ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΑΝΤΙΜΕΤΩΠΙΣΗ ΣΦΑΛΜΑΤΩΝ ΤΑΞΙΝΟΜΗΣΗΣ (MISCLASSIFICATION) ΣΤΗΝ ΑΝΑΛΥΣΗ ΚΑΤΗΓΟΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Γεώργιος Οικονόμου

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

Νοέμβριος 2010

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη 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:

- Κατέρη Μαρία (Επιβλέπουσα)
- Ηλιόπουλος Γεώργιος
- Τζαβελάς Γεώργιος

Η έγκριση της Διπλωματική 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

MISCLASSIFICATION TREATMENT IN CATEGORICAL DATA ANALYSIS

By

George Economou

MSc Dissertation

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

Piraeus, Greece

November 2010

Στον αγαπημένο μου πατέρα

Χρήστο

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά την κα. Κατέρη Μαρία, πρώην Επ. Καθηγήτρια του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς και νυν Αναπ. Καθηγήτρια του Μαθηματικού Τμήματος του Πανεπιστημίου Ιωαννίνων, για την καθοδήγηση και τη βοήθεια που μου παρείχε για την υλοποίηση της παρούσας εργασίας, για τη συνεργασία μας όλο αυτό το διάστημα, καθώς και για την κατανόηση που επέδειξε σε δυσκολίες που εμφανίστηκαν κατά την προσπάθεια εκπόνησης της.

Επίσης θα ήθελα να ευχαριστήσω τους κυρίους Ηλιόπουλο Γεώργιο, Αναπ. Καθηγητή του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, και Τζαβελά Γεώργιο, Επ. Καθηγητή του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, για τη συμμετοχή τους στην Τριμελή Επιτροπή. Παράλληλα, θα ήθελα να εκφράσω τις ευχαριστίες μου και στον κο. Κούτρα Μάρκο, Καθηγητή του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς και Πρόεδρο του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική, για τη βοήθεια του ως προς την ολοκλήρωση της παρούσας εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου για την αμέριστη συμπαράσταση που μου έδειξε όλο αυτό τον καιρό και για τη συμβολή της στην ολοκλήρωση των σπουδών μου, καθώς και τους φίλους μου για την υποστήριξη που μου παρείχαν.

ΟΙΚΟΝΟΜΟΥ ΓΕΩΡΓΙΟΣ

ΑΝΩ ΠΑΤΗΣΙΑ

Νοέμβριος 2010

Περίληψη

Η λάθος ταξινόμηση, η εσφαλμένη μέτρηση μίας ή περισσότερων κατηγορικών μεταβλητών, είναι ένα σημαντικό πρόβλημα που συναντάται στη διεξαγωγή ερευνών που αφορούν πολλούς επιστημονικούς τομείς, όπως η Ιατρική και η Επιδημιολογία. Τα παρατηρούμενα δεδομένα που συλλέγονται σε μελέτες υπόκεινται συχνά σε λάθη ταξινόμησης. Ακόμα και σε σχετικά απλά σενάρια, εκτός εάν οι πιθανότητες λάθους ταξινόμησης είναι πολύ μικρές, ένα σημαντικό ποσοστό μεροληψίας μπορεί να προκύψει κατά την εκτίμηση του βαθμού συσχέτισης που αφορά συνήθη μέτρα όπως η διαφορά κινδύνων, ο σχετικός λόγος κινδύνου και ο λόγος σχετικών πιθανοτήτων. Η λάθος ταξινόμηση μπορεί επίσης να οδηγήσει στη μείωση της αποτελεσματικότητας κατά την ανάλυση των πινάκων συνάφειας.

Ο κύριος στόχος αυτής της εργασίας είναι να παρουσιαστούν οι επιδράσεις της λάθους ταξινόμησης στη μονοδιάστατη, διδιάστατη και πολυδιάστατη ανάλυση κατηγορικών δεδομένων, καθώς και να καταδειχτούν τα αποτελέσματά της για την εκτίμηση παραμέτρων και για έλεγχους υποθέσεων. Θα παρουσιαστούν μέθοδοι διόρθωσης για τις συνέπειες της λάθους ταξινόμησης, όπως είναι οι απλές πινακικές μέθοδοι και οι μέθοδοι μέσω μοντελοποίησης χρησιμοποιώντας δεδομένα επικύρωσης, καθώς και μέθοδοι διόρθωσης που χρησιμοποιούν επαναλαμβανόμενες μετρήσεις. Αυτές οι μέθοδοι θα συγκριθούν προκειμένου να διαπιστωθεί ποια αποδίδει καλύτερα, ανάλογα με το διαθέσιμο σύνολο δεδομένων και τη δειγματοληψία διαδικασία. Τέλος, οι μέθοδοι διόρθωσης θα εφαρμοστούν σε ένα σύνολο δεδομένων που περιέχει λάθος ταξινομημένα δεδομένα από μία έρευνα ασφαλείας αυτοκινητοδρόμων, ώστε να αξιολογηθεί η αποτελεσματικότητα της χρήσης ζώνης ασφαλείας στη μείωση των τραυματισμών σε ατυχήματα αυτοκινήτων σε σχέση με παράγοντες όπως η σοβαρότητα ζημιών των αυτοκινήτων και το φύλο.

Abstract

Misclassification, the erroneous measurement of one or several categorical variables, is a major problem met in conducting surveys concerning many scientific fields such as Medicine and Epidemiology. The observed data which are collected in studies are often subject to misclassification errors. Even in rather simple scenarios, unless the misclassification probabilities are very small, a significant amount of bias can arise in estimating the degree of association concerning common measures like risk difference, risk ratio and odds ratio. Misclassification can also lead to the reduction of efficiency in the analysis of contingency tables.

The main aim of this MSc Dissertation is to present the effects of misclassification in univariate, bivariate and multivariate analysis of categorical data, as well as to demonstrate its effects to parameter estimation and hypothesis testing. Methods of adjusting for the effects of misclassification will be presented, including simple matrix and model-based methods using validation data, as well as adjustment methods using repeated data. These methods will be compared in order to find out which one performs better, depending on the available data set and sampling process. Finally, adjustment methods will be performed on a data set containing misclassified data from highway safety research in order to evaluate the effectiveness of seat belt use in reducing injuries in automobile accidents depending on factors such as car damage severity and sex.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κατάλογος Συντομογραφιών

EM algorithm	αλγόριθμος μεγιστοποίησης-αναμενόμενης τιμής (Expectation-Maximization algorithm).
IPFA	επαναληπτικός αλγόριθμος αναλογικής προσαρμογής (iterative proportional fitting algorithm).
LC model	μοντέλο λανθανουσών κατηγοριών (latent class model)
LR-test	έλεγχοι λόγου πιθανοφανειών (likelihood ratio test)
TSS	τριπλό σχέδιο δειγματοληψίας (triple-sampling scheme)

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Τι είναι η Λανθασμένη Ταξινόμηση ;

Η ταξινόμηση είναι μια στοιχειώδης εργασία που εκτελείται με φυσικό τρόπο από τους ανθρώπους. Πολλές δραστηριότητες έχουν την ταξινόμηση ως θεμελιώδη έννοια. Η ταξινόμηση των ατόμων ενός πληθυσμού είναι δυνατόν να πραγματοποιηθεί με πολλούς τρόπους. Τα παραδείγματα περιλαμβάνουν την ταξινόμηση των εξεταζόμενων υποκειμένων ως προς το φύλο (αρσενικό / θηλυκό), το καθεστώς καπνίσματος (καπνιστής / μη καπνιστής), την κατάσταση της υγείας τους (άρρωστοι / υγιείς), και τα λοιπά. Αυτά τα παραδείγματα είναι βασισμένα σε διχοτομικά στοιχεία (*dichotomous data*). Οι ταξινομήσεις σε περισσότερες από δύο το πλήθος κατηγορίες είναι επίσης δυνατές, π.χ. τα άτομα μπορούν να ταξινομηθούν στις ακόλουθες ομάδες: (A) αδέσμευτος, (B) παντρεμένος, (C) διαζευγμένος, και (D) χηρευάμενος. Ταξινομήσεις μπορούν επίσης να πραγματοποιηθούν από δύο ή περισσότερα κριτήρια.

Όταν οι πληροφορίες συλλέγονται στον πραγματικό κόσμο, τα δεδομένα δεν απεικονίζουν συχνά την αληθινή κατάσταση των στοιχείων του δείγματος, δηλαδή η διαδικασία παραγωγής δεδομένων είναι συχνά επιρρεπής σε σφάλματα . Αυτό το γεγονός μπορεί να συμβεί εξαιτίας διάφορων αιτιών. Στις καταναλωτικές έρευνες, οι καταναλωτές συχνά δεν μπορούν να θυμηθούν τις προηγούμενες καταναλωτικές συμπεριφορές και συνήθειες τους με ακρίβεια, μπορούν να παρανοήσουν τις ερωτήσεις διεξαγόμενων ερευνών ή μπορούν σκόπιμα να δηλώσουν ψευδή στοιχεία (*misreport*). Στις ιατρικές διαγνώσεις, οι αποτυχίες-λανθασμένες διαγνώσεις εξετάσεων και οι λανθασμένα κωδικοποιημένες πληροφορίες γίνονται αιτίες διαστρέβλωσης. Στις εκλογικές έρευνες, οι ψηφοφόροι εμφανίζονται συχνά απρόθυμοι να εκφράσουν τις αληθινές απόψεις τους. Αυτές και άλλες καταστάσεις εμφανίζονται σε πολλές άλλες εφαρμογές. Η βασική συνέπεια είναι ότι τέτοιες διαστρεβλώσεις ή επιρρέπεια σε σφάλματα μπορούν να έχουν μια σημαντική επίδραση στα εξαγόμενα συμπεράσματα επειδή το ποσοστό των ωφέλιμων και ορθά τεκμηριωμένων πληροφοριών που λαμβάνεται από το

δείγμα μειώνεται σε σημαντικό βαθμό. Εάν τα σφάλματα σε μία διαδικασία παραγωγής δεδομένων δεν μοντελοποιούνται κατάλληλα, οι πληροφορίες που λαμβάνουμε μπορούν να θεωρηθούν πιο ακριβείς απ' ότι πραγματικά είναι, οδηγώντας σε πολλές περιπτώσεις σε μη βέλτιστη λήψη αποφάσεων.

Κατά την περίπτωση όπου ο διαχωρισμός των κατηγοριών δεν επιδέχεται αμφισβήτησης (είναι για παράδειγμα της μορφής 'εν ζωή' και 'θανόντες'), ο κίνδυνος εμφάνισης λανθασμένης ταξινόμησης είναι πρακτικά αδύνατος. Μολαταύτα, ακόμα και στο πεδία της Ιατρικής ή της Επιδημιολογίας ελλοχεύει μία αξιοσημείωτη πιθανότητα σφάλματος σε ορισμένα είδη σύνθετων διαγνώσεων. Σχετικά αναφέρει ο Bross (1954)

" Στις πιο πολύπλοκες διαγνώσεις, ο κλινικός ιατρός συνειδητοποιεί ότι υπάρχει ένας αξιοσημείωτος κίνδυνος σφάλματος, ένας κίνδυνος που μπορεί να ποικίλλει σε μεγάλο βαθμό ανάλογα με την υπό μελέτη ασθένεια, τη διαθεσιμότητα και την ύπαρξη διαγνωστικών ελέγχων και άλλων παραγόντων. "

1.2 Λανθασμένη Ταξινόμηση στα Κατηγορικά Δεδομένα

Στα πρώιμα στάδια μελέτης των κατηγορικών δεδομένων, ανακαλύφθηκε ότι η λανθασμένη ταξινόμηση των κατηγορικών μεταβλητών επέφερε προβλήματα στην ανάλυση και στην ερμηνεία των αποτελεσμάτων. Στην Επιδημιολογία έχει παρατηρηθεί διαχρονικό ενδιαφέρον ως προς την εκτίμηση της επίδρασης της λανθασμένης ταξινόμησης σε μελέτες που συνδέουν ασθένειες και παράγοντες έκθεσης σε αυτές. Προσφάτως, έχει δοθεί μεγαλύτερη βαρύτητα στη μεθοδολογία εκτίμησης της δομής της λανθασμένης ταξινόμησης, καθώς και την προσαρμογή μοντέλων για την προκύπτουσα μεροληψία. Σε αυτά περιλαμβάνεται η συγκέντρωση βοηθητικών (*auxiliary*) δεδομένων διαμέσου δειγμάτων επικύρωσης (*validation samples*) και η μέθοδος των επαναλαμβανόμενων μετρήσεων (*repeated measures*). Εδώ όμως θα πρέπει να επισημανθεί το εξής : Μολονότι υπάρχουν άμεσοι παραλληλισμοί μεταξύ της μελέτης του φαινομένου της λανθασμένης ταξινόμησης και του φαινομένου των σφαλμάτων μέτρησης (*measurement errors*) στις συνεχείς μεταβλητές έκθεσης, τα δύο αυτά φαινόμενα διαφοροποιούνται και ως προς την ιστορική τους εξέλιξη και ως προς τη στατιστική μεθοδολογία και τεχνικές που εξετάζονται για την αντιμετώπιση τους.

Το πρόβλημα της λανθασμένης ταξινόμησης έχει αναπτυχθεί φιλολογικά αλλά κατά κανόνα μόνο για δεδομένα κατηγορικής φύσεως. Οι πρώιμες αναφορές έγιναν από τον Bross (1954). Ο Bross εξέτασε την επίδραση της λανθασμένης ταξινόμησης σε ένα ποσοστό και τη διαφορά μεταξύ δύο ποσοστών προερχόμενων από ανεξάρτητους πληθυσμούς λαμβάνοντας όμως ως δεδομένο ότι τα ποσοστά λανθασμένης ταξινόμησης των δύο δειγμάτων ήταν πανομοιότυπα, μία υπόθεση που αμφισβητήθηκε μελλοντικά από αρκετούς ερευνητές. Το αντικείμενο αυτό εξετάστηκε και από τους Cochran (1968) και Goldberg (1975), οι οποίοι έδωσαν επίσης τη μεροληψία και για μια περίπτωση δύο άνισων ποσοστών σφαλμάτων (*error rates*). Τα συμπεράσματα στα οποία κατέληξαν αυτές οι δύο περιπτώσεις προκύπτουν υπό τη δέσμευση ότι η λανθασμένη ταξινόμηση είναι πιθανή προς μία και μοναδική διάσταση ή κατεύθυνση. Τοιουτοτρόπως, υποθέτουν ότι οι δύο πληθυσμοί που αντιστοιχούν στα δύο δείγματα δεν συγχέονται (*not confounded*).

Όταν οι παρατηρήσεις προέρχονται από ένα τυχαίο δείγμα που δεν υπόκειται σε περιορισμούς και ταξινομούνται σε διάφορες κατηγορικές μεταβλητές, το να προβούμε σε συμπεράσματα για λανθασμένη ταξινόμηση προς μία μόνο κατεύθυνση είναι συχνά πολύ περιοριστικό. Αρκετοί ερευνητές προσέγγισαν αυτό το πρόβλημα για 2×2 πίνακες συνάφειας, μία περίπτωση που μπορεί να θεωρηθεί ότι μελετούμε μία διχοτομική ταξινομική μεταβλητή και μία διχοτομική μεταβλητή απόκρισης. Οι Keys and Kihlberg (1963) και οι Gullen et al. (1968) έδωσαν τις προϋποθέσεις βάσει των οποίων η προφανής διαφορά ανάμεσα στα δύο ποσοστά της μεταβλητής απόκρισης είναι μικρότερη από την πραγματική διαφορά. Αναφορά για τους $r\times c$ πίνακες συνάφειας έχει γίνει και από τους Assakul and Proctor (1967), με βασικό πεδίο ενδιαφέροντος τους X^2 ελέγχους. Στο τελευταίο σύγγραμμα, σε αντίθεση με τους άλλους ερευνητές, γίνεται αναφορά στην περίπτωση των εξαρτημένων σφαλμάτων στις διαφορετικές μεταβλητές.

Το πρόβλημα της λανθασμένης ταξινόμησης σε δύο κατευθύνσεις έχει επίσης γίνει αντικείμενο μελέτης και από άλλους ερευνητές (βλέπε Giesbrecht 1967, Goldberg 1975, Barron 1977 και Hochberg 1977). Ο Koch (1969) ερεύνησε την περίπτωση των επαναλαμβανόμενων ταξινομήσεων από έναν επιρρεπή σε σφάλματα μηχανισμό. Οι Rubin et al. (1956) και Rogot (1961) ασχολήθηκαν κυρίως με τη λανθασμένη ταξινόμηση προς μία κατεύθυνση, όπως και οι Mote and Anderson (1965), οι οποίοι και ερεύνησαν πολυωνυμικά δεδομένα. Ο Schwartz (1985) παρουσίασε συμπεράσματα για διαστήματα εμπιστοσύνης.

Ο Tenenbein (1970, 1971, 1972) υλοποίησε σχεδιασμούς με διπλή δειγματοληψία (*double sampling schemes*) για την εκτίμηση πολυωνυμικών ποσοστών σε έναν πίνακα συνάφειας. Αυτός ο σχεδιασμός εφαρμόστηκε για ελέγχους ανεξαρτησίας από τους Chiacchierini and Arnold (1977). Οι Hochberg (1977) και Chen (1979) επέκτειναν τη χρήση της διπλής δειγματοληψίας, δημιουργώντας μια βάση για μοντελοποίηση και ελέγχους γενικότερα για τα κατηγορικά δεδομένα, το δεύτερο μάλιστα με χρήση λογαριθμογραμμικών μοντέλων για τη λανθασμένη ταξινόμηση. Μία σχετική εργασία, με έμφαση στους πίνακες συνάφειας $2\times2\times2$, έχει γραφεί από τον Chen (1989). Οι Kuha και Skinner (1997) και ο S. Greenland (2007) παρουσιάζουν μια πιο πρόσφατη θεώρηση. Εξαιρετικά βιβλία έχουν γραφεί και από τους Gustafson (2003) και Buonaccorsi (2009), που περιγράφουν αναλυτικά την επίδραση της λανθασμένης ταξινόμησης με περισσότερη έμφαση στη διμεταβλητή περίπτωση.

1.3 Αντικείμενο της Εργασίας

Σε αυτήν την εργασία θα γίνει μια κριτική ανασκόπηση της υπάρχουσας βιβλιογραφίας για τις επιδράσεις της λανθασμένης ταξινόμησης στην ανάλυση κατηγορικών δεδομένων. Στο δεύτερο Κεφάλαιο θα παρουσιαστούν οι επιδράσεις της λανθασμένης ταξινόμησης κατά τη μονομεταβλητή, διμεταβλητή και πολυμεταβλητή περίπτωση και θα αναλυθούν διεξοδικά οι περιπτώσεις ύπαρξης διαφορίσιμης, μη διαφορίσιμης και ανεξάρτητης λανθασμένης ταξινόμησης. Στο τρίτο Κεφάλαιο θα εξεταστεί η επίδραση της λανθασμένης ταξινόμησης για βασικά κριτήρια ελέγχου υποθέσεων και θα προταθούν μέθοδοι για τη διόρθωση των κριτηρίων αυτών. Στο τέταρτο Κεφάλαιο θα παρουσιαστούν μέθοδοι για την εξαγωγή συμπερασμάτων για το μηχανισμό λανθασμένης ταξινόμησης, όπως η επιλογή δειγμάτων επικύρωσης ή η χρήση επαναλαμβανόμενων μετρήσεων. Στο πέμπτο Κεφάλαιο, θα παρουσιαστούν δύο βασικές πινακικές μέθοδοι διόρθωσης της λανθασμένης ταξινόμησης και θα προβούμε σε συγκρίσεις για την εύρεση της βέλτιστης εκ των δύο μεθόδων. Στο έκτο Κεφάλαιο, θα παρουσιαστούν μέθοδοι διόρθωσης μέσω μοντελοποίησης με χρήση λογαριθμογραμμικών μοντέλων και βασιζόμενες στην ύπαρξη δειγμάτων επικύρωσης. Στο έβδομο Κεφάλαιο, ερευνάται η μεθοδολογία αντιμετώπισης της λανθασμένης ταξινόμησης με χρήση επαναλαμβανόμενων μετρήσεων και επιχειρείται η σύγκριση των μεθόδων αυτών με προηγούμενες μεθόδους. Τέλος, στο όγδοο Κεφάλαιο γίνεται εφαρμογή ορισμένων μεθόδων διόρθωσης σε ένα σύνολο δεδομένων που περιέχει λανθασμένα ταξινομημένα δεδομένα από

μία έρευνα ασφαλείας αυτοκινητοδρόμων, ώστε να αξιολογηθεί η αποτελεσματικότητα της χρήσης ζώνης ασφαλείας στη μείωση των τραυματισμών σε ατυχήματα αυτοκινήτων σε σχέση με παράγοντες όπως η σοβαρότητα ζημιών των αυτοκινήτων και το φύλο.

ΚΕΦΑΛΑΙΟ 2

Επίδραση της λανθασμένης ταξινόμησης στην κατηγορική ανάλυση δεδομένων

2.1 Μονομεταβλητή Ανάλυση Κατηγορικών Δεδομένων

Ας εξετάσουμε πρώτα το βασικό πρόβλημα εύρεσης τρόπου για να παραστήσουμε τη λανθασμένη ταξινόμηση σε έναν πεπερασμένο πληθυσμό. Υποθέτουμε ότι ο τυπικός 'μηχανισμός' μέτρησης (*standard measurement device*) υπόκειται σε λανθασμένη ταξινόμηση, με αποτέλεσμα την εξαγωγή μεροληπτικών (*biased*) αποτελεσμάτων. Μη μεροληπτικές εκτιμήσεις μπορούν να γίνουν με χρήση πιο εκλεπτυσμένων εργαλείων μέτρησης, που συνήθως αναφέρονται ως προτιμώμενες διαδικασίες (*preferred procedures* – αναλυτικότερη αναφορά σε αυτές ακολουθεί σε επόμενο Κεφάλαιο). Για να παραστήσουμε τη λανθασμένη ταξινόμηση των πληθυσμιακών μονάδων, ορίζουμε ως Y' τη μεταβλητή ταξινόμησης που υπόκειται σε σφάλμα (*fallible variable*) και Y την αληθή μεταβλητή (*true variable*) που θέλουμε να μετρήσουμε μέσω της μεταβλητής ταξινόμησης. Στην Επιδημιολογία συνήθως αναφερόμαστε στην Y' ως υποκατάστατη μεταβλητή (*surrogate*) της Y. Για ένα άτομο του εξεταζόμενου πληθυσμού, έστω ξ, έχουμε τις τιμές των μεταβλητών Y'_ξ και Y_ξ αντίστοιχα. Για κάθε άτομο της έρευνας, οι τιμές των Y'_ξ και Y_ξ κατατάσσονται σε r αμοιβαίως αποκλειόμενες κατηγορίες. Υποθέτουμε επίσης ότι τα άτομα είναι ανεξάρτητα μεταξύ τους και ότι οι τιμές Y_ξ είναι καλά ορισμένες και σταθερές. Σύμφωνά με τον Bross (1954), οι πιθανότητες λανθασμένης ταξινόμησης γράφονται μέσω της ακόλουθης τυχαίας διαδικασίας ως εξής

$$\Pr(Y'_\xi = j \mid Y_\xi = i) = \alpha_{ji}, \qquad \text{με} \quad j,i = 1,....,r \tag{2.1}$$

δηλ. α_{ji} είναι η πιθανότητα ένα άτομο να καταγραφεί εσφαλμένα στην κατηγορία j ενώ στην πραγματικότητα ανήκει στην κατηγορία i.

Εάν θεωρήσουμε ότι οι πληθυσμιακές μονάδες που ανήκουν στην πραγματική κατηγορία *i* προέρχονται από έναν άπειρο υπερπληθυσμό (*infinite superpopulation*) των εν λόγω μονάδων, η ποσότητα α_{ji} μπορεί να ερμηνευτεί ως το ποσοστό των μονάδων αυτών που θα μπορούσαν να έχουν ταξινομηθεί σαν κατηγορίας j.

Στη σχετική βιβλιογραφία έχουν δοθεί και εναλλακτικές προσεγγίσεις για τη μορφή της λανθασμένης ταξινόμησης. Ο Koch (1969) θεωρεί μία αρκετά γενική προσέγγιση βάση της οποίας η μεταβλητή Y'_{ξ} αντιμετωπίζεται σαν το αποτέλεσμα μίας μόνο από ένα σύνολο επαναλαμβανόμενων μετρήσεων. Οι Lessler and Kalsbeek (1992, βλ. Παρ. 10.3) θεωρούν αντίθετα μία μη στοχαστική προσέγγιση όπου οι πιθανότητες α_{ji} αποτελούνται από πεπερασμένα πληθυσμιακά ποσοστά.

Για το σχηματισμό δομής στο μηχανισμό της λανθασμένης ταξινόμησης, οι τιμές των παραμέτρων α_{ji} αναπαριστώνται σε έναν $r \times r$ πίνακα λανθασμένης ταξινόμησης (*misclassification matrix*) $\mathbf{A} = [\alpha_{ji}]$, αποτελούμενο από μη αρνητικά στοιχεία και στήλες που τα στοιχεία τους αθροίζουν στη μονάδα. Για μία δυαδική (*binary*) μεταβλητή απόκρισης, όπου $r = 2$ και οι κατηγορίες της υποδεικνύουν την παρουσία ($Y_{\xi} = 2$) ή απουσία ($Y_{\xi} = 1$) του προς εξέταση χαρακτηριστικού, ο πίνακας λανθασμένης ταξινόμησης περιλαμβάνει δύο μόνο παραμέτρους, τις α_{11} και α_{22}

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} = \begin{pmatrix} \alpha_{11} & 1-\alpha_{22} \\ 1-\alpha_{11} & \alpha_{22} \end{pmatrix} \tag{2.2}$$

Στην ιατρική βιβλιογραφία, όπου το εξεταζόμενο χαρακτηριστικό είναι συνήθως ασθένεια, ορίζουμε τις παραμέτρους ευαισθησία (*sensitivity*) και ειδικότητα (*specificity*), με συμβολισμούς *sen* και *sp* αντίστοιχα. Οι πιθανότητες ταξινόμησης α_{11} και α_{22} μπορούν να εκφραστούν μέσω της ειδικότητας και ευαισθησίας μέσω των ακόλουθων σχέσεων

$$\alpha_{11} = sp \qquad \text{και} \qquad \alpha_{22} = sen\,.$$

Τα εξεταζόμενα χαρακτηριστικά διαφορετικών ερευνητικών μεθόδων που ακολουθούνται μπορούν να συγκριθούν βάσει των δύο αυτών ποσοτήτων (βλ. Rogan and Gladen, 1978 και King and Lu, 2008). Έστω για παράδειγμα μία ερευνητική μέθοδος που ταξινομεί ένα άτομο ως 'ανάπηρο'. Εάν κάποια από μία μεγάλη σειρά ερωτήσεων από το χρησιμοποιούμενο ερωτηματολόγιο έχει μία θετική απάντηση μπορεί να έχει 'καλή' ευαισθησία, δηλαδή τα περισσότερα άτομα με κάποια μορφή αναπηρίας θα επιλεγούν σωστά, αλλά μπορεί να έχει

‘κακή’ ειδικότητα, δηλαδή πολλά άτομα που δεν έχουν αναπηρία να ταξινομηθούν ως ανάπηρα. Χρησιμοποιώντας ορολογία που αντιστοιχεί σε ελέγχους υποθέσεων, η ποσότητα $1-sen$ είναι ουσιαστικά η πιθανότητα σφάλματος τύπου I, δηλαδή η πιθανότητα απόρριψης μίας μηδενικής υπόθεσης όταν ως μηδενική υπόθεση θεωρηθεί η ύπαρξη ασθένειας. Αφ’ ετέρου, η ποσότητα $1-sp$ αντιστοιχεί στην ισχύ του ελέγχου, δηλαδή στην πιθανότητα σωστής απόρριψης μίας λανθασμένης μηδενικής υπόθεσης. Βάσει των προλεχθέντων, οι πιθανότητες λανθασμένης ταξινόμησης μπορούν να εκφραστούν με χρήση της ευαισθησίας και της ειδικότητας μέσω των ακόλουθων πιθανοτήτων

$$\Pr(Y'_{\xi}=1\,|\,Y_{\xi}=2)=\alpha_{12}=1-sen$$

και

$$\Pr(Y'_{\xi}=2\,|\,Y_{\xi}=1)=\alpha_{21}=1-sp$$

που συμβολίζουν τις πιθανότητες λανθασμένα αρνητικού (*false negative*) και λανθασμένα θετικού (*false positive*) αντίστοιχα. Έτσι, ο πίνακας λανθασμένης ταξινόμησης της σχέσης (2.2) παίρνει την ακόλουθη μορφή

$$\mathbf{A}=\begin{pmatrix}\alpha_{11} & \alpha_{12}\\ \alpha_{21} & \alpha_{22}\end{pmatrix}=\begin{pmatrix} sp & 1-sen\\ 1-sp & sen\end{pmatrix},$$

όπου ευαισθησία και ειδικότητα είναι οι δύο μοναδικές παράμετροι.

Οι έννοιες της ευαισθησίας και ειδικότητας είναι βασικές έννοιες για την ανάλυση επίδρασης λανθασμένης ταξινόμησης σε κατηγορικά δεδομένα, χρησιμοποιούνται όμως κυρίως σε πεδία όπως η ιατρική και η επιδημιολογία. Λόγω αυτού και θέλοντας να διευρύνουμε τα εξαγόμενα συμπεράσματα στο γενικότερο πλαίσιο εφαρμογών των κατηγορικών δεδομένων, για την ανάλυση του σφάλματος ταξινόμησης στη διμεταβλητή ή πολυμεταβλητή ανάλυση κατηγορικών δεδομένων ορίζουμε ως ενιαίους συμβολισμούς των πιθανοτήτων λανθασμένης ταξινόμησης τους εξής

πιθανότητα λανθασμένα αρνητικού : $\theta_Y=\Pr(Y'_{\xi}=1\,|\,Y_{\xi}=2)=\alpha_{12}=1-sen$ (2.3)

και

πιθανότητα λανθασμένα θετικού : $\varphi_Y=\Pr(Y'_{\xi}=2\,|\,Y_{\xi}=1)=\alpha_{21}=1-sp$ (2.4)

Επανερχόμενοι στην ανάλυση για επίδραση της λανθασμένης ταξινόμησης για τη μονοδιάστατη ανάλυση κατηγορικών δεδομένων, προχωράμε στον ορισμό βασικών εννοιών στηριζόμενοι στο ερευνητικό πλαίσιο που δόθηκε από τους Kuha and Skinner (1997). Έστω ότι Y είναι η αποκριτική μεταβλητή του εξεταζόμενου πληθυσμού, για την οποία θέλουμε να

προβούμε σε συμπερασματολογία. Ορίζουμε $N_Y(i)$ το πλήθος των μονάδων του πληθυσμού που έχουν απόκριση στην Y και ανήκουν στην i κατηγορία αυτής, ισχύει δηλαδή $Y_\xi = i$, και την πιθανότητα σωστής ταξινόμησης στην κατηγορία αυτή ως

$$P_i = \Pr(Y_\xi = i) = N_Y(i)/N \tag{2.5}$$

με $\sum_{i=1}^{r} N(i) = N$ το σύνολο του πληθυσμού. Αντίστοιχα ορίζουμε και την πιθανότητα ταξινόμησης με απόκριση Y' στην j κατηγορία ως

$$P'_J = \Pr(Y'_\xi = j) = N_{Y'}(j)/N \tag{2.6}$$

με $\sum_{j=1}^{r} N(j) = N$. Τότε η σχέση που συνδέει την παρατηρούμενη πιθανότητα ταξινόμησης, μέσω της Y', με την ορθή ταξινόμηση των πληθυσμιακών μονάδων, μέσω της Y, δίνεται από την ακόλουθη σχέση

$$\Pr(Y'_\xi = j) = \sum_{i=1}^{r} \Pr(Y'_\xi = j \mid Y_\xi = i) \cdot \Pr(Y_\xi = i) \tag{2.7}$$

Υποθέτουμε ότι σκοπός μας είναι να εκτιμήσουμε το διάνυσμα των «ορθών» πιθανοτήτων $\mathbf{P_Y} = (P_{Y(1)}, ..., P_{Y(r)})^c$. Ένας γνωστός εκτιμητής του είναι το διάνυσμα των πληθυσμιακών ποσοστών

$$\hat{\mathbf{P}} = (\hat{P}_{Y(1)}, ..., \hat{P}_{Y(r)})^c, \quad \text{όπου} \quad \hat{P}_{Y(i)} = \sum_{\xi \in s} w_\xi I(Y_\xi = i) \tag{2.8}$$

και αντίστοιχα εκτιμητής του διανύσματος των «παρατηρούμενων» πιθανοτήτων $\mathbf{P'_Y} = (P'_{Y(1)}, ..., P'_{Y(r)})^c$ είναι το διάνυσμα

$$\hat{\mathbf{P}}' = (\hat{P}'_{Y(1)}, ..., \hat{P}'_{Y(r)})^c, \quad \text{όπου} \quad \hat{P}'_{Y(j)} = \sum_{\xi \in s} w_\xi I(Y'_\xi = j) \tag{2.9}$$

με s να δηλώνει το δείγμα, w_ξ το δειγματικό βάρος (*sample weight*)[1] για τη μονάδα ξ και $I(\cdot)$ τη δείκτρια συνάρτηση. Υποθέτουμε ότι τα βάρη w_ξ είναι ορισμένα έτσι ώστε το διανύσματα $\hat{\mathbf{P}}$ και $\hat{\mathbf{P}}'$ είναι κατά προσέγγιση αμερόληπτοι εκτιμητές των $\mathbf{P}$ και $\mathbf{P}'$ αντίστοιχα για το επιλεγμένο δειγματοληπτικό σχήμα. Ορίζοντας E_r την αναμενόμενη τιμή υπό το μοντέλο της λανθασμένης ταξινόμησης, με τη βοήθεια της σχέσης (2.7) προκύπτει ότι

[1] Η χρήση δειγματικών βαρών δίνει σε όλες τις μονάδες του χρησιμοποιούμενου δείγματος την ίδια βαρύτητα στην επιλογή τους στο δείγμα, διορθώνοντας τη δυσαναλογία του επιλεγμένου δείγματος ως προς την πιθανότητα επιλογής τους. Αποτελούν αντισταθμιστικό παράγοντα σε προβλήματα δειγματοληπτικών σχεδίων και μεθόδων συλλογής δεδομένων που οδηγούν σε μεροληπτικούς σχεδιασμούς, για παράδειγμα ελλιπής κάλυψη του συνόλου του πληθυσμού (*under-coverage*), "non-ignorable" δειγματοληπτικά σχέδια και άλλα.

$$E_r[I(Y'_\xi = j)] = \sum_{i=1}^{r} \alpha_{ji} I(Y_\xi = i)$$

έτσι ώστε

$$E_r[\hat{\mathbf{P}}'] = \mathbf{A} \cdot E_r[\hat{\mathbf{P}}] \quad , \tag{2.10}$$

από όπου και καταλήγουμε στην σχέση υπολογισμού των ορθών πιθανοτήτων ταξινόμησης

$$\mathbf{P} = \mathbf{A}^{-1}\mathbf{P}' \tag{2.11}$$

με $\mathbf{A} = [\alpha_{ji}]_{r\times r}$ να είναι ο πίνακας λανθασμένης ταξινόμησης. Για να ισχύει η τελευταία σχέση, θα πρέπει να υποθέσουμε ότι ο πίνακας $\mathbf{A}$ είναι μη ιδιάζων, δηλαδή αντιστρέψιμος. Στην πράξη, η υπόθεση αντιστρεψιμότητας του πίνακα $\mathbf{A}$ δεν επιβάλλει μεγάλους περιορισμούς στην επιλογή του σχήματος λανθασμένης ταξινόμησης. Ο πίνακας $\mathbf{A}^{-1}$ υπάρχει όταν τα στοιχεία της κύριας διαγωνίου του υπερτερούν σε σχέση με τα υπόλοιπα στοιχεία του πίνακα, με άλλα λόγια όταν τα στοιχεία $\{\alpha_{ii},\ \iota = 1,...,r\} > 1/2$. Η υπόθεση αυτή είναι λογική, δεδομένου ότι οι πιθανότητες αυτές είναι οι πιθανότητες ορθής ταξινόμησης των δεδομένων.

Εάν χρησιμοποιήσουμε το εκτιμώμενο διάνυσμα $\hat{\mathbf{P}}'$ σαν εκτιμητής του διανύσματος $\mathbf{P}$, η μεροληψία που προκύπτει από τη λανθασμένη ταξινόμηση εκφράζεται ως εξής

$$Bias(\hat{\mathbf{P}}') = (\mathbf{A} - \mathbf{I})\mathbf{P} \tag{2.12}$$

όπου $\mathbf{I}$ είναι ο $r \times r$ ταυτοτικός πίνακας. Με άλλα λόγια, η μεροληψία της λανθασμένης ταξινόμησης των υπό εξέταση μεταβλητών προσμετράται σε σχέση με την απόκλιση του πίνακα λανθασμένης ταξινόμησης $\mathbf{A}$ από τον αντίστοιχο μοναδιαίο πίνακα ίδιων διαστάσεων. Έτσι, όσο μεγαλύτερη είναι η διαφορά των δύο πινάκων, τόσο πιο έντονο είναι το φαινόμενο της λανθασμένης ταξινόμησης και η υπεισερχόμενη μεροληψία στις εκτιμήσεις. Επισημαίνεται ότι η μεροληψία αυτή εκτιμάται λαμβάνοντας υπ' όψιν και τη λανθασμένη ταξινόμηση αλλά και το δειγματοληπτικό σχέδιο που εφαρμόζεται για τον εξεταζόμενο πληθυσμό.

Η φύση της μεροληψίας είναι διαισθητικά ευκολότερη να ερμηνευθεί κατά τη δυαδική (*binary*) περίπτωση. Ορίζοντας τις πιθανότητες $P_1 = \Pr(Y = 1)$ και $P_2 = \Pr(Y = 2)$ ως πιθανότητες παρουσίας ή απουσίας του εξεταζόμενου χαρακτηριστικού αντίστοιχα[2], γνωρίζουμε ότι για το ζεύγος πιθανοτήτων (P_1, P_2) ισχύει $(P_1, P_2) = (1 - P_2, P_2)$. Η σχέση αυτή

[2] Σε μεγάλο αριθμό ερευνητικών βιβλίων και συγγραμμάτων αναφέρονται ως πιθανότητες αποτυχίας και επιτυχίας αντίστοιχα

σε συνδυασμό με τις σχέσεις (2.5), (2.6), (2.7) δίνουν τις ακόλουθες εκτιμήσεις των πληθυσμιακών ποσοστών κατά την ύπαρξη λανθασμένης ταξινόμησης

$$\hat{P}_1' = (1-\varphi_Y)(1-P_2) + \theta_Y P_2$$

$$\hat{P}_2' = \varphi_Y (1-P_2) + (1-\theta_Y) P_2$$

από όπου προκύπτουν οι ακόλουθες ισότητες για την προκύπτουσα μεροληψία των εκτιμήσεων

$$Bias(\hat{P}_1') = -Bias(\hat{P}_2') \tag{2.13}$$

και

$$Bias(\hat{P}_2') = \varphi_Y (1-P_2) - \theta_Y P_2 \tag{2.14}$$

(βλ. Cochran (1968), Schwartz (1985)). Από τις σχέσεις (2.3), (2.4) και (2.13) και λαμβάνοντας υπ'όψιν πως οι πιθανότητες λανθασμένης ταξινόμησης φ_Y και θ_Y είναι συνήθως άγνωστες και εκτιμούνται μέσω του δείγματος, η διαφορά των πληθυσμιακών ποσοστών $\hat{P}_2' - \hat{P}_2$ υπολογίζεται ως εξής

$$\hat{P}_2' - \hat{P}_2 = \hat{\varphi}_Y (1-P_2) - \hat{\theta}_Y P_2 ,$$

από όπου προκύπτει και ο διορθωμένος εκτιμητής $\hat{P}_2$ του πραγματικού ποσοστού

$$\hat{P}_2 = \frac{\hat{P}_2' - \hat{\phi}_Y}{1-\hat{\theta}_Y - \hat{\phi}_Y} \tag{2.15}$$

Οι δυνατές τιμές που μπορεί να πάρει ο παραπάνω εκτιμητής είναι δυνατόν να είναι μικρότερες του 0 και μεγαλύτερες του 1. Σε τέτοιες περιπτώσεις, και υπό το πρίσμα της σημειακής εκτίμησης, ορίζουμε ο εκτιμητής να παίρνει τις τιμές 0 και 1 αντίστοιχα. Αξίζει εδώ να σημειώσουμε εάν η εκτιμώμενη πιθανότητα λανθασμένα θετικού $\hat{\phi}_Y$ είναι μεγαλύτερη του $\hat{P}_2$, τότε ο διορθωμένος εκτιμητής $\hat{P}_2$ είναι αρνητικός (υπό την προϋπόθεση ότι η ποσότητα $1-\hat{\theta}_Y - \hat{\phi}_Y$ είναι θετική, το οποίο και ισχύει σχεδόν για όλες τις περιπτώσεις). Για να παρακαμφθεί το πρόβλημα αυτό, καταφεύγουμε στην κατασκευή κατάλληλων διαστημάτων εμπιστοσύνης για τον εκτιμητή $\hat{P}_2$ (βλ. Buonaccorsi (2009)).

Σε αυτό το σημείο, αξίζει επίσης να επισημανθεί κάτι. Ακόμα και εάν ο πίνακας λανθασμένης ταξινόμησης διαφέρει από τον ταυτοτικό πίνακα, είναι δυνατόν να μην υπάρχει καθαρή μεροληψία (*net bias*) στον υπολογισμό του $\hat{P}_2'$. Αυτό συμβαίνει όταν τα δύο

σφάλματα λανθασμένης ταξινόμησης είναι αμοιβαία συμψηφιζόμενα (*mutually compensated*), δηλαδή όταν ισχύει $\varphi_Y P_1 = \theta_Y P_2$ (παραδείγματος χάριν όταν $P_2 = 0.6$, $\phi_Y = 0.03, \theta_Y = 0.02$). Η πιθανότητα τέτοιου αμοιβαίου συμψηφισμού έχει εξεταστεί από τους Chua and Fuller (1987). Ας επισημάνουμε εδώ ότι η μη ύπαρξη μεροληψίας δεν εξαρτάται μόνο από τις παραμέτρους λανθασμένης ταξινόμησης α_{ji} αλλά και από τις πραγματικές παραμέτρους $\mathbf{P}$. Κατά αυτόν τον τρόπο, μία μεθοδολογία μέτρησης που χρησιμοποιεί προκαθορισμένο πίνακα λανθασμένης ταξινόμησης μπορεί να οδηγήσει σε μεροληπτικές εκτιμήσεις κατά την εξέταση ενός πληθυσμού και σε μη μεροληπτικές εκτιμήσεις για άλλους πληθυσμούς.

Συνοψίζοντας, η λανθασμένη ταξινόμηση μπορεί να οδηγήσει σε αυθαίρετες μορφές μεροληψίας για την εκτίμηση των πληθυσμιακών ποσοστών που αντιστοιχούν σε κατηγορίες μίας δοσμένης μεταβλητής που εξετάζεται για το δείγμα . Σε κάποιες περιπτώσεις, το μέγεθος της λανθασμένης ταξινόμησης που προκύπτει για μία κατηγορία της μεταβλητής μπορεί να εξισωθεί από το μέγεθος της λανθασμένης ταξινόμησης για άλλες κατηγορίες, με αποτέλεσμα να μην προκύψει καθαρή μεροληψία.

2.2 Διμεταβλητή Ανάλυση

2.2.1 Εκτιμήσεις για 2×2 πίνακες συνάφειας

Η απλούστερη μορφή ανάλυσης που μπορούμε να θεωρήσουμε είναι η σύγκριση των ποσοστών μεταξύ δύο υποομάδων του εξεταζόμενου πληθυσμού, δηλ. κατά την περίπτωση όπου η μεταβλητή απόκρισης και οι επεξηγηματικές μεταβλητές είναι δυαδικές. Έστω Y η δυαδική μεταβλητή απόκρισης, τα ποσοστά της οποίας και ενδιαφερόμαστε να εκτιμήσουμε, και X η δυαδική μεταβλητή που καθορίζει τις δύο υποομάδες εξέτασης του πληθυσμού. Θεωρούμε πρώτα την περίπτωση κατά την οποία η Y ταξινομείται εσφαλμένα ως Y'. Η περίπτωση αυτή έχει γίνει αντικείμενο μελέτης από πολλούς ερευνητές στο πεδίο της Επιδημιολογίας (για παράδειγμα Rubin et al. (1956)) για περιπτώσεις όπου η Y καθορίζει την ύπαρξη ή μη μίας ασθένειας και η X δύο ομάδες έκθεσης του πληθυσμού σε αυτήν, για παράδειγμα καπνιστές ή μη καπνιστές. Ορίζουμε $P_{Y(i)|X(k)} = \Pr(Y_\xi = i, X_\xi = k)$ το ποσοστό των μονάδων για τις οποίες έχουμε $Y_\xi = i$ για τον υποπληθυσμό όπου $X_\xi = k$, δηλαδή για τις

μονάδες του πληθυσμού που κατατάσσονται στην i κατηγορία της Y και ανήκουν στον k υποπληθυσμό, καθώς και $\hat{P}_{Y'(i)|X(k)}$ τον εκτιμητή του $P_{Y(i)|X(k)}$ με τρόπο ανάλογο της $\hat{P}'_Y$ της σχέσης (2.9). Θεωρούμε τη μεροληπτική διαφορά $\hat{P}_{Y'(2)|X(2)} - \hat{P}_{Y'(2)|X(1)}$ σαν εκτιμητή της διαφοράς ανάμεσα στα πραγματικά πληθυσμιακά ποσοστά $P_{Y(2)|X(2)} - P_{Y(2)|X(1)}$. Χάριν ευκολίας ως προς τον συμβολισμό λόγω της ύπαρξης περισσότερων από μία κατηγορικών μεταβλητών, η πιθανότητα λανθασμένης ταξινόμησης α_{ji} θα συμβολίζεται καθ' εξής ως $\alpha_{j|i}$.

Μη διαφορίσιμη λανθασμένη ταξινόμηση

Η έκφραση της μεροληψίας απλοποιείται υπό την προϋπόθεση ότι έχουμε την ίδια ευαισθησία *sen* και ειδικότητα *sp* και για τις δύο κατηγορίες της μεταβλητής X. Μία πιο ακριβής διατύπωση για την περίπτωση αυτή είναι η εξής : Η λανθασμένη ταξινόμηση της μεταβλητής Y ονομάζεται μη διαφορίσιμη (*non differential*) ως προς τη μεταβλητή (*with respect to*) X όταν οι πιθανότητες λανθασμένης ταξινόμησης για την Y δεν διαφοροποιούνται για κάθε επίπεδο της μεταβλητής X, δηλαδή

$$\Pr(Y'_\xi = j \mid Y_\xi = i, X_\xi = k) = \Pr(Y'_\xi = j \mid Y_\xi = i) \Leftrightarrow \alpha_{j|ik} = \alpha_{j|i} \quad \text{για κάθε } k \quad (2.16)$$

Συνεπάγεται έτσι από τη σχέση (13) ότι σε περίπτωση μη διαφορίσιμης λανθασμένης ταξινόμησης έχουμε

$$Bias[\hat{P}_{Y'(2)|X(2)} - \hat{P}_{Y'(2)|X(1)}] = -(\theta_Y + \varphi_Y)[P_{Y(2)|X(2)} - P_{Y(2)|X(1)}] \quad (2.17)$$

ή εναλλακτικά

$$E[\hat{P}_{Y'(2)|X(2)} - \hat{P}_{Y'(2)|X(1)}] = (1 - \theta_Y - \varphi_Y)[P_{Y(2)|X(2)} - P_{Y(2)|X(1)}]. \quad (2.18)$$

Στην πράξη, αναμένουμε οι πιθανότητες λανθασμένης ταξινόμησης θ_Y και φ_Y να είναι μικρότερες του $1/2$, απαίτηση λογική λόγω της φύσεως των πιθανοτήτων αυτών, ούτως ώστε ο συντελεστής $1-\theta_Y-\varphi_Y$ της σχέσης (2.18) να είναι μία θετική σταθερά με τιμές μεταξύ 0 και 1. Με τον τρόπο αυτόν, η διαφορά που προσμετράται από την εσφαλμένη μεταβλητή Y' είναι πάντα μικρότερη από την διαφορά που βασίζεται στους υπολογισμούς με χρήση της μεταβλητής Y και το αποτέλεσμα της λανθασμένης ταξινόμησης είναι να 'μειώνεται'- να 'εξασθενεί' (*attenuate*) η διαφορά μεταξύ των ποσοστών των υποκατηγοριών (βλέπε Rubin et al. (1956), Newell (1963), Buell and Dunn (1964), White (1986)). Με άλλα λόγια, η επίδραση

της μεταβλητής (ή του παράγοντα) X στη μεταβλητή Y παρουσιάζεται να είναι μικρότερη από ότι πραγματικά είναι.

Ομοίως, μπορούμε να δείξουμε παρόμοια αποτελέσματα και για τον λόγο σχετικών πιθανοτήτων (*odds ratio*), ένα από τα σημαντικότερα εργαλεία που χρησιμοποιούνται για την εξαγωγή συμπερασμάτων στην μελέτη των κατηγορικών δεδομένων. Ο εκτιμώμενος λόγος σχετικών πιθανοτήτων της τιμής του εξεταζόμενου χαρακτηριστικού $Y=2$ για την υποομάδα $X=2$ σε σχέση με την υποομάδα $X=1$ κατά την ύπαρξη λανθασμένης ταξινόμησης δίνεται από τον ακόλουθο τύπο

$$\hat{OR}_{Y'(2)|X(k)} = \frac{\hat{P}_{Y'(2)|X(2)} \big/ (1-\hat{P}_{Y'(2)|X(2)})}{\hat{P}_{Y'(2)|X(1)} \big/ (1-\hat{P}_{Y'(2)|X(1)})} \tag{2.19}$$

και ως γνωστόν αποτελεί έναν συνεπή εκτιμητή του σχετικών πιθανοτήτων λανθασμένης ταξινόμησης

$$OR_{Y'(2)|X(k)} = \frac{P_{Y'(2)|X(2)} \big/ (1-P_{Y'(2)|X(2)})}{P_{Y'(2)|X(1)} \big/ (1-P_{Y'(2)|X(1)})} \tag{2.20}$$

Επίσης, ο πραγματικός λόγος σχετικών πιθανοτήτων όταν δεν υφίσταται λανθασμένη ταξινόμηση είναι ο

$$OR_{Y(2)|X(k)} = \frac{P_{Y(2)|X(2)} \big/ (1-P_{Y(2)|X(2)})}{P_{Y(2)|X(1)} \big/ (1-P_{Y(2)|X(1)})} \tag{2.21}$$

Εφόσον έχουμε μη διαφορίσιμη λανθασμένη ταξινόμηση της μεταβλητής Y ως προς την X, ισχύουν οι ακόλουθες ισότητες για τα πραγματικά και εκτιμώμενα πληθυσμιακά ποσοστά επιτυχίας

$$P_{Y(2)|X(k)} = P_{Y(2)} \equiv P_2, \qquad \text{για } k=1,2$$

και

$$P_{Y'(2)|X(k)} = P_{Y'(2)} \equiv P_2', \qquad \text{για } k=1,2$$

Εισάγοντας τις ανωτέρω ισότητες στους λόγος σχετικών πιθανοτήτων $OR_{Y(2)|X(k)}$ και $OR_{Y'(2)|X(k)}$ αντίστοιχα, αποδεικνύεται ότι ο $OR_{Y'(2)|X(k)}$ είναι μεροληπτικός εκτιμητής του $OR_{Y(2)|X(k)}$, ενώ συγχρόνως τείνει στην 'μηδενική τιμή' (*null value*) της μονάδας, ικανοποιώντας τις σχέσεις $1 \le OR_{Y'(2)|X(k)} \le OR_{Y(2)|X(k)}$ ή $OR_{Y(2)|X(k)} \le OR_{Y'(2)|X(k)} \le 1$ (βλ. Gustafsson (2004, Παρ. 3.3), Buonaccorsi (2009)). Αυτό ισχύει λαμβάνοντας υπ'οψιν ότι ισχύει η σχέση $1-\theta_Y-\varphi_Y>0$, το οποίο συμβαίνει σχεδόν σε όλες τις περιπτώσεις. Στο ίδιο

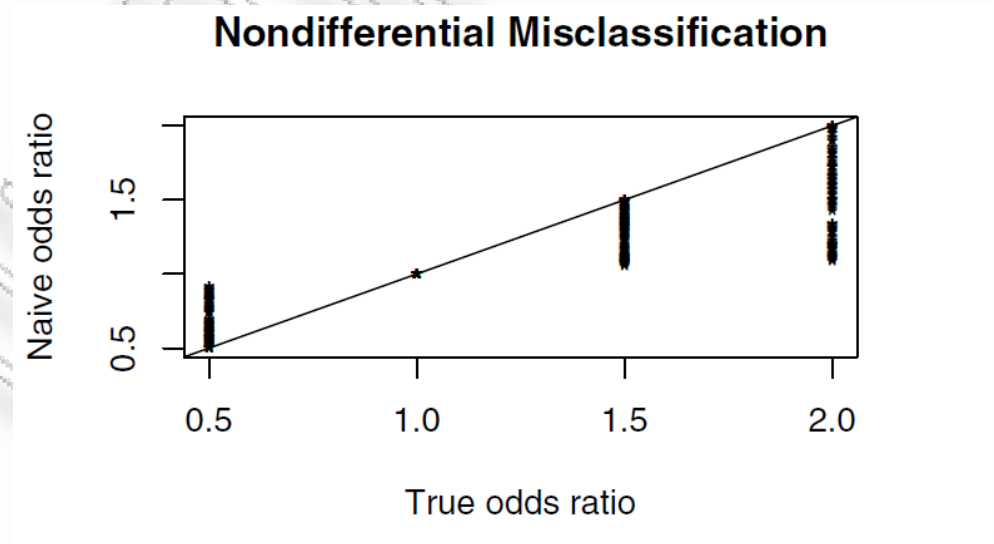
συμπέρασμα μπορούμε να καταλήξουμε χρησιμοποιώντας τον λόγο σχετικού κινδύνου (*relative risk*)

$$RR_{Y(2)|X(k)} = \frac{P_{Y(2)|X(2)}}{P_{Y(2)|X(1)}}$$

και τις εκτιμήσεις αυτού μέσω του λόγου $\hat{RR}_{Y'(2)|X(k)} = \hat{P}_{Y'(2)|X(2)} \Big/ \hat{P}_{Y'(2)|X(1)}$ ((Copeland et al. (1977), White (1986)), αφού ο λόγος σχετικών πιθανοτήτων και $OR_{Y(2)|X(k)}$ είναι ουσιαστικά ο λόγος σχετικού κινδύνου $RR_{Y(2)|X(k)}$ εάν οι πιθανότητες $P_{Y(2)|X(2)}$ και $P_{Y(2)|X(1)}$ είναι σχετικά μικρές. Ας σημειωθεί εδώ ότι τα παραπάνω συμπεράσματα χαρακτηρίζουν την ασυμπτωτική μεροληψία για μεγάλα μεγέθη δειγμάτων. Για μικρά δείγματα, είναι πιθανόν η μεροληψία να μην τείνει στην 'μηδενική τιμή', με αποτέλεσμα να υπάρχει σοβαρό ενδεχόμενο οι εκτιμήσεις που προκύπτουν να μην ανταποκρίνονται στα πραγματικά δεδομένα και να οδηγούν σε εντελώς λανθασμένα συμπεράσματα (βλ. Jurek et al. (2005)). Η επίδραση της μη διαφορίσιμης λανθασμένης ταξινόμησης στους λόγους σχετικών πιθανοτήτων μπορεί να δοθεί διαισθητικά στο παρακάτω σχήμα (Σχήμα 2.1)

ΣΧΗΜΑ 2.1

Διάγραμμα εκτιμώμενου λόγου σχετικών πιθανοτήτων έναντι πραγματικού λόγου σχετικών πιθανοτήτων υπό μη διαφορίσιμη λανθασμένη ταξινόμηση.
Η γραμμή δηλώνει την απουσία μεροληψίας ('μηδενική τιμή')
[ΠΗΓΗ : Buonaccorsi, 2009]



Στο σημείο αυτό ας επισημάνουμε κάτι σημαντικό. Τα δύο είδη λανθασμένης ταξινόμησης, που δίνονται από τις ποσότητες θ_Y και φ_Y, έχουν το ίδιο πρόσημο στη σχέση (2.18) ενώ αντίθετο πρόσημο στη σχέση (2.14). Συνεπώς, η λανθασμένη ταξινόμηση ανάμεσα σε ζεύγη κατηγοριών που μπορεί στη μονοδιάστατη ανάλυση να είναι αμοιβαίως συμψηφιζόμενα δε θα είναι η ίδια για τη διδιάστατη ανάλυση. Έτσι μπορούμε να προβούμε στην υπόθεση ότι η λανθασμένη ταξινόμηση θα έχει πιθανότατα μεγαλύτερη σχετική επίδραση στα αποτελέσματα τέτοιων αναλύσεων.

Για την περίπτωση όπου η μεταβλητή *X*, που καθορίζει τις υποομάδες του πληθυσμού, υπόκειται σε λανθασμένη ταξινόμηση ενώ η μεταβλητή απόκρισης είναι σωστά ταξινομημένη και επιπλέον η λανθασμένη ταξινόμηση της *X* είναι μη διαφορίσιμη ως προς την *Y*, προκύπτουν παρόμοιου τύπου 'εξασθενήσεις' (Rogot (1971), Shy et al. (1978), Greenland (1982), Flegal et al. (1986)).

Στην περίπτωση που έχουμε λανθασμένη ταξινόμηση και για τις δύο μεταβλητές *Y* και *X*, υποθέτουμε γενικά ότι οι εσφαλμένα ταξινομημένες Y' και X' καθορίζονται από κοινού από τις Y και X αντίστοιχα από την ακόλουθη στοχαστική διαδικασία

$$\Pr(Y'_\xi = j, X'_\xi = l \mid Y_\xi = i, X_\xi = k) = a_{jl|ik} \qquad \text{για κάθε } i, j, k, l$$

Η λανθασμένη ταξινόμηση των Y και X ονομάζεται ανεξάρτητη (*independent*) εάν ισχύει

$$\Pr(Y'_\xi = j, X'_\xi = l \mid Y_\xi = i, X_\xi = k) = \Pr(Y'_\xi = j \mid Y_\xi = i, X_\xi = k)\Pr(X'_\xi = l \mid Y_\xi = i, X_\xi = k)$$

για κάθε i, j, k, l και μη διαφορίσιμη ως προς τις Y και X όταν η λανθασμένη ταξινόμηση της *Y* είναι μη διαφορίσιμη ως προς την *X* και η λανθασμένη ταξινόμηση της *X* είναι μη διαφορίσιμη ως προς την *Y*, δηλαδή

$$\Pr(Y'_\xi = j \mid Y_\xi = i, X_\xi = k) = \Pr(Y'_\xi = j \mid Y_\xi = i) \Leftrightarrow a_{j|ik} = a_{j|i}$$

και

$$\Pr(X'_\xi = l \mid Y_\xi = i, X_\xi = k) = \Pr(X'_\xi = l \mid X_\xi = k) \Leftrightarrow a_{l|ik} = a_{l|k} \quad \text{για κάθε } i, j, k, l.$$

Για την περίπτωση που έχουμε ανεξάρτητη μη διαφορίσιμη λανθασμένη ταξινόμηση (*independent non differential misclassification*) για τις μεταβλητές *Y* και *X*, ισχύει με άλλα λόγια η σχέση

$$\Pr(Y'_\xi = j, X'_\xi = l \mid Y_\xi = i, X_\xi = k) = \Pr(Y'_\xi = j \mid Y_\xi = i)\Pr(X'_\xi = l \mid X_\xi = k) \Leftrightarrow a_{jl|ik} = a_{j|i}a_{l|k}$$

οι Gullen et al. (1968) έδειξαν ότι ο εκτιμητής για την πραγματική διαφορά πληθυσμιακών ποσοστών $P_{Y(2)|X(2)} - P_{Y(2)|X(1)}$ 'εξασθενεί' με τρόπο όμοιο με αυτόν που παρουσιάστηκε στην παραπάνω ανάλυση (βλ. επίσης Keys and Kihlberg (1963) και Barron (1977)), ούτως ώστε να

μπορεί να θεωρηθεί αμερόληπτος εκτιμητής της. Σε αυτό το σημείο πρέπει να λάβουμε σοβαρά υπ' όψιν τα πρόσφατα συμπεράσματα στα οποία κατέληξαν οι Greenland and Gustafson (2004), που θέτουν εν μέρει υπό αμφισβήτηση την αμεροληψία των εκτιμήσεων κάτω από την υπόθεση της μη διαφορίσιμης λανθασμένης ταξινόμησης. Συμπεραίνουν ότι η προκύπτουσα μεροληψία μπορεί να αποκλίνει από τη μηδενική τιμή που ορίζεται από τις πραγματικές τιμές των δεδομένων, όταν για τα εξεταζόμενα δεδομένα είναι δύσκολο να πληρούνται οι προϋποθέσεις για την ύπαρξη μη διαφορίσιμης λανθασμένης ταξινόμησης (όπως για παράδειγμα η συσχέτιση σφαλμάτων στα χαρακτηριστικά που εξετάζονται μέσω ερωτηματολογίων ή η μη ελεγχόμενη επίδραση συμμεταβλητών στους λόγους σφαλμάτων) ή ακόμα και σε ορισμένες περιπτώσεις που αυτές πληρούνται.

Μία ειδική περίπτωση εμφανίζεται σε διαχρονικά διεξαγόμενες μελέτες όταν Y και X είναι οι τιμές μίας μεταβλητής κατά τις διαδοχικές μετρήσεις της σε διαφορετικά χρονικά σημεία σε μία διεξαγόμενη έρευνα (*consecutive waves of a survey*) και οι παράμετροι $\Pr(Y_\xi = i, X_\xi = k)$ συμβολίζουν τις 'μικτές ροές' (*gross flows*) ανάμεσα στις κατηγορίες i και k (Chua and Fuller (1987)). Στην περίπτωση αυτή, η λανθασμένη ταξινόμηση μπορεί να επιφέρει σοβαρή σχετική μεροληψία για τις εκτιμήσεις των ροών που δεν βρίσκονται στην κύρια διαγώνιο (που ισχύει $i \neq k$).

Σε αυτό το σημείο πρέπει να τονίσουμε κάτι πολύ σημαντικό. Η μη διαφορίσιμη λανθασμένη ταξινόμηση δεν συνεπάγεται πάντα ότι οι εκτιμήσεις που προκύπτουν εξασθενούν προς τη μηδενική τιμή, είναι δηλαδή ασυμπτωτικά αμερόληπτες εκτιμήσεις των πραγματικών τιμών των εξεταζόμενων χαρακτηριστικών (Jurek et al. (2005)). Ακόμα και εάν θεωρήσουμε ότι οι εκτιμήσεις είναι πιθανότερα πιο κοντά στην μηδενική τιμή σε σχέση με τη μη ύπαρξη λανθασμένης ταξινόμησης, η εύρεση ασυμπτωτικά αμερόληπτων εκτιμητών όταν έχουμε μη διαφορίσιμη λανθασμένη ταξινόμηση προϋποθέτει να ισχύουν και άλλες υποθέσεις, όπως π.χ. την ανεξαρτησία του σφάλματος ταξινόμησης από άλλους τύπους σφαλμάτων (Chavance et al. (1992), Kristensen (1992)). Οι Greenland and Gustafson (2006) εκφράζουν την αβεβαιότητα τους για την ασυμπτωτική αμεροληψία των εκτιμήσεων και κατά την περίπτωση ύπαρξη ανεξάρτητης μη διαφορίσιμης λανθασμένης ταξινόμησης. Παραδείγματα που δείχνουν ότι η κατά προσέγγιση μη διαφορίσιμη λανθασμένη ταξινόμηση μπορεί να μην οδηγεί στη μηδενική τιμή δίνονται από τους Jurek et al. (2008).

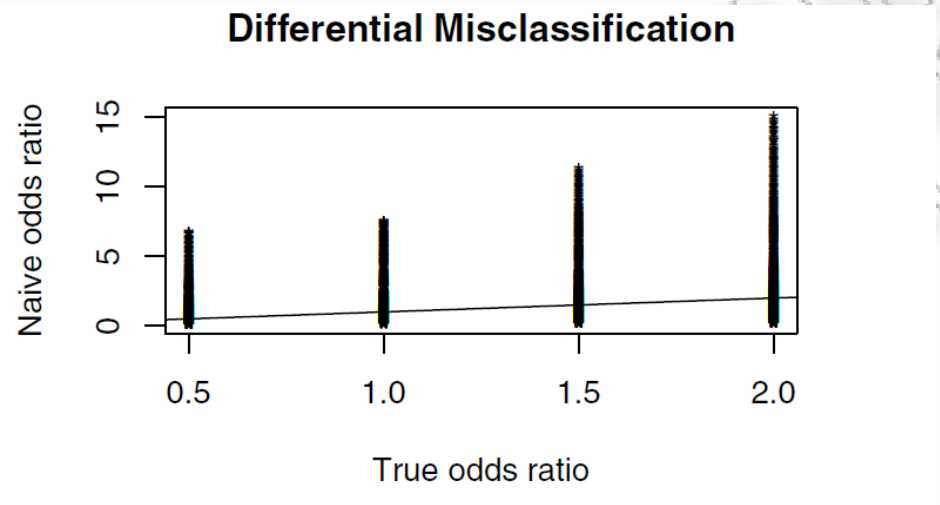
Διαφορίσιμη λανθασμένη ταξινόμηση

Ας εξετάσουμε τώρα την περίπτωση όπου η λανθασμένη ταξινόμηση δεν είναι μη διαφορίσιμη, αλλά διαφορίσιμη (*differential*). Κατά την περίπτωση αυτή, η μεροληψία της εκτίμησης της διαφοράς των πραγματικών ποσοστών $P_{Y(2)|X(2)} - P_{Y(2)|X(1)}$ μπορεί να πάρει οποιαδήποτε αυθαίρετη τιμή που να τείνει ή να αποκλίνει από τη μηδενική τιμή που σχετίζεται τα πραγματικά ποσοστά, όπως και στη μονοδιάστατη περίπτωση, συνεπώς δε χρειάζεται να συνεπάγεται εξασθένηση της σχέσης προς τη μηδενική τιμή αυτή. Την πιθανότητα μη εξασθενημένων προς τη μηδενική τιμή αποτελεσμάτων τονίζουν στην έρευνα τους οι Diamond and Lilienfeld (1962a, 1962b) αλλά προβαίνουν στην ομολογουμένως αφύσικη υπόθεση ότι η δεσμευμένη πιθανότητα της μεταβλητής Y δεδομένων των μεταβλητών Y' και X δεν εξαρτάται από τις τιμές της μεταβλητής X (και όχι για τη δεσμευμένη πιθανότητα της Y' δεδομένων των Y και X όπως αυτή δίνεται στη σχέση (2.18)). Άλλοι ερευνητές (Newell (1962), Keys and Kihlberg (1963), Buel and Dunn (1964)) επισημαίνουν αυτή την πηγή σύγχυσης. Ένας σαφής απολογισμός των πιθανών επιδράσεων για τη διαφορίσιμη λανθασμένη ταξινόμηση δίνεται από την Goldberg (1975) (βλ. επίσης και τους Copeland et al. (1977), Shy et al. (1978)). Ο Chyou (2007) με χρήση προσομοίωσης και βάσει των συνεπειών της διαφορίσιμης λανθασμένης ταξινόμησης στους λόγους σχετικών πιθανοτήτων (ένας μεγάλος αριθμός τείνει μακριά από τη μηδενική τιμή της μονάδας και ή προς το 0 ή προς το άπειρο) οδηγείται και αυτός στο συμπέρασμα της αυθαίρετης συμπεριφοράς των εκτιμητών ως προς τις τιμές που παίρνουν σε σχέση με τη μηδενική τιμή. Η επίδραση της διαφορίσιμης λανθασμένης ταξινόμησης στους λόγους σχετικών πιθανοτήτων μπορεί να δοθεί διαισθητικά στο παρακάτω σχήμα (Σχήμα 2.2)

ΣΧΗΜΑ 2.2

Διάγραμμα εκτιμώμενου λόγου σχετικών πιθανοτήτων έναντι πραγματικού λόγου σχετικών πιθανοτήτων υπό διαφορίσιμη λανθασμένη ταξινόμηση. Η γραμμή δηλώνει την απουσία μεροληψίας (‘μηδενική τιμή’)

[ΠΗΓΗ : Buonaccorsi, 2009]



Αλλαγές στις κατηγορίες μίας λανθασμένα ταξινομημένης μεταβλητής μπορούν να μετατρέψουν τη μη διαφορίσιμη λανθασμένη ταξινόμηση της σε διαφορίσιμη. Για παράδειγμα, ας υποθέσουμε ότι μία μεταβλητή Y με τρεις κατηγορίες είναι υποκείμενη σε μη διαφορίσιμη λανθασμένη ταξινόμηση ως προς μία δυαδική μεταβλητή X. Αν δύο από τις κατηγορίες της Y συνδυαστούν, η λανθασμένη ταξινόμηση του 2×2 πίνακα συνάφειας που προκύπτει δεν είναι απαραίτητα μη διαφορίσιμη (Wachholder et al., (1991)). Ομοίως, διαφορίσιμη λανθασμένη ταξινόμηση μπορεί να προκύψει όταν μία εσφαλμένα μετρημένη συνεχής μεταβλητή X διχοτομείται, ακόμα και αν το σφάλμα μέτρησης για την X είναι μη διαφορίσιμο (Flegal et al. 1991).

2.2.2 Εκτιμήσεις για r×2 πίνακες συνάφειας

Υποθέτουμε ότι θέλουμε να συγκρίνουμε τα πληθυσμιακά ποσοστά που ορίζονται από μία δυαδική μεταβλητή Y για τις $r>2$ υποομάδες που ορίζουν οι κατηγορίες της μεταβλητής X. Εάν μόνο η Y είναι λανθασμένα ταξινομημένη και η λανθασμένη ταξινόμηση αυτή είναι μη διαφορίσιμη ως προς την X, για κάθε εκτιμώμενη διαφορά

$$\hat{P}_{Y'(2)|X(k)} - \hat{P}_{Y'(2)|X(l)}, \qquad \text{με} \quad k,l = 1,2,...,r$$

έχουμε μία σχέση παρόμοια με αυτήν της σχέσης (2.18) με συντελεστή εξασθένησης $(1-\theta_Y-\varphi_Y)$. Κατά αυτόν τον τρόπο, η διάταξη των εκτιμώμενων ποσοστών είναι ίδια με τη διάταξη των πληθυσμιακών ποσοστών $P_{Y(2)|X(1)},...,P_{Y(2)|X(r)}$.

Εάν αντί αυτού η μεταβλητή Y είναι ορθά ταξινομημένη αλλά η μεταβλητή X που ορίζει τις υποομάδες του πληθυσμού είναι μη διαφορίσιμα λανθασμένα ταξινομημένη, μπορεί να προκύψουν διαφορετικά αποτελέσματα. Ορίζουμε τις εξής παραμέτρους λανθασμένης ταξινόμησης για τη μεταβλητή X

$$\beta_{lk} = \Pr(X'_{\xi} = l \mid X_{\xi} = k) \qquad (2.22)$$

με $k,l = 1,2,...,r$. Τότε για μεγάλα δείγματα η αναμενόμενη τιμή για το εκτιμώμενο ποσοστό $\hat{P}_{Y(2)|X'(l)}$ μπορεί με τη βοήθεια της σχέσης (2.22) και του θεωρήματος του Bayes να προσεγγιστεί από την παρακάτω σχέση

$$E[\hat{P}_{Y(2)|X'(l)}] = \frac{\sum_k P_{X(k)} P_{Y(2)|X(k)} \beta_{lk}}{\sum_{\kappa=1}^{r} P_{X(k)} \beta_{lk}}, \quad \text{για κάθε } l = 1,2,...,r \quad .$$

Υποθέτουμε ότι οι υποομάδες είναι διατεταγμένες έτσι ώστε

$$P_{Y(2)|X(1)} \le P_{Y(2)|X(2)} \le ... \le P_{Y(2)|X(r)}$$

Τότε ισχύει

$$P_{Y(2)|X(1)} \le E[\hat{P}_{Y(2)|X'(l)}] \le P_{Y(2)|X(r)}$$

(Kuha and Skinner (1997)) και για την αναμενόμενη τιμή της διαφοράς των εκτιμώμενων ποσοστών των ακραίων κατηγοριών της μεταβλητής X έχουμε

$$-[P_{Y(2)|X(r)} - P_{Y(2)|X(1)}] \le E[\hat{P}_{Y(2)|X'(l)} - \hat{P}_{Y(2)|X'(k)}] \le P_{Y(2)|X(r)} - P_{Y(2)|X(1)}$$

για κάθε $k,l = 1,2,...,r$.

Βάσει της παραπάνω σχέσης, η εκτιμώμενη διαφορά των ποσοστών μεταξύ των δύο ακραίων (*extreme*) υποομάδων της μεταβλητής X ($X_{\xi} = 1$ και $X_{\xi} = m$) είναι και πάλι 'εξασθενημένη' σε σχέση με την πραγματική διαφορά των ποσοστών. Μολαταύτα, οι εκτιμήσεις για άλλες διαφορές μεταξύ των ποσοστών των υποομάδων μπορεί να τείνουν μεροληπτικά είτε προς είτε μακριά από το μηδέν. Μάλιστα, η εκτίμηση αυτή μπορεί να έχει και διαφορετικό πρόσημο από αυτό της πραγματικής διαφοράς. Αυτό σημαίνει ότι η λανθασμένη ταξινόμηση μπορεί μερικές φορές να αλλάξει την προφανή διάταξη των ποσοστών των υποομάδων και να στρεβλώσει τις υπάρχουσες τάσεις (*trends*) ανάμεσα στα

ποσοστά των υποομάδων. Σε παρόμοια αποτελέσματα μπορούμε να καταλήξουμε ερευνώντας και άλλα μέτρα συσχέτισης (*measures of association*), όπως ο σχετικός λόγος κινδύνου (*relative risk*)

$$RR_{Y(2)|X(l)} = \frac{P_{Y(2)|X(l)}}{P_{Y(2)|X(k)}}$$

και ο λόγος σχετικών πιθανοτήτων (*odds ratio*)

$$OR_{Y(2)|X(l,k)} = \frac{P_{Y(2)|X(l)} / [1 - P_{Y(2)|X(l)}]}{P_{Y(2)|X(k)} / [1 - P_{Y(2)|X(k)}]},$$

που χρησιμοποιούνται ευρέως στην Επιδημιολογία (βλέπε Gladen and Rogan 1979, Dosemeci et al. 1990, Birkett 1992 καθώς και Blettner and Wahrendorf 1984).

Μολονότι υπάρχουν και άλλες πιθανές μορφές μεροληψίας, τα πιο συνηθισμένα αποτελέσματα για μία λανθασμένη ταξινόμηση με μη ακραίες περιπτώσεις τυχόν εσφαλμένα ταξινομημένων μονάδων, ακόμα και για μία πολυδιάστατη μεταβλητή X που διαχωρίζει τον πληθυσμό σε υποομάδες, φαίνονται να είναι τα εξής

- ✓ οι εκτιμήσεις των πληθυσμιακών ποσοστών διατηρούν τη διάταξη τους
- ✓ όλες οι εκτιμώμενες διαφορές των ποσοστών 'εξασθενούν'- τείνουν προς τη μηδενική τιμή.

Αυτά τα αποτελέσματα παρατηρούνται ειδικότερα όταν εμφανίζεται λανθασμένη ταξινόμηση μόνο για γειτονικές κατηγορίες της μεταβλητής X (βλέπε Marshall et al. 1981, Birkett 1992).

2.2.3 Έλεγχοι υποθέσεων για διδιάστατους πίνακες

Μία σημαντική συνέπεια των αποτελεσμάτων της παρατηρούμενης 'εξασθένησης' είναι ότι εάν δεν υπάρχει συσχέτιση ανάμεσα στις μεταβλητές Y και X δεν θα υπάρξει και συσχέτιση ανάμεσα στις μεταβλητές Y' και X' όταν έχουμε μη διαφορίσιμη λανθασμένη ταξινόμηση για μία από τις δύο μεταβλητές (Mote and Anderson (1975), Bross (1954), Rubin et al. (1956), Rogot (1961), Walsh (1963), Gladen and Rogan, (1979)) ή μη διαφορίσιμη και ανεξάρτητη λανθασμένη ταξινόμηση και τις δύο μεταβλητές (Assakul and Proctor (1967), Giesbrecht (1967)). Ως εκ τούτου, ένας έλεγχος για τη μη συσχέτιση ανάμεσα στις Y και X θα έχει το σωστό επίπεδο σημαντικότητας, αλλά γενικά η ισχύς του ελέγχου θα παρουσιάζεται μειωμένη. Οι Marshall et al. (1981), Walkner and Blettner (1985) και Freudenheim et al.

(1989) καταδεικνύουν παρόμοια αποτελέσματα για ελέγχους μη ύπαρξης τάσης στα ποσοστά των υποομάδων για πίνακες συνάφειας όπου η μεταβλητή απόκρισης Y είναι δίτιμη και η μεταβλητή που ορίζει τις υποομάδες στον πληθυσμό είναι μη διαφορίσιμα λανθασμένα ταξινομημένη.

2.3 Πολυμεταβλητή ανάλυση

2.3.1 Τριδιάστατοι πίνακες συνάφειας

Η απλούστερη μορφή πολυμεταβλητής ανάλυσης που μπορούμε να θεωρήσουμε παριστάνεται με τη βοήθεια ενός $2\times2\times2$ πίνακα συνάφειας. Έστω Y μία δυαδική μεταβλητή απόκρισης και X, Z δύο μεταβλητές που ορίζουν υποομάδες στον πληθυσμό. Ορίζουμε $P_{Y(i)|X(k),Z(l)}$ το ποσοστό των μονάδων του πληθυσμού για τις οποίες ισχύει $Y_\xi = i$ και ανήκουν στην υποομάδα για την οποία έχουμε $X_\xi = k$ και $Z_\xi = l$. Άλλες περιθώριες και δεσμευμένες πιθανότητες (παραδείγματος χάριν οι $P_{X(k)}$ και $P_{X(k)|Z(l)}$) ορίζονται με τρόπο παρόμοιο. Υποθέτουμε ότι επιθυμούμε να εξετάσουμε τη σχέση μεταξύ των μεταβλητών Y και X για μία δεδομένη κατηγορία της Z. Συγκεκριμένα, θέλουμε να εκτιμήσουμε τα ακόλουθα πληθυσμιακά ποσοστά

$$D_1 = P_{Y(2)|X(2),Z(1)} - P_{Y(2)|X(1),Z(1)} \qquad \text{και} \qquad D_2 = P_{Y(2)|X(2),Z(2)} - P_{Y(2)|X(1),Z(2)}.$$

Ας θεωρήσουμε σε πρώτη φάση ότι έχουμε μη διαφορίσιμη λανθασμένη ταξινόμηση για μία μόνο μεταβλητή. Εάν η μεταβλητή Z έχει ταξινομηθεί ορθά, μπορούμε να θεωρήσουμε τους διδιάστατους πίνακες συνάφειας των Y και X ξεχωριστά για κάθε υποομάδα που ορίζεται από την Z. Με αυτή τη θεώρηση, τα αποτελέσματα στα οποία καταλήξαμε για τη διμεταβλητή ανάλυση σε προηγούμενη ενότητα μπορούν να εφαρμοστούν κι εδώ. Εάν η Y είναι λανθασμένα ταξινομημένη, τότε οι διαφορές D_1 και D_2 εξασθενούν σύμφωνα με την ίδιο παράγοντα εξασθένησης που δίνεται στη σχέση (2.18). Οι διαφορές παρουσιάζουν επίσης εξασθένηση όταν έχουμε λανθασμένη ταξινόμηση της μεταβλητής X. Επειδή όμως η μεροληψία αυτή εξαρτάται επιπλέον και από τα ποσοστά $P_{X(2)|Z(2)}$ και $P_{X(2)|Z(2)}$, οι D_1 και D_2 εξασθενούν σε διαφορετικό βαθμό (όχι βάσει ενός κοινού συντελεστή) εάν υπάρχει συσχέτιση μεταξύ των X και Z. Κατά συνέπεια η μη διαφορίσιμη λανθασμένη ταξινόμηση για τη μεταβλητή X μπορεί να δημιουργήσει κίβδηλη ετερογένεια (*spurious heterogeneity*) για τις

εκτιμήσεις των D_1 και D_2 ή ακόμα και να καλύψει την ύπαρξη πραγματικής ετερογένειας (Greenland 1980).

Ας λάβουμε τώρα υπ' όψιν την περίπτωση όπου η Z ταξινομείται μη διαφορίσιμα ως Z' αλλά οι μεταβλητές Y και X είναι ορθά ταξινομημένες. Ορίζουμε τις πιθανότητες λανθασμένης ταξινόμησης με τρόπο ανάλογο αυτού της μονομεταβλητής ανάλυσης ως εξής

$$\gamma_Z = \Pr(Z'_\xi = 1 \mid Z_\xi = 2) \quad \text{και} \quad \delta_Z = \Pr(Z'_\xi = 2 \mid Z_\xi = 1). \qquad .$$

Χρησιμοποιώντας την παρακάτω σχέση

$$P_{Y(2)|X(k),Z(1)} = P_{Y(2)|X(k)} + [1 - P_{Z(1)|X(k)}][P_{Y(2)|X(k),Z(1)} - P_{Y(2)|X(k),Z(2)}]$$

μπορεί να αποδειχτεί ότι η αναμενόμενη τιμή της εκτιμήτριας $\hat{P}_{Y(2)|X(k),Z(1)}$ του πληθυσμιακού ποσοστού $P_{Y(2)|X(k),Z(1)}$ προσεγγίζεται για μεγάλα δείγματα από την ακόλουθη σχέση

$$E[\hat{P}_{Y(2)|X(k),Z'(1)}] = P_{Y(2)|X(k)} + \lambda_k \cdot [1 - P_{Z(1)|X(k)}][P_{Y(2)|X(k),Z(1)} - P_{Y(2)|X(k),Z(2)}]$$

όπου

$$\lambda_k = \frac{\varepsilon \cdot P_{Z(1)|X(k)}}{1 + \varepsilon \cdot P_{Z(1)|X(k)}} \quad \text{και} \quad \varepsilon = (1 - \delta_Z - \gamma_Z)/\gamma_Z$$

(βλ. Kuha and Skinner (1997)).Σε πρακτικό επίπεδο, περιμένουμε οι τιμές των πιθανοτήτων λανθασμένης ταξινόμησης δ_Z και γ_Z να είναι μικρότερες του 1/2 . Σε αυτήν την περίπτωση, ισχύει $0 \le \lambda_k \le 1$ και η αναμενόμενη τιμή του $\hat{P}_{Y(2)|X(k),Z'(1)}$ βρίσκεται ανάμεσα στις πραγματικές τιμές των ποσοστών $P_{Y(2)|X(k),Z(1)}$ και $P_{Y(2)|X(k)}$, με το τελευταίο ποσοστό να προκύπτει από την άθροιση των πραγματικών ποσοστών ως προς τις τιμές της μεταβλητής *Z*. Σε παρόμοια αποτελέσματα μπορούμε να προβούμε και για τις εκτιμήσεις άλλων ποσοστών της μορφής $P_{Y(2)|X(k),Z(l)}$ με $k,l = 1,2$. Η μεροληψία αυτής της μορφής είναι γνωστή στην επιδημιολογική βιβλιογραφία ως σύγχυση υπολοίπων (*residual confounding*) (βλ. Savitz and Barron (1989)). Προκύπτει λόγω του γεγονότος ότι η ανάλυση περιορίζεται σε λανθασμένα υποσύνολο μονάδων (σε αυτές για τις οποίες ισχύει $Z'_\xi = 1$ αντί για $Z_\xi = 1$, δηλαδή στη λανθασμένα ταξινομημένη μεταβλητή Z' αντί της πραγματικής Z) και λόγω αυτού η ετερογένεια των ποσοστών που οφείλεται στην μεταβλητή Z δεν μπορεί να ελεγχθεί πλήρως.

Λόγω της σύγχυσης κατάλοιπων, οι εκτιμήσεις για τις διαφορές των ποσοστών D_1 και D_2 και τα μέτρα συσχέτισης μεταξύ των μεταβλητών Y και X που είναι προσαρμοσμένα στα επίπεδα της Z μπορεί να παρουσιάζουν μεροληψία. Οι Savitz and Barron (1989) (βλ. επίσης

Greenland 1980) κατέληξαν στο ακόλουθο συμπέρασμα όσον αφορά το λόγο σχετικών πιθανοτήτων : σε έναν $2\times2\times2$ πίνακα συνάφειας η αναμενόμενη τιμή για ένα εκτιμώμενο 'περιληπτικό' μέτρο συσχέτισης ανάμεσα στις μεταβλητές Y και X δεδομένης της μεταβλητής Z έχει πεδίο τιμών ανάμεσα στο 'αδρό' και χωρίς περιορισμούς (στα επίπεδα της Z) μέτρο συσχέτισης και στο πραγματικό και υποκείμενο σε περιορισμούς 'περιληπτικό' μέτρο συσχέτισης. Ο Brenner (1993) δείχνει ότι το παραπάνω γεγονός δεν είναι γενικά αληθές όταν η μεταβλητή Z έχει περισσότερες από δύο κατηγορίες.

Η μεροληψία που οφείλεται στη σύγχυση υπολοίπων για τον υπολογισμό ενός εκτιμητή των διαφορών των ποσοστών D_1 και D_2 μπορεί να τείνει είτε προς είτε μακριά από τη 'μηδενική' τιμή. Ο εκτιμητής αυτός μπορεί συν τοις άλλοις να έχει διαφορετικό πρόσημο από τη πραγματική διαφορά, το οποίο και θα δείχνει μία συσχέτιση μεταξύ των Y και X που είναι αντίθετη της πραγματικής συσχέτισης μεταξύ των δύο μεταβλητών. Οι εκτιμήσεις αυτές μπορούν ακόμα να δείξουν ότι υπάρχει συσχέτιση όταν στην πραγματικότητα δεν υφίσταται ($D_1 = D_2 = 0$). Επομένως, τόσο το μέγεθος όσο και η ισχύς ενός ελέγχου μη συσχέτισης ανάμεσα στις Y και X είναι εσφαλμένα όταν η Z είναι λανθασμένα ταξινομημένη.

Το ανωτέρω απλό παράδειγμα λανθασμένης ταξινόμησης για έναν τριδιάστατο πίνακα συνάφειας μπορεί να επεκταθεί με πολλούς τρόπους. Κάποιες από τις μεταβλητές μπορεί να είναι πολυεπίπεδες αντί για διχοτομικές. Μπορεί να έχουμε ανεξάρτητη και μη-διαφορίσμη λανθασμένη ταξινόμηση για περισσότερες από μία μεταβλητές, περίπτωση κατά την οποία το αποτέλεσμα της λανθασμένης ταξινόμησης θα είναι ένας συνδυασμός 'εξασθένησης' και σύγχυσης υπολοίπων (Fung and Howe 1984). Εν τέλει, η λανθασμένη ταξινόμηση που είναι ταυτόχρονα μη ανεξάρτητη και διαφορίσιμη μπορεί να δημιουργήσει κάθε μορφής μεροληψία στις εκτιμήσεις (κάποια σχετικά παραδείγματα δίνονται από τους Greenland and Robins 1985).

2.3.2 Γενική Περίπτωση

Τα αποτελέσματα που προκύπτουν μπορούν να επεκταθούν σε πολυμεταβλητή ανάλυση με μεγαλύτερο αριθμό μεταβλητών. Όταν μερικές από τις μεταβλητές είναι λανθασμένα ταξινομημένες, δεν είναι γενικά εύκολο να δώσουμε ακόμα και ποσοτικές σχέσεις που να παριστούν την προκύπτουσα μεροληψία. Εξαίρεση αποτελεί μία χρήσιμη μελέτη του Korn (1981) που πραγματεύεται ελέγχους υποθέσεων για πίνακες συνάφειας με ανεξάρτητη και μη

διαφορίσιμη λανθασμένη ταξινόμηση για κάποιες από τις μεταβλητές. Ας υποθέσουμε ότι τα πληθυσμιακά ποσοστά ανά κελί ικανοποιούν ένα ιεραρχικό λογαριθμογραμμικό μοντέλο (εκτενέστερη αναφορά στην απεικόνιση ενός μοντέλου λανθασμένης ταξινόμησης μέσω λογαριθμογραμμικών μοντέλων και τρόποι διόρθωσης της λανθασμένης ταξινόμησης θα δοθούν στο Κεφάλαιο 6). Ένα τέτοιο μοντέλο μπορεί να προσδιοριστεί βάσει του καθορισμού των αλληλεπιδράσεων μεγαλύτερης τάξης των μεταβλητών που περιλαμβάνει (βλ. Bishop et al. 1975, Κεφ. 2). Για παράδειγμα, ένα μοντέλο τριών μεταβλητών για το οποίο δεν έχουμε συσχέτιση μεταξύ των Y και X δεδομένης μίας σταθερής τιμής της Z μπορεί να γραφεί ως $H(YZ, XZ)$ και ένα μοντέλο πλήρους ανεξαρτησίας ως $H(Y, X, Z)$. Τα αποτελέσματα της μελέτης του Korn αναφέρουν ότι ένα μοντέλο διατηρείται υπό την προϋπόθεση ύπαρξης λανθασμένης ταξινόμησης (*is preserved by misclassification*) της μεταβλητής που εμφανίζεται μόνο μία φορά στο σχηματιζόμενο μοντέλο. Ένα μοντέλο λέμε ότι διατηρείται όταν τα λανθασμένα ταξινομημένα δεδομένα ικανοποιούν επίσης το μοντέλο, με άλλα λόγια όταν η υπάρχουσα λανθασμένη ταξινόμηση δεν επιφέρει κίβδηλες συσχετίσεις ανάμεσα στις ταξινομημένες μεταβλητές. Κατά αυτόν τον τρόπο, το μοντέλο $H(YZ, XZ)$ διατηρείται κάτω την υπόθεση ύπαρξης λανθασμένης ταξινόμησης για τις μεταβλητές Y ή X ή και για τις δύο αλλά όχι κάτω από λανθασμένη ταξινόμηση για τη Z. Το συμπέρασμα αυτό συμφωνεί με τα άνωθεν αποτελέσματα που προέκυψαν κατά τη μελέτη των τριδιάστατων πινάκων συνάφειας. Ας υποθέσουμε στη συνέχεια ότι η μεταβλητή X ταξινομείται εσφαλμένα ως X'. Αφού δεν υπάρχει συσχέτιση μεταξύ των Y και X δεδομένης της Z, με άλλα λόγια δεν έχουμε αλληλεπιδράσεις που να περιέχουν τον όρο YX στο μοντέλο, δε θα υπάρξει συσχέτιση μεταξύ των μεταβλητών X' και Y δεδομένης της Z (δηλαδή καμία αλληλεπίδραση που να περιέχει τον όρο YX'). Αφ' ετέρου, μη διαφορίσιμη λανθασμένη ταξινόμηση για τη Z μπορεί να οδηγήσει στη δημιουργία κίβδηλης συσχέτισης μεταξύ των Y και X δεδομένης της Z' (μία αλληλεπίδραση της μορφής YX) και κίβδηλη ετερογένεια στις συσχετίσεις αυτές για τα επίπεδα της Z' (μία αλληλεπίδραση της μορφής YXZ').

Εάν ένα μοντέλο διατηρείται, μπορεί να ελεγχθεί για καλή προσαρμογή (*goodness of fit*) με τη χρήση των λανθασμένα ταξινομημένων τιμών. Το αποτέλεσμα του ελέγχου που προκύπτει έχει το σωστό επίπεδο σημαντικότητας αλλά μειωμένη ισχύ σε σχέση με τον έλεγχο που χρησιμοποιεί τις ορθά ταξινομημένες τιμές. Ένας τύπος που δίνεται από τον Korn (1982) μπορεί να χρησιμοποιηθεί για την εκτίμηση της ισχύος για ένα έλεγχο λόγου πιθανοφανειών (*likelihood ratio test*) για δύο εμφυτευμένα (*nested*) διατηρούμενα μοντέλα.

ΚΕΦΑΛΑΙΟ 3

Έλεγχοι υποθέσεων

3.1 Εισαγωγή

Πριν προχωρήσουμε στην παρουσίαση των μεθόδων για την αντιμετώπιση και διόρθωση της λανθασμένης ταξινόμησης σε έναν πληθυσμό που κατανέμεται σε υποομάδες βάσει κατηγορικών μεταβλητών, θα ήταν ουσιώδες να ερευνήσουμε το κατά πόσον το φαινόμενο αυτό επηρεάζει την εξαγωγή ορθών συμπερασμάτων για τα ζητούμενα χαρακτηριστικά του εξεταζόμενου πληθυσμού. Αυτό έχει βαρύνουσα σημασία, γιατί μπορεί να οδηγηθούμε σε εσφαλμένα συμπεράσματα που να ανταποκρίνονται ελάχιστα έως καθόλου στα πραγματικά δεδομένα. Ένας από τους πιο γνωστούς τρόπους διερεύνησης για τη προσαρμογή ενός μοντέλου, εν προκειμένω του μοντέλου της λανθασμένης ταξινόμησης, είναι η χρήση των ελέγχων υποθέσεων, συνήθως μέσω των ελέγχων καλής προσαρμογής.

Η χρήση των ελέγχων καλής προσαρμογής (*goodness of fit tests*) στην ανάλυση των κατηγορικών δεδομένων αποτελεί ένα πολύ σημαντικό εργαλείο της ερευνητικής διαδικασίας που αφορά τα κατηγορικά δεδομένα και έχει ερευνηθεί διεξοδικά από ένα μεγάλο σύνολο της επιστημονικής κοινότητας. Αποτελεί βασικό κομμάτι των ελέγχων υποθέσεων (*hypothesis testing*) για την προσαρμογή των παρατηρούμενων τιμών σε ένα δεδομένο μοντέλο που θέλουμε να υπόκειται το εξεταζόμενο δείγμα, για παράδειγμα το μοντέλο ανεξαρτησίας των παρατηρήσεων ή καλής προσαρμογής τους σε μία δεδομένη κατανομή[3]. Όπως γνωρίζουμε, Μία από τις δυσκολίες που συχνά αντιμετωπίζεται στην πρακτική εφαρμογή της θεωρίας των ελέγχων υποθέσεων είναι η πιθανότητα σφάλματος κατά την ταξινόμηση των υποκειμένων ενός πληθυσμού στις αντίστοιχες κατηγορίες που ανήκουν.

Στα πρώιμα στάδια της έρευνας για το ρόλο των διαγνωστικών ελέγχων για τη λανθασμένη ταξινόμηση κατηγορικών δεδομένων, η επίδραση της λανθασμένης ταξινόμησης για την εκτίμηση ενός δειγματικού δυωνυμικού ποσοστού ερευνήθηκε αρχικά από τον Bross

[3] Για τα κατηγορικά δεδομένα οι συνήθεις κατανομές που ακολουθούνται είναι η Δυωνυμική, η Υπεργεωμετρική, η Πολυωνυμική, η Αρνητική Δυωνυμική και η κατανομή Poisson.

(1954) υπό την προϋπόθεση ενός στοχαστικού μοντέλου λανθασμένης ταξινόμησης. Μέσω της μελέτης του, κατέληξε στα ακόλουθα συμπεράσματα :

- ο δειγματικός εκτιμητής της είναι μεροληπτικός και ότι η μεροληψία του, εφόσον εξαρτάται από το μέγεθος της λανθασμένης ταξινόμησης, είναι σημαντική.
- εάν ο ίδιος μηχανισμός λανθασμένης ταξινόμησης χρησιμοποιηθεί σε δύο διαφορετικούς πληθυσμούς η εγκυρότητα του ελέγχου σημαντικότητας στους 2×2 πίνακες συνάφειας δεν θα επηρεαστεί, όμως η ισχύς του ελέγχου θα μειωθεί.

Επίσης οι Diamond and Lilienfeld (1962 a, b) και Newell (1962) ασχολήθηκαν με τις επιδράσεις των σφαλμάτων λανθασμένης ταξινόμησης στις επιδημιολογικές μελέτες.

Κατά την αντιμετώπιση περιπτώσεων λανθασμένης ταξινόμησης στις παρατηρούμενες τιμές μίας κατηγορικής μεταβλητής, προκύπτουν τα ακόλουθα ερωτήματα :

- Με ποιον τρόπο μπορούμε να λάβουμε υπ'όψιν τα σφάλματα λανθασμένης ταξινόμησης στους 'συνήθεις' ελέγχους υποθέσεων ;
- Αν τα σφάλματα λανθασμένης ταξινόμησης αγνοηθούν, με ποιον τρόπο θα επηρεαστεί το επίπεδο σημαντικότητας του ελέγχου;
- Ποια θα είναι η επίδραση των σφαλμάτων αυτών στην ισχύ του ελέγχου;

Σε αυτά τα βασικά ερωτήματα πρώτοι οι Mote and Anderson (1965) και βασιζόμενοι στη μεθοδολογία τους μια πλειάδα άλλων συγγραφέων προσπάθησαν να δώσουν απαντήσεις. Χωρίς βλάβη της γενικότητας και χάριν ευκολίας, θα παρατεθεί η μεθοδολογία για την περίπτωση ύπαρξης λανθασμένης ταξινόμησης ενός πληθυσμού για μία κατηγορική μεταβλητή, δηλαδή για τη μονομεταβλητή περίπτωση. Η ίδια συμπερασματολογία ισχύει σε γενικές γραμμές και κατά τη διμεταβλητή (βλ. Mote and Anderson (1965)) και πολυμεταβλητή περίπτωση. Χάριν ευκολίας, οι συμβολισμοί που θα χρησιμοποιηθούν είναι γενικοί (π.χ. για το δειγματικό ποσοστό της μεταβλητής Y χρησιμοποιούμε το συμβολισμό p αντί να του συμβολισμού P_Y που ακολουθείται σε όλο το σύγγραμμα).

3.2 Επίδραση λανθασμένης ταξινόμησης στους ελέγχους υποθέσεων

Έστω ότι ο εξεταζόμενος πληθυσμός ταξινομείται σε r κατηγορίες, για τον οποίο και σχηματίζουμε τον αντίστοιχο πίνακα συχνοτήτων των παρατηρήσεων. Ορίζουμε ως p_{oi} την τιμή της πιθανότητας μίας παρατήρησης του πληθυσμού να ταξινομείται στην i κατηγορία, με τις ακόλουθες ιδιότητες

$$p_{oi} > 0 \quad \text{και} \quad \sum_{i=1}^{r} p_{oi} = 1. \tag{3.1}$$

Για ένα δείγμα του πληθυσμού μεγέθους N συμβολίζουμε με n_i την παρατηρούμενη συχνότητα για την i κατηγορία, με ιδιότητα

$$\sum_{i=1}^{r} n_i = N .$$

Για την περίπτωση μη ύπαρξης σφαλμάτων στις παρατηρούμενες τιμές, το κριτήριο του X^2 - ελέγχου είναι : απορρίπτουμε την H_0 εάν

$$X^2 = \sum_{i=1}^{r} x_i^2 > C_1 = c_a(r-1),$$

όπου

$$x_i = (n_i - Np_{oi}) \Big/ \sqrt{(Np_{oi})}$$

και η ποσότητα $C_1 = c_a(r-1)$ έχει επιλεχθεί έτσι ώστε να ικανοποιείται η σχέση $\Pr[\chi^2_{(r-1)} \geq C_1] = a$, με $\chi^2_{(r-1)}$ να είναι η συνήθης κατανομή χ^2 με $(r-1)$ βαθμούς ελευθερίας.

Η ύπαρξη λανθασμένης ταξινόμησης τροποποιεί το ανωτέρω κριτήριο ελέγχου. Η τροποποίηση αυτή οδηγεί στον ορισμό των ακόλουθων πιθανοτήτων

- πιθανότητα λανθασμένης ταξινόμησης α_{ji} $(j \neq i)$, την πιθανότητα δηλαδή μίας παρατήρησης να έχει ταξινομηθεί στην j κατηγορία ενώ πραγματικά ανήκει στην i
- πιθανότητα ορθής ταξινόμησης α_{ii}, την πιθανότητα δηλαδή μίας παρατήρησης που πραγματικά ανήκει στην i κατηγορία να έχει ταξινομηθεί ορθώς στην κατηγορία αυτή,

και ισχύει $\sum_{j=1}^{r} \alpha_{ji} = 1$ για κάθε $i = 1, \ldots, r$.

Ορίζουμε επίσης ως p_i την πραγματική πιθανότητα ταξινόμησης μίας μονάδας του εξεταζόμενου δείγματος στην i κατηγορία και p_i' την πιθανότητα ταξινόμησης μίας μονάδας στην i κατηγορία κατά την περίπτωση που υπάρχουν σφάλματα ταξινόμησης, με $\sum_{i=1}^{r} p_i = 1$ και $\sum_{i=1}^{r} p_i' = 1$. Από τις παραπάνω σχέσεις προκύπτει ότι

$$p_i' = \sum_{k=1}^{r} \alpha_{jk} p_k = 1, \quad i = 1, \ldots, r . \tag{3.2}$$

Αναπαριστώντας τις πιθανότητες λανθασμένης ταξινόμησης α_{ji} σε έναν $r \times r$ πίνακα λανθασμένης ταξινόμησης **A,** όπου $\mathbf{A} = [\alpha_{ji}]_{(r \times r)}$, και θεωρώντας τα διανύσματα $\mathbf{p} = [p_1, p_2, ..., p_r]$ και $\mathbf{p}' = [p_1', p_2', ..., p_r']$, η σχέση (4.1) γίνεται

$$\mathbf{p}' = \mathbf{A}\mathbf{p} \Leftrightarrow [p_1', p_2', ..., p_r'] = [\alpha_{ji}]_{(r \times r)} [p_1, p_2, ..., p_r]$$

Για τις περιπτώσεις που δεν υφίστανται (ή αγνοούνται) σφάλματα ταξινόμησης, καθώς και όταν αυτά λαμβάνονται υπ' όψιν, θεωρούμε τους ακόλουθους έλεγχους υποθέσεων αντίστοιχα

$$H_0 : p_i = p_{oi} \qquad \text{και} \qquad H_0' : p_i' = p_{oi}' \qquad (i = 1, ..., r),$$

όπου $$p_{0i}' = \sum_{k=1}^{r} \alpha_{jk} p_{ok} = 1, \quad i = 1, ..., r.$$

Ως εκ τούτου, η απόρριψη ή αποδοχή της μηδενικής υπόθεσης H_0 ισοδυναμεί με την απόρριψη ή αποδοχή της μηδενικής υπόθεσης H_0' εάν ο πίνακας **A** έχει αντίστροφο (είναι μη ιδιάζων).

Εξετάζουμε τις ακόλουθες περιπτώσεις

A. Ο πίνακας **A** είναι μη ιδιάζων και γνωστός

B. Ο πίνακας **A** είναι μη ιδιάζων και άγνωστος.

Κατά τη δεύτερη περίπτωση, οι r πιθανότητες $\{p_1', p_2', ..., p_r'\}$ εκτιμώνται μέσω $r(r\text{-}1)$ άγνωστων παραμέτρων. Επειδή ο αριθμός των αγνώστων παραμέτρων είναι μεγαλύτερος του αριθμού των κατηγοριών, οδηγούμαστε σε εκφυλισμένες λύσεις και το γενικό πρόβλημα δεν μπορεί να λυθεί. Για τη μείωση του αριθμού των αγνώστων παραμέτρων, θα καταφύγουμε στην εξέταση ορισμένων ειδικών μορφών του πίνακα λανθασμένης ταξινόμησης **A**.

A. Ο πίνακας λανθασμένης ταξινόμησης A είναι μη ιδιάζων και γνωστός

Το κριτήριο ελέγχου της μηδενικής υπόθεσης H_0' (και συνεπώς της μηδενικής υπόθεσης H_0) είναι : απορρίπτουμε την H_0' εάν

$$X'^2 = \sum_{i=1}^{r} [(n_i - N p_{oi}')^2 / (N p_{oi}')] \geq C_1.$$

Α.1 *Παράβλεψη λανθασμένης ταξινόμησης*

Έστω ότι οι παρατηρούμενες τιμές για τις εξεταζόμενες μονάδες του πληθυσμού υπόκεινται σε σφάλματα ταξινόμησης, τα οποία όμως επιλέγουμε να μη λάβουμε υπ' όψιν. Τότε έχουμε τον εξής έλεγχο καλής προσαρμογής για την ταξινόμηση των δεδομένων : απορρίπτουμε την $H_0 : p_i = p_{oi}$, $(i = 1,...,r)$ εάν $X^2 \geq C_1$. Για την εκτίμηση του πραγματικού μεγέθους του ελέγχου αυτού, θα βασιστούμε στη μεθοδολογία που αναπτύχθηκε από τους Pitman (1948), Cochran (1952) και Mitra (1958).

Έστω ένα σύνολο παραμέτρων απόκλισης, $\{\nu_i$, με $\nu_i \neq 0$ έστω για ένα $1 \leq k \leq r\}$ τέτοιο ώστε $\sum_{i=1}^{r} \nu_i = 0$. Ορίζουμε επίσης την ενάντια υπόθεση H_{1N}, με

$$H_{1N} : p_i = p_{oi} + \nu_i / \sqrt{N} .$$

Τότε η ασυμπτωτική ισχύς του ελέγχου χωρίς λανθασμένη ταξινόμηση είναι

$$\beta = \lim_{N \to \infty} \Pr\{X^2 \geq c_a(r-1) \mid p_i = p_{oi} + \nu_i / \sqrt{N}\} = 1 - F(C_1, r-1, \lambda) ,$$

όπου
$$\lambda = \sum_{i=1}^{r} (\nu_i^2 / p_{oi}) = N \sum_{i=1}^{r} [(n_i - p_{oi})^2 / p_{oi}] \qquad (3.3)$$

και $F(C_1, r-1, \lambda)$ είναι η αθροιστική μη κεντρική συνάρτηση κατανομής χ^2 με $r-1$ β.ε. και παράμετρο εκκεντρότητας λ.

Βάσει των ανωτέρω, το πραγματικό μέγεθος του ελέγχου (σφάλμα τύπου Ι) κατά την ύπαρξη λανθασμένης ταξινόμησης είναι

$$a' = 1 - F(C_1, r-1, \lambda_0') ,$$

όπου
$$\lambda_0' = N \sum_{i=1}^{r} [(p_{oi}' - p_{oi})^2 / p_{oi}] \equiv \sum_{i=1}^{r} (\nu_{oi}'^2 / p_{oi}) .$$

Οι τιμές των παραμέτρων λ_0' έχουν υπολογιστεί από τον Fix (1949) για δεδομένες τιμές των a, a' και $r-1$.

Πολλοί τύποι σφαλμάτων ταξινόμησης που παρουσιάζουν πρακτικό ενδιαφέρον μπορούν να εκφραστούν μέσω της εισαγωγής μίας άγνωστης παραμέτρου θ , τέτοια ώστε οι $r(r-1)$ ανεξάρτητες πιθανότητες a_{ji} του πίνακα λανθασμένης ταξινόμησης **Α** να συμβολίζονται ως εξής

$$a_{ji} = \begin{cases} \theta & (i \neq j) \\ 1-(r-1)\theta & (i = j, 0 < \theta < 1/r) \end{cases} . \qquad (3.4)$$

Στην περίπτωση αυτή, ορίζουμε ως ποσοστό λανθασμένης ταξινόμησης (*misclassification rate*) την ποσότητα $(r-1)\theta$. Βάσει της ανωτέρω σχέσης, έχουμε

$$p'_{oi} = p_{oi} + (1 - rp_{oi})\theta \qquad \text{και} \qquad \nu'_{oi} = (1 - rp_{oi})\theta\sqrt{N} \tag{3.5}$$

και η παράμετρος εκκεντρότητας λ'_0 γίνεται

$$\lambda'_0 = N\theta^2 \sum_{i=1}^{r} [(1 - rp_{oi})^2 / p_{oi}] = N\theta^2 \left[\sum_{i=1}^{r} (1/p_{oi}) - r^2 \right]$$

Εάν ισχύει $p_{oi} = 1/r$ για κάθε πιθανότητα p_{oi} προκύπτει ότι $\alpha = \alpha'$, δηλαδή το μέγεθος του ελέγχου παραμένει ίδιο ακόμα και όταν υπάρχει λανθασμένη ταξινόμηση. Στην περίπτωση όμως που δεν ισχύει αυτό, προκύπτει ότι $\lambda'_0 > 0$ και $\alpha' > \alpha$, το οποίο και σημαίνει ότι η ύπαρξη σφαλμάτων ταξινόμησης οδηγεί σε αύξηση του μεγέθους του ελέγχου. Αυτό καταδεικνύεται και στο παράδειγμα που παρουσιάζεται από τους Mote and Anderson (1965) (Πίνακας 3.1)

ΠΙΝΑΚΑΣ 3.1

Μέγεθος της παραμέτρου λανθασμένης ταξινόμησης $\theta\sqrt{N}$ που απαιτείται για την αύξηση του μεγέθους ελέγχου από α σε α' με μέγεθος δείγματος $r = 4$ και $N = 100$

[ΠΗΓΗ : Mote and Anderson, 1965]

		α	0·05			0·01
		α'	0·10	0·20	0·30	0·10
p_{oj}	$\lambda'_0/(N\theta^2)$	λ'_0	·779	2·096	3·302	2·763
$\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}$	$\frac{112}{9}$		0·250	0·410	0·515	0·471
$\frac{3}{10}, \frac{3}{10}, \frac{3}{10}, \frac{1}{10}$	4	$\theta\sqrt{N}$	·441	·724	·909	·831
$\frac{2}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}$	$\frac{3}{2}$		·721	1·182	1·484	1·357

όπου βλέπουμε ότι ένα ποσοστό λανθασμένης ταξινόμησης της τάξεως του 7.5% οδηγεί σε αύξηση του μεγέθους του ελέγχου από 0.05 σε 0.10, ενώ για ποσοστό 14.5% παρατηρούμε αύξηση του μεγέθους ελέγχου από 0.01 σε 0.10 . Θα πρέπει να σημειωθεί πως το μέγεθος του επιλεγομένου δείγματος του πληθυσμού θα πρέπει να είναι τέτοιο ώστε να ισχύει $\theta < 1/r$.

A.2 Επίδραση της λανθασμένης ταξινόμησης στην ασυμπτωτική ισχύ των ελέγχων

Βάσει της υπόθεσης ότι ο πίνακας λανθασμένης ταξινόμησης είναι γνωστός ούτως ώστε ο στατιστικός έλεγχος να γίνεται μέσω του X'^2, ορίζουμε ως ασυμπτωτική ισχύ του ελέγχου κατά την ύπαρξη λανθασμένης ταξινόμησης την ποσότητα

$$\beta' = \lim_{N\to\infty} \Pr\{X'^2 \geq C_1 \mid p_i' = p_{oi}' + \nu_i'/\sqrt{N}\} = 1 - F(C_1, r-1, \lambda'),$$

όπου $$\lambda' = \sum_{i=1}^{r} (\nu_i'^2 / p_{oi}') = N \sum_{i=1}^{r} [(p_i' - p_{oi}')^2 / p_{oi}'].$$

Σκοπός μας είναι να συγκριθεί η ασυμπτωτική ισχύς β' με την ασυμπτωτική ισχύ β χωρίς λανθασμένη ταξινόμηση $\beta = 1 - F(C_1, r-1, \lambda)$, όπου η παράμετρος λ δίνεται από τη σχέση (3.3).

Λόγω του ότι οι συναρτήσεις β και β' είναι αυστηρά μονοτονικές ως προς τις παραμέτρους λ και λ' αντίστοιχα, αρκεί να εξετάσουμε τη σχέση μεταξύ των παραμέτρων για να αποφανθούμε. Ο Mote (1957) με τη χρήση της ανισότητας Schwartz απέδειξε ότι $\lambda' \leq \lambda$. Συνεπώς αποδεικνύεται ότι η ύπαρξη λανθασμένης ταξινόμησης μειώνει την ισχύ του ελέγχου. Το ανωτέρω συμπέρασμα μπορεί να γίνει κατανοητό μέσω ενός παραδείγματος που χρησιμοποιούν οι Mote and Anderson (1965, §2.1.2) (Πίνακας 3.2), όπου οι πιθανότητες του πίνακα λανθασμένης ταξινόμησης **A** εκφράζονται μέσω της σχέσης (4.3) και ισχύει $p_{oi} = 1/r$ για κάθε πιθανότητα p_{oi}. Παρατηρούμε ότι ένα ποσοστό λανθασμένης ταξινόμησης της τάξεως του 9% για μέγεθος ελέγχου ίσο με $a = 0.05$ οδηγεί σε μείωση της ισχύος του ελέγχου από 0.90 σε 0.80 ή από 0.50 σε 0.40, ενώ για μέγεθος ελέγχου ίσο με $a = 0.01$ όμοια μείωση ισχύος παρουσιάζεται για ποσοστά λανθασμένης ταξινόμησης της τάξης του 8% και 7% αντίστοιχα. Βάσει των ανωτέρω, μπορούμε να συμπεράνουμε ότι η κλασματική απώλεια της ισχύος για ένα δεδομένο ποσοστό λανθασμένης ταξινόμησης αυξάνεται όσο μειώνεται το μέγεθος ελέγχου a και η ασυμπτωτική ισχύς β.

ΠΙΝΑΚΑΣ 3.2

Επίδραση του ποσοστού λανθασμένης ταξινόμησης *(r-1)θ* στην ασυμπτωτική ισχύ του έλεγχου όταν έχουμε $H_0 : p_{0i} = 1/r$ και $r = 5$ κατηγορίες

[ΠΗΓΗ : Mote and Anderson, 1965]

β'	λ'	4θ	λ'	4θ
	$\beta = 0{\cdot}9$			
	$\alpha = 0{\cdot}05; \lambda = 15{\cdot}405$		$\alpha = 0{\cdot}01; \lambda = 20{\cdot}737$	
0·8	11·935	0·096	16·749	0·081
0·7	9·683	·166	14·121	·140
0·6	7·924	·226	12·039	·190
0·5	6·420	·284	10·231	·238
0·4	5·050	·342	8·557	·286
0·3	3·737	·406	6·914	·338
0·2	2·401	·484	5·188	·400
0·1	0·910	·606	3·149	·488
	$\beta = 0{\cdot}5$			
	$\alpha = 0{\cdot}05; \lambda = 6{\cdot}420$		$\alpha = 0{\cdot}01; \lambda = 10{\cdot}231$	
0·4	5·050	0·090	8·557	0·068
0·3	3·737	·190	6·914	·142
0·2	2·401	·311	5·188	·230
0·1	0·910	·499	3·149	·356

Β. Ο πίνακας λανθασμένης ταξινόμησης Α είναι μη ιδιάζων και άγνωστός

Θεωρούμε την περίπτωση όπου οι πιθανότητες λανθασμένης ταξινόμησης δίνονται από τη σχέση (4.3), με την παράμετρο θ να είναι άγνωστη. Εύκολα μπορεί να διαπιστώσει κανείς πως ο πίνακας λανθασμένης ταξινόμησης **A** είναι μη ιδιάζων για κάθε τιμή της θ με $0 < \theta < 1/r$. Έστω ότι δύο από τα πραγματικά πληθυσμιακά ποσοστά p_0 είναι διάφορα του $1/r$. Αν θεωρηθεί απαραίτητο και μπορεί να γίνει για το επιλεγμένο δειγματοληπτικό σχήμα, αναδιατάσσουμε τις κατηγορίες που κατατάσσονται οι μονάδες του δείγματος ούτως ώστε να ισχύει $p_0 \neq 1/r$ για την πρώτη r_0 κατηγορία, με $2 \leq r_0 \leq r$. Από τις σχέσεις (3.1), (3.4) και (3.5) προκύπτουν τα ακόλουθα

- ✓ $p'_{oi} > 0$ και $\sum_{i=1}^{r} p'_{oi} = 1$
- ✓ p'_{oi} είναι μία συνεχής συνάρτηση της παραμέτρου θ
- ✓ οι μερικές παράγωγοι πρώτης και δεύτερης τάξης $\partial p'_{oi}/\partial\theta$ και $\partial^2 p'_{oi}/\partial\theta^2$ υπάρχουν και είναι συνεχείς
- ✓ το διάνυσμα των μερικών παραγώγων $(\partial p'_{o1}/\partial\theta, ..., \partial p'_{or}/\partial\theta)$ είναι τάξης 1

Ως εκ τούτου, βάσει του θεωρήματος του Cramer (1946), ισχύει ότι το στατιστικό

$$X'^2 = \sum_{i=1}^{r} [(n_i - N\hat{p}'_{oi})^2 / (N\hat{p}'_{oi})]$$

ακολουθεί μία ασυμπτωτική κατανομή χ^2 με $r-2$ β.ε.. Τα εκτιμώμενα ποσοστά $\hat{p}'_{oi}$ προκύπτουν εάν στη σχέση (3.5) αντικαταστήσουμε την παράμετρο θ με μία εκτίμηση της $\hat{\theta}$ που προκύπτει είτε μέσω του τροποποιημένου ελαχίστου χ^2 του Cramer είτε με το ελάχιστο χ^2 του Neyman. Με χρήση του $\hat{\theta}$ ο έλεγχος υποθέσεων διαμορφώνεται ως εξής

- απορρίπτουμε την H'_0 (και συνεπώς την H_0) εάν ισχύει $X'^2 \geq C_2$, με C_2 μία σταθερά τέτοια ώστε να ισχύει $\Pr(\chi^2_{r-2} \geq C_2) = \alpha$.

Όμοια με προηγουμένως, η ασυμπτωτική ισχύς του ελέγχου είναι

$$\beta' = 1 - F(C_2, r-1, \lambda')$$

με

$$\lambda' = (1-r\theta)^2 \left[\sum_{i=1}^{r} (v_i'^2 / p'_{oi}) - \left(\sum_{i=1}^{r_o} v_i (1-rp_{oi}) / p'_{oi} \right)^2 \Bigg/ \left(\sum_{i=1}^{r_o} \left[v_i (1-rp_{oi})^2 / p'_{oi} \right] \right) \right]$$

Εάν είναι γνωστό ότι δεν υπάρχει λανθασμένη ταξινόμηση, η ασυμπτωτική ισχύς είναι

$$\beta = 1 - F(C_1, r-1, \lambda)$$

Ο Mote (1957) αποδεικνύει ότι ισχύει $\lambda' \leq \lambda$. Συνεπώς, σε γενικές γραμμές η ασυμπτωτική ισχύς του ελέγχου μειώνεται κατά την ύπαρξη λανθασμένης ταξινόμησης.

Στις περισσότερες των περιπτώσεων ύπαρξης λανθασμένης ταξινόμησης δεν είναι πρακτικά λογικό ούτε ρεαλιστικό να υποθέσουμε ότι οι πιθανότητες λανθασμένης ταξινόμησης α_{ji} είναι σταθερές και ίδιες για κάθε στοιχείο του δείγματος, δηλαδή δεν μπορεί να ισχύει $\alpha_{ji} = \theta,$ για $i \neq j$. Οι Mote και Anderson (1965) εξετάζοντας την περίπτωση που

παρατηρείται λανθασμένη ταξινόμηση σε δεδομένα γειτονικών κατηγοριών με τις πιθανότητες λανθασμένης ταξινόμησης να εξαρτώνται από την απόσταση μεταξύ των κατηγοριών, καταλήγουν στα ίδια αποτελέσματα για την επιρροή της στο μέγεθος και την ισχύ του ελέγχου.

Επεκτείνοντας τη ανάλυση για την επίδραση της λανθασμένης ταξινόμησης από την μονομεταβλητή στη διμεταβλητή περίπτωση, ακολουθώντας την ίδια μεθοδολογία καταλήγουμε στο ίδιο συμπέρασμα για την ισχύ του ελέγχου προσαρμογής των δεδομένων : η ύπαρξη εσφαλμένα ταξινομημένων δεδομένων οδηγεί γενικότερα στη μείωση της ισχύος του ελέγχου. Για το μέγεθος του ελέγχου θα πρέπει να διακρίνουμε τις παρακάτω περιπτώσεις, ορίζοντας Y και X τις κατηγορικές μεταβλητές που ταξινομούνται εσφαλμένα από τις Y' και X' αντίστοιχα :

- εάν υπάρχει μη διαφορίσιμη λανθασμένη ταξινόμηση ως προς την Y ή X μεταβλητή, το μέγεθος του ελέγχου για τα εσφαλμένα ταξινομημένα δεδομένα δεν μεταβάλλεται σε σχέση με το μέγεθος του ελέγχου απουσία της λανθασμένης ταξινόμησης (βλ. Mote and Anderson (1965), Gladen and Rogan (1979))
- εάν υπάρχει μη διαφορίσιμη και ανεξάρτητη λανθασμένη ταξινόμηση και στις δύο μεταβλητές, το μέγεθος του ελέγχου επίσης δεν μεταβάλλεται (βλ. Assakul and Proctor (1967))
- εάν δεν ισχύει κάποια από τις ανωτέρω προϋποθέσεις, το μέγεθος του ελέγχου κατά την ύπαρξη λανθασμένης ταξινόμησης μεταβάλλεται και συνήθως αυξάνει

Γενικότερα, μπορούμε να πούμε ότι για κάθε περίπτωση λανθασμένης ταξινόμησης με πιθανότητες λανθασμένης ταξινόμησης που είναι διαφορετικές για κάθε στοιχείο των εξεταζόμενων μεταβλητών, η επίδραση της λανθασμένης ταξινόμησης οδηγεί γενικότερα σε μείωση της ισχύος του ελέγχου, ενώ η επίδραση του στο μέγεθος του ελέγχου εξαρτάται από τις σχέσεις που υπάρχουν ανάμεσα στις προς εξέταση μεταβλητές.

3.3 Διόρθωση ελέγχων υποθέσεων με χρήση διπλής δειγματοληψίας για πολυωνυμική κατανομή

Όπως είδαμε, κατά την περίπτωση ύπαρξης σφαλμάτων των παρατηρήσεων στην ταξινόμηση η αγνόηση τους επηρεάζει το μέγεθος και η ισχύς του ελέγχου επηρεάζονται σε

σημαντικό βαθμό. Θεωρώντας εφεξής ότι τα παρατηρούμενα δεδομένα υπόκεινται σε λανθασμένη ταξινόμηση, θα καταφύγουμε στη χρήση της μεθόδου της διπλής δειγματοληψίας (Tenenbein (1970, 1971, 1972), Hochberg (1977)), αποσκοπώντας στην καλύτερη προσαρμογή των ελέγχων για τα δεδομένα αυτά.

Έστω ότι ο εξεταζόμενος πληθυσμός ταξινομείται σε r κατηγορίες, για τον οποίο και σχηματίζουμε τον αντίστοιχο πίνακα συχνοτήτων. Ορίζουμε ως p_{oj} την τιμή της πιθανότητας ενός κελιού που προσδιορίζεται από την μηδενική υπόθεση H_0 για μία παρατήρηση του πληθυσμού που ανήκει στην j κατηγορία, με ιδιότητες όπως αυτές ορίζονται στην Παράγραφο 3.2. Αν για ένα δείγμα του πληθυσμού μεγέθους N συμβολίζουμε με n_j την παρατηρούμενη συχνότητα για την j κατηγορία, για την περίπτωση μη ύπαρξης σφαλμάτων στις παρατηρούμενες τιμές, το κριτήριο του X^2 ελέγχου είναι να απορρίψουμε την H_0 εάν

$$X^2 = \sum_{j=1}^{r} x_j^2 > c_a(r-1),$$

όπου

$$x_j = (n_j - Np_{oj}) \Big/ \sqrt{Np_{oj}}$$

και η ποσότητα $c_a(r-1)$ έχει επιλεχθεί έτσι ώστε να ικανοποιείται η σχέση $\Pr[\chi^2_{(r-1)} \geq c_a(r-1)] = a$, με $\chi^2_{(r-1)}$ να είναι η συνήθης κατανομή χ^2 με $(r-1)$ βαθμούς ελευθερίας.

Αρχικά λαμβάνεται ένα δείγμα N μονάδων από τον εξεταζόμενο πληθυσμό το οποίο και ταξινομείται σε r διακριτές κατηγορίες με μονοσήμαντο τρόπο. Για κάθε δειγματική μονάδα ορίζονται οι τυχαίες μεταβλητές Y και Y^0 ως εξής :

- $Y = i$, εάν η δειγματική μονάδα ανήκει πραγματικά στην κατηγορία i, $i = 1,...,r$.
- $Y^0 = j$, εάν η δειγματική μονάδα ταξινομείται εσφαλμένα στην κατηγορία j, $j = 1,...,r$.

Οι περιθώριες κατανομές των Y και Y^0 είναι

$$p_i = \Pr(Y = i) \quad \text{και} \quad p'_j = \Pr(Y^0 = j)$$

με $\sum_{i=1}^{r} p_i = 1$ και $\sum_{j=1}^{r} p'_j = 1$. Για την περιγραφή της λανθασμένης ταξινόμησης ορίζουμε την πιθανότητα λανθασμένης ταξινόμησης α_{ji}, την πιθανότητα δηλαδή μίας μονάδας να έχει ταξινομηθεί στην j κατηγορία ενώ πραγματικά ανήκει στην i σύμφωνα με τη σχέση

$$\alpha_{ji} = \Pr(Y^0 = j \mid Y = i)$$

με ιδιότητες $\sum_{j=1}^{r} \alpha_{ji} = 1$ και $p'_j = \sum_{i=1}^{r} \alpha_{ji} p_i = 1, \quad j = 1,...,r$.

Βάσει της μεθόδου διπλής δειγματοληψίας, για όλο το επιλεγμένο δείγμα είναι γνωστές οι τιμές εσφαλμένης ταξινόμησης Y^0. Θεωρώντας ένα εσωτερικό δείγμα επικύρωσης μεγέθους n $(n < N)$ για το οποίο γνωρίζουμε τις πραγματικές τιμές ταξινόμησης της μεταβλητής Y, προκύπτει το ακόλουθο δείγμα

$$\{(Y_i^0, Y_i)\ ,\ i = 1, 2, ..., n\}$$

Έτσι, μία αναπαράσταση των διαθέσιμων δεδομένων προς επεξεργασία είναι η ακόλουθη

$$(Y_1^0, Y_1), ..., (Y_n^0, Y_n), Y_{n+1}^0, ..., Y_N^0$$

Ακολούθως, συμβολίζουμε με n_{ji} τις μονάδες του εσωτερικού δείγματος επικύρωσης με πραγματική κατηγορία i και εσφαλμένη κατηγορία j. Τα n_{ji} ικανοποιούν τις σχέσεις $n_{j+} = \sum_{i=1}^{r} n_{ji} = 1$ και $n_{+i} = \sum_{j=1}^{r} n_{ji} = 1$. Για τις m_k μονάδες του κυρίως δείγματος μεγέθους $N - n$ που ταξινομούνται μόνο μέσω της εσφαλμένης διαδικασίας ταξινόμησης ισχύει η ακόλουθη σχέση

$$m_k = \sum_{j=n+1}^{N} I(Y^0 = k),$$

όπου $I(\cdot)$ η δείκτρια συνάρτηση. Κατά αυτόν τον τρόπο, ισχύουν οι σχέσεις $\sum_{j=1}^{r} \sum_{i=1}^{r} n_{ji} = 1$ και $\sum_{k=1}^{r} m_k = N - n$.

ΠΙΝΑΚΑΣ 3.3

Ταξινόμηση δειγματικών μονάδων μέσω διπλής δειγματοληψίας

Ορθή ταξινόμηση

Εσφαλμένη ταξινόμηση		*Δείγμα επικύρωσης*					*Κυρίως δείγμα*
		1	***2***	***…***	***r***		
	1	n_{11}	n_{12}	…	n_{1r}	$n_{1\cdot}$	m_1
	2	n_{21}	n_{ij}	…	n_{ij}	$n_{2\cdot}$	m_2
	⋮	⋮				⋮	⋮
		n_{r1}	n_{r2}	…	n_{ij}	$n_{r\cdot}$	m_r
	r	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot r}$	n	$N-n$

Βάσει των παρατηρούμενων δεδομένων, η από κοινού συνάρτηση πιθανοφάνειας είναι

$$L(\underset{\sim}{p}, \mathbf{A}) = \{\prod_{i=1}^{r}\prod_{j=1}^{r}(\alpha_{ji}p_i)^{n_{ji}}(p'_j - \alpha_{ji}p_i)^{n_{j\cdot}-n_{ji}}\}\cdot\{\prod_{k=1}^{r}p_k'^{m_k}\},$$

με $\underset{\sim}{p} = (p_1, p_2, ..., p_r)'$ και $\mathbf{A} = [\alpha_{ji}]_{r\times r}$. Σύμφωνα με τον Tenenbein (1972), οι εκτιμητές μέγιστης πιθανοφάνειας είναι

$$\hat{p}_i = \sum_{j=1}^{r}\{(m_j + n_{j\cdot})n_{ji}/Nn_{j\cdot}\}$$

και

$$\hat{\alpha}_{ji} = (m_j + n_{j\cdot})n_{ji}/(Nn_{j\cdot}\hat{p}_i).$$

Οι εκτιμητές μέγιστης πιθανοφάνειας μπορεί να έχουν ενδιαφέρουσα ερμηνεία (βλ. Cheng et al (1998). Ας υποθέσουμε ότι θεωρούνται οι ακόλουθες δεσμευμένες αναμενόμενες τιμές

$$\begin{aligned} E[I(Y=i)\,|\,Y^0=j] &= \Pr(Y=i\,|\,Y^0=j) \\ &= \alpha_{ji}p_i/p'_j \\ &= \lambda_{ji}\ . \end{aligned}$$

Τότε γενικά μπορεί να γραφεί

$$E[I(Y=i)\,|\,Y^0]=\sum_{j=1}^{r}\lambda_{ji}I(Y^0=j)\,.$$

Εφόσον ο εκτιμητής μέγιστης πιθανοφάνειας των λ_{ji} είναι $\hat{\lambda}_{ji}=n_{ji}/n_{j.}$, η αναμενόμενη τιμή $E[I(Y=i)\,|\,Y^0]$ μπορεί να εκτιμηθεί από την ακόλουθη σχέση

$$\hat{E}[I(Y=i)\,|\,Y^0]=\sum_{j=1}^{r}\lambda_{ji}I(Y^0=j)\,,\quad i=1,2,...,r\,.$$

Συνέπεια αυτού είναι ότι για δεδομένα που υπόκεινται σε λανθασμένη ταξινόμηση, οι 'παρατηρούμενες συχνότητες' για την *i* κατηγορία εκτιμώνται ως εξής

$$\hat{n}_i=\sum_{j=1}^{n}I(Y_j=i)+\sum_{k=n+1}^{N}\hat{E}[I(Y_k=i)\,|\,Y_k^0]$$

και τα 'δειγματικά ποσοστά' $\hat{n}_i/N$ είναι ουσιαστικά οι εκτιμητές μέγιστης πιθανοφάνειας $\hat{p}_i$, $i=1,2,...,r$.

Χρησιμοποιώντας τον τύπο για τα $\hat{n}_i$ και υποθέτοντας ότι $n/N\to f$, $f>0$ καθώς $N\to\infty$, η ασυμπτωτική κατανομή των δειγματικών ποσοστών $(\hat{n}_1,\hat{n}_2,...,\hat{n}_{r-1})$ μπορεί να δοθεί από τον παρακάτω τύπο

$$\frac{1}{\sqrt{N}}(\hat{n}_1-Np_1,\hat{n}_2-Np_2,...,\hat{n}_{r-1}-Np_{r-1})\xrightarrow{d}MVN(0,\sum\nolimits_A),\ \text{καθώς}\ N\to\infty$$

όπου *MVN* είναι η Πολυδιάστατη Κανονική Κατανομή (*Multivariate Normal Distribution*). Ο ασυμπτωτικός πίνακας συνδιακύμανσης $\sum_A=[\sigma_{ji}]_{(r-1)\times(r-1)}$, όπου

$$\sigma_{ji}=\begin{cases}\dfrac{p_iq_i}{f}[1-(1-f)K_i] & ,i=j\\[2ex] (1-\dfrac{1}{f})\displaystyle\sum_{k=1}^{r}\lambda_{kj}\lambda_{ki}p_k'-p_jp_i & ,i\neq j\end{cases}$$

με $q_i=1-p_i$ και συντελεστή αξιοπιστίας για την *i* κατηγορία (Tenenbein, 1972)

$$K_i=\Big[Corr\{I(Y=i),E[I(Y=i)\,|\,Y^0]\}\Big]^2=\frac{p_i}{q_i}\{\sum_{k=1}^{r}(\alpha_{ki}^2/p_k')-1\}\,.$$

Εφόσον ο πίνακας $\sum_A$ είναι πλήρους τάξης, ο αντίστροφος του $\sum_A^{-1}=[\tau_{ji}]_{(r-1)\times(r-1)}$ υπάρχει και τα στοιχεία του τ_{ji} μπορούν να εκτιμηθούν από τον εκτιμητή μέγιστης πιθανοφάνειας $\hat{\tau}_{ji}$. Κατά συνέπεια, καθώς $N\to\infty$ και κάτω από τη μηδενική υπόθεση H_0

$$\hat{X}^2 = \frac{1}{N}\sum_{j=1}^{r}\sum_{i=1}^{r}(\hat{n}_j - Np_{0j})\hat{\tau}_{ji}(\hat{n}_i - Np_{0i}) \xrightarrow{d} \chi^2_{(r-1)}.$$

Βασιζόμενοι στην ανωτέρω σχέση, απορρίπτουμε τη μηδενική υπόθεση H_0 αν για την τιμή του προσαρμοσμένου X^2 ελέγχου ισχύει

$$\hat{X}^2 > c_a(r-1).$$

3.4 Εφαρμογή διόρθωσης ελέγχων με χρήση διπλής δειγματοληψίας - δυωνυμική κατανομή

Για την περαιτέρω μελέτη της ασυμπτωτικής συμπεριφοράς του προσαρμοσμένου $\hat{X}^2$ ελέγχου, θα θεωρήσουμε τους ελέγχους καλής προσαρμογής για τη δυωνυμική κατανομή σε περίπτωση ύπαρξης σφαλμάτων λανθασμένης ταξινόμησης. Κατά συνέπεια, η μηδενική υπόθεση του ελέγχου είναι $H_0 : p = p_0$. Επίσης, θα επικεντρώσουμε τη μελέτη στη χρήση εσωτερικών δειγμάτων επικύρωσης. Εδώ με p συμβολίζεται η πιθανότητα αποτελέσματος ΄1’ για τη δυωνυμική παρατήρηση και με $q = 1 - p$ η πιθανότητα αποτελέσματος ‘0’. Οι πιθανότητες λανθασμένης ταξινόμησης εκφράζονται ως εξής

$$\theta = \Pr(Y^0 = 0 \mid Y = 1)$$

και

$$\phi = \Pr(Y^0 = 1 \mid Y = 0).$$

Έτσι προκύπτει ότι

$$p' = \Pr(Y^0 = 1) = p(1-\theta) + q\phi.$$

Στη συνέχεια, ορίζουμε ως n_{ji} με $i, j = 0,1$ τον αριθμό των μονάδων στο δείγμα επικύρωσης με αληθή κατηγορία i και εσφαλμένη κατηγορία j και

$$m_k = \sum_{j=n+1}^{N} I(Y^0 = k), \quad k = 0,1.$$

Κατά συνέπεια, οι εκτιμητές μέγιστης πιθανοφάνειας είναι

$$\hat{p} = \frac{n_{11}}{n_{1\cdot}} \cdot \frac{m_1 + n_{1\cdot}}{N} + \frac{n_{01}}{n_{0\cdot}} \cdot \frac{m_0 + n_{0\cdot}}{N}$$

$$\hat{\theta} = \left(\frac{n_{01}}{n_{0\bullet}} \cdot \frac{m_0 + n_{0\bullet}}{N}\right) \Big/ \hat{p}$$

και

$$\hat{\phi} = \left(\frac{n_{10}}{n_{1\bullet}} \cdot \frac{m_1 + n_{1\bullet}}{N}\right) \Big/ (1 - \hat{p}) .$$

Ομοίως, ας υποθέσουμε ότι θεωρούμε την ακόλουθη έκφραση για την δεσμευμένη αναμενόμενη τιμή

$$E(Y \mid Y^0) = \lambda_1 Y^0 + \lambda_2 (1 - Y^0) ,$$

με

$$\lambda_1 = \frac{p(1-\theta)}{p'} = \frac{p(1-\theta)}{p(1-\theta) + q\phi} \quad \text{και} \quad \lambda_2 = \frac{p\theta}{1-p'} = \frac{p\theta}{p\theta + q(1-\phi)} .$$

Τότε ο εκτιμητής της είναι η

$$\hat{E}(Y \mid Y^0) = \hat{\lambda}_1 Y^0 + \hat{\lambda}_2 (1 - Y^0) ,$$

με

$$\hat{\lambda}_1 = \frac{\hat{p}(1-\hat{\theta})}{\hat{p}(1-\hat{\theta}) + \hat{q}\hat{\phi}} \quad \text{και} \quad \hat{\lambda}_2 = \frac{\hat{p}\hat{\theta}}{\hat{p}\hat{\theta} + \hat{q}(1-\hat{\phi})} .$$

Εάν θεωρήσουμε την μεταβλητή Y' με τιμές

$$Y_i' = \begin{cases} Y_i & , i = 1, 2, \ldots, n \\ \hat{\lambda}_1 Y^0 + \hat{\lambda}_2 (1 - Y^0) & , i = n+1, \ldots, N \end{cases}$$

με

$$\bar{Y}_N' = \frac{1}{N} \sum_{i=1}^{N} Y_i' ,$$

μπορεί εύκολα να αποδειχτεί ότι $\bar{Y}_N' = \hat{p}$. Κατά αυτόν τον τρόπο, οι τιμές της μεταβλητής Y_i^* μπορούμε να πούμε ότι παριστούν τις προσαρμοσμένες ‘παρατηρήσεις’ για τα σφάλματα λανθασμένης ταξινόμησης, καθώς επίσης και ότι ο εκτιμητής μέγιστης πιθανοφάνειας για το p είναι το προσαρμοσμένο δειγματικό ποσοστό. Ας σημειωθεί σε αυτό το σημείο πως σε περίπτωση μη ύπαρξης σφαλμάτων λανθασμένης ταξινόμησης ισχύει $\theta = \phi = 0$, το οποίο και συνεπάγεται $\lambda_1 = 1,\ \lambda_2 = 0$ και $Y = Y^0$. Τότε ισχύει

$$\bar{Y}_N' = \frac{1}{N} \sum_{i=1}^{N} Y_i = \hat{p}$$

δηλαδή το προσαρμοσμένο δειγματικό ποσοστό είναι το σύνηθες δειγματικό ποσοστό, εφόσον $\hat{\lambda}_1 = 1 - \hat{\lambda}_2 = 1$.

Για $n/N \to f$ καθώς $N \to \infty$ και με τη χρήση των αναπτυγμάτων Taylor και του θεωρήματος σύγκλισης του Slutsky, μπορεί να δειχτεί ότι ισχύει

$$\sqrt{N}(\bar{Y}'_N - p) \xrightarrow{d} (0, \sigma^2)$$

με

$$\sigma^2 = \frac{pq}{f}(1-(1-f)K)$$

και συντελεστή αξιοπιστίας για τη δυωνυμική κατανομή (Tenenbein, 1970)

$$K = [Corr(Y, Y^0)]^2 = \frac{pq(1-\theta-\phi)^2}{p'(1-p')} = \frac{pq(1-\theta-\phi)^2}{[p(1-\theta)+q\phi][p\theta+q(1-\phi)]}.$$

Ως συνέπεια των ανωτέρω, η προσαρμοσμένη τιμή του X^2 ελέγχου υπό τη μηδενική υπόθεση H_0 δίνεται από τη σχέση

$$\hat{X}^2 = \frac{n(\bar{Y}'_N - p_0)^2}{p_0 q_0 [1-(1-\frac{n}{N})\hat{K}_0]},$$

όπου

$$\hat{K}_0 = \frac{p_0 q_0 (1-\hat{\theta}-\hat{\phi})^2}{[p_0(1-\hat{\theta})+q_0\hat{\phi}][p_0\hat{\theta}+q_0(1-\hat{\phi})]}.$$

Η μηδενική υπόθεση H_0 απορρίπτεται εάν ισχύει

$$\hat{X}^2 > c_a(1)$$

ή

$$|\bar{Y}'_N - p_0| > z_{a/2}\sqrt{\frac{p_0 q_0}{n}[1-(1-\frac{n}{N})\hat{K}_0]}$$

όπου $z_{a/2}$ είναι το σύνηθες $100(1-a/2)$ εκατοστημόριο της τυποποιημένης κανονικής κατανομής $N(0,1)$.

Μπορούμε επίσης να μελετήσουμε την ασυμπτωτική συμπεριφορά του ελέγχου θεωρώντας την ακολουθία των εναλλακτικών υποθέσεων

$$H_{1N} = p_0 + \frac{d}{\sqrt{N}} + o(1/\sqrt{N})$$

για κάθε σταθερά d. Μέσω της εφαρμογής του θεωρήματος του Noether (βλ. Randles and Wolfe (1979)), μπορεί να αποδειχτεί ότι η αποτελεσματικότητα (*efficacy*) του ελέγχου βάσει του $\bar{Y}'_N$ είναι

$$eff(\bar{Y}^*_N) = \lim_{N\to\infty} \frac{\frac{d}{dp}[E(\bar{Y}'_N)]\Big|_{p=p_0}}{\sqrt{NVar(\bar{Y}'_N)\Big|_{p=p_0}}} = \frac{1}{\sqrt{p_0 q_0[1-(1-f)K_0]/f}}.$$

Η αποτελεσματικότητα του ελέγχου που βασίζεται στο δειγματικό ποσοστό $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ για το οποίο λαμβάνουμε υπ' όψιν μόνο τα δεδομένα που έχουν ταξινομηθεί ορθά, η αποτελεσματικότητα είναι

$$eff(\bar{Y}_n) = \frac{1}{\sqrt{p_0 q_0 / f}}.$$

Για τη σύγκριση των δύο ελέγχων, βασιζόμαστε στην ασυμπτωτική σχετική αποτελεσματικότητα (*asymptotic relative efficiency* ή ARE), η οποία δίνεται από τη σχέση

$$ARE(\bar{Y}'_N, \bar{Y}_n) = \frac{eff(\bar{Y}'_N)}{eff(\bar{Y}_n)} = \frac{1}{1-(1-f)K_0}.$$

Λαμβάνοντας επίσης υπ' όψιν ότι $K_0 = [Corr(Y, Y^0)]^2\Big|_{p=p_0} \geq 0$, οδηγούμαστε στα ακόλουθα συμπεράσματα

- εφόσον $n/N \to f < 1$, λόγω δηλαδή του ότι το δείγμα επικύρωσης είναι μικρότερο του συνολικού εξεταζόμενου δείγματος του πληθυσμού, ισχύει πάντα ότι $ARE(\bar{Y}'_N, \bar{Y}_n) \geq 1$.
- για $Corr(Y, Y^0) \neq 0$, όταν δηλαδή η ορθή και η εσφαλμένη ταξινόμηση του δείγματος είναι συσχετισμένες σε έναν βαθμό, ισχύει $ARE(\bar{Y}'_N, \bar{Y}_n) > 1$. Κατά την περίπτωση της απόλυτης θετικής συσχέτισης όπου $Corr(Y, Y^0) = 1$, τότε έχουμε $ARE(\bar{Y}'_N, \bar{Y}_n) = 1/f$ και $ARE(\bar{Y}'_N, \bar{Y}_N) = 1$. Αυτό σημαίνει ότι η απλή δυωνυμική και η διπλή δειγματοληψία δίνουν τις ίδιες πληροφορίες όσον αφορά τον έλεγχο της μηδενικής υπόθεσης H_0 έναντι της ακολουθίας των εναλλακτικών υποθέσεων H_{1N} που έχουν οριστεί ως άνω.
- για $Corr(Y, Y^0) = 0$, όταν δηλαδή ορθή και εσφαλμένη ταξινόμηση είναι ασυσχέτιστες, ισχύει $ARE(\bar{Y}'_N, \bar{Y}_n) > 1$, το οποίο και δηλώνει ότι οι εσφαλμένα ταξινομημένες

παρατηρήσεις δεν παρέχουν καμία πληροφορία σχετικά με το πραγματικό πληθυσμιακό ποσοστό p.

Βάσει των ανωτέρω, οδηγούμαστε στο γενικό συμπέρασμα ότι η ασυμπτωτική σχετική αποτελεσματικότητα $ARE(\bar{Y}_N', \bar{Y}_n)$ είναι αυστηρώς αύξουσα ως συνάρτηση του συντελεστή αποτελεσματικότητας K_0.

Είναι φανερό πως οι τιμές ορθής ταξινόμησης $\{Y_1,...,Y_n\}$ της μεταβλητής Y είναι απαραίτητες για την απόκτηση πληροφοριών κατά την περίπτωση όπου οι πιθανότητες λανθασμένης ταξινόμησης θ και ϕ είναι άγνωστες. Αντικείμενο ενδιαφέροντος αποτελεί να εκτιμηθεί η επίδραση της ύπαρξης των πραγματικών ταξινομήσεων των δεδομένων (Cheng et al. (1998)). Ας υποθέσουμε ακολούθως ότι έχουμε τις εσφαλμένα ταξινομημένες παρατηρήσεις $\{Y_1^0,...,Y_N^0\}$ καθώς και ότι οι πιθανότητες λανθασμένης ταξινόμησης θ και ϕ είναι γνωστές. Κατά αυτόν τον τρόπο, οι προσαρμοσμένες παρατηρήσεις για την εσφαλμένη ταξινόμηση είναι

$$Y_i^* = \tilde{\lambda}_1 Y_i^0 + \tilde{\lambda}_2 (1 - Y_i^0), \quad i = 1,..., N$$

με

$$\tilde{\lambda}_1 = \frac{\tilde{p}(1-\theta)}{\tilde{p}(1-\theta) + \tilde{q}\phi} \quad , \quad \tilde{\lambda}_2 \frac{\tilde{p}\theta}{\tilde{p}\theta + \tilde{q}(1-\phi)}.$$

και εκτιμητή μέγιστης πιθανοφάνειας του p

$$\tilde{p} = \frac{\frac{1}{N}\sum_{i=1}^{N} Y_i^0 - \phi}{1 - \theta - \phi}.$$

Μπορούμε να δούμε ότι $\bar{Y}_N^* = \frac{1}{N}\sum_{i=1}^{N} Y_i^* = \tilde{p}$, άρα ο εκτιμητής $\bar{Y}_N^*$ μπορεί να θεωρηθεί ως το προσαρμοσμένο πληθυσμιακό ποσοστό. Επιπρόσθετα, η αποτελεσματικότητα του $\bar{Y}_N^*$ λαμβάνοντας υπ' όψιν την ακολουθία των εναλλακτικών υποθέσεων H_{1N}, όπως αυτήν έχει οριστεί προηγουμένως, είναι

$$eff(\bar{Y}_N^*) = \sqrt{\frac{(1-\phi-\theta)^2}{\pi_0(1-\pi_0)}} = \sqrt{\frac{K_0}{p_0 q_0}},$$

με $p_0' = p_0(1-\theta-\phi)+\phi$. Τοιουτοτρόπως, η ασυμπτωτική σχετική αποτελεσματικότητα του $\bar{Y}_N'$ ως προς τον $\bar{Y}_N^*$ είναι

$$ARE(\bar{Y}_N', \bar{Y}_N^*) = \frac{f}{K_0[1-(1-f)K_0]}.$$

Παρατηρούμε ότι για $n/N \to f > 0$ ισχύει $ARE(\bar{Y}_N', \bar{Y}_N^*) = 1$ για $K_0 = 1$ και $ARE(\bar{Y}_N', \bar{Y}_N^*) = \infty$ για $K_0 = 0$. Επίσης, για κάθε $0 < K_0 < 1$ μπορεί να αποδειχθεί πως ως συνάρτηση του $f > 0$ η ασυμπτωτική αποτελεσματικότητα $ARE(\bar{Y}_N', \bar{Y}_N^*)$ είναι αυστηρώς αύξουσα συνάρτηση και ισχύει $ARE(\bar{Y}_N', \bar{Y}_N^*) = 1$ για $f = K_0/(1+K_0)$. Με άλλα λόγια, εάν έχουμε $f = K_0/(1+K_0)$, τότε η γνώση του συνόλου των πραγματικών μεταβλητών $\{Y_1, ..., Y_n\}$ Είναι ισοδύναμη με το να ξέρουμε ακριβώς τις πιθανότητες λανθασμένης ταξινόμησης φ και θ. Αντίθετα, εάν $f < K_0/(1+K_0)$, ο εκτιμητής $\bar{Y}_N'$ είναι πιο αποτελεσματικός από τον $\bar{Y}_N^*$. Τέλος, για $0 < K_0 < 1$, δηλαδή όταν $f < 1/2$, ο εκτιμητής $\bar{Y}_N'$ είναι πάντα πιο αποτελεσματικός από τον $\bar{Y}_N^*$ για κάθε τιμή του K_0.

3.5 Το κριτήριο λόγου πιθανοφανειών (likelihood ratio test)

Στην ενότητα αυτή θα εξετάσουμε τη μηδενική υπόθεση H_0 με τη χρήση του συνήθους ελέγχου του λόγου πιθανοφανειών. Ας σημειωθεί ότι υπό την H_0 η λογαριθμική πιθανοφάνεια (log-likelihood) είναι

$$L(\underset{\sim}{p}_0, \mathbf{A}) = \sum_{j=1}^{r}\sum_{i=1}^{r}[n_{ji}(\log p_{0i} + \log \alpha_{ji}) + (n_{j\cdot} - n_{ji})\log(p_{0j}' - \alpha_{ji}p_{0i})] + \sum_{k=1}^{r} m_k \log p_{0k}'.$$

Επιλύοντας την εξίσωση της λογαριθμικής πιθανοφάνειας υπό τη μηδενική υπόθεση H_0 οδηγούμαστε στο συμπέρασμα ότι οι εκτιμητές μέγιστης πιθανοφάνειας $\tilde{\alpha}_{ji}$ δεν είναι μονοσήμαντα ορισμένοι. Για παράδειγμα, κατά την περίπτωση όπου $r = 2$, οι αντίστοιχες εξισώσεις πιθανοφάνειας είναι

$$\frac{n_{11}}{(1-\phi)q_0} - \frac{n_{21}}{\phi q_0} + \frac{m_1}{(1-\phi)q_0 + \theta p_0} - \frac{m_2}{\phi q_0 + (1-\theta)p_0} = 0$$

$$\frac{n_{12}}{\theta p_0} - \frac{n_{22}}{(1-\theta)p_0} + \frac{m_1}{(1-\phi)q_0 + \theta p_0} - \frac{m_2}{\phi q_0 + (1-\theta)p_0} = 0.$$

Η εύρεση προσεγγιστικών αριθμητικών λύσεων για τις εξισώσεις πιθανοφάνειας απαιτεί τη χρήση αριθμητικών μεθόδων, όπως για παράδειγμα τη μέθοδο Newton – Raphson. Είναι σύνηθες η πολυπλοκότητα των εξισώσεων να είναι μεγαλύτερη όταν το πλήθος των κατηγοριών των μεταβλητών αυξάνεται. Συνέπεια αυτού είναι να αυξάνεται ο βαθμός δυσκολίας της εύρεσης των δυνατών λύσεων. Ακολουθώντας τις συνήθεις μεθόδους επίλυσης, μπορεί να αποδειχθεί ότι ο λόγος πιθανοφανειών (*LR – Likelihood Ratio*) με τύπο

$$LR = 2\{\log L(\hat{\underset{\sim}{p}} - \hat{\mathbf{A}}) - \log L(\hat{\underset{\sim}{p}}_0 - \hat{\mathbf{A}})\}$$

$$= 2\sum_{j=1}^{r}\sum_{i=1}^{r}\{[n_{ji}(\log \hat{p}_i - \log p_{0i}) + (\log \hat{\alpha}_{ji} - \log \tilde{\alpha}_{ji})] +$$

$$+(n_{j.} - n_{ji})[\log(\hat{p}'_j - \hat{\alpha}_{ji}\hat{p}_i) - \log(\tilde{p}'_{0j} - \tilde{\alpha}_{ji}p_{0i})]\} +$$

$$+2\sum_{k=1}^{r} m_k(\log \hat{p}'_k - \log \tilde{p}'_{0k})$$

όπου $\hat{p}'_j = \sum_{i=1}^{r} \hat{\alpha}_{ji}\hat{p}_i$ και $\tilde{p}'_{0j} = \sum_{i=1}^{r} \tilde{p}_{0i}\tilde{\alpha}_{ji}$,

συγκλίνει κατά κατανομή στην κατανομή $\chi^2_{(r-1)}$ για $N \to \infty$ και $n/N \to f > 0$.

3.6 Σχετική βιβλιογραφία

Πολλοί ερευνητές ασχολήθηκαν με την επίδραση της λανθασμένης ταξινόμησης στις εκτιμήσεις των δεδομένων πληθυσμών στους οποίους και έχουμε ύπαρξη της. Οι Bross (1954), Mote and Anderson (1965), Assakul and Proctor (1967), Koch (1969), Cheng et al. (1998) είναι ενδεικτικά μερικοί ερευνητές που πραγματεύτηκαν το πρόβλημα αυτό και προέβησαν σε παρόμοια συμπεράσματα για το μέγεθος και την ισχύ του ελέγχου Ο Koch (1969) ειδικότερα πραγματεύτηκε την επίδραση της λανθασμένης ταξινόμησης στους ελέγχους κατά την ύπαρξη επαναλαμβανόμενων μετρήσεων στο δειγματοληπτικό σχήμα του υπό εξέταση πληθυσμού. Ενδιαφέρον παρουσιάζει και η προσέγγιση που δίνεται από τους Pardo and Zografos (2000), οι οποίοι αφού πρώτα δείχνουν ότι ο έλεγχος χ^2 και το κριτήριο λόγου πιθανοφανειών ανήκουν στην οικογένεια εκθετικών αποκλίσεων (*power divergence family*) και είναι υποπεριπτώσεις αυτής για διαφορετικές τιμές της χρησιμοποιούμενης

παραμέτρου λ (για $\lambda = 1$ και $\lambda \to 0$ αντίστοιχα), προχωρούν σε ελέγχους καλής προσαρμογής των δεδομένων και διόρθωση της υπάρχουσας λανθασμένης ταξινόμησης με κριτήρια ελέγχου που βασίζονται στις φ-αποκλίσεις (*φ-divergencies*). Βάσει αυτής της μεθόδου προσέγγισης, καταλήγουν και αυτοί με τη σειρά τους στο γενικότερο συμπέρασμα επίδρασης της λανθασμένης ταξινόμησης : η παρουσία λανθασμένης ταξινόμησης οδηγεί σε αύξηση του μεγέθους του ελέγχου και σε μείωση της ισχύος του εν συγκρίσει με την περίπτωση που τα δεδομένα του πληθυσμού είναι ορθά ταξινομημένα στις κατηγορίες όπου και ανήκουν.

ΚΕΦΑΛΑΙΟ 4

Συμπερασματολογία για το μηχανισμό της λανθασμένης ταξινόμησης

4.1 Εισαγωγή

Στην προηγούμενη ενότητα είδαμε ότι η λανθασμένη ταξινόμηση οδηγεί γενικά σε μεροληπτικές εκτιμήσεις. Για την προσαρμογή των δεδομένων και τη διόρθωση αυτής της μεροληψίας θα βασιστούμε στην χρήση υποθέσεων ή πληροφοριών για το μηχανισμό λανθασμένης ταξινόμησης.

Σε σπάνιο αριθμό περιπτώσεων ο μηχανισμός λανθασμένης ταξινόμησης θα είναι γνωστός. Συγκεκριμένα, για τη μεθοδολογία τυχαιοποιημένης απόκρισης (*randomized response* ή *RR technique*) (Warner (1965), Chen (1979a)) τα άτομα του πληθυσμού που ερευνάται εξετάζονται βάσει ενός τυχαίου μηχανισμού που εξασφαλίζει τον εμπιστευτικό χαρακτήρα των απαντήσεων τους σε ερωτήσεις σε ευαίσθητα ζητήματα. Ο ερευνητής δεν γνωρίζει τις πραγματικές απαντήσεις αλλά έχει γνώση του πίνακα λανθασμένης ταξινόμησης για τα δεδομένα και βάσει αυτού μπορεί να προβεί σε αμερόληπτες εκτιμήσεις των παραμέτρων του πληθυσμού όπως αυτές θα περιγραφούν στο Κεφάλαιο 6 (βλ. επίσης Press (1968)).

Ακόμα πιο συνηθισμένη είναι η περίπτωση κατά την οποία ο μηχανισμός λανθασμένης ταξινόμησης δεν είναι γνωστός. Στο επόμενο Κεφάλαιο θα εξεταστούν δύο μέθοδοι που μπορούν να χρησιμοποιηθούν για την εξαγωγή συμπερασμάτων για αυτό το μηχανισμό λανθασμένης ταξινόμησης. Ακόμα και στην περίπτωση που δεδομένα αυτής της μορφής είναι μη διαθέσιμα για τον μελετητή, οι Tweedie et al. (1994, σελ. 22) επιχειρηματολογούν υπέρ του γεγονότος πως «δεν πρέπει να παραιτηθούμε, να δηλώσουμε ότι τα δεδομένα είναι πιθανότατα λανθασμένα και τα συμπεράσματα μεροληπτικά και να αμφισβητήσουμε την εγκυρότητα των ισχυρισμών μας». Αντίθετα, συστήνουν μία ανάλυση ευαισθησίας για τα δεδομένα της μορφής ‘τι θα συμβεί εάν’, κατά την οποία πιθανές μεροληψίες μπορούν να

προσεγγιστούν και να αντιμετωπιστούν μέσω της θεώρησης ενός ρεαλιστικού φάσματος υποθέσεων λανθασμένης ταξινόμησης.

4.2 Μελέτες επικύρωσης με προτιμώμενη διαδικασία

Στην πλειονότητα των περιπτώσεων, η συνήθης διαδικασία καταγραφής και μέτρησης των δεδομένων του εξεταζόμενου πληθυσμού χαρακτηρίζεται από εσφαλμένες καταγραφές των εξεταζόμενων χαρακτηριστικών. Με άλλα λόγια, υπόκειται σε λανθασμένη ταξινόμηση. Αυτό έχει ως αποτέλεσμα την εξαγωγή μεροληπτικών αποτελεσμάτων και λανθασμένων συμπερασμάτων για τη διεξαγόμενη έρευνα. Αμερόληπτές εκτιμήσεις οι οποίες καθιστούν εφικτή την εκτίμηση των παραμέτρων του μηχανισμού λανθασμένης ταξινόμησης μπορούν να εξασφαλιστούν με τη χρήση περισσότερο εκλεπτυσμένων μεθόδων καταγραφής-μέτρησης μεγαλύτερης ακρίβειας. Οι διαδικασίες αυτές ονομάζονται προτιμώμενες διαδικασίες (*preferred procedures*) (Forsman and Schreiner (1991), Kuha and Skinner (1997)) . Ο Deming (1950) περιγράφει την προτιμώμενη διαδικασία ως «μία διαδικασία που χαρακτηρίζεται από το γεγονός ότι υποτίθεται πως δίνει ή θα μπορούσε να δώσει τα αναγκαία αποτελέσματα που χρειαζόμαστε για μία συγκεκριμένη έκβαση». Συχνά η προτιμώμενη διαδικασία μπορεί να είναι είτε μεγάλου κόστους είτε με κάποιον άλλο τρόπο ακατάλληλη προς εφαρμογή για όλες τις μονάδες της διεξαγόμενης έρευνας, καθιστώντας έτσι αναγκαία τη χρήση μίας επιρρεπούς σε λάθη διαδικασίας. Ωστόσο, μπορεί να είναι εφικτό να εφαρμόσουμε την προτιμώμενη διαδικασία για ένα δείγμα επικύρωσης n^{v} περιορισμένου μεγέθους σε σχέση με το συνολικά εξεταζόμενο δείγμα του πληθυσμού. Λαμβάνοντας υπ' όψιν ότι η πραγματική μεταβλητή Y αντιστοιχεί σε μετρήσεις που έχουν γίνει μέσω της προτιμώμενης διαδικασίας και η υποκείμενη σε σφάλμα μεταβλητή Y' σε μετρήσεις μέσω της συνήθους διαδικασίας που χρησιμοποιείται σε μία έρευνα, οδηγούμαστε στο ακόλουθο συμπέρασμα : εάν οι Y' και Y λαμβάνονται βάσει του δείγματος επικύρωσης n^{v}, τότε οι πιθανότητες λανθασμένης ταξινόμησης α_{ji} μπορούν να εκτιμηθούν άμεσα, δεδομένου ότι το δείγμα επικύρωσης έχει επιλεχθεί από τον εξεταζόμενο πληθυσμό από ένα σχεδιασμό με γνωστή πιθανότητα δειγματοληψίας.

4.2.1 Παραδείγματα Προτιμώμενων Διαδικασιών

Η προτιμώμενη διαδικασία μπορεί να περιλαμβάνει τεχνικές όπως η έρευνα με εκ νέου συνεντεύξεις (*re-interview surveys*) (Bailar (1968)), κρίσεις ειδικών εμπειρογνωμόνων, π.χ. αξιολόγηση ζημιών σε σπίτια από ειδικούς επιθεωρητές (βλ. Swires-Hennessy and Thomas (1987)) ή την εξέταση ατόμων που εμφανίζουν συμπτώματα ασθενειών από γιατρό (βλ. Cobb and Rosenbaum (1956)). Οι συγκρίσεις με διαχειριστικές καταγραφές μπορούν επίσης να παρέχουν προτιμώμενες διαδικασίες, π.χ. συγκρίσεις εργασιακού ιστορικού με αρχεία επιτροπών συνταξιοδότησης (βλ. Baumgarten et al. (1983)) ή σύγκριση δεδομένων μιας έρευνας για τη χρήση αντιβιοτικών με ιατρικά αρχεία (Greenland (1988a)). Στην Επιδημιολογία χρησιμοποιείται περισσότερο ο όρος 'χρυσός κανόνας' (gold standard). Για παραδείγματα βλ. Bauman and Koch (1983) (βιοχημικές μετρήσεις των επιδράσεων του καπνίσματος) και Willett et al. (1985) (μετρήσεις κατανάλωσης φαγητού ως δείκτες διατροφής). Άλλα παραδείγματα δίνονται από τους Copeland et al. (1977).

4.3 Επιλογή δείγματος επικύρωσης

Έστω η πραγματική μεταβλητή Y και λανθασμένα ταξινομημένη μεταβλητή Y'. Αυτές οι μεταβλητές, πιθανότατα μαζί με άλλες μεταβλητές, καταμετρούνται για κάθε μονάδα του πληθυσμού σε ένα δείγμα επικύρωσης. Το γεγονός αυτό εγείρει δύο σημαντικά προβλήματα

- Πως μπορεί να γίνει καταμέτρηση της πραγματικής μεταβλητής Y ;
- Με ποιον τρόπο πρέπει να επιλεγούν οι μονάδες του δείγματος επικύρωσης ;

Μια μέτρηση της μεταβλητής Y μπορεί να είναι δυνατή χρησιμοποιώντας μία μεθοδολογία που αναφέρεται ως χρυσός κανόνας (*gold standard*). Μολαταύτα, η μέθοδος αυτή μπορεί να είναι πολύ ακριβή στο κόστος εφαρμογής του για όλες τις μονάδες του συνολικού δείγματος της έρευνας ή να είναι διαθέσιμο μόνο για ένα υποσύνολο των μονάδων του εξεταζόμενου πληθυσμού. Ο χρυσός κανόνας είναι μία έννοια-κλειδί για τις μελέτες επικύρωσης και η υπόθεση που ακολουθείται, ότι καταμετρά την Y με ακρίβεια ή έστω με αμελητέο σφάλμα, είναι καίρια (βλ. Wachholder et al. (1993) για μία μελέτη πάνω στη μεροληψία που εισάγεται κατά τη χρήση ενός εσφαλμένου ή υποβαθμισμένης ποιότητας (*alloyed*) χρυσού κανόνα).

Υπό ιδανικές συνθήκες, το δείγμα επικύρωσης n^{v} πρέπει να είναι ένα υπο-δείγμα του κυρίου δείγματος των πρωταρχικών δεδομένων που λαμβάνεται μέσω ενός γνωστού

τυχαιοποιημένου σχεδίου διπλής δειγματοληψίας (*randomized double sampling scheme*), γνωστού επίσης και ως εσωτερική μελέτη επικύρωσης (*internal validation study*). Για τα εσωτερικά δείγματα επικύρωσης (*internal validation samples*) τα δειγματικά ποσοστά για τις κατηγορίες της μεταβλητής Y στο δείγμα επικύρωσης είναι αμερόληπτοι εκτιμητές των αντίστοιχων πληθυσμιακών ποσοστών. Με αυτόν τον τρόπο, οι πιθανότητες λανθασμένης ταξινόμησης $\Pr(Y|Y')$ και οι προγνωστικές αξίες (*predictive values*) $\Pr(Y|Y')$ στον εξεταζόμενο πληθυσμό μπορούν να εκτιμηθούν επαρκώς μέσω δεδομένων προερχόμενων από εσωτερικά δείγματα επικύρωσης. Χρήσιμη επίσης για εσωτερικές μελέτες επικύρωσης είναι και η επιλογή ενός προκαθορισμένου ποσοστού από τα δεδομένα του δείγματος επικύρωσης για κάθε κατηγορία της μεταβλητής Y', το οποίο και αυξάνει την αποτελεσματικότητα της εκτίμησης των προγνωστικών αξιών $\Pr(Y|Y')$ (Haitovsky et al., 1992).

Υπάρχουν ωστόσο κάποιες περιπτώσεις όπου πρακτικοί λόγοι δεν μας επιτρέπουν τη χρήση διπλής δειγματοληψίας. Εάν χρησιμοποιούνται δεδομένα επικύρωσης προερχόμενα από πρότερες μελέτες, ή εάν τα δεδομένα επικύρωσης συλλέγονται αργότερα από τη συλλογή των πρωτογενών δεδομένων, τότε μπορούμε, λογικά να θεωρήσουμε ότι η κατανομή των δεδομένων για τις κατηγορίες της εξεταζόμενης μεταβλητής δεν είναι ίδια για το δείγμα επικύρωσης και για τον πρωταρχικά εξεταζόμενο πληθυσμό. Για παράδειγμα, για μία δειγματική διαχρονική έρευνα (*panel survey*) μπορεί να είναι επιθυμητό να αποφύγουμε το επιπλέον φορτίο της διεξαγωγής μελέτης επικύρωσης για τα άτομα της έρευνας έτσι ώστε να μην υπάρξει ο κίνδυνος αλλοίωσης μελλοντικών ποσοστών απόκρισης (βλ. Hill 1992). Τότε καταφεύγουμε στην επιλογή δείγματος επικύρωσης, ξένου προς το κύριο δείγμα. Τέτοιου είδους μελέτες ονομάζονται εξωτερικές μελέτες επικύρωσης (*external validation studies*) και στις περισσότερες των περιπτώσεων δεν είναι λογικό να υποθέσουμε ότι τα δειγματικά ποσοστά των κατηγοριών της Y είναι αμερόληπτες εκτιμήσεις των πραγματικών ποσοστών $P_{Y(i)}$ του ερευνώμενου πληθυσμού. Για τα εξωτερικά δείγματα επικύρωσης (*external validation samples*) ισχύει ότι μόνο οι πιθανότητες λανθασμένης ταξινόμησης μπορούν να θεωρηθούν μεταβιβάσιμες ανάμεσα στα σύνολα των δεδομένων. Μπορεί να είναι πιο ρεαλιστικό να υποθέσουμε για τις μελέτες αυτές ότι ο πίνακας λανθασμένης ταξινόμησης **A** μπορεί να εκτιμηθεί αμερόληπτα από το δείγμα επικύρωσης. Αυτή η υπόθεση θα μπορούσε να γίνει εάν ο μηχανισμός λανθασμένης ταξινόμησης ήταν ομοιογενής για το δείγμα

επικύρωσης και για τον εξεταζόμενο πληθυσμό στο σύνολο του. Την ύπαρξη ή μη ομοιογένειας μπορούμε να ερευνήσουμε, ως ένα βαθμό, αξιολογώντας τη σχέση ανάμεσα στις μεταβλητές Y' και Y ως προς συμμεταβλητές (*covariates*) όπως π.χ. η ηλικία και το φύλο. Εάν για παράδειγμα βρεθεί ότι τα λανθασμένα ταξινομημένα ποσοστά εξαρτώνται από το φύλο, τότε καθίσταται απαραίτητο να ορίσουμε διαφορετικούς πίνακες λανθασμένης ταξινόμησης για τους άντρες και τις γυναίκες. Στην περίπτωση αυτή, η επιλογή δείγματος επικύρωσης ενός μόνου φύλου δεν είναι επαρκής για να προβούμε σε σωστά συμπεράσματα. Στο σύγγραμμα αυτό θα στρέψουμε την προσοχή μας περισσότερο στις περιπτώσεις εσωτερικών μελετών επικύρωσης, ερευνώντας όμως παράλληλα αλλά σε μικρότερο βαθμό και την περίπτωση χρήσης εξωτερικών δειγμάτων επικύρωσης.

Για ένα δείγμα προερχόμενο από διπλή δειγματοληψία θα είναι γενικά δυνατόν να λάβουμε άμεσες εκτιμήσεις όχι μόνο για τις πιθανότητες λανθασμένης ταξινόμησης α_{ji}, αλλά και για τα πραγματικά πληθυσμιακά ποσοστά $\mathbf{P_Y}$. Επιπλέον, είναι δυνατόν να εκτιμήσουμε δεσμευμένες πιθανότητες της ακόλουθης μορφής

$$c_{ij} = \Pr(Y_\xi = i \,|\, Y'_\xi = j) \quad \text{για κάθε} \;\; i, j = 1, ..., r \qquad (4.1)$$

που δηλώνουν ότι μία μονάδα που έχει ταξινομηθεί στην κατηγορία j ανήκει στην πραγματικότητα στην κατηγορία i. Βάσει ενός ορισμού που έχει δώσει ο Carroll (1992) σε μία εργασία του για τα σφάλματα μέτρησης συνεχών μεταβλητών, οι πιθανότητες c_{ij} ονομάζονται πιθανότητες βαθμονόμησης (*calibration probabilities*). Σε άλλες περιπτώσεις, μπορούν να αναφέρονται και ως πιθανότητες επαναταξινόμησης (*reclassification probabilities*) (βλ. Buonaccorsi (2009)). Όταν η Y είναι δυωνυμική, οι όροι θετική προγνωστική αξία (*positive predictive value*) για την πιθανότητα $\Pr(Y_\xi = 2 \,|\, Y'_\xi = 2)$ και αρνητική προγνωστική αξία (*negative predictive value*) για την πιθανότητα $\Pr(Y_\xi = 1 \,|\, Y'_\xi = 1)$ χρησιμοποιούνται συχνά στην ιατρική βιβλιογραφία. Κάποιες μέθοδοι εκτιμήσεων που σκοπό έχουν τη διόρθωση της λανθασμένης ταξινόμησης χρησιμοποιούν τις πιθανότητες βαθμονόμησης αντί των πιθανοτήτων λανθασμένης ταξινόμησης.

Μία επιλογή κατά τη χρήση ενός εσωτερικού δείγματος επικύρωσης είναι η στρωματοποίηση αναφορικά με τη μεταβλητή Y' και η επιλογή ενός προκαθορισμένου ποσοστού μονάδων εντός κάθε κατηγορίας της Y' (Deming 1977, Haitovsky and Rapp 1992). Μία ακραία μορφή του σχεδιασμού αυτού είναι η επικύρωση όλων των μονάδων σε κάποιες

κατηγορίες της μεταβλητής Y' και καμία από τις μονάδες σε άλλες κατηγορίες. Αυτό μπορεί να είναι αρκετά αποτελεσματικό εάν ο σκοπός είναι να ελεγχθεί η υπόθεση της μη συσχέτισης μεταξύ δύο μεταβλητών (Zelen and Haitovsky (1991)). Αντίθετα, για κάποιες εξωτερικές μελέτες επικύρωσης μπορεί να είναι πιθανή η στρωματοποίηση ως προς τη μεταβλητή Y εάν για παράδειγμα επιλέξουμε ένα δείγμα ατόμων από ένα σύνολο διαχειριστικών δεδομένων (*administrative records*) στους οποίους και θα χορηγηθεί ακολούθως ένα ερωτηματολόγιο της διεξαγόμενης έρευνας για την απόκτηση της μεταβλητής Y'. Σε τέτοιες περιπτώσεις, μόνο οι πιθανότητες λανθασμένης ταξινόμησης α_{ji} και όχι το διάνυσμα των πραγματικών πληθυσμιακών ποσοστών $\mathbf{P_Y}$ μπορούν να εκτιμηθούν. Η στρωματοποίηση είναι ιδιαιτέρως χρήσιμη όταν κάποια ποσοστά του $\mathbf{P_Y}$ είναι μικρά, και αυτό διότι με τον τρόπο διασφαλίζεται η λήψη δείγματος με επαρκή αριθμό μονάδων από όλες τις κατηγορίες της Y. Ο Marshall (1990) συγκρίνει σχεδιασμούς μελετών επικύρωσης με διαφορετικά περιθώρια αθροίσματα σταθερά.

Το ερώτημα που επίσης προκύπτει είναι η βέλτιστη επιλογή του δείγματος επικύρωσης και του κυρίως δείγματος. Η επιλογή αυτή εξαρτάται τα μέγιστα από την αναλογία των κοστών της χρήσης της βασικής μεθοδολογίας μέτρησης των μονάδων του πληθυσμού, που συνήθως είναι επιρρεπής σε σφάλματα, και της προτιμώμενης διαδικασίας, που θεωρείται ότι μας εξασφαλίζει τις πραγματικές μετρήσεις των δεδομένων, όπως επίσης και από την ακρίβεια των μετρήσεων της έρευνας. Σε κάποιες περιπτώσεις μπορεί να είναι αποτελεσματικό ως προς το κόστος να κατανείμουμε όλα τα δεδομένα σε ένα μικρότερο δείγμα μετρήσεων υψηλότερης ποιότητας χρησιμοποιώντας την προτιμώμενη διαδικασία σε σχέση με το συνδυασμό ενός μεγάλου και λανθασμένα ταξινομημένου κυρίως δείγματος και μίας μικρού μεγέθους μελέτη επικύρωσης (Tenenbein (1970, 1972), Palmgren (1987), Greenland (1988a)).

4.4 Επαναλαμβανόμενες μετρήσεις

Μερικές φορές δεν είναι εφικτή η χρήση της προτιμώμενης διαδικασίας ακόμα και για ένα μικρού μεγέθους δείγμα επικύρωσης. Αυτό μπορεί να οφείλεται στη δυσκολία συλλογής δείγματος επικύρωσης από τον εξεταζόμενο πληθυσμό, αλλά μπορεί να οφείλεται και στο σχεδιασμό συλλογής των δεδομένων από αυτόν. Για παράδειγμα, σε μελέτες που είναι από

σχεδιασμό επαναλαμβανόμενες και βασίζονται σε επαναλαμβανόμενες μετρήσεις των ίδιων ατόμων ενός πληθυσμού σε ανά τακτά ή μεγάλα χρονικά διαστήματα (όπως οι διαχρονικές μελέτες (*longitudinal studies*), μελέτες κοορτής (*cohort studies*) και μελέτες πλαισίου (*panel studies*)) δεν είναι δυνατή η επιλογή ενός τέτοιου δείγματος. Μία εναλλακτική λύση είναι η συλλογή επαναλαμβανόμενων μετρήσεων (*repeated measurements*) των λανθασμένα ταξινομημένων μεταβλητών. Σε μία διεξαγόμενη έρευνα αυτό μπορεί να σημαίνει τη λήψη δεδομένων για κάποια από τα εξεταζόμενα άτομα του πληθυσμού τουλάχιστον μία φορά επιπλέον. (Forsman and Schreiner (1991)). Παρόλο που είναι δυνατή η ύπαρξη σφαλμάτων λανθασμένης ταξινόμησης σε όλες τις επαναλαμβανόμενες μετρήσεις, μπορούμε να εξάγουμε πληροφορίες για τις παραμέτρους λανθασμένης ταξινόμησης με τη χρήση στατιστικών μεθόδων κάτω από τις απαραίτητες προϋποθέσεις για τη σχέση ανάμεσα στις επαναλαμβανόμενες μετρήσεις και τις πραγματικές μεταβλητές. Στο 7^{ο} Κεφάλαιο θα δειχτεί ότι συχνά είναι απαραίτητη η χρήση τουλάχιστον τριών επαναλαμβανόμενων μετρήσεων, παρόλο που σε ορισμένες περιπτώσεις δύο επαναλήψεις είναι αρκετές. Συχνά θεωρούμε ότι αυτές οι μετρήσεις είναι ανεξάρτητες μεταξύ τους δεδομένης της πραγματικής τιμής της μεταβλητής την οποία και έχουν σκοπό να μετρήσουν. Το ευτυχές είναι ότι δεν είναι απαραίτητο οι επαναλαμβανόμενες μετρήσεις να έχουν την ίδια κατανομή και για το λόγο αυτό μπορούν να ληφθούν χρησιμοποιώντας διαφορετικές μεθόδους. Τέτοιες μέθοδοι, σε επίπεδο ανάλυσης ερευνών, δίνονται μέσω παραδειγμάτων από τους Chua and Fuller (1987) (διεξαγωγή επαναλαμβανόμενων συνεντεύξεων από διαφορετικά άτομα) και Harper (1964) (επιλογή ερωτηματολογίου με διαφορετικής μορφής και δομής ερωτήσεις για μία συγκεκριμένη εξεταζόμενη μεταβλητή). Μπορούμε ακόμα να υποθέσουμε ότι η προτιμώμενη διαδικασία, που θεωρητικά καταγράφει τις πραγματικές τιμές των εξεταζόμενων μεταβλητών για τον πληθυσμό, δεν είναι ακριβής και περιέχει εσωτερικά σφάλματα λανθασμένης ταξινόμησης (σφάλματα που έχουν να κάνουν με την ίδια τη διαδικασία και όχι τα συλλεγόμενα δεδομένα) (Watcholder et al. (1993)). Σε αυτές τις περιπτώσεις κάποιες από τις μετρήσεις δεν προέρχονται από τα εξεταζόμενα άτομα του πληθυσμού αλλά από άλλες πηγές. Τότε εύλογο είναι να θεωρήσουμε ότι οι μετρήσεις αυτές είναι υπό συνθήκη ανεξάρτητες δεδομένης της πραγματικής μεταβλητής.

Όπως και σε μία μελέτη επικύρωσης, το δείγμα για το οποίο λαμβάνουμε τις επαναλαμβανόμενες μετρήσεις μπορεί να είναι είτε εσωτερικό είτε εξωτερικό, οπότε και ισχύουν όσα αναφέρθηκαν στην Παράγραφο 4.3. Ο Fuller ((1990), Section 5) πραγματεύεται

της επιλογής ενός δείγματος επικύρωσης n_v για δύο επαναλαμβανόμενες μετρήσεις για μία δυαδική μεταβλητή απόκρισης.

ΚΕΦΑΛΑΙΟ 5

Διόρθωση της επίδρασης λανθασμένης ταξινόμησης με χρήση πινακικών μεθόδων

5.1 Απλές μέθοδοι πινάκων με χρήση δειγμάτων επικύρωσης

Η χρήση ενός εκτιμητή μέγιστης πιθανοφάνειας αποτελεί την πιο αποτελεσματική μέθοδο διόρθωσης της λανθασμένης ταξινόμησης, παρουσιάζει συγχρόνως σχετική υπολογιστική δυσκολία εφόσον απαιτεί έναν επαναληπτικό αλγόριθμο επίλυσης για ένα σύνολο μη γραμμικών εξισώσεων που υπόκεινται σε περιορισμούς. Λόγω της πολυπλοκότητας των υπολογισμών, οι ερευνητές κατέφυγαν στη χρήση άλλων μεθόδων, πιο ευκολονόητων και υπολογιστικά απλούστερων. Έτσι οδηγήθηκαν στον πιο ξεκάθαρο και διαισθητικά φυσικό τρόπο για τη διόρθωση της λανθασμένης ταξινόμησης, μέσω απλών μεθόδων ανάστροφου υπολογισμού (*back-calculation methods*). Οι διορθωμένες μετρήσεις για κάθε κελί λαμβάνονται μέσω πολλαπλασιασμού ενός διανύσματος που περιέχει τις παρατηρούμενες μετρήσεις με έναν πίνακα διόρθωσης και η ανάλυση των δεδομένων γίνεται βάσει των μετασχηματισμένων τιμών. Κατά τον τρόπο αυτό, οι εκτιμήσεις που λαμβάνουμε έχουν μειωμένη μεροληψία Τέτοιες μέθοδοι ονομάζονται πινακικοί μέθοδοι διόρθωσης (*matrix methods of adjustment*) και έχουν αποτελέσει αντικείμενο έρευνας πολλών ερευνητών, βρίσκοντας εφαρμογές σε επιστημονικά πεδία όπως η ιατρική και η επιδημιολογία (βλ. Barron (1977), Marshall (1990), Lyles (2002)). . Σε αυτό το Κεφάλαιο, θα εξετάσουμε πινακικές μεθόδους διόρθωσης που βασίζονται στη χρήση δειγμάτων επικύρωσης. Θα αναφερθούμε πρώτα σε μεθόδους όπου ο πίνακας διόρθωσης είναι ο αντίστροφος πίνακας του πίνακα λανθασμένης ταξινόμησης. Ακολούθως, θα θεωρήσουμε τη μέθοδο διόρθωσης όπου τα στοιχεία του πίνακα διόρθωσης είναι οι εκτιμώμενες πιθανότητες βαθμονόμησης των πραγματικών μεταβλητών δεδομένων των εσφαλμένα ταξινομημένων πιθανοτήτων. Τέλος, θα προβούμε σε συγκρίσεις μεταξύ των δύο μεθόδων για την εύρεση του καλύτερου εκτιμητή.

5.2 Διορθώσεις με τη χρήση πιθανοτήτων λανθασμένης ταξινόμησης

Παρόλο που η χρήση τους στην πράξη δεν είναι και τόσο συχνή, οι μέθοδοι που χρησιμοποιούν ως πίνακα διόρθωσης τον αντίστροφο πίνακα του πίνακα λανθασμένης ταξινόμησης ίσως αποτελούν την πιο διαδεδομένη μέθοδο για τη διόρθωση εσφαλμένα ταξινομημένων δεδομένων. Αρκετά επιδημιολογικά εγχειρίδια προτείνουν τη χρήση της μεθόδου αυτής (Kleinbaum, Kupper, and Morgenstern (1982), Schlesselman (1982), Rothman and Greenland (1998)). Εξαιρετικές περιγραφές των μεθόδων αυτών δίνονται από τους Greenland (1988a, Appendix 1) και Selen (1986). Στη συνέχεια παραθέτουμε τη μέθοδο διόρθωσης λανθασμένης ταξινόμησης που δίνεται από τον Greenland (1988, Παράρτημα I). Η μέθοδος είναι όμοια ως προς τη μέθοδο που παρουσιάζει ο Selen, με τη διαφορά ότι ο Greenland επικεντρώνεται στις μετρήσεις των κελιών των υποομάδων και όχι στους μέσους των υποομάδων και η πολυπλοκότητα υπολογισμού του πίνακα λανθασμένης ταξινόμησης είναι συνεπώς μικρότερη.

Ας υποθέσουμε ότι κάθε μονάδα ξ ενός μεγάλου πληθυσμού κατατάσσεται μονοσήμαντα σε μία από r υποομάδες ή κατηγορίες, ξένων μεταξύ τους. Το κυρίως δείγμα του πληθυσμού αποτελείται από n^p παρατηρούμενες μονάδες που έχουν ταξινομηθεί μέσω μίας επιρρεπούς σε σφάλματα διαδικασίας ταξινόμησης, δηλαδή n^p μονάδες για τις οποίες μόνο η λανθασμένα ταξινομημένη κατηγορία είναι γνωστή. Οι τιμές των μεταβλητών για την ορθή και λανθασμένη ταξινόμηση τους στην κατηγορία i της αποκριτικής μεταβλητής Y είναι $Y_\xi = i$ και $Y'_\xi = j$ αντίστοιχα, όπως αυτές ορίζονται στο 2ο Κεφάλαιο. Ας σημειωθεί εδώ ότι οι r κατηγορίες της μεταβλητής Y μπορεί να αναπαριστούν διατεταγμένα, ως προς κάποιο κριτήριο, κελιά μίας διασταυρούμενης ταξινόμησης (*cross-classification*) αρκετών μεταβλητών. Η κατανομή του συνόλου των εξεταζόμενων μονάδων στα r^2 κελιά, που ορίζεται από τις πραγματικές και τις εσφαλμένα καταγεγραμμένες κατηγορίες, είναι πολυωνυμική. Ορίζουμε τα διανύσματα $\mathbf{Y} = (Y_\xi(1),...,Y_\xi(r))^c$ με $i = 1,...,r$ και $\mathbf{Y'} = (Y'_\xi(1),...,Y'_\xi(r))^c$ με $j = 1,...,r$, καθώς και τον πίνακα λανθασμένης ταξινόμησης $\mathbf{A} = [\alpha_{ji}]_{r\times r}$, ο οποίος περιέχει τις πιθανότητες των μονάδων του πληθυσμού που ανήκουν στην i κατηγορία της μεταβλητής Y να έχουν ταξινομηθεί εσφαλμένα στην j κατηγορία (να ταξινομούνται με άλλα λόγια στην j κατηγορία της μεταβλητής Y'). Επιπροσθέτως, εάν θέσουμε τις αναμενόμενες τιμές για τα ορθά και εσφαλμένα ταξινομημένα δεδομένα ως

$m(i) = \mathrm{E}(Y_\xi(i))$ και $m'(j) = \mathrm{E}(Y'_\xi(j))$ αντίστοιχα, το διάνυσμα $\mathbf{m} = (m(1),...,m(r))^c$ των αναμενόμενων ορθά ταξινομημένων δεδομένων και το διάνυσμα $\mathbf{m}' = (m'(1),...,m'(r))^c$ των αναμενόμενων εσφαλμένα ταξινομημένων δεδομένων συνδέονται μέσω της σχέσης

$$\mathbf{m}' = \mathbf{A}\mathbf{m} \tag{5.1}$$

Κάτω από την υπόθεση ότι ο **A** είναι μη ιδιάζων, το οποίο και σημαίνει ότι η λανθασμένη ταξινόμηση δεν είναι εντελώς τυχαία, η σχέση (5.1) γίνεται

$$\mathbf{m} = \mathbf{A}^{-1}\mathbf{m}' \tag{5.2}$$

Το σύνηθες πρόβλημα που αντιμετωπίζουμε είναι πως ο πίνακας λανθασμένης ταξινόμησης **A** συνήθως δεν είναι γνωστός. Λόγω αυτού, οδηγούμαστε στην εύρεση ενός εκτιμητή $\hat{\mathbf{A}}$ του πίνακα με τη χρήση δεδομένων επικύρωσης από ένα ανεξάρτητο του κυρίως δείγματος εσωτερικού ή εξωτερικού δείγματος επικύρωσης με πλήθος μονάδων n^v. Για ένα τέτοιο δείγμα είναι γνωστό ότι ένας συνεπής εκτιμητής της πιθανότητας λανθασμένης ταξινόμησης είναι ο

$$\hat{\alpha}_{ji} = \frac{n^v_{ji}}{n^v}, \tag{5.3}$$

ο όποιος όταν το δείγμα επικύρωσης είναι εσωτερικό αποτελεί και τον εκτιμητή μέγιστης πιθανοφάνειας της πιθανότητας α_{ji} (βλ. Selen (1986) και θεωρία Μαρκοβιανών αλυσίδων). Θέτοντας $\hat{\mathbf{A}} = [\hat{\alpha}_{ji}]_{r\times r}$, και θεωρώντας ότι είναι ένας μη ιδιάζων επαρκής εκτιμητής του πίνακα $\hat{\mathbf{A}}$, είναι δυνατή η εύρεση μίας διορθωμένης εκτίμησης του διανύσματος **Y** μέσω του διανύσματος των παρατηρούμενων στοιχείων $\mathbf{Y}' = (Y'_\xi(1),...,Y'_\xi(r))^c$ που δίνεται από τον τύπο

$$\hat{\mathbf{Y}}^r = \hat{\mathbf{A}}^{-1}\mathbf{Y}'. \tag{5.4}$$

Ας σημειωθεί εδώ ότι η παραπάνω σχέση δίνει έναν συνεπή εκτιμητή του διανύσματος **Y** των πραγματικών μεταβλητών και κατά την περίπτωση χρήσης εξωτερικού δείγματος επικύρωσης. Επίσης, θεωρούμε πως η πιθανότητα ο πίνακας $\hat{\mathbf{A}}$ να είναι ιδιάζων (μη αντιστρέψιμος) είναι αμελητέα, απαίτηση ελάχιστα περιοριστική αν η μέθοδος ταξινόμησης είναι λογική και το μέγεθος του δείγματος αρκετά μεγάλο.

Εάν το δείγμα επικύρωσης n^v είναι εσωτερικό, ορίζουμε n^v_{ji} τις τιμές των i στοιχείων που έχουν ταξινομηθεί ως j στοιχεία. Η περιθώρια κατανομή των στοιχείων για τις πραγματικές κατηγορίες δίνεται μέσω του διανύσματος $\mathbf{n}^{\mathbf{v}}_{\mathbf{t}} = (n^v_{t,1},...,n^v_{t,r})^c$ με $n^v_{t,i} = \sum_{j=1}^{r} n^v_{ji}$.

Η αντίστοιχη περιθώρια κατανομή των στοιχείων στις εσφαλμένα καταγεγραμμένες κατηγορίες δίνεται από το διάνυσμα $\mathbf{n_e^v} = (n_{e,1}^v, ..., n_{e,r}^v)^c$ με $n_{e,j}^v = \sum_{i=1}^r n_{ji}^v$. Λαμβάνοντας υπ'όψιν ότι γνωρίζουμε τις πραγματικές κατηγορίες για τα στοιχεία που έχουμε λάβει μέσω του δείγματος επικύρωσης, ένας εκτιμητής των πραγματικών τιμών $\mathbf{n_t^v}$ δίνεται από τον τύπο

$$\hat{\mathbf{n}}_\mathbf{t}^\mathbf{v} = \hat{\mathbf{A}}^{-1}\mathbf{n}_e^\mathbf{v} + \mathbf{n}_\mathbf{t}^\mathbf{v} \tag{5.5}$$

(Selen (1986)). Συνδυάζοντας τους τύπους (5.4) και (5.5) προκύπτει ο εξής τροποποιημένος εκτιμητής

$$\hat{\mathbf{Y}}^d = \frac{1}{n^p + n^v}[n^p(\hat{\mathbf{A}}^{-1}\mathbf{Y}') + n^v\mathbf{Y}^v], \tag{5.6}$$

όπου $\mathbf{Y}^v$ είναι ο εκτιμητής του διανύσματος $\mathbf{Y}$ που προέρχεται από τις πραγματικές τιμές του δείγματος επικύρωσης.

Θέλοντας να ανάγουμε τα ανωτέρω συμπεράσματα σε ότι αφορά πληθυσμιακά ποσοστά, μέσω των σχέσεων (5.4) και (5.6) εύκολα καταλήγουμε στους αντίστοιχους εκτιμητές $\hat{\mathbf{P}}_\mathbf{Y}^r$ και $\hat{\mathbf{P}}_\mathbf{Y}^d$ του διανύσματος $\mathbf{P_Y} = (P_Y(1), ..., P_Y(r))^c$ που αντιστοιχεί στα πραγματικά ποσοστά των μονάδων του πληθυσμού που ανήκουν στη μεταβλητή Y

$$\hat{\mathbf{P}}_\mathbf{Y}^r = \hat{\mathbf{A}}^{-1}\mathbf{P}_{\mathbf{Y}'} \tag{5.7}$$

$$\hat{\mathbf{P}}_\mathbf{Y}^d = \frac{1}{n^p + n^v}[n^p(\hat{\mathbf{A}}^{-1}\mathbf{P}_{\mathbf{Y}'}) + n^v\mathbf{P}_\mathbf{Y}^v], \tag{5.8}$$

με $\mathbf{P}_{\mathbf{Y}'}$ έναν εκτιμητή του $\mathbf{P_Y}$ που υπολογίζεται από το συγκεκριμένο δείγμα, όπως αυτός ορίζεται στη σχέση (2.2), και $\mathbf{P}_\mathbf{Y}^v$ τον εκτιμητή του $\mathbf{P_Y}$ που προέρχεται από τις πραγματικές τιμές του δείγματος επικύρωσης n^v.

Όσον αφορά τη μεταβλητότητα των παραπάνω εκτιμητών, αυτή μπορεί να υπολογιστεί μέσω της εύρεσης των αντίστοιχων πινάκων διακύμανσης – συνδιακύμανσης για κάθε μία εξ αυτών. Λαμβάνοντας υπ'όψιν πως ο πίνακας $\hat{\mathbf{A}}$ έχει θεωρηθεί επαρκής εκτιμητής του $\mathbf{A}$ και υποθέτοντας ανεξαρτησία μεταξύ των $\mathbf{Y}'$ και $\hat{\mathbf{A}}$, ο Greenland (1988a) έδειξε ότι με τη χρήση της πολυμεταβλητής μεθόδου δέλτα (βλ. Bishop et. al (1975), Selen (1986, σχέση 4.7)) ο ασυμπτωτικός πίνακας διακύμανσης-συνδιακύμανσης για τον εκτιμητή $\hat{\mathbf{P}}_\mathbf{Y}^r$ είναι ο εξής

$$\text{cov}(\hat{\mathbf{P}}_\mathbf{Y}^r) = \text{cov}[\hat{\mathbf{A}}^{-1}\text{E}(\mathbf{P}_{\mathbf{Y}'})] + \hat{\mathbf{A}}^{-1}\,\text{cov}(\mathbf{P}_{\mathbf{Y}'})(\hat{\mathbf{A}}^{-1})^c\ . \tag{5.9}$$

Μέσω μίας άλλης εφαρμογής της μεθόδου δέλτα, μία προσέγγιση του h στοιχείου του πίνακα $\hat{\mathbf{A}}^{-1}\mathrm{E}(\mathbf{P}_{\mathbf{Y}'})$, έστω g_h, δίνεται από τον τύπο

$$\sum_{ijkl}\left(\frac{\partial g_h}{\partial \hat{\alpha}_{ij}}\right)\left(\frac{\partial g_h}{\partial \hat{\alpha}_{kl}}\right)\mathrm{cov}(\hat{\alpha}_{ij},\hat{\alpha}_{kl}),$$

όπου $\partial g_h/\partial\hat{\alpha}_{ij}$ είναι η μερική παράγωγος του g_h δεδομένου ότι τα στοιχεία $\hat{\alpha}_{ij}$ είναι εκτιμήσεις των α_{ij} και ομοίως για $\partial g_h/\partial\hat{\alpha}_{kl}$. Ένας συνεπής εκτιμητής του πίνακα $\mathrm{cov}(\hat{\mathbf{P}}_{\mathbf{Y}}^{r})$ δίνεται αντικαθιστώντας τις παραμέτρους της σχέσης (5.9) με συνεπείς εκτιμητές των παραμέτρων αυτών.

Αντίστοιχα, για τον εκτιμητή $\hat{\mathbf{P}}_{\mathbf{Y}}^{d}$ ο ασυμπτωτικός πίνακας διακύμανσης-συνδιακύμανσης δίνεται από την ακόλουθη σχέση

$$\mathrm{cov}(\hat{\mathbf{P}}_{\mathbf{Y}}^{d})=\mathrm{cov}[\hat{\mathbf{A}}^{-1}\mathrm{E}(\mathbf{P}_{\mathbf{Y}'})]+(\hat{\mathbf{A}}^{-1})^{c}\,\mathrm{cov}(\mathbf{P}_{\mathbf{Y}'})\hat{\mathbf{A}}^{-1}+\mathrm{cov}(\mathbf{P}_{\mathbf{Y}}^{v})+2(\hat{\mathbf{A}}^{-1})^{c}\,\mathrm{cov}(\mathbf{P}_{\mathbf{Y}'},\mathbf{P}_{\mathbf{Y}}^{v})$$

(βλ. Greenland (1988a)). Ομοίως, για τον υπολογισμό ενός συνεπή εκτιμητή του πίνακα $\mathrm{cov}(\hat{\mathbf{P}}_{\mathbf{Y}}^{d})$ ισχύουν παρόμοιες προϋπόθεσης με αυτές που ισχύουν για τον πίνακα $\mathrm{cov}(\hat{\mathbf{P}}_{\mathbf{Y}}^{r})$.

5.3 Διορθώσεις με τη χρήση πιθανοτήτων βαθμονόμησης

Στο προηγούμενο Κεφάλαιο οι μέθοδοι διόρθωσης με χρήση πινάκων ορίστηκαν βάσει του αντίστροφου πίνακα λανθασμένης ταξινόμησης $\mathbf{A}=[\alpha_{ji}]_{r\times r}$, όπου α_{ji} είναι οι πιθανότητες λανθασμένης ταξινόμησης. Μία εναλλακτική μέθοδος διόρθωσης είναι ο πολλαπλασιασμός των παρατηρούμενων πληθυσμιακών ποσοστών με έναν εκτιμητή $\hat{\mathbf{C}}$ του πίνακα βαθμονόμησης $\mathbf{C}=[c_{ij}]_{r\times r}$, με στοιχεία τις πιθανότητες βαθμονόμησης $c_{ij}=\Pr(Y_{\xi}=i\,|\,Y'_{\xi}=j)$.

Ορίζουμε καταρχάς τις βασικές προϋποθέσεις που πρέπει να πληρούνται ώστε να χρησιμοποιήσουμε πιθανότητες βαθμονόμησης. Ας υποθέσουμε ότι το δείγμα επικύρωσης n^{v} είναι εσωτερικό και ότι η μέθοδος δειγματοληψίας που χρησιμοποιείται είναι τέτοια ώστε η κατανομή των πραγματικών μετρήσεων ανά κελί για κάθε δείγμα είναι πολυωνυμική. Μπορούμε να αντιληφθούμε καλύτερα τη φύση της υπεισερχόμενης μεροληψίας εάν θεωρήσουμε για τα εξεταζόμενα δεδομένα το πολυωνυμικό σχήμα που εκτιμά $r-1$ πληθυσμιακά ποσοστά μέσω της εκτίμησης $r-1$ ανεξάρτητων δυωνυμικών κατανομών.

Για την εύρεση ενός εκτιμητή με τη βοήθεια των πιθανοτήτων βαθμονόμησης θα βασιστούμε στην μεθοδολογία που παρουσιάζεται από τον Tenenbein (1970, 1971, 1972). Οι υποθέσεις που κάνουμε για το κυρίως δείγμα του πληθυσμού μεγέθους n^p, καθώς και για τις μεταβλητές ορθής και λανθασμένης ταξινόμησης Y και Y' είναι ίδιες με αυτές της προηγούμενης Παραγράφου. Ας υποθέσουμε ότι το δείγμα επικύρωσης n^v που χρησιμοποιείται είναι εσωτερικό και ότι η μέθοδος δειγματοληψίας είναι τέτοια ώστε η κατανομή των πραγματικών μετρήσεων ανά κελί για κάθε δείγμα είναι πολυωνυμική. Για τα δείγματα n^v και n^p ορίζουμε επίσης τα ακόλουθα

- n_{ij}^v, το πλήθος των μονάδων του δείγματος επικύρωσης n^v που ανήκουν στην i κατηγορία και έχουν ταξινομηθεί εσφαλμένα στην j κατηγορία, με $\sum_{i=1}^{r}\sum_{j=1}^{r} n_{ij}^v = n^v$
- n_{i+}^v, το πλήθος των μονάδων του δείγματος επικύρωσης n^v που ανήκουν στην i κατηγορία, με $\sum_{j=1}^{r} n_{ij}^v = n_{i+}^v$
- n_{+j}^v, το πλήθος των μονάδων του δείγματος επικύρωσης n^v που έχουν ταξινομηθεί εσφαλμένα στην j κατηγορία, με $\sum_{i=1}^{r} n_{ij}^v = n_{+j}^v$
- n_j^p, το πλήθος των μονάδων του κυρίως δείγματος n^p που έχουν ταξινομηθεί εσφαλμένα στην j κατηγορία, με $\sum_{i=1}^{r} n_j^p = n^p$

Θεωρώντας τις πιθανότητες λανθασμένης ταξινόμησης $\alpha_{ji} = \Pr(Y'_\xi = j \mid Y_\xi = i)$ και n_j^p τον αριθμό των ξ μονάδων του πληθυσμού που έχουν εσφαλμένα ταξινομηθεί μέσω της μεταβλητής Y' στην j κατηγορία, οι εκτιμητές μέγιστης πιθανοφάνειας των πραγματικών ποσοστών P_Y και οι εκτιμήσεις των πιθανοτήτων λανθασμένης ταξινόμησης α_{ji} δίνονται από τους ακόλουθους τύπους

$$\hat{P}_{Y(i)}^c = \sum_{j=1}^{r} n_{ij}^v \frac{(n_j^p + n_{+j}^v)}{n_{+j}^v (n^p + n^v)} \tag{5.10}$$

$$\hat{\alpha}_{ji} = n_{ij}^v \frac{(n_j^p + n_{+j}^v)}{n_{+j}^v (n^p + n^v) P_{Y(i)}^c} \tag{5.11}$$

(βλ. Tenenbein (1972), Παράρτημα Α)

Εξ ορισμού, οι πιθανότητες βαθμονόμησης είναι $c_{ij} = \Pr(Y_\xi = i \mid Y'_\xi = j)$. Μέσω του θεωρήματος του Bayes, οι πιθανότητες λανθασμένης ταξινόμησης συνδέονται με τις πιθανότητες βαθμονόμησης με τον ακόλουθο τύπο

$$c_{ij} = \Pr(Y_\xi = i \mid Y'_\xi = j) = \frac{P_Y(i)\alpha_{ji}}{\sum_{l=1}^{r} P_Y(l)\alpha_{jl}} \qquad (5.12)$$

για κάθε $i, j, l = 1, ..., r$. Συνδυάζοντας τις σχέσεις (5.10), (5.11) και (5.12), ο εκτιμητής μέγιστης πιθανοφάνειας του $\mathbf{P_Y}$ είναι

$$\hat{\mathbf{P}}^c_{\mathbf{Y}} = \frac{1}{n^p + n^v}[n^p(\hat{\mathbf{C}}\mathbf{P}'_{\mathbf{Y}}) + n^v\mathbf{P}^v_{\mathbf{Y}}] \qquad (5.13)$$

(βλ. Kuha and Skinner (1997)), όπου $\hat{\mathbf{C}} = [\hat{c}_{ij}]_{r\times r}$ είναι ο εκτιμητής του πίνακα $\mathbf{C} = [c_{ij}]_{r\times r}$ που υπολογίζεται βάσει του δείγματος επικύρωσης και $\mathbf{P}'_{\mathbf{Y}}$, $\mathbf{P}^v_{\mathbf{Y}}$ οι εκτιμητές του $\mathbf{P_Y}$ όπως αυτοί ορίζονται και για τον εκτιμητή $\hat{\mathbf{P}}^d_{\mathbf{Y}}$. Οι εκτιμώμενες πραγματικές μετρήσεις για κάθε κατηγορία του κυρίως δείγματος λαμβάνονται κατά αυτόν τον τρόπο με την κατανομή των μονάδων από τις λανθασμένα ταξινομημένες κατηγορίες στις πραγματικές κατηγορίες της προς εξέταση μεταβλητής σύμφωνα με τις εκτιμώμενες πιθανότητες βαθμονόμησης. Με τις κατάλληλες τροποποιήσεις, τα εξαγόμενα αποτελέσματα ισχύουν επίσης όταν τα δείγματα είναι στρωματοποιημένα (*stratified*) ως προς μία μεταβλητή που δεν είναι λανθασμένα ταξινομημένη. Κατά την περίπτωση αυτή, η από κοινού κατανομή των μεταβλητών είναι γινόμενο διάφορων πολυωνυμικών κατανομών. Ο ασυμπτωτικός πίνακας συνδιακύμανσης $\text{cov}(\hat{\mathbf{P}}^c_{\mathbf{Y}}) = \text{cov}(\hat{P}^c_{Y(1)}, ..., \hat{P}^c_{Y(i)}, ..., \hat{P}^c_{Y(r)})^c$ του εκτιμητή $\hat{\mathbf{P}}^c_{\mathbf{Y}}$ δίνεται μέσω του υπολογισμού της ασυμπτωτικής διακύμανσης του πληθυσμιακού ποσοστού $\hat{P}^c_{Y(i)}$, όπου

$$V(\hat{P}^c_{Y(i)}) = \frac{P_{Y(i)}(1-P_{Y(i)})}{n^v}(1-K_i) + \frac{P_{Y(i)}(1-P_{Y(i)})}{n^v + n^p}K_i, \quad i = 1, .., r$$

ή μετά από πράξεις

$$V(\hat{P}^c_{Y(i)}) = \frac{P_{Y(i)}(1-P_{Y(i)})}{n^v}[1-(1-f)K_i], \qquad i = 1, .., r \qquad (5.14)$$

με $f = n^v/(n^v + n^p)$ το ποσοστό των δειγματικών μονάδων που έχουν κατανεμηθεί (που ανήκουν) στο δείγμα επικύρωσης και

$$K_i = \frac{P_{Y(i)}(\sum_{j=1}^{r} a_{ji}/P_{Y'(j)} - 1)}{1 - P_{Y(i)}}$$

τον συντελεστή αξιοπιστίας της *i* κατηγορίας της ορθά ταξινομημένης μεταβλητής *Y*, όπως αυτός έχει οριστεί στο Κεφάλαιο 3 (βλ. Tenenbein (1972), Παράρτημα Β).

5.4 Συγκρίσεις μεθόδων

Όπως διαπιστώσαμε, όταν χρησιμοποιούμε πολυωνυμικά δεδομένα ο εκτιμητής $\hat{\mathbf{P}}_{\mathbf{Y}}^{c}$ της σχέσης (5.13) είναι ο εκτιμητής μέγιστης πιθανοφάνειας του $\hat{\mathbf{P}}_{\mathbf{Y}}$. Αυτό συνεπάγεται ότι είναι ένας ασυμπτωτικά πιο αποτελεσματικός εκτιμητής από τον εκτιμητή $\hat{\mathbf{P}}_{\mathbf{Y}}^{d}$ που δίνεται μέσω της σχέσης (5.8). Το κατά πόσον περισσότερο αποτελεσματικός είναι, αποτελεί το αντικείμενο έρευνας αυτής της παραγράφου. Η σύγκριση των δύο εκτιμητών θα γίνει για την απλή δυωνυμική περίπτωση και όπως θα δούμε ο εκτιμητής $\hat{\mathbf{P}}_{\mathbf{Y}}^{d}$ είναι εμφανώς ανώτερος του $\hat{\mathbf{P}}_{\mathbf{Y}}^{d}$.

Έστω *Y* μια διχοτομική μεταβλητή με πληθυσμιακά ποσοστά $\mathbf{P}_{\mathbf{Y}} = (P_{Y(1)}, P_{Y(2)}) = (1-P, P)$, με το ποσοστό $P_{Y(2)}$ που δηλώνει την παρουσία του εξεταζόμενου χαρακτηριστικού να αποτελεί την απαιτούμενη προς εκτίμηση ποσότητα. Υποθέτουμε ότι η μεταβλητή *Y* ταξινομείται λανθασμένα μέσω της μεταβλητής Y' με ευαισθησία $sen = \Pr(Y'_{\xi} = 2 \mid Y_{\xi} = 2)$ και ειδικότητα $sp = \Pr(Y'_{\xi} = 1 \mid Y_{\xi} = 1)$ για κάθε μονάδα ξ του πληθυσμού, με τη βοήθεια των οποίων μπορούμε να ορίσουμε τις αντίστοιχες πιθανότητες λανθασμένης ταξινόμησης $\theta_Y = \Pr(Y'_{\xi} = 1 \mid Y_{\xi} = 2) = 1 - sen$ και $\phi_Y = \Pr(Y'_{\xi} = 2 \mid Y_{\xi} = 1) = 1 - sp$. Ας θεωρήσουμε ότι έχουμε ένα κυρίως δείγμα αποτελούμενο από n_p μονάδες και ένα ανεξάρτητο εσωτερικό δείγμα επικύρωσης n_v μονάδων. Και τα δύο δείγματα λαμβάνονται από τον εξεταζόμενο πληθυσμό με χρήση απλής τυχαίας δειγματοληψίας. Βάσει των παραπάνω, τα δεδομένα αναπαριστώνται στον πίνακα που ακολουθεί (Πίνακας 5.1)

ΠΙΝΑΚΑΣ 5.1

Πίνακας απλών πινακικών μεθόδων διόρθωσης λανθασμένης ταξινόμησης για την 'εσφαλμένη' μεταβλητή και την 'πραγματική' μεταβλητή

			Y'		
			1	2	
Δείγμα επικύρωσης	Y	1	n_{11}^{ν}	n_{12}^{ν}	n_{1+}^{ν}
		2	n_{21}^{ν}	n_{22}^{ν}	n_{2+}^{ν}
			n_{+1}^{ν}	n_{+2}^{ν}	n^{ν}
Κυρίως δείγμα	Y	$1+2$	n_1^p	n_2^p	n^p

Ο μεροληπτικός εκτιμητής του P βάσει της λανθασμένα ταξινομημένης μεταβλητής Y' για το κυρίως δείγμα είναι $\hat{P}' = n_2^p / n^p$. Η αναμενόμενη τιμή του είναι $P' = \phi_Y(1-P) + (1-\theta_Y)P$. Εφόσον το δείγμα επικύρωσης είναι εσωτερικό, μπορούμε να μέσω αυτού να εκτιμήσουμε το πληθυσμιακό ποσοστό P χρησιμοποιώντας τον αμερόληπτο εκτιμητή $\hat{P}^{\nu} = n_{2+}^{\nu} / n^{\nu}$. Η διακύμανση του εκτιμητή είναι $Var(\hat{P}^{\nu}) = [P(1-P)]/n^{\nu}$ αγνοώντας τον παράγοντα διόρθωσης για πεπερασμένους πληθυσμούς. Θα χρησιμοποιήσουμε τον $\hat{P}^{\nu}$ ως σημείο αναφοράς για την αξιολόγηση της αποτελεσματικότητας των εκτιμητών των σχέσεων (5.8) και (5.13).

Έστω f το ποσοστό των δειγματικών μονάδων που έχουν που ανήκουν στο δείγμα επικύρωσης όπως αυτό ορίστηκε στην Παράγραφο 5.3 . Από τη σχέση (5.13), ο εκτιμητής μέγιστης πιθανοφάνειας για το ποσοστό P είναι η εξής

$$\hat{P}^c = (1-f)\tilde{P}^c + f \cdot \hat{P}^{\nu} \tag{5.15}$$

όπου

$$\tilde{P}^c = \frac{1}{n^p}\left(\frac{n_{21}}{n_{+1}} n_1^p + \frac{n_{22}}{n_{+2}} n_2^p\right). \tag{5.16}$$

Η ασυμπτωτική διακύμανση του είναι

$$Var(\hat{P}^c) = Var(\hat{P}^{\nu})[1-(1-f)K] \tag{5.17}$$

όπου

$$K = \frac{P(1-P)(1-\theta_Y - \phi_Y)^2}{P'(1-P')} \quad . \tag{5.18}$$

είναι ο συντελεστής αξιοπιστίας. Επειδή είναι το τετράγωνο της συσχέτισης μεταξύ των μεταβλητών Y και Y', αποτελεί ουσιαστικά ένα περιληπτικό μέτρο ένδειξης της ποιότητας των μετρήσεων (Tenenbein 1970). Εφόσον $0 \le K \le 1$, ο εκτιμητής $\hat{P}^c$ δεν είναι ποτέ λιγότερο ασυμπτωτικά αποτελεσματικός σε σχέση με τον $\hat{P}^v$. Εάν αυξηθεί το πλήθος των μονάδων του κυρίως δείγματος n^p ενώ το πλήθος των μονάδων του δείγματος επικύρωσης n^v παραμένει σταθερό, τότε η ασυμπτωτική σχετική αποτελεσματικότητα (*asymptotic relative efficiency*) $Var(\hat{P}^v)/Var(\hat{P}^c)$ τείνει στο $1/(1-K)$.

Ο εκτιμητής του P βάσει της σχέσης (14) δίνεται από τον ακόλουθο τύπο

$$\hat{P}^d = (1-f)\tilde{P}^d + f \cdot \hat{P}^v \tag{5.19}$$

όπου

$$\tilde{P}^d = \frac{\hat{P}' - \hat{\phi}_Y}{1 - \hat{\theta}_Y - \hat{\phi}_Y} \tag{5.20}$$

και $\hat{\theta}_Y = n_{12}/n_{+2}$, $\hat{\varphi}_Y = n_{21}/n_{+1}$ είναι οι εκτιμώμενες πιθανότητες λανθασμένης ταξινόμησης $\hat{\theta}_Y$ και $\hat{\phi}_Y$ αντίστοιχα υπολογισμένες βάσει του δείγματος επικύρωσης. Η ασυμπτωτική του διακύμανση (Mashall 1990) είναι

$$\begin{aligned} Var(\hat{P}^d) &= Var(\hat{P}^v)\left[1-(1-f)\cdot\frac{2K-1}{K}\right] \\ &= Var(\hat{P}^c) + Var(\hat{P}^v)\left[(1-f)\cdot\frac{(K-1)^2}{K}\right] \quad . \end{aligned} \tag{5.21}$$

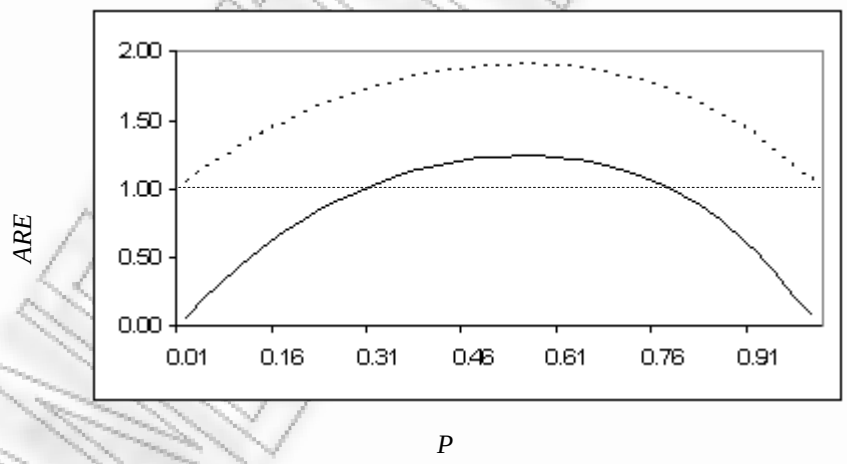
Προφανώς ισχύει $Var(\hat{P}^d) \ge Var(\hat{P}^c)$ για όλες τις τιμές των παραμέτρων. Αξιοσημείωτο είναι επίσης ότι έχουμε $Var(\hat{P}^d) > Var(\hat{P}^v)$ όταν $K < 1/2$. Κατά συνέπεια, ο εκτιμητής $\hat{P}^d$ της σχέσης (14), για την εύρεση του οποίου χρησιμοποιούνται και οι n^p παρατηρήσεις του κυρίως δείγματος, δίνει χειρότερες εκτιμήσεις ως αναφορά στη διακύμανση απ' ότι ο εκτιμητής $\hat{P}^v$ που κάνει χρήση μόνο του δείγματος επικύρωσης. Επιπλέον, για σταθερό δείγμα επικύρωσης n^v η διακύμανση του $\hat{P}^d$ αυξάνεται καθώς αυξάνεται το πλήθος των παρατηρήσεων n^p.

Κατά την περίπτωση όπου $K = 1/2$, η συσχέτιση μεταξύ της πραγματικής μεταβλητής Y και της λανθασμένα ταξινομημένης μεταβλητής Y' είναι της τάξεως του 0.71, τιμή που δεν είναι ιδιαιτέρως χαμηλή. Στον Πίνακα 5.2 παριστώνται κάποιοι συνδυασμοί των παραμέτρων (P, θ_Y, ϕ_Y) για τους οποίους οι τιμές του συντελεστή αξιοπιστίας K είναι περίπου 0.5. Όπως αναμενόταν, ο K είναι φθίνουσα συνάρτηση των θ_Y και ϕ_Y, με $K = 1$ όταν ισχύει $\theta_Y = \phi_Y = 0$, όταν δηλαδή δεν έχουμε λανθασμένη ταξινόμηση. Ωστόσο, για οποιεσδήποτε τιμές και των δύο πιθανοτήτων λανθασμένης ταξινόμησης θ_Y και ϕ_Y μεγαλύτερες του μηδενός, ισχύει $K < 1/2$ όταν το πληθυσμιακό ποσοστό P ή $1-P$ έχει τιμή κοντά στο μηδέν. Προς επαλήθευση των παραπάνω συμπερασμάτων, στο Σχήμα 5.1 απεικονίζονται οι ασυμπτωτικές σχετικές αποτελεσματικότητες $Var(\hat{P}^v)/Var(\hat{P}^c)$ και $Var(\hat{P}^v)/Var(\hat{P}^d)$ για ένα πεδίο τιμών του P όταν $\theta_Y = 0.1$, $\phi_Y = 0.15$ και $f = 1/6$.

ΣΧΗΜΑ 5.1

Ασυμπτωτικές σχετικές αποτελεσματικότητες των εκτιμητών $\hat{P}^d$ (συμπαγής γραμμή) και $\hat{P}^c$ (διακεκομμένη γραμμή) σε σχέση με τον $\hat{P}^v$ για $P \in [0.01, 0.99]$, $\theta_Y = 0.1$, $\phi_Y = 0.15$ και $f = 1/6$.

[ΠΗΓΗ : Kuha and Skinner, 1997]



ΠΙΝΑΚΑΣ 5.2

Τιμές του Συντελεστή Αξιοπιστίας Κ και των Σχετικών Αποτελεσματικοτήτων $ARE(\hat{P}^d)$ και $ARE(\hat{P}^c)$ για $f = 0$ και συνδυασμούς παραμέτρων

[ΠΗΓΗ : Kuha and Skinner, 1997]

P	θ_Y	φ_Y	K	$ARE(\hat{P}^d)$	$ARE(\hat{P}^c)$
0.50	0.33	0.00	0.50	1.02	2.02
0.50	0.15	0.15	0.49	0.96	1.96
0.30	0.20	0.10	0.48	0.93	1.93
0.10	0.07	0.07	0.51	1.02	2.02
0.70	0.20	0.05	0.48	0.94	1.94
0.90	0.05	0.15	0.51	1.04	2.04

Ένα επιπλέον πρόβλημα με τον εκτιμητή $\hat{P}^d$ είναι ότι το πεδίο τιμών του δεν είναι μεταξύ του 0 και του 1. Βάσει μελετών προσομοίωσης από τους Kuha and Skinner (1997) προκύπτει ότι για μικρότερες τιμές του *K* ένα ποσοστό των προσομοιωμένων δειγμάτων που δεν μπορεί να θεωρηθεί αμελητέο παράγει εκτιμητές που είναι είτε αρνητικοί είτε μεγαλύτεροι από τη μονάδα. Κυρίως λόγω αυτών των εκτός ορίων τιμών (εφόσον γνωρίζουμε ότι οι πιθανότητες παίρνουν τιμές στο διάστημα $[0,1]$), κατέληξαν στο συμπέρασμα ότι ο ασυμπτωτικός τύπος διακύμανσης της σχέσης (5.21) μπορεί να υποεκτιμήσει την πραγματική διακύμανση του $\hat{P}^d$. Εν αντιθέσει, ανακάλυψαν ότι ο αντίστοιχος τύπος για τον εκτιμητή $\hat{P}^c$ (Σχέση (5.17)) παίρνει τιμές αρκετά συναφείς με την πραγματική του διακύμανση ακόμα και για δείγματα μέτριου μεγέθους.

Στο παράδειγμα το οποίο και χρησιμοποιήθηκε στην ενότητα αυτή ο εκτιμητής μέγιστης πιθανοφάνειας $\hat{P}^c$ αποδείχτηκε να είναι σαφώς ανώτερος του εκτιμητή $\hat{P}^d$. Ωστόσο, ο $\hat{P}^c$ είναι γενικότερα κατάλληλος μόνο σε περιπτώσεις όπου το δείγμα επικύρωσης είναι εσωτερικό, αφού για εξωτερικά δείγματα επικύρωσης είναι πιο εύλογο οι πιθανότητες λανθασμένης ταξινόμησης α_{ji} να 'μεταφέρονται' (να είναι *transportable*) στο κυρίως δείγμα σε σχέση με τις πιθανότητες βαθμονόμησης c_{ij}. Όταν είναι διαθέσιμα μόνο δεδομένα προερχόμενα από εξωτερικό δείγμα επικύρωσης, τότε η φτωχή επίδοση του εκτιμητή $\hat{P}^d$ είναι ένα σοβαρής φύσεως πρόβλημα.

Ο Marshall (1990) συγκρίνει επίσης στο σύγγραμμα του τους εκτιμητές $\hat{P}^d$ και $\hat{P}^c$ χρησιμοποιώντας το ίδιο δυωνυμικό παράδειγμα που εφαρμόζεται στην ενότητα αυτή. Και αυτός εκφράζει την άποψη ότι ο εκτιμητής μέγιστης πιθανοφάνειας $\hat{P}^c$ είναι εμφανώς καλύτερος, αν και δεν επισημαίνει την ενδεχόμενη φτωχή απόδοση του $\hat{P}^d$ εν συγκρίσει με τον $\hat{P}^v$. Ο Selen (1986) δίνει με τη σειρά του κάποια αποτελέσματα προσομοίωσης συγκρίνοντας έναν πινακικό εκτιμητή και έναν εκτιμητή μέγιστης πιθανοφάνειας για έναν 2×2 πίνακα συνάφειας κατά την περίπτωση όπου μία μεταβλητή είναι λανθασμένα ταξινομημένη. Μολαταύτα, στην πλειονότητα των παραδειγμάτων που χρησιμοποιεί, οι πιθανότητες λανθασμένης ταξινόμησης θεωρούνται γνωστές. Οι Kuha et. al (1998) επισημαίνουν ότι ο $\hat{P}^d$ είναι μη συνεπής εκτιμητής όταν το χρησιμοποιούμενο δείγμα επικύρωσης είναι εξωτερικό και, όπως δείχθηκε και προηγουμένως, λιγότερο αποτελεσματικός του $\hat{P}^c$ όταν το δείγμα επικύρωσης είναι εσωτερικό.

Οι Morrissey and Spiegelman (1999) αναφερόμενοι στις πινακικές μεθόδους με χρήση πιθανοτήτων λανθασμένης ταξινόμησης και πιθανοτήτων βαθμονόμησης ως ‘πινακική μέθοδο’ (*matrix method*) και ‘αντίστροφη πινακική μέθοδο’ (*inverse matrix method*) αντίστοιχα και με το επιλεγόμενο δείγμα επικύρωσης να είναι εσωτερικό, προβαίνουν στα ακόλουθα συμπεράσματα για τη διμεταβλητή περίπτωση :

- ✓ όταν υπάρχει μη διαφορίσιμη λανθασμένη ταξινόμηση, δεν υπάρχει σαφής ένδειξη ότι οι εκτιμητές που προκύπτουν μέσω της πινακικής μεθόδου είναι πιο αποτελεσματικοί από αυτούς που προκύπτουν μέσω της αντίστροφης πινακικής μεθόδου· η αποτελεσματικότητα εξαρτάται από τις υποκείμενες παραμέτρους.
- ✓ όταν υπάρχει διαφορίσιμη λανθασμένη ταξινόμηση, οι εκτιμητές που προκύπτουν μέσω της αντίστροφης πινακικής μεθόδου είναι σχεδόν πάντα πιο αποτελεσματικοί από αυτούς που προκύπτουν μέσω της πινακικής μεθόδου· μάλιστα, παραμετροποιώντας την πιθανοφάνεια με χρήση των θετικών και αρνητικών προνωστικών αξιών, με άλλα λόγια με χρήση πιθανοτήτων βαθμονόμησης, οι εκτιμητές μέσω της αντίστροφης πινακικής μεθόδου είναι κατ’ ουσία εκτιμητές μέγιστης πιθανοφάνειας (Lyles (2002)).

Ενδιαφέρον θα είναι να εξεταστεί το κατά πόσον τα συμπεράσματα στα οποία προβήκαμε παραπάνω ισχύουν επίσης και για τις πολυμεταβλητές περιπτώσεις.

5.5 Σχετική βιβλιογραφία

Υπάρχει μία μεγάλη σε εύρος βιβλιογραφία όσον αφορά εφαρμογές μεθόδων πινάκων για διάφορα προβλήματα λανθασμένης ταξινόμησης. Οι Rogan and Gladen (1978) και Quade et. al (1980) περιγράφουν τη χρήση τέτοιων μεθόδων για την εκτίμηση ενός μόνο ποσοστού. Ο Tenenbein (1970) θεωρεί το ίδιο πρόβλημα, χρησιμοποιώντας όμως πιθανότητες βαθμονόμησης και όχι πιθανότητες λανθασμένης ταξινόμησης. Ο ίδιος (Tenenbein (1972)) επεκτείνει την ανωτέρω μεθοδολογία για προβλήματα όπου εκτιμώνται τα ποσοστά για μεταβλητές με περισσότερες των δύο κατηγοριών. Οι Tenenbein (1971), Hochberg and Tenenbein (1983) και Haitovsky and Rapp (1992) περιγράφουν διάφορους τρόπους βελτίωσης της ανάλυσης των δεδομένων με χρήση της μεθόδου διπλής δειγματοληψίας.

Οι πινακικοί μέθοδοι διόρθωσης για διδιάστατους πίνακες συνάφειας αποτελούν συχνά κομμάτι της έρευνας για τις επιδράσεις της λανθασμένης ταξινόμησης σε πίνακες τέτοιου είδους (Copeland et al. 1977, Shy et al. 1978, Barron 1977, Greenland 1982). Οι ερευνητές αυτοί θεωρούν την ύπαρξη λανθασμένης ταξινόμησης σε μία ή και στις δύο εξεταζόμενες μεταβλητές, καθώς επίσης διαφορίσιμη και μη διαφορίσιμη λανθασμένη ταξινόμηση. Ωστόσο θεωρούν ότι οι πιθανότητες λανθασμένης ταξινόμησης είναι γνωστές και σταθερές και δεν δίνουν εκτιμητές για τη διακύμανση των διορθωμένων πινάκων. Οι Greenland and Kleinbaum (1983) δίνουν μία σαφή εικόνα των πινακικών μεθόδων διόρθωσης (με σταθερό πίνακα λανθασμένης ταξινόμησης **A** και χωρίς τύπο για τη διακύμανση) για τους 2×2 πίνακες συνάφειας για περιπτώσεις διαφορίσιμης ή μη διαφορίσιμης λανθασμένης ταξινόμησης για μία ή και για τις δύο εξεταζόμενες μεταβλητές. Τονίζουν επίσης ότι αυτές οι μέθοδοι πρέπει να τροποποιηθούν ελάχιστα όταν το κυρίως δείγμα είναι στρωματοποιημένο ως προς μία λανθασμένα ταξινομημένη μεταβλητή, όταν δηλαδή το δείγμα επιλέγεται με τρόπο τέτοιο ώστε ο αριθμός των μονάδων για κάθε κατηγορία μίας λανθασμένης ταξινομημένης μεταβλητής να είναι σταθερός. Στην περίπτωση αυτή, η ευαισθησία και η ειδικότητα του δείγματος, συνεπώς και οι πιθανότητες λανθασμένης ταξινόμησης, εξαρτώνται από δειγματοληπτικά κλάσματα που λαμβάνονται (βλ. επίσης Chen 1989). Επίσης οι Morrissey and Spiegelman (1999) και Lyles (2002) ερευνούν τη σχέση των πινακικών εκτιμητών με τους εκτιμητές μέγιστης πιθανοφάνειας και συγκρίνουν την αποτελεσματικότητά τους κατά την ύπαρξη διαφορίσιμης ή όχι λανθασμένης ταξινόμησης. Οι Tzavidis and Lyn (2006) εξάγουν επίσης αρκετά ενδιαφέροντα συμπεράσματα για τη σχέση των δύο πινακικών μεθόδων διόρθωσης. Για διασταυρούμενες (*cross-sectional*) μελέτες η

χρήση πιθανοτήτων λανθασμένης ταξινόμησης ή πιθανοτήτων βαθμονόμησης οδηγεί σε πανομοιότυπα αποτελέσματα, εν αντιθέσει με τις διαχρονικές μελέτες όπου είναι προτιμότερη η χρήση πιθανοτήτων λανθασμένης ταξινόμησης για τον υπολογισμό εκτιμητών. Επισημαίνουν ακόμα ότι ένα πρόβλημα λανθασμένης ταξινόμησης μπορεί να αντιμετωπιστεί πρόβλημα ελλείπων δεδομένων (*missing data problem*) και προτείνουν παράλληλα έναν εκτιμητή ημι-πιθανοφάνειας (*quasi-likelihood estimator*), που μέσω προσομοιώσεων υποστηρίζουν πως είναι το ίδιο αποτελεσματικός με τον εκτιμητή μέγιστης πιθανοφάνειας.

ΚΕΦΑΛΑΙΟ 6

Διόρθωση της επίδρασης λανθασμένης ταξινόμησης μέσω μοντελοποίησης

6.1 Εισαγωγή

Στο προηγούμενο Κεφάλαιο έγινε περιγραφή των πινακικών μεθόδων διόρθωσης κατά την ύπαρξη λανθασμένης ταξινόμησης σε έναν υπό εξέταση πληθυσμό. Οι μέθοδοι αυτοί δίνουν εκτιμητές μέγιστης πιθανοφάνειας των ποσοστών ανά κελί κατά την χρήση ενός εσωτερικού δείγματος επικύρωσης. Οι εκτιμητές αυτοί είναι απλοί και εύκολοι στον υπολογισμό τους, αλλά ταυτόχρονα παρουσιάζουν και αρκετά μειονεκτήματα. Κατά πρώτον, δεν μπορούμε να προβούμε σε ελέγχους υποθέσεων ή να προσαρμόσουμε κατάλληλα μοντέλα για τους προκύπτοντες πραγματικούς πίνακες με τη χρήση των κλασικών τυποποιημένων μεθόδων, αφού οι μέθοδοι αυτές δε λαμβάνουν υπ' όψιν την επιπρόσθετη αβεβαιότητα που προκύπτει από τη διόρθωση της λανθασμένης ταξινόμησης. Δεύτερον, για πίνακές μεγάλων διαστάσεων, κυρίως πίνακες που αποτελούνται από αρκετές το πλήθος μεταβλητές, οι εκτιμητές του εκάστοτε πίνακα λανθασμένης ταξινόμησης είναι συχνά αρκετά ανακριβείς λόγω της διασποράς των εξεταζόμενων δεδομένων. Λόγω των μειονεκτημάτων αυτών, το ζητούμενο είναι η προσαρμογή κατάλληλων μοντέλων στο μηχανισμό λανθασμένης ταξινόμησης που να μας δίνουν εκτιμήσεις πιο ακριβείς και πλησιέστερες στις πραγματικές τιμές των εξεταζόμενων χαρακτηριστικών.

Στο Κεφάλαιο αυτό θα εξετάσουμε μεθόδους διόρθωσης για τη λανθασμένη ταξινόμηση ενός πληθυσμού που οδηγούν σε εκτιμητές μέγιστης πιθανοφάνειας για μία ευρεία κατηγορία μοντέλων και δομών των στοιχειών του πληθυσμού. Θα επικεντρωθούμε κυρίως σε μεθόδους που χρησιμοποιούν δείγματα επικύρωσης για την εκτίμηση των πιθανοτήτων λανθασμένης ταξινόμησης, που προέρχονται κυρίως από εσωτερικές μελέτες επικύρωσης. Λόγω της φύσης τους, οι μέθοδοι αυτές είναι υπολογιστικά πιο πολύπλοκοι και απαιτητικοί από τις πινακικές μεθόδους διόρθωσης. Για το λόγο αυτό, θα καταφύγουμε στη χρήση επαναληπτικών

μεθόδων, όπως για παράδειγμα ο αλγόριθμος μεγιστοποίησης-αναμενόμενης τιμής (*Expectation-Maximization* ή *EM algorithm*).

6.2 Μοντελοποίηση λανθασμένης ταξινόμησης μέσω λογαριθμογραμμικών μοντέλων

Στα προηγούμενα Κεφάλαια είδαμε πως ένα πρόβλημα λανθασμένης ταξινόμησης για ένα επιλεγμένο δείγμα του πληθυσμού μπορεί να αναπαρασταθεί μέσω ενός πίνακα συνάφειας με διαστάσεις που αντιστοιχούν στις διαστάσεις των πραγματικών και λανθασμένα ταξινομημένων μεταβλητών. Λόγω της κατανομής που ακολουθούν τα δεδομένα του δείγματος[4], ένα μεγάλο ποσοστό μοντέλων λανθασμένης ταξινόμησης μπορεί να αναπαρασταθεί μέσω λογαριθμογραμμικών μοντέλων (*log-linear models*) (βλ. Bishop et al. (1975)). Η εφαρμογή λογαριθμογραμμικών μοντέλων αντιμετωπίζει την εύρεση εκτιμητών και τους ελέγχους υποθέσεων σε ένα ενιαίο πλαίσιο και μπορεί να γίνει και σε πολυδιάστατους πίνακες συνάφειας. Στα λογαριθογραμμικά μοντέλα που χρησιμοποιούνται συμπεριλαμβάνονται και τα λογιστικά μοντέλα (*logistic models*), λόγω της κατηγορικής φύσης των δεδομένων, και για το λόγο αυτό θα γίνει εκτενέστερη αναφορά στα πρώτα. Βάσει αυτής της μεθόδου προσέγγισης, πρώτα γίνεται επιλογή ενός σχήματος για τη δομή της λανθασμένης ταξινόμησης και στη συνέχεια ερευνάται το σχήμα της δομής για τις ορθά ταξινομημένες μεταβλητές.

Για να γίνει περισσότερο κατανοητή η αναπαράσταση ενός πίνακα συνάφειας μέσω λογαριθμικών μοντέλων, θεωρούμε την εξής απλή περίπτωση ταξινόμησης δύο μεταβλητών χωρίς την ύπαρξη λανθασμένης ταξινόμησης. Έστω η μεταβλητή απόκρισης Y με r το πλήθος κατηγορίες και η επεξηγηματική μεταβλητή X με t κατηγορίες. Εάν συμβολίσουμε τις αναμενόμενες συχνότητες των κελιών του $r \times t$ πίνακα συνάφειας με m_{YX} ή m_{ik}, τότε το κορεσμένο (*saturated*) μοντέλο που περιέχει τις αλληλεπιδράσεις κάθε τάξης μεταξύ των X και Y είναι το εξής

$$\log m_{ik} = u + u_{Y(i)} + u_{X(j)} + u_{YX(ij)} \quad , \quad i = 1,...,r \quad , \quad j = 1,...,t$$

[4] Όπως είναι γνωστό, οι πιο συνηθισμένες κατανομές για κατηγορικά δεδομένα είναι η δυωνυμική, η πολυωνυμική, και η κατανομή Poisson. Η δυωνυμική κατανομή και η κατανομή Poisson υπό ορισμένες προϋποθέσεις υπάγονται στην πολυωνυμική κατανομή.

με τον περιορισμό τα περιθώρια αθροίσματα των *u*-όρων να αθροίζουν στο μηδέν για κάθε υποδείκτη, δηλαδή $\sum_i u_{Y(i)} = \sum_j u_{X(j)} = \sum_i u_{YX(ij)} = \sum_j u_{YX(ij)} = 0$. Ας τονίσουμε σε αυτό το σημείο ότι για την εφαρμοζόμενη μεθοδολογία γίνεται χρήση μόνο των ιεραρχικών (*hierarchical*) μοντέλων, που προσδιορίζονται από τους μεγαλύτερης τάξης *u*-όρους και περιέχουν όλους τους *u*-όρους μικρότερης τάξης. Εφόσον ο μεγαλύτερος τη τάξη *u*-όρος είναι ο YX, το μοντέλο συμβολίζεται με $H(YX)$.

Για τα λογαριθμογραμμικά μοντέλα, η διαδικασία εύρεσης εκτιμητών γίνεται μέσω χρήσης μεθόδων μέγιστης πιθανοφάνειας (*maximum likelihood* ή *ML methods*). Οι αναμενόμενες συχνότητες των κελιών μπορούν να υπολογιστούν μέσω ενός επαναληπτικού αλγορίθμου προσαρμογής ποσοστών (*iterative proportional fitting algorithm*) στα κελιά του πίνακα συνάφειας. Ο έλεγχος καλής προσαρμογής γίνεται μέσω του X^2-ελέγχου του Pearson (*Pearson's* X^2*-test*) ή του ελέγχου λόγου πιθανοφανειών (*likelihood ratio* ή *LR test*). Τα τυπικά σφάλματα των εκτιμητών μπορούν να υπολογιστούν μέσω της μεθόδου Newton-Raphson.

6.2.1. Λογαριθμογραμμικά μοντέλα με γνωστές πιθανότητες λανθασμένης ταξινόμησης

Οι περιπτώσεις κατά τις οποίες οι πιθανότητες λανθασμένης ταξινόμησης των ατόμων ενός πληθυσμού είναι γνωστές δεν απαντώνται συνήθως σε διεξαγόμενες έρευνες υπό πραγματικές συνθήκες. Στις περιπτώσεις που είναι όμως γνωστές, οι εκτιμήσεις τους μπορούν να γίνουν με απευθείας εφαρμογή του EM αλγόριθμου. Η μεθοδολογία που ακολουθεί περιγράφεται αναλυτικά από τον Chen (1978, 1979a, 1989) και επεκτείνεται από τους Kuha and Skinner (1997). Μία από τις κυριότερες εφαρμογές της είναι για σχεδιασμούς τυχαιοποιημένης απόκρισης. Οι Whitemore and Grosser (1986) εξετάζουν την περίπτωση όπου είναι γνωστές οι πιθανότητες βαθμονόμησης.

Χάριν απλότητας ως προς την παρουσίαση της προτεινόμενης μεθοδολογίας που θα ακολουθήσει, θα εξεταστεί η περίπτωση ύπαρξης δύο ορθά ταξινομημένων κατηγορικών μεταβλητών Y και E που ταξινομούνται εσφαλμένα στον πληθυσμό από τις Y' και E' αντίστοιχα. Παρόλα αυτά, η μεθοδολογία μπορεί να επεκταθεί και στη γενικότερη περίπτωση. Υποθέτουμε επίσης ότι τα δεδομένα προέρχονται από διπλή δειγματοληψία και η κατανομή τους, χωρίς βλάβη της γενικότητας, είναι πολυωνυμική. Οι Rao and Thomas (1991)

ερευνούν την επέκταση τους σε ορισμένες μορφές σύνθετων σχεδιασμών (*complex designs*). Σκοπός μας εδώ είναι να μοντελοποιήσουμε τη σχέση συσχέτισης ανάμεσα στις ορθά ταξινομημένες μεταβλητές.

Έστω $\mathbf{P}_{YE}$ τα πληθυσμιακά ποσοστά των ατόμων που ανήκουν στις κατηγορίες που ορίζονται από τις πιθανές τιμές των ορθά ταξινομημένων (πραγματικών) μεταβλητών και $\alpha_{Y'E'|YE}$ οι πιθανότητες λανθασμένης ταξινόμησης των ατόμων. Ορίζουμε ως $\mathbf{n}_{Y'E'}$ το διάνυσμα των παρατηρούμενων αριθμών δειγματικών μονάδων που ταξινομούνται στους δυνατούς συνδυασμούς των μεταβλητών Y' και E'. Ας υποτεθεί ότι είναι επιθυμητή η προσαρμογή ενός μοντέλου για τις μεταβλητές Y και E χρησιμοποιώντας το διάνυσμα $\mathbf{n}_{Y'E'}$ και τις γνωστές πιθανότητες λανθασμένης ταξινόμησης $\alpha_{Y'E'|YE}$, για παράδειγμα ενός μοντέλου ανεξαρτησίας των μεταβλητών (δηλαδή του μοντέλου $H(Y,E)$). Όπως είναι γνωστό, ο ΕΜ αλγόριθμος αποτελείται από τα εξής δύο βήματα, το βήμα αναμενόμενης τιμής ή Έ-βήμα και το βήμα μεγιστοποίησης ή Μ-βήμα, τα οποία και επαναλαμβάνονται μέχρις ότου επιτευχθεί σύγκλιση. Στο Ε-βήμα, μία αναμενόμενη τιμή του πλήρη πίνακα συνάφειας των μεταβλητών Y', E', Y και E υπολογίζεται μέσω της σχέσης

$$\mathbf{n}^{(i)}_{Y'E'YE} = \frac{\alpha_{Y'E'|YE}\hat{\mathbf{P}}^{(i)}_{YE}}{\sum_{Y,E}\alpha_{Y'E'|YE}\hat{\mathbf{P}}^{(i)}_{YE}} \cdot \mathbf{n}_{Y'E'}, \qquad (6.1)$$

όπου $\hat{\mathbf{P}}^{(i)}_{YE}$ είναι ο εκτιμητής του διανύσματος των πραγματικών πληθυσμιακών ποσοστών στο (i) βήμα της επαναληπτικής διαδικασίας (με $i = 0,1,2,....$). Η τιμή $\hat{\mathbf{P}}^{(0)}_{YE} = \mathbf{P}^{(0)}_{YE}$ είναι η τιμή του διανύσματος των πραγματικών πληθυσμιακών ποσοστών του αρχικού πίνακα συνάφειας που περιέχει τις ορθά ταξινομημένες μεταβλητές. Ας σημειωθεί πως η διαίρεση και ο πολλαπλασιασμός στο δεξιό μέλος της σχέσης (6.1) συνεπάγεται τη διαίρεση και τον πολλαπλασιασμό των στοιχείων των σχετικών πινάκων ή διανυσμάτων. Στη συνέχεια, υπολογίζεται ένας πίνακας με τις αναμενόμενες τιμές των πραγματικών μεταβλητών μέσω της σύμπτυξης του διανύσματος $\mathbf{n}^{(i)}_{Y'E'YE}$ ως προς τις εσφαλμένα ταξινομημένες μεταβλητές

$$\mathbf{n}^{(i)}_{YE} = \sum_{Y',E'} \mathbf{n}^{(i)}_{Y'E'YE}$$

Στο Μ-βήμα, υπολογίζεται ένας νέος εκτιμητής $\hat{\mathbf{P}}^{(i+1)}_{YE}$ του $\mathbf{P}_{YE}$ μέσω της προσαρμογής του ερευνώμενου μοντέλου για τις τιμές του $\mathbf{n}^{(i)}_{YE}$ και η διαδικασία επαναλαμβάνεται μέχρι την

επίτευξη σύγκλισης. Συνοπτικά, τα ακολουθούμενα βήματα του ΕΜ αλγόριθμου είναι τα ακόλουθα

Αρχική Συνθήκη : $\hat{\mathbf{P}}_{YE}^{0} = \mathbf{P}_{YE}^{0}$

Ε-βήμα : $$\mathbf{n}_{Y'E'YE}^{(i)} = \frac{\alpha_{Y'E'|YE}\hat{\mathbf{P}}_{YE}^{(i)}}{\sum_{Y,E}\alpha_{Y'E'|YE}\hat{\mathbf{P}}_{YE}^{(i)}} \cdot \mathbf{n}_{Y'E'}$$

$$\mathbf{n}_{YE}^{(i)} = \sum_{Y',E'} \mathbf{n}_{Y'E'YE}^{(i)}$$

Μ-βήμα : Προσαρμογή του ερευνώμενου μοντέλου για τις τιμές του $\mathbf{n}_{YE}^{(i)}$ και υπολογισμός του εκτιμητή των πραγματικών ποσοστών $\hat{\mathbf{P}}_{YE}^{(i+1)}$

Τέλος, ο έλεγχος καλής προσαρμογής του ζητούμενου μοντέλου γίνεται χρησιμοποιώντας κριτήρια ελέγχου όπως ο X^2-έλεγχος και ο έλεγχος λόγου πιθανοφανειών.

6.2.2. Λογαριθμογραμμικά μοντέλα με άγνωστες πιθανότητες λανθασμένης ταξινόμησης

Κατά την περίπτωση που οι πιθανότητες λανθασμένης ταξινόμησης δεν είναι γνωστές αλλά εκτιμώνται μέσω ενώ εσωτερικού δείγματος επικύρωσης, η διαδικασία διόρθωσης της λανθασμένης ταξινόμησης χρειάζεται να επεκταθεί. Για το λόγο αυτό ορίζουμε μία νέα μεταβλητή, έστω *L*, η οποία ταυτοποιεί το δείγμα από το οποίο προέρχεται μία μονάδα του πληθυσμού, είτε είναι το βασικό δείγμα είτε το δείγμα επικύρωσης. Η μεταβλητή *L* είναι δίτιμη όταν υπάρχει ένα κυρίως δείγμα και ένα δείγμα επικύρωσης. Αυτό δεν καθιστά απαγορευτική τη δυνατότητα επιλογής άλλων δειγματοληπτικών σχεδιασμών για τον εξεταζόμενο πληθυσμό. Για παράδειγμα, ο Chen (1984) θεωρεί ένα σχέδιο τριπλής δειγματοληψίας, όπου το τρίτο δείγμα που χρησιμοποιείται περιέχει μόνο ορθά ταξινομημένα (πραγματικά) στοιχεία, το οποίο και θα παρουσιαστεί αναλυτικά στην επόμενη Παράγραφο.

Για να καταδειχθεί η δομή της συσχέτισης ανάμεσα στις μεταβλητές, παρουσιάζουμε στη συνέχεια το λογαριθμογραμμικό μοντέλο που δόθηκε από τους Espeland and Odoroff (1985). Όλα τα μοντέλα που παρουσιάζονται είναι λογαριθμογραμμικά και για το λόγο αυτό μπορούν να προσδιοριστούν μέσω των υψηλότερης τάξης αλληλεπιδράσεων του κάθε μοντέλου.

Έστω οι ορθά ταξινομημένες κατηγορικές μεταβλητές Y και E που ταξινομούνται εσφαλμένα στον πληθυσμό από τις Y' και E' αντίστοιχα. Στη συνέχεια, θεωρούμε τα ακόλουθα μοντέλα, που ενυπάρχουν στην διπλή δειγματοληψία.

1. **Μοντέλο δειγματοληψίας** (*sampling model*)

   Το μοντέλο αυτό περιγράφει τις υποθέσεις δειγματοληψίας μεταξύ των μεταβλητών Y, E, Y', E' και L. Καταδεικνύει με άλλα λόγια τις διαφορές ανάμεσα στο κυρίως δείγμα και το δείγμα επικύρωσης μέσω των αλληλεπιδράσεων ανάμεσα στη μεταβλητή L και ένα μέρος των πραγματικών μεταβλητών (Y και E). Εάν το δείγμα επικύρωσης είναι εσωτερικό, και τα δύο δείγματα είναι τυχαία δείγματα προερχόμενα από τον ίδιο πληθυσμό και ο 'δειγματικός δείκτης' L είναι ανεξάρτητος των άλλων μεταβλητών. Το ιεραρχικό μοντέλο που αντιστοιχεί είναι το $H(YEY'E', L)$. Συχνά, το πρώτο βήμα της ανάλυσης είναι να ελεγχθεί εάν τα δεδομένα είναι συνεπή (*consistent*) ως προς αυτό το μοντέλο.

2. **Μοντέλο λανθασμένης ταξινόμησης ή λανθασμένης κατηγοριοποίησης** (*misclassification or miscategorization model*)

   Το μοντέλο αυτό περιγράφει τη σχέση ανάμεσα στις ορθά ταξινομημένες και τις εσφαλμένα ταξινομημένες μεταβλητές. Τα μοντέλα λανθασμένης ταξινόμησης εφαρμόζονται στο πλήρες μοντέλο δειγματοληψίας $H(YEY'E', L)$, αλλά λόγω των περιορισμών που εισάγει η ανεξαρτησία της μεταβλητής L, μόνο συγκεκριμένα μοντέλα είναι κατάλληλα. Μία ελάχιστη απαίτηση για το μοντέλο αυτό είναι να περιέχει έναν όρο αλληλεπίδρασης ανάμεσα σε κάθε ορθά ταξινομημένη μεταβλητή και τις αντίστοιχες σε αυτές λανθασμένα ταξινομημένες μεταβλητές, γιατί σε διαφορετική περίπτωση η λανθασμένη ταξινόμηση θα είναι εντελώς τυχαία. Υπό αυτούς τους περιορισμούς, διάφορες μορφές ιεραρχικών μοντέλων μπορούν να θεωρηθούν. Για παράδειγμα, στο μοντέλο $H(YE, YY', EE', L)$ η λανθασμένη ταξινόμηση των μεταβλητών Y και E είναι μη διαφορίσιμη και ανεξάρτητη. Στο μοντέλο $H(YEY', EE', L)$ είναι ανεξάρτητη και μη διαφορίσιμη ως προς τη μεταβλητή E, αλλά διαφορίσιμη για την Y ως προς την E. Τέλος, για το μοντέλο $H(YEY'E', L)$, η λανθασμένη ταξινόμηση δεν είναι ανεξάρτητη ούτε διαφορίσιμη για καμία από τις πραγματικές μεταβλητές.

3. Πειραματικό μοντέλο (*experimental model*)

Το μοντέλο αυτό περιγράφει τη σχέση ανάμεσα στις ορθά ταξινομημένες μεταβλητές. Δεν περιέχει σχέσεις που περιλαμβάνουν τις λανθασμένα ταξινομημένες μεταβλητές. Σε αυτό το μοντέλο, η μεταβλητή L μπορεί να απαλειφθεί εάν είναι ανεξάρτητη των ορθά ταξινομημένων μεταβλητών. Στη διμεταβλητή περίπτωση, τα δύο πιθανά πειραματικά μοντέλα είναι το κορεσμένο μοντέλο $H(YE)$ και το μοντέλο ανεξαρτησίας $H(Y,E)$. Οι εκτιμητές που περιγράφουν τη σχέση ανάμεσα στις μεταβλητές Y και E χρησιμοποιούν πληροφορίες από τον πλήρη πίνακα $YEY'E'L$.

Ας τονιστεί στο σημείο αυτό πως παρόλο που τα τρία υπομοντέλα είναι λογαριθμογραμμικά, το από κοινού παραγόμενο μοντέλο δεν χρειάζεται να είναι.

Από τα ανωτέρω μοντέλα, τα μοντέλα δειγματοληψίας και λανθασμένης ταξινόμησης πρέπει να προσαρμοστούν σε πίνακες συνάφειας που εμπεριέχουν μεικτά (*mixed-up*) κελιά. Έστω $\mathbf{x}$ το διάνυσμα συχνοτήτων των κελιών του πλήρους πολυδιάστατου πίνακα συνάφειας των στοιχείων του πληθυσμού και $\mathbf{m}$ το διάνυσμα των μέσων για το διάνυσμα $\mathbf{x}$. Ένα λογαριθμογραμμικό μοντέλο για το $\mathbf{m}$ είναι το

$$\log \mathbf{m} = \mathbf{Mb},$$

όπου $\mathbf{M}$ είναι ένας πίνακας σχεδιασμού και $\mathbf{b}$ είναι ένα διάνυσμα παραμέτρων. Θα χρησιμοποιήσουμε την εναλλακτική εκδοχή του ΕΜ αλγόριθμου που προτάθηκε από τους Espeland and Odoroff (1984a) για την προσαρμογή του μοντέλου.

Έστω $\mathbf{b}^{(0)}$ το διάνυσμα των αρχικών εκτιμήσεων των λογαριθμογραμμικών παραμέτρων και $\mathbf{m}^{(0)}$ το αρχικό διάνυσμα που εκφράζει τις αναμενόμενες τιμές των κελιών, με $\log \mathbf{m}^{(0)} = \mathbf{Mb}^{(0)}$. Για τον προτεινόμενο ΕΜ αλγόριθμο, στο Ε-βήμα παράγεται έναν πλήρως κατηγοριοποιημένο διάνυσμα $\mathbf{x}^{(1)}$, όπου

$$\mathbf{M}'\mathbf{x}^{(1)} = \mathbf{M}'E(\mathbf{x} \mid \mathbf{b}^{(0)}, \mathbf{S}'\mathbf{x} = \mathbf{y}),$$

και με $\mathbf{y} = \mathbf{S}'\mathbf{x}$ δηλώνεται η σχέση που συνδέει τις τιμές του διανύσματος συχνοτήτων $\mathbf{y}$ που αντιστοιχεί στον ελλιπή πίνακα με τις τιμές διανύσματος συχνοτήτων $\mathbf{x}$ του πλήρους πίνακα μέσω ενός πίνακα σχεδιασμού $\mathbf{S}$. Στο Μ-βήμα, για την εύρεση των εκτιμητών μέγιστης πιθανοφάνειας $\mathbf{b}^{(1)}$ και $\mathbf{m}^{(1)} = \exp(\mathbf{Mb}^{(1)})$ μέσω του διανύσματος $\mathbf{x}^{(1)}$, χρησιμοποιούμε

αλγόριθμους μέγιστης πιθανοφάνειας, όπως ο αλγόριθμος Newton-Raphson ή ο επαναληπτικός αλγόριθμος αναλογικής προσαρμογής (*iterative proportional fitting algorithm* ή *IPFA*). Κάθε Ε-βήμα χρησιμοποιεί τις μεταβλητές $\mathbf{b}^{(i)}$ και $\mathbf{m}^{(i)}$ από το προηγούμενο Μ-βήμα για να παραχθεί το διάνυσμα $\mathbf{x}^{(i+1)}$. Κάθε Μ-βήμα χρησιμοποιεί το διάνυσμα $\mathbf{x}^{(i+1)}$ για να παραχθούν τα διανύσματα $\mathbf{b}^{(i+1)}$ και $\mathbf{m}^{(i+1)}$. Συνοπτικά, με $i = 1, 2, \ldots$ να είναι ο δείκτης των επαναληπτικών βημάτων του αλγόριθμου, τα ακολουθούμενα βήματα του ΕΜ αλγόριθμου είναι τα ακόλουθα

Αρχική Συνθήκη : $i = 0$

$\mathbf{b}^{(0)}$, $\mathbf{m}^{(0)} = \exp(\mathbf{M}\mathbf{b}^{(0)})$

Ε-βήμα : $\mathbf{M}'\mathbf{x}^{(i+1)} = \mathbf{M}'E(\mathbf{x} \mid \mathbf{b}^{(i)}, \mathbf{S}'\mathbf{x} = \mathbf{y})$

Μ-βήμα : $\mathbf{b}^{(i+1)}$, $\mathbf{m}^{(i+1)} = \exp(\mathbf{M}\mathbf{b}^{(i+1)})$

Οι Espeland and Odoroff (1984b) έδειξαν ότι το Ε-βήμα μπορεί να εκτελεστεί μέσω μίας απλής αναλογικής προσαρμογής του Μ-βήματος που προηγείται μέσω της σχέσης

$$\mathbf{x}^{(i+1)} = \mathbf{S}\left(\frac{\mathbf{y}}{\mathbf{S}'\mathbf{m}^{(i)}}\right)\mathbf{m}^{(i)},$$

Κι εδώ ισχύει ότι η διαίρεση και ο πολλαπλασιασμός στο δεξιό μέλος της σχέσης συνεπάγεται τη διαίρεση και τον πολλαπλασιασμό των στοιχείων των σχετικών πινάκων ή διανυσμάτων.

Για το μοντέλο που περιγράφηκε είναι επίσης δυνατή και η εύρεση εκτιμώμενων διακυμάνσεων για τις εκτιμήσεις των παραμέτρων του (Espeland and Odoroff (1985)).

6.2.3. Τριπλή δειγματοληψία – μία ενδιαφέρουσα προσέγγιση διόρθωσης της λανθασμένης ταξινόμησης

Στις προηγούμενες Παραγράφους εξετάσαμε την εφαρμογή λογαριθμογραμμικών μοντέλων για την αντιμετώπιση της λανθασμένης ταξινόμησης κατά την περίπτωση της διπλής δειγματοληψίας, όπου το κύριο δείγμα ταξινομείται στον πληθυσμό μέσω των

εσφαλμένα ταξινομημένων μεταβλητών και το δείγμα επικύρωσης ταυτόχρονα από τις εσφαλμένες και ορθά ταξινομημένες (πραγματικές) μεταβλητές. Σε πολλές περιπτώσεις μπορεί να υπάρχει διαθέσιμο ένα επιπλέον τρίτο δείγμα, με τις μονάδες του να είναι ταξινομημένες μόνο μέσω των ορθά ταξινομημένων μεταβλητών. Σε άλλες περιπτώσεις κατά την εφαρμογή διπλής δειγματοληψίας μπορεί να προκύψουν δεδομένα που είναι διαθέσιμες μόνο οι πραγματικές τους τιμές. Και στις δύο περιπτώσεις καταλήγουμε στην εφαρμογή ενός σχεδίου διπλής δειγματοληψίας του πληθυσμού με διαθέσιμο ένα επιπλέον τρίτο και ανεξάρτητο δείγμα πραγματικών μεταβλητών, το οποίο και ονομάζουμε τριπλό σχέδιο δειγματοληψίας (*triple-sampling scheme* ή *TSS*).

Η δομή του TSS που θα θεωρηθεί στη συνέχεια είναι μία ειδική περίπτωση ελλιπούς πληροφόρησης (*missing information*). Το μοντέλο που χρησιμοποιείται αποτελείται από τη διασταύρωση δύο λογαριθμογραμμικών υπομοντέλων. Το πρώτο υπομοντέλο είναι μοντέλο συχνοτήτων των κελιών στον διασταυρούμενο πίνακα και ουδεμία σχέση έχει με τη μοντελοποίηση της δομής των σφαλμάτων λανθασμένης ταξινόμησης. Το δεύτερο υπομοντέλο είναι ένα μοντέλο για το 'πραγματικό περιθώριο' μόνο, δηλαδή για τον πίνακα που λαμβάνεται μέσω της σύμπτυξης ως προς τις μεταβλητές της εσφαλμένης ταξινόμησης. Ας τονιστεί σε αυτό το σημείο ότι το μοντέλο που προκύπτει δεν είναι απαραίτητα λογαριθμογραμμικό. Η μεθοδολογία που θα ακολουθήσει παρουσιάζεται διεξοδικά από τους Chen et al. (1984).

Έστω διάθεση μας $e+f$ κατηγορικές μεταβλητές, με e μεταβλητές να είναι επιρρεπείς σε σφάλματα και f μεταβλητές να είναι ελεύθερες σφαλμάτων. Έστω επίσης τα σύνολα των επιρρεπών σε σφάλματα μεταβλητών $Y_1,...,Y_e$ και $X_1,...,X_e$ που περιέχουν τις ορθά ταξινομημένες και λανθασμένα ταξινομημένες μεταβλητές αντίστοιχα και $Z_1,...,Z_f$ το σύνολο των μεταβλητών που είναι ελεύθερες σφαλμάτων και (συνεπώς) ταξινομούν ορθά τον εξεταζόμενο πληθυσμό. Θεωρώντας τα διανύσματα $\mathbf{Y}=(Y_1,...,Y_e)^c$, $\mathbf{X}=(X_1,...,X_e)^c$, $\mathbf{Z}=(Z_1,...,Z_f)^c$ με I, J, K το αντίστοιχο πλήθος των μεταβλητών τους, ισχύει $M=IJK$. Σημειώνεται ότι το πλήθος των μεταβλητών του **Y** δεν είναι απαραίτητο να είναι ίδιο με το πλήθος μεταβλητών του **X**.

Ας θεωρήσουμε στη συνέχεια τρία ανεξάρτητα πολυωνυμικά δείγματα n, n_1 και n_2 με συχνότητες κελιών $n(\mathbf{Y},\mathbf{X},\mathbf{Z})$, $n_1(\mathbf{Y},\mathbf{Z})$ και $n_2(\mathbf{X},\mathbf{Z})$ πραγματικά πληθυσμιακά ποσοστά

$p(\mathbf{Y},\mathbf{X},\mathbf{Z})$, $p(\mathbf{X},\mathbf{Z})$, $p(\mathbf{Y},\mathbf{Z})$ αντίστοιχα. Κατά την περίπτωση που και τα τρία δείγματα προέρχονται από την ίδια κατανομή, ισχύουν $p(\mathbf{Y},\mathbf{Z}) = \sum_{\mathbf{X}} p(\mathbf{Y},\mathbf{X},\mathbf{Z})$ και $p(\mathbf{X},\mathbf{Z}) = \sum_{\mathbf{Y}} p(\mathbf{Y},\mathbf{X},\mathbf{Z})$. Σκοπός μας είναι η προσαρμογή ιεραρχικών λογαριθμογραμμικών μοντέλων στα πραγματικά ποσοστά $p(\mathbf{Y},\mathbf{Z})$ για διάφορα μοντέλα που αφορούν τα σφάλματα ταξινόμησης $p(\mathbf{X}|\mathbf{Y},\mathbf{Z})$, δηλαδή τις δεσμευμένες πιθανότητες λανθασμένης ταξινόμησης μίας τυχαίας μονάδας του πληθυσμού όταν η πραγματική της ταξινόμηση στον πληθυσμό είναι $(\mathbf{Y},\mathbf{Z})$. Λόγω του ότι τέτοιες δομές παρουσιάστηκαν στις προηγούμενες Παραγράφους αυτού του Κεφαλαίου, θα αναφερθούμε στη συνέχεια στο γενικότερο πλαίσιο που στηρίζεται η παρούσα μέθοδος.

Προσεγγίζοντας το πρόβλημα αντιμετώπισης της λανθασμένης ταξινόμησης ως πρόβλημα ελλιπούς πληροφόρησης, γίνεται χρήση του ΕΜ αλγόριθμου για την εύρεση εκτιμητών μέγιστης πιθανοφάνειας. Στο Ε-βήμα, υπολογίζεται η αναμενόμενη λογαριθμική πιθανοφάνεια (*log-likelihood*) του συνόλου των δεδομένων, εξαρτώμενο από τα διαθέσιμα δεδομένα υπό τις τιμές των παραμέτρων στο συγκεκριμένο επαναληπτικό βήμα. Έστω $N = n + n_1 + n_2$ και $\mathbf{m}$ το διάνυσμα που περιέχει τις τιμές όλων των $m(\mathbf{Y},\mathbf{X},\mathbf{Z}) = Np(\mathbf{Y},\mathbf{X},\mathbf{Z})$. Συμβολίζοντας με $m^{(v)}(\mathbf{Y},\mathbf{X},\mathbf{Z})$ τις τιμές των παραμέτρων κατά το (ν) επαναληπτικό βήμα, ο πυρήνας της αναμενόμενης λογαριθμικής πιθανοφάνειας κατά το βήμα εκείνο δίνεται από τον τύπο

$$Q(\mathbf{m}|\mathbf{m}^{(v)}) = \sum_{\mathbf{Y},\mathbf{X},\mathbf{Z}} [N^{(v)}(\mathbf{Y},\mathbf{X},\mathbf{Z}) \cdot \log m(\mathbf{Y},\mathbf{X},\mathbf{Z})] \qquad (6.2)$$

με

$$N^{(v)}(\mathbf{Y},\mathbf{X},\mathbf{Z}) = n(\mathbf{Y},\mathbf{X},\mathbf{Z}) + \frac{m^{(v)}(\mathbf{Y},\mathbf{X},\mathbf{Z})}{\sum_{\mathbf{X}} m^{(v)}(\mathbf{Y},\mathbf{X},\mathbf{Z})} n(\mathbf{Y},\mathbf{Z}) + \frac{m^{(v)}(\mathbf{Y},\mathbf{X},\mathbf{Z})}{\sum_{\mathbf{Y}} m^{(v)}(\mathbf{Y},\mathbf{X},\mathbf{Z})} n(\mathbf{X},\mathbf{Z}).$$

Στο Μ-βήμα, μεγιστοποιείται ο πυρήνας $Q(\mathbf{m}|\mathbf{m}^{(v)})$ σε σχέση με το $\mathbf{m}$ υπό τους δεδομένους περιορισμούς. Στη συνέχεια, η εξαγόμενη τιμή του $\mathbf{m}$ χρησιμοποιείται ως η $(v+1)$ τιμή ($\mathbf{m}^{(v+1)}$) του επόμενου επαναληπτικού βήματος που βασίζεται στη σχέση (6.2).

Επειδή το μοντέλο που χρησιμοποιείται αποτελείται από τη διασταύρωση δύο λογαριθμογραμμικών υπομοντέλων και δεν είναι απαραίτητα λογαριθμογραμμικό, οι Chen et al. (1984) προτείνουν την ακόλουθη διαδικασία που, όπως υποστηρίζουν, είναι και αυτή ΕΜ αλγόριθμος που συγκλίνει στις εκτιμήσεις μέγιστης πιθανοφάνειας. Έστω $\mathbf{m}_1$ το διάνυσμα

μήκους M του οποίου τα στοιχεία είναι τα $m(\mathbf{Y},\mathbf{Z})$ με κάθε τιμή να επαναλαμβάνεται J φορές και $\mathbf{m}_2$ το αντίστοιχο διάνυσμα όλων των πιθανοτήτων λανθασμένης ταξινόμησης $p(\mathbf{X}|\mathbf{Y},\mathbf{Z})$. Ισχύει ότι $\mathbf{m}=\mathbf{m}_1\odot\mathbf{m}_2$ (με $\odot$ να συμβολίζει τον στοιχείο προς στοιχείο πολλαπλασιασμό των δύο διανυσμάτων), το οποίο και συνεπάγεται

$$Q(\mathbf{m}|\mathbf{m}^{(v)})=Q(\mathbf{m}_1|\mathbf{m}^{(v)})+Q(\mathbf{m}_2|\mathbf{m}^{(v)}).$$

Για την επίτευξη της μέγιστης τιμής για το $\mathbf{m}$ τον $Q(\mathbf{m}|\mathbf{m}^{(v)})$ ακολουθούνται τα εξής βήματα:

1. Παράγουμε το $\tilde{\mathbf{m}}^{(v+1)}$, το οποίο μεγιστοποιεί τον $Q(\mathbf{m}|\mathbf{m}^{(v)})$ σε σχέση με το $\mathbf{m}$, υπό το λογαριθμογραμμικό μοντέλο για τον διασταυρούμενα ταξινομημένο πίνακα. (Αυτό είναι το μοντέλο για τις πιθανότητες λανθασμένης ταξινόμησης). Σε αυτό το βήμα χρησιμοποιείται ένας IPFA, υπό τον περιορισμό το $\mathbf{m}$ να έχει τις ίδιες τιμές του $\mathbf{m}_1^{(v)}$ όπως αυτές που το $\mathbf{m}^{(v)}$ έχει.

2. Παράγουμε το $\tilde{\mathbf{m}}_1^{(v+1)}$ μέσω της μεγιστοποίησης του $Q(\mathbf{m}_1|\mathbf{m}_1^{(v)})$ σε σχέση με το $\mathbf{m}_1$. Αυτό επιτυγχάνεται μέσω της εφαρμογής ενός συνήθους IPFA στον $\{\mathbf{Y},\mathbf{Z}\}$ πίνακα με πλήθος στοιχείων $\sum_{\mathbf{X}} N^{(v)}(\mathbf{Y},\mathbf{X},\mathbf{Z})$.

3. Παράγουμε το $\tilde{\mathbf{m}}_2^{(v+1)}$ μέσω της σχέσης $\tilde{\mathbf{m}}^{(v+1)}=\tilde{\mathbf{m}}_1^{(v+1)}\odot\tilde{\mathbf{m}}_2^{(v+1)}$.

Σύμφωνα με τον Goodman (1973), το $\mathbf{m}^{(v+1)}$ μεγιστοποιεί τον πυρήνα της αναμενόμενης λογαριθμικής πιθανοφάνειας $Q(\mathbf{m}|\mathbf{m}^{(v)})$ και ως εκ τούτου θα χρησιμοποιηθεί για το $(v+1)$ επαναληπτικό βήμα του EM αλγόριθμου μέσω της σχέσης (6.2). Η σύγκλιση των τιμών του $\mathbf{m}^{(v)}$ οδηγεί σε εκτιμήσεις μέγιστης πιθανοφάνειας

6.3 Σχετική βιβλιογραφία

Πολλοί από τους ερευνητές που ασχολήθηκαν με τις μεθόδους διόρθωσης λανθασμένης ταξινόμησης μέσω μοντελοποίησης επέλεξαν το ίδιο σύνολο δεδομένων για να

παρουσιάσουν τις μεθόδους τους. Ο Hochberg (1977) εισήγαγε ένα παράδειγμα που προερχόταν από μία έρευνα για την ασφάλεια των αυτοκινητοδρόμων και εμπεριείχε τέσσερις μεταβλητές, δύο εκ των οποίων είναι λανθασμένα ταξινομημένες. Τα δεδομένα αυτά αποτέλεσαν αντικείμενο ανάλυσης και από τους Chen et al. (1984, 1989), Espeland and Odoroff (1985) και Ekholm and Palmgren (1987) (βλ. επίσης Ekholm (1991)). Οι παραπάνω συγγραφείς επέλεξαν διαφορετικά μοντέλα λανθασμένης ταξινόμησης, κάθε ένα εκ των οποίων είχε καλή προσαρμογή στα δεδομένα. Μολαταύτα, το κάθε ένα από αυτά οδήγησε στην εξαγωγή διαφορετικών εκτιμητών για την από κοινού κατανομή των πραγματικών μεταβλητών. Ο Chen (1992) συνοψίζει και επεκτείνει τα αποτελέσματα όλων αυτών των μεθόδων. Επίσης, οι Kuroda and Geng (2002), υπό το Μπεϋζιανό πρίσμα εξέτασης της λανθασμένης ταξινόμησης, χρησιμοποιούν αυτό το σύνολο δεδομένων για να δείξουν ότι ο αλγόριθμος επαύξησης δεδομένων (*data augmentation* ή *DA algorithm*) που χρησιμοποιούν είναι ισοδύναμης απόδοσης με τους EM και Fisher-score αλγόριθμους ως προς την εύρεση κατάλληλων εκτιμητών.

Εκτός από τις μεθόδους που ήδη παρουσιάστηκαν, έχουν προταθεί και άλλοι τρόποι για τη μοντελοποίηση πινάκων συνάφειας των οποίων τα στοιχεία υπόκεινται σε λανθασμένη ταξινόμηση. Ο Hochberg (1979) χρησιμοποιεί τη μέθοδο ελαχίστων τετραγώνων (*least square method*) που εισήγαγαν οι Grizzle et al. (1969). Οι Chiaccherini and Arnold (1977) εφαρμόζουν απλή μεγιστοποίηση (*straightforward maximization*) μέσω εκτιμητών μέγιστης πιθανοφάνειας για την προσαρμογή ενός μοντέλου ανεξαρτησίας σε έναν διδιάστατο πίνακα συνάφειας που υπόκειται σε λανθασμένη ταξινόμηση. Οι Espeland and Hui (1987) (βλ. επίσης Ekholm (1991) προτείνουν τη χρήση ενός Fisher-scoring αλγόριθμου για την προσαρμογή λογαριθμογραμμικών μοντέλων λανθασμένης ταξινόμησης χρησιμοποιώντας δεδομένα από διαφορετικά δείγματα επικύρωσης. Οι Ekholm and Palmgren (1987) και Palmgren and Ekholm (1987) περιγράφουν μία εναλλακτική προσέγγιση για την αντιμετώπιση της λανθασμένης ταξινόμησης. Σύμφωνα με τη μέθοδο που προτείνουν, οι πιθανότητες των κελιών των πινάκων συνάφειας των δειγμάτων καθορίζονται ευθέως μέσω των ποσοστών των κελιών περιθωρίου (*marginal cells*) για τις πραγματικές (ορθά ταξινομημένες) μεταβλητές και μέσω των πιθανοτήτων λανθασμένης ταξινόμησης. Οι προς εκτίμηση παράμετροι μπορούν να υπολογιστούν χρησιμοποιώντας μια τροποποιημένη μορφή

του επαναληπτικού αλγόριθμου επανασταθμισμένων ελαχίστων τετραγώνων (*iterative reweighted least squares algorithm* ή *IRLS*).

Μία πολύ ενδιαφέρουσα μέθοδο αντιμετώπισης της λανθασμένης ταξινόμησης παρουσιάζεται από τους Van der Hout and Van der Heijden (2002). Προτείνουν τη χρήση λογαριθμογραμμικών μοντέλων λανθανουσών μεταβλητών κατά την περίπτωση που οι πιθανότητες λανθασμένης ταξινόμησης είναι γνωστές, επικεντρώνοντας το ενδιαφέρον τους σε δεδομένα τυχαιοποιημένης απόκρισης (*randomized response* ή *RR data*). Παράλληλα, εξετάζουν τις περιπτώσεις που εμφανίζονται οριακές λύσεις (*boundary solutions*), όταν δηλαδή στον πίνακα συνάφειας του μοντέλου προκύπτουν μηδενικές αναμενόμενες συχνότητες κελιών. Επισημαίνουν δε ότι η μέθοδος που χρησιμοποιούν μπορεί να χρησιμοποιηθεί σε πεδία όπως η επιδημιολογία όταν η ευαισθησία και η ειδικότητα είναι γνωστές, ή στην εξόρυξη δεδομένων (*data mining*) όταν τα προσωπικά δεδομένα των ατόμων προστατεύονται μέσω εσκεμμένης στατιστικής διατάραξης τους (*intentional statistical perturbation*). Για τα μοντέλα λανθανουσών κατηγοριών και τη χρήση τους στην αντιμετώπιση της λανθασμένης ταξινόμησης θα αναφερθούμε στο διεξοδικότερα στο επόμενο Κεφάλαιο.

ΚΕΦΑΛΑΙΟ 7

Διόρθωση της επίδρασης λανθασμένης ταξινόμησης με χρήση επαναλαμβανόμενων μετρήσεων

7.1 Εισαγωγή

Όπως είδαμε στα προηγούμενα Κεφάλαια, μπορούμε να λάβουμε πληροφορίες για την επίδραση της λανθασμένης ταξινόμησης για τον εξεταζόμενο πληθυσμό χρησιμοποιώντας μεθόδους όπως οι πινακικές μέθοδοι ή η εφαρμογή μοντέλων προσαρμογής, συνήθως λογαριθμογραμικών, που απαιτούν την επιλογή και χρήση δειγμάτων επικύρωσης. Αντί των μεθόδων αυτών, μπορούμε να καταφύγουμε στην χρήση μεθόδων που βασίζονται στις επαναλαμβανόμενες μετρήσεις (*repeated measurements*) των εσφαλμένα ταξινομημένων μεταβλητών για την εκτίμηση παραμέτρων. Διάφορες μέθοδοι εκτίμησης έχουν αναπτυχθεί για την αντιμετώπιση τέτοιων περιπτώσεων. Η εκάστοτε επιλογή μίας μεθόδου εξαρτάται εν μέρει από τα δεδομένα που βρίσκονται στη διάθεση μας, όπως για παράδειγμα ο αριθμός των επαναλήψεων ανά μονάδα πληθυσμού, εάν οι μετρήσεις λαμβάνονται με τη χρήση μίας ή περισσότερων μεθόδων μέτρησης και εάν οι επαναλαμβανόμενες μετρήσεις είναι διαθέσιμες για όλο τον πληθυσμό ή μόνο για ένα μέρος του. Μία άλλη σημαντική παράμετρος είναι η πολυπλοκότητα του εξεταζόμενου προβλήματος, όπως για παράδειγμα το πλήθος των μεταβλητών και ο αριθμός των κατηγοριών τους, ποιες είναι οι παράμετροι που ενδιαφερόμαστε να εξετάσουμε, ποια είναι τα μοντέλα που θα προσαρμόσουμε στα δεδομένα και ποιες υποθέσεις θα ελέγξουμε.

Για την ανάλυση των δεδομένων που προέρχονται από επαναλαμβανόμενες μετρήσεις μπορούμε να χρησιμοποιήσουμε μοντέλα λανθανουσών κατηγοριών (*latent class models* ή *LC models*) (Goodman (1978), Hagenaars (1990) & (1993)). Οι μη παρατηρούμενες λανθάνουσες μεταβλητές είναι ουσιαστικά οι πραγματικές τιμές των λανθασμένα

ταξινομημένων μεταβλητών. Λαμβάνοντας υπ'όψιν τις λανθάνουσες τιμές, οι επαναλαμβανόμενες μετρήσεις μπορούν να θεωρηθούν ανεξάρτητες μεταξύ τους καθώς και με άλλες μεταβλητές. Οι εκτιμήσεις δίνονται κατά κανόνα με χρήση του ΕΜ αλγόριθμου, ο οποίος και έχει πολύ καλή προσαρμογή όσον αφορά την αντιμετώπιση προβλημάτων λανθανουσών κατηγοριών. Είναι επίσης δυνατό να καταφύγουμε στη χρήση του αλγόριθμου μεγιστοποίησης της πιθανοφάνειας Newton-Raphson (*Newton-Raphson likelihood maximization algorithm*), όπως και στην εύρεση κλειστού τύπου εκτιμήσεων (*closed-form estimates*) των παραμέτρων για απλούστερα μοντέλα.

7.2 Ταυτοποιησιμότητα των Μοντέλων

Όπως ισχύει και γενικότερα στην ανάλυση λανθανουσών κατηγοριών, έτσι και στα μοντέλα λανθανουσών κατηγοριών που χρησιμοποιούνται για τη διόρθωση λανθασμένης ταξινόμησης με επαναλαμβανόμενες κατηγορίες μία από τις βασικές έννοιες είναι το κατά πόσον το χρησιμοποιούμενο μοντέλο είναι ταυτοποιήσιμο (*identifiable*). Με την έννοια ταυτοποιήσιμο εννοούμε να υπάρχει 'ένα προς ένα' αντιστοιχία μεταξύ της κατανομής πιθανοτήτων των δεδομένων και των τιμών των παραμέτρων του μοντέλου. Κατά την εφαρμογή ενός μη ταυτοποιήσιμου μοντέλου, διαφορετικά άτομα μπορούν να προβούν σε διαφορετικά συμπεράσματα για τα παρατηρούμενα δεδομένα του ίδιου μοντέλου. Άρα η έννοια της ταυτοποιησιμότητας (*identifiability*) είναι εξέχουσας σημασίας.

Είναι λογικό να θεωρήσουμε ότι δεν είναι δυνατόν όλα τα μοντέλα να είναι ταυτοποιήσιμα και ότι πρέπει να ικανοποιούν ορισμένες συνθήκες ή να υπόκεινται σε περιορισμούς. Οι Liu and Liang (1991) δίνουν κάποιες βασικές προϋποθέσεις που πρέπει να πληροί ένα μοντέλο ώστε να θεωρείται ταυτοποιήσιμο για προβλήματα διόρθωσης λανθασμένης ταξινόμησης, επισημαίνοντας όμως ότι ισχύουν μόνο κατά την περίπτωση που η λανθασμένη ταξινόμηση του μοντέλου είναι μη διαφορίσιμη ή αν αυξηθεί εναλλακτικά ο αριθμός των μετρήσεων.

Για την εύρεση της σχέσης του αριθμού των επαναλαμβανόμενων μετρήσεων και των κατηγοριών ή υπο-κατηγοριών στις οποίες διαχωρίζεται ο εξεταζόμενος πληθυσμός με την ταυτοποίηση ενός μοντέλου που χρησιμοποιείται για τη διόρθωση της λανθασμένης ταξινόμησης, βασιζόμαστε στην διαδικασία που περιγράφεται αναλυτικά από τους Walter

and Irwig (1988). Χάριν ευκολίας, θα παραθέσουμε αποτελέσματα για δίτιμες αποκριτικές μεταβλητές, τα οποία μπορούν να γενικευτούν και για πολυεπίπεδες αποκριτικές μεταβλητές.

Ας θεωρήσουμε τη δίτιμη αποκριτική μεταβλητή *Y*, η οποία υπόκειται σε λανθασμένη ταξινόμηση, με πιθανότητες απουσίας και παρουσίας του εξεταζόμενου χαρακτηριστικού στο επιλεγμένο δείγμα $P_1 = \Pr(Y = 1)$ και $P_2 = \Pr(Y = 2)$ αντίστοιχα. Έστω *R* ο αριθμός των επαναλαμβανόμενων μετρήσεων που παρατηρούνται για όλα ή ένα μέρος των ατόμων του εξεταζόμενου δείγματος του πληθυσμού και *S* ο αριθμός των κατηγοριών στις οποίες διαχωρίζεται το δείγμα. Επίσης θεωρούμε την πιθανότητα λανθασμένα αρνητικού $\theta_Y(k)$ και την πιθανότητα λανθασμένα θετικού $\varphi_Y(k)$ της *k*-οστής μέτρησης, με $k = 1,...,R$. Οι ποσότητες *R* και *S* καθορίζουν τον αριθμό των διασταυρούμενων ταξινομήσεων (*cross-classifications*) στις οποίες ομαδοποιούνται τα δεδομένα, άρα και τον αριθμό των βαθμών ελευθερίας (β.ε.) που είναι διαθέσιμοι για την εκτίμηση των ζητούμενων παραμέτρων. Για κάθε μοντέλο που χαρακτηρίζεται από τις δύο αυτές ποσότητες πρέπει να υπολογιστεί ο αριθμός των ανεξάρτητων παραμέτρων που περιλαμβάνει, καθώς και οι β.ε. που το χρησιμοποιούμενο δειγματοληπτικό σχέδιο απαιτεί.

Σκοπός μας είναι να εκτιμήσουμε τις πιθανότητες $P_2^{(i)}$ $(i = 1,...,S)$ για κάθε κατηγορία *S* του δείγματος. Το συνολικό πλήθος των άγνωστων προς εκτίμηση παραμέτρων είναι $S(2R+1)$, εφόσον έχουμε να εκτιμήσουμε *R* πιθανότητες λανθασμένα αρνητικού, *R* πιθανότητες λανθασμένα θετικού και την πιθανότητα $P_2^{(i)}$ για *S* το πλήθος κατηγορίες. Εφόσον η μεταβλητή *Y* είναι δίτιμη, έχουμε $2^R - 1$ β.ε. για κάθε κατηγορία και συνολικά $S(2^R - 1)$ β.ε. για όλο το δείγμα. Συνεπώς η εκτίμηση όλων των παραμέτρων εξαρτάται αποκλειστικά από τον αριθμό των επαναλαμβανόμενων μετρήσεων *R*, αφού το *S* είναι κοινός παράγοντας. Ο αριθμός των παραμέτρων και των β.ε. για κάθε επαναλαμβανόμενη μέτρηση *R* δίνονται στον Πίνακα 7.1

ΠΙΝΑΚΑΣ 7.1

Πίνακας συσχέτισης αριθμού επαναλαμβανόμενων μετρήσεων R, αριθμού εκτιμώμενων παραμέτρων και β.ε. για υποκείμενα σε λανθασμένη ταξινόμηση μοντέλα
[ΠΗΓΗ : Walter and Irwig,1988]

Αριθμός επαναλαμβανόμενων μετρήσεων	R	1	2	3	4	5	.....
Αριθμός παραμέτρων	$2R+1$	3	5	7	9	11	.....
Αριθμός β.ε.	2^R-1	1	3	7	15	31	.....

Άρα για $R=3$ είναι ο μικρότερος αριθμός επαναλαμβανόμενων μετρήσεων για τον οποίο όλες οι παράμετροι μπορούν να εκτιμηθούν χωρίς επιπρόσθετες υποθέσεις και για οποιονδήποτε αριθμό κατηγοριών. Με άλλα λόγια, για $R\geq 3$ όλες οι παράμετροι είναι ταυτοποιήσιμες.

Ενδιαφέρον παρουσιάζει η εξέταση των περιπτώσεων όπου $R=2$. Σε αυτήν την περίπτωση, οι προς εκτίμηση παράμετροι δεν είναι πάντα ταυτοποιήσιμοι και είναι αναγκαία η εισαγωγή κάποιων περιορισμών. Για $S=1$, που αντιστοιχεί στη μονομεταβλητή περίπτωση, μία συνήθης υπόθεση και πρακτική είναι να θεωρήσουμε ότι όλες οι πιθανότητες λανθασμένης ταξινόμησης είναι ίδιες, δηλαδή έχουμε $\theta_Y(1)=\theta_Y(2)=\varphi_Y(1)=\varphi_Y(2)=\delta$. Οι εκτιμήσεις των ποσοτήτων $P_2^{(1)}=P_2$ και δ μπορούν να γίνουν μέσω κλειστών τύπων. Σημειώνουμε ότι οι λιγότερο αυστηροί περιορισμοί $\theta_Y(1)=\theta_Y(2)$ και $\varphi_Y(1)=\varphi_Y(2)$ δεν είναι επαρκείς σε αυτήν την περίπτωση, διότι η συμμετρία που διακρίνει τις επαναλαμβανόμενες μετρήσεις συνεπάγεται ότι υπάρχουν δύο μόνο βαθμοί ελευθερίας για την εκτίμηση των τριών παραμέτρων. Στην περίπτωση όπου $S=2$ κατηγορίες τίθονται υπό σύγκριση, είναι απαραίτητο να προβούμε στην υπόθεση ότι οι πιθανότητες λανθασμένης ταξινόμησης είναι οι ίδιες για όλες τις κατηγορίες. Κατά αυτόν τον τρόπο, το σύνολο των προς εκτίμηση παραμέτρων $\{P_2^{(1)}, P_2^{(2)}, \theta_Y(1), \theta_Y(2), \varphi_Y(1), \varphi_Y(2)\}$ είναι ταυτοποιήσιμο.

Ακόμα και για μοντέλα που μπορούν να θεωρηθούν ταυτοποιήσιμα, ο ορισμός των λανθανουσών κατηγοριών δεν είναι μονοσήμαντα ορισμένος. Εξαίρεση αποτελεί η περίπτωση που όλες οι πιθανότητες είναι ίσες με 1/2. Αυτό είναι ένα εγγενές πρόβλημα στα

μοντέλα λανθανουσών κατηγοριών. Για παράδειγμα, η ενδεχόμενη αυτή πολυσημία συνεπάγεται ότι κατά την περίπτωση όπου οι ποσότητες $\hat{P}_2^{(1)}, \hat{\theta}_Y, \hat{\varphi}_Y$ είναι εκτιμητές μέγιστης πιθανοφάνειας των ποσοτήτων $P_2^{(1)}$, θ_Y, φ_Y, τότε και οι ποσότητες $\hat{P}_2^{(1)\prime} = 1 - \hat{P}_2^{(1)}$, $\hat{\theta}'_Y = 1 - \hat{\theta}_Y = sen$, $\hat{\varphi}'_Y = 1 - \hat{\varphi}_Y = sp$ είναι εκτιμητές μέγιστης πιθανοφάνειας των ποσοτήτων αυτών. Σε αυτήν την περίπτωση, μία φυσική επιλογή εκτιμητών αποτελούν οι εκτιμητές εκείνοι που ικανοποιούν τη σχέση $\theta_Y + \varphi_Y \leq 1$ (βλ. Kuha and Skinner (1997)).

Οι συνθήκες κάτω από τις οποίες ένα μοντέλο είναι ταυτοποιήσιμο είναι περισσότερο πολύπλοκες όταν η προς εκτίμηση μεταβλητή έχει περισσότερες από δύο κατηγορίες ή όταν υπάρχουν αρκετές λανθασμένα ταξινομημένες μεταβλητές (Liu and Liang (1991)). Ωστόσο, η εφαρμογή μοντέλων με τρεις επαναλήψεις θεωρείται γενικότερα επαρκής, ενώ μοντέλα με δύο επαναλήψεις δεν είναι πάντα ταυτοποιήσιμα. Ακόμα όμως και στις πιο απλές περιπτώσεις εξέτασης ενός πληθυσμού, γίνεται χρήση μοντέλων με μεγαλύτερο αριθμό επαναλήψεων των μετρήσεων από τον αριθμό που θα μπορούσε να δώσει ικανοποιητικές εκτιμήσεις (*overidentified models*). Για την αντιμετώπιση τέτοιων μοντέλων, μπορούμε να 'θυσιάσουμε' έναν αριθμό βαθμών ελευθερίας ώστε να προβούμε στους απαραίτητους για τη συμπερασματολογία ελέγχους υποθέσεων και την προσπάθεια εύρεσης ενός όσο το δυνατόν 'οικονομικότερου' (*parsimonious*) μοντέλου.

7.3 Διόρθωση λανθασμένης ταξινόμησης μέσω μοντέλων λανθανουσών κατηγοριών για επαναλαμβανόμενες μετρήσεις

Όπως αναφέρθηκε και προηγουμένως, τα LC μοντέλα μπορούν να χρησιμοποιηθούν για την ανάλυση των δεδομένων που προέρχονται από επαναλαμβανόμενες μετρήσεις. Για την αντιμετώπιση των σφαλμάτων λόγω της λανθασμένης ταξινόμησης των δεδομένων και δεδομένου ότι τα εξεταζόμενα δεδομένα είναι κατηγορικά, θα καταφύγουμε στη χρήση ενός λογαριθμογραμμικού LC μοντέλου (*log-linear LC model*) για την περιγραφή της λανθασμένης ταξινόμησης καθώς και την αντιμετώπιση αυτής. Το μοντέλο που θα αναλυθεί στη συνέχεια βασίζεται στο λογιστικό LC μοντέλο που προτάθηκε από τους Kaldor and Clayton (1985) και αποτελεί μία εξαιρετική προσέγγιση του πλήρους LC μοντέλου. Στη

συνέχεια και προς καλύτερη κατανόηση του προτεινόμενου μοντέλου θα δοθεί και ένα παράδειγμα εφαρμογής του LLC μοντέλου για δίτιμες μεταβλητές.

7.3.1. Λογαριθμογραμικό μοντέλο λανθανουσών κατηγοριών

Έστω η μεταβλητή απόκρισης *Y*, η οποία προσμετράται χωρίς σφάλμα. Ορίζουμε στη συνέχεια τις ακόλουθες μεταβλητές

- $X_1,...,X_r$ οι μεταβλητές του που δεν υπόκεινται σε σφάλματα μέτρησης,
- $C_1,...,C_t$ οι λανθάνουσες μεταβλητές που δεν παρατηρούνται άμεσα στο δείγμα και
- $W_1,...,W_s$ οι μεταβλητές που προσμετρούν τις t λανθάνουσες μεταβλητές, με $s>t$ και υπόκεινται σε σφάλματα λανθασμένης ταξινόμησης.

Δεδομένων των μεταβλητών $\{C_k\ , k=1,...,t\}$, οι μεταβλητές $\{W_j\ , j=1,...,s\}$ είναι ανεξάρτητες μεταξύ τους και ανεξάρτητες της αποκριτικής μεταβλητής Y και των μεταβλητών $\{X_i\ , i=1,...,r\}$. Σχηματίζοντας τον $(1+r+s+t)$-διάστατο πίνακα που προκύπτει από τη διασταυρούμενη ταξινόμηση όλων των μεταβλητών του εξεταζόμενου προβλήματος, συμπεριλαμβανομένης και της αποκριτικής μεταβλητής, κάθε πιθανό αποτέλεσμα μπορεί να απεικονιστεί με τη βοήθεια ενός διανύσματος $\mathbf{Z} = (D, \mathbf{X}, \mathbf{W}, \mathbf{C})$, όπου D είναι η μεταβλητή απόκρισης, και **X**, **W**, **C** είναι τα διανύσματα-γραμμές *r*, *s* και *t* διάστασης αντίστοιχα που αντιπροσωπεύουν τις κατηγορίες των μεταβλητών $\{X_i\}$, $\{W_j\}$ και $\{C_k\}$ αντίστοιχα. Έστω $m_{\mathbf{Z}}$ το διάνυσμα που περιέχει τον αριθμό των ατόμων του πληθυσμού που ανήκουν στο διάνυσμα **Z**. Οι τιμές του $m_{\mathbf{Z}}$ είναι κατ' ουσίαν μη παρατηρήσιμες, εφόσον οι πραγματικές τιμές των λανθανουσών μεταβλητών είναι άγνωστες. Οι μόνες τιμές που μπορούν να παρατηρηθούν είναι οι $m_{D\mathbf{X}\mathbf{W}+}$, με το δείκτη (+) να υποδηλώνει το άθροισμα όλες τις κατηγοριών των t λανθανουσών μεταβλητών. Μολαταύτα, το διάνυσμα $m_{\mathbf{Z}}$ μπορεί να παραχθεί από ένα λογαριθμογραμμικό μοντέλο της ακόλουθης μορφής

$$\log \mu_{\mathbf{Z}} = \sum_A u_a^A , \qquad (7.1)$$

με $\mu_{\mathbf{Z}} = E(m_{\mathbf{Z}})$ το διάνυσμα των αναμενόμενων τιμών του $m_{\mathbf{Z}}$ και A να δηλώνει τα υποσύνολα των μεταβλητών D, $\{X_i\}$, $\{W_j\}$ και $\{C_k\}$. Για κάθε υποσύνολο A, το α είναι ένα υποσύνολο των στοιχείων του διανύσματος $\mathbf{Z}$ που αντιστοιχεί στις μεταβλητές του υποσυνόλου μεταβλητών A και οι παράγοντες της μορφής u_a^A είναι άγνωστοι. Στη συνέχεια,

επιλέγουμε μία τιμή του διανύσματος $\mathbf{Z}$ ως διάνυσμα αναφοράς, έστω $\mathbf{Z}_0$, και θέτουμε τον περιορισμό $u_a^A = 0$ για κάθε υποσύνολο στοιχείων α που περιέχει κοινά στοιχεία με ένα από τα στοιχεία του $\mathbf{Z}_0$. Σε τομείς όπως η επιδημιολογία που η αποκριτική μεταβλητή Y δηλώνει τα στάδια μίας νόσου, μία λογική επιλογή για κατηγορία αναφοράς είναι η κατηγορία $y = 0$ που αντιπροσωπεύει συνήθως τη μη ύπαρξη νόσου. Υποθέτουμε ότι το μοντέλο είναι ιεραρχικό, εννοώντας ότι η ύπαρξη του παράγοντα u_a^A συνεπάγεται την ύπαρξη όλων των παραγόντων u_b^B για όλα τα υποσύνολα Β του Α.

Λόγω της υπόθεσης δεσμευμένης ανεξαρτησίας στην οποία έχουμε προβεί, οι μεταβλητές $\{W_j\}$ που προσμετρούν τις λανθάνουσες μεταβλητές εμφανίζονται στις παραμέτρους u_a^A με τις μορφές

$$u_{w_j}^{W_j} \qquad \text{και} \qquad u_{w_j c_{k(j)}}^{W_j C_{k(j)}},$$

με $C_{k(j)}$ τη λανθάνουσα μεταβλητή που προσμετράται από την W_j μεταβλητή. Κατά την περίπτωση όπου οι $C_{k(j)}$ και W_j είναι διχοτομικές, οι μεταβλητές αυτές συνδέονται με απλές αριθμητικές σχέσεις μεταξύ τους και παράμετροι όπως η ευαισθησία και η ειδικότητα χρησιμοποιούνται για να δείξουν το μέγεθος της λανθασμένης ταξινόμησης των δεδομένων (βλ. Kaldor and Clayton (1985), Παράρτημα). Οι υπόλοιποι όροι στη σχέση (7.1) αντιπροσωπεύουν τις σχέσεις συσχέτισης μεταξύ των μεταβλητών D, $\{X_i\}$ και $\{C_k\}$ και είναι πανομοιότυποι με τις σχέσεις που εμφανίζονται σε μια τυπική αναπαράσταση ενός λογαριθμογραμμικού μοντέλου και τις έννοιες που αυτές αντιπροσωπεύουν.

Για να κατανοηθούν καλύτερα τα ανωτέρω, δίνεται ένα παράδειγμα από το χώρο της επιδημιολογίας. Εδώ η αποκριτική μεταβλητή δηλώνει τα στάδια μίας νόσου και είναι συνήθως δίτιμη. Σε αυτήν την περίπτωση, όροι της μορφής $u_{1x_i}^{YX_i}$ αντιπροσωπεύουν το λογάριθμο του λόγου σχετικών πιθανοφανειών ανάμεσα στη νόσο και τη x_i κατηγορία της X_i μεταβλητής εν συγκρίσει με την κατηγορία αναφοράς, υπό την προϋπόθεση ότι δεν υπάρχουν αλληλεπιδράσεις ανάμεσα στην κατηγορία X_i και στους άλλους παράγοντες κινδύνου $\{X_j,\ j \neq i\}$ και $\{C_k\}$. Παρόμοιες υποθέσεις μπορούμε να κάνουμε και για όρους της μορφής $u_{1c_k}^{YC_k}$. Επίσης, στις περιπτώσεις όπου εμφανίζονται όροι της μορφής

$$u^{X_1,\ldots,X_r\ ,\ C_1,\ldots,C_k}_{x_1,\ldots,x_r\ ,\ c_1,\ldots,c_k}$$

το λογαριθμικό μοντέλο μπορεί να μετασχηματιστεί σε ένα λογιστικό μοντέλο που περιγράφει τη σχέση ανάμεσα στη μεταβλητή απόκρισης Y και τις μεταβλητές $\{X_i\}$ και $\{C_k\}$ (Plackett (1974)), για τις οποίες η συμπερασματολογία που αφορά τη μεταξύ τους σχέση δεν εξαρτάται από το αν τα δεδομένα προέρχονται από μελέτες όπως ασθενών-μαρτύρων ή προοπτικές.

Οι εκτιμητές μέγιστης πιθανοφάνειας για τις παραμέτρους του εφαρμοζόμενου μοντέλου μπορούν εύκολα να επιτευχθούν με τη βοήθεια του EM αλγόριθμου (Dempster et al. (1977)). Με αρχική τιμή μία λογική εκτίμηση των $\{m_{\mathbf{Z}}\}$, προσαρμόζουμε το μοντέλο της σχέσης (7.1) χρησιμοποιώντας έναν επαναληπτικό κλιμακούμενο αλγόριθμο (βλ. Bishop et al. (1974)) για τον υπολογισμό των εκτιμητών μέγιστης πιθανοφάνειας $\hat{\mu}_{\mathbf{Z}}$ των $\mu_{\mathbf{Z}}$ του μοντέλου. Ακολούθως, τα 'δεδομένα' $m_{\mathbf{Z}}$ επανεκτιμούνται για το βήμα αναμενόμενης τιμής (E-step) του EM αλγόριθμου μέσω της σχέσης

$$\hat{m}_{\mathbf{Z}} = m_{D\mathbf{X}\mathbf{W}+}\frac{\hat{\mu}_{\mathbf{Z}}}{\hat{\mu}_{D\mathbf{X}\mathbf{W}+}}$$

και το μοντέλο επανεκτιμάται από τα νέα δεδομένα, διαδικασία που συνεχίζεται έως ότου επιτευχθεί σύγκλιση του αλγόριθμου.

Για να γίνει περισσότερο κατανοητή η χρήση του προτεινόμενου λογαριθμογραμμικού LC μοντέλου για τη διόρθωση της λανθασμένης ταξινόμησης μέσω επαναλαμβανόμενων μετρήσεων, δίνεται η ακόλουθη εφαρμογή (βλ. Kaldor and Clayton (1985))

Έστω η διχοτομική αποκριτική μεταβλητή Y που συμβολίζει τα στάδια μίας νόσου σε μία μελέτη μαρτύρων-ασθενών (*case-control study*) χρόνιων νοσημάτων, όπως ο καρκίνος. Για τις περισσότερες μεταβλητές ενδιαφέροντος σε μία τέτοιου είδους μελέτη, οι επαναλαμβανόμενες μετρήσεις είναι διαθέσιμες μόνο για την ομάδα των μαρτύρων (*controls*). Αυτό μπορεί να οφείλεται είτε στην μικρή πιθανότητα επιβίωσης των ασθενών (*case*) είτε στις αλλαγές που η νόσος προκαλεί στους εξεταζόμενους παράγοντες κινδύνου (*risk factors*). Έστω ότι εξετάζουμε την συσχέτιση της εμφάνισης καρκίνου στο εξεταζόμενο δείγμα του πληθυσμού με έναν παράγοντα κινδύνου όπως ο δείκτης ανθρώπινων λευκοκυτταρικών αντιγόνων (HLA δείκτης) και ο παράγοντας κινδύνου αυτός υπόκειται σε

εσφαλμένη ταξινόμηση. Για την αντιμετώπιση της λανθασμένης ταξινόμησης του παράγοντα κινδύνου, προχωρούμε σε επαναλαμβανόμενο προσδιορισμό του (*repeat determination*), καταφεύγουμε δηλαδή στη συλλογή s το πλήθος επαναλαμβανόμενων μετρήσεων του HLA δείκτη. Ας υποθέσουμε ότι έχουμε επαναλαμβανόμενες μετρήσεις του εξεταζόμενου πληθυσμού για το σύνολο του εξεταζόμενου πληθυσμού, δηλαδή και για ασθενείς και για μάρτυρες. Τότε, σύμφωνα με τους ορισμούς που δόθηκαν προηγουμένως στην Παράγραφο αυτή ορίζουμε

- C τον δείκτη HLA, δηλαδή την μη παρατηρούμενη λανθάνουσα μεταβλητή του δείγματος
- $W_1,...,W_s$ τις υποκείμενες σε σφάλματα λανθασμένης ταξινόμησης μεταβλητές που προσμετρούν την λανθάνουσα μεταβλητή C και συμβολίζουν τις s το πλήθος επαναλαμβανόμενες μετρήσεις του HLA δείκτη.

Εφόσον δεν υπάρχουν μεταβλητές χωρίς λανθασμένη ταξινόμηση στο πληθυσμό ισχύει $r=0$ και η ύπαρξη μία μόνο λανθάνουσας μεταβλητής συνεπάγεται ότι $t=1$. Θεωρώντας ότι η εξέταση του δείκτη HLA γίνεται ως προς τις κρίσιμες τιμές του που καταδεικνύουν την ύπαρξη ή μη ύπαρξη καρκίνου στα άτομα του πληθυσμού, η κατηγορική μεταβλητή C τον αντιπροσωπεύει είναι διχοτομική. Το ίδιο συμβαίνει και με τις μεταβλητές $W_1,...,W_s$, εφόσον προσμετράν τη λανθάνουσα μεταβλητή C. Σύμφωνα με τα ανωτέρω, το λογιστικό μοντέλο λανθανουσών κατηγοριών που μπορεί να εφαρμοστεί μέσω της σχέσης (7.1) είναι το ακόλουθο

$$\log \mu_{\mathbf{z}} = u + u_y^Y + u_c^C + u_{yc}^{YC} + \sum_{j=1}^{s} \left(u_{w_j}^{W_j} + u_{cw_j}^{CW_j} \right)$$

και επιλύεται μέσω της μεθοδολογίας που αναφέρθηκε προηγουμένως στην Παράγραφο αυτή.

7.4 Σχετική βιβλιογραφία

Η πιο απλή εφαρμογή επαναλαμβανόμενων μετρήσεων είναι η εκτίμηση πληθυσμιακών ποσοστών σε έναν υποπληθυσμό ($S=1$). Ο Harper (1964) δίνει εκτιμητές κλειστής μορφής για πληθυσμιακά ποσοστά προερχόμενα από δίτιμες μεταβλητές για δύο περιπτώσεις, με $R=2$ και $R=3$ επαναλήψεις. Οι Quade et al. (1980) εφαρμόζουν τον EM αλγόριθμο στο δεύτερο παράδειγμα του Harper. Οι Dawid and Skewe (1979) περιγράφουν τη διαδικασία

εύρεσης εκτιμητών ποσοστών για πολυεπίπεδες μεταβλητές με χρήση $R = 3$ επαναλαμβανόμενων μετρήσεων.

Κατά την εκτίμηση σε δύο υποπληθυσμούς ($S = 2$), για $R = 2$ επαναλαμβανόμενες μετρήσεις οι Hui and Walter (1980), προτείνοντας ένα μοντέλο για την εκτίμηση της ευαισθησίας και της ειδικότητας δύο διαγνωστικών ελέγχων απουσία του 'χρυσού κανόνα', δίνουν εκτιμητές κλειστής μορφής και πίνακες πληροφορίας για δυωνυμικά ποσοστά και πιθανότητες λανθασμένης ταξινόμησης όταν οι πιθανότητες λανθασμένης ταξινόμησης είναι διαφορετικές σε κάθε μέτρηση. Σε αυτή την περίπτωση, οι διαφορές ανάμεσα στα ποσοστά των υποπληθυσμών, καθώς και άλλα μέτρα συσχέτισης, μπορούν να υπολογιστούν από τις εκτιμώμενες τιμές. Ο Clayton (1985) ερευνά μια παρόμοια περίπτωση, με τη διαφορά ότι η ειδικότητες και οι ευαισθησίες παραμένουν σταθερές στις δύο μετρήσεις και ότι ένα μόνο μέρος των ατόμων του εξεταζόμενου πληθυσμού προσμετράται και στις δύο επαναλήψεις. Επίσης, ασχολείται διεξοδικά με την εκτίμηση διαστημάτων εμπιστοσύνης για τις παραμέτρους χρησιμοποιώντας με χρήση μεθόδων profile-πιθανοφανειών. Οι Ekholm and Palmgren (1982) θεωρούν ένα μοντέλο με ένα μόνο αριθμό επαναλήψεων, το οποίο είναι όμως ταυτοποιήσιμο λόγω της μορφής των δειγματικών ποσοστών $P_2^{(j)}$ και του γεγονότος πως οι πιθανότητες λανθασμένης ταξινόμησης είναι ίδιες σε κάθε υποπληθυσμό.

Ένα πιο σύνθετο παράδειγμα μοντέλου δίνεται από τους Palmgren and Ekholm (1987), οι οποίοι προσαρμόζουν μοντέλα με $R = 2$ επαναλήψεις στη μέτρηση δύο εσφαλμένα ταξινομημένων μεταβλητών μέσω της μεθόδου Newton-Raphson. Οι Duffy et al. (1989) θεωρούν μία περίπτωση όπου δύο από τις τρεις εξεταζόμενες μεταβλητές είναι λανθασμένα ταξινομημένες. Η εύρεση των εκτιμήσεων τους για τις πιθανότητες λανθασμένης ταξινόμησης γίνεται μέσω ενός συνόλου επαναλαμβανόμενων μετρήσεων. Οι εκτιμήσεις αυτές χρησιμοποιούνται για τον υπολογισμό ενός εκτιμητή του πίνακα συνάφειας με στοιχεία τα στοιχεία τα εσφαλμένα ταξινομημένων μεταβλητών, ο οποίος χρησιμοποιείται στη συνέχεια για την περαιτέρω ανάλυση των δεδομένων. Κατά συνέπεια, η μέθοδος αυτή είναι ουσιαστικά ίδια με τις πινακικές μεθόδους διόρθωσης που αναλύθηκαν στο 5^{ο} Κεφάλαιο, με τη διαφορά ότι οι πιθανότητες λανθασμένης ταξινόμησης εκτιμήθηκαν μέσω επαναλαμβανόμενων μετρήσεων και όχι μέσω δειγμάτων επικύρωσης.

Δύο βασικές μεθοδολογίες εκτίμησης οιουδήποτε αριθμού μεταβλητών με επαναλαμβανόμενες μετρήσεις μέσω μοντελοποίησης παρουσιάζονται από τους Liu and Liang (1991) και Kaldor and Clayton (1985). Οι πρώτοι πραγματεύονται μία μέθοδο που

σχετίζεται με πινακικές μεθόδους διόρθωσης λανθασμένης ταξινόμησης, η οποία είναι πιο ευέλικτη και γενική. Προτείνουν τη χρήση ενός γενικευμένου γραμμικού μοντέλου (*generalized linear model*), με αποκριτική μεταβλητή μία ορθά ταξινομημένη μεταβλητή που είναι ταυτοποιήσιμη. Το πρώτο βήμα της μεθόδου που προτείνουν είναι η εκτίμηση των παραμέτρων λανθασμένης ταξινόμησης μέσω ενός συνόλου επαναλαμβανόμενων μετρήσεων με μεθόδους λανθανουσών κατηγοριών. Δεδομένων αυτών των εκτιμήσεων, οι παράμετροι του μοντέλου παλινδρόμησης εκτιμούνται στη συνέχεια από τα υπόλοιπα δεδομένα χρησιμοποιώντας τεχνικές επίλυσης ημι-πιθανοφάνειας (*quasi-likelihood techniques*). Οι δεύτεροι παρουσιάζουν μία εξαιρετική προσέγγιση για το πλήρες μοντέλο λανθανουσών κατηγοριών. Αρχικά, υποθέτουν ότι όλα τα άτομα του εξεταζόμενου πληθυσμού προσμετρώνται μέσω επαναλαμβανόμενων μετρήσεων. Στη συνέχεια, ορίζουν ένα λογαριθμογραμμικό μοντέλο που συνδέει όλες τις παρατηρούμενες και λανθάνουσες μεταβλητές και η εκτίμηση των παραμέτρων του μοντέλου γίνεται με τη χρήση του ΕΜ αλγόριθμου. Είναι ακόμα δυνατή και η εκτίμηση διαστημάτων εμπιστοσύνης για της παραμέτρους αυτούς μέσω μίας συνάρτησης profile-πιθανοφάνειας.

Σε κάποιες περιπτώσεις, είναι εφικτή η χρήση μεθόδων απλούστερης μορφής και ταχύτερης σύγκλισης σε σχέση με τις μεθόδους εκτίμησης μέγιστης πιθανοφάνειας. Οι McClish and Quade (1985) προτείνουν τη χρήση δύο διαδοχικών σχεδιασμών για την εκτίμηση ποσοστών. Και στους δύο αυτούς σχεδιασμούς ένα άτομο του πληθυσμού ταξινομείται σε μία από δύο κατηγορίες σύμφωνα με το που το κατατάσσει η πλειοψηφία ενός συνόλου επαναλαμβανόμενων μετρήσεων. Στην περίπτωση που κατατάσσεται ισοδύναμα και στις δύο κατηγορίες, γίνεται εφαρμογή μίας ακόμα μέτρησης. Οι Marshall and Graham (1984) εκτιμούν παραμέτρους συσχέτισης χρησιμοποιώντας μόνο τα άτομα του πληθυσμού για τα οποία δύο επαναλαμβανόμενες μετρήσεις είναι σύμφωνες. Αυτές οι δύο μέθοδοι είναι περισσότερο εμπειρικές, εφαρμόζονται σε συγκεκριμένες περιπτώσεις και είναι σαφέστερα λιγότερο αποτελεσματικές από τις μεθόδους εκτίμησης μεγίστων πιθανοφανειών. Θα μπορούσαν όμως να χρησιμοποιηθούν για την εύρεση αρχικών τιμών για επαναληπτικούς αλγόριθμους εκτίμησης.

7.5 Συγκρίσεις μεταξύ πινακικών μεθόδων και μεθόδων επαναλαμβανόμενων μετρήσεων

Όπως εξετάστηκε στα προηγούμενα Κεφάλαια, υπάρχουν δύο τρόποι προσέγγισης για την αντιμετώπιση ύπαρξης λανθασμένης ταξινόμησης σε έναν εξεταζόμενο πληθυσμό, η διόρθωση με χρήση δειγμάτων επικύρωσης, που διακρίνεται σε χρήση πινακικών μεθόδων ή μεθόδων μοντελοποίησης, και η διόρθωση μέσω χρήσης επαναλαμβανόμενων μετρήσεων. Η επιλογή μεταξύ της χρήσης των δύο σχεδιασμών αντιμετώπισης λανθασμένης ταξινόμησης είναι συνήθως διακριτή. Και αυτό γιατί τα διαθέσιμα δεδομένα και το ερευνητικό πλαίσιο κάτω από το οποίο συλλέγονται (το είδος της διεξαγόμενης έρευνας) συνήθως καθορίζει και τη μορφή των δεδομένων αυτών. Για παράδειγμα, δεδομένα τα οποία προέρχονται από διαχρονικές μελέτες, από μελέτες πλαισίου ή μελέτες κοορτής αποτελούνται από επαναλαμβανόμενες μετρήσεις που γίνονται στα ίδια άτομα ενός πληθυσμού, είναι με άλλα λόγια από σχεδιασμό επαναλαμβανόμενα. Σε αυτές τις περιπτώσεις περιοριζόμαστε εκ των πραγμάτων σε χρήση μεθόδων επαναλαμβανόμενων μετρήσεων. Υπάρχουν όμως και μελέτες στις οποίες είναι δυνατόν να χρησιμοποιηθούν και οι δύο μέθοδοι διόρθωσης λανθασμένης ταξινόμησης. Στις περιπτώσεις αυτές κατά τις οποίες μπορεί να γίνει επιλογή της προς χρήση μεθόδου διόρθωσης, θεωρείται απαραίτητο να εξεταστεί ποια από τις δύο αποτελεί και τη βέλτιστη μέθοδο για τον εξεταζόμενο πληθυσμό. Οι Duffy et al. (1992) σε σχετική έρευνα τους για 2×2 πίνακες με λανθασμένη ταξινόμηση στην επεξηγηματική μεταβλητή κατέφυγαν στη χρήση δύο σχεδιασμών για τη σύγκριση εκτιμητών : μίας μελέτης με εξωτερικό δείγμα επικύρωσης και μίας μελέτης με εσωτερικό δείγμα επικύρωσης που έκανε χρήση δύο επαναλαμβανόμενων μετρήσεων της επεξηγηματικής μεταβλητής. Παρουσιάζοντας μία σειρά από παραδείγματα, προέβησαν στο γενικό συμπέρασμα ότι οι επαναλαμβανόμενες μετρήσεις δίνουν πιο ακριβείς εκτιμήσεις σε σχέση με τη χρήση δειγμάτων προερχόμενων από μία μελέτη επικύρωσης.

Ο Marshall (1989) συγκρίνει τα δύο αυτά είδη σχεδιασμών σε ένα γενικότερο πλαίσιο. Υποστηρίζει και αυτός έμμεσα τη χρήση επαναλαμβανόμενων μετρήσεων έναντι της χρήσης μεθόδων επικύρωσης, εκφράζοντας αμφιβολίες για την ‘ευαισθησία’ της ανίχνευσης σφαλμάτων ταξινόμησης μέσω της χρήσης πιθανοτήτων λανθασμένης ταξινόμησης. Σε αυτό το σημείο όμως πρέπει να τονιστεί ότι εν συγκρίσει με τη χρήση πιθανοτήτων λανθασμένης ταξινόμησης, η εύρεση καλύτερων εκτιμήσεων μπορεί να γίνει δυνατή χρησιμοποιώντας πιθανότητες βαθμονόμησης, όπως είδαμε και σε προηγούμενο Κεφάλαιο.

ΚΕΦΑΛΑΙΟ 8

Εφαρμογή μεθόδων διόρθωσης λανθασμένης ταξινόμησης

8.1 Εισαγωγή

Στο Κεφάλαιο αυτό θα ερευνήσουμε τα αποτελέσματα εφαρμογής κάποιων από τις προτεινόμενες μεθόδους για την αντιμετώπιση της λανθασμένης ταξινόμησης που παρουσιάστηκαν σε προηγούμενα Κεφάλαια της παρούσας εργασίας. Συγκεκριμένα, θα εφαρμόσουμε τη μέθοδο διόρθωσης λανθασμένης ταξινόμησης μέσω χρήσης λογαριθμικών μοντέλων[5] με εφαρμογή σχεδιασμού τριπλής και διπλής δειγματοληψίας στον ερευνώμενο πληθυσμό και θα παραθέσουμε τη συμπερασματολογία που δίνεται από τους Chen et al. (1984) και Chen (1989) αντίστοιχα.

8.1.1 Δεδομένα του εξεταζόμενου προβλήματος

Τα δεδομένα τα οποία θα χρησιμοποιηθούν στην παρούσα ανάλυση προέρχονται από μία έρευνα για την ασφάλεια των αυτοκινητοδρόμων που έλαβε χώρα στη Βόρεια Καρολίνα των Ηνωμένων Πολιτειών τα έτη 1974-1975. Τα προβλήματα που τίθονται προς εξέταση στο συγκεκριμένο σύνολο δεδομένων είναι η εκτίμηση της αποτελεσματικότητας της χρήσης ζώνης ασφαλείας στη μείωση τραυματισμών σε αυτοκινητιστικά ατυχήματα, καθώς και η εξέταση της φύσης των σφαλμάτων λανθασμένης ταξινόμησης που οφείλονται στις αστυνομικές αναφορές.

Σύμφωνα με σχετική έρευνα του Hochberg (1977), μία από τις δυσκολίες στην αξιολόγηση της αποτελεσματικότητας της χρήσης ζώνης έγκειται στο γεγονός ότι οι περισσότερες πληροφορίες για την πρόκληση τραυματισμών και τη χρήση ζώνης από τους

[5] Ας σημειωθεί τα χρησιμοποιούμενα λογαριθμογραμμικά μοντέλα θα αναπαριστώνται μέσω των υψηλότερης τάξης *u*-όρων τους.

επιβάτες των αυτοκινήτων λαμβάνεται από αστυνομικές αναφορές για τα αντίστοιχα ατυχήματα. Όπως έχει σημειωθεί από εργαζόμενους στη έρευνα ασφάλειας των αυτοκινητοδρόμων, υπάρχουν συστηματικά σφάλματα λανθασμένης ταξινόμησης στις αστυνομικές αναφορές σε σχέση με τους τραυματισμούς των επιβατών και τη χρήση ζώνης ασφαλείας. Μία άλλη δυσκολία που μπορεί να αντιμετωπιστεί σε μία τέτοιου είδους έρευνα είναι το γεγονός ότι η λήψη των πραγματικών μετρήσεων για την αποτελεσματικότητα της χρήσης ζώνης ασφαλείας ως προς τη μείωση των τραυματισμών σχετίζεται με το γεγονός πως οι επιβάτες των αυτοκινήτων που φορούν ή δεν φορούν ζώνη δεν έχουν παρόμοιες κατανομές για τα επίπεδα ορισμένων σχετικών παραγόντων, όπως για παράδειγμα ο τύπος του αυτοκινήτου (κατασκευή, μέγεθος κλπ.), το περιβάλλον και ο τύπος του ατυχήματος (ανατροπή αυτοκινήτου, μετωπική σύγκρουση κλπ.).

Προκειμένου να επιλυθούν οι δυσκολίες αυτές, προχωρούμε αρχικά στο διαχωρισμό των μηχανισμών ταξινόμησης των ατόμων του πληθυσμού. Ο μηχανισμός ορθής (πραγματικής) ταξινόμησης βασίζεται σε λεπτομερείς αναφορές σχετικά με τους τραυματισμούς, τη χρήση ζώνης και λοιπών παραγόντων από νοσοκομεία για τα άτομα που νοσηλεύτηκαν ή μέσω τηλεφωνικής επικοινωνίας για τα άτομα που δεν νοσηλεύτηκαν. Ως μηχανισμός λανθασμένης ταξινόμησης θεωρούνται οι αστυνομικές αναφορές (p -αναφορές). Για το λόγο αυτό, εφεξής θα αναφερόμαστε στον μηχανισμό ορθής ταξινόμησης ως "μη αστυνομικές αναφορές" (np -αναφορές).

Στη συνέχεια, ορίζουμε τις προς εξέταση μεταβλητές. Έστω Y και E οι μεταβλητές που δηλώνουν τις np -αναφορές τραυματισμών και χρήσης ζώνης αντίστοιχα, Y' και E' οι μεταβλητές που δηλώνουν τις p -αναφορές τραυματισμών και χρήσης ζώνης αντίστοιχα, C η μεταβλητή που δηλώνει το φύλο και D η μεταβλητή που δηλώνει τη ζημιά του αυτοκινήτου. Για τις μεταβλητές C και D υποθέτουμε ότι είναι ελεύθερες σφαλμάτων ταξινόμησης. Επίσης, όλες οι μεταβλητές που εξετάζονται είναι δίτιμες, εφόσον οι τιμές που μπορούν να πάρουν οι Y, E, Y' και E' είναι "ναι" ή "όχι", οι τιμές της C είναι "άντρας" ή "γυναίκα" και οι τιμές της D είναι "χαμηλή" ή "υψηλή". Το δείγμα που θα χρησιμοποιηθεί παρουσιάζεται σε μία έρευνα του Hochberg (1977) και αποτελείται από $n_2 = 80084$ άτομα που καταγράφηκαν πλήρως από την αστυνομία το 1974 λαμβάνοντας υπ' όψιν τις ως άνω ορισμένες μεταβλητές. Οι συχνότητες αυτών των p -αναφορών δίνονται στον ακόλουθο $2\times2\times2\times2$ $(Y'E'CD)$ πίνακα (Πίνακας 8.1)

ΠΙΝΑΚΑΣ 8.1

Συχνότητες ατυχημάτων ως προς τη ζημιά των αυτοκινήτων (D), το φύλο των οδηγών (C) ,τη χρήση ζώνης (E'), και τους τραυματισμούς (Y'), με βάση 80,084 (n_2) αστυνομικές αναφορές (p)
[ΠΗΓΗ : Hochberg ,1977]

Car damage	**Low**				**High**			
Driver's sex	**Male**		**Female**		**Male**		**Female**	
p - belt use	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
p - not injured	22536	3006	11199	1262	17476	2155	6964	728
p - injured	1687	199	1422	117	6746	583	3707	297

Το δεύτερο δείγμα που θα χρησιμοποιηθεί, και αποτελεί το δείγμα επικύρωσης της έρευνας, αποτελείται από $n = 1796$ άτομα που καταγράφηκαν το 1975 ταυτόχρονα από μη αστυνομικές και αστυνομικές αναφορές, δηλαδή από τον ορθό και λανθασμένα μηχανισμό ταξινόμησης αντίστοιχα. Οι συχνότητες αυτών των αναφορών δίνονται στον $2\times2\times2\times2\times2\times2$ πίνακα (*YY'EE'CD*) (Πίνακας 8.2).

ΠΙΝΑΚΑΣ 8.2

Συχνότητες ατυχημάτων ως προς τη ζημιά των αυτοκινήτων (D), το φύλο των οδηγών (C) και ως προς την ταξινόμηση από αστυνομικές (p) και μη αστυνομικές αναφορές για τη χρήση ζώνης και τους τραυματισμούς με βάση 1796 (n) ατυχήματα

[ΠΗΓΗ : Hochberg ,1977]

Driver's sex	**Male**				**Female**			
np-injury	**No**		**Yes**		**No**		**Yes**	
np-belt use	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
Car damage: low								
p-not injured/unbelted	407	62	45	7	206	18	37	5
p-not injured/belted	6	47	1	6	1	17	0	1
p-injured/unbelted	5	1	32	4	4	0	29	0
p-injured/belted	0	1	1	2	3	0	1	0
Car damage: high								
p-not injured/unbelted	299	20	59	9	102	7	53	4
p-not injured/belted	4	30	1	6	2	6	1	3
p-injured/unbelted	11	2	118	5	5	1	79	1
p-injured/belted	1	2	0	9	0	0	1	6

8.2 Διόρθωση λανθασμένης ταξινόμησης μέσω μεθόδου τριπλής δειγματοληψίας

Για την αντιμετώπιση της λανθασμένης ταξινόμησης θα χρησιμοποιήσουμε τη μέθοδο τριπλής δειγματοληψίας που δίνεται από τους Chen et. al (1984) ως μας μέθοδο. Παρόλο που οι περισσότεροι ερευνητές επικεντρώνουν το ενδιαφέρον τους σε μεθόδους διόρθωσης λανθασμένης ταξινόμησης με χρήση σχεδίων διπλής δειγματοληψίας, η εισαγωγή ενός ακόμα δείγματος, το οποίο μάλιστα αποτελείται μόνο από ορθά ταξινομημένα δεδομένα, θεωρούμε πως καθιστά δυνατή την επιλογή ενός βέλτιστου μοντέλου που έχει καλύτερη προσαρμογή στα δεδομένα και ταυτόχρονα αντιμετωπίζει αποτελεσματικότερα το φαινόμενο της λανθασμένης ταξινόμησης. Όπως αναφέρουν στην έρευνα τους οι Chen et. al, υπήρξαν αρκετές εκατοντάδων περιπτώσεων στην μελέτη του Hochberg για τη Βόρεια Καρολίνα για

τις οποίες ήταν διαθέσιμες μόνο np-αναφορές (αναφορές νοσοκομείων). Επειδή ένα τέτοιο δείγμα δεν είναι διαθέσιμο, δημιουργήθηκε ένα σύνολο τεχνητών δεδομένων μεγέθους $n_1 = 905$ παίρνοντας τις μισές περίπου συχνότητες του Πίνακα 8.2 και αθροίζοντας τες ως προς τις p-αναφορές. Το δείγμα αυτό θα χρησιμοποιηθεί στη μέθοδο τριπλής δειγματοληψίας. Οι συχνότητες αυτών των p-αναφορών δίνονται στον ακόλουθο $2\times2\times2\times2$ $(YECD)$ πίνακα (Πίνακας 8.3)

ΠΙΝΑΚΑΣ 8.3

Συχνότητες ατυχημάτων ως προς τη ζημιά των αυτοκινήτων (D), το φύλο των οδηγών (C) ,τη χρήση ζώνης (E), και τους τραυματισμούς (Y), με βάση 905 (n_1) μη αστυνομικές αναφορές (np)
[ΠΗΓΗ : Chen et al. 1984]

Car damage	**Low**				**High**			
Driver's sex	**Male**		**Female**		**Male**		**Female**	
np - belt use	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
np - not injured	200	55	100	13	150	30	54	8
np - injured	50	10	40	3	100	15	70	7

Με βάση τα ανωτέρω, τα τρία δείγματα που θα χρησιμοποιηθούν για την εφαρμογή της TSS είναι $n \equiv n(\mathbf{Y},\mathbf{X},\mathbf{Z}) = 1796$, $n_1 \equiv n_1(\mathbf{Y},\mathbf{X},\mathbf{Z}) = 905$ και $n_2 \equiv n_2(\mathbf{X},\mathbf{Z}) = 80084$, με $\mathbf{Y} = (Y,E)$, $\mathbf{X} = (Y',E')$ και $\mathbf{Z} = (C,D)$.

Το κορεσμένο μοντέλο που αντιστοιχεί στα εξεταζόμενα δεδομένα είναι το $H(YY'EE'CD)$. Αρχικά, για να διαπιστωθεί εάν τα τρία δείγματα της TSS συνάδουν το ένα με το άλλο, ελέγχουμε την προσαρμογή του κορεσμένου μοντέλου στα δεδομένα. Ο έλεγχος λόγου πιθανοφανειών (LR-test) δίνει $G^2 = 20.77$ με 30 β.ε. Ως εκ τούτου, αποδεχόμαστε την προσαρμογή του $H(YY'EE'CD)$ και προχωρούμε στη περαιτέρω διερεύνηση της δομής της λανθασμένης ταξινόμησης.

Για την εύρεση της κατάλληλης δομής ακολουθούνται δύο διαφορετικές αλληλουχίες εξεταζόμενων μοντέλων.

1η Αλληλουχία μοντέλων

$$H(YY'EE'D, YECD) \rightarrow H(YY'EE', Y'EE'D, YECD) \rightarrow H(YY'EE', Y'D, YECD)$$

Για αυτή την αλληλουχία μοντέλων σταματάμε στο μοντέλο $H(YY'EE', Y'EE'D, YECD)$ εφόσον το μοντέλο $H(YY'EE', Y'D, YECD)$ απορρίπτεται μέσω των LR-ελέγχων. (Για τους υπό συνθήκη LR-ελέγχους και γενικά για τους LR-ελέγχους των μοντέλων που προσαρμόζουμε στα δεδομένα βλ. Πίνακα 8.4)

2η Αλληλουχία μοντέλων

$$H(YY'EE', Y'E'CD, YECD) \rightarrow H(YY'EE', E'C, Y'D, YECD) \rightarrow H(YY'EE', Y'D, YECD)$$

Για αυτή την αλληλουχία μοντέλων σταματάμε στο μοντέλο $H(YY'EE', Y'E'CD, YECD)$ εφόσον το μοντέλο $H(YY'EE', E'C, Y'D, YECD)$ απορρίπτεται μέσω των υπό συνθήκη LR-ελέγχων.

ΠΙΝΑΚΑΣ 8.4

Λογαριθμογραμμικά μοντέλα LR-test, και β.ε. (Ι)
[ΠΗΓΗ : Chen et al. 1984]

Models		L.R. G^2	d.f.
$H(YY'EE'CD)$		20.77	30
$H(YY'EE'D, YECD)$		53.37	54
$H(YY'EE', Y'EE'D, YECD)$		56.05	60
$H(YY'EE', Y'E'CD, YECD)$		50.20	57
$H(YY'EE', E'C, Y'D, YECD)$		65.46	64
$H(YY'EE', Y'D, YECD)$		70.30	65
$H(YY'EE', YECD)$		187.98	66
$H(YY'EE', Y'E'CD, YECD)$	$H^*(YEC, YCD, YED, ECD)$	50.696	58
$H(YY'EE', Y'E'CD, YECD)$	$H^*(YE, YCD, ECD)$	52.408	60
$H(YY'EE', Y'E'CD, YECD)$	$H^*(YCD, ECD)$	56.915	61

Η επιλογή ανάμεσα μοντέλα $H(YY'EE', Y'EE'D, YECD)$ και $H(YY'EE', Y'E'CD, YECD)$ είναι κάπως αυθαίρετη. Το πρώτο μοντέλο υποδεικνύει ότι η λανθασμένη ταξινόμηση είναι

συνάρτηση των ζημιών του αυτοκινήτου και δεν εξαρτάται από το φύλο. Το δεύτερο μοντέλο υποδεικνύει ότι η λανθασμένη ταξινόμηση εξαρτάται από τις ζημιές του αυτοκινήτου και από το φύλο μαζί. Με βάση την καλύτερη προσαρμογή που δείχνει να έχει το δεύτερο μοντέλο επιλέγουμε αυτό, δηλαδή το $H(YY'EE', Y'E'CD, YECD)$. Οι εκτιμώμενοι πίνακες λανθασμένης ταξινόμησης για κάθε επίπεδο των ζημιών αυτοκινήτου και φύλου δίνονται στον πίνακα 8.5

ΠΙΝΑΚΑΣ 8.5

Πίνακας λανθασμένης ταξινόμησης γα κάθε κατηγορία ζημιών αυτοκινήτου και φύλου υπό το μοντέλο $H(YY'EE', Y'E'CD, YECD)$

[ΠΗΓΗ : Chen et al. 1984]

Car Damage	**Low**							
Driver's sex	**Male**				**Female**			
np-injury	**No**		**Yes**		**No**		**Yes**	
np-belt use	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
p-not injured/unbelted	0.972	0.495	0.615	0.475	0.963	0.414	0.567	0.382
p-not injured/belted	0.014	0.486	0.012	0.303	0.019	0.563	0.015	0.337
p-injured/unbelted	0.013	0.011	0.369	0.088	0.015	0.011	0.412	0.086
p-injured/belted	0.002	0.007	0.004	0.134	0.003	0.011	0.006	0.194
Car Damage	**High**							
Driver's sex	**Male**				**Female**			
np-injury	**No**		**Yes**		**No**		**Yes**	
np-belt use	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
p-not injured/unbelted	0.934	0.403	0.321	0.287	0.941	0.390	0.380	0.283
p-not injured/belted	0.018	0.545	0.009	0.252	0.020	0.563	0.011	0.265
p-injured/unbelted	0.042	0.031	0.664	0.184	0.033	0.023	0.601	0.138
p-injured/belted	0.006	0.021	0.006	0.277	0.007	0.023	0.009	0.314

Εξετάζοντας τις προσαρμοσμένες υπό το επιλεγμένο μοντέλο τιμές (βλ. Παράρτημα, Πίνακες Α.1, Α2, Α3), διαπιστώνονται τα ακόλουθα : Η γενική τάση για τη μεροληψία που προέρχεται από τις αστυνομικές αναφορές είναι να υπερεκτιμά τις τιμές των κελιών των πινάκων συνάφειας (μη τραυματίες, δεν φορούν ζώνη). Όταν η πραγματική ταξινόμηση είναι (τραυματίες, φορούν ζώνη), η μεροληψία μειώνεται από τις συνοδευτικές μεροληψίες υπέρ των κελιών (τραυματίες, δεν φορούν ζώνη) και (μη τραυματίες, φορούν ζώνη). Αυτή η δομή λανθασμένης ταξινόμησης ισχύει για όλες τις κατηγορίες φύλου και ζημιών αυτοκινήτου. Ωστόσο, η μεροληψία ως προς τους μη τραυματίες δείχνει να είναι μεγαλύτερη στις κατηγορίες υψηλών ζημιών αυτοκινήτου. Επίσης, η μεροληψία προς τους χρήστες ζώνης μοιάζει να είναι μεγαλύτερη στις κατηγορίες των γυναικών και η μεροληψία προς τους μη χρήστες μοιάζει μεγαλύτερη στις κατηγορίες των ανδρών.

Υπό το μοντέλο λανθασμένης ταξινόμησης $H(YY'EE', Y'E'CD, YECD)$ και για να απλοποιηθεί ακόμα περισσότερο η δομή της λανθασμένης ταξινόμησης, προσαρμόζουμε διαδοχικά τα μοντέλα

$H^*(YEC, YCD, YED, ECD)$, $H^*(YE, YCD, ECD)$ και $H^*(YCD, ECD)$. Οι τιμές των LR-ελέγχων των μοντέλων δίνονται στον Πίνακα 8.4. Σύμφωνα με τις τιμές αυτές, και απορρίπτοντας το μοντέλο $H^*(YCD, ECD)$ λόγω της τιμής του υπό συνθήκη LR-ελέγχου, επιλέγουμε το μοντέλο $H^*(YE, YCD, ECD)$. Έτσι, καταλήγουμε στην επιλογή του μοντέλου $H(YY'EE', Y'E'CD, YECD)H^*(YE, YCD, ECD)$ ως το τελικό μας μοντέλο. Η επιλογή αυτού του μοντέλου σημαίνει ότι για κάθε κατηγορία ζημιών αυτοκινήτου και φύλου, ο σχετικός λόγος κινδύνου για τους τραυματισμούς κατά τη μη χρήση ζώνης σε σχέση με τη χρήση ζώνης είναι σταθερός για όλες τις κατηγορίες, με την καλύτερη εκτίμηση του να είναι 1.296. Ως εκ τούτου, συμφωνά με τα αποτελέσματα που προκύπτουν ο επιπλέον κίνδυνος που προέρχεται από τη μη χρήση ζώνης είναι σημαντικά μεγαλύτερος του 0 και πλησιάζει το 30%.

8.3 Μέθοδος διπλής δειγματοληψίας

Προτού προχωρήσουμε στην αντιμετώπιση της λανθασμένης ταξινόμησης μέσω της χρήσης διπλής δειγματοληψίας, ενδιαφέρον παρουσιάζει η ανάλυση για τη εύρεση κατάλληλου λογαριθμογραμμικού μοντέλου που να περιγράφει όσον το δυνατόν καλύτερα

και με την απλούστερη δυνατή δομή τα εξεταζόμενα δεδομένα για το κάθε δείγμα ξεχωριστά, πέραν του κορεσμένου μοντέλου της κάθε περίπτωσης. Βάσει αυτού, διακρίνουμε τις ακόλουθες περιπτώσεις

1. Ανάλυση δεδομένων του βασικού δείγματος

Σε αυτήν την περίπτωση εξετάζουμε τα δεδομένα που προέρχονται μόνο από τις αστυνομικές αναφορές. Το λογαριθμογραμμικό μοντέλο που αντιστοιχεί στο δείγμα n_2 των p-αναφορών είναι το $H(Y'E'CD)$. Θεωρώντας τα λογαριθμογραμμικά μοντέλα $H(Y'E'D, Y'CD, E'CD)$, $H(Y'E', Y'CD, E'CD)$, $H(Y'CD, E'CD)$, τα LR-test δίνονται στον Πίνακα 8.6.

ΠΙΝΑΚΑΣ 8.6

Λογαριθμογραμμικά μοντέλα LR-test, β.ε. και P-value για δεδομένα που προέρχονται μόνο από αστυνομικές αναφορές (p)

Models	L.R. G^2	d.f.
$H(Y'E'D, Y'CD, E'CD)$	3.39	2
$H(Y'E', Y'CD, E'CD)$	6.47	3
$H(Y'CD, E'CD)$	83.02	4

Σύμφωνα με τις τιμές του Πίνακα 8.6, το τελικά αποδεκτό μοντέλο είναι το $H(Y'E', Y'CD, E'CD)$. Το μοντέλο αυτό υποδεικνύει την ύπαρξη ενός σταθερού λόγου σχετικών πιθανοτήτων για τους τραυματισμούς και τη χρήση ζώνης δεσμεύοντας ως προς το φύλο και τη ζημία του αυτοκίνητου. Προσαρμόζοντας το μοντέλο, Ο Chen (1989) υπολογίζει ότι ο αυτός ο λόγος σχετικών πιθανοτήτων είναι 0.75, από όπου και συμπεραίνουμε ότι η χρήση ζώνης συμβάλλει στη μείωση των τραυματισμών των επιβατών σε σχέση με τους τραυματισμούς κατά τη μη χρήση ζώνης.

2. Ανάλυση δεδομένων δείγματος επικύρωσης

Σε αυτήν την περίπτωση εξετάζουμε τα δεδομένα που προέρχονται από τη διασταύρωση των αστυνομικών και μη-αστυνομικών αναφορών. Το λογαριθμογραμμικό μοντέλο που αντιστοιχεί στο δείγμα n των p-αναφορών και np-αναφορών είναι το $H(YY'E E'CD)$.

Εφόσον τα δεδομένα προέρχονται μόνο από αυτό το δείγμα μπορούμε να προβούμε σε σύμπτυξη του πίνακα $(YY'EE'CD)$ ως προς τις μεταβλητές Y' και E', αφού ο έλεγχος των μοντέλων για τον πίνακα $(YECD)$ είναι ανεξάρτητος των μοντέλων λανθασμένης ταξινόμησης (βλ. Chen (1989)). Εξετάζοντας λοιπόν τον πίνακα $(YECD)$ αντί του $(YY'EE'CD)$ και θεωρώντας τα ακόλουθα λογαριθμογραμμικά μοντέλα $H(YEC,YCD,YED,ECD)$, $H(YE,YCD,ECD)$, $H(YCD,ECD)$, τα LR-test δίνονται στον Πίνακα 8.7 .

ΠΙΝΑΚΑΣ 8.7

Λογαριθμογραμμικά μοντέλα LR-test, β.ε. και P-value για δεδομένα που προέρχονται από αστυνομικές (p) και μη αστυνομικές (np) αναφορές

Models	**L.R. G^2**	**d.f.**
$H(YEC,YCD,YED,ECD)$	0.24	1
$H(YE,YCD,ECD)$	1.24	3
$H(YCD,ECD)$	2.33	4

Σύμφωνα με τις τιμές του Πίνακα 8.7, την καλύτερη προσαρμογή έχει το μοντέλο $H(YCD,ECD)$. Θέλοντας όμως να προβούμε σε συγκρίσεις με το επιλεγμένο μοντέλο για το βασικό δείγμα, χρησιμοποιούμε το μοντέλο $H(YE,YCD,ECD)$. Το μοντέλο αυτό υποδεικνύει την ύπαρξη ενός σταθερού λόγου σχετικών πιθανοτήτων, δεσμεύοντας ως προς το φύλο και τη ζημία του αυτοκίνητου, για τους τραυματισμούς και τη χρήση ζώνης. Προσαρμόζοντας το μοντέλο, ο Chen (1989) υπολογίζει ότι ο αυτός ο λόγος σχετικών πιθανοτήτων είναι 0.84, που σημαίνει ότι η χρήση ζώνης συμβάλλει στη μείωση των τραυματισμών των επιβατών σε σχέση με τους τραυματισμούς κατά τη μη χρήση ζώνης. Θα πρέπει όμως να επισημάνουμε ότι η διαφορά του με τη μονάδα δεν είναι στατιστικά σημαντική.

Αναλύοντας το βασικό δείγμα και το δείγμα επικύρωσης μαζί, εφαρμόζουμε τη μέθοδο διόρθωσης της λανθασμένης ταξινόμησης με χρήση διπλής δειγματοληψίας. Η μέθοδος αυτή μπορεί να θεωρηθεί υποπερίπτωση της TSS, όταν εξετάζουμε μόνο τα σύνολα n και n_2 ατόμων που υπάρχουν διαθέσιμα. Ακολουθώντας παρόμοια βήματα με αυτά που έγιναν για

την TSS, τα προς προσαρμογή στο $H(YY'EE'CD)$ μοντέλα και τα αντίστοιχα LR-test τους δίνονται στον Πίνακα 8.8

ΠΙΝΑΚΑΣ 8.8

Λογαριθμογραμμικά μοντέλα LR-test, και β.ε. (ΙΙ)

Models		L.R. G^2	d.f.
$H(YY'EE'CD)$		12.02	15
$H(YY'EE'D,YECD)$		41.60	39
$H(YY'EE',Y'EE'D,YECD)$		44.20	45
$H(YY'EE',Y'E'CD,YECD)$		42.21	49
$H(YY'EE',E'C,Y'D,YECD)$		55.85	50
$H(YY'EE',Y'D,YECD)$		58.32	51
$H(YY'EE',YECD)$		143.02	52
$H(YY'EE',Y'D,YECD)$	$H^*(YEC,YCD,YED,ECD)$	61.41	51
$H(YY'EE',Y'D,YECD)$	$H^*(YE,YCD,ECD)$	62.64	53
$H(YY'EE',Y'D,YECD)$	$H^*(YCD,ECD)$	66.14	54

Σύμφωνα με τα παραπάνω, το μοντέλο $H(YY'EE',Y'D,YECD)$ επιλέγεται ως μοντέλο λανθασμένης ταξινόμησης. Προσαρμόζοντας στη συνέχεια σε αυτό τα μοντέλα $H^*(YEC,YCD,YED,ECD)$, $H^*(YE,YCD,ECD)$ και $H^*(YCD,ECD)$, καταλήγουμε στην επιλογή του $H(YY'EE',Y'D,YECD)H^*(YCD,ECD)$ ως τελικό μοντέλο λανθασμένης ταξινόμησης. Προσαρμόζοντας το μοντέλο, ο Chen (1989) υπολογίζει ότι ο λόγος σχετικών πιθανοτήτων για τους τραυματισμούς και τη χρήση ζώνης δεσμεύοντας ως προς το φύλο και τη ζημία του αυτοκίνητου είναι 0.75. Άρα βάσει του επιλεγμένο μοντέλου, η χρήση ζώνης συμβάλλει στη μείωση των τραυματισμών των επιβατών σε σχέση με τους τραυματισμούς κατά τη μη χρήση ζώνης. Θα πρέπει όμως να επισημάνουμε ότι η διαφορά του με τη μονάδα δεν είναι στατιστικά σημαντική.

Τα συμπεράσματα που προκύπτουν από την επιλογή του συγκεκριμένου μοντέλου είναι τα ακόλουθα

- ✓ η λανθασμένη ταξινόμηση δεν σχετίζεται με το φύλο

- ✓ το χαμηλό επίπεδο των ζημιών αυτοκινήτου θα προκαταλάβει τις αστυνομικές αναφορές προς τη μη ύπαρξη τραυματισμών
- ✓ η ύπαρξη τραυματισμών και η χρήση ζώνης είναι υπό συνθήκη ανεξάρτητες δεδομένων των ζημιών του αυτοκινήτου και του φύλου.

Τέλος, αξίζει να σημειωθεί ότι εν συγκρίσει με τα συμπεράσματα στα οποία προβήκαμε μέσω της εφαρμογής της μεθόδου τριπλής δειγματοληψίας, η χρήση διπλού δειγματοληπτικού σχήματος στα συγκεκριμένα δεδομένα παρουσιάζει ένα μειονέκτημα : δεν μπορεί να υποστηρίξει την υπόθεση ότι η χρήση ζώνης μπορεί να "αναγνώσει" με αποτελεσματικό τρόπο την πρόκληση ή όχι τραυματισμών των επιβαινόντων στα αυτοκίνητα κατά την πρόκληση ενός ατυχήματος.

ΠΑΡΑΡΤΗΜΑ

Α. Πίνακες προσαρμοσμένων τιμών υπό το μοντέλο των Chen et. al (1984)

ΠΙΝΑΚΑΣ Α.1

Συχνότητες ατυχημάτων ως προς τη ζημιά των αυτοκινήτων (D), το φύλο των οδηγών (C) ,τη χρήση ζώνης (Ε'), και τους τραυματισμούς (Υ'), με βάση 80,084 (n_2) αστυνομικές αναφορές (p) (με τις προσαρμοσμένες υπό το τελικό μοντέλο τιμές κάτω από τις παρατηρούμενες τιμές)
[ΠΗΓΗ : Chen et. al (1984)]

Car damage	**Low**				**High**			
Driver's sex	**Male**		**Female**		**Male**		**Female**	
p - belt use	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
p - not injured	22536	3006	11199	1262	17476	2155	6964	728
	22552.2	*2998.7*	*11213.3*	*1249.1*	*17462.3*	*2147.8*	*6977.3*	*723.4*
p - injured	1687	199	1422	117	6746	583	3707	297
	1694.7	*198.4*	*1424.6*	*118.7*	*6731.7*	*581.2*	*3712.8*	*297.4*

ΠΙΝΑΚΑΣ Α.2

Συχνότητες ατυχημάτων ως προς τη ζημιά των αυτοκινήτων (D), το φύλο των οδηγών (C) και ως προς την ταξινόμηση από αστυνομικές (p) και μη αστυνομικές αναφορές για τη χρήση ζώνης και τους τραυματισμούς με βάση 1796 (n) ατυχήματα (με τις προσαρμοσμένες υπό το τελικό μοντέλο τιμές κάτω από τις παρατηρούμενες τιμές)
[ΠΗΓΗ : Chen et. al (1984)]

Driver's sex	**Male**				**Female**			
np-injury	**No**		**Yes**		**No**		**Yes**	
np-belt use	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
Car damage: low								
p-not injured/unbelted	407	62	45	7	206	18	37	5
	389.8	*56.8*	*50.7*	*8.5*	*196.4*	*14.3*	*36.9*	*3.3*
p-not injured/belted	6	47	1	6	1	17	0	1
	5.6	*55.2*	*1.0*	*5.5*	*4.0*	*19.9*	*1.0*	*3.1*
p-injured/unbelted	5	1	32	4	4	0	29	0
	5.1	*1.3*	*30.1*	*1.6*	*3.2*	*0.4*	*27.6*	*0.8*
p-injured/belted	0	1	1	2	3	0	1	0
	0.8	*0.9*	*0.3*	*2.5*	*0.6*	*0.3*	*0.3*	*1.4*
Car damage: high								
p-not injured/unbelted	299	20	59	9	102	7	53	4
	294.6	*25.1*	*63.4*	*8.4*	*97.5*	*6.0*	*48.7*	*4.2*
p-not injured/belted	4	30	1	6	2	6	1	3
	5.7	*33.3*	*1.7*	*7.5*	*2.1*	*8.7*	*1.4*	*4.0*
p-injured/unbelted	11	2	118	5	5	1	79	1
	13.2	*1.9*	*130.4*	*5.4*	*3.4*	*0.4*	*77.5*	*2.1*
p-injured/belted	1	2	0	9	0	0	1	6
	1.9	*1.4*	*1.3*	*8.4*	*0.7*	*0.4*	*1.1*	*4.5*

ΠΙΝΑΚΑΣ Α.3

Συχνότητες ατυχημάτων ως προς τη ζημιά των αυτοκινήτων (D), το φύλο των οδηγών (C) ,τη χρήση ζώνης (E), και τους τραυματισμούς (Y), με βάση 905 (n_1) μη αστυνομικές αναφορές (np) (με τις προσαρμοσμένες υπό το τελικό μοντέλο τιμές κάτω από τις παρατηρούμενες τιμές)

[ΠΗΓΗ : Chen et al. 1984]

Car damage	**Low**				**High**			
Driver's sex	**Male**		**Female**		**Male**		**Female**	
np - belt use	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
np - not injured	200	55	100	13	150	30	54	8
	202.2	*57.5*	*103.2*	*17.6*	*159.0*	*31.1*	*52.3*	*7.8*
np - injured	50	10	40	3	100	15	70	7
	41.4	*9.1*	*33.2*	*4.4*	*99.2*	*15.0*	*64.9*	*7.4*

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ξένη

Assakul, K. & Proctor, C.H. (1967). Testing independence in two-way contingency tables with data subject to misclassification, *Psychometrika* **32**, pp. 67–76.

Bailar, B.A. (1968). Recent research in reinterview procedures, Journal of the American Statistical Association **63,** No. 321, pp. 41–63.

Barron, B.A. (1977). The effects of misclassification on the estimation of relative risk, *Biometrics* **33**, pp. 414-418

Bauman K.E. and Koch G.G. (1983). *Validity of self-reports and descriptive and analytical conclusions: the case of cigarette smoking by adolescents and their mothers*, American Journal of Epidemiology **118** (1983), no. 1, pp. 90–98.

Baumgarten M., Siemiatycki J. and Gibbs G.W. (1983). Validity of work histories obtained by interview for epidemiologic purposes. *Am J Epidemiol.* **118**, pp. 583–591.

Begg, C.B. (1984), "Estimation of Risks When Verification of Disease Status Is Obtained,in a Selected Group of Subjects," *American Journal of Epidemiology*, 120, pp.328-330

Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association*, Vol. **45**, No. 250, pp. 164-180

Birkett, N.J. (1992). Effects of nondifferential misclassification on estimates of odds ratios with multiple levels of exposure, *American Journal of Epidemiology* **136**, pp. 356–362.

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

Blettner, M., and Wahrendod, I. (1984), "What Does an Observed Relative Risk Convey about Possible Misclassification," *Methods of Information in Medicine*, 23, pp.37-40.

Brenner, N. (1993), "Bias Due to Non-differential Misclassification of Polytomous Confounders," *Journal of Clinical Epidemiology*, 46, pp.57-63.

Bross, I. (1954). Misclassification in 2 X 2 tables, *Biometrics* **10**, pp. 488–495.

Buell, P., and Dunn, I.E. (1964), "The Dilution Effect of Misclassification," *American Journal of Public Health*, 54, pp.598-602.

Buonaccorsi, J.P. (2009). Measurement Error: Models, Methods and Applications. *Chapman & Hall.*

Carroll, R.I. (1992), 'Approaches to Estimation with Error in Predictors," in L. Fahrmeir, B. Francis, R. Gi1christ, and G. Tutz (eds.), *Advαnces in GLIM αnd Stαtistical Modelling* (Proceedings of the GLIM92 Conference and the 7th InternationaI Workshop on Statistical Modelling, Miinchen, July 1992), New York: Springer- Verlag, pp.40-47.

Chavance M, Dellatolas G, Lellouch J. (1992). Correlated nondifferential misclassifications of disease and exposure: application to a cross-sectional study of the relation between handedness and immune disorders. *Int J Epidemiol* , **21**, pp.537–46.

Chen, T., and Fienberg, S.E. (1974), "Two-dimensional Contingency Tables with Both Completely and Partially Cross-classified Data," *Biometrics.* 30, ρρ. 629-642.

Chen, T., and Fienberg, S.E. (1976), "The Analysis of Contingency Tables with Incompletely Classified Data," *Biometrics,* **32**, pp. 133-144.

Chen, T.T. (1978), "Log-linear Models for the Categorical Data Obtained from Randomized Response Techniques," *Proceedings of the Section on Socιal Statistics,* American Statistical Association, pp.284-288.

Chen, T.T. (1979a), "Analysis of Randomized Response as Purposively Misclassified Data," *Proceedings ofthe Section on Survey Reseαrch Methods,* American Statistical Association, pp.158-163.

Chen, T.T. (1979b), "Log-linear Models for Categorical Data with Misclassification and Double Sampling," *Journαl of the Americαn StatisticαI Associαtion,* 74, pp.481-488.

Chen, T.T., Hochberg, Y. & Tenenbein, A. (1984). Analysis of multivariate categorical data with misclassification errors by triple sampling schemes, *Journal of Statistical Planning and Inference* **9**, pp. 177–184.

Chen, T.T. (1989), "A Review of Methods for Misclassified Categorical Data in Epidemiology," *Stαtistics in Medicine,* 8, pp.1095-1106.

Chen, T.T. (1992), Reply to Ekholm, *Stαtistics in Medicine,* **11**, ρρ. 272-275.

Cheng KF, Hsueh HM and Chien TH (1998). Goodness of fit tests with misclassified data. Communications in Statistics, Theory and. Methods, 27, 1379–93.

Chiacchierini, R. P., and Arnold, J.C. (1977), 'A Two-sample Test for Independence in 2 x 2 Contingency Tables with Both Margins Subject to Misclassification," *Journal of the American StatisticaI Association,* 72, pp.170-174.

Chua T., Fuller W.A. (1987). A model for multinomial response errors applied to labor flows, *Journal of the American Statistical Association* 82, pp. 46–51

Chyou PH. (2007). Patterns of bias due to differential misclassification by case-control status in a case-control study, Eur J Epidemiol. 22(1):7-17.

Clayton, D. (1985), "Using Test-Retest Reliability Data to Irnprove Estimates of ReIative Risk: An Application of Latent CIass Analysis," *Statistics in Medicine,* 4, pp.445-455.

Cobb, S. and Rosenbaum, J.(1965). A comparison of specific symptom data obtained by nonmedical interviewers and by physicians, *J. chron. Dis*. **4**, pp. 245-252

Cochran, W.G. (1952). The $\chi 2$ test of goodness-of-fit. Ann. math. Statist. 23, pp. 315–345.

Cochran, W.G. (1968). Errors of measurement in statistics, *Technometrics* **10**, pp. 637–666.

Copeland, K.T., Checkoway, H., McMichael, A.J. & Holbrook, R. (1977). Bias due to misclassification in the estimation of relative risk, *American Journal of Epidemiology* **105**, pp. 488–495.

Cramér, H. (1946). *Mathematical* Methods of Statistics, Princeton University Press.

Dalenius, T. (1977). Bibliography of non-sampling errors in surveys, *International Statistical Review* **45**, pp. 71–89, 181–197, 303–317.

Dawid, A.P., and Skene, A.M. (1979), "Maximum Likelihood Estimation of Observer Error-rates Using the EM Algorithm," *Applied Statistics*, 28, pp.20-28.

Deming, W.E. (1950), *Some Theory OI Sampling*, New York: Wiley.

Deming, W.E. (1977), "An Essay on Screening, or on Two-phase Sampling, Applied to Surveys of a Community," *International Statistical Review*, 45, pp.29-37.

Dempster, A.P., Laird, N.W., and Rubin, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm" (with discussion), *Journαl OI the Royal Statistical Society*, Series B, 39, pp.1-38.

Diamond, E.L., and Lilienfeld, A.M. (1962a), 'Effects of Errors in Classification and Diagnosis in Various Types of Epidemiological Studies," *Americαn Journal OI Public Health*, 52, pp.1137-1144.

Diamond, E.L., and Lilienfeld, A.M. (1962b), "Misclassification Errors in 2 X 2 TabIes with One Margin Fixed: Some Further Comments," *American Journal OI Public Health*, 52, pp.2106-2110.

Dosemeci, M., Wacholder, S., and Lubin, J.H. (1990), "Does Nondifferential Misclassifi-cation of Exposure Always Bias a True Effect toward the Null Value?" *Americαn Journal OI Epidemiology*, 132, pp.746-748.

Duffy, S.W., Rohan, T.E. & Day, N.E. (1989). Misclassification in more than one factor in a case-control study: A combination of Mantel-Haenszel and maximum likelihood approaches, *Statistics in Medicine* **8**, pp. 1529–1536.

Duffy, S.W., Maximowitch, D.M., and Day, N.E. (1992), "External Validation, Repeat Determination, and Precision of Risk Estimation in Misclassified Exposure Data in Epidemiology," *Journal OI Epidemiology and Community Heαlth*, 46, pp.620-624.

Ekholm, A., and Palmgren, J. (1982), 'A Model for a Binary Response with Misclassifica-tions," in R. Gilchrist (ed.), *GLIM* 82: *Proceedings of the Internαtionαl Conference on*

Generalized Linear Models, Heidelberg: Springer, pp.128-143.

Ekholm, A. & Palmgren, J. (1987). Correction for misclassification using doubly sampled data, *Journal of Official Statistics* **3**, pp. 419–429.

Ekholm, A. (1991), "Algorithms Versus Models for Analyzing Data that Contain Misclassification Errors," *Biometrics,* 47, pp.1171-1182.

Ekholm, A. (1992), Letter to the Editor concerning the paper by Chen, *Statistics in Medicine,* 11, pp.271-272.

Elton, R.A., and Duffy, S.W. (1983), "Correcting for the Effect of Misclassification Bias in a Case-Control Study Using Data from Two Different Questionnaires", *Biometrics,* 39, pp.659-665.

Espeland M., and Odoroff C. (1984a). "*Algorithms* for *Computing* Maximum Likelihood Estimates From Incomplete Discrete Data," Technical Report 01/84, University of Rochester, Statistics Dept.

Espeland, M.A. & Odoroff, C.L. (1985). Log-linear models for doubly sampled categorical data fitted by the EM algorithm, *Journal of the American Statistical Association* **80**, pp. 663–670.

Espeland, M.A. & Hui, S.L. (1987). A general approach to analyzing epidemiologic data that contain misclassification errors, *Biometrics* **43**, pp. 1001–1012.

Espeland, MA, and Handelman, S.L. (1989), "Using Latent CIass Models to Characterize and Assess Relative Error in Discrete Measurements," *Biometrics,* 45, pp.587-599.

Flegal, K.M., Brownie, C., and Haas, J.D. (1986), "The Effects of Exposure Misclassification on Estimates of Relative Risk," *American Journal oı Epidemiology,* 123, pp.736-751.

Flegal, K.M., Keyl, P.M. & Nieto, F.J. (1991). Differential misclassification arising from nondifferential errors in exposure measurement, *American Journal of Epidemiology* **134**, pp. 1233–1244.

Fleiss, J.L. (1981), *StatisticaI Methods for Rates and Proportions,* New York: WiIey.

Forsman, G. and Schreiner, I. (1991), *The design and analysis of reinterview: an overview,* Measurement Error in Surveys (P. Biemer, R.M. Groves, L.E. Lyberg, N.A.Mathiewetz, and S. Sudman, eds.),JohnWiley & Sons, New York, pp. 279–302.

Fung, K.Y. & Howe, G.R. (1984). Methodological issues in case-control studies. III: The effect of joint misclassification of risk factors and confounding factors upon estimation and power, *International Journal of Epidemiology* **13**, pp. 366–370.

Giesbrecht, F.G. (1967). Classification Errors and Measures of Association in Contingency Tables, *Proceedings of the Social Statistics Section, American Statistical* Association, pp. pp. 271-273.

Gladen, B. & Rogan, W.J. (1979). Misclassification and the design of environmental studies, *American Journal of Epidemiology* **109**, pp. 607–616.

Goldberg, J.D. (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table, *Journal of the American Statistical Association* **70**, pp. 561–567.

Goodman, I.A. (1973), "The Analysis of Multidimensional Contingency Tables When Some Variables Are Posterior to Others: A Modified Path Analysis Approach," *Biometrika*, 60, pp.179- 192.

Goodman, I.A. (1978), *Analyzing Qualitatlve Categorical Data: Log-Linear Models and Latent-Structure Analysis*, Reading, MA: Addison-Wesley.

Green, M.S. (1983), "Use of Predictive Value to Adjust Relative Risk Estimates Biased by Misclassification of Outcome Status," *American Journal of Epidemiology,* 117, pp.98-105.

Greenland, S. (1980), "The Effect of Misclassification in the Presence of Covariates," *American Journal of Epidemiology,* 112, pp.564-569.

Greenland, S. (1982), "The Effect of Misclassificatioıı in Matched-Pair Case-Control Studies," *American Journal of Epidemiology,* 120,pp. 643-648.

Greenland, S. (1988a), "Variance Estimation for Epidemiologic Effect Estimates Under Misclassification," *Statistics in Medicine,* 7, pp.745-757.

Greenland, S. (1988b), "Statistical Uncertainty Due to Misc1assification: Implications for Validation Substudies," *Journal OI Clinical Epidemiology,* 41, pp. 1167-1174

Greenland, S. (1989), "On Correcting for Misclassification in Twin Studies and Other Matched-Pair Studies," *Statistics in Medicine,* 8, pp.825-829.

Greenland, S., and Kleinbaum, D.G. (1983), "Correcting for Misclassification in Two-Way Tables and Matched-Pair Studies," *International Journal OI Epidemiology,*12, pp.93-97.

Greenland, S., and Robins, J.M. (1985), "Confounding and Misclassification," *American Journal OI Epidemiology,* 122, pp.49?-506.

Greenland S., Gustafson P. (2006). Accounting for independent nondifferential misclassification does not increase certainty that an observed association is in the correct direction. *Am J Epidemiology* 2006 , **164:**63-68.

Greenland, S. (2007). Maximum-likelihood and closed-form estimators of epidemiologic measures under misclassification. *J Stat Plan Inference* **138**, pp. 528–38

Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969), "Analysis of Categorical Data by Linear Models," *Biometrics,* 25, pp.489-504

Gullen, W.H., Bearman, J.E. & Johnson, E.A. (1968) Effects of misclassification in epidemiologic studies, *Public Health Reports* **83**, pp. 914–918.

Gullen, W. N., Bearman, J. K, and Johnson, E.A. (1968), "Effects of Misclassification in Epidemiological Studies," *Public Health Reports*, 83, pp.914-918

Gustafson P. (2004) Measurement error and misclassification in statistics and epidemiology*:* impacts and Bayesian adjustments. *New York: Chapman & Hall*

Hagenaars, J.A. (1990). Categorical Longitudinal Data – Loglinear Analysis of Panel, Trend and Cohort Data. Newbury Park: Sage.

Hagenaars, J. A. (1993). *Loglinear models with latent variables*. Sage Publications.

Haitovsky, Y. & Rapp, J. (1992). Conditional resampling for misclassified multinomial data with applications to sampling inspection, *Technometrics* **34**, pp. 473–483.

Harper, D. (1964). Misclassification in epidemiological surveys, *American Journal of Public Health*, **54**, pp. 1882–1886.

Hochberg, Y. (1977). On the Use of Double Sampling Schemes in Analyzing Categorical Data *with* Misclassification Errors, *J. Am. Stat. Association* **72**, pp. 914–921

Hochberg Y. and Reinfurt D.W. (1975). An investigation of safety belt usage and effectiveness, A technical report by NHTSA, Dept. of Transportation, Washington, *Project (DOT-HS-801-805)*

Jurek A.M., Greenland S., Maldonado G. and Church T.R. (2005). Proper interpretation of non-differential misclassification effects: Expectations vs observations. *International Journal of Epidemiology* , **34**, 680-687.

Jurek A, Maldonado GM, Greenland S. (2008). How far from nondifferential exposure or disease does misclassification have to be to bias measures of association away from the null? *Int J Epidemiol* , **37**:382-385.

Kaldor, J. & Clayton, D. (1985). Latent class analysis in chronic disease epidemiology, *Statistics in Medicine* **4**, pp. 327–335.

Keys, A. and Khilberg, J.K. (1963). Effects of Misclassification on Estimated Relative Prevalence of Characteristics. *American Journal of Public Health*, **53,** pp. 1956-1965.

King, G. and Lu Ying. (2008). Verbal Autopsy Methods with Multiple Causes of Death, *Statistical Science* **23(1)**, pp. 78–91.

Kleinbaum D., Kupper L., and Morgenstern H. (1982). Epidemiologic Research: Principles and Quantitative Methods. Belmont, California: Lifetime Learning.

Korn, E.L. (1981). Hierarchical log-linear models not preserved by classification error, *Journal of the American Statistical Association* **76**, pp. 110–113.

Korn, E.L. (1982). The asymptotic efficiency of tests using misclassified data in contingency tables, *Biometrics* **38**, pp. 445–450.

Kristensen P. (1992). Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology* , **3**, pp. 210–15.

Kuha, J. & Skinner, C. (1997). Categorical data analysis and misclassification, in *Survey Measurement and Process Quality*, L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz & D. Trewin, eds. Wiley, New York, pp. 633–670.

Kuha, J., C. Skinner, and J. Palmgren (1998). Misclassification error. In P. Armitage and T. Colton (Eds.), Encyclopedia of Biostatistics, Volume 4, pp. 2615–2621. New York:Wiley. Reprinted in M. Gail and J. Benichou (eds.), Encyclopedia of Epidemiologic Methods, Wiley,2000.

Kuroda, M. & Geng, Z. (2002). Bayesian inference for categorical data with misclassification errors, ADVANCES IN STATISTICS, COMBINATORICS AND RELATED AREAS (World Scientific Publishing）

Lessler, J., & Kalsbeek, W. (1992). *Nonsampling error in surveys*. New York, NY: John Wiley & Sons.

Liu, X. & Liang, K.-Y. (1991). Adjustment for nondifferential misclassification error in the generalized linear model, *Statistics in Medicine* **10**, pp. 1197–1211.

Lyles, R.H., 2002. A note on estimating crude odds ratios in case–control studies with differentially misclassified exposure. *Biometrics* **58**,1034–1037.

Marshall, J.R., Priore, R., Graham, S. & Brasure, J. (1981). On the distortion of risk estimates in multiple exposure level case-control studies, *American Journal of Epidemiology* ,**113**, pp. 464–473.

Marshall, J.R., and Graham, S. (1984), "Use of Dual Responses to Increase Validity of Case-Control Studies," *Journal of Chronic Studies*, 37, pp.125-136.

Marshall, R.J., 1990. Validation study methods for estimating exposure proportions and odds ratios with misclassified data. *J. Clin. Epidemiol*. **43**,941–947.

McClish, D., and Quade, D. (1985), "Improving Estimates of Prevalence by Repeated Testing," *Biometrics*, 41, pp.81-89.

Mitra, S.K. (1958). On the limiting power function of the frequency chi-square test. Ann. Math. Stat. 29,1221–1233

Morrissey, M.J., Spiegelman, D., 1999. Matrix methods for estimating odds ratios with misclassified exposure data. *Biometrics* **55**, 338–344.

Mote, V.L. & Anderson, R.L. (1965). An investigation of the effect of misclassification on the properties of $\chi 2$- tests in the analysis of categorical data, *Biometrika* **52**, pp. 95–109.

Palmgren, J. (1987), "Precision of Double Sampling Estimators for Comparing Two Probabilities," *Biometrikα*, 74, pp.687-694.

Palmgren, J., and Ekholm, A. (1987), "Exponential Family Non-Linear Models for Categorical Data with Errors of Observation," *Applied Stochαstic Models αnd Dαtα Anαlysis*, 3, pp.111-124.

Pardo, L., Zografos K., (2000). Goodness of fit tests with misclassified data based *on* φ-divergences. *Biometrical Journal*, **42**, 223—237.

Pitman, E.J.G. (1948), Notes on nonparametric statistical inference, unpublished notes.

Plackett, R. C. (1974). The Analysis of Categorical Data. Griffin and Hall, London, pp. 75-76.

Rogan, W. J. and Gladen. B. (1978). Estimating prevalence from the results of a screening test. *American Journal of Epidemiology* **107**, pp. 71–76.

Rogot, E. (1961). A Note on Measurement Errors and Detecting Real Differences, *Journal of the American Statistical Association*, **56**, pp. 314-319.

Rothman K. J. and Greenland S. (1998). Modern Epidemiology. Philadelphia: Lippincott-Raven.

Rubin, T., Rosenbaum, A.B. & Cobb, S. (1956). The use of interview data for the detection of association in field studies, *Journal of Chronic Diseases* **4**, pp. 253–266.

Savitz, D.A. & Barron, A.E. (1989). Estimating and correcting for confounder misclassification, *American Journal of Epidemiology* **129**, pp. 1062–1071.

Selen, I. (1986), 'Adjusting for Errors in CIassification and Measurement in the Analysis of Partly and Purely Categorical Data," *Journαl of the Americαn StαtisticαlAssociαtion*, 81, pp.75-81

Schlesselman, J. (1982). Case-Control Studies: Design, Conduct, Analysis. New York: Oxford University Press.

Schwartz, J.E. (1985), The Neglected Problem of Measurement Error in Categorical Data, *Sociological Methods and Research*, **13**, pp. 435-466.

Swires-Hennessy, E. & Thomas, G.: *The good, bad and the ugly: multiple stratified sampling in the 1986 Welsh House Condition Survey*, Statistical News no 79, HMSO November 1987

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassification. *J. Amer. Statist. Assoc.* **65**, pp. 1350–1361.

Tenenbein, A. (1971). A Double Sampling Scheme for Estimating From Binomial Data with Misclassifications; Sample Size Determination, *Biometrics* **27**, pp. 935–944

Tenenbein, A. (1972). A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection, *Technometrics* **14**, pp. 187–202.

Tzavidis N, Lin Y-X. (2006) Estimating from cross-sectional categorical data subject to misclassification and double sampling*:* Moment-based, maximum likelihood and quasi-likelihood approaches*. Journal of Applied Mathematics and Decision*, Sciences 2006(1):1–13. 20.

Wachholder, S., Armstrong, B. & Hartge, P. (1993). Validation studies using an alloyed gold standard, *American Journal of Epidemiology* **137**, pp. 1251–1258.

Wachholder, S., Dosemeci, M. & Lubin, J.H. (1991). Blind assignment of exposure does not always prevent differential misclassification, *American Journal of Epidemiology* **134**, pp. 433–437.

Walter, S.D. & Irwig, L.M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review, *Journal of Clinical Epidemiology* **41**, pp. 923–937.

Warner, S.L. (1965), "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association,* 60, pp.63-69.

White, E. (1986), "The Effects of Misclassification of Disease Status in Follow-up Studies: Implication for Selecting Disease Classification Criteria," *American Journal of Epidemiology,* 124, pp.816-825.

Whittaker, J.W. (1990), *Graphical Models in Applied Multivariate Statistics,* Chichester: Wiley.

Whittemore, A.S., and Grosser, S. (1986), "Regression Methods for Data with Incomplete Covariates," in S.H. Moolgavkar, and R.L. Prentice (eds.), *Modern Statistical Methods in Chronic Disease Epidemiology,* New York: Wiley.

Zelen, M., and Haitovsky, U. (1991), "Testing Hypotheses with Binary Data Subject to Misclassification Errors: Analysis and Experimental Design," *Biometrika,* 78, pp.857-865.