Μη-Παραμετρική Στατιστική

Φροντιστήριο #7

Μέτρα συσχέτισης τάξης μεγέθους (Rank Correlation)

Εισαγωγή Μέτρο συσχέτισης είναι μία τ.μ. που χρησιμοποιείται όταν τα δεδομένα αποτελούνται από ζεύγη τιμών δηλαδή όταν έχουμε ένα διμεταβλητό τ.δ. μεγέθους n

$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$.

Υποθέτουμε ότι οι παρατηρήσεις του δείγματος (X_i, Y_i), $i = 1, 2, \dots, n$ είναι ισόνομες και η κοινή διμεταβλητή τους κατανομή είναι ίδια μ' αυτήν του τυχαίου διανύσματος (X, Y).

π.χ. $X_i :=$ ύψος του i-οστού ατόμου ενός δείγματος

$Y_i :=$ ύψος του πατέρα του i-οστού ατόμου

Οι τ.μ. X και Y μπορεί να είναι ανεξάρτητες, όπως π.χ.

$X_i :=$ μέση βαθμολογία ενός παίκτη ποδοσφαίρου

$Y_i :=$ βαθμολογία της φίλης του σε ένα μάθημα.

Προϋποθέσεις για τον ορισμό ενός μέτρου συσχέτισης:

(1) Η τιμή του θα πρέπει να είναι πάντα μεταξύ -1 και $+1$.

(2) Αν οι μεγαλύτερες τιμές της μεταβλητής X τείνουν να αντιστοιχούν στις μεγαλύτερες τιμές της μεταβλητής Y και άρα, οι μικρότερες τιμές της μεταβλητής X τείνουν να αντιστοιχούν στις μικρότερες τιμές της μεταβλητής Y, τότε το μέτρο συσχέτισης πρέπει

να είναι θετικό και να πλησιάζει την τιμή +1, αν η τάση είναι ισχυρή. Τότε, έχουμε θετική συσχέτιση μεταξύ των μεταβλητών X και Y.

③ Αν οι μεγαλύτερες τιμές της μεταβλητής X τείνουν να αντιστοιχούν στις μικρότερες τιμές της μεταβλητής Y και αντίστροφα, τότε το μέτρο συσχέτισης θα έχει μία αρνητική τιμή, κοντά στην τιμή −1, αν η τάση είναι ισχυρή. Τότε, έχουμε αρνητική συσχέτιση μεταξύ των μεταβλητών X και Y.

④ Αν οι τιμές της τ.μ. X φαίνονται να αντιστοιχούν με τυχαίο τρόπο σε τιμές της τ.μ. Y, το μέτρο συσχέτισης θα πρέπει να έχει μία τιμή κοντά στο 0. Τότε οι X και Y είναι ανεξάρτητες ή έχουν κάποιου άλλου είδους εξάρτηση (π.χ. καρπυλόγραμμη). Σε τις περιπτώσεις αυτές οι τ.μ. X και Y είναι ασυσχέτιστες ή δεν συσχετίζονται ή έχουν συσχέτιση 0.

Το πιο συχνά χρησιμοποιούμενο μέτρο συσχέτισης είναι ο συντελεστής συσχέτισης του Pearson, r. Ορίζεται ως εξής:

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\left[\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2\right]^{1/2}}, \text{ όπου } \bar{X}=\frac{\sum_{i=1}^{n} X_i}{n}$$

είναι ο μέσος των τιμών $X_1, X_2, \ldots, X_n$ και $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$ -3-
είναι ο μέσος των τιμών $Y_1, Y_2, \ldots, Y_n$.

Μία έκφραση του r που προσφέρεται περισσότερο για ταχύτερους υπολογισμούς είναι η ακόλουθη:

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}\right]^{1/2}\left[\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right]^{1/2}}$$

Μπορούμε να γράψουμε:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right]^{1/2}},$$

δηλαδή, προκύπτει ο λόγος της συνδιασποράς του δείγματος προς το γινόμενο των τυπικών αποκλίσεων των δύο δειγμάτων. Ο συντελεστής r ικανοποιεί τις προϋποθέσεις ① - ④.

Η κατανομή του r εξαρτάται από τη διπεταβλητή κατανομή του τυχαίου διανύσματος (X, Y). Συνεπώς, ο r δεν προσφέρεται ως σ.σ. ελέγχου για μη-παραμετρικούς ελέγχους, αν η κατανομή του διανύσματος (X, Y) είναι γνωστή.

Απαραίτητη προϋπόθεση για τη διενέργεια του παραμετρικού ελέγχου με σ.σ. τον συντελεστή συσχέτισης ρ είναι οι μεταβλητές X και Y να ακολουθούν τη διμεταβλητή κανονική κατανομή.

Διάφορα μέτρα συσχέτισης έχουν κατά καιρούς προταθεί τα οποία δεν εξαρτώνται από την κατανομή του διανύσματος (X,Y) αν οι μεταβλητές X και Y είναι ανεξάρτητες και μπορούν να χρησιμοποιηθούν ως ελεγχοσυναρτήσεις σε μη-παραμετρικούς ελέγχους ανεξαρτησίας. Τα μέτρα αυτά εξαρτώνται μόνο από τις τάξεις μεγέθους των παρατηρήσεων των δειγμάτων και έχουν κατανομές που είναι ανεξάρτητες από την κατανομή του διανύσματος (X,Y) αν οι μεταβλητές X και Y είναι ανεξάρτητες και συνεχείς.

Συντελεστής Συσχέτισης του Spearman

Δεδομένα $(X_1,Y_1),(X_2,Y_2),\dots,(X_n,Y_n)$ ένα δείγμα n παρατηρήσεων πάνω στο τυχαίο διάνυσμα (X,Y).

$R(X_i)$: βαθμός ή τάξη μεγέθους της X, $i=1,\dots,n$

$R(Y_i)$: βαθμός ή τάξη μεγέθους της Y, $i=1,2,\dots,n$

Τα δεδομένα αποτελούνται από μη-αριθμητικές παρατηρήσεις που εμφανίζονται σε ζεύγη αν οι παρατηρήσεις αυτές μπορούν να διαταχθούν κατά αύξουσα σειρά μεγέθους. Η διάταξη μπορεί να βασίζεται στην ποιότητα των παρατηρήσεων π.χ. από τη χειρότερη

στην καλύτερη παρατήρηση ή στον βαθμό προτί-
μησης που μπορεί να αντιστοιχηθεί στις παρατηρήσεις.

Ισοπαλίες (Ties). Αν δύο ή περισσότερες τιμές ταυτίζονται αντιστοιχίζουμε σε κάθε μία από τις ίσες αυτές τιμές τον μέσο των βαθμών που θα είχαν αν δεν ταυτίζονταν.

Μέτρο συσχέτισης Spearman. Προτάθηκε το 1904.
Δεν είναι άλλο από τον συντελεστή r του Pearson υπολογιζόμενο όμως με βάση τις τάξεις μεγέθους των παρατηρήσεων και όχι αυτές καθεαυτές τις παρατηρήσεις.

$$\rho = \frac{\sum_{i=1}^{n} [R(X_i) - \overline{R(X)}][R(Y_i) - \overline{R(Y)}]}{\left(\sum_{i=1}^{n} [R(X_i) - \overline{R(X)}]^2 \right)^{1/2} \left(\sum_{i=1}^{n} [R(Y_i) - \overline{R(Y)}]^2 \right)^{1/2}}$$

όπου $\overline{R(X)} = \frac{\sum_{i=1}^{n} R(X_i)}{n}$ και $\overline{R(Y)} = \frac{\sum_{i=1}^{n} R(Y_i)}{n}$.

Αν δεν υπάρχουν περιπτώσεις ίσων τιμών π.χ. στις X τιμές, τότε: $\overline{R(X)} = \frac{1}{n} \sum_{i=1}^{n} R(X_i) = \frac{1}{n} \sum_{i=1}^{n} i$

$$= \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}.$$

Τότε:

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

$$= \sum_{i=1}^{n} \left[i^2 + \left(\frac{n+1}{2}\right)^2 - 2\, i\, \frac{(n+1)}{2} \right]$$

$$= \frac{n(n+1)(2n+1)}{6} + \frac{n(n+1)^2}{4} - \frac{(n+1)^2 n}{2}$$

$$= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4}$$

$$= \frac{n(n+1)}{2}\left(\frac{2n+1}{3} - \frac{n+1}{2} \right) = \frac{n(n+1)}{12}(n-1)$$

$$= \frac{n(n^2-1)}{12}.$$

Άρα, αν όλες οι παρατηρήσεις είναι διαφορετικές, ο συντελεστής ρ του Spearman μπορεί να γραφεί:

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

Μία ισοδύναμη μορφή για τον ρ που προσφέρεται για ταχύτερους υπολογισμούς δίνεται παρακάτω:

$$\rho = 1 - \frac{6T}{n(n^2-1)}, \text{ όπου } T = \sum_{i=1}^{n} [R(X_i) - R(Y_i)]^2.$$

Αν οι X τιμές (ή οι Y τιμές) δεν είναι όλες διαφορετικές, δηλαδή υπάρχουν περιπτώσεις ίσων τιμών, τότε χρησιμοποιείται η εξής έκφραση:

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

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

Παρατήρηση: Ο συντελεστής ρ του Spearman, χρησιμοποιείται συχνά ως σ.σ. ελέγχου για τον έλεγχο της ανεξαρτησίας μεταξύ δύο τ.μ. Ο ρ δεν είναι εναίσθητος σε ορισμένες μορφές εξάρτησης. Ο ερευνητής μαζί είναι να προσδιορίζει τη μορφή της εξάρτησης των μεταβλητών που επιθυμεί να ελέγξει.

Ενδιαφερόμαστε να ελέγξουμε τις ακόλουθες υποθέσεις: (Α. (Αμφίπλευρος Έλεγχος)).

Α. H_0: οι X και Y είναι αμοιβαία ανεξάρτητες.

H_1: είτε υπάρχει τάση οι μεγαλύτερες τιμές της μεταβλητής X να αντιστοιχούν στις μεγαλύτερες τιμές της μεταβλητής Y, είτε υπάρχει τάση στις μικρότερες τιμές της μεταβλητής X να αντιστοιχούν στις μεγαλύτερες τιμές της Y.

Β. (Μονόπλευρος έλεγχος για θετική συσχέτιση).

H_0: οι μεταβλητές X και Y είναι αμοιβαία ανεξάρτητες.

H_1: υπάρχει τάση οι μεγαλύτερες τιμές της X να αντιστοιχούν στις μεγαλύτερες τιμές της Y και αντίστροφα.

Γ. (Μονόπλευρος έλεγχος για αρνητική συσχέτιση).

H_0: οι X και Y είναι αμοιβαία ανεξάρτητες.

H_1: υπάρχει τάση οι μικρότερες τιμές της X να αντιστοιχούν στις μεγαλύτερες τιμές της Y και αντίστροφα.

Οι εναλλακτικές υποθέσεις διατυπώνουν την ύπαρξη συσχέτισης μεταξύ X και Y. Μία μηδενική υπόθεση της μορφής "μη ύπαρξη συσχέτισης μεταξύ X και Y" θα ήταν περισσότερο ακριβής από την υπόθεση H_0 όπως δόθηκε παραπάνω χρησιμοποιείται περισσότερο και είναι ευκολότερο να ερμηνευθεί.

"ύπαρξη ανεξαρτησίας μεταξύ X και Y". Όμως η

Ο συντελεστής ρ του Spearman χρησιμοποιείται ως σ.σ. ελέγχου για τις παραπάνω υποθέσεις.

Η ακριβής κατανομή του ρ κάτω από την H_0 της ανεξαρτησίας των X και Y δίνεται από κατάλληλους πίνακες τιμών (π.χ. βλέπε βιβλίο Ξεναξάκη, Πίνακας 10, Παράρτημα). Για $n > 30$, χρησιμοποιείται κατάλληλη κανονική προσέγγιση.

Συχνά, για τον έλεγχο των υποθέσεων Α, Β και Γ, αντί -9- να χρησιμοποιηθεί ο συντελεστής ρ του Spearman, χρησιμοποιείται η σ.σ.

$$T = \sum_{i=1}^{n} [R(X_i) - R(Y_i)]^2.$$

Οποτεδήποτε, όμως υπάρχουν αριστές περιπτώσεις ταύτισης τιμών, θα πρέπει να χρησιμοποιείται ο συντελεστής ρ.

Ο έλεγχος που στηρίζεται στην T, είναι γνωστός ως έλεγχος των Hotelling and Pabst. Τα ποσοστιαία σημεία της αριβούς κατανομής της σ.σ. T δίνονται από κατάλληλους πίνακες τιμών (π.χ. βλέπε Πίνακα 11, βιβλίο Ξεμαλάμη). Κατάλληλη κανονική προσέγγιση έχουμε για $n > 30$.

Έτσι, για τον έλεγχο με τη σ.σ. ρ, η κρίσιμη περιοχή:

Β. H_0 σε επ. α αν $\rho > w_{1-\alpha}$ (Ισχύει: $w_p = -w_{1-p}$)

Γ. H_0 σε επ. α αν $\rho < w_\alpha$

Α. H_0 σε επ. α αν $\rho > w_{1-\alpha/2}$ ή $\rho < w_{\alpha/2}$

Για $n > 30$, ισχύει $w_p = \frac{z_p}{\sqrt{n-1}}$, όπου w_p το p-ποσοστιαίο σημείο της W και z_p το p-ποσοστιαίο σημείο της $N(0,1)$.

Για τον έλεγχο με τη σ.σ. T, η κρίσιμη περιοχή:

Β. H_0 σε ε.σ. α αν $T < w_\alpha$

Γ. H_0 σε ε.σ. α αν $T > w_{1-\alpha}$

Α. H_0 σε επ. α αν $T < w_{\alpha/2}$ ή $T > w_{1-\alpha/2}$.

Για $n > 30$, $w_p \cong \dfrac{\left[n(n^2-1) + 2p\dfrac{n(n^2-1)}{\sqrt{n-1}}\right]}{6}$

και $w_{1-p} = \dfrac{n(n^2-1)}{3} - w_p$.

Η τιμή της σ.σ. T είναι μεγάλη, όταν η τιμή της σ.σ. p είναι μικρή και αντιστρόφως.

Ας δούμε ένα παράδειγμα.

Παράδειγμα: 12 ζεύγη διδύμων υποβλήθηκαν σε ένα ψυχολογικό τεστ για να μετρηθεί η επιθετικότητά τους. Η έρευνα ήταν στην εξέταση του βαθμού ομοιότητας μεταξύ των διδύμων του ίδιου ζεύγους. Τα δεδομένα παριστάνουν μετρήσεις της επιθετικότητας και αναφέρονται στον πίνακα:

Ζεύγος διδύμων i	1	2	3	4	5	6	7	8	9	10	11	12
Πρωτότοκος X_i	86	71	77	68	91	72	77	91	70	71	88	87
Δευτερότοκος Y_i	88	77	76	64	96	72	65	90	65	80	81	72

Οι πρωτότοκοι όλων των ζευγαριών διδύμων διαταχθηκαν ως προς την επιθετικότητά τους κατά αύξουσα τάξη μεγέθους, όπως και οι δευτερότοκοι των ζευγαριών με τα εξής αποτελέσματα:

Ζεύγος διδύμων i	1	2	3	4	5	6	7	8
$R(X_i)$	8	3.5	6.5	1	11.5	5	6.5	11.5
$R(Y_i)$	10	7	6	1	12	4.5	2.5	11
$[R(X_i) - R(Y_i)]^2$	4	12.25	0.25	0	0.25	0.25	16	0.25

i	9	10	11	12
$R(X_i)$	2	3.5	10	9
$R(Y_i)$	2.5	8	9	4.5
$[R(X_i)-R(Y_i)]^2$	0.25	20.25	1	20.25

Η τιμή της σ.σ. T είναι

$$\tau = \sum_{i=1}^{12} [R(X_i)-R(Y_i)]^2 = 75.$$

Άρα, ο συντελεστής συσχέτισης ρ του Spearman είναι:

$$\rho = 1 - \frac{6\tau}{n(n^2-1)} = 1 - \frac{6\cdot 75}{12(144-1)}$$

$$= 0.7378.$$

Αν επιθυμούμε να ελέγξουμε τις υποθέσεις της περίπτωσης A σε ε.σ.σ. $\alpha = 5\%$, τότε θα πρέπει να συγκρίνουμε την παρατηρηθείσα τιμή του συντελεστή ρ με τις τιμές των 0.025 και 0.975 ποσοστιαίων σημείων του Πίνακα 10 του Παραρτήματος (βιβλίο Σπυράκη).

Από τον πίνακα προκύπτει ότι:

$$W_{0.975} = 0.5804 \text{ και } W_{0.025} = -W_{0.975} = -0.5804$$

Επειδή $\rho = 0,7378 > W_{0.975}$ η H_0 απορρίπτεται σε ε.σ.σ. $\alpha = 5\%$.

Αν κάναμε τον έλεγχο με τη σ.σ. T από τον πίνακα 11 του Παραρτήματος, για $n=12$, $W_{0.025}=120$,

$W_{0.975} = \frac{1}{3} n(n^2-1) - W_{0.025} = 452$. Η τιμή $\tau = 75$

ανήκει στην κρίσιμη περιοχή με επίπεδο $\alpha = 0.05$ (αφού είναι μικρότερη του $w_{0.025} = 120$) άρα η H_0 απορρίπτεται σε επ $\alpha = 5\%$ και συνεπώς φαίνεται να υπάρχει στατιστικά σημαντική θετική συσχέτιση στην επιθετικότητα ανάμεσα στα δίδυμα αδέλφια.

Υλοποίηση με την R:

Με τη συνάρτηση cor.test(x, y, method="pearson") ή cor.test(x, y, method="spearman") μπορούμε να λάβουμε τα αποτελέσματα για τα δύο tests.

Παραμετρικό: Pearson

Μη-παραμετρικό ανάλογο: Spearman.

Έλεγχος κανονικότητας των x, y υλοποιείται με το K-S test και γραφικές ενδείξεις για την κανονικότητα των x, y παρέχονται από τα qq-plots. Απαιτήται η εντολή library("ggpubr").

Οι τιμές των συντελεστών συσχέτισης δίνονται στα outputs για (a) Pearson: cor = 0.7344383

(b) Spearman: rho = 0.7354509.

-13-

```
#DATA
x=c(86, 71, 77, 68, 91, 72, 77, 91, 70, 71, 88, 87) #First born
y=c(88, 77, 76, 64, 96, 72, 65, 90, 65, 80, 81, 72) #Second born

plot(x,y) #Scatterplot of  x,y

#Kolmogorov-Smirnov test for normality
ks.test(x, "pnorm")
ks.test(y, "pnorm")

library("ggpubr")

#qqplots for x,y
ggqqplot(x, ylab = "aggression of firstborn")
ggqqplot(y, ylab = "aggression of secondborn")

#Pearson correlation test

res1 <- cor.test(x, y, method = "pearson")
res1

#Spearman correlation test
res00 <- cor(x,y, method= "spearman")
res2 <-cor.test(x, y,  method = "spearman")
res2
```

```
One-sample Kolmogorov-Smirnov test

data:  x
D = 1, p-value = 7.55e-11
alternative hypothesis: two-sided

Warning message:
In ks.test(x, "pnorm") :
  ties should not be present for the Kolmogorov-Smirnov test
> ks.test(y, "pnorm")

        One-sample Kolmogorov-Smirnov test

data:  y
D = 1, p-value = 7.55e-11
alternative hypothesis: two-sided

Warning message:
In ks.test(y, "pnorm") :
  ties should not be present for the Kolmogorov-Smirnov test
>
> library("ggpubr")
Loading required package: ggplot2
>
> #qqplots for x,y
> ggqqplot(x, ylab = "aggression of firstborn")
> ggqqplot(y, ylab = "aggression of secondborn")
>
> #Pearson correlation test
>
> res1 <- cor.test(x, y, method = "pearson")
> res1

        Pearson's product-moment correlation

data:  x and y
t = 3.4221, df = 10, p-value = 0.006524
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2775029 0.9203969
sample estimates:
      cor
0.7344383

>
> #Spearman correlation test
> res00 <- cor(x,y, method= "spearman")
> res2 <-cor.test(x, y,  method = "spearman")
Warning message:
In cor.test.default(x, y, method = "spearman") :
  Cannot compute exact p-value with ties
> res2

        Spearman's rank correlation rho

data:  x and y
S = 75.661, p-value = 0.006413
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7354509

>
>
```

Παρατήρηση: Η τιμή του ρ είναι διαφορετική από αυτή που βγάλαμε στο παράδειγμα ($\rho = 0.7378$). Η R χρησιμοποιεί μία "διόρθωση" του ρ λαμβάνοντας υπόψη τις ισοβαθμίες. Το ίδιο θα συμβεί και με τον συντελεστή τ του Kendall (βλ. #8).