ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΜΕΤΡΑ ΣΥΝΑΦΕΙΑΣ ΚΑΙ ΜΕΤΡΑ ΣΥΜΜΕΤΡΙΑΣ - ΑΣΥΜΜΕΤΡΙΑΣ ΓΙΑ ΠΙΝΑΚΕΣ ΣΥΝΑΦΕΙΑΣ

Γεώργιος Ι. Κουτσοχέρας

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Σεπτέμβριος 2010

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΜΕΤΡΑ ΣΥΝΑΦΕΙΑΣ ΚΑΙ ΜΕΤΡΑ ΣΥΜΜΕΤΡΙΑΣ - ΑΣΥΜΜΕΤΡΙΑΣ ΓΙΑ ΠΙΝΑΚΕΣ ΣΥΝΑΦΕΙΑΣ

Γεώργιος Ι. Κουτσοχέρας

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Σεπτέμβριος 2010

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

Τα μέλη της Επιτροπής ήταν:

- Παπαιωάννου Τάκης (Επιβλέπων)
- Κατέρη Μαρία
- Τζαβελάς Γεώργιος

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

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

MEASURES OF ASSOCIATION AND MEASURES OF SYMMETRY – ASYMMETRY FOR CONTINGENCY TABLES

By

George I. Koutsocheras

MSc Dissertation

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

Piraeus, Greece
September 2010

…Πληροφορία είναι η σημασία που
δίνει ο άνθρωπος στα δεδομένα…

Claude Shannon, (1948)

Ευχαριστίες

Ευχαριστώ θερμά τον επιβλέποντα καθηγητή μου κ. Παπαιωάννου Τ., για την υπομονή και την αμέριστη υποστήριξη και βοήθεια που μου προσέφερε κατά την διάρκεια της συγγραφής της παρούσας εργασίας. Με την ενθάρρυνση και καθοδήγησή του κατάφερα να εμβαθύνω στο γνωστικό αντικείμενο, το οποίο η εργασία διαπραγματεύεται και να αναπτύξω περαιτέρω το ενδιαφέρον μου για την Στατιστική.

Επιπλέον, ευχαριστώ ιδιαιτέρως την κα. Κατέρη Μ. και τον κ. Τζαβελά Γ., για την συμμετοχή τους στην Τριμελή Εξεταστική Επιτροπή και τα εποικοδομητικά τους σχόλια.

Στο σημείο αυτό, θα ήθελα επίσης να εκφράσω τις ευχαριστίες μου στον Professor Mr Tomizawa S. για το ενδιαφέρον που έδειξε, ικανοποιώντας ανιδιοτελώς το αίτημά μου.

Τέλος, θέλω να ευχαριστήσω τον Γεράσιμο, τον Γιώργο, την Κυριακή, την Ζωή, τον Μάνο, τον Alex, την Ευδοκία και τον John D., για όσα χωρίς να γνωρίζουν προσέφεραν.

Περίληψη

Υποκειμενικές εκτιμήσεις της ποιότητας ζωής, του πόνου, της ικανότητας κτλ, είναι συχνές στις κλινικές έρευνες. Στην πραγματικότητα, σκοπός κάθε ερευνητικής προσπάθειας, είναι να ανακαλύψει σημαντικές σχέσεις ή *συνάφειες*, μεταξύ των μεταβλητών. Όμως, σπανίως είναι αρκετό να γνωρίζουμε μόνο την ύπαρξη κάποιας σχέσης, κι έτσι οι έρευνες επικεντρώνονται στην ποσοτικοποίηση της συνάφειας, απαντώντας στο ερώτημα, «*Σε τι βαθμό η μεταβλητή X σχετίζεται με την μεταβλητή Y ;* »

Πάρα πολλά μέτρα ή συντελεστές έχουν προταθεί για το σκοπό αυτό. Τα μέτρα συνάφειας διανύουν μια διαδρομή 3 αιώνων και το πρόβλημα που αντιμετωπίζουν κυρίως οι ερευνητές, είναι η επιλογή του καταλληλότερου μέτρου για κάθε περίπτωση, καθώς τα μέτρα συνήθως αποδίδουν διαφορετικές τιμές, ακόμα και στην περίπτωση που αναλύουμε το ίδιο σύνολο δεδομένων. Η επιλογή ενός κατάλληλου μέτρου, βασίζεται μεταξύ άλλων, στις υποθέσεις που κάνει για τον τρόπο υπολογισμού του, καθώς και στα χαρακτηριστικά των υπό εξέταση μεταβλητών.

Στην παρούσα εργασία, κάνουμε μια γενική ανασκόπηση των «παραδοσιακών» μέτρων συνάφειας, που αναπτύχθηκαν κατά την έναρξη του $20^{ου}$ αιώνα, καθώς και όσων αναπτύχθηκαν μεταγενέστερα, βάσει πιο σύγχρονων υποθέσεων, γνωστά και ως μέτρα προγνωστικής συνάφειας. Επιπλέον, παρουσιάζουμε την σχετικά πρόσφατη ανάπτυξη των νεοσύστατων μέτρων συμμετρίας - ασυμμετρίας για τετραγωνικούς πίνακες συνάφειας, τα οποία στηρίζονται στην θεωρία της πληροφορίας. Συγκεκριμένα, παρουσιάζουμε 18 μέτρα συνάφειας ονοματικής ταξινόμησης, 7 μέτρα συνάφειας διατακτικής ταξινόμησης και 34 μέτρα συμμετρίας - ασυμμετρίας.

Ο σκοπός της εργασίας, είναι η λεπτομερή επεξήγηση των υποθέσεων, που κρύβονται στον τρόπο υπολογισμού κάθε μέτρου, παρέχοντας μια ξεκάθαρη ερμηνεία στον αναγνώστη. Εξετάζουμε τις ιδιότητες των μέτρων, περιγράφουμε τα πλεονεκτήματα και τα μειονεκτήματα αυτών, καθώς και τις ενδεχόμενες μεταξύ τους σχέσεις και καταμετρούμε τις δυνατές περιπτώσεις που μπορεί να αντιμετωπίσει ένας ερευνητής, φιλοδοξώντας να διευκολύνουμε την διαδικασία της επιλογής του.

Abstract

Assessment of subjective phenomena such as pain, quality of life, ability etc., is common in clinical researches. Even though, initial scope of each research is to identify significant underlying dependence or association between the variables, in most cases it is by no means enough. Thus, it is often of interest to quantify the statistical dependence, answering the question: «*What is the strength of the association between the variables* X *and* Y ?»

The literature on measures of association or coefficients is now vast and a big variety of them has been proposed, for every instance. The contingency tables analysis and the development of new measures, remain an active area of research today and thus our history covers activities that span 3 centuries. Despite recent advances in the area, many researchers are facing the problem of how to select the most appropriate measure suited to their purposes. Indeed, for any given situation, there may be several different measures that are valid, yielding different resulting values for the same data set. To decide on the appropriate measure, one must consider the assumptions behind measure specifications and to identify the characteristics of the variables being studied.

In this paper we give an overview of the most well-known and widely used measures of association. Some of them, usually called «traditional measures», were developed around the dawn of the 20^{th} century and other, known as «measures of predictive association», were developed later, in the mid of the 20^{th} century and are based on modern assumptions. In addition, we present the measures of symmetry-asymmetry for square contingency tables, which have only recently been developed and are based on Information Theory. More specifically, we present 18 measures of association for nominal variables, 7 measures of association for ordinal variables and 34 measures of symmetry-asymmetry for nominal and ordinal variables.

This paper attempts to clarify the assumptions and the underlying rules for the selected measures, providing the reader with clear interpretations of each measure. In that sense, the properties and the advantages or disadvantages of the measures are also included, aiming at illustrating potential links among them. Further, this paper enumerates possible cases that can be encountered by the researcher, hopefully having the effect of facilitating the selection process.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Συντομογραφιών

PRE	Proportionate Reduction in Error
S	Symmetry
GS	Global Symmetry
QS	Quasi Symmetry
MH	Marginal Homogeneity
TS	Triangular symmetry
CS	Conditional Symmetry
DS	Diagonal Symmetry

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Κατηγορικά δεδομένα

Προκειμένου να αξιολογήσουν την αποτελεσματικότητα μιας νέας επαναστατικής θεραπείας ή να προσδιορίσουν τους παράγοντες που επηρεάζουν τη γνώμη και τη συμπεριφορά ενός κοινωνικού συνόλου, οι ερευνητές συχνά ανατρέχουν σε στατιστικές μεθόδους ανάλυσης κατηγορικών δεδομένων. Σε αντίθεση με τις συνεχείς μεταβλητές, το βασικό γνώρισμα των κατηγορικών μεταβλητών, ονοματικών ή διατακτικών, είναι ότι δεν μπορεί να μετρηθεί η απόσταση μεταξύ των τιμών τους. Η ανάλυση κατηγορικών δεδομένων παρέχει τις κατάλληλες τεχνικές για την εξαγωγή της επιθυμητής πληροφορίας, όπως στη περίπτωση των συνεχών μεταβλητών, με την ανάλυση παλινδρόμησης. Η πιο κοινή έννοια στην ανάλυση κατηγορικών δεδομένων είναι οι πίνακες συνάφειας.

Όταν ένα σύνολο παρατηρήσεων αποτελείται από δυο ή περισσότερες μεταβλητές και στη συνέχεια οι κατηγορίες κάθε μεταβλητής διασταυρώνονται (*cross-classified*) σε ένα πίνακα συνάφειας, συχνά εγείρονται ερωτήματα, σχετικά με την ύπαρξη σχέσης ή συνάφειας, μεταξύ των μεταβλητών, ενώ παράλληλα το ενδιαφέρον επικεντρώνεται στην ένταση αυτής. Ειδικότερα, οι ερευνητές συνήθως επιθυμούν να προσδιορίσουν το βαθμό της σχέσης μεταξύ δυο μεταβλητών, χρησιμοποιώντας έναν απλό συντελεστή ή ένα μέτρο συνάφειας (*measure of association*), δηλαδή ένα καθαρό αριθμό που κυμαίνεται στο διάστημα $[-1,+1]$ ή $[0,1]$ και δείχνει πόσο έντονα οι δύο μεταβλητές συσχετίζονται. Για παράδειγμα, πως μπορούμε να μετρήσουμε τη συνάφεια μεταξύ δυο κατηγορικών μεταβλητών, όπως η θρησκεία, το επάγγελμα, ή η προτίμηση μεταξύ διάφορων επιλογών;

Σε τέτοιες περιπτώσεις, είναι σημαντικό να χρησιμοποιηθεί ένα κατάλληλο μέτρο, για να εκτιμηθεί ο βαθμός της συνάφειας. Δεν είναι ασυνήθιστο το φαινόμενο, ένας ερευνητής να επιλέγει έναν ακατάλληλο συντελεστή για να μετρήσει μια δοσμένη συνάφεια, εξάγοντας έτσι εσφαλμένα ή παραπλανητικά συμπεράσματα.

1.2 Πεδία εφαρμογών κατηγορικών δεδομένων

Μια κατηγορική μεταβλητή έχει ως κλίμακα μέτρησης ένα σύνολο αμοιβαία αποκλειόμενων κατηγοριών, με την έννοια ότι ένα υποκείμενο δεν μπορεί να ταξινομηθεί σε δυο κατηγορίες. Για παράδειγμα, οι πολιτικές πεποιθήσεις ενός ατόμου θα μπορούσαν να θεωρηθούν ως φιλελεύθερες, μετριοπαθείς ή συντηρητικές. Η διάγνωση μιας μαστογραφίας ενδεχομένως να είναι κανονική, καλοήθης, ύποπτη και κακοήθης. Η ανάπτυξη μεθόδων ανάλυσης κατηγορικών μεταβλητών, παρακινήθηκε από τις έρευνες κοινωνικών και βιοιατρικών επιστημών.

Κατηγορικές κλίμακες είναι διάχυτες στις κοινωνικές επιστήμες για την εκτίμηση υποκειμενικών φαινόμενων και τη μέτρηση της συμπεριφοράς και της γνώμης των ατόμων. Στις ιατρικές επιστήμες χρησιμοποιούνται για τη μέτρηση του αποτελέσματος μιας απόκρισης, όπως για παράδειγμα, εάν μια ιατρική θεραπεία είναι επιτυχής. Χρησιμοποιούνται στην ψυχιατρική για την διάγνωση του τύπου μιας ψυχικής νόσου (σχιζοφρένεια, κατάθλιψη, νεύρωση), στην ψυχολογία για την αξιολόγηση του επιπέδου άγχους (καθόλου, λίγο, αρκετά, πολύ, πάρα πολύ), στην επιδημιολογία και σε έρευνες για την δημόσια υγεία (η ενημέρωση για τον ιό του Aids έχει εντείνει την χρήση μέτρων προφύλαξης; ναι, όχι).

Επίσης, τις συναντούμε στην Ζωολογία για την ταξινόμηση των ζώων ανάλογα με τις διατροφικές τους συνήθειες, στην Γενετική για την αξιολόγηση του είδος του γονότυπου που κληρονομείται, στο Marketing για την μέτρηση της προτίμησης των καταναλωτών. Επεκτείνονται στην Μηχανολογία και στον βιομηχανικό έλεγχο ποιότητας, όταν διάφορα είδη ταξινομούνται ανάλογα με το αν ακολουθούν συγκεκριμένα πρότυπα ποιότητας (πόσο γευστικό είναι ένα συγκεκριμένο φαγητό ή πόσο εύκολη θεωρείται από έναν εργάτη μια συγκεκριμένη εργασία), στην Δημογραφία για την αξιολόγηση των μεταναστευτικών τάσεων και του κοινωνικοοικονομικού status, στην Πολιτική (κοινωνική αποδοχή των κυβερνητικών επιλογών ή στροφή του εκλογικού σώματος μετά από σημαντικά πολιτικά γεγονότα), στην Εκπαίδευση, την Οικολογία και στα Αθλητικά [*Agresti*, (2002)].

1.3 Ιστορική Αναδρομή

Ένα μεγάλο κομμάτι της πρώιμης ανάπτυξης των κατηγορικών δεδομένων, δηλαδή των πινάκων συνάφειας και των σχετικών μέτρων, έλαβε χώρα στην Αγγλία και καλυπτόταν από ένα πέπλο αντιπαραθέσεων, σχετικά με το ποια μέθοδος ήταν καταλληλότερη για να συνοψίσει την συνάφεια των μεταβλητών. Συγκεκριμένα το 1900 στο Λονδίνο, ο *Karl Pearson* εισήγαγε το $\chi^2 - test$ της ανεξαρτησίας, για την σύγκριση των παρατηρούμενων και αναμενόμενων (*θεωρητικών*) συχνοτήτων.

Ο *Pearson* (1857-1936), ήταν ήδη γνωστός στην στατιστική κοινότητα και οι εργασίες του τις προηγούμενες δεκαετίες αφορούσαν, την ανάπτυξη μιας οικογένειας επικλινών (*skewed*) κατανομών πιθανότητας (γνωστών ως *Pearson curves*), τον υπολογισμό του εκτιμητή του συντελεστή συσχέτισης και του τυπικού του σφάλματος καθώς επίσης την επέκταση των εργασιών του *Galton* αναφορικά με την γραμμική παλινδρόμηση. Ο *Pearson* υπέθετε, ότι μια συνεχής δυαδική κατανομή διέπει τους πίνακες διασταυρούμενων συχνοτήτων (*cross-classification tables*). Υποστήριζε ότι κάποιος θα μπορούσε να περιγράψει τη συνάφεια (*association*), προσεγγίζοντας ένα μέτρο σαν αυτό της συσχέτισης (*correlation*), για την υποκείμενη συνέχεια. Η άποψη αυτή οδήγησε αργότερα (1904) τον *Pearson*, να αναπτύξει τον τετραχωρικό συντελεστή συσχέτισης για 2×2 πίνακες. Επίσης, το 1904, ο *Pearson* εισήγαγε την έννοια της συνάφειας (*contingency*), καθώς επίσης και διάφορα μέτρα που βασίζονταν στο $\chi^2 - test$ και περιέγραφαν την ένταση της συνάφειας.

Ο *George U. Yule* (1871-1951) επίσης Άγγλος, είχε μια διαφορετική προσέγγιση. Έχοντας ήδη αναπτύξει μοντέλα πολλαπλής παλινδρόμησης και πολλαπλούς ή μερικούς συντελεστές συσχέτισης, ο *Yule* έστρεψε το ενδιαφέρον του, μεταξύ του 1900 και 1912, στην συνάφεια κατηγορικών μεταβλητών. Πίστευε ότι πολλές κατηγορικές μεταβλητές, όπως επιβίωση ή όχι, απασχολούμενος ή άνεργος, είναι εγγενείς διακριτές. Όρισε κάποιους δείκτες με έναν άμεσο τρόπο, χρησιμοποιώντας τις μετρήσεις των κελιών του πίνακα, χωρίς να υποθέτει μια υποκείμενη συνέχεια. Επιπλέον, διέδωσε τον λόγο σχετικών πιθανοτήτων (*odds ratio*), μια έννοια η οποία, όπως ο *Goodman* (2000) αναφέρει, ίσως πρώτα να προτάθηκε από τον Ούγκρο στατιστικό *Korosy J.* και στην συνέχεια, παρουσίασε κάποιους μετασχηματισμούς αυτού, γνωστούς ως

συντελεστές Q και Y του *Yule*. Το 1912, εισήγαγε τον συντελεστή φ, ο οποίος βασίζεται στο $\chi^2 - test$. Το 1903, ο *Yule* επίσης έδειξε την ενδεχόμενη ασυμφωνία μεταξύ της περιθώριας (*marginal*) και δεσμευμένης (*conditional*) συνάφειας, σε πίνακες συνάφειας. Πολλά χρόνια αργότερα, το 1951, η ασυμφωνία αυτή καταγράφηκε από τον *Simpson E. H.* και είναι πλέον γνωστή ως παράδοξο του *Simpson* (*Simpson's Paradox*).

Συνοψίζοντας, ο *Yule* υποστήριζε, ότι συχνά είναι παραπλανητικό και οδηγεί σε εσφαλμένα αποτελέσματα, όταν αναγκαζόμαστε να υποθέτουμε ότι τα δεδομένα προέρχονται από μια συνεχή κατανομή πιθανότητας, η οποία είναι κανονική. Ο *Pearson* υποστήριζε ότι τα μέτρα του *Yule* καταλήγουν σε διαφορετικές τιμές όταν ένας $I \times J$ πίνακας συμπτυχθεί σε έναν 2×2 πίνακα [βλέπε *Pearson & Heron,* (1913)]. Κάνοντας μια ανασκόπηση, οι *Pearson* και *Yule* είχαν από κοινού δίκιο. Κάποιες ταξινομήσεις όπως αυτές των ονοματικών μεταβλητών, δεν διέπονται από κάποια εμφανή συνεχή κατανομή. Από την άλλη μεριά, σε πολλές εφαρμογές τα δεδομένα διέπονται από μια υποκείμενη συνέχεια και το γεγονός αυτό θα μπορούσε να αποτελεί κίνητρο για την ανάπτυξη μοντέλων και συμπερασματολογίας. Ο *Goodman* (1981a,b) αναφέρει ότι τα διατακτικά μοντέλα παρέχουν ένα είδος συμφωνίας μεταξύ του *Yule* και του *Pearson*, καθώς το *odds ratio* χαρακτηρίζει μοντέλα τα οποία προσαρμόζονται καλά, όταν η κατανομή των δεδομένων είναι κατά προσέγγιση κανονική.

Η διαμάχη των *Pearson* και *Yule* ήταν μικρής σημασίας, συγκριτικά με αυτή μεταξύ των *Pearson* και *Fisher*. Ο *Ronald A. Fisher* (1890-1962) επίσης Άγγλος, χρησιμοποιώντας μια γεωμετρική απεικόνιση, εισήγαγε την έννοια των βαθμών ελευθερίας (*degrees of freedom*), για να χαρακτηρίσει την οικογένεια κατανομών χ^2 [*Fisher*, (1922)]. Ο *Fisher* υποστήριξε ότι οι βαθμοί ελευθερίας του $\chi^2 - test$ της ανεξαρτησίας, για ένα $I \times J$ πίνακα συνάφειας, ισούνται με $df = (I-1) \times (J-1)$, σε αντίθεση με ότι αρχικά (1900) ο *Pearson* είχε υπολογίσει, ο οποίος θεωρούσε ότι οι βαθμοί ελευθερίας ισούνται με $df = IJ - 1$. Η πρόταση του *Fisher* για διόρθωση των βαθμών ελευθερίας του $\chi^2 - test$ της ανεξαρτησίας, οδήγησε σε μια νέα ιστορική αντιπαράθεση. Ο *Fisher* τελικά απέδειξε τους ισχυρισμούς του το 1926. Το 1935, εισήγαγε το ακριβές τεστ (*Fisher's exact test*), για τον έλεγχο της ανεξαρτησίας σε 2×2 πίνακες, με συχνότητες (< 5) σε κάθε κελί.

Παράλληλα, το 1918, ο *Tschuprow* εισήγαγε τον συντελεστή T, το 1938 ο *Kendal* εισήγαγε τους συντελεστές *tau*, το 1946 ο *Cramer* εισήγαγε το συντελεστή V και το 1948 ο *Pearson* πρότεινε τον συντελεστή C. Σημειώνουμε ότι το 1977, ο *Sakoda* πρότεινε μια διόρθωση του συντελεστή C. Το 1935, ο Βρετανός στατιστικός *Maurice Bartlett* εισήγαγε τον ορισμό της ομοιογενούς συνάφειας (*homogeneous association*), χωρίς αλληλεπιδράσεις, για $2\times 2\times 2$ πίνακες συνάφειας. Ο *Norton* (1945) επέκτεινε τα αποτελέσματα του *Bartlett* σε $2\times 2\times k$ πίνακες. Το 1951, ο *Jerome Cornfield*, ένας στατιστικός με δεσμούς στην ιατρική επιστήμη, χρησιμοποίησε το *odds ratio* για να προσεγγίσει τον σχετικό κίνδυνο σε *case-control* μελέτες.

Οι περισσότερες εργασίες και στατιστικά βιβλία, όταν ανατρέχουν σε πηγές σχετικές με την ανάλυση πινάκων συνάφειας, αναφέρονται κυρίως στις ανωτέρω εργασίες των *Pearson* και *Yule* στις αρχές του 20ου αιώνα. Όμως, όπως ο *Stigler* (2002) αναφέρει, η ιδέα της ανάλυσης πινάκων συνάφειας χρονολογείται πολύ νωρίτερα, κατά την διάρκεια του 19ου, από τον Βελγικής καταγωγής *Quetelet* (1849), αναφορικά με την μέτρηση της συνάφειας και την έννοια του σχετικού κινδύνου και από τον *Bienayme* αναφορικά με την υπεργεωμετρική ανάλυση ενός 2×2 πίνακα συνάφειας [βλέπε *Heyde & Seneta*, (1977)]. Επίσης, ο *Francis Galton* (1892), εισήγαγε την έννοια της αναμενόμενης μέτρησης, ως την βάση για την μέτρηση της συνάφειας, σύμφωνα με τον τύπο $Expected\ Count(i,j)=\dfrac{(Row\ total\ i)\times(Column\ total\ j)}{(Grand\ total)}$ ο οποίος αργότερα, θα διαδραμάτιζε ένα σπουδαίο ρόλο για τον υπολογισμό του χ^2-test, ελέγχου της ανεξαρτησίας [βλέπε *Fienberg & Rinaldo*, (2007)]. Επίσης, όπως οι *Goodman & Kruskal* (1959) αναφέρουν, ο *M. H. Doolittle* με άρθρα του το 1887, προσπαθούσε να εξηγήσει την έλλειψη ακρίβειας ακόμη και για έναν 2×2 πίνακα, σε μια πρώτη προσπάθεια να ποσοτικοποιήσει την συνάφεια.

Μισό αιώνα αργότερα (1954), μετά την αντιπαράθεση των *Pearson* και *Yule*, οι *Leo Goodman* και *William Kruskal*, από το Πανεπιστήμιο του Chicago, εισήγαγαν έναν αριθμό εναλλακτικών μέτρων συνάφειας, γνωστά ως μέτρα προγνωστικής συνάφειας, που βασίζονται σε ένα πιθανοθεωρητικό μοντέλο και όχι στο γνωστό χ^2-test ή στην υπόθεση ότι τα δεδομένα προέρχονται από μια κοινή κανονική κατανομή, . Το βιβλίο τους το (1979), παρουσιάζει τέσσερα σημαντικά άρθρα, κατά την διάρκεια του 1950 και αποτελεί μια πολύ καλή αναφορά. Επιπλέον,

ο 1962, ο *Somers* εισήγαγε το συντελεστή συνάφειας *D* και το 1972, ο *Theil* εισήγαγε τον συντελεστή αβεβαιότητας *U* , που βασίζεται σε μέτρα στατιστικής πληροφορίας.

Οι πίνακες συνάφειας, αν και θεωρούνται η πιο γνωστή μέθοδος ανάλυσης κατηγορικών δεδομένων , δεν είναι η μοναδική. Στα μέσα του 1930, αναπτύχθηκαν τα πρώτα μοντέλα για κατηγορικές μεταβλητές. Η ανάλυση των κατηγορικών δεδομένων αποτελεί ακόμη και σήμερα, μια επιστημονική περιοχή που προσφέρεται για έρευνα, καλύπτοντας μια διαδρομή 3 αιώνων. Η βιβλιογραφία αναφορικά με την ανάλυση κατηγορικών δεδομένων είναι ανεξάντλητη και υπάρχουν διάφορες πτυχές που περιλαμβάνουν διαφορετικά μοντέλα και μεθόδους. Έτσι, οι κατηγορικές μεταβλητές χρησιμοποιούνται σε κάθε πεδίο γνώσης και δραστηριότητας. Ο *Agresti* (2002), παρέχει μια αναλυτική επισκόπηση.

1.4 Συντελεστές παλινδρόμησης και μέτρα συνάφειας

Η ένταση μεταξύ έμμεσων ή άμεσων σχέσεων μεταξύ συνεχών μεταβλητών συνοψίζεται με τους συντελεστές παλινδρόμησης. Για σχέσεις μεταξύ κατηγορικών μεταβλητών, η διαδικασία είναι λιγότερο εμφανής. Η πρόσφατη ανάπτυξη συμπερασματικής ανάλυσης κατηγορικών δεδομένων, κυριαρχείται από παραμετρικά μοντέλα. Όμως, μεταξύ των προτεινόμενων μοντέλων, κανένα δεν παρέχει μονούς (*single*) δείκτες των άμεσων ή έμμεσων σχέσεων μεταξύ πολύτομων μεταβλητών. Τα λογαριθμογραμμικά μοντέλα επικεντρώνονται μεταξύ άλλων, στην ανίχνευση δομών συνάφειας (*association patterns*). Ομοίως, η παραγοντική ανάλυση αντιστοιχιών (*factorial correspondence analysis*), επικεντρώνεται σε δεσμούς ή σχέσεις μεταξύ των κατηγοριών παρά μεταξύ των μεταβλητών, ενώ μοντέλα όπως αυτά της λογιστικής παλινδρόμησης ή της *Poisson* παλινδρόμησης, προσπαθούν να εξηγήσουν την πιθανότητα μια παρατήρηση να ανήκει σε μια δοσμένη κατηγορία. Οι παράμετροι της λογιστικής παλινδρόμησης, ίσως να δίνουν πληροφόρηση σχετικά με τους δεσμούς μεταξύ δίτιμων μεταβλητών. Όμως για πολύτομες μεταβλητές δεν αποκτούμε σύνθετους δείκτες παρά ένα σύνολο παραμέτρων. Όπως για παράδειγμα, το μοντέλο συνάφειας γραμμής - στήλης του *Goodman* περιέχει παραμέτρους συνάφειας. Έχουν όμως εφαρμογή σε διατακτικές μεταβλητές μόνο. Επιπλέον, ο αριθμός των παραμέτρων είναι μεγαλύτερος των ανα δυο σχέσεων και το μοντέλο δεν είναι λυσιτελές (*parsimonious*). Κατά συνέπεια, η παραμετρική προσέγγιση είναι σε

γενικές γραμμές μικρής σημασίας για την αξιοποίηση και έλεγχο ενός αιτιατού γραφήματος κατηγορικών μεταβλητών. Τα παραδοσιακά μέτρα συνάφειας και οι μερικοί δείκτες συνάφειας, παραμένουν τα πιο κατάλληλα εργαλεία για τον σκοπό αυτό [*Olszak & Ritschard*, (1995)].

1.5 Συνάφεια και Ασυμμετρία

Ο όρος συνάφεια (*contingency*), φαίνεται να προέρχεται από τον *Pearson* (1904), ο οποίος όρισε την συνάφεια για έναν $I \times J$ πίνακα σαν «ένα μέτρο της συνολικής απόκλισης της ταξινόμησης από την ανεξάρτητη πιθανότητα» (*Fienberg & Rinaldo* 2007). Η πιο γνωστή μέθοδος που εφαρμόζεται στην ανάλυση ενός πίνακα συνάφειας είναι το $\chi^2 - test$ της ανεξαρτησίας, που αναπτύχθηκε επίσης από τον *Pearson* (1900). Αν με X και Y συμβολίσουμε δυο κατηγορικές μεταβλητές με I και J κατηγορίες αντίστοιχα, τότε η διασταύρωσή τους οδηγεί σε έναν $I \times J$ πίνακα. Από την στιγμή που απορριφθεί η υπόθεση της ανεξαρτησίας, τότε η πληροφόρηση αναφορικά με την στατιστικά σημαντική συνάφεια των υπό εξέταση μεταβλητών, παρέχεται από μια ποικιλία μέτρων συνάφειας. Όταν η ένταση της συνάφειας μεταξύ δυο μεταβλητών, περιγράφεται από τον βαθμό απομάκρυνσης από την ανεξαρτησία, τότε αναφερόμαστε στα μέτρα συνάφειας (*measures of association*), ενώ όταν περιγράφεται σε όρους απομάκρυνσης από την συμμετρία, τότε αναφερόμαστε στα μέτρα συμμετρίας - ασυμμετρίας (*measures of symmetry - asymmetry*).

1.6 Διάρθρωση της εργασίας

Η παρούσα εργασία αποτελεί μια γενική επισκόπηση των πιο γνωστών μέτρων συνάφειας, που χρησιμοποιούνται στην ανάλυση πινάκων συνάφειας, για την μελέτη και ποσοτικοποίηση της έντασης της σχέσης μεταξύ δυο μεταβλητών, αλλά και πιο σύγχρονων προσεγγίσεων, που ίσως να μην έχουν την ίδια αναγνώριση, όπως των νεοσύστατων μέτρων συμμετρίας - ασυμμετρίας. Σκοπός της εργασίας είναι, η συνοπτική παρουσίαση των μέτρων και η ταξινόμησή τους με βάση τις υποθέσεις και τις ιδιότητές τους. Παράλληλα, εφαρμόζουμε τα μέτρα σε επιλεγμένα παραδείγματα, με στόχο την απόδοση της ερμηνεία τους και την καταγραφή των

ενδεχόμενων περιορισμών τους. Επιπλέον, συγκρίνουμε τα μέτρα για το ίδιο σύνολο δεδομένων, φιλοδοξώντας να εξάγουμε χρήσιμα συμπεράσματα, αναφορικά με τα κριτήρια επιλογής τους.

Συγκεκριμένα, στο Κεφάλαιο 2, θα αναφερθούμε στις βασικές έννοιες που χρησιμοποιούνται στην ανάλυση πινάκων συνάφειας, έτσι ώστε να διευκολύνουμε την παρουσίαση των μέτρων στα επόμενα Κεφάλαια. Θα περιγράψουμε τα είδη των κατηγορικών μεταβλητών και τις διάφορες κλίμακες μέτρησης, που αποτελούν βασικό κριτήριο επιλογής ενός μέτρου, καθώς και τα είδη των πινάκων συνάφειας.

Στο Κεφάλαιο 3, θα αναφερθούμε στις βασικές ιδιότητες των μέτρων, στα κριτήρια επιλογής τους, ενώ παράλληλα, θα περιγράψουμε τις προϋποθέσεις που θα πρέπει να πληρούνται για την κατασκευή ενός μέτρου συνάφειας. Επιπλέον, θα παρουσιάσουμε τις διάφορες κλάσεις μέτρων πληροφορίας και απόστασης, που προέρχονται από τον χώρο της θεωρίας της πληροφορίας (*Information Theory*) και είναι χρήσιμες για τον υπολογισμό των μέτρων συμμετρίας – ασυμμετρίας.

Στο Κεφάλαιο 4, θα παρουσιάσουμε τα κυριότερα μέτρα για ονοματικές μεταβλητές, τις βασικές ιδιότητες και την ερμηνεία τους. Συγκεκριμένα, θα αναφερθούμε στα «*παραδοσιακά*» μέτρα συνάφειας, όπως συνηθίζεται να αποκαλούνται, δηλαδή σε μέτρα που βασίζονται στο $\chi^2 - test$ του *Pearson* και στο *odds ratio*, καθώς και στα μέτρα προγνωστικής συνάφειας ή αναλογικής μείωσης του σφάλματος πρόβλεψης (*PRE – Proportionate Reduction in Error*), τα οποία εμφανίστηκαν μεταγενέστερα. Επιπλέον, θα εξηγήσουμε την κεντρική ιδέα που κρύβεται πίσω από την κατασκευή κάθε μέτρου, ώστε να είναι ξεκάθαροι οι λόγοι για τους οποίους κάθε μέτρο θα πρέπει να χρησιμοποιείται.

Στο Κεφάλαιο 5, θα παρουσιάσουμε τα κυριότερα μέτρα για διατακτικές μεταβλητές. Στην περίπτωση, αυτή ο τρόπος υπολογισμού των μέτρων διαφοροποιείται, καθώς αξιοποιούν την επιπρόσθετη πληροφορία που παρέχει η διάταξη των κατηγοριών κάθε μεταβλητής.

Στο Κεφάλαιο 6, θα παρουσιάσουμε τα βασικότερα μοντέλα συμμετρίας – ασυμμετρίας και τα αντίστοιχα μέτρα συμμετρίας - ασυμμετρίας για τετραγωνικούς πίνακες συνάφειας.

ΚΕΦΑΛΑΙΟ 2

ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΑΝΑΛΥΣΗΣ ΚΑΤΗΓΟΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

2.1 Εισαγωγή

Στο κεφάλαιο αυτό θα αναφερθούμε σε βασικές έννοιες που χρησιμοποιούνται στην ανάλυση πινάκων συνάφειας, έτσι ώστε να διευκολύνουμε την επεξήγηση των μέτρων στα επόμενα κεφάλαια. Συγκεκριμένα, θα περιγράψουμε τα είδη των κατηγορικών μεταβλητών και τις κλίμακες μέτρησής τους, που αποτελούν βασικό κριτήριο επιλογής ενός μέτρου, καθώς και τα είδη των πινάκων συνάφειας.

2.2 Πίνακες Συνάφειας

Ας υποθέσουμε ότι έχουμε δυο κατηγορικές μεταβλητές X,Y. Έστω I ο αριθμός των κατηγοριών της μεταβλητής X και J ο αριθμός των κατηγοριών της μεταβλητής Y. Η ταξινόμηση των υποκειμένων των δυο μεταβλητών στις διάφορες κατηγορίες έχει IJ δυνατούς συνδυασμούς. Η απόκριση (X,Y) ενός τυχαίως επιλεγμένου υποκειμένου από κάποιο πληθυσμό, έχει μια κατανομή πιθανότητας. Ένας ορθογώνιος πίνακας με I γραμμές για τις κατηγορίες της μεταβλητής X και J στήλες για τις κατηγορίες της μεταβλητής Y, έχει IJ κελιά, τα οποία περιγράφουν τα πιθανά αποτελέσματα αυτών των συνδυασμών και προσδιορίζουν την από κοινού κατανομή τους. Όταν τα κελιά περιέχουν τις συχνότητες για κάθε IJ πιθανό αποτέλεσμα ενός δείγματος, τότε έχουμε έναν διδιάστατο $I \times J$ πίνακα συνάφειας [*Agresti*, (2002)]. Ο ακόλουθος Πίνακας 2-1 (*σελ.10*) περιγράφει τον τρόπο με τον οποίο ταξινομούνται δυο μεταβλητές.

ΠΙΝΑΚΑΣ 2-1

$I \times J$ Πίνακας Συνάφειας

	Μεταβλητή Y				
Μεταβλητή X	Y_1	Y_2	. . .	Y_j	**Σύνολο**
X_1	N_{11} π_{11}	N_{12} π_{12}	. . . . . .	N_{1j} π_{1j}	$\mathbf{N_{1.}}$ $\boldsymbol{\pi_{1.}}$
X_2	N_{21} π_{21}	N_{22} π_{22}	. . . . . .	N_{2j} π_{2j}	$\mathbf{N_{2.}}$ $\boldsymbol{\pi_{2.}}$
. . .	. . .	. . .	. . .	. . .	. . .
X_i	N_{i1} π_{i1}	N_{i2} π_{i2}	. . . . . .	N_{ij} π_{ij}	$\mathbf{N_{i.}}$ $\boldsymbol{\pi_{i.}}$
Σύνολο	$\mathbf{N_{.1}}$ $\boldsymbol{\pi_{.1}}$	$\mathbf{N_{.2}}$ $\boldsymbol{\pi_{.2}}$	. . . . . .	$\mathbf{N_{.j}}$ $\boldsymbol{\pi_{.j}}$	**N** **1**

2.2.1 Είδη Πινάκων Συνάφειας

Η απλούστερη μορφή ενός πίνακα συνάφειας είναι αυτή, ενός 2×2 πίνακα και προέρχεται από την διασταύρωση δυο διχοτομημένων μεταβλητών. Κυρίως εμφανίζονται σε βιοιατρικές εφαρμογές και σε επιδημιολογικές μελέτες (Επιβίωση/Θεραπεία: επιτυχία-αποτυχία Group: case-control) και χρησιμοποιούνται όταν θέλουμε να συγκρίνουμε δυο ομάδες ως προς μια δίτιμη αποκριτική μεταβλητή.

Η συνηθέστερη μορφή ενός πίνακα που συναντάται σε διάφορες μελέτες και πειράματα, είναι ο διδιάστατος $I \times J$ πίνακας συνάφειας, για $I \neq J$, όπου μπορούμε να αναλύσουμε την συνάφεια δυο μεταβλητών, ανεξάρτητα από το μέγεθος των κατηγοριών τους. Η ειδική περίπτωση ενός τετραγωνικού $I \times I$ πίνακα συνάφειας, με σύμμετρη (*ίδια*) ταξινόμηση των μεταβλητών εμφανίζεται στις βιοιατρικές, παιδαγωγικές και κοινωνικές επιστήμες, στην ψυχολογία και σε άλλα επιστημονικά πεδία. Χαρακτηριστικά παραδείγματα είναι η σύγκριση της διάγνωσης ή της θεραπείας στο ίδιο υποκείμενο, αλλά από δυο διαφορετικούς εξεταστές (*συμφωνία*

βαθμολογητών), οι πίνακες που περιγράφουν την γεωγραφική κινητικότητα κοινωνικών στρωμάτων, η ανάλυση της προτίμησης της κοινής γνώμης, μεταξύ δυο περιόδων κ.α.

Στην περίπτωση περισσότερων από δυο μεταβλητών, αναφερόμαστε σε πολυδιάστατους πίνακες συνάφειας $I \times J \times ... \times R$. Πολλές μελέτες ιδιαίτερα στις κοινωνικές επιστήμες, έχουν να κάνουν με δεδομένα που περιγράφονται από πολλές μεταβλητές. Για παράδειγμα, ένας ενήλικας που ζει σε ένα μεγάλο αστικό κέντρο, θα μπορούσε να ταξινομηθεί με βάση τις εξής κατηγορίες ανά μεταβλητή: Δημοτικό διαμέρισμα $I = 5$, Αγαπημένη εφημερίδα $J = 6$, Ύπαρξη τηλεόρασης $K = 2$ (ναι-όχι), Επίπεδο εκπαίδευσης $L = 4$, Ηλικία $R = 10$.

Για λόγους ευκολίας της παρουσίασης και καλύτερης ερμηνείας των αποτελεσμάτων, θα περιοριστούμε σε εφαρμογές δυο διασταυρούμενων μεταβλητών, αν και τα σχόλιά μας, θα μπορούσαν να επεκταθούν για περιπτώσεις περισσότερων μεταβλητών.

2.2.2 Πλεονεκτήματα Πινάκων Συνάφειας

Τα κατηγορικά δεδομένα προσφέρονται για απεικόνιση σε πίνακα (*tabulation*), καθώς ένας πίνακας διατηρεί όλη την πληροφορία των δεδομένων αναλλοίωτη και κάνει την δομή των δεδομένων πιο ξεκάθαρη. Συγκεκριμένα, μας επιτρέπει να δούμε 4 χαρακτηριστικά τα οποία δεν θα ήταν εμφανή αν τα δεδομένα ήταν ακατέργαστα:

1. Την συνολική κατανομή της μεταβλητής X
2. Την συνολική κατανομή τη μεταβλητής Y
3. Πως η κατανομή της μεταβλητής X διαφοροποιείται κατά μήκος της μεταβλητής Y
4. Την διαφορετική αναλογία της μεταβλητής Y σε κάθε επίπεδο της μεταβλητής X.

Τα δυο τελευταία χαρακτηριστικά εκφράζουν την συνάφεια μεταξύ των μεταβλητών [http://teaching.sociology.ul.ie/SSS/lugano/lugano.html, *Brendan Halpin*, (2002)].

2.2.3 Συμβολισμοί ενός Πίνακα Συνάφειας

Έστω ότι ένα υποκείμενο επιλέγεται τυχαία από τον πληθυσμό και κατατάσσεται σε έναν από τους IJ πιθανούς συνδυασμούς, των κατηγοριών των μεταβλητών X, Y. Συμβολίζουμε με $i = 1, 2, ..., I$ και $j = 1, 2, ..., J$ τον αριθμό των κατηγοριών των μεταβλητών X, Y αντίστοιχα,

N_{ij} την συχνότητα εμφάνισης του (ij) συνδυασμού.

$N_{i.} = \sum_{j=1}^{J} N_{ij}$ το περιθώριο άθροισμα συχνοτήτων της $i-$γραμμής του πίνακα.

$N_{.j} = \sum_{i=1}^{I} N_{ij}$ το περιθώριο άθροισμα συχνοτήτων της $j-$στήλης του πίνακα.

$N = \sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}$ το συνολικό μέγεθος του πληθυσμού.

$\pi_{ij} = P(X = i, Y = j) = N_{ij}/N$ την πιθανότητα ένα υποκείμενο να ταξινομηθεί στην (ij) συντεταγμένη του πίνακα συνάφειας. Ισχύει ότι $\sum_{i=1}^{I}\sum_{j=1}^{J} \pi_{ij} = 1$.

$\pi_{i.} = P(X = i) = N_{i.}/N$ την περιθώρια πιθανότητα ένα υποκείμενο να ταξινομηθεί στην $i-$γραμμή του πίνακα.

$\pi_{.j} = P(Y = j) = N_{.j}/N$ την περιθώρια πιθανότητα ένα υποκείμενο να ταξινομηθεί στην $j-$στήλη του πίνακα.

$\pi_{i|j} = P(X = i \mid Y = j) = \pi_{ij}/\pi_{.j}$ η δεσμευμένη πιθανότητα ένα υποκείμενο να ταξινομηθεί στην $i-$γραμμή του πίνακα, δοθέντος ότι το υποκείμενο ταξινομείται στην $j-$στήλη του πίνακα.

$\pi_{j|i} = P(Y = j \mid X = i) = \pi_{ij}/\pi_{i.}$ η δεσμευμένη πιθανότητα ένα υποκείμενο να ταξινομηθεί στην $j-$στήλη του πίνακα, δοθέντος ότι το υποκείμενο ταξινομείται στην $i-$γραμμή του πίνακα.

$\{\pi_{ij}\}$ η από κοινού κατανομή πιθανότητας (*joint distribution*) των μεταβλητών X, Y.

$\{\pi_{i.}\}, \{\pi_{.j}\}$ οι περιθωριακές κατανομές πιθανότητες (*marginal distributions*) για κάθε κατηγορία των μεταβλητών X, Y.

$\{\pi_{1|j}, \pi_{2|j}, ..., \pi_{I|j}\}, \{\pi_{1|i}, \pi_{2|i}, ..., \pi_{J|i}\}$ οι δεσμευμένες κατανομές πιθανότητας (*conditional distributions*) των μεταβλητών X, Y, για κάθε γνωστό επίπεδο των μεταβλητών Y, X, αντίστοιχα.

Σημειώνουμε ότι οι δειγματικές συχνότητες συμβολίζονται με n_{ij}, οι δειγματικές πιθανότητες με p_{ij} και το συνολικό μέγεθος του δείγματος με n.

2.3 Μεταβλητή απόκρισης και επεξηγηματική μεταβλητή

Σε πολλούς πίνακες συνάφειας, η μια μεταβλητή θεωρείται μεταβλητή απόκρισης (*response variable*) ή αλλιώς εξαρτημένη μεταβλητή (συνήθως η μεταβλητή στήλη Y) και η άλλη επεξηγηματική μεταβλητή (*explanatory variable*) ή αλλιώς ανεξάρτητη (συνήθως η μεταβλητή γραμμή X). Όταν η μεταβλητή X είναι γνωστή και όχι τυχαία, η έννοια της από κοινού κατανομής, που περιγράψαμε στην προηγούμενη παράγραφο δεν έχει νόημα. Όμως, για μια γνωστή κατηγορία της X, η μεταβλητή Y έχει μια κατανομή πιθανότητας. Στην περίπτωση αυτή, έχει νόημα να μελετήσουμε πως η κατανομή πιθανότητας της Y μεταβάλλεται, σε κάθε επίπεδο της μεταβλητής X. Πρωταρχικός σκοπός πολλών μελετών είναι να συγκρίνουν τις δεσμευμένες κατανομές της μεταβλητής απόκρισης Y, για διάφορα επίπεδα της επεξηγηματικής μεταβλητής [*Agresti,* (2002)].

Στην περίπτωση που η ύπαρξη μιας αιτιώδους σχέσης μεταξύ των μεταβλητών προσδιορίζεται και στις δυο κατευθύνσεις, δηλαδή και οι δυο μεταβλητές είναι μεταβλητές απόκρισης (η μια μεταβλητή δεν προηγείται της άλλης, είτε χρονολογικά, είτε αιτιατά, είτε με οποιοδήποτε άλλο τρόπο), τότε οι μεταβλητές αντιμετωπίζονται συμμετρικά. Η διαφοροποίηση αυτή παίζει σπουδαίο ρόλο για την επιλογή ενός μέτρου συνάφειας, όπως θα δούμε στην συνέχεια.

2.4 Ανεξαρτησία

Όταν και οι δυο μεταβλητές είναι μεταβλητές απόκρισης, για να περιγράψουμε την μεταξύ τους συνάφεια, μπορούμε να χρησιμοποιήσουμε την από κοινού κατανομή τους και την δεσμευμένη κατανομή της Y ως προς X ή την δεσμευμένη κατανομή της X ως προς Y. Η δεσμευμένη κατανομή της Y δοθέντος της X, συνδέεται με την από κοινού κατανομή μέσω της σχέσης

$$\pi_{j|i} = \pi_{ij} / \pi_{i.} \text{ , για όλα τα } i, j \tag{2.1}$$

Δυο κατηγορικές μεταβλητές απόκρισης θεωρούνται ανεξάρτητες, εάν όλες οι από κοινού πιθανότητες ισούνται με το γινόμενο των περιθωριακών πιθανοτήτων τους, δηλαδή όταν

$$\pi_{ij} = \pi_{i.}\pi_{.j} \text{ , για } i = 1, 2, ..., I \text{ και } j = 1, 2, ..., J \tag{2.2}$$

Δυο μεταβλητές είναι στατιστικά ανεξάρτητες, εάν οι δεσμευμένες κατανομές της μεταβλητής Y, είναι ίδιες σε κάθε επίπεδο της μεταβλητής X. Όταν οι μεταβλητές X, Y είναι ανεξάρτητες τότε ισχύει

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i.}} = \frac{\pi_{i.}\pi_{.j}}{\pi_{i.}} = \pi_{.j} \text{ , για } i = 1, 2, ..., I \tag{2.3}$$

Με άλλα λόγια, κάθε δεσμευμένη κατανομή της Y είναι ίδια με την περιθωριακή κατανομή της Y. Επομένως, δυο μεταβλητές είναι ανεξάρτητες όταν η πιθανότητα οποιασδήποτε στήλης είναι ίδια σε κάθε γραμμή, δηλαδή όταν ισχύει ότι

$$\pi_{j|1} = \pi_{j|2} = ... = \pi_{j|I} \tag{2.4}$$

Όταν η Y είναι μεταβλητή απόκρισης και η X επεξηγηματική μεταβλητή, μπορούμε με πιο απλό και λογικό τρόπο να ορίσουμε την ανεξαρτησία, από ότι όταν και οι δυο είναι μεταβλητές απόκρισης. Η ανεξαρτησία στην περίπτωση αυτή, συχνά αναφέρεται και ως ομοιογένεια (*homogeneity*) της δεσμευμένης κατανομής. Δηλαδή όταν η δεσμευμένη κατανομή της Y είναι ίδια για κάθε επίπεδο της μεταβλητής X [*Agresti*, (2002)].

2.5 Συμμετρία

Στην περίπτωση ενός τετραγωνικού $R \times R$ πίνακα συνάφειας, ο οποίος έχει την ίδια (*ονοματική ή διατακτική*) ταξινόμηση γραμμών και στηλών, το ενδιαφέρον επικεντρώνεται στην μελέτη της συμμετρίας γύρω από τα στοιχεία της κυρίας διαγώνιου, παρά για την ανεξαρτησία μεταξύ των μεταβλητών γραμμής και στήλης. Γενικά, δυο μεταβλητές θεωρείται ότι έχουν σύμμετρη ταξινόμηση αν

$$\pi_{ij} = \pi_{ji} \text{ για } i, j = 1, 2, ..., R \text{ και } i \neq j \tag{2.5}$$

Στο Κεφάλαιο 6, αναφερόμαστε αναλυτικότερα στην έννοια της συμμετρίας ενός πίνακα συνάφειας, καθώς και άλλων συναφών μοντέλων.

2.6 Είδη κατηγορικών μεταβλητών

Η κλίμακα μέτρησης μιας μεταβλητής μπορεί να είναι συνεχής (*continuous*) ή διακριτή (*discrete*). Μια συνεχής μεταβλητή, θεωρητικά μπορεί να πάρει τιμές από όλο το εύρος του διαστήματος στο οποίο ανήκει, όπως για παράδειγμα το βάρος ή η αρτηριακή πίεση. Αντίθετα μια διακριτή μεταβλητή λαμβάνει μεμονωμένες τιμές, όπως για παράδειγμα το φύλλο ή το επίπεδο εκπαίδευσης. Τα κατηγορικά δεδομένα αποτελούνται από μεταβλητές οι οποίες επιδέχονται έναν περιορισμένο αριθμό διακριτών τιμών. Μπορεί να είναι ονοματικές, διατακτικές ή διαστηματικές, αλλά δεν μπορούν να είναι συνεχείς. Οι κατηγορίες των ονοματικών μεταβλητών δεν έχουν μια φυσική διάταξη. Για παράδειγμα, η μεταβλητή «θρησκευτική ιδεολογία» έχει τις κατηγορίες, Ορθόδοξος, Καθολικός, Προτεστάντης, Εβραίος κ.α.

Οι διατακτικές μεταβλητές αντιθέτως, αποτελούνται από κατηγορίες οι οποίες μπορούν να διαταχθούν. Για παράδειγμα, η μεταβλητή κοινωνική τάξη έχει κατηγορίες όπως, υψηλή τάξη, μεσαία τάξη, χαμηλή τάξη. Οι μεταβλητές αυτές μπορούν μεν να διαταχθούν, αλλά η απόσταση μεταξύ των κατηγοριών είναι άγνωστη. Για παράδειγμα, ένα άτομο με μετριοπαθείς πολιτικές αντιλήψεις είναι πιο φιλελεύθερο από ένα άτομο με συντηρητικές πολιτικές πεποιθήσεις, αλλά δεν υπάρχει μια αριθμητική τιμή που να περιγράφει πόσο πιο φιλελεύθερο είναι.. Όταν η απόσταση μεταξύ των κατηγοριών μπορεί να μετρηθεί, τότε οι μεταβλητές ονομάζονται διαστηματικές (*interval variables*). Για παράδειγμα, η μεταβλητή της αρτηριακής πίεσης, είναι μια συνεχής μεταβλητή, η οποία μπορεί να κατηγοριοποιηθεί ομαδοποιώντας τις τιμές της σε συγκεκριμένα διαστήματα τιμών, γνωστών αποστάσεων. Ο τρόπος με τον οποίο μια μεταβλητή μετριέται, καθορίζει και το είδος της. Για παράδειγμα, η μεταβλητή «Εκπαίδευση», είναι ονοματική όταν μετριέται με βάση το είδος της εκπαίδευσης, δημόσιο ή ιδιωτικό σχολείο, είναι διατακτική, όταν μετριέται με βάση το επίπεδο της εκπαίδευσης, Βασική εκπαίδευση, Λύκειο, Πανεπιστήμιο, Μεταπτυχιακό και τέλος είναι διαστηματική, όταν μετριέται με βάση τα χρόνια εκπαίδευσης, $1, 2, 3, \ldots$ Η κλίμακα μέτρησης μιας μεταβλητής καθορίζει και το είδος της στατιστικής μεθόδου που θα χρησιμοποιηθεί.

Διάφοροι μέθοδοι ανάλυσης διατακτικών μεταβλητών αξιοποιούν την πληροφορία που παρέχει η διάταξη των κατηγοριών. Ιεραρχικά, με την έννοια της πληροφορίας που παρέχουν,

πρώτες είναι οι διαστηματικές μεταβλητές, ακολουθούν οι διατακτικές και τέλος οι ονοματικές. Στατιστικές μέθοδοι για την ανάλυση μιας μεταβλητής ενός τύπου, μπορούν να χρησιμοποιηθούν και για την ανάλυση μεταβλητών υψηλότερης ιεραρχίας, αλλά ποτέ χαμηλότερης. Για παράδειγμα, στατιστικές μέθοδοι για την ανάλυση ονοματικών μεταβλητών μπορούν να χρησιμοποιηθούν και για διατακτικές, αγνοώντας την διάταξη. Οι ονοματικές μεταβλητές είναι ποιοτικές μεταβλητές (*qualitative*), οι κατηγορίες διαφέρουν μεταξύ τους ποιοτικά και όχι ποσοτικά, ενώ οι διαστηματικές μεταβλητές είναι ποσοτικές (*quantitative*), τα χαρακτηριστικά των επιπέδων των κατηγοριών διαφέρουν ποσοτικά. Ο χαρακτηρισμός των διατακτικών μεταβλητών σε ποιοτικές ή ποσοτικές, δεν είναι εύκολος. Οι αναλυτές συχνά τις χειρίζονται ως ποιοτικές, χρησιμοποιώντας μεθόδους ονοματικών μεταβλητών, αλλά μπορούμε να πούμε ότι μοιάζουν περισσότερο με διαστηματικές μεταβλητές παρά με ονοματικές. Αν και δεν μπορούν να μετρηθούν, κρύβουν μέσα τους μια συνεχή μεταβλητή. Για παράδειγμα, η ταξινόμηση των πολιτικών πεποιθήσεων (φιλελεύθερες, μετριοπαθείς ή συντηρητικές), μετρά αδέξια μια έμφυτη συνεχή μεταβλητή. Οι αναλυτές συχνά αξιοποιούν την ποσοτική φύση των διατακτικών μεταβλητών, αναθέτοντας σκορ στις διάφορες κατηγορίες ή υποθέτοντας ότι ακολουθούν συνεχή κατανομή [*Agresti*, (2002)].

2.7 Κατανομές κατηγορικών δεδομένων

Η συμπερασματική ανάλυση, απαιτεί υποθέσεις αναφορικά με τον τυχαίο μηχανισμό που παράγει τα δεδομένα. Στα μοντέλα παλινδρόμησης με συνεχείς μεταβλητές απόκρισης, η κανονική κατανομή έχει τον κυριότερο ρόλο. Οι κυριότερες κατανομές που περιγράφουν κατηγορικές αποκρίσεις είναι η διωνυμική, η πολυωνυμική και η *Poisson*.

2.7.1 Διωνυμική Κατανομή

Πολλές εφαρμογές αναφέρονται σε έναν γνωστό αριθμό n δίτιμων παρατηρήσεων. Έστω $y_1, y_2, ..., y_n$, οι αποκρίσεις για n ανεξάρτητες και πανομοιότυπες δοκιμές, έτσι ώστε $P(Y_i = 1) = \pi$ και $P(Y_i = 0) = 1 - \pi$, όπου $1 = επιτυχία$ και $0 = αποτυχία$.

Σημειώνουμε ότι με τον όρο «*πανομοιότυπες*» δοκιμές εννοούμε ότι η πιθανότητα της επιτυχίας π, είναι ίδια για κάθε δοκιμή και με τον όρο «*ανεξάρτητες*» δοκιμές εννοούμε ότι οι αποκρίσεις $\{Y_i\}$, είναι ανεξάρτητες, τυχαίες μεταβλητές. Οι δοκιμές αυτές είναι γνωστές και ως δοκιμές *Bernoulli*. Ο συνολικός αριθμός των επιτυχιών, $Y = \sum_{i=1}^{n} Y_i$, ακολουθεί την διωνυμική κατανομή παραμέτρου π, δηλαδή $Y \sim Bin(n, \pi)$. Η συνάρτηση πυκνότητας πιθανότητας για την πιθανή τιμή y της μεταβλητής Y, είναι

$$P(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \text{ για } y = 0,1,2,...,n \qquad (2.4)$$

όπου $\binom{n}{y} = \frac{n!}{y!(n-y)!}$.

Δεν υπάρχει εγγύηση ότι οι διαδοχικές διωνυμικές παρατηρήσεις είναι πάντα ανεξάρτητες και πανομοιότυπες. Έτσι κάποιες φορές χρησιμοποιούνται άλλες κατανομές. Μια τέτοια περίπτωση είναι όταν παίρνουμε διωνυμικό δείγμα από έναν γνωστό πληθυσμό χωρίς επανάθεση, για παράδειγμα, όταν παρατηρούμε το φύλλο μιας τάξης μαθητών, παίρνοντας δείγμα 10 μαθητών από μια τάξη με σύνολο 20 μαθητών. Στην περίπτωση αυτή η υπεργεωμετρική κατανομή είναι καταλληλότερη.

2.7.2 Πολυωνυμική Κατανομή

Κάποιες δοκιμές έχουν περισσότερα από δυο πιθανά αποτελέσματα. Έστω ότι κάθε μια από τις n ανεξάρτητες και πανομοιότυπες δοκιμές έχει αποτέλεσμα που ανήκει σε κάποια από τις $c-$κατηγορίες. Έστω, $y_{ij} = 1$, αν η $i-$δοκιμή έχει το j αποτέλεσμα και $y_{ij} = 0$, διαφορετικά. Τότε το αποτέλεσμα $y_i = (y_{i1}, y_{i2}, ..., y_{ic})$, αναπαριστά μια πολυωνυμική δοκιμή με $\sum_{j=1}^{c} y_{ij} = 1$. Για παράδειγμα, το αποτέλεσμα $(0,0,1,0)$, σημαίνει ότι ανάμεσα σε 4 διαδοχικές κατηγορίες, είχαμε αποτέλεσμα που ανήκει στην $3^{η}$ κατηγορία. Αξίζει να σημειώσουμε ότι το αποτέλεσμα y_{ic} είναι γραμμικώς εξαρτημένο από τα υπόλοιπα και επομένως η παρατήρησή του δεν είναι

απαραίτητη. Έστω ότι $n_j = \sum_i y_{ij}$, συμβολίζει τον αριθμό των δοκιμών, που φέρουν το $j-$αποτέλεσμα. Τότε οι μετρήσεις $(n_1, n_2, ..., n_c)$ ακολουθούν την πολυωνυμική κατανομή. Έστω, $\pi_j = P(Y_{ij} = 1)$ συμβολίζει την πιθανότητα να έχουμε αποτέλεσμα στην $j-$κατηγορία, για κάθε δοκιμή. Η πολυωνυμική συνάρτηση πυκνότητας πιθανότητας είναι

$$P(n_1, n_2, ..., n_{c-1}) = \left(\frac{n!}{n_1! n_2! ... n_c!} \right) \pi_1^{n_1} \pi_2^{n_2} ... \pi_c^{n_c} \qquad (2.5)$$

Καθώς $\sum_j n_j = n$, έχουμε $(c-1)-$διαστάσεις, με $n_c = n - (n_1 + n_2 + ... + n_{c-1})$. Η διωνυμική κατανομή είναι ειδική περίπτωση για $c = 2$.

2.7.3 Κατανομή Poisson

Μερικές φορές οι μετρήσεις δεν είναι αποτέλεσμα ενός καθορισμένου αριθμού δοκιμών. Για παράδειγμα, εάν y είναι ο αριθμός των θανάτων σε αυτοκινητιστικά δυστυχήματα κατά την διάρκεια της επόμενης εβδομάδος, δεν υπάρχει καθορισμένο άνω όριο n για το y. Καθώς y ένας μη αρνητικός αριθμός, η κατανομή του θα πρέπει να έχει τον όγκο της σε αυτό το εύρος. Μια τέτοια κατανομή είναι η *Poisson*. Οι πιθανότητες βασίζονται σε μια μόνο παράμετρο, τον μέσο μ. Η συνάρτηση πυκνότητας πιθανότητας της *Poisson* είναι

$$P(y) = \frac{e^{-\mu} \mu^y}{y!} \text{ με } y = 0, 1, 2, ... \qquad (2.6)$$

Όσο το μ αυξάνει, η κατανομή *Poisson* συγκλίνει στην κανονική κατανομή. Η κατανομή *Poisson* χρησιμοποιείται για να μετρήσει γεγονότα που συμβαίνουν τυχαία στον χρόνο ή τον χώρο, όταν τα αποτελέσματα σε περιόδους ή περιοχές χωρίς συνοχή, είναι ανεξάρτητα. Επίσης εφαρμόζεται σαν μια προσέγγιση της διωνυμικής κατανομής, όταν το n είναι μεγάλο και το π είναι μικρό, με $\mu = n\pi$. Άρα, όταν κάθε ένας οδηγός μιας χώρας με πληθυσμό $n = 50.000.000$, είναι ανεξάρτητες δοκιμές, με πιθανότητα θανάτου σε αυτοκινητιστικό δυστύχημα την επόμενη εβδομάδα $\pi = 0.000002$, τότε ο αριθμός των θανάτων Y ακολουθεί την διωνυμική κατανομή

$Y \sim Bin(50.000.000, 0.000002)$ ή την κατά προσέγγιση *Poisson* κατανομή, $Y \sim Poisson(100)$, καθώς $\mu = 50.000.000 \times 0.000002 = 100$.

Ένα κύριο χαρακτηριστικό της κατανομής *Poisson*, είναι ότι η διακύμανση ισούται με τον μέσο. Οι δειγματικές μετρήσεις ποικίλλουν περισσότερο όταν ο μέσος έχει μεγαλύτερη τιμή. Έτσι, όταν ο μέσος όρος των εβδομαδιαίων ατυχημάτων είναι 100, έχουμε μεγαλύτερη διακύμανση από ότι όταν είναι 10.

2.7.4 Πολυωνυμική, Διωνυμική και Poisson Δειγματοληψία

Οι κατανομές πιθανότητας που περιγράψαμε προηγουμένως, επεκτείνονται και στην περιγραφή των κελιών ενός πίνακα συνάφειας. Για παράδειγμα, ένα *Poisson* μοντέλο δειγματοληψίας, θεωρεί τις μετρήσεις $\{Y_{ij}\}$, ως ανεξάρτητες *Poisson* τυχαίες μεταβλητές παραμέτρου $\{\mu_{ij}\}$. Η από κοινού συνάρτηση πυκνότητας πιθανότητας των πιθανών αποτελεσμάτων $\{n_{ij}\}$, ισούται με το γινόμενο των πιθανοτήτων $\{P(Y_{ij} = n_{ij})\}$, για τα κελιά $\{IJ\}$, οι οποίες ακολουθούν την κατανομή *Poisson*, δηλαδή

$$f(x, y) = \prod_i \prod_j \exp(-\mu_{ij}) \mu_{ij}^{n_{ij}} / n_{ij}! \qquad (2.7)$$

Όταν το μέγεθος του δείγματος n είναι γνωστό, αλλά τα αθροίσματα των γραμμών και των στηλών όχι, τότε ένα πολυωνυμικό δειγματοληπτικό σχέδιο εφαρμόζεται. Τα $\{IJ\}$ κελιά είναι τώρα τα πιθανά αποτελέσματα. Η συνάρτηση πυκνότητας πιθανότητας των πιθανών αποτελεσμάτων ισούται με

$$\frac{n!}{n_{11}!...n_{ij}!} \prod_i \prod_j \pi_{ij}^{n_{ij}} \qquad (2.8)$$

Συχνά οι παρατηρήσεις της μεταβλητής απόκρισης Y, συμβαίνουν ξεχωριστά σε κάθε επίπεδο της επεξηγηματικής μεταβλητής X. Στην περίπτωση αυτή, θεωρούμε ότι τα αθροίσματα των γραμμών $(n_{i.})$ είναι γνωστά. Έστω ότι οι $\{n_i\}$ παρατηρήσεις της Y, για κάθε κατηγορία i της

X είναι ανεξάρτητες, με κατανομή πιθανότητας $\left\{\pi_{1|i} = \pi_{2|i} = ... = \pi_{J|i}\right\}$. Οι μετρήσεις $\left\{n_{i.}\right\}$, για $j = 1, 2, ..., J$, ικανοποιούν την σχέση $\sum_j n_{ij} = n_{i.}$ και έχουν την πολυωνυμική μορφή

$$\frac{n_{i.}!}{\prod_j n_{ij}!} \prod_j \pi_{j|i}^{n_{ij}} \tag{2.9}$$

Όταν τα δείγματα στα διάφορα επίπεδα της μεταβλητής X είναι ανεξάρτητα, το δειγματοληπτικό σχέδιο ονομάζεται γινόμενο πολυωνυμικής δειγματοληψίας (*product multinomial sampling*), καθώς η από κοινού συνάρτηση πιθανότητας των δεδομένων, είναι το γινόμενο πολυωνυμικών συναρτήσεων, για κάθε επίπεδο της X. Τέλος, όταν τα αθροίσματα των γραμμών και των στηλών είναι γνωστά, τότε η καταλληλότερη δειγματοληπτική κατανομή είναι η υπεργεωμετρική.

2.7.5 Παραδείγματα

Ερευνητές σχεδιάζουν να μελετήσουν την σχέση μεταξύ της χρήσης ζώνης ασφαλείας και θανητοφόρου ή όχι ατυχήματος. Στην περίπτωση που αποφασίσουν να καταγράψουν όλα τα ατυχήματα που θα συμβούν εντός ενός έτους, τότε το συνολικό μέγεθος δείγματος είναι μια τυχαία μεταβλητή. Τότε θα θεωρήσουν ότι ο αριθμός των παρατηρήσεων μεταξύ των τεσσάρων συνδυασμών, χρήση ζώνης και αποτέλεσμα ατυχήματος, ως ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την κατανομή Poisson με άγνωστη μέση τιμή $\left\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\right\}$.

Αντιθέτως, αν οι ερευνητές αποφασίσουν να πάρουν από το αρχείο της αστυνομίας, τυχαίο δείγμα 200 ατυχημάτων που συνέβησαν το προηγούμενο έτος και ταξινομήσουν κάθε ατύχημα σύμφωνα με τις κατηγορίες, «Χρήση ζώνης» (Ναι / Όχι), για την μεταβλητή X γραμμή και «Ατύχημα» (Θανατηφόρο / μη – θανητηφόρο), για την μεταβλητή Y στήλη, τότε το μέγεθος του δείγματος είναι γνωστό. Για την μελέτη αυτή, θα θεωρήσουν ότι τα 4 κελιά του πίνακα είναι πολυωνυμικές τυχαίες μεταβλητές, με $n = 200$ δοκιμές και άγνωστη παράμετρο την από κοινού πιθανότητα $\left\{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\right\}$.

Στην περίπτωση που τα θανητοφόρα ατυχήματα ήταν καταγεγραμμένα σε ξεχωριστά αρχεία της αστυνομίας, τότε οι ερευνητές θα έπρεπε να πάρουν τυχαίο δείγμα 100 ατυχημάτων από το αρχείο των θανατηφόρων γεγονότων και τυχαίο δείγμα 100 ατυχημάτων από το αρχείο των μη – θανητηφόρων γεγονότων. Αυτή η προσέγγιση θεωρεί το άθροισμα των στηλών γνωστό.

Στην περίπτωση αυτή κάθε στήλη του πίνακα, θα πρέπει να θεωρηθεί ως ένα ανεξάρτητο διωνυμικό δείγμα, αφού πλέον τα αποτελέσματα είναι χρήση ή μη χρήση ζώνης ασφαλείας. Τέλος η παραδοσιακή προσέγγιση ενός πειραματικού σχεδιασμού είναι να τυχαιοποιήσουμε 200 υποκείμενα, 100 να επιλεγούν να κάνουν χρήση ζώνης και 100 να μην κάνουν χρήση ζώνης και να τα υποβάλλουμε να υποστούν ατύχημα. Τα αποτελέσματα τότε θα είναι ανεξάρτητα διωνυμικά δείγματα σε κάθε μια γραμμή του πίνακα, με γνωστό άθροισμα κάθε γραμμής 100. Είναι προφανές ότι υπάρχουν ηθικοί λόγοι για την αποφυγή τέτοιου είδους σχεδιασμών, ειδικά στους ανθρώπους. Κάτι τέτοιο, είναι εντονότερο κυρίως σε ιατρικές μελέτες.

ΚΕΦΑΛΑΙΟ 3

ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΚΑΙ ΒΑΣΙΚΕΣ ΙΔΙΟΤΗΤΕΣ ΤΩΝ ΜΕΤΡΩΝ

3.1 Εισαγωγή

Πως μπορούμε να μετρήσουμε την συνάφεια μεταξύ δυο κατηγορικών μεταβλητών, όπως η θρησκεία, το επάγγελμα, ή η προτίμηση μεταξύ διαφορετικών επιλογών; Με την χρησιμοποίηση ενός μέτρου συνάφειας, δηλαδή μιας στατιστικής συνάρτησης που συνοψίζει τον βαθμό της σχέσης μεταξύ δυο μεταβλητών. Σε μια εκτεταμένη ανασκόπηση της βιβλιογραφίας, οι *Goodman & Kruskal* (1979), βρήκαν πάρα πολλά μέτρα για τον σκοπό αυτό. Εάν ο λόγος, για τον οποίο χρειαζόμαστε ένα μέτρο συνάφειας είναι γνωστός, τότε είναι πιο εύκολη η επιλογή του. Για παράδειγμα, μια εταιρία κατασκευής τηλεοράσεων ενδιαφέρεται να διαφημίσει ένα νέο μοντέλο μέσω εφημερίδας. Ο πίνακας συνάφειας διασταυρώνοντας τις μεταβλητές, «αγαπημένη εφημερίδα» και «χρήση τηλεόρασης», μπορεί να μας δώσει την πληροφορία: ποια εφημερίδα διαβάζουν περισσότερο όσοι έχουν τηλεόραση. Ένα λογικό μέτρο συνάφειας, θα ήταν απλά η αναλογία των ατόμων στον πληθυσμό που έχουν τηλεόραση και διαβάζουν την συγκεκριμένη εφημερίδα. Όμως, είναι σπάνιες οι περιπτώσεις για τις οποίες ο σκοπός μιας έρευνας μπορεί να ορισθεί. Συνήθως μια έρευνα είναι επεξηγηματική και έχει πολλούς στόχους. Μερικές φορές χρειαζόμαστε ένα μέτρο συνάφειας απλά για να συνοψίσουμε ένα μεγάλο σύνολο δεδομένων.

3.2 Κριτήρια επιλογής μέτρων συνάφειας

Καθώς τα διάφορα μέτρα συνάφειας, που μπορούν να υιοθετηθούν για την ανάλυση ενός πίνακα συνάφειας, δεν στηρίζονται στα ίδια κριτήρια για τον υπολογισμό της έντασης της σχέσης μεταξύ των δυο μεταβλητών, τότε εάν δυο ή περισσότερα μέτρα εφαρμοστούν στο ίδιο σύνολο δεδομένων, ίσως να μην αποδώσουν συγκρίσιμους συντελεστές συνάφειας. Αν και στην ανάλυσή μας, θα εξηγήσουμε τους παράγοντες που πρέπει να λαμβάνονται υπόψη, αναφορικά με το πιο από τα διάφορα μέτρα θα πρέπει να χρησιμοποιήσουμε, στην πλειοψηφία των

περιπτώσεων, ένα μέτρο δεν είναι απαραίτητα ανώτερο κάποιου άλλου, με την έννοια της πληροφόρησης που παρέχει για έναν πίνακα συνάφειας [*Goodman & Kruskal*, (1954)].

3.2.1 Είδος Δεδομένων (Συνέχεια)

Η υπόθεση που κάνουμε για την συνέχεια ή όχι των μεταβλητών καθορίζει και το είδος του μέτρου που θα επιλέξουμε. Όπως έχουμε αναφέρει, το θέμα αυτό απασχόλησε από πολύ νωρίς, τους *Pearson* και *Yule*, δημιουργώντας μια ιστορική αντιπαράθεση. Αν μια μεταβλητή προέρχεται από συνεχή δεδομένα, τότε θα μπορούσαμε να υποθέσουμε ότι ο πληθυσμός ακολουθεί συγκεκριμένου είδους κατανομή, την από κοινού πολυμεταβλητή κανονική κατανομή. Σε αυτήν την περίπτωση, είναι λογικό να υιοθετήσουμε μέτρα που βασίζονται στον συντελεστή συσχέτισης (*Person correlation coefficient*) r [*Goodman & Kruskal*, (1954)].

3.2.2 Διάσταση του Πίνακα

Η διάσταση του πίνακα συνάφειας παίζει επίσης σημαντικό ρόλο στην επιλογή ενός μέτρου συνάφειας καθώς κάποια μέτρα είναι σχεδιασμένα μόνο για 2×2 πίνακες, ενώ άλλα σχεδιασμένα μόνο για τετραγωνικούς $I\times I$ πίνακες.

3.2.3 Είδος μεταβλητών (Διάταξη)

Το είδος των κατηγορικών μεταβλητών που εξετάζουμε, παίζει καθοριστικό ρόλο στην επιλογή ενός κατάλληλου μέτρου. Το ενδεχόμενο ύπαρξης ή όχι μιας φυσικής διάταξης των μεταβλητών, μας οδηγεί στην επιλογή διαφορετικού μέτρου, καθώς κάποια μέτρα έχουν κατασκευαστεί για να αξιοποιούν την επιπρόσθετη πληροφορία της διάταξης, παρέχοντας έτσι και την κατεύθυνση της συνάφειας. Γενικά, οι 6 δυνατοί συνδυασμοί μεταβλητών που αντιμετωπίζουν συνήθως οι ερευνητές είναι: Συνεχής - Συνεχής, Συνεχής - Διατακτική, Συνεχής - Ονοματική, Διατακτική - Διατακτική, Διατακτική - Ονοματική, Ονοματική - Ονοματική.

3.2.4 Συμμετρικότητα

Κάποιος θα μπορούσε να εξετάσει δυο μεταβλητές συμμετρικά. Ένα μέτρο που ικανοποιεί την ιδιότητα της συμμετρίας, αποδίδει το ίδιο αποτέλεσμα, ανεπηρέαστο από το ποια είναι η εξαρτημένη μεταβλητή. Αν όμως, υπάρχει μια εκ των προτέρων γνωστή αιτιώδης σχέση μεταξύ των μεταβλητών, η οποία προσδιορίζεται προς μια κατεύθυνση (*και όχι και στις δυο*) ή όταν η μια μεταβλητή μπορεί να προβλεφθεί με βάση την πληροφορία που έχουμε για την άλλη, τότε το συμπέρασμα θα είναι ασύμμετρο [*Goodman & Kruskal*, (1954)]. Στην περίπτωση αυτή, η πρόβλεψη είναι συχνά ο στόχος της μελέτης και συγκεκριμένα μέτρα συνάφειας προτείνονται. Αν ένα μέτρο χειρίζεται δυο μεταβλητές ασύμμετρα, τότε η τιμή του μέτρου θα μεταβληθεί, αν ο πίνακας αντιμεταθεί. Αξίζει να σημειώσουμε ότι η έννοια της συμμετρίας ως ιδιότητα, διαφοροποιείται από την έννοια των μέτρων συμμετρίας – ασυμμετρίας που θα συναντήσουμε στο Κεφάλαιο 6. Η διαφορά έγκειται στο ότι τα μέτρα συμμετρίας αναφέρονται στην σύμμετρη ταξινόμηση των παρατηρήσεων ως προς την κύρια διαγώνιο ενός τετραγωνικού $I \times I$ πίνακα συνάφειας.

3.3 Βασικές ιδιότητες ενός μέτρου

Κάθε μέτρο συνάφειας που δημιουργείται έχει κάποια συγκεκριμένα χαρακτηριστικά. Όπως θα διαπιστώσουμε, όταν τα μέτρα αυτά εφαρμοστούν στο ίδιο σύνολο δεδομένων, τα αποτελέσματα που προκύπτουν συχνά διαφοροποιούνται. Για λόγους ευκολίας και σύγκρισης, τα μέτρα συνάφειας θα πρέπει ικανοποιούν κάποιες κοινές παραδοχές. Όπως οι *Goodman & Kruskal* (1954) αναφέρουν, ένα μέτρο συνάφειας θα πρέπει να πληροί τις εξής παραδοχές:

1. Να κυμαίνεται στο διάστημα $[-1,+1]$, όταν τα δεδομένα έχουν μια φυσική διάταξη, με τιμές ± 1 στην περίπτωση της πλήρης συνάφειας, όπου το πρόσημο δείχνει την κατεύθυνση της σχέσης και με τιμή 0 στην περίπτωση της ανεξαρτησίας
2. Να κυμαίνεται στο διάστημα $[0,1]$ όταν δεν υπάρχει κάποια φυσική διάταξη, με τιμή 1 στην περίπτωση της πλήρους συνάφειας και τιμή 0 στην περίπτωση της ανεξαρτησίας.

Τα περισσότερα μέτρα που παρουσιάζονται στην παρούσα εργασία, έχουν μια λειτουργική καθιέρωση, με την έννοια ότι έχουν παρουσιασθεί με βάση τις ιδιότητές τους. Βασίζονται περισσότερο στις συγκεκριμένες ανάγκες και σκοπούς για τους οποίους έχουν κατασκευαστεί, πληρώντας κάποιες βασικές παραδοχές, παρά σε κάποια αξιωματική θεωρία στα πλαίσια ενός συστηματοποιημένου συνόλου αξιωμάτων.

3.4 Τα αξιώματα του Renyi

Σε κάθε πεδίο εφαρμογών της στατιστικής, ένα σύνηθες πρόβλημα που συχνά αντιμετωπίζουν οι ερευνητές, είναι να καταφέρουν να αποδώσουν την ένταση της εξάρτησης, μεταξύ δυο μεταβλητών με μια αριθμητική τιμή. Φυσικά μια τέτοια τιμή εξυπηρετεί μόνο στην σύγκριση και επομένως το εύρος της είναι αυθαίρετο. Όπως ο *Renyi* (1959) αναφέρει, είναι λογικό να επιλέξουμε το διάστημα $[0,1]$ και να αντιστοιχήσουμε την τιμή 1 σε αυστηρή εξάρτηση και την τιμή 0 σε πλήρης ανεξαρτησία. Με αυτές τις παραδοχές και στα πλαίσια της αναζήτησης ενός γενικότερου πλαισίου για την αξιολόγηση ενός μέτρου εξάρτησης, έστω $\delta(\xi,\eta)$, ο *Renyi* διατύπωσε τα ακόλουθα αξιώματα:

R1. Το μέτρο $\delta(\xi,\eta)$ ορίζεται για κάθε ζεύγος τυχαίων μεταβλητών ξ και η.

R2. $\delta(\xi,\eta)=\delta(\eta,\xi)$, δηλαδή το μέτρο είναι συμμετρικό.

R3. $0\le\delta(\xi,\eta)\le 1$

R4. $\delta(\xi,\eta)=0$ αν και μόνον αν οι μεταβλητές ξ και η είναι ανεξάρτητες

R5. $\delta(\xi,\eta)=1$ αν υπάρχει μια αυστηρή εξάρτηση μεταξύ των ξ και η, για παράδειγμα, είτε $\xi=g(\eta)$, ή $\eta=f(\xi)$, όπου $g(x)$ και $f(x)$ είναι μετρήσιμες κατά *Borel* συναρτήσεις (*Borel-measurable functions*).

R6. Αν οι μετρήσιμες κατά Borel συναρτήσεις $g(x)$ και $f(x)$, είναι αμφιμονοσήμαντες, δηλαδή ένα-προς-ένα, τότε $\delta(f(\xi),g(\eta))=\delta(\xi,\eta)$.

R7. Εάν η από κοινού κατανομή των ξ και η είναι κανονική, τότε $\delta(\xi,\eta)=|R(\xi,\eta)|$, όπου $R(\xi,\eta)$, είναι ο συντελεστής συσχέτισης των ξ και η.

Αξίζει να σημειώσουμε ότι τα αξιώματα αυτά, αφορούν κυρίως τα μέτρα πληροφορίας και απόκλισης ή απόστασης (*divergence type mesures*) μεταξύ δυο κατανομών πιθανότητας και προέρχονται από τον χώρο της στατιστικής πληροφορίας (*Information Theory*), όπως για

παράδειγμα, η εντροπία *Shannon* στην οποία βασίζεται ο συντελεστής αβεβαιότητας U του *Theil*. Παρόλαυτα, προσφέρουν ένα ικανοποιητικό πλαίσιο, για την αξιολόγηση των παραδοσιακών μέτρων συνάφειας (Κεφάλαια 4 και 5) και σίγουρα ένα σημείο αναφοράς για την καταλληλότητα των νεοσύστατων μέτρων συμμετρίας – ασυμμετρίας (Κεφάλαιο 6).

3.5 Παρατηρήσεις

Ένας σημαντικός περιοριστικός όρος στην ερμηνεία κάθε μέτρου συνάφειας που θα πρέπει να τονίσουμε, είναι ότι όταν υπάρχει συνάφεια μεταξύ δυο ή περισσοτέρων μεταβλητών, με όποιο τρόπο και αν μετριέται, δεν θα πρέπει ποτέ να ερμηνεύεται σαν μια σχέση αιτίας και αποτελέσματος (*cause-effect relationship*) μεταξύ των μεταβλητών. Αν και η ύπαρξη ισχυρής συσχέτισης είναι αναγκαία για να θεμελιώσει μια σχέση αιτίας - αποτελέσματος, δεν είναι και επαρκής. Συνήθως, τα άτομα έχουν την τάση να εξάγουν τέτοιου είδους συμπεράσματα. Ένα κλασσικό παράδειγμα, εμφανίστηκε όταν ένα κράτος στην Αμερική αύξησε τους μισθούς των δασκάλων στα σχολεία και αργότερα βρέθηκε υψηλή συσχέτιση μεταξύ της αύξησης των μισθών και της αύξησης στην κατανάλωση λικέρ στο κράτος αυτό. Οι κριτικοί υποστήριξαν ότι οι δάσκαλοι χρησιμοποιούσαν την αύξηση του μισθού τους για να αγοράζουν περισσότερο λικέρ και το γεγονός αυτό προκάλεσε την αύξηση στην κατανάλωση λικέρ. Όμως, θα ήταν το ίδιο λογικό να ισχυριστούμε ότι η αύξηση στην κατανάλωση λικέρ, προκάλεσε την αύξηση των μισθών των δασκάλων. Κανένα τέτοιο συμπέρασμα δεν δικαιολογείται από την ύπαρξη συσχέτισης.

Όταν οι μεταβλητές είναι ονοματικές ή διατακτικές, τότε ίσως να υπάρχουν ισοτιμίες (*ties*) μεταξύ των δεδομένων, δηλαδή πολλά υποκείμενα μπορεί να ταξινομούνται στο ίδιο επίπεδο, για τις κατηγορίες των μεταβλητών X και Y. Υπάρχουν αρκετές διαφορετικές διαδικασίες για να χειριστούμε τους δεσμούς και επομένως είναι δυνατό για το ίδιο σύνολο δεδομένων, να έχουμε συντελεστές με ελαφρώς διαφοροποιημένες τιμές.

Όταν υπολογίζουμε την ένταση της συνάφειας μεταξύ δυο μεταβλητών, είναι σημαντικό να ελέγξουμε για την ύπαρξη μεταβλητών σύγχυσης (*confounder variables*), οι οποίες μπορεί να έχουν επίδραση στα αποτελέσματα μας. Αυτό μπορεί να επιτευχθεί με την χρησιμοποίηση μερικών συντελεστών συσχέτισης (*partial correlation coefficients*).

3.6 Μέτρα πληροφορίας και απόστασης ή απόκλισης

Η έννοια της απόστασης είναι θεμελιακή στα μαθηματικά και την στατιστική. Στην ανάλυση δεδομένων και την εφαρμοσμένη έρευνα παίζει καθοριστικό ρόλο για την μελέτη της ομοιότητας – ανομοιότητας μεταξύ φυσικών αντικειμένων και πληθυσμών. Από την άλλη πλευρά, οι έννοιες της πληροφορίας, εντροπίας και απόκλισης εμφανίζονται συχνά στο πεδίο των πιθανοτήτων και της στατιστικής και παίζουν σπουδαίο ρόλο στην θεωρία επικοινωνιών και πληροφοριών. Χρησιμοποιούνται ως μέτρα της απόστασης ή απόκλισης, της ομοιότητας – ανομοιότητας μεταξύ δυο κατανομών πιθανότητας, με την έννοια ότι όσο μικρότερη είναι η τιμή του μέτρου, τόσο πιο δύσκολη είναι η διάκριση και ο διαχωρισμός των κατανομών στις οποίες αναφέρεται [*Karagrigoriou & Papaioannou*, (2008)]. Στην παρούσα εργασία τα μέτρα απόστασης ή απόκλισης (*divergence type measures*) καθώς και η έννοια της εντροπίας (*entropy*), χρησιμοποιούνται ως βάση για την κατασκευή των σχετικά πρόσφατων, νεοσύστατων μέτρων συμμετρίας – ασυμμετρίας για τετραγωνικούς πίνακες συνάφειας. Αξίζει να σημειώσουμε, ότι η έννοια της εντροπίας, έχει ήδη χρησιμοποιηθεί για την κατασκευή του συντελεστή αβεβαιότητας U του *Theil* (1972), ως μέτρο συνάφειας για ονοματικές μεταβλητές.

3.6.1 Σύντομη ιστορική αναδρομή των μέτρων πληροφορίας

Τα μέτρα πληροφορίας έχουν μια μακρά ιστορία που ξεκινά από τις εργασίες των *Fisher* (1925), *Shannon* (1948) και *Kullback - Leibler* (1951). Υπάρχουν αρκετές φάσεις στην ιστορία της θεωρίας της πληροφορίας. Αρχικά έχουμε

(i) την ανάπτυξη των γενικευμένων μέτρων πληροφορίας και απόκλισης, όπως την f - *divergence*, την $(h-f)$ - *divergence*, την υποεντροπία (*hypoentropy*) κ.α.

(ii) την συλλογή και σύνθεση των ιδιοτήτων που θα πρέπει να ικανοποιούν και

(iii) την προσπάθεια ενοποίησης τους.

Όλες αυτές οι εργασίες αναφέρονται σε πληθυσμούς και κατανομές. Αργότερα, έχουμε την εμφάνιση ελεγχοσυναρτήσεων (*statistics*) πληροφορίας ή απόκλισης, που βασίζονται σε δεδομένα ή δείγματα και την χρησιμοποίησή τους στην στατιστική συμπερασματολογία, κυρίως για την εκτίμηση ελαχίστων αποστάσεων και για την ανάπτυξη ασυμπτωματικών ελέγχων καλής

προσαρμογής. Πρόσφατα έχουμε την αναβίωση του ενδιαφέροντος σε μέτρα πληροφορίας και απόκλισης τα οποία χρησιμοποιούνται σε πολλές εφαρμογές. Η ανάπτυξη της θεωρίας της στατιστικής πληροφορίας (*Information Theory*) έχει κάνει αρκετή πρόοδο, αλλά δεν έχει ακόμη αποκτήσει ευρεία αποδοχή και εφαρμογή [*Karagrigoriou & Papaioannou*, (2008)]. Υπάρχουν πολλές εργασίες που συζητούν τα παραπάνω θέματα. Ενδεικτικά αναφέρουμε, *Kendall* (1973), *Csiszar* (1977), *Kapur* (1984), *Aczel* (1986), *Papaioannou* (1981, 2010), *Zografos et al.* (1998, 2000), *Ferentinos & Papaioannou* (1979, 1982), *Soofi* (1994, 2000).

3.6.2 Κλάσεις μέτρων πληροφορίας και απόστασης

Γενικά, υπάρχουν 3 κλάσεις μέτρων πληροφορίας και απόκλισης: τα μέτρα τύπου *Fisher*, τα μέτρα τύπου απόκλισης (*divergence*) και τα μέτρα τύπου εντροπίας (*entropy*). Μερικά από αυτά έχουν αναπτυχθεί αξιωματικά, όπως η εντροπία κατά *Shannon* και οι γενικεύσεις της, για τα περισσότερα όμως η προσέγγιση είναι περισσότερο λειτουργική, με την έννοια ότι έχουν παρουσιασθεί με βάση τις ιδιότητές τους. Στην συνέχεια θεωρούμε ότι, $f(x,\theta)$ είναι μια συνάρτηση πυκνότητας πιθανότητας μια τυχαίας μεταβλητής X η οποία εξαρτάται από την παράμετρο θ.

Α. Μέτρα τύπου *Fisher*

Τα μέτρα τύπου *Fisher* είναι παραμετρικά μέτρα πληροφορίας και μετρούν το ποσό της πληροφορίας που πηγάζει από τα δεδομένα σχετικά με μια άγνωστη παράμετρο θ και είναι συναρτήσεις του θ. Στην περίπτωση αυτή, το πιο γνωστό μέτρο είναι το μέτρο πληροφορίας *Fisher* (1925), το οποίο ορίζεται ως

$$\textit{Fisher (1925)}: I_X^F(\theta) = \begin{cases} E_\theta\left[\dfrac{\partial}{\partial\theta}\log f(X,\theta)\right]^2 = -E_\theta\left[\dfrac{\partial^2}{\partial\theta^2}\log f(X,\theta)\right]^2, & \theta\ \textit{μονομεταβλητό} \\ \left\| E_\theta\left[\dfrac{\partial}{\partial\theta_i}\log f(X,\theta)\dfrac{\partial}{\partial\theta_j}\log f(X,\theta)\right]\right\|_{k\times k}, & \theta\ k-\textit{μεταβλητό} \end{cases} \tag{3.1}$$

Όπου $\|\ \|_{k\times k}$ συμβολίζει έναν $k\times k$ πίνακα. Αν θ $k-$μεταβλητό, ο πίνακας πληροφορίας του *Fisher* είναι το μόνο διαθέσιμο παραμετρικό μέτρο πληροφορίας.

Ο *Vajda* (1973) επέκτεινε τον παραπάνω ορισμό, στην περίπτωση που θ μονομεταβλητό, υψώνοντας τον βαθμό της συνάρτησης στην δύναμη a, όπου $a \geq 1$

$$\textit{Vajda}\ (1973):\ I_X^V(\theta) = E_\theta \left| \frac{\partial}{\partial \theta} \log f(X,\theta) \right|^\alpha,\ \alpha \geq 1 \tag{3.2}$$

Στην περίπτωση που η παράμετρος θ είναι διάνυσμα, οι *Ferentinos & Papaioannou* (1981) πρότειναν ως μέτρο πληροφορίας, οποιαδήποτε ιδιοτιμή ή ειδικές συναρτήσεις των ιδιοτιμών του πίνακα πληροφορίας *Fisher*, όπως το ίχνος (*trace*) ή το *determinant*.

Έστω, $\lambda_i(\theta)$ $i = 1, 2, ..., k$ οι ιδιοτιμές του πίνακα πληροφορίας *Fisher*, τότε

$$\textit{Ferentinos \& Papaioannou}\ (1981):\ I_X^{FP}(\theta) = tr\left(I_X^F(\theta)\right) = \sum_{i=1}^{k} \lambda_i(\theta) \tag{3.3}$$

$$\textit{Ferentinos \& Papaioannou}\ (1981):\ I_X^{FP}(\theta) = \det\left(I_X^F(\theta)\right) = \prod_{i=1}^{k} \lambda_i(\theta) \tag{3.4}$$

Οι *Tukey* (1965) και οι *Chandrasekar* και *Balakrishnan* (2002), εξέτασαν το ακόλουθο μέτρο πληροφορίας

$$I_X^{TB}(\theta) = \begin{cases} \dfrac{(\partial\mu/\partial\theta)^2}{\sigma^2}, & \text{X} - \textit{univariate} \sim f(x,\theta) \\ (\partial\mu/\partial\theta)' \sum^{-1} (\partial\mu/\partial\theta), & \text{X} - \textit{vector} \end{cases} \tag{3.5}$$

Άλλα παραμετρικά μέτρα πληροφορίας που συναντάμε στην βιβλιογραφία είναι

$$\textit{Mathai}\ (1967):\ I_X^{Mat}(\theta) = \left[E_\theta \left| \frac{\partial}{\partial \theta} \log f(X,\theta) \right|^\alpha \right]^{1/a},\ \alpha \geq 1 \tag{3.6}$$

$$\textit{Boekee}\ (1977):\ I_X^{Bo}(\theta) = \left[E_\theta \left| \frac{\partial}{\partial \theta} \log f(X,\theta) \right|^{\frac{s}{s-1}} \right]^{s-1},\ 1 < s < \infty \tag{3.7}$$

Β.1 Μέτρα τύπου απόκλισης (*divergence*)

Τα μέτρα τύπου *divergence* είναι μη – παραμετρικά μέτρα και εκφράζουν το ποσό της πληροφορίας που πηγάζει από τα δεδομένα, διακρίνοντας την κατανομή f_1 έναντι της f_2, ή μετρούν την απόσταση ή την συγγένεια μεταξύ των f_1 και f_2. Έστω f_1 και f_2 δυο συναρτήσεις

πυκνότητας πιθανότητας, οι οποίες ίσως να εξαρτώνται από μια άγνωστη παράμετρο, ορισμένης και πεπερασμένης διάστασης. Το πιο γνωστό μέτρο τύπου απόκλισης (*divergence*) είναι η *Kullback-Leibler divergence.*

$$\textit{Kullback-Leibler}\ (1951):\ I_X^{KL}(f_1, f_2) = \int f_1 \log(f_1/f_2) d\mu \tag{3.8}$$

Σημειώνουμε ότι αν η f_1 είναι η πυκνότητα της $X = (U,V)$ και f_2 είναι το γινόμενο των περιθώριων πυκνοτήτων των U και V, τότε το μέτρο I_X^{KL} είναι η πολύ γνωστή αμοιβαία πληροφορία (*mutual information*).

Οι προσθετικές (*additive*) και μη προσθετικές (*non-additive*) άμεσες αποκλίσεις τάξης a, παρουσιάστηκαν στην δεκατία 1960 και 1970 [βλέπε *Renyi* (1961), *Csiszar* (1963), *Rathie-Kannappan*, (1972)]. Το πολύ γνωστό μέτρο πληροφορίας τάξης a, του *Renyi* ισούται με

$$\textit{Renyi}\ (1961):\ I_X^R(f_1, f_2) = \frac{1}{a-1}\int f_1^a - f_2^{1-a} d\mu,\ \alpha > 0,\ \alpha \neq 1 \tag{3.9}$$

Πρέπει να σημειωθεί ότι όταν $\alpha \to 1$ το παραπάνω μέτρο ισούται με την απόκλιση *Kullback-Leibler*.

Άλλα μέτρα απόκλισης που συναντάμε στην βιβλιογραφία είναι

$$\textit{Kagan}\ (1963):\quad I_X^{Ka}(f_1, f_2) = \int \left(1 - \frac{f_2}{f_1}\right)^2 f_1 d\mu \tag{3.10}$$

$$\textit{Matusita}\ (1967):\ I_X^M(f_1, f_2) = \left[\int \left(f_1^{1/2} - f_2^{1/2}\right)^2 d\mu\right]^{1/2} \tag{3.11}$$

$$\textit{Vajda}\ (1973):\quad I_X^V(f_1, f_2) = \int \left|1 - \frac{f_2}{f_1}\right|^a f_1 d\mu \tag{3.12}$$

Β.2 Μέτρα τύπου φ – divergence

Το μέτρο πληροφορίας του *Csiszar* είναι ένα γενικευμένο μέτρο απόκλισης, γνωστής και ως φ – απόκλισης, η οποία βασίζεται σε μια κυρτή συνάρτηση φ και ορίζεται ως

$$\textit{Csiszar}\ (1963):\ I_X^C(f_1, f_2) = \int f_2 \varphi(f_1/f_2) d\mu \tag{3.13}$$

όπου φ μια κυρτή συνάρτηση στο $[0,\infty]$, έτσι ώστε $0\varphi(0/0)=0$, $\varphi(u)_{u\to 0}\to 0$ και $0\varphi(u/0)=u\varphi_\infty$ με $\varphi_\infty=\lim_{u\to\infty}[\varphi(u)/u]$.

Μπορούμε να παρατηρήσουμε ότι το μέτρο του *Csiszar*, γίνεται το μέτρο *Kullback-Leibler* εάν $\varphi(u)=u\ln u$. Όταν $\varphi(u)=(1-u)^2$ ή $\varphi(u)=\mathrm{sgn}(a-1)u^a$, $\alpha>0$, $\alpha\neq 1$, το μέτρο του *Csiszar* γίνεται το μέτρο *Kagan* ($\chi^2-Pearson$) και η απόκλιση του *Renyi* αντίστοιχα.

Β.3 Μέτρα τύπου *power - divergence*

Μια άλλη γενίκευση μέτρων απόκλισης είναι η οικογένεια μέτρων *power - divergence* που παρουσιάστηκαν από τους *Cressie & Read* και ισούνται με

$$\textit{Cressie \& Read}\ (1984):\ I_X^{CR}(f_1,f_2)=\frac{1}{\lambda(\lambda+1)}\int f_1(z)\left[\left(\frac{f_1(z)}{f_2(z)}\right)^{\lambda}-1\right]dz,\ \lambda\in R \qquad (3.14)$$

Για $\lambda=0,-1$ το μέτρο ορίζεται από συνέχεια. Το μέτρο *Kullback-Leibler* υπολογίζεται για $\lambda\to 0$. Σημειώνουμε ότι, τα παραπάνω μέτρα ορίζονται και στην περίπτωση διακριτών κατανομών. Έστω ότι $P=(\pi_1,\pi_2,...,\pi_m)$ και $Q=(q_1,q_2,...,q_m)$, δυο διακριτές πεπερασμένες κατανομές πιθανότητας. Η διακριτή εκδοχή του μέτρου *Csiszar* δίνεται από τον τύπο

$$I_X^C(P,Q)=\sum_{i=1}^{m}q_i\varphi(\pi_i/q_i) \qquad (3.15)$$

και του μέτρου *Cressie & Read* από τον τύπο

$$I_X^{CR}(P,Q)=\frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{m}\pi_i\left[\left(\frac{\pi_i}{q_i}\right)^{\lambda}-1\right],\ \lambda\in R \qquad (3.16)$$

όπου πάλι για $\lambda=0,-1$ το μέτρο ορίζεται από συνέχεια.

Γ. Μέτρα τύπου εντροπίας (*entropy*)

Τα μέτρα εντροπίας εκφράζουν το ποσό της πληροφορίας που εμπεριέχεται σε μια κατανομή, δηλαδή το ποσό της αβεβαιότητας σε σχέση με το αποτέλεσμα ενός πειράματος. Τα κλασσικά μέτρα τέτοιου τύπου είναι των *Shannon* και *Renyi*.

Έστω $P=(\pi_1,\pi_2,...,\pi_n)$ μια πεπερασμένη διακριτή κατανομή πιθανότητας, μιας τυχαίας μεταβλητής X. Η εντροπία κατά *Shannon* ορίζεται ως

$$\textit{Shannon}\ (1948):\ H_X^S(P)=-\sum_{i=1}^{n}\pi_i\log(\pi_i) \tag{3.17}$$

Η εντροπία *Shannon*, γενικεύθηκε αργότερα από τον *Renyi* ως εντροπία τάξης a και δίνεται απο τον τύπο

$$\textit{Renyi}\ (1961):\ H_X^S(P)=\frac{1}{1-a}\ln\sum_{i=1}^{n}\pi_i^a\ \text{ όπου }\ a>0\ \text{και}\ a\neq 1, \tag{3.18}$$

Μια περαιτέρω γενίκευση, στα πλαίσια του μέτρου *Csiszar*, που βασίζεται σε μια κυρτή συνάρτηση φ, γνωστή και ως $\varphi-$εντροπία, προτάθηκε από τους *Burbea-Rao* και δίνεται από τον τύπο

$$\textit{Burbea-Rao}\ (1982):\ H_X^{\varphi}(P)=-\sum_{i=1}^{n}\varphi(\pi_i) \tag{3.19}$$

Τέλος αξίζει να αναφέρουμε το μέτρο εντροπίας των *Havrda & Charvat*

$$\textit{Havrda \& Charvat}\ (1967):\ H_X^{HC}(P)=\frac{1-\sum\pi_i^a}{a-1} \tag{3.20}$$

όπου $a>0$ και $a\neq 1$ και για $a=2$ γίνεται ο δείκτης *Gini Concentration*.

Άλλα μέτρα εντροπίας που συμπεριλαμβάνουν την $\gamma-$εντροπία, δίνονται από τον τύπο

$$H_X^{\gamma}(P)=\frac{1-\left(\sum\pi_i^{1/\gamma}\right)^{\gamma}}{1-2^{\gamma-1}},\ \text{όπου}\ \gamma>0\ \text{και}\ \gamma\neq 1 \tag{3.21}$$

Η εντροπία ζευγών δίνεται από τον τύπο

$$H_X^{p}=-\sum\pi_i\ln\pi_i-\sum(1-\pi_i)\ln(1-\pi_i) \tag{3.22}$$

όπου το ζευγάρωμα γίνεται υπό την έννοια $(\pi_i,1-\pi_i)$ [βλέπε *Burbea & Rao* (1982)].

Τα μέτρα πληροφορίας και απόκλισης έχουν αρκετές ιδιότητες, όπως μεταξύ άλλων τις «*non-negativity*», «*maximal information*», «*sufficiency*». Υπάρχει ένα κομμάτι γνώσης, γνωστό ως θεωρία της στατιστικής πληροφορίας, το οποίο έχει κάνει προόδους, αλλά δεν έχει ευρεία

αποδοχή και εφαρμογή. Η προσέγγιση είναι περισσότερο λειτουργική παρά αξιωματική, όπως στην περίπτωση της εντροπίας κατά *Shannon*. Στο Κεφάλαιο 6 παρουσιάζουμε έναν αριθμό μέτρων πληροφορίας και απόκλισης.

ΚΕΦΑΛΑΙΟ 4

ΜΕΤΡΑ ΣΥΝΑΦΕΙΑΣ ΓΙΑ ΟΝΟΜΑΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ

4.1 Εισαγωγή

Στο Κεφάλαιο αυτό θα παρουσιάσουμε τα κυριότερα μέτρα, που συνοψίζουν την συνάφεια για ονοματικές μεταβλητές. Τα μέτρα συνάφειας για ονοματικά δεδομένα δεν στηρίζονται στην διάταξη των μεταβλητών και επομένως, τα μέτρα που θα εξετάσουμε μοιράζονται την ιδιότητα αυτή. Αν και κάποια από αυτά, θα μπορούσαν να χρησιμοποιηθούν και για διατακτικές μεταβλητές ή υψηλότερου επιπέδου, κάτι τέτοιο δεν προτείνεται, καθώς τα αντίστοιχα μέτρα για μεταβλητές υψηλότερου επιπέδου έχουν μεγαλύτερη ισχύ. Οι τιμές των μέτρων αυτών είναι πάντα θετικές και κυμαίνονται στο διάστημα $[0,1]$ καθώς η κατεύθυνση της συνάφειας δεν έχει νόημα για ονοματικές μεταβλητές. Η τιμή 0 υποδηλώνει ανυπαρξία κάποιας σχέσης μεταξύ των δυο μεταβλητών και η τιμή 1 υποδηλώνει πλήρης συνάφεια των μεταβλητών. Κάποια μέτρα εφαρμόζονται μόνο σε 2×2 πίνακες, τα περισσότερα όμως μπορούν να χρησιμοποιηθούν για διδιάστατους $I \times J$ πίνακες συνάφειας. Η ανάλυση πινάκων συνάφειας διάστασης > 2, ξεφεύγει από τον σκοπό της παρούσας εργασίας. Γενικά, τα ονοματικά μέτρα συνάφειας διακρίνονται, βάση του τρόπου υπολογισμού τους, σε τρεις κύριες κατηγορίες:

a. μέτρα που βασίζονται στο *odds ratio*
b. μέτρα που βασίζονται στο $Pearson's\ \chi^2 - test$ της ανεξαρτησίας
c. μέτρα προγνωστικής συνάφειας (*measures of predictive association*)

Οι δυο πρώτες κατηγορίες αναφέρονται στα παλαιότερα, κλασσικά ή παραδοσιακά μέτρα συνάφειας, ενώ η τελευταία κατηγορία αναφέρεται σε πιο σύγχρονες προσεγγίσεις. Πριν ξεκινήσουμε την παρουσίαση των μέτρων συνάφειας για ονοματικές μεταβλητές, θα αναφερθούμε σε μια ειδική κατηγορία μέτρων, γνωστά και ως μέτρα λόγου (*ratio measures*), δηλαδή κάποιον δεικτών που χρησιμοποιούνται για την σύγκριση ομάδων, στην περίπτωση δυο διχοτομημένων κατηγορικών μεταβλητών ενός 2×2 πίνακα συνάφειας.

4.2 Μέτρα συνάφειας για δίτιμες κατηγορικές μεταβλητές

Πολλές μελέτες είναι σχεδιασμένες να συγκρίνουν ομάδες με βάση μια δυαδική μεταβλητή απόκρισης. Η παράγραφος αυτή παρουσιάζει τις κύριες παραμέτρους για την σύγκριση ομάδων. Η εξαρτημένη μεταβλητή Y έχει μόνο δυο κατηγορίες, που περιγράφουν το αποτέλεσμα ή την επίδραση (*outcome or effect*), όπως για παράδειγμα, «αποτυχία» ή «επιτυχία». Η ανεξάρτητη μεταβλητή X , περιγράφεται από δυο ή περισσότερες κατηγορίες, τις ομάδες ή αγωγές (*group ή treatments*). Στην περίπτωση δυο ομάδων, ο ακόλουθος 2×2 πίνακας συνάφειας, Πίνακας 4-1, παρουσιάζει κατάλληλα τα αποτελέσματα.

ΠΙΝΑΚΑΣ 4-1

2×2 Πίνακας Συνάφειας

		ΕΠΙΔΡΑΣΗ		
		Απόκριση 1	Απόκριση 2	**Σύνολο**
ΟΜΑΔΑ	Ομάδα Α	N_{11} π_{11}	N_{12} π_{12}	$\mathbf{N_{1.}}$ $\boldsymbol{\pi_{1.}}$
	Ομάδα Β	N_{21} π_{21}	N_{22} π_{22}	$\mathbf{N_{2.}}$ $\boldsymbol{\pi_{2.}}$
	Σύνολο	$\mathbf{N_{.1}}$ $\boldsymbol{\pi_{.1}}$	$\mathbf{N_{.2}}$ $\boldsymbol{\pi_{.2}}$	**N** **1**

Οι δειγματικές συχνότητες κάθε κελιού συμβολίζονται με $\{n_{ij}\}$, οι περιθώριες συχνότητες με $\{n_{i.}\}$ και $\{n_{.j}\}$, ενώ το συνολικό μέγεθος του δείγματος ισούται με $n = \sum_i \sum_j n_{ij}$. Οι δειγματικές κατανομές πιθανότητας συμβολίζονται με p_{ij} . Για παράδειγμα, τα $\{p_{ij}\}$ συμβολίζουν την από κοινού δειγματική κατανομή, με $p_{ij} = n_{ij}/n$ και περιθώριες κατανομές $p_{i.} = n_{i.}/n$, $p_{.j} = n_{.j}/n$.

4.2.1 Σχετικός Κίνδυνος

Ο σχετικός κίνδυνος (*Relative Risk - RR*) είναι ένα πολύ γνωστό μέτρο συνάφειας για διχοτομημένες κατηγορικές μεταβλητές. Χρησιμοποιείται ευρέως σε ιατρικές μελέτες παραγόντων κινδύνου, όπως η σχέση μια θεραπείας με τα καρδιακά εμφράγματα. Ο σχετικός κίνδυνος, ορίζεται ως

$$RR = \frac{r_1}{r_2} = \frac{\pi_{11}}{\pi_{11}+\pi_{12}} \Big/ \frac{\pi_{21}}{\pi_{21}+\pi_{22}} = \frac{\pi_{11}\left(\pi_{21}+\pi_{22}\right)}{\pi_{21}\left(\pi_{11}+\pi_{12}\right)} \tag{4.1}$$

όπου $r_1 = \dfrac{\pi_{11}}{\pi_{11}+\pi_{12}} = \Pr\left(\text{Απόκριση } 1 \mid \text{Ομάδα } A\right)$ είναι ο κίνδυνος εμφάνισης ενός αποτελέσματος, δοθείσης της ομάδος στην οποία ανήκει το υποκείμενο,

και $r_2 = \dfrac{\pi_{21}}{\pi_{21}+\pi_{22}} = \Pr\left(\text{Απόκριση } 1 \mid \text{Ομάδα } B\right)$ ο κίνδυνος εμφάνισης του ίδιου αποτελέσματος, δοθείσης της άλλης ομάδος στην οποία ανήκει το υποκείμενο.

Η δειγματική εκτίμηση του σχετικού κινδύνου ισούται με

$$\hat{RR} = \frac{\hat{r}_1}{\hat{r}_2} = \frac{p_{11}}{p_{11}+p_{12}} \Big/ \frac{p_{21}}{p_{21}+p_{22}} \tag{4.2}$$

αντικαθιστώντας τις πληθυσμιακές πιθανότητες π_{ij} , με τις αντίστοιχες δειγματικές p_{ij} .

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 2×2 πίνακα (*σελ. 165, Παράδειγμα 1*), η εξαρτημένη μεταβλητή (*στήλη*) περιγράφει την επιβίωση ενός ατόμου (*Ναι, Όχι*), σε σχέση με το αν λαμβάνει κάποια θεραπεία (*Θεραπεία Α*) ή αν λαμβάνει ψευδοφάρμακο (*Θεραπεία Β*). Συγκρίνοντας τις γραμμές, με βάση την πρώτη κατηγορία "*Όχι*" της μεταβλητής απόκρισης «Επιβίωση», έχουμε:

Ο κίνδυνος θανάτου για κάποιο άτομο που παίρνει την *Θεραπεία Α* είναι:

$$\hat{r}_1 = \frac{p_{11}}{p_{11}+p_{12}} = \frac{0.08}{0.5} = \frac{80}{500} = 0.16$$

Ο κίνδυνος θανάτου για κάποιο άτομο που παίρνει την *Θεραπεία Β* είναι:

$$\hat{r}_2 = \frac{p_{21}}{p_{21}+p_{22}} = \frac{0.1}{0.5} = \frac{100}{500} = 0.20$$

Επομένως, ο σχετικός κίνδυνος ισούται με

$$\hat{RR} = \frac{0.16}{0.20} = 0.80 .$$

Κατά ανάλογο τρόπο, συγκρίνοντας τις γραμμές με βάση την δεύτερη κατηγορία "*Ναι*" της μεταβλητής απόκρισης «*Επιβίωση*», έχουμε διαφορετική τιμή για τον σχετικό κίνδυνο, ο οποίος ορίζεται ως

$$\hat{RR} = \frac{1-\hat{r}_1}{1-\hat{r}_2} = \frac{0.84}{0.80} = 1.05 .$$

Ερμηνεία

Σύμφωνα με το παράδειγμά μας, τα άτομα που βρίσκονται υπό αγωγή, έχουν το 80% του «*κινδύνου*» μη-επιβίωσης, σε σχέση με τα άτομα που παίρνουν ψευδοφάρμακο (επομένως μικρότερο κίνδυνο). Παρόμοια, αν ορίσουμε ως «*κίνδυνο*» το ενδεχόμενο της επιβίωσης, τότε τα άτομα που βρίσκονται υπό αγωγή έχουν το 105% του «*κινδύνου*» ή της «*ευκαιρίας επιβίωσης*» των ατόμων που παίρνουν ψευδοφάρμακο (επομένως μεγαλύτερο κίνδυνο, δηλαδή ευκαιρία επιβίωσης).

Πεδίο Ορισμού

Ο σχετικός κίνδυνος RR είναι ένας μη αρνητικό77ς αριθμός με τιμές $[0,\infty)$. Όταν $RR = 1$ οι μεταβλητές (*επιβίωση – θεραπεία*) είναι ανεξάρτητες, δηλαδή δεν υπάρχει διαφορά στον κίνδυνο μεταξύ των ομάδων (*θεραπεία – ψευδοφάρμακο*), ενώ όταν $RR \neq 1$ οι μεταβλητές είναι εξαρτημένες ή συσχετισμένες.

Συγκεκριμένα, γα τιμές <1 το ενδεχόμενο είναι λιγότερο πιθανό να συμβεί στην ομάδα θεραπείας από ότι στην ομάδα ελέγχου (*ψευδοφάρμακο*) και επομένως έχουμε αρνητική συνάφεια (η θεραπεία σχετίζεται με λιγότερες περιπτώσεις θανάτου).

Για τιμές >1 το ενδεχόμενο είναι περισσότερο πιθανό να συμβεί στην ομάδα θεραπείας από ότι στην ομάδα ελέγχου (*ψευδοφάρμακο*) και επομένως έχουμε θετική συνάφεια (η θεραπεία σχετίζεται με περισσότερες περιπτώσεις θανάτου).

Επίπεδο Δεδομένων

Εφαρμόζεται σε 2×2 πίνακες συνάφειας, με διχοτομημένες ονοματικές και διατακτικές μεταβλητές ή άλλης ανώτερης κατηγορίας.

Συμμετρία

Ο σχετικός κίνδυνος αποτελεί ένα ασύμμετρο μέτρο συνάφειας, με την έννοια ότι διαφορετικά αποτελέσματα θα προκύψουν ανάλογα με το ποια θα είναι η ανεξάρτητη μεταβλητή (*γραμμή*).

Σχόλια

Η έννοια του σχετικού κινδύνου, χρησιμοποιείται ευρέως στις ιατρικές επιστήμες ως παράγοντας κινδύνου (*risk factor*), όπως για παράδειγμα στην μελέτη της σχέσης μεταξύ μιας θεραπείας και μιας ασθένειας και επιτρέπει σε έναν ερευνητή να συγκρίνει τις σχετικές πιθανότητες επιβίωσης ενός ατόμου, εάν ανήκει σε μια ομάδα (συνήθως σε αυτή του υψηλότερου κινδύνου) σε σχέση με την επιβίωση ενός ατόμου που ανήκει σε μια ομάδα χαμηλότερου κινδύνου. Όταν τα r_i είναι πολύ κοντά στο $0,1$, ο σχετικός κίνδυνος παρέχει καλύτερη πληροφόρηση από την σύγκριση ποσοστών καθώς η τιμή $r_1 - r_2$ είναι μεγαλύτερης σημασίας. Για παράδειγμα, σε μελέτη που συγκρίνει δυο θεραπείες A και B με βάση την απόκριση 1, δηλαδή το ποσοστό των υποκειμένων που πεθαίνουν, η διαφορά $\hat{r}_1 = 0.010$ και $\hat{r}_2 = 0.001$ είναι πολύ πιο αξιοσημείωτη, από ότι η διαφορά $\hat{r}_1 = 0.410$ και $\hat{r}_2 = 0.401$, αν και στις δυο περιπτώσεις $\hat{r}_1 - \hat{r}_2 = 0.009$. Στην περίπτωση αυτή, ο σχετικός κίνδυνος είναι $\hat{R}\hat{R} = \frac{\hat{r}_1}{\hat{r}_2} = \frac{0.010}{0.001} = 10$ και $\hat{R}\hat{R} = \frac{\hat{r}_1}{\hat{r}_2} = \frac{0.410}{0.401} = 1.02$, αντίστοιχα. Με άλλα λόγια στην πρώτη περίπτωση, το ποσοστό θανάτου $\hat{r}_1 = 0.010$ (όταν ένα υποκείμενο λαμβάνει την θεραπεία A) είναι 900% μεγαλύτερο από το ποσοστό θανάτου $\hat{r}_2 = 0.001$ (όταν ένα υποκείμενο λαμβάνει την θεραπεία B). Στην δεύτερη περίπτωση, το ποσοστό θανάτου $\hat{r}_1 = 0.410$ για την θεραπεία A είναι μόλις 2% μεγαλύτερο του ποσοστού θανάτου της θεραπείας B.

Αξίζει να σημειώσουμε και την έννοια της μείωσης του σχετικού κινδύνου $\left(RRR\right)$ (*Relative Risk Reduction*), η οποία ορίζεται ως $RRR = 1 - RR$. Στο παράδειγμά μας, $RRR = 1 - 0.80 = 0.20$ και εκφράζει ότι για τα δεδομένα μας, η Θεραπεία A μειώνει τον κίνδυνο κατά 20%. Αν και ο σχετικός κίνδυνος RR είναι διαισθητικά ευκολότερος στην κατανόηση, το *odds ratio* είναι πιο χρήσιμο, όταν χρησιμοποιείται στα πλαίσια μιας στατιστικής ανάλυσης.

4.2.2 Λόγος πιθανοτήτων (*odds ratio*) ή λόγος διαγώνιων γινομένων

Ο λόγος πιθανοτήτων $\left(odds\,ratio\right)$ είναι ένα μέτρο που εισήγαγε ο *Cornfield* (1951) και χρησιμοποιείται ευρέως στις επιδημιολογικές μελέτες για να δείξει τον κίνδυνο ενός ατόμου να πάσχει από μια ασθένεια. Εκφράζει τον βαθμό συνάφειας μεταξύ δυο μεταβλητών, με ένα διαφορετικό αριθμητικό τρόπο από τα υπόλοιπα μέτρα συνάφειας, παρέχοντας έναν πιο σαφή τρόπο ερμηνείας ενός 2×2 πίνακα συνάφειας.

Ο τύπος υπολογισμού του είναι

$$\theta=\frac{\Omega_1}{\Omega_2}=\frac{\pi_{11}/\pi_{1.}}{1-\pi_{11}/\pi_{1.}}\Bigg/\frac{\pi_{21}/\pi_{2.}}{1-\pi_{21}/\pi_{2.}}=\frac{r_1}{1-r_1}\Bigg/\frac{r_2}{1-r_2}=\frac{\pi_{11}}{\pi_{12}}\Bigg/\frac{\pi_{21}}{\pi_{22}}=\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}\,, \tag{4.3}$$

όπου Ω_i τα *odds*.

Η δειγματική εκτίμηση ορίζεται ως

$$\hat{\theta}=\frac{\hat{\Omega}_1}{\hat{\Omega}_2}=\frac{p_{11}p_{22}}{p_{12}p_{21}} \tag{4.4}$$

αντικαθιστώντας τις πληθυσμιακές πιθανότητες π_{ij}, με τις αντίστοιχες δειγματικές p_{ij}.

Γενικά, το *odds* ενός ενδεχομένου E είναι το πηλίκο $P(E)/1-P(E)$. Στην ιατρική ή στην επιδημιολογία, περιγράφει τον λόγο της επιτυχίας ως προς την αποτυχία, συνήθως αναφορικά με την επιβίωση ενός ασθενή. Συμβολίζεται με Ω και σύμφωνα με τον Πίνακα 4-1 (*σελ. 36*),

$$\Omega_1=odds\left(\text{Απόκριση 1}\,|\,\text{Ομάδα }A\right)=\frac{\left(\text{Απόκριση 1}\,|\,\text{Ομάδα }A\right)}{1-\left(\text{Απόκριση 1}\,|\,\text{Ομάδα }A\right)}=\frac{r_1}{1-r_1}=\frac{\pi_{11}/\pi_{1.}}{1-\pi_{11}/\pi_{1.}}=\frac{\pi_{11}}{\pi_{12}}$$

και

$$\Omega_2=odds\left(\text{Απόκριση 1}\,|\,\text{Ομάδα }B\right)=\frac{\left(\text{Απόκριση 1}\,|\,\text{Ομάδα }B\right)}{1-\left(\text{Απόκριση 1}\,|\,\text{Ομάδα }B\right)}=\frac{r_2}{1-r_2}=\frac{\pi_{21}/\pi_{2.}}{1-\pi_{21}/\pi_{2.}}=\frac{\pi_{21}}{\pi_{22}}$$

Αποδεικνύεται αλγεβρικά ότι το $odds\,ratio$ για $2x2$ πίνακες συνάφειας, ισούται με το γινόμενο των συχνοτήτων των διαγώνιων κελιών του πίνακα, δηλαδή $\theta=\frac{n_{11}n_{22}}{n_{12}n_{21}}$ και είναι γνωστό και ως λόγος διαγώνιων γινομένων (*Cross - Product ratio*), [*Yule*, (1900, 1912)].

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 2×2 πίνακα, (*σελ. 165, Παράδειγμα 1*), έχουμε

$$\hat{\Omega}_1 = \frac{p_{11}}{p_{12}} = \frac{0.08}{0.42} = 0.19,$$

δηλαδή ο λόγος της επιτυχίας «*όχι-επιβίωση*» ως προς την αποτυχία «*επιβίωση*», δοθέντος ότι το υποκείμενο ανήκει στην ομάδα θεραπείας A.

$$\hat{\Omega}_2 = \frac{p_{21}}{p_{22}} = \frac{0.1}{0.4} = 0.25,$$

δηλαδή ο λόγος της επιτυχίας «*όχι-επιβίωση*» ως προς την αποτυχία «*επιβίωση*», δοθέντος ότι το υποκείμενο ανήκει στην ομάδα θεραπείας B. Επομένως, το *odds ratio* ισούται με

$$\hat{\theta} = \frac{\hat{\Omega}_1}{\hat{\Omega}_2} = \frac{0.19}{0.25} = 0.76.$$

Ερμηνεία

Όταν το $\Omega_i > 1$ τότε η πιθανότητα ενός γεγονότος να συμβεί είναι > 0.5, ενώ όταν $\Omega_i < 1$ η πιθανότητα ενός γεγονότος να συμβεί είναι < 0.5. Κατά συνέπεια, όσο μεγαλύτερο είναι το *odds* ενός γεγονότος να συμβεί, τόσο υψηλότερη η πιθανότητα ότι το γεγονός θα συμβεί και όσο μικρότερο είναι το *odds* ενός γεγονότος να συμβεί, τόσο χαμηλότερη η πιθανότητα ότι το γεγονός θα συμβεί. Όταν $\Omega_i = 1$, τότε η πιθανότητα το γεγονός να συμβεί ισούται με την πιθανότητα το γεγονός να μην συμβεί και όταν το $\Omega_i = 0$, τότε η πιθανότητα το γεγονός να συμβεί είναι 0.

Επομένως, η ένταση της σχέσης μεταξύ δυο μεταβλητών μπορεί να εκφραστεί, με τον βαθμό με τον οποίο τα δυο *odds* διαφέρουν. Η διαφορά αυτή συνοψίζεται στον λόγο των Ω_i. Στο παράδειγμά μας, η τιμή $\hat{\theta} = 0.76$ (< 1), υποδηλώνει αρνητική συνάφεια και σημαίνει ότι το Ω_1, ένας ασθενής να μην επιβιώσει υπό θεραπεία, ισούται με το 76% του Ω_2 αν δεν υποβάλλεται σε θεραπεία. Με άλλα λόγια, όταν κάποιος ανήκει στην χαμηλή ομάδα (Θεραπεία A), σχετίζεται με το να ανήκει στην υψηλότερη μεταβλητή αποτελέσματος (Επιβίωση: «Ναι»).

Πεδίο Ορισμού

Η τιμή του θ κυμαίνεται στο διάστημα $(-\infty,+\infty)$. Όταν $\theta=1$ οι μεταβλητές (Επιβίωση – Θεραπεία) είναι ανεξάρτητες, δηλαδή $\Omega_1=\Omega_2$, δεν υπάρχει διαφορά μεταξύ των *odds*, ενώ όταν $\theta\neq 1$ οι μεταβλητές είναι εξαρτημένες ή συσχετισμένες.

Συγκεκριμένα, για $\theta<1$ το ενδεχόμενο είναι λιγότερο πιθανό να συμβεί στην ομάδα θεραπείας A, από ότι στην ομάδα ελέγχου B (*ψευδοφάρμακο*) και επομένως έχουμε αρνητική συνάφεια.

Για $\theta>1$ το ενδεχόμενο είναι περισσότερο πιθανό να συμβεί στην ομάδα θεραπείας A από ότι στην ομάδα ελέγχου B (*ψευδοφάρμακο*) και επομένως έχουμε θετική συνάφεια.

Στην περίπτωση που ένα κελί έχει μηδενική πιθανότητα τότε $\theta=0$ ή $\theta=\infty$.

Η τιμή του $\log\theta$, που θα συναντήσουμε στην συνέχεια, κυμαίνεται στο $(-\infty,+\infty)$, με $\log\theta=0$, στην περίπτωση που οι μεταβλητές είναι ανεξάρτητες, δηλαδή $\theta=1$.

Επίπεδο Δεδομένων

Εφαρμόζεται σε 2×2 πίνακες συνάφειας, με διχοτομημένες ονοματικές και διατακτικές μεταβλητές ή άλλης ανώτερης κατηγορίας. Αν και επεκτείνεται και σε πίνακες μεγαλύτερης από 2×2 διάστασης, η ερμηνεία του γίνεται δυσκολότερη. Στις περιπτώσεις όμως, όπου έχουμε πολλές γραμμές αλλά δυο μόνο στήλες, η ερμηνεία του παραμένει ξεκάθαρη.

Συμμετρία

Το θ είναι ένα συμμετρικό μέτρο συνάφειας καθώς δεν μεταβάλλεται σε αλλαγές στην διάταξη γραμμών και στηλών. Όταν ο πίνακας αλλάξει προσανατολισμό, δηλαδή οι γραμμές γίνουν στήλες και οι στήλες γραμμές, τότε $\theta=1/\theta$. Επίσης, δεν είναι απαραίτητο να προσδιορίσουμε την μεταβλητή απόκρισης για να χρησιμοποιήσουμε το θ, διότι ορίζεται μέσω των δεσμευμένων κατανομών οποιασδήποτε κατεύθυνσης, δηλαδή

$$\theta=\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}=\frac{P(Y=1|X=1)P(Y=2|X=1)}{P(Y=1|X=2)P(Y=2|X=2)}=\frac{P(X=1|Y=1)P(X=2|Y=1)}{P(X=1|Y=2)P(X=2|Y=2)}$$

Άλλα χαρακτηριστικά

Η σχέση που συνδέει το *odds ratio* και τον σχετικό κίνδυνο *RR* είναι

$$\theta=RR\times\frac{(1-r_2)}{(1-r_1)}$$

Μπορούμε να παρατηρήσουμε, ότι όταν η πιθανότητα του αποτελέσματος που μας ενδιαφέρει είναι πολύ μικρή, δηλαδή όταν τα $r_i \to 0$, τότε τα δυο μέτρα συνάφειας θ και RR έχουν την ίδια βαρύτητα. Επομένως, στην περίπτωση αυτή, η εκτίμηση του θ αποτελεί μια αδρή εκτίμηση RR, όταν αυτός δεν είναι δυνατό να υπολογιστεί, όπως στην περίπτωση *case - control* μελετών. Όπως οι *Pagano & Gauvreau* (1993) σημειώνουν, όταν η πιθανότητα ενός γεγονότος να συμβεί είναι πολύ μικρή, τότε οι τιμές των κελιών n_{11} και n_{21} θα είναι πολύ μικρές. Όταν συμβαίνει αυτό, οι τιμές των θ και RR είναι σχεδόν ίδιες.

Σχόλια

Γενικά, όσο περισσότερο το θ απομακρύνεται από το 1, τόσο μεγαλύτερη επίδραση έχει η ανεξάρτητη μεταβλητή στην εξαρτημένη, υποδηλώντας είτε θετική $\theta > 1$, είτε αρνητική $\theta < 1$ συνάφεια. Για παράδειγμα, όταν $\theta = 4$, το *odds* της γραμμής 1 είναι 4 φορές μεγαλύτερο από το *odds* της γραμμής 2. Αυτό βέβαια δεν σημαίνει, ότι η πιθανότητα ή ο κίνδυνος θανάτου είναι τετραπλάσιος, δηλαδή $r_1 = 4r_2$. Στην ουσία, αυτό είναι η ερμηνεία του σχετικού κινδύνου. Επίσης, δυο τιμές του θ δείχνουν την ίδια συνάφεια αλλά σε διαφορετική κατεύθυνση, όταν η μία είναι αντίστροφη της άλλης. Για παράδειγμα, όταν $\hat{\theta} = 0.76$, το *odds* της επιτυχίας της γραμμής 1, είναι 0.76 φορές μεγαλύτερο της γραμμής 2, ή ισοδύναμα το *odds* της επιτυχίας της γραμμής 2 είναι $\frac{1}{\hat{\theta}} = 1.31$ φορές μεγαλύτερο από ότι αυτό της γραμμής 1. Τέλος, αξίζει να σημειώσουμε, ότι το *odds ratio* δεν είναι συμμετρικό γύρω από την τιμή 1. Δηλαδή, ένα *odds ratio* > 1 για ένα συγκεκριμένο βαθμό, δείχνει μικρότερη επίδραση από ότι ένα *odds ratio* < 1, για τον ίδιο βαθμό. Όταν $\theta < 1$ $\theta \in [0,1)$, ενώ όταν $\theta > 1$ τα *odds ratios* μπορούν να πάρουν οποιαδήποτε τιμή. Όμως, χρησιμοποιώντας το $\log\theta$, τότε το *odds ratio* κυμαίνεται συμμετρικά γύρω από το 1, με $\log\theta$ να κυμαίνεται συμμετρικά γύρω από το 0 (*ανεξαρτησία*) και η αντιστροφή των γραμμών ή των στηλών, συνεπάγεται την αλλαγή του πρόσημου. Στο παράδειγμά μας, $\hat{\theta} = 0.76$ και $\frac{1}{\hat{\theta}} = 1.31$, δηλαδή δυο διαφορετικές τιμές του θ δείχνουν την ίδια συνάφεια, αλλά σε διαφορετική κατεύθυνση, χωρίς να κυμαίνονται συμμετρικά γύρω από το 1. Όμως, $\log 0.76 = -0.11$ και $\log 1.31 = 0.11$, δηλαδή δυο ίδιες τιμές του $\log\theta$, που κυμαίνονται

συμμετρικά γύρω από το 0, δείχνουν τον ίδιο βαθμό συνάφειας, με διαφορετικό πρόσημο [*Hershberger, Fisher,* (2005)].

4.3 Μέτρα συνάφειας που βασίζονται στο *odds ratio*

Όπως είδαμε το $\theta \in (-\infty, +\infty)$. Ο *Yule* (1900), εισήγαγε τον συντελεστή Q και τον συντελεστή Y, σκοπεύοντας να περιορίσει τις τιμές του *odds ratio* στο διάστημα $[-1, +1]$. Οι συντελεστές βασίζονται στην έννοια του λόγου των διαγώνιων γινομένων και προτείνονται λιγότερο συχνά ως μέτρα συνάφειας για 2×2 πίνακες, από ότι ο συντελεστής *phi*, που θα συναντήσουμε στην συνέχεια, καθώς χρησιμοποιούν λιγότερη πληροφορία.

4.3.1 Συντελεστής συνάφειας Q του *Yule*

Ο συντελεστής συνάφειας Q του *Yule* (1900) υπολογίζεται από τον τύπο

$$Q = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}} = \frac{\theta - 1}{\theta + 1}, \tag{4.5}$$

όπου $\{\pi_{ij}\}$ οι πιθανότητες των κελιών ενός 2×2 πίνακα συνάφειας και θ η τιμή του *odds ratio*. Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε τις συχνότητες των κελιών $\{n_{ij}\}$.

Η δειγματική εκτίμηση του συντελεστή ισούται με

$$\hat{Q} = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}} = \frac{\hat{\theta} - 1}{\hat{\theta} + 1} \tag{4.6}$$

αντικαθιστώντας τις πληθυσμιακές πιθανότητες π_{ij}, με τις αντίστοιχες δειγματικές p_{ij}.

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 2×2 πίνακα, (*σελ. 165, Παράδειγμα 1*), έχουμε

$$\hat{Q} = \frac{0.08 \times 0.4 - 0.42 \times 0.1}{0.08 \times 0.4 + 0.42 \times 0.1} = -0.14 \left(= \frac{0.76 - 1}{0.76 + 1} \right)$$

Ερμηνεία

Η τιμή του συντελεστή $\hat{Q} = -0.14$, δείχνει αρνητική συνάφεια μεταξύ των δυο μεταβλητών, πολύ μικρής όμως έντασης. Μπορούμε να πούμε, ότι ο συντελεστής Q, βασίζεται στην διαφορά μεταξύ των «σύμφωνων» (*Concordant*) ζευγών παρατηρήσεων (C) και των «ασύμφωνων» (*Discordant*) ζευγών παρατηρήσεων (D) και εκφράζει την διαφορά $(C-D)$, ως ποσοστό επί του συνόλου των ζευγών παρατηρήσεων χωρίς δεσμούς $(C+D)$. Σημειώνουμε ότι το σύνολο των «σύμφωνων» ζευγών παρατηρήσεων για ένα 2×2 πίνακα, είναι $C = n_{11}n_{22}$ και το σύνολο των «ασύμφωνων» ζευγών είναι $C = n_{12}n_{21}$ (βλέπε *Παράγραφο 5.2, σελ. 77*). Κατά συνέπεια, μια ερμηνεία του αποτελέσματος θα μπορούσε να είναι ότι το έλλειμμα «σύμφωνων» ζευγών παρατηρήσεων έναντι των «ασύμφωνων» ισούται με το 14% επί του συνόλου των ζευγών παρατηρήσεων χωρίς δεσμούς.

Πεδίο Ορισμού

Ο συντελεστής *Yule's Q* κυμαίνεται στο διάστημα $[-1,+1]$, με $Q = 0$ να υποδηλώνει ανεξαρτησία μεταξύ των μεταβλητών.

Επίπεδο Δεδομένων

Εφαρμόζεται σε 2×2 πίνακες συνάφειας, με διχοτομημένες ονοματικές και διατακτικές μεταβλητές ή άλλης ανώτερης κατηγορίας.

Συμμετρία

Ο συντελεστής *Yule's Q* είναι ένα συμμετρικό μέτρο συνάφειας καθώς δεν μεταβάλλεται σε αλλαγές στην διάταξη γραμμών και στηλών, δηλαδή όταν ο πίνακας αλλάξει προσανατολισμό και οι γραμμές γίνουν στήλες και οι στήλες γραμμές. Με άλλα λόγια, δεν έχει σημασία ποια είναι η ανεξάρτητη μεταβλητή ή αλλιώς η μεταβλητή στήλη.

Άλλα χαρακτηριστικά

Όπως αναφέραμε, ο συντελεστής *Yule's Q* βασίζεται στην διαφορά μεταξύ των «σύμφωνων» και των «ασύμφωνων» ζευγών παρατηρήσεων και εκφράζει την διαφορά $(C-D)$, ως ποσοστό επί του συνόλου των ζευγών παρατηρήσεων χωρίς δεσμούς $(C+D)$. Κατά συνέπεια, ο

συντελεστής Q μπορεί να εκφραστεί σε όρους «σύμφωνων» και «ασύμφωνων» ζευγών παρατηρήσεων, ως εξής

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} = \frac{C - D}{C + D}.$$

Δηλαδή αποτελεί μια ειδική περίπτωση του μέτρου συνάφειας *gamma* των *Goodman & Kruskal* (βλέπε *παράγραφος 5.2.1, σελ. 80*), αν και ο συντελεστής *Yule's* Q χρησιμοποιείται για ονοματικές και διατακτικές μεταβλητές. Στην πραγματικότητα, ο συντελεστής *Yule's* Q αλγεβρικά ισοδυναμεί με τον 2×2 συντελεστή *Goodman & Kruskal gamma* και άρα μετράει και τον βαθμό εναρμόνισης ή μη, μεταξύ δυο μεταβλητών [*Hershberger, Fisher,* (2005)].

Οι *Ott et al*. (1992), σημειώνουν ότι η σημαντικότητα του συντελεστή Q μπορεί να εκτιμηθεί από την σχέση

$$z = \frac{Q}{\sqrt{\frac{1}{4}\left(1 - Q^2\right)^2 \left[\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right]}}.$$

Σχόλια

Στην βιβλιογραφία αναφέρεται ότι ο συντελεστής *Yule's* Q έχει την τάση να αυξάνει τον βαθμό συνάφειας του υπό μελέτη πληθυσμού. Οι *Ott et al*. (1992), σημειώνουν ότι ένας επιπρόσθετος περιορισμός του συντελεστή είναι, ότι για $|Q| = 1$, δεν σημαίνει απαραίτητα ότι υπάρχει πλήρης συνάφεια μεταξύ των δυο μεταβλητών. Συγκεκριμένα, για έναν 2×2 πίνακα, αν τουλάχιστον για μια παρατηρούμενη συχνότητα, ισχύει ότι $n_{ij} = 0$, τότε $Q = \pm 1$, δηλαδή ο συντελεστής θα ισούται με το άνω ή το κάτω όριό του. Στις περιπτώσεις αυτές, η ερμηνεία του συντελεστή Q μπορεί να είναι παραπλανητική και γι' αυτό ο συντελεστής Q δεν συνιστάται, όταν η συχνότητα για κάποιο κελί του πίνακα είναι πολύ μικρή.

4.3.2 Συντελεστής συνάφειας Y του *Yule*

Ο *Yule* (1900), πρότεινε ως εναλλακτικό του συντελεστή Q, τον συντελεστή Y (*Coefficient of Colligation*), που ορίζεται ως

$$Y = \frac{\sqrt{\pi_{11}\pi_{22}} - \sqrt{\pi_{12}\pi_{21}}}{\sqrt{\pi_{11}\pi_{22}} + \sqrt{\pi_{12}\pi_{21}}} = \frac{\sqrt{\theta} - 1}{\sqrt{\theta} + 1}, \tag{4.7}$$

όπου $\{p_{ij}\}$ οι πιθανότητες των κελιών ενός 2×2 πίνακα συνάφειας και θ η τιμή του *odds ratio*. Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε τις συχνότητες των κελιών $\{n_{ij}\}$.

Η δειγματική εκτίμηση του συντελεστή ισούται με

$$\hat{Y} = \frac{\sqrt{p_{11}p_{22}} - \sqrt{p_{12}p_{21}}}{\sqrt{p_{11}p_{22}} + \sqrt{p_{12}p_{21}}} = \frac{\sqrt{\hat{\theta}} - 1}{\sqrt{\hat{\theta}} + 1} \tag{4.8}$$

αντικαθιστώντας τις πληθυσμιακές πιθανότητες π_{ij} , με τις αντίστοιχες δειγματικές p_{ij} .

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 2×2 πίνακα (*σελ. 165, Παράδειγμα 1*), έχουμε

$$\hat{Y} = \frac{\sqrt{0.08 \times 0.4} - \sqrt{0.42 \times 0.1}}{\sqrt{0.08 \times 0.4} + \sqrt{0.42 \times 0.1}} = -0.07 \left(= \frac{\sqrt{0.76} - 1}{\sqrt{0.76} + 1} \right)$$

Ερμηνεία

Η τιμή του συντελεστή $\hat{Y} = -0.07$, δείχνει αρνητική συνάφεια μεταξύ των δυο μεταβλητών, ασήμαντης όμως έντασης. Ο συντελεστής *Yule's Y* δεν έχει μια εύκολη ερμηνεία. Ερμηνεύεται διαφορετικά από τον συντελεστή συσχέτισης Q, καθώς χρησιμοποιεί τον γεωμετρικό μέσο των διαγώνιων και μη διαγώνιων ζευγών παρατηρήσεων, αντί του αριθμού των ζευγών. Στην ουσία, εκφράζει την διαφορά μεταξύ των πιθανοτήτων των διαγώνιων και των μη διαγώνιων κελιών, όπου οι πιθανότητες γραμμής και στήλης έχουν τυποποιηθεί ως προς $0,5$. Σύμφωνα με το παράδειγμά μας, μπορούμε να πούμε ότι ο γεωμετρικός μέσος του ελλείμματος «σύμφωνων» ζευγών παρατηρήσεων έναντι των «ασύμφωνων», ως ποσοστό επί όλων των ζευγών χωρίς δεσμούς, είναι 7%.

Πεδίο Ορισμού

Ο συντελεστής *Yule's Y* κυμαίνεται στο διάστημα $[-1, +1]$, με $Y = 0$ να υποδηλώνει ανεξαρτησία μεταξύ των μεταβλητών.

Επίπεδο Δεδομένων

Εφαρμόζεται σε 2×2 πίνακες συνάφειας, με διχοτομημένες ονοματικές και διατακτικές μεταβλητές ή άλλης ανώτερης κατηγορίας.

Συμμετρία

Ο συντελεστής *Yule's* Y είναι ένα συμμετρικό μέτρο συνάφειας καθώς δεν μεταβάλλεται σε αλλαγές στην διάταξη γραμμών και στηλών, δηλαδή όταν ο πίνακας αλλάξει προσανατολισμό και οι γραμμές γίνουν στήλες και οι στήλες γραμμές. Με άλλα λόγια, δεν έχει σημασία ποια είναι η ανεξάρτητη μεταβλητή ή αλλιώς η μεταβλητή στήλη.

Σχόλια

Γενικά, ο συντελεστής *Yule's* Y θα είναι μικρότερος από ότι ο *Yule's* Q, κυρίως διότι ο Y λαμβάνει ποινές για σχεδόν αδύναμες μονότονες σχέσεις, παρόμοια με τον συντελεστή *Somers'* D (βλέπε *Παράγραφο 5.2.3, σελ. 85*) , ο οποίος όμως χρησιμοποιείται πολύ πιο συχνά, λόγω της εύκολης ερμηνείας του. Επίσης, ο *Yule's* Y είναι λιγότερο ευαίσθητος από ότι ο *Yule's* Q, στις διαφορές των περιθώριων κατανομών των δυο μεταβλητών. Ερμηνεύεται όπως ο συντελεστής συσχέτισης r του *Pearson*, αλλά αλγεβρικά δεν ισούται με 1.

4.4 Μέτρα συνάφειας που βασίζονται στο $\chi^2 - test$ του *Pearson*

Μια κοινά αποδεκτή ερμηνεία της ύπαρξης συνάφειας, σε έναν διδιάστατο πίνακα είναι ότι οι μεταβλητές γραμμής και στήλης είναι εξαρτημένες. Ο κλασσικός έλεγχος της υπόθεσης της ανεξαρτησίας, βασίζεται στο $\chi^2 - test$. Έστω ο Πίνακας 2-1 (*σελ. 10*), που περιγράφει έναν $I \times J$ πίνακα συνάφειας. Τα $\{\pi_{ij}\}$ είναι οι πληθυσμιακές πιθανότητες και είναι συνήθως άγνωστες στην πράξη. Η πιθανότητα ένα τυχαία επιλεγμένο υποκείμενο του πληθυσμού να ταξινομηθεί στο κελί (ij) ισούται με π_{ij}, έτσι ώστε $E(N_{ij}) = E_{ij} = N\pi_{ij}$ η αναμενόμενη συχνότητα για το κελί (ij). Μια πολύ γνωστή μέθοδος, που επιτρέπει να αποφασίσουμε αν οι αναμενόμενες συχνότητες $\{N\pi_{ij}\}$ των κελιών ενός $I \times J$ πίνακα συνάφειας ταυτίζονται με τις

παρατηρούμενες $\{n_{ij}\}$, είναι η ελεγχοσυνάρτηση καλής προσαρμογής των δεδομένων χ^2-test του *Pearson* (1904), η οποία στην γενική μορφή της ορίζεται ως

$$\chi^2=\sum_i\sum_j\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}=\sum_i\sum_j\frac{\left(N_{ij}-N\pi_{ij}\right)^2}{N\pi_{ij}}, \tag{4.9}$$

όπου O_{ij} οι παρατηρούμενες συχνότητες του πληθυσμού και E_{ij} οι αναμενόμενες συχνότητες.

Στην περίπτωση που οι μεταβλητές X,Y είναι ανεξάρτητες, δηλαδή κάτω από την μηδενική υπόθεση $\mathrm{H}_0:\pi_{ij}=\pi_{i.}\pi_{.j}$, για κάθε $i=1,2,...,I$ και $j=1,2,...,J$ το χ^2-test παίρνει την μορφή

$$\chi^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij}-n_{i.}n_{.j}/n\right)^2}{n_{i.}n_{.j}/n} \tag{4.10}$$

αντικαθιστώντας τις αναμενόμενες πληθυσμιακές πιθανότητες $\{\pi_{ij}\}$ και συχνότητες $\{N_{ij}\}$ με τις αντίστοιχες δειγματικές $\{p_{ij}\}$, $\{n_{ij}\}$, όπου $p_{ij}=p_{i.}p_{.j}=\frac{n_{i.}}{n}\frac{n_{.j}}{n}$.

Χρησιμοποιώντας τα δεδομένα του 2×2 πίνακα (*σελ. 165, Παράδειγμα 1*), δοθέντος ότι το 50% των ασθενών λαμβάνουν την Θεραπεία A και το 50% την Θεραπεία B, κάτω από την μηδενική υπόθεση της ανεξαρτησίας, ότι δηλαδή δεν υπάρχει σχέση μεταξύ αγωγής και αποτελέσματος, αναμένουμε $90\left(=\frac{180}{2}\right)$ ασθενείς να μην επιβιώσουν και $210\left(=\frac{420}{2}\right)$ να επιβιώσουν. Σημειώνουμε, ότι ο ακριβής τύπος υπολογισμού των αναμενόμενων συχνοτήτων $\{n_{ij}\}$ είναι $n_{ij}=\frac{n_{i.}n_{.j}}{n}$. Όσο μεγαλύτερη είναι η διαφορά μεταξύ των παρατηρούμενων και αναμενόμενων μετρήσεων, τόσο λιγότερο πιθανό να είναι αληθής, η μηδενική υπόθεση της ανεξαρτησίας. Σημειώνουμε για μελλοντική αναφορά, ότι η σχέση (4.10) για έναν 2×2 πίνακα συνάφειας, αλγεβρικά ισοδυναμεί με

$$\chi^2=\frac{n\left(n_{11}n_{22}-n_{21}n_{12}\right)^2}{n_{1.}n_{.1}n_{2.}n_{.2}}\ \text{[}\textit{Liebetrau}\text{, (1983)]}.$$

Η μέγιστη τιμή του $\chi^2 - test$ για έναν $I \times J$ πίνακα, είναι $\chi^2_{\max} = n(q-1)$, όπου $q = \min\{I, J\}$, όταν κάθε γραμμή με $I \geq J$ ή κάθε στήλη με $I \leq J$ του πίνακα περιέχει έναν μη μηδενικό αριθμό. Επομένως, το $\chi^2 - test$ κυμαίνεται στο διάστημα $\left[0, n(q-1)\right]$. Η τιμή του $\chi^2 - test$ εξαρτάται όχι μόνο από την ένταση της σχέσης μεταξύ των μεταβλητών, αλλά και από το μέγεθος του δείγματος. Επιπλέον, το $\chi^2 - test$ δεν περιγράφει την ένταση ή την φύση της συνάφειας. Χρησιμοποιώντας τα λόγια του *Sir Austin Bradford Hill* (1965), «*Όπως η φωτιά, το* $\chi^2 - test$ *είναι ένας εξαίρετος υπηρέτης αλλά ένας κακός αφέντης*» [*Scheaffer*, (1999)]. Το $\chi^2 - test$ μπορεί να μετασχηματισθεί σε αρκετά μέτρα συνάφειας, τα οποία όμως δεν χρήζουν εύκολης ερμηνείας. Τα μέτρα αυτά, περιγράφουν την ένταση της σχέσης μεταξύ των μεταβλητών και αποτελούν στην ουσία μια προσαρμογή της ελεγχοσυνάρτησης, προκειμένου να εξαλειφθεί η επίδραση του μεγέθους του δείγματος.

4.4.1 Συντελεστής φ του *Yule*

Ο συντελεστής φ (*Coefficient phi*, *Yule* 1912), εξαλείφει την επίδραση του μεγέθους του δείγματος, διαιρώντας το $\chi^2 - test$ με το μέγεθος του δείγματος n και στην συνέχεια υπολογίζοντας την τετραγωνική του ρίζα. Ο τύπος υπολογισμού του συντελεστή είναι

$$\varphi = \sqrt{\frac{\chi^2}{n}} \tag{4.11}$$

Στην βιβλιογραφία ο συντελεστής αναφέρεται και ως συντελεστής μέσης τετραγωνικής συνάφειας του *Pearson* (*Pearson's coefficient of mean square contingency*). Πράγματι, η πληθυσμιακή αναλογία της σχέσης [4.10], ονομάζεται μέση τετραγωνική συνάφεια του *Pearson* και ορίζεται ως

$$\varphi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(\pi_{ij} - \pi_{i.}\pi_{.j}\right)^2}{\pi_{i.}\pi_{.j}} \tag{4.12}$$

Αντικαθιστώντας τα $\{\pi_{ij}\}$ με τις δειγματικές εκτιμήσεις $\{p_{ij}\}$, έχουμε την δειγματική εκτίμηση

$$\hat{\varphi}^2 = \frac{\chi^2}{n} \tag{4.13}$$

[*Liebetrau, Measures of Association,* (1983)].

Εναλλακτικά, στην περίπτωση ενός 2×2 πίνακα, ο συντελεστής ορίζεται ως

$$\varphi = \frac{n_{11}n_{22} - n_{21}n_{12}}{\sqrt{n_{1.}n_{.1}n_{2.}n_{.2}}} \tag{4.14}$$

και αλγεβρικά ισοδυναμεί με την σχέση (4.11) [*Liu,* (1980)].

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 2×2 πίνακα (*σελ. 165, Παράδειγμα 1*), έχουμε

$$\hat{\varphi} = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{2.71}{1000}} = 0.05 \text{ ή αλλιώς } \hat{\varphi} = \frac{n_{11}n_{22} - n_{21}n_{12}}{\sqrt{n_{1.}n_{.1}n_{2.}n_{.2}}} = \frac{32000 - 42000}{192094} = -0.05\,.$$

Ερμηνεία

Η τιμή του συντελεστή $\hat{\varphi} = -0.05$ δείχνει αρνητική συνάφεια, ασήμαντης όμως σημασίας. Παρατηρούμε, ότι ανάλογα με τον τύπο υπολογισμού προκύπτει η ίδια απόλυτη τιμή. Δεν υπάρχει μια απλή διαισθητική ερμηνεία για τον συντελεστή. Βάσει του τύπου υπολογισμού του, που εμπεριέχει το $\chi^2 - test$, θα μπορούσαμε να πούμε ότι εκφράζει την συνάφεια μεταξύ δυο μεταβλητών ως ποσοστό της μέγιστης δυνατής μεταβλητότητάς τους. Σύμφωνα με τον *Darlington* [*Measures of Association for Crosstab Tables, Cornell University, New York* (2001)], για 2×2 πίνακες, ο συντελεστής φ, μπορεί να ερμηνευθεί ως η συμμετρική ποσοστιαία

διαφορά, μετρώντας το ποσοστό συγκέντρωσης των περιπτώσεων πάνω στην διαγώνιο, ή αλλιώς εκφράζει την μέση ποσοστιαία διαφορά μεταξύ των δυο μεταβλητών, θεωρώντας ότι η μια προκαλεί την άλλη. Πράγματι, υπολογίζοντας την ποσοστιαία διαφορά pd_1 ($pd-percentage\ difference$), θεωρώντας την μεταβλητή (*γραμμή*) «Αγωγή» ως ανεξάρτητη και την μεταβλητή (*στήλη*) «Επιβίωση» ως εξαρτημένη έχουμε: $pd_1 = n_{11}/n_{1.} - n_{21}/n_{2.} = 0.16 - 0.20 = -0.04$. Αντιστρέφοντας τις γραμμές και τις στήλες, θεωρώντας την μεταβλητή «Επιβίωση» ως ανεξάρτητη μεταβλητή και επομένως την μεταβλητή «Αγωγή» ως εξαρτημένη, τότε η ποσοστιαία διαφορά είναι $pd_2 = n_{11}/n_{1.} - n_{21}/n_{2.} = 0.44 - 0.51 = -0.07$. Έτσι, η τιμή του $\hat{\varphi} = -0.05$ εκφράζει την μέση ποσοστιαία διαφορά. Επιπλέον, ο *Liu* (1980) στην εργασία του σημειώνει, ότι οι *Hernes* (1970) και *Davis* (1971) αναφέρουν για τον συντελεστή φ, οτι εκφράζει τον γεωμετρικό μέσο της ποσοστιαίας διαφοράς κατά μήκος των γραμμών και των στηλών, στην περίπτωση ονοματικών μεταβλητών. Αυτό σημαίνει ότι μπορεί να ερμηνευθεί ως μια συμμετρική παραλλαγή της ποσοστιαίας διαφοράς.

Πεδίο Ορισμού

Ο συντελεστής φ για 2×2 πίνακες κυμαίνεται στο διάστημα $[-1, +1]$, αν και αυτό δεν συμβαίνει πάντα, καθώς το άνω και κάτω όριο του συντελεστή εξαρτάται από συγκεκριμένες συνθήκες. Οι *Carroll* (1961) και *Guilford* (1965) σημειώνουν, ότι αναγκαία συνθήκη για να ισούται ο συντελεστής φ με -1 ή 1 (*τέλεια συνάφεια*), σε έναν 2×2 πίνακα, είναι τα περιθώρια αθροίσματα κάθε γραμμής και στήλης να είναι ίσα, δηλαδή $[n_{11} + n_{12} = n_{21} + n_{22}]$ και $[n_{11} + n_{21} = n_{12} + n_{22}]$. Επιπλέον, ο *Liu* (1980) σημειώνει ότι το ίδιο ισχύει όταν δυο συμμετρικά αντίθετα διαγώνια κελιά είναι 0. Όταν οι τιμές στα κελιά n_{11} ή / και n_{22} είναι μεγαλύτερες από όσο αναμένεται, τότε $\varphi \to 1$ και όταν συμβαίνει το αντίθετο τότε $\varphi \to -1$.

Για $I \times J$ πίνακες με $I, J > 2$, ο συντελεστής φ κυμαίνεται στο διάστημα $\left[0, \sqrt{q-1}\right]$, όπου $q = \min\{I, J\}$. Αυτό σημαίνει ότι μπορεί να ισχύει $\varphi > 1$, με θεωρητική τιμή $\varphi \to \infty$ και να διαφοροποιείται ανάλογα με το μέγεθος του πίνακα.

Επίπεδο Δεδομένων

Η χρήση του συντελεστή φ περιορίζεται μόνο σε 2×2 πίνακες συνάφειας, που περιέχουν διχοτομημένες ονοματικές μεταβλητές, διότι για πίνακες μεγαλύτερης διάστασης η τιμή του συντελεστή μπορεί να ξεπεράσει την τιμή 1. Οι *Siegel* και *Castellan* (1988) σημειώνουν, ότι όταν οι δυο μεταβλητές που συσχετίζονται είναι διατακτικές, τότε χρησιμοποιώντας τον συντελεστή φ χάνεται πληροφόρηση και εξ' αιτίας αυτού, κάτω από αυτές τις συνθήκες είναι προτιμότερο να χρησιμοποιούμε εναλλακτικά μέτρα συνάφειας, σχεδιασμένα για διατεταγμένους πίνακες.

Συμμετρία

Ο συντελεστής φ είναι ένα συμμετρικό μέτρο συνάφειας καθώς δεν μεταβάλλεται σε αλλαγές στην διάταξη γραμμών και στηλών, δηλαδή όταν ο πίνακας αλλάξει προσανατολισμό και οι γραμμές γίνουν στήλες και οι στήλες γραμμές. Με άλλα λόγια, δεν έχει σημασία ποια είναι η ανεξάρτητη μεταβλητή ή αλλιώς η μεταβλητή στήλη. Σημειώνουμε ότι ο συντελεστής φ έχει την τάση να αποκρύπτει ασυμμετρικές σχέσεις.

Σημαντικότητα

Η σημαντικότητα του συντελεστή φ είναι η ίδια με αυτή του $\chi^2 - test$. Αξίζει να σημειώσουμε, ότι αν το μέγεθος του δείγματος n αυξηθεί, αλλά η αναλογία των κελιών (*effect size*) διατηρηθεί, κάποια στιγμή το $\chi^2 - test$ δεν θα είναι στατιστικά σημαντικό. Έτσι στην περίπτωση αυτή, η τιμή του συντελεστή θα παραμένει η ίδια, αλλά δεν θα είναι στατιστικά σημαντική.

Άλλα χαρακτηριστικά

Ο συντελεστής φ αλγεβρικά ισούται με $|r|$, όπου r ο συντελεστής συσχέτισης του *Pearson* (*Pearson product-moment correlation coefficient*). Για την ακρίβεια, είναι ο συντελεστής συσχέτισης του *Pearson*, ο οποίος υπολογίζεται όταν οι τιμές 0 και 1 υιοθετούνται για να περιγράψουν τα επίπεδα δυο διχοτομημένων μεταβλητών.

Σχόλια

Το εύρος τιμών του συντελεστή φ εξαρτάται από τις διαστάσεις του πίνακα και μπορεί να υπερβεί την τιμή 1, κάνοντας τον συντελεστή ένα όχι και τόσο κατάλληλο μέτρο συνάφειας. Επιπλέον, η σύγκριση μεταξύ δυο ή περισσότερων συντελεστών φ που προέρχονται από πίνακες με διαφορετικά περιθώρια αθροίσματα, μπορεί να είναι παραπλανητική, καθώς ο συντελεστής

είναι πολύ ευαίσθητος σε μεταβολές των περιθώριων κατανομών. (βλέπε *Liu, A note on phi-coefficient comparison*, 1980). Μεταξύ άλλων, οι *Guilford* (1965) και *Fleiss* (1981), σημειώνουν ότι μια από τις πιο χρήσιμες εφαρμογές του συντελεστή φ είναι να ορίσει την εσωτερική συσχέτιση (*intercorrelation*), μεταξύ των αποκρίσεων δυο υποκειμένων σε δυο διχοτομημένα είδη. Για μια πιο σφαιρική θεώρηση, ο αναγνώστης μπορεί να ανατρέξει στους *Guilford* (1965) και *Fleiss* (1981), οι οποίοι μεταξύ άλλων, συζήτησαν διάφορους λόγους οι οποίοι επιχειρηματολογούν υπέρ της χρήσης άλλων μέτρων πέρα από τον συντελεστή φ, για 2×2 πίνακες συνάφειας (βλέπε *Handbook of Parametric and Non-Parametric Statistical Procedures, 3rd Edition, Chapman & Hall*, 2004).

4.4.2 Συντελεστής συνάφειας T του *Tschuprow*

Ο συντελεστής συνάφειας T (*Tschuprow*, 1918), ορίζεται ως

$$T=\sqrt{\frac{\varphi^2}{\sqrt{(I-1)(J-1)}}}, \tag{4.15}$$

όπου I, J το πλήθος των γραμμών και των στήλων, αντίστοιχα. Η ποσότητα $\{(I-1)(J-1)\}$ είναι οι βαθμοί ελευθερίας του πίνακα. Η δειγματική εκτίμηση του συντελεστή ισούται με

$$\hat{T}=\sqrt{\frac{\chi^2}{n\sqrt{(I-1)(J-1)}}} \tag{4.16}$$

η οποία υπολογίζεται αντικαθιστώντας την ποσότητα φ^2 της σχέσης (4.12), με την δειγματική της εκτίμηση $\hat{\varphi}^2=\frac{\chi^2}{n}$, σχέση (4.13).

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 2×2 πίνακα (*σελ. 165, Παράδειγμα 1*), έχουμε

$$\hat{T}=\sqrt{\frac{\chi^2}{n\sqrt{(I-1)(J-1)}}}=\sqrt{\frac{2.71}{1000\sqrt{(2-1)(2-1)}}}=0.05$$

Ερμηνεία

Η τιμή του συντελεστή $\hat{T} = 0.05$ υποδηλώνει συνάφεια μεταξύ των μεταβλητών, ασήμαντης όμως βαρύτητας. Δεν υπάρχει μια απλή, λειτουργική ερμηνεία για τον συντελεστή T. Θα μπορούσαμε να πούμε ότι εκφράζει την συνάφεια μεταξύ δυο μεταβλητών ως ποσοστό της μέγιστης δυνατής μεταβλητότητάς τους.

Πεδίο Ορισμού

Ο συντελεστής T κυμαίνεται στο διάστημα $\left[0,1\right]$ και $T = 1$, μόνο όταν οι δυο μεταβλητές έχουν ίσα περιθώρια αθροίσματα και ο πίνακας είναι τετραγωνικός. Η μέγιστη τιμή του συντελεστή υπολογίζεται από τον τύπο $T_{\max} = \sqrt[4]{\frac{q-1}{l}}$, όπου $q = \min\left\{I, J\right\}$ και $l = \min\left\{I-1, J-1\right\}$. Όσο λιγότερο τετραγωνικός είναι ένας πίνακας και όσο περισσότερο ανόμοιες είναι οι περιθώριες κατανομές των γραμμών και των στηλών, τόσο περισσότερο ο συντελεστής θα γίνεται μικρότερος του 1. Για παράδειγμα, το άνω όριο του συντελεστή T για έναν 3×10 πίνακα συνάφειας θα είναι $T_{\max} = \sqrt[4]{\frac{q-1}{l}} = \sqrt[4]{\frac{3-1}{9}} = 0.69$ [*Liebetrau*, (1983)].

Επίπεδο Δεδομένων

Η χρήση του συντελεστή T περιορίζεται μόνο σε τετραγωνικούς $I \times I$ πίνακες με $i = 1, 2, ..., I$. Μπορεί να χρησιμοποιηθεί για ονοματικά ή και άλλης υψηλότερης κατηγορίας (*διατακτικά*) δεδομένα, αν και δεν προτείνεται η χρήση του στην περίπτωση διατακτικών μεταβλητών καθώς χάνεται πληροφόρηση.

Συμμετρία

Ο συντελεστής T είναι ένα συμμετρικό μέτρο συνάφειας καθώς δεν μεταβάλλεται σε αλλαγές στην διάταξη γραμμών και στηλών, δηλαδή όταν ο πίνακας αλλάξει προσανατολισμό και οι γραμμές γίνουν στήλες και οι στήλες, γραμμές. Με άλλα λόγια, δεν έχει σημασία ποια είναι η ανεξάρτητη μεταβλητή ή αλλιώς η μεταβλητή στήλη.

Σημαντικότητα

Η σημαντικότητα του συντελεστή T είναι η ίδια με αυτή του $\chi^2 - test$, όπως συμβαίνει με όλα τα μέτρα που βασίζονται στο $\chi^2 - test$. Όπως έχουμε αναφέρει για τον συντελεστή φ, όσο αυξάνεται το δείγμα n, αλλά η αναλογία των κελιών (*effect size*) διατηρηθεί, κάποια στιγμή το

$\chi^2 - test$ παύει να είναι στατιστικά σημαντικό. Έτσι στην περίπτωση αυτή, η τιμή του συντελεστή T θα παραμένει η ίδια, αλλά δεν θα είναι στατιστικά σημαντική.

Άλλα χαρακτηριστικά

Για 2×2 πίνακες συνάφειας ισχύει ότι $T = \varphi$.

Σχόλια

Η χρήση του συντελεστή T είναι περιορισμένη και δεν υποστηρίζεται από τα γνωστά στατιστικά πακέτα, καθώς για μη τετραγωνικούς πίνακες $T < 1$.

4.4.3 Συντελεστής συνάφειας V του *Cramer*

Ο συντελεστής συνάφειας V (*Cramer*, 1946), είναι μια επέκταση του συντελεστή φ, για $I \times J$ πίνακες συνάφειας με $I, J > 2$. Στην βιβλιογραφία τον συναντάμε και με τον συμβολισμό φ_C.

Ο τύπος υπολογισμού του συντελεστή είναι

$$V = \sqrt{\frac{\varphi^2}{q-1}}, \tag{4.17}$$

όπου $q = \min\{I, J\}$ και η δειγματική εκτίμηση του συντελεστή ισούται με

$$\hat{V} = \sqrt{\frac{\chi^2}{n(q-1)}} \tag{4.18}$$

η οποία υπολογίζεται αντικαθιστώντας την ποσότητα φ^2 της σχέσης (4.12), με την δειγματική της εκτίμηση, $\hat{\varphi}^2 = \frac{\chi^2}{n}$, σχέση (4.13).

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 2×2 πίνακα (*σελ. 165, Παράδειγμα 1*), έχουμε

$$\hat{V} = \sqrt{\frac{\chi^2}{n(q-1)}} = \sqrt{\frac{2.71}{1000(2-1)}} = 0.05$$

Χρησιμοποιώντας τα δεδομένα του 3×4 πίνακα (*σελ. 166, Παράδειγμα 2*), έχουμε

$$\hat{V} = \sqrt{\frac{\chi^2}{n(q-1)}} = \sqrt{\frac{1073.5}{6800(3-1)}} = 0.28$$

Ερμηνεία

Η προέλευση του συντελεστή V, βασίζεται στο γεγονός ότι $\chi^2_{\max} = n(q-1)$. Έτσι, η τιμή του V, είναι η τετραγωνική ρίζα της αναλογίας, που αναπαριστά τον υπολογισμένη τιμή του $\chi^2 - test$, διαιρεμένη με την μέγιστη δυνατή τιμή του $\left(\chi^2_{\max}\right)$. Δεν υπάρχει μια απλή, λειτουργική ερμηνεία για τον συντελεστή V. Σύμφωνα με τον *Darlington* [*Measures of Association for crosstab tables, Cornell University, New York* (2001)], παρόμοια με τον συντελεστή συσχέτισης r του *Pearson*, ο συντελεστής V εκφράζει την συνάφεια μεταξύ δυο μεταβλητών, ως ποσοστό της μέγιστης δυνατής μεταβλητότητάς τους. Για παράδειγμα, το 8% της διακύμανσης $\left[8\% = 100\left(\hat{V}^2\right) = 100\left(0.28^2\right)\right]$, οφείλεται στην σχέση που ανιχνεύθηκε από το $\chi^2 - test$, μεταξύ των δυο μεταβλητών του Παραδείγματος 2.

Πεδίο Ορισμού

Ο συντελεστής V κυμαίνεται στο διάστημα $[0,1]$, με $V = 1$, μόνο όταν οι δυο μεταβλητές έχουν ίσα περιθώρια αθροίσματα. Δηλαδή, όσο πιο άνισα τα περιθώρια αθροίσματα γραμμών και στηλών, τόσο ο συντελεστής θα είναι μικρότερος της μονάδος.

Επίπεδο Δεδομένων

Ο συντελεστής V, χρησιμοποιείται με ονοματικά ή και άλλης υψηλότερης κατηγορίας (*διατακτικά*) δεδομένα, για οποιοδήποτε μέγεθος ενός $I \times J$ πίνακα. Όπως και στην περίπτωση του συντελεστή φ, όταν οι κατηγορίες των μεταβλητών είναι διατεταγμένες, δεν προτείνεται να χρησιμοποιούμε τον συντελεστή V, καθώς χάνεται πληροφόρηση.

Συμμετρία

Ο συντελεστής V είναι ένα συμμετρικό μέτρο συνάφειας καθώς δεν μεταβάλλεται σε αλλαγές στην διάταξη γραμμών και στηλών, δηλαδή όταν ο πίνακας αλλάξει προσανατολισμό και οι γραμμές γίνουν στήλες και οι στήλες, γραμμές. Με άλλα λόγια, δεν έχει σημασία ποια είναι η ανεξάρτητη μεταβλητή ή αλλιώς η μεταβλητή στήλη.

Σημαντικότητα

Η σημαντικότητα του συντελεστή V είναι η ίδια με αυτή του $\chi^2 - test$. Όπως έχουμε αναφέρει για τον συντελεστή φ, όσο αυξάνεται το δείγμα n, αλλά η αναλογία των κελιών (*effect size*) διατηρηθεί, κάποια στιγμή το $\chi^2 - test$ παύει να είναι στατιστικά σημαντικό. Έτσι στην

περίπτωση αυτή, η τιμή του συντελεστή V θα παραμένει η ίδια, αλλά δεν θα είναι στατιστικά σημαντική.

Άλλα χαρακτηριστικά

Για 2×2 πίνακες, είναι προφανές ότι $V = T = \varphi = \sqrt{\chi^2/n}$, ενώ σε πίνακες μεγαλύτερης διάστασης, οι τιμές των μέτρων διαφοροποιούνται και επομένως, κάποια στατιστικά πακέτα υπολογίζουν την τιμή του V, μόνο στην περίπτωση πινάκων μεγαλύτερης διάστασης. Ο *Daniel* (1990) σημειώνει ότι σε 2×2 πίνακες με το ίδιο σύνολο δεδομένων, $V = \tau^2$ (βλέπε *Handbook of Parametric and Non-Parametric Statistical Procedures, 3rd Edition, Chapman & Hall,* 2004), όπου τ το μέτρο συνάφειας διατακτικής ταξινόμησης του *Kendall*, για διατακτικές μεταβλητές, που θα συναντήσουμε στην συνέχεια (βλέπε *Παράγραφος 5.2.2, σελ. 82*).

Σχόλια

Ο συντελεστής V, θεωρείται το καλύτερο και πιο διαδεδομένο μέτρο μεταξύ αυτών που βασίζονται $\chi^2 - test$, διότι ανεξάρτητα από το μέγεθος του πίνακα, κυμαίνεται πιο κανονικά στο διάστημα $[0,1]$, όταν τα περιθώρια αθροίσματα γραμμών και στηλών είναι ίδια. Ο *Conover* (1980, 1999), σημειώνει ότι υπάρχει μια τάση για την τιμή του $\chi^2 - test$ (και κατά συνέπεια για την τιμή του V), να αυξάνει όσο η διάσταση του πίνακα αυξάνεται. Για τον λόγο αυτό, ο *Conover* συστήνει ότι ο συντελεστής V, ίσως να μην είναι τελείως ακριβής για την σύγκριση του βαθμού συνάφειας, σε πίνακες με διαφορετική διάσταση. Επίσης σημειώνει ότι, αν και κάτω από όλες τις συνθήκες, η τιμή του V θα κυμαίνεται μεταξύ $[0,1]$, η ερμηνεία του εξαρτάται από την διάσταση του πίνακα (βλέπε H*andbook of Parametric and Non-Parametric Statistical Procedures, 3rd Edition, Chapman & Hall, 2004*).

4.4.4 Συντελεστής συνάφειας C του *Pearson*

Ο συντελεστής συνάφειας C (*Pearson's contingency coefficient,* 1948), αποτελεί μια βελτίωση του συντελεστή φ, προκειμένου να μπορεί να εφαρμοσθεί σε έναν διδιάστατο $I\times J$ πίνακα, οποιουδήποτε μεγέθους. Ο τύπος υπολογισμού του συντελεστή είναι

$$C = \sqrt{\frac{\varphi^2}{1+\varphi^2}} \tag{4.19}$$

και η δειγματική εκτίμηση του συντελεστή ισούται με

$$\hat{C} = \sqrt{\frac{\varphi^2}{1+\varphi^2}} = \sqrt{\frac{\chi^2/n}{1+\chi^2/n}} = \sqrt{\frac{\chi^2}{\chi^2+n}} \tag{4.20}$$

η οποία υπολογίζεται αντικαθιστώντας την ποσότητα φ^2 της σχέσης (4.12), με την δειγματική της εκτίμηση $\hat{\varphi}^2 = \frac{\chi^2}{n}$, σχέση (4.13).

Παρατηρούμε ότι καθώς $n \neq 0$ τότε $0 \leq C < 1$, δηλαδή ο συντελεστής δεν μπορεί να πάρει την τιμή 1. Επίσης, ένας δεύτερος περιορισμός είναι ότι ο συντελεστής εξαρτάται από την διάσταση του πίνακα. Πράγματι, το άνω όριο του συντελεστή είναι συνάρτηση του αριθμού των γραμμών και των στηλών του $I \times J$ πίνακα, καθώς $C_{\max} = \sqrt{\frac{q-1}{q}}$, όπου $q = \min\{I, J\}$. Μια πρόταση για να εξουδετερώσουμε το ανωτέρω πρόβλημα, είναι να χρησιμοποιήσουμε τον προσαρμοσμένο συντελεστή συνάφειας C_{adj} του *Sakoda* (1977), ο οποίος ορίζεται ως

$$C_{adj} = \frac{C}{C_{\max}} \tag{4.21}$$

Επομένως, αν υπάρχει τέλεια σχέση μεταξύ των μεταβλητών, η τιμή του C_{adj} θα αντανακλά την σχέση αυτή και $C_{adj} = 1$.

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 2×2 πίνακα (*σελ. 165, Παράδειγμα 1*) έχουμε

$$\hat{C} = \sqrt{\frac{\chi^2}{\chi^2+n}} = \sqrt{\frac{2.71}{2.71+1000}} = 0.05 \text{ και } \hat{C}_{adj} = \frac{\hat{C}}{\hat{C}_{\max}} = \frac{0.05}{0.71} = 0.07$$

Χρησιμοποιώντας τα δεδομένα του 3×4 πίνακα (*σελ. 166, Παράδειγμα 2*), έχουμε

$$\hat{C} = \sqrt{\frac{\chi^2}{\chi^2+n}} = \sqrt{\frac{1073.5}{1073.5+6800}} = 0.37 \text{ και } \hat{C}_{adj} = \frac{\hat{C}}{\hat{C}_{\max}} = \frac{0.37}{0.82} = 0.45$$

Ερμηνεία

Δεν υπάρχει μια εύκολη, διαισθητική ερμηνεία για τους συντελεστές C και C_{adj}. Θα μπορούσαμε να πούμε ότι εκφράζουν την συνάφεια μεταξύ δυο μεταβλητών ως ποσοστό της μέγιστης δυνατής μεταβλητότητάς τους. Για παράδειγμα, το 14% της διακύμανσης $\left[14\% = 100\left(\hat{C}^2\right) = 100\left(0.37^2\right)\right]$ ή το 20% της διακύμανσης $\left[20\% = 100\left(\hat{C}_{adj}^2\right) = 100\left(0.45^2\right)\right]$, οφείλεται στην σχέση που ανιχνεύθηκε από το $\chi^2 - test$, μεταξύ των δυο μεταβλητών. Ο *Pearson* θεωρούσε τον συντελεστή C, ως μια ονοματική προσέγγιση του συντελεστή γραμμικής συσχέτισης r [*Darlington*, (2001)].

Πεδίο Ορισμού

Ο συντελεστής C, για 2×2 πίνακες κυμαίνεται στο διάστημα $[0, 0.71]$, αλλά προσεγγίζει την μονάδα όσο ο αριθμός των γραμμών και των στηλών αυξάνει. Ο προσαρμοσμένος συντελεστής C_{adj} κυμαίνεται στο διάστημα $[0,1]$, ανεξάρτητα από το μέγεθος του πίνακα.

Επίπεδο Δεδομένων

Οι συντελεστές C και C_{adj}, χρησιμοποιούνται με ονοματικά ή και άλλης υψηλότερης κατηγορίας (*διατακτικά*) δεδομένα, για οποιοδήποτε μέγεθος ενός $I \times J$ πίνακα. Όπως και στην περίπτωση του συντελεστή φ, όταν οι κατηγορίες των μεταβλητών είναι διατεταγμένες, είναι προτιμότερο να χρησιμοποιούμε εναλλακτικά μέτρα συνάφειας, σχεδιασμένα για διατεταγμένους πίνακες, καθώς χάνεται πληροφόρηση.

Συμμετρία

Οι συντελεστές C και C_{adj} είναι συμμετρικά μέτρα συνάφειας καθώς δεν μεταβάλλονται σε αλλαγές στην διάταξη γραμμών και στηλών, δηλαδή όταν ο πίνακας αλλάξει προσανατολισμό και οι γραμμές γίνουν στήλες και οι στήλες, γραμμές. Με άλλα λόγια, δεν έχει σημασία ποια είναι η ανεξάρτητη μεταβλητή ή αλλιώς η μεταβλητή στήλη.

Σημαντικότητα

Η σημαντικότητα των συντελεστών C και C_{adj} είναι η ίδια με αυτή του $\chi^2 - test$. Όπως έχουμε αναφέρει για τον συντελεστή φ, όσο αυξάνεται το δείγμα n, αλλά η αναλογία των κελιών (*effect size*) διατηρηθεί, κάποια στιγμή το $\chi^2 - test$ παύει να είναι στατιστικά σημαντικό. Έτσι στην

περίπτωση αυτή, η τιμή των συντελεστών θα παραμένει η ίδια, αλλά δεν θα είναι στατιστικά σημαντική.

Σχόλια

Οι *Ott et al.* (1992), αναφέρουν ότι μεταξύ των μειονεκτημάτων που σχετίζονται με τον συντελεστή συνάφειας C, είναι ότι θα είναι πάντα μικρότερος του 1, ακόμα και αν οι δυο μεταβλητές είναι πλήρως εξαρτημένες. Επιπλέον, για να γίνουν συγκρίσεις μεταξύ των συντελεστών συνάφειας, που έχουν υπολογιστεί για δυο ή περισσότερους πίνακες, θα πρέπει οι πίνακες να έχουν τον ίδιο αριθμό γραμμών και στηλών. Μια πρόταση για να εξουδετερώσουμε τους ανωτέρω περιορισμούς, είναι ο προσαρμοσμένος συντελεστής C_{adj}. Όμως, αν και ο συντελεστής C_{adj} επιτρέπει καλύτερη σύγκριση μεταξύ ανόμοιων πινάκων, οι συγκρίσεις δεν θα είναι απόλυτα ακριβής, καθώς και αυτός εξαρτάται από τις διαστάσεις και το μέγεθος του πίνακα [βλέπε *Handbook of Parametric and Non-Parametric Statistical Procedures, 3rd Edition, Chapman & Hall,* 2004]. Αξίζει να σημειώσουμε, ότι σε αντίθεση με τον συντελεστή V του *Cramer*, που εξετάσαμε προηγουμένως, ο συντελεστής C δεν ισούται πάντα με τον συντελεστή φ, όταν υπολογίζεται για 2×2 πίνακες [*Hershberger & Fisher,* (2005)]. Οι ερευνητές προτείνουν για τον συντελεστή C να χρησιμοποιείται για 5×5 ή μεγαλύτερης διάστασης πίνακες, όπου προσεγγίζει την τιμή 1. Για πίνακες μικρότερης διάστασης, ο συντελεστής C θα υποεκτιμά το μέγεθος της συνάφειας, ακόμα και όταν όλες οι παρατηρήσεις βρίσκονται στην διαγώνιο του πίνακα [*Darlington,* (2001)].

4.5 Μέτρα προγνωστικής συνάφειας

Μια άλλη κατηγορία μέτρων συνάφειας για ονοματικές μεταβλητές, είναι τα μέτρα πρόβλεψης ή προγνωστικής συνάφειας (*predictive association*), ανάλογης φιλοσοφίας με τον συντελεστή πολλαπλής συσχέτισης (*coefficient of determination*) που χρησιμοποιείται στην ανάλυση παλινδρόμησης [*Hershberger & Fisher,* (2005)]. Όταν υπάρχει σχέση μεταξύ δυο ονοματικών μεταβλητών X και Y, τότε η γνώση της X, μας επιτρέπει να αποκτήσουμε πληροφόρηση για την Y και η πληροφόρηση αυτή, είναι μεγαλύτερη από αυτή που θα είχαμε, εάν δεν γνωρίζαμε τίποτα για την μεταβλητή X. Έστω Δ_Y η διασπορά της Y και $\Delta_{Y|X}$ η

δεσμευμένη διασπορά της Y, δοθείσης της X. Ένα μέτρο πρόβλεψης, έστω M, ορίζεται ως $M_{Y|X} = 1 - \frac{\Delta_{Y|X}}{\Delta_Y}$ και συγκρίνει την δεσμευμένη διασπορά της Y, δοθείσης της X, με την μη - δεσμευμένη διασπορά της Y, παρόμοια όπως ο συντελεστής πολλαπλής συσχέτισης συγκρίνει την δεσμευμένη διακύμανση της εξαρτημένης μεταβλητής, με την μη - δεσμευμένη διακύμανσή της. Όταν $M_{Y|X} = 0$, οι μεταβλητές X και Y είναι ανεξάρτητα κατανεμημένες, δηλαδή η γνώση της X δεν προσφέρει καμία πληροφόρηση για την Y $\left(\Delta_{Y|X} = \Delta_Y\right)$, ενώ όταν $M_{Y|X} = 1$, η μεταβλητή X είναι ένας τέλειος προγνώστης της Y.

Στην συνέχεια θα περιγράψουμε 4 μέτρα που κάνουν λειτουργική την ιδέα που κρύβεται στην σχέση $M_{Y|X} = 0$. Τα μέτρα αυτά, ανήκουν σε μια ειδική κατηγορία και εναλλακτικά ονομάζονται, μέτρα αναλογικής μείωσης του σφάλματος PRE (*Proportionate Reduction in Error*). Αυτό σημαίνει, ότι οι τιμές τους αντανακλούν την ποσοστιαία μείωση του σφάλματος από την πρόβλεψη της εξαρτημένης μεταβλητής, δοθείσης της πληροφορίας που έχουμε για την ανεξάρτητη [*Goodman & Kruskal*, (1954)]. Έστω, ότι θέλουμε να μελετήσουμε την συνάφεια μεταξύ του χρώματος των ματιών X και του χρώματος των μαλλιών Y μιας ομάδας ατόμων (*σελ. 166, Παράδειγμα 2*). Ας υποθέσουμε ότι το χρώμα ματιών είναι η ανεξάρτητη μεταβλητή και θέλουμε να προβλέψουμε το χρώμα των μαλλιών, είτε χωρίς να έχουμε καμμία πληροφόρηση (*περίπτωση 1*), είτε γνωρίζοντας το χρώμα των ματιών (*περίπτωση 2*). Τότε ένα μέτρο αναλογικής μείωσης σφάλματος M_{PRE} εκφράζεται από την ποσότητα

$$M_{PRE} = \frac{(\pi\iota\theta\alpha\nu\acute{o}\tau\eta\tau\alpha\ \sigma\phi\acute{\alpha}\lambda\mu\alpha\tau o\varsigma\ \pi\varepsilon\rho\acute{\iota}\pi\tau\omega\sigma\eta\ 1) - (\pi\iota\theta\alpha\nu\acute{o}\tau\eta\tau\alpha\ \sigma\phi\acute{\alpha}\lambda\mu\alpha\tau o\varsigma\ \pi\varepsilon\rho\acute{\iota}\pi\tau\omega\sigma\eta\ 2)}{(\pi\iota\theta\alpha\nu\acute{o}\tau\eta\tau\alpha\ \sigma\phi\acute{\alpha}\lambda\mu\alpha\tau o\varsigma\ \pi\varepsilon\rho\acute{\iota}\pi\tau\omega\sigma\eta\ 1)}$$

4.5.1 Συντελεστής συνάφειας λ των *Goodman- Kruskal*

Ο συντελεστής λ ή *lambda* [*Goodman & Kruskal*, (1954)] είναι ένα μέτρο αναλογικής μείωσης του σφάλματος. Υπάρχουν 3 τύποι υπολογισμού για τον συντελεστή. Οι δυο λ_b, λ_a είναι ασυμμετρικοί , ανάλογα με ποια μεταβλητή θεωρείται ανεξάρτητη και ορίζονται ως

$$\lambda_b\left(Y \mid X\right)=\frac{\sum_i \max_j \pi_{ij}-\max_j \pi_{.j}}{1-\max_j \pi_{.j}}, \tag{4.22}$$

όπου

$\max_j \pi_{ij}$ η μέγιστη πιθανότητα για κάθε κατηγορία i της ανεξάρτητης μεταβλητή X και

$\max_j \pi_{.j}$ η μέγιστη περιθώρια πιθανότητα της εξαρτημένης μεταβλητής Y.

$$\lambda_a\left(X \mid Y\right)=\frac{\sum_j \max_i \pi_{ij}-\max_i \pi_{i.}}{1-\max_i \pi_{i.}}, \tag{4.23}$$

όπου

$\max_i \pi_{ij}$ η μέγιστη πιθανότητα για κάθε κατηγορία j της ανεξάρτητης μεταβλητή Y και

$\max_i \pi_{i.}$ η μέγιστη περιθώρια πιθανότητα της εξαρτημένης μεταβλητής X.

Η συμμετρική εκδοχή του συντελεστή λ, αποτελεί στην ουσία τον μέσο όρο των δυο ασύμμετρων συντελεστών λ_b, λ_a και ορίζεται ως

$$\lambda=\frac{\sum_i \max_j \pi_{ij}+\sum_j \max_i \pi_{ij}-\max_j \pi_{.j}-\max_i \pi_{i.}}{2-\max_j \pi_{.j}-\max_i \pi_{i.}} \tag{4.24}$$

Οι δειγματικές εκτιμήσεις των συντελεστών $\hat{\lambda}_b$, $\hat{\lambda}_a$ και $\hat{\lambda}$ υπολογίζονται αντικαθιστώντας τις πληθυσμιακές πιθανότητες $\left\{\pi_{ij}\right\}$ με τις δειγματικές εκτιμήσεις $\left\{p_{ij}\right\}$.

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 3×4 πίνακα (*σελ. 166, Παράδειγμα 2*), έχουμε

$$\hat{\lambda}_b=\frac{\sum_i \max_j p_{ij}-\max_j p_{.j}}{1-\max_j p_{.j}}=\frac{\left(0.26+0.20+0.06\right)-0.42}{1-0.42}=0.19$$

$$\hat{\lambda}_a=\frac{\sum_j \max_i p_{ij}-\max_i p_{i.}}{1-\max_i p_{i.}}=\frac{\left(0.26+0.20+0.11+0.01\right)}{1-0.46}=0.22$$

$$\hat{\lambda} = \frac{\sum_i \max_j p_{ij} + \sum_j \max_i p_{ij} - \max_j p_{.j} - \max_i p_{i.}}{2 - \max_j p_{.j} - \max_i p_{i.}} = 0.208$$

Ερμηνεία

Η ερμηνεία των συντελεστών εξαρτάται από τι ακριβώς θέλουμε να εξετάσουμε. Για παράδειγμα, θα μπορούσαμε να μελετήσουμε την αποτελεσματικότητα του προσδιορισμού του χρωματότυπου των ανδρών, όταν γνωρίζουμε το χρώμα των ματιών τους, αλλά όχι το χρώμα των μαλλιών. Η τιμή του συντελεστή λ_b μπορεί να ερμηνευθεί ως εξής: γνωρίζοντας το χρώμα ματιών, έχουμε μείωση στο σφάλμα από την πρόβλεψη του χρώματος μαλλιών κατά 19%. Οι ποσότητες στους αριθμητές των εξισώσεων, ερμηνεύονται ως μέτρα της διακύμανσης για ονοματικές μεταβλητές και η μέτρηση της διακύμανσης ονομάζεται *Gini Concentration* [για περισσότερες λεπτομέρειες αναφορικά με τον δείκτη *Gini*, βλέπε *Haberman* (1982)]. Τα σφάλματα που συμβαίνουν όταν δεν γνωρίζουμε το χρώμα ματιών (*ανεξάρτητη μεταβλητή* X), υπολογίζονται, αν από το σύνολο των παρατηρήσεων αφαιρέσουμε την πιο συνηθισμένη κατηγορία της εξαρτημένης μεταβλητής Y (*χρώμα μαλλιών*), δηλαδή, η πιθανότητα σφάλματος είναι $1-0.42=0.58$. Αν κάποιος δεν ήξερε την κατανομή της ανεξάρτητης μεταβλητής, θα μπορούσε να πει ότι όλοι οι άνδρες έχουν ξανθά μαλλιά και θα ήταν $2.829(42\%)$ φορές σωστός και $3.971(58\%)$ λάθος. Η αναλογική μείωση του σφάλματος που επιτυγχάνεται όταν γνωρίζουμε την κατανομή της ανεξάρτητης μεταβλητής, είναι το άθροισμα των πιο συχνών κατηγοριών της εξαρτημένης μεταβλητής για κάθε γραμμή, μείον τις σωστές προβλέψεις που κάποιος θα έκανε ούτως ή άλλως, δηλαδή $(0.26+0.20+0.06)-0.42=0.11$, διαιρεμένο με τον αριθμό των σφαλμάτων (0.58) που κάποιος θα έκανε ούτως ή άλλως . Ο συντελεστής λ_a έχει παρόμοια ερμηνεία. Άλλο παράδειγμα θα ήταν, αν θέλαμε να μελετήσουμε την γνώμη του κοινού αναφορικά με την σχέση μεταξύ χρώματος μαλλιών και ματιών. Στην περίπτωση αυτή ο συμμετρικός συντελεστής $\hat{\lambda}=0.208$ περιγράφει μια μέτρια συνάφεια μεταξύ των δυο μεταβλητών.

Πεδίο Ορισμού

Η τιμή των συντελεστών λ_b, λ_a και λ κυμαίνονται στο διάστημα $[0,1]$, με το λ να κυμαίνεται μεταξύ των λ_a, λ_b. Ο συντελεστής λ_b δεν ορίζεται, όταν όλες οι παρατηρήσεις συγκεντρώνονται μόνο σε μια από τις στήλες της εξαρτημένης μεταβλητής. Όταν $\lambda_b = 0$ τότε οι μεταβλητές είναι ανεξάρτητα κατανεμημένες, δηλαδή η γνώση της X δεν προσφέρει καμία πληροφόρηση για την Y (χωρίς όμως να ισχύει πάντα το αντίθετο), ενώ όταν $\lambda_b = 1$, η μεταβλητή X είναι ένας τέλειος προγνώστης της Y. Ομοίως για τον συντελεστή λ_a. Ο συντελεστής λ, δεν ορίζεται όταν όλες οι παρατηρήσεις συγκεντρώνονται σε ένα μοναδικό κελί του πίνακα. Επίσης, $\lambda = 1$ μόνο όταν όλες οι παρατηρήσεις συγκεντρώνονται σε κελιά, δυο εκ των οποίων δεν ανήκουν στην ίδια γραμμή ή στήλη και $\lambda = 0$ όταν υπάρχει στατιστική ανεξαρτησία, χωρίς όμως να ισχύει πάντα το αντίθετο [*Goodman & Kruskal, Measures of Association for Cross Classifications,* (1954)].

Επίπεδο Δεδομένων

Οι συντελεστές λ_b, λ_a και λ μπορούν να χρησιμοποιηθούν με ονοματικά δεδομένα ή και άλλης υψηλότερης κατηγορίας (*διατακτικά*). Όταν οι κατηγορίες των μεταβλητών είναι διατεταγμένες, είναι προτιμότερο να χρησιμοποιούμε εναλλακτικά μέτρα συνάφειας, σχεδιασμένα για διατεταγμένους πίνακες.

Συμμετρία

Οι συντελεστές λ_b, λ_a είναι ασυμμετρικοί, δηλαδή διαφορετικά αποτελέσματα θα προκύψουν, ανάλογα με το ποια θεωρείται ανεξάρτητη μεταβλητή. Όμως όπως είδαμε, η συμμετρική εκδοχή των λ_b, λ_a είναι ο συντελεστής λ, ο οποίος στην ουσία εκφράζει τον μέσο όρο των δυο ασύμμετρων συντελεστών.

Σημαντικότητα

Επειδή η δειγματική κατανομή των συντελεστών λ_b, λ_a και λ είναι γνωστή (*κανονική ασυμπτωματική κατανομή*), μπορούμε να υπολογίσουμε το ασυμπτωτικό τυπικό τους σφάλμα και την σημαντικότητα. Καθώς οι τύποι υπολογισμού της διακύμανσης είναι πολύπλοκοι και ξεφεύγουν από τους σκοπούς της παρούσας εργασίας, παραπέμπουμε τον αναγνώστη στο

εγχειρίδιο SAS/STAT User's Guide για αναφορά [*SAS/STAT User's Guide, Version 8, Chapter 28, p. 1296-1297*, (1999)].

Σχόλια

Οι τιμές των λ_b, λ_a και λ μπορεί να είναι 0, χωρίς να ισχύει στατιστική ανεξαρτησία και αυτό είναι ένα από τα μειονεκτήματα του μέτρου. Όπως οι *Goodman & Kruskall* (1954) αναφέρουν, όλα τα μέτρα συνάφειας μπορεί να υπόκεινται σε παρόμοιες κριτικές, χωρίς όμως να μειώνεται η αξία τους. Αν θα πρέπει να ορίσουμε πιο αυστηρά τον ορισμό της συνάφειας, τότε αρκετά μέτρα θα πρέπει να απορριφθούν. Ένα άλλο μειονέκτημα είναι ότι τα λ_b, λ_a και λ είναι ευαίσθητα στις ανισότητες των περιθωρίων αθροισμάτων των γραμμών (*ανεξάρτητη μεταβλητή*). Για παράδειγμα, αν τα αθροίσματα των γραμμών στα δεδομένα του Παραδείγματος 2 (*σελ. 166*), ήταν ίσα, δηλαδή είχαμε ίδιο πλήθος ατόμων για κάθε χρώμα ματιών και η κατανομή συχνοτήτων παρέμενε η ίδια, όπως περιγράφεται στο Παράδειγμα 2α (*σελ. 166*), τότε $\hat{\lambda} = 0.350$ ενώ στα δεδομένα του Παραδείγματος 2, $\hat{\lambda} = 0.208$. Επομένως, οι συντελεστές λ_b, λ_a και λ ίσως να μην είναι κατάλληλοι, όταν χρειάζεται να συγκρίνουμε πίνακες διαφορετικών πληθυσμών, καθώς εξαρτώνται από τις περιθώριες κατανομές. Για να αντιμετωπίσουν το πρόβλημα αυτό οι *Goodman & Kruskal* πρότειναν τα μέτρα *tau* (τ).

4.5.2 Συντελεστής συνάφειας τ των *Goodman & Kruskal*

Όπως αναφέραμε, τα *lambda* βασίζονται στην επιλογή των πιο συχνών κατηγοριών της εξαρτημένης μεταβλητής, για κάθε επίπεδο της ανεξάρτητης μεταβλητής, δοθέντος ότι γνωρίζουμε την κατανομή της τελευταίας. Ο συντελεστής τ ή *tau* [*Goodman & Kruskal*, (1954)], μπορεί να θεωρηθεί καλύτερος από τον *lambda*, καθώς υποθέτει ότι η πρόβλεψη της εξαρτημένης μεταβλητής από την γνώση της ανεξάρτητης, βασίζεται στην πραγματική κατανομή της πρώτης. Δηλαδή, αντί να υποθέτει ότι η πρόβλεψη βασίζεται στην μεγαλύτερη κατηγορία κάθε στήλης (*υπόθεση του συντελεστή lambda*), ο συντελεστής *tau* υποθέτει ότι κατά την διάρκεια της πρόβλεψης, επιλέγονται περιπτώσεις βάσει της πραγματικής κατανομής της

εξαρτημένης μεταβλητής (*στήλη*). Οι συντελεστές τ_b και τ_a είναι ασυμμετρικοί , ανάλογα με ποια μεταβλητή θεωρείται ανεξάρτητη και ορίζονται ως

$$\tau_b\left(Y \mid X\right) = \frac{\sum_j \pi_{.j}\left(1-\pi_{.j}\right) - \sum_i \sum_j \pi_{ij}\left(1-\pi_{ij}/\pi_{i.}\right)}{\sum_j \pi_{.j}\left(1-\pi_{.j}\right)} \tag{4.25}$$

$$\tau_a\left(X \mid Y\right) = \frac{\sum_i \pi_{i.}\left(1-\pi_{i.}\right) - \sum_i \sum_j \pi_{ij}\left(1-\pi_{ij}/\pi_{.j}\right)}{\sum_i \pi_{i.}\left(1-\pi_{i.}\right)} \tag{4.26}$$

$$\tau = \frac{\sum_j \pi_{.j}\left(1-\pi_{.j}\right) + \sum_i \pi_{i.}\left(1-\pi_{i.}\right) - \sum_i \sum_j \pi_{ij}\left(1-\pi_{ij}/\pi_{.j}\right) - \sum_i \sum_j \pi_{ij}\left(1-\pi_{ij}/\pi_{i.}\right)}{\sum_j \pi_{.j}\left(1-\pi_{.j}\right) + \sum_i \pi_{i.}\left(1-\pi_{i.}\right)} \tag{4.27}$$

Οι δειγματικές εκτιμήσεις των συντελεστών $\hat{\tau}_b$, $\hat{\tau}_a$ και $\hat{\tau}$ υπολογίζονται αντικαθιστώντας τις πληθυσμιακές πιθανότητες $\left\{\pi_{ij}\right\}$ με τις αντίστοιχες δειγματικές $\left\{p_{ij}\right\}$.

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 3×4 πίνακα (*σελ. 166, Παράδειγμα 2*), έχουμε

$$\hat{\tau}_b = \frac{0.644-0.592}{0.644} = 0.081$$

$$\hat{\tau}_a = \frac{0.601-0.548}{0.601} = 0.089$$

$$\tau = \frac{0.644+0.601-0.592-0.548}{0.644+0.601} = 0.085$$

Ερμηνεία

Η ερμηνεία των συντελεστών *tau* ταυτίζεται με αυτή των συντελεστών *lambda* .

Πεδίο Ορισμού

Η τιμή των συντελεστών τ_b , τ_a και τ κυμαίνεται στο $\left[0,1\right]$.

Επίπεδο Δεδομένων

Οι συντελεστές τ_b, τ_a και τ μπορούν να χρησιμοποιηθούν με ονοματικά δεδομένα ή και άλλης υψηλότερης κατηγορίας (*διατακτικά*). Όταν οι κατηγορίες των μεταβλητών είναι διατεταγμένες, είναι προτιμότερο να χρησιμοποιούμε εναλλακτικά μέτρα συνάφειας, σχεδιασμένα για διατεταγμένους πίνακες.

Συμμετρία

Οι συντελεστές τ_b, τ_a είναι ασυμμετρικοί, δηλαδή διαφορετικά αποτελέσματα θα προκύψουν, ανάλογα με το ποια θεωρείται ανεξάρτητη μεταβλητή. Όμως όπως είδαμε, η συμμετρική εκδοχή τους είναι ο συντελεστής τ, ο οποίος στην ουσία εκφράζει τον μέσο όρο των δυο ασύμμετρων συντελεστών.

4.5.3 Συντελεστής αβεβαιότητας *U* του *Theil*

Ο συντελεστής αβεβαιότητας *U* (*Theil's Uncertainty Coefficient*, 1972), γνωστός και ως συντελεστής εντροπίας (*entropy coefficient*), ανήκει επίσης στην ομάδα μέτρων αναλογικής μείωσης του σφάλματος πρόβλεψης, αλλά διαφοροποιείται από τον συντελεστή *lambda*, με την έννοια ότι λαμβάνει υπόψη του ολόκληρη την κατανομή της εξαρτημένης μεταβλητής και όχι μόνο την πιο συχνή κατηγορία αυτής (όπου στηρίζεται ο συντελεστής *lambda*).

Έστω,

$H(X) = -\sum_i \pi_{i.} \ln(\pi_{i.})$, η εντροπία του *Shannon* για την περιθώρια κατανομή πιθανότητας της μεταβλητής X

$H(Y) = -\sum_j \pi_{.j} \ln(\pi_{.j})$, η εντροπία του *Shannon* για την περιθώρια κατανομή πιθανότητας της μεταβλητής Y

$H(X,Y) = -\sum_i \sum_j \pi_{ij} \ln(\pi_{ij})$, η από κοινού εντροπία του *Shannon* για την από κοινού κατανομή πιθανότητας των μεταβλητών X, Y.

Υπάρχουν 3 τύποι υπολογισμού για τον συντελεστή. Οι δυο $U_{Y|X}$ και $U_{X|Y}$ είναι ασυμμετρικοί, ανάλογα με το ποια μεταβλητή θεωρείται ανεξάρτητη.

Ο συντελεστής αβεβαιότητας $U_{Y|X}$, ή αλλιώς η αναλογική μείωση της αβεβαιότητας (*εντροπίας*) της μεταβλητής Y, λόγω της γνώσης της X, ορίζεται ως

$$U_{Y|X} = \frac{H(Y)+H(X)-H(X,Y)}{H(X)} = \frac{I(X,Y)}{H(X)}, \tag{4.28}$$

όπου $I(X,Y)$ η αμοιβαία πληροφορία του *Shannon*, ο οποία δίνεται από το μέτρο *Kullback - Leibler*, που θα μελετήσουμε στο Κεφάλαιο 6.

Ο συντελεστής αβεβαιότητας $U_{X|Y}$, ή αλλιώς η αναλογική μείωση της αβεβαιότητας (*εντροπίας*) της μεταβλητής X, λόγω της γνώσης της Y, ορίζεται ως

$$U_{X|Y} = \frac{H(X)+H(Y)-H(X,Y)}{H(Y)} = \frac{I(X,Y)}{H(Y)} \tag{4.29}$$

Ο συμμετρικός συντελεστής αβεβαιότητας U_{SYM}, ορίζεται ως

$$U_{SYM} = \frac{2\left[H(X)+H(Y)-H(X,Y)\right]}{H(X)+H(Y)} \tag{4.30}$$

Οι δειγματικές εκτιμήσεις των συντελεστών $\hat{U}_{Y|X}$, $\hat{U}_{X|Y}$ και $\hat{U}_{SYM}$ υπολογίζονται αντικαθιστώντας τις πληθυσμιακές πιθανότητες $\{\pi_{ij}\}$ με τις αντίστοιχες δειγματικές $\{p_{ij}\}$.

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 3×4 πίνακα (*σελ. 166, Παράδειγμα 2*), έχουμε

$$\hat{U}_{Y|X} = \frac{-(-1.11)+\left[-(-0.98)\right]-\left[-(-2.01)\right]}{-(-0.98)} = 0.085$$

$$\hat{U}_{X|Y} = \frac{-(-0.98)+\left[-(-1.11)\right]-\left[-(-2.01)\right]}{-(-1.11)} = 0.075$$

$$\hat{U}_{SYM} = 0.080$$

Ερμηνεία

Οι ασύμμετρες εκδοχές του συντελεστή U εκφράζουν το ποσοστό ελάττωσης του σφάλματος, που οφείλεται στην διακύμανση της εξαρτημένης μεταβλητής, όπου η διακύμανση ορίζεται σε όρους της εντροπίας *Shannon* και η ερμηνεία της βασίζεται στην θεωρία της πληροφορίας. Εναλλακτικά, μπορούμε να πούμε ότι η τιμή $\hat{U}_{Y|X} = 0.085$ εκφράζει την αναλογία της αβεβαιότητας για την μεταβλητή Y, που εξηγείται από την μεταβλητή X. Η ερμηνεία του συντελεστή $U_{X|Y}$ είναι παρόμοια.

Οι ποσότητες στους αριθμητές των εξισώσεων ερμηνεύονται ως μέτρα της διακύμανσης για ονοματικές μεταβλητές και η μέτρηση της διακύμανσης ονομάζεται εντροπία (*ενώ για τον συντελεστή lambda ονομάζεται Gini Concentration*). Ενώ οι ασύμμετρες εκδοχές του συντελεστή έχουν μια ξεκάθαρη ερμηνεία, με την έννοια της αναλογικής ελάττωσης του σφάλματος, η ερμηνεία του συμμετρικού συντελεστή δεν είναι άλλη από την απλή έκφραση του μέσου όρου των δυο ασύμμετρων συντελεστών. Όπως και στην περίπτωση του συντελεστή *lambda*, η ερμηνεία στηρίζεται στο τι θέλουμε να εξετάσουμε. Έτσι, αν θέλαμε να μελετήσουμε την γνώμη του κοινού αναφορικά με την σχέση μεταξύ χρώματος μαλλιών και ματιών, μπορούμε να πούμε ότι ο $\hat{U}_{SYM} = 0.080$ περιγράφει μια ασήμαντη συνάφεια μεταξύ των μεταβλητών.

Πεδίο Ορισμού

Ο συντελεστής U κυμαίνεται στο διάστημα $[0,1]$. Όταν $U = 0$ η γνώση της ανεξάρτητης μεταβλητής δεν προσφέρει καμιά προβλεπτική αξία στην εκτίμηση της εξαρτημένης.

Επίπεδο Δεδομένων

Ο συντελεστής U μπορεί να χρησιμοποιηθεί με ονοματικά δεδομένα ή και άλλης υψηλότερης κατηγορίας (*διατακτικά*), για οποιοδήποτε $I \times J$ πίνακα. Όταν οι κατηγορίες των μεταβλητών είναι διατεταγμένες, δεν προτείνεται να χρησιμοποιούμε τον συντελεστή U, καθώς χάνεται πληροφόρηση.

Συμμετρία

Ο συντελεστής U είναι ένα ασύμμετρο μέτρο καθώς παίρνει διαφορετικές τιμές ανάλογα με το ποια μεταβλητή ορίζεται ως εξαρτημένη και ποια ως ανεξάρτητη. Η συμμετρική εκδοχή του εκφράζεται με τον μέσο όρο των δυο ασύμμετρων συντελεστών.

Σημαντικότητα

Επειδή η δειγματική κατανομή του U είναι γνωστή (*κανονική ασυμπτωματική κατανομή*), μπορούμε να υπολογίσουμε το ασυμπτωτικό τυπικό του σφάλμα και την σημαντικότητα. Καθώς οι τύποι υπολογισμού της διακύμανσης είναι πολύπλοκοι και ξεφεύγουν από τους σκοπούς της παρούσας εργασίας, παραπέμπουμε τον αναγνώστη στο εγχειρίδιο SAS/STAT User's Guide για αναφορά [*SAS/STAT User's Guide, Version 8, Chapter 28, p. 1297-1298*, (1999)].

Σχόλια

Όπως αναφέραμε, ο συντελεστής αβεβαιότητας U διαφοροποιείται από τον συντελεστή *lambda*, με την έννοια ότι λαμβάνει υπόψη του ολόκληρη την κατανομή της εξαρτημένης μεταβλητής και όχι μόνο την πιο συχνή κατηγορία αυτής (*όπου στηρίζεται ο συντελεστής* λ). Για τον λόγο αυτό κάποιες φορές είναι προτιμότερη η χρήση του. Αξίζει να σημειώσουμε ότι για τον συντελεστή U υπάρχει μια θετική σχέση μεταξύ του αριθμού των κατηγοριών και της βαρύτητας της διακύμανσης, εισάγοντας έτσι διφορούμενες ερμηνείες κατά την αξιολόγηση και της διακύμανσης των μεταβλητών και της μεταξύ τους σχέσης [*Hershberger & Fisher*, (2005)].

4.6 Συμπεράσματα

Στο Κεφάλαιο αυτό εξετάσαμε 18 μέτρα συνάφειας τα οποία χρησιμοποιούνται για ονοματικές μεταβλητές. Τα μέτρα αυτά καλύπτουν μια διαδρομή μισού αιώνα και στηρίζονται σε διαφορετικές υποθέσεις αναφορικά με τον τρόπο υπολογισμού τους. Επομένως, οι τιμές θα πρέπει να ερμηνεύονται διαφορετικά. Είναι γεγονός, ότι για κάθε εξεταζόμενη περίπτωση, πολλά από τα προτεινόμενα μέτρα είναι έγκυρα και το πρόβλημα που αντιμετωπίζουν πολλοί ερευνητές είναι πιο μέτρο συνάφειας να επιλέξουν. Συνοψίζοντας τις τιμές των εξεταζόμενων μέτρων, βασιζόμενοι στο Παράδειγμα 1 (*σελ. 165)*, ενός 2×2 πίνακα και στο Παράδειγμα 2 (*σελ. 166*) ενός 3×4 πίνακα, θα προσπαθήσουμε να καταλήξουμε σε κάποια συμπεράσματα. Μελετώντας τον παρακάτω συνοπτικό Πίνακα 4-2 (*σελ. 72*), όλων των εξεταζόμενων μέτρων συνάφειας για το Παράδειγμα 1, συμπεραίνουμε ότι δεν υπάρχει συνάφεια μεταξύ των μεταβλητών, καθώς όλα τα μέτρα τείνουν στο 0. Δηλαδή η «*Επιβίωση*» δεν εξαρτάται από την «*Θεραπεία*».

Σημειώνουμε ότι για τον Σχετικό Κίνδυνο (RR) και το *odds ratio* η ανεξαρτησία υποδηλώνεται όταν οι τιμές ισούνται με 1.

ΠΙΝΑΚΑΣ 4-2

Συνοπτικός πίνακας μέτρων συνάφειας για ονοματικές μεταβλητές

	Παράδειγμα 1 (2x2 πίνακας)		
	Μέτρο συνάφειας	Αποτέλεσμα	Στρογγυλοποίηση
1	Relative Risk	0,800	0,80
2	Odds Ratio	0,762	0,76
3	Yule's Q	-0,135	-0,14
4	Yule's Y	-0,068	-0,07
5	Yule's φ	-0,052 0,052	-0,05 0,05
6	Tshuprow's T	0,052	0,05
7	Cramer's V	0,052	0,05
8	Pearson's C	0,052	0,05
9	Sakoda's C_{adj}	0,074	0,07
10	Goodman-Kruskal's λ_{α}	0,040	0,04
11	Goodman-Kruskal's λ_{β}	0,000	0,00
12	Goodman-Kruskal's λ	0,029	0,03
13	Goodman-Kruskal's τ_{α}	0,003	0,00
14	Goodman-Kruskal's τ_{β}	0,003	0,00
15	Goodman-Kruskal's τ	0,003	0,00
16	Theil's U(Y|X)	0,002	0,00
17	Theil's U(X|Y)	0,003	0,00
18	Theil's U_{sym}	0,002	0,00

Παρατηρούμε ότι, αν και τα μέτρα δεν πετυχαίνουν την ίδια ακριβώς τιμή (ποικίλουν με εύρος απόλυτων τιμών από 0 έως 0.14), μπορούμε να πούμε ότι αρκετά από αυτά ταυτίζονται και τα περισσότερα πετυχαίνουν πολύ κοντινές τιμές. Συγκεκριμένα, τα μέτρα $4-12$ πετυχαίνουν πολύ κοντινές τιμές, εκ των οποίων τα μέτρα $5-9$ που βασίζονται στο χ^2-test έχουν ίδιες τιμές, με εξαίρεση τον προσαρμοσμένο συντελεστή συνάφειας C_{adj} του *Sakoda*. Αξίζει να σημειώσουμε ότι, αν και τα μέτρα 10 και 12 έχουν μια διαφορετική φιλοσοφία, με την έννοια της αναλογικής

μείωσης του σφάλματος πρόβλεψης, οι τιμές τους βρίσκονται πολύ κοντά με τα παραδοσιακά μέτρα συνάφειας. Τα μέτρα 11 και 13−18 που είναι επίσης μέτρα αναλογικής μείωσης του σφάλματος πρόβλεψης, επίσης ταυτίζονται, αλλά απομακρύνονται περισσότερο από τα υπόλοιπα μέτρα. Τέλος, το μέτρο 10 (*Yule's Q*) παίρνει την μεγαλύτερη τιμή, καθώς όπως έχουμε αναφέρει έχει την τάση να αυξάνει τον βαθμό της συνάφειας.

Μελετώντας τον παρακάτω συνοπτικό Πίνακα 4-3 (*σελ. 73*), όλων των εξεταζόμενων μέτρων συνάφειας για το Παράδειγμα 2 (*σελ. 166*), συμπεραίνουμε ότι τα δεδομένα διέπονται από συνάφεια μεγαλύτερης έντασης.

ΠΙΝΑΚΑΣ 4-3

Συνοπτικός πίνακας μέτρων συνάφειας για ονοματικές μεταβλητές

Παράδειγμα 2 (3x4 πίνακας)			
A/A	Μέτρο συνάφειας	Αποτέλεσμα	Στρογγυλοποίηση
1	Relative Risk	-	-
2	Odds Ratio	-	-
3	Yule's Q	-	-
4	Yule's Y	-	-
5	Yule's φ	0,397	0,40
6	Tshuprow's T	0,254	0,25
7	Cramer's V	0,281	0,28
8	Pearson's C	0,369	0,37
9	Sakoda's C_{adj}	0,452	0,45
10	Goodman-Kruskal's λ_{α}	0,224	0,22
11	Goodman-Kruskal's λ_{β}	0,192	0,19
12	Goodman-Kruskal's λ	0,208	0,21
13	Goodman-Kruskal's τ_{α}	0,089	0,09
14	Goodman-Kruskal's τ_{β}	0,081	0,08
15	Goodman-Kruskal's τ	0,085	0,08
16	Theil's U(Y|X)	0,085	0,09
17	Theil's U(X|Y)	0,075	0,08
18	Theil's U_{sym}	0,080	0,08

Όμως, οι τιμές των μέτρων διαφοροποιούνται αρκετά (ποικίλουν με εύρος τιμών από 0.08 έως 0.45), προκαλώντας κάποιο εντονότερο προβληματισμό, ως προς το αν υπάρχει συνάφεια και τι έντασης. Σύμφωνα με τις παραδεκτές (αν και αυθαίρετες κατά μια έννοια) παραδοχές του Cohen (1977, 1988), που περιγράφονται στον παρακάτω Πίνακα 4-4 (*σελ. 74*), βλέπε [*Harry Khamis, Measures of association: How to Choose?,* (2008)] και [*Handbook of Parametric and Non-Parametric Statistical Procedures, 3rd Edition, Chapman & Hall,* (2004)], ισχύει ότι

ΠΙΝΑΚΑΣ 4-4

Ένταση της συνάφειας σύμφωνα με τις παραδοχές του Cohen

Τιμή μέτρου (m)		Ένταση συνάφειας
m < 0.3	→	αδύναμη
0.3 < m ≤ 0.5	→	μέτρια
m > 0.5	→	ισχυρή

Cohen (1977, 1988)

Αρχικά, παρατηρούμε ότι τα μέτρα $5-9$ που βασίζονται στο χ^2-test διαφοροποιούνται καθώς δεν αναφερόμαστε σε 2×2 πίνακες, εκ των οποίων τα 5 και 9 πετυχαίνουν τις ανώτερες τιμές 0.40 και 0.45, αντίστοιχα. Το μέτρο 5 όμως, γενικά δεν κρίνεται τόσο κατάλληλο καθώς μπορεί να πάρει τιμές >1. Επίσης, το μέτρο 6 δεν κρίνεται κατάλληλο καθώς χρησιμοποιείται για τετραγωνικούς πίνακες. Αξίζει να σημειώσουμε, ότι τα μέτρα προβλεπτικής συνάφειας $10-12$, παίρνουν τιμές κοντινές του μέτρου 7, που θεωρείται το πιο διαδεδομένο και κατανέμεται πιο κανονικά στο $[0,1]$, επιτρέποντας μας να ισχυριστούμε ότι υπάρχει μια αδύναμη εως μέτρια συνάφεια μεταξύ των μεταβλητών χρώματος ματιών και χρώματος μαλλιών. Τέλος, τα μέτρα $13-18$ υποδηλώνουν μια ασήμαντη συνάφεια, θέτοντας έτσι κάποιους προβληματισμούς για το τι τελικά θα συμπεράνουμε. Ακολουθεί ο συγκεντρωτικός Πίνακας 4-5 (*σελ. 75*), με τα βασικά χαρακτηριστικά και κάποια βασικά σχόλια για τα ονοματικά μέτρα συνάφειας που εξετάσαμε στο Κεφάλαιο αυτό.

ΠΙΝΑΚΑΣ 4-5

Συγκεντρωτικός πίνακας μέτρων συνάφειας για ονοματικές μεταβλητές

	Μέτρο συνάφειας	Πεδίο Ορισμού	Συμμετρία	Είδος δεδομένων	Διάσταση πίνακα	Σχόλια
1.	Relative Risk (RR)	[0,+oo)	ΟΧΙ	ονοματικά και διατακτικά	2x2	Χρησιμοποιείται ευρέως στις ιατρικές επιστήμες ως παράγοντας κινδύνου, για την σύγκριση ομάδων. Για RR=1 τότε ανεξαρτησία
2.	Odds Ratio (θ)	(-oo,+oo)	ΝΑΙ	ονοματικά και διατακτικά	2x2	Συνδέεται με τον σχετικό κίνδυνο μέσω σχέσης και μπορεί να χρησιμοποιηθεί για πίνακες μεγαλύτερης διάστασης. Για θ = 1 τότε ανεξαρτησία.
3.	Yule's Q	[-1,+1]	ΝΑΙ	ονοματικά και διατακτικά	2x2	Βασίζεται στο odds ratio, αποτελεί ειδική περίπτωση του διατακτικού μέτρου gamma. Δεν συνιστάται η χρήση του όταν η συχνότητα των κελιών είναι πολύ μικρή. Έχει την τάση να αυξάνει τον βαθμό της συνάφειας.
4.	Yule's Y	[-1,+1]	ΝΑΙ	ονοματικά και διατακτικά	2x2	Γενικά θα είναι μικρότερος του Yule's Q, είναι λιγότερο ευαίσθητος από τις διαφορές των περιθώριων κατανομών από ότι ο Q. Ερμηνεύεται όπως ο συντελεστής συσχέτισης r του Pearson (για συνεχή δεδομένα).
5.	Yule's φ	[-1,+1] ή [0, sqrt(q-1)]	ΝΑΙ	ονοματικά	2x2 ή *I* x *J*	Δεν θεωρείται το πιο κατάλληλο μέτρο καθώς η τιμή του μπορεί να ξεπεράσει την τιμή 1. Είναι πολύ ευαίσθητος στις περιθώριες κατανομές και η σύγκριση μεταξύ πινάκων μπορεί να είναι παραπλανητική. Αλγεβρικά ισούται με τον συντελεστή συσχέτισης r του Pearson.
6.	Tshuprow's T	[0,1]	ΝΑΙ	ονοματικά	*I* x *I*	Χρησιμοποιείται μόνο σε τετραγωνικούς πίνακες. T = φ για 2x2 πίνακες. Η χρήση του είναι περιορισμένη.
7.	Cramer's V	[0,1]	ΝΑΙ	ονοματικά	*I* x *J*	V = T = φ για 2x2 πίνακες. Θεωρείται το πιο διαδεδομένο μέτρο, διότι ανεξάρτητα από το μέγεθος του πίνακα κατανέμεται πιο κανονικά στο [0,1]. Εξαρτάται από την διάσταση του πίνακα και χρειάζεται προσοχή όταν συγκρίνουμε πίνακες. Επίσης $V = \tau^2$.
8.	Pearson's C	[0,1] ή [0,0.71]	ΝΑΙ	ονοματικά	*I* x *J*	Για 2x2 πίνακες κυμαίνεται στο [0,0.71]. Θα είναι <1 ακόμα και όταν οι μεταβλητές είναι πλήρως εξαρτημένες. Εξαρτάται από την διάσταση του πίνακα και χρειάζεται προσοχή όταν συγκρίνουμε πίνακες. Προτείνεται να χρησιμοποιείται για πίνακες > 5x5.
9.	Sakoda's C_{adj}	[0,1]	ΝΑΙ	ονοματικά	*I* x *J*	-
10.	Goodman-Kruskal's λ_α	[0,1]	ΟΧΙ	ονοματικά	*I* x *J*	Δεν ορίζεται όταν οι παρατηρήσεις συγκεντρώνονται σε μια από τις στήλες της εξαρτημένης μεταβλητής. Είναι ευαίσθητος στις ανισότητες των περιθώριων αθροισμάτων γραμμών και στηλών. Μπορεί $\lambda_\alpha = 0$ χωρίς να υπάρχει στατιστική ανεξαρτησία.
11.	Goodman-Kruskal's λ_β	[0,1]	ΟΧΙ	ονοματικά	I x *J*	Δεν ορίζεται όταν οι παρατηρήσεις συγκεντρώνονται σε μια από τις στήλες της εξαρτημένης μεταβλητής. Είναι ευαίσθητος στις ανισότητες των περιθώριων αθροισμάτων γραμμών και στηλών. Μπορεί $\lambda_\beta = 0$ χωρίς να υπάρχει στατιστική ανεξαρτησία.
12.	Goodman-Kruskal's λ	[0,1]	ΝΑΙ	ονοματικά	I x *J*	Δεν ορίζεται όταν οι παρατηρήσεις συγκεντρώνονται σε ένα μοναδικό κελί του πίνακα. Είναι ευαίσθητος στις ανισότητες των περιθώριων αθροισμάτων γραμμών και στηλών. Μπορεί λ = 0 χωρίς να υπάρχει στατιστική ανεξαρτησία.
13.	Goodman-Kruskal's τ_α	[0,1]	ΟΧΙ	ονοματικά	*I* x *J*	-
14.	Goodman-Kruskal's τ_β	[0,1]	ΟΧΙ	ονοματικά	*I* x *J*	-
15.	Goodman-Kruskal's τ	[0,1]	ΟΧΙ	ονοματικά	*I* x *J*	-
16.	Theil's U(Y|X)	[0,1]	ΟΧΙ	ονοματικά	*I* x *J*	Η ερμηνεία τους βασίζεται στην θεωρία της πληροφορίας και στην εντροπία Shannon. Διαφοροποιούνται από τον συντελεστή λ με την έννοια ότι λαμβάνει υπόψη του όλη την κατανομή της εξαρτημένης μεταβλητής. Ένα μειονέκτημα είναι ότι υπάρχει μια θετική σχέση μεταξύ του αριθμού των κατηγοριών και της βαρύτητας της διακύμανσης, εισάγοντας διφορούμενες ερμηνείες κατα την αξιολόγηση της διακύμανσης των μεταβλητών και της μεταξύ τους σχέσης.
17.	Theil's U(X|Y)	[0,1]	ΟΧΙ	ονοματικά	*I* x *J*	
18.	Theil's U_{sym}	[0,1]	ΝΑΙ	ονοματικά	*I* x J	

ΚΕΦΑΛΑΙΟ 5

ΜΕΤΡΑ ΣΥΝΑΦΕΙΑΣ ΓΙΑ ΔΙΑΤΑΚΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ

5.1 Εισαγωγή

Στο κεφάλαιο αυτό θα παρουσιάσουμε τα κυριότερα μέτρα διατακτικής συνάφειας, για διδιάστατους $I \times J$ πίνακες συνάφειας. Τα διατακτικά μέτρα συνάφειας αξιοποιούν την επιπρόσθετη πληροφορία της διάταξης των μεταβλητών και εξετάζουν εάν η μεταβλητή Y έχει την τάση να αυξάνεται, όταν μεταβλητή X αυξάνεται και το αντίστροφο. Τα μέτρα αυτά είναι κατάλληλα για διατακτικές μεταβλητές και ταξινομούν τα ζεύγη των παρατηρήσεων, σε «*σύμφωνα*» (*Concordant*) και σε «*ασύμφωνα*» (*Discordant*) ζεύγη καθώς και σε ζεύγη με ισοβαθμίες (*Ties*). Οι τιμές των μέτρων κυμαίνονται στο διάστημα $[-1,+1]$, προσδιορίζοντας την κατεύθυνση της συνάφειας. Η τιμή 0 υποδηλώνει ανυπαρξία συνάφειας μεταξύ των δυο μεταβλητών και οι τιμές ± 1, υποδηλώνουν πλήρης αρνητική ή θετική συνάφεια. Γενικά, τα διατακτικά μέτρα συνάφειας, διακρίνονται βάσει του τρόπου υπολογισμού τους, σε δυο κύριες κατηγορίες:

Α. μέτρα που βασίζονται στις έννοιες «*σύμφωνα*» και «*ασύμφωνα*» ζεύγη και ζεύγη που ισοβαθμούν ή αλλιώς ζεύγη με «*δεσμούς*»

Β. μέτρα που βασίζονται στην ανάθεση σκορ (*score*)

5.2 Διατακτικά μέτρα συνάφειας «σύμφωνων» και «ασύμφωνων» ζευγών

Έστω ένα δείγμα n παρατηρήσεων. Επιλέγουμε τυχαία 2 ανεξάρτητα ζεύγη παρατηρήσεων (X_1,Y_1) και (X_2,Y_2), με επανάθεση. Τα ζεύγη αυτά μπορούν να σχηματίσουν 2 συνδυασμούς $\{(X_1,Y_1),(X_2,Y_2)\}$ και $\{(X_2,Y_2),(X_1,Y_1)\}$, οι οποίοι διαφέρουν στην διάταξή τους αλλά είναι πανομοιότυποι. Συνολικά υπάρχουν $\binom{n}{2} = n(n-1)/2$ τέτοια ξεχωριστά ζεύγη.

Σύμφωνα ζεύγη

Ένα ζεύγος παρατηρήσεων θεωρείται «*σύμφωνο*» ή «*εναρμονισμένο*» (*Concordant*) αν $X_1 < X_2$ και $Y_1 < Y_2$ ή αν $X_1 > X_2$ και $Y_1 > Y_2$.

Επομένως, ένα ζεύγος είναι «*σύμφωνο*» όταν μια παρατήρηση που ταξινομείται σε μια χαμηλότερη (ή υψηλότερη) κατηγορία της μεταβλητής X, ταξινομείται επίσης χαμηλότερα (ή υψηλότερα) για την κατηγορία της μεταβλητής Y, δηλαδή όταν οι τιμές του ζεύγους έχουν την ίδια διάταξη.

Ο διπλάσιος αριθμός των σύμφωνων ζευγών συμβολίζεται με C (*Concordant*) και υπολογίζεται από τον τύπο

$$C = \sum_i \sum_j n_{ij} A_{ij}, \tag{5.1}$$

όπου $A_{ij} = \sum_{i>i'} \sum_{j>j'} n_{i'j'} + \sum_{i<i'} \sum_{j<j'} n_{i'j'}$, ο αριθμός των σύμφωνων ζευγών, με $i, i' = 1, 2, ..., I$ και $j, j' = 1, 2, ..., J$.

Ασύμφωνα ζεύγη

Ένα ζεύγος παρατηρήσεων θεωρείται «*ασύμφωνο*» ή «*μη – εναρμονισμένο*» (*Discordant*) αν $X_1 < X_2$ και $Y_1 > Y_2$ ή αν $X_1 > X_2$ και $Y_1 < Y_2$.

Επομένως, ένα ζεύγος παρατηρήσεων είναι «*ασύμφωνο*» όταν μια παρατήρηση που ταξινομείται σε μια υψηλότερη (ή χαμηλότερη) κατηγορία της μεταβλητής X, ταξινομείται σε μια χαμηλότερη (ή υψηλότερη) κατηγορία της μεταβλητής Y, δηλαδή όταν οι τιμές του μέτρου έχουν αντίστροφη διάταξη.

Ο διπλάσιος αριθμός των «*ασύμφωνων*» ζευγών συμβολίζεται με D (*Discordant*) και υπολογίζεται από τον τύπο

$$D = \sum_i \sum_j n_{ij} Q_{ij} \tag{5.2}$$

όπου $Q_{ij} = \sum_{i>i'} \sum_{j<j'} n_{i'j'} + \sum_{i<i'} \sum_{j>j'} n_{i'j'}$, ο αριθμός των ασύμφωνων ζευγών, με $i, i' = 1, 2, ..., I$ και $j, j' = 1, 2, ..., J$.

Ζεύγη που ισοβαθμούν

Ένα ζεύγος παρατηρήσεων «*ισοβαθμεί*», ή αλλιώς θεωρείται ότι έχει «*ισοβαθμίες*» (*Ties*) ως προς την μεταβλητή X, όταν $X_1 = X_2$ και $Y_1 > Y_2$ ή όταν $X_1 = X_2$ και $Y_1 < Y_2$. Ο αριθμός των ζευγών με «*ισοβαθμίες*» ως προς την μεταβλητή X, συμβολίζεται με T_X και υπολογίζεται από τον τύπο

$$T_X = \sum_i n_{i.}\left(n_{i.} - 1\right)/2 \tag{5.3}$$

Ένα ζεύγος παρατηρήσεων έχει «*ισοβαθμίες*» ως προς την μεταβλητή Y όταν $Y_1 = Y_2$ και $X_1 > X_2$ ή όταν $Y_1 = Y_2$ και $X_1 < X_2$. Ο αριθμός των ζευγών με «*ισοβαθμίες*» ως προς την μεταβλητή Y, συμβολίζεται με T_Y και υπολογίζεται από τον τύπο

$$T_Y = \sum_j n_{.j}\left(n_{.j} - 1\right)/2 \tag{5.4}$$

Τέλος ένα ζεύγος παρατηρήσεων έχει «*ισοβαθμίες*» ως προς τις μεταξύ του μεταβλητές όταν $X_1 = X_2$ και $Y_1 = Y_2$. Ο αριθμός των ζευγών με δεσμούς ως προς την μεταβλητή X και Y, δηλαδή ο αριθμός των κοινών κελιών, συμβολίζεται με T_{XY} και υπολογίζεται από τον τύπο

$$T_{XY} = \sum_i \sum_j n_{ij}\left(n_{ij} - 1\right)/2 \tag{5.5}$$

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 3×3 πίνακα συνάφειας (*σελ. 167, Παράδειγμα 3*) έχουμε

$$\hat{C} = \sum_i \sum_j n_{ij} A_{ij} = 302\left(331 + 250 + 155 + 185\right) + 105\left(250 + 185\right) + 409\left(155 + 185\right) + 331\left(185\right) =$$

$$= 524.112$$

Σημειώνουμε, ότι η διαδικασία προσδιορισμού των «*σύμφωνων*» ζευγών είναι, αφού επιλέξουμε ένα κελί , για παράδειγμα $n_{11} = 302$, πολλαπλασιάζουμε την συχνότητά του με την συχνότητα κάθε ενός κελιού που βρίσκεται νοτιοανατολικά αυτού.

$$\hat{D} = \sum_i \sum_j n_{ij} Q_{ij} = 23\left(409 + 331 + 15 + 155\right) + 105\left(409 + 15\right) + 250\left(15 + 155\right) + 331\left(15\right) =$$

$$= 112.915$$

Σημειώνουμε ότι η διαδικασία προσδιορισμού των «*ασύμφωνων*» ζευγών είναι, αφού επιλέξουμε ένα κελί, για παράδειγμα $n_{13}=23$, πολλαπλασιάζουμε την συχνότητά του με την συχνότητα κάθε ενός κελιού που βρίσκεται νοτιοδυτικά αυτού.

$$\hat{T}_X=\sum_i n_{i.}\left(n_{i.}-1\right)/2=\frac{\left(430\times 429\right)+\left(990\times 989\right)+\left(355\times 354\right)}{2}=644.625$$

$$\hat{T}_Y=\sum_j n_{j.}\left(n_{j.}-1\right)/2=\frac{\left(726\times 725\right)+\left(591\times 590\right)+\left(458\times 457\right)}{2}=542.173$$

$$\hat{T}_{XY}=\sum_i\sum_j n_{ij}\left(n_{ij}-1\right)/2=\frac{\left(302\times 301\right)+\left(105\times 104\right)+\ldots+\left(185\times 184\right)}{2}=249.400$$

Όπως έχουμε αναφέρει, ο συνολικός αριθμός των ζευγών των παρατηρήσεων είναι $\left\{n\left(n-1\right)/2\right\}$, ο οποίος μπορεί να εκφρασθεί μέσω των παραπάνω σχέσεων και ως $n\left(n-1\right)/2=C+D+T_X+T_Y-T_{XY}$.

Σημειώνουμε ότι η ποσότητα T_{XY} αφαιρείται καθώς τα ζεύγη με «*ισοβαθμίες*» στις μεταβλητές X και Y, έχουν μετρηθεί δύο φορές, μία κατά τον υπολογισμό των T_X και μια για τον υπολογισμό των T_Y.

5.2.1 Συντελεστής γ των *Goodman - Kruskal*

Ο συντελεστής διατακτικής συνάφειας γ ή αλλιώς *gamma* [*Goodman & Kruskal*, (1954)], είναι ένα μέτρο που βασίζεται στην διαφορά των «*σύμφωνων*» και «*ασύμφωνων*» ζευγών, αγνοώντας τις αναμεταξύ τους «*ισοβαθμίες*». Ο τύπος υπολογισμού του μέτρου είναι

$$\gamma=\frac{C-D}{C+D} \tag{5.6}$$

και η δειγματική εκτίμηση υπολογίζεται από τον τύπο

$$\hat{\gamma}=\frac{\hat{C}-\hat{D}}{\hat{C}+\hat{D}} \tag{5.7}$$

όπου $\hat{C}$, $\hat{D}$ ο αριθμός των «*σύμφωνων*» και «*ασύμφωνων*» ζευγών του δείγματος.

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 3×3 πίνακα συνάφειας (*σελ. 167, Παράδειγμα 3*)

$$\hat{\gamma} = \frac{524.112 - 112.915}{524.112 - 112.915} = 0.645$$

Ερμηνεία

Ο συντελεστής γ εκφράζει το πλεόνασμα των «*σύμφωνων*» ζευγών έναντι των «*ασύμφωνων*», ως ποσοστό επί όλων των ζευγών χωρίς «*ισοβαθμίες*». Θετικές τιμές υποδηλώνουν ότι $C > D$, δηλαδή υποκείμενα που ανήκουν σε υψηλότερες κατηγορίες της μεταβλητής X, ανήκουν επίσης σε υψηλότερες κατηγορίες της μεταβλητής Y, αποδίδοντας μια θετική συνάφεια. Αρνητικές τιμές υποδηλώνουν ότι $C < D$, δηλαδή υποκείμενα που ανήκουν σε υψηλότερες κατηγορίες της μεταβλητής X, ανήκουν σε χαμηλότερες κατηγορίες της μεταβλητής Y, αποδίδοντας μια αρνητική συνάφεια. Το μέτρο γ παρέχει επίσης, μια ερμηνεία αναλογικής μείωσης του σφάλματος. Αν αγνοήσουμε τα ζεύγη με «*ισοβαθμίες*» και προσπαθήσουμε να προβλέψουμε την τάξη (*rank*) δυο ζευγών, δοθείσης της ανεξάρτητης μεταβλητής X, τότε γνωρίζοντας ότι $X_1 < X_2$, θα μπορούσαμε να πούμε ότι $Y_1 < Y_2$. Σύμφωνα με το παράδειγμά μας, $\hat{\gamma} = 0.645$ και επομένως μπορούμε να πούμε ότι, γνωρίζοντας την ανεξάρτητη μεταβλητή X, έχουμε μείωση του σφάλματος για την πρόβλεψη της τάξης (και όχι της τιμής) της εξαρτημένης μεταβλητής, κατά 64.5%.

Πεδίο Ορισμού

Ο συντελεστής γ κυμαίνεται στο διάστημα $[-1, +1]$. Για $\gamma = 1$ δεν υπάρχουν «*ασύμφωνα*» ζεύγη και η διάταξη των X είναι σε πλήρης συμφωνία με την διάταξη των Y, δηλαδή, μεγαλύτερα (ή μικρότερα) X αντιστοιχούν σε μεγαλύτερα (ή μικρότερα) Y, αντίστοιχα. Τα δεδομένα συγκεντρώνονται στα άνω αριστερά και κάτω δεξιά σημεία της διαγώνιου του πίνακα. Για $\gamma = -1$, δεν υπάρχουν «*σύμφωνα*» ζεύγη και η διάταξη των X είναι σε πλήρης συμφωνία με την αντίθετη διάταξη των Y, δηλαδή μεγαλύτερα (ή μικρότερα) X αντιστοιχούν σε μικρότερα (ή μεγαλύτερα) Y, αντίστοιχα. Τα δεδομένα συγκεντρώνονται στα κάτω αριστερά και άνω δεξιά σημεία της διαγώνιου του πίνακα. Όταν $\gamma = 0$, οι δυο μεταβλητές είναι ανεξάρτητες. Όμως δυο μεταβλητές μπορεί να είναι πλήρως εξαρτημένες αλλά η τιμή του *gamma* να είναι μικρότερη της μονάδος.

Επίπεδο Δεδομένων

Ο συντελεστής γ χρησιμοποιείται για διατακτικές μεταβλητές και εφαρμόζεται σε οποιοδήποτε $I \times J$ πίνακα συνάφειας.

Συμμετρία

Ο συντελεστής γ είναι συμμετρικό μέτρο συνάφειας καθώς δεν μεταβάλλεται σε αλλαγές στην διάταξη γραμμών και στηλών, δηλαδή όταν ο πίνακας αλλάξει προσανατολισμό και οι γραμμές γίνουν στήλες και οι στήλες, γραμμές. Δεν έχει σημασία για το αποτέλεσμα ποια θα είναι η ανεξάρτητη μεταβλητή.

Άλλα χαρακτηριστικά

Όπως έχουμε αναφέρει, στην περίπτωση 2×2 πινάκων, $\gamma = Q$, όπου Q ο συντελεστής συνάφειας του *Yule* (*Παράγραφος 4.3.1 σελ. 44*). Επίσης, το μέτρο D του *Somers*, που θα εξετάσουμε στην συνέχεια μπορεί να θεωρηθεί η ασυμμετρική εκδοχή του μέτρου γ.

Σχόλια

Ο συντελεστής γ δεν ορίζεται όταν όλες οι παρατηρήσεις βρίσκονται σε μια γραμμή ή μια στήλη του πίνακα. Επίσης, δυο μεταβλητές μπορεί να είναι πλήρως εξαρτημένες, αλλά να ισχύει $\gamma < 1$.

5.2.2 Συντελεστές *tau* του *Kendall*

Τα μέτρα συνάφειας τ ή *tau* [*Kendall*, (1938)], συγκαταλέγονται στα πιο γνωστά διατακτικά μέτρα συνάφειας. Υπάρχουν 3 διαφορετικές παραλλαγές των μέτρων και συμβολίζονται με τ_a, τ_b και τ_c. Το μέτρο τ_a υποθέτει ότι δεν υπάρχουν ζεύγη παρατηρήσεων με «*ισοβαθμίες*», ενώ οι άλλες δυο παραλλαγές διαφέρουν μεταξύ τους ως προς τον τρόπο που διαχειρίζονται τις «*ισοβαθμίες*». Οι τύποι υπολογισμού των μέτρων είναι

$$\tau_a = \frac{C - D}{n(n-1)/2} \tag{5.8}$$

$$\tau_b = \frac{C - D}{\sqrt{\left[\left(n(n-1)/2\right) - T_X\right]\left[\left(n(n-1)/2\right) - T_Y\right]}} \tag{5.9}$$

$$\tau_c = \frac{2q(C-D)}{n^2(q-1)} \tag{5.10}$$

όπου $q = \min(I, J)$

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 3×3 πίνακα συνάφειας (*σελ. 167, Παράδειγμα 3*) έχουμε

$$\hat{\tau}_a = \frac{542.112 - 112.915}{1.574.425} = 0.261$$

$$\hat{\tau}_b = \frac{542.112 - 112.915}{\sqrt{(1.574.425 - 644.625)*(1.574.425 - 542.173)}} = 0.420$$

$$\hat{\tau}_c = \frac{2\times3\times(542.112 - 112.915)}{1775^2(2-1)} = 0.392$$

Ερμηνεία

Σύμφωνα με τα δεδομένα του παραδείγματός μας, διαπιστώνουμε ότι υπάρχει θετική συνάφεια μέτριας έντασης. Ο συντελεστής τ_a εκφράζει το πλεόνασμα των «*σύμφωνων*» ζευγών έναντι των «*ασύμφωνων*», ως ποσοστό του συνολικού αριθμού των ζευγών του δείγματος. Το μέτρο δεν κάνει διορθώσεις για τις «*ισοβαθμίες*» μεταξύ των παρατηρήσεων.

Ο συντελεστής τ_b δεν έχει μια εύκολη, διαισθητική ερμηνεία. Μπορούμε να πούμε ότι εκφράζει το πλεόνασμα των «*σύμφωνων*» ζευγών έναντι των «*ασύμφωνων*», ως ποσοστό του γεωμετρικού μέσου του αριθμού των ζευγών χωρίς «*ισοβαθμίες*» για την μεταβλητή X και χωρίς «*ισοβαθμίες*» για την μεταβλητή Y. Αξίζει να σημειώσουμε ότι το μέτρο κάνει διορθώσεις για τις «*ισοβαθμίες*» μεταξύ των παρατηρήσεων.

Ο συντελεστής τ_c επίσης δεν έχει μια εύκολη, διαισθητική ερμηνεία. Μπορούμε να πούμε ότι εκφράζει το πλεόνασμα των «*σύμφωνων*» ζευγών έναντι των «*ασύμφωνων*», ως ποσοστό μιας ποσότητας που βασίζεται στο μέγεθος του πίνακα.

Πεδίο Ορισμού

Η τιμή των συντελεστών τ_a, τ_b και τ_c κυμαίνονται στο διάστημα $[-1+1]$. Πετυχαίνουν την τιμή 1, όταν δεν υπάρχουν «*ασύμφωνα*» ζεύγη και η διάταξη των X βρίσκεται σε πλήρης

συμφωνία με την διάταξη των Y, ενώ πετυχαίνουν την τιμή -1 όταν δεν υπάρχουν «*σύμφωνα*» ζεύγη και η διάταξη των X είναι σε πλήρης συμφωνία με την αντίθετη διάταξη των Y.

Επίπεδο Δεδομένων

Οι συντελεστές τ_a, τ_b και τ_c χρησιμοποιούνται με διατακτικές μεταβλητές και εφαρμόζονται σε οποιοδήποτε $I \times J$ πίνακα συνάφειας.

Συμμετρία

Οι συντελεστές τ_a, τ_b και τ_c είναι συμμετρικά μέτρα συνάφειας καθώς δεν μεταβάλλονται σε αλλαγές στην διάταξη γραμμών και στηλών, δηλαδή όταν ο πίνακας αλλάξει προσανατολισμό και οι γραμμές γίνουν στήλες και οι στήλες, γραμμές. Δεν έχει σημασία για το αποτέλεσμα ποια θα είναι η ανεξάρτητη μεταβλητή. Το μέτρο D του *Somers*, που θα εξετάσουμε στην συνέχεια μπορεί να θεωρηθεί η ασυμμετρική εκδοχή τους.

Σημαντικότητα

Επειδή η δειγματική κατανομή των συντελεστών τ_a, τ_b και τ_c είναι γνωστή, μπορούμε να υπολογίσουμε το ασυμπτωτικό τυπικό τους σφάλμα και την σημαντικότητα.

Καθώς οι τύποι υπολογισμού της διακύμανσης είναι πολύπλοκοι και ξεφεύγουν από τους σκοπούς της παρούσας εργασίας, παραπέμπουμε τον αναγνώστη στο εγχειρίδιο SAS/STAT User's Guide για αναφορά [*SAS/STAT User's Guide, Version 8, Chapter 28, p. 1290-1291*, (1999)]. Επίσης, μια καλή αναφορά αποτελεί και το βιβλίο [*Measures of association, Liebetrau*, (1983)].

Άλλα χαρακτηριστικά

Ο συντελεστής τ_a είναι ισοδύναμος του συντελεστή *rho* του *Spearman*, που θα εξετάσουμε στην συνέχεια, αλλά ερμηνεύεται διαφορετικά. Ο συντελεστή *rho* του *Spearman* ερμηνεύεται ως η αναλογία της διακύμανσης που οφείλεται στην σχέση μεταξύ των μεταβλητών, ενώ ο συντελεστής τ_a αναπαριστά την πιθανότητα τα δεδομένα να έχουν την ίδια διάταξη, έναντι της πιθανότητας να μην έχουν την ίδια διάταξη. Ο συντελεστής τ_b είναι παρόμοιος με τον συντελεστή *gamma*, με την μόνη διαφορά ότι το μέτρο τ_b κάνει διορθώσεις για τις «*ισοβαθμίες*» μεταξύ των ζευγών. Επίσης, για 2×2 πίνακες, το μέτρο τ_b απλοποιείται στον

συντελεστή συσχέτισης του *Pearson*, αναθέτοντας σκορ στις γραμμές και στις στήλες ανάλογα με την διάταξή τους.

Σχόλια

Το μέτρο τ_a δεν κάνει διορθώσεις για τις «*ισοβαθμίες*» των ζευγών παρατηρήσεων και επομένως δεν είναι κατάλληλο όταν υπάρχουν πολλές «*ισοβαθμίες*». Το μέτρο τ_b πετυχαίνει την τιμή ± 1 μόνο για τετραγωνικούς πίνακες, όταν όλες οι παρατηρήσεις βρίσκονται στην διαγώνιο. Το μέτρο τ_c είναι σχεδιασμένο έτσι ώστε να μπορεί να πετύχει τις τιμές ± 1 για μη τετραγωνικούς πίνακες. Όμως, επειδή οι τιμές του εξαρτώνται από το μέγεθος του πίνακα, σύμφωνα με τον *Somers*, δεν θεωρείται και τόσο κατάλληλο μέτρο.

5.2.3 Συντελεστής D του Somers

Ο συντελεστής D [*Somers*, (1962)] προτάθηκε ως ένα εναλλακτικό ασύμμετρο μέτρο, για την πρόβλεψη της εξαρτημένης μεταβλητής, γνωρίζοντας την ανεξάρτητη. Στην ουσία αποτελεί μια τροποποίηση του μέτρου *gamma*, λαμβάνοντας υπόψη του τις «*ισοβαθμίες*» των ζευγών παρατηρήσεων και κάνοντας την υπόθεση ότι η ανεξάρτητη μεταβλητή μπορεί να χρησιμοποιηθεί για την πρόβλεψη της εξαρτημένης. Υπάρχουν 2 τύποι υπολογισμού για τον συντελεστή D, ανάλογα με το ποια μεταβλητή θεωρείται ανεξάρτητη.

Ο τύπος υπολογισμού του ασύμμετρου μέτρου $D_{Y|X}$, θεωρώντας την μεταβλητή γραμμή X ως ανεξάρτητη και την μεταβλητή στήλη Y ως εξαρτημένη, είναι ο εξής

$$D_{Y|X} = \frac{C - D}{\left(n(n-1)/2\right) - T_X} \tag{5.11}$$

Ο τύπος υπολογισμού του ασύμμετρου μέτρου $D_{Y|X}$, θεωρώντας την μεταβλητή στήλη Y ως ανεξάρτητη και την μεταβλητή γραμμή X ως εξαρτημένη, είναι ο εξής

$$D_{X|Y} = \frac{C - D}{\left(n(n-1)/2\right) - T_Y} \tag{5.12}$$

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 3×3 πίνακα συνάφειας (*σελ. 167, Παράδειγμα 3*) έχουμε

$$\hat{D}_{Y|X} = \frac{524.112 - 112.915}{1.574.425 - 644.625} = 0.442$$

και

$$\hat{D}_{X|Y} = \frac{524.112 - 112.915}{1.574.425 - 542.173} = 0.398$$

Ερμηνεία

Ο συντελεστής D εκφράζει το πλεόνασμα των «*σύμφωνων*» ζευγών έναντι των «*ασύμφωνων*», χρησιμοποιώντας μόνο τα ζεύγη παρατηρήσεων χωρίς «ισοβαθμίες». Το αποτέλεσμα του παραδείγματος για το μέτρο, $\hat{D}_{Y|X} = 0.442$, υποδηλώνει θετική συνάφεια μετρίου βαθμού, με την έννοια ότι οι παρατηρήσεις που βρίσκονται σε υψηλότερη κατηγορία εκπαίδευσης, βρίσκονται επίσης σε υψηλότερη κατηγορία εισοδήματος. Παρόμοια, είναι και η ερμηνεία για το $\hat{D}_{X|Y} = 0.398$.

Πεδίο Ορισμού

Ο συντελεστής D κυμαίνεται στο διάστημα $[-1+1]$.

Επίπεδο Δεδομένων

Ο συντελεστής D χρησιμοποιείται με διατακτικές μεταβλητές και εφαρμόζεται σε οποιοδήποτε $I \times J$ πίνακα συνάφειας.

Συμμετρία

Ο συντελεστής D είναι ένα ασύμμετρο μέτρο, καθώς διαφορετικά αποτελέσματα προκύπτουν ανάλογα με το ποια μεταβλητή θεωρείται ανεξάρτητη

Σχόλια

Το μέτρο D είναι στενά συνδεδεμένο με το μέτρο τ_b και έχει παρόμοια ερμηνεία. Διαφέρει από το τ_b στο ότι χρησιμοποιεί στον τύπο υπολογισμού του, διόρθωση μόνο για τα ζεύγη παρατηρήσεων που έχουν «ισοβαθμίες» στην ανεξάρτητη μεταβλητή. Για τετραγωνικούς πίνακες το μέτρο τ_b είναι ο γεωμετρικός μέσος των ασύμμετρων μέτρων $D_{X|Y}$ και $D_{Y|X}$. Επίσης σημειώνουμε ότι, όπως και στην περίπτωση του *gamma*, $D = 0$ όταν οι δυο μεταβλητές είναι ανεξάρτητες, αλλά δεν είναι απαραίτητο να ισούται με 1, στην περίπτωση που οι μεταβλητές

είναι πλήρως εξαρτημένες. Επίσης, έχει ερμηνεία αναλογικής μείωσης του σφάλματος πρόβλεψης.

5.3 Διατακτικά μέτρα συνάφειας που βασίζονται στην ανάθεση *σκορ*

Στην περίπτωση συνεχών δεδομένων, τα οποία προέρχονται από την κανονική κατανομή, προκειμένου να μελετήσουμε την ένταση της συνάφειας μεταξύ των μεταβλητών, χρησιμοποιούμε τον πολύ γνωστό συντελεστή συσχέτισης $\rho_{X,Y}$ (*Pearson*, 1904). Ο συντελεστής συσχέτισης $\rho_{X,Y}$ (*Pearson correlation coefficient*), που αποτέλεσε την βάση για την περαιτέρω ανάπτυξη αρκετών συντελεστών συνάφειας, μετρά την ένταση και την κατεύθυνση της γραμμικής σχέσης μεταξύ δυο μεταβλητών, περιγράφοντας τον βαθμό για τον οποίο η μια μεταβλητή σχετίζεται με την άλλη. Ο τύπος υπολογισμού του συντελεστή στην πληθυσμιακή του μορφή είναι

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E(X-\mu_X)(Y-\mu_Y)}{\sigma_X \sigma_Y} \tag{5.13}$$

Η δειγματική εκτίμηση του συντελεστή συμβολίζεται με r και ισούται με

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{5.14}$$

5.3.1 Συντελεστής ιεραρχικής συσχέτισης *rho* του Spearman

Κάποιοι μέθοδοι για την μέτρηση της συνάφειας μεταξύ διατακτικών μεταβλητών, βασίζονται στον συντελεστή r, απαιτούν όμως την ανάθεση σκορ (*score*) στα επίπεδα των μεταβλητών. Όταν αναλύουμε δεδομένα με πίνακες συνάφειας, τιμές κλίμακας αναθέτονται στις κατηγορίες των γραμμών και των στηλών και τα δεδομένα θεωρούμε ότι προέρχονται από μια ομαδική κατανομή συχνότητας. Έστω R_i και R_j οι τιμές των βαθμών (*rank*) που ανατέθηκαν

στις μεταβλητές X , Y , αντίστοιχα. Ο συντελεστής ιεραρχικής συσχέτισης r_S (*Spearman's rank correlation coefficient*), ισούται με

$$r_S = \frac{Cov(X,Y)}{\sqrt{S_X S_Y}} = \frac{\sum_i \sum_j R_i R_j n_{ij} - \frac{\left(\sum_i R_i n_{i.}\right)\left(\sum_j R_j n_{j.}\right)}{n}}{\sqrt{\left[\sum_i R_i^2 n_{i.} - \frac{\left(\sum_i R_i n_{i.}\right)^2}{n}\right]\left[\sum_j R_j^2 n_{.j} - \frac{\left(\sum_j R_j n_{.j}\right)^2}{n}\right]}} \qquad (5.15)$$

Ο αριθμητής εκφράζει το άθροισμα των γινομένων (*sum of cross product*) και ο παρανομαστής το άθροισμα των τετραγώνων (*sum of square*).

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 3×3 πίνακα συνάφειας (*σελ. 167, Παράδειγμα 3*) έχουμε

$\hat{r}_S = 0.462$, με $R_i = \{1,2,3\}$ και $R_j = \{1,2,3\}$

$$Cov(X,Y) = \sum_i \sum_j R_i R_j n_{ij} - \frac{\left(\sum_i R_i n_{i.}\right)\left(\sum_j R_j n_{j.}\right)}{n} = 6863 - \frac{3475\times 3282}{1775} = 437,68$$

$$S_X = \sum_i R_i^2 n_{i.} - \frac{\left(\sum_i R_i n_{i.}\right)^2}{n} = 7585 - \frac{3475^2}{1775} = 781,83$$

$$S_Y = \sum_j R_j^2 n_{.j} - \frac{\left(\sum_j R_j n_{.j}\right)^2}{n} = 7212 - \frac{3282^2}{1775} = 1143,53$$

Ερμηνεία

Το αποτέλεσμα του παραδείγματος, $\hat{r}_S = 0.462$, υποδηλώνει θετική συνάφεια μετρίου βαθμού, με την έννοια ότι οι παρατηρήσεις που βρίσκονται σε υψηλότερη κατηγορία εκπαίδευσης,

βρίσκονται επίσης σε υψηλότερη κατηγορία εισοδήματος. Γενικά, όταν η ανεξάρτητη μεταβλητή έχει τον ίδιο βαθμό (*rank*) με την εξαρτημένη, τότε $\hat{r}_S = 1$.

Πεδίο Ορισμού

Ο συντελεστής r_S κυμαίνεται στο διάστημα $[-1+1]$.

Επίπεδο Δεδομένων

Ο συντελεστής r_S χρησιμοποιείται με διατακτικές μεταβλητές και εφαρμόζεται σε οποιοδήποτε $I \times J$ πίνακα συνάφειας.

Συμμετρία

Ο συντελεστής r_S είναι ένα συμμετρικό μέτρο, καθώς δεν μεταβάλλεται σε αλλαγές στην διάταξη γραμμών και στηλών, δηλαδή όταν ο πίνακας αλλάξει προσανατολισμό και οι γραμμές γίνουν στήλες και οι στήλες, γραμμές.

Σχόλια

Το μέτρο r_S χρησιμοποιείται κυρίως όταν μια από τις διατακτικές μεταβλητές έχει πολλές κατηγορίες και μοιάζει σαν συνεχής μεταβλητή. Όταν δεν έχουμε πολλές κατηγορίες, τα μέτρα *gamma* και *tau* είναι καταλληλότερα.

5.4 Συμπεράσματα

Στο Κεφάλαιο αυτό, εξετάσαμε 7 μέτρα συνάφειας, τα οποία χρησιμοποιούνται για διατακτικές μεταβλητές. Τα μέτρα αυτά στηρίζονται στις έννοιες των «*σύμφωνων*» και «*ασύμφωνων*» ζευγών και διαφέρουν μεταξύ τους ως προς τον τρόπο με τον οποίο διαχειρίζονται τις «ισοβαθμίες» μεταξύ των ζευγών παρατηρήσεων. Σημειώνουμε ότι ο συντελεστής ιεραρχικής συσχέτισης r_S, διαφοροποιείται καθώς οι υποθέσεις στον τρόπο υπολογισμού του βασίζονται στην ανάθεση σκορ στις κατηγορίες των μεταβλητών και στην ουσία προέρχεται από τον πολύ διαδεδομένο συντελεστή συσχέτισης r του *Pearson*. Συνοψίζοντας τις τιμές των εξεταζόμενων μέτρων, βασιζόμενοι στο Παράδειγμα 3 (*σελ. 167*) ενός 3×3 πίνακα συνάφειας, θα προσπαθήσουμε να καταλήξουμε σε κάποια συμπεράσματα. Μελετώντας τον παρακάτω συνοπτικό Πίνακα 5-1 (*σελ. 90*), συμπεραίνουμε ότι υπάρχει μέτρια συνάφεια, σύμφωνα με τα

κριτήρια του Cohen (*βλέπε σελ. 74*) μεταξύ των μεταβλητών. Δηλαδή μεγαλύτερο επίπεδο εκπαίδευσης, σχετίζεται με υψηλότερο εισόδημα.

ΠΙΝΑΚΑΣ 5-1

Συνοπτικός πίνακας μέτρων συνάφειας για διατακτικές μεταβλητές

Παράδειγμα 3 (3x3 πίνακας)			
	Μέτρο συνάφειας	Αποτέλεσμα	Στρογγυλοποίηση
1.	Goodman-Kruskal's γ	0,645	0,65
2.	Kendall's τ_a	0,261	0,26
3.	Kendall's τ_b	0,420	0,42
4.	Kendall's τ_c	0,392	0,39
5.	Somers $D_{Y\|X}$	0,442	0,44
6.	Somers $D_{X\|Y}$	0,398	0,40
7.	Spearman r_s	0,463	0,46

Παρατηρούμε, ότι αν και τα μέτρα δεν πετυχαίνουν την ίδια ακριβώς τιμή (ποικίλουν με εύρος απόλυτων τιμών από 0.26 έως 0.65), μπορούμε να πούμε ότι τα περισσότερα πετυχαίνουν πολύ κοντινές τιμές. Συγκεκριμένα, τα μέτρα $3-7$ βρίσκονται πολύ κοντά, ενώ τα μέτρα $1-2$ παίρνουν την μέγιστη και την ελάχιστη τιμή, αντίστοιχα. Όπως έχουμε αναφέρει, ο λόγος που συμβαίνει αυτό, είναι ότι το μέτρο *gamma* έχει την τάση να αυξάνει την συνάφεια, ενώ το τ_α βασίζεται στο μέγεθος του πίνακα. Επίσης και τα δυο αυτά μέτρα δεν κάνουν διορθώσεις για τις «ισοβαθμίες» μεταξύ των ζευγών παρατηρήσεων και επομένως είναι προτιμότερο να αποφεύγονται. Αντιθέτως, τα μέτρα $3-7$ κάνουν διορθώσεις ως προς τις «ισοβαθμίες» και επομένως είναι καταλληλότερα, όταν τα δεδομένα που αναλύουμε έχουν μεγάλο πλήθος «ισοβαθμιών».

Μελετώντας τον παρακάτω συνοπτικό Πίνακα 5-2 (*σελ. 91*) όλων των εξεταζόμενων διατακτικών μέτρων συνάφειας, για το Παράδειγμα 4 (*σελ. 167*), συμπεραίνουμε ότι τα δεδομένα διέπονται από συνάφεια μεγαλύτερης έντασης. Δηλαδή το βάρος ενός ατόμου σχετίζεται με την

σειρά με την οποία γεννήθηκε και η συνάφεια αυτή είναι ισχυρή, σύμφωνα με τα κριτήρια του Cohen.

ΠΙΝΑΚΑΣ 5-2

Συνοπτικός πίνακας μέτρων συνάφειας για διατακτικές μεταβλητές

Παράδειγμα 4 (3x4 πίνακας)			
	Μέτρο συνάφειας	Αποτέλεσμα	Στρογγυλοποίηση
1.	Goodman-Kruskal's γ	0,699	0,70
2.	Kendall's τ_a	0,383	0,38
3.	Kendall's τ_b	0,544	0,54
4.	Kendall's τ_c	0,573	0,57
5.	Somers $D_{Y\|X}$	0,573	0,57
6.	Somers $D_{X\|Y}$	0,516	0,52
7.	Spearman r_s	0,585	0,59

Παρατηρούμε ότι και στο παράδειγμα αυτό οι τιμές των μέτρων ακολουθούν το ίδιο πρότυπο. Δηλαδή, τα μέτρα $1-2$ παίρνουν την μέγιστη και την ελάχιστη τιμή αντίστοιχα, ενώ τα μέτρα $3-7$ βρίσκονται πολύ κοντά, εκ των οποίων δυο από αυτά ταυτίζονται. Σημειώνουμε ότι το μέτρο τ_c είναι καταλληλότερο για μη-τετραγωνικούς πίνακες, ενώ το τ_b είναι καταλληλότερο για τετραγωνικούς πίνακες συνάφειας. Ακολουθεί ο συγκεντρωτικός Πίνακας 5-3 (*σελ. 92*), με τα βασικά χαρακτηριστικά και κάποια βασικά σχόλια για τα διατακτικά μέτρα συνάφειας που εξετάσαμε στο Κεφάλαιο αυτό.

ΠΙΝΑΚΑΣ 5-3

Συγκεντρωτικός πίνακας μέτρων συνάφειας για διατακτικές μεταβλητές

	Μέτρο συνάφειας	Πεδίο Ορισμού	Συμμετρία	Είδος δεδομένων	Διάσταση πίνακα	Σχόλια
1.	Goodman-Kruskal's γ	[-1,+1]	ΝΑΙ	Διατακτικά	*I* x *J*	Έχει την τάση να αυξάνει την συνάφεια και γενικά οι τιμές του θα είναι μεγαλύτερες από α υπόλοιπα διατακτικά μέτρα. Δεν κάνει διορθώσεις για τους δεσμούς των ζευγών παρατηρήσεων.
2.	Kendall's τ_a	[-1,+1]	ΝΑΙ	Διατακτικά	*I* x *J*	Βασίζεται στο μέγεθος του πίνακα. Δεν κάνει διορθώσεις για τους δεσμούς των ζευγών παρατηρήσεων.
3.	Kendall's τ_b	[-1,+1]	ΝΑΙ	Διατακτικά	*I* x *J*	Το μέτρο τ_b είναι παρόμοιο με το μέτρο gamma, με την διαφορά ότι κάνει διορθώσεις για τους δεσμούς. Είναι πιο συντηρητικό από το gamma για το ίδιο σύνολο δεδομένων. Είναι πιο αξιόπιστο όταν χρησιμοποιείται για τετραγωνικούς πίνακες καθώς έχει την τάση να μειώνει την συνάφεια, για πίνακες με ανόμοιες διαστάσεις.
4.	Kendall's τ_c	[-1,+1]	ΝΑΙ	Διατακτικά	*I* x *J*	Είναι κατάλληλο για οποιαδήποτε διάσταση του πίνακα, σε αντίθεση με το τ_b.
5.	Somers $D_{Y\|X}$	[-1,+1]	ΟΧΙ	Διατακτικά	*I* x *J*	Το μέτρο D συνδέεται με το μέτρο τ_b και έχει παρόμοια ερμηνεία, με την διαφορά ότι κάνει διορθώσεις για τα ζεύγη παρατηρήσεων που έχουν δεσμούς στην ανεξάρτητη μεταβλητή. Μπορεί οι μεταβλητές να είναι πλήρως εξαρτημένες αλλά D<1, όπως συμβαίνει με το γ. Έχει ερμηνεία αναλογικής μείωσης του σφάλματος πρόβλεψης.
6.	Somers $D_{X\|Y}$	[-1,+1]	ΟΧΙ	Διατακτικά	*I* x *J*	
7.	Spearman r_s	[-1,+1]	ΝΑΙ	Διατακτικά	*I* x *J*	Χρησιμοποιείται όταν μια μεταβλητή έχει πολλές κατηγορίες και μοιάζει σαν συνεχής. Αν τετραγωνιστεί αποκτά ερμηνεία αναλογικής μείωσης του σφάλματος πρόβλεψης.

ΚΕΦΑΛΑΙΟ 6

ΜΕΤΡΑ ΣΥΜΜΕΤΡΙΑΣ – ΑΣΥΜΜΕΤΡΙΑΣ

6.1 Εισαγωγή

Όπως είδαμε στα προηγούμενα κεφάλαια, για έναν $I \times J$ πίνακα συνάφειας (ή άλλον μεγαλύτερης διάστασης), μια μεγάλη ποικιλία μέτρων συνάφειας είναι διαθέσιμη, ανάλογα με το πείραμα και το είδος των μεταβλητών που εξετάζονται, για την μελέτη του βαθμού της συνάφειας, μεταξύ ονοματικών και διατακτικών μεταβλητών. Στην περίπτωση όμως ενός τετραγωνικού $I \times I$ πίνακα συνάφειας, ο οποίος έχει την ίδια (ονοματική ή διατακτική) ταξινόμηση γραμμών και στηλών, το ενδιαφέρον επικεντρώνεται στην μελέτη της συμμετρίας γύρω από τα στοιχεία της κυρίας διαγώνιου και στην ισότητα των περιθωρίων πιθανοτήτων της γραμμής και της αντίστοιχης στήλης, παρά για την ανεξαρτησία μεταξύ των μεταβλητών. Αν και το γενικό θέμα της συμμετρίας και της μοντελοποίησής της, έχει μελετηθεί από πολλούς στατιστικούς, ξεκινώντας το 1947 με τον *McNemar*, αντίστοιχα μέτρα συμμετρίας – ασυμμετρίας, που εκφράζουν τον βαθμό απομάκρυνσης από την πλήρης συμμετρία έχουν πρόσφατα αναπτυχθεί, σε σημαντικά μικρό βαθμό και σίγουρα δεν έχουν την αναγνώριση που έχουν τα κλασσικά μέτρα συνάφειας. Στο Κεφάλαιο 6, θα παρουσιάσουμε και θα σχολιάσουμε τα κυριότερα μέτρα συμμετρίας – ασυμμετρίας που έχουν προταθεί στην βιβλιογραφία και είναι χρήσιμα για την ανάλυση τετραγωνικών πινάκων συνάφειας και μελέτης της συμμετρίας τους.

6.2 Βασικά πεδία εφαρμογών

Τετραγωνικοί πίνακες συνάφειας με τις ίδιες ονοματικές ή διατακτικές κατηγορίες εμφανίζονται συχνά, κυρίως σε μελέτες κατά ζεύγη (*matched pairs studies*) ή σε μελέτες συγκεκριμένων ομάδων με ίδια χαρακτηριστικά, σε διαφορετικές χρονικές στιγμές (*panel studies*). Επίσης, χρησιμοποιούνται συχνά και σε εφαρμογές μοντέλων συμφωνίας βαθμολογητών. Χαρακτηριστικά παραδείγματα που οδηγούν σε τέτοιου είδους πίνακες είναι:

1. όταν η κατάσταση ενός υποκειμένου εξετάζεται σε δυο διαφορετικές χρονικές στιγμές

2. όταν συγκρίνουμε δυο παρόμοια χαρακτηριστικά ενός υποκειμένου σε ένα δείγμα (σύγκριση όρασης δεξιού και αριστερού ματιού)
3. όταν συγκρίνουμε το ίδιο χαρακτηριστικό για ένα συγκεκριμένο ζεύγος υποκειμένων (πατέρας – παιδί, σύζυγοι, αδέλφια κτλ.)
4. όταν συγκρίνουμε την διαγνωστική ικανότητα δυο ειδικών, χρησιμοποιώντας το ίδιο σύνολο υποκειμένων.

6.3 Ταξινόμηση των μέτρων συμμετρίας - ασυμμετρίας

Όπως και στην περίπτωση των μέτρων συνάφειας, τα μέτρα συμμετρίας – ασυμμετρίας διακρίνονται, ανάλογα με το είδος των μεταβλητών που εξετάζουν, σε αυτά που είναι κατάλληλα για ονοματικές μεταβλητές και άλλα κατάλληλα για διατακτικές μεταβλητές. Ένας άλλος διαχωρισμός των μέτρων, βασίζεται στο είδος της υποβόσκουσας δομής συμμετρίας του πίνακα. Έτσι, ανάλογα με το μοντέλο συμμετρίας ή εκτεταμένης συμμετρίας (*extended symmetry*) ή ασυμμετρίας που ισχύει, ορίζουμε και τα αντίστοιχα μέτρα συμμετρίας – ασυμμετρίας που έχουν προταθεί. Οι συντελεστές συμμετρίας βασίζονται κυρίως στα μοντέλα συμμετρίας $\left(S-Symmetry\right)$, ψευδοσυμμετρίας $\left(QS-Quasi\,Symmetry\right)$ και περιθώριας ομοιογένειας $\left(MH-marginal\,homogeneity\right)$, ενώ οι συντελεστές ασυμμετρίας βασίζονται στα μοντέλα διαγώνιας συμμετρίας $\left(DS-Diagonal\,Symmetry\right)$, τριγωνικής συμμετρίας $\left(TS-Triangular\,Symmetry\right)$ και δεσμευμένης συμμετρίας $\left(CS-Conditional\,Symmetry\right)$. Τέλος τα μέτρα συμμετρίας – ασυμμετρίας, μπορούν περαιτέρω να κατηγοριοποιηθούν με βάση το είδος της «απόστασης» ή «απόκλισης», που χρησιμοποιούν στον τύπο υπολογισμού τους. Όπως θα διαπιστώσουμε στην συνέχεια, οι συντελεστές συμμετρίας – ασυμμετρίας βασίζονται στα γνωστά μέτρα απόστασης ή απόκλισης (*divergence*) που παρουσιάστηκαν στην παράγραφο 3.6.2 και προέρχονται από τον χώρο της θεωρίας της πληροφορίας (*Information Theory*). Στα πλαίσια αυτά, δεν είναι απίθανο να συναντήσουμε συντελεστές, που μετρούν τον βαθμό απομάκρυνσης από την συμμετρία - ασυμμετρία ίδιας δομής, αλλά ορίζονται με διαφορετικό τρόπο, καθώς βασίζονται σε διαφορετικά μέτρα απόστασης (ή απόκλισης). Αξίζει να

σημειώσουμε ότι σε αντίθεση με τα μέτρα συνάφειας, η ανάπτυξη μέτρων συμμετρίας - ασυμμετρίας είναι πρωτίστως λειτουργική, με την έννοια ότι βασίζονται περισσότερο στις ιδιότητές τους, παρά σε αξιώματα [βλέπε *Renyi* (1959), *Kimeldorf & Sampson*, (1989)].

Οι κύριοι εισηγητές μέτρων συμμετρίας - ασυμμετρίας, από όσο γνωρίζουμε, είναι οι *Tomizawa* (1994, 1995), *Tomizawa et al.* (1998, 2001, 2003, 2005, 2007) και οι *Kateri & Papaioannou* [*Technical Report* (TR07 – 3), University of Pireaus, (2007)]. Οι πρώτοι βασίστηκαν στην απόκλιση τύπου *power – divergence* των *Cressie & Read* (1984) και στον δείκτη ανομοιότητας ή ποικιλότητας (*diversity index*) των *Patil & Taillie*. Σημειώνουμε ότι η απόκλιση τύπου *power – divergence* των *Cressie & Read*, εμπεριέχει σε ειδικές περιπτώσεις της, την πληροφορία *Kullback - Leibler* και την απόκλιση χ^2 *– Pearson*, ενώ ο δείκτης ανομοιότητας (ή ποικιλότητας) των *Patil & Taillie*, περιλαμβάνει σε ειδικές περιπτώσεις του, την εντροπία του *Shannon* και τον δείκτη συγκέντρωσης (*gini concentration*), ή αλλιώς τον δείκτη *Simpson*. Οι *Kateri & Papaioannou* βασίστηκαν στην απόκλιση φ *– divergence* των *Csiszar* (1963) και *Ali & Silvey* (1966). Η απόκλιση αυτή, είναι ένα γνωστό μέτρο της κατευθυνόμενης απόστασης (*direct divergence*) ή της ψευδό-απόστασης, η οποία έχει καλές ιδιότητες και έχει ενοποιήσει πολλούς άλλους δείκτες πληροφορίας (*information numbers*), όπως *Kullback-Leibler*, *Renyi*, *Cressie & Read*, *Pearson* [βλέπε *Papaioannou* (1985)]. Ο αναγνώστης μπορεί να ανατρέξει σε εργασίες των *Agresti* (1984), *Tomizawa* (1984, 1985, 1987, 1989, 1990, 1992), *Havranek & Lienert* (1986), *Rosenstein* (1989), *Chino* (1990) και *Becker* (1990), για μια πιο εκτεταμένη ανασκόπηση της βιβλιογραφίας σε μοντέλα συμμετρίας – ασυμμετρίας για τετραγωνικούς πίνακες συνάφειας.

Στην παράγραφο 6.4 παρουσιάζουμε τα μοντέλα συμμετρίας για τετραγωνικούς πίνακες συνάφειας, τα οποία είναι απαραίτητα για τον ορισμό των προτεινόμενων μέτρων συμμετρίας που περιγράφονται στην παράγραφο 6.5. Εν συνεχεία, στην παράγραφο 6.6 παρουσιάζουμε τα μοντέλα ασυμμετρίας για τετραγωνικούς πίνακες συνάφειας, τα οποία είναι απαραίτητα για τον ορισμό των προτεινόμενων μέτρων ασυμμετρίας που περιγράφονται στην παράγραφο 6.7.

6.4 Ορισμοί των μοντέλων Συμμετρίας

Αρκετά μοντέλα έχουν παρουσιαστεί για την μελέτη της συμμετρίας σε διδιάστατους τετραγωνικούς πίνακες συνάφειας, βλέπε *Agresti* (1990), για μια ανασκόπηση της βιβλιογραφίας και τα οποία σε κάποιες περιπτώσεις, μπορούν να επεκταθούν σε πίνακες μεγαλύτερης διάστασης. Ο *Bowker* (1948) εξέτασε το μοντέλο της συμμετρίας $\left(S-Symmetry\right)$ και πρότεινε ένα χ^2-test. Όταν το μοντέλο της συμμετρίας δεν ισχύει, τότε μας ενδιαφέρει να δούμε αν υπάρχει κάποια άλλη δομή συμμετρίας, που περιγράφεται από τα μοντέλα της περιθώριας ομοιογένειας $\left(MH-Marginal\ Homogeneity\right)$ που προτάθηκε από τους *Bishop et al*. (1975) και της ψευδοσυμμετρίας $\left(QS-Quasi\ Symmetry\right)$ που προτάθηκε από τον *Caussinus* (1965). Επίσης, όταν το μοντέλο S δεν ισχύει, είναι λογικό να εξετάσουμε κάποιους ιδιαίτερους τύπους ασυμμετρίας, οι οποίοι περιγράφονται από τα μοντέλα $\left(DS-Diagonal\ Symmetry\right)$, $\left(TS-Triangular\ Symmetry\right)$ και $\left(CS-Conditional\ Symmetry\right)$ στην παράγραφο 6.7 και αναφέρονται σε διατακτικές μεταβλητές ταξινόμησης.

Έστω ένας $R\times R$ τετραγωνικός πίνακα συνάφειας $\Pi=\left(\pi_{ij}\right)_{R\times R}$, όπου $\pi_{ij}=\Pr(X=i,Y=j)$ η πιθανότητα μια παρατήρηση να βρίσκεται στο κελί (i,j) και X,Y οι μεταβλητές σύμμετρης ταξινόμησης γραμμής και στήλης αντίστοιχα, με $i,j=1,2,...,R$ και $\sum_{i=1}^{R}\sum_{j=1}^{R}\pi_{ij}=1$. Οι αντίστοιχες περιθώριες κατανομές πιθανότητας συμβολίζονται με $\pi_{i.}$, $\pi_{.j}$, ενώ οι δειγματικές πιθανότητες συμβολίζονται με p_{ij}.

6.4.1 Μοντέλο Συμμετρίας (*S – Symmetry*)

Το μοντέλο συμμετρίας $\left(S\right)$ προτάθηκε από τον *Bowker* (1948) και ορίζεται ως

$$\pi_{ij}=\pi_{ji},\ \text{για}\ i,j=1,...,R\ \text{και}\ i\neq j \qquad \left(S_1\right)$$

Το μοντέλο αυτό δείχνει ότι η πιθανότητα μια παρατήρηση να βρεθεί στο κελί (i, j) του πίνακα, είναι ίση με την πιθανότητα η παρατήρηση να βρεθεί στο κελί (j, i). Δηλαδή το μοντέλο περιγράφει μια δομή συμμετρίας των πιθανοτήτων των μη διαγώνιων κελιών, γύρω από την κεντρική διαγώνιο του πίνακα. Θέτοντας $\pi_{ij}^{S} = \frac{(\pi_{ij} + \pi_{ji})}{2}$, το μοντέλο συμμετρίας μπορεί να ορισθεί και ως

$$\pi_{ij} = \pi_{ij}^{S} \text{ για } i, j = 1, ..., R \qquad (S_2)$$

6.4.2 Μοντέλο Περιθώριας Ομοιογένειας (MH – Marginal Homogeneity)

Το μοντέλο (MH) προτάθηκε από τους *Bishop et al*. (1975) και ορίζεται ως

$$\pi_{i.} = \pi_{.i} \text{ ή αλλιώς } \Pr(X = i) = \Pr(Y = i), \text{ για } i = 1, ..., R \qquad (MH_1)$$

όπου $\pi_{i.} = \sum_{k=1}^{R} \pi_{ik}$ και $\pi_{.i} = \sum_{k=1}^{R} \pi_{ki}$, οι περιθώριες κατανομές πιθανότητας γραμμής και στήλης.

Το μοντέλο (MH) μπορεί επίσης να εκφραστεί [βλέπε *Tomizawa et al*. (2003)] και ως

$$\pi_{i.}^{c} = \pi_{.i}^{c} \text{ ή αλλιώς } \Pr(X = i \mid X \neq Y) = \Pr(Y = i \mid X \neq Y), \text{ για } i = 1, 2, ..., R \qquad (MH_2)$$

όπου $\pi_{i.}^{c} = \frac{(\pi_{i.} - \pi_{ii})}{\delta}$, $\pi_{.i}^{c} = \frac{(\pi_{.i} - \pi_{ii})}{\delta}$ και $\delta = \sum\sum_{i \neq j} \pi_{ij}$.

Αυτό σημαίνει ότι η δεσμευμένη περιθώρια κατανομή της μεταβλητής γραμμής, είναι πανομοιότυπη με την δεσμευμένη περιθώρια κατανομή της μεταβλητής στήλης, δοθέντος ότι μια παρατήρηση θα βρεθεί σε ένα από τη μη-διαγώνια κελιά του πίνακα.

Επιπλέον, *στην περίπτωση διατακτικών μεταβλητών*, παίρνοντας την διαφορά των αθροιστικών περιθώριων πιθανοτήτων, $F_i^X - F_i^Y$ για $i = 1, 2, ..., R-1$, όπου $F_i^X = \Pr(X \leq i)$ και $F_i^Y = \Pr(Y \leq i)$, παρατηρούμε ότι το μοντέλο (MH) μπορεί να εκφραστεί [βλέπε *Tomizawa et al.* (2003)] και ως

$$G_{1(i)} = G_{2(i)}, \text{ ή αλλιώς } F_i^X = F_i^Y \text{ για } i = 1, 2, ..., R-1 \qquad (MH_3)$$

όπου $G_{1(i)} = \sum_{s=1}^{i}\sum_{t=i+1}^{R} \pi_{st}$ $\left[= \Pr\left(X \le i, Y \ge i+1\right)\right]$, $G_{2(i)} = \sum_{s=i+1}^{R}\sum_{t=1}^{i} \pi_{st}$ $\left[= \Pr\left(X \ge i+1, Y \le i\right)\right]$

για $i = 1, 2, ..., R-1$.

Απόδειξη

Εστώ $F_i^X = P\left(X \le i\right)$ η αθροιστική περιθώρια κατανομή πιθανότητας της μεταβλητής X, τότε

$$F_i^X = P\left(X \le i\right) = P\left(X \le i, Y \le i\right) + P\left(X \le i, Y \ge i+1\right) = P\left(X \le 1, Y \le 1\right) + P\left(X \le 1, Y \ge 2\right) +$$
$$+P\left(X \le 2, Y \le 2\right) + P\left(X \le 2, Y \ge 3\right) + ... + P\left(X \le R-1, Y \le R-1\right) + P\left(X \le R-1, Y \ge R\right)$$

Εστώ $F_i^Y = P\left(Y \le i\right)$ η αθροιστική περιθώρια κατανομή πιθανότητας της μεταβλητής Y, τότε

$$F_i^Y = P\left(Y \le i\right) = P\left(Y \le i, X \le i\right) + P\left(Y \le i, X \ge i+1\right) = P\left(Y \le 1, X \le 1\right) + P\left(Y \le 1, X \ge 2\right) +$$
$$+P\left(Y \le 2, X \le 2\right) + P\left(Y \le 2, X \ge 3\right) + ... + P\left(Y \le R-1, X \le R-1\right) + P\left(Y \le R-1, X \ge R\right)$$

Παίρνοντας την διαφορά των αθροιστικών περιθώριων πιθανοτήτων $F_i^X - F_i^Y$, έχουμε

$$F_i^X - F_i^Y = \left\{P\left(X \le 1, Y \le 1\right) + P\left(X \le 1, Y \ge 2\right) + P\left(X \le 2, Y \le 2\right) + P\left(X \le 2, Y \ge 3\right) + ... + \right.$$
$$\left. +P\left(X \le R-1, Y \le R-1\right) + P\left(X \le R-1, Y \ge R\right)\right\} -$$
$$-\left\{P\left(Y \le 1, X \le 1\right) + P\left(Y \le 1, X \ge 2\right) + P\left(Y \le 2, X \le 2\right) + P\left(Y \le 2, X \ge 3\right) + ... + \right.$$
$$\left. +P\left(Y \le R-1, X \le R-1\right) + P\left(Y \le R-1, X \ge R\right)\right\} =$$
$$= \left\{P\left(X \le 1, Y \ge 2\right) + P\left(X \le 2, Y \ge 3\right) + ... + P\left(X \le R-1, Y \ge R\right)\right\} -$$
$$-\left\{P\left(Y \le 1, X \ge 2\right) + P\left(Y \le 2, X \ge 3\right) + ... + P\left(Y \le R-1, X \ge R\right)\right\} =$$
$$= P\left(X \le i, Y \ge i+1\right) - P\left(X \ge i+1, Y \le i\right) = \sum_{s=1}^{i}\sum_{t=i+1}^{R} \pi_{st} - \sum_{s=i+1}^{R}\sum_{t=1}^{i} \pi_{st} = G_{1(i)} - G_{2(i)}$$

Αυτό σημαίνει δηλαδή ότι, η πιθανότητα μια παρατήρηση να πέσει στην i κατηγορία μιας γραμμής ή μικρότερη της i και στην $i+1$ κατηγορία μιας στήλης ή μεγαλύτερη της $i+1$, ισούται με την πιθανότητα μια παρατήρηση να πέσει στην i κατηγορία μιας στήλης ή μικρότερη της i και στην $i+1$ κατηγορία μιας γραμμής ή μεγαλύτερη της $i+1$.

6.4.3 Μοντέλο Ψευδοσυμμετρίας (QS – Quasi Symmetry)

Το μοντέλο $\left(QS\right)$, το οποίο προτάθηκε από τον *Caussinus* (1965), είναι μια επέκταση του μοντέλου $\left(S\right)$ και ορίζεται ως

$$\pi_{ij} = \alpha_i \beta_j \rho_{ij}, \text{ για } i, j = 1,...,R, \ i \neq j \text{ και } \rho_{ij} = \rho_{ji} \qquad \left(QS_1\right)$$

Υπάρχουν αρκετοί ισοδύναμοι ορισμοί για το μοντέλο $\left(QS\right)$. Ένας εξ'αυτών, προτάθηκε από τον *McCullagh* (1982), μέσω του εξής λογαριθμο-γραμμικού μοντέλου

$$\log \mu_{ij} = \theta_i + \eta_j + \varphi_{ij} \text{ για } i, j = 1,...,R \text{ και } \varphi_{ij} = \varphi_{ji} \qquad \left(QS_2\right)$$

όπου $\mu_{ij} = E\left(n_{ij}\right)$ και n_{ij} οι συχνότητες του πίνακα.

Εύκολα διαπιστώνουμε ότι $\mu_{ij} = E\left(n_{ij}\right) = n_{ij}/n = \pi_{ij}$ και επομένως το μοντέλο του *McCullagh*, είναι ισοδύναμο με το λογάριθμο του μοντέλου του *Caussinus*. Όπως ο *McCullagh* (1982) αναφέρει, τα βασικά κίνητρα ανάπτυξης του μοντέλου αυτού ήταν η μαθηματική και υπολογιστική του απλότητα.

Επιπλέον, το μοντέλο του $\left(QS\right)$ του *Causinnus* μπορεί να εκφρασθεί [βλέπε *Menendez et al.* (2005)] και ως

$$D_{ijk} = D_{kji} \text{ για } i, j, k = 1, 2,..., R \qquad \left(QS_3\right)$$

όπου $D_{ijk} = \pi_{ij}\pi_{jk}\pi_{ki}$ και $D_{kji} = \pi_{kj}\pi_{ji}\pi_{ik}$, *ως προς την κατηγορία* R

ή ισοδύναμα σε όρους *odds – ratios* ως

$$s_{ij} = s_{ji}, \text{ για } i \neq j \qquad \left(QS_4\right)$$

όπου $s_{ij} = \dfrac{\pi_{ij}\pi_{RR}}{\pi_{iR}\pi_{Rj}}$, για $i, j = 1, 2,..., R-1$.

Επίσης, ένας εναλλακτικός ορισμός του μοντέλου [βλέπε *Krampe et al.* (2009)], σε όρους των *τοπικών ή γειτονικών odds – ratios* είναι

$$\theta_{ij} = \theta_{ji} \qquad \left(QS_5\right)$$

όπου $\theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}}$ τα $odds-ratios$ των πιθανοτήτων γειτονικών κατηγοριών, για $i, j = 1,...,R-1$.

Τέλος, το μοντέλο (QS) μπορεί να εκφρασθεί [βλέπε *Kateri & Papaioannou*, (1997)] και ως

$$\pi_{ij} = \pi_{ij}^{QS},\ i, j = 1,...,R \qquad (QS_6)$$

όπου $\pi_{ij}^{QS} = \pi_{ij}^{S}\frac{2a_i}{a_i + a_j}$, $\pi_{ij}^{S} = \frac{\pi_{ij} + \pi_{ji}}{2}$ και a_i θετικοί παράμετροι.

Το μοντέλο (QS_6) έχει το πλεονέκτημα να θεωρεί την (QS) ως την απομάκρυνση από την πλήρης συμμετρία. Γενικά, το μοντέλο της (QS) χαρακτηρίζεται από την ιδιότητα των συμμετρικών αλληλεπιδράσεων [*Bishop et al.* (1975)], αντί των συμμετρικών πιθανοτήτων στα κελιά του πίνακα ή αλλιώς από την ιδιότητα των συμμετρικών γειτονικών $odds-ratios$ [*Goodman*, (1971b)]. Όπως ο *Agresti* (2002) αναφέρει, το μοντέλο (QS) είναι πιο ρεαλιστικό σε σχέση με το περιοριστικό μοντέλο (S), μειονεκτεί όμως σε απλότητα και ευκολία μιας φυσικής ερμηνείας.

6.5 Μέτρα Συμμετρίας

Στην παράγραφο αυτή θα εξετάσουμε διάφορα μέτρα συμμετρίας, που μετρούν τον βαθμό απομάκρυνσης από την (S), και άλλων δομών, όπως αυτή της (QS) και της (MH), ενός τετραγωνικού πίνακα συνάφειας με ονοματικές ή με διατακτικές μεταβλητές ταξινόμησης. Τα προτεινόμενα μέτρα βασίζονται στην πληροφορία *Kullback – Leibler*, στην απόκλιση $\chi^2 - Pearson$, στην εντροπία *Shannon*, στην απόκλιση *Gauss* και στην απόκλιση τύπου $power-divergence$ των *Cressie & Read* και είναι χρήσιμα για την σύγκριση του βαθμού απομάκρυνσης από την (S), την (QS) και την (MH), σε πολλούς πίνακες στους οποίους έχουν προσαρμοστεί τα ανωτέρω μοντέλα.

6.5.1 Μέτρα απομάκρυνσης από την Συμμετρία (S) για ονοματικές μεταβλητές

Υποθέτουμε ότι $\{\pi_{ij}+\pi_{ji}\}>0$, για κάθε $i \neq j$. Ο *Tomizawa* (1994) εξέτασε δυο ειδών μέτρα ϕ_S, ψ_S χρησιμοποιώντας την πληροφορία *Kullback-Leibler* και την απόκλιση $\chi^2 - Pearson$, αντίστοιχα. Επιπλέον, εξέφρασε ισοδύναμα τα μέτρα αυτά, χρησιμοποιώντας την εντροπία *Shannon* και την απόκλιση *Gauss* (ή αλλιώς το τετράγωνο της Ευκλείδειας απόστασης). Αξίζει να σημειώσουμε ότι οι πιθανότητες των διαγώνιων κελιών δεν συνεισφέρουν στα αθροίσματα, δηλαδή οι πιθανότητες των μη διαγώνιων κελιών δεν αθροίζουν στην μονάδα $\left(\sum\sum_{i\neq j}\pi_{ij}<1\right)$.

Ο *Tomizawa* τυποποίησε τις πιθανότητες π_{ij}, έτσι ώστε να αθροίζουν στην μονάδα, με την βοήθεια των μετασχηματισμών

$$\pi_{ij}^{*}=\frac{\pi_{ij}}{\delta} \text{ και } \pi_{ij}^{*S}=\frac{\pi_{ij}^{*}+\pi_{ji}^{*}}{2}, \text{ για } i,j=1,2,...,R \text{ και } i\neq j,$$

όπου $\delta=\sum\sum_{i\neq j}\pi_{ij}$ το άθροισμα των πιθανοτήτων των μη-διαγώνιων κελιών.

Σημειώνουμε ότι, η πληροφορία *Kullback-Leibler* μεταξύ δυο κατανομών $\{a_i\}$ και $\{b_i\}$, ισούται με

$$I\left(\{a_i\};\{b_i\}\right)=\sum_{i=1}^{R}a_i\log\left(a_i/b_i\right) \tag{6.1}$$

και η απόκλιση $\chi^2 - Pearson$ ισούται με

$$D\left(\{a_i\},\{b_i\}\right)=\sum_{i=1}^{R}\frac{\left(a_i-b_i\right)^2}{b_i} \tag{6.2}$$

A. Μέτρο φ_S τύπου πληροφορίας Kullback-Leibler και ψ_S τύπου απόκλισης Pearson

Τα μέτρα ορίζονται ως

A1. $$\varphi_S=\frac{1}{\log 2}I\left(\pi^{*};\pi^{*S}\right),$$

όπου $I\left(\pi^{*};\pi^{*S}\right)=\sum\sum_{i\neq j}\pi_{ij}^{*}\log\frac{\pi_{ij}^{*}}{\pi_{ij}^{*S}}\leq\log 2$

Απόδειξη

$$\frac{\pi_{ij}^{*}}{\pi_{ij}^{*S}}=\frac{2\pi_{ij}^{*}}{\pi_{ij}^{*}+\pi_{ji}^{*}}\leq 2\Leftrightarrow\log\frac{\pi_{ij}^{*}}{\pi_{ij}^{*S}}\leq\log 2\Leftrightarrow \mathrm{I}(\pi^{*};\pi^{*S})=\sum\sum_{i\neq j}\pi_{ij}^{*}\log\frac{\pi_{ij}^{*}}{\pi_{ij}^{*S}}\leq\log 2$$

A2. $\psi_S=D\left(\pi^{*};\pi^{*S}\right)$

όπου $D\left(\pi^{*};\pi^{*S}\right)=\sum\sum_{i\neq j}\frac{\left(\pi_{ij}^{*}-\pi_{ij}^{*S}\right)^2}{\pi_{ij}^{*S}}$

Σημειώνουμε ότι οι ποσότητες $I\left(\pi^{*};\pi^{*S}\right)$ και $D\left(\pi^{*};\pi^{*S}\right)$ είναι η πληροφορία *Kullback-Leibler* και η απόκλιση $\chi^2-Pearson$ αντίστοιχα, μεταξύ των κατανομών πιθανότητας $\left\{\pi_{ij}^{*}\right\}$ και $\left\{\pi_{ij}^{*S}\right\}$. Επιπλέον, τα π_{ij}^{*} εκφράζουν την πιθανότητα μια παρατήρηση να βρίσκεται στο κελί (i,j), δοθέντος ότι η παρατήρηση θα βρίσκεται σε ένα (ανεξαρτήτου ποιο) από τα μη-διαγώνια κελιά του πίνακα, ενώ τα π_{ij}^{*S} εκφράζουν το ήμισυ της πιθανότητας μια παρατήρηση να βρίσκεται στο κελί (i,j) ή (j,i), δοθείσης της ίδιας συνθήκης. Όταν $\pi_{ij}^{*}=\pi_{ij}^{*S}$ για όλα τα i,j τότε ο τετραγωνικός πίνακας είναι συμμετρικός και στην περίπτωση αυτή, τα π_{ij}^{*S} εκφράζουν την πιθανότητα μια παρατήρηση να βρίσκεται στο κελί (i,j), δοθέντος ότι μια παρατήρηση θα πέσει σε ένα από τα μη διαγώνια κελιά του τετραγωνικού πίνακα, όταν ισχύει το μοντέλο (S).

Β. Μέτρο φ_S τύπου εντροπίας Shannon και ψ_S απόκλισης Gausss

Τα παραπάνω μέτρα μπορούν περαιτέρω να εκφρασθούν, χρησιμοποιώντας τον μέσο όρο της δεσμευμένης εντροπίας *Shannon* (*Conditional Shannon Entropy*) και τον μέσο όρο της δεσμευμένης απόκλισης *Gauss* (*Conditional Gauss Discrepancy*), υπό την συνθήκη ότι μια παρατήρηση βρίσκεται σε ένα από τα μη διαγώνια κελιά (i,j) ή (j,i), για $i\neq j$. Πράγματι,

έστω π_{ij}^{c} η δεσμευμένη πιθανότητα ότι μια παρατήρηση βρίσκεται στο κελί (i, j), δοθέντος ότι η παρατήρηση θα βρίσκεται σε ένα από τα μη διαγώνια κελιά (i, j) ή (j,i), με $\pi_{ij}^{c} = \frac{\pi_{ij}}{\pi_{ij}+\pi_{ji}}$, για $i, j = 1, 2, ..., R$ με $i \neq j$ και $\pi_{ij}^{c} + \pi_{ji}^{c} = 1$. Σημειώνουμε ότι $\pi_{ij}^{c} = \frac{1}{2}, \forall i, j$, αν και μόνο αν ο πίνακας είναι συμμετρικός.

Έχουμε ότι

B1. $$\varphi_S = 1 - \frac{1}{\log 2} \sum\sum_{i<j} \left(\pi_{ij}^{*} + \pi_{ji}^{*}\right) H_{ij}\left(\pi^{c}\right) = 1 - \frac{1}{\log 2} \sum\sum_{i<j} \left(\pi_{ij}^{*} + \pi_{ji}^{*}\right) H_{ij}\left(\pi^{*c}\right),$$

όπου $H_{ij}\left(\pi^{c}\right) = H\left(\pi_{ij}^{c}\right) + H\left(\pi_{ji}^{c}\right) = -\pi_{ij}^{c} \log \pi_{ij}^{c} - \pi_{ji}^{c} \log \pi_{ji}^{c}$, η εντροπία *Shannon* μεταξύ των κατανομών $\left\{\pi_{ij}^{c}, \pi_{ji}^{c}\right\}$ δοθέντων των ζευγών (ij, ji). Ας σημειωθεί ότι $\pi_{ij}^{c} = \pi_{ij}^{*c}$

Επίσης, σημειώνουμε ότι ο τύπος υπολογισμού Α1 του μέτρου φ_S είναι ισοδύναμος με τον Β1.

<u>Απόδειξη</u>

$$\varphi_S = 1 - \frac{1}{\log 2} \sum\sum_{i<j} \left(\pi_{ij}^{*} + \pi_{ji}^{*}\right) H_{ij}\left(\pi^{c}\right) =$$

$$= 1 - \frac{1}{\log 2} \left\{ - \sum\sum_{i<j} \left(\pi_{ij}^{*} + \pi_{ji}^{*}\right) \pi_{ij}^{c} \log \pi_{ij}^{c} - \sum\sum_{i<j} \left(\pi_{ij}^{*} + \pi_{ji}^{*}\right) \pi_{ji}^{c} \log \pi_{ji}^{c} \right\}$$

$$= 1 - \frac{1}{\log 2} \left\{ - \sum\sum_{i<j} \left(\frac{\pi_{ij} + \pi_{ji}}{\delta} \right) \frac{\pi_{ij}}{\pi_{ij} + \pi_{ji}} \log \frac{\pi_{ij}}{\pi_{ij} + \pi_{ji}} - \sum\sum_{i<j} \left(\frac{\pi_{ij} + \pi_{ji}}{\delta} \right) \frac{\pi_{ji}}{\pi_{ij} + \pi_{ji}} \log \frac{\pi_{ji}}{\pi_{ij} + \pi_{ji}} \right\} =$$

$$= 1 - \frac{1}{\delta \log 2} \left\{ - \sum\sum_{i<j} \pi_{ij} \log \frac{\pi_{ij}}{\pi_{ij} + \pi_{ji}} - \sum\sum_{i<j} \pi_{ji} \log \frac{\pi_{ji}}{\pi_{ij} + \pi_{ji}} \right\} =$$

$$= 1 - \frac{1}{\delta \log 2} \left\{ - \sum\sum_{i<j} \pi_{ij} \log \frac{\pi_{ij}}{\pi_{ij} + \pi_{ji}} - \sum\sum_{i>j} \pi_{ij} \log \frac{\pi_{ij}}{\pi_{ij} + \pi_{ji}} \right\} = 1 + \frac{1}{\delta \log 2} \sum\sum_{i \neq j} \pi_{ij} \log \frac{\pi_{ij}}{\pi_{ij} + \pi_{ji}} =$$

$$= \frac{1}{\delta \log 2} \left(\log 2 \sum\sum_{i \neq j} \pi_{ij} + \sum\sum_{i \neq j} \pi_{ij} \log \frac{\pi_{ij}}{\pi_{ij} + \pi_{ji}} \right) = \frac{1}{\delta \log 2} \left[\sum\sum_{i \neq j} \pi_{ij} \left(\log 2 + \log \frac{\pi_{ij}}{\pi_{ij} + \pi_{ji}} \right) \right] =$$

$$= \frac{1}{\delta \log 2} \left[\sum\sum_{i \neq j} \pi_{ij} \left(\log \frac{2\pi_{ij}}{\pi_{ij} + \pi_{ji}} \right) \right] = \frac{1}{\delta \log 2} \left[\sum\sum_{i \neq j} \pi_{ij} \log \frac{\pi_{ij}}{\pi_{ij}^{S}} \right] = \frac{1}{\log 2} \sum\sum_{i \neq j} \pi_{ij}^{*} \log \frac{\pi_{ij}^{*}}{\pi_{ij}^{*S}}.$$

B2. $$\psi_S = 2\sum\sum_{i<j}\left(\pi_{ij}^{*}+\pi_{ji}^{*}\right)\Delta_{ij}\left(\pi^{c},1/2\right)$$

όπου $\Delta_{ij}\left(\pi^{c},1/2\right)=\left(\pi_{ij}^{c}-1/2\right)^{2}+\left(\pi_{ji}^{c}-1/2\right)^{2}$, η απόκλιση *Gauss* μεταξύ των κατανομών $\left\{\pi_{ij}^{c},\pi_{ji}^{c}\right\}$ και $\left\{\frac{1}{2},\frac{1}{2}\right\}$, ή αλλιώς το τετράγωνο της Ευκλείδειας απόστασης [για περισσότερες λεπτομέρειες αναφορικά με την απόκλιση *Gauss*, βλέπε *Linhart & Zucchini*, (1986)].

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 3×3 πίνακα συνάφειας (*σελ. 168, Παράδειγμα 5α και 5β*) έχουμε τα κάτωθι αποτελέσματα. Σημειώνουμε ότι οι 2 διαφορετικοί τύποι υπολογισμού Α1, Β1 του μέτρου φ_S είναι ισοδύναμοι. Ομοίως, οι τύποι υπολογισμού για το μέτρο ψ_S.

Παράδειγμα 5		**Παράδειγμα 5α** ***(ΑΥΓ vs ΟΚΤ 1971)***	***Παράδειγμα 5β (ΟΚΤ '71 vs ΔΕΚ '73)***
A1. Kullback-Leibler	**φS**	0,035	0,207
A2. Pearson discrepancy	**ψS**	0,048	0,268
B1. Shannon entropy	**φS**	0,035	0,207
B2. Gauss discrepancy	**ψS**	0,048	0,268

Ερμηνεία

Σύμφωνα με την πληροφορία *Kullback-Leibler* και την απόκλιση $\chi^2-Pearson$, οι ποσότητες φ_S, ψ_S αναπαριστούν τον βαθμό απομάκρυνσης από την (S). Ο βαθμός αυτός αυξάνεται καθώς οι τιμές των φ_S, ψ_S αυξάνονται. Επομένως, συγκρίνοντας τις απαντήσεις για την περίοδο Αύγουστος – Οκτώβριος 1971, οι τιμές των μέτρων τείνουν στο 0, δηλαδή ο πίνακας είναι συμμετρικός, με την έννοια ότι $\pi_{ij}=\pi_{ji}$. Αντιθέτως, συγκρίνοντας τις απαντήσεις για την περίοδο Οκτώβριος 1971 – Δεκέμβριος 1973, παρατηρούμε ότι ο βαθμός απομάκρυνσης από την συμμετρία είναι πολύ μεγαλύτερος, με την έννοια ότι οι $\pi_{ij}\neq\pi_{ji}$. Πιο συγκεκριμένα, μελετώντας τον πίνακα συνάφειας, παρατηρούμε ότι $p_{31}>p_{13}$ και $p_{32}>p_{23}$. Επομένως, τα άτομα δείχνουν να είναι πιο αναποφάσιστα κατά την δημοσκόπηση Οκτώβριος 1971 και πιο αποφασισμένα

(δηλαδή Ναι ή Όχι) κατά την δημοσκόπηση Δεκέμβριος 1973. Αξίζει να σημειώσουμε, ότι όπως ο *Tomizawa* (1994) αναφέρει, η εκτίμηση του βαθμού απομάκρυνσης από την συμμετρία, πρέπει να εξετάζεται σε όρους, ενός κατά προσέγγιση διαστήματος εμπιστοσύνης για τις εκτιμήσεις $\hat{\varphi}_S$, $\hat{\psi}_S$ των μέτρων. Τα συμπεράσματα παραμένουν τα ίδια, καθώς το διάστημα εμπιστοσύνης για το $\hat{\varphi}_S$ εμπεριέχει το 0, ενώ για το $\hat{\psi}_S$ όχι.

Πεδίο Ορισμού

Τα μέτρα φ_S, ψ_S κυμαίνονται στο διάστημα $[0,1]$. Σημειώνουμε ότι η ποσότητα $\left\{I\left(\pi^{*};\pi^{*S}\right)\right\}$ στον τύπο υπολογισμού του φ_S κυμαίνεται στο διάστημα $[0,\log 2]$ και κατά συνέπεια το μέτρο φ_S κυμαίνεται στο διάστημα $[0,1]$. Ο πίνακας συνάφειας είναι συμμετρικός, δηλαδή $\pi_{ij}=\pi_{ji}$ για $i, j=1,2,...,R$ και $i\neq j$, αν και μόνο αν $\varphi_S=0\ (\psi_S=0)$, ενώ υπάρχει πλήρης ασυμμετρία, υπό την έννοια ότι ο βαθμός απομάκρυνσης από την (S) μεγιστοποιείται, δηλαδή $\pi_{ij}=0$ ή $\pi_{ji}=0$, για $i, j=1,2,...,R$ και $i\neq j$, αν και μόνο αν $\varphi_S=1\ (\psi_S=1)$.

Επίπεδο Δεδομένων

Τα μέτρα φ_S, ψ_S είναι κατάλληλα μόνο για ονοματικές μεταβλητές, καθώς παραμένουν αναλλοίωτα κάτω από τις ίδιες αναδιατάξεις των γραμμών και των στηλών.

Διάστημα Εμπιστοσύνης

Χρησιμοποιώντας την μέθοδο Δέλτα, οι ποσότητες $\sqrt{n}\left(\hat{\varphi}_S-\varphi_S\right)$ και $\sqrt{n}\left(\hat{\psi}_S-\psi_S\right)$, ασυμπτωτικά $(n\to\infty)$ ακολουθούν την κανονική κατανομή με μέση τιμή $\mu=0$ και διακύμανση

B1. $$\sigma_{\varphi}^2=\frac{\sum\sum_{i\neq j}\pi_{ij}\Omega_{ij}^2-\delta\varphi_S^2}{\delta^2},\ \text{όπου}\ \Omega_{ij}=\frac{1}{\log 2}\log\left(\frac{2\pi_{ij}}{\pi_{ij}+\pi_{ji}}\right)$$

B2. $$\sigma_{\psi}^2=\frac{\sum\sum_{i\neq j}\pi_{ij}\Gamma_{ij}^2-\delta\psi_S^2}{\delta^2},\ \text{όπου}\ \Gamma_{ij}=\frac{\left(\pi_{ij}-\pi_{ji}\right)\left(\pi_{ij}+3\pi_{ji}\right)}{\left(\pi_{ij}+\pi_{ji}\right)^2}$$

Για περισσότερες λεπτομέρειες αναφορικά με την μέθοδο Δέλτα [βλέπε *Bishop et al*. (1975) και *Agresti* (1984)].

Σημειώνουμε ότι οι εκτιμήσεις των μέτρων φ_S, ψ_S συμβολίζονται με $\hat{\varphi}_S$, $\hat{\psi}_S$ και υπολογίζονται αντικαθιστώντας τις πληθυσμιακές πιθανότητες $\{\pi_{ij}\}$ με τις παρατηρούμενες πιθανότητες του δείγματος $\{p_{ij}\}$, όπου $p_{ij} = \frac{n_{ij}}{n}$ και $n = \sum_{i=1}^{R}\sum_{j=1}^{R} n_{ij}$. Έστω, $\hat{\sigma}_{\varphi}^2$ συμβολίζει την εκτίμηση της διακύμανσης σ_{φ}^2 από τις παρατηρήσεις του δείγματος, τότε η ποσότητα $\hat{\sigma}_{\varphi}/\sqrt{n}$ είναι η κατά προσέγγιση εκτιμήτρια του τυπικού σφάλματος και το κατά προσέγγιση $100\% \times (1-a)$ διάστημα εμπιστοσύνης για την εκτιμήτρια του μέτρου φ_S είναι

$$\hat{\varphi}_S \pm z_{a/2}\hat{\sigma}_{\varphi}/\sqrt{n} \tag{6.3}$$

όπου $z_{a/2}$ το ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής.

Κατά τον ίδιο τρόπο υπολογίζουμε και το διάστημα εμπιστοσύνης για την εκτιμήτρια του μέτρου ψ_S. Η ασυμπτωτική κανονική κατανομή των ποσοτήτων $\sqrt{n}(\hat{\varphi}_S - \varphi_S)$ και $\sqrt{n}(\hat{\psi}_S - \psi_S)$ είναι εφαρμόσιμη μόνο όταν $0 < \varphi_S < 1$ και $0 < \psi_S < 1$ αντίστοιχα, και κατά συνέπεια

$\sigma_{\varphi}^2 = 0$ $(\sigma_{\psi}^2 = 0)$, όταν $\varphi_S = 0$ $(\psi_S = 0)$ και $\varphi_S = 1$ $(\psi_S = 1)$

$\sigma_{\varphi}^2 > 0$ $(\sigma_{\psi}^2 > 0)$ όταν $0 < \varphi_S < 1$ $(0 < \psi_S < 1)$

Σχόλια

Τα μέτρα φ_S, ψ_S θα πρέπει να χρησιμοποιούνται όταν κάποιος θέλει να δει τον βαθμό απομάκρυνσης από την (S), χωρίς να γίνονται υποθέσεις ότι υπάρχει κάποια άλλη δομή συμμετρίας, όπως αυτή της (QS) και (MH). Είναι γνωστό ότι το μοντέλο (S) ισχύει αν και μόνο αν τα μοντέλα της (QS) και (MH) ισχύουν (βλέπε *Bishop et al*. 1975). Κατά συνέπεια, όταν ισχύει η (MH), ο βαθμός απομάκρυνσης από την (S) θα πρέπει να θεωρείται ένα μέτρο, που παίρνει την ελάχιστη τιμή όταν ισχύει η (S) και την μέγιστη τιμή, όταν ο βαθμός απομάκρυνσης από την (S) είναι ο μέγιστος, δοθέντος ότι ισχύει η (MH). Όπως ο *Tomizawa*

(1994) αναφέρει, υπό αυτές τις συνθήκες τα μέτρα φ_S, ψ_S, δεν είναι κατάλληλα. Επίσης αναφέρει, ότι είναι δύσκολο να αποσαφηνιστεί ποιο από τα δυο μέτρα είναι προτιμότερο. Ο αναλυτής για δοσμένους πίνακες, θα πρέπει να υπολογίσει και τις δυο τιμές των μέτρων φ_S, ψ_S και μετά να αποφασίσει για τον βαθμό απομάκρυνσης από την (S). Όπως αναφέραμε, η εκτίμηση του βαθμού απομάκρυνσης από την συμμετρία, πρέπει να εξετάζεται σε όρους, ενός κατά προσέγγιση διαστήματος εμπιστοσύνης για τα φ_S, ψ_S και όχι σε όρους μόνο των εκτιμητών τους $\hat{\varphi}_S$, $\hat{\psi}_S$ [για περισσότερες λεπτομέρειες βλέπε *Tomizawa* (1994)].

6.5.2 Γενίκευση των μέτρων απομάκρυνσης από την Συμμετρία (S) για ονοματικές μεταβλητές

Σε συνέχεια των ονοματικών μέτρων φ_S, ψ_S της προηγούμενης Παραγράφου 6.5.1, οι *Tomizawa et al.* (1998), πρότειναν μια γενίκευση αυτών. Υποθέτουμε ότι $\{\pi_{ij}+\pi_{ji}\}\geq 0$ για κάθε $i\neq j$.

A. Μέτρο απόκλισης τύπου power divergence Cressie & Read

Ένα γενικευμένο μέτρο που μετρά την απόσταση από την (S) για ονοματικές μεταβλητές, μπορεί να ορισθεί ως εξής

A1. $$\Phi_S^{(\lambda)}=\frac{\lambda(\lambda+1)}{2^{\lambda}-1}I^{(\lambda)}\left(\{\pi_{ij}^{*}\};\{\pi_{ij}^{*S}\}\right),\ \text{για}\ \lambda>-1$$

όπου $I^{(\lambda)}\left(\{\pi_{ij}^{*}\};\{\pi_{ij}^{*S}\}\right)=\frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{R}\sum_{\substack{j=1\\ j\neq i}}^{R}\pi_{ij}^{*}\left[\left(\frac{\pi_{ij}^{*}}{\pi_{ij}^{*S}}\right)^{\lambda}-1\right]$ η απόκλιση τύπου *power divergence Cressie & Read* μεταξύ των δεσμευμένων κατανομών $\{\pi_{ij}^{*}\}$ και $\{\pi_{ij}^{*S}\}$, δοθέντος ότι μια παρατήρηση θα πέσει σε ένα από τα μη – διαγώνια κελιά του τετραγωνικού πίνακα για $i\neq j$ και λ μια πραγματική τιμή που ορίζεται από τον χρήστη [για περισσότερες λεπτομέρειες αναφορικά με την απόκλιση *power divergence*, βλέπε *Cressie & Read* (1984) και *Read & Cressie* (1988)].

B. Μέτρο απόκλισης τύπου Patil & Taillie diversity index

Το γενικευμένο μέτρο $\Phi_S^{(\lambda)}$ μπορεί να εκφρασθεί ισοδύναμα και ως

B1. $$\Phi_S^{(\lambda)} = \sum_{i<j}\sum\left(\pi_{ij}^* + \pi_{ji}^*\right)\left[1-\frac{\lambda 2^{\lambda}}{2^{\lambda}-1}\right]H_{ij}^{(\lambda)}\left(\left\{\pi_{ij}^c, \pi_{ji}^c\right\}\right),\ \lambda > -1$$

όπου $H_{ij}^{(\lambda)}\left(\left\{\pi_{ij}^c, \pi_{ji}^c\right\}\right) = \frac{1}{\lambda}\left[1-\left(\pi_{ij}^c\right)^{\lambda+1}-\left(\pi_{ji}^c\right)^{\lambda+1}\right]$, ο δείκτης ποικιλότητας ή ανομοιότητας *Patil & Taillie* (*Patil & Taillie diversity index*) τάξης λ, για την δεσμευμένη κατανομή $\left\{\pi_{ij}^c, \pi_{ji}^c\right\}$, δοθέντων των ζευγών (ij, ji), ο οποίος περιλαμβάνει σε ειδικές περιπτώσεις, δηλαδή όταν $\lambda = 0$ και όταν $\lambda = 1$, την εντροπία κατά *Shannon* και τον δείκτη συγκέντρωσης *Gini concentration* (ή αλλιώς τον δείκτη *Simpson*), αντίστοιχα. Επομένως για κάθε λ, δοθέντος ότι μια παρατήρηση θα πέσει σε ένα από τα μη – διαγώνια κελιά του τετραγωνικού πίνακα, το γενικευμένο μέτρο $\Phi_S^{(\lambda)}$ αναπαριστά τον μέσο όρο του δείκτη ανομοιότητας $H_{ij}^{(\lambda)}\left(\left\{\pi_{ij}^c, \pi_{ji}^c\right\}\right)$ [για περισσότερες λεπτομέρειες αναφορικά με τον δείκτη ποικιλότητας ή ανομοιότητας (*diversity index*), βλέπε *Patil & Taillie* (1982), *Read & Cressie* (1988)].

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 3×3 πίνακα συνάφειας (*σελ. 168, Παράδειγμα 5α και 5β*), υπολογίζουμε το γενικευμένο μέτρο απόκλισης τύπου *power divergence Cressie & Read*, για διάφορες τιμές του λ.

λ	**$\Phi_S^{(\lambda)}$ Cressie and Read**	
	Παράδειγμα 5α *(ΑΥΓ vs ΟΚΤ 1971)*	*Παράδειγμα 5β (ΟΚΤ '71 vs ΔΕΚ '73)*
-0,8	0,009	0,060
-0,6	0,017	0,109
-0,4	0,024	0,149
0	0,035	0,207
1	0,048	0,268
1,4	0,049	0,273
1,6	0,049	0,273
2	0,048	0,268

Ερμηνεία

Παρατηρούμε, ότι για $\lambda = 0$ και $\lambda = 1$, οι τιμές του γενικευμένου μέτρου $\Phi_S^{(\lambda)}$, ταυτίζονται με αυτές των μέτρων φ_S (ψ_S) της παραγράφου 6.5.1. Η ερμηνεία των αποτελεσμάτων, το πεδίο ορισμού και το είδος των δεδομένων είναι τα ίδια.

Διάστημα Εμπιστοσύνης

Χρησιμοποιώντας την μέθοδο Δέλτα, η ποσότητα $\sqrt{n}\left(\hat{\Phi}_S^{(\lambda)} - \Phi_S^{(\lambda)}\right)$, ασυμπτωτικά $(n \to \infty)$ ακολουθεί την κανονική κατανομή με μέση τιμή $\mu = 0$ και διακύμανση

$$\sigma^2\left(\hat{\Phi}_S^{(\lambda)}\right) = \frac{1}{\delta^2}\left(\sum_{i=1}^{R}\sum_{\substack{j=1 \\ i\neq j}}^{R} \pi_{ij}\left(\Delta_{ij}^{(\lambda)}\right)^2 \delta\left(\Phi_S^{(\lambda)}\right)^2\right), \text{ για } \lambda > -1$$

όπου $\Delta_{ij}^{(\lambda)} = \begin{cases} \dfrac{1}{2^\lambda - 1}\left(2\pi_{ij}^c\right)^\lambda - 1 + \lambda\pi_{ji}^c\left[\left(2\pi_{ij}^c\right)^\lambda - \left(2\pi_{ji}^c\right)^\lambda\right], \ \lambda > -1, \ \lambda \neq 0 \\ \dfrac{1}{\log 2}\log\left(2\pi_{ij}^c\right), \ \lambda = 0 \end{cases}$

Για περισσότερες λεπτομέρειες αναφορικά με την μέθοδο Δέλτα [βλέπε *Bishop et al*. (1975) και *Agresti* (1984)].

Η δειγματική εκτίμηση του γενικευμένου μέτρου $\Phi_S^{(\lambda)}$ συμβολίζεται με $\hat{\Phi}_S^{(\lambda)}$ και υπολογίζεται αντικαθιστώντας τις πληθυσμιακές πιθανότητες $\{\pi_{ij}\}$ με τις παρατηρούμενες πιθανότητες του δείγματος $\{p_{ij}\}$, όπου $p_{ij} = \dfrac{n_{ij}}{n}$ και $n = \sum_{i=1}^{R}\sum_{j=1}^{R} n_{ij}$. Έστω $\hat{\sigma}^2\left(\hat{\Phi}_S^{(\lambda)}\right)$, συμβολίζει την εκτίμηση της διακύμανσης $\sigma^2\left(\Phi_S^{(\lambda)}\right)$ από τις παρατηρήσεις του δείγματος, τότε η ποσότητα $\hat{\sigma}\left(\hat{\Phi}_S^{(\lambda)}\right)\Big/\sqrt{n}$ είναι η κατά προσέγγιση εκτιμήτρια του τυπικού σφάλματος και το κατά προσέγγιση $100\% \times (1-a)$ διάστημα εμπιστοσύνης για την εκτιμήτρια του μέτρου $\Phi_S^{(\lambda)}$ είναι

$$\hat{\Phi}_S^{(\lambda)} \pm z_{a/2}\hat{\sigma}\left(\hat{\Phi}_S^{(\lambda)}\right)\Big/\sqrt{n} \qquad (6.4)$$

όπου $z_{a/2}$ το ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής.

Σχόλια

Σημειώνουμε ότι, $\Phi_S^{(0)} = \varphi_S$ και $\Phi_S^{(1)} = \psi_S$. Δηλαδή τα μέτρα Β1. φ_S και Β2. ψ_S που περιγράφονται στην παράγραφο 6.5.1, αποτελούν ειδικές περιπτώσεις του γενικευμένου μέτρου $\Phi_S^{(\lambda)}$ τύπου *power divergence Cressie & Read,* για $\lambda = 0$ και $\lambda = 1$. Επίσης, τα μέτρα Γ1. φ_S και Γ2. ψ_S, αποτελούν ειδικές περιπτώσεις του γενικευμένου μέτρου $\Phi_S^{(\lambda)}$ τύπου απόκλισης *Patil & Taillie,* για $\lambda = 0$ και $\lambda = 1$, αντίστοιχα. Και στις δυο κατηγορίες, για $\lambda = 0$, υπολογίζουμε το $\lim_{\lambda \to 0} \Phi_S^{(\lambda)}$. Σύμφωνα με την απόκλιση τύπου *power divergence Cressie & Read* και τύπου *Patil & Taillie,* το γενικευμένο μέτρο $\Phi_S^{(\lambda)}$ αναπαριστά τον βαθμό απομάκρυνσης από την (S) και ο βαθμός αυξάνεται καθώς το $\Phi_S^{(\lambda)}$ αυξάνεται [για περισσότερες λεπτομέρειες, βλέπε *Tomizawa et al*. (1998)].

6.5.3 Μέτρα απομάκρυνσης από την Συνολική Συμμετρία (GS) για διατακτικές μεταβλητές

Το μοντέλο (S) $(\pi_{ij} = \pi_{ji})$ εφαρμόζεται σε ονοματικές κατηγορικές μεταβλητές καθώς παραμένει αναλλοίωτο, κάτω από τις ίδιες διατάξεις των γραμμών ή στηλών του πίνακα. Στην περίπτωση διατακτικών μεταβλητών, καταλληλότερο είναι το μοντέλο της συνολικής συμμετρίας $(GS - Global\ Symmetry)$, το οποίο δεν παραμένει αναλλοίωτο στις ίδιες αναδιατάξεις των γραμμών και των στηλών, καθώς βασίζεται στην διάταξη των κατηγοριών. Ο *Read* (1977) όρισε το μοντέλο (GS) ως

$$\delta_U = \delta_L \tag{6.5}$$

όπου $\delta_U = \sum\sum_{i<j} \pi_{ij}$ $\left[= \Pr(X < Y)\right]$ και $\delta_L = \sum\sum_{i>j} \pi_{ij}$ $\left[= \Pr(X > Y)\right]$

Το μοντέλο αυτό εκφράζει το γεγονός ότι η πιθανότητα μια παρατήρηση να βρεθεί σε ένα από τα κελιά του άνω δεξιού τριγώνου του τετραγωνικού πίνακα, είναι ίση με την πιθανότητα ότι μια παρατήρηση θα βρεθεί σε ένα από τα κελιά του κάτω αριστερού τριγώνου.

Υποθέτουμε ότι $\delta_U + \delta_L > 0$ και έστω

$$\delta_U^* = \frac{\delta_U}{\delta_U + \delta_L} = \left[\Pr\left(X < Y \mid X \neq Y \right) \right] \tag{6.6}$$

η δεσμευμένη πιθανότητα μια παρατήρηση να βρίσκεται στο άνω δεξιό τρίγωνο του τετραγωνικού πίνακα, δοθέντος ότι η παρατήρηση θα βρίσκεται σε ένα (ανεξαρτήτου ποιο) από τα μη-διαγώνια κελιά του πίνακα, και

$$\delta_L^* = \frac{\delta_L}{\delta_U + \delta_L} = \left[\Pr\left(X > Y \mid X \neq Y \right) \right] \tag{6.7}$$

με $\delta_U^* + \delta_L^* = 1$

η πιθανότητα μια παρατήρηση να βρίσκεται στο κάτω αριστερό τρίγωνο, δοθέντος ότι η παρατήρηση θα βρίσκεται σε ένα (ανεξαρτήτου ποιο) από τα μη-διαγώνια κελιά του πίνακα. Σημειώνουμε ότι οι ανωτέρω μετασχηματισμοί είναι απαραίτητοι, προκειμένου τα αθροίσματα των πιθανοτήτων των μη διαγώνιων κελιών να αθροίζουν στην μονάδα, καθώς τα μέτρα δεν βασίζονται στις πιθανότητες της κεντρικής διαγώνιου.

Α. Μέτρο φ_{GS} τύπου πληροφορίας Kullback-Leibler και ψ_{GS} τύπου απόκλισης Pearson

Ο *Tomizawa* (1995) εξέτασε δυο ειδών μέτρα, με την ίδια διατακτική ταξινόμηση γραμμών και στηλών, τα οποία εκφράζονται χρησιμοποιώντας την πληροφορία *Kullback - Leibler* και την απόκλιση $\chi^2 - Pearson$. Τα μέτρα αυτά ορίζονται ως:

Α1. $$\varphi_{GS} = \frac{1}{\log 2} I\left(\left\{ \delta_U^*, \delta_L^* \right\}; \left\{ \frac{1}{2}, \frac{1}{2} \right\} \right) = 1 + \frac{1}{\log 2} \left(\delta_U^* \log \delta_U^* + \delta_L^* \log \delta_L^* \right)$$

όπου $I\left(\left\{ \delta_U^*, \delta_L^* \right\}; \left\{ \frac{1}{2}, \frac{1}{2} \right\} \right) = \delta_U^* \log \frac{\delta_U^*}{1/2} + \delta_L^* \log \frac{\delta_L^*}{1/2}$, η πληροφορία *Kullback - Leibler* μεταξύ των δυο κατανομών $\left\{ \delta_U^*, \delta_L^* \right\}$ και $\left\{ \frac{1}{2}, \frac{1}{2} \right\}$.

A2. $$\psi_{GS} = D\left(\left\{\delta_U^*, \delta_L^*\right\}; \left\{\frac{1}{2}, \frac{1}{2}\right\}\right) = 2\left(\delta_U^* + \delta_L^*\right) - 1$$

όπου $D\left(\left\{\delta_U^*, \delta_L^*\right\}; \left\{\frac{1}{2}, \frac{1}{2}\right\}\right) = \frac{\left(\delta_U^* - \frac{1}{2}\right)^2}{1/2} + \frac{\left(\delta_L^* - \frac{1}{2}\right)^2}{1/2}$, η απόκλιση $\chi^2 - Pearson$, μεταξύ των δυο κατανομών $\left\{\delta_U^*, \delta_L^*\right\}$ και $\left\{\frac{1}{2}, \frac{1}{2}\right\}$.

Σημειώνουμε ότι $\delta_U^* = \delta_L^* = \frac{1}{2}$, όταν ο πίνακας είναι συνολικά συμμετρικός. Επίσης παρατηρούμε ότι στην ακραία περίπτωση, όπου όλες οι παρατηρήσεις συγκεντρώνονται στο άνω τρίγωνο, τότε για το μέτρο φ_{GS} έχουμε ότι $\delta_U^* = 1$ (και $\delta_L^* = 0$), με $\delta_U^* \log \frac{\delta_U^*}{1/2} = \log 2$ και $\varphi_{GS} = 1$.

Β. Μέτρα τύπου εντροπίας Shannon και δείκτης Gini Concentration

Τα μέτρα Α1, Α2, μπορούν περαιτέρω να εκφρασθούν ισοδύναμα, χρησιμοποιώντας την εντροπία *Shannon* και τον δείκτη *Gini Concentration*, για την δεσμευμένη κατανομή $\left\{\delta_U^*, \delta_L^*\right\}$, ως εξής:

Β1. $$\varphi_{GS} = 1 - \frac{1}{\log 2} H\left(\left\{\delta_U^*, \delta_L^*\right\}\right),$$

όπου $H\left(\left\{\delta_U^*, \delta_L^*\right\}\right) = H\left(\delta_U^*\right) + H\left(\delta_L^*\right) = -\delta_U^* \log \delta_U^* - \delta_L^* \log \delta_L^*$ η εντροπία *Shannon* μεταξύ των κατανομών $\left\{\delta_U^*, \delta_L^*\right\}$.

Β2. $$\psi_{GS} = 1 - 2C\left(\left\{\delta_U^*, \delta_L^*\right\}\right),$$

όπου $C\left(\left\{\delta_U^*, \delta_L^*\right\}\right) = 1 - \left(\delta_U^{*2} + \delta_L^{*2}\right)$, ο δείκτης *Gini Concentration* μεταξύ των κατανομών $\left\{\delta_U^*, \delta_L^*\right\}$.

Για περισσότερες λεπτομέρειες αναφορικά με τον δείκτη *Gini Concentration* [βλέπε *Haberman*, (1982)]. Εύκολα διαπιστώνουμε, ότι στην ακραία περίπτωση όπου όλες οι παρατηρήσεις

συγκεντρώνονται στο άνω τρίγωνο, τότε για το μέτρο φ_{GS}, έχουμε ότι $H\left(\left\{\delta_U^*, \delta_L^*\right\}\right) = 0$ και $\varphi_{GS} = 0$.

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 8×8 πίνακα συνάφειας (*σελ.169, Παράδειγμα 6*), έχουμε τα κάτωθι αποτελέσματα. Σημειώνουμε ότι οι 2 διαφορετικοί τύποι υπολογισμού A1, B1 του μέτρου φ_{GS} είναι ισοδύναμοι. Ομοίως, οι τύποι A2, B2 για το μέτρο ψ_{GS}.

Παράδειγμα 6

		1955	***1965***	***1975***
A1. Kullback-Leibler	**φ_{GS}**	0,087	0,177	0,159
A2. Pearson discrepancy	**ψ_{GS}**	0,118	0,235	0,212
B1. Shannon entropy	**φ_{GS}**	0,087	0,177	0,159
B2. Gini concentration	**ψ_{GS}**	0,118	0,235	0,212

Ερμηνεία

Σε έναν πίνακα συνάφειας υπάρχει πλήρης ασυμμετρία, δηλαδή ο βαθμός απομάκρυνσης από την (GS) μεγιστοποιείται, υπό την έννοια ότι, όταν $\delta_U = 0$, τότε $\delta_L > 0$ ή όταν $\delta_L = 0$, τότε $\delta_U > 0$, αν και μόνο αν $\varphi_{GS} = 1$ $(\psi_{GS} = 1)$. Ο τετραγωνικός πίνακας είναι συνολικά συμμετρικός, για $i, j = 1, 2, ..., R$ και $i \neq j$, αν και μόνο αν $\varphi_{GS} = 0$ $(\psi_{GS} = 0)$. Επομένως, σύμφωνα με τα αποτελέσματα για τα έτη 1955, 1965, 1975, το μέτρο φ_{GS} εκφράζει ότι ο βαθμός απομάκρυνσης από την συμμετρία είναι 0.087, 0.177 και 0.159. Μπορούμε να πούμε ότι το έτος 1955, το επάγγελμα του υιού, σχεδόν ταυτίζεται με του πατέρα, ενώ κατά τα έτη 1965 και 1975 υπάρχει μεγαλύτερη ασυνέπεια.

Πεδίο Ορισμού

Τα μέτρα φ_{GS} και ψ_{GS} κυμαίνονται στο διάστημα $[0,1]$.

Επίπεδο Δεδομένων

Τα μέτρα φ_{GS} και ψ_{GS} είναι κατάλληλα για διατακτικές μεταβλητές, καθώς δεν παραμένουν αναλλοίωτα κάτω από τους ίδιες διατάξεις γραμμών ή στηλών του πίνακα.

Διάστημα Εμπιστοσύνης

Χρησιμοποιώντας την μέθοδο Δέλτα, οι ποσότητες $\sqrt{n}\left(\hat{\varphi}_{GS}-\varphi_{GS}\right)$ και $\sqrt{n}\left(\hat{\psi}_{GS}-\psi_{GS}\right)$, ασυμπτωτικά $(n\to\infty)$ ακολουθούν την κανονική κατανομή με μέση τιμή $\mu=0$ και διακύμανση

A1. $$\sigma^2\left(\varphi_{GS}\right)=\frac{\delta_U\delta_L}{\left(\log 2\right)^2\left(\delta_U+\delta_L\right)^3}\left\{\log\frac{\delta_U}{\delta_L}\right\}^2$$

και

A2. $$\sigma^2\left(\psi_{GS}\right)=\frac{16\delta_U\delta_L\left(\delta_U-\delta_L\right)^2}{\left(\delta_U+\delta_L\right)^5}$$

Για περισσότερες λεπτομέρειες αναφορικά με την μέθοδο Δέλτα [βλέπε *Bishop et al*. (1975) και *Agresti* (1984)].

Οι δειγματικές εκτιμήσεις των μέτρων φ_{GS} και ψ_{GS} συμβολίζονται με $\hat{\varphi}_{GS}$, $\hat{\psi}_{GS}$ και υπολογίζονται αντικαθιστώντας τις πληθυσμιακές πιθανότητες $\left\{\pi_{ij}\right\}$ με τις παρατηρούμενες πιθανότητες του δείγματος $\left\{p_{ij}\right\}$, όπου $p_{ij}=\frac{n_{ij}}{n}$ και $n=\sum_{i=1}^{R}\sum_{j=1}^{R}n_{ij}$. Έστω, $\hat{\sigma}^2\left(\hat{\varphi}_{GS}\right)$ συμβολίζει την εκτίμηση της διακύμανσης $\sigma^2\left(\varphi_{GS}\right)$ από τις παρατηρήσεις του δείγματος, τότε η ποσότητα $\hat{\sigma}\left(\hat{\varphi}_{GS}\right)/\sqrt{n}$ είναι η κατά προσέγγιση εκτιμήτρια του τυπικού σφάλματος και το κατά προσέγγιση $100\%\times(1-a)$ διάστημα εμπιστοσύνης για την εκτιμήτρια του μέτρο φ_{GS} είναι

$$\hat{\varphi}_{GS}\pm z_{a/2}\hat{\sigma}\left(\hat{\varphi}_{GS}\right)/\sqrt{n} \tag{6.8}$$

όπου $z_{a/2}$ το ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής.

Κατά τον ίδιο τρόπο υπολογίζουμε και το διάστημα εμπιστοσύνης για το μέτρο $\hat{\psi}_{GS}$.

Σχόλια

Σύμφωνα με την πληροφορία *Kullback-Leibler* και την ανακολουθία τύπου $\chi^2 - Pearson$, οι ποσότητες φ_{GS} (ψ_{GS}) αναπαριστούν τον βαθμό απομάκρυνσης από την (GS), υπό την συνθήκη ότι μια παρατήρηση θα πέσει σε ένα από τα μη-διαγώνια κελιά του πίνακα και ο βαθμός αυξάνεται καθώς οι τιμές των φ_{GS} (ψ_{GS}) αυξάνονται [βλέπε *Tomizawa* (1995)].

6.5.4 Μέτρα απομάκρυνσης από την Ψευδοσυμμετρία (QS) για ονοματικές μεταβλητές

Όπως είδαμε στην παράγραφο 6.4.3, το μοντέλο (QS) μπορεί να εκφρασθεί ως

$$D_{ijk} = D_{kji} \text{ για } i, j, k = 1, 2, ..., R,$$

όπου $D_{ijk} = \pi_{ij}\pi_{jk}\pi_{ki}$ και $D_{kji} = \pi_{kj}\pi_{ji}\pi_{ik}$

και σε όρους των *τοπικών odds ratio* ως

$$\theta_{ij} = \theta_{ji}, \text{ όπου } \theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}} \text{ για } i, j = 1, ..., R-1.$$

Σύμφωνα με τους *Tahata et al*. (2001), το μοντέλο της (QS) μπορεί να εκφρασθεί ως

$$D_{ijk} = D_{kji} \text{ για } i < j < k \tag{6.9}$$

όπου $D_{ijk} = \pi_{ij}\pi_{jk}\pi_{ki}$ και $D_{kji} = \pi_{kj}\pi_{ji}\pi_{ik}$

και σε όρους των *τοπικών odds ratio* ως

$$\theta_{(i<j;\, j<k)} = \theta_{(j<k;\, i<j)}, \; i < j < k \tag{6.10}$$

όπου $\theta_{(i<j;\, j<k)} = \frac{\pi_{ij}\pi_{jk}}{\pi_{jj}\pi_{ik}}$ και $\theta_{(j<k;\, i<j)} = \frac{\pi_{ji}\pi_{kj}}{\pi_{ki}\pi_{jj}}$ [βλέπε επίσης *Menendez et al*. (2005)].

Χρησιμοποιώντας τις δεσμευμένες πιθανότητες $\pi_{ij}^{C} = \frac{\pi_{ij}}{\pi_{ij} + \pi_{ji}}$, δηλαδή την πιθανότητα ότι μια παρατήρηση βρίσκεται στο κελί (i, j), δοθέντος ότι η παρατήρηση θα βρίσκεται σε ένα από τα μη διαγώνια κελιά (i, j) ή (j, i), για $i \neq j$, τότε το μοντέλο (QS) μπορεί να εκφρασθεί ως

$$Q_{ijk} = Q_{kji}\text{, για } i < j < k \tag{6.11}$$

όπου $Q_{ijk} = \pi_{ij}^{C}\pi_{jk}^{C}\pi_{ki}^{C}$ και $Q_{kji} = \pi_{kj}^{C}\pi_{ji}^{C}\pi_{ik}^{C}$

Οι *Tahata et al.* (2001), πρότειναν ένα γενικευμένο μέτρο $\Phi_{QS}^{(\lambda)}$, που αναπαριστά τον βαθμό απομάκρυνσης από την (QS), για έναν RxR πίνακα με ονοματικές μεταβλητές ταξινόμησης, τα οποία εκφράζονται χρησιμοποιώντας την απόκλιση τύπου *power divergence Cressie & Read* και τον δείκτη απόκλισης *Patil & Taillie.*

Έστω $\Delta = \sum_{i<j<k}\left(Q_{ijk} + Q_{kji}\right)$ και $Q_{ijk}^{*} = \frac{Q_{ijk}}{\Delta}$, $Q_{kji}^{*} = \frac{Q_{kji}}{\Delta}$, $C_{ijk}^{*} = C_{kji}^{*} = \frac{1}{2}\left(Q_{ijk}^{*} + Q_{kji}^{*}\right)$, για $i < j < k$

A. Μέτρο απόκλισης τύπου power divergence Cressie & Read

Υποθέτοντας ότι $Q_{ijk} + Q_{kji} \neq 0$, για $i < j < k$, ένα γενικευμένο μέτρο που μετρά την απόσταση από την (QS) για ονοματικές μεταβλητές, μπορεί να ορισθεί ως εξής

A1. $$\Phi_{QS}^{(\lambda)} = \frac{\lambda(\lambda+1)}{2^{\lambda}-1} I^{(\lambda)}\left(\left\{Q_{ijk}^{*}\right\};\left\{C_{ijk}^{*}\right\}\right), \text{ για } \lambda > -1$$

όπου λ μια πραγματική τιμή που ορίζεται από τον χρήστη και όπου $I^{(\lambda)}\left(\left\{Q_{ijk}^{*}\right\};\left\{C_{ijk}^{*}\right\}\right) = \frac{1}{\lambda(\lambda+1)}\sum_{i<j<k}^{R}\left[Q_{ijk}^{*}\left\{\left(\frac{Q_{ijk}^{*}}{C_{ijk}^{*}}\right)^{\lambda}-1\right\} + Q_{kji}^{*}\left\{\left(\frac{Q_{kji}^{*}}{C_{kji}^{*}}\right)^{\lambda}-1\right\}\right]$ η απόκλιση τύπου *power divergence Cressie & Read* μεταξύ των κατανομών $\left\{Q_{ijk}^{*}\right\}$ και $\left\{C_{ijk}^{*}\right\}$, για $i < j < k$ ή για $i > j > k$.

Για $\lambda = 0$, υπολογίζουμε το $\lim_{\lambda\to 0}\Phi_{QS}^{(\lambda)}$ και έχουμε

$$\Phi_{QS}^{(0)} = \frac{1}{\log(2)} I^{(0)}\left(\left\{Q_{ijk}^{*}\right\};\left\{C_{ijk}^{*}\right\}\right),$$

όπου $I^{(0)}\left(\left\{Q_{ijk}^{*}\right\};\left\{C_{ijk}^{*}\right\}\right) = \sum_{i<j<k}^{R}\left[Q_{ijk}^{*}\log\left(\frac{Q_{ijk}^{*}}{C_{ijk}^{*}}\right) + Q_{kji}^{*}\log\left(\frac{Q_{kji}^{*}}{C_{kji}^{*}}\right)\right]$.

Σημειώνουμε ότι $I^{(0)}\left(\left\{Q^*_{ijk}\right\};\left\{C^*_{ijk}\right\}\right)$ είναι η πληροφορία *Kullback - Leibler* μεταξύ των κατανομών $\left\{Q^*_{ijk}\right\}$ και $\left\{C^*_{ijk}\right\}$, για $i<j<k$ ή για $i>j>k$. Επίσης, όταν ισχύει το μοντέλο της $\left(QS\right)$ τότε $I^{(\lambda)}=0$.

B. Μέτρο απόκλισης τύπου Patil & Taillie diversity index

Χρησιμοποιώντας τις δεσμευμένες πιθανότητες $Q^C_{ijk}=\dfrac{Q_{ijk}}{Q_{ijk}+Q_{kji}}$, $Q^C_{kji}=\dfrac{Q_{kji}}{Q_{ijk}+Q_{kji}}$, για $i<j<k$, τότε το μέτρο μπορεί να εκφρασθεί ως εξής

B1. $$\Phi^{(\lambda)}_{QS}=1-\frac{\lambda 2^{\lambda}}{2^{\lambda}-1}\sum_{i<j<k}\left(Q^*_{ijk}+Q^*_{kji}\right)H^{(\lambda)}_{ijk}\left(\left\{Q^C_{ijk},Q^C_{kji}\right\}\right),\ \lambda>-1,$$

όπου $H^{(\lambda)}_{ijk}\left(\left\{Q^C_{ijk},Q^C_{kji}\right\}\right)=\dfrac{1}{\lambda}\left[1-\left(Q^C_{ijk}\right)^{\lambda+1}-\left(Q^C_{kji}\right)^{\lambda+1}\right]$, ο δείκτης ανομοιότητας *Patil & Taillie* τάξης λ, για την δεσμευμένη κατανομή $\left\{Q^C_{ijk},Q^C_{kji}\right\}$, ο οποίος περιλαμβάνει σε ειδικές περιπτώσεις, δηλαδή όταν $\lambda=0$ και όταν $\lambda=1$, την εντροπία κατά *Shannon* και τον δείκτη *Gini Concentration*, αντίστοιχα. Για $\lambda=0$, υπολογίζουμε το $\lim\limits_{\lambda\to 0}\Phi^{(\lambda)}_{QS}$ και έχουμε

$$\Phi^{(0)}_{QS}=1-\frac{1}{\log(2)}\sum_{i<j<k}\left(Q^*_{ijk}+Q^*_{kji}\right)H^{(0)}_{ijk}\left(\left\{Q^C_{ijk},Q^C_{kji}\right\}\right),$$

όπου $H^{(0)}_{ijk}\left(\left\{Q^C_{ijk},Q^C_{kji}\right\}\right)=-Q^C_{ijk}\log\left(Q^C_{ijk}\right)-Q^C_{kji}\log\left(Q^C_{kji}\right)$.

Το γενικευμένο μέτρο $\Phi^{(\lambda)}_{QS}$ αναπαριστά το σταθμισμένο άθροισμα του δείκτη ανομοιότητας $H^{(\lambda)}_{ijk}\left(\left\{Q^C_{ijk},Q^C_{kji}\right\}\right)$.

Διάστημα Εμπιστοσύνης του γενικευμένου μέτρου $\Phi^{(\lambda)}_{QS}$

Χρησιμοποιώντας την μέθοδο Δέλτα, η ποσότητα $\sqrt{n}\left(\hat{\Phi}^{(\lambda)}_{QS}-\Phi^{(\lambda)}_{QS}\right)$, ασυμπτωτικά $(n\to\infty)$ ακολουθεί την κανονική κατανομή με μέση τιμή $\mu=0$ και διακύμανση

A1. $$\sigma^2\left(\Phi^{(\lambda)}_{QS}\right)=d_2\left(\pi\right)\Sigma_1\left(\pi\right)d_2\left(\pi\right)'$$

όπου π το $1\times R^2$- διάνυσμα των πιθανοτήτων $\pi=\left(\pi_{(12)},\pi_{(13)},...,\pi_{(R-1,R)},\pi_{(11)},\pi_{(22)},...,\pi_{(RR)}\right)$, το οποίο ασυμπτωτικά $(n\to\infty)$ ακολουθεί την κανονική κατανομή με $N\left(\pi,\Sigma_1(\pi)\right)$ και $\Sigma_1(\pi)=\frac{1}{n}\left(D(\pi)-\pi\pi'\right)$ ο $R^2\times R^2$ πίνακας, με $D(\pi)$ τον διαγώνιο πίνακα $D(\pi)=\begin{vmatrix}\pi_{11} & 0 & \dots & 0\\ 0 & \pi_{22} & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & \pi_{RR}\end{vmatrix}$, $d_2(\pi)=\partial\Phi_{QS}^{(\lambda)}/\partial\pi'$ το $1\times R^2$- διάνυσμα και $d_2(\pi)'$ η αντιμετάθεση του διανύσματος.

Για περισσότερες λεπτομέρειες αναφορικά με την μέθοδο Δέλτα [βλέπε *Bishop et al*. (1975) και *Agresti* (1984)].

Σημειώνουμε ότι η δειγματική εκτίμηση του γενικευμένου μέτρου $\Phi_{QS}^{(\lambda)}$ συμβολίζεται με $\hat{\Phi}_{QS}^{(\lambda)}$ και υπολογίζεται αντικαθιστώντας τις πληθυσμιακές πιθανότητες $\{\pi_{ij}\}$ με τις παρατηρούμενες πιθανότητες του δείγματος $\{p_{ij}\}$, όπου $p_{ij}=\frac{n_{ij}}{n}$ και $n=\sum_{i=1}^{R}\sum_{j=1}^{R}n_{ij}$. Έστω $\hat{\sigma}^2\left(\hat{\Phi}_{QS}^{(\lambda)}\right)$ συμβολίζει την εκτίμηση της διακύμανσης $\sigma^2\left(\Phi_{QS}^{(\lambda)}\right)$ από τις παρατηρήσεις του δείγματος, τότε η ποσότητα $\hat{\sigma}\left(\hat{\Phi}_{QS}^{(\lambda)}\right)/\sqrt{n}$ είναι η κατά προσέγγιση εκτιμήτρια του τυπικού σφάλματος και το κατά προσέγγιση $100\%\times(1-a)$ διάστημα εμπιστοσύνης για την εκτιμήτρια του μέτρου $\Phi_{QS}^{(\lambda)}$ είναι

$$\hat{\Phi}_{QS}^{(\lambda)}\pm z_{a/2}\hat{\sigma}\left(\hat{\Phi}_{QS}^{(\lambda)}\right)/\sqrt{n} \tag{6.12}$$

όπου $z_{a/2}$ το ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής.

Ιδιότητες του γενικευμένου μέτρου $\Phi_{QS}^{(\lambda)}$, ονοματικής ταξινόμησης

1. Παρατηρώντας ότι $I^{(\lambda)}\left(\{Q_{ijk}^*\};\{C_{ijk}^*\}\right)\geq 0$ και $H_{ijk}^{(\lambda)}\left(\{Q_{ijk}^C,Q_{kji}^C\}\right)\geq 0$, συμπεραίνουμε ότι το γενικευμένο μέτρο $\Phi_{QS}^{(\lambda)}$ κυμαίνεται στο διάστημα $[0,1]$.
2. Για κάθε $\lambda>-1$, ο τετραγωνικός πίνακας είναι συμμετρικός, για $i,j=1,2,...,R$ και $i\neq j$, αν και μόνο αν $\Phi_{QS}^{(\lambda)}=0$.

3. Σε έναν πίνακα συνάφειας υπάρχει πλήρης ασυμμετρία, υπο την έννοια ότι ο βαθμός απομάκρυνσης από την (QS) μεγιστοποιείται, δηλαδή όταν $Q_{ijk}^{C}=0$ (οπότε $Q_{kji}^{C}=1$) ή όταν $Q_{kji}^{C}=0$ (οπότε $Q_{ijk}^{C}=1$), για κάθε $i<j<k$, αν και μόνο αν $\Phi_{QS}^{(\lambda)}=1$.
4. Όταν $\Phi_{QS}^{(\lambda)}=1$ τότε για κάθε $i<j<k$, ισχύει ότι $\pi_{ij}\pi_{jk}\pi_{ki}=0$ ή $\pi_{ji}\pi_{kj}\pi_{ik}=0$. Επομένως, για κάθε $i<j<k$, τουλάχιστον μια από τις πιθανότητες π_{ij}^{C}, π_{jk}^{C}, π_{ki}^{C} ισούνται με 0, ή τουλάχιστον μια από τις πιθανότητες π_{ji}^{C}, π_{kj}^{C}, π_{ik}^{C} ισούνται με 0. Δηλαδή για κάθε $i<j<k$, υπάρχει πλήρης ασυμμετρία, όταν για παράδειγμα $\pi_{ij}^{C}=0$ και $\pi_{ji}^{C}=1$, για τουλάχιστον ένα ζεύγος συμμετρικών κελιών, γεγονός που εκφράζει την μερική πλήρης ασυμμετρία των πιθανοτήτων των κελιών.
5. Σε όρους των *odds ratio*, για $\Phi_{QS}^{(\lambda)}=1$ αυτό σημαίνει ότι όταν $Q_{ijk}=0$ (οπότε $Q_{kji}>0$) ή όταν $Q_{kji}=0$ (οπότε $Q_{ijk}>0$) και κατ' επέκταση όταν $D_{ijk}=0$ (οπότε $D_{kji}>0$) ή όταν $D_{kji}=0$ (οπότε $D_{ijk}>0$), που σημαίνει ότι $D_{ijk}/D_{kji}=0$ ή $D_{ijk}/D_{kji}=\infty$, για $i<j<k$ και υποδεικνύει την πλήρης ασυμμετρία των *odds ratio*. Υπενθυμίζουμε ότι το μοντέλο (QS) ισχύει όταν $\dfrac{\theta_{(i<j;j<k)}}{\theta_{(j<k;i<j)}}=\dfrac{D_{ijk}}{D_{kji}}=1$
6. Σύμφωνα με την απόκλιση τύπου *power divergence* και τον δείκτη ανομοιότητας *Patil & Taillie*, το γενικευμένο μέτρο $\Phi_{QS}^{(\lambda)}$ αναπαριστά τον βαθμό απομάκρυνσης από την (QS) και ο βαθμός αυξάνεται καθώς το $\Phi_{QS}^{(\lambda)}$ αυξάνεται.

6.5.5 Μέτρα απομάκρυνσης από την Ομοιογένεια Περιθωρίου (MH) για ονοματικές μεταβλητές

Ο *Tomizawa* (1995) πρότεινε 4 ειδών μέτρα, που αναπαριστούν τον βαθμό απομάκρυνσης από την (MH), για έναν *RxR* πίνακα με ονοματικές μεταβλητές ταξινόμησης. Δυο από τα προτεινόμενα μέτρα είναι συναρτήσεις των περιθώριων πιθανοτήτων $\{\pi_{i.}\}$ και $\{\pi_{.i}\}$ και τα άλλα

δυο είναι συναρτήσεις των δεσμευμένων περιθώριων κατανομών $\left\{\pi_{i.}^{C}\right\}$ και $\left\{\pi_{.i}^{C}\right\}$, δοθέντος οτι $X_1 \neq X_2$

6.5.5.1 Μέτρα που βασίζονται στις αδέσμευτες περιθώριες κατανομές πιθανότητας

Τα δυο μέτρα φ_{MH} και ψ_{MH} χρησιμοποιούνται όταν το μοντέλο περιθώριας ομοιογένειας εκφράζεται ως $\pi_{i.} = \pi_{.i}$ και μετρούν τον βαθμό απομάκρυνσης των αδέσμευτων περιθώριων κατανομών από την (MH). Τα μέτρα αυτά εκφράζονται μέσω των ακόλουθων Α, Β, Γ ισοδύναμων ορισμών. Έστω ότι $\left\{\pi_{i.} + \pi_{.i} > 0\right\}$, έχουμε

Α. Μέτρο φ_{MH} τύπου πληροφορίας Kullback-Leibler και ψ_{MH} τύπου απόκλισης Pearson

Α1.
$$\varphi_{MH} = \frac{1}{2\log 2}\left[I\left(\left\{\pi_{i.}\right\},\left\{\pi_{i}^{*}\right\}\right) + I\left(\left\{\pi_{.i}\right\},\left\{\pi_{i}^{*}\right\}\right)\right],$$

όπου $\pi_i^* = \left(\pi_{i.} + \pi_{.i}\right)/2$ και $I\left(\left\{\pi_{i.}\right\},\left\{\pi_{i}^{*}\right\}\right)$, $I\left(\left\{\pi_{.i}\right\},\left\{\pi_{i}^{*}\right\}\right)$ η πληροφορία *Kullback-Leibler* μεταξύ των κατανομών $\left\{\pi_{i.}\right\}$ και $\left\{\pi_{i}^{*}\right\}$ και μεταξύ των κατανομών $\left\{\pi_{.i}\right\}$ και $\left\{\pi_{i}^{*}\right\}$, αντίστοιχα. Υπενθυμίζουμε ότι η πληροφορία *Kullback-Leibler* μεταξύ δυο κατανομών πιθανότητας $\left\{a_i\right\}$ και $\left\{b_i\right\}$, ορίζεται ως $I\left(\left\{a_i\right\};\left\{b_i\right\}\right) = \sum_{i=1}^{R} a_i \log\left(a_i/b_i\right)$.

Α2.
$$\psi_{MH} = \frac{1}{2}\left[D\left(\left\{\pi_{i.}\right\},\left\{\pi_{i}^{*}\right\}\right) + D\left(\left\{\pi_{.i}\right\},\left\{\pi_{i}^{*}\right\}\right)\right]$$

όπου $\pi_i^* = \left(\pi_{i.} + \pi_{.i}\right)/2$ και $D\left(\left\{\pi_{i.}\right\},\left\{\pi_{i}^{*}\right\}\right)$, $D\left(\left\{\pi_{.i}\right\},\left\{\pi_{i}^{*}\right\}\right)$ η απόκλιση $\chi^2 - Pearson$ μεταξύ των κατανομών $\left\{\pi_{i.}\right\}$ και $\left\{\pi_{i}^{*}\right\}$ και μεταξύ των κατανομών $\left\{\pi_{.i}\right\}$ και $\left\{\pi_{i}^{*}\right\}$, αντίστοιχα. Υπενθυμίζουμε ότι η απόκλιση $\chi^2 - Pearson$ μεταξύ δυο κατανομών πιθανότητας $\left\{a_i\right\}$ και $\left\{b_i\right\}$, ορίζεται ως $D\left(\left\{a_i\right\},\left\{b_i\right\}\right) = \sum_{i=1}^{R} \frac{\left(a_i - b_i\right)^2}{b_i}$.

Το μέτρο φ_{MH} εκφράζει στην ουσία το άθροισμα της πληροφορίας *Kullback-Leibler* μεταξύ των κατανομών $\{\pi_{i.}\}$ και $\{\pi_i^*\}$ και αυτής μεταξύ των κατανομών $\{\pi_{.i}\}$ και $\{\pi_i^*\}$, ενώ το μέτρο ψ_{MH} εκφράζει το άθροισμα της απόκλισης $\chi^2 - Pearson$, αντίστοιχα.

Β. Μέτρο φ_{MH} τύπου εντροπίας του Shannon και ψ_{MH} τύπου Gini Concentration

Τα προηγούμενα μέτρα μπορούν ισοδύναμα να εκφρασθούν και ως

Β1. $$\varphi_{MH} = 1 - \frac{1}{\log 2}\sum_{i=1}^{R} \pi_i^* H_i\left(\{\pi_{k(i)}\}\right),$$

όπου $\pi_i^* = \left(\pi_{i.} + \pi_{.i}\right)/2$ και $k = 1,2$

Β2. $$\psi_{MH} = 1 - 2\sum_{i=1}^{R} \pi_i^* C_i\left(\{\pi_{k(i)}\}\right)$$

όπου $\pi_{1(i)} = \frac{\pi_{i.}}{\pi_{i.} + \pi_{.i}}$, $\pi_{2(i)} = \frac{\pi_{.i}}{\pi_{i.} + \pi_{.i}}$, $H_i\left(\{\pi_{k(i)}\}\right) = -\sum_{k=1}^{2} \pi_{k(i)} \log \pi_{k(i)}$ είναι η εντροπία *Shannon* και $C_i\left(\{\pi_{k(i)}\}\right) = 1 - \sum_{k=1}^{2} \pi_{k(i)}^2$ είναι ο δείκτης συγκέντρωσης *Gini* (*Gini Concentration*).

Το μέτρο φ_{MH} αναπαριστά το σταθμισμένο άθροισμα της εντροπίας *Shannon*, ενώ το μέτρο ψ_{MH} αναπαριστά το σταθμισμένο άθροισμα του δείκτη *Gini concentration*. Για περισσότερες λεπτομέρειες αναφορικά με τον δείκτη *Gini Concentration* [βλέπε *Haberman*, (1982)].

Γ. Μέτρο φ_{MH} σταθμισμένης πληροφορίας Kullback-Leibler και ψ_{MH} σταθμισμένης απόκλισης Pearson

Γ1. $$\varphi_{MH} = \frac{1}{\log 2}\sum_{i=1}^{R} \pi_i^* I_i\left(\{\pi_{k(i)}\}; \left\{\frac{1}{2}\right\}\right)$$

όπου $I_i\left(\{\pi_{k(i)}\}; \left\{\frac{1}{2}\right\}\right) = \sum_{k=1}^{2} \pi_{k(i)} \log\left(\frac{\pi_{k(i)}}{1/2}\right)$

Γ2. $$\psi_{MH} = \sum_{i=1}^{R} \pi_i^* D_i \left(\left\{ \pi_{k(i)} \right\}; \left\{ \frac{1}{2} \right\} \right)$$

όπου $D_i \left(\left\{ \pi_{k(i)} \right\}; \left\{ \frac{1}{2} \right\} \right) = \sum_{k=1}^{2} \frac{\left(\pi_{k(i)} - 1/2 \right)^2}{1/2}$

Το μέτρο φ_{MH} εκφράζει στην ουσία το σταθμισμένο άθροισμα της πληροφορίας *Kullback-Leibler* μεταξύ των δυο κατανομών $\left\{ \pi_{1(i)}, \pi_{2(i)} \right\}$ και $\left\{ \frac{1}{2}, \frac{1}{2} \right\}$, ενώ το μέτρο ψ_{MH} εκφράζει το σταθμισμένο άθροισμα της ανακολουθίας $\chi^2 - Pearson$, αντίστοιχα. Σημειώνουμε ότι οι περιθώριες κατανομές $\left\{ \pi_{1(i)}, \pi_{2(i)} \right\}$ είναι ταυτόσημες των $\left\{ \frac{1}{2}, \frac{1}{2} \right\}$ όταν το μοντέλο της (MH) ισχύει.

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 3×3 πίνακα συνάφειας (*σελ. 168, Παράδειγμα 5α και 5β*), έχουμε τα κάτωθι αποτελέσματα. Σημειώνουμε ότι οι 3 διαφορετικοί τύποι υπολογισμού Α1, Β1, Γ1, του μέτρου φ_{MH} είναι ισοδύναμοι. Ομοίως, οι τύποι Α2, Β2, Γ2, για το μέτρο ψ_{MH}.

Παράδειγμα 5		**Παράδειγμα 5α** ***(ΑΥΓ vs ΟΚΤ 1971)***	***Παράδειγμα 5β (ΟΚΤ '71 vs ΔΕΚ '73)***
Α1. Kullback-Leibler	φ_{MH}	0,003	0,030
Α2. Pearson discrepancy	ψ_{MH}	0,004	0,040
Β1. Shannon entropy	φ_{MH}	0,003	0,030
Β2. Gini Concentration	ψ_{MH}	0,004	0,040
Γ1. Weighted Kullback-Leibler	φ_{MH}	0,003	0,030
Γ2. Weighted Pearson discrepancy	ψ_{MH}	0,004	0,040

Note: The results are based on unconditional probabilities

Ερμηνεία

Ο βαθμός απομάκρυνσης από την (MH) γίνεται μέγιστος, υπό την έννοια ότι, όταν $\pi_{i.} = 0$ τότε $\pi_{.i} > 0$ ή όταν $\pi_{.i} = 0$ τότε $\pi_{i.} > 0$, για όλα τα $i = 1, 2, ..., R$, αν και μόνον αν $\varphi_{MH} = 1$ (ή

$\psi_{MH}=1$). Τα μέτρα αυτά χρησιμοποιούνται για να μετρήσουν πόσο οι περιθώριες κατανομές διαφέρουν ή απέχουν από αυτές με δομή (MH). Ο βαθμός απομάκρυνσης από την (MH) αυξάνεται όσο οι τιμές των μέτρων αυξάνονται. Επομένως σύμφωνα με τα αποτελέσματα, ο βαθμός απομάκρυνσης από την (MH) είναι σχεδόν 0, για όλες τις δημοσκοπήσεις.

Πεδίο Ορισμού

Τα μέτρα φ_{MH} και ψ_{MH} κυμαίνονται στο διάστημα $[0,1]$ και υπάρχει δομή (MH) σε έναν $R\times R$ πίνακα συνάφειας, αν και μόνο αν $\varphi_{MH}=0$ (ή $\psi_{MH}=0$).

Επίπεδο Δεδομένων

Τα μέτρα φ_{MH}, ψ_{MH} είναι κατάλληλα για ονοματικές μεταβλητές, καθώς παραμένουν αναλλοίωτα κάτω από τους ίδιες αναδιατάξεις γραμμών και στηλών.

Διάστημα Εμπιστοσύνης

Χρησιμοποιώντας την μέθοδο Δέλτα, οι ποσότητες $\sqrt{n}\left(\hat{\varphi}_{MH}-\varphi_{MH}\right)$ και $\sqrt{n}\left(\hat{\psi}_{MH}-\psi_{MH}\right)$, ασυμπτωτικά $(n\to\infty)$ ακολουθούν την κανονική κατανομή με μέση τιμή $\mu=0$ και διακύμανση

A1. $$\sigma^2\left(\varphi_{MH}\right)=\sum_{i=1}^{R}\sum_{j=1}^{R}\pi_{ij}\omega_{ij}^2-\varphi_{MH}^2$$

όπου $\omega_{ij}=1+\dfrac{1}{2\log 2}\left(\log\pi_{1(i)}+\log\pi_{2(j)}\right)$

B2. $$\sigma^2\left(\psi_{MH}\right)=\sum_{i=1}^{R}\sum_{j=1}^{R}\pi_{ij}\gamma_{ij}^2-\psi_{MH}^2$$

όπου $\gamma_{ij}=\dfrac{1}{2}\left(\dfrac{\left(\pi_{i.}-\pi_{.i}\right)\left(\pi_{i.}+3\pi_{.i}\right)}{\left(\pi_{i.}-\pi_{.i}\right)^2}-\dfrac{\left(\pi_{j.}-\pi_{.j}\right)\left(\pi_{j.}+3\pi_{.j}\right)}{\left(\pi_{j.}-\pi_{.j}\right)^2}\right)$.

Για περισσότερες λεπτομέρειες αναφορικά με την μέθοδο Δέλτα [βλέπε *Bishop et al*. (1975) και *Agresti* (1984)].

Σημειώνουμε, ότι οι εκτιμήσεις των μέτρων φ_{MH} και ψ_{MH} συμβολίζονται με $\hat{\varphi}_{MH}$, $\hat{\psi}_{MH}$ και υπολογίζονται αντικαθιστώντας τις πληθυσμιακές πιθανότητες $\{\pi_{ij}\}$ με τις παρατηρούμενες πιθανότητες του δείγματος $\{p_{ij}\}$, όπου $p_{ij}=\frac{n_{ij}}{n}$ και $n=\sum_{i=1}^{R}\sum_{j=1}^{R}n_{ij}$.

Έστω $\hat{\sigma}^2(\hat{\varphi}_{MH})$ συμβολίζει την εκτίμηση της διακύμανσης $\sigma^2(\varphi_{MH})$ από τις παρατηρήσεις του δείγματος, τότε η ποσότητα $\hat{\sigma}(\hat{\varphi}_{MH})/\sqrt{n}$ είναι η κατά προσέγγιση εκτιμήτρια του τυπικού σφάλματος και το κατά προσέγγιση $100\%\times(1-a)$ διάστημα εμπιστοσύνης για την εκτιμήτρια του μέτρου φ_{MH} είναι

$$\hat{\varphi}_{MH}\, z_{a/2} \pm \hat{\sigma}(\hat{\varphi}_{MH})/\sqrt{n} \tag{6.13}$$

όπου $z_{a/2}$ το ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής.

Κατά τον ίδιο τρόπο υπολογίζουμε και το διάστημα εμπιστοσύνης για την εκτιμήτρια του μέτρου ψ_{MH} .

6.5.5.2 Μέτρα που βασίζονται στις δεσμευμένες περιθώριες κατανομές πιθανότητας

Τα δυο μέτρα φ_{MH}^{C} και ψ_{MH}^{C} χρησιμοποιούνται όταν το μοντέλο περιθώριας ομοιογένειας εκφράζεται ως

$$\pi_{i.}^{C}=\pi_{.i}^{C} \tag{6.14}$$

δηλαδή $P(X_1=i\,|\,X_1\neq X_2)=P(X_2=i\,|\,X_1\neq X_2)$,

όπου $\pi_{i.}^{C}=\frac{\pi_{i.}-\pi_{ii}}{\delta}$, $\pi_{.i}^{C}=\frac{\pi_{.i}-\pi_{ii}}{\delta}$, $\pi_{i}^{*C}=\frac{\pi_{i.}^{C}+\pi_{.i}^{C}}{2}$ και $\delta=\sum\sum_{i\neq j}\pi_{ij}$.

Αυτό σημαίνει ότι η δεσμευμένη περιθώρια κατανομή της γραμμής είναι πανομοιότυπη με την δεσμευμένη περιθώρια κατανομή της στήλης, δοθέντος ότι μια παρατήρηση θα βρίσκεται σε ένα από τα μη διαγώνια κελιά του πίνακα. Αντικαθιστώντας τις πιθανότητες $\{\pi_{i.}\}$, $\{\pi_{.i}\}$, και $\{\pi_{i}^{*}\}$ στους ισοδύναμους A, B, ορισμούς της παραγράφου 6.5.1, με τις δεσμευμένες πιθανότητες

$\left\{\pi_{i.}^{C}\right\}$, $\left\{\pi_{.i}^{C}\right\}$ και $\left\{\pi_{i}^{*C}\right\}$ αντίστοιχα, προκύπτουν ανάλογα μέτρα συμμετρίας, που αναπαριστούν τον βαθμό απομάκρυνσης των δεσμευμένων περιθώριων κατανομών από την (MH). Έστω ότι $\left\{\pi_{i.}^{C}+\pi_{.i}^{C}>0\right\}$, τότε έχουμε

Α. Μέτρο φ_{MH}^{C} τύπου πληροφορίας *Kullback-Leibler* και ψ_{MH}^{C} τύπου απόκλισης *Pearson*

Α1. $$\varphi_{MH}^{C}=\frac{1}{2\log 2}\left[I\left(\left\{\pi_{i.}^{C}\right\},\left\{\pi_{i}^{*C}\right\}\right)+I\left(\left\{\pi_{.i}^{C}\right\},\left\{\pi_{i}^{*C}\right\}\right)\right]$$

Α2. $$\psi_{MH}^{C}=\frac{1}{2}\left[D\left(\left\{\pi_{i.}^{C}\right\},\left\{\pi_{i}^{*C}\right\}\right)+D\left(\left\{\pi_{.i}^{C}\right\},\left\{\pi_{i}^{*C}\right\}\right)\right]$$

Β. Μέτρο φ_{MH}^{C} τύπου εντροπίας του *Shannon* και ψ_{MH}^{C} τύπου *Gini Concentration*

Τα παραπάνω μέτρα μπορούν ισοδύναμα να εκφρασθούν ως

Β1. $$\varphi_{MH}^{C}=1-\frac{1}{\log 2}\sum_{i=1}^{R}\pi_{i}^{*C}H_{i}\left(\left\{\pi_{k(i)}^{C}\right\}\right)$$

Β2. $$\psi_{MH}^{C}=1-2\sum_{i=1}^{R}\pi_{i}^{*C}C_{i}\left(\left\{\pi_{k(i)}^{C}\right\}\right)$$

όπου $\pi_{1(i)}^{C}=\frac{\pi_{i.}^{C}}{\pi_{i.}^{C}+\pi_{.i}^{C}}$, $\pi_{2(i)}^{C}=\frac{\pi_{.i}^{C}}{\pi_{i.}^{C}+\pi_{.i}^{C}}$, $H_{i}\left(\left\{\pi_{k(i)}\right\}\right)=-\sum_{k=1}^{2}\pi_{k(i)}\log\pi_{k(i)}$ είναι η εντροπία *Shannon* και $C_{i}\left(\left\{\pi_{k(i)}\right\}\right)=1-\sum_{k=1}^{2}\pi_{k(i)}^{2}$ είναι ο δείκτης συγκέντρωσης *Gini* (*Gini Concentration*).

Για περισσότερες λεπτομέρειες αναφορικά με τον δείκτη *Gini Concentration* [βλέπε *Haberman,* (1982)].

Γ. Μέτρο φ_{MH}^{C} σταθμισμένης πληροφορίας *Kullback-Leibler* και ψ_{MH}^{C} σταθμισμένης απόκλισης *Pearson*

Τέλος τα παραπάνω μέτρα μπορούν να δοθούν και στην ισοδύναμη μορφή

Γ1. $$\varphi_{MH}^{C} = \frac{1}{\log 2}\sum_{i=1}^{R} \pi_i^{*C} I_i\left(\left\{\pi_{k(i)}^{C}\right\};\left\{\frac{1}{2}\right\}\right)$$

Γ2. $$\psi_{MH}^{C} = \sum_{i=1}^{R} \pi_i^{*C} D_i\left(\left\{\pi_{k(i)}^{C}\right\};\left\{\frac{1}{2}\right\}\right)$$

Σημειώνουμε ότι

$$\pi_{1(i)}^{C} = \frac{\pi_{i.}^{C}}{\pi_{i.}^{C} + \pi_{.i}^{C}}\ \left(\Pr\left(X_1 = i \mid X_1 = i, X_2 = i\right)\right) \text{ και } \pi_{2(i)}^{C} = \frac{\pi_{.i}^{C}}{\pi_{i.}^{C} + \pi_{.i}^{C}}\ \left(\Pr\left(X_2 = i \mid X_1 = i, X_2 = i\right)\right)$$

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 3×3 πίνακα συνάφειας (*σελ. 168, Παράδειγμα 5α και 5β*), έχουμε τα κάτωθι αποτελέσματα. Σημειώνουμε ότι οι 3 διαφορετικοί τύποι υπολογισμού Α1, Β1, Γ1, του μέτρου φ_{MH}^{C} είναι ισοδύναμοι. Ομοίως, οι τύποι Α2, Β2, Γ2, για το μέτρο ψ_{MH}^{C}.

Παράδειγμα 5 **Παράδειγμα 5α *(ΑΥΓ vs ΟΚΤ 1971)*** ***Παράδειγμα 5β (ΟΚΤ '71 vs ΔΕΚ '73)***

A1. Kullback-Leibler	φ_{MH}		0,025	0,154
A2. Pearson discrepancy	ψ_{MH}		0,034	0,202
B1. Shannon entropy	φ_{MH}		0,025	0,154
B2. Gini Concentration	ψ_{MH}		0,034	0,202
Γ1. Weighted Kullback-Leibler	φ_{MH}		0,025	0,154
Γ2. Weighted Pearson discrepancy	ψ_{MH}		0,034	0,202

Note: The results are based on conditional probabilities

Ερμηνεία

Αναλύοντας τα αποτελέσματα παρατηρούμε ότι ο βαθμός απομάκρυνσης από την (MH) είναι μεγαλύτερος στην δεύτερη δημοσκόπηση (ΟΚΤ ’71 vs ΔΕΚ ’73). Δηλαδή η περιθώρια κατανομή του ΟΚΤ’ 71 είναι διαφορετική από αυτή του ΔΕΚ ’73, ενώ είναι ίδια όταν συγκρίνουμε τον ΑΥΓ ’71 vs ΟΚΤ ’71.

Πεδίο Ορισμού

Τα μέτρα φ_{MH}^{C} και ψ_{MH}^{C} κυμαίνονται στο διάστημα $[0,1]$ και υπάρχει δομή (MH) σε έναν $R \times R$ πίνακα συνάφειας, αν και μόνο αν $\varphi_{MH}^{C} = 0$ (ή $\psi_{MH}^{C} = 0$).

Επίπεδο Δεδομένων

Τα μέτρα φ_{MH}^{C}, ψ_{MH}^{C} είναι κατάλληλα για ονοματικές μεταβλητές, καθώς παραμένουν αναλλοίωτα κάτω από τους ίδιες αναδιατάξεις γραμμών και στηλών.

Διάστημα Εμπιστοσύνης

Χρησιμοποιώντας την μέθοδο Δέλτα, οι ποσότητες $\sqrt{n}\left(\hat{\varphi}_{MH}^{C} - \varphi_{MH}^{C}\right)$ και $\sqrt{n}\left(\hat{\psi}_{MH}^{C} - \psi_{MH}^{C}\right)$ ασυμπτωτικά $(n \to \infty)$ ακολουθούν την κανονική κατανομή με μέση τιμή $\mu = 0$ και διακύμανση

A1.
$$\sigma^2\left(\varphi_{MH}^{C}\right) = \frac{1}{\delta^2}\left(\sum_{i=1}^{R}\sum_{\substack{j=1 \\ i \neq j}}^{R} \pi_{ij}\Omega_{ij}^2 - \delta\left(\varphi_{MH}^{C}\right)^2\right)$$

όπου $\Omega_{ij} = 1 + \frac{1}{2\log 2}\left(\log \pi_{1(i)}^{C} + \log \pi_{2(j)}^{C}\right)$

B2.
$$\sigma^2\left(\psi_{MH}^{C}\right) = \frac{1}{\delta^2}\left(\sum_{i=1}^{R}\sum_{\substack{j=1 \\ i \neq j}}^{R} \pi_{ij}\Gamma_{ij}^2 - \delta\left(\psi_{MH}^{C}\right)^2\right)$$

όπου $\Gamma_{ij} = \frac{1}{2}\left(\frac{\left(\pi_{i.}^{C} - \pi_{.i}^{C}\right)\left(\pi_{i.}^{C} + 3\pi_{.i}^{C}\right)}{\left(\pi_{i.}^{C} - \pi_{.i}^{C}\right)^2} - \frac{\left(\pi_{j.}^{C} - \pi_{.j}^{C}\right)\left(3\pi_{j.}^{C} - \pi_{.j}^{C}\right)}{\left(\pi_{j.}^{C} - \pi_{.j}^{C}\right)^2}\right)$

Για περισσότερες λεπτομέρειες αναφορικά με την μέθοδο Δέλτα [βλέπε *Bishop et al*. (1975) και *Agresti* (1984)].

Σχόλια

Όλα τα παραπάνω μέτρα που περιγράψαμε, κυμαίνονται στο διάστημα $[0,1]$ ανεξάρτητα από την διάσταση R του πίνακα και το μέγεθος του δείγματος. Επομένως, είναι κατάλληλα για την σύγκριση μεταξύ διαφόρων πινάκων. Τα μέτρα φ_{MH} και ψ_{MH} (*Παράγραφος 6.5.5.1*) μετρούν τον βαθμό που οι μη-δεσμευμένες περιθώριες κατανομές, διαφέρουν από αυτές με δομή (MH), λαμβάνοντας υπόψη τις διαγώνιες πιθανότητες. Για παράδειγμα, σε τι βαθμό η περιθώρια

κατανομή της πρώτης δημοσκόπησης διαφέρει από αυτή της δεύτερης (ή αλλιώς, σε τι βαθμό η πιθανότητα ένα υποκείμενο να απαντήσει «Ναι», στην πρώτη δημοσκόπηση, διαφέρει από το να απαντήσει επίσης «Ναι», στην δεύτερη). Τα μέτρα φ_{MH}^{C} , ψ_{MH}^{C} (*Παράγραφος 6.5.5.2*) μετρούν τον βαθμό που οι δεσμευμένες περιθώριες κατανομές, διαφέρουν από τις αντίστοιχες δεσμευμένες με δομή (MH), χωρίς να λαμβάνουν υπόψη τις διαγώνιες πιθανότητες. Για παράδειγμα, σε τι βαθμό η πιθανότητα ένα υποκείμενο να απαντήσει «Ναι», στην πρώτη δημοσκόπηση (δοθέντος ότι δεν απάντησε «Ναι» στην δεύτερη δημοσκόπηση), διαφέρει από το να απαντήσει «Ναι», στην δεύτερη δημοσκόπηση (δοθέντος ότι δεν απάντησε «Ναι» στην πρώτη). Από τα αποτελέσματα παρατηρούμε, ότι τα μέτρα που βασίζονται στις δεσμευμένες πιθανότητες, δείχνουν εντονότερο βαθμό απομάκρυνσης από την (MH). Ο αναγνώστης ίσως ενδιαφέρεται ποιο από τα μέτρα είναι καταλληλότερο για ένα πίνακα, δηλαδή για το αν θα πρέπει ή όχι να λάβει υπόψη του τις διαγώνιες πιθανότητες. Κάτι τέτοιο εξαρτάται, από το ερώτημα που η έρευνα καλείται να απαντήσει, βάσει των παραπάνω εξηγήσεων. Επίσης, είναι δύσκολο να απαντήσουμε ποιο από τα μέτρα φ ή ψ είναι καταλληλότερο για ένα πίνακα. Όπως ο *Tomizawa* (1995) προτείνει, ο βαθμός απομάκρυνσης από την (MH), θα πρέπει να περιγράφεται σε όρους και των δυο τιμών. Επίσης, οι εκτιμήσεις του βαθμού απομάκρυνσης από την (MH), θα πρέπει να εξετάζονται σε όρους ενός κατά προσέγγιση διαστήματος εμπιστοσύνης για κάθε μέτρο, παρά βασιζόμενοι μόνο στις τιμές των μέτρων. Τα μέτρα φ_{MH} και φ_{MH}^{C} (ή ψ_{MH} και ψ_{MH}^{C}) δεν είναι ισοδύναμα. Τέλος, στην περίπτωση ενός 2×2 πίνακα τα μέτρα φ_{MH} και φ_{MH}^{C} είναι ισοδύναμα με το μέτρο φ_{S} (απομάκρυνσης από την συμμετρία), που εξετάστηκε στην παράγραφο 6.5.1.

6.5.6 Γενίκευση των μέτρων απομάκρυνσης από την (MH) για ονοματικές μεταβλητές

Σε συνέχεια των ονοματικών μέτρων φ_{MH} (ψ_{MH}) και φ_{MH}^{C} (ψ_{MH}^{C}), της προηγούμενης παραγράφου 6.5.5, οι *Tomizawa & Makii* (2001), πρότειναν μια γενίκευση αυτών. Το προτεινόμενο μέτρο εκφράζεται μέσω του μέσου όρου της απόκλισης τύπου *power divergence* των *Cressie & Read* (1984) και σε ειδικές περιπτώσεις του, εμπεριέχει τα ονοματικά μέτρα φ_{MH}

$\left(\psi_{MH}\right)$. Επίσης, αντικαθιστώντας τις πιθανότητες $\left\{\pi_{i.}\right\}$, $\left\{\pi_{.i}\right\}$, με τις δεσμευμένες πιθανότητες $\left\{\pi_{i.}^{C}\right\}$, $\left\{\pi_{.i}^{C}\right\}$, τότε σε ειδικές περιπτώσεις του εμπεριέχει τα ονοματικά μέτρα φ_{MH}^{C} $\left(\psi_{MH}^{C}\right)$. Το μέτρο είναι χρήσιμο για την σύγκριση του βαθμού της απομάκρυνσης από την $\left(MH\right)$ μεταξύ αρκετών πινάκων.

6.5.6.1 Γενικευμένο μέτρο $\Phi_{MH}^{(\lambda)}$

Υποθέτοντας ότι $\left\{\pi_{i.}+\pi_{.i}\neq 0\right\}$, για έναν RxR πίνακα με ονοματικές κατηγορίες, το γενικευμένο μέτρο $\Phi_{MH}^{(\lambda)}$ που μετρά τον βαθμό απομάκρυνσης από την $\left(MH\right)$, δίνεται ως ακολούθως

Α1. $$\Phi_{MH}^{(\lambda)}=\frac{\lambda(\lambda+1)}{2^{\lambda}-1}\sum_{i=1}^{R}\pi_{i}^{*}I_{i}^{(\lambda)}\left(\left\{\pi_{1(i)},\pi_{2(i)}\right\};\left\{\frac{1}{2},\frac{1}{2}\right\}\right),\ \text{για}\ \lambda\geq -1$$

όπου $\pi_{i}^{*}=\frac{\pi_{i.}+\pi_{.i}}{2}$, $\pi_{1(i)}=\frac{\pi_{i.}}{\pi_{i.}+\pi_{.i}}$, $\pi_{2(i)}=\frac{\pi_{.i}}{\pi_{i.}+\pi_{.i}}$

και $I_{i}^{(\lambda)}\left(\left\{a,b\right\};\left\{\frac{1}{2},\frac{1}{2}\right\}\right)=\frac{1}{\lambda(\lambda+1)}\left[a\left\{\left(\frac{a}{1/2}\right)^{\lambda}-1\right\}+b\left\{\left(\frac{b}{1/2}\right)^{\lambda}-1\right\}\right]$, $a+b=1$

Για $\lambda=0$ παίρνουμε $\lim_{\lambda\to 0}\Phi_{MH}^{(\lambda)}$ και επομένως

$$I_{i}^{(0)}\left(\left\{a,b\right\};\left\{\frac{1}{2},\frac{1}{2}\right\}\right)=a\log\left(\frac{a}{1/2}\right)+b\log\left(\frac{b}{1/2}\right)$$

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 3×3 πίνακα συνάφειας (*σελ. 168, Παράδειγμα 5α και 5β*), έχουμε τα κάτωθι αποτελέσματα.

	$\Phi_{MH}^{(\lambda)}$ **Cressie and Read**	
λ	Παράδειγμα 5α *(ΑΥΓ vs ΟΚΤ 1971)*	*Παράδειγμα 5β (ΟΚΤ '71 vs ΔΕΚ '73)*
-0,8	0,001	0,008
-0,6	0,001	0,012
-0,4	0,002	0,021
0	0,003	0,030
1	0,004	0,040
1,4	0,004	0,041
1,6	0,004	0,041
2	0,004	0,040

Ερμηνεία

Η ερμηνεία των αποτελεσμάτων είναι παρόμοια με τα μέτρα της παραγράφου 6.5.5. Παρατηρούμε ότι στην περίπτωση που $\lambda = 0$ και $\lambda = 1$, τα μέτρα $\Phi_{MH}^{(\lambda)}$, $\Phi_{MH}^{C\ (\lambda)}$, είναι στην ουσία τα μέτρα απομάκρυνσης από την (MH), που εξετάστηκαν από τον *Tomizawa* (1995) και περιγράφονται στην προηγούμενη παράγραφο 6.5.5.

6.5.6.2 Γενικευμένο μέτρο $\Phi_{MH}^{C\ (\lambda)}$

Επίσης, υποθέτοντας ότι $\left\{\pi_{i.}^{C} + \pi_{.i}^{C} \neq 0\right\}$, το γενικευμένο μέτρο $\Phi_{MH}^{C\ (\lambda)}$ απομάκρυνσης από την MH, αντικαθιστώντας τις πιθανότητες $\{\pi_{i.}, \pi_{.i}\}$, με τις αντίστοιχες δεσμευμένες $\left\{\pi_{i.}^{C}, \pi_{.i}^{C}\right\}$ δίνεται ως ακολούθως

Α1. $$\Phi_{MH}^{C\ (\lambda)} = \frac{\lambda(\lambda+1)}{2^{\lambda}-1}\sum_{i=1}^{R}\pi_i^{C*} I_i^{(\lambda)}\left(\left\{\pi_{1(i)}^{C}, \pi_{2(i)}^{C}\right\}; \left\{\frac{1}{2}, \frac{1}{2}\right\}\right), \text{ για } \lambda \geq -1$$

όπου $\pi_{i.}^{C} = \frac{\pi_{i.} - \pi_{ii}}{\delta}$, $\pi_{.i}^{C} = \frac{\pi_{.i} - \pi_{ii}}{\delta}$ οι δεσμευμένες πιθανότητες, με $\pi_{1(i)}^{C} = \frac{\pi_{i.}^{C}}{\pi_{i.}^{C} + \pi_{.i}^{C}}$, $\pi_{2(i)}^{C} = \frac{\pi_{.i}^{C}}{\pi_{i.}^{C} + \pi_{.i}^{C}}$, $\pi_i^{C*} = \frac{\pi_{i.}^{C} + \pi_{.i}^{C}}{2}$ και $\delta = \sum\sum_{i \neq j}\pi_{ij}$. Για $\lambda = 0$ παίρνουμε $\lim_{\lambda \to 0}\Phi_{MH}^{C\ (\lambda)}$.

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 3×3 πίνακα συνάφειας (*σελ. 168, Παράδειγμα 5α και 5β*), έχουμε τα κάτωθι αποτελέσματα.

	$\Phi^{c}{}_{MH}{}^{(\lambda)}$ **Cressie and Read**	
λ	Παράδειγμα 5α *(ΑΥΓ vs ΟΚΤ 1971)*	*Παράδειγμα 5β (ΟΚΤ '71 vs ΔΕΚ '73)*
-0,8	0,007	0,044
-0,6	0,010	0,067
-0,4	0,017	0,110
0	0,025	0,154
1	0,034	0,202
1,4	0,035	0,205
1,6	0,035	0,205
2	0,034	0,202

Ερμηνεία

Παρόμοια με τα αποτελέσματα της παραγράφου 6.5.5, παρατηρούμε ότι τα μέτρα που βασίζονται στις δεσμευμένες πιθανότητες, δείχνουν εντονότερο βαθμό απομάκρυνσης από την (MH) και ο βαθμός αυτός είναι μεγαλύτερος, συγκρίνοντας την 2^{η} και 3^{η} δημοσκόπηση, σε σχέση με την 1^{η} και 2^{η}.

6.5.7 Μέτρα απομάκρυνσης από την Ομοιογένεια Περιθωρίου (MH) για διατακτικές μεταβλητές

Στην περίπτωση διατακτικών μεταβλητών, όπου μπορούμε να αξιοποιήσουμε την πληροφορίας της διάταξης, ενδιαφερόμαστε για ένα μέτρο, το οποίο εξαρτάται από την διάταξη των κατηγοριών των μεταβλητών. Το προτεινόμενο μέτρο [βλέπε *Tomizawa et al*. (2003)] είναι συνάρτηση των αθροιστικών πιθανοτήτων $\{G_{1(i)}\}$ και $\{G_{2(i)}\}$ (αντί των $\{\pi_{i.}\}$ και $\{\pi_{.i}\}$ ή των $\{\pi^{C}_{i.}\}$ και $\{\pi^{C}_{.i}\}$), καθώς οι αθροιστικές πιθανότητες $\{G_{1(i)}\}$ και $\{G_{2(i)}\}$ ορίζονται μόνο για διατακτικές μεταβλητές.

Έστω, $G_{1(i)}=\sum_{s=1}^{i}\sum_{t=i+1}^{R}\pi_{st}\ \left[=\Pr\left(X\le i, Y\ge i+1\right)\right]$ και $G_{2(i)}=\sum_{s=i+1}^{R}\sum_{t=1}^{i}\pi_{st}\ \left[=\Pr\left(X\ge i+1, Y\le i\right)\right]$

για $i=1,2,...,R-1$. Εξετάζοντας την διαφορά μεταξύ των αθροιστικών (*cumalative*) περιθώριων πιθανοτήτων, $\left\{F_i^X - F_i^Y\right\}$, για $i=1,2,...,R-1$, όπου $F_i^X = P\left(X \le i\right)$ και $F_i^Y = P\left(Y \le i\right)$, το μοντέλο ομοιογένειας περιθωρίου $\left(MH\right)$ μπορεί να εκφραστεί ως $G_{1(i)} = G_{2(i)}$.

Έστω $\Delta = \sum_{i=1}^{R-1}\left(G_{1(i)} + G_{2(i)}\right)$ και $G_{1(i)}^* = \frac{G_{1(i)}}{\Delta}$, $G_{2(i)}^* = \frac{G_{2(i)}}{\Delta}$, $Q_i^* = \frac{1}{2}\left(G_{1(i)}^* + G_{2(i)}^*\right)$, για $i=1,2,...,R-1$. Σημειώνουμε ότι όταν ισχύει το μοντέλο της $\left(MH\right)$ τότε $\left\{G_{1(i)}^* = G_{2(i)}^* = Q_i^*\right\}$ καθώς και ότι $\sum_{i=1}^{R-1}\left(G_{1(i)}^* + G_{2(i)}^*\right) = 1$ και $\sum_{i=1}^{R-1}\left(2Q_i^*\right) = 1$.

6.5.7.1 Γενικευμένο μέτρο $\Gamma_{MH}^{(\lambda)}$ που βασίζεται στις αδέσμευτες περιθώριες κατανομές πιθανότητας

Υποθέτοντας ότι $\left\{G_{1(i)} + G_{2(i)} \ne 0\right\}$ για $i=1,2,...,R-1$ θα εξετάσουμε ένα γενικευμένο μέτρο που ορίζεται ως

Α1. $$\Gamma_{MH}^{(\lambda)} = \frac{\lambda(\lambda+1)}{2^\lambda - 1} I^{(\lambda)}\left(\left\{G_{1(i)}^*, G_{2(i)}^*\right\}; \left\{Q_i^*, Q_i^*\right\}\right) \text{ για } \lambda > -1$$

όπου $I^{(\lambda)}\left(\left\{G_{1(i)}^*, G_{2(i)}^*\right\}; \left\{Q_i^*, Q_i^*\right\}\right) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{R-1}\left[G_{1(i)}^*\left\{\left(\frac{G_{1(i)}^*}{Q_i^*}\right)^\lambda - 1\right\} + G_{2(i)}^*\left\{\left(\frac{G_{2(i)}^*}{Q_i^*}\right)^\lambda - 1\right\}\right]$.

Για την τιμή $\lambda = 0$, έχουμε το $\lim_{\lambda \to 0}\Gamma_{MH}^{(\lambda)}$ και κατά συνέπεια

$$\Gamma_{MH}^{(0)} = \frac{1}{\log 2} I^{(0)}\left(\left\{G_{1(i)}^*, G_{2(i)}^*\right\}; \left\{Q_i^*, Q_i^*\right\}\right), \text{ όπου } I^{(0)} = \sum_{i=1}^{R-1}\left[G_{1(i)}^* \log\frac{G_{1(i)}^*}{Q_i^*} + G_{2(i)}^* \log\frac{G_{2(i)}^*}{Q_i^*}\right].$$

Σημειώνουμε ότι το $I^{(\lambda)}\left(\left\{G_{1(i)}^*, G_{2(i)}^*\right\}; \left\{Q_i^*, Q_i^*\right\}\right)$ είναι η απόκλιση τύπου *power-divergence Cressie & Read* (1984), ανάμεσα στις ποσότητες $\left\{G_{1(i)}^*, G_{2(i)}^*\right\}$ και $\left\{Q_i^*, Q_i^*\right\}$ για $i=1,2,...,R-1$. Ειδικότερα η ποσότητα $\left\{I^{(0)}\right\}$, είναι η πληροφορία *Kullback-Leibler*, ενώ η ποσότητα $\left\{I^{(1)}\right\}$

είναι η απόκλιση $\chi^2 - Pearson$ μεταξύ τους. Όταν $I^{(\lambda)}\left(\left\{G^*_{1(i)}, G^*_{2(i)}\right\}; \left\{Q^*_i, Q^*_i\right\}\right) = 0$ τότε ισχύει το μοντέλο της (MH).

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 8×8 πίνακα συνάφειας (*σελ.169, Παράδειγμα 6*), έχουμε τα κάτωθι αποτελέσματα.

	$\Gamma^{(\lambda)}_{MH}$ **Cressie & Read (*power divergence*)**		
λ	***1955***	***1965***	***1975***
-0,8	0,047	0,104	0,103
-0,6	0,086	0,182	0,180
-0,4	0,118	0,241	0,237
0	0,165	0,319	0,312
1	0,216	0,392	0,382
1,4	0,221	0,397	0,387
1,6	0,220	0,397	0,386
2	0,216	0,392	0,382
2	0,216	0,392	0,382

Ερμηνεία

Το μοντέλο ομοιογένειας περιθωρίου (MH) για διατακτικές μεταβλητές, όπως είδαμε, εκφράζεται ως $G_{1(i)} = G_{2(i)}$. Αυτό σημαίνει ότι η αθροιστική πιθανότητα μια παρατήρηση να βρίσκεται στην $i-$ κατηγορία της μεταβλητής γραμμή ή χαμηλότερα και στην $i+1$ κατηγορία της μεταβλητής στήλης ή υψηλότερα, ισούται με την αθροιστική πιθανότητα μια παρατήρηση να βρίσκεται στην $i-$ κατηγορία της μεταβλητής στήλης ή χαμηλότερα και στην $i+1$ κατηγορία της μεταβλητής γραμμής ή υψηλότερα. Υπό αυτή την έννοια, τα αποτελέσματα του μέτρου, για διάφορες τιμές του λ, εκφράζουν τον βαθμό απομάκρυνσης από την (MH), ως προς την μέγιστη απομάκρυνση, της οποίας ο βαθμός ισούται με 1. Επομένως, τα αποτελέσματα $\Gamma^{(1)}_{MH(1955)} = 0.216$, $\Gamma^{(1)}_{MH(1965)} = 0.392$ και $\Gamma^{(1)}_{MH(1975)} = 0.382$, εκφράζουν ότι για το 1955, τα δεδομένα πετυχαίνουν το 21.6% του βαθμού της μέγιστης απομάκρυνσης, για το 1965 το 39.2%

και για το 1975 το 38.2%. Επομένως, το επάγγελμα του υιού διαφοροποιείται εντονότερα κατά τα έτη 1965, 1975.

Πεδίο Ορισμού

Το γενικευμένο μέτρο $\Gamma_{MH}^{(\lambda)}$ κυμαίνεται στο διάστημα $[0,1]$ και υπάρχει δομή (MH) σε έναν $R \times R$ πίνακα συνάφειας, αν και μόνο αν $\Gamma_{MH}^{(\lambda)} = 0$.

Επίπεδο Δεδομένων

Το γενικευμένο μέτρο $\Gamma_{MH}^{(\lambda)}$ εξαρτάται από την διάταξη των κατηγοριών των μεταβλητών και δεν παραμένει αναλλοίωτο κάτω από τους ίδιες αναδιατάξεις γραμμών και στηλών, εκτός από την αντίστροφη διάταξη. Σημειώνουμε ότι οι αθροιστικές πιθανότητες $\{G_{1(i)}\}$ και $\{G_{2(i)}\}$ δεν ορίζονται για ονοματικές μεταβλητές.

6.5.7.2 Γενικευμένο μέτρο $\Gamma_{MH}^{C\ (\lambda)}$ που βασίζεται στις δεσμευμένες περιθώριες κατανομές πιθανότητας

Έστω $G_{1(i)}^{C} = \dfrac{G_{1(i)}}{G_{1(i)} + G_{2(i)}}$ και $G_{2(i)}^{C} = \dfrac{G_{2(i)}}{G_{1(i)} + G_{2(i)}}$ για $i = 1, 2, ..., R-1$. Οι ποσότητες αυτές είναι οι δεσμευμένες αθροιστικές πιθανότητες. Για παράδειγμα,

$G_{1(i)}^{C} = \Pr(X \le i \mid X \le i, Y \ge i+1)$ (ή $G_{1(i)}^{C} = \Pr(X \le i \mid Y \le i, X \ge i+1)$)

$G_{2(i)}^{C} = \Pr(Y \le i \mid X \le i, Y \ge i+1)$ (ή $G_{2(i)}^{C} = \Pr(Y \le i \mid Y \le i, X \ge i+1)$)

Επιπλέον, $\{G_{1(i)}^{C} + G_{2(i)}^{C} = 1\}$ και το μοντέλο της (MH) μπορεί να εκφραστεί ως $G_{1(i)}^{C} = G_{2(i)}^{C} = \left(\dfrac{1}{2}\right)$ για $i = 1, 2, ..., R-1$. Δηλαδή, η δεσμευμένη πιθανότητα ότι η μεταβλητή X βρίσκεται στην i κατηγορία ή μικρότερη της i, δοθέντος ότι μια από τις μεταβλητές X, Y είναι στην i κατηγορία ή μικρότερη της i και η άλλη μεταβλητή είναι στην $i+1$ κατηγορία ή μεγαλύτερη της $i+1$, ισούται με την δεσμευμένη πιθανότητα ότι η μεταβλητή Y είναι στην i κατηγορία ή μικρότερη της i, δοθείσης της ίδιας δέσμευσης.

Χρησιμοποιώντας τις δεσμευμένες πιθανότητες $G^C_{1(i)}$ και $G^C_{2(i)}$, το γενικευμένο μέτρο $\Gamma^{(\lambda)}_{MH}$ μπορεί ισοδύναμα να εκφρασθεί ως

A1. $$\Gamma^{C}_{MH}{}^{(\lambda)} = \frac{\lambda(\lambda+1)}{2^{\lambda}-1}\sum_{i=1}^{R-1}\left(G^{*}_{1(i)}+G^{*}_{2(i)}\right)I_i^{(\lambda)}\left(\left\{G^C_{1(i)},G^C_{2(i)}\right\};\left\{\frac{1}{2},\frac{1}{2}\right\}\right) \text{ για } \lambda>-1,$$

όπου $$I_i^{(\lambda)}\left(\left\{G^C_{1(i)},G^C_{2(i)}\right\};\left\{\frac{1}{2},\frac{1}{2}\right\}\right)=\frac{1}{\lambda(\lambda+1)}\left[G^C_{1(i)}\left\{\left(\frac{G^C_{1(i)}}{1/2}\right)^{\lambda}-1\right\}+G^C_{2(i)}\left\{\left(\frac{G^C_{2(i)}}{1/2}\right)^{\lambda}-1\right\}\right]$$

Για την τιμή $\lambda=0$ έχουμε το $\lim_{\lambda\to 0}\Gamma^{C}_{MH}{}^{(\lambda)}$ και κατά συνέπεια

$$\Gamma^{C}_{MH}{}^{(0)} = \frac{1}{\log 2}\sum_{i=1}^{R-1}\left(G^{*}_{1(i)}+G^{*}_{2(i)}\right)I_i^{(0)}\left(\left\{G^C_{1(i)},G^C_{2(i)}\right\};\left\{\frac{1}{2},\frac{1}{2}\right\}\right),$$

όπου $I_i^{(0)}=G^C_{1(i)}\log\frac{G^C_{1(i)}}{1/2}+G^C_{2(i)}\log\frac{G^C_{2(i)}}{1/2}$.

Επομένως, το $\Gamma^{C}_{MH}{}^{(\lambda)}$ αναπαριστά στην ουσία, το σταθμισμένο άθροισμα της *power-divergence* $I_i^{(\lambda)}\left(\left\{G^C_{1(i)},G^C_{2(i)}\right\};\left\{\frac{1}{2},\frac{1}{2}\right\}\right)$. Σημειώνουμε ότι η ποσότητα $I_i^{(\lambda)}\left(\left\{G^C_{1(i)},G^C_{2(i)}\right\};\left\{\frac{1}{2},\frac{1}{2}\right\}\right)$ δείχνει (εξηγεί) πόσο μακριά βρίσκονται οι δεσμευμένες κατανομές $\left\{G^C_{1(i)},G^C_{2(i)}\right\}$, από τις αντίστοιχες, που έχουν όμως μια δομή ομοιογένειας περιθωρίου (MH), για παράδειγμα από τις $\left\{\frac{1}{2},\frac{1}{2}\right\}$.

Επιπλέον το $\Gamma^{C}_{MH}{}^{(\lambda)}$ μπορεί να εκφραστεί μέσω του δείκτη *Patil & Taillie* (1982), ως

$$\Gamma^{C}_{MH}{}^{(\lambda)} = 1-\frac{\lambda 2^{\lambda}}{2^{\lambda}-1}\sum_{i=1}^{R-1}\left(G^{*}_{1(i)}+G^{*}_{2(i)}\right)\mathrm{H}_i^{(\lambda)}\left(\left\{G^C_{1(i)},G^C_{2(i)}\right\}\right) \text{ για } \lambda>-1,$$

όπου $\mathrm{H}_i^{(\lambda)}\left(\left\{G^C_{1(i)},G^C_{2(i)}\right\}\right)=\frac{1}{\lambda}\left[1-\left(G^C_{1(i)}\right)^{\lambda+1}-\left(G^C_{2(i)}\right)^{\lambda+1}\right]$.

Για την τιμή $\lambda=0$ έχουμε $\lim_{\lambda\to 0}\Gamma^{C}_{MH}{}^{(\lambda)}$ και κατά συνέπεια

$$\Gamma^{C}_{MH}{}^{(0)} = 1-\frac{1}{\log 2}\sum_{i=1}^{R-1}\left(G^{*}_{1(i)}+G^{*}_{2(i)}\right)\mathrm{H}_i^{(0)}\left(\left\{G^C_{1(i)},G^C_{2(i)}\right\}\right),$$

όπου $\mathrm{H}_i^{(0)}\left(\left\{G_{1(i)}^C, G_{2(i)}^C\right\}\right) = -G_{1(i)}^C \log G_{1(i)}^C - G_{2(i)}^C \log G_{2(i)}^C$.

Σημειώνουμε ότι η ποσότητα $\mathrm{H}_i^{(\lambda)}\left(\left\{G_{1(i)}^C, G_{2(i)}^C\right\}\right)$ είναι ο δείκτης *Patil & Taillie's diversity index* λ - τάξης, για την ποσότητα $\left\{G_{1(i)}^C, G_{2(i)}^C\right\}$, ο οποίος περιλαμβάνει την εντροπία *Shannon* (όταν $\lambda = 0$) και τον δείκτη *Gini concentration* (όταν $\lambda = 1$). Το μέτρο ${\Gamma_{MH}^C}^{(\lambda)}$ αναπαριστά στην ουσία το σταθμισμένο άθροισμα του *diversity index* $\mathrm{H}_i^{(\lambda)}\left(\left\{G_{1(i)}^C, G_{2(i)}^C\right\}\right)$. Για περισσότερες λεπτομέρειες αναφορικά με τον δείκτη *Gini Concentration* [βλέπε *Haberman*, (1982)].

Παράδειγμα

Τα αποτελέσματα είναι ισοδύναμα με αυτά της προηγούμενης παραγράφου 6.5.7.1

Πεδίο Ορισμού

Το γενικευμένο μέτρο ${\Gamma_{MH}^C}^{(\lambda)}$ κυμαίνεται στο διάστημα $[0,1]$ και υπάρχει δομή (MH) σε έναν $R \times R$ πίνακα συνάφειας, αν και μόνο αν ${\Gamma_{MH}^C}^{(\lambda)} = 0$.

Διάστημα Εμπιστοσύνης

Χρησιμοποιώντας την μέθοδο Δέλτα η ποσότητα $\left\{\sqrt{n}\left({\hat{\Gamma}_{MH}^C}^{(\lambda)} - {\Gamma_{MH}^C}^{(\lambda)}\right)\right\}$, ακολουθεί ασυμπτωματικά την κανονική κατανομή (όσο $n \to \infty$), με μέση τιμή $\mu = 0$ και διακύμανση

$$\sigma^2\left[{\Gamma_{MH}^C}^{(\lambda)}\right] = \frac{1}{\Delta^2}\sum_{s=1}^{R-1}\sum_{t=s+1}^{R}\left[p_{st}\left(w_{st}^{(\lambda)}\right)^2 + p_{ts}\left(v_{ts}^{(\lambda)}\right)^2\right],$$

όπου για $\lambda > -1,\ \lambda \neq 0$

$$w_{st}^{(\lambda)} = \frac{2^\lambda}{2^\lambda - 1}\left[\sum_{i=s}^{t-1}\left\{\left(G_{1(i)}^C\right)^\lambda + \lambda\left(\left(G_{1(i)}^C\right)^\lambda - \left(G_{2(i)}^C\right)^\lambda\right)G_{2(i)}^C\right\} - (t-s)\frac{\left(2^\lambda - 1\right)\Gamma_{MH}^{(\lambda)} + 1}{2^\lambda}\right]$$

$$v_{ts}^{(\lambda)} = \frac{2^\lambda}{2^\lambda - 1}\left[\sum_{i=s}^{t-1}\left\{\left(G_{2(i)}^C\right)^\lambda + \lambda\left(\left(G_{2(i)}^C\right)^\lambda - \left(G_{1(i)}^C\right)^\lambda\right)G_{1(i)}^C\right\} - (t-s)\frac{\left(2^\lambda - 1\right)\Gamma_{MH}^{(\lambda)} + 1}{2^\lambda}\right]$$

και για $\lambda = 0$

$$w_{st}^{(0)} = \frac{1}{\log 2}\left[\sum_{i=s}^{t-1}\log G_{1(i)}^C - (t-s)\left(\Gamma_{MH}^{(0)} - 1\right)\log 2\right]$$

$$v_{st}^{(0)} = \frac{1}{\log 2}\left[\sum_{i=s}^{t-1} \log G_{2(i)}^{C} - (t-s)\left(\Gamma_{MH}^{(0)} - 1\right)\log 2\right].$$

Για περισσότερες λεπτομέρειες αναφορικά με την μέθοδο Δέλτα [βλέπε *Bishop et al*. (1975) και *Agresti* (1984)].

Σημειώνουμε ότι η ασυμπτωματική κατανομή της ποσότητας $\left\{\sqrt{n}\left(\hat{\Gamma}_{MH}^{C\ (\lambda)} - \Gamma_{MH}^{C\ (\lambda)}\right)\right\}$, δεν μπορεί να εφαρμοστεί όταν $\Gamma_{MH}^{C\ (\lambda)} = 0$ και $\Gamma_{MH}^{C\ (\lambda)} = 1$, διότι τότε $\sigma^2\left[\Gamma_{MH}^{C\ (\lambda)}\right] = 0$.

Η εκτίμηση της διακύμανσης συμβολίζεται με $\hat{\sigma}^2\left[\hat{\Gamma}_{MH}^{C\ (\lambda)}\right]$ και υπολογίζεται από την ποσότητα $\sigma^2\left[\Gamma_{MH}^{C\ (\lambda)}\right]$, αντικαθιστώντας τις πληθυσμιακές πιθανότητες $\{\pi_{ij}\}$ με τις δειγματικές πιθανότητες $\{p_{ij}\}$. Η ποσότητα $\hat{\sigma}\left[\hat{\Gamma}_{MH}^{C\ (\lambda)}\right]\Big/\sqrt{n}$ είναι μια κατά προσέγγιση εκτίμηση του τυπικού σφάλματος της εκτιμήτριας του μέτρου $\Gamma_{MH}^{C\ (\lambda)}$ και το κατά προσέγγιση $100\% \times (1-a)$ διάστημα εμπιστοσύνης για το μέτρο $\Gamma_{MH}^{C\ (\lambda)}$ είναι

$$\hat{\Gamma}_{MH}^{C\ (\lambda)} \pm z_{a/2}\,\hat{\sigma}\left[\hat{\Gamma}_{MH}^{C\ (\lambda)}\right]\Big/\sqrt{n} \tag{6.15}$$

όπου $z_{a/2}$ το ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής.

Σχόλια

Τα μέτρα $\Gamma_{MH}^{(\lambda)}$ και $\Gamma_{MH}^{C\ (\lambda)}$ είναι ισοδύναμα. Υπάρχει δομή ομοιογένειας περιθωρίου (MH) σε έναν RxR πίνακα, για παράδειγμα, όταν $F_i^X = F_i^Y$ και κατά συνέπεια $G_{1(i)} = G_{2(i)}$ $\left(G_{1(i)}^c = G_{2(i)}^c = \frac{1}{2}\right)$, για όλα τα $i = 1, 2, ..., R-1$), αν και μόνον αν $\Gamma_{MH}^{(\lambda)} = 0$. Ο βαθμός απομάκρυνσης από την (MH) είναι ο μέγιστος, (με την έννοια ότι για $G_{1(i)}^c = 0$ (τότε $G_{2(i)}^c = 1$) ή για $G_{2(i)}^c = 0$ τότε ($G_{1(i)}^c = 1$) για όλα $i = 1, 2, ..., R-1$), αν και μόνο αν $\Gamma_{\mathrm{M}}^{(\lambda)} = 1$. Δηλαδή, η δεσμευμένη πιθανότητα ότι η μεταβλητή X είναι στην i κατηγορία ή μικρότερη της i, δοθέντος ότι μια από τις μεταβλητές X, Y είναι στην i κατηγορία ή μικρότερη της i και η άλλη μεταβλητή είναι στην $i+1$ κατηγορία ή μεγαλύτερη της $i+1$, ισούται με 0 ή 1 για όλα τα $i = 1, 2, ..., R-1$.

6.6 Ορισμοί των μοντέλων Ασυμμετρίας για πίνακες συνάφειας

Στην παράγραφο 6.4 αναφερθήκαμε στο μοντέλο συμμετρίας $(S-Symmetry)$ και στα μοντέλα άλλης δομής, όπως αυτό της ψευδοσυμμετρίας $(QS-Quasi\ Symmetry)$ και της περιθώριας ομοιογένειας $(MH-Marginal\ Homogeneity)$. Στην παράγραφο 6.5 παρουσιάσαμε τα κυριότερα μέτρα συμμετρίας για ονοματικές και διατακτικές μεταβλητές, που υιοθετούνται για να μετρήσουν τον βαθμό απομάκρυνσης από την (S), την (QS) και την (MH), όταν το μοντέλο της συμμετρίας δεν ισχύει.

Στην παράγραφο αυτή, θα εξετάσουμε κάποιους ιδιαίτερους τύπους ασυμμετρίας, για διατακτικές μεταβλητές, που περιγράφονται από το μοντέλο της δεσμευμένης συμμετρίας $(CS-Conditional\ Symmetry)$, το οποίο προτάθηκε από τον *McCullagh* (1978) ή ισοδύναμα της τριγωνικής συμμετρίας $(TS-Triangular\ Symmetry)$, το οποίο προτάθηκε από τον *Goodman* (1979) και από το μοντέλο της διαγώνιας συμμετρίας $(DS-Diagonal\ Symmetry)$ ή $(DPS-Diagonal\ Parameter\ Symmetry)$, που επίσης προτάθηκε από τον *Goodman* (1979). Επιπλέον, θα παρουσιάσουμε τα κυριότερα μέτρα ασυμμετρίας, που χρησιμοποιούνται για να μετρήσουν τον βαθμό απομάκρυνσης από τα μοντέλα (CS), (DS) και (TS). Τα μοντέλα αυτά βασίζονται στην διάταξη των μεταβλητών και επομένως τα προτεινόμενα μέτρα χρησιμοποιούνται για για διατακτικές μεταβλητές.

Αξίζει να σημειώσουμε, ότι ο *Read* (1977) αναφέρθηκε στο μοντέλο (CS), ως το μοντέλο αναλογικής συμμετρίας και το οποίο είναι ισοδύναμο με ένα λογαριθμο γραμμικό μοντέλο που εξετάζεται από τον *Bishop et al.* (1975. Επίσης, ο *McCullagh* (1978), εξέτασε το γενικευμένο μοντέλο παλίνδρομης συμμετρίας $(GPS-Generilized\ Palindromic\ Symmetry)$, το οποίο εμπεριέχει το μοντέλο (CS), ως ειδική περίπτωση. Ο *Tomizawa* (1989b) εξέτασε το τροποποιημένο μοντέλο περιθώριας ομοιογένειας $(MMH-Modified\ Marginal\ Homogeneity)$, δίνοντας την διάσπαση του μοντέλου (CS), στα μοντέλα (GPS) και (MMH).

6.6.1 Μοντέλο Δεσμευμένης Συμμετρίας $(CS - Conditional\ Symmetry)$

Το μοντέλο (CS) προτάθηκε από τον *Goodman* (1979) και εξετάζεται όταν οι μεταβλητές ταξινόμησης είναι διατακτικές και το μοντέλο (S) δεν ισχύει. Είναι ιδιαίτερα απλό, απλούστερο του μοντέλου (S) και χαίρει εύκολης και άμεσης ερμηνείας.

Ορίζεται ως

$$\mathrm{P}(X = i, Y = j \mid X > Y) = \mathrm{P}(X = j, Y = i \mid X < Y), \text{ για } i > j \qquad (CS_1)$$

και μπορεί να εκφρασθεί μέσω της σχέσης

$$\pi_{ij} = \delta\pi_{ji} \text{ για } i, j = 1, \ldots, R,\ i > j \text{ και } \delta \in R^+ \qquad (CS_2)$$

Το μοντέλο αυτό υποθέτει ότι οι πιθανότητες των κελιών δεν είναι συμμετρικές, αλλά μιας σταθερής αναλογίας. Για παράδειγμα, στα πλαίσια της συμφωνίας βαθμολογητών, αυτό θα σήμαινε ότι ο ένας βαθμολογητής είναι σταθερά περισσότερο ή λιγότερο αυστηρός από τον άλλο, ανάλογα με το αν η σταθερά c είναι μεγαλύτερη ή μικρότερη του 1.

Επίσης, το μοντέλο (CS) έχει ισοδύναμα ορισθεί από τον *McCullagh* (1978) ως εξής

$$\pi_{ij} = \begin{cases} \Delta\psi_{ij} & i < j \\ \psi_{ij} & i \geq j \end{cases}, \text{ όπου } \psi_{ij} = \psi_{ji} \qquad (CS_3)$$

Ειδική περίπτωση για $\Delta = 1$, είναι το γνωστό μοντέλο συμμετρίας (S), της παραγράφου 6.4.1.

Οι *Tomizawa et al*. (1999) εξέφρασαν το ανωτέρω μοντέλο ως

$$\frac{\pi_{ij}}{\pi_{ji}} = \Delta \text{ για } i < j \qquad (CS_4)$$

6.6.2 Μοντέλο Τριγωνικής Συμμετρίας $(TS - Triangular\ Symmetry)$

Εναλλακτικά, το μοντέλο (CS) έχει ισοδύναμα εκφρασθεί από τον *Goodman* (1979) ως

$$\pi_{ij} = \begin{cases} \tau\pi_{ij}^{S} & i > j \\ (2-\tau)\pi_{ij}^{S} & i < j \end{cases} \qquad (TS_1)$$

όπου $\tau \in (0,2)$ και π_{ij}^{S} οι πιθανότητες των κελιών σε συνθήκες πλήρης συμμετρίας με $\pi_{ij}^{S} = (\pi_{ij} + \pi_{ji})/2$. Ο *Goodman* θεώρησε το μοντέλο (CS) σαν ένα μοντέλο που απομακρύνεται από την συμμετρία και το ονόμασε μοντέλο τριγωνικής συμμετρίας (TS) για ευνόητους λόγους. Σύμφωνα με το μοντέλο (TS), όλα τα μη-διαγώνια κελιά του κάτω τριγώνου ενός τετραγωνικού πίνακα, είναι σταθερά μικρότερα (ή μεγαλύτερα) συγκρινόμενα με τα άνω συμμετρικά κελιά τους (άνω τρίγωνο), χωρίς να εξετάζουμε την απόσταση ανάμεσα στις κατηγορίες γραμμής και στήλης των ζευγών των συμμετρικών κελιών.

6.6.3 Μοντέλο Διαγώνιας Συμμετρίας $(DS - Diagonal\ Symmetry)$

Στην περίπτωση διατακτικών μεταβλητών ταξινόμησης και καθώς στα πλαίσια αυτά, οι μεταβλητές ταξινόμησης του πίνακα είναι ανάλογου μέτρου, έχει νόημα να εξετάσουμε ένα μοντέλο που είναι ευαίσθητο στην απόσταση μεταξύ των κατηγοριών από την κύρια διαγώνιο. Το μοντέλο διαγώνιας συμμετρίας (DS), το οποίο προτάθηκε από τον *Goodman* (1979), έχει αυτή την ευαισθησία και ορίζεται ως

$$\pi_{ij} = \delta_t \pi_{ji}, \text{ για } i, j = 1,...,R,\ i > j \text{ και } t = i - j \qquad (DS_1)$$

Συνεπώς τα δ_t εξαρτώνται μόνο από την διαφορά $t = i - j$, δηλαδή την απόσταση από την κύρια διαγώνιο. Σημειώνουμε ότι τα δ_t μπορούν να θεωρηθούν σαν τα *odds* ότι μια παρατήρηση θα πέσει σε κάποιο από τα (i, j) κελιά , παρά σε κάποιο από τα (j, i) κελιά, όπου $t = i - j$ και $1 \le t \le R-1$ [βλέπε *Bhattacharya*, (1998)]. Συχνά, όπως ο *Bhattacharya* αναφέρει για το μοντέλο (DS), διάφοροι διατακτικοί περιορισμοί συμπεριλαμβανομένων των δ_t έχουν αρκετό ενδιαφέρον και είναι κατάλληλοι για την ανάλυση πολλών ειδών τετραγωνικών πινάκων συνάφειας διατεταγμένων κατηγοριών. Για παράδειγμα, όταν όλα τα δ_t είναι μεγαλύτερα (μικρότερα) του 1, τότε η μεταβλητή X είναι στοχαστικά μικρότερη (μεγαλύτερη) της Y.

Εναλλακτικά, το μοντέλο (DS) ορίζεται ως

$$\pi_{ij} = \begin{cases} \delta_{i-j}\pi_{ij}^{S} & i > j \\ (2-\delta_{j-i})\pi_{ij}^{S} & i < j \end{cases} \qquad (DS_2)$$

όπου $\delta_{i-j} \in (0,2)$ για $i > j$ και π_{ij}^{S} οι πιθανότητες των κελιών σε συνθήκες πλήρης συμμετρίας με $\pi_{ij}^{S} = (\pi_{ij} + \pi_{ji})/2$.

Σημειώνουμε ότι ο αριθμός των δ - παραμέτρων είναι $I-1$ και περιορίζοντας τους να ισούνται όλοι με $\delta_{i-j} = \delta$ για $i > j$, το μοντέλο (DS) ελαττώνεται στο μοντέλο (TS) με $\tau = \delta$. Επομένως, τα μοντέλα (CS) (ή (TS)) αποτελούν μια ειδική περίπτωση του μοντέλου (DS).

Οι *Tomizawa et al*. (2005), χρησιμοποίησαν τον ακόλουθο ισοδύναμο συμβολισμό, για να περιγράψουν το μοντέλο (DS) του *Goodman* (1979)

$$\pi_{ij} = \begin{cases} \Delta_{j-i}\varphi_{ij} & i < j \\ \varphi_{ij} & i \geq j \end{cases}, \text{ όπου } \varphi_{ij} = \varphi_{ji} \qquad (DS_3)$$

Παρόμοια, ειδικές περιπτώσεις του μοντέλου αυτού, για $\Delta_{j-i} = \Delta$ και για $\Delta_{j-i} = 1$, είναι το μοντέλο δεσμευμένης συμμετρίας (CS) του *McCullagh* (1978) και το μοντέλο συμμετρίας (S) των *Bishop et al*. (1975), αντίστοιχα.

6.7 Μέτρα Ασυμμετρίας

Στην παράγραφο αυτή θα εξετάσουμε διάφορα μέτρα ασυμμετρίας, που μετρούν τον βαθμό απομάκρυνσης από την (DS), την (CS) και την (TS), ενός τετραγωνικού πίνακα συνάφειας με διατακτικές μεταβλητές ταξινόμησης. Τα προτεινόμενα μέτρα βασίζονται στην απόκλιση τύπου *power-divergence* των *Cressie & Read* (1984), τα οποία προτάθηκαν από τον *Tomizawa* (1999, 2005) και στην απόκλιση φ – *divergence* του *Czisar* (1963), τα οποία προτάθηκαν από τους *Kateri & Papaioannou* [Technical Report (2007), University of Pireaus (TR07-3)].

6.7.1 Μέτρο απομάκρυνσης από την Διαγώνια Συμμετρία (DS)

Οι *Tomizawa et al.* (2005), χρησιμοποιώντας τις δεσμευμένες πιθανότητες $\left\{\pi^{U}_{i,i+k}\right\}$ και $\left\{\pi^{L}_{i+k,i}\right\}$, όρισαν εναλλακτικά το μοντέλο $\left(DS-Diagonal\ Symmetry\right)$ ως

$$\pi^{U}_{i,i+k}=\pi^{L}_{i+k,i} \tag{6.16}$$

για $k=1,2,...,R-2$ και $i=1,2,...,R-k$, όπου

$$\pi^{U}_{i,i+k}=\frac{\pi_{i,i+k}}{\delta^{U}_{k}}\left[=\Pr\left(X=i,Y=i+k\middle\|Y-X\right|=k,X<Y\right)\right],$$

$$\pi^{L}_{i+k,i}=\frac{\pi_{i+k,i}}{\delta^{L}_{k}}\left[=\Pr\left(X=i+k,Y=i\middle\|Y-X\right|=k,X>Y\right)\right],$$

$$\delta^{U}_{k}=\sum_{i=1}^{R-k}\pi_{i,i+k}\left[=\Pr\left(\left|Y-X\right|=k,X<Y\right)\right] \text{ και}$$

$$\delta^{L}_{k}=\sum_{i=1}^{R-k}\pi_{i+k,i}\left[=\Pr\left(\left|Y-X\right|=k,X>Y\right)\right]$$

Κατά συνέπεια, το μοντέλο αυτό δείχνει ότι για κάθε απόσταση k από την διαγώνιο, υπάρχει μια δομή συμμετρίας μεταξύ των δυο δεσμευμένων κατανομών $\left\{\pi^{U}_{i,i+k}\right\}$ και $\left\{\pi^{L}_{i+k,i}\right\}$.

Έστω ότι, $\delta^{U}_{k},\delta^{L}_{k}>0$ και $\pi_{i,i+k}+\pi_{i+k,i}>0$. Θέτοντας $\pi^{*}_{i,i+k}=\frac{\pi^{U}_{i,i+k}+\pi^{L}_{i+k,i}}{2}$, τότε το μοντέλο $\left(DS\right)$, μπορεί επίσης να εκφρασθεί ως

$$\pi^{U}_{i,i+k}=\pi^{*}_{i,i+k} \text{ και } \pi^{L}_{i+k,i}=\pi^{*}_{i,i+k}, \text{ για } k=1,2,...,R-2 \text{ και } i=1,2,...,R-k \tag{6.17}$$

Οι *Tomizawa et al.* (2005) πρότειναν το ακόλουθο γενικευμένο μέτρο τύπου *power-divegence* των *Cressie & Read* (1984, 1988), που δείχνει τον βαθμό απομάκρυνσης από την $\left(DS\right)$

A1. $$\Phi^{(\lambda)}_{DS}=\frac{1}{\Gamma}\sum_{k=1}^{R-2}\left(\delta^{U}_{k}+\delta^{L}_{k}\right)\Phi^{(\lambda)}_{k} \text{ για } \lambda>-1$$

όπου $\Gamma = \sum_{t=1}^{R-2}\left(\delta_t^U + \delta_t^L\right)$,

$$\Phi_k^{(\lambda)} = \frac{\lambda(\lambda+1)}{2^\lambda - 1}\frac{1}{2}\left[I_k^{(\lambda)}\left(\left\{p_{i,i+k}^U\right\};\left\{\pi_{i,i+k}^*\right\}\right) + I_k^{(\lambda)}\left(\left\{p_{i+k,i}^L\right\};\left\{\pi_{i+k,i}^*\right\}\right)\right]$$

και $I_k^{(\lambda)}\left(\left\{a_{i,i+k}\right\};\left\{b_{i,i+k}\right\}\right) = \frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{R-k} a_{i,i+k}\left[\left(\frac{a_{i,i+k}}{b_{i,i+k}}\right)^\lambda - 1\right]$, η απόκλιση *power-divergence* μεταξύ των δυο κατανομών $\left\{a_{i,i+k}\right\}$ και $\left\{b_{i,i+k}\right\}$.

Η τιμή του λ επιλέγεται από τον χρήστη και για $\lambda = 0$ υπολογίζουμε το όριο της συνάρτησης $\lim_{\lambda\to 0}\Phi_k^{(\lambda)}$ και έχουμε

$$\Phi_k^{(0)} = \frac{1}{2\log 2}\left[I_k^{(0)}\left(\left\{\pi_{i,i+k}^U\right\};\left\{\pi_{i,i+k}^*\right\}\right) + I^{(0)}\left(\left\{\pi_{i+k,i}^L\right\};\left\{\pi_{i,i+k}^*\right\}\right)\right]$$

όπου

$I^{(0)}\left(\left\{a_{i,i+k}\right\};\left\{b_{i,i+k}\right\}\right) = \sum_{i=1}^{R-k} a_{i,i+k}\log\left(\frac{a_{i,i+k}}{b_{i,i+k}}\right)$ η πληροφορία *Kullback-Leibler* μεταξύ των κατανομών $\left\{a_{i,i+k}\right\}$ και $\left\{b_{i,i+k}\right\}$.

Χρησιμοποιώντας τις δεσμευμένες πιθανότητες $\pi_{i,i+k}^C = \frac{\pi_{i,i+k}^U}{\pi_{i,i+k}^U + \pi_{i+k,i}^L}$ και $\pi_{i+k,i}^C = \frac{\pi_{i+k,i}^L}{\pi_{i,i+k}^U + \pi_{i+k,i}^L}$, το μοντέλο (DS), μπορεί επίσης να εκφραστεί ως

$$\pi_{i,i+k}^C = \pi_{i+k,i}^C = \frac{1}{2},\ \text{για}\ k = 1,2,...,R-2\ \text{και}\ i = 1,2,...,R-k \qquad (6.18)$$

Επομένως, το μέτρο $\Phi_{DS}^{(\lambda)}$ εκφράζεται ισοδύναμα ως

A2. $$\Phi_{DS}^{(\lambda)} = \frac{1}{\Gamma}\sum_{k=1}^{R-2}\left(\delta_k^U + \delta_k^L\right)\Phi_k^{(\lambda)},\ \text{για}\ \lambda > -1$$

όπου

$$\Phi_k^{(\lambda)} = \frac{\lambda(\lambda+1)}{2^\lambda - 1}\sum_{i=1}^{R-k}\pi_{i,i+k}^* I_{i,i+k}^{(\lambda)}$$

με

$$I_{i,i+k}^{(\lambda)}\left(\left\{\pi_{i,i+k}^{C},\pi_{i+k,i}^{C}\right\};\left\{\frac{1}{2},\frac{1}{2}\right\}\right)=\frac{1}{\lambda\left(\lambda+1\right)}\left[\pi_{i,i+k}^{C}\left\{\left(\frac{\pi_{i,i+k}^{C}}{1/2}\right)^{\lambda}-1\right\}+\pi_{i+k,i}^{C}\left\{\left(\frac{\pi_{i+k,i}^{C}}{1/2}\right)^{\lambda}-1\right\}\right]$$

Η τιμή του λ επιλέγεται από τον χρήστη και για $\lambda=0$ υπολογίζουμε το όριο της συνάρτησης $\lim_{\lambda\to 0}\Phi_k^{(\lambda)}$ και έχουμε

$$\Phi_k^{(0)}=\frac{1}{\log 2}\sum_{i=1}^{R-k}\pi_{i,i+k}^{*}I_{i,i+k}^{(0)}$$

όπου

$$I_{i,i+k}^{(0)}\left(\left\{\pi_{i,i+k}^{C},\pi_{i+k,i}^{C}\right\};\left\{\frac{1}{2},\frac{1}{2}\right\}\right)=\pi_{i,i+k}^{C}\log\frac{\pi_{i,i+k}^{C}}{1/2}+\pi_{i+k,i}^{C}\log\frac{\pi_{i+k,i}^{C}}{1/2}.$$

Περαιτέρω, το μέτρο μπορεί να εκφραστεί και ως

Α3. $$\Phi_{DS}^{(\lambda)}=\frac{1}{\Gamma}\sum_{k=1}^{R-2}\left(\delta_k^{U}+\delta_k^{L}\right)\Phi_k^{(\lambda)},\ \text{για}\ \lambda>-1$$

όπου

$$\Phi_k^{(\lambda)}=\sum_{i=1}^{R-k}\pi_{i,i+k}^{*}\left(1-\frac{\lambda 2^{\lambda}}{2^{\lambda}-1}H_{i,i+k}^{(\lambda)}\right)$$

με

$H_{i,i+k}^{(\lambda)}\left(\left\{\pi_{i,i+k}^{C}\right\};\left\{\pi_{i+k,i}^{C}\right\}\right)=\frac{1}{\lambda}\left[1-\left(\pi_{i,i+k}^{C}\right)^{\lambda+1}-\left(\pi_{i+k,i}^{C}\right)^{\lambda+1}\right]$, ο δείκτης ανομοιότητας ή ποικιλότητας (*diversity index*) των *Patil & Taillie* (1982), των κατανομών $\left\{\pi_{i,i+k}^{C},\pi_{i+k,i}^{C}\right\}$, όποιος περιλαμβάνει ως ειδική περιπτώση την εντροπία του *Shannon* (για $\lambda=0$).

Η τιμή του λ επιλέγεται από τον χρήστη και για $\lambda=0$ υπολογίζουμε το όριο της συνάρτησης $\lim_{\lambda\to 0}\Phi_k^{(\lambda)}$ και έχουμε

$$\Phi_k^{(0)}=\sum_{i=1}^{R-k}\pi_{i,i+k}^{*}\left[1-\frac{1}{\log 2}H_{i,i+k}^{(0)}\right]$$

όπου $H_{i,i+k}^{(0)}=-\pi_{i,i+k}^{C}\log\pi_{i,i+k}^{C}-\pi_{i+k,i}^{C}\log\pi_{i+k,i}^{C}$, η εντροπία του *Shannon*.

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 8×8 πίνακα συνάφειας (*σελ.169, Παράδειγμα 6*) έχουμε τα κάτωθι αποτελέσματα, βασιζόμενοι στον τύπο **A1**. Σημειώνουμε ότι οι **A2** και **A3** τύποι υπολογισμού του μέτρου $\Phi_{DS}^{(\lambda)}$, είναι ισοδύναμοι με τον **A1** και προς ευκολία της παρουσίασης παραλείπονται.

	$\Phi_{DS}^{(\lambda)}$ **Cressie and Read (power divergence)**		
λ	***1955***	***1965***	***1975***
-0,80	0,036	0,085	0,065
-0,40	0,088	0,197	0,155
-0,20	0,107	0,234	0,185
0,00	0,122	0,262	0,209
1,00	0,158	0,323	0,264
1,40	0,160	0,327	0,268
1,60	0,160	0,327	0,268
1,80	0,159	0,325	0,267
2,00	0,158	0,323	0,264

Ερμηνεία

Για κάθε $\lambda > -1$, σε έναν $R\times R$ πίνακα συνάφειας, υπάρχει δομή (DS), για παράδειγμα $\left\{\pi_{i,i+k}^{C}=\pi_{i+k,i}^{C}\right\}$, αν και μόνο αν $\Phi_{DS}^{(\lambda)}=0$. Ο βαθμός απομάκρυνσης από την (DS) μεγιστοποιείται (πλήρης δεσμευμένη ασυμμετρία), υπο την έννοια ότι, όταν $\pi_{i,i+k}^{C}=0$ τότε $\pi_{i+k,i}^{C}=1$ ή όταν $\pi_{i+k,i}^{C}=0$ τότε $\pi_{i,i+k}^{C}=1$, αν και μόνο αν $\Phi_{DS}^{(\lambda)}=1$, για $k=1,2,...,R-2$ και $i=1,2,...,R-k$. Σημειώνουμε ότι όταν $\Phi_{DS}^{(\lambda)}=1$, αυτό σημαίνει ότι για $\pi_{i,i+k}^{U}/\pi_{i+k,i}^{L}=\infty$ για κάποια i και $\pi_{i,i+k}^{U}/\pi_{i+k,i}^{L}=0$ για τα υπόλοιπα i και επομένως είναι λογικό να θεωρήσουμε ότι ο βαθμός απομάκρυνσης από την (DS) (για παράδειγμα από την κατάσταση όπου $\pi_{ij}^{U}/\pi_{ji}^{L}=l$), είναι ο μέγιστος. Ο βαθμός απομάκρυνσης από την (DS) μεγαλώνει όσο η τιμή του μέτρου $\Phi_{DS}^{(\lambda)}$ μεγαλώνει.

Πεδίο Ορισμού

Το μέτρο $\Phi_{DS}^{(\lambda)}$ κυμαίνεται στο διάστημα $[0,1]$, καθώς για κάθε k, η ποσότητα $\Phi_{k}^{(\lambda)}$ κυμαίνεται στο διάστημα $[0,1]$, διότι ισχύει ότι $I_{i,i+k}^{(\lambda)}\geq 0$, $H_{i,i+k}^{(\lambda)}\geq 0$.

Επίπεδο Δεδομένων

Το μέτρο $\Phi_{DS}^{(\lambda)}$ χρησιμοποιείται σε πίνακες συνάφειας με διατακτικές μεταβλητές, καθώς εξαρτάται από την διάταξη των κατηγοριών και δεν παραμένει αμετάβλητο κάτω από τις ίδιες αναδιατάξεις των γραμμών και των στηλών.

Διάστημα Εμπιστοσύνης

Η δειγματική εκτίμηση του μέτρου $\Phi_{DS}^{(\lambda)}$ συμβολίζεται με $\hat{\Phi}_{DS}^{(\lambda)}$ και υπολογίζεται αντικαθιστώντας τις πληθυσμιακές πιθανότητες $\{\pi_{ij}\}$ με τις παρατηρούμενες πιθανότητες του δείγματος $\{p_{ij}\}$, όπου $p_{ij} = n_{ij}/n$ και $n = \sum_{i=1}^{R}\sum_{j=1}^{R} n_{ij}$. Χρησιμοποιώντας την μέθοδο Δέλτα, η ποσότητα $\sqrt{n}\left(\hat{\Phi}_{DS}^{(\lambda)} - \Phi_{DS}^{(\lambda)}\right)$, ασυμπτωτικά $(n \to \infty)$ ακολουθεί την κανονική κατανομή με μέση τιμή $\mu = 0$ και διακύμανση

$$\sigma^2\left[\Phi_{DS}^{(\lambda)}\right] = \frac{1}{\Gamma^2}\sum_{k=1}^{R-2}\sum_{i=1}^{R-k}\left[\pi_{i,i+k}\left(\omega_{i,i+k}^{(\lambda)}\right)^2 + \pi_{i+k,i}\left(\omega_{i+k,i}^{(\lambda)}\right)^2\right]$$

όπου για $\lambda > -1$

$$\omega_{i,i+k}^{(\lambda)} = \Phi_k^{(\lambda)} - \Phi_{DS}^{(\lambda)} + \frac{\delta_k^U + \delta_k^L}{2\delta_k^U}\left[\upsilon_{i,i+k}^{(\lambda)} - \sum_{m=1}^{R-k}\pi_{m,m+k}^U \upsilon_{m,m+k}^{(\lambda)}\right]$$

$$\omega_{i+k,i}^{(\lambda)} = \Phi_k^{(\lambda)} - \Phi_{DS}^{(\lambda)} + \frac{\delta_k^U + \delta_k^L}{2\delta_k^L}\left[\upsilon_{i+k,i}^{(\lambda)} - \sum_{m=1}^{R-k}\pi_{m+k,m}^L \upsilon_{m+k,m}^{(\lambda)}\right]$$

και

$$\upsilon_{st}^{(\lambda)} = \frac{1}{2^\lambda - 1}\left[\left(2\pi_{st}^C\right)^\lambda + \lambda\pi_{ts}^C\left[\left(2\pi_{st}^C\right)^\lambda - \left(2\pi_{ts}^C\right)^\lambda\right]\right], \text{ για } \lambda \neq 0$$

$$\upsilon_{st}^{(0)} = \frac{\log\left(2\pi_{st}^C\right)}{\log 2}, \text{ για } \lambda = 0$$

Η εκτίμηση της διακύμανσης $\sigma^2\left[\Phi_{DS}^{(\lambda)}\right]$ του μέτρου $\Phi_{DS}^{(\lambda)}$, συμβολίζεται με $\hat{\sigma}^2\left[\hat{\Phi}_{DS}^{(\lambda)}\right]$ και υπολογίζεται αντικαθιστώντας τις πληθυσμιακές πιθανότητες $\{\pi_{ij}\}$ με τις δειγματικές $\{p_{ij}\}$. Η ποσότητα $\hat{\sigma}\left[\hat{\Phi}_{DS}^{(\lambda)}\right]/\sqrt{n}$ είναι η κατά προσέγγιση εκτιμήτρια του τυπικού σφάλματος του

μέτρου $\Phi_{DS}^{(\lambda)}$ και το κατά προσέγγιση $100\% \times (1-a)$ διάστημα εμπιστοσύνης της εκτιμήτριας του μέτρου $\Phi_{DS}^{(\lambda)}$ είναι

$$\hat{\Phi}_{DS}^{(\lambda)} \pm z_{a/2}\, \hat{\sigma}\left[\Phi_{DS}^{(\lambda)}\right] \Big/ \sqrt{n} \tag{6.19}$$

όπου $z_{a/2}$ το ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής.

Για περισσότερες λεπτομέρειες αναφορικά με την μέθοδο Δέλτα [βλέπε *Bishop et al*. (1975) και *Agresti* (1984)].

Σχόλια

Το μέτρο $\Phi_{DS}^{(\lambda)}$ κυμαίνεται στο διάστημα $[0,1]$ ανεξάρτητα από το μέγεθος n του δείγματος και από την διάσταση R του πίνακα και επομένως είναι κατάλληλο για τον σύγκριση του βαθμού απομάκρυνσης από την (DS), για διάφορους πίνακες (με διαφορετικό μέγεθος δείγματος). Το μέτρο $\Phi_{DS}^{(\lambda)}$ είναι χρήσιμο, όταν κάποιος θέλει να δει το βαθμό απομάκρυνσης από την (DS), ως προς τον μέγιστο βαθμό απομάκρυνσης από την (DS), όπως περιγράφηκε στην ερμηνεία του μέτρου (*σελ. 145*). Όπως αναφέραμε η τιμή του λ επιλέγεται από τον χρήστη και είναι αδύνατο να προσδιορίσουμε την καταλληλότερη τιμή του λ για έναν πίνακα. Είναι σκόπιμο και ασφαλές, για την σύγκριση του βαθμού απομάκρυνσης από την (DS), μεταξύ διάφορων πινάκων, ο αναλυτής να υπολογίσει το μέτρο $\Phi_{DS}^{(\lambda)}$ για διάφορες τιμές του λ. Η εκτίμηση του μέτρου $\Phi_{CS}^{(\lambda)}$ θα πρέπει να υπολογίζεται σε όρους ενός κατά προσέγγιση διαστήματος εμπιστοσύνης και όχι απλά εκτιμώντας μια τιμή του.

6.7.2 Μέτρο απομάκρυνσης από την Δεσμευμένη Συμμετρία (CS)

Όπως έχουμε αναφέρει (*παράγραφος 6.6.1, σελ. 139*), το μοντέλο δεσμευμένης συμμετρίας $(CS-Conditional\ \ Symmetry)$, ορίζεται ως $\dfrac{\pi_{ij}}{\pi_{ji}} = \Delta$, για $i < j$ και εκφράζει ότι ο λόγος της πιθανότητας μιας παρατήρησης να βρίσκεται στο κελί (i, j), του άνω δεξιού τριγώνου του

πίνακα συνάφειας, ως προς την πιθανότητα μια παρατήρηση να βρίσκεται στο κελί (j,i), του κάτω αριστερού τριγώνου, είναι σταθερός για κάθε $i < j$.

Εναλλακτικά, οι *Tomizawa et al*. (1999), εξέφρασαν το μοντέλο αυτό ως

$$\pi_{ij}^{U} = \pi_{ji}^{L} \text{ για } i < j \tag{6.20}$$

όπου

$\pi_{ij}^{U} = \frac{\pi_{ij}}{\delta_U}$, $\pi_{ji}^{L} = \frac{\pi_{ji}}{\delta_L}$ και $\delta_U = \sum\sum_{s<t} \pi_{st}$, $\delta_L = \sum\sum_{s>t} \pi_{st}$.

Δηλαδή, η πιθανότητα μια παρατήρηση να βρίσκεται στο κελί (i, j), δοθέντος ότι θα βρίσκεται σε ένα από τα κελιά του άνω δεξιού τριγώνου του πίνακα συνάφειας, ισούται με την πιθανότητα μια παρατήρηση να βρίσκεται στο κελί (j,i), δοθέντος ότι θα βρίσκεται σε ένα από τα κελιά του κάτω αριστερού τριγώνου. Με άλλα λόγια, το μοντέλο της (CS) εκφράζει ότι υπάρχει δομή συμμετρίας μεταξύ των δεσμευμένων πιθανοτήτων $\{\pi_{ij}^{U}\}$ και $\{\pi_{ji}^{L}\}$, για $i < j$.

Όταν το μοντέλο της (CS) δεν ισχύει (και επομένως το μοντέλο (S), επίσης δεν ισχύει), ενδιαφερόμαστε να μετρήσουμε τον βαθμό απομάκρυνσης από την (CS). Έστω ότι, $\delta_U, \delta_L > 0$ και $\pi_{ij} + \pi_{ji} > 0$, για $i, j = 1, 2, ..., R$ με $i \neq j$. Θέτοντας $\pi_{ij}^{*} = \frac{\pi_{ij}^{U} + \pi_{ji}^{L}}{2}$ για $i < j$ (σημειώνουμε ότι $\sum\sum_{i<j} \pi_{ij}^{*} = 1$), τότε το μοντέλο (CS), μπορεί επίσης να εκφρασθεί ως

$$\pi_{ij}^{U} = \pi_{ij}^{*} \text{ και } \pi_{ji}^{L} = \pi_{ij}^{*}, \text{ για } i < j \tag{6.21}$$

Οι *Tomizawa et al*. (1999) πρότειναν το ακόλουθο μέτρο τύπου *power-divegence* των *Cressie & Read* (1984, 1988)

A1. $$\Phi_{CS}^{(\lambda)} = \frac{\lambda(\lambda+1)}{2^{\lambda}-1} \frac{1}{2}\left[I^{(\lambda)}\left(\{\pi_{ij}^{U}\};\{\pi_{ij}^{*}\}\right) + I^{(\lambda)}\left(\{\pi_{ji}^{L}\};\{\pi_{ij}^{*}\}\right)\right],$$

για $\lambda > -1$, όπου

$I^{(\lambda)}\left(\{a_{ij}\};\{b_{ij}\}\right)=\frac{1}{\lambda(\lambda+1)}\sum\sum_{i<j}a_{ij}\left[\left(\frac{a_{ij}}{b_{ij}}\right)^{\lambda}-1\right]$ η απόκλιση *power-divergence* μεταξύ των δυο κατανομών $\{a_{ij}\}$ και $\{b_{ij}\}$.

Η τιμή του λ επιλέγεται από τον χρήστη και για $\lambda=0$ υπολογίζουμε το όριο της συνάρτησης $\lim_{\lambda\to 0}\Phi_{CS}^{(\lambda)}$ και έχουμε

$$\Phi_{CS}^{(0)}=\frac{1}{2\log 2}\left[I^{(0)}\left(\{\pi_{ij}^{U}\};\{\pi_{ij}^{*}\}\right)+I^{(0)}\left(\{\pi_{ji}^{L}\};\{\pi_{ij}^{*}\}\right)\right]$$

όπου

$I^{(0)}\left(\{a_{ij}\};\{b_{ij}\}\right)=\sum\sum_{i<j}a_{ij}\log\left(\frac{a_{ij}}{b_{ij}}\right)$ η πληροφορία *Kullback-Leibler* μεταξύ των κατανομών $\{a_{ij}\}$ και $\{b_{ij}\}$.

Χρησιμοποιώντας τις δεσμευμένες πιθανότητες $\pi_{ij}^{C}=\frac{\pi_{ij}^{U}}{\pi_{ij}^{U}+\pi_{ji}^{L}}$ και $\pi_{ji}^{C}=\frac{\pi_{ji}^{L}}{\pi_{ij}^{U}+\pi_{ji}^{L}}$, για $i<j$, το μοντέλο (CS), μπορεί επίσης να εκφραστεί ως

$$\pi_{ij}^{C}=\pi_{ji}^{C}=\frac{1}{2},\ \text{για}\ i<j \tag{6.22}$$

Επομένως, το μέτρο εκφράζεται ισοδύναμα ως

A2. $$\Phi_{CS}^{(\lambda)}=\frac{\lambda(\lambda+1)}{2^{\lambda}-1}\sum\sum_{i<j}\pi_{ij}^{*}I_{ij}^{(\lambda)}\left(\left\{\pi_{ij}^{C},\pi_{ji}^{C}\right\};\left\{\frac{1}{2},\frac{1}{2}\right\}\right),$$

για $\lambda>-1$, όπου

$$I_{ij}^{(\lambda)}\left(\left\{\pi_{ij}^{C},\pi_{ji}^{C}\right\};\left\{\frac{1}{2},\frac{1}{2}\right\}\right)=\frac{1}{\lambda(\lambda+1)}\left[\pi_{ij}^{C}\left\{\left(\frac{\pi_{ij}^{C}}{1/2}\right)^{\lambda}-1\right\}+\pi_{ji}^{C}\left\{\left(\frac{\pi_{ji}^{C}}{1/2}\right)^{\lambda}-1\right\}\right]$$

Επομένως, το μέτρο **A2** αναπαριστά το σταθμισμένο άθροισμα της απόκλισης *power-divergence* $I_{ij}^{(\lambda)}\left(\left\{\pi_{ij}^{C},\pi_{ji}^{C}\right\};\left\{\frac{1}{2},\frac{1}{2}\right\}\right)$.

Η τιμή του λ επιλέγεται από τον χρήστη και για $\lambda = 0$ υπολογίζουμε το όριο της συνάρτησης $\lim_{\lambda \to 0} \Phi_{CS}^{(\lambda)}$ και έχουμε

$$\Phi_{CS}^{(0)} = \frac{1}{\log 2} \sum \sum_{i<j} \pi_{ij}^{*} I_{ij}^{(0)} \left(\left\{ \pi_{ij}^{C}, \pi_{ji}^{C} \right\}; \left\{ \frac{1}{2}, \frac{1}{2} \right\} \right)$$

όπου

$$I_{ij}^{(0)} \left(\left\{ \pi_{ij}^{C}, \pi_{ji}^{C} \right\}; \left\{ \frac{1}{2}, \frac{1}{2} \right\} \right) = \pi_{ij}^{C} \log \frac{\pi_{ij}^{C}}{1/2} + \pi_{ji}^{C} \log \frac{\pi_{ji}^{C}}{1/2}.$$

Περαιτέρω, το μέτρο μπορεί να εκφραστεί και ως

Α3. $$\Phi_{CS}^{(\lambda)} = \sum \sum_{i<j} \pi_{ij}^{*} \left[1 - \frac{\lambda 2^{\lambda}}{2^{\lambda} - 1} H_{ij}^{(\lambda)} \left(\left\{ \pi_{ij}^{C}, \pi_{ji}^{C} \right\} \right) \right],$$

για $\lambda > -1$, όπου

$H_{ij}^{(\lambda)} \left(\left\{ \pi_{ij}^{C}, \pi_{ji}^{C} \right\} \right) = \frac{1}{\lambda} \left[1 - \left(\pi_{ij}^{C} \right)^{\lambda+1} - \left(\pi_{ji}^{C} \right)^{\lambda+1} \right]$, ο δείκτης ανομοιότητας ή ποικιλότητας (*diversity index*) των *Patil & Taillie* (1982), ο οποίος περιλαμβάνει ως ειδικές περιπτώσεις την εντροπία του *Shannon* (για $\lambda = 0$) και τον δείκτη συγκέντρωσης *Gini Concentration* (ή δείκτη *Simpson*) (για $\lambda = 0$). Για περισσότερες λεπτομέρειες αναφορικά με τον δείκτη *Gini Concentration*, βλεπε *Haberman*, (1982).

Η τιμή του λ επιλέγεται από τον χρήστη και για $\lambda = 0$ υπολογίζουμε το όριο της συνάρτησης $\lim_{\lambda \to 0} \Phi_{CS}^{(\lambda)}$ και έχουμε

$$\Phi_{CS}^{(0)} = \sum \sum_{i<j} \pi_{ij}^{*} \left[1 - \frac{1}{\log 2} H_{ij}^{(0)} \left(\left\{ \pi_{ij}^{C}, \pi_{ji}^{C} \right\} \right) \right]$$

όπου

$H_{ij}^{(0)} \left(\left\{ \pi_{ij}^{C}, \pi_{ji}^{C} \right\} \right) = -\pi_{ij}^{C} \log \pi_{ij}^{C} - \pi_{ji}^{C} \log \pi_{ji}^{C}$, η εντροπία του *Shannon*.

Παράδειγμα

Χρησιμοποιώντας τα δεδομένα του 8×8 πίνακα συνάφειας (*σελ.169, Παράδειγμα 6*), έχουμε τα κάτωθι αποτελέσματα, βασιζόμενοι στον τύπο **Α1**. Σημειώνουμε ότι οι **Α2** και **Α3** τύποι

υπολογισμού του μέτρου $\Phi_{CS}^{(\lambda)}$, είναι ισοδύναμοι με τον **A1** και προς ευκολία της παρουσίασης παραλείπονται.

	$\Phi_{CS}^{(\lambda)}$ **Cressie and Read (power divergence)**		
λ	***1955***	***1965***	***1975***
-0,80	0,177	0,096	0,077
-0,40	0,102	0,225	0,186
-0,20	0,123	0,267	0,224
0,00	0,141	0,299	0,254
1,00	0,186	0,370	0,322
1,40	0,177	0,375	0,327
1,60	0,186	0,374	0,327
1,80	0,185	0,373	0,325
2,00	0,183	0,370	0,322

Ερμηνεία

Για κάθε $\lambda > -1$ υπάρχει δομή (CS), σε έναν $R \times R$ πίνακα συνάφειας, αν και μόνο αν $\Phi_{CS}^{(\lambda)} = 0$. Ο βαθμός απομάκρυνσης από την (CS) μεγιστοποιείται, υπο την έννοια ότι, όταν $\pi_{ij}^{C} = 0$ τότε $\pi_{ji}^{C} = 1$ ή όταν $\pi_{ji}^{C} = 0$ τότε $\pi_{ij}^{C} = 1$, δηλαδή, όταν $\pi_{ij} = 0$ τότε $\pi_{ji} > 0$ και όταν $\pi_{ji} = 0$ τότε $\pi_{ij} > 0$, αν και μόνο αν $\Phi_{CS}^{(\lambda)} = 1$, για $i, j = 1, 2, ..., R$ με $i < j$. Σημειώνουμε ότι όταν $\Phi_{CS}^{(\lambda)} = 1$, αυτό σημαίνει ότι για $\pi_{ij}^{U} / \pi_{ji}^{L} = \infty$ για κάποια $i < j$ και $\pi_{ij}^{U} / \pi_{ji}^{L} = 0$ για τα υπόλοιπα $i < j$ (πλήρης δεσμευμένη ασυμμετρία) και επομένως είναι λογικό να θεωρήσουμε ότι ο βαθμός απομάκρυνσης από την (CS) (για παράδειγμα από την κατάσταση όπου $\pi_{ij}^{U} / \pi_{ji}^{L} = 1$, για $i < j$), είναι ο μέγιστος. Ο βαθμός απομάκρυνσης από την (CS) μεγαλώνει όσο η τιμή του μέτρου $\Phi_{CS}^{(\lambda)}$ μεγαλώνει.

Πεδίο Ορισμού

Το μέτρο $\Phi_{CS}^{(\lambda)}$ κυμαίνεται στο διάστημα $[0,1]$, καθώς $I_{ij}^{(\lambda)}\left(\left\{\pi_{ij}^{C}, \pi_{ji}^{C}\right\}; \left\{\frac{1}{2}, \frac{1}{2}\right\}\right) \geq 0$ και $H_{ij}^{(\lambda)}\left(\left\{\pi_{ij}^{C}, \pi_{ji}^{C}\right\}\right) \geq 0$

Επίπεδο Δεδομένων

Το μέτρο $\Phi_{CS}^{(\lambda)}$ χρησιμοποιείται σε πίνακες συνάφειας με διατακτικές μεταβλητές, καθώς εξαρτάται από την διάταξη των κατηγοριών και δεν παραμένει αμετάβλητο κάτω από τις ίδιες αναδιατάξεις των γραμμών και των στηλών.

Διάστημα Εμπιστοσύνης

Η δειγματική εκτίμηση του μέτρου $\Phi_{CS}^{(\lambda)}$ συμβολίζεται με $\hat{\Phi}_{CS}^{(\lambda)}$ και υπολογίζεται αντικαθιστώντας τις πληθυσμιακές πιθανότητες $\{\pi_{ij}\}$ με τις παρατηρούμενες πιθανότητες του δείγματος $\{p_{ij}\}$, όπου $p_{ij} = n_{ij}/n$ και $n = \sum_{i=1}^{R}\sum_{j=1}^{R} n_{ij}$. Χρησιμοποιώντας την μέθοδο Δέλτα, η ποσότητα $\sqrt{n}\left(\hat{\Phi}_{CS}^{(\lambda)} - \Phi_{CS}^{(\lambda)}\right)$, ασυμπτωτικά $(n \to \infty)$ ακολουθεί την κανονική κατανομή με μέση τιμή $\mu = 0$ και διακύμανση

$$\sigma^2\left[\Phi_{CS}^{(\lambda)}\right] = \sum_{i=1}^{R}\sum_{j=1}^{R} \pi_{ij}\left(\omega_{ij}^{(\lambda)}\right)^2$$

όπου για $\lambda > -1$ και $\lambda \neq 0$

$$\omega_{ij}^{(\lambda)} = \begin{cases} \dfrac{1}{2\left(2^{\lambda}-1\right)\delta_U}\left(\Delta_{ij}^{(\lambda)} - \sum\sum_{s<t} \pi_{st}^{U}\Delta_{st}^{(\lambda)}\right), & i < j \\ \dfrac{1}{2\left(2^{\lambda}-1\right)\delta_L}\left(\Delta_{ij}^{(\lambda)} - \sum\sum_{s>t} \pi_{st}^{L}\Delta_{st}^{(\lambda)}\right), & i > j \end{cases}$$

με $\Delta_{ij}^{(\lambda)} = \left(2\pi_{ij}^{C}\right)^{\lambda} - 1 + \lambda\pi_{ji}^{C}\left[\left(2\pi_{ij}^{C}\right)^{\lambda} - \left(2\pi_{ji}^{C}\right)^{\lambda}\right]$

και όπου για $\lambda = 0$

$$\omega_{ij}^{(0)} = \begin{cases} \dfrac{1}{2(\log 2)\delta_U}\left(\Delta_{ij}^{(0)} - \sum\sum_{s<t} \pi_{st}^{U}\Delta_{st}^{(0)}\right), & i < j \\ \dfrac{1}{2(\log 2)\delta_L}\left(\Delta_{ij}^{(0)} - \sum\sum_{s>t} \pi_{st}^{L}\Delta_{st}^{(0)}\right), & i > j \end{cases}$$

με $\Delta_{ij}^{(0)} = \log\left(2\pi_{ij}^{C}\right)$

Η εκτίμηση της διακύμανσης $\sigma^2\left[\Phi_{CS}^{(\lambda)}\right]$ του μέτρου $\Phi_{CS}^{(\lambda)}$, συμβολίζεται με $\hat{\sigma}^2\left[\hat{\Phi}_{CS}^{(\lambda)}\right]$ και υπολογίζεται αντικαθιστώντας τις πληθυσμιακές πιθανότητες $\left\{\pi_{ij}\right\}$ με τις παρατηρούμενες πιθανότητες $\left\{p_{ij}\right\}$. Η ποσότητα $\hat{\sigma}\left[\hat{\Phi}_{CS}^{(\lambda)}\right]/\sqrt{n}$ είναι η κατά προσέγγιση εκτιμήτρια του τυπικού σφάλματος του μέτρου $\Phi_{CS}^{(\lambda)}$ και το κατά προσέγγιση $100\% \times (1-a)$ διάστημα εμπιστοσύνης της εκτίμησης του μέτρου $\Phi_{CS}^{(\lambda)}$ είναι

$$\hat{\Phi}_{CS}^{(\lambda)} \pm z_{a/2}\, \hat{\sigma}\left[\hat{\Phi}_{CS}^{(\lambda)}\right]/\sqrt{n} \tag{6.23}$$

όπου $z_{a/2}$ το ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής.

Για περισσότερες λεπτομέρειες αναφορικά με την μέθοδο Δέλτα [βλέπε *Bishop et al*. (1975) και *Agresti* (1984)].

Σχόλια

Το μέτρο $\Phi_{CS}^{(\lambda)}$ κυμαίνεται στο διάστημα $[0,1]$ ανεξάρτητα από το μέγεθος n του δείγματος και από την διάσταση R του πίνακα και επομένως είναι κατάλληλο για τον σύγκριση του βαθμού απομάκρυνσης από την (CS), για διάφορους πίνακες (με διαφορετικό μέγεθος δείγματος). Το μέτρο $\Phi_{CS}^{(\lambda)}$ είναι χρήσιμο, όταν κάποιος θέλει να δει το βαθμό απομάκρυνσης από την (CS), ως προς τον μέγιστο βαθμό απομάκρυνσης από την (CS), όπως περιγράφηκε στην ερμηνεία του μέτρου (*σελ. 151*). Όπως αναφέραμε η τιμή του λ επιλέγεται από τον χρήστη και είναι αδύνατο να προσδιορίσουμε την καταλληλότερη τιμή του λ για έναν πίνακα. Είναι σκόπιμο και ασφαλές, για την σύγκριση του βαθμού απομάκρυνσης από την (CS) μεταξύ διάφορων πινάκων, ο αναλυτής να υπολογίσει το μέτρο $\Phi_{CS}^{(\lambda)}$ για διάφορες τιμές του λ. Η εκτίμηση του μέτρου $\Phi_{CS}^{(\lambda)}$ θα πρέπει να υπολογίζεται σε όρους ενός κατά προσέγγιση διαστήματος εμπιστοσύνης και όχι απλά εκτιμώντας μια τιμή του μέτρου. Η ασυμπτωτική κανονική κατανομή της ποσότητας $\sqrt{n}\left(\hat{\Phi}_{CS}^{(\lambda)} - \Phi_{CS}^{(\lambda)}\right)$ δεν εφαρμόζεται όταν $\Phi_{CS}^{(\lambda)} = 0$ και $\Phi_{CS}^{(\lambda)} = 1$, διότι τότε $\sigma^2\left[\Phi_{CS}^{(\lambda)}\right] = 0$. Για τιμές του $\lambda \le -1$,το μέτρο $\Phi_{CS}^{(\lambda)}$ δεν είναι κατάλληλο, καθώς ισχύει πάντα ότι $\Phi_{CS}^{(\lambda)} = 0$, για $\lambda = -1$ και για $\lambda < -1$ το μέτρο είναι ένας μη θετικός αριθμός.

6.7.3 Μέτρο απομάκρυνσης απο την Τριγωνική Συμμετρία (TS)

Ένα διάνυσμα πιθανότητας π για ένα RxR πίνακα συνάφειας έχει δομή (TS), αν $\pi = \pi^{\mathrm{T}}$, με $\pi^{\mathrm{T}} = (\pi_{11}, \pi_{12}^{\mathrm{T}}, ..., \pi_{ij}^{\mathrm{T}}, ..., \pi_{RR}^{\mathrm{T}})$ και

$$\pi_{ij}^{T} = \begin{cases} \tau\pi_{ij}^{S}, & i < j \\ (2-\tau)\pi_{ij}^{S}, & i > j \end{cases} \tag{6.24}$$

για $i, j = 1, ..., R$ και $\tau = 2\sum_{i<j}\pi_{ij}^{T} \Big/ \sum_{i\neq j}\pi_{ij}^{T} = 2\sum_{i<j}\pi_{ij}^{T} \Big/ \delta_{\pi}$ με $\delta_{\pi} = \sum_{i\neq j}\pi_{ij}$

Η $\varphi - divergence$ του *Csiszar* (1963) μεταξύ των διανυσμάτων π και π^{T} ορίζεται ως $I_C(\pi\ , \pi^{T}) = \sum_{i\neq j}\pi_{ij}^{T}\phi\left(\pi_{ij}/\pi_{ij}^{\mathrm{T}}\right)$

και ισχύει ότι $I_C(\pi\ , \pi^{T}) = \sum_{i\neq j}\pi_{ij}^{T}\phi\left(\pi_{ij}/\pi_{ij}^{\mathrm{T}}\right) \le \phi_0^{\mathrm{T}}\delta_{\pi}/2 \le \phi_0^{\mathrm{T}}/2$, δηλαδή φράσσεται από την ποσότητα $\phi_0^{\mathrm{T}}/2$, με $\phi_0^{\mathrm{T}} = \max\left(\xi_{\mathrm{T}}(0), \xi_{\mathrm{T}}(2)\right)$ και $\xi_{T(x)} = \tau\phi\left(\frac{x}{\tau}\right) + (2-\tau)\phi\left(\frac{2-x}{2-\tau}\right)$.

Οι *Kateri & Papaioannou* (2007), πρότειναν το ακόλουθο γενικευμένο μέτρο, που δείχνει τον βαθμό απομάκρυνσης από την (TS) και βασίζεται στην $\varphi -$ *divergence*

$$\tau_{\phi}^{T}\left(\pi, \pi^{T}\right) = \frac{2}{\phi_0^{T}\delta_{\pi}}\sum_{i\neq j}\pi_{ij}^{T}\phi\left(\frac{\pi_{ij}}{\pi_{ij}^{T}}\right) \tag{6.25}$$

ή ισοδύναμα

$$\tau_{\phi}^{T}\left(\pi, \pi^{T}\right) = \frac{2}{\phi_0^{T}\delta_{\pi}}\sum_{i<j}\frac{\pi_{ij}+\pi_{ji}}{2}\left\{\tau\phi\left(\frac{2\tau^{-1}\pi_{ij}}{\pi_{ij}+\pi_{ji}}\right) + (2-\tau)\phi\left(\frac{2(2-\tau)^{-1}\pi_{ji}}{\pi_{ij}+\pi_{ji}}\right)\right\} \tag{6.26}$$

Το άνω όριο ϕ_0^{T} εξαρτάται από την τιμή τ (για παράδειγμα εξαρτάται από την πιθανότητα π) και δυσκολεύει την εκτίμηση και τις ασυμπτωματικές πλευρές αυτού του μέτρου. Για να ξεπεράσουν το πρόβλημα αυτό, οι *Kateri & Papaioannou,* βρήκαν μια εναλλακτική έκφραση της (TS), η οποία δεν προκύπτει άμεσα από το διάνυσμα π, αλλά από έναν κατάλληλο μετασχηματισμό αυτού.

Ένας RxR πίνακας συνάφειας με πιθανότητα π_{ij} σε κάθε κελί, για $i, j = 1,...,R$, έχει δομή (TS), αν και μόνο αν ο πίνακας $\left(\left(\psi_{ij}\right)\right)_{RxR}$ είναι συμμετρικός, όπου

$$\psi_{ij} = \begin{cases} \delta_\pi \pi_{ij} \left[\left(\sum_{i<j} \pi_{ij} \right)^{-1} I_{i<j} + \left(\sum_{i>j} \pi_{ij} \right)^{-1} I_{i>j} \right] \Big/ 2, & i \neq j \\ \pi_{ij}, & i = j \end{cases} \quad (6.27)$$

για $i, j = 1,...,R$ και $I_{i<j}$, $I_{i>j}$ είναι οι συναρτήσεις $I(i<j)$ και $I(i>j)$, αντίστοιχα.

Επομένως, ένα εναλλακτικό μέτρο της (TS), δίνεται από τον τύπο

$\tau_\phi^{T1}\left(\pi, \pi^T\right) = \tau_\phi^S\left(\psi, \psi^S\right)$, ο οποίος ισούται με

$$\tau_\phi^{T1}\left(\pi, \pi^T\right) = \frac{2}{\phi_0 \delta_\pi} I^C\left(\psi, \psi^S\right) = \frac{2}{\phi_0 \delta_\pi} \sum_{i<j} \frac{\psi_{ij} + \psi_{ji}}{2} \left\{ \phi\left(\frac{2\psi_{ij}}{\psi_{ij} + \psi_{ji}} \right) + \phi\left(\frac{2\psi_{ji}}{\psi_{ij} + \psi_{ji}} \right) \right\} \quad (6.28)$$

και σε όρους του διανύσματος π

$$\tau_\phi^{T1}\left(\pi, \pi^T\right) = \frac{2}{\phi_0 \delta_\pi} \sum_{i<j} \frac{1}{2} \left(\frac{\delta_\pi \pi_{ij}}{2\sum_{s<t} \pi_{st}} + \frac{\delta_\pi \pi_{ji}}{2\sum_{s>t} \pi_{st}} \right) \left\{ \phi\left(u_{ij}\right) + \phi\left(2 - u_{ij}\right) \right\},$$

όπου $u_{ij} = \dfrac{2\left(\sum_{s>t} \pi_{st} \right) \pi_{ij}}{\left(\sum_{s>t} \pi_{st} \right) \pi_{ij} + \left(\sum_{s<t} \pi_{st} \right) \pi_{ji}}$.

6.8 Μέτρα Συμμετρίας τύπου *φ – divergence* του *Cziszar*

Στην παράγραφο αυτή θα εξετάσουμε διάφορα γενικευμένα μέτρα συμμετρίας, που μετρούν τον βαθμό απομάκρυνσης από την συμμετρία (S), και την εκτεταμένη συμμετρία (QS) και (MH), ενός τετραγωνικού πίνακα συνάφειας με ονοματικές ή διατακτικές μεταβλητές ταξινόμησης. Τα προτεινόμενα μέτρα βασίζονται στην απόκλιση *φ-divergence* του *Cziszar* (1963) και προτάθηκαν από τους *Kateri & Papaioannou* [Technical Report (2007), University of

Pireaus (TR07 – 3)]. Τα μέτρα αυτά, παρέχουν μια ευρύτερη τάξη μέτρων απόκλισης, είναι πιο γενικά από όσα έχουμε προαναφέρει και με κατάλληλη επιλογή της κυρτής συνάρτησης φ_0, προκύπτουν τα προαναφερόμενα μέτρα συμμετρίας – ασυμμετρίας, που προτάθηκαν από τους *Tomizawa et al*. (1994, 2005). Ειδικότερα, η απόκλιση του *Csiszar*, $I^C\left(p_{ij},q_{ij}\right)=\sum_{i\neq j}p_{ij}\varphi\left(\frac{p_{ij}}{q_{ij}}\right)$

για $\varphi(u)=u\log u$ ισούται με την πληροφορία *Kullback-Leibler*

για $\varphi(u)=(1-u)^2$ ισούται με την απόκλιση χ^2 *– Pearson* και

για $\varphi(u)=\frac{u^{a+1}-u}{a(a+1)}$, $a\neq -1,0$ ισούται με την απόκλιση *Cressie & Read* (1984) [βλέπε *Kateri & Papaioannou*, (2007)].

6.8.1 Μέτρο απομάκρυνσης από την Συμμετρία (S)

Στα πλαίσια της συμμετρίας, η *φ-divergence* γίνεται $I_C(\pi,\pi^S)=\sum_{i\neq j}\pi_{ij}^S\phi\left(\pi_{ij}/\pi_{ij}^S\right)$

καθώς τα διαγώνια κελιά δεν συνεισφέρουν στην τιμή της $I_C(\pi,\pi^S)$, διότι $\phi(1)=0$. Η ποσότητα $\phi_0/2$ είναι το άνω φράγμα της *φ-divergence*, καθώς αποδεικνύεται ότι

$$I_C(\pi,\pi^S)=\sum_{i\neq j}\pi_{ij}^S\phi\left(\pi_{ij}/\pi_{ij}^S\right)\leq\phi_0\delta_\pi/2\leq\phi_0/2\,.$$

Επομένως, το γενικευμένο μέτρο συμμετρίας που βασίζεται στην *φ-divergence*, ορίζεται ως

$$\tau_\phi^S\left(\pi,\pi^S\right)=\frac{2}{\phi_0\delta_\pi}I_C(\pi,\pi^S)=\frac{2}{\phi_0\delta_\pi}\sum_{i\neq j}\pi_{ij}^S\phi\left(\pi_{ij}/\pi_{ij}^S\right)\qquad(6.29)$$

όπου $\delta_\pi=\sum_{i\neq j}\pi_{ij}$, $\pi_{ij}^S=\left(\pi_{ij}+\pi_{ji}\right)/2$ και $\varphi_0=\varphi(0)+\varphi(2)$.

Για ευκολία θα συμβολίζουμε το μέτρο $\tau_\phi^S\left(\pi,\pi^S\right)$ με τ_ϕ^S. Εύκολα αποδεικνύεται ότι η τιμή τ_ϕ^S κυμαίνεται μεταξύ $[0,1]$, με την γνωστή ερμηνεία των ορίων. Η τιμή 1 επιτυγχάνεται στην

ακραία περίπτωση για την οποία $\pi_{ij}=0$ ή $\pi_{ji}=0$ για όλα τα $i<j$. Το γενικευμένο μέτρο συμμετρίας μπορεί ισοδύναμα να παρουσιασθεί εξετάζοντας το διάνυσμα πιθανότητας $\tilde{\pi}=(\tilde{\pi}_{11},\tilde{\pi}_{12},...\tilde{\pi}_{ij},...\tilde{\pi}_{R.R-1})$ με $\tilde{\pi}_{ij}=\tilde{\pi}_{ij}/\delta_{\pi}$ για $i\neq j$, το οποίο δεν περιέχει τις διαγώνιες πιθανότητες. Είναι προφανές ότι $I_C(\tilde{\pi},\tilde{\pi}^S)=I_C(\pi,\pi^S)/\delta$ και ένας εναλλακτικός ορισμός είναι

$$\tau_{\phi}^{S}=\frac{2}{\phi_0}I_C(\tilde{\pi},\tilde{\pi}^S) \tag{6.30}$$

Η απόκλιση *power-divergence* των *Read & Cressie* (1988), προκύπτει από την *φ-divergence* μέσω της σχέσης $\phi(u,\lambda)=\left[\lambda(\lambda+1)\right]^{-1}\left[u^{\lambda+1}-u+\lambda(1-u)\right]$, για $\lambda\neq -1,0$, όπου $\phi(u,0)=\lim_{\lambda\to 0}\phi(u,\lambda)$. Επιπλέον, από την σχέση $\phi(u,-1)=\lim_{\lambda\to -1}\phi(u,\lambda)$, το γενικευμένο μέτρο συμμετρίας τ_{ϕ}^{S} ισούται με $\tau_{CR}^{S}=\frac{1}{\delta_{\pi}\left(2^{\lambda}-1\right)}\sum_{i\neq j}\pi_{ij}\left[\left(\pi_{ij}/\pi_{ij}^{S}\right)^{\lambda}-1\right]$, $\lambda>-1$, $\lambda\neq 0$. Το μέτρο αυτό προτάθηκε από τους *Tomizawa et al*. (1998) και παρουσιάστηκε στην παράγραφο 6.5.2.

6.8.2 Μέτρο απομάκρυνσης από την Ψευδοσυμμετρία (QS)

Είναι γνωστό ότι οι παράμετροι α_i $i=1,...,R$ του μοντέλου (QS), δεν μπορούν να δοθούν σε κλειστή μορφή, αλλά μπορούν να εκτιμηθούν μέσω επαναληπτικής διαδικασίας [*Kateri & Papaioannou*, (1997)]. Κατά συνέπεια, προκειμένου να έχουμε ασυμπτωματικά αποτελέσματα για το μέτρο της (QS), το οποίο βασίζεται στην απομάκρυνση της πιθανότητας π από αυτή της ψευδοσυμμετρίας π^{QS}, οι *Kateri & Papaioannou* (2007), εξέφρασαν το μοντέλο της (QS) μέσω των *odds ratios* διαδοχικών κατηγοριών [βλέπε *Agresti* (1990)]. Έτσι, το μοντέλο (QS) ορίζεται ως

$$r_{ij}=r_{ji} \text{ για } i,j=1,...,R-1 \tag{6.31}$$

ή ισοδύναμα

$$r_{ij}=r_{ji}^{S} \text{ για } i,j=1,...,R-1 \tag{6.32}$$

όπου $r_{ij}=\left(\pi_{ij}\pi_{i+1.j+1}\right)/\left(\pi_{i.j+1}\pi_{i+1.j}\right)$ και $r_{ji}^{S}=\left(r_{ij}+r_{ji}\right)/2$.

Από την παραπάνω σχέση, η υπόθεση της (QS) για $\pi = (\pi_{11}, \pi_{12}, ...\pi_{ij}, ...\pi_{RR})$ είναι ισοδύναμη με την υπόθεση της πλήρης συμμετρίας για το διάνυσμα πιθανότητας $\theta = (\theta_{11}, \theta_{12}, ...\theta_{ij}, ...\theta_{RR})$ όπου $\theta_{ij} = r_{ij} \Big/ \sum_{i,j} r_{ij}$, Κατά συνέπεια, ένα λογικό μέτρο της (QS), που βασίζεται στην *φ-divergence*, θα μπορούσε να ορισθεί ως ακολούθως

$$\tau_{\phi}^{QS}\left(\pi, \pi^{QS}\right) = \tau_{\phi}^{S}\left(\pi, \pi^{S}\right) = \frac{2}{\phi_0 \delta_{\theta}} \sum_{i \neq j} \theta_{ij}^{S} \phi\left(\theta_{ij} \big/ \theta_{ij}^{S}\right) \qquad (6.33)$$

όπου $\theta^S = (\theta_{11}^S, ...\theta_{ij}^S, ...\theta_{(R-1)(R-1)}^S)$, $\theta_{ij}^S = r_{ij}^S \Big/ \sum_{i,j} r_{ij} = \left(\theta_{ij} + \theta_{ji}\right) \Big/ 2$ και $\delta_{\theta} = 1 - \sum_{i=1}^{R-1} \theta_{ii}$.

6.8.3 Μέτρο απομάκρυνσης από την Περιθώρια Ομοιογένεια (MH)

Προκειμένου να υπολογίσουμε το μέτρο της Ομοιογένειας Περιθωρίου (MH) για την πιθανότητα π, χρειάζεται να ορίσουμε το διάνυσμα των περιθωρίων πιθανοτήτων γραμμής και στήλης. Έχουμε $\pi^R = (\pi_{1.}, \pi_{2.}, ..., \pi_{R.})$ και $\pi^C = (\pi_{.1}, \pi_{.2}, ..., \pi_{.R})$, αντίστοιχα. Για να υπολογίσουμε την απομάκρυνση από την (MH), χρησιμοποιούμε την *φ-divergence*, $I_C(\pi^R, \pi^{MH})$ και $I_C(\pi^C, \pi^{MH})$, όπου $\pi^{MH} = (\pi_1^{MH}, \pi_2^{MH}, ..., \pi_R^{MH})$ είναι το διάνυσμα πιθανότητας, με την ιδιότητα ότι αν $\pi^{MH} = \pi^R$ ή αν $\pi^{MH} = \pi^C$ τότε $\pi_{i.} = \pi_{.i}$ για $i = 1, ..., R$. Είναι προφανές ότι το διάνυσμα αυτό θα πρέπει να ισούται με $\pi^{MH} = \left(\pi_{i.} + \pi_{.i}\right) \Big/ 2$ για $i = 1, ..., R$. Το συμπέρασμα αυτό οδηγεί στον ακόλουθο ορισμό για το γενικευμένο μέτρο της (MH)

$$\tau_{\phi}^{MH}\left(\pi, \pi^{MH}\right) = \frac{1}{\phi_0}\left[I_C(\pi^R, \pi^{MH}) + I_C(\pi^C, \pi^{MH}) \right] \qquad (6.34)$$

Προφανώς, αυτό σημαίνει ότι οι *φ-divergences* $I_C(\pi^R, \pi^{MH})$ και $I_C(\pi^C, \pi^{MH})$ αναταξινομούνται (*rescaled*) κατάλληλα, ώστε να κυμαίνονται στο διάστημα $[0,1]$. Το μοντέλο (MH), σε όρους του διανύσματος πιθανότητας $\tilde{\pi}$, μπορεί να εκφραστεί ως

$$\tilde{\pi}_{i.} = \tilde{\pi}_{.i}, \text{ για } i = 1, ..., R \qquad (6.35)$$

όπου $\tilde{\pi}_{i.} = \left(\pi_{i.} - \pi_{ii}\right)/\delta_{\pi}$.

Το μέτρο τ_{ϕ}^{MH} μπορεί αναλόγως να ορισθεί με βάση την πιθανότητα $\tilde{\pi}$ και οδηγεί στον τύπο

$$\tau_{\phi}^{MH}\left(\tilde{\pi}, \tilde{\pi}^{MH}\right) = \frac{1}{\phi_0}\left[I_C(\tilde{\pi}^{R}, \tilde{\pi}^{MH}) + I_C(\tilde{\pi}^{C}, \tilde{\pi}^{MH}) \right] \qquad (6.36)$$

6.9 Συμπεράσματα

Στο Κεφάλαιο αυτό παρουσιάσαμε τα κυριότερα μοντέλα συμμετρίας – ασυμμετρίας για τετραγωνικούς πίνακες συνάφειας και υπολογίσαμε τα αντίστοιχα μέτρα συμμετρίας - ασυμμετρίας, που έχουν εμφανισθεί στην βιβλιογραφία απο όσο γνωρίζουμε, και τα οποία δείχνουν τον βαθμό απομάκρυνσης από την εκάστοτε δομή συμμετρίας (S), (QS) ,(MH), (DS), (CS) και (TS), ενός τετραγωνικού πίνακα συνάφειας. Συγκεκριμένα, υπολογίσαμε 20 μέτρα (16 απλά και 4 γενικευμένα) για ονοματικές μεταβλητές και 10 μέτρα (4 απλά και 6 γενικευμένα) για διατακτικές μεταβλητές. Τα μέτρα αυτά παρουσιάσθηκαν από τους *Tomizawa et al*. και βασίζονται στην πληροφορία *Kullback-Leibler*, στην απόκλιση $\chi^2 - Pearson$, στην εντροπία *Shannon*, στην απόκλιση *Gauss*, καθώς και στην απόκλιση τύπου *power-divergence* των *Cressie & Read* (1984, 1988) και στον δείκτη ανομοιότητας ή ποικιλότητας των *Patil & Taillie* (1982), στην περίπτωση γενίκευσης αυτών. Επιπλέον, παρουσιάσαμε 4 γενικευμένα μέτρα των *Kateri & Papaioannou*, τα οποία βασίζονται στην απόκλιση $\varphi - divergence$ των *Csiszar* και *Ali & Silvey* (1963). Τα μέτρα αυτά είναι ακόμη πιο γενικά, καθώς έχουν ενοποιήσει αρκετούς δείκτες πληροφορίας (όπως *Kullback-Leibler*, *Cressie & Read*, *Pearson*) και εμπεριέχουν τα μέτρα των *Tomizawa et al*. με κατάλληλη επιλογή της κυρτής συνάρτησης φ [βλέπε *σελ. 156*].

Συνοψίζοντας τα αποτελέσματα και εξετάζοντας τους παρακάτω συνοπτικούς Πίνακες 6-1 και 6-2 (*σελ. 161* και *162*), αντίστοιχα, παρατηρούμε ότι για κάθε μοντέλο ξεχωριστά, έχουν προταθεί διάφορα μέτρα, εκ των οποίων κάποια είναι ισοδύναμα, ενώ κάποια άλλα διαφοροποιούνται ελαφρώς, μολονότι υπολογίζονται για το ίδιο σύνολο δεδομένων. Συγκεκριμένα, παρατηρούμε ότι τα μέτρα που βασίζονται στην πληροφορία *Kullback-Leibler* ισοδυναμούν με αυτά που βασίζονται στην εντροπία *Shannon*, ενώ αυτά που βασίζονται στην

απόκλιση $\chi^2 - Pearson$ ισοδυναμούν με αυτά που βασίζονται στην απόκλιση *Gauss* και στον δείκτη συγκέντρωσης *Gini Concentration*. Τα μέτρα που βασίζονται στην πληροφορία *Kullback-Leibler*, διαφοροποιούνται ελαφρώς, σε σχέση με αυτά που βασίζονται στην απόκλιση $\chi^2 - Pearson$ και επομένως είναι σκόπιμο ο αναλυτής να υπολογίσει και τα δυο, πριν καταλήξει σε συμπεράσματα. Όπως ο *Tomizawa* αναφέρει, είναι δύσκολο να αποσαφηνιστεί πιο μέτρο είναι καταλληλότερο. Επομένως, μπορούμε να πούμε ότι από το σύνολο των 20 απλών μέτρων, ο αναλυτής μπορεί να περιοριστεί σε 8 μέτρα, ανάλογα με το μοντέλο που εξετάζει.

Τα γενικευμένα μέτρα, βασίζονται στην απόκλιση *power-divergence* των *Cressie & Read* (1984, 1988) και στην δείκτη ανομοιότητας ή ποικιλότητας *Patil & Taillie* (1982) και υπολογίζονται για διάφορες τιμές του $\lambda > -1$. Για συγκεκριμένες τιμές του λ, εμπεριέχουν τα απλά μέτρα. Ειδικότερα όπως παρατηρούμε, στην περίπτωση της *power-divergence*, για $\lambda = 0$ και $\lambda = 1$, προκύπτει η πληροφορία *Kulback-Leibler* και η απόκλιση $\chi^2 - Pearson$, αντίστοιχα, ενώ στην περίπτωση του *Patil & Taillie's diversity index*, για $\lambda = 0$ και $\lambda = 1$, προκύπτει η εντροπία *Shannon* και ο δείκτης *Gini Concentration*. Τα γενικευμένα μέτρα που βασίζονται στην *power-divergence* ισοδυναμούν με αυτά που βασίζονται στον *Patil & Taillie's diversity index* και επομένως μπορούμε να πούμε ότι από το σύνολο των 10 γενικευμένων μέτρων, ο αναλυτής μπορεί να περιοριστεί σε 7 μέτρα, ανάλογα με το μοντέλο που εξετάζει.

Αναλύοντας τα αποτελέσματα για το Παράδειγμα 5, συμπεραίνουμε ότι για την περίοδο Αύγουστος – Οκτώβριος 1971, οι τιμές των μέτρων τείνουν στο 0 και επομένως υπάρχει δομή (S) και δομή (MH) στον πίνακα συνάφειας, με την έννοια ότι οι απαντήσεις δεν διαφοροποιούνται, δηλαδή $\pi_{ij} = \pi_{ji}$ και $\pi_{i.} = \pi_{.i}$, δηλαδή όσοι απάντησαν «Ναι» στην πρώτη δημοσκόπηση και «Όχι» στην δεύτερη, ισούνται με όσους απάντησαν «Όχι» στην πρώτη δημοσκόπηση και «Ναι» στην δεύτερη, κοκ. Επίσης, παρατηρούμε ότι για την περίοδο Οκτώβριος 1971 – Δεκέμβριος 1973, οι τιμές των μέτρων απομακρύνονται αρκετά από το 0, και επομένως συμπεραίνουμε ότι υπάρχει απομάκρυνση από την (S) συμμετρία, περίπου 20% - 25% του μέγιστου βαθμού απομάκρυνσης και απομάκρυνση από την (MH), περίπου 15% - 20%

του μέγιστου βαθμού απομάκρυνσης. Με άλλα λόγια, οι απαντήσεις των ατόμων διαφοροποιούνται μεταξύ των δυο περιόδων.

ΠΙΝΑΚΑΣ 6-1

Συνοπτικός πίνακας μέτρων συμμετρίας – ασυμμετρίας για ονοματικές μεταβλητές

Μέτρα συμμετρίας - ασυμμετρίας για ονοματικές μεταβλητές

	Παράδειγμα 5		Παράδειγμα 5 α (Αυγ '71-Οκτ '71)	Παράδειγμα 5 β (Οκτ '71-Δεκ '73)
	Μέτρο *(S)*			
1	A1. Kullback-Leibler	Φ_S	0,035	0,207
2	A2. Pearson discrepancy	Ψ_S	0,048	0,268
3	B1. Shannon entropy	Φ_S	0,035	0,207
4	B2. Gauss discrepancy	Ψ_S	0,048	0,268
	Μέτρο *(MH)* ** Based on unconditional probabilities*			
5	A1. Kullback-Leibler	Φ_{MH}	0,003	0,030
6	A2. Pearson discrepancy	Ψ_{MH}	0,004	0,040
7	B1. Shannon entropy	Φ_{MH}	0,003	0,030
8	B2. Gini Concentration	Ψ_{MH}	0,004	0,040
9	Γ1. Weighted Kullback-Leibler	Φ_{MH}	0,003	0,030
10	Γ2. Weighted Pearson discrepancy	Ψ_{MH}	0,004	0,040
	Μέτρο *(MH)* ** Based on conditional probabilities*			
11	A1. Kullback-Leibler	Φ_{MH}	0,025	0,154
12	A2. Pearson discrepancy	Ψ_{MH}	0,034	0,202
13	B1. Shannon entropy	Φ_{MH}	0,025	0,154
14	B2. Gini Concentration	Ψ_{MH}	0,034	0,202
15	Γ1. Weighted Kullback-Leibler	Φ_{MH}	0,025	0,154
16	Γ2. Weighted Pearson discrepancy	Ψ_{MH}	0,034	0,202

			Παράδειγμα 5 α (Αυγ '71-Οκτ '71)	Παράδειγμα 5 β (Οκτ '71-Δεκ '73)
	Γενικευμένο *μέτρο* *(S)*			
17	A1. Cressie & Read power divergence	$\Phi_S^{(\lambda)}$		
18	B1. Patil & Taillie diversity index	$\Phi_S^{(\lambda)}$		
	για λ			
	-0,8		0,009	0,060
	-0,6		0,017	0,109
	-0,4		0,024	0,149
	0		0,035	0,207
	1		0,048	0,268
	1,4		0,049	0,273
	1,6		0,049	0,273
	2		0,048	0,268
	Γενικευμένο μέτρο *(MH)* ** Based on unconditional probabilities*			
19	A1. Cressie & Read power divergence	$\Phi_{MH}^{(\lambda)}$		
	για λ			
	-0,8		0,001	0,008
	-0,6		0,001	0,012
	-0,4		0,002	0,021
	0		0,003	0,030
	1		0,004	0,040
	1,4		0,004	0,041
	1,6		0,004	0,041
	2		0,004	0,040
	Γενικευμένο μέτρο *(MH)* ** Based on conditional probabilities*			
20	A1. Cressie & Read power divergence	$\Phi_{MH}^{(\lambda)}$		
	για λ			
	-0,8		0,007	0,044
	-0,6		0,010	0,067
	-0,4		0,017	0,110
	0		0,025	0,154
	1		0,034	0,202
	1,4		0,035	0,205
	1,6		0,035	0,205
	2		0,034	0,202

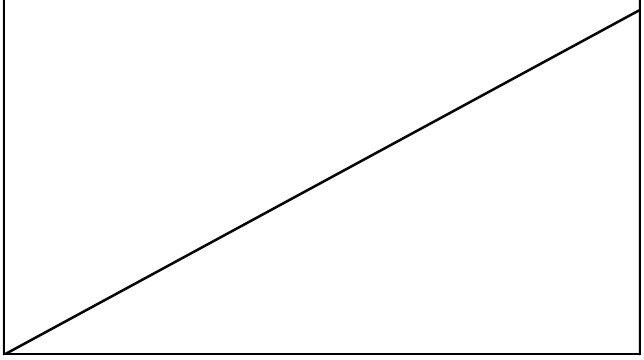
ΠΙΝΑΚΑΣ 6-2

Συνοπτικός πίνακας μέτρων συμμετρίας – ασυμμετρίας για ονοματικές μεταβλητές

Μέτρα συμμετρίας - ασυμμετρίας για διατακτικές μεταβλητές

	Παράδειγμα 6		1955	1965	1975
	Μέτρο *(S)*				
1	A1. Kullback-Leibler	Φ_S	0,087	0,177	0,159
2	A2. Pearson discrepancy	Ψ_S	0,118	0,235	0,212
3	B1. Shannon entropy	Φ_S	0,087	0,177	0,159
4	B2. Gini Concentration	Ψ_S	0,118	0,235	0,212

Γενικευμένο *μέτρο (S)*



Γενικευμένο μέτρο *(MH)*

** Based on unconditional probabilities*

5 A1. Cressie & Read power divergence $\Gamma_{MH}^{(\lambda)}$

για **λ**	1955	1965	1975
-0,8	0,047	0,104	0,103
-0,6	0,086	0,182	0,180
-0,4	0,118	0,241	0,237
0	0,165	0,319	0,312
1	0,216	0,392	0,382
1,4	0,221	0,397	0,387
1,6	0,220	0,397	0,386
2	0,216	0,392	0,382

Γενικευμένο μέτρο *(MH)*

** Based on conditional probabilities*

6 A1. Cressie & Read power divergence $\Gamma_{MH}^{(\lambda)}$

για **λ**	1955	1965	1975
-0,8	0,047	0,104	0,103
-0,6	0,086	0,182	0,180
-0,4	0,118	0,241	0,237
0	0,165	0,319	0,312
1	0,216	0,392	0,382
1,4	0,221	0,397	0,387
1,6	0,220	0,397	0,386
2	0,216	0,392	0,382

Γενικευμένο μέτρο *(DS)*

7 A1. Cressie & Read power divergence $\Phi_{DS}^{(\lambda)}$

8 A3. Patil & Taillie diversity index $\Phi_{DS}^{(\lambda)}$

για **λ**	1955	1965	1975
-0,8	0,036	0,085	0,065
-0,6	0,065	0,149	0,115
-0,4	0,088	0,197	0,155
0	0,122	0,262	0,209
1	0,158	0,323	0,264
1,4	0,160	0,327	0,268
1,6	0,160	0,327	0,268
2	0,158	0,323	0,264

Γενικευμένο μέτρο *(CS)*

9 A1. Cressie & Read power divergence $\Phi_{CS}^{(\lambda)}$

10 A3. Patil & Taillie diversity index $\Phi_{CS}^{(\lambda)}$

για **λ**	1955	1965	1975
-0,8	0,183	0,096	0,077
-0,6	0,075	0,169	0,138
-0,4	0,102	0,225	0,186
0	0,141	0,299	0,254
1	0,183	0,370	0,322
1,4	0,186	0,375	0,327
1,6	0,186	0,374	0,327
2	0,183	0,370	0,322

6.10 Ανακεφαλαίωση – Γενική Σύνοψη

Η παρούσα εργασία αποτελεί μια ανασκόπηση και μελέτη των μέτρων συνάφειας και μέτρων συμμετρίας – ασυμμετρίας, για πίνακες συνάφειας με ονοματικές ή διατακτικές μεταβλητές. Συγκεκριμένα, παρουσιάσαμε και υπολογίσαμε 18 μέτρα συνάφειας ονοματικής ταξινόμησης, 7 μέτρα συνάφειας διατακτικής ταξινόμησης και 34 μέτρα συμμετρίας – ασυμμετρίας. Είναι σημαντικό, να επιλέγουμε το κατάλληλο μέτρο για να εκτιμήσουμε τον βαθμό απομάκρυνσης από την ανεξαρτησία ή την συμμετρία, ανάλογα με το είδος των δεδομένων που αναλύουμε. Στα Κεφάλαια 4 και 5, αναφερθήκαμε στα κυριότερα μέτρα συνάφειας και περιγράψαμε τις ιδιότητές τους. Επιπλέον, με την εφαρμογή αυτών σε επιλεγμένα παραδείγματα, καταμετρήσαμε τις δυνατές περιπτώσεις που μπορεί να αντιμετωπίσει ένας αναλυτής και παρουσιάσαμε τα αποτελέσματα, προσδίδοντας την ερμηνεία των μέτρων, καθώς και βασικά συμπεράσματα. Οι τιμές των μέτρων δεν είναι πάντοτε ισοδύναμες και επομένως κρίνεται σκόπιμο ο αναλυτής, να μην περιορίζεται στην τιμή ενός μόνο μέτρου, αλλά να υπολογίζει τις τιμές διάφορων μέτρων, κατάλληλων για την εκάστοτε έρευνα, προτού καταλήξει σε συμπεράσματα. Επίσης, η χρήση των μέτρων για την μελέτη και σύγκριση του βαθμού συνάφειας, μεταξύ διαφορετικών πινάκων, θα πρέπει να γίνεται με ιδιαίτερη προσοχή, καθώς πολλά μέτρα συνάφειας εξαρτώνται από το μέγεθος και την διάσταση του πίνακα.

Στο Κεφάλαιο 6, παρουσιάσαμε την πρόσφατη ανάπτυξη των μέτρων συμμετρίας – ασυμμετρίας για τετραγωνικούς πίνακες συνάφειας, τα οποία βασίζονται σε γνωστά μέτρα στατιστικής πληροφορίας, τύπου απόκλισης (*divergence-type measures of Information*) και τύπου εντροπίας (*entropy-type measures of Information*). Τα μέτρα αυτά έχουν προταθεί από τους *Tomizawa et al*. (1994 – 2005) και από τους *Kateri & Papaioannou* (2007) και δείχνουν το βαθμό απομάκρυνσης από διάφορα μοντέλα συμμετρίας – ασυμμετρίας. Αρχικά, οι *Tomizawa et al*. πρότειναν μέτρα που βασίζονται στην πληροφορία *Kullback-Leibler*, στην απόσταση χ^2 *– Pearson* και στην εντροπία *Shannon*. Στην συνέχεια, πρότειναν γενικευμένα μέτρα, που για ειδικές περιπτώσεις τους, εμπεριέχουν τα προγενέστερα, χρησιμοποιώντας τον δείκτη ανομοιότητας ή ποικιλότητας των *Patil & Taillie*, καθώς επίσης χρησιμοποιώντας την απόκλιση

των *Cressie & Read*, η οποία αποτελεί μια γενίκευση των μέτρων πληροφορίας τύπου απόκλισης και ανήκει στην οικογένεια των *power – divergences*.

Οι *Kateri & Papaioannou* (2007), γενίκευσαν περαιτέρω τα μέτρα αυτά, χρησιμοποιώντας την απόκλιση *Csiszar*, η οποία είναι επίσης ένα γενικό μέτρο πληροφορίας τύπου απόκλισης και ανήκει στην οικογένεια των $\varphi - divergences$ καθώς βασίζεται σε μια κυρτή συνάρτηση φ. Επιπλέον, πρότειναν κάποια νέα μέτρα, όπως αυτό της τριγωνικής συμμετρίας (TS). Η απόκλιση *Csiszar* παρέχει μια ευρύτερη τάξη μέτρων απόκλισης και εμπεριέχει όλα τα μέτρα των *Tomizawa et al.*, με κατάλληλη επιλογή της κυρτής συνάρτησης φ (βλέπε *σελ. 156*). Επιπλέον, εμπεριέχει και άλλες σημαντικές αποκλίσεις, οι οποίες δεν καλύπτονται από την απόκλιση *Cressie & Read*.

Στο σημείο αυτό, αξίζει να αναφέρουμε μια από τις πιο πρόσφατα προταθείσες αποκλίσεις [*Basu et al*. (1998)], η οποία ανήκει στην οικογένεια των *power – divergences* και ορίζεται ως $I^{BHHJ}\left(p_{ij}, q_{ij}\right) = \sum_i \sum_j \left(q_{ij}^{1+a} - \left(1+\frac{1}{a}\right) p_{ij} q_{ij}^{a} + \frac{1}{a} p_{ij}^{1+a} \right)$, $a > 0$. Η απόκλιση I^{BHHJ} έχει εξετασθεί και γενικευθεί από τους *Mattheou & Karagrigoriou* (2009) και αποτελεί μια πολύ ενδιαφέρουσα περίπτωση για περαιτέρω έρευνα στο μέλλον, αναφορικά με την ανάπτυξη νέων μέτρων συμμετρίας – ασυμμετρίας, που ίσως να έχουν καλύτερες ιδιότητες. Επιπλέον, σημειώνουμε ότι σε αντίθεση με τα περισσότερα μέτρα συνάφειας, τα μέτρα συμμετρίας – ασυμμετρίας δεν υποστηρίζονται από γνωστά στατιστικά πακέτα. Ο υπολογισμός των μέτρων έχει πραγματοποιηθεί με την χρήση του υπολογιστικού πακέτου (*Excel*). Η ανάπτυξη αλγορίθμων για την ταχύτερη και ακριβέστερη μέθοδο υπολογισμού τους, είναι μια επίσης ενδιαφέρουσα ιδέα, που θα πραγματοποιηθεί στο μέλλον.

ΠΑΡΑΡΤΗΜΑ Α

Παράδειγμα 1

Ένας ερευνητής θέλει να προσδιορίσει αν υπάρχει σχέση μεταξύ της αγωγής που λαμβάνει ένας ασθενής, μεταβλητή X και του αποτελέσματος μεταβλητή Y, δηλαδή «Επιβίωση» ή όχι. 1000 ασθενείς τυχαιοποιήθηκαν, λαμβάνοντας την Θεραπεία Α ή την Θεραπεία Β (ψευδοφάρμακο) και καταγράφηκαν τα αποτελέσματα. Ο ακόλουθος 2×2 πίνακας συνάφειας συνοψίζει κατάλληλα τα αποτελέσματα.

Παράδειγμα 1

		ΕΠΙΒΙΩΣΗ		
		ΟΧΙ	ΝΑΙ	**Σύνολο**
ΑΓΩΓΗ	Θεραπεία Α	80 0,08	420 0,42	**500** **0,5**
	Θεραπεία Β	100 0,1	400 0,4	**500** **0,5**
	Σύνολο	**180** **0,18**	**820** **0,82**	**1000** **1**

Παράδειγμα 2

Ένας ερευνητής θέλει να προσδιορίσει, αν υπάρχει σχέση μεταξύ του χρώματος μαλλιών, μεταβλητή X, και του χρώματος ματιών Y. Ένα τυχαίο δείγμα 6800 ανδρών, ταξινομήθηκε σε μια από τις 3 κατηγορίες της μεταβλητής X και σε μια από τις 4 κατηγορίες της μεταβλητής Y. Ο ακόλουθος 3×4 πίνακας συνάφειας συνοψίζει κατάλληλα τα αποτελέσματα.

Παράδειγμα 2

		ΧΡΩΜΑ ΜΑΛΛΙΩΝ				
		Ξανθά	Καστανά	Μαύρα	Κόκκινα	**Σύνολο**
ΧΡΩΜΑ ΜΑΤΙΩΝ	Μπλε	1768 0,26	807 0,12	189 0,03	47 0,01	**2811** **0,41**
	Πράσινα ή Γκρι	946 0,14	1387 0,20	746 0,11	53 0,01	**3132** **0,46**
	Καστανά	115 0,02	438 0,06	288 0,04	16 0,002	**857** **0,13**
	Σύνολο	**2829** **0,42**	**2632** **0,39**	**1223** **0,18**	**116** **0,02**	**6800** **1,00**

Goodman - Kruskal (1954)

Παράδειγμα 2α

		ΧΡΩΜΑ ΜΑΛΛΙΩΝ				
		Ξανθά	Καστανά	Μαύρα	Κόκκινα	**Σύνολο**
ΧΡΩΜΑ ΜΑΤΙΩΝ	Μπλε	1970 0,29	899 0,10	211 0,02	52 0,01	**3132** **0,33**
	Πράσινα ή Γκρι	946 0,10	1387 0,15	746 0,08	53 0,01	**3132** **0,33**
	Καστανά	420 0,04	1601 0,17	1053 0,11	58 0,006	**3132** **0,33**
	Σύνολο	**3336** **0,36**	**3887** **0,41**	**2009** **0,21**	**164** **0,02**	**9396** **1,00**

Goodman - Kruskal (1954)

Παράδειγμα 3

Ένας ερευνητής θέλει να προσδιορίσει αν υπάρχει σχέση μεταξύ του επιπέδου εκπαίδευσης, μεταβλητή X και του ύψους εισοδήματος, μεταβλητή Y. Ένα τυχαίο δείγμα 1175 παρατηρήσεων, ταξινομήθηκε σε μια από τις 3 κατηγορίες της μεταβλητής X και σε μια από τις 3 κατηγορίες της μεταβλητής Y. Ο ακόλουθος 3×3 πίνακας συνάφειας συνοψίζει κατάλληλα τα αποτελέσματα.

Παράδειγμα 3

		ΕΠΙΠΕΔΟ ΕΙΣΟΔΗΜΑΤΟΣ			
		Χαμηλό	Μεσαίο	Υψηλό	**Σύνολο**
ΕΠΙΠΕΔΟ ΕΚΠΑΙΔΕΥΣΗΣ	Βασική	302 0,17	105 0,06	23 0,01	**430** **0,24**
	Πρωτοβάθμια	409 0,23	331 0,19	250 0,14	**990** **0,56**
	Δευτεροβάθμια	15 0,01	155 0,09	185 0,10	**355** **0,20**
	Σύνολο	**726** **0,41**	**591** **0,33**	**458** **0,26**	**1775** **1,00**

Παράδειγμα 4

Ένας ερευνητής θέλει να προσδιορίσει αν υπάρχει σχέση μεταξύ του βάρους ενός ατόμου, μεταβλητή X και της σειράς κατά την οποία γεννήθηκε, μεταβλητή Y. Ένα τυχαίο δείγμα 300 παρατηρήσεων, ταξινομήθηκε σε μια από τις 3 κατηγορίες της μεταβλητής X και σε μια από τις 4 κατηγορίες της μεταβλητής Y. Ο ακόλουθος 3×4 πίνακας συνάφειας συνοψίζει κατάλληλα τα αποτελέσματα.

Παράδειγμα 4

		ΣΕΙΡΑ ΓΕΝΝΗΣΗΣ				
		Πρωτότοκος	Δευτερότοκος	Τριτότοκος	Τεταρτότοκος +	**Σύνολο**
ΒΑΡΟΣ	Κάτω του Μ.Ο.	70 0,23	15 0,05	10 0,03	5 0,02	**100** **0,33**
	Μέσος όρος	10 0,03	60 0,20	20 0,07	10 0,03	**100** **0,33**
	Άνω του Μ.Ο.	10 0,03	15 0,05	35 0,12	40 0,133	**100** **0,33**
	Σύνολο	**90** **0,30**	**90** **0,30**	**65** **0,22**	**55** **0,18**	**300** **1,00**

Handbook of parametric and nonparametric statistical procedures

Παράδειγμα 5

Τα παραδείγματα 5α και 5β αφορούν τρεις συνεχόμενες δημοσκοπήσεις στην χώρα της Δανίας, κατά την περίοδο Αύγουστος 1971, Οκτώβριος 1971 και Δεκέμβριος 1973, αναφορικά με την απόφαση για το αν θα πρέπει η χώρα να αποτελεί μέλος της κοινής Ευρωπαϊκής αγοράς. Το δείγμα αποτελείται από 493 πολίτες, οι οποίοι ταξινομήθηκαν με βάση την απάντησή τους, «Ναι», «Όχι», «Δεν ξέρω». Ο ακόλουθος τετραγωνικός 3×3 πίνακας συνάφειας συνοψίζει κατάλληλα τα αποτελέσματα, προκειμένου να αποφασίσουμε, αν υπάρχει στροφή της κοινής γνώμης κατά το πέρασμα του χρόνου.

Παράδειγμα 5α **Οκτώβριος 1971**

		Ναι	Όχι	Αναποφάσιστοι	Σύνολο
Αύγουστος 1971	Ναι	176 0,36	33 0,07	40 0,08	249 0,51
	Όχι	21 0,04	94 0,19	32 0,06	147 0,30
	Αναποφάσιστοι	21 0,04	33 0,07	43 0,09	97 0,20
	Συνολο	218 0,44	160 0,32	115 0,23	493 1,00

Source: Andersen (1980) [Tomizawa et al. (1994)]

Παράδειγμα 5b **Δεκέμβριος 1973**

		Ναι	Όχι	Αναποφάσιστοι	Σύνολο
Οκτώβριος 1971	Ναι	167 0,34	36 0,07	15 0,03	218 0,44
	Όχι	19 0,04	131 0,27	10 0,02	160 0,32
	Αναποφάσιστοι	45 0,09	50 0,10	20 0,04	115 0,23
	Συνολο	231 0,47	217 0,44	45 0,09	493 1,00

Source: Andersen (1980) [Tomizawa et al. (1994)]

Παράδειγμα 6

Το παράδειγμα αφορά την μελέτη του το επαγγελματικού status πατέρα και υιού, τα έτη 1955, 1965 και 1975, στην Ιαπωνία, με δείγμα 1886, 1925 και 2238 παρατηρήσεων, αντίστοιχα. Οι κατηγορίες των μεταβλητών X (επάγγελμα πατέρα) και Y (επάγγελμα υιού), ορίζονται ως

1: Professional, 2: Manager, 3: Clerical, 4: Sales, 5: Skilled manual, 6: Semiskilled manual, 7: Unskilled manual, 8: Farmer. Οι ακόλουθοι 8×8 πίνακες συνάφειας, συνοψίζουν κατάλληλα τα αποτελέσματα.

Παράδειγμα 6α (1955)

Father's status	**Son's status** 1	2	3	4	5	6	7	8	**Σύνολο**
1	36 *0,02*	4 *0,00*	14 *0,01*	7 *0,00*	8 *0,00*	2 *0,00*	3 *0,00*	8 *0,00*	**82** ***0,04***
2	20 *0,01*	20 *0,01*	27 *0,01*	24 *0,01*	11 *0,01*	11 *0,01*	2 *0,00*	11 *0,01*	**126** ***0,07***
3	9 *0,00*	6 *0,00*	23 *0,01*	12 *0,01*	9 *0,00*	5 *0,00*	3 *0,00*	16 *0,01*	**83** ***0,04***
4	15 *0,01*	14 *0,01*	39 *0,02*	81 *0,04*	17 *0,01*	16 *0,01*	11 *0,01*	15 *0,01*	**208** ***0,11***
5	6 *0,00*	7 *0,00*	22 *0,01*	13 *0,01*	72 *0,04*	20 *0,01*	6 *0,00*	13 *0,01*	**159** ***0,09***
6	3 *0,00*	2 *0,00*	5 *0,00*	12 *0,01*	18 *0,01*	19 *0,01*	9 *0,00*	7 *0,00*	**75** ***0,04***
7	5 *0,00*	3 *0,00*	10 *0,01*	11 *0,01*	21 *0,01*	15 *0,01*	38 *0,02*	25 *0,01*	**128** ***0,07***
8	39 *0,02*	30 *0,02*	76 *0,04*	80 *0,04*	69 *0,04*	52 *0,03*	45 *0,02*	614 *0,33*	**1005** ***0,54***
Σύνολο	**133** ***0,07***	**86** ***0,05***	**216** ***0,12***	**240** ***0,13***	**225** ***0,12***	**140** ***0,08***	**117** ***0,06***	**709** ***0,38***	**1866** ***1,00***

Occupational status (exmined in 1955) for Japanese father-son pairs, Tomizawa (1979)

Παράδειγμα 6β (1965)

Father's status	Son's status 1	2	3	4	5	6	7	8	Σύνολο
1	**27**	10	16	3	6	6	1	2	**71**
	0,01	*0,01*	*0,01*	*0,00*	*0,00*	*0,00*	*0,00*	*0,00*	***0,04***
2	15	**38**	30	20	8	4	3	7	**125**
	0,01	***0,02***	*0,02*	*0,01*	*0,00*	*0,00*	*0,00*	*0,00*	***0,06***
3	13	17	**32**	17	7	16	6	5	**113**
	0,01	*0,01*	***0,02***	*0,01*	*0,00*	*0,01*	*0,00*	*0,00*	***0,06***
4	12	36	40	**132**	22	30	13	6	**291**
	0,01	*0,02*	*0,02*	***0,07***	*0,01*	*0,02*	*0,01*	*0,00*	***0,15***
5	8	22	38	41	**91**	42	22	9	**273**
	0,00	*0,01*	*0,02*	*0,02*	***0,05***	*0,02*	*0,01*	*0,00*	***0,14***
6	2	2	7	12	13	**16**	3	2	**57**
	0,00	*0,00*	*0,00*	*0,01*	*0,01*	***0,01***	*0,00*	*0,00*	***0,03***
7	3	2	11	11	13	26	**30**	6	**102**
	0,00	*0,00*	*0,01*	*0,01*	*0,01*	*0,01*	***0,02***	*0,00*	***0,05***
8	38	44	95	101	132	114	60	**309**	**893**
	0,02	*0,02*	*0,05*	*0,05*	*0,07*	*0,06*	*0,03*	***0,16***	***0,46***
Σύνολο	**118**	**171**	**269**	**337**	**292**	**254**	**138**	**346**	**1925**
	0,06	***0,09***	***0,14***	***0,18***	***0,15***	***0,13***	***0,07***	***0,18***	***1,00***

Occupational status (exmined in 1965) for Japanese father-son pairs, Tomizawa (1979)

Παράδειγμα 6γ (1975)

Father's status	Son's status 1	2	3	4	5	6	7	8	Σύνολο
1	**44**	18	28	8	6	8	1	5	**118**
	0,02	*0,01*	*0,01*	*0,00*	*0,00*	*0,00*	*0,00*	*0,00*	***0,05***
2	15	**50**	45	20	18	17	4	7	**176**
	0,01	***0,02***	*0,02*	*0,01*	*0,01*	*0,01*	*0,00*	*0,00*	***0,08***
3	18	25	**47**	30	24	18	5	7	**174**
	0,01	*0,01*	***0,02***	*0,01*	*0,01*	*0,01*	*0,00*	*0,00*	***0,07***
4	16	27	53	**77**	40	29	9	6	**257**
	0,01	*0,01*	*0,02*	***0,03***	*0,02*	*0,01*	*0,00*	*0,00*	***0,11***
5	18	25	42	31	**122**	43	17	13	**311**
	0,01	*0,01*	*0,02*	*0,01*	***0,05***	*0,02*	*0,01*	*0,01*	***0,13***
6	12	15	21	15	36	**33**	3	8	**143**
	0,01	*0,01*	*0,01*	*0,01*	*0,02*	***0,01***	*0,00*	*0,00*	***0,06***
7	3	5	8	7	26	21	**9**	3	**82**
	0,00	*0,00*	*0,00*	*0,00*	*0,01*	*0,01*	***0,00***	*0,00*	***0,04***
8	44	65	114	92	184	195	58	**325**	**1077**
	0,02	*0,03*	*0,05*	*0,04*	*0,08*	*0,08*	*0,02*	***0,14***	***0,46***
Σύνολο	**170**	**230**	**358**	**280**	**456**	**364**	**106**	**374**	**2338**
	0,07	***0,10***	***0,15***	***0,12***	***0,20***	***0,16***	***0,05***	***0,16***	***1,00***

Occupational status (exmined in 1975) for Japanese father-son pairs, Tomizawa (1979)

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΕΛΛΗΝΙΚΗ

[1] Παπαιωάννου Τ. & Φερεντίνος Κ. (2004), Ιατρική Στατιστική και Στοιχεία Βιομαθηματικών, Β' Έκδοση, *Εκδόσεις Σταμούλη*.

[2] Κατέρη Μ. (2006), Βιοστατιστική και Στατιστικές Μέθοδοι στην Επιδημιολογία, *Σημειώσεις Μαθήματος Πανεπιστημίου Πειραιώς*.

[3] Κατέρη Μ. (2006), Ανάλυση Διακριτών Δεδομένων, *Σημειώσεις Μαθήματος Πανεπιστημίου Πειραιώς*.

ΞΕΝΗ

[1] Adi Raveh (1986), On Measures of Monotone Association, *The American Statistician*, Vol.40, No. 2, pp. 117-123.

[2] Agresti A. (1983), A Simple Diagonals-Parameter Symmetry and Quasi-Symmetry Model, *Statistics and Probability Letters*, Vol. 1, Issue 6, pp. 313-316.

[3] Agresti A. (1984), Analysis of Ordinal Categorical Data, New York: John Wiley & Sons.

[4] Agresti A. (1990), Categorical Data Analysis, New York: John Wiley & Sons.

[5] Agresti A. (2002), Categorical Data Analysis, 2nd ed., New York: John Wiley & Sons.

[6] Bassu A., Harris I., Hjort N., Jones M. (1998), Robust and efficient estimation by minimising a density power divergence, *Biometrica*, Vol. 85, No. 1, pp. 549-559.

[7] Bhattacharya B. (1998), Testing Conditional Symmetry Against One-Sided Alternatives in Square Contingency Tables, *Metrica*, Vol. 47, No. 1, pp. 71-84.

[8] Bishop Y., Fienberg S., Holland P., (1975), Discrete Multivariate Testing Analysis: Theory and Practice, Cambridge: Massachusetts Institute of Technology.

[9] Bowker A. (1948), A test for symmetry in contingency tables, *Journal of American Statistical Association*, Vol. 43, pp. 572-574.

[10] Burbea J. & Rao C. (1982), On the convexity of some divergence measures based on entropy functions, *IEEE Transactions on Information Theory*, Vol. 28, pp. 489-495.

[11] Carroll J. (1961), The nature of data, or how to choose a correlation coefficient, *Psychometrika*, Vol. 26, pp. 347-372.

[12] Cressie N, & Read T. (1984), Multinomial goodness-of-fit-test, *Journal of the Royal Statistical Society, Series* B, Vol. 46, pp. 440-464.

[13] Caussinus H., (1965), Contribution a l' analyse statistique de tableaux de correlation, *Annales de la Faculte des Sciences de Toulouse*, Vol. 29, pp. 77-182.

[14] Cziszar I. (1963), Eine informationstheoretische Ungleichung und ihre Anwendungen auf den Beweis der Ergozitat von Markoffschen Ketten, *A Mayar Tudomanyos Academia Mathematikai Kutato Intezelent Kozlemezyri*, Vol. 8, pp. 85-108.

[15] Cohen J. (1977), Statistical power analysis for the behavioural scienses, New York: Academic Press.

[16] Conover W. (1980), Practical nonparametric statistics (2nd ed.), New York: John Wiley & Sons.

[17] Conover W. (1999), Practical nonparametric statistics (3rd ed.), New York: John Wiley & Sons.

[18] Costner H. (1965), Criteria for Measures of Association, *American Sociological Review*, Vol. 30, No. 3 pp. 341-353.

[19] Daniel W. (1990), Applied nonparametric statistics (2nd ed.), Boston: PWS-Kent Publishing Company.

[20] Darlington R. (2001), Measures of Association for Crosstab tables, *Cornell University*, New York.

[21] Davis J. (1971), Elementary survey analysis, Englewood Cliffs N., Prentice-Hall.

[22] Ferentinos K. & Papaioannou T. (1981), New Parametric Measures of Information, *Information and Controls*, Vol. 51, Issue 3, pp. 193-208.

[23] Fienberg S. & Rinaldo A. (2007), Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation, *Journal of Statistical Planning and Inference*, Vol. 137, Issue 11, pp. 3430-3445.

[24] Fisher R. (1925), Theory of statistical estimation, *Proc. Cambridge Philos. Soc.*, Vol. 22, pp. 700-725.

[25] Fleiss J. (1981), Statistical methods for rates and proportions (2nd ed.), New York: John Wiley & Sons.

[26] Goodman L. (1979), Multiplicative models for square contingency tables with ordered categories, *Biometrika*, Vol. 66, pp. 413-418.

[27] Goodman L. (1981a), Association models and canonical correlation in the analysis of cross-classifications having ordered categories, *Journal of the American Statistical Association*, Vol. 74, pp. 320-334.

[28] Goodman L. (1981b), Association models and the bivariate normal for contingency tables with ordered categories, *Biometrika*, Vol. 68, pp. 347-355.

[29] Goodman L. (2000), The analysis of cross-classified data: Notes on a century of progress in contingency table analysis and some comments on its prehistory and its future, *Statistics for*

[30] Goodman L. & Kruskal W. (1954), Measures of Association for Cross Classifications, *Journal of the American Statistical Association*, Vol. 49, No. 268, pp. 732-764.

[31] Goodman L. & Kruskal W. (1958), Ordinal Measures of Association for Cross Classifications, *Journal of the American Statistical Association*, Vol. 53, No. 284, pp. 814-861.

[32] Guilford J. (1965), Fundamental statistics in psychology and education (4^{th} ed.), New York: McGraw-Hill Book Company.

[33] Haberman S. (1982), Analysis of dispersion of multinominal responses, *Journal of the American Statistical Association*, Vol. 77, pp. 568-580.

[34] Hershberger S. & Fisher D. (2005), Measures of Association, *Encyclopedia of Statistics in Behavioral Science*, Vol. 3, pp. 1183-1192.

[35] Hernes G., (1970), A Markovian approach to measures of association, *Eamerican Journal of Sociology*, Vol. 75, pp. 997-1011.

[36] Heyde C. & Seneta E. (1977), I. J. Bienayme: Statistical Theory Anticipated, New York: Springer-Verlag.

[37] Hunter A. (1973), Validity of Measures of Association: The Nominal-Nominal, Two-by-Two case, *The American Journal of Sociology*, Vol. 79, No. 1, pp. 99-109.

[38] Joe H. (1989), Relative Entropy Measures of Multivariate Dependence, *Journal of the American Statistical Association*, Vol. 84, No. 405, pp. 157-164.

[39] Karagrigoriou A. & Papaioannou T. (2008), On Measures of Information and Divergence and Model Selection Criteria, *Statistics for Industry and Technology*, pp. 503-518, Birkhauser Boston.

[40] Kateri M. & Agresti A. (2006), A Class of Ordinal Quasi-Symmetry Models for Square Contingency Tables, *Statistics and Probability Letters*, Vol. 77, pp. 598-603.

[41] Kateri M. & Papaioannou T. (1996), Symmetry and Asymmetry models for Rectangular Tables, *Biometrical Journal*, Vol. 38, Issue 2, pp. 203-220.

[42] Kateri M. & Papaioannou T. (1997), Asymmetry Models for Contingency Tables, *Journal of the American Statistical Association*, Vol. 92, No. 439, pp. 1124-1131.

[43] Kateri M. & Papaioannou T. (2007), Measures of Symmetry-Asymmetry for Square Contingency Tables, [*Technical Report, University of Pireaus (TR07 – 3)*].

[44] Khamis H. (2008), Measures of Association: How to choose?, *Journal of Diagnostic Medical Sonography*, Vol. 24, No. 3, pp. 155-162.

[45] Kimeldorf G. & Sampson A. (1989), A framework for positive dependence, *Ann. Inst. Statist. Math*, Vol. 41, pp. 31-45.

[46] Krampe A. - Kateri M. - Kuhnt. S (2009), Asymmetry Models for Square Contingency Tables: exact tests via algebraic statistics, *Statistics and Computing*, DOI 10.1007/s11222-009-9146-7.

[47] Kullback S. & Leibler R. (1951), On Information and Sufficiency, *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp. 79-86.

[48] Liebetrau A. (1983), Measures of Association, *SAGE Publications, Inc*.

[49] Linhart H. & Zucchini W. (1986), Model Selection, New York: John Wiley.

[50] Liu R. (1980), A note on phi coefficient comparison, *Research in Higher Education*, Vol. 13, No. 1, pp. 3-8.

[51] Mattheou K. & Karagrigoriou A. (2009), A new family of divergence measures for test of fit, *Journal of Statistics*, Australian, New Zealand, Vol. 52, Issue 2, pp. 187-200.

[52] McCullagh P. (1978), A class of parametric models for the analysis of square contingency tables with ordered categories, *Biometrika*, Vol. 65, pp. 413-418.

[53] McCullagh P. (1982), Some Applications of Quasi Symmetry, *Biometrika*, Vol. 69, Issue 2, pp. 303-308.

[54] Menendez M., Pardo J., Pardo L. (2001), Tests based φ-divergence for Bivariate Symmetry, *Metrika*, Vol. 53, pp. 15-19.

[55] Menendez M. & Pardo J. (2004), Tests of Symmetry in Three-dimensional Contingency Tables Based on Phi-divergence statistics, *Journal of Applied Statistics*, Vol. 31, No. 9, pp. 1095-1114.

[56] Menendez M., Pardo J., Pardo L., Zografos K. (2005), On tests of Symmetry, Marginal Homogeneity and Quasi-Symmetry in two-ways Contingency Tables based on minimum φ-divergence Estimator with constraints, *Journal of Statistical Computation and Simulation*, Vol. 75, No. 7, pp. 555-580.

[57] Micheas A. & Zografos K. (2006), Measuring Stochastic Dependence Using φ-divergence, *Journal of Multivariate Analysis*, Vol. 97, Issue 3, pp. 765-784.

[58] Olszak M. & Ritschard G., (1995), The Behaviour of Nominal and Ordinal Partial Association Measures, *The Statistician*, Vol. 44, No. 2, pp. 195-212.

[59] Ott R., Larson R., Rexroat C., Mendenhall W. (1992), Statistics: A tool for the social sciences (5^{th} ed.), Boston: PWS-Kent Publishing Company.

[60] Pagano M. & Gauvreau K. (1993), Principles of Biostatistics, Belmont, CA: Duxbury Press.

[61] Papaioannou T. (1985), Measures of Information, *Encyclopedia of Statistical Sciences (Eds., Kotz S. & Johnson N.)*, Vol. 5, pp. 391-397, John Wiley & Sons, New York.

[62] Papaioannou T. & Ferentinos K. (2005), On Two Forms of Fisher's Measure of Information, *Communications in Statistics* - Theory and Methods, Vol. 34, pp. 1461-1470.

[63] Pardo L. & Menendez M. (2006), Phi-Divergence-Type Test for Positive Dependence Alternatives in $2 \times k$ Contingency Tables, *Statistics for Industry and Technology*, part V, pp. 417-431, DOI: 10.1007/0-8176-4487-3 27.

[64] Patil G. & Taillie C. (1982), Diversity as a concept and its measurements, *Journal of the American Statistical Association*, Vol. 77, pp. 548-561.

[65] Pearson K. & Heron D. (1913), On theories of association, *Biometrika* Vol. 9, pp. 159-315

[66] Rathie N. & Kannappan P. (1972), A directed-divergence function of type-β, *Information and Control*, Vol. 20, pp. 38-45.

[67] Read T. & Cressie N, (1988), Goodness-of-fit Statistics for Discrete Multivariate Data, New York: Springer.

[68] Read C. (1977), Partitioning chi-square in contingency tables: A teaching approach, *Communications in Statistics-Theory and Methods*, Vol. 6, pp. 533-562.

[69] Renyi A. (1959), On Measures of Dependence, *Acta Mathematica Hungarica*, Vol. 10, No 3-4, pp. 441-451.

[70] Renyi A. (1961), On Measures of Entropy and Information, *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 767.

[71] SAS/STAT User's guide (1999), Online doc., Version 8, Chapter 28.

[72] Shannon C. (1948), A mathematical theory of communication, *Bell System Technical Journal*, Vol. 27, pp.379-423, pp. 623-656.

[73] Siegel S. & Castellan N. (1988), Nonparametric statistics for the behavioral sciences (2^{nd} ed.), New York: McGraw-Hill Book Company.

[74] Scheaffer R. (1999), Categorical Data Analysis, *NCSSM Statistics Leadership Institute*, University of Florida 2nd ed. Wiley, New York.

[75] Schweizer B. & Wolff E. (1981), On Nonparametric Measures of Dependence for Random Variables, *The Annals of Mathematical Statistics*, Vol. 9, No. 4, pp. 879-885.

[76] Sheskin D. (2004), Handbook of Parametric and Nonparametric Statistical Precedures, 3rd ed. Chapman & Hall.

[77] Silvey S. (1964), On a Measure of Association, *The Annals of Mathematical Statistics*, Vol. 35, No. 3 pp. 1157-1166.

[78] Stigler S. (2002), The missing early history of contingency tables, *Annales de la Faculte des Sciences de Toulouse*, Vol.4, pp. 563-573.

[79] Tahata K., Miyamoto N., Tomizawa S. (2001), Measures of Departure from Quasi-Symmetry And Bradley-Terry Models for Square Contingency Tables with Nominal Categories, *Journal of the Korean Society*, Vol. 334, No. 1, pp. 129-147.

[80] Tomizawa S. (1989b), Decompositions for conditional symmetry model into palindromic symmetry and modified marginal homogeneity models, *Australian Journal of Statistics Sinica*, Vol. 31, pp. 287-296.

[81] Tomizawa S. (1994), Two Kinds of Measures of Departure from Symmetry In Square Contingency Tables Having Nominal Categories, *Statistica Sinica*, Vol. 4, pp. 325-334.

[82] Tomizawa S. (1995), Measures of Departure from Marginal Homogeneity for Contingency Tables with Nominal Categories, *The Statistician*, Vol. 44, No. 4, pp. 425-439.

[83] Tomizawa S. (1995), Measures of Departure from Global Symmetry for Square Contingency Tables with Ordered Categories, *Behaviormetrika*, Vol. 22, No. 1, pp. 91-98.

[84] Tomizawa S. (1998), A Decomposition of the Marginal Homogeneity Model into Three Models for Square Contingency Tables with Ordered Categories, *The Indian Journal of Statistics*, Series B, Vol. 60, No. 2, pp. 293-300.

[85] Tomizawa S. (2006), Decompositions of Symmetry Model Into Marginal Homogeneity αnd Distance Subsymmetry In Square Contingency Tables with Ordered Categories, REVSTAT - Statistical Journal, Vol. 4, No. 2, pp. 153-161.

[86] Tomizawa S., Seo T., Yamamoto H. (1998), Power-divergence-type Measure of Departure from Symmetry for Square Contingency Tables that Have Nominal Categories, *Statistics*, Vol. 39, No. 2, pp. 107-115.

[87] Tomizawa S., Saitoh K. (1999), Measure of Departure from Conditional Symmetry for Square Contingency Tables with Ordered Categories, *The Japan Statistical Society*, Vol. 29, No. 1, pp. 65-78.

[88] Tomizawa S., Miyamoto N., Hatanaka Y. (2001), Measures of Asymmetry for Square Contingency Tables Having Ordered Categories, *Australian and New Zealand Journal of Statistics*, Vol. 43, Issue 3, pp. 335-349.

[89] Tomizawa S., Miyamoto N., Ashihara N. (2003), Measures of Departure from Marginal Homogeneity for Square Contingency Tables Having Ordered Categories, *Behaviormetrika*, Vol. 30, No. 2, pp. 173-194.

[90] Tomizawa S., Miyamoto N., Yamane S., (2005), Power-divergence-type measure of Departure from Diagonals-Parameter Symmetry for Square Contingency Tables with Ordered Categories, *Journal of Applied Statistics*, Vol. 25, No. 2, pp. 387-398.