Μη-Παραμετρική Στατιστική

Φροντιστήριο #8

Rank correlation by Kendall

Εισαγωγή: Ο συντελεστής συσχέτισης τ του Kendall είναι γνωστός ως συντελεστής εναρμόνισης του Kendall. Μοιάζει με τον συντελεστή ρ του Spearman διότι υπολογίζεται με βάση τις τάξεις μεγέθους των παρατηρήσεων και η κατανομή του δεν εξαρτάται από την κατανομή των μεταβλητών X και Y, όταν αυτές είναι ανεξάρτητες και συνεχείς.

Η σ.σ. ελέγχου κατά Kendall πλεονεκτεί αυτής του Spearman διότι τείνει στην κανονική κατανομή σχετικά γρήγορα. Έτσι, η προσέγγιση της κατανομής του συντελεστή τ από την κανονική κατανομή είναι καλύτερη από την αντίστοιχη προσέγγιση της κατανομής του συντελεστή ρ του Spearman όταν ισχύει η μηδενική υπόθεση της ανεξαρτησίας των μεταβλητών X και Y. Επίσης, ο συντελεστής τ του Kendall μπορεί άμεσα και απλά να ερμηνευθεί μέσω των πιθανοτήτων με τις οποίες παρατηρούμε εναρμονισμένα ή συσχετισμένα ζεύγη τιμών (concordant) και μη-εναρμονισμένα ή μη-συσχετισμένα (discordant) ζεύγη τιμών.

Δεδομένα: Αποτελούνται από ένα διμεταβλητό τ.δ. μεγέθους n παρατηρήσεων (X_i, Y_i), $i=1,2,\dots,n$, πάνω στο τυχαίο διάνυσμα (X,Y).

<u>Ορισμός:</u> Δύο παρατηρήσεις, (X_j, Y_j) και (X_k, Y_k) καλούνται εναρμονισμένες ή συσχετισμένες (concordant) αν και τα δύο μέλη της μίας παρατήρησης είναι μεγαλύτερα (ή μικρότερα) από τα αντίστοιχα μέλη της άλλης παρατήρησης. Δηλαδή, αν $X_j > X_k$ (αντίστοιχα, $X_j < X_k$), τότε $Y_j > Y_k$ (αντίστοιχα, $Y_j < Y_k$). Οι παρατηρήσεις (X_j, Y_j) και (X_k, Y_k) θα καλούνται μη-εναρμονισμένες ή μη-συσχετισμένες (discordant), αν η διάταξη των πρώτων μελών τους είναι αντίθετη από τη διάταξη των δεύτερων μελών τους δηλαδή αν $X_j > X_k$ (αντίστοιχα, $X_j < X_k$), τότε $Y_j < Y_k$ (αντίστοιχα, $Y_j > Y_k$).

Ισοδύναμα, δύο ζεύγη παρατηρήσεων (X_j, Y_j) και (X_k, Y_k) θα ονομάζονται <u>εναρμονισμένα</u> αν οι διαφορές $X_j - X_k$ και $Y_j - Y_k$ έχουν το ίδιο πρόσημο δηλ. αν $(X_j - X_k)(Y_j - Y_k) > 0$. Τα ζεύγη (X_j, Y_j) και (X_k, Y_k) θα ονομάζονται <u>μη-εναρμονισμένα</u> αν οι διαφορές $X_j - X_k$ και $Y_j - Y_k$ έχουν αντίθετο πρόσημο δηλ. αν $(X_j - X_k)(Y_j - Y_k) < 0$.

Έστω N_c και N_d οι αριθμοί των εναρμονισμένων και των μη-εναρμονισμένων ζευγών παρατηρήσεων, αντίστοιχα.

Τα ζεύγη (X_j, Y_j) και (X_k, Y_k) για τα οποία $X_j = X_k$ ή/και $Y_j = Y_k$, δεν είναι ούτε εναρμονισμένα ούτε μη-εναρμονισμένα. Τα ζεύγη αυτά καλούνται ισοβαθμούντα (tied).

Έστω $N_o := \#$ ισοβαθμούντων ζευγών παρατηρήσεων

Οι n παρατηρήσεις μπορούν να συνδυασθούν ανά δύο με $\binom{n}{2} = \frac{n(n-1)}{2}$ διαφορετικούς τρόπους.

Άρα,
$$N_c + N_d + N_o = \binom{n}{2}.$$

Τα δεδομένα μπορούν να αποτελούνται από μη-αριθμητικές παρατηρήσεις που εμφανίζονται κατά ζεύγη, με την προϋπόθεση ότι οι παρατηρήσεις είναι τέτοιες ώστε να μπορούν να ορισθούν εναρμονισμένα και μη-εναρμονισμένα ζεύγη παρατηρήσεων και να είναι δυνατός ο υπολογισμός των N_c και N_d.

Ο Kendall (1938) πρότεινε το ακόλουθο μέτρο συσχέτισης:
$$\tau = \frac{N_c - N_d}{\binom{n}{2}} = \frac{N_c - N_d}{n(n-1)/2}$$

που αναπαριστά τη διαφορά μεταξύ των ποσοστών των εναρμονισμένων και μη-εναρμονισμένων ζευγών παρατηρήσεων.

Είναι $-1 \le \tau \le +1$.

Ο υπολογισμός του συντελεστή τ γίνεται απλός, αν οι παρατηρήσεις (X_i, Y_i), $i=1,\dots,n$, διαταχθούν σε μία στήλη κατά αύξουσα τάξη μεγέθους των τιμών των παρατηρήσεων πάνω στη μεταβλητή X. Τότε κάθε τιμή Y χρειάζεται να συγκριθεί μόνο με τις τιμές Y που είναι "κάτω" από αυτήν. Έτσι, κάθε ζεύγος παρατηρήσεων εξετάζεται μόνο μία φορά και ο αριθμός των συσχετισμένων και μη-συσχετισμένων ζευγών προσδιορίζεται ευκολότερα.

Ο συντελεστής τ μπορεί επίσης να χρησιμοποιηθεί ως σ.σ. ελέγχου για τον έλεγχο της H_0 που υποθέτει ανεξαρτησία μεταξύ των μεταβλητών X και Y, όπως συμβαίνει και με τον συντελεστή συσχέτισης ρ του Spearman. Συχνή είναι και η χρήση της διαφοράς $N_c - N_d$ ως σ.σ. ελέγχου, για τον έλεγχο αυτής της υπόθεσης. Δηλ. ως σ.σ. ελέγχου, χρησιμοποιείται συχνά η συνάρτηση: $T = N_c - N_d$ η οποία καλείται

<u>ελεγχοσυνάρτηση</u> <u>του Kendall.</u>

Τα ποσοστιαία σημεία της κατανομής της σ.σ. T δίνονται από κατάλληλους πίνακες τιμών (βλέπε π.χ. βιβλίο Ξεκαλάκη, Πίνακας 12, Παράρτημα).

Για τους ελέγχους Α, Β, Γ του Φρ. #7, όπου:

Α. (Αμφίπλευρος έλεγχος συσχέτισης)

Β. (Μονόπλευρος έλεγχος θετικής συσχέτισης)

Γ. (Μονόπλευρος έλεγχος αρνητικής συσχέτισης)

ο κανόνας απόφασης διαμορφώνεται ως εξής:

Α. $\cancel{H}_0$ σε ε.σ.σ. α αν $T > W_{1-\alpha/2}$ ή $T < W_{\alpha/2} = -W_{1-\alpha/2}$

Β. $\cancel{H}_0$ σε εσσ α αν $T > W_{1-\alpha}$

Γ. $\cancel{H}_0$ σε εσσ α αν $T < W_\alpha = -W_{1-\alpha}$.

Εν γένει, ο συντελεστής ρ του Spearman τείνει να είναι μεγαλύτερος κατά απόλυτη τιμή από τον συντελεστή τ του Kendall. Δεν υπάρχουν επαρκείς λόγοι για τους οποίους πρέπει ο ένας έλεγχος να προτιμάται έναντι του άλλου.

Γενικά σχόλια για συσχέτιση τάξης μεγέθους (rank correlation)

① Η ακριβής κατανομή των σ.σ. ρ και τ μπορεί να προσδιοριστεί, αν και πρακτικά η διαδικασία υπολογισμού είναι χρονοβόρα ακόμη και για μέτριο μέγεθος δείγματος n, κάτω από την υπόθεση ότι οι X και Y είναι ανεξάρτητες και ισόνομες.

② Στην περίπτωση μεγάλων δειγμάτων, επειδή οι σ.σ. ρ και τ είναι αθροίσματα τ.μ. μπορεί να χρησιμοποιηθεί το ΚΟΘ για να προσεγγίσει τις κατανομές τους. Και οι δύο κατανομές είναι συμμετρικές γύρω από την τιμή 0. Η προσέγγιση μέσω ΚΟΘ είναι καλύτερη στην περίπτωση της σ.σ. τ για $n \geq 8$. Δεν είναι όμως εξ ίσου ικανοποιητική όταν χρησιμοποιείται ~~στην~~ προσέγγιση της κατανομής της σ.σ. ρ.

(3) Αν τα ζεύγη (X_i, Y_i), $i=1,\ldots,n$ προέρχονται από τη διδιάστατη κανονική τρ. τότε ο παραμετρικός έλεγχος με τον συντελεστή συσχέτισης r του Pearson μπορεί να χρησιμοποιηθεί για την ανάλυση της συσχέτισης των μεταβλητών X και Y.

Θεωρούμε το παράδειγμα του Φρ. #7.

<u>Παράδειγμα</u> Θεωρούμε τα δεδομένα πάνω στην επιθετικότητα των διδύμων. Διατάσσοντας τις παρατηρήσεις (X_i, Y_i), $i=1,2,\ldots,n$ κατά αύξουσα τάξη μεγέθους των τιμών των παρατηρήσεων X_i, $i=1,\ldots,n$ καταλήγουμε στον ακόλουθο πίνακα, όπου:

$X^{(i)}$, $i=1,\ldots,n$ είναι η διατεταγμένη ακολουθία των παρατηρήσεων X_i,

Y_i^*, $i=1,\ldots,n$ η προκύπτουσα αναδιάταξη των αντίστοιχων σ' αυτές τιμών των Y_i.

Η δεύτερη στήλη του πίνακα, δίνει τον αριθμό των ζευγών $(X^{(i+1)}, Y^*_{i+1})$ για τα οποία $Y_i^* < Y_{i+1}^*$

$X^{(i)} < X^{(i+1)}$, $i=1,\dots,n-1$.

Η τρίτη στήλη δίνει τον αριθμό των ζευγών $(X^{(i+1)}, Y^*_{i+1})$ για τα οποία $Y^*_i > Y^*_{i+1}$ όταν

$X^{(i)} < X^{(i+1)}$, $i=1,2,\dots,n-1$.

	$(X^{(i)}, Y^*_i)$	Εναρμονισμένα ζεύγη κάτω από $(X^{(i)}, Y^*_i)$	Μη-Εναρμονισμένα ζεύγη κάτω από $(X^{(i)}, Y^*_i)$
	(68,64)	11	0
	(70,65)	9	0
tied	(71,77)	4	4
	(71,80)	4	4
	(72,72)	5	1
tied	(77,65)	5	0
	(77,76)	4	1
	(86,88)	2	2
	(87,72)	3	0
	(88,81)	2	0
tied	(91,90)	0	0
	(91,96)	0	0
Σύνολο		$N_c = 49$	$N_d = 12$

Με βάση τα στοιχεία του πίνακα, η τιμή του συντελεστή συσχέτισης τ του Kendall είναι:

$$\tau = \frac{N_c - N_d}{n(n-1)/2} = \frac{49-12}{(12)(11)/2} = 0.5606$$

Άρα, υπάρχει θετική συσχέτιση τάξης μεγέθους μεταξύ των μετρήσεων της επιθετικότητας των διδύμων, όπως προκύπτει από τον συντελεστή συσχέτισης τ του Kendall.

Έστω ότι ενδιαφερόμαστε να ελέγξουμε την

H_0: οι X και Y είναι ασυσχέτιστες

vs

H_1: οι X και Y είναι συσχετισμένες

(αμφίπλευρος έλεγχος).

Η σ.σ. T είναι: $T = N_c - N_d = 49 - 12 = 37$.

Για τον αμφίπλευρο έλεγχο μεγέθους $\alpha = 5\%$, για $n=12$, $W_{0.975} = 28$, $W_{0.025} = -W_{0.975} = -28$ (πίνακας 12). Άρα H_0 απορρίπτεται σε εσσ $\alpha = 5\%$. Όπως έχουμε ήδη αναφέρει προέκυψε ότι: $\rho > \tau$.

Υλοποίηση με την R

Πάλι με την εντολή cor.test βάζοντας ως method = "kendall."

Η τιμή του τ που βγάζει η R είναι διαφορετική διότι χρησιμοποιείται μία "διόρθωση" της τ που λαμβάνει υπόψη τις ισοβαθμίες. Την ίδια "διόρθωση" κάνουν (σχεδόν) όλα τα στατιστικά πακέτα στον υπολογισμό τους, (βγάζει $\tau = 0.5826$ και όχι $\tau = 0.5606$ δηλ. την τιμή που βγάλαμε στο παράδειγμα).

-10-

```
#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

#Spearman correlation test

res2 <-cor.test(x, y,  method = "kendall")
res2
```

```
 Kendall's rank correlation tau

data:  x and y
z = 2.567, p-value = 0.01026
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.5826952
```

(χρησιμοποιεί κανονική προσέγγιση, $n \geq 8$)