Εισαγωγή στην Αλγεβρική Στατιστική

Δανάη Δεληγεωργάκη

Εργασία στα πλαίσια του μαθήματος
Ειδικά Θέματα
Καθηγήτρια: Χαραλάμπους Χαρά
Τμήμα Μαθηματικών, Α.Π.Θ.

Θεσσαλονίκη, Μάρτιος 2018

Περιεχόμενα

ΠΡΟΛΟΓΟΣ

Η εργασία αυτή στην πρώτη της μορφή ξεκίνησε από κοινού με την Ελένη Παππά το χειμερινό εξάμηνο του ακαδημαϊκού έτους 2016-2017 και ολοκληρώθηκε το Μάρτιο του 2018. Η Ελένη Παππά, η οποία είναι πλέον μεταπτυχιακή φοιτήτρια του τμήματος Μαθηματικών του Α.Π.Θ. στην ειδίκευση της Στατιστικής, συμμετείχε σε αυτήν έως την αποφοίτησή της, το Μάρτιο του 2017.

Η Αλγεβρική Στατιστική είναι ένας κλάδος των μαθηματικών που παρουσιάζει ραγδαία εξέλιξη τα τελευταία 30 χρόνια. Αφορά κυρίως στην εφαρμογή μεθόδων της Αλγεβρικής Γεωμετρίας, της Αντιμεταθετικής Άλγεβρας και της Συνδυαστικής για την επίλυση προβλημάτων της Στατιστικής. Από τη μία πλευρά, η Άλγεβρα προσφέρει ισχυρά εργαλεία για τη λύση προβλημάτων της Στατιστικής. Από την άλλη, σπανίως οι ήδη υπάρχουσες αλγεβρικές μέθοδοι είναι επαρκείς για να αντιμετωπίσουν τις προκλήσεις της Στατιστικής και συνήθως χρειάζεται να αναπτυχθούν νέες τεχνικές στην Άλγεβρα. Συνεπώς, ο διάλογος μεταξύ Άλγεβρας και Στατιστικής ωφελεί και τους δύο τομείς.

Στο πρώτο κεφάλαιο της εργασίας παρουσιάζονται τα βασικά στοιχεία της θεωρίας διωνυμικών ιδεωδών. Δίδεται έμφαση στην έννοια της βάσης Groebner (απλή, ελαχιστική, ανάγωγη, καθολική) και της βάσης Graver. Επίσης, μελετώνται τα τορικά ιδεώδη και οι Μαρκοβιανές βάσεις για αυτά.

Στο δεύτερο κεφάλαιο ορίζονται τα πλέγματα ακεραίων και οι αντίστοιχες βάσεις Groebner και Graver για αυτά. Επιπλέον, παρουσιάζεται η έννοια της ίνας ενός διανύσματος φυσικών αριθμών και ορισμένα προβλήματα που προκύπτουν για τις ίνες. Τα προβλήματα αυτά μπορούν να λυθούν με χρήση των Μαρκοβιανών βάσεων, που ορίζονται στη συνέχεια, για πλέγματα ακεραίων. Κατόπιν, παρουσιάζεται η σύνδεση που υπάρχει μεταξύ των διωνυμικών ιδεωδών και των πλεγμάτων ακεραίων, η οποία οδηγεί στο Θεμελιώδες Θεώρημα των Μαρκοβιανών βάσεων.

Για να γίνει αντιληπτή η σχέση του παραπάνω Θεωρήματος με τη Στατιστική είναι απαραίτητο να προηγηθούν κάποιες βασικές έννοιες της Στατιστικής, οι οποίες υπάρχουν στο τρίτο κεφάλαιο. Πρωταρχικό ρόλο παίζουν οι πίνακες συνάφειας και οι αντίστοιχοι πίνακες από κοινού πιθανότητας. Ακόμη, ο έλεγχος Χ-τετράγωνο και ο έλεγχος Fisher, οι οποίοι περιγράφονται συνοπτικά.

Η κεντρική ιδέα του ελέγχου Fisher γενικεύεται για περισσότερες από 2 μεταβλητές στο τέταρτο κεφάλαιο. Ο έλεγχος πραγματοποιείται με τη χρήση Μαρκοβιανών βάσεων, οι οποίες εισάγονται για μία κλάση μοντέλων που ονομάζονται λογαριθμογραμμικά (log-linear models). Οι αλγεβρικές έννοιες που αναπτύχθηκαν στα πρώτα κεφάλαια χρησιμοποιούνται προκειμένου να επιλυθεί το σημαντικότερο πρόβλημα της εργασίας, που αφορά έναν έλεγχο υποθέσεως.

Στο πέμπτο κεφάλαιο υπάρχει μία σύνοψη των κυριότερων θεμάτων της εργασίας και ορισμένα παραδείγματα. Αξίζει να σημειωθεί ότι η διαδικασία αυτή συναντά πολυάριθμες εφαρμογές σε κλάδους της βιολογίας και της ιατρικής.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά την Ελένη Παππά για την συνδρομή της, όπως και όλους τους προπτυχιακούς και μεταπτυχιακούς φοιτητές, υποψήφιους Διδάκτορες

και Καθηγητές που συνέβαλαν σε αυτήν την προσπάθεια. Ιδιαίτερα οφείλω να ευχαριστήσω τους απόφοιτους του τμήματος Κώστα Κάρτα και Φώτη Πατενίδη για την συνεχή και πολύτιμη βοήθειά τους. Θα ήθελα επίσης να ευχαριστήσω τον κύριο Χρήστο Τατάκη, του οποίου το αντικείμενο με ενέπνευσε να συνεχίσω αυτήν την προσπάθεια. Τέλος, ευχαριστώ από καρδιάς την κυρία Χαραλάμπους για την πρόθυμη συνεργασία και τις εύστοχες παρατηρήσεις της.

Κεφάλαιο 1

Διωνυμικά Ιδεώδη

Σε αυτό το κεφάλαιο παρουσιάζονται συνοπτικά τα διωνυμικά ιδεώδη και κάποιες βάσεις για αυτά όπως οι βάσεις Groebner, οι βάσεις Graver και οι Μαρκοβιανές βάσεις. Στόχος είναι ο υπολογισμός των παραπάνω βάσεων για μία υποκατηγορία των διωνυμικών ιδεωδών, τα λεγόμενα τορικά ιδεώδη. Στα παρακάτω θεωρούμε ότι τα πολυώνυμα που μελετώνται έχουν συντελεστές από το σώμα $\mathbb{R}$ των πραγματικών αριθμών.

1.1 Εισαγωγή στα Διωνυμικά Ιδεώδη

Ας θυμηθούμε ότι ένα μη κενό υποσύνολο $\mathcal{I}$ ενός αντιμεταθετικού δακτυλίου R με μοναδιαίο στοιχείο λέγεται **ιδεώδες** του R αν για κάθε $a, b \in \mathcal{I}$ ισχύει ότι $a - b \in \mathcal{I}$ και για κάθε $a \in \mathcal{I}$ και $r \in R$ ισχύει ότι $ra \in \mathcal{I}$.

Στο εξής θα συμβολίζουμε το τυχαίο σώμα ως $\mathbb{K}$.

Ορισμός 1.1.1. Ένα **μονώνυμο** στον πολυωνυμικό δακτύλιο $\mathbb{K}[p] := \mathbb{K}[p_1, p_2, ..., p_k]$ είναι ένα γινόμενο της μορφής $p^u := p_1^{u_1} p_2^{u_2} \cdots p_k^{u_k}$, όπου $u := (u_1, u_2, ..., u_k) \in \mathbb{N}^k$. Η διαφορά δύο μονωνύμων ορίζεται ως **διώνυμο**. Ένα ιδεώδες που παράγεται αποκλειστικά από διώνυμα καλείται **διωνυμικό ιδεώδες**.

Παράδειγμα 1.1.2. Το ιδεώδες $\mathcal{I} =< p^a - p^b \ : \ a, b \in \mathbb{N}^k >$ είναι ένα διωνυμικό ιδεώδες του $\mathbb{K}[p]$.

Ένα εύλογο ερώτημα που προκύπτει για τα διωνυμικά ιδεώδη είναι το αν μπορούν πάντα να περιγραφούν από ένα πεπερασμένο σύνολο· αν είναι, δηλαδή, πεπερασμένα παραγόμενα. Αυτό μπορεί εύκολα να απαντηθεί μέσω της θεωρίας των δακτυλίων της Noether.

Ορισμός 1.1.3. Ένας δακτύλιος R λέγεται **δακτύλιος της Noether** αν κάθε ιδεώδες του, $\mathcal{I}$, είναι πεπερασμένα παραγόμενο, δηλαδή υπάρχουν $f_1, f_2, ..., f_n$, ούτως ώστε $\mathcal{I} =< f_1, f_2, ..., f_n >$.

Θεώρημα 1.1.4. *Αν το σώμα $\mathbb{K}$ είναι δακτύλιος της Noether, τότε και ο πολυωνυμικός δακτύλιος $\mathbb{K}[p]$ είναι δακτύλιος της Noether.*

Απόδειξη. Είναι άμεση συνέπεια του Θεωρήματος βάσης του Hilbert (βλέπε (9), Θεώρημα 3.2.1). ■

Παράδειγμα 1.1.5. Κάθε σώμα $\mathbb{K}$ είναι δακτύλιος της Noether, αφού τα μοναδικά ιδεώδη του είναι το $< 1 >= \mathbb{K}$ και το $< 0 >= 0$. Επομένως, ο πολυωνυμικός δακτύλιος $\mathbb{K}[p]$ είναι δακτύλιος της Noether.

Κάθε διωνυμικό ιδεώδες είναι ιδεώδες του δακτυλίου $\mathbb{K}[p]$, ο οποίος είναι δακτύλιος της Noether. Επομένως, κάθε διωνυμικό ιδεώδες είναι πεπερασμένα παραγόμενο.

Το πεπερασμένο σύνολο από το οποίο παράγεται δεν είναι μοναδικό. Ειδικότερα, συνήθως υπάρχουν άπειρα τέτοια σύνολα.

Ορισμός 1.1.6. Καλούμε **βάση** ενός διωνυμικού ιδεώδους $\mathcal{I}$ οποιοδήποτε σύνολο γεννητόρων του, δηλαδή, οποιοδήποτε υποσύνολο του $\mathcal{I}$, το οποίο το παράγει. Καλούμε **ελαχιστοτική βάση** του $\mathcal{I}$ κάθε ελαχιστοτικό σύνολο γεννητόρων του $\mathcal{I}$, δηλαδή κάθε σύνολο γεννητόρων το οποίο δεν έχει γνήσιο υποσύνολο που να παράγει το $\mathcal{I}$.

Υπάρχουν διάφορες βάσεις για τα διωνυμικά ιδεώδη του δακτυλίου $\mathbb{K}[p]$, όμως κάποιοι τύποι βάσεων είναι πιο διαδεδομένοι λόγω ορισμένων ιδιοτήτων τους, όπως οι βάσεις Groebner και οι βάσεις Graver.

1.2 Βάσεις Groebner

Η θεωρία των βάσεων Groebner παρουσιάστηκε από τον Bruno Buchberger μαζί με έναν αλγόριθμο για τον υπολογισμό τους (αλγόριθμος Buchberger) το 1965 και έδωσε λύση σε πολλά προβλήματα της Αντιμεταθετικής Άλγεβρας και της Αλγεβρικής Γεωμετρίας. Συνοπτικά, μία βάση Groebner είναι ένα σύνολο πολυωνύμων πολλών μεταβλητών που έχει κάποιες επιθυμητές αλγοριθμικές ιδιότητες.

Κάθε σύνολο πολυωνύμων μπορεί να μετασχηματισθεί σε μια βάση Groebner, χάρη στον αλγόριθμο Buchberger και μάλιστα, αποδεικνύεται ότι μια βάση Groebner ενός διωνυμικού ιδεώδους αποτελείται αποκλειστικά από διώνυμα. Έτσι, αν έχουμε ένα σύνολο γεννητόρων ενός διωνυμικού ιδεώδους μπορούμε πάντα να υπολογίσουμε μια βάση Groebner για αυτό. Ο υπολογισμός αυτός είναι γενικά χρονοβόρα διαδικασία, για αυτό χρησιμοποιούνται συνήθως προγράμματα εξειδικευμένα σε τέτοιους υπολογισμούς, όπως το $4ti2$ και το CoCoA. Παρακάτω παρουσιάζονται ορισμένα στοιχεία από τη θεωρία των βάσεων Groebner.

Συμβολίζουμε με $T[n]$ το σύνολο των μονωνύμων $p^u = p_1^{u_1} p_2^{u_2} \cdots p_n^{u_n}$ στο δακτύλιο $\mathbb{K}[p]$.

Ορισμός 1.2.1. Μια **μονωνυμική διάταξη** στο $T[n]$ είναι μια ολική διάταξη του $T[n]$, όπου
(i) $1 < p^u$, για κάθε $p^u \in T[n]$ με $p^u \neq 1$ και
(ii) αν $p^a < p^b$ τότε $p^a p^c < p^b p^c$, για κάθε $p^c \in T[n]$.

Αποδεικνύεται ότι για $n \geq 2$ υπάρχουν άπειρες μονωνυμικές διατάξεις στο $\mathcal{T}[n]$ (βλέπε (11), Κεφάλαιο 2). Για $n = 1$ υπάρχει μοναδική διάταξη και είναι η εξής:

$$1 < p_1 < p_1^2 < \cdots < p_1^k \cdots .$$

Παράδειγμα 1.2.2. Ας υποθέσουμε ότι $p_n < p_{n-1} < \cdots < p_1$. Μία μονωνυμική διάταξη του $T[n]$, που είναι γνωστή ως **λεξικογραφική διάταξη**, είναι η εξής: $p^a <_{lex} p^b$ αν και μόνο αν η πρώτη μη μηδενική συντεταγμένη του διανύσματος $b - a$ είναι θετική. Εφόσον ισχύουν οι προϋποθέσεις του ορισμού (1.2.1) η λεξικογραφική διάταξη είναι μία μονωνυμική διάταξη του $T[n]$.

Παρατήρηση 1.2.3. Μέσω των μονωνυμικών διατάξεων καθίσταται εφικτή η διάταξη των όρων των πολυωνύμων του $\mathbb{K}[p]$. Στο εξής, αν $<$ είναι μια μονωνυμική διάταξη του $T[n]$ και f ένα μη μηδενικό πολυώνυμο του $\mathbb{K}[p]$, θα γράφουμε $f = a_1 p^{u_1} + a_2 p^{u_2} + \cdots a_s p^{u_s}$, όταν $a_i \neq 0$ για κάθε $i \in \{1, 2, ..., s\}$ και $p^{u_1} > p^{u_2} > \cdots > p^{u_s}$.

Το μονώνυμο p^{u_1} ονομάζεται **αρχικό μονώνυμο του** f και συμβολίζεται ως $\text{in}_<(f)$. Ο συντελεστής a_1 καλείται **αρχικός συντελεστής του** f και συμβολίζεται με $\text{lc}(f)$.

Χρησιμοποιώντας αυτούς τους συμβολισμούς είμαστε σε θέση να δώσουμε τον παρακάτω ορισμό. Έστω $<$ μία διάταξη των στοιχείων του $\mathbb{K}[p]$.

Ορισμός 1.2.4. Ένα πεπερασμένο υποσύνολο $G = \{g_1, ..., g_t\}$ μη-μηδενικών πολυωνύμων ενός ιδεώδους $\mathcal{I}$, ονομάζεται **βάση Groebner** του $\mathcal{I}$ και συμβολίζεται με $G_<$ αν και μόνο αν για κάθε μη μηδενικό πολυώνυμο $f \in \mathcal{I}$ υπάρχει κάποιο $i \in \{1, ..., t\}$, ώστε το αρχικό μονώνυμο $in_<(g_i)$ να διαιρεί το $in_<(f)$.

Μία βάση Groebner $G_< = \{g_1, g_2, ..., g_t\}$ καλείται **ελαχιστική βάση Groebner** (minimal) αν $\text{lc}(g_i) = 1$ και το αρχικό μονώνυμο $\text{in}_<(\text{g}_i)$ δεν διαιρεί το $\text{in}_<(\text{g}_j)$, για κάθε $i, j \in \{1, 2, ..., t\}$ με $i \neq j$.

Μια βάση Groebner $G_< = \{g_1, ..., g_t\}$ ονομάζεται **ανάγωγη βάση Groebner** (reduced) αν $\text{lc}(g_i) = 1$ και δεν υπάρχει μη μηδενικός όρος του g_i, ο οποίος να διαιρείται από κάποιο αρχικό μονώνυμο του g_j, για κάθε $i, j \in \{1, 2, ..., t\}$ με $j \neq i$.

Σε αντίθεση με τα παραπάνω, η λεγόμενη καθολική βάση Groebner είναι ανεξάρτητη της διάταξης που έχουμε επιλέξει.

Ορισμός 1.2.5. Καλούμε **καθολική βάση Groebner** ενός ιδεώδους $\mathcal{I}$ την ένωση όλων των ανάγωγων βάσεων Groebner $G_<$ του $\mathcal{I}$ που προκύπτουν για κάθε δυνατή διάταξη $<$ των στοιχείων του $\mathcal{I}$.

Εξ ορισμού μια ανάγωγη βάση Groebner είναι και ελαχιστική. Το αντίθετο δεν ισχύει. Μπορεί να υπάρχουν άπειρες ελαχιστικές βάσεις Groebner για το $\mathcal{I}$ ως προς μία διάταξη $<$, αλλά μόνο μία από αυτές είναι ανάγωγη.

Παράδειγμα 1.2.6. Με χρήση του αλγορίθμου Buchberger (βλέπε (10), ενότητα 2.4) υπολογίζεται μία βάση Groebner του ιδεώδους

$$\mathcal{I} =< p_1 - p_2, p_2 p_4 - p_3^2, p_2^3 - p_3 p_4, p_2^2 p_3 - p_4^2 > \subset \mathbb{R}[p_1, p_2, p_3, p_4]$$

ως προς τη λεξικογραφική διάταξη $<_{lex}$. Είναι το σύνολο

$$\mathrm{Gr}_< = \{p_1-p_2,\ p_1^4p_2^6-p_4^6,\ p_2p_4-p_3^2,\ p_1^{100}-p_2^{100},\ p_2^3-p_3p_4,\ p_2^2p_3-p_4^2,\ p_2p_3^3-p_4^3,\ p_3^5-p_4^4,\ p_1p_2^3-p_3^3\}.$$

Αυτή η βάση μπορεί να μετασχηματισθεί στην αντίστοιχη ανάγωγη βάση Groebner του $\mathcal{I}$, που είναι το σύνολο $\{p_1-p_2, p_2p_4-p_3^2, p_2^3-p_3p_4, p_2^2p_3-p_4^2, p_2p_3^3-p_4^3, p_3^5-p_4^4\}$.

Αποδεικνύεται (βλέπε (11), Παρατήρηση 2.1.9) ότι η καθολική βάση Groebner ενός ιδεώδους $\mathcal{I}$ είναι πάντοτε πεπερασμένο σύνολο, δηλαδή το πλήθος των ανάγωγων βάσεων Groebner κάθε ιδεώδους $\mathcal{I}$ είναι πεπερασμένο. Η καθολική βάση Groebner του αποτελεί βάση Groebner του $\mathcal{I}$ ως προς οποιαδήποτε μονωνυμική διάταξη του $\mathbb{K}[p]$.

Παράδειγμα 1.2.7. Η καθολική βάση Groebner του ιδεώδους $\mathcal{I}$ του Παραδείγματος 1.2.6 αποτελείται από τα στοιχεία της ανάγωγης βάσης Groebner ως προς τη λεξικογραφική διάταξη και 6 επιπλέον στοιχεία. Συγκεκριμένα, είναι το σύνολο

$$U = \{p_1-p_2, p_2p_4-p_3^2, p_2^3-p_3p_4, p_2^2p_3-p_4^2, p_2p_3^3-p_4^3, p_3^5-p_4^4, p_1p_4-p_3^2, p_1^3-p_3p_4,$$
$$p_1^2p_3-p_4^2, p_1p_3^3-p_4^3, p_1^4-p_3^3, p_2^4-p_3^3, p_1^5-p_4^3, p_2^5-p_4^3\}.$$

1.3 Βάσεις Graver

Οι βάσεις Graver είναι ένα πολύτιμο εργαλείο της Αντιμεταθετικής Άλγεβρας. Ένας από τους λόγους που τις καθιστούν σημαντικές είναι ότι στη βάση Graver ενός διωνυμικού ιδεώδους περιέχεται η καθολική βάση Groebner, καθώς και άλλα ιδιαίτερα σύνολα διωνύμων που θα εξεταστούν παρακάτω. Υπολογίζονται και αυτές με τη χρήση προγραμμάτων, όπως τα 4*ti*2 και CoCoA ή αλγορίθμων (βλέπε (11), Αλγόριθμος 3.3.6).

Ορισμός 1.3.1. Ένα διώνυμο $p^{u^+}-p^{u^-}\in\mathcal{I}$ καλείται **πρωταρχικό**, εάν δεν υπάρχει άλλο μη μηδενικό διώνυμο της μορφής $p^{v^+}-p^{v^-}\in\mathcal{I}$ τέτοιο ώστε: $p^{v^+}\mid p^{u^+}$ και $p^{v^-}|p^{u^-}$.

Το σύνολο όλων των πρωταρχικών διωνύμων ενός διωνυμικού ιδεώδους $\mathcal{I}$ καλείται **βάση Graver** αυτού.

Στο επόμενο θεώρημα παρουσιάζεται η σχέση της καθολικής βάσης Groebner με τη βάση Graver.

Θεώρημα 1.3.2. *Κάθε στοιχείο $p^u := p^{u^+}-p^{u^-}$ της καθολικής βάσης Groebner ενός διωνυμικού ιδεώδους $\mathcal{I}$ είναι πρωταρχικό.*

Απόδειξη. Βλέπε (8), Θεώρημα 2.2.4. ■

Επομένως, ως προς οποιαδήποτε μονωνυμική διάταξη των στοιχείων του $\mathcal{I}$, ισχύει ο εγκλεισμός:

$$\text{ανάγωγη βάση Groebner} \subset \text{καθολική βάση Groebner} \subset \text{βάση Graver}.$$

Η βάση Graver του διωνυμικού ιδεώδους $\mathcal{I}$ είναι βάση Groebner για το $\mathcal{I}$ ως προς κάθε μονωνυμική διάταξη, αφού αποτελείται από την καθολική βάση Groebner και πεπερασμένου πλήθους στοιχεία του $\mathcal{I}$.

Παράδειγμα 1.3.3. Η βάση Graver του ιδεώδους $\mathcal{I}$ του Παραδείγματος 1.2.6 αποτελείται από τα στοιχεία του συνόλου

$$\begin{aligned}\mathrm{Gr}_{\mathcal{I}} = \{&p_1-p_2,\ p_2p_4-p_3^2,\ p_2^3-p_3p_4,\ p_2^2p_3-p_4^2,\ p_2p_3^3-p_4^3,\ p_3^5-p_4^4, p_1p_4-p_3^2,\\ &p_1^3-p_3p_4,\ p_1^2p_3-p_4^2,\ p_1p_3^3-p_4^3,\ p_1^4-p_3^3,\ p_2^4-p_3^3,\ p_1^5-p_4^3, p_2^5-p_4^3,\ p_1p_2p_3-p_4^2,\ p_1p_2^2\\ &-p_3p_4,\ p_1^2p_2-p_3p_4,\ p_1p_2^3-p_3^3,\ p_1^2p_2^2-p_3^3,\ p_1^3p_2^1-p_3^3,\ p_1p_2^4-p_4^3,\ p_1^4p_2-p_4^3,\ p_1^2p_2^3-p_4^3,\\ &p_1^3p_2^2-p_4^3\},\end{aligned}$$

αν για κάθε στοιχείο συμπεριληφθεί και το αντίθετό του, ως προς το πρόσημο, στοιχείο.

1.4 Τορικά Ιδεώδη

Παρακάτω εισάγεται η έννοια του τορικού ιδεώδους, το οποίο έχει πρωταρχικό ρόλο στην αλγεβρική στατιστική.

Ορισμός 1.4.1. Έστω $\mathcal{A}$ ένας πίνακας διάστασης $m \times n$ με συντελεστές φυσικούς αριθμούς. Καλούμε **τορικό ιδεώδες** που αντιστοιχεί στον πίνακα $\mathcal{A}$ το διωνυμικό ιδεώδες

$$\mathcal{I}_{\mathcal{A}} \ := \ \langle p^u - p^v \ : \ u, v \in \mathbb{N}^n \text{ και } \mathcal{A}u = \mathcal{A}v \rangle.$$

Ένα πεπερασμένο σύνολο γεννητόρων του τορικού ιδεώδους $\mathcal{I}_{\mathcal{A}}$ καλείται **Μαρκοβιανή βάση** του $\mathcal{I}_{\mathcal{A}}$. Μία Μαρκοβιανή βάση ονομάζεται **ελαχιστοτική** αν δεν υπάρχει γνήσιο υποσύνολό της που να είναι Μαρκοβιανή βάση του $\mathcal{I}_{\mathcal{A}}$.

Οι βάσεις Groebner και Graver είναι υποπεριπτώσεις των Μαρκοβιανών βάσεων για τα τορικά ιδεώδη. Αποδεικνύεται ότι κάθε ελαχιστοτικό σύστημα γεννητόρων του $\mathcal{I}_{\mathcal{A}}$ ανήκει στη βάση Graver, διότι κάθε στοιχείο του είναι πρωταρχικό. Δηλαδή, κάθε ελαχιστοτική Μαρκοβιανή βάση είναι υποσύνολο της βάσης Graver. Επιπλέον, εξ ορισμού σε κάθε ανάγωγη βάση Groebner περιέχεται τουλάχιστον μια ελαχιστοτική Μαρκοβιανή βάση. Ο εγκλεισμός της προηγούμενης ενότητας μπορεί να συμπληρωθεί τώρα για τις διάφορες βάσεις του $\mathcal{I}_{\mathcal{A}}$ ως εξής:

$$\text{ελαχιστοτική Μαρκοβιανή βάση} \subset \text{ανάγωγη βάση Groebner}$$

$$\subset \text{ καθολική βάση Groebner } \subset \text{ βάση Graver} \tag{1.1}$$

Στην παραπάνω σχέση θα πρέπει να σημειωθεί ότι ο πρώτος εγκλεισμός δεν ισχύει απαραιτήτως για κάθε ελαχιστοτική Μαρκοβιανή βάση, αλλά για τουλάχιστον μία.

Παράδειγμα 1.4.2. Θεωρούμε τον πίνακα $\mathcal{A} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$.
Υπολογίζουμε το τορικό ιδεώδες $\mathcal{I}_\mathcal{A}$ που αντιστοιχεί στον πίνακα $\mathcal{A}$:

$$\mathcal{A}u = \mathcal{A}v \iff u_1 - v_1 = u_2 - v_2 \text{ και } u_3 - v_3 = u_4 - v_4.$$

Τα διώνυμα της μορφής $p^u - p^v$ που ικανοποιούν τις παραπάνω προϋποθέσεις παράγονται από το σύνολο $\mathcal{B} = \left\{p_1p_2 - 1, p_3p_4 - 1\right\}$.

Η βάση Graver του $\mathcal{I}_\mathcal{A}$ είναι το σύνολο $\mathcal{B}$ (βλέπε (8), Παράδειγμα 2.2.24). Επομένως, μία Μαρκοβιανή βάση του $\mathcal{I}_\mathcal{A}$ είναι το σύνολο $\mathcal{B}$. Ειδικότερα, το $\mathcal{B}$ είναι μια ελαχιστοτική Μαρκοβιανή βάση, καθώς $p_1p_2 - 1 \notin < p_3p_4 - 1 >$ και $p_3p_4 - 1 \notin < p_1p_2 - 1 >$.

Από την (1.1) έπεται ότι η ανάγωγη βάση Groebner του $\mathcal{I}_\mathcal{A}$ ως προς οποιαδήποτε μονωνυμική διάταξη του $T[4]$, όπως και η καθολική βάση Groebner του $\mathcal{I}_\mathcal{A}$ είναι το $\mathcal{B}$.

Κεφάλαιο 2

Πλέγματα ακεραίων και πλεγματικά ιδεώδη

Σε αυτήν την ενότητα παρουσιάζονται τα πλέγματα ακεραίων και η σύνδεσή τους με τα διωνυμικά ιδεώδη. Μέσω αυτής της σύνδεσης, ορίζονται και υπολογίζονται για τα πλέγματα ακεραίων οι βάσεις που είδαμε παραπάνω.

2.1 Πλέγματα ακεραίων

Ορισμός 2.1.1. Έστω $v_1, v_2, ..., v_m \in \mathbb{Z}^n$ ένα σύνολο γραμμικά ανεξάρτητων διανυσμάτων. Το **ακέραιο πλέγμα** ή **πλέγμα ακεραίων** (integral/integer lattice) $\mathcal{L}$ που παράγεται από τα διανύσματα $v_1, v_2, ..., v_m$ είναι το σύνολο των $\mathbb{Z}$-γραμμικών συνδυασμών τους, δηλαδή

$$\mathcal{L} = \{a_1v_1 + a_2v_2 + ... + a_mv_m \ : \ a_1, a_2, ..., a_m \in \mathbb{Z}\}.$$

Γενικότερα, καλούμε **πλέγμα ακεραίων** μια υποομάδα της προσθετικής ομάδας $\mathbb{Z}^n$, για κάποιο $n \in \mathbb{N}, n \geq 1$.

Καλούμε **υποπλέγμα** ενός πλέγματος ακεραίων $\mathcal{L}$, ένα υποσύνολο του $\mathcal{L}$ που είναι επίσης πλέγμα ακεραίων.

Ένα υποσύνολο $\mathcal{B} = \{b_1, b_2, ..., b_r\}$ ενός πλέγματος $\mathcal{L}$ ονομάζεται **πλεγματική βάση** του $\mathcal{L}$, εάν κάθε διάνυσμα v του $\mathcal{L}$ έχει μία μοναδική αναπαράσταση της μορφής

$$v \quad = \quad \lambda_1 b_1 \ + \ \lambda_2 b_2 \ + \ ... \ + \lambda_r b_r, \quad \text{όπου } \lambda_i \in \mathbb{Z}.$$

Όλες οι πλεγματικές βάσεις του $\mathcal{L}$ έχουν τον ίδιο πληθάριθμο r και η καθεμία προσδιορίζει έναν ισομορφισμό μεταξύ του $\mathcal{L}$ και του $\mathbb{Z}^r$. Ο αριθμός r ονομάζεται **βαθμίδα του πλέγματος** $\mathcal{L}$. Περισσότερα στοιχεία για τα πλέγματα ακεραίων υπάρχουν στο (13) (ενότητα 6.4).

2.2 Ίνα διανύσματος και Μαρκοβιανές βάσεις

Ορισμός 2.2.1. Έστω $\mathcal{L} \subset \mathbb{Z}^k$ ένα πλέγμα ακεραίων και $u \in \mathbb{N}^k$ ένα διάνυσμα φυσικών αριθμών. Καλούμε **ίνα** (fiber) **του u** το σύνολο των μη-αρνητικών διανυσμάτων που ανήκουν στην ίδια κλάση ισοδυναμίας modulo $\mathcal{L}$. Δηλαδή,

$$\mathcal{F}(u) \quad := \quad (u + \mathcal{L}) \cap \mathbb{N}^k \quad = \quad \Big\{ v \in \mathbb{N}^k \ : \ u - v \in \mathcal{L} \Big\}.$$

Έστω $\mathcal{L}$ ένα οποιοδήποτε υποπλέγμα του $\mathbb{Z}^k$ με την ιδιότητα το μόνο μη-αρνητικό διάνυσμά του να είναι το μηδενικό. Δηλαδή, το $\mathcal{L}$ είναι μια υποομάδα της ομάδας $(\mathbb{Z}^k, +)$ για την οποία ισχύει η σχέση

$$\mathcal{L} \cap \mathbb{N}^k = \{0\}.$$

Η παραπάνω υπόθεση εξασφαλίζει ότι η ίνα οποιουδήποτε σημείου $u \in \mathbb{N}^k$ είναι πεπερασμένο σύνολο. Υπάρχουν 4 θεμελιώδη προβλήματα που αφορούν στις ίνες αυτής της μορφής.
Ένα από αυτά είναι το πρόβλημα της βελτιστοποίησης επί της ίνας. Πρόκειται για το ακόλουθο πρόβλημα του ακέραιου προγραμματισμού για κάποιο διάνυσμα πραγματικών αριθμών w:

$$\text{να ελαχιστοποιηθεί το } w \cdot v \text{ όταν } v \ \in \ \mathcal{F}(u).$$

Εξίσου σημαντικό πρόβλημα είναι το πρόβλημα της δειγματοληψίας, το οποίο απαιτεί ένα τυχαίο σημείο της ίνας $\mathcal{F}(u)$, που λαμβάνεται μέσω μιας κατανομής στην $\mathcal{F}(u)$.

Τα 4 θεμελιώδη προβλήματα για τις ίνες της παραπάνω μορφής είναι:
(1) ο υπολογισμός του πλήθους των στοιχείων τους,
(2) η απαρίθμησή τους,
(3) η βελτιστοποίηση επί της ίνας και
(4) η λήψη δείγματος από αυτήν.

Τα 4 αυτά προβλήματα μπορούν να λυθούν αν είμαστε σε θέση να πραγματοποιήσουμε τυχαίους περιπάτους (random walks) που συνδέουν τα στοιχεία των ινών $\mathcal{F}(u)$ μέσω των στοιχείων του $\mathcal{L}$. Αυτό μπορεί να γίνει καλύτερα αντιληπτό μέσω της έννοιας της Μαρκοβιανής βάσης, της βάσης Groebner και της βάσης Graver για ένα πλέγμα ακεραίων.

Με τον όρο **βάση** ενός πλέγματος ακεραίων $\mathcal{L}$ εννοούμε οποιοδήποτε σύνολο γεννητόρων του $\mathcal{L}$.

Ορισμός 2.2.2. Μαρκοβιανή βάση ενός πλέγματος $\mathcal{L} \subseteq \mathbb{Z}^n$ ονομάζεται ένα *πεπερασμένο* σύνολο διανυσμάτων $\mathcal{M} \subseteq \mathcal{L}$ τέτοιο ώστε για κάθε a,$b \in \mathbb{N}^n$ για τα οποία $a - b \in \mathcal{L}$, να υπάρχει μια πεπερασμένη ακολουθία διανυσμάτων $\{z_1, z_2, ..., z_k\} \subset \mathbb{N}^n$ με στοιχεία μη αρνητικούς ακέραιους, όπου $z_1 = a$, $z_k = b$ και είτε $z_i - z_{i+1} \in \mathcal{M}$ για κάθε $i \in \{1, ..., k-1\}$, είτε $z_{i+1} - z_i \in \mathcal{M}$ για κάθε $i \in \{1, ..., k-1\}$.

Καλούμε **ελαχιστοτική Μαρκοβιανή βάση** του $\mathcal{L}$ οποιαδήποτε Μαρκοβιανή βάση του $\mathcal{L}$, που δεν είναι γνήσιο υπερσύνολο κάποιας Μαρκοβιανής βάσης του $\mathcal{L}$.

Μια βάση ενός πλέγματος ακεραίων $\mathcal{L}$ δεν είναι, εν γένει, Μαρκοβιανή βάση για το $\mathcal{L}$. Ωστόσο, μια Μαρκοβιανή βάση του $\mathcal{L}$ περιέχει πάντα μια πλεγματική βάση για το $\mathcal{L}$. Περισσότερες λεπτομέρειες σχετικά με τις Μαρκοβιανές και τις πλεγματικές βάσεις ενός πλέγματος ακεραίων υπάρχουν στο (1) (p.26-30).

Ορισμός 2.2.3. Ο **φορέας** $supp(u)$ ενός διανύσματος $u = (u_1, u_2, ..., u_n) \in \mathbb{Z}^n$, ορίζεται ως το σύνολο:

$$\big\{i \in \{1, 2, ..., n\} : u_i \neq 0\big\}.$$

Κάθε διάνυσμα $b \in \mathcal{L}$ μπορεί να γραφεί μοναδικά στη μορφή $\mathbf{b} = \mathbf{b}^+ - \mathbf{b}^-$, ως διαφορά δύο μη-αρνητικών διανυσμάτων με $supp(b^+) \cap supp(b^-) = \emptyset$.

Παράδειγμα 2.2.4. Για το πλέγμα ακεραίων $\mathcal{L} = \mathbb{Z}^4$, έχουμε

(1,1,1,-2)=(1,1,1,0)-(0,0,0,2),

(3,0,-1,-1)=(3,0,0,0)-(0,0,1,1),

(4,1,0,3)=(4,1,0,3)-(0,0,0,0).

Ορισμός 2.2.5. Έστω $b \in \mathcal{L}$. Ονομάζουμε **ίνα** (fiber) **του b** την κλάση ισοδυναμίας του $\mathbb{N}^k$ modulo $\mathcal{L}$ που περιέχει τα b^+ και b^-. Δηλαδή,

$$\text{fiber(b)} \ := \ \mathcal{F}(b^+) \ = \ \mathcal{F}(b^-).$$

Το επόμενο θεώρημα παρέχει χρήσιμες πληροφορίες για τον πληθάριθμο των Μαρκοβιανών βάσεων. Αποδεικνύεται με χρήση της θεωρίας γραφημάτων στο (1).

Θεώρημα 2.2.6. *Για μια ελαχιστοτική Μαρκοβιανή βάση $\mathcal{B}$ ενός πλέγματος $\mathcal{L}$, το σύνολο*

$$\big\{\textit{fiber(b)} \ : \ b \in \mathcal{B}\big\}$$

είναι μια αναλλοίωτος του πλέγματος $\mathcal{L} \subset \mathbb{Z}^k$ και συνεπώς το ίδιο ισχύει για το πλήθος στοιχείων της $\mathcal{B}$.

Απόδειξη. Βλέπε (1), Theorem 1.3.2. ■

Επομένως, παρόλο που οι Μαρκοβιανές βάσεις και οι ελαχιστοτικές Μαρκοβιανές βάσεις ενός πλέγματος $\mathcal{L}$ δεν είναι μοναδικές, όλες οι ελαχιστοτικές Μαρκοβιανές βάσεις του $\mathcal{L}$ έχουν τον ίδιο πληθάριθμο.

2.3 Τα πλεγματικά ιδεώδη

Υπάρχει η παρακάτω αντιστοιχία μεταξύ των διωνυμικών ιδεωδών και των πλεγμάτων ακεραίων.

Ορισμός 2.3.1. Το πλέγμα $\mathcal{L} \subset \mathbb{Z}^k$ αντιστοιχίζεται στο διωνυμικό ιδεώδες

$$I_{\mathcal{L}} \ := \ \big\langle p^u - p^v \ : \ u, v \in \mathbb{N}^k \text{ και } u - v \in \mathcal{L} \big\rangle \ \subset \ \mathbb{R}[p].$$

Το ιδεώδες $I_{\mathcal{L}}$ που προκύπτει ονομάζεται **πλεγματικό ιδεώδες** που αντιστοιχεί στο $\mathcal{L}$.

Μέσω των πλεγματικών ιδεωδών μπορούμε να ορίσουμε και να υπολογίσουμε τις βάσεις Groebner ενός πλέγματος ακεραίων $\mathcal{L}$.

Ορισμός 2.3.2. Έστω $<$ μία διάταξη των μονωνύμων του δακτυλίου $\mathbb{R}[p]$, ένα πλέγμα $\mathcal{L} \subset \mathbb{Z}^k$ και ένα οποιοδήποτε υποπλέγμα $\mathcal{L}'$ του $\mathcal{L}$ με την ιδιότητα $\mathcal{L}' \cap \mathbb{N}^k = \{0\}$. Ονομάζουμε **βάση Groebner** του $\mathcal{L}$ ένα σύνολο $G_< \subset \mathcal{L}'$, τέτοιο ώστε το σύνολο $\{p^{u^+} - p^{u^-} : u \in G_<\}$ να αποτελεί βάση Groebner για το αντίστοιχο πλεγματικό ιδεώδες $I_\mathcal{L}$.

Αντίστοιχα, καλούμε μια βάση Groebner $G_<$ του $\mathcal{L}$ **ανάγωγη** αν το σύνολο $\{p^{u^+} - p^{u^-} : u \in G_<\}$ συνιστά ανάγωγη βάση Groebner για το $I_\mathcal{L}$.

Η ένωση των ανάγωγων βάσεων Groebner $G_<$ του $\mathcal{L}$, που προκύπτουν για κάθε δυνατή διάταξη $<$ των στοιχείων του $I_\mathcal{L}$, ονομάζεται **καθολική βάση Groebner** του $\mathcal{L}$.

Οι βάσεις Graver ενός πλέγματος ακεραίων ορίζονται επίσης ανάλογα με τις βάσεις Graver των διωνυμικών ιδεωδών.

Ορισμός 2.3.3. Έστω $\mathcal{L}$ ένα πλέγμα ακεραίων και $\mathcal{L}'$ ένα οποιοδήποτε υποπλέγμα του $\mathcal{L}$ με την ιδιότητα $\mathcal{L}' \cap \mathbb{N}^k = \{0\}$. Ονομάζουμε **βάση Graver** του $\mathcal{L}$ ένα σύνολο $\mathrm{Gr}_\mathcal{L} \subset \mathcal{L}'$, τέτοιο ώστε το σύνολο $\{p^{u^+} - p^{u^-} : u \in \mathrm{Gr}_\mathcal{L}\}$ να αποτελεί βάση Graver για το αντίστοιχο πλεγματικό ιδεώδες $I_\mathcal{L}$.

Συνοψίζοντας, οι βάσεις που ορίστηκαν για τα διωνυμικά ιδεώδη είναι σε αντιστοιχία με τις βάσεις των ακέραιων πλεγμάτων και ισχύει ο εγκλεισμός:

$$\text{πλεγματική βάση} \subset \text{ελαχιστοτική Μαρκοβιανή βάση} \subset \text{ανάγωγη βάση Groebner}$$

$$\subset \text{καθολική βάση Groebner} \subset \text{βάση Graver}$$

Ακολουθεί ένα παράδειγμα για τις διάφορες βάσεις ενός πλέγματος ακεραίων, όπως αυτές υπολογίζονται στο (1) (Example 1.3.1), που δείχνει ότι ο παραπάνω εγκλεισμός μπορεί να είναι αυστηρός.

Παράδειγμα 2.3.4. Θεωρούμε το πλέγμα ακεραίων

$$\mathcal{L} = \Big\{(u_1, u_2, u_3, u_4) \in \mathbb{Z}^4 : 3u_1 + 3u_2 + 4u_3 + 5u_4 = 0\Big\}.$$

Μία πλεγματική βάση για το $\mathcal{L}$ είναι το σύνολο

$$\mathcal{B}_1 = \{(1, -1, 0, 0), (0, 1, -2, 1), (0, 3, -1, -1)\}.$$

Επομένως, η βαθμίδα του $\mathcal{L}$ είναι 3 και $\mathcal{L} \simeq \mathbb{Z}^3$.
Μία ελαχιστοτική Μαρκοβιανή βάση του $\mathcal{L}$ που περιέχει την πλεγματική βάση $\mathcal{B}_1$ είναι το σύνολο

$$\mathcal{B}_2 = \mathcal{B}_1 \cup \{(0, 2, 1, -2)\}.$$

Όλες, λοιπόν, οι ελαχιστοτικές Μαρκοβιανές βάσεις του $\mathcal{L}$ έχουν διάσταση 4. Μία ανάγωγη βάση Groebner που περιέχει τις παραπάνω βάσεις είναι το σύνολο

$$\mathcal{B}_3 = \mathcal{B}_2 \cup \{(0,1,3,-3),(0,0,5,-4)\}.$$

Υπάρχουν πολλές ανάγωγες βάσεις Groebner που δεν είναι απαραίτητο να αποτελούνται κι αυτές από 6 στοιχεία. Το σύνολο αυτών συνιστά την καθολική βάση Groebner

$$\mathcal{B}_4 = \mathcal{B}_3 \cup \{(1,0,-2,1),(3,0,-1,-1),(2,0,1,-2),(1,0,3,-3),(0,4,-3,0),$$

$$(4,0,-3,0),(0,5,0,-3),(5,0,0,-3)\}.$$

Η καθολική βάση Groebner του $\mathcal{L}$ είναι μοναδική. Το ίδιο ισχύει και για τη βάση Graver, που στην προκειμένη περίπτωση είναι το σύνολο

$$\mathrm{Gr}_{\mathcal{L}} = \mathcal{B}_4 \cup \{(1,1,1,-2),(1,2,-1,-1),(2,1,-1,-1),(1,3,-3,0),$$

$$(2,2,-3,0),(3,1,-3,0),(1,4,0,-3),(2,3,0,-3),(3,2,0,-3),(4,1,0,-3)\},$$

όπου για κάθε στοιχείο συμπεριλαμβάνεται και το αντίθετό του.

Το επόμενο θεώρημα διατυπώθηκε και αποδείχθηκε από τους Diaconis και Sturmfels στο (3) και συντέλεσε στη δημιουργία του κλάδου της Αλγεβρικής Στατιστικής.

Θεώρημα 2.3.5. *(Θεμελιώδες Θεώρημα των Μαρκοβιανών Βάσεων). Ένα υποσύνολο $\mathcal{B}$ ενός πλέγματος ακεραίων $\mathcal{L}$ συνιστά μια Μαρκοβιανή βάση για το $\mathcal{L}$ αν και μόνο αν το αντίστοιχο σύνολο διωνύμων $\left\{p^{b^+} - p^{b^-} \mid b \in \mathcal{B}\right\}$ παράγει το πλεγματικό ιδεώδες $\mathcal{I}_{\mathcal{L}}$.*

Άμεση συνέπεια αυτού είναι η παρακάτω πρόταση.

Πρόταση 2.3.6. *Μία Μαρκοβιανή βάση ενός πλέγματος ακεραίων $\mathcal{L}$ είναι ελαχιστοτική αν και μόνο αν το αντίστοιχο σύνολο διωνύμων $\left\{p^{b^+} - p^{b^-} \mid b \in \mathcal{B}\right\}$ αποτελεί ελαχιστοτικό σύνολο γεννητόρων για το πλεγματικό ιδεώδες $\mathcal{I}_{\mathcal{L}}$.*

Απόδειξη. Έστω $\mathcal{B}$ μια ελαχιστοτική Μαρκοβιανή βάση του $\mathcal{L}$ και $\mathcal{I}_{\mathcal{B}} = \left\{p^{b^+} - p^{b^-} \mid b \in \mathcal{B}\right\}$ το αντίστοιχο σύνολο διωνύμων για το $\mathcal{B}$. Σύμφωνα με το προηγούμενο Θεώρημα, το $\mathcal{I}_{\mathcal{B}}$ είναι ένα σύνολο γεννητόρων του $\mathcal{I}_{\mathcal{L}}$. Έστω $\mathcal{I}_{\mathcal{B}'}$ υποσύνολο του $\mathcal{I}_{\mathcal{B}}$ που είναι και αυτό σύνολο γεννητόρων του $\mathcal{I}_{\mathcal{L}}$. Το σύνολο $\mathcal{B}' = \{b \in \mathcal{B} \mid \ p^{b^+} - p^{b^-} \in \mathcal{B}'\}$ έχει ως αντίστοιχο σύνολο διωνύμων $\left\{p^{b^+} - p^{b'^-} \mid b' \in \mathcal{B}'\right\}$ το σύνολο $\mathcal{I}_{\mathcal{B}'}$, άρα, σύμφωνα το προηγούμενο Θεώρημα, είναι μια Μαρκοβιανή βάση για το $\mathcal{L}$. Από την υπόθεση ότι η Μαρκοβιανή βάση $\mathcal{B}$ είναι ελαχιστοτική και το γεγονός ότι $\mathcal{B}' \subset \mathcal{B}$ συμπεραίνουμε ότι $\mathcal{B}' = \mathcal{B}$ και άρα $\mathcal{I}_{\mathcal{B}'} = \mathcal{I}_{\mathcal{B}}$. Καθώς το σύνολο $\mathcal{I}_{\mathcal{B}'}$ είναι ένα τυχαίο υποσύνολο του $\mathcal{I}_{\mathcal{B}}$, κανένα γνήσιο υποσύνολο του $\mathcal{I}_{\mathcal{B}}$ δεν παράγει το πλεγματικό ιδεώδες $\mathcal{I}_{\mathcal{L}}$. Επομένως, το σύνολο $\mathcal{I}_{\mathcal{B}}$ είναι ένα ελαχιστοτικό σύνολο γεννητόρων του $\mathcal{I}_{\mathcal{L}}$.
Όμοια αποδεικνύεται ο αντίστροφος ισχυρισμός. ■

Κεφάλαιο 3

Στοιχεία Στατιστικής

Ένας από τους κυριότερους στόχους της επιστήμης της Στατιστικής είναι η μελέτη και κατανόηση των ιδιοτήτων ενός πληθυσμού μέσω ενός δείγματος, δηλαδή, ενός μέρους αυτού. Τα στοιχεία του δείγματος ονομάζονται παρατηρήσεις. Προκειμένου να μελετηθεί ένας πληθυσμός ως προς ορισμένα χαρακτηριστικά αρκεί να ληφθεί ένα τυχαίο δείγμα και να κατηγοριοποιηθούν οι παρατηρήσεις ως προς αυτά τα χαρακτηριστικά. Για τη μελέτη ενός τέτοιου δείγματος που αντιστοιχεί σε πεπερασμένα χαρακτηριστικά με πεπερασμένες κατηγορίες είθιστε να χτησιμοποιούνται οι λεγόμενοι πίνακες συνάφειας. Στόχος αυτού του κεφαλαίου είναι επίλυση ορισμένων προβλημάτων που προκύπτουν για τους πίνακες συνάφειας.

3.1 Εισαγωγικές Έννοιες

Ας ξεκινήσουμε με μία ανασκόπηση των βασικών εννοιών.
Κάθε διαδικασία που εκτελείται ή παρατηρείται και το αποτέλεσμα είναι τυχαίο, ονομάζεται **πείραμα τύχης**.

Έστω Ω ένας **διακριτός δειγματικός χώρος**, δηλαδή ένα πεπερασμένο σύνολο που περιέχει τα δυνατά αποτελέσματα ενός πειράματος τύχης. **Διακριτή τυχαία μεταβλητή** καλείται μία απεικόνιση X του Ω σε ένα πεπερασμένο υποσύνολο των πραγματικών αριθμών. Γράφουμε $\mathrm{P}(X \in A)$, εννοώντας την πιθανότητα $\mathrm{P}(\{\omega \in \Omega : X(\omega) \in A\})$.

Έστω X διακριτή τυχαία μεταβλητή που απεικονίζει το δειγματικό χώρο Ω στο σύνολο A, το οποίο έχει r στοιχεία. Χωρίς περιορισμό της γενικότητας, μπορούμε να υποθέσουμε ότι $\mathrm{A} = \{1, 2, ..., r\}$. Στο εξής θα συμβολίζουμε ως $[\mathbf{r}] := \{\mathbf{1}, \mathbf{2}, ..., \mathbf{r}\}$ το σύνολο A.

Ορισμός 3.1.1. Έστω X, Y δύο διακριτές τυχαίες μεταβλητές που απεικονίζουν το δειγματικό χώρο στα σύνολα $[r], [c]$, αντίστοιχα. Ορίζουμε ως **από κοινού συνάρτηση πιθανότητας** (joint probability distribution) των X και Y τη συνάρτηση $\mathrm{P} : [r] \times [c] \to \mathbb{R}$, όπου

$$p_{ij} = \mathrm{P}(i, j) := \mathrm{P}(X = i, Y = j),$$

όπου $\mathrm{P}(X=i, Y=j)$ είναι η πιθανότητα η μεταβλητή X να πάρει την τιμή i και η μεταβλητή Y να πάρει την τιμή j, συγχρόνως.
Οι από κοινού συναρτήσεις πιθανότητας ικανοποιούν τις σχέσεις:
1) $p_{ij} \geq 0$, για κάθε $(i,j) \in [r] \times [c]$ και
2) $\sum_{i\in[r]} \sum_{j\in[c]} p_{ij} = 1$.

Οι πιθανότητες

$$p_{i+} := \sum_{j\in[c]} p_{ij} = P(X=i), \qquad i \in [r],$$

$$p_{+j} := \sum_{i\in[r]} p_{ij} = P(Y=j), \qquad j \in [c]$$

καλούνται **περιθώριες πιθανότητες** και εκφράζουν την πιθανότητα η μεταβλητή X να πάρει την τιμή i και την πιθανότητα η μεταβλητή Y να πάρει την τιμή j, αντίστοιχα.

Παράδειγμα 3.1.2. Από μία κανονική τράπουλα 52 φύλλων χωρίς μπαλαντέρ επιλέγουμε τυχαία ένα φύλλο και παρατηρούμε αν το χρώμα του είναι κόκκινο (Κ) ή μαύρο (Μ) και εάν απεικονίζει φιγούρα (Φ) ή αριθμό (Α). Ο δειγματικός χώρος για αυτό το πείραμα είναι το σύνολο Ω={ΚΦ,ΚΑ,ΜΦ,ΜΑ}. Ας θεωρήσουμε τώρα τις διακριτές τυχαίες μεταβλητές $X : \Omega \to \{1,2\}$, $Y : \Omega \to \{1,2\}$, όπου

$$X : \ \mathrm{K\Phi} \to 1, \mathrm{KA} \to 1, \mathrm{M\Phi} \to 2, \mathrm{MA} \to 2,$$
$$Y : \ \mathrm{K\Phi} \to 1, \mathrm{KA} \to 2, \mathrm{M\Phi} \to 1, \mathrm{MA} \to 2.$$

Θα υπολογίσουμε την από κοινού συνάρτηση πιθανότητας και τις περιθώριες πιθανότητες.

Από τα 52 φύλλα της τράπουλας τα 26 είναι κόκκινα και από αυτά τα 6 είναι φιγούρες. Επομένως, η πιθανότητα το φύλλο που τραβήξαμε να είναι κόκκινο και φιγούρα είναι $p_{11} = \frac{6}{52} = \frac{3}{26}$, ενώ η πιθανότητα να είναι κόκκινο και αριθμός είναι $p_{12} = \frac{20}{52} = \frac{5}{13}$. Όμοια βρίσκουμε ότι $p_{21} = \frac{3}{26}$ και $p_{22} = \frac{5}{13}$.
Οι περιθώριες πιθανότητες είναι

$$p_{1+} = p_{11} + p_{12} = \frac{1}{2}, \qquad p_{2+} = p_{21} + p_{22} = \frac{1}{2},$$

$$p_{+1} = p_{11} + p_{21} = \frac{3}{13}, \qquad p_{+2} = p_{12} + p_{22} = \frac{10}{13}$$

και αντιστοιχούν στην πιθανότητα το φύλλο να είναι κόκκινο (p_{1+}), μαύρο (p_{2+}), φιγούρα (p_{+1}) ή αριθμός (p_{+2}).

Παρατηρούμε ότι το χρώμα του φύλλου είναι ανεξάρτητο από το αν απεικονίζει φιγούρα, δηλαδή η πιθανότητα να είναι κόκκινο είναι $\frac{1}{2}$, είτε απεικονίζει φιγούρα, είτε όχι. Το ίδιο συμβαίνει και με την πιθανότητα το φύλλο να είναι φιγούρα. Τέτοιες μεταβλητές ονομάζονται ανεξάρτητες. Συγκεκριμένα,

Ορισμός 3.1.3. δύο τυχαίες μεταβλητές X και Y με πεδίο τιμών $[r]$ και $[c]$, αντίστοιχα, καλούνται **ανεξάρτητες** αν οι από κοινού πιθανότητες p_{ij} παραγοντοποιούνται ως $p_{ij} = p_{i+}p_{+j}$ για κάθε $i \in [r],\ j \in [c]$. Το σύμβολο $X \perp\!\!\!\perp Y$ χρησιμοποιείται για να δηλώσει την ανεξαρτησία των X, Y.

Πρόταση 3.1.4. *Έστω X, Y δύο τυχαίες μεταβλητές με πεδίο τιμών $[r]$ και $[c]$, αντίστοιχα, και $p = (p_{ij})$ ο πίνακας διάστασης $r \times c$, που στη θέση (i, j) έχει την πιθανότητα p_{ij}, για κάθε $i \in [r], j \in [c]$. Οι X και Y είναι ανεξάρτητες αν και μόνον αν ο p έχει βαθμίδα 1.*

Απόδειξη. ($\Rightarrow$) Η παραγοντοποίηση στον Ορισμό 3.1.3 εκφράζει τον πίνακα p ως γινόμενο του διανύσματος στήλης με στοιχεία τις περιθώριες πιθανότητες p_{i+} και του διανύσματος γραμμής με στοιχεία τις πιθανότητες p_{j+}. Δηλαδή,

$$p_{ij} = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1c} \\ p_{21} & p_{22} & \dots & p_{2c} \\ \vdots & \vdots & \vdots & \vdots \\ p_{r1} & p_{r2} & \dots & p_{rc} \end{pmatrix} = \begin{pmatrix} p_{1+}p_{+1} & p_{1+}p_{+2} & \dots & p_{1+}p_{+c} \\ p_{2+}p_{+1} & p_{2+}p_{+2} & \dots & p_{2+}p_{+c} \\ \vdots & \vdots & \vdots & \vdots \\ p_{r+}p_{+1} & p_{r+}p_{+2} & \dots & p_{r+}p_{+c} \end{pmatrix}$$

$$\Rightarrow p_{ij} = \begin{pmatrix} p_{1+} \\ p_{2+} \\ \vdots \\ p_{r+} \end{pmatrix} \begin{pmatrix} p_{+1} & p_{+2} & \dots & p_{+c} \end{pmatrix}.$$

Εφόσον όλες οι γραμμές του p είναι πολλαπλάσια του ίδιου διανύσματος, η βαθμίδα του είναι 1.

($\Leftarrow$) Ας υποθέσουμε ότι ο πίνακας p έχει βαθμίδα 1. Αυτό σημαίνει ότι όλες οι γραμμές του είναι πολλαπλάσια του ίδιου διανύσματος, ειδάλλως 2 από αυτές θα ήταν γραμμικά ανεξάρτητες και η βαθμίδα θα ήταν μεγαλύτερη από 1. Δηλαδή, ο p μπορεί να γραφεί ως $p = ab^T$ για κάποια $a = (a_1, a_2, ..., a_r) \in \mathbb{R}^r$ και $b = (b_1, b_2, ..., b_c) \in \mathbb{R}^c$. Επειδή όλα τα στοιχεία του p είναι μη αρνητικά, μπορούμε να επιλέξουμε τα διανύσματα a και b ώστε να έχουν και αυτά μη αρνητικά στοιχεία. Ας είναι a_+ το άθροισμα των στοιχείων του a και b_+ το άθροισμα των στοιχείων του b. Τότε, έχουμε

$$p_{ij} = a_i b_j,\ \text{για κάθε } i \in [r], j \in [c],$$

$$p_{i+} = a_i b_+,\ \text{για κάθε } i \in [r] \qquad \text{και} \qquad p_{+j} = a_+ b_j,\ \text{για κάθε } j \in [c].$$

Επιπλέον, $\quad a_+ b_+ = (\sum_{i=1}^{r} a_i) b_+ = \sum_{i=1}^{r} a_i b_+ = \sum_{i=1}^{r} p_{i+} = \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} = 1.$

Επομένως, $\qquad p_{ij} = a_i b_j = a_i b_+ a_+ b_j = p_{i+}p_{+j}, \quad$ για κάθε $i \in [r], j \in [c]$.

■

Παρακάτω παρουσιάζεται συνοπτικά η έννοια του συμπλέγματος πιθανοτήτων. Περισσότερα στοιχεία σχετικά με τα συμπλέγματα πιθανοτήτων και τις αλγεβρικές τους ιδιότητες υπάρχουν στο (5) (Κεφάλαιο 2).

Ορισμός 3.1.5. Καλούμε **σύμπλεγμα πιθανοτήτων** (probability simplex) διάστασης n το σύνολο

$$\Delta_n = \Big\{ q = (q_i) \in \mathbb{R}^{n+1} \ : \ q_i \geq 0 \text{ για κάθε } i \in \{1, 2, ..., n+1\} \text{ και } \sum_{i=1}^{n+1} = 1 \Big\}.$$

Ένα υποσύνολο M ενός συμπλέγματος πιθανοτήτων ονομάζεται **στατιστικό μοντέλο** (statistical model).

Κάθε στοιχείο q ενός n-διάστατου στατιστικού μοντέλου M παραπέμπει σε μία κατανομή πιθανοτήτων για ένα σύνολο $n+1$ στοιχείων, όπου η πιθανότητα να συμβεί το i-οστό στοιχείο είναι $q_i,\ i \in \{1, 2, ..., n+1\}$. Από την άλλη, ένα διάνυσμα $(q_1, q_2, ..., q_{n+1})$ αυτής της μορφής, είναι πάντα στοιχείο του Δ_n, καθώς οι πιθανότητες μίας κατανομής πιθανοτήτων ενός συνόλου είναι εξ ορισμού μη αρνητικές και αθροίζουν στο 1. Επομένως, μπορούμε να εργασθούμε χρησιμοποιώντας στατιστικά μοντέλα προκειμένου να μελετήσουμε τις δυνατές κατανομές πιθανοτήτων για ένα σύνολο.

Παράδειγμα 3.1.6. Η από κοινού συνάρτηση πιθανότητας δύο διακριτών τυχαίων μεταβλητών X και Y με πεδία τιμών $[r]$ και $[c]$, αντιστοιχίζει μία πιθανότητα σε κάθε δυνατό αποτέλεσμα για τις X και Y, δηλαδή είναι μία κατανομή πιθανοτήτων για το σύνολο $[r] \times [c]$. Ο πίνακας $p = (p_{ij})$ ανήκει στο $(rc-1)$-διάστατο σύμπλεγμα πιθανοτήτων

$$\Delta_{rc-1} := \Big\{ q = (q_{ij}) \in \mathbb{R}^{r \times c} \ : \ q_{ij} \geq 0 \text{ για κάθε } i \in [r], j \in [c] \text{ και } \sum_{i=1}^{r} \sum_{j=1}^{c} q_{ij} = 1 \Big\}.$$

Ένα στατιστικό μοντέλο για τις X, Y είναι το σύνολο

$$\mathrm{M} = \Big\{ q = (q_{ij}) \in \mathbb{R}^{r \times c} \ : \ q_{ij} = \frac{1}{rc} \text{ για κάθε } i \in [r], j \in [c] \Big\},$$

δηλαδή ο πίνακας που σε μορφή διανύσματος γράφεται ως $q = (\frac{1}{rc}, \frac{1}{rc}, ..., \frac{1}{rc})$ και περιγράφει την κατανομή πιθανοτήτων όταν όλα τα δυνατά αποτελέσματα είναι ισοπίθανα.

Ορισμός 3.1.7. Το σύνολο των δυνατών κατανομών πιθανοτήτων που καθιστούν τις X, Y ανεξάρτητες είναι υποσύνολο του Δ_{rc-1}. Το στατιστικό αυτό μοντέλο ονομάζεται **μοντέλο ανεξαρτησίας** των X και Y και συμβολίζεται ως $M_{X \perp\!\!\!\perp Y}$.

Από την Πρόταση 3.1.4 γνωρίζουμε ότι οι X και Y είναι ανεξάρτητες αν και μόνο αν ο πίνακας $p = (p_{ij})$ της από κοινού συνάρτησης πιθανότητας έχει βαθμίδα 1. Επομένως,

$$M_{X \perp\!\!\!\perp Y} := \big\{ p \ \in \ \Delta_{rc-1} : \ rank(p) = 1 \big\}.$$

Παρατήρηση: Όλες οι 2×2-υποορίζουσες ενός πίνακα p του μοντέλου ανεξαρτησίας $M_{X \perp\!\!\!\perp Y}$ είναι ίσες με μηδέν. Διαφορετικά η βαθμίδα του θα ήταν μεγαλύτερη

από 1 διότι, όπως είναι γνωστό από τη γραμμική άλγεβρα, αν υπάρχει μη μηδενική $n \times n$-υποορίζουσα ενός πίνακα p, τότε η βαθμίδα του είναι μεγαλύτερη ή ίση του n. Συνεπώς, ισχύει ότι

$$p_{ij}p_{kl} - p_{il}p_{kj} = 0 \tag{3.1}$$

για κάθε $1 \leq i < k \leq r$ και $1 \leq j < l \leq c$. Το σύνολο των λύσεων αυτού του συστήματος τετραγωνικών εξισώσεων είναι γνωστό ως πολλαπλότητα Segre (Segre variety) στην αλγεβρική γεωμετρία. Εάν όλες οι πιθανότητες είναι θετικές, από την (3.1) προκύπτει ότι

$$\frac{p_{ij}/p_{il}}{p_{kj}/p_{kl}} = 1. \tag{3.2}$$

Ο λόγος στην (3.2) συναντάται ως λόγος συμπληρωματικών πιθανοτήτων ή κλάσμα λόγου πιθανοτήτων (odds ratio) στη στατιστική βιβλιογραφία.

3.2 Πίνακες Συνάφειας

Συχνά χρειάζεται να κατηγοριοποιήσουμε τις παρατηρήσεις ενός δείγματος ως προς ορισμένα χαρακτηριστικά. Συνήθως χρησιμοποιούμε πίνακες για να παραστήσουμε αυτήν την κατηγοριοποίηση.

Ένας πίνακας που περιέχει το πλήθος των μετρήσεων/παρατηρήσεων που λαμβάνονται από τα δεδομένα κάποιου δείγματος όταν κατηγοριοποιηθούν ως προς ορισμένα διακριτά χαρακτηριστικά, δηλαδή χαρακτηριστικά με πεπερασμένο αριθμό κατηγοριών, ονομάζεται **πίνακας συνάφειας**.

Ας υποθέσουμε ότι διαλέγουμε τυχαία n παρατηρήσεις από τις οποίες προκύπτουν n ανεξάρτητα ζεύγη από διακριτές τυχαίες μεταβλητές

$$\begin{pmatrix} X^{(1)} \\ Y^{(1)} \end{pmatrix}, \begin{pmatrix} X^{(2)} \\ Y^{(2)} \end{pmatrix}, ..., \begin{pmatrix} X^{(n)} \\ Y^{(n)} \end{pmatrix}, \tag{3.3}$$

οι οποίες ακολουθούν την ίδια κατανομή, όπου

$$P(X^{(k)} = i,\ Y^{(k)} = j) = p_{ij},\ \text{για κάθε } i \,\epsilon\, [r],\ j \,\epsilon\, [c],\ k \,\epsilon\, [n].$$

Η διάταξη των ζευγών στην (3.3) δεν παρέχει πληροφορίες για τον πίνακα $p = (p_{ij})$. Έτσι, μπορούμε να συνοψίσουμε τις παρατηρήσεις σε έναν πίνακα συνάφειας

$$U_{ij} = \sum_{k=1}^{n} 1_{\{X^{(k)}=i, Y^{(k)}=j\}}, \qquad i \in [r],\ j \in [c], \tag{3.4}$$

όπως στο επόμενο παράδειγμα.

Παράδειγμα 3.2.1. Στον Πίνακα 1 παρουσιάζεται μία κατηγοριοποίση των 271 φοιτητών του Α.Π.Θ. που εξετάστηκαν το Φεβρουάριο του 2016 στο μάθημα ¨Εισαγωγή στην Άλγεβρα¨. Τα δύο χαρακτηριστικά εδώ είναι το αν πέρασαν ή όχι το μάθημα και το αν ήταν πρωτοετείς ή μεγαλύτερου έτους φοιτητές.

Πίνακας 1

	ΠΡΩΤΟΕΤΕΙΣ		
ΑΠΟΤΕΛΕΣΜΑ	NAI	OXI	$ΣΥΝΟΛΟ$
ΠΕΡΑΣΑΝ	94	69	163
ΚΟΠΗΚΑΝ	39	69	108
ΣΥΝΟΛΟ	133	138	271

Συμβολίζουμε το σύνολο όλων των πινάκων συνάφειας που μπορεί να προκύψουν για συγκεκριμένο μέγεθος δείγματος n με

$$T(n) := \left\{ u \in \mathbb{N}^{r\times c} : \sum_{i=1}^{r}\sum_{j=1}^{c} u_{ij} = n \right\}.$$

Πρόταση 3.2.2. *Ο τυχαίος πίνακας $U = (U_{ij})$ ακολουθεί πολυωνυμική κατανομή. Δηλαδή, για ένα δείγμα μεγέθους n, εάν $u \in T(n)$, τότε*

$$\begin{aligned} P(U = u) &= \binom{n}{u_{11}} \binom{n-u_{11}}{u_{12}} \cdots \binom{u_{rc}}{u_{rc}} \prod_{i=1}^{r}\prod_{j=1}^{c} {p_{ij}}^{u_{ij}} \\ &= \frac{n!}{u_{11}!u_{12}!...u_{rc}!} \prod_{i=1}^{r}\prod_{j=1}^{c} {p_{ij}}^{u_{ij}}. \end{aligned}$$

Απόδειξη. Βλέπε (1), Proposition 1.1.4. ■

Παράδειγμα 3.2.3. Η πιθανότητα να λάβουμε τον πίνακα συνάφειας του Παραδείγματος 3.2.1 από το σύνολο $T(271)$ είναι

$$\frac{271!}{94!69!39!69!} p_{11}^{94} p_{12}^{69} p_{21}^{39} p_{22}^{69},$$

όπου p_{ij} είναι η πιθανότητα κάποιος φοιτητής να περάσει (για $i = 1$) ή όχι (για $i = 2$) το μάθημα στη συγκεκριμένη εξέταση και να είναι πρωτοετής (για $j = 1$) ή μεγαλύτερου έτους φοιτητής (για $j = 2$).

Ένα ενδιαφέρον ερώτημα για αυτόν τον πίνακα είναι το κατά πόσο το να περάσει κάποιος φοιτητής το μάθημα στη συγκεκριμένη εξέταση ήταν ανεξάρτητο από το αν είναι, ή όχι, πρωτοετής.

Τις περισσότερες φορές μπορούμε να αποφανθούμε για τέτοιου είδους ερωτήματα μέσω στατιστικών ελέγχων υποθέσεως ανεξαρτησίας. Οι έλεγχοι υποθέσεων είναι ένα από τα κύρια αντικείμενα μελέτης της επιστήμης της Στατιστικής. Οι βασικοί ορισμοί, όπως και πλήθος παραδειγμάτων υπάρχουν στο (6) (Κεφάλαιο 5).

Έστω U ένας τυχαίος πίνακας συνάφειας για μέγεθος δείγματος n και $p \in \Delta_{rc-1}$, ο αντίστοιχος πίνακας της από κοινού συνάρτησης πιθανότητας. Θεωρούμε το πρόβλημα ελέγχου υποθέσεων

$$H_0 : p \in M_{X \perp\!\!\!\perp Y} \quad , \quad H_1 : p \notin M_{X \perp\!\!\!\perp Y}. \tag{3.5}$$

Εν ολίγοις, προσπαθούμε να αποφασίσουμε αν ο πίνακας συνάφειας U περιέχει πληροφορίες που απορρίπτουν τη μηδενική υπόθεση ή την κάνουν αποδεκτή, πράγμα που σημαίνει ότι η άγνωστη από κοινού κατανομή του p ανήκει στο μοντέλο ανεξαρτησίας $M_{X \perp\!\!\!\perp Y}$. Παρουσιάζουμε δύο συνήθεις προσεγγίσεις αυτού του προβλήματος.

3.3 Χ-τετράγωνο Έλεγχος Ανεξαρτησίας

Αν η H_0 είναι αληθής, τότε $p_{ij} = p_{i+}p_{+j}$ και ο αναμενόμενος αριθμός εμφανίσεων του από κοινού συμβάντος $\{X = i, Y = j\}$ είναι $np_{i+}p_{+j}$. Τα δύο σύνολα των περιθώριων πιθανοτήτων μπορούν να εκτιμηθούν από τις αντίστοιχες εμπειρικές αναλογίες

$$\hat{p}_{i+} = \frac{U_{i+}}{n} \quad \text{και} \quad \hat{p}_{+j} = \frac{U_{+j}}{n},$$

όπου το άθροισμα της i γραμμής

$$U_{i+} = \sum_{j=1}^{c} U_{ij}$$

μετράει πόσο συχνά το γεγονός $\{X = i\}$ εμφανίζεται στις μετρήσεις μας και, ομοίως ορισμένο, το άθροισμα των στηλών U_{+j} δηλώνει τις εμφανίσεις του γεγονότος $\{Y = j\}$. Ουσιαστικά εκτιμούμε τις άγνωστες ποσότητες p_{i+} και p_{+j} από τα δειγματικά δεδομένα, κάτω από την υπόθεση ότι οι X, Y είναι ανεξάρτητες και συνεπώς τον αναμενόμενο αριθμό $np_{i+}p_{+j}$ από το $\hat{u}_{ij} := n\hat{p}_{i+}\hat{p}_{+j}$. Το X^2 **-στατιστικό**

$$X^2(U) := \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(U_{ij} - \hat{u}_{ij})^2}{\hat{u}_{ij}} \tag{3.6}$$

συγκρίνει τον αναμενόμενο αριθμό $\hat{u}_{ij}$ με τον παρατηρούμενο αριθμό U_{ij} λαμβάνοντας υπ' όψιν πόσο πιθανό έχουμε εκτιμήσει ότι είναι το κάθε κοινό γεγονός να συμβεί.

Ορισμός 3.3.1. Λέμε ότι ένας πίνακας συνάφειας διάστασης $r \times c$ έχει $(r-1)(c-1)$ **βαθμούς ελευθερίας** και συμβολίζουμε $\mathrm{df} = (r-1)(c-1)$.

Μέσω του X^2-στατιστικού και του πλήθους df των βαθμών ελευθερίας του πίνακα U, υπολογίζεται η λεγόμενη p-τιμή του ελέγχου. Σε κάθε τιμή του df αντιστοιχεί μία X^2-κατανομή (βλ. (14)). Η πιθανότητα που αντιστοιχίζεται στην τιμή του X^2-στατιστικού μέσω αυτής της κατανομής είναι η p-τιμή του ελέγχου. Ο Πίνακας 2 δίνει την τιμή της συνάρτησης πυκνότητας πιθανότητας της X^2-κατανομής, ανάλογα με την p-τιμή του ελέγχου και τους βαθμούς ελευθερίας, όταν οι τελευταίοι κυμαίνονται από 1 εώς 5. Μπορούμε μέσω αυτού να βρούμε ένα άνω και κάτω φράγμα για την p-τιμή του ελέγχου όταν $1 \leq \mathrm{df} \leq 5$ παρατηρώντας που κυμαίνεται το $X^2(U)$.

Πίνακας 2

$df \setminus p-\tau.$	0.995	0.900	0.500	0.100	0.050	.005
1	0.00004	0.01579	0.45494	2.70554	3.84146	7.87944
2	0.01003	0.21072	1.38629	4.60517	5.99146	10.59663
3	0.07172	0.58437	2.36597	6.25139	7.81473	12.83816
4	0.20699	1.06362	3.35669	7.77944	9.48773	14.86026
5	0.41174	1.61031	4.35146	9.23636	11.07050	16.74960

Παράδειγμα 3.3.2. Χρησιμοποιώντας το στατιστικό λογισμικό R για τα δεδομένα του Παραδείγματος 3.2.1 προκύπτει ότι $X^2 = 12.08$. Οι βαθμοί ελευθερίας είναι df$= (r-1)(c-1) = 1$. Παρατηρώντας την πρώτη γραμμή του Πίνακα 2 και λαμβάνοντας υπόψιν το διάγραμμα της X^2-κατανομής για έναν βαθμό ελευθερίας (βλ. (14)) συμπεραίνουμε ότι η p-τιμή του ελέγχου είναι μικρότερη από 0.005, καθώς $X^2 = 12.08 > 7.87944$.

Διαισθητικά, εαν η μηδενική υπόθεση είναι αληθής, αναμένουμε το $X^2(U)$ να έχει μικρή τιμή, αφού το U πρέπει να είναι κοντά στο $\hat{u}$.

Ας είναι $u \in T(n)$ ένας πίνακας συνάφειας που ανήκει στην ίδια ίνα με τον U (στη συγκεκριμένη περίπτωση αυτό σημαίνει ότι τα επιμέρους αθροίσματα των στηλών και των γραμμών των 2 πινάκων ταυτίζονται) και

$$X^2(u) = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(u_{ij} - \hat{u}_{ij})^2}{\hat{u}_{ij}} \tag{3.7}$$

η αντίστοιχη αριθμητική εκτίμηση του X^2-στατιστικού. Θέλουμε να υπολογίσουμε την πιθανότητα η τυχαία μεταβλητή $X^2(U)$ που ορίστηκε στην (3.7) να έχει τιμή μεγαλύτερη ή ίση του $X^2(u)$, δεδομένου ότι η H_0 είναι αληθής. Αυτή η πιθανότητα είναι η **p-τιμή** του ελέγχου. Αν η p-τιμή είναι πολύ μικρή, τότε είναι απίθανη η παρατήρηση ενός πίνακα με p-τιμή του στατιστικού X^2 ίση ή μεγαλύτερη της τιμής του $X^2(u)$, όταν λαμβάνουμε δεδομένα από την κατανομή του μοντέλου ανεξαρτησίας $M_{X\perp Y}$. Επομένως, μία μικρή p-τιμή παρέχει στοιχεία εναντίον της H_0 και οδηγεί στο συμπέρασμα ότι η υπόθεση ότι ο πίνακας p ανήκει στο $M_{X\perp Y}$ είναι λανθασμένη.

Από την άλλη πλευρά, εάν η p-τιμή είναι μεγάλη, τότε ο έλεγχος X^2 δεν αποφαίνεται. Σε αυτήν την περίπτωση, λέμε ότι ο έλεγχος X^2 δεν παρέχει στοιχεία εναντίον της μηδενικής υπόθεσης.

Μπορούμε να αποφασίσουμε πότε μία p-τιμή είναι μεγάλη ή μικρή με τον παρακάτω τρόπο. Έστω ότι η p-τιμή για τα δεδομένα μας είναι πολύ μικρή, ας πούμε 0.003. Τότε, θεωρώντας ότι το μοντέλο που προσδιορίζεται από τη μηδενική υπόθεση H_0 είναι σωστό, η πιθανότητα να παρατηρηθούν δεδομένα όπως αυτά που παρουσιάσαμε είναι 3 στα 1000. Από εδώ προκύπτουν 2 δυνατά συμπεράσματα. Είτε ότι αυτό το σπάνιο αυτό γεγονός με πιθανότητα 0.003 πράγματι πραγματοποιήθηκε, είτε ότι η μηδενική υπόθεση ήταν εσφαλμένη. Το ποιο συμπέρασμα θέλει ο καθένας να υιοθετήσει είναι δική του απόφαση. Ωστόσο, είθισται να απορρίπτεται η μηδενική υπόθεση αν η p-τιμή είναι μικρότερη από ένα κατώφλι($threshold$)

με στάθμη σημαντικότητας από το 0.01 έως το 0.05. Η επιλογή του 0.05 έχει επικρατήσει στην επιστημονική βιβλιογραφία.

Υπάρχουν στατιστικές μέθοδοι και προγράμματα με τα οποία μπορούμε να υπολογίσουμε μία p-τιμή και να πραγματοποιήσουμε τον έλεγχο X^2, όπως το στατιστικό λογισμικό R.

Παράδειγμα 3.3.3. Για τα δεδομένα του Παραδείγματος (3.2.1), όπως υπολογίστηκε στο Παράδειγμα (3.3.2), η p-τιμή του ελέγχου είναι μικρότερη από 0.05. Επομένως, μπορούμε να συμπεράνουμε ότι η υπόθεση ότι ο πίνακας p ανήκει στο μοντέλο ανεξαρτησίας $M_{X \perp\!\!\!\perp Y}$ είναι λανθασμένη.

Σε περίπτωση που ο έλεγχος X^2 δεν επαρκεί για να απαντήσουμε στο ερώτημα της ανεξαρτησίας υπάρχει μία εναλλακτική προσέγγιση για το πρόβλημα ελέγχου (3.5), που παρέχει μεγαλύτερη ακρίβεια.

3.4 Έλεγχος Ακρίβειας του Fisher

Πρόταση 3.4.1. *Έστω $r = c = 2$ και u ένας πίνακας συνάφειας για μέγεθος δείγματος n. Αν $p = (p_{ij}) \in M_{X \perp\!\!\!\perp Y}$, τότε η δεσμευμένη κατανομή της U_{11}, για $U_{1+} = u_{1+}, U_{+1} = u_{+1}$ είναι η υπεργεωμετρική κατανομή $(HypGeo(n, u_{1+}, u_{+1}))$. Δηλαδή,*

$$P(U_{11} = u_{11} \mid U_{1+} = u_{1+}, U_{+1} = u_{+1}) = \frac{\begin{pmatrix} u_{1+} \\ u_{11} \end{pmatrix} \begin{pmatrix} n - u_{1+} \\ u_{+1} - u_{11} \end{pmatrix}}{\begin{pmatrix} n \\ u_{+1} \end{pmatrix}}$$

για $u_{11} \in \{a_1, ..., a_s\}$ και μηδέν σε κάθε άλλη περίπτωση, όπου $a_1 = max(0, u_{1+} + u_{+1} - n)$ είναι η ελάχιστη τιμή που μπορεί να πάρει το στοιχείο U_{11} όταν $U_{1+} = u_{1+}, U_{+1} = u_{+1}$ και $a_s = min(u_{1+}, u_{+1})$ η μέγιστη.

Απόδειξη. Βλέπε (1), Proposition 1.1.8. ■

Ας υποθέσουμε ότι οι τυχαίες μεταβλητές X και Y που χρησιμοποιούνται για την κατηγοριοποίηση των παρατηρήσεων έχουν 2 κατηγορίες έκαστη, δηλαδή $r = c = 2$. Τότε ο πίνακας συνάφειας είναι ένας πίνακας διάστασης 2×2.

Έστω $r = c = 2$ και $u \in T(n)$ ένας πίνακας συνάφειας διάστασης 2×2 για μέγεθος δείγματος n. Σύμφωνα με την Πρόταση 3.4.1, μπορούμε να βασίσουμε την απόρριψη της H_0 στην (3.5) στη (δεσμευμένη) p-τιμή

$$P(X^2(U) \geq X^2(u) \mid U_{1+} = u_{1+}, U_{+1} = u_{+1}). \tag{3.8}$$

Τα παραπάνω παραπέμπουν στο στατιστικό έλεγχο που είναι γνωστός ως **έλεγχος ακρίβειας του Fisher** (βλ.(7)). Ο υπολογισμός της p-τιμής στην (3.8) ανάγεται στο άθροισμα των υπεργεωμετρικών πιθανοτήτων

$$\frac{\begin{pmatrix} u_{1+} \\ u_{11} \end{pmatrix} \begin{pmatrix} n - u_{1+} \\ u_{+1} - u_{11} \end{pmatrix}}{\begin{pmatrix} n \\ u_{+1} \end{pmatrix}},$$

για κάθε τιμή $u_{11} \in \{a_1, ..., a_s\}$, όπου $a_1 = max(0, u_{1+} + u_{+1} - n)$, $a_s = min(u_{1+}, u_{+1})$, με την ιδιότητα το X^2-στατιστικό για τον πίνακα με στοιχεία τα u_{11}, $u_{12} = u_{1+} - u_{11}$, $u_{21} = u_{+1} - u_{11}$, $u_{22} = n - u_{1+} - u_{+1} + u_{11}$ να είναι μεγαλύτερο ή ίσο της αριθμητικής εκτίμησης $X^2(u)$.

Παρατήρηση 3.4.2. Ο έλεγχος ακρίβειας του Fisher μπορεί να βασιστεί σε κριτήρια διαφορετικά του X^2-στατιστικού. Για παράδειγμα, κάποιος θα μπορούσε να συγκρίνει έναν τυχαίο πίνακα U με τον πίνακα παρατηρήσεων u, υπολογίζοντας ποιο από τα U_{11} και u_{11} είναι πιθανότερο να εμφανιστεί κάτω από την υπεργεωμετρική κατανομή σύμφωνα με την Πρόταση 3.4.1. Με αυτόν τον τρόπο πραγματοποιείται ο έλεγχος Fisher στη γλώσσα R μέσω της εντολής command fisher.test(u). Μία συζήτηση για τις διαφορές των 2 κριτηρίων σχετικά με τη σύγκριση του τυχαίου πίνακα U με τον πίνακα δεδομένων u μπορεί να βρεθεί στο (4).

Παράδειγμα 3.4.3. Με χρήση του στατιστικού λογισμικού R για τον πίνακα συνάφειας του Παραδείγματος 3.2.1 προκύπτει ότι η p-τιμή είναι 0.00053, τιμή επαρκώς μικρή για να απορριφθεί η μηδενική υπόθεση. Δηλαδή, εξάγεται και πάλι το συμπέρασμα ότι η υπόθεση πως ο πίνακας p ανήκει στο μοντέλο ανεξαρτησίας $M_{X \perp\!\!\!\perp Y}$ είναι λανθασμένη.

Ο έλεγχος ακρίβειας του Fisher εφαρμόζεται μόνο σε 2×2-πίνακες συνάφειας. Ωστόσο, η κεντρική ιδέα μπορεί να γενικευθεί με τον τρόπο που παρουσιάζεται παρακάτω.

Κεφάλαιο 4

Αλγεβρική Στατιστική

Σε αυτό το Κεφάλαιο αναπτύσσονται περισσότερο οι έννοιες του προηγούμενου Κεφαλαίου, με τη συμβολή των αλγεβρικών στοιχείων που παρουσιάστηκαν στα δύο πρώτα Κεφάλαια.

4.1 Πίνακες Συνάφειας Πολλών Παραγόντων

Έστω $X_1, ..., X_m$ διακριτές τυχαίες μεταβλητές, όπου το X_i παίρνει τιμές στο $[r_i]$, για $i = \{1, 2, ..., m\}$. Έστω $\mathcal{R} := \prod_{i=1}^{m}[r_i] = [r_1] \times [r_2] \times \cdots \times [r_m]$ ένα διατεταγμένο σύνολο με την εξής σχέση διάταξης:

Αν $(i_1, i_2, ..., i_m), (j_1, j_2, ..., j_m) \in \mathcal{R}$ και $(i_1, i_2, ..., i_m) \neq (j_1, j_2, ..., j_m)$, τότε
$(i_1, i_2, ..., i_m) > (j_1, j_2, ..., j_m) \iff$ η πρώτη από αριστερά μη μηδενική συντεταγμένη του διανύσματος $(i_1 - j_1, i_2 - j_2, ..., i_m - j_m)$ είναι θετική.

Ας θεωρήσουμε τώρα το σύνολο μετρήσεων

$$U_i = \sum_{k=1}^{n} 1_{\left\{X_1^{(k)}=i_1,...,X_m^{(k)}=i_m\right\}}, \quad i = (i_1, ..., i_m) \in \mathcal{R}, \tag{4.1}$$

που βασίζονται στο τυχαίο δείγμα μεγέθους n από τα ανεξάρτητα και ισόνομα διανύσματα

$$\begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_m^{(1)} \end{pmatrix}, \begin{pmatrix} X_1^{(2)} \\ \vdots \\ X_m^{(2)} \end{pmatrix}, \ldots, \begin{pmatrix} X_1^{(n)} \\ \vdots \\ X_m^{(n)} \end{pmatrix}.$$

Οι μετρήσεις U_i σχηματίζουν έναν πίνακα πίνακα συνάφειας ως προς m χαρακτηριστικά $U = (U_i)$.

Παρατήρηση 4.1.1. Σε κάθε παρατήρηση που κατηγοριοποιείται αντιστοιχίζεται ένα διάνυσμα του $\mathcal{R}$, ανάλογα με το σε ποια κατηγορία ανήκει ως προς το κάθε

χαρακτηριστικό. Επιπλέον, κάθε διάνυσμα του $\mathcal{R}$ παραπέμπει σε μία συγκεκριμένη κατηγοριοποίηση ως προς το κάθε χαρακτηριστικό. Έτσι, υπάρχει μία ένα προς ένα αντιστοιχία μεταξύ των κελιών του πίνακα συνάφειας που αντιστοιχεί στις μεταβλητές $X_1, ..., X_m$ και των διανυσμάτων του $\mathcal{R}$. Στο εξής θα θεωρούμε τους πίνακες συνάφειας είτε σε μορφή πίνακα, είτε σε μορφή διανύσματος, βασιζόμενοι στην ισοδυναμία μεταξύ του συνόλου των πινάκων διάστασης $(r_1 \times r_2 \times \cdots \times r_m)$ και του συνόλου των διανυσμάτων διάστασης $(r_1 r_2 \cdots r_m \times 1)$, με στοιχεία από τους φυσικούς αριθμούς.

Ορισμός 4.1.2. Συμβολίζουμε με $\#\mathcal{R}$ το πλήθος των στοιχείων του $\mathcal{R}$.

Ορίζουμε ως

$$T(n) = \left\{ u \in \mathbb{N}^{(\#\mathcal{R})} : \sum_{i \in \mathcal{R}} u_i = n \right\}.$$

το σύνολο των πινάκων συνάφειας που μπορεί να προκύψουν για μέγεθος δείγματος n.

Ορίζουμε το $(\#\mathcal{R}-1)$-**διάστατο σύμπλεγμα πιθανοτήτων** ως

$$\Delta_{(\#\mathcal{R}-1)} = \Big\{ q = (q_i) = (q_{i_1 i_2 \ldots i_m}) \in \mathbb{R}^{(\#\mathcal{R}-1)} \ : \ q_i \geq 0, \text{ για κάθε } i = (i_1, i_2, ..., i_m) \in \mathcal{R} \text{ και } \sum_{i \in \mathcal{R}} q_i = 1 \Big\}.$$

Καλούμε **εσωτερικό του** $\Delta_{(\#\mathcal{R}-1)}$ και συμβολίζουμε ως $\mathrm{int}(\Delta_{(\#\mathcal{R}-1)})$ το σύνολο

$$\mathrm{int}(\Delta_{(\#\mathcal{R}-1)}) = \Big\{ q = (q_i) = (q_{i_1 i_2 \ldots i_m}) \in \mathbb{R}^{(\#\mathcal{R}-1)} \ : \ q_i > 0, \text{ για κάθε } i = (i_1, i_2, ..., i_m) \in \mathcal{R} \text{ και } \sum_{i \in \mathcal{R}} q_i = 1 \Big\}.$$

Ορίζουμε ως **από κοινού συνάρτηση πιθανότητας** των $X_1, ..., X_m$ τη συνάρτηση $\mathrm{P} : \mathcal{R} \to \mathbb{R}$, όπου

$$p_i = P(X_1 = i_1, ..., X_m = i_m), \quad \text{για κάθε } \ i = (i_1, i_2 ..., i_m) \in \mathcal{R}.$$

Οι από κοινού συναρτήσεις πιθανότητας ικανοποιούν τις σχέσεις:
1) $p_i \geq 0$, για κάθε $i \in \mathcal{R}$ και
2) $\sum_{i \in \mathcal{R}} p_i = 1$. Από την κοινή συνάρτηση πιθανότητας των $X_1, ..., X_m$ προκύπτει ένας **πίνακας από κοινού πιθανότητας** $p = (p_i \mid i \in \mathcal{R})$ που ανήκει στο σύμπλεγμα $\Delta_{(\#\mathcal{R}-1)}$.

Το σύμπλεγμα $\Delta_{(\#\mathcal{R}-1)}$ αποτελείται από όλες τις κατανομές πιθανότητας των $X_1, ..., X_m$, ενώ το εσωτερικό του από αυτές που είναι αυστηρά θετικές.

Η ακόλουθη κλάση μοντέλων παρέχει μία χρήσιμη γενίκευση του μοντέλου ανεξαρτησίας του Ορισμού 3.1.7.

Έστω $p = (p_1, p_2, ..., p_n)$ ένα διάνυσμα με στοιχεία από τους θετικούς πραγματικούς αριθμούς. Συμβολίζουμε ως log$p$ το λογάριθμο του διανύσματος p ανά συντεταγμένη. Δηλαδή, log p:=(log p_1,log p_2,...,log p_n). Αντίστοιχα, συμβολίζουμε με e^p το διάνυσμα $(e^{p_1}, e^{p_2}, ..., e^{p_n})$.

Ορισμός 4.1.3. Έστω $A \in \mathbb{Z}^{d\times(\#\mathcal{R})}$, όπου $d \in \mathbb{N}$, ένας πίνακας με κοινό άθροισμα στοιχείων σε κάθε στήλη. Το **λογαριθμογραμμικό μοντέλο** (log-linear model) που αντιστοιχεί στον A είναι το σύνολο όλων των θετικών πινάκων πιθανότητας

$$\mathcal{M}_A = \left\{ p = (p_i) \in \text{int}(\Delta_{(\#\mathcal{R}-1)}) : \log p \in \text{rowspan}(A) \right\},$$

όπου rowspan(A) ο γραμμικός χώρος που παράγεται από τις γραμμές του A.

Σημειώνουμε ότι στη βιβλιογραφία συχνά χρησιμοποιείται ο όρος τορικό μοντέλο (toric model) για το $\mathcal{M}_A$.

Πρόταση 4.1.4. *Αν $p \in \mathcal{M}_A$, τότε $p = e^{A^T b}$, για κάποιο $b \in \mathbb{Z}^d$.*

Απόδειξη. Έστω $p \in \mathcal{M}_A$, δηλαδή $\log p \in \text{rowspan}(A)$. Εάν $a_1, a_2, ..., a_d$ είναι τα διανύσματα των γραμμών του πίνακα A και $A_1, A_2, ..., A_{(\#\mathcal{R})}$ τα διανύσματα των στήλων του, τότε υπάρχει $b = (b_1, b_2, ..., b_d) \in \mathbb{Z}^d$, ώστε
log p=(log p_1,log p_2,...,log $p_{(\#\mathcal{R})}$) $= b_1 a_1 + b_2 a_2 + \cdots b_d a_d$. Θέτοντας
$a_i = (a_{i1}, a_{i2}, ..., a_{i(\#\mathcal{R})})$ για κάθε $i \in \{1, 2, ..., d\}$, λαμβάνουμε

$$\begin{aligned}(\log p_1, \log p_2, ..., \log p_{(\#\mathcal{R})}) &= (b_1 a_{11} + b_2 a_{21} + \cdots b_d a_{d1}, b_1 a_{12} + b_2 a_{22} + \cdots b_d a_{d2}, \\ &..., b_1 a_{1(\#\mathcal{R})} + b_2 a_{2(\#\mathcal{R})} + \cdots b_d a_{d(\#\mathcal{R})}). \text{ Άρα, } (p_1, p_2, ..., p_{(\#\mathcal{R})}) = \\ &\left(e^{b_1 a_{11} + b_2 a_{21} + \cdots b_d a_{d1}}, e^{b_1 a_{12} + b_2 a_{22} + \cdots b_d a_{d2}}, ..., e^{b_1 a_{1(\#\mathcal{R})} + b_2 a_{2(\#\mathcal{R})} + \cdots b_d a_{d(\#\mathcal{R})}} \right) \\ &= \left(e^{A_1 b}, e^{A_2 b}, ..., e^{A_{(\#\mathcal{R})} b} \right) = e^{A^T b}.\end{aligned}$$

■

Ορισμός 4.1.5. Έστω $u \in T(n)$ ένας πίνακας συνάφειας για δείγμα μεγέθους n και $A \in \mathbb{Z}^{d\times(\#\mathcal{R})}$, ένας πίνακας που τα στοιχεία κάθε στήλης του αθροίζουν στον ίδιο αριθμό k, όπου $k \neq 0$. Το διάνυσμα Au ονομάζεται **ελάχιστο επαρκές στατιστικό** για το μοντέλο $\mathcal{M}_A$, και το σύνολο πινάκων

$$\mathcal{F}_A(u) := \mathcal{F}(u) = \left\{ v \in \mathbb{N}^{(\#\mathcal{R})} : Av = Au \right\}$$

καλείται **ίνα** (fiber) του πίνακα συνάφειας u στο μοντέλο $\mathcal{M}_A$.

Στην επόμενη πρόταση παρουσιάζεται μία γενίκευση της Πρότασης 3.4.1, η οποία οδήγησε στη γενίκευση του ελέγχου ακρίβειας του Fisher.

Πρόταση 4.1.6. *Αν* $p = e^{A^T b} \in M_A$ *και* $u \in T(n)$, *τότε*

$$P(U = u) = \frac{n!}{\prod_{i \in \mathcal{R}} u_i!} e^{b^T (Au)},$$

και η δεσμευμένη πιθανότητα $P(U = u \mid AU = Au)$ *δεν εξαρτάται από τον πίνακα* p.

Απόδειξη. Γενικεύοντας το συμπέρασμα ότι ο τυχαίος πίνακας $U = (U_{ij})$ ακολουθεί πολυωνυμική κατανομή, έχουμε

$$P(U = u) = \frac{n!}{\prod_{i \in \mathcal{R}} u_i!} \prod_{i \in \mathcal{R}} p_i^{u_i} = \frac{n!}{\prod_{i \in \mathcal{R}} u_i!} \prod_{i \in \mathcal{R}} e^{(A^T b)_i u_i} = \frac{n!}{\prod_{i \in \mathcal{R}} u_i!} e^{b^T (Au)}.$$

Επιπλέον,

$$P(U = u \mid AU = Au) = \frac{P(U = u)}{P(AU = Au)},$$

όπου

$$P(AU = Au) = \sum_{v \in \mathcal{F}(u)} P(U = v)$$

$$= \sum_{v \in \mathcal{F}(u)} \frac{n!}{\prod_{i \in \mathcal{R}} v_i!} e^{b^T (Av)} = n! e^{b^T (Au)} \sum_{v \in \mathcal{F}(u)} \left(\prod_{i \in \mathcal{R}} v_i! \right)^{-1}.$$

Δηλαδή,

$$P(U = u \mid AU = Au) = \frac{1/\left(\prod_{i \in \mathcal{R}} u_i!\right)}{\sum_{v \in \mathcal{F}(u)} 1/\left(\prod_{i \in \mathcal{R}} v_i!\right)}. \tag{4.2}$$

Αυτή η παράσταση είναι ανεξάρτητη από το b και άρα ανεξάρτητη από τον πίνακα p. ■

Ας θεωρήσουμε το πρόβλημα ελέγχου υποθέσεως

$$H_0 : p \in \mathcal{M}_A \quad \text{και} \quad H_1 : p \notin \mathcal{M}_A. \tag{4.3}$$

Βασιζόμενοι στην Πρόταση 4.1.6, μπορούμε να γενικεύσουμε τον έλεγχο ακρίβειας του Fisher υπολογίζοντας την p-τιμή

$$P(X^2(U) \geq X^2(u) \mid AU = Au). \tag{4.4}$$

Εδώ

$$X^2(U) = \sum_{i \in \mathcal{R}} \frac{(U_i - \hat{u}_i)^2}{\hat{u}_i} \tag{4.5}$$

είναι η φυσική γενίκευση του X^2-στατιστικού στην (3.7). Η εκτίμηση του $X^2(U)$ απαιτεί τον υπολογισμό των αναμενόμενων μετρήσεων με βάση το μοντέλο $\hat{u}_i = n\hat{p}_i$, όπου $\hat{p}_i$ είναι οι εκτιμητές μέγιστης πιθανοφάνειας(βλ.(6) ,4.2.2).

Ο ακριβής υπολογισμός της p-τιμής στην (4.4) χρησιμοποιεί το άθροισμα όλων των μη-αρνητικών ακέραιων λύσεων του συστήματος γραμμικών εξισώσεων στον Ορισμό (4.1.5). Πράγματι, η p-τιμή ισούται με

$$\frac{\sum_{v\in\mathcal{F}(u)} 1_{X^2(v)\geq X^2(u)} \Big/ \left(\prod_{i\in\mathcal{R}} u_i!\right)}{\sum_{v\in\mathcal{F}(u)} 1 \Big/ \left(\prod_{i\in\mathcal{R}} v_i!\right)}.$$

Ακόμη και σε μετρίου μεγέθους πίνακες συνάφειας ο ακριβής υπολογισμός του αθροίσματος μπορεί να είναι αρκετά επίπονος. Ωστόσο, η p-τιμή μπορεί να εκτιμηθεί χρησιμοποιώντας ορισμένους αλγορίθμους, όπως ο αλγόριθμος Metropolis-Hastings που θα δούμε στη συνέχεια, για δείγματα πινάκων από τη δεσμευμένη κατανομή του U για $AU = Au$. Για να γίνει αυτό είναι αναγκαία η χρήση Μαρκοβιανών βάσεων για ένα λογαριθμογραμμικό μοντέλο, οι οποίες ορίζονται παρακάτω.

Ορισμός 4.1.7. Έστω A ένας πίνακας ακέραιων αριθμών διάστασης $d \times k$. Ονομάζουμε **ακέραιο πυρήνα** (integer kernel) του A το σύνολο $ker_{\mathbb{Z}}(A) = \{u \in \mathbb{Z}^k : \ Au = 0\}$.

Ο ακέραιος πυρήνας $\ker_{\mathbb{Z}}(A)$ ενός πίνακα $A \in \mathbb{Z}^{d\times(\#\mathcal{R})}$ είναι μια υποομάδα της προσθετικής ομάδας $\mathbb{Z}^{(\#\mathcal{R})}$. Συνεπώς συνιστά πλέγμα ακεραίων, σύμφωνα με τον Ορισμό 2.1.1.

Ορισμός 4.1.8. Έστω $A \in \mathbb{Z}^{d\times(\#\mathcal{R})}$, όπου $d \in \mathbb{N}$, ένας πίνακας με κοινό άθροισμα στοιχείων σε κάθε στήλη και $\mathcal{M}_A$ το λογαριθμογραμμικό μοντέλο που αντιστοιχεί στον A. Ένα πεπερασμένο υποσύνολο $\mathcal{B}$ του ακέραιου πυρήνα του A, $ker_{\mathbb{Z}}(A)$, είναι μία ***Μαρκοβιανή βάση*** του $\mathcal{M}_A$, αν για κάθε πίνακα συνάφειας u και κάθε ζεύγος $v, v' \in \mathcal{F}(u)$ υπάρχει μία ακολουθία $u_1, .., u_L \in \mathcal{B}$, τέτοια ώστε

$$v' = v + \sum_{k=1}^{L} u_k \quad \text{και} \quad v + \sum_{k=1}^{l} u_k \geq 0 \quad \text{για } \ l = 1, ..., L.$$

Τα στοιχεία της Μαρκοβιανής βάσης ονομάζονται **κινήσεις** (moves).

Από τον παραπάνω ορισμό προκύπτει ότι σε μία Μαρκοβιανή βάση ορισμένα στοιχεία μπορεί να περιέχονται με θετικό και αρνητικό πρόσημο. Στα επόμενα, δε θα επικεντρωθούμε στο πρόσημο των στοιχείων μίας Μαρκοβιανής βάσης, αλλά θα **θεωρούμε Μαρκοβιανή βάση** του λογαριθμογραμμικού μοντέλου $\mathcal{M}_A$ οποιοδήποτε σύνολο του οποίου τα στοιχεία μαζί με τα αντίθετά τους συνιστούν μια Μαρκοβιανή βάση για το $\mathcal{M}_A$. (Με τον όρο αντίθετα εννοούμε τα συμμετρικά στοιχεία στην προσθετική ομάδα $ker_{\mathbb{Z}}(A) \subset \mathbb{Z}^{(\#\mathcal{R})}$.) Με αυτή την έννοια, θεωρούμε ότι

Παρατήρηση 4.1.9. ένα σύνολο $\mathcal{B}$ είναι Μαρκοβιανή βάση του λογαριθμογραμμικού μοντέλου $\mathcal{M}_A$ αν και μόνο αν είναι Μαρκοβιανή βάση του πλέγματος $\ker_{\mathbb{Z}}(A)$, δηλαδή ικανοποιεί τον ορισμό 2.2.2.

Στην περίπτωση που $A \in \mathbb{N}^{d\times(\#\mathcal{R})}$ ισχύει ότι $\ker_{\mathbb{Z}}(A)\cap\mathbb{N}^{(\#\mathcal{R})} = \{0\}$. Έτσι, αν πάρουμε έναν πίνακα συνάφειας $u \in \mathbb{N}^{(\#\mathcal{R})}$ και έναν πίνακα μη αρνητικών ακεραίων A, η ίνα του u modulo $\ker_{\mathbb{Z}}(A)$, ή αλλιώς ίνα του u στο λογαριθμογραμμικό μοντέλο $\mathcal{M}_A$, που ορίστηκε ως $\mathcal{F}(u) = \{v \in \mathbb{N}^{(\#\mathcal{R})} : Av = Au\} = \{v \in \mathbb{N}^{(\#\mathcal{R})} : u - v \in \ker_{\mathbb{Z}}(A)\}$, είναι πάντα πεπερασμένο σύνολο. Διαισθητικά, αυτό συμβαίνει διότι τα στοιχεία του πίνακα u είναι φυσικοί αριθμοί και έτσι το κάθε μη μηδενικό στοιχείο της Μαρκοβιανής βάσης, η οποία είναι πεπερασμένο σύνολο, μπορεί να προστεθεί πεπερασμένου πλήθους φορές στον πίνακα u, επειδή περιέχει και στοιχεία μικρότερα του μηδενός.

Το γεγονός ότι η ίνα ενός πίνακα συνάφειας $u \in \mathbb{N}^{(\#\mathcal{R})}$ στο μοντέλο $\mathcal{M}_A$ συμπίπτει με την ίνα του u modulo $\ker_{\mathbb{Z}}(A)$ μας επιτρέπει να εφαρμόσουμε τα αποτελέσματα που έχουμε δει για τα πλέγματα ακεραίων, τις Μαρκοβιανές βάσεις και τις ίνες σε αυτά, στον πυρήνα $\ker_{\mathbb{Z}}(A)$, τις Μαρκοβιανές βάσεις του μοντέλου $\mathcal{M}_A$ και την ίνα $\mathcal{F}(u)$ του πίνακα συνάφειας.
Συμπεραίνουμε, λοιπόν, ότι μέσω μιας Μαρκοβιανής βάσης μπορούμε να πάρουμε δείγμα από την ίνα $\mathcal{F}(u)$. Το αποτέλεσμα αυτό είναι πολύ σημαντικό διότι κάθε στοιχείο της ίνας ανήκει στο σύνολο $T(n)$, δηλαδή στο σύνολο όλων των πινάκων συνάφειας που μπορεί να προκύψουν για το συγκεκριμένο μέγεθος δείγματος n του πίνακα συνάφειας u.

Έστω $\mathcal{B}$ μία Μαρκοβιανή βάση για ένα λογαριθμογραμμικό μοντέλο $\mathcal{M}_A$. Ο επόμενος αλγόριθμος πραγματοποιεί έναν τυχαίο περίπατο σε μια ίνα $\mathcal{F}(u)$.

Αλγόριθμος 1.1.13. (Metropolis-Hastings)
Είσοδος: Ένας πίνακας συνάφειας $u \in T(n)$ και μία Μαρκοβιανή βάση $\mathcal{B}$ για το μοντέλο $\mathcal{M}_A$.
Έξοδος: Μία ακολουθία τιμών για το X^2-στατιστικό $(X^2(v_t))_{t=1}^{\infty}$ για πίνακες v_t στην ίνα $\mathcal{F}(u)$.
Βήμα 1: Αρχική συνθήκη: $v_1 = u$.
Βήμα 2: Για $t = 1, 2, ...$ επαλαμβάνουμε τα παρακάτω βήματα:
(i) Επιλέγουμε μία τυχαία κίνηση $b_t \in \mathcal{B}$.
(ii) Αν $min(v_t + b_t) < 0$, τότε θέτουμε $v_{t+1} = v_t$, διαφορετικά θέτουμε

$$v_{t+1} = \begin{cases} v_t + b_t \\ v_t \end{cases} \quad \text{με πιθανότητα} \quad \begin{cases} q \\ 1-q \end{cases},$$

όπου

$$q = min\left\{1, \frac{P(U = v_t + b_t \mid AU = Au)}{P(U = v_t \mid AU = Au)}\right\}.$$

(iii) Υπολογίζουμε την τιμή $X^2(v_t)$.

Μία σημαντική ιδιότητα του αλγορίθμου Metropolis-Hastings είναι ότι η πιθανότητα q στο Βήμα $2(ii)$ έχει οριστεί ως ο λόγος δύο δεσμευμένων πιθανοτήτων. Έτσι, δεν χρειάζεται να υπολογίσουμε το άθροισμα στον παρονομαστή στην εξίσωση (4.2).

Ο Αλγόριθμος Metropolis-Hastings είναι απλά η πιο βασική μέθοδος για να λάβουμε δειγματικούς πίνακες από μια ίνα. Στο (1) (σελ. 17) υπάρχουν περισσότερες λεπτομέρειες για τον Αλγόριθμο και περιγράφεται ο τρόπος με τον οποίο η

έξοδός του οδηγεί στην εκτέλεση της γενικευμένης μορφής του ελέγχου ακρίβειας του Fisher για το πρόβλημα ελέγχου υποθέσεως 4.3.

4.2 Το μοντέλο ανεξαρτησίας ως λογαριθμογραμμικό μοντέλο

Παρατήρηση 4.2.1. Ενας $r \times c$ πίνακας πιθανοτήτων $p = (p_{ij})$ ανήκει στο μοντέλο αεξαρτησίας $M_{X \perp Y}$ αν και μόνο αν κάθε παράγοντας p_{ij} γράφεται ως γινόμενο των περιθώριων πιθανοτήτων p_{i+} και p_{+j}. Αν ο p έχει μόνο θετικά στοιχεία, τότε

$$\log p_{ij} = \log p_{i+} + \log p_{+j}, \qquad i \in [r],\ j \in [c]. \tag{4.6}$$

Για παράδειγμα, ας υποθέσουμε ότι $r = 2$ και $c = 3$. Ο πίνακας $\log p$ είναι ένας 2×3 πίνακας, αλλά θα τον γράφουμε ως διάνυσμα με έξι συντεταγμένες. Τότε η εξίσωση (4.6) φανερώνει ότι το διάνυσμα $\log p$ ανήκει στο χώρο που παράγεται από τις γραμμές του πίνακα

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix},$$

διότι

$$\log p = (\log p_{11}, \log p_{12}, \log p_{13}, \log p_{21}, \log p_{22}, \log p_{23}) = A^T \begin{pmatrix} p_{1+} \\ p_{2+} \\ p_{+1} \\ p_{+2} \\ p_{+3} \end{pmatrix}.$$

Επιπλέον, η μορφή του πίνακα A σε συνδυασμό με τον περιορισμό τα στοιχεία των πινάκων p του λογαριθμογραμμικού μοντέλου $\mathcal{M}_A$ να αθροίζουν στη μονάδα εξασφαλίζει ότι κάθε στοιχείου του λογαριθμογραμμικού μοντέλου ανήκει στο μοντέλο ανεξαρτησίας. Δηλαδή, το θετικό μέρος του μοντέλου ανεξαρτησίας ισούται με το λογαριθμογραμμικό μοντέλο $\mathcal{M}_A$. Γενικά, ο A είναι ένας $(r+c) \times rc$ πίνακας.

Ας είναι u ένας $r \times c$ πίνακας συνάφειας (τον οποίο θα σκεφτόμαστε σε μορφή διανύσματος όπως προηγουμένως). Ο πίνακας A που αντιπροσωπεύει το μοντέλο ανεξαρτησίας καθορίζεται από την εξίσωση

$$Au = \begin{pmatrix} u_{\cdot +} \\ u_{+ \cdot} \end{pmatrix},$$

όπου $u_{\cdot +}$, $u_{+ \cdot}$ είναι τα διανύσματα των αθροισμάτων των γραμμών και στηλών του

πίνακα u. Στην περίπτωση που r=2 και c=3 η εξίσωση αυτή δίνει

$$Au = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \\ u_{13} \\ u_{21} \\ u_{22} \\ u_{23} \end{pmatrix} = \begin{pmatrix} u_{1+} \\ u_{2+} \\ u_{+1} \\ u_{+2} \\ u_{+3} \end{pmatrix}.$$

Ας είναι e_{ij} ο πίνακας διάστασης $r \times c$ που έχει το στοιχείο 1 στη θέση (i, j) και 0 αλλού. Για ένα διάνυσμα ή πίνακα u, συμβολίζουμε με $|| u ||_1 = \sum_{i=1}^{\mathcal{R}} | u_i |$ την 1-νόρμα του u. Χρησιμοποιώντας αυτούς τους συμβολισμούς μπορούμε να αποδείξουμε την επόμενη Πρόταση.

Πρόταση 4.2.2. *Η μοναδική ελάχιστοτική Μαρκοβιανή βάση για το μοντέλο ανεξαρτησίας $\mathcal{M}_{X \perp Y}$ αποτελείται από τις ακόλουθες $2 \cdot \binom{r}{2}\binom{c}{2}$ κινήσεις, που καθεμία έχει 1-νόρμα 4:*

$$\mathcal{B} = \big\{ \pm (e_{ij} + e_{kl} - e_{il} - e_{kj}) : \ 1 \le i \le k \le r, 1 \le j \le l \le c \big\}.$$

Απόδειξη. Έστω $u \neq w$ δύο πίνακες μη αρνητικών ακεραίων που κάθε στήλη και κάθε γραμμή τους αθροίζει στον ίδιο αριθμό. Αρκεί να δείξουμε ότι υπάρχει ένα στοιχείο $b \in \mathcal{B}$, τέτοιο ώστε $u + b \ge 0$ και $|| u - w ||_1 > || u + b - w ||_1$, διότι αυτό αποδεικνύει ότι μπορούμε να χρησιμοποίησουμε στοιχεία από το B για να φέρουμε κοντά μεταξύ τους τα στοιχεία μιας ίνας. Εφόσον οι u και w δεν είναι ίσοι μεταξύ τους και $Au = Aw$, υπάρχει τουλάχιστον ένα θετικό στοιχείο στον πίνακα $u - w$. Χωρίς περιορισμό της γενικότητας, έστω $u_{11} - w_{11} > 0$. Δεδομένου ότι $u - w \in ker_{\mathbb{Z}} A$, υπάρχει ένα στοιχείο στην πρώτη γραμμή του $u - w$ που είναι αρνητικό, ας πούμε ότι $u_{12} - w_{12} < 0$. Ομοίως, ας θεωρήσουμε ότι το θετικό στοιχείο στη δεύτερη γραμμή του $u - w$ είναι το $u_{22} - w_{22} > 0$. Εάν θέσουμε $b = e_{12} + e_{21} - e_{11} - e_{22}$ προκύπτει το επιθυμητό αποτέλεσμα, δηλαδή ότι $|| u - w ||_1 > || u + b - w ||_1$ και $u + b \ge 0$.

Η Μαρκοβιανή βάση $\mathcal{B}$ είναι ελάχιστοτική επειδή αν κάποιο από τα στοιχεία της παραλειφθεί, η ίνα $fiber(b)$ που περιέχει τα θετικά και αρνητικά τμήματά του θα αποκοπεί και το σύνολο $\big\{ fiber(b) \ : \ b \in \mathcal{B} \big\}$ θα αλλάξει. Η μοναδικότητά της είναι και αυτή συνέπεια του Θεωρήματος 2.2.6.

■

Παρατήρηση 4.2.3. Ας σημειώσουμε ότι για πιο σύνθετα λογαριθμογραμμικά μοντέλα συνηθίζεται η χρήση διαφορετικών μοναδιαίων αναπαραστάσεων (unary representations) για τα στοιχεία της Μαρκοβιανής βάσης. Αρκετά διαδεδομένη είναι η περιγραφή τους σημειώνοντας τους δείκτες των μη μηδενικών στοιχείων τους. Αυτός ο τρόπος γραφής ονομάζεται **συμβολισμός ταμπλό**(tableau).

Για παράδειγμα, ο συμβολισμός ταμπλό για την κίνηση $(e_{ij} + e_{kl}) - (e_{il} + e_{kj})$ μιας Μαρκοβιανής βάσης του μοντέλου ανεξαρτησίας είναι:

$$\begin{bmatrix} i & j \\ k & l \end{bmatrix} - \begin{bmatrix} i & l \\ k & j \end{bmatrix},$$

Στην κίνηση $e_{11} + e_{12} - 2e_{13} - e_{21} - e_{22} + 2e_{23}$ αντιστοιχεί ο συμβολισμός ταμπλό

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 2 & 3 \\ 2 & 3 \end{bmatrix} - \begin{bmatrix} 1 & 3 \\ 1 & 3 \\ 2 & 1 \\ 2 & 2 \end{bmatrix}.$$

Προσέξτε ότι οι δείκτες [1 3] και [2 3] επαναλαμβάνονται 2 φορές, εφόσον τα e_{13} και e_{23} εμφανίζονται με πολλαπλότητα 2 στην κίνηση.

Κεφάλαιο 5

Ανακεφαλαίωση

Συνοψίζοντας τα παραπάνω, είδαμε ότι δεδομένου ενός πίνακα συνάφειας $u \in \mathbb{N}^{(\#\mathcal{R})}$ και ενός πίνακα $A \in \mathbb{Z}^{d\times(\#\mathcal{R})}$ του οποίου οι στήλες αθροίζουν στον ίδιο αριθμό, μπορούμε να ορίσουμε το λογαριθμογραμμικό μοντέλο του A ως $\mathcal{M}_A = \left\{p = (p_i) \in \text{int}(\Delta_{(\#\mathcal{R}-1)}) : \text{logp} \in \text{rowspan}(A)\right\}$, το οποίο μάλιστα ταυτίζεται με το θετικό μέρος του μοντέλου ανεξαρτησίας $\mathcal{M}_{X\perp\!\!\!\perp Y}$ όταν $\mathcal{R} = r\times c$. Με χρήση του αλγορίθμου Metropolis-Hastings και της γενίκευσης του ελέγχου ακρίβειας του Fisher μπορούμε να εξετάσουμε αν ένας πίνακας p ανήκει στο $\mathcal{M}_A$. Συγκεκριμένα, το πρόβλημα ελέγχου υποθέσεως (4.3),

$$H_0 : p \in \mathcal{M}_A \quad \text{και} \quad H_1 : p \notin \mathcal{M}_A,$$

ανάγεται στον υπολογισμό μίας Μαρκοβιανής βάσης για το λογαριθμογραμμικό μοντέλο $\mathcal{M}_A$. Αυτός ο υπολογισμός είναι πάντα εφικτός χάρη στις αλγεβρικές ιδιότητες του ακέραιου πυρήνα $\ker_{\mathbb{Z}}(A) = \left\{u \in \mathbb{Z}^{(\#\mathcal{R})} : Au = 0\right\}$, στον οποίο ανήκουν τα στοιχεία των Μαρκοβιανών βάσεων του $\mathcal{M}_A$.

Ειδικότερα, είδαμε ότι ο πυρήνας $\ker_{\mathbb{Z}}(A)$ είναι ένα πλέγμα ακεραίων. Επομένως, η ζητούμενη Μαρκοβιανή βάση υπάρχει αν και μόνο αν το αντίστοιχο πλεγματικό ιδεώδες $I_{\ker_{\mathbb{Z}}(A)}$ παράγεται από ένα πεπερασμένο σύνολο διωνύμων. Όμως, το $I_{\ker_{\mathbb{Z}}(A)}$ είναι πεπερασμένα παραγόμενο ως διωνυμικό ιδεώδες. Με χρήση προγραμμάτων όπως το 4ti2 ή το CoCoA μπορεί να υπολογιστεί μία βάση του, η οποία μάλιστα δύναται να έχει και επιπλέον επιθυμητές ιδιότητες που θα διευκολύνουν τη διαδικασία, όπως το να είναι ελαχιστοτική.

Παράδειγμα 5.0.1. Ας υποθέσουμε ότι ο παρακάτω πίνακας συνάφειας έχει προκύψει από την καταγραφή της πίεσης 680 ασθενών μετά τη χορήγηση φαρμάκων 3 τύπων, τα οποία ονομάζουμε Φάρμακο 1, Φάρμακο 2, και Placebo. Η πίεση των ασθενών σημείωσε άνοδο, κάθοδο ή παρέμεινε σταθερή.

Πίνακας 2

	ΠΟΡΕΙΑ ΠΙΕΣΗΣ			
ΤΥΠΟΣ	$ΑNΟΔΟΣ$	$ΚΑΘΟΔΟΣ$	$ΣΤΑΘΕΡΗ$	$ΣΥΝΟΛΟ$
ΦΑΡΜΑΚΟ.1	160	180	23	363
ΦΑΡΜΑΚΟ.2	82	91	15	188
PLACEBO	54	60	15	129
ΣΥΝΟΛΟ	296	331	53	680

Ας πάρουμε τον πίνακα $A \in \mathbb{Z}^{6\times 9}$, όπου

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

Τα στοιχεία της κάθε στήλης αθροίζουν στο 2, επομένως μπορούμε να εφαρμόσουμε τη θεωρία που αναπτύχθηκε. Αρχικά, υπολογίζουμε τη βάση Graver του τορικού ιδεώδους $\mathcal{I}_{\mathcal{L}}$ που προκύπτει από το πλέγμα $\mathcal{L} = \ker_{\mathbb{Z}}(A)$. Αποδεικνύεται ότι αυτή είναι το σύνολο

$$\mathcal{G}r = \{p_1p_8 - p_2p_9,\ p_3p_5 - p_4p_8,\ p_1p_3p_5 - p_2p_4p_9, p_1p_4 - p_6p_7,\ p_1p_3p_5 - p_6p_7p_8,$$
$$p_2p_4p_9 - p_6p_7p_8,\ p_2p_3p_5p_9 - p_6p_7p_8^2,\ p_1^2p_3p_5 - p_2p_6p_7p_9,\ p_2p_4^2p_9 - p_3p_5p_6p_7\}.$$

Επομένως, μία Μαρκοβιανή βάση του $\mathcal{L}$ είναι το σύνολο διανυσμάτων

$$\{(1,-1,0,0,0,0,0,1,-1),(0,0,1,-1,1,0,0,-1,0),(1,-1,1,-1,1,0,0,0,-1),$$
$$(1,0,0,1,0,-1,-1,0,0),(1,0,1,0,1,-1,-1,-1,0),(0,1,0,1,0,-1,-1,-1,1),$$
$$(0,1,1,0,1,-1,-1,-2,1),(2,-1,1,0,1,-1,-1,0,-1),(0,1,-1,2,-1,-1,-1,0,1)\}.$$

Από τη βάση Graver μπορούμε να οδηγηθούμε σε μία ελαχιστοτική Μαρκοβιανή βάση του $\mathcal{L}$. Προκύπτει ότι η μοναδική ελαχιστοτική Μαρκοβιανή βάση του $\mathcal{L}$ είναι το σύνολο

$$\mathcal{B} = \{(1,-1,0,0,0,0,0,1,-1),(0,0,1,-1,1,0,0,-1,0),(1,0,0,1,0,-1,-1,0,0)\}.$$

Παρατήρηση 5.0.2. Μπορούμε να γράψουμε τα στοιχεία της βάσης $\mathcal{B} = \{b_1, b_2, b_3\}$ σε μορφή πινάκων όπου

$$b_1 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & -1 \end{pmatrix}, b_2 = \begin{pmatrix} 0 & 0 & 1 \\ -1 & 1 & 0 \\ 0 & -1 & 0 \end{pmatrix}, b_3 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & -1 \\ -1 & 0 & 0 \end{pmatrix}.$$

Κάθε πίνακας συνάφειας που ο αντίστοιχος πίνακας πιθανοτήτων του ανήκει στο μοντέλο ανεξαρτησίας $\mathcal{M}_A$ γράφεται ως άθροισμα του αρχικού πίνακα συνάφειας, εάν θεωρηθεί ως 3×3 πίνακας, και ενός γραμμικού συνδυασμού των πινάκων b_1, b_2, b_3.

Δεδομένου ότι ασχολούμαστε με πίνακες συνάφειας, που είναι πίνακες με στοιχεία από τους φυσικούς αριθμούς, βολεύει να επιλέγουμε τον πίνακα A ώστε να έχει και αυτός φυσικούς αριθμούς ως στοιχεία. Αυτή η συνθήκη εξασφαλίζει ότι το άθροισμα n των στοιχείων ενός πίνακα συνάφειας διατηρείται όταν προσθέτουμε στοιχεία της Μαρκοβιανής βάσης και έτσι ο νέος πίνακας που προκύπτει (αν δεν έχει στοιχεία μικρότερα του μηδενός) ανήκει στο σύνολο $T(n)$.

Παράδειγμα 5.0.3. Ας θεωρήσουμε και πάλι τον πίνακα συνάφειας του προηγούμενου παραδείγματος και τον πίνακα

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

Ο πίνακας A αντιπροσωπεύει το μοντέλο ανεξαρτησίας και είναι ανάλογος αυτού που είδαμε στο παράδειγμα 4.2.1. Αναλυτικότερα,

$$Au = A \begin{pmatrix} u_{11} \\ u_{12} \\ u_{13} \\ u_{21} \\ u_{22} \\ u_{23} \\ u_{31} \\ u_{32} \\ u_{33} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 160 \\ 180 \\ 23 \\ 82 \\ 91 \\ 15 \\ 54 \\ 60 \\ 15 \end{pmatrix} = \begin{pmatrix} 363 \\ 188 \\ 129 \\ 296 \\ 331 \\ 53 \end{pmatrix} = \begin{pmatrix} u_{1+} \\ u_{2+} \\ u_{3+} \\ u_{+1} \\ u_{+2} \\ u_{+3} \end{pmatrix}.$$

Η μοναδική ελαχιστοτική Μαρκοβιανή βάση για το λογαριθμογραμμικό μοντέλο $\mathcal{M}_A$, το οποίο ταυτίζεται με το θετικό μέρος του μοντέλου ανεξαρτησίας είναι το σύνολο $\mathcal{B} = \{(e_{11} + e_{22} - e_{12} - e_{21}), (e_{11} + e_{32} - e_{12} - e_{31}), (e_{11} + e_{23} - e_{13} - e_{21}), (e_{11} + e_{33} - e_{13} - e_{31}), (e_{21} + e_{32} - e_{22} - e_{31}), (e_{21} + e_{33} - e_{23} - e_{31}), (e_{12} + e_{23} - e_{13} - e_{22}), (e_{12} + e_{33} - e_{13} - e_{32}), (e_{22} + e_{33} - e_{23} - e_{32})\}$ (αν δε λάβουμε υπόψη το πρόσημο των στοιχείων). Δηλαδή, η Μαρκοβιανή βάση είναι το σύνολο

$$\mathcal{B} = \{B_1, B_2, ...B_9\},$$

όπου

$$B_1 = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, B_2 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 0 \\ -1 & 1 & 0 \end{pmatrix}, B_3 = \begin{pmatrix} 1 & 0 & -1 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},$$
$$B_4 = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}, B_5 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & -1 & 0 \\ -1 & 1 & 0 \end{pmatrix}, B_6 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & -1 \\ -1 & 0 & 1 \end{pmatrix}$$
$$B_7 = \begin{pmatrix} 0 & 1 & -1 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{pmatrix}, B_8 = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & -1 & 1 \end{pmatrix}, B_9 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

Για να πραγματοποιήσουμε τον γενικευμένο έλεγχο ανεξαρτησίας Fisher για το πρόβλημα ελέγχου υποθέσεως

$$H_0 : p \in \mathcal{M}_A \quad \text{και} \quad H_1 : p \notin \mathcal{M}_A,$$

που εξετάζει αν η πορεία των ασθενών είναι ανεξάρτητη από το φάρμακο που πήραν, αρκεί να εκτελέσουμε τη διαδικασία που περιγράφηκε, χρησιμοποιώντας ως είσοδο στον αλγόριθμο Metropolis-Hastings τη Μαρκοβιανή βάση $\mathcal{B}$.

Ευρετήριο Όρων

Βιβλιογραφία

[1] M.Drton, B.Sturmfels and S.Sullivant, Lectures on Algebraic Statistics (2008), [online] Available: https://math.berkeley.edu/ bernd/owl.pdf

[2] A.Bogdanov, Optimal Control of a Double Inverted Pendulum on a Cart. Technical Report

[3] P.Diaconis and B.Sturmfels, Algebraic algorithms for sampling from conditional distributions, Ann. Statist. 26 (1998), no. 1, 363–397

[4] L.J.Davis, Exact tests for 2 × 2 contingency tables, The American Statistician 40 (1986), no. 2, 139–141.

[5] S.Boyd and L.Vandenberghe, Convex Optimization. Cambridge university press, 2004.

[6] Φ.Κολυβά-Μαχαίρα και Ε.Μπόρα-Σέντα, Στατιστική: Θεωρία-Εφαρμογές. Εκδόσεις Ζήτη (2013), $2^{η}$ Έκδοση.

[7] J.V.Freeman and M.J.Campbell, THE ANALYSIS OF CATEGORICAL DATA: FISHER'S EXACT TEST [online] Available: https://www.sheffield.ac.uk/polopolyfs/1.43998!/file/tutorial-9-fishers.pdf

[8] Α.Πλιάτσικα, Μεταπτυχιακή Διατριβή: Πολυπλοκότητα Βάσεων Markov και Graver, Τμήμα Μαθηματικών, Πανεπιστήμιο Ιωαννίνων, 2013

[9] Χ.Χαραλάμπους, Μία Εισαγωγή στην Αντιμεταθετική Άλγεβρα [online] Διαθέσιμο: http://users.auth.gr/ hara/Books-Notes/commutative.pdf

[10] Α.Θωμά, Εισαγωγή στην Υπολογιστική Άλγεβρα: Βάσεις Groebner [online] Διαθέσιμο: https://sites.google.com/site/apostolosthomamath/teaching/-groebner

[11] Μ.Μοιρασγεντή, Διπλωματική Εργασία: Υπολογιστική Άλγεβρα και Συνδυαστική Τορικών Ιδεωδών, Τμήμα Μαθηματικών, Α.Π.Θ., 2017

[12] R.Hemmecke, Computation of Hilbert bases and Graver bases (2005) [online] Available *http* : *//people.math.sfu.ca/ tamon/Seminar/IMO/slides*$_0$51124*.pdf*

[13] J.Hoffstein, J.Pipher and J.H.Silvermanfile, An Introduction to Mathematical Cryptography, Springer, 2008

[14] Statistics How To, WordPress [online] Available http://www.statisticshowto.com/probability-and-statistics/chi-square/

[15] Συναρτήσεις Τυχαίων Μεταβλητών [online], http://www2.stat-athens.aueb.gr/ jpan/statistiki-skepsi-II/chapter6.pdf

[16] Χ.Τατάκης, Διδακτορική Διατριβή: Τορικά Ιδεώδη και Θεωρία Γραφημάτων στη Συνδυαστική Μεταθετική Άλγεβρα, Πανεπιστήμιο Ιωαννίνων, 2011