ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ ΓΙΑ ΑΡΑΙΟΥΣ ΠΙΝΑΚΕΣ ΣΥΝΑΦΕΙΑΣ

Γεώργιος Α. Μακρυνιώτης

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
Ιούνιος 2004

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ ΓΙΑ ΑΡΑΙΟΥΣ ΠΙΝΑΚΕΣ ΣΥΝΑΦΕΙΑΣ

Γεώργιος Α. Μακρυνιώτης

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
Ιούνιος 2004

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:

- Μ.Κατέρη (Επιβλέπων)
- Γ.Ηλιόπουλος
- Τ.Παπαϊωάννου

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

INFERENCE FOR SPARSE CONTINGENCY TABLES

By

George A. Makriniotis

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
June 2004

Στους γονείς μου

Αναστάσιο και Δέσποινα

Ε Υ Χ Α Ρ Ι Σ Τ Ι Ε Σ

Στο σημείο αυτό θα ήθελα να εκφράσω τις θερμές ευχαριστίες μου στην κα. Κατέρη Μαρία, επ. Καθηγήτρια του τμήματος Στατιστικής και Ασφαλιστικής επιστήμης του Πανεπιστημίου Πειραιά, για την απλόχερη υποστήριξη, ενθάρρυνση και βοήθεια που μου πρόσφερε καθόλη την διάρκεια υλοποίησης της παρούσας εργασίας.

Ιδιαίτερες ευχαριστίες οφείλω στον κ. Ηλιόπουλο Γεώργιο, επ. Καθηγητή του τμήματος Στατιστικής και Ασφαλιστικής επιστήμης του Πανεπιστημίου Πειραιά, για την πολύτιμη συνεισφορά του στην κατανόηση εννοιών σχετικά με τις MCMC μεθόδους, και στον κ. Παπαιωάννου Τάκη, Καθηγητή του τμήματος Στατιστικής και Ασφαλιστικής επιστήμης του Πανεπιστημίου Πειραιά, για την συμμετοχή του στην τριμελή εξεταστική επιτροπή.

Τέλος ευχαριστώ τους φίλους και συμφοιτητές μου Αναστασία Ελευθεράκη και Σταυρούλα Πουλοπούλου για τις εύστοχες παρατηρήσεις τους, καθώς και την φίλη μου Ιουλία Χιωτάκη για την υπομονή της στην μετάφραση ξένων ορολογιών και κειμένων.

ΜΑΚΡΥΝΙΩΤΗΣ ΓΙΩΡΓΟΣ

Ν. ΦΙΛΑΔΕΛΦΕΙΑ

ΙΟΥΝΙΟΣ 2004

Π Ε Ρ Ι Λ Η Ψ Η

Τα τελευταία χρόνια το ενδιαφέρον για την ανάλυση αραιών πινάκων έχει αυξηθεί. Η αντιμετώπιση τους με την κλασσική ασυμπτωτική προσέγγιση είναι ακατάλληλη. Ως εναλλακτική προσέγγιση προτείνεται η χρήση ακριβών (*exact*) μεθόδων. Στην εργασία παρουσιάζουμε την θεωρητική και υπολογιστική εξέλιξη που γνώρισαν οι ακριβείς μέθοδοι.

Κυρίως εστιάζουμε το ενδιαφέρον μας στην ακριβή προσέγγιση με δέσμευση (*exact conditional*), καλύπτοντας τις περιπτώσεις ελέγχου ανεξαρτησίας και δεσμευμένης ανεξαρτησίας των μεταβλητών ταξινόμησης σε διδιάστατους και τριδιάστατους αραιούς πίνακες συνάφειας. Επιπλέον, κάνουμε μια συνοπτική αναφορά στις ακριβείς μεθόδους για την ανάλυση τετραγωνικών πινάκων όταν ελέγχουμε ψευδοανεξαρτησία (*quasi-independence*), συμμετρία (*symmetry*) και ψευδοσυμμετρία (*quasi-symmetry*), ενώ πραγματοποιούμε και μια κριτική ανασκόπηση των ακριβών μεθόδων με δέσμευση και χωρίς δέσμευση (*exact conditional* vs *exact unconditional*).

Για πολλά χρόνια η ευρεία χρήση των ακριβών μεθόδων παρεμποδιζόταν από την έλλειψη κατάλληλων υπολογιστικών προγραμμάτων. Με την ραγδαία όμως αύξηση της ισχύος των ηλεκτρονικών υπολογιστών δόθηκε τεράστια ώθηση και στην εφαρμογή των ακριβών μεθόδων. Ακόμα όμως και με την σύγχρονη υπολογιστική ισχύ, η υλοποίηση των ακριβών μεθόδων συχνά είναι ανέφικτη.

Στην εργασία περιγράφουμε δύο αντιπροσωπευτικές εναλλακτικές μεθόδους, από τις τελευταίες και πιο σύγχρονες εξελίξεις στον τομέα της ακριβούς συμπερασματολογίας. Πρόκειται για προσεγγιστική υλοποίηση ακριβών δεσμευμένων τέστ, με χρήση των τεχνικών Importance Sampling και Markov Chain Monte Carlo μεθόδων (Gibbs Sampling). Στο Παράρτημα παραθέτουμε τους αντίστοιχους αλγορίθμους των προσεγγίσεων αυτών, τους οποίους υλοποιήσαμε στις γλώσσες προγραμματισμού Mathematica και Fortran, αντίστοιχα.

ABSTRACT

Lately, the interest on the analysis of sparse contingency tables has increased. The confrontation of those tables by the asympotic approach is inadequate. As an alternative approach the use of exact methods is suggested. In this project the theoretical and computational progress which the exact methods met, is reviewed.

The interest is mostly focused on the exact conditional approach, covering the cases of testing independence and conditional independence of the classification variables in two dimensional and three dimensional sparse contingency tables. In addition, a concise report is made about the exact methods of the analysis of square tables when we test quasi-independence, symmetry and quasi-symmetry, while we make a critical survey of the exact conditional methods in cotroversy with exact unconditional methods.

For many years the wide use of the exact methods was hindered by the lack of suitable software. However, because of the rapid increasement of computing power, there was a huge thrust at the enforcement of the exact methods. Even though the help of the modern computing power, the realization of the exact methods is usually infeasible.

In this project two representative alternative methods are illustrated by the latest and most modern evolutions of exact inference. The first approach provides exact conditional tests using Importance Sampling, while the second using Gibbs Sampling. Finally the corresponding algorithms of the two approaches are implemented through Mathematica and Fortran respectively, and are provided in Appendices.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1

Αραιοί πίνακες συνάφειας

1.1 Εισαγωγή

Έστω ένας *I*x*J* πίνακας συνάφειας με μεταβλητές ταξινόμησης X και Y με επίπεδα 1,…,*I* και 1,…,*J* αντίστοιχα. Έστω y_{ij} η παρατηρούμενη συχνότητα στο κελλί (i, j), μ_{ij} η αναμενόμενη συχνότητα και $\hat{\mu}_{ij}$ η εκτίμηση της αναμενόμενης συχνότητας του κελλιού (i, j) κάτω από το θεωρούμενο μοντέλο. Για τους τριδιάστατους πίνακες συνάφειας *I*x*J*x*K* η τρίτη μεταβλητή ταξινόμησης είναι η *Z* με επίπεδα 1,…,*K* , και θα υποδηλώνεται από τον τρίτο δείκτη στον συμβολισμό των παρατηρούμενων y_{ijk} και αναμενόμενων μ_{ijk} συχνοτήτων.

Ένας πίνακας συνάφειας λέγεται αραιός (*sparse*) όταν ο λόγος του μεγέθους του δείγματος ($y_{++} = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}$) προς τον αριθμό των κελλιών ($I \cdot J$) είναι σχετικά μικρός. Θα υποθέσουμε ότι όταν $y_{++} \to \infty$ και $I \cdot J \to \infty$, ο λόγος $y_{++} / I \cdot J$ θα παραμένει πεπερασμένος.

Συχνά στους αραιούς πίνακες συνάφειας εμφανίζονται και αρκετά κενά (μηδενικά) κελλιά. Τα μηδενικά αυτά κελλιά διαχωρίζονται σε δύο κατηγορίες. Στα κελλιά που περιέχουν δομικά (*structural*) μηδενικά και σε αυτά που περιέχουν τυχαία (*random* ή *sampling*) μηδενικά.

Τα κελλιά με δομικά μηδενικά είναι αυτά που πρέπει να παραμείνουν κενά εξαιτίας της δομής του προβλήματος. Για παράδειγμα σε έναν πίνακα συνάφειας που διασταυρώνει κατηγορίες καρκίνου (προστάτη, πνευμόνων, στομάχου) με το φύλο του ασθενούς (άντρας, γυναίκα) δεν έχει λογική να εμφανιστούν παρατηρήσεις στα κελλιά που διασταυρώνεται, ο καρκίνος προστάτη με το γυναικείο φύλο. Τέτοια λοιπόν κελλιά που περιέχουν δομικά μηδενικά, έχουν αναμενόμενες συχνότητες $\mu_{ij} = 0$ και κατά συνέπεια και οι εκτιμήσεις τους θα είναι $\hat{\mu}_{ij} = 0$.

Αντίθετα τα τυχαία μηδενικά εμφανίζονται σε κελλιά για τα οποία δεν καταγράψαμε καμία παρατήρηση, παρόλο που η πιθανότητα να συμβεί κάτι τέτοιο ήταν θετική. Τα κελλιά

αυτά έχουν αναμενόμενες συχνότητες $\mu_{ij} > 0$ και ανάλογα με το μοντέλο που θα εφαρμόσουμε οι εκτιμήσεις τους θα είναι $\hat{\mu}_{ij} = 0$ ή $\hat{\mu}_{ij} > 0$.

Συχνά όταν θέλουμε να αναλύσουμε ένα πίνακα συνάφειας δεσμεύουμε ως προς κάποιο περιθώριο άθροισμα που δεν ήταν σταθερό από την αρχή του προβλήματος. Αν το περιθώριο αυτό άθροισμα προκύψει μηδέν, τα τυχαία μηδενικά που συνεισφέρουν σε αυτό το μηδενικό δεσμευμένο περιθώριο άθροισμα, θα τα μετατρέψουμε σε δομικά μηδενικά για την μελέτη μας. Είναι πολύ σημαντικό λοιπόν, να έχουμε ξεκαθαρίσει από την αρχή πόσα κελλιά με δομικά μηδενικά υπάρχουν, διότι αυτά δεν είναι παρατηρήσεις, δεν αποτελούν μέρος των δεδομένων μας και συνεπώς θα πρέπει να εξαιρεθούν από την ανάλυση. Η εξαίρεση τους γίνεται με το να μην αντιστοιχίσουμε βαθμούς ελευθερίας στα κελλιά αυτά (βλ. *Baker et. al.* [7], *Bishop, Fienberg, Holland* [14] § 3.8). Για παράδειγμα οι βαθμοί ελευθερίας για τον έλεγχο ανεξαρτησίας σε έναν *I*x*J* πίνακας συνάφειας που περιέχει δομικά μηδενικά είναι $df = IJ - 1 - (\text{πλήθος ανεξάρτητων παραμέτρων}) - z_C$, όπου z_C είναι το πλήθος όλων των δομικών μηδενικών του πίνακα.

1.2 Προβλήματα από την ύπαρξη αραιών πινάκων συνάφειας

Στην παράγραφο αυτή θα αναπτύξουμε τα προβλήματα και τις δυσκολίες που είναι πιθανό να συναντήσει κάποιος στην προσπάθεια του να αναλύσει έναν αραιό πίνακα συνάφειας.

1.2.1 Αδυναμία υπολογισμού εκτιμητών μέγιστης πιθανοφάνειας

Τα κελλιά με τυχαία μηδενικά σε έναν πίνακα συνάφειας μπορεί να επηρρεάσουν την ύπαρξη πεπερασμένων εκτιμητών μέγιστης πιθανοφάνειας (*Maximum Likelihood Estimators*). Πράγματι για ιεραρχικά μή-κορεσμένα λογαριθμογραμμικά μοντέλα, οι ε.μ.π. υπάρχουν όταν όλα τα κελλιά έχουν παρατηρήσεις ($y_{ij} > 0$), ενώ δεν υπάρχουν όταν κάποιο από τα επαρκή στατιστικά (τα οποία συνήθως αντιστοιχούν στα περιθώρια αθροίσματα) είναι μηδέν (*Fienberg* [21], *Haberman* [30], [31]).

Αν σε τουλάχιστον ένα κελλί υπάρχει $y_{ij} = 0$, αλλά τα επαρκή περιθώρια αθροίσματα είναι όλα θετικά, τότε υπάρχουν πεπερασμένοι ε.μ.π., αρκεί τα μοντέλα να έχουν εκτιμητές που υπολογίζονται άμεσα (*direct* MLE) , βλ. *Glonek G. et.al.* [28].

Για τα μοντέλα όμως που περιέχουν όλα τα ζεύγη συσχετίσεων των μεταβλητών τους, τα πράγματα γίνονται πιο πολύπλοκα. Ας δούμε αυτή την περίπτωση καλύτερα μέσω ενός παραδείγματος.

Έστω τρείς μεταβλητές X,Y,Z οι οποίες έχουν από δύο κατηγορίες, και έστω ότι επιθυμούμε να προσαρμόσουμε το μοντέλο χωρίς την αλληλεπίδραση των τριών παραγόντων $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ}$, i,j,k=1,2, το οποίο συμβολίζουμε ως (XY,XZ,YZ). Οι εκτιμητές μέγιστης πιθανοφάνειας για αυτό το μοντέλο θα υπάρχουν, μόνο αν στον πίνακα συνάφειας εμφανιστεί ένα μόνο μηδενικό κελλί ($y_{ij} = 0$), ενώ μπορεί να μην υπάρχουν όταν δύο κελλιά είναι κενά, παρόλο που τα επαρκή περιθώρια αθροίσματα μπορεί να είναι όλα θετικά. Ας θεωρήσουμε ότι ο 2x2x2 πίνακας συνάφειας έχει δύο μηδενικά κελλιά, με την εξής διάταξη :

	Z:	1		2	
X	Y:	1	2	1	2
1		0	y_{121}	y_{112}	y_{122}
2		y_{211}	y_{221}	y_{212}	0

*όπου y_{ijk} οι θετικές συχνότητες των κελλιών

Τα επαρκή στατιστικά του μοντέλου (XY,XZ,YZ) είναι $\{y_{ij+}\},\{y_{i+k}\},\{y_{+jk}\}$, και αντιστοιχούν σε 2x2 πίνακες, με συχνότητες τα αντίστοιχα περιθώρια αθροίσματα που προκύπτουν από τον αρχικό πίνακα συνάφειας αν αθροίσουμε ως προς κάθε μια μεταβλητή. Οι συχνότητες αυτών των περιθώριων πινάκων είναι όλες θετικές.

Αν συμβολίσουμε με Δ τον εκτιμητή για το πρώτο κελλί ($\hat{\mu}_{111} = \Delta > 0$), οι εκτιμήσεις των άλλων κελλιών προκύπτουν από τις εξισώσεις πιθανοφάνειας (*likelihood equations*) $\{\hat{\mu}_{ij+} = y_{ij+}\},\{\hat{\mu}_{i+k} = y_{i+k}\},\{\hat{\mu}_{+jk} = y_{+jk}\}$ ως συνάρτηση του Δ.

Βρίσκουμε λοιπόν τις εκτιμήσεις για κάθε κελλί

	Z:	1		2	
X	Y:	1	2	1	2
1		Δ	y_{121} -Δ	y_{112} -Δ	y_{122} +Δ
2		y_{211} -Δ	y_{221} +Δ	y_{212} +Δ	-Δ

Προκύπτει όμως $\hat{\mu}_{222} = -\Delta < 0$ αρνητική. Είναι αδύνατον λοιπόν να βρούμε τιμή για τον εκτιμητή Δ τέτοια ώστε να προκύψουν θετικές εκτιμήσεις και για τα δύο μηδενικά κελλιά. Είναι φανερό λοιπόν ότι σε έναν τέτοιο πίνακα συνάφειας δεν μπορούμε να προσαρμόσουμε το μοντέλο (XY,XZ,YZ) που επιθυμούμε.

Παρατήρηση

Αν στο προηγούμενο πίνακα θεωρήσουμε ότι τα μηδενικά κελλιά είχαν διαφορετική διάταξη, όπως παρακάτω :

	Z:	1		2	
X	Y:	1	2	1	2
1		0	y_{121}	0	y_{122}
2		y_{211}	y_{221}	y_{212}	y_{222}

τότε θα εμφανιζόταν μηδενική συχνότητα στο πρώτο κελλί του πίνακα με τα XY περιθώρια αθροίσματα. Αυτό θα είχε ως συνέπεια να προκύψουν μη πεπερασμένοι εκτιμητές $\{\hat{\lambda}_{ij}^{XY}\}$ για τις παραμέτρους του λογαριθμογραμμικού μοντέλου που αντιστοιχούν στην αλληλεπίδραση XY. Σε αυτόν τον πίνακα και πάλι δεν μπορούμε να προσαρμόσουμε το μοντέλο (XY,XZ,YZ), μπορούμε όμως να προσαρμόσουμε κάποιο άλλο λογαριθμογραμμικό μοντέλο, αρκεί να μην περιέχει την αλληλεπίδραση των όρων XY.

1.2.2 Επίδραση στην ασυμπτωτική κατανομή των στατιστικών καλής προσαρμογής

Γνωρίζουμε από την ασυμπτωτική θεωρία ότι η προσεγγιστική κατανομή που ακολουθούν τα γνωστά στατιστικά καλής προσαρμογής X^2 του Pearson $X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(y_{ij}-\hat{\mu}_{ij}\right)^2}{\hat{\mu}_{ij}}$, και likelihood-ratio $G^2 = -2\cdot\log\Lambda = 2\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}\cdot\log\frac{y_{ij}}{\hat{\mu}_{ij}}$, είναι η χ^2-κατανομή. Μάλιστα η σύγκλιση σε αυτήν την κατανομή γίνεται καλύτερη όσο μεγαλύτερο είναι το μέγεθος του δείγματος ($y_{++}\to\infty$) και ο αριθμός των κελλιών ($I\cdot J$) παραμένει σταθερός.

Η καταλληλότητα λοιπόν της χ^2-κατανομής εξαρτάται από τα $y_{++}, I \cdot J$. Σε αραιούς πίνακες συνάφειας οδηγούμαστε στην μή ικανοποιητική σύγκλιση των στατιστικών καλής προσαρμογής στην χ^2-κατανομή.

Παραθέτουμε συνοπτικά κάποια γενικά αποδεκτά συμπεράσματα για το πότε είναι αξιόπιστη η προσέγγιση μέσω χ^2-κατανομής, καθώς και ποιό στατιστικό καλής προσαρμογής είναι προτιμότερο κάτω από συγκεκριμένες ειδικές περιπτώσεις αραιών πινάκων συνάφειας.

- Όταν ο λόγος $y_{++}/I \cdot J < 5$ το στατιστικό G^2 δεν προσεγγίζεται ικανοποιητικά από την χ^2-κατανομή. Μάλιστα όταν οι περισσότερες αναμενόμενες συχνότητες είναι $\mu_{ij} < 0{,}5$, η προσέγγιση αυτή μας οδηγεί σε ένα πολύ συντηρητικό (*conservative*) τέστ, όπου το p-value που εμείς υπολογίζουμε θα είναι αρκετά μεγαλύτερο από το πραγματικό p-value, ενώ όταν οι περισσότερες αναμενόμενες συχνότητες είναι $0{,}5 < \mu_{ij} < 4$ οδηγούμαστε σε ένα προοδευτικό (*liberal*) τέστ, όπου το p-value που εμείς θα υπολογίζουμε είναι αρκετά μικρότερο από το πραγματικό.
- Όταν η ελάχιστη τιμή που μπορεί να πάρουν οι αναμενόμενες συχνότητες των κελλιών του πίνακα είναι $\mu_{ij} \approx 1$, και το πολύ το 20% από αυτές είναι $\mu_{ij} < 5$, τότε το στατιστικό X^2 προσεγγίζεται ικανοποιητικά από την χ^2-κατανομή. Μάλιστα η προσέγγιση θα είναι αρκετά ακριβής αν τα μ_{ij} είναι σχεδόν όλα ίσα. Αν όμως ο πίνακας συνάφειας περιέχει μικρές αλλά και μεγάλες αναμενόμενες συχνότητες τότε η προσέγγιση δεν είναι ικανοποιητική.
- Το στατιστικό X^2, σε σύγκριση με το G^2, συμπεριφέρεται πιο ικανοποιητικά σε αραιούς πίνακες με μικρό μέγεθος δείγματος.

Για την θεωρητική τεκμηρίωση των παραπάνω συμπερασμάτων βλ. *Fienberg* [22], *Haberman* [29] και σχετικές αναφορές στους *Cressie & Read* [17], κεφ. *Historical Perpective*.

1.2.3 Προβλήματα στην απόδοση των στατιστικών προγραμμάτων

Αρκετά λογαριθμογραμμικά μοντέλα δεν έχουν εκτιμητές που να υπολογίζονται άμεσα, αλλά μόνο μετά την εφαρμογή επαναληπτικών μεθόδων (π.χ. *Newton-Raphson*).

Όταν σε έναν πίνακα συνάφειας κάποιες από τις εκτιμήσεις των αναμενόμενων συχνοτήτων των κελλιών προκύψουν μηδέν, είμαστε σίγουροι ότι κάποιοι ε.μ.π. για τις

παραμέτρους του μοντέλου που προσαρμόζουμε, θα προκύψουν ∞ (ή $-\infty$). Τότε ελοχεύει ο κίνδυνος, τα στατιστικά προγράμματα που θα χρησιμοποιήσουμε για τον υπολογισμό των εκτιμητών να μην μας δώσουν τις σωστές εκτιμήσεις (βλ. *Agresti* [1] σελ. 393-394).

Κατά την εκτέλεση της επαναληπτικής μεθόδου για τον υπολογισμό αυτών των ε.μ.π., αναμένουμε να μην οδηγηθούμε σε σύγκλιση διότι κάποιος εκτιμητής προφανώς θα συνεχίζει να αυξάνεται σε κάθε επανάληψη. Το πρόβλημα είναι ότι τα στατιστικά προγράμματα δεν καταφέρνουν να ανιχνεύσουν την συνεχιζόμενη αυτή αύξηση, εξαιτίας της πολύ μικρής καμπυλότητας του λογαρίθμου της συνάρτησης πιθανοφάνειας, με συνέπεια ύστερα από κάποιο σημείο να παρουσιάζουν σύγκλιση με λανθασμένες εκτιμήσεις των παραμέτρων και αντίστοιχα υπερβολικά μεγάλα τυπικά σφάλματα.

Χρειάζεται λοιπόν ιδιαίτερη προσοχή κάθε φορά από τον ερευνητή ώστε να μην «ξεγελαστεί» από την λανθασμένη σύγκλιση των στατιστικών πακέτων και αναφέρει ως εκτιμήσεις των παραμέτρων τιμές λανθασμένες και άκυρες.

1.3 Εναλλακτικές προσεγγίσεις

Στην συνέχεια θα μιλήσουμε για μερικές εναλλακτικές προσεγγίσεις που έχουν προταθεί με στόχο την αντιμετώπιση των δυσκολίων που συχνά συναντάμε στην ανάλυση των αραιών πινάκων συνάφειας.

1.3.1 Προσθήκη σταθεράς στα κελλιά του πίνακα συνάφειας

Για να αντιμετωπιστούν τα προβλήματα της μή ύπαρξης πεπερασμένων εκτιμητών, καθώς και για να εξασφαλιστεί η σύγκλιση των αλγορίθμων που χρησιμοποιούν τα στατιστικά πακέτα έχει προταθεί η πρόσθεση μιας σταθεράς σε όλα τα κελλιά του πίνακα (*Yates* [62]).

Ας δούμε ένα απλό παράδειγμα στο οποίο η χρήση της σταθεράς είναι πολύ βοηθητική. Έστω ο 2x2 πίνακας συνάφειας

X	Y: 1	2
1	y_{11}	0
2	y_{21}	y_{22}

* y_{ij} οι θετικές συχνότητες των κελλιών

όπου θέλουμε να εκτιμήσουμε τον λόγο πιθανοτήτων (*odds ratio*). Η εκτίμηση μέγιστης πιθανοφάνειας όμως είναι μή πεπερασμένη $\hat{\theta} = \frac{y_{11} \cdot y_{22}}{y_{12} \cdot y_{21}} = \frac{y_{11} \cdot y_{22}}{0 \cdot y_{21}} = \infty$. Σε αυτό το σημείο μπορεί να μας βοηθήσει η προσθήκη μιας σταθεράς. Έστω ότι προσθέτουμε την σταθερά 1/2 σε κάθε κελλί, τότε βρίσκουμε την πεπερασμένη εκτίμηση

$$\tilde{\theta} = \frac{(y_{11} + 0{,}5) \cdot (y_{22} + 0{,}5)}{(y_{12} + 0{,}5) \cdot (y_{21} + 0{,}5)} = \frac{(y_{11} + 0{,}5) \cdot (y_{22} + 0{,}5)}{0{,}5 \cdot (y_{21} + 0{,}5)} \ .$$

Βέβαια με την προσθήκη κάποιας σταθεράς στα κελλιά και εν συνεχεία με την προσαρμογή ενός μή-κορεσμένου μοντέλου στον πίνακα, υπάρχει ο κίνδυνος να ομαλοποιηθούν πολύ τα δεδομένα μας αχρηστεύοντας την δειγματική κατανομή. Πιθανή συνέπεια θα είναι να οδηγηθούμε σε κάποια αρκετά συντηρητικά (*conservative*) τέστ. Μάλιστα όσο μεγαλύτερος ο αριθμός των μηδενικών κελλιών, τόσο μεγαλύτερος είναι αυτός ο κίνδυνος. Συνεπώς η χρήση σταθεράς δεν είναι «πανάκεια», είναι αρκετά ριψοκίνδυνη μέθοδος και η επιλογή μας να την χρησιμοποιήσουμε ή όχι πρέπει να γίνεται μετά από σκέψη.

Έστω ότι στο προηγούμενο παράδειγμα θέλουμε να υπολογίσουμε ένα διάστημα εμπιστοσύνης (δ.ε.) για το θ. Πρίν την προσθήκη της σταθεράς, το δ.ε. που βασίζεται στον λόγο πιθανοφανειών θα έχει την μορφή (L, ∞), όπου L ένα πεπερασμένο κάτω άκρο. Μετά την προσθήκη της σταθεράς, το διάστημα εμπιστοσύνης που θα πάρουμε θα έχει και τα δύο άκρα πεπερασμένα. Όμως αυτό δεν έχει λογική, αφού κανένα δείγμα δεν θα μας δώσει ένδειξη ότι το θ βρίσκεται κάτω από κάποια δεδομένη τιμή. Είναι λοιπόν πιο σωστό να διατηρήσουμε ως άνω άκρο του διαστήματος εμπιστοσύνης το ∞. Το παράδειγμα αυτό είναι μια χαρακτηριστική περίπτωση όπου η προσθήκη της σταθεράς μας οδηγεί σε μή αποδεκτά αποτελέσματα.

1.3.2 Έμμεσος έλεγχος της καλής προσαρμογής ενός μοντέλου

Όταν επιθυμούμε να ελέγξουμε την καλή προσαρμογή ενός μοντέλου (M_1) σε έναν πίνακα συνάφειας, συνήθως χρησιμοποιούμε κάποιο από τα συνήθη στατιστικά (G^2, X^2) τα οποία προσεγγίζονται, όπως έχουμε προαναφέρει, από την χ^2-κατανομή. Αν όμως ο πίνακας συνάφειας είναι αραιός, τότε η ασυμπτωτική σύγκλιση των στατιστικών στην χ^2-κατανομή δεν είναι ικανοποιητική.

Μια εναλλακτική πρόταση είναι ο έμμεσος έλεγχος της καλής προσαρμογής του M_1, δοθέντος ότι ένα λιγότερο πολύπλοκο μοντέλο M_2 ($M_2 \supset M_1$) έχει καλή προσαρμογή στα δεδομένα μας. Για τον σκοπό αυτό τα στατιστικά που χρησιμοποιούμε είναι

$$G^2\left(M_1 \middle| M_2\right) = 2\sum_{i=1}^{I}\sum_{j=1}^{J} \hat{\mu}_{ij(2)} \cdot \log\left(\frac{\hat{\mu}_{ij(2)}}{\hat{\mu}_{ij(1)}}\right) \quad \text{και} \quad X^2\left(M_1 \middle| M_2\right) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(\hat{\mu}_{ij(2)} - \hat{\mu}_{ij(1)}\right)^2}{\hat{\mu}_{ij(1)}} \qquad (1.1)$$

όπου $\hat{\mu}_{ij(k)}$ είναι οι εκτιμήσεις των αναμενόμενων συχνοτήτων για το M_k μοντέλο, k=1,2. Το πλεονέκτημά τους είναι ότι προσεγγίζουν ασυμπτωτικά την χ^2-κατανομή γρηγορότερα και αποτελεσματικότερα σε σύγκριση με τα $G^2\left(M_1\right)$ και $X^2\left(M_1\right)$, βλ. *Agresti & Yang* [5].

Διαισθητικά αυτό μπορεί να εξηγηθεί ως εξής. Παρατηρούμε ότι τα στατιστικά (1.1) εξαρτώνται από τα δεδομένα μόνο μέσω των $\{\hat{\mu}_{ij(2)}\}$, ως εκ τούτου μόνο μέσω των ελαχίστων επαρκών στατιστικών για το μοντέλο M_2, και όχι απευθείας από τις συχνότητες $\{y_{ij}\}$ όπως συμβαίνει στα $G^2\left(M_1\right)$ και $X^2\left(M_1\right)$. Επιπλέον γνωρίζουμε ότι τα στατιστικά G^2, X^2 συγκλίνουν καλύτερα στην χ^2-κατανομή όσο οι αναμενόμενες συχνότητες των επαρκών στατιστικών του μοντέλου αυξάνονται. Για τα περισσότερα λογαριθμογραμμικά (*loglinear*) μοντέλα τα επαρκή στατιστικά τους αντιστοιχούν στα περιθώρια αθροίσματα του πίνακα. Επειδή όμως τα περιθώρια αυτά αθροίσματα είναι λιγότερο αραιά από ότι τα κελλιά του πίνακα, ως εκ τούτου αναμένουμε τα στατιστικά (1.1) να συγκλίνουν στην χ^2-κατανομή ταχύτερα, σε σύγκριση με τα $G^2\left(M_1\right)$, $X^2\left(M_1\right)$ τα οποία εξαρτώνται από τα κελλιά του πίνακα.

Πρακτικά, αν τα περιθώρια αθροίσματα στα οποία αντιστοιχούν τα ελάχιστα επαρκή στατιστικά του M_2, έχουν τιμές τουλάχιστον από 5 έως 10, η χ^2-κατανομή είναι καλή προσέγγιση για τα στατιστικά (1.1), ακόμα και αν οι εκτιμήσεις $\{\hat{\mu}_{ij}\}$ είναι μικρές. Βέβαια αν το μοντέλο M_2 δεν έχει καλή προσαρμογή στα δεδομένα, είναι προφανές ότι δεν έχει νόημα να χρησιμοποιήσουμε κάποιο από τα στατιστικά $G^2\left(M_1 \middle| M_2\right)$, $X^2\left(M_1 \middle| M_2\right)$.

Μια πολύ καλή συγκριτική μελέτη για την συμπεριφορά και την καταλληλότητα της χ^2-κατανομής ως προσέγγιση των στατιστικών X^2 και G^2 έχουν κάνει οι *Agresti* και *Yang* [5].

1.3.3 Ασυμπτωτική προσέγγιση μέσω κανονικής κατανομής

Εναλλακτικά, αντί της χ^2-κατανομής έχει προταθεί η χρηση της κανονικής κατανομής ως προσέγγιση των X^2 και G^2.

Συγκεκριμένα, για το στατιστικό G^2 προτείνεται η προσέγγιση μέσω κανονικής κατανομής όταν οι περισσότερες αναμενόμενες συχνότητες είναι < 5, $y_{++} \geq 15$ και $y_{++}^2 / I \cdot J > 10$ (*Kohler & Larntz* [38], *Kohler* [37]) . Η προσέγγιση αυτή είναι πιο ακριβής και προτιμότερη σε σύγκριση με την αντίστοιχη προσέγγιση του X^2 μέσω κανονικής κατανομής (*Zelterman* [65]), η οποία είναι πολύ ευαίσθητη αν ο πίνακας συνάφειας έχει κάποιες πολύ μικρές αναμενόμενες συχνότητες.

Γενικά όμως όταν $10 \leq y_{++} \leq 20$ και $2 \leq I \cdot J \leq 6$ τότε η προσέγγιση των X^2 και G^2 μέσω κανονικής κατανομής δεν αποτελεί καθόλου καλή επιλογή. Σε αυτές τις περιπτώσεις προτιμότερη είναι η προσέγγιση μέσω χ^2-κατανομής (βλ. *Cressie & Read* [17], κεφ. *Historical Perpective*).

1.3.4 Εφαρμογή ακριβών (*exact*) μεθόδων

Όταν η εφαρμογή της ασυμπτωτικής θεωρίας είναι ακατάλληλη, μια καλή εναλλακτική πρόταση είναι οι ακριβείς (*exact*) μέθοδοι. Με τις μεθόδους αυτές μπορούμε να κάνουμε πιθανοτικούς υπολογισμούς (π.χ. p-value) βασιζόμενοι στις ακριβείς κατανομές και όχι σε ασυμπτωτικά προσεγγιστικές κατανομές. Το χαρακτηριστικό τους είναι ότι σε μια υπόθεση ελέγχου, το μέγεθος (*size*) (α = *P*[*σφάλμα τύπου Ι*]) είναι πάντα μικρότερο ή ίσο του ονομαστικού (*nominal*) μεγέθους του τέστ.

Υπάρχουν δύο διαφορετικές προσεγγίσεις των ακριβών μεθόδων. Οι ακριβείς μέθοδοι με δέσμευση (*exact conditional*) και χωρίς δέσμευση (*exact unconditional*). Πολλοί στατιστικοί άσκησαν έντονη κριτική στις ακριβείς μεθόδους. Η βάση της κριτικής τους, τις περισσότερες φορές είχε να κάνει με την συντηρητικότητά (*conservatism*) τους, ότι δηλαδή οι πραγματικές πιθανότητες σφαλμάτων είναι μικρότερες από τα ονομαστικά επίπεδα. Η συντηρητικότητα βέβαια πηγάζει από την διακριτότητα (*discreteness*) των ακριβών κατανομών. Για μια πολύ καλή κριτική επισκόπηση των ακριβών μεθόδων βλέπε *Agresti* [6].

Στο επόμενο κεφάλαιο θα μιλήσουμε αναλυτικότερα για τις ακριβείς μεθόδους, θα δούμε την εφαρμογή τους σε διδιάστατους και τριδιάστατους πίνακες συνάφειας, και θα μιλήσουμε για τις διενέξεις που έχουν προκαλέσει οι δύο διαφορετικές ακριβείς προσεγγισείς. Το

ενδιαφέρον μας θα εστιαστεί στο πλαίσιο ελέγχου ανεξαρτησίας και δεσμευμένης ανεξαρτησίας.

ΚΕΦΑΛΑΙΟ 2

Ακριβής (*exact*) συμπερασματολογία

2.1 Εισαγωγή

Ιστορικά η πιό συνηθισμένη προσέγγιση της ακριβούς συμπερασματολογίας στους πίνακες συνάφειας είναι μέσω δέσμευσης. Συνήθως επιθυμούμε να εξάγουμε συμπεράσματα για κάποια παράμετρο ενός λογαριθμογραμμικού μοντέλου. Οι δεσμευμένες ακριβείς μέθοδοι κάνουν χρήση της κατανομής του επαρκούς στατιστικού για την παράμετρο αυτή, δεσμεύοντας όμως ως προς τα επαρκή στατιστικά των άλλων παραμέτρων του μοντέλου που δεν μας ενδιαφέρουν, τις λεγόμενες και οχληρές (*nuisance*) παραμέτρους.

Παραδείγματος χάριν έστω ότι θέλουμε να ελέγξουμε την υπόθεση H_0 : το μοντέλο M_0 έχει καλή προσαρμογή στα δεδομένα, έναντι της εναλλακτικής H_1 : ένα πιο πολύπλοκο μοντέλο M_1 έχει καλή προσαρμογή. Αν με $\mathbf{z}_0$, $\mathbf{z}_1$ συμβολίζουμε τα ελάχιστα επαρκή στατιστικά των M_0 και M_1 αντίστοιχα, η δεσμευμένη ακριβής μέθοδος θα χρησιμοποιήσει την κατανομή του $\mathbf{z}_1$, δοθέντος του $\mathbf{z}_0$. Λόγω της δέσμευσης, η κατανομή του $\mathbf{z}_1$ δεν θα εξαρτάται πια από τις παραμέτρους που δεν μας ενδιαφέρουν, και έτσι ο ακριβής υπολογισμός των πιθανοτήτων της κατανομής αυτής καθίσταται δυνατός.

Εκτός όμως της παραπάνω μεθόδου αναπτύχθηκε και η θεωρία των ακριβών μεθόδων χωρίς δέσμευση. Η προσέγγιση αυτή είναι πιο γενική αφού δεν απαιτεί μοντέλα με μειωμένα (δεσμευμένα) επαρκή στατιστικά. Στην §2.6 θα αναφερθούμε εκτενέστερα για την μέθοδο αυτή και θα μιλήσουμε και για τις διενέξεις που που έχουν δημιουργηθεί ανάμεσα στους υποστηρικτές των δύο αυτών διαφορετικών προσεγγίσεων των ακριβών μεθόδων (*conditional* vs *unconditional*).

Στις επόμενες παραγράφους θα παρουσιάσουμε την εφαρμογή των δεσμευμένων ακριβών μεθόδων σε διδιάστατους και τριδιάστατους πίνακες συνάφειας, όταν στόχος μας είναι ο έλεγχος της ανεξαρτησίας (*independence*) καθώς και της δεσμευμένης ανεξαρτησίας (*conditional independence*) μεταξύ των μεταβλητών ταξινόμησης, ενώ στην §2.4 θα κάνουμε μια συνοπτική αναφορά στην ανάλυση διδιάστατων τετραγωνικών πινάκων συνάφειας.

Εκμεταλλευόμενοι την προσαρμοστικότητα των δεσμευμένων ακριβών μεθόδων, αφού μπορούν να εφαρμοστούν σε οποιαδήποτε εκθετική οικογένεια μοντέλων με κανονικό σύνδεσμο (*canonical link*), εμείς θα τις εφαρμόσουμε σε πλαίσιο Poisson λογαριθμογραμμικών μοντέλων τα οποία μας επιτρέπουν την «απαλοιφή» των οχληρών παραμέτρων, απλά δεσμεύοντας ως προς τα αντίστοιχα επαρκή στατιστικά (βλ. *McCullagh & Nelder* [45] σελ.32).

2.2 Ανεξαρτησία σε *2x2* πίνακες συνάφειας

Έστω το Poisson λογαριθμογραμμικό μοντέλο M_0 : $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$, i=1,2 , j=1,2. Το μοντέλο αυτό εκφράζει την ανεξαρτησία των μεταβλητών ταξινόμησης X,Y. Υπό τους περιορισμούς $\lambda_2^X = \lambda_2^Y = 0$ οι μή πλεονάζουσες (*non-redundant*) παράμετροι του μοντέλου είναι $\{\lambda, \lambda_1^X, \lambda_1^Y\}$. Με βάση τα συμπεράσματα του *Birch* [12] για τις εξισώσεις πιθανοφάνειας (*likelihood equations*), έχουμε ότι τα επαρκή στατιστικά θα είναι το y_{++} για την παράμετρο λ, το y_{1+} για την παράμετρο λ_1^X και το y_{+1} για την παράμετρο λ_1^Y.

Θα ελέγξουμε την καλή προσαρμογή του μοντέλου αυτού έναντι του κορεσμένου M_1 : $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$, i,j=1,2. Υπό τους περιορισμούς $\lambda_2^X = \lambda_2^Y = 0$ και $\lambda_{2j}^{XY} = \lambda_{i2}^{XY} = 0$, i,j=1,2 , οι μή πλεονάζουσες παράμετροι του μοντέλου είναι $\{\lambda, \lambda_1^X, \lambda_1^Y, \lambda_{11}^{XY}\}$

Η παράμετρος που προφανώς μας ενδιαφέρει είναι η λ_{11}^{XY}, διότι αν λ_{11}^{XY}=0 τότε αυτό συνεπάγεται την ανεξαρτησία των X,Y. Τα επαρκή στατιστικά του M_1 είναι y_{++} για την παράμετρο λ, y_{1+} για την παράμετρο λ_1^X, y_{+1} για την παράμετρο λ_1^Y και y_{11} για την παράμετρο λ_{11}^{XY}. Για να πραγματοποιήσουμε λοιπόν ακριβή συμπερασματολογία για το λ_{11}^{XY}, θα υπολογίσουμε την δεσμευμένη κατανομή του y_{11}, δοθέντων των y_{++}, y_{1+}, y_{+1} που είναι τα επαρκή στατιστικά των οχληρών παραμέτρων.

Η κατανομή αυτή εξαρτάται από την παράμετρο λ_{11}^{XY} και όπως έδειξε ο *Fisher* [23], είναι η μη κεντρική υπεργεωμετρική κατανομή, με συνάρτηση μάζας πιθανότητας

$$P(y_{11}=t)=f(t\,|\,y_{++},y_{1+},y_{+1};\theta)=\frac{\binom{y_{1+}}{t}\cdot\binom{y_{++}-y_{1+}}{y_{+1}-y_{11}}\cdot\theta^{y_{11}}}{\sum_{u=m_-}^{m_+}\binom{y_{1+}}{u}\cdot\binom{y_{++}-y_{1+}}{y_{+1}-u}\cdot\theta^{u}} \qquad (2.1)$$

όπου $m_-=\max(0,y_{1+}+y_{+1}-y_{++})$ και $m_+=\min(y_{1+},y_{+1})$ και $\theta=e^{\lambda_{11}^{XY}}$, το odds ratio $\theta=\frac{\mu_{11}\cdot\mu_{22}}{\mu_{12}\cdot\mu_{21}}$.

Υπό την μηδενική υπόθεση της ανεξαρτησίας H_0 : $\lambda_{11}^{XY}=0$ (ή ισοδύναμα $\theta=1$), η παραπάνω κατανομή είναι η υπεργεωμετρική κατανομή και δεν εξαρτάται από καμία παράμετρο. Είναι συνεπώς μια ακριβής κατανομή αφού μας επιτρέπει να υπολογίσουμε ακριβώς οποιαδήποτε πιθανότητά της.

Για την ολοκλήρωση του ελέγχου απαιτείται και ο υπολογισμός της p-value. Η p-value για το μονόπλευρο τέστ H_0 : $\theta=1$ κατά H_1 : $\theta>1$ είναι:

$$\text{p-value}=P_{H_0}[\,y_{11}\geq y_{11,obs}\,|\,y_{++},y_{1+},y_{+1};\theta=1\,]=\sum_S f(t\,|\,y_{++},y_{1+},y_{+1};\theta=1)$$

όπου S είναι το σύνολο των πινάκων που έχουν ακριβώς τις ίδιες τιμές των ποσοτήτων y_{++}, y_{1+}, y_{+1} με τον παρατήρουμενο πίνακα, και επιπλέον έχουν $y_{11}\geq y_{11,obs}$. Δηλαδή $S=\{t:t\geq y_{11,obs}\}$. Αυτό το τέστ είναι ευρέως γνωστό ως ακριβές τέστ του Fisher (*Fisher's exact test*).

Αν η εναλλακτική υπόθεση ήταν H_1 : $\theta<1$ τότε το μόνο που θα άλλαζε στον υπολογισμό της p-value θα ήταν το σύνολο των πινάκων ως προς τους οποίους θα αθροίζαμε τις πιθανότητες (2.1). Δηλαδή θα είχαμε $S=\{t:t\leq y_{11,obs}\}$. Για την p-value του αμφίπλευρου ελέγχου H_0 : $\theta=0$ κατά H_1 : $\theta\neq1$ έχουν προταθεί διάφορες προσεγγίσεις (*Yates* [63], *Davis* [19], *Dupont* [20], *Mantel* [43b], *Lloyd* [42]). Οι πιο διαδεδομένες είναι:

a. $S=\{t:f(t\,|\,y_{++},y_{1+},y_{+1})\leq f(y_{11,obs}\,|\,y_{++},y_{1+},y_{+1})\}$

b. $S=\{t:|t-E(y_{11})|\geq|y_{11,obs}-E(y_{11})|\}$ όπου $E(y_{11})=\frac{y_{1+}\cdot y_{+1}}{y_{++}}$ η αναμενόμενη συχνότητα στο κελλί (1,1), δηλαδή η μέση τιμή της υπεργεωμετρικής κατανομής. Η μέθοδος αυτή είναι όμοια με το αν υπολογίζαμε την p-value βασιζόμενοι στο X^2 του Pearson. Δηλαδή $\text{p-value}=P_{H_0}[\,X^2\geq X^2_{obs}\,]=\sum_S f(t\,|\,y_{++},y_{1+},y_{+1};\theta=1)$ όπου $S=\{t:X^2\geq X^2_{obs}\}$.

Συνήθως προτιμούμε την μέθοδο αυτή διότι γενικεύεται εύκολα σε *I*x*J* πίνακες συνάφειας.

c. Διπλασιασμός της p-value του μονόπλευρου ελέγχου. Δηλαδή p-value = $2\min\{ P_{H_0}[\, y_{11} \geq y_{11,obs}\,],\ P_{H_0}[\, y_{11} \leq y_{11,obs}\,]\}$. Η μέθοδος αυτή μας παρέχει έναν απλό και εύκολο τρόπο υπολογισμού της p-value έχει όμως το μειονέκτημα ότι μπορεί να προκύψει p-value >1.

2.2.1 Παράδειγμα

Ένα κλασσικό παράδειγμα που θα βοηθήσει να γίνουν περισσότερο κατανοητά όσα αναφέραμε στην προηγούμενη παράγραφο, είναι το πείραμα που πραγματοποίησε ο *Fisher* γνωστό με το όνομα "Η κυρία με το τσάι".

Μια κυρία που πίνει τσάι με γάλα, ισχυριζόταν ότι μπορεί να καταλάβει ποιό από τα δύο ρίχτηκε πρώτο στην κούπα. Της δόθηκαν οκτώ κούπες τσάι όπου στις τέσσερις το τσάι προστέθηκε πρώτο, ενώ στις υπόλοιπες τέσσερις το γάλα ήταν αυτό που προστέθηκε πρώτο. Στην συνέχεια καταγράφηκαν οι απαντήσεις της σε έναν πίνακα όπως ο παρακάτω :

	Απάντηση (πρόβλεψη) κυρίας		
Πραγματική εικόνα	Γάλα 1ο	Τσάι 1ο	Σύνολο
Γάλα 1ο	3	1	4
Τσάι 1ο	1	3	4
Σύνολο	4	4	8

Πηγή : Βασισμένο σε πείραμα του *Fisher* [23]

Στόχος μας είναι να ελέγξουμε αν οι απαντήσεις της κυρίας είναι ανεξάρτητες από την πραγματική εικόνα (H_0:θ=1) έναντι της υπόθεσης ότι η κυρία έχει ικανότητα πρόβλεψης (H_1:θ>1). Αρχικά θα βρούμε όλους τους *2*x*2* πίνακες που έχουν τις ίδιες τιμές στις ποσότητες y_{++}, y_{1+}, y_{+1} με τον παρατηρούμενο πίνακα ($y_{++}=8,\ y_{1+}=4,\ y_{+1}=4$). Από αυτούς τους πίνακες (πέντε στο πλήθος) μόνο δύο έχουν τιμή $y_{11} \geq y_{11,obs}=3$. Αυτό συμβαίνει για τους πίνακες με συχνότητες (3,1/1,3) και (4,0/0,4) κατά γραμμή.

Συνεπώς

$$\text{p-value} = f(y_{11}=3)+f(y_{11}=4) = \frac{\binom{4}{3}\cdot\binom{8-4}{4-3}}{\sum_{u=0}^{4}\binom{4}{u}\cdot\binom{8-4}{4-u}} + \frac{\binom{4}{4}\cdot\binom{8-4}{4-4}}{\sum_{u=0}^{4}\binom{4}{u}\cdot\binom{8-4}{4-u}} =$$

$$= 0{,}228 + 0{,}014 = 0{,}243 >> 0{,}05.$$

Άρα καταλήγουμε στο συμπέρασμα ότι με επίπεδο σημαντικότητας 5% δεν απορρίπτουμε την H_0. Δηλαδή η κυρία έκανε τις προβλέψεις της στην τύχη.

Για τον αμφίπλευρο έλεγχο H_0:θ=1 κατά H_1:θ≠1, δουλεύοντας με παρόμοια λογική υπολογίζουμε τις p-values και για τις τρείς διαφορετικές προσεγγίσεις που αναφέραμε στο τέλος της §2.2.

a. p-value = $f(y_{11}=3)+f(y_{11}=4)+f(y_{11}=1)+f(y_{11}=0)$=0,486

b. p-value = $f(y_{11}=3)+f(y_{11}=4)+f(y_{11}=0)$=0,257

c. p-value = 2 x 0,243 =0,486

Οι τρείς παραπάνω προσεγγίσεις καταλήγουν στο ίδιο συμπέρασμα, της μή απόρριψης της μηδενικής υπόθεσης σε επίπεδο σημαντικότητας 5%. Δηλαδή η κυρία έκανε τις προβλέψεις της στην τύχη.

Το γεγονός ότι η μέθοδος (b) μας δίνει διαφορετικό p-value σε σχέση με τις άλλες δύο μεθόδους, οφείλεται στην διακριτότητα (*discreteness*) και στην έντονη λοξότητα της υπεργεωμετρικής κατανομής που χρησιμοποιούμε. Η διακριτότητα της δεσμευμένης κατανομής είναι ένα επιχείρημα που χρησιμοποιούν εναντίον των δεσμευμένων ακριβών μεθόδων οι μή υποστηρικτές τους. Αναλυτικότερη παρουσίαση των μακροχρόνιων αυτών διενέξεων θα κάνουμε στην §2.7.

2.3 Ανεξαρτησία σε *I*x*J* πίνακες συνάφειας

Στους *I*x*J* πίνακες συνάφειας θα γενικεύσουμε τον τρόπο με τον οποίο εργαστήκαμε για τους *2*x*2* πίνακες συνάφειας. Η ανεξαρτησία των μεταβλητών X,Y αντιστοιχεί στο λογαριθμογραμμικό μοντέλο :

$$M_0 : \log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y \text{ , } i=1,\ldots,I \text{ και } j=1,\ldots,J \qquad (2.2)$$

Υπό τους περιορισμούς $\lambda_I^X = \lambda_J^Y = 0$ το M_0 έχει 1+(*I*-1)+(*J*-1) μή πλεονάζουσες παραμέτρους, τις $\{\lambda, \lambda_i^X, \lambda_j^Y\}$, *i*=1,…,*I*-1 , *j*=1,…, *J*-1 και τα επαρκή στατιστικά για τις παραμέτρους αυτές είναι τα y_{++}, $\{y_{i+}\}$ και $\{y_{+j}\}$, *i*=1,…,*I*-1 , *j*=1,…, *J*-1.

Θα ελέγξουμε την καλή προσαρμογή του M_0, έναντι του κορεσμένου μοντέλου M_1 : $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$, *i*=1,…,*I* , *j*=1,…,*J*. Υπό τους περιορισμούς $\lambda_I^X = \lambda_J^Y = 0$ και $\lambda_{iJ}^{XY} = \lambda_{Ij}^{XY} = 0$ το M_1 θα έχει 1+(*I*-1)+(*J*-1)+(*I*-1)(*J*-1) μή πλεονάζουσες παραμέτρους, τις $\{\lambda, \lambda_i^X, \lambda_j^Y, \lambda_{ij}^{XY}\}$, *i*=1,…,*I*-1 , *j*=1,…, *J*-1 και τα επαρκή στατιστικά για τις παραμέτρους αυτές είναι τα y_{++}, $\{y_{i+}\}$,$\{y_{+j}\}$και $\{y_{ij}\}$, *i*=1,…,*I*-1 , *j*=1,…, *J*-1.

Οι παράμετροι που μας ενδιαφέρουν είναι οι όροι αλληλεπίδρασης $\{\lambda_{ij}^{XY}\}$, *i*=1,…,*I*-1 , *j*=1,…, *J*-1. Για να πραγματοποιήσουμε λοιπόν ακριβή συμπερασματολογία για την παράμετρο αυτή, θα χρησιμοποιήσουμε την δεσμευμένη κατανομή των κελιών $\{y_{ij}\}$, δοθέντων των y_{++}, $\{y_{i+}\}$, $\{y_{+j}\}$, *i*=1,…,*I*-1 , *j*=1,…, *J*-1 που είναι τα επαρκή στατιστικά των οχληρών παραμέτρων λ, $\{\lambda_i^X\}$, $\{\lambda_j^Y\}$ αντίστοιχα.

Βέβαια όταν σε έναν πίνακα συνάφειας γνωρίζουμε τις ποσότητες y_{++}, $\{y_{i+}\}$, $\{y_{+j}\}$, *i*=1,…,*I*-1 , *j*=1,…, *J*-1 , τότε εκ των πραγμάτων γνωρίζουμε και τα τελευταία περιθώρια αθροίσματα y_{I+} και y_{+J}. Για ευκολία λοιπόν από εδώ και πέρα θα αναφέρουμε την δεσμευμένη κατανομή του $\{y_{ij}\}$, δοθέντων όλων των περιθώριων αθροισμάτων $\{y_{i+}\},\{y_{+j}\}$, *i*=1,…,*I* , *j*=1,…, *J* , τα οποία είναι και τα ελάχιστα επαρκή στατιστικά του M_0.

Ο *Cornfield* [16] έδειξε ότι η κατανομή αυτή, υπό την H_0: $\{\lambda_{ij}^{XY}\} = 0$, *i*=1,…,*I*-1 , *j*=1,…, *J*-1 , είναι η πολυδιάστατη υπεργεωμετρική. Συνεπώς η πιθανότητα ένας πίνακας $\{t_{ij}\}$ να έχει περιθώρια αθροίσματα ίδια με τα παρατηρούμενα είναι

$$P_{H_0}\left(\{t_{ij}\}\right) = f\left(\{t_{ij}\} \middle| \{y_{i+}\}, \{y_{+j}\}\right) = \frac{\prod_{i=1}^{I} y_{i+}! \cdot \prod_{j=1}^{J} y_{+j}!}{y_{++}! \cdot \prod_{i=1}^{I}\prod_{j=1}^{J} t_{ij}!} \qquad (2.3)$$

Η p-value για τον αμφίπλευρο αυτόν έλεγχο, είναι p-value = $\sum_S f\left(\{t_{ij}\} \middle| \{y_{i+}\}, \{y_{+j}\}\right)$ όπου $S = \left\{\{t_{ij}\} : f\left(\{t_{ij}\} \middle| \{y_{i+}\}, \{y_{+j}\}\right) \le f\left(\{y_{ij}\} \middle| \{y_{i+}\}, \{y_{+j}\}\right)\right\}$. Η p-value αυτή προτάθηκε από τους

Freeman & Halton [26]. Πολλοί στατιστικοί όμως υποστήριξαν ότι η p-value είναι προτιμότερο να βασίζεται στην ακριβή κατανομή ενός στατιστικού (π.χ. του *score* στατιστικού), έτσι ώστε να μπορεί να ποσοτικοποιηθεί η «απόσταση» των δεδομένων μας από την υπόθεση H_0 που ελέγχουμε. Ο *Yates* [62] για παράδειγμα, χρησιμοποιήσε το X^2 του Pearson το οποίο είναι το score στατιστικό για αυτόν τον έλεγχο ανεξαρτησίας. Τότε έχουμε p-value = $\sum_S f\left(\{t_{ij}\}\middle|\{y_{i+}\},\{y_{+j}\}\right)$ όπου $S=\left\{\{t_{ij}\}:X^2\geq X^2_{obs}\right\}$.

Βέβαια σε όλους τους παραπάνω ελέγχους αντιμετωπίζαμε τις μεταβλητές X, Y ως ονοματικές (*nominal*). Αν όμως κάποια από τις μεταβλητές αυτές είναι διατάξιμη (*ordinal*) ή ακόμα αν και οι δύο είναι διατάξιμες, και εμείς εφαρμόσουμε το προηγούμενο τέστ, τότε ελοχεύει ο κίνδυνος να οδηγηθούμε σε λάθος συμπεράσματα. Στην §2.3.4 θα το δούμε αυτό μέσω ενός παραδείγματος. Στην συνέχεια θα εξετάσουμε τις περιπτώσεις διαταξιμότητας των μεταβλητών ταξινόμησης X,Y.

2.3.1 Οι μεταβλητές X και Y είναι διατάξιμες

Έστω ότι και οι δύο μεταβλητές είναι διατάξιμες (*ordinal*). Σε αυτές τις περιπτώσεις είναι προτιμότερο να ελέγξουμε την καλή προσαρμογή του μοντέλου της ανεξαρτησίας M_0 έναντι ενός πιο περιοριστικού μοντέλου από το M_1, το οποίο θα περιγράφει την συνάφεια των X,Y χρησιμοποιώντας λιγότερες παραμέτρους και θα λαμβάνει υπόψην του και την διαταξιμότητα των μεταβλητών. Με αυτόν τον τρόπο επιτυγχάνουμε αύξηση της ισχύος του τέστ για την ανίχνευση πιθανής τάσης στην συνάφεια των X,Y , αφού εστιάζουμε την αλληλεπίδραση τους σε λιγότερους βαθμούς ελευθερίας.

Ως εναλλακτικό λοιπόν μοντέλο του (2.2), θα χρησιμοποιήσουμε το :

$$M_2:\ \log(\mu_{ij})=\lambda+\lambda_i^X+\lambda_j^Y+\beta u_i\upsilon_j\ ,\ i=1,\ldots,I\ \text{και}\ j=1,\ldots,J. \qquad (2.4)$$

Το μοντέλο M_2 έχει μία επιπλέον παράμετρο από την ανεξαρτησία, την β, σε αντίθεση με το κορεσμένο μοντέλο M_1 το οποίο χρησιμοποιούσε $(I-1)(J-1)$ επιπλέον παραμέτρους. Με $u_1<\ldots<u_I$ και $\upsilon_1<\ldots<\upsilon_J$ συμβολίζουμε τα μονότονα scores που αναθέτουμε στις γραμμές και τις στήλες του πίνακα αντίστοιχα, τα οποία συνήθως ισαπέχουν για διαδοχικές κατηγορίες των μεταβλητών ταξινόμησης. Το μοντέλο M_2 είναι γνωστό και ως μοντέλο ομοιόμορφης (*uniform*) ή *linear-by-linear* συνάφειας. Προφανώς η υπόθεση H_0: β=0 όταν ισχύει,

συνεπάγεται την ανεξαρτησία των X,Y. Τα ελάχιστα επαρκή στατιστικά του M_2 είναι τα $\{y_{i+}\}$, $\{y_{+j}\}$ και $\sum_{i=1}^{I}\sum_{j=1}^{J} u_i \cdot \upsilon_j \cdot y_{ij}$.

Για να πραγματοποιήσουμε ακριβή συμπερασματολογία για την παράμετρο β που μας ενδιαφέρει, θα χρησιμοποιήσουμε την δεσμευμένη κατανομή του επαρκούς στατιστικού $T=\sum_{i=1}^{I}\sum_{j=1}^{J} u_i \cdot \upsilon_j \cdot y_{ij}$, δοθέντων των ελαχίστων επαρκών στατιστικών του μοντέλου της ανεξαρτησίας M_0, ή πιο απλά δοθέντων όλων των περιθώριων αθροισμάτων. Ο *Cornfield* [16] έδειξε ότι όταν δεσμεύουμε ως προς τα περιθώρια αθροίσματα, η κατανομή του $\{y_{ij}\}$ είναι ανάλογη της ποσότητας

$$\frac{\prod_{i=1}^{I-1}\prod_{j=1}^{J-1}\alpha_{ij}^{y_{ij}}}{\prod_{i=1}^{I}\prod_{j=1}^{J} y_{ij}!} \qquad (2.5)$$

όπου $\alpha_{ij}=\frac{\mu_{ij}\cdot\mu_{IJ}}{\mu_{iJ}\cdot\mu_{Ij}}$, i=1,…,I-1 , j=1,…,J-1 είναι το τοπικό (*local*) odds ratio με κελλί αναφοράς το (I, J). Το μοντέλο M_2 μπορεί να γραφεί ισοδύναμα σε όρους odds ratios, παίρνοντας την μορφή: $\log(\alpha_{ij})=\beta\cdot(u_1-u_i)\cdot(\upsilon_J-\upsilon_j)$, i=1,…,I-1 , j=1,…,J-1. Η (2.5) τότε απλοποιείται και γίνεται $e^{\beta T}\Big/\prod_{i=1}^{I}\prod_{j=1}^{J} y_{ij}!$. Η κατανομή λοιπόν του T θα εξαρτάται από το β, και θα είναι

$$f\left(\mathrm{T}=t\,\middle|\,\{y_{i+}\},\{y_{+j}\};\beta\right)=\frac{C_t\cdot e^{\beta T_{obs}}}{\sum_{S} C_k\cdot e^{\beta k}},\ i=1,\ldots,I,\ j=1,\ldots,J\ ,\ \text{όπου}\ C_k=\sum_{S_k}\left(\prod_i\prod_j y_{ij}!\right)^{-1}.$$

Ο παρανομαστής αθροίζει ως προς το σύνολο S των πινάκων που έχουν τα ίδια περιθώρια αθροίσματα με τον παρατηρούμενο πίνακα (έστω R στο πλήθος οι πίνακες αυτοί). Επίσης με $T_1, T_2,\ldots,T_k,\ldots,T_K$ συμβολίζουμε τις K δυνατές τιμές που παίρνει το στατιστικό T για τους R πίνακες που έχουν τα ίδια περιθώρια αθροίσματα με τα παρατηρούμενα, και S_k = {το σύνολο των πινάκων οι οποίοι έχουν τα ίδια περιθώρια αθροίσματα με τα παρατηρούμενα και επίσης έχουν $T=T_k$}. Προφανώς δεν είναι απαραίτητο να ισχύει R=K, αλλά ισχύει R≥K.

Η p-value για το μονόπλευρο τέστ H_0 : β=0 κατά H_1 : β>0 είναι

$$\text{p-value} = P_{H_0}\left(\mathrm{T} \geq \mathrm{T}_{obs} \middle| \{y_{i+}\}, \{y_{+j}\}; \beta = 0\right) = \frac{\sum_{S_1}\left(\prod_i \prod_j y_{ij}!\right)^{-1}}{\sum_{S}\left(\prod_i \prod_j y_{ij}!\right)^{-1}}$$

όπου $S_1 = \{$το σύνολο των πινάκων που ανήκουν στο S και επιπλέον έχουν $\mathrm{T} \geq \mathrm{T}_{obs}\}$

Για την p-value του αμφίπλευρου ελέγχου H_0: $\beta=0$ κατά H_1: $\beta \neq 0$ προτείνεται η προσέγγιση p-value = $\sum_{S_2} f\left(\mathrm{T} \middle| \{y_{i+}\}, \{y_{+j}\}; \beta\right)$, με $S_2 = \{$το σύνολο των πινάκων που ανήκουν στο S και επιπλέον έχουν $\left|\mathrm{T} - E(\mathrm{T})\right| \geq \left|\mathrm{T}_{obs} - E(\mathrm{T})\right|\}$, όπου $E(\mathrm{T})$ είναι η αναμενόμενη τιμή του στατιστικού T όταν $\beta=0$, $E(\mathrm{T}) = \sum_{k=1}^{K} f\left(\mathrm{T}_k \middle| \{y_{i+}\}, \{y_{+j}\}; \beta = 0\right) \cdot \mathrm{T}_k$.

2.3.2 Η μεταβλητή X ονοματική και η Y διατάξιμη

Αν η μεταβλητή X είναι ονομαστική και η Y διατάξιμη, τότε ως εναλλακτικό μοντέλο του M_0 θα χρησιμοποιήσουμε το

$$M_3 : \log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + c_i \upsilon_j \ , \ i=1,\ldots,I \ , \ j=1,\ldots,J.$$

Το μοντέλο M_3 έχει I-1 επιπλέον παραμέτρους από εκείνο της ανεξαρτησίας, τις $\{c_i\}$. Με $\upsilon_1 < \ldots < \upsilon_J$ συμβολίζουμε τα μονότονα scores που αναθέτουμε για τις στήλες του πίνακα. Το μοντέλο M_3 είναι γνωστό και ως μοντέλο επίδρασης γραμμών (*row effect model*) και προκύπτει αν αντικαταστήσουμε τις διατάξιμες τιμές $\{\beta u_i\}$ στον όρο $\{\beta u_i \upsilon_j\}$ του μοντέλου (2.4) με τις παραμέτρους $\{c_i\}$.

Προφανώς η αλήθεια της υπόθεσης H_0: $c_1 = c_2 = \ldots = c_I$ συνεπάγεται την ανεξαρτησία των μεταβλητών X,Y. Τα ελάχιστα επαρκή στατιστικά του μοντέλου είναι $\{y_{i+}\}, \{y_{+j}\}$ και

$\sum_{j=1}^{J} \upsilon_j \cdot y_{ij}$, $i=1,\ldots,I$

Θα χρησιμοποιήσουμε το στατιστικό :

$$H = \frac{12}{n(n+1)\cdot\left[1 - \frac{\gamma}{n^3 - n}\right]} \cdot \sum_{i=1}^{I}\left(R_i - y_{i+} \cdot \frac{(n+1)}{2}\right)^2 \Big/ y_{i+}$$

όπου $\gamma = \sum_{l=1}^{J} \left(y_{+l}^3 - y_{+l} \right)$ και

$$R_i = y_{i1} \cdot \left(\frac{y_{+1}+1}{2} \right) + y_{i2} \cdot \left(y_{+1} + \frac{y_{+2}+1}{2} \right) + \ldots + y_{iJ} \cdot \left(y_{+1} + y_{+2} + \ldots + y_{+,J-1} + \frac{y_{+J}+1}{2} \right)$$

Το στατιστικό *H* είναι γνωστό ως το στατιστικό των *Kruskal-Wallis* ρυθμισμένο για δεσμούς (*ties*) (βλ. *Klotz & Teng* [36])

Θα έχουμε λοιπόν για τον αμφίπλευρο έλεγχο ότι p-value = $P_{H_0}\left(H \geq H_{obs} \middle| \{y_{i+}\}, \{y_{+j}\} \right) =$ $\sum_{S} f\left(\{t_{ij}\} \middle| \{y_{i+}\}, \{y_{+j}\} \right)$ με $S = \left\{ \{t_{ij}\} : H \geq H_{obs} \right\}$, όπου $f\left(\{t_{ij}\} \middle| \{y_{i+}\}, \{y_{+j}\} \right)$ από την (2.3).

Παρατήρηση

Εναλλακτικά, εκτός του *H* των *Kruskal-Wallis*, έχει προταθεί ως στατιστικό και το μέτρο συνάφειας γ (*gamma*) των *Goodman & Kruskal* (βλ. *Kruskal* [39]). Η δειγματική τιμή αυτού του στατιστικού είναι $\gamma = \frac{\mathrm{C-D}}{\mathrm{C+D}}$ όπου $\mathrm{C} = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij} \left(\sum_{h>i} \sum_{k>j} y_{hk} \right)$ είναι το σύνολο των σύμφωνων ζευγών (*concordant pairs*), ενώ $\mathrm{D} = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij} \left(\sum_{h>i} \sum_{k<j} y_{hk} \right)$ είναι το σύνολο των ασύμφωνων ζευγών (*disconcordant pairs*).

Θα έχουμε λοιπόν ότι p-value = $P_{H_0}\left(\gamma \geq \gamma_{obs} \middle| \{y_{i+}\}, \{y_{+j}\} \right) = \sum_{S} f\left(\{t_{ij}\} \middle| \{y_{i+}\}, \{y_{+j}\} \right)$ με $S = \left\{ \{t_{ij}\} : \gamma \geq \gamma_{obs} \right\}$, όπου $f\left(\{t_{ij}\} \middle| \{y_{i+}\}, \{y_{+j}\} \right)$ από την (2.3).

Το γ (*gamma*) αντιμετωπίζει τις X,Y συμμετρικά. Δηλαδή δεν ξεχωρίζει ποιά από τις X,Y είναι η διατάξιμη και επομένως μπορούμε να το εφαρμόσουμε και στην περίπτωση που η X είναι διατάξιμη και η Y ονοματική, όπως στην επόμενη παράγραφο.

2.3.3 Η μεταβλητή X διατάξιμη και η Y ονοματική

Αν η μεταβλητή X ήταν διατάξιμη και η Y ονοματική, θα εργαζόμασταν όπως και στην §2.3.2, με την μόνη διαφορά ότι ως εναλλακτικό του μοντέλου ανεξαρτησίας (2.2) θα ορίζαμε το μοντέλο M_4 : $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + u_i v_j$, $i=1,\ldots,I$, $j=1,\ldots,J$, με $u_1 < \ldots < u_I$ τα μονότονα scores που αναθέτουμε στις γραμμές του πίνακα, και $\{v_j\}$ οι παράμετροι. Το μοντέλο αυτό είναι γνωστό και ως μοντέλο επίδρασης στηλών (*column effect model*).

Προφανώς η αλήθεια της υπόθεσης H_0: $\nu_1 = \nu_2 = \ldots = \nu_J$ συνεπάγεται την ανεξαρτησία των μεταβλητών X, Y.

Για αναλυτικότερη παρουσίαση των μοντέλων συνάφειας linear-by-linear, row effect and column effect βλ. *Agresti* [1] §§9.4 και 9.5.

2.3.4 Παράδειγμα

Έστω ο 2x3 πίνακας συνάφειας που διασταυρώνει επίπεδο καπνίσματος με εμφάνιση εμφράγματος του μυοκαρδίου, σε ένα δείγμα νεαρών γυναικών για μια case-control μελέτη.

	Επίπεδο καπνίσματος (τσιγάρα ανα ημέρα)		
	0	1-24	>24
Control	25	25	12
Εμφράγμα μυοκαρδίου	0	1	3

Πηγή : S.Shapiro et.al., *Lancet*; **8119**: 743-746 (1979)

Το επίπεδο καπνίσματος (Y) είναι διατάξιμη μεταβλητή ενώ η X (control-έμφραγμα) είναι ονομαστική. Στόχος μας είναι να ελέγξουμε την ανεξαρτησία των μεταβλητών X,Y αλλά θα πρέπει να λάβουμε υπόψην μας και την διαταξιμότητα της Y. Για τον σκοπό αυτό θα χρησιμοποιήσουμε το στατιστικό γ (*gamma*) των *Goodman & Kruskal*. Η τιμή του στατιστικού γ (*gamma*) για τον παρατηρούμενο πίνακα είναι :

$$\gamma_{obs} = \frac{C - D}{C + D} = 0{,}872 \ , \text{ όπου } C = 25 \cdot (1+3) + 25 \cdot 3 = 175 \text{ και } D = 12 \cdot (0+1) + 25 \cdot 0 = 12$$

Δεσμεύοντας στην συνέχεια ως προς τα περιθώρια αθροίσματα y_{1+}, y_{2+}, y_{+1}, y_{+2} του παρατηρούμενου πίνακα, προκύπτουν 15 διαφορετικοί πίνακες με τα ίδια αυτά περιθώρια αθροίσματα. Από αυτούς μόνο δύο έχουν τιμή στατιστικού $\gamma \geq \gamma_{obs} = 0{,}872$. Αυτό συμβαίνει για τους πίνακες με συχνότητες κατά γραμμή (25,26,11/0,0,4) όπου γ =1 και (25,25,12/0,1,3) όπου γ = 0,872 .

Συνεπώς η p-value θα προκύψει από το άθροισμα των πιθανοτήτων (2.3) για τους δύο αυτούς πίνακες. Μετά από υπολογισμούς καταλήγουμε ότι p-value = $1{,}893 \cdot 10^{-3} + 0{,}016$ = 0,018. Συμπεραίνουμε λοιπόν ότι το επίπεδο καπνίσματος επηρρεάζει θετικά την εμφάνιση εμφράγματος μυοκαρδίου σε επίπεδο σημαντικότητας 5%.

Αν αδιαφορούσαμε για την διαταξιμότητα της μεταβλητής Y, και επιλέγαμε να χρησιμοποιήσουμε ως στατιστικό το X^2 του Pearson το οποίο δεν λαμβάνει υπόψην του την διαταξιμότητα, αυτό θα είχε ως συνέπεια να οδηγηθούμε σε λαναθασμένα συμπεράσματα.

Πράγματι ο παρατηρηθείς πίνακας έχει X^2_{obs} = 6,96. Από τους δεκαπέντε συνδυασμούς πινάκων με περιθώρια αθροίσματα ίδια με τα παρατηρούμενα, μόνο τέσσερεις δίνουν $X^2 \geq X^2_{obs} = 6{,}96$. Αυτό συμβαίνει για τους πίνακες με συχνότητες κατά γραμμή (25,26,11/0,0,4) όπου X^2=14,47 , (24,26,12/1,0,3) όπου X^2=6,983 , (21,26,15/4,0,0) όπου X^2=6,983 και (25,25,12/0,1,3) όπου X^2=6,96

Συνεπώς η p-value θα προκύψει από το άθροισμα των πιθανοτήτων (2.3) για τους τέσσερεις αυτούς πίνακες. Μετά από υπολογισμούς βρίσκουμε ότι p-value = $1{,}893 \cdot 10^{-3}$ + 0,01578+0,0175+0,016 = 0,052. Καταλήγουμε λοιπόν στο λανθασμένο συμπέρασμα της οριακής αποδοχής της υπόθεσης ανεξαρτησίας των X,Y, σε επίπεδο σημαντικότητας 5%.

2.4 Τετραγωνικοί *I*x*I* πίνακες συνάφειας

Όταν οι παρατηρήσεις μας καταγράφονται σε έναν διδιάστατο πίνακα συνάφειας όπου οι γραμμές (i=1,…I) και οι στήλες (j=1,…,I) έχουν το ίδιο πλήθος κατηγοριών, τότε μιλάμε για ανάλυση τετραγωνικών *I*x*I* πινάκων συνάφειας (*square contingency tables*). Αν επιπλέον οι μεταβλητές ταξινόμησης κατά γραμμές και στήλες είναι σύμμετρες, τότε έχουν και τις ίδιες κατηγορίες. Τα δεδομένα της κεντρικής διαγωνίου ενός τέτοιου πίνακα αντιστοιχούν σε περιπτώσεις με ταυτόσημες αποκρίσεις στην γραμμή και την στήλη του πίνακα, οπότε το ενδιαφέρον επικεντρώνεται στα μή διαγώνια κελλιά του πίνακα, δηλαδή σε αυτά που αντιπροσωπεύουν αλλαγή κατάστασης. Τα διαγώνια κελλιά αποτελούν δομικά μηδενικά για την μελέτη μας.

Ένα παράδειγμα τετραγωνικού πίνακα συνάφειας είναι ο παρακάτω 8x8 πίνακας, στον οποίο καταγράφονται οι γάμοι που έγιναν μεταξύ ευρωπαίων μεταναστών στις Ηνωμένες Πολιτείες το 1910 και διασταυρώνονται οι εθνικότητες των δύο συζύγων.

Εθνικότητα του συζύγου	Εθνικότητα της συζύγου							
	Αγγλική	Ιρλανδική	Σκανδιναβική	Γερμανική	Ιταλική	Πολωνική	Εβραική	
							κεντρική Ευρώπη	ανατολ. Ευρώπη
Αγγλική	314	63	10	15	0	1	1	0
Ιρλανδική	27	625	2	5	0	0	0	0
Σκανδιναβική	4	9	835	20	1	0	0	0
Γερμανική	26	26	10	1096	0	4	0	0
Ιταλική	3	6	0	4	477	1	0	0
Πολωνική	1	0	0	7	0	421	0	0
Εβραική (κεντρική Ευρώπη)	1	0	0	1	0	1	112	11
Εβραική (ανατολική Ευρώπη)	1	0	0	1	0	1	30	347

Πηγή : Pagnini & Morgan. *American Journal of Sociology* 1990; **96**: 405-432

Προφανώς, όλα τα λογαριθμογραμμικά μοντέλα που περιγράψαμε στην §2.3 για τον έλεγχο ανεξαρτησίας μπορούν να εφαρμοστούν σε τετραγωνικούς πίνακες, όμως επιπλέον ιδιαίτερο ενδιαφέρον παρουσιάζουν και τα μοντέλα ψευδοανεξαρτησίας (*quasi-independence*), συμμετρίας (*symmetry*) και ψευδοσυμμετρίας (*quasi-symmetry*).

Ένας τετραγωνικός πίνακας συνάφειας ικανοποιεί το μοντέλο της ψευδοανεξαρτησίας (QI) όταν οι μεταβλητές ταξινόμησης είναι ανεξάρτητες, δοθέντος όμως ότι η γραμμή και η στήλη θα έχουν διαφορετικό αποτελεσμα. Η λογαριθμογραμμική μορφή του είναι :

$$\log\left(\mu_{ij}\right)=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_{ij}^{XY}\cdot \mathrm{I}(i=j)\ \ ,\ \ i,j=1,...,I$$

όπου I(.) είναι η δείκτρια συνάρτηση με $\mathrm{I}(i=j)=\begin{cases}1 & ,i=j\\ 0 & ,i\neq j\end{cases}$.

Το QI μοντέλο έχει I περισσότερες παραμέτρους από τό μοντέλο της απλής ανεξαρτησίας, αφού μόνο οι όροι αλληλεπίδρασης για τα μή διαγώνια κελλιά θεωρούνται μηδέν, και έχει συνολικά $1+(I-1)+(I-1)+I$ μή πλεονάζουσες παραμέτρους.

Από τις εξισώσεις πιθανοφάνειας του QI μοντέλου

$$\hat{\mu}_{i+}=y_{i+},\ \ \hat{\mu}_{+j}=y_{+j}\ \ \text{και}\ \ \hat{\mu}_{ii}=y_{ii},\ \ i,j=1,...,I$$

παρατηρούμε ότι έχουμε τέλεια προσαρμογή στην κεντρική διαγώνιο, ενώ η ανεξαρτησία υφίσταται στα εναπομείναντα κελλιά. Δηλαδή η QI είναι μια μορφή ανεξαρτησίας, «δεσμευμένη» υπό τον περιορισμό της προσοχής μας, στα μή διαγώνια κελλιά του πίνακα.

Στην ουσία επιτρέπει στα $\{\mu_{ii}\}$ να παρεκλίνουν από το μοντέλο της απλής ανεξαρτησίας, και να παίρνουν αυθαίρετες θετικές τιμές.

Ένα ακριβές τέστ για τον έλεγχο της υπόθεσης της ψευδοανεξαρτησίας $H_0 : \lambda_{ij}^{XY} = 0$, $i \neq j$, $i, j = 1,...,I$, βασίζεται στην δεσμευμένη κατανομή των κελιών $\{y_{ij}\}$, δοθέντων των περιθώριων αθροισμάτων $\{y_{i+}\},\{y_{+j}\}$, $i, j = 1,...,I$ καθώς και των παρατηρούμενων τιμών στα διαγώνια κελλιά $\{y_{ii}\}$, $i = 1,...,I$. Οι ποσότητες ως προς τις οποίες δεσμεύουμε, είναι τα ελάχιστα επαρκή στατιστικά του μοντέλου QI.

Το μοντέλο της συμμετρίας (S) ικανοποιείται όταν $\mu_{ij} = \mu_{ji}$ για όλα τα $i \neq j$. Η ισοδύναμη λογαριθμογραμμική του έκφραση είναι :

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \quad , \quad i, j = 1,...,I$$

όπου $\lambda_i^X = \lambda_i^Y$, $i = 1,...,I$, και επιπλέον οι όροι αλληλεπίδρασης για τα συμμετρικά κελλιά γύρω από την κεντρική διαγώνιο είναι ίσοι $\lambda_{ij}^{XY} = \lambda_{ji}^{XY}$, $i, j = 1,...,I$. Έχει $1+(I-1)+\frac{I \cdot (I-1)}{2}$ μή πλεονάζουσες παραμέτρους, ενώ η επίλυση των εξισώσεων πιθανοφάνειας

$$\hat{\mu}_{ij} + \hat{\mu}_{ji} = y_{ij} + y_{ji}, \; i<j \text{ και } \hat{\mu}_{ii} = y_{ii} \;,$$

μας δίνει $\hat{\mu}_{ij} = \frac{y_{ij} + y_{ji}}{2}$, $i, j = 1,...,I$.

Το μοντέλο της συμμετρίας είναι ένα πολύ απλό μοντέλο, αλλά σε σπάνιες περιπτώσεις έχει καλή προσαρμογή. Συνήθως έχει καλή προσαρμογή ένα λιγότερο περιοριστικό μοντέλο το οποίο επιτρέπει στους όρους κύριας επίδρασης να διαφέρουν. Το μοντέλο που προκύπτει είναι γνωστό ως μοντέλο ψευδοσυμμετρίας (QS) και η λογαριθμογραμμική του έκφραση είναι :

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \quad , \quad i, j = 1,...,I$$

όπου $\lambda_{ij}^{XY} = \lambda_{ji}^{XY}$, $i, j = 1,...,I$. Έχει $1+(I-1)+(I-1)+\frac{I \cdot (I-1)}{2}$ μή πλεονάζουσες παραμέτρους, ενώ οι εξισώσεις πιθανοφάνειας είναι $\hat{\mu}_{i+} = y_{i+}$, $\hat{\mu}_{+j} = y_{+j}$, $\hat{\mu}_{ij} + \hat{\mu}_{ji} = y_{ij} + y_{ji}$, $i<j$ και $\hat{\mu}_{ii} = y_{ii}$, $i, j = 1,...,I$. Το QS μοντέλο είναι γνωστό και ως μοντέλο συμμετρικής συνάφειας, λόγω της ιδιότητας της συμμετρικότητας των τοπικών (*local*) odds ratios $\theta_{ij} = \theta_{ji}$, όπου

$\theta_{ij} = \frac{\mu_{ij} \cdot \mu_{i+1,j+1}}{\mu_{i+1,j} \cdot \mu_{i,j+1}}$, $i, j = 1,...,I-1$.

Ένα ακριβές τέστ για τον έλεγχο της υπόθεσης της ψευδοσυμμετρίας $H_0 : \lambda_{ij}^{XY} = \lambda_{ji}^{XY}$, $i \neq j$, $i, j = 1,...,I$, βασίζεται στην δεσμευμένη κατανομή των κελιών $\{y_{ij}\}$, δοθέντων των περιθώριων αθροισμάτων $\{y_{i+}\},\{y_{+j}\}$, $i, j = 1,...,I$, των παρατηρούμενων τιμών στα διαγώνια κελλιά $\{y_{ii}\}$, $i = 1,...,I$, καθώς και των αθροισμάτων των τιμών των συμμετρικών κελιών γύρω από την κεντρική διαγώνιο $\{y_{ij} + y_{ji}\}$, $i, j = 1,...,I$. Οι ποσότητες ως προς τις οποίες δεσμεύουμε είναι τα ελάχιστα επαρκή στατιστικά του μοντέλου QS.

Η υλοποίηση των ακριβών δεσμευμένων τέστ για τα προαναφερθέντα μοντέλα δεν είναι εύκολη υπόθεση. Η πλήρης απαρίθμηση των πινάκων που έχουν επαρκή στατιστικά για τις οχληρές παραμέτρους ίσα με αυτά του παρατηρούμενου πίνακα είναι συχνά αδύνατη, ενώ επίσης και η δειγματική δεσμευμένη κατανομή εξαρτάται από μια σταθερά κανονικοποίησης (*normalizing constant*) η οποία είναι δύσκολο να καθοριστεί. Στο θέμα αυτό θα επανέλθουμε αργότερα στο κεφάλαιο 3, όπου εκεί θα μιλήσουμε για τις εναλλακτικές προσεγγίσεις που έχουν προταθεί.

Για περισσότερες τεχνικές λεπτομέρειες σχετικά με τα μοντέλα QI, S, QS σε διδιάστατους τετραγωνικούς πίνακες, καθώς και το πώς αυτά γενικεύονται σε τριδιάστατους *I*x*I*x*I* τετραγωνικούς πίνακες συνάφειας, βλέπε *Bishop et. al.* [14] κεφ. 5 & 8.

2.5 Ακριβής συμπερασματολογία σε τριδιάστατους πίνακες συνάφειας

Προχωρόντας ένα ακόμα βήμα, θα μιλήσουμε για τριδιάστατους πίνακες συνάφειας και πώς μπορούμε να πραγματοποιήσουμε ελέγχους ανεξαρτησίας με χρήση των ακριβών μεθόδων.

Αρχικά θα αναφερθούμε στους τριδιάστατους 2x2x*K* πίνακες συνάφειας οι οποίοι συναντώνται για παράδειγμα σε πολυκλινικές μελέτες (*multiclinical trials*) όταν συγκρίνουμε μια δίτιμη απόκριση (Y) για δύο θεραπείες (X), χρησιμοποιώντας τα δεδομένα από *K* κλινικές (Z). Στην συνέχεια θα γενικεύσουμε τις μεθόδους και για τους τριδιάστατους *I*x*J*x*K* πίνακες.

Η πιο σημαντική υπόθεση που συνήθως επιθυμούμε να ελέγξουμε σε πίνακες αυτής της μορφής είναι η υπόθεση της δεσμευμένης ανεξαρτησίας των μεταβλητών X,Y δοθείσης της μεταβλητής Z. Δηλαδή ελέγχουμε αν οι μεταβλητές X,Y είναι ανεξάρτητες σε κάθε επίπεδο της μεταβλητής Z.

2.5.1 Δεσμευμένη ανεξαρτησία σε *2x2xK* πίνακες συνάφειας

Η δεσμευμένη ανεξαρτησία εκφράζεται μέσω του λογαριθμογραμμικού μοντέλου :

$$\mathrm{M}_0 : \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \ , \ i=1,2 \ , \ j=1,2 \ , \ k=1,\ldots,K$$

Τα ελάχιστα επαρκή στατιστικά του μοντέλου είναι οι περιθώριοι πίνακες $\{y_{i+k}\},\{y_{+jk}\}$ $\forall\ i,j,k$.

Θα ελέγξουμε το μοντέλο της δεσμευμένης ανεξαρτησίας έναντι του μοντέλου που δεν περιλαμβάνει τον όρο της τριπλής αλληλεπίδρασης και άρα υπονοεί ομοιογενή συνάφεια για κάθε ζεύγος των μεταβλητών ταξινόμησης

$$\mathrm{M}_1 : \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \ , \ \forall\ i,j,k.$$

Τα ελάχιστα επαρκή στατιστικά του M_1 είναι οι περιθώριοι πίνακες $\{y_{i+k}\},\{y_{+jk}\}$, $\{y_{ij+}\}$, $\forall\ i,j,k.$

Η δεσμευμένη κατανομή που μας ενδιαφέρει είναι η κατανομή του $\{y_{ij+}\}$ περιθώριου πίνακα, δοθέντων των $\{y_{i+k}\},\{y_{+jk}\}$ περιθώριων πινάκων. Στην περίπτωση των 2x2x*K* πινάκων η κατανομή αυτή απλοποιείται στην κατανομή του $\mathrm{T}=\sum_{k=1}^{K} y_{11k}$, δοθέντων των $\{y_{1+k}, y_{2+k}, y_{+1k}, y_{+2k}\}$ για k=1,…,K (βλ. *Zelen* [64]).

Τα $\{y_{11k},\ k=1,\ldots,K\}$ έχουν ανεξάρτητες μη κεντρικές υπεργεωμετρικές κατανομές, καθεμία της μορφής (2.1). Συνεπώς η δεσμευμένη κατανομή του αθροίσματός τους $f\left(\mathrm{T}=\sum_{k=1}^{K} y_{11k} \,\middle|\, \{y_{1+k}\},\{y_{+1k}\},\{y_{2+k}\},\{y_{+2k}\};\theta\right)$, θα καθορίζεται απο το γινόμενο των *K* αυτών συναρτήσεων πυκνότητας πιθανότητας, και δίνεται απο την σχέση :

$$\sum_{S_T}\left[\prod_{k=1}^{K} P\left(y_{11k}=t_k \,\middle|\, \{y_{1+k}\},\{y_{+1k}\},\{y_{2+k}\},\{y_{+2k}\};\theta\right)\right] =$$

$$= \sum_{S_T}\left[\prod_{k=1}^{K}\left(\frac{\binom{y_{1+k}}{t_k}\cdot\binom{y_{++k}-y_{1+k}}{y_{+1k}-t_k}\cdot\theta^{t_k}}{\sum_{u_k=u_{k(\min)}}^{u_{k(\max)}}\binom{y_{1+k}}{u_k}\cdot\binom{y_{++k}-y_{1+k}}{y_{+1k}-u_k}\cdot\theta^{u_k}}\right)\right] \qquad (2.6)$$

όπου S_T ={το σύνολο των πινάκων που έχουν τα ίδια $\{y_{1+k}\},\{y_{+1k}\},\{y_{2+k}\},\{y_{+2k}\}$ με τον παρατηρούμενο πίνακα, και επιπλέον έχουν $\sum_{k=1}^{K} t_k = \mathrm{T}\}$, $u_{k(\min)} = \max\left(0, y_{+1k} - y_{2+k}\right)$ και $u_{k(\max)} = \min\left(y_{1+k}, y_{+1k}\right)$.

Για τον έλεγχο λοιπόν της υπόθεσης της δεσμευμένης ανεξαρτησίας, η οποία σε όρους *odds ratios* εκφράζεται ώς $\mathrm{H}_0 : \theta_{XY(1)} = \theta_{XY(2)} = \ldots = \theta_{XY(K)} = \theta = 1$ κατά H_1: $\theta > 1$ η p-value θα είναι ίση με $P_{H_0}\left(\mathrm{T} = \sum_k y_{11k} \geq \mathrm{T}_{\mathrm{obs}} \,\middle|\, \{y_{1+k}\},\{y_{+1k}\},\{y_{2+k}\},\{y_{+2k}\}; \theta = 1\right) =$ $\sum_S f\left(\mathrm{T} \middle| \{y_{1+k}\},\{y_{+1k}\},\{y_{2+k}\},\{y_{+2k}\}\right)$, όπου *S*={το σύνολο των πινάκων που έχουν τα ίδια $\{y_{1+k}\},\{y_{+1k}\},\{y_{2+k}\},\{y_{+2k}\}$ με τον παρατηρούμενο πίνακα, και επιπλέον έχουν $\mathrm{T} \geq \mathrm{T}_{\mathrm{obs}}$}.Το τέστ αυτό είναι γνωστό και ως η exact εκδοχή του *Cohran-Mantel-Haenszel* τέστ (βλ.[43a],[44]).

Βέβαια σε μερικούς πίνακες είναι μη ρεαλιστικό να περιμένουμε ότι τα odds ratios $\{\theta_{XY(k)}$, k=1,…,$K\}$ θα είναι ίσα σε κάθε επίπεδο της μεταβλητής Z. Σε τέτοιες περιπτώσεις είναι προτιμότερο ως εναλλακτικό μοντέλο του M_0 να επιλέξουμε το κορεσμένο μοντέλο, το οποίο εκφράζει ετερογενή συνάφεια των μεταβλητών X,Y (δηλαδή τα *K* το πλήθος θ_{XY} odds ratios διαφέρουν) :

$$\mathrm{M}_2 : \log\left(\mu_{ijk}\right) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \quad , \forall\, i,j,k$$

Στην περίπτωση αυτή το score στατιστικό που θα χρησιμοποιήσουμε είναι το $\mathrm{T}_1 = \sum_{k=1}^{K} \frac{y_{++k} - 1}{y_{++k}} \cdot X_k^2$ όπου X_k^2 είναι το X^2 του Pearson για τον έλεγχο της ανεξαρτησίας των μεταβλητών X,Y στο επίπεδο *k* της μεταβλητής Z.

Συνεπώς p-value = $P_{H_0}\left(\mathrm{T}_1 \geq \mathrm{T}_{1,\mathrm{obs}}\right) = \sum_S f\left(\mathrm{T} \middle| \{y_{1+k}\},\{y_{+1k}\},\{y_{2+k}\},\{y_{+2k}\}\right)$ όπου *S*={το σύνολο των πινάκων που έχουν τα ίδια $\{y_{1+k}\},\{y_{+1k}\},\{y_{2+k}\},\{y_{+2k}\}$ με τον παρατηρούμενο πίνακα, και επιπλέον έχουν $\mathrm{T}_1 \geq \mathrm{T}_{1,\mathrm{obs}}$}.

2.5.1.1 Παράδειγμα

Ο παρακάτω πίνακας συνάφειας αναφέρεται σε υπαλλήλους της αμερικανικής κυβέρνησης οι οποίοι είχαν τις ίδιες προοπτικές για προαγωγή, και καταγράψαμε πόσοι από αυτούς προάχθηκαν στην διάρκεια τριών μηνών. Συγκεκριμένα διασταυρώνονται η απόφαση για προαγωγή (μεταβλητή Y) και η φυλή του υπαλλήλου (μεταβλητή X) για τρείς διαφορετικούς μήνες (μεταβλητή Z).

	Προαγωγές Ιουλίου		Προαγωγές Αυγούστου		Προαγωγές Σεπτεμβριού	
Φυλή	Ναι	Όχι	Ναι	Όχι	Ναι	Όχι
Μαύρος	0	7	0	7	0	8
Λευκός	4	16	4	13	2	13

Πηγή : J.Gastwirth, *Statistical Reasoning in Law and Public Policy* 1988; **1**: 266

Υποθέτοντας ομοιόγενεια των odds ratios, θα ελέγξουμε την υπόθεση της δεσμευμένης ανεξαρτησίας της φυλής με την απόφαση για προαγωγή. Δηλαδή θα ελέγξουμε την υπόθεση $H_0: \theta_{XY(1)} = \theta_{XY(2)} = \theta_{XY(3)} = \theta = 1$ έναντι της εναλλακτικής $H_1: \theta < 1$ ότι η πιθανότητα προαγωγής ήταν χαμηλότερη για τους μαύρους υπαλλήλους από ότι ήταν για τους λευκούς υπαλλήλους.

Δοθέντων των περιθώριων αθροισμάτων των τριών υποπινάκων, η συχνότητα y_{111} μπορεί να κυμαίνεται μεταξύ 0 και 4, η y_{112} μπορεί να κυμαίνεται μεταξύ 0 και 4, και η συχνότητα y_{113} μπορεί να κυμαίνεται μεταξύ 0 και 2. Ως εκ τούτου το στατιστικό $T=\sum_{k=1}^{3} y_{11k}$ που μας ενδιαφέρει μπορεί να πάρει τιμές μεταξύ 0 και 10. Η παρατηρούμενη τιμή του στατιστικού είναι $T_{obs} = 0$.

Η p-value λοιπόν θα είναι :

$$P_{H_0}\left(T \le 0 \,\middle|\, \{y_{1+k}\}, \{y_{2+k}\}, \{y_{+1k}\}, \{y_{+2k}\}\right) = \sum_{S} f\left(T \,\middle|\, \{y_{1+k}\}, \{y_{2+k}\}, \{y_{+1k}\}, \{y_{+2k}\}\right) \quad (2.7)$$

όπου *S*={το σύνολο των πινάκων που έχουν τα ίδια περιθώρια αθροίσματα $y_{1+k}, y_{2+k}, y_{+1k}, y_{+2k}, k = 1,2,3$, με τον παρατηρούμενο πίνακα, και επιπλέον έχουν $T \le 0$ }, και η $f\left(T \,\middle|\, \{y_{1+k}\}, \{y_{2+k}\}, \{y_{+1k}\}, \{y_{+2k}\}\right)$ υπολογίζεται από την (2.6).

Παρατηρούμε όμως ότι εκτός από τον παρατηρούμενο πίνακα, κανένας άλλος πίνακας δεν έχει τιμή στατιστικού T≤0. Άρα η p-value από την (2.7) θα είναι :

$$\text{p-value} = \frac{\binom{7}{0}\cdot\binom{20}{4}}{\sum_{k=0}^{4}\binom{7}{k}\cdot\binom{20}{4-k}} \cdot \frac{\binom{7}{0}\cdot\binom{17}{4}}{\sum_{k=0}^{4}\binom{7}{k}\cdot\binom{17}{4-k}} \cdot \frac{\binom{8}{0}\cdot\binom{15}{2}}{\sum_{k=0}^{2}\binom{8}{k}\cdot\binom{15}{2-k}} = 0{,}026$$

δηλαδή σε ε.σ. 5% συμπεραίνουμε ότι οι λευκοί υπάλληλοι είχαν μεγαλύτερη πιθανότητα προαγωγής από τους μαύρους συναδέλφους τους.

2.5.2 Δεσμευμένη ανεξαρτησία σε *I*x*J*x*K* πίνακες συνάφειας

Στην παράγραφο αυτή θα παρουσιάσουμε τα στατιστικά που μπορούν να χρησιμοποιηθούν για να ελέγξουμε την υπόθεση της δεσμευμένης ανεξαρτησίας (βλ. *Kim & Agresti* [34]). Πρόκειται για τα score στατιστικά που πρώτος εισήγαγε ο *Birch* [13].

Πρακτικά την υπόθεση της δεσμευμένης ανεξαρτησίας η οποία αντιστοιχεί στο λογαριθμογραμμικό μοντέλο

$$\mathrm{M}_0 : \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \ , \ i=1,\ldots,I \ , \ j=1,\ldots,J \ , \ k=1,\ldots,K \quad (2.8)$$

την ελέγχουμε έναντι του εναλλακτικού μοντέλου

$$\mathrm{M}_1 : \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \ , \forall\, i,j,k.$$

Το μοντέλο M_1 δεν περιλαμβάνει τον όρο της τριπλής αλληλεπίδρασης, και συνεπώς υπονοεί ομοιογενή συνάφεια των μεταβλητών X,Y , δηλαδή $\theta_{ij(1)} = \theta_{ij(2)} = \ldots = \theta_{ij(K)}$, όπου

$\theta_{ij(k)} = \dfrac{\mu_{ijk}\cdot\mu_{i+1,j+1,k}}{\mu_{i,j+1,k}\cdot\mu_{i+1,j,k}}$, $i=1,\ldots I\text{-}1$, $j=1,\ldots,J\text{-}1$, $k=1,\ldots,K$ τα τοπικά (*local*) odds ratios.

Στις επόμενες τρεις παραγράφους θα παρουσιάσουμε τρία score στατιστικά για τον έλεγχο αυτό. Το ένα αντιμετωπίζει τις μεταβλητές X και Y ως ονοματικές, το άλλο ως διατάξιμες και το τελευταίο τέστ αντιμετωπίζει την X ως ονοματική και την Y ως διατάξιμη.

Στην §2.5.3 θα αναφερθούμε σε εναλλακτικά τέστ τα οποία χρησιμοποιούνται περιστασιακά όταν η συνάφεια των X,Y είναι αρκετά ετερογενής. Στην περίπτωση αυτή ως εναλλακτικό μοντέλο του M_0 είναι προτιμότερο το κορεσμένο μοντέλο.

2.5.2.1 Οι μεταβλητές X και Y είναι ονοματικές

Η δεσμευμένη ανεξαρτησία όπως είδαμε αντιστοιχεί στο λογαριθμογραμμικό μοντέλο (2.8). Θα ελέγξουμε το μοντέλο αυτό έναντι του μοντέλου M_1 το οποίο εκφράζει την ομοιογένεια των odds ratios στα K επίπεδα της μεταβλητής Z. Τα ελάχιστα επαρκή στατιστικά του (2.8) είναι τα $\{y_{i+k}\},\{y_{+jk}\},\forall\, i,j,k,$ ενώ για το μοντέλο M_1 είναι τα $\{y_{i+k}\},\{y_{+jk}\},\{y_{ij+}\},\forall\, i,j,k.$

Η δεσμευμένη κατανομή που μας ενδιαφέρει, είναι η κατανομή του περιθώριου πίνακα $\{y_{ij+}\}$ δοθέντων των $\{y_{i+k}\},\{y_{+jk}\}$, υπό την μηδενική υπόθεση της δεσμευμένης ανεξαρτησίας η οποία ισοδυναμεί με την υπόθεση $H_0: \theta_{ij(1)} = \theta_{ij(2)} = \ldots = \theta_{ij(K)} = \theta = 1.$

Η δεσμευμένη κατανομή αποτελεί γινόμενο από ανεξάρτητες γενικευμένες υπεργεωμετρικές κατανομές για τα διάφορα στρώματα της Z, και δεν εξαρτάται από οχληρές παραμέτρους. Ο *Birch*[13] έδειξε ότι η κατανομή αυτή έχει πιθανότητες ανάλογες της ποσότητας $\left(\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} y_{ijk}!\right)^{-1}$.

Το στατιστικό που θα χρησιμοποιήσουμε για να πραγματοποιήσουμε τον έλεγχο της υπόθεσης $H_0: \theta_{ij(1)} = \theta_{ij(2)} = \ldots = \theta_{ij(K)} = \theta = 1$ θα είναι το score στατιστικό το οποίο ορίζεται ως εξής :

Έστω $\mathbf{y}_k = \left(y_{11k}, y_{12k}, \ldots, y_{1,J-1,k}, \ldots, y_{I-1,J-1,k}\right)'$, k=1,…,K , το διάνυσμα των συχνοτήτων των non-redundant κελιών του πίνακα για το k επίπεδο της Z, και $\boldsymbol{\mu}_k = E\left(\mathbf{y}_k\right)$ οι αντίστοιχες αναμενόμενες συχνότητες των κελιών αυτών. Συγκεκριμένα θα είναι

$$\boldsymbol{\mu}_k = \left(y_{1+k}\cdot y_{+1k}, y_{1+k}\cdot y_{+2k}, \ldots, y_{I-1,+,k}\cdot y_{+,J-1,k}\right)' \Big/ y_{++k}\ ,\ k=1,\ldots,K.$$

Αν με $\mathbf{V}_k$ συμβολίσουμε τον πίνακα συνδιακυμάνσεων του $\mathbf{y}_k$, υπό την μηδενική υπόθεση, με στοιχεία

$$\operatorname{cov}\left(y_{ijk}, y_{i'j'k}\right) = \frac{y_{i+k}\left(\delta_{ii'}\cdot y_{++k} - y_{i'+k}\right)\cdot y_{+jk}\cdot\left(\delta_{jj'}\cdot y_{++k} - y_{+j'k}\right)}{y_{++k}^2\cdot\left(y_{++k}-1\right)}$$

όπου $\delta_{ab} = \begin{cases} 1, \text{αν a = b} \\ 0, \text{ διαφορ.} \end{cases}$ είναι το δέλτα του *Kronecker*, και τέλος αθροίσουμε ως προς τα K επίπεδα της μεταβλητής Z έτσι ώστε $\mathbf{y} = \sum_{k=1}^{K}\mathbf{y}_k$, $\boldsymbol{\mu} = \sum_{k=1}^{K}\boldsymbol{\mu}_k$, $\mathbf{V} = \sum_{k=1}^{K}\mathbf{V}_k$, τότε το score

στατιστικό είναι $T_N = (\mathbf{y}-\boldsymbol{\mu})' \cdot \mathbf{V}^{-1} \cdot (\mathbf{y}-\boldsymbol{\mu})$. Ασυμπτωτικά το στατιστικό αυτό ακολουθεί χ^2-κατανομή με $(I-1)(J-1)$ βαθμούς ελευθερίας.

Αν $K=1$ τότε το στατιστικό αυτό απλοποιείται στο $T_N = \frac{y_{++}-1}{y_{++}} \cdot X^2_{Pearson}$. Η ακριβής p-value για τον δεξιόπλευρο έλεγχο $H_0: \theta_{ij(1)} = \theta_{ij(2)} = \ldots = \theta_{ij(K)} = \theta = 1$ κατα $H_1: \theta > 1$ θα είναι :

$$\text{p-value} = P_{H_0}\left(T_N \geq T_{N,obs} \,\middle|\, \{y_{i+k}\}, \{y_{+jk}\}\right) = \frac{\sum_{S_1}\left(\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} y_{ijk}!\right)^{-1}}{\sum_{S}\left(\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} y_{ijk}!\right)^{-1}} \qquad (2.9)$$

όπου S={το σύνολο των πινάκων οι οποίοι έχουν τα ίδια περιθώρια αθροίσματα $\{y_{i+k}\}, \{y_{+jk}\}$ με τον παρατηρούμενο πίνακα } και S_1= {το σύνολο των πινάκων που ανήκουν στο S και επιπλέον έχουν $T_N \geq T_{N,obs}$ }.

2.5.2.2 Οι μεταβλητές X και Y είναι διατάξιμες

Όταν οι μεταβλητές X,Y είναι διατάξιμες, θα επιλέξουμε ως εναλλακτικό του μοντέλου δεσμευμένης ανεξαρτησίας (2.8) ένα πιο περιοριστικό μοντέλο από το M_1, το οποίο θα λαμβάνει υπόψην του την διάταξη των μεταβλητών, όπως κάναμε και στους IxJ πίνακες συνάφειας. Με αυτόν τον τρόπο επιτυγχάνουμε αύξηση της ισχύς του τέστ, αφού εστιάζουμε την αλληλεπίδραση των μεταβλητών X,Y σε λιγότερους βαθμούς ελευθερίας.

Το μοντέλο λοιπόν που θα χρησιμοποιήσουμε θα είναι το M_2: $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta \cdot u_i \cdot \upsilon_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$, $\forall i,j,k$, το οποίο προκύπτει από το μοντέλο M_1, αν ο γενικός όρος της αλληλεπίδρασης $\{\lambda_{ij}^{XY}\}$ αντικατασταθεί από τον linear-by-linear όρο $\{\beta \cdot u_i \cdot \upsilon_j\}$. Με $u_1 < \ldots < u_I$ και $\upsilon_1 < \ldots < \upsilon_J$ συμβολίζουμε τα μονότονα scores που αναθέτουμε στις γραμμές και τις στήλες του πίνακα αντίστοιχα.

Η υπόθεση της δεσμευμένης ανεξαρτησίας των μεταβλητών X,Y δοθείσης της Z, ταυτίζεται με την υπόθεση H_0: $\beta=0$ (αφού τα local odds ratios ικανοποιούν την σχέση $\log\theta_{ij(k)} = \beta \cdot (u_{i+1} - u_i) \cdot (\upsilon_{j+1} - \upsilon_j)$, $k = 1, \ldots, K$). Τα ελάχιστα επαρκή στατιστικά του M_2 είναι τα $\{y_{i+k}\}, \{y_{+jk}\}, \left\{\sum_{i=1}^{I}\sum_{j=1}^{J} u_i \cdot \upsilon_j \cdot y_{ij+}\right\}, \forall\, i,j,k.$

Η δεσμευμένη κατανομή που μας ενδιαφέρει, είναι η κατανομή του $\left\{\sum_{i=1}^{I}\sum_{j=1}^{J} u_i \cdot \upsilon_j \cdot y_{ij+}\right\}$ δοθέντων των $\{y_{i+k}\},\{y_{+jk}\}$, υπο την μηδενική υπόθεση της δεσμευμένης ανεξαρτησίας H_0: $\beta=0$.

Ο *Mantel* [43a] έδειξε ότι η μέση τιμή και η διακύμανση είναι

$$E\left(\sum_{i=1}^{I}\sum_{j=1}^{J} u_i \cdot \upsilon_j \cdot y_{ij+}\right) = \frac{\left(\sum_{i=1}^{I} u_i \cdot y_{i+k}\right)\cdot\left(\sum_{j=1}^{J} \upsilon_j \cdot y_{+jk}\right)}{y_{++k}}$$

$$Var\left(\sum_{i=1}^{I}\sum_{j=1}^{J} u_i \cdot \upsilon_j \cdot y_{ij+}\right) = \frac{1}{y_{++k}-1}\cdot\left[\sum_{i=1}^{I} u_i^2 \cdot y_{i+k} - \frac{\left(\sum_{i=1}^{I} u_i \cdot y_{i+k}\right)^2}{y_{++k}}\right]$$

$$\cdot\left[\sum_{j=1}^{J} \upsilon_j^2 \cdot y_{+jk} - \frac{\left(\sum_{j=1}^{J} \upsilon_j \cdot y_{+jk}\right)^2}{y_{++k}}\right]$$

και πρότεινε ως στατιστικό για τον έλεγχο H_0: $\beta=0$ κατά H_1: $\beta>0$, το score στατιστικό :

$$T_O = \frac{\left\{\sum_{k=1}^{K}\left[\sum_{i=1}^{I}\sum_{j=1}^{J} u_i \cdot \upsilon_j \cdot y_{ijk} - E\left(\sum_{i=1}^{I}\sum_{j=1}^{J} u_i \cdot \upsilon_j \cdot y_{ijk}\right)\right]\right\}^2}{\sum_{k=1}^{K} Var\left(\sum_{i=1}^{I}\sum_{j=1}^{J} u_i \cdot \upsilon_j \cdot y_{ijk}\right)}$$

το οποίο ασυμπτωτικά ακολουθεί την χ^2-κατανομή με έναν βαθμό ελευθερίας. Η ακριβής p-value για τον έλεγχο υπολογίζεται από την σχέση (2.9) μόνο που παίρνουμε ως S_1 ={το σύνολο των πινάκων που ανήκουν στο S και επιπλέον έχουν $T_O \geq T_{O,\,obs}$ }.

2.5.2.3 Η μεταβλητή X ονοματική και η Y διατάξιμη

Στην περίπτωση αυτή ως εναλλακτικό του μοντέλου δεσμευμένης ανεξαρτησίας (2.8), θα χρησιμοποιήσουμε το

$$\mathrm{M}_3: \ \log\left(\mu_{ijk}\right)=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+c_i\cdot\upsilon_j+\lambda_{ik}^{XZ}+\lambda_{jk}^{YZ}\ ,\ \forall\, i,j,k\ ,$$

το οποίο προκύπτει αντικαθιστώντας τις διατάξιμες τιμές $\{\beta\cdot u_i\}$ στον όρο $\{\beta\cdot u_i\cdot\upsilon_j\}$ του μοντέλου M_2, με τις παραμέτρους $\{c_i\}$. Με $\upsilon_1<\ldots<\upsilon_J$ συμβολίζουμε τα μονότονα scores που αναθέτουμε στις στήλες του πίνακα. Τα ελάχιστα επαρκή στατιστικά του μοντέλου είναι τα $\{y_{i+k}\},\{y_{+jk}\},\forall\, i,j,k$, και $\left\{\sum_{j=1}^{J}\upsilon_j\cdot y_{ij+}\ ,\ i=1,\ldots.I\right\}$.

Για να ελέγξουμε λοιπόν την υπόθεση της δεσμευμένης ανεξαρτησίας, η οποία ταυτίζεται με την υπόθεση H_0: $c_1=c_2=\ldots=c_I$ (αφού τα τοπικά odds ratios ικανοποιούν την σχέση $\log\theta_{ij(k)}=\left(c_{i+1}-c_i\right)\cdot\left(\upsilon_{j+1}-\upsilon_j\right),\ k=1,\ldots,K$), θα χρησιμοποιήσουμε το score στατιστικό το οποίο προκύπτει ως εξής:

Έστω $\boldsymbol{\zeta}$ το (I-1)x1 διάνυσμα με στοιχεία $\zeta_i=\sum_{k=1}^{K}y_{i+k}\cdot\left(\bar{y}_{ik}-\bar{y}_k\right),\ i=1,\ldots,I-1$ όπου $\bar{y}_{ik}=\sum_{j=1}^{J}y_{ijk}\cdot\upsilon_j\Big/y_{i+k}$ και $\bar{y}_k=\sum_{i=1}^{I}\sum_{j=1}^{J}y_{ijk}\cdot\upsilon_j\Big/y_{++k}$.

Αν με **A** συμβολίσουμε τον πίνακα συνδιακυμάνσεων του $\boldsymbol{\zeta}$, υπό την μηδενική υπόθεση, με στοιχεία :

$$\mathrm{cov}\left(\zeta_i,\zeta_{i'}\right)=\sum_{k=1}^{K}\left[\frac{y_{i+k}\cdot\left(\delta_{ii'}\cdot y_{++k}-y_{i'+k}\right)}{y_{++k}\cdot\left(y_{++k}-1\right)}\cdot\sum_{j=1}^{J}y_{+jk}\cdot\left(\upsilon_j-\bar{y}_k\right)^2\right],\ \text{όπου}\ \delta_{ab}=\begin{cases}1,\ \text{αν a = b}\\0,\ \text{διαφορ.}\end{cases}$$

τότε το score στατιστικό είναι $\mathrm{T_{NO}}=\boldsymbol{\zeta}'\cdot\mathbf{A}^{-1}\cdot\boldsymbol{\zeta}$. Ασυμπτωτικά το στατιστικό αυτό ακολουθεί χ^2-κατανομή με (I-1) βαθμούς ελευθερίας.

Η ακριβής p-value για τον έλεγχο H_0: $c_1=\ldots=c_I$ κατά H_1: *διαφορετικά*, υπολογίζεται από την σχέση (2.9) μόνο που παίρνουμε ως S_1 = {το σύνολο των πινάκων που ανήκουν στο S και επιπλέον έχουν $\mathrm{T_{NO}}\geq\mathrm{T_{NO,\,obs}}$ }.

2.5.3 Δεσμευμένη ανεξαρτησία σε *I*x*J*x*K* πίνακα συνάφειας, όταν υποθέτουμε ετερογένεια της συνάφειας των X,Y μεταβλητών

Σε μερικές εφαρμογές, συχνά αναμένουμε την συνάφεια των μεταβλητών X,Y να διαφέρει σημαντικά στα διάφορα επίπεδα της μεταβλητής Z. Αν αυτό αληθεύει, είναι λάθος να εφαρμόσουμε τα στατιστικά που παρουσιάσαμε στην προηγούμενη ενότητα. Πρέπει να

βασίσουμε την μελέτη μας σε μοντέλα που περιλαμβάνουν τον όρο αλληλεπίδρασης τριών παραγόντων.

Λαμβάνοντας λοιπόν υπόψην μας της ετερογένεια της XY συνάφειας, θα παρουσιάσουμε τρία score στατιστικά όπου το ένα αντιμετωπίζει τις μεταβλητές X και Y ως ονοματικές, το άλλο ως διατάξιμες και το τελευταίο στατιστικό αντιμετωπίζει την X ως ονοματική και την Y ως διατάξιμη. Τα στατιστικά αυτά έχουν την ίδια δομή $\mathrm{T}^* = \sum_{k=1}^{K} \mathrm{T}(k)$, όπου $\mathrm{T}(k)$ είναι ένα από τα τρία στατιστικά που αναλύσαμε στην προηγούμενη παράγραφο §2.5.2, εφαρμοσμένο στο επίπεδο *k* της μεταβλητής Z. Οι ακριβείς p-values θα υπολογίζονται από την (2.9), μόνο που παίρνουμε ως S_1 = {το σύνολο των πινάκων που ανήκουν στο *S* και επιπλέον έχουν $\mathrm{T}^* \geq \mathrm{T}^*_{obs}$}.

Όταν οι X,Y είναι ονοματικές, ως εναλλακτική υπόθεση του μοντέλου δεσμευμένης ανεξαρτησίας (2.8), θα χρησιμοποιήσουμε το κορεσμένο μοντέλο

$$\mathrm{M}_4 : \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \quad \forall\, i,j,k.$$

Αν με X_k^2 συμβολίσουμε το X^2 του Pearson για τον έλεγχο ανεξαρτησίας στο επίπεδο *k* της μεταβλητής Z, τότε το δεσμευμένο score στατιστικό είναι $\mathrm{T}_N^* = \sum_{k=1}^{K} \frac{y_{++k} - 1}{y_{++k}} \cdot X_k^2$. Ασυμπτωτικά το στατιστικό αυτό ακολουθεί χ^2-κατανομή με df=$K(I\text{-}1)(J\text{-}1)$

Στην περίπτωση που οι X,Y είναι διατάξιμες αναμένουμε μια μονότονη συνάφεια μεταξύ τους, η οποία θα αλλάζει στα διάφορα επίπεδα της μεταβλητής Z. Ως εναλλακτικό του μοντέλου (2.8), προτείνεται ένα πιο περιοριστικό μοντέλο από το κορεσμένο, το οποίο θα λαμβάνει υπόψην του εκτός από την ετερογένεια και την διάταξη των μεταβλητών. Ένα τέτοιο μοντέλο είναι το

$$\mathrm{M}_4 : \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_k \cdot u_i \cdot \upsilon_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \ \forall\, i,j,k,$$

το οποίο είναι γνωστό και ως ετερογενές linear-by-linear μοντέλο συνάφειας. Το μοντέλο αυτό είναι σαν να προσαρμόζουμε το linear-by-linear μοντέλο (2.4) σε κάθε επίπεδο της Z, χωριστά. Το επαρκές στατιστικό για την παράμετρο $\{\beta_k, k = 1,\ldots,K\}$ που μας ενδιαφέρει, είναι το $\left\{ \sum_{i=1}^{I} \sum_{j=1}^{J} u_i \cdot \upsilon_j \cdot y_{ijk},\ k = 1,\ldots,K \right\}$.

Το score στατιστικό που θα χρησιμοποιήσουμε για τον έλεγχο της δεσμευμένης ανεξαρτησίας H_0 : $\beta_1=\beta_2=\ldots=\beta_k=0$ κατά H_1: *διαφορετικά*, προκύπτει ώς μια τεραγωνική

μορφή, βασισμένη στο διάνυσμα $\mathbf{r}_{Kx1}$ με στοιχεία $r_k = \sum_{i=1}^{I}\sum_{j=1}^{J} u_i \cdot \upsilon_j \cdot \left(y_{ijk} - \frac{y_{i+k} \cdot y_{+jk}}{y_{++k}} \right)$, και παίρνει την απλουστευμένη μορφή $\mathrm{T}_{\mathrm{O}}^{*} = \sum_{k=1}^{K} \mathrm{T}_{\mathrm{O}}(k)$, όπου $\mathrm{T}_{\mathrm{O}}(k)$ συμβολίζει το στατιστικό T_{O} υπολογισμένο μόνο για το στρώμα *k*.

Τέλος όταν η X είναι ονοματική και η Y διατάξιμη, ως εναλλακτικό του μοντέλου (2.8) θα χρησιμοποιήσουμε ένα μοντέλο το οποίο να επιτρέπει την ετερογένεια των X,Y στα διάφορα στρώματα της Z. Ενα τέτοιο είναι το M_5 : $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + c_{ik} \cdot \upsilon_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$ $\forall\, i,j,k$ όπου οι παράμετροι των γραμμών $\{c_{ik}, k = 1,\ldots,K\}$ διαφέρουν σε κάθε επίπεδο της μεταβλητής Z, ενώ τα μονότονα scores των στηλών $\upsilon_1 < \ldots < \upsilon_J$ είναι γνωστά.

Για να ελέγξουμε λοιπόν την υπόθεση της δεσμευμένης ανεξαρτησίας, η οποία ταυτίζεται με την υπόθεση $\mathrm{H}_0: c_{1k} = c_{2k} = \ldots = c_{Ik}$ κατά H_1: *διαφορετικά*, $k = 1,\ldots,K$, θα χρησιμοποιήσουμε το score στατιστικό το οποίο προκύπτει ως μία τετραγωνική μορφή, βασισμένη στο διάνυσμα $\mathbf{q}_{K(I-1)x1}$ με στοιχεία $q_{ik} = \sum_{j=1}^{J} \upsilon_j \cdot \left(y_{ijk} - \frac{y_{i+k} \cdot y_{+jk}}{y_{++k}} \right)$, i=1,…,I-1, k=1,…,K, και παίρνει την απλουστευμένη μορφή $\mathrm{T}_{\mathrm{NO}}^{*} = \sum_{k=1}^{K} \mathrm{T}_{\mathrm{NO}}(k)$, όπου $\mathrm{T}_{\mathrm{NO}}(k)$ συμβολίζει το στατιστικό T_{NO} υπολογισμένο για το *k* στρώμα μόνο.

Ανάλογα θα εργαζόμασταν αν η X ήταν διατάξιμη και η Y ονοματική. Τότε το εναλλακτικό μοντέλο που επιλέγουμε, εκφράζει την XY αλληλεπίδραση με τον όρο $\{u_i \cdot v_{jk}\}$, όπου οι παράμετροι των στηλών $\{v_{jk}, k$=1,…,$K\}$ διαφέρουν σε κάθε επίπεδο της Z, ενώ τα μονότονα scores των γραμμών $u_1 < \ldots < u_I$ είναι γνωστά.

2.6 Ακριβής συμπερασματολογία χωρίς δέσμευση (*exact unconditional*)

Η ακριβής προσέγγιση χωρίς δέσμευση (*unconditional*) αναπτύχθηκε κυρίως για 2x2 πίνακες συνάφειας. Πρόκειται για μια μέθοδο που χρησιμοποιεί τις γνήσιες (*unconditional*) κατανομές, απαλοίφοντας τις οχληρές παραμέτρους με χρήση του χειρότερου (*worst-case*) σεναρίου. Δηλαδή η p-value υπολογίζεται ως μια πιθανότητα ουράς, μεγιστοποιημένη για όλες τις πιθανές τιμές των οχληρών παραμέτρων. Επειδή λοιπόν δεν βασίζεται σε οχληρές

παραμέτρους, και επιπλέον εξασφαλίζει ότι το μέγεθος του ελέγχου είναι το πολύ ίσο με το ονομαστικό (*nominal*) μέγεθος α, η προσέγγιση αυτή μπορεί να χαρακτηριστεί ως ακριβής μέθοδος.

Για να γίνουν πιο σαφή τα παραπάνω ας δούμε την εφαρμογή της ακριβής προσέγγισης χωρίς δέσμευση στο πλαίσιο ελέγχου ανεξαρτησίας δύο πληθυσμών με δίτιμη μεταβλητή απόκρισης. Έστω τα δεδομένα μας, τα οποία μπορούν να γραφούν στην παρακάτω μορφή πίνακα :

	Επιτυχίες	Αποτυχίες	
Πληθυσμός 1	y_{11}	y_{12}	y_{1+}
Πληθυσμός 2	y_{21}	y_{22}	y_{2+}
			y_{++}

Το σύνηθες δειγματοληπτικό σχήμα σε αυτές τις περιπτώσεις είναι ότι κάθε γραμμή του πίνακα ακολουθεί διωνυμική κατανομή. Από την φύση λοιπόν του προβλήματος τα περιθώρια αθροίσματα $\{y_{i+}\}$, $i=1,2$, ως δειγματικά μεγέθη των δύο πληθυσμών, θα είναι σταθερά εξ αρχής.

Αν με π_1 και π_2 συμβολίσουμε τις πιθανότητες επιτυχίας των δύο πληθυσμών, μας ενδιαφέρει ο έλεγχος της υπόθεσης $H_0: \pi_1 = \pi_2 = \pi_0$, όπου η κοινή τιμή π_0 είναι η άγνωστη οχληρή παράμετρος.

Για τον έλεγχο αυτόν θα χρησιμοποιήσουμε κάποιο στατιστικό T (συνήθως επιλέγεται το X^2 του Pearson). Το στατιστικό T, για σταθερά αθροίσματα γραμμών, θα πάρει κάποιες συγκεκριμένες διακριτές τιμές, μια από τις οποίες είναι και η παρατηρούμενη t_{obs}. Υπό την μηδενική υπόθεση, η ακριβής p-value για τον αμφίπλευρο έλεγχο υπολογίζεται ως

$$\text{p-value} = P_{H_0}\left(\mathrm{T} \geq t_{obs}\right) = \sum_{S} f\left(\pi_0\right)$$

όπου
$$f\left(\pi_0\right) = \binom{y_{1+}}{y_{11}} \cdot \left[{\pi_0}^{y_{11}} \cdot \left(1-\pi_0\right)^{y_{12}}\right] \cdot \binom{y_{2+}}{y_{21}} \cdot \left[{\pi_0}^{y_{12}} \cdot \left(1-\pi_0\right)^{y_{22}}\right] \qquad (2.10)$$

η από κοινού κατανομή των δύο πληθυσμών, και S={το σύνολο των πινάκων που έχουν τα ίδια περιθώρια γραμμών με τον παρατηρούμενο πίνακα, και επιπλέον έχουν $\mathrm{T} \geq t_{obs}$}.

Επειδή όμως η οχληρή παραμετρος π_0 είναι άγνωστη, για να την απαλείψουμε θα της δώσουμε την τιμή $\pi_0 = \rho$ η οποία μεγιστοποιεί την p-value. Δηλαδή θα έχουμε ότι p-value = $\sup_{0 \le \pi_0 \le 1} P_{H_0}\left(\mathrm{T} \ge t_{obs}\right) = \sum_S f(\rho)$.

Ας το δούμε μέσω ενός απλού αριθμητικού παραδείγματος. Έστω ο 2x2 πίνακας με συχνότητες κελλίων (3,0 / 0,3) κατά γραμμή, και σταθερά αθροίσματα γραμμών (3,3) ως διωνυμικά δειγματικά μεγέθη. Θα χρησιμοποιήσουμε το στατιστικό X^2 του Pearson, το οποίο για τον παρατηρούμενο πίνακα είναι $X^2_{obs} = 6$. Παρατηρούμε ότι μόνο ένας ακόμα πίνακας έχει $X^2 \ge X^2_{obs}$ και είναι ο (0,3 / 3,0).

Η πιθανότητα (2.10) για τους δύο αυτούς πίνακες είναι $\pi_0^3 \cdot (1-\pi_0)^3$ και $(1-\pi_0)^3 \cdot \pi_0^3$ αντίστοιχα. Συνεπώς p-value = $P_{H_0}\left(X^2 \ge 6\right) = 2\,\pi_0^3 \cdot (1-\pi_0)^3$ η οποία όμως μεγιστοποιείται όταν $\pi_0 = 1/2$. Καταλήγουμε λοιπόν ότι η ακριβής p-value θα πάρει την τιμή p-value $= 2 \cdot \left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^3 =$ 0,031. Σε αντίθεση, η ακριβής p-value με δέσμευση για τον αμφίπλευρο έλεγχο του συγκεκριμένου παραδείγματος είναι p-value = 0,100.

2.7 Σύγκριση των ακριβών μεθόδων με δέσμευση και χωρίς δέσμευση

Ο *Barnard* [8],[9] ήταν ο πρώτος που εισήγαγε το uncoditional τεστ για σύγκριση διωνυμικών παραμέτρων, και αρκετοί μετέπειτα στατιστικοί το υποστήριξαν, όπως οι *Berkson* [11], *Kempthorne* [33], *Upton* [60], *Suissa & Shuster* [59], *D'Agostino, Chese & Belanger* [18].

Το γεγονός ότι οι δύο διαφορετικές ακριβής προσεγγίσεις (*conditional, uncoditional*) μπορεί να δώσουν αρκετά αντιφατικά αποτελέσματα (κυρίως στους 2x2 πίνακες) τροφοδότησε μια 50χρονη αντιπαράθεση για το ποιά από τις δύο προσεγγίσεις είναι καταλληλότερη (βλ. *Agresti* [3]).

Για πολλούς η conditional προσέγγιση είναι αρκετά τεχνητή, αφού δεσμεύουμε και ως προς περιθώρια αθροίσματα τα οποία δεν είναι σταθερά από την φύση του προβλήματος. Για τους υποστηρικτές της όμως είναι αφύσικο να βασίζουμε τον υπολογισμό της p-value σε πίνακες πολύ διαφορετικούς από τον παρατηρούμενο, ιδιαίτερα όταν καμία ουσιώδης

απώλεια πληροφορίας σχετικά με την μηδενική υπόθεση δεν προκύπτει με το να δεσμεύουμε ως προς τα περιθώρια αθροίσματα.

Οι υποστηρικτές της προσέγγισης χωρίς δέσμευση ισχυρίζονται ότι το χωρίς δέσμευση τέστ στο πλαίσιο των 2x2 πινάκων είναι λίγοτερο συντηρητικό και πιο ισχυρό από το αντίστοιχο ακριβές τέστ με δέσμευση του *Fisher* (βλ. *Shuissa & Shuster* [58]).

Πράγματι το γεγονός ότι στην προσέγγιση με δέσμευση δεσμεύουμε ως προς όλα τα περιθώρια αθροίσματα, έχει ως συνέπεια το σύνολο αναφοράς (σύνολο πινάκων με τα ίδια περιθώρια αθροίσματα) για την δειγματική κατανομή του εκάστοτε στατιστικού T, να περιλαμβάνει μικρότερο αριθμό πινάκων σε σύγκριση με το αντίστοιχο σύνολο αναφοράς της προσέγγισης χωρίς δέσμευση. Κατά συνέπεια η κατανομή χωρίς δέσμευση είναι λιγότερο διακριτή, αφού σε αυτήν τα μόνα δεσμευμένα περιθώρια αθροίσματα είναι αυτά που ήταν σταθερά από την φύση του προβλήματος. Σε αυτό το σημείο πρέπει να καταστήσουμε σαφές, πώς το πρόβλημα της συντηρητικότητας (εξαιτίας της διακριτότητας της δειγματικής κατανομής), είναι αναπόφευκτο όποια από τις δύο ακριβείς προσεγγίσεις επιλέξουμε. Απλά το πρόβλημα γίνεται πιο έντονο στην προσέγγιση με δέσμευση εξαιτίας της επιπλέον δέσμευσης.

Για παράδειγμα στο πείραμα του *Fisher* "Η κυρία με το τσάι" (§ 2.2.1), το y_{11} μπορεί να πάρει μόνο τις τιμές 4, 3, 2, 1, 0. Για τον μονόπλευρο έλεγχο $H_0 : \theta = 1$ κατά $H_1: \theta > 1$, βρίσκουμε p-value = 0,243. Διατηρώντας σταθερά τα περιθώρια αθροίσματα, οι μόνες πιθανές p-values για παρατηρούμενες τιμές $y_{11} = 4, 3, 2, 1, 0$ είναι 0,014, 0,243, 0,757, 0,986, 1,00 αντίστοιχα.

Αν για ονομαστικό μέγεθος του ελέγχου επιλέξουμε το $\alpha = 0{,}05$, απορρίπτουμε τη H_0 μόνο όταν $y_{11} = 4$. Υπό την μηδενική υπόθεση αυτό συμβαίνει με πιθανότητα 0,014. Έτσι η πραγματική πιθανότητα σφάλματος τύπου I (απορρίψη της H_0, όταν H_0 ορθή) του τέστ θα είναι P (*σφαλμα τύπου I*) = 0,014 και όχι 0,05 που επιθυμούμε. Συμπεραίνουμε λοιπόν ότι η προσέγγιση με δέσμευση είναι αρκετά συντηρητική.

Εν μέρει το πρόβλημα αυτό μπορεί να αντιμετωπιστεί κάνοντας χρήση συμπληρωματικής τυχαιοποίησης (*supplementary randomization*) προκειμένου να επιτύχουμε το επιθυμητό μέγεθος του ελέγχου.

Στο παράδειγμά μας ο τυχαιοποιημένος έλεγχος έχει ελεγχοσυνάρτηση

$$\varphi(y_{11}) = \begin{cases} 1, & y_{11} > 3 \ \text{(απόρριψη της } H_0\text{)} \\ \gamma, & y_{11} = 3 \ \text{(απόρριψη της } H_0 \text{ με πιθανότητα } \gamma\text{)} \\ 0, & y_{11} < 3 \ \text{(αποδοχή της } H_0\text{)} \end{cases}$$

Για α=0,05 το γ θα είναι :

$$\alpha = E_{H_0}\varphi(y_{11}) = 1 \cdot P(y_{11} > 3) + \gamma \cdot P(y_{11} = 3) + 0 \cdot P(y_{11} < 3) = P(y_{11} = 4) + \gamma \cdot P(y_{11} = 3)$$

$$\Rightarrow 0{,}05 = 0{,}014 + \gamma \cdot 0{,}228 \Rightarrow \gamma = 0{,}157$$

Συνεπώς η p-value θα είναι $P_{rand} = P_{H_0}(y_{11} > y_{11,obs}) + U \cdot P_{H_0}(y_{11} = y_{11,obs})$, όπου U είναι ένας τυχαίος αριθμός από την *Uniform*(0,1). Πρακτικά αν ο $U < 0{,}157$ τότε απορρίπτουμε την H_0 ενώ αν $U > 0{,}157$ την αποδεχόμαστε.

Στην θεωρία αυτή η προσέγγιση μας ικανοποιεί, στην πράξη όμως μια τέτοια αυθαίρετη τυχαιοποίηση είναι μη αποδεκτή. Προτείνεται λοιπόν η χρήση μιάς βελτιωμένης p-value της λεγόμενης mid-p-value (*Lancaster* [40]), η οποία προκύπτει από την p-value του τυχαιοποιημένου ελέγχου, αν αντικαταστήσουμε την ομοιόμορφη τ.μ.U με την αναμενόμενη τιμή της. Δηλαδή mid-p-value = $P_{H_0}(y_{11} > y_{11,obs}) + \frac{1}{2} \cdot P_{H_0}(y_{11} = y_{11,obs})$. Στο παράδειγμα "Η κυρία με το τσάι" βρίσκουμε mid-p-value = $\frac{1}{2} \cdot P(y_{11} = 3) + P(y_{11} > 3) = \frac{1}{2} \cdot 0{,}228 + 0{,}014$ = 0,129.

Το πλεονέκτημα της mid-p-value εκτός από την απλότητά της, είναι και το ότι συμπεριφέρεται σχεδόν σαν ομοιόμορφη (0,1) τυχαία μεταβλητή, μια ιδιότητα που χαρακτηρίζει τις p-values όταν το εκάστοτε στατιστικό T που χρησιμοποιούμε έχει συνεχή κατανομή (*Kim & Agresti* [35]). Αποτελεί λοιπόν έναν λογικό συμβιβασμό μεταξύ της συντηρητικότητας των ακριβών μεθόδων (*conditional,unconditional*) και της αβεβαιότητας που προκαλεί η χρήση των ασυμπτωτικών μεθόδων.

Όταν πηγαίνουμε σε μεγαλύτερα πλαίσια πινάκων (περισσότερα κελλιά ή μεγαλύτερο δείγμα), το πρόβλημα της διακριτότητας ελαττώνεται, με αποτελέσμα τα τέστ με δέσμευση να προκύπτουν αρκετές φορές ισχυρότερα από τα τέστ χωρίς δέσμευση (βλ. *Mehta & Hilton* [48]). Απλά αυτό που συμβαίνει στους 2x2 πίνακες συνάφειας είναι ότι η συντηρητικότητα που οφείλεται στην διακριτότητα της κατανομής με δέσμευση του στατιστικού T, υπερισχύει της συντηρητικότητας που προκαλείται από τον τρόπο εξάλειψης της οχληρής παραμέτρου στην κατανομή χωρίς δέσμευση του T. Όταν το πρόβλημα της διακριτότητας ξεπερασθεί,

τότε το πλεονέκτημα της μεγαλύτερης ισχύος των τέστ χωρίς δέσμευση δεν είναι πια εξασφαλισμένο.

Πρακτικά το μεγαλύτερο μειονέκτημα των τέστ χωρίς δέσμευση είναι η πολυπλοκότητα τους όταν τα εφαρμόζουμε σε πίνακες μεγάλυτερων διαστάσεων, στους οποίους θα πρέπει να αντιμετωπίσουμε αρκετές οχληρές παραμέτρους. Μάλιστα όσο μεγαλύτερος ο αριθμός των οχληρών παραμέτρων που πρέπει να αντιμετωπίσουμε, τόσο πιο δύσκολη γίνεται η εξάλειψή τους με χρήση του supremum. Αντίθετα τα conditional τέστ, εφαρμόζονται εύκολα παρέχοντας μας έναν τρόπο εξάλειψης των οχληρών παραμέτρων, απλά δεσμεύοντας ως προς τα περιθώρια αθροίσματα.

Οι *Berger & Boos* [10] αντιμετώπισαν εν μέρη την κριτική αυτή, περιορίζοντας την αναζήτηση του supremum της p-value σε ένα διάστημα εμπιστοσύνης τιμών για την οχληρή παράμετρο π_0, και όχι να αναζητήσουν το μέγιστο ως προς όλες τις πιθανές τιμές της οχληρής παραμέτρου. Η p-value για το δειγματοληπτικό πλαίσιο της παραγράφου §2.6, που πρότειναν οι *Berger & Boos* είναι :

$$\text{p-value} = \sup_{\pi_0 \in C_\gamma} [P_{H_0}(\mathrm{T} \geq t_{obs})] + \gamma \ ,$$

όπου C_γ συμβολίζει ένα 100 (1-γ)% δ.ε. για την παράμετρο π_0, και γ είναι πολύ μικρό (π.χ. 0,001) και αυθαίρετο.

ΚΕΦΑΛΑΙΟ 3

Ακριβείς μέθοδοι και υπολογιστικά προγράμματα

3.1 Ιστορική αναδρομή

Για πολλά χρόνια η ευρεία χρήση των ακριβών μεθόδων παρεμποδιζόταν από την έλλειψη κατάλληλων υπολογιστικών προγραμμάτων. Για να κάνουμε υπολογισμούς στην ακριβή προσέγγιση με δέσμευση, όπως είδαμε χρειάζεται να δουλέψουμε με το σύνολο αναφοράς των πινάκων που έχουν επαρκή στατιστικά ίδια με αυτά του παρατηρούμενου πίνακα. Το ενδεχομένως όμως μεγάλο μέγεθος του συνόλου αναφοράς αποτελούσε συχνά ανυπέρβλητο εμπόδιο. Για παράδειγμα, ένας 4x4 πίνακας με y_{++}= 20 μπορεί να έχει έως και 40.176 πίνακες με τα ίδια περιθώρια αθροίσματα, ενώ με y_{++}=100 μπορεί να φτάσει να έχει έως και $7{,}2 \cdot 10^9$ τέτοιους πίνακες.

Παράλληλα με την ανάπτυξη των υπολογιστών, αναπτύχθηκαν και χρησιμοποιήθηκαν πολλοί αλγόριθμοι για τον υπολογισμό των δειγματικών ακριβών δεσμευμένων κατανομών. Μια πολύ καλή ιστορική επισκόπηση των πρώτων αυτών αλγορίθμων έκαναν οι *Verbeek* & *Kroonenberg* [61]. Το μειονέκτημα όμως των περισσοτέρων από αυτούς τους πρώτους αλγορίθμους, είναι ότι πραγματοποιούν πλήρη απαρίθμηση του συνόλου αναφοράς κάτι που καθιστά την εφαρμογή τους χρονοβόρα και ακατάλληλη για μεγάλα προβλήματα.

Επειδή τις περισσότερες φορές μας ενδιαφέρει μόνο ο υπολογισμός του p-value και όχι ολόκληρη η κατανομή του εκάστοτε στατιστικού, ιδιαίτερη άνθηση γνώρισαν τα προγράμματα που χρησιμοποιούσαν τον δικτυωτό αλγόριθμο (*Network algorithm*). Ο αλγόριθμος αυτός βρήκε εφαρμογή σε πολλά προβλήματα, αφού είχε το πλεονέκτημα να είναι γρήγορος εξαιτίας τού ότι δεν απαιτεί πλήρη απαρίθμηση των πινάκων του συνόλου αναφοράς. Ενδιαφέροντα άρθρα με εφαρμογές του δικτυωτού αλγορίθμου γράφτηκαν από τους *Mehta, Patel* και συνεργάτες τους [4],[32],[49-52]. Για λόγους ιστορικού ενδιαφέροντος

στην §3.2 θα σκιαγραφήσουμε την κεντρική ιδέα του δικτυωτού αλγορίθμου στο πλαίσιο του ελέγχου ανεξαρτησίας σε *I*x*J* πίνακα συνάφειας.

Την δεκαετία του '90 η ραγδαία αύξηση της ισχύος των ηλεκτρονικών υπολογιστών, έδωσε τεράστια ώθηση και στην εφαρμογή των ακριβών μεθόδων. Όμως όσο και αν αυξηθεί η υπολογιστική ισχύς των H/Y, πάντα θα υπάρχουν μεγάλοι «προβληματικοί» αραιοί πίνακες συνάφειας, για τους οποίους είναι αδύνατον να υλοποιήσουμε τις ακριβείς μεθόδους. Το κυριότερο πρόβλημα είναι το μεγάλο μέγεθος του συνόλου αναφοράς, το οποίο μάλιστα αυξάνεται εκθετικά όσο αυξάνονται οι διαστάσεις του πίνακα.

Μια ικανοποιητική εναλλακτική αντιμετώπιση τέτοιων πινάκων είναι να εκτιμήσουμε το χαρακτηριστικό που μας ενδιαφέρει (π.χ. p-value) χρησιμοποιώντας Monte Carlo προσομοίωση της ακριβούς δεσμευμένης κατανομής (*exact conditional distribution*).

Ας υποθέσουμε ότι έχουμε έναν *I*x*J*x*K* πίνακα συνάφειας και θέλουμε να εκτιμήσουμε την ακριβή p-value για ένα στατιστικό T. Με T_{obs} συμβολίζουμε την παρατηρούμενη τιμή του στατιστικού, ενώ με *S* το σύνολο αναφοράς των πινάκων που έχουν επαρκή στατιστικά των οχληρών παραμέτρων ίδια με αυτά του παρατηρούμενου πίνακα. Προσομοιώνουμε παρατηρήσεις y_{ijk} , i=1,…I , j=1,…,J , k=1,…,K , από την ακριβή δεσμευμένη κατανομή, και για κάθε σετ προσομοιωμένων δεδομένων $\mathbf{y}^{(t)} = \left(y_{111}^{(t)}, \ldots, y_{IJK}^{(t)}\right)$, t=1,…,N , υπολογίζουμε την τιμή του στατιστικού T. Κάθε σετ $\mathbf{y}^{(t)}$ που παράγουμε αντιστοιχεί σε κάποιον πίνακα του συνόλου αναφοράς *S*. Αν είναι $T^{(t)} \geq T_{obs}$ τότε αντιστοιχίζουμε στον πίνακα αυτόν την τιμή z=1 διαφορετικά την τιμή z=0.

Η σημειακή λοιπόν εκτίμηση της ακριβούς p-value είναι $\hat{P} = \frac{1}{N}\sum_{t=1}^{N} z_t$, δηλαδή ο δειγματικός μέσος των Bernoulli τυχαίων μεταβλητών z_t , t=1,..,N. Η ακρίβεια της εκτίμησης καθορίζεται από την εκτιμώμενη δειγματική διακύμανση $\frac{\hat{P}\cdot\left(1-\hat{P}\right)}{N}$. Συνεπώς ανάλογα με το πλήθος των πινάκων που θα επιλέξουμε μπορούμε να επιτύχουμε τον επιθυμητό βαθμό ακρίβειας (*Agresti et.al.* [2]). Διαισθητικά, είναι σαν να πραγματοποιούμε δειγματοληψία με επανάθεση *N* πινάκων από το σύνολο αναφοράς *S*.

Οι *Mehta, Patel* & *Senchaudhuri* [50] περιέγραψαν μια βελτιωμένη και γρηγορότερη τροποποίηση της παραπάνω Monte Carlo προσέγγισης, κάνοντας χρήση της μεθόδου Importance Sampling (βλ. Παράρτημα A1 για μια σύντομη περιγραφή της).

Η εφαρμογή της Monte Carlo προσέγγισης όταν ελέγχουμε ανεξαρτησία και δεσμευμένη ανεξαρτησία γίνεται εύκολα, διότι η δειγματική δεσμευμένη κατανομή βασίζεται στην πολυδιάστατη υπεργεωμετρική κατανομή. Όταν όμως θέλουμε να ελέγξουμε πιο πολύπλοκες υποθέσεις (π.χ. ψευδοσυμμετρία (*quasi-symmetry*) ή ψευδοανεξαρτησία (*quasi-independence*)) η υλοποίηση είναι δύσκολη, διότι η δεσμευμένη κατανομή εξαρτάται από μια σταθερά κανονικοποίησης (*normalizing constant*) που δεν είναι εύκολο να καθοριστεί.

Μια ομάδα Βρεττανών στατιστικών (*Forster, Smith* & *McDonald*) ανέπτυξαν εναλλακτικές μεθόδους με χρήση Markov Chain Monte Carlo (*MCMC*) (βλ. Παράρτημα Α.2), με τις οποίες καλύπτουν την εκτίμηση των ακριβών p-values για υπόθεσεις ανεξαρτησίας, δεσμευμένης ανεξαρτησίας καθώς και πιο πολύπλοκων υποθέσεων [24-25],[46-47],[55-56]. Η MCMC προσέγγιση [24] που πρότειναν εφαρμόζεται γενικά σε loglinear και logit μοντέλα.

Οι *Booth* & *Butler* [15] παρουσίασαν μια γρηγορότερη και πιο αποτελεσματική υπολογιστική προσέγγιση για ακριβή τέστ καλής προσαρμογής σε λογαριθμογραμμικά μοντέλα, στην οποία αξιοποιείται η τεχνική Importance Sampling βασισμένη στην προσέγγιση της Poisson κατανομής από την κανονική κατανομή. Η μέθοδος των *Booth* & *Butler* βρίσκει εφαρμογή σε πληθώρα λογαριθμογραμμικών μοντέλων, αλλά έχει τον περιορισμό ότι δίνει καλά αποτελέσματα μόνο αν οι βαθμοί ελευθερίας του τέστ είναι το πολύ 19. Για περισσότερους βαθμούς ελευθερίας, θα πρέπει να προτιμηθεί η MCMC μέθοδος.

Στα κεφάλαια 4 και 5 θα αναλύσουμε διεξοδικότερα τις μεθόδους των *Booth* & *Butler*[15] καθώς και των *Forster, Smith* & *McDonald* [24], διότι αντιπροσωπεύουν ένα μέρος των τελευταίων και πιο σύγχρονων εξελίξεων στον τομέα της ακριβούς συμπερασματολογίας.

3.2 Ο Δικτυωτός αλγόριθμος (*Network algorithm*)

3.2.1 Τυποποίηση

Έστω $\{y_{ij}\}$ ο παρατηρούμενος *I*x*J* πίνακας, $R_i = \sum_{j=1}^{J} y_{ij}$ και $C_j = \sum_{i=1}^{I} y_{ij}$ τα περιθώρια αθροίσματα της *i*-γραμμής και της *j*-στήλης του πίνακα αντίστοιχα, και

$S_1 = \left\{ \{t_{ij}\} : \sum_{j=1}^{J} t_{ij} = R_i \, , \sum_{i=1}^{I} t_{ij} = C_j \right\}$ το σύνολο που περιέχει όλους τους πιθανούς πίνακες με περιθώρια αθροίσματα ίδια με τα παρατηρούμενα.

Η πιθανότητα να παρατηρήσουμε έναν τέτοιο πίνακα $\{t_{ij}\} \in S_1$ είναι :

$$P_{H_0}\left(\{t_{ij}\}\right) = f\left(\{t_{ij}\} \,\middle|\, R_i \, , C_j\right) = \left(\prod_{j=1}^{J} \frac{C_j!}{t_{1j}! \cdot t_{2j}! \cdot \ldots \cdot t_{Ij}!} \right) \Big/ \frac{T!}{R_1! \cdot R_2! \cdot \ldots \cdot R_I!} \tag{3.1}$$

όπου $T = \sum_{i=1}^{I} R_i$. Η ακριβής p-value για τον αμφίπλευρο έλεγχο της ανεξαρτησιάς είναι

$$\text{p-value} = \sum_{S_2} f\left(\{t_{ij}\} \,\middle|\, R_i \, , C_j\right) \tag{3.2}$$

όπου $S_2 = \{ \{t_{ij}\} : \{t_{ij}\} \in S_1$ και επιπλέον $f\left(\{t_{ij}\} \,\middle|\, R_i \, , C_j\right) \le f\left(\{y_{ij}\} \,\middle|\, R_i \, , C_j\right) \}$

Η σχέση (3.1) αλλά και η p-value είναι ισοδύναμες με τις αντίστοιχες σχέσεις της παραγράφου §2.3 (σελ.16) για τον ακριβή ελέγχο ανεξαρτησίας σε *I*x*J* πίνακα συνάφειας. Απλά εδώ ακολουθήσαμε διαφορετική τυποποίηση η οποία όμως θα μας διευκολύνει στην περιγραφή του ακριβούς ελέγχου ανεξαρτησίας ως δικτυωτό πρόβλημα.

3.2.2 Κατασκευή του δικτύου και περιγραφή του ελέγχου ανεξαρτησίας ως δικτυωτό πρόβλημα

Η δικτυωτή απεικόνιση του συνόλου αναφοράς S_1 αποτελεί την "καρδιά" του δικτυωτού αλγορίθμου. Θα περιγράψουμε πώς κατασκευάζεται αυτή η δικτυωτή απεικόνιση και θα δώσουμε και ένα παράδειγμα.

Το δίκτυο αποτελείται από κόμβους (*nodes*) και τόξα (*arcs*) σε *J*+1 επίπεδα, τα οποία αριθμούνται ως *J*, *J*-1, …, 0. Σε κάθε επίπεδο *k* υπάρχει ένα σύνολο κόμβων, κάθε έναν από τους οποίους θα τον συμβολίζουμε ως $(k, R_{1k}, \ldots, R_{Ik})$.

Τα τόξα πηγάζουν από κάθε κόμβο στο επίπεδο *k*, και το καθένα καταλήγει σε έναν μόνο κόμβο στο επίπεδο *k*-1. Στο αρχικό επίπεδο *J* υπάρχει μόνο ένας κόμβος, ο $(J, R_{1J}, \ldots, R_{IJ})$ όπου $R_{iJ} \equiv R_i$, $i = 1, \ldots, I$.

Το εύρος τιμών που μπορεί να πάρουν τα $R_{i,k-1}$, $i=1,2,\ldots,I$ δίνονται από την σχέση :

$$\max\left(0, R_{ik} - C_k + \sum_{l=1}^{i-1}\left(R_{lk} - R_{l,k-1}\right)\right) \le R_{i,k-1} \le \min\left(R_{ik}, S_{k-1} - \sum_{l=1}^{i-1} R_{l,k-1}\right) \qquad (3.3)$$

όπου $S_j = \sum_{l=1}^{j} C_l$. Επιπλέον αν το κάτω όριο ενός αθροίσματος είναι μεγαλύτερο από το πάνω όριο, τότε το άθροισμα αυτό είναι μηδέν.

Εφαρμόζοντας σταδιακά την (3.3) σε όλα τα επίπεδα *J*, *J*-1, …, 0 ολοκληρώνουμε το δίκτυο. Στο τελευταίο επίπεδο 0, υπάρχει μόνο ένας κόμβος τον οποίο συμβολίζουμε (0,0,…,0) και είναι ο τερματικός κόμβος.

Το μήκος κάθε τόξου που ξεκινά από τον κόμβο $(k, R_{1k}, \ldots, R_{Ik})$ και καταλήγει στον κόμβο $(k-1, R_{1k-1}, \ldots, R_{Ik-1})$ είναι ίσο με

$$\frac{C_k!}{\left(R_{1k} - R_{1,k-1}\right)! \cdot \ldots \cdot \left(R_{Ik} - R_{I,k-1}\right)!}$$

Μια πλήρης διαδρομή-μονοπάτι (*path*) στο δίκτυο, από τον αρχικό κόμβο έως τον τερματικό κόμβο, έχει μήκος το γινόμενο των μηκών των τόξων που αποτελούν το μονοπάτι αυτό. Δηλαδή είναι

$$\prod_{j=1}^{J} \frac{C_j!}{\left(R_{1j} - R_{1,j-1}\right)! \cdot \ldots \cdot \left(R_{Ij} - R_{I,j-1}\right)!} \qquad (3.4)$$

Από τον τρόπο κατασκευής του δικτύου είναι προφανές ότι κάθε μονοπάτι της μορφής $\left(J, R_{1J}, \ldots, R_{IJ}\right) \to \left(J-1, R_{1J-1}, \ldots, R_{IJ-1}\right) \to \ldots \to \left(0,0,\ldots,0\right)$ αντιστοιχεί σε έναν *I*x*J* πίνακα $\{t_{ij}\} \in S_1$ όπου $t_{ij} = R_{ij} - R_{i,j-1}$, $i=1,\ldots I$, $j=1,\ldots,J$.

Αν ρίξουμε μια πιο προσεκτική ματιά στις σχέσεις (3.1) και (3.4) παρατηρούμε ότι το μήκος του κάθε μονοπατιού ισοδυναμεί με την ποσότητα

$$P_{H_0}\left(\{t_{ij}\}\right) \cdot D \quad \text{όπου} \quad D = \frac{T!}{R_1! \cdot R_2! \cdot \ldots \cdot R_I!}$$

Άρα ο υπολογισμός του p-value από την (3.2) μετατρέπεται σε πρόβλημα εύρεσης και άθροισης των μονοπατιών του δικτύου, που έχουν μήκος μικρότερο ή ίσο του $P_{H_0}\left(\{y_{ij}\}\right) \cdot D$. Υπενθυμίζουμε ότι $\{y_{ij}\}$ είναι ο παρατηρούμενος πίνακας.

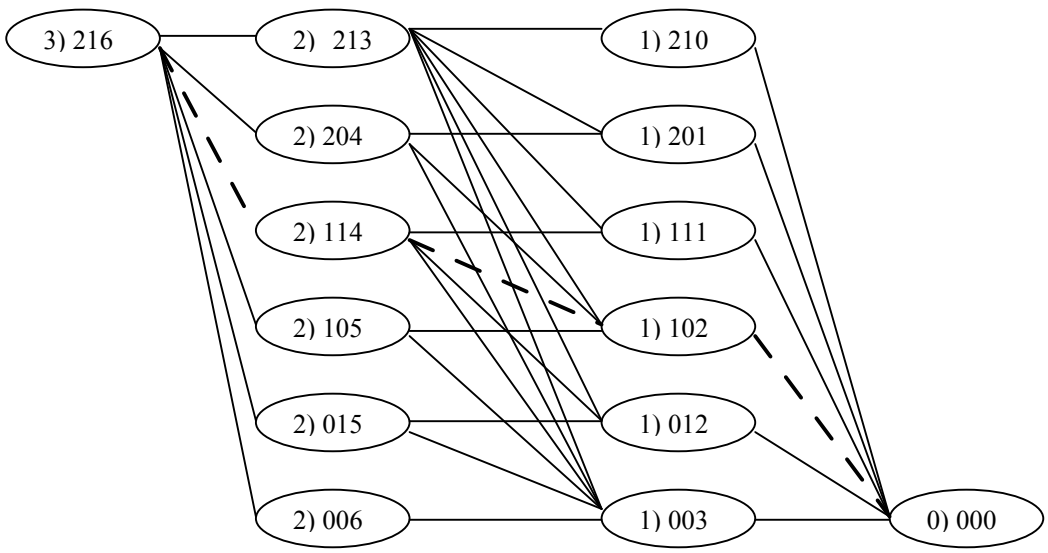
Το πλεονέκτημα με την δικτυωτή απεικόνιση είναι ότι δεν χρειαζόμαστε πια την εξαντλητική και χρονοβόρα απαρίθμηση κάθε μονοπατιού (δηλ. κάθε πίνακα στο S_1), αλλά

μπορούμε με την βοήθεια κάποιων ειδικών συνθηκών να ανιχνεύσουμε από την αρχή τα μήκη των μονοπατιών που συνεισφέρουν στην p-value, και να υπολογίσουμε μόνο αυτά.

Δεν θα αναφέρουμε περισσότερες τεχνικές λεπτομέρειες για τις ειδικές συνθήκες και την περαιτέρω δομή του αλγορίθμου, διότι κάτι τέτοιο ξεφεύγει από τους στόχους αυτής της εργασίας. Περισσότερες πληροφορίες μπορεί κάποιος να αντλήσει από τα άρθρα [4],[49],[52].

3.2.3 Παράδειγμα

Έστω ένας 3x3 πίνακας συνάφειας με περιθώρια αθροίσματα γραμμών R_1=2, R_2=1, R_3=6 και περιθώρια αθροίσματα στηλών C_1=3, C_2=3, C_3=3. Το σύνολο S_1 των πιθανών πινάκων με περιθώρια αθροίσματα ίδια με τα παρατηρούμενα μπορεί να απεικονιστεί σε δικτυωτή μορφή ως :



Κάθε μονοπάτι του δικτύου αντιστοιχεί σε έναν πίνακα του συνόλου S_1. Το μονοπάτι με την διακεκομμένη γραμμή αντιστοιχεί στον πίνακα

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 2 & 2 & 2 \end{bmatrix}$$

και έχει μήκος $\frac{3!}{0!\cdot 0!\cdot 3!}\cdot\frac{3!}{0!\cdot 0!\cdot 3!}\cdot\frac{3!}{2!\cdot 1!\cdot 0!}$.

ΚΕΦΑΛΑΙΟ 4

Η μέθοδος Importance Sampling για ακριβή δεσμευμένα τέστ σε πίνακες συνάφειας

4.1 Εισαγωγή

Όπως έχουμε ήδη αναφέρει, όταν μας ενδιαφέρει η πραγματοποίηση ενός ελέγχου υποθέσεων για τις παραμέτρους ενός μοντέλου και υπάρχουν οχληρές παράμετροι, τότε χρησιμοποιούμε την δεσμευμένη κατανομή των επαρκών στατιστικών για τις παραμέτρους που μας ενδιαφέρουν, δεσμεύοντας ως προς τα επαρκή στατιστικά των οχληρών παραμέτρων. Αυτό δεν είναι τίποτε άλλο από ένα δεσμευμένο ακριβές τέστ.

Σε προσεγγίσεις όπως της παραπάνω μορφής, το πρόβλημα είναι ο καθορισμός της δεσμευμένης κατανομής και ο υπολογισμός της p-value. Όταν η εφαρμογή της ασυμπτωτικής θεωρίας είναι αναξιόπιστη (μικρό δείγμα ή *sparse data*) ή δεν μπορούμε να καθορίσουμε πλήρως την μορφή της δεσμευμένης κατανομής, τότε εναλλακτικά μπορούμε να προσεγγίσουμε το πρόβλημα με χρήση της μεθόδου Markov Chain Monte Carlo (*MCMC*).

Οι *Booth & Butler* [15] πρότειναν μια μέθοδο προσομοίωσης χρησιμοποιώντας την τεχνική του Importance Sampling (βλ. Παράρτημα.Α.1). Η μέθοδος τους βασίζεται στην κανονική προσέγγιση της Poisson κατανομής και βρίσκει εφαρμογή όταν μας ενδιαφέρει να ελέγξουμε την καλή προσαρμογή μοντέλων όπως μοντέλα ανεξαρτησίας, ομοιόμορφης συνάφειας, ψευδο-ανεξαρτησίας, και ψευδο-συμμετρίας σε διδιάστατους πίνακες συνάφειας. Σε πολυδιάστατους πίνακες η μέθοδος μπορεί να χρησιμοποιηθεί για τον έλεγχο της σημαντικότητας των όρων αλληλεπίδρασης μεγαλύτερης τάξης.

Είναι σημαντικό να αναφέρουμε ότι η μέθοδος αυτή εκτός του ότι είναι απλούστερη, είναι και 3 με 4 φορές ταχύτερη από την κλασσική MCMC διαδικασία (χρησιμοποιώντας π.χ. Gibbs Sampling). Όμως η αποτελεσματικότητά της περιορίζεται σε μικρούς πίνακες συνάφειας. Για την ακρίβεια, η μέθοδος των *Booth & Butler* «δουλεύει» ικανοποιητικά όταν οι βαθμοί ελευθερίας, για το τέστ καλής προσαρμογής του υπό έλεγχο μοντέλου, κυμαίνονται

από 1 εως 19. Συνεπώς, για πολύ μεγάλους πίνακες συνάφειας η επιλογή μας θα είναι η MCMC διαδικασία που αντιμετωπίζει αποτελεσματικότερα πολυδιάστατα προβλήματα.

Το υπόλοιπο κεφάλαιο είναι οργανωμένο ως εξής. Στην §4.2 θα περιγράψουμε την προσαρμογή της ιδέας του Importance Sampling στο πλαίσιο των πινάκων συνάφειας, από τους *Booth & Butler*. Στις §§4.3 και 4.4 παραθέτουμε τα βήματα του σχετικού αλγορίθμου, καθώς και κάποιες τεχνικές λεπτομέρειες σχετικές με τον πίνακα σχεδιασμού (*design matrix*).

Ο αλγόριθμος υλοποιήθηκε χρησιμοποιώντας Mathematica, και παρατείθεται αναλυτικά στο Παράρτημα Δ.1. Μια εφαρμογή του για τον έλεγχο ανεξαρτησίας σε έναν 4x4 πίνακα συνάφειας παρουσιάζεται στην §4.5

4.2 Το μοντέλο και η μέθοδος Importance Sampling

Ας δούμε αναλυτικότερα την μέθοδο Importance Sampling που πρότειναν οι *Booth & Butler*. Για ευκολία στην περιγραφή θα χρησιμοποιήσουμε τον διανυσματικό συμβολισμό των λογαριθμογραμμικών μοντέλων.

Έστω τα δεδομένα, $\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \dots & y_n \end{pmatrix}'$, προερχόμενα από ένα κορεσμένο Poisson λογαριθμογραμμικό μοντέλο με μέσες τιμές $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 & \mu_2 & \dots & \mu_n \end{pmatrix}'$. Το μοντέλο μας θα είναι :

$$\log(\mu_i) = \mathbf{x}_i' \cdot \boldsymbol{\beta}$$

όπου $\mathbf{x}_i' = \begin{pmatrix} x_{i1} & x_{i2} & \dots & x_{ip} \end{pmatrix}$ είναι το p-διάνυσμα των γνωστών συμμεταβλητών, ενώ $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 & \beta_2 & \dots & \beta_p \end{pmatrix}'$ είναι το p-διάνυσμα των αγνώστων παραμέτρων.

Αν με $\hat{\boldsymbol{\beta}}$ συμβολίσουμε τον εκτιμητή μέγιστης πιθανοφάνειας (MLE) του $\boldsymbol{\beta}$ και αντίστοιχα με $\hat{\boldsymbol{\mu}}$ τον MLE του $\boldsymbol{\mu}$, τότε $\hat{\mu}_i = \exp\left(\mathbf{x}_i' \cdot \hat{\boldsymbol{\beta}} \right)$.

Τα πιο γνωστά στατιστικά που χρησιμοποιούνται για τον έλεγχο καλής προσαρμογής (*goodness of fit*) ενός μοντέλου είναι η απόκλιση (*Deviance*)

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\{ l(\mathbf{y}; \mathbf{y}) - l(\hat{\boldsymbol{\mu}}; \mathbf{y}) \} \text{ όπου } l(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} \left(y_i \cdot \log \mu_i - \mu_i \right)$$

και το X^2 του Pearson

$$X^2(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

Εμείς για την περιγραφή της μεθόδου, θα χρησιμοποιήσουμε την απόκλιση *D*. Αν συμβολίσουμε με $\mathbf{s} = \mathbf{X}' \cdot \mathbf{y}$ το διάνυσμα των επαρκών στατιστικών για το $\boldsymbol{\beta}$, τότε η δεσμευμένη ακριβής p-value ορίζεται ως :

$$\text{p-value} = P\left(D(\mathbf{y};\hat{\boldsymbol{\mu}}) \geq D_{obs} \,|\mathbf{s}\right) = \sum_{\mathbf{X}'\cdot\mathbf{y}=\mathbf{s}} I\left(D(\mathbf{y};\hat{\boldsymbol{\mu}}) \geq D_{obs}\right) \cdot f\left(\mathbf{y}|\mathbf{s}\right) \tag{4.1}$$

όπου :

- D_{obs} είναι η παρατηρούμενη τιμή για την απόκλιση
- $I\left(D(\mathbf{y};\hat{\boldsymbol{\mu}}) \geq D_{obs}\right) = \begin{cases} 1, & \text{όταν } D(\mathbf{y};\hat{\boldsymbol{\mu}}) \geq D_{obs} \\ 0, & \text{διαφορετικά} \end{cases}$, η δείκτρια συνάρτηση
- **X** είναι ο *n* x *p* πίνακας συμμεταβλητών
- $f\left(\mathbf{y}|\mathbf{s}\right)$ είναι η από κοινού κατανομή των δεδομένων μας δεσμεύοντας ως προς τα επαρκή στατιστικά.

Η από κοινού συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) $f\left(\mathbf{y};\boldsymbol{\mu}\right)$ των δεδομένων μας , παραγοντοποιείται ως εξής : $f\left(\mathbf{y};\boldsymbol{\mu}\right) = f\left(\mathbf{y}|\mathbf{s}\right) \cdot f\left(\mathbf{s};\boldsymbol{\mu}\right)$. Συνεπώς η σχέση (4.1) γράφεται :

$$\text{p} - \text{value} = \frac{\sum_{\mathbf{X}'\cdot\mathbf{y}=\mathbf{s}} I\left(D(\mathbf{y};\hat{\boldsymbol{\mu}}) \geq D_{obs}\right) \cdot f\left(\mathbf{y};\boldsymbol{\mu}\right)}{f\left(\mathbf{s};\boldsymbol{\mu}\right)} = \frac{\sum_{\mathbf{X}'\cdot\mathbf{y}=\mathbf{s}} I\left(D(\mathbf{y};\hat{\boldsymbol{\mu}}) \geq D_{obs}\right) \cdot f\left(\mathbf{y};\boldsymbol{\mu}\right)}{\sum_{\mathbf{X}'\cdot\mathbf{y}=\mathbf{s}} f\left(\mathbf{y};\boldsymbol{\mu}\right)}$$

Η τελευταία σχέση δεν εξαρτάται από την συγκεκριμένη τιμή του $\boldsymbol{\mu}$ που ικανοποιεί την $\log\left(\mu_i\right) = \mathbf{x}_i' \cdot \boldsymbol{\beta}$, και επομένως μπορούμε να επιλέξουμε $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$. Όμως ο υπολογισμός της p-value με την παραπάνω σχέση απαιτεί τον ξεχωριστό υπολογισμό των $D(\mathbf{y};\hat{\boldsymbol{\mu}})$ και $f\left(\mathbf{y};\hat{\boldsymbol{\mu}}\right)$ για κάθε έναν διαφορετικό συνδυασμό του $\mathbf{y} = \left(y_1 \quad y_2 \quad \ldots \quad y_n\right)'$ που ικανοποιεί την $\mathbf{s} = \mathbf{X}' \cdot \mathbf{y}$. Κάτι τέτοιο προφανώς είναι χρονοβόρο και πολλές φορές είναι και πρακτικά αδύνατο.

Εναλλακτικά, αν μπορούσαμε να προσομοιώσουμε τιμές απ' ευθείας από την $f\left(\mathbf{y}|\mathbf{s}\right)$, τότε πολύ εύκολα θα υπολογίζαμε την Monte Carlo προσέγγιση της p-value

$$\tilde{\text{p}} - \text{value} = \frac{1}{N}\sum_{k=1}^{N} I\left(D_k^* \geq D_{obs}\right) \tag{4.2}$$

όπου $D_k^* = D\left(\mathbf{y}_k^*; \boldsymbol{\mu}\right)$ είναι η τιμή της απόκλισης που υπολογίστηκε χρησιμοποιώντας το σετ *k*, των δεδομένων $\mathbf{y}_k^* = \left(y_1^*, y_2^*, \ldots, y_n^*\right)$ που παργάγαμε από την $f\left(\mathbf{y}|\mathbf{s}\right)$. Τις περισσότερες φορές όμως αυτό είναι ανέφικτο, άρα και πάλι θα πρέπει να εργαστούμε διαφορετικά.

Ως εναλλακτική μέθοδο επίλυσης του προβλήματος, οι *Booth & Butler* πρότειναν την Monte Carlo προσέγγιση της p-value με χρήση του Importance Sampling. Προσομοιώνοντας λοιπόν σετ δεδομένων από μια βοηθητική κατανομή $g\left(\mathbf{y}|\mathbf{s}\right)$ που έχει το ίδιο στήριγμα (*support*) με την $f\left(\mathbf{y}|\mathbf{s}\right)$ υπολογίζουμε ότι

$$\tilde{\text{p}} - \text{value} = \frac{\sum_{k=1}^{N} I\left(D\left(\mathbf{y}_k^*; \hat{\boldsymbol{\mu}}\right) \geq D_{obs}\right) \cdot \dfrac{f\left(\mathbf{y}_k^*; \hat{\boldsymbol{\mu}}\right)}{g\left(\mathbf{y}_k^* | \mathbf{s}\right)}}{\sum_{k=1}^{N} \dfrac{f\left(\mathbf{y}_k^*; \hat{\boldsymbol{\mu}}\right)}{g\left(\mathbf{y}_k^* | \mathbf{s}\right)}} = \frac{\sum_{k=1}^{N} I\left(D\left(\mathbf{y}_k^*; \hat{\boldsymbol{\mu}}\right) \geq D_{obs}\right) \cdot w_k^*}{\sum_{k=1}^{N} w_k^*} \qquad (4.3)$$

όπου $w_k^* = \dfrac{f\left(\mathbf{y}_k^*; \hat{\boldsymbol{\mu}}\right)}{g\left(\mathbf{y}_k^* | \mathbf{s}\right)}$ το βάρος του *k*-σετ δεδομένων $\mathbf{y}_k^* = \left(y_1^*, y_2^*, \ldots, y_n^*\right)$ που παραγάγαμε από την $g(\cdot)$.

Στην συνέχεια για να απαντηθεί το ερώτημα ποιά θα είναι η βοηθητική κατανομή $g(\cdot)$, οι *Booth & Butler* στηρίχτηκαν στην κανονική προσέγγιση της Poisson. Δηλαδή $y_i \sim N\left(\mu_i, \mu_i\right)$ και επειδή $y_1, y_2, \ldots y_n$ είναι ανεξάρτητες προκύπτει ότι η από κοινού κατανομή τους θα είναι μια πολυμεταβλητή κανονική, με μέσο **μ** και πίνακα συνδιακυμάνσεων **V**=diag(**μ**).

Για την ακρίβεια στηρίχτηκαν στην παρατήρηση ότι αν διαμερίσουμε τον *n*x*p* πίνακα σχεδιασμού **X** σε δύο υποπίνακες, τον $\mathbf{X}_1$ χρησιμοποιώντας τις πρώτες *n-p* γραμμές και τον $\mathbf{X}_2$ χρησιμοποιώντας τις *p* τελευταίες γραμμες του **X**, έτσι ώστε $\mathbf{X}' = \left(\mathbf{X}_1' \quad \mathbf{X}_2'\right)$, και όμοια διαμερίσουμε και το **y** έτσι ώστε $\mathbf{Y} = \left(\mathbf{Y}_1' \quad \mathbf{Y}_2'\right)'$, τότε προφανώς ισχύει $\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{s} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{X}_1' & \mathbf{X}_2' \end{pmatrix} \cdot \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}$ όπου **I** είναι ο (*n-p*)x(*n-p*) μοναδιαίος πίνακας και **0** ο (*n-p*)x*p* μηδενικός πίνακας.

Προκύπει λοιπόν ότι η κατανομή του $\mathbf{Y}_1$ δοθέντος του **s** είναι μια πολυμεταβλητή κανονική κατανομή με μέση τιμή

$$E\left(\mathbf{Y}_1|\mathbf{s}\right) = E\left(\mathbf{Y}_1\right) - \mathbf{V}_{11}\mathbf{X}_1\left(\mathbf{X}'\mathbf{V}\mathbf{X}\right)^{-1}\left\{\mathbf{s} - E\left(\mathbf{s}\right)\right\} \qquad (4.4)$$

και πίνακα συνδιακυμάνσεων

$$\mathbf{V}_{11\bullet\mathbf{s}}(\boldsymbol{\mu}) = \mathbf{V}_{11} - \mathbf{V}_{11}\mathbf{X}_1\left(\mathbf{X}'\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}_1' \cdot \mathbf{V}_{11} \tag{4.5}$$

όπου $\mathbf{V}_{11} = \text{var}(\mathbf{Y}_1) = \begin{pmatrix} \mu_1 & 0 & \cdots & 0 \\ 0 & \mu_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \mu_{n-p} \end{pmatrix}$ και $\mathbf{V}_{nxn} = diag(\boldsymbol{\mu})$

Επειδή όμως η πολυμεταβλητή κανονική κατανομή του $\mathbf{Y}_1$ δοθέντος του **s** δεν εξαρτάται από το **β**, μπορούμε να επιλέξουμε $\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}$, κάτι που συνεπάγεται $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$ και ως εκ τούτου $\text{E}(\mathbf{s}) = \mathbf{X}'\boldsymbol{\mu} = \mathbf{X}'\hat{\boldsymbol{\mu}} = \mathbf{s}$. Η σχέση (4.4) θα πάρει τότε την απλούστερη μορφή $E(\mathbf{Y}_1|\mathbf{s}) = (\mathbf{I} \quad \mathbf{0}) \cdot \hat{\boldsymbol{\mu}}$, ενώ η δεσμευμένη διακύμανση (4.5) θα υπολογίζεται ως $\hat{\mathbf{V}}_{11\bullet\mathbf{s}} = \mathbf{V}_{11\bullet\mathbf{s}}(\hat{\boldsymbol{\mu}})$.

Βασιζόμενοι λοιπόν στην παραπάνω παρατήρηση, οι *Booth & Butler* όρισαν $\mathbf{Z} = (z_1, z_2, \ldots, z_{n-p})'$ ως μια πολυμεταβλητή κανονική τ.μ. με μέσο το διάνυσμα $(\mathbf{I} \quad \mathbf{0}) \cdot \hat{\boldsymbol{\mu}}$ και πίνακα συνδιακυμάνσεων $\hat{\mathbf{V}}_{11\bullet\mathbf{s}} = (\hat{v}_{ij})$ με διαστάσεις (*n*-*p*)x(*n*-*p*) . Προφανώς όμως η κατανομή της **Z** δεν έχει το ίδιο στήριγμα (*support*) με την $f(\mathbf{y}|\mathbf{s})$ και άρα δεν μπορεί να χρησιμοποιηθεί ως $g(\mathbf{y}|\mathbf{s})$. Η λύση είναι να παράγουμε τιμές από την πολυμεταβλητή κανονική κατανομή και στην συνέχεια να τις στρογγυλοποιήσουμε στον πλησιέστερο ακέραιο.

4.3 Ο αλγόριθμος

Ας δούμε αναλυτικά πώς εφαρμόζονται όλα τα προαναφερθέντα κατά την υλοποίηση του αλγορίθμου:

<u>Βήμα 1</u> : *Προσομοίωση των* $y_1^*, y_2^*, \ldots, y_{n-p}^*$

Παράγουμε $z_1 = z_1^*$ από την περιθώρια κανονική κατανομή με μέσο $m_1 = \hat{\mu}_1$, δηλαδή το 1ο στοιχείο του διανύσματος $(\mathbf{I} \quad \mathbf{0}) \cdot \hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_{n-p} \end{pmatrix}$, και διακύμανση $s_{11} = \hat{v}_{11}$, δηλαδή το (1,1) στοιχείο του $\hat{\mathbf{V}}_{11\bullet\mathbf{s}}$.

Ορίζουμε ως y_1^* την τιμή του z_1^* στρογγυλοποιημένη (*rounding*) στον πλησιέστερο ακέραιο, με αντίστοιχη πιθανότητα :

$$P_Y\left(Y_1^* = z_1^*\right) = P_Z\left(z_1^* - \frac{1}{2} \le Z_1^* \le z_1^* + \frac{1}{2}\right) = P_Z\left(Z_1^* \le \frac{z_1^* + \frac{1}{2} - m_1}{\sqrt{s_{11}}}\right) - P_Z\left(Z_1^* \le \frac{z_1^* - \frac{1}{2} - m_1}{\sqrt{s_{11}}}\right) =$$

$$= \Phi\left(\frac{z_1^* + \frac{1}{2} - m_1}{\sqrt{s_{11}}}\right) - \Phi\left(\frac{z_1^* - \frac{1}{2} - m_1}{\sqrt{s_{11}}}\right) \qquad (4.6)$$

Στην συνέχεια, όμοια ακολουθεί η προσομοίωση των $y_2^*, \ldots, y_{n-p}^*$. Παράγουμε λοιπόν τα $z_k = z_k^*$, $k = 2, 3, \ldots, n-p$, όπου ο μέσος και η διακύμανση των περιθώριων κανονικών κατανομών είναι αντίστοιχα :

$$m_k = \hat{\mu}_k + \sum_{i=1}^{k-1} \frac{\mathrm{x}_i \cdot \hat{v}_{ik}^*}{s_{ii}} \quad \text{και} \quad s_{kk} = \hat{v}_{kk} - \sum_{i=1}^{k-1} \frac{\left(v_{ik}^*\right)^2}{s_{ii}}$$

όπου $\hat{v}_{i,k+1}^* = \begin{cases} \hat{v}_{i,k+1} & , \ i = 1 \\ \hat{v}_{i,k+1} - \sum_{j=1}^{i-1} \frac{v_{ji}^* \cdot v_{j,k+1}}{s_{jj}} & , \ i \ge 2 \end{cases}$ και $\mathrm{x}_i = y_i^* - \hat{\mu}_i$.

Ορίζουμε ως y_k^* την τιμή του z_k^* στρογγυλοποιημένη στον πλησιέστερο ακέραιο, με αντίστοιχη πιθανότητα που ορίζεται ανάλογα της (4.6).

<u>Βήμα 2</u> : *Υπολογισμός των* $y_{n-p+1}^*, y_{n-p+2}^*, \ldots, y_n^*$

Στο προηγούμενο βήμα έχουμε παραγάγει τις τιμές $y_1^*, y_2^*, \ldots, y_{n-p}^*$ και σε αυτό το βήμα θα παράγουμε τις υπόλοιπες τιμές $y_{n-p+1}^*, y_{n-p+2}^*, \ldots, y_n^*$. Επειδή όμως δεσμεύουμε ως προς τα επαρκή στατιστικά **s**, πρέπει οι *n* στο πλήθος τιμές που συνολικά θα παραχθούν, να ικανοποιούν την σχέση $\mathbf{s} = \mathbf{X}' \cdot \mathbf{y}^*$

Συνεπώς τις υπόλοιπες *p* στο πλήθος τιμές $y_{n-p+1}^*, y_{n-p+2}^*, \ldots, y_n^*$ θα τις ορίσουμε από την επίλυση του παρακάτω συστήματος

$$\sum_{i=n-p+1}^{n} x_{ij} \cdot y_i^* = s_j - \sum_{i=1}^{n-p} x_{ij} \cdot y_i^* \quad , \ j=1,2,\ldots,p \qquad (4.7)$$

Το (4.7) έχει μοναδική λύση. Την ύπαρξη της μοναδικής λύσης θα μας την εξασφαλίσει το γεγονός ότι όταν διαμερίσουμε τον πίνακα σχεδιασμού $\mathbf{X}$ θα το κάνουμε με τέτοιον τρόπο ώστε ο υποπίνακας $\mathbf{X}_2$ να είναι πλήρους τάξης (*full rank*). Αυτό γίνεται περισσότερο κατανοητό αν δούμε το (4.7) με μορφή πινάκων. Δηλαδή θα είναι $\mathbf{X}_2' \cdot \mathbf{Y}_2^* = \mathbf{s} - \mathbf{X}_1' \cdot \mathbf{Y}_1^*$ όπου $\mathbf{Y}_2^* = \left(y_{n-p+1}^* \;\; y_{n-p+2}^* \;\; \ldots \;\; y_n^*\right)'$ είναι η ζητούμενη λύση του (4.7), ενώ $\mathbf{Y}_1^* = \left(y_1^* \;\; y_2^* \;\; \ldots \;\; y_{n-p}^*\right)'$ είναι το διάνυσμα των τιμών που έχουμε ήδη παραγάγει από τα προηγούμενα βηματα.

Για να έχει λοιπόν το σύστημα λύση, θα πρέπει ο $\mathbf{X}_2'$ να αντιστρέφεται, δηλαδή να έχει μή μηδενική ορίζουσα. Αυτό μας το εξασφαλίζει η ιδιότητα της πλήρους τάξης (*full rank*) που θα έχει ο $\mathbf{X}_2$ εκ κατασκευής. Στο θέμα αυτό θα επανέλθουμε στην §4.4 όπου θα αναλύσουμε λεπτομερέστερα τον τρόπο υπολογισμού του πίνακα σχεδιασμού.

Βήμα 3 : *Υπολογισμός του βάρους (weight)*

Στην συνέχεια υπολογίζουμε το βάρος w^* για το σέτ των δεδομένων $y_1^*, y_2^*, \ldots, y_n^*$ που παραγάγαμε.

Θα είναι $w^* = 0$, όταν το σέτ των δεδομένων μας περιλαμβάνει έστω και μία αρνητική τιμή, διαφορετικά $w^* = \dfrac{f\left(\mathbf{y}^*; \hat{\boldsymbol{\mu}}\right)}{\prod_{i=1}^{n-p}\left\{\Phi\left(\dfrac{y_i^* + \frac{1}{2} - m_i}{\sqrt{s_{ii}}}\right) - \Phi\left(\dfrac{y_i^* - \frac{1}{2} - m_i}{\sqrt{s_{ii}}}\right)\right\}}$

Βήμα 4 : *Υπολογισμός της απόκλισης*

Υπολογίζουμε την απόκλιση $D^* = D\left(\mathbf{y}^*, \hat{\boldsymbol{\mu}}\right)$. Σημαντικό είναι να παρατηρήσουμε ότι οι προσαρμοσμένες τιμές $\hat{\boldsymbol{\mu}}$ είναι οι ίδιες για όλα τα προσομοιωμένα δεδομένα , αφού στο Βήμα 2 τα επαρκή στατιστικά $\mathbf{s}$ δεν άλλαξαν.

Βήμα 5 : *Έλεγχος σύγκλισης*

Επαναλαμβάνουμε τα βήματα 1 έως 4 μέχρι την τελική σύγκλιση του αλγορίθμου, και υπολογίζουμε την ζητούμενη $\tilde{\mathrm{p}}$ -value (4.3).

Για τον έλεγχο της σύγκλισης οι *Booth & Butler* πρότειναν να συνεχιστούν οι επαναλήψεις του αλγορίθμου μέχρι το απόλυτο σφάλμα (*absolute error*) να εκτιμηθεί ότι είναι μικρότερο από ένα προκαθορισμένο επίπεδο ε , με 100(1-α)% εμπιστοσύνη. Δηλαδή

$$AE := \left| z_{\alpha/2} \right| \cdot \frac{\widetilde{\sigma}}{\sqrt{N}} \le \varepsilon$$

όπου

$$\widetilde{\sigma} = \frac{1}{\overline{w}^*} \cdot \sqrt{\frac{1}{N} \sum_{k=1}^{N} \left(u_k^* - w_k^* \cdot \widetilde{P} \right)^2} \tag{4.8}$$

είναι η προσέγγιση της τυπικής απόκλισης της $\widetilde{P} \equiv \widetilde{p}$-value (4.3), βασισμένη στις N επαναλήψεις του αλγορίθμου, με $u_k^* = w_k^* \cdot I\left(D_k^* \ge D_{obs}\right)$ και $\overline{w}^* = \frac{1}{N} \sum_{k=1}^{N} w_k^*$. Η τυπική απόκλιση $\widetilde{\sigma}$ της $\widetilde{p}$-value υπολογίζεται με την μέθοδο Δέλτα. Στο Παράρτημα Β παραθέτουμε μια σκιαγράφηση της απόδειξης αυτής.

Εναλλακτικά θα μπορούσαμε να σταματήσουμε τις επαναλήψεις του αλγορίθμου με το κριτήριο του σχετικού σφάλματος (*relative error*), δηλαδή όταν

$$RE := \frac{AE}{\widetilde{P}} \le \varepsilon.$$

Στο παράδειγμα της §4.5 κατά την υλοποίηση του αλγορίθμου, χρησιμοποιούμε το 5% κριτήριο σχετικού σφάλματος, ε=0,05, με 99% επίπεδο σημαντικότητας, α=0,01. Επιπλέον στον αλγόριθμο προσθέσαμε και το μέτρο του συντελεστή μεταβλητότητας (*coefficient of variation*) το οποίο επιθυμούμε να είναι πολύ μικρό (βλ. *Liu* [41], σελ.35)

$$c.v.^2\left(\overline{w}^*\right) = \frac{\sum_{i=1}^{N} \left(w_i^*\right)^2}{\left(\sum_{i=1}^{N} w_i^*\right)^2} - \frac{1}{N}.$$

4.4 Ο πίνακας σχεδιασμού

Ο αλγόριθμος που περιγράψαμε στην προηγούμενη παράγραφο είναι ένας γενικός αλγόριθμος. Δηλαδή μπορεί να βρεί εφαρμογή για οποιοδήποτε μοντέλο για το οποίο μας ενδιαφέρει ο έλεγχος καλής προσαρμογής του (π.χ. ανεξαρτησίας, ψευδοσυμμετρίας, ομοιόμορφης συνάφειας κ.α.).

Το μόνο που πρέπει να εισάγουμε κάθε φορά στον αλγόριθμο είναι ο κατάλληλος πίνακας σχεδιασμού **X** και η κατάλληλη σ.π.π. $f(\mathbf{y};\boldsymbol{\mu})$. Στα δικά μας παραδείγματα έχουμε υποθέσει από την αρχή ότι η $f(\mathbf{y};\boldsymbol{\mu})$ θα είναι η Poisson.

Το ενδιαφέρον λοιπόν θα εστιάζεται κάθε φορά στον προσεκτικό υπολογισμό του πίνακα σχεδιασμού. Ας το δούμε αυτό μέσω δύο παραδειγμάτων για τον έλεγχο ανεξαρτησίας και ψευδοσυμμετρίας.

4.4.1 Ανεξαρτησία

Ας υποθέσουμε ότι στόχος μας είναι να ελέγξουμε την καλή προσαρμογή ενός μοντέλου ανεξαρτησίας σε έναν *I*x*J* πίνακα συνάφειας.

Πίνακας 1

	Y				
X	1	2	…	*J*-1	*J*
1	y_{11}	y_{12}	…	$y_{1,J\text{-}1}$	y_{1J}
2	y_{21}	y_{22}	…	$y_{2,J\text{-}1}$	y_{2J}
…	…	…	…	…	…
I-1	$y_{I\text{-}1,1}$	$y_{I\text{-}1,2}$	…	$y_{I\text{-}1,J\text{-}1}$	$y_{I\text{-}1,J}$
I	y_{I1}	y_{I2}	…	$y_{I,J\text{-}1}$	y_{IJ}

Το λογαριθμογραμμικό μοντέλο είναι $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$, i=1,2,…I , j=1,2,…,J με περιορισμούς $\lambda_I^X = \lambda_J^Y = 0$ ή ισοδύναμα εκφρασμένο σε διανυσματική μορφή ως $\log(\boldsymbol{\mu}) = \mathbf{X}\cdot\boldsymbol{\beta}$, όπου $\boldsymbol{\mu}_{nx1} = (\mu_{11}\ \ \mu_{12}\ \ldots\ \mu_{IJ})'$ το διάνυσμα των αναμενόμενων συχνοτήτων, $\boldsymbol{\beta}_{px1} = (\lambda\ \ \lambda_1^X\ \ \ldots\ \ \lambda_{I-1}^X\ \ \lambda_1^Y\ \ \ldots\ \ \lambda_{J-1}^Y)'$ το διάνυσμα των αγνώστων παραμέτρων και **X** ο *nxp* πίνακας σχεδιασμού, με $p=I+J-1$, $n=IJ$ και $n-p=(I-1)(J-1)$.

Ο πίνακας σχεδιασμού συγκεκριμένα είναι

$$\mathbf{X}=\begin{pmatrix} \mathbf{1} & \mathbf{R_1} & \mathbf{C} \\ \mathbf{1} & \mathbf{R_2} & \mathbf{C} \\ \vdots & \vdots & \vdots \\ \mathbf{1} & \mathbf{R_{I-1}} & \mathbf{C} \\ \mathbf{1} & \mathbf{0} & \mathbf{C} \end{pmatrix}$$

όπου $\mathbf{1}_{Jx1}=\begin{pmatrix}1 & 1 & \dots & 1\end{pmatrix}'$, $\mathbf{0}_{Jx(I-1)}=\begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix}$, $\mathbf{C}_{Jx(J-1)}=\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \end{pmatrix}=\begin{pmatrix} \mathbf{I}_{J-1} \\ \mathbf{0}_{1x(J-1)} \end{pmatrix}$

και $\mathbf{R}_1=\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}$ με διαστάσεις Jx(I-1), $\mathbf{R}_2=\begin{pmatrix} 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \end{pmatrix}$ κ.ο.κ.

Όπως έχουμε δεί από την παράγραφο §4.3, στο βήμα 2 του αλγορίθμου πρέπει να επιλύσουμε το σύστημα (4.7)

$$\mathbf{X}_2' \cdot \mathbf{Y}_2^* = \mathbf{s} - \mathbf{X}_1' \cdot \mathbf{Y}_1^*$$

Αν ως $\mathbf{X}_2$ επιλέξουμε τον υποπίνακα του $\mathbf{X}$ που σχηματίζεται από τις τελευταίες p γραμμές του, τότε υπάρχει το πρόβλημα ότι ο $\mathbf{X}_2$ έχει μηδενική ορίζουσα και συνεπώς δεν αντιστρέφεται ώστε να λυθεί το σύστημα. Πρέπει λοιπόν να αναδιατάξουμε τις γραμμές του $\mathbf{X}$, ώστε οι τελευταίες p γραμμές του να σχηματίσουν έναν υποπίνακα $\mathbf{X}_2$ πλήρους τάξης (*full rank*). Συγκεκριμένα λοιπόν ο πίνακας σχεδιασμού θα αναδιαταχθεί και θα γίνει $\mathbf{X}=\begin{pmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{pmatrix}$, όπου ο $\mathbf{X}_1$ αποτελείται από τις n-p γραμμές του $\mathbf{X}$, που αντιστοιχούν στα κελιά του υποπίνακα (I-1)x(J-1) του πίνακα 1, επιλεγμένα κατά γραμμή, ενώ ο $\mathbf{X}_2$ αποτελείται από τις υπόλοιπες p γραμμές του $\mathbf{X}$, που αντιστοιχούν στα εναπομείναντα κελιά της στήλης J του πίνακα 1, εκτός του κελιού (I, J), και ακολουθούν τα κελιά της τελευταίας γραμμής. Είναι λοιπόν :

$$\mathbf{X}_1=\begin{pmatrix} \mathbf{1}_{(J-1)x1} & \mathbf{R_1} & \mathbf{I}_{J-1} \\ \mathbf{1}_{(J-1)x1} & \mathbf{R_2} & \mathbf{I}_{J-1} \\ \vdots & \vdots & \vdots \\ \mathbf{1}_{(J-1)x1} & \mathbf{R}_{I-1} & \mathbf{I}_{J-1} \end{pmatrix}_{(n-p)xp} \quad , \quad \mathbf{X}_2=\begin{pmatrix} \mathbf{1}_{(I+J-2)x1} & \mathbf{I}_{I+J-2} \\ 1 & \mathbf{0}_{1x(I+J-2)} \end{pmatrix}_{pxp}$$

όπου $\mathbf{R}_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}$ με διαστάσεις $(J\text{-}1)\text{x}(I\text{-}1)$, $\mathbf{R}_2 = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \end{pmatrix}$ κ.ο.κ.

Έτσι πετυχαίνουμε ο $\mathbf{X}_2$ να είναι ένας πίνακας πλήρους τάξης, και το σύστημα (4.7) επιλύεται άμεσα.

Βέβαια αντίστοιχα πρέπει να αναδιατάξουμε και το διάνυσμα **Y.** Πρώτα αναπτύσσουμε κατά γραμμή, τον υποπίνακα $(I\text{-}1)\text{x}(J\text{-}1)$ του πίνακα 1, ακολουθεί η τελευταία στήλη εκτός από το y_{IJ}, και τέλος η τελευταία γραμμή. Θα είναι λοιπόν :

$$\mathbf{Y} = (y_{11} \ \dots \ y_{1(J-1)} \ \ y_{21} \ \dots \ y_{2(J-1)} \ \dots \ y_{(I-1)1} \ \dots \ y_{(I-1)(J-1)} \ \ y_{1J} \ \dots \ y_{(I-1)J} \ \ y_{I1} \ \dots \ y_{IJ})'$$

4.4.2 Ψευδοσυμμετρία (*quasi-symmetry*)

Ας υποθέσουμε ότι στόχος μας είναι ο έλεγχος της καλής προσαρμογής του μοντέλου ψευδοσυμμετρίας (*quasi-symmetry*) σε έναν τεραγωνικό *I*x*I* πίνακα συνάφειας.

Πίνακας 2

	Y				
X	1	2	…	*I*-1	*I*
1	y_{11}	y_{12}	…	$y_{1,I\text{-}1}$	y_{1I}
2	y_{21}	y_{22}	…	$y_{2,I\text{-}1}$	y_{2I}
…	…	…	…	…	…
I-1	$y_{I\text{-}1,1}$	$y_{I\text{-}1,2}$	…	$y_{I\text{-}1,I\text{-}1}$	$y_{I\text{-}1,I}$
I	y_{I1}	y_{I2}	…	$y_{I,I\text{-}1}$	y_{II}

Όπως έχουμε ήδη αναφέρει στην §2.4, το λογαριθμογραμμικό μοντέλο της ψευδοσυμμετρίας (QS) είναι $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$, $i, j = 1,...,I$, όπου $\lambda_{ij}^{XY} = \lambda_{ji}^{XY}$, υπο τους περιορισμούς

$$\lambda_I^X = \lambda_I^Y = 0 \text{ και } \lambda_{iI}^{XY} = \lambda_{Ij}^{XY} = 0 \ , \ i, j = 1,\dots,I \qquad (4.9)$$

Ισοδύναμα, σε διανυσματική μορφή γράφεται ως

$$\log(\boldsymbol{\mu}) = \mathbf{X} \cdot \boldsymbol{\beta}$$

όπου $\boldsymbol{\mu}_{nx1} = (\mu_{11} \ \ \mu_{12} \ \dots \ \mu_{II})'$ το διάνυσμα των αναμενόμενων συχνοτήτων,

$$\boldsymbol{\beta}_{px1} = (\lambda \quad \lambda_1^X \quad \dots \quad \lambda_{I-1}^X \quad \lambda_1^Y \quad \dots \quad \lambda_{I-1}^Y \quad \lambda_{11}^{XY} \quad \lambda_{12}^{XY} \quad \dots \quad \lambda_{1,I-1}^{XY}$$
$$\lambda_{22}^{XY} \quad \lambda_{23}^{XY} \quad \dots \quad \lambda_{2,I-1}^{XY} \quad \dots \quad \lambda_{I-1,I-1}^{XY})'$$

το διάνυσμα των αγνώστων μή πλεοναζουσών (*non-redundant*) παραμέτρων, και **X** ο *n*x*p* πίνακας σχεδιασμού, με $n = I^2$ και $p = 1+(I\text{-}1)+(I\text{-}1)+\frac{I \cdot (I-1)}{2}$.

Πριν διαμερίσουμε τον πίνακα σχεδιασμού **X** στους δύο υποπίνακες $\mathbf{X}_1$, $\mathbf{X}_2$ με διαστάσεις (*n*-*p*)x*p* και *p*x*p* αντίστοιχα, θα πρέπει να αναδιατάξουμε τις γραμμές του **X**, έτσι ώστε οι τελευταίες *p* γραμμές του να σχηματίσουν έναν υποπίνακα $\mathbf{X}_2$ πλήρους τάξης (*full rank*). Με αυτόν τον τρόπο εξασφαλίζουμε ότι ο υποπίνακας $\mathbf{X}_2$ αντιστρέφεται και άρα το σύστημα (4.7) έχει λύση.

Συγκεκριμένα, ο $\mathbf{X}_1$ αποτελείται από τις *n*-*p* γραμμές του **X** οι οποίες αντιστοιχούν στα κελιά του τριγώνου που σχηματίζεται στον πίνακα 2 κάτω από την κεντρική διαγώνιο, επιλεγμένα κατά γραμμή, ενώ ο $\mathbf{X}_2$ αποτελείται από τις υπόλοιπες *p* γραμμές του **X**, οι οποίες αντιστοιχούν στα εναπομείναντα κελιά του πίνακα 2, επιλεγμένα κατά γραμμή.

Με την ίδια λογική θα αναδιατάξουμε και το διάνυσμα **Y**. Πρώτα αναπτύσουμε κατά γραμμή το τρίγωνο που σχηματίζεται κάτω από την κεντρική διαγώνιο, και στην συνέχεια ακολουθούν τα υπόλοιπα κελιά με ανάπτυξη κατά γραμμή. Θα είναι λοιπόν :

$$\mathbf{Y} = (y_{21} \quad y_{31} \quad y_{32} \quad y_{41} \quad y_{42} \quad y_{43} \quad \dots \quad y_{I-1,1} \quad y_{I-1,2} \quad \dots \quad y_{I-1,I-2} \quad y_{11} \quad y_{12} \quad \dots \quad y_{1I}$$
$$y_{22} \quad y_{23} \quad \dots \quad y_{2I} \quad \dots \quad y_{I-1,I-1} \quad y_{I-1,I} \quad y_{I1} \quad y_{I2} \quad \dots \quad y_{II})'$$

Παρατήρηση

Η αναδιάταξη που κάναμε στους **X** και **Y**, έγινε με την προϋπόθεση ότι *«το άθροισμα των συμμετρικών κελιών είναι μεγαλύτερο του μηδενός*, $y_{ij} + y_{ji} > 0$, $i, j = 1,...,I$ *»*.

Η προϋπόθεση αυτή είναι πολύ σημαντική, διότι αν το άθροισμα κάποιων συμμετρικών κελιών προκύψει μηδέν ($y_{ij} + y_{ji} = 0$), από τις εξισώσεις πιθανοφάνειας του QS μοντέλου (βλ. §2.4) θα προκύψουν μηδενικές αναμενόμενες συχνότητες για τα κελιά αυτά. Πράγματι

$$\left.\begin{array}{c} y_{ij} + y_{ji} = 0 \\ \mu_{ij} + \mu_{ji} = y_{ij} + y_{ji} \end{array}\right\} \Rightarrow \mu_{ij} + \mu_{ji} = 0 \Rightarrow \mu_{ij} = \mu_{ji} = 0$$

Συμπεραίνουμε λοιπόν ότι αν υπάρχουν μηδενικά σε συμμετρικές θέσεις, τότε αυτά θα θεωρούνται δομικά μηδενικά (*structural zeros*) και θα πρέπει να εξαιρεθούν από την μελέτη

μας. Στην περίπτωση αυτή, ιδιαίτερη προσοχή χρειάζεται στον επαναπροσδιορισμό του συνόλου (n) των κελιών καθώς και του πλήθους (p) των παραμέτρων.

Για παράδειγμα στον 8x8 πίνακα συνάφειας που είδαμε στην §2.4, (για ευκολία στην παρουσίαση δεν χρησιμοποιούμε τον πίνακα με τα δεδομένα, αλλά τον παρακάτω πίνακα 3)

Πίνακας 3

	Y							
X	1	2	3	4	5	6	7	8
1	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8
2	2,1	2,2	2,3	2,4	2,5	**2,6**	**2,7**	**2,8**
3	3,1	3,2	3,3	3,4	3,5	**3,6**	**3,7**	**3,8**
4	4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8
5	5,1	5,2	5,3	5,4	5,5	5,6	**5,7**	**5,8**
6	6,1	**6,2**	**6,3**	6,4	6,5	6,6	6,7	6,8
7	7,1	**7,2**	**7,3**	7,4	**7,5**	7,6	7,7	7,8
8	8,1	**8,2**	**8,3**	8,4	**8,5**	8,6	8,7	8,8

παρατηρούμε ότι υπάρχουν οκτώ ζεύγη συμμετρικών μηδενικών κελιών (τα μαύρα κελιά στον πίνακα 3) :

$$(2,6)\leftrightarrow(6,2),\ (2,7)\leftrightarrow(7,2),\ (2,8)\leftrightarrow(8,2),\ (3,6)\leftrightarrow(6,3),$$
$$(3,7)\leftrightarrow(7,3),\ (3,8)\leftrightarrow(8,3),\ (5,7)\leftrightarrow(7,5),\ (5,8)\leftrightarrow(8,5).$$

Αυτό έχει ως συνέπεια να δημιουργηθούν 16 δομικά μηδενικά στην μελέτη μας, τα οποία και πρέπει να εξαιρεθούν. Οπότε n = 64-16 = 48.

Επιπλέον και οι μή πλεονάζουσες παράμετροι θα ελαττωθούν σε p = 43-8 = 35. Η αιτία της αφαίρεσης οκτώ παραμέτρων θα γίνει αντιληπτή αν εστιάσουμε την προσοχή μας στο τμήμα του πίνακα 3, πάνω από την κεντρική διαγώνιο.

Αρχικά αφαιρούμε τις πέντε παραμέτρους αλληλεπίδρασης που αντιστοιχούν στα μαύρα κελιά (2,6), (2,7), (3,6), (3,7) και (5,7) του πίνακα 3, διότι τα κελιά αυτά περιλαμβάνουν δομικά μηδενικά. Επιπλέον αφαιρούμε τις τρείς παραμέτρους αλληλεπίδρασης που αντιστοιχούν στα κελιά (2,5), (3,5) και (5,6) του πίνακα 3, αφού η ύπαρξη δομικών μηδενικών στις γραμμές 2, 3 και 5 μετατρέπει τους περιορισμούς (4.9) σε $\lambda_{25}^{XY}=0$, $\lambda_{35}^{XY}=0$

και $\lambda_{56}^{XY} = 0$ αντίστοιχα. Δηλαδή στον πίνακα 3, με γκρί χρωματίζουμε τα redundant κελιά, λόγω των περιορισμών (4.9).

Οι παράμετροι λοιπόν θα είναι

$$\boldsymbol{\beta}_{35x1} = (\lambda \quad \lambda_1^X \quad \ldots \quad \lambda_7^X \quad \lambda_1^Y \quad \ldots \quad \lambda_7^Y \quad \lambda_{11}^{XY} \quad \lambda_{12}^{XY} \quad \ldots \quad \lambda_{17}^{XY}$$
$$\lambda_{22}^{XY} \quad \lambda_{23}^{XY} \quad \lambda_{24}^{XY} \quad \lambda_{33}^{XY} \quad \lambda_{34}^{XY} \quad \lambda_{44}^{XY} \quad \lambda_{45}^{XY} \quad \lambda_{46}^{XY} \quad \lambda_{47}^{XY} \quad \lambda_{55}^{XY} \quad \lambda_{66}^{XY} \quad \lambda_{67}^{XY} \quad \lambda_{77}^{XY})'$$

Έχοντας καθορίσει το πλήθος των κελιών (n) και των παραμέτρων (p), κατασκευάζουμε τον πίνακα σχεδιασμού $\mathbf{X}_{48x35}$ του οποίου τις γραμμές αναδιατάσσουμε, όπως έχουμε ήδη περιγράψει, έτσι ώστε ο υποπίνακας $\mathbf{X}_1$ να έχει διαστάσεις 13x35, και ο διαστάσεων 35x35 υποπίνακας $\mathbf{X}_2$ να αντιστρέφεται.

Με την ίδια λογική θα αναδιατάξουμε και το διάνυσμα **Y**. Συγκεκριμένα για τον 8x8 πίνακα συνάφειας της §2.4 θα είναι :

Y = (27, 4, 9, 26, 26, 10, 3, 4, 1, 7, 1, 1, 1,
314, 63, 10, 15, 0, 1, 1, 0, 625, 2, 5, 0, 835, 20, 1, 1096, 0, 4, 0, 0,
477, 1, 421, 0, 0, 112, 11, 1, 6, 0, 1, 0, 1, 30, 347)′

4.5 Παράδειγμα

Στον παρακάτω 4x4 πίνακα συνάφειας καταγράφηκαν οι απαντήσεις από 91 παντρεμένα ζευγάρια στην Αριζόνα των Ηνωμένων Πολιτειών, όταν τους τέθηκε η ερώτηση "Πόσο συχνά για εσάς, η ερωτική επαφή είναι ευχάριστη;".

	Απάντηση της συζύγου (Y)			
Απάντηση του συζύγου (X)	Ποτέ ή σπάνια	Αρκετά συχνά	Πολύ συχνά	Σχεδόν πάντα
Ποτέ ή σπάνια	7	7	2	3
Αρκετά συχνά	2	8	3	7
Πολύ συχνά	1	5	4	9
Σχεδόν πάντα	2	8	9	14

Πηγή *: Agresti* [1], σελ.65

Στόχος μας είναι να ελέγξουμε κατά πόσο οι απαντήσεις των δύο συζύγων είναι ανεξάρτητες μεταξύ τους, δηλαδή θα ελέγξουμε την ανεξαρτησία των X,Y μεταβλητών. Πρακτικά ελέγχουμε την καλή προσαρμογή του λογαριθμογραμμικού μοντέλου $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$, i=1,..4 , j=1,…,4 , έναντι του κορεσμένου μοντέλου.

Από την §4.4.1 ο πίνακας σχεδιασμού (μετά την αναδιάταξη) θα είναι

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

ενώ και το διάνυσμα των δεδομένων θα είναι

$$\mathbf{Y} = \{7, 7, 2, 2, 8, 3, 1, 5, 4, 3, 7, 9, 2, 8, 9, 14\}$$

Υλοποιώντας σε Mathematica τον αλγόριθμο που περιγράψαμε στις προηγούμενες παραγράφους (στο Παράρτημα Δ.1 τον παραθέτουμε ολόκληρο), και ύστερα από 20000 επαναλήψεις, λαμβάνουμε τα εξής αποτελέσματα :

```
PValueImpSamplingSimulation[20000];
Estimated p value = 0.109095
Estimated standard deviation = 0.00247283
99% confidence interval: [0.102725, 0.115465]
wstar mean coeff. of variation =7.40148 x 10^-6
Rejected samples (some ystar<0) = 4.275%
```

Βρήκαμε ότι $\tilde{p}$ – value =0,109095 > 0,01 , και συνεπώς οδηγούμαστε στο συμπέρασμα ότι σε ε.σ. 99% οι απαντήσεις των δύο συζύγων είναι ανεξάρτητες μεταξύ τους. Η προσέγγιση της τυπικής απόκλισης είναι $\tilde{\sigma} = 0{,}0247283$ ενώ ένα 99% δ.ε. για το $\tilde{p}$ – value είναι [0,102725, 0,115465].

Από την πολύ μικρή τιμή του συντελεστή μεταβλητότητας $c.v.^2\left(\overline{w}^*\right)$=7,40148·10^{-6}, έχουμε καλή ένδειξη για την αποτελεσματικότητα του αλγορίθμου, ενώ από τα 20000 δείγματα που παραγάγαμε, χρειάστηκε να απορρίψουμε μόνο το 4,275% από αυτά, αφού είχαν μηδενικά βάρη (*weights*).

Βέβαια στο παράδειγμα που επιλέξαμε είναι εφικτός ο υπολογισμός της ακριβούς p-value με τις μεθόδους που αναλύσαμε στο κεφάλαιο 2 ή και με την εναλλακτική Monte Carlo προσέγγιση (4.2), αφού μπορούμε να προσομοιώσουμε τιμές απευθείας από την ακριβή δεσμευμένη κατανομή $f\left(\mathbf{y} \mid \mathbf{s}\right)$.

Πράγματι οι *Booth & Butler* με χρήση του στατιστικού πακέτου *StatXact* [57] υπολόγισαν την ακριβή p-value=0,1137. Το μειονέκτημα είναι ότι απαιτήθηκαν περίπου 13 λεπτά και 20 δευτερόλεπτα για το αποτέλεσμα αυτό, σε σύγκριση με τα μόλις 14 δευτερόλεπτα που χρειάστηκαν με την χρήση του αλγορίθμου.

ΚΕΦΑΛΑΙΟ 5

Η μέθοδος Gibbs Sampling για ακριβή δεσμευμένα τέστ σε πίνακες συνάφειας

5.1 Εισαγωγή

Όταν έχουμε να αντιμετωπίσουμε μοντέλα για μεγάλους πίνακες συνάφειας, η Monte Carlo προσέγγιση της p-value με χρήση Importance Sampling δεν είναι αποτελεσματική. Η εναλλακτική λύση, όπως έχουμε ήδη αναφέρει, είναι η αντιμετώπιση του προβλήματος με χρήση μιας μεθόδου MCMC. Επειδή δεσμεύουμε ως προς τα επαρκή στατιστικά των οχληρών παραμέτρων, συνήθως υπάρχει δυσκολία στο να παράγουμε τιμές από την πολυδιάστατη δεσμευμένη κατανομή των επαρκών στατιστικών για τις παραμέτρους που μας ενδιαφέρουν. Οι *Forster, McDonald & Smith* [24] πρότειναν μια εναλλακτική προσέγγιση με χρήση της τεχνικής Gibbs Sampling (βλ.Παράρτημα Α.2).

Στην επόμενη παράγραφο θα περιγράψουμε την προσαρμογή της ιδέας των *Forster, McDonald & Smith* στο πλαίσιο των πινάκων συνάφειας. Στην §5.3 περιγράφουμε τα βήματα του σχετικού αλγορίθμου, για έναν 2x2x2 πίνακα συνάφειας, ενώ η υλοποίηση του αλγορίθμου με χρήση της γλώσσας προγραμματισμού Fortran, παρατίθεται ολόκληρη στο Παράρτημα Δ.2.

5.2 Το μοντέλο και η μέθοδος Gibbs Sampling

Έστω τα δεδομένα, $\mathbf{y} = (y_1 \;\; y_2 \;\; \ldots \;\; y_n)'$, προερχόμενα από ένα κορεσμένο Poisson λογαριθμογραμμικό μοντέλο με μέσες τιμές $\boldsymbol{\mu} = (\mu_1 \;\; \mu_2 \;\; \ldots \;\; \mu_n)'$. Για λόγους απλότητας θα χρησιμοποιήσουμε την διανυσματική έκφραση του λογαριθμογραμμικού μοντέλου, το οποίο γράφεται :

$$\log(\mu_i) = \mathbf{x}_i' \cdot \boldsymbol{\beta} = \sum_{j=1}^{p} x_{ij} \cdot \beta_j \;\; , \;\; i = 1,2,\ldots n$$

όπου $\boldsymbol{\beta} = \left(\beta_1 \;\; \beta_2 \;\; \ldots \;\; \beta_p\right)'$ είναι το p-διάνυσμα των αγνώστων παραμέτρων, με αντίστοιχο διάνυσμα συμμεταβλητών για κάθε μία από αυτές, το $\mathbf{x}_i' = \left(x_{i1} \;\; x_{i2} \;\; \ldots \;\; x_{ip}\right)$.

Έστω $\boldsymbol{\beta}_R$ το διάνυσμα των r-παραμέτρων ($r<p$) που μας ενδιαφέρει ο έλεγχος της σημαντικότητάς τους και $\boldsymbol{\beta}_{\setminus R}$ το διάνυσμα των υπολοίπων οχληρών παραμέτρων. Δηλαδή $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_R' \;\; \boldsymbol{\beta}_{\setminus R}'\right)'$. Συμβολίζουμε επίσης με $\mathbf{z}_R$ το επαρκές στατιστικό για το διάνυσμα $\boldsymbol{\beta}_R$, και αντίστοιχα με $\mathbf{z}_{\setminus R}$, το επαρκές στατιστικό για το διάνυσμα $\boldsymbol{\beta}_{\setminus R}$. Στόχος μας είναι ο έλεγχος της υπόθεσης H_0: $\boldsymbol{\beta}_R = 0$ έναντι της εναλλακτικής H_1: $\boldsymbol{\beta}_R \neq 0$ η οποία αντιστοιχεί στο κορεσμένο μοντέλο.

Η r-διάστατη κατανομή του $\mathbf{z}_R$,δεσμεύοντας ως προς $\mathbf{z}_{\setminus R}$, υπό την μηδενική υπόθεση, είναι

$$f\left(\mathbf{z}_R \middle| \mathbf{z}_{\setminus R}\right) \propto \left\{ \prod_{i=1}^{n} \left(\sum_{j=1}^{p} x^{ji} \cdot z_j \right)! \right\}^{-1} \quad \text{(βλ. Παράρτημα Γ)} \qquad (5.1)$$

όπου

- $z_j = \sum_{i=1}^{n} y_i \cdot x_{ij}$, $j=1,2,\ldots,p$ είναι το επαρκές στατιστικό για την β_j παράμετρο.
- $\left(\mathbf{X}^{-1}\right)' = \left(x^{ji}\right)$ είναι ο ανάστροφος του αντιστρόφου του πίνακα σχεδιασμού $\mathbf{X}_{nxp} = \left(x_{ij}\right)$.

Όμως η προσομοίωση τιμών απευθείας από την πολυδιάστατη κατανομή (5.1) δεν είναι εφικτή, αφού εξαρτάται από μια σταθερά κανονικοποίησης (*normalizing constant*) που είναι δύσκολο να καθοριστεί. Εξαιρείται η περίπτωση που έχουμε πίνακα συνάφειας δύο διαστάσεων και ελέγχουμε ανεξαρτησία, αφού τότε η κατανομή θα ήταν η πολυδιάστατη υπεργεωμετρική κατανομή και συνεπώς εύκολα θα μπορούσαμε να προσομοιώσουμε παρατηρήσεις από αυτήν.

Η μέθοδος Gibbs Sampling μας επιτρέπει να παράγουμε τιμές από μια πολυδιάστατη δεσμευμένη κατανομή, όπως η (5.1), προσομοιώνοντας διαδοχικά τιμές από τις μονομεταβλητές πλήρεις δεσμευμένες κατανομές της (*full conditional distributions*).

Συγκεκριμένα στο δικό μας πρόβλημα στόχος είναι να παράγουμε τιμές από την r-διάστατη κατανομή $\mathbf{z}_R = \left(z_1, z_2, \ldots, z_r\right) \sim f\left(\mathbf{z}_R \middle| \mathbf{z}_{\setminus R}\right)$. Όμως είναι ευκολότερο να παράγουμε τιμές από τις πλήρεις δεσμευμένες κατανομές της.

Πράγματι έστω z_k ένα στοιχείο του $\mathbf{z}_R$. Τότε η μονομεταβλητή κατανομή του, δεσμεύοντας ως προς τα υπόλοιπα στοιχεία του διανύσματος $\mathbf{z}=(z_1, z_2, \ldots, z_p)$, είναι :

$$f(z_k|\mathbf{z}_{\setminus k}) = f(z_k|z_1,\ldots,z_{k-1},z_{k+1},\ldots,z_p) \propto \left\{\prod_{i=1}^{n}(c_i^k + x^{ki}\cdot z_k)!\right\}^{-1},\ c_i^k = \sum_{j\neq k} x^{ji}\cdot z_j \quad (5.2)$$

Το τελευταίο ζήτημα που θέλει προσοχή για να μπορέσουμε να παράγουμε τιμές από αυτές τις μονομεταβλητές δεσμευμένες κατανομές, είναι ο καθορισμός του στηρίγματος (*support*), καθώς και η εύρεση της σταθεράς κανονικοποίησής τους .

Για την καλύτερη κατανόηση των παραπάνω, στην επόμενη παράγραφο παραθέτουμε την περιγραφή της μεθόδου σε 2x2x2 πίνακα συνάφειας, για τον έλεγχο καλής προσαρμογής ιεραρχικών λογαριθμογραμμικών μοντέλων.

5.3 Περιγραφή της μεθόδου Gibbs Sampling σε 2x2x2 πίνακα συνάφειας

Έστω ένας 2x2x2 πίνακας συνάφειας και X, Y, Z οι τρείς μεταβλητές ταξινόμησης αντίστοιχα. Συνήθως στόχος μας είναι ο έλεγχος καλής προσαρμογής κάποιου ιεραρχικού μοντέλου, έναντι κάποιου μεγαλύτερου ιεραρχικού μοντέλου (π.χ. κορεσμένο (XYZ)).

Ας θεωρήσουμε ότι ως εναλλακτικό μοντέλο έχουμε το κορεσμένο μοντέλο

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ},\quad i=1,2,\ j=1,2,\ k=1,2.$$

Υπο τον περιορισμό ότι οι δεύτερης κατηγορίας παράμετροι είναι ίσοι με μηδέν, έχει οκτώ μή πλεονάζουσες παραμέτρους, τις $\{\lambda, \lambda_1^X, \lambda_1^Y, \lambda_1^Z, \lambda_{11}^{XY}, \lambda_{11}^{XZ}, \lambda_{11}^{YZ}, \lambda_{111}^{XYZ}\}$, ενώ σε διανυσματική μορφή γράφεται $\log(\boldsymbol{\mu}) = \mathbf{X}\cdot\boldsymbol{\beta}$, όπου

$$\boldsymbol{\mu} = (\mu_{111}\quad \mu_{121}\quad \mu_{112}\quad \mu_{122}\quad \mu_{211}\quad \mu_{221}\quad \mu_{212}\quad \mu_{222})'$$

$$\boldsymbol{\beta} = (\lambda_{111}^{XYZ}\quad \lambda_{11}^{XZ}\quad \lambda_{11}^{XY}\quad \lambda_1^X\quad \lambda_{11}^{YZ}\quad \lambda_1^Z\quad \lambda_1^Y\quad \lambda)'$$

και $\mathbf{X}$ ο πίνακας σχεδιασμού

$$\mathbf{X}=\begin{pmatrix}1&1&1&1&1&1&1&1\\0&1&0&1&0&1&0&1\\0&0&1&1&0&0&1&1\\0&0&0&1&0&0&0&1\\0&0&0&0&1&1&1&1\\0&0&0&0&0&1&0&1\\0&0&0&0&0&0&1&1\\0&0&0&0&0&0&0&1\end{pmatrix}$$

Τα επαρκή στατιστικά για το διάνυσμα των παραμέτρων **β** είναι αντίστοιχα $\mathbf{z}=(z_1\,,z_2\,,z_3\,,z_4\,,z_5\,,z_6\,,z_7\,,z_8)=\left(y_{111},y_{1+1},y_{11+},y_{1++},y_{+11},y_{++1},\ y_{+1+},y_{+++}\right).$

Αν το υπο έλεγχο μοντέλο είναι το

$$\text{(XZ, YZ)}:\ \log\left(\mu_{ijk}\right)=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ik}^{XZ}+\lambda_{jk}^{YZ},\quad i=1,2\ ,\ j=1,2\ ,\ k=1,2\ ,$$

τότε οι παράμετροι που μας ενδιαφέρει η σημαντικότητά τους είναι οι λ_{111}^{XYZ} και λ_{11}^{XY}, και ως εκ τούτου $\mathbf{z}_R=\left(z_1,z_3\right).$

Ανάλογα αν το υπο έλεγχο μοντέλο είναι το

$$\text{(XY, XZ)}:\ \log\left(\mu_{ijk}\right)=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ij}^{XY}+\lambda_{ik}^{XZ},\quad i=1,2\ ,\ j=1,2\ ,\ k=1,2\ ,$$

τότε οι παράμετροι που μας ενδιαφέρει η σημαντικότητά τους είναι οι λ_{111}^{XYZ} και λ_{11}^{YZ}, και ως εκ τούτου $\mathbf{z}_R=\left(z_1,z_5\right).$

Η δεσμευμένη κατανομή $f\left(z_1\,\middle|\,\mathbf{z}_{\setminus 1}\right)$, υπολογίζεται από την (5.2), και είναι

$$f\left(z_1\,\middle|\,\mathbf{z}_{\setminus 1}\right)\propto\{z_1!\cdot(z_2-z_1)!\cdot(z_3-z_1)!\cdot(z_4-z_3-z_2+z_1)!\cdot(z_5-z_1)!\cdot(z_6-z_5-z_2+z_1)!\cdot(z_7-z_5-z_3+z_1)!\cdot(z_8-z_7-z_6+z_5-z_4+z_3+z_2-z_1)!\}^{-1}\qquad(5.3)$$

Το στήριγμα της κατανομής αυτής προκύπτει εύκολα, αφού πρέπει όλοι οι παραγοντικοί όροι να είναι μή αρνητικοί. Δηλαδή

$$z_1\geq 0\,,\ \ z_1\leq z_2\ ,\ \ z_1\leq z_3\ ,\ \ z_1\geq z_2+z_3-z_4\ ,\ \ z_1\leq z_5\ ,\ \ z_1\geq z_2+z_5-z_6\ ,$$

$$z_1\geq z_3+z_5-z_7\ \ \text{και}\ \ z_1\leq z_2+z_3-z_4+z_5-z_6-z_7+z_8$$

Συνεπώς θα είναι :

$$\max[0,\ z_2+z_3-z_4\,,\ z_2+z_5-z_6\,,\ z_3+z_5-z_7\,]\leq z_1\leq$$

$$\min[\,z_2\,,\ z_3\,,\ z_5\,,\ z_2+z_3-z_4+z_5-z_6-z_7+z_8\,]\qquad(5.4)$$

Οι υπόλοιπες δεσμευμένες κατανομές $f\left(z_i \mid \mathbf{z}_{\setminus i}\right)$, i=2,3,…,8 προκύπτουν εύκολα από την (5.3) αν διατηρήσουμε τους όρους των παραγοντικών που περιέχουν το z_i, ενώ τα αντίστοιχα στηρίγματα δίνονται στον παρακάτω πίνακα.

	Supports
z_2	$\max[\, z_1 \,,\, z_1 - z_3 + z_4 - z_5 + z_6 + z_7 - z_8 \,] \le z_2 \le \min[\, z_1 - z_3 + z_4 \,,\, z_1 - z_5 + z_6 \,]$
z_3	$\max[\, z_1 \,,\, z_1 - z_2 + z_4 - z_5 + z_6 + z_7 - z_8 \,] \le z_3 \le \min[\, z_1 - z_2 + z_4 \,,\, z_1 - z_5 + z_7 \,]$
z_4	$z_3 + z_2 - z_1 \le z_4 \le z_8 - z_7 - z_6 + z_5 + z_3 + z_2 - z_1$
z_5	$\max[\, z_1 \,,\, z_1 - z_2 - z_3 + z_4 + z_6 + z_7 - z_8 \,] \le z_5 \le \min[\, z_6 - z_2 + z_1 \,,\, z_7 - z_3 + z_1 \,]$
z_6	$z_5 + z_2 - z_1 \le z_6 \le z_8 - z_7 + z_5 - z_4 + z_3 + z_2 - z_1$
z_7	$z_5 + z_3 - z_1 \le z_7 \le z_8 - z_6 + z_5 - z_4 + z_3 + z_2 - z_1$
z_8	$z_1 - z_2 - z_3 + z_4 - z_5 + z_6 + z_7 \le z_8 < \infty$

Έχοντας λοιπόν καθορίσει τα στηρίγματα για τις πλήρεις δεσμευμένες κατανομές, το τελευταίο που απομένει είναι ο υπολογισμός της σταθεράς κανονικοποίησης για κάθε μια από αυτές. Το πώς γίνεται ο υπολογισμός αυτός θα το δούμε μέσα από την αναλυτική παρουσίαση των βημάτων του αλγορίθμου.

<u>Βήμα 1</u>: *Προσομοίωση του* $\mathbf{z}_R$

- Δίνουμε κάποιες αρχικές τιμές στα επαρκή στατιστικά $\mathbf{z}^{(0)} = \left(z_1^{(0)}, z_2^{(0)}, \ldots, z_8^{(0)}\right) = \left(\mathbf{z}_R^{(0)}, \mathbf{z}_{\setminus R}^{(0)}\right)$. Είναι οι τιμές που λαμβάνουν τα επαρκή στατιστικά με βάση τον παρατηρούμενο πίνακα.

- Η σταθερά κανονικοποίησης C_{p} της $f\left(z_1 \mid z_{\setminus 1}\right)$ υπολογίζεται ως εξής :

$$\sum_{z_1=z_{1(\min)}}^{z_{1(\max)}} f\left(z_1 \,\middle|\, z_{\setminus 1}\right)=1 \Rightarrow \sum_{z_1=z_{1(\min)}}^{z_{1(\max)}} C_p \cdot A(z_1)=1 \Rightarrow C_p=\frac{1}{\sum_{z_1=z_{1(\min)}}^{z_{1(\max)}} A(z_1)}$$

Όπου $A(z_1)=\{z_1!\cdot\left(z_2^{(0)}-z_1\right)!\cdot\left(z_3^{(0)}-z_1\right)!\cdot\left(z_4^{(0)}-z_3^{(0)}-z_2^{(0)}+z_1\right)!\cdot\left(z_5^{(0)}-z_1\right)!\cdot(z_6^{(0)}-z_5^{(0)}-z_2^{(0)}+z_1)!\cdot\left(z_7^{(0)}-z_5^{(0)}-z_3^{(0)}+z_1\right)!\cdot\left(z_8^{(0)}-z_7^{(0)}-z_6^{(0)}+z_5^{(0)}-z_4^{(0)}+z_3^{(0)}+z_2^{(0)}-z_1\right)!\}^{-1}$

Τα άκρα του αθροίσματος προκύπτουν από το στήριγμα (5.4), και είναι

$$z_{1(\min)}=\max[0,\ z_2^{(0)}+z_3^{(0)}-z_4^{(0)},\ z_2^{(0)}+z_5^{(0)}-z_6^{(0)},\ z_3^{(0)}+z_5^{(0)}-z_7^{(0)}]$$

$$z_{1(\max)}=\min[z_2^{(0)},\ z_3^{(0)},\ z_5^{(0)},\ z_2^{(0)}+z_3^{(0)}-z_4^{(0)}+z_5^{(0)}-z_6^{(0)}-z_7^{(0)}+z_8^{(0)}]$$

- Έχοντας υπολογίσει την σταθερά C_p γνωρίζουμε πλήρως την μορφή της $f\left(z_1 \,\middle|\, \mathbf{z}_{\setminus 1}\right)$, και συνεπώς μπορούμε να προσομοιώσουμε μια τιμή $z_1^{(1)}$ από την κατανομή αυτή. Η προσομοίωση θα γίνει με χρήση της μεθόδου της αντιστροφής (*inverse transform method*), υλοποιώντας την παρακάτω ανακύκλωση

> **a.** Παράγουμε έναν τυχαίο αριθμό U ~ *Uniform* (0,1)
>
> **b.** Υπολογίζουμε την ποσότητα $F_{dist}=C_\text{p}\cdot A(z_1)$ για $z_1=z_{1(\min)}$
>
> **c.** Αν U < F_{dist} , τότε $z_1^{(1)}=z_1$ και σταματάμε την ανακυκλωση
>
> **d.** Αν U > F_{dist} , τότε θέτουμε $z_1=z_1+1$ και $F_{dist}=F_{dist}+C_\text{p}\cdot A(z_1)$
>
> {όπου F_{dist} είναι η αθροιστική συνάρτηση κατανομής)
>
> **e.** Πηγαίνουμε στο *βήμα c*.

Για περισότερες λεπτομέρειες σχετικά με την μέθοδο της αντιστροφής βλ. *Robert & Casella* [54] σελ.35-49.

Παραγάγαμε λοιπόν την τιμή $z_1^{(1)}\sim f\left(z_1 \,\middle|\, z_2^{(0)}, z_3^{(0)}, \ldots, z_r^{(0)}, \mathbf{z}_{\setminus R}^{(0)}\right)$, και στην συνέχεια με τον ίδιο τρόπο παράγουμε διαδοχικά τιμές για τα υπόλοιπα z_i του $\mathbf{z}_R=(z_1, z_2, \ldots, z_r)$,

$$z_2^{(1)}\sim f\left(z_2 \,\middle|\, z_1^{(1)}, z_3^{(0)}, \ldots, z_r^{(0)}, \mathbf{z}_{\setminus R}^{(0)}\right)$$

$$z_3^{(1)}\sim f\left(z_3 \,\middle|\, z_1^{(1)}, z_2^{(1)}, z_4^{(0)}, \ldots, z_r^{(0)}, \mathbf{z}_{\setminus R}^{(0)}\right)$$

$$\vdots$$

$$z_r^{(1)}\sim f\left(z_r \,\middle|\, z_1^{(1)}, z_2^{(1)}, \ldots, z_{r-1}^{(1)}, \mathbf{z}_{\setminus R}^{(0)}\right)$$

Δηλαδή μέχρις στιγμής ολοκληρώσαμε την προσομοίωση μιας τιμής του διανύσματος $\mathbf{z}_R$ των επαρκών στατιστικών των παραμέτρων που μας ενδιαφέρει η σημαντικότητά τους, $\mathbf{z}_R^{(1)} = \left(z_1^{(1)}, z_2^{(1)}, \ldots, z_r^{(1)}\right)$, ώστε $\mathbf{z}^{(1)} = \left(\mathbf{z}_R^{(1)}, \mathbf{z}_{\setminus R}^{(0)}\right)$.

Συγκεκριμένα, αν το υπο έλεγχο μοντέλο είναι το (XZ, YZ) με $\mathbf{z}_R = \left(z_1, z_3\right)$, τότε

$z_1^{(1)} \sim f\left(z_1 \mid z_3^{(0)}, \mathbf{z}_{\setminus R}^{(0)}\right)$

$z_3^{(1)} \sim f\left(z_3 \middle| z_1^{(1)}, \mathbf{z}_{\setminus R}^{(0)}\right)$

έτσι ώστε $\mathbf{z}_R^{(1)} = \left(z_1^{(1)}, z_3^{(1)}\right)$

Παρατήρηση

Στην πράξη δεν θα χρειαστεί ποτέ να προσομοιώσουμε το z_8, αφού είναι το επαρκή στατιστικό που αντιστοιχεί στην σταθερά λ, αξίζει όμως να αναφέρουμε μια ιδιαιτερότητα της προσομοίωσής του, διότι το στήριγμα της $f\left(z_8 \mid \mathbf{z}_{\setminus 8}\right)$ δεν φράσσεται από πάνω. Έχουμε δηλαδή $z_8 \sim f\left(z_8 \mid \mathbf{z}_{\setminus 8}\right) \propto \frac{1}{(z_8 - A)!}$ με στήριγμα $\mathrm{A} \le z_8 < \infty$, όπου $A = z_1^{(1)} - z_2^{(1)} - z_3^{(1)} + z_4^{(1)} - z_5^{(1)} + z_6^{(1)} + z_7^{(1)}$. Απο ιδιότητες των σειρών γνωρίζουμε ότι $\sum_{k=0}^{\infty} \frac{1}{k!} = e \Rightarrow \sum_{k=0}^{\infty} \frac{1}{k! \cdot e} = 1$, όποτε αν θέσουμε $k = (z_8 - A)$ θα έχουμε ότι

$$k = (z_8 - A) \sim f(k) = \frac{1}{k! \cdot e} \text{ , με στήριγμα } [0, \infty) \qquad (5.5)$$

Γνωρίζοντας λοιπόν πλήρως την μορφή της (5.5) μπορούμε να προσομοιώσουμε έμμεσα την τιμή $z_8^{(1)}$. Η προσομοίωση θα γίνει με χρήση της μεθόδου της αντιστροφής, υλοποιώντας την παρακάτω ανακύκλωση.

a. Παράγουμε έναν τυχαίο αριθμό U ~ *Uniform* (0,1)

b. Υπολογίζουμε την ποσότητα $f(k) = \frac{1}{k! \cdot e}$ για k=0

και ορίζουμε $f_{dens} = f(0)$ και $F_{dist} = f_{dens}$

c. Αν $\mathrm{U} < F_{dist}$, τότε $z_8^{(1)} = k + A$ και σταματάμε την ανακυκλωση

d. Αν $\mathrm{U} > F_{dist}$, τότε θέτουμε $k = k + 1$, $f_{dens} = \frac{f_{dens}}{k}$ και $F_{dist} = F_{dist} + f_{dens}$

e. Πηγαίνουμε στο *βήμα c*.

Βήμα 2: *Υπολογισμός του στατιστικού*

Ως στατιστικό επιλέξαμε να χρησιμοποιήσουμε το $X^2\left(\mathbf{y}^*;\hat{\boldsymbol{\mu}}\right)$ του Pearson. Με $\mathbf{y}^*$ συμβολίζουμε τις συχνότητες των κελιών που προκύπτουν από το διάνυσμα των επαρκών στατιστικών που παραγάγαμε, $\mathbf{y}^* = \left(\mathbf{X}^{-1}\right)' \cdot \mathbf{z}^{(1)}$ όπου $\mathbf{z}^{(1)} = \left(\mathbf{z}_R^{(1)}, \mathbf{z}_{\setminus R}^{(0)}\right)$, ενώ με $\hat{\boldsymbol{\mu}}$ συμβολίζουμε τις εκτιμήσεις των αναμενόμενων συχνοτήτων κάτω από το υπό έλεγχο μοντέλο.

Είναι σημαντικό να διευκρινήσουμε ότι οι εκτιμήσεις $\hat{\boldsymbol{\mu}}$ θα είναι σταθερές σε κάθε ανακύκλωση, αφού υπολογίζονται συναρτήσει των επαρκών στατιστικών των οχληρών παραμέτρων $\mathbf{z}_{\setminus R}^{(0)}$, τα οποία δεν αλλάζουν.

Συγκεκριμένα αν το υπο έλεγχο μοντέλο είναι το (XZ, YZ), τότε υπολογίζουμε τις συχνότητες των οκτώ κελιών του πίνακα ως εξής :

$$\begin{pmatrix} y^*_{111} \\ y^*_{121} \\ y^*_{112} \\ y^*_{122} \\ y^*_{211} \\ y^*_{221} \\ y^*_{212} \\ y^*_{222} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\ -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} z_1^{(1)} \\ z_2^{(0)} \\ z_3^{(1)} \\ z_4^{(0)} \\ z_5^{(0)} \\ z_6^{(0)} \\ z_7^{(0)} \\ z_8^{(0)} \end{pmatrix}$$

ενώ οι εκτιμήσεις των αναμενόμενων συχνοτήτων που είναι $\hat{\mu}_{ijk} = \dfrac{y_{i+k} \cdot y_{+jk}}{y_{++k}}$, συναρτήσει των $\mathbf{z}_{\setminus R}^{(0)} = \left(z_2^{(0)}, z_4^{(0)}, z_5^{(0)}, z_6^{(0)}, z_7^{(0)}, z_8^{(0)}\right)$ υπολογίζονται ως εξής :

$$E\left(y_{111}\right) = \frac{z_2^{(0)} \cdot z_5^{(0)}}{z_6^{(0)}},\ E\left(y_{121}\right) = \frac{z_2^{(0)} \cdot \left(z_6^{(0)} - z_5^{(0)}\right)}{z_6^{(0)}},\ E\left(y_{112}\right) = \frac{\left(z_4^{(0)} - z_2^{(0)}\right) \cdot \left(z_7^{(0)} - z_5^{(0)}\right)}{z_8^{(0)} - z_6^{(0)}},$$

$$E\left(y_{122}\right) = \left(z_4^{(0)} - z_2^{(0)}\right) - E\left(y_{112}\right),\ E\left(y_{211}\right) = \frac{z_5^{(0)} \cdot \left(z_6^{(0)} - z_2^{(0)}\right)}{z_6^{(0)}},\ E\left(y_{221}\right) = \frac{\left(z_6^{(0)} - z_5^{(0)}\right) \cdot \left(z_6^{(0)} - z_2^{(0)}\right)}{z_6^{(0)}},$$

$$E\left(y_{212}\right) = \left(z_7^{(0)} - z_5^{(0)}\right) - E\left(y_{112}\right),\ E\left(y_{222}\right) = z_8^{(0)} - z_7^{(0)} - z_6^{(0)} + z_5^{(0)} - z_4^{(0)} + z_2^{(0)} + E\left(y_{112}\right)$$

Βήμα 3

Δοθείσης της τιμής $\mathbf{z}^{(1)} = \left(z_1^{(1)}, z_2^{(1)}, \ldots, z_r^{(1)}, \mathbf{z}_{\setminus R}^{(0)}\right)$ παράγουμε διαδοχικά, όμοια με πρίν, τις τιμές για όλα τα z_i του $\mathbf{z}_R = \left(z_1, z_2, \ldots, z_r\right)$

$$z_1^{(2)} \sim f\left(z_1 \middle| z_2^{(1)}, z_3^{(1)}, \ldots, z_r^{(1)}, \mathbf{z}_{\setminus R}^{(0)}\right)$$

$$z_2^{(2)} \sim f\left(z_2 \middle| z_1^{(2)}, z_3^{(1)}, \ldots, z_r^{(1)}, \mathbf{z}_{\setminus R}^{(0)}\right)$$

$$\vdots$$

$$z_r^{(2)} \sim f\left(z_r \middle| z_1^{(2)}, z_2^{(2)}, \ldots, z_{r-1}^{(2)}, \mathbf{z}_{\setminus R}^{(0)}\right)$$

Παραγάγαμε λοιπόν μια ακόμα τιμή $\mathbf{z}^{(2)} = \left(z_1^{(2)}, z_2^{(2)}, \ldots, z_r^{(2)}, \mathbf{z}_{\setminus R}^{(0)}\right) = \left(\mathbf{z}_R^{(2)}, \mathbf{z}_{\setminus R}^{(0)}\right)$

Συνεχίζοντας με την ίδια λογική και ύστερα από N βήματα θα παράγουμε συνολικά N τιμές $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(N)}$. Σε κάθε βήμα υπολογίζουμε και την τιμή του στατιστικού X^2 του Pearson.

Η προσέγγιση της p-value τελικά θα είναι :

$$\tilde{\text{p}} - \text{value} = \frac{1}{N} \cdot \sum_{t=1}^{N} I\left(X^2\left(\mathbf{y}_t^*; \hat{\boldsymbol{\mu}}\right) \geq X_{obs}^2\left(\mathbf{y}; \hat{\boldsymbol{\mu}}\right)\right)$$

όπου $\mathbf{y}_t^*$ είναι το t-σετ συχνοτήτων, που προκύπτουν από το παραγόμενο $\mathbf{z}^{(t)} = \left(\mathbf{z}_R^{(t)}, \mathbf{z}_{\setminus R}^{(0)}\right)$.

5.3.1 Παράδειγμα

Στον παρακάτω 2x2x2 πίνακα συνάφειας καταγράφονται οι καταδίκες σε θανατική ποινή που επέβαλε το δικαστήριο της Florida των Ηνωμένων Πολιτειών, για φόνους που διαπράχθηκαν, μεταξύ 1976-1987. Συγκεκριμένα διασταυρώνονται η ετυμηγορία του δικαστηρίου (Y), η φυλή του θύτη (X) και η φυλή του θύματος (Z).

Φυλή θύματος	Φυλή θύτη	Θανατική ποινή	
		Ναί	Όχι
Λευκός	Λευκός	53	414
	Μαύρος	11	37
Μαύρος	Λευκός	0	16
	Μαύρος	4	139

Πηγή : M.L.Radelet and G.L.Pierce, *Florida Law Rev.*(1991); **43** : 1-34

Στόχος μας είναι να ελέγξουμε αν η ετυμηγορία του δικαστηρίου είναι ανεξάρτητη της φυλής του θύτη, δοθέντος της φυλής του θύματος. Πρακτικά θα ελέγξουμε την καλή προσαρμογή του μοντέλου της δεσμευμένης ανεξαρτησίας των X,Y δοθέντος της Z,

$$(\text{XZ, YZ}) : \log\left(\mu_{ijk}\right) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \quad i=1,2\ ,\ j=1,2\ ,\ k=1,2$$

έναντι του κορεσμένου μοντέλου.

Υλοποιώντας σε Fortran τον αλγόριθμο που περιγράψαμε στην προηγούμενη παράγραφο (στο Παράρτημα Δ.2 τον παραθέτουμε ολόκληρο), και ύστερα από 100000 επαναλήψεις, λαμβάνουμε τα εξής αποτελέσματα :

```
NUMBER OF ITERATIONS =  100000
P-value =  0.919800  +/-  0.002212
SOBS =  5.810875
```

Βρίσκουμε $\tilde{\text{p}} - \text{value} = 0{,}9198 >> 0{,}01$, οπότε οδηγούμαστε στο συμπέρασμα ότι σε ε.σ. 99% η ετυμηγορία του δικαστηρίου είναι ανεξάρτητη της φυλής του θύτη, δοθέντος της φυλής του θύματος. Επίσης ένα 99% δ.ε για την $\tilde{\text{p}} - \text{value}$ είναι [0,917, 0,972], ενώ $X_{obs}^2 = 5{,}810875$.

ΠΑΡΑΡΤΗΜΑΤΑ

Παράρτημα Α : Υπολογιστικές μέθοδοι

Α.1 Η μέθοδος Importance Sampling

Ένα συχνό αλλά δύσκολο πρόβλημα που συναντάμε στην στατιστική συμπερασματολογία είναι ο υπολογισμός αναμενόμενων τιμών σε πολυμεταβλητές κατανομές. Συγκεκριμένα αν **X** είναι ένα διάνυσμα *k*-συνεχών τυχαίων μεταβλητών με κατανομή $\pi(\cdot)$, το πρόβλημα θα είναι ο υπολογισμός της αναμενόμενης τιμής

$$E_{\pi}\left[h(\mathbf{X})\right] = \frac{\int h(\mathbf{x}) \cdot \pi(\mathbf{x}) d\mathbf{x}}{\int \pi(\mathbf{x}) d\mathbf{x}} \tag{1}$$

όπου $h(\cdot)$ μια οποιαδήποτε συνάρτηση που μας ενδιαφέρει. Για ευκολία στους συμβολισμούς θεωρήσαμε ότι το διάνυσμα **X** αποτελείται από *k*-συνεχής τυχαίες μεταβλητές. Φυσικά η μέθοδος που θα περιγράψουμε μπορεί να γενικευτεί και για περιπτώσεις όπου το διάνυσμα **X** αποτελείται από διακριτές τυχαίες μεταβλητές, μόνο που τότε τα ολοκληρώματα στις σχέσεις θα πρέπει να αντικατασταθούν από αθροίσματα. Επιπρόσθετα το **X** μπορεί να είναι και ένας συνδιασμός διακριτών και συνεχών τυχαίων μεταβλητών.

Η πιο συνηθισμένη αντιμετώπιση του προβλήματος υπολογισμού της $E_{\pi}\left[h(\mathbf{X})\right]$ είναι με χρήση της Monte Carlo ολοκλήρωσης. Σύμφωνα με την προσέγγιση αυτή αν παράγουμε *N* ακολουθίες τιμών $\mathbf{x}^{(t)} = \left\{x_i^{(t)}, i = 1,2,\ldots,k\right\}$, $t = 1,2,\ldots,N$, που πήραν οι ανεξάρτητες τ.μ. $X_1, X_2, \ldots, X_k$ με κατανομή $\pi(\cdot)$, τότε προσεγγιστικά θα ισχύει :

$$\bar{h}_N = \frac{1}{N}\sum_{t=1}^{N} h\left(\mathbf{x}^{(t)}\right) \rightarrow E_{\pi}\left[h(\mathbf{X})\right]$$

Το γεγονός ότι ο μέσος $\bar{h}_N$ θα συγκλίνει στην ζητούμενη αναμενόμενη τιμή μας το εξασφαλίζει ο Ισχυρός Νόμος των Μεγάλων Αριθμών.

Συχνά όμως η κατανομή $\pi(\cdot)$ είναι γνωστή μόνο μέχρι κάποια σταθερά κανονικοποίησης. Δηλαδή γνωρίζουμε μόνο ότι $\pi(\mathbf{x}) \propto \tilde{\pi}(\mathbf{x})$ που σημαίνει απλά $\pi(\mathbf{x}) = C_p \cdot \tilde{\pi}(\mathbf{x})$, όπου $C_p = \dfrac{1}{\int \tilde{\pi}(\mathbf{x}) d\mathbf{x}}$ είναι η σταθερά κανονικοποίησης και δεν είναι εύκολο να υπολογιστεί.

Σε τέτοιες περιπτώσεις δεν μπορούμε να παράγουμε σύνολα $\mathbf{x}^{(t)}$, $t = 1,2,\ldots,N$, απευθείας από την $\pi(\cdot)$ ώστε να εφαρμόσουμε την κλασσική Monte Carlo ολοκλήρωση. Μπορούμε όμως να εφαρμόσουμε μια παραλλαγή της η οποία καθιστά την απευθείας προσομοίωση τιμών της $\pi(\cdot)$ περιττή. Η μέθοδος αυτή είναι γνωστή με το όνομα Importance Sampling και βασίζεται στην παραγωγή συνόλων $\mathbf{x}^{(t)}$, $t = 1,2,\ldots,N$ από μια άλλη κατανομή διαφορετική της $\pi(\cdot)$, αλλά γνωστή. Έστω $g(\cdot)$ αυτή η κατανομή.

Η ζητούμενη αναμενόμενη τιμή (1) μπορεί προφανώς να γραφεί ισοδύναμα και ως

$$E_\pi[h(\mathbf{X})] = \frac{\int h(\mathbf{x}) \cdot \dfrac{\pi(\mathbf{x})}{g(\mathbf{x})} \cdot g(\mathbf{x}) d\mathbf{x}}{\int \dfrac{\pi(\mathbf{x})}{g(\mathbf{x})} \cdot g(\mathbf{x}) d\mathbf{x}} = E_g\left[\frac{\pi(\mathbf{X})}{g(\mathbf{X})} \cdot h(\mathbf{X})\right]$$

Παράγοντας λοιπόν N ακολουθίες τιμών $\mathbf{x}^{(t)}$, $t = 1,2,\ldots,N$ που πήραν οι ανεξάρτητες τ.μ. $X_1, X_2, \ldots, X_k$ με κατανομή $g(\cdot)$, προσεγγιστικά θα ισχύει :

$$\bar{h}'_N = \frac{\dfrac{1}{N}\sum_{t=1}^{N} h(\mathbf{x}^{(t)}) \cdot \dfrac{\pi(\mathbf{x}^{(t)})}{g(\mathbf{x}^{(t)})}}{\dfrac{1}{N}\sum_{t=1}^{N} \dfrac{\pi(\mathbf{x}^{(t)})}{g(\mathbf{x}^{(t)})}} = \frac{\sum_{t=1}^{N} w_t \cdot h(\mathbf{x}^{(t)})}{\sum_{t=1}^{N} w_t} \to E_g\left[\frac{\pi(\mathbf{X})}{g(\mathbf{X})} \cdot h(\mathbf{X})\right] = E_\pi[h(\mathbf{X})]$$

όπου $w_t = \dfrac{\pi(\mathbf{x}^{(t)})}{g(\mathbf{x}^{(t)})}$ το βάρος (*weight*) της t ακολουθίας τιμών $\mathbf{x}^{(t)}$.Το γεγονός ότι ο εκτιμητής $\bar{h}'_N$ θα συγκλίνει στην ζητούμενη αναμενόμενη τιμή μας το εξασφαλίζει και πάλι ο Ισχυρός Νόμος των Μεγάλων Αριθμών. Μάλιστα η σύγκλιση επιτυγχάνεται οποιαδήποτε συνάρτηση και αν επιλέξουμε ως $g(\cdot)$, αρκεί $\text{support}(g) \supseteq \text{support}(\pi)$.

Η μέθοδος λοιπόν του Importance Sampling είναι αρκετά πρακτική αφού δεν μας περιορίζει ως προς την κατανομή $g(\cdot)$ που θα επιλέξουμε, και ως εκ τούτου μας παρέχει την δυνατότητα να επιλέγουμε κατανομή από την οποία είναι εύκολο να προσομοιώσουμε.

Παρόλη την ελευθερία που έχουμε ως προς την επιλογή της $g(\cdot)$ είναι αναμενόμενο να υπάρχουν κάποιες επιλογές που είναι καλύτερες από κάποιες άλλες. Ένα μέτρο σύγκρισης

είναι η επιλογή κατανομής $g(\cdot)$ τέτοια ώστε να ελαχιστοποιείται η διακύμανση του εκτιμητή $\bar{h}'_N$. Αυτό συμβαίνει όταν $g^*(\mathbf{x}) = \dfrac{|h(\mathbf{x})| \cdot \pi(\mathbf{x})}{\int |h(\mathbf{z})| \cdot \pi(\mathbf{z}) d\mathbf{z}}$. Πράγματι η διακύμανση του εκτιμητή είναι :

$$\mathrm{Var}\left(\frac{\pi(\mathbf{x}) \cdot h(\mathbf{x})}{g(\mathbf{x})} \right) = E_g\left(\left[\frac{\pi(\mathbf{x}) \cdot h(\mathbf{x})}{g(\mathbf{x})} \right]^2 \right) - \left(E_g\left[\frac{\pi(\mathbf{x}) \cdot h(\mathbf{x})}{g(\mathbf{x})} \right] \right)^2 =$$

$$= E_g\left(\frac{\pi^2(\mathbf{x}) \cdot h^2(\mathbf{x})}{g^2(\mathbf{x})} \right) - \left(\int \frac{\pi(\mathbf{x}) \cdot h(\mathbf{x})}{g(\mathbf{x})} \cdot g(\mathbf{x}) d\mathbf{x} \right)^2 = E_g\left(\frac{\pi^2(\mathbf{x}) \cdot h^2(\mathbf{x})}{g^2(\mathbf{x})} \right) - \left(\int \pi(\mathbf{x}) \cdot h(\mathbf{x}) d\mathbf{x} \right)^2$$

Παρατηρούμε ότι ο δεύτερος όρος της παραπάνω ισότητας δεν εξαρτάται από την $g(\cdot)$. Συνεπώς για να ελαχιστοποιήσουμε την διακύμανση αρκεί να ελαχιστοποιήσουμε μόνο τον πρώτο όρο. Όμως ισχύει η ανισότητα

$$E_g\left(\frac{\pi^2(\mathbf{x}) \cdot h^2(\mathbf{x})}{g^2(\mathbf{x})} \right) \geq \left(E_g\left[\frac{\pi(\mathbf{x}) \cdot |h(\mathbf{x})|}{g(\mathbf{x})} \right] \right)^2 = \left(\int \pi(\mathbf{x}) \cdot |h(\mathbf{x})| d\mathbf{x} \right)^2 \quad (\text{ανισότητα } \textit{Jensen})$$

η οποία μας δίνει ένα κατώτατο όριο για την ποσότητα που θέλουμε να ελαχιστοποιήσουμε.

Είναι λοιπόν προφανές ότι το κατώτατο αυτό όριο μπορεί να επιτευχθεί μόνο αν επιλέξουμε

$$g(\mathbf{x}) = g^*(\mathbf{x}) = \frac{|h(\mathbf{x})| \cdot \pi(\mathbf{x})}{\int |h(\mathbf{z})| \cdot \pi(\mathbf{z}) d\mathbf{z}}.$$

Πρακτικά λοιπόν θα προσπαθούμε να επιλέγουμε κατανομές $g(\cdot)$ για τις οποίες η ποσότητα $\dfrac{|h(\cdot)|\,\pi(\cdot)}{g(\cdot)}$ θα είναι σχεδόν σταθερή και με πεπερασμένη διακύμανση. Δηλαδή πιο απλά καλό είναι να αποφεύγονται κατανομές $g(\cdot)$ για τις οποίες $\int \dfrac{\pi^2(\mathbf{x})}{g(\mathbf{x})} d\mathbf{x} = +\infty$, διότι τότε η μέθοδος Importance Sampling θα είναι πολύ χειρότερη από την κλασσική άμεση Monte Carlo προσέγγιση (π.χ. ο εκτιμητής έχει πολύ αργή σύγκλιση, το Κεντρικό Οριακό Θεώρημα χωρίς πεπερασμένη διακύμανση δεν ισχύει). Για περισσότερες λεπτομέρειες βλέπε *Robert* & *Casella* [54] §3.3.

Στην συνέχεια παραθέτουμε την αντιστοίχηση των συμβολισμών που χρησιμοποιήσαμε παραπάνω, σε σχέση με αυτούς που χρησιμοποιήσαμε στο κεφάλαιο 4.

$$h(\mathbf{x}) \Rightarrow I\left(D(\mathbf{y};\hat{\boldsymbol{\mu}}) \geq D_{obs}\right)\ ,\ \ \pi(\mathbf{x}) \Rightarrow f(\mathbf{y}|\mathbf{s}) = \frac{f(\mathbf{y};\boldsymbol{\mu})}{f(\mathbf{s};\boldsymbol{\mu})}\ ,$$

$$g(\mathbf{x}) \Rightarrow g(\mathbf{y}|\mathbf{s}) \ \text{ και } \ \mathbf{x}^{(t)} = \left\{x_i^{(t)}, i = 1,2,\ldots,k\right\} \Rightarrow \mathbf{y}_k^* = \left(y_1^*, y_2^*, \ldots, y_n^*\right)$$

A.2 Μαρκοβιανές αλυσίδες και η μέθοδος Gibbs Sampling

Έστω μια ακολουθία τυχαίων μεταβλητών $\{X_0, X_1, X_2, \ldots\}$, τέτοια ώστε σε κάθε στιγμή $n \geq 0$ η επόμενη κατάσταση X_{n+1} παράγεται από μια κατανομή $P(X_{n+1}|X_n)$ η οποία εξαρτάται μόνο από την κατάσταση X_n. Δηλαδή δοθέντος της X_n η επόμενη κατάσταση δεν εξαρτάται από το προηγούμενο παρελθόν $\{X_0, X_1, X_2, \ldots, X_{n-1}\}$. Μια τέτοια ακολουθία ονομάζεται Μαρκοβιανή αλυσίδα (*Markov Chain*).

Τις αλυσίδες αυτές μπορούμε να τις φανταστούμε σαν την ακολουθία των τυχαίων μεταβλητών να εξελίσεται όσο περνάει ο χρόνος, με την πιθανότητα μετάβασης στην επόμενη κατάσταση να εξαρτάται μόνο από την τωρινή κατάσταση που βρίσκεται η αλυσίδα. Η κατασκευή τέτοιων αλυσίδων στηρίζεται στον πυρήνα μετάβασης (*transition kernel*), δηλαδή σε μια δεσμευμένη πιθανότητα τέτοια ώστε $X_{n+1} \sim P(X_{n+1}|X_n) \equiv K(X_n, X_{n+1})$. Αν ο πυρήνας μετάβασης είναι ανεξάρτητος του n η αλυσίδα ονομάζεται ομογενής.

Το βασικό χαρακτηριστικό των Μαρκοβιανών αλυσίδων είναι ότι «ξεχνάνε» την αρχική τους κατάσταση X_0, και όσο μεγαλώνει το n τελικά συγκλίνουν σε μια μοναδική αναλλοίωτη κατανομή η οποία δεν εξαρτάται από το n ή το X_0. Η κατανομή αυτή είναι γνωστή ως στάσιμη (*stationary*) κατανομή της αλυσίδας και την συμβολίζουμε με $\pi(\cdot)$. Δηλαδή $\lim_{n\to\infty} P(X_n|X_0) = \pi(\cdot)$, το οποίο πρακτικά σημαίνει ότι αν $X_n \sim \pi(\cdot)$ τότε $X_{n+1} \sim \pi(\cdot)$.

Βέβαια για να συγκλίνει η κατανομή της X_{n+1} στην στάσιμη κατανομή $\pi(\cdot)$, η αλυσίδα θα πρέπει να ικανοποιεί τρείς βασικές ιδιότητες. Αρχικά θα πρέπει να είναι αμείωτη (*irreducible*), που σημαίνει ότι ο πυρήνας μετάβασης $K(\cdot,\cdot)$ πρέπει να επιτρέπει στην αλυσίδα να «φτάνει» σε όλα τα σημεία του χώρου καταστάσεων $\mathcal{X}$ (*state-space*) που έχουν θετική πιθανότητα. Δηλαδή για κάθε $A \in \mathcal{X}$ με $\pi(A) > 0$, να ισχύει $P(X_n \in A | X_0) > 0$.

Κατά δεύτερον η αλυσίδα χρειάζεται να είναι απεριοδική, δηλαδή να μην ταλαντεύεται ανάμεσα σε σημεία του χώρου κατάστασεων ακολουθώντας μια περιοδική συστηματική κίνηση, και τέλος η πιο σημαντική ιδιότητα που πρέπει να χαρακτηρίζει την αλυσίδα είναι η θετική επαναληπτικότητα (*positive recurrent*). Αυτή η ιδιότητα μας εξασφαλίζει ότι αν η αρχική κατάσταση $X_0 \sim \pi(\cdot)$, τότε και όλες οι επακόλουθες καταστάσεις θα προέρχονται από την κατανομή $\pi(\cdot)$.

Όταν μια Μαρκοβιανή αλυσίδα $\{X_0, X_1, X_2, \ldots\}$ πληρεί τις παραπάνω ιδιότητες και έχει στάσιμη κατανομή $\pi(\cdot)$, τότε ισχύει ο Ισχυρός Νόμος των Μεγάλων Αριθμών για Μαρκοβιανές αλυσίδες (Εργοδικό Θεώρημα) ο οποίος μας εξασφαλίζει την ασυμπτωτική σύγκλιση $\bar{h}_n := \frac{1}{n}\sum_{i=0}^{n-1} h(X_i) \to E_\pi[h(X)] = \int h(x)\cdot\pi(x)$ για οποιαδήποτε συνάρτηση $h(\cdot)$ μας ενδιαφέρει.

Αν λοιπόν επιθυμούμε να προσομοιώσουμε τιμές από μια κατανομή $f(\cdot)$, αρκεί να κατασκευάσουμε μια αλυσίδα Markov έτσι ώστε η στάσιμη κατανομή της να είναι η συγκεκριμένη κατανομή $f(\cdot)$ που μας ενδιαφέρει. Οι μέθοδοι κατασκευής εργοδικών αλυσίδων Markov με στόχο την εκμετάλλευση του Εργοδικού Θεωρήματος καθώς και της σύγκλισης στην στάσιμη κατανομή είναι γενικότερα γνωστοί ως Markov Chain Monte Carlo (*MCMC*) μέθοδοι.

Μια από αυτές τις μεθόδους είναι και η Gibbs Sampling μέθοδος η οποία βρίσκει ιδιαίτερη εφαρμογή όταν επιθυμούμε να προσομοιώσουμε τιμές από κάποια πολυδιάστατη κατανομή. Για την υλοποίηση της μεθόδου απαιτείται μια επιπλέον γνώση της ζητούμενης κατανομής ώστε να μπορούμε να αντλήσουμε από αυτήν τις πλήρεις δεσμευμένες πυκνότητες πιθανότητας.

Για παράδειγμα έστω η πολυδιάστατη κατανομή $\mathbf{X} = (X_1, X_2, \ldots, X_p) \sim f(x_1, x_2, \ldots, x_p)$, όπου τα X_i μπορεί να είναι μονοδιάστατα ή πολυδιάστατα και από την οποία επιθυμούμε να παράγουμε τιμές. Είναι όμως ευκολότερο να παράγουμε παρατηρήσεις από τις πλήρεις δεσμευμένες κατανομές (*full conditional distributions*) της $f(\cdot)$:

$$X_i \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p \sim f(x_i \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p) \text{ για } i = 1,2,\ldots,p.$$

Η μέθοδος Gibbs Sampling για την μετάβαση της αλυσίδας από την κατάσταση $\mathbf{X}^{(n)}$ στην κατάσταση $\mathbf{X}^{(n+1)}$ βασίζεται στην παρακάτω ανακύκλωση:

Δοθείσης της $\mathbf{X}^{(n)} = \left(X_1, X_2, \ldots, X_p\right) = \left(x_1^{(n)}, x_2^{(n)}, \ldots, x_p^{(n)}\right),\ n \geq 0$, παράγουμε διαδοχικά τις τιμές :

$$X_1^{(n+1)} \sim f\left(x_1 \left| X_2 = x_2^{(n)}, X_3 = x_3^{(n)}, \ldots, X_p = x_p^{(n)}\right.\right)$$

$$X_2^{(n+1)} \sim f\left(x_2 \left| X_1 = x_1^{(n+1)}, X_3 = x_3^{(n)}, \ldots, X_p = x_p^{(n)}\right.\right)$$

$$X_3^{(n+1)} \sim f\left(x_3 \left| X_1 = x_1^{(n+1)}, X_2 = x_2^{(n+1)}, X_4 = x_4^{(n)}, \ldots, X_p = x_p^{(n)}\right.\right)$$

$$\vdots$$

$$X_p^{(n+1)} \sim f\left(x_p \left| X_1 = x_1^{(n+1)}, X_2 = x_2^{(n+1)}, \ldots, X_{p-1} = x_{p-1}^{(n+1)}\right.\right)$$

Παραγάγαμε λοιπόν μια τιμή $\mathbf{X}^{(n+1)} = \left(x_1^{(n+1)}, x_2^{(n+1)}, \ldots, x_p^{(n+1)}\right)$ από την πολυδιάστατη κατανομή $f\left(x_1, x_2, \ldots, x_p\right)$.

Τα διανύσματα $\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)}, \mathbf{X}^{(p+1)}, \ldots$ που τελικά παράγουμε με συνεχή χρήση του Gibbs sampler, αποτελούν πραγμάτωση μια Μαρκοβιανής αλυσίδας, με πυρήνα μετάβασης από την κατάσταση $\mathbf{X}^{(i)}$ στην κατάσταση $\mathbf{X}^{(i+1)}$ το γινόμενο :

$$K\left(\mathbf{X}^{(i)}, \mathbf{X}^{(i+1)}\right) = \prod_{m=1}^{p} f\left(x_m^{(i+1)} \left| \left\{x_n^{(i+1)}, n < m\right\}, \left\{x_n^{(i)}, n > m\right\}\right.\right)$$

Το κυριότερο πλεονέκτημα της μεθόδου είναι ότι ακόμα και σε μεγάλα πολυδιάστατα προβλήματα όλες οι προσομοιώσεις μας μπορεί να προκύψουν από μονοδιάστατες κατανομές (π.χ. αν οι full conditional κατανομές είναι μονοδιάστατες), γεγονός που κάνει την υλοποίηση του αλγορίθμου ευκολότερη.

Βέβαια το γεγονός ότι οι μόνες πυκνότητες που χρησιμοποιούμε είναι οι πλήρεις δεσμευμένες πυκνότητες της $f(\cdot)$, προυποθέτει καλή γνώση των πιθανοτικών ιδιοτήτων της $f(\cdot)$, και έτσι η χρήση του Gibbs Sampling είναι κάπως περιοριστική ως προς την επιλογή της συμβάλλουσας κατανομής.

Για περισσότερες τεχνικές λεπτομέρειες σχετικά με τις Μαρκοβιανές αλυσίδες και τις MCMC μεθόδους βλ. *Gilks et.al.* [27], *Robert & Casella* [54].

Τέλος οι προσομοιωμένες τιμές του **X** δεν είναι ανεξάρτητες μεταξύ τους και αυτό πρέπει να ληφθεί υπόψην όταν υπολογίζουμε για παράδειγμα την ακρίβεια της εκτιμώμενης p-value για κάποιον έλεγχο υπόθεσης. Οι *Raftery & Lewis* [53] περιέγραψαν μια μέθοδο υπολογισμού του πλήθους των επαναλήψεων που απαιτούνται ώστε να επιτύχουμε ένα επιθυμητό επίπεδο ακριβείας με την μέθοδο Gibbs Sampling.

Παράρτημα Β : Απόδειξη της σχέσης (4.8)

(εφαρμογή Δέλτα μεθόδου)

Υστερα απο N επαναλήψεις του αλγορίθμου έχουμε παραγάγει τα εξής σετ τιμών

$$\left(u_1^*, w_1^*\right), \left(u_2^*, w_2^*\right), \ldots, \left(u_N^*, w_N^*\right)$$

Αν ορίσουμε ως g την συνάρτηση g(u,w)=$\frac{u}{w}$ τότε $\widetilde{P}$-value = $\frac{\sum_{k=1}^{N} u_k^*}{\sum_{k=1}^{N} w_k^*} = \frac{\overline{u}}{\overline{w}}$=E(g).

Επίσης οι μερικές παράγωγοι της g(u,w) είναι $\nabla g(u,w) = \left(\frac{1}{w} \quad -\frac{u}{w^2}\right)$.

Με χρήση της μεθόδου Δέλτα βρίσκουμε προσεγγιστικά :

(οι όροι μεγαλύτερης τάξης έχουν παραληφθεί, για αυτό και το αποτέλεσμα θα είναι προσεγγιστικό)

$$\sigma^2_{\widetilde{P}-value} \approx \left(\frac{1}{\overline{w}} \quad -\frac{\overline{u}}{\overline{w}^2}\right) \cdot \begin{pmatrix} \mathrm{var}(u) & \mathrm{cov}(u,w) \\ \mathrm{cov}(u,w) & \mathrm{var}(w) \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{\overline{w}} \\ -\frac{\overline{u}}{\overline{w}^2} \end{pmatrix} =$$

$$= \frac{1}{\overline{w}^2} \cdot \mathrm{var}(u) + \frac{\overline{u}^2}{\overline{w}^4} \cdot \mathrm{var}(w) - 2\frac{\overline{u}}{\overline{w}^2} \cdot \frac{1}{\overline{w}} \cdot \mathrm{cov}(u,w) = \ldots$$

και στην συνέχεια μετά απο αλγεβρικές πράξεις καταλήγουμε οτι

$$\widetilde{\sigma}_{\widetilde{P}-value} = \frac{1}{\overline{w}} \cdot \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(u_k - w_k \cdot \widetilde{P}\right)^2}$$

Παράρτημα Γ : Απόδειξη της σχέσης (5.1)

(Ακριβής δεσμευμένη κατανομή γενικευμένων γραμμικών μοντέλων με κανονικό σύνδεσμο)

Έστω $\mathbf{y}=(y_1, y_2, \ldots, y_n)'$ ανεξάρτητες τ.μεταβλητές, με κάθε y_i να προέρχεται από εκθετική οικογένεια κατανομών με παραμέτρους (θ_i, ϕ_i). Η από κοινού συνάρτηση πυκνότητας πιθανότητας (σ.π.π) των y_i, i=1,2,…,n είναι

$$f(\mathbf{y};\boldsymbol{\theta},\boldsymbol{\varphi})=\exp\left[\sum_{i=1}^{n}\left\{\frac{y_i\cdot\theta_i-c(\theta_i)}{\phi_i}+h(y_i,\phi_i)\right\}\right] \qquad (1)$$

όπου $\mathbf{y},\boldsymbol{\theta},\boldsymbol{\varphi}$ είναι διανύσματα διάστασης nx1. Επειδή έχουμε κανονικό σύνδεσμο προκύπτει $\theta_i=\theta_i(\boldsymbol{\beta})=\mathbf{x}_i'\cdot\boldsymbol{\beta}=\sum_{j=1}^{p}x_{ij}\cdot\beta_j$, i=1,2,…,n , όπου $\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_p)'$ είναι το p-διάνυσμα των αγνώστων παραμέτρων, με αντίστοιχο διάνυσμα συμμεταβλητών για κάθε μία από αυτές, το $\mathbf{x}_i'=(x_{i1},x_{i2},\ldots,x_{ip})$.

Η από κοινού σ.π.π. (1) γίνεται

$$f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\varphi})=\exp\left[\sum_{j=1}^{p}\beta_j\cdot\sum_{i=1}^{n}\frac{y_i\cdot x_{ij}}{\phi_i}-\sum_{i=1}^{n}\frac{c\{\theta_i(\beta)\}}{\phi_i}+\sum_{i=1}^{n}h(y_i,\phi_i)\right] \qquad (2)$$

Αν το $\boldsymbol{\varphi}$ είναι γνωστό τότε $\left\{z_j=\sum_{i=1}^{n}\frac{y_i\cdot x_{ij}}{\phi_i}, j=1,\ldots,p\right\}$ είναι το σύνολο των επαρκών στατιστικών για τις παραμέτρους $\{\beta_j, j=1,\ldots,p\}$. Συνεπώς $\mathbf{z}=\mathbf{X}'\cdot\mathbf{W}\cdot\mathbf{y}$ όπου $\mathbf{X}_{nxp}=(x_{ij})$ είναι ο πίνακας σχεδιασμού και $\mathbf{W}_{nxn}=\begin{pmatrix}1/\phi_1 & 0 & \cdots & 0\\ 0 & 1/\phi_2 & \cdots & 0\\ \vdots & \vdots & \cdots & \vdots\\ 0 & 0 & \cdots & 1/\phi_n\end{pmatrix}$. Ο πίνακας σχεδιασμού είναι αντιστρέψιμος, οπότε $\mathbf{y}(\mathbf{z})=\mathbf{W}^{-1}\cdot(\mathbf{X}')^{-1}\cdot\mathbf{z}$.

Η συνάρτηση πυκνότητας των επαρκών στατιστικών θα είναι τότε :

$$f(\mathbf{z};\boldsymbol{\beta},\boldsymbol{\varphi})\propto\exp\left[\sum_{j=1}^{p}z_j\cdot\beta_j+\sum_{i=1}^{n}h(y_i(\mathbf{z}),\phi_i)\right]$$

αφού η ποσότητα $\exp\left(-\sum_{i=1}^{n}\frac{c\{\theta_i(\beta)\}}{\phi_i}\right)$ δεν εξαρτάται από το **z.**

Έστω ότι χωρίζουμε το διάνυσμα των παραμέτρων **β** σε δύο υποδιανύσματα $\boldsymbol{\beta}_R$ και $\boldsymbol{\beta}_{\setminus R}$, όπου το πρώτο περιλαμβάνει τις *r* παραμέτρους που μας ενδιαφέρουν, ενώ το δεύτερο περιλαμβάνει τις οχληρές (*nuisance*) παραμέτρους.

Η *r*-διάστατη δεσμευμένη κατανομή των επαρκών στατιστικών του $\boldsymbol{\beta}_R$, δοθέντος των επαρκών στατιστικών του $\boldsymbol{\beta}_{\setminus R}$ είναι

$$f\left(\mathbf{z}_R \mid \mathbf{z}_{\setminus R};\boldsymbol{\beta}_R,\boldsymbol{\varphi}\right) \propto \exp\left[\sum_{j\in R} z_j \cdot \beta_j + \sum_{i=1}^{n} h\left(y_i(\mathbf{z}),\phi_i\right)\right] \tag{3}$$

αφού η ποσότητα $\exp\left[\sum_{j\in\setminus R} z_j \cdot \beta_j\right]$ δεν εξαρτάται από το $\mathbf{z}_R$.

Όταν $\boldsymbol{\beta}_R = 0$ η (3) γίνεται :

$$f\left(\mathbf{z}_R \mid \mathbf{z}_{\setminus R};\boldsymbol{\varphi}\right) \propto \exp\left[\sum_{i=1}^{n} h\left(y_i(\mathbf{z}),\phi_i\right)\right] = \prod_{i=1}^{n}\left[\mathbf{H}\left(\phi_i \cdot \sum_{j=1}^{p} x^{ji} \cdot z_j,\phi_i\right)\right] \tag{4}$$

όπου $\mathbf{H}(\cdot,\cdot) = \exp[h(\cdot,\cdot)]$ και $\left(\mathbf{X}'\right)^{-1} = \left(x^{ji}\right)$.

Συγκεκριμένα τώρα στην περίπτωση του κορεσμένου log-linear Poisson μοντέλου που μας ενδιαφέρει, έχουμε ότι

- $\phi_i = 1$
- $\theta_i = \log(\mu_i) = \sum_{j=1}^{p} x_{ij} \cdot \beta_j$ και
- $h(y_i,\phi_i) = -\log y_i! \;\Rightarrow\; \mathbf{H}(y_i,\phi_i) = \exp(-\log y_i!) = \frac{1}{y_i!}$

Συνεπώς η δεσμευμένη κατανομή για τον έλεγχο της υπόθεσης H_0 : $\boldsymbol{\beta}_R = 0$ έναντι της εναλλακτικής H_1 : $\boldsymbol{\beta}_R \neq 0$ που αντιστοιχεί στο κορεσμένο μοντέλο, από την σχέση (4) θα είναι

$$f\left(\mathbf{z}_R \mid \mathbf{z}_{\setminus R}\right) \propto \left\{\prod_{i=1}^{n}\left(\sum_{j=1}^{p} x^{ji} \cdot z_j\right)!\right\}^{-1} .$$

Παράρτημα Δ : Οι αλγόριθμοι

Δ.1 Ο αλγόριθμος του Importance Sampling

Παραθέτουμε την πλήρη υλοποίηση του αλγορίθμου σε γλώσσα προγραμματισμού Mathematica. Είναι ένας αλγόριθμος που μπορεί να εφαρμοστεί για τον έλεγχο ανεξαρτησίας σε οποιονδήποτε διδιάστατο *I*x*J* πίνακα συνάφειας.

Το μόνο που πρέπει να εισάγουμε κάθε φορά, είναι οι συχνότητες $\{y_{ij}\}$ απο τον παρατηρούμενο πίνακα συνάφειας. Στο παράδειγμα της §4.5.1 έχουμε

```
Ydata := {{7, 7, 2, 3}, {2, 8, 3, 7}, {1, 5, 4, 9}, {2, 8, 9, 14}}
```

Ο αλγόριθμος λοιπόν είναι :

- Mathematica Libraries

```
<< LinearAlgebra`MatrixManipulation`
<< Statistics`ContinuousDistributions`
```

- Standard normal CDF

```
Φ[x_] := (1 + Erf[x / Sqrt[2]]) / 2 // N
```

- Poisson density

```
PoiLogDens[x_, λ_] := If[x + λ == 0, 0, -λ + x Log[λ]];
```

- OneColumn[n] is the n\times 1 column of ones

```
OneColumn[n_] := Table[1, {i, 1, n}, {j, 1, 1}];
```

- ZeroRow[n] is the 1\times n row of zeros

```
ZeroRow[n_] := Table[0, {i, 1, 1}, {j, 1, n}];
```

- Rmatrix[n,m,k] is a matrix with ones in the k-th column and zeros elsewhere

```
Rmatrix[n_, m_, k_] := If[And[k ≥ 1, k ≤ m], Table[If[j == k, 1, 0], {i, 1, n}, {j, 1, m}],
   Print["R Error"]];
```

- Testing independence in contingency tables :

NR = Number of rows, NC = Number of columns

```
X1matrix[NR_, NC_] :=
  Last[{
    X1local = AppendRows[OneColumn[NC - 1], Rmatrix[NC - 1, NR - 1, 1],
      IdentityMatrix[NC - 1]];
    For[i = 2, i ≤ NR - 1, i++,
     {X1localnew = AppendRows[OneColumn[NC - 1], Rmatrix[NC - 1, NR - 1, i],
        IdentityMatrix[NC - 1]];
      X1local = AppendColumns[X1local, X1localnew];}];
    X1local
   }];
X2matrix[NR_, NC_] :=
  Last[{
    X2local1 = AppendRows[OneColumn[NR + NC - 2], IdentityMatrix[NR + NC - 2]];
    X2local2 = AppendRows[{{1}}, ZeroRow[NR + NC - 2]];
    X2local = AppendColumns[X2local1, X2local2];
    X2local
   }];
```

- The data and the MLE

```
Ydata := {{7, 7, 2, 3}, {2, 8, 3, 7}, {1, 5, 4, 9}, {2, 8, 9, 14}};
NC = Length[Ydata[[1]]]; NR = Length[Ydata];
Y = {};
Do[For[j = 1, j ≤ NC - 1, j++, Y = Append[Y, Ydata[[i]][[j]]]], {i, 1, NR - 1}];
For[i = 1, i ≤ NR - 1, i++, Y = Append[Y, Ydata[[i]][[NC]]]];
For[j = 1, j ≤ NC, j++, Y = Append[Y, Ydata[[NR]][[j]]]];
Print["Y = ", Y];
sk1 = {}; sk2 = {};
Do[skr = Sum[Ydata[[i]][[j]], {j, 1, NC}]; sk1 = Append[sk1, skr], {i, 1, NR}];
Do[skc = Sum[Ydata[[i]][[j]], {i, 1, NR}]; sk2 = Append[sk2, skc], {j, 1, NC}];
Yft = Flatten[Ydata];
S = Sum[Yft[[i]], {i, 1, Length[Yft]}];
mm1 = {};
Do[
  Do[mm = sk1[[i]] * sk2[[j]] / S // N; mm1 = Append[mm1, mm], {j, 1, NC}], {i, 1, NR}];
mm2 = Partition[mm1, NC];
Mhat = {};
Do[For[j = 1, j ≤ NC - 1, j++, Mhat = Append[Mhat, mm2[[i]][[j]]]], {i, 1, NR - 1}];
For[i = 1, i ≤ NR - 1, i++, Mhat = Append[Mhat, mm2[[i]][[NC]]]];
For[j = 1, j ≤ NC, j++, Mhat = Append[Mhat, mm2[[NR]][[j]]]];
Print["the mle = ", Mhat]
```

```
Y = {7, 7, 2, 2, 8, 3, 1, 5, 4, 3, 7, 9, 2, 8, 9, 14}
the mle = {2.50549, 5.84615, 3.75824, 2.63736, 6.15385, 3.95604, 2.50549,
  5.84615, 3.75824, 6.89011, 7.25275, 6.89011, 4.35165, 10.1538, 6.52747, 11.967}
```

```
p = NR + NC - 1; n = NR * NC;
```

- Design matrix

```
X1 = X1matrix[NR, NC];
X2 = X2matrix[NR, NC];
X = AppendColumns[X1, X2];
```

- Sufficient statistics

```
s = Transpose[X].Y;
```

- Observed Deviance

```
DevObs = 2 Sum[PoiLogDens[Y[[i]], Y[[i]]] - PoiLogDens[Y[[i]], Mhat[[i]]], {i, 1, n}] // N;
```

```
V = DiagonalMatrix[Mhat];
EY1 = AppendRows[IdentityMatrix[n - p], ZeroMatrix[n - p, p]].Mhat;
V11 = TakeRows[TakeColumns[V, n - p], n - p];
V11givens = V11 - V11.X1.Inverse[Transpose[X].V.X].Transpose[X1].V11;
```

The importance sampling scheme

1. *Conditional variances for univariate normal simulations*

```
CovMatrixForSimulation :=
  {
   Clear[sigmatable, sigma, vstartable, vstar];
   sigmatable = Table[sigma[i], {i, 1, n - p}];
   vstartable = Table[vstar[i, j], {i, 1, n - p}, {j, 1, n - p}];
   sigma[1] = V11givens[[1]][[1]]; For[k = 1, k ≤ n - p, k++, vstar[1, k] = V11givens[[1]][[k]]];
   For[k = 2, k ≤ n - p, k++,
    {
     sigma[k] = V11givens[[k]][[k]] - Sum[vstar[i, k]^2 / sigma[i], {i, 1, k - 1}];
     For[i = 2, i ≤ k, i++,
      vstar[i, k] = V11givens[[i]][[k]] - Sum[vstar[j, i] V11givens[[j]][[k]] / sigma[j], {j, 1, i - 1}];];
    }
   ];
  };
```

2. The simulation algorithm

```
PValueImpSamplingSimulation[Iterations_] :=
{
 count = 0;
 Swstar = 0; SwstarSQ = 0; Swstartimesind = 0; SwstartimesindSQ = 0; SwstartimesindCP = 0;
 CovMatrixForSimulation; Clear[ystartable, ystar]; ystartable = Table[ystar[i],
   {i, 1, n - p}];
 m[1] = EY1[[1]];
 For[iter = 1, iter ≤ Iterations, iter++,
  {
   neg1 = 0; neg2 = 0;
```

♦ Simulation

$$z = \text{Random}\left[\text{NormalDistribution}\left[m[1], \sqrt{\text{sigma}[1]}\right]\right]; \text{ystar}[1] = \text{Round}[z]; x[1] = \text{ystar}[1] - m[1];$$

$$\text{For}\left[k = 2, k \le n - p, k{+}{+},\right.$$

$$m[k] = \text{EY1}[[k]] + \text{Sum}\left[\frac{x[i]\,\text{vstar}[i, k]}{\text{sigma}[i]}, \{i, 1, k - 1\}\right];$$

$$\left.z = \text{Random}\left[\text{NormalDistribution}\left[m[k], \sqrt{\text{sigma}[k]}\right]\right]; \text{ystar}[k] = \text{Round}[z]; x[k] = \text{ystar}[k] - m[k];\right];$$

♦ Checks if negative ystars have been generated

```
For[i = 1, i ≤ n - p, i++, If[ystar[i] < 0, neg1 = 1]];
```

♦ Solves the linear system

```
If[neg1 == 0,
  {
   ystar2 = LinearSolve[Transpose[X2], s - Transpose[X1].ystartable];
   For[i = 1, i ≤ p, i++, If[ystar2[[i]] < 0, neg2 = 1]];
  }
 ];
```

♦ Updates provided simulated values have been accepted

$$\text{If}\left[\text{neg1} + \text{neg2} == 0,\right.$$

```
 {
  logf = Sum[PoiLogDens[ystar[i], Mhat[[i]]] - LogGamma[ystar[i] + 1], {i, 1, n - p}]
    + Sum[PoiLogDens[ystar2[[i]],
       Mhat[[n - p + i]]] - LogGamma[ystar2[[i]] + 1], {i, 1, p}];
```

$$\text{logg} = \text{Sum}\left[\text{Log}\left[\Phi\left[\frac{\text{ystar}[i] + 0.5 - m[i]}{\sqrt{\text{sigma}[i]}}\right] - \Phi\left[\frac{\text{ystar}[i] - 0.5 - m[i]}{\sqrt{\text{sigma}[i]}}\right]\right], \{i, 1, n - p\}\right];$$

```
      wstar = Exp[logf - logg];
      Swstar = Swstar + wstar;
      SwstarSQ = SwstarSQ + wstar^2;
      DevSim =
       2 (Sum[PoiLogDens[ystar[i], ystar[i]] - PoiLogDens[ystar[i], Mhat[[i]]],
           {i, 1, n - p}] +
         Sum[PoiLogDens[ystar2[[i]], ystar2[[i]]] - PoiLogDens[ystar2[[i]],
           Mhat[[n - p + i]]], {i, 1, p}]);
      If[DevSim ≥ DevObs, wstartimesind = wstar, wstartimesind = 0];
      Swstartimesind = Swstartimesind + wstartimesind;
      SwstartimesindSQ = SwstartimesindSQ + wstartimesind^2;
      SwstartimesindCP = SwstartimesindCP + wstar * wstartimesind;
     },
     {count = count + 1;}
    ];
  }];
```

♦ Results

```
  EstPVvalue = Swstartimesind / Swstar;
  StDevEstPValue =
   √(SwstartimesindSQ + EstPValue^2 SwstarSQ - 2 EstPValue SwstartimesindCP) / Swstar // N;
  CVMeanwstar = SwstarSQ / Swstar^2 - 1 / Iterations // N;

  Print["Estimated p value = ", EstPValue];
  Print["Estimated standard deviation = ", StDevEstPValue];
  Print["99% confidence interval: [", EstPValue - 2.576 StDevEstPValue, ",",
   EstPValue + 2.576 StDevEstPValue, "]"];
  Print["wstar mean coeff. of variation = ", CVMeanwstar];
  Print["Rejected samples (some ystar<0) = ", 100 count / Iterations // N, "%"];

 };

PValueImpSamplingSimulation[20000];
```

```
Estimated p value = 0.109095
Estimated standard deviation = 0.00247283
99% condidence interval: [0.102725,0.115465]
wstar mean coeff. of variation = 7.40148×10^-6
Rejected samples (some ystar<0) = 4.275%
```

Δ.2 Ο αλγόριθμος του Gibbs Sampling

Παραθέτουμε την πλήρη υλοποίηση του αλγορίθμου σε γλώσσα προγραμματισμού Fortran. Ο αλγόριθμος μπορεί να εφαρμοστεί σε 2x2x2 πίνακα συνάφειας για τον έλεγχο καλής προσαρμογής οποιουδήποτε ιεραρχικού λογαριθμογραμμικού μοντέλου, έναντι του κορεσμένου μοντέλου.

Προσοχή χρειάζεται κάθε φορά στην εισαγωγή των δεδομένων μας, στον υπολογισμό των στηριγμάτων (*supports*) των επαρκών στατιστικών $\mathbf{z}_R$ των παραμέτρων που μας ενδιαφέρει ο έλεγχος της σημαντικότητάς τους, και τέλος στην τροποποίηση της υπορουτίνας `EXPECT1`, ώστε οι εκτιμήσεις των αναμενόμενων συχνοτήτων να υπολογίζονται συναρτήσει του $\mathbf{z}_{\setminus R}$.

```
    USE NUMERICAL_LIBRARIES
    IMPLICIT NONE

      INTEGER                                                          &
                      N,K,I,J,LL,ITERATIONS,IFL,SIFL
      REAL(8)                                                          &
        PVAL,SOBS,STS,FF,FFD,U,COUNT,MEANSTS,MEANSTSSQ,MINSTS,MAXSTS
      INTEGER,ALLOCATABLE::                                            &
            Z(:),A(:,:),ZMIN(:),ZMAX(:),Y(:),                          &
            XT(:,:),IND(:,:),AA(:),Z1(:),Z2(:),Z3(:),                  &
            Z4(:),Z5(:),Z6(:),Z7(:),Z8(:)
        REAL(8),ALLOCATABLE::                                          &
                STSVALUES(:),ZR(:),YR(:),EY(:)


!   ********************************************************************
!     RESULTS FILE
!   ********************************************************************
      OPEN(9, FILE='HIST.TXT')

      K = 8

      ITERATIONS = 100000

      ALLOCATE    (                                                    &

            Z(1:K),A(1:K-1,1:K),ZMIN(1:K),ZMAX(1:K-1),Y(1:K),EY(1:K),  &
                  XT(1:K,1:K),IND(1:K,1:K),AA(1:K),ZR(1:K),YR(1:K),    &
                  STSVALUES(ITERATIONS),                               &
                  Z1(ITERATIONS),Z2(ITERATIONS),Z3(ITERATIONS),        &
                  Z4(ITERATIONS),Z5(ITERATIONS),Z6(ITERATIONS),        &
                  Z7(ITERATIONS),Z8(ITERATIONS))
```

```
!   ************************************************************************
!     K= number of non-redundant parameters
!     XT  IS THE TRANSPOSE OF THE DESIGN MATRIX of the saturated model
!   ************************************************************************

      DO I=1, K
            XT(I,1)=1
            DO J=2, K
                  IF (I == J) THEN
                        XT(I,J)=1
                  ELSE
                        XT(I,J)=0
                  END IF
            END DO
      END DO
      XT(4,2)=1
      XT(6,2)=1
      XT(8,2)=1
      XT(4,3)=1
      XT(7,3)=1
      XT(8,3)=1
      XT(8,4)=1
      XT(6,5)=1
      XT(7,5)=1
      XT(8,5)=1
      XT(8,6)=1
      XT(8,7)=1

!   ************************************************************************
!       Y ARE THE DATA IN VECTOR FORM:
!
!       Y(1)=Y111, Y(2)=Y121, Y(3)=Y112, Y(4)=Y122,
!       Y(5)=Y211, Y(6)=Y221, Y(7)=Y212, Y(8)=Y222
!
!   ************************************************************************

      Y(1)=53
      Y(2)=414
      Y(3)=0
      Y(4)=16
      Y(5)=11
      Y(6)=37
      Y(7)=4
      Y(8)=139

!   ************************************************************************
!      INITIAL  VALUES  ASSIGNMENT to the sufficient statistics
!      Z(1),...,Z(8)
!   ************************************************************************

      STS=0.0d0
      SOBS=0.0d0
      IFL=0
      SIFL=0

      Z = MATMUL(XT,Y)

      DO I=1,K
            ZR(I)=Z(I)
```

```
            YR(I)=Y(I)
      END DO
      CALL EXPECT1(K,ZR,EY,IFL)
      SIFL=SIFL+IFL


!     ***************************************************************************
!       Find the observed X^2 Pearson (SOBS)
!     ***************************************************************************

      CALL PEARS(K,YR,EY,SOBS)

!     ***************************************************************************
!        INDEX  FOR TERMS OF THE FULL LIKELIHOOD PRESENT IN EACH CONDITIONAL
!     ***************************************************************************

      DO J=1, 8
            IND(1,J)=1
      END DO
      DO I=2,8
            DO J=1,7
                  IND(I,J)=0
            END DO
      END DO
      DO I=2, 7
            IND(I,8)=1
      END DO
      DO I=2, 4
            IND(I,2)=1
      END DO
      IND(2,3)=1
      IND(2,5)=1
      IND(3,4)=1
      IND(3,6)=1
      IND(5,3)=1
      IND(5,4)=1
      IND(5,7)=1
      IND(6,3)=1
      IND(7,4)=1


      COUNT = 0.0d0
      MEANSTS = 0.0d0
      MEANSTSSQ =  0.0d0
      MINSTS = 10000.0d0
      MAXSTS = 0.0d0

!     ***************************************************************************
!       FIND THE SUPPORTS  for Z(1) and Z(3)
!     ***************************************************************************

      DO LL = 1, ITERATIONS
            CALL COMPA1(Z,AA)
            ZMIN(1)=MAX(AA(1),AA(2),AA(3),AA(4))
            ZMAX(1)=MIN(AA(5),AA(6),AA(7),AA(8))
```

```
!    ************************************************************************
!      AA(I), (I=1,..,8) ARE THE TERMS OF THE LIKELIHOOD, FOR WHICH WE
!      COMPUTE ()!
!    ************************************************************************

              CALL COMPA2(Z(1),AA)
              CALL SIM1T7(1,ZMIN(1),ZMAX(1),Z(1),IND,AA)
              Z1(LL)=Z(1)

              A(2,1)=Z(1)

!             A(2,2)=Z(1)-Z(3)+Z(4)-Z(5)+Z(6)+Z(7)-Z(8)
!             A(2,3)=Z(1)-Z(3)+Z(4)
!             A(2,4)=Z(1)-Z(5)+Z(6)

!             ZMIN(2)=MAX(A(2,1),A(2,2))
!             ZMAX(2)=MIN(A(2,3),A(2,4))

!             CALL COMPA3(Z, AA)
!             CALL SIM1T7(2,ZMIN(2),ZMAX(2),Z(2),IND,AA)

              A(3,1)=Z(1)-Z(2)+Z(4)-Z(5)+Z(6)+Z(7)-Z(8)
              A(3,2)=Z(1)-Z(2)+Z(4)
              A(3,3)=Z(1)-Z(5)+Z(7)

              ZMIN(3)=MAX(A(2,1),A(3,1))
              ZMAX(3)=MIN(A(3,2),A(3,3))

              CALL COMPA3(Z, AA)
              CALL SIM1T7(3,ZMIN(3),ZMAX(3),Z(3),IND,AA)

!             ZMIN(4)=Z(3)+Z(2)-Z(1)
!             ZMAX(4)=Z(8)-Z(7)-Z(6)+Z(5)+Z(3)+Z(2)-Z(1)

!             CALL COMPA3(Z, AA)
!             CALL SIM1T7(4,ZMIN(4),ZMAX(4),Z(4),IND,AA)

!             A(4,1)=Z(1)-Z(2)-Z(3)+Z(4)+Z(6)+Z(7)-Z(8)
!             A(4,2)=Z(6)+Z(1)-Z(2)
!             A(4,3)=Z(7)+Z(1)-Z(3)

!             ZMIN(5)=MAX(A(2,1),A(4,1))
!             ZMAX(5)=MIN(A(4,2),A(4,3))

!             CALL COMPA3(Z, AA)
!             CALL SIM1T7(5,ZMIN(5),ZMAX(5),Z(5),IND,AA)

!             ZMIN(6)=Z(5)+Z(2)-Z(1)
!             ZMAX(6)=-Z(1)+Z(2)+Z(3)-Z(4)+Z(5)-Z(7)+Z(8)

!             CALL COMPA3(Z, AA)
!             CALL SIM1T7(6,ZMIN(6),ZMAX(6),Z(6),IND,AA)

!             ZMIN(7)=Z(5)+Z(3)-Z(1)
!             ZMAX(7)=-Z(1)+Z(2)+Z(3)-Z(4)+Z(5)-Z(6)+Z(8)

!             CALL COMPA3(Z, AA)
!             CALL SIM1T7(7,ZMIN(7),ZMAX(7),Z(7),IND,AA)
```

```
!   ************************************************************************
!     SIMULATE Z(8)
!   ************************************************************************

!           ZMIN(8)=Z(1)-Z(2)-Z(3)+Z(4)-Z(5)+Z(6)+Z(7)
!           N=0
!           FF=DEXP(-1.0d0)
!           FFD=FF
!           CALL DRNUN(1,U)
!           DO WHILE (U > FFD)
!                 N=N+1
!                 FF=FF/N
!                 FFD=FFD+FF
!           END DO
!           Z(8)=N+ZMIN(8)


!   ************************************************************************
!     WHAT FOLLOWS IS COMPUTED AT THE END OF EACH STEP
!   ************************************************************************
            DO I=1,K
                  ZR(I)=Z(I)
            END DO
            CALL EXPECT2(K,ZR,YR,IFL)
            SIFL=SIFL+IFL
            CALL PEARS(K,YR,EY,STS)
            STSVALUES(LL) = STS
            MINSTS = MIN(MINSTS,STS)
            MAXSTS = MAX(MAXSTS,STS)
            MEANSTS = MEANSTS + (STS-MEANSTS)/LL
            MEANSTSSQ = MEANSTSSQ + (STS**2-MEANSTSSQ)/LL
            IF (STS >= SOBS) THEN
                  COUNT = COUNT + 1
            END IF

      END DO


!   ************************************************************************
!     Print results
!   ************************************************************************

      PVAL = COUNT/ITERATIONS

      WRITE(9,10) PVAL,2.576*DSQRT(PVAL*(1-PVAL)/ITERATIONS)
      WRITE(9,11) ITERATIONS
      WRITE(9,13) SOBS

      WRITE(*,10) PVAL,2.576*DSQRT(PVAL*(1-PVAL)/ITERATIONS)
      WRITE(*,11) ITERATIONS
      WRITE(*,13) SOBS
      WRITE(*,*) 'TABLES WITH EY(7)<0',IFL

10    FORMAT ('P-value = ',F12.6,'  +/-',F12.6)
11    FORMAT ('NUMBER OF ITERATIONS = ',I10)
13    FORMAT ('SOBS = ',F12.6)


END PROGRAM
```

```
!**************************************************************************
      SUBROUTINE COMPA1(Z,AA)

            IMPLICIT NONE

            INTEGER Z(8), AA(8)

            AA(1)=0
            AA(2)=Z(2)+Z(3)-Z(4)
            AA(3)=Z(2)+Z(5)-Z(6)
            AA(4)=Z(3)+Z(5)-Z(7)
            AA(5)=Z(2)
            AA(6)=Z(3)
            AA(7)=Z(5)
            AA(8)=Z(2)+Z(3)-Z(4)+Z(5)-Z(6)-Z(7)+Z(8)

      END SUBROUTINE

!**************************************************************************
      SUBROUTINE COMPA2(Z, AA)

            IMPLICIT NONE

            INTEGER  I, Z, AA(8)

            DO I = 1, 4
            AA(I)=Z-AA(I)
            AA(I+4)=AA(I+4)-Z
            END DO

      END SUBROUTINE

!**************************************************************************
      SUBROUTINE COMPA3(Z, AA)

            IMPLICIT NONE

            INTEGER  Z(8), AA(8)

            AA(1)=Z(1)
            AA(2)=Z(1)-(Z(2)+Z(3)-Z(4))
            AA(3)=Z(1)-(Z(2)+Z(5)-Z(6))
            AA(4)=Z(1)-(Z(3)+Z(5)-Z(7))
            AA(5)=Z(2)-Z(1)
            AA(6)=Z(3)-Z(1)
            AA(7)=Z(5)-Z(1)
            AA(8)=Z(2)+Z(3)-Z(4)+Z(5)-Z(6)-Z(7)+Z(8)-Z(1)

      END SUBROUTINE

!**************************************************************************
!     THE SUBROUTINE SIM1T7 (T=1,2,…,7), DETERMINES WHICH SUFFICIENT
!     STATISTIC Z(I) WE SIMULATE
!**************************************************************************

SUBROUTINE SIM1T7(T,ZMIN,ZMAX,Z,IND,AA)

            IMPLICIT NONE
```

```
            INTEGER ZMIN,ZMAX,Z,T,IND(8,8),AA(8), J, I
            REAL(8) BMAX,CINV,LOGC,FDIST,U, DLNGAM, B(ZMIN:ZMAX)

!   ************************************************************************
!     find the constant
!   ************************************************************************

      DO J = ZMIN, ZMAX
            B(J)=0.0d0
            DO I = 1, 8

                  IF (IND(T,I) == 1) THEN
                        B(J)=B(J)+DLNGAM(AA(I)+1.0d0)
                  END IF
            END DO
      END DO

         BMAX = 0.0d0
            DO J = ZMIN, ZMAX
                  BMAX = MAX(B(J), BMAX)
            END DO

!   ************************************************************************
!     calculate constant C
!   ************************************************************************

      CINV = 0.0d0
            DO J = ZMIN, ZMAX
                  CINV = CINV + DEXP(BMAX-B(J))
            END DO
      LOGC = -DLOG(CINV)

!   ************************************************************************
!     simulate Z(I), I<8
!   ************************************************************************

      Z = ZMIN
      IF (ZMIN <> ZMAX) THEN
            FDIST = DEXP(LOGC+BMAX-B(Z))
            CALL DRNUN(1,U)
            DO WHILE ( U > FDIST)
                  Z=Z+1
                  FDIST = FDIST + DEXP(LOGC+BMAX-B(Z))
            END DO
      END IF

END SUBROUTINE

!************************************************************************
!     THE SUBROUTINE  PEARS CALCULATES THE X^2 PEARSON
!************************************************************************
      SUBROUTINE PEARS(K,Y,EY,STS)

            IMPLICIT NONE

            INTEGER  I,K
            REAL(8)  EY(8),STS,Y(8)

            STS=0.0d0
```

```
            DO I=1,K
                  IF (EY(I)==0.0) THEN
                        EY(I)=0.001
                  END IF
            STS=STS+((Y(I)-EY(I))**2/EY(I))
            END DO

      END SUBROUTINE


!**********************************************************************
!    THE SUBROUTINE  EXPECT1 FINDS CELL ESTIMATES UNDER THE MODEL (XZ, YZ).
!    IS APPLIED ONLY ONCE AT THE BEGINNING,  AND DOES NOT CHANGE DURING
!    SIMULATIONS, BECAUSE THE ESTIMATES DEPEND ONLY ON  Z\R.
!**********************************************************************

      SUBROUTINE EXPECT1(K,ZR,EY,IFL)

            IMPLICIT NONE

            INTEGER   I, K,IFL,II
            REAL(8)  EY(8), SUMEY, ZR(8)
            II=0
            IFL=0

            EY(1)=ZR(5)*ZR(2)/ZR(6)
            EY(2)=(ZR(6)-ZR(5))*ZR(2)/ZR(6)
            EY(3)=(ZR(7)-ZR(5))*(ZR(4)-ZR(2))/(ZR(8)-ZR(6))
            EY(4)=ZR(4)-ZR(2)-EY(3)
            EY(5)=(ZR(6)-ZR(2))*ZR(5)/ZR(6)
            EY(6)=(ZR(6)-ZR(2))*(ZR(6)-ZR(5))/ZR(6)
            EY(7)=ZR(7)-ZR(5)-EY(3)
            SUMEY=0.0d0
                  DO I=1,7
                  SUMEY=SUMEY+EY(I)
                  END DO
            EY(8)=ZR(8)-SUMEY
                  DO I=1,K
                        IF (EY(I)<0.) THEN
                        II=1
                        END IF
                  IFL=IFL+II
                  END DO


      END SUBROUTINE


!**********************************************************************
!     THE SUBROUTINE  EXPECT2 FINDS CELL FREQUENCIES FOR THE NEW Z(I)'s,
!     BASED ON THE NEW Z_R AND THE FIXED Z_\R.
!**********************************************************************

      SUBROUTINE EXPECT2(K,ZR,YR,IFL)

            IMPLICIT NONE

            INTEGER   I, K,IFL,II
```

```
      REAL(8)  YR(8), SUMYR, ZR(8)

      II=0
      IFL=0

      YR(1)=ZR(1)
      YR(2)=ZR(2)-ZR(1)
      YR(3)=ZR(3)-ZR(1)
      YR(4)=ZR(4)-ZR(3)-ZR(2)+ZR(1)
      YR(5)=ZR(5)-ZR(1)
      YR(6)=ZR(6)-ZR(5)-ZR(2)+ZR(1)
      YR(7)=ZR(7)-ZR(5)-ZR(3)+ZR(1)
      SUMYR=0.0d0
            DO I=1,7
            SUMYR=SUMYR+YR(I)
            END DO
      YR(8)=ZR(8)-SUMYR
            DO I=1,K
                  IF (YR(I)<0.) THEN
                  II=1
                  END IF
            IFL=IFL+II
            END DO


END SUBROUTINE
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Agresti,A. (2001). *Categorical Data Analysis* (2nd edition). New York : Wiley
2. Agresti,A. D.Wackerly, and J.Boyett. Exact conditional tests for cross-classifications : approximations of attained significance levels. *Psychometrika* 1979; **44** : 75-83
3. Agresti,A. Exact inference for categorical data : recent advances and continuing controversies. *Statistics in Medicine* 2001; **20** : 2709-2722
4. Agresti,A., C.R.Mehta, N.R.Patel. Exact inference for contingency tables with ordered categories. *Journal of the American Statistical Association* 1990; **85** : 453-458
5. Agresti,A., M.Yang. An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics and Data Analysis* 1987; **5** : 9-21
6. Agresti,A. A survey of exact inference for contingency tables. *Statistical Science* 1992; **7**: 131-153
7. Baker,R.J., M.R.B.Clarke, P.W.Lane. Zero entries in contingency tables. *Computational Statistics and Data Analysis* 1985; **3** : 33-45
8. Barnard,G.A. A new test for 2x2 tables. *Nature* 1945; **156** : 177
9. Barnard,G.A. Significance tests for 2x2 tables. *Biometrika* 1947; **34** : 123-138
10. Berger,R.L., D.D.Boos. P-values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* 1994; **89** : 1012-1016
11. Berkson,J. In dispaise of the exact test. *Journal of Statistical Planning Inference* 1978; **2** : 27-42
12. Birch,M.W. Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society* 1963; Series B **25** : 220-233
13. Birch,M.W. The detection of partial association II : The general case. *Journal of the Royal Statistical Society*, Series B 1965; **27** : 111-124
14. Bishop,Y.M.M., S.E.Fienberg, and P.W.Holland. (1975). *Discrete Multivariate Analysis*. Cabridge, MA : MIT Press
15. Booth,J., R.Butler. Monte Carlo approximation of exact conditional tests for log-linear models. *Biometrika* 1999; **86** : 321-332
16. Cornfield,J. A statistical problem arising from retrospective studies. *Proc. Third Berkeley Symp. Math. Statist. Probab.* 1956; **4** : 135-148

17. Cressie,N., and T.R.C.Read. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*. New York : Springer
18. D'Agostino,R.B., W.Chase, and Belanger. The appropriateness of some common procedures for testing the equality of two independent binomial populations. *American Statistician* 1988; **42** : 198-202
19. Davis,L.J. Exact tests for 2 by 2 contingency tables. *American Statistician* 1986; **40** : 139-141
20. Dupont,W.D. Sensitivity of Fisher's exact test to minor perturbations in 2x2 contingency tables. *Statistics in Medicine* 1986; **5** ; 629-635
21. Fienberg,S.E. Quasi-independence and maximum likelihood estimation in incomplete contingency tables. *Journal of the American Statistical Association* 1970; **65** : 1610-1616
22. Fienberg,S.E. The use of chi-squared statistics for categorical data problems. *Journal of the Royal Statistical Society* 1979; **41** : 54-64
23. Fisher,R.A. (1935). *The Design of Experiments* (8th ed. 1966). Edinburgh : Oliver & Boyd
24. Forster,J.J., J.W.McDonald, P.W.F.Smith. Monte Carlo exact conditional tests for log-linear and logistic models. *Journal of the Royal Statistical Society*, Series B, Methodological 1996; **58** : 445-453
25. Forster,J.J., J.W.McDonald, P.W.F.Smith. Monte Carlo exact tests for square contingency tables. *Journal of the Royal Statistical Society*, Series A, General 1996; **159** : 309-321
26. Freeman,G.H., and J.H.Halton. Note on an exact treatment of contingency, goodness-of-fit and other problems of significance. *Biometrika* 1951; **38** : 141-149
27. Gilks,W.R., S.Ritchardson, and D.J.Spiegelhalter. (1996). *Markov Chain Monte Carlo in practice*. London : Chapman & Hall
28. Glonek,G., J.N.Darroch, and T.P.Speed. On the existence of maximum likelihood estimators for hierarchical log-linear models. *Scandinavian Journal of Statistics* 1988; **15** : 187-193
29. Haberman,S.J. A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *Journal of the American Statistical Association* 1988; **83** ; 555-560
30. Haberman,S.J. Loglinear models and frequency tables with small expected cell counts. *Annals of Statistics* 1977; **5** : 1148-1169

31. Haberman,S.J. Loglinear models for frequency data : Sufficient statistics and likelihood equations. *Annals of Statistics* 1973; **1** : 617-632

32. Hirji, K.F., C.R.Mehta, N.R.Patel. Computing distributions for exact logistic regression. *Journal of the American Statistical Association* 1987; **82** : 1110-1117

33. Kempthorne,O. In dispraise of the exact test : Reactions. *Journal of Statistical Planning Inference* 1979; **3** : 199-213

34. Kim,D., A.Agresti. Nearly exact tests of conditional independence and marginal homogeneity for sparse contingency tables. *Computational and Graphical Statistics* 1992; **1** : 21-40

35. Kim,D., and A.Agresti. Improved exact inference about conditional association in three-way contingency tables. *Journal of the American Statistical Association* 1995; **90** : 632-639

36. Klotz,J., J.Teng. One-Way layout for counts and the exact enumeration of the Kruskal-Wallis H distribution with ties. *Journal of the American Statistical Association* 1977; **72** : 165-169

37. Koehler,K.J. Goodness-of-fit tests for log-linear models in sparse contingency tables. *Journal of the American Statistical Assosiation* 1986; **81** : 483-493

38. Koehler,K.J., K.Larntz. An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association* 1980; **75** : 336-344

39. Kruskal,W.H. Ordinal measures of association. *Journal of the American Statistical Association* 1958; **53** : 814-861

40. Lancaster,H.O. Significance test in discrete distributions (corrections **57** : 919). *Journal of the American Statistical Association* 1961; **56** : 223-234

41. Liu,J.S. (1999). *Monte Carlo Strategies in Scientific Computing*. New York : Springer

42. Lloyd,C.J. Doubling the one-sided P-value in testing independence in 2x2 tables against a two- sided alternative. *Statistics in Medicine* 1988; **7** : 1297-1306

43a.Mantel,N. Chi-Square tests with one degree of freedom : extensions of the Mantel-Haenszel procedure. Journal of the American Statistical Association 1963; 58 : 690-700

43b.Mantel,N. Exact tests for 2x2 contingency tables (Letter). *American Statistician* 1987; **42**: 159

44. Mantel,N., and W.Haenszel. Statistical aspects of the analysis of retrospective studies of disease. Journal of Natural Cancer Institute 1959; 22 : 719-748

45. McCullagh,P., and J.A.Nelder. (1989). *Generalized Linear Models*, (2nd edition). Chapman and Hall, London
46. McDonald,J.W., D.C.DeRoure, D.T.Michaelides. Exact tests for two-way symmetric contingency tables. *Statistics and Computing* 1998; **8** : 391-399
47. McDonald,J.W., P.W.F.Smith. Exact conditional tests of quasi-independence for triangular contingency tables : estimating attained significance levels. *Applied Statistics* 1995; **44** : 143-151
48. Mehta,C.R., J.F.Hilton. Exact power of conditional and unconditional tests : Going beyond the 2x2 contingency tables. *American Statistician* 1993; **47** : 91-98
49. Mehta,C.R., N.R.Patel, and R.Gray. Computing an exact confidence interval for the common odds ratio in several 2 by 2 contingency tables. *Journal of the American Statistical Association* 1985; **80** : 969-973
50. Mehta,C.R., N.R.Patel, and Senchaudhuri. Importance Sampling for estimating exact probabilities in permutational inference. *Journal of the American Statistical Association* 1988; **83** : 999-1005
51. Mehta,C.R., N.R.Patel. A Fortran subroutine for Fisher's exact test in unordered *r*x*c* contingency tables. *ACM Trans. Math. Softwar*e 1986; **12** : 154-161
52. Mehta,C.R., N.R.Patel. A Network algorithm for perfoming Fisher's exact test in *r*x*c* contingency tables. *Journal of the American Statistical Association* 1983; **78** : 427-434
53. Raftery,A.E., and S.Lewis. (1992). *How many iterations in the Gibbs Sampler?* Bayesian Statistics 4. pp. 777-784. Oxford : Oxford University Press
54. Robert,P., G.Casella. (1999). *Monte Carlo Statistical Methods*. New York : Springer
55. Smith,P.W.F., J.W.McDonald. Exact conditional tests for incomplete contingency tables : estimating attained significance levels. *Statistics and Computing* 1995; **5**: 253-256
56. Smith,P.W.F., J.W.McDonald. Simulate and reject Monte Carlo exact conditional tests for quasi-independence. *COMPSTAT. Proceedings in Computational Statistics, 11th Symposium* 1994; 509-514
57. StatXact (1998). *A Statistical Package for Exact Nonparametric Inference* (version 4.0). Cambridge, MA : CYTEL Software Corporation
58. Suissa,S., J.J.Shuster. Are uniformly most powerful unbiased tests really best? *American Statistisian* 1984; **38** : 204-206

59. Suissa,S., J.J.Shuster. Exact unconditional sample sizes for the 2 by 2 binomial trials. *Journal of the Royal Statistical Society* 1985; **148** : 317-327

60. Upton,G.J.G. A comparison of alternative tests for the 2x2 comparative trial. *Journal of the Royal Statistical Society*, Series A 1982; **145** : 86-105

61. Verbeek,A., and P.M.Kroonenberg. A survey of algorithms for exact distributions of test statistics in *rxc* contingency tables with fixed margins. *Computational Statistics and Data Analysis* 1985; **3** : 159-185

62. Yates,F. Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society* 1934; Suppl. **1**: 217-235

63. Yates,F. Tests of significance for 2x2 contingency tables. *Journal of the Royal Statistical Society* 1984; Series A **147** : 426-463

64. Zelen,M. The analysis of several 2x2 contingency tables. *Biometrika* 1971; **58** : 129-137

65. Zelterman,D. Goodness-of-fit tests for large sparse multinomial disributions. *Journal of the American Statistical Association* 1987; **82** : 624-629