ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Γενικά

Η μελέτη της συμφωνίας αξιολογητών (βαθμολογητών, ερευνητών, ιατρών κλπ.) είναι ένα πεδίο στην ανάλυση κατηγορικών δεδομένων με ενδιαφέρουσες εφαρμογές στην ιατρική (κλινικές μελέτες, επιδημιολογία, χειρουργική κλπ.), στη ψυχολογία (περιπτώσεις κατάθλιψης, σχιζοφρένειας κλπ.) και στις κοινωνικές επιστήμες (δημογραφία, εκπαίδευση, κοινωνιολογία κλπ.).

Σε πολλές μελέτες αντικείμενα (ή υποκείμενα ή ασθενείς κλπ.) ταξινομούνται σε κατηγορίες από δύο βαθμολογητές. Για παράδειγμα γιατροί ταξινομούν ασθενείς ανάλογα με το στάδιο της ασθένειάς τους ή ανώτερα στελέχη επιχειρήσεων αξιολογούν την επίδοση και δραστηριότητα των υπαλλήλων τους χρησιμοποιώντας μια διακριτή κλίμακα.

Ενδιαφέρον επίσης παρουσιάζουν οι περιπτώσεις όπου περισσότεροι των δύο ερευνητών ταξινομούν τα αντικείμενα προς αξιολόγηση, οπότε και η εξέταση της συμφωνίας μεταξύ πολλών βαθμολογητών γίνεται πιο πολύπλοκη. Συναντάται επίσης το φαινόμενο κάθε αντικείμενο να αξιολογείται και να ταξινομείται από διαφορετικούς αριθμούς βαθμολογητών ή και σε περισσότερες από μία κατηγορίες. Δηλαδή ο ίδιος βαθμολογητής να ταξινομεί το αντικείμενο σε δύο ή και περισσότερες κατηγορίες. Στις περιπτώσεις αυτές, όπως γίνεται αντιληπτό, η ανάλυση της συμφωνίας γίνεται ακόμη πιο πολύπλοκη, είτε αυτή πραγματοποιείται με χρήση μέτρων είτε με χρήση μοντέλων.

Για τη μελέτη της συμφωνίας μεταξύ δύο βαθμολογητών, καταλήγουμε στη μελέτη πινάκων διδιάστατων (γιατί θεωρούμε δύο βαθμολογητές) και τετραγωνικών IxI (αν θεωρήσουμε ότι η κλίμακα αξιολόγησης αποτελείται από I κατηγορίες). Σχηματικά, ο πίνακας συχνοτήτων είναι ο Πίνακας 1.1. Αυτό που αναμένουμε είναι στα κελιά της κυρίας διαγωνίου οι παρατηρούμενες συχνότητες να είναι αυξημένες γιατί πρόκειται για τα κελιά της πλήρους συμφωνίας, ενώ όσο απομακρυνόμαστε από την κύρια διαγώνιο, δηλαδή στα πάνω δεξιά και στα κάτω αριστερά κελιά του πίνακα συχνοτήτων αναμένουμε μικρές έως μηδενικές συχνότητες.

Αν στη μελέτη λαμβάνουν μέρος περισσότεροι των δύο βαθμολογητών, έστω R>2, τότε αναφερόμαστε σε R-διάστατους ($IxIx..xI$) πίνακες συχνοτήτων.

1 4 2 43
R

Πίνακας 1.1

Υπόδειγμα πίνακα συχνοτήτων για μελέτη της συμφωνίας μεταξύ δύο βαθμολογητών

		Α βαθμολογητής			
		1η κατηγορία	2η κατηγορία	…	$I^{η}$ κατηγορία
Β βαθμολογητής	1η κατηγορία				
	2η κατηγορία				
	…				
	$I^{η}$ κατηγορία				

Στην προσπάθεια ανάλυσης της συμφωνίας μεταξύ βαθμολογητών, βοηθητικό και επεξηγηματικό ως προς τα συμπεράσματα της ανάλυσης ρόλο παίζει η μελέτη της διαφωνίας μεταξύ των βαθμολογητών.

Οι βαθμολογητές, αν και ταξινομούν τα αντικείμενα στις διάφορες κατηγορίες με βάση κάποιες αρχές, κριτήρια και νόμους, προβαίνουν σε καταχωρήσεις από τις οποίες δεν μπορεί να εκλείψει πλήρως το υποκειμενικό στοιχείο (προσωπικές εμπειρίες, διαφορετική ικανότητα αντίληψης γεγονότων και καταστάσεων και άλλοι εσωτερικοί και εξωτερικοί παράγοντες). Θα πρέπει, δηλαδή, να εξετάζεται το κατά πόσο οι αξιολογήσεις των βαθμολογητών στηρίζονται στα ίδια ή διαφορετικά κριτήρια και να εκτιμάται η μεροληψία τους, η τάση π.χ. κάποιων βαθμολογητών να κάνουν υψηλότερες ή χαμηλότερες αξιολογήσεις από τους άλλους. Και τέλος, πρέπει να λαμβάνεται υπόψη το πώς ερμηνεύει και αντιλαμβάνεται κάθε κατηγορία ταξινόμησης κάθε βαθμολογητής.

Σημαντικό είναι η κλίμακα αξιολόγησης, βάσει της οποίας γίνονται οι καταχωρήσεις από τους βαθμολογητές, να είναι σωστά κατασκευασμένη ώστε να αποτρέπει στο μέτρο του δυνατού τα σφάλματα. Χαρακτηριστικά, η εκτίμηση της αξιοπιστίας ενός συστήματος – κλίμακας αξιολόγησης έχει προσεγγιστεί από την οπτική γωνία της συμφωνίας των αξιολογητών που τη χρησιμοποιούν. Με άλλα λόγια, ο έλεγχος της αντικειμενικότητας μιας κλίμακας αξιολόγησης γίνεται και μέσω των επιπέδων συμφωνίας των βαθμολογητών που τη χρησιμοποιούν. Η εκτενής ενασχόληση με κλίμακες αξιολόγησης είναι εκτός των πλαισίων της παρούσας εργασίας. Για λόγους πληρότητας, στο Παράρτημα παραθέτουμε κάποιες

κλίμακες αξιολόγησης καθώς και κάποια γενικά συνοπτικά στοιχεία κατασκευής και διάκρισής τους.

1.2 Σύντομη ιστορική αναδρομή

Οι πρώτες προσπάθειες προσέγγισης και εξέτασης της συμφωνίας μεταξύ βαθμολογητών εστίασαν στα παρατηρούμενα ποσοστά συμφωνίας (Goodman and Kruskal, 1954), οπότε σύμφωνα με τον τρόπο αυτό η τυχαία συμφωνία μπορούσε να αγνοηθεί. Εναλλακτικά, ο Cohen (1960) εισήγαγε έναν συντελεστή (τον kappa) για τη μέτρηση της συμφωνίας, της διορθωμένης από τυχαιότητες, σε ονομαστικές κλίμακες αξιολόγησης. Ανάλογα μέτρα συμφωνίας έχουν προταθεί από τους Cohen (1968), Fleiss et al. (1969), Fleiss (1971), Fleiss and Cohen (1973).

Πέρα από τον εντοπισμό και τη μέτρηση της συμφωνίας, το ενδιαφέρον των ερευνητών εστιάστηκε και στην ανάλυση της δομής της συμφωνίας που προκύπτει από τα δεδομένα, με αποτέλεσμα να κατασκευαστούν και να ορίσουν μοντέλα συμφωνίας [Birch (1963), Goodman (1972), Haberman (1974), Bishop et al (1975), Goodman (1978), Haberman (1978)].

Οι Fleiss et al. (1972) και η Kraemer (1980) μελέτησαν περιπτώσεις όπου κάθε βαθμολογητής μπορεί να κάνει περισσότερες από μία αξιολογήσεις. Επίσης, οι Landis and Koch (1977c) ασχολήθηκαν με το ενδεχόμενο κάθε αντικείμενο να ταξινομείται από διαφορετικό αριθμό βαθμολογητών.

Οι Tanner and Young (1985b) ασχολήθηκαν με λογαριθμογραμμικά μοντέλα συμφωνίας. Οι Agresti (1988) και Becker (1989, 1990) περιέγραψαν σχετικές μεθόδους βασισμένοι στα μοντέλα συνάφειας (association models) του Goodman (1986). Οι Darroch and McCloud (1986) εξέτασαν quasi-συμμετρικά μοντέλα συμφωνίας, ενώ οι Agresti and Lang (1993) ασχολήθηκαν με μια μέθοδο που συνδυάζει quasi-συμμετρικά και μοντέλα λανθάνουσας τάξης (latent class).

Τα μοντέλα λανθάνουσας τάξης για πολλαπλούς βαθμολογητές έχουν προσελκύσει το ενδιαφέρον των ερευνητών. Οι Gelfand and Solomon (1975), Walter and Irwig (1988), Espeland and Handleman (1989), Uebersax and Grove (1990) μεταξύ άλλων έχουν ασχοληθεί με τέτοια μοντέλα για αξιολογήσεις της μορφής 0 – 1. Οι Uebersax and Grove (1989, 1993)

πρότειναν ένα μοντέλο λανθάνουσας κατανομής για πολλαπλούς βαθμολογητές για διατάξιμα κατηγορικά δεδομένα.

1.3 Διάρθρωση της Εργασίας

Στο δεύτερο κεφάλαιο της εργασίας περιγράφονται κάποια γνωστά μέτρα συμφωνίας μεταξύ δύο βαθμολογητών, όπως το kappa του Cohen, ο εντός των τάξεων συντελεστής συσχέτισης καθώς και ο τετραχωρικός συντελεστής συσχέτισης. Ενδιαφέρον παρουσιάζουν κάποιες γενικεύσεις των μέτρων συμφωνίας για ζεύγη δεδομένων, για περιπτώσεις μελετών όπου υπάρχουν συμμεταβλητές, και για περιπτώσεις όπου τα δεδομένα προέρχονται από πολυκεντρική κλινική δοκιμή.

Στο τρίτο κεφάλαιο αναφέρονται αναλυτικά μοντέλα για τη μελέτη της συμφωνίας μεταξύ δύο βαθμολογητών. Περιγράφεται το μοντέλο ανεξαρτησίας, συμμετρίας και οι αντίστοιχες quasi μορφές τους, το μοντέλο επιδράσεων γραμμών και στηλών, καθώς και το μοντέλο του Agresti, των Tanner and Young και εισάγεται και ένα νέο λογαριθμογραμμικό μοντέλο. Ένα παράδειγμα με δεδομένα των Graham and Jackson (1993) μας βοηθά να κατανοήσουμε τα προηγούμενα μοντέλα και το επίπεδο προσαρμογής τους στα δεδομένα.

Στο τέταρτο κεφάλαιο παρουσιάζουμε μέτρα και μοντέλα για τη μελέτη της συμφωνίας περισσότερων από δύο βαθμολογητών. Πιο συγκεκριμένα μελετάται η περίπτωση όπου η αξιολόγηση έχει δύο κατηγορίες ταξινόμησης και διαφορετικός αριθμός βαθμολογητών ταξινομεί στις δύο αυτές κατηγορίες τα αντικείμενα και στη συνέχεια η περίπτωση πολλών βαθμολογητών που ταξινομούν τα αντικείμενα σε πολλαπλές κατηγορίες ταξινόμησης. Στη συνέχεια παρουσιάζονται κάποια μοντέλα για τη μελέτη της συμφωνίας μεταξύ πολλών βαθμολογητών. Τέλος μελετάμε τη διαφωνία μεταξύ πολλών βαθμολογητών, επικεντρώνοντας το ενδιαφέρον στην αμεροληψία μεταξύ τους δηλαδή στη συστηματική τους ή όχι διαφωνία. Ακολουθεί ένα παράδειγμα για την ανάλυση της διαφωνίας σε ψυχιατρικά δεδομένα (Fleiss (1971), Landis and Koch (1977c), Kraemer (1980)).

Στο πέμπτο κεφάλαιο της εργασίας αναφέρουμε συνοπτικά κάποια θέματα σχετικά με τη συμφωνία μεταξύ βαθμολογητών, τα οποία έχουν απασχολήσει τα τελευταία χρόνια τους ερευνητές. Τα θέματα αυτά αφορούν γενικεύσεις μέτρων και επεκτάσεις μοντέλων, επαναλαμβανόμενες μετρήσεις, κλίμακες ταξινόμησης, μέγεθος δείγματος κλπ.

ΚΕΦΑΛΑΙΟ 2

Βασικά μέτρα συμφωνίας

2.1 Εισαγωγή

Η αρχική προσπάθεια μελέτης της συμφωνίας μεταξύ βαθμολογητών ήταν να υπολογιστεί απλά και μόνο το ποσοστό στο οποίο οι βαθμολογητές συμφωνούσαν. Βέβαια η προσπάθεια αυτή γρήγορα φάνηκε ανεπαρκής, εφόσον δεν εξετάζεται το ενδεχόμενο κάποιες από τις περιπτώσεις συμφωνίας να οφείλονται μόνο σε τυχαίους παράγοντες. Άλλη μελέτη της συμφωνίας μεταξύ βαθμολογητών βασίστηκε στο X^2 στατιστικό μέτρο που μπορεί να υπολογιστεί σε κάθε πίνακα συνάφειας. Και αυτό βέβαια κρίθηκε μη ικανοποιητικό εφόσον όταν εφαρμόζεται σε πίνακες συνάφειας αποτελεί μέτρο συσχέτισης των δύο βαθμολογητών και όχι απαραίτητα μέτρο συμφωνίας – συσχέτιση μεταξύ των δύο βαθμολογητών μπορεί να προέρχεται από τυχαία απόκλιση μεταξύ των βαθμολογητών, είτε οφειλόμενη σε συμφωνία είτε σε ασυμφωνία. Ισχυρή συμφωνία απαιτεί ισχυρή συσχέτιση χωρίς όμως να ισχύει και το αντίστροφο. Αν ο Α βαθμολογητής συστηματικά κατατάσσει τα αντικείμενα προς βαθμολόγηση μία κατηγορία ψηλότερα από ότι ο Β βαθμολογητής, τότε η συμφωνία είναι φτωχή αν και η συσχέτιση είναι ισχυρή. Η ανάλυση σε τέτοιες περιπτώσεις εστιάζεται στην περιγραφή της έντασης της συμφωνίας και στην ανακάλυψη δομών ασυμφωνίας.

Το γνωστότερο μέτρο συμφωνίας είναι ο συντελεστής kappa του Cohen, που περιγράφεται στην παράγραφο 2.2, τόσο στην περίπτωση που τα δεδομένα προς ταξινόμηση λαμβάνονται σαν ισότιμα όσο και όταν λαμβάνεται υπόψη η διαφορετική βαρύτητα των κατηγοριών ταξινόμησης οπότε και καταλήγουμε στην παράγραφο 2.2.1 στον σταθμισμένο συντελεστή kappa του Cohen. Στην παράγραφο 2.3 γίνεται αναφορά στον εντός των τάξεων (*intraclass*) συντελεστή συσχέτισης που εφαρμόζεται σε δεδομένα της μορφής 0 – 1, ενώ στην παράγραφο 2.4 περιγράφεται ο τετραχωρικός συντελεστής συσχέτισης που αφορά δύο κατηγορίες ταξινόμησης συνεχούς μεταβλητής. Στην παράγραφο 2.5 αναφέρονται γενικεύσεις των μέτρων συμφωνίας στην περίπτωση πάντα δύο αξιολογητών. Έτσι στην παράγραφο 2.5.1 περιγράφεται ο συντελεστής kappa για ζεύγη δεδομένων, στην παράγραφο 2.5.2 η προσοχή εστιάζεται στην αντιμετώπιση περιπτώσεων που υπάρχουν συμμεταβλητές, ενώ στην παράγραφο 2.5.3 περιγράφεται η μελέτη του συντελεστή kappa όταν τα δεδομένα

προέρχονται από μια πολυκεντρική κλινική δοκιμή. Τέλος, στην παράγραφο 2.6 αναφέρονται εφαρμογές και παραδείγματα των μέτρων που περιγράφηκαν στις προηγούμενες παραγράφους.

Για τη διευκόλυνση στο συμβολισμό, στην περιγραφή και ανάλυση των προαναφερόμενων μέτρων συμφωνίας, ας θεωρήσουμε τον δισδιάστατο πίνακα συνάφειας *I*x*I*. Έστω n_{ij}, $i, j = 1,2,\ldots I$, η παρατηρούμενη συχνότητα στο κελί (i,j) και m_{ij}, $i,j = 1,2,\ldots I$, η αναμενόμενη συχνότητα στο κελί (i,j) κάτω από το μοντέλο που θεωρώ. Το δειγματικό ποσοστό του κελιού (i,j) είναι $p_{ij} = n_{ij}/n$, όπου n είναι ο συνολικός αριθμός των παρατηρήσεων στον πίνακα συνάφειας, ενώ π_{ij} είναι η πιθανότητα στο (i,j) κελί.

2.2 Συντελεστής kappa του Cohen

Πρόκειται για ένα μέτρο συμφωνίας που παρουσίασε ο Scott (1955) και επεκτάθηκε από τον Cohen (1960) και έμεινε γνωστό σαν μέτρο kappa του Cohen. Το μέτρο αυτό βασίζεται στην ιδέα ότι μεταξύ των παρατηρούμενων περιπτώσεων συμφωνίας υπάρχουν και κάποιες οι οποίες τυχαία και μόνο τυχαία παρουσιάζουν συμφωνία.

Υποθέτουμε, λοιπόν, ότι δύο αξιολογητές βαθμολογούν n αντικείμενα με βάση μια κλίμακα I κατηγοριών, από το 1 έως το I, ο ένας ανεξάρτητα από τον άλλο.

Έτσι κατασκευάζεται ένας διδιάστατος *I*x*I* πίνακας συνάφειας με άγνωστες περιθώριες κατανομές για κάθε έναν από τους δύο αξιολογητές. Αν p_{ij} είναι το δειγματικό ποσοστό των αντικειμένων στο (i,j) κελί, το ποσοστό δηλαδή των αντικειμένων που βαθμολογήθηκαν στην i-κατηγορία από τον πρώτο βαθμολογητή και στην j-κατηγορία από τον δεύτερο, τότε $p_{i.} = \sum_{j=1}^{I} p_{ij}$ είναι το ποσοστό των αντικειμένων που βαθμολογήθηκαν στην i-κατηγορία (i-γραμμή) από τον πρώτο βαθμολογητή. Αντίστοιχα $p_{.j} = \sum_{i=1}^{I} p_{ij}$ είναι το ποσοστό των αντικειμένων που βαθμολογήθηκαν στην j-κατηγορία (j-στήλη) από τον δεύτερο βαθμολογητή.

Αν με $p_o = \sum_{i=1}^{I} p_{ii}$ θεωρήσουμε το παρατηρηθέν ποσοστό συμφωνίας και με $p_c = \sum_{i=1}^{I} p_{i.} p_{.i}$ το ποσοστό συμφωνίας που οφείλεται σε τυχαίους παράγοντες, τότε το μέτρο kappa του Cohen δίνεται από τον τύπο $\hat{k} = \frac{p_o - p_c}{1 - p_c}$.

Το μέτρο kappa του Cohen θεωρείται επέκταση του δείκτη που αρχικά είχε εισάγει ο Scott. Ο Scott καθόρισε το p_c υποθέτοντας ότι η κατανομή των ποσοστών στις *I* κατηγορίες είναι γνωστή για τον πληθυσμό και κοινή και για τους δύο βαθμολογητές. Τότε αν όντως πρόκειται για ίδιες περιθώριες κατανομές, στον πίνακα που κατασκευάσαμε πιο πάνω, τα μέτρα του Scott και Cohen ταυτίζονται.

Για να προσδιορίσουμε αν ο $\hat{k}$ είναι στατιστικά σημαντικός, μπορούμε να χρησιμοποιήσουμε τον τύπο της ασυμπτωτικής διακύμανσης των Fleiss et al. (1969) για τον *I*x*I* πίνακα. Αν ο αριθμός των παρατηρούμενων αντικειμένων *n* είναι μεγάλος, τότε ο παραπάνω τύπος είναι ισοδύναμος της ακριβούς διακύμανσης του Everitt (1968) που βασίστηκε στην κεντρική υπεργεωμετρική κατανομή. Κάτω από την υπόθεση της τυχαίας συμφωνίας, ισχύει

$$\widehat{\mathrm{var}}_0(\hat{k}) = \frac{p_c + p_c^{\,2} - \sum_{i=1}^{I} p_{i.} p_{.i} (p_{i.} + p_{.i})}{n(1 - p_c)^2}.$$

Υποθέτοντας, επίσης, ότι $\frac{\hat{k}}{\sqrt{\widehat{\mathrm{var}}_0(\hat{k})}}$ ακολουθεί την κανονική κατανομή μπορούμε να ελέγξουμε την υπόθεση της τυχαίας συμφωνίας με αναφορά στην τυπική κανονική κατανομή. Βέβαια η εξέταση της τυχαιότητας δεν παρουσιάζει ενδιαφέρον, στο πεδίο της έρευνας της αξιοπιστίας μεθόδων ή αξιολογητών, γιατί γενικά οι αξιολογητές εκπαιδεύονται να είναι αξιόπιστοι. Έτσι, στην περίπτωση αυτή ταιριάζει ένα μικρότερο όριο του kappa, που δίνεται από τους Fleiss et al. (1969) με ασυμπτωτική έκφραση:

$$\widehat{\mathrm{var}}(\hat{k}) = \frac{1}{n(1 - p_c)^2} \left\{ \sum_{i=1}^{I} p_{ii} [1 - (p_{i.} + p_{.i})(1 - \hat{k})]^2 + (1 - \hat{k})^2 \sum_{i \neq j}^{I} p_{ij} (p_{i.} + p_{.j})^2 - [\hat{k} - p_c (1 - \hat{k})]^2 \right\}$$

Μελέτη της ακρίβειας του τυπικού σφάλματος του k' για μεγάλα δείγματα δίνεται και από τους Cicchetti και Fleiss (1977) και από τους Fleiss και Cicchetti (1978) με χρήση Monte Carlo προσομοίωσης.

Δεν υπάρχουν αυστηρά όρια χαρακτηρισμού της έντασης της συμφωνίας μεταξύ βαθμολογητών. Σύμφωνα με τους Landis και Koch (1977a) μπορούμε να χαρακτηρίσουμε το επίπεδο συμφωνίας μεταξύ δύο βαθμολογητών με βάση το δείκτη κ του Cohen ως εξής:

- $0,75 < k'$, πολύ καλή συμφωνία
- $0,40 < k' < 0,75$, καλή συμφωνία
- $k' < 0,40$, φτωχή συμφωνία.

Η χρήση και η ερμηνεία του kappa του Cohen προκάλεσαν αντιπαράθεση – και πιο συγκεκριμένα όσον αφορά 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 στον kappa.

Ένας άλλος παράγοντας που επηρεάζει τον συντελεστή kappa είναι το κατά πόσο μια τάση ταξινόμησης επικρατεί στον κάθε πληθυσμό. Οι ίδιοι βαθμολογητές μπορούν να καταλήξουν σε διαφορετικές τιμές του kappa όταν εξετάζουν δύο διαφορετικούς πληθυσμούς [Feinstein and Cicchetti (1990), Byrt et al (1993)]. Δηλαδή η συμφωνία των βαθμολογητών εξαρτάται και από τον πληθυσμό που εξετάζουν.

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

Μερικοί συγγραφείς (Hutchinson 1993) αντιμετωπίζουν το γεγονός ότι το kappa του Cohen εμπεριέχει δύο είδη διαφωνίας: τη διαφωνία που οφείλεται στην μεροληψία των βαθμολογητών και τη διαφωνία που προκύπτει από την διαφορετική αντίληψη περί κατάταξης των αντικειμένων μεταξύ των βαθμολογητών. Λύση αντιμετώπισης των παραπάνω διαφορών αποτελεί ο εντός των τάξεων kappa συντελεστής (Bloch και Kraemer 1989) που θα εξετάσουμε σε επόμενη ενότητα. Ωστόσο ο Zwick (1988) επισημαίνει ότι είναι προτιμότερο από το να αγνοούμε τις διαφορές στις περιθώριες κατανομές ή να προσπαθούμε να τις διορθώσουμε, οι ερευνητές να διευκρινίζουν αν οι

διαφορές αυτές οφείλονται σε σημαντική ασυμφωνία μεταξύ των βαθμολογητών ή κυρίως σε τυχαία σφάλματα. Έτσι, οποιοδήποτε αξιολόγηση της συμφωνίας μεταξύ των βαθμολογητών θα πρέπει να αρχίζει με έρευνα της περιθώριας ομοιογένειας.

2.2.1 Σταθμισμένος συντελεστής kappa του Cohen

Επειδή ο κ δεν λαμβάνει υπόψη τις διαφορές μεταξύ των κατηγοριών ταξινόμησης ή την απόσταση που έχουν μεταξύ τους, και τις αντιμετωπίζει σαν ισότιμες, τα αποτελέσματα συχνά δεν ανταποκρίνονται στην πραγματικότητα. Ο Cohen το 1968 εισάγει μια γενίκευση του κ, τον σταθμισμένο κ_w, ως μέτρο του ποσοστού της σταθμισμένης συμφωνίας διορθωμένου από τυχαίους παράγοντες. Αν θεωρήσουμε ότι w_{ij} δηλώνει το βάρος του (i,j) κελιού, $i, j = 1,\ldots I$, τότε ο κ_w δίνεται από τον τύπο:

$$\kappa_w = \frac{\sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}\, p_{ij} - \sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}\, p_{i.}\, p_{.j}}{1 - \sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}\, p_{i.}\, p_{.j}} .$$

Στις περιπτώσεις όπου $w_{ij}= 1$, για $i = j$ και $w_{ij}= 0\ \forall\ i \neq j$, τότε ο κ_w είναι ο συντελεστής κ που είδαμε στην προηγούμενη ενότητα. Επίσης, όταν $w_{ij} = \frac{1-(i-j)^2}{(I-1)^2}$, τότε ο κ_w μπορεί να ερμηνευθεί ως ένας εντός των τάξεων συντελεστής συσχέτισης για ανάλυση διακύμανσης κατά δύο παράγοντες, υπολογισμένος κάτω από την υπόθεση ότι τα n αντικείμενα και οι βαθμολογητές είναι τυχαία δείγματα από τους αντίστοιχους πληθυσμούς αντικειμένων και βαθμολογητών (Fleiss και Cohen 1973).

Οι Fleiss et al (1969) κατέληξαν στον τύπο της ασυμπτωτικής διακύμανσης για το κ_w. Ο τύπος που χρησιμοποιήθηκε για την κατασκευή διαστημάτων εμπιστοσύνης και για τεστ σημαντικότητας από τους Cicchetti και Fleiss (1977) και Fleiss και Cicchetti (1978). Βασισμένοι σε Monte Carlo μελέτες, αναφέρουν ότι μόνο μετρίου μεγέθους δείγματα είναι ικανά να εξετάσουν την υπόθεση ότι δύο κ_w προερχόμενοι από ανεξάρτητα δείγματα είναι ίσοι. Ωστόσο το ελάχιστο μέγεθος δείγματος που απαιτείται για τον καθορισμό διαστημάτων εμπιστοσύνης γύρω από μια τιμή του κ_w είναι $n = 16I^2$, το οποίο τις περισσότερες φορές είναι υπερβολικά μεγάλο.

2.3 Εντός των τάξεων συντελεστής kappa

Οι Bloch και Kraemer (1989) εισήγαγαν τον εντός των τάξεων συντελεστή συσχέτισης ως εναλλακτική εκδοχή του kappa του Cohen, υποθέτοντας ότι για κάθε βαθμολογητή ισχύει η ίδια περιθώρια κατανομή . Υποθέτουν ότι τα *n* δεδομένα ταξινομούνται σε δύο κατηγορίες από δύο σταθερούς βαθμολογητές. 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 $E(p_i) = P$, $P' = 1-P$ και $\text{var}(p_i) = \sigma_p^2$ και ο εντός των τάξεων συντελεστής kappa δίνεται από τον τύπο

$$k_I = \frac{s_p^2}{PP'} \qquad (1.1)$$

Πίνακας 2.1

Μοντελοποίηση για την εκτίμηση του εντός των τάξεων συντελεστή kappa.

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}	$P'^2 + k_I PP'$

Ένας εκτιμητής του εντός των τάξεων συντελεστή kappa προκύπτει από το μοντέλο πιθανοτήτων του Πίνακα 2.1 για την από κοινού απόκριση, με τον συντελεστή kappa σαφώς ορισμένο στο παραμετρικό μοντέλο. Έτσι ο λογάριθμος της συνάρτησης πιθανοφάνειας είναι:

$$\ln L(P, k_I / n_{11}, n_{12}, n_{21}, n_{22}) = n_{11}\ln(P^2 + k_I PP') + (n_{12} + n_{21})\ln[PP'(1-k_I)] + n_{22}\ln(P'^2 + k_I PP')$$

Οι εκτιμητές μεγίστης πιθανοφάνειας $\hat{P}$ και $\hat{k}_I$ για το P και k_I, αντίστοιχα, προκύπτουν ως:

$$\hat{P} = \frac{2n_{11} + n_{12} + n_{21}}{2n},$$

$$\hat{k}_I = \frac{4(n_{11}n_{22} - n_{12}n_{21}) - (n_{12} - n_{21})^2}{(2n_{11} + n_{12} + n_{21})(2n_{22} + n_{12} + n_{21})}, \qquad (1.2)$$

με

$$SE(\hat{k}_I) = \left[\frac{1-\hat{k}_I}{n} \left((1-\hat{k}_I)(1-2\hat{k}_I) + \frac{\hat{k}_I(2-\hat{k}_I)}{2\hat{p}(1-\hat{p})} \right) \right]^{\frac{1}{2}}. \qquad (1.3)$$

Ο εκτιμητής $\hat{k}_I$, ο εκτιμητής μέγιστης πιθανοφάνειας του k_I, που δίνεται από την (1.1), ταυτίζεται με τον εντός των τάξεων συντελεστή συσχέτισης για δεδομένα της μορφής 0 – 1. Αν υποθέσουμε ότι ο $\hat{k}_I$ κατανέμεται κανονικά με μέσο k_I και τυπικό σφάλμα SE($\hat{k}_I$), το 100(1-α)% διάστημα εμπιστοσύνης δίνεται από τον τύπο $\hat{k}_I \pm z_{1-\alpha/2} SE(\hat{k}_I)$,όπου $z_{1-\alpha/2}$ το 100(1-α) ποσοστιαίο σημείο της τυπικής κανονικής κατανομής. Το παραπάνω διάστημα εμπιστοσύνης έχει αξία για μελέτη μόνο σε πολύ μεγάλα δείγματα που δεν συναντάμε συχνά σε μελέτες για συμφωνία μεταξύ βαθμολογητών.

Οι Bloch και Kraemer (1989) απέδειξαν επίσης μια σταθεροποιημένη διακύμανση για τον $\hat{k}_I$ η οποία βελτιώνει την ακρίβεια για την εκτίμηση διαστημάτων εμπιστοσύνης, τον υπολογισμό της ισχύος ή την κατασκευή στατιστικών τεστ.

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

$$k_L = (\frac{1}{9}y_3^{\,2} - \frac{1}{3}y_2)^{\frac{1}{2}} (\cos\frac{q+2p}{3} + \sqrt{3}\sin\frac{q+2p}{3}) - \frac{1}{3}y_3$$

$$k_U = 2(\frac{1}{9}y_3^{\,2} - \frac{1}{3}y_2)^{\frac{1}{2}} \cos\frac{q+5p}{3} - \frac{1}{3}y_3, \ \pi = 3{,}14.$$

Όπου

$$q = \arccos\frac{V}{W}, \quad V = \frac{1}{27}y_3^3 - \frac{1}{6}(y_2y_3 - 3y_1), \quad W = (\frac{1}{9}y_3^2 - \frac{1}{3}y_2)^{\frac{3}{2}},$$

και

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

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

$$y_3 = \frac{n_{12} + n_{21} + \{1 - 2P(1-P)\}c^2_{1,1-a}}{P(1-P)\left(c^2_{1,1-a} + n\right)} - 1$$

Με τη μέθοδο αυτή η ακρίβεια, σε όλες τις τιμές τόσο του k_I όσο και του P, σε μικρά δείγματα, έχει βελτιωθεί.

Οι Donner και Eliasziw (1997) κατέληξαν σε γενίκευση της μεθόδου στην περίπτωση των τριών ή περισσοτέρων κατηγοριών ταξινόμησης για κάθε αντικείμενο. Η μέθοδός τους βασίζεται σε μία σειρά από στατιστικά ανεξάρτητες αναφορές, που η κάθε μία αναφέρεται σε δίτιμη μεταβλητή απόκρισης που προέρχεται συνδυάζοντας ένα ουσιώδες υποσύνολο των γνήσιων κατηγοριών.

2.4 Ο τετραχωρικός συντελεστής συσχέτισης (TTC)

Ο τετραχωρικός συντελεστής συσχέτισης χρησιμοποιείται στις περιπτώσεις όπου δύο είναι οι πιθανές καταστάσεις: π.χ. φυσιολογικό – μη φυσιολογικό, που απορρέουν από συνεχή μεταβλητή. Το πλαίσιο αυτό το συναντάμε συνήθως στις ιατρικές επιστήμες. Αναφέρουμε πιο κάτω χαρακτηριστικά του TTC συντελεστή:

1. Αγνοεί το γεγονός ότι οι δύο παρατηρητές διαφέρουν ως προς το «κατώφλι» τιμών του μη φυσιολογικού. Παρουσιάζει διαφορές στην οπτική γωνία ή τη στάση απόφασης, αν και υπάρχουν δεδομένα κριτήρια για το πώς θα κρίνεται μια τιμή φυσιολογική ή μη.
2. Η πιθανότητα για λανθασμένη ταξινόμηση σε σχέση με το «κατώφλι» τιμών εξαρτάται από την πραγματική τιμή της μεταβλητής. Όσο πιο απομακρυσμένη είναι η τιμή τόσο πιο βέβαιο είναι ότι θα ταξινομηθεί σωστά σαν φυσιολογική ή μη τιμή. Αν όμως βρίσκεται κοντά στο όριο που διαχωρίζει τις φυσιολογικές από τις μη τιμές τόσο πιο πιθανό είναι να μην ταξινομηθεί σωστά.

3. Επίσης, επειδή τα παραπάνω ισχύουν και για τους δύο βαθμολογητές, οι πιθανότητες της λανθασμένης ταξινόμησης δεν είναι ανεξάρτητες για τον καθένα. Για το λόγο αυτό, ο σταθμισμένος και μη kappa του Cohen, καθώς και ο intraclass kappa, δεν συνιστώνται για τις περιπτώσεις αυτές.

Όταν η διάγνωση αντιμετωπίζεται ως η απόφαση για «φυσιολογικό» ή «μη φυσιολογικό» μιας συνεχούς μεταβλητής και πιο συγκεκριμένα τυπικής κανονικής, ο TTC είναι ο κατάλληλος δείκτης για εξέταση συμφωνίας μεταξύ βαθμολογητών. Πιο συγκεκριμένα, ο TTC εκτιμά τη συσχέτιση μεταξύ των πραγματικών μη παρατηρούμενων μεταβλητών που χαρακτηρίζουν την πιθανότητα διάγνωσης ως «μη φυσιολογικό» κάθε βαθμολογητή και βασίζεται στην υπόθεση ότι αυτές ακολουθούν την διδιάστατη κανονική. Οι διαφορές του TTC με τους τύπου kappa συντελεστές δεν είναι μόνο ως προς τις περιπτώσεις που εφαρμόζεται αλλά και ως προς το γεγονός ότι εκτιμούν ποσοτικά δύο διαφορετικές, παρότι σχετιζόμενες οντότητες όπως αναφέρει η Kraemer (1997).

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

2.5 Επεκτάσεις και γενικεύσεις στην περίπτωση δύο αξιολογητών.

2.5.1 Ο συντελεστής kappa για ζεύγη δεδομένων

Ας υποθέσουμε ότι δύο βαθμολογητές ταξινομούν μαζί και το δεξί και το αριστερό μάτι *n* ασθενών, εξετάζοντας την παρουσία ή όχι μιας συγκεκριμένης ανωμαλίας στην όραση. Τα μέτρα συμφωνίας μεταξύ των δύο βαθμολογητών που θα χρησιμοποιηθούν στην περίπτωση αυτή θα πρέπει να επιτρέπουν η θετική συσχέτιση γενικά να παρουσιάζεται μεταξύ παρατηρήσεων που έγιναν στα «ταιριασμένα» όργανα του ίδιου ασθενή. Αν τα δεδομένα αυτά αντιμετωπιστούν σαν να προήλθαν από ένα τυχαίο δείγμα *2n* οργάνων, τότε οδηγούμαστε σε μη πραγματικά «στενά» διαστήματα εμπιστοσύνης για το k και σε πλαστά υψηλά επίπεδα απόρριψης της μηδενικής υπόθεσης H_0: $k=0$. Αν και συχνά τα παραπάνω προβλήματα αντιμετωπίζονται υπολογίζοντας διαφορετικά k για τα δύο όργανα, ωστόσο

εξακολουθεί και η μέθοδος αυτή να είναι ανεπαρκής και να μην είναι ακριβής ως προς την παρουσίαση των αποτελεσμάτων.

Ο Oden (1991) πρότεινε μία μέθοδο εκτίμησης ενός από κοινού *k* μεταξύ των δύο βαθμολογητών όταν και οι δύο βαθμολογητές αξιολογούν το ίδιο σύνολο ζευγών ματιών. Η μέθοδός του υποθέτει ότι οι πραγματικές τιμές των k για το δεξί και για το αριστερό μάτι είναι ίσες και κάνει χρήση των συσχετισμένων δεδομένων για να εκτιμήσει διαστήματα εμπιστοσύνης για το κοινό kappa. Ο από κοινού εκτιμητής του *k* είναι ο σταθμισμένος μέσος των kappa για τα δεξιά και αριστερά μάτια και δίνεται από τον τύπο:

$$\hat{k}_{pooled} = \frac{(1-\sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}\, r_{i.}\, r_{.j})\hat{k}_{right} + (1-\sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}\, l_{i.}\, l_{.j})\hat{k}_{left}}{(1-\sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}\, r_{i.}\, r_{.j}) + (1-\sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}\, l_{i.}\, l_{.j})},$$

όπου ρ_{ij} είναι το ποσοστό των ασθενών των οποίων το δεξί μάτι ταξινομήθηκε στην *i*-κατηγορία από τον πρώτο και στην *j*-κατηγορία από τον δεύτερο βαθμολογητή, λ_{ij} είναι το αντίστοιχο ποσοστό για το αριστερό μάτι και w_{ij} είναι ο συντελεστής συμφωνίας που δείχνει το επίπεδο συμφωνίας μεταξύ των δύο βαθμολογητών αν χρησιμοποιούν αντίστοιχα την *i* και *j* κατηγορία για το ίδιο μάτι.

Χρησιμοποιώντας τη μέθοδο δέλτα, ο Oden κατέληξε σε μία προσέγγιση του τυπικού σφάλματος για το $\hat{k}_{pooled}$. Ο $\hat{k}_{pooled}$ αποδείχτηκε σχεδόν αμερόληπτος (το μέσο σφάλμα, οφειλόμενο στις μεθόδους προσομοίωσης είναι της τάξης του 10^{-3}) κι έχει καλύτερη εφαρμογή τόσο από τον απλοϊκό εκτιμητή kappa του Cohen που αντιμετωπίζει τα δεδομένα σαν τυχαίο δείγμα από 2*n* μάτια, όσο και από τον εκτιμητή που βασίζεται στο κάθε μάτι ξεχωριστά, στα πλαίσια της correct coverage πιθανότητας του 95% διαστήματος εμπιστοσύνης για το πραγματικό kappa (Oden 1991).

Ο Schouten (1993) παρουσίασε μια εναλλακτική προσέγγιση του παραπάνω που παρουσιάζεται στον Πίνακα 2.2. Οι συμβολισμοί των τάξεων και η επεξήγησή τους δίνονται παρακάτω:

R+L+: παρουσία ανωμαλίας και στα δύο μάτια (ασθενές άτομο),

R+L-: παρουσία ανωμαλίας στο δεξί μόνο μάτι και όχι στο αριστερό,

R-L+: παρουσία ανωμαλίας μόνο στο αριστερό μάτι και όχι στο δεξί,

R-L-: απουσία ανωμαλίας και στα δύο μάτια (υγιές άτομο).

Ο Schouten χρησιμοποίησε το σταθμισμένο kappa για να καταλήξει σε ένα μέτρο συμφωνίας. Χρησιμοποίησε τα βάρη έτσι ώστε:

$$w_{ij} = \begin{cases} 0, & \text{όταν οι δύο βαθμολογητές διαφωνούν και στα δύο μάτια, πλήρης ασυμφωνία} \\ 0.5, & \text{όταν οι δύο βαθμολογητές διαφωνούν στο ένα μάτι, μερική ασυμφωνία} \\ 1, & \text{όταν οι δύο βαθμολογητές συμφωνούν στα δύο μάτια, πλήρης συμφωνία} \end{cases}$$

Πίνακας 2.2

Συχνότητες εντοπισμού ανωμαλίας κατά την εξέταση και των δύο ματιών από δύο βαθμολογητές.

	$2^{ος}$ Βαθμολογητής				
$1^{ος}$ Βαθμ/τής	R+L+	R+L-	R-L+	R-L-	ΣΥΝΟΛΟ
R+L+	f_{11} (1,0)	f_{12} (0,5)	f_{13} (0,5)	f_{14} (0,0)	$f_{1.}$
R+L-	f_{21} (0,5)	f_{22} (1,0)	f_{23} (0,0)	f_{24} (0,5)	$f_{2.}$
R-L+	f_{31} (0,5)	f_{32} (0,0)	f_{33} (1,0)	f_{34} (0,5)	$f_{3.}$
R-L-	f_{41} (0,0)	f_{42} (0,5)	f_{43} (0,5)	f_{44} (1,0)	$f_{4.}$
ΣΥΝΟΛΟ	$f_{.1}$	$f_{.2}$	$f_{.3}$	$f_{.4}$	n

(εντός των παρενθέσεων δίνονται τα αντίστοιχα βάρη)

Έτσι το μέτρο συμφωνίας κατά Schouten δίνεται από τον τύπο:

$$\hat{k}_w = \frac{p_o - p_c}{1 - p_c},$$

όπου

$$p_o = \frac{\sum_{i=1}^{4}\sum_{j}^{4} w_{ij} f_{ij}}{n},$$

$$p_c = \frac{\sum_{i=1}^{4}\sum_{j=1}^{4} w_{ij} f_{i.} f_{.j}}{n^2}$$

και w_{ij} τα βάρη όπως ορίστηκαν παραπάνω.

Ο $\hat{k}_w$ μπορεί εύκολα να επεκταθεί και για πέραν των δύο κατηγοριών ταξινόμησης αναπροσαρμόζοντας απλά τα αντίστοιχα βάρη w_{ij}. Επιπλέον, τόσο η μέθοδος του Oden όσο και του Schouten μπορούν να γενικευθούν και για την περίπτωση στην οποία περισσότερες των δύο όμοιων οργάνων εξετάζονται, π.χ. αδένες ή αγγεία.

Οι Shoukri et al. (1995) πρότειναν ένα διαφορετικό τρόπο αντιμετώπισης όμοιων περιπτώσεων, στις οποίες οι βαθμολογητές ταξινομούν τα αντικείμενα τυφλά, σύμφωνα με δύο διαφορετικά πρωτόκολλα ταξινόμησης σε μία από τις δύο κατηγορίες. Σκοπός είναι να συμφωνηθεί σύγχρονη εγκυρότητα των δύο πρωτοκόλλων. Για παράδειγμα δύο πρόσφατα τεστ για τη διάγνωση μιας μορφής φυματίωσης (paratuberculosis) στα βοοειδή είναι το DIA (immunobinding assay) και το ELISA (enzyme linked immunosorbent assay). Σύγκριση των αποτελεσμάτων των δύο τεστ εξαρτάται από δείγματα ορού που συλλέχθηκαν από τα βοοειδή. Εξετάζει, λοιπόν, κάθε εξεταστής, το ίδιο δείγμα ορού χρησιμοποιώντας και τα δύο τεστ – μια διαδικασία που καταλήγει σαφώς σε πραγματικό «ταίριασμα».

Ας υποθέσουμε ότι $X_i = 1$ ή 0 ($i=1, 2, \ldots, n$) ανάλογα με το αν το i-δείγμα ορού που εξετάσθηκε με το DIA τεστ ήταν θετικό ή αρνητικό, και ας θέσουμε $Y_i = 1$ ή 0 ανάλογα με το αν το i-δείγμα ορού που εξετάσθηκε με το ELISA τεστ ήταν θετικό ή όχι. Έτσι αν π_{kl} ($k,l=0,1$) είναι η πιθανότητα $X_i = k$ και $Y_i = l$, τότε $p_1 = \pi_{11} + \pi_{01}$ είναι η πιθανότητα το δείγμα ορού να είναι θετικό με βάση το ELISA τεστ και $p_2 = \pi_{11} + \pi_{10}$ είναι η πιθανότητα το «ταιριασμένο» δείγμα ορού να είναι θετικό με βάση το DIA τεστ. Με βάση το μοντέλο αυτό,

$$k = \frac{2r(p_1q_1p_2q_2)^{\frac{1}{2}}}{p_1q_2 + p_2q_1},$$

όπου $q_1=1-p_1$, $q_2= 1-p_2$ και ρ είναι ο συντελεστής συσχέτισης μεταξύ X και Y. Οι Shoukri et al. βάσει ενός τυχαίου δείγματος n ζευγών τέτοιων correlated binary αποκρίσεων κατέληξαν στον εκτιμητή μέγιστης πιθανοφάνειας

$$k = \frac{2(\bar{t} - \bar{x}\,\bar{y})}{\bar{y}(1-\bar{x}) + \bar{x}(1-\bar{y})},$$

όπου $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ και $\bar{t} = \frac{\sum_{i=1}^{n} x_i y_i}{n}$. Χρησιμοποιώντας την έκφραση της διακύμανσης για μεγάλα δείγματα, μπορεί να εξεταστεί και η υπόθεση ότι δύο διαγνωστικά τεστ είναι ασυσχέτιστα.

2.5.2 Παρουσία συμμεταβλητών

Υπάρχουν περιπτώσεις όπου η περιθώρια πιθανότητα ταξινόμησης μπορεί να εξαρτάται από μία ή περισσότερες συμμεταβλητές. Για παράδειγμα είναι περισσότερο πιθανό να κατατάξουμε μία ανωμαλία σε μαστογραφία σαν καρκίνο του στήθους όταν είναι γνωστό ότι

η ασθενής έχει οικογενειακό ιστορικό καρκίνου του στήθους ή όταν η ασθενής είναι κάποιας μεγαλύτερης ηλικίας, γιατί το οικογενειακό ιστορικό και η ηλικία είναι γνωστοί παράγοντες κινδύνου για τον καρκίνο του στήθους. Αν δεν ληφθούν υπόψη οι παράγοντες αυτοί μπορεί να οδηγηθούν οι βαθμολογητές σε αυξημένη συμφωνία.

Οι Barlow et al (1991) θεώρησαν τα n αντικείμενα χωρισμένα σε στρώματα και υπέθεσαν ότι ο k συντελεστής για κάθε στρώμα είναι ο ίδιος, αλλά η περιθώρια πιθανότητα ταξινόμησης μπορεί να διαφέρει για τους πίνακες συνάφειας των αντίστοιχων στρωμάτων. Πρότειναν επίσης ένα στρωματοποιημένο kappa, βασισμένο σε σταθμισμένους των ανεξάρτητων kappa των στρωμάτων. Ο καλύτερος (στα πλαίσια ελαχιστοποίησης του τετραγωνικού σφάλματος) στρωματοποιημένος kappa εκτιμητής της συνολικής συμφωνίας είναι ο σταθμισμένος μέσος των ανεξάρτητων kappa εκτιμητών, με στάθμες ανάλογων των δειγματικών μεγεθών των στρωμάτων.

Όσο ο αριθμός των συμμεταβλητών μεγαλώνει, πληθαίνουν τα στρώματα και συνεπώς ελαττώνεται το πλήθος των παρατηρήσεων σε κάθε στρώμα. Έτσι ο στρωματοποιημένος συντελεστής kappa μπορεί να βασιστεί και σε λίγες μόνο παρατηρήσεις σε κάθε στρώμα – πίνακα και θα έχει φτωχή ασυμπτωτική συμπεριφορά.

2.5.3 Συγκρίνοντας τα kappa από πολλαπλές μελέτες

Σε μια πολυκεντρική κλινική δοκιμή, προκαταρτικές μελέτες αξιοπιστίας διεξάγονται ανεξάρτητα σε καθένα από τα διάφορα κέντρα. Έτσι προκύπτουν και οι αντίστοιχες τιμές του kappa συντελεστή και συχνά ο μελετητής ενδιαφέρεται να εξετάσει αν το επίπεδο της συμφωνίας κατανέμεται ομοιόμορφα στα διάφορα κέντρα της κλινικής δοκιμής, π.χ. H_0: $k_1 = k_2 = \ldots = k_N$, όπου k_h η τιμή του k του πληθυσμού που αντιστοιχεί στο h κέντρο. Εναλλακτικά, οι N kappa συντελεστές μπορούν να προκύψουν από μια απλή μελέτη στην οποία τα αντικείμενα είναι χωρισμένα σε N στρώματα. Επίκεντρο της μελέτης και στις δύο περιπτώσεις είναι να διαπιστωθεί η ομοιογένεια του επιπέδου συμφωνίας και στους N πληθυσμούς.

Οι Donner et al (1996) ανέπτυξαν μια μεθοδολογία για να εξετάσουν την ομοιογένεια N ανεξάρτητων συντελεστών kappa (τιμές για τον πληθυσμό) της εντός των τάξεων μορφής. Υποθέτουν ότι N ανεξάρτητες μελέτες των $n = \sum_{h=1}^{N} n_h$ αντικειμένων συνολικά, έχουν διεξαχθεί, όπου κάθε αντικείμενο μπορεί να ταξινομηθεί σε δύο κατηγορίες (επιτυχία-

αποτυχία) από τον κάθε βαθμολογητή. Επιπλέον, υποθέτουν ότι η περιθώρια πιθανότητα να ταξινομηθεί ένα αντικείμενο ως «επιτυχία» είναι σταθερή για τους βαθμολογητές στη συγκεκριμένη *h*-μελέτη κι έστω είναι p_h. Ωστόσο αυτή η πιθανότητα μπορεί να κυμαίνεται μεταξύ των *N* μελετών. Με άλλα λόγια, δεν υπάρχει μεροληψία των βαθμολογητών σε κάθε μελέτη. Με αυτές τις προϋποθέσεις, οι πιθανότητες των από κοινού αποκρίσεων στην h-μελέτη προκύπτουν από ένα τριωνυμικό μοντέλο κι έχουν ως εξής:

$$\text{Δύο επιτυχίες: } p_{1h}(k_h) = P_h^2 + P_h(1-P_h)k_h$$

$$\text{Μία επιτυχία και μία αποτυχία: } p_{2h}(k_h) = 2P_h(1-P_h)(1-k_h)$$

$$\text{Δύο αποτυχίες: } p_{3h}(k_h) = (1-P_h)^2 + P_h(1-P_h)k_h,$$

Όπου P_h και k_h δίνονται από τους αντίστοιχους εκτιμητές για το *h*-στρώμα,

$$\hat{P}_h = \frac{2n_{1h} + n_{2h}}{2n_h}$$

και

$$\hat{k}_h = 1 - \frac{n_{2h}}{2n_h\hat{P}_h(1-\hat{P}_h)},$$

όπου n_{1h} είναι ο αριθμός των αντικειμένων στην *h*- μελέτη που ταξινομήθηκαν ως «επιτυχίες» και από τους δύο βαθμολογητές, n_{2h} είναι ο αριθμός αντικειμένων της *h*- μελέτης τα οποία ταξινομήθηκαν ως «επιτυχία» από τον ένα βαθμολογητή και ως «αποτυχία» από τον άλλο, n_{3h} είναι ο αριθμός των αντικειμένων που ταξινομήθηκαν ως «αποτυχίες» και από τους δύο βαθμολογητές και $n_h = n_{1h} + n_{2h} + n_{3h}$. Ένα μέτρο συμφωνίας για όλες τις μελέτες εκτιμάται υπολογίζοντας ένα σταθμισμένο μέσο των ανεξάρτητων k_h :

$$\hat{k} = \frac{\sum_{h=1}^{N} n_h\hat{P}_h(1-\hat{P}_h)k_h}{\sum_{h=1}^{N} n_h\hat{P}_h(1-\hat{P}_h)}.$$

Για να ελέγξουμε τη μηδενική υπόθεση H_0: $k_1 = k_2 = \ldots = k_N$, οι Donner et al. προτείνουν ένα τεστ καλής προσαρμογής χρησιμοποιώντας το στατιστικό

$$X_G^2 = \sum_{h=1}^{N}\sum_{l=1}^{3} \frac{\{n_{lh} - n_h\hat{p}_{lh}(\hat{k})\}^2}{n_h\hat{p}_{lh}(\hat{k})},$$

όπου η ποσότητα $\hat{p}_{1h}(\hat{k})$ βρίσκεται αντικαθιστώντας τα P_h και k_h από τους αντίστοιχους εκτιμητές τους στον τύπο της $\pi_{lh}(\kappa_h)$, $l= 1, 2, 3$, $h= 1, 2, \ldots N$. Κάτω από τη μηδενική υπόθεση το στατιστικό X_G^2 ακολουθεί προσεγγιστικά χ^2 κατανομή με N-1 βαθμούς ελευθερίας.

Οι Donner et al. (1996) μελέτησαν και άλλη μια μέθοδο ελέγχου της μηδενικής υπόθεσης χρησιμοποιώντας μια μέθοδο διακύμανσης μεγάλου δείγματος. Η εκτίμηση της διακύμανσης του μεγάλου δείγματος, σύμφωνα με τους Bloch και Kraemer (1989) και τους Fleiss και Davies (1982) δίνεται από τον τύπο

$$\hat{Var}\hat{k}_h = \frac{1-\hat{k}_h}{n_h}\left((1-\hat{k}_h)(1-2\hat{k}_h)+\frac{\hat{k}_h(2-\hat{k}_h)}{2P_h(1-P_h)}\right).$$

Θέτοντας

$$\hat{W}_h = 1/\hat{Var}k_h$$

και

$$\tilde{k} = \frac{\sum_{h=1}^{N}\hat{W}_h\hat{k}_h}{\sum_{h=1}^{N}\hat{W}_h},$$

ένα προσεγγιστικό τεστ για τη μηδενική υπόθεση βρίσκεται χρησιμοποιώντας το στατιστικό $X_V^2 = \sum_{h=1}^{N}\hat{W}_h(\hat{k}_h-\tilde{k})^2$ που ακολουθεί χ^2 κατανομή με N-1 βαθμούς ελευθερίας.

Όταν $\hat{k}_h = 1$ για κάποιο h, το X_n^2 δεν ορίζεται. Δυστυχώς, αυτό συμβαίνει αρκετά συχνά σε δείγματα μικρού ή μεσαίου μεγέθους. Αντίθετα το τεστ καλής προσαρμογής X_G^2 μπορεί να υπολογιστεί σε όλες τις περιπτώσεις εκτός από εκείνη στην οποία $\hat{k}_h = 1$ για όλα τα στρώματα (όταν δηλαδή για όλα τα h, με $h=1, 2, \ldots, N$, τα αντίστοιχα k_h=1). Κανένα από τα δύο στατιστικά τεστ δεν μπορεί να υπολογιστεί όταν $\hat{P}_h = 0$ ή 1 για κάποιο h, εφόσον το $\hat{k}_h$ είναι απροσδιόριστο. Διάφοροι συγγραφείς βασισμένοι σε μελέτες Monte Carlo κατέληξαν ότι τα δύο στατιστικά τεστ έχουν παρόμοιες ιδιότητες για μεγάλα δείγματα (n_h> 100 για κάθε στρώμα h). Για μικρότερου μεγέθους δείγματα το τεστ καλής προσαρμογής X_G^2 είναι προτιμότερο.

ΚΕΦΑΛΑΙΟ 3

Μοντέλα συμφωνίας

3.1 Εισαγωγή

Στις βιοϊατρικές και κοινωνικές επιστήμες είναι συχνό φαινόμενο οι βαθμολογητές να ταξινομούν ένα δείγμα αντικειμένων σε κατηγορίες ονοματικές ή διατάξιμες. Ένα χαρακτηριστικό παράδειγμα κατηγορικής ταξινόμησης είναι η συχνότητα μιας συνήθειας: «συνέχεια», «τις περισσότερες φορές», πολύ συχνά», «περιστασιακά», «σπάνια», «ποτέ». Όταν χρησιμοποιούμε ένα τέτοιο σχέδιο ταξινόμησης, οι βαθμολογητές δεν δείχνουν τέλεια συμφωνία αλλά ταξινομούν μέρος του δείγματος σε διαφορετικές κατηγορίες. Η ταξινόμηση, λοιπόν, σε διαφορετικές κατηγορίες οδηγεί τελικά στην ετερογένεια των ταξινομήσεων.

Στο προηγούμενο κεφάλαιο αναλύσαμε μέτρα με τα οποία μελετάμε την συμφωνία μεταξύ βαθμολογητών, με πιο γνωστό το δείκτη kappa του Cohen (k) ο οποίος λαμβάνει την τιμή 1 όταν πρόκειται για πλήρη συμφωνία. Επίσης είδαμε και τον σταθμισμένο kappa του Cohen όταν πρόκειται για ταξινόμηση διατακτικής κλίμακας, αφού αναλύει πιο πιστά τη συμφωνία μεταξύ διαδοχικών κατηγοριών. Πολλά άρθρα επισημαίνουν τα ανεπιθύμητα χαρακτηριστικά των δύο ανωτέρω δεικτών (πχ Brennan and Prediger (1981), Darroch and McCloud (1986), Feinstein and Cicchetti (1990), Spitznagel and Helzer (1985)).

Όταν εξετάζουμε δεδομένα ως προς τη συμφωνία βαθμολογητών, είναι προτιμότερο να χρησιμοποιούμε στατιστικά μοντέλα, που ποικίλλουν από απλά σε σύνθετα, εφόσον συμπεράσματα σχετικά με τη δομή της συμφωνίας μπορούν να εξαχθούν μόνο από ένα καλά προσαρμοσμένο μοντέλο. Επιπλέον, αν τα μοντέλα είναι και ιεραρχικά συνδεδεμένα, η αναζήτηση μοντέλου μπορεί να βασιστεί στις διαφορές στα τεστ καλής προσαρμογής μεταξύ των μοντέλων.

Στο παρόν κεφάλαιο αναφέρονται τα μοντέλα ανεξαρτησίας και quasi-ανεξαρτησίας, συμμετρίας και quasi-συμμετρίας. Περιγράφονται επίσης τα μοντέλα ομοιογένειας περιθωρίου, τριγωνικής ασυμμετρίας και διαγώνιας, καθώς επίσης και τα μοντέλα ομοιόμορφης συσχέτισης, επίδρασης γραμμών, επίδρασης στηλών, επίδρασης γραμμών και στηλών. Στη συνέχεια αναλύονται τα μοντέλα των Tanner and Young (1985a, 1985b) που είναι πιο αποτελεσματικά για δεδομένα ονοματικής κλίμακας, του Agresti (1988) που

χρησιμοποιούνται συνήθως για ταξινομήσεις σε διατάξιμη κλίμακα, ένα λογαριθμο-πολλαπλασιαστικό μοντέλο και εισάγεται ένα νέο λογαριθμο-γραμμικό μοντέλο.

3.2 Μοντέλο ανεξαρτησίας

Ας υποθέσουμε την περίπτωση δύο βαθμολογητών, όπου ο καθένας ταξινομεί n αντικείμενα σε I διαφορετικές κατηγορίες. Οι συχνότητες με τις οποίες παρατηρήθηκε καθένας από τους I^2 πιθανούς συνδυασμούς ταξινομήθηκαν σε έναν IxI πίνακα συνάφειας.

Ας θεωρήσουμε τον δισδιάστατο πίνακα συνάφειας IxI. Έστω n_{ij}, i,j =1,2,…I, η παρατηρούμενη συχνότητα στο κελί (i,j) και m_{ij}, i,j =1,2,…I, η αναμενόμενη συχνότητα στο κελί (i,j) κάτω από το μοντέλο που θεωρώ. Με $p_{ij} = n_{ij}/n$, όπου n ο συνολικός αριθμός των παρατηρήσεων στον πίνακα συνάφειας, ας συμβολίσουμε το δειγματικό ποσοστό, και με π_{ij} την πιθανότητα στο (i,j) κελί.

Οι αναμενόμενες συχνότητες κάτω από την ανεξαρτησία είναι:

$$m_{ij} = np_{i}p_{.j}\,,\ i, j=1, \ldots, I \tag{3.1}$$

όπου "." στο δείκτη δηλώνει άθροιση ως προς το δείκτη αυτό (π.χ. $p_{i.} = \sum_j p_{ij}$).

Ισοδύναμη είναι η λογαριθμογραμμική έκφραση

$$\log(m_{ij}) = l + l_i^X + l_j^Y, \qquad i, j=1, \ldots, I. \tag{3.2}$$

Οι κύριοι παράγοντες λ^X, λ^Y ικανοποιούν για λόγους προσδιορισιμότητας τους περιορισμούς:

$$\sum_{i=1}^{I} l_i^X = \sum_{j=1}^{I} l_j^Y = 0 \tag{3.3}$$

(ή ισοδύναμα τους $l_1^X = l_1^Y = 0$ ή $l_I^X = l_J^Y = 0$).

Από την (3.2) και την (3.3), ισχύει ότι,

$$l_i^X = \frac{1}{J}\sum_j \log(m_{ij}) - l\ ,\ \text{i= 1, …,I}$$

$$l_i^Y = \frac{1}{I}\sum_i \log(m_{ij}) - l\ ,\quad \text{j=1, …, I} \tag{3.4}$$

$$l = \frac{1}{IJ}\sum_{i,j} \log(m_{ij})$$

Παρατηρούμε ότι σε κάθε γραμμή ο λόγος πιθανοτήτων (odds) της στήλης j ως προς την k ισούται με $e^{l^Y_j - l^Y_k}$. Η προσαρμογή του μοντέλου στα δεδομένα , καθώς και των μοντέλων που ακολουθούν, ελέγχεται με το X^2 στατιστικό του Pearson $\left(X^2 = \sum_{i,j} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \right)$ και το G^2 στατιστικό πηλίκο πιθανοφάνειας $\left(G^2 = \sum_{i,j} n_{ij} \log \frac{n_{ij}}{m_{ij}} \right)$, όπου m_{ij} είναι ο εκτιμητής μέγιστης πιθανοφάνειας της αναμενόμενης συχνότητας m_{ij} κάτω από το υπό θεώρηση μοντέλο. Τα στατιστικά αυτά ακολουθούν ασυμπτωτικά την χ^2 κατανομή με τους ίδιους βαθμούς ελευθερίας $(I-1)^2$.

3.3 Πλήρες μοντέλο

Αν στο μοντέλο της ανεξαρτησίας (3.2) προσθέσουμε τον όρο της αλληλεπίδρασης λ^{XY} καταλήγουμε στο πλήρες μοντέλο:

$$\log(m_{ij}) = l + l^X_i + l^Y_j + l^{XY}_{ij},\ i, j = 1, \ldots, I \qquad (3.5)$$

Εκτός των περιορισμών (3.3), που ικανοποιούν οι κύριοι παράγοντες, ισχύουν ακόμη:

$$\sum_j l^{XY}_{ij} = 0 \ \text{(i=1, …, I) και} \ \sum_i l^{XY}_{ij} = 0 \ (j = 1, \ldots, I) \qquad (3.6)$$

$$[\text{ή}\ \ l^{XY}_{1j} = l^{XY}_{i1} = 0 \ \text{ή}\ \ l^{XY}_{Ij} = l^{XY}_{iJ} = 0\,]$$

όπου τα λ, l^X_i, l^Y_j δίνονται στις εξισώσεις (3.4), ενώ

$$l^{XY}_{ij} = \log(m_{ij}) - (l + l^X_i + l^Y_j),\ i, j = 1, \ldots, I.$$

Οι όροι της αλληλεπίδρασης μεταξύ των μεταβλητών X και Y ορίζουν συνάφεια. Πράγματι, αν θ_{ij} είναι τα odds ratio που ορίζονται για τους 2x2 πίνακες των διαδοχικών κατηγοριών, τότε:

$$\log(\theta_{ij}) = \log\left(\frac{m_{ij} m_{i+1,j+1}}{m_{i+1,j} m_{i,j+1}}\right) = l^{XY}_{ij} + l^{XY}_{i+1,j+1} - l^{XY}_{i+1,j} - l^{XY}_{i,j+1},\ i, j = 1, \ldots, I-1.$$

Το πλήθος των παραμέτρων του μοντέλου (3.5) ισούται με $1+(I-1)+(I-1)+(I-1)(I-1)=I^2$, δηλαδή με το πλήθος των κελιών του πίνακα. Συνεπώς το μοντέλο έχει df=0, όλα τα υπόλοιπα του πίνακα είναι μηδενικά και τελικά $X^2=G^2=0$.

Το πλήρες μοντέλο είναι ιεραρχικό μοντέλο. [παράδειγμα μη-ιεραρχικού μοντέλου: $\log(m_{ij}) = l + l_i^X + l_{ij}^{XY}$]

3.4 Λογαριθμο-γραμμικά μοντέλα για πίνακες με δομικά μηδενικά

Το γεγονός ότι δύο βαθμολογητές ταξινομούν ως προς κάποια μεταβλητή κάποια αντικείμενα (ή περιπτώσεις) συνεπάγεται τη συγκέντρωση των περισσότερων στοιχείων του *I*x*I* πίνακα ταξινόμησης στα κελιά της διαγωνίου ενώ στα πάνω δεξιά και στα κάτω αριστερά κελιά του πίνακα αναμένουμε, αν όχι μηδενικές, πολύ μικρές συχνότητες. Επειδή, λοιπόν, τα κελιά της διαγωνίου δεν μας βοηθούν στην εξέταση της ανεξαρτησίας των δύο βαθμολογητών, αντίθετα μάλιστα την επηρεάζουν, θα θεωρήσουμε τα κελιά της διαγωνίου ως τεχνητά δομικά μηδενικά.

Γενικά, τα δομικά μηδενικά μπορεί να είναι πραγματικά (π.χ. σε έναν πίνακα συσχέτισης με μεταβλητές το φύλο – άνδρας, γυναίκα – και την εμφάνιση καρκίνο του προστάτη – ναι, όχι – ή σε έναν πίνακα όπου εμφανίζονται τα εντός και τα εκτός έδρας παιχνίδια αθλητικών σωματείων, όπου τα κελιά της διαγωνίου θα είναι μηδενικά εφόσον δεν ορίζεται σαν έννοια το ταυτόχρονο παιχνίδι εντός και εκτός έδρας για την ίδια ομάδα), μπορεί να είναι και τεχνητά. Συνήθως όταν κάποια κελιά παρουσιάζουν ειδική συμπεριφορά και έχουν μεγαλύτερες ή μικρότερες τιμές, τότε θέτουμε τεχνητά δομικά μηδενικά, όπως στην περίπτωσή μας, της εξέτασης συμφωνίας μεταξύ βαθμολογητών ή ακόμα και σε πίνακες κοινωνικής εξέλιξης και κινητικότητας που συναντάμε στη Δημογραφία (π.χ. κοινωνική κατάσταση- ευρωστία πατέρα και γιου, όπου αναμένεται τα κελιά της διαγωνίου να συγκεντρώνουν τα μεγαλύτερα ποσοστά παρατηρήσεων, ενώ τα πιο απομακρυσμένα από τη διαγώνιο να είναι σχεδόν μηδενικά).

Ας ορίσουμε S το σύνολο των κελιών του πίνακα που δεν περιέχει δομικά μηδενικά, δηλαδή $m_{ij} = 0$ για $(i,j) \notin S$.

Έτσι, το μοντέλο έχει ως εξής:

$$\log(m_{ij}) = l + l_i^X + l_j^Y + l_{ij}^{XY}, \ (i,j) \in S \qquad (3.7)$$

Οι περιορισμοί (3.3) και (3.6) αντικαθίστανται από τους:

$$\sum_i d_i^X \, l_i^X = \sum_j d_j^Y \, l_j^Y = 0 \qquad (3.8\alpha)$$

$$\sum_i d_{ij}\, l_{ij}^{XY} = \sum_j d_{ij}\, l_{ij}^{XY} = 0 \qquad (3.8\beta)$$

όπου

$$d_{ij} = \begin{cases} 1, (i,j) \in S \\ 0, (i,j) \notin S \end{cases},\ d_i^{\,X} = \begin{cases} 1, \text{ όταν } \exists\, j : d_{ij} = 1 \\ 0, \quad \text{διαφορετικά} \end{cases} \text{ και } d_j^{\,Y} = \begin{cases} 1, \text{ όταν } \exists\, i : d_{ij} = 1 \\ 0, \quad \text{διαφορετικά} \end{cases}.$$

3.5 Μοντέλο Quasi-ανεξαρτησίας (QI)

Το μοντέλο της quasi-ανεξαρτησίας (QI) ορίζεται ως:

$$m_{ij} = np_{i}p_{.j},$$

ή

$$\log(m_{ij}) = l + l_i^X + l_j^Y,\ (i,j) \in S.$$

Δηλαδή $l_{ij}^{XY} = 0, (i,j) \in S$ στο (3.7), ενώ οι περιορισμοί που ισχύουν για την προσδιορισιμότητα των παραμέτρων είναι οι (3.8α). Η φυσική ερμηνεία της QI είναι ανάλογη με της ανεξαρτησίας σε έναν πλήρη πίνακα.

Η QI συνεπάγεται ότι τα σχετικά ποσοστά υποκειμένων ταυτίζονται στα αντίστοιχα κελιά δύο οποιωνδήποτε γραμμών (στηλών) του πίνακα, δοθέντος ότι δεν υπολογίζουμε τις στήλες (γραμμές) που έχουν δομικά μηδενικά σε τουλάχιστον ένα από τα εν λόγω κελιά. Δηλαδή η QI είναι μια μορφή ανεξαρτησίας, δεσμευμένη υπό τον περιορισμό της προσοχής μας σε ένα μη-πλήρες τμήμα του πίνακα.

Οι βαθμοί ελευθερίας (d.f.) για το μοντέλο QI δίνονται από τη σχέση df=$(I\text{-}1)^2$-z_c, όπου z_c είναι το πλήθος όλων των δομικών μηδενικών του πίνακα

Ο έλεγχος καλής προσαρμογής του μοντέλου QI μπορεί να γίνει με τα στατιστικά X^2 ή G^2, όπου αθροίζουμε τους όρους τους για όλα τα κελιά του *S*. Κάτω από τη μηδενική υπόθεση της QI και τα δύο αυτά στατιστικά ακολουθούν την χ^2-κατανομή με τους βαθμούς ελευθερίας που δώσαμε παραπάνω.

3.6 Μοντέλο Συμμετρίας (S)

Ένας *I*x*I* πίνακας συνάφειας θα ακολουθεί το μοντέλο της συμμετρίας αν οι πιθανότητες στα συμμετρικά κελιά είναι ίσες, δηλαδή:

$$\pi_{ij} = \pi_{ij}^{S},\ \pi_{ij}^{S} = \pi_{ji}^{S},\ i \neq j,\ i, j = 1, \ldots, I$$

Η ισοδύναμη λογαριθμογραμμική έκφραση είναι:

$$\log(p_{ij}) = l + l_i^X + l_j^Y + l_{ij}^{XY},\ i, j = 1, \ldots, I,$$

με τους περιορισμούς:

$$l_i^X = l_i^Y,\ i, j = 1, \ldots, I \qquad (3.9\alpha)$$

$$l_{ij}^{XY} = l_{ji}^{XY},\ i, j = 1, \ldots, I \qquad (3.9\beta)$$

3.7 Μοντέλο Quasi-συμμετρίας (QS)

Αν χαλαρώσουμε τους περιορισμούς (3.9α), δηλαδή αν θεωρήσουμε το παραπάνω λογαριθμογραμμικό μοντέλο με τους περιορισμούς (3.9β), τότε καταλήγουμε στο μοντέλο της quasi-συμμετρίας (QS).

Ενώ η φυσική ερμηνεία του μοντέλου *S* είναι εμφανής, δε συμβαίνει το ίδιο και με τη φυσική ερμηνεία του QS, καθώς δεν είναι εύκολο να αποδοθεί ερμηνευτικά η συμμετρικότητα των όρων της αλληλεπίδρασης.

Το μοντέλο QS είναι ακόμη γνωστό και ως μοντέλο συμμετρικής συνάφειας, λόγω της ιδιότητας της συμμετρικότητας των odds ratio:

$$\theta_{ij} = \theta_{ji},\ i \neq j,\ i, j = 1, \ldots, I$$

Άλλες ισοδύναμες εκφράσεις του (QS) είναι η πολλαπλασιαστική

$$p_{ij}^{QS} = a_i b_j g_{ij},\ \ g_{ij} = g_{ji},\ i \neq j,\ i, j = 1, \ldots, I$$

και η

$$p_{ij}^{QS} = \frac{2c_i}{c_i + c_j} p_{ij}^{S},\ i \neq j,\ i, j = 1, \ldots, I \qquad (3.10)$$

που έχει τη μορφή απόκλισης από τη συμμετρία.

Ισχύει

$$c_i = \exp\left[\frac{1}{I}\sum_{j(j\neq i)} \ln\left(\frac{p_{ij}^{QS}}{p_{ji}^{QS}}\right) - 1\right].$$

Αν $c_i = 1$ (i=1, …, I) τότε το μοντέλο QS απλοποιείται στο μοντέλο πλήρους συμμετρίας (S).

Ας δούμε τώρα πως σχετίζεται το μοντέλο quasi-συμμετρίας με άλλα μοντέλα:

Η συμμετρία (S) συνεπάγεται τη quasi-συμμετρία (QS).

Όταν I=3, τότε το (QS) είναι ισοδύναμο με το μοντέλο quasi-ανεξαρτησίας (QI): $\pi_{ij}= \alpha_i\beta_j$, $i \neq j$. Για I>3, το μοντέλο QI συνεπάγεται το QS (Caussinus, 1965).

Οι Darroch and MacCloud (1986) επιχειρηματολόγησαν ότι κάτω από λογικές υποθέσεις οι αναμενόμενες συχνότητες ενός μοντέλου συμφωνίας μεταξύ βαθμολογητών θα πρέπει να έχουν quasi-συμμετρική δομή, δηλαδή οι αναμενόμενες συχνότητες δίνονται από το γινόμενο των επιδράσεων γραμμών, επιδράσεων στηλών και παραμέτρους από μια συμμετρική σειρά από αλληλεπιδράσεις. Έτσι, τα μοντέλα quasi-συμμετρίας, για την αναμενόμενη συχνότητα στο κάθε κελί, m_{ij}, θα πρέπει να έχουν τη γενική μορφή:

$$m_{ij}= a_i \beta_j \psi_{ij}, \psi_{ij} = \psi_{ji}, \tag{3.11}$$

όπου a_i>0 και β_j>0 είναι οι επιδράσεις γραμμών και στηλών και ψ_{ij}>0 είναι συμμετρικές αλληλεπιδράσεις.

Η προσέγγιση των Darroch και MacCloud βασίζεται σε δύο υποθέσεις με αναφορά στη συγκεκριμένη πιθανότητα του βαθμολογητή π_{hri}, με την οποία το αντικείμενο h ταξινομείται από τον r βαθμολογητή στην κατηγορία i. Πρώτα, υποτίθεται ότι ο κάθε βαθμολογητής ταξινομεί τα αντικείμενα ανεξάρτητα από τους άλλους βαθμολογητές - δεν έχει πληροφόρηση δηλαδή για την ταξινόμηση των αντικειμένων από τους άλλους βαθμολογητές. Δεύτερον, υποτίθεται ότι οι πιθανότητες π_{hri} ικανοποιούν την υπόθεση ότι δεν υπάρχει δεύτερης τάξης αλληλεπίδραση. Έτσι, για κάποιες τιμές των a_{hi} >0, β_{ti} >0 και c_{ht} >0 παραμέτρων,

$$\pi_{hri} = a_{hi} \beta_{ti} c_{ht} \tag{3.12}$$

Αυτό σημαίνει ότι οι διαφορές μεταξύ των αντικειμένων, οι οποίες αντικατοπτρίζονται στις παραμέτρους a_{hi} και οι διαφορές μεταξύ των βαθμολογητών, οι οποίες αντικατοπτρίζονται στις παραμέτρους β_{ti} , συνδυάζονται χωρίς αλληλεπίδραση.

Αν και οι δύο υποθέσεις ισχύουν και n είναι ο αριθμός των αντικειμένων που ταξινομούνται από τους δύο βαθμολογητές, τότε οι πιθανότητες κάτω από κάθε παρατηρούμενη συχνότητα σε κάθε κελί είναι:

$$p_{ij} = n^{-1} \sum_{h=1}^{N} p_{h1i} p_{h2j} \tag{3.13}$$

Αντικαθιστώντας την εξίσωση (3.12) στην (3.13) προκύπτει η quasi-συμμετρική δομή για τις πιθανότητες p_{ij} ή ισοδύναμα για τις $m_{ij}= \alpha_i \beta_j \psi_{ij}$ της σχέσης (3.11).

3.8 Μοντέλο Ομοιογένειας Περιθωρίου (MH)

Το μοντέλο ομοιογένειας περιθωρίου (MH) ορίζεται ως:

$$\pi_{i.} = \pi_{.i},\ i = 1, \ldots, I.$$

Το μοντέλο αυτό δεν έχει ισοδύναμη λογαριθμογραμμική έκφραση. Είναι προφανές ότι η συμμετρία S συνεπάγεται την ομοιογένεια περιθωρίου MH.

Η πιο χαρακτηριστική ιδιότητα που συνδέει το μοντέλο MH με τα μοντέλα S και QS, που παρουσιάσαμε παραπάνω, είναι

$$\mathrm{S} = \mathrm{MH} \cap \mathrm{QS}.$$

Στην ιδιότητα αυτή στηρίζεται ο υπό συνθήκη έλεγχος καλής προσαρμογής της MH, δοθέντος ότι το μοντέλο QS ισχύει. Ο εν λόγω έλεγχος γίνεται μέσω του στατιστικού $G^2(\mathrm{S}|\mathrm{QS}) = G^2(\mathrm{S}) - G^2(\mathrm{QS})$, που ασυμπτωτικά ακολουθεί τη χ^2- κατανομή με I-1 βαθμούς ελευθερίας και δεν απαιτεί την εκτίμηση των πιθανοτήτων κάτω από το μοντέλο MH. Τέλος, όταν ισχύει το μοντέλο QS, τότε η κακή προσαρμογή της S οφείλεται στην περιθώρια ανομοιογένεια. Η δε κατεύθυνση και ένταση της περιθώριας ανομοιογένειας, για κάποιο i (i=1, ..., I), προσδιορίζονται από την τιμή της παραμέτρου c_i [μοντέλο (3.12β)].

3.9 Μοντέλο Τριγωνικής Ασυμμετρίας (T)

Το μοντέλο τριγωνικής ασυμμετρίας (T) ορίζεται ως:

$$p_{ij} = \begin{cases} t p^{S}_{ij}, i > j \\ (2-t) p^{S}_{ij}, i < j \end{cases},\ i,j = 1, \ldots, I,$$

όπου π_{ij} είναι η πιθανότητα του κελιού (i,j) κάτω από το μοντέλο T.

Το μοντέλο αυτό δηλώνει ότι υπάρχει μια «σταθερή» προτίμηση των κελιών του πάνω (κάτω) έναντι του κάτω (πάνω) τριγώνου αν τ<1 (τ>1). Ισχύει S = T ∩ QS.

1. Όταν τ=1, τότε το μοντέλο T ισοδυναμεί με το (S).
2. Το μοντέλο T εκφράζεται ισοδύναμα ως

$$p_{ij} = \frac{t}{2-t} p_{ji}, i > j,\ i,j = 1, \ldots, I$$

Ο τελευταίος ορισμός αναδεικνύει καλύτερα την φυσική ερμηνεία του μοντέλου: Κάτω από το μοντέλο T, η πιθανότητα μια παρατήρηση να πέσει στο κελί (i,j) με i>j είναι τ/(2-τ) φορές μεγαλύτερη από την πιθανότητα να πέσει στο συμμετρικό της κελί (j,i).

3.10 Μοντέλο Διαγώνιας Ασυμμετρίας (D)

Αν η απόκλιση από τη συμμετρία δε μετριέται με βάση τα τρίγωνα που ορίζει η κεντρική διαγώνιος (βάσει του μοντέλου T) αλλά με βάση τις συμμετρικές δευτερεύουσες διαγωνίους του πίνακα, τότε ορίζεται το μοντέλο της διαγώνιας ασυμμετρίας (Goodman, 1979b)

$$p_{ij} = d_k p_{ij}^{S}, k = i - j \text{ , } i,j = 1, \ldots, I.$$

Ισχύουν ότι:

1. Όταν δ_k=1, k=1, …, I-1, τότε το μοντέλο) ισοδυναμεί με το S.
2. Όταν $\delta_k = \tau$, k= 1, …, I-1, τότε το μοντέλο D ισοδυναμεί με το T.
3. Όταν το μοντέλο D ισχύει, αν $d_k \leq 1$, για κάθε k>0, τότε η μεταβλητή ταξινόμησης των γραμμών είναι στοχαστικά μικρότερη αυτής των στηλών.
4. Εύκολα μπορεί να δειχθεί ότι $\delta_k + \delta_{-k} = 2$ και το μοντέλο διαγώνιας ασυμμετρίας εκφράζεται ισοδύναμα ως

$$p_{ij} = \frac{d_k}{2 - d_k} p_{ji}, k = i - j, i > j \text{ , } i,j = 1, \ldots, I$$

παρέχοντάς μας την ανάλογη φυσική ερμηνεία.

Πρέπει να σημειώσουμε ότι κάτω από όλα τα μοντέλα συμμετρίας και ασυμμετρίας, για τα διαγώνια κελιά ισχύει $\hat{p}_{ii} = p_{ii}$, i=1,…,I.

Στον Πίνακα 3.1 που ακολουθεί φαίνονται οι βαθμοί ελευθερίας των μοντέλων που περιγράφηκαν πιο πάνω.

Πίνακας 3.1

Βαθμοί ελευθερίας μοντέλων

Μοντέλο	Βαθμοί ελευθερίας (d.f.)
S (συμμετρίας)	I(I-1)/2
QS (quasi-συμμετρίας)	(I-1)(I-2)/2
MH (ομοιογένειας περιθωρίου)	(I-1)
T (τριγωνικής ασυμμετρίας)	(I+1)(I-2)/2
D (διαγώνιας ασυμμετρίας)	(I-1)(I-2)/2

3.11 Παράδειγμα

Τα δεδομένα του Πίνακα 3.2, που έχουν παρουσιαστεί και αναλυθεί από τους Graham and Jackson (1993), έχουν προέλθει από μια case – control μελέτη για τη στεφανιαία νόσο και αφορούν 456 ασθενείς. Ένα τυχαία επιλεγμένο δείγμα αποκρινόμενων που είχαν υποστεί μη θανατηφόρο έμφραγμα του μυοκαρδίου (ασθενείς) δηλώνουν πόσο συχνά έκαναν χρήση αλκοόλ. Επιπρόσθετα οι άμεσοι συγγενείς τους (controls) παρείχαν πληροφόρηση για τις συνήθειες των ασθενών όσον αφορά την κατανάλωση αλκοόλ.

Πίνακας 3.2

Παρατηρούμενες συχνότητες 456 ζευγών ασθενών και συγγενών τους ως προς τη συχνότητα χρήσης αλκοόλ σε μια μελέτη case-control για τη στεφανιαία νόσο.

	Ασθενείς					
Άμεσοι συγγενείς	«Δεν πίνω αλκοόλ»	«Το έχω κόψει»	>1/μήνα <1/εβδομάδα	>1/εβδομάδα <1/μέρα	>1/μέρα	σύνολο
«Δεν πίνω αλκοόλ»	47	13	19	4	0	83
«Το έχω κόψει»	5	6	2	1	2	16
>1/μήνα <1/εβδομάδα	15	6	76	19	4	120
>1/εβδομάδα <1/μέρα	1	1	23	54	22	101
>1/μέρα	0	0	4	33	99	136
σύνολο	68	26	124	111	127	456

Όπως είναι αναμενόμενο παρατηρούμε μεγάλη συγκέντρωση στη διαγώνιο (61.8%), που αντιστοιχεί στις περιπτώσεις που συμφωνεί η απάντηση του ασθενή με αυτή του συγγενή του. Το ενδιαφέρον επικεντρώνεται στη μοντελοποίηση της συμφωνίας. Τα μοντέλα συμμετρίας – ασυμμετρίας που παρουσιάσαμε, δεν προσαρμόζονται άμεσα από κάποιο στατιστικό πακέτο. Για το λόγο αυτό χρησιμοποιήσαμε κώδικα του SPSS (syntax) και καταλήξαμε στον επόμενο Πίνακα 3.3. Το αρχείο syntax είναι από την Κατέρη (2005) ενώ το σχετικό αρχείο αποτελεσμάτων δίνεται:

Matrix

```
Run MATRIX procedure:

DATA TABLE
     47,000     13,000     19,000      4,000       ,000
      5,000      6,000      2,000      1,000      2,000
     15,000      6,000     76,000     19,000      4,000
      1,000      1,000     23,000     54,000     22,000
       ,000       ,000      4,000     33,000     99,000

MLE OF THE EXPECTED UNDER SYMMETRY (S) TABLE
     47,000      9,000     17,000      2,500       ,000
      9,000      6,000      4,000      1,000      1,000
     17,000      4,000     76,000     21,000      4,000
      2,500      1,000     21,000     54,000     27,500
       ,000      1,000      4,000     27,500     99,000

MLE OF THE EXPECTED UNDER QUASI SYMMETRY (QS) TABLE
     47,000     12,610     20,332      3,090       ,000
      5,390      6,000      3,110       ,818       ,689
     13,668      4,890     76,000     21,879      3,619
      1,910      1,182     20,121     54,000     23,747
       ,000      1,311      4,381     31,253     99,000

MLE OF THE PARAMETER VECTOR C
       ,728       ,311       ,490       ,450       ,593

ITERATIONS
  16

MLE OF THE EXPECTED UNDER TRIANGULAR ASYMMETRY (T) TABLE
     47,000      8,897     16,805      2,471       ,000
      9,103      6,000      3,954       ,989       ,989
     17,195      4,046     76,000     20,759      3,954
      2,529      1,011     21,241     54,000     27,184
       ,000      1,011      4,046     27,816     99,000

MLE OF THE PARAMETER T
      1,011

MLE OF THE EXPECTED UNDER DIAGONAL ASYMMETRY (D) TABLE
     47,000      8,195     18,545      4,286       ,000
      9,805      6,000      3,642      1,091      1,714
     15,455      4,358     76,000     19,122      4,364
       ,714       ,909     22,878     54,000     25,041
       ,000       ,286      3,636     29,959     99,000

MLE OF THE PARAMETER VECTOR D
      1,089       ,909       ,286       ,000


LIKELIHOOD RATIO TEST FOR S
         G2         DF       SIG.
   13.5441     9.0000      .1395

PEARSON'S X2 FOR S
         X2         DF       SIG.
   12.4071     9.0000      .1913
```

```
LIKELIHOOD RATIO TEST FOR QS
         G2          DF       SIG.
     7.1374      5.0000      .2106

PEARSON'S X2 FOR QS
         X2          DF       SIG.
     6.5728      5.0000      .2544

LIKELIHOOD RATIO TEST FOR T
         G2          DF       SIG.
    13.5211      8.0000      .0951

PEARSON'S X2 FOR T
         X2          DF       SIG.
    12.3857      8.0000      .1348

LIKELIHOOD RATIO TEST FOR D
         G2          DF       SIG.
     8.2325      5.0000      .1439

PEARSON'S X2 FOR D
         X2          DF       SIG.
     7.7849      5.0000      .1685

------ END MATRIX -----
```

Πίνακας 3.3

Συγκεντρωτικά αποτελέσματα από το SPSS

Μοντέλο	G^2	p-value	X^2	p-value	Βαθμοί ελευθερίας (d.f.)
S	13.5441	0.2593	12.4071	0.1395	I(I-1)/2 -1 =10
QS	7.1374	0.4147	6.5728	0.2106	(I-1)(I-2)/2 -1 =6
T	13.5211	0.1960	12.3857	0.0951	(I+1)(I-2)/2 -1 =9
D	8.2325	0.3125	7.7849	0.1439	(I-1)(I-2)/2 -1 =6

Στον Πίνακα 3.3 παρατηρούμε ότι οι βαθμοί ελευθερίας των μοντέλων έχουν ελαττωθεί κατά έναν και αυτό συμβαίνει γιατί όταν για τα συμμετρικά κελιά ($n_{15} = n_{51} = 0$) $n_{15} + n_{51} = 0$, οπότε είναι σαν να έχουμε ένα δομικό μηδενικό και διορθώνονται κατάλληλα οι βαθμοί ελευθερίας των μοντέλων.

3.12 Μοντέλα Συνάφειας

Στην περίπτωση των διδιάστατων πινάκων συνάφειας, που το μοντέλο της ανεξαρτησίας απορρίπτεται είναι δυνατή η εισαγωγή μοντέλων που βασίζονται στην ανάθεση σκορ στις

κατηγορίες των μεταβλητών ταξινόμησης. Τα μοντέλα αυτά εκφράζουν τη συνάφεια με λιγότερες από $(I\text{-}1)^2$ παραμέτρους (που αντιστοιχούν στο πλήρες μοντέλο). Τα σκορ μπορεί να είναι γνωστά η παράμετροι προς εκτίμηση.

Ισχυρή συμφωνία απαιτεί ισχυρή συνάφεια, όμως ισχυρή συνάφεια μπορεί να υφίσταται και χωρίς την ύπαρξη ισχυρής συμφωνίας. Αν π.χ. ο Α βαθμολογητής κατατάσσει τα υποκείμενα μια κατηγορία υψηλότερα από ότι ο Β βαθμολογητής, τότε η ισχύς της συμφωνίας είναι φτωχή αν και η συνάφεια είναι ισχυρή.

Η θεμελίωση και ανάπτυξή τους έγινε κυρίως από τους Simon(1974), Haberman (1974), Goodman (1979b, 1981, 1985, 1986).

Ας ορίσουμε αρχικά τα μοντέλα:

(α) Λογαριθμική έκφραση

$$\log(m_{ij}) = l + l_i^X + l_j^Y + fm_i n_j,\ i,j=1,\ \ldots,I$$

με τους περιορισμούς:

$$\sum_{i=1}^{I} l_i^X = \sum_{j=1}^{I} l_j^Y = 0$$

$$\sum_{i=1}^{I} w_{1i} m_i = \sum_{j=1}^{I} w_{2j} n_j = 0 \qquad (3.14\alpha)$$

$$\sum_{i=1}^{I} w_{1i} m_i^2 = \sum_{j=1}^{I} w_{2j} n_j^2 = 1 \qquad (3.14\beta)$$

όπου$\{w_{1i},\ w_{2j},\ i,j=1,\ \ldots,I\}$ συστήματα βαρών. Συνήθως $w_{1i} = w_{2j}=1$ *ή* $w_{1i}=\pi_{i.}$ *και* $w_{2j}=\pi_{.j}$ $(i,j=1,\ \ldots,\ I)$.

Πίνακας 3.4

Γνωστά και παραμετρικά μέρη των μοντέλων συνάφειας

Μοντέλο	$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_I)$	$\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_I)$
Ομοιόμορφης συνάφειας (U)	γνωστό	γνωστό
Επίδρασης γραμμών (R)	παράμετρος	γνωστό
Επίδρασης στηλών (C)	γνωστό	παράμετρος
Επίδρασης γραμμών-στηλών (RC)	παράμετρος	παράμετρος

Ας δούμε τώρα πως αναθέτουμε τιμές στα γνωστά σκορ. Η λογικότερη επιλογή είναι $\mu_i = i$ και $\nu_j = j$ $(i,j = 1,\ \ldots,\ I)$ ή γενικότερα η ανάθεση τιμών που ισαπέχουν για διαδοχικές

κατηγορίες των μεταβλητών ταξινόμησης. Δηλαδή η επιλογή τέτοιων σκορ ώστε να ισχύει $\mu_{i+1} - \mu_i = \Delta_1$ και $\nu_{j+1} - \nu_j = \Delta_2$ (Δ_1 και Δ_2 γνωστά).

Σε περίπτωση που η μεταβλητή ταξινόμησης είναι διαστηματική ή υπάρχει πληροφόρηση για άνισες αποστάσεις μεταξύ των ομάδων τότε αναθέτουμε στα σκορ τιμές που δεν ισαπέχουν για διαδοχικές κατηγορίες.

Στον ακόλουθο Πίνακα 3.5 δίνονται το πλήθος των παραμέτρων επιπλέον της ανεξαρτησίας καθώς και οι βαθμοί ελευθερίας για κάθε μοντέλο.

Πίνακας 3.5

Βαθμοί ανεξαρτησίας των μοντέλων συνάφειας

Μοντέλο	Πλήθος παραμέτρων επιπλέον της ανεξαρτησίας	Βαθμοί ελευθερίας (d.f.)
U	1	$(I\text{-}1)^2\text{-}1$
R	$1+(I\text{-}2) = I\text{-}1$	$(I\text{-}1)(I\text{-}2)$
C	$1+(I\text{-}2) = I\text{-}1$	$(I\text{-}1)(I\text{-}2)$
RC	$1+(I\text{-}2)+(I\text{-}2) = 2I\text{-}3$	$(I\text{-}2)^2$

Οι βαθμοί ελευθερίας προκύπτουν από το γενικό κανόνα:

d.f. = I^2-1-(πλήθος ανεξάρτητων παραμέτρων).

(β) πολλαπλασιαστική έκφραση

$$m_{ij} = a_i b_j e^{f m_i n_j}\ ,\ i, j = 1, \ldots, I$$

όσο για τους περιορισμούς, ισχύουν οι περιορισμοί (3.14α) και (3.14β) ενώ στα α_i και β_j δεν τίθενται περιορισμοί.

(γ) Έκφραση ως προς τα πηλίκα πιθανοτήτων (odds ratio)

$$q_{ij} = \frac{m_{ij} m_{i+1,\,j+1}}{m_{i,\,j+1} m_{i+1,\,j}} = e^{f(m_i - m_{i+1})(n_j - n_{j+1})}$$

ή

$$\Phi_{ij} = \log(q_{ij}) = f(m_i - m_{i+1})(n_j - n_{j+1})\ ,\ i, j=1, \ldots, I\text{-}1 \qquad (3.15)$$

Η φ είναι παράμετρος εγγενούς συνάφειας (intrinsic association) και ισχύει

$$f = \sum_{i,x} w_{1i} w_{2j} m_i n_j \log(m_{ij}) \qquad (3.16)$$

Η (3.16) δηλώνει ότι η φ είναι ένα σταθμικό μέσο συσχέτισης μεταξύ των γραμμών και των στηλών του πίνακα.

Άλλωστε από τον τύπο (3.15) είναι εμφανές ότι η φ ισούται με το λογάριθμο του πηλίκου πιθανοτήτων όταν τα σκορ των διαδοχικών γραμμών απέχουν μεταξύ τους κατά μία μονάδα και το ίδιο ισχύει για τα σκορ των στηλών.

Στην περίπτωση του μοντέλου (U), η (3.15) παίρνει τη μορφή $\Phi_{ij} = f\Delta_1\Delta_2$, $i,j=1, \ldots, I$, δηλαδή το Φ_{ij} είναι σταθερό για κάθε δυνατή τιμή των i και j. Διαπίστωση που δικαιώνει βέβαια το όνομα του (U) ως μοντέλο «ομοιόμορφης συνάφειας».

Ακολουθούν κάποιες ιδιότητες της συμπεριφοράς των μοντέλων (U), (R), (C), (RC):

1. Τα μοντέλα (U), (R), (C) είναι λογαριθμο-γραμμικά ενώ το (RC) είναι λογαριθμικό, μη-γραμμικό.
2. Το μοντέλο (U) είναι ευαίσθητο σε αναδιατάξεις των γραμμών ή των στηλών του πίνακα. Συνεπώς απαιτεί τη διαταξιμότητα και των δύο μεταβλητών ταξινόμησης του πίνακα.
3. Το μοντέλο (R) (ή C) είναι ευαίσθητο μόνο σε αναδιατάξεις των στηλών (ή γραμμών). Άρα απαιτεί τη διαταξιμότητα μόνο της μεταβλητής ταξινόμησης κατά στήλες (ή γραμμές) του πίνακα.
4. Το μοντέλο (RC) είναι ανεπηρέαστο τόσο από αναδιατάξεις των γραμμών όσο και των στηλών του πίνακα. Συνεπώς μπορεί να χρησιμοποιηθεί και σε πίνακες με μη-διατάξιμες μεταβλητές ταξινόμησης.
5. Από την παραπάνω παρατήρηση απορρέει ότι τα παραμετρικά σκορ των μοντέλων (R), (C) και (RC) δεν είναι απαραίτητο να είναι μονότονα. Έλλειψη της μονοτονίας τους συνεπάγεται μη μονότονη συνάφεια, με την έννοια ότι η τοπική συνάφεια θα είναι θετική σε κάποιες περιοχές και αρνητική αλλού.
6. Όταν τα σκορ και των γραμμών και των στηλών είναι διατεταγμένα κατά αύξουσα τάξη, τότε θετικές τιμές της παραμέτρου φ δηλώνουν θετική συνάφεια μεταξύ των μεταβλητών ταξινόμησης.
7. Οι αναμενόμενες συχνότητες (άρα και η προσαρμογή) των μοντέλων συνάφειας παραμένουν αναλλοίωτες σε γραμμικούς μετασχηματισμούς των σκορ.
8. Ο ρόλος του συστήματος στάθμισης: παρόλο που η επιλογή των βαρών δεν επηρεάζει την προσαρμογή του μοντέλου, έχει αντίκτυπο στη φυσική ερμηνεία και απόδοση του μοντέλου. Τα ομοιόμορφα βάρη είναι προτιμότερα όταν οι κατανομές

περιθωρίου δεν είναι σταθερές και ενδιαφερόμαστε να συγκρίνουμε πίνακες με άνισες κατανομές περιθωρίου. Οι κατανομές περιθωρίου χρησιμοποιούνται ως βάρη προκειμένου να είναι δυνατή η σύγκριση των μοντέλων συνάφειας με την ανάλυση αντιστοιχιών καθώς και στη περίπτωση που μέσω του κατάλληλου μοντέλου συνάφειας οδηγούμαστε σε συγχωνεύσεις γραμμών ή/ και στηλών. [Goodman (1985), Becker and Clogg (1989)]

9. Μεγαλύτερη ισχύς στην ανίχνευση της ανεξαρτησίας: το υπό συνθήκη τεστ ανεξαρτησίας, δοθέντος ότι το μοντέλο (U), (R), ή (C) ισχύει, παρουσιάζει ασυμπτωτικά μεγαλύτερη ισχύ από το κλασικό τεστ ανεξαρτησίας (Agresti 1983). Παράδειγμα: ο έλεγχος ανεξαρτησίας, δοθέντος ότι το μοντέλο (R) είναι αποδεκτό, επιτυγχάνεται δια του ελέγχου της σημαντικότητας της διαφοράς $G^2(I \mid R) = G^2(I) - G^2(R)$ που ακολουθεί ασυμπτωτικά χ^2-κατανομή με d.f.(I|R)=d.f.(I) –d.f.(R).
10. Το δεσμευμένο τεστ ανεξαρτησίας δοθέντος ότι το μοντέλο (RC) ισχύει δεν μπορεί να γίνει, καθώς η αντίστοιχη διαφορά G^2(I|RC) δεν ακολουθεί ασυμπτωτικά χ^2-κατανομή Goodman(1986). Η ίδια λογική μπορεί να υιοθετηθεί και για τον έλεγχο προσαρμογής ενός μοντέλου συνάφειας, δοθέντος ότι κάποιο πολυπλοκότερο ισχύει [π.χ. (R|RC), (C|RC) ή (U|R)].

3.13 Στοιχεία των μοντέλων για συμφωνία μεταξύ των βαθμολογητών

Μετά την παραπάνω ανάλυση μπορούμε να καταλήξουμε ότι τα μοντέλα για συμφωνία μεταξύ βαθμολογητών σε δεδομένα διατάξιμων μεταβλητών αποτελούνται από δύο χαρακτηριστικά. Πρώτον, υπάρχει αξιοσημείωτη θετική συσχέτιση μεταξύ των ταξινομήσεων των δύο βαθμολογητών. Δεύτερον, τα κελιά της κυρίας διαγωνίου έχουν κύρια σημασία εφόσον παρουσιάζουν τις περιπτώσεις συμφωνίας μεταξύ των βαθμολογητών. Έτσι τα μοντέλα αυτά λαμβάνουν υπόψη τους τα δύο παραπάνω στοιχεία περιέχοντας στη μορφή τους μια συνιστώσα για τη συσχέτιση και μια άλλη για τη συμφωνία.

3.13.1 Η συνιστώσα της συσχέτισης

Αναλύοντας τη συσχέτιση σε πίνακες συνάφειας με διατάξιμες μεταβλητές, ο Goodman (1979b) πρότεινε να επικεντρώσουμε την προσοχή μας στα odds ratio των 2x2 πινάκων που

σχηματίζονται από γειτονικά κελιά (πχ. Από τη γραμμές i, i+1 και από τις στήλες j, j+1). Έτσι, αν συμβολίσουμε με m_{ij} τις αναμενόμενες συχνότητες στα κελιά, τα odds ratio των 2x2 πινάκων διαδοχικών κατηγοριών ορίζονται ως εξής:

$$\theta_{ij} = (m_{ij}\, m_{i+1,j+1}) / (m_{i+1,j}\, m_{i,j+1}),\ \ i., j=1, \ldots, I\text{-}1 \qquad (3.17)$$

Για έναν τετραγωνικό πίνακα συνάφειας, η ολική συσχέτιση που παρουσιάζει in the cross-classification μπορεί να εκφραστεί σε όρους των $(I\text{-}1)^2$ odds ratios. Ο Goodman (1979b) παρουσίασε μοντέλα σε όρους odds ratios που ποικίλουν από απλά σε σύνθετα και είναι natural candidates για να περιγράψουν συσχέτιση μεταξύ βαθμολογητών. Ωστόσο, σύμφωνα με τους Darroch και MacCloud (1986), μόνο τα μοντέλα του Goodman που ικανοποιούν μια quasi-συμμετρική δομή είναι κατάλληλα για τη συνιστώσα συσχέτισης. Στον Πίνακα 3.6 καθώς και στην παρουσίαση που ακολουθεί τα odds ratio και οι αναμενόμενες συχνότητες που βασίζονται σε μοντέλα συσχέτισης σημειώνονται με τόνο. Δηλαδή αν το μοντέλο συσχέτισης είναι το πραγματικό μοντέλο που προσαρμόζεται στις συχνότητες του πίνακα, τότε $m'_{ij} = m_{ij}$ και $\theta'_{ij} = \theta_{ij}$. Το εύρος των παραμέτρων για τα μοντέλα στον Πίνακα 3.6, είτε σε όρους odds ratio είτε σε όρους αναμενόμενων συχνοτήτων, είναι: θ_i>0, α_i>0, β_i>0, γ_i>0, $\tilde{g}_i > 0$, $\infty < f_i < \infty$, $\infty < m_i < \infty$, $\infty < \tilde{m}_i < \infty$, όπου i = 1,……, I. Τα δύο είδη των γαμμα-παραμέτρων του μοντέλου I* σε όρους odds ratio και αναμενόμενων συχνοτήτων συνδέονται ως εξής: $\tilde{g}_i = g_{i+1} / g_i$. Τα δύο είδη των μ-παραμέτρων του μοντέλου II* σε όρους odds ratio και αναμενόμενων συχνοτήτων συνδέονται ως εξής: $\tilde{m}_i = m_{i+1} - m_i$.

Πίνακας 3.6

Quasi-συμμετρικά μοντέλα συσχέτισης του Goodman σε όρους odds ratio και multiplicative μοντέλα για τις αναμενόμενες συχνότητες.

Μοντέλο συσχέτισης	Odds ratio	Αναμενόμενη συχνότητα	Βαθμοί ελευθερίας
Μοντέλο Μηδενικής συσχέτισης (ανεξαρτησίας)	$\theta'_{ij\,=\,1}$	$m'_{ij} = \alpha_i \beta_j$	$(I\text{-}1)^2$
Μοντέλο Ομοιόμορφης συσχέτισης	$\theta'_{ij\,=\,\theta}$	$m'_{ij} = \alpha_i \beta_j e^{\varphi ij}$	$(I\text{-}1)^2$ - 1
Μοντέλο I*	$\theta'_{ij} = q\tilde{g}_i\tilde{g}_j$	$m'_{ij} = \alpha_i \beta_j \theta^{ij} \gamma_i^j \gamma_j^i$	$(I\text{-}1)(I\text{-}2)$
Μοντέλο II*	$\theta'_{ij} = e^{f\tilde{m}_i\tilde{m}_j}$	$m'_{ij} = \alpha_i \beta_j e^{f m_i m_j}$	$(I\text{-}1)(I\text{-}2)$

Θεωρούμε τα μοντέλα του Πίνακα 3.6. Το μοντέλο της μηδενικής συσχέτισης είναι ταυτόσημο του μοντέλου ανεξαρτησίας μεταξύ γραμμών και στηλών ταξινόμησης. Το μοντέλο ομοιόμορφης συσχέτισης υποθέτει ότι η ένταση της συσχέτισης σε όρους γειτονικών odds ratio είναι σταθερή σε όλο τον πίνακα συνάφειας. Και τα δύο μοντέλα I* και II* επιτρέπουν διαφορετικό βαθμό συσχέτισης στα γειτονικά κελιά εισάγοντας πρόσθετες παραμέτρους (γ_i ή $\tilde{g}_i$ στο μοντέλο I*, μ_i ή $\tilde{m}_i$ στο μοντέλο II*). Τα μοντέλα I* και II* ονομάζονται επίσης και μοντέλα ομογενούς αλληλεπίδρασης γραμμών –στηλών, εφόσον οι πρόσθετες παράμετροι είναι ίδιες για την ταξινόμηση κατά στήλες ή κατά γραμμές, αν και οι τιμές τους είναι δυνατό να διαφέρουν. Πιο συγκεκριμένα, το μοντέλο II* είναι το (RC) μοντέλο με περιορισμό $\mu_i = \nu_i$, i= 1, …, I. Ωστόσο, μοντέλα που έχουν διαφορετικές παραμέτρους γραμμών και στηλών δεν ικανοποιούν τη quasi-συμμετρική δομή της (3.11). Για το λόγο αυτό τα μοντέλα αυτά δεν συμπεριλαμβάνονται στον Πίνακα 3.6.

Τα μοντέλα σε όρους odds ratio μπορούν να προκύψουν από τα μοντέλα αναμενόμενων συχνοτήτων χρησιμοποιώντας την ερμηνεία των γειτονικών odds ratio (βλ. εξίσωση (3.17)). Αντίθετα, τα μοντέλα σε όρους odds ratio μπορούν να εξελιχθούν σε μοντέλα αναμενόμενων συχνοτήτων χρησιμοποιώντας τη σχέση:

$$m_{ij} = a_i b_j \prod_{s=1}^{i-1} \prod_{t=1}^{j-1} q_{st} \,, \tag{3.18}$$

όπου α_i >0 και β_i >0 είναι οι επιδράσεις γραμμών και στηλών αντίστοιχα και θ_{st} = 1 αν s=1 ή t=1 (Goodman, 1979a)

3.13.2 Η συνιστώσα της συμφωνίας.

Επειδή τα κελιά της κύριας διαγωνίου έχουν ιδιαίτερη σημασία όταν μοντελοποιούμε δεδομένα για εξέταση συμφωνίας, τα μοντέλα πρέπει να περιέχουν παραμέτρους που σκοπό έχουν να εξηγήσουν συμφωνία η οποία δεν εξηγείται από τη συνιστώσα της συσχέτισης. Η πιο απλή σκέψη είναι να υποθέσουμε μια σταθερή τιμή της δ_{ij} για όλα τα διαγώνια κελιά

$$d_{ij} = \begin{cases} d, & \text{i = j} \\ 0, & \text{αλλιώς} \end{cases} \tag{3.19}$$

Μια δεύτερη οπτική είναι να εισάγουμε μια ξεχωριστή παράμετρο για κάθε κελί στη κύρια διαγώνιο, δηλαδή , $\delta_{ij} = \delta_i$ για $i = j$ και $\delta_{ij} = 0$ αλλιώς. Αυτό σημαίνει ότι το μοντέλο θα προσαρμόζεται τέλεια στη κύρια διαγώνιο το οποίο είναι ισοδύναμο με το να παραλείπονται αυτά τα κελιά και να θεωρούμε έτσι έναν ημιτελή πίνακα. Αυτή η θεώρηση προέρχεται από

το μοντέλο quasi-συμμετρίας. Μια τρίτη θεώρηση είναι ένας συμβιβασμός μεταξύ της παραμέτρου δ για μοντελοποίηση των κελιών της κύριας διαγωνίου και των παραμέτρων δ_i για την ακριβή προσαρμογή της κυρίας διαγωνίου. Συγκεκριμένα, θεωρούμε μια παράμετρο για ένα υποσύνολο των κελιών της κύριας διαγωνίου και τα υπόλοιπα διαγώνια κελιά προσαρμόζονται ακριβώς. Έτσι, τα δ_{ij} προσδιορίζονται ως $\delta_{ij} = \delta$ για $i = j$, $i \in$ S και $\delta_{ij} = \delta_i$ για $i = j$, $i \notin$ S και $\delta_{ij} = 0$ αλλιώς, όπου S είναι ένα υποσύνολο του {1, 2, …, I}. Επειδή η (3.19) είναι το πιο φειδωλό μοντέλο για τη συνιστώσα συμφωνίας, μόνο αυτή η θεώρηση χρησιμοποιείται σαν μέρος των μοντέλων συμφωνίας που περιγράφονται στις επόμενες ενότητες.

3.14 Μοντέλα συμφωνίας για δεδομένα διατάξιμων μεταβλητών

Σε όρους αναμενόμενων συχνοτήτων και οι δύο συνιστώσες που είδαμε στην παράγραφο 3.13 υποθέτουμε ότι περιγράφονται από την σχέση:

$$m_{ij} = m'_{ij} e^{d_{ij}} \tag{3.20}$$

η οποία υποθέτει για τα odds ratio ότι

$$q_{ij} = \frac{m_{ij} m_{i+1,j+1}}{m_{i+1,j} m_{i,j+1}} = q'_{ij} e^{d_{ij} - d_{i+1,j} - d_{i,j+1} + d_{i+1,j+1}} \tag{3.21}$$

Τα μοντέλα των σχέσεων (3.20) και (3.21) θα αναφέρονται στο εξής σαν μοντέλα συμφωνίας. Εφόσον βασιστήκαμε στη σχέση (3.19), η γενικότερη έκφραση για τα odds ratio των μοντέλων συμφωνίας που θα δούμε παρακάτω απλοποιείται στην εξής:

$$q_{ij} = \begin{cases} q'_{ij} e^{2d}, & |i - j| = 0 \\ q'_{ij} e^{-d}, & |i - j| = 1 \\ q'_{ij}, & |i - j| > 1 \end{cases} \tag{3.22}$$

Οι βαθμοί ελευθερίας των καταλοίπων των μοντέλων συμφωνίας μπορούν να υπολογιστούν με απλό τρόπο. Εφόσον ο όρος συμφωνίας των μοντέλων που περιγράφονται παρακάτω περιέχει μόνο την παράμετρο δ, ο αριθμός των βαθμών ελευθερίας για ένα συγκεκριμένο μοντέλο είναι πάντα ένας λιγότερος από τους βαθμούς ελευθερίας του μοντέλου συσχέτισης που δίνεται στον Πίνακα 3.6. Ωστόσο αν χρησιμοποιείται άλλου τύπου όρου συμφωνίας, ο αριθμός των διαφορετικών δ παραμέτρων πρέπει να αφαιρεθεί από τους

βαθμούς ελευθερίας του αντίστοιχου μοντέλου συσχέτισης ώστε να καταλήξουμε στους βαθμούς ελευθερίας του συγκεκριμένου μοντέλου συμφωνίας.

3.14.1 Το μοντέλο των Tanner και Young.

Το πιο απλό μοντέλο σε όρους odds ratios, το μοντέλο μηδενικής συσχέτισης, είναι ισοδύναμο με το μοντέλο ανεξαρτησίας μεταξύ των μεταβλητών γραμμών και στηλών. Εφόσον για το μοντέλο μηδενικής συσχέτισης θ'_{ij} =1, προκύπτει από τη σχέση (3.22) ότι τα odds ratio για το συγκεκριμένο μοντέλο συμφωνίας είναι:

$$q_{ij} = \begin{cases} e^{2d}, |\, i-j\,| = 0 \\ e^{-d}, |\, i-j\,| = 1 \\ 1, |\, i-j\,| > 1 \end{cases}$$

Οι Tanner και Young (1985a) πρότειναν αυτό το μοντέλο σε λογαριθμογραμμική μορφή

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \delta I(i=j)$$

το οποίο μπορεί να προκύψει από την εξίσωση (3.13). όπου $I(i=j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$, είναι μια δείκτρια συνάρτηση. Το μοντέλο έχει $(I\text{-}1)^2\text{-}1$ βαθμούς ελευθερίας καταλοίπων και εφαρμόζεται καλύτερα σε κατηγορικά δεδομένα εφόσον οι μεταθέσεις γραμμών και των αντίστοιχων στηλών δεν διαφοροποιούν την προσαρμογή του μοντέλου και την εκτίμηση της παραμέτρου δ.

3.14.2 Το μοντέλο του Agresti

Το μοντέλο μηδενικής ανεξαρτησίας δεν ταιριάζει καλά για ordinal δεδομένα συμφωνίας γιατί συνήθως οι βαθμολογήσεις των παρατηρητών σχετίζονται θετικά. Ένα απλό μοντέλο που ταιριάζει σε τέτοιες περιπτώσεις, υποθέτει ότι αυτή η συσχέτιση είναι ομοιόμορφη σε όλη την επιφάνεια του πίνακα, δηλαδή $\theta_{ij}' = \theta$. τότε το μοντέλο συμφωνίας έχει ως εξής:

$$q_{ij} = \begin{cases} qe^{2d}, |\, i-j\,| = 0 \\ qe^{-d}, |\, i-j\,| = 1 \\ q, |\, i-j\,| > 1 \end{cases}$$

το οποίο είναι μια ειδική περίπτωση της περίπτωσης (3.22). Χρησιμοποιώντας την εξίσωση (3.18), το μοντέλο σε λογαριθμογραμμική μορφή είναι

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \varphi ij + \delta I(i=j) \qquad (3.23)$$

Τα ακέραια σκορ *i* και *j* στο γινόμενο *φij* της εξίσωσης (3.23) είναι ισοδύναμα με τις δείκτριες γραμμών και στηλών, αντίστοιχα. Εύκολα αποδεικνύεται ότι $\varphi = \log\theta$. Το μοντέλο έχει $(I\text{-}1)^2\text{-}2$ βαθμούς ελευθερίας (εδώ έχουμε δύο παραμέτρους επιπλέον της ανεξαρτησίας τις φ και δ). Αν αντικαταστήσουμε τα σκορ *i, j,* με τα γενικά σκορ μ_i και μ_j καταλήγουμε στο μοντέλο

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \varphi\mu_i\mu_j + \delta I(i=j) \qquad (3.24)$$

Παρόλο που το μοντέλο (3.24) είναι πολύ πιο πολύπλοκο από το (3.23), παραμένει λογαριθμογραμμικό όσο τα μ_i είναι γνωστά. Ωστόσο, είναι σπάνια περίπτωση, αφού η έλλειψη μιας καθαρής ερμηνείας των κατηγοριών δυσκολεύει τη δικαιολόγηση ανάθεσης εκ των προτέρων σκορ με πολύπλοκη δομή. Για το λόγο αυτό, τα σκορ που ισαπέχουν για διαδοχικές κατηγορίες της μεταβλητής ταξινόμησης χρησιμοποιούνται πιο συχνά.

3.14.3 Το λογαριθμο-πολλαπλασιαστικό μοντέλο συμφωνίας

Είναι επίσης πιθανό να θεωρήσουμε τα scores του μοντέλου (3.24) ως άγνωστες παραμέτρους που πρέπει να εκτιμηθούν από τα δεδομένα. Αυτό μελετήθηκε από τον Agresti (1988), τον Becker (1989, 1990b) και τους Graham and Jackson (1993). Στην περίπτωση αυτή το μοντέλο δεν είναι λογαριθμογραμμικό και θα το αναφέρουμε στο εξής ως λογαριθμο-πολλαπλασιαστικό. Έχει $(I\text{-}1)^2\text{-}I$ βαθμούς ελευθερίας. Το μοντέλο αυτό μπορεί να παραχθεί από το μοντέλο II* του Πίνακα 3.6. έτσι, σε όρους odds ratios το μοντέλο είναι

$$q_{ij} = \begin{cases} e^{2d + f\tilde{m}_i^2}, |\, i-j\,| = 0 \\ e^{-d + f\tilde{m}_i\tilde{m}_j}, |\, i-j\,| = 1 \\ e^{f\tilde{m}_i\tilde{m}_j}, |\, i-j\,| > 1 \end{cases}$$

3.14.4 Ένα νέο λογαριθμογραμμικό μοντέλο για συμφωνία διατάξιμων δεδομένων.

Εφόσον το λογαριθμο-πολλαπλασιαστικό μοντέλο συμφωνίας μπορεί να εξαχθεί από το μοντέλο II* του Πίνακα 3.6, φαίνεται φυσικό να εξάγουμε και ένα μοντέλο συμφωνίας από το μοντέλο I* του Πίνακα 3.6. Δεν υπάρχει σημαντικός λόγος για να θεωρήσουμε το ένα μοντέλο ανώτερο από το άλλο αφού και τα δύο έχουν τον ίδιο βαθμό πολυπλοκότητας όσον αφορά τα odds ratios. Σε όρους odds ratios το νέο μοντέλο συμφωνίας είναι

$$q_{ij} = \begin{cases} e^{2d} q\tilde{g}_i^2, |\, i-j\,| = 0 \\ e^{-d} q\tilde{g}_i\tilde{g}_j, |\, i-j\,| = 1 \\ q\tilde{g}_i\tilde{g}_j, |\, i-j\,| > 1 \end{cases} \qquad (3.25)$$

χρησιμοποιώντας την εξίσωση (3.18), το μοντέλο μπορεί να γραφτεί σε λογαριθμογραμμική μορφή ως,

$$\log m_{ij} = l + l_i^A + l_j^B + fij + jz_i + iz_j + \delta I(i=j) \qquad (3.26)$$

όπου φ και ζ είναι παράμετροι που ορίζονται ως εξής: $\varphi=\log\theta$ και $\zeta_i=\log\gamma_i$. Οι βαθμοί ελευθερίας αυτού του μοντέλου είναι $(I\text{-}1)^2\text{-}I$. Ο περιορισμός $\zeta_1=\zeta_2=\ldots=\zeta_I=0$ μειώνει το μοντέλο στο μοντέλο του Agresti με integer-scores.

Για να είναι προσδιορίσιμοι οι παράμετροι αυτού του μοντέλου, πρέπει να θέσουμε περιορισμούς σε αυτές. Έτσι οι παράμετροι των κύριων επιδράσεων μπορούν να περιοριστούν χρησιμοποιώντας τους περιορισμούς $\sum_i l_i^A = 0$ και $\sum_i l_i^B = 0$. Για να ορίσουμε τις ζ- παραμέτρους ομοιόμορφα, πρέπει να θέσουμε δύο περιορισμούς, όπως για παράδειγμα $\zeta_1 = \zeta_I = 0$ ή $\sum_i z_i = 0$.

Ας επανέλθουμε στο παράδειγμα με τα δεδομένα του Πίνακα 3.2 που αφορούσαν την κατανάλωση αλκοόλ όπως τη δηλώνουν οι ίδιοι οι ασθενείς που έχουν υποστεί μη θανατηφόρο έμφραγμα του μυοκαρδίου και όπως τη δηλώνουν οι άμεσοι συγγενείς τους. Ο Πίνακας 3.7 περιέχει τα δεδομένα του Πίνακα 3.2 και τις αντίστοιχες εκτιμήσεις των αναμενόμενων συχνοτήτων για κάθε κελί σύμφωνα με το νέο λογαριθμογραμμικό μοντέλο.

Στον Πίνακα 3.8 φαίνονται τα αποτελέσματα από την προσπάθεια να προσαρμόσουμε τα δεδομένα στα μοντέλα ανεξαρτησίας, Tanner and Young, Agresti, το νέο μοντέλο συμφωνίας, το λογαριθμο-πολλαπλασιαστικό μοντέλο και το μοντέλο quasi-συμμετρίας.

Εφόσον στον πίνακα υπάρχει ένας αξιοσημείωτος αριθμός κελιών με λίγες παρατηρήσεις μπορεί το στατιστικό X^2 του Pearson και το likelihood ratio στατιστικό G^2 που ακολουθούν χ^2 κατανομές να μην είναι κατάλληλα στατιστικά για να χρησιμοποιηθούν στην ανάλυση. Ωστόσο το G^2 μπορεί να χρησιμοποιηθεί για να συγκρίνει nested μοντέλα (Agresti and Yang, 1986, Haberman, 1977). Είναι ενδιαφέρον ότι η προσαρμογή του μοντέλου σε όρους likelihood ratio βελτιώνεται σημαντικά όσο τα μοντέλα γίνονται πιο πολύπλοκα. Για το λογαριθμο-πολλαπλασιαστικό μοντέλο και το νέο λογαριθμογραμμικό μοντέλο η προσαρμογή είναι περίπου ίδια.

Πίνακας 3.7

Παρατηρούμενες και προσαρμοσμένες συχνότητες (εντός των παρενθέσεων) με βάση το μοντέλο (3.26) 456 ζευγών ασθενών και συγγενών τους ως προς τη συχνότητα χρήσης αλκοόλ σε μια μελέτη case-controlγια τη στεφανιαία νόσο.

	Ασθενείς					
Άμεσοι συγγενείς	«Δεν πίνω αλκοόλ»	«Το έχω κόψει»	>1/μήνα <1/εβδομάδα	>1/εβδομάδα <1/μέρα	>1/μέρα	Σύνολο
«Δεν πίνω αλκοόλ»	47 (46.970)	13 (12.603)	19 (20.547)	4 (2.657)	0 (0.206)	83
«Το έχω κόψει»	5 (5.577)	6 (4.325)	2 (4.986)	1 (1.004)	2 (0.108)	16
>1/μήνα <1/εβδομάδα	15 (13.653)	6 (7.487)	76 (73.172)	19 (20.288)	4 (5.300)	120
>1/εβδομάδα <1/μέρα	1 (1.631)	1 (1.383)	23 (18.706)	54 (57.714)	22 (21.566)	101
>1/μέρα	0 (0.170)	0 (0.202)	4 (6.589)	33 (29.220)	99 (99.819)	136
Σύνολο	68	26	124	111	127	456

Πίνακας 3.8

Τιμές των X^2 και G^2 στατιστικών με τους αντίστοιχους βαθμούς ελευθερίας για τα μοντέλα.

Μοντέλο	G^2	X^2	Βαθμοί ελευθερίας
Μηδενικής Συσχέτισης (ανεξαρτησίας)	470.78	481.84	16
Μηδενικής συσχέτισης + δ	156.93	134.34	15
Ομοιόμορφης συσχέτισης + δ	41.61	41.81	14
Μοντέλο I* + δ (λογαριθμο-πολλαπλασιαστικό)	16.92	40.56	11
Μοντέλο II* + δ	17.06	48.67	11
Quasi-συμμετρίας	7.14	6.58	7

Ωστόσο, οι τιμές του X^2 δίνουν διαφορετικά αποτελέσματα. Μεγάλες διαφορές μεταξύ των στατιστικών G^2 και X^2 έχουν αναφερθεί και από τον Becker (1989). Όπως οι Koehler and Larntz (1980) σημειώνουν ότι υπάρχουν διαφορές στη χρήση των δύο στατιστικών G^2 και X^2, με το δεύτερο να είναι πιο ευαίσθητο στην ανάλυση πινάκων με κελιά με μικρό αριθμό παρατηρήσεων. Αυτό επεξηγεί και τη μεγάλη τιμή του X^2 για το λογαριθμο-πολλαπλασιαστικό μοντέλο. Όπως επισήμαναν και οι Graham and Jackson (1993, p.1060) το κελί με τη μικρότερη αναμενόμενη συχνότητα είναι εκείνο που οδηγεί στην έλλειψη προσαρμογής του X^2 – στην συγκεκριμένη περίπτωση το κελί (2,5). (Παρόλο που ο Πίνακας 3.7 έχει βασιστεί στο νέο λογαριθμογραμμικό μοντέλο, οι προσαρμοσμένες τιμές για το λογαριθμο-πολλαπλασιαστικό είναι σχεδόν ίδιες).

Εφόσον το νέο λογαριθμογραμμικό μοντέλο είναι ιεραρχικά συνδεδεμένο με το συνηθισμένο μοντέλο quasi-συμμετρίας, η διαφορά στις τιμές των G^2 στατιστικών των δύο μοντέλων μπορεί να εξετασθεί στατιστική σημαντικότητα. Σύμφωνα με τον Πίνακα 3.8, ΔG^2=9,78 βασισμένο σε βαθμούς ελευθερίας df=5, με p-value = 0.0817.

Ο Πίνακας 3.9 δίνει τις παραμέτρους που εκτιμήθηκαν για το νέο λογαριθμο-γραμμικό μοντέλο συμφωνίας και που είναι σημαντικό να ερμηνευθούν. Εκτός από την ζ_4, όλες οι εκτιμήσεις των παραμέτρων έχουν τιμές μεγάλες σε σχέση με το τυπικό τους σφάλμα. Από τον Πίνακα 3.9 οι τιμές των παραμέτρων του μοντέλου σε όρους odds ratios μπορούν εύκολα να βρεθούν.

Πίνακας 3.9

Εκτιμήσεις παραμέτρων και το τυπικό τους σφάλμα

Παράμετροι	Εκτιμήσεις	s(est.)	exp(est.)
$\hat{f}$	0.657	0.103	1.930
$\hat{z}_2$	-0.446	0.164	0.640
$\hat{z}_3$	-0.355	0.081	0.701
$\hat{z}_4$	-0.129	0.108	0.879
$\hat{d}$	0.649	0.152	1.910

Πιο συγκεκριμένα, εφόσον $\hat{q} = \exp(\hat{f}) = 1.930$, $\hat{g}_2 = \exp(\hat{z}_2) = 0.640$, $\hat{g}_3 = \exp(\hat{z}_3) = 0.701$, $\hat{g}_4 = \exp(\hat{z}_4) = 0.879$, οι εκτιμήσεις αυτές μαζί με τους περιορισμούς $g_1 = \exp(z_1) = 1$ και $g_5 = \exp(z_5) = 1$ και τη σχέση $\tilde{g}_i = g_{i+1}/g_i$ μας δίνουν τις τιμές των $\hat{\tilde{g}}_1 = 0.64$, $\hat{\tilde{g}}_2 = 1.10$, $\hat{\tilde{g}}_3 = 1.25$, $\hat{\tilde{g}}_4 = 1.14$.

Κάνοντας χρήση της εξίσωσης (3.25), οι εκτιμήσεις των τοπικών odds ratios $q_{ij} = \frac{p_{ij}p_{i+1,j+1}}{p_{i,j+1}p_{i+1,j}}$ (local odds ratios) μπορούν να υπολογιστούν με βάση τις εκτιμήσεις των θ, δ και $\tilde{g}$ παραμέτρων. Για παράδειγμα ισχύει

$$\hat{q}_{11} = \exp(2\hat{d})q\,\hat{\tilde{g}}_1^2 = 3.67 * 1.93 * 0.41 = 2.90$$

και η τιμή αυτή σχεδόν ταυτίζεται με το αντίστοιχο local odds ratio που υπολογίζεται από τις προσαρμοσμένες τιμές του μοντέλου που δίνονται στον Πίνακα 3.2:

$$\hat{q}_{11} = (46.97/12.603)/(5.577/4.325) = 2.89.$$

Η ερμηνεία του μοντέλου μπορεί να βασιστεί στις παραμέτρους στις οποίες τα τοπικά odds ratio, σύμφωνα με την εξίσωση (3.25), αναλύονται. Πρώτον, η τιμή του $\hat{q} = 1.93$, δείχνει ότι υπάρχει σημαντική θετική συσχέτιση μεταξύ των ασθενών και των συγγενών τους. Δεύτερον, η τιμή της παραμέτρου $\hat{d} = 0.649$ δηλώνει ότι μια ουσιώδης αναλογία της συνολικής πιθανότητας περιέχεται στα διαγώνια κελιά του πίνακα, εφόσον η τιμή της παραμέτρου $\hat{q}$ αυξάνεται κατά έναν παράγοντα $\exp(2\hat{d}) = 3.67$ αν τα local odds ratio περιέχουν δύο διαγώνια κελιά του πίνακα. Τρίτον, οι εκτιμήσεις των παραμέτρων $\tilde{g}_i$ δείχνουν ένα ενδιαφέρον υπόδειγμα. Ενώ η παράμετρος $\hat{\tilde{g}}_1$ φαίνεται να έχει τιμή αρκετά μικρότερη της μονάδας, οι άλλες τρεις παράμετροι $\hat{\tilde{g}}_2$, $\hat{\tilde{g}}_3$, $\hat{\tilde{g}}_4$ έχουν εκτιμώμενες τιμές του ίδιου περίπου μεγέθους μεγαλύτερες όμως της μονάδας. Έτσι, σε όσα τοπικά odds ratio υπάρχει η παράμετρος $\tilde{g}_1$, που αντιπροσωπεύει την σχέση μεταξύ των κατηγοριών «δεν πίνω αλκοόλ» και «το έχω κόψει», τότε αυτά παρουσιάζονται σημαντικά μειωμένα. Για παράδειγμα, το θ_{11} local odds ratio, που εμπεριέχει την παράμετρο $\tilde{g}_1$, έχει εκτιμηθεί 2.89, ενώ τα θ_{22}, θ_{33} και θ_{44} τα οποία δεν εμπεριέχουν την $\tilde{g}_1$, έχουν εκτιμηθεί 8.48, 11.07 και 9.14 αντίστοιχα. Με την

διαπίστωση αυτή καταλήγουμε στο συμπέρασμα ότι οι ίδιοι οι ασθενείς όσο και οι συγγενείς τους συγχέονται κατά τον διαχωρισμό των δύο πρώτων κατηγοριών, δηλαδή μεταξύ των κατηγοριών «δεν πίνω αλκοόλ» που σημαίνει ποτέ δεν έπινα και της κατηγορίας «το έχω κόψει» (δεν πίνω αλκοόλ, αλλά κάποτε έπινα).

Εφόσον το νέο μοντέλο είναι λογαριθμογραμμικό, είναι εύκολο να ελέγξουμε την υπόθεση αυτή. Δεν είναι δύσκολο να δούμε ότι οι περιορισμοί $z_3 = (2/3)z_2$ και $z_4 = (1/3)z_2$ δηλώνουν ότι $\tilde{g}_2 = \tilde{g}_3 = \tilde{g}_4$. Επαναπροσαρμόζοντας το μοντέλο με αυτούς τους δύο πρόσθετους περιορισμούς καταλήγουμε σε ένα πηλίκο πιθανοφάνειας G^2=17.38 με 13 βαθμούς ελευθερίας. Η διαφορά μεταξύ αυτού του μοντέλου και του μοντέλου χωρίς τους περιορισμούς είναι μόνο 0.46 με 2 βαθμούς ελευθερίας. Έτσι φαίνεται καθαρά ότι η υπόθεση μιας κοινής τιμής για τις παραμέτρους $\hat{\tilde{g}}_2$, $\hat{\tilde{g}}_3$, $\hat{\tilde{g}}_4$ είναι βάσιμη. Οι εκτιμήσεις για τις παραμέτρους $\tilde{g}_i$ στο μοντέλο με τους περιορισμούς είναι $\hat{\tilde{g}}_1 = 0.75$ και $\hat{\tilde{g}}_2 = \hat{\tilde{g}}_3 = \hat{\tilde{g}}_4 = 1.10$.

ΚΕΦΑΛΑΙΟ 4

ΑΝΑΛΥΣΗ ΣΥΜΦΩΝΙΑΣ ΚΑΙ ΔΙΑΦΩΝΙΑΣ ΜΕΤΑΞΥ ΠΟΛΛΩΝ ΒΑΘΜΟΛΟΓΗΤΩΝ

4.1 Συμφωνία μεταξύ πολλαπλών βαθμολογητών σε κατηγοριοποίηση 0 – 1

4.1.1. Ανταλλάξιμες βαθμολογήσεις

Στην ενότητα αυτή θεωρούμε την έννοια της συμφωνίας σε μια κατάσταση όπου διαφορετικά υποκείμενα (ή αντικείμενα ή περιπτώσεις) αξιολογούνται από διαφορετικούς βαθμολογητές και ο αριθμός των βαθμολογητών για κάθε αντικείμενο ποικίλει. Επικεντρωνόμαστε στην περίπτωση όπου η αξιολόγηση έχει δύο κατηγορίες. (π.χ. ναι – όχι, άρρωστο – υγιές, κλπ.). Ο Fleiss και ο Cuzick (1979) παρείχαν ένα παράδειγμα όπου υποκείμενα θεωρήθηκαν νοσηλευόμενοι ψυχικά άρρωστοι ασθενείς, το χαρακτηριστικό υπό μελέτη ήταν η παρουσία ή η απουσία σχιζοφρένειας, και οι βαθμολογητές ήταν οι ψυχίατροι που διέμεναν στην κλινική, οι οποίοι επιλέχτηκαν από ένα μεγαλύτερο σύνολο ψυχιάτρων οι οποίοι καλούνται στην κλινική κατά την εισαγωγή ασθενούς. Στη συγκεκριμένη περίπτωση κάθε ψυχίατρος μπορεί να είναι υπεύθυνος και να αξιολογεί συγκεκριμένους ασθενείς και επίσης διαφορετικός αριθμός ψυχιάτρων μπορεί να αξιολογεί κάθε ασθενή.

Για διευκόλυνση, ας θεωρήσουμε n τον αριθμό των υπό μελέτη υποκειμένων, K_i τον αριθμό των βαθμολογητών που αξιολογεί το i υποκείμενο. Ας θεωρήσουμε επίσης y_{ih} την βαθμολόγηση του h βαθμολογητή για το i υποκείμενο, με $y_{ih} = 1$, αν υπάρχει παρουσία της υπόθεσης και $y_{ih} = 0$ όταν η υπόθεση δεν υφίσταται, και τέλος, ας θεωρήσουμε

$$y_i = \sum_{h=1}^{K_i} y_{ih}$$

τον αριθμό των θετικών καταχωρήσεων για το i αντικείμενο. Ας δούμε τον Πίνακα 4.1 για διευκόλυνσή μας:

Πίνακας 4.1

Μορφή δεδομένων για πολλαπλούς βαθμολογητές και δύο κατηγορίες ταξινόμησης

	Υποκείμενα				
	1	2	3	…	n
Αριθμός θετικών αξιολογήσεων	y_1	y_2	y_3	…	y_n
Αριθμός βαθμολογητών	K_1	K_2	K_3	…	K_n

Όταν $K_1 = K_2 = \ldots = K_n = K$, ο Fleiss (1971) για να εκτιμήσει το ποσοστό παρατηρούμενης και αναμενόμενης συμφωνίας κάτω από την ανεξαρτησία της αξιολόγησης των βαθμολογητών, χρησιμοποίησε τους εκτιμητές αντίστοιχα,

$$\hat{P}_o = 1 - \frac{2}{n}\sum_{i=1}^{n}\frac{y_i(K - y_i)}{K(K-1)}$$

$$\hat{P}_e = 1 - 2\hat{p}(1-2\hat{p}),$$

όπου

$$\hat{p} = \frac{1}{nK}\sum_{i=1}^{n} y_i ,$$

για να ορίσει ένα δειγματικό μέτρο συμφωνίας

$$\hat{k}_f = \frac{\hat{P}_o - \hat{P}_E}{1 - \hat{P}_e} = 1 - \frac{\sum_{i=1}^{n} y_i(K_i - y_i)}{nK(K-1)\hat{p}(1-\hat{p})}.$$

Για διαφορετικούς αριθμούς βαθμολογητών ανά αντικείμενο, οι Fleiss και Cuzick (1979) επέκτειναν την ερμηνεία του k_f, ως εξής:

$$\hat{k}_f = 1 - \frac{1}{n(\bar{K}-1)\hat{p}(1-\hat{p})}\sum_{i=1}^{k}\frac{y_i(K_i - y_i)}{K_i},$$

όπου

$$\bar{K} = \frac{1}{n}\sum_{i=1}^{n} K_i$$

είναι ο μέσος αριθμός βαθμολογητών ανά αντικείμενο, και

$$\hat{p} = \frac{1}{n\bar{K}}\sum_{i=1}^{n} y_i .$$

Έδειξαν, επίσης, ότι όταν το $n \to \infty$, ο εκτιμητής $\hat{k}_f$ είναι ασυμπτωτικά ισοδύναμος της εκτιμημένης εντός των τάξεων συσχέτισης $\hat{r}$ που δίνεται από τον τύπο:

$$\hat{\rho} = \frac{MSB - MSW}{MSB + (n_o - 1)MSW},$$

όπου,

$$MSB = \frac{1}{n-1}\left[\sum_i \frac{y_i^2}{K_i} - \frac{(\Sigma y_i)^2}{K'}\right]$$

$$MSW = \frac{1}{K'-n}\left[\sum_i y_i - \sum_i \frac{y_i^2}{K_i}\right]$$

$$n_o = \frac{1}{n-1}\left[K' - \frac{\sum_i K_i^2}{K'}\right]$$

και $K' = \sum_i K_i$. Ένας εκτιμητής για την ασυμπτωτική διακύμανση του $\hat{\rho}$ δόθηκε από τον Mak (1988):

$$Var(\hat{\rho}) = \frac{1}{n}\left[V_{11}C_1^2 + 2C_1C_2V_{12} + V_{22}C_2^2\right],$$

όπου,

$$C_1 = \frac{1}{(\bar{K}-1)\hat{p}(1-\hat{p})}$$

$$C_2 = \frac{1 + (\bar{K}-1)[\hat{\rho} + 2\hat{p}(1-\hat{\rho})]}{\bar{K}(\bar{K}-1)\hat{p}(1-\hat{p})}$$

$$V_{11} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i^4}{K_i^2} - f_i^2\right]$$

$$V_{12} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i^3}{K_i^2} - \hat{p}K_i f_i\right]$$

$$V_{22} = \frac{1}{n}\sum_{i=1}^{n}\left[y_i^2 - \hat{p}^2 K_i^2\right]$$

και

$$f_i = \hat{p}(1-\hat{p})[1 + (K_i - 1)\hat{\rho}] + K_i\hat{p}^2.$$

4.1.2. Τεστ για τα σφάλματα μεταξύ βαθμολογητών

Γνωρίζουμε ότι σε έναν 2x2 πίνακα συχνοτήτων, όπου δύο βαθμολογητές ταξινομούν n αντικείμενα ανάλογα με την ύπαρξη ή όχι ενός χαρακτηριστικού σε δύο κατηγορίες (0 – 1) ταξινόμησης, το τεστ του McNemar (1947) χρησιμοποιείται για τον έλεγχο σφαλμάτων μεταξύ των δύο βαθμολογητών. Αν το $n \to \infty$, τότε το X^2 στατιστικό του McNemar ($X^2 = \frac{(n_{10} - n_{01})^2}{n_{10} + n_{01}}$) ακολουθεί την χ^2 κατανομή με έναν βαθμό ελευθερίας. Πολλοί ερευνητές προχώρησαν σε μια γενίκευση του τεστ του McNemar προκειμένου να ελέγξουν τα περιθώρια ποσοστά σε *I*x*I* πίνακες συχνοτήτων και κατέληξαν στο στατιστικό των Stuart – Maxwell (SM).

Ανάλογα, στην περίπτωση πολλών βαθμολογητών K >2, μας ενδιαφέρει η ένταση της ομοιογένειας των ποσοστών των θετικών καταχωρίσεων από τους βαθμολογητές. Ας δούμε το παράδειγμα που ακολουθεί των Shoukri and Pause (1999).

Στα πλαίσια ενός προγράμματος στο Πανεπιστήμιο Κτηνιατρικής του Οντάριο, ζητήθηκε από τέσσερις τελειόφοιτους φοιτητές να ανιχνεύσουν την ύπαρξη ή όχι αυχενικής σπονδυλικής δυσμορφίας σε 20 πουλάρια. Ανεξάρτητα ο ένας από τον άλλον, ταξινόμησαν τις αντίστοιχες 20 ακτινογραφίες σε δύο κατηγορίες: ύπαρξη ασθένειας («1») και μη ύπαρξη ασθένειας («0»). Τα δεδομένα παρουσιάζονται στον Πίνακα 4.2.

Προφανώς, τα περιθώρια αθροίσματα δείχνουν τις διαφορές στις πιθανότητες ταξινόμησης των τεσσάρων φοιτητών για τις ίδιες 20 ακτινογραφίες. Ο έλεγχος για τη σημαντικότητα των διαφορών μεταξύ των περιθώριων πιθανοτήτων (ή ο έλεγχος για σφάλματα μεταξύ των βαθμολογητών) μπορεί να γίνει χρησιμοποιώντας το στατιστικό Q του Cochran.

Πίνακας 4.2

Ταξινόμηση των ακτινογραφιών από τέσσερις φοιτητές για ανίχνευση της αυχενικής σπονδυλικής δυσμορφίας.

	Βαθμολογητές (φοιτητές)				
Ακτινογραφία	Α	Β	Γ	Δ	Σύνολο
1	0	0	0	0	0
2	0	0	1	0	1
3	1	1	1	1	4
4	1	1	1	1	4

5	1	1	0	1	3
6	0	0	0	0	0
7	1	0	0	0	1
8	0	0	0	0	0
9	1	1	1	1	4
10	1	0	1	1	3
11	1	1	1	1	4
12	1	1	0	1	3
13	1	1	0	0	2
14	1	0	1	0	2
15	1	0	0	0	1
16	0	0	1	0	1
17	1	1	1	1	4
18	1	0	0	0	1
19	1	1	1	1	4
20	1	1	1	1	4
Σύνολο	15	10	11	10	46

Συνοπτικά, τα δεδομένα του Πίνακα 4.2 δίνονται στον 2x2x2x2 Πίνακα 4.3 που διασταυρώνει τις δίτιμες απαντήσεις των τεσσάρων βαθμολογητών.

Πίνακας 4.3

Παρουσίαση των 20 ακτινογραφιών σε 2x2x2x2 πίνακα

			Α βαθμολογητής «0»		Α βαθμολογητής «1»		
			Γ βαθμολογητής «0»	Γ βαθμολογητής «1»	Γ βαθμολογητής «0»	Γ βαθμολογητής «1»	σύνολο
Β βαθμολογητής «0»	Δ βαθμ)της	«0»	3	2	3	1	9
		«1»	0	0	0	1	1
Β βαθμολογητής «1»	Δ βαθμ)της	«0»	0	0	1	0	1
		«1»	0	0	2	7	9
		σύνολο	3	2	6	8	20

Μια γενικότερη μορφή του Πίνακα 4.2 δίνεται στον επόμενο Πίνακα 4.4.

Πίνακας 4.4

Γενική μορφή δεδομένων για πολλαπλούς βαθμολογητές και δύο κατηγορίες ταξινόμησης.

	Βαθμολογητής				
Αντικείμενο	1	2	…	*k*	Σύνολο
1	y_{11}	y_{12}	…	y_{1k}	$y_{1.}$
2	y_{21}	y_{22}	…	y_{2k}	$y_{2.}$
…	…	…	…	…	…
n	y_{n1}	Y_{n2}	…	y_{nk}	$y_{n.}$
Σύνολο	$y_{.1}$	$y_{.2}$	…	$y_{.k}$	$y_{..}$

Ας υποθέσουμε ότι y_{ih} είναι το σκορ του *h* βαθμολογητή στο *i* αντικείμενο ($i = 1, 2, \ldots, n$, $h = 1, 2, \ldots, K$), όπου $y_{ih} = 1$, όταν ο *h* βαθμολογητής έκρινε ότι το *i* αντικείμενο φέρει την ασθένεια και $y_{ih} = 0$ σε κάθε άλλη περίπτωση. Ας υποθέσουμε επίσης ότι $y_{i.}$ είναι το σύνολο των βαθμολογητών που έκριναν το *i* αντικείμενο σαν ασθενές και $y_{.h}$ το άθροισμα των αντικειμένων που κρίθηκαν από τον *h* βαθμολογητή σαν ασθενή. Το στατιστικό *Q* του Cochran δίνεται από τον τύπο:

$$Q = \frac{K(K-1)\sum_{h=1}^{K}\left(y_{.h} - \frac{y_{..}}{K}\right)^2}{Ky_{..} - \sum_{i=1}^{n} y_{i.}^2}$$

όπου

$$y_{..} = \sum_{h=1}^{K} y_{.h} = \sum_{i=1}^{n} y_{i.} \,.$$

Κάτω από τη μηδενική υπόθεση της μη ύπαρξης σφαλμάτων μεταξύ των βαθμολογητών, το *Q* προσεγγίζει την X^2 κατανομή με *n-1* βαθμούς ελευθερίας.

Σημειώνουμε δε, ότι στην περίπτωση δύο βαθμολογητών (*n=2*) το *Q* στατιστικό του Cochran είναι ισοδύναμο του τεστ του McNemar.

Στην συγκεκριμένη περίπτωση, και με βάση τα δεδομένα του πίνακα, *Q=6.375* με 3 βαθμούς ελευθερίας. Συνεπώς σε επίπεδο σημαντικότητας $\alpha = 0,05$, δεν απορρίπτουμε τη μηδενική υπόθεση, δηλαδή δεν υπάρχουν σφάλματα μεταξύ των βαθμολογητών.

4.2. Πολλαπλές κατηγορίες και πολλαπλοί βαθμολογητές

Ο Fleiss (1971) εισήγαγε μια γενίκευση του δείκτη kappa του Cohen για τη μέτρηση συμφωνίας μεταξύ ενός σταθερού αριθμού από *K* βαθμολογητές. Καθένα από τα *n* υποκείμενα ταξινομούνται από *K>2* ανεξάρτητους βαθμολογητές σε μία από τις *I* mutually αποκλειστικές και εξαντλητικές ονοματικές κατηγορίες. Το αρχικό παράδειγμα ήταν μια μελέτη στην οποία καθένας από τους 30 ασθενείς ταξινομούνταν από 6 ψυχιάτρους (οι οποίοι ήταν τυχαία επιλεγμένοι από ένα σύνολο 43 ψυχιάτρων) σε μία από πέντε κατηγορίες. Ας θεωρήσουμε K_{ij} τον αριθμό των βαθμολογητών – ψυχιάτρων που κατέταξαν το *i* υποκειμενο – ασθενή στην *j* κατηγορία ($i=1, 2, \ldots, n$, $j=1, 2, \ldots, I$) . Τότε

$$p_j = \frac{1}{nK}\sum_{i=1}^{n} K_{ij}$$

Εδώ p_j είναι η αναλογία των καταχωρήσεων στην *j* κατηγορία. Το διορθωμένο από τυχαιότητες μέτρο της συνολικής συμφωνίας που πρότεινε ο Fleiss (1971) είναι

$$\hat{k}_{mr} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{I} K_{ij}^2 - Kn\left\{1+(K-1)\sum_{j=1}^{I} p_j^2\right\}}{Kn(K-1)\left(1-\sum_{j=1}^{I} p_j^2\right)} \qquad (4.1)$$

Για ευκολία στους υπολογισμούς, γράφουμε την εξίσωση (4.1) ως

$$\hat{k}_{mr} = \frac{p_0 - p_e}{1-p_e} \qquad (4.2)$$

όπου

$$p_0 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{I} K_{ij}^2 - nK}{Kn(K-1)} \qquad (4.3)$$

και

$$p_e = \sum_{j=1}^{I} p_j^2 \tag{4.4}$$

όπου

$$p_j = \frac{1}{nK} \sum_{i=1}^{n} K_{ij} \tag{4.5}$$

Κάτω από τη μηδενική υπόθεση της μη συμφωνίας πέραν της τυχαίας, οι *K* καταχωρίσεις κάθε αντικειμένου είναι πολυωνυμικές κατανομές με πιθανότητες p_1, p_2, …, p_I. Χρησιμοποιώντας αυτό, ο Fleiss (1971) κατέληξε σε μια περίπου ασυμπτωτική διακύμανση του $\hat{k}_{mr}$ κάτω από τη μηδενική υπόθεση ότι δεν υπάρχει άλλου είδους συμφωνία πέρα από την τυχαία συμφωνία:

$$Var(\hat{k}_{mr}) = \frac{2}{Kn(n-1)} x \frac{\sum_{j=1}^{I} p_j^2 - (2n-3)(\sum_{j=1}^{I} p_j^2)^2 + 2(n-2)\sum_{j=1}^{I} p_j^3}{(1-\sum_{j=1}^{I} p_j^2)^2}$$

Η κατασκευή των διαστημάτων εμπιστοσύνης του $\hat{k}_{mr}$ είναι δύσκολη εξαιτίας του γεγονότος ότι μόνο η ασυμπτωτική διακύμανση, κάτω από τη μηδενική υπόθεση της μη ύπαρξης συμφωνίας πέρα από την τυχαία συμφωνία είναι γνωστή. Οι Davis and Fleiss (1982) συζητάνε μερικές ενδιαφέρουσες εφαρμογές όπου η υπόθεση $\hat{k}_{mr} = 0$ μπορεί να παρουσιάζει ενδιαφέρον. Συγκεκριμένα, στην εκτίμηση της κληρονομικότητας κάποιων ψυχιατρικών διαταραχών, όπου διάφορα μέλη των οικογενειών ρωτήθηκαν να αναφέρουν δεδομένα για τα υπόλοιπα μέλη ανεξάρτητα από αυτά, η αποτυχία απόρριψης της μηδενικής υπόθεσης μπορεί να σημαίνει ότι η εμπιστοσύνη στις αναφορές πληροφοριών διαθέσιμων συγγενών για μη διαθέσιμους συγγενείς τους μπορεί να οδηγήσει σε τελείως αναξιόπιστα δεδομένα.

Επιπρόσθετα, ο Fleiss (1971) πρότεινε και ένα άλλο στατιστικό μέτρησης του μεγέθους της συμφωνίας καταχώρησης ενός αντικειμένου σε μια συγκεκριμένη κατηγορία. Το προτεινόμενο μέτρο της μη τυχαίας συμφωνίας καταχώρησης στην κατηγορία *j* δίνεται από τον τύπο:

$$\hat{k}_j = \frac{\sum_{i=1}^{n} K_{ij}^2 - Knp_j\{1+(K-1)p_j\}}{Kn(K-1)p_j(1-p_j)}.$$

Σημειώνουμε ότι ο $\hat{k}_{mr}$ της σχέσης (4.1) είναι ο σταθμισμένος μέσος των $\hat{k}_j$ με αντίστοιχα βάρη τα $p_j(1-p_j)$, όταν τα αντικείμενα ταξινομούνται από βαθμολογητές διαφορετικού

πλήθους. Κάτω από τη μηδενική υπόθεση της μη ύπαρξης συμφωνίας πέρα από την τυχαία συμφωνία, μια περίπου ασυμπτωτική διακύμανση του $\hat{k}_j$ είναι η

$$\text{var}(\hat{k}_j) = \frac{\{1+2(n-1)p_j\}^2 + 2(n-1)p_j(1-p_j)}{nk(n-1)^2 p_j(1-p_j)}$$

Η Kraemer (1980) εισήγαγε την περίπτωση διαφορετικών αριθμών καταχωρίσεων για κάθε αντικείμενο. Χαλάρωσε επίσης τις προϋποθέσεις των αμοιβαία αποκλειόμενων κατηγοριών, επιτρέποντας ένα αντικείμενο να ταξινομηθεί σε περισσότερες από μία κατηγορίες από τον ίδιο βαθμολογητή. Για παράδειγμα, ένα αντικείμενο θα μπορούσε να ταξινομηθεί στην κατηγορία Α ή ισότιμα στην κατηγορία Β (A/B) ή στην κατηγορία Α αρχικά και στην κατηγορία Β στη συνέχεια (AB). Αν και τα δύο σχέδια κατηγοριοποιήσεων εμπεριέχουν και τις δύο κατηγορίες Α και Β, αντιμετωπίζονται διαφορετικά. Η επέκταση του $\hat{k}_{mr}$ στην περίπτωση αυτή – ας την ονομάσουμε $\hat{k}_0$- παράγεται αντιστοιχίζοντας μια παρατήρηση σαν rank ordering από τις m κατηγορίες. Στο παραπάνω παράδειγμα, μια βαθμολόγηση του τύπου A/B θα impose a rank of 1.5 στις κατηγορίες Α και Β και a rank of 1/2(*I*+3) στις άλλες *I-2* κατηγορίες. Χρησιμοποιώντας τον Spearman rank συντελεστή συσχέτισης ο $\hat{k}_0$ δίνεται από τον τύπο,

$$\hat{k}_0 = \frac{r_I - r_T}{1 - r_T},$$

όπου r_I είναι ο αστάθμητος μέσος των εντός των αντικειμένων συντελεστών συσχέτισης και r_T ο μέσος Spearman συντελεστής συσχέτισης τάξεων ανάμεσα σε όλα τα ζεύγη παρατηρήσεων του δείγματος. Ισοδύναμα, αυτό το kappa στατιστικό μπορεί να υπολογιστεί χρησιμοποιώντας μια ανάλυση διακύμανσης on the ranks.

4.2.1. Εφαρμογή

Τα δεδομένα που παρείχε ο Williams (1976), αν και για διαφορετικούς σκοπούς, μπορούν να χρησιμοποιηθούν σαν ένα παράδειγμα που αποδεικνύει την εκτίμηση του $\hat{k}_{mr}$.

Σαν μέρος του κλινικού – εργαστηριακού ποιοτικής εξέλιξης προγράμματος, το College of American Pathologists (CAP) διεξάγει ένα επαρκές πειραματικό πρόγραμμα για syphilis serology. Ο Πίνακας 4.5 αντιπροσωπεύει μια καταγραφή από 28 syphilis serology δείγματα που εξετάστηκαν ανεξάρτητα από τέσσερα κλινικά εργαστήρια χρησιμοποιώντας το FTA –

ABS τεστ. Οι καταγραφές αναφέρονται σε τρεις κατηγορίες: μη αντιδρώντα (nonreactive – NR),οριακά (borderline – BL) και αντιδρώντα (reactive – R).

Πίνακας 4.5

Καταχώρηση των 28 δειγμάτων από τέσσερα εργαστήρια

	Εργαστήριο			
Δείγμα	0	1	2	3
1	R	R	R	R
2	R	R	R	R
3	BL	NR	NR	NR
4	BL	NR	NR	NR
5	BL	NR	NR	NR
6	R	R	R	R
7	BL	NR	NR	NR
8	R	R	R	R
9	NR	NR	NR	NR
10	NR	NR	NR	NR
11	R	R	R	R
12	R	R	BL	BL
13	R	R	R	R
14	R	R	BL	BL
15	R	R	R	R
16	R	R	NR	BL
17	R	R	NR	BL
18	R	R	R	R
19	R	R	R	R
20	BL	BL	NR	NR
21	R	R	R	R
22	BL	NR	NR	NR
23	BL	BL	NR	NR
24	BL	BL	NR	NR
25	R	R	R	R
26	NR	NR	NR	NR
27	R	R	R	R
28	NR	NR	NR	NR

Για την εκτίμηση του k'_{mr} της σχέσης (4.2) πρώτα κατασκευάζουμε έναν πίνακα που μας δείχνει τον αριθμό των καταχωρήσεων σε κάθε κατηγορία (Πίνακας 4.6).

Ο Fleiss (1971) ερμήνευσε το p_0 της σχέσης (4.3) ως ακολούθως: ας θεωρήσουμε ένα αντικείμενο ότι επιλέγεται τυχαία και γίνεται διάγνωση πάνω σε αυτό από ένα τυχαία επιλεγμένο εργαστήριο.

Πίνακας 4.6

Οι τιμές των k_{ij} για τα τέσσερα εργαστήρια

	(j)			
(i)	NR	BL	R	$\sum_{j=1}^{I} k_{ij}^2$
Δείγμα	(1)	(2)	(3)	
1	-	-	4	16
2	-	-	4	16
3	3	1	-	10
4	3	1	-	10
5	3	1	-	10
6	-	-	4	16
7	3	1	-	10
8	-	-	4	16
9	4	-	-	16
10	4	-	-	16
11	-	-	4	16
12	-	2	2	8
13	-	-	4	16
14	-	2	2	8
15	-	-	4	16
16	1	1	2	6
17	1	1	2	6
18	-	-	4	16
19	-	-	4	16
20	2	2	-	8
21	-	-	4	16
22	3	1	-	10
23	2	2	-	8
24	2	2	-	8
25	-	-	4	16
26	4	-	-	16
27	-	-	4	16
28	4	-	-	16
Σύνολο	39	17	56	358

$$p_0 = \frac{1}{(28)(4)(3)}[358 - (28)(4)] = 0.732$$

Αν πάνω στο αντικείμενο αυτό γινόταν διάγνωση και από ένα δεύτερο επιλεγμένο εργαστήριο, η δεύτερη διάγνωση θα συμφωνούσε με την πρώτη πάνω από 73% τη φορά. Σύμφωνα με τις σχέσεις (4.4), (4.5), έχουμε,

$$p_1 = \frac{39}{(28)(4)} = 0.348,$$

$$p_2 = \frac{17}{(28)(4)} = 0.152$$

$$p_3 = \frac{56}{(28)(4)} = 0.5$$

$$p_e = \sum_{j=1}^{3} p_j^2 = (0.348)^2 + (0.152)^2 + (0.5)^2 = 0.394$$

Έτσι, από τη σχέση (4.2),

$$k_{mr} = \frac{0.732 - 0.394}{1 - 0.394} = 0.558.$$

4.3. Μοντέλα συμφωνίας

4.3.1. Λογαριθμογραμμικά μοντέλα

Οι Tanner και Young (1985 a) θεώρησαν λογαριθμογραμμικά μοντέλα για να εκφράσουν τη συμφωνία μεταξύ βαθμολογητών και διέκριναν τα συστατικά της, όπως τυχαία συμφωνία και μη τυχαία συμφωνία. Χρησιμοποιώντας, λοιπόν, τα λογαριθμογραμμική μοντελοποίηση, κανείς μπορεί να δείξει υποδείγματα συμφωνίας μεταξύ διαφόρων ερευνητών ή να συγκρίνει τα υποδείγματα συμφωνίας όταν τα αντικείμενα είναι στρωματοποιημένα με βάση τις τιμές κάποιας συμμεταβλητής. Υποθέτοντας ότι υπάρχουν n αντικείμενα τα οποία έχουν αξιολογηθεί από τους ίδιους *K* βαθμολογητές *(K>2)* σε *I* ονομαστικές κατηγορίες, οι Tanner και Young εξέφρασαν τη τυχαία συμφωνία ή την στατιστική ανεξαρτησία των αξιολογήσεων, χρησιμοποιώντας το ακόλουθο λογαριθμογραμμικό μοντέλο:

$$\log m_{ij\ldots l} = l + l_i^{R_1} + l_j^{R_2} + \ldots + l_l^{R_K}, \quad i, j, \ldots, l = 1, 2, \ldots, I$$

όπου $m_{ij\ldots l}$ είναι η αναμενόμενη συχνότητα του *ij…l* κελιού της από κοινού *K* – διάστατης ταξινόμησης των αξιολογήσεων, λ είναι η ολική επίδραση, $l_c^{R_h}$ είναι η επίδραση λόγω της κατηγοριοποίησης του *h* βαθμολογητή στην *c* κατηγορία *(h = 1, …,K, c = 1, …, I)* και

$\sum_{i=1}^{I} u_i^{R_1} = ... = \sum_{l=1}^{I} u_l^{R_K} = 0$. Μια χρήσιμη γενίκευση του μοντέλου της ανεξαρτησίας ενσωματώνει μη τυχαία συμφωνία με τον ακόλουθο τρόπο:

$$\log m_{ij...l} = l + l_i^{R_1} + l_j^{R_2} + ... + l_l^{R_K} + d_{ij...l}.$$

Ο πρόσθετος όρος $\delta_{ij...l}$ παρουσιάζει τη μη τυχαία συμφωνία για το *ij…l* κελί. Για να ελέγξουμε μια δοσμένη υπόθεση που αφορά το μοντέλο συμφωνίας, οι παράμετροι που αντιστοιχούν στο συστατικό της συμφωνίας $\delta_{ij...l}$ ανατίθενται σε συγκεκριμένα κελιά ή σύνολα κελιών στον πίνακα συνάφειας. Ο όρος $\delta_{ij...l}$ μπορεί να προσδιοριστεί ανάλογα με το τι είδους μοντέλο συμφωνίας εξετάζεται. Όπως είδαμε και στο Κεφάλαιο 3, για παράδειγμα, κατά την εξέταση της ομοιογενούς συμφωνίας μεταξύ $K=2$ βαθμολογητών, κάποιος μπορεί να ορίσει το δ_{ij} να είναι ίσο με δ όταν $i = j$, και ίσο με μηδέν για τις υπόλοιπες περιπτώσεις. Από την άλλη πλευρά, για να εξετάσουμε ένα είδος πιθανής μη ομοιογενούς συμφωνίας (π.χ. διαφορική συμφωνία ανά κατηγόρια απόκρισης), κάποιος θα μπορούσε να θέσει $\delta_{ij} = \delta_i I(i=j)$, $i,j = 1, \ldots, I$, όπου η δείκτρια συνάρτηση $I(i=j)$ ισούται με 1 όταν $i=j$ και με 0 σε κάθε άλλη περίπτωση. Γενικότερα, όταν εμπλέκονται $K>2$ βαθμολογητές, η μέθοδος αυτή απαιτεί υψηλότερης τάξης συμφωνία καθώς και κατά ζεύγη συμφωνία (Tanner και Young 1985a). Οι παράμετροι τότε περιγράφουν υποθετική συμφωνία: για παράδειγμα, η συμφωνία μεταξύ 2 βαθμολογητών για fixed βαθμολογήσεις από τους άλλους βαθμολογητές.

4.3.2. Μοντέλα λανθάνουσας τάξης

Διάφοροι ερευνητές έχουν προτείνει μοντέλα λανθάνουσας τάξης για να εξετάσουν τη συμφωνία μεταξύ βαθμολογητών (Aickin 1990, Uebersax και Grove 1990, Agresti 1992). Τα λανθάνουσας τάξης μοντέλα εκφράζουν την από κοινού κατανομή των βαθμολογήσεων σαν ένα «μείγμα» από κατανομές των τάξεων μιας ανεξερεύνητης (λανθάνουσας) μεταβλητής. Κάθε κατανομή του «μείγματος» αναφέρεται σε μια συστάδα αντικειμένων που αντιπροσωπεύουν μια ξεχωριστή τάξη μιας κατηγορικής λανθάνουσας μεταβλητής. Τα αντικείμενα αυτά είναι ομοιογενή κατά μια έννοια. Σύμφωνα με τον Agresti (1992), περιγράφουμε ένα βασικό λανθάνον ταξικό μοντέλο για την εξέταση συμφωνίας μεταξύ βαθμολογητών.

Ας υποθέσουμε ότι υπάρχουν τρεις βαθμολογητές, ο Α, ο Β και ο Γ, ο καθένας από τους οποίους ταξινομεί καθένα από τα n αντικείμενα σε m ονομαστικές κατηγορίες. Το λανθάνον ταξικό μοντέλο υποθέτει ότι υπάρχει μια μη παρατηρηθείσα κατηγορική διαβάθμιση *Y*, με *L*

κατηγορίες, έτσι ώστε τα αντικείμενα σε κάθε κατηγορία της *Y* να είναι ομοιογενή. Εξαιτίας αυτής της ομοιογένειας, με δοσμένο το επίπεδο της *Y*, οι από κοινού καταχωρήσεις των Α, Β, Γ βαθμολογητών υποτίθεται ότι είναι στατιστικά ανεξάρτητες. Αυτό χαρακτηρίζεται ως «τοπική ανεξαρτησία». Για ένα τυχαία επιλεγμένο αντικείμενο, ας θέσουμε ως π_{ijlk} την πιθανότητα ταξινόμησης στις κατηγορίες *(i, j, l)* από τους βαθμολογητές (Α, Β, Γ) και την κατηγοριοποίηση στην τάξη *c* της διαβάθμισης *Y*. Συνεχίζοντας, ας θέσουμε ως m_{ijlk} τις αναμενόμενες συχνότητες για την Α-Β-Γ-Υ ταξινόμηση. Με τον τρόπο αυτό, τα παρατηρούμενα δεδομένα αποτελούν έναν τριδιάστατο περιθώριο πίνακα ενός μη παρατηρημένου πίνακα τεσσάρων διαστάσεων. Το λανθάνον ταξικό μοντέλο που αντιστοιχεί στο λογαριθμογραμμικό μοντέλο (ΑΧ, ΒΧ, ΓΧ) είναι το μη γραμμικό μοντέλο που έχει τη μορφή:

$$\log m_{ijl+} = l + l_i^A + l_j^B + l_l^\Gamma + \log \sum_c \exp(l_c^Y + l_{ic}^{AY} + l_{jc}^{BY} + l_{lc}^{\Gamma Y}).$$

Με δοσμένη την λανθάνουσα τάξη, μπορεί η προσαρμογή του μοντέλου να χρησιμοποιηθεί για την εκτίμηση υποθετικών πιθανοτήτων καταχώρησης των αντικειμένων από τους βαθμολογητές. Μπορεί επίσης να εκτιμηθούν πιθανότητες των αντικειμένων να ταξινομηθούν σε διάφορες λανθάνουσες τάξεις, υποθετικά σε ένα συγκεκριμένο υπόδειγμα παρατηρούμενων καταχωρήσεων, και να χρησιμοποιηθούν για να γίνουν προβλέψεις σχετικά με τη λανθάνουσα τάξη στην οποία ανήκει ένα συγκεκριμένο αντικείμενο. Με την έννοια αυτή, τα λανθάνοντα ταξικά μοντέλα επικεντρώνονται λιγότερο στη συμφωνία μεταξύ των βαθμολογητών και περισσότερο στη συμφωνία κάθε βαθμολογητή με τη σωστή αξιολόγηση. Αυτή είναι χρήσιμη πληροφόρηση αν οι λανθάνουσες τάξεις αντιστοιχούν στις πραγματικές κατηγορίες ταξινόμησης. Αυτό, βέβαια, κανείς δεν ξέρει αν συμβαίνει. Γι' αυτό φαίνεται ότι ο συνδυασμός της λογαριθμογραμμικής και της μοντελοποίησης λανθάνουσας τάξης αποτελεί μια χρήσιμη στρατηγική για τη μελέτη της συμφωνίας μεταξύ βαθμολογητών.

4. 4 Ανάλυση της διαφωνίας μεταξύ πολλών βαθμολογητών

4.4.1. Εισαγωγή

Στην παράγραφο αυτή, σκοπός μας είναι να διατυπώσουμε μια αντίληψη της αμεροληψίας των καταχωρήσεων των βαθμολογητών που παρουσιάζουν διαφωνία. Η αντίληψη αυτή δίνει έμφαση στις κατηγορίες ταξινόμησης οι οποίες φαίνεται να συγχέονται από τους

βαθμολογητές, με την έννοια ότι μεταξύ αυτών των κατηγοριών εμφανίζεται μεγαλύτερη αναλογία διαφωνιών από ότι αναμένεται. Η αμεροληψία ορίζεται σε όρους συγκεκριμένων πιθανοτήτων. Επίσης αναπτύσσονται η απαραίτητη θεωρία ασυμπτωτικών κατανομών καθώς και οι διαδικασίες ελέγχου. Τέλος, εξετάζουμε ένα σετ δεδομένων που έχει αναλυθεί στη βιβλιογραφία και αφορά την ταξινόμηση αντικειμένων από ψυχιάτρους (Fleiss (1971), Landis and Koch (1977b), Kraemer (1980)).

Οι Fleiss et al. (1972) και ο Kraemer (1980) υπέθεσαν περιπτώσεις όπου κάθε βαθμολογητής μπορεί αν κάνει πολλές καταχωρήσεις για το ίδιο αντικείμενο. Τέτοιου είδους επεκτάσεις δεν μελετώνται στην παρούσα παράγραφο, καθώς επίσης ούτε και περιπτώσεις όπου διαφορετικοί αριθμοί βαθμολογητών αξιολογούν κάθε αντικείμενο (Landis and Koch, 1977c).

4.4.2 Αμεροληψία των περιπτώσεων διαφωνίας

Ας δούμε τον συμβολισμό που χρησιμοποιούν οι Landis and Koch (1977c): ας ορίσουμε το y_{ihj} να παίρνει την τιμή 1 αν το i αντικείμενο καταχωρείται στην κατηγορία j από τον h βαθμολογητή και να παίρνει την τιμή 0 σε κάθε άλλη περίπτωση, όπου $i = 1, \ldots, n, j = 1, \ldots, I$ και το h δείχνει το σύνολο των βαθμολογητών που λαμβάνουν μέρος στη διαδικασία ταξινόμησης, R από τους οποίους αξιολογούν κάθε αντικείμενο. Υποθέτουμε ότι δεν υπάρχουν σφάλματα μεταξύ των βαθμολογητών, συνεπώς οι πιθανότητες,

$$p_j = p(y_{ihj} = 1)$$

και

$$p_{uj} = p(y_{ihj} = 1, y_{ih'u} = 1), \quad h \neq h'$$

δεν εξαρτώνται από τα h και h'.

p_j είναι η περιθώρια πιθανότητα με την οποία κάθε αντικείμενο καταχωρείται από έναν βαθμολογητή στην κατηγορία j και p_{uj} είναι η από κοινού πιθανότητα διαφορετικών βαθμολογητών να καταχωρούν το ίδιο αντικείμενο i στις κατηγορίες u και j. Η υπόθεση μη ύπαρξης σφαλμάτων μεταξύ των βαθμολογητών είναι κατάλληλη αν, για παράδειγμα, οι R βαθμολογητές για κάθε αντικείμενο έχουν επιλεγεί τυχαία από ένα μεγαλύτερο σύνολο βαθμολογητών, αλλά μπορεί να μην είναι και η κατάλληλη αν οι ίδιοι R βαθμολογητές αξιολογούν κάθε αντικείμενο. Μέθοδοι για τον έλεγχο της ισχύος της παραπάνω υπόθεσης δίνονται από τους Landis and Koch (1977a, b).

Αν $p_{jj} = p_j^2$, λέμε ότι δεν υπάρχει άλλου είδους συμφωνία στην κατηγορία *j*, εκτός από εκείνη που οφείλεται στην τύχη.

Προκειμένου να μελετήσουμε υποδείγματα διαφωνίας για την κατηγορία *j*, ας θεωρήσουμε την υποθετική πιθανότητα ότι ένας δεύτερος βαθμολογητής καταχωρεί ένα αντικείμενο στην κατηγορία *u*, δεδομένου ότι ένας πρώτος βαθμολογητής έχει καταχωρήσει το ίδιο αντικείμενο στην κατηγορία *x*, και δεδομένου ότι οι βαθμολογητές διαφωνούν. Έτσι,

$$h_{u|j} = p(y_{ihu} = 1 \mid y_{ih'j} = 1, y_{ihj} = 0) = \frac{p_{uj}}{p_j - p_{jj}},$$

όπου υποθέτουμε ότι $p_j - p_{jj} > 0$. Επιπρόσθετα, ας θεωρήσουμε

$$q_{u|j} = p(y_{ihu} = 1 \mid y_{ihj} = 0) = \frac{p_u}{1 - p_j}$$

την υποθετική πιθανότητα ότι ένας βαθμολογητής καταχωρεί ένα αντικείμενο στην κατηγορία *u*, δεδομένου ότι η καταχώρηση δεν ήταν στην κατηγορία *j*. Σημειώνουμε ότι $\sum_{u \neq j} h_{u|j} \sum_{u \neq j} q_{u|j} = 1$. Λέμε ότι οι διαφωνίες με την κατηγορία *j* είναι αμερόληπτες αν $h_{u|j} = q_{u|j}$ για κάθε $u \neq j$. Για παράδειγμα, αν οι αναλογίες των καταχωρήσεων διαφωνίας που πηγαίνουν σε άλλες κατηγορίες κατοπτρίζουν τις σχετικές περιθώριες αναλογίες. Μια μικρή επανατοποθέτηση μας δείχνει ότι $h_{u|j} = q_{u|j}$, αν και μόνο αν

$$p(y_{ih'j} = 1 \mid y_{ihu} = 1) = p(y_{ih'j} = 1 \mid y_{ihj} = 0).$$

Μια εναλλακτική ερμηνεία είναι ότι οι διαφωνίες με την κατηγορία *j* είναι αμερόληπτες αν η πιθανότητα ένας βαθμολογητής να καταχωρεί το αντικείμενο στην κατηγορία *j*, δεδομένου ότι ένας άλλος βαθμολογητής έχει ήδη καταχωρήσει το ίδιο αντικείμενο σε μια διαφορετική κατηγορία, είναι το ίδιο ανεξάρτητη από την οποιαδήποτε κατηγορία έχει επιλεγεί.

Αν, αναφερόμενοι στην κατηγορία *j*, δεν υπάρχει αμεροληψία, τότε ενδιαφέρον παρουσιάζουν οι κατηγορίες *u* εκείνες για τις οποίες ισχύει $\eta_{u|j} > \theta_{u|j}$. Αυτές οι κατηγορίες είναι αποδέκτες περισσότερων καταχωρήσεων που παρουσιάζουν διαφωνία από ότι τυχαία θα αναμενόταν, γεγονός το οποίο δείχνει ότι οι βαθμολογητές έχουν ιδιαίτερη δυσκολία στο να αποφασίσουν μεταξύ της καταχώρησης στην κατηγορία *j* ή *u*. Αν η αναλογία των διαφωνιών είναι μεγάλη, θεμιτή είναι και η συγχώνευση των κατηγοριών, αν αυτή έχει έννοια, στην προσπάθεια απόκτησης περισσότερο αξιόπιστων ταξινομήσεων. Ένα ενδιαφέρον παράδειγμα αυτής της προσπάθειας χρησιμοποιήθηκε από την Kraemer (1979) και αφορούσε την συμπεριφορά χιμπατζήδων.

Σημειώνουμε ότι αν $\eta_{u|j}>\theta_{u|j}$, τότε απαραίτητα υπάρχει τουλάχιστον μία άλλη κατηγορία u', για την οποία ισχύει ότι $\eta_{u'|j}<\theta_{u'|j}$, εφόσον οι πιθανότητες αθροίζουν στο 1. Συνεπώς, πρέπει να είμαστε προσεκτικοί στην ερμηνεία τέτοιων αρνητικών συσχετίσεων. Επιπρόσθετα, είναι πιθανό να έχουμε $\eta_{u|j}>\theta_{u|j}$ ενώ $\eta_{j|u}<\theta_{j|u}$. Για να δούμε πως προκύπτει αυτό, ας υποθέσουμε ότι η κατηγορία *j* έχει διακριτικά χαρακτηριστικά, γεγονός το οποίο εξασφαλίζει ότι αν ένας βαθμολογητής καταχωρεί ένα αντικείμενο στην κατηγορία *j* και ένας δεύτερος βαθμολογητής διαφωνεί, τότε η δεύτερη καταχώρηση θα είναι σχεδόν σίγουρα στην κατηγορία *u*, δίνοντας έτσι στην κατηγορία *u* μια θετική συσχέτιση με την κατηγορία *j*. Ωστόσο, αν η συμφωνία στην κατηγορία *u* είναι μικρή, είναι πιθανόν ότι άλλα χαρακτηριστικά εξασφαλίζουν ότι οι περισσότερες διαφωνίες με την *u* καταχωρούνται σε μια τρίτη κατηγορία, έτσι η κατηγορία *j* προφανώς είναι αρνητικά συσχετισμένη με την κατηγορία *u*.

4.4.3 Αμεροληψία και μέτρα συμφωνίας

Ο Fleiss (1971) πρότεινε δείκτες για τη μέτρηση της έκτασης της συμφωνίας μεταξύ βαθμολογητών σε μια συγκεκριμένη κατηγορία και σε όλες τις κατηγορίες. Η θεωρητική σκοπιά αυτών των δεικτών, που τα είδαμε σαν *kappa* στατιστικά, είναι

$$k_j = \frac{p_{jj} - p_j^2}{p_j(1-p_j)}$$

για τη συμφωνία στην κατηγορία *j*, και

$$k = \frac{\sum_j p_j(1-p_j)k_j}{\sum_j p_j(1-p_j)}$$

$$= \frac{\sum_j (p_{jj} - p_j^2)}{1-\sum_j p_j^2}$$

για τη συνολική συμφωνία. Και τα δύο αυτά στατιστικά παίρνουν τη μέγιστη τιμή 1 όταν υπάρχει πλήρης συμφωνία μεταξύ των βαθμολογητών, ενώ παίρνει την τιμή 0 αν δεν υπάρχει συμφωνία, πέρα από την τυχαία συμφωνία. Τα *kappa* στατιστικά είναι πρακτικά ίδια με τις μεταξύ των τάξεων συσχετίσεις που περιγράφονται από τους Landis and Koch (1977c). Στην ενότητα αυτή μελετούμε κάποια από τα συμπεράσματα από την ερμηνεία της αμεροληψίας στα πλαίσια των *k* και k_j.

Σημειώνουμε ότι οι διαφωνίες με την κατηγορία j είναι αμερόληπτες αν $p_{uj}=p_u p_j(1-k_j)$ για κάθε $u \neq j$. Αν $k_j=0$, τότε η ερμηνεία είναι ισοδύναμη της ανεξαρτησίας κατά ζεύγη των καταχωρήσεων στην κατηγορία j και στην κατηγορία u από διαφορετικούς βαθμολογητές. Η υπόθεση $p_{uj} = p_u p_j$ χαρακτηρίστηκε σαν μη ύπαρξη συσχέτισης μεταξύ των κατηγοριών j και u από τους Holman et al (1982), ανεξάρτητα από το αν $k_j=0$. Ωστόσο, αν $k_j>0$, όπως συνήθως συμβαίνει, μη ύπαρξη συσχέτισης ουσιαστικά σημαίνει ότι οι περισσότερες διαφωνίες με την κατηγορία j (από ότι αναμέναμε να κατανέμονται σε όλες τις υπόλοιπες κατηγορίες ανάλογα με τις σχετικές περιθώριες αναλογίες) έχουν καταχωρηθεί στην κατηγορία u. Η ερμηνεία μας της αμεροληψίας μπορεί να θεωρηθεί σαν τροποποίηση της μη ύπαρξης συσχέτισης , η οποία δείχνει τη συμφωνία στην κατηγορία j πέρα από την τυχαία συμφωνία. Οι Landis and Koch (1977c) συνήγαγαν μια ερμηνεία για τη μη ύπαρξη συσχέτισης παρόμοια αυτής των Holman et al. (1982) αλλά βασισμένη σε εντός των τάξεων συσχετίσεις.

Αν οι διαφωνίες με τις κατηγορίες j και u είναι αμερόληπτες, τότε από τη σχέση $p_{uj}=p_{ju}$ συνεπάγεται ότι $k_j=k_u$. Η ανάγκη της «ισοδύναμης αξιοπιστίας» των δύο κατηγοριών εμφανίζεται εκπληκτική στην αρχή αλλά μειώνεται όταν σημειώνουμε ότι

$$1-k_j = \frac{p(y_{ih'j}=1 \mid y_{ihj}=0)}{p_j},$$

ο αριθμητής είναι η πιθανότητα διαφωνίας που είδαμε στην εναλλακτική ερμηνεία της αμεροληψίας στην παράγραφο 4.4.2. Είναι φανερό ότι οι περισσότερες υποθέσεις σχετικά με τον τρόπο που κατανέμονται οι διαφωνίες θα έχουν συμπεράσματα για τους δείκτες συμφωνίας.

Αν οι διαφωνίες σε όλες τις κατηγορίες είναι αμερόληπτες τότε θα έχουμε $k_1= k_2= \ldots= k_I = k$ όταν

$$p_{uj} = p_u p_j (1-k) \text{ για κάθε } u \neq j.$$

Ας υποθέσουμε τώρα ότι έχουμε αμερόληπτες διαφωνίες με όλες τις κατηγορίες και επιθυμούμε να συνδυάσουμε τις κατηγορίες j και u για να κατασκευάσουμε μια νέα κατηγορία, έστω w. Έχει αποδειχθεί ότι οι διαφωνίες μεταξύ όλων των καινούριων κατηγοριών είναι αμερόληπτες, έτσι $k_w= k$ και ο δείκτης ολικής συμφωνίας παραμένει ο ίδιος. Αν βέβαια δεν υπάρχει πλήρης αμεροληψία, τότε η συγχώνευση των δύο κατηγοριών μπορεί να διαφοροποιήσει το k, ακόμα και αν οι δύο κατηγορίες που συνδυάστηκαν έχουν τον ίδιο δείκτη *kappa* με εκείνον της κατηγορίας που παρήγαγαν.

Στην πράξη, βέβαια, η πλήρης αμεροληψία είναι σπάνιο φαινόμενο, αν και μπορεί περίπου να επιτευχθεί μετά από συγχωνεύσεις κατηγοριών. Από ένα σημείο και μετά οι συγχωνεύσεις κατηγοριών δεν θα οδηγούν σε σημαντικά υψηλότερους δείκτες συμφωνίας.

4.4.4 Εφαρμογή σε ψυχιατρικά δεδομένα

Το παράδειγμά μας αφορά την καταχώρηση 30 ασθενών στις κατηγορίες: (1) «Κατάθλιψη», (2) «Διαταραχή προσωπικότητας», (3) «Σχιζοφρένια», (4) «Νεύρωση» και (5) «άλλη περίπτωση». Τα δεδομένα δόθηκαν από τον Fleiss (1971) και αναλύθηκαν ξανά από τους Landis και Koch (1977c) και από την Kraemer (1980). Αν και οι ασθενείς είχαν αρχικά εξετασθεί από διάφορους αριθμούς ψυχιάτρων (βαθμολογητών), ο αριθμός των καταχωρήσεων για κάθε έναν ασθενή περιορίστηκαν στις έξι από τον Fleiss για λόγους διευκόλυνσης. Τα δεδομένα μπορούν να βρεθούν στον Fleiss (1971) ή στους Landis και Koch (1977c) και δεν θα παρουσιαστούν εδώ. Υπάρχει σημαντική συμφωνία – πέρα από την τυχαία συμφωνία – σε κάθε κατηγορία, αν και στην κατηγορία «Κατάθλιψη» και στην κατηγορία «Διαταραχή προσωπικότητας»η συμφωνία είναι σε χαμηλά επίπεδα. Για τις πέντε κατηγορίες που αναφέρθηκαν πιο πάνω, οι εκτιμημένες υποθετικές πιθανότητες $p(y_{ihj}=1|y_{ih'j}=1)$ είναι 0.35, 0.35, 0.60, 0.63 και 0.67 αντίστοιχα με εκτιμημένους δείκτες *kappa* ίσους με 0.24, 0.24, 0.52, 0.47 και 0.57 αντίστοιχα. Ο εκτιμημένος δείκτης *kappa* για την ολική συμφωνία είναι 0.43.

Οι εκτιμημένες υποθετικές πιθανότητες $h'_{u|j}$ και $q'_{u|j}$ για τα δεδομένα αυτά δίνονται στον Πίνακα 4.7, ενώ οι τιμές των X^2 τεστ με τρεις βαθμούς ελευθερίας για την αμεροληψία κάθε κατηγορίας δίνονται στον Πίνακα 4.8, μαζί με τις κατηγορίες για τις οποίες $h'_{u|j}$ υπερέχει της $q'_{u|j}$. Να υπενθυμίσουμε ότι διεξάγονται πολλαπλά τεστ στα ίδια δεδομένα, έτσι τα επίπεδα σημαντικότητας χρησιμοποιούνται υποκειμενικά.

Οι κατηγορίες 3 και 5 αξίζουν ιδιαίτερης μνείας. Σε καμία περίπτωση ασθενούς καμία κατηγορία δεν φαίνεται να δέχεται μεγάλο αριθμό διαφωνιών από αυτές τις δύο κατηγορίες, αλλά και στις δύο περιπτώσεις η κατηγορία 4 δέχεται πολύ λιγότερες από τις αναμενόμενες. Έτσι η διαφωνία με τη «Σχιζοφρένια» ή την κατηγορία «άλλη περίπτωση» είναι πολύ λιγότερο πιθανό να οφείλεται σε μια δεύτερη καταχώρηση στην κατηγορία «Νεύρωση» από ότι αναμενόταν, οι υπερβολικές διαφωνίες να κατανέμονται εξίσου σε όλες τις υπόλοιπες κατηγορίες.

Πίνακας 4.7

Εκτιμημένες υποθετικές πιθανότητες $\eta'_{u|j}$ και $q'_{u|j}$ για την ταξινόμηση των ψυχιατρικών δεδομένων

	Κατηγορία				
	1	2	3	4	5
$\eta'_{u\|1}$	-	0.07	0.25	0.46	0.21
$q'_{u\|1}$	-	0.17	0.19	0.36	0.28
$\eta'_{u\|2}$	0.07	-	0.15	0.56	0.21
$q'_{u\|1}$	0.17	-	0.19	0.36	0.28
$\eta'_{u\|3}$	0.35	0.22	-	0.05	0.38
$q'_{u\|3}$	0.17	0.17	-	0.37	0.29
$\eta'_{u\|4}$	0.39	0.47	0.03	-	0.12
$q'_{u\|4}$	0.21	0.21	0.24	-	0.34
$\eta'_{u\|5}$	0.25	0.25	0.32	0.17	-
$q'_{u\|5}$	0.19	0.19	0.22	0.4	-

Πίνακας 4.8

Τιμές X^2 τεστ

Κατηγορία x	Τιμή του τεστ αμεροληψίας (3 βαθμοί ελευθερίας)	Κατηγορίες για τις οποίες $\eta_{u\|j} > \theta_{u\|j}$ σημαντικά
1	3.4	-
2	6.1	4(P=0.09)
3	11.4	-
4	32.5	1(P=0.07), 2(P=0.01)
5	6.0	-

Αν κάποιος επιθυμούσε να συνδυάσει τις κατηγορίες προκειμένου να αυξήσει την αξιοπιστία, η συγχώνευση των κατηγοριών 2 και 4 θα φαινόταν ένα πρωταρχικό λογικό βήμα. Αυτή η συγχώνευση οδηγεί σε μια τιμή του X^2 στατιστικού ίση με 5.4 για το τεστ αμεροληψίας της συνδυασμένης κατηγορίας {2,4}, 0.5 για την κατηγορία 1, 3.8 για την κατηγορία 3, 1.6 για την κατηγορία 5, με δύο πάντα βαθμούς ελευθερίας. Ο εκτιμημένος

δείκτης ***kappa*** για το συνδυασμό {2,4} είναι 0.59 και το ολικό ***kappa*** αυξήθηκε στο 0.51. Μια περαιτέρω συγχώνευση των κατηγοριών 1 και {2,4} δίνει τιμή του X^2 τεστ ίση με 0.04 για το συνδυασμό κατηγοριών {1,2,4}, 0.41 για την κατηγορία 3 και 0.91 για την κατηγορία 5, με έναν πλέον βαθμό ελευθερίας, γεγονός που δείχνει σχεδόν αμεροληψία μεταξύ των τριών τελικά κατηγοριών. Ο εκτιμημένος ***kappa*** δείκτης για την κατηγορία {1,2,4}είναι 0.61, ενώ ο εκτιμημένος δείκτης ολικής συμφωνίας είναι 0.57. Σημειώνουμε ότι οι εκτιμημένοι δείκτες ***kappa*** είναι τώρα σχεδόν ίσοι, και μόνο μια ελάχιστη περαιτέρω αύξηση στην τιμή 0.61 μπορεί να επιτευχθεί με τη συγχώνευση των κατηγοριών 3 και 5. Η Kraemer (1980) επίσης συνέστησε τη συγχώνευση των κατηγοριών 1, 2 και 4 και έδειξε εμπειρικά ότι αυξήθηκε σημαντικά το μέτρο ολικής συμφωνίας. Η μεθοδολογία που αναπτύχθηκε εδώ ουσιαστικά τυποποιεί την εμπειρική του προσπάθεια.

ΚΕΦΑΛΑΙΟ 5

Σύγχρονες τάσεις στη μέτρηση και ανάλυση της συμφωνίας μεταξύ βαθμολογητών

Έως στο σημείο αυτό, είδαμε μέτρα και μοντέλα συμφωνίας μεταξύ βαθμολογητών τα οποία εισήχθηκαν από ερευνητές τα τελευταία 40 χρόνια, καθώς και θέματα γύρω από τη συμφωνία και τη διαφωνία μεταξύ βαθμολογητών με τα οποία έχουν κατά καιρούς ασχοληθεί οι ερευνητές. Ο Agresti (1999) ερευνά πρόσφατες επεκτάσεις αυτών των μοντέλων και τη σχετική μεθοδολογία για ειδικού τύπου εφαρμογές όπως επαναλαμβανόμενες μετρήσεις ή γενικότερα συσχετισμένων δεδομένων (clustered data). Ερευνά επίσης άλλες πτυχές μοντελοποίησης κατηγορικών δεδομένων, όπως ανάλυση μικρών δειγμάτων, την ισχύ και το μέγεθος του δείγματος καθώς και τη διαθεσιμότητα αντίστοιχου λογισμικού. Εντοπίζει επίσης και περιοχές οι οποίες προσφέρονται για έρευνα και μελλοντική μελέτη.

Η βιβλιογραφία για την συμφωνία μεταξύ βαθμολογητών είναι πολύπλοκη εξαιτίας του γεγονότος ότι έχει να κάνει με τουλάχιστον δύο διαφορετικά χαρακτηριστικά της αξιολόγησης δεδομένων: τον αριθμό των βαθμολογητών (δύο ή και περισσότεροι) και την κλίμακα αξιολόγησης (κατηγορική ή αριθμητική). Ενώ τα kappa στατιστικά χρησιμοποιούνται εκτενώς σε κατηγορικές κλίμακες, ο μεταξύ των τάξεων συντελεστής συσχέτισης προτιμάται στις αριθμητικές κλίμακες αξιολόγησης. Οι Shuster and Smith (2005) προτείνουν ένα πλαίσιο σταθμισμένου kappa διασποράς για πολλαπλούς βαθμολογητές ο οποίος ολοκληρώνει μερικά στατιστικά συμφωνίας χρησιμοποιώντας βάρη για να εκφράσει τη διαφωνία.

Υπάρχουν διάφοροι τύποι κλιμάκων αξιολόγησης στις οποίες οι αποκρίσεις μπορεί να είναι συνεχείς ή διακριτές. Σε πολλές μελέτες έχουν συγκριθεί η οπτική αναλογική κλίμακα (visual analogue scale VAS) με διάφορους τύπους διακριτών κλιμάκων, όπως για παράδειγμα οι προφορικές περιγραφικές κλίμακες και οι αριθμητικές κλίμακες αξιολόγησης. Αυτές οι μελέτες έχουν παρουσιάσει αντικρουόμενα αποτελέσματα, χωρίς σταθερές ενδείξεις για υπεροχή κάποιου από τα δύο είδη κλίμακας. Ωστόσο, τα συμπεράσματα έχουν συχνά βασιστεί σε παραμετρικές στατιστικές μεθόδους. Η Svensson (2000) χρησιμοποίησε μια μέθοδο αμετάβλητων τάξεων για συγκρίσεις μεταξύ των κλιμάκων, προκειμένου να

αξιολογήσει τη συνοχή των τάξεων μεταξύ της VAS, μιας γραφικής κλίμακας αξιολόγησης και μιας προφορικής περιγραφικής κλίμακας πέντε σημείων. Η συγκριτική αξιολόγηση έδειξε ότι και η προφορική κλίμακα και η γραφική ήταν ανώτερες της VAS και ότι η διακριτή κλίμακα είχε το υψηλότερο επίπεδο σταθερότητας.

Ο συντελεστής kappa του Fleiss για περισσότερους των δύο βαθμολογητών είναι γνωστό ότι επηρεάζεται από τα σφάλματα, γεγονός που μπορεί να οδηγήσει στο παράδοξο υψηλής συμφωνίας με χαμηλό kappa. Υποθέτει επίσης ότι οι βαθμολογητές είναι περιορισμένοι για το πώς θα κατανέμουν τις περιπτώσεις μεταξύ όλων των κατηγοριών, γεγονός το οποίο όμως δε συμβαίνει συνήθως στις μελέτες συμφωνίας. Ο Randolph (2005) εισάγει έναν συντελεστή ελεύθερου περιθωρίου, εναλλακτικό εκείνου του Fleiss για πολλούς βαθμολογητές. Ο εναλλακτικός αυτός συντελεστής δεν επηρεάζεται από τον kappa και είναι κατάλληλος για τυπικές μελέτες συμφωνίας, στις οποίες οι κατανομές των βαθμολογητών δεν είναι περιορισμένες.

Οι Janson and Olsson (2004) ασχολήθηκαν με το πρόβλημα της μέτρησης της ολικής πολυμεταβλητότητας διορθωμένης από τυχαιότητες συμφωνίας μεταξύ βαθμολογητών, όταν τα αντικείμενα αξιολογήθηκαν από διαφορετικές ομάδες βαθμολογητών (και όχι απαραίτητα του ίδιου αριθμού). Η προτεινόμενη από αυτούς μέθοδος στηρίζεται στη γενίκευση του kappa του Cohen αλλά ενσωματώνει κιόλας βάρη για τον αριθμό των βαθμολογητών ανά αντικείμενο και εφαρμόζει μια έκφραση για την αναμενόμενη διαφωνία ταιριαστή με την περίπτωση διαφορετικού αριθμού βαθμολογητών ανά αντικείμενο. Συστήνουν ότι η ελκυστικότητα της μεθόδου έγκειται στο γεγονός ότι μπορεί εύκολα να ερμηνεύσει την αναμενόμενη και την παρατηρούμενη διαφωνία σαν μέσες αποστάσεις μεταξύ των παρατηρήσεων και στο γεγονός ότι μελετώντας τη συμφωνία χωρίς αναφορές στη συνδιακύμνση των μεταβλητών έχει πλεονεκτήματα ως προς την απλότητα και ως προς την ερμηνεία.

Ο Schuster (2004) παρουσιάζει ένα τύπο για τον σταθμισμένο kappa σε όρους μέσων, διακυμάνσεων και συνδιακύμασης των βαθμολογητών που είναι ιδιαίτερα χρήσιμη, τονίζοντας ότι ο σταθμισμένος kappa είναι ένα απόλυτο μέτρο συμφωνίας με την έννοια ότι είναι ευαίσθητο στις διαφορές των περιθώριων κατανομών των βαθμολογητών. Πιο συγκεκριμένα, οι διαφορές στους μέσους των βαθμολογητών θα μειώσει την τιμή του σταθμισμένου kappa τη σχετική με την τιμή της μεταξύ των τάξεων συσχέτισης η οποία αγνοεί τις διαφορές στους μέσους. Αν επίσης διαφέρουν οι διακυμάνσεις των βαθμολογητών,

τότε η τιμή του σταθμισμένου kappa θα μειωθεί ανάλογα με την τιμή της product-moment συσχέτισης (ο product-moment συντελεστής συσχέτισης δίνεται από τον τύπο $r = \frac{s_{12}}{\sqrt{s_1^2 s_2^2}}$, όπου s_{12}, s_1^2, s_2^2 είναι η συνδιακύμανση και οι διακυμάνσεις των δύο βαθμολογητών. Δίνονται περιορισμοί στους μέσους και στις διακυμάνσεις των βαθμολογητών προκειμένου να περιγράψουν τις σχέσεις μεταξύ σταθμισμένου, μεταξύ των τάξεων συσχέτισης και της product-moment συσχέτισης. Επιπρόσθετα, η έκφραση για τον σταθμισμένο kappa του Schuster μας δείχνει ότι ανήκει στην οικογένεια των διορθωμένων ως προς την τυχαιότητα συντελεστών συσχέτισης Zegers-ten Berge.

Στις μελέτες συμφωνίας, όταν αξιολογούνται αντικείμενα ανεξάρτητα από δύο βαθμολογητές, προκύπτει μια συσχέτιση ανάμεσα στις αξιολογήσεις τους σε δύο κατηγορίες, η οποία αντανακλά τη δυσκολία στη διάκριση των δύο κατηγοριών από τους δύο βαθμολογητές. Όταν οι αξιολογήσεις αφορούν διατάξιμη μεταβλητή, αυτή η συσχέτιση μεταξύ των αξιολογήσεων σε δύο κατηγορίες αυξάνει όταν η απόσταση μεταξύ των δύο κατηγοριών αυξάνει. Τα λογαριθμογραμμικά μοντέλα του Goodman που προέκυψαν από την ανάλυση συμφωνίας μεταξύ δύο βαθμολογητών σε κατηγορική διάταξη, υποθέτουν ότι η μη διάκριση μεταξύ γειτονικών κατηγοριών είναι είτε σταθερή είτε εκ των προτέρων γνωστή. Λογαριθμικά μη-γραμμικά μοντέλα που επιτρέπουν μεταβολές της μη διακρισης μεταξύ γειτονικών κελιών σε όλο το μήκος της κλίμακας, μπορεί να οδηγήσουν σε δυσκολίες στην εκτίμηση παραμέτρων. Οι Valet, Guinot, Mary (2006) περιγράφουν μια καινούρια τάξη λογαριθμογραμμικών μοντέλων μη ομοιόμορφης συνάφειας. Τα μοντέλα αυτά επεκτείνουν το λογαριθμογραμμικό μοντέλο ομοιόμορφης συνάφειας επιτρέποντας μεταβολές στη μη διάκριση μεταξύ γειτονικών κατηγοριών σε όλη την κλίμακα.

Στις περιπτώσεις εκτίμησης της συμφωνίας μεταξύ ποικίλων βαθμολογητών για την αξιολόγηση διατάξιμων κατηγορικών δεδομένων, οι Williamson and Manatunga (1997) ερευνούν τη χρήση ενός λανθάνοντος μοντέλου που προτάθηκε από τους Qu, Piedmonte and Medendrop (1995) για να εκτιμήσουν τη συσχέτιση μεταξύ των βαθμολογητών για κάθε μέθοδο και να ελέγξουν την ισοδυναμία τους. Για την κάθε μέθοδο αξιολόγησης, αυτές οι συσχετίσεις μπορούν να ερμηνευθούν ως οι συνιστώσες διακύμανσης των τυχαίων επιδράσεων που αναπαριστούν αντικείμενο και βαθμολογητή.

Οι συντελεστές συμφωνίας ποσοτικοποιούν το κατά πόσο ένα σύνολο οργάνων, ερευνητών κλπ. συμφωνούν κατά την αξιολόγηση κάποιου χαρακτηριστικού ενός

πληθυσμού. Πολλοί συντελεστές συμφωνίας (όπως π.χ. ο kappa για ονοματικές, ο σταθμισμένος kappa για κατηγορικές και ο συντελεστής συσχέτισης συμφωνίας για συνεχείς αποκρίσεις) μπορούν να υποδείξουν αυξημένη συμφωνία ενώ οι περιθώριες κατανομές των δύο βαθμολογητών διαφοροποιούνται αν και το επίπεδο της διαφωνίας παραμένει το ίδιο ή αυξάνεται. Ο Fay (2005) περιγράφει το πρόβλημα αυτό και για το συντελεστή συσχέτισης – συμφωνίας. Προτείνει, επίσης, μια λύση για όλα τα είδη των αποκρίσεων στη μορφή των συντελεστών τυχαίας περιθώριας συμφωνίας, οι οποίοι χρησιμοποιούν μια διαφορετική αντιμετώπιση της τυχαιότητας από ότι οι προαναφερθέντες συντελεστές. Οι συντελεστές συσχέτισης – συμφωνίας έχουν το πλεονέκτημα οι διαφορές μεταξύ των περιθώριων κατανομών δεν καταλήγουν σε αυξημένη συμφωνία, αν και δεν ισχύει. Όπως και οι συντελεστές συμφωνίας kappa, και οι συντελεστές συσχέτισης – συμφωνίας δεν απαιτούν υποθέσεις διδιάστατων κατανομών των τυχαίων μεταβλητών που σχετίζονται με τους δύο βαθμολογητές. Έτσι περιγράφονται οι συντελεστές συμφωνίας – συσχέτισης για ονοματικά, κατηγορικά και συνεχή δεδομένα.

Οι κλινικές μελέτες συχνά δείχνουν ενδιαφέρον στο αν διαφορετικοί βαθμολογητές παράγουν παρόμοιες τιμές κατά τη μέτρηση ποσοτικών μεταβλητών. Η χρήση του συντελεστή συσχέτισης – συμφωνίας σαν ένα μέτρο αναπαραγωγιμότητας, κερδίζει έδαφος από την πρώτη εισαγωγή του από τον Lin (1989). Η μέθοδός του μπορεί να χρησιμοποιηθεί σε μελέτες αξιολόγησης δύο βαθμολογητών χωρίς επανάληψη. Chinchilli et al. (1996) επέκτειναν την μέθοδο του Lin σε σχεδιασμούς επαναλαμβανόμενων μετρήσεων χρησιμοποιώντας έναν σταθμισμένο συντελεστή συσχέτισης – συμφωνίας μέσω τριών σετ εξισώσεων εκτίμησης. Η μέθοδος που προτείνουν οι Barnhart and Williamson (2001) είναι ελαστική στο ότι μπορεί να χειριστεί περισσότερες από δύο ενδείξεις και να ελέγξει την ισότητα των εξαρτημένων εκτιμήσεων συσχέτισης – συμφωνίας. Είναι επίσης ελαστική γιατί μπορεί να ενσωματώνει συμμεταβλητές που προβλέπουν τις περιθώριες κατανομές και γιατί μπορεί να χρησιμοποιηθεί για να αναγνωρίσει συμμεταβλητές που προβλέπουν τη συσχέτιση συμφωνία. Τέλος είναι ελαστική γιατί απαιτεί ελάχιστες υποθέσεις κατανομών.

Η Goodwin (2001) ασχολήθηκε με τη διάκριση ανάμεσα στη «συμφωνία» και την «αξιοπιστία» μεταξύ βαθμολογητών. Παρουσιάζει τρεις μεθόδους ή τεχνικές για την εκτίμηση της συμφωνίας και της αξιοπιστίας μεταξύ βαθμολογητών, τη μέθοδο απλών ποσοστών συμφωνίας και του kappa, απλές τεχνικές συσχέτισης και τέλος γενικευμένης θεωρίας τεχνικές. Μελετώντας τα σχετικά πλεονεκτήματα και μειονεκτήματα των διαφόρων

μεθόδων, δίνεται έμφαση στις γενικευμένης θεωρίας τεχνικές γιατί είναι πιο αντιληπτές και ελαστικές, ώστε να επιτρέπουν στον ερευνητή να απομονώνει πολλές πηγές σφαλμάτων.

Το πρόβλημα του ορισμού και της εκτίμησης της εντός των βαθμολογητών, μεταξύ των βαθμολογητών και της αξιοπιστίας των επαναλαμβανόμενων μετρήσεων αναφέρεται από τους Rousson, Gasser and Seifert (2002). Συστήνουν ότι η συνήθης ερμηνεία της product-moment συσχέτισης εφαρμόζεται ικανοποιητικά στις περιπτώσεις επαναλαμβανόμενων μετρήσεων, ενώ η εντός των τάξεων συσχέτιση θα πρέπει να χρησιμοποιείται στην εξέταση της εντός και μεταξύ των βαθμολογητών αξιοπιστίας. Η σημαντική διαφορά ανάμεσα σε αυτές τις δύο μεθόδους είναι η αντιμετώπιση του συστηματικού σφάλματος. Επιπλέον συγκρίνουν αυτές τις μεθόδους αξιοπιστίας εισάγοντας την έννοια των ορίων συμφωνίας που έχουν προταθεί για την αξιολόγηση της συμφωνίας μεταξύ δύο μεθόδων κλινικών μετρήσεων από τους Bland and Altman (1986) και αποδεικνύουν πως η product-moment συσχέτιση συνδέεται με αυτά τα όρια. Στη συνέχεια προτείνουν νέα όρια συμφωνίας τα οποία συνδέονται με την εντός των τάξεων συσχέτιση.

Ένα άλλο θέμα με το οποίο πρόσφατα έχουν ασχοληθεί οι ερευνητές είναι η κατασκευή διαστημάτων εμπιστοσύνης για τα kappa στατιστικά. Το στατιστικό kappa του Cohen είναι το πιο γνωστό μέτρο της συμφωνίας μεταξύ δύο βαθμολογητών σε δεδομένα της μορφής 0 – 1. Έχουν αναφερθεί διάφορες εκφράσεις για την ασυμπτωτική του διακύμανση και η κανονική προσέγγιση της κατανομής του χρησιμοποιείται για την κατασκευή διαστημάτων εμπιστοσύνης. Ωστόσο, η πληροφόρηση για την ακρίβεια αυτών των διαστημάτων δεν είναι κατανοητή. Κάτω από το γνωστό μοντέλο συσχέτισης για δεδομένα της μορφής 0 – 1, οι Blackman and Koval (2000) αξιολογούν ένα 95% διάστημα εμπιστοσύνης χρησιμοποιώντας τέσσερις εκφράσεις ασυμπτωτικής διακύμανσης. Χρησιμοποίησαν ακριβείς υπολογισμούς παρά μεθόδους προσομοίωσης και προσδιόρισαν συγκεκριμένες υποθέσεις κάτω από τις οποίες η χρήση της ασυμπτωτικής διακύμανσης έχει νόημα.

Αν και οι διαδικασίες οι βασισμένες σε μοντέλα για το kappa στατιστικό έχουν αναπτυχθεί με ταχείς ρυθμούς την τελευταία δεκαετία, δεν έχει αναπτυχθεί μέθοδος για την κατασκευή διαστημάτων εμπιστοσύνης για τη διαφορά μεταξύ δύο ανεξάρτητων kappa στατιστικών, εργαλείο χρήσιμο σε μικρού έως μεσαίου μεγέθους δείγματα. Οι Donner and Zou (2002) προτείνουν και αξιολογούν δύο τέτοιες μεθόδους, βασισμένοι σε μια ιδέα εισηγμένη από τον Newcombe (1998) για την κατασκευή ενός διαστήματος εμπιστοσύνης μεταξύ ανεξάρτητων αναλογιών. Οι παραπάνω μέθοδοι φαίνεται να παρέχουν ικανοποιητικά αποτελέσματα σε

δείγματα μεγέθους τάξης 25 αντικειμένων ανά ομάδα. Παρουσιάζουν επίσης τις απαιτήσεις του μεγέθους του δείγματος με τις οποίες επιτυγχάνεται ένα προκαθορισμένο αναμενόμενο εύρος για το διάστημα εμπιστοσύνης για τη διαφορά των kappa στατιστικών.

Ας θεωρήσουμε το πρόβλημα κατασκευής διαστημάτων εμπιστοσύνης για τον μεταξύ των τάξεων συντελεστή συσχέτισης σε μια μελέτη συμφωνίας μεταξύ δύο βαθμολογητών όταν τόσο οι βαθμολογητές όσο και τα αντικείμενα υπό αξιολόγηση είναι τυχαίες επιδράσεις σε ένα μοντέλο two-way random effects. Οι Cappelleri and Ting (2003) διεξάγουν μια μελέτη προσομοίωσης για να ερευνήσουν και να συγκρίνουν τα διαστήματα εμπιστοσύνης που δίνονται από τρεις διαφορετικές μεθόδους, εκ των οποίων η τροποποιημένη μεγάλου δείγματος μέθοδος παρέχει την πιο ικανοποιητική προσέγγιση των διαστημάτων.

Όσο για το μέγεθος του δείγματος , ο Donner (1998) παρέχει τύπους και πίνακες για το σχεδιασμό των μελετών στις οποίες γίνεται σύγκριση δύο ή περισσότερων συντελεστών συμφωνίας μεταξύ βαθμολογητών. Τέτοιες μελέτες μπορούν π.χ. να προκύψουν όταν το ενδιαφέρον επικεντρώνεται στο πως τα μέτρα συμφωνίας μεταξύ βαθμολογητών ποικίλουν στο διάφορα υποσύνολα ασθενών.

Η αναλυτική παρουσίαση των παραπάνω θεμάτων είναι εκτός των στόχων της συγκεκριμένης εργασίας για αυτό περιοριστήκαμε σε μια σύντομη αναφορά τους.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Ανδρουλάκης, Ν. (1986). *Ποινικόν Δίκαιον γεν. μέρος ΙΙ*, Αντ. Ν. Σάκκουλα, Αθήνα – Κομοτηνή.

Γενά, Αγγελική. (2002). *Αυτισμός και Διάχυτες αναπτυξιακές διαταραχές*, Εκδόσεις: ιδιωτική, Αθήνα.

Γεωργάς, Παρασκευόπουλος, Μπεζεβέγκης, Γιαννιτσάς. (1998). *Ελληνικό WISC-III: Οδηγός εξεταστή*, Ελληνικά Γράμματα.

Κατέρη, Μαρία. (2005). *Σημειώσεις «Ανάλυση Διακριτών Δεδομένων»*.

Κυριόπουλος Γιάννης, Γείτονα Μαίρη, Σκουρολιάκου Μαρία. (1996). *Φαρμακοοικονομία Αρχές και Μέθοδοι Αξιολόγησης*, Εξάντας Εκδόσεις ΑΕ, σελ 24 - 33,36-46,79,88.

Μόττη –Στεφανίδη, Φρόσω. (1999). *Αξιολόγηση της νοημοσύνης Παιδιών Σχολικής Ηλικίας και εφήβων*, Ελληνικά Γράμματα.

Παρασκευόπουλος, Ι.Ν. (1991). *Στατιστική εφαρμοσμένη στις επιστήμες της συμπεριφοράς. Τόμος Α': Περιγραφική Στατιστική*, Ελληνικά Γράμματα, σελ.178.

Ξένη

Adams J. (2001). *Evolution of a classification scale: Medical evaluation of suspected child sexual abuse*, Child Maltreatment, **6**, 31-36.

Agresti, A. and Kezouh, A. (1983). Association models for multi-dimensional cross-classifications of ordinal variables, *Commun. Statist. – Theor. Meth.*, **12**, 1261-1276.

Agresti, A. (1988). A model for agreement between ratings an an ordinal scale, *Biometrics*, **44**, 539-548.

Agresti A. (1999). Modelling ordered categorical data: recent advances and future challenges, *Statistics in medicine*, **18**, 2191-2207.

Agresti, A., and Lang, J. B. (1993). Quasi-Symmetric Latent Class Models, with Application to Rater Agreement, *Biometrics*, **49**, 131-139.

Agresti, A., and Yang, M. (1986). An empirical investigation of some effects of sparseness in contingency tables, *Computational Statistics and Data Analysis*, **5**, 9-21.

Agresti, A. (1992). Modeling patterns of agreement and disagreement, *Statist. Methods Med. Res.*, **1**, 201-218.

Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive model, and its relation to Cohen's kappa, *Biometrics*, **46**, 293-302.

Barlow, W., Lai, M.Y., and Azen, S.P. (1991). A comparison of methods for calculating a stratified kappa, *Statist. Med.,* **10**, 1465-1472.

Barnhart H., Williamson J. (2001). Modeling concordance correlation via GEE to evaluate reproducibility, *Biometrics*, **57**, 931-940.

Becker, M. P. (1989). Using association models to analyze agreement data: two examples, *Statistics in Medicine*, **8**, 1199-1207.

Becker, M. P. (1990a). Algorithm AS 253. Maximum likelihood estimation of the RC(M) association model, *Applied Statistics,* **39**, 152-167.

Becker, M. P. (1990b). Quasi-symmetric models for the analysis of square contingency tables, *Journal of the Royal Statistical Society* **52**(2), 369-378.

Becker, M. and Clogg, C.C. (1989). Analysis of sets of two-way contingency tables using association models, *Journal of the American Statistical Association*, **84**, 142-151.

Becker, M. P. Chambers, J. M. and Wilks, A. R. (1988). The *new S language: a programming environment for data analysis and graphics*, CA: Wadsworth 7Brooks Cole, Pacific Grove.

Birch, M. W. (1963). Maximum Likelihood in Three-way Contigency Tables, *Journal of the Royal Statistical Society*, **26**, 313-324.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). Discrete *Multivariate Analysis: Theory and Practice*, MA: MIT Press, Cambridge,.

Blackman N., Koval J. (2000). Interval estimation for Cohen's kappa as a measure of agreement, *Statistics in medicine*, **19**, 723-741.

Bland JM, Altman DG. (1986). *Statistical methods for assessing agreement between two methods of clinical measurement*, Lancet, **1**(8476), 307-10.

Bloch, D.A., and Kraemer, H.C. (1989). 2x2 kappa coefficients: Measures of agreement or association, *Biometrics,* **45**, 269-287.

Brennan, R. L. and Prediger, J. D. (1981). Coefficient kappa: some uses, misuses and alternatives, *Educational and Psychological Measurement,* **41**, 687-699.

Byrt, T., Bishop, J., and Carlin, J.B. (1993). Bias, prevalence and kappa, *J.Clin. Epidemiol.*, **46**, 423-429.

Cappeleri J, Ting N. (2003) A modified large-sample approach to approximate interval estimation for a particular intraclass correlation coefficient, *Statistics in Medicine*, **22**, 1861-1877.

Caussinus, H. (1965). Contribution a l' anlyse statistique des tableaux de correlation, *Annals de la Faculte des Sciences de l' Universite de Toulouse,* **29**, 77-182.

Chinchilli et al. (1996). A weighted concordance correlation coefficient for repeated measurement designs, *Biometrics*, **52**, 341-353.

Cicchetti, D.V., and Fleiss, J.L. (1977). Comparison of the null distributions of weighted kappa and the C ordinal statistic, *Appl. Psych. Meas.,* **1**, 195-201.

Cohen, J. (1960). A coefficient of agreement for nominal scales, *Edu. and Psych. Meas.*, **20**, 37-46.

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

Darroch, J. N. and McCloud, P. I. (1986). Category distyinguishability and observer agreement, *Australian Journal of Statistics,* **28**(3), 371-388.

Davies A. R., Ware J. E. (1981). *Measuring Health Perceptions in the Health Insurance Experiment*, Publication R 1987/2, DHEN, Rand Corporation, Santa Monica.

Davies, M., and Fleiss, J.L. (1982). Measuring agreement for multinomial data, *Biometrics,* **38**, 1047-1051.

Donner A. (1998). Sample size requirements for the comparison of two or more coefficients of inter-observer agreement, *Statistics in Medicine*, **17**, 1157-1168.

Donner, A., and Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic:Confidence interval construction, significance-testing and sample size estimation, *Statist. Med.,* **11**, 1511-1519.

Donner, A., and Eliasziw, M. (1997). A hierarchical approach to inferences concerning intrerobserver agreement for multinomial data, *Statist. Med.,* **16**, 1097-1106.

Donner,A., Eliasziw, M., and Klar, N. (1996). Testinghomogeneity of kappa statistics, *Biometrics,* **52**, 176-183.

Donner A., Zou G. (2002). Interval estimation for a difference between intraclass kappa statistics, *Biometrics*, **58**, 209.

Espeland, M. A., and Handleman, S. L. (1989). Using Lattent Class Model to Characterize and Assess Relative Error in Discrete Measurements, *Biometrics*, **45**, 587-599.

Fay M. (2005). Random marginal agreement coefficients: rethinking the adjustment for chance when measuring agreement, *Biostatistics,* **6**, 171-180.

Everitt, B.S. (1968). Moments of the statistics kappa and weighted kappa, *British J. Math. Statist. Psych.,* **21**, 97-103.

Feinstein, A.R., and Cicchetti, D.V. (1990). High agreement but low kappa: I The problems of two paradoxes, *J. Clin. Epidemiol.,* **43**(6), 543-548.

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

Fleiss, J. L., Spitzer, R. L., Endicott, J. and Cohen, J. (1972). Quantification of agreement in multiple psychiatric diagnosis, *Archives of General Psychiatry,* **26**, 168-171.

Fleiss, J.L., and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, *Educ. and Psych. Meas.,* **33**, 613-619.

Fleiss, J. L., Cohen, J., and Everitt, B. S. (1969). Large Sample Standard Errors of Kappa and Weighted Kappa, *Psychological Bulletin,* **72**, 323-324.

Fleiss, J.L., and Cicchetti, D.V. (1978). Inference about weighted kappa in the non-null case. *Appl. Psych. Meas.,* **2**, 113-117.

Fleiss, J. and Cuzick, J. (1979). The reliability of dichotomous judgment: unequal number of judgments per subject, *Applied Psychological Measurement,* **3**, 537-542.

Fleiss, J.L., and Davies, M. (1982). Jackknifing functions of multinomial frequencies, with an application to a measure of concordance, *Amer. J. Epidemiol.,* **115**, 841-845.

Gelfand, A.E. and Solomon, H. (1975). Analyzing the decision making process of the American Jury, *Journal of the American Statistical Association,* **70**, 305-310.

Graham, P. (1995). Modelling covariate effects in observer agreement studies : The case of nominal scale agreement, *Statist. Med.,* **14**, 299-310.

Goodman, L. A. (1972). Some multiplicative models for the analysis of cross-classified data, In L. LeCam et al. (Ed.). *Proceedings of the 6th Berkley Symposium,* **1**, 649-696. Berkley: University of California Press.

Goodman, L. A. (1978). *Analysing Qualitative/Categorical Data*, MA: Abt Books, Cambridge.

Goodman, L. A. (1979a). Multiplicative models for the analysis of occupational mobility tables and other Kinds of cross-classification tables, *American Journal of Sociology,* **84**, 804-819.

Goodman, L. A. (1979b). Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories, *Journal of the American Statistical Association,* **74**(367), 537-552.

Goodman, L. A. (1981). Criteria for determining whether certain categories in a cross-classification table should be combined, with special reference to occupational categories in an occupational mobility table, *American Journal of Sociology,* **87**, 612-650.

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models and asymmetry models for contingency tables with or without missing entries, *Ann. Statist.*, **13**, 10-69.

Goodman, L. A. (1986). Some useful extensions of the usual log-linear models approach in the analysis of contingency tables (with discussion), *International Statistical Review,* **54**, 243-309.

Goodman, L. A., and Kruskal, W. H. (1954). Measures of Association for Cross Classifications, *Journal of the Americaj Statistical Association,* **49**, 732-764.

Goodwin L. (2001). Interrater agreement and reliability, *Measurement in physical education and exercise science*, **5**, 13-34.

Graham, P. and Jackson, R. (1993). The analysis of ordinal agreement data: beyond weighted kappa, *Journal of Clinical Epidemology* **46**(9), 1055-1062.

Haberman, S. J. (1974). *The analysis of frequency Data*, IL: University of Chicago Press, Chicago.

Haberman, S. J. (1977). Loglinear models and frequency tables with small expected counts, *The Annals of Statistics,* **5**, 148-169.

Haberman, S. J. (1978). *Analysis of quantitive data:vol.1 – Introductory Topics,* Academic Press, New York.

Hamdan, M. (1970). The equivalence of tetrchoric and maximum likelihood estimates of ρ in 2x2 tables, *Biometrika*, **57**, 212-215.

Holman, C. D. J., James, I. R., Heenan, P. J., Matz, L. R., Blackwell, J. B., Kelsall, g. r. h., Singh, A. and Ten Seldam, R. E. J. (1982). An improved method of analysis of observer variation between pathologists, *Histopathology,* **6**, 581-590.

Hunt S.M. et al. (1981). *The Nottingham health profile: subjective health status and medical consultations*, Social Science and Medicine,**15** A, 221-229.

Hutchinson, T.P. (1993). Kappa muddles together two sources of disagreement: tetrachoric correlation is preferable, *Res. Nursing and Health,* **16**, 313-315.

Jackson, R., Scragg, R. and Beaglehole, R. (1991). Alcohol consumption and risk of coronary heart disease, *British Medical Journal,* **303**, 211-216.

Janson H., 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.

Johnson, N.L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions.* Wiley, New York, 117-122.

Koehler, K. J. and Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics in sparse multinomials, *Journal of the American Statistical Association*, **75**, 336-344.

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

Kraemer, H. C. (1979). A study of reliability and its hierarchical structure in observed chimpanzee behaviour, *Primates,* **20**, 553-561.

Kraemer, H. C. (1980). Extension of the kappa coefficient, *Biometrics,* **36**, 207-216.

Kraemer, H. C. (1983). Kappa coefficient. In Johnson, N. L. and Kotz, S., *Encyclopedia of Statistical Sciences, 352-354*, New York: John Wiley & Sons.

Landis, R.J., and Koch, G.G. (1975a). A review of statistical methods in the analysis of data arising from observer reliability studies (Part I), *Statistica Neerlandica,* **29**, 101-123.

Landis, R.J., and Koch, G.G. (1975b). A review of statistical methods in the analysis of data arising from observer reliability studies (Part II), *Statistica Neerlandica,* **29**, 151-161.

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

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

Landis, R.J., and Koch, G.G. (1977c). A one-way components of variance model for categorical data, *Biometrics,* **33**, 671-679.

Lin, (1989). Concordance correlation coefficient to evaluate reproducibility, *Biometrics,* **45**, 255-268.

Mak, T.K. (1988). Analysing intraclass correlation for dichotomous variables, *Applied Statistics,* **20**, 37-46.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika,* **12**, 153-157.

Newcombe R.. (1998). Interval estimation for the difference between independent proportions: coparison of eleven methods, *Statistics in Medicine*, **17**, 873-890.

Oden, N.L. (1991). Estimating kappa from binocular data, *Statist. Med.*, **10**, 1303-1311.

Qu, Piedmonte, Medendorp. (1995). Latent variable models for clusteral ordinal data, *Biometrics*, **51**, 268-275.

Randolph J. (2005). Free-marginal multirater kappa (multirater K[free]):an alternative to Fleiss' fixed-marginal multirater kappa, Joensuu Learning and Instruction Symposium, 2005

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed., Wiley, New York.

Rousson V., Gasser T., Seifert B. (2002). Assessing intrarater, interrater and test-retest reliability of continuous measurements, *Statistics in Medicine*, **21**, 3431-3446.

Schouten, H.J.A. (1993). Estimating kappa from binocular data and comparing marginal probabilities, *Statist. Med.*, **12**, 2207-2217.

Schuster C. (2004). A note on the interpretation of weighted kappa and its relation to other rater agreement statistics for metric scales, *Educational and Psychological Measurement*, **64**, 243-253.

Schuster C., Smith D. (2005). Dispersion-weighted kappa: an integrative framework for metric and nominal scale agreement coefficients, *Biometrika*, **70**, 135-146.

Scott, W.A. (1955). Reliability of content analysis: The case of nominal scale coding, *Public Opinion Quart.*, **19**, 321-325.

Shoukri, M.M., Martin, S.W., and Mian, I.U.H. (1995). Maximum likelihood estimation of the kappa coefficient from models of matched binary responses, *Statist. Med.*, **14**, 83-99.

Shoukri, M.M. and Pause, C.A. (1999). Statistical methods for health Sciences, 2nd edition, CRC Press, Boca Raton, Florida.

Simon, G. (1974). Alternative analysis for the singly-ordered contingency table, *Journal of the American Statistical Association*,**69**, 971-976.

Spitznagel, E. L. and Helzer, J. E. (1985). A proposed solution to the base rate problem in the Kappa statistic, *Archives of General Psychiatry*, **42**, 25-28.

Svensson E. (2000). Comparison of the quality of assessments using continuous and discrete ordinal ratingscales, *Biometrical Journal*, **42**, 417-434.

Tallis, G.M. (1962). The maximum likelihood estimation of correlations from contingency tables, *Biometrics*, **18**, 342-353.

Tanner, M. A. and Young, M. A. (1985a). Modeling agreement among raters, *Journal of the American Statistical Association*, **80**(389), 175-180.

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

Uebersax, J. S. (1988). Validity inferences from interobserver agreement, *Psychological Bulletin*, **104**, 405-416.

Uebersax, J.s., and Grove, W.M. (1989). *Latent Structure Agreement Analysis*, CA: The RAND Corporation, Note N-3029-RC,Santa Monica.

Uebersax, J.s., and Grove, W.M. (1990). Latent class analysis of diagnostic agreement, *Statist. Med.*, **9**, 559-572.

Uebersax, J.s., and Grove, W.M. (1993). Alatent Trait Finite Mixture Model for the Analysis of Rating Agreement, *Biometrics*, **49**, 823-835.

Valet F., Guinot C., Mary J. (Mar. 2006). Log-linear non-uniform association models for agreement between two ratings on an ordinal scale, *Statistics in Medicine*.

Walter, S. D., and Irwig, L. M. (1988). Estimation of Test Error Rates, Disease Prevalence, and Relative Risk from Misclassified Data: A Review, *Journal of Clinical Epidemiology*, **41**, 923-937.

Williams, G.W. (1976). Comparing the joint agreement of several raters with another rater, *Biometrics*, **32**, 619-627.

Williamson J., Manatunga A. (1997). Assessing interrater agreement from dependent data, *Biometrics*, **53**, 707-714.

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