3.4 ΜΕΤΡΑ ΣΥΣΧΕΤΙΣΗΣ ΤΑΞΗΣ ΜΕΓΕΘΟΥΣ

(Measures of Rank Correlation)

Μέτρο Συσχέτισης είναι μία τυχαία μεταβλητή η οποία χρησιμοποιείται σε περιπτώσεις όπου τα δεδομένα αποτελούνται από ζεύγη τιμών, δηλαδή, σε περιπτώσεις που έχουμε ένα διμεταβλητό τυχαίο δείγμα μεγέθους n, έστω (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n). Συχνά, θα χρησιμοποιούμε τον συμβολισμό (X, Y) όταν αναφερόμαστε στο ζεύγος (X_i, Y_i), i = 1, 2, ..., n, εν γένει. Δηλαδή, θα υποθέτουμε ότι οι παρατηρήσεις του δείγματος (X_i, Y_i), i = 1, 2, ..., n είναι ισόνομες και ότι η κοινή διμεταβλητή κατανομή τους είναι η ίδια με την διμεταβλητή κατανομή του τυχαίου διανύσματος (X, Y).

Παραδείγματα διμεταβλητών τυχαίων μεταβλητών περιλαμβάνουν την περίπτωση όπου X_i παριστάνει το ύψος του i ατόμου ενός δείγματος και Y_i παριστάνει το ύψος του πατέρα του, ή την περίπτωση όπου X_i παριστάνει το αποτέλεσμα ενός τεστ για το i άτομο και Y_i παριστάνει το βαθμό κατάρτισης του ατόμου αυτού στο συγκεκριμένο θέμα. Οι τυχαίες μεταβλητές X και Y μπορεί ακόμη να είναι και ανεξάρτητες, όπως θα μπορούσαν να είναι στην περίπτωση που X_i παριστάνει την μέση βαθμολογία ενός παίκτη του basketball και Y_i παριστάνει την βαθμολογία της φίλης του σε ένα συγκεκριμένο μάθημα.

Ένα μέτρο συσχέτισης μεταξύ των μεταβλητών X και Y πρέπει να ικανοποιεί τις εξής προϋποθέσεις για να είναι αποδεκτό.

1. Η τιμή του μέτρου συσχέτισης 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,

επομένως, οι μικρότερες τιμές της μεταβλητής X τείνουν να αντιστοιχούν στις μικρότερες τιμές της μεταβλητής Y, τότε το μέτρο συσχέτισης θα πρέπει να είναι θετικό και να πλησιάζει την τιμή +1, αν η τάση αυτή είναι ισχυρή. Στην περίπτωση αυτή, θα μιλάμε για *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 πρέπει να έχει μία τιμή αρνητική, η οποία να είναι κοντά στην τιμή +1, αν η τάση είναι ισχυρή. Στην περίπτωση αυτή, θα λέμε ότι οι μεταβλητές X και Y είναι *αρνητικά συσχετισμένες*.

4. Αν οι τιμές της τυχαίας μεταβλητής X φαίνονται να αντιστοιχούν με τυχαίο τρόπο σε τιμές της τυχαίας μεταβλητής Y, το μέτρο συσχέτισης θα πρέπει να έχει μία τιμή αρκετά κοντά στο 0. Αυτή θα ήταν η περίπτωση όπου οι τυχαίες μεταβλητές X και Y είναι ανεξάρτητες και, ενδεχομένως, κάποιες άλλες περιπτώσεις όπου οι μεταβλητές X και Y δεν είναι ανεξάρτητες. Θα λέμε στις περιπτώσεις αυτές ότι οι τυχαίες μεταβλητές X και Y είναι *ασυσχέτιστες*, ή ότι *δεν σχετίζονται*, ή ότι *έχουν συσχέτιση* 0.

Το πιο συχνά χρησιμοποιούμενο μέτρο συσχέτισης είναι *ο συντελεστής συσχέτισης του Pearson*, ο οποίος συμβολίζεται με r και ορίζεται ως εξής:

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

όπου $\overline{X}$ είναι ο μέσος των τιμών $X_1, X_2, ..., X_n$ και $\overline{Y}$ είναι ο μέσος των τιμών $Y_1, Y_2, ..., Y_n$.

Μία αλλη μορφή του συντελεστή συσχέτισης r του Pearson, με την οποία είναι ευρύτατα γνωστός και η οποία προσφέρεται πολύ περισσότερο για ταχύτερους υπολογισμούς, είναι η εξής:

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

Αν ο αριθμητής και ο παρονομαστής στο δεξί μέλος της πρώτης σχέσης διαιρεθούν με n, ο συντελεστής συσχέτισης r μπορεί να γραφεί με τη μορφή

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

η οποία μπορεί να ερμηνευθεί ως ο λόγος της συνδιασποράς του δείγματος προς το γινόμενο των τυπικών αποκλίσεων των δύο περιθωρίων δειγμάτων. Είναι προφανές ότι ο συντελεστής αυτός ικανοποιεί τις προϋποθέσεις 1–4. Όμως, η κατανομή του συντελεστή αυτού εξαρτάται από την διμεταβλητή κατανομή του διανύσματος (X, Y). Επομένως, ο συντελεστής r δεν προσφέρεται ως ελεγχοσυνάρτηση για μη παραμετρικούς ελέγχους ή ως στατιστική συνάρτηση για την κατασκευή διαστημάτων εμπιστοσύνης εκτός, βέβαια, αν η κατανομή του διανύσματος (X, Y) είναι γνωστή.

Πολλά άλλα μέτρα συσχέτισης έχουν προταθεί, τα οποία ικανοποιούν τις προαναφερθείσες προϋποθέσεις για να είναι αποδεκτά

μέτρα συσχέτισης. Μερικά από αυτά τα μέτρα έχουν κατανομές οι οποίες δεν εξαρτώνται από την κατανομή του διανύσματος (X, Y) αν οι μεταβλητές X και Y είναι ανεξάρτητες και, επομένως, μπορούν να χρησιμοποιηθούν ως ελεγχοσυναρτήσεις σε μη παραμετρικούς ελέγχους ανεξαρτησίας. Τα μέτρα συσχέτισης, τα οποία εξετάζονται στην συνέχεια, εξαρτώνται μόνο απο τις τάξεις μεγέθους των παρατηρήσεων των δειγμάτων και έχουν κατανομές, οι οποίες είναι ανεξάρτητες από την κατανομή του διανύσματος (X, Y), αν οι μεταβλητές X και Y είναι ανεξάρτητες και συνεχείς.