ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΜΕΤΡΑ ΣΥΜΦΩΝΙΑΣ ΑΞΙΟΛΟΓΗΤΩΝ

Παντελής Ε. Γρηγοράσκος

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής

Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των

απαιτήσεων για την απόκτηση του Μεταπτυχιακού

Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς,

Σεπτέμβριος 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΜΕΤΡΑ ΣΥΜΦΩΝΙΑΣ ΑΞΙΟΛΟΓΗΤΩΝ

Παντελής Ε. Γρηγοράσκος

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής

Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των

απαιτήσεων για την απόκτηση του Μεταπτυχιακού

Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς,

Σεπτέμβριος 2022

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστική Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμώ …… συνεδρίαση του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της επιτροπή ήταν:

- Σ. Μπερσίμης (Αναπληρωτής Καθηγητής) (επιβλέπων)
- Δ. Παναγιωτάκος (Καθηγητής)
- Γ. Τζαβελάς (Αναπληρωτής Καθηγητής)

Η έγκριση της Διπλωματικής Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστική Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

INTERRATER AGREEMENT MEASURES

By

Pantelis E. Grigoraskos

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
September 2022

Στη μητέρα μου και στην κοπέλα μου …

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα αναπληρωτή καθηγητή κύριο Μπερσίμη Σωτήριο για την καθοδήγηση του σε όλη τη διάρκεια της εκπόνησης της παρούσας διπλωματικής εργασίας, όπως επίσης για τις εύστοχες υποδείξεις που μου έκανε για να πετύχω το καλύτερο δυνατό αποτέλεσμα.

Περίληψη

Στα πλαίσια της παρούσας εργασίας πραγματοποιείται μια κριτική ανασκόπηση των μέτρων συμφωνίας αξιολογητών. Το πιο διαδεδομένο μέτρο είναι το μέτρο k του Cohen. Ωστόσο, πολλοί συγγραφείς – ερευνητές παρουσίασαν εναλλακτικούς τρόπους και επεκτάσεις με σκοπό της προσπάθεια καταπολέμησης των παραδόξων που επιφέρει ο δείκτης αυτός. Η μέτρηση της συμφωνίας κατηγοριοποιείται ανάλογα με το πλήθος των βαθμολογητών, την ύπαρξη ελλειπουσών τιμών καθώς και το είδος των δεδομένων προς αξιολόγηση. Επίσης, παρουσιάζεται μια μη παραμετρική δοκιμή ως ένας εναλλακτικός τρόπος προσέγγισης της μέτρησης της συμφωνίας. Εκτός των δεικτών, στη βιβλιογραφία προτείνονται και μοντέλα με γνωστότερα τα log-linear.

Abstract

In the context of this bachelor's thesis, a critical review of rater agreement measures is carried out. The most common measure is Cohen's k measure. However, many authors – researchers presented alternative ways and extensions in order to try to combat the paradoxes brought about this index. The measurement of agreement is categorized according to the numbers of raters, the existence of missing values as well as the type of data to be valued. Furthermore, a non-parametric way of approaching the measurement of agreement. In addition to the indicators, in the literature, models are also proposed. The most well-known are the log - linear models.

Περιεχόμενα

Εισαγωγή

Γενικά

Η βιομηχανία της υγείας βασίζεται σε πλήθος ανθρώπων που συλλέγουν ερευνητικά ή κλινικού εργαστηρίου δεδομένα. Γεννιέται, λοιπόν το ερώτημα αν όλοι αυτοί οι άνθρωποι συμφωνούν καθώς είναι λογικό να υπάρχουν διαφοροποιήσεις ανάμεσα τους. Πρέπει να υπάρξουν έρευνες οι οποίες μετρούν τη συμφωνία μεταξύ τους. Πέρα από το κλάδο της ιατρικής υπάρχουν και άλλα επιστημονικά πεδία στα οποία είναι χρήσιμη η μελέτη συμφωνίας. Γίνεται χρήση στην ψυχολογία όπου για παράδειγμα άτομα ταξινομούνται σε κατηγορίες όπως κατάθλιψη, σχιζοφρένεια κλπ. Ευρέως γνωστή είναι στις κοινωνικές επιστήμες όπως στη δημογραφία, την εκπαίδευση, την κοινωνιολογία κλπ. Η τέλεια συμφωνία σπανίως επιτυγχάνεται ακόμα και αν οι ιατροί που αξιολογούν κάποια υποκείμενα είναι υψηλά καταξιωμένοι. Η έκταση της συμφωνίας μεταξύ των ανθρώπων που συλλέγουν δεδομένα καλείται «interrater reliability».

Σε διάφορες έρευνες τα αντικείμενα ταξινομούνται σε κατηγορίες από δύο βαθμολογητές. Για παράδειγμα, δύο ιατροί ταξινομούν τους ασθενείς ανάλογα με το στάδιο της ασθένειας τους. Η μελέτη γίνεται ακόμα πιο πολύπλοκη όταν οι κατηγορίες που μπορεί να ταξινομηθεί ένας ασθενής είναι περισσότερες από δύο. Σε όλα αυτά πρέπει να συνυπολογίσουμε ότι μπορεί ένας αξιολογητής να μην κατατάξει κάποιον ασθενή (missing data) ή να τον κατατάξει σε περισσότερες από μία κατηγορίες. Η μελέτη της συμφωνίας μπορεί να επεκταθεί και στη περίπτωση όπου έχουμε περισσότερους από δύο αξιολογητές. Πάνω σε όλα αυτά μπορεί να προστεθεί και ποιο είναι το είδος των δεδομένων που μελετάμε (nominal, ordinal, interval). Λαμβάνοντας υπόψη όλες αυτές τις περιπτώσεις, το πρόβλημα της ανάλυσης της συμφωνίας γίνεται ακόμα πιο πολύπλοκο. Για τη διερεύνηση γίνεται χρήση είτε μέτρων είτε μοντέλων.

Διάρθρωση της εργασίας

Στο κεφάλαιο 1 περιγράφουμε τα πιο γνωστά μέτρα που αφορούν την περίπτωση δύο αξιολογητών σε ονομαστικά δεδομένα. Οι δείκτες είναι το percent agreement, το kappa του Cohen, ο σταθμισμένος kappa του Cohen, ο εντός των τάξεων συντελεστής συσχέτισης, ο τετραχωρικός συντελεστής συσχέτισης, το Pi του Scott, το alpha του Krippendorff, ο συντελεστής AC_1 του Gwet και ο δείκτης G. Επίσης, υπάρχει αριθμητικό παράδειγμα και μια παράγραφο που αναλύει τη περίπτωση των ελλειπουσών βαθμολογιών (missing rates).

Στο δεύτερο κεφάλαιο παρουσιάζονται μέτρα σχετικά με τρείς ή περισσότερους αξιολογητές σε ονομαστικά δεδομένα, συνοδευόμενα από ένα αριθμητικό παράδειγμα. Οι επεκτάσεις αυτές είναι το γενικευμένο μέτρο του Fleiss, του Conger, του Gwet και το alpha του Krippendorff.

Στο τρίτο κεφάλαιο γίνεται μια ξεχωριστή αναφορά σε ένα μη παραμετρικό τεστ που προτάθηκε από τους Μπερσίμης, Σαχλάς και Chakraborti. Είναι ένα μέτρο που μπορεί να χρησιμοποιηθεί όταν θέλουμε μια άμεση ανάλυση συμφωνίας σε δύο ή περισσότερους αξιολογητές.

Στο τέταρτο κεφάλαιο μελετάται η περίπτωση που τα δεδομένα είναι είτε διατάξιμα (ordinal), είτε σε διάστημα (interval), είτε είναι δεδομένα αναλογίας (ratio). Επιπροσθέτως υπάρχει ένα υπολογιστικό παράδειγμα, καθώς και μία παράγραφο που κάνει λόγο για τη χρήση συντελεστών βαρύτητας και με ποιο κριτήριο τους επιλέγουμε.

Στο πέμπτο κεφάλαιο της εργασίας γίνεται λόγος για χρήση στατιστικών μοντέλων και συγκεκριμένα τα λεγόμενα log-linear. Τέλος, υπάρχει ένα παράδειγμα χρήση αυτών και ποια είναι η κατάλληλη επιλογή του μοντέλου που τελικά προσαρμόζεται καλά στα δεδομένα.

Τέλος, στο έκτο κεφάλαιο της εργασίας δίνεται ο κώδικας στην γλώσσα προγραμματισμού R, με σκοπό να στελεχώσει τα αποτελέσματα που δόθηκαν στα παραδείγματα των κεφαλαίων 2 και 3. Περιέχονται δύο παραδείγματα. Το πρώτο περιλαμβάνει κώδικα από υπαρκτά δεδομένα για δυο αξιολογητές και το δεύτερο για τέσσερις.

Κεφάλαιο 1ο: Μέτρα συμφωνίας αξιολογητών και επεκτάσεις με δυο βαθμολογητές

1.1 Percent Agreement

Στη παρούσα παράγραφο θα εισάγουμε το μέτρο της ποσοστιαίας συμφωνίας που όπως είναι γνωστή στη διεθνή βιβλιογραφία ως «*percent agreement*». Για να υπολογιστεί αυτό το μέτρο θα χρειαστεί ένας πίνακας όπου στις στήλες του θα περιλαμβάνει τους διαφορετικούς βαθμολογητές και στις γραμμές του τις μεταβλητές για τις οποίες οι βαθμολογητές έχουν συλλέξει δεδομένα. Στα κελιά περιέχονται τα σκορ με τα οποία κάθε βαθμολογητής έχει αξιολογήσει τη κάθε μεταβλητή. Στον Πίνακα 1. υπάρχει ένα παράδειγμα τέτοιου πίνακα. Πιο συγκεκριμένα έχουμε δύο βαθμολογητές (*X* και *Y*) και τη δίτιμη (0-1) καταγραφή των αξιολογήσεων τους στις δέκα μεταβλητές. Για τον υπολογισμό του συγκεκριμένου μέτρου καταγράφεται η διαφορά που έχουν οι δύο βαθμολογητές. Αν υπάρχει ομοφωνία η διαφορά τους είναι μηδενική, ενώ αν υπάρχει διαφωνία τότε στη στήλη *Difference* καταγράφεται η τιμή 1 ή -1 (αναλόγως με τον ποιον έχουμε θεωρήσει ως αφαιρετέο). Στο τέλος του πίνακα μετράμε το πλήθος των μηδενικών και για τον υπολογισμό του μέτρου κάνουμε χρήση του πηλίκου:

$$\frac{\textit{πλήθος μηδενικών}}{\textit{πλήθος των μεταβλητών}}$$

Στο παράδειγμα του Πίνακα 1. έχει βρεθεί ποσοστό 80%. Αυτό σημαίνει ότι το 20% των δεδομένων της έρευνας είναι εσφαλμένο καθώς μόνο ένας από τους δύο βαθμολογητές μπορεί να είναι ο «σωστός».

Το ίδιο μέτρο μπορεί να υπολογιστεί και όταν έχουμε περισσότερους από δύο αξιολογητές και στις μεταβλητές να εισάγεται η αξιολόγηση 1 ή 0. Στο Πίνακα 2. υπάρχει ένα αντιπροσωπευτικό παράδειγμα. Σε κάθε γραμμή υπολογίζεται το ποσοστό συμφωνίας και στο τέλος προκύπτει ο μέσος όρος όλων των μεταβλητών. Για παράδειγμα στη μεταβλητή 4 παρατηρούμε ότι συμφωνούν οι 4 από τους 5 αξιολογητές το οποίο μεταφράζεται σε ποσοστό 80%. Πλεονέκτημα της μεθόδου είναι ότι επιτρέπει στον ερευνητή να ανακαλύψει αν τα σφάλματα είναι τυχαία και αν διανέμονται ισοδύναμα ανάμεσα στους αξιολογητές και στις μεταβλητές ή αν ένας συγκεκριμένος αξιολογητής βαθμολογεί συστηματικά διαφορετικά συγκριτικά με τους άλλους. Επιπρόσθετο θετικό στοιχείο είναι ότι μπορούμε να αναγνωρίσουμε αν υπάρχει κάποια "προβληματική" μεταβλητή με την έννοια ότι υπάρχει σε αυτή μεγάλη διχογνωμία των βαθμολογητών (π.χ μεταβλητή 10). Να σημειωθεί ότι τα δεδομένα των παρακάτω πινάκων είναι πλασματικά με

μοναδικό σκοπό να γνωρίσουμε τη μορφή ενός πίνακα δεδομένων και τον τρόπο υπολογισμού του μέτρου "*percent agreement*".

Πίνακας 1: Υπολογισμός ποσοστιαίας συμφωνίας (Πλασματικά Δεδομένα)

Μεταβλητή	Αξιολογητές		Διαφορά
	X	Y	
1	1	1	0
2	1	0	1
3	1	1	0
4	0	1	-1
5	1	1	0
6	0	0	0
7	1	1	0
8	1	1	0
9	0	0	0
10	1	1	0
Πλήθος μηδενικών			8
Πλήθος αντικειμένων			10
Ποσοστό συμφωνίας			80

Πίνακας 2: Ποσοστιαία συμφωνία μεταξύ πολλών (5) βαθμολογητών (Πλασματικά Δεδομένα)

Μεταβλητή	Αξιολογητές					Συμφωνία %
	X	Y	Z	W	V	
1	1	1	1	1	1	1.00
2	1	1	1	1	1	1.00
3	1	1	1	1	1	1.00
4	0	1	1	1	1	0.80
5	0	1	0	0	0	0.80
6	0	0	0	0	0	1.00
7	1	1	1	1	1	1.00
8	1	1	1	1	0	0.80
9	0	0	0	0	0	1.00
10	1	1	0	0	1	0.60
Interrater Reliability						0.9

Μέχρι τώρα, έχουμε σιωπηρά υποθέσει ότι η πλειοψηφία των αξιολογήσεων ήταν ορθή και ότι οι βαθμολογητές έχουν κάνει μια συνειδητή αξιολόγηση. Στο μέτρο του Cohen που περιγράφεται σε επόμενη παράγραφο λαμβάνονται υπόψη αυτοί οι προβληματισμοί. Δηλαδή το μέτρο του Cohen προϋποθέτει ότι πολλοί αξιολογητές έχουν κάνει μια τυχαία επιλογή για να χαρακτηρίσουν μια μεταβλητή. Σε αυτή τη περίπτωση, η τυχαία συμφωνία λαμβάνεται ως λανθασμένη συμφωνία.

1.2 Μέτρο k του Cohen

Αξίζει μια περιληπτική αναφορά στο πιο γνωστό μέτρο συμφωνίας που είναι ο συντελεστής k του Cohen (1960). Ο σκοπός του ήταν να λάβει υπόψη τη τυχαία συμφωνία μεταξύ των αξιολογητών.

Για τη δημιουργία αυτού του μέτρου έχουν υιοθετηθεί οι παρακάτω υποθέσεις. Αρχικά, έχουμε υπόψη δυο αξιολογητές, που βαθμολογούν n αντικείμενα (ανεξάρτητα) με βάση μια κλίμακα *I* κατηγοριών (ανεξάρτητες), από το 1 έως το *I*. Να τονιστεί ότι οι αποκλίσεις ανάμεσα στις κατηγορίες αντιμετωπίζονται ισοδύναμα. Δηλαδή η απόκλιση ανάμεσα στη πρώτη με δεύτερη κατηγορία είναι ισοδύναμη με την απόκλιση ανάμεσα στη πρώτη με τρίτη κατηγορία κ.ο.κ.

Κατασκευάζεται, λοιπόν, ένας δισδιάστατος *IxI* πίνακας συνάφειας, έχοντας άγνωστες τις περιθώριες κατανομές για καθένα από τους δυο αξιολογητές. Συμβολίζουμε με p_{ij} το δειγματικό ποσοστό των αντικειμένων στο *(i,j)*-οστό κελί, όπου *i*-κατηγορία προκύπτει από τον πρώτο βαθμολογητή και *j*-κατηγορία από τον δεύτερο. Επίσης, με $p_{i.} = \sum_{j=1}^{I} p_{ij}$ συμβολίζουμε το ποσοστό των αντικειμένων που βαθμολογήθηκαν στην *i*-κατηγορία από τον πρώτο βαθμολογητή και με $p_{.j} = \sum_{i=1}^{I} p_{ij}$ το ποσοστό των αντικειμένων που βαθμολογήθηκαν στην *j*-κατηγορία από το δεύτερο βαθμολογητή. Τέλος, με $p_e = \sum_{i=1}^{I} p_{ij}$ θα θεωρήσουμε το παρατηρηθέν ποσοστό συμφωνίας και με $p_a = \sum_{i=1}^{I} p_{i.} p_{.i}$ το ποσοστό συμφωνίας που οφείλεται σε τυχαίους παράγοντες.

Ο τύπος υπολογισμού του είναι:

$$\hat{k} = \frac{p_a - p_e}{1 - p_e} \tag{1.1}$$

Το k του Cohen είναι μια επέκταση του δείκτη του Scott (1955). Ο τελευταίος θεώρησε τη ποσότητα p_α, χρησιμοποιώντας την υπόθεση ότι οι *I* κατηγορίες έχουν γνωστή κατανομή και ίση για τους δύο αξιολογητές. Συνεπώς, αν πρόκειται για τις ίδιες περιθώριες κατανομές τότε τα μέτρα των Scott και Cohen ταυτίζονται.

Για την υπόθεση $k = 0$, δηλαδή αν ο συντελεστής k είναι στατιστικά σημαντικός ή όχι, θα μπορούσαμε να χρησιμοποιήσουμε το τύπο της ασυμπτωτικής διακύμανσης των Fleiss et al. (1969). Για μεγάλο δείγμα n, ο παραπάνω τύπος είναι ισοδύναμος της διακύμανσης του

Everitt (1968) που βασίστηκε στην υπεργεωμετρική κατανομή. Κάτω από την υπόθεση της τυχαίας συμφωνίας παίρνουμε τον τύπο:

$$\widehat{Var_0}(\hat{k}) = \frac{p_c + p_c^2 - \sum_{i=1}^{I} p_{i.}p_{.i}(p_{i.} + p_{.i})}{n(1-p_c)^2} \quad (1.2)$$

Υποθέτοντας ότι η ποσότητα $\frac{\hat{k}}{\sqrt{\widehat{Var_0}(\hat{k})}}$ ακολουθεί τη κανονική κατανομή, μπορούμε να ελέγξουμε την υπόθεση της τυχαίας συμφωνίας με αναφορά στην τυπική κανονική. Ωστόσο, η εξέταση της συμφωνίας δε παρουσιάζει κάποιο ενδιαφέρον, με την έννοια ότι η αξιοπιστία των μεθόδων ή αξιολογητών θεωρείται δεδομένη, αφού θεωρητικά εκπαιδεύονται για να είναι αξιόπιστοι. Συνεπώς, προτείνεται ένα μικρότερο όριο του kappa από τους Fleiss et al. (1969) με ασυμπτωτική έκφραση.

$$\widehat{Var}\left(\hat{k}\right) = \frac{1}{n(1-p_c)}\Big(\sum_{i=1}^{I} p_{ii}\{1-(p_{i.}+p_{.i})(1-\hat{k})\}^2 + (1-\hat{k})^2 \sum_{i\neq j}^{I} p_{ij}(p_{i.}+p_{.j})^2 - \{\hat{k} - p_c(1-\hat{k})\}^2\Big) \quad (1.3)$$

Για τη μελέτη της ακρίβειας του τυπικού σφάλματος του $\hat{k}$ δίνεται από τους Cicchetti και Fleiss (1977) και από τους Fleiss και Cicchetti (1978) με χρήση Monte Carlo προσομοίωσης.

Σαφή όρια για το χαρακτηρισμό της συμφωνίας ανάλογα με τη τιμή του $\hat{k}$ δεν υπάρχουν. Οι Landis και Koch (1977a) έχουν δώσει τους εξής χαρακτηρισμούς για τα επίπεδα συμφωνίας.

$\hat{k} > 0.75$ πολύ καλή συμφωνία

$0.4 < \hat{k} < 0.75$ καλή συμφωνία

$\hat{k} < 0.4$ φτωχή συμφωνία

Όρια του συντελεστή kappa του Cohen:

Όταν έχουμε απόλυτη ταύτιση της συμφωνίας με τη τυχαία συμφωνία προκύπτει $k = 0$. Όταν έχουμε μεγαλύτερες τιμές από τη τυχαία συμφωνία οδηγούμαστε σε θετικές τιμές του $\hat{k}$ και για μικρότερες τιμές προκύπτουν αρνητικές τιμές του $\hat{k}$. Το ανώτερο όριο είναι η τιμή 1, στη μοναδική περίπτωση που θα έχουμε τη τέλεια συμφωνία μεταξύ των αξιολογητών. Η τιμή του κατώτερου ορίου εξαρτάται από τις περιθώριες κατανομές. Σύμφωνα με τον Cohen (1960), η θεωρητικά μικρότερη τιμή είναι $\frac{-p_e}{1-p_e}$. Στη πράξη ωστόσο, θεωρούμε ότι η συμφωνία δε μπορεί να είναι μικρότερη (χειρότερη) απ' ότι κάτω από την ανεξαρτησία (κ. Μπερσίμης Πανεπιστημιακές Σημειώσεις), με αποτέλεσμα η μικρότερη τιμή να είναι το 0.

Η χρήση και η ερμηνεία του kappa του Cohen προκάλεσαν αντιπαραθέσεις κυρίως για τη σχέση του με τις περιθώριες κατανομές. Οι περιθώριες κατανομές περιγράφουν πως ο κάθε βαθμολογητής κατανέμει τα n αντικείμενα στις I κατηγορίες. Όσο η μεροληψία του κάθε

βαθμολογητή μειώνεται, τόσο οι περιθώριες κατανομές τείνουν να ταυτιστούν. Οι Feinstein και Cicchetti (1990) και οι Byrt et al. (1993) ερεύνησαν την επίδραση της μεροληψίας των βαθμολογητών του kappa.

Ακόμα ένας παράγοντας που επηρεάζει τον kappa είναι το κατά πόσο μια τάση ταξινόμησης επικρατεί στον κάθε πληθυσμό. Οι ίδιοι οι βαθμολογητές μπορούν να καταλήξουν σε διαφορετικές τιμές του kappa όταν εξετάζουν δύο διαφορετικούς πληθυσμούς (Feinstein and Ciccetti 1990, Byrt et al 1993). Δηλαδή, η συμφωνία εξαρτάται και από τον ίδιο τον πληθυσμό που εξετάζουν.

Με βάση τα παραπάνω, είναι σημαντικό να καταλαβαίνει κανείς ότι οι μελέτες συμφωνίας μεταξύ βαθμολογητών που διεξάγονται σε δείγματα που είναι «βολικά» ή σε πληθυσμούς με υψηλή επικράτηση μιας διάγνωσης δεν αντανακλά τη συμφωνία μεταξύ των αξιολογητών.

Μερικοί συγγραφείς (Hutchinson 1993) υποστηρίζουν ότι ο συντελεστής kappa περιέχει δύο είδη διαφωνίας: τη διαφωνία που οφείλεται στη μεροληψία των βαθμολογητών και τη διαφωνία που προκύπτει από τη διαφορετική αντίληψη περί της κατάταξης των αντικειμένων μεταξύ των βαθμολογητών. Μια απάντηση, έρχεται από τον εντός των τάξεων kappa συντελεστή (Bloch και Kraemer 1989) που περιγράφεται παρακάτω. Ωστόσο, ο Zwick (1988) επισημαίνει ότι είναι προτιμότερο να προσπαθούμε να διορθώσουμε τις περιθώριες κατανομές από το να τις αγνοούμε και να διευκρινίζεται από τους ερευνητές αν οι διαφορές αυτές προέρχονται από τυχαία σφάλματα ή από σημαντική διαφωνία των αξιολογητών.

1.3 Σταθμισμένος kappa του Cohen

Σε πολλές περιπτώσεις, υπάρχουν διαφωνίες μεταξύ των αξιολογητών πιο σοβαρές από άλλες. Για παράδειγμα, σε ψια ψυχιατρική διάγνωση με κατηγορίες "διαταραχή προσωπικότητας", "νεύρωση" και "ψύχωση", ο ερευνητής θα επιθυμούσε να διαβαθμίσει τη διαφορά μεταξύ νεύρωσης και ψύχωσης ως πιο σοβαρή συγκριτικά με τη διαφορά νεύρωση και διαταραχής προσωπικότητας. Ο Cohen (1968) πρότεινε μια επέκταση του kappa με την ονομασία σταθμισμένοο kappa ($\widehat{k_w}$), με σκοπό να μετρά την αναλογία σταθμισμένης συμφωνίας διορθωμένη από τυχαιότητα.

Θεωρώντας ως w_{ij} το βάρος του (i,j) κελιού με $i, j = 1,2, \dots, I$ τότε ο $\widehat{k_w}$ δίνεται από τον τύπο:

$$\widehat{k_w} = \frac{\sum_{i=1}^{I} w_{ij} p_{ij} - \sum_{i=1}^{I} \sum_{j=1}^{I} w_{ij} p_{i.} p_{.j}}{1 - \sum_{i=1}^{I} \sum_{j=1}^{I} w_{ij} p_{.} p_{.j}} \qquad (1.4)$$

Παρατηρούμε ότι ο απλός συντελεστής kappa είναι ειδική περίπτωση του σταθμισμένου συντελεστή όταν $w_{ij} = 1$ για $i = j$ και $w_{ij} = 0$ για $i \neq j$. Επίσης, όταν $w_{ij} = \frac{1-(i-j)^2}{(I-1)^2}$, τότε ο $\widehat{k_w}$ μπορεί να ερμηνευτεί ως ένας εντός των τάξεων συντελεστής συσχέτισης για ανάλυση

διακύμανσης κατά δύο παράγοντες υπό την υπόθεση ότι τα n αντικείμενα και οι βαθμολογητές είναι τυχαία δείγματα από τους πληθυσμούς αντικειμένων και βαθμολογητών αντίστοιχα (Fleiss and Cohen 1973).

Ο Fleiss (1969) καθιέρωσε το τύπο υπολογισμού της ασυμπτωτικής διακύμανσης του $\widehat{k_w}$. Ο τύπος αυτός έχει αξιολογηθεί για τη χρησιμότητά του σε τεστ σημαντικότητας και για κατασκευή διαστημάτων εμπιστοσύνης σύμφωνα με τους Cicchetti και Fleiss (1977) και Fleiss και Ciccheti (1978). Έπειτα από Monte Carlo προσομοιώσεις κατέληξαν ότι μόνο μετρίου μεγέθους δείγματα είναι ικανά να εξετάσουν την υπόθεση ότι δύο $\widehat{k_w}$ είναι ίσα προερχόμενοι από ανεξάρτητα δείγματα. Ωστόσο, το ελάχιστο μέγεθος δείγματος που απαιτείται για τον καθορισμό διαστημάτων εμπιστοσύνης γύρω από μια τιμή του $\widehat{k_w}$ είναι $n = 16I^2$, το οποίο τις περισσότερες φορές είναι υπερβολικά μεγάλο.

1.4 Εντός των τάξεων συντελεστής kappa

Το 1989, οι Bloch και Kraemer εισήγαγαν τον εντός των τάξεων συντελεστή συσχέτισης (interclass kappa) ως εναλλακτική εκδοχή του kappa του Cohen, χρησιμοποιώντας την υπόθεση ότι για τον κάθε αξιολογητή ισχύει η ίδια περιθώρια κατανομή. Έχει αποδειχτεί ότι είναι αλγεβρικά ισοδύναμος με το δείκτη του Scott (1955).

Η υπόθεση μας εδώ είναι ότι τα n δεδομένα ταξινομούνται σε δύο κατηγορίες από δύο σταθερούς βαθμολογητές. Επίσης, υποθέτουμε ότι οι αξιολογήσεις των υποκειμένων είναι αμετάβλητες. Θεωρούμε ότι X_{ij} είναι η ταξινόμηση του *i*-αντικειμένου από τον *j*-βαθμολογητή (*i=1,…n , j=1,2*) και ότι για κάθε αντικείμενο ισχύει $p_i = P(X_{ij} = 1)$ η πιθανότητα η ταξινόμηση να είναι επιτυχής. Τότε έχουμε, $E(p_i) = P$, $P' = 1 - P$ και $var(p_i) = \sigma_p^2$. Ο τύπος υπολογισμού του δείκτη είναι:

$$\hat{k}_I = \frac{\sigma_p^2}{PP'} \tag{1.5}$$

Ο λογάριθμος της συνάρτησης πιθανοφάνειας δίνεται ως εξής:

$$\begin{aligned} &\ln L(P, k_I | n_{11}, n_{12}, n_{21}, n_{22}) \\ &\quad = n_{11} \ln(P^2 + k_I PP') + (n_{12} + n_{21}) \ln\{\mathrm{PP}'(1 - \mathrm{k_I})\} \\ &\quad + n_{22} \ln(P'^2 + k_I PP'). \end{aligned} \tag{1.6}$$

Οι προαναφερθείσες ποσότητες φαίνονται στον παρακάτω πίνακα για την από κοινού απόκριση.

Είδος απόκρισης			
X_{i1}	X_{i2}	Παρατηρούμενη συχνότητα	Αναμενόμενη πιθανότητα
1	1	n_{11}	$P^2 + k_I PP'$
1	0	n_{12}	$PP'(1 - k_I)$
0	1	n_{21}	$PP'(1 - k_I)$
0	0	n_{22}	$P'^2 + k_I PP'$

Οι εκτιμητές μέγιστης πιθανοφάνειας $\hat{P}$, $\widehat{k_I}$ για το P και k_I προκύπτουν ως εξής:

$$\hat{P} = \frac{2n_{11} + m_{12} + n_{21}}{2n} \tag{1.7}$$

$$\widehat{k_I} = \frac{4(n_{11}n_{22} - n_{12}n_{21}) - (n_{12} - n_{21})^2}{(2n_{11} + n_{12} + n_{21})(2n_{22} + n_{12} + n_{21})} \tag{1.8}$$

με εκτίμηση του σφάλματος,

$$SE(\widehat{k_I}) = \left\{ \frac{1 - \widehat{k_I}}{n} \left((1 - \widehat{k_I})(1 - 2\widehat{k_I}) + \frac{\widehat{k_I}(2 - \widehat{k_I})}{2\hat{p}(1 - \hat{p})} \right) \right\}^{\frac{1}{2}} \tag{1.9}$$

Υποθέτοντας ότι ο $\widehat{k_I}$ κατανέμεται κανονικά με μέσο k_I και τυπική απόκλιση $SE(\hat{k})$, τότε, το $100(1 - \alpha)\%$ διάστημα εμπιστοσύνης δίνεται από τον τύπο $\widehat{k_I} \pm z_{1-\frac{\alpha}{2}} SE(\widehat{k_I})$, όπου το $z_{1-\frac{\alpha}{2}}$ το $100(1 - \alpha)$ ποσοστιαίο σημείο της τυπικής κανονικής κατανομής. Το παραπάνω διάστημα εμπιστοσύνης έχει νόημα σε μεγάλα δείγματα που δύσκολα συναντάμε στη πράξη.

Οι Bloch και Kraemer (1989) απέδειξαν μια σταθεροποιημένη διακύμανση για τον $\widehat{k_I}$ η οποία βελτιώνει την ακρίβεια των εν λόγω διαστημάτων εμπιστοσύνης, την υπολογιστική ισχύς ή την κατασκευή στατιστικών τεστ. Αυτή η προσέγγιση βασίζεται στον εκτιμητή $\widehat{k_J}$ του k_I (Bloch και Kraemer 1989, Fleiss και Davies 1982). Ο εκτιμητής αυτός λαμβάνεται ως ο μέσος όρος των $\widehat{k_{-\iota}}$, , όπου $\widehat{k_{-\iota}}$ είναι η τιμή των $\widehat{k_I}$ συμπεριλαμβανομένων όλων των αντικειμένων εκτός της *i-οστής*. Ωστόσο, οι συγγραφείς τονίζουν ότι ο εκτιμητής αυτός είναι σχετικά υψηλός σε μικρά δείγματα, ειδικότερα όταν η P πλησιάζει είτε το 0 είτε το 1.

Για κατασκευή διαστημάτων εμπιστοσύνης σε μικρότερα δείγματα, οι Donner και Eliasziw (1992) προτείνουν μια διαδικασία βασισμένη σε X^2 τεστ καλής προσαρμογής. Η προσπάθεια τους βασίζεται στην εξίσωση του υπολογισμένου X^2 στατιστικού με ένα βαθμό ελευθερίας με μια κατάλληλα επιλεγμένη κριτική τιμή και επιλύοντας ως προς k καταλήγουμε σε δύο ρίζες: τη τιμή του άνω (k_U) και κάτω (k_L) ορίου του $100(1 - \alpha)\%$ διαστήματος εμπιστοσύνης για το k_I:

$$k_L = \left(\frac{1}{9}y_3^2 - \frac{1}{3}y_2\right)^{\frac{1}{2}} \left(\cos\frac{\theta + 2\pi}{3} + \sqrt{3}\sin\frac{\theta + 2\pi}{3} - \frac{1}{3}y_3\right. \quad (1.10)$$

$$k_U = 2\left(\frac{1}{9}y_3^2 - \frac{1}{3}y_2\right)^{\frac{1}{2}} \cos\frac{\theta + 5\pi}{3} - \frac{1}{3}y_3\,, \pi = 3{,}14 \quad (1.11)$$

όπου

$$\theta = \arccos\frac{V}{W} \quad , \quad V = \frac{1}{27}y_3^3 - \frac{1}{6}(y_2 y_3 - 3y_1) \quad , \quad W = \left(\frac{1}{9}y_3^2 - \frac{1}{3}y_2\right)^{\frac{3}{2}} ,$$

και

$$y_1 = \frac{\left\{n_{12} + n_{21} - 2n\hat{P}\left(1-\hat{P}\right)\right\}^2 + 4n^2\hat{P}^2\left(1-\hat{P}\right)^2}{4n\hat{P}^2(1-\hat{P})(\chi^2_{1,1-\alpha} + n)} - 1,$$

$$y_2 = \frac{(n_{12} + 1_{21})^2 - 4n\hat{P}\left(1-\hat{P}\right)\left\{1 - 4\hat{P}\left(1-\hat{P}\right)\right\}\chi^2_{1,1-\alpha}}{4n\hat{P}^2(1-\hat{P})(\chi^2_{1,1-\alpha} + n)} - 1,$$

$$y_3 = \frac{n_{12} + n_{21} + \left\{1 - 2\hat{P}\left(1-\hat{P}\right)\right\}\chi^2_{1,1-\alpha}}{\hat{P}\left(1-\hat{P}\right)\left(\chi^2_{1,1-\alpha} + n\right)} - 1.$$

Με αυτή τη διαδικασία η ακρίβεια τόσο στις τιμές του k_I και του P , σε μικρά δείγματα, έχει βελτιωθεί.

1.5 Τετραχωρικός συντελεστής συσχέτισης

Η χρήση του τετραχωρικού συντελεστής συσχέτισης αναφέρεται κυρίως στις ιατρικές επιστήμες. Πολλές κλινικές ανωμαλίες, στις οποίες οι γνώμες των ιατρών είναι διχασμένες, δε μπορούν να μετρηθούν είτε για τεχνικούς λόγους είτε εξαιτίας των περιορισμών της ανθρώπινης αντιληπτικής ικανότητας. Ένα παράδειγμα είναι η ακτινολογική αξιολόγηση της πνευμονοκονίασης, όπου αξιολογείται από ακτινογραφίες θώρακος προβάλλοντας μια πληθώρα από μικρές ακανόνιστες αδιαφάνειες. Με λίγα λόγια, ο συγκεκριμένος συντελεστής χρησιμοποιείται στις περιπτώσεις των δύο καταστάσεων π.χ φυσιολογικό-μη φυσιολογικό, που απορρέουν από συνεχή μεταβλητή.

Ένα πρώτο χαρακτηριστικό του είναι ότι αξιολογήσεις από δύο βαθμολογητές μπορεί να διαφέρουν ως προς το κατώφλι που έχει ορίσει ο καθένας. Με τον όρο κατώφλι εννοούμε τη τιμή η οποία ξεχωρίζει τις καταστάσεις ανωμαλίας ή μη. Οι δύο αξιολογητές χρησιμοποιούν διαφορετικά κατώφλια λόγω της δικής τους προσωπικής οπτικής γωνίας, παρά το γεγονός ότι υπάρχουν προκαθορισμένα ιατρικά κριτήρια για το καθορισμό της τιμής του κατωφλιού. Επιπροσθέτως, η πιθανότητα λανθασμένης ταξινόμησης μια κατάστασης ενός ασθενούς, εξαρτάται από τη πραγματική τιμή της συνεχούς μεταβλητής που εξετάζεται. Η τιμή αυτή όσο πιο κοντά είναι στο λεγόμενο κατώφλι, τόσο πιο πιθανό είναι να γίνει λανθασμένη ταξινόμηση. Αυτός, λοιπόν είναι και ο λόγος που τα προηγούμενα προαναφερθέντα μέτρα

(όπως ο μη σταθμισμένος και ο σταθμισμένος kappa, ο εντός των τάξεων kappa) δεν είναι κατάλληλοι για τέτοιες καταστάσεις.

Όταν η διάγνωση αντιμετωπίζεται ως η απόφαση για «φυσιολογικό» ή «μη φυσιολογικό» μιας συνεχούς μεταβλητής και πιο συγκεκριμένα μιας τυπικής κανονικής κατανομής, ο τετραχωρικός συντελεστής συσχέτισης είναι ο κατάλληλος δείκτης για εξέταση συμφωνίας μεταξύ δύο αξιολογητών. Πιο συγκεκριμένα, ο TTC εκτιμά τη συσχέτιση μεταξύ των πραγματικών μη παρατηρούμενων τιμών που χαρακτηρίζουν τη πιθανότητα διάγνωσης ως «μη φυσιολογικό» κάθε βαθμολογητή και βασίζεται στην υπόθεση ότι αυτές ακολουθούν τη διδιάστατη κανονική κατανομή. Ο TTC διαφέρει με τους τύπου kappa συντελεστές όχι μόνο ως προς τις περιπτώσεις όπου εφαρμόζεται, αλλά και ως προς το γεγονός ότι εκτιμούν ποσοτικά δυο διαφορετικές (παρότι συσχετιζόμενες) οντότητες όπως αναφέρει η Kraemer (1997).

Ο TTC προκύπτει ως ένας εκτιμητής μέγιστης πιθανοφάνειας του συντελεστή συσχέτισης για τη διδιάστατη κανονική κατανομή όταν ως διαθέσιμη πληροφορία έχουμε μόνο τα στοιχεία του πίνακα συσχέτισης [Tallies (1962), Hamdan (1970)]. Ο υπολογισμός της στηρίζεται στην επαναληπτική μέθοδο, χρησιμοποιώντας πίνακες για το δισδιάστατο κανονικό ολοκλήρωμα [Johnson and Kotz (1972)].

1.6 Scott's Pi Coefficient

Περίπου πέντε χρόνια προτού εκδοθεί ο kappa του Cohen, Scott (1955) πρότεινε τη χρήση ενός συντελεστή με το όνομα Pi ή στα ελληνικά *π*. Ο δείκτης αυτός βασίζεται στη ποσοστιαία συμφωνία με τύπο,

$$p_a = \frac{n_{11} + n_{22}}{n} \tag{1.12}$$

αν δεν έχουμε ελλείπουσες τιμές (missing values). Να σημειωθεί ότι n_{11} αντιστοιχεί στη κοινή βαθμολογία των δύο αξιολογητών για τη πρώτη κατηγορία και το n_{22} αντιστοιχεί στη κοινή βαθμολογία τους για τη δεύτερη κατηγορία.

Ενώ, αν υπάρχουν ελλείπουσες τιμές βασίζεται στο τύπο,

$$p_a = \frac{n_{11} + n_{22}}{n - (n_{+x} + n_{x+})} \tag{1.13}$$

όπου n_{+x} και n_{x+} δηλώνουν το πλήθος των παρατηρήσεων που δεν αξιολογήθηκαν από τον Α΄ και Β΄ βαθμολογητή αντίστοιχα.

Επιπροσθέτως, στον υπολογισμό του δείκτη του Scott, γίνεται χρήση του τύπου

$$p_e = \hat{\pi}_1^2 + (1 - \hat{\pi}_1)^2, \tag{1.14}$$

όπου $\widehat{\pi_1} = \frac{p_{1+} + p_{+1}}{2}$

και ορίζεται από τον Scott ως η συχνότητα με την οποία η κατηγορία 1 χρησιμοποιείται από τους ερευνητές. Η μέθοδός του αυτή, αμφισβητήθηκε από τον Cohen (1960) επειδή σιωπηρά υποθέτει μια μοναδική κλίση για ταξινόμηση σε μια συγκεκριμένη κατηγορία για όλους τους ερευνητές. Ο Scott όμως συνειδητά δεν έκανε μια τέτοια υπόθεση, καθώς τον ενδιέφερε περισσότερα η συχνότητα της χρήσης κάθε κατηγορίας από κάθε αξιολογητή.

Ο συντελεστής συμφωνίας του Scott στην επίσημη βιβλιογραφία ορίζεται ως εξής:

$$\widehat{\kappa_s} = \frac{p_a - p_e}{1 - p_e}, \quad \text{όπου } p_e = \widehat{\pi_1}^2 + (1 - \widehat{\pi_1})^2 \qquad (1.15)$$

Ας δώσουμε ένα αριθμητικό παράδειγμα χρήσης του παραπάνω τύπου. Για το σκοπό αυτό δίνεται ο παρακάτω πίνακας δεδομένων.

Πίνακας 3: Κατανομή 100 αντικειμένων από δύο αξιολογητές (Α' και Β') σε 2 κατηγορίες (1 & 2)

Α΄ Αξιολογητής	Β΄ Αξιολογητής		
	1	2	Σύνολο
1	n_{11} =35	n_{21} =20	n_{1+} =55
2	n_{21} =5	n_{22} =40	n_{2+} =45
Σύνολο	n_{+1} =40	n_{+2} =60	n =100

Κάνοντας χρήση του τύπου προκύπτει:

$$p_a = \frac{35+40}{100} = 0.75\ ,\ p_e = \left(\frac{\left(\left(\frac{55}{100}\right)+\left(\frac{40}{100}\right)\right)}{2}\right)^2 + \left(\frac{\left(\frac{45}{100}\right)+\left(\frac{60}{100}\right)}{2}\right)^2 = 0.5013 \text{ και}$$

$$\widehat{\kappa_s} = \frac{0.75 - 0.5013}{1 - 0.5013} = 0.4988$$

Ωστόσο, υπάρχουν πολλοί περιορισμοί στο μέτρο αυτό, που έχουν αιτιολογηθεί και τεκμηριωθεί πλήρως στη βιβλιογραφία και είναι εφάμιλλοι με εκείνους του μέτρου k του Cohen. Το μεγαλύτερο όμως πρόβλημα με το δείκτη π του Scott, είναι στον υπολογισμό του ποσοστού της τυχαίας συμφωνίας p_e. Το ερώτημα είναι αν η ποσότητα p_e, η οποία παρατηρούμε ότι αφαιρείται από τον υπολογισμό του δείκτη του Scott, περιγράφει όντως ένα φαινόμενο το οποίο έλαβε χώρα κατά τη διαδικασία της αξιολόγησης ή όχι. Είναι εύλογο και κατανοητό να θεωρήσουμε ότι κάποιοι αξιολογητές συμφώνησαν τυχαία, αλλά σίγουρα όχι σε όλα. Αυτό ήταν ένα από τα σημαντικά γεγονότα, που οδήγησαν πολλούς συγγραφείς να αμφιβάλλουν για την ποιότητα του συντελεστή του Scott.

1.7 Krippendorff's alpha

Ο Klaus Krippendorf (1970,2012) πρότεινε ένα συντελεστή συμφωνίας με την ονομασία a ("*alpha*"), ο οποίος χρησιμοποιείται από ερευνητές στο τομέα των επικοινωνιών. Για τη διαδικασία υπολογισμού του δείκτη βασίστηκε στις βασικές ιδέες του Cohen (1960,1968).

Το alpha του Krippendorff βασίζεται σε αντικείμενα που αξιολογούνται από δύο ή και περισσότερους αξιολογητές. Τα αντικείμενα που αξιολογούνται μόνο από έναν βαθμολογητή πρέπει να εξαλειφθούν πριν την έναρξη των απαραίτητων υπολογισμών. Θεωρούμε τη ποσότητα $\varepsilon_n = 1/(2n)$, όπου n είναι το πλήθος των αντικειμένων προς αξιολόγηση και από τους δύο βαθμολογητές. Ο τύπος υπολογισμού του δείκτη περιγράφεται παρακάτω:

$$\widehat{\alpha_K} = \frac{p'_a - p_e}{1 - p_e} \tag{1.16}$$

Όπου $p_e = \hat{\pi}_1^2 + (1 - \hat{\pi}_1)^2$, $p'_a = (1 - \varepsilon_n)p_a + \varepsilon_n$ και $p_a = \frac{n_{11}+n_{22}}{n}$

Να σημειωθεί ότι η ποσότητα ε_n θα είναι διαφορετική όταν εμπλέκονται 3 ή περισσότεροι αξιολογητές. Έχει ειπωθεί, ότι η ο δείκτης αυτός είναι ισοδύναμος με το συντελεστή π του Scott. Κάτι τέτοιο είναι αβάσιμο, καθώς η ισοδυναμία αυτή προκύπτει μόνο εάν στα δεδομένα δεν περιέχονται ελλείπουσες τιμές (missing values).

Η εξίσωση που περιγράφεται παραπάνω παραπέμπει στο ότι ο δείκτης του Scott και το Krippendorff's alpha είναι σχεδόν ταυτόσημοι, με τη μόνη διαφορά να βρίσκεται στη ποσοστιαία συμφωνία. Η εκδοχή του Krippendorff για τη ποσοστιαία συμφωνία οδηγεί σε ένα σταθμισμένο μέσο των παρατηρούμενων ποσοστιαίων συμφωνιών και η μεγαλύτερη τιμή του είναι 1, το οποίο είναι πάντα μεγαλύτερο από το παρατηρούμενο ποσοστό συμφωνίας.

1.8 Συντελεστής AC_1 του Gwet

Ο Gwet (2008) πρότεινε ένα συντελεστή συμφωνίας με το όνομα AC_1. Αναπτύχθηκε για να ξεπεράσει του πολλούς περιορισμούς του μέτρου k του Cohen. Ο δείκτης αυτός βασίζεται στο ίδιο ποσοστό συμφωνίας p_a (καθώς και με το τύπο όπου έχουμε ελλείπουσες τιμές) που συζητήθηκε σε προηγούμενη παράγραφο. Επίσης γίνεται χρήση ενός νέου ποσοστού τυχαίας συμφωνίας p_e.

Ο συντελεστής συμφωνίας του Gwet AC_1 ορίζεται στην επίσημη βιβλιογραφία ως εξής:

$$\widehat{\kappa_G} = \frac{p_a - p_e}{1 - p_e} \text{ όπου } p_e = 2\widehat{\pi_1}(1 - \widehat{\pi_1}) \tag{1.17}$$

Κάνοντας χρήση του Πίνακα 3. έχουμε τα εξής αποτελέσματα:

$$p_e = 2\left(\frac{\left(\left(\frac{55}{100}\right)+\left(\frac{40}{100}\right)\right)}{2}\right)\left(1-\left(\frac{\left(\left(\frac{55}{100}\right)+\left(\frac{40}{100}\right)\right)}{2}\right)\right)$$
$$= (0.75 - 0.49875)(1 - 0.49875) = 0.5012$$

Να σημειωθεί ότι το ποσοστό της τυχαίας συμφωνίας p_e στην πραγματικότητα υπολογίζεται ως $[\widehat{\pi_1}(1-\widehat{\pi_1}) + \widehat{\pi_2}(1-\widehat{\pi_2})]/(2-1)$ και λαμβάνει υπόψη το γεγονός ότι $[\widehat{\pi_1}(1-\widehat{\pi_1}) = \widehat{\pi_2}(1-\widehat{\pi_2})]$. Ο αριθμός 2 στον παρονομαστή δηλώνει το πλήθος των κατηγοριών.

Στο σημείο αυτό είναι απαραίτητο να αναλύσουμε το σκεπτικό του μέτρου που προτείνει ο Gwet. Σε δημοσίευση των (Grove et al., 1981) είχε αναφερθεί στο τι πραγματικά συμβαίνει στον ιατρικό χώρο σχετικά με τους ιατρούς και τον τρόπο με τον οποίο κάνουν τις διαγνώσεις σε ασθενείς. "*Προσδιορίζουν τις εύκολες περιπτώσεις ή αλλιώς τις περιπτώσεις που συμφωνούν απόλυτα απο τη θεωρία στα πραγματικά γεγονότα (textbook cases) σε διαγνώσεις με ελάχιστο ή καθόλου σφάλμα και αντιθέτως σε δύσκολες περιπτώσεις (non-textbook cases) τα αξιολογούν με πολύ ρίσκο ή ακόμα και εντελώς στην τύχη. Αν κάποιος ήξερε ποιες από τις αξιολογήσεις ανήκουν σε όποια από τις δύο κατηγορίες, τότε θα τις αντιμετώπιζε ξεχωριστά. Αλλά αυτό είναι ανέφικτο να συμβεί*." Συνεπώς, ο Gwet πιστεύει ότι πρέπει να γίνει μια πολύ καλή διάκριση ανάμεσα σε αυτές τις δύο περιπτώσεις παρά τη δυσκολία του εγχειρήματος αυτού. Στον υπολογισμό του συντελεστή kappa του Cohen ή στον *Pi* του Scott δεν ενσωματώνεται μια εκτίμηση των αβέβαιων περιπτώσεων που προαναφέρθηκαν.

Επιπροσθέτως, οι συντελεστές *kappa* και *Pi* βασίζονται στην ποσοστιαία τυχαία συμφωνία που είναι έγκυρη μόνο κάτω από την σχεδόν απίθανη προϋπόθεση ότι όλες οι αξιολογήσεις είναι ανεξάρτητες πριν την εκτέλεση του πειράματος. Κάτι τέτοιο στην πραγματικότητα είναι δύσκολο να συμβεί, αφού οι αξιολογητές συχνά αξιολογούν τα ίδια άτομα και αναμένεται να προκύψουν εξαρτημένες αξιολογήσεις με ίσως ελάχιστες εξαιρέσεις όταν είναι σε αμφιβολία.

Ο Gwet θεωρεί λοιπόν, ότι η ανεξαρτησία συμβαίνει όταν μια μη ντετερμινιστική [1]αξιολόγηση αποδίδεται σε κάποιο άτομο που είναι δύσκολο να το κατατάξει ο αξιολογητής. Αυτές οι μη ντετερμινιστικές αξιολογήσεις αποτελούν ένα μικρό κομμάτι των παρατηρημένων αξιολογήσεων. Ο δείκτης AC_1 προϋποθέτει ότι μόλις μια μερίδα των παρατηρημένων αξιολογήσεων θα οδηγήσει σε τυχαία συμφωνία.

[1] μη ντετερμινιστική ορίζεται η αξιολόγηση που δεν έχει εμφανή σύνδεση με τα χαρακτηριστικά του υποκειμένου που αξιολογείται.

Η ιδέα του μέτρου AC_1 για δύο αξιολογητές ξεκινάει από την παραδοχή ότι κάποια περιστατικά που έχουν αξιολογηθεί στην ίδια ή σε πανομοιότυπη κατηγορία έχουν εντοπιστεί και έχουν αφαιρεθεί από τον συνολικό πληθυσμό. Αυτή η διαδικασία δημιουργεί έναν νέο πληθυσμό απαλλαγμένο από αυτά τα υποκείμενα και συνεπώς είναι απίθανο να βρεθεί στον νέο αυτό πληθυσμό κάποια τυχαία συμφωνία.

Ο Gwet (Gwet,2008a), ορίζει τον συντελεστή AC_1 ως τη πιθανότητα ότι οι δύο αξιολογητές συμφωνούν με δεδομένο ότι τα υποκείμενα σίγουρα δεν έχουν συμφωνήσει αμιγώς τυχαία. Ο ορισμός αυτός έρχεται πολύ κοντά στον στόχο του Cohen (1960). Ο Cohen υποστήριζε ότι το μέτρο kappa αντιπροσωπεύει "*το ποσοστό της συμφωνίας αφού η τυχαία συμφωνία έχει αποκλειστεί.*" Η ονομασία του μέτρου AC_1 αποδίδεται στο A για τη λέξη "*Agreement*" , το C για τη λέξη "*Coefficient*" και ο αριθμός 1 ως υπο δείκτης ότι η μοναδική συμφωνία μεταξύ των δυο αξιολογητών που λαμβάνεται υπόψη προκύπτει από τα στοιχεία της κύριας διαγωνίου του πίνακα συνάφειας και μόνο. Στη συνέχεια της παρούσας εργασίας θα γίνει και μικρή αναφορά στον δείκτη AC_2, ο οποίος είναι ένα άλλο μέτρο το οποίο λαμβάνει υπόψη συμφωνίες «δεύτερου επιπέδου», κοινώς μερικές συμφωνίες.

Το στατιστικό AC_1

Βρισκόμαστε στη περίπτωση των δύο αξιολογητών σε ονομαστική κλίμακα q κατηγοριών. Ο τύπος υπολογισμού του δείκτη περιγράφεται παρακάτω ως εξής:

$$\widehat{\gamma_1} = \frac{p_a - p_e}{1 - p_e}, \text{ με } p_a = \sum_{k=1}^{q} p_{kk} \ , p_e = \frac{1}{q-1} \sum_{k=1}^{q} \pi_\kappa (1 - \pi_\kappa) \qquad (1.18)$$

όπου $\pi_\kappa = \frac{p_k + p_{+k}}{2}$ και p_{k+} και p_{+k} είναι σχετικές ποσότητες των υποκειμένων που αποδίδονται στην κατηγορία k από τους αξιολογητές A και B αντίστοιχα. Η ποσότητα p_{kk} αντιπροσωπεύει το σχετικό πλήθος των υποκειμένων που έχουν ταξινομηθεί στην κατηγορία k και από τους δύο βαθμολογητές. Η ποσότητα p_e αποτελείται από δύο γινόμενα. Το πρώτο γινόμενο αντιστοιχεί στην πιθανότητα συμφωνίας των δυο αξιολογητών όταν το υποκείμενο προς αξιολόγηση έχει ταξινομηθεί στην μη ντετερμινιστική περίπτωση. Η πιθανότητα αυτή είναι $\frac{1}{q}$, καθώς οι μη ντετερμινιστικές αξιολογήσεις θεωρούνται τυχαία και συνεπώς έχοντας ισοπίθανες επιλογές για τις q κατηγορίες. Το δεύτερο γινόμενο αφορά την τάση του αξιολογητή να προσδιορίσει μια μη ντετερμινιστική αξιολόγηση η οποία υπολογίζεται από το κλάσμα $\sum_{k=1}^{q} \pi_\kappa (1 - \pi_\kappa)/(1 - \frac{1}{q})$. Το ιδιαίτερο που θα πρέπει να κρατηθεί από αυτή την έκφραση είναι ότι μια κατανομή των υποκειμένων που είναι "λοξή" προς λίγες κατηγορίες αξιολόγησης, θα ελαττώσει τη τάση των μη ντετερμινιστικών αξιολογήσεων.

1.9 G-Index

Ο δείκτης *G* είναι η πιο απλή εκδοχή του συντελεστή συμφωνίας διορθωμένου από τυχαιότητα. Αρχικά προτάθηκε από τους Holley και Guilford (1964) και στη συνέχεια γενικεύτηκε σε τρείς ή περισσότερες κατηγορίες από τους Brennan και Prediger (1981). Βασίζεται και αυτός στα προαναφερθέντα ποσοστά συμφωνίας p_a (χωρίς και με ελλείπουσες τιμές). Ωστόσο, εδώ το ποσοστό τυχαίας συμφωνίας είναι απλά ίσο με $1/2$, όπου 2 δηλώνει τον αριθμό των κατηγοριών που χρησιμοποιούνται στη πειραματική διαδικασία.

Ο δείκτης G των Holley-Guilford ορίζεται στην επίσημη βιβλιογραφία ως εξής:

$$\widehat{\kappa_2} = \frac{p_a - 0.5}{1 - 0.5} \tag{1.19}$$

Από τα δεδομένα του Πίνακα 3. ο δείκτης $\widehat{\kappa_2}$ υπολογίζεται ως εξής:

$$\widehat{\kappa_2} = \frac{0.75 - 0.5}{1 - 0.5} = 0.5$$

1.10 Αριθμητικά αποτελέσματα από πραγματικά δεδομένα

Παρακάτω δίνεται ένας πίνακας που θα μας βοηθήσει στον υπολογισμό όλων των προαναφερθέντων μέτρων συμφωνίας.

Πίνακας 4: Αξιολογήσεις ασθενών με πόνους στη σπονδυλική στήλη από δύο αξιολογητές με τρεις κατηγορίες ταξινόμησης

	Ιατρός Β			
Ιατρός Α	Σύνδρομο Διαταραχής	Σύνδρομο Δυσλειτουργίας	Σύνδρομο στάσης σώματος	Σύνολο
Σύνδρομο Διαταραχής	22	10	2	34
Σύνδρομο Δυσλειτουργίας	6	27	11	44
Σύνδρομο στάσης σώματος	2	5	17	24
Σύνολο	30	42	30	102

Πίνακας 5: Ποσοστά αξιολογήσεων ασθενών με πόνους στη σπονδυλική στήλη από δύο αξιολογητές με τρεις κατηγορίες ταξινόμησης

	Ιατρός Β			
Ιατρός Α	Σύνδρομο Διαταραχής	Σύνδρομο Δυσλειτουργίας	Σύνδρομο στάσης σώματος	Σύνολο
Σύνδρομο Διαταραχής	0,2157	0,0935	0,0196	0,3288
Σύνδρομο Δυσλειτουργίας	0,0588	0,2647	0,1078	0,4313
Σύνδρομο στάσης σώματος	0,0196	0,049	0,1666	0,2352
Σύνολο	0,2941	0,4072	0,294	1

Ξεκινώντας τους υπολογισμούς έχουμε:

$$p_a = 0.2157 + 0.2647 + 0.1666 = 0.647$$

- Το *kappa του Cohen*.

$$p_e = 0.3288 \cdot 0.2941 + 0.4314 \cdot 0.4072 + 0.2352 \cdot 0.294 = 0.34152$$

$$\hat{k} = \frac{0.647 - 0.34152}{1 - 0.34152} = 0.4639$$

- Το *Pi του Scott*

  Αρχικά θα υπολογίσουμε τις ποσότητες $\widehat{\pi_1}, \widehat{\pi_2}$, και $\widehat{\pi_3}$.

  $\widehat{\pi_1} = \frac{0.3288+0.2941}{2} = 0.31145$, $\widehat{\pi_2} = \frac{0.4313+0.4072}{2} = 0.4192$, $\widehat{\pi_3} = \frac{0.2352+0.294}{2} = 0.2647$

  Έπειτα, $\mathrm{p_e} = 0.31145^2 + 0.4192^2 + 0.2647^2 = 0.3428$,

  Επομένως, $\widehat{\kappa_s} = \frac{0.647-0.3428}{1-0.3428} = 0.4629$

- Το AC_1 του Gwet.

  Κάνοντας χρήση των περιθώριων πιθανοτήτων $\widehat{\pi_1}, \widehat{\pi_2}$, και $\widehat{\pi_3}$ υπολογίζουμε την $\mathrm{p_e} = \frac{[0.31145\cdot(1-0.31145)+0.4192\cdot(1-0.4192)+0.2647\cdot(1-0.2647)]}{3-1} = 0.3262$,

  Επομένως, $\widehat{\kappa_G} = \frac{0.647-0.3262}{1-0.3262} = 0.4761$

- Το *Alpha* του Krippendorff

  Έχουμε $\varepsilon_n = \frac{1}{2n} = \frac{1}{2\cdot 102} = 0.004902$ και $p'_a = (1 - 0.004902) \cdot 0.647 + 0.004902 = 0.6487304$.

  Επομένως, $\widehat{\alpha_\kappa} = \frac{0.6487304-0.3428}{1-0.3428} = 0.4655$

Συγκεντρωτικά αποτελέσματα:

***Kappa* του Cohen**	***Pi* του Scott**	AC_1	***Alpha***
0.4639	0.4629	0.4761	0.4655

1.11 Ύπαρξη ελλειπουσών βαθμολογιών

Στη περίπτωση των δυο βαθμολογητών των n αντικειμένων προς αξιολόγηση, μπορεί να υπάρξουν περιπτώσεις όπου ο Α΄ βαθμολογητής να μην αξιολογήσει κάποιο αντικείμενο (ή αντίστροφα ο Β΄). Συνεπώς, σε έναν πίνακα αξιολόγησης θα υπάρχουν ελλείπουσες τιμές. Είναι απαραίτητο στο πίνακα να προστεθούν μια στήλη και μία γραμμή ονόματι «Χ» στην οποία θα αναγράφονται το πλήθος των αντικειμένων που δεν αξιολόγησε ο Α΄ και ο Β΄ βαθμολογητής αντίστοιχα. Θα έχει δηλαδή τη παρακάτω μορφή:

Πίνακας 6: Αξιολόγηση n αντικειμένων από δύο βαθμολογητές με τη προσθήκη της στήλης και γραμμής «Χ»

Αξιολογητής Α	**Αξιολογητής Β**					**Σύνολο**
	1	2	…	Q	X	
1	n_{11}	n_{12}	…	n_{1q}	n_{1X}	n_{1+}
2	n_{21}	n_{22}	…	n_{2q}	n_{2X}	n_{2+}
.					.	.
.			…		.	.
.					.	.
Q	n_{q1}	n_{q2}	…	n_{qq}	n_{qX}	n_{q+}
X	n_{X1}	n_{X2}	…	n_{Xq}	0	n_{X+}
Σύνολο	n_{+1}	n_{+2}	…	n_{+q}	n_{+X}	n

Για τον υπολογισμό των μέτρων συμφωνίας Percent Agreement, τα μέτρα των Cohen, Scott, και Krippendorff, θα γίνει ξανά χρήση των τύπων που έχουν αναφερθεί σε προηγούμενες παραγράφους. Η διαφορά έγκειται στον υπολογισμό των πιθανοτήτων που υπάρχουν σε αυτές τις εξισώσεις.

Για τον υπολογισμό του μέτρου ποσοστιαίας συμφωνίας p_a μεταξύ των αξιολογητών Α΄ και Β΄, θα πρέπει να αφαιρεθούν από τον υπολογισμό το άθροισμα των n_{+X} και n_{X+} που εκφράζουν το πλήθος των ελλειπουσών τιμών. Ο τύπος υπολογισμού διαμορφώνεται ως εξής:

$$p_a = \sum_{k=1}^{q} \frac{p_{kk}}{1 - (p_{+X} + p_{X+})} \qquad (1.10)$$

όπου $p_{+X} = \frac{n_{+X}}{n}$, $p_{X+} = \frac{n_{X+}}{n}$ και το κλάσμα $\frac{p_{kk}}{1-(p_{+X}+p_{X+})}$ εκφράζει το σχετικό πλήθος των αντικειμένων που αξιολογήθηκαν και από τις δύο αξιολογητές και μάλιστα στην ίδια κατηγορία *k*.

Παρακάτω, δίνονται δύο πίνακες αναφερόμενες στο ίδιο πρόβλημα "πόνοι στη σπονδυλικοί στήλη", με τη προσθήκη της στήλης και γραμμής με τίτλο «Χ», όπου αναφέρονται περιπτώσεις ασθενών που δεν έχουν αξιολογηθεί είτε από τον Α είτε από τον Β ιατρό αντίστοιχα.

Πίνακας 7: Αξιολογήσεις ασθενών με πόνους στη σπονδυλική στήλη από δύο αξιολογητές με προσθήκη της στήλης και γραμμής Χ

Ιατρός Α'	Ιατρός Β'				
	Σύνδρομο Διαταραχής	Σύνδρομο Δυσλειτουργίας	Σύνδρομο Στάσης Σώματος	Χ	Σύνολο
Σύνδρομο διαταραχής	22	10	2	3	37
Σύνδρομο Δυσλειτουργίας	6	27	11	2	46
Σύνδρομο Στάσης σώματος	2	5	17	3	27
Χ	3	1	6	0	10
Σύνολο	33	43	36	8	120

Πίνακας 8: Ποσοστά αξιολογήσεων ασθενών με πόνους στη σπονδυλική στήλη από δύο αξιολογητές με τη προσθήκη της στήλης και γραμμής Χ

Ιατρός Α'	Ιατρός Β'				Σύνολο1 (%)	Σύνολο2 (%)	Σύνολο3 (%)	Σύνολο! (%)
	Σύνδρομο Διαταραχής	Σύνδρομο Δυσλειτουργίας	Σύνδρομο Στάσης Σώματος	Χ				
Σύνδρομο Διαταραχής	0,183333	0,083333	0,016667	0,025	0,2833	0,33636	0,33333	0.308
Σύνδρομο Δυσλειτουργίας	0,05	0,225	0,091667	0,016667	0,3667	0,41818	0,43137	0.383
Σύνδρομο Στάσης Σώματος	0,016667	0,041667	0,141667	0,025	0,2	0,24545	0,23529	0.225

X	0,025	0,008333	0,05		*0,0833*	-	-	-
Σύνολο4(%)	0,250	0,350	0,250	*0,067*	1,0	1,0	1,0	
Σύνολο5(%)	0,295	0,384	0,321	-	1,0			
Σύνολο6(%)	0,294	0,412	0,294	-	1,0			
Σύνολο!(%)	0.275	0.3583	0.3	-				

Η ποσοστιαία συμφωνίας p_a υπολογίζεται τώρα ως εξής:

$$p_a = (0.1833 + 0.2250 + 0.1417)/(1 - 0.0667 -$$

$$0.0833)) = 0.6471$$

- Το kappa του Cohen

  Υπολογίζουμε το $p_e = 0.275 \cdot 0.308 + 0.358 \cdot 0.383 + 0.300 \cdot 0.225 + 0.067 \cdot 0.083 = 0.295$

  Επομένως, $\widehat{\kappa_C} = \frac{0.6471-0.295}{1-0.295} = 0.4994$

- Το Pi του Scott

  Αρχικά υπολογίζουμε τις ποσότητες $\widehat{\pi_1}, \widehat{\pi_2}, \text{και } \widehat{\pi_3}$. Σύμφωνα με τον τύπο υπολογισμού των $\widehat{\pi_\iota}$ χρησιμοποιούμε το μέσο όρο του στοιχείου του Συνόλου2 και του Συνόλου5.

  Έχουμε $\widehat{\pi_1} = \frac{0.33636+0.295}{2} = 0.3155$, ομοίως $\widehat{\pi_2} = 0.4011$ και $\widehat{\pi_3} = 0.2834$

  Συνεπώς, $p_e = 0.3155^2 + 0.4011^2 + 0.2834^2 = 0.3407$.

  Επομένως, $\widehat{\kappa_S} = \frac{0.6471-0.3407}{1-0.3407} = 0.4647$

- Το AC_1 του Gwet

  Κάνοντας χρήση των $\widehat{\pi_1}, \widehat{\pi_2}, \text{και } \widehat{\pi_3}$ που υπολογίστηκαν στον συντελεστή του Scott έχουμε $p_e = \frac{[0.3155 \cdot (1-0.3155) + 0.4011 \cdot (1-0.4011) + 0.2834 \cdot (1-0.2834)]}{3-1} = 0.3296$

  Επομένως, $\widehat{\kappa_G} = \frac{0.6471-0.3296}{1-0.3296} = 0.4735$

- Το *Alpha* του Krippendorff

  Σε αυτή τη περίπτωση οι ποσότητες $\widehat{\pi_1}, \widehat{\pi_2}, \text{και } \widehat{\pi_3}$ διαφέρουν σε σχέση με τις προηγούμενες περιπτώσεις. Εδώ, θα χρησιμοποιηθούν μόνο τα υποκείμενα που αξιολογήθηκαν και από τους δύο αξιολογητές. Για παράδειγμα για τον υπολογισμό της ποσότητας $\widehat{\pi_1}$ θα βρούμε το μέσο όρο του στοιχείου της στήλης Σύνολο3 και της γραμμής Σύνολο6.

  Συνεπώς, έχουμε $\widehat{\pi_1} = \frac{0.333+0.294}{2} = 0.3137$, ομοίως $\widehat{\pi_2} = 0.4216$ και $\widehat{\pi_3} = 0.2647$

  Αντίστοιχα, $p_e = 0.3137^2 + 0.4216^2 + 0.2647^2 = 0.3462$ και

  $$p'_a = 0.6488$$

Επομένως,

$$\hat{\alpha}_{\kappa} = \frac{0.6488 - 0.3462}{1 - 0.3462} = 0.4628$$

Κεφάλαιο 2:
Μέτρα συμφωνίας αξιολογητών με τρεις ή περισσότερους βαθμολογητές

2.1 Εισαγωγή

Στο κεφάλαιο αυτό θα σχολιαστούν μέτρα που αφορούν τρείς ή περισσότερους αξιολογητές σε ονομαστικά δεδομένα. Πιο συγκεκριμένα, τα μέτρα που μελετώνται είναι το γενικευμένο μέτρο kappa του Fleiss (1971), του Conger (1980), του Gwet (2008a) και το alpha του Krippendorff (1970, 1978, 2004). Για την ανάλυσή τους θα χρησιμοποιηθεί ένα παράδειγμα με πραγματικά δεδομένα [Rowland (1984)]. Στο παράδειγμα δεν υπάρχουν missing data.

2.2 Εφαρμογή σε παράδειγμα με πραγματικά δεδομένα

Στο σημείο αυτό είναι απαραίτητο να δούμε μέσω ενός παραδείγματος τον υπολογισμό των παρακάτω μέτρων:

A. Fleiss' Generalized Kappa Coefficient
B. Conger's Generalized Kappa Coefficient
Γ. Gwet's AC_1 Coefficient
Δ. Krippendorff's Alpha Coefficient

Ο παρακάτω πίνακας μας δείχνει τα δεδομένα από μια πειραματική έρευνα με σκοπό την ανίχνευση αλλαγών χρωματισμού πάνω σε μια ομάδα ψαριών ονόματι "Stickleback" [Rowland (1984)].

Πίνακας 9: Κατανομή των ψαριών "Stickleback" κατά βαθμολογητή και κατηγορία χρωματισμού

	Χρώμα					
Βαθμολογητής	1	2	3	4	5	Σύνολο
1	10	0	11	1	7	29
2	10	2	11	1	5	29
3	10	1	9	3	6	29
4	12	0	6	3	8	29
Μέσος Όρος	1.448	0.103	1.276	0.276	0.897	29

Πίνακας 10: Κατανομή των βαθμολογητών κατά ψάρι και κατηγορία χρωματισμού

			Χρώμα			
Ψάρι	**1**	**2**	**3**	**4**	**5**	**Σύνολο**
1	0	0	0	0	4	4
2	2	0	2	0	0	4
3	0	0	0	0	4	4
4	2	0	2	0	0	4
5	0	0	0	1	3	4
6	1	1	2	0	0	4
7	3	0	1	0	0	4
8	3	0	1	0	0	4
9	0	0	2	2	0	4
10	3	0	1	0	0	4
11	0	0	0	0	4	4
12	4	0	0	0	0	4
13	4	0	0	0	0	4
14	4	0	0	0	0	4
15	0	0	3	1	0	4
16	1	0	2	1	0	4
17	0	0	0	2	2	4
18	0	0	0	0	4	4
19	0	0	3	0	1	4
20	0	1	3	0	0	4
21	0	0	1	0	3	4
22	0	0	3	1	0	4
23	4	0	0	0	0	4
24	4	0	0	0	0	4
25	2	0	2	0	0	4
26	1	0	3	0	0	4
27	2	0	2	0	0	4
28	2	0	2	0	0	4
29	0	1	2	0	1	4

2.2.1 Fleiss' Generalized Kappa Coefficient

Ο Fleiss (1971) θεωρεί ένα σενάριο όπου ένας βαθμολογητής που επιλέγεται τυχαία (με επανατοποθέτηση) από μια δεξαμενή από r σε πλήθος βαθμολογητές, πρέπει να αξιολογήσει τυχαία ένα υποκείμενο που επιλέγεται από μια δεξαμενή n σε πλήθος υποκειμένων. Η πιθανότητα επιλογής ενός υποκειμένου και ενός βαθμολογητή που το ταξινομεί σε μια από τις k κατηγορίες συμβολίζεται με π_k και υπολογίζεται από τον Fleiss (1971) ως εξής:

$$\widehat{\pi_k} = \frac{1}{n}\sum_{i=1}^{n}\frac{r_{ik}}{r_i} \tag{2.1}$$

Θα πρέπει να σημειωθεί ότι τα υποκείμενα που αξιολογούνται μόνο από έναν βαθμολογητή συμπεριλαμβάνονται στον υπολογισμό του $\widehat{\pi_k}$. Επειδή ο Fleiss ορίζοντας την ποσοστιαία συμφωνία, δε λαμβάνει υπόψη τις ελλείπουσες τιμές, εδώ θα παρουσιαστεί ένας συντελεστής kappa που το πετυχαίνει αυτό. Γενικότερα, ο Fleiss ορίζει ως ποσοστιαία συμφωνία τη πιθανότητα πως κάθε ζευγάρι αξιολογητών ταξινομεί ένα υποκείμενο στην ίδια ακριβώς κατηγορία. Ο τύπο υπολογισμού είναι:

$$\widehat{\kappa_F} = \frac{p_a - p_e}{1 - p_e}, \text{όπου } p_e = \sum_{k=1}^{q}\pi_k^2 \tag{2.2}$$

Κάνοντας χρήση του Πίνακα 9 υπολογίζουμε τα εξής:

$$\widehat{\pi_1} = \frac{1}{29}\left(\frac{1}{4}+\frac{2}{4}+\frac{0}{4}+\cdots+\frac{2}{4}+\frac{0}{4}\right) = 0.36209$$

Ομοίως προκύπτει: $\widehat{\pi_2} = 0.025862$, $\widehat{\pi_3} = 0.318966$, $\widehat{\pi_4} = 0.068966$, $\widehat{\pi_5} = 0.224138$

Επίσης, $p_e = \widehat{\pi_1}^2 + \widehat{\pi_2}^2 + \cdots + \widehat{\pi_5}^2 = 0.288496$

Για τον υπολογισμό του p_a προκύπτει ότι :

$$p_a = 0.58046 \ (\theta\alpha\ \chi\rho\eta\sigma\iota\mu o\pi o\iota\eta\theta\epsilon\acute{\iota}\ \kappa\alpha\iota\ \sigma\tau o\upsilon\varsigma\ \upsilon\pi\acute{o}\lambda o\iota\pi o\upsilon\ \sigma\upsilon\nu\tau\epsilon\lambda\epsilon\sigma\tau\acute{\epsilon}\varsigma)$$

Συνεπώς,

$$\widehat{\kappa_F} = \frac{0.58046 - 0.288496}{1 - 0.288496} = 0.410347$$

2.2.2 Conger's Generalized Kappa Coefficient

Για τον υπολογισμό του μέτρου του Conger θα χρειαστούμε τον Πίνακα 10, από τον οποίο προκύπτει ο επόμενος πίνακας.

Πίνακας 11: Ποσοστά των υποκειμένων που ταξινόμησαν οι αξιολογητές στην κάθε κατηγορία χρωματισμού

Βαθμολογητής	1	2	3	4	5	Σύνολο
1	0,344828	0	0,37931	0,034483	0,241379	
2	0,344828	0,068966	0,37931	0,034483	0,172414	
3	0,344828	0,034483	0,310345	0,103448	0,206897	
4	0,413793	0	0,206897	0,103448	0,275862	
$\overline{p_{.k}}$	0,362069	0,025862	0,318966	0,068966	0,224138	
$\overline{p_{.k}}^2$	0,131094	0,000669	0,101739	0,004756	0,050238	0,288496

Στα κελιά του παραπάνω πίνακα έχουν υπολογιστεί οι πιθανότητες $p_{gk} = n_{gk}/n_g$. Για τον υπολογισμό τη ποσοστιαίας τυχαίας συμφωνίας p_e θα χρειαστούμε τη δειγματική διακύμανση s_k^2 των r ποσοστών $p_{1k}, \ldots, p_{rk}$. Ο τύπος υπολογισμού της διακύμανσης είναι:

$$s_k^2 = \frac{1}{r-1}\sum_{g=1}^{r}\left(p_{gk} - \overline{p_{.k}}\right)^2 \qquad (2.3)$$

όπου $\overline{p_{.k}}$ είναι η μέση τιμή των πιθανοτήτων $p_{1k}, \ldots, p_{rk}$.

Έπειτα από πράξεις υπολογίζουμε:

$s_1^2 = 0.001189\,, s_2^2 = 0.00109, s_3^2 = 0.006639, s_4^2 = 0.001585\,, s_5^2 = 0.001982$

Η ποσοστιαία τυχαία συμφωνία θα βρεθεί από τον τύπο:

$$p_e = \sum_{k=1}^{q}\overline{p_{.k}} - \sum_{k=1}^{q} s_k^2/r \qquad (2.4)$$

Έπειτα από πράξεις προκύπτει $p_e = 0.295375$

Ο τύπος υπολογισμού του μέτρου του Conger είναι:

$$\widehat{\kappa_C} = \frac{p_a - p_e}{1 - p_e}, \qquad (2.5)$$

όπου $p_a = 0.58046$ (όπως και στο μέτρο του Fleiss). Συνεπώς, προκύπτει ότι

$\widehat{\kappa_C} = 0.412923$

2.2.3 Gwet' AC_1 Coefficient

Ο συντελεστής αυτός βασίζεται στο ποσοστό συμφωνίας p_a που έχει ήδη χρησιμοποιηθεί στους συντελεστές των Fleiss και Conger. Όσων αφορά το ποσοστό τυχαίας συμφωνίας υπολογίζεται ως το γινόμενο της πιθανότητας συμφωνίας όταν η αξιολόγηση έχει γίνει τυχαία με τη πιθανότητα επιλογής ενός δύσκολου προς αξιολόγηση ("hard-to-score[2]") υποκείμενο.

[2] Είναι εκείνα τα υποκείμενα που αξιολογούνται τυχαία και οποιαδήποτε συμφωνία οφείλεται καθαρά στην τύχη.

Ο τύπος υπολογισμού του μέτρου του Gwet έχει ως εξής:

$$\widehat{\kappa_G} = \frac{p_a - p_e}{1 - p_e}, \qquad \text{όπου } p_e = \frac{1}{q(q-1)} \sum_{k=1}^{q} \widehat{\pi_k}(1 - \widehat{\pi_k}) \tag{2.6}$$

Να σημειωθεί ότι ο υπολογισμός του $\widehat{\pi_k}$ έχει ήδη αναφερθεί σε προηγούμενη ενότητα.

Με χρήση του Πίνακα 10, προκύπτουν τα παρακάτω αποτελέσματα.

$\widehat{\pi_1} = 0.362069$, $\widehat{\pi_2} = 0.025862$, $\widehat{\pi_3} = 0.318966$, $\widehat{\pi_4} = 0.068966$, $\widehat{\pi_5} = 0.224138$

Επιπλέον $p_e = 0.035575$ και τελικά $\widehat{\kappa_G} = 0.564984$

2.2.4 Krippendforff's Alpha Coefficient

Όπως έχει ήδη περιγραφεί σε προηγούμενη παράγραφο ο συντελεστής που προτείνει ο Krippendorff (1970,1978,2004), βασίζεται στη αξιολόγηση των υποκειμένων από δύο ή περισσότερους αξιολογητές. Τα υποκείμενα που αξιολογούνται από έναν μόνο βαθμολογητή αφαιρούνται από τους υπολογισμούς. Το ιδιαίτερο με αυτό το μέτρο είναι ο υπολογισμός της ποσοστιαίας συμφωνίας p_a. Η ποσοστιαία συμφωνία στη συγκεκριμένη περίπτωση συμβολίζεται ως p'_a και υπολογίζεται ως εξής:

$$p'_a = (1 - \varepsilon_n)p_a + \varepsilon_n \tag{2.7}$$

όπου $\varepsilon_n = 1/(n'\bar{r})$ με n' το πλήθος των υποκειμένων που αξιολογήθηκαν από τουλάχιστον δύο βαθμολογητές και $\bar{r}$ το μέσο πλήθος των βαθμολογητών που αξιολόγησαν ένα υποκείμενο. Ο συντελεστής alpha του Krippendorff υπολογίζεται ως:

$$\widehat{\alpha_k} = \frac{p'_a - p_e}{1 - p_e} \;\; \text{όπου } p_e = \sum_{k=1}^{q} \widehat{\pi_k} \tag{2.8}$$

με $\widehat{\pi_k} = \frac{1}{n'} \sum_{k=1}^{n'} r_{ik}/\bar{r}$.

Το μέτρο του Krippendorff είναι αρκετά κοντά με το γενικευμένο μέτρο kappa του Fleiss γι' αυτό και οι τιμές τους είναι αρκετά κοντά. Αυτό συμβαίνει ιδιαίτερα στην περίπτωση που δεν έχουμε ελλείπουσες τιμές και το πλήθος των βαθμολογητών ξεπερνά τους πέντε.

Με τη βοήθεια του Πίνακα 10 υπολογίζουμε τα εξής:

$\bar{r} = 4$, $p_e = 0.288496$ (ακριβώς ίδια τιμή με το γενικευμένο Kappa του Fleiss) και $p'_a = 0.584076$. Τελικά,

$$\widehat{\alpha_k} = 0.415431$$

2.3 Συμπεράσματα

Συγκεντρωτικά τα αποτελέσματα που προέκυψαν από τους τέσσερις ανωτέρω δείκτες συμφωνίας που συζητήθηκαν προκύπτει ο τελικός Πίνακας 12.

Πίνακας 12: Συγκεντρωτικά αποτελέσματα δεικτών

Fleiss	Conger	Gwet	Krippendorff
0.410347	0.412923	0.564984	0.415431

Καθώς φαίνεται οι δείκτες των Fleiss, Conger και Krippendorff διαφέρουν αμελητέα. Παρατηρείται μεγάλη διαφορά στο δείκτη του Gwet. Σε κάθε περίπτωση η συμφωνία των αξιολογητών μπορεί να χαρακτηριστεί ως μέτρια.

Κεφάλαιο 3ο : Μη παραμετρικό τεστ αξιολόγησης συμφωνίας

3.1 Sequential test

Στη κεφάλαιο αυτό προτείνεται μια μη παραμετρική διαδοχική δοκιμή για αξιολόγηση συμφωνίας μεταξύ βαθμολογητών από τους Μπερσίμης, Σαχλάς και Chakraborti (2017). Οι υπάρχοντες δείκτες συμφωνίας για τα κατηγορικά δεδομένα απαιτούν πίνακες έκτακτης ανάγκης, σε αντίθεση με τη πρόταση των Μπερσίμης κ.α. Το γεγονός αυτό βοηθά στην εξοικονόμηση χρόνου, κόπου, ειδικότερα όταν ενδιαφερόμαστε για μια γρήγορη αξιολόγηση της συμφωνίας.

Το τεστ αυτό βασίζεται στο άθροισμα του πλήθους των διαφωνιών μεταξύ δύο βαθμολογητών καθώς και ένα στατιστικό μέτρο που αντιπροσωπεύει το χρόνο αναμονής μέχρι το άθροισμα του πλήθους των διαφωνιών να ξεπεράσει ένα κατώφλι που ορίζει ο ερευνητής. Θα παρουσιαστούν περιπτώσεις δύο βαθμολογητών ως προς ένα χαρακτηριστικό προς αξιολόγηση σε δύο κατηγορίες, η περίπτωση δύο βαθμολογητών ως προς ένα χαρακτηριστικό σε περισσότερες από δύο κατηγορίες αξιολόγησης, η περίπτωση μεγάλης διαφωνίας μεταξύ των δύο αξιολογητών και τέλος η περίπτωση με περισσότερους από τρείς αξιολογητές.

Πιο αναλυτικά, η μηδενική υπόθεση του τεστ είναι:

$$H_0: \pi = \sum_i \pi_{ii} = \pi_0$$

με $i = 1,2, \dots, r$ και π είναι η πιθανότητα της συμφωνίας μεταξύ των βαθμολογητών. Η πιθανότητα αυτή θέλουμε να ισούται με ένα προκαθορισμένο π_0.

Η εναλλακτική υπόθεση είναι:

$$H_1: \pi < \pi_0$$

Αυτό σημαίνει ότι αξιολογείται ακολουθητέα η παρατηρούμενη συμφωνία , όπου σε κάθε χρονική στιγμή εξετάζεται αν έχει μειωθεί σε σχέση με την προκαθορισμένη π_0. Είναι αναμενόμενο ότι εύκολες υποθέσεις αξιολόγησης θα ακολουθούνται από ποσοστά συμφωνίας υψηλότερα. Ωστόσο, στην πραγματικότητα δε θα είναι όλες οι περιπτώσεις εύκολες. Στην εναλλακτική υπόθεση δεν επιλέχτηκε το σύμβολο ≠ , καθώς στόχος είναι να ελέγχουμε αν η παρατηρούμενη συμφωνία είναι μικρότερη από το επιθυμητό.

Ας συμβολίσουμε με Y_t^l την αξιολόγηση του l βαθμολογητή στον χρόνο t. Δηλαδή,

$$Y_t^l = \begin{cases} 1, & \text{αν ο βαθμολογητής ταξινομήσει το υποκείμενο στην κατηγορία } 1 \\ 2, & \text{αν ο βαθμολορητής ταξινομήσει το υποκείμενο στην κατηγορία } 2 \end{cases}$$

όπου $l = A, B$ και $t = 1,2, \dots$

Οι αντίστοιχες πιθανότητες θα τις συμβολίσουμε ως εξής:

$$p_{ij} = P(Y_t^A = i\ , Y_t^B = j)\ , \qquad i,j = 1,2$$

Για την αξιολόγηση της συμφωνίας θα χρησιμοποιήσουμε τη ποσότητα

$$AD = \sum_{t=1}^{t_0} |Y_t^A - Y_t^B| \tag{3.1}$$

η οποία περιγράφει το άθροισμα των απολύτων τιμών των διαφορών των αξιολογήσεων από δύο αξιολογητές. Επίσης, ορίζουμε μια τυχαία μεταβλητή T_{AD} που μετρά το πλήθος των αξιολογήσεων μέχρι η ποσότητα AD να ξεπεράσει ένα κατώφλι f για πρώτη φορά. Θα απορρίψουμε τη μηδενική υπόθεση με σφάλμα τύπου I, $\alpha = P(T_{AD} \leq \kappa | f, H_0)$, όπου $t_0 \leq \kappa$.

Αρχικά, επιλέγουμε το επιθυμητό a σφάλμα τύπου I. Έπειτα, υπολογίζουμε μέσω της εξίσωσης $\alpha = P(T_{AD} \leq \kappa | f, H_0)$ μερικές τιμές για το c για κάποιες τιμές των f. Η επιλογή του f σχετίζεται με το επιθυμητό επίπεδο ισχύς γ, για την εναλλακτική υπόθεση και αυτό διορθώνει τη τιμή του κ. Στη συνέχεια μέσω πίνακα θα παρουσιαστούν δύο περιπτώσεις. Η πρώτη θα αφορά αξιολογήσεις μεταξύ βαθμολογητών που συμφωνούν στις περισσότερες περιπτώσεις και στη δεύτερη περίπτωση αφορά παράδειγμα συχνών διαφωνιών. Αυτό έχει ως αποτέλεσμα, στην πρώτη περίπτωση να έχουμε μικρές τιμές για την ποσότητα AD, ενώ στη δεύτερη περίπτωση μεγάλες. Είναι ευνόητο ότι στην περίπτωση απόλυτη συμφωνίας η τιμή AD θα πλησιάζει το 0 ενώ σε περίπτωση συχνών διαφωνιών η τιμή AD θα ξεπεράσει πολύ γρήγορα το κατώφλι f. Αν συμβεί η δεύτερη περίπτωση το αποτέλεσμα θα είναι σημαντική απόκλιση από τη μηδενική υπόθεση.

Για την προσδιορισμό της ακριβής κατανομής T_{AD} που αναφέρθηκε προηγουμένως, χρησιμοποιείται η τεχνική των Μαρκοβιανών αλυσίδων. Η τεχνική αυτή χρησιμοποιείται κατά κόρον τη σημερινή εποχή, για την εύρεση κατανομών από πολλά είδη στατιστικών πάνω σε ακολουθίες διακριτών τυχαίων μεταβλητών. Η τεχνική αυτή έχει χρησιμοποιηθεί από τους Κούτρας, Μπερσίμης και Αντζουλάκος (2008) και από πολλούς άλλους ερευνητές. Λόγω της πολυπλοκότητας του αλγορίθμου δε θα αναφερθούμε εκτενώς στα παρακάτω παραδείγματα. Θα σχολιαστούν μόνο τα αποτελέσματα που προκύπτουν από τη δημοσίευση των Μπερσίμης κ.α (2017).

3.2 Συμφωνία μεταξύ δυο βαθμολογητών με βάση ένα χαρακτηριστικό και ως προς δυο ή περισσότερες κατηγορίες ταξινόμησης

Ο παρακάτω πίνακας μας περιγράφει την αξιολόγηση δύο ακτινολόγων που ταξινομούν κάθε μαστογραφία ως προς ένα χαρακτηριστικό σε τρείς κατηγορίες ταξινόμησης σε 19 διαφορετικούς χρόνους.

Πίνακας 13: Αθροίσματα απολύτων διαφορών των αξιολογήσεων για έλεγχο της συμφωνίας δύο αξιολογητών ως προς ένα χαρακτηριστικό

Αξιολόγηση (t)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
Y_t^A	1	2	3	3	1	3	1	3	3	1	3	1	2	3	1	2	1	1	1
Y_t^B	1	3	3	1	1	3	1	3	2	2	1	2	2	3	1	1	1	1	1
$\lvert Y_t^A - Y_t^B \rvert$	0	1	0	2	0	0	0	0	1	1	2	1	0	0	0	1	0	0	0
AD	0	1	1	3	3	3	3	3	4	5	7	8	8	8	8	9	9	9	9

Το στατιστικό που θα χρησιμοποιηθεί για την συμφωνίας των δύο ακτινολόγων είναι το AD. Αν το AD ξεπεράσει το κατώφλι f που έχουμε ορίσει, έστω πριν την $\kappa - οστή$ αξιολόγηση, τότε θα υπάρχει σημαντική διαφωνία μεταξύ των δυο ακτινολόγων. Να σημειωθεί, ότι οι αξιολογητές αυτοί επιλέγονται με τυχαίο τρόπο από πληθυσμό με ισοδύναμα έμπειρους ακτινολόγους, ώστε να εξασφαλίσουμε την εγκυρότητα της σύγκρισης.

3.3 Περίπτωση ισχυρής διαφωνίας

Σε αυτή την υποενότητα θα σχολιαστεί η περίπτωση που οι δύο βαθμολογητές έχουν μεγάλη απόκλιση ως προς τις αξιολογήσεις τους. Συχνές διαφωνίες μπορεί να υπάρξουν σε κλίμακα Likert (1932) όπου οι κατηγορίες ταξινόμησης φτάνουν τις 7. Στην περίπτωση ακραίας διαφωνίας, δηλαδή ο ένας να ταξινομήσει το υποκείμενο στην πρώτη κατηγορία και ο άλλος στην έβδομη, τότε προτείνεται αυτόματα να σταματάει η διαδικασία της σύγκρισης, να απορρίπτεται η H_0.

Οι συμβολισμοί των πιθανοτήτων που χρειάζονται είναι ίδιοι με αυτούς της προηγούμενης υποενότητας. Προστίθεται ωστόσο, η πιθανότητα της ακραίας διαφωνίας. Ως ακραία διαφωνίας ορίζεται ως η περίπτωση όπου $|i - j| \geq \xi$, όπου ξ είναι ένας ακέραιος αριθμός που καθορίζεται από τον ερευνητή. Συνεπώς, οι περιπτώσεις όπου δεχόμαστε τη διαφωνία δεν είναι μόνο η περίπτωση όπου το AD ξεπεράσει το κατώφλι f, αλλά και η περίπτωση της ακραίας διαφωνίας.

3.4 Εξετάζοντας τη συμφωνία ως προς δύο χαρακτηριστικά

Πολλές φορές οι βαθμολογητές χρειάζεται να ταξινομήσουν ένα υποκείμενο σε κάποια κατηγορία ως προς δύο εξαρτώμενα χαρακτηριστικά. Για παράδειγμα, στη περίπτωση των ακτινολόγων χρειάζεται να ταξινομήσουν έναν όγκο ως προς το σχήμα τους και ως προς τον όγκο τους. Η αξιολόγηση ορίζεται ως:

$$Y_{i,t}^l = \begin{cases} 1, & μαστογραφία\ ως\ κανονική\ ως\ προς\ το\ χαρακτηριστικό\ i \\ 2, & μαστογραφία\ ως\ μη\ κανονική\ ως\ προς\ το\ χαρακτηριστικό\ i \end{cases},$$

όπου $i = 1,2\ ,\ l = A, B$

Η αλγεβρική έκφραση του στατιστικού AD θα είναι:

$$AD = \sum_{i=1}^{t_0} \left(\left| Y_{1,t}^{A} - Y_{1,t}^{B} \right| + \left| Y_{2,t}^{A} - Y_{2,t}^{B} \right| \right) \tag{3.3}$$

Οι τιμές που μπορεί να πάρει η ποσότητα $\left| Y_{1,t}^{A} - Y_{1,t}^{B} \right| + \left| Y_{2,t}^{A} - Y_{2,t}^{B} \right|$ είναι 0,1 ή 2. Όπως και στις προηγούμενες περιπτώσεις εξετάζουμε αν το AD ξεπεράσει το κατώφλι που έχουμε εξ' αρχής ορίσει.

3.5 Εξέταση συμφωνίας αξιολογητών ως προς ένα χαρακτηριστικό σε δύο κατηγορίες ταξινόμησης

Στη περίπτωση αυτή θα ορίσουμε τις εξής τυχαίες μεταβλητές:

$$AD_{1,t} = |Y_t^A - Y_t^B|\ ,\ \ AD_{2,t} = |Y_t^A - Y_t^C|\ \ ,\ \ AD_{3,t} = |Y_t^B - Y_t^C|$$

Οι μεταβλητές αυτές παίρνουν τις τιμές 0 ή 1. Αυτή τη φορά ορίζεται το στατιστικό,

$$AD = \sum_{i=1}^{t_0} \max\left\{ AD_{1,t}\ , AD_{2,t}\ , AD_{3,t} \right\} \tag{3.4}$$

το οποίο παίρνει τιμές από το σύνολο των ακεραίων αριθμών. Όπως και στις προηγούμενες περιπτώσεις, αν το AD ξεπεράσει το κατώφλι f πριν την $\kappa - οστή$ αξιολόγηση, τότε θα λέμε ότι υπάρχει απόκλιση από την H_0.

3.5 Σύγκριση του sequential test με το μέτρο k του Cohen και συμπεράσματα

Στη δημοσίευση των Μπερσίμης κ.α (2017) πραγματοποιείται σύγκριση του προτεινόμενου sequential test με το δείκτη k του Cohen (1960). Η μελέτη του Cantor (1996) , στην περίπτωση των δύο βαθμολογητών και δύο κατηγορίες ταξινόμησης, έδειξε ότι για τον έλεγχο της μηδενικής υπόθεσης $H_0: \kappa = 0.700$ και $H_1: \kappa = 0.500$ με $\alpha = 5\%$, το μέγεθος του δείγματος που απαιτείται είναι 214 ασθενείς με ισχύ ελέγχου τουλάχιστον 80%. Στη περίπτωση του ακολουθιακού τεστ με $H_0: p_{11} = 0.500\ , p_{12} = 0.072\ , p_{21} = 0.075\ , p_{22} = 0.353$ (το οποίο ισοδυναμεί με $\kappa = 0.700$) , επιλέχθηκε $f = 19$, που είναι το μικρότερο το οποίο πετυχαίνει ισχύ τουλάχιστον 80% ($\gamma = 0.814$). Το μέγιστο πλήθος των ασθενών που αξιολογούνται είναι 93 το οποίο αντιστοιχεί σε $\alpha = 0.049$, ενώ το αναμενόμενο πλήθος αξιολογήσεων είναι 80. Συνεπώς, το προτεινόμενο τεστ προαπαιτεί μικρότερο δείγμα για να επιτύχει ισοδύναμο επίπεδο ισχύς. Θα πρέπει όμως να σημειωθεί, ότι η σύγκριση των δύο μεθόδων είναι έγκυρη μόνο στην απλή περίπτωση (δηλαδή ως αξιολόγηση ως προς ένα χαρακτηριστικό και δύο κατηγορίες ταξινόμησης).

Συμπερασματικά, θα λέγαμε πως το ακολουθιακό τεστ υπερτερεί έναντι των κλασσικών μέτρων που προτείνονται στη βιβλιογραφία. Αρχικά, δε χρειάζεται να περιμένουμε για

μεγάλο χρονικό διάστημα για να συλλέξουμε δεδομένα. Επιπροσθέτως, το τεστ βασίζεται σε ένα απλό στατιστικό που είναι το άθροισμα των απόλυτων διαφορών των διαφωνιών των δυο βαθμολογητών. Το τρίτο πλεονέκτημα είναι ότι ανιχνεύονται ακόμα και μικρές αποκλίσεις από τη συμφωνία με μεγάλη ισχύ του ελέγχου. Τέλος, το προτεινόμενο τεστ γενικεύεται στη περίπτωση με περισσότερους αξιολογητές και ως προς περισσότερα από ένα χαρακτηριστικά προς αξιολόγηση. Τα μειονεκτήματα της μεθόδου είναι ότι δε λαμβάνει υπόψη τη τυχαία συμφωνία και δεν υπάρχουν βάρη. Η μέθοδος αυτή προτείνεται από τους συγγραφείς για χρήση ως εργαλείο ελέγχου ποιότητας, στην ιατρική επιστήμη καθώς και στην αξιολόγηση της συμφωνίας των βαθμολογητών στις Πανελλαδικές Εξετάσεις.

Κεφάλαιο 4: Μέτρα συμφωνίας σε ordinal data, interval data και ratio data

4.1 Εισαγωγή

Έως τώρα έχουν συζητηθεί οι περιπτώσεις ονομαστικών δεδομένων. Τι συμβαίνει όμως αν τα δεδομένα είναι διατάξιμα; Μια απλή περίπτωση που συναντάται είναι ο χαρακτηρισμός ενός υποκειμένου από "χαμηλή κλίμακα" έως "υψηλή κλίμακα" με ποικίλες ενδιάμεσες διατάξιμες κατηγορίες. Σε αυτή τη περίπτωση το μέτρο Kappa του Cohen δε μπορεί να εφαρμοστεί. Η αιτία είναι ότι ο Cohen θεωρεί ως ολική διαφωνία κάθε αξιολόγηση του βαθμολογητή (π.χ πολύ λίγο – λίγο). Το 1968 έγινε μια προσπάθεια από τον Cohen να προσθέσει "βάρη" για να διορθώσει αυτό το πρόβλημα. Χρειάζεται όμως, μια συστηματική και λογική προσέγγιση για επέκταση των συντελεστών συμφωνίας με απώτερο σκοπό τη διαχείριση διατάξιμων (ordinal data) καθώς και δεδομένα διαστήματος (interval data) και δεδομένα αναλογίας (ratio data). Οι Berry and Mielke (1988), οι Janson and Olsson (2001 και 2004) έχουν προτείνει επεκτάσεις του Kappa για ordinal, interval and ratio data. Παρά το γεγονός ότι οι Berry and Mielke (1988) ήταν πρωτοστάτες με την εισαγωγή των ιδεών τους, οι Janson and Olsson (2001) δίνουν με μεγαλύτερη σαφήνεια τις προτάσεις τους και επιπλέον τις επεκτείνουν λαμβάνοντας υπόψη τις ελλείπουσες τιμές (2004).

4.2 Γενικεύοντας τους συντελεστές στην περίπτωση των δυο βαθμολογητών

Στην παράγραφο αυτή θα ασχοληθούμε με τη γενίκευση των ήδη προαναφερθέντων συντελεστών συμφωνίας με σκοπό τη χρήση τους σε ordinal, interval και ratio data. Αρχικά, τα δεδομένα αυτά θα χρησιμοποιηθούν για τον υπολογισμό των "βαρών" ("weights"). Θα μας βοηθήσουν για να υπολογίσουμε τους σταθμισμένους συντελεστές: *Pi* του Scott (1955), το *Alpha* του Krippendorff και το AC_1 του Gwet (2008a). Έστω ότι οι βαθμολογητές Α και Β πρέπει να προσδιορίσουν κάθε υποκείμενο σε μια από τις q κατηγορίες (τύπου interval), δηλαδή $(x_1, \dots, x_k, \dots, x_q)$.

- *Weighted Scotts's Pi Coefficient*

Ο συντελεστής αυτός θα τον συμβολίσουμε ως $\widehat{\kappa_s}'$. Ορίζεται ως εξής:

$$\widehat{\kappa_s}' = \frac{p_a - p_e}{1 - p_e}\ , όπου\ p_a = \sum_{k,l}^{q} w_{kl}\, p'_{kl}\ και\ \ p_e = \sum_{k,l}^{q} w_{kl} \pi'_k \pi'_l \qquad (4.1)$$

όπου $\pi'_k = (p'_{k+} + p'_{+k})/2$. Η ποσότητα p'_{kl} προκύπτει ως το πηλίκο $\frac{p_{kl}}{p_{AB}}$. Με p_{AB} συμβολίζεται το σχετικό πλήθος των υποκειμένων που αξιολογήθηκαν και από τους δύο

βαθμολογητές και p_{kl} αντιπροσωπεύει το σχετικό πλήθος των υποκειμένων που ταξινομήθηκαν στις κατηγορίες k, l από τους βαθμολογητές A και B αντίστοιχα. Επίσης, $p'_{k+} = p_{k+}/p_A$, με p_{k+} είναι το σχετικό πλήθος των υποκειμένων που ταξινόμησε ο βαθμολογητής A στην κατηγορία k και p_A είναι το σχετικό πλήθος των υποκειμένων που έχει αξιολογήσει ο A. Αντίστοιχες ποσότητες ορίζονται για την ποσότητα π'_l. Για την ποσότητα w_{kl} θα γίνει ξεχωριστεί αναφορά σε επόμενη παράγραφο πιο σχολαστικά. Επειδή όμως, θα κάνουμε χρήση αριθμητικού παραδείγματος ας ορίζουμε για αρχή ως:

$$w_{kl} = \begin{cases} 1 - \dfrac{(x_k - x_l)^2}{(x_{\max} - x_{\min})^2}, & \alpha\nu\ k \neq l \\ 1, & \alpha\nu\ k = l \end{cases}$$

όπου x_{max} και x_{min} είναι η μεγαλύτερη και η μικρότερη τιμή αντίστοιχα. Θα αναφέρουμε επίσης ότι τα συγκεκριμένα βάρη είναι τα λεγόμενα "*Quadratic weights*".

Ας χρησιμοποιήσουμε τα παρακάτω παράδειγμα για τον υπολογισμό του σταθμισμένου συντελεστή Pi του Scott όπως και για τους υπόλοιπους συντελεστές.

Πίνακας 14: Αξιολόγηση 12 υποκειμένων από δύο βαθμολογητές με ύπαρξη ελλειπουσών τιμών

Unit	Rater1	Rater2
1	a	
2	b	c
3	c	c
4	c	c
5	b	b
6	b	
7	a	a
8	a	b
9	b	b
10	b	b
11		c

Πίνακας 15: Quadratic Weights (w_{kl})

	A	B	C
A	1	0.75	0
B	0.75	1	0.75
C	0	0.75	1

Έπειτα από υπολογισμούς όλων των προαναφερθέντων ποσοτήτων βρέθηκε ότι:

$p_a = 0.9375$ και $p_e = 0.7429$. Συνεπώς, $\widehat{\kappa'_S} = 0.7569$

Αν εφαρμόζαμε το μη σταθμισμένο δείκτη του Scott όπως αναφέρθηκε στην παράγραφο 1.6 θα βρίσκαμε $\widehat{\kappa_S} = 0.6038$.

- *Weighted Krippendorff's Alpha Coefficient*

Όπως έχουμε ήδη αναφέρει στην παράγραφο 1.7, ο δείκτης αυτός λαμβάνει υπόψη του μόνο τα υποκείμενα που αξιολογήθηκαν και από τους δύο αξιολογητές, ενώ τα υπόλοιπα αποκλείονται από τον υπολογισμό. Ο τύπος υπολογισμού του είναι:

$$\widehat{\alpha_K'} = \frac{p_a^* - p_e}{1 - p_e} \text{, όπου } \begin{cases} p_a^* = (1 - \varepsilon_n)p_a, & \varepsilon_n = 1/(2n) \\ p_a = \sum_{k,l}^{q} w_{kl}p_{kl}, & p_e = \sum_{k,l}^{q} w_{kl}\pi_{kl}\pi_l \end{cases} \quad (4.2)$$

Να σημειωθεί ότι $\pi_k = (p_{k+} + p_{+k})/2$, με p_{+k} και p_{k+} όπως ορίστηκαν προηγουμένως. Επίσης, σημαντικό είναι ότι ο σταθμισμένος *alpha* είναι παρόμοιος με τον δείκτη Pi του Scott όταν δεν υπάρχουν ελλείπουσες βαθμολογίες.

Στο παράδειγμα που προαναφέραμε, έπειτα από υπολογισμούς βρέθηκαν τα εξής:

$p_a = 0.9414$, $p_e = 0.7578$, και $\widehat{\alpha_K'} = 0.75806$. Αντίστοιχα το μη σταθμισμένο $alpha$ βρίσκουμε ότι είναι ίσο με $\widehat{\alpha_K} = 0.6203$

- *Weighted AC_1 Coefficient*

Συχνά αναφέρεται στη βιβλιογραφία ως AC_2 (Gwet 2008a). Ο τύπος υπολογισμού του ορίζεται ως εξής:

$$\widehat{\kappa_G'} = \frac{p_a - p_e}{1 - p_e} \text{, όπου } \begin{cases} p_a = \sum_{k,l}^{q} w_{kl}p'_{kl} \\ p_e = \frac{T_w}{q(q-1)} \sum_{k=1}^{q} \pi'_\kappa(1 - \pi'_\kappa) \end{cases} \quad (4.3)$$

όπου, $\pi'_\kappa = (p'_{k+} + p'_{+k})/2$. Η ποσότητα T_W παριστάνει το συνολικό άθροισμα όλων των σταθμισμένων τιμών w_{kl}.

Στο παράδειγμά μας έπειτα από υπολογισμούς προκύπτει ότι $p_a = 0.9375$ (όπως και στη περίπτωση του Pi του Scott), $p_e = 0.7429$ και τελικά $\widehat{\kappa_G}' = 0.8307$. Το μη σταθμισμένο μέτρο $\widehat{\kappa_G} = 0.6348$.

4.3 Μέτρα συμφωνίας τριών ή περισσότερων αξιολογητών

Έννοιες – Συμβολισμοί

Ας συμβολίσουμε με r_{ik} το πλήθος των βαθμολογητών που εκχώρησαν κάποιο σκορ x_k σε ένα υποκείμενο i. Το πλήθος των βαθμολογητών που αξιολόγησαν ένα υποκείμενο i θα συμβολίζεται ως r_i. Σε όλα τα μέτρα που θα παρουσιαστούν στη συνέχεια θα λαμβάνονται υπόψη αξιολογήσεις από τουλάχιστον έναν βαθμολογητή. Με q συμβολίζεται το πλήθος των κατηγοριών προς αξιολόγηση. Να σημειωθεί ότι σε κάθε έρευνα που γίνεται υπάρχουν κατηγορίες που είναι σχετικές με το αντικείμενο που αξιολογείται. Τέλος, με n_q έχουμε το

πλήθος των υποκειμένων που σχετίζονται με τον βαθμολογητή g και το σκορ x_k και με n_g το πλήθος των υποκειμένων που αξιολογούνται από τον g βαθμολογητή.

- *Conger's Kappa*

  Το μέτρο Kappa του Conger (Conger A,J. 1980) ορίζεται ως εξής:

$$\widehat{\kappa_C} = \frac{p_a - p_e}{1 - p_e}\text{, όπου } \begin{cases} p_a = \frac{1}{n'}\sum_{i=1}^{n'}\sum_{k=1}^{q}\frac{r_{ik}(r_{ik}^* - 1)}{r_i(r_i - 1)} \\ p_e = \sum_{k,l} w_{kl}(\bar{p}_{+k}\,\bar{p}_{+l} - s_{kl}^2/r). \end{cases} \quad (4.4)$$

  με n' παριστάνει το πλήθος των υποκειμένων που αξιολογήθηκαν από δυο ή περισσότερους βαθμολογητές. Επίσης,

$$r_{ik}^* = \sum_{l=1}^{q} w_{kl} r_{il}\,, \qquad \bar{p}_{+k} = \frac{1}{r}\sum_{g=1}^{r} p_{gk}\,, \quad (4.5)$$

$$p_{gk} = \frac{n_{gk}}{n_g} \;\; και \;\; s_{kl}^2 = \frac{1}{r-1}\left(\sum_{g=1}^{r} p_{gk}p_{gl} - r\bar{p}_{+k}\bar{p}_{+l}\right) \quad (4.6)$$

  Η συμφωνία σχετίζεται όχι μόνο εξετάζοντας το πλήθος των αξιολογητών r_{ik}, αλλά κοιτάζοντας και τα σκορ x_l που αντιπροσωπεύουν τη μερική συμφωνία με το x_k. Τα σκορ x_k και x_l $(k \neq l)$ είναι σε μερική συμφωνία όταν το βάρος w_{kl} είναι διάφορο του μηδενός.

- *Fleiss' Kappa*

  Ο συντελεστής Kappa του Fleiss (Fleiss, 1971) ορίζεται ως:

$$\widehat{\kappa_F}' = \frac{p_a - p_e}{1 - p_e}\text{, όπου } p_e = \sum_{k,l} w_{kl}\pi_k\pi_l \quad (4.7)$$

  και $\pi_k = \frac{1}{n}\sum_{i=1}^{n}\frac{r_{ik}}{r_i}$.

  Στο σημείο αυτό θα πρέπει να τονιστούν δυο πράγματα. Πρώτον, οτι ο συντελεστής *kappa* του Fleiss για δυο αξιολογητές με αυτόν που χρησιμοποιείται για τρείς ή περισσότερους διαφέρει μόνο ως προς τον υπολογισμό του π_k. Και δεύτερον, ο Fleiss (1971) πρότεινε ένα μη σταθμισμένο *kappa* ως επέκταση για τρεις ή περισσότερους αξιολογητές και αναφέρεται μόνο σε nominal data και χωρίς να μπορεί να διαχειριστεί missing rates. Ο Gwet (2014) επεκτείνει αυτό το συντελεστή στα interval data, με σκοπό τη διαχείριση και των missing data.

- *Krippendorff's Alpha Coefficient*

  Ο Gwet (2014) προτείνει έναν πιο απλοποιημένο τύπο για τον υπολογισμό του $alpha$ με σκοπό την πιο εύκολη υπολογιστική χρήση. Για να γίνει πιο αντιληπτή η κατανόηση του τύπου δίνονται οι παρακάτω έννοιες-συμβολισμοί.

  1. n' ορίζεται το πλήθος των υποκειμένων που αξιολογήθηκαν από τουλάχιστον δυο βαθμολογητές και με n το συνολικό πλήθος των υποκειμένων που έλαβαν μέρος στο πείραμα.
  2. r_i είναι το πλήθος των βαθμολογητών που αξιολόγησαν το υποκείμενο i.
  3. $\bar{r}$ είναι η μέση τιμή των r_i.
  4. $\varepsilon_n = 1/(n'\bar{r})$.

  Ο τύπος υπολογισμού του $alpha$ του Krippendorff είναι:

$$\widehat{\alpha_\kappa} = \frac{p_a - p_e}{1 - p_e}, \text{όπου} \begin{cases} p_a = (1-\varepsilon_n)p'_a + \varepsilon_n \\ p'_a = 1/n' \sum_{i=1}^{n'} \sum_{k=1}^{q} \frac{r_{ik}(r^*_{ik} - 1)}{\bar{r}(r_i - 1)} \\ p_e = \sum_{k,l} w_{kl}\pi_k\pi_l \end{cases} \quad (4.8)$$

  με $r^*_{ik} = \sum_{l=1}^{q} w_{kl} r_{il}$ και $\pi_k = \frac{1}{n'}\sum_{i=1}^{n'} \frac{r_{ik}}{\bar{r}}$

Ο ανωτέρω τύπος μοιάζει με το τύπο του Fleiss ($\widehat{\kappa_F}$). Διαφέρουν ως προς το πώς αντιμετωπίζουν την ύπαρξη missing rates καθώς και στη ποσοστιαία συμφωνία p_a.

- *Gwet's AC_2 Coefficient*

  Ο δείκτης αυτός είναι η σταθμισμένη έκδοση του AC_1 (Gwet 2008a). Ορίζεται ως:

$$\widehat{\kappa_G} = \frac{p_a - p_e}{1 - p_e}, \text{όπου} \begin{cases} p_a = \frac{1}{n'} \sum_{i=1}^{n'} \sum_{k=1}^{q} \frac{r_{ik}(r^*_{ik} - 1)}{r_i(r_i - 1)} \\ p_e = \frac{T_w}{q(q-1)} \sum_{k,l} \pi_\kappa(1 - \pi_\kappa). \end{cases} \quad (4.9)$$

  Ο υπολογισμός κάθε ποσότητας που αναφέρεται έχει ήδη οριστεί.

4.4 Υπολογιστικό παράδειγμα στη περίπτωση των 3 ή περισσότερων αξιολογητών

Για τους συντελεστές *Conger'sWeighted Kappa, Fleiss' Weighted Kappa, Krippednorff's Weighted Kappa* και *Gwet's Weighted Kappa* που αναφέρθηκαν θα χρησιμοποιηθεί ένα παράδειγμα με πραγματικά δεδομένα.

Πίνακας 16: Κατανομή σκορ σε 16 υποκείμενα από 4 αξιολογητές

Subject	L	K	W	B
a.lycia	1	1,5	1	
a.milbe	2	2	2	2
a.hegon	0,5	1	1,5	1,5
a.oslar	1	1	1	1
a.viali	1	1	1	1,5
a.logan		1	2,5	
a.numit	2,5	2,5	2,5	2,5
a.saraa	1	1		1
a.sarat		1	2	1
a.mormo	1	1	0,5	1
a.celti	1,5	1,5	1,5	1,5
a.clyto	1	1,5	1	
a.hiann	1	1	1,5	
b.phile	1	2	2,5	2
b.alask		1	1,5	1
b.taad	0,5	0,5	0,5	0,5

Στο παράδειγμα αυτό θα χρησιμοποιηθούν quadratic weights τα οποία φαίνονται στον Πίνακα 17.

Πίνακας 17: Quadratic Weights

	0,5	1	1,5	2	2,5
0,5	1	0,9375	0,75	0,4375	0
1	0,9375	1	0,9375	0,75	0,4375
1,5	0,75	0,9375	1	0,9375	0,75
2	0,4375	0,75	0,9375	1	0,9375
2,5	0	0,4375	0,75	0,9375	1

Έπειτα από υπολογισμούς βρέθηκε ότι:

Για όλα τα μέτρα πέρα από το *alpha* του *Krippendorff* προκύπτει ότι $p_a = 0.9206$, ενώ για το *alpha προκύπτει* $p_a = 0.9364$. Επίσης, για τις ποσοστιαίες τυχαίες συμφωνίες προέκυψαν τέσσερις διαφορετικές που απεικονίζονται στον Πίνακα 18.

Πίνακας 18: Υπολογισμός των ποσοστών τυχαίας συμφωνίας

Conger's Weighted	Fleiss' Weighted	Krippendorff's Weighted	Gwet's Weighted
0.8314	0.8377	0.8336	0.6462

Έπειτα από εφαρμογή των τύπων του κάθε μέτρου προέκυψαν τα παρακάτω αποτελέσματα που συνοψίζονται στον Πίνακα 19.

Πίνακας 19: Εκτιμήσεις των μέτρων συμφωνίας

Conger's Weighted Kappa	Fleiss' Weighted Kappa	Krippendorff's Weighted Kappa	Gwet's Weighted Kappa
0.5290	0.5107	0.6180	0.7755

Από τα αποτελέσματα του παραπάνω Πίνακα 17, προκύπτει ότι ο συντελεστής του Gwet είναι αρκετά μεγαλύτερος από τους άλλους τρείς. Αυτό συμβαίνει, καθώς οι μέθοδοι των Conger, Fleiss και Krippendorff έχουν τη τάση να υπολογίζουν πολύ μεγάλα ποσοστά τυχαίας συμφωνίας (βλ. Πίνακα 18).

4.5 Επιλογές συντελεστών βαρύτητας

Σε προηγούμενο κεφάλαιο έχει ήδη γίνει αναφορά για το σταθμισμένο *Kappa* του Cohen (1968). Τα βάρη χρησιμοποιήθηκαν για να διαβαθμίσουν τη σημασία δύο ή περισσότερων κατηγοριών. Στη παράγραφο αυτή προτείνονται τα Ordinal, Linear, Radical, Ratio, Circular και Bipolar Weights. Να σημειωθεί ότι δεν υπάρχει συγκεκριμένος λόγος για την επιλογή του είδους του βάρους που κάποιος μπορεί να επιλέξει στην ανάλυση του. Ωστόσο, όλα τα είδη στοχεύουν στην ενσωμάτωση μερικών συμφωνιών στον υπολογισμό του εκάστοτε συντελεστή συμφωνίας. Το δεύτερο κοινό στους στοιχείο είναι ότι για "τέλειες" συμφωνίες απονέμεται η τιμή 1, για μερικές συμφωνίες μια τιμή μεταξύ 0 και 1 και για "πλήρης" διαφωνίες λαμβάνουν τη τιμή 0.

Στη περίπτωση που έχουμε non-numeric ordinal ratings, η χρήση των quadratic weights που χρησιμοποιήθηκαν σε παράδειγμα της προηγούμενης παραγράφου ίσως να μην εφαρμόζονται καλά. Η αιτία είναι ότι, αν και συνήθως αριθμούμε τη πρώτη κατηγορία με τη τιμή 1, τη δεύτερη με τη τιμή 2 κ.ο.κ, η επιλογή αυτή είναι υποκειμενική. Γι' αυτό προτείνονται τα ordinal weights. Τα ordinal weights δεν επηρεάζονται από τις πραγματικές τιμές των βαθμολογιών, παρά μόνο η σειρά τους (rank).

Για τα linear weights οι τιμές τους εξαρτώνται από τα αν οι αξιολογήσεις είναι alphabetic ή numeric type. Αν είναι alphabetic, τότε αριθμούνται από το 1 έως το q (όπου q το πλήθος των κατηγοριών). Αν είναι ήδη numeric, τότε αυτές οι τιμές αξιολόγησης θα πρέπει να χρησιμοποιηθούν. Τα βάρη αυτά προτείνονται από τον Cohen (1968). Αν ο ερευνητής χαρακτηρίζει τα προηγούμενα δυο βάρη ως "πολύ μεγάλα", τότε ο Gwet (2014) προτείνει τα Radical Weights. Είναι παρόμοια με τα Linear διαφέρουν μόνο ως προς το τύπο υπολογισμού τους.

Τα Ratio Weights έχουν χρησιμοποιηθεί από τον Krippednforff (1970, 1978, 2004) και μπορούν να χρησιμοποιηθούν σε rating data. Εδώ, χρειάζεται προσοχή καθώς όταν πρόκειται για non-numeric ordinal data θα πρέπει ο ερευνητής να αποδώσει μια ακέραια τιμή σε κάθε κατηγορία πριν τον υπολογισμό των βαρών. Τα Circular Weights , που χρησιμοποιήθηκαν ξανά από τον Krippendorff (1970, 1978, 2004), προτείνονται όταν η αξιολόγηση αντιπροσωπεύει το μέγεθος μιας γωνίας εκφρασμένης είτε σε μοίρες είτε σε ακτίνια. Τέλος, τα Bipolar Weights, συμπεριφέρονται όπως τα Ratio στο κέντρο της κλίμακας και όπως τα Quadratic προς το τέλος της.

Κεφάλαιο 5: Μοντέλα συμφωνίας

5.1 Εισαγωγή

Στις κοινωνικές επιστήμες είτε στις επιστήμες υγείας, οι βαθμολογητές καλούνται να αξιολογήσουν ένα δείγμα υποκειμένων σε ονοματικές ή διατάξιμες κατηγορίες. Σε έρευνες βλέπουμε συχνά τις επιλογές «καθόλου», «ελάχιστα», «λίγο», «αρκετά», «πολύ», «πάρα πολύ». Σε τέτοιες περιπτώσεις οι αξιολογητές είναι αναμενόμενο να μη δείχνουν τέλεια συμφωνία, με αποτέλεσμα να υπάρχει ετερογένεια των ταξινομήσεων τους.

Μέχρι τώρα, έχουμε αναφέρει τα μέτρα συμφωνίας βαθμολογητών, με δημοφιλέστερο το δείκτη *kappa* του Cohen. Έχουμε αναφέρει πολλές γνωστές επεκτάσεις του με σκοπό την αποφυγή των ανεπιθύμητων χαρακτηριστικών που έχει ο δείκτης *kappa*.

Στο παρών κεφάλαιο, θα ασχοληθούμε με log-linear μοντέλα συμφωνίας. Τα μοντέλα αυτά αναλύουν τη δομή της συμφωνίας στα δεδομένα (Tanner and Young, 1985a). Γενικότερα, οι μελέτες που ασχολούνται με μοντέλα δίνουν περισσότερες πληροφορίες για τον πίνακα συνάφειας. Εκτός από την ανάλυση της συμφωνίας, θα υπολογιστούν και odds ratios κάτω από το προσαρμοσμένο μοντέλο για να συμπεράνουν τον βαθμό της συμφωνίας.

5.2 Μοντέλα συμφωνίας για ονοματικά δεδομένα

Τα μοντέλα συμφωνίας προτείνονται για χρήση σε τετραγωνικούς πίνακες συνάφειας με ονοματικά δεδομένα. Τα μοντέλα αυτά είναι τα agreement (Tanner and Young, 1985a), disagreement, symmetric band disagreement (Tanner and Young, 1985b) και agreement plus disagreement (Saracbasi, 2011b) models. Ας υποθέσουμε ότι έχουμε έναν τετραγωνικό πίνακα συνάφειας $R\ x\ R$ όπου ο πρώτος αξιολογητής εκπροσωπείται από το σύμβολο X και ο δεύτερος από το σύμβολο Y. Στον πίνακα αυτό διπλής εισόδου, n σε πλήθος υποκείμενα ταξινομούνται σε δύο κατηγορίες απόκρισης. Το log-linear μοντέλο δίνεται ως εξής:

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \delta_{ij}, \qquad (5.1)$$

όπου λ είναι η παράμετρος ολικής επίδρασης, λ_i^X είναι η επίδραση του αξιολογητή X στο i υποκείμενο και λ_i^Y είναι η επίδραση του Y στο j υποκείμενο. Οι περιορισμό είναι: $\sum_{i=1}^{R} \lambda_i^X = \sum_{j=1}^{C} \lambda_j^Y = 0$. Η ποσότητα m_{ij} είναι οι αναμενόμενες τιμές και δ_{ij} είναι η παράμετρος συμφωνίας ανάμεσα στις X και Y, με $i = 1,2,\dots,R$ και $j = 1,2,\dots,C$. Οι παράμετροι των agreement, disagreement και symmetric band disagreement models δίνονται παρακάτω:

$$\delta_{ij} = \begin{cases} \delta, & \alpha\nu\ i = j \\ 0, & \alpha\nu\ i \neq j \end{cases} \qquad \text{(agreement)} \qquad (5.2)$$

$$\delta_{ij} = \begin{cases} \delta, & \alpha\nu\ i \neq j \\ 0, & \alpha\nu\ i = j \end{cases} \qquad \text{(disagreement)} \qquad (5.3)$$

$$\delta_{ij} = \begin{cases} \delta_1, & \alpha\nu\ |i-j| = 1 \\ \delta_2, & \alpha\nu\ |i-j| = 2 \\ \vdots & \\ \delta_{R-1}, & \alpha\nu\ |i-j| = R-1 \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \qquad (5.4)$$

Τα agreement και disagreement models έχουν $(R-1)^2 - 1$ βαθμούς ελευθερίας, ενώ τα symmetric band disagreement models έχουν $(R-1)^2 - R + 1$ βαθμούς ελευθερίας.

Τα agreement plus disagreement models έχουν τη μορφή:

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \gamma_{ij} + \delta_{ij}, \qquad (5.5)$$

όπου γ_{ij} είναι η παράμετρος της (5.2) και δ_{ij} είναι η παράμετρος της (5.4).

Τα odds ratios θ_{ij} των επιτυχημένα προσαρμοσμένων μοντέλων μπορούν να χρησιμοποιηθούν για να συμπεράνουν τη συμφωνία. Όταν τα odds ratio αποκλίνουν από τη τιμή 1, σημαίνει ότι οι αποφάσεις των αξιολογητών είναι παραπλήσιες, ενώ όταν συγκλίνει στο 0, σημαίνει ότι οι αποφάσεις τους είναι περισσότερο διαφορετικές από όμοιες. Η ομοιότητα υποδεικνύει συμφωνία μεταξύ τους. Τα odds ratios (θ_{ij}) υπολογίζονται ως εξής:

$$\theta_{ij} = \frac{m_{ij}m_{i+1,j+1}}{m_{i+1,j}m_{i,j+1}}, \qquad \mu\varepsilon\ i,j = 1,2,\dots,R \qquad (5.6)$$

5.3 Παράδειγμα με αποτελέσματα και σχόλια

Παρακάτω δίνεται ένας πίνακας δεδομένων που αντλήθηκε από τον Gwet (2012).

Πίνακας 20: Ταξινόμηση σπονδυλικών πόνων από 2 αξιολογητές

	Clinician B		
Clinician A	**Derangement Syndrome**	**Dysfunctional Syndrome**	**Postural Syndrome**
Derangement Syndrome	55	10	2
Dysfunctional Syndrome	6	4	10
Postural Syndrome	2	5	6

Οι Ayfer Ezgi Yilmaz and Tulay Saracbasi (2017) προσάρμοσαν agreement models στα παραπάνω δεδομένα. Σύμφωνα με τα αποτελέσματα του παρακάτω Πίνακα 21 συμπεραίνουμε τα εξής:

Τα agreement και disagreement model δεν προσαρμόζουν τα δεδομένα, αλλά το symmetric band disagreement model προσαρμόζει τα δεδομένα σε επίπεδο σημαντικότητας 1%. Η παράμετρος συμφωνίας στο model agreement model είναι $\delta > 0$ και είναι σημαντική σε επίπεδο 5%. Συνεπώς, υπάρχει περισσότερη συμφωνία από ότι αναμενόταν κατά τύχη. Στα disagreement models, υπάρχει λιγότερη διαφωνία από ότι αναμενόταν κατά τύχη καθώς $\delta < 0$. Στο symmetric band model, οι παράμετροι διαφωνίας είναι και οι δυο σημαντικοί κάτι που υποδεικνύει συμφωνία μεταξύ των αξιολογητών.

Πίνακας 21: Αποτελέσματα των agreement models

Models	G^2	**Df**	**p-value**	**Parameter Estimations**
Agreement	24.959	3	<0.01	$\hat{\delta} = 0.974$
Disagreement	24.959	3	<0.01	$\hat{\delta} = -0.974$
Symmetric band disagreement	6.756	2	0.034	$\hat{\delta}_1 = -0.297$, $\hat{\delta}_2 = -2.477$

Επίσης, το καλύτερο μοντέλο που προσαρμόζεται στα δεδομένα είναι το symmetric band disagreement model. Τα odds ratios μπορούν να ερμηνευθούν από τις εκτιμώμενες παραμέτρους. Η πιθανότητα να αποδοθεί derangement syndrome σε σχέση με dysfunctional syndrome του αξιολογητή Α΄ είναι 1.81 φορές μεγαλύτερη από το να αποδοθεί derangement syndrome σε σχέση με dysfunctional syndrome από τον αξιολογητή Β΄. Ακόμα, η πιθανότητα να αποδοθεί derangement syndrome σε σχέση με dysfunctional syndrome από τον αξιολογητή Α΄ είναι 6.57 φορές μεγαλύτερη από να αποδοθεί dysfunctional syndrome σε σχέση με postural syndrome από τον αξιολογητή Β΄. Συνεπώς, οι αποφάσεις των αξιολογητών είναι περισσότερο παρόμοιες από κατηγορία σε κατηγορία και υπάρχει συμφωνία μεταξύ τους.

5.4 Μοντέλα συμφωνίας για διατάξιμα δεδομένα για 2 αξιολογητές

Μια μέθοδος εφαρμογής των μοντέλων συμφωνίας για πίνακες με διατάξιμα δεδομένα είναι να αγνοηθεί η ιεραρχία ανάμεσα σε γειτονικές κατηγορίες των διατάξιμων. μεταβλητών. Κάτι τέτοιο όμως θα οδηγούσε σε έλλειψη πληροφορίας. Η κατάλληλη ανάλυση για τετραγωνικούς πίνακες συσχέτισης έχοντας διαταγμένες κατηγορίες είναι τα *association models* με παράμετρο συμφωνίας.

Το *linear-by-linear association plus agreement mode* για δύο τακτικές μεταβλητές δίνεται ως εξής:

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_l^Y + \beta u_i v_i + \delta_{ij}, \qquad (5.7)$$

όπου $u_1 \leq u_2 \leq \cdots \leq u_R$ είναι τα σκορ γραμμών και $v_1 \leq v_2 \leq \cdots \leq v_C$ είναι τα σκορ στηλών και β είναι η παράμετρος συσχέτισης. Το μοντέλο αυτό έχει $2(R-1)-2$ βαθμούς ελευθερίας. Ο Goodman (1979) αναφέρθηκε στην ειδική περίπτωση του *uniform association plus agreement model (UAA),* όπου $u_i = i$ και $v_j = j$. Οι Bagheban and Zayeri (2010) αναφέρθηκαν στο *model exponential scores association plus agreement (EAA),* όπου $u_i = i^a$ και $v_j = j^b$. Οι Aktas and Saracbasi (2009) αναφέρθηκαν στο *model symmetric disagreement plus uniform association (DUA),* το οποίο έχει $(R+1)(R-3)$ βαθμούς ελευθερίας και η παράμετρο συμφωνίας δ_{ij} ορίζεται ως:

$$\delta_{ij} = \begin{cases} \delta_1 & \alpha\nu\ |i-j| = 1 \\ \delta_2 & \alpha\nu\ |i-j| = 2 \\ \delta & \alpha\nu\ |i-j \geq 3 \\ 0 & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \tag{5.8}$$

Επιπροσθέτως, οι Valet, Guinot και Mary (2007) πρότειναν το *non-uniform association plus agreement model (NUAA),* με σκοπό να περιγράψουν τη διαφορά της διακριτικότητας ανάμεσα σε γειτονικές κατηγορίες. Σε αντίθεση με το *UAA model,* το προτεινόμενο αυτό μοντέλο περιλαμβάνει $(R-1)$ παραμέτρους συσχέτισης $\beta_{k,k+1}$ και επεκτείνει το *UAA model* επιτρέποντας αποκλίσεις διακριτικότητας ανάμεσα στις γειτονικές κατηγορίες. Για το λόγο αυτό, το μη ομοιόμορφο αυτό μοντέλο περιγράφει τη ποιότητα μιας διατάξιμης κλίμακας με περισσότερη ακρίβεια. Η αλγεβρική μορφή του μοντέλου είναι:

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_l^Y - \frac{|i-j|}{2} \times \sum_{k=\min(i,j)}^{\max(i,j)-1} \beta_{k,k+1} + \delta_{ij} \tag{5.9}$$

με δ_{ij} δίνεται από την (5.8).

Το μοντέλο αυτό έχει $\mathrm{R}^2 - 3R + 1$ βαθμούς ελευθερίας.

Οι Fu, Gao, Tang και Shi (2012) πρότειναν ένα μοντέλο συνδέοντας τη πληροφορία της τακτικής κλίμακας και τη διακριτικότητα ανάμεσα στις διαταγμένες κατηγορίες. Επίσης, για το μοντέλο αυτό δεν απαιτείται εκχώρηση σκορ για της διατάξιμες κατηγορίες. Η μορφή του μοντέλου είναι:

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_l^Y + \lambda_{|i-j|}, \tag{5.10}$$

με $0 = \lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{R-1}$. Οι βαθμοί ελευθερίας τους είναι $(R-1)(R-2)$.

5.5 Παράδειγμα με αποτελέσματα και σχόλια

Παρακάτω δίνεται ένα παράδειγμα 149 ασθενών από το Winnipeg του Καναδά που ταξινομήθηκαν από δυο νευρολόγους και κατατάχθηκαν σε τέσσερις διαγνωστικές κατηγορίες με σκοπό να ερευνήσουν τη πιθανότητα η ασθένεια κατανεμήθηκε διαφορετικά γεωγραφικά. Τα δεδομένα ελήφθησαν από του Westlund και Kurland (1953) και

συζητήθηκαν από τους Landis και Koch (1977a), Gwet (2012) και από τους Bangdiwala και Shankar (2013). Τα δεδομένα συγκεντρώνονται στον Πίνακα 20.

Πίνακας 22: Αξιολόγηση 149 ασθενών από δυο ανεξάρτητους νευρολόγους ως προς τη πιθανότητα διάγνωσης σκλήρωσης

	Winnipeg Neurologist			
New Orleans Neurologist	**Certain**	**Probable**	**Possible**	**No**
Certain	38	5	0	1
Probable	33	11	3	0
Possible	10	14	5	6
No	3	7	3	10

Έπειτα από ανάλυση προκύπτει ο Πίνακας 23.

Πίνακας 23: Αποτελέσματα από τα μοντέλα συμφωνίας που προκύπτουν από τα δεδομένα του Πίνακα 20.

Μοντέλα	G^2	**Df**	**p-value**	**Parameter**	**Estimations**	**AIC**
UAA	9.416	7	0.224	$\hat{\beta} = 0.804^{*}$	$\hat{\delta} = 0.028^{*}$	-4.584
DUA	6.956	5	0.224	$\hat{\beta} = 0.429$ $\hat{\delta}_2 = -0.627$	$\hat{\delta}_1 = -0.195$ $\hat{\delta}_3 = -1.348$	-3.044
NUAA	7.968	5	0.158	$\hat{\beta}_{12} = 1.094^{*}$ $\hat{\beta}_{34} = 0.492$	$\hat{\beta}_{23} = -0.856^{**}$ $\hat{\delta} = -0.099$	-2.032
Fu	8.140	6	0.228	-		-3.860
*: Η παράμετρο είναι σημαντική σε επίπεδο σημαντικότητας $\alpha = 0.05$ **: Η παράμετρος είναι σημαντική σε επίπεδο σημαντικότητας $\alpha = 0.01$						

Από τα αποτελέσματα του Πίνακα 23 συμπεραίνουμε ότι όλα τα μοντέλα προσαρμόζονται στα δεδομένα. Επίσης, οι παράμετροι συμφωνίας του μοντέλου δεν είναι σημαντικοί. Σύμφωνα με το κριτήριο του Akaike (Akaike, 1974), το μοντέλο που προσαρμόζεται καλύτερα στα δεδομένα είναι εκείνο με τη μικρότερη τιμή AIC, δηλαδή το *uniform association plus agreement model (UAA).*

Με χρήση της (5.9) η πιθανότητα να εξάγουν ίδια απόφαση οι νευρολόγοι από τη Νέα Ορλεάνη και από το Winnipeg είναι 2.36 φορές μεγαλύτερη από το να δοθεί απόφαση ενός μεγαλύτερου επιπέδου. Αυτό σημαίνει, ότι υπάρχει συμφωνία μεταξύ τους. Η πιθανότητα να δοθεί (“certain”) απόφαση σε σχέση με (“possible”) πιθανή απόφαση από τον νευρολόγο

από την Νέα Ορλεάνη είναι 2.17 φορές μεγαλύτερη από το να δοθεί ("probable") απόφαση σε σχέση με "possible" από το νευρολόγο του Winnipeg. Ο πίνακας με τα odds ratios (θ_{ij}) δίνεται παρακάτω:

$$\hat{\theta}_{UAA} = \begin{bmatrix} 2.36 & 2.17 & 2.23 \\ 2.17 & 2.36 & 2.17 \\ 2.23 & 2.17 & 2.36 \end{bmatrix}$$

5.6 Μοντέλα συμφωνίας με περισσότερους από 2 αξιολογητές για τακτικές κατηγορίες

Για ονομαστικά δεδομένα προτείνονται από τη βιβλιογραφία τα global, global and partial, global and partial σύμφωνα με τις κατηγορίες βαθμολόγησης και τα global and heterogeneous partial models (Rogel, Boelle, and Mary, 1998; Kastango,2006). Για τα τακτικά δεδομένα αντίστοιχα προτείνονται τα Association plus agreement models.

Στην περίπτωση των τακτικών δεδομένων, έστω ότι έχουμε τρεις αξιολογητές τους X, Y, Z. Με $u_i = i, v_j = j$ και $\omega_k = k$ συμβολίζονται οι τιμές σκορ των μεταβλητών X, Y, Z αντίστοιχα. Με β_1 συμβολίζεται η παράμετρος συσχέτισης ανάμεσα στον X και Y, β_2 ανάμεσα στον X και Z και β_3 ανάμεσα στον Y και Z. Η παράμετρος δ αντιπροσωπεύει την παράμετρο ολικής συμφωνίας *(global parameter)*, η οποία υποδεικνύει τη συμφωνία ανάμεσα στους τρεις βαθμολογητές. Στον παρακάτω πίνακα απεικονίζονται association plus agreement models όπου $i = j = k = 1,2, \dots R$ (Melia and Dienaer-West, 1994; Lawal, 2003; Saracbasi, 2011a).

Πίνακας 24: Uniform Association plus Agreement models για μελέτες με περισσότερους από δύο αξιολογητές

Μοντέλο	Εξίσωση
M1	$log(m_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \delta_1 + \delta_2 + \delta_3 + \delta_4$
M2	$log(m_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 u_i v_j + \beta_2 u_i w_k + \beta_3 v_j w_k + \beta_4 u_i v_j w_k$
M3	$log(m_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 u_i v_j + \beta_2 u_i w_k + \beta_3 v_j w_k + \delta_1 + \delta_2 + \delta_3 + \delta_4$
M4	$log(m_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 u_i v_j + \beta_2 u_i w_k + \beta_3 v_j w_k + \delta_1 + \delta_2 + \delta_3$
M5	$log(m_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 u_i v_j + \beta_2 u_i w_k + \beta_3 v_j w_k + \delta_4$
M6	$log(m_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 u_i v_j + \beta_2 u_i w_k + \beta_3 v_j w_k + \beta_4 u_i v_j w_k + \delta_1 + \delta_2 + \delta_3$
M7	$log(m_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 u_i v_j + \beta_2 u_i w_k + \beta_3 v_j w_k + \beta_4 u_i v_j w_k + \delta_1 + \delta_2 + \delta_3 + \delta_4$

Η μεταβλητή $\beta_{l,l+1}$ με $l = 1,2,\dots,(R-1)$, είναι η παράμετρος συσχέτισης ανάμεσα στις γειτονικές κατηγορίες l και $l+1$ των αξιολογητών X, Y. Αντίστοιχα, η παράμετρος $\varphi_{l,l+1}$ δηλώνει τη συσχέτιση ανάμεσα στους X, Z και η παράμετρος $\omega_{l,l+1}$ τη συσχέτιση ανάμεσα στους Y, Z. Από τον Yilmaz (2013) διατίθενται τα μη ομοιόμορφα Association plus Agreement models τα οποία απεικονίζονται στον παρακάτω Πίνακα.

Πίνακας 25: Non-Uniform Association plus Agreement models για περισσότερους από δυο αξιολογητές

Μοντέλο	Εξίσωση
M8	$\log(m_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert}{2} \times \sum_{l=\min(i,j)}^{\max(i,j)-1} \beta_{l,l+1} - \frac{\lvert i-k\rvert}{2} \times \sum_{l=\min(i,k)}^{\max(i,k)-1} \varphi_{l,l+1} - \frac{\lvert j-k\rvert}{2} \times \sum_{l=\min(i,j)}^{\max(j,k)-1} \omega_{l,l+1} - \frac{\lvert i-j\rvert+\lvert i-k\rvert+\lvert j-k\rvert}{2(R-1)} \beta$
M9	$\log(m_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert}{2} \times \sum_{l=\min(i,j)}^{\max(i,j)-1} \beta_{l,l+1} - \frac{\lvert i-k\rvert}{2} \times \sum_{l=\min(i,k)}^{\max(i,k)-1} \varphi_{l,l+1} - \frac{\lvert j-k\rvert}{2} \times \sum_{l=\min(i,j)}^{\max(j,k)-1} \omega_{l,l+1} - \frac{\lvert i-j\rvert+\lvert i-k\rvert+\lvert j-k\rvert}{2(R-1)} \beta + \delta_4$
M10	$\log(m_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert+\lvert i-k\rvert+\lvert j-k\rvert}{2(R-1)} + \delta_4$
M11	$\log(m_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert+\lvert i-k\rvert+\lvert j-k\rvert}{2(R-1)} + \delta_1 + \delta_2 + \delta_3$
M12	$\log(m_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert+\lvert i-k\rvert+\lvert j-k\rvert}{2(R-1)} + \delta_1 + \delta_2 + \delta_3 + \delta_4$

5.7 Παράδειγμα με αποτελέσματα και σχόλια

Στον παρακάτω πίνακα απεικονίζονται δεδομένα που συζητήθηκαν για πρώτη φορά απο τους Holmquist, McMahon και Williams (1967) και αργότερα από τους Landis and Koch (1977b), Becker and Agresti (1992) και τον Saracbasi (2011a). Με σκοπό τη διερεύνηση της μεταβλητότητας στη ταξινόμηση του καρκινώματος in Situ[3], τρεις παθολόγοι ταξινόμησαν 118 slides σε 5 κατηγορίες. Επειδή τα δεδομένα περιέχουν σε κάποιες κατηγορίες αρκετά μηδενικά, οι αυθεντικές κατηγορίες έχουν συμπτυχτεί στις εξής κατηγορίες:

(1) Negative (αρνητικό)
(2) Atypical Squamous Hyperplasia (Άτυπη πλακώδης υπερπλασία)
(3) Carcinoma in Situ + Squamous Carcinoma with Early Stromal Invasion +Invasive Carcinoma (Καρκίνωμα in Situ + Πλακώδες καρκίνωμα με πρώιμη στρωματική εισβολή + διηθητικό καρκίνωμα)

[3] In situ: Χρησιμοποιείται στη βιολογία και σημαίνει να εξεταστεί ένα φαινόμενο ακριβώς στο μέρος που εμφανίζεται.

Πίνακας 26: Ταξινόμηση 118 ασθενών από τρεις ανεξάρτητους παθολόγους

X	Y	Z 1	2	3
1	1	12	10	0
	2	1	1	0
	3	0	2	0
2	1	2	3	0
	2	1	4	2
	3	0	5	9
3	1	0	0	0
	2	0	2	1
	3	0	4	59

Έπειτα από ανάλυση όλων των μοντέλων M1 ως M10 των Yilmaz and Saracbasi (2013) βρέθηκαν τα εξής αποτελέσματα:

Πίνακας 27: Αποτελέσματα του παραδείγματος των 118 ασθενών

Μοντέλα	G^2	df	P-Value	AIC
M1	52.374	16	< 0.001	–
M2	7.222	16	0.969	-24.778
M3	5.983	13	0.947	-20.017
M4	5.983	14	0.967	-22.017
M5	9.581	16	0.888	-22.419
M6	3.453	13	0.996	-22.547
M7	3.453	12	0.991	-20.547
M8	7.106	13	0.897	-18.894
M9	3.306	10	0.973	-16.694
M10	12.870	18	0.799	-23.130
M11	9.882	16	0.873	-22.118
M12	8.478	15	0.903	-21.522

Συμπεραίνουμε ότι όλα τα μοντέλα προσαρμόζονται στα δεδομένα πέρα από το M1. Το καλύτερο μοντέλο από τα ομοιόμορφα είναι το M2 και από τα μη ομοιόμορφα είναι το M10 καθώς έχουν τη μικρότερη τιμή AIC.

Στον Πίνακα 28 απεικονίζονται οι εκτιμήσεις των παραμέτρων, τα τυπικά τους σφάλματα καθώς και τα αντίστοιχα p-value των δύο καλύτερων μοντέλων, δηλαδή του M2 και M10.

Πίνακας 28: Εκτιμήσεις των παραμέτρων των μοντέλων M2 και M10

Μοντέλα	**Εκτιμήσεις Παραμέτρων**	**Standard error**	**P-Value**
M2	$\hat{\beta}_1 = 0.177$ $\hat{\beta}_2 = 0.312$ $\hat{\beta}_3 = 0.686$ $\hat{\beta}_4 = 0.578$	0.842 0.962 0.867 0.371	0.834 0.745 0.428 0.119
M10	$\hat{\beta}_4 = 7.383$ $\hat{\delta} = -2.041$	1.514 0.863	0.000 0.018

Συμπεραίνουμε ότι οι εκτιμήσεις των παραμέτρων του μοντέλου M10 είναι σημαντικές σε επίπεδο σημαντικότητας $\alpha = 5\%$. Αντιθέτως, καμία από τις παραμέτρους του μοντέλου M2 δεν είναι σημαντικές. Για τη παράμετρο συμφωνίας δ, η οποία είναι σημαντική, παρατηρούμε ότι έχει αρνητική τιμή. Αυτό μας υποδεικνύει ότι υπάρχει λιγότερη συμφωνία από ότι αναμενόταν κατά τύχη. Αρνητική τιμή του δ μας δείχνει ότι υπάρχει διαφωνία μεταξύ των βαθμολογητών.

Για το μοντέλο M10 θα υπολογιστούν τα odds ratios. Τα odds ratios για τους πολλαπλούς πίνακες ονομάζονται "conditional odds ratios", όπου ο ένας αξιολογητής θεωρείται σταθερός (fixed). Για τους πίνακες με τα conditional odds ratios θεωρείται $\hat{\theta}_{(i)jk} = \hat{\theta}_{i(j)k} = \hat{\theta}_{ij(k)}$ και δίνονται με τη βοήθεια της εξίσωσης (5.8). Ο εν λόγω πίνακας απεικονίζεται παρακάτω.

$$\hat{\theta}_{(i)jk} = \hat{\theta}_{i(j)k} = \hat{\theta}_{ij(k)} =$$

5.21	1.00
1.00	40.11
– – –	– – –
5.21	7.70
7.70	5.21
– – –	– – –
7.70	5.21
40.11	1.00

Ερμηνεία του πίνακα:

Η πιθανότητα να δοθεί απόφαση για "άτυπη πλακώδης υπερπλασία" σε σχέση με "αρνητική" του παθολόγου Y είναι 5.21 φορές μεγαλύτερη από το να δοθεί απόφαση για "άτυπη πλακώδης υπερπλασία" σε σχέση με την "αρνητική" από τον παθολόγο Z, για σταθερά επίπεδα του παθολόγου A. Επίσης, να σημειωθεί ότι επειδή τα odds ratios στη

κύρια διαγώνιο διαφέρουν από τη μονάδα, οι αποφάσεις των παθολόγων είναι περισσότερο ομοειδής από ότι ένα επίπεδο πάνω στη κατηγορία “καρκίνωμα in Situ”. Συνεπώς, υπάρχει συμφωνία μεταξύ των αποφάσεων τους.

Κεφάλαιο 6: Μέτρα συμφωνίας στην R

6.1 Εισαγωγή

Στο σημείο αυτό, θα παρουσιαστούν κάποιες εντολές στην R, χρήσιμες για κάθε αναγνώστη. Στα κεφάλαια 2, 3, 4 τα παραδείγματα που δόθηκαν πραγματοποιήθηκαν μεσω Excell (Gwet 2014). Ένα πιο εύχρηστο εργαλείο όπως η R θα μας δώσει πιο άμεσα τα εξής μέτρα συμφωνίας αξιολογητών:

- k του Cohen
- Pi του Scott
- AC_1 του Gwet
- Alpha του Krippendorff

Η επιπρόσθετη δυνατότητα που μας δίνει η R είναι ότι μπορεί να μας δώσει 95% διαστήματα εμπιστοσύνης (ή σε όποιο επίπεδο εμείς επιθυμούμε), καθώς και μπορούμε να προσαρμόσουμε τα διάφορα βάρη όπως συζητήθηκαν στη παράγραφο 4.5. Θα χρησιμοποιηθούν τα δεδομένα του πίνακα 22 της παραγράφου 5.5.

6.2 Περίπτωση 2 βαθμολογητών

Code

```
ratings=as.table(rbind(c(38,5,0,1),c(33,11,3,0),c(10,14,5,6),c(3,7,3,10)))
categories=c("Certain","Probable","Possible","No")
dimnames(ratings)=list(New_Orlean_Neurologist=categories,
Winnipeg_Neurologist=cat  egories)
ratings

library(irrCAC)
kappa2.table(ratings,weights=diag(ncol(ratings)),conflev=0.95,N=Inf)
scott2.table(ratings,weights=diag(ncol(ratings)),conflev=0.95,N=Inf)
gwet.ac1.table(ratings,weights=diag(ncol(ratings)),conflev=0.95,N=Inf)
krippen2.table(ratings,weights=diag(ncol(ratings)),conflev=0.95,N=Inf)
quadratic.weights(1:4)
kappa2.table(ratings,quadratic.weights(1:4))
scott2.table(ratings,quadratic.weights(1:4))
gwet.ac1.table(ratings,quadratic.weights(1:4))
krippen2.table(ratings,quadratic.weights(1:4))
```

Συγκεντρωτικά Αποτελέσματα

Table 1: Unweighted kappa Coefficients

Coeff. name	Coeff. val	Coeff. se	Coeff. ci	Coeff. pval
Cohen's Kappa	0.2079425	0.05045537	(0.108,0.308)	6.249e-05
Scott's Pi	0.1782377	0.05651824	(0.067,0.29)	1.953e-03
Gwet's AC_1	0.2577797	0.05441219	(0.15,0.365)	5.026e-06
Krippendorff's Alpha	0.1809953	0.05651824	(0.069,0.293)	1.669e-03

Table 2: Quadratic Weights

1.0000000	0.8888889	0.5555556	0.0000000
0.8888889	1.0000000	0.8888889	0.5555556
0.5555556	0.8888889	1.0000000	0.8888889
0.0000000	0.5555556	0.8888889	1.0000000

Table 3: Quadratic weights coefficients

Coeff. name	Coeff. val	Coeff. se	Coeff. Ci	Coeff. pval
Cohen's Kappa	0.5245765	0.0600551	(0.406,0.643)	4.885e-15
Scott's Pi	0.4969858	0.06870114	(0.361,0.633)	2.343e-11
Gwet's AC_1	0.6220919	0.05529571	(0.513,0.731)	0e+00
Krippendorff's Alpha	0.4986737	0.06870114	(0.363,0.634)	2.049e-11

6.3 Περίπτωση 3 ή περισσότερων βαθμολογητών

Τέλος, θα παρουσιαστεί ο κώδικας και τα αποτελέσματα κάνοντας χρήση του παραδείγματος με την αξιολόγηση των Stickleback fishes από 4 αξιολογητές όπως παρουσιάζονται στον Πίνακα 10 της παραγράφου 2.2. Θα χρησιμοποιηθούν οι δείκτες του Fleiss, του Gwet και του Krippendorff.

Code

```
ratings2=matrix(c(0,0,0,0,4,2,0,2,0,0,0,0,0,0,4,2,0,2,0,0,0,0,0,1,3,1,1,2,0,0,3,0,1,0,0,3,0,1,
0,0,0,0,2,2,0,3,0,1,0,0,0,0,0,0,4,4,0,0,0,0,4,0,0,0,0,4,0,0,0,0,0,0,3,1,0,1,0,2,1,0,0,0,0,2,2,0,
0,0,0,4,0,0,3,0,1,0,1,3,0,0,0,0,1,0,3,0,0,3,1,0,4,0,0,0,0,4,0,0,0,0,2,0,2,0,0,1,0,3,0,0,2,0,2,0,
0,2,0,2,0,0,0,1,2,0,1),ncol=5,byrow=T)

fleiss.kappa.dist(ratings2)
gwet.ac1.dist(ratings2)
krippen.alpha.dist(ratings2)
quadratic.weights(1:5)
fleiss.kappa.dist(ratings2, quadratic.weights(1:5))
gwet.ac1.dist(ratings2,quadratic.weights(1:5))
krippen.alpha.dist(ratings2,quadratic.weights(1:5))
```

Συγκεντρωτικά αποτελέσματα

Table 4: Quadratic Weights

1.0000	0.9375	0.7500	0.4375	0.0000
0.9375	1.0000	0.9375	0.7500	0.4375
0.7500	0.9375	1.0000	0.9375	0.7500
0.4375	0.7500	0.9375	1.0000	0.9375
0.0000	0.4375	0.7500	0.9375	1.0000

Table 5: Unweighted kappa coefficients

coeff.name	coeff	Stderr	conf.int	p.value	p_a	p_e
Fleiss	0.4103475	0.07867581	(0.249,0.572)	1.538146e05	0.5804598	0.2884958
Gwet	0.4896874	0.06941578	(0.347,0.632)	1.129416e07	0.5804598	0.177876
Krippendorff	0.4154307	0.07769675	(0.256,0.575)	1.075314e05	0.5840765	0.2884958

Table 6: Quadratic kappa coefficients

coeff.name	coeff	stderr	conf.int	p.value	p_a	p_e
Fleiss	0.7337819	0.06692514	(0.597,0.871)	1.214495e-11	0.9206178	0.7018152
Gwet	0.7615899	0.04026596	(0.679,0.844)	0	0.9206178	0.6670352
Krippendorff	0.7360769	0.05459699	(0.624,0.848)	9.103829e-14	0.9213021	0.7018152

Συμπεράσματα Εργασίας

Στην παρούσα εργασία έγινε κριτική ανασκόπηση διάφορων μέτρων συμφωνίας και επεκτάσεις του γνωστότερου στο ευρύ κοινό μέτρου kappa του Cohen. Έγινε αναφορά σε γνωστούς δείκτες, σε μη παραμετρικό τεστ καθώς και σε log-linear models. Η απόφαση για το ποια μέθοδο να επιλέξουμε εξαρτάται από πολλούς παράγοντες. Συνοπτικά, θα υποστηρίζαμε πως αν θέλουμε μια γρήγορη και εύκολη απόφαση για την αξιολόγηση της συμφωνίας είναι προτιμότερο να επιλέξουμε το sequential test. Ειδάλλως, μπορούμε να επιλέξουμε από τα μέτρα συμφωνίας που προκύπτουν από τη βιβλιογραφία. Από αυτά, θα πρέπει να ξεκαθαρίσουμε για το είδος των βαθμολογίων δηλαδή αν είναι nominal, ordinal, interval ή ratio καθώς και για το πλήθος των αξιολογητών. Αν τα δεδομένα μας είναι nominal με δύο βαθμολογητές, το ποιο διαδεδομένο είναι το kappa του Cohen. Εναλλακτικά μπορεί να χρησιμοποιηθεί το AC_1 του Gwet. Το Pi του Scott και το alpha του Krippendorff είναι ισοδύναμα. Στην περίπτωση των τριών ή περισσότερων αξιολογητών μπορεί να γίνει χρήση ακόμα και το kappa του Conger, παρά το γεγονός ότι κι αυτό όπως του Cohen και του Scott έχει τα μειονεκτήματά του. Στην περίπτωση των ordinal δεδομένων και των δύο ή και τριών βαθμολογητών η λογική είναι η ίδια. Δηλαδή, είναι καλύτερα να επιλέξουμε το δείκτη AC_1 καθώς είναι πιο ανθεκτικό σε παράδοξα συγκριτικά με τα προαναφερθέντα μέτρα. Αν τα δεδομένα μας είναι interval είτε ratio προτείνεται μια λίστα με όλα τα βάρη που μπορούμε να επιλέξουμε έτσι ώστε να έχουμε επιθυμητά αποτελέσματα. Τέλος, πέρα από τα μέτρα συμφωνίας προτείνονται και τα log-linear μοντέλα που μπορούν να χρησιμοποιηθούν. Ασφαλώς, η χρήση μοντέλων απαιτεί και γνώσεις στατιστικής που δεν είναι πάντα εύκολο για το ευρύ κοινό.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Αντζουλάκος, Δ. (2019-2020). *Πανεπιστημιακές σημειώσεις «Ανάλυση Δεδομένων με τη χρήση Στατιστικών Πακέτων»*

Μπερσίμης, Ν. (2020-2021). *Πανεπιστημιακές σημειώσεις «Βιοστατιστική»*

Πολίτης, Κ. (2020-2021). *Πανεπιστημιακές σημειώσεις «Γενικευμένα Γραμμικά Μοντέλα»*

Ξένη

Agresti, A (1988). "*A model Agreement Between Ratings on an Ordinal Scale." Biometrics, **44**, 539-548*

Berry, K. J., and Mielke, Jr., P. W. (1988), "*A Generalization of Cohen's Kappa Agreement Measure To Interval Measurement and Multiple Raters", Educational and Psychological Measurement, **48**, 921-933.*

Bersimis S., Sachlas A., Chakraborti S., (2017). *"A seuantial test for assessing observed agreement between raters.", Biometrical Journal, 60, 128-145.*

Cicchetti, D. V., and Feinstein, A. R. (1990). "*High Agreement but low Kappa: II. Resolving the paradoxes." Journal of Clinical Epidemiology, **43**, 551-558*

Cohen, J. (1968). " *Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit." Psychological Bulletin, 70, 213-220.*

Cohen. J. (1960). " *A coefficient of agreement for nominal scales." Educational and Psychological Measurement, **20**, 37-46.*

Conger, A. J. (1980), *" Integration and Generalization of Kappas for Multiple Raters," Psychological Bulletin, **88**, 322-328.*

Daniel A. Bloch and Helena Chmura Kraemer (1989). *"2 × 2 Kappa Coefficinets: Measures of Agreement or Association" Biometrics, **45**, 269-287.*

Everitt, B. S. (1992). *"Boostrap methods: another look at the jackknife." Annals of statistics, **7**, 1-26*

Feinstein, A. R., and Cicchetti, D, V, (1990), *"High agreement but low kappa: I. The problems of two paradoxes," Journal of Clinical Epidemiology, 43, 543-549.*

Fleiss, J. L. (1971). *"Measuring nominal scale agreement among many raters", Psychological Bulletin, **76**, 378-382.*

Fleiss, J. L., and Davies, M. (1982). *"Jackknifing Functions of Multinomial Frequencies, with an Application to a Measure of Concordance," American Journal of Epidemiology, 115, 841-845.*

Fleiss, J. L., Cohen, J., and Everitt, B. S. (1969). *"Large sample standard errors of kappa and weighted kappa," Psychological Bulletin, **72**, 323-327.*

Gwet, K. L. (2008a). *"Computing inter-rater reliability and its variance in the presence of high agreement. " British Journal of Mathematical and Statistical Psychology,* ***61****, 29-48.*

Gwet, K. L. (2008b). *"Variance estimation of nominal – scale inter – rater reliability with random selection of raters." British Journal of Mathematical and Statistical Psychology,* ***61****, 29-48.*

Gwet, K. L. (2010a). *How to compute Intraclass Correlation Using Excel: A Practical Guide of Inter-Rater Reliability Assessment for Quantitative Data, Advanced Analytics, LLC.*

Gwet, K. L. (2010b). *The Practical Guide to Statistics: Basic Conscepts, Methods, and Meaning, Advanced Analytics, LLC.*

Gwet, K. L. (2014). *Handbook of Inter-Rater Reliability, "The Definitice Guide to Measuring the Extent of Agreement Among Raters., 4th Edition.*

Holley, J.W., and Guilford, J. P. (1964), *"A note on the G index of agreement." Educational and Psychological Measurement,* ***24****, 749-753.*

Janson, H., and Olsson, U. (2001). *"A Measure of Agreement for Interval or Nominal Multivariate Observations, " Educational and Psychological Measurement,* ***61****, 277-289.*

Janson, H., and Olsson, U. (2004). *"A Measure of Agreement for Interval or Nominal Multivariate Observations by Different Sets of Judges," Educational and Psychological Measurement,* ***64****, 62-70.*

Kastango, K. B. (2006). *"Assessing agreement among raters and identifying atypical raters using a log-linear modeling approach.", Doctor of Philosophy, University of Pittsburg Graduate School of Public Health, Department of Biostatistics, Pitssburgh, 2006.*

Kraemer, H.C. (1997). *"What is the right statistical measure of twin concordance (or diagnostic reliability and validity)?. Arch. Gen. Psychiatric,* ***54****, 1121-1124.*

Krippendorff, K. (1970). *"Estimatind the reliability, systematic error, and random error of interval data," Educational an Pshychological Measurement,* ***30****, 61-70.*

Krippendorff, K. (2012). *Content analysis: An Indroduction to its Methodology, 3rd. Edition, Thousans Oaks, CA: SAGE Publications, Inc.*

Landis, J. R and Koch G (1977a). *"The measurement of observer agreement for categorical dara," Biometrics,* ***33****, 159-174.*

Landis, R.J., and Koch G.G. (1977b). *"An application od Hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers", Biometrics, 33, 363-374.*

Likert, R. (1932). *"A Technique for the Measurement of Attitudes." Archives of Psychology,* ***140****, 1-55.*

Mary L. McHugh (2012). "*Interrater reliability: the kappa statistic", **22**, 276-282*

Scott, W. A. (1955). *"Reliability of content analysis: the case of nominal scale coding." Public Opinion Quarterly, **XIX**, 321-325.*

Tanner, M. A. and Young, M. A. (1985b). *"Modeling ordinal scale disagreement", Psychological Bulletin, **98**(2), 408-415.*

Tanner, M.A and Young, M.A. (1985a). *"Modeling agreement among raters, " Journal of American Statistical Association, **80**, 175-180.*

Yilmaz A. E. (2017). *"Assessing Agreement between Raters from the Point of Coefficinents and Log-Linear Models.", Journal of Data Science, 15(1), 1-24.*

Yilmaz, A. E. (2013). "*Association models with agreement parameter for square contingency tables with ordered categories." , Master of Science, Hacettepe University, Department of Statistics, Ankara.*

Zwick, R. (1988). *"Another look at interrater agreement.", Psychological Bulletin, **103**(3), 374-378, 1988.*

Zwick, Rebecca (1986). "*Another Look at Inter-Rater Agreement. Research Report", **2**, 1-23*