ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΑΚΡΙΒΕΙΣ ΚΑΙ ΠΡΟΣΕΓΓΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΓΙΑ ΤΗ ΜΕΛΕΤΗ ΣΥΣΤΗΜΑΤΩΝ ΑΞΙΟΠΙΣΤΙΑΣ ΚΑΙ ΠΡΟΒΛΗΜΑΤΩΝ ΕΛΕΓΧΟΥ ΠΟΙΟΤΗΤΑΣ

Φώτιος Σ. Μηλιένος

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ
Υποβλήθηκε στο
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης
του Πανεπιστήμιου Πειραιώς

Πειραιάς
Μάρτιος 2009

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS

AND INSURANCE SCIENCE

EXACT AND APPROXIMATE METHODS FOR THE STUDY OF RELIABILITY SYSTEMS AND QUALITY CONTROL PROBLEMS

Fotios S. Milienos

PhD Thesis

Submitted to

Department of Statistics and Insurance Science of

the University of Piraeus

Piraeus

March 2009

Στους γονείς μου
Στέργο και Δέσποινα,
και στην αδερφή μου Παρασκευή

Ευχαριστίες

Αισθάνομαι την ανάγκη να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή Μάρκο Β. Κούτρα, ο οποίος με τη στάση του και την πολύπλευρη βοήθειά του, συνέβαλε καθοριστικά στο να ολοκληρωθεί η συγκεκριμένη διατριβή. Η στήριξή του, η έμπειρη και αποτελεσματική καθοδήγησή του, αποτελούσε τη λύση για όλα τα προβλήματα που ανέκυπταν. Μπορούσε με ιδανικό τρόπο, να με εισάγει σε άγνωστες έννοιες και πεδία, και ήταν πάντα πρόθυμος, να επιδείξει την απαιτούμενη εμπιστοσύνη και υπομονή, προς εμένα.

Θα ήθελα επίσης να ευχαριστήσω ειλικρινά τον επίκουρο καθηγητή Μιχαήλ Β. Μπούτσικα, για την άψογη συνεργασία μας, και την άμεση και εποικοδομητική ανταπόκρισή του, κάθε φορά που χρειαζόταν. Ακόμη, θα ήθελα να εκφράσω τις ευχαριστίες μου στον καθηγητή και μέλος της τριμελούς συμβουλευτικής επιτροπής Βασίλειο Ζησιμόπουλο, ο οποίος με συνεχή ενθάρρυνση και περισσή προθυμία, προσέφερε τη βοήθειά του.

Οφείλω να ευχαριστήσω τους γονείς μου και την αδερφή μου, για την αμέριστη συμπαράστασή τους όλα αυτά τα χρόνια. Ήταν παρόντες σε οποιαδήποτε ευχάριστη ή δυσάρεστη στιγμή, δίνοντάς μου το κίνητρο και τη δύναμη, που πάντα χρειάζεσαι.

Τέλος, να ευχαριστήσω και το Ίδρυμα Κρατικών Υποτροφιών για την οικονομική βοήθεια που μου παρείχε, κατά τη διάρκεια εκπόνησης της διδακτορικής μου διατριβής.

Περίληψη

Στη διατριβή αυτή μελετώνται προβλήματα που σχετίζονται με τη θεωρία αξιοπιστίας και το στατιστικό έλεγχο ποιότητας. Συγκεκριμένα, εξετάζονται οι ιδιότητες τυχαίων μεταβλητών που σχετίζονται με συστήματα αξιοπιστίας μονάδων με πολλαπλά επίπεδα αποτυχίας, με στατιστικές συναρτήσεις σάρωσης και με πίνακες πλήρους κάλυψης. Τα αποτελέσματα που εξάγονται, αφορούν την προσέγγιση ή την εύρεση της ακριβούς κατανομής, των υπό μελέτη τυχαίων μεταβλητών, και παράλληλα, παρουσιάζονται εφαρμογές σε διάφορα επιστημονικά πεδία. Βασικά εργαλεία στη μελέτη μας, είναι η θεωρία των στοχαστικών διατάξεων, η προσέγγιση κατανομών από μια σύνθετη ή απλή κατανομή Poisson και οι τεχνικές της εμφύτευσης τυχαίων μεταβλητών, σε Μαρκοβιανή αλυσίδα.

Abstract

The present Phd thesis deals with problems related to reliability theory and statistical quality control. Specifically, we study the properties of random variables which are useful for investigating multiple failure mode reliability systems, scan statistics and covering arrays. The new results established in this study, concern the approximation or the exact evaluation of the distribution, of the aforementioned random variables; also, several applications of the theoretical results in various fields are presented. The main tools exploited in this study are fetched from the theory of stochastic ordering, the Poisson approximation theory and the Markov chain embedding method.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Πρόλογος

Μέσα από τη μελέτη της βιβλιογραφίας των τελευταίων δεκαετιών, και όχι μόνο, γίνεται αντιληπτό ότι πάρα πολλά προβλήματα, στα οποία η θεωρία πιθανοτήτων καλείται να δώσει απαντήσεις, σχετίζονται με την εξέταση των χαρακτηριστικών μιας ακολουθίας αποτελεσμάτων, η οποία σχηματίζεται από διακριτές ή συνεχείς τυχαίες μεταβλητές (τ.μ.), είτε αυτές είναι πολυδιάστατες είτε μονοδιάστατες (εξαρτημένες ή μη). Χαρακτηριστικό παράδειγμα αποτελεί η περίπτωση των τ.μ. που αναφέρονται στην εμφάνιση ροών από όμοια αποτελέσματα (runs), σε μια ακολουθία διακριτών τ.μ., η οποία έχει τις ρίζες της στη μελέτη του de Moivre το 1738.

Κάτω απ' αυτό το πρίσμα μπορούν να υπαχθούν και τα συστήματα αξιοπιστίας (reliability systems). Με τον όρο συστήματα αξιοπιστίας εννοούμε σύνολα από μονάδες (για παράδειγμα, υπολογιστές, δέκτες/πομποί, μηχανές κ.α.) οι οποίες είναι συνδεδεμένες με τέτοιο τρόπο, ώστε να φέρνουν εις πέρας μια συγκεκριμένη διεργασία (βλ. για παράδειγμα, Barlow and Proschan (1981)). Η κατάσταση στην οποία βρίσκεται κάθε μονάδα, αλλά και ολόκληρο το σύστημα, εκφράζεται μέσα από τ.μ., οι οποίες μπορεί να είναι δίτιμες ή πλειότιμες. Συγκεκριμένα σύνολα από μονάδες, καθορίζουν πλήρως την κατάσταση στην οποία θα βρίσκεται ολόκληρο το σύστημα, και χωρίς περιορισμό της γενικότητας, μπορούμε να θεωρήσουμε ότι οι καταστάσεις των μονάδων θα εκφράζονται μέσω μιας ακολουθίας τ.μ. (δυνητικά εξαρτημένων). Παραδείγματα συστημάτων, τα οποία έχουν εκτενώς μελετηθεί στη βιβλιογραφία είναι τα συνεχόμενα-k-από-τα-n (consecutive-k-out-of-n, βλ. Chao et al (1995)), όπως και η γενίκευση τους, τα συνεχόμενα-r-μεταξύ-k-από-τα-n (r-within-consecutive-k-out-of-n, Griffith (1986)), το σύστημα k-από-τα-n (k-out-of-n), το σύστημα σε μορφή γέφυρας ή το consecutively-connected system (βλ. π.χ. Barlow and Proschan (1981) ή Kuo and Zuo (2003)).

Μπορεί να γίνει εύκολα αντιληπτό ότι στα προαναφερθέντα συστήματα, οι στατιστικές συναρτήσεις ροών και σάρωσης (scan statistics), παίζουν πρωταρχικό ρόλο. Επιπλέον, το σύνολο των τ.μ. που σχετίζονται με τα παραπάνω συστήματα, έχουν άμεση σχέση με διαδικασίες που αναδύονται και μέσα από το στατιστικό έλεγχο ποιότητας (Balakrishnan and

Koutras (2002)). Στη δειγματοληψία αποδοχής (acceptance sampling), για παράδειγμα, οι κανόνες ροών χρησιμοποιούνται από πολύ παλιά σε δειγματικά σχέδια (Mosteller (1941), Wolfowitz (1943)), με βάση τα οποία ένας σωρός γίνεται αποδεκτός εάν κατα τον έλεγχο βρεθεί ένα πλήθος από συνεχόμενα, μη ελαττωματικά προϊόντα (Vance and McDonald (1979), Govindaraju and Lai (1999)). Επιπλέον, στα διαγράμματα ελέγχου μιας διεργασίας, κριτήρια βασισμένα σε συναρτήσεις ροών και σάρωσης, προσφέρουν λύσεις (ενδείκνυται η χρήση τους) σε αρκετές περιπτώσεις (Page (1955), Champ and Woodall (1987), Klein (2000), Koutras et al (2006), Rakitzis (2008)). Αξίζει ακόμη ενδεικτικά να αναφέρουμε τις περιοχές των τεστ εκκίνησης (start-up demonstration test, Hahn and Gage (1983), Viveros and Balakrishnan (1993), Koutras and Balakrishnan (1999)), των ελέγχων τυχαιότητας (randomness test, Gibbons and Chakraborti (1992), Agin and Godbole (1992), Lou (1997), Koutras and Alexandrou (1997)) αλλά και της σύγκρισης αλυσίδων DNA (Arratia et al (1990), Dembo and Karlin (1992)).

Μια συνήθης πρακτική που συναντάμε και στα δυο επιστημονικά πεδία-της θεωρίας αξιοπιστίας και του ελέγχου ποιότητας-είναι η έκφραση των υπό μελέτη ενδεχομένων, μέσα από δείκτριες (βοηθητικές) τ.μ. Το γεγονός αυτό έχει ως αποτέλεσμα, ο υπολογισμός και η μελέτη των πιθανοτήτων που μας απασχολούν, να μας οδηγούν, στην πλειοψηφία των περιπτώσεων, στην εξέταση αθροισμάτων και γινομένων δίτιμων τ.μ. (εν δυνάμει εξαρτημένων). Επομένως, είναι λογικό να υποθέσει κάποιος ότι γενικές μέθοδοι που αποσκοπούν στην προσέγγιση αλλά, και τον ακριβή υπολογισμό της κατανομής ενός αθροίσματος ή γινομένου τ.μ., να αποτελούν σημαντικό κομμάτι και χρήσιμο εργαλείο στην έρευνα.

Πάνω σ' αυτές τις βάσεις στηρίχθηκε και η συγκεκριμένη διατριβή, καθώς κύριος σκοπός της είναι να μελετήσει τυχαίες μεταβλητές, που σχετίζονται με τη θεωρία των συστημάτων αξιοπιστίας και προβλημάτων ελέγχου ποιότητας. Κύρια εργαλεία στη μελέτη των παραπάνω τ.μ., αποτέλεσαν η θεωρία των στοχαστικών διατάξεων, ανάμεσα σε τυχαία διανύσματα (multivariate stochastic ordering, ενδεικτικά αναφέρουμε το βιβλίο των Muller and Stoyan (2002)), η προσέγγιση της κατανομής ενός αθροίσματος τ.μ. μέσω κατανομής Poisson ή σύνθετης κατανομής Poisson (Poisson or compound Poisson approximations, Arratia et al (1990), Barbour et al (1992), Barbour and Chryssaphinou (2001), Boutsikas and Koutras (2001)), η θεωρία των ακραίων παρατηρήσεων (extreme value theory, βλ. Embrechts et al (1997), Reiss and Thomas (1997)), η μέθοδος της εμφύτευσης τ.μ. σε Μαρκοβιανή αλυσίδα (Fu and Koutras (1994), Koutras and Alexandrou (1995), Fu and Lou (2003), Koutras (2003)) και στοιχεία από τη θεωρία της συνδυαστικής (βλ. Charalambides (2002)).

Πιο συγκεκριμένα, η παρούσα διατριβή μπορεί χωριστεί σε δυο μέρη. Το πρώτο μέ-

ρος, απαρτιζόμενο από τα Κεφάλαια 1 και 2, περιλαμβάνει προσεγγιστικά αποτελέσματα, για συστήματα αξιοπιστίας και για τις στατιστικές συναρτήσεις σάρωσης. Στο δεύτερο μέρος, Κεφάλαια 4 και 5, εξετάζεται μια συγκεκριμένη κλάση τυχαίων πινάκων, και μελετάμε συγκεκριμένες τ.μ. για τις οποίες δίνουμε και ακριβή και προσεγγιστικά αποτελέσματα. Αναλυτικά, το περιεχόμενο των κεφαλαίων συνοψίζεται στα παρακάτω:

Στο πρώτο κεφάλαιο θα ασχοληθούμε με μια συγκεκριμένη κλάση συστημάτων αξιοπιστίας. Τα συστήματα αυτά αποτελούνται από μονάδες, οι οποίες μπορούν να βρίσκονται είτε σε κατάσταση (πλήρους) λειτουργίας είτε να αντιμετωπίζουν ένα από m διαφορετικούς τύπους βλάβης (όπου, $m \geq 1$). Επομένως, οι καταστάσεις των μονάδων εκφράζονται μέσα από ανεξάρτητες διακριτές τ.μ. (υποθέτουμε επιπλέον ότι οι μονάδες λειτουργούν ανεξάρτητα η μια από την άλλη), με πεδίο τιμών το σύνολο $\{0, 1, \ldots, m\}$. Η κατάσταση ολόκληρου του συστήματος (λειτουργία ή μη), καθορίζεται από m συγκεκριμένες οικογένειες συνόλων.

Για τα συστήματα αυτά έχει επικρατήσει η ονομασία *Συστήματα Αξιοπιστίας Μονάδων με Πολλαπλά Επίπεδα Αποτυχίας* (Multiple Failure Mode systems, *MFM*) και έχουν μελετηθεί από τους Barlow et al (1963), Ben-Dov (1980), Satoh et al (1993), Koutras (1997), Boutsikas and Koutras (2002a) κ.α. Αποτελούν γενίκευση των απλών συστημάτων αξιοπιστίας (βλ. Barlow and Proschan (1981)), στα οποία τόσο το σύστημα όσο και οι μονάδες, μπορούν είτε να λειτουργούν είτε να βρίσκονται σε κατάσταση αποτυχίας (δηλαδή, $m = 1$). Στη βιβλιογραφία υπάρχει και μια άλλη κλάση συστημάτων μονάδων, με πολλαπλά επίπεδα αποτυχίας (Συστήματα με Πολλαπλές καταστάσεις, Multistate Systems), η οποία όμως διαφέρει από την προαναφερθείσα, καθώς στη συγκεκριμένη υιοθετείται μια διάταξη ανάμεσα στα διαφορετικά επίπεδα αποτυχίας (Barlow and Wu (1978), Ross (1979) ή το βιβλίο των Kuo and Zuo (2003)). Σ' αυτά οι μονάδες λειτουργούν πλήρως όταν βρίσκονται στην κατάσταση 0, ενώ η μετάβαση τους από την κατάσταση s στην $s + 1$ ($s \in \{0, 1, \ldots, m\}$), υποδηλώνει περαιτέρω μείωση της λειτουργικής τους ικανότητας. Στην κατάσταση m, θεωρούμε ότι έχουμε την «πλήρη» αποτυχία/μη-λειτουργία της μονάδος.

Στο συγκεκριμένο κομμάτι της διατριβής ασχολούμαστε με την προσέγγιση της αξιοπιστίας (της πιθανότητας λειτουργίας) ενός *MFM* συστήματος, όπου προτείνουμε ένα νέο κάτω φράγμα, πολλαπλασιαστικού τύπου. Κύριο εργαλείο στη μελέτη μας είναι η θεωρία των στοχαστικών διατάξεων, ανάμεσα σε τυχαία διανύσματα. Επιπλέον, γίνεται μια αριθμητική μελέτη διαφόρων συστημάτων, τα οποία βρίσκουν εφαρμογή σε διάφορα επιστημονικά πεδία, όπως στον έλεγχο ποιότητας (για παράδειγμα, η γενίκευση των συστημάτων συνεχόμενα-k-από-τα-n, σε περιβάλλον μονάδων με πολλαπλά επίπεδα αποτυχίας, τα οποία σχετίζονται άμεσα και με την εμφάνιση ροών και στατιστικών συναρτήσεων σάρωσης, σε ακολουθίες

πλειότιμων τ.μ.).

Στο Κεφάλαιο 2, μελετάμε αρχικά την κατανομή των στατιστικών συναρτήσεων σάρωσης, οριζόμενων επάνω σε μια ακολουθία από ανεξάρτητες και ισόνομες δίτιμες τ.μ. Συγκεκριμένα, γίνεται μια ανασκόπηση των αποτελεσμάτων της βιβλιογραφίας, που αφορούν την προσέγγιση των παραπάνω κατανομών, μέσω της κατανομής Poisson ή της σύνθετης κατανομής Poisson, με την αρωγή της θεωρίας των αποστάσεων μεταξύ κατανομών, οπού δίδεται ταυτόχρονα και συγκεκριμένο φράγμα, για το σφάλμα της προσέγγισης. Τα φράγματα αυτά βοηθάνε στη διατύπωση ασυμπτωτικών αποτελεσμάτων, τα οποία αποδεικνύονται πολύ χρήσιμα στο μεγάλο εύρος εφαρμογών, που παρουσιάζουν οι συναρτήσεις σάρωσης (στη θεωρία αξιοπιστίας, στον ποιοτικό έλεγχο, στη ασφαλιστική επιστήμη, τη βιολογία κ.α.). Ασχολούμαστε τόσο με τη «κλασική» συνάρτηση σάρωσης, που ορίζεται ως το μέγιστο πλήθος όμοιων συμβόλων, ανάμεσα σε k συνεχόμενες δοκιμές, από τις συνολικά n, όπως και με την απαριθμήτρια τ.μ. των k συνεχόμενων δοκιμών, με τουλάχιστον r όμοια σύμβολα.

Στη συνέχεια, επιθυμώντας να εξετάσουμε την εμφάνιση ακραίων τιμών (δηλαδή, τιμών που υπερβαίνουν ένα «κατώφλι» u), σε μια ακολουθία από ανεξάρτητες και ισόνομες συνεχείς τ.μ., ορίζουμε τις παραπάνω στατιστικές συναρτήσεις σάρωσης, κάτω από ένα γενικότερο μοντέλο (ως ειδική περίπτωση αυτού του μοντέλου, προκύπτουν οι στατιστικές συναρτήσεις σάρωσης, ορισμένες σε μια ακολουθία δοκιμών Bernoulli). Αρχικώς γίνεται προσέγγιση της κατανομής των συναρτήσεων σάρωσης, από μια σύνθετη κατανομή Poisson, και επιπλέον, διατύπωση κάποιων νέων οριακών αποτελεσμάτων, κάτω από την υπόθεση ότι οι συνεχείς τ.μ. ανήκουν σ' ένα από τα μέγιστα πεδία έλξης (maximum domain of attraction) των κατανομών Frechet, Weibull ή Gumbel. Ακόμη, αποδεικνύεται ότι τα προηγούμενα αποτελέσματα έχουν άμεση σχέση με τις κινούμενες διατεταγμένες παρατηρήσεις (moving order statistics, βλ. David and Rogers (1983)) και παρουσιάζονται διάφορα αριθμητικά αποτελέσματα, που φανερώνουν την ποιότητα των προσεγγίσεων.

Στο Κεφάλαιο 3 θα εστιάσουμε το ενδιαφέρον μας σ' ένα πρόβλημα του ποιοτικού ελέγχου, και ειδικότερα, σε διαδικασίες που έχουν άμεσες εφαρμογές (μεταξύ άλλων) στον έλεγχο «λογισμικού» και «υλικού» (software and hardware testing, βλ. π.χ. Dalal and Mallows (1998), Colbourn (2004), Hartman (2006)). Είναι γνωστό μέσα από αρκετές μελέτες ότι η διαδικασία του ελέγχου ενός νέου λογισμικού, αποτελεί ένα από τα σημαντικότερα στάδια στη διαδικασία της παράγωγης, και μια από τις πιο απαιτητικές διεργασίες στον οικονομικό σχεδιασμό (μελέτες υποστηρίζουν ότι η φάση αυτή της παραγωγής, απαιτεί από το 1/3 έως το 1/2 του προϋπολογισμού, Beizer (1990), Carroll (2003a, b)). Ταυτόχρονα,

τα προβλήματα που μπορούν να ανακύψουν από ένα ελλειπή έλεγχο, μπορεί να προκαλέσουν μεγάλη, και ίσως ανεπανόρθωτη οικονομική ζημιά, σε μια εταιρεία (Hartman (2006)).

Για να σκιαγραφήσουμε το πρόβλημα το οποίο μελετάμε, ας θεωρήσουμε ένα σύστημα αποτελούμενο από k μονάδες (π.χ. ένα δίκτυο από υπολογιστές, τερματικά, εκτυπωτές κτλ), όπου για κάθε μια μονάδα, υπάρχουν διαθέσιμα, και μπορούν να χρησιμοποιηθούν, q διαφορετικά λογισμικά. Ας υποθέσουμε ακόμη, ότι κατασκευάζεται ένα νέο λογισμικό (προϊόν) , για ένα συγκεκριμένο είδος μονάδων (για μία συγκεκριμένη διεργασία). Τότε στόχος μας είναι να ελεγχθεί κατά πόσο ομαλά «επικοινωνεί» (συνεργάζεται) το νέο προϊόν, με όλα τα υπόλοιπα διαφορετικά λογισμικά. Επομένως, για να εξετάσουμε όλες τις δυνατές περιπτώσεις ή αλλιώς, να εξετάσουμε όλες τις πιθανές αλληλεπιδράσεις ανάμεσα στα διαφορετικά λογισμικά των k μονάδων (ώστε να εξακριβώσουμε εάν υπάρχει κάποιο πρόβλημα, σε κάποιο συνδυασμό από λογισμικά), θα πρέπει να γίνουν q^k διαφορετικές δοκιμές. Πρόβλημα δημιουργείται όταν το q και το k πάρουν μεγάλες τιμές (όπως συμβαίνει στα περισσότερα προβλήματα), οπότε το σενάριο αυτό καθίσταται πρακτικά μη εφαρμόσιμο, και από πλευράς κόστους, αλλά και χρόνου. Επομένως, ένας αποτελεσματικός τρόπος για να αντιμετωπίσουμε αυτή την κατάσταση, είναι να σχεδιάσουμε μια διαδικασία ελέγχου η όποια θα εξασφαλίζει τον έλεγχο όλων των αλληλεπιδράσεων ανάμεσα σε t (και όχι k) διαφορετικά λογισμικά, όπου $t \leq k$, με το ελάχιστο δυνατό πλήθος δοκιμών (Dalal and Mallows (1998), Hartman (2006)).

Το πρόβλημα αυτό ισοδυναμεί με την κατασκευή ενός $k \times n$ πίνακα, με στοιχεία από ένα αλφάβητο με q γράμματα, ο οποίος θα έχει την ιδιότητα, κάθε ένας από τους $\binom{k}{t}$ υποπίνακές του διάστασης $t \times n$, έχει ως στήλες και τις q^t διαφορετικές λέξεις μήκους t (το n στην περίπτωσή μας, παίζει το ρόλο του πλήθους των δοκιμών που πρέπει να γίνουν). Οι πίνακες με την παραπάνω ιδιότητα ονομάζονται *πίνακες πλήρους κάλυψης* (covering arrays , συμβ. t-CA, βλ. π.χ. Colbourn (2004)). Στο κεφάλαιο αυτό, εισάγουμε και μελετάμε μια νέα κλάση $k \times n$ πινάκων, που έχουν άμεση σχέση με τους t-CA. Οι πίνακες αυτοί έχουν στοιχεία από ένα αλφάβητο με q γράμματα, και χαρακτηρίζονται από την ιδιότητα, κάθε $t \times n$ υποπίνακάς τους, αποτελούμενος από t συνεχόμενες γραμμές (υπάρχουν $k - t + 1$ τέτοιοι υποπίνακες), έχει ως στήλες και τις q^t διαφορετικές λέξεις μήκους t. Τους τελευταίους πίνακες τους ονομάζουμε *πίνακες συνεχόμενης πλήρους κάλυψης* (consecutive covering arrays, συμβ. t-CCA). Στα πλαίσια της παρούσης διατριβής, αναφέρουμε πλεονεκτήματα και μειονεκτήματα των t-CCA, σε σχέση με τους t-CA, καθώς επίσης και διάφορες χρήσιμες εφαρμογές τους. Να σημειώσουμε ότι στους πίνακες πλήρους κάλυψης, οι $t \times n$ υποπίνακες με τους οποίους ασχολούμαστε, αποτελούνται από οποιεσδήποτε t γραμμές, και όχι

αναγκαστικά συνεχόμενες.

Για τη μελέτη των πινάκων t-CCA μεταβαίνουμε από το ντετερμινιστικό περιβάλλον, όπου απευθύνεται η μεγάλη πλειοψηφία των εργασιών στα t-CA, σε στοχαστικό περιβάλλον, θεωρώντας ότι τα στοιχεία του πίνακα είναι ανεξάρτητες και ισόνομες διακριτές τ.μ. Μέσω των τεχνικών εμφύτευσης τ.μ. σε Μαρκοβιανές αλυσίδες και με τη βοήθεια στοιχείων από τη θεωρία της συνδυαστικής, προσδιορίζουμε την ακριβή κατανομή της τ.μ. που εκφράζει το πλήθος των υποπινάκων, από τους οποίους απουσιάζει τουλάχιστον μια λέξη μήκους t. Αρχικά μελετάμε με λεπτομέρεια την περίπτωση όπου τα στοιχεία του πίνακα είναι ανεξάρτητες και ισόνομες δοκιμές Bernoulli (δηλαδή, $q = 2$). Στη συνέχεια, εστιάζουμε την προσοχή μας στα πολλά κοινά στοιχεία που εμφανίζουν οι διαδικασίες που αναφέραμε, με τους πειραματικούς σχεδιασμούς, και παρουσιάζουμε μια εφαρμογή των νέων αποτελεσμάτων.

Στο Κεφάλαιο 4, αρχικά επεκτείνουμε τη μέθοδο που εφαρμόσαμε στο προηγούμενο κεφάλαιο, για την περίπτωση $q > 2$. Έπειτα, διαπιστώνοντας την ανάγκη για εύρεση ποιοτικών και υπολογιστικά εύχρηστων προσεγγίσεων, χρησιμοποιούμε τη θεωρία που αναφέρθηκε και στο Κεφάλαιο 2 (μέθοδος Chen-Stein), ώστε να διατυπώσουμε αντίστοιχα αποτελέσματα, για την υπό μελέτη τ.μ. Τέλος, εισάγουμε και μελετάμε τις ιδιότητες ενός νέου ελέγχου τυχαιότητας (randomness test), που βασίζεται στα προηγούμενα αποτελέσματα.

Κεφάλαιο 1

Συστήματα αξιοπιστίας μονάδων, με πολλαπλά επίπεδα αποτυχίας

Με τον όρο *Συστήματα Αξιοπιστίας*, εννοούμε σύνολα από μονάδες (για παράδειγμα, υπολογιστές, δέκτες/πομποί, μηχανές κ.α.) οι οποίες είναι συνδεδεμένες με τέτοιο τρόπο, ώστε να φέρνουν εις πέρας μια συγκεκριμένη διεργασία. Στο πρώτο κεφάλαιο της παρούσης διατριβής, θα επικεντρωθούμε σε μια συγκεκριμένη κατηγορία συστημάτων αξιοπιστίας, τα οποία συναντάμε στη βιβλιογραφία με την ονομασία, *Συστήματα Αξιοπιστίας Μονάδων με Πολλαπλά Επίπεδα Αποτυχίας* (Multiple Failure Mode systems, *MFM*). Τα συστήματα *MFM* αποτελούνται από μονάδες οι οποίες, μπορούν είτε να λειτουργούν, είτε να βρίσκονται σε κατάσταση αποτυχίας τύπου s, όπου $s \in \{1, 2, \ldots, m\}$ (βλ. Barlow et al (1963), Ben-Dov (1980), Satoh et al (1993), Koutras (1997) και Boutsikas and Koutras (2002a)). Σε κάθε τύπο αποτυχίας, αντιστοιχείται μια οικογένεια συνόλων (από μονάδες) $\mathbf{C}_s$, έτσι ώστε το σύστημα να μη λειτουργεί αν και μόνο αν υπάρχει τουλάχιστον ένα σύνολο, από κάποια οικογένεια $\mathbf{C}_s$, με όλες τις μονάδες του σε κατάσταση αποτυχίας τύπου s ($s \in \{1, 2, \ldots, m\}$).

Είναι φανερό ότι τα παραπάνω συστήματα είναι μια γενίκευση των απλών συστημάτων (Single Failure Mode systems, *SFM*), στα οποία τόσο οι μονάδες, όσο και ολόκληρο το σύστημα, έχουν δυο μόνο δυνατές καταστάσεις (λειτουργίας ή μη λειτουργίας, βλ. π.χ. Barlow and Proschan (1981)). Για κάθε σύστημα *SFM*, προσδιορίζεται μία οικογένεια συνόλων από μονάδες («ελάχιστα σύνολα διακοπής», minimal cut sets), με τέτοιο τρόπο ώστε ολόκληρο το σύστημα να αποτυγχάνει, εάν και μόνο εάν όλες οι μονάδες ενός τουλάχιστον συνόλου, από την παραπάνω οικογένεια, να είναι σε κατάσταση αποτυχίας. Η γενίκευση των απλών συστημάτων, σε συστήματα *MFM*, επιτυγχάνεται εάν σε κάθε τύπο αποτυχίας, αντιστοιχήσουμε μια ξεχωριστή οικογένεια από ελάχιστα σύνολα διακοπής.

Αξίζει να αναφέρουμε, ότι η θεμελίωση της θεωρίας των συστημάτων αξιοπιστίας, ξεκίνησε από τις εργασίες των Moore and Shannon (1956) και Birnbaum et al (1961). Έκτοτε έχει επιδειχθεί μεγάλο ενδιαφέρον για διάφορα ζητήματα, τα οποία προκύπτουν συνεχώς μέσα από πραγματικά προβλήματα, και ως απώτερο στόχο έχουν, την κατασκευή όσο το δυνατόν αξιόπιστων συστημάτων (βλ. π.χ. το εξαιρετικό βιβλίο των Kuo and Zuo (2003)). Βασικό κριτήριο για την αξιολόγηση ενός συστήματος, δεν είναι άλλο από την αξιοπιστία του, δηλαδή, την πιθανότητα λειτουργίας του.

Την επιστημονική κοινότητα έχει απασχολήσει και μια άλλη κλάση συστημάτων, με πολλαπλά επίπεδα αποτυχίας: τα Συστήματα με Πολλαπλές καταστάσεις (Multistate Systems). Η σημαντική διαφορά ανάμεσα στις δυο κλάσεις (που προκαλεί και τη διαφοροποίηση στην ερευνητική δραστηριότητα) έγκειται στο γεγονός ότι στα προαναφερθέντα συστήματα, υιοθετείται μια διάταξη ανάμεσα στα διαφορετικά επίπεδα αποτυχίας (Barlow and Wu (1978), Ross (1979) ή Kuo and Zuo (2003)). Οι μονάδες λειτουργούν πλήρως όταν βρίσκονται στην κατάσταση 0, ενώ η μετάβαση τους από την κατάσταση s στην $s+1$ ($s \in \{0, 1, \ldots, m-1\}$), υποδηλώνει περαιτέρω μείωση της λειτουργικής τους ικανότητας (μέχρι την κατάσταση m, όπου έχουμε την «πλήρη» αποτυχία/μη-λειτουργία της μονάδος). Αξίζει όμως να επισημάνουμε ότι, δεν είναι δυνατόν για κάθε σύστημα αξιοπιστίας μονάδων, με πολλαπλά επίπεδα αποτυχίας, να υποθέσουμε μια διάταξη ανάμεσα στους διαφορετικούς τύπους βλαβών.

Έτσι, η μετάβαση από τα συστήματα *SFM* στα *MFM*, θεωρείται απαραίτητη και χρήσιμη, βλέποντας τις εφαρμογές και τα προβλήματα που μπορεί να εκφραστούν μέσα από μοντέλα *MFM*. Για παράδειγμα, οι βλάβες που μπορεί να παρουσιάσει ένα σύστημα ασφαλείας είναι, είτε αδυναμία να σημάνει ορθά ένα συναγερμό (failure to detect a breakdown), είτε να σημάνει λανθασμένα (false alarm). Σ' ένα σύστημα ελέγχου ροής υγρών, μια βαλβίδα μπορεί να παραμείνει κολλημένη κλειστή (stuck closed) ή κολλημένη ανοιχτή (stuck open) και πολλά άλλα. Στα δύο παραπάνω συστήματα, είναι δύσκολο να διατυπωθεί μια διάταξη ανάμεσα στις βλάβες που παρουσιάζουν (ανάλογης αυτής των συστημάτων με πολλαπλές καταστάσεις).

Συστήματα *MFM* (ή *DFM*, για την περίπτωση $m = 2$) απασχολούν τους συγγραφείς για μισό περίπου αιώνα (ως οι παλαιότερες εργασίες στην περιοχή, θα μπορούσαν να θεωρηθούν αυτές των Moore and Shannon (1956), Barlow and Hunter (1960a,b), Barlow et al (1963)). Αξίζει να αναφέρουμε ότι την ίδια εποχή ξεκίνησε και το μεγάλο ενδιαφέρον για τη θεωρία αξιοπιστίας, και η προσπάθεια πολλών επιστημόνων να εφαρμόσουν αποτελέσματα από τη θεωρία των πιθανοτήτων, στο συγκεκριμένο πεδίο (βλ. σχετικές αναφορές στο βιβλίο των Barlow and Proschan (1981) ή την ιστορική αναδρομή των Rueda and Pawlak (2004)).

Ένα μεγάλο πλήθος από τις εργασίες στα συστήματα *MFM*, όπως αυτές των Moore and Shannon (1956), Barlow and Hunter (1960a,b), Barlow et al (1963), Ben-Dov (1980), Page and Perry (1988), Pham and Malon (1994), Levitin and Lisnianski (2001), Zhang et al (2002), Levitin (2002), ασχολούνται με συστήματα αξιοπιστίας με συγκεκριμένη δομή (π.χ. παράλληλα-σειριακά, parallel-series) και κυρίως *DFM*. Ένα χαρακτηριστικό των συστημάτων που αναφέρονται στις προηγούμενες εργασίες είναι ότι, η προσθήκη νέων μονάδων μπορεί να μειώσει ή να αυξήσει, την πιθανότητα λειτουργίας ολόκληρου του συστήματος. Έτσι, η προσπάθειά τους επικεντρώνεται, στην εύρεση του βέλτιστου αριθμού μονάδων στο σύστημα, αλλά και σε ζητήματα που έχουν να κάνουν με την πολιτική που πρέπει να ακολουθηθεί, σε περίπτωση μη λειτουργίας κάποιας μονάδος (replacement problem).

Όμως, ένα από τα κυριότερα χαρακτηριστικά ενός συστήματος, όπως έχει ήδη αναφερθεί, είναι η αξιοπιστία του. Λογική και πολύ χρήσιμη θα είναι λοιπόν, και η έρευνα επάνω στις μεθόδους για τον υπολογισμό της ακριβούς τιμής ή προσέγγισης, της αξιοπιστίας ενός συστήματος *MFM*. Οι εργασίες που έχουμε στη διάθεση μας, για τον ακριβή υπολογισμό της αξιοπιστίας, αφορούν είτε συγκεκριμένα συστήματα (π.χ. Satoh et al (1993), Barlow and Heidtmann (1984), Koutras (1997)), είτε οποιαδήποτε συστήματα (Dhillon and Rayapati (1986), Boutsikas and Koutras (2002a), Levitin (2003)). Για παράδειγμα, οι Satoh et al (1993) ασχολούνται με τον υπολογισμό της αξιοπιστίας ενός *DFM* συστήματος, με δομή σειριακή-παράλληλη ή παράλληλη-σειριακή. Οι Barlow and Heidtmann (1984) υπολογίζουν την αξιοπιστία ενός k-από-τα-n, *DFM* συστήματος, με μια μέθοδο βασισμένη σε γεννήτριες συναρτήσεις. Στην εργασία Koutras (1997) , εισάγεται ένα νέο σύστημα μονάδων, με δύο είδη αποτυχιών (συνεχόμενο-k-από-τα-n, *DFM*), και δίδεται ένας αναδρομικός τύπος για τον υπολογισμό της αξιοπιστίας του (τον οποίο και θα χρησιμοποιήσουμε στους αριθμητικούς υπολογισμούς, της Παραγράφου 1.4). Από την άλλη, οι Dhillon and Rayapati (1986) προσφέρουν μία μέθοδο για τον υπολογισμό της αξιοπιστίας ενός *DFM* συστήματος, που όμως, η εφαρμογή της περιορίζεται σε συστήματα με πολύ μικρο αριθμό μονάδων (4 ή 5). Στην εργασία Levitin (2003), μελετώνται διάφορα μέτρα αξιοπιστίας για *DFM* συστήματα (με γεννήτριες συναρτήσεις), και εφαρμόζονται σε συγκεκριμένα συστήματα (όπως, η γέφυρα, βλ. και Σχήμα 1.2.1, σελίδα 10). Οι Boutsikas and Koutras (2002a) αναπτύσσουν δυο μεθόδους για τον υπολογισμό της αξιοπιστίας (οποιωνδήποτε) *MFM* συστημάτων: μέσω κατάλληλα ορισμένου απλού συστήματος, ή μέσω ενός αναπτύγματος σε αθροίσματα γινομένων τ.μ. μίας δίτιμης συνάρτησης που εκφράζει τη λειτουργία του συστήματος (για τις δυο τελευταίες μεθόδους θα μιλήσουμε και στην Παράγραφο 1.2).

Εύκολα γίνεται κατανοητό, ότι οι μέθοδοι που χρησιμοποιούνται από τους παραπάνω

συγγραφείς, ποικίλουν, ενώ η ανάγκη για την εύρεση πιο απλών υπολογιστικά διαδικασιών (για τον υπολογισμό ή την προσέγγιση, της αξιοπιστίας), παραμένει σημαντική. Ειδικότερα, σε συστήματα όπου το πλήθος των μονάδων είναι αρκετά μεγάλο (ή το σύστημα έχει πολύπλοκη δομή), ο υπολογισμός της ακριβούς τιμής της αξιοπιστίας (με τις υπάρχουσες τεχνικές), γίνεται πρακτικά αδύνατος (όπως θα διαπιστώσουμε και στις επόμενες παραγράφους). Επομένως, η εύρεση «ποιοτικών» προσεγγίσεων (με τιμές κοντά στην τιμή της αξιοπιστίας), μέσα από απλές διαδικασίες, είναι επιβεβλημένη, και προς αυτή την κατεύθυνση κινείται και η δική μας μελέτη.

Στο παρόν κεφάλαιο, εισάγουμε ένα νέο κάτω φράγμα, πολλαπλασιαστικού τύπου, για τη συνάρτηση αξιοπιστίας ενός *MFM* συστήματος. Το νέο φράγμα, στηρίζεται στην εύρεση της αξιοπιστίας, κατάλληλα ορισμένων απλών συστημάτων, και βελτιώνει ένα από τα φράγματα που είχαν προτείνει οι Boutsikas and Koutras (2002a). Βασικό ρόλο στην απόδειξη των νέων αποτελεσμάτων, παίζει η θεωρία των στοχαστικών διατάξεων, ανάμεσα σε τυχαία διανύσματα (βλ. π.χ. Muller and Stoyan (2002)). Ο τρόπος με τον οποίο χρησιμοποιείται η προαναφερθείσα θεωρία, διαφέρει με τον έως τώρα ρόλο της, σε θέματα που σχετίζονται με τα συστήματα αξιοπιστίας. Στην κλασσική βιβλιογραφία, οι στοχαστικές διατάξεις προσφέρουν (κυρίως) εργαλεία για τη μελέτη ενός συστήματος, σε συνάρτηση με το χρόνο, τη μελέτη για τη διατήρηση διαφόρων ιδιοτήτων γήρανσης κ.α.-δείτε π.χ. Boland et al (1994). Αντίθετα εδώ, η θεωρία των στοχαστικών διατάξεων μας δίνει τη δυνατότητα να μελετήσουμε την αξιοπιστία ενός συστήματος *MFM*, σε μία συγκεκριμένη χρονική στιγμή, όπου η πιθανότητα μια μονάδα να βρίσκεται σε κατάσταση αποτυχίας τύπου s ($s \in \{0, 1, \ldots, m-1\}$), θεωρείται σταθερή.

Στην πρώτη παράγραφο του κεφαλαίου αυτού, παρουσιάζονται κάποια στοιχεία από τη θεωρία των στοχαστικών διατάξεων. Στην επόμενη παράγραφο, εισάγονται όλοι οι απαραίτητοι συμβολισμοί και οι ιδιότητες, ενός συστήματος *MFM*. Έπειτα, προχωρούμε στην απόδειξη των νέων αποτελεσμάτων (Παράγραφος 1.3), ενώ στην τελευταία παράγραφο, εστιάζουμε σε αριθμητικούς υπολογισμούς και συγκρίσεις, ανάμεσα στα φράγματα που έχουν ήδη εμφανιστεί στη βιβλιογραφία.

1.1 Στοχαστικές διατάξεις

Ο στόχος της πρώτης παραγράφου είναι να εισάγει τον αναγνώστη σε βασικές έννοιες της θεωρίας των στοχαστικών διατάξεων, προσφέροντας παράλληλα, όλα εκείνα τα εργαλεία που θα φανούν χρήσιμα στην πορεία. Με το σκεπτικό αυτό, επιχειρείται όχι μια απλή διατύπωση

των εννοιών και των αποτελεσμάτων, αλλά μια πιο λεπτομερής ανάπτυξη. Ειδικότερα, σε αποτελέσματα που παρουσιάζουν γενικότερο ενδιαφέρον, και αναδεικνύουν τόσο το εύρος των εφαρμογών των στοχαστικών διατάξεων (σε πολλά και διαφορά επιστημονικά πεδία), αλλά όσο και τη δυναμική τους, έχουν δοθεί για λόγους πληρότητας, και οι αντίστοιχες αποδείξεις. Για το σύνολο της θεωρίας που υπάρχει στη συγκεκριμένη παράγραφο, ο αναγνώστης μπορεί να ανατρέξει στα βιβλία, π.χ., των Shaked and Shanthikumar (1994) ή Muller and Stoyan (2002).

Οι στοχαστικές διατάξεις ανάμεσα σε τυχαίες μεταβλητές, είναι μερικές διατάξεις, πάνω στο χώρο των συναρτήσεων κατανομής. Στη βιβλιογραφία υπάρχουν πολλά είδη στοχαστικών διατάξεων, όπως η διάταξη λόγου πιθανοφάνειας (likelihood ratio order), η κυρτή διάταξη (convex order), και η διάταξη με βάση τη βαθμίδα αποτυχίας (hazard rate order). Ωστόσο, για τις ανάγκες της ανάπτυξης του παρόντος κεφαλαίου, μας αρκεί να επικεντρωθούμε στη διάταξη που εισάγεται από τον Ορισμό 1.1.1 (τη συνήθη στοχαστική διάταξη, usual stochastic order).

Πριν προχωρήσουμε στη διατύπωση των αποτελεσμάτων, είναι απαραίτητο να αποσαφηνίσουμε τη χρήση ορισμένων βασικών εννοιών. Για δυο διανύσματα $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ και $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, του $\Re^n$ $(n \geq 1)$, θα λέμε ότι είναι $\mathbf{z} \leq \mathbf{y}$, εάν $z_i \leq y_i$, για κάθε $i = 1, 2, \ldots, n$ (όμοια, $\mathbf{z} < \mathbf{y}$ εάν $z_i < y_i$, για κάθε $i = 1, 2, \ldots, n$). Στο εξής, μια συνάρτηση $f : \Re^n \to \Re$ θα καλείται αύξουσα, εάν για κάθε $\mathbf{z} \leq \mathbf{y}$, ισχύει $f(\mathbf{z}) \leq f(\mathbf{y})$ (μη φθίνουσα κατά συντεταγμένη). Ακόμη, ένα σύνολο $D \subseteq \Re^n$ θα λέγεται *άνω σύνολο* (upper set), εάν για κάθε $\mathbf{z} \in D$ και $\mathbf{y} \geq \mathbf{z}$, ισχύει $\mathbf{y} \in D$ (εάν το D είναι Borel μετρήσιμο, τότε είναι άνω σύνολο εάν και μόνο εάν, η δείκτρια συνάρτησή του είναι αύξουσα).

Ορισμός 1.1.1 *Η τ.μ. Z είναι μικρότερη από την τ.μ. Y, με βάση τη συνήθη στοχαστική διάταξη (συμβ., $Z \leq_{st} Y$), εάν για κάθε $u \in \Re$, ισχύει*

$$P(Z > u) \leq P(Y > u),$$

ή ισοδύναμα

$$P(Z \leq u) \geq P(Y \leq u).$$

Η συνήθης στοχαστική διάταξη, έχει εμφανιστεί στη βιβλιογραφία και με άλλες ονομασίες όπως, ισχυρή στοχαστική διάταξη (strong stochastic order, Szekli (1995)), στοχαστική διάταξη πρώτης τάξης (first order stochastic dominance, Quirk and Saposnik (1962)) ή απλά στοχαστική διάταξη. Οι εφαρμογές που παρουσιάζει είναι αρκετές, ενώ αξίζει να αναφέρουμε τη σχέση της με τους χάρτες ποσοστιαίων σημείων (*Q-Q plot*). Συγκεκριμένα, για δυο τ.μ.

Z, Y (με συναρτήσεις κατανομής F_Z, F_Y, αντιστοίχως) ισχύει $Z \leq_{st} Y$, εάν και μόνο εάν το γράφημα της F_Z^{-1} έναντι της F_Y (δηλαδή, τα σημεία $(x, F_Z^{-1}(F_Y(x))))$, βρίσκεται κάτω από τη διχοτόμο των αξόνων (με F_Z^{-1} συμβολίζουμε τη γενικευμένη αντίστροφη συνάρτηση, της F_Z).

Ενδιαφέρον είναι και το επόμενο θεώρημα, διότι αποτελεί σημαντικό κρίκο, στην αλυσίδα των αποτελεσμάτων που ακολουθούν.

Θεώρημα 1.1.1 *Έστω δυο τ.μ. Z και Y, με συναρτήσεις κατανομής F_Z, F_Y, αντιστοίχως. Τότε $Z \leq_{st} Y$ εάν και μόνο εάν, υπάρχουν δυο άλλες τ.μ. Z' και Y' (σε κάποιο χώρο πιθανότητας), με συναρτήσεις κατανομής F_Z, F_Y, αντιστοίχως, τέτοιες ώστε $Z' \leq Y'$.*

Απόδειξη. Ας θεωρήσουμε ότι $Z \leq_{st} Y$, και την (γενικευμένη) αντίστροφη συνάρτηση

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}, u \in (0,1),$$

μίας συνάρτησης κατανομής F. Έστω ακόμη μια τ.μ. U, με ομοιόμορφη συνάρτηση κατανομής στο (0,1), και οι τ.μ. $Z' = F_Z^{-1}(U)$ και $Y' = F_Y^{-1}(U)$. Τότε, εύκολα διαπιστώνουμε ότι οι Z' και Y', έχουν συναρτήσεις κατανομής F_Z, F_Y, αντιστοίχως. Επιπλέον, αφού $Z \leq_{st} Y$, δηλαδή

$$F_Z(x) \geq F_Y(x), \text{ για κάθε } x \in (-\infty, +\infty),$$

ισχύει και

$$F_Z^{-1}(u) \leq F_Y^{-1}(u), \text{ για κάθε } u \in (0,1).$$

Άρα $Z' \leq Y'$ και η απόδειξη ολοκληρώνεται άμεσα, ακολουθώντας και την αντίστροφη, πορεία.

■

Με βάση το προηγούμενο θεώρημα, μπορούμε να προχωρήσουμε στην απόδειξη του επόμενου αποτελέσματος, το οποίο (όπως θα φανεί και στη συνέχεια) μας προσφέρει ένα κοινό σημείο μεταξύ της στοχαστικής διάταξης τ.μ. και της διάταξης τυχαίων διανυσμάτων.

Θεώρημα 1.1.2 *Για δυο τ.μ. Z και Y, ισχύει $Z \leq_{st} Y$ εάν και μόνο εάν*

$$E(f(Z)) \leq E(f(Y)),$$

για κάθε αύξουσα συνάρτηση $f : \Re \to \Re$, για την οποία υπάρχουν και οι δυο μέσες τιμές.

Απόδειξη. Αρχικά υποθέτουμε ότι $Z \leq_{st} Y$. Τότε, με βάση το Θεώρημα 1.1.1 και χωρίς περιορισμό της γενικότητας, θεωρούμε ότι $Z \leq Y$. Όποτε για κάθε αύξουσα συνάρτηση f, ισχύει $f(Z) \leq f(Y)$ και από τις ιδιότητες της μέσης τιμής προκύπτει, $E(f(Z)) \leq E(f(Y))$. Αντίστροφα, εάν $E(f(Z)) \leq E(f(Y))$ για κάθε αύξουσα συνάρτηση f, τότε η προηγούμενη ανισότητα θα ισχύει και για την αύξουσα συνάρτηση $f_u(x)$, όπου

$$f_u(x) = \mathbf{1}_{(u,\infty)}(x) = \begin{cases} 1, & \text{για } x > u \\ 0, & \text{διαφορετικά .} \end{cases}$$

Η απόδειξη ολοκληρώνεται, αν λάβουμε υπόψη ότι, $P(Z > u) = E(f_u(Z))$ και $P(Y > u) = E(f_u(Y))$, για κάθε $u \in \Re$.

■

Αξίζει να αναφέρουμε, ότι η συνήθης στοχαστική διάταξη ανάμεσα στις τ.μ. Z, Y ($Z \leq_{st} Y$) μπορεί ισοδύναμα να διατυπωθεί και ως

$$P(Z \in D) \geq P(Y \in D), \text{ για κάθε άνω σύνολο } D \subseteq \Re$$

(ανάλογη σχέση, συναντάμε και στην περίπτωση των τυχαίων διανυσμάτων). Επίσης, το επόμενο πόρισμα, φαίνεται να παρουσιάζει αρκετές εφαρμογές και αξίζει να προχωρήσουμε στη διατύπωση του.

Πόρισμα 1.1.1 *Έστω $Z_1, Z_2, \ldots, Z_n$ και $Y_1, Y_2, \ldots, Y_n$, ανεξάρτητες τ.μ. με $Z_i \leq_{st} Y_i$, για κάθε $i = 1, 2, \ldots, n$. Τότε,*

$$Z_{i:n} \leq Y_{i:n}, \text{ για } i = 1, 2, \ldots, n,$$

όπου με $Z_{i:n}$ συμβολίζεται η $i-$οστή μεγαλύτερη παρατήρηση στο δείγμα $Z_1, Z_2, \ldots, Z_n$ (όμοια για την $Y_{i:n}$).

Ο στόχος όσων προηγήθηκαν, είναι διττός: αφενός αποτελούν μια πολύ μικρή εισαγωγή στη θεωρία των στοχαστικών διατάξεων, και αφετέρου, να μας εισάγουν στο σημαντικό πεδίο των διατάξεων, ανάμεσα σε τυχαία διανύσματα. Η συνήθης στοχαστική διάταξη, γενικεύεται σε περιβάλλον διανυσμάτων, με τον παρακάτω ορισμό.

Ορισμός 1.1.2 *Έστω δυο τυχαία διανύσματα $\mathbf{Z}$ και $\mathbf{Y}$, του $\Re^n$. Θα λέμε ότι το τυχαίο διάνυσμα $\mathbf{Z}$ είναι μικρότερο από το τυχαίο διάνυσμα $\mathbf{Y}$, με βάση τη συνήθη (πολυδιάστατη) στοχαστική διάταξη (συμβ. $\mathbf{Z} \leq_{st} \mathbf{Y}$), εάν ισχύει*

$$E(f(\mathbf{Z})) \leq E(f(\mathbf{Y})),$$

για κάθε αύξουσα συνάρτηση $f : \Re^n \to \Re$.

Μια ισοδύναμη συνθήκη, διαπιστώσαμε στο Θεώρημα 1.1.2 και για τη διάταξη ανάμεσα σε τ.μ. Η συνήθης πολυδιάστατη διάταξη (Lehmann (1955)), μελετήθηκε και εφαρμόστηκε, από πολλούς συγγραφείς σε διάφορα πεδία (βλ. π.χ. Marshall and Olkin (1979)). Η επόμενη πρόταση, αποτελεί τη γενίκευση σε πολλές διαστάσεις, αντίστοιχου αποτελέσματος για μονοδιάστατες τ.μ.

Πρόταση 1.1.1 *Για δυο τυχαία διανύσματα* $\mathbf{Z}$ *και* $\mathbf{Y}$*, του* $\Re^n$*, ισχύει* $\mathbf{Z} \leq_{st} \mathbf{Y}$*, αν και μόνο αν*

$$P(\mathbf{Z} \in D) \geq P(\mathbf{Y} \in D), \text{ για κάθε άνω σύνολο } D \subseteq \Re^n.$$

Απ' όσα προηγήθηκαν, μπορεί κάποιος να κατανοήσει ότι είναι πολύ δύσκολο να εξακριβώσουμε τη συνήθη στοχαστική διάταξη, ανάμεσα σε δυο τυχαία διανύσματα, με βάση τις παραπάνω σχέσεις. Σημαντική βοήθεια σ' ένα τέτοιο εγχείρημα, μας προσφέρει το επόμενο Θεώρημα 1.1.3 (Veinott (1965)), όπου ουσιαστικά η εξέταση για την ύπαρξη της πολυδιάστατης συνήθης στοχαστικής διάταξης, ανάγεται σε πρόβλημα στοχαστικής διάταξης, ανάμεσα σε δύο μονοδιάστατες τυχαίες μεταβλητές.

Πριν το επόμενο θεώρημα, να σημειώσουμε ότι ο συμβολισμός $[Z_s|Z_1 = z_1, Z_2 = z_2, \ldots, Z_{s-1} = z_{s-1}]$ χρησιμοποιείται για να υποδηλώσει την τ.μ. με κατανομή, τη δεσμευμένη κατανομή του Z_s, δοθέντος $Z_1 = z_1, Z_2 = z_2, \ldots, Z_{s-1} = z_{s-1}$ (όμοια και για την $[Y_s|Y_1 = y_1, Y_2 = y_2, \ldots, Y_{s-1} = y_{s-1}]$).

Θεώρημα 1.1.3 *Ας θεωρήσουμε δύο τυχαία διανύσματα* $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$ *και* $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$*, του* $\Re^n$*. Εάν* $Z_1 \leq_{st} Y_1$*, και για κάθε* $s = 2, 3, \ldots, n$ *ισχύει*

$$[Z_s|Z_1 = z_1, Z_2 = z_2, \ldots, Z_{s-1} = z_{s-1}] \leq_{st} [Y_s|Y_1 = y_1, Y_2 = y_2, \ldots, Y_{s-1} = y_{s-1}] \quad (1.1.1)$$

για όλα τα $z_j \leq y_j$ *(για* $j = 1, 2, \ldots, s-1$*), τότε*

$$\mathbf{Z} \leq_{st} \mathbf{Y}.$$

Για την απόδειξη του παραπάνω θεωρήματος (το οποίο θα παίξει σημαντικό ρόλο, στο κεφάλαιο αυτό), χρησιμοποιείται μια «κατασκευαστική» μέθοδος (με πολλές εφαρμογές-κυρίως στην προσομοίωση, standard construction method, Rubinstein and Melamed (1998)), και ένα αποτέλεσμα για τυχαία διανύσματα, αντίστοιχο μ' αυτό του Θεωρήματος 1.1.1.

1.2 Συστήματα μονάδων, με πολλαπλά επίπεδα αποτυχίας

Ας υποθέσουμε ότι $I = \{1, 2, \ldots, n\}$ είναι το σύνολο των μονάδων, ενός συστήματος MFM. Για κάθε μονάδα $i \in I$ συμβολίζουμε με 0, την κατάσταση λειτουργίας της, ενώ το σύνολο $S = \{1, 2, \ldots, m\}$, περιλαμβάνει όλους τους διαφορετικούς τύπους αποτυχίας. Σε μια συγκεκριμένη χρονική στιγμή t, μια μονάδα $i \in I$ μπορεί να βρεθεί μόνο σε μια, από τις $m+1$ διαφορετικές καταστάσεις του συνόλου $S \cup \{0\}$. Επομένως, συμβολίζοντας με $p_i = q_{0i}$ την πιθανότητα λειτουργίας της i μονάδος και με q_{si}, $s \in S$ την πιθανότητα η ίδια μονάδα να βρίσκεται σε κατάσταση αποτυχίας τύπου s, έχουμε

$$p_i + \sum_{s=1}^{m} q_{si} = \sum_{s=0}^{m} q_{si} = 1, \quad i = 1, 2, \ldots, n$$

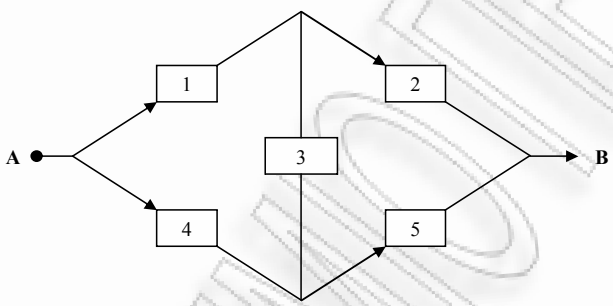
(να σημειώσουμε ότι η παράμετρος t, έχει απαλειφθεί από τους συμβολισμούς μας, διότι περιορίζουμε τη μελέτη του συστήματος, σε μια συγκεκριμένη χρονική στιγμή).

Έχουμε ήδη αναφέρει ότι τα συστήματα αξιοπιστίας που μελετάμε, βρίσκονται σε κατάσταση μη λειτουργίας, εάν συγκεκριμένα σύνολα μονάδων, αντιμετωπίζουν τον ίδιο τύπο αποτυχίας. Εξετάζοντας τη δομή ενός MFM συστήματος, είναι δυνατόν να προσδιοριστούν m οικογένειες συνόλων $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_m$, από το δυναμοσύνολο του I, τέτοιες ώστε το σύστημα να αποτυγχάνει, εάν και μόνο εάν υπάρχει τουλάχιστον ένα $s \in S$ και ένα τουλάχιστον σύνολο $C \in \mathbf{C}_s$, με όλες τις μονάδες του σε κατάσταση αποτυχίας τύπου s (σε μια δεδομένη χρονική στιγμή t). Θα αναφερόμαστε στο τελευταίο γεγονός με τον όρο «αποτυχία συστήματος τύπου s», ενώ τα στοιχεία του συνόλου $\mathbf{C}_s$ (υποσύνολα του I), θα ονομάζονται «ελάχιστα σύνολα διακοπής τύπου s». Να σημειώσουμε ότι, η οικογένεια $\mathbf{C}_s$ (για κάποιο $s \in S$), περιλαμβάνει όλα εκείνα τα υποσύνολα του I, για τα οποία η στιγμιαία αποτυχία των μονάδων τους, προκαλεί αποτυχία συστήματος τύπου s, άλλα και ταυτόχρονα, για κάθε σύνολο του $\mathbf{C}_s$, δεν είναι δυνατόν να υπάρξει υποσύνολο του, με την παραπάνω ιδιότητα (εκεί αποδίδεται και η χρησιμοποίηση του όρου «ελάχιστα σύνολα διακοπής»).

Πολλά από τα συστήματα που εντοπίζουμε στη βιβλιογραφία ή σε πραγματικές εφαρμογές, ικανοποιούν τις παραπάνω προϋποθέσεις. Για παράδειγμα, σε ηλεκτρονικά κυκλώματα όπου συναντάμε συχνά συστήματα σε μορφή γέφυρας (ένα από τα πιο γνωστά συστήματα στη θεωρία αξιοπιστίας, δείτε και Σχήμα 1.2.1), θα μπορούσε κάποιος να κάνει τη ρεαλιστική υπόθεση ότι οι μονάδες παρουσιάζουν αδυναμία να κλείσουν ή να ανοίξουν το κύκλωμα (οπότε έχουμε δυο ειδών αποτυχίες). Επίσης, σε συστήματα τηλεπικοινωνιών ή αναμετάδοσης σήματος, μπορεί ένας πομπός είτε να επίδειξη αδυναμία στη μετάδοση του σήματος ή να

προβεί σε λανθασμένη μετάδοση (αντίστοιχα προβλήματα μπορεί να υπάρξουν και για ένα δέκτη). Στο Παράδειγμα 1.1 θα δούμε πώς ένα σύστημα k-από-τα-n (το οποίο έχει μελετηθεί από πολλούς συγγραφείς), μπορεί να να μετασχηματιστεί σε περιβάλλον πολλαπλών αποτυχιών. Το ίδιο μπορεί να συμβεί και για τα συνεχόμενα-k-από-τα-n, και τις γενικεύσεις τους (ο αναγνώστης μπορεί να ανατρέξει στο βιβλίο των Kuo and Zuo (2003), για μια ανασκόπηση στη θεωρία των παραπάνω συστημάτων).

Σχήμα 1.2.1: Σύστημα αξιοπιστίας: Γέφυρα



Παράδειγμα 1.1 Ας θεωρήσουμε ένα σύστημα με n μονάδες, όπου κάθε μια μπορεί είτε να λειτουργεί, είτε να βρίσκεται σε κατάσταση αποτυχίας τύπου s, με $s = 1, 2, 3$ (δηλαδή, $m = 3$). Το σύστημα αυτό δε θα λειτουργεί εάν και μόνο εάν, ανάμεσα στις n μονάδες, υπάρχουν τουλάχιστον k_1 σε κατάσταση αποτυχίας τύπου 1, ή τουλάχιστον k_2 σε κατάσταση αποτυχίας τύπου 2, ή τουλάχιστον k_3 σε κατάσταση αποτυχίας τύπου 3 ($k_s \leq n, s = 1, 2, 3$). Με το παραπάνω σύστημα (και για συγκεκριμένες κυρίως τιμές των παραμέτρων), έχουν ασχοληθεί πολλοί συγγραφείς όπως οι Ben-Dov (1980), Barlow and Heidtmann (1984) κ.α.

Είναι φανερό ότι το συγκεκριμένο σύστημα αποτελεί μια γενίκευση του κλασικού-απλού συστήματος k-από-τα-n:F. Το απλό σύστημα απαρτίζεται από n μονάδες, οι οποίες μπορούν να βρεθούν μόνο σε κατάσταση λειτουργίας ή αποτυχίας ($m = 1$)-το σύστημα αποτυγχάνει αν και μόνο αν τουλάχιστον k από τις n μονάδες του, δε λειτουργούν.

Στο σύστημα *MFM* του παραδείγματός μας, ορίζονται τρεις οικογένειες ελάχιστων συνόλων διακοπής, $\mathbf{C}_1$,$\mathbf{C}_2$,$\mathbf{C}_3$, όπου καθεμία περιλαμβάνει όλα τα υποσύνολα του $I = \{1, 2, \ldots, n\}$, με k_s στοιχεία, για $s = 1, 2, 3$, αντιστοίχως. Δηλαδή,

$$\mathbf{C}_s = \{C \subseteq I : |C| = k_s\}, \quad s = 1, 2, 3,$$

όπου με $|C|$ συμβολίζουμε τον πληθάριθμο του συνόλου C (οπότε, $|\mathbf{C}_s| = \binom{n}{k_s}, s = 1, 2, 3$).

■

Παράδειγμα 1.2 Ο Shanthikumar το 1987 εισήγαγε ένα σύστημα το οποίο εμφανίζει πολυάριθμες εφαρμογές, στα πεδία των τηλεπικοινωνιών, των δικτύων μεταφοράς υγρών, δικτύων αναμετάδοσης σήματος και αλλού. Το σύστημα αυτό αποτελείται από ένα «πομπό» (συμβ. 0), ένα «δέκτη» (συμβ. $n+1$) και n μονάδες $I = \{1, 2, \ldots, n\}$, συνδεδεμένες σε μία σειρά. Ο πομπός είναι συνδεδεμένος (επικοινωνεί) με τις μονάδες $\{1, 2, \ldots, \epsilon_0\}$, και κάθε μια μονάδα $i \in I$, με τις μονάδες $\{i+1, i+2, \ldots, i+\epsilon_i\}$, $1 \leq \epsilon_i \leq n, i = 1, 2, \ldots, n$. Κάθε μονάδα (όπως και ολόκληρο το σύστημα) μπορεί είτε να λειτουργεί είτε να αποτυγχάνει. Ο πομπός, ο δέκτης και οι συνδέσεις μεταξύ τους, θεωρούνται ότι είναι πάντα σε κατάσταση λειτουργίας. Το σύστημα λειτουργεί, εάν και μόνο εάν, υπάρχει μια σύνδεση του πομπού με το δέκτη, μέσω μονάδων που λειτουργούν. Για το προηγούμενο σύστημα έχει χρησιμοποιηθεί η ονομασία consecutively-connected system, (CCS).

Μεταφράζοντας όλα τα παραπάνω, σ' ένα πρόβλημα μετάδοσης σήματος, μπορούμε να θεωρήσουμε, ότι από τον πομπό μεταδίδεται ένα σήμα σ' ένα από τους ϵ_0 πλησιέστερους σταθμούς αναμετάδοσης (μονάδες). Ο σταθμός $i \in I$ που λαμβάνει το σήμα, εκπέμπει με τη σειρά του το σήμα σ' ένα από τους ϵ_i αναμεταδότες, με τους οποίους είναι συνδεδεμένος· και η διαδικασία συνεχίζεται, έως ότου το σήμα φθάσει στο δέκτη.

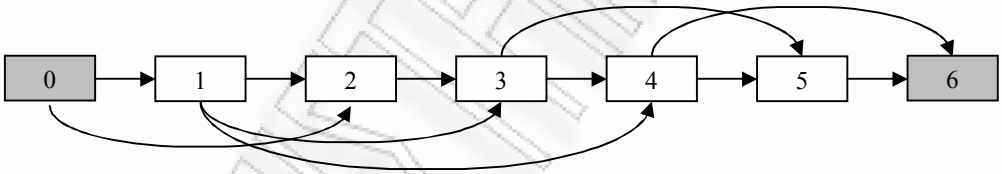
Είναι όμως ρεαλιστικό κάποιος να υποθέσει, για παράδειγμα, ότι οι αναμεταδότες (μονάδες) παρουσιάζουν όχι μόνο ένα είδος βλάβης, αλλά m διαφορετικά (π.χ. σαν σφάλμα τύπου 1, μπορεί να είναι μια συγκεκριμένη αλλοίωση στο σήμα, σαν σφάλμα τύπου 2, αδυναμία μετάδοσης κτλ). Έτσι, η γενίκευση του παραπάνω συστήματος σε περιβάλλον μονάδων, με πολλαπλά είδη αποτυχίας, φαίνεται αρκετά χρήσιμη και παρουσιάζει ιδιαίτερο πρακτικό ενδιαφέρον.

Προς αυτή την κατεύθυνση έχουν γίνει ήδη κάποιες προσπάθειες, και μια απ' αυτές είναι των Hwang and Yao (1989). Στη συγκεκριμένη εργασία, η κατάσταση κάθε μονάδος εκφράζει το πλήθος των πλησιέστερων μονάδων, με τις οποίες επικοινωνεί, σε κάποια δεδομένη χρονική στιγμή. Μ' άλλα λόγια, οι δυνατές καταστάσεις μιας μονάδος $i \in I$ είναι οι $\{0, 1, \ldots, \epsilon_i\}$. Επιπλέον, μια μονάδα θεωρείται χαλασμένη εάν δεν μπορεί να επικοινωνήσει με καμία άλλη μονάδα (δηλαδή, είναι στην κατάσταση μηδέν). Παράλληλα, το σύστημα είναι σε κατάσταση λειτουργίας, εάν μπορούμε να μεταβούμε από τον πομπό στο δέκτη, διαμέσου μονάδων με καταστάσεις διαφορετικές από τη μηδενική (ο αναγνώστης μπορεί να

ανατρέξει στο βιβλίο των Kuo and Zuo (2003), για τη μελέτη του προηγούμενου συστήματος, όπως και για τις γενικεύσεις του-circular consecutively-connected systems, two-way consecutively-connected systems κ.α.).

Ο τρόπος που επιχειρούμε να μεταφέρουμε το CCS, σε σύστημα με μονάδες με πολλαπλά επίπεδα αποτυχίας, είναι διαφορετικός, από τον προηγούμενο. Θεωρούμε (σε μια συγκεκριμένη χρονική στιγμή) ότι κάθε μονάδα μπορεί να είναι σε κατάσταση λειτουργίας ή να είναι, σε κατάσταση αποτυχίας τύπου $s, (s \in S = \{1, 2, \ldots, m\})$. Το σύστημα λειτουργεί εάν και μόνο εάν, υπάρχει τρόπος να φθάσουμε από τον πομπό στο δέκτη, μέσω μονάδων, από τις οποίες το πολύ n_s (όπου $n_s \in \{0, 1, \ldots, n\}$) απ' αυτές είναι σε κατάσταση αποτυχίας τύπου s, για κάθε $s \in S$. Αλλιώς, μπορούμε να πούμε ότι το σύστημα αποτυγχάνει, εάν και μόνο εάν, σε κάθε πιθανή διαδρομή από τον πομπό στο δέκτη, υπάρχουν περισσότερες από n_s μονάδες (δηλαδή, $n_s + 1, n_s + 2, \ldots, n$), σε κατάσταση αποτυχίας τύπου s (για κάποιο $s \in S$). Να σημειώσουμε ότι, εάν $n_s = 0$ για κάθε s, τότε η αξιοπιστία του συστήματος, ταυτίζεται μ' αυτή του απλού CCS.

Σχήμα 1.2.2: Σύστημα CCS με $\epsilon_0 = 2, \epsilon_1 = 3, \epsilon_2 = 1, \epsilon_3 = 2, \epsilon_4 = 2, \epsilon_5 = 1$



Για παράδειγμα, όταν $m = 2$, θεωρούμε ότι η μονάδα i είναι σε κατάσταση αποτυχίας τύπου 1, εάν π.χ. υπάρχει αδυναμία αποστολής του σήματος, ενώ είναι σε κατάσταση αποτυχίας τύπου 2, εάν π.χ. μεταδίδει το σήμα λανθασμένα. Ας θεωρήσουμε ότι έχουμε 5 μονάδες, με $n_1 = 0, n_2 = 1$ και $\epsilon_0 = 2, \epsilon_1 = 3, \epsilon_2 = 1, \epsilon_3 = 2, \epsilon_4 = 2, \epsilon_5 = 1$ (βλ. Σχήμα 1.2.2). Μελετώντας το σύστημα, οι δυο οικογένειες ελάχιστων συνόλων διακοπής είναι,

$$\mathbf{C}_1 = \{\{1, 2\}, \{1, 3\}, \{3, 4\}, \{4, 5\}\}, \quad \mathbf{C}_2 = \{\{1, 2, 3, 4\}, \{1, 2, 4, 5\}, \{1, 3, 4, 5\}\}.$$

■

Με βάση τα προηγούμενα καταλαβαίνουμε ότι η κατάσταση μιας μονάδος $i \in I$, μπορεί να εκφραστεί με τη βοήθεια μιας ακέραιας τ.μ. V_i, με πεδίο τιμών το $S \cup \{0\}$, με τον εξής

τρόπο

$$V_i = \begin{cases} 0, & \text{εάν η } i \text{ μονάδα λειτουργεί,} \\ 1, & \text{εάν η } i \text{ μονάδα είναι σε κατάσταση αποτυχίας τύπου } 1, \\ \vdots & \\ m, & \text{εάν η } i \text{ μονάδα είναι σε κατάσταση αποτυχίας τύπου } m. \end{cases}$$

Παρ' όλα αυτά, θα φανεί αρκετά γόνιμο για τη συνέχεια, εάν για την περιγραφή των καταστάσεων μίας μονάδος, χρησιμοποιηθεί ένα τυχαίο διάνυσμα (στήλη) $\mathbf{X}_{.i} = (X_{1i}, X_{2i}, \ldots, X_{mi})'$ (και όχι η τ.μ. V_i), τέτοιο ώστε η συντεταγμένη του X_{si}, να παίρνει την τιμή 0 εάν η i μονάδα βρίσκεται σε κατάσταση αποτυχίας τύπου s, και 1 διαφορετικά ($s \in S$ και $i \in I$). Επομένως, $\mathbf{X}_{.i} = (1, 1, \ldots, 1)'$ εάν και μόνο εάν η μονάδα λειτουργεί. Με την πρακτική αυτή, η κατάσταση ολόκληρου του συστήματος, μπορεί εύκολα να περιγραφεί μέσω μιας δίτιμης (αύξουσας) συνάρτησης, όμοιας με τη συνάρτηση δομής ενός κλασικού συστήματος αξιοπιστίας (συστήματος *SFM*, βλ. π.χ. Barlow and Proschan (1981)).

Ας θεωρήσουμε στη συνέχεια τον παρακάτω τυχαίο πίνακα, ο όποιος έχει ως στήλες τα τ.δ. $\mathbf{X}_{.1}, \mathbf{X}_{.2} \ldots, \mathbf{X}_{.n}$, δηλαδή

$$\mathbf{X} = (X_{si})_{m\times n} = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1n} \\ X_{21} & X_{22} & \ldots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \ldots & X_{mn} \end{pmatrix} \tag{1.2.1}$$

και ας συμβολίσουμε με $\mathbf{X}_{1.}, \mathbf{X}_{2.} \ldots, \mathbf{X}_{m.}$, τις γραμμές του. Τότε θα έχουμε αποτυχία συστήματος τύπου s, εάν και μόνο εάν η διτιμη (0-1) συνάρτηση δομής

$$\varphi_s(\mathbf{X}_{s.}) = \prod_{C\in\mathbf{C}_s} (1 - \prod_{i\in C}(1 - X_{si}))$$

πάρει την τιμή 0. Επιπρόσθετα η πιθανότητα να μην έχουμε αποτυχία του συστήματος τύπου s, είναι ίση με $E(\varphi_s(\mathbf{X}_{s.}))$. Η τελευταία μέση τιμή, είναι η αξιοπιστία ενός συστήματος *SFM*, αποτελούμενου από n ανεξάρτητες μονάδες, με πιθανότητες αποτυχίας $q_{s1}, q_{s2}, \ldots, q_{sn}$ (ή αλλιώς, με πιθανότητες λειτουργίας $1 - q_{s1}, 1 - q_{s2}, \ldots, 1 - q_{sn}$) και ελάχιστα σύνολα διακοπής $\mathbf{C}_s$.

Έτσι, αν $\mathbf{q}_{s.} = (q_{s1}, q_{s2}, \ldots, q_{sn})$ είναι το διάνυσμα πιθανοτήτων αποτυχίας τύπου s, των μονάδων, τότε

$$R_s(\mathbf{q}_{s.}) = E(\varphi_s(\mathbf{X}_{s.})) \tag{1.2.2}$$

είναι η πιθανότητα το σύστημα να μη χαλάσει, λόγω αποτυχίας τύπου s. Επομένως, το σύστημα αποτύγχάνει εάν και μόνο εάν $\varphi_s(\mathbf{X}_{s.}) = 0$, για τουλάχιστον ένα $s \in S$. Έτσι, η δίτιμη (και αύξουσα) συνάρτηση $\varphi : \{0,1\}^{mn} \to \{0,1\}$, με

$$\varphi = \varphi(\mathbf{X}_{1.}, \mathbf{X}_{2.}, \ldots, \mathbf{X}_{m.}) = \prod_{s=1}^{m} \varphi_s(\mathbf{X}_{s.}) = \prod_{s=1}^{m} \left(\coprod_{C \in \mathbf{C}_s} (1 - \prod_{i \in C} (1 - X_{si})) \right)$$

εκφράζει την κατάσταση ολόκληρου του συστήματος, καθώς: $\varphi = 1$ εάν το σύστημα λειτουργεί και $\varphi = 0$, εάν αποτύχει. Άρα, η αξιοπιστία του συστήματος θα δίνεται από τη σχέση

$$\begin{aligned} R(\mathbf{q}) &= E(\varphi(\mathbf{X}_{1.}, \mathbf{X}_{2.}, \ldots, \mathbf{X}_{m.})) \\ &= E\left(\prod_{s=1}^{m} \varphi_s(\mathbf{X}_{s.}) \right) = E\left(\prod_{s=1}^{m} \left(\coprod_{C \in \mathbf{C}_s} (1 - \prod_{i \in C} (1 - X_{si})) \right) \right) \end{aligned} \tag{1.2.3}$$

όπου $\mathbf{q}$ είναι ο $m \times n$ πίνακας (πιθανοτήτων αποτυχίας)

$$\mathbf{q} = \begin{pmatrix} \mathbf{q}_{1.} \\ \mathbf{q}_{2.} \\ \vdots \\ \mathbf{q}_{m.} \end{pmatrix} = \begin{pmatrix} q_{11} & q_{12} & \ldots & q_{1n} \\ q_{21} & q_{22} & \ldots & q_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ q_{m1} & q_{m2} & \ldots & q_{mn} \end{pmatrix}. \tag{1.2.4}$$

Η παραπάνω προσέγγιση για τη μελέτη ενός συστήματος *MFM* (χρησιμοποιώντας τα προαναφερθέντα τυχαία διανύσματα, για να εκφράσουμε τις καταστάσεις των μονάδων, και δίνοντας την αξιοπιστία του συστήματος μέσω της (1.2.3)), προτάθηκε από τους Boutsikas and Koutras (2002a). Στην ίδια εργασία, δόθηκαν δυο τρόποι για τον ακριβή υπολογισμό της $R(\mathbf{q})$: ο ένας εκφράζοντας τη συνάρτηση αξιοπιστίας ως ανάπτυγμα αθροισμάτων από γινόμενα πιθανοτήτων αποτυχίας και ο δεύτερος (που δίδεται από το παρακάτω θεώρημα), μέσω της αξιοπιστίας ενός κατάλληλα ορισμένου συστήματος *SFM*.

Θεώρημα 1.2.1 *Η συνάρτηση αξιοπιστίας ενός MFM συστήματος, με n ανεξάρτητες μονάδες και πίνακα πιθανοτήτων αποτυχίας τον* (1.2.4) *(με $p_i > 0$, για κάθε $i \in I$), δίδεται από τη σχέση*

$$R(\mathbf{q}) = R_{SFM}\left(\prod_{i=1}^{n} \frac{p_i^{m-1}}{\prod_{s=1}^{m}(p_i + q_{si})} \right)^{-1}.$$

Η R_{SFM} είναι η συνάρτηση αξιοπιστίας ενός απλού συστήματος, με τα εξής χαρακτηριστικά:

a. το σύστημα αποτελείται από mn ανεξάρτητες μονάδες (τις $(s,i), s \in S, i \in I$), με πιθανότητες αποτυχίας $q_{si}(p_i + q_{si})^{-1}$,

β. τα $\sum_{s=1}^{m} |\mathbf{C}_s| + n\binom{m}{2}$ *ελάχιστα σύνολα διακοπής του, είναι τα παρακάτω*

- $\{(s,i), i \in C\}, C \in \mathbf{C}_s$, *για κάθε* $s \in S$
- $\{(s,i),(t,i)\}, s,t \in S, s < t$, *για κάθε* $i \in I$.

Στην εργασία Koutras (1997), δόθηκε ένας αναδρομικός τύπος, για τη συνάρτηση αξιοπιστίας ενός συγκεκριμένου συστήματος *MFM* (για το οποίο θα μιλήσουμε στην τελευταία παράγραφο, του παρόντος κεφαλαίου), για την περίπτωση $m = 2$.

Ωστόσο, καθώς αυξάνονται οι τιμές των παραμέτρων n, m, ή και ο πληθάριθμος των $\mathbf{C}_s$, $s \in S$, ο υπολογισμός της ακριβούς τιμής της αξιοπιστίας, γίνεται πρακτικά αδύνατος. Έτσι, οι Boutsikas and Koutras (2002a), πρότειναν κάποια φράγματα για την $R(\mathbf{q})$ (μερικά από τα οποία θα συναντήσουμε και στην Παράγραφο 1.4), και μελέτησαν την ασυμπτωτική της συμπεριφορά. Στην επόμενη παράγραφο, θα εισάγουμε ένα νέο κάτω φράγμα, πολλαπλασιαστικού τύπου (για τη συνάρτηση αξιοπιστίας), το οποίο βελτιώνει ένα από τα κάτω φράγματα που είχαν προτείνει οι παραπάνω συγγραφείς.

1.3 Κάτω φράγμα για τη συνάρτηση αξιοπιστίας

Ας θεωρήσουμε ότι έχουμε ένα σύστημα αξιοπιστίας *MFM*, με σύνολο μονάδων I και διανύσματα πιθανοτήτων αποτυχίας τύπου s, το $\mathbf{q}_{s\cdot} = (q_{s1}, q_{s2}, \ldots, q_{sn})$, με $s \in S$. Επιπλέον, ας εισάγουμε τα τυχαία διανύσματα (στήλες) $\mathbf{T}_{\cdot i}$, $i = 1, 2, \ldots, n$, όπου

$$\mathbf{T}_{\cdot i} = (T_{1i}, T_{2i}, \ldots, T_{mi})'$$

τα οποία αποτελούνται από ανεξάρτητες και δίτιμες (0-1) τ.μ., με

$$P(T_{1i} = 0) = q_{1i}, \quad P(T_{1i} = 1) = 1 - q_{1i}$$

και

$$P(T_{si} = 0) = q_{si}/(1 - \sum_{j=1}^{s-1} q_{ji}), \quad P(T_{si} = 1) = (1 - \sum_{j=1}^{s} q_{ji})/(1 - \sum_{j=1}^{s-1} q_{ji}),$$

για $s = 2, 3, \ldots, m$. Η πρόταση που ακολουθεί, αποτελεί το πρώτο βήμα για την απόδειξη του κάτω φράγματος, καθώς εξακριβώνει μια χρήσιμη στοχαστική διάταξη ανάμεσα στα $\mathbf{T}_{\cdot i}$, που εισήχθησαν παραπάνω, και τα τυχαία διανύσματα $\mathbf{X}_{\cdot i}$, που εκφράζουν τις καταστάσεις των μονάδων (Milienos and Koutras (2008)).

Πρόταση 1.3.1 *Έστω ένα σύστημα αξιοπιστίας MFM, για το οποίο οι καταστάσεις των ανεξάρτητων μονάδων του και οι πιθανότητες αποτυχίας τους, περιγράφονται από τους πίνακες* (1.2.1) *και* (1.2.4)*, αντιστοίχως. Τότε,*

$$\mathbf{T}'_{\cdot i} \leq_{st} \mathbf{X}'_{\cdot i}$$

για κάθε $i \in I$.

Απόδειξη. Με βάση το Θεώρημα 1.1.3 αρκεί να αποδείξουμε ότι, για κάθε $u \in (-\infty, +\infty)$ ισχύει,

$$P(T_{1i} \leq u) \geq P(X_{1i} \leq u),$$

και για $s = 2, 3, \ldots, m$,

$$P(T_{si} \leq u | T_{1i} = t_1, \ldots, T_{s-1,i} = t_{s-1}) \geq P(X_{si} \leq u | X_{1i} = x_1, \ldots, X_{s-1,i} = x_{s-1}), \quad (1.3.1)$$

για οποιαδήποτε $t_j \leq x_j$, $j = 1, 2, \ldots, s-1$. Επειδή και οι δύο τ.μ. T_{si} και X_{si}, είναι δίτιμες (0-1), η πρώτη συνθήκη ισχύει ως ισότητα, αφού για $u \geq 1$ έχουμε

$$P(T_{1i} \leq u) = P(X_{1i} \leq u) = 1,$$

για $u < 0$, είναι προφανές ότι

$$P(T_{1i} \leq u) = P(X_{1i} \leq u) = 0,$$

και για $0 \leq u < 1$, προκύπτει ότι

$$P(T_{1i} \leq u) = P(T_{1i} = 0) = q_{1i} = P(X_{1i} = 0) = P(X_{1i} \leq u).$$

Για να προχωρήσουμε στην απόδειξη της συνθήκης (1.3.1), παρατηρούμε αρχικά ότι, λόγω της ανεξαρτησίας των τ.μ. $T_{1i}, T_{2i}, \ldots, T_{si}$, η δεσμευμένη πιθανότητα που υπάρχει στο αριστερό μέρος της ανισότητας, είναι ίση με

$$P(T_{si} \leq u) = \begin{cases} 1, & \text{εάν } u \geq 1 \\ \frac{q_{si}}{1 - \sum_{j=1}^{s-1} q_{ji}}, & \text{εάν } 0 \leq u < 1 \\ 0, & \text{εάν } u < 0. \end{cases}$$

Παράλληλα, για $u \geq 1$ έχουμε

$$P(T_{si} \leq u | T_{1i} = t_1, \ldots, T_{s-1,i} = t_{s-1}) = P(X_{si} \leq u | X_{1i} = x_1, \ldots, X_{s-1,i} = x_{s-1}) = 1,$$

ενώ για $u < 0$, και οι δύο παραπάνω πιθανότητες, μηδενίζονται. Επομένως, η απόδειξη της προτάσεως θα ολοκληρωθεί, εάν δείξουμε ότι

$$P(X_{si} \leq u | X_{1i} = x_1, X_{2i} = x_2, \ldots, X_{s-1,i} = x_{s-1}) \leq \frac{q_{si}}{1 - \sum_{j=1}^{s-1} q_{ji}}$$

για $0 \leq u < 1$ και $t_j \leq x_j$, $j = 1, 2, \ldots, s-1$, ή ισοδύναμα

$$P(X_{si} = 0 | X_{1i} = x_1, X_{2i} = x_2, \ldots, X_{s-1,i} = x_{s-1}) \leq \frac{q_{si}}{1 - \sum_{j=1}^{s-1} q_{ji}}$$

για κάθε $x_j \in \{0, 1\}$, $j = 1, 2, \ldots, s-1$.

Λαμβάνοντας υπόψιν ότι, το πολύ μία από τις $X_{1i}, X_{2i}, \ldots, X_{mi}$ μπορεί να πάρει την τιμή 0 (με τις υπόλοιπες να είναι ίσες με 1), καταλήγουμε στα εξής:

α. όταν τουλάχιστον μια από τις $x_1, x_2, \ldots, x_{s-1}$ είναι ίση με 0, τότε η προηγούμενη ανισότητα ισχύει, καθώς το αριστερό της μέρος μηδενίζεται,

β. όταν οι $x_1, x_2, \ldots, x_{s-1}$ είναι όλες ίσες με 1, τότε έχουμε

$$\begin{aligned} P(X_{si} = 0 | X_{1i} = 1, X_{2i} = 1, \ldots, X_{s-1,i} = 1) &= \frac{P(X_{si} = 0, X_{1i} = 1, \ldots, X_{s-1,i} = 1)}{P(X_{1i} = 1, X_{2i} = 1, \ldots, X_{s-1,i} = 1)} \\ &= \frac{P(X_{si} = 0)}{1 - \sum_{j=1}^{s-1} P(X_{ji} = 0)} = \frac{q_{si}}{1 - \sum_{j=1}^{s-1} q_{ji}}. \end{aligned}$$

Επομένως, και σ' αυτήν την περίπτωση η ανισότητα που εξετάζουμε είναι αληθής.

Έτσι, τα διανύσματα $\mathbf{T}'_{.i}$ και $\mathbf{X}'_{.i}$ $(i \in I)$ ικανοποιούν τις προϋποθέσεις του Θεωρήματος 1.1.3, και η απόδειξη της προτάσεως έχει ολοκληρωθεί.

■

Είμαστε πλέον σε θέση να αναφέρουμε το κύριο αποτέλεσμα του συγκεκριμένου κεφαλαίου, που αφορά το κάτω φράγμα για τη συνάρτηση αξιοπιστίας, ενός *MFM* συστήματος. Το φράγμα αυτό υπολογίζεται ως γινόμενο από συναρτήσεις αξιοπιστίας, κατάλληλα ορισμένων απλών συστημάτων, και σημαντικό ρόλο στην απόδειξη του, παίζει η στοχαστική διάταξη που προκύπτει από την Πρόταση 1.3.1.

Θεώρημα 1.3.1 *Έστω ένα MFM σύστημα αξιοπιστίας, για το οποίο οι καταστάσεις των ανεξάρτητων μονάδων του και οι πιθανότητες αποτυχίας τους, περιγράφονται από τους πίνακες* (1.2.1) *και* (1.2.4)*, αντιστοίχως. Για κάθε $i \in I$ και $s \in S$, ας θεωρήσουμε τις ποσότητες*

$$Q_{si} = \begin{cases} q_{1i}, & \text{εάν } s = 1 \\ \frac{q_{si}}{1 - \sum_{j=1}^{s-1} q_{ji}}, & \text{εάν } s \geq 2. \end{cases} \tag{1.3.2}$$

Τότε για την αξιοπιστία $R(\mathbf{q})$, του συστήματος MFM ισχύει

$$R(\mathbf{q}) \geq \prod_{s=1}^{m} R_s(Q_{s1}, Q_{s2}, \ldots, Q_{sn}) = L(\mathbf{q}),$$

όπου οι πιθανότητες $R_s(Q_{s1}, Q_{s2}, \ldots, Q_{sn})$, δίνονται μέσω της (1.2.2).

Απόδειξη. Ας εκφράσουμε αρχικά τη συνάρτηση δομής φ του συστήματος, μέσω της συνάρτησης $\psi : \{0,1\}^{mn} \to \{0,1\}$, της οποίας το όρισμα είναι το διάνυσμα $(\mathbf{X}'_{.1}, \mathbf{X}'_{.2}, \ldots, \mathbf{X}'_{.n})$, διαστάσεως $1 \times (mn)$, και ορίζεται από τον τύπο

$$\psi(\mathbf{X}'_{.1}, \mathbf{X}'_{.2}, \ldots, \mathbf{X}'_{.n}) = \prod_{s=1}^{m} \varphi_s(\mathbf{X}_{s.}) = E\left(\prod_{s=1}^{m}\left(\prod_{C\in\mathbf{C}_s}(1-\prod_{i\in C}(1-X_{si}))\right)\right).$$

Είναι εύκολο να διαπιστώσουμε ότι και η ψ, είναι μια δίτιμη και αύξουσα συνάρτηση. Ακόμη, λόγω του ότι οι συντεταγμένες των τυχαίων διανυσμάτων $\mathbf{T}'_{.i}$ $(i \in I)$, είναι ανεξάρτητες τ.μ., τα τυχαία διανύσματα $\mathbf{T}'_{.1}, \mathbf{T}'_{.2}, \ldots, \mathbf{T}'_{.n}$ είναι ανεξάρτητα μεταξύ τους. Επιπλέον, από την Πρόταση 1.3.1, έχουμε

$$\mathbf{T}'_{.i} \leq_{st} \mathbf{X}'_{.i}, \text{ για } i = 1, 2, \ldots, n,$$

όπου και τα $\mathbf{X}'_{.1}, \mathbf{X}'_{.2}, \ldots, \mathbf{X}'_{.n}$ είναι ανεξάρτητα τυχαία διανύσματα, αφού το σύστημα αποτελείται από ανεξάρτητες μονάδες. Επομένως, ισχύει (λόγω ανεξαρτησίας των τυχαίων διανυσμάτων $\mathbf{T}'_{.i}$ και $\mathbf{X}'_{.i}, i = 1, 2, \ldots, n$)

$$(\mathbf{T}'_{.1}, \mathbf{T}'_{.2}, \ldots, \mathbf{T}'_{.n}) \leq_{st} (\mathbf{X}'_{.1}, \mathbf{X}'_{.2}, \ldots, \mathbf{X}'_{.n}),$$

και από τη μονοτονία της ψ, παίρνουμε

$$E(\psi(\mathbf{T}'_{.1}, \mathbf{T}'_{.2}, \ldots, \mathbf{T}'_{.n})) \leq E(\psi(\mathbf{X}'_{.1}, \mathbf{X}'_{.2}, \ldots, \mathbf{X}'_{.n}))$$

ή ισοδύναμα

$$E(\varphi(\mathbf{T}_{1.}, \mathbf{T}_{2.}, \ldots, \mathbf{T}_{m.})) \leq E(\varphi(\mathbf{X}_{1.}, \mathbf{X}_{2.}, \ldots, \mathbf{X}'_{m.})).$$

Η μέση τιμή στο δεξί μέρος της ανισότητας, είναι η αξιοπιστία $R = R(\mathbf{q})$ του συστήματος *MFM*. Από την άλλη, κάνοντας χρήση του τύπου (1.2.3) (για τα διανύσματα $\mathbf{T}_{1.}, \mathbf{T}_{2.}, \ldots, \mathbf{T}_{m.}$), μπορούμε να γράψουμε ότι

$$E(\varphi(\mathbf{T}_{1.}, \mathbf{T}_{2.}, \ldots, \mathbf{T}_{m.})) = E(\prod_{s=1}^{m} \varphi_s(\mathbf{T}_{s.}))$$

και λόγω της ανεξαρτησίας των T_{si}, να πάρουμε

$$E(\prod_{s=1}^{m}\varphi_s(\mathbf{T}_{s.})) = \prod_{s=1}^{m}E(\varphi_s(\mathbf{T}_{s.})).$$

Η απόδειξη ολοκληρώνεται, με την παρατήρηση ότι για τους παράγοντες του παραπάνω γινομένου, ισχύει

$$E(\varphi_s(\mathbf{T}_{s.})) = R_s(Q_{s1}, Q_{s2}, \ldots, Q_{sn}), \text{ για κάθε } s = 1, 2, \ldots, m.$$

Αξίζει να σημειώσουμε ότι η $R_s = R_s(Q_{s1}, Q_{s2}, \ldots, Q_{sn})$, είναι η αξιοπιστία ενός *SFM* συστήματος, αποτελούμενα από n ανεξάρτητες μονάδες, με πιθανότητες αποτυχίας Q_{si}, $i \in I$, και ελάχιστα σύνολα διακοπής $C \in \mathbf{C}_s$.

■

Ένα σημαντικό χαρακτηριστικό του $L(\mathbf{q})$ είναι ότι ο υπολογισμός του, βασίζεται σε μια διάταξη των m οικογενειών $\mathbf{C}_s$, $s = 1, 2, \ldots, m$. Αυτό έχει ως αποτέλεσμα η τιμή του $L(\mathbf{q})$ να εξαρτάται από την αντίστοιχη διάταξη (βλ. τη μορφή των πιθανοτήτων Q_{si} στην έκφραση (1.3.2)). Το γεγονός αυτό οφείλεται στην απόδειξη που ακολουθήσαμε στην Πρόταση 1.3.1, όπου για να εξακριβώσουμε τη στοχαστική διάταξη, ανάμεσα στα τυχαία διανύσματα που μας ενδιέφεραν, έπρεπε να υπολογίσουμε κάποιες δεσμευμένες πιθανότητες. Η σειρά με την οποία «εισέρχονταν» τα γεγονότα στη δέσμευση, έπαιζε ρόλο και ουσιαστικά, ήταν η σειρά με την οποία είχαμε διατάξει τις οικογένειες $\mathbf{C}_s$, $s = 1, 2, \ldots, m$. Άμεσο επακόλουθο είναι η ύπαρξη, εν δυνάμει $m!$ διαφορετικών τιμών για το κάτω φράγμα. Σίγουρα, ανάμεσα σ' αυτές τις τιμές κάποιος μπορεί, και πρέπει να επιλέξει εκείνη τη διάταξη, η οποία μεγιστοποιεί το $L(\mathbf{q})$. Θα δούμε στην παράγραφο που ακολουθεί, πόσο αλλάζουν οι τιμές του $L(\mathbf{q})$, από διάταξη σε διάταξη (σ' ένα συγκεκριμένο σύστημα), και αν πραγματικά ο επιπλέον κόπος της αναζήτησης της μέγιστης τιμής, ισοσταθμίζεται από τη βελτίωση της προσέγγισης, που επιτυγχάνουμε.

Με το προηγούμενο θεώρημα εισάγαμε ένα νέο φράγμα για τη συνάρτηση αξιοπιστίας, η εύρεση του οποίου, απαιτεί τον υπολογισμό της αξιοπιστίας m απλών συστημάτων (διαδικασία εύκολη, από τη στιγμή που έχουμε προσδιορίσει τις m οικογένειες ελάχιστων συνόλων διακοπής). Οι μέθοδοι που υπάρχουν, για να φέρουμε εις πέρας μια τέτοια διαδικασία, είναι πάρα πολλές (π.χ. αναδρομικές σχέσεις, εμφύτευση τ.μ. σε Μαρκοβιανή αλυσίδα κ.α.) και μπορούν τις περισσότερες φορές να δώσουν αποτελεσματικές και υπολογιστικά γρήγορες λύσεις, στο πρόβλημα που μελετάμε. Παρ' όλα αυτά, θα είναι χρήσιμο αν καταφέρουμε μέσω του παραπάνω φράγματος, να προσεγγίσουμε την αξιοπιστία του συστήματος, μέσα από μια πιο απλή σχέση, διατηρώντας όμως την ακρίβεια της προσέγγισης. Εκεί στοχεύει και

το επόμενο πόρισμα, όπου κάνοντας χρήση του κάτω φράγματος των Esary and Proschan (1963), για τη συνάρτηση αξιοπιστίας ενός συστήματος *SFM*, μπορούμε να βρούμε ένα νέο φράγμα και για την $R(\mathbf{q})$.

Πόρισμα 1.3.1 *Έστω ένα MFM σύστημα με n ανεξάρτητες μονάδες, και m διαφορετικούς τύπους αποτυχίας. Τότε για τη συνάρτηση αξιοπιστίας του $R(\mathbf{q})$, ισχύει*

$$R(\mathbf{q}) \geq \prod_{s=1}^{m} \left(\prod_{C \in \mathbf{C}_s} (1 - \prod_{i \in C} Q_{si}) \right)$$

όπου τα Q_{si} δίδονται από την (1.3.2).

Απόδειξη. Από το Θεώρημα 1.3.1, παίρνουμε ένα κάτω φράγμα για τη συνάρτηση αξιοπιστίας $R(\mathbf{q})$, με τη μορφή

$$L(\mathbf{q}) = \prod_{s=1}^{m} R_s(Q_{s1}, Q_{s2}, \ldots, Q_{sn}).$$

Όμως, για τη συνάρτηση αξιοπιστίας R_s ξέρουμε ότι (Esary and Proschan (1963))

$$R_s(Q_{s1}, Q_{s2}, \ldots, Q_{sn}) \geq \prod_{C \in \mathbf{C}_s} (1 - \prod_{i \in C} Q_{si}).$$

Επομένως,

$$L(\mathbf{q}) \geq \prod_{s=1}^{m} \prod_{C \in \mathbf{C}_s} (1 - \prod_{i \in C} Q_{si}),$$

και

$$R(\mathbf{q}) \geq L(\mathbf{q}) \geq \prod_{s=1}^{m} \prod_{C \in \mathbf{C}_s} (1 - \prod_{i \in C} Q_{si}).$$

■

Ένα παρόμοιο φράγμα με το $L(\mathbf{q})$, είχαν προτείνει και οι Boutsikas and Koutras (2002a), οι οποίοι, εκφράζοντας την $R(\mathbf{q})$ μέσω της συνάρτησης αξιοπιστίας, ενός κατάλληλα ορισμένου απλού συστήματος και χρησιμοποιώντας στοιχεία από τη θεωρία των συναφών τ.μ., απέδειξαν ότι

$$R(\mathbf{q}) \geq \prod_{s=1}^{m} R_s(Q'_{s1}, Q'_{s2}, \ldots, Q'_{sn}) = L'(\mathbf{q}) \tag{1.3.3}$$

όπου

$$Q'_{si} = \frac{q_{si}}{p_i + q_{si}}, i = 1, 2, \ldots, n.$$

Άρα, για κάθε $s \in S$ και $i \in I$, ισχύει

$$Q'_{si} \geq Q_{si},$$

και δεδομένου ότι οι R_s (αξιοπιστίες ενός SFM συστήματος), είναι αύξουσες συναρτήσεις ως προς τις πιθανότητες λειτουργίας των μονάδων (π.χ. Barlow and Proschan (1981)), το φράγμα από το Θεώρημα 1.3.1, είναι πάντοτε καλύτερο από το (1.3.3), καθώς

$$L(\mathbf{q}) \geq L'(\mathbf{q}), \text{ για κάθε } \mathbf{q} = (q_{si})_{m \times n}.$$

1.4 Εφαρμογές και αριθμητικά αποτελέσματα

Με βάση όλα τα προηγούμενα, καταλαβαίνουμε ότι πάρα πολλά από τα γνωστά συστήματα μονάδων, με δυο δυνατές καταστάσεις, μπορούν κάτω από κατάλληλες υποθέσεις, να γενικευτούν σε συστήματα MFM.

Στη συγκεκριμένη παράγραφο μελετάμε τη συμπεριφορά του νέου κάτω φράγματος, σε μια κατηγορία συστημάτων, τα συνεχόμενα-$k_1, k_2, \ldots, k_m$-από-τα-n: MFM. Πραγματοποιούμε αριθμητικούς υπολογισμούς, για διάφορες τιμές των παραμέτρων, και συγκρίνουμε το νέο φράγμα, με τα φράγματα που έχουν ήδη εμφανιστεί στη βιβλιογραφία. Έπειτα, δίνουμε αριθμητικά αποτελέσματα για το σύστημα που εισάγαμε στο Παράδειγμα 1.2.

1.4.1 Σύστημα συνεχόμενα-$k_1, k_2, \ldots, k_m$-από-τα-n: *MFM*

Το σύστημα συνεχόμενα-$k_1, k_2, \ldots, k_m$-από-τα-n: MFM, αποτελείται από n μονάδες συνδεδεμένες γραμμικά, όπου η κάθε μια μπορεί να αντιμετωπίσει m διαφορετικά είδη αποτυχίας. Το σύστημα δε λειτουργεί λόγω αποτυχίας τύπου s, εάν και μόνο εάν τουλάχιστον k_s συνεχόμενες μονάδες, βρεθούν σε κατάσταση αποτυχίας τύπου s ($s = 1, 2, \ldots, m$). Η ειδική περίπτωση με δυο είδη αποτυχίας ($m = 2$), μελετήθηκε στη εργασία Koutras (1997), ενώ φράγματα για τη γενική περίπτωση ($m \geq 2$) έχουν δοθεί από τους Boutsikas and Koutras (2002a) και Chryssaphinou and Vaggelatou (2002). Αξίζει να αναφέρουμε την άμεση σχέση που έχει η αξιοπιστία ενός συστήματος, όπως το προηγούμενο, με τους χρόνους αναμονής μέχρι την εμφάνιση ροών από όμοια σύμβολα, σε μια ακολουθία πλειότιμων τ.μ. (Aki (1992) ή Fu and Lou (2003)).

Εύκολα μπορούμε να διαπιστώσουμε ότι οι οικογένειες των m ελάχιστων συνόλων διακοπής, είναι οι

$$\mathbf{C}_s = \{\{i, i+1, \ldots, i+k_s-1\} : i = 1, 2, \ldots, n-k_s+1\}, s = 1, 2, \ldots, m.$$

Υποθέτουμε ότι οι μονάδες, ή πιο σωστά οι τ.μ. που εκφράζουν τις καταστάσεις των μονάδων είναι ανεξάρτητες και ισόνομες τ.μ., με $q_{0i} = p, q_{si} = q_s, s = 1, 2, \ldots, m$. Επιπλέον, το σύμβολο $R_s(q)$ θα χρησιμοποιείται στο εξής για να δηλώνει την αξιοπιστία ενός απλού συστήματος, συνεχόμενα-k_s-από-τα-n, με την πιθανότητα αποτυχίας των ανεξάρτητων και ισόνομων μονάδων του, να είναι ίση με q. Όπως έχουμε ήδη αναφέρει, η $R_s(q)$ μπορεί να υπολογιστεί μέσω αναδρομικών τύπων, μέσων τεχνικών συνδυαστικής, εμφύτευσης τ.μ. σε Μαρκοβιανή αλυσίδα κ.α. (βλ. π.χ. Chao et al (1995), Balakrishnan and Koutras (2002), Fu and Lou (2003)). Για τις δικές μας ανάγκες, χρησιμοποιήσαμε την αναδρομική σχέση

$$R_{s,n}(q) = R_{s,n-1}(q) - (1-q)q^{k_s} R_{s,n-k_s-1}(q),$$

με αρχικές συνθήκες,

$$R_{s,n}(q) = 1, \text{ για } n < k_s,$$
$$R_{s,k}(q) = 1 - q^{k_s}, \text{ για } n = k_s$$

όπου με $R_{s,n'}(q)$, για οποιοδήποτε ακέραιο $n' \geq 1$, συμβολίσαμε την αξιοπιστία του συνεχόμενα-k_s-από-τα-n'.

Πίνακας 1.4.1: Ακριβή τιμή της αξιοπιστίας και σχετική βελτίωση

n	m	k_1	k_2	k_3	q_1	q_2	q_3	R	L	L'	$(L-L')/R(\%)$
100	2	2	2	–	0.08	0.05	–	0.4348	0.4186	0.3935	5.79
		2	3	–	0.05	0.05	–	0.7800	0.7786	0.7594	2.47
		3	5	–	0.10	0.06	–	0.9152	0.9151	0.8994	1.73
		5	4	–	0.20	0.15	–	0.9355	0.8844	0.8595	2.66
100	3	2	2	2	0.10	0.05	0.05	0.2502	0.2168	0.1705	18.52
		2	3	4	0.09	0.03	0.06	0.4811	0.4726	0.4068	13.68
		3	4	4	0.15	0.05	0.05	0.7559	0.7506	0.6793	9.43
		6	2	4	0.16	0.06	0.08	0.7104	0.6151	0.5544	8.55
250	3	2	3	2	0.06	0.05	0.02	0.3705	0.3648	0.3198	12.15
		2	2	2	0.10	0.05	0.05	0.0300	0.0214	0.0117	32.34
		6	6	6	0.20	0.15	0.15	0.9853	0.9515	0.8585	9.44
		5	4	5	0.12	0.12	0.12	0.9437	0.9052	0.8433	6.56

Κάτω απ' αυτές τις υποθέσεις, το νέο κάτω φράγμα (Θεώρημα 1.3.1) και το φράγμα (1.3.3), παίρνουν τη μορφή

$$L = L(q_1, q_2, \ldots, q_m) = \prod_{s=1}^{m} R_s(Q_s), \quad L' = L'(q_1, q_2, \ldots, q_m) = \prod_{s=1}^{m} R_s(Q'_s)$$

όπου

$$Q_s = \begin{cases} q_1, & \text{εάν } s = 1 \\ \frac{q_s}{1-\sum_{j=1}^{s-1} q_j}, & \text{εάν } s \geq 2 \end{cases} \quad \text{και } Q'_s = \frac{q_s}{p+q_s}, \text{ για } s = 1, 2, \ldots, m.$$

Όπως έχουμε επισημάνει στη συζήτηση που ακολούθησε μετά το Πόρισμα 1.3.1, το $L(\mathbf{q})$ είναι πάντα καλύτερο (μεγαλύτερο) από το $L'(\mathbf{q})$. Στον Πίνακα 1.4.1, μπορούμε να παρατηρήσουμε τη βελτίωση της προσέγγισης που επιτυγχάνουμε με το νέο φράγμα. Συγκεκριμένα, για διάφορες τιμές των παραμέτρων δίνεται η τιμή της αξιοπιστίας του συστήματος MFM, $R = R(q_1, q_2, \ldots, q_m)$ (για $m = 2$ έχουμε υπολογίσει την ακριβή της τιμή, ενώ για $m = 3$ την εκτίμηση μέσω προσομοίωσης), οι τιμές των L, L' και η σχετική βελτίωση της προσέγγισης $(L - L')/R$, όταν χρησιμοποιείται το L και όχι το L', για την προσέγγιση της R.

Επίσης στο Σχήμα 1.4.1, βλέπουμε πως συμπεριφέρεται το κάτω φράγμα L (και το L'), για διάφορες τιμές του p (έχουμε πάρει $q_1 = q_2 = (1-p)/2$ και $n = 300$), έχοντας συμπεριλάβει και την ακριβή τιμή της αξιοπιστίας R. Αξίζει να παρατηρήσουμε ότι όσο το k_2 αυξάνει (ενώ τα n και k_1, παραμένουν σταθερά), το L προσεγγίζει καλύτερα την R, και η απόσταση του από το L', μεγαλώνει. Γενικώς, το L προσεγγίζει καλύτερα την R, όταν το k_1 είναι αρκετά μικρότερο σε σχέση με το k_2 (οι προηγούμενες παρατηρήσεις, μπορούν να δικαιολογηθούν και με χρήση του τύπου που δόθηκε για το L).

Δυο επιπρόσθετα φράγματα για τη συνάρτηση αξιοπιστίας R, είναι τα παρακάτω (βλ. Boutsikas and Koutras (2002a))

$$L_B = L_B(\mathbf{q}_{1.}, \mathbf{q}_{2.}, \ldots, \mathbf{q}_{m.}) = \sum_{s=1}^{m} R_s(\mathbf{q}_{s.}) - m + 1,$$

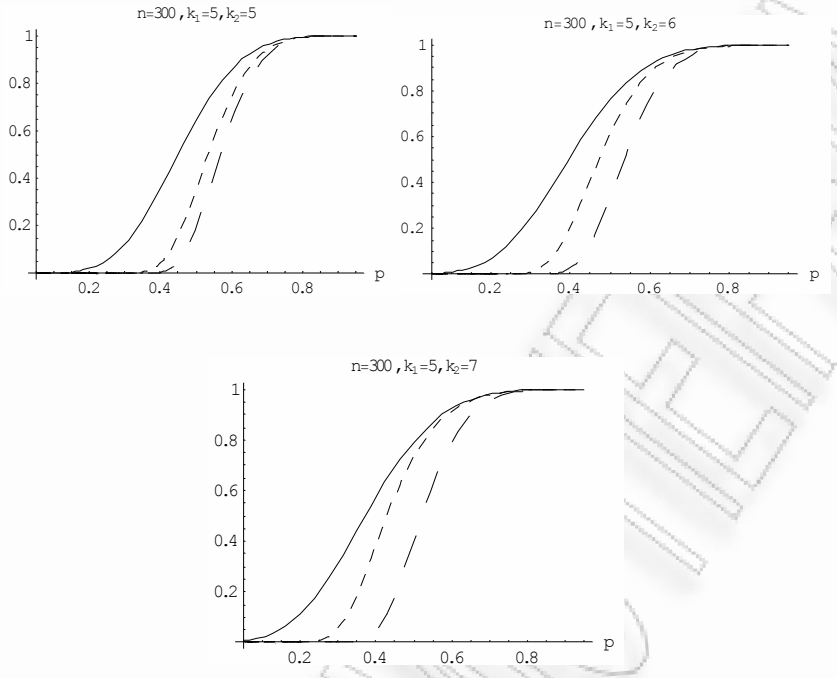
$$L_C = L_C(\mathbf{q}_{1.}, \mathbf{q}_{2.}, \ldots, \mathbf{q}_{m.}) = \prod_{s=1}^{m} R_s(\mathbf{q}_{s.}) - \sum_{1 \leq s < t \leq m} c_{st}$$

όπου

$$c_{st} = \sum_{C \in \mathbf{C}_s} \sum_{\substack{D \in \mathbf{C}_t \\ D \cap C \neq \emptyset}} \prod_{i \in C} q_{si} \prod_{j \in D} q_{tj}.$$

Το πρώτο φράγμα αποδεικνύεται με χρήση των ανισοτήτων Bonferroni, ενώ το δεύτερο, μέσω ενός γενικότερου αποτελέσματος που αφορά την απόσταση μεταξύ ενός αθροίσματος από συναφείς τ.μ. και ενός αθροίσματος από ανεξάρτητες τ.μ., με τις ίδιες περιθώριες κατανομές, με τις προηγούμενες (Boutsikas and Koutras (2000)).

Αξίζει να αναφέρουμε ότι αγνοώντας το άθροισμα από το L_C, παίρνουμε το παρακάτω



Σχήμα 1.4.1: Απεικόνιση της R και των L, L', για την περίπτωση: $m = 2$ και $q_1 = q_2 = (1-p)/2$

άνω φράγμα, για τη συνάρτηση αξιοπιστίας (Boutsikas and Koutras (2002a)),

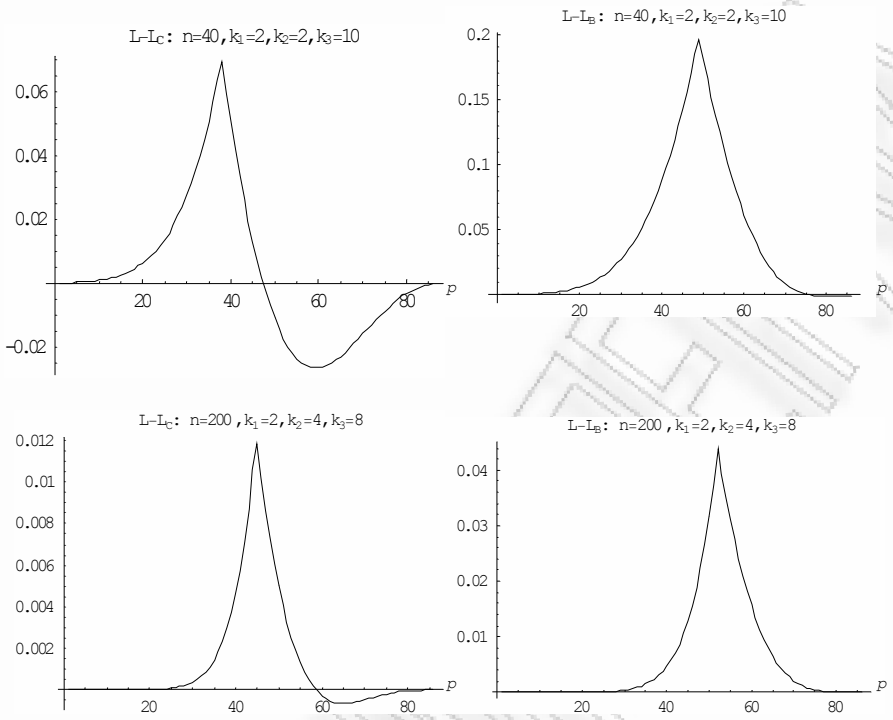
$$U_C = U_C(\mathbf{q}_{1.}, \mathbf{q}_{2.}, \ldots, \mathbf{q}_{m.}) = \prod_{s=1}^{m} R_s(\mathbf{q}_{s.}).$$

Το συγκεκριμένο φράγμα είναι ακριβώς της ίδιας μορφής με το νέο κάτω φράγμα L (αλλά και με το L'), εάν εξαιρέσουμε τις πιθανότητες αποτυχίας, οι οποίες είναι διαφορετικές. Ακόμη, επειδή ισχύει

$$L'(\mathbf{q}) \leq L(\mathbf{q}) \leq R(\mathbf{q}) \leq U_c(\mathbf{q})$$

για κάθε $\mathbf{q} = (q_{si})_{m\times n}$, και τα φράγματα $L'(\mathbf{q}), U_c(\mathbf{q})$ τείνουν ασυμπτωτικά στην ακριβή τιμή της αξιοπιστίας $R(\mathbf{q})$ (κάτω από συγκεκριμένες συνθήκες, Boutsikas and Koutras (2002a)), το νέο κάτω φράγμα $L(\mathbf{q})$, τείνει και αυτό στην ακριβή τιμή της $R(\mathbf{q})$ (κάτω από τις ίδιες συνθήκες). Επομένως, τα φράγματα $L(\mathbf{q})$ και $U_c(\mathbf{q})$ μπορούν να χρησιμοποιηθούν ταυτόχρονα, ώστε να πάρουμε διαστήματα (σ' αρκετές περιπτώσεις, μικρού μήκους) στα οποία ανήκει η $R(\mathbf{q})$.

Ένα ακόμη κάτω φράγμα για τη συνάρτηση αξιοπιστίας (όταν $\max_{1\leq s\leq m} q_s \leq 1/2$), το οποίο λαμβάνεται με προσέγγιση της κατανομής, μιας κατάλληλα ορισμένης απαριθμήτριας



Σχήμα 1.4.2: Απεικόνιση της $L - L_B$ και $L - L_C$, για την περίπτωση: $m = 3$ και $q_1 = q_2 = q_3 = (1 - p)/3$

τ.μ., από μια σύνθετη κατανομή Poisson, είναι το παρακάτω (Chryssaphinou and Vaggelatou (2002))

$$L_{CP}(\mathbf{q}) = e^{-l^*} - B_2 - \sum_{s=1}^{m}(k_s - 1)q_s^{k_s},$$

όπου $l^* = n\sum_{s=1}^{m}(1 - q_s)q_s^{k_s}$. Για τις λεπτομέρειες που αφορούν την ποσότητα B_2, ο αναγνώστης μπορεί να ανατρέξει στην αντίστοιχη εργασία.

Στους Πίνακες 1.4.2 και 1.4.3, παραθέτουμε αριθμητικές συγκρίσεις ανάμεσα στο νέο φράγμα και τα L_B, L_C, L_{CP}, για διάφορες τιμές των παραμέτρων n, k_1, k_2, k_3 και $q_s, s = 1, 2, 3$ (με έντονα γράμματα, σημειώνουμε τη μέγιστη τιμή, απ' όλα τα κάτω φράγματα). Μελετώντας τα αποτελέσματα, διαπιστώνουμε ότι τις περισσότερες φορές το νέο φράγμα είναι καλύτερο από τα υπόλοιπα, ειδικά στις περιπτώσεις, όπου μια τουλάχιστον πιθανότητα αποτυχίας, παίρνει σχετικά μεγάλες τιμές (το L_{CP} τις περισσότερες φορές είναι αρνητικό). Σε μία τέτοια κατάσταση (όπως η προηγούμενη), τα αποτελέσματα φαίνονται να γίνονται ακόμη καλύτερα, εάν στη διάταξη που χρησιμοποιούμε, τοποθετήσουμε πρώτη την οικογένεια, στην οποία αντιστοιχεί η υψηλότερη πιθανότητα αποτυχίας. Στο Σχήμα 1.4.2, υπάρ-

Πίνακας 1.4.2: Συνεχόμενα-2, 3, 4-από-τα-n:MFM

n	q_1	q_2	q_3	L	L_C	L_B	L_{CP}	L^*
100	0.01	0.28	0.30	**0.02549**	0.00237	−0.24982	−1.05747	0.00454
	0.16	0.01	0.02	**0.10472**	**0.10472**	0.10463	−0.52320	0.07672
	0.20	0.02	0.02	**0.03132**	0.03122	0.03059	−1.22554	0.01755
	0.15	0.08	0.10	**0.12348**	0.12334	0.08232	−0.40383	0.09421
	0.16	0.15	0.15	**0.05532**	0.03463	−0.18390	−0.6865	0.03530
	0.01	0.30	0.27	**0.02869**	0.00603	−0.19282	−1.22821	0.00624
	0.12	0.20	0.03	**0.10874**	0.09797	−0.20122	−0.38110	0.07475
	0.15	0.10	0.01	**0.11824**	0.11579	0.051478	−0.42038	0.08958
	0.13	0.18	0.13	**0.09815**	0.09084	−0.18356	−0.40014	0.06846
500	0.02	0.20	0.09	**0.02406**	0.02254	−0.16878	−0.24006	0.01069
	0.01	0.01	0.01	0.95130	**0.95131**	0.95129	0.95087	0.95083
	0.09	0.02	0.01	**0.02343**	0.02333	0.01965	−0.16570	0.01719
	0.08	0.04	0.04	**0.04861**	0.04821	0.01929	−0.09031	0.03889
	0.09	0.09	0.08	**0.01463**	0.00394	−0.27695	−0.19542	0.01020
	0.03	0.18	0.18	**0.01620**	0.00614	−0.61443	−0.18942	0.00685
	0.05	0.17	0.05	**0.02777**	0.01388	−0.57212	−0.15502	0.01626
	0.07	0.10	0.11	**0.05036**	0.04789	−0.32475	−0.06898	0.03980
	0.07	0.07	0.02	**0.08230**	0.08210	−0.04693	−0.01940	0.06969
1.000	0.02	0.15	0.15	**0.01315**	0.00773	−0.61831	−0.09490	0.00642
	0.06	0.03	0.01	**0.03203**	0.03181	0.00724	−0.04450	0.02637
	0.08	0.02	0.01	**0.00253**	0.00233	−0.00525	−0.1558	0.00162
	0.06	0.01	0.06	**0.03248**	0.03237	0.01999	−0.04377	0.02675
	0.08	0.05	0.02	**0.00219**	−0.00093	−0.10941	−0.15929	0.00140
	0.03	0.14	0.15	**0.01303**	0.00478	−0.83919	−0.08340	0.00695
	0.03	0.15	0.07	**0.01695**	0.00989	−0.54934	−0.09218	0.00956
	0.05	0.08	0.15	**0.02559**	0.02435	−0.63310	−0.04010	0.01872
	0.05	0.10	0.03	**0.03226**	0.02729	−0.50277	−0.03828	0.02556

χουν τα διαγράμματα των συναρτήσεων $L-L_B$ και $L-L_C$, για διάφορες τιμές του p (θέσαμε $q_1 = q_2 = q_3 = (1-p)/3$), απ' όπου διαπιστώνουμε ότι για τις περισσότερες περιπτώσεις, το νέο φράγμα υπερτερεί των υπολοίπων (στην πραγματικότητα, απεικονίσαμε γραφικά τις συναρτήσεις $L - \max(L_B, 0)$, $L - \max(L_C, 0)$).

Ο Πίνακας 1.4.4 περιέχει επιπλέον τις τιμές του $L = L(q_1, q_2, q_3)$, για το σύστημα συνεχόμενα-k_1, k_2, k_3 -από-τα-n: MFM, για την περίπτωση $k_1 = k_2 = k_3 = k$. Οι τιμές των παραμέτρων q_1, q_2, q_3, έχουν επιλεχθεί (για κάθε k, ξεχωριστά) έτσι ώστε να δημιουργούνται όλες οι πιθανές διατάξεις των τριών οικογενειών (έξι στο πλήθος). Επομένως, μπορούμε να πάρουμε μια ιδέα για το κατά πόσο η τιμή του κάτω φράγματος, αλλάζει από διάταξη σε διάταξη (ενώ, προφανώς η τιμή της αξιοπιστίας παραμένει αμετάβλητη). Παρατηρούμε, ότι και στις τρεις περιπτώσεις που εξετάσαμε, η μέγιστη τιμή επιτυγχάνεται, όταν οι οικογένειες διαταχθούν σε φθίνουσα σειρά, με βάση τις αντίστοιχες πιθανότητες αποτυχίας. Πρέπει επίσης να σημειώσουμε, πως το νέο φράγμα δεν απαιτεί καμία υπολογιστική διαδικασία παραπάνω, σε σχέση με τα L_B, L_C (και το L_{CP}). Ο υπολογισμός όλων των παραπάνω φραγμάτων,

Πίνακας 1.4.3: Συνεχόμενα-k_1, k_2, k_3-από-τα-5.000:MFM

q_1	q_2	q_3	k_1	k_2	k_3	L	L_C	L_B	L_{CP}	L^*
0.10	0.01	0.02	3	2	5	**0.00596**	0.00469	−0.37949	−0.02652	0.00363
			3	4	3	**0.01038**	0.01035	−0.02751	−0.01316	0.00636
			5	5	6	**0.95603**	**0.95603**	**0.95603**	0.95593	0.95127
0.02	0.09	0.03	2	8	7	**0.14064**	0.14064	0.14062	0.12959	0.13533
			2	3	6	**0.00414**	−0.00076	−0.82333	−0.03126	0.00282
			9	6	5	0.99707	**0.99747**	**0.99747**	0.99746	0.99679
0.03	0.02	0.04	2	4	4	**0.01241**	0.01238	−0.00038	−0.00618	0.01091
			2	3	5	**0.01207**	0.01197	−0.02631	−0.00987	0.01061
			4	6	9	**0.99608**	**0.99608**	**0.99608**	0.99608	0.99596

Πίνακας 1.4.4: Συνεχόμενα-k, k, k-από-τα-1.000:MFM

$k=3$				$k=4$				$k=5$			
q_1	q_2	q_3	L	q_1	q_2	q_3	L	q_1	q_2	q_3	L
0.10	0.05	0.01	**0.34491**	0.10	0.05	0.01	**0.90596**	0.10	0.05	0.01	**0.99058**
0.10	0.01	0.05	0.34314	0.10	0.01	0.05	0.90559	0.10	0.01	0.05	0.99055
0.05	0.10	0.01	0.31173	0.05	0.10	0.01	0.89090	0.05	0.10	0.01	0.98826
0.05	0.01	0.10	0.30190	0.05	0.01	0.10	0.88682	0.05	0.01	0.10	0.98765
0.01	0.10	0.05	0.33425	0.01	0.10	0.05	0.90236	0.01	0.10	0.05	0.99011
0.01	0.05	0.10	0.30088	0.01	0.05	0.10	0.88661	0.01	0.05	0.10	0.98764

προϋποθέτει την εύρεση της αξιοπιστίας απλών συνεχόμενων-k-από-τα-n, συστημάτων (για το παράδειγμα που μελετήσαμε), ενώ ακόμη περισσότερο, για το L_C χρειαζόμαστε και τον προσδιορισμό των όρων c_{st}. Βέβαια, με βάση το Πόρισμα 1.3.1, μπορούμε να δώσουμε ένα φράγμα, ακόμη πιο «ελκυστικό» υπολογιστικά (χάνοντας σε ακρίβεια προσέγγισης), με τη μορφή

$$L^* = \prod_{s=1}^{m} (1 - Q_s^{k_s})^{n-k_s+1}.$$

Στους υπολογισμούς του Πίνακα 1.4.2 και 1.4.3 έχουμε συμπεριλάβει στις τελευταίες στήλες και το L^*, έτσι ώστε να διαπιστώσουμε πόσο πολύ χάνουμε σε ακρίβεια (σε σχέση με τη χρησιμοποίηση του L) και να συγκρίνουμε το L^*, με τα L_B, L_C. Σε μερικές περιπτώσεις, όπως οι πίνακες φανερώνουν, το L^* παραμένει και αυτό καλύτερο από τα L_B, L_C. Αξίζει να αναφέρουμε ότι το L και το L^*, παίρνουν μόνο θετικές τιμές, σε αντίθεση με τα L_B και L_C, τα οποία για συγκεκριμένες επιλογές παραμέτρων, μπορούν να πάρουν και αρνητικές τιμές. Φυσικά στην περίπτωση αυτή, θα θεωρούμε ότι είναι ίσα με μηδέν, μη δίνοντάς μας καμία πληροφορία για την αξιοπιστία του συστήματος.

1.4.2 Σύστημα CCS, με πολλαπλά επίπεδα αποτυχίας

Στο Παράδειγμα 1.2 είχαμε γενικεύσει το σύστημα CCS, σε περιβάλλον μονάδων με πολλαπλά επίπεδα αποτυχίας. Επίσης, αναφέρθηκε ότι ο τρόπος που ορίζουμε την αποτυχία του συστήματος, διαφέρει απ' αυτόν που πολλοί συγγραφείς είχαν χρησιμοποιήσει, για ανάλογο σκοπό (βλ. Kuo and Zuo (2003)).

Υπενθυμίζουμε ότι, το σύστημα αυτό αποτελείται από ένα «πομπό», ένα «δέκτη» και n μονάδες $I = \{1, 2, \ldots, n\}$, συνδεδεμένες σε μία σειρά. Ο πομπός είναι συνδεδεμένος (επικοινωνεί) με τις μονάδες $\{1, 2, \ldots, \epsilon_0\}$, και μια μονάδα $i \in I$, επικοινωνεί με τις μονάδες $\{i+1, i+2, \ldots, i+\epsilon_i\}$, $1 \leq \epsilon_i \leq n, i = 1, 2, \ldots, n$. Κάθε μονάδα μπορεί να είναι σε κατάσταση λειτουργίας ή σε κατάσταση αποτυχίας τύπου $s, (s \in S = \{1, 2, \ldots, m\})$.

Το σύστημα αποτυγχάνει, εάν και μόνο εάν, σε κάθε πιθανή διαδρομή από τον πομπό στο δέκτη, υπάρχουν περισσότερες από n_s μονάδες (δηλαδή, $n_s + 1, n_s + 2, \ldots, n$), σε κατάσταση αποτυχίας τύπου s, για κάποιο $s \in S$. Εάν $n_s = 0$ για κάθε s, τότε η αξιοπιστία του συστήματος, ταυτίζεται μ' αυτή του απλού CCS. Επιπλέον, εάν $\epsilon_i = 1, i = 0, 1, \ldots, n$ και $n_s = k_s$ για κάθε $s \in S$, τότε προκύπτει το σύστημα του Παραδείγματος 1.1, το οποίο αποτυγχάνει εάν και μόνο εάν ανάμεσα στις n μονάδες, υπάρχουν τουλάχιστον k_s, σε κατάσταση αποτυχίας τύπου s (η γενίκευση του απλού συστήματος k-από-τα-n).

Αξίζει να σημειώσουμε ότι στο απλό CCS, τα ελάχιστα σύνολα διακοπής του συστήματος, ανήκουν όλα στην οικογένεια $\mathbf{C}^* = \{C_j^* : j = \epsilon_0, \epsilon_0 + 1, \ldots, n\}$, όπου

$$C_j^* = \{i : i \leq j \text{ και οι μονάδες } i \text{ και } j+1 \text{ είναι άμεσα/απευθείας συνδεδεμένες}\}.$$

Ουσιαστικά, το σύνολο C_j^* αποτελείται απ' όλες τις μονάδες, οι οποίες «επικοινωνούν» άμεσα, με την $j+1$. Επομένως, τα ελάχιστα σύνολα διακοπής, στο απλό CCS, μπορούν εύκολα να αναζητηθούν, ανάμεσα στα C_j^*.

Ας πάρουμε την περίπτωση όπου το σύστημα αποτελείται από $n = 7$ ανεξάρτητες και ισόνομες μονάδες, με $m = 3$, $n_1 = 0, n_2 = 1, n_3 = 2$ και $\epsilon_0 = 2, \epsilon_1 = 3, \epsilon_2 = 1, \epsilon_3 = \epsilon_4 = \epsilon_5 = \epsilon_6 = 2, \epsilon_7 = 1$ (βλ. Σχήμα 1.4.3).

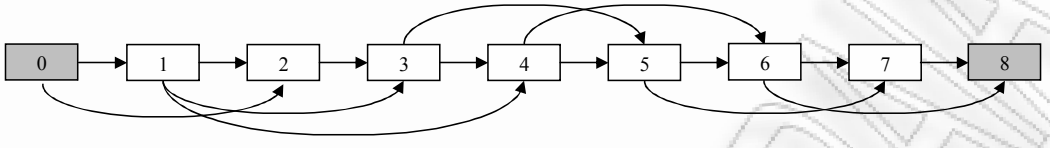
Εξετάζοντας το σύστημα, οι οικογένειες ελάχιστων συνόλων διακοπής είναι οι εξής

$\mathbf{C}_1 = \{\{1,2\},\{1,3\},\{3,4\},\{4,5\},\{5,6\},\{6,7\}\}$,

$\mathbf{C}_2 = \{\{1,2,3,4\},\{1,2,4,5\},\{1,2,5,6\},\{1,2,6,7\},\{1,3,4,5\},\{1,3,5,6\},\{1,3,6,7\},$
$\{3,4,5,6\},\{3,4,6,7\},\{4,5,6,7\}\}$,

$\mathbf{C}_3 = \{\{1,2,3,4,5,6\},\{1,2,3,4,6,7\},\{1,2,4,5,6,7\},\{1,3,4,5,6,7\}\}$.

Σχήμα 1.4.3: Σύστημα CCS με: $\epsilon_1 = 3, \epsilon_0 = \epsilon_3 = \epsilon_4 = \epsilon_5 = \epsilon_6 = 2, \epsilon_2 = \epsilon_7 = 1$

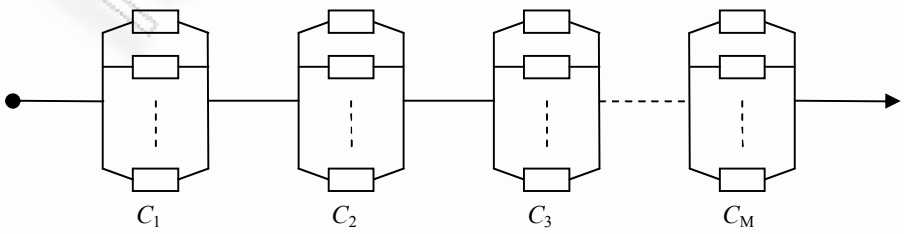


Η οικογένεια συνόλων $\mathbf{C}_1$ είναι ίδια με το σύνολο των ελάχιστων συνόλων διακοπής του απλού CCS (τα οποία με τη σειρά τους ταυτίζονται με το $\mathbf{C}^*$). Η οικογένεια $\mathbf{C}_2$, είναι στην πραγματικότητα όλες οι ανά δυο ενώσεις, των συνόλων της $\mathbf{C}_1$, οι οποίες δεν έχουν κοινά στοιχεία. Όμοια, η $\mathbf{C}_3$, είναι οι ανά τρεις ενώσεις των συνόλων της $\mathbf{C}_1$, οι οποίες δεν έχουν κοινά στοιχεία. Είναι φανερό ότι η γνώση των ελάχιστων συνόλων διακοπής του απλού συστήματος, μας οδήγησε στον προσδιορισμό των τριών οικογενειών $\mathbf{C}_1, \mathbf{C}_2$ και $\mathbf{C}_3$. Αυτό δεν αποτελεί τυχαίο γεγονός, καθώς ισχύει το παρακάτω.

Πόρισμα 1.4.1 *Έστω ένα MFM σύστημα CCS, με n μονάδες, οπού η λειτουργία του προσδιορίζεται από τις παραμέτρους n_s, ϵ_s, για $s = 1, 2, \ldots, m$. Τότε, ένα σύνολο ανήκει στην οικογένεια $\mathbf{C}_s$, εάν και μόνο εάν, γράφεται ως ένωση n_s+1, ξένων ανά δυο, ελάχιστων συνόλων διακοπής του απλού συστήματος.*

Απόδειξη. Για την κατανόηση της απόδειξης, θα βοηθήσει εάν έχουμε υπόψιν μας, την «ιδεατή» αναπαράσταση ενός απλού συστήματος. Για να γίνουμε πιο συγκεκριμένοι, ας υποθέσουμε ότι ένα απλό σύστημα έχει M ελάχιστα σύνολα διακοπής. Τότε, για κάθε ελάχιστο σύνολο διακοπής, κατασκευάζουμε ένα υποσύστημα, συνδέοντας παράλληλα τις μονάδες που ανήκουν σ' αυτό. Στη συνέχεια, τα M παράλληλα υποσυστήματα τα συνδέουμε σε σειρά (Σχήμα 1.4.4), και προφανώς, το σύστημα που προκύπτει μ' αυτόν τον τρόπο, έχει την ίδια αξιοπιστία με το αρχικό.

Σχήμα 1.4.4: Ιδεατή αναπαράσταση ενός απλού συστήματος



Να θυμίσουμε ότι ένα απλό σύστημα αποτυγχάνει εάν και μόνο εάν, ένα τουλάχιστον ελάχιστο σύστημα διακοπής, έχει όλες τις μονάδες του σε κατάσταση αποτυχίας.

Είναι εύκολο να διαπιστωθεί ότι, κάθε σύνολο C που ανήκει στην οικογένεια $\mathbf{C}_s$, γράφεται ως ένωση κάποιων (έστω στο πλήθος $\rho \geq n_s + 1$) ελάχιστων συνόλων διακοπής. Εάν αυτά δεν είναι ξένα ανά δυο, τότε υπάρχουν δυο τουλάχιστον ελάχιστα σύνολα διακοπής, με κοινές μονάδες. Τότε η ύπαρξη και των δυο στην ένωση, δεν προσφέρει παραπάνω αποτυχίες (τύπου s), σε κάθε πιθανή διαδρομή.

Επομένως, αφαιρώντας από το C, ένα από τα σύνολα που έχουν τομή διάφορη του κενού, εξακολουθεί να προκαλεί τον ίδιο αριθμό αποτυχιών τύπου s, σε κάθε διαδρομή. Συνεχίζοντας με τα ίδια επιχειρήματα, καταλήγουμε ότι ένα σύνολο C για να ανήκει στην οικογένεια $\mathbf{C}_s$, πρέπει να γράφεται ως ένωση $n_s + 1$, ξένων ανά δυο ελάχιστων συνόλων διακοπής του απλού συστήματος (ασφαλώς, δεν μπορεί να είναι ένωση λιγότερων από n_s+1).

■

Πίνακας 1.4.5: Κάτω φράγμα για τη συνάρτηση αξιοπιστίας του CCS με: $n_1 = 0, n_2 = 1, n_3 = 2$ και $\epsilon_1 = 3, \epsilon_2 = \epsilon_7 = 1, \epsilon_0 = \epsilon_3 = \epsilon_4 = \epsilon_5 = \epsilon_6 = 2$.

q_1	q_2	q_3	L	L^*	q_1	q_2	q_3	L	L^*	q_1	q_2	q_3	L	L^*
0.05	0.05	0.05	0.98559	0.98502	0.05	0.05	0.30	0.98153	0.97962	0.07	0.60	0.25	0.10809	0.00335
0.10	0.05	0.05	0.94543	0.94139	0.05	0.30	0.05	0.92319	0.89140	0.07	0.55	0.25	0.19246	0.02219
0.05	0.10	0.01	0.98460	0.98389	0.30	0.05	0.05	0.62377	0.56772	0.01	0.50	0.30	0.19066	0.02785
0.20	0.02	0.02	0.80691	0.78276	0.10	0.10	0.10	0.94425	0.94003	0.05	0.60	0.22	0.15797	0.00996
0.20	0.10	0.10	0.80521	0.78082	0.20	0.20	0.20	0.78132	0.74859	0.08	0.50	0.30	0.18848	0.02739
0.10	0.20	0.20	0.92685	0.91677	0.30	0.30	0.30	0.35425	0.18397	0.01	0.59	0.25	0.11023	0.00425

Το νέο κάτω φράγμα $L(\mathbf{q})$, για τη συνάρτηση αξιοπιστίας του παραπάνω συστήματος ($n_1 = 0, n_2 = 1, n_3 = 2$ και $\epsilon_1 = 3, \epsilon_2 = \epsilon_7 = 1, \epsilon_0 = \epsilon_3 = \epsilon_4 = \epsilon_5 = \epsilon_6 = 2$), θα έχει τη μορφή

$$L = L(q_1, q_2, q_3) = R_1(Q_1)R_2(Q_2)R_3(Q_3),$$

όπου

$$R_1(Q_1) = R_1(q_1) = E\left(\prod_{C\in\mathbf{C}_1}(1 - \prod_{i\in C}(1 - T_{1i}))\right) = 1 - 6q_1^2 + 5q_1^3 + 6q_1^4 - 9q_1^5 + 3q_1^6$$

$$R_2(Q_2) = R_2\left(\frac{q_2}{1-q_1}\right) = 1 - \frac{10q_2^4}{(1-q_1)^4} + \frac{12q_2^5}{(1-q_1)^5} - \frac{q_2^6}{(1-q_1)^6} - \frac{2q_2^7}{(1-q_1)^7}$$

$$R_3(Q_3) = R_3\left(\frac{q_3}{1-q_1-q_2}\right) = 1 - \frac{4q_3^6}{(1-q_1-q_2)^6} + \frac{3q_3^7}{(1-q_1-q_2)^7},$$

ενώ

$$L^* = (1-q_1^2)^6 \left(1-\left(\frac{q_2}{1-q_1}\right)^4\right)^{10} \left(1-\left(\frac{q_3}{1-q_1-q_2}\right)^6\right)^4 .$$

Ο Πίνακας 1.4.5 περιέχει τις τιμές των δυο παραπάνω φραγμάτων, για διαφορές τιμές των πιθανοτήτων αποτυχίας. Αξιοπρόσεχτο είναι ότι στις περισσότερες περιπτώσεις, οι διαφορές ανάμεσα στα δύο φράγματα, είναι πολύ μικρές (να θυμίσουμε ότι $L(\mathbf{q}) \geq L^*(\mathbf{q})$, για κάθε $\mathbf{q}$).

Κεφάλαιο 2

Οριακά αποτελέσματα για τις συναρτήσεις σάρωσης

Η διακριτή συνάρτηση σάρωσης $S_{n,k}$, ορισμένη σε μια ακολουθία από n δίτιμες δοκιμές (1:επιτυχία, 0:αποτυχία) εκφράζει το μέγιστο αριθμό επιτυχιών, ανάμεσα σε κάθε k συνεχόμενες δοκιμές (όπου n και k είναι θετικοί ακέραιοι αριθμοί, με $k \leq n$). Η $S_{n,k}$, έχει προκαλέσει έντονο ερευνητικό ενδιαφέρον τις τελευταίες δεκαετίες, με αφορμή τις εφαρμογές που παρουσιάζει σε διάφορα επιστημονικά πεδία, όπως στον έλεγχο ποιότητας, στη θεωρία αξιοπιστίας, στην ασφαλιστική επιστήμη, τη βιολογία, την πολυμεταβλητή ανάλυση δεδομένων κ.α. (βλ. π.χ. Naus (1974), Huntington and Naus (1975), Glaz and Balakrishnan (1999), Glaz, Naus and Wallenstein (2001) και Balakrishnan and Koutras (2002)).

Μία από τις περιπτώσεις όπου συναντάμε την $S_{n,k}$ είναι στους ελέγχους τυχαιότητας (randomness tests), και συγκεκριμένα, όταν σε μια ακολουθία από δίτιμες τ.μ. X_i, $i = 1, 2, \ldots, n$, θέλουμε να ελέγξουμε την υπόθεση ότι οι X_i είναι ανεξάρτητες και ισόνομες τ.μ., με $P(X_i = 1) = 0.5$. Η υπόθεση αυτή μπορεί να παραβιάζεται, εάν υπάρχει ένα είδος εξάρτησης ανάμεσα στα X_i ή ασφαλώς, όταν $P(X_i = 1) \neq 0.5$, για κάποια υπακολουθία της $X_1, X_2, \ldots, X_n$. Οι Glaz and Naus (1991) είχαν αποδείξει ότι ο έλεγχος που προκύπτει με χρήση του γενικευμένου λόγου πιθανοφανειών (generalized likelihood ratio test), απορρίπτει την υπόθεση των ισοπίθανων αποτελεσμάτων ($P(X_i = 1) = 0.5$), όταν $S_{n,k} \geq c$, όπου το c προσδιορίζεται από το επιθυμητό επίπεδο σημαντικότητας του ελέγχου (το σφάλμα τύπου I). Επομένως, ο υπολογισμός της παραμέτρου c, απαιτεί την εύρεση της ακριβούς κατανομής της $S_{n,k}$. Επειδή τις περισσότερες φορές οι έλεγχοι τυχαιότητας, εφαρμόζονται σε μεγάλα σύνολα δεδομένων (με σκοπό, π.χ. να μειωθεί το σφάλμα τύπου II), τα θεωρητικά αποτελέσματα που σχετίζονται με την ασυμπτωτική συμπεριφορά της $S_{n,k}$ (καθώς τα $n, k \rightarrow$

∞), προβάλουν εξίσου σημαντικά και χρήσιμα.

Επίσης, στην ασφαλιστική επιστήμη για τις n καθημερινές απαιτήσεις ενός χαρτοφυλάκιου, μπορούμε (για παράδειγμα) να χρησιμοποιήσουμε τις δίτιμες τ.μ. X_i, $i = 1, 2, \ldots, n$, όπου η X_i θα παίρνει την τιμή ένα εάν η απαίτηση υπερβαίνει κάποιο όριο u («κατώφλι», threshold), και μηδέν ($X_i = 0$) διαφορετικά. Τότε, η $S_{n,k}$ θα περιγράφει το μέγιστο αριθμό από «υψηλές» απαιτήσεις (δηλαδή, απαιτήσεις οι οποίες υπερβαίνουν το κατώφλι u), σε περιόδους k συνεχόμενων ημερών. Ασφαλώς, είναι έκδηλο το ενδιαφέρον για τη συμπεριφορά της $S_{n,k}$, όχι μόνο στο άμεσο «μέλλον», αλλά και σε μακροχρόνιο ορίζοντα (όταν το n και το k, τείνουν στο ∞).

Στη θεωρία αξιοπιστίας, τα συστήματα συνεχόμενα-k-από-τα-n αποτελούνται από n μονάδες συνδεδεμένες σε μια σειρά, οι οποίες μπορούν είτε να λειτουργούν, είτε να βρίσκονται σε κατάσταση αποτυχίας. Ολόκληρο το σύστημα αποτυγχάνει εάν και μόνο εάν, k τουλάχιστον συνεχόμενες μονάδες είναι σε κατάσταση αποτυχίας (βλ. π.χ. Chao et al (1995)). Η γενίκευση των παραπάνω συστημάτων, τα r-μεταξύ-k-συνεχόμενων-από-τα-n ($r \leq k \leq n$), αποτυγχάνουν εάν ανάμεσα σε k συνεχόμενες μονάδες (από τις n), βρεθούν τουλάχιστον r, σε κατάσταση αποτυχίας (Griffith (1986)). Είναι εύκολο να διαπιστώσουμε ότι στην πρώτη περίπτωση (στα συνεχόμενα-k-από-τα-n) το σύστημα αποτυγχάνει εάν και μόνο εάν $S_{n,k} = k$, ενώ στη δεύτερη (στα r-μεταξύ-k-συνεχόμενων-από-τα-n), εάν ισχύει $S_{n,k} \geq r$.

Είναι γνωστό ότι μια αλυσίδα DNA μπορεί να θεωρηθεί ως μια ακολουθία από τ.μ., με τέσσερα δυνατά αποτελέσματα. Ένα από τα προβλήματα που αντιμετωπίζουν στο χώρο της βιολογίας, είναι η σύγκριση δυο αλυσίδων DNA, από διαφορετικά είδη οργανισμών, με σκοπό την εύρεση γενετικών ομοιοτήτων (βλ. π.χ. Arratia, Gordon and Waterman (1990), Goldstein and Waterman (1992)). Έτσι, αυτό που μας ενδιαφέρει είναι ο εντοπισμός μεγάλων τμημάτων από τις αλυσίδες, στις οποίες έχουμε ταύτιση ή σχεδόν ταύτιση. Όμως, για κάθε σύγκριση μεταξύ δυο αλυσίδων, μπορούμε να δημιουργήσουμε μια νέα ακολουθία με αποτελέσματα 1 ή 0, ανάλογα με το εάν στην αντίστοιχη θέση οι δυο αλυσίδες ταυτίζονται ή όχι. Έτσι, αυτό που αναζητούμε μεταφράζεται, στην εύρεση υπακολουθιών με πολλές επιτυχίες (επιτυχία είναι η ταύτιση των αλυσίδων, σε κάποιο συγκεκριμένο σημείο), ώστε να εξετάσουμε τυχόν γενετικές ομοιότητες. Εν τέλει, ασχολούμαστε με τις τιμές που θα πάρει η $S_{n,k}$, και τις αντίστοιχες πιθανότητες, κάτω από διάφορες προϋποθέσεις.

Ένα άλλο επιστημονικό πεδίο, από το οποίο αναδύεται η χρήση των στατιστικών συναρτήσεων σάρωσης, είναι και ο ποιοτικός έλεγχος. Για παράδειγμα, είναι πολύ δημοφιλής τις τελευταίες δεκαετίες, η μελέτη των διαγραμμάτων ελέγχου μιας διεργασίας, με τη χρήση κανόνων ροών ή συναρτήσεων σάρωσης (Page (1955), Champ and Woodall (1987), Klein

(2000), Koutras et al (2006), Rakitzis (2008)). Οι κανόνες αυτοί έχουν ως κύριο στόχο τους την ευαισθητοποίηση των διαγραμμάτων, για την εύρεση μικρών αλλαγών στις τιμές των παραμέτρων μιας διεργασίας (π.χ. της μέσης τιμής ή της διασποράς, του χαρακτηριστικού που μελετάμε/ελέγχουμε). Συγκεκριμένα, έχοντας προσδιορίσει τα προειδοποιητικά όρια ελέγχου ενός διαγράμματος, αποφασίζεται ότι η διεργασία είναι εκτός ελέγχου (δηλαδή, κάποια από τις παραμέτρους που μας ενδιαφέρουν, έχει αλλάξει τιμή, χωρίς αυτό να έπρεπε να συμβεί), εάν εντοπίσουμε k συνεχόμενα σημεία πάνω από κάποιο προειδοποιητικό όριο, ή βρούμε r μεταξύ k συνεχόμενων μετρήσεων, πάνω από κάποιο άλλο όριο.

Ακριβή αποτελέσματα για την κατανομή της συνάρτησης σάρωσης, μπορούμε να βρούμε στις εργασίες (μεταξύ άλλων) Naus (1974), Fu (2001) ή στα βιβλία των Balakrishnan and Koutras (2002) και Fu and Lou (2003). Οι μέθοδοι που έχουν προταθεί γι αυτό το σκοπό (όπως γίνεται κατανοητό και μέσα από τη μελέτη της προηγούμενης βιβλιογραφίας), καθίστανται πρακτικά μη χρήσιμες, στις περιπτώσεις όπου τα n, k πάρουν μεγάλες τιμές. Για το λόγο αυτό, οι προσεγγίσεις και τα φράγματα, έχουν αποσπάσει την προσοχή της ερευνητικής κοινότητας, τις τελευταίες δεκαετίες. Έτσι πολλοί συγγραφείς, μέσα από διάφορες μεθόδους επιχείρησαν να δώσουν προσεγγίσεις για τη συνάρτηση κατανομής της $S_{n,k}$, όπως πολλαπλασιαστικού τύπου προσεγγίσεις, φράγματα τύπου Bonferroni, προσεγγίσεις μέσω της θεωρίας των martingales, προσεγγίσεις μέσω κατανομών Poisson κ.α. (βλ. π.χ. Glaz and Balakrishnan (1999), Glaz, Naus and Wallenstein (2001), Pozdnyakov et al (2005)).

Μια άλλη τ.μ. άμεσα συνδεδεμένη με την $S_{n,k}$, είναι η απαριθμήτρια των k συνεχόμενων δοκιμών (ή αλλιώς, των παραθύρων μήκους k), με τουλάχιστον r επιτυχίες, ανάμεσά τους. Εάν σε μια ακολουθία από n δίτιμες τ.μ., συμβολίσουμε με $W_{n,k,r}$ την παραπάνω απαριθμήτρια, τότε εύκολα μπορούμε να διαπιστώσουμε την ισχύ της παρακάτω σχέσης

$$P(W_{n,k,r} = 0) = P(S_{n,k} < r).$$

Για τη στατιστική συνάρτηση $W_{n,k,r}$, χρησιμοποιείται η ονομασία *πολλαπλή συνάρτηση σάρωσης* (multiple scan statistic). Οι εφαρμογές στις οποίες μπορούμε να συναντήσουμε την πολλαπλή συνάρτηση σάρωσης είναι παρόμοιες μ' αυτές της $S_{n,k}$ («απλή» συνάρτηση σάρωσης), όπου ουσιαστικά βελτιώνουμε ή επεκτείνουμε τα αντίστοιχα αποτελέσματα.

Οι Balakrishnan and Koutras (2002) εισήγαγαν και δύο άλλες τ.μ. οι οποίες αναφέρονται στην καταμέτρηση των παραθύρων (μήκους k), με r τουλάχιστον επιτυχίες. Η διαφοροποίηση ανάμεσά τους, έγκειται στον τρόπο με τον οποίο επιλέγουμε τα παράθυρα που μας ενδιαφέρουν-για παράδειγμα, αυτά μπορεί να είναι επικαλυπτόμενα ή μη. Έτσι, χρησιμοποίησαν τους όρους καταμέτρηση «τύπου ΙΙΙ», για την περίπτωση που ενδιαφερόμαστε για επικαλυπτόμενα παράθυρα (όπως συμβαίνει στο συγκεκριμένο κεφάλαιο) και καταμέτρηση

«τύπου Ι» και «τύπου ΙΙ», για κατάλληλα ορισμένες διαδικασίες απαρίθμησης, σε παράθυρα χωρίς κοινά σημεία (τ.μ.).

Για την προσέγγιση της συνάρτησης πιθανότητας της $W_{n,k,r}$ στο σημείο μηδέν, υπάρχουν στη βιβλιογραφία αρκετά αποτελέσματα, προσφέροντας ακριβείς προσεγγίσεις και χρήσιμα φράγματα (βλ. π.χ. Chen and Glaz (1999)). Όταν όμως εστιάσουμε στις προσεγγίσεις που αφορούν ολόκληρη την κατανομή της $W_{n,k,r}$, τότε διαπιστώνουμε ότι οι καταστάσεις γίνονται εξαιρετικά πολύπλοκες, και τα διαθέσιμα αποτελέσματα, δεν προσφέρουν (τις περισσότερες φορές) τα αναμενόμενα. Οι Koutras and Alexandrou (1995) παρουσίασαν μία μέθοδο βασισμένη στην εμφύτευση τ.μ. σε Μαρκοβιανή αλυσίδα, με την οποία υπολογίζεται η ακριβής κατανομή της $W_{n,k,r}$. Δυστυχώς, εύκολα κάποιος μπορεί να διαπιστώσει ότι για μεγάλες τιμές των παραμέτρων k, r και n, και η παραπάνω διαδικασία, είναι μη εφαρμόσιμη, λόγω των υπολογιστικών της απαιτήσεων. Αυτό αποτελεί ένα ακόμη στοιχείο, που συνηγορεί στην εύρεση και τη μελέτη των ασυμπτωτικών ιδιοτήτων της $W_{n,k,r}$, κάτω από ρεαλιστικές (γενικές) συνθήκες, με στόχο την κάλυψη όσο το δυνατόν περισσότερων περιπτώσεων.

Στη βιβλιογραφία έχουν εμφανισθεί και διάφορα αλλά αποτελέσματα, που στη πραγματικότητα αφορούν γενικεύσεις, των συναρτήσεων σάρωσης. Για παράδειγμα, έχουμε καταλάβει από τα προηγούμενα, ότι σε μια ακολουθία από ανεξάρτητες και ισόνομες δίτιμες τ.μ. $X_1, X_2, \ldots, X_n$, η συνάρτηση $S_{n,k}$ ορίζεται ως ο μέγιστος αριθμός επιτυχιών σε παράθυρα μήκους k, ή ισοδύναμα, η μέγιστη τιμή από τα αθροίσματα, $X_i + X_{i+1} + \ldots + X_{i+k-1}, i = 1, 2, \ldots, n-k+1$. Τι γίνεται όμως, εάν θεωρήσουμε ότι οι τ.μ. X_i, $i = 1, 2, \ldots, n$ είναι τοποθετημένες σ' ένα κύκλο (δηλαδή, η X_1 με την X_n είναι γειτονικές, Chen and Glaz (1999)), ή σ' ένα ορθογώνιο (βλ. π.χ. Chen and Glaz (1996)); Προφανώς ανακύπτουν παρόμοια προς μελέτη ζητήματα, μ' αυτά που έχουμε ήδη αναφέρει για τις απλές συναρτήσεις σάρωσης. Ενδιαφέρον παρουσιάζει και η περίπτωση όπου μπορούμε να θεωρήσουμε ότι το μήκος του παράθυρου μπορεί να πάρει διάφορες τιμές -όχι μόνο μια συγκεκριμένη τιμή- και ανάμεσα στις επιλογές αυτές, να αξιολογήσουμε τις αντίστοιχες τιμές των $S_{n,k}$ και $W_{n,k,r}$ (π.χ. να εντοπίσουμε αυτή που φαίνεται λιγότερο αναμενόμενη, κάτω από τις προϋποθέσεις μας, δείτε Loader (1991) και Kulldorff (1997)).

Στο συγκεκριμένο κεφάλαιο, θα παρουσιάσουμε αρχικά όλα τα αποτελέσματα που συναντάμε στη βιβλιογραφία, και αφορούν την προσέγγιση των συναρτήσεων σάρωσης (της $S_{n,k}$ και $W_{n,k,r}$), μέσα από διαδικασίες που προσφέρουν ταυτόχρονα και φράγματα, για τα σφάλματα των αντίστοιχων προσεγγίσεων. Αυτό, όπως θα διαπιστώσουμε, επιτυγχάνεται μέσα από την προσέγγιση της κατανομής των συναρτήσεων σάρωσης, από κατανομές Poisson (απλές ή σύνθετες), και την εύρεση φραγμάτων, για τις αντίστοιχες αποστάσεις που

χρησιμοποιούνται για να εκτιμήσουμε την εγγύτητα μεταξύ κατανομών.

Στη συνέχεια, εισάγουμε τις συναρτήσεις σάρωσης κάτω από ένα γενικότερο μοντέλο, βασισμένο σε μια ακολουθία ανεξάρτητων και ισόνομων συνεχών τυχαίων μεταβλητών. Κάτω από κατάλληλες συνθήκες θα εξετάσουμε την ασυμπτωτική συμπεριφορά των υπερβάσεων (πάνω από ένα κατώφλι) της παραπάνω ακολουθίας, σε κινούμενα-επικαλυπτόμενα παράθυρα. Βασικό ρόλο στα προηγούμενα αποτελέσματα, θα παίξει και η θεωρία των ακραίων τιμών (extreme value theory), καθώς κάτω από την υπόθεση ότι η συνάρτηση κατανομής των τ.μ. ανήκει σε κάποιο από τα μέγιστα πεδία έλξης (maximum domain of attraction) των κατανομών ακραίων τιμών, εξασφαλίζουμε κάποιες βασικές προϋποθέσεις για την ανάπτυξη χρήσιμων οριακών αποτελεσμάτων. Ενδιαφέρον θα παρουσιάσει και η σύνδεση των νέων (γενικευμένων) συναρτήσεων σάρωσης, με τις κινούμενες διατεταγμένες παρατηρήσεις (moving order statistics, βλ. π.χ. David and Rogers (1983)). Ως ειδική περίπτωση του γενικότερου μοντέλου, προκύπτουν οι κλασικές συναρτήσεις σάρωσης, $S_{n,k}$ και $W_{n,k,r}$, ορισμένες σε ακολουθίες από δίτιμες τ.μ.

Έτσι, στην Παράγραφο 2.1 παρουσιάζουμε κάποια βασικά στοιχεία από τη θεωρία (κατανομές Poisson, αποστάσεις μεταξύ κατανομών, προσεγγίσεις από κατανομές Poisson και θεωρία ακραίων τιμών), τα οποία θα μας βοηθήσουν στην ανάπτυξη των αποτελεσμάτων που ακολουθούν. Στη συνέχεια (Παράγραφος 2.2), εισάγουμε τους απαραίτητους συμβολισμούς και τις βασικές ιδιότητες, των συναρτήσεων σάρωσης. Στην Παράγραφο 2.3, θα συμπεριλάβουμε όλα τα αποτελέσματα που υπάρχουν στη βιβλιογραφία, και αφορούν την προσέγγιση των $S_{n,k}$ και $W_{n,k,r}$, μέσω κατανομής Poisson (απλής ή σύνθετης), τα οποία ταυτόχρονα προσφέρουν, και φράγματα για τα σφάλματα της προσέγγισης. Τέλος, στην Παράγραφο 2.4 μελετούμε τις γενικευμένες συναρτήσεις σάρωσης, και δίνουμε κάποια νέα αποτελέσματα μαζί με αριθμητικούς υπολογισμούς και συγκρίσεις.

2.1 Προσεγγίσεις μέσω απλής ή σύνθετης κατανομής Poisson, και στοιχεία από τη θεωρία των ακραίων τιμών

Όπως έχει ήδη αναφερθεί, το συγκεκριμένο κεφάλαιο διαπραγματεύεται την προσέγγιση των στατιστικών συναρτήσεων σάρωσης, από κατανομές Poisson (απλές ή σύνθετες). Θα λέμε ότι μία τ.μ. N ακολουθεί την κατανομή Poisson, με μέση τιμή λ (συμβ. $Po(\lambda)$), εάν η

συνάρτηση πιθανότητας της N δίδεται από τη σχέση

$$P(N=\nu)=e^{-\lambda}\frac{\lambda^{\nu}}{\nu!},\ \text{με}\ \lambda\geq 0\ \text{και}\ \nu=0,1,\ldots$$

Με τον όρο σύνθετη κατανομή Poisson, εννοούμε την κατανομή της τ.μ.

$$U=\sum_{i=1}^{N}Z_i,$$

όπου τα $Z_i, i=1,2,\ldots,n$, είναι ανεξάρτητες και ισόνομες τ.μ. με συνάρτηση κατανομής F, και N είναι μία τ.μ. με κατανομή Poisson, ανεξάρτητη από τις Z_i, με μέση τιμή λ (συμβ. $U\sim CP(\lambda,F)$). Η πιθανογεννήτρια συνάρτηση της U, έχει τη μορφή

$$P_U(t)=E(t^U)=e^{-\lambda(1-E(t^Z))}=e^{-\lambda(1-P_Z(t))}, \tag{2.1.1}$$

όπου $P_Z(t)$, είναι η πιθανογεννήτρια των Z_i. Επομένως, στην περίπτωση που γνωρίζουμε την πιθανογεννήτρια $P_Z(t)$, η συνάρτηση πιθανότητας της U, μπορεί να υπολογιστεί από το ανάπτυγμα Taylor της $P_U(t)$, ή μέσα από αναδρομικές σχέσεις (Bowers et al (1997)), όπως θα δούμε και στη συνέχεια.

Ισοδύναμα, η σύνθετη κατανομή Poisson μπορεί να θεωρηθεί ως η κατανομή της παρακάτω τ.μ.,

$$U=\sum_{j=1,2,\ldots}jN_j,$$

όπου N_j, είναι ανεξάρτητες τ.μ., με $N_j\sim Po(\lambda_j)$ (σ' αυτήν περίπτωση, θα γράφουμε ότι $U\sim CP(\{\lambda_i\})$). Η συνάρτηση πιθανότητας της U, θα είναι ίση με

$$P(U=i)=\begin{cases}\exp(-\sum_{j=1}^{\infty}\lambda_j), & \text{εάν } i=0\\ \exp(-\sum_{j=1}^{\infty}\lambda_j)v(i), & \text{εάν } i=1,2,\ldots\end{cases} \tag{2.1.2}$$

όπου

$$v(i)=\sum_{s_1+2s_2+\ldots+is_i=i,s_i\geq 0}\frac{\lambda_1^{s_1}\lambda_2^{s_2}\cdots\lambda_i^{s_i}}{s_1!s_2!\cdots s_i!}$$

(δηλαδή, το άθροισμα κινείται σ' όλες τις ακέραιες λύσεις της εξίσωσης $s_1+2s_2+\ldots+is_i=i$, με $s_i\geq 0$).

Για να αξιολογήσουμε την προσέγγιση της κατανομής μίας τ.μ. (την οποία συνήθως δυσκολευόμαστε να προσδιορίσουμε ή να εκφράσουμε), από μια άλλη κατανομή, χρειαζόμαστε την έννοια της απόστασης, ανάμεσα σε κατανομές. Οι δύο αποστάσεις που θα παίξουν σημαντικό ρόλο στα αποτελέσματά μας, είναι η ομοιόμορφη απόσταση (ή απόσταση Kolmogorov,

d_K) και η απόσταση ολικής κύμανσης (total variation distance, d_{TV}). Για δυο τ.μ. U, V (ορισμένες στον ίδιο χώρο πιθανότητας) με συναρτήσεις κατανομής F_U, F_V, αντίστοιχα, οι παραπάνω αποστάσεις δίνονται από τις σχέσεις

$$d_K(\mathcal{L}(U), \mathcal{L}(V)) = d_K(F_U, F_V) = \sup_{-\infty<x<\infty} |F_U(x) - F_V(x)|,$$

και

$$d_{TV}(\mathcal{L}(U), \mathcal{L}(V)) = d_{TV}(F_U, F_V) = \sup_A |P(U \in A) - P(V \in A)|$$

όπου το supremum υπολογίζεται για όλα τα σύνολα A, της σ-άλγεβρας του χώρου πιθανότητας των U, V. Για λόγους ευκολίας, θα χρησιμοποιούμε πολλές φορές μέσα στο κείμενο, και το συμβολισμό $d(U, V)$ εννοώντας, $d(F_U, F_V)$ (για $d = d_K$ ή για $d = d_{TV}$).

Μια ιδιότητα των παραπάνω αποστάσεων, που θα επικαλούμαστε αρκετά συχνά στις επόμενες παραγράφους, είναι ότι εάν για μία ακολουθία τ.μ. U_n, οι αντίστοιχες ακολουθίες των αποστάσεων $d_K(U_n, V)$ ή $d_{TV}(U_n, V)$, τείνουν στο μηδέν, τότε η U_n συγκλίνει κατά κατανομή στην τ.μ. V (ασθενής σύγκλιση). Αξίζει επίσης να αναφέρουμε, την ισχύ της ανισότητας, $d_K(U, V) \leq d_{TV}(U, V)$, για οποιεσδήποτε τ.μ. U, V.

Το βασικό αντικείμενο του συγκεκριμένου κεφαλαίου, εντάσσεται σ' ένα γενικότερο πρόβλημα, που αφορά την προσέγγιση της κατανομής ενός αθροίσματος από τ.μ., και ειδικότερα, την προσέγγιση ενός αθροίσματος από δίτιμες τ.μ. (συνήθως εξαρτημένες), από μια κατανομή Poisson ή μία σύνθετη κατανομή Poisson. Η ενασχόληση με τέτοιου είδους προβλήματα, ξεκινάει σχεδόν δυο αιώνες πριν, με την προσέγγιση ενός αθροίσματος από ανεξάρτητες και ισόνομες δοκιμές Bernouli (διωνυμική κατανομή), από μία κατανομή Poisson (Poisson (1837)). Τα τελευταία χρόνια, για την εύρεση υπολογιστικά εύχρηστων φραγμάτων, για την απόσταση μεταξύ της κατανομής ενός αθροίσματος από δίτιμες τ.μ., και μιας κατανομής Poisson (και όχι μόνο), χρησιμοποιείται ευρέως η μέθοδος Chen-Stein (βλ. π.χ. Arratia et al (1989, 1990), Barbour et al (1992)). Η μέθοδος αυτή βασίστηκε στη μεθοδολογία που εισήγαγε ο C. Stein το 1972 για συνεχείς κατανομές, και προσαρμόστηκε από τον L. Chen, στην κατανομή Poisson, το 1975 (Chen (1975)). Στην περίπτωση όπου η εμφάνιση των γεγονότων που μας ενδιαφέρουν, είναι σπάνια, η παραπάνω μέθοδος αποδεικνύεται ιδιαίτερα αποτελεσματική.

Έτσι μέσα από τη μελέτη της βιβλιογραφίας, που αφορά την προσέγγιση των στατιστικών συναρτήσεων σάρωσης, μέσω κατανομής Poisson (βλ. Παράγραφος 2.3), διαπιστώσαμε ότι το επόμενο θεώρημα αποτελεί ένα από τα βασικότερα εργαλεία (Arratia et al (1989)).

Θεώρημα 2.1.1 *Έστω Γ ένα τυχαίο (απαριθμήσιμο) σύνολο, και I_i μία τ.μ. Bernoulli με μέση τιμή $\pi_i = E(I_i) = P(I_i = 1)$, για κάθε $i \in \Gamma$. Εάν $W = \sum_{i\in\Gamma} I_i$ και $\lambda = E(W) = \sum_{i\in\Gamma} \pi_i$ $(\lambda \in (0,\infty))$ τότε*

$$d_{TV}(\mathcal{L}(W), Po(\lambda)) \leq \frac{1-e^{-\lambda}}{\lambda}(b_1+b_2) + b_3 \min\{1, 1.4\lambda^{-1/2}\}$$

όπου

$$b_1 = \sum_{i\in\Gamma}\sum_{j\in\Gamma_i} \pi_i\pi_j, \quad b_2 = \sum_{i\in\Gamma}\sum_{j\in\Gamma_i\setminus\{i\}} \pi_{ij}, \quad b_3 = \sum_{i\in\Gamma} E\left|E\left(I_i - \pi_i|\sigma(I_j : j \notin \Gamma_i)\right)\right|$$

με $\pi_{ij} = E(I_iI_j)$ και $\{\Gamma_i : \Gamma_i \subset \Gamma, i \in \Gamma\}$ να είναι οποιαδήποτε οικογένεια υποσυνόλων του Γ.

Αξίζει να σημειώσουμε ότι ο όρος b_3, είναι ουσιαστικά ένα μέτρο για το βαθμό της εξάρτησης των I_i, από τ.μ. (γεγονότα) που είναι έξω από την «τοπική» περιοχή, που ορίζουν τα σύνολα Γ_i. Αυτό συμβαίνει διότι η σ-άλγεβρα, ως προς την οποία γίνεται η δέσμευση (η $\sigma(I_j : j \notin \Gamma_i)$), παράγεται από τις $\{I_j : j \notin \Gamma_i\}$. Έτσι, όταν τα Γ_i επιλεγούν με τέτοιο τρόπο ώστε να περιλαμβάνουν όλες τις τ.μ., με τις οποίες η I_i δεν είναι ανεξάρτητη, τότε ισχύει $b_3 = 0$, καθώς $E(I_i - \pi_i|\sigma(I_j : j \notin \Gamma_i)) = E(I_i - \pi_i) = 0$. Οπότε, βλέποντας και τη μορφή των b_1, b_2 καταλαβαίνουμε γιατί τα αποτελέσματα μέσω της μεθόδου Chen-Stein είναι ικανοποιητικά, όταν οι μέσες τιμές των I_i είναι μικρές (σπάνια ενδεχόμενα), και έχουμε ασθενή τοπική εξάρτηση μεταξύ των $I_i, i \in \Gamma$.

Σε αρκετές περιπτώσεις, και κυρίως όταν τα γεγονότα που μελετάμε, έχουν μια τάση να εμφανίζονται κατά ομάδες (συστάδες, clusters/clumps), οι προσεγγίσεις μέσα από μια κατανομή Poisson, αποδεικνύονται όχι και τόσο αποτελεσματικές (μεγάλες τιμές στα φράγματα, και αργή σύγκλιση στο μηδέν). Ένας τρόπος να αντιμετωπίσουμε τέτοιες καταστάσεις, είναι να προσεγγίσουμε την κατανομή του αθροίσματος, μέσα από μια σύνθετη κατανομή Poisson (ειδικά, εάν το πλήθος των ομάδων φαίνεται να ακολουθεί μια κατανομή Poisson, ενώ το «μέγεθος» κάθε ομάδας, μια κατανομή F). Το ακόλουθο θεώρημα, που αφορά την προσέγγιση ενός αθροίσματος μη αρνητικών τ.μ., Boutsikas and Koutras (2001), έχει ήδη χρησιμοποιηθεί σε προσεγγίσεις, των στατιστικών συναρτήσεων σάρωσης (βλ. Παράγραφος 2.3), αλλά αποτελεί και τη βάση των νέων αποτελεσμάτων (Παράγραφος 2.4), του παρόντος κεφαλαίου.

Θεώρημα 2.1.2 *Έστω $Z_i, i = 1, 2, \ldots, n$ μία ακολουθία μη αρνητικών τ.μ., με $p_i = P(Z_i > 0)$, για κάθε $i = 1, 2, \ldots, n$. Εάν $\lambda = \sum_{i=1}^{n} p_i$ τότε*

$$d_K(\mathcal{L}(\sum_{i=1}^{n} Z_i), CP(\lambda, F)) \leq \sum_{i=2}^{n} \left(P(Z_i > 0, \sum_{j \in B_i} Z_j > 0) + P(Z_i > 0) P(\sum_{j \in B_i} Z_j > 0) \right) + \frac{1}{2} \sum_{i=1}^{n} P(Z_i > 0)^2 + \sum_{i=2}^{n} d_k(\sum_{j \in B_i'} Z_j + Z_i, \sum_{j \in B_i'} Z_j + Z_i')$$

όπου $F(x) = \frac{1}{\lambda} \sum_{i=1}^{n} p_i P(Z_i \leq x | Z_i > 0), x \in \Re$, $B_i, i = 2, 3, \ldots, n$ μια οικογένεια συνόλων, με $B_i \subseteq \{1, 2, \ldots, i-1\}$ (για κάθε $i = 2, 3, \ldots, n$) και Z_i' μια ακολουθία από ανεξάρτητες τ.μ., με τις ίδιες κατανομές, με τις Z_i $(i = 1, 2, \ldots, n)$.

Τα προηγούμενα θεωρήματα, ναι μεν αποτελούν τα θεμέλια για τα αποτελέσματα που ακολουθούν, όμως εξίσου ενδιαφέρουσα είναι και η σύνδεση των στατιστικών συναρτήσεων σάρωσης, με τις κινούμενες διατεταγμένες παρατηρήσεις (moving order statistics). Για να γίνουμε πιο σαφείς, με τον όρο κινούμενες διατεταγμένες παρατηρήσεις, εννοούμε τις διατεταγμένες παρατηρήσεις της ακολουθίας τ.μ. $Y_1, Y_2, \ldots, Y_n$, σε κάθε παράθυρο $Y_i, Y_{i+1}, \ldots, Y_{i+k-1}$, μήκους k $(i \in \{1, 2, \ldots, n-k+1\})$. Αναλυτικά, σε κάθε παράθυρο μήκους k, διατάσσουμε τις παρατηρήσεις σε φθίνουσα σειρά, και συμβολίζουμε με $Y_{r:k}^{(i)}$ την r-οστή μεγαλύτερη παρατήρηση, δηλαδή

$$Y_{1:k}^{(i)} \geq Y_{2:k}^{(i)} \geq \ldots \geq Y_{k:k}^{(i)}.$$

Τότε οι τ.μ. $Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}$ (όπου $1 \leq k \leq n, 1 \leq r \leq k$), αποτελούν τις κινούμενες διατεταγμένες παρατηρήσεις (μήκους k) του αρχικού δείγματος. Οι κινούμενες διατεταγμένες παρατηρήσεις έχουν απασχολήσει διάφορους συγγραφείς, καθώς παρουσιάζει ενδιαφέρον ο προσδιορισμός των από κοινού κατανομών τους, ο υπολογισμός της συνδιακύμανσης δυο κινούμενων διατεταγμένων παρατηρήσεων, όπως και οι εφαρμογές τους στον ποιοτικό έλεγχο, και στα διαγράμματα ελέγχου που στόχο έχουν την εξακρίβωση της εξάρτησης ανάμεσα σε δυο τ.μ. (δείτε π.χ. David (1955), Cleveland and Kleiner (1975), David and Rogers (1983)).

Στις επόμενες παραγράφους, θα δούμε πως αποδεικνύεται εύκολα ότι σε μία ακολουθία από ανεξάρτητες και ισόνομες συνεχείς τ.μ., η ασυμπτωτική συμπεριφορά των γενικευμένων συναρτήσεων σάρωσης, ταυτίζεται μ' αυτή των κινούμενων διατεταγμένων παρατηρήσεων. Ταυτόχρονα, με τη αρωγή της θεωρίας των ακραίων τιμών (extreme value theory), επιτυγχάνεται η εξασφάλιση των κατάλληλων προϋποθέσεων, για την επίτευξη χρήσιμων οριακών αποτελεσμάτων. Έτσι, πριν κλείσουμε την παρούσα παράγραφο, είναι απαραίτητο να συμπεριλάβουμε μερικά βασικά στοιχεία, από την παραπάνω θεωρία.

Η μελέτη των ακραίων παρατηρήσεων, πήρε μεγάλη ώθηση από το χώρο της αστρονομίας, λόγω του πλήθους των θεωρητικών αποτελεσμάτων που σχετίζονται με την αντιμετώπιση προβλημάτων που εμφανίζονται στο συγκεκριμένο επιστημονικό πεδίο. Από το 1920 έως το 1950, δημιουργήθηκε ένα μεγάλο ρεύμα στο χώρο των πιθανοτήτων, που οφείλεται κυρίως στις εφαρμογές της παραπάνω θεωρίας σε τομείς της μετεωρολογίας, στον έλεγχο της ανθεκτικότητας των υλικών, στη σεισμολογία, στην ασφαλιστική και αναλογιστική επιστήμη κ.α. Για μια σφαιρική ανασκόπηση της θεωρίας, μπορεί κάποιος να ανατρέξει στα βιβλία, π.χ., των Embrechts et al (1997), Kotz and Nadarajah (2000), Reiss and Thomas (2000), Coles (2001).

Η στοχαστική συμπεριφορά της μέγιστης παρατήρησης $M_n = \max\{Y_1, Y_2, \ldots, Y_n\}$, σε μία ακολουθία από ανεξάρτητες και ισόνομες τ.μ. $Y_1, Y_2, \ldots, Y_n$, με συνάρτηση κατανομής F, αποτελεί σημείο αναφοράς για τη θεωρία των ακραίων τιμών. Οι ασυμπτωτικές ιδιότητες των διατεταγμένων παρατηρήσεων, και συνάμα, το πλήθος των τ.μ. που υπερβαίνουν ένα κατώφλι (number of exceedances), καθορίζονται από τη συμπεριφορά της ουράς της κατανομής, από την οποία προέρχεται το δείγμα που εξετάζουμε. Η ακριβής κατανομή της τ.μ. M_n, υπολογίζεται εύκολα, καθώς

$$P(M_n \leq x) = F(x)^n.$$

Όμως, πολλά προβλήματα εξακολουθούν να υφίστανται, διότι αφενός, η συνάρτηση κατανομής F, είναι συνήθως άγνωστη, και αφετέρου,

$$\lim_{n\to\infty} P(M_n \leq x) = \lim_{n\to\infty} F(x)^n = 0,$$

για κάθε x, με $F(x) < 1$. Επομένως, η τελευταία σχέση δε μας δίνει καμία πληροφορία, για την ασυμπτωτική συμπεριφορά της μέγιστης παρατήρησης.

Λειτουργώντας στα ίδια πλαίσια με το κεντρικό οριακό θεώρημα, και βρίσκοντας κάποιες σταθερές με σκοπό την «κανονικοποίηση» της μέγιστης παρατήρησης, στοχεύουμε σε πιο εύχρηστες οριακές ιδιότητες. Έτσι, το σημαντικό θεώρημα Fisher-Tippett (βλ. π.χ. Embrechts et al (1997)), μας λέει ότι εάν υπάρχουν σταθερές $c_n > 0$ και $d_n \in \Re$, τέτοιες ώστε

$$\lim_{n\to\infty} P\left(\frac{M_n - d_n}{c_n} \leq x\right) = H(x) \tag{2.1.3}$$

όπου H είναι μια μη-εκφυλισμένη συνάρτηση κατανομής, τότε η H ανήκει σε μία από τις τρεις ακόλουθες οικογένειες (θέσης-διασποράς) κατανομών:

$$Frechet: \qquad \Phi_a(x) = \begin{cases} 0, & x \leq 0 \\ \exp(-x^{-a}), & x > 0 \end{cases}, \quad a > 0$$

$$(Reversed)Weibull : \Psi_a(x) = \begin{cases} \exp(-(-x)^a), & x \leq 0 \\ 1, & x > 0 \end{cases}, \quad a > 0$$

$$Gumbel : \qquad \Lambda(x) = \exp(-e^{-x}), \quad x \in \Re.$$

Πιο συγκεκριμένα, θα ισχύει

$$H(x) = \Phi_a((x-\mu)/\sigma) \quad \text{ή} \quad H(x) = \Psi_a((x-\mu)/\sigma) \quad \text{ή} \quad H(x) = \Lambda((x-\mu)/\sigma),$$

όπου $\mu \in \Re$ είναι μια παράμετρος θέσης και $\sigma > 0$ παράμετρος διασποράς. Οι παραπάνω οικογένειες είναι γνωστές ως *κατανομές ακραίων τιμών* (extreme value distributions), ενώ οι αντίστοιχες ακολουθίες c_n, d_n, ως *σταθερές κανονικοποίησης* (norming constants).

Εάν ισχύει η (2.1.3) για μια κατανομή F, θα λέμε ότι η ουρά της F (συμβ. $\overline{F}$) ανήκει στο μέγιστο πεδίο έλξης (maximum domain of attraction) της H (συμβ., $\overline{F} \in MDA(H)$). Τέλος, σημείο κλειδί για τη συνέχεια, είναι και το παρακάτω θεώρημα (βλ. π.χ., Embrechts et al (1997)), το οποίο μας προσφέρει μια ικανή και αναγκαία συνθήκη ώστε μια κατανομή F, να ανήκει σε κάποιο πεδίο έλξης, με συγκεκριμένες σταθερές κανονικοποίησης.

Θεώρημα 2.1.3 *Η συνάρτηση κατανομής F ανήκει στο μέγιστο πεδίο έλξης της κατανομής ακραίων τιμών H, με σταθερές κανονικοποίησης $c_n > 0, d_n \in \Re$ εάν και μόνο εάν*

$$\lim_{n\to\infty} n\overline{F}(c_n x + d_n) = -\ln H(x), \quad x \in \Re.$$

Όταν $H(x) = 0$, τότε το όριο θεωρείται $+\infty$.

2.2 Στατιστικές συναρτήσεις σάρωσης

Έστω $Y_1, Y_2, \ldots, Y_n$ μία ακολουθία από ανεξάρτητες και ισόνομες, συνεχείς τ.μ., με συνάρτηση κατανομής F και $X_i(u)$ η δείκτρια τ.μ., που ορίζεται ως εξής

$$X_i(u) = I_{(u,\infty)}(Y_i) = \begin{cases} 1, & \text{εάν } Y_i > u \\ 0, & \text{εάν } Y_i \leq u \end{cases}, \quad i = 1, 2, \ldots, n.$$

Η τ.μ. Y_i, υπερβαίνει την τιμή $u \in \Re$ με πιθανότητα

$$p = P(X_i(u) = 1) = E(X_i(u)) = P(Y_i > u) = \overline{F}(u).$$

Ας θεωρήσουμε στη συνέχεια, όλα τα κινούμενα παράθυρα μήκους k, της ακολουθίας $Y_1, Y_2, \ldots, Y_n$, δηλαδή,

$$Y_i, Y_{i+1}, \ldots, Y_{i+k-1}, \quad i = 1, 2, \ldots, n-k+1$$

και ας εισάγουμε την ακολουθία του πλήθους των υπερβάσεων από το κατώφλι u, σε παράθυρα μήκους k (k-scan exceedance process), με τον εξής τρόπο

$$S_k^{(i)}(u) = \sum_{j=i}^{i+k-1} X_j(u) = \sum_{j=i}^{i+k-1} I_{(u,\infty)}(Y_j), \quad i = 1, 2, \ldots, n-k+1.$$

Προφανώς, η $S_k^{(i)}(u)$ εκφράζει τον αριθμό των τ.μ. Y_i, $Y_{i+1}, \ldots, Y_{i+k-1}$, οι οποίες έχουν τιμή μεγαλύτερη από u, ενώ η

$$S_{n,k}(u) = \max_{1\le i\le n-k+1} S_k^{(i)}(u) = \max_{1\le i\le n-k+1} \sum_{j=i}^{i+k-1} X_j(u)$$

εκφράζει το μέγιστο πλήθος τέτοιων «ακραίων» τιμών (πάνω από u), ανάμεσα σ' όλα τα παράθυρα μήκους k, της ακολουθίας $Y_1, Y_2, \ldots, Y_n$. Άμεση σχέση με την προηγούμενη τ.μ., έχει και η τ.μ.

$$W_{n,k,r}(u) = \sum_{i=1}^{n-k+1} I_{[r,\infty)}(S_k^{(i)}(u))$$

η οποία, απαριθμεί το πλήθος των παραθύρων μήκους k, στα οποία οι τ.μ. που έχουν ξεπεράσει το κατώφλι u, είναι περισσότερες από r. Είναι φανερό πως ισχύει

$$P(W_{n,k,r}(u) = 0) = P(S_{n,k}(u) < r).$$

Οι τ.μ. $W_{n,k,r}(u)$ και $S_{n,k}(u)$ θα αποκαλούνται, *γενικευμένες συναρτήσεις σάρωσης*.

Όπως έχουμε ήδη αναφέρει, υπάρχει μια ενδιαφέρουσα σύνδεση των γενικευμένων συναρτήσεων σάρωσης, με τις κινούμενες διατεταγμένες παρατηρήσεις, της αρχικής ακολουθίας $Y_1, Y_2, \ldots, Y_n$. Αρχικά ας θυμηθούμε ότι, διατάσσοντας σε φθίνουσα σειρά τις Y_i, Y_{i+1}, …, Y_{i+k-1}, σε κάθε παράθυρο μήκους k ($i \in \{1, 2, \ldots, n-k+1\}$), παίρνουμε

$$Y_{1:k}^{(i)} \ge Y_{2:k}^{(i)} \ge \ldots \ge Y_{k:k}^{(i)}.$$

Δηλαδή, με $Y_{r:k}^{(i)}$ συμβολίζουμε την r-οστή μεγαλύτερη παρατήρηση, στο i παράθυρο. Επιπλέον, για κάποιο συγκεκριμένο r, ας διατάξουμε τις τ.μ. $Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}$, σε φθίνουσα σειρά, και με $Y_{m:r:k}$ να συμβολίσουμε την m-οστή μεγαλύτερη παρατήρηση, ανάμεσα στις προηγούμενες $n-k+1$ τ.μ. Αυτό σημαίνει ότι

$$Y_{1:r:k} \ge Y_{2:r:k} \ge \ldots \ge Y_{n-k+1:r:k}.$$

Στις ειδικές περιπτώσεις όπου $m = 1$ και $m = n-k+1$, έχουμε

$$\begin{aligned} Y_{1:r:k} &= \max\{Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}\}, \\ Y_{n-k+1:r:k} &= \min\{Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}\}. \end{aligned}$$

Σ' αυτό το σημείο αναφέρουμε ότι η παράμετρος n, έχει αφαιρεθεί από τους τελευταίους συμβολισμούς, αφού τη θεωρήσαμε ως σταθερά. Στην περίπτωση όμως, που το n μεταβάλλεται (π.χ. στην εξέταση των ασυμπτωτικών ιδιοτήτων, όπου το $n \to \infty$) θα χρησιμοποιούμε το συμβολισμό $Y_{m:r:k}(n)$, αντί $Y_{m:r:k}$. Η συνάρτηση κατανομής της $Y_{m:r:k}(n)$ μπορεί να εκφραστεί μέσω της αντίστοιχης της $W_{n,k,r}(u)$, ως εξής:

$$\begin{aligned} P\left(Y_{m:r:k}(n) \le u\right) &= P(\text{ το πολύ } m-1 \text{ από τα } Y_{r:k}^{(1)}, \ldots, Y_{r:k}^{(n-k+1)} \text{ υπερβαίνουν το } u) \\ &= P(\sum_{i=1}^{n-k+1} I_{(u,\infty)}(Y_{r:k}^{(i)}) < m) \\ &= P(\sum_{i=1}^{n-k+1} I_{[r,\infty)}(\sum_{j=i}^{i+k-1} I_{(u,\infty)}(Y_j)) < m) \\ &= P(\sum_{i=1}^{n-k+1} I_{[r,\infty)}(S_k^{(i)}(u)) < m) \\ &= P(W_{n,k,r}(u) < m). \end{aligned} \tag{2.2.1}$$

Για παράδειγμα, για $m = 1$, παίρνουμε

$$P(\max\{Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}\} \le u) = P(W_{n,k,r}(u) = 0), \tag{2.2.2}$$

δηλαδή, η συνάρτηση κατανομής της μέγιστης, από τις κινούμενες διατεταγμένες παρατηρήσεις $Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}$, προσδιορίζεται από την τιμή που θα πάρει η συνάρτηση πιθανότητας της $W_{n,k,r}(u)$, στο μηδέν.

Αξίζει να επισημάνουμε, ότι η μέγιστη παρατήρηση που αναφέραμε προηγουμένως, απευθύνεται σ' ένα σύνολο από εξαρτημένες τ.μ. (καθώς οι $Y_{r:k}^{(i)}$ ορίζονται σε επικαλυπτόμενα παράθυρα). Επίσης, οι συμβολισμοί που χρησιμοποιούμε εδώ για τις διατεταγμένες παρατηρήσεις, είναι διαφορετικοί απ' αυτούς που χρησιμοποιούνται συνήθως στα βιβλία (βλ. π.χ. Arnold and Balakrishnan (1989) και David and Nagaraja (2003)). Συγκεκριμένα, για εμάς η r-οστή διατεταγμένη αναφέρεται στην r-οστή μεγαλύτερη παρατήρηση, και όχι στην r-οστή μικρότερη.

Ας θεωρήσουμε στη συνέχεια, την ειδική περίπτωση όπου οι τ.μ. Y_i ακολουθούν την Ομοιόμορφη κατανομή στο (0, 1), και ας πάρουμε $u = 1 - p$, με $0 < p < 1$. Τότε η $X_i(u)$ γίνεται μια τ.μ. Bernoulli (την οποία θα συμβολίζουμε με X_i), με πιθανότητα επιτυχίας p, αφού

$$P(X_i = 1) = P(Y_i > u) = P(Y_i > 1 - p) = 1 - P(Y_i \le 1 - p) = p.$$

Επίσης, σ' αυτή την περίπτωση η $S_{n,k}(u)$ και η $W_{n,k,r}(u)$, μετασχηματίζονται στις κλασικές

συναρτήσεις σάρωσης, $S_{n,k}$ και $W_{n,k,r}$, αντίστοιχα, ορισμένες επάνω σε ανεξάρτητες και ισόνομες δοκιμές Bernoulli.

Για την τελευταία περίπτωση χρησιμοποιούμε τους ακόλουθους συμβολισμούς

$$\begin{aligned} S_k^{(i)} &= \sum_{j=i}^{i+k-1} X_j, \ \ i = 1, 2, \ldots, n-k+1, \\ S_{n,k} &= \max_{1 \le i \le n-k+1} S_k^{(i)}, \\ W_n &= W_{n,k,r} = \sum_{i=1}^{n-k+1} I_{[r,\infty)}(S_k^{(i)}). \end{aligned}$$

Στη συνέχεια θα συμβολίζουμε με $b(x; l, p)$ και $B(x; l, p)$, τη συνάρτηση πιθανότητας και την αθροιστική συνάρτηση κατανομής μιας διωνυμικής τ.μ. X, αντιστοίχως, δηλαδή

$$\begin{aligned} b(x; l, p) &= P(X = x) = \binom{l}{x} p^x (1-p)^{l-x}, \quad x = 0, 1, \ldots, n, \\ B(x; l, p) &= P(X \le x) = \sum_{j=0}^{\lfloor x \rfloor} b(j; l, p), \quad x \in \Re. \end{aligned}$$

όπου με $\lfloor x \rfloor$ αναφερόμαστε στο ακέραιο μέρος του x. Επιπλέον, στις επόμενες παραγράφους, κάνουμε χρήση και των πιθανοτήτων

$$\begin{aligned} f(s; k, p) &= P(S_k^{(1)} < s, S_k^{(2)} < s, \ldots, S_k^{(k)} < s, S_k^{(k+1)} \ge s) \\ G(s; k, p) &= P(S_k^{(1)} < s, S_k^{(2)} < s, \ldots, S_k^{(k+1)} < s) \end{aligned}$$

οι οποίες μπορούν να εκφραστούν μέσω των $b(x; l, p)$ και $B(x; l, p)$, με τον εξής τρόπο (Glaz and Naus (1991))

$$\begin{aligned} f(s; k, p) = \frac{p}{s} b(s-1; k-1, p)[s(1-p)b(s-1; k-1, p) \\ + (s - kp)B(s-2; k-1, p)], \qquad (2.2.3) \\ G(s; k, p) = B(s-1; k, p)^2 - b(s; k, p)[(s-1)B(s-2; k, p) \\ - kpB(s-3; k-1, p)], \end{aligned}$$

με $1 \le s \le k$ (εάν $s > k$ ή $s < 0$ τότε $f(s; k, p) = 0$).

Ακόμη, με $H(\theta, p)$ συμβολίζουμε την απόσταση Kullback-Leibler (ή σχετική εντροπία, relative entropy)

$$H(\theta, p) = \theta \ln\left(\frac{\theta}{p}\right) + (1-\theta) \ln\left(\frac{1-\theta}{1-p}\right) = \ln \frac{\theta^\theta (1-\theta)^{1-\theta}}{p^\theta (1-p)^{1-\theta}},$$

και με $h(\theta, p)$ τη μερική παράγωγο της $H(\theta, p)$ ως προς θ, δηλαδή

$$h(\theta, p) = \frac{d}{d\theta} H(\theta, p) = \ln \frac{\theta(1-p)}{p(1-\theta)},$$

όπου $0 < p < \theta < 1$.

Τέλος, τα σύμβολα $\sim, O(\cdot)$ χρησιμοποιούνται με το συνήθη τρόπο, δηλαδή,

$$f(t) \sim g(t) \text{ καθώς } t \to t_0 \text{ εάν } \lim_{t \to t_0} \frac{f(t)}{g(t)} = 1,$$

$$f(t) = O(g(t)) \text{ εάν η συνάρτηση } \frac{f(t)}{g(t)}, \text{ είναι φραγμένη.}$$

2.3 Προσεγγίσεις για τις στατιστικές συναρτήσεις σάρωσης, μέσω απλής ή σύνθετης κατανομής Poisson

Η συγκεκριμένη παράγραφος, αποτελεί μια ανασκόπηση των αποτελεσμάτων που αφορούν την προσέγγιση της «απλής» στατιστικής συνάρτησης σάρωσης $S_{n,k}$ και της απαριθμήτριας τ.μ. $W_{n,k,r}$. Ειδικότερα, στην Παράγραφο 2.3.1 συμπεριλαμβάνουμε όλα τα αποτελέσματα για την προσέγγιση της κατανομής των κλασικών συναρτήσεων σάρωσης, μέσω σύνθετης ή απλής κατανομής Poisson. Περιοριζόμαστε σε αποτελέσματα τα οποία προσφέρουν εκτιμήσεις για το σφάλμα της προσέγγισης (άνω φράγμα για την απόσταση της προσεγγιστικής κατανομής, από την ακριβή), τα οποία συνήθως είναι αρκετά αποτελεσματικά, όταν το n παίρνει μεγάλες τιμές.

Παρόλο που στη βιβλιογραφία υπάρχει και μια άλλη κατηγορία προσεγγίσεων, πολλαπλασιαστικού τύπου, δε σκοπεύουμε να αναφερθούμε σ' αυτές διότι, συνήθως δεν προσφέρουν φράγματα για τα σφάλματα των προσεγγίσεων, με αποτέλεσμα να μην μπορούν να χρησιμοποιηθούν για ασυμπτωτικά αποτελέσματα. Ο αναγνώστης μπορεί να ανατρέξει στο βιβλίο των Glaz, Naus and Wallenstein (2001), για μια λεπτομερή παρουσίαση των τελευταίων προσεγγίσεων.

Όπως έχουμε ήδη αναφέρει, μια δημοφιλής μέθοδος, η οποία μας δίνει τη δυνατότητα μέσα από εύχρηστα εργαλεία, να προσεγγίζουμε την κατανομή ενός αθροίσματος από δίτιμες τ.μ., από μια κατανομή Poisson, είναι η μέθοδος Chen-Stein. Επομένως, το Θεώρημα 2.1.1 παίζει κυρίαρχο ρόλο στα αποτελέσματα που ακολουθούν, κυρίως όταν προσεγγίζονται οι συναρτήσεις σάρωσης, από μια απλή κατανομή Poisson. Όμως, επειδή η πολλαπλή συνάρτηση σάρωσης $W_{n,k,r}$, απαριθμεί γεγονότα τα οποία έχουν την τάση να παρουσιάζονται

κατά συστάδες, μια προσέγγιση μέσω μιας σύνθετης κατανομής Poisson, είναι πιθανότατα μια πιο ενδεδειγμένη λύση. Για τη διαμόρφωση των αποτελεσμάτων που αναφέρονται στην τελευταία προσέγγιση, αποδεικνύεται ιδιαίτερα χρήσιμο το Θεώρημα 2.1.2.

2.3.1 Προσεγγίσεις και φράγματα

Θα ξεκινήσουμε μ' ένα αποτέλεσμα για την απλή συνάρτηση σάρωσης $S_{n,k}$, το οποίο οφείλεται στους Arratia, Gordon and Waterman (1990). Ας εισάγουμε αρχικά τις τ.μ.

$$C_i = \begin{cases} 1, & \text{εάν } \sum_{j=i}^{i+k-1} X_j = r \\ 0, & \text{διαφορετικά} \end{cases}, i = 1, 2, \ldots, n-k+1.$$

(σύμβαση: $C_i = 0$ για $i \leq 0$) και έπειτα, ας θεωρήσουμε τη βοηθητική τ.μ. W, αθροίζοντας τις ποσότητες

$$D_i = C_i \prod_{j=1}^{k} (1 - C_{i-j}), \quad i = 1, 2, \ldots, n-k+1$$

δηλαδή

$$W = \sum_{i=1}^{n-k+1} C_i \prod_{j=1}^{k} (1 - C_{i-j}).$$

Οι Arratia, Gordon and Waterman (1990) με τη βοήθεια του Θεωρήματος 2.1.1, απέδειξαν το επόμενο ενδιαφέρον αποτέλεσμα.

Θεώρημα 2.3.1 *Εάν $p < r/k < 1$ τότε*

$$\left| P(S_{n,k} < r) - e^{-E(W)} \right| \leq 7kb(r; k, p) + (1 - B(r; k, p)).$$

Δηλαδή, η πιθανότητα $P(S_{n,k} < r)$ φράσσεται από την ποσότητα $e^{-E(W)} \pm UB$, όπου το UB είναι ίσο με το δεξιό μέλος της τελευταίας ανισότητας, ενώ

$$E(W) = (n-k+1)E(D_i).$$

Οι ίδιοι συγγραφείς, απέδειξαν ότι για τη μέση τιμή $E(W)$, ισχύει

$$\frac{r}{k} - p \leq \frac{E(W)}{(n-k+1)b(r; k, p)} \leq (\frac{r}{k} - p) + 2(1 - \frac{r}{k})(1 - B(r; k, p)) \tag{2.3.1}$$

ενώ ένα εναλλακτικό άνω φράγμα δίδεται από

$$\frac{E(W)}{(n-k+1)b(r; k, p)} \leq (\frac{r}{k} - p) + 2(1 - \frac{r}{k})e^{-kH(r/k,\, p)}$$

όπου $H(\theta, p)$ είναι η απόσταση Kullback-Leibler. Για μεγάλες τιμές της παραμέτρου k (με το πηλίκο r/k, να διατηρείται σταθερό) ο δεύτερος όρος του αθροίσματος, στο δεξί μέλος της ανισότητας (2.3.1), παίρνει αμελητέες τιμές και επομένως, μπορεί να χρησιμοποιηθεί ο προσεγγιστικός τύπος

$$E(W) \approx (\frac{r}{k} - p)(n - k + 1)b(r; k, p).$$

Οι Dembo and Karlin (1992), χρησιμοποιώντας με τη σειρά τους τη μέθοδο Chen-Stein (Θεώρημα 2.1.1), μελέτησαν την πολλαπλή συνάρτηση σάρωσης, ορισμένη πάνω σε μία ακολουθία ανεξάρτητων και ισόνομων θετικών τ.μ. (όχι αναγκαστικά δίτιμων). Συγκεκριμένα, έδωσαν ένα άνω φράγμα για την απόσταση ολικής κύμανσης, μεταξύ της κατανομής της συναρτήσεως σάρωσης, και μίας απλής κατανομής Poisson, το οποίο για την περίπτωση ακολουθίας ανεξάρτητων και ισόνομων δίτιμων τ.μ., περιγράφεται στο επόμενο θεώρημα.

Θεώρημα 2.3.2 *Εάν*

$$\lambda = (n - k + 1)B(r - 1; k, p), \quad \mu = (n - k + 1)(1 - B(r - 1; k, p)).$$

τότε

$$\textbf{\textit{α}}.\ d_{TV}(\mathcal{L}(W_{n,k,r}), Po(\mu)) \leq (1 - e^{-\mu})[(2k - 1)(1 - B(r - 1; k, p)) + 2\sum_{i=1}^{k-1} P(S_k^{(i+1)} \geq r | S_k^{(1)} \geq r)]$$

$$\textbf{\textit{β}}.\ d_{TV}(\mathcal{L}(n - k + 1 - W_{n,k,r}), Po(\lambda)) \leq (1 - e^{-\lambda})[(2k - 1)B(r - 1; k, p) + 2\sum_{i=1}^{k-1} B(r - 1; i, p)]$$

Τις δεσμευμένες πιθανότητες $P(S_k^{(i+1)} \geq r | S_k^{(1)} \geq r)$, που εμφανίζονται στο πρώτο άνω φράγμα, μπορούμε να τις υπολογίσουμε μέσω των ακόλουθων σχέσεων

$$P(S_k^{(i+1)} \geq r | S_k^{(1)} \geq r) = \frac{1}{1 - B(r - 1; k, p)} \cdot \sum_{s=0}^{k-i} (1 - B(r - s - 1; i, p))^2 b(k - i; s, p)$$

με $B(x; l, p) = b(x; l, p) = 0$, για $x < 0$.

Επομένως, το αποτέλεσμα από το Θεώρημα 2.3.2, μπορεί να χρησιμοποιηθεί για να δώσουμε φράγματα για τη συνάρτηση πιθανότητας ή την αθροιστική συνάρτηση κατανομής της

τ.μ. $W_{n,k,r}$. Τα φράγματα αυτά, σχηματίζουν διαστήματα, με κέντρο τις αντίστοιχες συναρτήσεις (συνάρτηση πιθανότητας ή αθροιστική συνάρτηση κατανομής), μιας απλής κατανομής Poisson, και το μήκος τους θα ισούται με δύο φορές το άνω φράγμα, που εισήχθηκε από το παραπάνω θεώρημα.

Όπως έχουμε επισημάνει, λόγω του γεγονότος ότι τα παράθυρα με μεγάλο πλήθος επιτυχιών, έχουν την τάση να εμφανίζονται κατά ομάδες, είναι λογικό οι προσεγγίσεις μέσω απλής κατανομής Poisson, όπως αυτή από το Θεώρημα 2.3.2, να μην είναι και τόσο ακριβείς. Προς επίρρωσιν των λεγομένων μας, αναφέρουμε ότι το άνω φράγμα της (a) τείνει στο μηδέν (όταν το n τείνει στο άπειρο) μόνο όταν $r = k$ και $p \to 0$, ενώ το φράγμα από την (b) μόνο εάν $r = 1$ και $p \to 1$

Παρακινούμενοι από ένα πρόβλημα σύγκρισης ακολουθιών, όπου είναι απαραίτητη η ανάγκη για ακριβείς προσεγγίσεις, οι Goldstein and Waterman (1992) χρησιμοποίησαν τις βοηθητικές τ.μ.

$$E_i = I_{[r,\infty)}(S_k^{(i)}) \prod_{j=1}^{\min(s,i-1)} (1 - I_{[r,\infty)}(S_k^{(i-j)})), \quad i = 1, 2, \ldots, n-k+1$$

(όπου s είναι ένας σταθερός ακέραιος αριθμός [1]) ώστε να υπολογίσουν ένα άνω φράγμα για την απόσταση ολικής κύμανσης, μεταξύ της κατανομής της πολλαπλής συνάρτησης σάρωσης και μίας σύνθετης κατανομής Poisson. Ουσιαστικά, οι τ.μ. E_i ορίζουν μια ομάδα (συστάδα), με την εμφάνιση ενός παραθύρου μήκους k (με αρχικό σημείο την X_i), με r τουλάχιστον επιτυχίες, το οποίο απέχει από το τελευταίο παράθυρο που έχει την ίδια ιδιότητα, τουλάχιστον $\min(s, i-1)$ δοκιμές. Η κατανομή του πλήθους C, των εμφανίσεων των γεγονότων $S_k^{(j)} \geq r$, μέσα σε μία συστάδα, δίδεται από τον τύπο

$$P(C = c) = P(\sum_{j=i}^{\beta} I_{[r,\infty)}(S_k^{(j)}) = c | E_i = 1), \quad c = 1, 2, \ldots$$

όπου

$$\beta = \min(\gamma \geq i : I_{[r,\infty)}(S_k^{(\gamma)}) = 1, I_{[r,\infty)}(S_k^{(\gamma+1)}) = 0, \ldots, I_{[r,\infty)}(S_k^{(\gamma+s)}) = 0).$$

Θέτοντας $s = k$, αποδεικνύεται το επόμενο αποτέλεσμα (Goldstein and Waterman (1992)) με τη βοήθεια της μεθόδου Chen-Stein (μέσω ενός ανάλογου αποτελέσματος, μ' αυτό του Θεωρήματος 2.1.1, για την περίπτωση των προσεγγίσεων μέσω διαδικασίας Poisson, Arratia et al (1989)).

[1] Γινόμενα της μορφής $\prod_{i=i_1}^{i_2} f(i)$ με $i_1 > i_2$, θεωρούνται ίσα με 1.

Θεώρημα 2.3.3 *Η απόσταση ολικής κύμανσης μεταξύ της $W_{n,k,r}$ και της σύνθετης κατανομής Poisson $CP(\lambda, G)$, με $\lambda = (n-k+1)(1-B(r-1;k,p))/E(C)$ και συνθέτουσα κατανομή $G(x) = P(C \leq x)$, φράσσεται άνω ως εξής*

$$d_{TV}(\mathcal{L}(W_{n,k,r}), CP(\lambda, G)) \leq 6\lambda^2(1+E(C))\frac{k}{n-k} + 2\lambda P(C > k).$$

Να σημειώσουμε ότι, στην παραπάνω έκφραση για το λ, αγνοήθηκαν οι περιορισμοί που οφείλονται στο εύρος της ακολουθίας των τ.μ. (boundary effects). Επειδή η ακριβής κατανομή της C, είναι πολύ δύσκολο να προσδιοριστεί, οι Goldstein and Waterman (1992) είχαν προτείνει μια απλή προσέγγιση για την κατανομή της C (για περισσότερες λεπτομέρειες, μπορούμε να ανατρέξουμε στη συγκεκριμένη εργασία). Εάν όμως ενδιαφερόμαστε μόνο για την κατανομή της $S_{n,k}$, και όχι για ολόκληρη την κατανομή της $W_{n,k,r}$ (να θυμίσουμε ότι, $P(S_{n,k} < r) = P(W_{n,k,r} = 0)$), το επόμενο απλό φράγμα για την $E(C)$ (δοθέντος ότι $s = k$) μπορεί να φανεί χρήσιμο

$$\frac{r}{k} - p \leq \frac{1}{E(C)} \leq (\frac{r}{k} - p) + 2(1-\frac{r}{k})e^{-kH(r/k,\, p)}.$$

Κάνοντας χρήση παρόμοιων βοηθητικών τ.μ., οι Boutsikas and Koutras (2002b) εισήγαγαν και υπολόγισαν ένα φράγμα για την απόσταση ανάμεσα στην κατανομή της $W_{n,k,r}$ και μίας σύνθετης κατανομής Poisson. Συγκεκριμένα, με τη χρήση των παρακάτω (περικεκομμένων) βοηθητικών μεταβλητών ("truncated" declumping variables)

$$E_i' = (1 - I_{[r,\infty)}(S_k^{(i-1)})) \sum_{j=1}^{k} \prod_{l=i}^{i+j-1} (I_{[r,\infty)}(S_k^{(l)})), \quad i = 1, 2, \ldots, n-k+1$$

και επικαλούμενοι το Θεώρημα 2.1.2, έφθασαν στο Θεώρημα 2.3.4, που διατυπώνεται στη συνέχεια.

Για την παρουσίαση των αποτελεσμάτων από δω και στο εξής, θεωρούμε ότι για τον υπολογισμό των $S_k^{(i)}$, η ακολουθία των τ.μ. $X_i, i = 1, 2, \ldots, n$, ορίζεται και για $i < 1$ ή $i > n$ (ώστε να εξαλείψουμε τα boundary effects). Τονίζουμε επίσης ότι, η απόσταση που χρησιμοποιείται στη συνέχεια, είναι η Kolmogorov και όχι η ολικής κύμανσης.

Θεώρημα 2.3.4 *Έστω* $\lambda = (n-k+1)P(E_1' > 0)$ *και* $G(x) = P(E_1' \leq x | E_1' > 0)$ $x = 0, 1, \ldots, k$. *Τότε,*

$$\begin{aligned} d_K(\mathcal{L}(W_{n,k,r}), CP(\lambda, G)) \leq\ & (\lambda+1)\sum_{i=r}^{k}\binom{k}{i}p^i(1-p)^{k-i} + \\ & (\lambda(3k-1)+k-1)\binom{k-1}{r-1}p^r(1-p)^{k-r+1} + \\ & (n-k)\sum_{b=2}^{k-1}\sum_{i=\max\{0,r-k+b-1\}}^{\min\{b-2,r-2\}}\binom{k-b}{r-i-1}\binom{b-2}{i} \\ & \binom{k-b}{r-i-2}p^{2r-i-1}(1-p)^{2k-b-2r+i+3}. \end{aligned}$$

Το πλεονέκτημα του παραπάνω αποτελέσματος, σε σχέση μ' αυτό που περιγράφεται στο Θεώρημα 2.3.3 είναι ότι, τόσο το άνω φράγμα όσο και οι παράμετροι λ, G της σύνθετης κατανομής Poisson, δίνονται από πιο απλές εκφράσεις. Έτσι, μπορεί εύκολα να αποδειχθεί ότι,

$$P(E_1' > 0) = \binom{k-1}{r-1}p^r q^{k-r+1}$$

και επομένως

$$\lambda = (n-k+1)P(E_1' > 0) = (n-k+1)\binom{k-1}{r-1}p^r q^{k-r+1}.$$

Οι Boutsikas and Koutras (2002b) έδειξαν ότι η συνθέτουσα αθροιστική συνάρτηση κατανομής $G(x)$, έχει τη μορφή

$$\begin{aligned} G(x) =\ & 1 - \sum_{j=\max\{0,r-x-1\}}^{\min\{k-x-1,r-1\}}\left(\frac{\binom{x}{x-r+j+1}\binom{k-x-1}{j}}{\binom{k-1}{r-1}}\right) \cdot \\ & \cdot\left(\binom{x}{x-r+j+1}p^{r-j-1}(1-p)^{x-r+j+1} + \right. \\ & \left. + \left(1-\frac{(x+1)(1-p)}{x-r+j+2}\right)\left(\sum_{i=0}^{x-r+j}\binom{x}{i}(1-p)^i p^{x-i-1}\right)\right) \end{aligned}$$

για $x = 1, 2, \ldots, k-1$, και $G(0) = 0, G(k) = 1$. Το άνω φράγμα του Θεωρήματος 2.3.4 έχει τάξη σύγκλισης $O(p)$ (για $r < k$) και επομένως, προσφέρει αρκετά ακριβείς προσεγγίσεις για την κατανομή της $W_{n,k,r}$, τουλάχιστον για την περίπτωση που $p \to 0$. Βέβαια, εάν το p διατηρείται σταθερό η ποιότητα των προσεγγίσεων δε θα είναι αρκετά καλή, και επομένως δεν μπορούν να εξαχθούν με τη βοήθεια του Θεωρήματος 2.3.4, χρήσιμα ασυμπτωτικά αποτελέσματα καθώς $n, k \to \infty$ (για p σταθερό).

Λόγω της τελευταίας παρατήρησης (και με σκοπό να αντιμετωπιστεί το προηγούμενο πρόβλημα), οι Boutsikas and Koutras (2006) εισήγαγαν την ακόλουθη οικογένεια από βοηθητικές μεταβλητές

$$E_i'' = \left(\prod_{j=i-k}^{i-1} (1 - I_{[r,\infty)}(S_k^{(j)})) \right) I_{[r,\infty)}(S_k^{(i)}) \left(\sum_{l=i}^{i+k} I_{[r,\infty)}(S_k^{(l)}) \right), \quad i = 1, 2, \ldots$$

Η τελευταία παρένθεση απαριθμεί το πλήθος των παραθύρων μήκους k, τα οποία ξεκινάνε από τις θέσεις $i, i+1, \ldots, i+k$, και περιέχουν τουλάχιστον r επιτυχίες. Παράλληλα, η πρώτη παρένθεση εξασφαλίζει ότι στις προηγούμενες k θέσεις, $i-k, i-k+1, \ldots, i-1$, όλα τα παράθυρα μήκους k περιέχουν λιγότερες από r επιτυχίες. Ο όρος αυτός είναι που κάνει δυνατή (για την περίπτωση που μας ενδιαφέρει) την κατασκευή ακριβών προσεγγίσεων. Έτσι οι Boutsikas and Koutras (2006), χρησιμοποιώντας τις $E_i'', i = 1, 2, \ldots$, κατέληξαν στο ακόλουθο θεώρημα.

Θεώρημα 2.3.5 *Έστω* $\lambda = (n-k+1)P(E_1'' > 0) = (n-k+1)f(r;k,p)$, *και*

$$\begin{aligned} G(x) &= P(E_1'' \le x | E_1'' > 0) \\ &= P\left(\sum_{l=k+1}^{2k+1} I_{[r,\infty)}(S_k^{(l)}) \le x \,\middle|\, I_{[r,\infty)}(S_k^{(l)}) = 0, l = 1, 2, \ldots, k, I_{[r,\infty)}(S_k^{(k+1)}) = 1 \right), \\ &\qquad\qquad x = 0, 1, \ldots, k. \end{aligned}$$

Τότε

$$\begin{aligned} d_K(\mathcal{L}(W_{n,k,r}), CP(\lambda, G)) \le{}& (2k-1)\lambda p(1-p)b(r-1;k-1,p) \\ &+ 3\lambda k f(r;k,p) + (\lambda+2)(1 - G(r;k,p)) \end{aligned}$$

όπου οι ποσότητες $f(r;k,p), G(r;k,p)$ *δίνονται από* (2.2.3).

Ο υπολογισμός της συνθέτουσας συνάρτησης κατανομής $G(x)$, είναι μια επίπονη διαδικασία. Παρόλα αυτά, το προηγούμενο αποτέλεσμα είναι εξαιρετικά χρήσιμο για τη μελέτη της ασυμπτωτικής συμπεριφοράς της $W_{n,k,r}$, για p σταθερό και $n, k \to \infty$.

Ακόμη, εάν το ενδιαφέρον μας επικεντρωθεί στην τ.μ. $S_{n,k}$ και όχι σ' ολόκληρη την κατανομή της $W_{n,k,r}$, προκύπτει εύκολα η επόμενη πρόταση.

Πόρισμα 2.3.1 *Η συνάρτηση κατανομής της* $S_{n,k}$ *μπορεί να προσεγγιστεί από την ποσότητα* $e^{-\lambda}$, $\lambda = (n-k+1)f(r;k,p)$, *ενώ για το σφάλμα προσέγγισης, ισχύει*

$$\begin{aligned} |P(S_{n,k} < r) - e^{-\lambda}| \le{}& (2k-1)\lambda p(1-p)b(r-1;k-1,p) + 3\lambda k f(r;k,p) \\ &+ (\lambda+2)(1 - G(r;k,p)). \end{aligned}$$

2.3.2 Ασυμπτωτικά αποτελέσματα

Στην παρούσα παράγραφο, θα παρουσιάσουμε ένα αριθμό ασυμπτωτικών αποτελεσμάτων, σχετικά με τη σύγκλιση στην απλή ή σύνθετη κατανομή Poisson, των συναρτήσεων σάρωσης $S_{n,k}$ και $W_{n,k,r}$. Οι λεπτομέρειες των αποδείξεων δε θα συμπεριληφθούν, παρόλο που είναι άμεσα επακόλουθα των φραγμάτων (επάνω στις αποστάσεις ολικής κύμανσης ή Kolmogorov), που μελετήσαμε στην προηγούμενη παράγραφο.

Με την αρωγή του Θεωρήματος 2.3.1 (και τη συζήτηση που ακολούθησε), προκύπτει εύκολα το επόμενο αποτέλεσμα (Arratia, Gordon and Waterman (1990)).

Πόρισμα 2.3.2 *Εάν οι n, k, r είναι θετικοί ακέραιοι αριθμοί, με $p < r/k < 1$ και*

$$\lambda = (n-k+1)(\frac{r}{k} - p)b(r; k, p)$$

τότε η $P(S_{n,k} < r)$ μπορεί να προσεγγιστεί από την $e^{-\lambda}$ με το σφάλμα της προσέγγισης να έχει τάξη σύγκλισης $O(\frac{lnn}{n})$.

Στην επόμενη πρόταση, η οποία αποδεικνύεται εύκολα με τη βοήθεια του Θεωρήματος 2.3.4 (για λεπτομέρειες, βλέπε Boutsikas and Koutras (2002b)), διατυπώνεται μία προσέγγιση της $W_{n,k,r}$, μέσω σύνθετης κατανομής Poisson.

Πόρισμα 2.3.3 *Ας υποθέσουμε ότι τα k, r παραμένουν σταθερά, ενώ $n \to \infty, p_n \to 0$ έτσι ώστε*

$$\lambda_n = (n-k+1)\binom{k-1}{r-1}p_n^r(1-p_n)^{k-r+1} \to \lambda \in (0, \infty).$$

Τότε η κατανομή της $W_{n,k,r}$ συγκλίνει σε μια σύνθετη κατανομή Poisson, με παραμέτρους λ και

$$G(x) = \begin{cases} 0, & x \leq 0 \\ 1 - \frac{\binom{k-x-1}{r-1}}{\binom{k-1}{r-1}}, & x = 1, 2, \ldots, k-r, \\ 1, & x \geq k-r+1. \end{cases}$$

Κάτω από τις ίδιες προϋποθέσεις, για την αθροιστική συνάρτηση κατανομής της $S_{n,k}$ ισχύει

$$P(S_{n,k} < r) \sim e^{-\lambda}.$$

Η συνάρτηση πιθανότητας $g(x)$, της συνθέτουσας κατανομής $G(x)$, δίδεται από τη σχέση

$$g(x) = G(x) - G(x-1) = \frac{\binom{k-x-1}{r-2}}{\binom{k-1}{r-1}}, \quad x = 1, 2, \ldots, k-r+1.$$

Αξίζει να αναφέρουμε ότι, στην ειδική περίπτωση $r = 2 < k$, η $g(x)$ είναι μία ομοιόμορφη διακριτή κατανομή στο σύνολο $1, 2, \ldots, k-1$. Επίσης, για $k = r$ η συνθέτουσα κατανομή εκφυλίζεται σε μια σταθερή κατανομή (παίρνει μόνο την τιμή 1) και η οριακή σύνθετη κατανομή Poisson $CP(\lambda, G)$, καταλήγει σε μια απλή κατανομή Poisson.

Η πιθανογεννήτρια συνάρτηση της σύνθετης κατανομής Poisson της Προτάσεως 2.3.3, είναι (βλ. (2.1.1))

$$\begin{aligned} P(t) = E(t^{W_{n,k,r}}) = e^{-\lambda(1-E(t^Z))} &= e^{-\lambda\left(1-\sum_{x=1}^{k-r+1} t^x g(x)\right)} \\ &= \exp\left(-\lambda\left(1-\sum_{x=1}^{k-r+1} t^x \frac{\binom{k-x-1}{r-2}}{\binom{k-1}{r-1}}\right)\right) \end{aligned}$$

και επομένως, οι πιθανότητες $P(W_{n,k,r} = i)$ μπορούν εύκολα να υπολογιστούν από την εύρεση του i όρου, της δυναμοσειράς της $P(t)$, $i = 0, 1, \ldots$ Αυτό μπορεί να γίνει είτε αριθμητικά (για συγκεκριμένες τιμές των παραμέτρων) ή αναλυτικά για όρους μικρής τάξεως.

Για παράδειγμα,

$$\begin{aligned} P(W_{n,k,r} = 0) &= \frac{1}{0!}P(0) = e^{-\lambda}, \\ P(W_{n,k,r} = 1) &= \frac{1}{1!}\left.\frac{dP(t)}{dt}\right|_{t=0} = \lambda g(1)e^{-\lambda} = \lambda\frac{r-1}{k-1}e^{-\lambda} \\ P(W_{n,k,r} = 2) &= \frac{1}{2!}\left.\frac{d^2P(t)}{dt^2}\right|_{t=0} = \lambda\frac{(r-1)(k-r)}{(k-1)(k-2)}e^{-\lambda} + \frac{\lambda^2}{2!}(\frac{r-1}{k-1})^2e^{-\lambda} \end{aligned} \tag{2.3.2}$$

κτλ. Εναλλακτικά, μπορεί να χρησιμοποιηθεί η αναδρομική σχέση (βλ. Bowers et al (1997))

$$\begin{aligned} &P(W_{n,k,r} = 0) = e^{-\lambda}, \\ &P(W_{n,k,r} = i) = \frac{\lambda k}{ri}\binom{k}{r}^{-1}\sum_{j=1}^{\min\{k-r+1,i\}} j\binom{k-j-1}{r-2}P(W_{n,k,r} = i-j), \quad i = 1, 2, \ldots \end{aligned}$$

Για $r < k$, ο ρυθμός σύγκλισης, που εξάγουμε από την Πρόταση 2.3.3, για την προσέγγιση της $P(S_{n,k} < r) \approx e^{-\lambda}$ είναι τάξεως $O(p)$. Άμεσο επακόλουθο είναι τα «φτωχά» οριακά αποτελέσματα, για την περίπτωση που το n και το k τείνουν στο ∞, ενώ το p παραμένει σταθερό. Η τελευταία περίπτωση, αντιμετωπίζεται από το επόμενο αποτέλεσμα (βλ. Θεώρημα 2.3.5 και για περισσότερες λεπτομέρειες, Boutsikas and Koutras (2006)).

Πόρισμα 2.3.4 *Ας υποθέσουμε ότι το p παίρνει μια σταθερή τιμή, και ότι $\theta \in (p, 1)$, ενώ οι k_n, r_n είναι δυο ακολουθίες θετικών ακέραιων αριθμών, με την ιδιότητα*

$$\lim_{n\to\infty}\frac{r_n - \theta k_n}{\sqrt{k_n}} = 0.$$

Εάν $\rho_n = r_n - \theta k_n$ *και η ακολουθία*

$$l_n = n\frac{(\theta - p)e^{-k_n H(\theta,p)-\rho_n h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k_n}}, \quad n = 1, 2, \ldots$$

είναι άνω φραγμένη, τότε

$$P(S_{n,k} < r) \sim e^{-l_n}$$

με το ρυθμό σύγκλισης να είναι τάξεως $O(\frac{\rho_n^2+1}{k_n})$.

Από την τελευταία πρόταση μπορούμε να συμπεράνουμε ότι για μεγάλες τιμές των παραμέτρων n, k, r και $p < r/k \neq 1$ η αθροιστική συνάρτηση κατανομής της $S_{n,k}$ προσεγγίζεται από τον παρακάτω τύπο (στον τύπο του l_n, αντικαταστήσαμε τα r_n, k_n,θ, με $r, k, r/k$, αντιστοίχως)

$$P(S_{n,k} < r) \approx \exp\left(-n\frac{(r-kp)e^{-kH(r/k,p)-\rho h(r/k,p)}}{\sqrt{2\pi rk(k-r)}}\right).$$

2.3.3 Αποτελέσματα ακραίων τιμών

Αντικείμενο έρευνας για πολλές δεκαετίες, έχει αποτελέσει η ασυμπτωτική συμπεριφορά κινούμενων αθροισμάτων από ανεξάρτητες και ισόνομες (όχι απαραίτητα δίτιμες) τυχαίες μεταβλητές. Τέτοιου είδους αποτελέσματα, αναφέρονται ως νόμοι Erdös-Rényi και αφορούν την τ.μ.

$$U_n = \max_{1\leq i\leq n-k+1} \sum_{j=i}^{i+k-1} Y_j$$

όπου $Y_1, Y_2, \ldots$ είναι μια ακολουθία από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές. Το κλασικό θεώρημα από την εργασία Erdös-Rényi (1970), εξασφαλίζει τη σχεδόν βέβαιη σύγκλιση στη μονάδα, της ακολουθίας $U_n/(ak_n)$. Πιο συγκεκριμένα αποδεικνύεται ότι ισχύει $P(\lim_{n\to\infty} U_n/(ak_n) = 1) = 1$, για μια μεγάλη κλάση κατανομών των Y_i, όπου $k = k_n = \lfloor c\ln(n)\rfloor$ για κάποια θετική σταθερά c, ενώ το $a > 0$ εξαρτάται από την τιμή της c, και την κατανομή των Y_i. Ακόμη, οι Deheuvels and Devroye (1987), απέδειξαν ότι κάτω από την υπόθεση ότι οι Y_i ακολουθούν μια non-lattice κατανομή με μηδενική μέση τιμή, ισχύει

$$\lim_{n\to\infty} P\left(\frac{U_n - b_n}{a_n} \leq x\right) = \Lambda(x)$$

όπου $\Lambda(x) = \exp(-e^{-x})$, $x \in \Re$ είναι η συνάρτηση κατανομής Gumbel και $a_n, b_n \in \Re$ είναι κατάλληλες σταθερές κανονικοποίησης.

Τα επόμενα δύο αποτελέσματα (ακραίων τιμών), αφορούν την απλή συνάρτηση σάρωσης $S_{n,k}$. Το γεγονός ότι έχουμε μια ακολουθία δίτιμων τ.μ. (και επομένως κατανομή lattice), μας επιτρέπει να εκφράσουμε, με συγκεκριμένους κλειστούς τύπους, τις σταθερές κανονικοποίησης. Έτσι, το ακόλουθο θεώρημα προέρχεται από μία επαναδιατύπωση ενός αποτελέσματος, των Arratia, Gordon and Waterman (1990).

Θεώρημα 2.3.6 *Έστω $k > -(\ln p)^{-1} \ln n$, και ας συμβολίσουμε με $\theta = \theta(n,k,p) \in (p,1)$ τη μοναδική λύση[2] της εξίσωσης*

$$H(\theta, p) = \frac{\ln n}{k}$$

και με b_n τις σταθερές κανονικοποίησης, όπου

$$b_n = \theta \frac{\ln n}{H(\theta,p)} - \frac{1}{2h(\theta,p)} \ln(\ln n) - \frac{1}{2h(\theta,p)} \ln(\frac{2\pi\theta(1-\theta)}{H(\theta,p)}) + \frac{\ln(\theta - p)}{h(\theta,p)}.$$

Τότε, για κάθε $\varepsilon > 0$ τέτοιο ώστε $1 + \varepsilon \leq -(\ln n)^{-1} k \ln p \leq 1/\varepsilon$ ισχύει (καθώς $n, k \to \infty$)

$$\sup_x |P(S_{n,k} - b_n < x) - \Lambda(h(\theta,p)x)| \to 0$$

Το supremum υπολογίζεται επάνω σ' όλα τα $x \in \Re$, τέτοια ώστε το $x + b_n$ να είναι ένας ακέραιος θετικός αριθμός.

Επιπρόσθετα, από την Πρόταση 2.3.4 προκύπτει άμεσα ένα ανάλογο αποτέλεσμα για την $S_{n,k}$, με οριακή συνάρτηση κατανομής την Gumbel, και κατάλληλα ορισμένες σταθερές κανονικοποίησης (Boutsikas and Koutras (2006)).

Θεώρημα 2.3.7 *Για σταθερό $p \in (0,1)$, και για κάθε θ που ανήκει στο διάστημα $(p,1)$ ορίζουμε*

$$\begin{aligned} k_n &= \lfloor \ln n / H(\theta,p) \rfloor, \\ b_n &= k_n \theta + \frac{1}{h(\theta,p)} \ln \frac{n(\theta-p)e^{-k_n H(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k_n}}, \\ \epsilon_n(y) &= \left(b_n + \frac{y}{h(\theta,p)} \right) - \left\lfloor b_n + \frac{y}{h(\theta,p)} \right\rfloor. \end{aligned}$$

Τότε

$$\lim_{n\to\infty} \left[P\left(\frac{S_{n,k} - b_n}{1/h(\theta,p)} < y \right) - \Lambda(y - \epsilon_n(y) h(\theta,p)) \right] = 0$$

με ρυθμό σύγκλισης, τάξεως $O(\frac{(\ln k_n)^2}{k_n})$.

[2] Επειδή, $\frac{d}{d\theta} H(\theta,p) = h(\theta,p) > 0$, η συνάρτηση $H(\theta,p)$ είναι αύξουσα στο διάστημα από 0 έως $-\ln p$. Άρα, η εξίσωση $H(\theta,p) = c$ έχει μοναδική λύση $\theta \in (p,1)$ για $0 < c < -\ln p$

2.3.4 Αριθμητικές συγκρίσεις

Στην παράγραφο αυτή θα επιχειρήσουμε μια αριθμητική σύγκριση των αποτελεσμάτων που προηγήθηκαν και αφορούν την προσέγγιση της συνάρτησης κατανομής της $S_{n,k}$. Ουσιαστικά, θα εξετάσουμε (αριθμητικά) την ποιότητα των φραγμάτων, που εξάγονται από τα παραπάνω θεωρήματα, ενώ για τις ασυμπτωτικές ιδιότητες μερικών απ' αυτών, είχαμε μιλήσει και στην Παράγραφο 2.3.2.

Με βάση τα προηγούμενα αποτελέσματα, για τη συνάρτηση κατανομής $P(S_{n,k} < r)$,$r = 1,2,\ldots,k$, ισχύει ότι

$$P(S_{n,k} < r) \in [e^{-b_2} - UB_1, \quad e^{-b_1} + UB_1], \tag{2.3.3}$$

(για κάθε n,r,k,p με $p<r/k< 1$) όπου UB_1 είναι το άνω φράγμα του Θεωρήματος 2.3.1 και (δείτε (2.3.1)),

$$b_1 = (n-k+1)b(r;k,p)\left((\frac{r}{k}-p) + 2(1-\frac{r}{k})(1-B(r;k,p))\right),$$
$$b_2 = (n-k+1)b(r;k,p)(\frac{r}{k}-p).$$

Δυο διαφορετικά διαστήματα (για κάθε n, r, k, p), στα οποία κινείται η $P(S_{n,k} < r)$, δίδονται από τις παρακάτω σχέσεις

$$P(S_{n,k} < r) \in [e^{-\mu} - UB_2, \quad e^{-\mu} + UB_2], \tag{2.3.4}$$
$$P(S_{n,k} < r) \in [e^{-\lambda_1} - UB_3, \quad e^{-\lambda_1} + UB_3], \tag{2.3.5}$$

όπου

$$\mu = (n-k+1)(1-B(r-1;k,p)), \quad \lambda_1 = (n-k+1)B(r-1;k,p),$$

και τα UB_2, UB_3 είναι τα άνω φράγματα που εμφανίζονται στο Θεώρημα 2.3.2.

Από το Θεώρημα 2.3.4, παίρνουμε

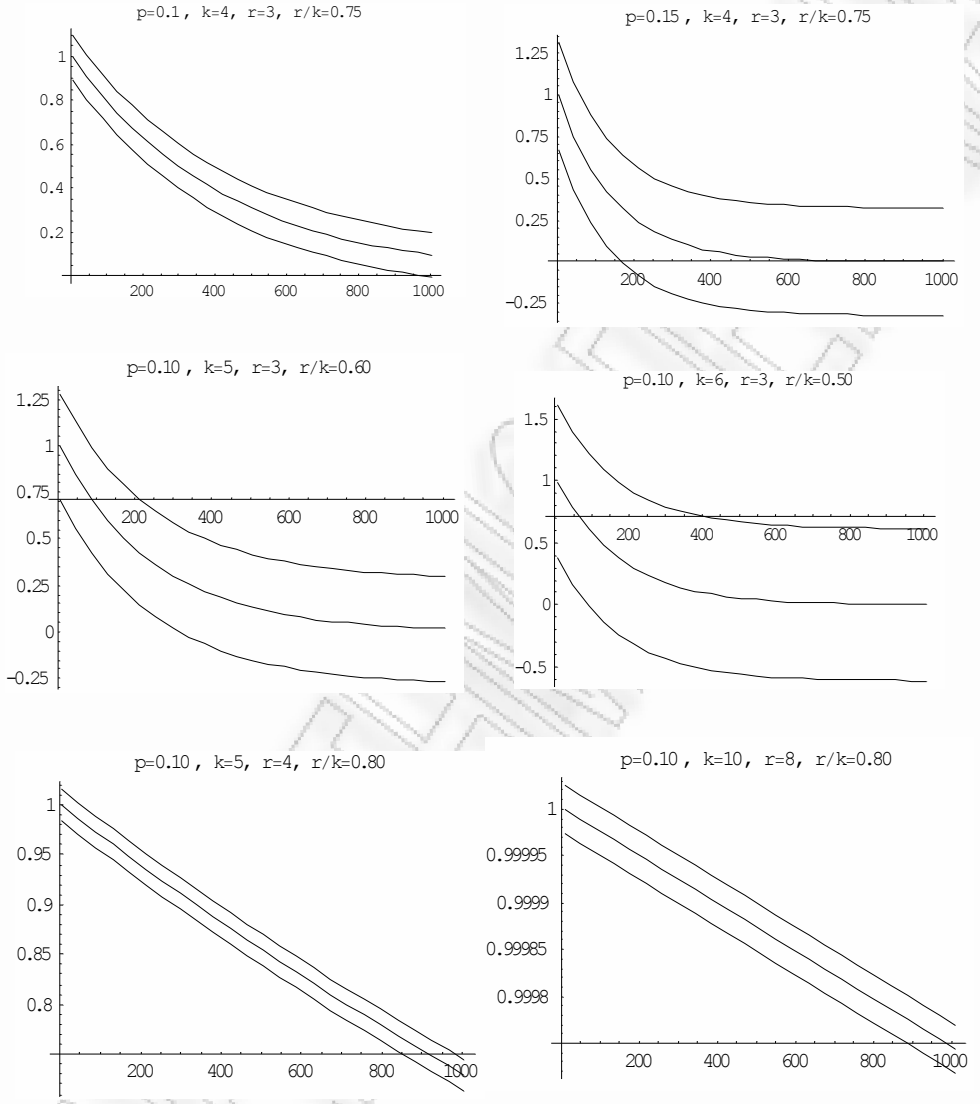
$$P(S_{n,k} < r) \in [e^{-\lambda_2} - UB_4, \quad e^{-\lambda_2} + UB_4], \tag{2.3.6}$$

όπου UB_4 να είναι το άνω φράγμα του τελευταίου θεωρήματος και

$$\lambda_2 = (n-k+1)\binom{k-1}{r-1}p^r q^{k-r+1}.$$

Τέλος, έχουμε και το παρακάτω διάστημα

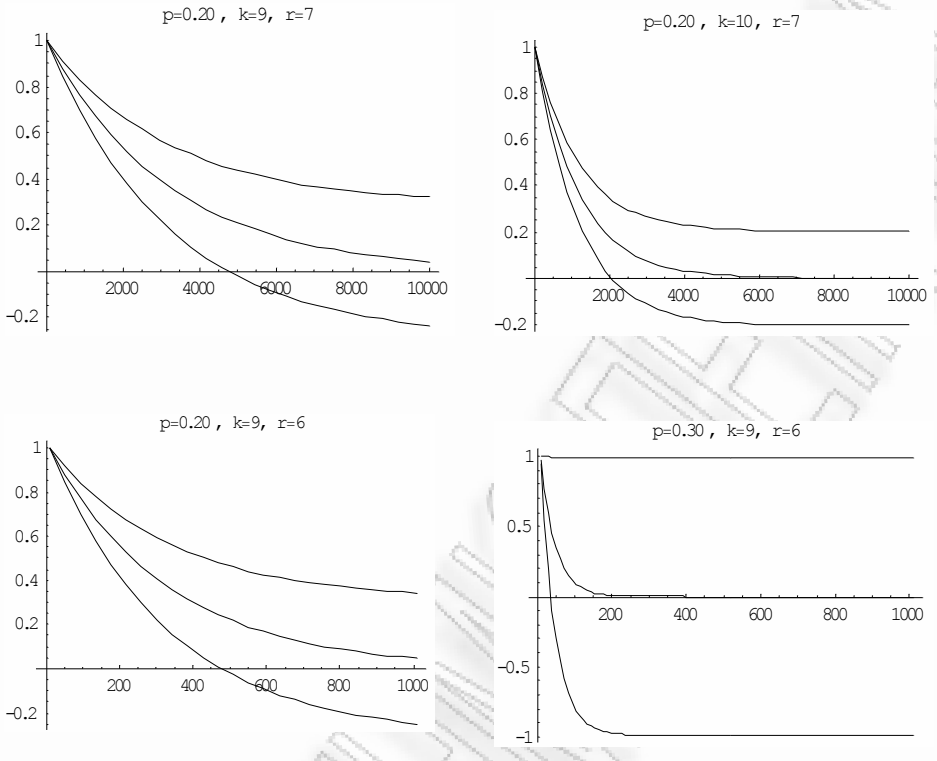
$$P(S_{n,k} < r) \in [e^{-\lambda_3} - UB_5, \quad e^{-\lambda_3} + UB_5], \tag{2.3.7}$$



Σχήμα 2.3.1: Γράφημα του διαστήματος (2.3.3), με κεντρική γραμμή την e^{-b_2}.

όπου το UB_5 δίδεται από το Πόρισμα 2.3.1 και $\lambda_3 = (n-k+1)f(r;k,p)$.

Να σημειώσουμε ότι δε θα αναφερθούμε στο διάστημα που προκύπτει από το Θεώρημα 2.3.3, καθώς η μορφή του δεν επιτρέπει τον άμεσο υπολογισμό του, και παράλληλα οι διαθέσιμες προσεγγίσεις, προσφέρουν κακής ποιότητας αποτελέσματα (πολύ μεγάλα διαστήματα). Όπως είδαμε και στις προηγούμενες παραγράφους, οι συνθήκες κάτω από τις οποίες τα παραπάνω φράγματα γίνονται αποτελεσματικά, διαφέρουν, κάτι που θα φανεί και από τους αριθμητικούς μας υπολογισμούς.

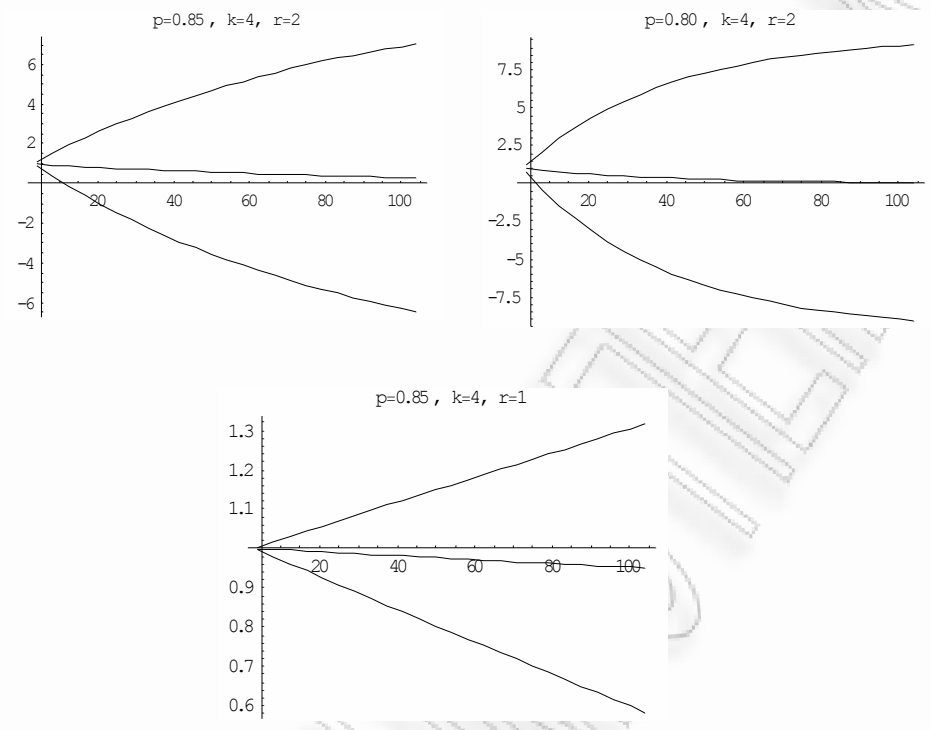


Σχήμα 2.3.2: Γράφημα του διαστήματος (2.3.4).

Στο Σχήμα 2.3.1 και στις δυο πρώτες γραφικές παραστάσεις, παρατηρούμε ότι για μια συγκεκριμένη επιλογή των k, r, καθώς αυξάνουμε την πιθανότητα επιτυχίας (από 0.10 σε 0.15) το διάστημα (2.3.3) γίνεται χειρότερο. Στα δυο επόμενα γραφήματα, του ίδιου σχήματος, καθώς αυξάνουμε το k (ενώ το r παρέμεινε σταθερό), το διάστημα απέκτησε μεγαλύτερο εύρος. Τέλος, αυξάνοντας και το k και το r με τέτοιο τρόπο ώστε το r/k να μείνει σταθερό (από $k = 5, r = 4$ σε $k = 10, r = 8$), το φράγμα βελτιώθηκε.

Από το Σχήμα 2.3.2, παρατηρούμε πως το διάστημα (2.3.4) βελτιώνεται αυξάνοντας το k, για δεδομένα p, r (δυο πρώτες γραφικές παραστάσεις), ενώ αυξάνοντας μόνο το p, έχουμε τα αντίστροφα αποτελέσματα.

Το διάστημα (2.3.5), μας δίνει προσεγγίσεις κακής ποιότητας, σ' όλες σχεδόν τις περιπτώσεις που μελετήσαμε, όπως φαίνεται και από το Σχήμα 2.3.3. Στο Σχήμα 2.3.4, όπου αναφερόμαστε στο διάστημα (2.3.6), διαπιστώνουμε πως αν μειώσουμε την τιμή του p, θα βελτιωθεί η προσέγγιση (δείτε τα δυο πρώτα γραφήματα), ενώ αντίθετα θα αυξηθεί το εύρος του διαστήματος, αν αυξήσουμε το k. Βέβαια, εάν αυξήσουμε και το k και το r, η προσέγγιση παραμένει το ίδιο καλή (δείτε το δεύτερο και το τέταρτο γράφημα).

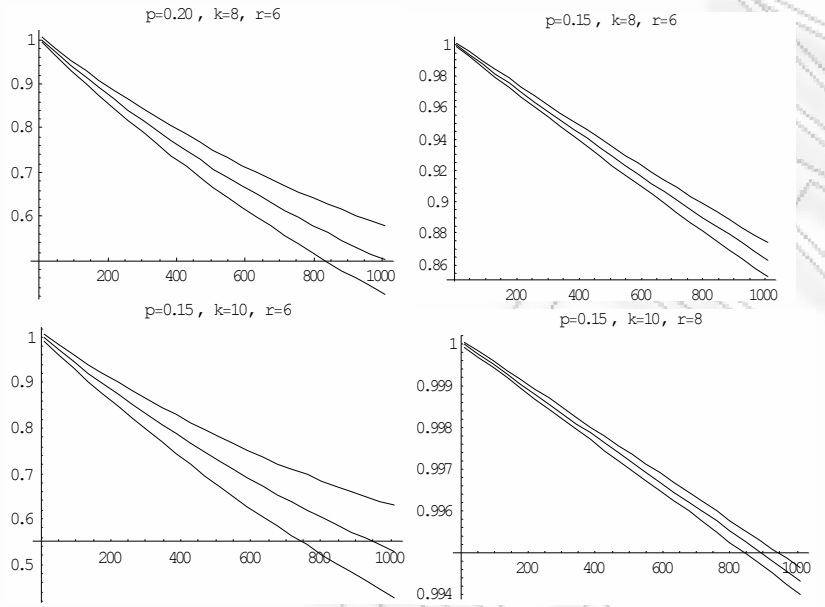


Σχήμα 2.3.3: Γράφημα του διαστήματος (2.3.5).

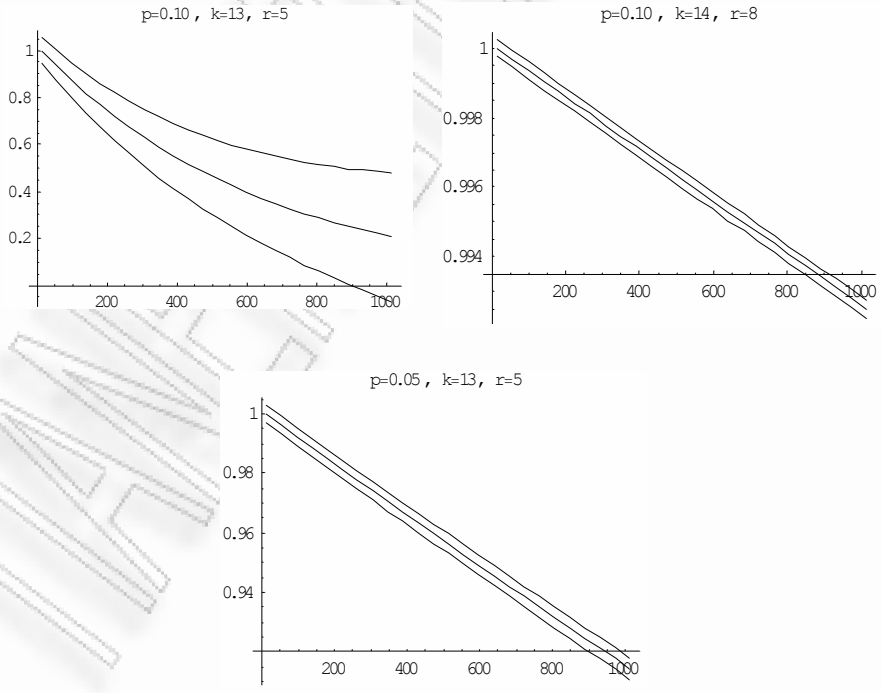
Τέλος, η ποιότητα του (2.3.7) (Σχήμα 2.3.5) φαίνεται να βελτιώνεται είτε αυξάνοντας τις τιμές των k, r είτε, μειώνοντας το p.

2.4 Γενικευμένες συναρτήσεις σάρωσης

Στην παρούσα ενότητα θα παρουσιάσουμε κάποια νέα αποτελέσματα σχετικά με τη γενικευμένη πολλαπλή συνάρτηση σάρωσης (μοντέλο υπερβάσεων, exceedance model), σε μια ακολουθία συνεχών τ.μ. Με βάση την Πρόταση 2.3.3, αποδεικνύουμε ένα θεώρημα για την προσέγγιση της κατανομής της $W_{n,k,r}(u_{a_n})$, μέσω μίας σύνθετης κατανομής Poisson, όπου a_n είναι μια κατάλληλη ακολουθία, θετικών πραγματικών αριθμών. Στη συνέχεια, εξετάζεται η ασυμπτωτική κατανομή της $W_{n,k,r}(u_{a_n})$ ή ισοδύναμα, της διατεταγμένης παρατήρησης $Y_{m:r:k}(n)$ (βλ. Παράγραφος 2.2), κάτω από την υπόθεση ότι η κατανομή των Y_i, ανήκει σ' ένα από τα τρία πεδία έλξης, των κλασικών κατανομών ακραίων τιμών, (Weibull, Frechet ή Gumbel). Τέλος, παρουσιάζονται ένα πλήθος από αριθμητικά αποτελέσματα, τα οποία φανερώνουν την ποιότητα των προσεγγίσεων μας.



Σχήμα 2.3.4: Γράφημα του διαστήματος (2.3.6).



Σχήμα 2.3.5: Γράφημα του διαστήματος (2.3.7).

2.4.1 Προσεγγίσεις μέσω σύνθετης κατανομής Poisson, για τη γενικευμένη πολλαπλή συνάρτηση σάρωσης

Ας θεωρήσουμε αρχικά μια ακολουθία $Y_1, Y_2, \ldots, Y_n$ από ανεξάρτητες και ισόνομες συνεχείς τ.μ., με συνάρτηση κατανομής F, και ένα κατώφλι $u = u_n$. Για να εκμεταλλευτούμε τα αποτελέσματα που προηγήθηκαν, υποθέτουμε ότι το κατώφλι u_n, μεταβάλλεται συναρτήσει του n, με τέτοιο τρόπο ώστε το ενδεχόμενο $Y_i > u_n$ να γίνεται ένα σπάνιο ενδεχόμενο, δηλαδή

$$\lim_{n\to\infty} P(Y_i > u_n) = 0$$

(προφανώς, μια ακολουθία u_n με την παραπάνω ιδιότητα, είναι δυνατόν να βρεθεί για οποιαδήποτε κατανομή F). Μια συνήθης συνθήκη που εξασφαλίζει κάτι τέτοιο, χωρίς να οδηγούμαστε σε τετριμμένα αποτελέσματα, είναι η ακόλουθη

$$\lim_{n\to\infty} nP(Y_i > u_n) = \lim_{n\to\infty} n\overline{F}(u_n) = \tau \in (0, \infty).$$

Χρησιμοποιώντας μια τέτοια ακολουθία πραγματικών αριθμών (u_n), μπορούμε να αποδείξουμε το επόμενο, ενδιαφέρον θεώρημα (Boutsikas et al (2008)), εκμεταλλευόμενοι το Θεώρημα 2.3.4 και το Πόρισμα 2.3.3.

Θεώρημα 2.4.1 *Έστω ανεξάρτητες και ισόνομες συνεχείς τ.μ. $Y_1, Y_2, \ldots, Y_n$ (με $Y_i \sim F$, για $i = 1, 2, \ldots, n$), u_n μια ακολουθία πραγματικών αριθμών, τέτοια ώστε*

$$\lim_{n\to\infty} n\overline{F}(u_n) = \tau > 0$$

και $a_n = n^{1/r}$. Τότε η κατανομή του πλήθους $W_{n,k,r}(u_{a_n})$ των κινούμενων παραθύρων μήκους k, τα οποία περιέχουν τουλάχιστον r υπερβάσεις του ορίου u_{a_n}, συγκλίνει σε μια σύνθετη κατανομή Poisson, με παράμετρο

$$\lambda = \binom{k-1}{r-1}\tau^r$$

και συνθέτουσα συνάρτηση πιθανότητας

$$g(x) = \frac{\binom{k-x-1}{r-2}}{\binom{k-1}{r-1}}, \quad x = 1, 2, \ldots, k-r+1. \tag{2.4.1}$$

Απόδειξη. Από τη σχέση

$$\lim_{n\to\infty} n\overline{F}(u_n) = \tau > 0,$$

προκύπτει άμεσα ότι

$$\lim_{n\to\infty} n\overline{F}(u_{a_n})^r = \lim_{n\to\infty} (a_n\overline{F}(u_{a_n}))^r = \tau^r.$$

Επιπλέον, οι τ.μ. $X_i(u_{a_n}) = I_{(u_{a_n},\infty)}(Y_i)$, $i = 1, 2, \ldots$, σχηματίζουν μια ακολουθία από δοκιμές Bernoulli, με πιθανότητα επιτυχίας $p_n = \overline{F}(u_{a_n})$. Τότε είναι φανερό ότι,

$$\lim_{n\to\infty} p_n = \lim_{n\to\infty} \overline{F}(u_{a_n})^r = 0.$$

Επίσης, για την ποσότητα

$$\lambda_n = (n-k+1)\binom{k-1}{r-1} p_n^r (1-p_n)^{k-r+1}$$

(δείτε και Θεώρημα 2.3.4) ισχύει

$$\begin{aligned}\lim_{n\to\infty} \lambda_n &= \lim_{n\to\infty} (n-k+1)\binom{k-1}{r-1} p_n^r (1-p_n)^{k-r+1} \\ &= \binom{k-1}{r-1} \lim_{n\to\infty} (n-k+1) p_n^r (1-p_n)^{k-r+1} = \binom{k-1}{r-1}\tau^r > 0\end{aligned}$$

αφού

$$\lim_{n\to\infty} (n-k+1)p_n^r = \lim_{n\to\infty} np_n^r = \tau^r, \quad \lim_{n\to\infty} (1-p_n)^{k-r+1} = 1.$$

Έτσι, μπορούμε να εφαρμόσουμε την Πρόταση 2.3.3, και να συμπεράνουμε ότι η $W_{n,k,r}(u_{a_n})$ συγκλίνει σε μια σύνθετη κατανομή Poisson, με συνθέτουσα συνάρτηση πιθανότητας την (2.4.1) και $\lambda = \lim_{n\to\infty} \lambda_n = \binom{k-1}{r-1}\tau^r$.

■

Σύμφωνα με το προηγούμενο θεώρημα, για μεγάλες τιμές του n η συνάρτηση κατανομής της $W_{n,k,r}(u_{a_n})$ (όπου $a_n = n^{1/r}$), προσεγγίζεται από μια σύνθετη κατανομή Poisson (με κατάλληλες παραμέτρους), για κάθε ακολουθία πραγματικών αριθμών u_n, με την ιδιότητα $\lim_{n\to\infty} n\overline{F}(u_n) = \tau \in (0,\infty)$. Σ' αυτήν την περίπτωση το σφάλμα της προσέγγισης παίρνει μικρές τιμές, και προσδιορίζεται ακριβώς, από το Θεώρημα 2.3.4.

Άμεση συνέπεια του προηγούμενου θεωρήματος, είναι το επόμενο πόρισμα, που αφορά την οριακή συμπεριφορά της διατεταγμένης παρατήρησης $Y_{m:r:k}(n)$.

Πόρισμα 2.4.1 *Εάν συμβολίσουμε με f_{CP} τη συνάρτηση πιθανότητας της σύνθετης κατανομής Poisson που περιγράφεται στο Θεώρημα 2.4.1, τότε ισχύει*

$$\lim_{n\to\infty} P(Y_{m:r:k}(n) \le u_{a_n}) = \sum_{i=0}^{m-1} f_{CP}(i).$$

Για την απόδειξη του Πορίσματος 2.4.1, αρκεί να ανακαλέσουμε τη σχέση ανάμεσα στην $Y_{m:r:k}(n)$ και την $W_{n,k,r}(u)$ (δείτε (2.2.1)).

Παρόλο που η ακολουθία u_{a_n} ορίζεται συνήθως στο σύνολο των θετικών ακέραιων n, από εδώ και στο εξής θα θεωρούμε την επέκταση της u_{a_n}, στο σύνολο θετικών πραγματικών $\Re^+$, δηλαδή $u_x = u(x)$, $x \in \Re^+$. Κάτω από αυτή την υπόθεση είναι δυνατόν να γράφουμε u_{a_n} όπου a_n είναι μία οποιαδήποτε ακολουθία πραγματικών αριθμών (όχι απαραίτητα ακέραιων)- κάτι που συναντάμε συχνά σ' αυτή την ενότητα. Αξίζει να αναφέρουμε ότι αυτή η τακτική, χρησιμοποιείται και στα παραδείγματα της Παραγράφου 2.4.3. Εάν όμως, δεν είναι εφικτή μια επέκταση της ακολουθίας u_{a_n}, στο $\Re^+$, τότε μπορούμε να χρησιμοποιήσουμε την ακολουθία $u_{\lfloor a_n \rfloor}$, στη θέση της u_{a_n}, με όλα τα θεωρητικά αποτελέσματα να παραμένουν ορθά.

Βλέποντας το Θεώρημα 2.4.1, διαπιστώνουμε ότι η μόνη δυσκολία για τον προσδιορισμό της ασυμπτωτικής κατανομής, είναι η εύρεση της ακολουθίας u_n, για την οποία πρέπει να ισχύει

$$\lim_{n\to\infty} n\overline{F}(u_n) = \tau > 0.$$

Στην επόμενη παράγραφο, θα αντιμετωπίσουμε το πρόβλημα αυτό με την υπόθεση ότι η κατανομή από την οποία προέρχεται το τυχαίο δείγμα, ανήκει σε κάποιο από τα τρία πεδία έλξης των κατανομών των ακραίων τιμών, εκμεταλλευόμενοι έτσι και το Θεώρημα 2.1.3.

Αξίζει να τονίσουμε ότι το όριο

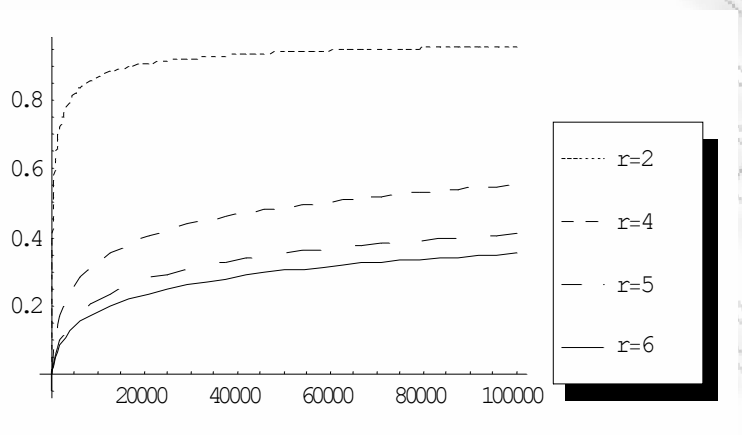
$$\lim_{n\to\infty} \lambda_n = \binom{k-1}{r-1}\tau^r > 0$$

υπολογίστηκε, χρησιμοποιώντας την ιδιότητα

$$(1-p_n)^{k-r+1} = \left(1 - \frac{a_n\overline{F}(u_{a_n})}{a_n}\right)^{k-r+1} \sim \left(1 - \frac{\tau}{a_n}\right)^{k-r+1} \underset{n\to\infty}{\to} 1$$

Η παραπάνω σύγκλιση όμως, είναι εξαιρετικά αργή, ακόμη και για μικρές τιμές του r. Για παράδειγμα, εάν $k = 4, r = 2, \tau = 2$ τότε $q_{100}^{k-r+1} \approx 0.512, q_{1000}^{k-r+1} \approx 0.822, q_{10000}^{k-r+1} \approx 0.941$. Τα πράγματα γίνονται ακόμη χειρότερα όταν το r παίρνει μεγαλύτερες τιμές (δείτε και Σχήμα 2.4.1). Επομένως, παρότι τα ασυμπτωτικά αποτελέσματα από το Θεώρημα 2.4.1 ισχύουν για $n \to \infty$, η κατανομή της $W_{r,k,n}(u_{a_n})$ δεν μπορεί να προσεγγιστεί ικανοποιητικά για μέτριες τιμές του n, μέσω της σύνθετης κατανομής Poisson $CP(\lambda_n, G)$. Σε τέτοιες περιπτώσεις είναι προτιμότερο η προσέγγιση να γίνει μέσω της σύνθετης κατανομής Poisson $CP(\lambda_n^*, G)$, με παράμετρο

$$\lambda_n^* = \binom{k-1}{r-1}\tau^r \left(1 - \frac{\tau}{a_n}\right)^{k-r+1} = \lambda\left(1 - \frac{\tau}{a_n}\right)^{k-r+1}. \tag{2.4.2}$$



Σχήμα 2.4.1: Γράφημα της ποσότητας $(1-\tau/a_n)^{k-r+1}$ για $k=8, \tau=2$ και διάφορες τιμές του r

Την προσέγγιση αυτή υιοθετούμε και στα παραδείγματα της Παραγράφου 2.4.3.

Ένα αποτέλεσμα παρόμοιο μ' αυτό του Θεωρήματος 2.4.1, δημοσιεύθηκε από τον Dudkiewicz (1998). Όμως, το αποτέλεσμα αυτό (Θεώρημα 2.4.2), όπως θα καταλάβουμε, αναφέρεται στην ειδική περίπτωση που μελετάμε κινούμενα ελάχιστα και εφαρμόζεται κάτω από διαφορετικές συνθήκες, για τις τιμές των παραμέτρων k, r.

Θεώρημα 2.4.2 *Έστω μια ακολουθία θετικών ακέραιων αριθμών k_n, τέτοια ώστε*

$$\lim_{n\to\infty} \frac{k_n}{\ln n} = d \geq 0,$$

και μια ακολουθία u_n, με την ιδιότητα

$$\lim_{n\to\infty} n\overline{F}(u_n)^{k_n} = \tau' > 0.$$

Τότε για την ασυμπωτική κατανομή της διατεταγμένης παρατήρησης $Y_{m,k_n,k_n}(n)$, ισχύει

$$\lim_{n\to\infty} P(Y_{m,k_n,k_n}(n) \leq u_n) = \sum_{i=0}^{m-1} f_{CP}(i).$$

όπου οι $f_{CP}(i)$ δίνονται από την (2.1.2), με

$$\lambda_1 = \tau', \quad \lambda_j = 0, \ \textit{για } j \geq 2$$

όταν $d = 0$, και

$$\lambda_j = \tau'(1-\exp(-1/d))^2 \exp(-(j-1)/d), \ \textit{για } j \geq 1,$$

όταν $d > 0$.

Στην ειδική περίπτωση της $Y_{n-k_n+1,k_n,k_n}(n)$, προκύπτει ένα αποτέλεσμα που είχαν αποδείξει οι Canfield and McCormick (1992). Επίσης, στην εργασία Dudkiewicz (1998) δε γίνεται καμία αναφορά για τον προσδιορισμό της ακολουθίας u_n.

Πριν κλείσουμε την παρούσα ενότητα, αξίζει να αναφερθούμε σ' ένα δείκτη για το βαθμό εξάρτησης ανάμεσα στις μεταβλητές που μελετάμε, και τον οποίο οι Embrechts et al (1997), ονόμασαν *extremal index*. Ο δείκτης αυτός παίρνει τιμές στο διάστημα από 0 έως 1, και εισάγεται μέσω του παρακάτω ορισμού.

Ορισμός 2.4.1 *Έστω μια ακολουθία (αυστηρώς στάσιμη) από ισόνομες τ.μ.* $Y_i, i = 1, 2, \ldots, n$ *(με* $Y_i \sim F, \forall i$*) και* θ *ένας μη αρνητικός αριθμός. Ας υποθέσουμε ότι για κάθε* $\tau > 0$*, υπάρχει μια ακολουθία πραγματικών αριθμών* u_n*, τέτοια ώστε*

$$\lim_{n\to\infty} n\overline{F}(u_n) = \tau,$$
$$\lim_{n\to\infty} P(M_n \leq u_n) = e^{-\theta\tau},$$

όπου $M_n = \max\{Y_1, Y_2, \ldots, Y_n\}$*. Τότε, η ποσότητα* θ *ονομάζεται extremal index της ακολουθίας* $Y_i, i = 1, 2, \ldots, n$*.*

Ο extremal index έχει αρκετά ενδιαφέρουσες ιδιότητες, όπως αυτή που περιγράφεται από την επόμενη πρόταση.

Πρόταση 2.4.1 *Έστω μια αυστηρώς στάσιμη ακολουθία* $Y_i, i = 1, 2, \ldots, n$*, με* $Y_i \sim F$ *και* $F \in MDA(H)$*, και σταθερές κανονικοποίησης* c_n, d_n*. Τότε*

$$\lim_{n\to\infty} P\left(\frac{M_n - d_n}{c_n} \leq x\right) = H^\theta(x)$$

εάν και μόνο εάν

$$\lim_{n\to\infty} P\left(\frac{M'_n - d_n}{c_n} \leq x\right) = H(x),$$

όπου με M'_n *συμβολίζουμε τη μέγιστη παρατήρηση, από τις* $Y_i, i = 1, 2, \ldots, n$*, με την επιπλέον προϋπόθεση ότι είναι ανεξάρτητες τ.μ.*

Αξίζει να σημειώσουμε ότι η κατανομή $H^\theta(x)$ ανήκει στην ίδια οικογένεια με την $H(x)$, δηλαδή υπάρχουν c και d, με την ιδιότητα, $H^\theta(x) = H(cx+d)$. Επίσης, μπορεί να αποδειχθεί πως, ο ορισμός του extremal index είναι ανεξάρτητος από την επιλογή της u_n. Ακόμη, εάν η ακολουθία $Y_i, i = 1, 2, \ldots, n$ έχει extremal index θ, τότε οι δύο συνθήκες του Ορισμού 2.4.1, και η σχέση

$$\lim_{n\to\infty} P\left(M'_n \leq u_n\right) = e^{-\tau},$$

είναι ισοδύναμες (να επισημάνουμε πως υπάρχουν ακολουθίες τυχαίων μεταβλητών, οι οποίες δεν έχουν extremal index).

Για την περίπτωση που μελετάμε, στην παρούσα παράγραφο, λαμβάνοντας υπόψιν το Πόρισμα 2.4.1 και τη σχέση (2.2.2), έχουμε

$$\lim_{n\to\infty} P(\max\{Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}\} \leq u_{a_n}) = \lim_{n\to\infty} P(W_{n,k,r}(u_{a_n}) = 0) = e^{-\lambda}.$$

Η παραπάνω σχέση αφορά την ασυμπτωτική συμπεριφορά της μέγιστης παρατήρησης, από ένα σύνολο από εξαρτημένες τ.μ., τις $Y_{r:k}^{(1)}$,$Y_{r:k}^{(2)}$,...,$Y_{r:k}^{(n-k+1)}$. Οι περιθώριες κατανομές των $Y_{r:k}^{(i)}$, μπορούν να εκφραστούν μέσω των διωνυμικών κατανομών, με παραμέτρους k και $\overline{F}(u_{a_n})$, με τον εξής τρόπο

$$nP(Y_{r:k}^{(i)} > u_{a_n}) = n\sum_{i=r}^{k} \binom{k}{i} \overline{F}(u_{a_n})^i F(u_{a_n})^{k-i}$$

και

$$n\sum_{i=r}^{k} \binom{k}{i} \overline{F}(u_{a_n})^i F(u_{a_n})^{k-i} \sim \binom{k}{r} (a_n \overline{F}(u_{a_n}))^r \to_{n\to\infty} \binom{k}{r} \tau^r.$$

Εάν υποθέσουμε τώρα ότι οι $Y_{r:k}^{(i)}$ ήταν ανεξάρτητες, η ασυμπτωτική συμπεριφορά της μέγιστης παρατήρησης θα ήταν

$$\lim_{n\to\infty} P(\max\{Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}\} \leq u_{a_n}) = \lim_{n\to\infty} \left(1 - \frac{nP(Y_{r:k}^{(i)} > u_{a_n})}{n}\right)^n = e^{-\lambda_{ind}}$$

όπου $\lambda_{ind} = \binom{k}{r}\tau^r$. Ο λόγος

$$\frac{\lambda}{\lambda_{ind}} = \frac{\binom{k-1}{r-1}\tau^r}{\binom{k}{r}\tau^r} = \frac{r}{k}$$

είναι ο extremal index της ακολουθίας $Y_{r:k}^{(i)}$, $i = 1, 2, \ldots, n-k+1$ και χαρακτηρίζει το βαθμό εξάρτησης ανάμεσα στις $Y_{r:k}^{(i)}$. Όπως έχουμε αναφέρει, ο extremal index παίρνει τιμές από 0 έως 1, και όσο μικρότερος είναι, τόσο μεγαλύτερη φαίνεται να είναι η εξάρτηση ανάμεσα στις τ.μ. της ακολουθίας. Έτσι, με βάση τα προηγούμενα, ο extremal index για την ακολουθία $Y_{r:k}^{(i)}$, $i = 1, 2, \ldots, n-k+1$ μειώνεται καθώς το k αυξάνεται ή το r μειώνεται, και επομένως, η εξάρτηση γίνεται μεγαλύτερη (γεγονός το οποίο συμβαδίζει και με τη διαίσθησή μας).

2.4.2 Ασυμπτωτική μελέτη της συνάρτησης σάρωσης, με βάση το πεδίο έλξης των συνεχών τ.μ.

Στην παρούσα ενότητα θα μελετήσουμε κάποια ασυμπτωτικά αποτελέσματα, για τη γενικευμένη πολλαπλή συνάρτηση σάρωσης $W_{n,k,r}(u)$, κάτω από την υπόθεση ότι η συνάρτηση

κατανομής F, των συνεχών τ.μ. Y_i, ανήκει σ' ένα από τα τρία μέγιστα πεδία έλξης (MDA), των κατανομών Φ_a, Ψ_a ή Λ.

Θεώρημα 2.4.3 *Εάν* $\overline{F} \in MDA(H)$ *με σταθερές κανονικοποίησης* $c_n > 0, d_n \in \Re$ *τότε*

$$\lim_{n\to\infty} P\left(\frac{Y_{m:r:k}(n) - d_{a_n}}{c_{a_n}} \leq x\right) = \lim_{n\to\infty} P(W_{n,k,r}(c_{a_n}x + d_{a_n}) < m) = \sum_{i=0}^{m-1} f_{CP}(x;i)$$

όπου $a_n = n^{1/r}$ *και* $f_{CP}(x;\cdot)$, *είναι η συνάρτηση πιθανότητας της σύνθετης κατανομής Poisson, με παράμετρο*

$$\lambda(x) = \binom{k-1}{r-1}(-\ln H(x))^r \tag{2.4.3}$$

και συνθέτουσα κατανομή την (2.4.1).

Απόδειξη. Επειδή $\overline{F} \in MDA(H)$, έχουμε (Θεώρημα 2.1.3)

$$\lim_{n\to\infty} n\overline{F}(c_n x + d_n) = -\ln H(x), \quad x \in \Re.$$

Εφαρμόζοντας το Θεώρημα 2.4.1 για

$$u_n = c_n x + d_n \text{ και } \tau = -\ln H(x),$$

καταλήγουμε ότι η $W_{n,k,r}(u_{a_n})$ συγκλίνει (ασθενώς) στη σύνθετη κατανομή Poisson, με παράμετρο

$$\binom{k-1}{r-1}\tau^r = \binom{k-1}{r-1}(-\ln H(x))^r = \lambda(x) \tag{2.4.4}$$

και αντίστοιχη συνθέτουσα κατανομή, την (2.4.1). Η απόδειξη ολοκληρώνεται με χρήση της ισότητας (δείτε (2.2.1))

$$P\left(Y_{m:r:k}(n) \leq c_n x + d_n\right) = P(W_{n,k,r}(c_n x + d_n) < m).$$

■

Με βάση το Θεώρημα 2.4.3, η κατανομή της $W_{n,k,r}(c_{a_n}x + d_{a_n})$, μπορεί να προσεγγιστεί για μεγάλες τιμές του n, από μία σύνθετη Poisson, με παράμετρο $\lambda(x)$ που δίδεται από την (2.4.4). Επαναλαμβάνοντας τα σχόλια όμως που ακολούθησαν το Θεώρημα 2.4.1, μπορούμε κάλλιστα να βελτιώσουμε την προσέγγιση, εάν αντικαταστήσουμε το $\lambda(x)$ με το (βλ. (2.4.2))

$$\lambda^*(x) = \lambda(x)\left(1 - \frac{\tau}{a_n}\right)^{k-r+1} = \binom{k-1}{r-1}(-\ln H(x))^r\left(1 + \frac{\ln H(x)}{n^{1/r}}\right)^{k-r+1}.$$

Κάποιες πιθανές επιλογές για τα c_n, d_n, όπως και οι τιμές των παραμέτρων $\lambda(x)$, $\lambda^*(x)$, για κάθε πεδίο έλξης ξεχωριστά, δίνονται παρακάτω (χρησιμοποιούμε το συμβολισμό x_F, για το άνω πέρας του στηρίγματος της κατανομής F, δηλαδή, $x_F = \sup\{x \in \Re : F(x) < 1\}$):

α. Μέγιστο πεδίο έλξης της Frechet

Εάν $\overline{F} \in MDA(\Phi_a)$ τότε $x_F = \infty$, και μία πιθανή επιλογή των c_n, d_n είναι η

$$c_n = F^{-1}(1 - n^{-1}) \text{ και } d_n = 0$$

όπου με F^{-1} συμβολίζουμε τη γενικευμένη αντίστροφη της F. Επειδή,

$$-\ln H(x) = -\ln \Phi_a(x) = x^{-a}$$

τα $\lambda(x)$, $\lambda^*(x)$ παίρνουν τη μορφή

$$\lambda(x) = \binom{k-1}{r-1} x^{-ra}, \lambda^*(x) = \binom{k-1}{r-1} x^{-ra} \left(1 - \frac{x^{-a}}{n^{1/r}}\right)^{k-r+1}, \quad x > 0.$$

Χαρακτηριστικές κατανομές, από το συγκεκριμένο πεδίο έλξης (με «βαριά» δεξιά ουρά) είναι οι Cauchy, Pareto και Loggamma.

β. Μέγιστο πεδίο έλξης της (Reversed) Weibull

Εάν $\overline{F} \in MDA(\Psi_a)$ τότε το x_F είναι πεπερασμένο, και μια δυνατή επιλογή των c_n, d_n είναι η

$$c_n = x_F - F^{-1}(1 - n^{-1}) \text{ και } d_n = x_F.$$

Οι παράμετροι της σύνθετης κατανομής Poisson

$$\lambda(x) = \binom{k-1}{r-1} (-x)^{ra}, \tag{2.4.5}$$

$$\lambda^*(x) = \binom{k-1}{r-1} (-x)^{ra} \left(1 - \frac{(-x)^a}{n^{1/r}}\right)^{k-r+1}, \quad x \leq 0. \tag{2.4.6}$$

Στο πεδίο έλξης της (Reversed) Weibull, ανήκουν π.χ. η Ομοιόμορφη κατανομή και η κατανομή Βήτα.

γ. Μέγιστο πεδίο έλξης της Gumbel

Εάν $\overline{F} \in MDA(\Lambda)$ τότε η F μπορεί να γραφεί ως (βλ. π.χ. Embrechts et al (1997))

$$\overline{F}(x) = c(x) e^{-\int_z^x \frac{g(t)}{a(t)} dt}, \quad z < x < x_F$$

όπου z είναι ένας πραγματικός αριθμός, με $z < x_F$ και c, g είναι (μετρήσιμες) συναρτήσεις τέτοιες ώστε $c(x) \to c_0 > 0$, και $g(x) \to 1$, όταν $x \uparrow x_F$. Η $a(\cdot)$, είναι μια συνάρτηση θετική, απολύτως συνεχής, με παράγωγο a', τέτοια ώστε $a'(x) \to 0$, καθώς $x \uparrow x_F$. Μία επιλογή για την a, είναι η

$$a(x) = \int_x^{x_F} \frac{\overline{F}(t)}{\overline{F}(x)} dt.$$

ενώ για τις σταθερές κανονικοποίησης, μπορούμε να πάρουμε

$$d_n = F^{-1}(1 - n^{-1}) \text{ και } c_n = a(d_n).$$

Οι παράμετροι $\lambda(x)$, $\lambda^*(x)$, ικανοποιούν τώρα τις σχέσεις

$$\lambda(x) = \binom{k-1}{r-1} e^{-rx}, \tag{2.4.7}$$

$$\lambda^*(x) = \binom{k-1}{r-1} e^{-rx} \left(1 - \frac{e^{-x}}{n^{1/r}}\right)^{k-r+1}. \tag{2.4.8}$$

Η Κανονική κατανομή, η Εκθετική και η Γάμμα, είναι κάποιες από τις κατανομές του συγκεκριμένου πεδίου έλξης.

Όπως έχουμε ήδη διαπιστώσει, η περίπτωση $m = 1$, του Θεωρήματος 2.4.3, αναφέρεται στην ασυμπτωτική συμπεριφορά της μέγιστης παρατήρησης, από συγκεκριμένες κινούμενες διατεταγμένες παρατηρήσεις, μήκους k, δηλαδή,

$$Y_{1:r:k} = \max\{Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}\}.$$

Στις εφαρμογές, τα Y_i μπορεί να είναι τιμές από μετρήσεις κάποιου μεγέθους (σε μια συγκεκριμένη κλίμακα), π.χ. ωριαίες μετρήσεις του επιπέδου της θάλασσας, καθημερινές απαιτήσεις σε κάποιο χαρτοφυλάκιο, μηνιαίες θερμοκρασίες κτλ. Τότε, η στατιστική συνάρτηση $Y_{r:k}^{(i)}$ είναι ένα μέτρο θέσης, για μια δεδομένη χρονική περίοδο (π.χ. η $Y_{r:k}^{(i)}$ μπορεί να είναι η διάμεσος από k συνεχόμενες ημέρες) και επομένως, η $Y_{1:r:k}$ αναφέρεται στη μέγιστη παρατήρηση από τα προηγούμενα μέτρα θέσης. Προβλήματα στα οποία τα παραπάνω θεωρητικά αποτελέσματα βρίσκουν εφαρμογές, συναντάμε στη συλλογή και την αποθήκευση δεδομένων (λόγω, π.χ., περιορισμένης χωρητικότητας), στη σάρωση ενός χώρου περιορισμένου εύρους κτλ. Μια άλλη περίπτωση, που μπορεί να χρησιμοποιηθεί η $Y_{1:r:k}$, είναι σε διαγράμματα απεικόνισης μέτρων θέσης (όπως είναι τα διαγράμματα απεικόνισης κινούμενης μέσης τιμής), όπου θέλοντας να έχουμε πιο ανθεκτικά (robust) διαγράμματα σε ακραίες μεταβολές, χρησιμοποιούμε ως απεικονιζόμενη ποσότητα, κάποια από τις κινούμενες διατεταγμένες παρατηρήσεις.

Η επόμενη Πρόταση αναφέρεται στην ασυμπτωτική συμπεριφορά της $Y_{1:r:k}$.

Πρόταση 2.4.2 *Αν* $\overline{F} \in MDA(H)$ *με* $c_n > 0, d_n \in \Re$, *τότε*

$$\lim_{n\to\infty} P\left(\frac{\max\{Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}\} - d_{a_n}}{c_{a_n}} \le x\right) = e^{-\lambda(x)}$$

όπου το $\lambda(x)$ *δίδεται από την* (2.4.3).

Απόδειξη. Το ζητούμενο προκύπτει άμεσα, από τη σχέση

$$\lim_{n\to\infty} P\left(\frac{\max\{Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}\} - d_{a_n}}{c_{a_n}} \le x\right)$$
$$= \lim_{n\to\infty} P(W_{n,k,r}(c_{a_n}x + d_{a_n}) < 1) = f_{CP}(x;0) = e^{-\lambda(x)}.$$

■

Αξίζει να σημειώσουμε ότι η μέγιστη παρατήρηση, από τις κινούμενες διατεταγμένες παρατηρήσεις, ανήκει στο ίδιο πεδίο έλξης, μ' αυτό των Y_i. Αυτά που αλλάζουν είναι η παράμετρος θέσης μ και η παράμετρος κλίμακας σ. Για την απόδειξη του παραπάνω ισχυρισμού, αρκεί να παρατηρήσουμε τα εξής:

α. Μέγιστο πεδίο έλξης της Frechet

Αν $H(x) = \Phi_a(x)$, τότε

$$\lambda(x) = \binom{k-1}{r-1}(-\ln \Phi_a(x))^{-r} = \binom{k-1}{r-1}(x)^{-ra} = \left(\frac{x-\mu}{\sigma}\right)^{-a'}, x > 0$$

όπου

$$a' = ar, \quad \mu = 0, \quad \sigma = \binom{k-1}{r-1}^{1/ra}.$$

Επομένως

$$e^{-\lambda(x)} = e^{-\left(\frac{x-\mu}{\sigma}\right)^{-a'}} = \Phi_{a'}\left(\frac{x-\mu}{\sigma}\right).$$

β. Μέγιστο πεδίο έλξης της (Reversed) Weibull

Εάν $H(x) = \Psi_a(x)$ τότε

$$\lambda(x) = \binom{k-1}{r-1}(-\ln \Psi_a(x))^{-r} = \binom{k-1}{r-1}(-x)^{ra} = \left(-\frac{x-\mu}{\sigma}\right)^{a'}, x \le 0$$

με

$$a' = ar, \quad \mu = 0, \quad \sigma = \binom{k-1}{r-1}^{-1/ra}.$$

Έτσι

$$e^{-\lambda(x)} = e^{-\left(\frac{x-\mu}{\sigma}\right)^{-a'}} = \Psi_{a'}\left(\frac{x-\mu}{\sigma}\right).$$

γ. Μέγιστο πεδίο έλξης της Gumbel

Εάν $H(x) = \Lambda(x)$ θα έχουμε

$$\lambda(x) = \binom{k-1}{r-1}(-\ln\Lambda(x))^{-r} = \binom{k-1}{r-1}e^{-rx} = e^{-\frac{x-\mu}{\sigma}}, x \in \Re$$

όπου

$$\mu = \frac{1}{r}\ln\binom{k-1}{r-1}, \sigma = \frac{1}{r}.$$

Συνεπώς

$$e^{-\lambda(x)} = e^{-e^{-\frac{x-\mu}{\sigma}}} = \Lambda\left(\frac{x-\mu}{\sigma}\right).$$

Σημειώνουμε και πάλι ότι, τα αποτελέσματα για την προσέγγιση της κατανομής της τ.μ. $Y_{1:r:k}$, χρησιμοποιώντας την Πρόταση 2.4.2, βελτιώνονται αρκετά, εάν αντικαταστήσουμε το $\lambda(x)$ με το $\lambda^*(x)$ (ειδικά για μικρές τιμές του n).

2.4.3 Αριθμητικές συγκρίσεις

Για να εξακριβώσουμε την ποιότητα των προσεγγίσεων που εισάγαμε, με χρήση του Θεωρήματος 2.4.3 και της Πρότασης 2.4.2, προχωρούμε στη συνέχεια σε γραφική σύγκριση της ακριβούς και προσεγγιστικής κατανομής, για ειδικές κατανομές από κάθε πεδίο έλξης.

α. Κατανομή Pareto. Ας υποθέσουμε ότι οι τ.μ. $Y_1, Y_2, \ldots$ ακολουθούν μια κατανομή Pareto, με συνάρτηση κατανομής

$$F(x) = 1 - \left(\frac{c}{x}\right)^a, x \geq c$$

όπου a και c, θετικοί αριθμοί. Η κατανομή Pareto είναι μία από τις πιο δημοφιλείς κατανομές με βαριά ουρά, και οι εφαρμογές που παρουσιάζει είναι πάρα πολλές, κυρίως στα πεδία

των ασφαλίσεων, των αναλογιστικών και κοινωνικό-οικονομικών μοντέλων κ.α. (βλ. π.χ. Johnson et al (1994) ή το εξαιρετικό βιβλίο του Arnold (1985)).

Θεωρώντας τις σταθερές κανονικοποίησης

$$c_n = F^{-1}(1-n^{-1}) = cn^{1/a}, \quad d_n = 0,$$

έχουμε

$$n\overline{F}(c_n x + d_n) = n\left(\frac{c}{cn^{1/a}x}\right)^a = x^{-a}, \quad x > 0,$$

οπότε,

$$\overline{F} \in MDA(\Phi_a).$$

Επομένως, από το Θεώρημα 2.4.3 προκύπτει

$$\lim_{n\to\infty} P\left(\frac{Y_{m:r:k}(n)}{cn^{1/ra}} \le x\right) = \lim_{n\to\infty} P(W_{n,k,r}(cn^{1/ra}x) < m) = \sum_{i=0}^{m-1} f_{CP}(x;i), x > 0$$

όπου $f_{CP}(x;\cdot)$ είναι η συνάρτηση πιθανότητας της σύνθετης κατανομής Poisson, με παράμετρο

$$\lambda(x) = \binom{k-1}{r-1} x^{-ra}, x > 0$$

και συνθέτουσα κατανομή την (2.4.1). Επιπρόσθετα, θα ισχύει

$$\lim_{n\to\infty} P\left(\frac{\max\{Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}\}}{cn^{1/ra}} \le x\right) = e^{-\lambda(x)}, x > 0$$

ενώ μια καλύτερη προσέγγιση επιτυγχάνουμε, χρησιμοποιώντας το

$$\lambda^*(x) = \binom{k-1}{r-1} x^{-ra}\left(1 - \frac{x^{-a}}{n^{1/r}}\right)^{k-r+1}, x > 0.$$

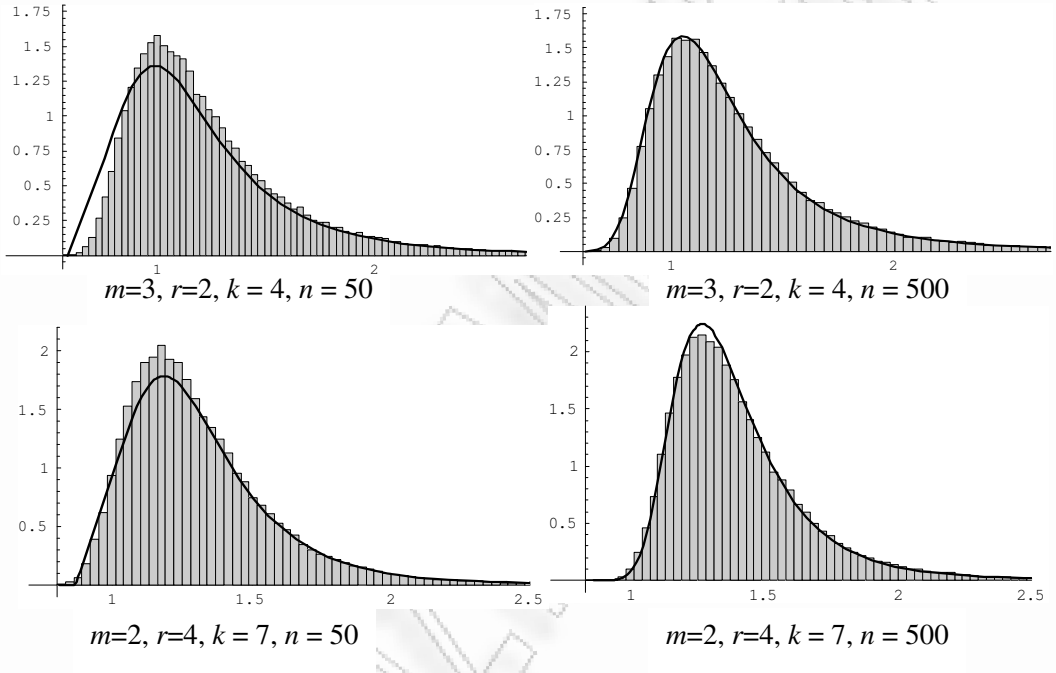
στη θέση του $\lambda(x)$.

Για παράδειγμα, εάν επιθυμούμε να μελετήσουμε την κατανομή της δεύτερης μεγαλύτερης παρατήρησης, ανάμεσα στα $Y_{r:k}^{(i)}$, $i = 1, 2, \ldots, n-k+1$, καταλήγουμε στη σχέση (εφαρμόζουμε το Θεώρημα 2.4.3, για $m = 2$, κάνοντας χρήση της (2.3.2))

$$P\left(Y_{2:r:k}(n) \le x\right) \approx e^{-\lambda^*\left(\frac{x}{cn^{1/ra}}\right)}\left(1 + \frac{r-1}{k-1}\lambda^*\left(\frac{x}{cn^{1/ra}}\right)\right).$$

Στο Σχήμα 2.4.2, παρουσιάζουμε την κατανομή της $Y_{m:r:k}(n)$ (κατάλληλα κανονικοποιημένης) για $n = 50$ και $n = 500$, και 2 επιλογές για τις τιμές των m, k, r. Η καμπύλη που

υπάρχει στα σχήματα, είναι η ασυμπτωτική κατανομή της $Y_{m:r:k}(n)$, ενώ το ιστόγραμμα αφορά τις τιμές της $Y_{m:r:k}(n)/\ (cn^{1/ra})$, από την προσομοίωση 100.000 επαναλήψεων (όπου τα Y_i, $i = 1, 2, \ldots, n$ ακολουθούν μια κατανομή Pareto με $c = 1, a = 2$). Αξιοσημείωτη είναι η πολύ καλή προσέγγιση της κατανομής, ακόμη και για μικρές τιμές του n (π.χ. για $n = 50$).



Σχήμα 2.4.2: Ακριβής (μέσω προσομοίωσης) και προσεγγιστική κατανομή, της $Y_{m:r:k}$, για την περίπτωση της κατανομής Pareto , με $F(x) = 1 - x^{-2}$, $x \geq 1$.

β. Ομοιόμορφη Κατανομή. Ας υποθέσουμε τώρα ότι, οι ανεξάρτητες και ισόνομες τ.μ. $Y_1, Y_2, \ldots$, ακολουθούν μία ομοιόμορφη κατανομή στο διάστημα $(0, 1)$, με

$$F(x) = x, \quad 0 < x < 1.$$

Αφού, $\overline{F} \in MDA(\Psi_a)$ (και πιο συγκεκριμένα, $\overline{F} \in MDA(\Psi_1)$) με $x_F = 1$, μπορούμε να πάρουμε

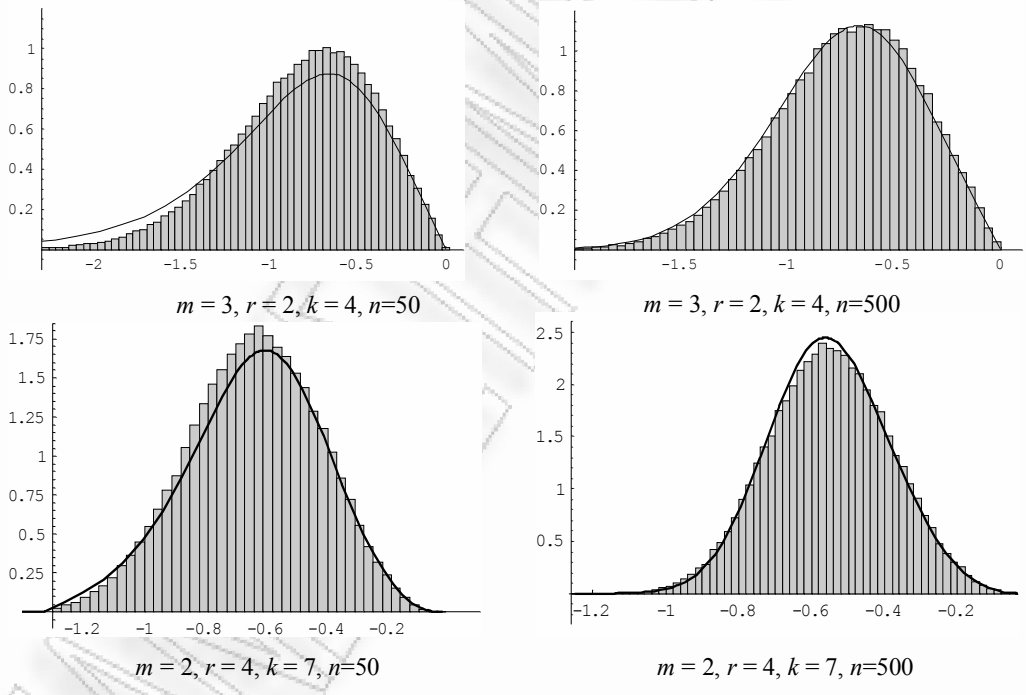
$$c_n = x_F - F^{-1}(1 - n^{-1}) = n^{-1}, \quad d_n = x_F = 1,$$

οπότε καταλήγουμε στην ασυμπτωτική κατανομή

$$\lim_{n\to\infty} P\left(\frac{Y_{m:r:k}(n)-1}{n^{-1/r}} \le x\right) = \lim_{n\to\infty} P(W_{n,k,r}(n^{-1/r}x+1)) < m)$$
$$= \sum_{i=0}^{m-1} f_{CP}(x;i), \;\; x>0$$

όπου η παράμετρος $\lambda(x)$ της σύνθετης κατανομής Poisson, να δίδεται από την (2.4.5).

Στο Σχήμα 2.4.3, συγκρίνεται η ακριβής κατανομή της $Y_{m:r:k}(n)$ (μέσω προσομοίωσης) με την ασυμπτωτική, για 2 διαφορετικές τιμές των παραμέτρων n, m, r, k (όπως έχουμε ήδη αναφέρρει, για καλύτερα προσεγγιστικά αποτελέσματα, χρησιμοποιήθηκε ο τύπος (2.4.6), αντί του (2.4.5)).



Σχήμα 2.4.3: Ακριβής (μέσω προσομοίωσης) και προσεγγιστική κατανομή, της $Y_{m:r:k}$, για την περίπτωση της Ομοιόμορφης κατανομής, $F(x) = x$, $0 < x < 1$.

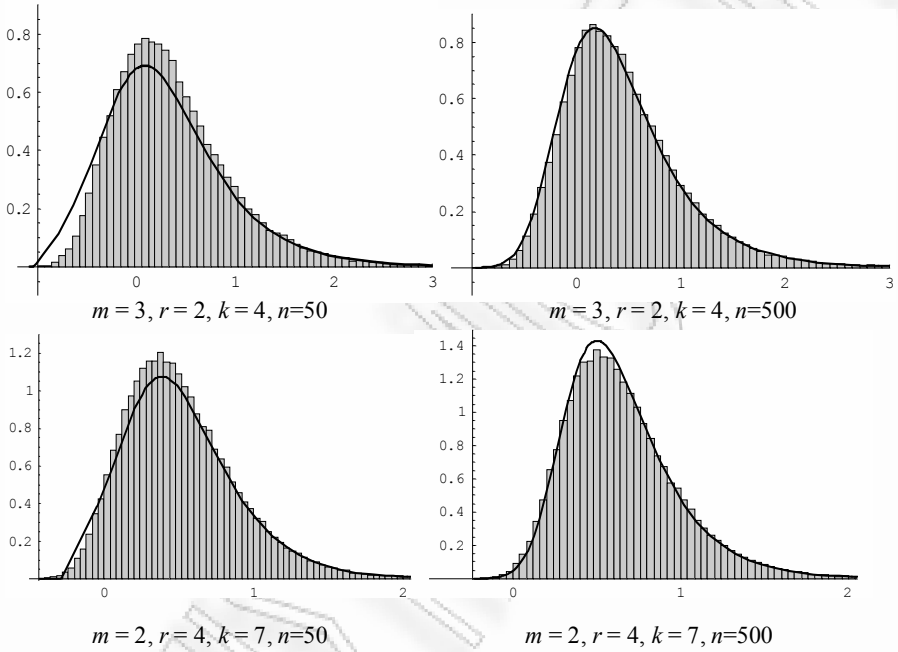
γ. Κανονική και εκθετική κατανομή. Δυο χαρακτηριστικά μέλη από το μέγιστο πεδίο έλξης της κατανομής Gumbel, είναι η Εκθετική κατανομή, με μέση τιμή $1/\beta$ και η τυπική Κανονική κατανομή. Μια επιλογή από σταθερές κανονικοποίησης, είναι η ακόλουθη (βλ. π.χ. τον Πίνακα 3.4.2, στο Embrechts et al (1997))

$$c_n = \beta^{-1}, \quad d_n = \beta^{-1}\ln n,$$

και

$$c_n = (2\ln n)^{-1/2}, \quad d_n = (2\ln n)^{1/2} - \frac{\ln 4\pi + \ln\ln n}{2(2\ln n)^{1/2}}$$

αντιστοίχως (η τελευταία επιλογή είναι ουσιαστικά, μια προσέγγιση των σταθερών κανονικοποίησης). Εφαρμόζοντας το Θεώρημα 2.4.3, συμπεραίνουμε άμεσα ότι η κατανομή της

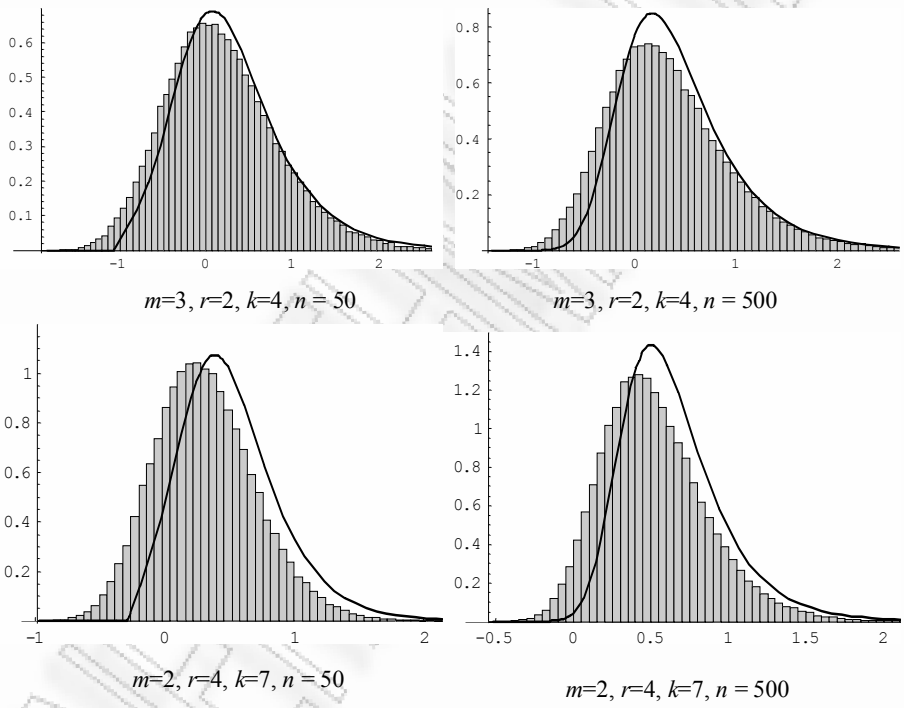


Σχήμα 2.4.4: Ακριβής (μέσω προσομοίωσης) και προσεγγιστική κατανομή, της $Y_{m:r:k}$, για την περίπτωση της Εκθετικής κατανομής, $F(x) = 1 - e^{-x}$, $x \geq 0$.

$Y_{m:r:k}(n)$, προσεγγίζεται από μία σύνθετη κατανομή Poisson, με το $\lambda(x)$ να δίνεται από την (2.4.7) (για ποιοτικότερη προσέγγιση, χρησιμοποιούμε τη σχέση (2.4.8) αντί της (2.4.1)).

Στα Σχήματα 2.4.4 και 2.4.5, δίνεται μια γραφική σύγκριση των προσεγγίσεων με τις ακριβείς κατανομές, για την τ.μ. $Y_{m:r:k}(n)$, στην περίπτωση που οι τ.μ. $Y_1, Y_2, \ldots$ ακολουθούν την Εκθετική και την Κανονική κατανομή, αντίστοιχα. Η χαμηλή ποιότητα της προσέγγισης που διαφαίνεται στο Σχήμα 2.4.5, αποδίδεται στο μικρο ρυθμό σύγκλισης, του μέγιστου από τ.μ. με κανονική κατανομή, στην κατανομή Gumbel (ο ρυθμός σύγκλισης της $n\overline{\Phi}(c_n x + d_n)$ στην e^{-x}, είναι της τάξεως $O((\ln n)^{-1})$).

Τέλος, αξίζει να αναφέρουμε ότι μπορούν να διατυπωθούν άμεσα, παρόμοια αποτελέσματα για την περίπτωση που μας ενδιαφέρουν οι ελάχιστες κινούμενες διατεταγμένες παρατηρήσεις (και όχι οι μέγιστες). Προς τούτο αρκεί να δουλέψουμε με τις μέγιστες παρατηρήσεις των $-Y_i$, στη θέση των Y_i.



Σχήμα 2.4.5: Ακριβής (μέσω προσομοίωσης) και προσεγγιστική κατανομή, της $Y_{m:r:k}$, για την περίπτωση της τυπικής κανονικής κατανομής, $\Phi(x)$, $x \in \Re$.

Κεφάλαιο 3

Δυαδικοί τυχαίοι πίνακες πλήρους κάλυψης

Ας θεωρήσουμε ένα πίνακα δυο διαστάσεων με k γραμμές και n στήλες, του οποίου τα kn στοιχεία προέρχονται από ένα αλφάβητο, με q γράμματα. Εάν επιλέξουμε t γραμμές από τον παραπάνω $k \times n$ πίνακα (όπου t είναι ένας ακέραιος αριθμός, με $1 < t \leq k$), τότε σε κάθε μία στήλη του υποπίνακα $t \times n$ που προκύπτει (από τις t επιλεγμένες γραμμές), σχηματίζεται μια λέξη μήκους t (έτσι παίρνουμε συνολικά n λέξεις μήκους t). Για παράδειγμα, ας θεωρήσουμε τον επόμενο 4×6 πίνακα (δηλαδή, $k = 4$ και $n = 6$), όπου κάθε στοιχείο του προέρχεται από ένα αλφάβητο με 2 γράμματα (τα {0,1})

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

Αν επιλέξουμε δυο γραμμές ($t = 2$) του παραπάνω πίνακα, π.χ. τη δεύτερη και την τέταρτη, τότε προκύπτει ο 2×6 υποπίνακας

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \end{pmatrix}$$

οι στήλες του οποίου, σχηματίζουν τις επόμενες 6 δυαδικές λέξεις

$$\binom{0}{1}, \binom{1}{1}, \binom{1}{0}, \binom{0}{1}, \binom{1}{1} \text{ και } \binom{1}{0}.$$

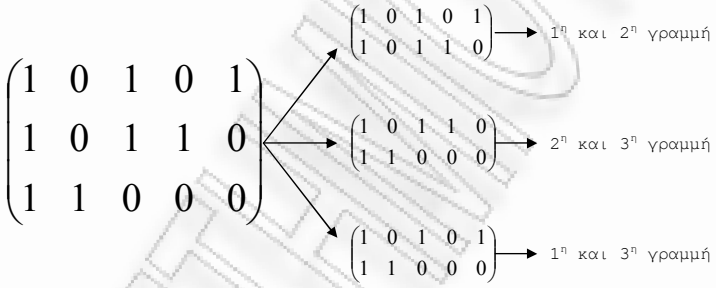
Ένας $k \times n$ πίνακας (με στοιχεία από ένα αλφάβητο με q γράμματα) θα ονομάζεται *πίνακας πλήρους κάλυψης ή πίνακας με πλήρη κάλυψη, μεγέθους* t (*t-covering array*, *t*-CA), εάν κάθε $t \times n$ υποπίνακας του, περιλαμβάνει στις στήλες του και τις q^t δυνατές λέξεις,

μήκους t (όπου, $n \geq q^t$). Ασφαλώς, το πλήθος των υποπινάκων, διαστάσεως $t \times n$, του αρχικού $k \times n$ πίνακα, είναι ίσο με τους συνδυασμούς των k ανά t (δηλαδή, $\binom{k}{t}$).

Ας πάρουμε την περίπτωση που έχουμε ένα πίνακα με $k = 3$ γραμμές, $n = 5$ στήλες (όπως αυτός του Σχήματος 3.0.1), με στοιχεία από ένα αλφάβητο με 2 γράμματα (έστω $\{0, 1\}$). Ο πίνακας αυτός θα είναι πίνακας πλήρους κάλυψης μεγέθους $t = 2$, εάν και μόνο εάν κάθε ένας από τους τρεις (καθώς, $\binom{3}{2} = 3$) 2×5 υποπίνακες, έχει στις στήλες του και τις 4 δυνατές λέξεις μήκους 2 (ενός αλφάβητου με δυο γράμματα), δηλαδή, τις λέξεις

$$\binom{0}{0}, \binom{0}{1}, \binom{1}{0} \text{ και } \binom{1}{1}.$$

Έτσι, για τον πίνακα του Σχήματος 3.0.1, έχουμε ότι ο πρώτος υποπίνακας ο οποίος



Σχήμα 3.0.1: Πίνακας πλήρης κάλυψης μεγέθους $t = 2$, με $k = 3$,$n = 5$ και $q = 2$.

σχηματίζεται από την πρώτη και τη δεύτερη γραμμή, έχει ως στήλες τις λέξεις (μήκους 2)

$$\binom{1}{1}, \binom{0}{0}, \binom{1}{1}, \binom{0}{1} \text{ και } \binom{1}{0}$$

ο δεύτερος υποπίνακας (αποτελείται από τη δεύτερη και τρίτη γραμμή) τις

$$\binom{1}{1}, \binom{0}{1}, \binom{1}{0}, \binom{1}{0} \text{ και } \binom{0}{0}$$

ενώ ο τρίτος υποπίνακας (δημιουργείται από την πρώτη και τρίτη γραμμή) τις

$$\binom{1}{1}, \binom{0}{1}, \binom{1}{0}, \binom{0}{0} \text{ και } \binom{1}{0}.$$

Εύκολα διαπιστώνουμε ότι κάθε 2×5 υποπίνακας, έχει και τις τέσσερις πιθανές λέξεις μήκους 2, που αναφέρθηκαν προηγουμένως. Επομένως, ο παραπάνω πίνακας είναι όντως ένας πίνακας πλήρους κάλυψης, μεγέθους 2 (ένας 2-CA).

Στη βιβλιογραφία συναντάμε διάφορες έννοιες, κυρίως από το χώρο της συνδυαστικής, που ουσιαστικά ο ορισμός τους είναι ισοδύναμος με τον ορισμό των t-CA. Πιο συγκεκριμένα,

στις ονομασίες t-qualitatively independent partitions (βλ. π.χ. Katona (1973), Poljak et al (1983), Poljak and Tuza (1989)), ή (k, t)-universal set (Bierbrauer and Schellwatt (2000)) ή t-surjective array (Gargano et al (1992)), αποδίδονται εναλλακτικοί ορισμοί των t-CA.

Ένα από τα βασικότερα αντικείμενα έρευνας της θεωρίας των t-CA, είναι η εύρεση των βέλτιστων τιμών για τις παραμέτρους n και k. Για να γίνουμε περισσότερο σαφείς, ας υποθέσουμε ότι θέλουμε να κατασκευάσουμε ένα t-CA πίνακα, με k γραμμές, του οποίου κάθε στοιχείο να προέρχεται από ένα αλφάβητο με q γράμματα. Τότε, όσες περισσότερες στήλες πάρουμε, τόσο πιο εύκολα μπορούμε να φέρουμε εις πέρας μια τέτοια διαδικασία. Στην περίπτωση αυτή ως βέλτιστη τιμή της παραμέτρου n, ορίζουμε το ελάχιστο πλήθος στηλών που μπορεί να έχει ένας t-CA πίνακας, για δεδομένα k, t, q. Αντίστροφα, αυξάνοντας το πλήθος των γραμμών και κρατώντας σταθερές τις τιμές των παραμέτρων n, t, q, η κατασκευή ενός t-CA πίνακα, γίνεται ολοένα και πιο «δύσκολη» (έως ότου γίνει αδύνατη). Έτσι, ως βέλτιστη τιμή για το k, ορίζουμε τη μέγιστη τιμή για την οποία υπάρχει ένας t-CA πίνακας (για δεδομένα n, t, q). Η βιβλιογραφία είναι πλούσια σε δημοσιεύσεις επι των παραπάνω θεμάτων και ενδεικτικά αναφέρουμε τα Renyi (1971), Katona (1973), Kleitman and Spencer (1973), Sloane (1993), Godbole et al (1996) και Colbourn (2004).

Απαντήσεις για τα δυο παραπάνω (ισοδύναμα) προβλήματα, έχουν δοθεί μόνο για την περίπτωση $t = 2$ και $q \geq 2$. Αναλυτικά, για την περίπτωση $t = 2$ και $q = 2$, οι Katona (1973) και Kleitman and Spencer (1973) απέδειξαν ανεξάρτητα, ότι η μέγιστη τιμή που μπορεί να πάρει το k (για δεδομένο n), δίδεται από τη σχέση

$$k = \binom{n-1}{\lfloor n/2 \rfloor},$$

(όπου $\lfloor n/2 \rfloor$ είναι το ακέραιο μέρος του $n/2$). Για την περίπτωση $t = 2$ και $q > 2$, οι Gargano et al (1994) απέδειξαν ότι η ελάχιστη τιμή του n (για δεδομένα k, q), έχει την παρακάτω μορφή

$$n = \frac{q}{2} \log k(1 + o(1)).$$

Οι Godbole et al (1996) ασχολήθηκαν με τη γενικότερη περίπτωση, δηλαδή για $t \geq 2$ και $q \geq 2$, και κατάφεραν να δώσουν ένα άνω φράγμα για το ελάχιστο n. Συγκεκριμένα, μέσα από μία «πιθανοθεωρητική» προσέγγιση (διαφορετική από τις έως τότε προσεγγίσεις), έδειξαν ότι η παρακάτω ποσότητα αποτελεί ένα άνω φράγμα, για το ελάχιστο n:

$$\frac{(t-1)\log k}{\log \frac{q^t}{q^t-1}}(1 + o(1)).$$

Γενικότερα, όταν το $t > 2$ έχουμε στη διάθεση μας διάφορα άνω φράγματα για το ελάχιστο n, τα οποία αποδεικνύονται συνήθως μέσα από «κατασκευαστικές» τεχνικές. Δηλαδή, καταφέρνοντας με κάποια διαδικασία να κατασκευάσουμε ένα t-CA πίνακα, για δεδομένα k, n, t, q, είμαστε σίγουροι ότι το ελάχιστο n, είναι μικρότερο (ή ίσο) από το n, που εμείς χρησιμοποιήσαμε (αντίστοιχα, για το μέγιστο k). Οι μέθοδοι και οι τεχνικές που χρησιμοποιούνται, είναι πάρα πολλές και για μια εξαιρετική ανασκόπηση τους, μπορούμε να ανατρέξουμε στην εργασία του Colbourn (2004). Επίσης, μέσω της προηγούμενης εργασίας (μεταξύ άλλων), γίνεται κατανοητό ότι ένα δεύτερο μεγάλο ζήτημα στη θεωρία των t-CA πινάκων, είναι η κατασκευή τους, με αποτελεσματικό τρόπο. Δηλαδή, ποιος είναι ο γρηγορότερος τρόπος, για να κατασκευαστεί ένας t-CA πίνακας, γνωρίζοντας ότι υπάρχει ένας τέτοιος πίνακας, για δεδομένες τιμές των παραμέτρων.

Οι πίνακες πλήρους κάλυψης έχουν στενή σχέση με προβλήματα ποιοτικού ελέγχου και πειραματικών σχεδιασμών (βλ. π.χ. Dalal and Mallows (1998), Roy (2001)). Ειδικότερα, παρουσιάζεται έντονο ενδιαφέρον για διαδικασίες που βρίσκουν άμεσες εφαρμογές στον ποιοτικό έλεγχο «λογισμικού» και «υλικού» (software and hardware testing, βλ. π.χ. Dalal and Mallows (1998), Roy (2001), Colbourn (2004), Hartman (2006)). Πριν εξηγήσουμε τον τρόπο με τον οποίο υπεισέρχονται οι t-CA στον έλεγχο ποιότητας ενός λογισμικού, αξίζει να αναφέρουμε πως το κομμάτι του ελέγχου, αποτελεί ένα από τα σημαντικότερα στάδια στη διαδικασία της παράγωγης νέου λογισμικού, και μια από τις πιο οικονομικά απαιτητικές διεργασίες (Beizer (1990), Hartman (2006)). Μελέτες υποστηρίζουν ότι η φάση αυτή της παραγωγής, απαιτεί από το 1/3 έως το 1/2 του προϋπολογισμού, ενώ ταυτόχρονα, τα προβλήματα που μπορούν να ανακύψουν από ένα ελλειπή έλεγχο, μπορεί να προκαλέσουν μεγάλη, και ίσως ανεπανόρθωτη, οικονομική ζημιά σε μια εταιρεία (Beizer (1990), Carroll (2003a, b), Hartman (2006)).

Για να εντοπίσουμε ένα από τα σημεία τομής ανάμεσα στη θεωρία των t-CA και τον έλεγχο ποιότητας, ας θεωρήσουμε ένα σύστημα αποτελούμενο από k μονάδες (π.χ. ένα δίκτυο από υπολογιστές, τερματικά, εκτυπωτές κτλ), όπου για κάθε μια μονάδα, υπάρχουν διαθέσιμα, και μπορούν να χρησιμοποιηθούν, q διαφορετικά λογισμικά. Ας υποθέσουμε ακόμη, ότι κατασκευάζεται ένα νέο λογισμικό (προϊόν), για ένα συγκεκριμένο είδος μονάδων. Τότε για να ελεγχθεί πόσο ομαλά «επικοινωνεί» ή αλλιώς συνεργάζεται το νέο προϊόν, με όλα τα υπόλοιπα διαφορετικά λογισμικά, θα εφοδιάσουμε μια μονάδα του συστήματος, με το συγκεκριμένο νέο λογισμικό. Είναι φανερό ότι για να εξετάσουμε όλες τις δυνατές περιπτώσεις ή αλλιώς, να εξετάσουμε όλες τις πιθανές αλληλεπιδράσεις ανάμεσα στα διαφορετικά λογισμικά των k μονάδων (ώστε να εξακριβώσουμε εάν υπάρχει κάποιο πρόβλημα,

σε κάποιο συνδυασμό από λογισμικά), θα πρέπει να γίνουν q^k διαφορετικές δοκιμές.

Κλασικό παράδειγμα αποτελεί ένα δίκτυο υπολογιστών, για το οποίο υπάρχει ένα μεγάλο πλήθος από διαφορετικές συνθήκες, κάτω από τις οποίες πρέπει να λειτουργεί ομαλά. Συγκεκριμένα, οι υπολογιστές ενός δικτύου μπορούν να χρησιμοποιούν, ένα από τρία διαφορετικά λειτουργικά συστήματα (π.χ. Windows, Linux, Macintosh), να έχουν τρεις δυνατές επιλογές για τον περιηγητή του διαδικτύου (Browser, όπως Internet Explorer, Netscape, Mozilla), τρία διαφορετικά πρωτόκολλα επικοινωνίας (ethernet, token ring, ARCnet), ή τρεις επιλογές για τα πρωτόκολλα του λειτουργικού συστήματος του δικτύου (APPC/ARRN, TCP/IP, IPX/SPX, βλ. και Σχήμα 3.0.2). Τότε για να ελέγξουμε όλες τις πιθανές αλληλεπιδράσεις, ανάμεσα σ' όλες τις παραπάνω επιλογές, με σκοπό βέβαια την εξασφάλιση της ομαλής λειτουργίας του δικτύου, πρέπει να κάνουμε $3^4 = 81$ ($k = 4, q = 3$) ελέγχους.

Λειτουργικό Σύστημα	Browser	Πρωτόκολλα Επικοινωνίας	Λειτουργικό Σύστημα Δικτύου
Windows	Internet Explorer	Ethernet	APPC/APPN
Linux	Netscape	Token ring	TCP/IP
Macintosh	Mozilla	ARCnet	IPX/SPX

Σχήμα 3.0.2: Διάφορα λογισμικά σ' ένα δίκτυο.

Ως ένα δεύτερο παράδειγμα, αναφέρουμε ένα κύκλωμα με k διακόπτες, για το οποίο θέλουμε να ελέγξουμε εάν υπάρχει κάποιο πρόβλημα στο άνοιγμα ή στο κλείσιμό τους, για κάποιο συγκεκριμένο συνδυασμό. Τότε το πλήθος των δοκιμών που πρέπει να κάνουμε είναι ίσο με 2^k (αφού $q = 2$).

Είναι φανερό (και από τα προηγούμενα παραδείγματα) ότι θα δημιουργείται έντονο πρόβλημα όταν το q και k πάρουν μεγάλες τιμές (όπως συμβαίνει στα περισσότερα προβλήματα), οπότε το σενάριο της εξέτασης όλων των συνδυασμών θα καθίσταται πρακτικά μη εφαρμόσιμο, και από πλευράς κόστους, αλλά και χρόνου (π.χ. όταν το $k = 10$ και $q = 5$, τότε χρειαζόμαστε περίπου 10 εκατομμύρια ελέγχους/δοκιμές). Ένας αποτελεσματικός τρόπος (όπως έχει αποδειχθεί, βλ. π.χ. Dalal and Mallows (1998), Hartman (2006)) να αντιμετωπίσουμε αυτή την κατάσταση, είναι να σχεδιάσουμε μια διαδικασία ελέγχου η όποια θα εξασφαλίζει τον έλεγχο όλων των αλληλεπιδράσεων ανάμεσα σε t (και όχι k) διαφορετικά

λογισμικά (όπου $t \leq k$), με το ελάχιστο δυνατό πλήθος δοκιμών. Το πρόβλημα αυτό ισοδυναμεί με την κατασκευή ενός $k \times n$ πίνακα πλήρους κάλυψης, μεγέθους t, με στοιχεία από ένα αλφάβητο με q γράμματα. Το n στην περίπτωση μας, παίζει το ρόλο του πλήθους των δοκιμών, που πρέπει να πραγματοποιήσουμε.

Μια άλλη κατηγορία πινάκων (άμεσα συνδεδεμένη με τους t-CA) που έχει επίσης να παρουσιάσει έντονο ενδιαφέρον, είναι η *ορθογώνιοι πίνακες πλήρους κάλυψης* (*orthogonal array*, βλ. Cheng (1995), Hedayat et al (1999)). Ένας πίνακας με k γραμμές, n στήλες και στοιχεία από ένα αλφάβητο με q γράμματα, ονομάζεται ορθογώνιος πίνακας πλήρους κάλυψης, μεγέθους t και συχνότητας c, εάν κάθε $t \times n$ υποπίνακάς του, έχει ως στήλες κάθε μία από τις q^t διαφορετικές λέξεις μήκους t, ακριβώς c φορές (όπου $c \in \{1, 2, \ldots\}$ και $n = cq^t$). Οι εφαρμογές των ορθογώνιων πινάκων πλήρους κάλυψης, είναι παρόμοιες μ' αυτές των t-CA (π.χ. στο πεδίο των πειραματικών σχεδιασμών), και εξίσου ενδιαφέρουσα είναι η σύνδεσή τους με κλασικές έννοιες από τη θεωρία της συνδυαστικής, όπως τα λατινικά τετράγωνα (latin squares, βλ. π.χ. Bush (1952), Hedayat et al (1999)).

Στο συγκεκριμένο κεφάλαιο θα εισάγουμε και θα μελετήσουμε μια νέα κλάση πινάκων, οι οποίοι έχουν άμεση σχέση με τους t-CA, τους οποίους θα αποκαλούμε *πίνακες συνεχόμενης πλήρους κάλυψης* (*consecutive covering array*). Ένας $k \times n$ πίνακας, με στοιχεία από ένα αλφάβητο με q γράμματα, θα ονομάζεται συνεχόμενης πλήρους κάλυψης, μεγέθους t (συμβ. t-CCA) εάν κάθε $t \times n$ υποπίνακάς του, αποτελούμενος από t *συνεχόμενες* γραμμές του αρχικού πίνακα, περιέχει στις στήλες του και τις q^t δυνατές λέξεις μήκους t. Το πλήθος των παραπάνω $t \times n$ υποπινάκων, είναι ίσο με $k - t + 1$, όσες δηλαδή και οι συνεχόμενες t-αδες, στην ακολουθία $1, 2, \ldots, k$. Επομένως, η διαφορά ανάμεσα στους t-CA και τους t-CCA, έγκειται στην επιλογή των $t \times n$ υποπινάκων, στους οποίους επιθυμούμε να υπάρχουν όλες οι δυνατές λέξεις, ως στήλες. Είναι φανερό πως, εάν ένας πίνακας είναι t-CA τότε είναι και t-CCA (το αντίστροφο δεν ισχύει).

Για παράδειγμα, ο πίνακας του Σχήματος 3.0.1 θα είναι 2-CCA εάν και μόνο εάν ο 2×5 υποπίνακας, αποτελούμενος από την πρώτη και τη δεύτερη γραμμή, δηλαδή ο

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix}$$

και ο 2×5 υποπίνακας, που σχηματίζεται από τη δεύτερη και την τρίτη γραμμή,

$$\begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

έχουν και τις 4 λέξεις μήκους 2, ως στήλες (κάτι που όπως είδαμε συμβαίνει). Ο υποπίνακας

που αποτελείται από την πρώτη και την τρίτη γραμμή, δε μας απασχολεί, καθώς οι δυο αυτές γραμμές, δεν είναι συνεχόμενες.

Επίσης, θα χρησιμοποιούμε τον όρο *μη πλήρης υποπίνακας* ή *υποπίνακας μη πλήρους κάλυψης*, για να δηλώσουμε τον $t \times n$ υποπίνακα (που αποτελείται από t συνεχόμενες γραμμές) από τον οποίο λείπει μια τουλάχιστον λέξη μήκους t-από τις 2^t διαφορετικές.

Η μετάβαση μας από τους t-CA στους t-CCA, είναι ανάλογη με την αντίστοιχη μετάβαση που πραγματοποιείται σε μια (μονοδιάστατη) ακολουθία δοκιμών Bernouli, όταν αντί να ενδιαφερόμαστε για τις συχνότητες εμφάνισης των δυο αποτελεσμάτων, κοιτάμε για συνεχόμενες ροές από όμοια αποτελέσματα (βλ. π.χ. Balakrishnan and Koutras (2002)). Αυτό έχει γίνει γιατί, πολλά στατιστικά ή πιθανοθεωρητικά προβλήματα ωθούν στην εξέταση των συνεχόμενων αποτελεσμάτων (εμφάνιση σχηματισμών, ροών κτλ), σε μια ακολουθία δοκιμών Bernouli. Παρόμοια, στους t-CCA μπορούμε να θεωρήσουμε ότι κάθε γραμμή του πίνακα είναι το αποτέλεσμα σε διαδοχικές χρονικές στιγμές, από δίτιμα διανύσματα διαστάσεως $1 \times n$. Επομένως, αντί να εξετάζουμε οποιοδήποτε υποπίνακα διαστάσεως $t \times n$, περιοριζόμαστε σε υποπίνακες οι οποίοι προέρχονται από συνεχόμενες γραμμές (ή αλλιώς συνεχόμενα αποτελέσματα), του αρχικού πίνακα. Βέβαια, οι (πολυδιάστατοι) σχηματισμοί που μας ενδιαφέρουν να εντοπίσουμε, είναι αυτοί που έχουν όλες τις δυνατές λέξεις μήκους t, του αντίστοιχου αλφάβητου. Έκδηλη είναι και η ομοιότητα με τα δισδιάστατα συστήματα, που συναντάμε στη θεωρία αξιοπιστίας, και στα οποία ελέγχουμε επίσης την εμφάνιση δισδιάστατων σχηματισμών (βλ. π.χ. Akiba and Yamamoto (2001), Boutsikas and Koutras (2003), Aki and Hirano (2004)).

Όπως θα φανεί καλύτερα και από τις επόμενες παραγράφους, οι πίνακες συνεχόμενης κάλυψης μπορεί να παρουσιάσουν παρόμοιες εφαρμογές με τους t-CA. Επίσης, αξίζει να επισημάνουμε πως οι t-CCA πίνακες είναι απαλλαγμένοι από ερωτήματα που αφορούν την εύρεση των βέλτιστων τιμών για τις παραμέτρους k και n (όπως αυτά που είδαμε στους t-CA), αφού μπορούμε εύκολα να διαπιστώσουμε ότι για οποιεσδήποτε τιμές των k, n, t, q (με $n \geq q^t$ και $t \leq k$), υπάρχει τουλάχιστον ένας t-CCA. Επιπλέον, η κατασκευή ενός τουλάχιστον t-CCA μπορεί να πραγματοποιηθεί με μια πολύ απλή διαδικασία (βλ. Παράγραφο 3.2.1).

Στα πλαίσια της παρούσης διατριβής θα ασχοληθούμε με τη στοχαστική διάσταση των t-CCA, θεωρώντας αρχικά ότι τα στοιχεία του $k \times n$ πίνακα $\mathbf{X} = (X_{ij})$, είναι ανεξάρτητες και ισόνομες (i.i.d.) δοκιμές Bernouli ($q = 2$), με

$$P(X_{ij} = 1) = p, \text{ και } P(X_{ij} = 0) = 1 - p$$

για κάθε $i = 1, 2, \ldots, k$ και $j = 1, 2, \ldots, n$. Στο επόμενο κεφάλαιο θα ασχοληθούμε με τη γενικότερη περίπτωση, όπου $t, q \geq 2$, και τα στοιχεία του πίνακα θα είναι i.i.d. τ.μ. με πεδίο

τιμών το $\{0, 1, \ldots, q-1\}$. Οι εργασίες οι οποίες μελετάνε τις ιδιότητες των στοχαστικών t-CA, προσπαθώντας παράλληλα να δώσουν κάποιες απαντήσεις και στα ερωτήματα που αναφέρθηκαν προηγουμένως, είναι λίγες σε σύγκριση με τις υπόλοιπες (θα αναφέρουμε τις Godbole and Janson (1996), Godbole et al (1996), Carey and Godbole (2008)).

Η τ.μ. που θα παίξει σημαντικό ρόλο στη μελέτη μας είναι η $T_{k,n,t}$, η οποία θα απαριθμεί το πλήθος των υποπινάκων διαστάσεως $t \times n$, από τους οποίους λείπει τουλάχιστον μια λέξη από τις 2^t (δηλαδή, η $T_{k,n,t}$ εκφράζει το πλήθος των υποπινάκων μη πλήρους κάλυψης). Με τις τεχνικές της εμφύτευσης τ.μ. σε Μαρκοβιανή αλυσίδα (Fu and Koutras (1994), Koutras and Alexandrou (1995)) και με την αρωγή στοιχείων από τη θεωρία της συνδυαστικής, θα υπολογίσουμε τη συνάρτηση πιθανότητας της $T_{k,n,t}$ ($T_{k,n,t} \in \{0, 1, \ldots, k-t+1\}$). Θα εξετάσουμε μια εφαρμογή των αποτελεσμάτων μας, στους πειραματικούς σχεδιασμούς, προσπαθώντας να δώσουμε απαντήσεις σε διάφορα πρακτικής σημασίας ερωτήματα.

Αναλυτικά, στην Παράγραφο 3.1 θα αναφερθούμε στις τεχνικές της εμφύτευσης τ.μ. σε Μαρκοβιανή αλυσίδα, οι οποίες θα μας βοηθήσουν στον υπολογισμό της συνάρτησης πιθανότητας της $T_{k,n,t}$. Στην Παράγραφο 3.2.1 εισάγουμε τις έννοιες και τις τυχαίες μεταβλητές, με τις οποίες θα ασχοληθούμε, ενώ στις Παραγράφους 3.2.2 και 3.2.3, θα υπολογίσουμε την πιθανότητα $P(T_{k,n,t} = 0)$. Έπειτα, βάσει της μεθόδου που χρησιμοποιήσαμε για τον υπολογισμό της $P(T_{k,n,t} = 0)$ και τη διαπίστωση ότι η $T_{k,n,t}$ είναι εμφυτεύσιμη τ.μ. διωνυμικού τύπου (βλ. Παράγραφο 3.1), θα προχωρήσουμε (Παράγραφο 3.3) στον προσδιορισμό της συνάρτησης πιθανότητας της $T_{k,n,t}$. Στην επόμενη παράγραφο, μας απασχολεί η περίπτωση που υπάρχει Μαρκοβιανή εξάρτηση ανάμεσα στις γραμμές του πίνακα (τα στοιχεία του πίνακα δεν είναι πλέον ανεξάρτητες τ.μ.), ενώ στην Παράγραφο 3.5 θα ορίσουμε και θα μελετήσουμε τους *ορθογώνιους πίνακες συνεχόμενης πλήρους κάλυψης* (*consecutive orthogonal arrays*), με τρόπο αντίστοιχο μ' αυτόν των t-CCA. Τέλος, θα αναφερθούμε σε πιθανές εφαρμογές και αριθμητικά αποτελέσματα, που αφορούν τους t-CCA (Παράγραφος 3.6).

3.1 Εμφύτευση τυχαίων μεταβλητών σε Μαρκοβιανή αλυσίδα

Οι Fu and Koutras (1994), έβαλαν τα θεμέλια για μία μέθοδο η οποία αρχικά στόχευε στη μελέτη τ.μ. που σχετίζονται με την καταμέτρηση ροών όμοιων συμβόλων, σε μια ακολουθία από διακριτές τ.μ. Η προσέγγιση που πρότειναν ονομάστηκε *μέθοδος εμφύτευσης σε πεπερασμένη Μαρκοβιανή αλυσίδα* (*finite Markov chain embedding technique*, MCET). Η παραπάνω τεχνική έχει τις ρίζες της, στις εργασίες των Fu (1986), Fu and Hu (1987), Chao

and Fu (1989, 1991), στις οποίες μελετώνται συγκεκριμένα προβλήματα, που σχετίζονται με τη θεωρία αξιοπιστίας (όπως ακριβής υπολογισμός της αξιοπιστίας και ασυμπτωτικές ιδιότητες, συγκεκριμένων συστημάτων). Στη συνέχεια, οι Koutras and Alexandrou (1995) εισήγαγαν και μελέτησαν την έννοια των εμφυτεύσιμων τ.μ. σε Μαρκοβιανές αλυσίδες, διωνυμικού τύπου (Markov chain embeddable variable of Binomial type, MVB), προσφέροντας αρκετά χρήσιμα αποτελέσματα, για τη μελέτη μιας ευρείας ομάδας τ.μ. (με συγκεκριμένα χαρακτηριστικά).

Για λόγους πληρότητας της παρούσας διατριβής και για να διευκολύνουμε τον αναγνώστη, θα ξεκινήσουμε με την παρουσίαση των βασικότερων εννοιών, από την MCET. Ο ορισμός μιας τυχαίας μεταβλητής, εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα, δίδεται παρακάτω (Koutras and Alexandrou (1995)).

Ορισμός 3.1.1 *Έστω ένας χώρος καταστάσεων* $\Omega = \{a_1, a_2, \ldots\}$. *Μια θετική ακέραια τ.μ.* T_v, $v = 0, 1, \ldots$, *με σύνολο τιμών*

$$\{0, 1, \ldots, l_v\}$$

όπου $l_v = \max\{m : P(T_v = m) > 0\}$, *θα λέγεται εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα, εάν ισχύουν τα ακόλουθα*

i. *υπάρχει μια Μαρκοβιανή αλυσίδα διακριτού χρόνου* $\{Y_r, r = 0, 1, \ldots\}$, *ορισμένη στο χώρο καταστάσεων* Ω,

ii. *υπάρχει διαμέριση* $\{C_m, m = 0, 1, \ldots\}$ *του χώρου* Ω, *και επιπλέον*

iii. $P(T_v = m) = P(Y_v \in C_m)$, *για κάθε* $m = 0, 1, \ldots, l_v$.

Ο Ορισμός 3.1.1 μας λέει ότι, για να εκμεταλλευτούμε τις ιδιότητες των Μαρκοβιανών αλυσίδων, με σκοπό να υπολογίσουμε τη συνάρτηση πιθανότητας μιας ακέραιας τ.μ. T_v, πρέπει αρχικώς να ορίσουμε κατάλληλα μια αλυσίδα διακριτού χρόνου $\{Y_r, r = 0, 1, \ldots\}$, (σε κάποιον πεπερασμένο χώρο Ω). Για την αλυσίδα αυτή, θα πρέπει να βρούμε μια διαμέριση του χώρου των καταστάσεών της ($\{C_m, m = 0, 1, \ldots\}$), έτσι ώστε να ισχύει

$$P(T_v = m) = P(Y_v \in C_m),$$

για κάθε $m = 0, 1, \ldots, l_v$. Δηλαδή, η πιθανότητα $P(T_v = m)$ θα είναι ίση με την πιθανότητα η Μαρκοβιανή αλυσίδα $\{Y_r, r = 0, 1, \ldots\}$, μετά από v βήματα να βρίσκεται σε μια από τις καταστάσεις του συνόλου C_m.

Το επόμενο θεώρημα είναι άμεση συνέπεια του Ορισμού 3.1.1

Θεώρημα 3.1.1 *Αν η τ.μ. T_v είναι εμφυτεύσιμη στην Μαρκοβιανή αλυσίδα $\{Y_r, r = 0, 1, \ldots\}$, με χώρο καταστάσεων $\Omega = \{a_1, a_2, \ldots\}$, τότε*

$$P(T_v = m) = \boldsymbol{\pi}_1 \left(\prod_{r=1}^{v} \Lambda_r \right) \sum_{i:a_i \in C_m} e_i',$$

όπου

$$\boldsymbol{\pi}_1 = (P(Y_0 = a_1), P(Y_0 = a_2), \ldots),$$

είναι το διάνυσμα αρχικών πιθανοτήτων της αλυσίδας, Λ_r ο πίνακας των πιθανοτήτων μετάβασης (πρώτης τάξης) της αλυσίδας και e_i το μοναδιαίο διάνυσμα (γραμμή), που έχει όλα τα στοιχεία του να είναι ίσα με μηδέν, εκτός της i-οστής συνιστώσας, η οποία είναι ίση με ένα.

Για να κάνουμε τις παραπάνω έννοιες περισσότερο κατανοητές, ας δούμε το επόμενο παράδειγμα, το οποίο αναφέρεται σε ένα πρόβλημα καταμέτρησης επικαλυπτωμενων ροών μήκους k, σε μια ακολουθία από i.i.d. δοκιμές Bernoulli και σχετίζεται άμεσα με το αντικείμενο του Κεφαλαίου 2 (για τη χρήση της MCET σε προβλήματα ροών, εμφάνιση σχηματισμών, συναρτήσεων σάρωσης και άλλα, ο αναγνώστης μπορεί να ανατρέξει στις εργασίες των Fu and Koutras (1994), Koutras and Alexandrou (1995), Koutras (2003) ή στις μονογραφίες Balakrishnan and Koutras (2002) και Fu and Lou (2003)).

Παράδειγμα 3.1 Έστω μία ακολουθία $X_1, X_2, \ldots, X_n$, από ανεξάρτητες και ισόνομες δοκιμές Bernoulli, με πιθανότητα επιτυχίας

$$P(X_i = 1) = p, \quad i = 1, 2, \ldots, n.$$

Στο Κεφάλαιο 2, είδαμε πως η τ.μ. $W_{n,k,r}$ (ορισμένη σε μια ακολουθία από i.i.d. δοκιμές Bernoulli) απαριθμεί το πλήθος των επικαλυπτόμενων παραθύρων μήκους k, με τουλάχιστον r επιτυχίες ($r \leq k \leq n$, βλ. Παράγραφο 2.2). Για παράδειγμα, για τις παρακάτω ακολουθίες

$$11100110111, \quad 00101110111$$

η $W_{11,2,2}$ παίρνει τις τιμές 5 και 4, αντίστοιχα. Να θυμηθούμε πως η $W_{n,k,r}$ είναι μια θετική ακέραια τ.μ. ($W_{n,k,r} \in \{0, 1, \ldots, n - k + 1\}$), και επομένως με βάση τον Ορισμό 3.1.1, αποτελεί εν δυνάμει εμφυτεύσιμη τ.μ. Πράγματι, στην ειδική περίπτωση όπου $r = k$, οι Fu and Koutras (1994) πρότειναν την επόμενη Μαρκοβιανή αλυσίδα, η οποία μπορεί να χρησιμοποιηθεί για τον προσδιορισμό της συνάρτησης πιθανότητας της $W_{n,k,k}$.

Ορίζουμε αρχικά το χώρο καταστάσεων Ω μιας Μαρκοβιανής αλυσίδας $\{Y_r, r = 0, 1, \ldots\}$ ως εξής:

$$\Omega = \{(m,i) : m = 0, 1, \ldots, n-k, i = -1, 0, 1, \ldots, k-1\} \cup \{(n-k+1, -1)\} \backslash \{(0,-1)\}.$$

Επιπλέον, ας θεωρήσουμε ότι για οποιαδήποτε χρονική στιγμή, η μεταβλητή δ, μετράει το πλήθος των συνεχόμενων επιτυχιών, ξεκινώντας από τη χρονική στιγμή που βρισκόμαστε και πηγαίνοντας προς τα πίσω και ας συμβολίσουμε με m το πλήθος των παραθύρων με k επιτυχίες που έχουμε ήδη παρατηρήσει (μέχρι τη χρονική στιγμή που βρισκόμαστε).

Ακόμη, θεωρούμε ότι ισχύει $Y_j = (m, \delta)$ εάν και μόνο εάν, κατά την j-οστή δοκιμή, έχουμε ήδη παρατηρήσει m παράθυρα (μήκους k), με k επιτυχίες και ταυτόχρονα ισχύει $\delta \leq k-1$. Τέλος, θέτουμε $Y_j = (m, -1)$ εάν και μόνο εάν κατά την j-οστή δοκιμή, έχουμε ήδη παρατηρήσει m παράθυρα (μήκους k), με k επιτυχίες και ταυτόχρονα $\delta \geq k$.

Εξετάζοντας τον τρόπο που ορίσαμε την Μαρκοβιανή αλυσίδα, καταλαβαίνουμε ότι μια κατάλληλη διαμέριση του χώρου καταστάσεων, είναι η ακόλουθη

$$\begin{aligned} C_0 &= \{(0,i) : i = 0, 1, \ldots, k-1\} \\ C_m &= \{(m,i) : i = -1, 0, 1, \ldots, k-1\}, \quad m = 1, 2, \ldots, n-k \\ C_{n-k+1} &= \{(n-k+1, -1)\}. \end{aligned}$$

Μπορούμε εύκολα να διαπιστώσουμε ότι

$$\begin{aligned} &P(Y_j = (m, i+1) | Y_{j-1} = (m,i)) = p, && \text{για } 0 \leq m \leq n-k \text{ και } 0 \leq i \leq k-2, \\ &P(Y_j = (m+1, -1) | Y_{j-1} = (m, k-1)) = p, && \text{για } 0 \leq m \leq n-k, \\ &P(Y_j = (m+1, -1) | Y_{j-1} = (m, -1)) = p, && \text{για } 1 \leq m \leq n-k, \end{aligned}$$

ενώ το αρχικό διάνυσμα πιθανοτήτων είναι το $\boldsymbol{\pi}_1 = (1, 0, \ldots, 0)$ (δηλαδή, $P(Y_0 = (0,0)) = 1$). Η διάσταση του πίνακα πιθανοτήτων μετάβασης είναι ίση με

$$\sum_{m=0}^{n-k+1} |C_m| = (n-k+1)(k+1).$$

Έτσι, π.χ. για την περίπτωση $n = 4$ και $k = 2$, ο πίνακας πιθανοτήτων μετάβασης (της

ομογενούς Μαρκοβιανής αλυσίδας) είναι ο ακόλουθος (όπου $q = 1 - p$)

$$\Lambda = \left(\begin{array}{ccc|ccc|ccc} (0,0) & (0,1) & (1,-1) & (1,0) & (1,1) & (2,-1) & (2,0) & (2,1) & (3,-1) \\ \hline q & p & 0 & & & & & & \\ q & 0 & p & & & & & & \\ 0 & 0 & 0 & q & 0 & p & & & \\ & & & q & p & 0 & & & \\ & & & q & 0 & p & & & \\ & & & & & & q & 0 & p \\ & & & & & & q & p & 0 \\ & & & & & & q & 0 & p \\ \hline & & & & & & 0 & 0 & 1 \end{array}\right)$$

■

Αξίζει επίσης να αναφέρουμε την παρακάτω ενδιαφέρουσα ιδιότητα (αποδεικνύεται άμεσα, από τις ιδιότητες των Μαρκοβιανών αλυσίδων), που αφορά τον υπολογισμό των ροπών και της πιθανογεννήτριας συνάρτησης των εμφυτεύσιμων τ.μ.

Θεώρημα 3.1.2 *Αν η τ.μ. T_v είναι εμφυτεύσιμη στην Μαρκοβιανή αλυσίδα $\{Y_r, r = 0, 1, \ldots\}$, με χώρο καταστάσεων $\Omega = \{a_1, a_2, \ldots\}$, τότε οι ροπές και η πιθανογεννήτρια συνάρτηση της T_v, δίδονται (αντίστοιχα) από τις παρακάτω σχέσεις*

$$E(T_v^{(i)}) = \boldsymbol{\pi}_1 \left(\prod_{r=1}^{v} \Lambda_r\right) \boldsymbol{\xi}'(i),$$

$$\varphi_T(z) = \sum_{m=0}^{l_v} P(T_v = m) z^m = \boldsymbol{\pi}_1 \left(\prod_{r=1}^{v} \Lambda_r\right) \boldsymbol{\psi}'(z)$$

όπου

$$\boldsymbol{\xi}(i) = \sum_{m=0}^{l_v} m^i \sum_{j:a_j \in C_m} e'_j, \quad \boldsymbol{\psi}(z) = \sum_{m=0}^{l_v} z^m \sum_{j:a_j \in C_m} e'_j.$$

Οι Koutras and Alexandrou (1995) παρατήρησαν ότι η εμφύτευση τ.μ., σε πολλά από τα προβλήματα που σχετίζονται με την εμφάνιση ροών, δημιουργεί πίνακες πιθανοτήτων μετάβασης, οι οποίοι όταν γραφούν στη μορφή διαμερισμένων πινάκων (blocked matrices) συγκεντρώνουν τους μη μηδενικούς πίνακες, στην κύρια διαγώνιο. Πιο συγκεκριμένα, τα μη μηδενικά στοιχεία των πινάκων αυτών, εμφανίζονται μόνο σε υποπίνακες, οι οποίοι βρίσκονται

γύρω από την κύρια διαγώνιο (όπως στο Παράδειγμα 3.1). Έτσι, με βασικό ερέθισμα την προηγούμενη παρατήρηση, εισήγαγαν και μελέτησαν μια νέα κατηγορία εμφυτεύσιμων τ.μ.

Πριν προχωρήσουμε στον ορισμό της παραπάνω κατηγορίας, είναι χρήσιμο να παρατηρήσουμε ότι για μια εμφυτεύσιμη τ.μ., μπορούμε (χωρίς περιορισμό της γενικότητας) να υποθέσουμε πως κάθε σύνολο της διαμέρισης του χώρου καταστάσεών της, αποτελείται από τον ίδιο αριθμό στοιχείων. Μ' άλλα λόγια, για τη διαμέριση $\{C_m, m = 0, 1, \ldots\}$, του χώρου Ω, θα ισχύει

$$C_m = \{c_{m0}, c_{m1}, \ldots, c_{m,s-1}\}, \text{ για κάθε } m = 0, 1, \ldots,$$

όπου s είναι ο κοινός πληθάριθμος των συνόλων C_m ($s = |C_m|$).

Ορισμός 3.1.2 *Μια θετική ακέραια τ.μ. T_v, $v = 0, 1, \ldots$, λέγεται εμφυτεύσιμη τ.μ. διωνυμικού τύπου (MVB), εάν ισχύουν τα ακόλουθα*

i. είναι εμφυτεύσιμη τ.μ. (Ορισμός 3.1.1), με διαμέριση $C_m = \{c_{m0}, c_{m1}, \ldots, c_{m,s-1}\}$, και

ii. $P(Y_r \in C_{m_1} | Y_{r-1} \in C_{m_2}) = 0$, για κάθε $m_1 \neq m_2, m_2 + 1$ και $r \geq 1$.

Σύμφωνα με τον παραπάνω ορισμό, μια MVB τ.μ. αν βρίσκεται σε μια κατάσταση ενός συνόλου C_{m_2} (της διαμέρισης του χώρου καταστάσεων), την επόμενη χρονική στιγμή μπορεί είτε να μεταβεί σε κάποιο στοιχείο της ίδιας κλάσης (C_{m_2}) είτε να μεταβεί στην αμέσως επόμενη κλάση (C_{m_2+1}). Η ιδιότητα αυτή, δημιουργεί δυο διαφορετικούς τύπους πινάκων πιθανοτήτων μετάβασης, διαστάσεως $s \times s$, με τη μορφή

$$A_r(m) = \left(P(Y_r \in C_m | Y_{r-1} \in C_m)\right), \quad B_r(m) = \left(P(Y_r \in C_{m+1} | Y_{r-1} \in C_m)\right), \quad (3.1.1)$$

όπου ο πίνακας $A_r(m) + B_r(m)$, είναι στοχαστικός. Στην περίπτωση της ομογενούς Μαρκοβιανής αλυσίδας, οι πίνακες $A_r(m), B_r(m)$ δεν εξαρτώνται από τα r, m και έτσι θα γράφουμε

$$A_r(m) = A, \quad B_r(m) = B, \text{ για κάθε } r \text{ και } m.$$

Επιπλέον, ας συμβολίσουμε με $\mathbf{f}_r(m)$, τα $(1 \times s)$ διανύσματα πιθανοτήτων

$$\mathbf{f}_r(m) = \left(P(Y_r = c_{m0}), P(Y_r = c_{m0}), \ldots, P(Y_r = c_{m,s-1})\right),$$

με $0 \leq r \leq v$ και $0 \leq m \leq l_v$. Τότε, για μια MVB τ.μ. ισχύει το παρακάτω, πολύ χρήσιμο θεώρημα.

Θεώρημα 3.1.3 *Η διπλή ακολουθία διανυσμάτων* $\mathbf{f}_r(m)$, *με* $0 \le r \le v, 0 \le m \le l_v$, *ικανοποιεί τις αναδρομικές σχέσεις*

$$\begin{aligned} \mathbf{f}_r(0) &= \mathbf{f}_{r-1}(0)A, \\ \mathbf{f}_r(m) &= \mathbf{f}_{r-1}(m)A + \mathbf{f}_{r-1}(m-1)B, \quad 1 \le m \le l_v, \end{aligned} \tag{3.1.2}$$

για κάθε $r = 1, 2, \ldots, v$. *Οι αρχικές συνθήκες είναι*

$$\mathbf{f}_0(m) = (P(Y_0 = c_{m0}), P(Y_0 = c_{m1}), \ldots, P(Y_0 = c_{m,s-1}))$$

για $0 \le m \le l_v$, *ενώ η συνάρτηση πιθανότητας της* T_v *δίδεται από τη σχέση*

$$P(T_v = m) = \mathbf{f}_v(m)\mathbf{1}', \quad 0 \le m \le l_v$$

όπου $\mathbf{1} = (1, 1, \ldots, 1)$ *είναι το διάνυσμα (γραμμή) του* $\Re^s$, *μ' όλες τις συνιστώσες του ίσες με ένα.*

Με βάση το Θεώρημα 3.1.3, αν διαπιστώσουμε ότι μια εμφυτεύσιμη τυχαία μεταβλητή, είναι ταυτόχρονα και MVB, οι υπολογισμοί (πολλαπλασιασμοί πινάκων) για τον προσδιορισμό της συνάρτησης πιθανότητας, απλοποιούνται αρκετά (καθώς οι πίνακες και τα διανύσματα που πρέπει να πολλαπλασιάσουμε, είναι μικρότερης διαστάσεως).

Παράδειγμα 3.2 Ας επανέλθουμε στο Παράδειγμα 3.1, όπου κοιτώντας τις πιθανότητες μετάβασης, μπορούμε εύκολα να διαπιστώσουμε ότι η τ.μ $W_{n,k,k}$ είναι μια MVB. Οι δυο πίνακες μετάβασης A, B έχουν τη μορφή

$$A = \left(\begin{array}{c|cccc|c} (\cdot,0) & (\cdot,1) & (\cdot,2) & \cdots & (\cdot,k-1) & (\cdot,-1) \\ \hline q & p & 0 & \cdots & 0 & 0 \\ q & 0 & p & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ q & 0 & 0 & \cdots & p & 0 \\ q & 0 & 0 & \cdots & 0 & 0 \\ \hline q & 0 & 0 & \cdots & 0 & 0 \end{array}\right)_{(k+1)\times(k+1)}$$

και

$$B = \left(\begin{array}{c|cccc|c} (\cdot,0) & (\cdot,1) & (\cdot,2) & \cdots & (\cdot,k-1) & (\cdot,-1) \\ \hline 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & p \\ \hline 0 & 0 & 0 & \cdots & 0 & p \end{array}\right)_{(k+1)\times(k+1)}$$

ενώ για τα διανύσματα πιθανοτήτων $\mathbf{f}_r(m)$ ισχύει

$$\mathbf{f}_0(0) = (1, 0, \ldots, 0),$$
$$\mathbf{f}_0(m) = (0, 0, \ldots, 0), \text{ για κάθε } m > 0.$$

■

Για την πιθανογεννήτρια συνάρτηση και τις ροπές των MVB, μπορούν να δοθούν ακόμη πιο «ελκυστικοί» τύποι, σε σχέση μ' αυτούς που περιγράφονται από το Θεώρημα 3.1.2 (βλ. Koutras and Alexandrou (1995)). Η ιδέα των εμφυτεύσιμων τ.μ. διωνυμικού τύπου, επεκτάθηκε τόσο σε εμφυτεύσιμες τ.μ. πολυωνυμικού τύπου (Antzoulakos et al (2003)), όσο και σε τυχαία διανύσματα (Koutras et al (2008)).

3.2 Υπολογισμός της πιθανότητας εμφάνισης πίνακα, συνεχόμενης πλήρους κάλυψης

Στην παρούσα ενότητα θα εισάγουμε αρχικά τις βασικές έννοιες και συμβολισμούς που είναι σχετικές με το πρόβλημα των t-CCA. Θα ορίσουμε την τ.μ. $T_{k,n,t}$ η οποία απαριθμεί τους υποπίνακες μη πλήρους κάλυψης, και θα υπολογίσουμε την πιθανότητα $P(T_{k,n,t} = 0)$, μέσω MCET. Με τον τρόπο αυτό θα εξασφαλίσουμε μια ομαλή μετάβαση, για τη μελέτη της γενικότερης περίπτωσης $P(T_{k,n,t} = m)$, για $m > 0$.

3.2.1 Έννοιες και συμβολισμοί

Ας θεωρήσουμε ένα $k \times n$ τυχαίο πίνακα $\mathbf{X} = (X_{ij})$, του οποίου τα kn στοιχεία είναι ανεξάρτητες και ισόνομες δοκιμές Bernoulli, με

$$P(X_{ij} = 1) = p, \text{ και } P(X_{ij} = 0) = q = 1 - p$$

για $i = 1, 2, \ldots, k$ και $j = 1, 2, \ldots, n$.

Όπως έχουμε ήδη αναφέρει, ο $k \times n$ τυχαίος πίνακας $\mathbf{X}$, θα ονομάζεται συνεχόμενης πλήρους κάλυψης, μεγέθους t εάν κάθε $t \times n$ υποπίνακάς του, αποτελούμενος από t συνεχόμενες γραμμές (του αρχικού πίνακα), περιέχει στις στήλες του και τις 2^t δυνατές λέξεις μήκους t. Το πλήθος των παραπάνω $t \times n$ υποπινάκων, είναι ίσο με $k - t + 1$, όσες δηλαδή και οι συνεχόμενες t-αδες, στην ακολουθία $1, 2, \ldots, k$.

Ας ορίσουμε στη συνέχεια τις δίτιμες τ.μ. $I_i, i = 1, 2, \ldots, k - t + 1$, με τον εξής τρόπο

$$I_i = \begin{cases} 1, & \text{εάν ο υποπίνακας που αποτελείται από τις γραμμές } i, i+1, \ldots, i+t-1, \\ & \text{δεν έχει τουλάχιστον μία από τις } 2^t \text{ λέξεις μήκους } t, \text{ ως στήλη} \\ 0, & \text{διαφορετικά.} \end{cases}$$

Είναι φανερό πως η τυχαία μεταβλητή $T_{k,n,t}$, που ορίζεται ως

$$T_{k,n,t} = \sum_{i=1}^{k-t+1} I_i, \tag{3.2.1}$$

εκφράζει το συνολικό πλήθος των $t \times n$ (συνεχόμενων) υποπινάκων του $\mathbf{X}$, από τους οποίους λείπει τουλάχιστον μία λέξη μήκους t. Οι τιμές της τ.μ. $T_{k,n,t}$ ανήκουν στο σύνολο $\{0, 1, \ldots, k-t+1\}$ και ο πίνακας $\mathbf{X}$ θα είναι t-CCA εάν και μόνο εάν $T_{k,n,t} = 0$. Επομένως, η πιθανότητα ο πίνακας $\mathbf{X}$ να είναι t-CCA, είναι ίση με $P(T_{k,n,t} = 0)$, ενώ αν τα στοιχεία του πίνακα είναι συμμετρικές δοκιμές Bernoulli, δηλαδή

$$P(X_{ij} = 1) = \frac{1}{2}, \text{ για κάθε } i = 1, 2, \ldots, k \text{ και } j = 1, 2, \ldots, n \ ,$$

τότε ο πληθάριθμος της οικογένειας των συνεχόμενων πινάκων πλήρους κάλυψης (συμβ. $C_{k,n,t}$) θα είναι ίσος με

$$C_{k,n,t} = P(T_{k,n,t} = 0)2^{kn}.$$

Παράλληλα, η πιθανότητα $P(T_{k,n,t} = 0)$ είναι μη μηδενική, για οποιεσδήποτε τιμές των παραμέτρων k, n, t, με $t \leq k$ και $n \geq 2^t$. Αν υποθέσουμε πως δουλεύουμε σε περιβάλλον μη τυχαίο (δηλαδή, οι Y_{ij} δεν είναι τ.μ.), τότε μπορεί εύκολα να κατασκευαστεί ένας τουλάχιστον t-CCA πίνακας, με τον εξής τρόπο

- στον $t \times n$ υποπίνακα που αποτελείται από τις πρώτες t γραμμές, θέτουμε ως στήλες, όλες τις δυνατές λέξεις μήκους t, τουλάχιστον μία φορά,
- έπειτα, ως $(t+1)$-οστή γραμμή, επαναλαμβάνουμε την πρώτη γραμμή, ως $(t+2)$-οστή γραμμή τη δεύτερη γραμμή κ.ο.κ.

Από την άλλη, γνωρίζουμε ότι ο πίνακας $\mathbf{X}$ θα ονομάζεται πίνακας πλήρους κάλυψης, μεγέθους t (t-CA), εάν κάθε $t \times n$ υποπίνακας του, έχει στις στήλες του και τις 2^t δυνατές λέξεις, μήκους t (όπου, $n \geq 2^t$). Η διαφορά ανάμεσα στους t-CA και t-CCA είναι ότι στους δεύτερους πίνακες μας ενδιαφέρει κάθε $t \times n$ υποπίνακας, ο οποίος αποτελείται από t συνεχόμενες γραμμές, και όχι οποιοσδήποτε, όπως συμβαίνει στους t-CA.

Εύκολα καταλαβαίνουμε πως, εάν ένας πίνακας είναι t-CA τότε είναι και t-CCA. Επομένως, η πιθανότητα $P(T_{k,n,t} = 0)$ αποτελεί ένα άνω φράγμα για την πιθανότητα ο πίνακας $\mathbf{X}$, να είναι t-CA, και αντίστοιχα, ο πληθάριθμος $C_{k,n,t}$ της οικογένειας των t-CCA, είναι μεγαλύτερος ή ίσος από τον αντίστοιχο πληθάριθμο των t-CA (η εύρεση του οποίου αποτελεί σημαντικό αντικείμενο έρευνας, τις τελευταίες δεκαετίες- βλ. π.χ. Renyi (1971), Sloane (1993), Godbole et al (1996), Colbourn (2004)).

3.2.2 Υπολογισμός της πιθανότητας εμφάνισης πίνακα συνεχόμενης πλήρους κάλυψης, μεγέθους 2

Στην παράγραφο αυτή θα ασχοληθούμε με τον υπολογισμό της πιθανότητας $P(T_{k,n,2} = 0)$, μέσω MCET. Πάνω στη μέθοδο που θα ακολουθήσουμε, θα βασιστεί και ο προσδιορισμός της $P(T_{k,n,t} = 0)$ αλλά και («ολόκληρης») της συνάρτησης πιθανότητας της $T_{k,n,t}$ (Godbole et al (2008a)).

Κρίσιμο σημείο για τη συνέχεια είναι η παρατήρηση ότι, ο σχηματισμός ενός $k \times n$ πίνακα μπορεί να θεωρηθεί ως μία διαδικασία, η οποία εξελίσσεται σταδιακά. Πιο συγκεκριμένα, μπορούμε να κάνουμε την υπόθεση πως η δημιουργία ενός πίνακα ξεκινάει με το πρώτο «κομμάτι» του, διάστασης $(t-1) \times n$, και ολοκληρώνεται σταδιακά ύστερα από $k - t + 1$ βήματα, όπου σε κάθε βήμα γίνεται η προσθήκη μιας γραμμής. Σε κάθε βήμα κρατάμε ως πληροφορία τον τελευταίο $(t-1) \times n$ υποπίνακα, και παράλληλα, πηγαίνουμε από το ένα βήμα στο άλλο, μόνο εάν η ιδιότητα που μας ενδιαφέρει, ικανοποιείται. Με τον τρόπο αυτό σχηματίζεται μία Μαρκοβιανή αλυσίδα, και εστιάζοντας στο $(k - t + 1)$-οστό βήμα της, μπορούμε να υπολογίσουμε την πιθανότητα που μας ενδιαφέρει.

Για να γίνουν όλα τα παραπάνω πιο ξεκάθαρα, ας ασχοληθούμε αρχικά με την πιο απλή περίπτωση όπου $t = 2$. Ας θεωρήσουμε λοιπόν, την Μαρκοβιανή αλυσίδα $\{Y_r, r = 0, 1, \ldots\}$,

με χώρο καταστάσεων Ω, όπου

$$\Omega = \Omega_1 \cup \{x_{abs}\}$$

με

$$\Omega_1 = \{(x_1, x_2, \ldots, x_n) : x_i \in \{0,1\} \text{ για } i = 1, 2, \ldots, n \text{ και } 2 \leq \sum_{i=1}^{n} x_i \leq n-2\}.$$

Το σύνολο Ω_1 περιέχει όλα τα διανύσματα $(x_1, x_2, \ldots, x_n)$ του $\{0,1\}^n$, για τα οποία το πλήθος των μονάδων είναι μεγαλύτερο ή ίσο από δύο, και μικρότερο ή ίσο από $n-2$ (όλα τα υπόλοιπα διανύσματα $(x_1, x_2, \ldots, x_n)$ που δεν περιέχονται στον Ω_1, έχουν τοποθετηθεί στην κατάσταση απορρόφησης x_{abs}). Να σημειώσουμε ότι δεν είναι δυνατόν για ένα διάνυσμα με πλήθος άσσων μικρότερο του δυο ή μεγαλύτερο του $n-2$, να υπάρξει κάποιο άλλο διάνυσμα ώστε αυτά τα δύο μαζί να σχηματίσουν $2 \times n$ πίνακα, με όλες τις λέξεις. Αυτός είναι ο λόγος που έχει τεθεί ο περιορισμός $2 \leq \sum_{i=1}^{n} x_i \leq n-2$, στον ορισμό του χώρου καταστάσεων. Επιπλέον, θεωρούμε ότι τη χρονική στιγμή r, η Μαρκοβιανή αλυσίδα $\{Y_r, r = 0, 1, \ldots\}$, θα βρίσκεται στην κατάσταση $(x_1, x_2, \ldots, x_n)$, δηλαδή,

$$Y_r = (x_1, x_2, \ldots, x_n), \quad r \geq 1,$$

εάν και μόνο εάν η $(r+1)$-οστή γραμμή του πίνακα $\mathbf{X}$ είναι η $(x_1, x_2, \ldots, x_n)$ και ταυτόχρονα, ο υποπίνακας του $\mathbf{X}$ ο οποίος αποτελείται από τις πρώτες $r+1$ γραμμές (τις γραμμές $1, 2, \ldots, r+1$), είναι ένας πίνακας συνεχόμενης πλήρους κάλυψης, μεγέθους 2. Σε οποιαδήποτε άλλη περίπτωση, η Μαρκοβιανή μας αλυσίδα, θα βρίσκεται στην κατάσταση απορρόφησης x_{abs}.

Το πλήθος των καταστάσεων του Ω_1 είναι ίσο με

$$s = 2^n - 2(n+1)$$

οπότε ο πληθάριθμος του χώρου καταστάσεων Ω της προηγούμενης Μαρκοβιανής αλυσίδας, ισούται με $|\Omega| = s + 1$. Οι μη μηδενικές πιθανότητες μετάβασης, μεταξύ των καταστάσεων του Ω_1 δίνονται ως εξής:

$$P(Y_r = (x_1, x_2, \ldots, x_n) | Y_{r-1} = (x'_1, x'_2, \ldots, x'_n)) = p^{\sum_{j=1}^{n} x_j} (1-p)^{n - \sum_{j=1}^{n} x_j} \qquad (3.2.2)$$

εάν και μόνο εάν, ο $2 \times n$ πίνακας

$$\begin{pmatrix} x'_1 & x'_2 & \ldots & x'_n \\ x_1 & x_2 & \ldots & x_n \end{pmatrix}$$

περιέχει (ως στήλες) και τις $2^2 = 4$ δυνατές (δίτιμες) λέξεις μήκους 2. Άρα, η πιθανότητα του ενδεχομένου $T_{k,n,2} = 0$ θα δίδεται από τη σχέση

$$P(T_{k,n,2} = 0) = P(Y_{k-1} \neq x_{abs}) = \boldsymbol{\pi}_1 \Lambda^{k-1} \begin{pmatrix} \mathbf{1} \\ 0 \end{pmatrix}$$

όπου

$$\Lambda = \begin{pmatrix} P & \mathbf{h}' \\ \mathbf{0} & 1 \end{pmatrix}$$

ενώ ο $s \times s$ πίνακας P αναφέρεται στις πιθανότητες μετάβασης, ανάμεσα στις καταστάσεις του Ω_1 (και επομένως οι πιθανότητες που περιλαμβάνει, περιγράφονται από την (3.2.2)), $\mathbf{0} = (0, 0, \ldots, 0)$ είναι το μηδενικό $1 \times s$ διάνυσμα, και

$$\mathbf{h}' = \mathbf{1}' - P\mathbf{1}' = (I - P\mathbf{1}'), \ \ \mathbf{1} = (1, 1, \ldots, 1),$$

με $\boldsymbol{\pi}_1$ να είναι το $(1 \times (s+1))$ διάνυσμα των αρχικών πιθανοτήτων της αλυσίδας. Επομένως, ο $k \times n$ πίνακας, θα είναι t-CCA εάν η Μαρκοβιανή αλυσίδα μετά από $k - t + 1$ βήματα, δε βρίσκεται στην κατάσταση απορρόφησης x_{abs}.

Αξίζει να σημειώσουμε ότι, η παραπάνω διαμέριση του πίνακα πιθανοτήτων μετάβασης Λ, απλοποιεί ως ένα βαθμό τους υπολογισμούς μας, καθώς

$$\Lambda^{k-1} = \begin{pmatrix} P^{k-1} & (I + P + \ldots + P^{k-2})\mathbf{h}' \\ \mathbf{0} & 1 \end{pmatrix}$$

και η πιθανότητα $P(T_{k,n,2} = 0)$ μπορεί να δοθεί τελικά μέσω της σχέσης

$$P(T_{k,n,2} = 0) = \boldsymbol{\pi}_0 P^{k-1} \mathbf{1}', \tag{3.2.3}$$

όπου $\boldsymbol{\pi}_0$ είναι το $1 \times s$ διάνυσμα, με συντεταγμένες τις s πρώτες αρχικές πιθανότητες της Μαρκοβιανής αλυσίδας, δηλαδή

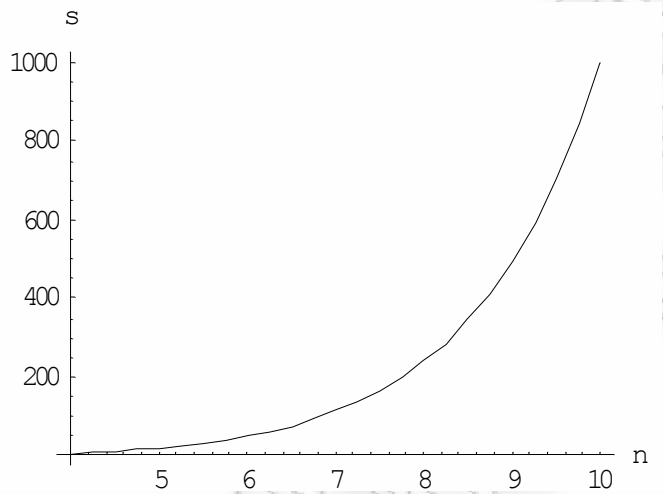
$$\boldsymbol{\pi}_0 = (p_1, p_2, \ldots, p_s).$$

Η πιθανότητα $p_i, i = 1, 2, \ldots, s$, είναι ίση με την πιθανότητα εμφάνισης της i κατάστασης (διανύσματος). Για παράδειγμα η πιθανότητα εμφάνισης της κατάστασης $(1, 1, 0, \ldots, 0)$ είναι ίση με $p^2(1-p)^{n-2}$. Πρέπει επίσης να επισημάνουμε ότι το $\boldsymbol{\pi}_0$ προκύπτει από το $\boldsymbol{\pi}_1$, αφαιρώντας την πιθανότητα που αντιστοιχεί στην κατάσταση απορρόφησης, και επομένως δεν αποτελεί διάνυσμα αρχικών πιθανοτήτων.

Το σημαντικότερο μειονέκτημα της μεθόδου που μόλις περιγράψαμε είναι το μέγεθος του χώρου καταστάσεων Ω. Συγκεκριμένα, είδαμε ότι

$$|\Omega| = s + 1 = 2^n - 2(n+1) + 1,$$

και επομένως, καθώς το n αυξάνεται το πλήθος των καταστάσεων αυξάνεται εκθετικά, λόγω του όρου 2^n (βλ. και Σχήμα 3.2.1).



Σχήμα 3.2.1: Ο πληθάριθμος του Ω_1, για την περίπτωση $t = 2$.

Ευτυχώς, μια κατάλληλη ομαδοποίηση του χώρου καταστάσεων της αλυσίδας, μπορεί να μας οδηγήσει σε ένα χώρο με πολύ μικρότερη διάσταση. Η μείωση αυτή επιτυγχάνεται μελετώντας πιο προσεχτικά τις πιθανότητες μετάβασης και παρατηρώντας ότι όλα τα διανύσματα που έχουν τον ίδιο αριθμό μονάδων (ή ισοδύναμα, μηδενικών), έχουν την ίδια πιθανότητα να βρεθεί ένα άλλο διάνυσμα, σε συνδυασμό με το οποίο θα φτιάχνουν έναν $2 \times n$ πίνακα πλήρους κάλυψης. Επομένως, από ένα διάνυσμα $(x_1, x_2, \ldots, x_n)$ του χώρου Ω_1 , η πληροφορία που αρκεί να κρατήσουμε, είναι το πλήθος των άσσων, και όχι η ακριβής θέση αυτών. Κάτω απ' αυτή τη λογική όλα τα διανύσματα $(x_1, x_2, \ldots, x_n)$ με

$$\sum_{i=1}^{n} x_i = x, \quad x \in \{0, 1, \ldots, n\}$$

(δηλαδή με το ίδιο πλήθος άσσων x), θα αποτελούν μια συγκεκριμένη κατάσταση, η οποία θα συμβολίζεται με x. Λαμβάνοντας υπόψη ότι οι περιπτώσεις με $\sum_{i=1}^{n} x_i < 2$ ή $\sum_{i=1}^{n} x_i > n - 2$, έχουν ενσωματωθεί στην κατάσταση απορρόφησης x_{abs}, ο νέος (μειωμένος) χώρος καταστάσεων γίνεται

$$\Omega = \Omega_2 \cup \{x_{abs}\}$$

όπου

$$\Omega_2 = \{2, 3, \ldots, n - 2\}$$

και $|\Omega| = s + 1$, με $s = n - 3$.

Όπως και στην προηγούμενη προσέγγιση, η Μαρκοβιανή αλυσίδα $\{Y_r, r = 0, 1, \ldots\}$, θα βρίσκεται τη χρονική στιγμή r, στην κατάσταση $x \in \Omega_2$, δηλαδή,

$$Y_r = x, \quad r \geq 1,$$

εάν και μόνο εάν η $(r+1)$-οστή γραμμή του πίνακα $\mathbf{X}$ έχει x άσσους, και ταυτόχρονα, ο υποπίνακας του $\mathbf{X}$ ο οποίος αποτελείται από τις πρώτες $r+1$ γραμμές (τις γραμμές $1, 2, \ldots, r+1$), είναι ένας πίνακας συνεχόμενης πλήρους κάλυψης, μεγέθους 2. Σε οποιαδήποτε άλλη περίπτωση, η Μαρκοβιανή μας αλυσίδα, θα βρίσκεται στην κατάσταση απορρόφησης x_{abs}.

Μέσα από τη θεωρία της συνδυαστικής, δεν είναι δύσκολο να αποδείξουμε ότι οι μη μηδενικές πιθανότητες μετάβασης, ανάμεσα στις καταστάσεις του Ω_2, για την αντίστοιχη Μαρκοβιανή αλυσίδα $\{Y_r, r = 0, 1, \ldots\}$, δίδονται από τον τύπο

$$P(Y_r = x \,|\, Y_{r-1} = x') = p^x (1-p)^{n-x} \sum_{j=\max\{1, x+x'+1-n\}}^{\min\{x-1, x'-1\}} \binom{x'}{j} \binom{n-x'}{x-j},$$

όπου $2 \leq x, x' \leq n-2$. Η τελευταία σχέση προέκυψε με τον εξής τρόπο: είναι πλέον κατανοητό ότι η $P(Y_r = x \,|\, Y_{r-1} = x')$, είναι η πιθανότητα για ένα διάνυσμα με x' άσσους, να βρεθεί ένα άλλο διάνυσμα με x άσσους, τέτοιο ώστε ο $2 \times n$ πίνακας που σχηματίζουν αυτά τα δύο διανύσματα, να είναι πλήρης. Έτσι, το πλήθος των διανυσμάτων για τα οποία έχουμε j άσσους (από τους συνολικά x), ανάμεσα στις συντεταγμένες που υπάρχουν οι x' άσσοι, του αρχικού διανύσματος, είναι

$$\binom{x'}{j} \binom{n-x'}{x-j}, \quad j = 0, 1, \ldots, x.$$

Όμως, για να είναι ο $2 \times n$ πίνακας που προκύπτει πλήρης, θα πρέπει να ισχύει

$$1 \leq j \leq \min\{x-1, x'-1\} \text{ και } 1 \leq x - j \leq n - x' - 1,$$

και λαμβάνοντας υπόψιν ότι έχουμε ανεξάρτητες και ισόνομες δοκιμές Bernoulli, καταλήγουμε στην προαναφερθείσα πιθανότητα μετάβασης.

Οι συντεταγμένες του διανύσματος $\boldsymbol{\pi}_0 = (p_1, p_2, \ldots, p_{n-3})$, θα είναι ίσες με

$$p_x = \binom{n}{x+1} p^{x+1} (1-p)^{n-x-1}, \quad x = 1, 2, \ldots, n-3.$$

Παράδειγμα 3.3 Ας θεωρήσουμε την περίπτωση $n = 6$ και $p = 1/2$ (να σημειώσουμε ότι μέχρι στιγμή έχουμε αναφερθεί μόνο στην περίπτωση που $t = 2$). Ο (μειωμένος) χώρος καταστάσεων Ω είναι

$$\Omega = \Omega_2 \cup \{x_{abs}\} = \{2, 3, 4, x_{abs}\}$$

ενώ οι πιθανότητες μετάβασης ανάμεσα στις καταστάσεις του Ω_2, δίδονται από τους τύπους

$$\begin{aligned}
&P(Y_r = 2|\, Y_{r-1} = 2) = 8p^6, \quad P(Y_r = 3|\, Y_{r-1} = 2) = 12p^6,\\
&P(Y_r = 4|\, Y_{r-1} = 2) = 8p^6, \quad P(Y_r = 2|\, Y_{r-1} = 3) = 9p^6,\\
&P(Y_r = 3|\, Y_{r-1} = 3) = 18p^6, \; P(Y_r = 4|\, Y_{r-1} = 3) = 9p^6,\\
&P(Y_r = 2|\, Y_{r-1} = 4) = 8p^6, \; P(Y_r = 3|\, Y_{r-1} = 4) = 12p^6,\\
&P(Y_r = 3|\, Y_{r-1} = 4) = 8p^6.
\end{aligned}$$

Επομένως, ο πίνακας μετάβασης έχει τη μορφή

$$P = \begin{pmatrix} 8p^6 & 12p^6 & 8p^6 \\ 9p^6 & 18p^6 & 9p^6 \\ 8p^6 & 12p^6 & 8p^6 \end{pmatrix}$$

ενώ για το διάνυσμα $\boldsymbol{\pi}_0$, παίρνουμε

$$\boldsymbol{\pi}_0 = \begin{pmatrix} 15p^6 & 20p^6 & 15p^6 \end{pmatrix}.$$

Εφαρμόζοντας τον τύπο (3.2.3) για διάφορα k, μπορούμε εύκολα να καταλήξουμε στις πιθανότητες που σχετίζονται με το πρόβλημα 2-CCA. Για παράδειγμα, για $k = 5$ και $k = 10$ έχουμε

$$P(T_{5,6,2} = 0) = \boldsymbol{\pi}_0 P^4 \mathbf{1}' = 0.0464, \quad P(T_{10,6,2} = 0) = \boldsymbol{\pi}_0 P^9 \mathbf{1}' = 0.0014$$

και οι αντίστοιχοι πληθάριθμοι $C_{k,6,2}$ της οικογένειας των 2-CCA, είναι

$$\begin{aligned}
C_{5,6,2} &= 2^{5\cdot 6} P(T_{5,6,2} = 0) = 49774080,\\
C_{10,6,2} &= 2^{10\cdot 6} P(T_{10,6,2} = 0) = 160109 \cdot 10^{10}.
\end{aligned}$$

■

3.2.3 Υπολογισμός της πιθανότητας εμφάνισης πίνακα συνεχόμενης πλήρους κάλυψης, μεγέθους $t \geq 2$

Θα προχωρήσουμε στη συνέχεια με τη γενικότερη περίπτωση όπου $t \geq 2$. Ας θέσουμε ότι $a = 2^{t-1}$ και ας συμβολίσουμε με

$$w_1, w_2, \ldots, w_a$$

τις a διαφορετικές λέξεις μήκους $t-1$ (ενός αλφάβητου με γράμματα $\{0,1\}$), διατεταγμένες σε λεξικογραφική σειρά. Για παράδειγμα αν $t = 4$, έχουμε $a = 2^3 = 8$ και

$$w_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, w_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, w_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, w_4 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix},$$

$$w_5 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, w_6 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, w_7 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, w_8 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Με επιχειρήματα, παρόμοια μ' αυτά της περίπτωσης $t = 2$, μπορούμε εύκολα να διαπιστώσουμε ότι για να έχουμε τον πλήρη έλεγχο των καταστάσεων της αλυσίδας, κατά τη μετάβαση από τη χρονική στιγμή (γραμμή) $r-1$ στην r, η μόνη πληροφορία που χρειαζόμαστε είναι ο αριθμός των εμφανίσεων x_i, της λέξης w_i, $i = 1, 2, \ldots, a$, στις τελευταίες $t-1$ γραμμές. Επομένως, ένας κατάλληλος χώρος καταστάσεων, για την περίπτωση $t \geq 2$, θα δίδεται από τον

$$\Omega = \Omega_2 \cup \{x_{abs}\}$$

όπου με Ω_2 συμβολίζουμε το σύνολο των ακέραιων λύσεων της γραμμικής εξίσωσης

$$x_1 + x_2 + \ldots + x_a = n, \text{ με } x_i \geq 2 \text{ και } i = 1, 2, \ldots, a. \tag{3.2.4}$$

Χρησιμοποιούμε τον περιορισμό $x_i \geq 2, i = 1, 2, \ldots, a$ διότι κάθε $(t-1) \times n$ υποπίνακας, ενός t-CCA πίνακα, θα πρέπει να περιέχει ανάμεσα στις στήλες του, τις λέξεις $w_1, w_2, \ldots, w_a$, τουλάχιστον δυο φορές (την κάθε μία). Έτσι, κάθε $(t-1) \times n$ υποπίνακας που δεν ικανοποιεί τον παραπάνω περιορισμό, θεωρείται πως υπάγεται στην κατάσταση απορρόφησης.

Είναι γνωστό από τη θεωρία της συνδυαστικής (βλ. π.χ. Charalambides (2002)), ότι το πλήθος των ακέραιων λύσεων της παρακάτω γραμμικής εξίσωσης,

$$x_1 + x_2 + \ldots + x_\gamma = \beta,$$

(οι λύσεις της εξίσωσης, θα συμβολίζονται με το διάνυσμα $(x_1, x_2, \ldots, x_\gamma)$) με τους περιορισμούς $x_i \geq c_i$ για κάθε $i = 1, 2, \ldots, \gamma$, όπου x_i, γ, β, c_i είναι μη αρνητικοί ακέραιοι αριθμοί, είναι ίσο με

$$\binom{\gamma + \beta - \sum_{i=1}^{\gamma} c_i - 1}{\gamma - 1} = \binom{\gamma + \beta - \sum_{i=1}^{\gamma} c_i - 1}{\beta - \sum_{i=1}^{\gamma} c_i}.$$

Επομένως, το πλήθος των ακέραιων μη αρνητικών λύσεων, της γραμμικής εξίσωσης (3.2.4) είναι

$$s = |\Omega_2| = \binom{a + (n - 2a) - 1}{n - 2a} = \binom{n - a - 1}{a - 1},$$

και έτσι έχουμε

$$|\Omega| = |\Omega_2| + 1 = \binom{n - a - 1}{a - 1} + 1.$$

Ακόμη, θα ισχύει $Y_r = (x_1, x_2, \ldots, x_a) \in \Omega_2$, εάν και μόνο εάν, το πλήθος των εμφανίσεων της λέξης w_i, στον $(t-1) \times n$ υποπίνακα, του οποίου η τελευταία γραμμή είναι η $r + t - 1$, ισούται με x_i (όπου $x_i \geq 2$ για $i = 1, 2, \ldots, a$) και ταυτόχρονα, ο πίνακας που αποτελείται από τις γραμμές $1, 2, \ldots, r+t-1$, είναι t-CCA. Σε οποιαδήποτε άλλη περίπτωση, θα θεωρούμε ότι βρισκόμαστε στην κατάσταση απορρόφησης.

Θα προχωρήσουμε τώρα στον υπολογισμό των πιθανοτήτων μετάβασης

$$P(Y_r = (x_1, x_2, \ldots, x_a) | Y_{r-1} = (x'_1, x'_2, \ldots, x'_a))$$

για $(x_1, x_2, \ldots, x_a), (x'_1, x'_2, \ldots, x'_a) \in \Omega_2$. Το επόμενο λήμμα, μας προσφέρει μια ικανή και αναγκαία συνθήκη, ώστε να είναι μη μηδενική μια πιθανότητα μετάβασης, της προηγούμενης μορφής.

Λήμμα 3.2.1 *Μια ικανή και αναγκαία συνθήκη ώστε να ισχύει*

$$P(Y_r = (x_1, x_2, \ldots, x_a) | Y_{r-1} = (x'_1, x'_2, \ldots, x'_a)) \neq 0$$

είναι η

$$x_{2i-1} + x_{2i} = x'_i + x'_{i+a/2}, \tag{3.2.5}$$

για κάθε $i = 1, 2, \ldots, a/2$.

Απόδειξη. Έχοντας διατάξει τις $a = 2^{t-1}$ διαφορετικές λέξεις μήκους $t-1$ σε λεξικογραφική σειρά $(w_1, w_2, \ldots, w_a)$, μπορούμε εύκολα να διαπιστώσουμε ότι για κάθε $i = 1, 2, \ldots, a/2$,

οι λέξεις w_{2i-1} και w_{2i}, έχουν ακριβώς το ίδιο αρχικό κομμάτι (τμήμα) μήκους $t-2$. Δηλαδή, εάν συμβολίσουμε με $w_i(j)$, $j = 1, 2, \ldots, t-1$ το γράμμα που βρίσκεται στη j θέση της λέξης w_i (j συντεταγμένη), παίρνουμε ότι

$$w_{2i-1}(j) = w_{2i}(j), \quad j = 1, 2, \ldots, t-2$$
$$w_{2i-1}(t-1) = 0, w_{2i}(t-1) = 1.$$

Παρόμοια, για κάθε $i = 1, 2, \ldots, a/2$, το τελευταίο τμήμα μήκους $t-2$ των λέξεων w_i και $w_{i+a/2}$ ταυτίζεται, και επομένως

$$w_i(j) = w_{i+a/2}(j), \quad j = 2, 3, \ldots, t-1$$
$$w_i(1) = 0, w_{i+a/2}(1) = 1.$$

Επιπρόσθετα, το αρχικό τμήμα των λέξεων w_{2i-1}, w_{2i} με το τελικό τμήμα των $w_i, w_{i+a/2}$, είναι ακριβώς το ίδιο.

Σύμφωνα με τον τρόπο ορισμού της Μαρκοβιανής αλυσίδας $\{Y_r, r = 0, 1, \ldots\}$, η πιθανότητα μετάβασης

$$P(Y_r = (x_1, x_2, \ldots, x_a) | Y_{r-1} = (x'_1, x'_2, \ldots, x'_a))$$

αναφέρεται στην περίπτωση στην οποία, από ένα $(t-1) \times n$ πίνακα στον οποίο οι λέξεις $w_i, w_{i+a/2}$, εμφανίζονται $x'_i + x'_{i+a/2}$ φορές, μεταβαίνουμε (αφαιρώντας την πρώτη του γραμμή, και προσθέτοντας μια νέα γραμμή στο τέλος) σ' ένα άλλο $(t-1) \times n$ πίνακα, ο οποίος περιέχει τώρα τις λέξεις w_{2i-1}, w_{2i}, $x_{2i-1} + x_{2i}$ φορές. Όμως, κάτι τέτοιο είναι δυνατόν να συμβεί εάν και μόνο εάν, οι αριθμοί $x'_i + x'_{i+a/2}$ και $x_{2i-1} + x_{2i}$, ταυτίζονται και η απόδειξη του λήμματος, συμπληρώθηκε.

■

Εκμεταλλευόμενοι και το αποτέλεσμα του Λήμματος 3.2.1 είμαστε πλέον σε θέση να αποδείξουμε το κύριο αποτέλεσμα της συγκεκριμένης παραγράφου, που αναφέρεται στον υπολογισμό των (μη μηδενικών) πιθανοτήτων μετάβασης, της αλυσίδας οι οποίες θα χρησιμοποιηθούν στη συνέχεια για τον υπολογισμό της $P(T_{k,n,t} = 0)$.

Θεώρημα 3.2.1 *Εάν* $(x_1, x_2, \ldots, x_a) \in \Omega_2$ *και* $(x'_1, x'_2, \ldots, x'_a) \in \Omega_2$, *με*

$$x_{2i-1} + x_{2i} = x'_i + x'_{i+a/2}, \text{ για κάθε } i = 1, 2, \ldots, a/2.$$

τότε

$$P(Y_r = (x_1, x_2, \ldots, x_a) | Y_{r-1} = (x'_1, x'_2, \ldots, x'_a))$$
$$= (p^{\sum_{i=1}^{a/2} x_{2i}} (1-p)^{n - \sum_{i=1}^{a/2} x_{2i}}) \prod_{i=1}^{a/2} c_i \qquad (3.2.6)$$

όπου

$$c_i = \binom{x'_{i+a/2} + x'_i}{x_{2i}} - \binom{x'_{i+a/2}}{m_i}\binom{x'_i}{x_{2i} - m_i} - \binom{x'_{i+a/2}}{m'_i}\binom{x'_i}{x_{2i} - m'_i}, \tag{3.2.7}$$
$$m_i = \max\{0, x_{2i} - x'_i\}, \quad m'_i = \min\{x_{2i}, x'_{i+a/2}\},$$

για $i = 1, 2, \ldots, a/2$.

Απόδειξη. Αρχικώς, να σημειώσουμε πάλι ότι ο νέος $(t-1) \times n$ πίνακας, του οποίου η δομή περιγράφεται από την κατάσταση $(x_1, x_2, \ldots, x_a)$, θα προκύπτει από τον προηγούμενο πίνακα-τον πίνακα που αναφέρεται στο ενδεχόμενο $Y_{r-1} = (x'_1, x'_2, \ldots, x'_a)$-αφαιρώντας την πρώτη του γραμμή και προσθέτοντας μια νέα στο τέλος. Το πλήθος των άσσων που περιέχει η νέα γραμμή, είναι

$$b = \sum_{i=1}^{a/2} x_{2i}$$

και ταυτόχρονα περιέχει,

$$\sum_{i=1}^{a/2} x_{2i-1} = n - \sum_{i=1}^{a/2} x_{2i} = n - b$$

μηδενικά, γεγονός που δικαιολογεί και την εμφάνιση του όρου $p^b(1-p)^{n-b}$, στην έκφραση (3.2.6).

Ας κοιτάξουμε τώρα στις στήλες του νέου $(t-1) \times n$ πίνακα, οι οποίες έχουν ακριβώς το ίδιο αρχικό τμήμα, μήκους $t-2$. Για συγκεκριμένο $i \in \{1, 2, \ldots, a/2\}$, οι $x_{2i} + x_{2i-1}$ στήλες οι οποίες περιέχουν τις λέξεις w_{2i-1}, w_{2i} θα προκύπτουν από τις $x'_i + x'_{i+a/2}$ στήλες, του προηγούμενου $(t-1) \times n$ πίνακα, οι οποίες θα περιέχουν τις λέξεις $w_i, w_{i+a/2}$. Επομένως, αυτό που χρειαζόμαστε (ώστε να εξασφαλίσουμε τη δημιουργία ενός $t \times n$ πίνακα πλήρους κάλυψης) είναι να κατανείμουμε τα x_{2i-1} μηδενικά και τους x_{2i} άσσους, στη νέα γραμμή, με τέτοιο τρόπο, ώστε να υπάρχει τουλάχιστον ένα μηδενικό και τουλάχιστον ένας άσσος, σε κάθε μία από τις δυο ομάδες στηλών, που αντιστοιχούν στις λέξεις w'_i και $w'_{i+a/2}$.

Το πλήθος των τρόπων με τους οποίους μπορούμε να πετύχουμε κάτι τέτοιο, είναι

$$c_i = \sum_{j=1}^{x_{2i}-1} \binom{x'_{i+a/2}}{j}\binom{x'_i}{x_{2i} - j},$$

θεωρώντας ότι

$$\binom{u}{v} = 0, \quad u \le v.$$

Εάν θέλουμε να απαλλαγούμε από την προηγουμένη παραδοχή, τα άνω και κάτω όρια του αθροίσματος, πρέπει να αλλάξουν με

$$\max\{1, x_{2i} - x'_i + 1\}, \text{ και } \min\{x_{2i} - 1, x'_{i+a/2} - 1\},$$

αντίστοιχα. Έτσι, το c_i θα γίνει

$$c_i = \sum_{j=\max\{1,x_{2i}-x'_i+1\}}^{\min\{x_{2i}-1,x'_{i+a/2}-1\}} \binom{x'_{i+a/2}}{j}\binom{x'_i}{x_{2i}-j}.$$

Επιπλέον, εφαρμόζοντας το γνωστό τύπο του Cauchy (βλ. π.χ. Charalambides (2002))

$$\binom{u_1+u_2}{v} = \sum_{j=\max\{0,v-u_2\}}^{\min\{u_1,v\}} \binom{u_1}{j}\binom{u_2}{v-j},$$

για

$$u_1 = x'_{i+a/2}, u_2 = x'_i, v = x_{2i}$$

καταλήγουμε ότι τα c_i μπορούν να υπολογιστούν μέσω της (3.2.7), αφού

$$\begin{aligned}
\binom{x'_{i+a/2}+x'_i}{x_{2i}} &= \sum_{j=\max\{0,x_{2i}-x'_i\}}^{\min\{x_{2i},x'_{i+a/2}\}} \binom{x'_{i+a/2}}{j}\binom{x'_i}{x_{2i}-j} \\
&= \sum_{j=\max\{1,x_{2i}-x'_i+1\}}^{\min\{x_{2i}-1,x'_{i+a/2}-1\}} \binom{x'_{i+a/2}}{j}\binom{x'_i}{x_{2i}-j} \\
&\quad + \binom{x'_{i+a/2}}{m_i}\binom{x'_i}{x_{2i}-m_i} + \binom{x'_{i+a/2}}{m'_i}\binom{x'_i}{x_{2i}-m'_i} \\
&= c_i + \binom{x'_{i+a/2}}{m_i}\binom{x'_i}{x_{2i}-m_i} + \binom{x'_{i+a/2}}{m'_i}\binom{x'_i}{x_{2i}-m'_i},
\end{aligned}$$

όπου, $m_i = \max\{0, x_{2i} - x'_i\}$, $\quad m'_i = \min\{x_{2i}, x'_{i+a/2}\}$.

Τέλος, είδαμε πως τα c_i δίνουν το πλήθος των τρόπων με τους οποίους μπορούμε από τις $x'_i + x'_{i+a/2}$ λέξεις $w_i, w_{i+a/2}$ (τη χρονική στιγμή $r-1$) να μεταβούμε στις $x_{2i-1} + x_{2i}$ λέξεις w_{2i-1}, w_{2i} (τη χρονική στιγμή r), κάτω από τις προϋποθέσεις που περιγράψαμε προηγουμένως (ώστε να ικανοποιείται το t-CCA κριτήριο). Άρα, η μετάβαση από την κατάσταση $(x'_1, x'_2, \ldots, x'_a)$ στην $(x_1, x_2, \ldots, x_a)$ (ώστε να πάρουμε ένα t-CCA πίνακα), μπορεί να γίνει με $\prod_{i=1}^{a/2} c_i$ διαφορετικούς τρόπους, και αυτό ολοκληρώνει την απόδειξή μας.

■

Έτσι, για να υπολογίσουμε την πιθανότητα ένας $k \times n$ πίνακας να είναι t-CCA, μπορούμε να κάνουμε χρήση της σχέσης

$$P(T_{k,n,t} = 0) = \boldsymbol{\pi}_0 P^{k-t+1} \mathbf{1}', \quad k \geq t \tag{3.2.8}$$

όπου το $\boldsymbol{\pi}_0 = (p_1, p_2, \ldots, p_s)$ είναι το $1 \times s$ διάνυσμα πιθανοτήτων, με τις συντεταγμένες του να αναφέρονται στην εμφάνιση των $(t-1) \times n$ (δυαδικών) πινάκων, από τους οποίους μπορεί να ξεκινήσει η «κατασκευή» ενός $k \times n$ πίνακα. Επομένως, οι $p_1, p_2, \ldots, p_s$ είναι οι πιθανότητες να πάρουμε ένα $(t-1) \times n$ πίνακα με δομή $(x_1, x_2, \ldots, x_a)$, όπου τα $(x_1, x_2, \ldots, x_a)$ αντιστοιχούν στις s ακέραιες λύσεις της εξίσωσης (3.2.4). Οι τελευταίες πιθανότητες ισούνται με

$$\frac{n!}{x_1! x_2! \cdots x_a!} p^{\sum_{i=1}^{a} x_i |w_i|} (1-p)^{n(t-1) - \sum_{i=1}^{a} x_i |w_i|} \tag{3.2.9}$$

όπου με $|w_i|$, $i = 1, 2, \ldots, a$ συμβολίζουμε το πλήθος των άσσων, στη λέξη i. Εναλλακτικά, το άθροισμα $\sum_{i=1}^{a} x_i |w_i|$ μπορεί να θεωρηθεί ως ο συνολικός αριθμός άσσων, στον $(t-1) \times n$ πίνακα, ο οποίος περιέχει x_i φορές, τη λέξη w_i, για $i = 1, 2, \ldots, a$.

Το επόμενο παράδειγμα, θα μας βοηθήσει να κατανοήσουμε καλύτερα, τη μέθοδο που μόλις περιγράψαμε, υπολογίζοντας όλες τις παραπάνω πιθανότητες, για συγκεκριμένες τιμές των παραμέτρων.

Παράδειγμα 3.4 Ας υποθέσουμε ότι ενδιαφερόμαστε για την περίπτωση όπου $n = 10$ και $t = 3$. Τότε $a = 2^{t-1} = 4$ και όλες οι διαφορετικές λέξεις (λεξικογραφικά διατεταγμένες) μήκους $t - 1 = 2$, είναι

$$w_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, w_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, w_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, w_4 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Ο χώρος καταστάσεων Ω είναι της μορφής $\Omega = \Omega_2 \cup \{x_{abs}\}$ όπου ο Ω_2 περιλαμβάνει τις ακέραιες λύσεις, της γραμμικής εξίσωσης

$$x_1 + x_2 + x_3 + x_4 = 10,$$

κάτω από τους περιορισμούς $x_i \geq 2$, για $i = 1, 2, 3, 4$. Ο πληθάριθμος του Ω_2 είναι ίσος με

$$s = |\Omega_2| = \binom{n-a-1}{a-1} = \binom{10-4-1}{4-1} = 10$$

και συμβολίζοντας με ω_i, $i = 1, 2, \ldots, 10$ τα στοιχεία του, παίρνουμε

$$\omega_1 = (4,2,2,2), \omega_2 = (2,2,4,2), \omega_3 = (2,4,2,2), \omega_4 = (2,2,2,4), \omega_5 = (3,2,3,2),$$
$$\omega_6 = (3,3,2,2), \omega_7 = (3,2,2,3), \omega_8 = (2,2,3,3), \omega_9 = (2,3,2,3), \omega_{10} = (2,3,3,2).$$

Με βάση το Λήμμα 3.2.1 έχουμε ότι οι πιθανότητες μετάβασης

$$P(Y_r = (x_1, x_2, \ldots, x_a)|Y_{r-1} = (x'_1, x'_2, \ldots, x'_a))$$

δε μηδενίζονται, εάν και μόνο εάν

$$(x_1, x_2, \ldots, x_a) \in C \text{ και } (x'_1, x'_2, \ldots, x'_a) \in D$$

όπου

α. $C = \{\omega_1, \omega_3, \omega_6\}$ και $D = \{\omega_1, \omega_2, \omega_5\}$, ή

β. $C = \{\omega_2, \omega_4, \omega_8\}$ και $D = \{\omega_3, \omega_4, \omega_9\}$, ή

γ. $C = \{\omega_5, \omega_7, \omega_9, \omega_{10}\}$ και $D = \{\omega_6, \omega_7, \omega_8, \omega_{10}\}$.

Για παράδειγα, αν θεωρήσουμε την κατάσταση $\omega_1 = (x_1, x_2, x_3, x_4) = (4, 2, 2, 2)$ και την $\omega_5 = (x'_1, x'_2, x'_3, x'_4) = (3, 2, 3, 2)$, θα έχουμε

$$x_{2i-1} + x_{2i} = \begin{cases} x_1 + x_2 = 4 + 2 = 6, & i = 1 \\ x_3 + x_4 = 2 + 2 = 4, & i = 2 \end{cases}$$

και

$$x'_i + x'_{i+a/2} = \begin{cases} x'_1 + x'_3 = 3 + 3 = 6, & i = 1 \\ x'_2 + x'_4 = 2 + 2 = 4, & i = 2 \end{cases}$$

το οποίο εξασφαλίζει ότι ικανοποιείται η συνθήκη (3.2.5), του Λήμματος 3.2.1. Έτσι για να υπολογίσουμε την αντίστοιχη πιθανότητα μετάβασης, αρκεί να υπολογίσουμε τις ποσότητες (3.2.7). Επειδή,

$$m_i = \begin{cases} \max\{0, 2-3\} = 0, & i = 1 \\ \max\{0, 2-2\} = 0, & i = 2 \end{cases}, m'_i = \begin{cases} \min\{2, 3\} = 2, & i = 1 \\ \min\{2, 2\} = 2, & i = 2 \end{cases},$$

παίρνουμε

$$c_1 = \binom{3+3}{2} - \binom{3}{0}\binom{3}{2-0} - \binom{3}{2}\binom{3}{2-2} = 9$$
$$c_2 = \binom{2+2}{2} - \binom{2}{0}\binom{2}{2-0} - \binom{2}{2}\binom{2}{2-2} = 4$$

και επομένως, ο τύπος (3.2.6) δίνει

$$P(Y_r = (4, 2, 2, 2)|Y_{r-1} = (3, 3, 2, 2)) = 36p^4(1-p)^6.$$

Δουλεύοντας παρόμοια, για τους υπόλοιπους συνδυασμούς καταστάσεων, καταλήγουμε στον επόμενο πίνακα πιθανοτήτων μετάβασης

$$P = \begin{pmatrix}
32b_1 & 0 & 32b_3 & 0 & 0 & 48b_2 & 0 & 0 & 0 & 0 \\
32b_1 & 0 & 32b_3 & 0 & 0 & 48b_2 & 0 & 0 & 0 & 0 \\
0 & 32b_1 & 0 & 32b_3 & 0 & 0 & 0 & 48b_2 & 0 & 0 \\
0 & 32b_1 & 0 & 32b_3 & 0 & 0 & 0 & 48b_2 & 0 & 0 \\
36b_1 & 0 & 36b_3 & 0 & 0 & 72b_2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 36b_1 & 0 & 36b_2 & 0 & 36b_3 & 36b_2 \\
0 & 0 & 0 & 0 & 36b_1 & 0 & 36b_2 & 0 & 36b_3 & 36b_2 \\
0 & 0 & 0 & 0 & 36b_1 & 0 & 36b_2 & 0 & 36b_3 & 36b_2 \\
0 & 36b_1 & 0 & 36b_3 & 0 & 0 & 0 & 72b_2 & 0 & 0 \\
0 & 0 & 0 & 0 & 36b_1 & 0 & 36b_2 & 0 & 36b_3 & 36b_2
\end{pmatrix}$$

όπου $b_i = p^{3+i}(1-p)^{7-i}, i = 1, 2, 3$. Το αρχικό διάνυσμα $\boldsymbol{\pi}_0$ το οποίο εμφανίζεται στον τύπο (3.2.8) μπορεί εύκολα να υπολογιστεί από την (3.2.9), καθώς $\boldsymbol{\pi}_0 = (p_1, p_2, \ldots, p_{10})$ όπου

$$p_1 = \alpha p^8(1-p)^{12}, \quad p_2 = p_3 = \alpha p^{10}(1-p)^{10}, \quad p_4 = \alpha p^{12}(1-p)^8,$$
$$p_5 = p_6 = \beta p^9(1-p)^{11}, \quad p_7 = p_{10} = \beta p^{10}(1-p)^1, \quad p_8 = p_9 = \beta p^{11}(1-p)^9,$$

με

$$\alpha = \frac{10!}{4!(2!)^3} = 18900, \quad \beta = \frac{10!}{(2!)^2(3!)^2} = 25200.$$

Τέλος, ο υπολογισμός των πιθανοτήτων που μας ενδιαφέρουν (για την περίπτωση $n = 10$ και $t = 3$), μπορεί να ολοκληρωθεί, με βάση τη σχέση

$$P(T_{k,10,3} = 0) = \boldsymbol{\pi}_0 P^{k-2}\mathbf{1}', \quad k \geq 3.$$

Αξίζει να επισημάνουμε ότι ο τελευταίος τύπος μπορεί ακόμη να χρησιμοποιηθεί ώστε να βρούμε μια αναδρομική σχέση για τις πιθανότητες $a_k = P(T_{k,10,3} = 0)$, $k = 3, 4, \ldots$ Για να καταφέρουμε κάτι τέτοιο, αρκεί να εκφράσουμε τη γεννήτρια συνάρτηση

$$\sum_{k=3}^{\infty} a_k z^k = \sum_{k=3}^{\infty} \boldsymbol{\pi}_0 z^2 (zP)^{k-2}\mathbf{1}' = z^3 \boldsymbol{\pi}_0 \left(I - zP\right)^{-1} P\mathbf{1}' = z^2(\boldsymbol{\pi}_0 \left(I - zP\right)^{-1}\mathbf{1}' - \boldsymbol{\pi}_0\mathbf{1}')$$

ως λόγο δυο πολυωνύμων του z, δηλαδή

$$\sum_{k=3}^{\infty} a_k z^k = P_1(z)/P_2(z),$$

και να γράψουμε την τελευταία ταυτότητα στη μορφή $P_2(z)\left(\sum_{k=3}^{\infty} a_k z^k\right) = P_1(z)$, ώστε να συγκρίνουμε του συντελεστές των δυναμοσειρών που προκύπτουν, στο αριστερό και στο δεξιό μέλος.

Στην ειδική περίπτωση των συμμετρικών δοκιμών Bernoulli ($p = 1/2$), ο P γίνεται

$$P = \frac{1}{2^{10}} \begin{pmatrix} 32 & 0 & 32 & 0 & 0 & 48 & 0 & 0 & 0 & 0 \\ 32 & 0 & 32 & 0 & 0 & 48 & 0 & 0 & 0 & 0 \\ 0 & 32 & 0 & 32 & 0 & 0 & 0 & 48 & 0 & 0 \\ 0 & 32 & 0 & 32 & 0 & 0 & 0 & 48 & 0 & 0 \\ 36 & 0 & 36 & 0 & 0 & 72 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 36 & 0 & 36 & 0 & 36 & 36 \\ 0 & 0 & 0 & 0 & 36 & 0 & 36 & 0 & 36 & 36 \\ 0 & 0 & 0 & 0 & 36 & 0 & 36 & 0 & 36 & 36 \\ 0 & 36 & 0 & 36 & 0 & 0 & 0 & 72 & 0 & 0 \\ 0 & 0 & 0 & 0 & 36 & 0 & 36 & 0 & 36 & 36 \end{pmatrix} \tag{3.2.10}$$

ενώ το $\boldsymbol{\pi}_0$, γράφεται στη μορφή

$$\boldsymbol{\pi}_0 = \frac{1}{2^{20}}(\alpha, \alpha, \alpha, \alpha, \beta, \beta, \beta, \beta, \beta, \beta). \tag{3.2.11}$$

Οι πιθανότητες $P(T_{k,10,3} = 0)$ για $k = 3, 4, 5, 6$, είναι ίσες με

$$P(T_{3,10,3} = 0) = \frac{3.024 \cdot 10^7}{2^{30}} = 0.02816,$$
$$P(T_{4,10,3} = 0) = \frac{4.08361 \cdot 10^9}{2^{40}} = 0.00371,$$
$$P(T_{5,10,3} = 0) = \frac{5.53977 \cdot 10^{11}}{2^{50}} = 0.00049.$$

Επιπλέον, πραγματοποιώντας τους απαραίτητους υπολογισμούς, μπορούμε εύκολα να καταλήξουμε στη σχέση

$$\sum_{k=3}^{\infty} a_k z^k = \frac{P_1(z)}{P_2(z)} = \frac{-4725(-409600z^3 + 384z^4 + 243z^5)}{65536(1048576 - 139264z - 576z^2 + 81z^3)}$$

απ' όπου προκύπτει ο ακόλουθος αναδρομικός τύπος

$$a_k = \frac{1}{1048576}(139264a_{k-1} + 576a_{k-2} - 81a_{k-3}), k \geq 6$$

ο οποίος χρησιμοποιείται για τον υπολογισμό των πιθανοτήτων $P(W_{k,10,3} = 0)$, για $k = 6, 7, \ldots$

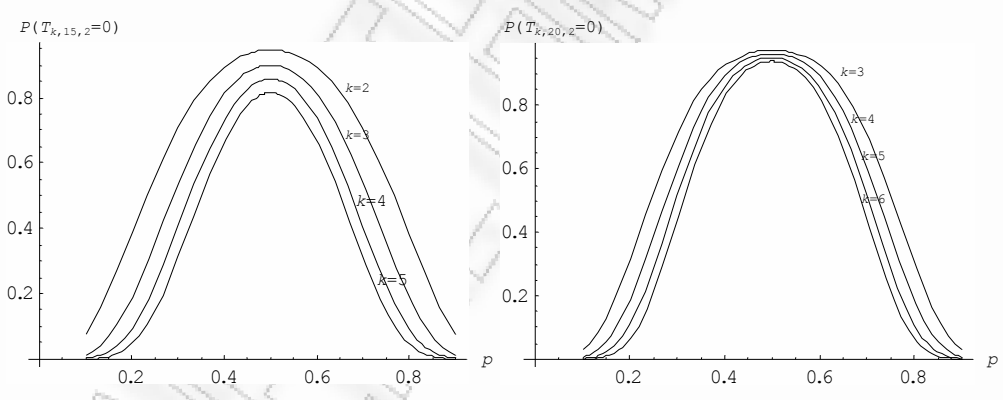
Δυο άμεσα συμπεράσματα, απ' όσα προηγήθηκαν, είναι τα ακόλουθα:

α. Η πιθανότητα $P(T_{k,10,3} = 0)$ μειώνεται καθώς το k αυξάνεται.

β. Οι αριθμητές των προηγούμενων πιθανοτήτων $P(T_{k,10,3} = 0)$, είναι απλά οι πληθάριθμοι $C_{k,10,3}$, για $n = 10$, $t = 3$ και $k = 3, 4, 5$.

■

Από το σύνολο της ανάλυσης που προηγήθηκε, μπορούμε εύκολα να διαπιστώσουμε ότι οι λέξεις μήκους t, θα έχουν την ίδια πιθανότητα εμφάνισης, εάν και μόνο εάν $p = 0.5$. Επομένως, είναι ορθό να αναμένουμε ότι η πιθανότητα $P(W_{k,n,t} = 0)$ θα παίρνει τη μέγιστη τιμή της, όταν ισχύει $p = 0.5$ (για δεδομένες τις τιμές των παραμέτρων k, n, t). Το συμπερασμά αυτό επιβεβαιώνεται και από τα γραφήματα του Σχήματος 3.2.2, για $t = 2$ και διάφορες επιλογές, για τις τιμές των παραμέτρων n, k.



Σχήμα 3.2.2: Η πιθανότητα $P(T_{k,n,2} = 0)$, συναρτήσει του p.

Τέλος, αξίζει να αναφέρουμε ότι για οποιουσδήποτε ακέραιους θετικούς αριθμούς n, k και t, με $t \leq k$ και $n \geq 2^t$, ισχύει

$$\lim_{k\to\infty} P(T_{k,n,t} = 0) = 0,$$

δηλαδή, η πιθανότητα η Μαρκοβιανή αλυσίδα (που χρησιμοποιήσαμε για την εμφύτευση της τ.μ.), να μη βρίσκεται στην κατάσταση απορρόφησης, μετά από άπειρα βήματα, τείνει στο μηδέν. Το παραπάνω αποτέλεσμα, αποδεικνύεται πολύ εύκολα μέσα από τις ιδιότητες των Μαρκοβιανών αλυσίδων, οι οποίες έχουν τουλάχιστον μία κατάσταση απορρόφησης.

3.3 Η κατανομή του πλήθους των υποπινάκων μη πλήρους κάλυψης, ενός τυχαίου δυαδικού πίνακα

Στην παράγραφο αυτή θα μας απασχολήσει ο προσδιορισμός της συνάρτησης πιθανότητας της $T_{k,n,t}$, όπως αυτή έχει οριστεί από την (3.2.1). Η μέθοδος που θα χρησιμοποιήσουμε βασίζεται σ' αυτήν που περιγράψαμε στην προηγούμενη παράγραφο, και στη πολύ σημαντική διαπίστωση, ότι η $T_{k,n,t}$ είναι εμφυτεύσιμη τ.μ., διωνυμικού τύπου (βλ. Ορισμό 3.1.2).

Για να κατανοήσουμε καλύτερα τι ακριβώς δηλώνει η $T_{k,n,t}$, ας θεωρήσουμε την παρακάτω πραγματοποίηση ενός $k \times n$ τυχαίου πίνακα, με $k = 6, n = 10$

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

Τότε, η τ.μ. $T_{6,10,2}$ παίρνει την τιμή 2, αφού

α. στον υποπίνακα διάστασης 2×10, που σχηματίζεται από τις γραμμές $\{1, 2\}$, δεν υπάρχει η λέξη $\binom{1}{1}$,

β. στον υποπίνακα που σχηματίζεται από τις γραμμές $\{5, 6\}$, δεν υπάρχουν οι λέξεις $\binom{0}{1}, \binom{1}{0}$,

ενώ, οι υποπίνακες με γραμμές $\{2, 3\}$ και $\{3, 4\}$, είναι πλήρους κάλυψης. Παρόμοια, έχουμε ότι $T_{6,10,3} = 3$ καθώς οι $t \times n = 3 \times 10$ υποπίνακες, που σχηματίζονται από τις γραμμές $\{1, 2, 3\}$, $\{3, 4, 5\}$ και $\{4, 5, 6\}$, είναι μη πλήρους κάλυψης.

Αποσκοπώντας στον υπολογισμό της συνάρτησης πιθανότητας της $T_{k,n,t}$, θα επεκτείνουμε τον χώρο καταστάσεων Ω_2 (βλ. (3.2.4)) έτσι ώστε να κρατάμε όχι μόνο την πληροφορία που αφορά τη δομή του τελευταίου $(t-1) \times n$ υποπίνακα, τη χρονική στιγμή r, άλλα και το πλήθος των υποπινάκων μη πλήρους κάλυψης, που θα έχουμε παρατηρήσει μέχρι τη στιγμή εκείνη. Τότε, διαπιστώνουμε πως η $T_{k,n,t}$ είναι μια εμφυτεύσιμη τ.μ. διωνυμικού τύπου, και τα εργαλεία που έχουμε στη διάθεση μας (βλ. Ορισμό 3.1.2 και Θεώρημα 3.1.3) μας οδηγούν άμεσα στην κατανομή της $T_{k,n,t}$.

Για την εμφύτευση της απαριθμήτριας τ.μ. $T_{k,n,t}$ σε μια Μαρκοβιανή αλυσίδα, θα χρησιμοποιήσουμε αρχικά εκτός από τις s καταστάσεις του Ω_2, που εισάγαμε στην προηγούμενη

παράγραφο, μία επιπλέον κατάσταση ω_{s+1}. Η νέα αυτή κατάσταση θα δηλώνει ότι στον $(t-1)\times n$ πίνακα που μελετάμε (μια δεδομένη χρονική στιγμή), υπάρχει μια λέξη μήκους $t-1$, η οποία έχει εμφανιστεί λιγότερες από 2 φορές. Χρησιμοποιώντας επιπλέον και μια απαριθμήτρια μεταβλητή, έστω m, η οποία θα κρατάει ως πληροφορία τον αριθμό των υποπινάκων μη πλήρους κάλυψης, καταλήγουμε στο νέο χώρο καταστάσεων

$$\Omega^* = (\Omega_2 \cup \{\omega_{s+1}\}) \times \{0, 1, \ldots, k-t+1\}$$

(το Ω^* είναι το καρτεσιανό γινόμενο των συνόλων $\Omega_2 \cup \{\omega_{s+1}\}$ και $\{0, 1, \ldots, k-t+1\}$) με πληθάριθμο

$$|\Omega^*| = (s+1)(k-t+2).$$

Ο ορισμός τη αλυσίδας $\{Y_r, r = 0, 1, \ldots\}$, με χώρο καταστάσεων Ω^*, ολοκληρώνεται με τις παρακάτω συνθήκες:

i. θα ισχύει $Y_r = (\omega, m)$, με $\omega = (x_1, x_2, \ldots, x_a) \in \Omega_2$ και $0 \le m \le k-t+1$, εάν και μόνο εάν το πλήθος των εμφανίσεων της λέξης w_i, στον $(t-1)\times n$ υποπίνακα, του οποίου η τελευταία γραμμή είναι η $r+t-1$, ισούται με x_i (όπου $x_i \ge 2$ για $i = 1, 2, \ldots, a$) και επιπλέον, στον πίνακα που αποτελείται από τις γραμμές $1, 2, \ldots, r+t-1$, υπάρχουν ακριβώς m υποπίνακες μη πλήρους κάλυψης,

ii. θα ισχύει $Y_r = (\omega, m)$ με $\omega = \omega_{s+1}$ και $0 \le m \le k-t+1$, εάν και μόνο εάν τουλάχιστον μία λέξη από τις $w_1, w_2, \ldots, w_a$ $(a = 2^{t-1})$ έχει εμφανιστεί λιγότερες από 2 φορές, στον $(t-1)\times n$ υποπίνακα, του οποίου η τελευταία γραμμή είναι η $r+t-1$ και επιπλέον, υπάρχουν ακριβώς m υποπίνακες μη πλήρους κάλυψης, στον πίνακα που αποτελείται από τις γραμμές $1, 2, \ldots, r+t-1$.

Θεωρώντας την παρακάτω διαμέριση, του χώρου καταστάσεων Ω^*

$$\Omega^* = \bigcup_{m\ge 0} C_m, \quad C_m = \{(\omega, m) : \omega \in \Omega_2 \cup \{\omega_{s+1}\}\}, \quad m = 0, 1, \ldots, k-t+1,$$

εύκολα διαπιστώνουμε πως η τ.μ. $T_{k,n,t}$ γίνεται μια MVB, για την οποία ο πίνακας πιθανοτήτων μετάβασης A (βλ. και (3.1.1)) έχει τη μορφή

$$A = \begin{pmatrix} P & \mathbf{0}' \\ \mathbf{0} & 0 \end{pmatrix}_{(s+1)\times(s+1)} \tag{3.3.1}$$

όπου P είναι ο πίνακας μετάβασης που χρησιμοποιήσαμε για τον υπολογισμό της $P(T_{k,n,t} = 0)$. Η μορφή του A δικαιολογείται και από το γεγονός ότι οι πρώτες s γραμμές και s στήλες

του, αναφέρονται στις πιθανότητες μετάβασης ανάμεσα στις καταστάσεις του χώρου Ω_2, απ' όπου δεν προκύπτουν υποπίνακες μη πλήρους κάλυψης. Από την άλλη, εάν $Y_{r-1} = (\omega_{s+1}, m)$, είναι φανερό ότι η Y_r δεν μπορεί να περάσει σε καμία από τις καταστάσεις (ω, m), με $\omega \in \Omega_2$, και αυτό έχει ως αποτέλεσμα οι πιθανότητες στην τελευταία γραμμή του A, να μηδενίζονται.

Πριν συνεχίσουμε με τον πίνακα μετάβασης B, πρέπει να αναφέρουμε ότι οι πρώτες s συντεταγμένες του αρχικού διανύσματος πιθανοτήτων

$$\mathbf{f}_0(0) = (P(Y_0 = (\omega_1, 0)), P(Y_0 = (\omega_2, 0)), \ldots, P(Y_0 = (\omega_{s+1}, 0))),$$

υπολογίζονται άμεσα από την (3.2.9), ενώ η $P(Y_0 = (\omega_{s+1}, 0))$ είναι ίση με

$$P(Y_0 = (\omega_{s+1}, 0)) = 1 - \sum_{i=1}^{s} P(Y_0 = (\omega_i, 0)). \tag{3.3.2}$$

Ο πίνακας μετάβασης B, μπορεί να γραφεί ως

$$B = \begin{pmatrix} Q & \mathbf{c}' \\ \mathbf{b} & \rho \end{pmatrix}_{(s+1)\times(s+1).}$$

Οι πιθανότητες μετάβασης που περιέχονται στον Q, αναφέρονται στις περιπτώσεις

$$P(Y_r = (\omega, m+1) | Y_{r-1} = (\omega', m)),$$

όπου

$$\omega = (x_1, x_2, \ldots, x_a) \in \Omega_2 \text{ και } \omega' = (x_1', x_2', \ldots, x_a') \in \Omega_2.$$

Σύμφωνα με το Λήμμα 3.2.1, εάν δεν ικανοποιείται η συνθήκη (3.2.5), η κατάσταση ω είναι μη προσβάσιμη (με ένα βήμα) από την ω' και επομένως, $P(Y_r = (\omega, m+1) | Y_{r-1} = (\omega', m)) = 0$. Αν όμως η (3.2.5) ικανοποιείται, η πιθανότητα $P(Y_r = (\omega, m) | Y_{r-1} = (\omega', m))$, η οποία ταυτίζεται με την (3.2.6), έχει συμπεριληφθεί στον υποπίνακα P, του A. Άρα, ο υποπίνακας Q του B θα περιλαμβάνει την «υπόλοιπη» πιθανότητα, η οποία μπορεί να εκφραστεί μέσω του τύπου

$$\begin{aligned} &P(Y_r = (\omega, m+1) | Y_{r-1} = (\omega', m)) \\ &\qquad = p^{\sum_{i=1}^{a/2} x_{2i}} (1-p)^{n - \sum_{i=1}^{a/2} x_{2i}} \left(\prod_{i=1}^{a/2} d_i - \prod_{i=1}^{a/2} c_i \right), \end{aligned} \tag{3.3.3}$$

όπου

$$d_i = \binom{x_{2i} + x_{2i-1}}{x_{2i}}, i = 1, 2, \ldots, a/2.$$

Έχοντας προσδιορίσει τις πιθανότητες του πίνακα Q, το διάνυσμα $\mathbf{c}$ υπολογίζεται άμεσα από τη σχέση

$$\mathbf{c}' = \mathbf{1}' - Q\mathbf{1}' - P\mathbf{1}' \tag{3.3.4}$$

η οποία προκύπτει από το γεγονός ότι ο πίνακας $A + B$ είναι ένας στοχαστικός πίνακας.

Ας προχωρήσουμε τώρα στο διάνυσμα $\mathbf{b}$, το οποίο αποτελείται από πιθανότητες μετάβασης, της μορφής

$$P(Y_r = (\omega, m+1)|Y_{r-1} = (\omega_{s+1}, m)) \tag{3.3.5}$$

όπου $\omega = (x_1, x_2, \ldots, x_a) \in \Omega_2$. Επιπλέον, θα συμβολίζουμε με $\omega_{s+1,j} = (x'_{1j}, x'_{2j}, \ldots, x'_{aj})$, $j = 1,2,\ldots,h$ τις ακέραιες λύσεις της εξίσωσης

$$x'_{1j} + x'_{2j} + \ldots + x'_{aj} = n,$$

κάτω από τους περιορισμούς

- $x'_i < 2$, για τουλάχιστον ένα $i \in \{1, 2, \ldots, a\}$ και,
- $x'_{ij} + x'_{i+a/2,j} = x_{2i-1} + x_{2i}$ για κάθε $i \in \{1, 2, \ldots, a\}$.

Τότε, το πλήθος των λύσεων της παραπάνω εξίσωσης, είναι ίσο με

$$h = \prod_{i=1}^{a/2}(x_{2i} + x_{2i-1} + 1) - \prod_{i=1}^{a/2}(x_{2i} + x_{2i-1} - 3),$$

και η (3.3.5) γίνεται

$$\begin{aligned}
&P(Y_r = (\omega, m+1)|Y_{r-1} = (\omega_{s+1}, m)) \\
&= P(Y_1 = (\omega, 1)|Y_0 = (\omega_{s+1}, 0)) \\
&= \frac{1}{P(Y_0 = (\omega_{s+1}, 0))} \sum_{j=1}^{h} P(Y_0 = (\omega_{s+1,j}, 0)) P(Y_1 = (\omega, 1)|Y_0 = (\omega_{s+1,j}, 0)), \\
&= \frac{1}{P(Y_0 = (\omega_{s+1}, 0))} \sum_{j=1}^{h} P(Y_0 = (\omega_{s+1,j}, 0)) p_0, \\
&= \frac{p_0}{P(Y_0 = (\omega_{s+1}, 0))} \sum_{j=1}^{h} p_{\omega_{s+1,j}},
\end{aligned} \tag{3.3.6}$$

όπου $p_{\omega_{s+1,j}}$ είναι πιθανότητες οι οποίες υπολογίζονται μέσω παρόμοιων σχέσεων με τις (3.2.9), η $P(Y_0 = (\omega_{s+1}, 0))$ δίδεται από την (3.3.2) και

$$p_0 = p^{\sum_{i=1}^{a/2} x_{2i}} (1-p)^{n - \sum_{i=1}^{a/2} x_{2i}} \prod_{i=1}^{a/2} d_i.$$

Αξίζει να επισημάνουμε πως το άθροισμα που εμφανίζεται στην (3.3.6), μπορεί εναλλακτικά να εκφραστεί ως

$$\sum_{\substack{0 \le y_i \le x_{2i-1}+x_{2i},\\ \text{για } i=1,2,\ldots,a/2}} \frac{n!}{\prod_{i=1}^{a/2} y_i!(x_{2i}+x_{2i-1}-y_i)!}\pi_\omega - \sum_{\substack{2 \le y_i \le x_{2i-1}+x_{2i}-2,\\ \text{για } i=1,2,\ldots,a/2}} \frac{n!}{\prod_{i=1}^{a/2} y_i!(x_{2i}+x_{2i-1}-y_i)!}\pi_\omega$$

όπου

$$\pi_\omega = p^{\sum_{i=1}^{a/2}(x_{2i}+x_{2i-1})|w_i|+n-\sum_{i=1}^{a/2} y_i}(1-p)^{n(t-1)-\sum_{i=1}^{a/2}(x_{2i}+x_{2i-1})|w_i|-n+\sum_{i=1}^{a/2} y_i}.$$

Στην περίπτωση ανεξάρτητων και συμμετρικών δοκιμών Bernoulli, το π_ω δεν εξαρτάται από το $\omega \in \Omega_2$ ($\pi_\omega = 2^{-n(t-1)}$ για κάθε $\omega \in \Omega_2$), ενώ για $t = 2$ η πιθανότητα μετάβασης (3.3.6) λαμβάνει την ακόλουθη απλή μορφή

$$P(Y_r = (\omega, m+1)|Y_{r-1} = (\omega_{s+1}, m)) = \binom{n}{x_2} p^{x_2}(1-p)^{n-x_2}$$

(διότι, $P(Y_0 = (\omega_{s+1}, 0)) = \sum_{j=1}^{h} p_{\omega_{s+1},j}$). Η παραπάνω προσέγγιση, θα γίνει περισσότερο κατανοητή, μέσω του επόμενου παραδείγματος.

Παράδειγμα 3.5 Ας υποθέσουμε ότι $n = 10$, $t = 3$ και $p = 1/2$ (i.i.d. συμμετρικές δοκιμές Bernoulli). Τότε, ανακαλώντας τους συμβολισμούς και την ανάλυση του Παραδείγματος 3.4, μπορούμε να γράψουμε το χώρο καταστάσεων Ω^* ως

$$\Omega^* = \{(\omega, m) : \omega = (x_1, x_2, x_3, x_4) \in (\Omega_2 \cup \{\omega_{s+1}\}) \text{ και } 0 \le m \le k-2\},$$

όπου $\Omega_2 = \{\omega_1, \omega_2, \ldots, \omega_{10}\}$. Οι πρώτες 10 συντεταγμένες του αρχικού διανύσματος πιθανοτήτων $\mathbf{f}_0(0)$, ταυτίζονται με αυτές του $\boldsymbol{\pi}_0$ (βλ. (3.2.11)) και επομένως, από τη σχέση

$$\mathbf{f}_0(0) = (\boldsymbol{\pi}_0, 1 - \boldsymbol{\pi}_0 \mathbf{1}')$$

παίρνουμε

$$\mathbf{f}_0(0) = \frac{1}{2^{20}}(\alpha, \alpha, \alpha, \alpha, \beta, \beta, \beta, \beta, \beta, \beta, \gamma),$$

με, $\alpha = 18900, \beta = 25200$ και $\gamma = 2^{20} - 4\alpha - 6\beta$. Ο πίνακας μετάβασης A, θα έχει τη μορφή που περιγράφεται από την (3.3.1), όπου ο P θα δίδεται από την (3.2.10).

Οι πιθανότητες του Q υπολογίζονται, μέσω της (3.3.3). Στη δική μας περίπτωση, ο τύπος (3.3.3) γίνεται

$$P(Y_r = (\omega, m+1)|Y_{r-1} = (\omega', m)) = \frac{1}{2^{10}}\binom{x_1 + x_2}{x_2}\binom{x_3 + x_4}{x_4} - p(\omega, \omega')$$

όπου $p(\omega, \omega')$ είναι οι αντίστοιχες πιθανότητες του P. Έτσι, για $\omega_1 = (4, 2, 2, 2)$, παίρνουμε

$$P(Y_r = (\omega_1, m+1)|Y_{r-1} = (\omega_1, m)) = \frac{1}{2^{10}}\left(\binom{6}{2}\binom{4}{2} - 32\right) = \frac{58}{2^{10}}$$

και δουλεύοντας με παρόμοιο τρόπο για τις υπόλοιπες μη μηδενικές πιθανότητες του Q, καταλήγουμε στον πίνακα

$$Q = \frac{1}{2^{10}}\begin{pmatrix} 58 & 0 & 58 & 0 & 0 & 72 & 0 & 0 & 0 & 0 \\ 58 & 0 & 58 & 0 & 0 & 72 & 0 & 0 & 0 & 0 \\ 0 & 58 & 0 & 58 & 0 & 0 & 0 & 72 & 0 & 0 \\ 0 & 58 & 0 & 58 & 0 & 0 & 0 & 72 & 0 & 0 \\ 54 & 0 & 54 & 0 & 0 & 48 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 64 & 0 & 64 & 0 & 64 & 64 \\ 0 & 0 & 0 & 0 & 64 & 0 & 64 & 0 & 64 & 64 \\ 0 & 0 & 0 & 0 & 64 & 0 & 64 & 0 & 64 & 64 \\ 0 & 54 & 0 & 54 & 0 & 0 & 0 & 48 & 0 & 0 \\ 0 & 0 & 0 & 0 & 64 & 0 & 64 & 0 & 64 & 64 \end{pmatrix}.$$

Το διάνυσμα $\mathbf{c}$, που εμφανίζεται στην τελευταία στήλη του B, υπολογίζεται μέσω της (3.3.4), και βρίσκεται ίσο με

$$\mathbf{c}' = \mathbf{1}' - Q\mathbf{1}' - P\mathbf{1}' = \frac{1}{2^{10}}(\delta, \delta, \delta, \delta, \epsilon, \epsilon, \epsilon, \delta, \epsilon)'$$

όπου $\delta = 724$, $\epsilon = 624$. Τέλος, το $\mathbf{b}$ προκύπτει από την (3.3.6), ενώ για $n = 10, t = 3$ έχουμε

$$\mathbf{b} = \frac{1}{2^{10}}(\zeta, \zeta, \zeta, \zeta, \eta, \theta, \eta, \theta, \eta, \eta)$$

με $\zeta = 16.6513$, $\eta = 19.1351$ και $\theta = 22.2017$. Προφανώς, η πιθανότητα ρ που υπάρχει στην τελευταία στήλη και γραμμή του πίνακα B, θα είναι ίση με

$$\rho = 1 - \mathbf{b}\mathbf{1}' = 1 - 2^{-10}(4\zeta + 4\eta + 2\theta) = \frac{1}{2^{10}}836451.$$

Μέσω των αναδρομικών σχέσεων (3.1.2) μπορούμε να προσδιορίσουμε ολόκληρη την κατανομή της $T_{5,10,3}$ και να καταλήξουμε στα εξής:

$$\begin{aligned}
P(T_{5,10,3}=0) &= \mathbf{f}_0(0)A^3\mathbf{1}' = 0.00049\\
P(T_{5,10,3}=1) &= \mathbf{f}_0(0)(BA^2+ABA+A^2B)\mathbf{1}' = 0.00722\\
P(T_{5,10,3}=2) &= \mathbf{f}_0(0)(B^2A+BAB+AB^2)\mathbf{1}' = 0.06858\\
P(T_{5,10,3}=3) &= \mathbf{f}_0(0)B^3\mathbf{1}' = 0.92371.
\end{aligned}$$

■

3.4 Η περίπτωση της Μαρκοβιανής εξάρτησης

Η μέθοδος που αναλύσαμε στις προηγούμενες παραγράφους, μπορεί εύκολα να προσαρμοστεί και στην περίπτωση όπου οι τ.μ. που βρίσκονται στην ίδια στήλη του πίνακα, έχουν μια Μαρκοβιανή εξάρτηση, ενώ οι στήλες είναι μεταξύ τους ανεξάρτητες. Στη συγκεκριμένη παράγραφο θα περιγράψουμε εν συντομία τον τρόπο με τον οποίο θα αντιμετωπίσουμε το πρόβλημα για την περίπτωση $t=2$, και όταν οι τ.μ. που βρίσκονται στην ίδια στήλη, έχουν μια Μαρκοβιανή εξάρτηση πρώτης τάξης.

Θα αναφερθούμε μόνο στην περίπτωση $t=2$, ώστε οι διάφορες έννοιες να εξηγηθούν καλύτερα. Η γενίκευση των αποτελεσμάτων για $t\geq 2$, είναι άμεση (για Μαρκοβιανή εξάρτηση τάξεως, το πολύ $t-1$), ωστόσο οι συμβολισμοί γίνονται αρκετά πολύπλοκοι. Για το λόγο αυτό δε θα επεκταθούμε περισσότερο στο θέμα αυτό.

Έστω, $\mathbf{X}=(X_{ij})_{k\times n}$, ένας (δυαδικός) τυχαίος πίνακας με ανεξάρτητες στήλες και ας υποθέσουμε πως για δεδομένο $j\in\{1,2,\ldots,n\}$, η ακολουθία $\{X_{ij}, i=1,2,\ldots,k\}$ είναι μία ομογενής Μαρκοβιανή αλυσίδα, με δυο καταστάσεις και πιθανότητες μετάβασης

$$\begin{aligned}
P(X_{ij}=1|X_{i-1,j}=0)=p_{01}, \quad P(X_{ij}=0|X_{i-1,j}=0)=p_{00},\\
P(X_{ij}=1|X_{i-1,j}=1)=p_{11}, \quad P(X_{ij}=0|X_{i-1,j}=1)=p_{10},
\end{aligned}$$

για $i=1,2,\ldots,k$. Τότε, η τ.μ. που απαριθμεί το πλήθος των $2\times n$ υποπινάκων, μη πλήρους κάλυψης, μπορεί να αντιμετωπιστεί ως MVB τ.μ. με χώρο καταστάσεων

$$\Omega^*=\{(x,m): x\in\{2,3,\ldots,n-2,\omega^*\}, m=0,1,\ldots,k-t+1\},$$

όπου η τ.μ. $x \in \{2,3,\ldots,n-2\}$ μετράει το πλήθος των άσσων στη γραμμή που βρισκόμαστε, ενώ όταν $x = \omega^*$, σημαίνει ότι το πλήθος των άσσων είναι ίσο με $0, 1, n-1$ ή n. Η απαριθμήτρια m αναφέρεται στο πλήθος των $2 \times n$ υποπινάκων μη πλήρους κάλυψης, που έχουν εντοπιστεί μέχρι τη χρονική στιγμή που βρισκόμαστε.

Οι πιθανότητες μετάβασης του πίνακα P (υποπίνακας του A) είναι τώρα ίσες με

$$\begin{aligned} &P(Y_r = (x,m)|Y_{r-1} = (x',m)) \\ &\quad = \sum_{j=\max\{1,x+x'+1-n\}}^{\min\{x-1,x'-1\}} \binom{x'}{j}\binom{n-x'}{x-j} p_{11}^{j} p_{01}^{x-j} p_{10}^{x'-j} p_{00}^{n-x'-x+j} \end{aligned}$$

για $x, x' \in \{2,3,\ldots,n-2\}$ ενώ οι υπόλοιπες πιθανότητες του A, μηδενίζονται. Οι πιθανότητες μετάβασης του πίνακα Q, δίδονται από την έκφραση

$$\begin{aligned} &P(Y_r = (x,m+1)|Y_{r-1} = (x',m)) \\ &\quad = \sum_{j=0}^{x} \binom{x'}{j}\binom{n-x'}{x-j} p_{11}^{j} p_{01}^{x-j} p_{10}^{x'-j} p_{00}^{n-x'-x+j} \\ &\quad - \sum_{j=\max\{1,x+x'+1-n\}}^{\min\{x-1,x'-1\}} \binom{x'}{j}\binom{n-x'}{x-j} p_{11}^{j} p_{01}^{x-j} p_{10}^{x'-j} p_{00}^{n-x'-x+j} \end{aligned}$$

για $x, x' \in \{2,3,\ldots,n-2\}$. Το διάνυσμα $\mathbf{b}$ (που εμφανίζεται στην τελευταία γραμμή του πίνακα B) υπολογίζεται από τη σχέση

$$\begin{aligned} &P(Y_r = (x,m+1)|Y_{r-1} = (\omega^*,m)) \\ &= \sum_{y\in\{0,1,n-1,n\}} \sum_{j=0}^{x} \binom{y}{j}\binom{n-y}{x-j} p_{11}^{j} p_{01}^{x-j} p_{10}^{y-j} p_{00}^{n-y-x+j}, x \in \{2,3,\ldots,n-2\}. \end{aligned}$$

Εάν υποθέσουμε ότι η πρώτη γραμμή του πίνακα, προέρχεται από i.i.d. δίτιμες τ.μ. με πιθανότητα (αποτυχίας) επιτυχίας p $(1-p)$, τότε το αρχικό διάνυσμα πιθανοτήτων $\boldsymbol{f}_0(0)$, το οποίο είναι απαραίτητο για τις αρχικές συνθήκες των (3.1.2), θα έχει συντεταγμένες

$$\binom{n}{i+1} p^{i+1}(1-p)^{n-i-1}, \quad i = 1,2,\ldots,n-2,$$

ενώ η τελευταία συντεταγμένη του, είναι η διαφορά από τη μονάδα, του αθροίσματος των υπολοίπων.

3.5 Ορθογώνιοι πίνακες, συνεχόμενης πλήρους κάλυψης

Έστω ένας πίνακας με k γραμμές, n στήλες και στοιχεία από ένα αλφάβητο με q γράμματα. Όπως έχουμε ήδη αναφέρει, ένας τέτοιος πίνακας θα ονομάζεται ορθογώνιος πίνακας πλήρους κάλυψης μεγέθους t, και συχνότητας c, εάν κάθε $t \times n$ υποπίνακάς του, περιέχει στις στήλες του κάθε μία από τις q^t διαφορετικές λέξεις μήκους t, ακριβώς c φορές (όπου $c \in \{1, 2, \ldots\}$ και $n = cq^t$).

Για παράδειγμα, ας θεωρήσουμε τον επόμενο πίνακα με $k = 3$ γραμμές και $n = 8$ στήλες (με $q = 2$):

$$\begin{pmatrix} 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

Τότε, οι 2×8 υποπίνακες, οι οποίοι σχηματίζονται από τις γραμμές $\{1, 2\}$, $\{2, 3\}$ και $\{1, 3\}$, είναι αντίστοιχα οι

$$\begin{pmatrix} 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{pmatrix} \text{ (πρώτη και δεύτερη γραμμή)}$$

$$\begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \text{ (δεύτερη και τρίτη γραμμή)}$$

$$\begin{pmatrix} 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \text{ (πρώτη και τρίτη γραμμή) .}$$

Παρατηρώντας πως καθένας απ' αυτούς, περιέχει ως στήλες, κάθε μια από τις λέξεις μήκους 2, ακριβώς δυο φορές, οδηγούμαστε στο συμπέρασμα ότι ο παραπάνω πίνακας είναι ένας ορθογώνιος πίνακας πλήρους κάλυψης, μεγέθους $t = 2$, συχνότητας $c = 2$.

Οι ορθογώνιοι πίνακες (όπως και οι t-CA) παρουσιάζουν σημαντικές εφαρμογές στο χώρο των πειραματικών σχεδιασμών, και έχουν άμεση σχέση με τη θεωρία της συνδυαστικής. Τα προβλήματα που απασχολούν την ερευνητική κοινότητα, είναι παρόμοια μ' αυτά των t-CA (η «ύπαρξη» ενός ορθογώνιου πίνακα, για δεδομένες τιμές των παραμέτρων και η ανάπτυξη αποτελεσματικών αλγόριθμων, για την κατασκευή του). Για μία ανασκόπηση στη θεωρία των παραπάνω πινάκων, και τη μελέτη των εφαρμογών τους, μπορούμε να ανατρέξουμε στη μονογραφία των Hedayat et al (1999), όπου θα διαπιστώσουμε και το μεγάλο πλήθος των ερωτημάτων στα οποία δεν έχουν δοθεί ακόμη απαντήσεις.

Με την ίδια λογική που εισάγαμε την έννοια των πινάκων συνεχόμενης πλήρους κάλυψης, βασισμένοι στους πίνακες πλήρους κάλυψης, μπορούμε να εισάγουμε μια αντίστοιχη κλάση ορθογώνιων πινάκων. Έτσι, ένας $k \times n$ πίνακας θα ονομάζεται *ορθογώνιος πίνακας συνεχόμενης πλήρους κάλυψης, μεγέθους t και συχνότητας c*, εάν κάθε $t \times n$ υποπίνακας του, που σχηματίζεται από t συνεχόμενες γραμμές, έχει ως στήλες κάθε μια από τις q^t δυνατές λέξεις, μήκους t, ακριβώς c φορές (όπου, $c \in \{1, 2, \ldots\}$ και $n = cq^t$). Για παράδειγμα, προκειμένου να εξετάσουμε εάν ο 3×8 πίνακας που αναφέραμε προηγουμένως είναι ορθογώνιος πίνακας συνεχόμενης πλήρους κάλυψης, μεγέθους 2 (και συχνότητας $c = 2$), θα πρέπει να ελέγξουμε τους υποπίνακες που σχηματίζονται από τις γραμμές $\{1, 2\}$ και $\{2, 3\}$ (δε μας ενδιαφέρει ο υποπίνακας που σχηματίζεται από τις γραμμές $\{1, 3\}$, καθώς η πρώτη και η τρίτη γραμμή δεν είναι συνεχόμενες).

Θα συμβολίζουμε με $COA(k, t, c)$, το ενδεχόμενο ένας τυχαίος $k \times n$ πίνακας (όπου $n = cq^t$), με στοιχεία i.i.d. δοκιμές Bernoulli, να είναι ορθογώνιος πίνακας συνεχόμενης πλήρους κάλυψης, μεγέθους t και συχνότητας c. Τότε για την πιθανότητα του ενδεχομένου $COA(k, t, c)$, παίρνουμε το επόμενο αποτέλεσμα.

Θεώρημα 3.5.1 *Έστω ένας $k \times n$ τυχαίος πίνακας, με στοιχεία i.i.d. δοκιμές Bernoulli. Τότε η πιθανότητα του ενδεχομένου $COA(k, t, c)$, δίνεται από τον τύπο ($a = 2^{t-1}$)*

$$P(COA(k,t,c)) = \frac{(2^t c)!}{((2c)!)^a} \binom{2c}{c}^{a(k-t+1)} ((1-p)p)^{cak}.$$

Απόδειξη. Για την απόδειξη του θεωρήματος θα ακολουθήσουμε παρόμοια μέθοδο μ' αυτή που χρησιμοποιήσαμε για τον υπολογισμό της πιθανότητας ένας $k \times n$ πίνακας (με $q = 2$), να είναι t-CCA (δηλαδή, της πιθανότητα $P(T_{k,n,t} = 0)$). Αρχικώς, παρατηρούμε ότι εάν ένας $t \times n$ πίνακας, έχει ως στήλες κάθε μια από τις λέξεις μήκους t, ακριβώς c φορές, τότε θα έχει και κάθε μια από τις λέξεις μήκους $t-1$, ακριβώς $2c$ φορές.

Ας ορίσουμε ένα νέο χώρο καταστάσεων Ω, ο οποίος θα είναι η ένωση των στοιχείων του Ω_3 και μίας κατάστασης απορρόφησης x_{abs}. Ο Ω_3 περιλαμβάνει τις ακέραιες λύσεις της εξίσωσης

$$x_1 + x_2 + \ldots + x_a = c2^t, \quad a = 2^{t-1}$$

κάτω από τους περιορισμούς $x_i = 2c$, δηλαδή, ο Ω_3 περιέχει τη μοναδική λύση,

$$(x_1, x_2, \ldots, x_a) = (2c, 2c, \ldots, 2c).$$

Επομένως, ο Ω αποτελείται από δυο στοιχεία, την κατάσταση $\omega = (2c, 2c, \ldots, 2c)$ και την x_{abs}. Η Μαρκοβιανή αλυσίδα $\{Y_r, r = 0, 1, \ldots\}$, ορίζεται επάνω στον Ω, με τον εξης τρόπο:

$$Y_r = (2c, 2c, \ldots, 2c), \quad r \geq 1,$$

εάν και μόνο εάν, το πλήθος των εμφανίσεων της λέξης w_i, στον $(t-1)\times n$ υποπίνακα, του οποίου η τελευταία γραμμή είναι η $r+t-1$, ισούται με $2c$ (για κάθε $i=1,2,\ldots,a$) και ταυτόχρονα, ο πίνακας που αποτελείται από τις γραμμές $1,2,\ldots,r+t-1$, είναι ένας ορθογώνιος πίνακας, συνεχόμενης πλήρους κάλυψης. Σε οποιαδήποτε άλλη περίπτωση, θεωρούμε ότι βρισκόμαστε στην κατάσταση απορρόφησης.

Η πιθανότητα μετάβασης,

$$P(Y_r=(2c,2c,\ldots,2c)|Y_{r-1}=(2c,2c,\ldots,2c))$$

είναι ίση με

$$P(Y_r=(2c,2c,\ldots,2c)|Y_{r-1}=(2c,2c,\ldots,2c))=\left(\binom{2c}{c}p^c(1-p)^c\right)^{2^{t-1}}=\pi_{11}.$$

Έτσι, ο πίνακας πιθανοτήτων μετάβασης της αλυσίδας είναι της μορφής

$$\Lambda=\begin{pmatrix}\pi_{11} & 1-\pi_{11}\\ 0 & 1\end{pmatrix}$$

οπότε

$$P(COA(k,t,c))=p_1(\pi_{11})^{k-t+1}=p_1\left(\binom{2c}{c}p^c(1-p)^c\right)^{2^{t-1}(k-t+1)}$$

όπου p_1 είναι η πιθανότητα να πάρουμε ένα $(t-1)\times n$ πίνακα, με τη δομή $(2c,2c,\ldots,2c)$. Η πιθανότητα p_1 αποδεικνύεται άμεσα ότι είναι ίση με

$$p_1=\frac{(2^tc)!}{((2c)!)^a}(1-p)^{2cs(t)}p^{c(t-1)2^t-2cs(t)}$$

όπου

$$s(t)=\sum_{i=0}^{t-1}\binom{t-1}{i}(t-1-i)=(t-1)2^{t-2}$$

και έτσι ολοκληρώνεται η απόδειξη του θεωρήματος.

■

Τελειώνοντας, αξίζει να αναφέρουμε πως η πιθανότητα $P(COA(k,t,c))$ αποτελεί ένα κάτω φράγμα για την $P(T_{k,c2^t,t}=0)$, δηλαδή,

$$P(T_{k,c2^t,t}=0)\geq P(COA(k,t,c)),$$

ενώ αντίστροφα, η $P(COA(k,t,c))$ είναι μεγαλύτερη ή ίση από την πιθανότητα ένας πίνακας να είναι ορθογώνιος πλήρους κάλυψης. Στην επόμενη παράγραφο, θα εξετάσουμε τον τρόπο με τον οποίο όλα τα παραπάνω αποτελέσματα, θα χρησιμοποιηθούν στη θεωρία των παραγοντικών σχεδιασμών.

3.6 Εφαρμογές και αριθμητικά αποτελέσματα

Μια ενδεχόμενη εφαρμογή των t-CCA προέρχεται από το πεδίο των παραγοντικών σχεδιασμών. Ας υποθέσουμε ότι ενδιαφερόμαστε για την επίδραση $t = 3$ παραγόντων A, B, C (με δυο επίπεδα ο καθένας), σε μια συνεχή μεταβλητή απόκρισης Z. Επίσης, ας θεωρήσουμε ότι χρησιμοποιούμε ένα τυχαίο σχεδιασμό, με τον εξής τρόπο: σε κάθε χρονική στιγμή αντιστοιχούνται τυχαία n μετρήσεις, που αφορούν το επίπεδο ενός συγκεκριμένου παράγοντα (βλ. π.χ. Dalal and Mallows (1998)). Ας συμβολίσουμε με $p, 1-p$ την πιθανότητα ένας παράγοντας να είναι στο επίπεδο 1, 0, αντιστοίχως.

Ακόμη, υποθέτουμε πως η τυχαία αντιστοίχηση των επιπέδων σε κάποιο παράγοντα, επαναλαμβάνεται κάθε $t = 3$ χρονικές στιγμές (π.χ. ημέρες), με ένα κυκλικό τρόπο. Αυτό σημαίνει ότι την πρώτη ημέρα παίρνουμε n μετρήσεις που αφορούν τα επίπεδα του παράγοντα A, τη δεύτερη ημέρα συλλέγουμε n μετρήσεις για τα επίπεδα του παράγοντα B (ενώ τα επίπεδα του παράγοντα A παραμένουν ίδια), την τρίτη ημέρα οι πληροφορίες αφορούν τα επίπεδα του παράγοντα C (τα επίπεδα των παραγόντων A, B παραμένουν), ενώ την τέταρτη ημέρα παίρνουμε μετρήσεις για τον παράγοντα A (επανερχόμαστε στον πρώτο παράγοντα), τη πέμπτη ημέρα ασχολούμαστε με τον B κ.ο.κ.

Εάν τώρα συμβολίσουμε με $Z_{i-2,j}, j = 1, 2, \ldots, n$ την τιμή της συνεχούς τ.μ. που αντιστοιχεί στην j-οστή μέτρηση, της i ημέρας, όπου $i = 3, 4, \ldots, k$, τότε το περιβάλλον του πειράματος μας, μπορεί να αναπαρασταθεί από τον Πίνακα 3.6.1 (το D είναι κάποιο από τα A, B, C).

Πίνακας 3.6.1: Πλήρης παραγοντικός σχεδιασμός.

		επίπεδα για την i μέτρηση					απόκριση για την i μέτρηση				
ημέρα	παράγοντας	1	2	3	...	n	1	2	3	...	n
1	A	0	1	0	...	0					
2	B	1	1	0	...	0					
3	C	1	0	1	...	1	z_{11}	z_{12}	z_{13}	...	z_{1n}
4	A	0	1	1	...	1	z_{21}	z_{22}	z_{23}	...	z_{2n}
5	B	1	0	0	...	0	z_{31}	z_{32}	z_{33}	...	z_{3n}
6	C	0	0	0	...	1	z_{41}	z_{42}	z_{43}	...	z_{4n}
7	A	1	1	0	...	1	z_{51}	z_{52}	z_{53}	...	z_{5n}
⋮	⋮	⋮	⋮	⋮	...	⋮	⋮	⋮	⋮	...	⋮
k	D	1	0	1	...	0	$z_{k-2,1}$	$z_{k-2,2}$	$z_{k-2,3}$	...	$z_{k-2,n}$

Έστω ότι επιθυμούμε για κάθε μέρα, να έχουμε δεδομένα για ένα πλήρη παραγοντικό σχεδιασμό, ώστε να μπορούμε να μελετήσουμε όλες τις κύριες επιδράσεις και τις αλληλεπιδράσεις, ανάμεσα σε δυο παράγοντες και όλες τις αλληλεπιδράσεις ανάμεσα σε τρεις παράγοντες. Αυτό σημαίνει ισοδύναμα ότι, στο δεύτερο block του πίνακα θα πρέπει να εμφανισθούν όλες οι λέξεις μήκους 3, τουλάχιστον μία φορά, δηλαδή θα πρέπει ο $k \times n$ πίνακας σχεδιασμού, να είναι ένας t-CCA, με $t = 3$.

Η πιθανότητα του ενδεχομένου $T_{k,n,t} = 0$, που μελετήθηκε στην προηγούμενη παράγραφο, αντιστοιχεί στην πιθανότητα να έχουμε ένα πλήρη παραγοντικό σχεδιασμό- με βάση τη διαδικασία που περιγράψαμε προηγουμένως-σε μία περίοδο k ημερών. Μερικά πρακτικής σημασίας (και όχι μόνο), ερωτήματα που ανακύπτουν, είναι τα εξής:

i. για μια δεδομένη χρονική περίοδο k ημερών, ποιος είναι ο ελάχιστος αριθμός μετρήσεων n που πρέπει να πάρουμε, ώστε η πιθανότητα να έχουμε ένα πλήρη παραγοντικό σχεδιασμό να είναι μεγαλύτερη από κάποια τιμή (επίπεδο σημαντικότητας);

ii. για ένα δεδομένο αριθμό μετρήσεων n, ποια είναι η μέγιστη χρονική περίοδος (η μέγιστη τιμή του k) που μπορούμε να μελετάμε το πρόβλημά μας, ώστε η πιθανότητα να έχουμε ένα πλήρη παραγοντικό σχεδιασμό να είναι μεγαλύτερη από κάποια προκαθορισμένη τιμή;

iii. ποιες είναι οι αντίστοιχες τιμές για τα n (για την ερώτηση (i)) ή το k (για την ερώτηση (ii)), εάν ενδιαφερόμαστε και για μη πλήρεις παραγοντικούς σχεδιασμούς; Για παράδειγμα, μπορούμε να ζητήσουμε η $P(T_{k,n,t} \leq 1)$ να είναι μεγαλύτερη από κάποιο επίπεδο, για συγκεκριμένο χρονικό ορίζοντα ή πλήθος δοκιμών.

Βέβαια, εάν έχουμε στη διάθεση μας t παράγοντες (και όχι τρεις), τότε όλα τα παραπάνω μπορούν εύκολα να μελετηθούν με τη βοήθεια του γενικού προβλήματος t-CCA, απαντώντας στις αντίστοιχες ερωτήσεις. Έτσι, ας θεωρήσουμε ότι ενδιαφερόμαστε για την επίδραση t παραγόντων,$A_1, A_2, \ldots, A_t$ (καθένας από τους οποίους έχει δυο επίπεδα), στη συνεχή τ.μ. απόκρισης Z. Τα επίπεδα των παραγόντων προσδιορίζονται από τυχαίο μηχανισμό, με βάση τον οποίο ένας παράγοντας βρίσκεται στο επίπεδο 1, με πιθανότητα p και με πιθανότητα $1 - p$, στο επίπεδο 0. Σε μια χρονική στιγμή, παίρνουμε n μετρήσεις για ένα μόνο από τους t παράγοντες, έστω τον A_1. Την επόμενη ημέρα, παίρνουμε πάλι n μετρήσεις, για τον παράγοντα A_2 και αυτό συνεχίζεται μέχρι την t-οστη ημέρα, όπου θα πάρουμε τα επίπεδα για τον παράγοντα A_t. Αφού πάρουμε τα επίπεδα για κάθε ένα παράγοντα μια φορά (με τη σειρά που αναφέρθηκε προηγουμένως), επαναλαμβάνουμε τη διαδικασία (με μια κυκλική σειρά), ξεκινώντας πάλι από τον πρώτο παράγοντα, και ουσιαστικά θα μελετάμε τους παράγοντες

(θα συλλέγουμε πληροφορίες για τη μεταβλητή απόκρισης Z, μεταβάλλοντας τυχαία τα επίπεδα τους), με τη σειρά: $A_1, A_2, \ldots, A_t, A_1, A_2, \ldots$ κ.ο.κ.

Με τον τρόπο που παίρνουμε τα δεδομένα, τις πρώτες $t-1$ ημέρες δεν είναι δυνατόν να συλλέγουμε πληροφορίες για τη συνεχή μεταβλητή απόκρισης Z, καθώς μέχρι τότε δεν έχουμε δεδομένα για όλους τους παράγοντες. Έτσι, οι τιμές της Z, έστω $z_{ij}, j = 1, 2, \ldots, n$ θα αναφέρονται στις t προηγούμενες ημέρες, από τη στιγμή i, δηλαδή, τις ημέρες $i-t+1, i-t+2, \ldots, i$.

Πίνακας 3.6.2: Περίπτωση $t = 2$ και $p = 1/2$ ($P(T_{k,n,2} = 0)$).

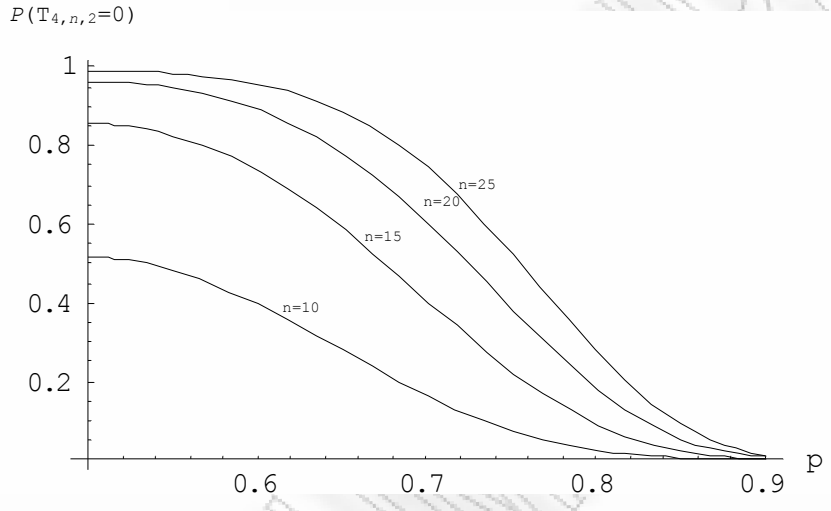
k	$n=12$	$n=13$	$n=14$	$n=15$	$n=16$	$n=17$	$n=18$	$n=19$	$n=20$
2	0.8748	0.9057	0.9291	0.9467	0.9600	0.9700	0.9775	0.9831	0.9873
3	0.7782	0.8293	0.8694	0.9004	0.9244	0.9427	0.9566	0.9673	0.9753
4	0.6923	0.7593	0.8135	0.8564	0.8900	0.9161	0.9363	0.9517	0.9634
5	0.6158	0.6953	0.7612	0.8145	0.8570	0.8903	0.9163	0.9363	0.9517
6	0.5478	0.6366	0.7122	0.7747	0.8252	0.8653	0.8968	0.9212	0.9401
7	0.4873	0.5829	0.6664	0.7368	0.7945	0.8409	0.8777	0.9064	0.9287
8	0.4335	0.5337	0.6236	0.7007	0.7650	0.8172	0.8589	0.8918	0.9174

Αν ο στόχος μας είναι να υλοποιούμε ένα πλήρη παραγοντικό σχεδιασμό, για τις ημέρες $t, t+1, \ldots, k$ είναι φανερό ότι ο $k \times n$ πίνακας σχεδιασμού, πρέπει να είναι ένας t-CCA (με $q = 2$). Τότε, για κάποιο k, ο ελάχιστος αριθμός μετρήσεων που πρέπει να πάρουμε ώστε να συμβεί το παραπάνω, με πιθανότητα μεγαλύτερη από $1-a$ $(0 < a < 1)$, εκφράζεται ως εξής:

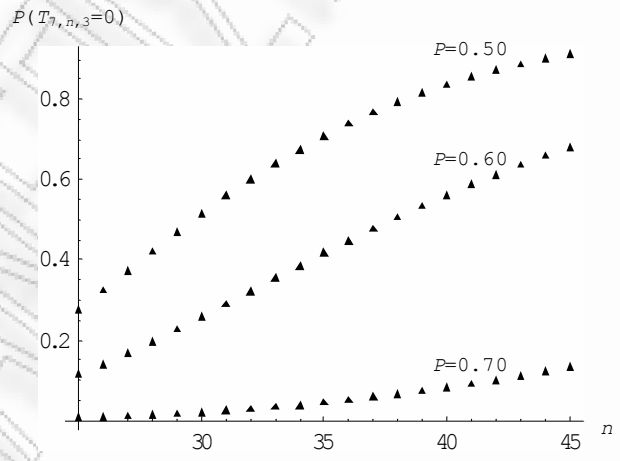
$$n_{min} = \min\{n : P(T_{k,n,t} = 0) \geq 1 - a\}.$$

Στην παρούσα ενότητα θα χρησιμοποιήσουμε τα αποτελέσματα των προηγούμενων παραγράφων, ώστε να δώσουμε αριθμητικά αποτελέσματα και απαντήσεις, που αφορούν το n_{min}, άλλα και ερωτήματα σαν αυτά που περιγράψαμε στην αρχή της παραγράφου. Έτσι, στον Πίνακα 3.6.2, υπάρχουν οι τιμές της πιθανότητας $P(T_{k,n,t} = 0)$ για $t = 2$, $p = 1/2$, οι οποίες μπορούν να χρησιμοποιηθούν ώστε να απαντήσουμε στο ερώτημα (i). Για παράδειγμα, εάν θέλουμε πιθανότητα τουλάχιστον 90%, για να έχουμε ενα πληρη παραγοντικό σχεδιασμό για $k = 5$, τότε θα πρεπει να πάρουμε τουλαχιστον $n_{min} = 18$ μετρησεις, κάθε ημέρα. Εαν ο χρονικος μας οριζοντας μεγαλώσει στις $k = 8$ ημέρες, τότε απαιτούνται $n_{min} = 20$ μετρήσεις, για να πετύχουμε τον ίδιο στόχο. Επίσης, παρόμοιες πληροφορίες μπορούν να εξαχθούν και από τα Σχήματα 3.6.1, 3.6.2 και 3.6.3, όπου γίνεται μια αριθμητική μελέτη

της $P(T_{k,n,t}=0)$, για διάφορες τιμές των παραμέτρων. Για παράδειγμα, στο Σχήμα 3.6.1 παρατηρούμε πως μεταβάλεται η πιθανότητα $P(T_{4,n,2}=0)$, ως συνάρτηση του p, και για διάφορα n.

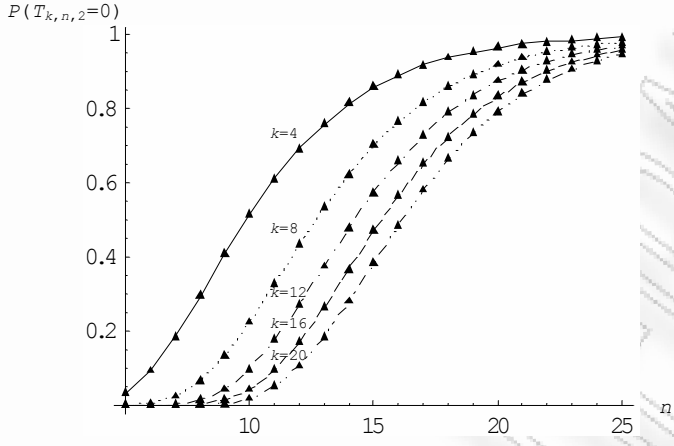


Σχήμα 3.6.1: Η πιθανότητα $P(T_{4,n,2}=0)$, για $p \geq 0.5$.



Σχήμα 3.6.2: Η πιθανότητα $P(T_{7,n,3}=0)$, για $p=0.50, 0.60, 0.70$.

Ο Πίνακας 3.6.3, μπορεί να χρησιμοποιηθεί για την εύρεση του ελάχιστου πλήθους μετρήσεων n, ώστε να ισχύει $P(T_{k,n,t}=0) \geq 1-a$, για $t=3, p=0.50$ και διάφορες επιλογές του a ($a=0.01, 0.025, 0.05, 0.10$) και του k ($k=3,4,\ldots,16$). Παράλληλα, ο Πίνακας 3.6.4 αναφέρεται σε αντίστοιχα αποτελέσματα, όταν δεν ισχύει $p=0.50$. Αξίζει



Σχήμα 3.6.3: Η πιθανότητα $P(T_{k,n,2}=0)$, για $p=0.50$.

να σημειώσουμε το πόσο πολύ αυξάνεται το ελάχιστο πλήθος μετρήσεων n (ώστε να ισχύει $P(T_{k,n,t}=0)\geq 1-a$), καθώς απομακρυνόμαστε από την τιμή $p=0.50$ (π.χ. για $k=4$ και $a=0.10$, ενώ χρειαζόμασταν 38 μετρήσεις, όταν $p=0.50$, οι μετρήσεις που απαιτούνται γίνονται 41, 51 και 70, για $p=0.55, 0.60, 0.65$, αντιστοίχως).

Η απάντηση στην ερώτηση (ii), δίδεται μέσω της εύρεσης του k_{max}, όπου

$$k_{max}=\max\{k : P(T_{k,n,t}=0)\geq 1-a\}.$$

Πίνακας 3.6.3: Ο ελάχιστος αριθμός n, για τον οποίο ισχύει $P(T_{k,n,3}=0)\geq 1-a$ $(p=0.50)$.

k	$a=0.10$	$a=0.05$	$a=0.025$	$a=0.01$	k	$a=0.10$	$a=0.05$	$a=0.025$	$a=0.01$
3	33	38	44	50	10	48	54	59	66
4	38	43	48	55	11	49	54	60	66
5	41	46	52	58	12	50	55	61	67
6	43	48	54	60	13	51	56	61	68
7	45	50	55	62	14	51	57	62	69
8	46	51	57	63	15	52	57	62	69
9	47	53	58	65	16	52	58	63	70

Από τον Πίνακα 3.6.2 $(t=2, p=1/2)$ παίρνουμε π.χ. ότι, εάν $n=15$ τότε ο μέγιστος αριθμός ημερών που μπορούμε να μελετήσουμε το πρόβλημα μας, ώστε να έχουμε ένα πλήρη παραγοντικό σχεδιασμό με πιθανότητα τουλάχιστον 90%, ειναι $k=3$ ημέρες. Εάν όμως συλλέγουμε $n=19$ μετρήσεις κάθε ημέρα, τότε το πλήθος των ημερών αυξάνεται σε $k=7$.

Ο Πίνακας 3.6.3 μπορεί επίσης να χρησιμοποιηθεί για τον υπολογισμό του μέγιστου k_{max} για $t = 3, p = 1/2$, και διάφορες τιμές του n και a.

Τέλος, στο ερώτημα (iii) ενδιαφερόμαστε π.χ., για την εύρεση του n έτσι ώστε να ισχύει

$$n_{min}(r) = \min\{n : P(T_{k,n,t} \leq r) \geq 1 - a\},$$

για προκαθορισμένες τιμές των παραμέτρων $r \in \{1, 2, \ldots\}$ και $a \in (0, 1)$. Ο Πίνακας 3.6.6 μας προσφέρει την αθροιστική συνάρτηση κατανομής (CDF) της $T_{k,n,t}$ για $t = 3$, $p = 1/2$ και για διάφορες επιλογές των n και k. Με βάση αυτά τα αριθμητικά αποτελέσματα, εάν θέλουμε με πιθανότητα τουλάχιστον 90% το πλήθος των υποπινάκων μη πλήρους κάλυψης, να είναι το πολυ 1, τότε θα πρέπει να παίρνουμε τουλάχιστον $n_{min}(1) = 26$ μετρήσεις, καθημερινώς, για χρονικό ορίζοντα $k = 4$, ή $n_{min}(1) = 29$ για $k = 5$ κ.ο.κ. Προφανώς, η τιμή του n που απαιτείται κατω απο το πλαίσιο αυτό, είναι πάντα μικρότερη από την αντίστοιχη, που αναφέρεται σ' ενα πληρη παραγοντικο σχεδιασμό (εφόσον οι υπόλοιπες παράμετροι παραμένουν αμετάβλητες).

Πίνακας 3.6.4: Ο ελάχιστος αριθμός n, για τον οποίο ισχύει $P(T_{k,n,3} = 0) \geq 1 - a$, για $p = 0.55, 0.60, 0.65$.

	$p = 0.55$		$p = 0.60$		$p = 0.65$	
k	$a = 0.10$	$a = 0.05$	$a = 0.10$	$a = 0.05$	$a = 0.10$	$a = 0.05$
3	35	41	43	51	58	70
4	41	47	51	59	70	84
5	44	49	55	64	78	91
6	46	53	59	68	84	97
7	48	55	62	71	88	102
8	50	56	65	74	92	107
9	52	58	66	75	96	111

Τέλος, έστω $n = c2^t$ και ότι για κάθε ημέρα δε θέλουμε απλά να έχουμε ένα πλήρη παραγοντικό σχεδιασμό, αλλά κάθε μία από τις 2^t θεραπείες, να έχει χρησιμοποιηθεί (για τη μέτρηση της μεταβλητής απόκρισης), ακριβώς c φορές (δηλαδή, θέλουμε να έχουμε ακριβώς c επαναλαμβανόμενες μετρήσεις, ανά θεραπεία). Τότε, η πιθανότητα να συμβαίνει κάτι τέτοιο, είναι ίση με την πιθανότητα του ενδεχομένου $COA(k, t, c)$. Έτσι από το Θεώρημα 3.5.1, προκύπτει άμεσα ο Πίνακας 3.6.5, ο οποίος περιέχει τις $P(COA(k, t, c))$, για διάφορες τιμές των c, t, k.

Πίνακας 3.6.5: Η πιθανότητα $P(COA(k,t,c))$.

	k	$c=1$	$c=2$	$c=3$	$c=4$
$t=2$	2	0.0938	0.0385	0.0220	0.0147
	3	0.0234	0.0054	0.0022	0.0011
	4	0.0059	0.0008	0.0002	0.0001
$t=3$	3	0.0024	0.0003	0.0001	0.00003
	4	0.0002	$5.742 \cdot 10^{-6}$	$7.450 \cdot 10^{-7}$	$1.687 \cdot 10^{-7}$
	5	$9.388 \cdot 10^{-6}$	$1.136 \cdot 10^{-7}$	$7.114 \cdot 10^{-9}$	$9.429 \cdot 10^{-10}$

Πίνακας 3.6.6: Περίπτωση $t=3$ και $p=1/2$ $(CDF : P(T_{k,n,3} \leq i))$.

k	i	$n=25$	$n=26$	$n=27$	$n=28$	$n=29$	$n=30$	$n=31$	$n=32$	$n=33$
4	0	0.5734	0.6157	0.6549	0.6910	0.7240	0.7541	0.7813	0.8058	0.8279
	1	0.8999	0.9184	0.9336	0.9461	0.9562	0.9645	0.9713	0.9767	0.9812
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	0	0.4485	0.4962	0.5419	0.5851	0.6255	0.6631	0.6978	0.7296	0.7586
	1	0.7995	0.8334	0.8622	0.8865	0.9067	0.9236	0.9376	0.9491	0.9585
	2	0.9619	0.9714	0.9787	0.9841	0.9881	0.9912	0.9935	0.9951	0.9963
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	0	0.3508	0.3999	0.4483	0.4953	0.5404	0.5831	0.6232	0.6606	0.6952
	1	0.7004	0.7470	0.7878	0.8230	0.8531	0.8786	0.9001	0.9180	0.9329
	2	0.9095	0.9308	0.9474	0.9603	0.9701	0.9775	0.9832	0.9875	0.9907
	3	0.9860	0.9903	0.9934	0.9955	0.9970	0.9979	0.9986	0.9991	0.9994
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	0	0.2743	0.3223	0.3709	0.4194	0.4668	0.5127	0.5566	0.5981	0.6370
	1	0.6065	0.6628	0.7134	0.7582	0.7973	0.8310	0.8599	0.8843	0.9048
	2	0.8468	0.8809	0.9081	0.9296	0.9464	0.9594	0.9695	0.9771	0.9828
	3	0.9608	0.9726	0.9809	0.9868	0.9909	0.9938	0.9958	0.9971	0.9980
	4	0.9948	0.9968	0.9980	0.9987	0.9992	0.9995	0.9997	0.9998	0.9998
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Κεφάλαιο 4

Πίνακες συνεχόμενης πλήρους κάλυψης και έλεγχος τυχαιότητας

Στο Κεφάλαιο 3, υπολογίσαμε (με τη μέθοδο της εμφύτευσης τ.μ., σε Μαρκοβιανή αλυσίδα) την ακριβή κατανομή της τ.μ. $T_{k,n,t}$, η οποία απαριθμούσε το πλήθος των $t \times n$ υποπινάκων μη πλήρους κάλυψης, για την περίπτωση $q = 2$. Επίσης, μελετήθηκε μια εφαρμογή των πινάκων πλήρους κάλυψης, στο πεδίο των παραγοντικών σχεδιασμών. Στο συγκεκριμένο κεφάλαιο θα ασχοληθούμε ξανά με τους πίνακες συνεχόμενης πλήρους κάλυψης. Θα μας απασχολήσει η γενικότερη περίπτωση, όπου τα kn στοιχεία του $k \times n$ πίνακα, θα είναι διακριτές τ.μ. με πεδίο τιμών το αλφάβητο $\mathcal{A} = \{0, 1, \ldots, q-1\}$, με $q \geq 2$. Με τεχνικές παρόμοιες μ' αυτές του προηγούμενου κεφαλαίου (MCET), θα υπολογίσουμε την ακριβή κατανομή της $T_{k,n,t}(q)$, η οποία είναι η αντίστοιχη της απαριθμήτριας $T_{k,n,t}$, για την περίπτωση $q \geq 2$.

Θα γίνει φανερό και μέσα από τα αποτελέσματα που θα ακολουθήσουν, ότι στην περίπτωση που το n παίρνει πολύ μεγάλες τιμές (σε σχέση με το q^t), ο προσδιορισμός των πινάκων μετάβασης (που απαιτούνται για τον ακριβή υπολογισμό της κατανομής της $T_{k,n,t}(q)$), καταλήγει σε μία πολύ απαιτητική διαδικασία (έως, πρακτικά μη εφικτή). Οπότε, σε τέτοιες περιπτώσεις η ανάγκη για εύρεση ποιοτικών προσεγγίσεων, για την υπό μελέτη κατανομή, προβάλει επιτακτική. Ως εκ τούτου, θα επιδιώξουμε μέσω της μεθόδου Chen-Stein, να προσεγγίσουμε την κατανομή της $T_{k,n,t}(q)$, από μια κατάλληλα ορισμένη κατανομή Poisson, δίνοντας ταυτόχρονα και ένα άνω φράγμα για το σφάλμα της προσέγγισής μας. Να σημειώσουμε ότι στη μέθοδο Chen-Stein και τη συνεισφορά της στην προσέγγιση των κατανομών, που σχετίζονται με τις συναρτήσεις σάρωσης, είχαμε αναφερθεί και στο Κεφάλαιο 2, της παρούσης διατριβής.

Η μέθοδος Chen-Stein έχει ήδη χρησιμοποιηθεί σε παρόμοια προβλήματα, όπως στην

εργασία των Godbole et al (1996), όπου προσεγγίζεται η αντίστοιχη απαριθμήτρια τ.μ., για την περίπτωση των t-CA (όταν τα στοιχεία του πίνακα είναι ανεξάρτητες και ισόνομες τ.μ., οι οποίες ακολουθούν την ομοιόμορφη διακριτή κατανομή στο $\mathcal{A}$). Επίσης, την παραπάνω μέθοδο τη συναντάμε και στην εργασία των Godbole and Janson (1996), όπου εξετάζεται ένα πρόβλημα από το χώρο της συνδυαστικής, που έχει άμεση σχέση με το t-CA.

Είναι γνωστό ότι πάνω σε μια ακολουθία δοκιμών, ορίζονται διάφορες τ.μ. όπως η συχνότητα εμφάνισης κάθε αποτελέσματος, η συχνότητα εμφάνισης διαφόρων σχηματισμών (patterns), το πλήθος των ροών, το μέγιστο μήκος μίας ροής κτλ (βλ. π.χ. Balakrishnan and Koutras (2002)). Να επισημάνουμε σ' αυτό το σημείο ότι σε μια ακολουθία διακριτών τ.μ., ως ροή ορίζουμε τα συνεχόμενα όμοια αποτελέσματα. Για παράδειγμα, στην ακολουθία 0011101, έχουμε αρχικά μια ροή μήκους 2, από μηδενικά, στη συνέχεια μια ροή μήκους 3, από άσσους, και τέλος, δύο ροές μήκους 1. Έτσι, έχουμε συνολικά 4 ροές και το μήκος της μεγαλύτερης ροής είναι τρία.

Μελετώντας τις εφαρμογές των παραπάνω τ.μ. διαπιστώνουμε, την άμεση σχέση τους με τους ελέγχους τυχαιότητας, οι οποίοι χρησιμοποιούνται για τον έλεγχο της υπόθεσης ότι η ακολουθία σχηματίζεται από i.i.d. τ.μ. (βλ. π.χ. Lehmann (1973), Gibbons and Chakraborti (1992)). Ισοδύναμα, ο έλεγχος τυχαιότητας σε μια ακολουθία δοκιμών Bernoulli, μπορεί να διατυπωθεί και ως εξής: δοθέντος του πλήθους των επιτυχιών, είναι κάθε μετάθεση των παραπάνω αποτελεσμάτων, ισοπίθανη;

Το ενδιαφέρον της ερευνητικής κοινότητας, για τους ελέγχους τυχαιότητας ξεκινάει το 1940, με την εργασία των Wald and Wolfowitz (1940) όπου προτείνεται ένας νέος έλεγχος, για την εξέταση της υπόθεσης κατά πόσο δύο δείγματα, προέρχονται από τον ίδιο πληθυσμό. Η στατιστική συνάρτηση του ελέγχου είναι το συνολικό πλήθος των ροών (σε μια κατάλληλα ορισμένη, δίτιμη ακολουθία αποτελεσμάτων), για την οποία υπολογίστηκε η δεσμευμένη της κατανομή, δοθέντος, του πλήθους των επιτυχιών (βλ. επίσης και την εργασία των Swed and Eisenhart (1943), η οποία δίνει συγκεκριμένα αριθμητικά αποτελέσματα, για την παραπάνω κατανομή).

Οι έλεγχοι τυχαιότητας, για μια ακολουθία δίτιμων τ.μ., είναι αυτοί που έχουν τραβήξει τη μεγαλύτερη προσοχή. Στην εργασία του Mosteller (1941), ως κριτήριο τυχαιότητας θεωρήθηκε και μελετήθηκε, το μήκος της μεγαλύτερης ροής (αλλιώς, το μέγιστο μήκος ροής). Παρόμοια αντιμετώπιση, συναντάμε και στις εργασίες Lou (1996, 1997), όπου μελετώνται οι έλεγχοι τυχαιότητας, βασισμένοι στο πλήθος των ροών επιτυχιών και στο μήκος της μεγαλύτερης ροής επιτυχιών.

Οι Koutras and Alexandrou (1997) εξετάζουν τις κατανομές τυχαίων μεταβλητών (σε

μια ακολουθία δοκιμών Bernoulli), όπως το πλήθος των επικαλυπτόμενων ροών (ενός συγκεκριμένου μήκους), το πλήθος των μη επικαλυπτόμενων ροών, το πλήθος των ροών τουλάχιστον κάποιου (συγκεκριμένου) μήκους κ.α. Εφαρμόζουν τα αποτελέσματά τους σε ελέγχους τυχαιότητας, και εξετάζουν τη συμπεριφορά των νέων ελέγχων σε σχέση με τους κλασικούς ελέγχους, στους οποίους έχουμε ήδη αναφερθεί. Επίσης, γίνεται σύγκριση και μ' ένα νέο έλεγχο που εισήγαγαν οι Agin and Godbole (1992), για τον οποίο η εύρεση της κρίσιμης περιοχής (critical region), βασίζεται στο πλήθος των μη επικαλυπτόμενων ροών (συγκεκριμένου μήκους).

Ενδιαφέρον παρουσιάζει και ο έλεγχος τυχαιότητας που προτάθηκε στην εργασία του O' Brien (1976). Ο έλεγχος αυτός βασίζεται στην (δειγματική) τυπική απόκλιση, του μήκους των ροών επιτυχιών, όπου όμως, έχει χρησιμοποιηθεί ένας διαφορετικός τρόπος για τον ορισμό των ροών (σε σχέση μ' αυτό που έχουμε αναφέρει-τον κλασικό ορισμό). Αργότερα, οι O'Brien and Dyck (1985), χρησιμοποιώντας τον κλασικό ορισμό μιας ροής και λαμβάνοντας υπόψη και τις ροές των αποτυχιών, εισήγαγαν ένα έλεγχο βασισμένο στο σταθμισμένο άθροισμα των τυπικών αποκλίσεων, του μήκους των ροών επιτυχιών και αποτυχιών. Οι Larsen et al (1973) μελέτησαν μια διαδικασία με την οποία μπορούμε να εντοπίσουμε αποκλίσεις από την τυχαιότητα, που προκαλούν τη συγκέντρωση των επιτυχιών, σε κάποιο συγκεκριμένο τμήμα της ακολουθίας (π.χ. στο κέντρο της ακολουθίας). Έτσι, θεώρησαν ένα κατάλληλο μέτρο που αποσκοπεί στην ποσοτικοποίηση της απόστασης των εμφανίσεων των επιτυχιών, από ένα συγκεκριμένο σημείο της ακολουθίας, και μελέτησαν την κατανομή του. Ο έλεγχος που εισήγαγαν προβλέπει τη χρησιμοποίηση πολλών ανεξάρτητων ακολουθιών, με σκοπό την εξαγωγή πιο ασφαλών συμπερασμάτων.

Ενώ όλοι οι προηγούμενοι έλεγχοι τυχαιότητας αναφέρονται σε δίτιμες τ.μ., είναι προφανές ότι ερευνητικό ενδιαφέρον παρουσιάζει και η γενικότερη περίπτωση, όπου έχουμε ακολουθίες από διακριτές τ.μ. με πεπερασμένο πεδίο τιμών, έστω, το $\mathcal{A}$ ($|\mathcal{A}| = q \geq 2$). Η διαθέσιμη βιβλιογραφία είναι σαφώς πιο περιορισμένη, και αντιπροσωπευτικές εργασίες, μπορεί να θεωρηθούν αυτές των Shaughnessy (1981) και Rubin et al (1990). Η στατιστική συνάρτηση του ελέγχου, και στις δύο περιπτώσεις είναι το άθροισμα του πλήθους των ροών, καθενός από τα q διαφορετικά αποτελέσματα.

Γενικά, αποτελέσματα που αναφέρονται στη θεωρία ροών, σε μια ακολουθία διακριτών τ.μ. με πεδίο τιμών το $\mathcal{A}$ ($|\mathcal{A}| = q \geq 2$), μπορούμε να βρούμε (μεταξύ άλλων), στις εργασίες Mood (1940) και Schwager (1983). Άμεση σχέση έχουν και οι χρόνοι αναμονής (waiting times) για την εμφάνιση σχηματισμών, σε μια ακολουθία όπως η προηγούμενη (βλ. Fu (1996), Antzoulakos (2001), Fu and Chang (2002) κ.α.), όπως και η θεωρία των

στατιστικών συναρτήσεων σάρωσης, με την οποία ασχοληθήκαμε στο Κεφάλαιο 2.

Ο σκοπός του συγκεκριμένου κεφαλαίου, πέρα από τον ακριβή υπολογισμό και την προσέγγιση της κατανομής της τ.μ. $T_{k,n,t}(q)$, είναι η χρησιμοποίηση των παραπάνω αποτελεσμάτων, για την εισαγωγή και τη μελέτη ενός νέου ελέγχου τυχαιότητας. Ο έλεγχος αυτός θα αφορά τυχαίους πίνακες διάστασης $k \times n$, όπου τα kn στοιχεία τους, θα είναι διακριτές τ.μ. με πεδίο τιμών το αλφάβητο $\mathcal{A}$. Η στατιστική συνάρτηση του τεστ, θα είναι η τ.μ. $T_{k,n,t}(q)$. Θα διαπιστώσουμε στη συνέχεια, πως υπάρχουν αρκετά προβλήματα, που μπορεί να μας οδηγήσουν στη χρήση ενός τέτοιου ελέγχου (π.χ. ζητήματα που προκύπτουν από τους κλασικούς/μονοδιάστατους ελέγχους, ή φυσικά, ζητήματα που αφορούν πολυδιάστατα δεδομένα).

Θα ξεκινήσουμε το συγκεκριμένο κεφάλαιο (Παράγραφο 4.1), με τον υπολογισμό της ακριβούς κατανομής της $T_{k,n,t}(q)$, για $q \geq 2$, με μια μέθοδο, παρόμοια μ' αυτή που χρησιμοποιήσαμε για την περίπτωση της $T_{k,n,t}$ (ουσιαστικά, μπορούμε να μιλήσουμε για τη γενίκευση της τελευταίας μεθόδου). Στην Παράγραφο 4.2, αναδεικνύουμε αρχικώς το πρόβλημα που δημιουργείται όταν το n πάρει μεγάλες τιμές (σε σχέση με το q^t), καθώς σε τέτοιες περιπτώσεις, η διάσταση των πινάκων πιθανοτήτων μετάβασης που απαιτούνται, γίνεται πολύ μεγάλη. Έπειτα, αποδεικνύουμε ένα άνω φράγμα για την απόσταση ολικής κύμανσης μεταξύ της κατανομής της $T_{k,n,t}(q)$ και μιας κατανομής Poisson, με μέση τιμή ίδια με αυτή της $T_{k,n,t}(q)$ (με την αρωγή της μεθόδου Chen-Stein, και όπως αυτή διατυπώνεται από τους Arratia et al (1989, 1990)). Με βάση το φράγμα αυτό, καταλήγουμε στην ασθενή σύγκλιση των δυο κατανομών, για την περίπτωση που το $n, k \to \infty$ (ενώ οι παράμετροι t, q παραμένουν σταθερές). Τέλος, η Παράγραφος 4.3 αφιερώνεται στις ιδιότητες και τα χαρακτηριστικά του νέου ελέγχου τυχαιότητας.

4.1 Η κατανομή του πλήθους των υποπινάκων μη πλήρους κάλυψης, σε ένα τυχαίο πίνακα με στοιχεία διακριτές τυχαίες μεταβλητές

Ας θεωρήσουμε ένα τυχαίο πίνακα $\mathbf{X} = (X_{ij})_{k\times n}$, όπου τα kn στοιχεία του X_{ij}, είναι ανεξάρτητες και ισόνομες τ.μ., οι οποίες ακολουθούν την ομοιόμορφη διακριτή κατανομή στο σύνολο $\mathcal{A} = \{0, 1, \ldots, q-1\}$. Δηλαδή,

$$P(X_{ij} = l) = 1/q, \quad l = 0, 1, \ldots, q-1,$$

για $i = 1, 2, \ldots, k$ και $j = 1, 2, \ldots, n$.

4.1 Η κατανομή του πλήθους των υποπινάκων μη πλήρους κάλυψης, σε ένα τυχαίο πίνακα με στοιχεία διακριτές τυχαίες μεταβλητές

Στο Κεφάλαιο 3 μας απασχόλησε η απαριθμήτρια τ.μ. $T_{k,n,t}$, η οποία είχε οριστεί σ' ένα τυχαίο πίνακα $\mathbf{X}$, με στοιχεία δίτιμες τ.μ. Στην παρούσα παράγραφο θα ασχοληθούμε με τη γενικότερη περίπτωση, όπου $q \geq 2$, και σε αναλογία με τους συμβολισμούς του προηγούμενου κεφαλαίου, μπορούμε να γράψουμε ότι

$$T_{k,n,t}(q) = \sum_{i=1}^{k-t+1} I_i, \tag{4.1.1}$$

όπου (όπως και στο Κεφάλαιο 3)

$$I_i = \begin{cases} 1, & \text{εάν ο υποπίνακας που αποτελείται από τις συνεχόμενες γραμμές} \\ & i, i+1, \ldots, i+t-1, \text{δεν περιέχει και τις } q^t \text{ λέξεις, ως στήλες} \\ 0, & \text{διαφορετικά} \end{cases} \tag{4.1.2}$$

με $i = 1, 2, \ldots, k-t+1$ $(t \leq k)$.

Θα περιγράψουμε στη συνέχεια τον τρόπο με τον οποίο μπορούμε να προχωρήσουμε στον υπολογισμό της ακριβούς κατανομής της $T_{k,n,t}(q)$, μέσω της MCET. Η μέθοδος που θα ακολουθήσουμε είναι όμοια μ' αυτή που περιγράψαμε στο Κεφάλαιο 3. Ο όρος μη πλήρης υποπίνακας ή υποπίνακας μη πλήρους κάλυψης, θα χρησιμοποιείται και στο συγκεκριμένο κεφάλαιο, για να δηλώσουμε τον $t \times n$ υποπίνακα (που αποτελείται από t συνεχόμενες γραμμές) από τον οποίο λείπει (από τις στήλες του) μια τουλάχιστον λέξη μήκους t, από τις q^t πιθανές λέξεις (του αλφάβητου $\mathcal{A}$).

Είναι ξεκάθαρο ότι η $T_{k,n,t}(q)$ απαριθμεί το πλήθος των υποπινάκων μη πλήρους κάλυψης, που υπάρχουν στον $k \times n$ (αρχικό) πίνακα. Πριν προχωρήσουμε στον υπολογισμό της συνάρτησης κατανομής της $T_{k,n,t}(q)$, θα αποδειχθεί χρήσιμο να ασχοληθούμε αρχικώς με τον υπολογισμό της πιθανότητας $P(T_{k,n,t}(q) = 0)$.

Όπως ήδη το κάναμε στο Κεφάλαιο 3, θα θεωρήσουμε και πάλι ότι ο σχηματισμός ενός $k \times n$ πίνακα, ο οποίος θα ικανοποιεί το κριτήριο των t-CCA, πραγματοποιείται σταδιακά. Πιο συγκεκριμένα, η δημιουργία ενός πίνακα θα ξεκινάει με το πρώτο $(t-1) \times n$ τμήμα του, και θα ολοκληρώνεται σταδιακά ύστερα από $k-t+1$ βήματα. Σε κάθε βήμα θα γίνεται η προσθήκη μιας γραμμής, και θα κρατάμε ως πληροφορία τη δομή του τελευταίου $(t-1) \times n$ υποπίνακα (δηλαδή, πόσες φορές εμφανίστηκε η κάθε λέξη). Παράλληλα, θα πηγαίνουμε από το ένα βήμα στο άλλο, μόνο εάν η ιδιότητα που μας ενδιαφέρει, ικανοποιείται. Τότε, θα σχηματίζουμε μία Μαρκοβιανή αλυσίδα, και κοιτώντας στο $(k-t+1)$-οστό βήμα της, θα μπορούμε να υπολογίσουμε την πιθανότητα $P(T_{k,n,t}(q) = 0)$.

Ένας κατάλληλος χώρος καταστάσεων Ω, για την εμφύτευση της $T_{k,n,t}(q)$ είναι ο

$$\Omega = \Omega_4 \cup \{x_{abs}\},$$

όπου το σύνολο Ω_4 περιλαμβάνει τις (ακέραιες) λύσεις της γραμμικής εξίσωσης ($a = q^{t-1}$)

$$x_1 + \ldots + x_a = n, \tag{4.1.3}$$

κάτω από τους περιορισμούς $x_i \geq q$, για κάθε $i = 1, 2, \ldots, a$ (το a είναι ουσιαστικά, το πλήθος των διαφορετικών λέξεων μήκους $t-1$). Δηλαδή,

$$\Omega_4 = \{(x_1, \ldots, x_a) : x_1 + \ldots + x_a = n, x_i \in Z \text{ και } x_i \geq q\}.$$

Ο πληθάριθμος του Ω_4 είναι ίσος με (βλ. και Παράγραφο 3.2.3)

$$|\Omega| = |\Omega_2| + 1 = \binom{n + a(1-q) - 1}{a-1} + 1 = s + 1.$$

Ο ορισμός της Μαρκοβιανής αλυσίδας $\{Y_r, r = 0, 1, \ldots\}$, που θα χρησιμοποιηθεί για τον υπολογισμό της $P(T_{k,n,t}(q) = 0)$, ολοκληρώνεται με τα εξής:

- $Y_r = (x_1, x_2, \ldots, x_a) \in \Omega_4, 1 \leq r \leq k - t + 1$ εάν και μόνο εάν ο αριθμός των εμφανίσεων της $i-$οστής λέξης μήκους $t-1$ (από το αλφάβητο με q γράμματα), στις γραμμές $r+1, r+2, \ldots, r+t-1$ είναι ίσος με x_i (για κάθε $i = 1, 2, \ldots, a$) και ταυτόχρονα, ο υποπίνακας που αποτελείται από τις πρώτες $r + t - 1$ γραμμές του πίνακα X, είναι t-CCA. Σε οποιαδήποτε άλλη περίπτωση, θα θεωρούμε ότι η Μαρκοβιανή αλυσίδα βρίσκεται στην κατάσταση απορρόφησης x_{abs}.

Δεν είναι πλέον δύσκολο να διαπιστώσουμε ότι ο πίνακας πιθανοτήτων μετάβασης, της $\{Y_r, r = 0, 1, \ldots\}$, θα έχει τη μορφη

$$\Lambda = \begin{pmatrix} P & \mathbf{h}' \\ \mathbf{0} & 1 \end{pmatrix}$$

όπου ο P είναι ένας $s \times s$ πίνακας που περιέχει τις πιθανότητες μετάβασης, ανάμεσα στις καταστάσεις του συνόλου Ω_4, δηλαδή, τις πιθανότητες

$$P(Y_r = (x_1, x_2, \ldots, x_n) | Y_{r-1} = (x'_1, x'_2, \ldots, x'_n))$$

με $(x_1, x_2, \ldots, x_n), (x'_1, x'_2, \ldots, x'_n) \in \Omega_4$. Το διάνυσμα $\mathbf{h}$, διαστάσεως $1 \times s$, δίδεται από τη σχέση

$$\mathbf{h}' = \mathbf{1}' - P\mathbf{1}' = (I - P)\mathbf{1}', \;\; \mathbf{1} = (1, 1, \ldots, 1),$$

ενώ $\mathbf{0} = (0, 0, \ldots, 0)_{1 \times s}$. Η τιμή της συνάρτησης πιθανότητας της τ.μ. $T_{k,n,t}(q)$, στο μηδέν, υπολογίζεται μέσω του τύπου

$$P(T_{k,n,t}(q) = 0) = P(Y_{k-t+1} \neq x_{abs}) = \boldsymbol{\pi}_0 P^{k-t+1} \mathbf{1}',$$

όπου το $\boldsymbol{\pi}_0$, είναι το $1 \times s$ διάνυσμα των αρχικών πιθανοτήτων της $\{Y_r, r = 0, 1, \ldots\}$. Συγκεκριμένα, το $\boldsymbol{\pi}_0$ έχει τη μορφή

$$\boldsymbol{\pi}_0 = (p_1, p_2, \ldots, p_s)$$

όπου οι πιθανότητες $p_1, p_2, \ldots, p_s$ αντιστοιχούν στις s λύσεις της εξίσωσης (4.1.3), και εκφράζουν την πιθανότητα να πάρουμε ένα $(t-1) \times n$ πίνακα, με δομή $(x_1, x_2, \ldots, x_a) \in \Omega_4$. Επομένως, τα $p_i, i = 1, 2, \ldots, s$ είναι της μορφής

$$\frac{1}{q^{(t-1)n}} \frac{n!}{x_1! x_2! \cdots x_a!}. \tag{4.1.4}$$

Για να υπολογίσουμε τις πιθανότητες μετάβασης του πίνακα P, ας συμβολίσουμε με

$$w_1, w_2, \ldots, w_a$$

τις $a = q^{t-1}$ διαφορετικές λέξεις μήκους $t-1$ (από ένα αλφάβητο με q γράμματα), σε λεξικογραφική διάταξη. Αρχικώς πρέπει να επισημάνουμε πως ο περιορισμός $x_i \geq q, i = 1, 2, \ldots, a$ (για τον προσδιορισμό των καταστάσεων του χώρου Ω_4) χρησιμοποιήθηκε διότι, κάθε υποπίνακας $(t-1) \times n$ ενός πίνακα $t \times n$ ο οποίος είναι t-CCA, θα πρέπει να περιέχει καθεμία από τις $w_1, w_2, \ldots, w_a$ τουλάχιστον q φορές.

Το επόμενο λήμμα, μας προσφέρει μια ικανή και αναγκαία συνθήκη, ώστε οι πιθανότητες,

$$P(Y_r = (x_1, x_2, \ldots, x_a) | Y_{r-1} = (x'_1, x'_2, \ldots, x'_a))$$

για $(x_1, x_2, \ldots, x_a) \in \Omega_4$ και $(x'_1, x'_2, \ldots, x'_a) \in \Omega_4$ να είναι μη μηδενικές. Το αποτέλεσμα αυτό, αποτελεί γενίκευση του Λήμματος 3.2.1, το οποίο αναφέρεται στο μη μηδενισμό των αντίστοιχων πιθανοτήτων μετάβασης, για την περίπτωση $q = 2$.

Λήμμα 4.1.1 *Μια ικανή και αναγκαία συνθήκη για να ισχύει*

$$P(Y_r = (x_1, x_2, \ldots, x_a) | Y_{r-1} = (x'_1, x'_2, \ldots, x'_a)) \neq 0$$

είναι η εξής

$$x_{q(i-1)+1} + x_{q(i-1)+2} + \ldots + x_{q(i-1)+q} = x'_i + x'_{i+a/q} + \ldots + x'_{i+(q-1)a/q}, \tag{4.1.5}$$

για κάθε $i = 1, 2, \ldots, a/q$.

Απόδειξη. Έχοντας διατάξει τις $a = q^{t-1}$ διαφορετικές λέξεις $w_1, w_2, \ldots, w_a$, μήκους $t-1$, σε λεξικογραφική σειρά, είναι εύκολο να διαπιστώσουμε ότι για κάθε $i = 1, 2, \ldots, a/q$ οι λέξεις

$$w_{q(i-1)+1}, w_{q(i-1)+2}, \ldots, w_{q(i-1)+q},$$

έχουν ακριβώς το ίδιο αρχικό τμήμα μήκους $t-2$. Π.χ. για $q=t=3$, οι $a=3^{3-1}=9$ διαφορετικές λέξεις, μήκους $t-1=2$, σε λεξικογραφική σειρά, είναι οι

$$w_1=\binom{0}{0}, w_2=\binom{0}{1}, w_3=\binom{0}{2}, w_4=\binom{1}{0}, w_5=\binom{1}{1},$$
$$w_6=\binom{1}{2}, w_7=\binom{2}{0}, w_8=\binom{2}{1}, w_9=\binom{2}{2}.$$

Στην περίπτωση αυτή, οι λέξεις w_1, w_2, w_3, έχουν το ίδιο αρχικό τμήμα μήκους $t-2=1$ (και συγκεκριμένα, το αρχικό τους τμήμα είναι ίσο με μηδέν), οι w_4, w_5, w_6, έχουν το ίδιο αρχικό τμήμα μήκους 1 (ίσο με ένα) και οι w_7, w_8, w_9, έχουν το ίδιο αρχικό τμήμα μήκους 1 (ίσο με δύο).

Επομένως, εάν με $w_i(j)$, $j=1,2,\ldots,t-1$ συμβολίσουμε το $i-$οστό γράμμα, της λέξης w_i, θα ισχύει

$$w_{q(i-1)+1}(j)=w_{q(i-1)+2}(j)=\ldots=w_{q(i-1)+q}(j), \quad j=1,2,\ldots,t-2$$

ενώ για το $(t-1)-$οστό γράμμα, θα έχουμε

$$w_{q(i-1)+l+1}(t-1)=l, \quad l=0,1,\ldots,q-1.$$

Παρόμοια, για $i=1,2,\ldots,a/q$, το τελικό τμήμα μήκους $t-2$, των λέξεων

$$w_i, w_{i+a/q}, \ldots, w_{i+(q-1)a/q}$$

ταυτίζεται, ήτοι

$$w_i(j)=w_{i+a/q}(j)=\ldots=w_{i+(q-1)a/q}(j), \quad j=2,3,\ldots,t-1,$$

ενώ

$$w_{i+la/q}(1)=l, \quad l=0,1,\ldots,q-1.$$

Επιπλέον, το αρχικό τμήμα των λέξεων

$$w_{q(i-1)+1}, w_{q(i-1)+2}, \ldots, w_{q(i-1)+q}$$

και το αντίστοιχο τελικό τμήμα των

$$w_i, w_{i+a/q}, \ldots, w_{i+(q-1)a/q}$$

είναι το ίδιο (για κάθε $i=1,2,\ldots,a/q$). Σ' αυτό το σημείο, επιβάλλεται να αναφέρουμε ότι η πιθανότητα μετάβασης

$$P(Y_r=(x_1,x_2,\ldots,x_a)|Y_{r-1}=(x'_1,x'_2,\ldots,x'_a))$$

αναφέρεται στην περίπτωση όπου, από ένα $(t-1)\times n$ πίνακα, στον οποίο το πλήθος των λέξεων w_i,$w_{i+a/q}$,...,$w_{i+(q-1)a/q}$, είναι ισο με

$$x'_i + x'_{i+a/q} + \ldots + x'_{i+(q-1)a/q}$$

πηγαίνουμε σ' ένα άλλο $(t-1)\times n$ πίνακα, αφαιρώντας την πρώτη του γραμμή, και προσθέτοντας μια νέα στο τέλος. Ο νέος πίνακας περιέχει $x_{q(i-1)+1} + x_{q(i-1)+2} + \ldots + x_{q(i-1)+q}$ φορές, τις λέξεις

$$w_{q(i-1)+1}, w_{q(i-1)+2}, \ldots, w_{q(i-1)+q}.$$

Με βάση την ανάλυση που προηγήθηκε προκύπτει ότι, κάτι τέτοιο μπορεί να συμβαίνει εάν και μόνο εάν, οι αριθμοί

$$x_{q(i-1)+1} + x_{q(i-1)+2} + \ldots + x_{q(i-1)+q}$$

και

$$x'_i + x'_{i+a/q} + \ldots + x'_{i+(q-1)a/q},$$

είναι ίσοι (ο.ε.δ.).

■

Είμαστε πλέον σε θέση να υπολογίσουμε τις πιθανότητες μετάβασης της Μαρκοβιανής αλυσίδας, που εισάγαμε προηγουμένως (Godbole et al (2008b)).

Θεώρημα 4.1.1 *Εάν* $(x_1, x_2, \ldots, x_a)$, $(x'_1, x'_2, \ldots, x'_a) \in \Omega_2$ *και*

$$x_{q(i-1)+1} + x_{q(i-1)+2} + \ldots + x_{q(i-1)+q} = x'_i + x'_{i+a/q} + \ldots + x'_{i+(q-1)a/q},$$

για κάθε $i = 1, 2, \ldots, a/q$, *τότε*

$$P(Y_r = (x_1, x_2, \ldots, x_a) | Y_{r-1} = (x'_1, x'_2 \ldots, x'_a)) = \\ = \frac{1}{q^n} \prod_{i=1}^{a/q} (\Sigma_1 \Sigma_2 \ldots \Sigma_{q-1} D_{1i} D_{2i} \cdots D_{q-1,i}) \,. \qquad (4.1.6)$$

Το $l-$*οστό άθροισμα* $(l = 1, 2, \ldots, q-1)$ *εκτείνεται σ' όλες τις (θετικές) ακέραιες λύσεις της γραμμικής εξίσωσης*

$$r_{l1} + r_{l2} \ldots + r_{lq} = x_{q(i-1)+l}, \quad l = 1, 2, \ldots, q-1$$

ενώ οι ποσότητες $D_{li}, l = 1, 2, \ldots, q-1$ *και* $i = 1, 2, \ldots, a/q$, *δίδονται από τον τύπο*

$$D_{li} = D_{li}(r_{l1}, r_{l2}, \ldots, r_{lq}) = \prod_{j=1}^{q} \binom{x'_{i+(j-1)a/q} - \sum_{v=1}^{l-1} r_{vj}}{r_{lj}},$$

με τη σύμβαση

$$\binom{u}{v} = 0, \quad \textit{για } u \leq v \textit{ ή } u \leq 0.$$

Απόδειξη. Όπως έχουμε ήδη αναφέρει, ο νέος $(t-1) \times n$ πίνακας, του οποίου η δομή περιγράφεται από το διάνυσμα $(x_1, x_2, \ldots, x_a)$, προκύπτει από τον $(t-1) \times n$ πίνακα, με τα χαρακτηριστικά $(x'_1, x'_2, \ldots, x'_a)$ (αφαιρώντας απ' αυτόν την πρώτη του γραμμή, και προσθέτοντας μια νέα, στο τέλος). Στη νέα γραμμή, το πλήθος των εμφανίσεων του γράμματος l $(l = 0, 1, \ldots, q-1)$ είναι ίσο με

$$\sum_{i=1}^{a/q} x_{q(i-1)+l+1}.$$

Ας κοιτάξουμε τώρα στις στήλες του νέου $(t-1) \times n$ πίνακα, οι οποίες έχουν το ίδιο αρχικό τμήμα μήκους $t-2$. Τη χρονική στιγμή r (για συγκεκριμένο $i \in \{1, 2, \ldots, a/q\}$), οι λέξεις

$$w_{q(i-1)+1}, w_{q(i-1)+2}, \ldots, w_{q(i-1)+q}$$

εμφανίζονται $x_{q(i-1)+1} + x_{q(i-1)+2} + \ldots + x_{q(i-1)+q}$ φορές, και προέρχονται από τις στήλες του $(t-1) \times n$ πίνακα (της χρονικής στιγμής $r-1$), που περιέχουν τις λέξεις

$$w_i, w_{i+a/q}, \ldots, w_{i+(q-1)a/q}$$

και οι οποίες είναι σε πλήθος $x'_i + x'_{i+a/q} + \ldots + x'_{i+(q-1)a/q}$. Επομένως, αυτό που πρέπει να γίνει ώστε να μην περάσουμε στην κατάσταση απορρόφησης, είναι να κατανείμουμε τα $x_{q(i-1)+l+1}$ γράμματα l $(l = 0, 1, \ldots, q-1)$, στη νέα γραμμή, με τέτοιο τρόπο ώστε, στις στήλες που υπάρχουν οι λέξεις w_i, να υπάρχει τουλάχιστον ένα l, στις στήλες που υπάρχουν οι λέξεις $w_{i+a/q}$, να υπάρχει τουλάχιστον ένα l, κτλ, και στις στήλες που υπάρχουν οι λέξεις $w_{i+(q-1)a/q}$, να υπάρχει και εκεί, τουλάχιστον ένα l.

Η ποσότητα $D_{1i}, i = 1, 2, \ldots, a/q$, με

$$D_{1i} = \prod_{j=1}^{q} \binom{x'_{i+(j-1)a/q}}{r_{1j}}$$

δίνει το πλήθος των τρόπων με τους οποίους μπορούμε να τοποθετήσουμε τα $x_{q(i-1)+1}$ μηδενικά, έτσι ώστε r_{1j} απ' αυτά να έχουν τοποθετηθεί στις στήλες που αντιστοιχούν στη λέξη $w_{i+(j-1)a/q}$, για $j = 1, 2, \ldots, q$. Όμοια, η ποσότητα

$$D_{2i} = \prod_{j=1}^{q} \binom{x'_{i+(j-1)a/q} - r_{1j}}{r_{2j}},$$

δίνει το πλήθος των τρόπων με τους οποίους μπορούμε να τοποθετήσουμε τους $x_{q(i-1)+2}$ άσσους, έτσι ώστε να υπάρχουν r_{21} άσσοι στις $x'_i - r_{11}$ στήλες, που αντιστοιχούν στη λέξη w_i, (να σημειώσουμε ότι στις στήλες που αντιστοιχούν στη λέξη w_i, ήδη έχουν τοποθετηθεί r_{11} μηδενικά), να υπάρχουν r_{22} άσσοι στις $x'_{i+a/q} - r_{12}$ στήλες, που αντιστοιχούν στη λέξη $w_{i+a/q}$, κ.ο.κ. Οι ποσότητες $D_{li}, l = 3, 4, \ldots, q-1$ αναφέρονται στην καταμέτρηση των αντίστοιχων διαδικασιών.

Να σημειώσουμε πως, τοποθετώντας τα γράμματα $0, 1, \ldots, q-2$ (με τη διαδικασία που περιγράψαμε), το γράμμα $q-1$, μπορεί να τοποθετηθεί στις υπόλοιπες (εναπομείναντες) κενές θέσεις, μόνο μ' ένα τρόπο. Έτσι, το γινόμενο $\prod_{l=1}^{q-1} D_{li}$, μας δίνει τους τρόπους με τους οποίους όλη η παραπάνω διαδικασία, μπορεί να ολοκληρωθεί.

Η σύμβαση

$$\binom{u}{v} = 0, \quad \text{για } u \leq v \text{ ή } u \leq 0.$$

εξασφαλίζει την εμφάνιση q διαφορετικών γραμμάτων, σε κάθε μία από ομάδες στηλών, που αντιστοιχούν στις λέξεις w_i, $w_{i+a/q}$, $\ldots$, $w_{i+(q-1)a/q}$. Η απόδειξη θα ολοκληρωθεί, λαμβάνοντας υπόψη ότι όλη η παραπάνω διαδικασία πρέπει να επαναληφθεί, για κάθε $i \in \{1, 2, \ldots, a/q\}$, και ότι τα στοιχεία του πίνακα είναι i.i.d. τ.μ., που ακολουθούν την ομοιόμορφη διακριτή κατανομή.

■

Αξίζει να επισημάνουμε ότι οι παραπάνω πιθανότητες μετάβασης, σχετίζονται άμεσα μ' ένα συγκεκριμένο μοντέλο κάλπης, το οποίο μπορεί να περιγραφεί από τα εξής:

> Ας θεωρήσουμε ένα πληθυσμό με $x'_i + x'_{i+a/q} + \ldots + x'_{i+(q-1)a/q} = s'_i$ αντικείμενα (άτομα). Τα αντικείμενα αυτά χωρίζονται σε q ομάδες, και τα $x'_{i+(l-1)a/q}$ απ' αυτά ανήκουν στην ομάδα h_l, με $l = 1, 2, \ldots, q$. Ας υποθέσουμε επιπλέον πως, παίρνουμε αρχικά ένα δείγμα από $x_{q(i-1)+1}$ αντικείμενα του πληθυσμού, χωρίς επανάθεση. Έπειτα, από τα υπόλοιπα $s'_i - x_{q(i-1)+1}$ αντικείμενα παίρνουμε ακόμη ένα δείγμα, μεγέθους τώρα $x_{q(i-1)+2}$. Συνεχίζουμε τη δειγματοληψία με τον τρόπο αυτό, παίρνοντας $x_{q(i-1)+l}$ αντικείμενα, από τα συνολικά $s'_i - x_{q(i-1)+1} - x_{q(i-1)+2} - \ldots - x_{q(i-1)+l-1}$, κατά την $l-$οστή λήψη δείγματος ($l \geq 2$). Τότε, το άθροισμα
>
> $$\sum_1 \sum_2 \ldots \sum_{q-1} D_{1i} D_{2i} \cdots D_{q-1,i}$$
>
> το οποίο εμφανίζεται στο Θεώρημα 4.1.1 (με τα ίδια όρια και περιορισμούς) εκφράζει τους τρόπους με τους οποίους μπορεί να πραγματοποιηθεί το παραπάνω δειγματοληπτικό σχέδιο, ώστε σε κάθε ένα από τα q διαδοχικά δείγματα, να υπάρχει τουλάχιστον ένα αντικείμενο, από κάθε μία από τις q διαφορετικές ομάδες.

Έτσι, μόλις υπολογίσουμε τις πιθανότητες μετάβασης, μπορούμε να προχωρήσουμε στον προσδιορισμό του πίνακα P, και έπειτα, μέσω της σχέσεως

$$P(T_{k,n,t}(q) = 0) = \boldsymbol{\pi}_0 P^{k-t+1}\mathbf{1}', \quad k \geq t \tag{4.1.7}$$

να πάρουμε την πιθανότητα ένας $k \times n$ τυχαίος πίνακας (με στοιχεία i.i.d. τ.μ., που ακολουθούν την ομοιόμορφη διακριτή κατανομή), να είναι t-CCA. Μέσα από το επόμενο παράδειγμα, θα κατανοήσουμε καλύτερα τον τρόπο με τον οποίο δουλεύει η παραπάνω μέθοδος.

Παράδειγμα 4.1 Έστω ότι ενδιαφερόμαστε για την περίπτωση $k \times n$ πινάκων, με $n = 28$, $t = 3$ και $q = 3$ (η κατανομή των $nk = 28k$ ανεξάρτητων και ισόνομων τ.μ., θα είναι η ομοιόμορφη διακριτή, στο $\{0, 1, 2\}$). Τότε, οι $a = 3^{3-1} = 9$ διαφορετικές λέξεις, μήκους $t - 1 = 2$, σε λεξικογραφική σειρά, είναι οι

$$w_1 = \begin{pmatrix}0\\0\end{pmatrix}, w_2 = \begin{pmatrix}0\\1\end{pmatrix}, w_3 = \begin{pmatrix}0\\2\end{pmatrix}, w_4 = \begin{pmatrix}1\\0\end{pmatrix}, w_5 = \begin{pmatrix}1\\1\end{pmatrix},$$
$$w_6 = \begin{pmatrix}1\\2\end{pmatrix}, w_7 = \begin{pmatrix}2\\0\end{pmatrix}, w_8 = \begin{pmatrix}2\\1\end{pmatrix}, w_9 = \begin{pmatrix}2\\2\end{pmatrix}.$$

Ο χώρος καταστάσεων Ω, είναι της μορφής $\Omega = \Omega_4 \cup \{x_{abs}\}$, όπου το σύνολο Ω_4 με πληθάριθμο

$$|\Omega_4| = \binom{n + a(1-q) - 1}{a - 1} = \binom{28 + 9(1-3) - 1}{9 - 1} = 9 = s,$$

περιλαμβάνει τις ακέραιες λύσεις της εξίσωσης

$$x_1 + x_2 + \ldots + x_9 = 28,$$

κάτω από τους περιορισμούς $x_i \geq 3$, για $i = 1, 2, \ldots, 9$. Συμβολίζοντας με ω_i, $i = 1, 2, \ldots, 9$ τα στοιχεία του Ω_4, δηλαδή,

$$\omega_1 = (3, 3, \ldots, 3, 4), \omega_2 = (3, 3, \ldots, 4, 3), \ldots, \omega_8 = (3, 4, \ldots, 3, 3), \omega_9 = (4, 3, \ldots, 3, 3)$$

και εφαρμόζοντας το Λήμμα 4.1.1, συμπεραίνουμε ότι οι πιθανότητες μετάβασης

$$P(Y_r = (x_1, x_2, \ldots, x_a) | Y_{r-1} = (x'_1, x'_2, \ldots, x'_a))$$

είναι μη μηδενικές, εάν $(x_1, x_2, \ldots, x_a) \in E$ και $(x'_1, x'_2, \ldots, x'_a) \in F$, όπου

a. $E = \{\omega_1, \omega_2, \omega_3\}$ και $F = \{\omega_1, \omega_4, \omega_7\}$, ή

b. $E = \{\omega_4, \omega_5, \omega_6\}$ και $F = \{\omega_2, \omega_5, \omega_8\}$, ή

c. $E = \{\omega_7, \omega_8, \omega_9\}$ και $F = \{\omega_3, \omega_6, \omega_9\}$.

Για να υπολογίσουμε, π.χ., την πιθανότητα μετάβασης από την κατάσταση $\omega_4 = (x'_1, x'_2, \ldots, x'_9) = (3,3,3,3,3,4,3,3,3)$ στην κατάσταση $\omega_1 = (x_1, x_2, \ldots, x_9) = (3,3,3,3,3,3,3,3,3)$, χρειαζόμαστε τα αθροίσματα $\sum_1 \sum_2 D_{1i} D_{2i}$, για $i = 1,2,3$, με

$$D_{1i} = \binom{x'_i}{r_{11}} \binom{x'_{i+3}}{r_{12}} \binom{x'_{i+6}}{r_{13}}, \quad D_{2i} = \binom{x'_i - r_{11}}{r_{21}} \binom{x'_{i+3} - r_{12}}{r_{22}} \binom{x'_{i+6} - r_{13}}{r_{23}}.$$

Τα όρια του εξωτερικού αθροίσματος $\sum_1$, εκτείνονται στις ακέραιες θετικές λύσεις της εξίσωσης,

$$r_{11} + r_{12} + r_{13} = x_{3(i-1)+1},$$

ενώ του εσωτερικού αθροίσματος $\sum_2$, αναφέρονται στις ακέραιες λύσεις της εξίσωσης

$$r_{21} + r_{22} + r_{23} = x_{3(i-1)+2}.$$

Για $i = 1$, παίρνουμε

$$r_{11} + r_{12} + r_{13} = 3, \quad r_{21} + r_{22} + r_{23} = 3$$

και επομένως

$$\begin{aligned} \Sigma_1 \Sigma_2 D_{11} D_{21} &= \binom{x'_1}{1} \binom{x'_4}{1} \binom{x'_7}{1} \binom{x'_1 - 1}{1} \binom{x'_4 - 1}{1} \binom{x'_7 - 1}{1} \\ &= \binom{3}{1} \binom{3}{1} \binom{3}{1} \binom{2}{1} \binom{2}{1} \binom{2}{1} = 216. \end{aligned}$$

Με τον ίδιο τρόπο, για $i = 2$ και $i = 3$, θα πάρουμε ότι

$$\Sigma_1 \Sigma_2 D_{12} D_{22} = 216, \quad \Sigma_1 \Sigma_2 D_{13} D_{23} = 432.$$

Άρα, η αντίστοιχη πιθανότητα μετάβασης γίνεται

$$P(Y_r = \omega_1 | Y_{r-1} = \omega_4) = \frac{1}{3^{28}} \prod_{i=1}^{3} (\Sigma_1 \Sigma_2 D_{1i} D_{2i}) = \frac{1}{3^{28}} 216^2 \times 432.$$

Δουλεύοντας με τον παρόμοιο τρόπο, για όλες τις πιθανότητες μετάβασης, διαπιστώνουμε ότι ο πίνακας P έχει την παρακάτω μορφή

$$P = \frac{512}{3^{28}} \begin{pmatrix} b_1 & b_1 & b_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & b_1 & b_1 & b_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & b_1 & b_1 & b_1 \\ b_1 & b_2 & b_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & b_1 & b_2 & b_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & b_1 & b_2 & b_1 \\ b_1 & b_2 & b_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & b_1 & b_2 & b_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & b_1 & b_2 & b_1 \end{pmatrix}$$

με $b_1 = 39366$ και $b_2 = 59049$. Το αρχικό διάνυσμα πιθανοτήτων $\boldsymbol{\pi}_0$, το οποίο εμφανίζεται στον τύπο (4.1.7), μπορεί εύκολα να υπολογιστεί, καθώς οι συντεταγμένες του είναι της μορφής (4.1.4), και συγκεκριμένα, $\boldsymbol{\pi}_0 = (p_1, p_2, \ldots, p_9)$ με

$$p_i = \frac{1}{3^{56}} \frac{28!}{3!3! \cdots 4!}, \quad i = 1, 2, \ldots, 9.$$

Τελικά, οι πιθανότητες $P(T_{k,28,3}(3) = 0)$, δίδονται μέσω της έκφρασης (βλ. και (4.1.7))

$$P(T_{k,28,3}(3) = 0) = \boldsymbol{\pi}_0 P^{k-2} \mathbf{1}', \quad k \geq 3,$$

και για διάφορες τιμές του k, μπορούμε εύκολα να πάρουμε τον Πίνακα 4.1.1.

Πίνακας 4.1.1: Η $P(T_{k,28,3}(3) = 0)$ για διάφορα k.

k	$P(T_{k,28,3}(3) = 0)$	k	$P(T_{k,28,3}(3) = 0)$
3	$3.820 \cdot 10^{-10}$	7	$2.885 \cdot 10^{-32}$
4	$1.122 \cdot 10^{-15}$	8	$8.517 \cdot 10^{-38}$
5	$3.311 \cdot 10^{-21}$	9	$2.514 \cdot 10^{-43}$
6	$9.772 \cdot 10^{-27}$	10	$7.423 \cdot 10^{-49}$

■

Από το προηγούμενο παράδειγμα μπορούμε να δούμε πως η $P(T_{k,28,3}(3) = 0)$ μειώνεται καθώς το k αυξάνεται. Δεν είναι δύσκολο να αποδείξουμε ότι για οποιαδήποτε n, t, q, ισχύει

$$\lim_{k \to \infty} P(T_{k,n,t}(q) = 0) = 0.$$

Προς τούτο, αρκεί να παρατηρήσουμε πως η κατάσταση απορρόφησης είναι προσβάσιμη από κάθε άλλη κατάσταση, και να επικαλεστούμε τις ιδιότητες τέτοιου είδους Μαρκοβιανών αλυσίδων.

Η μέθοδος που περιγράψαμε, μπορεί να τροποποιηθεί καταλλήλως, ώστε να καλύψει τη γενικότερη περίπτωση, όπου τα στοιχεία του πίνακα X είναι i.i.d. τ.μ., οι οποίες ακολουθούν οποιαδήποτε διακριτή κατανομή (με πεπερασμένο πεδίο τιμών) και όχι μόνο την ομοιόμορφη διακριτή. Έτσι, αν θεωρήσουμε ότι

$$P(X_{ij} = l) = p_l, \quad l = 0, 1, \ldots, q-1, \tag{4.1.8}$$

$(\sum_{l=0}^{q-1} p_l = 1)$ για $i = 1, 2, \ldots, k$ και $j = 1, 2, \ldots, n$, με βάση τα ίδια επιχειρήματα και την ίδια λογική, μπορεί να αποδειχθεί ότι οι πιθανότητες μετάβασης (4.1.6), γίνονται

$$\begin{aligned} &P(Y_{r+1} = (x_1, \ldots, x_a) | Y_r = (x'_1, \ldots, x'_a)) \\ &\quad = \prod_{l=0}^{q-1} p_l^{b_l} \prod_{i=1}^{a/q} (\Sigma_1 \Sigma_2 \ldots \Sigma_{q-1} D_{1i} D_{2i} \cdots D_{q-1,i}) \,, \end{aligned} \tag{4.1.9}$$

όπου

$$b_l = \sum_{j=0}^{a/q-1} x_{l+jq+1}, \quad l = 0, 1, \ldots, q-1,$$

ενώ οι υπόλοιποι συμβολισμοί, είναι όμοιοι μ' αυτούς του Θεωρήματος 4.1.1. Η πιθανότητα του ενδεχομένου $T_{k,n,t}(q) = 0$, μπορεί επίσης να δοθεί από την (4.1.7), όπου ο $s \times s$ πίνακας P θα αναφέρεται στις πιθανότητες μετάβασης (4.1.9) και το $\boldsymbol{\pi}_0$, είναι το διάνυσμα των αρχικών πιθανοτήτων. Συγκεκριμένα, το $\boldsymbol{\pi}_0$ θα έχει τώρα συντεταγμένες της μορφής

$$\frac{n!}{x_1! x_2! \cdots x_a!} \prod_{l=0}^{q-1} p_l^{\sum_{i=1}^{a} x_i |w_{li}|} \tag{4.1.10}$$

όπου τα $|w_{li}|$, δίνουν τον αριθμό των εμφανίσεων τους γράμματος l, στη λέξη w_i, για $i = 1, 2, \ldots, a$ και $l = 0, 1, \ldots, q-1$.

Θα συνεχίσουμε τη μελέτη μας, με τον προσδιορισμό της συνάρτησης πιθανότητας της $T_{k,n,t}(q)$ (όχι μόνο στο μηδέν, άλλα σε οποιαδήποτε σημείο του πεδίου τιμών της). Αυτό που θα κάνουμε είναι να επεκτείνουμε τον χώρο καταστάσεων Ω_4, με τέτοιο τρόπο ώστε εκτός από την πληροφορία που αφορά τη δομή του τελευταίου $(t-1) \times n$ πίνακα (δηλαδή, του πίνακα που βρισκόμαστε τη χρονική στιγμή r), να κρατάμε και την πληροφορία που αναφέρεται στο πλήθος των υποπινάκων μη πλήρους κάλυψης, που έχουν ήδη παρατηρηθεί

έως το χρόνο r. Τότε, η τ.μ. $T_{k,n,t}(q)$ μπορεί να θεωρηθεί ως μια εμφυτεύσιμη Μαρκοβιανή αλυσίδα, διωνυμικού τύπου (βλ. Ορισμό 3.1.2 και Θεώρημα 3.1.3, σελίδα 91).

Επομένως, για την εμφύτευση της $T_{k,n,t}(q)$, σε μια Μαρκοβιανή αλυσίδα, ας θεωρήσουμε εκτός από τις s καταστάσεις του Ω_4, μια επιπλέον κατάσταση ω_{s+1}, η οποία θα δηλώνει την περίπτωση που ο $(t-1)\times n$ πίνακας, στον οποίο βρισκόμαστε, περιέχει (ως στήλη) τουλάχιστον μια λέξη μήκους $t-1$, λιγότερες από q φορές. Κάνοντας χρήση και μιας απαριθμήτριας m, η οποία θα κρατάει την πληροφορία του πλήθους των υποπινάκων μη πλήρους κάλυψης, που έχουμε ήδη συναντήσει (μέχρι τη χρονική στιγμή που βρισκόμαστε), καταλήγουμε στο νέο χώρο καταστάσεων

$$\Omega^* = (\Omega_4 \cup \{\omega_{s+1}\}) \times \{0, 1, \ldots, k-t+1\}$$

με πληθάριθμο

$$|\Omega^*| = (s+1)(k-t+2).$$

Τέλος, η Μαρκοβιανή αλυσίδα $\{Y_r, r = 0, 1, \ldots\}$, στον Ω^*, θα ορίζεται ως εξής:

i. θέτουμε $Y_r = (\omega, m)$, με $\omega = (x_1, x_2, \ldots, x_a) \in \Omega_2$ και $0 \le m \le k-t+1$, εάν και μόνο εάν, στον $(t-1)\times n$ υποπίνακα, του οποίου η τελευταία γραμμή είναι η $r+t-1$, ο αριθμός των εμφανίσεων της λέξης w_i, ισούται με x_i (όπου $x_i \ge q$ για $i = 1, 2, \ldots, a$) και επιπλέον, στον πίνακα που αποτελείται από τις γραμμές $1, 2, \ldots, r+t-1$, υπάρχουν ακριβώς m υποπίνακες μη πλήρους κάλυψης,

ii. θέτουμε $Y_r = (\omega, m)$ με $\omega = \omega_{s+1}$ και $0 \le m \le k-t+1$, εάν και μόνο εάν, στον $(t-1)\times n$ υποπίνακα, του οποίου η τελευταία γραμμή είναι η $r+t-1$, μια τουλάχιστον από τις $a = q^{t-1}$ διαφορετικές λέξεις $w_1, w_2, \ldots, w_a$, έχει εμφανιστεί λιγότερες από q φορές και επιπλέον, στον πίνακα που αποτελείται από τις γραμμές $1, 2, \ldots, r+t-1$, υπάρχουν ακριβώς m υποπίνακες μη πλήρους κάλυψης.

Θεωρώντας την παρακάτω διαμέριση του Ω^*,

$$\Omega^* = \bigcup_{m \ge 0} C_m, \quad C_m = \{(\omega, m) : \omega \in \Omega_2 \cup \{\omega_{s+1}\}\}, m = 0, 1, \ldots, k-t+1,$$

είναι εύκολο να αποδείξουμε ότι η $T_{k,n,t}(q)$ είναι μια MVB τ.μ., για την οποία ο πίνακας πιθανοτήτων μετάβασης $A = (P(Y_r \in C_m | Y_{r-1} \in C_m))$, έχει τη μορφή

$$A = \begin{pmatrix} P & \mathbf{0}' \\ \mathbf{0} & 0 \end{pmatrix}_{(s+1)\times(s+1)}$$

όπου ο P είναι ο πίνακας πιθανοτήτων μετάβασης, που μελετήσαμε προηγουμένως, για την περίπτωση $T_{k,n,t}(q) = 0$. Ο πίνακας B της αλυσίδας

$$B = \left(P(Y_r \in C_{m+1} | Y_{r-1} \in C_m)\right),$$

μπορεί επίσης να γράφει στη μορφή

$$B = \begin{pmatrix} Q & \mathbf{c}' \\ \mathbf{b} & \rho \end{pmatrix}_{(s+1)\times(s+1)}$$

όπου ο πίνακας Q, αναφέρεται στις πιθανότητες μετάβασης

$$P(Y_r = (\omega, m+1) | Y_{r-1} = (\omega', m)),$$

με $\omega = (x_1, x_2, \ldots, x_a) \in \Omega_4$ και $\omega' = (x'_1, x'_2, \ldots, x'_a) \in \Omega_4$.

Επιπλέον, πρέπει να παρατηρήσουμε ότι εάν ικανοποιείται η συνθήκη (4.1.5), τότε η πιθανότητα $P(Y_r = (\omega, m) | Y_{r-1} = (\omega', m))$, η οποία ταυτίζεται με την (4.1.9), έχει ήδη συμπεριληφθεί στον υποπίνακα P, του A. Επομένως, στον πίνακα Q θα υπάρχει η «υπόλοιπη» πιθανότητα, δηλαδή

$$P(Y_r = (\omega, m+1) | Y_{r-1} = (\omega', m)) = \prod_{l=0}^{q-1} p_l^{b_l} \left(\prod_{i=1}^{a/q} d_i - \prod_{i=1}^{a/q} c_i \right)$$

όπου

$$c_i = \sum_1 \sum_2 \cdots \sum_{q-1} D_{1i} D_{2i} \cdots D_{q-1,i}, \quad d_i = \frac{(\sum_{j=1}^{q} x_{q(i-1)+j})!}{\prod_{j=1}^{q} (x_{q(i-1)+j})!},$$

για $i = 1, 2, \ldots, a/q$. Να σημειώσουμε ότι η τελευταία έκφραση αναφέρεται στη γενικότερη περίπτωση όπου τα στοιχεία του πίνακα $\mathbf{X}$ ακολουθούν οποιαδήποτε διακριτή κατανομή (βλ. (4.1.8)). Ο παραπάνω τύπος απλοποιείται ως ένα βαθμό, όταν δουλεύουμε με την περίπτωση που τα στοιχεία του πίνακα $\mathbf{X}$ είναι i.i.d. τ.μ., με ομοιόμορφη διακριτή κατανομή στο $\mathcal{A}$.

Λαμβάνοντας υπόψη ότι ο πίνακας $A + B$ είναι ένας στοχαστικός πίνακας, μπορούμε να πάρουμε άμεσα ότι

$$\mathbf{c}' = \mathbf{1}' - Q\mathbf{1}' - P\mathbf{1}'.$$

Για το διάνυσμα $\mathbf{b}$ μπορούμε να καταλάβουμε ότι περιέχει τις πιθανότητες μετάβασης

$$P(Y_r = (\omega, m+1) | Y_{r-1} = (\omega_{s+1}, m)) \tag{4.1.11}$$

με $\omega = (x_1, x_2, \ldots, x_a) \in \Omega_4$. Ας συμβολίσουμε επιπλέον, με

$$\omega_{s+1,j} = (x'_{1j}, x'_{2j}, \ldots, x'_{aj}),$$

$j = 1,2,\ldots,h$ τις ακέραιες λύσεις της εξίσωσης

$$x'_{1j} + x'_{2j} + \ldots + x'_{aj} = n,$$

με περιορισμούς: $x'_i < q$, για τουλάχιστον ένα $i \in \{1,2,\ldots,a\}$ και

$$x_{q(i-1)+1} + x_{q(i-1)+2} + \ldots + x_{q(i-1)+q} = x'_{ij} + x'_{i+a/q,j} + \ldots + x'_{i+(q-1)a/q,j}$$

για κάθε $i \in \{1,2,\ldots,a/q\}$. Τότε,

$$h = \prod_{i=1}^{a/q} \binom{q+s_i-1}{q-1} - \prod_{i=1}^{a/q} \binom{q+s_i-q^2-1}{q-1},$$

όπου $s_i = x_{q(i-1)+1} + x_{q(i-1)+2} + \ldots + x_{q(i-1)+q}$, $i = 1,2,\ldots,a/q$ και η πιθανότητα μετάβασης (4.1.11) μπορεί να εκφραστεί στη μορφή

$$P(Y_r = (\omega, m+1) | Y_{r-1} = (\omega_{s+1}, m)) = \frac{p_0}{P(Y_0 = (\omega_{s+1}, 0))} \sum_{j=1}^{h} p_{\omega_{s+1,j}},$$

όπου

α. οι πιθανότητες $p_{\omega_{s+1,j}}$ (που αναφέρονται στις λύσεις $\omega_{s+1,j}$) υπολογίζονται μέσω τύπων, όμοιων με τον (4.1.10),

β. η πιθανότητα $P(Y_0 = (\omega_{s+1}, 0))$ δίνεται από τον τύπο

$$P(Y_0 = (\omega_{s+1}, 0)) = 1 - \sum_{i=1}^{s} P(Y_0 = (\omega_i, 0)),$$

όπου οι s αρχικές πιθανότητες $P(Y_0 = (\omega_i, 0))$ ταυτίζονται με τις αντίστοιχες συντεταγμένες του $\boldsymbol{\pi}_0$ (άρα, υπολογίζονται μέσω της (4.1.10)),

γ. και τέλος

$$p_0 = \prod_{l=0}^{q-1} p_l^{b_l} \prod_{i=1}^{a/q} d_i.$$

Αξίζει να αναφέρουμε πως η πιθανότητα $\sum_{j=1}^{h} p_{\omega_{s+1,j}}$, στην περίπτωση που τα στοιχεία του πίνακα ακολουθούν την ομοιόμορφη διακριτή κατανομή, δίδεται εναλλακτικά και μέσω της επόμενης σχέσεως

$$\sum_{j=1}^{h} p_{\omega_{s+1,j}} = \frac{1}{q^{n(t-1)}} \left(\sum_{\mathbf{j}_1 \geq 0} \sum_{\mathbf{j}_2 \geq 0} \cdots \sum_{\mathbf{j}_{a/q} \geq 0} \frac{n!}{\prod_{i=1}^{a/q} j_{i1}! j_{i2}! \cdots j_{i,q-1}! (s_i - j_{i1} - \ldots - j_{i,q-1})!} \right.$$
$$\left. - \sum_{\mathbf{j}_1 \geq q} \sum_{\mathbf{j}_2 \geq q} \cdots \sum_{\mathbf{j}_{a/q} \geq q} \frac{n!}{\prod_{i=1}^{a/q} j_{i1}! j_{i2}! \cdots j_{i,q-1}! (s_i - j_{i1} - \ldots - j_{i,q-1})!} \right)$$

όπου τα παραπάνω αθροίσματα, αντιστοιχούν στα επόμενα πολλαπλά αθροίσματα

$$\sum_{\mathbf{j}_i \geq 0} = \sum_{j_{i1}=0}^{s_i} \sum_{j_{i2}=0}^{s_i - j_{i1}} \cdots \sum_{j_{i,q-1}=0}^{s_i - j_{1i} - \ldots - j_{i,q-2}},$$

$$\sum_{\mathbf{j}_i \geq q} = \sum_{j_{i1}=q}^{s_i - q(q-1)} \sum_{j_{i2}=q}^{s_i - j_{i1} - q(q-2)} \cdots \sum_{j_{i,q-1}=q}^{s_i - j_{1i} - \ldots - j_{i,q-2} - q}$$

για $i = 1, 2, \ldots, a/q$.

Έχοντας πλέον υπολογίσει τα στοιχεία των A, B, η συνάρτηση πιθανότητας της $T_{k,n,t}(q)$ δίδεται μέσω της

$$P(T_{k,n,t}(q) = m) = \mathbf{f}_{k-t+1}(m)\mathbf{1}', \quad k \geq t$$

όπου $\{\mathbf{f}_r(m), 0 \leq r, m \leq k - t + 1\}$, είναι η ακολουθία των διανυσμάτων

$$\mathbf{f}_r(m) = (P(Y_r = (\omega_1, m)), P(Y_r = (\omega_2, m)), \ldots, P(Y_r = (\omega_{s+1}, m)))$$

η οποία ικανοποιεί τις παρακάτω αναδρομικές σχέσεις

$$\begin{aligned} \mathbf{f}_r(0) &= \mathbf{f}_{r-1}(0)A, \\ \mathbf{f}_r(m) &= \mathbf{f}_{r-1}(m)A + \mathbf{f}_{r-1}(m-1)B, \quad 1 \leq m \leq k - t + 1, \end{aligned}$$

με αρχικές συνθήκες (ας θυμηθούμε ότι, $\boldsymbol{\pi}_1 = (\boldsymbol{\pi}_0, 1 - \boldsymbol{\pi}_0\mathbf{1}')$)

$$\mathbf{f}_0(0) = \boldsymbol{\pi}_1 \text{ και } \mathbf{f}_0(m) = \mathbf{0}, \text{ για } m > 0.$$

Παράδειγμα 4.2 Όπως και στο Παράδειγμα 4.1, θα ασχοληθούμε με την περίπτωση $n = 28$, $t = 3$ και $q = 3$, όπου τα $nk = 28k$ στοιχεία του πίνακα είναι ανεξάρτητες και ισόνομες τ.μ., οι οποίες ακολουθούν την ομοιόμορφη διακριτή κατανομή, στο $\{0, 1, 2\}$. Ο χώρος καταστάσεων της Μαρκοβιανής αλυσίδας, είναι ο

$$\Omega^* = (\Omega_4 \cup \{\omega_{10}\}) \times \{0, 1, \ldots, k-2\}$$

(βλ. και Παράδειγμα 4.1, σελ. 140), όπου Ω_4 είναι το σύνολο των ακέραιων λύσεων της εξίσωσης

$$x_1 + x_2 + \ldots + x_9 = 28,$$

κάτω από τους περιορισμούς $x_i \geq 3$, για $i = 1, 2, \ldots, 9$. Επομένως,

$$|\Omega_4| = \binom{n + a(1-q) - 1}{a - 1} = \binom{28 + 9(1-3) - 1}{9 - 1} = 9 = s,$$

και ας συμβολίσουμε με ω_i, $i = 1, 2, \ldots, s$ τα στοιχεία του Ω_4, δηλαδή,

$$\omega_1 = (3, 3, \ldots, 3, 4), \omega_2 = (3, 3, \ldots, 4, 3), \ldots, \omega_8 = (3, 4, \ldots, 3, 3), \omega_9 = (4, 3, \ldots, 3, 3).$$

Ο πίνακας πιθανοτήτων μετάβασης A, προκύπτει άμεσα ότι θα έχει τη μορφή

$$A = \begin{pmatrix} P & \mathbf{0}' \\ \mathbf{0} & 0 \end{pmatrix}_{10\times 10}$$

όπου

$$P = \frac{512}{3^{28}} \begin{pmatrix} b_1 & b_1 & b_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & b_1 & b_1 & b_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & b_1 & b_1 & b_1 \\ b_1 & b_2 & b_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & b_1 & b_2 & b_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & b_1 & b_2 & b_1 \\ b_1 & b_2 & b_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & b_1 & b_2 & b_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & b_1 & b_2 & b_1 \end{pmatrix}$$

με $b_1 = 39366$ και $b_2 = 59049$. Για να υπολογίσουμε τον πίνακα B

$$B = \begin{pmatrix} Q & \mathbf{c}' \\ \mathbf{b} & \rho \end{pmatrix}_{10\times 10}$$

θα ξεκινήσουμε με τις πιθανότητες μετάβασης, που υπάρχουν στον Q, δηλαδή με πιθανότητες της μορφής

$$P(Y_r = (\omega, m+1)|Y_{r-1} = (\omega', m)) = \frac{1}{3^{28}} \left(\prod_{i=1}^{3} d_i - \prod_{i=1}^{3} c_i \right)$$

με $\omega, \omega' \in \Omega_4$ και

$$c_i = \sum_1 \sum_2 D_{1i} D_{2i}, \quad d_i = \frac{(\sum_{j=1}^{3} x_{3(i-1)+j})!}{\prod_{j=1}^{3} (x_{3(i-1)+j})!}.$$

Ουσιαστικά, επειδή οι πιθανότητες $\frac{1}{3^{28}} \prod_{i=1}^{3} c_i$ έχουν ήδη υπολογιστεί, καθώς συμπεριλαμβάνονται στον πίνακα P, αυτό που απομένει είναι να υπολογίσουμε τις ποσότητες $\frac{1}{3^{28}} \prod_{i=1}^{3} d_i$. Έτσι, π.χ. για την περίπτωση

$$P(Y_r = (\omega_1, m+1)|Y_{r-1} = (\omega_4, m))$$

όπου

$$\omega_1 = (x_1, x_2, \ldots, x_9) = (3,3,3,3,3,3,3,3,4),$$

$$\omega_4 = (x'_1, x'_2, \ldots, x'_9) = (3,3,3,3,3,4,3,3,3)$$

παίρνουμε

$$d_1 = \frac{(x_1+x_2+x_3)!}{x_1!x_2!x_3!}, d_2 = \frac{(x_4+x_5+x_6)!}{x_4!x_5!x_6!}, d_3 = \frac{(x_7+x_8+x_9)!}{x_7!x_8!x_9!},$$

οπότε

$$P(Y_r = (\omega_1, m+1)|Y_{r-1} = (\omega_4, m)) = \frac{1}{3^{28}}\frac{9!9!10!}{(3!)^8 4!} - \frac{1}{3^{28}}216^2 \cdot 432 = 0.00051729.$$

Δουλεύοντας με παρόμοιο τρόπο για τις υπόλοιπες πιθανότητες μετάβασης, θα πάρουμε

$$Q = \begin{pmatrix} a_1 & a_1 & a_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & a_1 & a_1 & a_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_1 & a_1 & a_1 \\ a_1 & a_1 & a_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & a_1 & a_1 & a_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_1 & a_1 & a_1 \\ a_1 & a_1 & a_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & a_1 & a_1 & a_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_1 & a_1 & a_1 \end{pmatrix}$$

όπου $a_1 = 0.00051729$. Για το διάνυσμα $\mathbf{c}$, έχουμε

$$\mathbf{c} = \mathbf{1} - \mathbf{1}Q - \mathbf{1}P = (a_2, a_2, \ldots, a_2),$$

με $a_2 = 0.998445$. Τέλος, για τον υπολογισμό του διανύσματος $\mathbf{b}$, πρέπει να χειριστούμε πιθανότητες της μορφής

$$P(Y_r = (\omega, m+1)|Y_{r-1} = (\omega_{10}, m)).$$

Η κατάσταση ω_{10}, αντιστοιχεί σε μία από τις ακέραιες λύσεις της εξίσωσης

$$x'_1 + x'_2 + \ldots + x'_9 = 28,$$

με περιορισμούς: $x'_i < 3$, για τουλάχιστον ένα $i \in \{1, 2, \ldots, 9\}$ και

$$x_{3(i-1)+1} + x_{3(i-1)+2} + x_{3(i-1)+3} = x'_i + x'_{i+3} + x'_{i+6} = s_i$$

για κάθε $i \in \{1,2,3\}$. Π.χ. για $\omega_1 = (x_1, x_2, \ldots, x_9) = (3,3,3,3,3,3,3,3,4)$, έχουμε ότι

$$P(Y_r = (\omega_1, m+1) | Y_{r-1} = (\omega_{10}, m)) = \frac{p_0}{P(Y_0 = (\omega_{s+1}, 0))} \sum_{j=1}^{h} p_{\omega_{10,j}},$$

με

$$\sum_{j=1}^{h} p_{\omega_{10,j}} = \frac{1}{3^{56}} \left(\sum_{j_{11}=0}^{s_1} \sum_{j_{12}=0}^{s_1-j_{11}} \sum_{j_{21}=0}^{s_2} \sum_{j_{22}=0}^{s_2-j_{21}} \sum_{j_{31}=0}^{s_3} \sum_{j_{32}=0}^{s_3-j_{31}} \frac{n!}{\prod_{i=1}^{3} j_{i1}! j_{i2}! (s_i - j_{i1} - j_{i2})!} \right.$$
$$\left. - \sum_{j_{11}=3}^{s_1-6} \sum_{j_{12}=3}^{s_1-j_{11}-3} \sum_{j_{21}=3}^{s_2-6} \sum_{j_{22}=3}^{s_2-j_{21}-3} \sum_{j_{31}=3}^{s_3-6} \sum_{j_{32}=3}^{s_3-j_{31}-3} \frac{n!}{\prod_{i=1}^{3} j_{i1}! j_{i2}! (s_i - j_{i1} - j_{i2})!} \right),$$

και $s_1 = s_2 = 9, s_3 = 10$. Άρα,

$$\sum_{j=1}^{h} p_{\omega_{10,j}} = 0.0278472,$$

ενώ

$$p_0 = \frac{1}{3^{28}} \frac{9!9!10!}{(3!)^8 4!}$$

και $P(Y_0 = (\omega_{10}, 0)) = 1 - \boldsymbol{\pi}_0 \mathbf{1}'$, όπου $\boldsymbol{\pi}_0 = (p_1, p_2, \ldots, p_s)$ με

$$p_i = \frac{1}{3^{56}} \frac{28!}{3!3! \cdots 4!}, \quad i = 1, 2, \ldots, 9.$$

Με τον τρόπο αυτό καταλήγουμε στο διάνυσμα

$$\mathbf{b} = (\zeta, \zeta, \zeta, \zeta, \zeta, \zeta, \zeta, \zeta, \zeta)$$

με $\zeta = 0.144315 \cdot 10^{-4}$. Με βάση όσα προηγήθηκαν, μπορούμε πλέον να υπολογίσουμε την ακριβή κατανομή της $T_{k,28,3}(3)$. Για παράδειγμα, στην ειδική περίπτωση $k = 3$, θα πάρουμε

$$P(T_{3,28,3}(3) = 0) = 3.43787 \cdot 10^{-10}, P(T_{3,28,3}(3) = 1) = 1 - 3.43787 \cdot 10^{-10},$$

ενώ για $k = 4$, έχουμε

$$P(T_{4,28,3}(3) = 0) = 9.08671 \cdot 10^{-16}, P(T_{4,28,3}(3) = 1) = 6.87572 \cdot 10^{-10},$$
$$P(T_{4,28,3}(3) = 2) = 1 - 9.08671 \cdot 10^{-15} - 6.87572 \cdot 10^{-10}.$$

■

Πριν κλείσουμε τη συγκεκριμένη παράγραφο, είναι χρήσιμο να αναφέρουμε πως τα αποτελέσματα που προηγήθηκαν, μπορούν εύκολα να τροποποιηθούν ώστε να αντιμετωπίσουμε και την περίπτωση που έχουμε Μαρκοβιανή εξάρτηση ανάμεσα στις γραμμές του πίνακα. Ο τρόπος που μπορεί να γίνει κάτι τέτοιο, είναι ανάλογος μ' αυτόν που περιγράψαμε στην Παράγραφο 3.4.

4.2 Προσέγγιση της κατανομής του πλήθους των υποπινάκων μη πλήρους κάλυψης, μέσω της κατανομής Poisson

Στην προηγούμενη παράγραφο, περιγράψαμε μια μέθοδο η οποία μας επιτρέπει να υπολογίσουμε την ακριβή τιμή της συνάρτησης πυκνότητας της τ.μ. $T_{k,n,t}(q)$, μέσα από τεχνικές εμφύτευσης τ.μ. σε Μαρκοβιανή αλυσίδα. Όμως, όταν η παράμετρος n πάρει πολύ μεγαλύτερες τιμές από την q^t, ο ακριβής υπολογισμός γίνεται αρκετά δύσκολος, καθώς τότε οι διαστάσεις των πινάκων μετάβασης A και B, γίνονται πολύ μεγάλες (το k δεν επηρεάζει τη διάσταση των πινάκων, πάρα μόνο το πλήθος των πολλαπλασιασμών, μεταξύ πινάκων, που πρέπει να γίνουν). Για παράδειγμα, εάν $q = 3, t = 4$ τότε το n θα πρέπει να είναι $n \geq q^t = 81$ και οι πίνακες A, B θα έχουν διάσταση $(s+1) \times (s+1)$, όπου

$$s = \binom{n + q^{t-1}(1-q) - 1}{q^{t-1} - 1} = \binom{n-55}{26}, \quad n \geq 81.$$

Έτσι, εάν το n πάρει τιμές αρκετά μεγαλύτερες από το $q^t = 81$, τότε το s γίνεται πολύ μεγάλο (π.χ., έχουμε $s = 27, 378, 3654, 27405$, για $n = 82, 83, 84, 85$, αντιστοίχως) και ο ακριβής υπολογισμός της κατανομής της $T_{k,n,t}(q)$, γίνεται πρακτικά αδύνατος.

Άρα είναι απαραίτητο, και λόγω της φύσεως της εφαρμογής που θα μελετήσουμε στη συνέχεια (έλεγχος τυχαιότητας, ο οποίος εφαρμόζεται συνήθως σ' ένα μεγάλο πλήθος αποτελεσμάτων), να αναζητήσουμε ποιοτικές προσεγγίσεις για την κατανομή της $T_{k,n,t}(q)$. Ασφαλώς, θα ήταν πολύ χρήσιμο εάν ταυτόχρονα, υπάρχει δυνατότητα εκτίμησης για τα μεγέθη των σφαλμάτων που προκύπτουν, από τις παραπάνω προσεγγίσεις. Μια μέθοδος η οποία μπορεί να μας βοηθήσει, και για την οποία έχουμε μιλήσει στο Κεφάλαιο 2 (για τη συμβολή της, στις προσεγγίσεις των στατιστικών συναρτήσεων σάρωσης) είναι η Chen-Stein (βλ. και Παράγραφο 2.1).

Θα θεωρήσουμε για τη συνέχεια ότι τα στοιχεία του πίνακα X, είναι i.i.d. τ.μ., με τη συνάρτηση πιθανότητάς τους να δίδεται από την (4.1.8). Από τη μορφή που έχει η $T_{k,n,t}(q)$ (βλ. (4.1.1)), είναι εύκολο να διαπιστώσει κανείς ότι μια κατάλληλη επιλογή για το σύνολο Γ που εμφανίζεται στο Θεώρημα 2.1.1, είναι η

$$\Gamma = \{1, 2, \ldots, k - t + 1\}.$$

Επομένως, η παράμετρος λ της προσεγγιστικής κατανομής Poisson, γίνεται

$$\lambda_{k,n} = E(T_{k,n,t}(q)) = \sum_{i=1}^{k-t+1} \pi_i = (k - t + 1)\pi, \tag{4.2.1}$$

όπου π είναι η κοινή μέση τιμή των τ.μ. $I_i, i \in \Gamma$, που δίδονται από την (4.1.2), δηλαδή

$$E(I_i) = P(I_i = 1) = \pi_i = \pi, \quad i = 1, 2, \ldots, k-t+1.$$

Επίσης για τη συνέχεια θεωρούμε ότι, $I_i = 0$ για $i < 1$ ή $i > k-t+1$.

Πριν προχωρήσουμε στη διατύπωση του κύριου αποτελέσματος της συγκεκριμένης παραγράφου, θα αναφέρουμε κάποιους επιπλέον (απαραίτητους) συμβολισμούς. Με Γ_i θα συμβολίζουμε το υποσύνολο (δεικτών) του Γ, για το οποίο ισχύει

$$\Gamma_i = \Gamma \cap \{z \in Z : i-t+1 \leq z \leq i+t-1\} \backslash \{i\}, \tag{4.2.2}$$

ενώ με A_j, $j = 0, 1, \ldots, k-t$, το ενδεχόμενο ότι λείπουν ακριβώς j λέξεις, από ένα $t \times n$ πίνακα, που σχηματίζεται από t συνεχόμενες γραμμές. Η πιθανότητα του A_j, μπορεί να δοθεί μέσω του τύπου

$$P(A_j) = \sum_{v=1}^{aq-j} (-1)^{a-j-v} \binom{a-v}{j} \sum (p_{i_1}(1,t) + p_{i_2}(1,t) + \ldots + p_{i_v}(1,t))^n,$$
$$j = 0, 1, \ldots, k-t$$

όπου το εσωτερικό άθροισμα, κινείται σε όλους τους συνδυασμούς των v στοιχείων $\{i_1, i_2, \ldots, i_v\}$, του συνόλου $\{1,2,\ldots,aq\}$ και

$$p_i(x,y) = \prod_{j=x}^{y} p_{w_i(j)}, \quad i = 1, 2, \ldots, aq, \quad y, x \in \{1, 2, \ldots, t\}, y \geq x. \tag{4.2.3}$$

Άμεσα προκύπτει ότι $\pi = 1 - P(A_0)$, ενώ $\sum_{j=1}^{aq} p_j(1,t) = 1$.

Αποδεικνύουμε στη συνέχεια, το κύριο αποτέλεσμα της παρούσης παραγράφου.

Θεώρημα 4.2.1 *Η απόσταση ολικής κύμανσης μεταξύ της κατανομής $\mathcal{L}(T_{k,n,t}(q))$ της $T_{k,n,t}(q)$ και της κατανομής Poisson, με μέση τιμή $\lambda_{k,n}$ ($Po(\lambda_{k,n})$), φράσσεται ως εξής*

$$d_{TV}(\mathcal{L}(T_{k,n,t}(q)), Po(\lambda_{k,n})) \leq (1 - e^{-(k-t+1)\pi}) \left(\left(\frac{1}{k-t+1} \sum_{i=1}^{k-t+1} |\Gamma_i| + 1 \right) \pi \right.$$
$$\left. + \frac{1}{k-t+1} \sum_{i=1}^{k-t+1} \sum_{r=2}^{t} (I_{[i-r \geq 0]} + I_{[i+r \leq k-t+2]}) u(r) \right) = UB_{TV}$$

όπου

$$u(r) = UB_r \frac{P(A_1)}{1 - P(A_0)} + \frac{1 - P(A_0) - P(A_1)}{1 - P(A_0)}$$

και

$$UB_r = \sum_{j=1}^{aq} (1 - p_j(t-r+2, t))^{q^{r-1}} (1 - p_j(1,t))^{n-a+1}.$$

Απόδειξη. Λόγω της συγκεκριμένης επιλογής των συνόλων Γ_i, η ποσότητα b_3 που εμφανίζεται στο Θεώρημα 2.1.1, μηδενίζεται και η αντίστοιχη ανισότητα, γίνεται

$$d_{TV}(\mathcal{L}(T_{k,n,t}(q)), Po(\lambda_n)) \leq \frac{1-e^{-\lambda_{k,n}}}{\lambda_{k,n}} \left(\sum_{i=1}^{k-t+1} \pi_i^2 + \sum_{i=1}^{k-t+1} \sum_{j \in \Gamma_i} (E(I_i I_j) + \pi_i \pi_j) \right)$$
$$= (1-e^{-(k-t+1)\pi}) UB$$

όπου

$$UB = \pi + \frac{1}{k-t+1} \sum_{i=1}^{k-t+1} \sum_{j \in \Gamma_i} (P(I_j = 1 | I_i = 1) + \pi).$$

Επίσης,

$$UB = \pi + \frac{1}{k-t+1} \sum_{i=1}^{k-t+1} \sum_{j \in \Gamma_i} \pi + \frac{1}{k-t+1} \sum_{i=1}^{k-t+1} \sum_{j \in \Gamma_i} P(I_j = 1 | I_i = 1)$$
$$= UB^* + \frac{1}{k-t+1} \sum_{i=1}^{k-t+1} \sum_{j \in \Gamma_i} P(I_j = 1 | I_i = 1) \tag{4.2.4}$$

όπου

$$UB^* = \left(\frac{1}{k-t+1} \sum_{i=1}^{k-t+1} |\Gamma_i| + 1 \right) \pi,$$

και το γεγονός ότι ισχύει $I_i = 0$ για $i < 1$ ή $i > k-t+1$, μας βοηθάει στο να γράψουμε το τελευταίο άθροισμα της (4.2.4), στη μορφή

$$\sum_{j \in \Gamma_i} P(I_j = 1 | I_i = 1) = \sum_{r=1}^{t-1} (P(I_{i-r} = 1 | I_i = 1) + P(I_{i+r} = 1 | I_i = 1)).$$

Αντικαθιστώντας την παραπάνω σχέση στον τύπο του UB, παίρνουμε

$$UB = UB^* + \frac{1}{k-t+1} \sum_{i=1}^{k-t+1} \sum_{r=1}^{t-1} (P(I_{i-r} = 1 | I_i = 1) + P(I_{i+r} = 1 | I_i = 1)).$$

Προφανώς, ισχύει ότι εάν

$$P(I_{i-r} = 1 | I_i = 1) \neq 0 \text{ και } P(I_{i+r} = 1 | I_i = 1) \neq 0$$

τότε

$$P(I_{i-r} = 1 | I_i = 1) = P(I_{i+r} = 1 | I_i = 1)$$

για κάθε $r \in \{1, 2, \ldots, t-1\}$ και $i \in \{1, 2, \ldots, k-t+1\}$. Όμοια, εάν

$$P(I_{i_1-r} = 1 | I_{i_1} = 1) \neq 0 \text{ και } P(I_{i_2-r} = 1 | I_{i_2} = 1) \neq 0$$

τότε

$$P(I_{i_1-r}=1|I_{i_1}=1)=P(I_{i_2-r}=1|I_{i_2}=1)$$

για κάθε $i_1,i_2\in\{1,2,\ldots,k-t+1\}$ και $i_1\neq i_2$. Επομένως, το UB μπορεί να γραφεί ως εξής:

$$\begin{aligned}UB&=UB^*+\frac{1}{k-t+1}\sum_{i=1}^{k-t+1}\sum_{r=1}^{t-1}(I_{[i-r\geq 1]}+I_{[i+r\leq k-t+1]})P(I_{1+r}=1|I_1=1)\\&=UB^*+\frac{1}{(k-t+1)}\sum_{i=1}^{k-t+1}\sum_{r=2}^{t}(I_{[i-r\geq 0]}+I_{[i+r\leq k-t+2]})P(I_r=1|I_1=1).\end{aligned}\tag{4.2.5}$$

Συνεπώς, για τον ακριβή υπολογισμό του άνω φράγματος, το μόνο που θα πρέπει να μας απασχολήσει, είναι οι δεσμευμένες πιθανότητες $P(I_r=1|I_1=1)$. Έτσι, εφαρμόζοντας μία μέθοδο για τον υπολογισμό των προηγούμενων πιθανοτήτων, παρόμοια μ' αυτή των Godbole et al (1996) (η οποία απευθυνόταν στο t-CA πρόβλημα για την περίπτωση της ομοιόμορφης διακριτής κατανομής), παίρνουμε

$$\begin{aligned}P(I_r=1|I_1=1)&=P(I_r=1|A_1\cup A_2\cup\ldots A_{aq-1})\\&=\frac{P(I_r=1,A_1)}{1-P(A_0)}+\frac{\sum_{j=2}^{aq-1}P(I_r=1,A_j)}{1-P(A_0)}\\&=P(I_r=1|A_1)\frac{P(A_1)}{1-P(A_0)}+\frac{\sum_{j=2}^{aq-1}P(I_r=1,A_j)}{1-P(A_0)}.\end{aligned}\tag{4.2.6}$$

Ας συμβολίσουμε τώρα, με B_j το ενδεχόμενο ότι η λέξη w_j (μήκους t) δεν υπάρχει, στον $r-$οστό υποπίνακα που σχηματίζεται από t συνεχόμενες γραμμές ($j=1,2,\ldots,q^t$). Τότε, η πιθανότητα $P(I_r=1|A_1)$ φράσσεται με τη σειρά της (κάνοντας χρήση των ανισοτήτων Bonferroni), ως εξής

$$P(I_r=1|A_1)=P(B_1\cup\ldots\cup B_{aq}|A_1)\leq\sum_{j=1}^{aq}P(B_j|A_1).\tag{4.2.7}$$

Στη συνέχεια, θα ασχοληθούμε με τις πιθανότητες $P(B_j|A_1)$, οι οποίες όπως θα δούμε, εκφράζονται μέσω των ποσοτήτων $p_i(x,y)$, που εισάγαμε στην (4.2.3). Παρατηρούμε αρχικώς ότι, ο πρώτος και ο $r-$οστός υποπίνακας, έχουν ακριβώς $t-r+1$ κοινές γραμμές. Επιπλέον, ισχύει

$$P(B_j|A_1)=\sum_{c=0}^{n}P(B_j|Z_j=c,A_1)P(Z_j=c|A_1)$$

όπου Z_j είναι μια απαριθμήτρια τ.μ., η οποία μετράει το πλήθος των εμφανίσεων του αρχικού τμήματος της λέξης w_j, μήκους $t-r+1$, στις $t-r+1$ κοινές γραμμές του πρώτου και

του $r-$οστού υποπίνακα. Δοθέντος ότι μόνο μια λέξη δεν υπάρχει στον πρώτο υποπίνακα, αναμένουμε ότι:

- $q^{t-r+1}-1$ λέξεις μήκους $t-r+1$, να εμφανίζονται τουλάχιστον q^{r-1} φορές, στο κοινό κομμάτι των $t-r+1$ γραμμών, και
- μια λέξη μήκους $t-r+1$, να εμφανίζεται τουλάχιστον $q^{r-1}-1$ φορές, στο κοινό κομμάτι των $t-r+1$ γραμμών (αυτή η λέξη είναι στην πραγματικότητα, ίδια με το τελευταίο τμήμα μήκους $t-r+1$, της λέξης που δεν υπάρχει).

Επομένως, η πιθανότητα $P(Z_j=c|A_1)$ είναι μη μηδενική, εάν και μόνο εάν,

$$q^{r-1}-1\leq c\leq n-q^t+q^{r-1},$$

για την τελευταία περίπτωση, και

$$q^{r-1}\leq c\leq n-q^t+q^{r-1}+1$$

για κάθε μία από τις $q^{t-r+1}-1$ λέξεις, της πρώτης περίπτωσης.

Όταν το τελευταίο τμήμα μήκους $t-r+1$, της λέξης που δεν υπάρχει, είναι το ίδιο με το αντίστοιχο αρχικό τμήμα της λέξης w_j, παίρνουμε $c=q^{r-1}-1$, και η $P(B_j|A_1)$ γίνεται

$$\begin{aligned}P(B_j|A_1)&=\sum_{b=0}^{n-aq+1}P(B_j|Z_j=c+b,A_1)P(Z_j=c+b|A_1)\\&=\sum_{b=0}^{n-aq+1}(1-p_j(t-r+2,t))^{c+b}P(Z_j=c+b|A_1).\end{aligned}$$

Χρησιμοποιώντας τέλος την έκφραση

$$P(Z_j=c+b|A_1)=\binom{n-aq+1}{b}p_j(1,t-r+1)^b(1-p_j(1,t-r+1))^{n-aq+1-b}$$

καταλήγουμε στη σχέση

$$P(B_j|A_1)=(1-p_j(t-r+2,t))^{q^{r-1}-1}(1-p_j(1,t))^{n-aq+1}.$$

Όμοια με πριν, εάν το τελευταίο τμήμα μήκους $t-r+1$ της λέξης που δεν υπάρχει, δεν είναι το ίδιο με το αντίστοιχο αρχικό τμήμα της λέξης w_j, παίρνουμε

$$P(B_j|A_1)=(1-p_j(t-r+2,t))^{q^{r-1}}(1-p_j(1,t))^{n-aq+1}.$$

Άρα, η σχέση (4.2.7) μας οδηγεί στο άνω φράγμα

$$P(I_r = 1|A_1) \leq \sum_{j=1}^{aq}(1 - p_j(t - r + 2, t))^{q^{r-1}}(1 - p_j(1, t))^{n-aq+1} = UB_r$$

και η απόδειξη ολοκληρώνεται, συνδυάζοντας τις (4.2.4),(4.2.5),(4.2.6) και (4.2.7). ■

Με το επόμενο αποτέλεσμα, εξασφαλίζουμε την ασθενή σύγκλιση, της $T_{k,n,t}(q)$ σε μία κατάλληλα ορισμένη κατανομή Poisson.

Πόρισμα 4.2.1 *Έστω ότι $n, k \to \infty$, έτσι ώστε*

$$\lambda_{k,n} \to \lambda \in (0, \infty),$$

ενώ οι παράμετροι q, t και p_l (για $l = 0, 1, \ldots, q-1$), παραμένουν σταθερές. Τότε,

$$d_{TV}(\mathcal{L}(T_{k,n,t}(q)), Po(\lambda_{k,n})) \to 0$$

και επομένως η συνάρτηση κατανομής της $T_{k,n,t}(q)$ συγκλίνει (ασθενώς) στην κατανομή Poisson με παράμετρο λ.

Απόδειξη. Από τον τρόπο ορισμού των συνόλων Γ_i,

$$\Gamma_i = \Gamma \cap \{z \in Z : i - t + 1 \leq z \leq i + t - 1\} \backslash \{i\}$$

προκύπτει άμεσα ότι $|\Gamma_i| \leq 2(t-1)$. Κάνοντας χρήση των ανισοτήτων Bonferroni, αποδεικνύεται άμεσα ότι

$$\pi \leq \sum_{j=1}^{aq}(1 - p_j(1, t))^n$$

ενώ επιπλέον, ισχύει

$$UB_r \leq \sum_{j=1}^{aq}(1 - p_j(1, t))^{n-aq+1}.$$

Επίσης, εύκολα μπορούμε να αποδείξουμε ότι

$$1 - P(A_0) - P(A_1) \leq \sum_{i,j=1, i\neq j}^{aq}(1 - p_j(1, t) - p_i(1, t))^n$$

$$1 - P(A_0) \geq (1 - p_{min}(1, t))^{n-aq+1}\frac{\prod_{j=1}^{aq} p_j(1, t)}{p_{min}(1, t)}$$

όπου

$$p_{min}(1,t) = \min\{p_j(1,t) : j = 1, 2, \ldots, aq\}$$

και να καταλήξουμε στο παρακάτω άνω φράγμα, για το UB_{TV} του Θεωρήματος 4.2.1

$$UB_{TV} \leq (2t-1)\sum_{j=1}^{aq}(1-p_j(1,t))^n + 2(t-1)\left(\sum_{j=1}^{aq}(1-p_j(1,t))^{n-aq+1} + \frac{\sum_{i,j=1,i\neq j}^{aq}(1-p_j(1,t)-p_i(1,t))^n}{(1-p_{min}(1,t))^{n-aq+1}}\frac{p_{min}(1,t)}{\prod_{j=1}^{aq}p_j(1,t)}\right).$$

Η απόδειξη του πορίσματος, ολοκληρώνεται με την παρατήρηση ότι

$$\frac{\sum_{i,j=1,i\neq j}^{aq}(1-p_j(1,t)-p_i(1,t))^n}{(1-p_{min}(1,t))^{n-aq+1}} \to 0$$

καθώς $n \to \infty$.

■

4.2.1 Αριθμητικά αποτελέσματα για την προσέγγιση, μέσω κατανομής Poisson

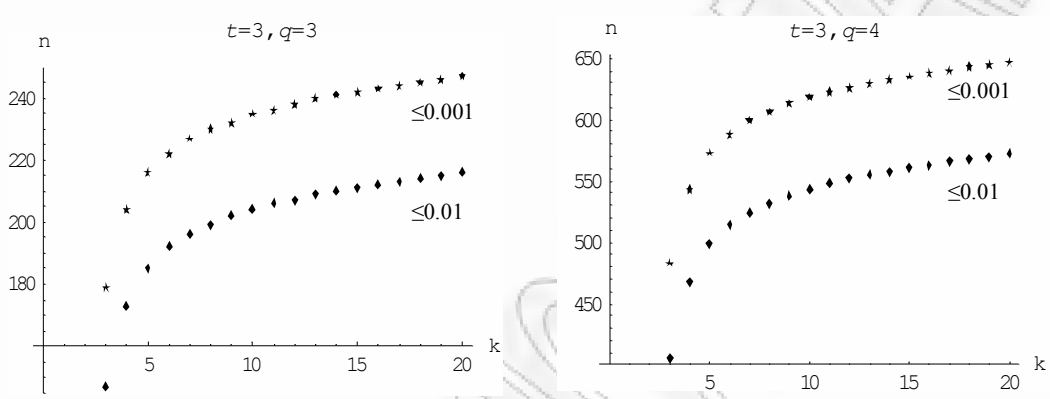
Στην παράγραφο αυτή, θα δώσουμε μερικά αριθμητικά αποτελέσματα, με τα οποία θα αξιολογήσουμε την ποιότητα της προσέγγισης που προκύπτει από το Θεώρημα 4.2.1. Πιο συγκεκριμένα, θα εξετάσουμε την περίπτωση που τα στοιχεία του πίνακα, ακολουθούν την ομοιόμορφη διακριτή κατανομή.

Πίνακας 4.2.1: Οι τιμές του UB_{TV}, για $t = q = 2$.

k	$n=5$	$n=10$	$n=20$	$n=30$	$n=40$	$n=50$
5	3.0834	0.5146	0.0024	$7.846 \cdot 10^{-6}$	$2.484 \cdot 10^{-8}$	$7.871 \cdot 10^{-11}$
10	3.6882	0.8640	0.0060	0.00002	$6.354 \cdot 10^{-8}$	$2.013 \cdot 10^{-10}$
15	3.8225	0.9902	0.0094	0.00003	$1.022 \cdot 10^{-7}$	$3.240 \cdot 10^{-10}$
20	3.8845	1.0386	0.0126	0.00004	$1.409 \cdot 10^{-7}$	$4.466 \cdot 10^{-10}$

Ο Πίνακας 4.2.1 δίνει τις τιμές του άνω φράγματος, για την περίπτωση $t = q = 2$ και για διάφορες τιμές των k, n. Για παράδειγμα, όταν το $n = 20$ και $k = 10$, το UB_{TV} είναι ίσο με 0.0060, ενώ όταν το n γίνει 30, είναι $UB_{TV} = 0.00002$. Επιπλέον, από τον προηγούμενο πίνακα παρατηρούμε ότι καθώς το n μεγαλώνει, το φράγμα γίνεται ολοένα και μικρότερο, για κάθε k (π.χ. για $k = 15$ το UB_{TV} παίρνει τις τιμές $3.8225, 0.9902, 0.0094$, για $n = 5, 10, 20$,

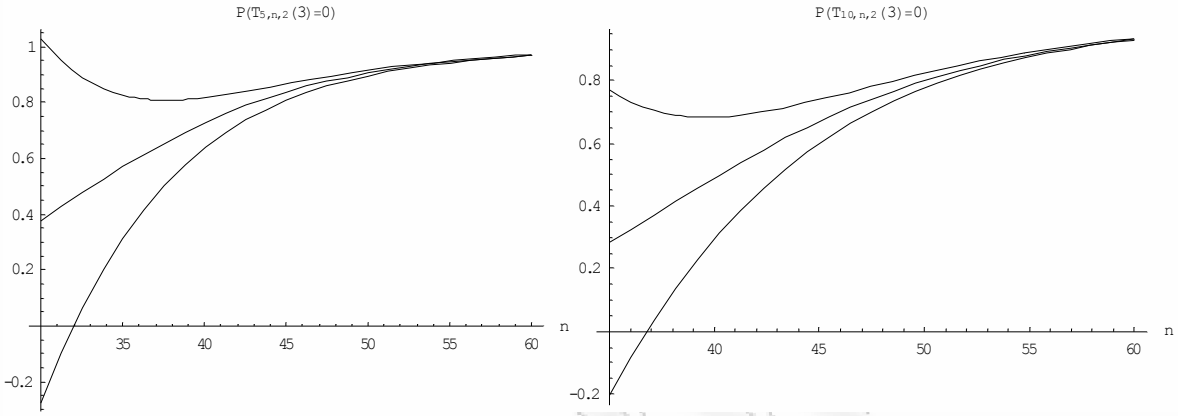
αντιστοίχως). Αντίστροφα, όσο το k μεγαλώνει, για συγκεκριμένο n, το φράγμα παίρνει μεγαλύτερες τιμές. Αξίζει επίσης να σημειώσουμε τις πολύ μικρές του UB_{TV}, ακόμη και για $n = 20$.



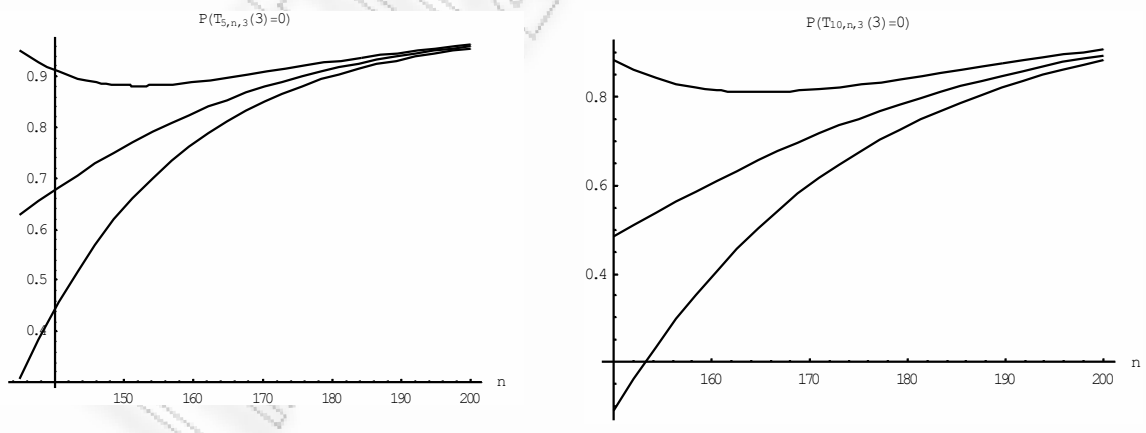
Σχήμα 4.2.1: Το ελάχιστο n, για το οποίο ισχύει $UB_{TV} \leq 0.01$ ή 0.001.

Αντίστοιχα συμπεράσματα μπορούν να εξαχθούν και για τις περιπτώσεις, $\{t = 2, q = 3\}$, $\{t = 3, q = 3\}$ και $\{t = 3, q = 4\}$ βλέποντας τον Πίνακα 4.2.2. Παράλληλα, στο Σχήμα 4.2.1 απεικονίζεται το ελάχιστο n που απαιτείται, ώστε το άνω φράγμα να πάρει τιμές μικρότερες από 0.01 ή 0.001, για δεδομένα t, q και k. Για παράδειγμα, όταν $t = q = 3$ και $k = 15$, τότε $UB_{TV} \leq 0.01$ για $n \geq 211$, ενώ $UB_{TV} \leq 0.001$ για $n \geq 242$ (οι αντίστοιχες τιμές για $t = 3, q = 4$ και $k = 15$, είναι 561 και 635).

Στα Σχήματα 4.2.2, 4.2.3, βλέπουμε τη γραφική παράσταση της $P(Z_{\lambda_{k,n}} = 0)$, μαζί με το άνω και κάτω φράγμα της εκτίμησης, δηλαδή, τις γραφικές παραστάσεις των $P(Z_{\lambda_{k,n}} = 0) \pm UB_{TV}$ (να θυμίσουμε ότι με $Z_{\lambda_{k,n}}$ συμβολίζουμε την τ.μ., που ακολουθεί κατανομή Poisson, με μέση τιμή $\lambda_{k,n}$). Παρατηρούμε και από τα δυο σχήματα ότι αρκετά γρήγορα (δηλαδή, για σχετικά μικρές τιμές του n) τα άνω και κάτω φράγματα συγκλίνουν.



Σχήμα 4.2.2: Περίπτωση: $P(Z_{\lambda_{k,n}} = 0) \pm UB_{TV}$ για $k = 5, 10$, $t = 2, q = 3$.



Σχήμα 4.2.3: Περίπτωση: $P(Z_{\lambda_{k,n}} = 0) \pm UB_{TV}$ για $k = 5, 10$, $t = q = 3$.

Πίνακας 4.2.2: Μελέτη του UB_{TV}.

$t=2,q=3$	k	$n=15$	$n=30$	$n=35$	$n=40$	$n=50$	$n=60$
	5	3.595	0.652	0.256	0.091	0.009	0.001
	10	4.225	1.061	0.487	0.194	0.023	0.002
	15	4.376	1.196	0.605	0.264	0.036	0.004
	20	4.447	1.245	0.665	0.311	0.046	0.005
$t=3,q=3$	k	$n=50$	$n=120$	$n=140$	$n=180$	$n=220$	$n=250$
	5	4.737	0.790	0.234	0.014	0.001	0.0001
	10	7.481	1.948	0.712	0.053	0.003	0.0003
	15	8.059	2.322	0.970	0.087	0.005	0.001
	20	8.315	2.460	1.110	0.116	0.007	0.001
$t=3,q=4$	k	$n=200$	$n=340$	$n=390$	$n=420$	$n=490$	$n=580$
	5	4.575	0.840	0.240	0.104	0.013	0.001
	10	7.291	2.057	0.732	0.349	0.049	0.003
	15	7.857	2.444	0.997	0.512	0.081	0.005
	20	8.108	2.585	1.141	0.622	0.108	0.008

4.3 Έλεγχος τυχαιότητας και αριθμητικά αποτελέσματα

Όπως έχουμε ήδη αναφέρει και στην αρχή του συγκεκριμένου κεφαλαίου, ένας από τους στόχους μας είναι να εισάγουμε και να μελετήσουμε, ένα νέο έλεγχο τυχαιότητας. Η στατιστική συνάρτηση του νέου ελέγχου, θα είναι η $T_{k,n,t}(q)$. Αρχικά θα περιγράψουμε τα χαρακτηριστικά του νέου ελέγχου, και έπειτα θα αναφέρουμε μερικά προβλήματα, μέσα από τα οποία μπορεί να προκύψει η ανάγκη για την εφαρμογή/χρησιμοποίησή του.

Ας υποθέσουμε ότι έχουμε ένα τυχαίο πίνακα διάστασης $k\times n$, και ότι η μηδενική υπόθεση H_0, την οποία θέλουμε να ελέγξουμε, είναι ότι τα kn στοιχεία του πίνακα, είναι ανεξάρτητες τ.μ., με κατανομή $P(X_{ij}=l)=1/q, l\in\mathcal{A}$. Δηλαδή,

$$H_0:\ X_{ij} \text{ ανεξάρτητες και ισόνομες με } P(X_{ij}=l)=1/q,\quad l\in\mathcal{A}$$
$$\text{για } i=1,2,\ldots,k, j=1,2,\ldots,n.$$

Η μηδενική υπόθεση, θα απορρίπτεται όταν ισχύει

$$T_{k,n,t}(q)<c_1 \text{ ή } T_{k,n,t}(q)>c_2$$

όπου $c_1,c_2\in\{0,1,\ldots,k-t+1\}$, με $c_1<c_2$. Άρα η κρίσιμη περιοχή του ελέγχου, είναι η

$$[0,c_1)\cup(c_2,k-t+1],$$

Πίνακας 4.3.1: Η πιθανότητα $P(T_{k,n,3}(3) = 0|H_0)$.

k	$n = 27$	$n = 28$	$n = 29$	$n = 30$
3	$2.455 \cdot 10^{-11}$	$3.820 \cdot 10^{-10}$	$3.053 \cdot 10^{-9}$	$1.675 \cdot 10^{-8}$
4	$3.245 \cdot 10^{-17}$	$1.122 \cdot 10^{-15}$	$1.883 \cdot 10^{-14}$	$2.057 \cdot 10^{-13}$
5	$4.289 \cdot 10^{-23}$	$3.311 \cdot 10^{-21}$	$1.175 \cdot 10^{-19}$	$2.576 \cdot 10^{-18}$
6	$5.668 \cdot 10^{-29}$	$9.772 \cdot 10^{-27}$	$7.333 \cdot 10^{-25}$	$3.227 \cdot 10^{-23}$

όπου τα c_1, c_2, για δεδομένο επίπεδο σημαντικότητας $a \in (0, 1)$, προσδιορίζονται από τη σχέση

$$P(T_{k,n,t}(q) < c_1 \text{ ή } T_{k,n,t}(q) > c_2 \mid H_0) \leq a,$$

ισοδύναμα,

$$P(T_{k,n,t}(q) < c_1|H_0) + P(T_{k,n,t}(q) > c_2|H_0) \leq a. \tag{4.3.1}$$

Η χρήση μιας κρίσιμης περιοχής, με την παραπάνω μορφή, είναι απαραίτητη καθώς για μικρές τιμές του n, αναμένουμε κάτω από την H_0, ένα μεγάλο πλήθος από υποπίνακες μη πλήρους κάλυψης. Οπότε κάτι αντίθετο, σίγουρα θα πρέπει να μας γεμίζει αμφιβολίες, για την ορθότητα της H_0. Βέβαια, για μεγάλα n, αυτό που είναι μη αναμενόμενο (κάτω από την H_0), θα είναι το μεγάλο πλήθος υποπινάκων μη πλήρους κάλυψης.

Παράλληλα, όταν το n παίρνει μικρές τιμές (κοντά στην τιμή q^t), τότε για κάθε $c_2 \in \{0, 1, \ldots, k - t\}$, οι πιθανότητες $P(T_{k,n,t}(q) > c_2|H_0)$ παίρνουν τιμές κοντά στη μονάδα (ισοδύναμα, η πιθανότητα $P(T_{k,n,t}(q) = 0|H_0)$ παίρνει πολύ μικρές τιμές). Αυτό είναι εμφανές και από τον Πίνακα 4.3.1, αν κοιτάξουμε τις τιμές της $P(T_{k,n,t}(q) = 0|H_0)$, όταν το n ανήκει στο σύνολο $\{q^t, q^t + 1, q^t + 2, q^t + 3\}$ (για $t = 3$, $q = 3$ και διάφορα k). Για να αντιμετωπίσουμε την περίπτωση αυτή, θέτουμε $c_2 = k - t + 1$, με αποτέλεσμα να ισχύει $P(T_{k,n,t}(q) > k - t + 1) = 0$. Η τιμή της παραμέτρου c_1, προσδιορίζεται τότε από τη σχέση

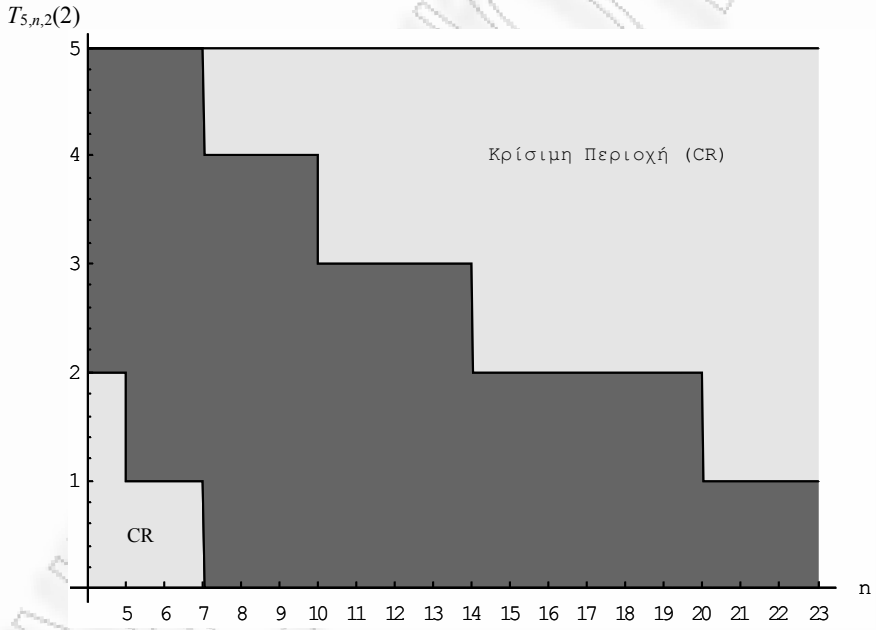
$$c_1 = \max\{r : P(T_{k,n,t}(q) < r) \leq a\} \in \{1, 2, \ldots, k - t + 1\}.$$

Αντίστροφα, εάν το n παίρνει πολύ μεγάλες τιμές σε σχέση με το q^t, τότε οι πιθανότητες $P(T_{k,n,t}(q) < c_1|H_0)$ είναι κοντά στη μονάδα, για κάθε $c_1 \in \{1, 2, \ldots, k-t+1\}$. Επομένως, στην περίπτωση αυτή μπορούμε να θέτουμε $c_1 = 0$ (οπότε, $P(T_{k,n,t}(q) < 0|H_0) = 0$), και το c_2, θα υπολογίζεται από τη σχέση

$$c_2 = \min\{r : P(T_{k,n,t}(q) > r) \leq a\} \in \{0, 1, \ldots, k - t\}.$$

Εάν όμως το n δεν πάρει ακραίες τιμές (πολύ μικρές ή πολύ μεγάλες), τότε σίγουρα μπορούμε να βρούμε $c_1, c_2 \in \{0, 1, \ldots, k-t+1\}$, με $c_1 \neq 0$ και $c_2 \neq k-t+1$, ώστε να ισχύει η (4.3.1). Από την άλλη μεριά όμως, η $T_{k,n,t}(q)$ είναι μια διακριτή τ.μ., και η χρήση μη τυχαιοποιημένων ελέγχων (randonmized test), δεν μπορεί να μας προσφέρει πάντοτε την ακριβή τιμή που επιδιώκουμε για το σφάλμα τύπου I.

Στα Σχήματα 4.3.1 και 4.3.2, απεικονίζεται η κρίσιμη περιοχή του ελέγχου, για τις περιπτώσεις $t = 2, q = 2$ και $t = 2, q = 3$, όταν $k = 5$ και $a = 0.05$. Έτσι, για $n = 6$ και $t = 2, q = 2$ (Σχήμα 4.3.1), η υπόθεση H_0 θα απορρίπτεται (σε επίπεδο σημαντικότητας $a = 0.05$), όταν $T_{5,6,2}(2) = 0$, ενώ για $n = 59$ και $t = 2, q = 3$ (Σχήμα 4.3.2), όταν $T_{5,59,2}(3) \geq 1$.

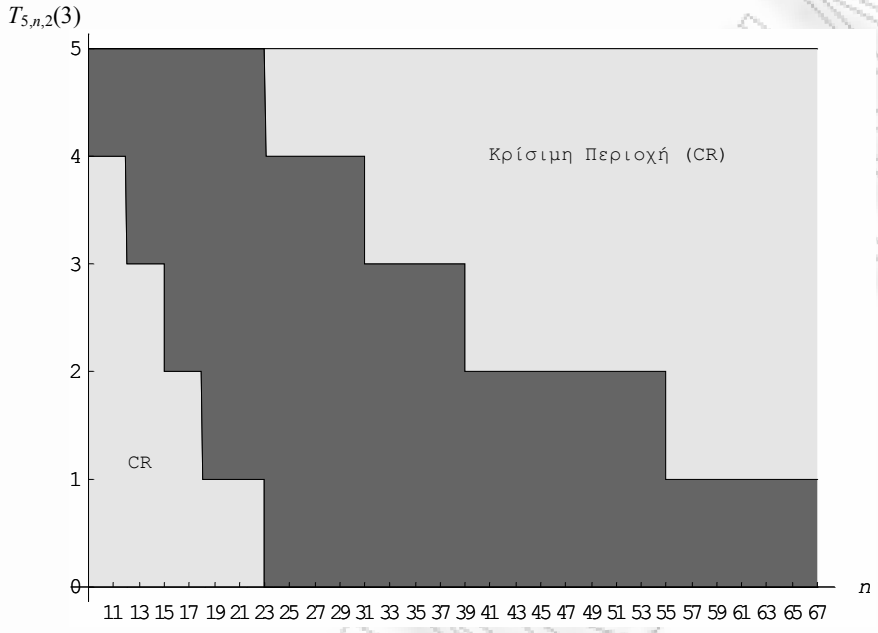


Σχήμα 4.3.1: Η κρίσιμη περιοχή για $t = q = 2, k = 5$, και $a = 0.05$.

Στα παραδείγματα που θα ακολουθήσουν, μας απασχολούν περιπτώσεις όπου το n παίρνει μεγάλες τιμές, και επομένως η κρίσιμη περιοχή του ελέγχου, θα ανάγεται στην $(c_2, k-t+1]$.

Υπάρχει μεγάλο πλήθος από εναλλακτικές υποθέσεις H_1, για τις οποίες ένας έλεγχος τυχαιότητας βασισμένος σε συνεχόμενες παρατηρήσεις (όπως αυτός που προτείνουμε), μπορεί να μας προσφέρει ικανοποιητικά αποτελέσματα (όσον αφορά βέβαια, την ισχύ του ελέγχου). Για παράδειγμα, ας υποθέσουμε ότι $q \geq 2$ και ότι οι τ.μ. της πρώτης γραμμής, ακολουθούν την ομοιόμορφη κατανομή, ήτοι

$$P(X_{1j} = l) = \frac{1}{q}, \quad j = 1, 2, \ldots, n$$



Σχήμα 4.3.2: Η κρίσιμη περιοχή για $t = 2, q = 3, k = 5$, και $a = 0.05$.

ενώ, για $i \geq 2$, έχουμε

$$P(X_{ij} = l_2 | X_{i-1,j} = l_1) = \vartheta_1, \text{ για } l_1 = l_2,$$

$$P(X_{ij} = l_2 | X_{i-1,j} = l_1) = \vartheta_2, \text{ για } l_1 \neq l_2,$$

με $l, l_1, l_2 \in \mathcal{A}$, $j = 1, 2, \ldots, n$, όπου

$$\vartheta_1 + (q-1)\vartheta_2 = 1.$$

Τότε, εύκολα διαπιστώνουμε ότι για κάθε $l \in \mathcal{A}$, ισχύει

$$P(X_{2j} = l) = \sum_{l_1=0}^{q-1} P(X_{2j} = l | X_{1j} = l_1) P(X_{1j} = l_1) = \frac{1}{q}(\vartheta_1 + (q-1)\vartheta_2) = \frac{1}{q},$$

με $j = 1, 2, \ldots, n$. Επομένως, για οποιεσδήποτε τιμές των ϑ_1, ϑ_2 (με $\vartheta_1, \vartheta_2 \in (0, 1)$ και $\vartheta_1 + (q-1)\vartheta_2 = 1$), η παραπάνω περίπτωση καταλήγει σε μια (στάσιμη) Μαρκοβιανή αλυσίδα, με $P(X_{ij} = 1) = 1/q$, για κάθε i, j. Τότε, είναι σχεδόν βέβαιο ότι έλεγχοι που βασίζονται σε στατιστικές συναρτήσεις σχετικές με το πλήθος των εμφανίσεων των γραμμάτων, όπως η $\sum_{i,j} X_{ij}$, δε θα είναι ικανοί να εντοπίσουν την απόκλιση ανάμεσα στην H_0 και την H_1. Από την άλλη, εάν το ϑ_1 πάρει μεγάλες τιμές, τότε σε κάθε στήλη θα συναντάμε μεγάλο πλήθος από συνεχόμενα όμοια σύμβολα, και δε θα υπάρχουν λέξεις όπως η $0101\ldots$ (π.χ. για την

Πίνακας 4.3.2: Περίπτωση: t=2, q=3.

	k	c	$P(Z_{\lambda_{k,n}} > c)$	UB_{TV}	$power_1$	$power_2$	$power_M$
$n = 60$	5	0	0.030	0.001	0.161(0.162)	0.359(0.359)	0.332
	10	1	0.002	0.002	0.066(0.061)	0.267(0.265)	0.199
	20	1	0.010	0.005	0.204(0.206)	0.624(0.624)	0.540
	40	1	0.037	0.010	0.521(0.516)	0.929(0.930)	0.885
$n = 80$	5	0	0.003	10^{-5}	0.041(0.040)	0.148(0.158)	0.088
	10	0	0.007	$2 \cdot 10^{-5}$	0.083(0.089)	0.328(0.322)	0.195
	20	0	0.014	$5 \cdot 10^{-5}$	0.178(0.178)	0.562(0.559)	0.367
	40	0	0.028	0.0001	0.325(0.331)	0.801(0.814)	0.613

$$power_1 : p_0 = 0.45, p_1 = 0.30, p_2 = 0.25$$

$$power_2 : p_0 = 0.40, p_1 = 0.40, p_2 = 0.20$$

$$power_M : p_{l_1 l_2} = 0.60, \text{ για } l_1 = l_2, p_{l_1 l_2} = 0.20, \text{ για } l_1 \neq l_2$$

περίπτωση $q = 2$). Έτσι το πλήθος των υποπινάκων μη πλήρους κάλυψης, θα είναι μεγάλο. Αντίστροφα, εάν το ϑ_2 πάρει μεγάλες τιμές, τότε λέξεις της μορφής $111\ldots$ ή $000\ldots$, θα είναι απούσες, δίνοντας και σ' αυτή την περίπτωση μεγάλες τιμές στην $T_{k,n,t}(q)$.

Η μηδενική υπόθεση που μελετάμε στους επόμενους πίνακες, είναι ότι τα στοιχεία του πίνακα είναι ανεξάρτητες και ισόνομες τ.μ., οι οποίες ακολουθούν την ομοιόμορφη διακριτή κατανομή. Ως εναλλακτικές υποθέσεις, θεωρούμε διαφορές περιπτώσεις, στις οποίες ο έλεγχος, όπως φαίνεται, δίνει αξιόλογα αποτελέσματα. Με βάση τη συζήτηση που προηγήθηκε, η μηδενική υπόθεση θα απορρίπτεται εάν

$$T_{k,n,t}(q) > c$$

όπου το c-για επίπεδο σημαντικότητας $a \in (0, 1)$-θα προσδιορίζεται από τη σχέση

$$c = \min\{r : P(T_{k,n,t}(q) > r) \leq a\}.$$

Πρέπει να αναφέρουμε, ότι στους Πίνακες 4.3.2 και 4.3.3, χρησιμοποιούμε τις προσεγγίσεις μέσω Poisson, που προκύπτουν από τα αποτελέσματα της Παραγράφου 4.2, ώστε να εκτιμήσουμε το c, από τη σχέση ($a = 0.05$)

$$c = \min\{r : P(Z_{\lambda_{k,n}} > r) + UB_{TV} \leq 0.05\}. \tag{4.3.2}$$

Επίσης, παρόλο που η ακριβής κατανομή της $T_{k,n,t}(q)$ έχει προσδιοριστεί και για την περίπτωση που περιγράφεται από την (4.1.8), η ισχύς του ελέγχου θα εκτιμάται μέσω προσομοίωσης ή από την προσέγγιση μέσω της κατανομής Poisson (λόγω των μεγάλων τιμών

που παίρνει το n). Στην περίπτωση της Μαρκοβιανής εξάρτησης, ανάμεσα στα στοιχεία μιας συγκεκριμένης στήλης (Μαρκοβιανή εξάρτηση ανάμεσα στις γραμμές του πίνακα), θα εκτιμήσουμε την ισχύ μόνο μέσω προσομοίωσης, αφού δεν υπάρχει διαθέσιμη προσέγγιση. Σε πολλές περιπτώσεις παρατηρούμε ότι η ισχύς παίρνει μεγάλες τιμές, ακόμη και της τάξεως του 0.999. Στον Πίνακα 4.3.2 και στις στήλες $power_1$ και $power_2$, δίνεται η προσέγγιση μέσω προσομοίωσης των πιθανοτήτων

$$P(T_{k,n,t}(q) > c|p_0, p_1, \ldots, p_{q-1})$$

ενώ στις παρενθέσεις έχουμε την προσέγγιση που προκύπτει από το Θεώρημα 4.2.1 ($P(Z_{\lambda_{k,n}} > c|p_0, p_1, \ldots, p_{q-1})$). Η στήλη $power_M$ (τιμές μέσω προσομοίωσης), αναφέρεται στην περίπτωση της Μαρκοβιανής εξάρτησης, ανάμεσα στις γραμμές του πίνακα, δηλαδή

$$P(T_{k,n,t}(q) > c|p_{00}, p_{01}, \ldots, p_{q-1,q-1})$$

όπου

$$p_{l_1 l_2} = P(X_{ij} = l_2|X_{i-1,j} = l_1), l_1, l_2 \in \{0, 1, \ldots, q-1\}$$

για $i = 2, 3, \ldots, k$ και $j = 1, 2, \ldots, n$ (Μαρκοβιανή εξάρτηση τάξεως 1), ενώ $P(X_{1j} = l) = 1/q, l \in \mathcal{A}$.

Μέσα και από τα αριθμητικά αποτελέσματα, είναι φανερό πως, καθώς απομακρυνόμαστε από την περίπτωση $p_l = 1/q$, η ισχύς του ελέγχου αυξάνεται. Για παράδειγμα, όταν $p_0 = 0.45, p_1 = 0.30, p_2 = 0.25$ (με $n = 60, t = 2, q = 3$, Πίνακας 4.3.2) η ισχύς είναι ίση με 0.1624, ενώ η ισχύς γίνεται 0.3593, όταν πηγαίνουμε στην περίπτωση $p_0 = 0.40, p_1 = 0.40, p_2 = 0.20$. Τέλος, αξίζει να αναφέρουμε ότι οι στήλες $power_M$ αντιστοιχούν σε περιπτώσεις σαν και αυτές που περιγράψαμε στην αρχή της παραγράφου, όπου κριτήρια βασισμένα στις συχνότητες εμφάνισης των γραμμάτων, δεν είναι δυνατόν να εντοπίσουν την «απομάκρυνση» μας, από τη μηδενική υπόθεση (αφού, οι μη δεσμευμένες πιθανότητες είναι ίσες με $1/q$).

Μία άλλη περίπτωση, από την οποία μπορούμε να οδηγηθούμε σ' ένα έλεγχο της παραπάνω μορφής, προκύπτει και από τους μονοδιάστατους ελέγχους τυχαιότητας, επάνω σε μια ακολουθία τ.μ. Πιο συγκεκριμένα, ας υποθέσουμε ότι στην υπό μελέτη ακολουθία, υπάρχει ένα είδος (ισχυρής) κυκλικής επίδρασης (cyclical effects) στα δεδομένα μας. Αυτό έχει ως αποτέλεσμα οι τιμές που θα παρατηρούμε από μια χρονική στιγμή και έπειτα, να επηρεάζονται από συγκεκριμένες στιγμές του παρελθόντος. Ας θεωρήσουμε το επόμενο παράδειγμα, όπου έχουμε μια ακολουθία από 30 δοκιμές Bernoulli

$$\underbrace{001110010001000}_{\text{οι πρώτες 15 δοκιμές}}101110010101100.$$

Πίνακας 4.3.3: Περίπτωση: $t = 3, q = 3$.

	k	c	$P(Z_{\lambda_{k,n}} > c)$	UB_{TV}	$power_1$	$power_2$	$power_M$
$n = 200$	5	0	0.042	0.003	0.380(0.369)	0.728(0.687)	0.533
	10	1	0.006	0.013	0.349(0.348)	0.848(0.814)	0.686
	20	1	0.028	0.030	0.765(0.763)	0.996(0.992)	0.977
	40	2	0.017	0.057	0.933(0.931)	0.999(0.999)	0.999
$n = 250$	5	0	0.006	10^{-4}	0.160(0.163)	0.504(0.492)	0.284
	10	0	0.017	$3 \cdot 10^{-4}$	0.380(0.378)	0.839(0.836)	0.663
	20	0	0.038	0.001	0.650(0.657)	0.983(0.983)	0.936
	40	1	0.003	0.002	0.654(0.659)	0.998(0.998)	0.979
$n = 280$	5	0	0.002	10^{-7}	0.096(0.096)	0.388(0.389)	0.184
	10	0	0.006	$3 \cdot 10^{-5}$	0.227(0.237)	0.729(0.731)	0.506
	20	0	0.012	10^{-4}	0.458(0.455)	0.958(0.948)	0.820
	40	0	0.026	$2 \cdot 10^{-4}$	0.730(0.723)	0.999(0.998)	0.990

$power_1 : p_0 = 0.45, p_1 = 0.30, p_2 = 0.25$, $power_2 : p_0 = p_1 = 0.40, p_2 = 0.20$

$power_M : p_{00} = p_{22} = p_{11} = 0.50, p_{01} = p_{10} = p_{20} = 0.30,$

$p_{12} = p_{21} = p_{02} = 0.20$

Είναι φανερό ότι, η ακολουθία των αποτελεσμάτων από την $15^η$ παρατήρηση και μετά, είναι σχεδόν η ίδια με την αρχική (λόγω, πιθανότατα, μίας ισχυρής κυκλικής επίδρασης). Οι περισσότεροι από τους ελέγχους τυχαιότητας που αναφέρονται σε ακολουθίες δίτιμων δοκιμών (για τους οποίους έχουμε μιλήσει στην αρχή του κεφαλαίου), βασίζονται ως επί το πλείστον, σε κριτήρια ροών (runs rules), όπως το συνολικό πλήθος ροών ή το μήκος της μέγιστης ροής (the length of the longest run), είτε γενικά σε απαριθμήτριες ροών, συγκεκριμένου είδους (π.χ. ροές συγκεκριμένου μήκους). Έτσι οι περισσότεροι από τους παραπάνω ελέγχους, είναι πολύ πιθανόν να μη διακρίνουν την κυκλικότητα που υπάρχει στην παραπάνω ακολουθία και ως εκ τούτου, δε θα απέρριπταν την υπόθεση των ανεξάρτητων και ισόνομων δοκιμών. Εάν όμως χωρίσουμε την αρχική μας ακολουθία σε δυο υπακολουθίες των 15 παρατηρήσεων, και τις τοποθετήσουμε τη μια κάτω από την άλλη, θα πάρουμε τον επόμενο 2×15 πίνακα

$$001110010001000$$
$$101110010101100$$

Από τον προηγούμενο πίνακα, παρατηρούμε ότι η λέξη $\binom{1}{0}$, δεν υπάρχει ως στήλη. Η πιθανότητα να συμβαίνει κάτι τέτοιο ($T_{2,15,2}(2) = 1$), δηλαδή, ένας 2×15 πίνακας να είναι μη πλήρους κάλυψης, κάτω από την υπόθεση ότι τα στοιχεία του είναι i.i.d. συμμετρικές

δοκιμές Bernoulli, είναι περίπου 0.05. Άρα, τα προηγούμενα δεδομένα θα πρέπει να μας δημιουργήσουν αρκετές αμφιβολίες, για την ορθότητα της υπόθεσής μας. Συνυπολογίζοντας το γεγονός ότι η παραπάνω διαδικασία ελέγχου, μπορεί να εφαρμοστεί σε οποιαδήποτε ακολουθία διακριτών τ.μ. (με $q > 2$), καταλαβαίνουμε πως ο νέος έλεγχος, παρουσιάζει μία αξιοπρόσεχτη δυναμική.

Σύνοψη

Η παρούσα διατριβή ξεκίνησε (Κεφάλαιο 1), με τη μελέτη των συστημάτων αξιοπιστίας μονάδων, με πολλαπλά επίπεδα αποτυχίας (στα οποία δεν υιοθετείται, καμία διάταξη ανάμεσα στις καταστάσεις αποτυχίας των μονάδων). Αρχικώς, εισάγαμε τις βασικές έννοιες από την προηγούμενη κλάση συστημάτων και αναφερθήκαμε στη θεωρία των στοχαστικών διατάξεων, ανάμεσα σε τυχαίες μεταβλητές και τυχαία διανύσματα. Τα εργαλεία που μας προσέφερε η θεωρία των στοχαστικών διατάξεων μας επέτρεψαν να προσεγγίσουμε την αξιοπιστία ενός συστήματος με πολλαπλά επίπεδα αποτυχίας, δηλαδή, την πιθανότητα το σύστημα να βρίσκεται σε κατάσταση λειτουργίας (ισοδύναμα, αποτυχίας).

Αυτό επιτεύχθηκε μέσα από ένα κάτω φράγμα, πολλαπλασιαστικού τύπου, για την προαναφερθείσα πιθανότητα. Ο υπολογισμός του νέου φράγματος, βασίζεται σε μια διάταξη των οικογενειών των ελάχιστων συνόλων διακοπής, με αποτέλεσμα, σε ένα σύστημα μονάδων με m διαφορετικά επίπεδα αποτυχίας ($m \in \{1, 2, \ldots\}$), να έχουμε εν γένει $m!$, διαφορετικές τιμές για το κάτω φράγμα. Παράλληλα, ο υπολογισμός του φράγματος μέσα από υπολογιστικά απλές διαδικασίες, μας έδωσε τη δυνατότητα να προσεγγίσουμε την αξιοπιστία συστημάτων με πολύ μεγάλο πλήθος μονάδων, της οποίας ο ακριβής προσδιορισμός, είναι πρακτικά αδύνατος (με τις υπάρχουσες μεθόδους). Η ακρίβεια και η ποιότητα της προσέγγισης, εξετάστηκε μέσα από διάφορα συστήματα που έχουν ήδη απασχολήσει τη βιβλιογραφία, αλλά και μέσα από ένα σύστημα, που για πρώτη φορά εισήχθηκε και μελετήθηκε, στη συγκεκριμένη διατριβή.

Στο Κεφάλαιο 2 ασχοληθήκαμε με τις συναρτήσεις σάρωσης, οι οποίες έχουν άμεση σχέση τόσο με τη θεωρία αξιοπιστίας, όσο και με το στατιστικό έλεγχο ποιότητας. Έγινε αρχικά μία ανασκόπηση της θεωρίας που αφορά την προσέγγιση των κατανομών της απλής ή πολλαπλής συνάρτησης σάρωσης. Επικεντρωθήκαμε σε διαδικασίες που προσφέρουν ταυτόχρονα και φράγματα για τα σφάλματα των αντίστοιχων προσεγγίσεων, καθώς στόχος μας ήταν να μελετήσουμε και τις ασυμπτωτικές ιδιότητες των προσεγγίσεων. Παρατηρήσαμε ότι το σύνολο των παραπάνω αποτελεσμάτων, αναφέρονται στην προσέγγιση των συναρτήσεων σάρωσης μέσα από μία κατανομή Poisson (απλή ή σύνθετη). Στη συνέχεια, ορίσαμε

τις συναρτήσεις σάρωσης μέσα από ένα γενικότερο μοντέλο, βασισμένο σε μια ακολουθία ανεξάρτητων και ισόνομων συνεχών τυχαίων μεταβλητών. Κάτω από την υπόθεση ότι η συνάρτηση κατανομής των τυχαίων μεταβλητών, ανήκει σε κάποιο από τα μέγιστα πεδία έλξης των κατανομών ακραίων τιμών, εξασφαλίσαμε τις προϋποθέσεις για την ανάπτυξη χρήσιμων αποτελεσμάτων, για την ασυμπτωτική συμπεριφορά των υπερβάσεων της παραπάνω ακολουθίας, σε κινούμενα-επικαλυπτόμενα παράθυρα. Ενδιαφέρον παρουσίασε και η σύνδεση των νέων (γενικευμένων) συναρτήσεων σάρωσης, με τις κινούμενες διατεταγμένες παρατηρήσεις.

Τέλος, στα Κεφάλαια 3 και 4, μας απασχόλησαν οι πίνακες πλήρους κάλυψης. Αναφερθήκαμε στις βασικότερες έννοιες, και σκιαγραφήσαμε τις εφαρμογές που παρουσιάζουν οι παραπάνω πίνακες, κυρίως, στον έλεγχο ποιότητας. Στη συνέχεια, εισάγαμε και μελετήσαμε μια νέα κλάση τυχαίων πινάκων (με άμεση σχέση με τους τελευταίους), τους πίνακες συνεχόμενης πλήρους κάλυψης. Συγκεκριμένα, υπολογίσαμε την κατανομή της τυχαίας μεταβλητής που απαριθμεί το πλήθος των υποπινάκων, από τους οποίους λείπει τουλάχιστον μια λέξη, στην περίπτωση που τα στοιχεία του πίνακα είναι διακριτές τυχαίες μεταβλητές (με πεπερασμένο πεδίο τιμών). Η μέθοδος που χρησιμοποιήθηκε για να φέρουμε εις πέρας την παραπάνω διαδικασία, ήταν η εμφύτευση τυχαίων μεταβλητών σε Μαρκοβιανή αλυσίδα. Ταυτόχρονα, η μελέτη που προηγήθηκε ανέδειξε και την ανάγκη για την εύρεση ποιοτικών προσεγγίσεων, με αποτέλεσμα να αναπτύξουμε ένα άνω φράγμα για την απόσταση ανάμεσα στην προηγούμενη απαριθμήτρια τυχαία μεταβλητή, και μία κατάλληλα ορισμένη κατανομή Poisson (μέσω της μεθόδου Chen-Stein). Επιπλέον, μελετήθηκαν δύο εφαρμογές στα πεδία των παραγοντικών σχεδιασμών και των ελέγχων τυχαιότητας.

Βιβλιογραφία

[1] Agin, M.A. and Godbole, A.P. (1992). A new exact runs test for randomness. In *Computing science and statistics* (Eds., Page,C. and Le Page, R.), Proceedings of the 22nd Symposium on the Interface, 281-285, Springer-Verlag, New York.

[2] Aki, S. and Hirano, K. (2004). Waiting time problems for a two-dimensional pattern. *Annals of the Institute of Statistical Mathematics*, 56, 169-182.

[3] Akiba, T. and Yamamoto, H. (2001). Reliability of a two-dimensional k-within-consecutive-$r \times s$-out-of-$m \times n$: F system. *Naval Research Logistics*, 48, 625-637.

[4] Antzoulakos, D.L. (2001). Waiting times for patterns in a sequence of multistate trials. *Journal of Applied Probability*, 38, 508-518.

[5] Antzoulakos, D.L., Bersimis, S. and Koutras, M.V. (2003). On the distribution of the total number of run lengths. *Annals of the Institute of Statistical Mathematics*, 55, 865-884.

[6] Arnold, B.C. (1985). Pareto distributions. In *Encyclopedia of Statistical Sciences* (Eds., Kotz, S. Johnson, N.L. and Read, C.B.), 568-574. John Wiley & Sons, New York.

[7] Arnold, B.C. and Balakrishnan, N. (1989). *Relations, Bounds, and Approximations for Order Statistics*. Springer, New York.

[8] Arratia, R.L, Goldstein, L. and Gordon, L. (1989). Two moments suffice for Poisson approximations: The Chen-Stein method. *Annals of Probability*, 17, 9-25.

[9] Arratia, R.L, Goldstein, L. and Gordon, L. (1990). Poisson approximation and the Chen-Stein method. *Statistical Science*, 5, 403-423.

[10] Arratia, R.L, Gordon, L. and Waterman, M. (1990). The Erdös-Rényi Law in distribution, for coin tossing and sequence matching. *Annals of Statistics*, 18, 539-570.

[11] Balakrishnan, N. and Koutras, M.V. (2002). *Runs and Scans with Applications*. John Wiley & Sons, New York.

[12] Barbour, A.D., Holst, L., and Janson, S. (1992). *Poisson Approximation*. Clarendon Press, Oxford.

[13] Barlow, R. and Heidtmann, K. (1984). Computing k-out-of-n system reliability. *IEEE Transactions on Reliability*, 33, 322-323.

[14] Barlow, R. and Hunter, L. (1960a). Mathematical models for system reliability, Part II. *The Sylvania Technology*, 13, 55-65.

[15] Barlow, R. and Hunter, L. (1960b). Criteria for determining optimum redundancy. *IRE Transactions on Reliability and Quality Control*, 9, 73-77.

[16] Barlow, R., Hunter, L. and Proschan, F. (1963). Optimum redundancy when components are subject to two kinds of failure. *Journal of the Society for Industrial and Applied Mathematics*, 11, 64-73.

[17] Barlow, R. and Proschan, F. (1981). *Statistical Theory of Reliability and Life Testing*. Holt, Reinhart and Winston, New York.

[18] Barlow, R. and Wu, A. (1978). Coherent systems with multi-state components. *Mathematics of Operations Research*, 3, 275-281.

[19] Beizer, B. (1990). *Software Testing Techniques* (Second ed.). International Thompson Computer Press, Boston.

[20] Ben-Dov, Y. (1980). Optimal reliability design of k-out-of-n systems subject to kinds of failure. *Journal of Operation Research Society*, 31, 743-748.

[21] Bhapkar, V. P. (1961). A nonparametric test for the problem of several samples. *Annals of Mathematical Statistics*, 32, 1108-1117.

[22] Bierbrauer, J. and Schellwatt, H. (2000). Almost independent and weakly biased arrays: efficient constructions and cryptologic applications. *Advances in Cryptology, Lecture Notes in Computer Science*, 533-543.

[23] Birnbaum, Z.W., Esary, J.D. and Saunders, S.C. (1961). Multi-component systems and structures and their reliability. *Technometrics*, 3, 55-77.

[24] Boland, P.J. El-Newehi, E. and Proschan, F. (1994). Application of the hazard rate ordering in reliability and order statistics. *Journal of Applied Probability*, 31, 180-192.

[25] Boutsikas, M.V. and Koutras, M.V. (2000). A bound for the distribution of the sum of discrete associated or NA random variables. *The Annals of Applied Probability*, 10, 1137-1150.

[26] Boutsikas, M.V. and Koutras, M.V. (2001). Compound Poisson approximation for sums of dependent random variables. In *Probability and statistical models with applications* (Eds., Charalambides, Ch.A., Koutras, M.V. and Balakrishnan, N.), 63-86. FL: Chapman & Hall, Boca Raton.

[27] Boutsikas, M.V. and Koutras, M.V. (2002a). On a class of multiple failure mode systems. *Naval Research Logistics*, 49, 167-185.

[28] Boutsikas, M.V. and Koutras, M.V. (2002b). Modeling claim exceedances over thresholds. *Insurance: Mathematics and Economics*, 30, 67-83.

[29] Boutsikas, M.V. and Koutras, M.V. (2003). Bounds for the destribution of two-dimensional binary scan statistics. *Probability in the Engineering and Informational Sciences*, 17, 509-525.

[30] Boutsikas, M.V. and Koutras, M.V. (2006). On the asymptotic distribution of the discrete scan statistic. *Journal of Applied Probability*, 43, 1137-1154.

[31] Boutsikas, M.V., Koutras, M.V. and Milienos, F.S. (2008). Extreme value results for scan statistics. In *Scan statistics methods and applications* (Eds., Glaz, J., Pozdnyakov, V. and Wallenstein, S.), 57-82, Birkhauser.

[32] Bowers, N.L., Gerber, H.U., Hickman, J., Jones, D.A. and Nesbitt, C.J. (1997). *Actuarial Mathematics* (Second ed.). The society of Actuaries, Illinois.

[33] Bush, K.A. (1952). Orthogonal arrays of index unity. *The Annals of Mathematical Statistics*, 23, 426-434.

[34] Carey, P.A. and Godbole, A.P. (2008). Partial covering arrays and a generalized Erdös-Ko-Rado property. under revision.

[35] Carroll, C.T. (2003a). The cost of poor testing: A U.S. government study (part 1). EDPACS, *The EDP Audit, Control and Security Newsletter*, 31 (1), 1-17.

[36] Carroll, C.T. (2003b). The cost of poor testing: A U.S. government study (part 2). EDPACS, *The EDP Audit, Control and Security Newsletter*, 31 (2), 1-16.

[37] Champ, C.W. and Woodall, W.H. (1987). Exact results for Shewhart control charts with supplementary runs rules. *Technometrics*, 29, 393-399.

[38] Chao, M.T. and Fu, J.C. (1989). A limit theorem of certain repairable systems. *Annals of the Institute of Statistical Mathematics*, 41, 809-818.

[39] Chao, M.T. and Fu, J.C. (1991). The reliability of large series system under a Markov chain structure. *Advances in Applied Probability*, 23, 894-908.

[40] Chao, M.T., Fu, J.C. and Koutras, M.V. (1995). Survey of reliability studies of consecutive-k-out-of-n: F and related systems, *IEEE Transactions on Reliability*, R44, 120-127.

[41] Charalambides, Ch. A. (2002). *Enumerative Combinatorics*. Chapman & Hall/CRC Press, Boca Raton, Florida.

[42] Cheng, C. (1995). Some projection properties of orthogonal arrays. *The Annals of Statistics*, 23, 1223-1233.

[43] Chryssaphinou, O. and Vaggelatou, E. (2002). Compound Poisson approximation for multiple runs in a Markov chain. *Annals of the Institute of Statistical Mathematics*, 54, 411-424.

[44] Chen, J. and Glaz, J. (1996). Two dimensional discrete scan statistics. *Statistics and Probability Letters*, 31, 59-68.

[45] Chen, J. and Glaz, J. (1999). Approximations for discrete scan statistics on the circle. *Statistics and Probability Letters*, 44, 167-176.

[46] Chen, J. and Glaz, J. (1999). Approximations for the distribution and the moments of discrete scan statistics. In *Scan Statistics and Applications* (Eds., Glaz, J. and Balakrishnan, N.), Birkhäuser, Boston.

[47] Cleveland, W.S. and Kleiner, B. (1975). A graphical technique for enhancing scatterplots with moving statistics. *Technometrics*, 17, 447-454.

[48] Colbourn, C.J. (2004). Combinatorial aspects of covering arrays. *Le Matematiche* (Catania), 58, 121-167.

[49] Colbourn, C.J. (2008). Strength two covering arrays: Existence tables and projection. *Discrete Mathematics*, 308, 772-786.

[50] Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer-Verlag, London.

[51] Dalal, S.R. and Mallows, C.L. (1998). Factor-covering designs for testing software. *Technometrics*, 40, 234-243.

[52] Canfield, E.R. and McCormick, W.P. (1992). Asymptotic reliability of consecurive k-out-of-n systems. *Journal of Applied Probability,* 29, 142-155.

[53] David, H.A. (1955). A note on moving ranges. *Biometrika*, 42, 512-515.

[54] David, H.A. and Nagaraja, H.N. (2003). *Order Statistics* (Third ed.). John Wiley & Sons, New York.

[55] David, H.A. and Rogers, M.P. (1983). Order statistics in overlapping samples, moving order statistics U-statistics. *Biometrika*, 70, 245-249.

[56] de Moivre, A. (1738). *The doctrine of chance* (Third ed.). Chelsea Publishing Co., New York.

[57] Dembo, A. and Karlin, S. (1992). Poisson approximations for r-scan processes. *Annals of Applied Probability*, 2, 329-357.

[58] Deheuvels, P. and Devroye, L.(1987). Limit laws of Erdös-Rényi-Shepp type. *The Annals of Probability*, 15, 1363-1386.

[59] Dhillon, B. and Rayapati, S. (1986). A method to evaluate reliability of three-state device networks. *Microelectronics and Reliability*, 26, 525-554.

[60] Dudkiewicz, J. (1998). Compound Poisson approximation for extremes for moving minima in arrays of independent random variables. *Applicationes Mathematicae*, 25, 19-28.

[61] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modeling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin.

[62] Erdös-Rényi (1970). On a new law of large numbers. *Journal Analyse Mathematics*, 23, 103-111.

[63] Esary, J.D. and Proschan, F. (1963). Coherent structures of non-identical components. *Technometrics*, 5, 191-209.

[64] Fu, J.C. (1986). Reliability of consecutive-k-out-of-n: F systems with (k-1)-step Markov dependence. *IEEE Transactions on Reliability*, 35, 602-606.

[65] Fu, J.C. (1996). Distribution theory of runs and patterns associated with a sequence of multi-state trials. *Statistica Sinica*, 6, 957-974.

[66] Fu, J.C. (2001). Distribution of the scan statistic for a sequence of bistate trials. *Journal of Applied Probability*, 38, 908-916.

[67] Fu, J.C. and Chang, Y.M. (2002). On probability generating functions for waiting time distributions of compound patterns in a sequence of multistate trials. *Journal of Applied Probability*, 39, 70-80.

[68] Fu, J.C. and Hu, B. (1987). On reliability of large consecutive-k-out-of-n: F systems with (k-1)-step Markov dependence. *IEEE Transactions on Reliability*, 36, 75-77.

[69] Fu, J.C. and Lou W.Y.W (2003). D*istribution Theory of Runs and Patterns and its Applications: A Finite Markov Chain Imbedding Approach*. World Scientific, Singapore.

[70] Fu, J.C. and Koutras M.V. (1994). Distribution theory of runs: a Markov chain approach. *Journal of the American Statistical Association* , 89, 1050-1058.

[71] Gargano, L., Korner, J. and Vaccaro, U. (1992). Sperner theorems on directed graphs and qualitative independence. *Journal of Combinatorial Theory, Series A*, 61, 173-192.

[72] Gargano, L., Korner, J., and Vaccaro, U. (1994). Capacities: from information theory to extremal set theory. *Journal of Combinatorial Theory, Series A*, 68, 296-316.

[73] Gibbons, J.D. and Chakraborti, S. (1992). *Non Parametric Statistical Inference* (Third ed.). Marcel Dekker, New York.

[74] Glaz, J. and Balakrishnan, N. (1999). *Scan Statistics and Applications*. Birkhäuser, Boston.

[75] Glaz, J. and Naus, J. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data. *The Annals of Applied Probability*, 1, 306-318.

[76] Glaz, J., Naus, J., and Wallenstein, S. (2001). *Scan Statistics*. Springer Verlag, New York.

[77] Godbole, A. and Janson, S. (1996). Random covering designs. *Journal of Combinatorial Theory, Series A*, 75, 85-98.

[78] Godbole, A.P., Koutras, M.V. and Milienos, F.S. (2008a). Binary consecutive covering arrays (accepted for publiacation).

[79] Godbole, A.P., Koutras, M.V. and Milienos, F.S. (2008b). Consecutive covering arrays and a new randomness Test. submitted for publication.

[80] Godbole, A.P., Skipper, D.E. and Sunley, R.A. (1996). t-covering arrays: upper bounds and Poisson approximations. *Combinatorics, Probability, and Computing*, 5, 105-117.

[81] Goldstein, L. and Waterman, M. (1992). Poisson, compound Poisson and process approximations for testing statistical significance in sequence comparisons. *Bulletin of Mathematical Biology*, 54, 785-812.

[82] Govindaraju, K. and Lai, C.D. (1999). Design of multiple sampling plan. *Communications in Statistics-Simulation and Computation*, 28, 1-11.

[83] Griffith, W.S. (1986). On consecutive k-out-n failure systems and their generalizations. *Reliability and quality control*, 157-165.

[84] Hahn, G.J. and Gage, J.B. (1983). Evaluation of a start-up demonstration test. *Journal of Quality Technology*, 15, 103-106.

[85] Hartman, A. (2006). Software and hardware testing using combinatorial covering suites. *In Interdisciplinary Applications of Graph theory, Combinatorics and Algorithms* (Eds., Golumbic, M.C. and Hartman, A.), Springer, 237-266.

[86] Hedayat, A.S., Sloane, N.J.A. and Stufken, J. (1999). *Orthogonal Arrays*. Springer, New York.

[87] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994). *Continuous univariate distributions*, Vol. 1. John Wiley & Sons, New York.

[88] Hwang, F.C. and Yao Y.C. (1989). Multistate consecutively-connected systems. *IEEE Transactions on Reliability*,4,472-474.

[89] Huntington, R. J. and Naus, J.I. (1975). A simpler expression for k-th nearest neighbor coincidence probabilities. *Annals of Probability*, 3, 894-896.

[90] Katona, G. (1973). Two applications (for search theory and truth functions) of Sperner type theorems. *Periodica Mathematica Hungarica*, 3, 19-26.

[91] Klein, M. (2000). Two alternatives to the Shewhart X control chart. *Journal of Quality Technology*, 32, 427-431.

[92] Kleitman, D. and Spencer, J. (1973). Families of k-independent sets. *Discrete Mathematics*, 6, 255-262.

[93] Kotz, S. and Nadarajah, S., (2000). *Extreme Value Distributions: Theory and Applications*. Imperial College Press, London.

[94] Koutras, M.V. (1997). Consecutive-k,r-out-of-n:DFM systems. *Microelectronics and Reliability*, 37, 597-603.

[95] Koutras, M.V. (2003). Applications of Markov chains to the distribution theory of runs and patterns. In Handbook of Statistics, *Stochastic Processes, Modeling and Simulation* (Eds., Shanbhag, D.N. and Rao, C.R.), North Holland Publ. Co.

[96] Koutras, M.V. and Alexandrou, V. A. (1995). Runs, scans and urn model distributions: a unified Markov chain approach. *Annals of the Institute of Statistical Mathematics*, 47, 743-766.

[97] Koutras, M.V. and Alexandrou, V.A. (1997). Non-parametric randomness tests based on success runs of fixed length. *Statistics and Probability Letters*, 32, 393-404.

[98] Koutras, M.V. and Balakrishnan, N. (1999). A start-up demonstration test using a simple scan-based statistic. In *Scan Statistics and Applications* (Eds., Glaz, J. and Balakrishnan, N.), Birkhauser, Boston, 251-267.

[99] Koutras, M.V., Bersimis, S. and Antzoulakos, D. L. (2006). Improving the performance of the chi-square control chart via runs rules. *Methodology and Computing in Applied Probability*, 8, 409-426.

[100] Koutras, M.V., Bersimis, S. and Antzoulakos, D. L. (2008). Bivariate Markov chain embeddable variables of polynomial type. *Annals of the Institute of Statistical Mathematics*, 60, 173-191.

[101] Kruskal, W. H. (1952). A nonparametric test for several sample problems. *Annals of Mathematical Statistics*, 23, 525-540.

[102] Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26, 1481-1496.

[103] Kuo, W. and Zuo, M.J. (2003). *Optimal Reliability Modeling: Principles and Applications*. John Wiley & Sons, NJ.

[104] Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

[105] Lehmann, E. (1955). Ordered families of distributions. *Annals of Mathematical Statistics*, 26, 399-419.

[106] Levitin, G. (2002). Optimal series-parallel topology of multi-state system with two failure modes. *Reliability Engineering and System Safety*, 77, 93-107.

[107] Levitin, G. (2003). Reliability of multi-state systems with two failure modes. *IEEE Transactions on Reliability*, 52, 340-348.

[108] Levitin, G. and Lisnianski, A. (2001). Structure optimization of multi-state system with two failure modes. *Reliability Engineering and System Safety*, 72, 75-89.

[109] Larsen, R.J., Holmes, C.L. and Heath, C.W. (1973). A Statistical Test for Measuring Unimodal Clustering: A Description of the Test and of Its Application to Cases of Acute Leukemia in Metropolitan Atlanta, Georgia. *Biometrics*, 29, 301-309.

[110] Loader, C. (1991). Large deviation approximations to distribution of scan statistics. *Advances in Applied Probability*, 23, 751-771.

[111] Lou, W.Y.W. (1996). On runs and longest run tests: a method of finite Markov chain imbedding. *Journal of the American Statistical Association*, 91, 1595-1601.

[112] Lou, W.Y.W. (1997). An application of the method of finite Markov chain imbedding to runs tests. *Statistics and Probability Letters*, 31, 155-161.

[113] Malinowski, J. and Preuss, W. (1996). Reliability evaluation for tree-structured systems with multistate components. *Microelectronics and Reliability*, 36, 9-17.

[114] Milienos, F.S. and Koutras, M.V. (2008). A lower bound for the reliability function of multiple failure mode systems. *Statistics and Probability Letters*, 78, 1639-1648.

[115] Mood, A.M. (1940). The distribution theory of runs. *Annals of Mathematical Statistics*, 11, 367-392.

[116] Moore, E. and Shannon, C. (1956). Reliable circuits using less reliable relays. *Journal of Franklin Institute*, 9, 191-208.

[117] Mosteller, F. (1941). Note on an application of runs to quality control charts. *Annals of Mathematical Statistics*, 12, 228-232.

[118] Muller, A. and Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*. John Willey & Sons, Chichester.

[119] Naus, J. (1974). Probabilities for a generalized birthday problem. *Journal of the American Statistical Association*, 69, 810-815.

[120] O'Brien, P. C. (1976). A test of randomness. *Biometrics*, 32, 391-401.

[121] O'Brien, P. C. and Dyck, P. J. (1985). A runs test based on run lengths. *Biometrics*, 41, 237-244.

[122] Page, E.S. (1955). Control charts with warning lines. *Biometrika*, 42, 243-257.

[123] Page, L. and Perry, J. (1988). Optimal series-parallel networks of 3-state devices. *IEEE Transactions on Reliability*, 37, 388-394.

[124] Pham, H. and Malon, M.(1994). Optimal design of systems with competing failure modes. *IEEE Transactions on Reliability*, 43, 251-254.

[125] Poisson, S.D. (1837). *Recherches sur la probabilite des jugements en matiere criminelle et en matiere civile, precedees des regles generales du calcul des probabilites*. Bachelier, Paris.

[126] Poljak, S. and Tuza, Z. (1989). On the maximum number of qualitatively independent partitions. *Journal of Combinatorial Theory, Series A*, 51, 111-116.

[127] Poljak, S., Pultr, A. and Rodl, V. (1983). On qualitatively independent partitions and related problems. *Discrete Applied Mathematics*, 6, 193-205.

[128] Pozdnyakov, V., Glaz, J. Kulldorff, M. and Steele, J.M. (2005). A martingale approach to scan statistics. *Annals of the Institute of Statistical Mathematics*, 57, 21-37.

[129] Quirk, P. and Saposnik, R. (1962). Admissibility and measurement utility function. *Review of Economic Studies*, 29, 140-146.

[130] Reiss, R.D. and Thomas, M. (1997). *Statistical Analysis of Extreme Values*. Birkhäuser Verlang, Basel.

[131] Rakitzis, A.C. (2008). Theory of runs and patterns in statistical quality control. Phd Thesis, Department of Statistics and Insurance Science, University of Piraeus, Greece.

[132] Renyi, A. (1971). *Foundations of Probability*. John Willey & Sons, New York.

[133] Ross, S. (1979). Multivalued state component systems. *Annals of Probability*, 7, 379-383.

[134] Roy, R.K (2001). *Design of Experiments Using the Taguchi Approach*. John Wiley & Sons, New York.

[135] Rubin, G., McCulloch, C.E. and Shapiro, M.A. (1990). Multinomial runs test to detect clustering in constrained free recall. *Journal of the American Statistical Association*, 85, 315-320.

[136] Rubinstein, R.Y. and Melamed, B (1998). *Modern Simulation and Modeling*. John Willey & Sons, New York.

[137] Rueda, A. and Pawlak, M. (2004). Pioneers of the reliability theories of the past 50 years. *Reliability and Maintainability, Annual Symposium*, 102-109.

[138] Satoh, N., Sasaki, M., Yuge, T. and Yanasi, S.(1993). Reliability of 3-state device systems with simultaneous failures. *IEEE Transactions on Reliability*, 42, 470-477.

[139] Schwager, S. J. (1983). Run probabilities in sequences of Markov-dependent trials. *Journal of the American Statistical Association*, 78, 168-175.

[140] Szekli, R. (1995). *Stochastic Ordering and Dependence in Applied Probability*. Springer, New York

[141] Shaked, M. and Shanthikumar, J.G. (1994). *Stochastic orders and their applications*. Academic Press, Boston.

[142] Shanthikumar, J.G. (1987). Reliability of systems with consecutive minimal cut sets. *IEEE Transactions on Reliability*, 36, 546-549.

[143] Shaughnessy P.W. (1981). Multiple runs distributions: recurrences and critical values. *Journal of the American Statistical Association*, 76, 732-736.

[144] Swed, F.S., Eisenhart, C. (1943). Tables for testing randomness of grouping in a sequence of alternatives. *The Annals of Mathematical Statistics*, 14, 66-87.

[145] Sloane, N.J.A. (1993). Covering arrays and intersecting codes. *Journal of Combinatorial Designs*, 1, 51-63.

[146] Vance, L.C. and McDonald, G.C. (1979). A class of multiple run sampling plans. *Technometrics*, 21, 141-146.

[147] Veinott, A.F. (1965). Optimal policy in a dynamic, single product, nonstationary inventory model with several demand classes. *Operations Research*, 13, 761-778.

[148] Viveros, R. and Balakrishnan, N. (1993). Statistical inference from start-up demonstration test data. *Journal of Quality Technology*, 25, 119-130.

[149] Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Annals of Mathematical Statistics*, 2, 147-162.

[150] Wolfowitz, J. (1943). On the theory of runs with some applications to quality control. *Annals of Mathematical Statistics*, 14, 280-288.

[151] Zhang, Y.L., Yam, R.C.M. and Zuo, M.J. (2002). Optimal replacement policy for a multistate repairable system. *Journal of the Operational Research Society*, 53, 336-341.