουμε ότι ο συντελεστής συσχέτισης του Spearman χρησιμοποιείται αρκετές φορές ως ελεγχοσυνάρτηση για τον έλεγχο ανεξαρτησίας μεταξύ δυο τυχαίων μεταβλητών. Στην περίπτωση αυτή έχουμε τις εξής δυνατές υποθέσεις ελέγχου:

**Αμφίπλευρος έλεγχος**

$H_0$: Οι μεταβλητές X και Y είναι αμοιβαία ανεξάρτητες (δεν υπάρχει συσχέτιση)

$H_1$: Υπάρχει είτε τάση για θετική συσχέτιση είτε τάση για αρνητική συσχέτιση

Στην περίπτωση του αμφίπλευρου ελέγχου απορρίπτουμε την μηδενική υπόθεση $H_0$ αν $\rho<w_{a/2}$ ή αν $\rho>w_{1-a/2}$.

**Μονόπλευρος έλεγχος για θετική συσχέτιση**

$H_0$: Οι μεταβλητές X και Y είναι αμοιβαία ανεξάρτητες

$H_1$: Οι μεταβλητές X και Y είναι θετικά συσχετισμένες

Στην περίπτωση αυτή απορρίπτουμε την μηδενική υπόθεση αν $\rho>w_{1-a}$.

**Μονόπλευρος έλεγχος για αρνητική συσχέτιση**

$H_0$: Οι μεταβλητές X και Y είναι αμοιβαία ανεξάρτητες

$H_1$: Οι μεταβλητές X και Y είναι αρνητικά συσχετισμένες

Στην περίπτωση αυτή απορρίπτουμε την μηδενική υπόθεση $H_0$ αν $\rho < w_a$

**Παράδειγμα 3.4.1 (σελ. 329)**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathbf{X_i}$ | 86 | 71 | 77 | 68 | 91 | 72 | 77 | 91 | 70 | 71 | 88 | 87 |
| $\mathbf{Y_i}$ | 88 | 77 | 76 | 64 | 96 | 72 | 65 | 90 | 65 | 80 | 81 | 72 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathbf{R(X_i)}$ | 8 | 3.5 | 6.5 | 1 | 11.5 | 5 | 6.5 | 11.5 | 2 | 3.5 | 10 | 9 |
| $\mathbf{R(Y_i)}$ | 10 | 7 | 6 | 1 | 12 | 4.5 | 2.5 | 11 | 2.5 | 8 | 9 | 4.5 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathbf{R(X_i)-R(Y_i)}$ | -2 | -3.5 | 0.5 | 0 | 0.5 | 0.5 | 4 | 0.5 | -0.5 | -4.5 | 1 | 4.5 |
| $\mathbf{[R(X_i)-R(Y_i)]^2}$ | 4 | 12.25 | 0.25 | 0 | 0.25 | 0.25 | 16 | 0.25 | 0.25 | 20.25 | 1 | 20.25 |

Στο συγκεκριμένο πρόβλημα θα θεωρήσουμε ως $H_0$: οι X και Y ανεξάρτητες και ως $H_1$: υπάρχει τάση μεταξύ των X και Y. Από τα δεδομένα του προβλήματος μας έχουμε ότι

$$\rho = 1 - \frac{6\sum_{t=1}^{T}\left[R(X_i) - R(Y_i)\right]^2}{n(n^2-1)} = 1 - \frac{6 \cdot 75}{12 \cdot (144-1)} = 1 - \frac{450}{1716} = 1 - 0,262 = 0,738$$

Γνωρίζουμε ότι απορρίπτουμε την $H_0$ αν $\rho < w_{a/2}$ ή αν $\rho > w_{1-a/2}$. Επίσης βρίσκουμε ότι $w_{0.975} = -0,5804$ και $w_{0.025} = -w_{0.975} = -0.5804$ αφού $w_p = -w_{1-p}$ (σε επίπεδο στατιστικής σημαντικότητας α=0,05).

**Συντελεστής τ του Kendall (συντελεστής εναρμόνισης του Kendall)**

Ο συγκεκριμένος συντελεστής μπορούμε να πούμε ότι τείνει σχετικά γρήγορα στην κανονική κατανομή. Στο σημείο αυτό να αναφέρουμε ότι δυο παρατηρήσεις $(X_j,Y_j)$ και $(X_κ,Y_κ)$ λέγονται εναρμονισμένες ή συσχετισμένες αν ισχύει $X_j>X_k$ και $Y_j>Y_k$ ή αν ισχύει $X_j<X_k$ και $Y_j<Y_k$ .

Επίσης δυο παρατηρήσεις $(X_j,Y_j)$ και $(X_\kappa,Y_\kappa)$ θα λέγονται μη εναρμονισμένες ή συσχετισμένες αν ισχύει $X_j>X_k$ και $Y_j<Y_k$ ή αν ισχύει $X_j<X_k$ και $Y_j>Y_k$ .

Όμοια δυο ζεύγη παρατηρήσεων $(X_j,Y_j)$ και $(X_\kappa,Y_\kappa)$ θα λέγονται εναρμονισμένα αν οι διαφορές τους έχουν το ίδιο πρόσημο και αντίστοιχα δυο ζεύγη παρατηρήσεων $(X_j,Y_j)$ και $(X_\kappa,Y_\kappa)$ θα λέγονται μη εναρμονισμένα αν οι διαφορές τους έχουν αντίθετο πρόσημο.

Έστω $N_c$ ο αριθμός των εναρμονισμένων ζευγών παρατηρήσεων και $N_d$ ο αριθμός των μη εναρμονισμένων ζευγών παρατηρήσεων. Επίσης έστω $N_0$ ο αριθμός των ισοβαθμούντων ζευγών παρατηρήσεων ($X_j=X_\kappa$ ή/και $Y_j=Y\kappa$). Τότε ορίζουμε ως μέτρο συσχέτισης τ του Kendall την αριθμητική ποσότητα $T=\frac{N_c-N_D}{\binom{n}{2}}=\frac{N_c-N_D}{\frac{n(n-1)}{2}}$ όπου το τ παίρνει τιμές από το -1 έως το +1 ($\tau\in(-1,+1)$). Ουσιαστικά η φιλοσοφία για τον συγκεκριμένο συντελεστή είναι να διατάσσουμε τις τιμές της μεταβλητής X και να συγκρίνουμε τις τιμές της μεταβλητής Y.
Σε ότι αφορά τώρα τους ελέγχους υποθέσεων χρησιμοποιούμε ως ελεγχοσυνάρτηση την στατιστική συνάρτηση $T=N_c-N_d$. Μπορούμε εδώ, όπως σε κάθε στατιστικό έλεγχο υποθέσεων, να διακρίνουμε τις τρεις γνωστές κατηγορίες ελέγχου υποθέσεων:

**Αμφίπλευρος έλεγχος συσχέτισης**

Στον έλεγχο αυτό απορρίπτουμε την μηδενική υπόθεση $H_0$ αν $T<w_{a/2}$ ή αν $T>w_{1-a/2}$ .

**Μονόπλευρος έλεγχος θετικής συσχέτισης**

Στην περίπτωση αυτή απορρίπτουμε την μηδενική υπόθεση αν $T>w_{1-a}$

**Μονόπλευρος έλεγχος αρνητικής συσχέτισης**

Στην περίπτωση αυτή απορρίπτουμε την μηδενική υπόθεση αν $T<w_a=-w_{1-a}$

**Παράδειγμα 3.4.3 (σελ.342)**

| $X_i$ | $Y_i$ | $X^{(i)}$ | $Y_i^*$ | Εναρμονισμένα ζεύγη | Μη εναρμονισμένα ζεύγη |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| 86 | 88 | 68 | 64 | 11 | 0 |
| 71 | 77 | 70 | 65 | 9 | 0 |
| 77 | 76 | 71 | 77 | 4 | 4 |
| 68 | 64 | 71 | 80 | 4 | 4 |
| 91 | 96 | 72 | 72 | 5 | 1 |
| 72 | 72 | 77 | 65 | 5 | 0 |
| 77 | 65 | 77 | 76 | 4 | 1 |
| 91 | 90 | 86 | 88 | 2 | 2 |
| 70 | 65 | 87 | 72 | 3 | 0 |
| 71 | 80 | 88 | 81 | 2 | 0 |
| 88 | 81 | 91 | 90 | 0 | 0 |
| 87 | 72 | 91 | 96 | 0 | 0 |
| | | | | $N_c$=49 | $N_d$=12 |

Από τον παραπάνω πίνακα μπορούμε να συμπεράνουμε ότι

$\tau = \frac{N_c - N_d}{\frac{n(n-1)}{2}} = \frac{49-12}{\frac{12\cdot 11}{2}} = 0.5606$ και $\mathrm{T} = N_c - N_d = 49-12 = 37$. Στο συγκεκριμένο πρόβλημα έχουμε ως $H_0$: οι μεταβλητές X και Y ασυσχέτιστες και ως $H_1$: οι μεταβλητές X και Y συσχετισμένες. Επίσης βρίσκουμε ότι $w_{0.975} = 28$ και $w_{0.025} = -w_{0.975} = -28$. Άρα λοιπόν μπορούμε να συμπεράνουμε ότι επειδή $T > w_{1-a/2}$ απορρίπτουμε την $H_0$ σε επίπεδο στατιστικής σημαντικότητας α=0,05 και συνεπώς οι δυο μεταβλητές του προβλήματος μας X και Y είναι συσχετισμένες.

## Κεφάλαιο 4
## ΕΛΕΓΧΟΙ ΚΑΤΑΝΟΜΩΝ

Με τους ελέγχους υποθέσεων ουσιαστικά ενδιαφερόμαστε στο να ελέγξουμε διαφορές ανάμεσα σε πληθυσμούς που αφορούν χαρακτηριστικά όπως η μέση τιμή, η διάμεσος, τα ποσοστιαία σημεία, η διασπορά. Οι έλεγχοι αυτοί δεν αποκαλύπτουν όμως διαφορές σε άλλα χαρακτηριστικά του πληθυσμού. Άρα μας ενδιαφέρει όταν ελέγχουμε υποθέσεις για την άγνωστη κατανομή πιθανότητας μιας τ.μ. να κάνουμε μια υπόθεση η οποία θα αναφέρεται

ταυτόχρονα σε όλα τα ποσοστιαία σημεία και όλες τις πιθανότητες. Ουσιαστικά μιλάμε για έναν έλεγχο ο οποίος δίνει απάντηση στο ερώτημα «Αποτελούν οι παρατηρήσεις μας δείγμα από κάποια συγκεκριμένη κατανομή;».Υποθέσεις όπως οι παραπάνω μπορούν να ελεγχθούν με ελέγχους καλής προσαρμογής (goodness of fit tests) όπως είναι οι ακόλουθοι:

- $X^2$ έλεγχος καλής προσαρμογής (Pearson)
- Kolmogorov test (Kolmogorov)

Με άλλα λόγια, μιλάμε για ελέγχους οι οποίοι θα μπορούσαν να εξετάσουν την υπόθεση ότι οι τιμές μιας μεταβλητής X ακολουθούν την U (0,1) , U (0,5) , Normal (0,3) κτλ.

## 4.1 $X^2$ έλεγχος καλής προσαρμογής

Στον συγκεκριμένο έλεγχο μιλάμε για υποθέσεις της μορφής

$H_0 : F_x(x) = F_0(x)$

$H_1 : F_x(x) \neq F_0(x)$ όπου $F_x(x) = P(X \leq x)$ είναι η συνάρτηση κατανομής της τυχαίας μεταβλητής X.

Έστω ότι έχουμε ένα τυχαίο δείγμα παρατηρήσεων μεγέθους $n'$ οι οποίες ταξινομούνται σε n κατηγορίες (κλάσεις) όπως φαίνεται στον ακόλουθο πίνακα

| Κλάση i | 1 | 2 | 3 | … | n | Σύνολο |
|---|---|---|---|---|---|---|
| $O_i$ | $O_1$ | $O_2$ | $O_3$ | … | $O_n$ | $\sum_{i=1}^{n} O_i = n'$ |

όπου ο συμβολισμός $O_i$ είναι ο αριθμός των παρατηρήσεων στην i κατηγορία.

Έστω επίσης $p_t^0$ να είναι η πιθανότητα μια παρατήρηση της τυχαίας μεταβλητής x να ανήκει στην i κατηγορία, κάτω από τον ($H_0$) μηδενική υπόθεση ότι η $F_0(x)$ είναι η συνάρτηση κατανομής της x. Τότε ισχύει ότι $p_i^0 = P(x_j \in i \text{ κατηγορία } / H_0)$ για κάθε $j, i$. Τότε μπορούμε να δούμε εύκολα ότι ο αναμενόμενος αριθμός των παρατηρήσεων στην κατηγορία i , κάτω από την $H_0$ είναι $E_i = n' \cdot p_i^0$. Στον συγκεκριμένο έλεγχο καλής προσαρμογής χρησιμοποιούμε ως στατιστική συνάρτηση ελέγχου την $T = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$, όπου n είναι ο αριθμός των κλάσεων. Η κατανομή της συγκεκριμένης στατιστικής συνάρτησης ελέγχου T είναι η $X_{n-1}^2$. Άρα μπορούμε να δούμε από τα παραπάνω ότι μεγάλες τιμές της T συνεπάγονται μεγάλες αποκλίσεις των παρατηρούμενων τιμών και των αναμενόμενων τιμών,

και έτσι απορρίπτουμε την μηδενική υπόθεση $H_0$ αν ισχύει $T > X^2_{n-1,1-a}$. Αξίζει στο σημείο αυτό να αναφέρουμε ότι η $F_0(x)$ μπορεί να έχει γνωστές παραμέτρους όπως π.χ $H_0 : F_x(x) = F_{Poisson(3)}(x)$ ή $H_0 : F_x(x) = F_{Normal(10,5)}(x)$. Επίσης μπορεί να έχει και άγνωστες παραμέτρους όπως π.χ $H_0 : F_x(x) = F_{Poisson(\lambda)}(x)$ ή $H_0 : F_x(x) = F_{Normal(\mu,\sigma^2)}(x)$. Από τα παραπάνω μπορεί να γίνει εμφανές ότι η $F_0(x)$ καθορίζει μια οικογένεια κατανομών. Όταν έχουμε τέτοιες περιπτώσεις εκτιμώνται οι άγνωστες παράμετροι και η κατανομή της Τ είναι $X^2_{n-1-k}$ όπου κ είναι ο αριθμός των παραμέτρων που εκτιμώνται. Άρα δηλαδή έχουμε ότι $T = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim X^2_{n-k-1}$. Σε όλα τα παραπάνω που αναφέραμε υπάρχει πρόβλημα αν κάποιες αναμενόμενες συχνότητες $E_i$ είναι μικρές όπου και η προσέγγιση της στατιστικής συνάρτησης ελέγχου Τ από την κατανομή $X^2$ δεν είναι ικανοποιητική.

**Παράδειγμα 4.11 (σελ.395)**

Σε αυτό το παράδειγμα ας υποθέσουμε ένα τυχαίο δείγμα 52 τιμών που αφορούν την ζήτηση Χ για ένα συγκεκριμένο προϊόν, όπου οι τιμές αυτές ταξινομούνται με βάση τον παρακάτω πίνακα.

| Κλάση | Τιμές ζήτησης $X_i$ | $O_i$ (συχνότητες) | $p_i^0$ | $E_i = n' p_i^0$ |
|---|---|---|---|---|
| 1 | $\leq 1$ | 4 | 0,0916 | 4,763 |
| 2 | 2 | 9 | 0,1465 | 7,618 |
| 3 | 3 | 11 | 0,1954 | 10,161 |
| 4 | 4 | 7 | 0,1954 | 10,161 |
| 5 | 5 | 8 | 0,1563 | 8,128 |
| 6 | 6 | 9 | 0,1042 | 5,418 |
| 7 | 7 | 1 | 0,0595 | 3,094 |
| 8 | $\geq 8$ | 3 | 0,0511 | 2,657 |
| | | Σύνολο=52 | | |

Ουσιαστικά στο συγκεκριμένο παράδειγμα ενδιαφερόμαστε για τον έλεγχο υποθέσεων: $H_0 : F_x(x) = F_{Poisson(\lambda)}(x)$ έναντι $H_1 : F_x(x) \neq F_{Poisson(\lambda)}(x)$ σε επίπεδο στατιστικής

σημαντικότητας α=0,05. Άρα συνεπώς για να ελέγξουμε κάτι τέτοιο θα πρέπει να υπολογίσουμε τις παρακάτω ποσότητες:

$$p_1^0 = P\left(X_j \le 1 / H_0\right) = P\left(X_{Poisson(\lambda)} \le 1\right)$$

$$p_2^0 = P\left(X_j = 2 / H_0\right) = P\left(X_{Poisson(\lambda)} = 2\right)$$

.

.

.

$$p_8^0 = P\left(X_j \ge 8 / H_0\right) = P\left(X_{Poisson(\lambda)} \ge 8\right) = 1 - P\left(X_{Poisson(\lambda)} \le 7\right) = 1 - \sum_{i=1}^{7} p_i^0$$

Γνωρίζουμε στην Στατιστική ότι η παράμετρος λ είναι η μέση τιμή της κατανομής Poisson. Άρα στην περίπτωση μας έχουμε ότι η εκτιμήτρια ης παραμέτρου λ είναι η ακόλουθη:

$$\hat{\lambda} = \bar{x} = \frac{\sum_{t=1}^{n'} x_t}{n} \text{ ή } \hat{\lambda} = \bar{x} = \frac{\sum_{i=1}^{n} m_i \cdot O_i}{\sum O_i} = \frac{\sum_{i=1}^{n} m_i \cdot O_i}{n} \simeq 4$$

Οπότε τώρα έχουμε ότι $P\left(X_{Poisson(4)} = x\right) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} = \frac{e^{-4} \cdot 4^x}{x!}$ και ότι

$T = \sum_{i=1}^{8} \frac{\left(O_i - E_i\right)^2}{E_i} \sim X^2_{n-k-1} = X^2_{8-1-1} = X^2_6$. Από τα παραπάνω βγάζουμε ότι T=5,256 και ότι $X^2_6$=12,59 άρα λοιπόν δεν απορρίπτουμε την μηδενική υπόθεση $H_0$. Στο σημείο αυτό θα υπολογίσουμε και το κρίσιμο επίπεδο του ελέγχου το οποίο είναι ίσο με $\hat{a} = P\left(T \ge 5.256 / H_0\right) = 1 - P\left(T < 5.256 / H_0\right) = 1 - P\left(X^2_6 < 5.256\right) = 1 - 0.4855 = 0.5055$. Άρα λοιπόν, ανακεφαλαιώνοντας, συμπεραίνουμε από τα παραπάνω αποτελέσματα του ελέγχου υποθέσεων που κάναμε ότι σε επίπεδο στατιστικής σημαντικότητας α=0,05 , τα δεδομένα δεν παρέχουν ενδείξεις ότι η κατανομή της ζήτησης διαφέρει από την κατανομή Poisson.

**<u>Συνεχής περίπτωση</u>**

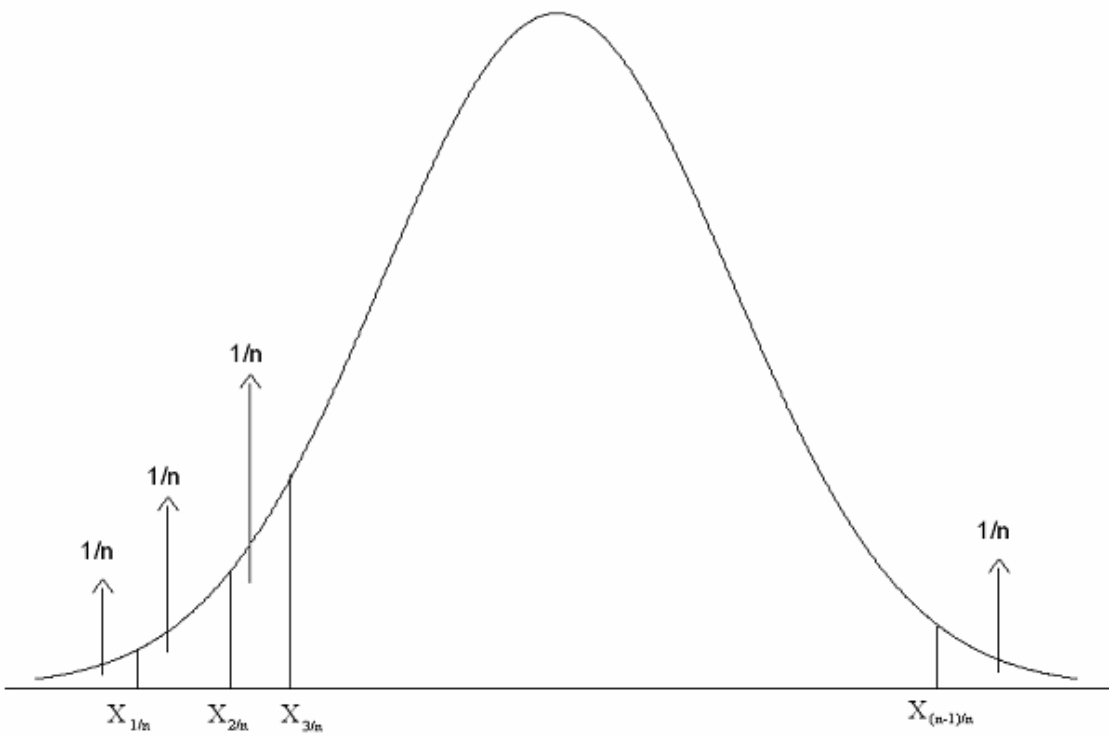
**Παράδειγμα με κανονική κατανομή**

Ο έλεγχος υποθέσεων που θέλουμε να κάνουμε έχει την μορφή

$H_0 : F_x\left(x\right) = F_{N(\mu,\sigma^2)}(x)$ για κάθε $-\infty < x < \infty$

$H_1 : F_x\left(x\right) \ne F_{N(\mu,\sigma^2)}(x)$

Σε αυτή την περίπτωση όπου η μεταβλητή μας είναι συνεχής χωρίζουμε το δείγμα μας σε n υποδιαστήματα (συνήθως όχι λιγότερα από 5) έτσι ώστε να ισχύει $p_i^0 = P(\text{η X να ανήκει στο i υποδιάστημα}) = 1/n$. Στο παρακάτω διάγραμμα βλέπουμε πως αναπαριστάται η μεθοδολογία που περιγράφουμε γραφικά.




Στη συνέχεια ορίζουμε ως $x_i$ να είναι το i ποσοστιαίο σημείο της $N(\mu,\sigma^2)$ και έχουμε ότι $P(X \le x_i) = i$ και επίσης $P(X \le x_{1/n}) = \frac{i}{n}$ πράγμα που συνεπάγεται ότι

$$P\left(\frac{X-\mu}{\sigma} \le \frac{x_{\frac{i}{n}} - \mu}{\sigma}\right) = \frac{i}{n}$$ και τελικά ότι

$P\left(X \le \frac{x_{i/n} - \mu}{\sigma}\right) = \frac{i}{n}$, $i = 1, 2, ..., n-1$. Επίσης από την τελευταία αυτή έκφραση συμπεραίνουμε ότι ισχύει η ακόλουθη σχέση

$\frac{x_{\frac{i}{n}} - \mu}{\sigma} = z_{i/n}$ όπου $z_{i/n}$ ορίζουμε το $i/n$ ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής. Άρα το τελικό συμπέρασμα στο οποίο καταλήγουμε είναι η έκφραση που ακολουθεί παρακάτω

$$x_{\frac{i}{n}} = \mu + \sigma \cdot z_{\frac{i}{n}}$$

Στην περίπτωση που τα μ και σ είναι άγνωστα τότε η ποσότητα $x_{\frac{i}{n}} = \mu + \sigma \cdot z_{\frac{i}{n}}$ εκτιμάται από την $X^*_{i/n}$ σύμφωνα με τον ακόλουθο τύπο:

$$X^*_{i/n} = \begin{cases} \overline{X} + s^* \cdot z_{\frac{i}{n}}, \text{ αν } \dfrac{i}{n} \geq 0.5 \\ \overline{X} - s^* \cdot z_{\frac{(n-i)}{n}}, \text{ αν } \dfrac{i}{\mathrm{n}} < 0.5 \end{cases}$$ όπου $\overline{x}$ και $s^*$ είναι αμερόληπτες εκτιμήτριες των μ και $\sigma^2$. Με βάση αυτή την νέα μεθοδολογία έχουμε ότι οι κλάσεις είναι της μορφής $\left[ x^*_{\frac{(i-1)}{n}}, x^*_{\frac{i}{n}} \right), i = 1,2,...n+1,$ όπου $x^*_0 = -\infty,\ x^*_{\frac{n+1}{n}} = +\infty$. Συγκεκριμένα μπορούμε να φτιάξουμε τον ακόλουθο πίνακα όπου συνοψίζουμε τα παραπάνω

| κλάση | $O_i$ | $p_i^0$ | $E_i = n' \cdot p_i^0$ |
|---|---|---|---|
| $\left(-\infty, x^*_{1/n}\right)$ | $O_1$ | $p_1 = 1/n$ | $n'/n$ |
| $\left[ x^*_{\frac{1}{n}}, x^*_{\frac{2}{n}} \right)$ | $O_2$ | $p_2 = 1/n$ | $n'/n$ |
| $\left[ x^*_{\frac{2}{n}}, x^*_{\frac{3}{n}} \right)$ | $O_3$ | $p_3 = 1/n$ | $n'/n$ |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| $\left[ x^*_{\frac{n-1}{n}}, +\infty \right)$ | $O_n$ | $p_n = 1/n$ | $n'/n$ |

**Παράδειγμα**

Έστω ότι έχουμε 40 δεδομένα τα οποία διαχωρίζουμε σε 8 κλάσεις. Έχουμε δηλαδή ότι $n'$=40 και n=8. Έστω επίσης ότι από τα δεδομένα μας βρίσκουμε ότι $\overline{x} = 15.96$ και ότι $s^* = 2.144$. Τότε έχουμε ότι $p_i^0 = P\left(x \in i \text{ υποδιάστημα}/\mathrm{H}_0\right) = 1/n = 1/8 = 0.125$ και τελικά συμπεραίνουμε ότι $x^*_{1/8} = x^*_{0.125} = \overline{x} - s^* \cdot z_{\frac{(n-1)}{n}} = 15.96 - 2.144 \cdot z_{\frac{7}{8}} = 15.96 - 2.144 \cdot z_{0.875} = 13.49$.

## 4.2 Έλεγχος Kolmogorov

Ο έλεγχος Kolmogorov χρησιμοποιεί την αθροιστική συνάρτηση κατανομής $F(x)=P(X\le x)$ και την εμπειρική συνάρτηση κατανομής S(x).Τα δεδομένα για τον συγκεκριμένο έλεγχο θα πρέπει να είναι δεδομένα σε διατεταγμένη κλίμακα σε αντίθεση με τον έλεγχο $X^2$ ο οποίος είναι έλεγχος για δεδομένα ονομαστικής κλίμακας. Η λογική του ελέγχου αυτού είναι να συγκρίνει την αθροιστική συνάρτηση κατανομής με την εμπειρική συνάρτηση κατανομής. Υπενθυμίζουμε στο σημείο ότι ως εμπειρική συνάρτηση κατανομής ορίζουμε την $S(x)=\dfrac{\text{αριθμός τιμών δείγματος που είναι το πολύ ίσες με x}}{n}$. Στον έλεγχο Kolmogorov ως στατιστική συνάρτηση ελέγχου ορίζουμε την $T=\sup_x\left|F(x)-S(x)\right|$. Στον συγκεκριμένο έλεγχο οι τρεις δυνατές μορφές των ελέγχων υποθέσεων είναι οι ακόλουθες:

| Α. | Β. | Γ. |
|---|---|---|
| $H_0:F(x)=F_0(x)$ | $H_0:F(x)\ge F_0(x)$ | $H_0:F(x)\le F_0(x)$ |
| $H_1:F(x)\ne F_0(x)$ | $H_1:F(x)<F_0(x)$ | $H_1:F(x)=F_0(x)$ |

$$T=\sup_x\left|F_0(x)-S(x)\right| \qquad T^+=\sup_{x:F_0(x)\ge S(x)}\left[F_0(x)-S(x)\right] \qquad T^-=\sup_{x:S(x)>F_0(x)}\left[S(x)-F_0(x)\right]$$

Από τα παραπάνω γίνεται προφανές ότι και στις τρεις περιπτώσεις απορρίπτουμε την μηδενική υπόθεση για μεγάλες τιμές της στατιστικής συνάρτησης ελέγχου. Στον πίνακα 10 του παραρτήματος μπορούμε να δούμε τα (1-α) ποσοστιαία σημεία των κατανομών των παραπάνω στατιστικών συναρτήσεων T. Ο συγκεκριμένος πίνακας είναι ακριβής όταν η $F_0(x)$ είναι συνεχής κατανομή και επίσης όταν έχουμε $n\le 20$ στους αμφίπλευρους ελέγχους. Για τις περιπτώσεις εκείνες όπου το μέγεθος του δείγματος είναι μεγαλύτερο από 20, όπως και στις περιπτώσεις των μονόπλευρων ελέγχων, οι πίνακες αυτοί δίνουν ικανοποιητικές προσεγγίσεις οι οποίες επί τω πλείστων συμπίπτουν με τις ακριβείς τιμές. Τέλος, από τα προαναφερθέντα είναι εμφανές ότι ισχύει $T=\max(T^+,T^-)$. Ας δώσουμε στο σημείο αυτό κάποια παραδείγματα βασισμένα στον παραπάνω έλεγχο.

**Παράδειγμα 1**

Έστω ότι έχουμε τα δεδομένα 0.6, 0.8, 1.1, 1.2, 1.4, 1.7, 1.8, 1.9, 2.2, 2.4, 2.5, 2.9, 3.1, 3.4, 3.4, 3.9, 4.4, 4.9, 5.2, 5.9 και θέλουμε να ελέγξουμε την υπόθεση της μορφής $\begin{matrix}H_0:F_x(x)=F_{U(0.6)}(x)\\H_1:F_x(x)\ne F_{U(0.6)}(x)\end{matrix}$. Από τα παραπάνω δεδομένα και εφαρμόζοντας την γνωστή μεθοδολογία υπολογίζουμε την αθροιστική συνάρτηση κατανομής καθώς και την εμπειρική συνάρτηση κατανομής, οι οποίες παρουσιάζονται στους παρακάτω πίνακες

| $S(x)$ | Διάστημα |
|---|---|
| 0 | $-\infty < x < 0.6$ |
| 1/20 | $0.6 < x < 0.8$ |
| 2/20 | $0.8 < x < 1.1$ |
| 3/20 | $1.1 < x < 1.2$ |

$$F(x) = \begin{cases} 0, & -\infty < x < 0 \\ x/6, & 0 \le x < 6 \\ 1, & -6 \le x < +\infty \end{cases}$$

Επίσης με βάση την παραπάνω αθροιστική συνάρτηση κατανομής έχουμε ότι $f(x) = \begin{cases} 0, & x < 0 \\ 1/6, & 0 < x \le 6 \\ 0, & x > 6 \end{cases}$. Άρα λοιπόν, μπορούμε πολύ εύκολα εφαρμόζοντας τον ορισμό για την στατιστική συνάρτηση ελέγχου του ελέγχου Kolmogorov να υπολογίσουμε ότι T=0.18.

Έστω τώρα ότι μας ενδιαφέρει η περίπτωση Α του ελέγχου Kolmogorov και έχουμε ότι T=0.18 , n=20 και $w_{0.95}$=0.294. Με βάση την μεθοδολογία που ξέρουμε μπορούμε να υπολογίσουμε ότι

$$\hat{a} = P(T \ge 0.18 / H_0) = 1 - P(T < 0.18) > 1 - P(T < 0.232) = 1 - 0.80 = 0.20 \text{ άρα } \hat{a} > 0.20$$

Αν μας ενδιέφερε η περίπτωση Γ του ελέγχου Kolmogorov θα είχαμε ότι $T_0^-$ =0.18, n=20 και $w_{0.95}$=0.265 (σε επίπεδο στατιστικής σημαντικότητας α=0,05). Άρα λοιπόν θα υπολογίζαμε ότι

$$\hat{a} = P(T^- \ge 0.18 / H_0) = 1 - P(T^- < 0.18) > 1 - P(T^- < 0.232) = 1 - 0.90 = 0.10 \quad \hat{a} > 0.10$$

Άρα, συμπεραίνουμε ότι και στις δυο περιπτώσεις ελέγχου δεν έχουμε επαρκή στοιχεία από το δείγμα μας για να απορρίψουμε την μηδενική υπόθεση.

***Παρατηρήσεις για τους ελέγχους Kolmogorov και $X^2$***

- Ο έλεγχος καλής προσαρμογής του Kolmogorov καλύπτει μόνο τις περιπτώσεις όπου η συνάρτηση κατανομής που υποθέτουμε είναι εξ' ολοκλήρου ορισμένη, δηλαδή όταν δεν υπάρχουν άγνωστες παράμετροι που πρέπει να εκτιμηθούν από το δείγμα.
- Ο έλεγχος καλής προσαρμογής $X^2$ επιτρέπει την εκτίμηση ορισμένων παραμέτρων με βάση το δείγμα. Όμως ο έλεγχος $X^2$ απαιτεί την ομαδοποίηση των δεδομένων του δείγματος και μια τέτοια ομαδοποίηση είναι πολλές φορές αυθαίρετη. Τέλος κάποιες

φορές η ισχύς του ελέγχου δεν είναι και τόσο καλή επειδή η κατανομή της στατιστικής συνάρτησης είναι μόνο κατά προσέγγιση γνωστή.

- Κοιτάζοντας την βιβλιογραφία θα παρατηρήσουμε ότι κατά καιρούς έχουν υπάρξει αρκετές παραλλαγές του έλεγχου Kolmogorov για την χρήση του σε περιπτώσεις όπου οι παράμετροι εκτιμώνται από τα δεδομένα. Μια τέτοια παραλλαγή είναι αυτή του ελέγχου της σύνθετης υπόθεσης της κανονικότητας, δηλαδή του ελέγχου ότι ο υπό εξέταση πληθυσμός ανήκει στην οικογένεια των κανονικών κατανομών.

### 4.3.1 Έλεγχος κανονικότητας του Lillefors

Έστω ότι έχουμε τις παρατηρήσεις $x_1, x_2 \ldots x_n$ από κάποιο άγνωστο πληθυσμό με άγνωστη συνάρτηση κατανομής F(x). Ενδιαφερόμαστε να ελέγξουμε την υπόθεση

$H_0$: τα δεδομένα μας προέρχονται από την κανονική κατανομή με άγνωστη μέση τιμή μ και άγνωστη διακύμανση $\sigma^2$

$H_1$: τα δεδομένα μας προέρχονται από μια μη κανονική κατανομή

Η συγκεκριμένη υπόθεση μπορεί να ελεγχθεί με την χρήση της αμφίπλευρης ελεγχοσυνάρτησης του Kolmogorov η οποία είναι η μέγιστη κατακόρυφη απόσταση μεταξύ της εμπειρικής συνάρτησης κατανομής των $x_i$ και της συνάρτησης κατανομής της κανονικής κατανομής με μέση τιμή ίση με τον μέσο του δείγματος και τυπική απόκλιση ίση με την αμερόληπτη εκτίμηση της μέσω του δείγματος. Σην περίπτωση αυτή θα έχουμε δηλαδή ότι

$\bar{x} = \frac{\sum x_i}{n}, s^* = \sqrt{\frac{1}{n-1} \cdot \sum \left(x_i - \bar{x}\right)^2}$ και τέλος $T = \sup_x \left| F(x) - S(x) \right|$, $F \downarrow N\left(\bar{x}, s^{*2}\right)$. Ισοδύναμα μπορούμε να υπολογίσουμε τις τυποποιημένες τιμές $Z_1,\ldots,Z_n$ του αρχικού μας δείγματος όπου $Z_i = \frac{x_i - \bar{x}}{s^*}$ και να ορίσουμε τις εξής υποθέσεις

$H_0$: τα δεδομένα μας $Z_1,\ldots,Z_n$ προέρχονται από την τυποποιημένη κανονική κατανομή

$H_1$: τα δεδομένα μας $Z_1,\ldots,Z_n$ δεν προέρχονται από την τυποποιημένη κανονική κατανομή

Ως στατιστική συνάρτηση ελέγχου σε αυτή την περίπτωση θα χρησιμοποιήσουμε την μέγιστη κατακόρυφη απόκλιση της εμπειρικής συνάρτησης κατανομής $S_z^*$ του τυποποιημένου δείγματος από την συνάρτηση κατανομής $F_0^*(z)$ της τυποποιημένης κανονικής κατανομής. Δηλαδή σε αυτή την περίπτωση ως ελεγχοσυνάρτηση θα έχουμε την $\mathrm{T}_1 = \sup_z \left| F_0^*(z) - S^*(z) \right|$. Όπως και σε προηγούμενου ελέγχους έτσι και εδώ, οι μεγάλες τιμές της στατιστικής συνάρτησης ελέγχου T είναι ένδειξη για απόρριψη της μηδενικής υπόθεσης.

### 4.3.2 Έλεγχος Lilliefors για την εκθετική κατανομή

Χρησιμοποιούμε τον συγκεκριμένο έλεγχο για τον έλεγχο της υπόθεσης ότι ο αρχικός μας πληθυσμός προέρχεται από την εκθετική κατανομή με συνάρτηση κατανομής $F(x) = 1 - e^{-x/\mu}$ , $x > 0$ .Ουσιαστικά ο έλεγχος αυτός χρησιμοποιείται για την περιγραφή της κατανομής του χρόνου μεταξύ δυο διαδοχικών γεγονότων όταν τα γεγονότα αυτά συμβαίνουν τυχαία στο χρόνο. Άρα θα χαρακτηρίζαμε τον έλεγχο αυτό ως έλεγχο τυχαιότητας. Οι υποθέσεις που μας ενδιαφέρουν στον συγκεκριμένο έλεγχο είναι της μορφής

$\mathrm{H}_0$: $F_x(x) = \begin{cases} 1 - e^{-x/\mu} & , x > 0 \\ 0 & \text{διαφορετικά} \end{cases}$

$\mathrm{H}_1$: η κατανομή της X δεν είναι εκθετική

Αρχικά για να υπολογίσουμε την στατιστική συνάρτηση ελέγχου θα αναφέρουμε ότι πρέπει να μετασχηματίσουμε τα δεδομένα μας με τον μετασχηματισμό $z_i = \frac{x_i}{\bar{x}}$ όπου $\bar{x} = \frac{\sum x_i}{n}$ .

Έτσι η στατιστική συνάρτηση ελέγχου θα είναι η $T_2 = \sup_z \cdot \left| F^*(z) - S^*(z) \right|$ όπου $F^*(z) = 1 - e^{-z}$ , $\mathrm{z} > 0$ θα είναι η συνάρτηση κατανομής των μετασχηματισμένων δεδομένων και $S_z^*$ θα είναι η εμπειρική συνάρτηση κατανομής των μετασχηματισμένων δεδομένων.

**Κεφάλαιο 8**

# ΑΠΑΡΙΘΜΗΣΗ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ

## 8.1 Πίνακες συνάφειας

Όπως γίνεται αντιληπτό από τον τίτλο του κεφαλαίου, όταν μιλάμε για πίνακα συνάφειας μιλάμε για έναν rxc πίνακα στον οποίο είναι ταξινομημένα δεδομένα σε r γραμμές και c στήλες. Αναφερόμαστε δηλαδή σε έναν πίνακα στον οποίο παρουσιάζονται δεδομένα που περιέχονται σε r δείγματα (γραμμές) και c στήλες (κατηγορίες) και ελέγχουμε αν οι πιθανότητες ένα τυχαία επιλεγμένο αντικείμενο να ανήκει στις κατηγορίες 1,2,3,…,c (στήλες) διαφέρουν ή όχι από δείγμα σε δείγμα.(από γραμμή σε γραμμή). Όταν όμως έχουμε ένα και μοναδικό δείγμα το οποίο μελετάμε τότε το κάθε στοιχείο του δείγματος αυτού μπορεί να ταξινομηθεί σε μια από τις r διαφορετικές κατηγορίες σύμφωνα με ένα χαρακτηριστικό και σε μια από τις c διαφορετικές κατηγορίες σύμφωνα με ένα άλλο χαρακτηριστικό. Στην περίπτωση αυτή μας ενδιαφέρει ο έλεγχος ότι οι κατηγορίες του ενός κριτηρίου δεν επηρεάζουν σημαντικά τις αναλογίες των αντικειμένων σε κάθε μια από τις κατηγορίες του άλλου χαρακτηριστικού.

### 8.1.1 $X^2$ έλεγχος για ύπαρξη διαφορών σε πιθανότητες ή στις αναλογίες εκπροσώπησης r πληθυσμών σε c κατηγορίες

Έστω ότι έχουμε r αμοιβαία ανεξάρτητα τυχαία δείγματα μεγέθους $n_1,n_2, \ldots,n_r$ των οποίων τα στοιχεία μπορούν να ταξινομηθούν σε c κατηγορίες σύμφωνα με τον ακόλουθο πίνακα

| | Κατηγορία | | | | | |
|---|---|---|---|---|---|---|
| Δείγμα | | 1 | 2 | … | c | |
| | 1 | $O_{11}$ | $O_{12}$ | … | $O_{1c}$ | $n_1$ |
| | 2 | $O_{21}$ | $O_{22}$ | … | $O_{2c}$ | $n_2$ |
| | … | … | … | … | … | … |
| | r | $O_{r1}$ | $O_{r2}$ | … | $O_{rc}$ | $n_r$ |
| | | $C_1$ | $C_2$ | … | $C_c$ | $N$ |

Έστω $O_{ij}$ να είναι ο αριθμός των παρατηρήσεων που προέρχονται από το i δείγμα και ανήκουν στην κατηγορία j. Τότε ισχύει ότι $n_i = O_{i1} + O_{i2} + \ldots + O_{ic}$ , $i = 1, 2, \ldots r$ . Επίσης ο αριθμός των παρατηρήσεων που ανήκουν στην j κατηγορία συμβολίζεται με $C_j$ και ισχύει ότι $C_i = O_{1j} + O_{2j} + \ldots + O_{rj}$ , $j = 1, 2, \ldots, c$ . Τέλος, ο συνολικός αριθμός των παρατηρήσεων

από όλα τα δείγματα συμβολίζεται με N και ισχύει ότι $N = n_1 + n_2 + n_3 + ... + n_r$. Οι διάφορες υποθέσεις που μας ενδιαφέρει να ελέγξουμε είναι της μορφής:

$H_0$: οι πληθυσμοί, από όταν προέλθουν τα τυχαία δείγματα, εκπροσωπούνται σε ίσες αναλογίες στις διάφορες κατηγορίες.
$H_1$: τουλάχιστον 2 από τους πληθυσμούς, εκπροσωπούνται με διαφορετικές αναλογίες στις διάφορες κατηγορίες.

Αν στο σημείο αυτό ορίσουμε ως $p_{ij}$: πιθανότητα μια τιμή από τον i πληθ. να ανήκει στην j κατηγορία, τότε ο παραπάνω έλεγχος υποθέσεων παίρνει την μορφή:

$H_0 : p_{1j} = p_{2j} = ... = p_{rj}$ , $j = 1, 2, ...c$

$H_1 : p_{ij} \neq p_{kj}$ για κάποια τιμή της κατηγορίας j και για κάποιο ζεύγος i,k.

Αν η $H_0$ είναι αληθής τότε έχουμε ότι ο αριθμός των στοιχείων που περιμένουμε να παρατηρήσουμε στο κελί (i,j) είναι ίσος με

$E_{ij} = (\text{μέγεθος i δείγματος}) \cdot (\text{ποσοστό συνολικών παρατηρήσεων που ανήκουν στην j κατηγορία}) = \frac{n_i \cdot C_j}{N}$

.

Ως ελεγχοσυνάρτηση στον συγκεκριμένο έλεγχο ορίζουμε την $T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ η οποία ακολουθεί την κατανομή $X^2_{(r-1)(c-1)}$. Άρα λοιπόν απορρίπτουμε την $H_0$ αν ισχύει $T > X^2_{(r-1)(c-1),1-a}$. Μια ισοδύναμη έκφραση της στατιστικής συνάρτησης T είναι η $T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} - N$. Στο σημείο αυτό να αναφέρουμε ότι επειδή για τον έλεγχο υποθέσεων χρησιμοποιείται η ασυμπτωτική κατανομή της T , η θεωρούμενη τιμή του επιπέδου στατιστικής σημαντικότητας αποτελεί μια καλή προσέγγιση της πραγματικής τιμής του επιπέδου στατιστικής σημαντικότητας αν οι αναμενόμενες συχνότητες $E_{ij}$ έχουν αρκετά μεγάλες τιμές. Ορίζουμε έτσι τον περιορισμό ότι θα πρέπει να ισχύει $E_{ij} > 5$ για να εφαρμόσουμε τον συγκεκριμένο έλεγχο υποθέσεων με την συγκεκριμένη στατιστική συνάρτηση ελέγχου.

### 8.1.2 $X^2$ έλεγχος ανεξαρτησίας

Στην περίπτωση αυτού του ελέγχου έχουμε ένα τυχαίο δείγμα μεγέθους N του οποίου κάθε παρατήρηση μπορεί να ταξινομηθεί σύμφωνα με δυο χαρακτηριστικά. Δηλαδή υπάρχουν r κατηγορίες (γραμμές) ως προς το ένα χαρακτηριστικό και c κατηγορίες (στήλες) ως προς το δεύτερο χαρακτηριστικό. Ο πίνακας συνάφειας αυτού του ελέγχου έχει την μορφή

| Χαρακτηριστικό B | | | | | | |
|---|---|---|---|---|---|---|
| Χαρακτηριστικό A | | 1 | 2 | … | c | |
| | 1 | $O_{11}$ | … | | | $R_1$ |
| | 2 | … | … | | | $R_2$ |
| | 3 | | | | | |
| | 4 | | | | | |
| | … | | | | | |
| | r | | | | | $R_r$ |
| | | $C_1$ | $C_2$ | | $C_c$ | $N$ |

Σε αυτή την μορφή του ελέγχου η μηδενική υπόθεση $H_0$ παίρνει την μορφή:

$H_0$: το ενδεχόμενο μια παρατήρηση να ανήκει στην i γραμμή είναι ανεξάρτητο από το ενδεχόμενο η ίδια παρατήρηση να ανήκει στην j στήλη

Εναλλακτικά μπορούμε να διατυπώσουμε τις υποθέσεις του συγκεκριμένου ελέγχου ως ακολούθως:

$H_0 : p_{ij} = p_i \cdot p_j$ για $i = 1,...,r$ , $j = 1,...c$

$H_1 : p_{ij} \neq p_i \cdot p_j$ για κάποια i,j

Στην περίπτωση που η $H_0$ είναι αληθής, η αναμενόμενη συχνότητα του (i,j) κελιού είναι ίση με

$$E_{ij} = (\text{μέγεθος τυχαίου δείγματος}) \cdot (\text{ποσοστό παρατηρήσεων i γραμμής}) \cdot$$
$$(\text{ποσοστό παρατηρήσεων j στήλης}) =$$
$$N \cdot \frac{R_i}{N} \cdot \frac{C_j}{N} = \frac{R_i \cdot C_j}{N}$$

Ως στατιστική συνάρτηση ελέγχου στον παραπάνω έλεγχο χρησιμοποιούμε την

ελεγχοσυνάρτηση $T^* = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} - N$ .