3.4.2 Ο Συντελεστής Συσχέτισης τ του Kendall

Ο συντελεστής συχέτισης τ του Kendall, γνωστός και ως συντελεστής *εναρμόνισης* του Kendall, μοιάζει με τον συντελεστή ρ του Spearman ως προς το ότι υπολογίζεται με βάση την τάξη μεγέθους των παρατηρήσεων και όχι με βάση τις παρατηρήσεις αυτές καθεαυτές και, επιπλέον, η κατανομή του δεν εξαρτάται από την κατανομή των μεταβλητών X και Y, όταν αυτές είναι ανεξάρτητες και συνεχείς. Το κύριο πλεονέκτημα του μέτρου αυτού σε σχέση με το μέτρο ρ του Spearman είναι ότι τείνει στην κανονική κατανομή σχετικά γρήγορα. Αποτέλεσμα αυτού είναι ότι 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 αληθεύει η μηδενική υπόθεση της ανεξαρτησίας μεταξύ των μεταβλητών X και Y. Ενα άλλο πλεονέκτημα του συντελεστή τ του Kendall βρίσκεται στο γεγονός ότι μπορεί άμεσα και απλά να ερμηνευθεί μέσω των πιθανοτήτων με τις οποίες παρατηρούμε *εναρμονισμένα ή συσχετισμένα (concordant)* ζεύγη τιμών και *μη εναρμονισμένα ή μη συσχετισμένα (discordant)* ζεύγη τιμών, όπως αυτά ορίζονται στην συνέχεια.

Τα δεδομένα αποτελούνται από ένα διμεταβλητό τυχαίο δείγμα μεγέθους n παρατηρήσεων (X_i, Y_i), $i = 1, 2, ..., n$, πάνω στο τυχαίο διάνυσμα (X, Y).

Ορισμός: Δύο παρατηρήσεις, έστω (X_j, Y_j) και (X_k, Y_k), ονομάζονται *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 ονομάζονται *μη εναρμονισμένες ή μη συσχετισμένες (discordant)*, αν η διάταξη των πρώτων μελών τους είναι αντίθετη από την 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 *εναρμονισμένα* αν οι διαφορές $X_j - X_k$ και $Y_j - Y_k$ έχουν το ίδο πρόσημο (αν $(X_j - X_k)(Y_j - Y_k) > 0$). Τα ζεύγη (X_j, Y_j) και (X_k, Y_k) θα ονομάζονται *μη εναρμονισμένα* αν οι διαφορές $X_j - X_k$ και $Y_j - Y_k$ έχουν αντίθετο πρόσημο (αν $(X_j - X_k)(Y_j - Y_k) < 0$).

Έστω N_c και N_d οι αριθμοί των εναρμονισμένων και μη εναρμονισμένων ζευγών παρατηρήσεων, αντίστοιχα. Τα ζεύγη των παρατηρήσεων (X_j, Y_j) και (X_k, Y_k), για τα οποία ισχύει ότι $X_j = X_k$ ή/και $Y_j = Y_k$, δεν είναι ούτε εναρμονισμένα ούτε μη εναρμονισμένα. Τα ζεύγη αυτά ονομάζονται *ισοβαθμούντα (tied)*.

Εστω N_0 ο αριθμός των ισοβαθμούντων ζευγών παρατηρήσεων. Επειδή οι n παρατηρήσεις μπορούν να συνδυασθούν ανά δύο με $\binom{n}{2} = n(n-1)/2$ διαφορετικούς τρόπους, έπεται ότι

$$N_c + N_d + N_0 = \binom{n}{2}$$

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

Το μέτρο συσχέτισης που προτάθηκε από τον Kendall το 1938 ορίζεται ως εξής:

$$\tau = \frac{N_c - N_d}{\binom{n}{2}} = \frac{N_c - N_d}{n(n-1)/2}.$$

Ο συντελεστής τ, δηλαδή, παριστάνει την διαφορά μεταξύ των ποσοστών των εναρμονισμένων και μη εναρμονισμένων ζευγών παρατηρήσεων.

Αν όλα τα ζεύγη παρατηρήσεων είναι εναρμονισμένα, τότε ο συντελεστής τ είναι ίσος με 1. Αν όλα τα ζεύγη είναι μη εναρμονισμένα, τότε η τιμή του συντελεστή τ είναι –1. Είναι, δηλαδή, οι τιμές του συντελεστή τ μεταξύ –1 και 1. Επιπλέον, ο συντελεστής τ ικανοποιεί όλες τις προϋποθέσεις που προαναφέρθηκαν.

Ο υπολογισμός του συντελεστή τ γίνεται απλούστερος, αν οι παρατηρήσεις (X_i, Y_i), i = 1, 2, ..., n διαταχθούν σε μία στήλη κατά αύξουσα τάξη μεγέθους των τιμών των παρατηρήσεων πάνω στην τυχαία μεταβλητή X. Τότε, κάθε Y τιμή χρειάζεται να συγκριθεί μόνο με τις Y τιμές που είναι *"κάτω"* από αυτήν. Έτσι, κάθε ζεύγος παρατηρήσεων εξετάζεται μόνο μία φορά και ο αριθμός των συσχετισμένων και μη συσχετισμένων ζευγών προσδιορίζεται γρηγορότερα.

Παράδειγμα 3.4.3: Ας θεωρήσουμε τα δεδομένα πάνω στην επιθετικότητα των διδύμων. Διατάσσοντας τις παρατηρήσεις (X_i, Y_i), i = 1, 2, ..., n κατά αύξουσα τάξη μεγέθους των τιμών των παρατηρήσεων X_i, i = 1, 2, ..., n, καταλήγουμε στον εξής πίνακα:

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

Εδώ $X^{(i)}$, $i = 1, \ldots, n$ είναι η διατεταγμένη ακολουθία των παρατηρήσεων X_i, και Y_i^*, $i = 1, \ldots, n$ η προκύπτουσα αναδιάταξη των αντιστοιχουσών 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 $Y_i^* < Y_{i+1}^*$ όταν $X^{(i)} < X^{(i+1)}$, $(i = 1, \ldots, n\text{-}1)$. Η τρίτη στήλη δίνει τον αριθμό των ζευγών $(X^{(i+1)}, Y_{i+1}^*)$ για τα οποία $Y_i^* > Y_{i+1}^*$ όταν $X^{(i)} < X^{(i+1)}$ $(i = 1, \ldots, n\text{-}1)$.

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

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

Υπάρχει, επομένως, θετική συσχέτιση τάξης μεγέθους μεταξύ των μετρήσεων της επιθετικότητας των διδύμων, όπως προκύπτει από την μέτρηση του συντελεστή συσχέτισης τ του Kendall.

Ο συντελεστής τ μπορεί, επίσης, να χρησιμοποιηθεί ως ελεγχοσυνάρτηση για τον έλεγχο της μηδενικής υπόθεσης της ανεξαρτησίας μεταξύ των τυχαίων μεταβλητών X και Y, με αμφίπλευρες ή μονόπλευρες εναλλακτικές, όπως εξάλλου και στην περίπτωση του συντελεστή συσχέτισης ρ του Spearman. Περισσότερο συχνή, όμως, είναι η χρήση της διαφοράς $N_c - N_d$ ως ελεγχοσυνάρτησης για τον έλεγχο των υποθέσεων αυτών. Χρησιμοποιούμε, δηλαδή, ως ελεγχοσυνάρτηση την στατιστική συνάρτηση

$$T = N_c - N_d,$$

την οποία ονομάζουμε *ελεγχοσυνάρτηση του Kendall*.

Τα ποσοστιαία σημεία της κατανομής της στατιστικής συνάρτησης T δίνονται από τον πίνακα 12 του παραρτήματος. Είναι προφανές ότι μεγάλες τιμές της στατιστικής συνάρτησης T αποτελούν ένδειξη εναντίον της υπόθεσης H_0 και υπέρ της μονόπλευρης εναλλακτικής υπόθεσης θετικής συσχέτισης μεταξύ των μεταβλητών X και Y. Αντίστοιχα, μικρές τιμές της στατιστικής συνάρτησης T αποτελούν ένδειξη εναντίον της υπόθεσης H_0 και υπέρ της μονόπλευρης εναλλακτικής υπόθεσης αρνητικής συσχέτισης μεταξύ των μεταβλητών X και Y. Επομένως, θεωρώντας την ταξινόμηση των ζευγών υποθέσεων που θα μπορούσαν να μας ενδιαφέρουν στο πλαίσιο του παρόντος προβλήματος ή παρόμοιων προβλημάτων, στις κατηγορίες

Α. (Αμφίπλευρος έλεγχος συσχέτισης)

Β. (Μονόπλευρος έλεγχος θετικής συσχέτισης)

Γ. (Μονόπλευρος έλεγχος αρνητικής συσχέτισης),

ο κανόνας απόφασης διαμορφώνεται ως εξής:

Α. Η υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας α, αν

$T > w_{1-\alpha/2}$ ή αν $T < w_{\alpha/2} = -w_{1-\alpha/2}$.

Β. Η υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας α, αν

$T > w_{1-\alpha}$.

Γ. Η υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας α, αν

$T < w_{\alpha} = -w_{1-\alpha}$.

Επιστρέφοντας στο παράδειγμά μας, ας υποθέσουμε ότι ενδιαφερόμαστε να ελέγξουμε την υπόθεση H_0: οι μεταβλητές Χ και Υ είναι ασυσχέτιστες, εναντίον της εναλλακτικής H_1: οι μεταβλητές Χ και Υ είναι συσχετισμένες (αμφίπλευρος έλεγχος).

Από τον πίνακα που κατασκευάσθηκε παραπάνω, έχουμε ότι η παρατηρούμενη τιμή της στατιστικής συνάρτησης Τ είναι

$$N_c - N_d = 49-12 = 37.$$

Από τον σχετικό πίνακα του παραρτήματος, προκύπτει ότι τα ποσοστιαία σημεία για έναν αμφίπλευρο έλεγχο μεγέθους α=0.05, για n=12, είναι $w_{0.975} = 28$ και $w_{0.025} = -w_{0.975} = -28$. Επομένως, η τιμή της στατιστικής συνάρτησης Τ υπερβαίνει το 0.975-ποσοστιαίο σημείο της κατανομής της και, κατά συνέπεια, η μηδενική υπόθεση ότι δεν υπάρχει συσχέτιση μεταξύ των Χ και Υ απορρίπτεται σε επίπεδο σημαντικότητας 0.05. Το κρίσιμο επίπεδο του ελέγχου αυτού είναι, όπως φαίνεται από τον σχετικό πίνακα του παραρτήματος, περίπου ίσο με

$$\hat{\alpha} \cong 2(0.005) = 0.01.$$

Λύση με το MINITAB: Το MINITAB δεν δίνει την δυνατότητα ελέγχου της υπόθεσης ύπαρξης συσχέτισης με βάση τον συντελεστή συσχέτισης τ του Kendall και δεν προσφέρεται για την διεξαγωγή του με έμμεσο υπολογισμό της τιμής του τ με εύκολο τρόπο.

Λύση με το SPSS: Ο συντελεστής συσχέτισης του Kendall είναι διαθέσιμος από το ίδιο πλαίσιο που χρησιμοποιήσαμε για τον συντελεστή συσχέτισης του Spearman. Ο υπολογισμός του από το SPSS δίνει:

Correlations

			Πρωτότοκοι	Δευτερότοκοι
Kendall's tau_b	Πρωτότοκοι	Correlation Coefficient	1.000	.583**
		Sig. (2-tailed)	.	.010
		N	12	12
	Δευτερότοκοι	Correlation Coefficient	.583**	1.000
		Sig. (2-tailed)	.010	.
		N	12	12

* Correlation is significant at the .05 level (2-tailed).

Ο συντελεστής συσχέτισης είναι 0.583 και η τιμή του κρίσιμου επιπέδου για αμφίπλευρο έλεγχο είναι 0.01. Όπως δηλώνει το πρόγραμμα, η τιμή που βρήκαμε είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας α=0.05 και, επομένως, η υπόθεση της έλλειψης συσχέτισης μεταξύ X και Y δεν είναι εύλογη σε επίπεδο σημαντικότητας 0.05.

Σημειώνεται ότι και στην περίπτωση τουν συντελεστή τ, το SPSS δεν παρέχει την τιμή της ελεγχοσυνάρτησης T.