ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΘΕΩΡΙΑ ΡΟΩΝ ΕΠΙΤΥΧΙΩΝ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

Σωτήριος Μπερσίμης

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Υποβλήθηκε στο

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

του Πανεπιστημίου Πειραιώς

Πειραιάς
Νοέμβριος 2005

UNIVERSITY OF PIRAEUS
DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

THEORY OF SUCCESS RUNS WITH APPLICATIONS

Sotirios Bersimis

PhD Thesis

Submitted to

Department of Statistics and Insurance Science

of the University of Piraeus

Piraeus
November 2005

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ολοκλήρωση αυτής της διατριβής αισθάνομαι την ανάγκη να ευχαριστήσω θερμά όσους με κάθε τρόπο με στήριξαν στην πορεία για τη σύνταξή της.

Θα επιθυμούσα να ευχαριστήσω μέσα από την καρδιά μου τον επιβλέποντα καθηγητή Μάρκο Β. Κούτρα, ο οποίος με έμπειρο χέρι, διακριτικότητα και υπομονή με βοήθησε να προσανατολιστώ στο τομέα *της θεωρίας ροών και σχηματισμών* που αρχικά μου ήταν ελάχιστα οικείος, υποδεικνύοντάς μου ταυτόχρονα πλευρές και κατευθύνσεις που διεύρυναν με πολύ γόνιμο τρόπο τη διαπραγμάτευση της διδακτορικής μελέτης. Για όλα αυτά, αλλά και για την εμπιστοσύνη με την οποία με περιέβαλε και στήριξε την προσπάθειά μου από την αρχή μέχρι το τέλος, καθώς και την αδιάλειπτη κατανόησή του, θέλω να του εκφράσω για ακόμη μια φορά τις πλέον θερμές ευχαριστίες μου.

Επίσης, θα επιθυμούσα να ευχαριστήσω πραγματικά τον επίκουρο καθηγητή Δημήτριο Λ. Αντζουλάκο για την αδιάκοπη προσφορά πολύτιμης βοήθειας και συνεργασίας σε όλα τα έτη της εκπόνησης της διδακτορικής διατριβής.

Καθώς, επίσης θα ήθελα να ευχαριστήσω το μέλος της τριμελούς συμβουλευτικής επιτροπής, αναπληρωτή καθηγητή Ευστάθιο Χατζηκωνσταντινίδη, για τη διαρκή του ενθάρρυνση σε όλη τη διάρκεια εκπόνησης της διδακτορικής διατριβής.

Επίσης, είμαι ευγνώμων σε όσους στάθηκαν δίπλα μου κατά τη διάρκεια όλων αυτών των ετών. Γρήγορες σκέψεις με ωθούν να κατονομάσω φίλους και συνεργάτες όπως, τον Δημήτρη Γκίνη, τον Βασίλη Μιχαλακόπουλο, τον Πέτρο Μαραβελάκη, τον Στέλιο Ψαράκη, τον Μιχάλη Σφακιανάκη, καθώς και τον αδελφό μου Φραγκίσκο, τους γονείς μου και φυσικά την Νατάσα για τη συνεχή και αδιάλειπτη στήριξή τους.

Τέλος, ευγνωμοσύνη θα ήθελα να αποδώσω στο Ελληνικό Κράτος που διαμέσου του Υπουργείου Ανάπτυξης και της Γενικής Γραμματείας Έρευνας και Τεχνολογίας, επιχορήγησε τη διδακτορική μου μελέτη μέσω του Προγράμματος Ενίσχυσης Ερευνητικού Δυναμικού (ΠΕΝΕΔ 2001).

ΠΕΡΙΛΗΨΗ

Στη διατριβή αυτή παρουσιάζεται η μελέτη τυχαίων μεταβλητών που σχετίζονται με προβλήματα ροών επιτυχιών, σε ακολουθίες πειραμάτων με δύο ή περισσότερα αποτελέσματα. Η μελέτη αυτή στηρίζεται στην τεχνική της Μαρκοβιανής εμφύτευσης. Συγκεκριμένα, ορίζεται μια νέα και πολύ γενική κατηγορία διακριτών τυχαίων μεταβλητών των οποίων η κατανομή μπορεί να μελετηθεί με τη χρήση κατάλληλης Μαρκοβιανής αλυσίδας και αναπτύσσονται κατάλληλα εργαλεία για τη μελέτη τυχαίων μεταβλητών που ανήκουν σε αυτή την κατηγορία. Στη συνέχεια μελετώνται μονοδιάστατες και πολυδιάστατες μεταβλητές που ανήκουν σε αυτή την κατηγορία. Τέλος, παρουσιάζονται εφαρμογές των τυχαίων μεταβλητών που μελετήθηκαν σε διάφορους επιστημονικούς τομείς.

SUMMARY

The present PhD thesis deals with the study of random variables that are related to the occurrence of successes runs in sequences of experiments with two or more outcomes in each trial. This study is based on the well known Markov chain embedding technique. Concretely, a new category of random variables is defined whose pmf can be evaluated by the aid of appropriate Markov chain models and suitable tools are established for the study of random variables belonging to this new category. Both univariate and multidimensional variables belonging in this class are investigated in detail. Finally, several applications of the general theoretical results in various fields are presented.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ

Ο κύριος σκοπός της διατριβής αυτής είναι η μελέτη τυχαίων μεταβλητών που σχετίζονται με προβλήματα ροών επιτυχιών, σε ακολουθίες πειραμάτων με δύο ή περισσότερα αποτελέσματα. Στο Κεφάλαιο 1 παρουσιάζονται εν συντομία μερικά ιστορικά στοιχεία της γένεσης και της ανάπτυξης της θεωρίας ροών επιτυχιών, μια ανασκόπηση της βιβλιογραφίας και οι τομείς εφαρμογής της θεωρίας ροών επιτυχιών. Στο Κεφάλαιο 2 παρουσιάζουμε τη μέθοδο της Μαρκοβιανής εμφύτευσης. Σύμφωνα με τη μέθοδο αυτή, η συνάρτηση πιθανότητας, οι ροπές και η γεννήτρια πιθανοτήτων μιας εμφυτευμένης τυχαίας μεταβλητής εκφράζονται μέσω των πινάκων μεταπήδησης μιας Μαρκοβιανής αλυσίδας. Στο Κεφάλαιο αυτό ορίζουμε έναν νέο και πολύ γενικό τύπο τυχαίων μεταβλητών και αναπτύσσουμε κατάλληλα εργαλεία για τη μελέτη τους. Στο Κεφάλαιο 3 εφαρμόζουμε τη μέθοδο της Μαρκοβιανής εμφύτευσης σε τυχαίες μεταβλητές που σχετίζονται με το πλήθος και το μήκος ροών επιτυχιών στη μονοδιάστατη περίπτωση, ενώ στο Κεφάλαιο 4, εφαρμόζουμε τη μέθοδο Μαρκοβιανής εμφύτευσης σε τυχαίες μεταβλητές που σχετίζονται με το πλήθος και το μήκος ροών επιτυχιών στην πολυδιάστατη περίπτωση. Στο Κεφάλαιο 5, εφαρμόζουμε τη μέθοδο Μαρκοβιανής εμφύτευσης, σε τυχαίες μεταβλητές που σχετίζονται με χρόνους αναμονής για ροές επιτυχιών στη μονοδιάστατη, αλλά και πολυδιάστατη περίπτωση. Στα Κεφάλαια 6 και 7 παρουσιάζονται υπάρχουσες αλλά και νέες εφαρμογές της θεωρίας ροών επιτυχιών σε διάφορους επιστημονικούς τομείς. Στο σημείο αυτό είναι απαραίτητο να διευκρινίσουμε ότι αν και στη συγκεκριμένη διατριβή παρουσιάζεται ένα μεγάλο μέρος της βιβλιογραφίας στον τομέα της θεωρίας ροών, ο κύριος σκοπός είναι να επικεντρωθούμε στα νέα αποτελέσματα που προέκυψαν στα πλαίσια της διδακτορικής μελέτης. Για μια συνολική παρουσίαση του τομέα αυτού, ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στη μονογραφία των Balakrishnan and Koutras (2002).

ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ

1.1. Εισαγωγή

Ο κύριος σκοπός της διατριβής αυτής είναι η μελέτη τυχαίων μεταβλητών που σχετίζονται με συγκεκριμένους σχηματισμούς, γνωστούς και ως ροές επιτυχιών, σε ακολουθίες πειραμάτων με δύο ή περισσότερα αποτελέσματα. Η μελέτη αυτή επιτυγχάνεται με τη μέθοδο εμφύτευσης τυχαίας μεταβλητής σε Μαρκοβιανή αλυσίδα η οποία εισήχθη από τους Fu and Koutras (1994), βελτιώθηκε από τους Koutras and Alexandrou (1995) και γενικεύεται στην παρούσα διατριβή. Στο παρόν Κεφάλαιο παρουσιάζονται στοιχεία απαραίτητα για την κατανόηση των μεθόδων που χρησιμοποιούνται στα πλαίσια της διατριβής. Πιο συγκεκριμένα, δίνουμε τους ορισμούς και παραδείγματα της έννοιας της ροής επιτυχιών (success run), της έννοιας της γενικευμένης ροής επιτυχιών (generalized success run), καλούμενης και συνάρτηση σάρωσης (scan) καθώς επίσης και της έννοιας του σχηματισμού (pattern). Επίσης, δίνεται μια σύντομη ανασκόπηση της βιβλιογραφίας, στοιχεία καθώς και εφαρμογές των ροών επιτυχιών και των γενικευμένων ροών επιτυχιών σε διάφορους επιστημονικούς κλάδους.

1.2. Η Έννοια της Ροής, της Γενικευμένης Ροής Επιτυχιών και του Σχηματισμού

Ο όρος ροή σε μια ακολουθία δοκιμών, από μία μη εξειδικευμένη οπτική γωνία, αφορά τη διαδοχή (χωρίς διακοπή) όμοιων αποτελεσμάτων. Από στατιστικής απόψεως μπορούμε να ορίσουμε ως ροή τη διαδοχή (χωρίς διακοπή) όμοιων αποτελεσμάτων τα οποία ακολουθούνται και έπονται από διαφορετικά αποτελέσματα σε n (n θετικός ακέραιος) εκτελέσεις ενός πειράματος τύχης, το οποίο σε κάθε επανάληψή του μπορεί

να δώσει δύο δυνατά αποτελέσματα, επιτυχία (1) ή αποτυχία (0). Ο αριθμός k των όμοιων στοιχείων μιας ροής αναφέρεται ως μήκος της ροής (k θετικός ακέραιος).

Για παράδειγμα, έστω ότι εκτελούμε $n=10$ διαδοχικές επαναλήψεις ενός πειράματος τύχης το οποίο αφορά τη ρίψη ενός νομίσματος και συμβολίζουμε με επιτυχία (1) την ένδειξη κεφαλή και με αποτυχία (0) την ένδειξη γράμματα. Αν υποθέσουμε ότι προέκυψε η ακολουθία αποτελεσμάτων 1100011110, μπορούμε εύκολα να παρατηρήσουμε ότι έχουμε διαδοχικά μια ροή μήκους $k=2$ από άσους (1), ακολουθεί μια ροή μήκους $k=3$ από μηδέν (0), στη συνέχεια ακολουθεί μια ροή μήκους $k=4$ από άσους (1), και τέλος ακολουθεί μια ροή μήκους $k=1$ από μηδέν (0).

Στηριζόμενοι στην έννοια της ροής μπορούμε να ορίσουμε μια σειρά από τυχαίες μεταβλητές που αναφέρονται σε n αποτελέσματα ενός πειράματος τύχης με δύο δυνατά αποτελέσματα. Τέτοιες τυχαίες μεταβλητές είναι (α) το μήκος L_n της ροής επιτυχιών που έχει το μέγιστο μήκος, (β) ο αριθμός $N_{n,k}$ των μη επικαλυπτόμενων ροών επιτυχιών μήκους k (Feller (1968)), (γ) ο αριθμός $M_{n,k}$ των επικαλυπτόμενων ροών επιτυχιών μήκους k (Ling (1988)), (δ) ο αριθμός $G_{n,k}$ των ροών επιτυχιών μήκους τουλάχιστον k (Mood (1940)) και (ε) ο αριθμός $E_{n,k}$ των ροών επιτυχιών μήκους ακριβώς k (Mood (1940)). Επίσης μπορούμε να ορίσουμε τυχαίες μεταβλητές που σχετίζονται με το χρόνο αναμονής μέχρι την εμφάνιση της $1^{ης}$ ή γενικότερα της $r^{οστης}$ ροής επιτυχιών οποιουδήποτε είδους.

Στο παράδειγμά μας, η ροή επιτυχιών μέγιστου μήκους, L_{10}, είναι ίση με 4. Ο αριθμός των μη επικαλυπτόμενων ροών επιτυχιών μήκους $k=2$, $N_{10,2}$, είναι ίσος με 3, ενώ, ο αριθμός των επικαλυπτόμενων ροών επιτυχιών μήκους $k=2$, $M_{10,2}$, είναι ίσος με 4. Επίσης, ο αριθμός των ροών επιτυχιών μήκους τουλάχιστον $k=2$, $G_{10,2}$, είναι ίσος με 2. Η $E_{10,2}$, είναι ίση με 1. Ο χρόνος αναμονής $T_{3,1}$ μέχρι την πρώτη εμφάνιση μιας μη επικαλυπτόμενης ροής επιτυχιών μήκους $k=3$, είναι ίσος με 8, ενώ ο χρόνος αναμονής $T_{2,2}$ μέχρι τη δεύτερη εμφάνιση μιας μη επικαλυπτόμενης ροής επιτυχιών μήκους $k=2$, είναι ίσος με 7. Τέλος, το άθροισμα $S_{10,2}$ των μηκών των ροών επιτυχιών μήκους τουλάχιστον $k=2$, είναι ίσο με 6 (η ακριβής κατανομή της $S_{n,k}$ μελετάται για πρώτη φορά στην παρούσα διατριβή).

Τροποποιώντας την έννοια της ροής, μπορούμε να ορίσουμε και να μελετήσουμε πολυπλοκότερα σχήματα, όπως είναι η έννοια της γενικευμένης ροής. Με τον όρο r-γενικευμένη ροή επιτυχιών μήκους k, εννοούμε ότι σε μια ακολουθία n δοκιμών, μας ενδιαφέρει ο αριθμός των παραθύρων μήκους k τα οποία περιέχουν τουλάχιστον r επιτυχίες. Το κύριο ερευνητικό ενδιαφέρον μέχρι σήμερα απέσπασαν οι τυχαίες μεταβλητές σάρωσης $N^{I}_{n,k,r}$, $N^{II}_{n,k,r}$ και $M^{I}_{n,k,r}$, όπου $N^{I}_{n,k,r}$ είναι ο αριθμός των μη επικαλυπτόμενων τμημάτων σάρωσης μεταβλητού μήκους (το πολύ k) που περιλαμβάνουν ακριβώς r επιτυχίες (non-overlapping scans of Type I), $N^{II}_{n,k,r}$ είναι ο αριθμός των μη επικαλυπτόμενων τμημάτων σάρωσης σταθερού μήκους k σε καθένα από τα οποία περιέχονται περισσότερες από r επιτυχίες (non-overlapping scans of Type II), και $M^{I}_{n,k,r}$ είναι ο αριθμός των επικαλυπτόμενων τμημάτων σάρωσης σταθερού μήκους k, σε καθένα από τα οποία έχουν εμφανισθεί περισσότερες από r επιτυχίες (overlapping scans of Type II). Επίσης, έχουν μελετηθεί τυχαίες μεταβλητές που σχετίζονται με το χρόνο αναμονής μέχρι την εμφάνιση του $r^{\text{οστου}}$ τμήματος σάρωσης κάθε είδους.

Στο παράδειγμά μας, αν συμβολίσουμε με X_t, τον αριθμό των επιτυχιών για το παράθυρο (scan) που ξεκινά από τη δοκιμή t, με $k=3$, έχουμε ότι X_1=2, X_2=1, X_3=0, X_4=1, X_5=2, X_6=3, X_7=3, X_8=2. Η τυχαία μεταβλητή $M^{I}_{10,3,2}$, η οποία δίνει τον αριθμό των scans μήκους 3 στα οποία εμφανίζονται τουλάχιστον 2 επιτυχίες, παίρνει την τιμή $M^{I}_{10,3,2}=5$.

Τέλος, γενικεύοντας κάποιες από τις έννοιες των ροών και των συναρτήσεων σάρωσης μπορούμε να διαμορφώσουμε και να μελετήσουμε ακόμα πιο πολύπλοκους σχηματισμούς (patterns). Τέτοιες περιπτώσεις εμφανίζονται συνήθως όταν μελετάμε ακολουθίες n εκτελέσεων ενός πειράματος τύχης το οποίο καταλήγει σε περισσότερα από δύο διαφορετικά αποτελέσματα. Για παράδειγμα, σε $n=10$ διαδοχικές επαναλήψεις ενός πειράματος τύχης με τρία δυνατά αποτελέσματα 0,1,2, προέκυψε η ακολουθία αποτελεσμάτων 1210212120. Παρατηρούμε ότι ο σχηματισμός 12 εμφανίζεται 3 φορές, και ο σχηματισμός 121 εμφανίζεται 2 φορές.

1.3. Ιστορική Αναδρομή

1.3.1. Η Γέννηση της Έννοιας της Ροής Επιτυχιών

Η έννοια της ροής επιτυχιών γεννήθηκε από την προσπάθεια διάσημων μαθηματικών του $17^{ου}$ αιώνα, όπως ο Fermat και ο Pascal, να απαντήσουν σε ερωτήσεις σχετικές με τα τυχερά παιχνίδια. Οι πρωτοπόρες μελέτες της εποχής εκείνης άνοιξαν το δρόμο για τη θεωρητική μελέτη μοντέλων που αναφέρονται σε ακολουθίες πειραμάτων. Έτσι, το $18^{ο}$ αιώνα, οι de Moivre (1738) και Simpson (1740), (βλέπε επίσης Laplace (1812) και Todhunter (1865)) μελέτησαν ανεξάρτητα το πρόβλημα «ποια είναι η πιθανότητα σε n δοκιμές να έχω μια ροή επιτυχιών μήκους τουλάχιστον k». Στα τέλη του 19ου αιώνα, άρχισε η συστηματική μελέτη της έννοιας της ροής επιτυχιών, με άμεσο αποτέλεσμα, η θεωρία αυτή να αναπτυχθεί σε τέτοιο σημείο ώστε το 1920 να μελετάται η ασυμπτωτική συμπεριφορά του μήκους της ροής επιτυχιών.

Όμως, η πρώτη σημαντική εφαρμογή της θεωρίας ροών επιτυχιών εμφανίστηκε με καθυστέρηση δύο αιώνων από την πρώτη μελέτη προβλήματος σχετιζόμενου με τις ροές επιτυχιών (de Moivre (1738)).

Συγκεκριμένα, οι Wald and Wolfowitz (1940) απέδειξαν ότι ο αριθμός των τρόπων που μπορεί να εμφανισθεί συγκεκριμένος αριθμός ροών (ανεξαρτήτως μήκους) σε ακολουθία με δύο είδη συμβόλων (αποτυχία 0 και επιτυχία 1) ακολουθεί ασυμπτωτικά την κανονική κατανομή. Στηριζόμενοι σε αυτό το ασυμπτωτικό αποτέλεσμα όρισαν και μελέτησαν το γνωστό και ευρέως χρησιμοποιούμενο απαραμετρικό κριτήριο τυχαιότητας των ροών.

Στις αρχές της δεκαετίας του 1980, ανανεώνεται το ενδιαφέρον για τυχαίες μεταβλητές σχετικές με ροές επιτυχιών και πολλοί ερευνητές επεκτείνουν και εξετάζουν διεξοδικά τις έννοιες αυτές. Συγκεκριμένα, σε μια προσπάθεια ομαδοποίησης των κατανομών που σχετίζονται με τις ροές επιτυχιών εισάγεται ο όρος διωνυμικές κατανομές k τάξης (Binomial Distributions of Order k). Έτσι παρουσιάζονται, υπό την κοινή αυτή ονομασία, οι τυχαίες μεταβλητές $N_{n,k}$, η $M_{n,k}$, η $G_{n,k}$, καθώς και η $E_{n,k}$.

1.3.2. Μελέτη Κατανομών που Σχετίζονται με Ροές Επιτυχιών

Η μελέτη των διωνυμικών κατανομών k τάξης γίνεται, στις αρχές της δεκαετίας του 1980, μέσω συνδυαστικών τεχνικών και οι τύποι που εξάγονται για τις συναρτήσεις πιθανότητας περιλαμβάνουν πολλαπλά αθροίσματα σε περιοχές που προκύπτουν από τη λύση διοφαντικών εξισώσεων. Έτσι, οι Hirano (1986) και Philippou and Makri (1986) δουλεύοντας ανεξάρτητα έδωσαν τον ίδιο περίπου κλειστό τύπο για τον υπολογισμό της κατανομής της τυχαίας μεταβλητής $N_{n,k}$. Η δυσκολία υπολογισμού της τιμής των πιθανοτήτων μέσω εκφράσεων της μορφής αυτής, οδήγησε τους ερευνητές σε αναζήτηση εναλλακτικών τρόπων προσέγγισης του θέματος. Στις περισσότερες μετέπειτα εργασίες (Aki and Hirano (1988), Chryssaphinou and Papastavridis (1988, 1990), Chryssaphinou et al. (1993), Godbole (1990,1991,1992), Hirano et al. (1991), Hirano and Aki (1993), Antzoulakos and Chadjiconstantinidis (2001)), δόθηκε έμφαση στην εξαγωγή απλών, συνήθως αναδρομικών εκφράσεων για τις συναρτήσεις πιθανότητας.

Οι Chao and Fu (1989) μελετώντας το πρόβλημα της αξιοπιστίας ενός συστήματος, το οποίο παύει να λειτουργεί όταν τουλάχιστον k διαδοχικά στοιχεία από τα n στοιχεία του δεν λειτουργούν, ανέπτυξαν έναν νέο τρόπο προσέγγισης προβλημάτων που σχετίζονται με ροές επιτυχιών χρησιμοποιώντας εργαλεία από τη θεωρία των αλυσίδων Markov. Οι Fu and Koutras (1994), γενικεύοντας την τεχνική αυτή ώστε να προκύπτει ολόκληρη η κατανομή (αντί της ουράς, που απαιτούσε η θεωρία αξιοπιστίας) πρότειναν ακριβείς εκφράσεις για τις διωνυμικές κατανομές k τάξης. Η τεχνική αυτή αποτέλεσε σημαντικό εργαλείο για τη μελέτη μιας ευρύτατης οικογένειας τυχαίων μεταβλητών, την οικογένεια των τυχαίων μεταβλητών διωνυμικού τύπου εμφυτεύσιμων σε Μαρκοβιανή αλυσίδα που εισήχθησαν από τους Koutras and Alexandrou (1995).

Στην παρούσα διατριβή, εισάγεται για πρώτη φορά η έννοια της τυχαίας μεταβλητής εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα πολυωνυμικού τύπου (ως γενίκευσης του διωνυμικού τύπου). Η έννοια αυτή επιτρέπει τον εύκολο υπολογισμό της ακριβούς κατανομής της τυχαίας μεταβλητής $S_{n,k}$ η οποία καταγράφει το άθροισμα των μηκών των ροών επιτυχιών των οποίων το μήκος είναι τουλάχιστον k (βλέπε επίσης και Antzoulakos et al. (2003)). Μέχρι σήμερα η συγκεκριμένη τυχαία μεταβλητή

έχει μελετηθεί μόνο σε επίπεδο ασυμπτωτικής κατανομής (Wang (2001), Fu et al (2002)).

1.3.3. Πολυδιάστατες Γενικεύσεις Διωνυμικών Κατανομών *k* Τάξης

Οι Philippou et al. (1989), Philippou et al. (1990), Philippou and Antzoulakos (1990) καθώς και άλλοι ερευνητές, στην προσπάθειά τους να γενικεύσουν τις σχετικές με ροές επιτυχιών κατανομές, εισάγουν τις πολυδιάστατες κατανομές k τάξης. Οι μέθοδοι που εφαρμόζουν είναι συνήθως συνδυαστικές και σε αρκετές περιπτώσεις χρησιμοποιούν γενικευμένα πολυώνυμα Fibonacci.

Το συνεχώς αυξανόμενο ενδιαφέρον για τη ανάλυση συγκεκριμένων εφαρμογών (στατιστικός έλεγχος ποιότητας, θεωρία αξιοπιστίας, απαραμετρικά κριτήρια τυχαιότητας, οικολογία, μετεωρολογία, αλυσίδες DNA) ώθησε στη γενίκευση της έννοιας των πολυδιάστατων διωνυμικών κατανομών τάξης k προς πολλές κατευθύνσεις.

Μια κατεύθυνση πολυδιάστατης γενίκευσης των διωνυμικών κατανομών τάξης k αφορά διδιάστατες και πολυδιάστατες μεταβλητές οι οποίες απαριθμούν ροές επιτυχιών με διαφορετικά μήκη. Οι Godbole et al. (1997) γενικεύουν τις διωνυμικές κατανομές τάξης k, εισάγοντας τις πολυδιάστατες τυχαίες μεταβλητές που απαριθμούν τις ροές επιτυχιών με διάφορα μήκη.

Μια άλλη κατεύθυνση, αφορά διδιάστατες και πολυδιάστατες μεταβλητές οι οποίες απαριθμούν ροές τόσο επιτυχιών όσο και αποτυχιών. Έτσι, οι Ling and Tai (1990) εισάγουν για πρώτη φορά πολυδιάστατες κατανομές που έχουν τη μορφή αυτή μελετώντας μόνο την ειδική περίπτωση του αριθμού των (μη επικαλυπτόμενων και επικαλυπτόμενων) ροών αποτυχίας και επιτυχίας μήκους 2 σε ακολουθίες δίτιμων δοκιμών. Στην ίδια κατεύθυνση εργάζονται και οι Chadjiconstantinidis et al. (2000).

Μια άλλη κατεύθυνση, αφορά διδιάστατες και πολυδιάστατες μεταβλητές οι οποίες απαριθμούν ροές επιτυχιών σε ακολουθίες δοκιμών με περισσότερα των δύο αποτελεσμάτων. Στη μονοδιάστατη περίπτωση, οι Guibas and Odlyzko (1980, 1981) θεωρώντας ένα αλφάβητο με n γράμματα ανέπτυξαν μια γενική μέθοδο για τη μελέτη απαριθμητριών σχηματισμών γραμμάτων, στις οποίες δεν περιλαμβάνονται κάποια

συγκεκριμένα σχήματα (patterns). Τόσο οι Guibas and Odlyzko (1980, 1981) όσο και οι Aki (1992) και Ling and Low (1993) προτείνουν γενικές εκφράσεις για τις γεννήτριες πιθανοτήτων τυχαίων μεταβλητών που σχετίζονται με το χρόνο αναμονής μέχρι την εμφάνιση ενός σχηματισμού. Ο Fu (1996) επέκτεινε τη μεθοδολογία της εμφύτευσης τυχαίων μεταβλητών σε Μαρκοβιανή αλυσίδα και για τον υπολογισμό της ακριβούς κατανομής σχηματισμών (patterns), ενώ, οι Johnson and Fu (1999) μελετούν με χρήση Μαρκοβιανών αλυσίδων την κατανομή σχηματισμών ανοδικών ροών σε ακολουθίες πολλαπλών αποτελεσμάτων. Οι Aki and Hirano (2004) μελέτησαν χρόνους αναμονής σχετικούς με διδιάστατους σχηματισμούς (patterns). Επιστρέφοντας στην πολυδιάστατη περίπτωση, οι Alexandrou (1997) και οι Han and Aki (1999) μελετούν πολυδιάστατες κατανομές με χρήση της μεθοδολογίας της εμφύτευσης τυχαίων μεταβλητών σε Μαρκοβιανή αλυσίδα. Στην παρούσα διατριβή μελετώνται πολυδιάστατες τυχαίες μεταβλητές με τη γενίκευση της παραπάνω μεθόδου.

Ο Han (2001) δίνει μια νέα κατεύθυνση απαριθμώντας ροές επιτυχιών σε περισσότερες της μίας, εξαρτημένες ακολουθίες δοκιμών, με περισσότερα των δύο αποτελεσμάτων, κάνοντας χρήση της μεθοδολογίας της εμφύτευσης.

Στην παρούσα διατριβή, εισάγεται για πρώτη φορά η έννοια της πολυδιάστατης τυχαίας μεταβλητής εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα πολυωνυμικού τύπου (ως γενίκευσης του διωνυμικού τύπου). Η έννοια αυτή επιτρέπει τον εύκολο υπολογισμό της ακριβούς κατανομής της διδιάστατης τυχαίας μεταβλητής $\left(S_{n,k}^{(S)}, N_{n,r}^{(F)}\right)$ η οποία καταγράφει ταυτόχρονα το άθροισμα των μηκών των ροών επιτυχιών των οποίων το μήκος είναι τουλάχιστον k και τον αριθμό των μη επικαλυπτόμενων ροών μήκους r (βλέπε επίσης και Koutras et al. (2005b)).

1.3.4. Προβλήματα Χρόνου Αναμονής Σχετιζόμενα με Ροές Επιτυχιών

Ανατρέχοντας στη βιβλιογραφία, μπορούμε να διαπιστώσουμε ότι η κατανομή του χρόνου αναμονής (ή των γεωμετρικών κατανομών k τάξης όπως ονομάζονται αλλιώς) στη γενική της μορφή έχει απασχολήσει πολλούς ερευνητές λόγω της πληθώρας των εφαρμογών που βρίσκει.

Η θεωρητική μελέτη της κατανομής του χρόνου αναμονής μέχρι την εμφάνιση ροών επιτυχιών ξεκινά από τον Feller (1968) μέσω της θεωρίας των ανανεούμενων ενδεχόμενων. Οι Philippou and Muwafi (1982), οι Philippou et al. (1983) και οι Uppuluri and Patil (1983) όρισαν με γενικό τρόπο και μελέτησαν τις γεωμετρικές κατανομές k τάξης. Αργότερα, οι Ebneshahrashoob and Sobel (1990), προτείνουν τρόπους υπολογισμού των πιθανογεννητριών για τα προβλήματα της ταχύτερης και της βραδύτερης ροής (sooner/later problem). Συγκεκριμένα, ασχολούνται με το χρόνο αναμονής, α) μέχρι την πρώτη εμφάνιση μιας ροής επιτυχιών μήκους k ή μιας ροής αποτυχιών μήκους r, όποια εμφανιστεί πρώτη (πρόβλημα ταχύτερης ροής) ή β) μέχρι να εμφανισθούν ροές και των δύο ειδών (πρόβλημα βραδύτερης ροής).

Οι Aki and Hirano (1993) και οι Balasubramanian et al. (1993) εξετάζουν τα ίδια προβλήματα, όταν οι δοκιμές έχουν Μαρκοβιανή εξάρτηση πρώτης τάξης. Ακόμη, οι Aki (1992) και οι Ling and Low (1993) προτείνουν γενικές εκφράσεις για τις γεννήτριες πιθανοτήτων τυχαίων μεταβλητών σχετιζόμενων με χρόνους αναμονής, ενώ επιπλέον, οι Chryssaphinou et al. (1994) θεωρούν ένα αλφάβητο με n διαφορετικά γράμματα και μελετούν το χρόνο αναμονής μέχρι την εμφάνιση ενός συγκεκριμένου σχηματισμού (όπως και οι Guibas and Odlyzko (1980, 1981)). Οι Uchida and Aki (1995), μελετούν το χρόνο αναμονής μέχρι τη n-οστή εμφάνιση μιας ροής επιτυχιών μήκους k ή τη m-οστή εμφάνιση μιας ροής αποτυχιών μήκους r υποθέτοντας ότι οι υπό εξέταση ακολουθίες αποτελούνται από δίτιμες δοκιμές που παρουσιάζουν Μαρκοβιανή εξάρτηση πρώτης τάξης.

Η μεθοδολογία της Μαρκοβιανής εμφύτευσης χρησιμοποιήθηκε και για τον υπολογισμό της ακριβούς κατανομής τυχαίων μεταβλητών που αφορούν το χρόνο αναμονής. Έτσι, οι Koutras (1996b, 1997a, 1997b), Koutras and Alexandrou (1997b), Antzoulakos (1999), Antzoulakos (2001), έχουν ως αντικείμενο μελέτης το χρόνο αναμονής μέχρι την εμφάνιση συγκεκριμένων σχηματισμών σε ακολουθίες ανεξάρτητων ή Μαρκοβιανά εξαρτημένων δοκιμών.

Στην παρούσα διατριβή, μελετάται ο χρόνος αναμονής μιας τυχαίας μεταβλητής η οποία σχετίζεται με το άθροισμα των μηκών των ροών επιτυχιών των οποίων το μήκος είναι τουλάχιστον k (βλέπε επίσης και Antzoulakos et al. (2004)).

1.3.5. Ροές Επιτυχιών σε Μοντέλα Καταλήψεων

Οι ροές επιτυχιών σχετίζονται με μια σειρά από μοντέλα καταλήψεων. Για παράδειγμα, εάν από μια κάλπη η οποία περιέχει a λευκές και b μαύρες σφαίρες, εξάγουμε τυχαία n σφαίρες, μια προς μια, χωρίς επανάθεση τότε η κατανομή του αριθμού $N_{n,k}^{*}$ των μη επικαλυπτόμενων k-άδων από λευκές σφαίρες ονομάζεται υπεργεωμετρική κατανομή τάξης k (Hypergeometric distribution of order k). Τέτοια μοντέλα μελέτησαν, κύρια οι Panaretos and Xekalaki (1986), Aki and Hirano (1988), και Godbole (1990b).

Στη βιβλιογραφία (Tripsiannis (1993), Tripsiannis and Philippou (1997a,1997b)) εμφανίζεται επίσης, ο όρος κατανομή Polya τάξης k (Polya distribution of order k). Η κατανομή αυτή προκύπτει όταν στο παραπάνω μοντέλο, κάθε σφαίρα επιστρέφεται στην κάλπη μαζί με c σφαίρες του ίδιου χρώματος, πριν από την επόμενη εξαγωγή σφαίρας. Η ειδική περίπτωση $c=1$ αναφέρεται ως αρνητική υπεργεωμετρική κατανομή τάξης k (Negative Hypergeometric distribution of order k). Ενώ στην περίπτωση όπου εκτός από τις c σφαίρες του ίδιου χρώματος, προσθέτουμε στην κάλπη και d σφαίρες του αντίθετου χρώματος ορίζεται η κατανομή Friedman τάξης k, η οποία βασίζεται στην απλή κατανομή Friedman (Friedman (1949), Friedman (1965)).

1.3.6. Τυχαίες Μεταβλητές Σάρωσης

Ένα άλλο πεδίο όπου υπάρχει σημαντικό ερευνητικό ενδιαφέρον τα τελευταία χρόνια αποτελούν οι τυχαίες μεταβλητές σάρωσης με τις οποίες συνδέεται μεγάλος αριθμός εφαρμογών. Η επίλυση πολλών προβλημάτων βιολογίας, αλυσίδων DNA, στατιστικού έλεγχου ποιότητας, και άλλων, βασίζεται στη μελέτη του αριθμού των διακριτών τμημάτων σάρωσης (scans) μήκους k που περιλαμβάνουν περισσότερες από r επιτυχίες (Glaz (1989), Glaz and Naus (1991), Wallenstein et al. (1994)).

Η ακριβής κατανομή των στατιστικών συναρτήσεων $N_{n,k,r}^{I}$, $N_{n,k,r}^{II}$, $M_{n,k,r}^{I}$, εκτός από κάποιες ειδικές περιπτώσεις, είναι άγνωστη, ιδιαίτερα όταν οι δοκιμές της ακολουθίας που εξετάζουμε είναι μη ισόνομες. Αυτός είναι ο λόγος για τον οποίο

δόθηκε ιδιαίτερη έμφαση στην εύρεση προσεγγιστικών εκφράσεων των κατανομών των παραπάνω τυχαίων μεταβλητών. Έτσι, ο Glaz (1983) πρότεινε φράγματα για το χρόνο αναμονής μέχρι την εμφάνιση ενός τμήματος μήκους r με περισσότερες από k επιτυχίες (για ανεξάρτητες ή Μαρκοβιανά εξαρτημένες δοκιμές). Οι Dembo and Karlin (1992) και Karlin and Macken (1991), χρησιμοποιώντας τη μεθόδο Chen-Stein, ανέπτυξαν προσεγγίσεις της κατανομής των τυχαίων μεταβλητών σάρωσης από την κατανομή Poisson, όταν οι δοκιμές είναι ανεξάρτητες τυχαίες μεταβλητές που παίρνουν θετικές ακέραιες τιμές. Το συνεχές ανάλογο των μεταβλητών σάρωσης μελετήθηκε από τους Huntington (1978), Naus (1982) και άλλους.

1.4. Χρήσεις των Ροών Επιτυχιών και των Γενικευμένων Ροών

Στην παράγραφο αυτή δίνουμε μια συνοπτική ανασκόπηση των κυριοτέρων εφαρμογών της θεωρίας ροών επιτυχιών. Οι εφαρμογές αυτές αναλύονται εκτενέστερα στα Κεφάλαια 6 και 7.

1.4.1. Απαραμετρικοί Έλεγχοι Τυχαιότητας

Όπως προαναφέρθηκε, η πρώτη εφαρμογή της θεωρίας των ροών είναι οι απαραμετρικοί έλεγχοι τυχαιότητας, οι οποίοι αποτελούν ένα σημαντικό κομμάτι της στατιστικής. Από τους ελέγχους αυτούς, το πλέον γνωστό και εύκολα εφαρμόσιμο κριτήριο ελέγχου της τυχαιότητας μιας ακολουθίας με δύο τύπους συμβόλων είναι το κλασικό κριτήριο των ροών, το οποίο βασίζεται στον αριθμό των ροών ανεξαρτήτως του μήκους των (Gibbons (1971)).

Ο Mosteller (1941) πρότεινε έναν έλεγχο τυχαιότητας βασισμένο στο μήκος της μέγιστης ροής (μεγάλες ροές αποτελούν ισχυρές ενδείξεις έλλειψης τυχαιότητας). Ο Mood (1940) σε μια ιδιαίτερα σημαντική μελέτη πάνω στη θεωρία των ροών έδωσε μια πλήρη σειρά από τύπους σε σχέση με τις ροές επιτυχιών.

Αργότερα, οι O'Brian (1976), O'Brian and Dyck (1985) παρουσίασαν ένα κριτήριο τυχαιότητας που λαμβάνει υπόψη και τη διασπορά του μήκους των ροών, και οι Agin

and Godbole (1992) πρότειναν έναν έλεγχο τυχαιότητας που χρησιμοποιεί τον αριθμό των ροών επιτυχιών προκαθορισμένου μήκους.

Τα προαναφερθέντα κριτήρια βασίζονται στη δεσμευμένη κατανομή του αριθμού των ροών ή του μήκους της μέγιστης ροής επιτυχιών, όταν όμως είναι γνωστός ο συνολικός αριθμός των επιτυχιών στο δείγμα των n δοκιμών. Οι Koutras and Alexandrou (1997) πρότειναν μια σειρά από μη παραμετρικούς έλεγχους τυχαιότητας στηριζόμενοι στους γνωστότερους τύπους ροών επιτυχιών ($N_{n,k}, M_{n,k}, G_{n,k}$), συγκρίνοντας την ισχύ των ελέγχων αυτών.

Στην παρούσα διατριβή, δίνουμε έναν νέο μη παραμετρικό έλεγχο τυχαιότητας (βλέπε επίσης και Antzoulakos et al. (2003)) και ελέγχουμε την ισχύ του, συγκρίνοντάς την με την αντίστοιχη ισχύ του βέλτιστου ελέγχου των Koutras and Alexandrou (1997).

1.4.2. Θεωρία Αξιοπιστίας Συστημάτων

Οι διωνυμικές κατανομές τάξης k βρίσκουν άμεση εφαρμογή στη μελέτη της αξιοπιστίας των διαδοχικών-k-από-τα-n: F συστημάτων (consecutive-k-out-of-n: F systems) όπως παρατηρήθηκε από τον Philippou (1986). Ένα τέτοιο σύστημα αποτελείται από n μονάδες και παύει να λειτουργεί αν τουλάχιστον k διαδοχικές μονάδες του χαλάσουν.

Μπορούμε να εκφράσουμε την αξιοπιστία ενός τέτοιου συστήματος με τη βοήθεια μιας διωνυμικής κατανομής τάξης k, εάν θεωρήσουμε ως επιτυχία την εμφάνιση χαλασμένης μονάδας και ως αποτυχία την εμφάνιση μονάδας που λειτουργεί κανονικά (βλέπε Makri and Philippou (1996)). Έτσι, η αξιοπιστία του συστήματος εκφράζεται ως

$$R_{n,k} = P(N_{n,k} = 0) = P(M_{n,k} = 0) = P(G_{n,k} = 0).$$

Εκτενής αναφορά σε θέματα αξιοπιστίας δίνεται στην επισκόπηση των Chao et al. (1995) και στην εργασία του Koutras (1996a).

Στην παρούσα διατριβή περιγράφεται ένα σύνθετο μοντέλο αξιοπιστίας το οποίο βασίζεται σε παραλλαγή των διαδοχικών-k-από-τα-n: F συστημάτων (βλέπε επίσης και Koutras et al. (2005b)).

1.4.3. Πολυμεταβλητή Στατιστική Ανάλυση

Δύο από τις πλέον ενδιαφέρουσες περιοχές της πολυμεταβλητής στατιστικής ανάλυσης (κυρίως λόγω των πολλών εφαρμογών που βρίσκουν στις εφαρμοσμένες επιστήμες όπως ψυχολογία, βιοϊατρική, βιολογία, αρχαιολογία κ.α.) είναι η Διαχωριστική Ανάλυση (Discriminant analysis) και η Ανάλυση Συστάδων (Cluster Analysis). Το κύριο χαρακτηριστικό σημείο και των δύο αυτών μεθόδων είναι η χρησιμοποίηση μέτρων ομοιότητας ή απόστασης ως κριτήρια ομαδοποίησης ή διαχωρισμού.

Στην παρούσα διατριβή, δίνουμε ένα νέο μέτρο απόστασης μεταξύ πολυμεταβλητών παρατηρήσεων το οποίο κάνει χρήση των καμπύλων Andrews και τη θεωρία ροών επιτυχιών (βλέπε επίσης Vassiliou et al. (2004)). Στη συνέχεια, το νέο αυτό μέτρο απόστασης, ενσωματώθηκε σε ήδη υπάρχοντα αλγόριθμο Cluster Analysis ως εναλλακτική της Ευκλείδειας απόστασης, με αποτέλεσμα την κατά μεγάλο ποσοστό βελτίωση του αλγορίθμου. Από όσο γνωρίζουμε, στη βιβλιογραφία, δεν υπάρχουν άλλες αναφορές εφαρμογών της θεωρίας ροών επιτυχιών στην πολυμεταβλητή στατιστική ανάλυση.

1.4.4. Στατιστικός Έλεγχος Ποιότητας

Ένας πολύ μεγάλος αριθμός εφαρμογών της θεωρίας των ροών εμφανίζεται στο τομέα της βιομηχανίας και συγκεκριμένα στον Στατιστικό Έλεγχο Ποιότητας (Statistical Quality Control).

Οι ροές και οι γενικευμένες ροές επιτυχιών εμφανίζονται κατά κόρον στην βιβλιογραφία του Στατιστικού Ποιοτικού Ελέγχου (Western Electric Company (1956), Walker et al. (1991), Champ και Woodall (1987), Champ και Woodall (1990), Palm (1990), Lowry et al. (1995), Divoky and Taylor (1995), Champ και Woodall (1997), Klein (2000), Shmueli and Cohen (2000), Fu et al (2002), Rakitzis (2004), Aparisi et al. (2004)).

Στην παρούσα διατριβή δίνεται ιδιαίτερη έμφαση στο συγκεκριμένο θέμα και εισάγεται μια νέα έκδοση του διαγράμματος ελέγχου Chi-Square για πολυμεταβλητές

διεργασίες, κάνοντας χρήση θεωρίας ροών επιτυχιών με αποδοτικό τρόπο, το οποίο παρουσιάζει αυξημένη ευαισθησία στην ανίχνευση μικρών αλλαγών του διανύσματος των μέσων (βλέπε επίσης Koutras et al. (2005a)).

1.4.5. Ψυχολογία

Η ψυχολογία αποτελεί ένα ακόμη πεδίο εφαρμογής της θεωρίας ροών επιτυχιών. Η πρώτη εφαρμογή παρουσιάστηκε από τον Grant (1946), και αφορούσε ένα κριτήριο αξιολόγησης μαθησιακών δραστηριοτήτων. Άλλες αναφορές έχουμε από τους Child (1946), Grant (1946, 1947) και Bogartz (1965).

Στην παρούσα διατριβή περιγράφεται ένα σύνθετο μοντέλο αξιολόγησης μαθησιακών δραστηριοτήτων το οποίο βασίζεται σε παραλλαγή του αντίστοιχου κριτηρίου που προτάθηκε από τον Grant (1946).

1.4.6. Άλλες Εφαρμογές

Η θεωρία ροών επιτυχιών παρέχει μια μεγάλη ποικιλία από εργαλεία τα οποία χρησιμοποιούνται από ερευνητές διάφορων άλλων επιστημονικών τομέων.

Ένας τέτοιος επιστημονικός τομέας είναι αυτός που μελετά τις αλυσίδες DNA. Στη μελέτη ακολουθιών του γενετικού υλικού DNA (οι οποίες περιλαμβάνουν 4 είδη συμβόλων-βάσεων) εξετάζονται τμήματα της ακολουθίας. Ανάλογα με την περιεκτικότητα σε συγκεκριμένες βάσεις και την περιοδικότητα ή την τυχαιότητα με την οποία εμφανίζονται διάφοροι σχηματισμοί, εξάγονται συμπεράσματα για το χαρακτηρισμό του είδους ή τις ιδιότητες του είδους (Dembo and Karlin (1992), Goldstein (1990), Karlin and Cardon (1994), Karlin and Macken (1991)).

Άλλος επιστημονικός τομέας εφαρμογής της θεωρίας ροών επιτυχιών είναι ο τομέας της οικονομικής ανάπτυξης και της ανταγωνιστικότητας. Ένα από τα βασικότερα εργαλεία της οικονομικής ανάλυσης είτε σε κρατικό/διακρατικό επίπεδο είτε σε επίπεδο επιχείρησης είναι οι χρονοσειρές (Time Series). Στην περιοχή αυτή έχουν αναπτυχθεί διάφορες στατιστικές τεχνικές με στόχο τη μελλοντική πρόβλεψη των υπό μελέτη ποσοτήτων. Η ακολουθιακή ύφη μιας χρονοσειράς και η εμφανής σημασία της

ανίχνευσης πιθανής συγκέντρωσης κάποιων προκαθορισμένων κατηγοριών αποτελεσμάτων, δίνουν σαφείς ενδείξεις ότι η θεωρία ροών επιτυχιών μπορεί να χρησιμοποιηθεί αποτελεσματικά για οικονομικές αναλύσεις και διερεύνηση ύπαρξης συστηματικών τάσεων. Παρόμοιες τεχνικές φαίνονται να προσφέρουν ικανοποιητικά στατιστικά υποδείγματα για μοντέλα ιδιωτικής ή δημόσιας ασφάλισης όπου το ρόλο της επιτυχίας/αποτυχίας παίζει η υπέρβαση/μη υπέρβαση ενός προκαθορισμένου ορίου του ύψους αποζημίωσης/επιδόματος/σύνταξης που καταβάλλεται σε έναν ασφαλισμένο. Μερικά προκαταρκτικά αποτελέσματα στον τομέα αυτό αναλύονται σε μια πρόσφατη εργασία των Boutsikas and Koutras (2001), ενώ για μια γενική θεώρηση παραπέμπουμε στους Binswanger and Embrechts (1994) και τη μονογραφία των Embrechts, Kluppelberg and Mikosch (1997).

Εφαρμογές της θεωρίας ροών επιτυχιών στην οικολογία και ιδιαίτερα στον έλεγχο ή την πρόβλεψη της εξάπλωσης ασθενειών, έχουμε από τους Pielou (1962, 1963a, 1963b, 1977) και Knight (1974). Εφαρμογές των ροών στη μετεωρολογία έχουν γίνει από τους Cochran (1938), Nair (1942), Gabriel and Neumann (1962), και Sen (1980) και αφορούν μοντέλα συνάφειας των καιρικών συνθηκών σε γειτονικές περιοχές.

Τέλος, εφαρμογές της θεωρίας ροών επιτυχιών μπορούμε να βρούμε στη βιομηχανία, εκτός από το πεδίο του ποιοτικού ελέγχου, και στους λεγόμενους έλεγχους εκκινήσεως (Start-Up tests), με τα οποία οι μηχανικοί επιλέγουν εξοπλισμό προς αγορά. Αυτοί στηρίζονται στις επαναλαμβανόμενες απόπειρες για εκκίνηση του δοκιμαζόμενου μηχανήματος. Μια λογική σκέψη είναι, η αποδοχή ή απόρριψη ενός μηχανήματος να γίνεται με βάση τον αριθμό των συνεχόμενων επιτυχών εκκινήσεων του μηχανήματος (Hahn and Gage (1983), Viveros and Balakrishnan (1993), Balakrishnan et al. (1995, 1997)).

1.5. Ανακεφαλαίωση

Στο κεφάλαιο αυτό δώσαμε μια σύντομη ανασκόπηση της βιβλιογραφίας που αφορά τη θεωρία ροών επιτυχιών καθώς και ένα μεγάλο μέρος από τις εφαρμογές της θεωρίας ροών σε διάφορα επιστημονικά πεδία άλλα και σε άλλους τομείς της στατιστικής επιστήμης.

ΚΕΦΑΛΑΙΟ 2: Η ΜΕΘΟΔΟΣ ΤΗΣ ΜΑΡΚΟΒΙΑΝΗΣ ΕΜΦΥΤΕΥΣΗΣ

2.1. Εισαγωγή

Ο κύριος στόχος της διατριβής αυτής, όπως αναφέρθηκε και στο εισαγωγικό κεφάλαιο, είναι η μελέτη μιας οικογένειας τυχαίων μεταβλητών που σχετίζονται με συγκεκριμένους σχηματισμούς σε ακολουθίες πειραμάτων με δύο ή περισσότερα δυνατά αποτελέσματα. Η μελέτη αυτή στηρίζεται στη μέθοδο εμφύτευσης τυχαίας μεταβλητής σε Μαρκοβιανή αλυσίδα. Προκειμένου να είναι δυνατή η μελέτη των τυχαίων μεταβλητών που μας ενδιαφέρουν, κατέστη αναγκαία η κατάλληλη γενίκευση της έννοιας της τυχαίας μεταβλητής διωνυμικού τύπου εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα (MVB), η οποία εισήχθη από τους Koutras and Alexandrou (1995). Για την καλύτερη κατανόηση της μεθόδου εμφύτευσης τυχαίας μεταβλητής σε Μαρκοβιανή αλυσίδα θα παρουσιάσουμε τη μέθοδο διαχρονικά και την εξέλιξή της βήμα προς βήμα.

Στο κεφάλαιο αυτό, αρχικά, θα παρουσιάσουμε τη μέθοδο της Μαρκοβιανής εμφύτευσης όπως εισήχθη από τους Fu and Koutras (1994). Με την μέθοδο αυτή η συνάρτηση πιθανότητας, οι ροπές και η γεννήτρια πιθανοτήτων μιας τυχαίας μεταβλητής εκφράζονται μέσω πινάκων πιθανοτήτων μεταπήδησης μιας κατάλληλης Μαρκοβιανής αλυσίδας. Η υπολογιστική δυσκολία που παρουσιάζεται κατά την εφαρμογή των τύπων που προκύπτουν με την μέθοδο της Μαρκοβιανής εμφύτευσης όπως ορίστηκε από τους Fu and Koutras (1994), είναι οι πολλαπλασιασμοί μεταξύ πινάκων, οι οποίοι έχουν διάσταση που εξαρτάται από τον αριθμό των δοκιμών της ακολουθίας πειραμάτων. Αυτό έχει ως αποτέλεσμα τη χρήση πινάκων μεγάλης διάστασης όταν η ακολουθία αποτελείται από μεγάλο αριθμό δοκιμών.

Η δυσκολία αυτή αντιμετωπίσθηκε από τους Koutras and Alexandrou (1995) με τον ορισμό της έννοιας της τυχαίας μεταβλητής διωνυμικού τύπου εμφυτεύσιμης σε

Μαρκοβιανή αλυσίδα (MVB). Έτσι, στην συνέχεια του κεφαλαίου αυτού παρουσιάζεται η τεχνική της εμφύτευσης όπως προτάθηκε από τους Koutras and Alexandrou (1995) και η οποία αποτελεί πολύτιμο εργαλείο για την εύρεση της ακριβούς κατανομής τυχαίων μεταβλητών που σχετίζονται με μεγάλη κλάση προβλημάτων.

Τέλος, στο κεφάλαιο αυτό εισάγουμε την έννοια της τυχαίας μεταβλητής πολυωνυμικού τύπου εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα (MVP) (βλέπε επίσης και Antzoulakos et al. (2003)) προκειμένου να αντιμετωπισθεί μια ευρύτερη κλάση προβλημάτων σε σχέση με τις μεταβλητές MVB και στη συνέχεια τη γενικεύουμε δίνοντας τον ορισμό της διδιάστατης και της πολυδιάστατης τυχαίας μεταβλητής πολυωνυμικού τύπου εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα (MVP) (βλέπε επίσης και Koutras et al. (2005b)). Για τις τυχαίες μεταβλητές τύπου MVP δίνονται αναδρομικές σχέσεις για τον υπολογισμό της συνάρτησης πιθανότητας, μελετάται η μονή και η διπλή γεννήτρια πιθανοτήτων, καθώς και η ροπογεννήτρια συνάρτηση αυτών.

2.2. Μονοδιάστατες Μεταβλητές Εμφυτεύσιμες σε Μαρκοβιανή Αλυσίδα

Οι Fu and Koutras (1994) εισήγαγαν και παρουσίασαν μια νέα μέθοδο που μπορεί να χρησιμοποιηθεί αποτελεσματικά, μεταξύ άλλων, και για τη μελέτη τυχαίων μεταβλητών που απαριθμούν σχηματισμούς καθορισμένου μήκους (για παράδειγμα ροών επιτυχιών μήκους k) σε ακολουθίες δοκιμών Bernoulli. Η μέθοδος αυτή αναφέρεται ως μέθοδος Μαρκοβιανής εμφύτευσης και συνίσταται στη μελέτη της τυχαίας μεταβλητής δια μέσω της εμφύτευσής της σε μια κατάλληλα ορισμένη Μαρκοβιανή αλυσίδα, εκμεταλλευόμενη την ακολουθιακή φύση των μοντέλων που μελετώνται.

Με τον επόμενο ορισμό οι Fu and Koutras (1994) εισήγαγαν την έννοια της μεταβλητής εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα.

Ορισμός 2.1: *Έστω ένας χώρος καταστάσεων* $\Omega = \{a_1, a_2, ...\}$. *Μια θετική ακέραια τυχαία μεταβλητή* X_n *με σύνολο τιμών* $\{0,1,2,....,\mathbf{l}_n\}$ ($n \in N$, $\mathbf{l}_n = \max\{x : P(X_n = x) > 0\}$) *θα λέγεται* ***εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα*** *αν*

(i) υπάρχει μια Μαρκοβιανή αλυσίδα διακριτού χρόνου $\{Y_t : t \geq 0\}$ *ορισμένη στο χώρο καταστάσεων* $\Omega = \{a_1, a_2, ...\}$

(ii) υπάρχει μια διαμέριση $\{C_x,\ x = 0,1,2,...\}$ *του* Ω *και*

(iii) για κάθε $x \in \{0,1,2,....,\mathbf{l}_n\}$ *ισχύει* $P(X_n = x) = P(Y_n \in C_x)$.

Με χρήση του θεωρήματος που δίνεται στη συνέχεια είναι δυνατός ο υπολογισμός της συνάρτησης πιθανότητας μιας τυχαίας μεταβλητής εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα.

Θεώρημα 2.1: *Αν η τυχαία μεταβλητή* X_n *είναι εμφυτεύσιμη στην Μαρκοβιανή αλυσίδα* $\{Y_t : t \geq 0\}$ *με χώρο καταστάσεων* $\Omega = \{a_1, a_2, ...\}$, *τότε*

$$P(X_n = x) = \boldsymbol{\pi}_0 \left(\prod_{t=1}^{n} \boldsymbol{\Lambda}_t \right) \sum_{r: a_r \in C_x} \mathbf{e}'_r$$

όπου $\boldsymbol{\pi}_0 = (P(Y_0 = a_1), P(Y_0 = a_2), ...)$ *είναι το διάνυσμα των αρχικών πιθανοτήτων της αλυσίδας,* $\boldsymbol{\Lambda}_t$ *ο πίνακας των πιθανοτήτων μετάβασης πρώτης τάξης της αλυσίδας και* $\mathbf{e}_r$ *το μοναδιαίο διάνυσμα γραμμή με όλες τις συνιστώσες ίσες με* 0 *εκτός της συνιστώσας* r *που είναι ίση με* 1.

Ένα σημαντικό πρόβλημα που προκύπτει κατά τη χρήση του Θεωρήματος 2.1. για τον αριθμητικό υπολογισμό της συνάρτησης πιθανότητας της τυχαίας μεταβλητή X_n είναι η μεγάλη διάσταση του πίνακα $\boldsymbol{\Lambda}_t$ στην περίπτωση που το μήκος της ακολουθίας που μελετάμε είναι μεγάλο.

Αντίστοιχα, το πλεονέκτημα της μεθόδου αυτής είναι ότι σε κάθε βήμα της αλυσίδας ξέρουμε τι ακριβώς έχει προηγηθεί και έχουμε όλες τις απαραίτητες πληροφορίες για να προχωρήσουμε στο επόμενο βήμα. Επιπλέον η μέθοδος μπορεί να εφαρμοσθεί με ελάχιστες τροποποιήσεις στους πίνακες $\boldsymbol{\Lambda}_t$ και στην περίπτωση που οι

δοκιμές είναι μη ισόνομες, Μαρκοβιανά εξαρτημένες ή με περισσότερα των δύο πιθανά αποτελέσματα.

Στο πόρισμα που ακολουθεί δίνονται τύποι για τις ροπές και την πιθανογεννήτρια συνάρτηση των εμφυτεύσιμων μεταβλητών.

Πόρισμα 2.1: *Αν η τυχαία μεταβλητή* X_n *είναι εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα* $\{Y_t : t \geq 0\}$ *με χώρο καταστάσεων* $\Omega = \{a_1, a_2, ...\}$, *τότε οι ροπές και η πιθανογεννήτρια συνάρτηση εκφράζονται, αντίστοιχα, ως*

$$E[X_n^{(i)}] = \boldsymbol{\pi}_0 \left(\prod_{t=1}^{n} \boldsymbol{\Lambda}_t \right) \mathbf{v}_i' \quad \textit{και}$$

$$j_x(z) = \sum_{x=0}^{\mathbf{l}_n} P[X_n = x] z^x = \boldsymbol{\pi}_0 \left(\prod_{t=1}^{n} \boldsymbol{\Lambda}_t \right) \mathbf{w}_z',$$

όπου $\mathbf{v}_i = \sum_{x=0}^{\mathbf{l}_n} x^i \left(\sum_{r:a_r \in C_x} \mathbf{e}_r' \right)$ *και* $\mathbf{w}_z = \sum_{x=0}^{\mathbf{l}_n} z^x \left(\sum_{r:a_r \in C_x} \mathbf{e}_r' \right)$.

Προκειμένου να γίνει πιο κατανοητή η έννοια της εμφύτευσης σε Μαρκοβιανή αλυσίδα όπως εισήχθη από τους Fu and Koutras (1996) δίνουμε το ακόλουθο παράδειγμα.

Παράδειγμα 2.1: Υποθέτουμε ότι $Z_1, Z_2, ..., Z_n$ είναι μια ακολουθία ανεξάρτητων δοκιμών Bernoulli με δυνατά αποτελέσματα, επιτυχία (1) και αποτυχία (0), με αντίστοιχες πιθανότητες p_t και $q_t = 1 - p_t$ για $t = 1,2,...,n$. Θα εφαρμόσουμε τη μέθοδο της Μαρκοβιανής εμφύτευσης προκειμένου να υπολογίσουμε την κατανομή της τυχαίας μεταβλητής $N_{n,k}$ η οποία δηλώνει τον αριθμό των μη επικαλυπτόμενων ροών μήκους k στις n δοκιμές.

Βήμα 1: Θεωρούμε το χώρο καταστάσεων

$$\Omega = \{(x,i) : x = 0,1,2,...,\mathbf{l}_n, i = 0,1,2,...,k-1\} \text{ με } \mathbf{l}_n = \left[\frac{n}{k} \right].$$

Βήμα 2: Θεωρούμε τη Μαρκοβιανή αλυσίδα $\{Y_t : t \geq 0\}$ ορισμένη στον χώρο Ω. Ορίζουμε $Y_t = (x,i)$ εάν στον χρόνο $t \geq 0$ έχουν εμφανισθεί x μη επικαλυπτόμενες

ροές μήκους k και επίσης i επιτυχίες μετά από την τελευταία εμφάνιση αποτυχίας ή από τη συμπλήρωση της τελευταίας ροής μήκους k.

Βήμα 3: Για $0 \le x \le \mathbf{l}_n$ ορίζουμε τα σύνολα καταστάσεων $C_x = \{(x,i) : i = 0,1,...,k-1\}$. Τα σύνολα C_x, $x = 0,1,2,...,\mathbf{l}_n$ αποτελούν μια διαμέριση του χώρου

$$\Omega = \{(x,i) : x = 0,1,2,...,\mathbf{l}_n, i = 0,1,2,...,k-1\}$$

αφού είναι φανερό ότι ισχύει $\Omega = \bigcup_{x=0}^{\mathbf{l}_n} C_x$.

Βήμα 4: Προφανώς για κάθε $x = \{0,1,2,....,\mathbf{l}_n\}$ ισχύει $P(N_{n,k} = x) = P(Y_n \in C_x)$.

Από τα βήματα που προηγήθηκαν προκύπτει εύκολα ότι ικανοποιείται ο Ορισμός 2.1 και συνεπώς η τυχαία μεταβλητή $N_{n,k}$ εμφυτεύεται σε Μαρκοβιανή αλυσίδα.

Το επόμενο βήμα της μεθόδου είναι να καθορίσουμε τις αρχικές πιθανότητες της αλυσίδας και τον πίνακα των πιθανοτήτων μετάβασης πρώτης τάξης της αλυσίδας $\mathbf{\Lambda}_t$.

Το διάνυσμα των αρχικών πιθανοτήτων είναι ίσο με $\boldsymbol{\pi}_0 = (1,0,0,...,0)$. Οι μη-μηδενικές πιθανότητες μετάβασης πρώτης τάξης, στοιχεία του $\mathbf{\Lambda}_t$, δίνονται από τις σχέσεις:

$$P(Y_t = (x,0) \mid Y_{t-1} = (x,i)) = q_t = 1 - p_t,\ x = 0,1,2,...,\mathbf{l}_n,\ i = 0,1,2,...,k-1$$

$$P(Y_t = (x,i+1) \mid Y_{t-1} = (x,i)) = p_t,\ x = 0,1,2,...,\mathbf{l}_n,\ i = 0,1,2,...,k-2$$

$$P(Y_t = (x+1,0) \mid Y_{t-1} = (x,i)) = p_t,\ x = 0,1,2,...,\mathbf{l}_n - 1,\ i = k-1.$$

Με χρήση των παραπάνω σχέσεων οι οποίες και δίνουν τα μη μηδενικά στοιχεία του $m \times m$ πίνακα ($m = k \times (\mathbf{l}_n + 1)$) των πιθανοτήτων μετάβασης πρώτης τάξης της αλυσίδας $\mathbf{\Lambda}_t$ και την παρατήρηση ότι το τελευταίο στοιχείο της κύριας διαγωνίου είναι ίσο με 1, μπορούμε να κατασκευάσουμε τον πίνακα $\mathbf{\Lambda}_t$.

Έτσι, για $k = 2$ και $n = 10$ έχουμε ότι ο πίνακας $\mathbf{\Lambda}_t$ έχει την ακόλουθη μορφή (όλα τα κενά στοιχεία του πίνακα είναι ίσα με το 0):

$$\mathbf{\Lambda}_t = \begin{array}{c} \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \\ (2,0) \\ (2,1) \\ (3,0) \\ (3,1) \\ (4,0) \\ (4,1) \\ (5,0) \\ (5,1) \end{array} \begin{array}{c} \begin{array}{cccccccccccc} (0,0) & (0,1) & (1,0) & (1,1) & (2,0) & (2,1) & (3,0) & (3,1) & (4,0) & (4,1) & (5,0) & (5,1) \end{array} \\ \begin{bmatrix} q_t & p_t & 0 & 0 & & & & & & & & \\ q_t & 0 & p_t & 0 & & & & & & & & \\ & & q_t & p_t & 0 & 0 & & & & & & \\ & & q_t & 0 & p_t & 0 & & & & & & \\ & & & & q_t & p_t & 0 & 0 & & & & \\ & & & & q_t & 0 & p_t & 0 & & & & \\ & & & & & & q_t & p_t & 0 & 0 & & \\ & & & & & & q_t & 0 & p_t & 0 & & \\ & & & & & & & & q_t & p_t & 0 & 0 \\ & & & & & & & & q_t & 0 & p_t & 0 \\ & & & & & & & & & & q_t & p_t \\ & & & & & & & & & & 0 & 1 \end{bmatrix}_{12\times 12} \end{array}$$

Παρατηρούμε εύκολα ότι ο πίνακας $\mathbf{\Lambda}_t$ έχει διδιαγώνια μορφή, δηλαδή στην κύρια διαγώνιο εμφανίζεται και επαναλαμβάνεται ένας πίνακας της μορφής $\mathbf{A}_t = \begin{bmatrix} q_t & p_t \\ q_t & 0 \end{bmatrix}$ ενώ άνω της κυρίας διαγωνίου, επαναλαμβάνεται ένας πίνακας της μορφής $\mathbf{B}_t = \begin{bmatrix} 0 & 0 \\ p_t & 0 \end{bmatrix}$. Οι δύο πίνακες έχουν διάσταση 2×2. Επίσης εμφανίζεται ο πίνακας $\mathbf{A}_t^* = \begin{bmatrix} q_t & p_t \\ 0 & 1 \end{bmatrix}$.

Είναι φανερό ότι στον πίνακα $\mathbf{\Lambda}_t$ προσθέτουμε κάποιες ψευδείς καταστάσεις (dummy states) – με την έννοια ότι τις καταστάσεις αυτές δεν τις φτάνει ποτέ η Μαρκοβιανή αλυσίδα – προκειμένου να διατηρούμε τις διαστάσεις των πινάκων σταθερές.

Η γενική μορφή των παραπάνω πινάκων για οποιεσδήποτε τιμές των k και n έχει ως εξής:

$$\mathbf{\Lambda}_t = \begin{bmatrix} \mathbf{A}_t & \mathbf{B}_t & \mathbf{0} & \mathbb{M} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_t & \mathbf{B}_t & \mathbb{M} & \mathbf{0} & \mathbf{0} \\ \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{O} & \mathbf{L} & \mathbf{L} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbb{M} & \mathbf{A}_t & \mathbf{B}_t \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbb{M} & \mathbf{0} & \mathbf{A}_t^* \end{bmatrix}$$

όπου

$$\mathbf{A}_t = \begin{bmatrix} & (x,0) & (x,1) & (x,2) & \mathbb{M} & (x,k-1) \\ (x,0) & q_t & p_t & 0 & \mathbb{M} & 0 \\ (x,1) & q_t & 0 & p_t & \mathbb{M} & 0 \\ \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{O} & \mathbf{L} \\ (x,k-2) & q_t & 0 & 0 & \mathbb{M} & p_t \\ (x,k-1) & q_t & 0 & 0 & \mathbb{M} & 0 \end{bmatrix}_{k\times k}$$

$$\mathbf{B}_t = \begin{bmatrix} & (x+1,0) & (x+1,1) & (x+1,2) & \mathbb{M} & (x+1,k-1) \\ (x,0) & 0 & 0 & 0 & \mathbb{M} & 0 \\ (x,1) & 0 & 0 & 0 & \mathbb{M} & 0 \\ \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{O} & \mathbf{L} \\ (x,k-2) & 0 & 0 & 0 & \mathbb{M} & 0 \\ (x,k-1) & p_t & 0 & 0 & \mathbb{M} & 0 \end{bmatrix}_{k\times k}$$

Τέλος, με χρήση του Θεωρήματος 2.1 είναι δυνατός ο υπολογισμός της συνάρτησης πιθανότητας της $N_{n,k}$.

Το πρόβλημα του αριθμητικού υπολογισμού της συνάρτησης πιθανότητας στην περίπτωση που ο πίνακας $\mathbf{\Lambda}_t$ έχει μεγάλη διάσταση λύθηκε με τη μελέτη και τον ορισμό της μεταβλητής διωνυμικού τύπου εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα (MVB) από τους Koutras and Alexandrou (1995).

Οι Koutras and Alexandrou (1995) εξήγαγαν μια σειρά αποτελεσμάτων για τη συνάρτηση πιθανότητας, τη γεννήτρια πιθανοτήτων και τη μέση τιμή της MVB. Με βάση τους τύπους που προέκυψαν, ο υπολογισμός αυτών των ποσοτήτων ανάγεται στον υπολογισμό κάποιων στοιχείων του αντίστροφου πίνακα ή του γινομένου πινάκων μικρής διάστασης (με λίγα μη μηδενικά στοιχεία). Παράλληλα, χωρίς να διαφοροποιούνται ουσιαστικά τα βήματα της τεχνικής αυτής, παρέχεται η δυνατότητα μελέτης μη ισόνομων ή Μαρκοβιανά εξαρτημένων δοκιμών.

Οι Koutras and Alexandrou (1995) όρισαν τη νέα τεχνική παρατηρώντας ότι, συνήθως, ο $\mathbf{\Lambda}_t$ μπορεί να εκφραστεί σε μια διδιαγώνια μορφή. Δηλαδή, οι μη μηδενικοί υποπίνακες $\mathbf{A}_t$ και $\mathbf{B}_t$ εμφανίζονται μόνο στην κύρια διαγώνιο και στη διαγώνιο που βρίσκεται ακριβώς επάνω από την κύρια. Έτσι, διαμόρφωσαν μια κατάλληλη τροποποίηση της τεχνικής της Μαρκοβιανής εμφύτευσης η οποία χρησιμοποιεί μόνο τους μη μηδενικούς υποπίνακες, ξεπερνώντας το πρόβλημα της

μεγάλης διάστασης του πίνακα $\mathbf{\Lambda}_t$. Η τροποποιημένη τεχνική στηρίζεται στη χρήση των πινάκων $\mathbf{A}_t(x)$ και $\mathbf{B}_t(x)$, της γενίκευσης των πινάκων $\mathbf{A}_t$ και $\mathbf{B}_t$ αντίστοιχα, καθώς και στα διανύσματα πιθανότητας $\mathbf{f}_t(x)$ τα οποία σε κάθε δοκιμή στο χρόνο t περιγράφουν τον σχηματισμό της ακολουθίας που έχει προηγηθεί.

Ορισμός 2.2: *Μια θετική ακέραια τυχαία μεταβλητή X_n θα λέγεται* ***μεταβλητή διωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα*** *(MVB) αν*

- *Η X_n εμφυτεύεται σε Μαρκοβιανή αλυσίδα $\{Y_t : t \geq 0\}$ σύμφωνα με τον Ορισμό 2.1, και επιπλέον $C_x = \{c_{x,0}, c_{x,1}, \mathbf{K}, c_{x,s-1}\}$, $x \geq 0$, όπου $s = |C_x|$, ο κοινός πληθάριθμος των C_x.*
- *Ισχύει ότι $P(Y_t = c_{y,j} \mid Y_{t-1} = c_{x,i}) = 0$ για κάθε $y \neq x, x+1$, $t = 1,\ldots,n$.*

Για κάθε MVB ορίζουμε τους πίνακες μετάβασης διαστάσεων $s \times s$,

$$\mathbf{A}_t(x) = (P(Y_t = c_{x,j} \mid Y_{t-1} = c_{x,i})) \ ,$$

$$\mathbf{B}_t(x) = (P(Y_t = c_{x+1,j} \mid Y_{t-1} = c_{x,i})),$$

και τα διανύσματα πιθανότητας διαστάσεων $1 \times s$

$$\mathbf{f}_t(x) = (P(Y_t = c_{x,0}), P(Y_t = c_{x,1}),\ldots, P(Y_t = c_{x,s-1})), \ 0 \leq t \leq n .$$

Τα στοιχεία του πίνακα $\mathbf{A}_t(x)$ δίνουν τις πιθανότητες μετάβασης ανάμεσα στις υποκαταστάσεις ($c_{x,i}$, $i = 0,1,2,\ldots,s-1$) ενώ τα στοιχεία του πίνακα $\mathbf{B}_t(x)$ δίνουν τις πιθανότητες μετάβασης ανάμεσα στις καταστάσεις (C_x, $x \geq 0$). Είναι φανερό ότι ισχύει $[\mathbf{A}_t(x) + \mathbf{B}_t(x)]\mathbf{1}' = \mathbf{1}'$.

Στο Θεώρημα 2.2 δίνονται αναδρομικές σχέσεις για τη συνάρτηση πιθανότητας που διευκολύνουν σημαντικά τον αριθμητικό υπολογισμό της κατανομής μιας τυχαίας μεταβλητής διωνυμικού τύπου εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα.

Θεώρημα 2.2: *Η διπλή ακολουθία των διανυσμάτων* $\mathbf{f}_t(x),\ 0 \le x \le \mathbf{l}_n,\ 1 \le t \le n$ *ικανοποιεί τις αναδρομικές σχέσεις*

$$\mathbf{f}_t(0) = \mathbf{f}_{t-1}(0)\mathbf{A}_t(0),\ \ t = 1,2,....,n$$

$$\mathbf{f}_t(x) = \mathbf{f}_{t-1}(x)\mathbf{A}_t(x) + \mathbf{f}_{t-1}(x-1)\mathbf{B}_t(x-1),\quad 0 \le x \le \mathbf{l}_n,\ \ t \ge 1$$

με αρχικές συνθήκες $\mathbf{f}_0(x) = \boldsymbol{\pi}_x,\ 0 \le x \le \mathbf{l}_n$. *Η συνάρτηση πιθανότητας της* X_n *δίνεται από τον τύπο*

$$P(X_n = x) = \mathbf{f}_n(x)\mathbf{1}',\quad 0 \le x \le \mathbf{l}_n$$

όπου $\mathbf{1}$ *συμβολίζει το διάνυσμα διάστασης* $1 \times s$ *με όλες τις συνιστώσες του ίσες με* 1.

Η χρήση του ονόματος μεταβλητή διωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα (MVB) οφείλεται στο γεγονός ότι ανάλογες αναδρομικές σχέσεις, με τις αναδρομικές σχέσεις του Θεωρήματος 2.2 ισχύουν και για την κλασική διωνυμική κατανομή.

Σημειώνουμε εδώ ότι οι αποδείξεις των Θεωρημάτων της παραγράφου αυτής, οι οποίες οφείλονται στους Koutras and Alexandrou (1995), παραλείπονται αφού μπορούν να προκύψουν και ως ειδικές περιπτώσεις των αποδείξεων των θεωρημάτων της επόμενης παραγράφου (2.3).

Παράδειγμα 2.2: Υποθέτοντας ότι $Z_1, Z_2, ..., Z_n$ είναι μια ακολουθία ανεξάρτητων και ισόνομων δοκιμών Bernoulli με δυνατά αποτελέσματα επιτυχία (1) και αποτυχία (0) και αντίστοιχες πιθανότητες p_t και $q_t = 1 - p_t$ για $t = 1,2,...,n$, μπορούμε να εφαρμόσουμε το Θεώρημα 2.2 για να υπολογίσουμε τη κατανομή της τυχαίας μεταβλητής $N_{n,k}$ η οποία καταγράφει τον αριθμό των μη επικαλυπτόμενων ροών μήκους k, χρησιμοποιώντας τους πίνακες μετάβασης διαστάσεων $s \times s$, $\mathbf{A}_t = (P(Y_t = c_{x,j} \mid Y_t = c_{x,i}))$, $\mathbf{B}_t = (P(Y_t = c_{x+1,j} \mid Y_t = c_{x,i}))$, και τα $1 \times s$ διανύσματα πιθανότητας $\mathbf{f}_t(x) = (P(Y_t = c_{x,0}), P(Y_t = c_{x,1}), ..., P(Y_t = c_{x,s-1}))$, $0 \le t \le n$ του Παραδείγματος 2.1.

Οι Koutras and Alexandrou (1995) έδωσαν εύχρηστους τύπους συναρτήσει των πινάκων $\mathbf{A}_t$ και $\mathbf{B}_t$ για τον υπολογισμό της μονής γεννήτριας, της διπλής γεννήτριας καθώς και της γεννήτριας των μέσων τιμών της τυχαίας μεταβλητής X_n. Συγκεκριμένα, στην περίπτωση μιας αλυσίδας, με πίνακες μετάβασης $\mathbf{A}_t(x), \mathbf{B}_t(x)$ ανεξάρτητους του x, έχουμε το ακόλουθο θεώρημα για τη διανυσματική γεννήτρια.

Θεώρημα 2.3: *Αν* $\mathbf{A}_t(x) = \mathbf{A}_t$ *και* $\mathbf{B}_t(x) = \mathbf{B}_t$ *για* $t \geq 1$ *και* $x \geq 0$ *τότε η διανυσματική γεννήτρια πιθανοτήτων*

$$\boldsymbol{\varphi}_t(z) = \sum_{x=0}^{\infty} \mathbf{f}_t(x) z^x$$

δίνεται από τον τύπο

$$\boldsymbol{\varphi}_t(z) = \boldsymbol{\pi}_0 \prod_{r=1}^{t} \left(\mathbf{A}_r + z\mathbf{B}_r\right).$$

Στην περίπτωση που η αλυσίδα είναι και ομογενής, οι Koutras and Alexandrou (1995) διατύπωσαν το ακόλουθο θεώρημα για τη διπλή διανυσματική γεννήτρια.

Θεώρημα 2.4: *Αν* $\mathbf{A}_t(x) = \mathbf{A}$ *και* $\mathbf{B}_t(x) = \mathbf{B}$ *για* $t \geq 1$ *και* $x \geq 0$ *τότε η διπλή διανυσματική γεννήτρια πιθανοτήτων* $\boldsymbol{\Phi}(z, w) = \sum_{t=0}^{\infty} \boldsymbol{\varphi}_t(z) w^t$ *δίνεται από τον τύπο*

$$\boldsymbol{\Phi}(z, w) = \boldsymbol{\pi}_0 \left[\mathbf{I} - w(\mathbf{A} + z\mathbf{B})\right]^{-1},$$

όπου $\mathbf{I}$ *είναι ο μοναδιαίος πίνακας διάστασης* $s \times s$.

Τέλος, για την γεννήτρια των μέσων τιμών της τυχαίας μεταβλητής X_n έχουμε το επόμενο αποτέλεσμα.

Θεώρημα 2.5: *Αν* $\mathbf{A}_t(x) = \mathbf{A}$ *και* $\mathbf{B}_t(x) = \mathbf{B}$ *για* $t \geq 1$ *και* $x \geq 0$ *τότε η γεννήτρια των μέσων τιμών δίνεται από τον τύπο*

$$M(w) = \sum_{n=1}^{\infty} E(X_n) w^n = \frac{w}{1-w} \boldsymbol{\pi}_0 \left[\mathbf{I} - w(\mathbf{A} + z\mathbf{B})\right]^{-1} \mathbf{B}\mathbf{1}'.$$

Με χρήση των μεταβλητών διωνυμικού τύπου εμφυτεύσιμων σε Μαρκοβιανή αλυσίδα (MVB) επιτυγχάνεται η μελέτη αρκετών τυχαίων μεταβλητών που σχετίζονται με τον αριθμό των εμφανίσεων ροών επιτυχιών καθώς και των περισσότερων τυχαίων μεταβλητών οι οποίες σχετίζονται με χρόνους αναμονής για την πρώτη ή την r-οστή εμφάνιση μιας ροής επιτυχιών. Δηλαδή, με κοινή μεθοδολογία είναι δυνατή η μελέτη, της $N_{n,k}$, απαριθμήτριας των μη επικαλυπτόμενων (non-overlapping) ροών επιτυχιών μήκους k, της $M_{n,k}$, απαριθμήτριας των επικαλυπτόμενων (overlapping) ροών επιτυχιών μήκους k, της $G_{n,k}$, τυχαίας μεταβλητής που απαριθμεί τις ροές επιτυχιών μήκους τουλάχιστον k.

Τέλος, με τη μέθοδο της Μαρκοβιανής εμφύτευσης είναι δυνατόν να επιτευχθεί και η μελέτη της τυχαίας μεταβλητής $E_{n,k}$ που αντιστοιχεί στις ροές επιτυχιών μήκους ακριβώς k, χρησιμοποιώντας κατάλληλη παραλλαγή της (Han and Aki (1999)).

2.3. Μεταβλητές Εμφυτεύσιμες σε Μαρκοβιανή Αλυσίδα Πολυωνυμικού Τύπου

Στην παράγραφο αυτή ορίζεται η έννοια της μεταβλητής πολυωνυμικού τύπου εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα (MVP). Η ανάγκη της εισαγωγής της έννοιας αυτής ανέκυψε προκειμένου να αντιμετωπισθεί μια ευρύτερη κλάση προβλημάτων σε σχέση με τις μεταβλητές MVB. Έτσι, έχουμε τον ακόλουθο ορισμό:

Ορισμός 2.3: *Μια θετική ακέραια τυχαία μεταβλητή* X_n *θα λέγεται* ***μεταβλητή πολυωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα*** (*MVP*) *αν*

- *Η* X_n *εμφυτεύεται σε Μαρκοβιανή αλυσίδα* $\{Y_t : t \geq 0\}$ *σύμφωνα με τον Ορισμό 2.1, και επιπλέον* $C_x = \{c_{x,0}, c_{x,1}, \mathbf{K}, c_{x,s-1}\}$, $x \geq 0$, *με* $s = |C_x|$ *τον κοινό πληθάριθμο των συνόλων* $C_x = \{c_{x,0}, c_{x,1}, ..., c_{x,s-1}\}$ *που αποτελούν μια διαμέριση του χώρου* Ω.
- *Ισχύει* $P(Y_t = c_{y,j} \mid Y_{t-1} = c_{x,i}) = 0$ *για κάθε* $y \neq x, x+1, ..., x+m$, $t \geq 1$, *όπου* m *ένας θετικός ακέραιος αριθμός.*

Η μελέτη μιας τυχαίας μεταβλητής τύπου MVP επιτυγχάνεται μέσω των πινάκων πιθανοτήτων μετάβασης διαστάσεων $s \times s$

$$\mathbf{A}_{t,i}(x) = (P(Y_t = c_{x+i,j'} \mid Y_{t-1} = c_{x,j})), \quad 0 \le i \le m, \ t \ge 1, \ x \ge 0$$

με ($j, j' \in \{0,1,...,s-1\}$) και των διανυσμάτων πιθανότητας διαστάσεων $1 \times s$

$$\mathbf{f}_t(x) = (P(Y_t = c_{x,0}), P(Y_t = c_{x,1}), ..., P(Y_t = c_{x,s-1})), \ 0 \le t \le n.$$

Από τον Ορισμό 2.3, είναι φανερό ότι ισχύει $\left(\sum_{i=0}^{m} \mathbf{A}_{t,i}(x)\right)\mathbf{1}' = \mathbf{1}'$, δηλαδή το άθροισμα των πινάκων $\mathbf{A}_{t,i}(x)$ για όλα τα i με $0 \le i \le m,$ είναι στοχαστικός πίνακας.

Στην συνέχεια, παρουσιάζεται ένα παράδειγμα προκειμένου να διευκρινιστούν οι διαφορές και οι ομοιότητες των τυχαίων μεταβλητών τύπου MVP και MVB.

Παράδειγμα 2.3: Έστω μια ακολουθία δίτιμων δοκιμών μήκους $n = 16$, και ας συμβολίσουμε με *F* την αποτυχία και *S* την επιτυχία. Έστω επίσης, μια τυχαία μεταβλητή $N_{n,k}$ η οποία μετρά τον αριθμό των μη επικαλυπτόμενων ροών μήκους $k = 2$ και μια άλλη μεταβλητή $S_{n,k}$ η οποία μετρά το άθροισμα των ροών μήκους τουλάχιστον $k = 2$. Είναι εύκολο να συμπεράνουμε ότι η μεταβλητή $N_{n,k}$ ανήκει στη κλάση των τυχαίων μεταβλητών διωνυμικού τύπου εμφυτεύσιμων σε Μαρκοβιανή αλυσίδα και για τη δεδομένη ακολουθία δίτιμων δοκιμών, του επόμενου πίνακα, τα ζεύγη καταστάσεων – υποκαταστάσεων είναι αυτά που περιγράφονται στον επόμενο πίνακα. Όμως συμπεραίνουμε ότι η μεταβλητή $S_{n,k}$ δεν ανήκει στην κλάση των τυχαίων μεταβλητών διωνυμικού τύπου εμφυτεύσιμων σε Μαρκοβιανή αλυσίδα. Ανήκει όμως στην ευρύτερη κλάση των MVP. Ας περιγράψουμε για μια δεδομένη ακολουθία δίτιμων δοκιμών το πως δουλεύει η μεθοδολογία που περιγράψαμε. Η περιγραφή θα στηριχθεί στην ακολουθία που δίνεται στον παρακάτω πίνακα. Στον ίδιο πίνακα δίνονται και τα ζεύγη καταστάσεων – υποκαταστάσεων της αλυσίδας που αντιστοιχεί στις δύο μεταβλητές.

Συνοπτικά, τα ζεύγη καταστάσεων – υποκαταστάσεων (*x*,*y*) για την $N_{n,k}$ αντιστοιχούν στην εμφάνιση *x* ροών επιτυχιών μήκους *k* και *y* συνεχόμενων επιτυχιών από την τελευταία αποτυχία. Για την $S_{n,k}$ τα ζεύγη καταστάσεων –

υποκαταστάσεων (x,y) αντιστοιχούν στην εμφάνιση x επιτυχιών σε ροές μήκους τουλάχιστον k και y συνεχόμενων επιτυχιών από την τελευταία αποτυχία (με την παρατήρηση ότι το σύμβολο * αντιστοιχεί στην περίπτωση που συμπληρώθηκε μια ροή μήκους τουλάχιστον k και ακολουθεί επιτυχία).

Βήμα	1	2	3	4	5	6	7	8
Δίτιμη Ακολουθία	*F*	*F*	*S*	*F*	*F*	*S*	*S*	*S*
$S_{n,k}$ Καταστάσεις	0,0	0,0	0,1	0,0	0,0	0,1	2,*	3,*
$N_{n,k}$ Καταστάσεις	0,0	0,0	0,1	0,0	0,0	0,1	1,0	1,1

Βήμα	9	10	11	12	13	14	15	16
Δίτιμη Ακολουθία	*S*	*F*	*S*	*S*	*F*	*S*	*S*	*S*
$S_{n,k}$ Καταστάσεις	4,*	4,0	4,1	6,*	6,0	6,1	8,*	9,*
$N_{n,k}$ Καταστάσεις	2,0	2,0	2,1	3,0	3,0	3,1	4,0	4,1

Έτσι, για την $N_{n,k}$ βλέπουμε ότι στο $3^{ο}$ βήμα με την εμφάνιση της $1^{ης}$ επιτυχίας αυξάνει ο δεύτερος δείκτης κατά μια μονάδα καταγράφοντας την απαρχή μιας πιθανής ροής επιτυχιών. Στο επόμενο βήμα όμως με την εμφάνιση αποτυχίας μηδενίζεται ξανά. Στην $6^{η}$ δοκιμή αυξάνεται ξανά κατά μια μονάδα, ενώ με την εμφάνιση της δεύτερης στη σειρά επιτυχίας μηδενίζεται ξανά αλλά αυτήν την φορά αυξάνει ο δείκτης καταστάσεων καταγράφοντας τον σχηματισμό της πρώτης ροής επιτυχιών μήκους 2. Την ίδια χρονική στιγμή ($7^{η}$) και με την συμπλήρωση της πρώτης ροής αυξάνει ο δείκτης καταστάσεων της $S_{n,k}$ κατά δύο μονάδες αλλά δεν μηδενίζεται ο δείκτης υποκαταστάσεων ο οποίος και λαμβάνει τη συμβολική τιμή (*) η οποία ερμηνεύεται ως μια κατάσταση αναμονής. Συγκεκριμένα, βλέπουμε ότι με την επόμενη επιτυχία, ο δείκτης των καταστάσεων αυξάνει κατά μια μονάδα ακόμα ενώ στο $10^{ο}$ βήμα και ενώ ο δείκτης καταστάσεων έχει ήδη την τιμή 4, ο δείκτης υποκαταστάσεων μηδενίζεται ξανά.

Στη συνέχεια, στην παράγραφο αυτή αναπτύσσουμε βήμα προς βήμα τους ορισμούς και τη μεθοδολογία που σχετίζεται με τις μεταβλητές τύπου MVP, βρίσκοντας αναδρομικές σχέσεις για τον υπολογισμό της συνάρτησης πιθανότητας, μελετώντας τη μονή και τη διπλή γεννήτρια πιθανοτήτων, καθώς και τη ροπογεννήτρια συνάρτηση. Εδώ θα πρέπει να σημειώσουμε ότι στις περισσότερες εφαρμογές οι πίνακες $\mathbf{A}_{t,i}(x)$ είναι ανεξάρτητοι του x και του t, δηλαδή $\mathbf{A}_{t,i}(x) = \mathbf{A}_i$.

Στο Θεώρημα 2.6 δίνουμε αναδρομικές σχέσεις για τον υπολογισμό της συνάρτησης πιθανότητας της X_n. Στα Θεωρήματα 2.7, 2.8, 2.9, χρησιμοποιώντας τεχνικές αντίστοιχες με αυτές των Koutras and Alexandrou (1995) αναπτύσσουμε τα κατάλληλα εργαλεία προκειμένου στα δύο επόμενα κεφάλαια να μελετήσουμε σειρά προβλημάτων τα οποία δεν είναι δυνατόν να μελετηθούν με την τεχνική της μεταβλητής διωνυμικού τύπου εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα (MVB).

Συνεχίζουμε δίνοντας το πρώτο αποτέλεσμα το οποίο αφορά την συνάρτηση πιθανότητας. Συγκεκριμένα, έχουμε το ακόλουθο θεώρημα για τα διανύσματα $\mathbf{f}_t(x)$.

Θεώρημα 2.6: *Η ακολουθία των διανυσμάτων* $\mathbf{f}_t(x)$ *ικανοποιεί τις σχέσεις*

$$\mathbf{f}_t(x) = \sum_{i=0}^{\min(x,m)} \mathbf{f}_{t-1}(x-i)\mathbf{A}_{t,i}(x-i), \quad x \geq 0,\ t \geq 1$$

και η συνάρτηση πιθανότητας X_n *δίνεται από τον τύπο*

$$P(X_n = x) = \mathbf{f}_n(x)\mathbf{1}'$$

όπου $\mathbf{1}$ *συμβολίζει το διάνυσμα διάστασης* $1 \times s$ *με όλες τις συνιστώσες του ίσες με* 1.

Απόδειξη: Εάν ισχύει $t \geq 1, x \geq 0, 0 \leq j \leq s-1$, τότε από το Θεώρημα της ολικής πιθανότητας έχουμε

$$P(Y_t = c_{x,j}) = \sum_{i=0}^{\min(x,m)} \sum_{r=0}^{s-1} P(Y_t = c_{x,j} \mid Y_{t-1} = c_{x-i,r}) P(Y_{t-1} = c_{x-i,r}),$$

το οποίο μπορεί να γραφτεί ισοδύναμα

$$P(Y_t = c_{x,j}) = \sum_{i=0}^{\min(x,m)} \sum_{r=0}^{s-1} \mathbf{e}_{r+1}\mathbf{A}_{t,i}(x-i)\mathbf{e}'_{j+1}\, P(Y_{t-1} = c_{x-i,r}).$$

Όμως,

$$\sum_{r=0}^{s-1} \mathbf{e}_{r+1} P(Y_{t-1} = c_{x-i,r}) = \mathbf{f}_{t-1}(x-i)$$

και η τελευταία ισότητα παίρνει την μορφή

$$P(Y_t = c_{x,j}) = \sum_{i=0}^{\min(x,m)} \mathbf{f}_{t-1}(x-i)\mathbf{A}_{t,i}(x-i)\mathbf{e}'_{j+1},$$

όπου $\mathbf{e}_j = \begin{pmatrix} 0 & 0 & \ldots & 1 & \ldots & 0 \end{pmatrix}_{s\times s}$. □

Η χρήση του ονόματος μεταβλητή πολυωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα (MVP) οφείλεται στο γεγονός ότι ανάλογες αναδρομικές σχέσεις, με αυτές του Θεωρήματος 2.6 ισχύουν και για τη συνήθη πολυωνυμική κατανομή.

Με χρήση του γεγονότος ότι στις περισσότερες εφαρμογές, η Μαρκοβιανή αλυσίδα που αντιστοιχεί στην MVP είναι ομογενής (οι πίνακες μετάβασης είναι ανεξάρτητοι από τον χρόνο και από την τιμή της τυχαίας μεταβλητής) στα Θεωρήματα 2.7 και 2.8 δίνουμε τη μονή και τη διπλή γεννήτρια πιθανοτήτων.

Πιο συγκεκριμένα εκφράζουμε τη μονή και τη διπλή γεννήτρια με χρήση απλών γινομένων και αθροισμάτων των πινάκων μετάβασης $\mathbf{A}_{t,i}(x),\ \ i = 0,1,2,..,m$. Πριν όμως προχωρήσουμε ας δούμε ένα παράδειγμα MVP.

Παράδειγμα 2.4: Αν η τυχαία μεταβλητή $S_{n,k}$, συμβολίζει / καταγράφει το άθροισμα των ροών μήκους τουλάχιστον $k = 2$ σε μία ακολουθία δοκιμών μήκους n, με πιθανότητα επιτυχίας p_t, τότε οι πίνακες $\mathbf{A}_{t,i}, i = 0,1,2$ του Ορισμού 2.3 είναι οι εξής:

$$\mathbf{A}_{t,0} = \begin{bmatrix} q_t & p_t & 0 \\ q_t & 0 & 0 \\ q_t & 0 & 0 \end{bmatrix},\ \mathbf{A}_{t,1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & p_t \end{bmatrix},\ \mathbf{A}_{t,2} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & p_t \\ 0 & 0 & 0 \end{bmatrix}$$

με καταστάσεις για τους $\mathbf{A}_{t,0}, \mathbf{A}_{t,1}, \mathbf{A}_{t,2}$ τα ζεύγη $(x,0)$, $(x,1)$, $(x,2)$. Τέλος, μπορούμε να εφαρμόσουμε το Θεώρημα 2.6 για να υπολογίσουμε την κατανομή της τυχαίας μεταβλητής $S_{n,k}$, χρησιμοποιώντας τους πίνακες μετάβασης διαστάσεων $s\times s$, $\mathbf{A}_{t,i}$, και τα διανύσματα πιθανότητας διαστάσεων $1\times s$, $\mathbf{f}_t(x)$.

Θεωρώντας ότι η Μαρκοβιανή αλυσίδα που αντιστοιχεί στην MVP είναι ανεξάρτητη του *x* έχουμε το ακόλουθο θεώρημα.

Θεώρημα 2.7: *Αν* $\mathbf{A}_{t,i}(x) = \mathbf{A}_{t,i}$ *για* $t \geq 1$ *και* $x \geq 0$ *τότε η διανυσματική γεννήτρια πιθανοτήτων*

$$\boldsymbol{\varphi}_t(z) = \sum_{x=0}^{\infty} \mathbf{f}_t(x) z^x$$

δίνεται από τον τύπο

$$\boldsymbol{\varphi}_t(z) = \boldsymbol{\pi}_0 \prod_{r=1}^{t} \left(\sum_{i=0}^{m} z^i \mathbf{A}_{r,i} \right), \quad t \geq 1.$$

Απόδειξη: Με $t \geq 1$ και με εφαρμογή του Θεωρήματος 2.6 προκύπτει ότι

$$\begin{aligned}
\boldsymbol{\varphi}_t(z) &= \sum_{x=0}^{\infty} \mathbf{f}_t(x) z^x = \sum_{x=0}^{m} \sum_{i=0}^{x} \mathbf{f}_{t-1}(x-i) \mathbf{A}_{t,i} z^x + \sum_{x=m+1}^{\infty} \sum_{i=0}^{m} \mathbf{f}_{t-1}(x-i) \mathbf{A}_{t,i} z^x \\
&= \sum_{i=0}^{m} z^i \left(\sum_{x=i}^{m} \mathbf{f}_{t-1}(x-i) z^{x-i} \right) \mathbf{A}_{t,i} + \sum_{i=0}^{m} z^i \left(\sum_{x=m+1}^{\infty} \mathbf{f}_{t-1}(x-i) z^{x-i} \right) \mathbf{A}_{t,i} \\
&= \sum_{i=0}^{m} z^i \left(\sum_{y=0}^{\infty} \mathbf{f}_{t-1}(y) z^y \right) \mathbf{A}_{t,i} = \varphi_{t-1}(z) \left(\sum_{i=0}^{m} z^i \mathbf{A}_{t,i} \right),
\end{aligned}$$

και επαναλαμβάνοντας τη διαδικασία μέχρι να φτάσουμε στην

$$\boldsymbol{\varphi}_0(t) = \boldsymbol{\pi}_0$$

ολοκληρώνεται η απόδειξη. □

Η διπλή γεννήτρια αποτελεί πολύ χρήσιμο εργαλείο ενώ η υπόθεση της ομογένειας της αλυσίδας καθώς και οι υποθέσεις της ισονομίας και της ανεξαρτησίας των δοκιμών πληρούται στο μεγαλύτερο μέρος των εφαρμογών.

Έτσι, στην περίπτωση που η αλυσίδα είναι ομογενής μπορούμε να διατυπώσουμε το ακόλουθο θεώρημα για τη διπλή διανυσματική γεννήτρια.

Θεώρημα 2.8: *Αν* $\mathbf{A}_{t,i}(x) = \mathbf{A}_i$ *για* $t \geq 1$ *και* $x \geq 0$ *τότε η διπλή διανυσματική γεννήτρια πιθανοτήτων*

$$\mathbf{\Phi}(z,w) = \sum_{t=0}^{\infty} \boldsymbol{\varphi}_t(z) w^t$$

δίνεται από τον τύπο

$$\mathbf{\Phi}(z,w) = \boldsymbol{\pi}_0 \left[\mathbf{I} - w \left(\sum_{i=0}^{m} z^i \mathbf{A}_i \right) \right]^{-1}$$

όπου $\mathbf{I}$ *είναι ο μοναδιαίος πίνακας διάστασης* $s \times s$.

Απόδειξη: Με χρήση του Θεωρήματος 2.7 έχουμε

$$\mathbf{\Phi}(z,w) = \sum_{t=0}^{\infty} \boldsymbol{\varphi}_t(z) w^t = \boldsymbol{\pi}_0 \sum_{t=0}^{\infty} \left(\left(\sum_{i=0}^{m} z^i \mathbf{A}_i \right) w \right)^t = \boldsymbol{\pi}_0 \left[\mathbf{I} - w \left(\sum_{i=0}^{m} z^i \mathbf{A}_i \right) \right]^{-1},$$

περιορίζοντας το w σε μια κατάλληλη περιοχή του μηδενός, για να συγκλίνει η σειρά

$$\sum_{t=0}^{\infty} \left(\left(\sum_{i=0}^{m} z^i \mathbf{A}_i \right) w \right)^t,$$

η απόδειξη ολοκληρώνεται. □

Τέλος, στο επόμενο θεώρημα δίνεται μια έκφραση για τη γεννήτρια μέσων τιμών καθώς και για τη μέση τιμή μιας MVP.

Θεώρημα 2.9: *Αν* $\mathbf{A}_{t,i}(x) = \mathbf{A}_i$ *για* $t \geq 1$ *και* $x \geq 0$ *τότε,*

α) η μέση τιμή της X_n *μπορεί να εκφραστεί ως*

$$E(X_n) = \boldsymbol{\pi}_0 \left\{ \sum_{r=1}^{n} \left(\sum_{i=0}^{m} \mathbf{A}_i \right)^{r-1} \right\} \left(\sum_{i=1}^{m} i \mathbf{A}_i \right) \mathbf{1}',$$

β) η γεννήτρια των μέσων τιμών δίνεται από τον τύπο

$$M(w) = \sum_{n=1}^{\infty} E(X_n) w^n = \frac{w}{1-w} \boldsymbol{\pi}_0 \left[\mathbf{I} - w \sum_{i=0}^{m} \mathbf{A}_i \right]^{-1} \left(\sum_{i=1}^{m} i \mathbf{A}_i \right) \mathbf{1}'.$$

Απόδειξη:

α) Από τη γνωστή σχέση

$$E(X_n) = \left[\frac{d}{dz} \boldsymbol{\varphi}_t(z) \right] \Big|_{z=1}$$

έχουμε

$$E(X_n)=\frac{d}{dz}\left(\boldsymbol{\pi}_0\prod_{r=1}^{t}\left(\sum_{i=0}^{m}z^i\mathbf{A}_{r,i}\right)\right)\Bigg|_{z=1}=\boldsymbol{\pi}_0\sum_{r=1}^{t}\left[\left(\sum_{i=0}^{m}\mathbf{A}_i z^i\right)^{r-1}\left(\sum_{i=1}^{m}i\mathbf{A}_i\right)\left(\sum_{i=0}^{m}\mathbf{A}_i z^i\right)^{t-r}\right]\mathbf{1}'.$$

β) Η σχέση

$$M(w)=\sum_{n=1}^{\infty}E(X_n)w^n$$

μπορεί να γραφεί ως

$$M(w)=\boldsymbol{\pi}_0\sum_{t=1}^{\infty}\sum_{r=1}^{t}\left[\left(\sum_{i=1}^{m}\mathbf{A}_i\right)^{r-1}\left(\sum_{i=1}^{m}i\mathbf{A}_i\right)\right]w^t\mathbf{1}'=\boldsymbol{\pi}_0 w\sum_{r=1}^{\infty}\left(\sum_{i=1}^{m}\mathbf{A}_i\right)^{r-1}w^{r-1}\sum_{t=r}^{\infty}w^{t-r}\left(\sum_{i=1}^{m}i\mathbf{A}_i\right)\mathbf{1}'$$

και τέλος με χρήση της σχέσης

$$\sum_{r=1}^{\infty}\left(\sum_{i=1}^{m}\mathbf{A}_i\right)^{r-1}w^{r-1}=\left(I-w\sum_{i=1}^{m}\mathbf{A}_i\right)^{-1}$$

καταλήγουμε στην τελική μορφή. □

Αξίζει να σημειωθεί ότι, μελετώντας προσεκτικά τους Ορισμούς 2.2 και 2.3, προκύπτει πολύ εύκολα το συμπέρασμα ότι οι τυχαίες μεταβλητές διωνυμικού τύπου εμφυτεύσιμες σε Μαρκοβιανή αλυσίδα είναι η ειδική περίπτωση $m=1$ των τυχαίων μεταβλητών πολυωνυμικού τύπου εμφυτεύσιμες σε Μαρκοβιανή αλυσίδα. Συνέπεια του γεγονότος αυτού είναι ότι οι αποδείξεις των Θεωρημάτων 2.2 έως 2.5 μπορούν να προκύψουν ως πορίσματα των Θεωρημάτων 2.6 έως 2.9 που ήδη αναφέραμε.

2.4. Διδιάστατες και Πολυδιάστατες Μεταβλητές Εμφυτεύσιμες σε Μαρκοβιανή Αλυσίδα

Εργαζόμενοι ανεξάρτητα οι Alexandrou (1997) και οι Han and Aki (1999) γενίκευσαν την έννοια της τυχαίας μεταβλητής διωνυμικού τύπου εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα στην περίπτωση διδιάστατων και πολυδιάστατων τυχαίων μεταβλητών, δίνοντας και πάλι εύχρηστους τύπους συναρτήσει των πινάκων $\mathbf{A}_t$ και $\mathbf{B}_t$ για τον υπολογισμό της συνάρτησης πιθανότητας, της γεννήτριας μέσων τιμών, της μονής και της διπλής γεννήτριας μιας πολυδιάστατης τυχαίας μεταβλητής $\mathbf{X}_n$, σχετικής

με ροές επιτυχιών. Επιπλέον, επειδή σε αυτές τις περιπτώσεις έχει νόημα και η συνδιακύμανση (μεταξύ των τυχαίων μεταβλητών που αποτελούν την πολυδιάστατη μεταβλητή), δόθηκε και μια έκφραση για τη γεννήτρια της μέσης τιμής του γινομένου των επιμέρους μεταβλητών.

Στην παράγραφο αυτή αρχικά παραθέτουμε την διδιάστατη τυχαία μεταβλητή διωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα όπως δόθηκε από την Alexandrou (1997) και στην συνέχεια εισάγουμε και μελετάμε τις διδιάστατες και πολυδιάστατες τυχαίες μεταβλητές πολυωνυμικού τύπου εμφυτεύσιμων σε Μαρκοβιανή αλυσίδα.

Ορισμός 2.4: *Μια θετική ακέραια διδιάστατη τυχαία μεταβλητή* $(X_n^{(1)}, X_n^{(2)})$ *ορισμένη στο* $\{0,1,2,...,\mathbf{l}_n^1\}\times\{0,1,2,...,\mathbf{l}_n^2\}$ *με* $n\in N$ *θα λέγεται «**διδιάστατη μεταβλητή διωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα» (BMVB)** *αν*

- *Υπάρχει μια Μαρκοβιανή αλυσίδα διακριτού χρόνου* $\{Y_t : t\geq 0\}$ *ορισμένη στο χώρο καταστάσεων* $\Omega = \bigcup_{x_1,x_2} C_{x_1,x_2}$, $C_{x_1,x_2} = \{c_{x_1,x_2,0}, c_{x_1,x_2,1},.., c_{x_1,x_2,s-1}\}$.
- $P(Y_t \in C_{x_1,x_2} \mid Y_{t-1} \in C_{y_1,y_2}) = 0$ *για κάθε* $(x_1,x_2) \notin \{(y_1,y_2),(y_1+1,y_2),(y_1,y_2+1)\}$, $t\geq 1$.
- *Ισχύει ότι* $P(X_n^{(1)} = x_1, X_n^{(2)} = x_2) = P(Y_n \in C_{x_1,x_2})$, $n\geq 0$, $x_1, x_2 \geq 0$.

Θα ορίσουμε τώρα την έννοια της διδιάστατης μεταβλητής πολυωνυμικού τύπου εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα (BMVP). Η τυχαία μεταβλητή τύπου BMVP αποτελεί τη διδιάστατη γενίκευση της τυχαίας μεταβλητής τύπου MVP που ορίσαμε στην παράγραφο 2.3.

Είναι ιδιαίτερα σημαντικό το γεγονός, ότι, η γενίκευση από την μονοδιάστατη στην διδιάστατη περίπτωση και η μελέτη της διδιάστατης περίπτωσης δεν απαιτεί παρά ελάχιστες τροποποιήσεις στον ορισμό.

Ορισμός 2.5: *Μια θετική ακέραια διδιάστατη τυχαία μεταβλητή* $(X_n^{(1)}, X_n^{(2)})$ *ορισμένη στο* $\{0,1,2,...,\mathbf{l}_n^1\}\times\{0,1,2,...,\mathbf{l}_n^2\}$ *με* $n\in N$ *θα λέγεται* ***διδιάστατη μεταβλητή πολυωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα*** *(BMVP) αν*

- *Υπάρχει μια Μαρκοβιανή αλυσίδα διακριτού χρόνου* $\{Y_t : t\ge 0\}$ *ορισμένη στο χώρο καταστάσεων* $\Omega = \bigcup_{x_1,x_2} C_{x_1,x_2}$, $C_{x_1,x_2} = \{c_{x_1,x_2,0}, c_{x_1,x_2,1}, .., c_{x_1,x_2,s-1}\}$
- *Υπάρχουν δύο θετικοί αριθμοί* m_1, m_2 *τέτοιοι ώστε για* $t\ge 1$, $P(Y_t \in C_{x_1,x_2} \mid Y_{t-1}\in C_{y_1,y_2}) = 0$ *για* $(x_1,x_2)\notin\{(y_1+u, y_2), (y_1, y_2+v)\}$, $0\le u\le m_1$, $0\le v\le m_2$.
- *Ισχύει ότι* $P(X_n^{(1)} = x_1, X_n^{(2)} = x_2) = P(Y_n \in C_{x_1,x_2})$, $n\ge 0$, $x_1, x_2 \ge 0$.

Για την μελέτη μιας BMVP χρειαζόμαστε το διάνυσμα αρχικών πιθανοτήτων

$$\boldsymbol{\pi}_{x_1,x_2} = (P(Y_0 = c_{x_1,x_2,0}, P(Y_0 = c_{x_1,x_2,1}),...,P(Y_0 = c_{x_1,x_2,s-1})),\ x_1, x_2 \ge 0$$

τους «εσωτερικούς» πίνακες μετάβασης

$$\mathbf{A}_{t,0}(x_1,x_2) = (P(Y_t = c_{x_1,x_2,j'} \mid Y_{t-1} = c_{x_1,x_2,j}))_{s\times s},$$

και τους «εξωτερικούς» πίνακες μετάβασης

$$\mathbf{A}_{t,i}^{(1)}(x_1,x_2) = (P(Y_t = c_{x_1+i,x_2,j'} \mid Y_{t-1} = c_{x_1,x_2,j}),\quad 1\le i\le m_1,$$

και

$$\mathbf{A}_{t,i}^{(2)}(x_1,x_2) = (P(Y_t = c_{x_1,x_2+i,j'} \mid Y_{t-1} = c_{x_1,x_2,j}),\quad 1\le i\le m_2.$$

Εισάγοντας τα διανύσματα πιθανότητας

$$\mathbf{f}_t(x_1,x_2) = (P(Y_t = c_{x_1,x_2,0}), P(Y_t = c_{x_1,x_2,1}),...,P(Y_t = c_{x_1,x_2,s-1})),\ x_1,x_2\ge 0,\ t\ge 0$$

προκύπτει ότι

$$P(X_n^{(1)} = x_1, X_n^{(2)} = x_2) = P(Y_n \in C_{x_1,x_2}) = \mathbf{f}_n(x_1,x_2)\mathbf{1}'.$$

Στο Θεώρημα 2.10 δίνεται μια αναδρομική σχέση για τον υπολογισμό των διανυσμάτων $\mathbf{f}_t(x_1,x_2)$.

Είναι εμφανής η ομοιότητα ή ακόμα καλύτερα η αναλογία των αποτελεσμάτων με την μονοδιάστατη περίπτωση.

Θεώρημα 2.10: *Η ακολουθία των διανυσμάτων* $\mathbf{f}_t(x_1, x_2)$ *ικανοποιεί το αναδρομικό σχήμα*

$$\mathbf{f}_t(x_1,x_2) = \mathbf{f}_{t-1}(x_1,x_2)\mathbf{A}_{t,0}(x_1,x_2) + \sum_{i=1}^{m_1} \mathbf{f}_{t-1}(x_1-i,x_2)\mathbf{A}_{t,i}^{(1)}(x_1-i,x_2) + \sum_{i=1}^{m_2} \mathbf{f}_{t-1}(x_1,x_2-i)\mathbf{A}_{t,i}^{(2)}(x_1,x_2-i)$$

με $x_1, x_2 \geq 0,\ t \geq 1.$

Απόδειξη: Για $t \geq 1, x \geq 0, 0 \leq j \leq s-1,$ από το θεώρημα της ολικής πιθανότητας έχουμε

$$\begin{aligned} P(Y_t = c_{x_1,x_2,j}) &= \sum_{r=0}^{s-1} P(Y_t = c_{x_1,x_2,j} \mid Y_{t-1} = c_{x_1,x_2,r}) P(Y_{t-1} = c_{x_1,x_2,r}) \\ &+ \sum_{i=0}^{m_1}\sum_{r=0}^{s-1} P(Y_t = c_{x_1,x_2,j} \mid Y_{t-1} = c_{x_1-i,x_2,r}) P(Y_{t-1} = c_{x_1-i,x_2,r}) \\ &+ \sum_{i=0}^{m_2}\sum_{r=0}^{s-1} P(Y_t = c_{x_1,x_2,j} \mid Y_{t-1} = c_{x_1,x_2-i,r}) P(Y_{t-1} = c_{x_1,x_2-i,r}). \end{aligned}$$

Η παραπάνω σχέση μπορεί να γραφεί ως

$$\begin{aligned} P(Y_t = c_{x_1,x_2,j}) &= \sum_{r=0}^{s-1} \mathbf{e}_{r+1}\mathbf{A}_{t,0}(x_1,x_2)\mathbf{e}'_{j+1}\, P(Y_{t-1} = c_{x_1,x_2,r}) \\ &+ \sum_{i=0}^{m_1}\sum_{r=0}^{s-1} \mathbf{e}_{r+1}\mathbf{A}_{t,i}^{(1)}(x_1-i,x_2)\mathbf{e}'_{j+1}\, P(Y_{t-1} = c_{x_1-i,x_2,r}) \\ &+ \sum_{i=0}^{m_2}\sum_{r=0}^{s-1} \mathbf{e}_{r+1}\mathbf{A}_{t,i}^{(2)}(x_1,x_2-i)\mathbf{e}'_{j+1}\, P(Y_{t-1} = c_{x_1,x_2-i,r}) = \\ &= \mathbf{f}_{t-1}(x_1,x_2)\mathbf{A}_{t,0}(x_1,x_2)\mathbf{e}'_{j+1} \\ &+ \sum_{i=1}^{m_1} \mathbf{f}_{t-1}(x_1-i,x_2)\mathbf{A}_{t,i}^{(1)}(x_1-i,x_2)\mathbf{e}'_{j+1} + \sum_{i=1}^{m_2} \mathbf{f}_{t-1}(x_1,x_2-i)\mathbf{A}_{t,i}^{(2)}(x_1,x_2-i)\mathbf{e}'_{j+1} \end{aligned}$$

και η απόδειξη ολοκληρώνεται. □

Η μονή και η διπλή γεννήτρια συνάρτηση μιας BMVP δίνεται στα Θεωρήματα 2.11 και 2.12 αντίστοιχα.

Θεώρημα 2.11: *Αν* $\mathbf{A}_{t,0}(x_1,x_2)=\mathbf{A}_{t,0}$, $\mathbf{A}^{(1)}_{t,i}(x_1,x_2)=\mathbf{A}^{(1)}_{t,i}$, $i=1,2,...,m_1$, *και* $\mathbf{A}^{(2)}_{t,i}(x_1,x_2)=\mathbf{A}^{(2)}_{t,i}$, $i=1,2,...,m_2$, *για κάθε* (x_1,x_2), *τότε η (διανυσματική) γεννήτρια συνάρτηση των διανυσμάτων* $\mathbf{f}_t(x_1,x_2)$ *δίνεται από τον τύπο*

$$\boldsymbol{\varphi}_t(z_1,z_2)=\sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty}\mathbf{f}_t(x_1,x_2)z_1^{x_1}z_2^{x_2}=\boldsymbol{\varphi}_0(z_1,z_2)\prod_{r=1}^{t}\left(\mathbf{A}_{r,0}+\sum_{i=1}^{m_1}\mathbf{A}^{(1)}_{r,i}z_1^i+\sum_{i=1}^{m_2}\mathbf{A}^{(2)}_{r,i}z_2^i\right)$$

όπου

$$\boldsymbol{\varphi}_0(z_1,z_2)=\sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty}\mathbf{f}_0(x_1,x_2)z_1^{x_1}z_2^{x_2}\ .$$

Απόδειξη: Η $\boldsymbol{\varphi}_0(z_1,z_2)=\sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty}\mathbf{f}_0(x_1,x_2)z_1^{x_1}z_2^{x_2}$ μπορεί να γραφτεί ως

$$\boldsymbol{\varphi}_t(z_1,z_2)=\mathbf{f}_t(0,0)+\sum_{i=1}^{6}\sum_{S_i}\mathbf{f}_t(x_1,x_2)z_1^{x_1}z_2^{x_2}$$

όπου $S_1=\{(x_1,x_2):x_1=0,x_2\ge 1\}$,

$S_2=\{(x_1,x_2):x_1\ge 1,x_2=0\}$,

$S_3=\{(x_1,x_2):1\le x_1\le m_1,1\le x_2\le m_2\}$,

$S_4=\{(x_1,x_2):1\le x_1\le m_1,x_2\ge m_2+1\}$,

$S_5=\{(x_1,x_2):x_1\ge m_1+1,1\le x_2\le m_2\}$,

$S_6=\{(x_1,x_2):x_1\ge m_1+1,x_2\ge m_2+1\}$.

Εάν ισχύει $t\ge 1$, τότε με χρήση του Θεωρήματος 2.10 μπορούμε να εκφράσουμε τα έξι αθροίσματα σε μια μορφή όπως αυτή που ακολουθεί (για το τέταρτο άθροισμα):

$$\begin{aligned}\sum_{x_1=1}^{m_1}\sum_{x_2=m_2+1}^{\infty}\mathbf{f}_t(x_1,x_2)z_1^{x_1}z_2^{x_2}&=\\ &=\sum_{x_1=1}^{m_1}\sum_{x_2=m_2+1}^{\infty}\mathbf{f}_{t-1}(x_1,x_2)\mathbf{A}_{t,0}z_1^{x_1}z_2^{x_2}+\sum_{x_1=1}^{m_1}\sum_{x_2=m_2+1}^{\infty}\sum_{u=1}^{x_1}\mathbf{f}_{t-1}(x_1-u,x_2)\mathbf{A}^{(1)}_{t,u}z_1^{x_1}z_2^{x_2}+\\ &+\sum_{x_1=1}^{m_1}\sum_{x_2=m_2+1}^{\infty}\sum_{u=1}^{x_2}\mathbf{f}_{t-1}(x_1,x_2-u)\mathbf{A}^{(2)}_{t,u}z_1^{x_1}z_2^{x_2}\\ &=\sum_{x_1=1}^{m_1}\sum_{x_2=m_2+1}^{\infty}\mathbf{f}_{t-1}(x_1,x_2)\mathbf{A}_{t,0}z_1^{x_1}z_2^{x_2}+\sum_{u=1}^{m_1}z_1^u\left(\sum_{x_1=0}^{m_1-u}\sum_{x_2=m_2+1}^{\infty}\mathbf{f}_{t-1}(x_1,x_2)\mathbf{A}^{(1)}_{t,u}z_1^{x_1}z_2^{x_2}\right)+\\ &+\sum_{u=1}^{m_2}z_2^u\left(\sum_{x_1=1}^{m_1}\sum_{x_2=m_2-u+1}^{\infty}\mathbf{f}_{t-1}(x_1,x_2)\mathbf{A}^{(2)}_{t,u}z_1^{x_1}z_2^{x_2}\right)\end{aligned}$$

Δουλεύοντας με τον ίδιο τρόπο μπορούμε να εκφράσουμε με την ίδια μορφή και τα υπόλοιπα αθροίσματα και αντικαθιστώντας το $\mathbf{f}_t(0,0)$ με $\mathbf{f}_{t-1}(0,0)\mathbf{A}_{t,0}$ έχουμε

$$\sum_{x_1=1}^{m_1}\sum_{x_2=m_2+1}^{\infty}\mathbf{f}_{t-1}(x_1,x_2)\mathbf{A}_{t,0}z_1^{x_1}z_2^{x_2}+\sum_{u=1}^{m_1}z_1^u\left(\sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty}\mathbf{f}_{t-1}(x_1,x_2)\mathbf{A}_{t,u}^{(1)}z_1^{x_1}z_2^{x_2}\right)+$$

$$+\sum_{u=1}^{m_2}z_2^u\left(\sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty}\mathbf{f}_{t-1}(x_1,x_2)\mathbf{A}_{t,u}^{(2)}z_1^{x_1}z_2^{x_2}\right)$$

$$=\boldsymbol{\varphi}_{t-1}(z_1,z_2)\left(\mathbf{A}_{t,0}+\sum_{i=1}^{m_1}\mathbf{A}_{t,i}^{(1)}z_1^u+\sum_{i=1}^{m_2}\mathbf{A}_{t,i}^{(2)}z_2^u\right).$$

Επαναλαμβάνοντας την διαδικασία για κάθε σύνολο S_i, $i=1,2,3,5,6$ ολοκληρώνεται η απόδειξη □

Θεώρημα 2.12: *Αν* $\mathbf{A}_{t,0}(x_1,x_2)=\mathbf{A}_0$, $\mathbf{A}_{t,i}^{(1)}(x_1,x_2)=\mathbf{A}_i^{(1)}$, $i=1,2,...,m_1$, *και* $\mathbf{A}_{t,i}^{(2)}(x_1,x_2)=\mathbf{A}_i^{(2)}$, $i=1,2,...,m_2$, *για κάθε* (x_1,x_2) *και* t, *τότε η διπλή (διανυσματική) γεννήτρια συνάρτηση των διανυσμάτων* $\mathbf{f}_t(x_1,x_2)$ *δίνεται από τον τύπο*

$$\boldsymbol{\Phi}(z_1,z_2;w)=\sum_{t=0}^{\infty}\boldsymbol{\varphi}_t(z_1,z_2)w^t=\boldsymbol{\pi}_{0,0}\left(\mathbf{I}-w\left(\mathbf{A}_0+\sum_{i=1}^{m_1}\mathbf{A}_i^{(1)}z_1^i+\sum_{i=1}^{m_2}\mathbf{A}_i^{(2)}z_2^i\right)\right)^{-1}.$$

Απόδειξη: Από το Θεώρημα 2.11 έχουμε

$$\boldsymbol{\Phi}(z_1,z_2;w)=\sum_{t=0}^{\infty}\boldsymbol{\varphi}_t(z_1,z_2)w^t=$$

$$=\boldsymbol{\pi}_{0,0}\sum_{t=0}^{\infty}\left(\left(\mathbf{A}_0+\sum_{i=1}^{m_1}\mathbf{A}_i^{(1)}z_1^i+\sum_{i=1}^{m_2}\mathbf{A}_i^{(2)}z_2^i\right)w\right)^t$$

και περιορίζοντας το w σε μια κατάλληλη περιοχή του μηδενός, για να συγκλίνει η σειρά, έχουμε ότι

$$\boldsymbol{\Phi}(z_1,z_2;w)=\boldsymbol{\pi}_{0,0}\left(\mathbf{I}-w\left(\mathbf{A}_0+\sum_{i=1}^{m_1}\mathbf{A}_i^{(1)}z_1^i+\sum_{i=1}^{m_2}\mathbf{A}_i^{(2)}z_2^i\right)\right)^{-1}.$$

□

Στο Θεώρημα 2.13 δίνουμε σχέσεις για τον υπολογισμό των ποσοτήτων $E(X_t^{(j)})$ και $E(X_t^{(1)}X_t^{(2)})$. Η ποσότητα $E(X_t^{(1)}X_t^{(2)})$ είναι απαραίτητη για τον υπολογισμό της

συνδιακύμανσης και του συντελεστή συσχέτισης των δύο τυχαίων μεταβλητών $X_t^{(1)}$ και $X_t^{(2)}$.

Θεώρημα 2.13: *Αν* $\mathbf{A}_{t,0}(x_1,x_2)=\mathbf{A}_0$, $\mathbf{A}_{t,i}^{(1)}(x_1,x_2)=\mathbf{A}_i^{(1)}$, $i=1,2,...,m_1$, *και* $\mathbf{A}_{t,i}^{(2)}(x_1,x_2)=\mathbf{A}_i^{(2)}$, $i=1,2,...,m_2$, *για κάθε* (x_1,x_2) *και* t, *τότε*

(α) $E(X_t^{(j)})=\boldsymbol{\pi}_{0,0}\sum_{r=1}^{t}\mathbf{B}^{r-1}\mathbf{D}_j\mathbf{1}'$, $j=1,2$

(β) $E(X_t^{(1)}X_t^{(2)})=\boldsymbol{\pi}_{0,0}\sum_{r=1}^{t}\left(\sum_{i=1}^{r-1}\mathbf{B}^{i-1}\mathbf{D}_2\mathbf{B}^{r-i-1}\mathbf{D}_1+\mathbf{B}^{r-1}\mathbf{D}_1\sum_{i=1}^{t-r}\mathbf{B}^{i-1}\mathbf{D}_2\right)\mathbf{1}'$

(γ) $M_j(w)=\sum_{t=1}^{\infty}E(X_t^{(j)})w^t=\frac{w}{1-w}\boldsymbol{\pi}_{0,0}(1-wB)^{-1}D_j\mathbf{1}'$, $j=1,2$

(δ) $M_{1,2}(w)=\sum_{t=1}^{\infty}E(X_t^{(1)}X_t^{(2)})w^t=\frac{w^2}{1-w}\boldsymbol{\pi}_{0,0}(\mathbf{I}-w\mathbf{B})^{-1}\left[\mathbf{D}_1(\mathbf{I}-w\mathbf{B})^{-1}\mathbf{D}_2+\mathbf{D}_2(\mathbf{I}-w\mathbf{B})^{-1}\mathbf{D}_1\right]\mathbf{1}'$

με $j=1,2$

όπου

$$\mathbf{B}=\mathbf{A}_0+\sum_{i=1}^{m_1}\mathbf{A}_i^{(1)}+\sum_{i=1}^{m_2}\mathbf{A}_i^{(2)},\ \textit{και}\ \mathbf{D}_1=\sum_{i=1}^{m_1}i\mathbf{A}_i^{(1)},\ \mathbf{D}_2=\sum_{i=1}^{m_2}i\mathbf{A}_i^{(2)}$$

Απόδειξη:

(α) Με χρήση της παρακάτω ισότητας η οποία ισχύει για τετραγωνικούς πίνακες $\mathbf{M}_i$,

$$\frac{d}{dz}\left(\sum_{i=0}^{k}z^i\mathbf{M}_i\right)^t=\sum_{r=1}^{t}\left[\left(\sum_{i=0}^{k}z^i\mathbf{M}_i\right)^{r-1}\left(\sum_{i=1}^{k}i\mathbf{M}_iz^{i-1}\right)\left(\sum_{i=0}^{k}z^i\mathbf{M}_i\right)^{t-r}\right]$$

και του Θεωρήματος 2.11, καταλήγουμε ότι,

$$E(X_t^{(1)})=\left[\frac{d}{dz_1}\left[\boldsymbol{\varphi}_t(z_1,z_2)\mathbf{1}'\right]\right]\Big|_{z_1=z_2=1}=\boldsymbol{\pi}_{0,0}\sum_{r=1}^{t}\mathbf{B}^{r-1}\mathbf{D}_1\mathbf{1}',$$

όπου θέσαμε

$$\mathbf{B}=\mathbf{A}_0+\sum_{i=1}^{m_1}\mathbf{A}_i^{(1)}+\sum_{i=1}^{m_2}\mathbf{A}_i^{(2)},\ \text{ως}\ \mathbf{D}_1=\sum_{i=1}^{m_1}i\mathbf{A}_i^{(1)},\ \text{και}\ \mathbf{D}_2=\sum_{i=1}^{m_2}i\mathbf{A}_i^{(2)}.$$

(με $\mathbf{B}\mathbf{1}'=\mathbf{1}'$).

(γ) Η γεννήτρια συνάρτηση των μέσων τιμών μπορεί να γραφτεί ως

$$M_1(w) = \boldsymbol{\pi}_{0,0} \sum_{t=1}^{\infty} \sum_{r=1}^{t} \mathbf{B}^{r-1} \mathbf{D}_1 w^t \mathbf{1}' = \boldsymbol{\pi}_{0,0} \sum_{r=1}^{\infty} (w\mathbf{B})^{r-1} \sum_{t=r}^{w} w^{t-r+1} \mathbf{D}_1 \mathbf{1}'$$

και το τελικό αποτέλεσμα προκύπτει με χρήση της σχέσης

$$\sum_{r=1}^{\infty} (w\mathbf{B})^{r-1} = (\mathbf{I} - w\mathbf{B})^{-1}.$$

(β) Οι μέσες τιμές των γινομένων μπορούν να γραφτούν ως

$$E(X_t^{(1)} X_t^{(2)}) = \frac{d^2}{dz_1 dz_2} \left[\boldsymbol{\varphi}_t(z_1, z_2) \mathbf{1}' \right] \Big|_{z_1 = z_2 = 1}$$

και με χρήση της σχέσης

$$\frac{d}{dz} \left[\left(\sum_{i=0}^{k} z^i \mathbf{M}_i \right)^{r-1} \left(\sum_{i=1}^{k} i \mathbf{M}_i z^{i-1} \right) \left(\sum_{i=0}^{k} z^i \mathbf{M}_i \right)^{t-r} \right] =$$

$$= \sum_{i=1}^{r-1} \left(\sum_{i=0}^{k} z^i \mathbf{M}_i \right)^{i-1} \left(\sum_{i=1}^{k} i \mathbf{M}_i z^{i-1} \right) \left(\sum_{i=0}^{k} z^i \mathbf{M}_i \right)^{r-i-1} \left(\sum_{i=1}^{k} i \mathbf{M}_i z^{i-1} \right) \left(\sum_{i=0}^{k} z^i \mathbf{M}_i \right)^{t-r} +$$

$$+ \left(\sum_{i=0}^{k} z^i \mathbf{M}_i \right)^{r-1} \left(\sum_{i=1}^{k} i \mathbf{M}_i z^{i-1} \right) \sum_{i=1}^{t-r} \left(\sum_{i=0}^{k} z^i \mathbf{M}_i \right)^{i-1} \left(\sum_{i=1}^{k} i \mathbf{M}_i z^{i-1} \right) \left(\sum_{i=0}^{k} z^i \mathbf{M}_i \right)^{t-r-i}$$

προκύπτει τελικά

$$E(X_t^{(1)} X_t^{(2)}) = \boldsymbol{\pi}_{0,0} \sum_{r=1}^{t} \left(\sum_{i=1}^{r-1} \mathbf{B}^{i-1} \mathbf{D}_2 \mathbf{B}^{r-i-1} \mathbf{D}_1 + \mathbf{B}^{r-1} \mathbf{D}_1 \sum_{i=1}^{t-r} \mathbf{B}^{i-1} \mathbf{D}_2 \right) \mathbf{1}'.$$

(δ) Η γεννήτρια συνάρτηση των μέσων τιμών προκύπτει με χρήση των παρακάτω ισοτήτων

$$\sum_{t=1}^{\infty} \sum_{r=1}^{t} \sum_{i=1}^{r-1} \mathbf{B}^{i-1} \mathbf{D}_2 \mathbf{B}^{r-1-i} w^t = \frac{w^2}{1-w} (\mathbf{I} - w\mathbf{B})^{-1} \mathbf{D}_2 (\mathbf{I} - w\mathbf{B})^{-1}$$

$$\sum_{t=1}^{\infty} \sum_{r=1}^{t} \sum_{i=1}^{t-r} \mathbf{B}^{r-1} \mathbf{D}_1 \mathbf{B}^{i-1} w^t = \frac{w^2}{1-w} (\mathbf{I} - w\mathbf{B})^{-1} \mathbf{D}_1 (\mathbf{I} - w\mathbf{B})^{-1}$$

οι οποίες μπορούν εύκολα να επαληθευτούν με απλές πράξεις. □

Στη συνέχεια ορίζεται η έννοια της πολυδιάστατης μεταβλητής πολυωνυμικού τύπου εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα (MMVP) κατά τρόπον ανάλογο με την διδιάστατη περίπτωση.

Ορισμός 2.6: *Μια θετική ακέραια p-διάστατη τυχαία μεταβλητή* $\left(X_n^{(1)}, X_n^{(2)}, ..., X_n^{(p)}\right)$ *ορισμένη στο p-διάστατο σύνολο* $\{0,1,2,...,\mathbf{l}_n\}^p$ *με* $n \in N$ *θα λέγεται* ***πολυδιάστατη μεταβλητή πολυωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα (MMVP)*** *αν*

- *Υπάρχει μια Μαρκοβιανή αλυσίδα διακριτού χρόνου* $\{Y_t : t \geq 0\}$ *ορισμένη στο χώρο καταστάσεων*

$$\Omega = \bigcup_{x_1, x_2, ..., x_p} C_{x_1, x_2, ..., x_p}, \quad C_{x_1, x_2, ..., x_p} = \{c_{x_1, x_2, ..., x_p, 0}, c_{x_1, x_2, ..., x_p, 1}, .., c_{x_1, x_2, ..., x_p, s-1}\}$$

- *Υπάρχουν p θετικοί αριθμοί* $m_1, m_2, ..., m_p$ *τέτοιοι ώστε για* $t \geq 1$,

$$P(Y_t \in C_{x_1, x_2, ..., x_p} \mid Y_{t-1} \in C_{y_1, y_2, ..., y_p}) = 0$$

για κάθε

$(x_1, x_2, .., x_p) \notin \{(y_1 + u_1, y_2, .., y_p), (y_1, y_2 + u_2, .., y_p), ..., (y_1, y_2, .., y_p + u_p)\}$,

με $0 \leq u_i \leq m_i$.

- *Ισχύει ότι*

$$P(X_n^{(1)} = x_1, X_n^{(2)} = x_2, ..., X_n^{(p)} = x_p) = P(Y_n \in C_{x_1, x_2, ..., x_p}),$$

$n \geq 0, \ x_1, x_2, ..., x_p \geq 0$.

Για τη μελέτη μιας MMVP χρειαζόμαστε το διάνυσμα αρχικών πιθανοτήτων

$$\boldsymbol{\pi}_{x_1, x_2, ..., x_p} = (P(Y_0 = c_{x_1, x_2, ..., x_p, 0}, P(Y_0 = c_{x_1, x_2, ..., x_p, 1}), ..., P(Y_0 = c_{x_1, x_2, ..., x_p, s-1})), \ x_1, x_2, ..., x_p \geq 0$$

τους «εσωτερικούς» πίνακες μετάβασης

$$\mathbf{A}_{t,0}(x_1, x_2, ..., x_p) = (P(Y_t = c_{x_1, x_2, ..., x_p, j'} \mid Y_{t-1} = c_{x_1, x_2, ..., x_p, j}))_{s \times s},$$

και τους «εξωτερικούς» πίνακες μετάβασης

$$\mathbf{A}_{t,u_j}^{(j)}(x_1, x_2, ..., x_p) = (P(Y_t = c_{x_1, x_2, ..., x_p, y'} \mid Y_{t-1} = c_{x_1, x_2, ..., x_p, y}),$$

$1 \leq j \leq p, 0 \leq u_j \leq m_j$.

Εισάγοντας τα διανύσματα πιθανότητας

$$\mathbf{f}_t(x_1, x_2, ..., x_p) = (P(Y_t = c_{x_1, x_2, ..., x_{p_2}, 0}), P(Y_t = c_{x_1, x_2, ..., x_{p_2}, 1}), ..., P(Y_t = c_{x_1, x_2, ..., x_{p_2}, s-1})),$$

όπου $x_1, x_2, ..., x_p \geq 0, \ t \geq 0$

προκύπτει

$$P(X_n^{(1)} = x_1,...,X_n^{(p)} = x_p) = P(Y_n \in C_{x_1,x_2,...,x_p}) = \mathbf{f}_n(x_1,...,x_p)\mathbf{1}'.$$

Πόρισμα 2.2: *Για μια p-διάστατη MMVP τυχαία μεταβλητή* $(X_n^{(1)}, X_n^{(2)},...,X_n^{(p)})$ *έχουμε*

$$\mathbf{f}_t(x_1,x_2,...,x_p) = \mathbf{f}_{t-1}(x_1,x_2,...,x_p)\mathbf{A}_{t,0}(x_1,x_2,...,x_p) + \sum_{j=1}^{p}\sum_{u_j=1}^{\min(m_j,x_j)}\mathbf{f}_{t-1}(x_1,...,x_j-u_j,...,x_p)\mathbf{A}_{t,u_j}^{(j)}(x_1,x_2,...,x_j-u_j,...,x_p)$$

για $t \geq 1$ *και* $x_i \geq 0, 1 \leq i \leq p$.

Πόρισμα 2.3: *Αν* $\mathbf{A}_{t,0}(x_1,x_2,...,x_p) = \mathbf{A}_{t,0}$, $\mathbf{A}_{t,u_j}^{(j)}(x_1,x_2,...,x_p) = \mathbf{A}_{t,u_j}^{(j)}$, $j = 1,2,...,p$, $u_j = 1,2,...,m_j$ *για κάθε* $t \geq 1$ *και* $x_1,x_2,...,x_p \geq 0$, *τότε η (διανυσματική) γεννήτρια συνάρτηση των διανυσμάτων* $\mathbf{f}_t(x_1,x_2,...,x_p)$ *δίνεται από τον τύπο*

$$\boldsymbol{\varphi}_t(z_1,z_2,...,z_p) = \sum_{x_1,x_2,...,x_p} P(X_n^{(1)} = x_1, X_n^{(2)} = x_2,...,X_n^{(P)} = x_P,)\prod_{j=1}^{P} z_j^{x_j} = \boldsymbol{\pi}_{0,0,...,0}\prod_{r=1}^{t}\left(\mathbf{A}_{r,0} + \sum_{j=1}^{p}\sum_{u_j=1}^{m_j}\mathbf{A}_{r,u_j}^{(j)} z_j^{u_j}\right),$$

όπου $\boldsymbol{\pi}_{0,0,...,0} = \mathbf{f}_0(0,0,...,0)$.

Πόρισμα 2.4: *Αν* $\mathbf{A}_{t,0}(x_1,x_2,...,x_p) = \mathbf{A}_0$, $\mathbf{A}_{t,u_j}^{(j)}(x_1,x_2,...,x_p) = \mathbf{A}_{u_j}^{(j)}$ $(j = 1,2,...,p, u_j = 1,...,m_j)$ *για κάθε* $(x_1,x_2,...,x_p)$ *και* $t \geq 0$, *τότε η διπλή (διανυσματική) γεννήτρια συνάρτηση των διανυσμάτων* $\mathbf{f}_t(x_1,x_2,...,x_p)$ *δίνεται από τον τύπο*

$$\boldsymbol{\Phi}(z_1,...,z_p;w) = \sum_{t=0}^{\infty}\boldsymbol{\varphi}_t(z_1,...,z_p)w^t = \boldsymbol{\pi}_{0,0}\left(\mathbf{I} - w\left(\mathbf{A}_0 + \sum_{i=1}^{p}\sum_{u_j=1}^{m_i}\mathbf{A}_{u_j}^{(j)} z_j^{u_j}\right)\right)^{-1}.$$

Η αποδείξεις των Πορισμάτων 2.2, 2.3, και 2.4 προκύπτουν με τρόπο ανάλογο της διδιάστατης περίπτωσης.

2.5. *Συμπεράσματα*

Η μέθοδος της εμφύτευσης τυχαίας μεταβλητής πολυωνυμικού τύπου σε Μαρκοβιανή αλυσίδα είναι ένα πολύ χρήσιμο εργαλείο. Αποτελεί τη βάση για την εύρεση της (ακριβούς) κατανομής μιας ευρύτερης κλάσης τυχαίων μεταβλητών στην οποία συμπεριλαμβάνονται οι περισσότερες τυχαίες μεταβλητές που σχετίζονται με τις ροές επιτυχιών, τα τμήματα σάρωσης και άλλους σχηματισμούς που μπορεί να παρουσιασθούν σε μια ακολουθία δοκιμών. Η χρήση της μεθοδολογίας των μεταβλητών τύπου MVP επιτρέπει ακόμη και τον υπολογισμό δεσμευμένων κατανομών.

Χρήσιμη είναι η σύγκριση της μεθόδου που προτείνουμε με την αρχική μέθοδο των Fu and Koutras (1994) και την μεταγενέστερη μέθοδο της εμφύτευσης τυχαίας μεταβλητής διωνυμικού τύπου σε Μαρκοβιανη αλυσίδα των Koutras and Alexandrou (1995).

Εύκολα διαπιστώνει κανείς ότι με τη μέθοδο των Fu and Koutras (1994) δεν υπάρχει η δυνατότητα να μελετήσουμε μεγάλες ακολουθίες, διότι οι πίνακες που προκύπτουν με αυτή την προσέγγιση του προβλήματος είναι μεγάλης διάστασης. Μάλιστα, όσο ο αριθμός n των δοκιμών της ακολουθίας αυξάνει, η διάσταση των πινάκων γίνεται πολύ μεγάλη. Επιπλέον, αν και προτείνονται εκφράσεις για τις ροπές και την πιθανογεννήτρια, οι τύποι αυτοί εφαρμόζονται δύσκολα.

Από την άλλη μεριά, η μέθοδος της εμφύτευσης τυχαίας μεταβλητής διωνυμικού τύπου σε Μαρκοβιανή αλυσίδα των Koutras and Alexandrou (1995) μας παρέχει τη δυνατότητα να υπολογίσουμε την ακριβή κατανομή μιας ευρείας κλάσης τυχαίων μεταβλητών απαρίθμησης σε ακολουθίες δοκιμών, οι οποίες έχουν δύο (ή περισσότερα) δυνατά αποτελέσματα σε κάθε δοκιμή. Η μέθοδος της εμφύτευσης τυχαίας μεταβλητής πολυωνυμικού τύπου σε Μαρκοβιανή αλυσίδα αντιμετωπίζει μια πολύ ευρύτερη κλάση προβλημάτων. Επιπλέον, όπως και η μέθοδος των Koutras and Alexandrou (1995) μπορεί να εφαρμοσθεί στην επίλυση προβλημάτων, όπου οι δοκιμές της σειράς των πειραμάτων δεν είναι ανεξάρτητες και ισόνομες. Αυτό αποτελεί ένα ιδιαίτερο πλεονέκτημα, αφού στην περίπτωση που οι πιθανότητες επιτυχίας διαφέρουν από δοκιμή σε δοκιμή, πολλές στατιστικές συναρτήσεις είναι δύσκολο, αν όχι αδύνατο, να μελετηθούν με άλλους τρόπους.

Η μέθοδος της εμφύτευσης τυχαίας μεταβλητής πολυωνυμικού τύπου σε Μαρκοβιανή αλυσίδα δημιουργεί ένα πλαίσιο μέσα στο οποίο μια μεγάλη ομάδα τυχαίων μεταβλητών μπορεί να μελετηθεί θεωρητικά. Με τη βοήθεια της μεθόδου αυτής προκύπτουν γενικά συμπεράσματα που αφορούν εκφράσεις για τη γεννήτρια συνάρτηση, τη μέση τιμή και τη γεννήτρια συνάρτηση των μέσων τιμών.

2.6. Ανακεφαλαίωση

Στο κεφάλαιο αυτό κάναμε μια αναλυτική παρουσίαση της μεθόδου της Μαρκοβιανής εμφύτευσης τυχαίων μεταβλητών. Πιο συγκεκριμένα, δώσαμε μια σειρά από αποτελέσματα τα οποία οφείλονται στους Fu and Koutras (1994) και Koutras and Alexandrou (1995). Στη συνέχεια εισάγαμε την έννοια της μονοδιάστατης και της πολυδιάστατης μεταβλητής πολυωνυμικού τύπου εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα η οποία και αποτελεί ένα σημαντικό κομμάτι της παρούσας διατριβής.

ΚΕΦΑΛΑΙΟ 3: ΜΟΝΟΔΙΑΣΤΑΤΕΣ ΚΑΤΑΝΟΜΕΣ ΣΧΕΤΙΚΕΣ ΜΕ ΡΟΕΣ ΕΠΙΤΥΧΙΩΝ

3.1. Εισαγωγή

Στο Κεφάλαιο 3, εφαρμόζουμε τα αποτελέσματα του Κεφαλαίου 2, σε τυχαίες μεταβλητές που σχετίζονται με ροές επιτυχιών στη μονοδιάστατη περίπτωση. Οι διωνυμικές κατανομές τάξης k (όπως ονομάζονται οι κατανομές που σχετίζονται με ροές επιτυχιών σε πεπερασμένες ακολουθίες δίτιμων δοκιμών) μελετήθηκαν διεξοδικά από τους Koutras and Alexandrou (1995) με την τεχνική της εμφύτευσης MVB. Στο Κεφάλαιο αυτό, παρουσιάζονται οι περιπτώσεις των μη επικαλυπτόμενων, των επικαλυπτόμενων ροών επιτυχιών μήκους k και των ροών επιτυχιών μήκους τουλάχιστον k. Με τη μέθοδο MVB παρέχεται η δυνατότητα υπολογισμού της συνάρτησης πιθανότητας και της μέσης τιμής των μεταβλητών, με τη βοήθεια απλών αναδρομικών σχέσεων καθώς και πιθανογεννητριών συναρτήσεων.

Στη συνέχεια μελετάμε διεξοδικά μια νέα τυχαία μεταβλητή η οποία ισούται με το άθροισμα των μηκών των ροών επιτυχιών μήκους τουλάχιστον k σε μια πεπερασμένη ακολουθία ανεξάρτητων δοκιμών Bernoulli. Η μελέτη της συγκεκριμένης μεταβλητής γίνεται με τη μέθοδο MVP η οποία εισήχθη στο προηγούμενο κεφάλαιο ενώ δίνονται κατευθύνσεις για τη μελέτη της μεταβλητής αυτής και στην περίπτωση που η ακολουθία των δοκιμών είναι Μαρκοβιανά εξαρτημένη. Τέλος, μελετάμε τη δεσμευμένη κατανομή της ίδιας τυχαίας μεταβλητής, δοθέντος του αριθμού των επιτυχιών. Η μελέτη της δεσμευμένης αυτής κατανομής βασίστηκε πάνω στη γενική μεθοδολογία των Koutras and Alexandrou (1997) και παρουσιάζει ιδιαίτερα σημαντική

χρησιμότητα στην μη παραμετρική στατιστική ως έλεγχος τυχαιότητας (βλέπε επίσης και Antzoulakos et al. (2003)).

3.2. Μελέτη Κατανομών, Σχετικών με Απαρίθμηση Ροών Επιτυχιών, σε Ακολουθίες Ανεξάρτητων Δοκιμών Bernoulli

Οι Koutras and Alexandrou (1995) με χρήση της μεθοδολογίας εμφύτευσης διωνυμικού τύπου τυχαίων μεταβλητών σε Μαρκοβιανή αλυσίδα μελέτησαν τα κυριότερα είδη ροών. Συγκεκριμένα, μελέτησαν τις περιπτώσεις των μη επικαλυπτόμενων, των επικαλυπτόμενων ροών επιτυχιών μήκους k και των ροών επιτυχιών μήκους τουλάχιστον k. Προκειμένου να προχωρήσουν στην μελέτη αυτή έδωσαν έναν αυστηρό και κατά κάποιο τρόπο ενιαίο ορισμό των διαφορετικών ειδών ροών επιτυχιών.

Ορισμός 3.1: *Έστω $Z_1, Z_2, ..., Z_n$ μια ακολουθία ανεξάρτητων και ισόνομων δοκιμών Bernoulli με δυνατά αποτελέσματα επιτυχία (1) και αποτυχία (0) και αντίστοιχες πιθανότητες $p_t = P(Z_t = 1)$ και $q_t = 1 - p_t$ για $t = 1,2,...,n$ με αρχική συνθήκη $Z_0 = 0$. Θεωρούμε την τυχαία μεταβλητή*

$$W_t = \prod_{j=t}^{t+k-1} Z_j, \ t = 1,2,...,n-k+1$$

και έστω

$$W_t' = \begin{cases} W_t, & an \sum_{j=1}^{k-1} W_{t-j} = 0 \\ 0, & diaforetiká \end{cases}, \ t = 1,2,3,...$$

όπου θέτουμε ότι $W_t' = 0 \, gia \ t \le 0$. Τότε,

- *Η απαριθμήτρια τυχαία μεταβλητή $N_{n,k}$ των μη επικαλυπτόμενων ροών επιτυχιών μήκους k, μπορεί να εκφραστεί ως $N_{n,k} = \sum_{t=1}^{n-k+1} W_t'$*
- *Η απαριθμήτρια τυχαία μεταβλητή $M_{n,k}$ των επικαλυπτόμενων ροών επιτυχιών*

μήκους k, *μπορεί να εκφραστεί ως* $M_{n,k} = \sum_{t=1}^{n-k+1} W_t$

- *Η απαριθμήτρια τυχαία μεταβλητή* $G_{n,k}$ *των ροών επιτυχιών μήκους τουλάχιστον* k, *μπορεί να εκφραστεί ως* $G_{n,k} = \sum_{t=1}^{n-k+1} (1 - Z_{t-1}) W_t$.

Στη συνέχεια παρουσιάζουμε τα κυριότερα αποτελέσματα των Koutras and Alexandrou (1995), για τις τυχαίες μεταβλητές $N_{n,k}$, $M_{n,k}$ *και* $G_{n,k}$. Για όλα τα υπό μελέτη είδη απαριθμητριών πριν την παρουσίαση των κυριοτέρων αποτελεσμάτων, δίνουμε κατευθύνσεις, υπό την μορφή εφαρμογών για την κατασκευή των απαραίτητων πινάκων.

Στην Εφαρμογή 3.1 δίνουμε βήμα προς βήμα την μεθοδολογία εμφύτευσης της απαριθμήτριας τυχαίας μεταβλητής $N_{n,k}$ σε Μαρκοβιανή αλυσίδα χρησιμοποιώντας την τεχνική των Koutras and Alexandrou (1995).

Εφαρμογή 3.1: Τεχνική Εμφύτευσης της τυχαίας μεταβλητής $N_{n,k}$

Υποθέτουμε ότι $Z_1, Z_2, \ldots, Z_n$ είναι μια ακολουθία ανεξάρτητων δοκιμών Bernoulli με δυνατά αποτελέσματα επιτυχία (1) και αποτυχία (0) και αντίστοιχες πιθανότητες $p_t = P(Z_t = 1)$ και $q_t = 1 - p_t$ για $t = 1,2,\ldots,n$.

Θα εφαρμόσουμε τη μέθοδο της Μαρκοβιανής εμφύτευσης προκειμένου να υπολογίσουμε την κατανομή της τυχαίας μεταβλητής $N_{n,k}$ η οποία καταγράφει τον αριθμό των μη επικαλυπτόμενων ροών μήκους k.

Βήμα 1: Θεωρούμε τον χώρο καταστάσεων

$$\Omega = \{(x,i) : x = 0,1,2,\ldots,\mathbf{l}_n,\ i = 0,1,2,\ldots,k-1\} \text{ με } \mathbf{l}_n = \left[\frac{n}{k}\right].$$

Βήμα 2: Θεωρούμε τη Μαρκοβιανή αλυσίδα $\{Y_t : t \geq 0\}$ ορισμένη στον χώρο Ω. Ορίζουμε $Y_t = (x,i)$, εάν στον χρόνο $t \geq 1$ έχουν εμφανισθεί x μη επικαλυπτόμενες ροές μήκους k και επίσης i επιτυχίες από την τελευταία εμφάνιση αποτυχίας ή από την συμπλήρωση της τελευταίας ροής μήκους k.

Βήμα 3: Για κάποιο $0 \leq x \leq \mathbf{l}_n$ συμβολίζουμε με $C_x = \{(x,i) : i = 0,1,\ldots,k-1\}$

(δηλαδή το σύνολο των πιθανών υποκαταστάσεων της αλυσίδας). Το σύνολο των C_x, για κάθε x, αποτελεί μια διαμέριση του χώρου

$$\Omega = \{(x,i) : x = 0,1,2,...,\mathbf{l}_n, i = 0,1,2,...,k-1\}.$$

Βήμα 4: Ισχύει ότι για κάθε $x = \{0,1,2,....,\mathbf{l}_n\}$ ισχύει $P(N_{n,k} = x) = P(Y_n \in C_x)$.

Βήμα 5: Επίσης, ισχύει ότι $P(Y_t = c_{y,j} \mid Y_{t-1} = c_{x,i}) = 0$ για κάθε $y \neq x, x+1$, $t = 1,...,n$.

Από τα βήματα που προηγήθηκαν προκύπτει εύκολα ότι ικανοποιούνται τόσο ο Ορισμός 2.1 όσο και ο Ορισμός 2.2. Συνεπώς η τυχαία μεταβλητή $N_{n,k}$ εμφυτεύεται σε Μαρκοβιανή αλυσίδα.

Το επόμενο βήμα της μεθόδου είναι να ορίσουμε τις αρχικές πιθανότητες της αλυσίδας ίσες με $\boldsymbol{\pi}_0 = (1,0,0,...,0)$ και τους πίνακες $\mathbf{A}_t(x)$ και $\mathbf{B}_t(x)$. Οι πιθανότητες μετάβασης (στοιχεία των $\mathbf{A}_t(x)$ και $\mathbf{B}_t(x)$) πρώτης τάξης δίνονται από τις ακόλουθες σχέσεις:

$$P(Y_t = (x,i+1) \mid Y_{t-1} = (x,i)) = p_t,\ x = 0,1,2,...,\mathbf{l}_n,\ i = 0,1,2,...,k-2,$$

$$P(Y_t = (x+1,0) \mid Y_{t-1} = (x,i)) = p_t,\ x = 0,1,2,...,\mathbf{l}_n - 1,\ i = k-1.$$

Με χρήση των ανωτέρω σχέσεων οι οποίες και δίνουν τα μη μηδενικά στοιχεία των πινάκων $\mathbf{A}_t(x)$ και $\mathbf{B}_t(x)$, για οποιαδήποτε k και n, έχουμε:

$$\mathbf{A}_t(x) = \begin{bmatrix} & (x,0) & (x,1) & (x,2) & \cdots & (x,k-1) \\ (x,0) & q_t & p_t & 0 & \cdots & 0 \\ (x,1) & q_t & 0 & p_t & \cdots & 0 \\ \vdots & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{O} & 0 \\ (x,k-2) & q_t & 0 & 0 & \cdots & p_t \\ (x,k-1) & q_t & 0 & 0 & \cdots & 0 \end{bmatrix}_{k\times k}, x = 0,1,2,...,\mathbf{l}_n$$

$$\mathbf{B}_t(x) = \begin{bmatrix} & (x+1,0) & (x+1,1) & (x+1,2) & \cdots & (x+1,k-1) \\ (x,0) & 0 & 0 & 0 & \cdots & 0 \\ (x,1) & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{O} & 0 \\ (x,k-2) & 0 & 0 & 0 & \cdots & 0 \\ (x,k-1) & p_t & 0 & 0 & \cdots & 0 \end{bmatrix}_{k\times k}$$

Τα στοιχεία του πίνακα $\mathbf{A}_t(x)$ αποτελούν τις πιθανότητες μετάβασης ανάμεσα

στις καταστάσεις ($c_{x,i}$, $i = 0,1,2,...,s-1$), ενώ τα στοιχεία του πίνακα $\mathbf{B}_t(x)$ αποτελούν τις πιθανότητες μετάβασης ανάμεσα στα σύνολα καταστάσεων (C_x, $x \geq 0$). Είναι φανερό ότι ισχύει $[\mathbf{A}_t(x) + \mathbf{B}_t(x)]\mathbf{1}' = \mathbf{1}'$. Τέλος, με χρήση του Θεωρήματος 2.2 και των πινάκων $\mathbf{A}_t(x)$, $\mathbf{B}_t(x)$ είναι δυνατός ο υπολογισμός της συνάρτησης πιθανότητας της $N_{n,k}$.

Στις περισσότερες εφαρμογές οι πίνακες $\mathbf{A}_t(x)$ και $\mathbf{B}_t(x)$ είναι ανεξάρτητοι του x. Έτσι, στην ειδική περίπτωση $k = 2$ έχουμε:

$$\mathbf{A}_t = \begin{bmatrix} q_t & p_t \\ q_t & 0 \end{bmatrix} \text{ και } \mathbf{B}_t = \begin{bmatrix} 0 & 0 \\ p_t & 0 \end{bmatrix}.$$

Στο Πόρισμα 3.1 δίνονται εκφράσεις για την πιθανογεννήτρια συνάρτηση, τη διπλή γεννήτρια συνάρτηση καθώς και για την γεννήτρια μέσων τιμών της τυχαίας μεταβλητής $N_{n,k}$. Οι εκφράσεις αυτές οφείλονται στους Koutras and Alexandrou (1995) αλλά όπως έχουμε προαναφέρει αποτελούν ειδικές περιπτώσεις των εκφράσεων της Παραγράφου 3.3.

Πόρισμα 3.1: *Έστω $N_{n,k}$ ο αριθμός των μη επικαλυπτόμενων ροών επιτυχιών μήκους k σε μια ακολουθία ανεξάρτητων και ισόνομων δοκιμών Bernoulli. Τότε,*

- *Αν*

$$f_n(z) = \sum_{x=0}^{\mathbf{l}_n} P(N_{n,k} = x) z^x$$

είναι η πιθανογεννήτρια συνάρτηση, ισχύει ότι

$$f_n(z) = \begin{cases} 1, & n < k \\ (1 - p^k) + p^k z, & n = k \\ \sum_{i=0}^{k-1} p^i q f_{n-i-1}(z) + p^k z f_{n-k}(z), & n > k \end{cases}$$

- *Αν*

$$\Phi_n(z, w) = \sum_{n=0}^{\infty} \sum_{x=0}^{\mathbf{l}_n} P(N_{n,k} = x) z^x w^n$$

είναι η διπλή γεννήτρια συνάρτηση, ισχύει ότι

$$\Phi_n(z,w)=\frac{1-(pw)^k}{1-w+qw(pw)^k-(1-pw)(pw)^k z}$$

- *Αν*

$$M_n(w)=\sum_{n=0}^{\infty} m_n w^n$$

είναι η γεννήτρια συνάρτηση των μέσων τιμών $m_n=E(N_{n,k})$, *ισχύει ότι*

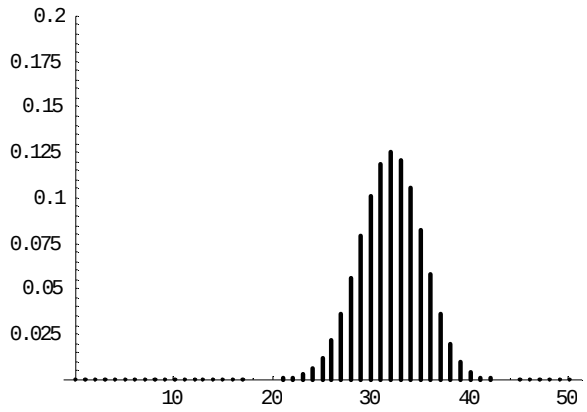
$$M_n(w)=\frac{(pw)^k(1-pw)}{(1-w)^2(1-(pw)^k)},$$

καθώς και το ακόλουθο αναδρομικό σχήμα

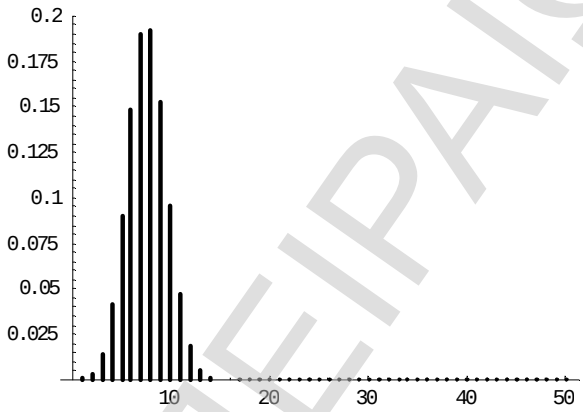
$$m_n=\begin{cases}0, & n<k\\ p^k, & n=k\\ p^k(1+q), & n=k+1\\ 2m_{n-1}-m_{n-2}+p^k(m_{n-k}-2m_{n-k-1}+m_{n-k-2}), & n>k+1.\end{cases}$$

Στο Σχήμα 3.1 δίνεται η συνάρτηση πιθανότητας της $N_{n,k}$ για διάφορες τιμές των n, k, p.

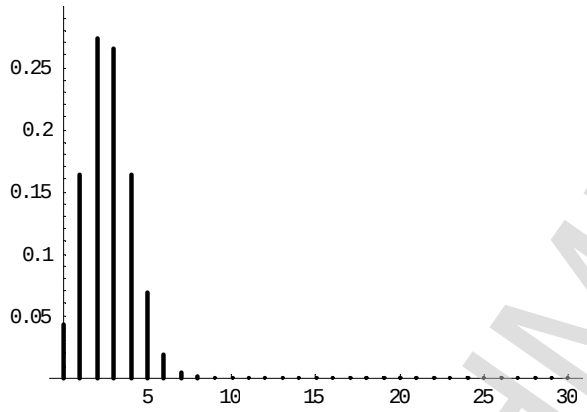
Σχήμα 3.1: Η συνάρτηση πιθανότητας της $N_{n,k}$ για διάφορες τιμές των n, k, p



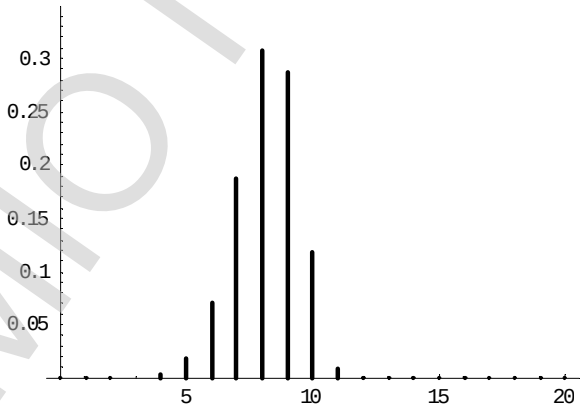
$n = 100,\ p = 0.75, k = 2.$



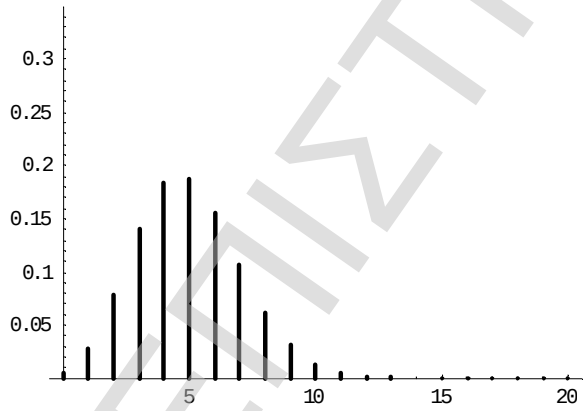
$n = 100,\ p = 0.75, k = 5.$



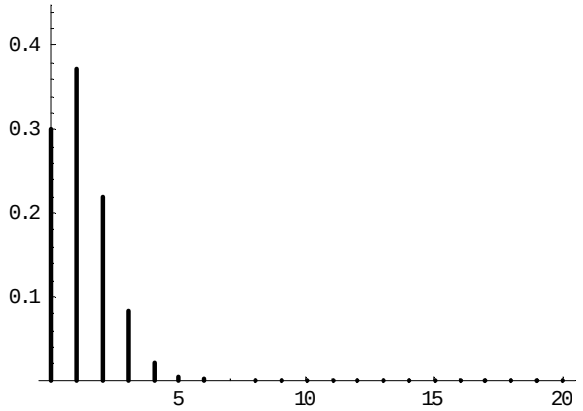
$n = 100,\ p = 0.75, k = 8$



$n = 100,\ p = 0.95, k = 9$



$n = 100,\ p = 0.25, k = 2$



$n = 100,\ p = 0.25, k = 3$

Συνεχίζοντας, στην Εφαρμογή 3.2 δίνουμε βήμα βήμα την μεθοδολογία εμφύτευσης της απαριθμήτριας τυχαίας μεταβλητής $M_{n,k}$ σε Μαρκοβιανή αλυσίδα, όπως δόθηκε από τους Fu and Koutras (1994) και Koutras and Alexandrou (1995).

Εφαρμογή 3.2: Τεχνική Εμφύτευσης της τυχαίας μεταβλητής $M_{n,k}$

Υποθέτουμε ότι $Z_1, Z_2, ..., Z_n$ είναι μια ακολουθία ανεξάρτητων και ισόνομων δοκιμών Bernoulli με δυνατά αποτελέσματα επιτυχία (1) και αποτυχία (0) και αντίστοιχες πιθανότητες $p_t = P(Z_t = 1)$ και $q_t = 1 - p_t$ για $t = 1,2,...,n$. Θα εφαρμόσουμε τη μέθοδο της Μαρκοβιανής εμφύτευσης προκειμένου να υπολογίσουμε την κατανομή της τυχαίας μεταβλητής $M_{n,k}$ η οποία καταγράφει τον αριθμό των επικαλυπτόμενων ροών μήκους k.

Βήμα 1: Θεωρούμε τον χώρο καταστάσεων

$$\Omega = \{(x,i) : x = 0,1,2,..., \mathbf{l}_n, i = -1,0,1,2,...,k-1\} \text{ με } \mathbf{l}_n = n - k + 1.$$

Βήμα 2: Θεωρούμε τη Μαρκοβιανή αλυσίδα $\{Y_t : t \geq 0\}$ ορισμένη στον χώρο Ω. Ορίζουμε,

- $Y_t = (x,i)$, $i = 0,1,...,k-1$ εάν στο χρόνο $t \geq 0$ έχουν εμφανισθεί x μη επικαλυπτόμενες ροές μήκους k και επίσης i επιτυχίες από την τελευταία εμφάνιση αποτυχίας ή από τη συμπλήρωση της τελευταίας ροής μήκους k.
- $Y_t = (x,-1)$ εάν στο χρόνο $t \geq 0$ έχουν εμφανισθεί x μη επικαλυπτόμενες ροές μήκους k, ενώ οι τελευταίες διαδοχικές δοκιμές είναι επιτυχίες και ξεπερνούν τις $k-1$.

Βήμα 3: Για κάποιο $0 \leq x \leq \mathbf{l}_n$ συμβολίζουμε με $C_x = \{(x,i) : i = -1,0,1,2,...\}$ (δηλαδή το σύνολο των πιθανών υποκαταστάσεων της αλυσίδας). Το σύνολο των C_x για κάθε x αποτελεί μια διαμέριση του χώρου

$$\Omega = \{(x,i) : x = 0,1,2,..., \mathbf{l}_n, i = -1,0,1,2,...,k-1\}.$$

Βήμα 4: Προφανώς ισχύει ότι για κάθε $x = \{0,1,2,....,\mathbf{l}_n\}$ ισχύει $P(X_n = x) = P(Y_n \in C_x)$.

Βήμα 5: Επίσης, ισχύει ότι $P(Y_t = c_{y,j} \mid Y_{t-1} = c_{x,i}) = 0$ για κάθε $y \neq x, x+1$,

$t = 1,...,n$.

Από τα βήματα που προηγήθηκαν προκύπτει εύκολα ότι ικανοποιούνται τόσο ο Ορισμός 2.1 όσο και ο Ορισμός 2.2. Συνεπώς η τυχαία μεταβλητή $M_{n,k}$ εμφυτεύεται σε Μαρκοβιανή αλυσίδα.

Το επόμενο βήμα της μεθόδου είναι να ορίσουμε τις αρχικές πιθανότητες της αλυσίδας ίσες με $\boldsymbol{\pi}_0 = (1,0,0,...,0)$ και τους πίνακες $\mathbf{A}_t(x)$ και $\mathbf{B}_t(x)$. Οι πιθανότητες μετάβασης πρώτης τάξης δίνονται από τις σχέσεις:

$$P(Y_t = (x,0) \mid Y_{t-1} = (x,i)) = q_t = 1 - p_t, \;\; x = 0,1,2,...,\mathbf{l}_n, \;\; i = -1,0,1,2,...,k-1$$

$$P(Y_t = (x,i+1) \mid Y_{t-1} = (x,i)) = p_t, \;\; x = 0,1,2,...,\mathbf{l}_n, \;\; i = 0,1,2,...,k-2$$

$$P(Y_t = (x+1,-1) \mid Y_{t-1} = (x,i)) = p_t, \;\; x = 0,1,2,...,\mathbf{l}_n - 1, \;\; i = k-1$$

$$P(Y_t = (x+1,-1) \mid Y_{t-1} = (x,-1)) = p_t, \;\; x = 0,1,2,...,\mathbf{l}_n$$

Με χρήση των παραπάνω σχέσεων οι οποίες και δίνουν τα μη μηδενικά στοιχεία των πινάκων μπορούμε να ορίσουμε τους $\mathbf{A}_t(x)$ και $\mathbf{B}_t(x)$ για οποιοδήποτε k ως εξής:

$$\mathbf{A}_t(x) = \begin{bmatrix} & (x,0) & (x,1) & (x,2) & \mathbb{M} & (x,-1) \\ (x,0) & q_t & p_t & 0 & \mathbb{M} & 0 \\ (x,1) & q_t & 0 & p_t & \mathbb{M} & 0 \\ \mathbb{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{O} & 0 \\ (x,k-2) & q_t & 0 & 0 & \mathbb{M} & 0 \\ (x,-1) & q_t & 0 & 0 & \mathbb{M} & 0 \end{bmatrix}_{k \times k}$$

$$\mathbf{B}_t(x) = \begin{bmatrix} & (x+1,0) & (x+1,1) & (x+1,2) & \mathbb{M} & (x+1,-1) \\ (x,0) & 0 & 0 & 0 & \mathbb{M} & 0 \\ (x,1) & 0 & 0 & 0 & \mathbb{M} & 0 \\ \mathbb{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{O} & 0 \\ (x,k-2) & 0 & 0 & 0 & \mathbb{M} & p_t \\ (x,-1) & 0 & 0 & 0 & \mathbb{M} & p_t \end{bmatrix}_{k \times k}$$

Τα στοιχεία του πίνακα $\mathbf{A}_t(x)$ και σε αυτήν την περίπτωση αποτελούν τις πιθανότητες μετάβασης ανάμεσα στις καταστάσεις $c_{x,i}$, $i = -1,0,1,2,...,s-1$, ενώ τα στοιχεία του πίνακα $\mathbf{B}_t(x)$ αποτελούν τις πιθανότητες μετάβασης ανάμεσα στα σύνολα καταστάσεων $C_x, x \geq 0$. Με χρήση του Θεωρήματος 2.2 και των πινάκων

$\mathbf{A}_t(x),\ \mathbf{B}_t(x)$, μπορούμε να υπολογίσουμε τη συνάρτηση πιθανότητας της $M_{n,k}$.

Όπως ήδη έχει αναφερθεί, στις περισσότερες εκ των εφαρμογών οι πίνακες $\mathbf{A}_t(x),\ \mathbf{B}_t(x)$ είναι ανεξάρτητοι του x (ή ακόμα και του χρόνου *t*). Έτσι, σε αυτήν την περίπτωση και για $k=2$ έχουμε δύο πίνακες διαστάσεων 3×3 **A** και **B** της μορφής

$$\mathbf{A}=\begin{bmatrix} q & p & 0 \\ q & 0 & 0 \\ q & 0 & 0 \end{bmatrix} \text{ και } \mathbf{B}=\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & p \\ 0 & 0 & p \end{bmatrix}.$$

Με εφαρμογή των Θεωρημάτων 2.2, 2.3, και 2.4, προκύπτουν οι εκφράσεις του ακόλουθου Πορίσματος 3.2 για την πιθανογεννήτρια, την διπλή πιθανογεννήτρια καθώς και για την γεννήτρια μέσων τιμών της τυχαίας μεταβλητής $M_{n,k}$.

Πόρισμα 3.2: Έστω $M_{n,k}$ *ο αριθμός των μη επικαλυπτόμενων ροών επιτυχιών μήκους* k *σε μια ακολουθία ανεξάρτητων και ισόνομων δοκιμών* Bernoulli. *Τότε,*

- Αν

$$\Phi(z,w)=\sum_{n=0}^{\infty}\sum_{x=0}^{\mathbf{1}_n} P(M_{n,k}=x)z^x w^n$$

είναι η διπλή γεννήτρια συνάρτηση, ισχύει ότι

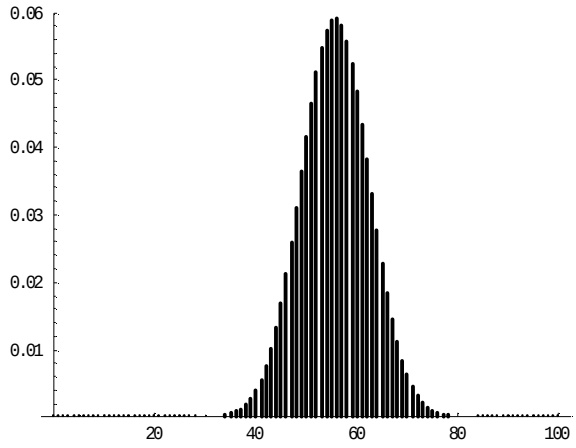
$$\Phi(z,w)=\frac{(1-pzw)(1-(pw)^k)+(pw)^k z(1-pw)}{(1-w)(1-pzw)+(1-z)qp^k w^{k+1}}$$

- Αν $M_n(w)=\sum_{n=0}^{\infty} m_n w^n$ *είναι η γεννήτρια συνάρτηση των μέσων τιμών με* $m_n=E(M_{n,k})$, *ισχύει ότι*

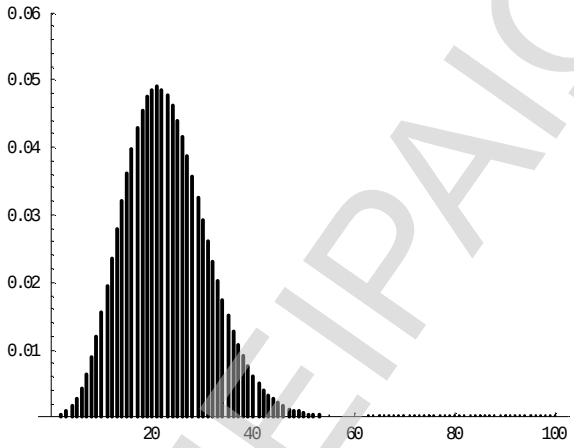
$$M_n(w)=\frac{(pw)^k}{(1-w)^2} \text{ και } E(M_{n,k})=(n-k+1)p^k.$$

Επίσης, στο Σχήμα 3.2 δίνεται η συνάρτηση πιθανότητας της $f(M_{n,k})$ για διάφορες τιμές των παραμέτρων n,k.

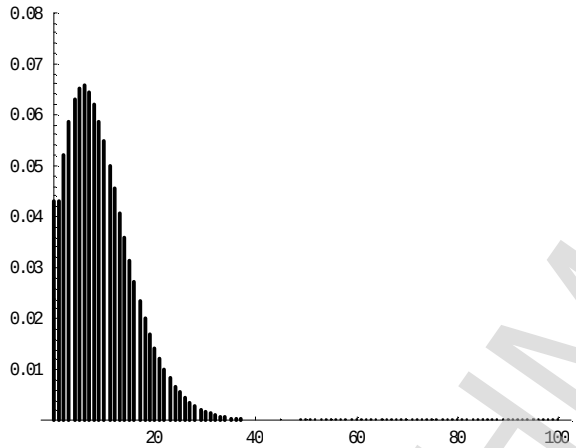
Σχήμα 3.2: Η συνάρτηση πιθανότητας της $M_{n,k}$ για διάφορες τιμές των n, k, p



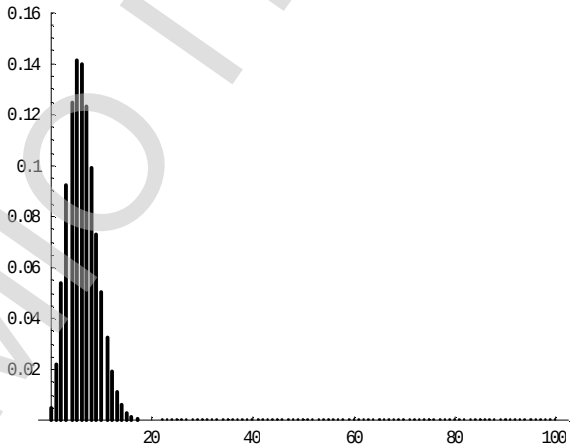
$n = 100,\ p = 0.75, k = 2.$



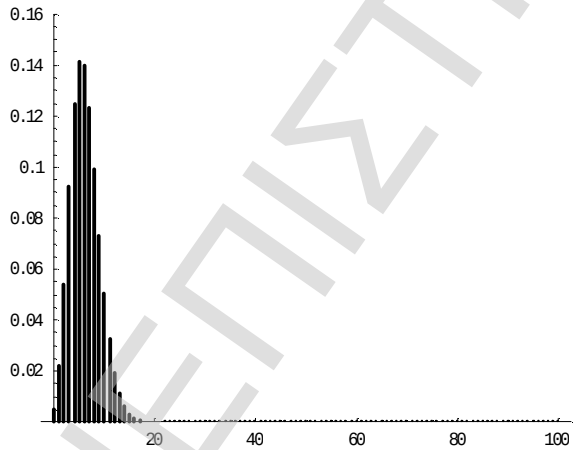
$n = 100,\ p = 0.75, k = 5.$



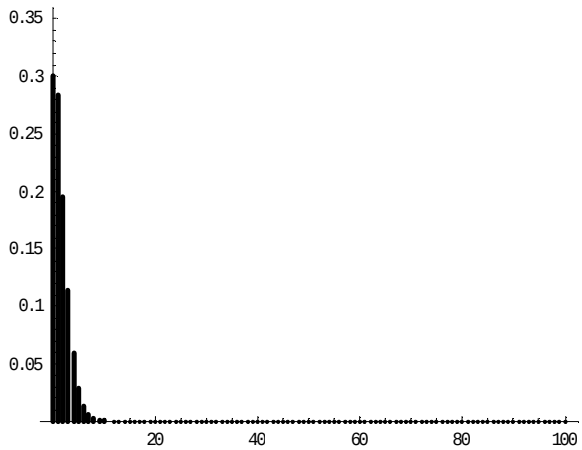
$n = 100,\ p = 0.75, k = 8$



$n = 100,\ p = 0.35, k = 2$



$n = 100,\ p = 0.25, k = 2$



$n = 100,\ p = 0.25, k = 3$

Η συνάρτηση πιθανότητας της $M_{n,k}$ του Σχήματος 3.2 υπολογίστηκε με χρήση της Μαρκοβιανής προσέγγισης και του αναδρομικού τύπου όπως ορίσθηκε στο Πόρισμα

3.2. Αξίζει κάποιος να παρατηρήσει την μορφή της συγκεκριμένης τυχαίας μεταβλητής για τις διάφορες τιμές των παραμέτρων n, k, p. Η συγκεκριμένη μεταβλητή για μικρές τιμές του k, έχει μορφή πολύ κοντά στην Διωνυμική κατανομή.

Στη συνέχεια και στην Εφαρμογή 3.3 δίνουμε αναλυτικά την μεθοδολογία εμφύτευσης της απαριθμήτριας τυχαίας μεταβλητής $G_{n,k}$.

Εφαρμογή 3.3: Τεχνική Εμφύτευσης της τυχαίας μεταβλητής $G_{n,k}$

Υποθέτουμε ότι $Z_1, Z_2, ..., Z_n$ είναι μια ακολουθία ανεξάρτητων και ισόνομων δοκιμών Bernoulli με δυνατά αποτελέσματα επιτυχία (1) και αποτυχία (0) και αντίστοιχες πιθανότητες $p_t = P(Z_t = 1)$ και $q_t = 1 - p_t$ για $t = 1,2,...,n$. Θα εφαρμόσομε την μέθοδο της Μαρκοβιανής εμφύτευσης προκειμένου να υπολογίσουμε την κατανομή της τυχαίας μεταβλητής $G_{n,k}$ η οποία καταγράφει τον αριθμό των ροών μήκους τουλάχιστον k.

Βήμα 1: Θεωρούμε τον χώρο καταστάσεων

$$\Omega = \{(x,i) : x = 0,1,2,...,\mathbf{l}_n, i = -1,0,1,2,...,k-1\} \text{ με } \mathbf{l}_n = \left[\frac{n+1}{k+1}\right].$$

Βήμα 2: Θεωρούμε την Μαρκοβιανή αλυσίδα $\{Y_t : t \geq 0\}$ ορισμένη στον χώρο Ω. Υποθέτουμε ότι στις πρώτες t δοκιμές και στο τέλος αυτών έχουμε i συνεχόμενες επιτυχίες. Ορίζουμε

- $Y_t = (x,i)$, αν $0 \leq i \leq k-1$ και στον χρόνο $t \geq 0$ έχουν εμφανισθεί $x \geq 0$ ροές επιτυχιών μήκους τουλάχιστον k στις δοκιμές που προηγούνται των i τελικών.
- $Y_t = (x,-1)$, αν $i \geq k$ και στον χρόνο $t \geq 0$ έχουν εμφανισθεί $x - 1 \geq 0$ ροές επιτυχιών μήκους τουλάχιστον k στις δοκιμές που προηγούνται των i τελικών.

Βήμα 3: Για κάποιο $0 \leq x \leq \mathbf{l}_n$ συμβολίζουμε με C_x το σύνολο των πιθανών υποκαταστάσεων της αλυσίδας. Το σύνολο των C_x για κάθε x αποτελεί μια διαμέριση του χώρου $\Omega = \{(x,i) : x = 0,1,2,...,\mathbf{l}_n, i = -1,0,1,2,...,k-1\}$.

Βήμα 4: Προφανώς για κάθε $x = \{0,1,2,....,\mathbf{l}_n\}$ ισχύει ότι $P(X_n = x) = P(Y_n \in C_x)$.

Από τα βήματα που προηγήθηκαν προκύπτει εύκολα ότι ικανοποιείται ο Ορισμός 2.1 και συνεπώς η τυχαία μεταβλητή $G_{n,k}$ εμφυτεύεται σε Μαρκοβιανή αλυσίδα.

Βήμα 5: Επίσης, ισχύει ότι $P(Y_t = c_{y,j} \mid Y_{t-1} = c_{x,i}) = 0$ για κάθε $y \neq x, x+1$, $t = 1,\ldots,n$.

Και στην περίπτωση της τυχαίας μεταβλητής $G_{n,k}$ επαληθεύονται οι συνθήκες των Ορισμών 2.1 και 2.2. Το επόμενο βήμα της μεθόδου είναι να ορίσουμε τις αρχικές πιθανότητες της αλυσίδας ίσες με $\boldsymbol{\pi}_0 = (1,0,0,\ldots,0)$, και τους πίνακες $\mathbf{A}_t(x)$ και $\mathbf{B}_t(x)$. Οι πιθανότητες μετάβασης πρώτης τάξης δίνονται από τις σχέσεις:

$$P(Y_t = (x,0) \mid Y_{t-1} = (x,i)) = q_t = 1 - p_t, \quad x = 0,1,2,\ldots,\mathbf{l}_n, \quad i = -1,0,1,2,\ldots,k-1,$$

$$P(Y_t = (x,i+1) \mid Y_{t-1} = (x,i)) = p_t, \quad x = 0,1,2,\ldots,\mathbf{l}_n, \quad i = \max(0,k-2),$$

$$P(Y_t = (x+1,-1) \mid Y_{t-1} = (x,i)) = p_t, \quad x = 0,1,2,\ldots,\mathbf{l}_n - 1, \quad i = k-1,$$

$$P(Y_t = (x,-1) \mid Y_{t-1} = (x,-1)) = p_t, \quad x = 1,2,\ldots,\mathbf{l}_n.$$

Με χρήση των ανωτέρων σχέσεων έχουμε:

$$\mathbf{A}_t(x) = \begin{bmatrix} & (x,0) & (x,1) & \cdots & (x,k-1) & (x,-1) \\ (x,0) & q_t & p_t & \cdots & 0 & 0 \\ (x,1) & q_t & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & 0 \\ (x,k-1) & q_t & 0 & \cdots & 0 & 0 \\ (x,-1) & q_t & 0 & \cdots & 0 & p_t \end{bmatrix}_{(k+1)\times(k+1)},$$

$$\mathbf{B}_t(x) = \begin{bmatrix} & (x+1,0) & (x+1,1) & \cdots & (x+1,k-1) & (x+1,-1) \\ (x,0) & 0 & 0 & \cdots & 0 & 0 \\ (x,1) & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & & 0 \\ (x,k-1) & 0 & 0 & \cdots & 0 & p_t \\ (x,-1) & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}_{(k+1)\times(k+1)}.$$

Τα στοιχεία του πίνακα $\mathbf{A}_t(x)$ αποτελούν τις πιθανότητες μετάβασης ανάμεσα στις καταστάσεις $(c_{x,i}, \quad i = -1,0,1,2,\ldots,s-1)$, ενώ τα στοιχεία του πίνακα $\mathbf{B}_t(x)$ αποτελούν τις πιθανότητες μετάβασης ανάμεσα στις καταστάσεις $(C_x, \quad x \geq 0)$. Η συνάρτηση πιθανότητας μπορεί να υπολογιστεί με χρήση του Πορίσματος 3.3 που ακολουθεί. Στην περίπτωση όπου οι πίνακες $\mathbf{A}_t(x)$, $\mathbf{B}_t(x)$ είναι ανεξάρτητοι του x και του χρόνου t και

για $k=2$ έχουμε πίνακες της μορφής

$$\mathbf{A}_t=\begin{bmatrix} q_t & p_t & 0 \\ q_t & 0 & 0 \\ q_t & 0 & p_t \end{bmatrix} \text{ και } \mathbf{B}_t=\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & p_t \\ 0 & 0 & 0 \end{bmatrix}.$$

Με εφαρμογή των Θεωρημάτων 2.2, 2.3, και 2.4, προκύπτουν οι εκφράσεις του Πορίσματος 3.3.

Πόρισμα 3.3: Έστω $G_{n,k}$ *ο αριθμός των ροών επιτυχιών μήκους τουλάχιστον* k *σε μια ακολουθία ανεξάρτητων και ισόνομων δοκιμών* Bernoulli. *Τότε,*

- Αν

$$\Phi_n(z,w)=\sum_{n=0}^{\infty}\sum_{x=0}^{\mathbf{l}_n}P(G_{n,k}=x)z^x w^n$$

είναι η διπλή γεννήτρια συνάρτηση, ισχύει ότι

$$\Phi_n(z,w)=\frac{1-(pw)^k(1-z)}{1-w+(1-z)qp^k w^{k+1}}.$$

- Αν

$$M_n(w)=\sum_{n=0}^{\infty}m_n w^n$$

είναι η γεννήτρια συνάρτηση των μέσων τιμών $m_n=E(G_{n,k})$, *ισχύει ότι*

$$M_n(w)=\frac{(pw)^k(1-pw)}{(1-w)^2},$$

καθώς και το ακόλουθο αναδρομικό σχήμα

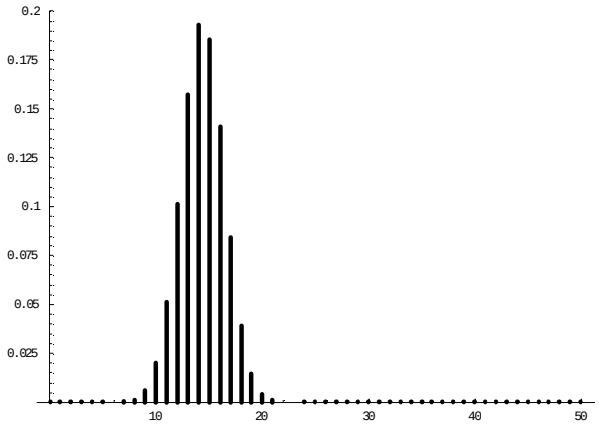
$$m_n=\begin{cases} 0 & n<k \\ p^k & n=k \\ p^k(1+q) & n=k+1 \\ 2m_{n-1}-m_{n-2} & n>k+1 \end{cases}$$

και

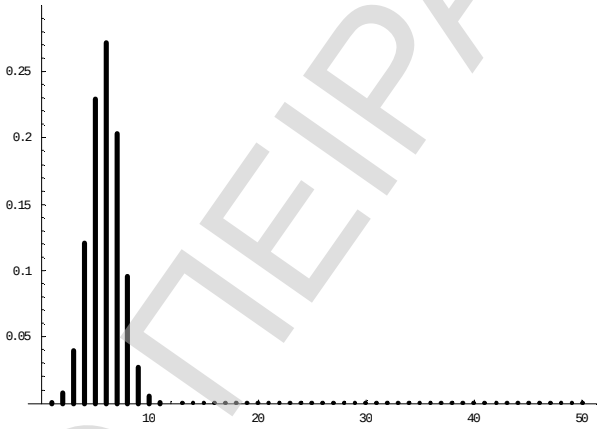
$$m_n=E(G_{n,k})=p^k((n-k)q+1)$$

Στο Σχήμα 3.3 δίνεται η συνάρτηση πιθανότητας f της $G_{n,k}$ για διάφορες τιμές των n, k, p. Οι υπολογισμοί έγιναν με χρήση του Πορίσματος 3.3.

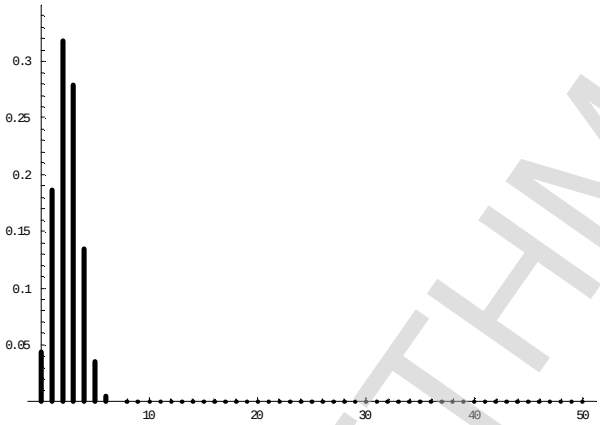
Σχήμα 3.3: Η συνάρτηση πιθανότητας της $G_{n,k}$ για διάφορες τιμές των n, k, p



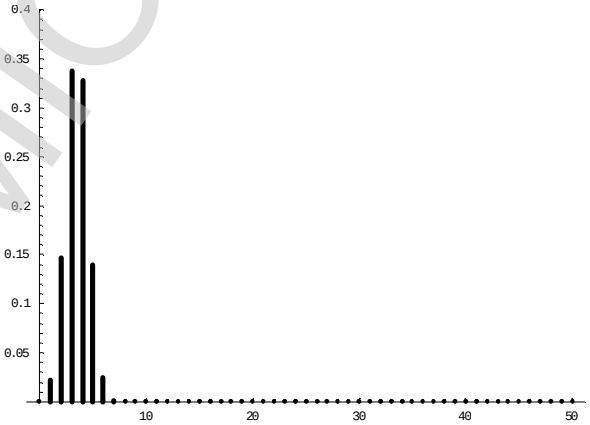
$n = 100,\ p = 0.75, k = 2.$



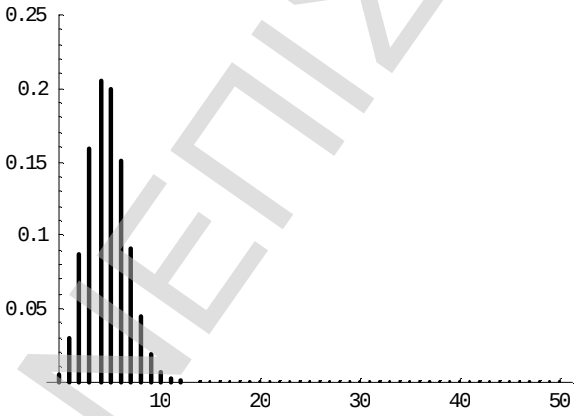
$n = 100,\ p = 0.75, k = 5.$



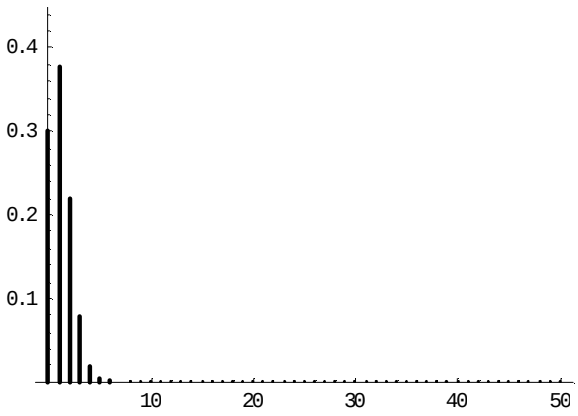
$n = 100,\ p = 0.75, k = 8$



$n = 100,\ p = 0.95, k = 9$



$n = 100,\ p = 0.25, k = 2$



$n = 100,\ p = 0.25, k = 3$

Στο σημείο αυτό θα πρέπει να σημειώσουμε ότι η προσέγγιση της Μαρκοβιανής εμφύτευσης όπως εισήχθη από τους Koutras and Alexandrou (1995) δεν καλύπτει την μελέτη του αριθμού $E_{n,k}$ των ροών επιτυχίας μήκους ακριβώς k. Για να ξεπεράσουν το εμπόδιο αυτό, οι Han and Aki (1999) εισήγαγαν την κλάση των τυχαίων μεταβλητών εμφυτεύσιμων σε Μαρκοβιανή αλυσίδα ανανεωτικού τύπου (returnable type) αντικαθιστώντας τη 2^{η} συνθήκη του Ορισμού 2.2 με τη συνθήκη

$$\Pr\left(Y_t \in C_u \mid Y_{t-1} \in C_x\right) = 0 \text{ για όλα τα } u \neq x-1, x, x+1.$$

Υπό αυτή τη συνθήκη, πέρα από τον πίνακα πιθανοτήτων εσωτερικών μεταπηδήσεων $\mathbf{A}_t(x)$, εμφανίζονται και δυο πίνακες πιθανοτήτων εξωτερικών μεταπηδήσεων οι οποίοι ελέγχουν τις προς τα άνω και κάτω μεταπηδήσεις, δηλαδή ο $\mathbf{B}(x) = (\Pr(Y_t = C_{x+1,j} \mid Y_{t-1} = C_{x,i}))_{s\times s}$ και ο $\mathbf{C}_t(x) = (\Pr(Y_t = C_{x-1,j} \mid Y_{t-1} = C_{x,i}))_{s\times s}$ αντίστοιχα.

Οι αναδρομικές σχέσεις του Θεωρήματος 2.2 γίνονται τώρα

$$\mathbf{f}_t(x) = \mathbf{f}_{t-1}(x)\mathbf{A}_t(x) + \mathbf{f}_{t-1}(x-1)\mathbf{B}_t(x-1) + \mathbf{f}_{t-1}(x+1)\mathbf{C}_t(x+1)$$

και στην ειδική περίπτωση, όπου $\mathbf{A}_t(x) = \mathbf{A}$, $\mathbf{B}_t(x) = \mathbf{B}$ και $\mathbf{C}_t(x) = \mathbf{C}$ για όλα τα x και t, η πιθανογεννήτρια συνάρτηση

$$f_n(z) = \sum_{x=0}^{l_n} \Pr(X_n = x) z^x$$

παίρνει τη μορφή

$$f_n(z) = \boldsymbol{\pi}_0 (\mathbf{A} + z\mathbf{B} + z^{-1}\mathbf{C})^n \mathbf{1}'.$$

Για περισσότερες λεπτομέρειες για την κλάση αυτή των τυχαίων μεταβλητών εμφυτεύσιμων σε Μαρκοβιανή αλυσίδα, ο ενδιαφερόμενος παραπέμπεται στους Han and Aki (1999).

Εδώ, μπορούμε να αναφέρουμε ότι η μεθοδολογία που αναπτύχθηκε για τις MVB μπορεί επίσης να χρησιμοποιηθεί αποτελεσματικά για τη μελέτη της ακριβούς κατανομής τυχαίων μεταβλητών που απαριθμούν σχηματισμούς σε ακολουθίες πολύτιμων δοκιμών.

Ο Fu (1996) γενίκευσε την μεθοδολογία της εμφύτευσης για τον υπολογισμό τόσο της κατανομής απαριθμητριών τυχαίων μεταβλητών όσο και για την κατανομή του χρόνου αναμονής σύνθετων σχηματισμών (patterns).

Αν ο πίνακας των πιθανοτήτων μεταπήδησης $\mathbf{\Lambda}$ που σχετίζεται με την διαδικασία εμφύτευσης αλυσίδας Markov για τη μεταβλητή X_n είναι διδιαγώνιος, τότε τα αντίστοιχα διανύσματα πιθανοτήτων θα ικανοποιούν μια τριγωνική επαναληπτική σχέση και συνεπώς η X_n θα είναι MVB.

Για παράδειγμα, η μελέτη του σχηματισμού $e = 2112$ οδηγεί σε έναν πίνακα πιθανοτήτων μεταπήδησης της μορφής

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{A} & \mathbf{B} & & \\ & \mathbf{A} & \mathbf{B} & \\ & & \mathbf{A} & \mathbf{B} \\ & & & \mathbf{O} \end{bmatrix}.$$

Στη συνέχεια, εισάγοντας την ακολουθία των διανυσμάτων

$$\mathbf{f}_t(x) = (\Pr(Y_t = (x,0)), \Pr(Y_t = (x,1)), \Pr(Y_t = (x,2)), \Pr(Y_t = (x,3)))$$

μπορούμε να γράψουμε το ακόλουθο σύνολο επαναληπτικών σχέσεων

$$\mathbf{f}_t(0) = \mathbf{f}_{t-1}(0)\mathbf{A},$$

$$\mathbf{f}_t(x) = \mathbf{f}_{t-1}(x)\mathbf{A} + \mathbf{f}_{t-1}(x-1)\mathbf{B}, \quad 1 \le x \le \lceil n/4 \rceil,$$

$$t = 1,2,\mathbf{K},n.$$

Η ακριβής κατανομή του αριθμού X_n των εμφανίσεων του $e = 2112$ σε n δοκιμές μπορεί να εκφραστεί ως

$$\Pr(X_n = x) = \mathbf{f}_n(x)\mathbf{1}', \; x = 0,1,\mathbf{K},\lceil n/4 \rceil.$$

Επιπλέον, η διπλή γεννήτρια συνάρτηση

$$\Phi(z,w) = \sum_{n=0}^{\infty} \sum_{x=0}^{\lceil n/4 \rceil} \Pr(X_n = x) z^x w^n$$

μπορεί να γραφεί ως

$$\Phi(z,w) = \mathbf{e}_1 \left[\mathbf{I} - w(\mathbf{A} + z\mathbf{B})\right]^{-1} \mathbf{1}' = \frac{1 + p_1^2 p_2 w^3}{1 - w + p_1^2 p_2 w^3 - p_1^2 p_2 w^4 (1 - p_2 + p_2 z)}.$$

Περισσότερές λεπτομέρειες για την μελέτη στατιστικών σχετιζόμενων με σχηματισμούς ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στον Fu (1996).

3.3. Μελέτης της Κατανομής του Αθροίσματος των μηκών των Ροών μήκους τουλάχιστον *k*

Στην παρούσα παράγραφο μελετάμε την κατανομή της τυχαίας μεταβλητής $S_{n,k}$ που καταγράφει το άθροισμα των μηκών των ροών επιτυχιών μήκους τουλάχιστον k σε μια ακολουθία n δίτιμων δοκιμών. Η μελέτη της $S_{n,k}$ επιτυγχάνεται με τη μέθοδο της εμφύτευσης τυχαίων μεταβλητών πολυωνυμικού τύπου σε Μαρκοβιανή αλυσίδα (MVP). Έτσι, αρχικά στην Εφαρμογή 3.4 δίνεται η διαδικασία εμφύτευσης της τυχαίας μεταβλητής πολυωνυμικού τύπου $S_{n,k}$ ενώ στη συνέχεια αποδεικνύονται μια σειρά από αποτελέσματα για την συνάρτηση πιθανότητας, τη διπλή και τη μονή γεννήτρια, καθώς και για την γεννήτρια μέσων τιμών της $S_{n,k}$. Ιδιαίτερη χρησιμότητα παρουσιάζει ο αναδρομικός τύπος υπολογισμού της συνάρτησης πιθανότητάς της.

Εφαρμογή 3.4: Τεχνική Εμφύτευσης της τυχαίας μεταβλητής $S_{n,k}$

Υποθέτουμε ότι $Z_1, Z_2, ..., Z_n$ είναι μια ακολουθία ανεξάρτητων δοκιμών Bernoulli με δυνατά αποτελέσματα επιτυχία (1) και αποτυχία (0) και αντίστοιχες πιθανότητες $p_t = P(Z_t = 1)$ και $q_t = 1 - p_t$ για $t = 1,2,...,n$ (με $Z_0 = 0$). Θεωρούμε την τυχαία μεταβλητή,

$$U_t = \begin{cases} k+\mathbf{1}, & \text{αν } Z_{t-k-\mathbf{1}+1} = Z_{t-k-\mathbf{1}+2} = ... = Z_t = 1, Z_{t-k-\mathbf{1}} = Z_{t+1} = 0 \\ 0, & \text{διαφορετικά.} \end{cases}$$

Το άθροισμα των μηκών των ροών μήκους τουλάχιστον k μπορεί εκφραστεί ως

$$S_{n,k} = \sum_{t=k}^{n} U_t \,,\ n \geq 1.$$

Θα εφαρμόσουμε τη μέθοδο της Μαρκοβιανής εμφύτευσης προκειμένου να υπολογίσουμε την κατανομή της τυχαίας μεταβλητής $S_{n,k}$ η οποία καταγράφει το άθροισμα των μηκών των ροών επιτυχιών μήκους τουλάχιστον k.

Βήμα 1: Θεωρούμε τον χώρο καταστάσεων

$$\Omega = \{(x,i) : x = 0,1,2,...,\mathbf{l}_n, i = 0,1,2,...,k\} \text{ με } \mathbf{l}_n = n.$$

Βήμα 2: Θεωρούμε την Μαρκοβιανή αλυσίδα $\{Y_t : t \geq 0\}$ ορισμένη στον χώρο Ω.

Ορίζουμε:

- $Y_t = (x, i)$, εάν στο χρόνο $t \geq 0$ το άθροισμα των μηκών των ροών μήκους τουλάχιστον k είναι ίσο με x και i επιτυχίες από την τελευταία εμφάνιση αποτυχίας, με $i = 1,2,...,k-1$.
- $Y_t = (x, k)$, εάν στον χρόνο $t \geq 0$ το άθροισμα των μηκών των ροών μήκους τουλάχιστον k είναι ίσο με x και i επιτυχίες από την τελευταία εμφάνιση αποτυχίας, με $i = k,...,n$.

Βήμα 3: Για κάποιο $0 \leq x \leq \mathbf{l}_n$ ορίζουμε $C_x = \{(x,i) : i = 0,1,...,k-1\}$. Το σύνολο των C_x, $0 \leq x \leq l_n$ για κάθε x αποτελεί μια διαμέριση του χώρου $\Omega = \{(x,i) : x = 0,1,2,..., \mathbf{l}_n, i = 0,1,2,...,k\}$. Είναι προφανές ότι σε κάθε C_x ο πίνακας μετάβασης μπορεί να οδηγήσει σε ένα από τα C_x, C_{x+1}, C_{x+k}.

Βήμα 4: Προφανώς για κάθε $x = \{0,1,2,....,\mathbf{l}_n\}$ ισχύει $P(X_n = x) = P(Y_n \in C_x)$.

Βήμα 5: Επίσης, ισχύει $P(Y_t = c_{y,j} \mid Y_{t-1} = c_{x,i}) = 0$ για κάθε $y \neq x, x+1, ..., x+k,\ t \geq 1$.

Από τα βήματα που προηγήθηκαν προκύπτει εύκολα ότι ικανοποιείται ο Ορισμός 2.3 και συνεπώς η τυχαία μεταβλητή $S_{n,k}$ αποτελεί μια τυχαία μεταβλητή πολυωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα.

Συνεπώς, το επόμενο βήμα της μεθόδου είναι να ορίσουμε τις αρχικές πιθανότητες της αλυσίδας ίσες με $\boldsymbol{\pi}_0 = (1,0,0,...,0)$ καθώς και τους πίνακες μετάβασης $\mathbf{A}_{t,i}(x), \quad i = 0,1,k$.

Η γενική μορφή των πινάκων $\mathbf{A}_{t,0}(x), \mathbf{A}_{t,1}(x), \mathbf{A}_{t,k}(x)$ έχει ως εξής:

$$\mathbf{A}_{t,0}(x) = \begin{bmatrix} & (x,0) & (x,1) & (x,2) & (x,3) & \text{M} & (x,k-1) & (x,k) \\ (x,0) & q_t & p_t & & & \text{M} & & \\ (x,1) & q_t & & p_t & & \text{M} & & \\ (x,2) & q_t & & & p_t & \text{M} & & \\ (x,3) & q_t & & & & \text{M} & & \\ \text{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{O} & p_t & \mathbf{L} \\ (x,k-1) & q_t & & & & & & \\ (x,k) & q_t & & & & \text{M} & & \end{bmatrix}_{(k+1)\times(k+1)}$$

$$\mathbf{A}_{t,1}(x)=\begin{bmatrix} & (x,0) & (x,1) & (x,2) & (x,3) & \text{M} & (x,k-1) & (x,k) \\ (x,0) & & & & & \text{M} & & \\ (x,1) & & & & & \text{M} & & \\ (x,2) & & & & & \text{M} & & \\ (x,3) & & & & & \text{M} & & \\ \text{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{O} & \mathbf{L} & \mathbf{L} \\ (x,k-1) & & & & & & & \\ (x,k) & & & & & \text{M} & & p_t \end{bmatrix}_{(k+1)\times(k+1)}$$

$$\mathbf{A}_{t,k}(x)=\begin{bmatrix} & (x,0) & (x,1) & ((x,2)) & (x,3) & \text{M} & (x,k-1) & (x,k) \\ (x,0) & & & & & \text{M} & & \\ (x,1) & & & & & \text{M} & & \\ (x,2) & & & & & \text{M} & & \\ (x,3) & & & & & \text{M} & & \\ \text{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{O} & \mathbf{L} & \mathbf{L} \\ (x,k-1) & & & & & & & p_t \\ (x,k) & & & & & \text{M} & & \end{bmatrix}_{(k+1)\times(k+1).} .$$

Τα στοιχεία του πίνακα $\mathbf{A}_{t,0}(x)$ αποτελούν τις πιθανότητες μετάβασης ανάμεσα στις καταστάσεις $(c_{x,i},\ \ i=0,1,2,...,k-1)$ ενώ τα στοιχεία των πινάκων $\mathbf{A}_{t,1}(x)$, $\mathbf{A}_{t,k}(x)$ δίνουν τις πιθανότητες μετάβασης ανάμεσα στα σύνολα καταστάσεων C_x, C_{x+1}, C_{x+k}.

Τέλος, μπορούμε να υπολογίσουμε την κατανομή της τυχαίας μεταβλητής $S_{n,k}$, με εφαρμογή του Θεωρήματος 2.6 και χρησιμοποιώντας τους πίνακες μετάβασης διαστάσεων $s\times s$, $\mathbf{A}_{t,0}(x),\mathbf{A}_{t,1}(x),\mathbf{A}_{t,k}(x)$ και τα διανύσματα πιθανότητας διαστάσεων $1\times s$ $\mathbf{f}_t(x)$.

Για παράδειγμα, εάν η τυχαία μεταβλητή μεταβλητή $S_{n,k}$, συμβολίζει / καταγράφει το άθροισμα των ροών μήκους τουλάχιστον $k=2$ σε μία ακολουθία δοκιμών μήκους $n=3$, με πιθανότητα επιτυχίας p_t, τότε ο πίνακας μετάβασης $\mathbf{\Lambda}_t$ και οι πίνακες $\mathbf{A}_{t,i}(x),\quad i=0,1,2$ του Ορισμού 2.3 είναι οι εξής:

$$\mathbf{\Lambda}_t = \begin{bmatrix} q_t & p_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & \\ q_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_t & & & \\ q_t & 0 & 0 & 0 & 0 & p_t & 0 & 0 & 0 & & & \\ & & & q_t & p_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & q_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_t \\ & & & q_t & 0 & 0 & 0 & 0 & p_t & 0 & 0 & 0 \\ & & & & & & q_t & p_t & 0 & 0 & 0 & 0 \\ & & & & & & q_t & 0 & 0 & 0 & 0 & 0 \\ & & & & & & q_t & 0 & 0 & 0 & 0 & p_t \\ & & & & & & & & & q_t & p_t & 0 \\ & & & & & & & & & q_t & 0 & 0 \\ & & & & & & & & & q_t & 0 & 1 \end{bmatrix},$$

και

$$\mathbf{A}_{t,0} = \begin{bmatrix} q_t & p_t & 0 \\ q_t & 0 & 0 \\ q_t & 0 & 0 \end{bmatrix}_{(k+1)\times(k+1)}, \quad \mathbf{A}_{t,1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & p_t \end{bmatrix}_{(k+1)\times(k+1)}, \quad \mathbf{A}_{t,2} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & p_t \\ 0 & 0 & 0 \end{bmatrix}_{(k+1)\times(k+1)}$$

Συνεπώς, με εφαρμογή του Θεωρήματος 2.6 και χρήση των πινάκων μετάβασης $\mathbf{A}_{t,0}, \mathbf{A}_{t,1}, \mathbf{A}_{t,k}$ και των διανυσμάτων πιθανότητας διαστάσεων 1×3 $\mathbf{f}_t(x)$, είναι δυνατόν να υπολογίσουμε την κατανομή της τυχαίας μεταβλητής $S_{n,3}$.

Στα επόμενα θεωρήματα δίνουμε μια σειρά από αποτελέσματα τα οποία αφορούν την τυχαία μεταβλητή $S_{n,k}$.

Συγκεκριμένα, στα Θεωρήματα 3.1 έως 3.6, αποδεικνύονται με χρήση των αποτελεσμάτων του Κεφαλαίου 2, μια σειρά αποτελεσμάτων για τη μονή και τη διπλή γεννήτρια, τη συνάρτηση πιθανότητας, καθώς και για τις μέσες τιμές (απλές και παραγοντικές).

Θεώρημα 3.1: *Έστω $S_{n,k}$ το άθροισμα των μηκών των ροών επιτυχιών μήκους τουλάχιστον k σε μια n ακολουθία ανεξάρτητων και ισόνομων δοκιμών* Bernoulli. *Τότε, αν*

$$\Phi_n(z,w) = \sum_{n=0}^{\infty} \sum_{x=0}^{\mathbf{l}_n} P(S_{n,k} = x) z^x w^n$$

είναι η διπλή γεννήτρια συνάρτηση της $S_{n,k}$, *ισχύει ότι*

$$\Phi_n(z,w)=\frac{P_1(z,w)}{P_2(z,w)},$$

όπου

$$P_1(z,w)=1-wpz-(wp)^k(1-z^k)-(wp)^{k+1}(z^k-z)$$

και

$$P_2(z,w)=1-w(1+pz)+w^2pz+w^{k+1}qp^k(1-z^k)+w^{k+2}qp^{k+1}(z^k-z).$$

Απόδειξη: Η απόδειξη προκύπτει άμεσα με εφαρμογή του Θεωρήματος 2.7. Συγκεκριμένα, από το Θεώρημα 2.7 γνωρίζουμε ότι

$$\mathbf{\Phi}(z,w)=\boldsymbol{\pi}_0\left[\mathbf{I}-w\left(\sum_{i=0}^{m}z^i\mathbf{A}_i\right)\right]^{-1}.$$

Στη συγκεκριμένη περίπτωση έχουμε,

$$\mathbf{\Phi}(z,w)=\boldsymbol{\pi}_0[\mathbf{I}-w[\mathbf{A}_{t,0}(x)+z\mathbf{A}_{t,1}(x)+z^k\mathbf{A}_{t,k}(x)]]^{-1}.$$

Αντιστρέφοντας τον πίνακα

$$[\mathbf{I}-w[\mathbf{A}_{t,0}(x)+z\mathbf{A}_{t,1}(x)+z^k\mathbf{A}_{t,k}(x)]],$$

εφαρμόζοντας απλές πράξεις πινάκων, και κάνοντας μια σειρά από σύνθετες αλγεβρικές πράξεις έχουμε το ζητούμενο αποτέλεσμα. □

Θεώρημα 3.2: *Έστω* $S_{n,k}$ *το άθροισμα των μηκών των ροών επιτυχιών μήκους τουλάχιστον* k *σε μια ακολουθία* n *ανεξάρτητων και ισόνομων δοκιμών* Bernoulli. *Τότε, για την πιθανογεννήτρια συνάρτηση* $f_n(z)=\sum_{x=0}^{\mathbf{l}_n}P(S_{n,k}=x)z^x$ *της* $S_{n,k}$, *ισχύει ότι*

$$f_n(z)=\begin{cases}1, & 0<k<n\\ 1-p^k+(pz)^k, & n=k\\ 1-p^k(1+q)+2qp^kz^k+(pz)^{k+1}, & n=k+1\\ (1-pz)f_{n-1}(z)-pzf_{n-1}(z)-qp^k(1-z^k)\times & \\ \times f_{n-k-1}(z)-qp^{k+1}(z^k-z)f_{n-k-2}(z), & n\geq k+1.\end{cases}$$

Απόδειξη: Το θεώρημα αποδεικνύεται γράφοντας την ισότητα

$$\sum_{n=0}^{\infty}\sum_{x=0}^{\mathbf{l}_n}P(S_{n,k}=x)z^xw^n=\frac{P_1(z,w)}{P_2(z,w)}$$

στη μορφή

$$P_2(z,w)\sum_{n=0}^{\infty}\sum_{x=0}^{\mathbf{l}_n}P(S_{n,k}=x)z^x w^n = P_1(z,w)$$

και εξισώνοντας τις δυνάμεις των συντελεστών του w^n για $n=0,1,2,3,....$ στα δυο μέλη της παραπάνω ισότητας. □

Θεώρημα 3.3: *Έστω* $S_{n,k}$ *το άθροισμα των μηκών των ροών επιτυχιών μήκους τουλάχιστον* k *σε μια ακολουθία* n *ανεξάρτητων και ισόνομων δοκιμών Bernoulli. Τότε, αν* $g_n(x)=P(S_{n,k}=x)$ *είναι η συνάρτηση πιθανότητας της* $S_{n,k}$, *ισχύει ότι*

$$g_n(x)=g_{n-1}(x)+pg_{n-1}(x-1)-pg_{n-2}(x-1)-qp^k(g_{n-k-1}(x)-g_{n-k-1}(x-k))$$
$$-qp^{k+1}(g_{n-k-2}(x-k)-g_{n-k-2}(x-1))$$

για $n\geq k+2, x\geq 0$, *με αρχικές συνθήκες*

$$g_n(x)=0, \text{ για } x<0 \text{ ή } x>n,$$

$$g_n(x)=\begin{cases}1, & x=0\\ 0, & x>0\end{cases} \text{ και } 0\leq n<k,$$

$$g_n(x)=\begin{cases}1-p^k, & x=0\\ p^k, & x=k\\ 0, & 1\leq x\leq k-1\end{cases} \text{ και } n=k,$$

$$g_n(x)=\begin{cases}1-p^k(1+q), & x=0\\ 2qp^k, & x=k\\ p^{k+1}, & x=k+1\\ 0, & 1\leq x\leq k+1\end{cases} \text{ και } n=k+1.$$

Απόδειξη: Το θεώρημα αποδεικνύεται αντικαθιστώντας το $f_n(z)$ με το

$$\sum_{x=0}^{\mathbf{l}_n}P(S_{n,k}=x)z^x$$

στο αναδρομικό σχήμα του Θεωρήματος 3.2. Ύστερα από αλγεβρικές πράξεις και εξισώνοντας τις δυνάμεις των συντελεστών του z^x για $x=1,2,3,....$ στην δυναμοσειρά που προκύπτει, έχουμε το παραπάνω αποτέλεσμα. □

Θεώρημα 3.4: *Για τις ροπές* $m_{n,r}, r \geq 1$ *της τυχαίας μεταβλητής* $S_{n,k}$, *ισχύει το ακόλουθο αναδρομικό σχήμα,*

$$m_{n,r} = m_{n-1,r} + p\sum_{i=0}^{r}\binom{r}{i}(m_{n-1,i} - m_{n-2,i}) - qp^k m_{n-k-1,r}$$
$$+ qp^k \sum_{i=0}^{r}\binom{r}{i}\left[k^{r-i}(m_{n-k-1,i} - m_{n-k-2,i}) + pm_{n-k-2,r}\right]$$

για $n \geq k+2$ *και αρχικές συνθήκες τις*

$$m_{n,r} = \begin{cases} 0, & 0 \leq n \leq k \\ k^r p^k, & n = k \\ 2k^r qp^k + (k+1)^r p^{k+1}, & n = k+1. \end{cases}$$

Απόδειξη: Το θεώρημα αποδεικνύεται αντικαθιστώντας το z με e^z στο αναδρομικό σχήμα του Θεωρήματος 3.2 και γνωρίζοντας ότι

$$m_{n,r} = \begin{cases} \dfrac{d^r}{dz^r} E(e^{zX})\Big|_{z=0}, & r \geq 1 \\ 1 & r = 0 \end{cases}$$

καθώς και ότι $\dfrac{d^r}{dz^r} E(e^{zX})\Big|_{z=0} = \sum_{i=0}^{r}\binom{r}{i}k^{r-i}m_{n,i}$ η απόδειξη ολοκληρώνεται. □

Θεώρημα 3.5: *Η γεννήτρια των μέσων τιμών της τυχαίας μεταβλητής* $S_{n,k}$ *δίνεται από την έκφραση,*

$$M(w) = \sum_{n=0}^{\infty} E(S_{n,k})w^n = \frac{(wp)^k(k - wp(k-1))}{(1-w)^2}.$$

Απόδειξη: Η απόδειξη προκύπτει με εφαρμογή του Θεωρήματος 2.9. □

Θεώρημα 3.6: *Η μέση τιμή της τυχαίας μεταβλητής* $S_{n,k}$ *δίνεται από την έκφραση,*

$$E(S_{n,k}) = p^k(k + (n-k)(kq+p)), \quad n \geq k.$$

Απόδειξη: Η απόδειξη προκύπτει γράφοντας την σχέση

$$\sum_{n=0}^{\infty} E(S_{n,k})w^n = \frac{(wp)^k(k - wp(k-1))}{(1-w)^2}$$

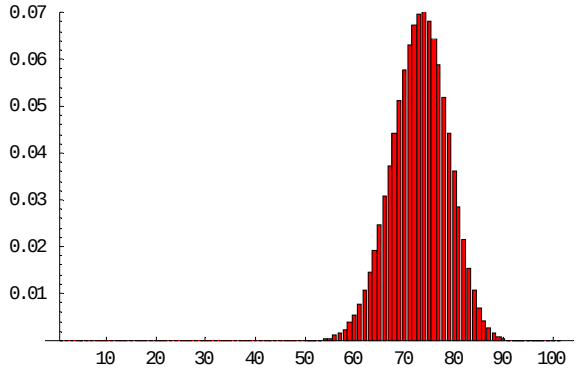
στην μορφή

$$(1-w)^2 \sum_{n=0}^{\infty} E(S_{n,k}) w^n = (wp)^k (k - wp(k-1))$$

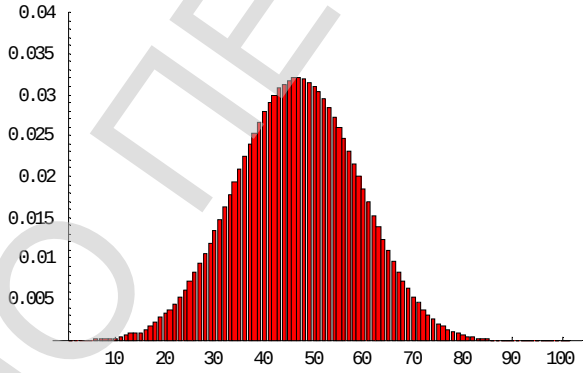
και εξισώνοντας τις δυνάμεις των συντελεστών του w^n για $n = 1,2,3,....$ στην δυναμοσειρά που προκύπτει, έχουμε το παραπάνω αποτέλεσμα. □

Στο Σχήμα 3.4 δίνεται η κατανομή της $S_{n,k}$ για διάφορες τιμές των n, k, p.

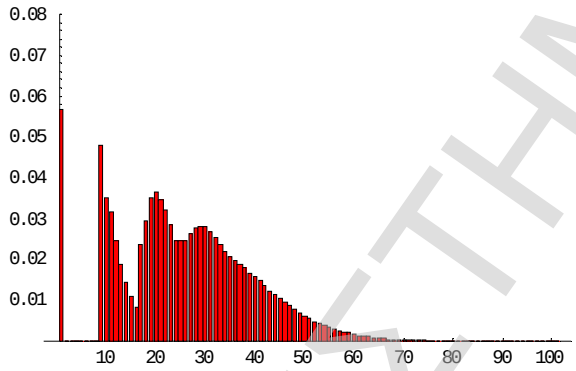
Σχήμα 3.4: Η συνάρτηση πιθανότητας $g(x) = P(S_{n,k} = x)$ για διάφορες τιμές των n, k, p



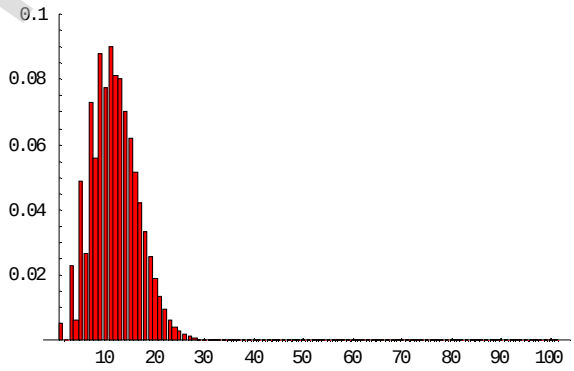
$n = 100, p = 0.75, k = 2.$



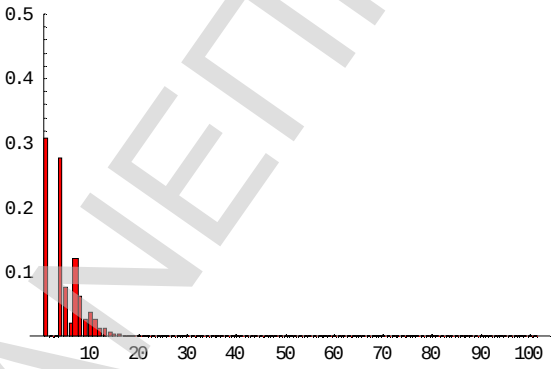
$n = 100, p = 0.75, k = 5.$



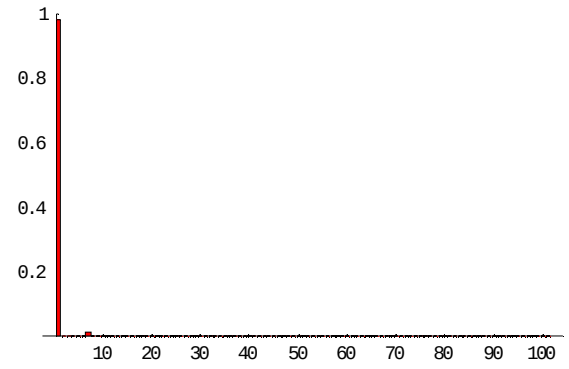
$n = 100, p = 0.75, k = 8$



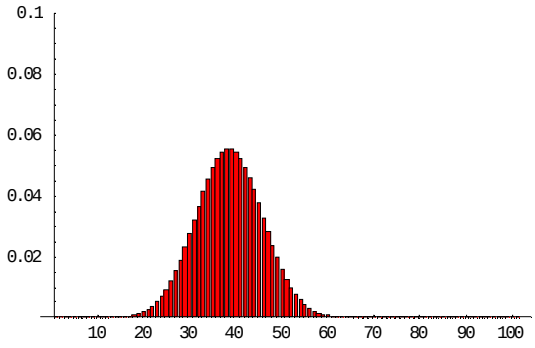
$n = 100, p = 0.25, k = 2$



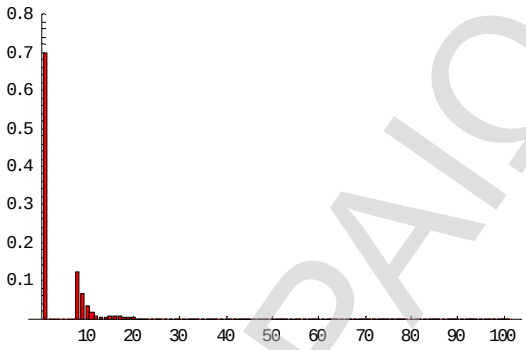
$n = 100, p = 0.25, k = 3$



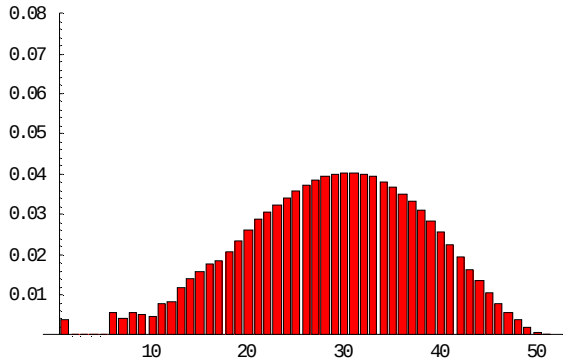
$n = 100, p = 0.25, k = 6$



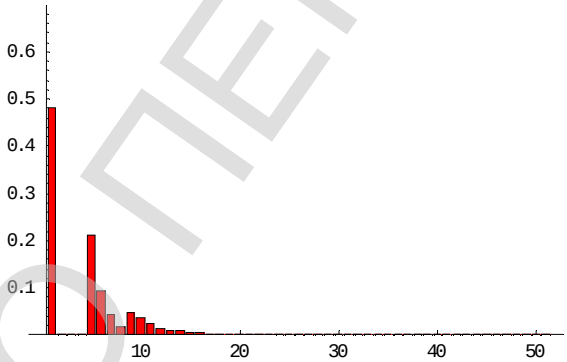
$n = 100, p = 0.5, k = 2$



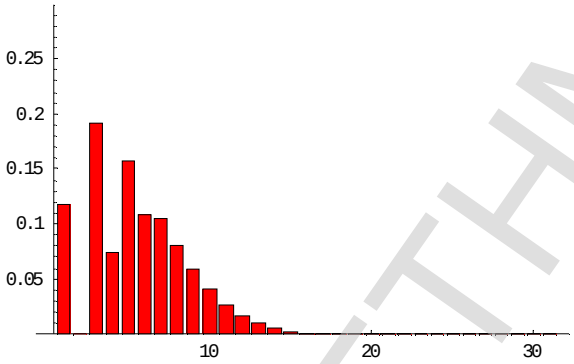
$n = 100, p = 0.5, k = 7$



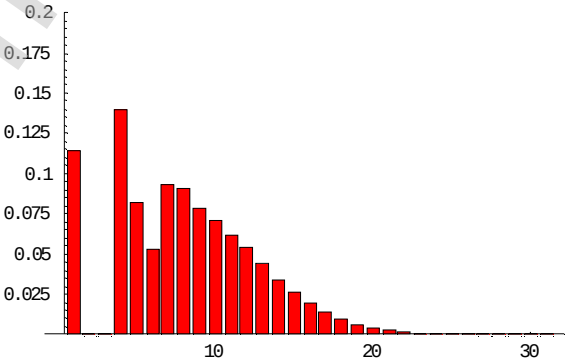
$n = 50, p = 0.8, k = 2$



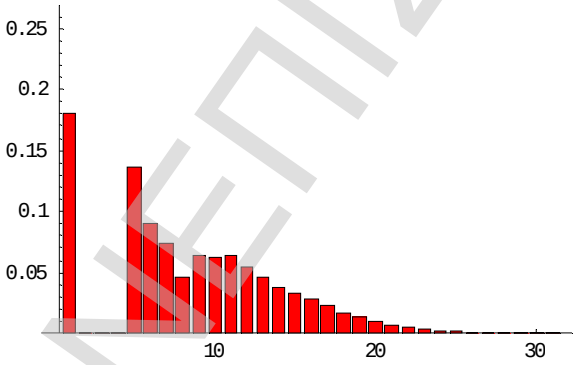
$n = 50, p = 0.4, k = 4$



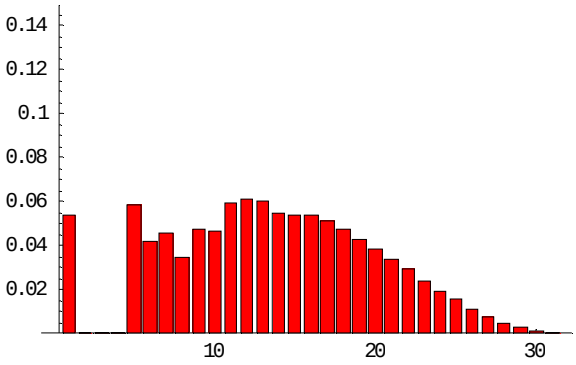
$n = 30, p = 0.3, k = 2$



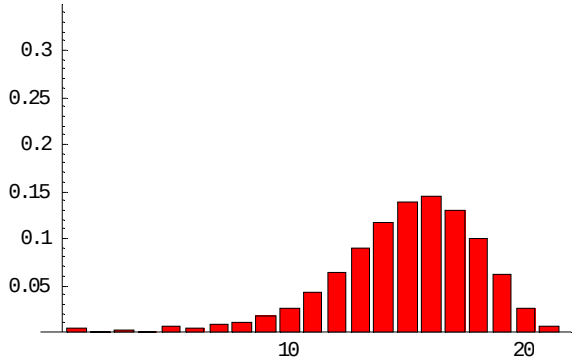
$n = 30, p = 0.5, k = 3$



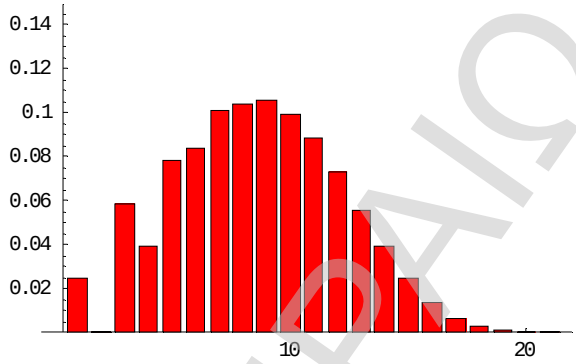
$n = 30, p = 0.6, k = 4$



$n = 30, p = 0.7, k = 4$



$n = 20, p = 0.7, k = 2$



$n = 20, p = 0.5, k = 2$

3.4. Μελέτη Δεσμευμένων Κατανομών Σχετικών με Ροές Επιτυχιών σε Ακολουθίες Ανεξάρτητων και Ισόνομων Δοκιμών Bernoulli

Ο Schuster (1994) ανέπτυξε μια μέθοδο με την οποία απέδειξε απλές σχέσεις για τη δεσμευμένη κατανομή του μήκους της μεγαλύτερης ροής δύο ειδών σε μια τυχαία διάταξη από y αποτυχίες και $n - y$ επιτυχίες. Η βασική ιδέα της μεθόδου είναι η μελέτη του δεσμευμένου μήκους της μεγαλύτερης ροής δεδομένου του αποτελέσματος της τελευταίας δοκιμής της ακολουθίας και η σύνδεση του με τη μη δεσμευμένη κατανομή με τη βοήθεια του θεωρήματος ολικής πιθανότητας. Έτσι στην εργασία αυτή, προτείνονται αναδρομικοί τύποι για τη δεσμευμένη και μη δεσμευμένη κατανομή που μπορούν εύκολα να εφαρμοσθούν.

Στην συνέχεια, οι Koutras and Alexandrou (1997) εφάρμοσαν την τεχνική της Μαρκοβιανής εμφύτευσης προκειμένου να μελετήσουν τις δεσμευμένες κατανομές των πλέον βασικών ειδών ροών. Συγκεκριμένα, μελέτησαν τις περιπτώσεις των δεσμευμένων κατανομών (όταν είναι γνωστός ο συνολικός αριθμός επιτυχιών) των μη επικαλυπτόμενων, των επικαλυπτόμενων ροών επιτυχιών μήκους k και των ροών επιτυχιών μήκους τουλάχιστον k.

Η ίδια μεθοδολογία χρησιμοποιείται και στη διατριβή αυτή προκειμένου να υπολογιστεί η δεσμευμένη κατανομή $S_{n,k}$ του αθροίσματος των ροών επιτυχιών μήκους τουλάχιστον k δοθέντος του αριθμού των επιτυχιών στις n δοκιμές.

Οι παραπάνω δεσμευμένες κατανομές κρίνονται κατάλληλες για χρήση ως ελεγχοσυναρτήσεις τυχαιότητας σε προβλήματα αντίστοιχα με αυτά που εξετάστηκαν από τους O'Brien and Dyck (1985), Agin and Godbole (1992), Koutras and Alexandrou (1997). Έτσι, στην παράγραφο αυτή παρουσιάζουμε αποτελέσματα των Koutras and Alexandrou (1995) ενώ στην επόμενη αναπτύσσουμε τα αντίστοιχα αποτελέσματα για την περίπτωση της τυχαίας μεταβλητής $S_{n,k}$ που μελετήθηκε στην προηγούμενη παράγραφο. Σημαντικό είναι να τονιστεί ότι στην παράγραφο αυτή υποτίθεται ότι η σύνθεση της παρατηρηθείσας ακολουθίας είναι γνωστή, δηλαδή, ο αριθμός επιτυχιών και οι αποτυχίες είναι σταθερές ποσότητες.

Συνεπώς, οι πιθανότητες που μελετάμε είναι δεσμευμένες στον συνολικό αριθμό των επιτυχιών. Υποθέτουμε δηλαδή, ότι η ακολουθία ανεξάρτητων και ισόνομων δοκιμών Bernoulli $Z_1, Z_2, ..., Z_n$ με δυνατά αποτελέσματα επιτυχία (1) και αποτυχία (0) και αντίστοιχες πιθανότητες p και $q = 1 - p$ ($\mu\varepsilon Z_0 = 0$) έχει πραγματοποιηθεί. Η πρόθεσή μας, αρχικά είναι να παρουσιάσουμε τη δεσμευμένη κατανομή της τυχαίας μεταβλητής $Y_{n,k}$ (μίας εκ των $N_{n,k}$, $M_{n,k}$, $G_{n,k}$) και στην επόμενη παράγραφο να μελετήσουμε την αντίστοιχη δεσμευμένη κατανομή της $S_{n,k}$, δεδομένου του αριθμού $n-y$ *των* επιτυχιών στις n δοκιμές. Τα Θεωρήματα 3.7 έως 3.9 οφείλονται στους Koutras and Alexandrou (1995). Συγκεκριμένα, το επόμενο θεώρημα αναφέρεται στην δεσμευμένη κατανομή της $N_{n,k}$, δεδομένου του αριθμού X_n των επιτυχιών σε n ανεξάρτητες και ισόνομες δίτιμες δοκιμές Bernoulli.

Θεώρημα 3.7: *Η δεσμευμένη πιθανότητα της τυχαίας μεταβλητής $N_{n,k}$ δοθέντος του αριθμού των επιτυχιών $X_n = n - y$ σε μια ακολουθία n ανεξάρτητων και ισόνομων δίτιμων δοκιμών Bernoulli δίνεται από τον τύπο*

$$P(N_{n,k} = x \mid X_n = n - y) = \frac{\binom{y+x}{x}}{\binom{n}{y}} \sum_{j=0}^{\left[\frac{n-y-kx}{k}\right]} (-1)^j \binom{y+1}{j} \binom{n-kx-kj}{y},$$

για $x \geq 0,\ n \geq kx + y.$

Στο επόμενο θεώρημα δίνεται η δεσμευμένη κατανομή της $M_{n,k}$, δεδομένου του αριθμού X_n των επιτυχιών σε n δίτιμες δοκιμές Bernoulli. Η συγκεκριμένη κατανομή αποδείχθηκε πολύ χρήσιμη στο πεδίο των ελέγχων τυχαιότητας. Συγκεκριμένα, οι Koutras and Alexandrou (1997) έδειξαν ότι η χρήση της $M_{n,k}$ οδηγεί στον έλεγχο τυχαιότητας με τη μεγαλύτερη ισχύ, ανάμεσα στους ελέγχους με χρήση μίας εκ των στατιστικών $N_{n,k}$, $M_{n,k}$, $G_{n,k}$.

Θεώρημα 3.8: *Η δεσμευμένη πιθανότητα της τυχαίας μεταβλητής $M_{n,k}$ δοθέντος του αριθμού των επιτυχιών $X_n = n - y$ σε μια ακολουθία n ανεξάρτητων και ισόνομων δίτιμων δοκιμών Bernoulli δίνεται από τον τύπο*

$$P(M_{n,k} = x \mid X_n = n - y) = \frac{1}{\binom{n}{y}} \left\{ \sum_{j=0}^{\min(x-1,y)} \sum_{i=0}^{\infty} A_n(x, y; i, j) + \binom{n-k(y+1)}{y} d_{n,x+(y+1)k-1},\ x \geq 1 \right\},$$

όπου $d_{i,j}$ είναι ο τελεστής δέλτα του Kronecker ο οποίος παίρνει την τιμή 1 ($d_{i,j} = 1$) αν και μόνο εάν $i = j$, και $A_n(x, y; i, j) = (-1)^i \binom{x-1}{j} \binom{y+1}{y-j} \binom{y-j}{i} \binom{n-k(i+j+1)-x}{y-j-1}$.

Το επόμενο θεώρημα αναφέρεται στη δεσμευμένη κατανομή της $G_{n,k}$, δεδομένου του αριθμού X_n των επιτυχιών σε n δίτιμες δοκιμές Bernoulli.

Θεώρημα 3.9: *Η δεσμευμένη πιθανότητα της τυχαίας μεταβλητής $G_{n,k}$ δοθέντος του αριθμού των επιτυχιών $X_n = n - y$ σε μια ακολουθία n ανεξάρτητων και ισόνομων δίτιμων δοκιμών Bernoulli δίνεται από τον τύπο*

$$P(G_{n,k} = x \mid X_n = n - y) = \frac{\binom{y+1}{x}}{\binom{n}{y}} \sum_{j=0}^{\left[\frac{n-y-kx}{k}\right]} (-1)^j \binom{y-x+1}{j} \binom{n-k(x+j)}{y},\ n \geq kx + y, x > 0.$$

3.5. Η Δεσμευμένη Κατανομή του Αθροίσματος των Μηκών των Ροών Επιτυχιών Μήκους Τουλάχιστον *k*

Το επόμενο θεώρημα αφορά τη διπλή γεννήτρια συνάρτηση των δεσμευμένων πιθανοτήτων.

Το αποτέλεσμα του θεωρήματος αυτού αποτελεί εργαλείο για την απόδειξη του Θεωρήματος 3.11 για την δεσμευμένη κατανομή της υπό μελέτης τυχαίας μεταβλητής $S_{n,k}$, δεδομένου του αριθμού $n-y$ των επιτυχιών στις n δοκιμές που ακολουθεί.

Στο εξής, με

$$f_n(z;p)=\sum_{x=0}^{\infty}P(S_{n,k}=x)z^x$$

συμβολίζουμε τη μονή γεννήτρια, με

$$\Phi(z,w;p)=\sum_{n=0}^{\infty}\sum_{x=0}^{\infty}P(S_{n,k}=x)z^x w^n$$

συμβολίζουμε τη διπλή γεννήτρια συνάρτηση αντίστοιχα.

Επίσης με

$$y_n(z;p)=\sum_{x=0}^{\infty}P(S_{n,k}=x\mid X_n=n-y)z^x$$

συμβολίζουμε τη γεννήτρια συνάρτηση των δεσμευμένων πιθανοτήτων δοθέντος του αριθμού των επιτυχιών σε μια ακολουθία n δίτιμων δοκιμών.

Έτσι, έχουμε το ακόλουθο θεώρημα για την ποσότητα $a_n(z;y)=\binom{n}{y}y_n(z;y)$, $y=0,1,2,\ldots,n$.

Θεώρημα 3.10: *Η διπλή γεννήτρια συνάρτηση των*

$$a_n(z;y)=\binom{n}{y}y_n(z;y),\quad y=0,1,2,\ldots,n$$

δίνεται από την έκφραση

$$\sum_{y=0}^{\infty}\left(\sum_{n=y}^{\infty}a_n(z;y)\right)t^y=\Phi\left(z,(1+t)w,\frac{1}{1+t}\right).$$

Απόδειξη: Αντικαθιστώντας στην $f_n(z;p)=\sum_{x=0}^{\infty}P(S_{n,k}=x)z^x$

την

$$P(S_{n,k}=x)=\sum_{y=0}^{n}P(S_{n,k}=x\mid X_n=n-y)P(X_n=n-y)$$

$$=\sum_{y=0}^{n}\binom{n}{y}p^n\left(\frac{q}{p}\right)^y P(S_{n,k}=x\mid X_n=n-y)$$

και χρησιμοποιώντας την έκφραση

$$y_n(z;p)=\sum_{x=0}^{\infty}P(S_{n,k}=x\mid X_n=n-y)z^x$$

έχουμε

$$f_n(z;p)=\sum_{y=0}^{n}\binom{n}{y}p^n\left(\frac{q}{p}\right)^y y_n(z;y)$$

ή ισοδύναμα την έκφραση

$$\Phi_n(z,w;p)=\sum_{n=0}^{\infty}\sum_{y=0}^{n}(pw)^n\left(\frac{q}{p}\right)^y a_n(z;y).$$

Θέτοντας $t=\frac{q}{p}$ και αντικαθιστώντας στην προηγούμενη έκφραση έχουμε

$$\Phi_n\left(z,w;\frac{1}{(1+t)}\right)=\sum_{n=0}^{\infty}\sum_{y=0}^{n}\left(\frac{w}{1+t}\right)^n t^y a_n(z;y).$$

Τέλος, θέτοντας όπου w το $(1+t)w$ στην

$$\Phi\left(z,(1+t)w,\frac{1}{1+t}\right)=\sum_{y=0}^{\infty}\left(\sum_{n=y}^{\infty}a_n(z;y)\right)t^y$$

καταλήγουμε στο ζητούμενο αποτέλεσμα □

Με το επόμενο θεώρημα, το οποίο αναφέρεται στη δεσμευμένη κατανομή της $S_{n,k}$, δεδομένου του αριθμού X_n των επιτυχιών σε n δίτιμες δοκιμές Bernoulli κλείνει η παράγραφος αυτή, η οποία και αφορά δεσμευμένες κατανομές τυχαίων μεταβλητών σχετιζόμενες με ροές επιτυχιών.

Θεώρημα 3.11: *Η δεσμευμένη πιθανότητα της τυχαίας μεταβλητής $S_{n,k}$ δοθέντος του αριθμού των επιτυχιών $X_n = n - y$ σε μια ακολουθία n ανεξάρτητων και ισόνομων δίτιμων δοκιμών Bernoulli δίνεται από τον τύπο*

$$P(S_{n,k} = x \mid X_n = n - y) = \binom{n}{y}^{-1} \sum_{r=0}^{y+1} \sum_{i=0}^{r} \sum_{j_1=0}^{r-i} \sum_{j_2=0}^{i} (-1)^d \binom{y+1}{r} \binom{r}{i} \binom{r-i}{j_1} \binom{i}{j_2} \binom{r+a-1}{a} \binom{y+b}{b}$$

όπου $a = x - i + j_2 - k(j_1 + j_2),\ b = n - y - kr - i - a,\ d = r + i + j_1 - j_2$.

Απόδειξη: Με χρήση του Θεωρήματος 3.10 έχουμε

$$\Phi\left(z, (1+t)w, \frac{1}{1+t}\right) = \sum_{y=0}^{\infty} \left(\frac{1 - wz - w^k(1 - z^k) - w^{k+1}(z^k - z)}{(1-w)(1-wz)} \right)^{y+1} (wt)^y ,$$

η οποία μπορεί να γραφεί ως

$$\sum_{n=y}^{\infty} a_n(z; y) w^n = w^y \left(\frac{1 - wz - w^k(1 - z^k) - w^{k+1}(z^k - z)}{(1-w)(1-wz)} \right)^{y+1}$$

ή ακόμη ως $\displaystyle\sum_{n=y}^{\infty} a_n(z; y) w^n = w^y \left(\frac{1}{(1-w)} \right)^{y+1} \left(1 - w^k \frac{1 - z^k + w(z^k - z)}{(1-wz)} \right)^{y+1}$.

Αναλύοντας το δεξί μέλος σε δυνάμεις του w και χρησιμοποιώντας τις συμβάσεις $\binom{n}{m} = 0,\ m < 0$ και $\binom{-1}{0} = 1$, μπορούμε να καταλήξουμε στην

$$a_n(z; y) = \sum_{m=0}^{\infty} \sum_{r=0}^{y+1} \sum_{i=0}^{r} (-1)^r \binom{y+s}{s} \binom{y+1}{r} \binom{r}{i} \binom{r+m-1}{m} (z^k - z)(1 - z^k)^{r-i} z^m,\ n \geq y$$

με $s = n - y - kr - i - m$.

Το τελικό αποτέλεσμα προκύπτει με περαιτέρω ανάλυση των δυνάμεων που εμφανίζονται στο άθροισμα με τη βοήθεια του διωνυμικού τύπου. □

Στο θέμα των δεσμευμένων κατανομών θα επιστρέψουμε στο Κεφάλαιο 6, κάνοντας χρήση των άνω αποτελεσμάτων σε ελέγχους τυχαιότητας.

3.6. Μελέτη της Κατανομής του Αθροίσματος των Μηκών των Ροών Μήκους Τουλάχιστον *k*, σε Ακολουθίες Μαρκοβιανά Εξαρτημένων Διτιμων Δοκιμών

Η μεθοδολογία της Μαρκοβιανής εμφύτευσης μπορεί εύκολα να επεκταθεί και στη περίπτωση που οι n δοκιμές παρουσιάζουν εξάρτηση κατά Markov. Δηλαδή, με χρήση της μεθοδολογίας εμφύτευσης μπορούν να μελετηθούν σε περίπτωση εξάρτησης, τυχαίες μεταβλητές που αφορούν τον αριθμό των μη επικαλυπτόμενων ροών, των επικαλυπτόμενων ροών επιτυχιών μήκους k, των ροών επιτυχιών μήκους τουλάχιστον k αλλά και του αθροίσματος των ροών επιτυχιών μήκους τουλάχιστον k.

Στην περίπτωση της εξάρτησης οι απαραίτητες τροποποιήσεις περιορίζονται στον πίνακα μεταπήδησης $\mathbf{\Lambda}_t$ και στους αντίστοιχους παραγόμενους πίνακες.

Στην συγκεκριμένη παράγραφο θα περιγράψουμε τις τροποποιήσεις του πίνακα μεταπήδησης $\mathbf{\Lambda}_t$ και στους αντίστοιχους παραγόμενους πίνακες για την περίπτωση της τυχαίας μεταβλητής $S_{n,k}$.

Εφαρμογή 3.5: Υποθέτουμε ότι $Z_1, Z_2, ..., Z_n$ είναι μια ακολουθία Μαρκοβιανά εξαρτημένων δοκιμών με δυνατά αποτελέσματα επιτυχία (1) και αποτυχία (0) και αντίστοιχες πιθανότητες $p_{00} = P(Z_t = 0 \mid Z_{t-1} = 0)$, $p_{01} = P(Z_n = 1 \mid Z_{n-1} = 0)$, $p_{10} = P(Z_t = 1 \mid Z_{t-1} = 0)$, $p_{11} = P(Z_t = 1 \mid Z_{t-1} = 1)$ $t = 1,2,...,n$ (με $Z_0 = 0$). Προκειμένου να υπολογίσουμε την κατανομή της τυχαίας μεταβλητής $S_{n,k}$ στην περίπτωση των δοκιμών που παρουσιάζουν Μαρκοβιανή εξάρτηση, ορίζουμε ως $\boldsymbol{\pi}_1 = (p_1, p_0, 0, ..., 0)$ τις αρχικές πιθανότητες της αλυσίδας ενώ, η γενική μορφή των πινάκων $\mathbf{A}_{t,0}(x), \mathbf{A}_{t,1}(x), \mathbf{A}_{t,k}(x)$ για οποιαδήποτε k και n έχει ως εξής:

$$\mathbf{A}_{t,0}(x)=\begin{bmatrix} & (x,0) & (x,1) & (x,2) & (x,3) & \mathbb{M} & (x,k-1) & (x,k) \\ (x,0) & p_{00} & p_{01} & & & \mathbb{M} & & \\ (x,1) & p_{10} & & p_{01} & & \mathbb{M} & & \\ (x,2) & p_{10} & & & p_{01} & \mathbb{M} & & \\ (x,3) & p_{10} & & & & \mathbb{M} & & \\ \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{O} & p_{01} & \mathbf{L} \\ (x,k) & p_{10} & & & & & & \\ (x,k+1) & p_{10} & & & & \mathbb{M} & & \end{bmatrix}_{(k+1)\times(k+1)}, \quad x=0,1,2,\ldots,\mathbf{l}_n$$

$$\mathbf{A}_{t,1}(x)=\begin{bmatrix} & (x,0) & (x,1) & (x,2) & (x,3) & \mathbb{M} & (x,k-1) & (x,k) \\ (x,0) & & & & & \mathbb{M} & & \\ (x,1) & & & & & \mathbb{M} & & \\ (x,2) & & & & & \mathbb{M} & & \\ (x,3) & & & & & \mathbb{M} & & \\ \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{O} & \mathbf{L} & \mathbf{L} \\ (x,k) & & & & & & & \\ (x,k+1) & & & & & \mathbb{M} & & p_{11} \end{bmatrix}_{(k+1)\times(k+1)}, \quad x=0,1,2,\ldots,\mathbf{l}_n$$

$$\mathbf{A}_{t,k}(x)=\begin{bmatrix} & (x,0) & (x,1) & (x,2) & (x,3) & \mathbb{M} & (x,k-1) & (x,k) \\ (x,0) & & & & & \mathbb{M} & & \\ (x,1) & & & & & \mathbb{M} & & \\ (x,2) & & & & & \mathbb{M} & & \\ (x,3) & & & & & \mathbb{M} & & \\ \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \mathbf{O} & \mathbf{L} & \mathbf{L} \\ (x,k) & & & & & & & p_{11} \\ (x,k+1) & & & & & \mathbb{M} & & \end{bmatrix}_{(k+1)\times(k+1)}, \quad x=0,1,2,\ldots,\mathbf{l}_n .$$

Στηριζόμενοι στην παραπάνω μορφή για τους πίνακες $\mathbf{A}_{t,0}(x),\mathbf{A}_{t,1}(x),\mathbf{A}_{t,k}(x)$ έχουμε το Θεώρημα 3.11 για την διπλή γεννήτρια της $S_{n,k}$ στην περίπτωση που η ακολουθία των δοκιμών παρουσιάζει Μαρκοβιανή εξάρτηση. Εδώ θα πρέπει να τονίσουμε ότι στη σχέση που δίνεται στο Θεώρημα 3.11 μπορεί να στηριχθεί η αντίστοιχη ανάπτυξη αποτελεσμάτων για την μονή γεννήτρια συνάρτηση, για την συνάρτηση πιθανότητας καθώς και μιας σειράς αποτελεσμάτων για τις ροπές της κατανομής.

Θεώρημα 3.11: *Υποθέτουμε ότι* $Z_1, Z_2, ..., Z_n$ *είναι μια ακολουθία Μαρκοβιανά εξαρτημένων δοκιμών* **Bernoulli** *με δυνατά αποτελέσματα επιτυχία* (1) *και αποτυχία* (0) *και αντίστοιχες πιθανότητες* $p_{00} = P(Z_t = 0 \mid Z_{t-1} = 0)$, $p_{01} = P(Z_n = 1 \mid Z_{n-1} = 0)$, $p_{10} = P(Z_t = 1 \mid Z_{t-1} = 0)$, $p_{11} = P(Z_t = 1 \mid Z_{t-1} = 1)$ $t = 1,2,...,n$ $(με Z_0 = 0)$.

Η διπλή γεννήτρια της $S_{n,k}$ *δίνεται από τον τύπο*

$$\Phi_n(z,w) = \frac{Q_1(z,w)}{Q_2(z,w)}$$

όπου

$$Q_1(z,w) = 1-(a + p_{11}z) + w^2 a p_{11} z - w^k p_1 p_{11}^{k-1}(1-z^k) - w^{k+1} p_{11}^k - w^{k+1} p_{11}^k [p_1(z^k - z) + g(1-z^k)] - w^{k+2} p_{11}^{k+1} g(z^k - z)$$

και

$$Q_2(z,w) = 1 - w(1 + a + p_{11}z) + w^2 [a + p_{11}z(1+a)] - w^3 a p_{11} z - w^{k+1} b(1-z^k) + w^{k+2} b p_{11}(z^k - z)$$

με $a = p_{11} - p_{01}$, $b = p_{10} p_{01} p_{11}^{k-1}$ *και* $g = p_{01} - p_1$.

Απόδειξη: Η απόδειξη προκύπτει ύστερα από μια σειρά μακροσκελών αλγεβρικών πράξεων με χρήση της σχέσης

$$\Phi(z,w) = 1 + f_1(z)[\mathbf{I} - w[\mathbf{A}_{t,0}(x) + z\mathbf{A}_{t,1}(x) + z^k \mathbf{A}_{t,k}(x)]]^{-1}\mathbf{1}'$$

με $f_1(z) = (p_0, p_1, 0, ..., 0)$. □

3.7. Ανακεφαλαίωση

Στο Κεφάλαιο 3, εφαρμόσαμε τα αποτελέσματα του Κεφαλαίου 2, σε μονοδιάστατες τυχαίες μεταβλητές που σχετίζονται με ροές επιτυχιών. Εφαρμόσαμε τη μεθοδολογία μελέτης των τυχαίων μεταβλητών πολυωνυμικού τύπου που αναπτύχθηκε στο Κεφάλαιο 2, προκειμένου να μελετήσουμε, μια στατιστική συνάρτηση που καταγράφει το άθροισμα των μηκών των ροών μήκους τουλάχιστον k σε ακολουθίες τόσο ανεξάρτητων όσο και εξαρτημένων (κατά Markov) δοκιμών Bernoulli, καθώς και τη δεσμευμένη (ως προς το συνολικό αριθμό επιτυχιών) κατανομή αυτής.

ΚΕΦΑΛΑΙΟ 4: ΠΟΛΥΔΙΑΣΤΑΤΕΣ ΚΑΤΑΝΟΜΕΣ ΣΧΕΤΙΚΕΣ ΜΕ ΡΟΕΣ ΕΠΙΤΥΧΙΩΝ

4.1. Εισαγωγή

Το Κεφαλαίο 4 αφιερώνεται στη μελέτη διδιάστατων μεταβλητών που αφορούν ροές. Όπως αναφέρθηκε στο Κεφάλαιο 1, ένας σημαντικός αριθμός ερευνητών (Philippou et al. (1989), Philippou and Antzoulakos (1990), Ling and Tai (1990), Alexandrou (1997), Godbole et al. (1997), Doi and Yamamoto (1998), Han and Aki (1999), Chadjiconstantinidis et al. (2000), Han (2001)) προχώρησαν σε πολυδιάστατες γενικεύσεις των μονοδιάστατων κατανομών που σχετίζονται με ροές επιτυχιών. Τόσο η ιδέα των MVB όσο και των MVP μπορεί εύκολα να προσαρμοστεί σε ένα πολυμεταβλητό πλαίσιο. Μια τέτοια επέκταση προσφέρει αποτελεσματικά εργαλεία για την εύρεση της από κοινού κατανομής τυχαίων μεταβλητών που απαριθμούν ροές και γενικότερα σχηματισμούς σε ακολουθίες δίτιμων και πολύτιμων δοκιμών. Η γενική θεωρία στην περίπτωση των διμεταβλητών και πολυμεταβλητών MVB και MVP παρουσιάστηκε στο Κεφάλαιο 2.

Το συνεχώς αυξανόμενο ενδιαφέρον για την ανάλυση εφαρμογών (στατιστικός έλεγχος ποιότητας, θεωρία αξιοπιστίας, απαραμετρικά κριτήρια τυχαιότητας, οικολογία, μετεωρολογία, αλυσίδες DNA) ώθησε στη γενίκευση της έννοιας των πολυδιάστατων διωνυμικών κατανομών τάξης k προς πολλές κατευθύνσεις.

Η πρώτη κατεύθυνση που παρουσιάστηκε αφορούσε διδιάστατες και πολυδιάστατες μεταβλητές οι οποίες απαριθμούν ροές επιτυχιών του ιδίου τύπου με διαφορετικά μήκη. Μια άλλη κατεύθυνση αφορούσε διδιάστατες και πολυδιάστατες μεταβλητές οι οποίες απαριθμούν και πάλι ροές επιτυχιών αλλά διαφορετικών τύπων και του ιδίου μήκους ενώ μια τρίτη κατεύθυνση, αφορούσε διδιάστατες και πολυδιάστατες μεταβλητές οι

οποίες απαριθμούν ροές επιτυχιών διαφορετικού τύπου και με διαφορετικά μήκη. Οι προαναφερθήσες κατευθύνσεις αφορούν την περίπτωση όπου η ακολουθία αποτελείται από δίτιμες δοκιμές Bernoulli. Έτσι, όλες οι παραπάνω περιπτώσεις ενοποιούνται υπό ένα πλαίσιο που αφορά διδιάστατες και πολυδιάστατες μεταβλητές οι οποίες απαριθμούν ροές επιτυχιών σε ακολουθίες δοκιμών με περισσότερα των δύο αποτελεσμάτων.

Μία άλλη κατεύθυνση αφορά διδιάστατες και πολυδιάστατες μεταβλητές οι οποίες απαριθμούν ή αθροίζουν τόσο ροές επιτυχιών όσο και ροές αποτυχιών σε ακολουθίες δίτιμων δοκιμών. Στο κεφάλαιο αυτό μελετάμε μια διδιάστατη περίπτωση τυχαίας μεταβλητής η οποία αθροίζει τα μήκη των ροών επιτυχιών ενώ καταγράφει συγχρόνως και τον αριθμό των επιτυχιών (βλέπε επίσης και Koutras et al. (2005b)).

Μια τελευταία κατεύθυνση αφορά διδιάστατες και πολυδιάστατες μεταβλητές οι οποίες απαριθμούν ροές επιτυχιών σε περισσότερες της μίας, εξαρτημένες, ακολουθίες δοκιμών, με περισσότερα των δύο αποτελεσμάτων.

Η μελέτη όλων των παραπάνω προβλημάτων είναι δυνατόν να γίνει με τη χρήση της τεχνικής της εμφύτευσης των τυχαίων μεταβλητών σε κατάλληλη Μαρκοβιανή αλυσίδα.

4.2. Μελέτη Διδιάστατων Τυχαίων Μεταβλητών που Απαριθμούν Ροές Επιτυχιών.

Οι Godbole et al. (1997) γενίκευσαν τις διωνυμικές κατανομές τάξης k, εισάγοντας τις πολυδιάστατες τυχαίες μεταβλητές που απαριθμούν τις ροές επιτυχιών με διάφορα μήκη. Συγκεκριμένα, έστω $Z_1, Z_2, ..., Z_n$ μια ακολουθία ανεξάρτητων και ισόνομων δοκιμών Bernoulli με δυνατά αποτελέσματα επιτυχία (1) και αποτυχία (0) και αντίστοιχες πιθανότητες $p_t = P(Z_t = 1)$ και $q_t = 1 - p_t$ για $t = 1,2,...,n$ με αρχική συνθήκη $Z_0 = 0$. Στη συγκεκριμένη ακολουθία είναι δυνατόν να ορίσουμε ένα σύνολο από διδιάστατες τυχαίες μεταβλητές που απαριθμούν ροές επιτυχιών (του ιδίου τύπου) με διαφορετικά μήκη, όπως οι $(N_{n,k}, N_{n,r})$, $(M_{n,k}, M_{n,r})$, $(G_{n,k}, G_{n,r})$. Για παράδειγμα, σε $n = 15$ διαδοχικές επαναλήψεις ενός πειράματος τύχης το οποίο αφορά την ρίψη

ενός νομίσματος με επιτυχία την ένδειξη κεφάλι (1) και αποτυχία την ένδειξη γράμματα (0), πήραμε την παρακάτω ακολουθία αποτελεσμάτων 110001110111011. Παρατηρούμε ότι έχουμε $(N_{n,2}, N_{n,3}) = (4,2)$, $(M_{n,2}, M_{n,3}) = (6,2)$ και τέλος $(G_{n,2}, G_{n,3}) = (4,2)$.

Σε ακολουθίες της ίδιας μορφής είναι δυνατόν να ορίσουμε ένα σύνολο από διδιάστατες τυχαίες μεταβλητές που απαριθμούν ροές επιτυχιών (διαφορετικού) τύπου με διαφορετικά μήκη, όπως οι $(N_{n,k}, M_{n,r})$, $(M_{n,k}, G_{n,r})$ και άλλες. Στο ίδιο παράδειγμα, παρατηρούμε ότι $(N_{n,3}, M_{n,2}) = (2,6)$ και $(M_{n,2}, G_{n,2}) = (6,4)$.

Οι Guibas and Odlyzko (1978, 1980, 1981) θεωρώντας ένα αλφάβητο με n γράμματα ανέπτυξαν μια γενική μέθοδο για τη μελέτη απαριθμητριών σχηματισμών γραμμάτων, στους οποίους δεν περιλαμβάνονται κάποια συγκεκριμένα σχήματα (patterns). Στην συνέχεια οι Alexandrou (1997) και Han and Aki (1999) μελέτησαν διδιάστατες τυχαίες μεταβλητές του ιδίου τύπου με χρήση της τεχνικής της εμφύτευσης σε Μαρκοβιανή αλυσίδα. Η συγκεκριμένη περίπτωση αποτελεί ουσιαστικά ειδική περίπτωση της γενικής μεθοδολογίας που εισήχθη από τον Fu (1996), ο οποίος μελέτησε την ακριβή κατανομή σχηματισμών (patterns), με χρήση της τεχνικής της εμφύτευσης σε Μαρκοβιανή αλυσίδα.

Έστω $Z_1, Z_2, ..., Z_n$ μια ακολουθία ανεξάρτητων και ισόνομων δοκιμών με δυνατά αποτελέσματα, επιτυχία $1^{ου}$ τύπου (1), επιτυχία $2^{ου}$ τύπου (2), και αποτυχία (0) και αντίστοιχες πιθανότητες $P(Z_t = 1) = p_1$, $P(Z_t = 2) = p_2$ και $P(Z_t = 0) = q = 1 - p_1 - p_2$ για $t = 1,2,...,n$ με αρχική συνθήκη $Z_0 = 0$. Στη συγκεκριμένη ακολουθία είναι δυνατόν να ορίσουμε ένα σύνολο από διδιάστατες τυχαίες μεταβλητές που απαριθμούν ροές επιτυχιών τύπου 1 και 2 με διαφορετικά μήκη, όπως οι $(N_{n,k}^{(1)}, M_{n,r}^{(2)})$, $(M_{n,k}^{(1)}, G_{n,r}^{(2)})$, και άλλες, όπου για παράδειγμα η $N_{n,k}^{(1)}$ καταγράφει τον αριθμό των εμφανίσεων μη επικαλυπτόμενων ροών επιτυχιών ($1^{ου}$ τύπου) μήκους k και η $M_{n,r}^{(2)}$ καταγράφει τον αριθμό των εμφανίσεων επικαλυπτόμενων ροών επιτυχιών ($2^{ου}$ τύπου) μήκους r. Για παράδειγμα, σε $n = 15$ διαδοχικές επαναλήψεις ενός πειράματος τύχης με επιτυχία $1^{ου}$ τύπου την ένδειξη (1), με επιτυχία $2^{ου}$ τύπου την ένδειξη (2), και αποτυχία την ένδειξη

(0), πήραμε την ακόλουθη ακολουθία αποτελεσμάτων 112201222110011. Παρατηρούμε ότι έχουμε $(N_{n,2}^{(1)}, M_{n,2}^{(2)}) = (3,3)$ και $(M_{n,2}^{(1)}, G_{n,2}^{(2)}) = (3,2)$.

Προκειμένου να γίνει κατανοητή η μεθοδολογία του υπολογισμού της ακριβούς κατανομής διδιάστατων ή και πολυδιάστατων τυχαίων μεταβλητών με χρήση της τεχνικής εμφύτευσης σε Μαρκοβιανή αλυσίδα, ακολουθούν μερικές ειδικές περιπτώσεις της γενικής μεθοδολογίας.

Στην Εφαρμογή 4.1 δίνουμε αναλυτικά τη μεθοδολογία εμφύτευσης της απαριθμήτριας τυχαίας μεταβλητής $(N_{n,k}^{(1)}, N_{n,r}^{(2)})$ σε Μαρκοβιανή Αλυσίδα. Η μεθοδολογία εμφύτευσης οφείλεται στην Alexandrou (1997).

Εφαρμογή 4.1: Τεχνική Εμφύτευσης της τυχαίας μεταβλητής $(N_{n,k}^{(1)}, N_{n,r}^{(2)})$

Υποθέτουμε ότι $Z_1, Z_2, ..., Z_n$ είναι μια ακολουθία ανεξάρτητων δοκιμών Bernoulli με δυνατά αποτελέσματα επιτυχία $1^{ου}$ τύπου (1), με επιτυχία $2^{ου}$ τύπου (2), και αποτυχία (0) και αντίστοιχες πιθανότητες $P(Z_t = 1) = p_1$, $P(Z_t = 2) = p_2$ και $q = 1 - p_1 - p_2$ για $t = 1,2,...,n$ με αρχική συνθήκη $Z_0 = 0$. Θα εφαρμόσουμε την μέθοδο της Μαρκοβιανής εμφύτευσης προκειμένου να υπολογίσουμε την κατανομή της τυχαίας μεταβλητής $(N_{n,k}^{(1)}, N_{n,r}^{(2)})$ η οποία καταγράφει τον αριθμό των μη επικαλυπτόμενων ροών επιτυχιών τύπου 1 μήκους k καθώς και τον αριθμό των μη επικαλυπτόμενων ροών επιτυχιών τύπου 2 μήκους r.

Βήμα 1: Θεωρούμε το χώρο καταστάσεων $\Omega = \bigcup_{x_1, x_2 \geq 0} C_{x_1,x_2}$ με

$C_{x_1,x_2} = \{c_{x_1,x_2;0,0}, c_{x_1,x_2;1,1}, ..., c_{x_1,x_2;k-1,1}, c_{x_1,x_2;1,2}, c_{x_1,x_2;2,2}, ..., c_{x_1,x_2;r-1,2}\}$, όπου ισχύει ότι $c_{x_1,x_2;m,j} = (x_1, x_2; m, j)$ με $1 \leq m \leq k-1$ για $j = 1$ και $1 \leq m \leq r-1$ για $j = 2$ και επιπλέον $c_{x_1,x_2;0,0} = (x_1, x_2; 0,0)$ για $x_1, x_2 \geq 0$.

Βήμα 2: Θεωρούμε τη Μαρκοβιανή αλυσίδα $\{Y_t : t \geq 0\}$ ορισμένη στον χώρο Ω.

Ορίζουμε

- $Y_t = (x_1, x_2, m, 1)$, εάν στο χρόνο $t \geq 0$ έχουν εμφανισθεί x_1 μη επικαλυπτόμενες ροές επιτυχιών $1^{ου}$ τύπου μήκους k, x_2 μη επικαλυπτόμενες ροές επιτυχιών $2^{ου}$ τύπου μήκους r, και επίσης m

επιτυχίες $1^{ου}$ τύπου από την τελευταία εμφάνιση αποτυχίας ή επιτυχίας $2^{ου}$ τύπου ή από την συμπλήρωση της τελευταίας ροής επιτυχιών $1^{ου}$ τύπου μήκους k.

- $Y_t = (x_1, x_2, m, 2)$, εάν στον χρόνο $t \geq 0$ έχουν εμφανισθεί x_1 μη επικαλυπτόμενες ροές επιτυχιών $1^{ου}$ τύπου μήκους k, x_2 μη επικαλυπτόμενες ροές επιτυχιών $2^{ου}$ τύπου μήκους r, και επίσης m επιτυχίες $2^{ου}$ τύπου από την τελευταία εμφάνιση αποτυχίας ή επιτυχίας $1^{ου}$ τύπου ή από την συμπλήρωση της τελευταίας ροής επιτυχιών $2^{ου}$ τύπου μήκους k.

Βήμα 3: Θέτουμε $\mathbf{l}_n^{(1)} = \left[\frac{n}{k}\right]$, $\mathbf{l}_n^{(2)} = \left[\frac{n}{r}\right]$ και συμβολίζουμε με

$C_{x_1,x_2} = \{c_{x_1,x_2;0,0}, c_{x_1,x_2;1,1}, ..., c_{x_1,x_2;k-1,1}, c_{x_1,x_2;1,2}, ..., c_{x_1,x_2;r-1,2}\}$ (δηλαδή το σύνολο των πιθανών υποκαταστάσεων της αλυσίδας). Το σύνολο των C_{x_1,x_2} για όλα τα δυνατά ζεύγη (x_1, x_2) αποτελεί μια διαμέριση του χώρου $\Omega = \bigcup_{x_1,x_2 \geq 0} C_{x_1,x_2}$.

Βήμα 4: Προφανώς ότι για κάθε (x_1, x_2) με $0 \leq x_1 \leq \mathbf{l}_n^{(1)}$ και $0 \leq x_2 \leq \mathbf{l}_n^{(2)}$ ισχύει $P(N_{n,k}^1 = x_1, N_{n,r}^2 = x_2) = P(Y_t \in C_x)$.

Από τα βήματα που προηγήθηκαν προκύπτει ότι ικανοποιείται ο Ορισμός 2.4 (δηλαδή η τυχαία μεταβλητή που μελετάμε είναι τύπου *BMVB*) και συνεπώς η τυχαία μεταβλητή $(N_{n,k}^{(1)} = x_1, N_{n,r}^{(2)} = x_2)$ εμφυτεύεται σε Μαρκοβιανή αλυσίδα. Το επόμενο βήμα της μεθόδου είναι να ορίσουμε τις αρχικές πιθανότητες της αλυσίδας $\boldsymbol{\pi}_{0,0} = (1,0,0,...,0)$, τον «εσωτερικό» $(k+r-1)\times(k+r-1)$ πίνακα μετάβασης $\mathbf{A}_t(x_1, x_2) = (P(Y_t = c_{x_1,x_2,j'} \mid Y_{t-1} = c_{x_1,x_2,j}))_{s\times s}$, ο οποίος είναι ίσος με

$$\mathbf{A}_t(x_1,x_2)=\begin{bmatrix}
 & (0,0) & (1,1) & (2,1) & \mathbf{L} & (k-2,1) & (k-1,1) & (1,2) & (2,1) & \mathbf{L} & (r-2.2) & (r-1,2)\\
(0,0) & q & p_1 & 0 & \mathbf{L} & 0 & 0 & p_2 & 0 & \mathbf{L} & 0 & 0\\
(1,1) & q & 0 & p_1 & \mathbf{L} & 0 & 0 & p_2 & 0 & \mathbf{L} & 0 & 0\\
(2,1) & q & 0 & 0 & \mathbf{L} & 0 & 0 & p_2 & 0 & \mathbf{L} & 0 & 0\\
\mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} & \mathbf{M}\\
(k-2,1) & 0 & 0 & 0 & \mathbf{L} & 0 & p_1 & p_2 & 0 & \mathbf{L} & 0 & 0\\
(k-1,1) & q & 0 & 0 & \mathbf{L} & 0 & 0 & p_2 & 0 & \mathbf{L} & 0 & 0\\
(1,2) & q & 0 & 0 & \mathbf{L} & 0 & 0 & 0 & p_2 & \mathbf{L} & 0 & 0\\
(2,2) & q & 0 & 0 & \mathbf{L} & 0 & 0 & 0 & 0 & \mathbf{L} & 0 & 0\\
\mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} & 0 & \mathbf{M} & \mathbf{M} & \mathbf{L} & \mathbf{M} & \mathbf{M}\\
(r-2,2) & q & 0 & 0 & \mathbf{L} & 0 & 0 & 0 & & \mathbf{L} & 0 & p_2\\
(r-1,2) & q & 0 & 0 & \mathbf{L} & 0 & 0 & 0 & 0 & \mathbf{L} & 0 & 0
\end{bmatrix}$$

και τους «εξωτερικούς» πίνακες μετάβασης $\mathbf{B}_t^{(1)}(x_1,x_2)=(P(Y_t=c_{x_1+i,x_2,j'}\mid Y_{t-1}=c_{x_1,x_2,j})$ οι οποίοι θα έχουν όλα τα στοιχεία ίσα με 0 πλην του $(k,1)$ στοιχείου και $\mathbf{B}_t^{(2)}(x_1,x_2)=(P(Y_t=c_{x_1,x_2+i,j'}\mid Y_{t-1}=c_{x_1,x_2,j})$ που θα έχουν όλα τα στοιχεία ίσα με 0 πλην του $(k+r-1,1)$ στοιχείου. Στην συνέχεια, ορίζοντας τα διανύσματα πιθανότητας

$$\mathbf{f}_t(x_1,x_2)=(P(Y_t=c_{x_1,x_2,0}),P(Y_t=c_{x_1,x_2,1}),\ldots,P(Y_t=c_{x_1,x_2,s-1})),\ x_1,x_2\ge 0,\ t\ge 0,$$

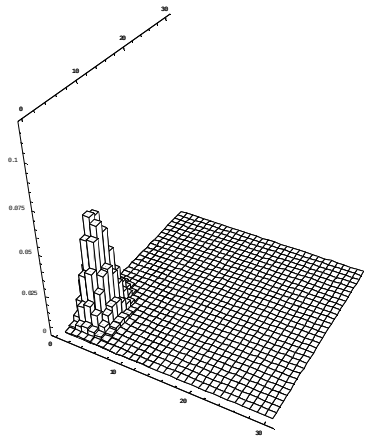
προκύπτει ότι

$$P(X_n^{(1)}=x_1,X_n^{(2)}=x_2)=P(Y_n\in C_{x_1,x_2})=\mathbf{f}_n(x_1,x_2)\mathbf{1}'.$$

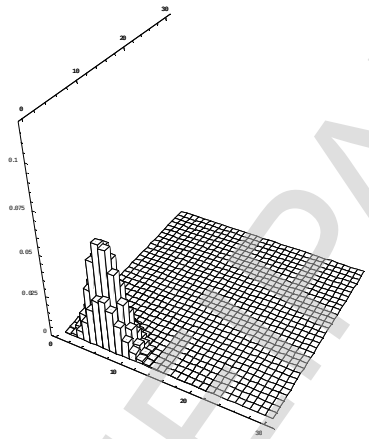
Τέλος, με χρήση του Θεωρήματος 2.10 είναι δυνατός ο υπολογισμός της συνάρτησης πιθανότητας της $(N_{n,k}^1=x_1,N_{n,r}^2=x_2)$.

Στο Σχήμα 4.1 δίνεται η συνάρτηση πιθανότητας f της διδιάστατης τυχαίας μεταβλητής $(N_{n,k}^{(1)},N_{n,r}^{(2)})$ για διάφορες τιμές των n,k,r,p_1,p_2. Η συνάρτηση πιθανότητας της $(N_{n,k}^{(1)}=x_1,N_{n,r}^{(2)}=x_2)$ υπολογίστηκε με χρήση της Μαρκοβιανής προσέγγισης.

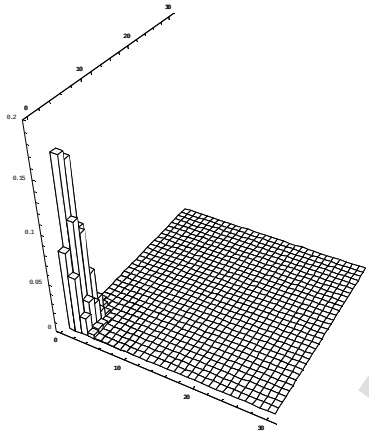
Σχήμα 4.1: Η συνάρτηση πιθανότητας της $(N_{n,k}^{(1)}, N_{n,r}^{(2)})$ για διάφορα n, k, r, p_1, p_2



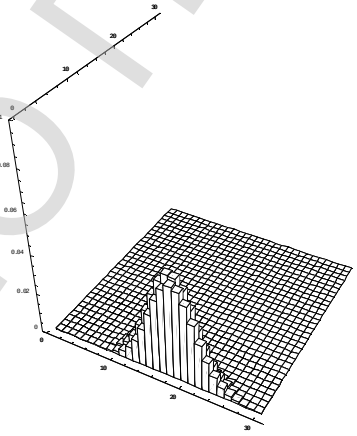
$n = 60, k = 3, r = 2, p_1 = 0.35, p_2 = 0.35.$



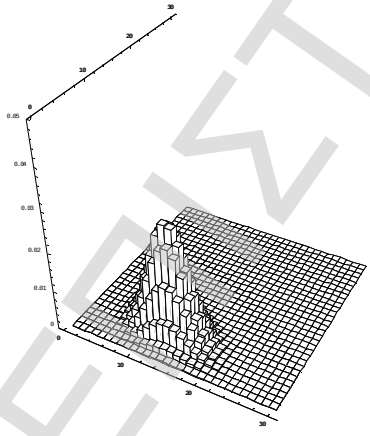
$n = 60, k = 2, r = 3, p_1 = 0.35, p_2 = 0.35.$



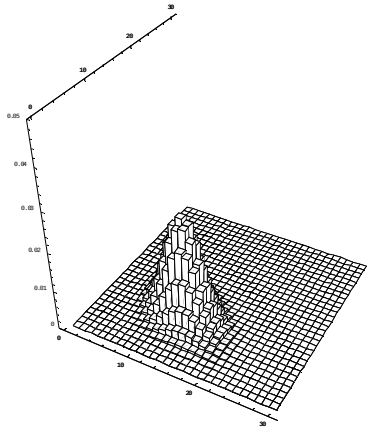
$n = 60, k = 4, r = 3, p_1 = 0.35, p_2 = 0.35.$



$n = 60, k = 2, r = 4, p_1 = 0.35, p_2 = 0.35.$



$n = 100, k = 2, r = 3, p_1 = 0.40, p_2 = 0.45.$



$n = 100, k = 2, r = 3, p_1 = 0.40, p_2 = 0.55.$

Στην Εφαρμογή 4.2 προχωράμε στην περιγραφή της μεθοδολογίας εμφύτευσης της απαριθμήτριας τυχαίας μεταβλητής $(M_{n,k}^{(1)}, M_{n,r}^{(2)})$ σε Μαρκοβιανή Αλυσίδα (βλέπε Alexandrou (1997)).

Εφαρμογή 4.2: Τεχνική Εμφύτευσης της τυχαίας μεταβλητής $(M_{n,k}^{(1)}, M_{n,r}^{(2)})$

Έστω ότι $Z_1, Z_2, ..., Z_n$ είναι μια ακολουθία ανεξάρτητων δοκιμών Bernoulli με δυνατά αποτελέσματα επιτυχία $1^{ου}$ τύπου (1), με επιτυχία $2^{ου}$ τύπου (2), και αποτυχία (0) και αντίστοιχες πιθανότητες $P(Z_t = 1) = p_1$, $P(Z_t = 2) = p_2$ και $q = 1 - p_1 - p_2$ για $t = 1,2,...,n$ με αρχική συνθήκη $Z_0 = 0$. Θα εφαρμόσουμε τη μέθοδο της Μαρκοβιανής εμφύτευσης προκειμένου να υπολογίσουμε την κατανομή της τυχαίας μεταβλητής $(M_{n,k}^{(1)}, M_{n,r}^{(2)})$ η οποία καταγράφει τον αριθμό των επικαλυπτόμενων ροών μήκους k.

Βήμα 1: Θεωρούμε το χώρο καταστάσεων $\Omega = \bigcup_{x_1, x_2 \geq 0} C_{x_1,x_2}$ με

$C_{x_1,x_2} = \{c_{x_1,x_2;0,0}, c_{x_1,x_2;1,1}, ..., c_{x_1,x_2;k-1,1}, c_{x_1,x_2;-1,1}, c_{x_1,x_2;1,2}, ..., c_{x_1,x_2;r-1,2}, c_{x_1,x_2;-1,2},\}$, όπου ισχύει ότι $c_{x_1,x_2;m,j} = (x_1, x_2; m, j)$ με $\{\{-1\} \cup 1\{\leq m \leq k-1\}\}$ για $j = 1$ και $\{\{-1\} \cup 1\{\leq m \leq r-1\}\}$ για $j = 2$ και επιπλέον $c_{x_1,x_2;0,0} = (x_1, x_2; 0, 0)$ για $x_1, x_2 \geq 0$.

Βήμα 2: Θεωρούμε την Μαρκοβιανή αλυσίδα $\{Y_t : t \geq 0\}$ ορισμένη στον χώρο Ω.

Ορίζουμε

- $Y_t = (x_1, x_2, m, 1)$, εάν στον χρόνο $t \geq 0$ έχουν εμφανισθεί x_1 επικαλυπτόμενες ροές επιτυχιών $1^{ου}$ τύπου μήκους k, x_2 επικαλυπτόμενες ροές επιτυχιών $2^{ου}$ τύπου μήκους r, και επίσης $m \leq k-1$ επιτυχίες $1^{ου}$ τύπου από τη τελευταία εμφάνιση αποτυχίας ή επιτυχίας $2^{ου}$ τύπου.
- $Y_t = (x_1, x_2, -1, 1)$, εάν στον χρόνο $t \geq 0$ έχουν εμφανισθεί x_1 επικαλυπτόμενες ροές επιτυχιών $1^{ου}$ τύπου μήκους k, x_2 επικαλυπτόμενες ροές επιτυχιών $2^{ου}$ τύπου μήκους r, και επίσης μετά από τη συμπλήρωση της τελευταίας ροής επιτυχιών $1^{ου}$ τύπου μήκους k εμφανίσθηκε ξανά επιτυχία $1^{ου}$ τύπου.
- $Y_t = (x_1, x_2, m, 2)$, εάν στον χρόνο $t \geq 0$ έχουν εμφανισθεί x_1

επικαλυπτόμενες ροές επιτυχιών $1^{ου}$ τύπου μήκους k, x_2 επικαλυπτόμενες ροές επιτυχιών $2^{ου}$ τύπου μήκους r, και επίσης $m \le r-1$ επιτυχίες $2^{ου}$ τύπου από την τελευταία εμφάνιση αποτυχίας ή επιτυχίας $1^{ου}$ τύπου ή από τη συμπλήρωση της τελευταίας ροής επιτυχιών $2^{ου}$ τύπου μήκους k.

- $Y_t = (x_1, x_2, -1, 2)$, εάν στον χρόνο $t \ge 0$ έχουν εμφανισθεί x_1 επικαλυπτόμενες ροές επιτυχιών $1^{ου}$ τύπου μήκους k, x_2 επικαλυπτόμενες ροές επιτυχιών $2^{ου}$ τύπου μήκους r, και επίσης μετά από τη συμπλήρωση της τελευταίας ροής επιτυχιών $2^{ου}$ τύπου μήκους k εμφανίσθηκε ξανά επιτυχία $2^{ου}$ τύπου.

Βήμα 3: Ορίζουμε $\mathbf{l}_n^{(1)} = n-k+1$, $\mathbf{l}_n^{(2)} = n-r+1$ και συμβολίζουμε με

$$C_{x_1,x_2} = \{c_{x_1,x_2;0,0}, c_{x_1,x_2;1,1}, \ldots, c_{x_1,x_2;k-1,1}, c_{x_1,x_2;-1,1}, c_{x_1,x_2;1,2}, \ldots, c_{x_1,x_2;r-1,2}, c_{x_1,x_2;-1,2}, \}$$

(δηλαδή το σύνολο των πιθανών υποκαταστάσεων της αλυσίδας). Το σύνολο των C_{x_1,x_2} για κάθε (x_1, x_2) αποτελεί μια διαμέριση του χώρου $\Omega = \bigcup_{x_1,x_2 \ge 0} C_{x_1,x_2}$.

Βήμα 4: Προφανώς ισχύει ότι για κάθε (x_1, x_2) με $0 \le x_1 \le \mathbf{l}_n^{(1)}$ και $0 \le x_2 \le \mathbf{l}_n^{(2)}$ ισχύει $P(N_{n,k}^{(1)} = x_1, N_{n,r}^{(2)} = x_2) = P(Y_t \in C_x)$.

Από τα βήματα που προηγήθηκαν προκύπτει εύκολα ότι ικανοποιείται ο Ορισμός 2.4 (δηλαδή η τυχαία μεταβλητή που μελετάμε είναι τύπου *BMVB*) και συνεπώς η τυχαία μεταβλητή $(N_{n,k}^{(1)} = x_1, N_{n,r}^{(2)} = x_2)$ εμφυτεύεται σε Μαρκοβιανή αλυσίδα.

Το επόμενο βήμα της μεθόδου είναι να ορίσουμε τις αρχικές πιθανότητες της αλυσίδας ίσες με $\boldsymbol{\pi}_{0,0} = (1,0,0,\ldots,0)$, τον «εσωτερικό» $(k+r+1)\times(k+r+1)$ πίνακα μετάβασης $\mathbf{A}_t(x_1, x_2) = (P(Y_t = c_{x_1,x_2,j'} \mid Y_{t-1} = c_{x_1,x_2,j}))_{s\times s}$, ίσο με

$$\mathbf{A}_t(x_1,x_2)=\begin{bmatrix}
 & (0,1) & (1,1) & (2,1) & \mathbf{L} & (k-1,1) & (-1,1) & (1,2) & (2,2) & \mathbf{L} & (r-1,2) & (-1,2)\\
(0,0) & q & p_1 & 0 & \mathbf{L} & 0 & 0 & p_2 & 0 & \mathbf{L} & 0 & 0\\
(1,1) & q & 0 & p_1 & \mathbf{L} & 0 & 0 & p_2 & 0 & \mathbf{L} & 0 & 0\\
\mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} & \mathbf{M}\\
\mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{L} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{L} & \mathbf{M} & \mathbf{M}\\
(k-1,1) & q & 0 & 0 & \mathbf{L} & 0 & p_1 & p_2 & 0 & \mathbf{L} & 0 & 0\\
(-1,1) & q & 0 & 0 & \mathbf{L} & 0 & 0 & p_2 & 0 & \mathbf{L} & 0 & 0\\
(1,2) & q & 0 & 0 & \mathbf{L} & 0 & 0 & 0 & p_2 & \mathbf{L} & 0 & 0\\
\mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} & \mathbf{M}\\
\mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{L} & \mathbf{M} & 0 & \mathbf{M} & \mathbf{M} & \mathbf{L} & \mathbf{M} & \mathbf{M}\\
(r-1,2) & q & 0 & 0 & \mathbf{L} & 0 & 0 & 0 & & \mathbf{L} & 0 & 0\\
(-1,2) & q & 0 & 0 & \mathbf{L} & 0 & 0 & 0 & 0 & \mathbf{L} & 0 & 0
\end{bmatrix}$$

και τους «εξωτερικούς» πίνακες μετάβασης $\mathbf{B}_t^{(1)}(x_1,x_2)=(P(Y_t=c_{x_1+i,x_2,j'}\,|\,Y_{t-1}=c_{x_1,x_2,j})$ με όλα τα στοιχεία ίσα με 0 πλην των $(k,k+1),(k+1,k+1)$ στοιχείων που είναι ίσο με p_1 και $\mathbf{B}_t^{(2)}(x_1,x_2)=(P(Y_t=c_{x_1,x_2+i,j'}\,|\,Y_{t-1}=c_{x_1,x_2,j})$ με όλα τα στοιχεία ίσα με 0 πλην του $(k+r,k+r+1),(k+r+1,k+r+1)$ στοιχείου που είναι ίσο με p_2. Τέλος, ορίζοντας τα διανύσματα πιθανότητας

$$\mathbf{f}_t(x_1,x_2)=(P(Y_t=c_{x_1,x_2,0}),P(Y_t=c_{x_1,x_2,1}),\ldots,P(Y_t=c_{x_1,x_2,s-1})),\ x_1,x_2\geq 0,\ t\geq 0,$$

προκύπτει ότι

$$P(X_n^{(1)}=x_1,X_n^{(2)}=x_2)=P(Y_n\in C_{x_1,x_2})=\mathbf{f}_n(x_1,x_2)\mathbf{1}'.$$

Τέλος, με χρήση του Θεωρήματος 2.10 είναι δυνατός ο υπολογισμός της συνάρτησης πιθανότητας της $(M_{n,k}^{(1)},M_{n,r}^{(2)})$.

Η συνάρτηση πιθανότητας της $(M_{n,k}^{(1)},M_{n,r}^{(2)})$ υπολογίστηκε με χρήση της Μαρκοβιανής προσέγγισης. Για την λεπτομερή μελέτη της συγκεκριμένης μεταβλητής (αναδρομικοί τύποι για την πιθανότητα, γεννήτριες συναρτήσεις, ροπές) ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στους Balakrishnan and Koutras (2002) και Alexandrou (1997).

4.3. Μελέτη Διδιάστατων Τυχαίων Μεταβλητών που Απαριθμούν Ροές Επιτυχιών ή Αποτυχιών και Ταυτόχρονα Καταγράφουν το Άθροισμα των Μηκών των Ροών

Οι Ling and Tai (1990) και οι Chadjiconstantinidis et al. (2000) εργάστηκαν σε παρόμοιες κατευθύνσεις ενδιαφερόμενοι για την από κοινού κατανομή ροών επιτυχιών και αποτυχιών.

Σε μια ακολουθία δίτιμων δοκιμών είναι δυνατόν να παρατηρήσουμε ένα σύνολο από διδιάστατες τυχαίες μεταβλητές που απαριθμούν ροές επιτυχιών (S) και ροές αποτυχιών (F) (του ιδίου τύπου) με διαφορετικά μήκη, όπως οι $(N_{n,k}^{(S)}, N_{n,r}^{(F)})$, $(M_{n,k}^{(S)}, M_{n,r}^{(F)})$, $(G_{n,k}^{(S)}, G_{n,r}^{(F)})$.

Για παράδειγμα, σε $n=15$ διαδοχικές επαναλήψεις ενός πειράματος τύχης το οποίο αφορά την ρίψη ενός νομίσματος με επιτυχία την ένδειξη κεφάλι (1) και αποτυχία την ένδειξη γράμματα (0), πήραμε την ακολουθία αποτελεσμάτων 110001110111011. Παρατηρούμε ότι έχουμε $(N_{n,2}^{(S)}, N_{n,3}^{(F)}) = (4,1)$, $(M_{n,2}^{(S)}, M_{n,2}^{(F)}) = (6,2)$ και τέλος $(G_{n,2}^{(S)}, G_{n,1}^{(F)}) = (4,3)$. Η περίπτωση αυτή αποτελεί ειδική περίπτωση της μελέτης διδιάστατων τυχαίων μεταβλητών που απαριθμούν ροές επιτυχιών ή αποτυχιών ενώ ταυτόχρονα καταγράφουν το άθροισμα των μηκών των ροών μήκους τουλάχιστον k σε ακολουθίες δίτιμων δοκιμών.

Στην παραπάνω ακολουθία είναι δυνατόν να παρατηρήσουμε ένα σύνολο από διδιάστατες τυχαίες μεταβλητές που απαριθμούν ροές αποτυχιών (ανεξαρτήτως τύπου) ενώ ταυτόχρονα καταγράφουν το άθροισμα των μηκών των ροών μήκους τουλάχιστον k, όπως οι $(N_{n,r}^{(F)}, S_{n,k}^{(S)})$, $(M_{n,r}^{(F)}, S_{n,k}^{(S)})$, και άλλες, για παράδειγμα, παρατηρούμε ότι $(N_{n,2}^{(F)}, S_{n,2}^{(S)}) = (1,10)$ και $(M_{n,2}^{(F)}, S_{n,2}^{(S)}) = (2,10)$.

Στην Εφαρμογή 4.2 περιγράφουμε αναλυτικά την μεθοδολογία εμφύτευσης της απαριθμήτριας τυχαίας μεταβλητής $(S_{n,k}^{(S)}, N_{n,r}^{(F)})$ σε Μαρκοβιανή αλυσίδα.

Εφαρμογή 4.2: Τεχνική Εμφύτευσης της τυχαίας μεταβλητής $(S_{n,k}^{(S)}, N_{n,r}^{(F)})$

Υποθέτουμε ότι $Z_1, Z_2, ..., Z_n$ είναι μια ακολουθία ανεξάρτητων δοκιμών Bernoulli με δυνατά αποτελέσματα επιτυχία (1), και αποτυχία (0) και αντίστοιχες πιθανότητες $P(Z_t = 1) = p$, $t = 1, 2, ..., n$ και $q = 1 - p$ με αρχική συνθήκη $Z_0 = 0$. Θα εφαρμόσουμε τη μέθοδο της Μαρκοβιανής εμφύτευσης προκειμένου να υπολογίσουμε την κατανομή της τυχαίας μεταβλητής $(S_{n,k}, N_{n,r})$ η οποία καταγράφει τον αριθμό $N_{n,r}$ των μη επικαλυπτόμενων ροών αποτυχιών μήκους r και ταυτόχρονα το άθροισμα $S_{n,k}$ των μηκών των ροών επιτυχιών με μήκος τουλάχιστον k.

Βήμα 1: Θεωρούμε τον χώρο καταστάσεων $\Omega = \bigcup_{x_1, x_2 \geq 0} C_{x_1, x_2}$ με

$C_{x_1, x_2} = \{(\{x_1, x_2; 0, 0\}) \cup \{(x_1, x_2; i, 1) : 1 \leq i \leq k\} \cup \{(x_1, x_2; j, 0) : 1 \leq j \leq r-1\}\}$ για $x_1, x_2 \geq 0$.

Βήμα 2: Θεωρούμε τη Μαρκοβιανή αλυσίδα $\{Y_t : t \geq 0\}$ ορισμένη στο χώρο Ω.

Ορίζουμε

- $Y_t = (x_1, x_2, i, 1)$, εάν στο χρόνο $t \geq 0$ το άθροισμα των ροών επιτυχιών μήκους τουλάχιστον k είναι ίσο με x_1 και ο αριθμός των μη επικαλυπτόμενων ροών αποτυχιών είναι ίσο με x_2 και επίσης έχουν εμφανισθεί i επιτυχίες από την τελευταία εμφάνιση αποτυχίας (θέτωντας $i=k$ εάν το $i>k$).
- $Y_t = (x_1, x_2, j, 0)$, εάν στο χρόνο $t \geq 0$ το άθροισμα των ροών επιτυχιών μήκους τουλάχιστον k είναι ίσο με x_1 και ο αριθμός των μη επικαλυπτόμενων ροών αποτυχιών είναι ίσο με x_2 και επίσης έχουν εμφανισθεί $j \leq r-1$ αποτυχίες από την τελευταία εμφάνιση επιτυχίας.

Βήμα 3: Θεωρούμε $\mathbf{l}_n^{(1)} = n$, $\mathbf{l}_n^{(2)} = \left[\frac{n}{r}\right]$ και συμβολίζουμε με C_{x_1, x_2} το σύνολο των πιθανών υποκαταστάσεων της αλυσίδας. Το σύνολο των C_{x_1, x_2} για κάθε (x_1, x_2) αποτελεί μια διαμέριση του χώρου $\Omega = \bigcup_{x_1, x_2 \geq 0} C_{x_1, x_2}$.

Βήμα 4: Προφανώς ισχύει ότι για κάθε (x_1, x_2) με $0 \leq x_1 \leq \mathbf{l}_n^{(1)}$ και

$0 \le x_2 \le \mathbf{l}_n^{(2)}$ ισχύει $P(S_{n,k} = x_1, N_{n,r} = x_2) = P(Y_t \in C_x)$.

Από τα βήματα που προηγήθηκαν προκύπτει εύκολα ότι ικανοποιείται ο Ορισμός 2.5 (δηλαδή η τυχαία μεταβλητή είναι BMVP) και συνεπώς η τυχαία μεταβλητή $(S_{n,k}^F, N_{n,r}^S)$ εμφυτεύεται σε Μαρκοβιανή αλυσίδα.

Το επόμενο βήμα της μεθόδου είναι να ορίσουμε τις αρχικές πιθανότητες της αλυσίδας $\boldsymbol{\pi}_{0,0} = (1,0,0,...,0)$, τον «εσωτερικό» $(k+r)\times(k+r)$ πίνακα μετάβασης $\mathbf{A}_t(x_1,x_2) = (P(Y_t = c_{x_1,x_2,j'} \mid Y_{t-1} = c_{x_1,x_2,j}))_{s\times s}$, ο οποίος είναι ίσος με

$$\mathbf{A}_t(x_1,x_2) = \begin{bmatrix}
 & (0,0) & (1,0) & (2,0) & \mathbf{L} & (k-2,0) & (k-1,0) & (0,1) & (0,2) & \mathbf{L} & (0,r-2) & (0,r-1) \\
(0,0) & 0 & p & 0 & \mathbf{L} & 0 & 0 & q & 0 & \mathbf{L} & 0 & 0 \\
(1,0) & 0 & 0 & p & \mathbf{L} & 0 & 0 & q & 0 & \mathbf{L} & 0 & 0 \\
(2,0) & 0 & 0 & 0 & \mathbf{L} & 0 & 0 & q & 0 & \mathbf{L} & 0 & 0 \\
\mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} & \mathbf{M} \\
(k-2,0) & 0 & 0 & 0 & \mathbf{L} & 0 & p & q & 0 & \mathbf{L} & 0 & 0 \\
(k-1,0) & 0 & 0 & 0 & \mathbf{L} & 0 & 0 & q & 0 & \mathbf{L} & 0 & 0 \\
(0,1) & 0 & p & 0 & \mathbf{L} & 0 & 0 & 0 & q & \mathbf{L} & 0 & 0 \\
(0,2) & 0 & p & 0 & \mathbf{L} & 0 & 0 & 0 & 0 & \mathbf{L} & 0 & 0 \\
\mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} & 0 & \mathbf{M} & \mathbf{M} & \mathbf{L} & \mathbf{M} & \mathbf{M} \\
(0,r-2) & 0 & p & 0 & \mathbf{L} & 0 & 0 & 0 & & \mathbf{L} & 0 & q \\
(0,r-1) & 0 & p & 0 & \mathbf{L} & 0 & 0 & 0 & 0 & \mathbf{L} & 0 & 0
\end{bmatrix}$$

και τους «εξωτερικούς» πίνακες μετάβασης: $\mathbf{A}_{t,1}^{(1)}$ με όλα τα στοιχεία ίσα με 0 πλην του $(k+1,k+1)$, $\mathbf{A}_{t,k}^{(1)}$ με όλα τα στοιχεία ίσα με 0 πλην του $(k,k+1)$, στοιχείου και $\mathbf{A}_{t,1}^{(2)}$ με όλα τα στοιχεία ίσα με 0 πλην του $(k+r,1)$ στοιχείου.

Τέλος, με χρήση του Θεωρήματος 2.10 είναι δυνατός ο υπολογισμός της συνάρτησης πιθανότητας της $(S_{n,k}^{(S)}, N_{n,r}^{(F)})$.

Η διπλή γεννήτρια συνάρτηση της $(S_{n,k}^{(S)}, N_{n,r}^{(F)})$, στην περίπτωση που ισχύει $\mathbf{A}_{t,0}(x_1,x_2) = \mathbf{A}_{t,0}$, $\mathbf{A}_{t,i}^{(1)}(x_1,x_2) = \mathbf{A}_{t,i}^{(1)}$, $i = 1,2,...,m_1$, και επιπλέον $\mathbf{A}_{t,i}^{(2)}(x_1,x_2) = \mathbf{A}_{t,i}^{(2)}$, $i = 1,2,...,m_2$, για κάθε (x_1,x_2), δίνεται στο Πόρισμα 4.1.

Πόρισμα 4.1: Η διανυσματική γεννήτρια συνάρτηση των διανυσμάτων $\mathbf{f}_t(x_1, x_2)$ δίνεται από τον τύπο

$$\boldsymbol{\varphi}_t(z_1, z_2) = \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty}\mathbf{f}_t(x_1, x_2)z_1^{x_1}z_2^{x_2} = \boldsymbol{\varphi}_0(z_1, z_2)\prod_{r=1}^{t}\left(\mathbf{A}_{r,0} + \mathbf{A}_{r,1}^{(1)}z_1^1 + \mathbf{A}_{r,k}^{(1)}z_1^k + \mathbf{A}_{r,1}^{(2)}z_2^1\right)$$

όπου

$$\boldsymbol{\varphi}_0(z_1, z_2) = \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty}\mathbf{f}_0(x_1, x_2)z_1^{x_1}z_2^{x_2}\ .$$

Απόδειξη: Από το Θεώρημα 2.12 γνωρίζουμε ότι η

$$\boldsymbol{\varphi}_t(z_1, z_2) = \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty}\mathbf{f}_t(x_1, x_2)z_1^{x_1}z_2^{x_2} = \boldsymbol{\varphi}_0(z_1, z_2)\prod_{r=1}^{t}\left(\mathbf{A}_{r,0} + \sum_{i=1}^{m_1}\mathbf{A}_{r,i}^{(1)}z_1^i + \sum_{i=1}^{m_2}\mathbf{A}_{r,i}^{(2)}z_2^i\right)$$

και κρατώντας μόνο τους παράγοντες που μας ενδιαφέρουν προκύπτει το ζητούμενο. □

Η διπλή γεννήτρια συνάρτηση της $(S_{n,k}^{(S)}, N_{n,r}^{(F)})$ όταν επιπλέον οι πίνακες $\mathbf{A}_{t,0}$, $\mathbf{A}_{t,i}^{(1)}$, $i = 1,2,...,m_1$, και $\mathbf{A}_{t,i}^{(2)}$, $i = 1,2,...,m_2$ (όπου m_1, m_2 δύο θετικοί ακέραιοι) δεν εξαρτώνται από τον χρόνο *t*, δίνεται στο Πόρισμα 4.2.

Πόρισμα 4.2: Η διπλή (διανυσματική) γεννήτρια συνάρτηση των διανυσμάτων $\mathbf{f}_t(x_1, x_2)$ δίνεται από τον τύπο

$$\boldsymbol{\Phi}(z_1, z_2; w) = \sum_{t=0}^{\infty} j_t(z_1, z_2)w^t = \frac{P(z_1, z_2; w)}{Q(z_1, z_2; w)},$$

με

$$\begin{aligned}P(z_1, z_2; w) = {} & 1 - (pw)z_1 - (pw)^k(1 - z_1^k) - (pw)^{k+1}(z_1^k - z_1) - (qw)^r + (qw)^r(pw)z_1 \\ & + (pw)^k(qw)^r(1 - z_1^k) + (pw)^{k+1}(qw)^r(z_1^k - z_1)\end{aligned}$$

και

$$\begin{aligned}Q(z_1, z_2; w) = {} & 1 - w(1 + pz_1) + w^2pz_1 - (qw)^r z_2 + (qw)^r w(p(1 + z_1z_2) + qz_2) \\ & - (qw)^r(pw)wz_1(p + qz_2) + (pw)^k(qw)(1 - z_1^k) + (pw)^{k+1}(qw)(z_1^k - z_1) \\ & - (pw)^k(qw)^r(1 - z_2)(1 - z_1^k) \\ & - (pw)^k(qw)^r w\left[\left(qz^2(1 - z_1^k)\right) + p(z_1^k - z_1)(1 - z_2)\right] \\ & - (pw)^{k+1}(qw)^{r+1}(z_1^k - z_1)z_2\end{aligned}$$

Αριθμητή και παρονομαστή, αντίστοιχα.

Απόδειξη: Από το Θεώρημα 2.12 έχουμε

$$\begin{aligned}\mathbf{\Phi}(z_1, z_2; w) &= \sum_{t=0}^{\infty} j_t(z_1, z_2) w^t \\ &= \boldsymbol{\pi}_{0,0} \sum_{t=0}^{\infty} ((\mathbf{A}_0 + \mathbf{A}_1^{(1)} z_1 + \mathbf{A}_k^{(1)} z_1^k + + \mathbf{A}_1^{(2)} z_2) w)^t \\ &= \boldsymbol{\pi}_{0,0} (I - w(\mathbf{A}_0 + \mathbf{A}_1^{(1)} z_1 + \mathbf{A}_k^{(1)} z_1^k + + \mathbf{A}_1^{(2)} z_2))^{-1}.\end{aligned}$$

Γνωρίζοντας ότι $\Phi(z_1, z_2; w) = \mathbf{\Phi}(z_1, z_2; w)\mathbf{1}'$ και ύστερα από αλγεβρικές πράξεις καταλήγουμε

$$\Phi(z_1, z_2; w) = \frac{1 + (pw)P_1(z_1; w) + (qw)P_2(z_2; w) + (pw)(qw)P_1(z_1; w)P_2(z_2; w)}{1 - (pw)(qw)P_1(z_1; w)P_2(z_2; w)}$$

με

$$P_1(z_1; w) = \frac{(1 - pwz_1)(1 - (pw)^{k-1}) + (1 - pw)z_1(pwz_1)^{k-1}}{(1 - pw)(1 - pwz_1)}$$

και

$$P_2(z_2; w) = \frac{1 - (qw)^{r-1}(1 - z2 + qwz_2)}{(1 - qw)(1 - z_2(qw)^r)}$$

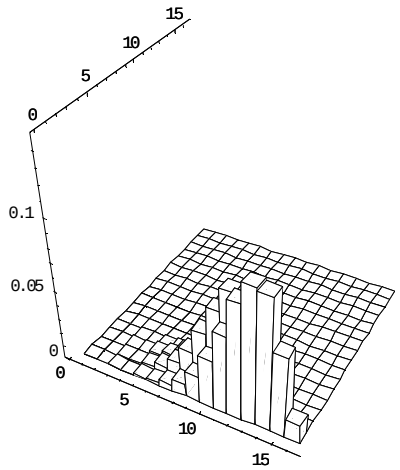
και ύστερα από αλγεβρικές πράξεις καταλήγουμε στη μορφή του πηλίκου πολυωνύμων.

□

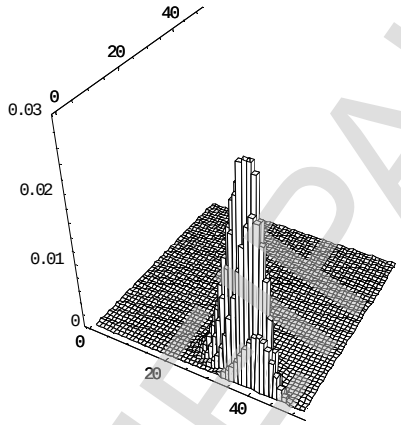
Στο Σχήμα 4.2 δίνεται η συνάρτηση πιθανότητας της διδιάστατης τυχαίας μεταβλητής $(S_{n,k}^{(S)}, N_{n,r}^{(F)})$ για διάφορες τιμές των n, k, r, p, q, ενώ, στον Πίνακα 4.1 δίνονται οι συσχετίσεις των $(S_{n,k}^{(S)}, N_{n,r}^{(F)})$ για διάφορες τιμές των n, k, r, p, q.

Χαρακτηριστικό είναι ότι οι δύο τυχαίες μεταβλητές παρουσιάζουν έντονη αρνητική συσχέτιση.

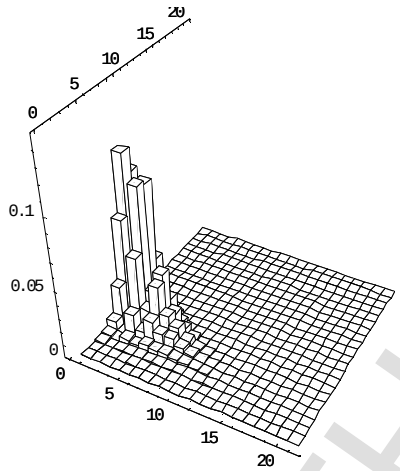
Σχήμα 4.2: Η συνάρτηση πιθανότητας της $(S_{n,k}^{(S)}, N_{n,r}^{(F)})$ για διάφορα n, k, r, p, q



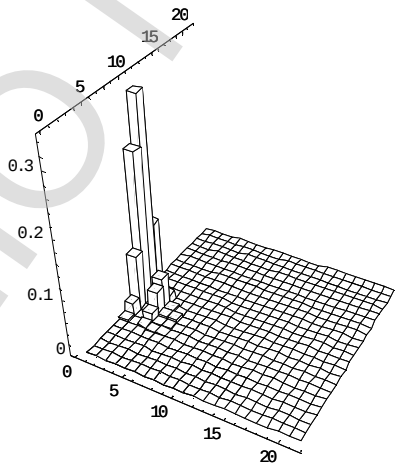
$n = 15,\ p = 0.75, k = 2, r = 2$



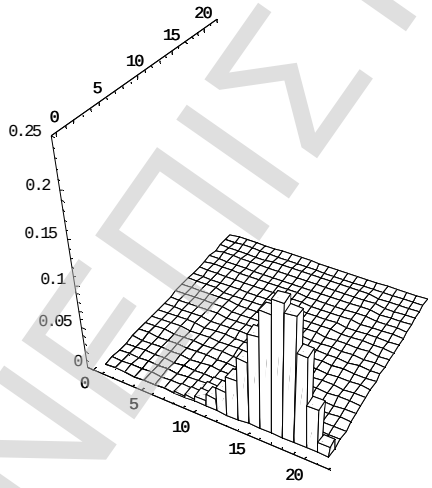
$n = 50,\ p = 0.75, k = 2, r = 2$



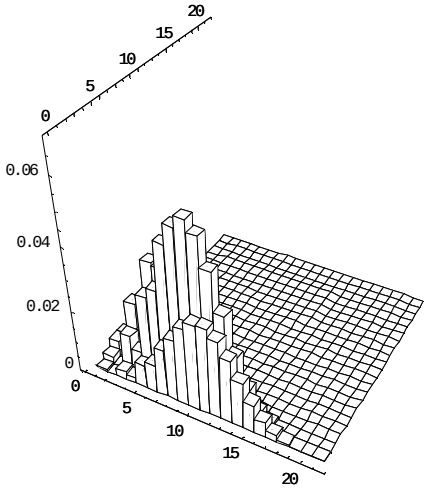
$n = 20,\ p = 0.25, k = 2, r = 2$



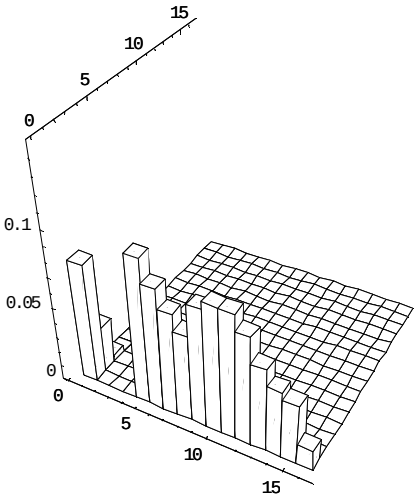
$n = 20,\ p = 0.15, k = 2, r = 2$



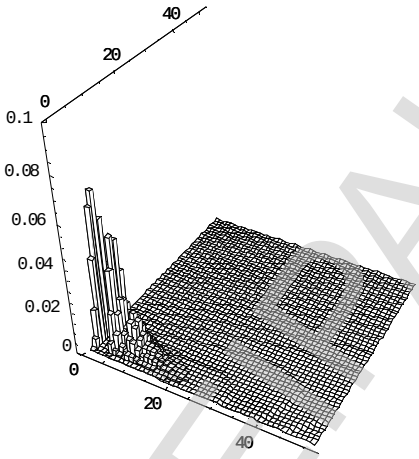
$n = 20,\ p = 0.8, k = 2, r = 3$



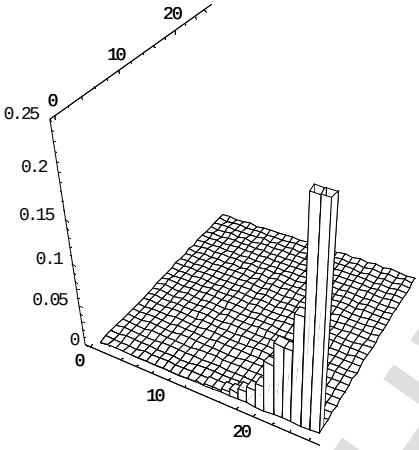
$n = 20,\ p = 0.5, k = 2, r = 3$



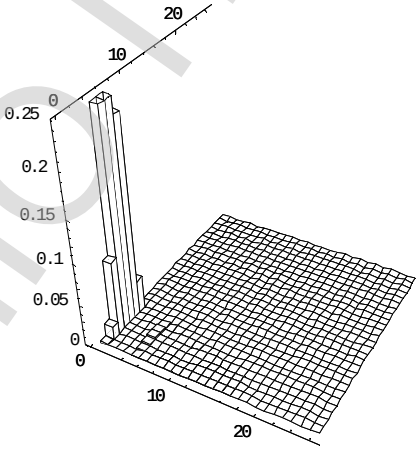
$n = 15,\ p = 0.75, k = 4, r = 3$



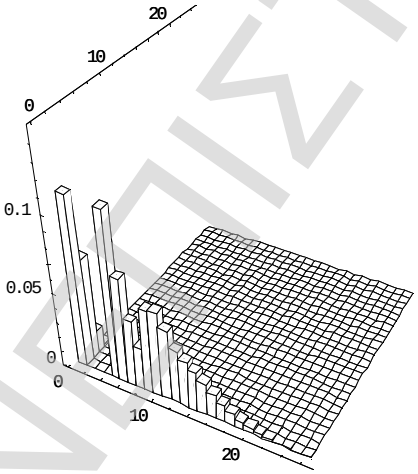
$n = 50,\ p = 0.45, k = 4, r = 3$



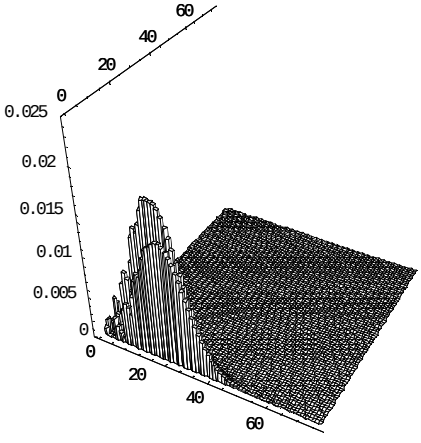
$n = 25,\ p = 0.95, k = 4, r = 4$



$n = 25,\ p = 0.15, k = 4, r = 4$



$n = 25,\ p = 0.60, k = 4, r = 4$



$n = 75,\ p = 0.60, k = 4, r = 4$

Πίνακας 4.1: Συντελεστές συσχέτισης των $(S_{n,k}^{(S)}, N_{n,r}^{(F)})$ για διάφορα n, k, r, p

n	k	r	p	$Corr(S_{n,k}^{(S)}, N_{n,r}^{(F)})$
30	2	2	0,25	-0,575472
30	2	2	0,50	-0,645813
30	2	3	0,25	-0,486567
30	2	3	0,50	-0,510352
30	2	3	0,75	-0,394063
30	2	3	0,95	-0,119666
30	4	3	0,25	-0,176318
30	4	3	0,50	-0,292866
30	4	3	0,75	-0,260741
30	4	3	0,95	-0,083848
50	4	3	0,50	-0,297753
100	4	3	0,50	-0,305279
100	2	2	0,50	-0,653370
100	2	2	0,95	-0,384958

4.4. Ανακεφαλαίωση

Στο Κεφάλαιο αυτό κάναμε μια αναλυτική παρουσίαση προβλημάτων σχετικών με ροές επιτυχιών τα οποία είναι δυνατόν να λυθούν με χρήση της τεχνικής της Μαρκοβιανής εμφύτευσης πολυδιάστατων τυχαίων μεταβλητών. Φυσικά, με χρήση της τεχνικής της Μαρκοβιανής εμφύτευσης πολυδιάστατων τυχαίων μεταβλητών δίνεται η δυνατότητα να λυθούν προβλήματα πολύ γενικότερα, δηλαδή τα προβλήματα που παρουσιάστηκαν αποτελούν ένα ελάχιστο υποσύνολο.

ΚΕΦΑΛΑΙΟ 5: ΚΑΤΑΝΟΜΕΣ ΧΡΟΝΩΝ ΑΝΑΜΟΝΗΣ ΣΧΕΤΙΚΕΣ ΜΕ ΡΟΕΣ ΕΠΙΤΥΧΙΩΝ

5.1. Εισαγωγή

Γενικά, το πρόβλημα του χρόνου αναμονής μέχρι την εμφάνιση ενός σχηματισμού συμβόλων σε μια ακολουθία πειραμάτων είναι συνυφασμένο με τη θεωρία των ροών. Η θεωρητική μελέτη της κατανομής του χρόνου αναμονής ροών επιτυχιών ξεκινά από τον Feller (1968) μέσω της θεωρίας των ανανεωτικών ενδεχόμενων (*recurrent events*).

Ανατρέχοντας στη βιβλιογραφία, μπορούμε να διαπιστώσουμε ότι η κατανομή χρόνων αναμονής μέχρι να συμβεί ένα ενδεχόμενο στη γενική της μορφή έχει απασχολήσει πολλούς ερευνητές (Philippou and Muwafi (1982), Philippou et al. (1983), Uppuluri and Patil (1983), Ebneshahrashoob and Sobel (1990), Aki and Hirano (1993), Balasubramanian et al (1993), Aki (1992), Ling and Low (1993), Chryssaphinou et al. (1994), Uchida and Aki (1995), Koutras (1996b, 1997a, 1997b), Koutras and Alexandrou (1997), Antzoulakos (1999), Antzoulakos (2001), και άλλοι.

Έτσι, στο Κεφάλαιο 5, μετά από μια σύντομη περιγραφή γενικών προβλημάτων χρόνων αναμονής, η μελέτη εστιάζεται στη τυχαία μεταβλητή T_r η οποία συμβολίζει τον χρόνο αναμονής μέχρι το άθροισμα των μηκών των ροών που εμφανίσθηκαν να είναι ίσο με r (βλέπε επίσης Antzoulakos et al. (2004)).

Η τυχαία μεταβλητή αυτή συνδέεται με τις τυχαίες μεταβλητές εμφυτεύσιμες σε Μαρκοβιανή αλυσίδα πολυωνυμικού τύπου (MVP), οι οποίες εισήχθησαν στα Κεφάλαια 2 και 3.

Η μελέτη της κατανομής της τυχαίας μεταβλητής T_r καθιστά δυνατή την ανάπτυξη νέων μεθοδολογιών στον τομέα του στατιστικού ποιοτικού ελέγχου και ειδικότερα στην

δειγματοληψία αποδοχής. Μέσω της εμφύτευσης σε Μαρκοβιανή αλυσίδα, εξάγονται απλοί τύποι για τις πιθανότητες και τη γεννήτρια πιθανοτήτων του χρόνου αναμονής για ισόνομες, μη ισόνομες και Μαρκοβιανά εξαρτημένες δοκιμές.

5.2. Εμφύτευση Τυχαίων Μεταβλητών σχετικών με Χρόνους Αναμονής, Ροών Επιτυχιών σε Μαρκοβιανή Αλυσίδα

Στην παρούσα παράγραφο ασχολούμαστε με την εμφύτευση σε Μαρκοβιανή αλυσίδα τυχαίων μεταβλητών, που αφορούν τον χρόνο αναμονής μέχρι την εμφάνιση της $1^{ης}$ ή της r-οστής ροής επιτυχιών (οποιουδήποτε είδους).

Αρχικά δίνουμε μερικά βασικά αποτελέσματα σχετικά με τυχαίες μεταβλητές που ανήκουν στην κλάση των τυχαίων μεταβλητών εμφυτεύσιμων σε Μαρκοβιανή αλυσίδα διωνυμικού τύπου (Koutras (1997a, 1997b) και Koutras and Alexandrou (1997b)).

Ενώ, στη συνέχεια αναπτύσσουμε τη διαδικασία εμφύτευσης τυχαίων μεταβλητών που ανήκουν στην κλάση των τυχαίων μεταβλητών εμφυτεύσιμων σε Μαρκοβιανή αλυσίδα πολυωνυμικού τύπου, καθώς και μια σειρά από αποτελέσματα σχετικά με αυτές.

Έστω $Z_1, Z_2, \mathbf{K}$ μια σειρά από ανεξάρτητες δίτιμες ή πλειότιμες δοκιμές και X_n μια τυχαία μεταβλητή που μετρά τον αριθμό των εμφανίσεων ενός απλού (ή σύνθετου) ενδεχόμενου e μεταξύ των $Z_1, Z_2, \mathbf{K}, Z_n$. Συμβολίζουμε με T_r, $r \geq 1$, τον χρόνο αναμονής για την r-οστή εμφάνιση του ενδεχόμενου που απαριθμείται με τη X_n. Αν υποθέσουμε ότι η X_n είναι μια ομογενής MVB και διατηρήσουμε το συμβολισμό που εισάχθηκε στο Κεφάλαιο 2, τότε σύμφωνα με τον Koutras (2003) ισχύει το ακόλουθο αποτέλεσμα.

Θεώρημα 5.1: *Η συνάρτηση πιθανότητας της T_r δίνεται από την σχέση*

$$\Pr(T_r = n) = \mathbf{f}_{n-1}(r-1)\mathbf{B1}', \; n = 1, 2, \ldots.$$

Το Θεώρημα 5.1 παρέχει ένα αποτελεσματικό εργαλείο για τον υπολογισμό της ακριβούς κατανομής του T_r, κάνοντας χρήση των διανυσμάτων πιθανοτήτων $\mathbf{f}_{n-1}(r-1)$, και των αθροισμάτων b_i των γραμμών του πίνακα $\mathbf{B}$, $i = 1,2,\mathbf{K},s$.

Θεώρημα 5.2: *Η διπλή γεννήτρια συνάρτηση του χρόνου αναμονής δίνεται από τον τύπο*

$$H(z,w) = wz\boldsymbol{\pi}_0[\mathbf{I} - w(\mathbf{A} + z\mathbf{B})]^{-1}\mathbf{B}\mathbf{1}'.$$

Πόρισμα 5.1: *Η γεννήτρια συνάρτηση του χρόνου αναμονής T_r δίνεται από την*

$$\sum_{n=1}^{\infty} \Pr(T_r = n)w^n = w^r\boldsymbol{\pi}_0[(\mathbf{I} - w\mathbf{A})^{-1}\mathbf{B}]^r\mathbf{1}', \; r \geq 1.$$

Πόρισμα 5.2: *Η γεννήτρια συνάρτηση του χρόνου αναμονής T_1 δίνεται από την*

$$H(w) = \sum_{n=1}^{\infty} \Pr(T_1 = n)w^n = w\boldsymbol{\pi}_0(\mathbf{I} - w\mathbf{A})^{-1}\mathbf{B}\mathbf{1}'.$$

Στην συνέχεια παρουσιάζουμε τη διαδικασία εμφύτευσης τυχαίων μεταβλητών που αφορούν χρόνους αναμονής σχετιζόμενων με τυχαίες μεταβλητές MVP.

Έστω V_n (n μη αρνητικός ακέραιος) μια μη αρνητική τυχαία μεταβλητή, η οποία λαμβάνει ακέραιες τιμές, με συνάρτηση πιθανότητας

$$f_n(v) = \Pr(V_n = v), \; n \geq 0, \; v = 0,1,\mathbf{K}$$

και

$$j_n(z) = E[z^{V_n}] = \sum_{v=0}^{\infty} \Pr(V_n = v)z^v, \; n \geq 0,$$

$$\Phi(z,w) = \sum_{n=0}^{\infty} j_n(z)w^n = \sum_{n=0}^{\infty}\sum_{v=0}^{\infty} \Pr(V_n = v)z^v w^n$$

η μονή και η αντίστοιχη διπλή πιθανογεννήτρια συνάρτηση. Θέτουμε $j_0(z) = 1$, δηλαδή θεωρούμε ότι

$$\Pr(V_0 = v) = \begin{cases} 1, & \text{αν } v = 0 \\ 0, & \text{αν } v = 1,2,\mathbf{K}. \end{cases}$$

Στη συνέχεια ας συμβολίσουμε με T_r (r μη αρνητικός ακέραιος) την τυχαία μεταβλητή του χρόνου αναμονής που ορίζεται ως $T_r = \min\{n \geq 0 : V_n \geq r\}$ και έστω ότι $h_r(n) = \Pr(T_r = n)$, $n = 0,1,\mathbf{K}$ είναι η συνάρτηση πιθανότητάς της. Η μονή και διπλή πιθανογεννήτρια συνάρτηση της T_r θα συμβολίζονται με $H_r(w)$ και $H_r(z,w)$ αντίστοιχα, δηλαδή

$$H_r(w) = E[w^{T_r}] = \sum_{n=0}^{\infty} h_r(n) w^n, \ r \geq 0$$

και

$$H_r(z,w) = \sum_{n=0}^{\infty} H_r(n) z^r = \sum_{r=0}^{\infty} \sum_{n=0}^{\infty} h_r(n) w^n z^r .$$

Σημειώνουμε ότι, με βάση τους προηγούμενους ορισμούς, μπορούμε να γράψουμε

$$\Pr(T_0 = n) = \begin{cases} 1, & \text{αν } n = 0 \\ 0, & \text{αν } n = 1,2,\mathbf{K} \end{cases}$$

από το οποίο προκύπτει ότι $H_0(w) = 1$.

Εάν η V_n ανήκει στην κλάση των τυχαίων μεταβλητών πολυωνυμικού τύπου εμφυτεύσιμων σε Μαρκοβιανή αλυσίδα, και T_r, $r \geq 1$, είναι ο χρόνος αναμονής μέχρι την πρώτη εμφάνιση του ενδεχόμενου $\{V_n \geq r\}$, τότε η συνάρτηση πιθανότητας της T_r υπολογίζεται με τη χρήση του επόμενου θεωρήματος.

Θεώρημα 5.3: *Η συνάρτηση πιθανότητας $h_r(n)$ της T_r, $r \geq 1$ δίνεται από τον τύπο*

$$h_r(n) = \sum_{i=1}^{m} \sum_{j=1}^{i} \mathbf{f}_{n-1}(r-j) \mathbf{A}_{n,i}(r-j) \mathbf{1}', \ n \geq 1.$$

Απόδειξη: Η συνάρτηση πιθανότητας $h_r(n)$ της T_r μπορεί να εκφραστεί ως

$$\Pr(T_r = n) = \sum_{i=1}^{m} \sum_{i'=1}^{m-i} \Pr(Y_n \in C_{r+i'}, Y_{n-1} \in C_{r-i})$$

$$= \sum_{i=1}^{m} \sum_{j=i}^{m} \Pr(Y_n \in C_{r-i+j}, Y_{n-1} \in C_{r-i})$$

και αντικαθιστώντας το $\Pr(Y_n \in C_{r-i+j}, Y_{n-1} \in C_{r-i})$ με το άθροισμα $\sum_{k=0}^{s-1} \Pr(Y_n \in C_{r-i+j} \mid Y_{n-1} \in c_{r-i,k}) \Pr(Y_{n-1} \in c_{r-i,k})$ παίρνουμε

$$\Pr(T_r = n) = \sum_{i=1}^{m}\sum_{j=i}^{m}\sum_{k=0}^{s-1} \Pr(Y_n \in C_{r-i+j} \mid Y_{n-1} \in c_{r-i,k}) \Pr(Y_{n-1} \in c_{r-i,k}) .$$

Παρατηρούμε στη συνέχεια ότι $\Pr(Y_{n-1} \in c_{r-i,k}) = \mathbf{f}_{n-1}(r-i)\mathbf{e}'_{k+1}$,

όπου το $\mathbf{e}_k = (0, \mathbf{K}, 1, \mathbf{K}, 0)$ συμβολίζει το $k-$οστό μοναδιαίο διάνυσμα - γραμμή του R^s

και $\Pr(Y_n \in C_{r-i+j} \mid Y_{n-1} \in c_{r-i,k}) = \sum_{l=1}^{s} \mathbf{e}_{k+1}\mathbf{A}_{n,j}(r-i)\mathbf{e}'_l = \mathbf{e}_{k+1}\mathbf{A}_{n,j}(r-i)\mathbf{1}'$.

Συνεπώς μπορούμε να γράψουμε

$$\Pr(T_r = n) = \sum_{i=1}^{m}\sum_{j=i}^{m}\sum_{k=0}^{s-1} \mathbf{e}_{k+1}\mathbf{A}_{n,j}(r-i)\mathbf{1}'\mathbf{f}_{n-1}(r-i)\mathbf{e}'_{k+1}$$

και λαμβάνοντας υπόψη ότι

$$\begin{aligned}\sum_{k=0}^{s-1} \mathbf{e}_{k+1}\mathbf{A}_{n,j}(r-i)\mathbf{1}'\mathbf{f}_{n-1}(r-i)\mathbf{e}'_{k+1} &= \sum_{k=0}^{s-1} \mathbf{f}_{n-1}(r-i)\mathbf{e}'_{k+1}\mathbf{e}_{k+1}\mathbf{A}_{n,j}(r-i)\mathbf{1}' \\ &= \mathbf{f}_{n-1}(r-i)\sum_{k=0}^{s-1} \mathbf{e}'_{k+1}\mathbf{e}_{k+1}\mathbf{A}_{n,j}(r-i)\mathbf{1}' \\ &= \mathbf{f}_{n-1}(r-i)\mathbf{I}\mathbf{A}_{n,j}(r-i)\mathbf{1}' \\ &= \mathbf{f}_{n-1}(r-i)\mathbf{A}_{n,j}(r-i)\mathbf{1}'\end{aligned}$$

(όπου $\mathbf{I}$ είναι ο μοναδιαίος $s \times s$ πίνακας), παίρνουμε

$$\Pr(T_r = n) = \sum_{i=1}^{m}\sum_{j=i}^{m} \mathbf{f}_{n-1}(r-i)\mathbf{A}_{n,j}(r-i)\mathbf{1}'$$

από το οποίο προκύπτει αμέσως το ζητούμενο. □

Αν οι εσωτερικοί και εξωτερικοί πίνακες πιθανοτήτων μετάβασης είναι ανεξάρτητοι των t, v, δηλαδή ισχύει $\mathbf{A}_{t,i}(v) = \mathbf{A}_i$ για όλα τα $t \geq 1$ και $v \geq 0$, η διπλή πιθανογεννήτρια συνάρτηση της V_n θα δίνεται από τον τύπο (βλέπε Θεώρημα 2.8)

$$\Phi(z,w) = \sum_{n=0}^{\infty} j_n(z)w^n = \boldsymbol{\pi}_0\left(\mathbf{I} - w\sum_{i=0}^{m}\mathbf{A}_i z^i\right)^{-1}\mathbf{1}' . \qquad (5.1)$$

Το επόμενο θεώρημα περιγράφει ένα ανάλογο αποτέλεσμα για τη διπλή πιθανογεννήτρια συνάρτηση του χρόνου αναμονής της τυχαίας μεταβλητής T_r.

Θεώρημα 5.4: *Αν* $\mathbf{A}_{t,i}(v) = \mathbf{A}_i$ *για όλα τα* $t \geq 1$ *και* $v \geq 0$, *η διπλή πιθανογεννήτρια συνάρτηση της* T_r *δίνεται από τον τύπο*

$$H(z,w) = 1 + w\boldsymbol{\pi}_0 \sum_{i=1}^{m} \sum_{j=1}^{i} z^j \left(\mathbf{I} - w \sum_{l=0}^{m} \mathbf{A}_l z^l \right)^{-1} \mathbf{A}_i \mathbf{1}'.$$

Απόδειξη: Η διπλή πιθανογεννήτρια συνάρτηση $H(z,w)$ της T_r μπορεί να εκφραστεί ως

$$H(z,w) = 1 + \sum_{r=1}^{\infty} \sum_{n=1}^{\infty} h_r(n) w^n z^r$$

και αντικαθιστώντας το $h_r(n)$ με τον τύπο που εξήχθη στο Θεώρημα 5.3, μπορούμε να γράψουμε

$$\begin{aligned} H(z,w) &= 1 + \sum_{r=1}^{\infty} \sum_{n=1}^{\infty} \sum_{i=1}^{m} \sum_{j=1}^{i} \mathbf{f}_{n-1}(r-j) \mathbf{A}_i \mathbf{1}' w^n z^r \\ &= 1 + \sum_{i=1}^{m} \sum_{j=1}^{i} \sum_{n=0}^{\infty} \sum_{r=j}^{\infty} \mathbf{f}_n(r-j) \mathbf{A}_i \mathbf{1}' w^{n+1} z^r \\ &= 1 + \sum_{i=1}^{m} \sum_{j=1}^{i} \sum_{n=0}^{\infty} \sum_{r=0}^{\infty} \mathbf{f}_n(r) \mathbf{A}_i \mathbf{1}' w^{n+1} z^{r+j} \\ &= 1 + \sum_{i=1}^{m} \sum_{j=1}^{i} wz^j \left(\sum_{n=0}^{\infty} \sum_{r=0}^{\infty} \mathbf{f}_n(r) z^r w^n \right) \mathbf{A}_i \mathbf{1}'. \end{aligned}$$

Ο τελικός τύπος για την $H(z,w)$ επαληθεύεται αμέσως από την ταυτότητα (βλέπε σχέση (5.1)

$$\left(\sum_{n=0}^{\infty} \sum_{r=0}^{\infty} \mathbf{f}_n(r) w^n z^r \right) \mathbf{1}' = \sum_{n=0}^{\infty} j_n(z) w^n = \boldsymbol{\pi}_0 \left(\mathbf{I} - w \sum_{l=0}^{m} \mathbf{A}_l z^l \right)^{-1} \mathbf{1}'.$$

□

5.3. Η Κατανομή του Χρόνου Αναμονής για την πρώτη φορά που το Άθροισμα των μηκών των Ροών μήκους τουλάχιστον *k* ξεπερνά μια τιμή *r*

Σ' αυτή την παράγραφο μελετούμε λεπτομερώς μια κατανομή χρόνου αναμονής που σχετίζεται με το άθροισμα των μηκών υποακολουθιών (strings) που αποτελούνται από k ή περισσότερες συνεχόμενες επιτυχίες (k μη αρνητικός ακέραιος). Αυτή η

τυχαία μεταβλητή θα μελετηθεί με χρήση της τεχνικής της εμφύτευσης τυχαίων μεταβλητών πολυωνυμικού τύπου σε Μαρκοβιανή αλυσίδα που είδαμε στην προηγούμενη παράγραφο. Έτσι, εκμεταλλευόμενοι τα αποτελέσματα της προηγούμενης παραγράφου, θα μελετήσουμε τα χαρακτηριστικά της κατανομής που μας ενδιαφέρει.

Θεωρούμε μια ακολουθία δοκιμών Bernoulli $Z_1, Z_2, \mathbf{K}$ με πιθανότητες επιτυχίας $p_t = \Pr(Z_t = 1)$ και πιθανότητες αποτυχίας $q_t = \Pr(Z_t = 0) = 1 - p_t$, $t \geq 1$ και έστω n, k δυο θετικοί ακέραιοι. Για $k \leq t \leq n$ ορίζουμε την

$$U_t = \begin{cases} k+l, & \text{αν } Z_{t-k-l+1} = Z_{t-k-l+2} = \mathbf{K} = Z_t = 1, Z_{t-k-l} = Z_{t+1} = 0 \\ 0, & \text{διαφορετικά} \end{cases}$$

όπου l είναι ένας μη αρνητικός ακέραιος (θεωρούμε: $Z_0 = Z_{n+1} = 0$). Το άθροισμα των μηκών των υποακολουθιών της ακολουθίας $Z_1, Z_2, \mathbf{K}, Z_n$ που περιέχει k ή περισσότερες συνεχόμενες επιτυχίες, μπορεί να γραφεί στη μορφή αθροίσματος τυχαίων μεταβλητών ως

$$V_n = \sum_{t=k}^{n} U_t, \ n \geq 1.$$

Είναι σαφές ότι το σύνολο τιμών της V_n (πρόκειται για την $S_{n,k}$ του Κεφαλαίου 3) είναι το $\{0\}$ αν $n \leq k$ και το $\{0, k+1, k, \mathbf{K}, n\}$ αν $n > k$. Αποδεικνύεται εύκολα ότι η V_n (βλέπε Κεφάλαιο 3) μπορεί να θεωρηθεί ως μια MVP με $\boldsymbol{\pi}_0 = (1, 0, \mathbf{K}, 0)$ και πίνακες πιθανοτήτων μεταπηδήσεων

$$\mathbf{A}_{t,0} = \begin{bmatrix} q_t & p_t & 0 & \mathbf{M} & 0 & 0 & 0 \\ q_t & 0 & p_t & \mathbf{M} & 0 & 0 & 0 \\ \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} & \mathbf{M} & \mathbf{M} \\ q_t & 0 & 0 & \mathbf{M} & 0 & p_t & 0 \\ q_t & 0 & 0 & \mathbf{M} & 0 & 0 & 0 \\ q_t & 0 & 0 & \mathbf{M} & 0 & 0 & 0 \end{bmatrix}_{(k+1)\times(k+1)}$$

$$\mathbf{A}_{t,1} = \begin{bmatrix} 0 & 0 & \mathbf{M} & 0 & 0 \\ 0 & 0 & \mathbf{M} & 0 & 0 \\ \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} & \mathbf{M} \\ 0 & 0 & \mathbf{M} & 0 & 0 \\ 0 & 0 & \mathbf{M} & 0 & p_t \end{bmatrix}_{(k+1)\times(k+1)}, \quad \mathbf{A}_{t,k} = \begin{bmatrix} 0 & 0 & \mathbf{M} & 0 & 0 \\ 0 & 0 & \mathbf{M} & 0 & 0 \\ \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} & \mathbf{M} \\ 0 & 0 & \mathbf{M} & 0 & p_t \\ 0 & 0 & \mathbf{M} & 0 & 0 \end{bmatrix}_{(k+1)\times(k+1)}$$

(οι πίνακες $\mathbf{A}_{t,2}, \mathbf{A}_{t,3}, \mathbf{K}, \mathbf{A}_{t,k-1}$ έχουν όλα τους τα στοιχεία ίσα με μηδέν). Η συνάρτηση πιθανότητας της V_n μπορεί εύκολα να βρεθεί μέσω του τύπου

$$\Pr(V_n = v) = \mathbf{f}_n(v)\mathbf{1}', \; n \geq 0, \; v \geq 0$$

και της αναδρομικής σχέσης

$$\mathbf{f}_n(v) = \mathbf{f}_{n-1}(v)\mathbf{A}_{n,0} + \mathbf{f}_{n-1}(v-1)\mathbf{A}_{n,1} + \mathbf{f}_{n-k}(v-k)\mathbf{A}_{n,k}, \; n \geq 1, \; v \geq 0.$$

Θεωρούμε στη συνέχεια την τυχαία μεταβλητή T_r, $r \geq 0$ που ορίζεται ως $T_r = \min\{n \geq 0 : V_n \geq r\}$. Η τυχαία μεταβλητή T_r συμβολίζει το χρόνο αναμονής μέχρι τη δοκιμή που για πρώτη φορά το άθροισμα των μηκών των μερών της ακολουθίας $Z_1, Z_2, \mathbf{K}$ που περιέχουν k ή περισσότερες συνεχόμενες επιτυχίες είναι ίσο ή υπερβαίνει το r. Στην ουσία, η T_r περιγράφει τον «χρόνο αναμονής μέχρι η τυχαία μεταβλητή V_n να ισούται ή να υπερβαίνει την τιμή r για πρώτη φορά». Οι παραπάνω ορισμοί εξηγούνται καλύτερα με το παρακάτω παράδειγμα.

Παράδειγμα 5.1: Έστω ότι έχουμε την παρακάτω ακολουθία από 30 αποτελέσματα δοκιμών Bernoulli (οι επιτυχίες συμβολίζονται με $\oplus$ ενώ οι αποτυχίες με $\otimes$). Για $k = 3$ έχουμε ότι $V_{30} = 12$, το οποίο σημαίνει ότι το άθροισμα των μηκών μιας ακολουθίας που αποτελείται από $k = 3$ ή περισσότερες συνεχόμενες επιτυχίες ισούται με 12.

6 7 8 64 7 48 6 4 7 4 8

$\oplus\otimes\oplus\otimes\otimes\otimes\oplus\oplus\oplus\otimes\oplus\oplus\oplus\oplus\otimes\oplus\otimes\otimes\oplus\oplus\oplus\oplus\oplus\otimes\otimes\oplus\oplus\otimes\oplus\oplus$

Είναι εύκολο να ελεγχθεί ότι $T_1 = T_2 = T_3 = 9$, $T_4 = T_5 = T_6 = 13$, $T_7 = 14$, $T_8 = T_9 = T_{10} = 21$, $T_{11} = 22$ και $T_{12} = 23$.

Η γενική προσέγγιση που παρουσιάστηκε εύκολα μπορεί να δώσει λύση στο ειδικό πρόβλημα που παρουσιάστηκε σε αυτή την παράγραφο. Ως εκ τούτου αποτελεί ένα λειτουργικό πλαίσιο για τη μελέτη της κατανομής των T_r. Η παρακάτω πρόταση δίνει έναν τύπο για την συνάρτηση πιθανότητας της T_r.

Πρόταση 5.3: *Η συνάρτηση πιθανότητας $h_r(n)$ της T_r, $r \geq 1$ δίνεται από τον τύπο*

$$h_r(n) = \Pr(T_r = n) = \mathbf{f}_{n-1}(r-i)\, p_n \mathbf{e}'_{k+1} + \sum_{i=1}^{k} \mathbf{f}_{n-1}(r-i)\, p_n \mathbf{e}'_{k+1}.$$

Απόδειξη: Άμεση εφαρμογή του Θεωρήματος 5.3 δίνει ότι

$$h_r(n) = \mathbf{f}_{n-1}(r-i)\mathbf{A}_{n,1}\mathbf{1}' + \sum_{i=1}^{k} \mathbf{f}_{n-1}(r-i)\mathbf{A}_{n,k}\mathbf{1}',\ n \geq 1$$

και η απόδειξη της πρότασης προκύπτει εύκολα. □

Από αυτό το σημείο και πέρα, υποθέτουμε ότι η ακολουθία $Z_1, Z_2, \mathbf{K}$ αποτελείται από ισόνομες και ανεξάρτητες δοκιμές Bernoulli, δηλαδή $p_t = p$ και $q_t = q$ για $t = 1,2,\mathbf{K}$. Στην περίπτωση αυτή οι πίνακες, $\mathbf{A}_{t,i}$, $i = 0,1,2,\mathbf{K}\,k$ δεν εξαρτώνται από το t, ισχύει δηλαδή $\mathbf{A}_{t,i} = \mathbf{A}_i$ και η μορφή τους μπορεί εύκολα να βρεθεί από τη γενική μορφή αντικαθιστώντας όλα τα p_t με p και όλα τα q_t με q. Η επόμενη πρόταση παρέχει ένα τύπο για τον υπολογισμό της διπλής πιθανογεννήτριας συνάρτησης της T_r.

Πρόταση 5.4: *Η διπλή πιθανογεννήτρια συνάρτηση* $H(\mathbf{z}, w)$ *της* T_r *μπορεί να εκφραστεί ως*

$$H(z,w) = 1 + wzp_0[\mathbf{I} - w(\mathbf{A}_0 + z\mathbf{A}_1 + z^k\mathbf{A}_k)]^{-1} p\mathbf{e}_k' + wz\left(\frac{z^k - 1}{z-1}\right) p_0[\mathbf{I} - w(\mathbf{A}_0 + z\mathbf{A}_1 + z^k\mathbf{A}_k)]^{-1} p\mathbf{e}_{k-1}'.$$

Απόδειξη: Η απόδειξη γίνεται άμεσα από το Θεώρημα 5.4 λαμβάνοντας υπόψη ότι $\mathbf{A}_i = \mathbf{0}$ για $i = 2,3,\mathbf{K}\,k-1$. □

Το αποτέλεσμα της Πρότασης 5.4 μπορεί να χρησιμοποιηθεί για να εκφράσουμε την $H(z,w)$ ως πηλίκο δυο πολυωνύμων. Για να το πετύχουμε αυτό θα χρησιμοποιήσουμε την πιθανογεννήτρια συνάρτηση μιας τυχαίας μεταβλητής που σχετίζεται με ροές και πιο συγκεκριμένα του χρόνου αναμονής X_k μέχρι την πρώτη εμφάνιση k συνεχόμενων επιτυχιών (ροές επιτυχίας μήκους k) σε μια ακολουθία δοκιμών Bernoulli $Z_1, Z_2, \mathbf{K}$. Η τυχαία μεταβλητή X_k μπορεί επίσημα να οριστεί ως $X_k = \min\{n : Z_{n-k+1} = \mathbf{K} = Z_n = 1\}$ και η πιθανογεννήτρια συνάρτησή της δίνεται από την

$$E[w^{X_k}] = G(w) = \sum_{n=0}^{\infty} g(n)w^n = \frac{(wp)^k(1-wp)}{1-w+(wq)(wp)^k}.$$

Η κατανομή της X_k είναι η γεωμετρική κατανομή τάξης k, στην οποία έχουμε ήδη αναφερθεί, η οποία έχει μελετηθεί εκτεταμένα τις δυο τελευταίες δεκαετίες (βλέπε Philippou and Muwafi (1982), Philippou, Georgiou and Philippou (1983), Hahn and Gage (1983), Aki et al. (1984), Uppuluri and Patil (1983)). Για περισσότερες πληροφορίες σχετικά με αυτή την κατανομή ο ενδιαφερόμενος αναγνώστης παραπέμπεται στους Balakrishnan and Koutras (2002).

Το παρακάτω πόρισμα δείχνει ότι η διπλή πιθανογεννήτρια συνάρτηση $H(z,w)$ μπορεί να εκφραστεί μέσω της πιθανογεννήτριας συνάρτησης $G(w)$ της X_k.

Πρόταση 5.5: *Η διπλή πιθανογεννήτρια συνάρτηση $H(z,w)$ της T_r μπορεί να γραφεί ως*

$$H(z,w)=\frac{1+z(G(w)-wp)+(1-wp)G(w)\sum_{i=2}^{k-1}z^i+(1-w)G(w)z^k}{1-z(wp)-z^k(wq)G(w)}$$

Απόδειξη: Η απόδειξη απορρέει από την Πρόταση 5.4 αντιστρέφοντας τον πίνακα $[\mathbf{I}-w(\mathbf{A}_0+z\mathbf{A}_1+z^k\mathbf{A}_k)]$. Αξίζει να σημειώσουμε ότι για τον υπολογισμό χρειάζονται μόνο ορισμένα στοιχεία του αντίστροφου πίνακα. □

Το αποτέλεσμα του Πορίσματος 5.5 μπορεί να χρησιμοποιηθεί για την εξαγωγή ενός επαναληπτικού σχήματος για τον υπολογισμό της απλής γεννήτριας συνάρτησης $H_r(w)$ της T_r.

Πρόταση 5.6: *Η πιθανογεννήτρια συνάρτηση $H_r(w)$ της T_r ικανοποιεί το επαναληπτικό σχήμα*

$$H_r(w)=(wp)H_{r-1}(w)+(wq)G(w)H_{r-k}(w),\ r\geq k+1$$

με αρχικές συνθήκες $H_i(w)=G(w)$ για $1\leq i\leq k$.

Απόδειξη: Χρησιμοποιώντας την Πρόταση 5.5 μπορούμε να γράψουμε

$$(1-z(wp)-z^k(wq)G(w))\sum_{r=0}^{\infty}H_r(w)z^r$$
$$=1+z(G(w)-wp)+(1-wp)G(w)\sum_{i=2}^{k-1}z^i+(1-w)G(w)z^k.$$

Το επιθυμητό αποτέλεσμα απορρέει εύκολα εξισώνοντας τους συντελεστές των z^r και

στα δυο μέρη της παραπάνω ισότητας. □

Αξίζει να σημειώσουμε ότι για $1 \le r \le k$, έχουμε $H_r(w) = G(w)$ και η κατανομή της T_r, $r = 1,2,\mathbf{K},k$, συμπίπτει με αυτή της X_k. Η επαναλαμβανόμενη εφαρμογή του επαναληπτικού σχήματος της Πρότασης 5.6, για $k+1 \le r \le 2k$, δίνει

$$H_r(w) = G(w)\left((wq)G(w)\sum_{i=0}^{r-k-1}(wp)^i + (wp)^{r-k} \right).$$

Η παραπάνω ταυτότητα αποκαλύπτει ότι η $H_{k+1}(w)$ συμπίπτει με την πιθανογεννήτρια συνάρτηση του χρόνου αναμονής για τη δεύτερη επικαλυπτόμενη ροή επιτυχίας μήκους k (Hirano et al (1991) ή Balakrishnan and Koutras (2002)).

Η παραπάνω σχέση μπορεί επίσης να αποδειχθεί άμεσα αν παρατηρήσουμε ότι το ενδεχόμενο $\{T_r = n\}$ για $k+1 \le r \le 2k$ αντιστοιχεί σε αποτελέσματα του χώρου δειγματοληψίας της μορφής

$$\underbrace{\mathbf{K}11\mathbf{K}111\mathbf{K}1}_{X_k} \text{ ή } \underbrace{\mathbf{K}11\mathbf{K}111\mathbf{K}10\mathbf{K}11\mathbf{K}1}_{X_k},\ 0 \le i \le r-k-1.$$

Μια άλλη ενδιαφέρουσα συνέπεια της Πρότασης 5.6 είναι ότι για $r \ge k+1$ οι χρόνοι αναμονής T_r ικανοποιούν το ακόλουθο επαναληπτικό σχήμα

$$T_r = \begin{cases} T_{r-1} + 1, & \text{με πιθανότητα } p \\ T_{r-k} + 1 + X_k, & \text{με πιθανότητα } q \end{cases} \qquad (5.2)$$

όπου T_{r-k} και X_k είναι ανεξάρτητες τυχαίες μεταβλητές.

Στη συνέχεια θα εκμεταλλευτούμε το αποτέλεσμα της Πρότασης 5.6 για να αποδείξουμε μια επαναληπτική σχέση για τη συνάρτηση πιθανότητας της T_r.

Πρόταση 5.7: *Η συνάρτηση πιθανότητας $h_r(n)$ της T_r ικανοποιεί το επαναληπτικό σχήμα*

$$\begin{aligned} h_r(n) = h_r(n-1) &+ p[h_{r-1}(n-1) - h_{r-1}(n-2)] \\ &- qp^k\left[h_r(n-k-1) - h_{r-k}(n-k-1)\right] \\ &- qp^{k+1}[h_{r-1}(n-k-2) - h_{r-k}(n-k-2)], \ r \ge k+1 \end{aligned}$$

με αρχικές συνθήκες $h_r(n) = g(n) = \Pr(X_k = n)$ για $1 \le r \le k$.

Απόδειξη: Η σχέση της Πρότασης 5.6 μπορεί ισοδύναμα να γραφεί ως

$$(1-w+w^{k+1}qp^k)\sum_{n=0}^{\infty}h_r(n)w^n =$$

$$(wp)(1-w+w^{k+1}qp^{k+1})\sum_{n=0}^{\infty}h_{r-1}(n)w^n+(wq)(wp)^k(1-w)\sum_{n=0}^{\infty}h_{r-k}(n)w^n$$

και εξισώνοντας τους συντελεστές των w^n και στα δυο μέρη της παραπάνω σχέσης μπορούμε εύκολα να δείξουμε το παραπάνω επαναληπτικό σχήμα. Οι αρχικές συνθήκες απορρέουν αμέσως από το γεγονός ότι $H_r(w)=G(w)$ για $1\le r\le k$. □

Έχοντας ένα επαναληπτικό σχήμα για τη συνάρτηση πιθανότητας της T_r δεν είναι δύσκολο να διαμορφώσουμε ένα επαναληπτικό σχήμα για την αθροιστική συνάρτηση κατανομής ή τις ουρές (*tail probabilities*) της ίδιας τυχαίας μεταβλητής.

Πρόταση 5.8: *Η συνάρτηση πιθανότητας ουράς* $\bar{h}_r(n)=\Pr(T_r>n)$ *της* T_r *ικανοποιεί την επαναληπτική σχέση*

$$\begin{aligned}\bar{h}_r(n) &= 2\bar{h}_r(n-1)-\bar{h}_r(n-2)-p[2\bar{h}_{r-1}(n-2)-\bar{h}_{r-1}(n-1)-\bar{h}_{r-1}(n-3)]\\ &+qp^k[\bar{h}_r(n-k-2)-\bar{h}_r(n-k-1)]+qp^k[\bar{h}_{r-k}(n-k-1)-\bar{h}_{r-k}(n-k-2)]\\ &-qp^{k+1}[\bar{h}_{r-1}(n-k-3)-\bar{h}_{r-1}(n-k-2)]-qp^{k+1}[\bar{h}_{r-k}(n-k-2)-\bar{h}_{r-k}(n-k-\end{aligned}$$

για $r\ge k+1$ *και με αρχικές συνθήκες* $\bar{h}_r(n)=\bar{g}(n)=\Pr(X_k>n)$ *για* $1\le r\le k$.

Απόδειξη: Η απόδειξη μπορεί να γίνει είτε από το αποτέλεσμα της Πρότασης 5.7 αντικαθιστώντας το $h_r(n)$ με το $\bar{h}_r(n-1)-\bar{h}_r(n)$ είτε από το αποτέλεσμα της Πρότασης 5.6 παρατηρώντας ότι η γεννήτρια συνάρτηση της $\bar{h}_r(n)$ μπορεί να γραφεί μέσω της $H_r(w)$ στη μορφή $\sum_{n=0}^{\infty}\bar{h}_r(n)w^n=\dfrac{1-H_r(w)}{1-w}$. □

Η τελευταία πρόταση είναι αρκετά χρήσιμη για τη μελέτη της λειτουργικής χαρακτηριστικής καμπύλης διάφορων δειγματοληπτικών σχεδίων. Περισσότερα για την δειγματοληψία αποδοχής δίνονται στο Κεφάλαιο 6. Αν και δεν φαίνεται δυνατόν να εξαχθεί μια σαφής έκφραση για τη μέση τιμή της T_r, ο αριθμητικός υπολογισμός του

$m_r = E[T_r]$ μπορεί εύκολα να γίνει με ένα επαναληπτικό σχήμα, όπως δείχνει η παρακάτω πρόταση.

Πρόταση 5.9: *Οι μέσες τιμές* $m_r = E[T_r]$, $r = 1,2,\mathbf{K}$ *των τυχαίων μεταβλητών* T_r *ικανοποιούν το επαναληπτικό σχήμα*

$$m_r = \frac{1}{p^k} + pm_{r-1} + qm_{r-k},\ r \geq k+1$$

με αρχικές συνθήκες $m_r = (1-p^k)/p^k$ *για* $1 \leq r \leq k$.

Απόδειξη: Εφαρμόζοντας τον τελεστή της μέσης τιμής στην σχέση (5.2) μπορούμε να πάρουμε την ισότητα $m_r = 1 + pm_{r-1} + q(m_{r-k} + E[X_k])$, $r \geq k+1$ και η απόδειξη ολοκληρώνεται παρατηρώντας ότι $m_r = E[X_k]$, $1 \leq r \leq k$ και ανακαλώντας το γνωστό αποτέλεσμα $E[X_k] = \frac{1-p^k}{p^k}$. □

Το Σχήμα 5.1 δίνει τη συνάρτηση πιθανότητας $h_r(n)$ της τυχαίας μεταβλητής T_r για διάφορους συνδυασμούς των παραμέτρων r, k και p.

Κλείνοντας την ενότητα σημειώνουμε ότι η έκφραση της Πρότασης 5.5 για την $H(z,w)$ μπορεί επίσης να εξαχθεί χρησιμοποιώντας τον τύπο

$$H(z,w) = \frac{z(1-w)\Phi(z,w) - 1}{z-1},$$

ο οποίος περιγράφει τη σχέση μεταξύ των διπλών γεννητριών συναρτήσεων της μεταβλητής T_r και της αντίστοιχης στατιστικής συνάρτησης απαρίθμησης V_n. Η ισότητα αυτή μπορεί να αποδειχθεί με τη βοήθεια της προφανούς σχέσης

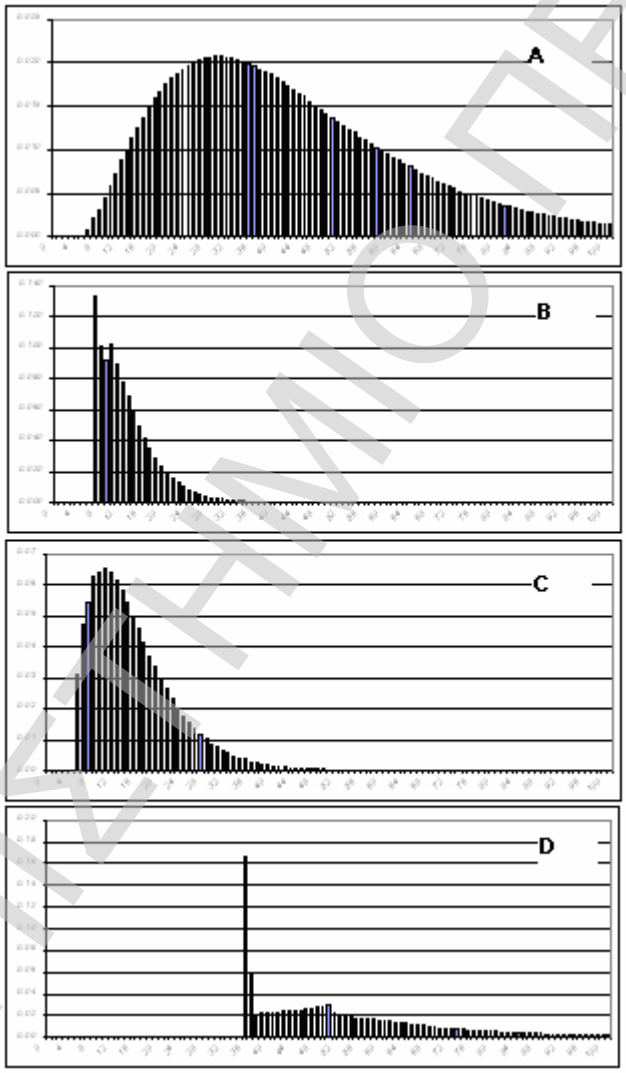
$$\Pr(T_r > n) = \Pr(V_n < r).$$

Μια ανάλογη ταυτότητα έχει αναφερθεί στον Feller (1968). Από το Κεφάλαιο 3 προκύπτει ότι $\Phi(z,w) = \frac{P_1(z,w)}{P_2(z,w)}$, με $P_1(z,w) = 1 - wpz - (wp)^k(1-z^k) - (wp)^{k+1}(z^k - z)$ και $P_2(z,w) = 1 - w(1+pz) + w^2 pz + w^{k+1}qp^k(1-z^k) + w^{k+2}qp^{k+1}(z^k - z)$.

Αντικαθιστώντας αυτές τις τελευταίες εκφράσεις στη παραπάνω σχέση μπορούμε να επανεξαγάγουμε, μετά από μακροσκελείς αλλά ευθείς υπολογισμούς, το αποτέλεσμα της Πρότασης 5.5.

Σχήμα 5.1: *Η συνάρτηση πιθανότητας* $h_r(n)$ *της* T_r *για διάφορες τιμές των παραμέτρων:*

$$A: p = 0.30, k = 2, r = 6,\ B: p = 0.30, k = 2, r = 6,$$
$$C: p = 0.50, k = 2, r = 5,\ D: p = 0.95, k = 15, r = 35$$



5.4. Ανακεφαλαίωση

Στο κεφάλαιο αυτό περιγράψαμε συνοπτικά ένα μεγάλο μέρος εκ των προβλημάτων που σχετίζονται με χρόνους αναμονής εμφάνισης ροών επιτυχιών που έχουν αναφερθεί στην διεθνή βιβλιογραφία. Επίσης στο κεφάλαιο αυτό μελετήσαμε διεξοδικά την τυχαία μεταβλητή T_r η οποία συμβολίζει τον χρόνο αναμονής μέχρι το άθροισμα των μηκών των ροών που εμφανίσθηκαν να είναι ίσο με r. Η τελευταία μπορεί να μελετηθεί με χρήση των μεθόδων και των τεχνικών που αναπτύχθηκαν στα Κεφάλαια 2 και 3 της διατριβής αυτής για την οικογένεια των τυχαίων μεταβλητών εμφυτεύσιμων σε Μαρκοβιανή αλυσίδα πολυωνυμικού τύπου (MVP).

ΚΕΦΑΛΑΙΟ 6: ΕΦΑΡΜΟΓΕΣ ΣΤΟΝ ΣΤΑΤΙΣΤΙΚΟ ΕΛΕΓΧΟ ΠΟΙΟΤΗΤΑΣ

6.1. Εισαγωγή

Στο κεφάλαιο αυτό θα παρουσιάσουμε αναλυτικά εφαρμογές της θεωρίας ροών επιτυχιών και γενικά σχηματισμών στον στατιστικό έλεγχο ποιότητας και ειδικότερα στον στατιστικό έλεγχο διεργασιών και στα διαγράμματα ελέγχου.

6.2. Στατιστικός Έλεγχος Ποιότητας

Η αγορά ή όχι ενός προϊόντος από κάποιον ενδιαφερόμενο καταναλωτή καθορίζεται κύρια από δύο παράγοντες, την ποιότητα του προϊόντος και την τιμή αυτού. Η αναγνώριση αυτών των δύο κύριων παραγόντων οι οποίοι καθορίζουν τις αγοραστικές συνήθειες των καταναλωτών, οδήγησε τις κατασκευάστριες επιχειρήσεις στην αναζήτηση νέων επιχειρηματικών τακτικών που θα οδηγήσουν στην βέλτιστη δυνατή ποιότητα των προϊόντων με το ελάχιστο δυνατό κόστος. Σημαντικό ρόλο στις τακτικές αυτές έχει ο στατιστικός έλεγχος ποιότητας. Αποτελεί την πιο παλιά και πιο ευρέως γνωστή από τις μεθόδους ελέγχου παραγωγικών διεργασιών και γενικά προαγωγής της ποιότητας του παραγόμενου προϊόντος.

Στη διεθνή βιβλιογραφία, διαχρονικά, η έννοια της ποιότητας βρήκε πολλούς και διάφορους ορισμούς. Από τη στατιστική πλευρά, μπορούμε να ορίσουμε την ποιότητα ως αντιστρόφως ανάλογη της μεταβλητότητας της ικανοποίησης που προσφέρει το προϊόν στον χρήστη του που καθορίζεται πλήρως από έναν μεγάλο αριθμό χαρακτηριστικών. Δηλαδή, βελτιώνουμε το προϊόν, όταν ελαχιστοποιούμε τη μεταβλητότητά της ικανοποίησης που προσφέρει στον χρήστη ή ελαχιστοποιώντας αντίστοιχα την μεταβλητότητα των χαρακτηριστικών τα οποία σχετίζονται με την

ικανοποίηση που προσφέρει το προϊόν. Κατ' επέκταση αυτό γίνεται ελαχιστοποιώντας αντίστοιχα την μεταβλητότητα των παραγόντων εκείνων της παραγωγικής διαδικασίας που προσδιορίζουν την μεταβλητότητα των χαρακτηριστικών που σχετίζονται με την ικανοποίηση που προσφέρει το προϊόν.

Η ιστορία της ποιότητας ξεκινά στις αρχές του 20ου αιώνα στη Μεγάλη Βρετανία. Τα βήματα στη συνέχεια είναι αλματώδη μέχρι τις μέρες μας που πλέον η ποιότητα έχει αναδειχθεί σε κύριο μέλημα τόσο της βιομηχανίας όσο και των υπηρεσιών. Σε αυτή την ανάπτυξη, μεγάλη ώθηση έδωσαν οι στατιστικές τεχνικές που εφαρμόστηκαν στις παραγωγικές διαδικασίες. Χαρακτηριστικό του ενδιαφέροντος για την ποιότητα στους στατιστικούς είναι ότι στο άρθρο του Woodall (2000) αναφέρεται ότι τα μέλη του στατιστικού τμήματος της American Society for Quality (ASQC) είναι 11000 δηλαδή περίπου το 60% των μελών ολόκληρης της American Statistical Association (ASA) που αριθμεί 18000 μέλη.

Ο στατιστικός έλεγχος ποιότητας αποτελείται από ένα σύνολο μεθόδων στατιστικής ανάλυσης. Το σύνολο αυτό μπορεί να διαχωριστεί σε τρία κύρια υποσύνολα, το καθ' ένα εκ των οποίων περιέχει στατιστικές μεθόδους προσανατολισμένες σε διαφορετικές φάσεις της παραγωγικής διαδικασίας: α) τον Σχεδιασμό και Ανάλυση Πειραμάτων (Design of Experiments), β) τον Στατιστικό Έλεγχο Διεργασιών (Statistical Process Control) και γ) τη Δειγματοληψία Αποδοχής (Acceptance Sampling). Έτσι, ο σχεδιασμός και η ανάλυση πειραμάτων εμπεριέχει όλες εκείνες τις τεχνικές οι οποίες κρίνονται απαραίτητες για τη βέλτιστη σχεδίαση της παραγωγικής διεργασίας αλλά και του τελικού προϊόντος. Ο στατιστικός έλεγχος διεργασιών εμπεριέχει τις τεχνικές εκείνες που είναι απαραίτητες για τον έλεγχο της παραγωγικής διαδικασίας κατά τη διάρκεια της παραγωγής των προϊόντων, ενώ η δειγματοληψία αποδοχής παρέχει τις τεχνικές εκείνες (κύρια αφορά δειγματοληπτικές τεχνικές) για τον έλεγχο του τελικού προϊόντος. Σημαντικότατες εφαρμογές της θεωρίας ροών επιτυχιών στον στατιστικό έλεγχο ποιότητας, βρίσκουμε, στον στατιστικό έλεγχο διεργασιών και στην δειγματοληψία αποδοχής. Στο παρόν κεφάλαιο της συγκεκριμένης διατριβής, θα παρουσιάσουμε εφαρμογές της θεωρίας ροών επιτυχιών στον στατιστικό έλεγχο διεργασιών και ιδιαίτερα στο κυριότερο εργαλείο του, τα διαγράμματα ελέγχου. Για περισσότερες πληροφορίες επάνω στον στατιστικό έλεγχο ποιότητας ο ενδιαφερόμενος αναγνώστης παραπέμπεται στον Antzoulakos (2003b).

6.3. Στατιστικός Έλεγχος Διεργασιών

Ο στατιστικός έλεγχος διεργασιών είναι μια συλλογή εργαλείων που στόχο έχει να επιτύχουμε σταθερότητα σε μια διαδικασία και να βελτιώσουμε την ικανότητά της, μειώνοντας τη μεταβλητότητα. Ο στατιστικός έλεγχος διεργασιών μπορεί να εφαρμοσθεί σε κάθε διεργασία (ακόμη και αν δεν είναι παραγωγική).

Τα επτά κυριότερα εργαλεία του (Montgomery (2000), Ryan (2000)) είναι: 1) το ιστόγραμμα ή διάγραμμα μίσχου-φύλλου (stem and leaf plot), 2) το φύλο ελέγχου, 3) το διάγραμμα Pareto, 4) το διάγραμμα αιτίας-αποτελέσματος, 5) το διάγραμμα συγκέντρωσης ελαττωματικών, 6) το διάγραμμα διασποράς (scatter plot), 7) το διάγραμμα ελέγχου. Από αυτά το ιστόγραμμα και το διάγραμμα μίσχου-φύλλου είναι τα πιο γνωστά, ενώ τα πιο χρήσιμα είναι τα διαγράμματα ελέγχου. Τα διαγράμματα ελέγχου είναι εκείνα τα εργαλεία που χρησιμοποιούνται για να διατηρήσουμε ένα χαρακτηριστικό ποιότητας υπό έλεγχο κατά την διάρκεια της παραγωγικής διεργασίας. Στην συνέχεια θα παρουσιάσουμε τα κύρια χαρακτηριστικά ενός διαγράμματος ελέγχου, πριν όμως θα δώσουμε κάποιες βασικές έννοιες που απαιτούνται για την κατανόηση των προβλημάτων που θα περιγράψουμε στην διατριβή αυτή.

Κάθε παραγωγική διαδικασία ανεξάρτητα από το πόσο καλά σχεδιασμένη είναι, έχει ένα ποσοστό φυσικής μεταβλητότητας. Δηλαδή, όσο καλορυθμισμένα και να είναι τα μηχανήματα, όσο ικανοί και να είναι οι χειριστές των μηχανημάτων, όσο ικανοποιητική και αν είναι η πρώτη ύλη, ποτέ το παραγόμενο προϊόν που παράγεται στον χρόνο *t*, δεν είναι το ίδιο με το προϊόν που παρήχθη στον χρόνο *t*-1. Αυτή η μεταβλητότητα, προέρχεται από το άθροισμα της μεταβλητότητας πολλών μικρών αιτιών, τυχαίων στην φύση τους, (για παράδειγμα εξωγενείς αιτίες που οφείλονται σε περιβαντολογικούς παράγοντες όπως η θερμοκρασία, η υγρασία κ.α.) οι οποίες είναι αδύνατο να ελεγχθούν ή να εξαλειφθούν. Η μεταβλητότητα αυτή αναφέρεται ως κοινή μορφή μεταβλητότητας και ένα σύστημα το οποίο λειτουργεί με την παρουσία μόνο τέτοιας μορφής μεταβλητότητας θεωρείται ότι είναι εντός στατιστικού ελέγχου (Statistically In-Control Process) ή ότι είναι σε σταθερή κατάσταση (stable state).

Όμως σε μια διεργασία μπορεί να εμφανίζονται και άλλες μορφές μεταβλητότητας, οι οποίες δεν οφείλονται σε τυχαία αίτια αλλά αφορούν τη συστηματική αλλαγή στο επίπεδο κάποιου ή κάποιων παραγόντων που καθορίζουν την ποιότητα του τελικού προϊόντος. Αυτές οι μορφές οφείλονται συνήθως σε ένα από τους παρακάτω λόγους: 1) Λανθασμένα ρυθμισμένες μηχανές, 2) Λάθη του χειριστή της μηχανής, 3) Κακής ποιότητας πρώτη ύλη. Οι παραπάνω μορφές της μεταβλητότητας είναι γενικά αυτές που οδηγούν μια διεργασία εκτός στατιστικού ελέγχου. Οι μορφές αυτής της μεταβλητότητας αναφέρονται ως ειδικές μορφές μεταβλητότητας. Όταν μια διαδικασία λειτουργεί με την παρουσία κάποιας ειδικής μορφής μεταβλητότητας τότε λέμε ότι η διαδικασία είναι εκτός στατιστικού ελέγχου (Statistically Out-of-Control Process) ή ότι είναι σε ασταθή κατάσταση (unstable state).

Βασικές είναι επίσης και οι έννοιες, ελαττωματικό (nonconforming) και μη ελαττωματικό (conforming) προϊόν οι οποίες είναι εύκολα αντιληπτές σε κάθε άνθρωπο. Γενικά, όμως μπορούμε να πούμε ότι ένα προϊόν μπορεί να χαρακτηριστεί ελαττωματικό όταν έχει έναν αριθμό σημαντικών ατελειών (defects). Η έννοια της ατέλειας (defect) γενικά είναι ποιοτική και όχι κατ' ανάγκη αντικειμενική και ορίζεται για κάθε προϊόν κατά την φάση του σχεδιασμού του προϊόντος. Γενικά, ένα προϊόν μετά την παραγωγή του μπορεί να χαρακτηριστεί ως ελαττωματικό εάν απλά το χαρακτηριστικό ποιότητας είναι εκτός κάποιων ορίων τα οποία ονομάζονται όρια προδιαγραφών (specification limits) ή όρια ανοχής (acceptance limits).

6.3.1. Τύποι Διαγραμμάτων Ελέγχου

Τα διαγράμματα ελέγχου (Control Charts) μπορούν να διαχωριστούν σε πολλές κατηγορίες ανάλογα με τα χαρακτηριστικά της διαδικασίας ή ανάλογα της στατιστικής θεωρίας που στηρίζει την κατασκευή κάθε διαγράμματος. Έτσι έχουμε:

Α) Διαγράμματα ελέγχου διεργασιών τύπου Shewhart (Shewhart Type Control Charts), τύπου CUSUM (Cumulative SUM – CUSUM Type Control Charts), και τύπου EWMA (Exponentially Weighted Moving Average - EWMA Type Control Charts).

Β) Η ποσότητα - στόχος ενός διαγράμματος ελέγχου μπορεί να διακριθεί σε δύο κατηγορίες: Εάν το διάγραμμα ελέγχου αφορά το μέσο επίπεδο (mean or nominal level)

της παραγωγικής διεργασίας μιλάμε για διάγραμμα ελέγχου για την μέση τιμή (mean), ενώ αν το διάγραμμα ελέγχου αφορά την διασπορά (dispersion) των μετρήσεων που λαμβάνονται από την παραγωγική διεργασία μιλάμε για διάγραμμα ελέγχου για την διασπορά.

Γ) Εάν από την παραγωγική διεργασία λαμβάνονται δείγματα μετρήσεων μεγέθους μεγαλύτερου της μονάδας, μιλάμε για διαγράμματα ελέγχου για ομάδες (Control Charts for Rational Subgroups), σε αντίθετη περίπτωση μιλάμε για αυτόνομες παρατηρήσεις (Control Charts for Individual Observations).

Δ) Εάν οι μετρήσεις που λαμβάνονται σε κάθε χρονική στιγμή *t* είναι εξαρτημένες από τις μετρήσεις που λαμβάνονται στον χρόνο *t*-1, μιλάμε για διαγράμματα ελέγχου για αυτοσυσχετιζόμενες διεργασίες (Control Charts for Autocorelated Process) ενώ διαφορετικά μιλάμε για μη αυτοσυσχετιζόμενες διεργασίες (Control Charts for non-Autocorelated Process).

Ε) Εάν οι μετρήσεις που λαμβάνονται σε κάθε χρονική στιγμή *t* αφορούν μια μεταβλητή ή ένα χαρακτηριστικό ποιότητας γενικότερα μιλάμε για μονομεταβλητά διαγράμματα (Univariate Control Charts) ελέγχου, διαφορετικά μιλάμε για πολυμεταβλητά διαγράμματα ελέγχου (Multivariate Control Charts).

ΣΤ) Εάν οι μετρήσεις που λαμβάνονται ακολουθούν μια γνωστή κατανομή τότε μιλάμε για παραμετρικά διαγράμματα ελέγχου (Parametric Control Charts). Σε αντίθετη περίπτωση μιλάμε για μη παραμετρικά διαγράμματα ελέγχου (Non Parametric Control Charts).

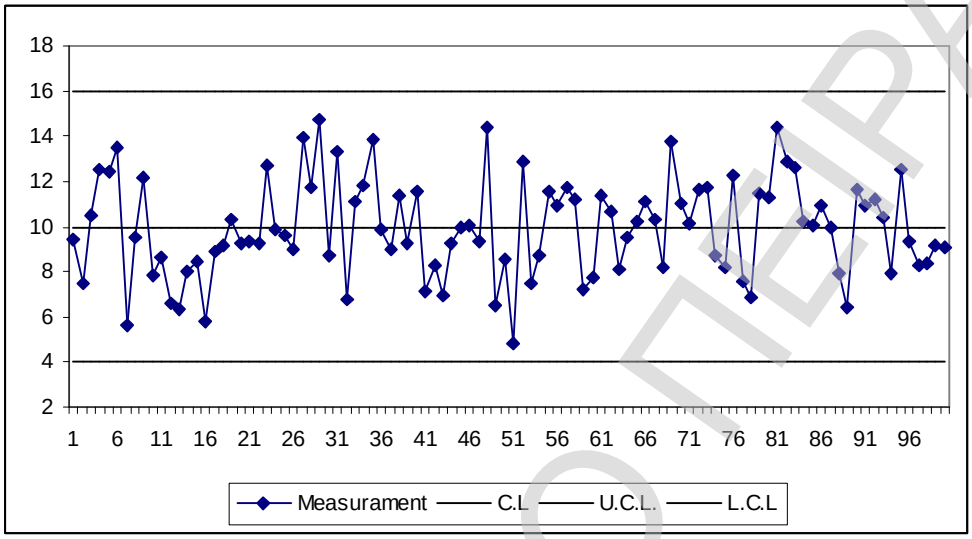
Ζ) Δευτερεύουσες κατηγορίες δημιουργούνται με βάση το μέγεθος του δείγματος και τα χρονικά διαστήματα ανάμεσα σε κάθε δειγματοληψία.

6.3.2. Βασικά Χαρακτηριστικά Διαγραμμάτων Ελέγχου

Στην παρούσα διατριβή θα ασχοληθούμε με εφαρμογές της θεωρίας ροών στα διαγράμματα ελέγχου τύπου Shewhart. Ένα τυπικό διάγραμμα ελέγχου τύπου Shewhart δεν είναι τίποτα περισσότερο από ένα απλό καρτεσιανό διάγραμμα, όπου στον άξονα των *X* καταγράφονται οι διαδοχικοί χρόνοι δειγματοληψίας ή ο αύξων αριθμός του δείγματος και στον άξονα των *Y* οι τιμές της στατιστικής συνάρτησης (συνήθως η

εκτιμήτρια της μέσης τιμής ή της διακύμανσης του μετρήσιμου χαρακτηριστικού στο δείγμα) που μας ενδιαφέρει.

Σχήμα 6.1: Τυπικό Διάγραμμα ελέγχου Shewhart για την μέση τιμή



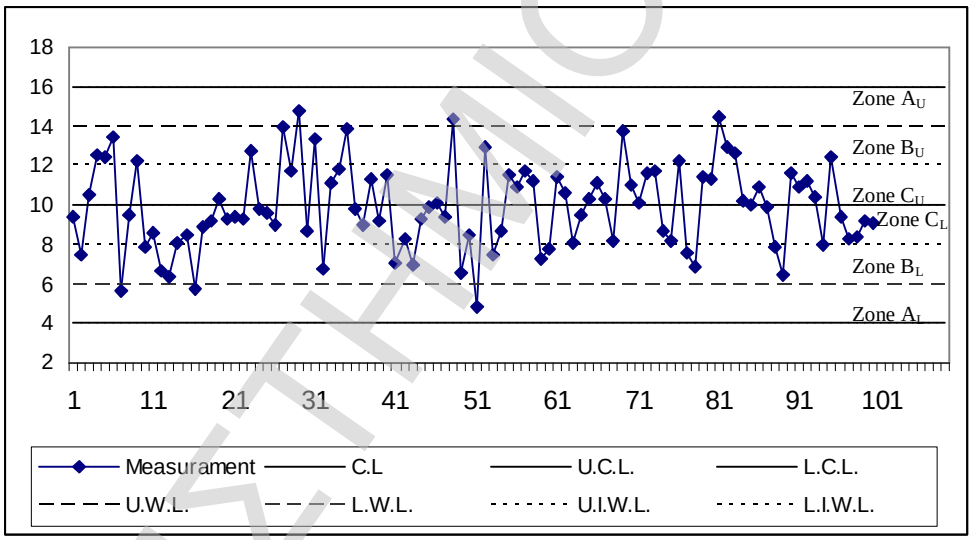
Ένα τυπικό διάγραμμα ελέγχου έχει τα ακόλουθα χαρακτηριστικά:

1. Τα Όρια Προδιαγραφών (Upper and Lower Specification Limits): Ως άνω και κάτω όρια προδιαγραφών (U.S.L. and L.S.L.) χαρακτηρίζονται τα προδιαγεγραμμένα όρια ανοχής μέσα στα οποία οφείλει να κινείται η ποσότητα που απεικονίζεται πάνω στο διάγραμμα. Τα όρια αυτά καθορίζονται κατά την φάση του σχεδιασμού του προϊόντος.
2. Τα Όρια Ελέγχου (Upper and Lower Control Limits): Ως άνω και κάτω όρια προδιαγραφών (U.C.L. and L.C.L.) χαρακτηρίζονται τα στατιστικά όρια μέσα στα οποία οφείλει να κινείται η ποσότητα που απεικονίζεται πάνω στο διάγραμμα. Τα όρια αυτά υπολογίζονται κατά την φάση υπολογισμού των παραμέτρων του διαγράμματος ελέγχου (Phase I). Πρόκειται για όρια που προκύπτουν με χρήση της κατανομής πιθανοτήτων που ακολουθεί η ποσότητα που απεικονίζεται πάνω στο διάγραμμα.
3. Τα Όρια Προειδοποίησης (Upper and Lower Warning Limits): Ως άνω και κάτω όρια προδιαγραφών (U.W.L. and L.W.L.) χαρακτηρίζονται τα προαιρετικά στατιστικά όρια μέσα στα οποία οφείλει να κινείται ένας συγκεκριμένος αριθμός διαδοχικών τιμών της ποσότητας που απεικονίζεται πάνω στο διάγραμμα. Τα όρια

αυτά υπολογίζονται κατά την φάση υπολογισμού των παραμέτρων του διαγράμματος ελέγχου (Phase I). Πρόκειται για όρια που προκύπτουν με χρήση της κατανομής πιθανοτήτων που ακολουθεί η ποσότητα που απεικονίζεται πάνω στο διάγραμμα. Επίσης, έχουμε και τα Εσωτερικά Όρια Προειδοποίησης (Upper Inner and Lower Inner Warning Limits):

4. Την κεντρική γραμμή ή μέσο επίπεδο λειτουργίας (Center Line or Target Value): Σαν κεντρική γραμμή (C.L.) χαρακτηρίζεται το μέσο επίπεδο πάνω στο οποίο οφείλει να κινείται η ποσότητα που απεικονίζεται πάνω στο διάγραμμα. Η ποσότητα αυτή καθορίζεται κατά την φάση του σχεδιασμού του προϊόντος και πρέπει να ταυτίζεται με τη μέση τιμή της διαδικασίας που υπολογίζεται κατά την φάση υπολογισμού των παραμέτρων του διαγράμματος ελέγχου (Phase I).

Σχήμα 6.2: Διάγραμμα ελέγχου Shewhart με Ζώνες



6.3.3. Διαγράμματα Ελέγχου Φάσης I και Φάσης II

Στη βιβλιογραφία έχουν οριστεί δύο αυστηρά διαχωρισμένες φάσεις στην διαδικασία ελέγχου μιας παραγωγικής, βιομηχανικής ή μη, διεργασίας με τη χρήση διαγραμμάτων ελέγχου.

Φάση 1η (Phase I - Off-Line Control Phase): Τα διαγράμματα ελέγχου χρησιμοποιούνται αναδρομικά για να ελέγξουν εάν η διεργασία ήταν εντός ή εκτός

ελέγχου κατά τη χρονική στιγμή που τα πρώτα δείγματα επιλέχθηκαν από τη διεργασία. Σ' αυτή τη φάση τα διαγράμματα ελέγχου βοηθούν τον διαχειριστή της διαδικασίας να φέρει τη διεργασία σε κατάσταση εντός στατιστικού ελέγχου. Όταν αυτό επιτευχθεί, τα διαγράμματα ελέγχου χρησιμοποιούνται για να ορίσουν αυτό που εννοούμε όταν λέμε εντός ελέγχου διεργασία. Αυτή η χρήση των διαγραμμάτων ελέγχου αναφέρεται ως αναδρομική χρήση των διαγραμμάτων ελέγχου. Γενικά, πολύ περισσότερες δράσεις και σκέψεις εξελίσσονται κατά τη διάρκεια αυτής της φάσης, εκτός της απλής διαγραμματοποίησης κάποιων δεδομένων. Κατά τη διάρκεια αυτής της φάσης ο διαχειριστής της διαδικασίας μελετά σε βάθος τη διεργασία, και αυτό γιατί δεν είναι εύκολο να αποφασίσει εάν η διεργασία ήταν όντως εντός στατιστικού ελέγχου κατά τη διάρκεια που τα δεδομένα καταγράφηκαν.

Μια γενική περιγραφή της $1^{ης}$ Φάσης είναι η εξής: Όταν χρησιμοποιήσουμε αρχικά δείγματα για να κατασκευάσουμε τα όρια στα διαγράμματα ελέγχου, τότε αυτά τα όρια θα τα θεωρούμε δοκιμαστικά, γιατί μας επιτρέπουν να καθορίσουμε αν τα αρχικά δείγματα που έχουμε χρησιμοποιήσει είναι ή όχι εντός ελέγχου. Για να εξετάσουμε την υπόθεση ότι τα m αρχικά δείγματα είναι εντός ελέγχου σχεδιάζουμε τα διαγράμματα ελέγχου και εάν όλα τα σημεία είναι εντός των ορίων ελέγχου και δεν υπάρχει κάποια συστηματική συμπεριφορά καταλήγουμε στο συμπέρασμα ότι η διεργασία ήταν υπό έλεγχο στο παρελθόν και άρα τα δοκιμαστικά όρια ελέγχου είναι κατάλληλα για να διατηρήσουμε υπό έλεγχο στο παρόν ή στο μέλλον την διεργασία. Αν δούμε ότι μια ή περισσότερες τιμές στα διαγράμματα ελέγχου είναι εκτός των ορίων ελέγχου τότε εξετάζουμε τα δοκιμαστικά όρια ελέγχου και κάθε σημείο που είναι εκτός ελέγχου ελέγχεται για πιθανή ύπαρξη ειδικού λόγου μεταβλητότητας. Αν βρεθεί κάποιος τέτοιος λόγος ξανά υπολογίζουμε τα όρια χωρίς να χρησιμοποιήσουμε αυτό το σημείο. Στην συνέχεια αλγοριθμικά επαναλαμβάνουμε την διαδικασία αυτή εξετάζοντας τα υπόλοιπα σημεία για το αν είναι εντός ελέγχου (και αυτό διότι πλέον μετά την αφαίρεση ένος σημείου από τα δεδομένα μας τα όρια θα είναι τώρα πιο στενά). Η διαδικασία αυτή επαναλαμβάνεται μέχρι όλα τα σημεία να βρεθούν εντός ελέγχου. Σε περίπτωση που έχουμε κάποιο σημείο εκτός ελέγχου και δεν μπορούμε να βρούμε κάποιο ειδικό λόγο μεταβλητότητας, υπάρχουν δύο εναλλακτικές αντιμετώπισης. Η πρώτη είναι να αφαιρέσουμε το σημείο όπως θα κάναμε αν είχαμε βρει κάποιο ειδικό λόγο μεταβλητότητας. Η μόνη εξήγηση που μπορεί να δοθεί για αυτή την ενέργεια είναι ότι

γενικά ένα σημείο εκτός των ορίων είναι πιθανό να προέρχεται από κατανομή πιθανότητας ενός χαρακτηριστικού που είναι εκτός ελέγχου. Η εναλλακτική ενέργεια είναι να θεωρήσουμε ότι τα όρια που υπολογίσαμε είναι κατάλληλα για να ελέγχουμε αν η διαδικασία είναι υπό έλεγχο. Αν βέβαια το σημείο που βρίσκεται εκτός ελέγχου πράγματι παρουσιάζει μια εκτός ελέγχου κατάσταση τότε τα όρια θα είναι μεγαλύτερα από ότι θα έπρεπε να είναι. Η ύπαρξη όμως μόνο ενός ή δύο τέτοιων σημείων δεν θα επηρεάσει σημαντικά τα αποτελέσματα μας. Αν επιπλέον τα μελλοντικά δείγματα είναι υπό έλεγχο μπορούμε να θεωρήσουμε ότι τα όρια είναι ικανοποιητικά.

Φάση 2η (Phase II - On-Line Control Phase): Τα διαγράμματα ελέγχου χρησιμοποιούνται προκειμένου να παρακολουθείται συνεχώς η διεργασία ούτως ώστε να ελέγξουμε εάν η διαδικασία παραμένει εντός ελέγχου. Δηλαδή, στην φάση αυτή ο διαχειριστής έχει στα χέρια του ένα πολύτιμο εργαλείο μέσω του οποίου είναι δυνατόν να παρακολουθεί συνεχώς την παραγωγική διεργασία και να ανιχνεύει γρήγορα μια πιθανή αλλαγή στο επίπεδο των χαρακτηριστικών που καθορίζουν την ποιότητα του παραγόμενου προϊόντος. Δηλαδή, σε κάθε χρονικό σημείο που ένα δείγμα μεγέθους ***n*** λαμβάνεται από την διεργασία ο υπεύθυνος παίρνει μια απάντηση στο ερώτημα «παραμένει η διεργασία εντός ελέγχου;». Σε αυτήν την φάση ο υπεύθυνος αδιαφορεί για το εάν το επίπεδο ελέγχου έχει εκτιμηθεί ή έχει δοθεί.

Μια γενική περιγραφή της 2ης Φάσης είναι η εξής: Τα διαδοχικά δείγματα που λαμβάνονται από την διεργασία απεικονίζονται σε ένα διάγραμμα ελέγχου, με στόχο τον συνεχή έλεγχο για το αν η διεργασία βρίσκεται εντός ελέγχου. Στόχος δηλαδή της λειτουργίας των διαγραμμάτων ελέγχου στην φάση αυτή, είναι να ανιχνεύσει το συντομότερο δυνατόν την ύπαρξη μιας μετατόπισης στο μέσο επίπεδο λειτουργίας της παραγωγικής διεργασίας ή μια αυξητική μετατόπιση της διασποράς της διεργασίας, που πλέον οδηγεί την παραγωγική διεργασία σε Out-of-Control κατάσταση.

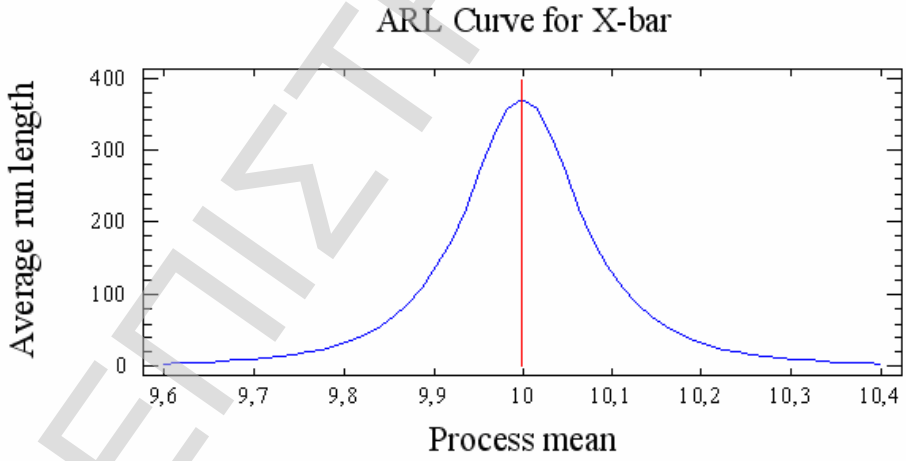
Ο Woodall (2000) υποστηρίζει ότι μεγάλη προσπάθεια εμβάθυνσης στο τρόπο που δουλεύει η διεργασία και βελτιωτικές κινήσεις πρέπει να γίνουν κατά το πέρασμα από την Φάση I στην Φάση II.

6.3.4. Μέσο Μήκος Ροής (Average Run Length (ARL))

Για κάθε διάγραμμα ελέγχου το ARL είναι ένα μέτρο της συμπεριφοράς του ή αλλιώς ένα μέτρο της ικανότητας του. Το ARL δίνει τον αναμενόμενο αριθμό δειγμάτων που λαμβάνονται μέχρι την πρώτη εμφάνιση ενός εκτός ελέγχου σήματος είτε στην περίπτωση που η διεργασία είναι εντός ελέγχου είτε στην περίπτωση που η διεργασία είναι εκτός ελέγχου. Μια σημαντική απόφαση που πρέπει να λάβουμε για την κατασκευή ενός διαγράμματος ελέγχου τύπου Shewhart, είναι η επιλογή της τιμής του εντός ελέγχου ARL (του μέσου μήκους ροής για όσο η διεργασία παραμένει εντός ελέγχου), δηλαδή το In-Control Average Run Length το οποίος στο εξής θα συμβολίζουμε με ARL_0.

Για το ARL_0 επιθυμούμε να έχει μεγάλη τιμή αφού μας δίνει το μέσο αριθμό των δειγμάτων εντός των οποίων αναμένεται να υπάρξει μια εσφαλμένη ένδειξη για εκτός ελέγχου διεργασία. Το ARL_0 συνδέεται με την πιθανότητα σφάλματος τύπου I, εάν δούμε τα διαγράμματα ελέγχου ως διαδοχικούς ελέγχους υποθέσεων.

Σχήμα 6.3: Γραφική παράσταση του μέσου μήκους ροής $ARL\ (N(10, (0.25)^2/2))$ ενός διαγράμματος τύπου Shewhart για τον έλεγχο της μέσης τιμής (με χρήση του Statgraphics)



Σημαντικότατο ρόλο διαδραματίζει και η πιθανότητα σφάλματος τύπου II, η οποία καθορίζει την ποσότητα του μέσου μήκους ροής για μια εκτός ελέγχου διεργασία,

δηλαδή το Out-of-Control Average Run Length το οποίο στο εξής θα συμβολίζουμε με ARL_1.

Για το ARL_1 επιθυμούμε να έχει μικρή τιμή αφού μας δίνει το μέσο αριθμό των δειγμάτων εντός των οποίων αναμένεται να υπάρξει μια ένδειξη ότι η διεργασία είναι εκτός ελέγχου από τη στιγμή που η διεργασία βρέθηκε πραγματικά εκτός ελέγχου. Το ARL_1 είναι ένας πολύ σημαντικός δείκτης ευαισθησίας ενός διαγράμματος ελέγχου ο οποίος μας επιτρέπει την σύγκριση δύο διαγραμμάτων ελέγχου με το ίδιο ARL_0.

6.4. Θεωρία Ροών Επιτυχιών και Διαγράμματα Ελέγχου τύπου Shewhart

Έχει διαπιστωθεί ότι για μικρές μετατοπίσεις (αυξήσεις ή μειώσεις) του μέσου επιπέδου της διεργασίας ή και της διασποράς της διεργασίας, το εκτός ελέγχου μέσο μήκος ροής του διαγράμματος ελέγχου τύπου Shewhart δεν είναι ικανοποιητικό (είναι αρκετά μεγάλος αριθμός).

Ως λύση στο πρόβλημα αυτό έχει προταθεί στην διεθνή βιβλιογραφία ένα ευρύ σύνολο από κανόνες οι οποίοι βασίζονται στην θεωρία ροών επιτυχιών προκειμένου τα διαγράμματα ελέγχου τύπου Shewhart να γίνουν ευαίσθητα στην ανίχνευση των μικρών μετατοπίσεων.

Οι σημαντικότεροι κανόνες που χρησιμοποιούνται για την ευαισθητοποίηση ενός διαγράμματος ελέγχου Shewhart σύμφωνα με τον Montgomery (2000), είναι οι ακόλουθοι (βλέπε Σχήμα 6.2 για αναγνώριση των ζωνών A, B, C):

1. Ένα σημείο εκτός των ορίων ελέγχου.

2. Δύο από τρία συνεχόμενα σημεία στην Ζώνη A (σε μια από τις δύο ζώνες A).

3. Τέσσερα από πέντε συνεχόμενα σημεία πέραν της Ζώνης C (στις δύο περιοχές).

4. Οκτώ συνεχόμενα σημεία στην ίδια μεριά (πάνω ή κάτω) της κεντρικής γραμμής.

5. Έξι συνεχόμενα σημεία σε αύξουσα ή φθίνουσα διάταξη.

6. Δεκαπέντε συνεχόμενα σημεία στην Ζώνη C.

7. Δεκατέσσερα συνεχόμενα σημεία σε εναλλασσόμενη μορφή πάνω / κάτω.

8. Οκτώ συνεχόμενα σημεία εκτός της ολικής Ζώνης C.

9. Οποιαδήποτε ασυνήθιστη ή μη τυχαία ακολουθία σημείων.

10. Ένα ή περισσότερα σημεία κοντά στα προειδοποιητικά όρια ή τα όρια ελέγχου.

Οι πρώτοι τέσσερις κανόνες είναι γνωστοί ως ***Western Electric rules*** λόγω του ότι προτάθηκαν για πρώτη φορά σε βιβλίο της συγκεκριμένης εταιρείας για τον ποιοτικό έλεγχο (Western Electric Company (1956)). Οι υπόλοιποι κανόνες έχουν προταθεί και μελετηθεί κατά περιόδους από άλλους ερευνητές.

Η χρήση πολλών κανόνων ταυτοχρόνως γίνεται με ιδιαίτερη προσοχή και αυτό διότι η ταυτόχρονη χρήση τους συνεπάγεται μεγάλο αριθμό λανθασμένων συναγερμών. Δηλαδή, μεγάλο αριθμό λανθασμένων διακοπών της παραγωγικής διεργασίας για την ανίχνευση ειδικών αιτιών μεταβλητότητας με αποτέλεσμα την αύξηση του κόστους παραγωγής. Η πρώτη εργασία που μελετά το πρόβλημα αυτό (με προσομοίωση) είναι των Walker et al. (1991).

Ένα άλλο σημαντικό πρόβλημα που δημιουργείται από την ταυτόχρονη χρήση πολλών κανόνων ροών είναι το ότι καθίσταται εξαιρετικά δύσκολος ο υπολογισμός του μέσου μήκους ροής του διαγράμματος ελέγχου. Ο Page (1954) πρότεινε ως λύση τη χρήση της τεχνικής της Μαρκοβιανής εμφύτευσης για τον υπολογισμό της κατανομής του μήκους ροής ενός διαγράμματος ελέγχου. Η πρώτη εργασία που μελετά την κατανομή του ARL με ένα ενοποιημένο τρόπο είναι αυτή των Champ και Woodall (1987). Η μελέτη του ARL στο άρθρο αυτό έγινε με την τεχνική της εμφύτευσης σε Μαρκοβιανή αλυσίδα.

Άλλες εργασίες που αντιμετωπίζουν το θέμα του υπολογισμού του ARL ενός διαγράμματος ελέγχου με κανόνες ροών ή προτείνουν νέους κανόνες ροών έχουν παρουσιαστεί από τους Mosteller (1941), Dudding and Jannet (1942), Weiler (1953), Moore (1958), Roberts (1958), Westgard and Groth (1977), Westgard et al. (1979), Bissel (1978), Nelson (1984, 1985), SAS Institute (1986), Coleman (1986), Palm (1990), Champ and Woodall (1990), Champ (1992), Alwan et al. (1994), Lowry et al. (1995), Divoky and Taylor (1995), Champ και Woodall (1997), Derman and Ross

(1997), Klein (2000), Shmueli and Cohen (2000), Fu et al (2002, 2003), Koutras et al. (2005).

6.5. Η Τεχνική της Μαρκοβιανής εμφύτευσης για τον υπολογισμό του ARL

Όπως είδαμε, τα διαγράμματα ελέγχου αποτελούν κοινό τόπο εφαρμογής, τόσο της θεωρίας ροών και σχηματισμών όσο και της τεχνικής της Μαρκοβιανής εμφύτευσης.

Προκειμένου να είναι πιο κατανοητή η μαθηματική διερεύνηση της κατανομής του ARL ενός διαγράμματος ελέγχου, παρατηρούμε ότι:

- μπορούμε να δούμε ένα διάγραμμα ελέγχου με ζώνες ως μια ακολουθία ανεξάρτητων πολύτιμων δοκιμών (δεδομένου ότι σε κάθε χρονική στιγμή η στατιστική συνάρτηση που απεικονίζουμε στο διάγραμμα και της οποίας γνωρίζουμε την κατανομή, έχει συγκεκριμένη πιθανότητα να βρεθεί εντός μιας ζώνης),
- μπορούμε να δούμε καθέναν από τους κανόνες ευαισθητοποίησης, ως σχηματισμούς στην ακολουθία των πολύτιμων δοκιμών, και τέλος
- μπορούμε να μελετήσουμε τον χρόνο αναμονής μέχρι την πρώτη εμφάνιση του υπό μελέτη σχηματισμού (άρα κανόνα ευαισθητοποίησης) με χρήση της τεχνικής της Μαρκοβιανής εμφύτευσης.

Συνεπώς, μπορούμε να παραστήσουμε το διάγραμμα ελέγχου ως μια ακολουθία ανεξάρτητων και ισόνομων τυχαίων μεταβλητών $\{X_t,\ t \geq 1\}$ με σύνολο τιμών $A=\{a_1, a_2, ..., a_l\}$, $l \geq 2$, και έστω ότι $P(X_t = a_i) = p_i,\quad 1 \leq i \leq l,\quad t \geq 1$. Ουσιαστικά το σύνολο τιμών $A=\{a_1, a_2, ..., a_l\}$ καθορίζεται από τις διακριτές ζώνες μέσα στις οποίες είναι δυνατόν να βρεθεί η στατιστική συνάρτηση που απεικονίζεται στο διάγραμμα ελέγχου σε κάθε χρονική στιγμή t. Δηλαδή, θα καλούμε την τυχαία μεταβλητή X_t ως t-οστή δοκιμή.

Επίσης, μπορούμε να θεωρήσουμε κάθε κανόνα απόφασης για το αν η παραγωγική διεργασία είναι εντός ή εκτός ελέγχου ως ένα σύνθετο ή απλό ενδεχόμενο Ε για το οποίο μπορούμε να απαντήσουμε στο ερώτημα αν αυτό έχει συμβεί στη t-οστή δοκιμή

της πεπερασμένης ακολουθίας $X_1 X_2,..., X_t$, υποθέτοντας ότι το ενδεχόμενο Ε συμβαίνει τουλάχιστον μια φορά με πιθανότητα 1 σε μια επαρκώς μεγάλη ακολουθία δοκιμών.

Τέλος, αν συμβολίσουμε με T την τυχαία μεταβλητή που δηλώνει το χρόνο αναμονής μέχρι την πρώτη εμφάνιση του ενδεχόμενου Ε, τότε η μελέτη της τυχαίας μεταβλητής T μπορεί να επιτευχθεί με χρήση της μεθόδου της εμφύτευσης των τυχαίων μεταβλητών X_t σε μια ομογενή διακριτή αλυσίδα Markov.

Την τεχνική αυτή χρησιμοποίησαν οι Champ and Woodall (1987) προκειμένου να μελετήσουν το ARL διαγραμμάτων τύπου Shewhart με κανόνες ροών. Σύμφωνα με τους Champ and Woodall (1987) ο συμβολισμός $T(k, m, a, b)$ δηλώνει ότι k από m διαδοχικά σημεία του διαγράμματος βρίσκονται στο διάστημα $(m + as\,, m + bs\,)$, $a < b$.

Έτσι στο σύνηθες διάγραμμα ελέγχου Shewhart, $L = 3$, ο Κανόνας 1 της Παραγράφου 6.4, που είναι ο κλασικός κανόνας λήψης ένδειξης εκτός ελέγχου διεργασίας, μπορεί να γραφεί στη μορφή $C_1 = \{T(1,1,-\infty,-3), T(1,1,3,\infty)\}$, και ο Κανόνας 2 της Παραγράφου 6.4 μπορεί να γραφεί ως $C_2 = \{T(2,3,-3,-2), T(2,3,2,3)\}$.

Οι Champ και Woodall (1987) μελέτησαν τους ακόλουθους κανόνες ευαισθητοποίησης

$1^{ος}$ Κανόνας: $C_1 = \{T(1,1,-\infty,-3), T(1,1,3,\infty)\}$

$2^{ος}$ Κανόνας: $C_2 = \{T(2,3,-3,-2), T(2,3,2,3)\}$

$3^{ος}$ Κανόνας: $C_3 = \{T(4,5,-3,-1), T(4,5,1,3)\}$

$4^{ος}$ Κανόνας: $C_4 = \{T(8,8,-3,0), T(8,8,0,3)\}$

$5^{ος}$ Κανόνας: $C_5 = \{T(2,2,-3,-2), T(2,2,2,3)\}$

$6^{ος}$ Κανόνας: $C_6 = \{T(5,5,-3,-1), T(5,5,1,3)\}$

$7^{ος}$ Κανόνας $C_7 = \{T(1,1,-\infty,-3.09), T(1,1,3.09,\infty)\}$

$8^{ος}$ Κανόνας: $C_8 = \{T(2,3,-3.09,-1.96), T(2,3,-1.96,3.09\}$

$9^{ος}$ Κανόνας: $C_9 = \{T(8,8,-3.09,0), T(8,8,0,3.09)\}$.

Με το συμβολισμό $C_{ij...k} = C_i \cup C_j \cup ... \cup C_k$ δηλώνεται ένα διάγραμμα ελέγχου Shewhart το οποίο δίνει ένδειξη εκτός ελέγχου διεργασίας όταν συμβεί τουλάχιστον ένα

ενδεχόμενο από αυτά που περιγράφουν οι κανόνες $C_i, C_j, ..., C_k$. Ο κανόνας $C_{ij...k}$ ονομάζεται σύνθετος κανόνας.

Στη συνέχεια παρουσιάζουμε τον πίνακα (Πίνακας 6.1) που έδωσαν οι Champ and Woodall (1987) (βλέπε επίσης Rakitzis (2004)) για το μέσο μήκος ροής ενός διαγράμματος ελέγχου Shewhart ($LCL = m - 3s$, $UCL = m + 3s$) υπό την παρουσία κανόνων ευαισθητοποίησης και υπό την υπόθεση ότι η απεικονιζόμενη στατιστική συνάρτηση ακολουθεί κατανομή $N(m + ds, s^2)$ (εντός ελέγχου μέσος και τυπική απόκλιση m και s αντιστοίχως, ενώ το d συμβολίζει την μετατόπιση από τον εντός ελέγχου μέσο m σε μονάδες τυπικής απόκλισης).

Γενικά, από τον πίνακα αυτό προκύπτει ότι το μέσο μήκος ροής μειώνεται για μικρές μετατοπίσεις του μέσου στην περίπτωση που χρησιμοποιούμε τον κανόνα C_1 (ή τον ισοδύναμο κανόνα C_7) μαζί με ένα τουλάχιστον επιπρόσθετο κανόνα ευαισθητοποίησης, σε σχέση με την αποκλειστική χρησιμοποίηση του κλασικού κανόνα C_1. Παρατηρούμε αντίστοιχα ότι στην ίδια περίπτωση μειώνεται το εντός ελέγχου μέσο μήκος ροής που ισοδυναμεί με αύξηση των λανθασμένων συναγερμών. Ωστόσο μπορούμε να επιτύχουμε οποιοδήποτε εντός ελέγχου μέσο μήκος ροής αυξάνοντας απλά το πλάτος (L) των ορίων ελέγχου.

Πίνακας 6.1: Μέσο μήκος ροής για διαγράμματα ελέγχου τύπου Shewhart με χρήση κανόνων ευαισθητοποίησης

δ	$ARL(d)$															
	C_1	C_7	C_{12}	C_{78}	C_{15}	C_{13}	C_{14}	C_{79}	C_{16}	C_{123}	C_{156}	C_{124}	C_{789}	C_{134}	C_{1456}	C_{1234}
0.0	370.40	499.62	225.44	239.75	278.03	166.05	152.73	170.41	349.38	132.89	266.82	122.05	122.05	105.78	133.21	91.75
0.2	308.43	412.01	177.56	185.48	222.59	120.70	110.52	120.87	279.53	97.86	208.82	89.14	89.14	76.01	96.37	66.80
0.4	200.08	262.19	104.46	106.15	134.17	63.88	59.76	63.80	165.48	52.93	119.47	48.71	48.71	40.95	51.94	36.61
0.6	119.67	153.86	57.92	57.80	75.27	33.99	33.64	35.46	89.07	28.70	63.70	27.49	27.49	23.15	29.01	20.90
0.8	71.55	90.41	33.12	32.75	42.96	19.78	21.07	22.09	48.40	16.93	34.96	17.14	17.14	14.62	17.94	13.25
1.0	43.89	54.55	20.01	19.70	25.61	12.66	15.58	15.26	27.74	10.95	20.43	11.73	11.73	10.19	12.19	9.22
1.2	27.82	34.03	12.81	12.62	16.06	8.84	10.90	11.42	17.05	6.78	12.83	8.61	8.61	7.66	8.90	6.89
1.4	18.25	21.97	8.69	8.58	10.60	6.62	8.60	9.05	11.28	5.76	8.65	6.63	6.63	6.08	6.84	5.41
1.6	12.38	14.68	6.21	6.16	7.36	5.24	7.03	7.44	7.98	4.54	6.22	5.27	5.27	5.01	5.42	4.41
1.8	8.69	10.15	4.66	4.64	5.36	4.33	5.85	6.24	5.97	3.73	4.71	4.27	4.27	4.24	4.39	3.68
2.0	6.30	7.25	3.65	3.65	4.07	3.68	4.89	5.25	4.67	3.14	3.72	3.50	3.50	3.65	3.61	3.13
2.2	4.72	5.36	2.96	2.98	3.22	3.18	4.08	4.41	3.78	2.70	3.04	2.91	2.91	3.17	3.01	2.70
2.4	3.65	4.08	2.48	2.51	2.64	2.78	3.38	3.67	3.14	2.35	2.55	2.47	2.47	2.77	2.54	2.35
2.6	2.90	3.20	2.13	2.17	2.22	2.43	2.81	3.05	2.64	2.07	2.19	2.13	2.13	2.43	2.19	2.07
2.8	2.38	2.59	1.87	1.91	1.93	2.14	2.35	2.54	2.26	1.85	1.91	1.87	1.87	2.14	1.91	1.85
3.0	2.00	2.15	1.68	1.71	1.70	1.89	1.99	2.14	1.95	1.67	1.70	1.68	1.68	1.89	1.70	1.67

6.6. Το Διάγραμμα Ελέγχου C^2 με κανόνες Ροών

6.6.1. Το Διάγραμμα Chi-Square

Σήμερα, το διάγραμμα ελέγχου c^2 αποτελεί την πιο διαδεδομένη μέθοδο για την παρακολούθηση και τον έλεγχο παραγωγικών διεργασιών στις οποίες η ποιότητα του τελικού προϊόντος εξαρτάται από την από κοινού συμπεριφορά p τυχαίων μεταβλητών. Το διάγραμμα αυτό χαρακτηρίζεται ως διάγραμμα ελέγχου τύπου Shewhart επειδή κάθε σημείο που απεικονίζεται σε αυτό χρησιμοποιεί πληροφορίες μόνο από το τρέχον δείγμα. Έτσι, το διάγραμμα παρουσιάζει αδυναμία στην ανίχνευση μικρών αλλαγών του διανύσματος των μέσων των μεταβλητών.

Στην παράγραφο αυτή της παρούσας διατριβής εισάγουμε μια τροποποίηση του διαγράμματος ελέγχου Chi-Square, κάνοντας επίσης χρήση θεωρίας ροών επιτυχιών με έναν διαφορετικό και πιο αποδοτικό τρόπο. Το προτεινόμενο διάγραμμα ελέγχου παρουσιάζει αυξημένη ευαισθησία στην ανίχνευση μικρών αλλαγών του διανύσματος των μέσων (βλέπε επίσης Koutras et al. (2005a)).

Όπως προαναφέραμε, η παρακολούθηση μια διεργασίας που αναφέρεται στον ταυτόχρονο έλεγχο δυο ή περισσοτέρων εξαρτημένων μεταβλητών (ποιοτικά χαρακτηριστικά) αναφέρεται συνήθως στη βιβλιογραφία ως *πολυμεταβλητός έλεγχος ποιότητας*. Η έρευνα στην περιοχή του πολυμεταβλητού ελέγχου ποιότητας ξεκίνησε με την πρωτοπόρο εργασία του Hotelling (1947). Από τότε έχουν εμφανισθεί στη βιβλιογραφία πολλές εργασίες σχετικές με προβλήματα πολυμεταβλητού ελέγχου ποιότητας όπως για παράδειγμα οι εργασίες των Alt and Smith (1988), Crosier (1988), Pignatiello and Runger (1990), Lowry and Montgomery (1995), Bersimis (2000) και Bersimis et al. (2005). Τα πιο γνωστά σχήματα πολυμεταβλητού ελέγχου ποιότητας είναι τα διαγράμματα ελέγχου Shewhart, CUSUM και EWMA. Για ειδικές εφαρμογές αυτών των διαγραμμάτων, καθώς και για εφαρμογές άλλων πολυμεταβλητών μεθόδων στη βελτίωση της ποιότητας, ο ενδιαφερόμενος αναγνώστης παραπέμπεται στους Crosier (1998), Pignatiello and Runger (1990),

Hawkins (1991), Lowry and Montgomery (1995), Ryan (2000) και Maravelakis et al. (2002).

Όπως έχει ήδη αναφερθεί, το ενδιαφέρον της παραγράφου αυτής έγκειται στην ταυτόχρονη παρακολούθηση m συσχετισμένων μεταβλητών $X_1, X_2, \mathbf{K}, X_m$ που χαρακτηρίζουν την ποιότητα μιας διεργασίας. Θα υποθέτουμε ότι η εντός ελέγχου από κοινού κατανομή πυκνότητας του διανύσματος $\mathbf{X} = (X_1, X_2, \mathbf{L}, X_m)'$ ακολουθεί την m-διάστατη Κανονική κατανομή, με διάνυσμα μέσου $\boldsymbol{\mu}_0$ και πίνακα διακυμάνσεων-συνδιακυμάνσεων $\boldsymbol{\Sigma}_0$, δηλαδή $\mathbf{X} \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.

Σε κάθε χρονική στιγμή i εξετάζεται ένα δείγμα μεγέθους $n > 1$ και υπολογίζεται το διάνυσμα δειγματικών μέσων $\overline{\mathbf{X}}_i$ του i-οστού δείγματος (subgroup).

Σε ένα διάγραμμα ελέγχου c^2 (από εδώ και πέρα θα συμβολίζεται με CSCC) για την παρακολούθηση του μέσου της διεργασίας, απεικονίζεται η στατιστική συνάρτηση ελέγχου $D_i^2 = n(\overline{\mathbf{X}}_i - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\overline{\mathbf{X}}_i - \boldsymbol{\mu}_0)$, $i \geq 1$ που είναι η σταθμισμένη απόσταση (απόσταση Mahalanobis) μεταξύ των $\overline{\mathbf{X}}_i$ και $\boldsymbol{\mu}_0$ στον Ευκλείδειο χώρο R^{m} και ακολουθεί μια (κεντρική) κατανομή c^2 με m βαθμούς ελευθερίας ($D_i^2 \sim c_m^2$). Αυτό σημαίνει ότι το άνω όριο ελέγχου του CSCC δίνεται από το $UCL = c_{m,a}^2$, όπου $c_{m,a}^2$ είναι το άνω a ποσοστιαίο σημείο της κατανομής c^2, δηλαδή $\Pr(D_i^2 > c_{m,a}^2) = a$. Αν $D_i^2 > c_{m,a}^2$ τότε το διάγραμμα ελέγχου δίνει σήμα ότι η διεργασία είναι πιθανώς εκτός ελέγχου λόγω κάποιας πραγματικής αιτίας, διαφορετικά η διεργασία θεωρείται εντός ελέγχου και δεν απαιτείται καμιά ενέργεια.

Αξίζει να σημειωθεί ότι σε ένα CSCC δεν υπάρχει ανάγκη για κάτω όριο ελέγχου αφού η διαφοροποίηση μεταξύ των εντός και εκτός ελέγχου καταστάσεων προσδιορίζεται από το $D_i^2 \geq 0$. Τιμές της στατιστικής συνάρτησης D_i^2 πολύ κοντά στο 0 δείχνουν ότι το σημείο $\overline{\mathbf{X}}_i$ είναι κοντά στο $\boldsymbol{\mu}_0$ και συνεπώς είναι λογικό να υποθέσουμε ότι η διεργασία είναι εντός ελέγχου. Προφανώς η τιμή της D_i^2 εξαρτάται μόνο από την απόσταση του εκτός ελέγχου από τον εντός ελέγχου μέσο και όχι από τη συγκεκριμένη κατεύθυνση της διαφοράς (directional invariance).

Για δείγματα μεγέθους 1 $(n = 1)$, στον υπολογισμό της D_i^2, η ποσότητα $\overline{\mathbf{X}}_i$ πρέπει να αντικατασταθεί από την $\mathbf{X}_i$. Ένα άλλο ενδιαφέρον σημείο είναι ότι, σε

πραγματικές εφαρμογές, οι παράμετροι $\boldsymbol{\mu}_0$ και $\boldsymbol{\Sigma}_0$ είναι συνήθως άγνωστες, οπότε πρέπει να εκτιμηθούν από την ανάλυση προκαταρκτικών δειγμάτων. Στη συνέχεια το κατάλληλο διάγραμμα ελέγχου οδηγεί στο λεγόμενο Hotelling T^2 διάγραμμα ελέγχου του οποίου το άνω όριο ελέγχου εξαρτάται από τα ποσοστημόρια μιας κατάλληλης F κατανομής (για περισσότερες λεπτομέρειες παραπέμπουμε στους Alt and Smith (1988) και Lowry and Montgomery (1995)).

Το CSCC είναι ένα διάγραμμα ελέγχου τύπου Shewhart αφού λαμβάνει υπόψη μόνο την πληροφορία που προκύπτει από το τελευταίο δείγμα που έχει ελεγχθεί. Το γεγονός αυτό μεταφέρει στο CSCC το κλασικό μειονέκτημα όλων των διαγραμμάτων ελέγχου Shewhart, που είναι η μη ευαισθησία στον εντοπισμό βαθμιαίων ή μικρών μετατοπίσεων στο διάνυσμα του μέσου της διεργασίας.

Στην παρούσα παράγραφο, θεωρούμε ένα CSCC που δίνει ένδειξη για εκτός ελέγχου διεργασία όταν k συνεχόμενες τιμές της D_i^2 υπερβαίνουν ένα κατάλληλο άνω όριο ελέγχου L_k (k θετικός ακέραιος). Η διαδικασία ονομάζεται $k\,|\,k$ CSCC και στην ειδική περίπτωση όπου $k=1$ συμπίπτει με το αντίστοιχο τυπικό διάγραμμα. Για $k \geq 2$ το νέο διάγραμμα ελέγχου έχει καλύτερο (μικρότερο) μέσο μήκος ροής από το αντίστοιχο τυπικό ($1\,|\,1$ CSCC). Επιπλέον, εισάγουμε ένα συνδυασμένο $r\,|\,r-k\,|\,k$ CSCC το οποίο είναι ένα διάγραμμα ελέγχου με δυο όρια ελέγχου L_r και L_k που δίνουν σήμα εκτός ελέγχου διεργασίας αν k συνεχόμενες τιμές της D_i^2 υπερβαίνουν το L_k ή r συνεχόμενες τιμές της D_i^2 υπερβαίνουν το L_r ($r<k$). Όπως αποδεικνύεται το συνδυασμένο $r\,|\,r-k\,|\,k$ CSCC με $k \geq 2$ είναι πιο αποτελεσματικό από τα $1\,|\,1$ και $k\,|\,k$ CSCC.

Η Παράγραφος 6.6 έχει οργανωθεί ως εξής: Στην υποπαράγραφο 6.6.2 εισάγουμε το $k\,|\,k$ CSCC ενώ στην υποπαράγραφο 6.6.3 μελετούμε θεωρητικά τα χαρακτηριστικά του, εστιάζοντας κυρίως στην λειτουργικότητά του συγκρινόμενη με αυτή του τυπικού CSCC. Στην υποπαράγραφο 6.6.4 εισάγουμε το συνδυασμένο $r\,|\,r-k\,|\,k$ CSCC και μελετούμε την ειδική περίπτωση $r=1$. Στην υποπαράγραφο 6.6.5 μελετούμε το μέσο μήκος ροής των δυο νέων CSCC. Τέλος, στην υποπαράγραφο 6.6.6 παρουσιάζουμε συνοπτικά ορισμένα συμπεράσματα.

6.6.2. Το *k*|*k* CSCC

Θεωρούμε ένα τυπικό διάγραμμα ελέγχου τύπου Shewhart με ένα άνω όριο ελέγχου L_1. Υποθέτουμε επίσης ότι οι τιμές μιας στατιστικής συνάρτησης W_i απεικονίζονται σε διάγραμμα ελέγχου και η διεργασία θεωρείται εκτός ελέγχου αν $W_i > L_1$. Είναι γνωστό ότι το εντός ελέγχου μέσο μήκος ροής (ARL_{in}) δίνεται από τον τύπο

$$ARL_{in} = \frac{1}{p_1}$$

όπου $p_1 = \Pr(W_i > L_1)$ (η τελευταία πιθανότητα υπολογίζεται υπό την υπόθεση ότι η διεργασία είναι εντός ελέγχου). Επιπλέον, το εκτός ελέγχου μέσο μήκος ροής (ARL_{out}) δίνεται από τον τύπο

$$ARL_{out} = \frac{1}{\Pr(W_i > L_1)}$$

με την πιθανότητα στον παρονομαστή να υπολογίζεται υπό την υπόθεση ότι η παράμετρος της διεργασίας έχει μετατοπιστεί σε μια τιμή διαφορετική από αυτή που έχει οριστεί ως τιμή εντός ελέγχου (τιμή στόχος).

Σαν παράδειγμα, έστω ένα CSCC με εντός ελέγχου διάνυσμα μέσων $\boldsymbol{\mu}_0$. Το στατιστικό ελέγχου δίνεται από τον τύπο $W_i = D_i^2 = n(\overline{\mathbf{X}}_i - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\overline{\mathbf{X}}_i - \boldsymbol{\mu}_0)$, $i \geq 1$ και ακολουθεί κατανομή c^2 με m βαθμούς ελευθερίας ($W_i \sim c^2_m$). Αν $E(\mathbf{X}) = \boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \boldsymbol{\delta}$, $\boldsymbol{\delta} \neq \mathbf{0}$, δηλώνει ένα εκτός ελέγχου διάνυσμα μέσων, το ίδιο στατιστικό ακολουθεί μη κεντρική κατανομή c^2 με m βαθμούς ελευθερίας ($W_i \sim c^2_m(l)$) και παράμετρο μη κεντρικότητας

$$l = l(\boldsymbol{\mu}_1) = n(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = n\boldsymbol{\delta}^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\delta}. \qquad (6.1).$$

Για να ξεχωρίζουμε την εντός από την εκτός ελέγχου κατάσταση, θα χρησιμοποιούμε στη συνέχεια το σύμβολο $D_i^2(l)$ για να δηλώνουμε το στατιστικό όταν το δείγμα προέρχεται από διεργασία που βρίσκεται εκτός ελέγχου και $D_i^2 = D_i^2(0)$ όταν προέρχεται από διεργασία που βρίσκεται εντός ελέγχου. Η συνάρτηση πυκνότητας (σ.π.) και η αθροιστική συνάρτηση κατανομής (α.σ.κ.) μιας μη κεντρικής c^2 κατανομής με m βαθμούς ελευθερίας και παράμετρο μη

κεντρικότητας l, θα συμβολίζονται με $f_m(x;l)$ και $F_m(x;l)$ αντίστοιχα. Οι αντίστοιχες σ.π. και α.σ.κ. μιας (κεντρικής) c^2 κατανομής θα συμβολίζονται με $f_m(x;0)=f_m(x)$ και $F_m(x;0)=F_m(x)$ αντίστοιχα. Οι ενδιαφερόμενοι αναγνώστες για τις παραπάνω κατανομές, παραπέμπονται στον Johnson et al. (1995). Χρησιμοποιώντας τους παραπάνω συμβολισμούς, τα ARL του CSCC παίρνουν τη μορφή

$$ARL_{in}=\frac{1}{\Pr(D_i^2>L_1)}=\frac{1}{1-F_m(L_1)},\ ARL_{out}=\frac{1}{\Pr(D_i^2(l)>L_1)}=\frac{1}{1-F_m(L_1;l)}.$$

Ας ξαναγυρίσουμε στο τυπικό Shewhart διάγραμμα ελέγχου με ένα άνω όριο ελέγχου L_1. Το κύριο μειονέκτημα των Shewhart διαγραμμάτων ελέγχου είναι η έλλειψη ευαισθησίας στην περίπτωση σταδιακής ή μικρής μετατόπισης της τιμής της εντός ελέγχου παραμέτρου. Το μειονέκτημα αυτό προέρχεται από το γεγονός ότι για να ληφθεί απόφαση λαμβάνεται υπόψη μόνο το πιο πρόσφατο δείγμα. Μια λύση είναι να παίρνουμε την απόφαση βασιζόμενοι στην πληροφορία που δίνει μια σειρά από δείγματα.

Μια λογική προσέγγιση είναι να θεωρήσουμε τροποποιημένα διαγράμματα ελέγχου που δηλώνουν μια εκτός ελέγχου διεργασία αν $k\geq 1$ συνεχόμενες τιμές της στατιστικής συνάρτησης υπερβαίνουν το άνω όριο ελέγχου, έστω L_k. Για $k=1$ ο παραπάνω κανόνας οδηγεί στο κλασικό διάγραμμα ελέγχου Shewhart χωρίς πρόσθετους κανόνες ευαισθησίας.

Για να μελετήσουμε το ARL_{in} του τροποποιημένου διαγράμματος ελέγχου ορίζουμε τις δίτιμες μεταβλητές

$$Y_i=\begin{cases}1, & W_i>L_k\\ 0, & W_i\leq L_k\end{cases}$$

για $i=1,2,\mathbf{K}$. Θεωρώντας ότι τα δείγματα είναι ανεξάρτητα και προέρχονται από την ίδια κατανομή, συμπεραίνουμε ότι τα $Y_1,Y_2,\mathbf{K}$ αποτελούν μια ακολουθία ανεξάρτητων δοκιμών Bernoulli με πιθανότητα επιτυχίας (αποτυχίας) $p=\Pr(W_i>L_k)$ $(q=1-p=\Pr(W_i\leq L_k))$. Έτσι το σημείο στο οποίο η διεργασία θα θεωρηθεί εκτός ελέγχου (ενώ στην πραγματικότητα είναι εντός ελέγχου), θα περιγράφεται από το χρόνο αναμονής T_k για την πρώτη εμφάνιση μιας ροής επιτυχιών μήκους k.

Η αναμενόμενη τιμή της T_k (βλέπε Balakrishnan and Koutras (2002) ή Fu and Lou (2003)) δίνεται από τον τύπο

$$E(T_k) = h_k(p) = \frac{1-p^k}{p(1-p)}$$

και είναι φθίνουσα συνάρτηση του p. Επιπλέον είναι φανερό ότι ισχύουν οι σχέσεις $\lim_{p\to 0} h_k(p) = +\infty$, $\lim_{p\to 1} h_k(p) = k$. Επομένως μπορούμε να πετύχουμε ένα προκαθορισμένο $ARL_{in} = c > k$ προσαρμόζοντας κατάλληλα το άνω όριο ελέγχου L_k του σχεδίου. Αυτό επιτυγχάνεται υπολογίζοντας πρώτα τη μοναδική ρίζα $p_k \in (0,1)$ της εξίσωσης

$$h_k(p_k) = \frac{1-(p_k)^k}{(p_k)^k(1-p_k)} = c\,,\ (c > k) \tag{6.2}$$

και βρίσκοντας στη συνέχεια το L_k με τη βοήθεια της συνθήκης $\Pr(W_i > L_k) = p_k$ (ο υπολογισμός της τελευταίας πιθανότητας γίνεται υπό την υπόθεση ότι η διεργασία είναι εντός ελέγχου).

Στη συνέχεια, το διάγραμμα που προκύπτει από την παραπάνω διεργασία θα αναφέρεται ως $k\,|\,k$ Shewhart διάγραμμα ελέγχου με $ARL_{in} = c$. Μια αλγοριθμική περιγραφή αυτού του διαγράμματος είναι:

Βήμα 1: Επιλέγουμε έναν θετικό ακέραιο αριθμό k.

Βήμα 2: Θέτουμε το επιθυμητό $ARL_{in} = c\,,\ c > k$.

Βήμα 3: Υπολογίζουμε τη μοναδική ρίζα p_k της εξίσωσης $c = \frac{1-(p_k)^k}{(p_k)^k(1-p_k)}$ στο διάστημα $(0,1)$.

Βήμα 4: Υπολογίζουμε το άνω όριο ελέγχου L_k με τη βοήθεια της ισότητας $p_k = \Pr(W_i > L_k)$.

Βήμα 5: Θεωρούμε τη διεργασία εκτός ελέγχου αν k συνεχόμενα σημεία απεικονίζονται πάνω από το L_k, δηλαδή αν στην εξέταση του i-οστού $(i \geq k)$ δείγματος, η ανισότητα $W_i > L_k$ ισχύει για όλα τα $j = i-k+1, i-k+2, \mathbf{K}, i$.

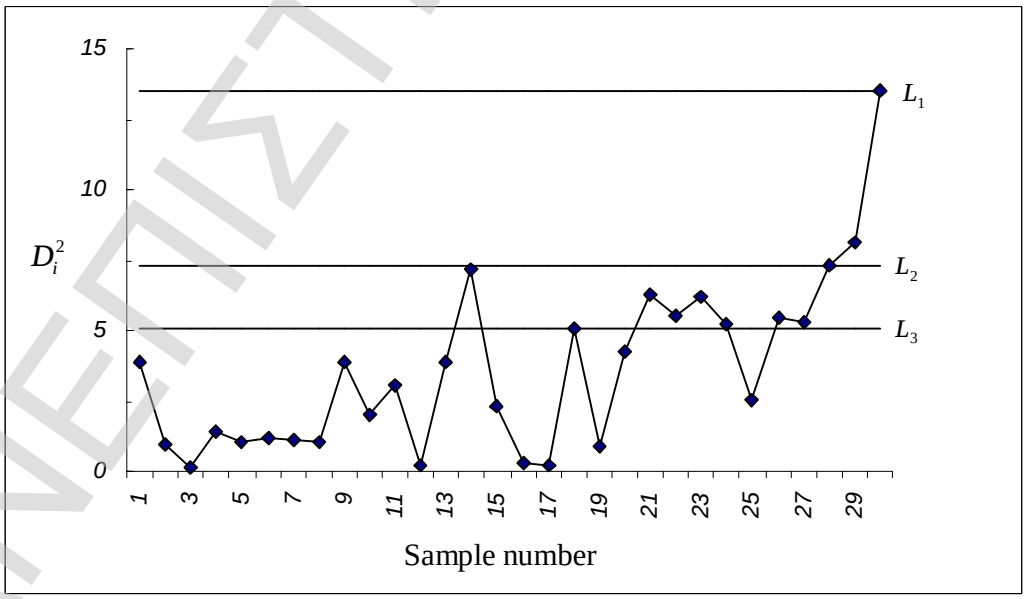
Αξίζει να σημειωθεί ότι το ARL_{out} του $k\,|\,k$ Shewhart διαγράμματος ελέγχου μπορεί επίσης να υπολογιστεί μέσω της $h_k(p)$ αντικαθιστώντας το p με την

πιθανότητα του ενδεχομένου $\{W_i > L_k\}$ υπό την υπόθεση ότι η παράμετρος της διεργασίας έχει μετατοπιστεί από την εντός ελέγχου τιμή της.

Ας εξετάσουμε τώρα την εφαρμογή του $k \mid k$ Shewhart διαγράμματος ελέγχου σε ένα CSCC. Στην περίπτωση αυτή έχουμε $ARL_{in} = E(T_k) = h_k(p_k)$ με την $h_k(\cdot)$ να δίνεται από τη σχέση (6.2) και $p_k = \Pr(D_i^2 > L_k) = 1 - F_m(L_k)$. Συνεπώς για να πετύχουμε ένα προκαθορισμένο ARL_{in}, έστω c, αρκεί να υπολογίσουμε τη μοναδική ρίζα $p_k \in (0,1)$ της εξίσωσης $h_k(p_k) = c$ και στη συνέχεια να θέσουμε $L_k = c^2_{m,p_k}$.

Ας θεωρήσουμε μια πολυμεταβλητή κανονική διεργασία με $m = 3$, πίνακα διακυμάνσεων-συνδιακυμάνσεων $\boldsymbol{\Sigma}_0$ και διάνυσμα μέσων $\boldsymbol{\mu}_0 = (100,100,100)$. Στο Σχήμα 6.4 έχουμε απεικονίσει τη στατιστική συνάρτηση ελέγχου $W_i = D_i^2$ για 30 διαφορετικές παρατηρήσεις από τη διεργασία, θεωρώντας ότι παραμένει εντός ελέγχου για τις πρώτες 20 μεμονωμένες παρατηρήσεις και μετατοπίζεται σε έναν εκτός ελέγχου μέσο $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \boldsymbol{\delta} = (105,105,105)$ στην 21^{η} παρατήρηση (ο πίνακας $\boldsymbol{\Sigma}_0$ παραμένει σταθερός για όλες τις παρατηρήσεις). Επίσης, έχουν απεικονιστεί τρία διαφορετικά όρια ελέγχου L_1, L_2, L_3 τα οποία αντιστοιχούν στα $1|1$, $2|2$, $3|3$ CSCC με κοινό $ARL_{in} = 250$.

Σχήμα 6.4: Τα 1|1, 2|2 και 3|3 CSCC



Από το Σχήμα 6.4, παρατηρούμε ότι το τυπικό $1|1$ CSCC δίνει ένα εκτός ελέγχου σήμα στο 30^{o} δείγμα ενώ τα $2|2$ και $3|3$ CSCC δίνουν σήμα γρηγορότερα στα δείγματα 29 και 23 αντίστοιχα.

Στην επόμενη υποπαράγραφο θα δώσουμε θεωρητικά αποτελέσματα που αφορούν τη συμπεριφορά και εμφάνιση του $k|k$ CSCC.

6.6.3. Ιδιότητες του *k*|*k* CSCC

Για τα άνω όρια ελέγχου L_1, L_2, L_3 των τριών CSCC του προηγούμενου παραδείγματος ισχύει $L_1 > L_2 > L_3$ (βλέπε Σχήμα 6.4). Μια τέτοια διάταξη ισχύει γενικά δοθέντος ότι το επίπεδο του ARL_{in} διατηρείται σταθερό. Για να το αποδείξουμε, έστω r, k δυο θετικοί ακέραιοι αριθμοί τέτοιοι ώστε $r < k$ και τα αντίστοιχα $r|r$ και $k|k$ CSCC καθένα από τα οποία έχει το ίδιο $ARL_{in} = c$. Μπορεί εύκολα να ελεγχθεί ότι ισχύει $h_r(x) < h_k(x)$ για όλα τα $x \in (0,1)$ και συνεπώς οι μοναδικές ρίζες p_r και p_k των εξισώσεων $c = h_r(p_r)$, $c = h_k(p_k)$, $(r < k)$, $c > k$ ικανοποιούν την ανισότητα $p_r < p_k$ που στη συνέχεια οδηγεί στην ανισότητα $L_r > L_k$. Επακόλουθο του προηγούμενου αποτελέσματος είναι ότι το άνω όριο ελέγχου L_k του $k|k$ CSCC με $k \geq 2$ είναι μικρότερο από το άνω όριο ελέγχου L_1 του τυπικού CSCC με το ίδιο ARL_{in}.

Στη συνέχεια θα εξετάσουμε πώς μπορεί να υπολογιστεί το ARL_{out} ενός $k|k$ CSCC. Όπως έχει ήδη αναφερθεί, το ARL_{out} είναι ο μέσος της τυχαίας μεταβλητής του χρόνου αναμονής T_k υπό την υπόθεση ότι η παράμετρος ελέγχου $\boldsymbol{\mu}$ έχει μετατοπιστεί σε μια εκτός ελέγχου τιμή $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \boldsymbol{\delta}$. Συνεπώς,

$$ARL_{out}(l) = h_k(p_k(l)) = \frac{1-(p_k(l))^k}{(p_k(l))^k(1-p_k(l))},$$

όπου $p_k(l) = \Pr(D_i^2(l) > L_k) = 1 - F_m(L_k; l)$ και η παράμετρος $l = l(\boldsymbol{\mu}_1)$ δίνεται από τη σχέση (6.1)

Στη συνέχεια υιοθετούμε το συμβολισμό $ARL_k(l)$ για το ARL_{out} ενός $k|k$ CSCC ενώ το αντίστοιχο $ARL_{in} = ARL_k(0)$ θα συμβολίζεται απλά με ARL_k, $k \geq 1$. Μπορούμε να γράψουμε

$$ARL_k(l) = \frac{1-[1-F_m(L_k;l)]^k}{F_m(L_k;l)[1-F_m(L_k;l)]^k} = \frac{1}{1-H_k(F_m(L_k;l))} \tag{6.3}$$

όπου

$$H_k(x) = \frac{1-(1+x)(1-x)^k}{1-(1-x)^k} = 1 - \frac{x(1-x)^k}{1-(1-x)^k}.$$

Έπειτα στρέφουμε την προσοχή μας στο πρόβλημα της σύγκρισης δυο $k|k$ CSCC με διαφορετικά k. Ο κύριος στόχος μας είναι, για οποιαδήποτε δυο $r|r$, $k|k$ CSCC με $r<k$ και κοινό $ARL_{in} = c$ (δηλαδή $ARL_r = ARL_k = c$), ο προσδιορισμός συνθηκών για το l που να εξασφαλίζουν ότι $ARL_r(l) < ARL_k(l)$ ή $(ARL_r(l) > ARL_k(l))$. Αν ισχύει μια τέτοια συνθήκη για όλα τα $l > 0$, τότε κάποιο από τα CSCC θα είναι προτιμότερο από τα υπόλοιπα, χωρίς να μας ενδιαφέρει πόσο απέχει η εκτός ελέγχου τιμή από την αντίστοιχη εντός ελέγχου. Ιδιαίτερο ενδιαφέρον παρουσιάζει η περίπτωση $r=1$ που οδηγεί σε συμπεράσματα για τη σύγκριση ενός $k|k$ CSCC με το τυπικό CSCC.

Παρατηρούμε πρώτα ότι η σχέση (6.3) οδηγεί στην έκφραση

$$ARL_r(l) - ARL_k(l) = \frac{S_{r,k}(l)}{[1-H_r(F_m(L_r;l))][1-H_k(F_m(L_k;l))]} \tag{6.4}$$

όπου $S_{r,k}(l)$ συμβολίζει τη διαφορά

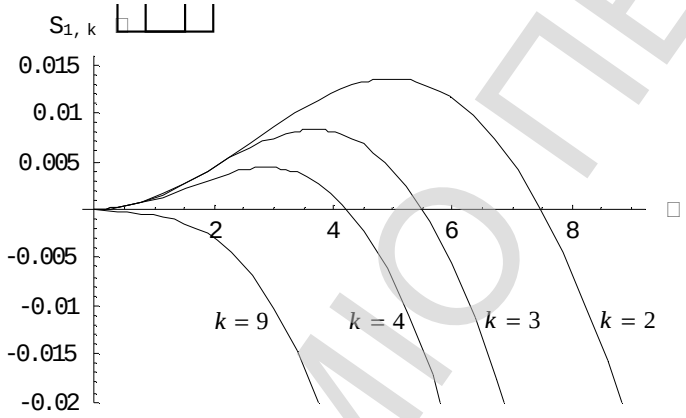
$$S_{r,k}(l) = H_r(F_m(L_r;l)) - H_k(F_m(L_k;l)). \tag{6.5}$$

Επειδή η ποσότητα $H_k(x)$ είναι μικρότερη του 1 για όλα τα $0<x<1$ και για οποιοδήποτε θετικό ακέραιο αριθμό k, από τη σχέση (6.4) προκύπτει ότι ο παρονομαστής της διαφοράς $ARL_r(l) - ARL_k(l)$ είναι πάντα θετικός και συνεπώς το πρόσημο της $ARL_r(l) - ARL_k(l)$ συμπίπτει με το πρόσημο της $S_{r,k}(l)$.

Στο Σχήμα 6.5, απεικονίζεται η ποσότητα $S_{1,k}(l)$ για $m=5$, $ARL_{in} = c = 200$ και διάφορες επιλογές για το $k \geq 2$. Από αυτά τα διαγράμματα συμπεραίνουμε ότι στις περισσότερες περιπτώσεις δεν υπάρχει ομοιόμορφη διάταξη μεταξύ των $ARL_k(l)$, $k \geq 2$ και $ARL_1(l)$. Είναι επίσης προφανές ότι για μικρές τιμές του k το

διάστημα (αν υπάρχει) του l στο οποίο το $k|k$ CSCC έχει καλύτερη συμπεριφορά από το $1|1$ CSCC μεγαλώνει ενώ για μεγάλες τιμές του l το $1|1$ CSCC παρουσιάζει καλύτερη συμπεριφορά από το $k|k$ CSCC. Αξίζει να σημειωθεί ότι και στη γενικότερη περίπτωση που θέλουμε να συγκρίνουμε τα $k|k$ με τα $r|r$ CSCC, τα διαγράμματα των $S_{r,k}(l)$ για διάφορες επιλογές των $r,k \geq 2$ είναι περίπου τα ίδια δοθέντος ότι η διάταξη $r < k$ ισχύει.

Σχήμα 6.5: Διαγράμματα της $S_{r,k}(l)$ για k=2, 3, 4, 9 και r=1



Οι παρακάτω προτάσεις δίνουν την δυνατότητα επιλογής του καλύτερου ανά περίπτωση CSCC με έναν πιο αυστηρό τρόπο. Το πρώτο αποτέλεσμα δηλώνει ότι για μεγάλες τιμές της παραμέτρου l το $r|r$ CSCC λειτουργεί πάντα καλύτερα από το $k|k$ CSCC $(r < k)$.

Πρόταση 6.1: *Έστω r, k δυο θετικοί ακέραιοι τέτοιοι ώστε $r < k$. Υπάρχει ένας πραγματικός αριθμός l_0 τέτοιος ώστε τα εκτός ελέγχου ARL των $r|r$ CSCC και $k|k$ CSCC με την ίδια τιμή c για το εντός ελέγχου ARL $(ARL_r = ARL_k = c)$ ικανοποιούν την ανισότητα $ARL_r(l) < ARL_k(l)$ για όλα τα $l > l_0$.*

Απόδειξη: Μπορεί εύκολα να αποδειχθεί ότι για κάθε θετικό ακέραιο k, έχουμε

$$\lim_{x \to 0} H_k(x) = 1 - \frac{1}{k}.$$

Συνεπώς, από το γνωστό αποτέλεσμα (βλέπε Johnson et al. (1995))

$$\lim_{l\to\infty} F_k(x;\, l) = 0$$

μπορούμε εύκολα να συμπεράνουμε ότι

$$\lim_{l\to\infty}(ARL_r(l) - ARL_k(l)) = rk \lim_{l\to\infty}(S_{r,k}(l)) = r - k < 0$$

που ολοκληρώνει την απόδειξη. □

Ένα άμεσο επακόλουθο της Πρότασης 6.1 είναι ότι για μεγάλες τιμές της παραμέτρου l (δηλαδή μεγάλες μετατοπίσεις από την τιμή της εντός ελέγχου παραμέτρου) το τυπικό CSCC $(r=1)$ είναι ανώτερο του $k\,|\,k$ CSCC με $k \geq 2$. Υπάρχει άλλη μια περίπτωση όπου το τυπικό CSCC είναι καλύτερο από το $k\,|\,k$ CSCC με $k \geq 2$. Ειδικότερα θα αποδείξουμε ότι αν η προκαθορισμένη τιμή c του ARL_{in} είναι αρκετά μικρή, τότε η καμπύλη του ARL_{out} για τον $1\,|\,1$ CSCC είναι ομοιόμορφα καλύτερη από την αντίστοιχη καμπύλη $k\,|\,k$ CSCC με $k \geq 2$. Πριν αποδείξουμε αυτόν τον ισχυρισμό, δίνουμε ένα χρήσιμο λήμμα που θα βοηθήσει στην διαδικασία απόδειξης.

Λήμμα 6.1:

(α) Η ποσότητα $S_{1,k}(l)$ είναι μια φθίνουσα συνάρτηση της παραμέτρου $l > 0$ αν και μόνο αν ισχύει η ανισότητα

$$H_k'(F_m(L_k;\, l)) < \frac{f_{m+2}(L_1;\, l)}{f_{m+2}(L_k;\, l)}. \qquad (6.6)$$

(β) Αν $k \geq 2$ έχουμε

$$H_k'(x) < 1$$

για όλα τα $0 < x < 1$.

Απόδειξη:

(α) Η γνωστή σχέση για την αθροιστική συνάρτηση κατανομής της μη κεντρικής κατανομής c^2

$$\frac{\partial}{\partial l}(F_m(L_k;\, l)) = -f_{m+2}(x;\, l),$$

(βλέπε Johnson et al. (1995)) μπορεί να χρησιμοποιηθεί σε συνδυασμό με τον κανόνα της αλυσίδας για να γράψουμε

$$\frac{\partial}{\partial l} H_k(F_m(L_k; l)) = - f_{m+2}(L_k; l) H_k'(x)\big|_{x=F_m(L_k;l)}.$$

Παραγωγίζοντας την (6.5) (για $r=1$) ως προς l παίρνουμε

$$\begin{aligned} S_{1,k}'(l) &= \frac{\partial}{\partial l}[H_r(F_m(L_r; l)) - H_k(F_m(L_k; l))] \\ &= f_{m+2}(L_k; l) H_k'(F_m(L_k; l)) - f_{m+2}(L_r; l) H_r'(F_m(L_r; l)) \end{aligned}$$

και λαμβάνοντας υπόψη ότι $H_1'(x)=1$ για όλα τα $0<x<1$, παίρνουμε την παρακάτω έκφραση για την $S_{1,k}'(l)$

$$S_{1,k}'(l) = f_{m+2}(L_k; l)\left(H_k'(F_m(L_k; l)) - \frac{f_{m+2}(L_1; l)}{f_{m+2}(L_k; l)} \right). \tag{6.7}$$

Λαμβάνοντας υπόψη την τελευταία έκφραση, η συνθήκη (6.6) δίνει ότι $S_{1,k}'(l)<0$, το οποίο εξασφαλίζει ότι η $S_{1,k}(l)$ είναι μονότονα φθίνουσα (και αντίστροφα). Έτσι ολοκληρώνεται η απόδειξη του (α).

(β) Έστω $k \geq 2$. Αφού ισχύει

$$H_k'(x) = \frac{(1-x)^{k-1}[(1-x)^{k-1} + x(k+1) - 1]}{(1-(1-x)^k)^2}$$

η συνθήκη $H_k'(x)<1$ για $0<x<1$ είναι ισοδύναμη με την

$$(1-x)^{k-1}[x(k-1)+1] < 1.$$

Η τελευταία ανισότητα ισχύει για $k=2$ και μπορεί να ελεγχθεί με επαγωγή ότι ισχύει και για κάθε θετικό ακέραιο $k>2$. Αυτό ολοκληρώνει την απόδειξη του (β). □

Είμαστε τώρα έτοιμοι να αποδείξουμε την ανωτερότητα του τυπικού CSCC έναντι του $k|k$ CSCC με $k \geq 2$ για μικρές ARL_{in} τιμές.

Πρόταση 6.2: *Έστω c ένας θετικός ακέραιος τέτοιος ώστε $c^2_{m,1/c} \leq m$. Τότε για όλα τα $k \geq 2$ τα εκτός ελέγχου ARL των $1|1$ CSCC και $k|k$ CSCC με την ίδια τιμή c για το εντός ελέγχου ARL ($ARL_r = ARL_k = c$) ικανοποιούν την ανισότητα*

$$ARL_1(l) < ARL_k(l)$$

για όλα τα $l>0$.

Απόδειξη: Έστω M_l η κορυφή (mode) της κατανομής $c^2_m(l)$. Η σ.π.π. $f_{m+2}(x; l)$ είναι αύξουσα για τα $x < M_l$ ενώ η M_l είναι αύξουσα συνάρτηση του l (βλέπε

Johnson et al. (1995)). Επιπλέον η κορυφή της κεντρικής κατανομής c^2 με $m+2$ βαθμούς ελευθερίας ισούται με $m = M_0$ ενώ η $f_{m+2}(x; l)$ είναι αύξουσα συνάρτηση του x για $x < m = M_0 < M_x$. Παρατηρώντας ότι η συνθήκη $c^2_{m,1/c} \le m$ οδηγεί στο $L_1 < m$ και χρησιμοποιώντας τη μονοτονία της $f_{m+2}(x; l)$ ως προς x καταλήγουμε στην ανισότητα $\frac{f_{m+2}(L_1; l)}{f_{m+2}(L_k; l)} > 1$ (ισχύει επίσης $L_k < L_1$). Χρησιμοποιώντας το Λήμμα 3.1 (β) παίρνουμε ότι $H_k'(F_m(L_k; l)) < 1 < \frac{f_{m+2}(L_1; l)}{f_{m+2}(L_k; l)}$ και συνεπώς η $S_{1,k}(l)$ είναι φθίνουσα ως προς l. Το επιθυμητό αποτέλεσμα μπορεί τώρα να προκύψει εύκολα αν λάβουμε υπόψη ότι η $S_{1,k}(l)$ είναι φθίνουσα συνάρτηση του l με, $\lim_{l \to 0} S_{1,k}(l) = \frac{1}{k} - 1 < 0$ (σημειώνουμε ότι $\lim_{l \to \infty} F_k(x; l) = 0$) και ότι το πρόσημο της $S_{1,k}(l)$ συμπίπτει με αυτό της διαφοράς $ARL_1(l) - ARL_k(l)$. □

Η συνθήκη $c^2_{m,1/c} \le m$ ικανοποιείται πρακτικά για πολύ μικρές τιμές του c (κοντά στο 2). Τιμές του c τόσο μικρές δεν έχουν πρακτική σημασία στον στατιστικό έλεγχο ποιότητας. Πρέπει επίσης να τονισθεί ότι η συνθήκη $c^2_{m,1/c} \le m$ δεν είναι αναγκαία και ικανή και συνεπώς κάποιος μπορεί να υποπτευθεί ότι η υπεροχή του τυπικού CSCC έναντι του $k \mid k$ CSCC με $k \ge 2$ μπορεί να εμφανιστεί ακόμα και για μεγάλες τιμές του c. Παρόλα αυτά, όπως έδειξε και η αριθμητική μελέτη, η κατάσταση αυτή δεν εμφανίζεται για τιμές του c που έχουν πρακτική σημασία $(c \ge 200)$.

Στην συνέχεια θα μελετήσουμε περαιτέρω τη συμπεριφορά της διαφοράς $ARL_1(l) - ARL_k(l)$ για μικρές τιμές του l και / ή μεγάλες ARL_{in} τιμές $k \mid k$ CSCC, προκειμένου να κατασκευάσουμε κατάλληλα $k \mid k$ CSCC με $k \ge 2$ τα οποία θα βελτιώνουν το τυπικό CSCC. Τα επόμενα δυο αποτελέσματα οδηγούν προς αυτή την κατεύθυνση.

Πρόταση 6.3: *Αν για προκαθορισμένο* $k \ge 2$ *ισχύει η ανισότητα*

$$H_k'(F_m(L_k)) > \left(\frac{L_1}{L_k}\right)^{m/2} \exp[(L_k - L_1)/2] \qquad (6.8)$$

τότε υπάρχει ένας θετικός πραγματικός αριθμός l_0 *τέτοιος ώστε*

$$ARL_1(l) > ARL_k(l)$$

για όλα τα $l < l_0$.

Απόδειξη: Εφαρμόζοντας τη σχέση (6.7) για $l = 0$ και αντικαθιστώντας την $f_{m+2}(x;0)$ με

$$f_{m+2}(x) = \frac{1}{2^{(m/2)+1}\Gamma((m/2)+1)} x^{m/2} e^{-x/2}$$

έχουμε

$$\begin{aligned} S'_{1,k}(0) &= f_{m+2}(L_k)\left(H'_k(F_m(L_k)) - \frac{f_{m+2}(L_1)}{f_{m+2}(L_k)} \right) \\ &= f_{m+2}(L_k)\left(H'_k(F_m(L_k)) - \left(\frac{L_1}{L_k}\right)^{m/2} \exp[(L_k - L_1)/2] \right) \end{aligned}$$

και χρησιμοποιώντας τη συνθήκη (6.8) συμπεραίνουμε ότι $S'_{1,k}(0) > 0$. Αφού η συνάρτηση $S'_{1,k}(l)$ είναι συνεχής, θα υπάρχει ένα διάστημα της μορφής $(0, l_0)$, $l_0 > 0$ τέτοιο ώστε για $l \in (0, l_0)$ να ισχύει $S'_{1,k}(1) > 0$. Επομένως η συνάρτηση $S_{1,k}(l)$ είναι αύξουσα συνάρτηση του l. Αφού $S_{1,k}(0) = 0$, θα ισχύει $S_{1,k}(l) > 0$ για $l \in (0, l_0)$ το οποίο σημαίνει ότι $ARL_1(l) > ARL_k(l)$ για όλα τα $l \in (0, l_0)$. Αυτό ολοκληρώνει την απόδειξη. □

Πόρισμα 6.1: *Έστω* c *ένας θετικός αριθμός τέτοιος ώστε*

$$1 - \frac{1}{(1+s)^2} > \left(\frac{c^2_{m,1/c}}{c^2_{m,s}} \right)^{m/2} \exp[(c^2_{m,s} - c^2_{m,1/c})/2]$$

όπου $s = (1+\sqrt{1+4c})/2c$. *Τότε υπάρχει πραγματικός θετικός αριθμός* l_0 *τέτοιος ώστε*

$$ARL_1(l) > ARL_2(l)$$

για όλα τα $l < l_0$.

Απόδειξη: Εφαρμόζοντας την Πρόταση 6.3 για $k = 2$ έχουμε την απόδειξη του πορίσματος. □

Το προφανές συμπέρασμα της παραπάνω ανάλυσης είναι ότι αν κάποιος επιθυμεί να εργαστεί με μεγάλες τιμές του ARL_{in} (όπως συνήθως συμβαίνει) και/ή περιμένει

μόνο μικρές μετατοπίσεις της παραμέτρου ελέγχου από το εντός ελέγχου επίπεδο, θα πρέπει να χρησιμοποιήσει ένα $k \mid k$ CSCC με $k \geq 2$ αντί του τυπικού CSCC. Στην αντίθετη περίπτωση το τυπικό CSCC είναι καλύτερη επιλογή.

6.6.4. Το *r|r-k|k* CSCC

Μια τεχνική που χρησιμοποιείται συχνά στον εφαρμοσμένο έλεγχο ποιότητας όταν είναι διαθέσιμες δυο μη ομοιόμορφα διατεταγμένες διαδικασίες, είναι να θεωρήσουμε ένα συνδυασμένο διάγραμμα που εκμεταλλεύεται τα όρια ελέγχου και των δυο διαδικασιών.

Θεωρούμε ένα τυπικό διάγραμμα ελέγχου Shewhart με ένα άνω και ένα κάτω όριο ελέγχου. Στο κλασικό διάγραμμα ελέγχου Shewhart μια στατιστική συνάρτηση ελέγχου W_i απεικονίζεται στο διάγραμμα και παράγεται ένα εκτός ελέγχου σήμα όταν ένα σημείο πέφτει έξω από τα όρια ελέγχου. Μια προφανής τροποποίηση αυτού του κανόνα είναι να δίνει εκτός ελέγχου τη διεργασία αν είτε k συνεχόμενα σημεία απεικονίζονται πάνω από το άνω όριο ελέγχου είτε r συνεχόμενα σημεία απεικονίζονται χαμηλότερα από το κάτω όριο ελέγχου. Αυτό το είδος διαγράμματος θα ονομάζεται «$r \mid r-k \mid k$ Shewhart διάγραμμα ελέγχου».

Στη συνέχεια εισάγουμε το απαιτούμενο μαθηματικό υπόβαθρο για τη μελέτη του $r \mid r-k \mid k$ Shewhart διαγράμματος ελέγχου. Θεωρούμε μια ακολουθία ανεξάρτητων δοκιμών $Y_1, Y_2, \mathbf{K}$ με τρία πιθανά αποτελέσματα, έστω 0, 1, 2 και υποθέτουμε ότι $q = \Pr(Y_i = 0)$, $p_U = \Pr(Y_i = 1)$ και $p_L = \Pr(Y_i = 2)$ για $i \geq 1$. Συμβολίζουμε με $T_{r,k}$ το χρόνο αναμονής μέχρι την πρώτη εμφάνιση μιας ροής από 1 μήκους k ή μιας ροής από 2 μήκους r, όποια και αν εμφανιστεί πρώτη. Η τυχαία μεταβλητή του χρόνου αναμονής $T_{r,k}$ έχει μελετηθεί από τους Koutras and Alexandrou (1997) (βλέπε επίσης Aki and Hirano (1993) και Han and Aki (2000)) οι οποίοι εξέφρασαν την πιθανογεννήτρια συνάρτησή της ως

$$E[z^{T_{r,k}}] = \frac{(p_U z)^k G_L(z) + (p_L z)^r G_U(z)}{G_L(z) + G_U(z) - G_L(z) G_U(z)(1+qz)}$$

όπου

$$G_L(z)=\frac{1-(p_L z)^r}{1-p_L z},\ G_U(z)=\frac{1-(p_U z)^k}{1-p_U z}.$$

Η μέση τιμή της $T_{r,k}$ δίνεται από

$$E[T_{r,k}]=h_{r,k}(p_L,p_U)=\frac{1}{C}\left(A+\frac{B}{C}\right) \tag{6.10}$$

όπου

$$\begin{aligned}
A &= p_U^k[k-(r+k)p_L^r](1-p_U)-p_U^{k+1}(1-p_L^r)\\
&\quad + p_L^r[r-(r+k)p_U^k](1-p_L)-p_L^{r+1}(1-p_U^k),\\
B &= [p_U^k(1-p_L^r)(1-p_U)+p_L^r(1-p_U^k)(1-p_L)]\times\\
&\quad [1-(k+1)p_U^k(1-p_U)-(r+1)p_L^r(1-p_L)+p_U^k p_L^r[(r+k)(1+q)+q]],\\
C &= p_U^k(1-p_U)+p_L^r(1-p_L)-p_U^k p_L^r(1+q).
\end{aligned}$$

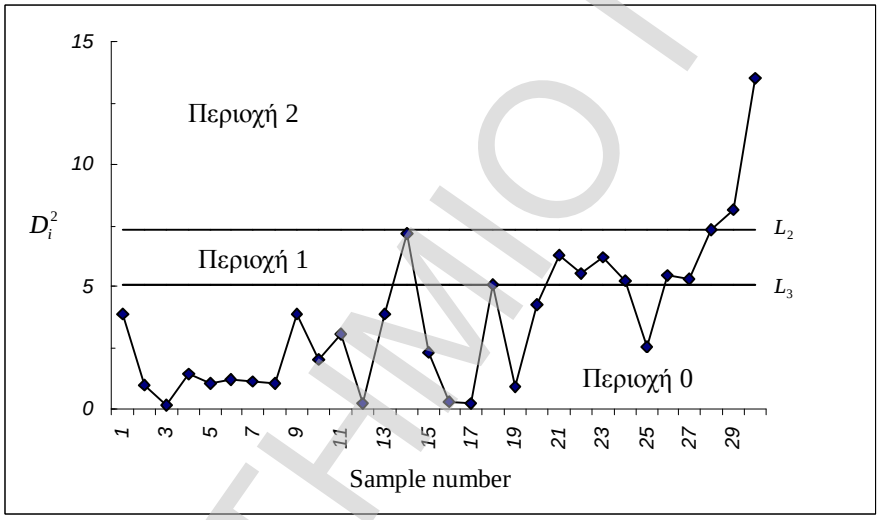
Εφαρμόζοντας την παραπάνω σχέση για $p_U=\Pr(W_i>UCL)$, $p_L=\Pr(W_i<LCL)$, $q=1-p_U-p_L$ μπορούμε να πάρουμε το ARL του $r\,|\,r-k\,|\,k$ Shewhart διαγράμματος ελέγχου. Για το τυπικό $\overline{X}$ διάγραμμα ελέγχου Shewhart, οι εδικές περιπτώσεις $k=r=2$ και $p_U=p_L$ έχουν μελετηθεί από τον Klein (2000).

Στη συνέχεια θα δείξουμε πώς η βασική αρχή του $r\,|\,r-k\,|\,k$ Shewhart διαγράμματος ελέγχου μπορεί να χρησιμοποιηθεί σε ένα CSCC. Θεωρούμε πρώτα δυο $r\,|\,r$ και $k\,|\,k$ CSCC $(r<k)$ με κοινό $ARL_{in}=c$ $(ARL_r=ARL_k=c)$ και αντίστοιχα όρια ελέγχου L_r και L_k $(L_r>L_k)$. Ο όρος $r\,|\,r-k\,|\,k$ χι-τετράγωνο διάγραμμα ελέγχου (το οποίο θα συμβολίζεται με $r\,|\,r-k\,|\,k$ CSCC) θα χρησιμοποιηθεί για να δηλώνει ένα διάγραμμα ελέγχου το οποίο δίνει την διεργασία εκτός ελέγχου αν είτε k συνεχόμενες τιμές του στατιστικού ελέγχου $W_i=D_i^2$ απεικονίζονται πάνω από το L_k είτε r συνεχόμενες τιμές απεικονίζονται κάτω από το L_r. Σε ένα $r\,|\,r-k\,|\,k$ CSCC ορίζονται τρεις περιοχές: μια αποτελούμενη από τα σημεία κάτω από το όριο ελέγχου L_k (περιοχή 0), μια που περιέχει τα σημεία πάνω από το όριο ελέγχου L_r (περιοχή 2) και μια κεντρική περιοχή μεταξύ των δυο ορίων (περιοχή 1). Για μια εντός ελέγχου διεργασία, η πιθανότητα ότι ένα σημείο πέφτει στις περιοχές 0, 1 και 2 είναι $1-p_k$, p_k-p_r και p_r αντίστοιχα, όπου $p_k=\Pr(D_i^2>L_k)$, $p_r=\Pr(D_i^2>L_r)$, ενώ για μια εκτός ελέγχου διεργασία, οι

αντίστοιχες πιθανότητες γίνονται $1-p_k(l)$, $p_k(l)-p_r(l)$ και $p_r(l)$, όπου $p_k(l)=\Pr(D_i^2(l)>L_k)$, $p_r(l)=\Pr(D_i^2(l)>L_r)$.

Ως παράδειγμα θεωρούμε πάλι τη διεργασία του Σχήματος 6.4 και θέτουμε τα όρια L_2 και L_3 όπως φαίνεται στο Σχήμα 6.6. Η χρήση του $2|2-3|3$ CSCC για την παρακολούθηση του μέσου της διεργασίας θα δώσει ένα εκτός ελέγχου σήμα όταν 2 συνεχόμενα τιμές του στατιστικού ελέγχου D_i^2 υπερβαίνουν το L_2 (περιοχή 2) ή 3 συνεχόμενες τιμές του D_i^2 υπερβαίνουν το L_2 (περιοχές 1 και 2). Για τα δεδομένα που απεικονίζονται στο Σχήμα 6.6 αυτό συμβαίνει στο δείγμα 23.

Σχήμα 6.6: Το 2|2-3|3 CSCC



Είναι φανερό ότι, το *ARL* του συνδυασμένου διαγράμματος συμπίπτει με τη μέση τιμή του χρόνου αναμονής για την εμφάνιση μιας ροής από 2 μήκους r ή την εμφάνιση μιας σειράς k συνεχόμενων δοκιμών αποτελούμενων από 1 ή 2, οποιαδήποτε και αν εμφανιστεί πρώτη. Αν και υπάρχουν στη βιβλιογραφία μερικά εργαλεία για τη μελέτη τυχαίων μεταβλητών χρόνου αναμονής αυτής της μορφής (βλέπε Aki (1992), Aki et al. (1996), Fu and Chang (2003), Antzoulakos (2001) και Koutras (1997b)), δεν θα ασχοληθούμε με αυτό το γενικό σχήμα. Αντίθετα θα περιοριστούμε στην ειδική περίπτωση $r=1$ ($1|1-k|k$ CSCC) του οποίου το *ARL* μπορεί να υπολογιστεί μέσω της σχέσης (6.10) αν αντικαταστήσουμε τα p_U και p_L με τα p_k-p_1 και p_1 αντίστοιχα (βλέπε και Page (1955)). Ειδικότερα, αν

συμβολίσουμε τα ARL_{out} και ARL_{in} του $1|1-k|k$ CSCC με $ARL_{1,k}(l)$ και $ARL_{1,k} = ARL_{1,k}(0)$ αντίστοιχα μπορούμε να γράψουμε $ARL_{1,k}(l) = h_{1,k}(p_1(l), p_k(l) - p_1(l))$, $ARL_{1,k} = h_{1,k}(p_1, p_k - p_1)$.

Ένα ενδιαφέρον σημείο είναι ότι, για την κατασκευή ενός $1|1-k|k$ CSCC μπορούμε να χρησιμοποιήσουμε διαφορετικές τιμές για τα εντός ελέγχου ARL των ανεξάρτητων $1|1$ και $k|k$ CSCC, έστω $c_1 = ARL_1$ και $c_k = ARL_k$, εφ' όσον βέβαια ισχύει η διάταξη $L_1 > L_k$ (η συνθήκη $k < c_k < h_k(1/c_1)$ εξασφαλίζει αυτή τη διάταξη). Μια αλγοριθμική περιγραφή δίνεται παρακάτω:

Βήμα 1: Επιλέγουμε έναν θετικό ακέραιο αριθμό $k \geq 2$.

Βήμα 2: Θέτουμε τα επιθυμητά $c_1 = ARL_1$ και $c_k = ARL_k$ για τα ανεξάρτητα $1|1$ και $k|k$ CSCC $(k < c_k < h_k(1/c_1))$ και υπολογίζουμε τα αντίστοιχα όρια ελέγχου L_1 και L_k.

Βήμα 3: Θεωρούμε τη διεργασία εκτός ελέγχου αν k συνεχόμενα σημεία απεικονίζονται χαμηλότερα από το L_k ή ένα σημείο απεικονίζεται πάνω από το L_1.

Κάποιος μπορεί να χρησιμοποιήσει το $\min\{c_1, c_k\}$ ως αδρό εκτιμητή του $ARL_{1,k}$ του $1|1-k|k$ CSCC. Στην πραγματικότητα, αυτό είναι ένα άνω φράγμα για το $ARL_{1,k}$. Παρόλα αυτά, μια καλύτερη προσέγγιση του $ARL_{1,k}$ δίνεται μέσω του κλασικού τύπου

$$\frac{1}{ARL_{1,k}} \cong \frac{1}{ARL_1} + \frac{1}{ARL_k} = \frac{1}{c_1} + \frac{1}{c_k} \tag{6.11}$$

που χρησιμοποιείται πολύ συχνά στον έλεγχο ποιότητας.

Πίνακας 6.2: Υπολογισμοί του ARL

	ARL			$ARL_{1,k}$	
k	ARL_1	ARL_k	$\min(ARL_1, ARL_k)$	Σχέση 6.11	Ακριβές
2	600	600	600	300	312
3	600	600	600	300	306
4	600	600	600	300	304
2	300	200	200	120	127
2	200	300	200	120	128
3	400	300	300	171	176
3	300	400	300	171	176
4	500	200	200	143	145
4	200	500	200	143	146
2	970	970	970	485	500
3	987	987	987	493	500
4	990	990	990	495	500

Η ανωτερότητα του παραπάνω τύπου (6.11) φαίνεται στον παραπάνω Πίνακα 6.2. Από τη σχέση (6.11), προκύπτει ότι, αν κάποιος επιθυμεί να δουλέψει με προκαθορισμένο $ARL_{1,k} = c$, θα μπορούσε απλά να υπολογίσει τις τιμές των ορίων ελέγχου L_1, L_k θεωρώντας τα ανεξάρτητα $1|1$ και $k|k$ CSCC με $ARL_1 = ARL_k = 2c$.

6.6.5. Αριθμητικές συγκρίσεις

Στην παρούσα υποπαράγραφο διευκρινίζουμε μερικά προβλήματα πρακτικής σημασίας. Ειδικότερα, διαμορφώνουμε, με τη βοήθεια αριθμητικών μεθόδων, μερικές διαδικασίες που διευκολύνουν την επιλογή του καταλληλότερου σχεδίου ελέγχου. Θεωρούμε πρώτα ότι θέλουμε να εργαστούμε με ένα $k|k$ CSCC και το πρόβλημα είναι ποια τιμή του k να χρησιμοποιήσουμε. Εφόσον δεν υπάρχει ομοιόμορφη διάταξη μεταξύ των $ARL_k(l)$ και $ARL_1(l)$, χρειαζόμαστε ένα κοινό διάγραμμα των

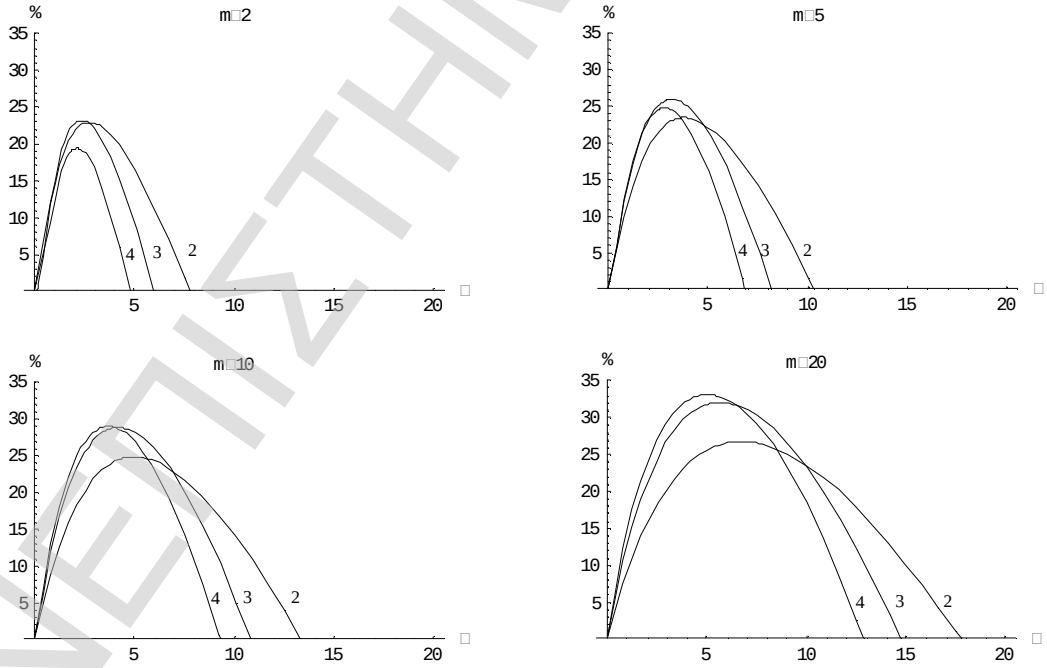
δυο ποσοτήτων (ως προς το l) καθώς και επιπλέον πληροφορία για το μέγεθος των μετατοπίσεων που είναι πιθανόν να εμφανιστούν.

Στο Σχήμα 6.7 έχουμε απεικονίσει το ποσοστό βελτίωσης του ARL_{out} διαφόρων $k \mid k$ CSCC έναντι του τυπικού CSCC $(k = 1)$, υπό το ίδιο $ARL_{in} = 500$. Η καμπύλη που αντιστοιχεί στο $k = 4$ είναι η γραφική παράσταση της ποσότητας

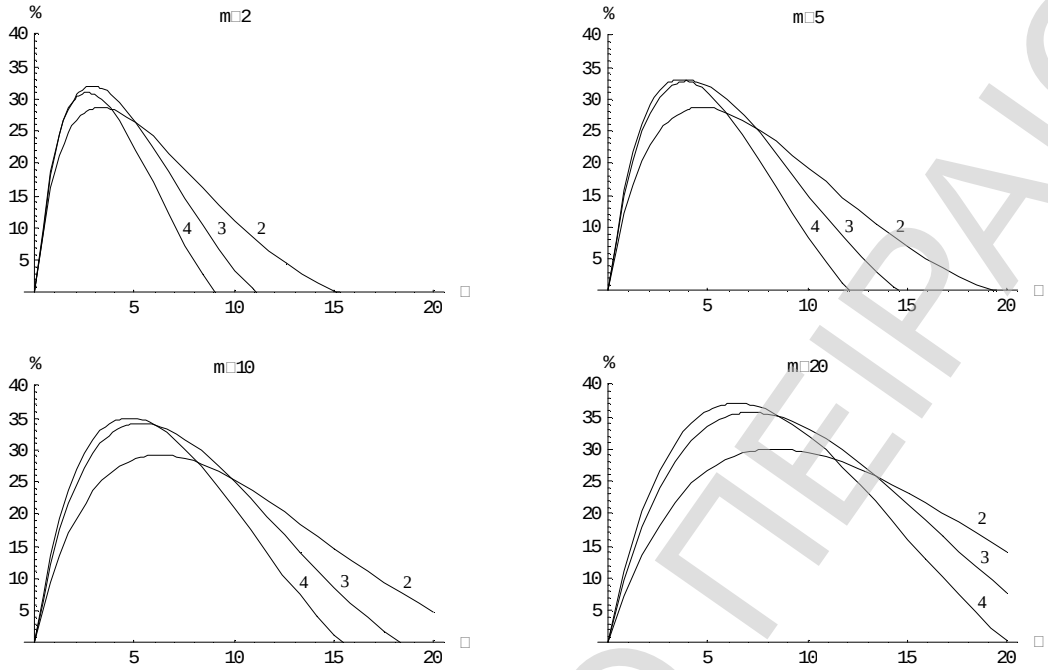
$$I(l) = \frac{ARL_1(l) - ARL_4(l)}{ARL_1(l)},\ l \geq 0.$$

Παρουσιάζονται τέσσερα διαφορετικά γραφήματα καθένα από τα οποία αναφέρεται σε διαφορετική διάσταση m των δεδομένων. Η σύγκριση, έδειξε ότι όταν το m αυξάνει, όχι μόνο διευρύνεται το εύρος των τιμών της l για τις οποίες η $I(l)$ παραμένει θετική, αλλά αυξάνεται επίσης και η μέγιστη τιμή της. Ένα άλλο σημαντικό σημείο είναι ότι, για μεγάλες τιμές του m, η καμπύλη με τη μέγιστη τιμή είναι αυτή που αντιστοιχεί στη μεγαλύτερη τιμή για το k (παρόλα αυτά, σημειώνουμε ότι αυτή η καμπύλη παίρνει αρνητικές τιμές γρηγορότερα από τις υπόλοιπες).

Σχήμα 6.7: Ποσοστό βελτίωσης του ARL_1 του $k \mid k$ CSCC στο κλασικό CSCC



Σχήμα 6.8: Ποσοστό βελτίωσης του ARL_1 του $1|1\text{-}k\,|k$ CSCC στο κλασικό CSCC



Είναι φανερό από τα προηγούμενα ότι, η ιδανική επιλογή για τη τιμή του k εξαρτάται πολύ από το μέγεθος της μετατόπισης που επιθυμούμε να ανακαλύψουμε. Για να κάνουμε μια σωστή επιλογή, μπορούμε να επιλέξουμε εκείνο το k που έχει την καλύτερη απόδοση για συγκεκριμένη μετατόπιση κοιτώντας το αντίστοιχο διάγραμμα. Ως παράδειγμα θεωρούμε την περίπτωση όπου $m = 20$ και $ARL_{in} = 500$. Στη συνέχεια, σύμφωνα με το τελευταίο γράφημα του Σχήματος 6.7, για $l < 6.52$ καλύτερα αποτελέσματα (μεταξύ των $k = 1,2,3,4$) πετυχαίνουμε επιλέγοντας $k = 4$, για $6.52 < l < 9.89$ η καλύτερη επιλογή είναι η $k = 3$, για $9.89 < l < 17.71$ είναι προτιμότερη η χρήση του $k = 2$ ενώ προτείνεται το τυπικό CSCC για $l > 17.71$. Σε παρόμοια συμπεράσματα καταλήγουμε και για άλλες τιμές του m που συμπεριλαμβάνονται στο Σχήμα 6.7.

Στο ίδιο πνεύμα, το Σχήμα 6.8 απεικονίζει διαγράμματα του ποσοστού βελτίωσης του ARL_{out} των $1|1-2|2$, $1|1-3|3$ και $1|1-4|4$ CSCC έναντι του τυπικού CSCC με $ARL_{in} = 500$.Για να επιτύχουμε την ίδια τιμή του ARL_{in} για το συνδυασμένο διάγραμμα χρησιμοποιήσαμε τα αποτελέσματα των τριών τελευταίων γραμμών του Πίνακα 6.2. Σημειώνουμε ότι το ARL του $1|1-k|k$ CSCC εμφανίζει την ίδια συμπεριφορά και παρόμοια χαρακτηριστικά με το $k|k$ CSCC. Παρόλα αυτά, συγκρίνοντας τα Σχήματα 6.7 και 6.8 βλέπουμε ότι το ποσοστό βελτίωσης που

επετεύχθη από το $1|1-k|k$ CSCC είναι αρκετά μεγαλύτερο από τη βελτίωση που επετεύχθη από το αντίστοιχο $k|k$ CSCC. Επιπλέον, το διάστημα της παραμέτρου μη κεντρικότητας l στο οποίο το $1|1-k|k$ CSCC έχει καλύτερη εμφάνιση είναι αρκετά πιο ευρύ. Πρέπει επίσης να τονισθεί ότι όταν είναι πιθανόν να εμφανιστούν τεράστιες μετατοπίσεις στο διάνυσμα των μέσων, το $1|1-k|k$ CSCC είναι μια πιο φυσική επιλογή από το $k|k$ CSCC με $k>1$. Κλείνοντας, σημειώνουμε ότι η μελέτη έδειξε πως, καθώς η τιμή του ARL_{in} αυξάνει, τα καινούρια CSCC γίνονται πιο ελκυστικά συγκρινόμενα με το τυπικό CSCC.

6.6.6. Συμπεράσματα

Στην παρούσα παράγραφο η συνδυασμένη χρήση της θεωρίας ροών και του κλασικού πολυμεταβλητού Shewhart CSCC οδήγησε σε μια διαδικασία που βελτιώνει τις αδυναμίες του CSCC στην περίπτωση σχετικά μικρών μετατοπίσεων στο διάνυσμα των μέσων. Η βελτιωμένη απόδοση της προτεινόμενης παραλλαγής μπορεί να αποδοθεί στην αυξανόμενη ευαισθησία του στατιστικού των ροών στην αποκάλυψη συστάδων παρόμοιων αποτελεσμάτων.

Σε ένα πρόσφατο άρθρο, οι Aparisi et al. (2004) εξέτασαν την εμφάνιση του CSCC με συμπληρωματικούς κανόνες ροών. Ειδικότερα, πέρα από το κλασικό κριτήριο εκτός ελέγχου (ένα σημείο πάνω από το UCL) πρότειναν τη χρήση τριών επιπλέον κανόνων βασιζόμενοι σε δυο ροές και συναρτήσεις σάρωσης. Όπως εκεί αποδείχθηκε, για μέτριες μετατοπίσεις, η συνδυασμένη χρήση όλων των συμπληρωματικών κανόνων ροών βελτιώνουν τις τιμές του ARL_{out} του CSCC περίπου κατά 25% (για $ARL_{in}=200$). Η μελέτη μας, έδειξε ότι η απλή προσέγγιση που προτάθηκε σε αυτό το κεφάλαιο οδηγεί σε τιμές του ARL_{out} πολύ κοντά σε αυτές που πέτυχαν οι Aparisi et al. (2004). Για $ARL_{in}=200$ και $m=2,3,10$ ο Πίνακας 6.3 παρουσιάζει τις τιμές του ARL του τυπικού CSCC, του $k|k$ CSCC, του $1|1-k|k$ CSCC και των αντίστοιχων τιμών που παρουσίασαν οι Aparisi et al. (2004) (στήλη SRR). Οι τιμές στις παρενθέσεις στις στήλες $k|k$ και $1|1-k|k$ δίνουν τις του k ($k=2,3$ ή 4) που παράγουν τη μεγαλύτερη μείωση του ARL.

Πίνακας 6.3: Υπολογισμοί του *ARL* και συγκρίσεις

k	$\sqrt{l}$	CSCC	$k \mid k$		$1 \mid 1-k \mid k$		SRR
2	0	200	200		200		200.6
	1	41.93	38.54	(2)	35.16	(3)	34.16
	2	6.88	6.36	(2)	5.49	(2)	5.68
	3	2.16	2.71	(2)	2.01	(2)	2.21
3	0	200	200		200		199.7
	1	52.64	48.25	(2)	44.49	(3)	42.01
	2	8.82	7.90	(2)	6.93	(3)	7.06
	3	2.55	3.01	(2)	2.30	(2)	2.52
10	0	200	200		200		199.77
	1	92.70	83.69	(4)	80.80	(4)	78.06
	2	20.62	17.27	(3)	15.44	(3)	14.65
	3	5.21	4.99	(2)	4.23	(2)	4.56

Όταν τα δεδομένα μας απαιτούν την ταυτόχρονη ανάλυση δυο ή περισσοτέρων ποιοτικών χαρακτηριστικών, μπορούμε επίσης να χρησιμοποιήσουμε ένα πολυμεταβλητό CUSUM ή πολυμεταβλητό EWMA διάγραμμα ελέγχου (βλέπε Alwan (1986), Lowry and Montgomery (1995)). Αυτά τα διαγράμματα, όπως και τα αντίστοιχα μονομεταβλητά, είναι αποτελεσματικότερα των διαγραμμάτων Shewhart στην ανίχνευση μικρών μετατοπίσεων στη διαδικασία αλλά δύσκολα στην εφαρμογή τους κι οι κανόνες που τα διέπουν μπορεί να είναι τόσο πολύπλοκοι που να αποκλείουν κάθε θεωρητική ανάλυση.

Σύμφωνα με τα παραπάνω, μπορούμε να συνοψίσουμε τα κύρια σημεία της νέας μεθόδου ως εξής: (α) διατηρεί την απλότητα του τυπικού CSCC και προσφέρει ένα εύχρηστο περιβάλλον για την εξαγωγή αποτελεσμάτων θεωρητικού ενδιαφέροντος, (β) βελτιώνει σημαντικά την εμφάνιση του τυπικού CSCC, (γ) αντίθετα από τις πολυμεταβλητές CUSUM και EWMA διαδικασίες δεν απαιτεί την εκτίμηση των παραμέτρων μέσω δύσκολων και χρονοβόρων μαθηματικών υπολογισμών.

6.7. Δειγματοληψία Αποδοχής

Ο Wolfowitz (1943) και αργότερα οι Praire et al. (1962) ανέπτυξαν ένα δειγματοληπτικό ακολουθιακό σχέδιο (κατά σωρούς), σύμφωνα με το οποίο αποδεχόμαστε ένα σωρό, αν σε ένα δείγμα αντικειμένων από το σωρό, υπάρχει μια ροή από k μη ελαττωματικά αντικείμενα. Η πιθανότητα αποδοχής του σωρού μπορεί να υπολογισθεί με βάση τα αποτελέσματα των Κεφαλαίων 3 και 4, που αφορούν τις διωνυμικές κατανομές τάξης k και τις πολυμεταβλητές γενικεύσεις αυτών. Το παραπάνω σχέδιο γενικεύεται στην περίπτωση που έχουμε περισσότερα από δύο δυνατά αποτελέσματα σε κάθε δοκιμή (επιλογή αντικειμένου). Ας υποθέσουμε ότι διαλέγουμε αντικείμενα από ένα σωρό, έτσι ώστε κάθε δοκιμή να είναι τρίτιμη (με σταθερές πιθανότητες). Κάθε αντικείμενο χαρακτηρίζεται ως πλήρως ικανοποιητικό, δηλαδή σύμφωνο με τις προδιαγραφές του κατασκευαστή (τύπος F), μερικώς ικανοποιητικό (τύπος S') ή ελαττωματικό (τύπος S). Σύμφωνα με το δειγματοληπτικό σχέδιο που θεωρούμε εδώ, ένας σωρός γίνεται αποδεκτός αν στο δείγμα των n αντικειμένων που επιλέγονται από αυτόν εμφανιστούν τουλάχιστον r διαδοχικά αντικείμενα τύπου S', ενώ απορρίπτεται αν παρατηρηθούν τουλάχιστον k διαδοχικά ελαττωματικά αντικείμενα (τύπου S).

Ένα νέο δειγματοληπτικό σχέδιο δειγματοληψίας αποδοχής μπορεί να βασισθεί στην τυχαία μεταβλητή T_r η οποία μελετήθηκε στο $5^{ο}$ Κεφάλαιο της παρούσας διατριβής.

6.8. Έλεγχος Εκκίνησης Μηχανημάτων στην Βιομηχανία

Ένας άλλος κλάδος εφαρμογών, όπου εμφανίζονται προβλήματα παρόμοια με τα προηγούμενα, είναι ο έλεγχος εκκίνησης μηχανημάτων (Start-Up Tests). Υποθέτουμε ότι ένα μηχάνημα δοκιμάζεται ως προς την ικανότητα του να τεθεί σε λειτουργία. Υπάρχουν τρία ενδεχόμενα: το μηχάνημα να λειτουργεί κανονικά (S'), να λειτουργήσει μόνο για περιορισμένο χρονικό διάστημα (S) ή να μην λειτουργήσει (F). Το μηχάνημα θεωρείται καλό όταν λειτουργήσει κανονικά σε r διαδοχικές εκκινήσεις και ελαττωματικό αν δε λειτουργήσει k διαδοχικές φορές. Η πιθανότητα να ολοκληρωθεί ο έλεγχος της συσκευής μέχρι τη n-οστή δοκιμή (εκκίνηση)

εκφράζεται μέσω μιας τυχαίας μεταβλητής που καταγράφει τον χρόνο αναμονής μέχρι την πρώτη εμφάνιση ροής επιτυχιών ή την πρώτη εμφάνιση ροής αποτυχιών.

Για περισσότερες λεπτομέρειες σε τέτοιους ελέγχους, ο ενδιαφερόμενος παραπέμπεται στους Balakrishnan et al (1993) και Viveros and Balakrishnan (1993).

6.9. Ανακεφαλαίωση

Στο Κεφάλαιο αυτό αναφερθήκαμε διεξοδικά στις εφαρμογές της θεωρίας ροών επιτυχιών και γενικά των σχηματισμών στην βιομηχανία και στον στατιστικό έλεγχο ποιότητας.

ΚΕΦΑΛΑΙΟ 7: ΕΦΑΡΜΟΓΕΣ ΣΕ ΔΙΑΦΟΡΑ ΕΠΙΣΤΗΜΟΝΙΚΑ ΠΕΔΙΑ

7.1. Εισαγωγή

Στο κεφάλαιο αυτό θα παρουσιάσουμε αναλυτικά κάποιες εξίσου σπουδαίες, με αυτές του προηγούμενου κεφαλαίου, εφαρμογές των τυχαίων μεταβλητών που μελετήθηκαν στα προηγούμενα κεφάλαια. Συγκεκριμένα, θα αναφερθούμε διεξοδικά στις εφαρμογές της θεωρίας ροών επιτυχιών και γενικά των σχηματισμών στα πεδία των ελέγχων τυχαιότητας, στη θεωρία αξιοπιστίας συστημάτων, στην πολυμεταβλητή στατιστική τεχνική της ανάλυσης συστάδων (cluster analysis), καθώς και περιληπτικά σε μία σειρά άλλων εφαρμογών όπως οι αλυσίδες DNA, η οικονομική ανάπτυξη και ανταγωνιστικότητα, η ψυχολογία, η οικολογία και η μετεωρολογία.

7.2. Έλεγχοι Τυχαιότητας

Στη δεκαετία του 1940, όταν το ενδιαφέρον για τη θεωρία των ροών επιτυχιών και των εφαρμογών τους ήταν ιδιαίτερα αυξημένο (Mood (1940), Wald and Wolfowitz (1940), Mosteller (1941), Grand (1946) και David (1947)) προτάθηκαν δύο διαφορετικοί στην φύση τους έλεγχοι τυχαιότητας. Ο πρώτος βασίζεται στο κλασικό κριτήριο των ροών, όπου γίνεται χρήση του συνολικού αριθμού R_n των ροών επιτυχιών και αποτυχιών ανεξαρτήτως μήκους, ενώ το δεύτερο κριτήριο χρησιμοποιεί το μήκος της μέγιστης ροής L_n.

Τρεις δεκαετίες αργότερα, οι O'Brien (1976) και O'Brien and Dyck (1985) ανέπτυξαν έναν έλεγχο τυχαιότητας που λαμβάνει υπόψη και τη διασπορά του μήκους των ροών.

Αργότερα, οι Agin and Godbole (1992) πρότειναν ένα νέο έλεγχο βασισμένο στη δεσμευμένη κατανομή του $N_{n,k}$, δοθέντος του αριθμού των επιτυχιών X_n της ακολουθίας.

Επειδή αυτό το κριτήριο αποδείχθηκε πολύ πιο ισχυρό από τους κλασικούς ελέγχους στο να ανακαλύπτει συγκεκριμένους τύπους μη τυχαιότητας (ομαδοποιήσεις), οι Koutras and Alexandrou (1997) προτείναν και μελέτησαν τη συμπεριφορά κριτηρίων που βασίζονται στις στατιστικές συναρτήσεις $M_{n,k}$ και $G_{n,k}$ με πολύ ενθαρρυντικά αποτελέσματα, ιδιαίτερα για το $M_{n,k}$.

Στην παράγραφο αυτή προχωράμε σε μια διεξοδική μελέτη ενός νέου κριτηρίου που βασίζεται στην στατιστική συνάρτηση $S_{n,k}$ (άθροισμα των μηκών των ροών επιτυχιών μήκους τουλάχιστον k) και συγκρίνουμε τα αποτελέσματά του με αυτά των στατιστικών συναρτήσεων $M_{n,k}$, $G_{n,k}$ που προτάθηκαν από τους Koutras and Alexandrou (1997).

Στα απαραμετρικά κριτήρια οι εναλλακτικές υποθέσεις είναι γενικές υποθέσεις που δεν περιλαμβάνουν παραμέτρους. Για το λόγο αυτό η αξιολόγηση των κριτηρίων τέτοιου είδους βασίζεται συνήθως σε τεχνικές Monte-Carlo, που εφαρμόζονται για συγκεκριμένες εναλλακτικές υποθέσεις. Έτσι, δημιουργούνται τεχνητά ακολουθίες για τις οποίες αληθεύει η εναλλακτική υπόθεση και εκτιμάται με προσομοίωση η ικανότητα του ελέγχου να οδηγήσει σε σωστή απόρριψη της μηδενικής υπόθεσης. Συνήθως, υπολογίζεται η p-value του ελέγχου, η οποία συγκρίνεται με το επίπεδο σημαντικότητας a. Το ποσοστό των ορθών απορρίψεων της μηδενικής υπόθεσης εκφράζει την εμπειρική ισχύ του ελέγχου.

Για την αξιολόγηση του κριτηρίου που βασίζεται στη δεσμευμένη κατανομή $S_{n,k} \mid X_n = n - y$ ακολουθήσαμε την εξής διαδικασία. Για τον έλεγχο της μηδενικής υπόθεσης,

H_0 : *Η πιθανότητα επιτυχίας στις* n *δοκιμές είναι ίση με 0.5,*

χρησιμοποιήσαμε ως κρίσιμη περιοχή την $S_{n,k} \le c_y \mid X_n = n - y$.

Για κάθε $y = 0,1,2,...,n$ υπολογίσαμε την τιμή του c_y για την οποία η πιθανότητα σφάλματος τύπου Ι είναι μικρότερη ή ίση από το επίπεδο σημαντικότητας a. Για τον υπολογισμό των c_y χρησιμοποιήθηκε το Θεώρημα 3.11.

Στη συνέχεια, δημιουργήθηκαν τεχνητά 100 ακολουθίες έτσι ώστε για τις πιθανότητες επιτυχίας των δοκιμών τους να αληθεύει μια συγκεκριμένη εναλλακτική υπόθεση.

Για κάθε ακολουθία υπολογίσθηκε η κρίσιμη τιμή (με χρήση των τύπων που δόθηκαν στην Παράγραφο 3.5) της αντίστοιχης στατιστικής συνάρτησης που χρησιμοποιήθηκε. Στην συνέχεια καταγράφηκε ο αριθμός των ακολουθιών στις οποίες η τιμή της στατιστικής

συνάρτησης που χρησιμοποιήθηκε ανήκε στην κρίσιμη περιοχή του ελέγχου. Με τη βοήθεια του αριθμού αυτού (που είναι ίσος με τον αριθμό των ακολουθιών στις οποίες η p-value του ελέγχου είναι μικρότερη ή ίση του a), υπολογίσθηκε το ποσοστό των ορθών απορρίψεων της μηδενικής υπόθεσης ή ισοδύναμα, η εμπειρική ισχύς του ελέγχου.

Οι υπολογισμοί πραγματοποιήθηκαν για μέγεθος ακολουθίας $n = 20, 50, 100, 200$ και επίπεδο σημαντικότητας $a = 0.01, 0.05, 0.10$. Χρησιμοποιήθηκαν δύο κατηγορίες εναλλακτικών της μηδενικής υποθέσεων (ακολουθιών):

1. Μαρκοβιανή εξάρτηση πρώτης τάξης με $p_1 = 0.5$ και

$$p_i = \begin{cases} p, & \text{άν η δοκιμή } (i-1) \text{ είναι επιτυχία} \\ p_1, & \text{άν η δοκιμή } (i-1) \text{ είναι αποτυχία} \end{cases}$$

όπου $p = 0.60, 0.65, ..., 0.99$. Η παραπάνω ακολουθία βρίσκει εφαρμογή σε κριτήρια εκμάθησης, αθλητικό συναγωνισμό, κλπ., όπου ο ερευνητής καλείται να διαπιστώσει αν ένα συγκεκριμένο αποτέλεσμα (π.χ. επίδοση σε μια δοκιμασία) δικαιολογεί την αρχή "η επιτυχία γεννά επιτυχία", δηλαδή, αν κάποιο θετικό αποτέλεσμα οδηγεί με μεγάλη πιθανότητα σε επίσης θετικό αποτέλεσμα στην επόμενη δοκιμή.

2. Κυκλική ομαδοποίηση (με κύκλο μήκους 10) με $p_1 = 0.5$ και

$$p_i = \begin{cases} p, & \text{άν } 10r+1 \le i \le 10r+c \\ p_1, & \text{διαφορετικά} \end{cases}$$

με $c \le 10$ και $p = 0.60, 0.65, ..., 0.99$.

Η εμπειρική ισχύς του νέου κριτηρίου (που στηρίζεται στην $S_{n,k} \mid X_n = n - y$), συγκρίθηκε με την εμπειρική ισχύ του κριτηρίου τυχαιότητας που προτάθηκε από τους Koutras and Alexandrou (1997) και το οποίο στηρίζεται στην κατανομή της στατιστικής $M_{n,k} \mid X_n = n - y$, εφαρμόζοντας τις δύο παραπάνω κατηγορίες δοκιμών.

Τα αποτελέσματα της μελέτης με χρήση προσομοίωσης για την περίπτωση Μαρκοβιανής εξάρτησης μεταξύ των διαδοχικών δοκιμών Bernoulli εμφανίζονται στον Πίνακα 7.1. Η εμπειρική ισχύς καταγράφηκε εφαρμόζοντας ελέγχους για $k = 2,3,...,9$ και επιλέγοντας την μέγιστη ισχύ σε κάθε περίπτωση. Όπως προκύπτει από την παρατήρηση του Πίνακα 7.1 οι έλεγχοι που στηρίζονται στην στατιστική $S_{n,k}$ είναι ισχυρότεροι από τους αντίστοιχους που στηρίζονται στην στατιστική $M_{n,k}$ στην περίπτωση που το επίπεδο σημαντικότητας a

πρέπει να είναι μικρό ($a \leq 0.01$) ενώ στις περιπτώσεις που το επίπεδο σημαντικότητας a ανήκει στο διάστημα $0.05 \leq a \leq 0.10$, τα αποτελέσματα είναι συγκρίσιμα.

Πίνακας 7.1: Εμπειρική ισχύς των ελέγχων για τη περίπτωση Μαρκοβιανής εξάρτησης

Παράμετρος		$a = 0.10$		$a = 0.05$		$a = 0.01$	
p	n	$S_{n,k}$	$M_{n,k}$	$S_{n,k}$	$M_{n,k}$	$S_{n,k}$	$M_{n,k}$
0,99	50	0,98	0,99	0,99	1,00	0,96	0,88
0,99	100	0,98	0,98	0,98	0,97	0,98	0,85
0,99	150	0,94	1,00	0,97	1,00	0,93	0,81
0,95	50	0,92	0,92	0,91	0,92	0,90	0,76
0,95	100	0,86	0,99	0,80	0,95	0,87	0,87
0,95	150	0,82	0,98	0,74	0,97	0,69	0,91
0,65	50	0,74	0,59	0,46	0,25	0,39	0,14
0,65	100	0,62	0,65	0,48	0,54	0,39	0,14
0,65	150	0,63	0,66	0,49	0,54	0,41	0,15

Τα αποτελέσματα της μελέτης, με χρήση προσομοίωσης, για την περίπτωση κυκλικής ομαδοποίησης εμφανίζονται στον Πίνακα 7.2. Παρατηρώντας προσεκτικά τον Πίνακα 7.2 προκύπτει ότι οι έλεγχοι που στηρίζονται στην στατιστική $S_{n,k}$ είναι ισχυρότεροι από τους ελέγχους που στηρίζονται στην στατιστική $M_{n,k}$ (ανεξαρτήτου του επίπεδου σημαντικότητας a). Στην περίπτωση της κυκλικής ομαδοποίησης, υπάρχουν περιπτώσεις όπου η ισχύς των ελέγχων που στηρίζονται στην στατιστική $M_{n,k}$ είναι εξαιρετικά χαμηλή.

Ένα πολύ ενδιαφέρον συμπέρασμα της μελέτης, στην περίπτωση της Μαρκοβιανής εξάρτησης των δοκιμών, είναι ότι τα αποτελέσματα ελέγχων που στηρίζονται στην στατιστική $S_{n,k}$ δεν είναι ευαίσθητα στην επιλογή της παραμέτρου k.

Συγκεκριμένα, όπως μπορούμε να παρατηρήσουμε στον Πίνακα 7.3 και στα Σχήματα 7.1 και 7.2, η επιλογή του k δεν επηρεάζει ιδιαίτερα την ισχύ των ελέγχων που στηρίζονται στην $S_{n,k}$, το οποίο σημαίνει ότι δεν είναι απαραίτητο να ορίσουμε επιπλέον κριτήρια επιλογής της τιμής του k.

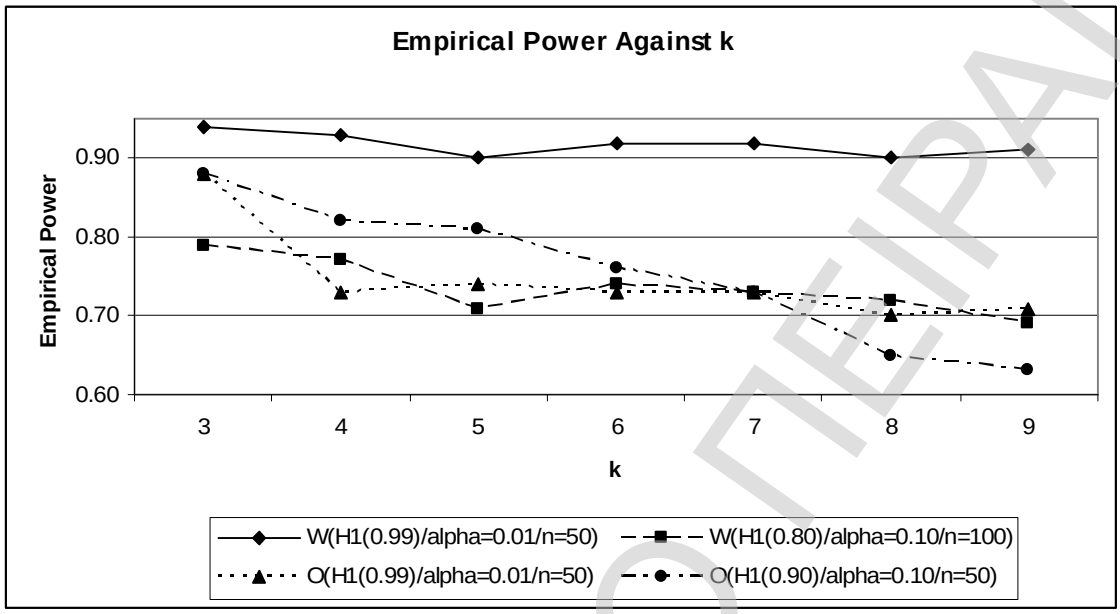
Πίνακας 7.2: Εμπειρική ισχύς των ελέγχων για την περίπτωση Κυκλικής εξάρτησης

Παράμετρος		$a = 0.10$		$a = 0.05$		$a = 0.01$	
p	n	$S_{n,k}$	$M_{n,k}$	$S_{n,k}$	$M_{n,k}$	$S_{n,k}$	$M_{n,k}$
0,99	50	0,65	0,32	0,68	0,35	0,66	0,04
0,99	100	0,78	0,38	0,51	0,26	0,50	0,11
0,90	50	0,68	0,27	0,69	0,14	0,65	0,02
0,90	100	0,49	0,29	0,47	0,18	0,48	0,29
0,80	50	0,62	0,16	0,55	0,10	0,46	0,09
0,80	100	0,51	0,17	0,45	0,11	0,38	0,02
0,65	50	0,77	0,16	0,59	0,08	0,56	0,02
0,65	100	0,51	0,13	0,49	0,07	0,42	0,01

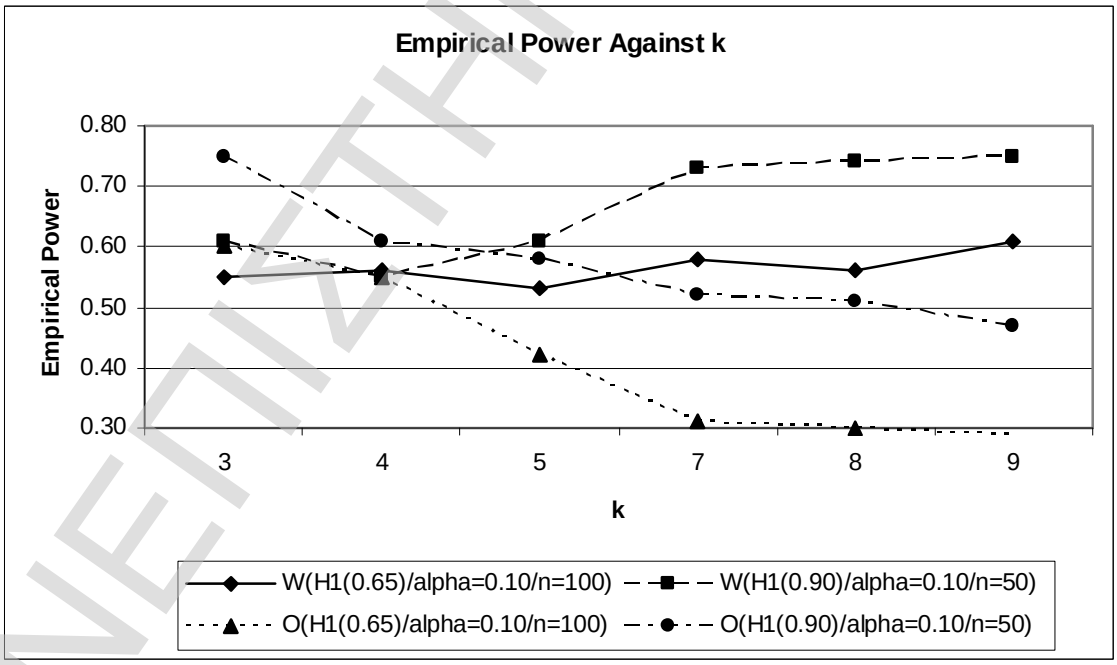
Πίνακας 7.3: Εμπειρική ισχύς των ελέγχων για τη περίπτωση Μαρκοβιανής εξάρτησης

Παράμετροι				k						
p	a	n	$X_{n,k}$	2	3	5	6	7	9	max − min
0,90	0,10	50	$S_{n,k}$	0,79	0,77	0,73	0,67	0,67	0,70	0,12
0,90	0,10	50	$M_{n,k}$	0,82	0,82	0,74	0,78	0,61	0,62	0,21
0,90	0,05	50	$S_{n,k}$	0,71	0,69	0,68	0,58	0,60	0,66	0,13
0,90	0,05	50	$M_{n,k}$	0,86	0,82	0,67	0,63	0,48	0,54	0,38
0,95	0,05	100	$S_{n,k}$	0,64	0,67	0,63	0,67	0,60	0,70	0,07
0,95	0,05	100	$M_{n,k}$	0,88	0,82	0,71	0,65	0,59	0,50	0,38
0,65	0,10	100	$S_{n,k}$	0,52	0,52	0,46	0,37	0,56	0,62	0,16
0,65	0,10	100	$M_{n,k}$	0,65	0,60	0,41	0,26	0,32	0,28	0,39

Σχήμα 7.1: Εμπειρική ισχύς των ελέγχων $S_{n,k}$ (W) και $M_{n,k}$ (O) για την περίπτωση Μαρκοβιανής εξάρτησης για διάφορα k, για $a = 0.01$ και $p = 0.99$



Σχήμα 7.2: Εμπειρική ισχύς των ελέγχων $S_{n,k}$ (W) και $M_{n,k}$ (O) για την περίπτωση Μαρκοβιανής εξάρτησης για διάφορα k, για $a = 0.10$ και $p = 0.65$



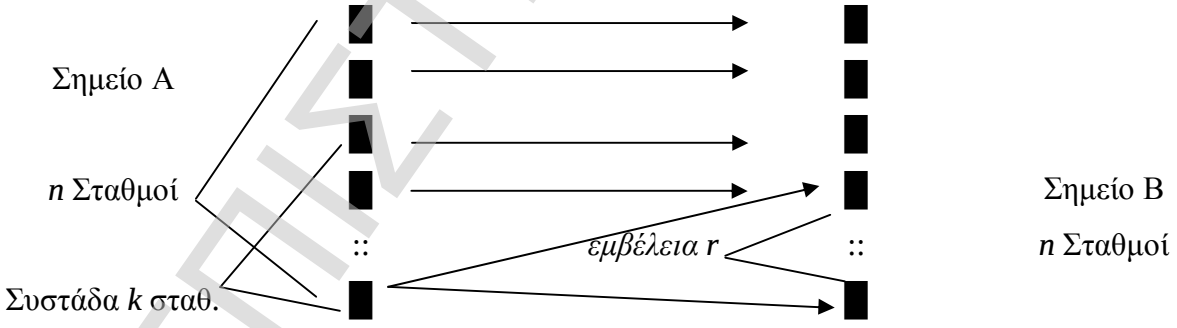
7.3. Θεωρία Αξιοπιστίας

Η θεωρία ροών επιτυχιών βρίσκει άμεση εφαρμογή στη μελέτη της αξιοπιστίας των διαδοχικών-k-από-τα-n: F συστημάτων (consecutive-k-out-of-n: F systems). Ένα τέτοιο σύστημα που αποτελείται από n μονάδες, παύει να λειτουργεί αν τουλάχιστον k διαδοχικές μονάδες του χαλάσουν. Αν θεωρήσουμε ως επιτυχία την εμφάνιση χαλασμένης μονάδας και ως αποτυχία την εμφάνιση μονάδας που λειτουργεί κανονικά, τότε η αξιοπιστία του συστήματος εκφράζεται ως $R_{n,k} = P(N_{n,k} = 0) = P(M_{n,k} = 0) = P(G_{n,k} = 0)$. Για περισσότερες λεπτομέρειες ο ενδιαφερόμενος αναγνώστης παραπέμπεται στο άρθρο των Chao et al. (1995). Επιπρόσθετα, ένα σύστημα διαδοχικών-r-μεταξύ-k-από-τα-n (consecutive-r-within-k-out-of-n: F system) ορίζεται ως το σύστημα που παύει να λειτουργεί όταν σε k διαδοχικές μονάδες του συστήματος εμφανίζονται συνολικά τουλάχιστον r επιτυχίες (r χαλασμένες μονάδες). Ο υπολογισμός της αξιοπιστίας του συστήματος αποτελεί άμεση εφαρμογή της θεωρίας των γενικευμένων ροών επιτυχιών (scans). Επίσης, στην βιβλιογραφία έχουν μελετηθεί διδιάστατα συστήματα (Chen and Glaz (1996)) κάθε μονάδα των οποίων θεωρείται τυχαία μεταβλητή Z_i που παίρνει θετικές ακέραιες τιμές. Το σύστημα αυτό παύει να λειτουργεί όταν σε ένα ορθογώνιο με διαστάσεις (k_1, k_2) το άθροισμα $\sum_{\in I_{k_1,k_2}} Z_i$ είναι μεγαλύτερο από έναν προκαθορισμένο αριθμό. Η μέθοδος εμφύτευσης σε Μαρκοβιανή αλυσίδα, μπορεί να χρησιμοποιηθεί για τον υπολογισμό της αξιοπιστίας αυτής. Μια άλλη μορφή συστημάτων αξιοπιστίας γεννάτε όταν ένα σύστημα αποτελείται από μονάδες τριών τύπων, καλή (good), μονάδα εκτός λειτουργίας τύπου 1 (failed-short) και μονάδα εκτός λειτουργίας τύπου 2 (failed-open), λέγεται σύστημα με δύο είδη αποτυχίας DFM (dual-failure mode system). Τέτοια συστήματα μπορεί να εμφανισθούν σε διάφορα προβλήματα σχετιζόμενα με την βιομηχανία. Η αξιοπιστία των DFM συστημάτων υπήρξε αντικείμενο έρευνας από τα μέσα της δεκαετίας του 1950. Εφαρμόζοντας την αρχή DFM στο γνωστό σύστημα διαδοχικό-k:-από-τα-n έχει εισαχθεί το διαδοχικό-k, r-από-τα-n: DFM σύστημα (consecutive-k, r-out-of-n: DFM system). Αυτό αποτελείται από n στοιχεία διατεταγμένα γραμμικά ή κυκλικά και κάθε στοιχείο του μπορεί να είναι καλό (F), να είναι εκτός λειτουργίας τύπου 1 (S) ή να είναι εκτός λειτουργίας τύπου 2 (S'). Το σύστημα θεωρείται ότι είναι εκτός λειτουργίας τύπου 1 (failed-short) αν τουλάχιστον k διαδοχικά στοιχεία του είναι εκτός λειτουργίας τύπου 1 ενώ είναι εκτός

λειτουργίας τύπου 2 (failed-open) αν τουλάχιστον r διαδοχικά στοιχεία του είναι εκτός λειτουργίας τύπου 2. Το διαδοχικό-k,r-από-τα-n: DFM σύστημα ($1<k<n$ και $1<r<n$) μελετήθηκε από τον Koutras (1996a) με τη βοήθεια αναγωγικών σχέσεων και φραγμάτων.

Μια ιδιαίτερα χρήσιμη εφαρμογή της θεωρίας ροών και σχηματισμών στην θεωρία αξιοπιστίας συστημάτων δίνεται στην διατριβή αυτή (βλέπε επίσης και Koutras et al. (2005b)). Όπως προαναφέραμε το διαδοχικό-k-από-τα-n: DFM σύστημα ($1<k<n$) αποτελείται από n διαδοχικά στοιχεία σε μια ευθεία και παύει να λειτουργεί εάν k διαδοχικά στοιχεία του αποτύχουν. Στην παρούσα παράγραφο δίνουμε μια τροποποίηση του διαδοχικού-k,r-από-τα-n: DFM συστήματος.

Έστω ότι n διαδοχικοί σταθμοί radar χρησιμοποιούνται προκειμένου να μεταδοθεί το σήμα από ένα σημείο Α σε ένα σημείο Β (οι σταθμοί radar ισαπέχουν στοιχισμένοι) Υποθέτουμε επίσης ότι κάθε σταθμός radar του σημείου Α έχει εμβέλεια r σταθμούς του σημείου Β (βλέπε και το σχήμα που ακολουθεί). Αυτό σημαίνει ότι για να διακοπεί η μετάδοση του σήματος πρέπει να εμφανιστούν r διαδοχικοί ελαττωματικοί σταθμοί. Δηλαδή, το σήμα μεταδίδεται επιτυχώς, όσο δεν εμφανίζονται r και άνω διαδοχικοί μη λειτουργικοί σταθμοί. Υποθέτουμε ακόμα ότι κατά τη διάρκεια της λειτουργίας του συστήματος, σταθμοί radar που βρίσκονται σε μεγάλες ομάδες χρησιμοποιούνται για την αποστολή επιπλέον σημάτων κωδικοποιημένων σε σήματα μήκους k και άνω πακέτων, που αποστέλλονται από k και άνω διαδοχικούς σταθμούς radar που βρίσκονται σε λειτουργία.



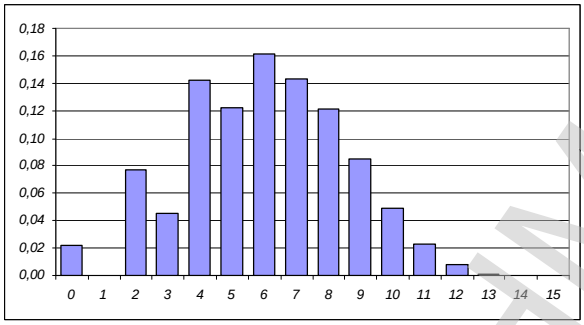
Συμβολίζουμε με $X_{n,k}^{(1)}$ το σύνολο των σταθμών που βρίσκονται σε λειτουργία σε συστάδες μήκους τουλάχιστον k ανάμεσα στους n σταθμούς (άθροισμα μηκών ροών επιτυχιών μήκους τουλάχιστον r), ενώ με $X_{n,r}^{(2)}$ των αριθμό των μη λειτουργικών σταθμών οι

οποίοι βρίσκονται σε συστάδες μήκους r (ροών αποτυχιών μήκους r). Η πιθανότητα το επιπλέον κωδικοποιημένο σήμα που προσφέρουν οι n σταθμοί να είναι ίσο με x δίνεται

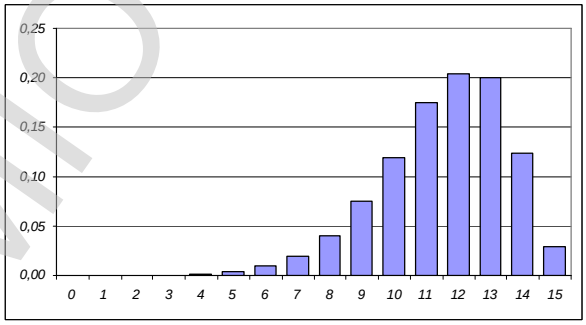
$$P[X_{n,k}^{(1)} = x \mid X_{n,r}^{(2)} = 0] = \frac{P[X_{n,k}^{(1)} = x, X_{n,r}^{(2)} = 0]}{P[X_{n,r}^{(2)} = 0]}$$

ενώ, η μέση τιμή της επιπλέον ποσότητας σήματος που εκπέμπεται είναι ίση με $E[X_{n,k}^{(1)} = x \mid X_{n,r}^{(2)} = 0]$. Η μελέτη του συστήματος αυτού μπορεί να γίνει με χρήση της δεσμευμένης κατανομής της διδιάστατης τυχαίας μεταβλητής $(S_{n,k}^{(S)}, N_{n,r}^{(F)})$ που μελετήθηκε στο Κεφάλαιο 4, (θέτοντας $S_{n,k}^{(S)} = X_{n,k}^{(1)}$ και $N_{n,r}^{(F)} = X_{n,r}^{(2)}$). Στα Σχήματα 7.3.α και 7.3.β, που ακολουθούν δίνονται οι δεσμευμένες κατανομές $P[X_{n,k}^{(1)} = x \mid X_{n,r}^{(2)} = 0]$ για διάφορα k, r, p.

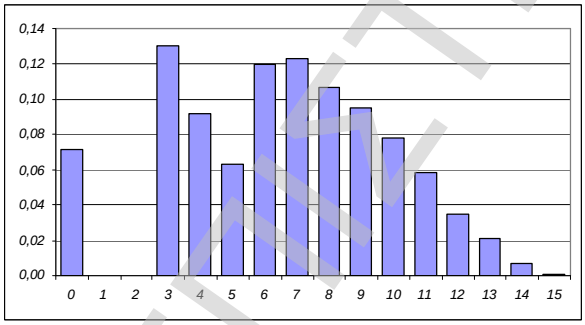
Σχήμα 7.3.α: Η $P[X_{n,k}^{(1)} = x \mid X_{n,r}^{(2)} = 0]$ για διάφορα k, r, p και $n = 15$



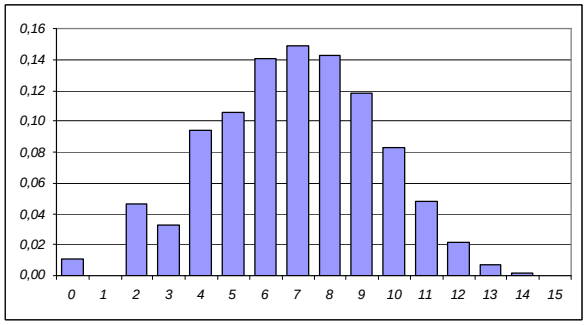
$n = 15, p = 0.25, k = 2, r = 2$



$n = 15, p = 0.75, k = 2, r = 2$



$n = 15, p = 0.50, k = 3, r = 2$



$n = 15, p = 0.50, k = 2, r = 3$

Με χρήση των αποτελεσμάτων του Κεφαλαίου 4, είναι δυνατόν να προκύψουν αναδρομικά σχήματα για τη συνάρτηση πιθανότητας και τη διπλή γεννήτρια των δεσμευμένων πιθανοτήτων. Για παράδειγμα, αντικαθιστώντας $z_2 = 0$ στο Πόρισμα 4.1 έχουμε το ακόλουθο Πόρισμα για τη διπλή γεννήτρια της $(X_{n,k}^{(1)} = x, X_{n,r}^{(2)} = 0)$.

Πόρισμα 7.1: Η διπλή γεννήτρια της από κοινού κατανομής των $(X_{n,k}^{(1)} = x, X_{n,r}^{(2)} = 0)$ δίνεται από τον τύπο

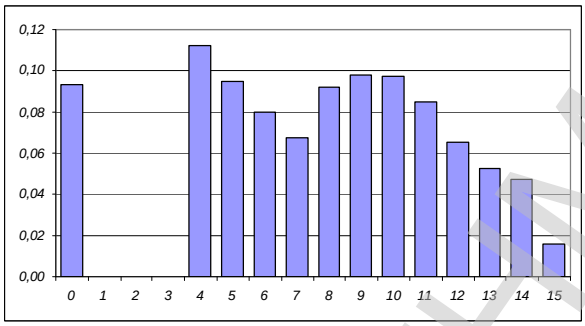
$$\Phi(z_1,0;w) = \sum_{t=0}^{\infty} j_{\,t}(z_1,0)w^t = \frac{P(z_1,0;w)}{Q(z_1,0;w)}$$

με
$$P(z_1,0;w) = 1-(pw)z_1-(pw)^k(1-z_1^k)-(pw)^{k+1}(z_1^k-z_1)-(qw)^r+(qw)^r(pw)z_1$$
$$+(pw)^k(qw)^r(1-z_1^k)+(pw)^{k+1}(qw)^r(z_1^k-z_1)$$

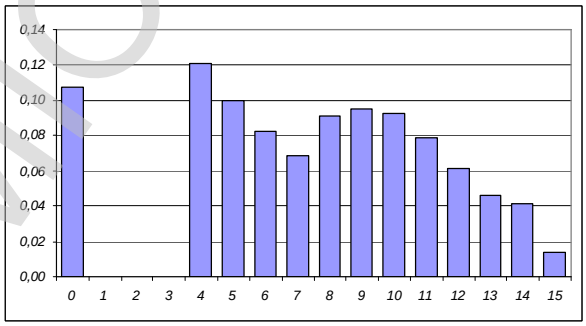
και

$$Q(z_1,z_2;w) = 1-w(1+pz_1)+w^2pz_1+(qw)^r wp-(qw)^r(pw)wz_1p+(pw)^k(qw)(1-z_1^k)$$
$$+(pw)^{k+1}(qw)(z_1^k-z_1)-(pw)^k(qw)^r(1-z_1^k)-(pw)^k(qw)^r w[(qz^2(1-z_1^k))+p(z_1^k-z_1)].$$

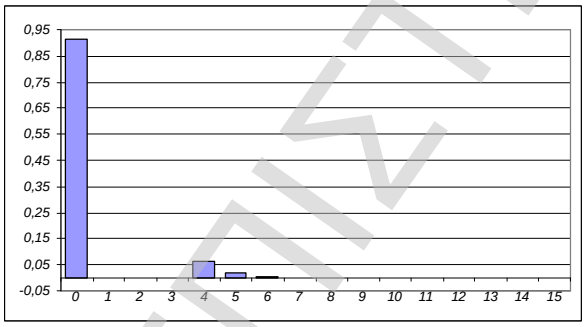
Σχήμα 7.3.β: Η $P[X_{n,k}^{(1)} = x \mid X_{n,r}^{(2)} = 0]$ για διάφορα k, r, p και $n = 15$



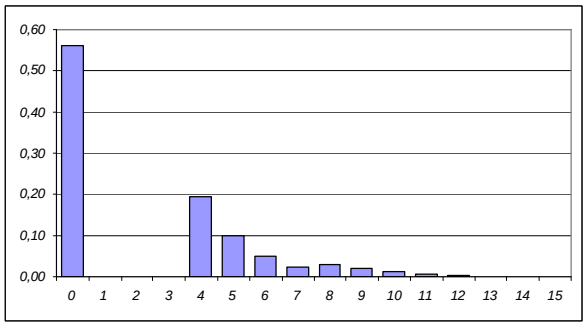
$n = 15, p = 0.75, k = 4, r = 3$



$n = 15, p = 0.75, k = 4, r = 4$



$n = 15, p = 0.50, k = 4, r = 4$



$n = 15, p = 0.25, k = 4, r = 4$

7.4. Πολυμεταβλητή Στατιστική Ανάλυση– Ανάλυση κατά Συστάδες (Cluster Analysis)

Στην εφαρμοσμένη στατιστική έρευνα, εμφανίζονται συχνά περιπτώσεις όπου χρειάζεται να γίνει ομαδοποίηση μεγάλου αριθμού αντικειμένων, συμβόλων ή ατόμων σε αμοιβαίως αποκλειόμενες ομάδες, κάθε μια από τις οποίες έχει μέλη που είναι όσο το δυνατόν πιο όμοια μεταξύ τους με βάση κάποια χαρακτηριστικά. Αυτό δεν είναι μόνο λογικό στο ανθρώπινο μυαλό αλλά είναι επίσης ελκυστικό ως αντικείμενο έρευνας καθώς διευκολύνει στη θεώρηση και κατανόηση αλληλοσυσχετίσεων σε μεγάλες συλλογές δεδομένων. Μια πολυμεταβλητή τεχνική που ασχολείται με την ταξινόμηση αντικείμενων ή ατόμων σε ομάδες με συναφή συμπεριφορά είναι η Ανάλυση κατά Συστάδες. Στην παρούσα διατριβή δίνεται μια εφαρμογή της θεωρίας ροών επιτυχιών στην Ανάλυση κατά Συστάδες (βλέπε επίσης και Vassiliou et al. (2004)).

7.4.1. Εισαγωγή

Ίσως το πιο δύσκολο σημείο στην Ανάλυση κατά Συστάδες είναι η αναγνώριση του πραγματικού αριθμού των συστάδων που υπάρχουν στο σύνολο των δεδομένων που έχουν συγκεντρωθεί. Η μη ιεραρχική διαδικασία συνήθως απαιτεί να καθορίσει ο χρήστης τις παραμέτρους πριν εφαρμόσει κάποιον αλγόριθμο ομαδοποίησης, ενώ οι ιεραρχικές μέθοδοι παράγουν μια σειρά από λύσεις που εκτείνονται από m συστάδες (m είναι ο αριθμός των ατόμων στο σύνολο δεδομένων) έως μια λύση με μια μόνο συστάδα. Με σκοπό να παρέχουν στον ερευνητή επαρκή πληροφορία σχετικά με τον αριθμό των ομάδων που υπάρχουν στα υπό μελέτη δεδομένα, έχουν προταθεί διάφοροι αριθμητικοί δείκτες που οδηγούν σε κατάλληλους «κανόνες διακοπής», από τους Everitt (1978), Jardine and Sibson (1971), Glasbey (1987) και Hardy (1996).

Σε μια μεγάλη μελέτη προσομοίωσης, οι Milligan and Cooper (1985) (και Milligan (1985)), εξέτασαν την ικανότητα 30 τέτοιων κανόνων ώστε να ανακαλύψουν το σωστό αριθμό ομάδων σε συγκεκριμένα τεχνητά σύνολα δεδομένων. Οι Krolak-Schwerdt and Eckes

(1992) διαμόρφωσαν ένα κριτήριο στηριζόμενοι στη θεωρία γραφημάτων (Graph Theory) που οδηγεί σε μια καλά ορισμένη λύση. Ο αλγόριθμος που στη συνέχεια θα αναφέρεται ως αλγόριθμος ή μέθοδος GRAPH πέρα από το πλεονέκτημα να οδηγεί σε σαφείς αποφάσεις, έχει και άλλα σημαντικά στοιχεία όπως ότι είναι αναλλοίωτος σε μονότονους μετασχηματισμούς των δεδομένων και προσφέρει αποτελεσματική αντιμετώπιση των συνόλων δεδομένων με ακραίες τιμές (outliers).

Αν και ο αλγόριθμος GRAPH βασίζεται σε ένα καλά ορισμένο θεωρητικό κριτήριο, αποτυγχάνει να αναγνωρίζει τη σωστή δομή των ομάδων όταν η διάσταση p των δεδομένων και ο πραγματικός αριθμός N των ομάδων που υπάρχουν στα δεδομένα ικανοποιούν μια συνθήκη ανισότητας. Όπως θα δειχθεί στη συνέχεια, ο αλγόριθμος GRAPH μπορεί να μη λειτουργεί σωστά όταν ισχύει η ανισότητα $N \geq p+2$. Παίρνοντας κίνητρο από αυτή την παρατήρηση, εισάγουμε ένα νέο μέτρο απόστασης που ξεπερνά αυτό το πρόβλημα και προσφέρει στη διαδικασία GRAPH την ικανότητα να συμπεριφέρεται σωστά ακόμα και όταν ισχύει η παραπάνω ανισότητα.

Η δομή της παραγράφου 7.4 έχει ως εξής: Στην υποπαράγραφο 7.4.2 παρουσιάζουμε τον τυπικό αλγόριθμο GRAPH και συζητάμε συνοπτικά τις αδυναμίες του. Στην υποπαράγραφο 7.4.3 προτείνεται ένα μέτρο απόστασης βασισμένο στις ροές και τις καμπύλες Andrews. Στην υποπαράγραφο 7.4.4 περιγράφουμε με λεπτομέρεια έναν αλγόριθμο που βασίζεται στο νέο μέτρο απόστασης και τη διαδικασία GRAPH ενώ στην υποπαράγραφο 7.4.5 εξετάζουμε τη λειτουργία του νέου αλγόριθμου μέσω προσομοιωμένων δεδομένων. Τέλος, στην υποπαράγραφο 7.4.6 συζητάμε κάποιες κατευθύνσεις για περαιτέρω έρευνα.

7.4.2. Η μέθοδος GRAPH

Έστω ότι τα δεδομένα μας αποτελούνται από m p-διάστατες παρατηρήσεις $\mathbf{x}_1, \mathbf{x}_2, \mathbf{K}, \mathbf{x}_m$. Επιπλέον, έστω ότι $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1,2,\mathbf{K},m$ συμβολίζει την Ευκλείδεια απόσταση μεταξύ των $\mathbf{x}_i = (x_{i1}, x_{i2}, \mathbf{K}, x_{ip})$ και $\mathbf{x}_j = (x_{j1}, x_{j2}, \mathbf{K}, x_{jp})$, δηλαδή

$$d_{ij} = \left[\sum_{r=1}^{p} (x_{ir} - x_{jr})^2 \right]^{1/2}.$$

Για να εφαρμόσουμε τον αλγόριθμο GRAPH πρέπει να μοντελοποιήσουμε το πρόβλημά μας ορίζοντας το γράφημα $G=(X,V)$, όπου $X=\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{K},\mathbf{x}_m\}$ είναι το σύνολο των κορυφών (vertex set) και $V=\{(\mathbf{x}_i,\mathbf{x}_j):i,j=1,2,\mathbf{K},m\}$ είναι το σύνολο των ακμών (edge set) και το βάρος (μήκος) που σχετίζεται με την ακμή $(\mathbf{x}_i,\mathbf{x}_j)$ ισούται με $d_{ij}=d(\mathbf{x}_i,\mathbf{x}_j)$.

Στη συνέχεια πρέπει να ορίσουμε το δέντρο ελαχίστων αποστάσεων (minimal spanning tree) T, το οποίο είναι υποσύνολο του V (που συνδέει το σύνολο των μελών του X), και του οποίου το συνολικό βάρος (μήκος) $\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in T} d(\mathbf{x}_i,\mathbf{x}_j)$ είναι ελάχιστο. Το υπογράφημα (X,T) θα συμβολίζεται με $S\min$. Τέλος, πρέπει να ορίσουμε το δέντρο μεγίστων αποστάσεων (maximal spanning tree) $S\max=(X,T')$ που αναφέρεται σε ένα υποσύνολο T' του V (που συνδέει όλες τις ακμές του X) του οποίου το συνολικό βάρος (μήκος) $\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in T'} d(\mathbf{x}_i,\mathbf{x}_j)$ είναι μέγιστο.

Η χρήση του δέντρου ελαχίστων αποστάσεων εξασφαλίζει ότι παρόμοια αντικείμενα ($\mathbf{x}_i,\mathbf{x}_j$ με μεγάλα βάρη $d(\mathbf{x}_i,\mathbf{x}_j)$) τοποθετούνται στην ίδια συστάδα, ενώ το δέντρο μεγίστων αποστάσεων χρησιμοποιείται για την αποφυγή της τοποθέτησης μαζί ανόμοιων αντικειμένων ($\mathbf{x}_i,\mathbf{x}_j$ με μικρά βάρη $d(\mathbf{x}_i,\mathbf{x}_j)$).

Ο υπολογισμός του αλγόριθμου GRAPH είναι σχετικά απλός και περιλαμβάνει τη δημιουργία μιας ακολουθίας διαμερίσεων $l_m,l_{m-1},l_{m-2},\mathbf{K}$ του $X=\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{K},\mathbf{x}_m\}$ με την παρακάτω διαδικασία:

Βήμα 1: Κατασκεύασε τα $S\min$ και $S\max$.

Βήμα 2: Δημιούργησε μια αρχική διαμέριση l_m θεωρώντας m συστάδες, η κάθε μια από τις οποίες περιέχει ένα στοιχείο.

Βήμα 3: Ένωσε τα στοιχεία (ακμή) του X που ανήκουν στο $S\min$ με το μικρότερο βάρος (μήκος) δημιουργώντας μια νέα συστάδα. (Αυτό οδηγεί σε μια διαμέριση l_{m-1} που περιλαμβάνει μια συστάδα με δυο αντικείμενα και $m-2$ συστάδες με ένα αντικείμενο η κάθε μια.)

Βήμα 4: Έστω j μετρητής με αρχική τιμή $j=m-1$.

Βήμα 5: Επέλεξε την ακμή από τις εναπομείνουσες ακμές του $S\min$ με το μικρότερο βάρος (μήκος) και δημιούργησε τη διαμέριση l_{j-1} από την προηγούμενη διαμέριση l_j συνδέοντας την επιλεγμένη ακμή. Αν υπάρχουν μια ή περισσότερες ομάδες στη l_{j-1}

που τα στοιχεία τους συνδέονται με μια ακμή του $S\max$, η προηγούμενη διαμέριση l_j είναι η τελική λύση και ο αλγόριθμος τερματίζει. Διαφορετικά, μείωσε το μετρητή j κατά 1 και επανάλαβε το βήμα (5).

Ο αλγόριθμος GRAPH είναι μαθηματικά καλά ορισμένος και οδηγεί πάντα σε μοναδική λύση. Παρόλα αυτά, υπάρχουν περιπτώσεις που δεν λειτουργεί σωστά. Για παράδειγμα δεν πρόκειται ποτέ να καταλήξει σε μια συστάδα ακόμα και αν όλα τα σημεία έχουν υψηλή συμπύκνωση.

7.4.3. Ένα μέτρο απόστασης βασισμένο στις ροές και τις καμπύλες Andrews

Ένα προφανές πρόβλημα με τα πολυμεταβλητά δεδομένα είναι η ταυτόχρονη απεικόνιση των τιμών πολυδιάστατων τυχαίων μεταβλητών. Ο Andrews (1972) περιέγραψε μια μέθοδο για τη γραφική απεικόνιση σε δυο διαστάσεις, πολυμεταβλητών δεδομένων. Κάθε σημείο $\mathbf{x} = (x_1, x_2, \mathbf{K}, x_p)$ αναπαρίσταται από μια αρμονική συνάρτηση της μορφής

$$f_x(t) = x_1/\sqrt{2} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \mathbf{K} \qquad (7.1)$$

η οποία διαγραμματοποιείται στο εύρος $-p \le t \le p$ (Σχήματα 7.4 και 7.5). Αν κάποιες καμπύλες δημιουργούν μια ζώνη παραμένοντας κοντά η μια στην άλλη, τότε τα αντίστοιχα σημεία είναι κοντά μεταξύ τους στον Ευκλείδειο χώρο και μια τέτοια ζώνη δίνει ένδειξη για ύπαρξη συμπαγούς συστάδας. Ο παραπάνω ισχυρισμός δικαιολογείται από το γεγονός ότι η L_2 απόσταση μεταξύ των συναρτήσεων Andrews $f_{\mathbf{x}}$, $f_{\mathbf{y}}$ είναι ανάλογη της γνωστής Ευκλείδειας απόστασης μεταξύ των αντίστοιχων σημείων $\mathbf{x} = (x_1, x_2, \mathbf{K}, x_p)$ και $\mathbf{y} = (y_1, y_2, \mathbf{K}, y_p)$, και πιο συγκεκριμένα

$$\int_{-p}^{p} [f_{\mathbf{x}}(t) - f_{\mathbf{y}}(t)]^2 \, dt = p \sum_{i=1}^{p} (x_i - y_i)^2 .$$

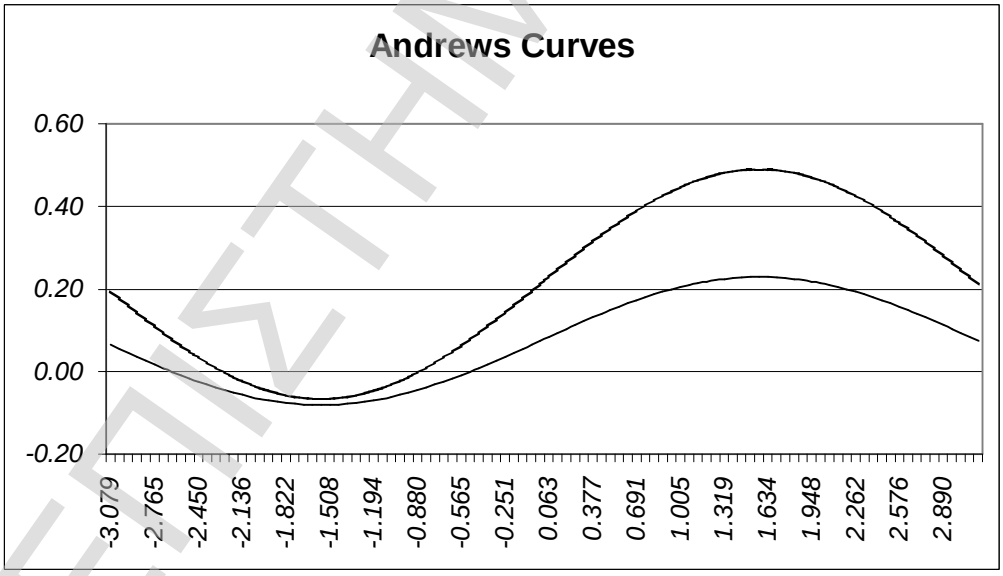
Δυο ενδιαφέροντα χαρακτηριστικά των καμπυλών Andrews είναι η διατήρηση των μέσων και διακυμάνσεων. Το κύριο πλεονέκτημα της τεχνικής αυτής είναι ότι οι ιδιότητές της βασίζονται στη μαθηματική θεωρία και συνεπώς, η ερμηνεία της δεν είναι τόσο υποκειμενική όσο άλλων γραφικών μεθόδων (Embrechts and Herzberg (1991)).

Παρακινούμενοι από την αξιοσημείωτη περιγραφική δύναμη της τεχνικής του Andrews, θα προχωρήσουμε στη διαμόρφωση ενός μέτρου ομοιότητας για τις καμπύλες της μορφής (7.1). Θεωρούμε έναν προκαθορισμένο θετικό αριθμό c και μια διαμέριση $-p < t_1 < t_2 < \mathbf{K} t_n = p$ του διαστήματος $[-p, p]$. Για κάθε ζεύγος x, y δημιουργούμε μια ακολουθία δίτιμων αποτελεσμάτων $Z_1, Z_2, \mathbf{K}, Z_n$ που ορίζεται ως

$$Z_r = \begin{cases} S, & \text{αν } \left| f_{\mathbf{x}}(t_r) - f_{\mathbf{y}}(t_r) \right| \le c \\ F, & \text{αν } \left| f_{\mathbf{x}}(t_r) - f_{\mathbf{y}}(t_r) \right| > c \end{cases} \qquad r = 1,2,\mathbf{K},n.$$

Είναι φανερό ότι, αν οι καμπύλες Andrews $f_{\mathbf{x}}$, $f_{\mathbf{y}}$ παραμένουν κοντά μεταξύ τους για μεγάλα υποδιαστήματα του $[-p, p]$, η αντίστοιχη ακολουθία των δίτιμων αποτελεσμάτων $Z_1, Z_2, \mathbf{K}, Z_n$ θα περιέχει μεγάλες ροές επιτυχιών. Συνεπώς, ο αριθμός των ροών επιτυχιών μεγάλου μήκους στην ακολουθία $Z_1, Z_2, \mathbf{K}, Z_n$ μπορεί να χρησιμοποιηθεί ως μέτρο ομοιότητας μεταξύ των $\mathbf{x}$ και $\mathbf{y}$.

Σχήμα 7.4: Οι καμπύλες του Andrews για δύο πολυδιάστατες παρατηρήσεις

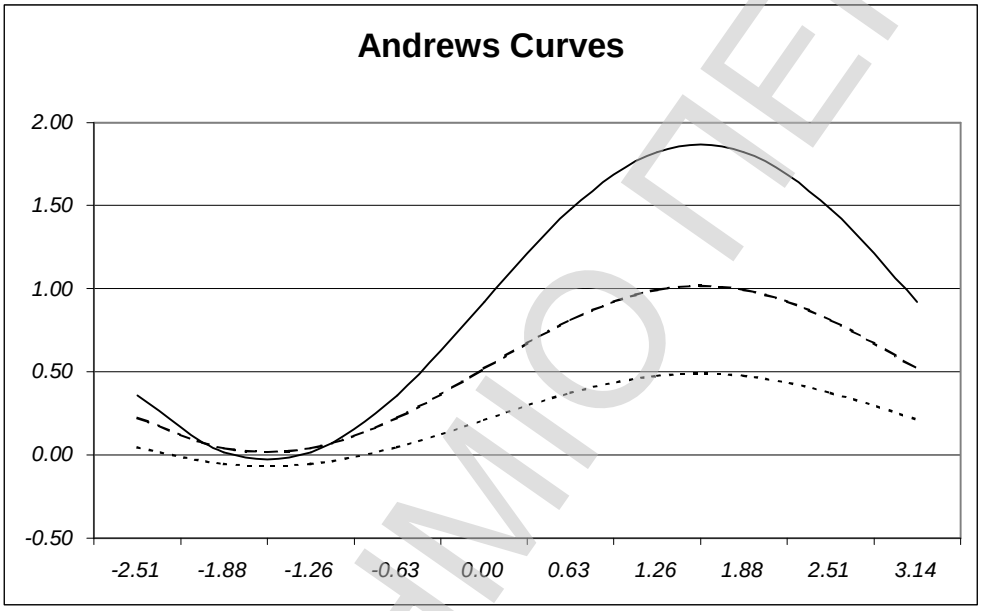


Προφανώς, η σειρά με την οποία είναι διατεταγμένες οι μεταβλητές επηρεάζει το σχήμα των καμπυλών και συνεπώς τη σειρά των δίτιμων αποτελεσμάτων. Ο Andrews (1972) πρότεινε πριν την εφαρμογή της μεθόδου να γίνεται ανάλυση κύριων συνιστωσών και η πρώτη κύρια συνιστώσα να χρησιμοποιείται ως πρώτη μεταβλητή, η δεύτερη κύρια

συνιστώσα ως δεύτερη μεταβλητή κ.ο.κ. Η διαδικασία αυτή σύμφωνα με τον Andrews (1972) βελτιώνει την περιγραφική ικανότητα των καμπυλών.

Για να δημιουργήσουμε ένα μέτρο ομοιότητας μεταξύ δυο πολυμεταβλητών παρατηρήσεων, θα χρησιμοποιήσουμε τον αριθμό των επικαλυπτόμενων ροών επιτυχίας προκαθορισμένου μήκους $k \geq 2$ $M_{n,k}$.

Σχήμα 7.5: Οι καμπύλες του Andrews για τρεις πολυδιάστατες παρατηρήσεις



Έστω $\mathbf{x}$, $\mathbf{y}$ δυο p-διάστατες παρατηρήσεις και ας συμβολίσουμε με $M_{n,k}$ τον αριθμό των επικαλυπτόμενων ροών επιτυχιών στη δίτιμη ακολουθία $Z_1, Z_2, \mathbf{K}, Z_n$ που δημιουργήθηκε όπως αναφέραμε πριν, με τη διαμέριση $P: -p < t_1 < t_2 < \mathbf{K} t_n = p$ και προκαθορισμένη τιμή για την παράμετρο c. Ως μέτρο ομοιότητας μεταξύ των $\mathbf{x}$ και $\mathbf{y}$ θα χρησιμοποιήσουμε την ποσότητα

$$S_p(\mathbf{x},\mathbf{y}) = M_{n,k} / (n-k+1).$$

Είναι προφανές ότι η $S_P(\mathbf{x},\mathbf{y})$ ικανοποιεί τις ακόλουθες ιδιότητες, που είναι οι κλασικές συνθήκες για όλα τα μέτρα ομοιότητας

$$0 \leq S_P(\mathbf{x},\mathbf{y}) \leq 1,\ S_P(\mathbf{x},\mathbf{x}) = 1,\ S_P(\mathbf{x},\mathbf{y}) = S_P(\mathbf{y},\mathbf{x}).$$

Μια κοινή πρακτική ώστε να δημιουργηθεί ένα μέτρο απόστασης από ένα μέτρο ομοιότητας είναι να χρησιμοποιηθεί ο τύπος

$$d_p(\mathbf{x},\mathbf{y}) = \sqrt{2(1-S_p(\mathbf{x},\mathbf{y}))}\,.$$

Είναι επίσης φανερό ότι ισχύουν οι σχέσεις

$$0 \le d_P(\mathbf{x},\mathbf{y}) \le \sqrt{2}\,,\ d_P(\mathbf{x},\mathbf{x}) = 0\,,\ d_P(\mathbf{x},\mathbf{y}) = d_P(\mathbf{y},\mathbf{x})\,.$$

Σημειώνουμε ότι η σχέση $d_P(\mathbf{x},\mathbf{x}) = 0$ δε σημαίνει απαραίτητα ότι ισχύει $\mathbf{x} = \mathbf{y}$, ούτε η d_P ικανοποιεί την τριγωνική ανισότητα. Όπως απέδειξαν οι Gower and Ross (1969), αν έχουμε m ανεξάρτητα $\mathbf{x}_i$, $i = 1,2,\mathbf{K},m$ και ο πίνακας $D_P = (S_P(\mathbf{x}_i,\mathbf{x}_j))_{m\times m}$ είναι θετικά ημιορισμένος, τότε η $d_P(\mathbf{x},\mathbf{y})$ μπορεί να λειτουργήσει ως μετρική απόστασης.

7.4.4. Ο αλγόριθμος GRAPH-RUNS

Στη συνέχεια θα χρησιμοποιήσουμε τη μέθοδο GRAPH σε συνδυασμό με το μέτρο απόστασης που εισαγάγαμε στην προηγούμενη ενότητα, με σκοπό να κατασκευάσουμε ένα νέο αλγόριθμο για τον προσδιορισμό του αριθμού των συστάδων σε ένα σύνολο δεδομένων. Επειδή ο πίνακας απόστασης D_p εξαρτάται από την επιλογή των παραμέτρων c, k και n, θα πρέπει πρώτα να ερευνήσουμε πώς αυτές οι παράμετροι επηρεάζουν τον τελικό αριθμό των συστάδων.

Αν δεν υπάρχει εκ των προτέρων πληροφορία για την απόσταση μεταξύ των συστάδων μπορεί να δημιουργηθεί ένας εμπειρικός κανόνας που να δίνει κατάλληλες τιμές στο c. Συμβολίζουμε με $\mathbf{x}_i = (x_{i1}, x_{i2}, \mathbf{K}, x_{ip})$, $i = 1,2,\mathbf{K},m$ τις m p-διάστατες παρατηρήσεις του σύνολο δεδομένων και με

$$\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i = (\bar{x}_1, \bar{x}_2, \mathbf{K}, \bar{x}_p)$$

το διάνυσμα των μέσων, όπου

$$\bar{\mathbf{x}}_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij}\ ,\ j = 1,2,\mathbf{K},p\,.$$

Η συνάρτηση Andrews για το $\bar{\mathbf{x}}$

$$\bar{f}_x(t) = \bar{x}_1/\sqrt{2} + \bar{x}_2\sin t + \bar{x}_3\cos t + \bar{x}_4\sin 2t + \bar{x}_5\cos 2t + \mathbf{K}$$

συμπίπτει με το μέσο των m ανεξάρτητων συναρτήσεων Andrews

$$f_i(t) = x_{i1}/\sqrt{2} + x_{i2}\sin t + x_{i3}\cos t + x_{i4}\sin 2t + x_{i5}\cos 2t + \mathbf{K}\,,$$

δηλαδή

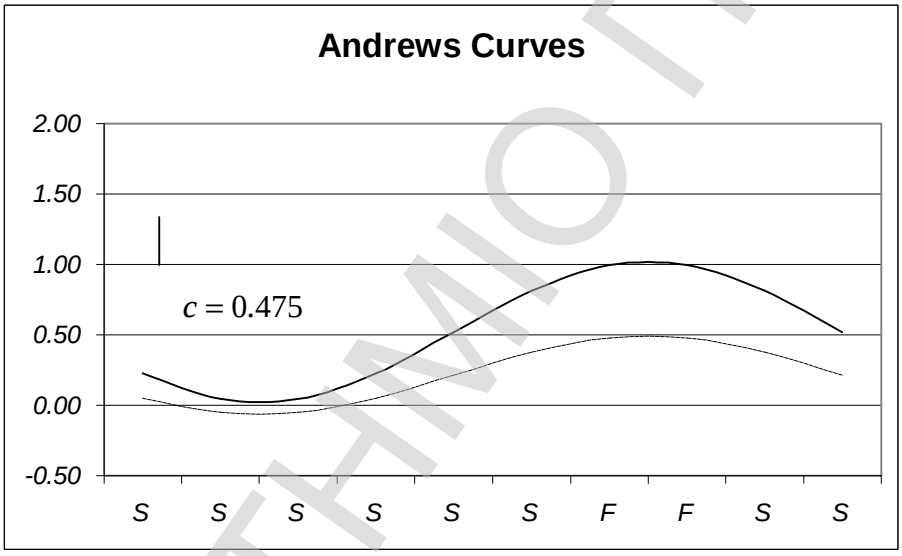
$$\bar{f}_x(t) = \frac{1}{m}\sum_{i=1}^{m} f_i(t).$$

Μια λογική επιλογή για το c (Σχήμα 7.6) φαίνεται να είναι η ποσότητα

$$\bar{c} = \frac{1}{mn}\sum_{r=1}^{n}\sum_{i=1}^{m}\left| f_i(t_r) - \bar{f}(t_r)\right|$$

η οποία μετρά τη μέση (απόλυτη) απόκλιση όλων των (m το πλήθος) συναρτήσεων Andrews από την $\bar{f}$ υπολογισμένη για τα n σημεία της διαμέρισης P.

Σχήμα 7.6: Απόσταση c δύο καμπύλων για δύο πολυδιάστατες παρατηρήσεις



Για συγκεκριμένο c, ο πίνακας απόστασης D_p εξαρτάται από την επιλογή του μήκους ροής k. Μικρές τιμές του k (σε σύγκριση με το n) παράγουν μεγάλο αριθμό επικαλυπτόμενων ροών επιτυχίας $M_{n,k}$ και αντίστροφα. Για προκαθορισμένο n, το επόμενο βήμα είναι να απαντηθεί το ερώτημα ποιο είναι το λογικό μήκος ροής που μπορεί να χρησιμοποιηθεί στην κατασκευή του πίνακα D_p. Η κατανομή του μέγιστου μήκους ροής L_n σε n όμοιες ανεξάρτητες τυχαίες δοκιμές $Z_1, Z_2, \mathbf{K}, Z_n$ με σταθερές πιθανότητες επιτυχίας p, πολύ γρήγορα προσεγγίζει την ασυμπτωτική μορφή της (Balakrishnan and Koutras (2001)). Συνεπώς, ακόμα και για μέτριες ή μικρές τιμές του n, οι ασυμπτωτικές μορφές τους

παρέχουν καλές προσεγγίσεις των ποσοτήτων που μας ενδιαφέρουν. Ο ασυμπτωτικός μέσος του μέγιστου μήκους ροής L_n δίνεται από τον τύπο

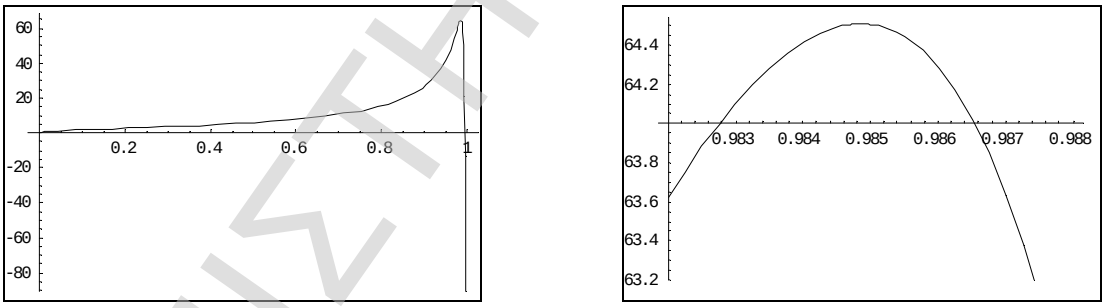
$$E(L_n) = \log_{1/p}[n(1-p)] + \frac{g}{\ln(1/p)} - \frac{1}{2} + r(n) + e(n) \quad (7.2)$$

όπου $g = 0.577\mathbf{K}$ είναι η σταθερά του Euler, $\lim_{n\to\infty} e(n) = 0$ και $r(n)$ είναι μια περιοδική συνάρτηση του $\log_{1/p} n$. Αν και για το πρόβλημα που μας ενδιαφέρει, οι n δοκιμές δεν είναι ανεξάρτητα κατανεμημένες, μπορούμε να εκμεταλλευτούμε τον προηγούμενο τύπο για να πάρουμε μια εκτιμήτρια για τη μέση τιμή του L_n. Έτσι, για $n = 100$ παίρνουμε

$$E(L_{100}) = \log_{1/p}[100(1-p)] + \frac{0.577}{\ln(1/p)} - 0.5$$

και μπορεί αμέσως να ελεγχθεί αριθμητικά ότι η μέγιστη τιμή που λαμβάνει η $E(L_{100})$, ως συνάρτηση του p, είναι μικρότερη του 65 (Σχήμα 7.7). Συνεπώς, είναι αρκετά απίθανο να παρατηρήσουμε ροή επιτυχίας μήκους μεγαλύτερου του 65, άσχετα με τη τιμή των πιθανοτήτων επιτυχίας. Άρα μπορούμε να απορρίψουμε όλες τις τιμές του k που υπερβαίνουν το 65.

Σχήμα 7.7: Το ασυμπτωτικό μήκος της $E(L_{100})$



Επίσης, γειτονικές τιμές του k συνήθως παράγουν τον ίδιο πίνακα απόστασης D_p και συνεπώς οδηγούν στην ίδια ομαδοποίηση των δεδομένων. Έτσι, αντί να δοκιμάζουμε όλες τις τιμές του k μπορούμε να τις μειώσουμε στο μισό, στο ένα τρίτο κ.ο.κ.

Οι παραπάνω παρατηρήσεις οδηγούν στον επόμενο αλγόριθμο:

Βήμα 1: Όρισε τους αριθμούς n_k και n_c των k και c που θα χρησιμοποιηθούν.

Βήμα 2: Θεώρησε τη διαμέριση $P: -p < t_1 < t_2 < \mathbf{K}\, t_n = p$ του $[-p, p]$, όπου $t_r = -p + (2p/99)(r-1)$, $r = 1,2,\mathbf{K},100$ $(n = 100)$.

Βήμα 3: Επανέλαβε τα βήματα (4) και (5) για όλα τα $i = 1,2,\mathbf{K},n_k$ και $j = 1,2,\mathbf{K},n_c$.

Βήμα 4: Υπολόγισε τα $k_i = i\left[\frac{65}{n_k}\right]$, $c_j = 0.5\bar{c} + (j-1)\frac{0.5\bar{c}}{n_c - 1}$.

Βήμα 5: Υπολόγισε τον πίνακα αποστάσεων D_p για τις τιμές k_i και c_j που υπολογίστηκαν στο βήμα (4) και εφάρμοσε τον αλγόριθμο GRAPH. Έστω N_{ij} ο αριθμός των συστάδων που προκύπτουν.

Βήμα 6: Υπολόγισε το $\max\{N_{ij} : 1 \le i \le n_k, 1 \le j \le n_c\}$. Αυτός είναι ο προτεινόμενος αριθμός συστάδων.

Ο αλγόριθμος αυτός δίνει ενδιαφέροντα αποτελέσματα και φαίνεται να βελτιώνει την ικανότητα του αλγόριθμου GRAPH. Ένα μειονέκτημα του αλγορίθμου μας είναι ότι είναι πιο χρονοβόρος από τον GRAPH, καθώς κατά την εφαρμογή του απαιτεί την επανάληψη της διαδικασίας GRAPH πολλές φορές μέχρι να φθάσει στο τελικό αποτέλεσμα.

7.4.5. Μελέτη με Χρήση Προσομοίωσης

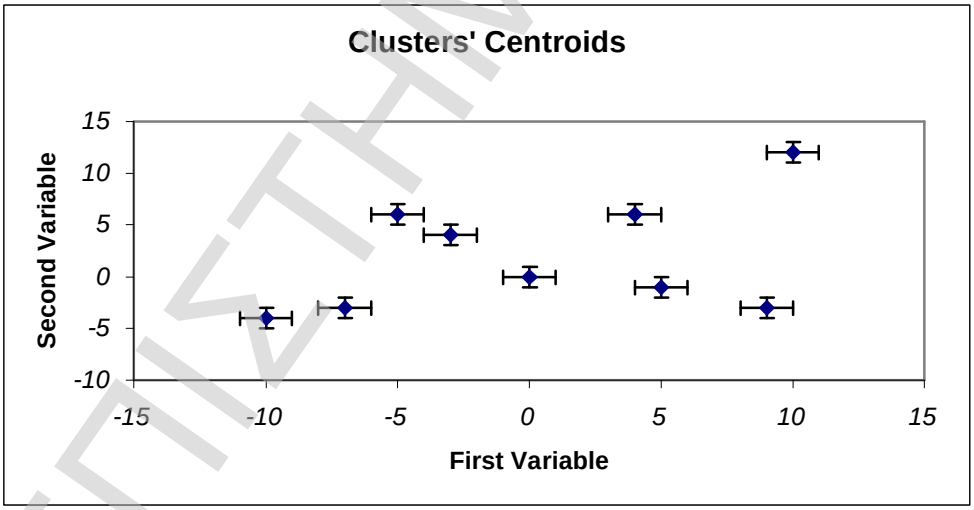
Για να δείξουμε την αποδοτικότητα του αλγόριθμου RUNS-GRAPH, συγκρίναμε την λειτουργία του με αυτή της κλασικής διαδικασίας GRAPH με τη βοήθεια δυο προσομοιωμένων σύνολο δεδομένων.

Τα τεχνητά δεδομένα αποτελούνταν από 9 διακριτές συστάδες (Πίνακας 7.4 και Σχήμα 7.8) στο χώρο των δυο, τριών και τεσσάρων διαστάσεων $(p = 2,3,4)$. Τα μέλη κάθε συστάδας ήταν κατανεμημένα σύμφωνα με μια πολυδιάστατη Κανονική κατανομή $N_p(\boldsymbol{\mu}_i, s^2\mathbf{I}_p)$, $i = 1,2,\mathbf{K},9$.

Πίνακας 7.4: Τα κεντροειδή των 9 διακριτών συστάδων

Cluster	*Var*1	*Var*2	*Var*3	*Var*4
1	0	0	0	0
2	4	6	6	4
3	-7	-3	-3	-1
4	-5	6	6	3
5	9	-3	-3	-5
6	-3	4	9	5
7	-10	-4	3	3
8	5	-1	-9	5
9	10	12	10	10

Σχήμα 7.8: Τα κεντροειδή των 9 διακριτών συστάδων στον διδιάστατο χώρο



Όταν επιλέγεται τυπική απόκλιση ίση με 0.5, τότε οι συστάδες είναι σχεδόν μη επικαλυπτόμενες ενώ για τυπική απόκλιση περίπου ίση με 1.0, οι συστάδες επικαλύπτονται. Παρήχθησαν τριάντα παρατηρήσεις για κάθε συστάδα και στη συνέχεια εφαρμόσθηκε ο αλγόριθμος GRAPH χρησιμοποιώντας είτε την Ευκλείδεια απόσταση είτε πίνακα απόστασης

D_p. Προκειμένου να υπολογιστεί η αποδοτικότητα της μεθόδου, στον προσδιορισμό του αληθινού αριθμού των συστάδων που είναι παρούσες στα δεδομένα, επαναλάβαμε ολόκληρη τη διαδικασία 100 φορές και καταγράφηκε ο αριθμός επιτυχών εφαρμογών.

Τα αποτελέσματα της μελέτης παρουσιάζονται στους Πίνακες 7.5 (για τυπική απόκλιση ίση με 0.5), 7.6 (για τυπική απόκλιση ίση με 1.0), και 7.7 (για διαφορετικές τυπικές αποκλίσεις σε κάθε διάσταση). Οι στήλες κάτω από τον τίτλο "GRAPH" παρέχουν τις λεπτομέρειες σχετικά με την απόδοση του αρχικού αλγορίθμου GRAPH (όπου χρησιμοποιείται η συνηθισμένη Ευκλείδεια απόσταση) ενώ οι στήλες κάτω από τον τίτλο "RUNS-GRAPH" παρέχουν τα αντίστοιχα αποτελέσματα όταν υιοθετείται ο νέος αλγόριθμος.

Η δεύτερη στήλη των πινάκων 7.5 και 7.6 δείχνει τον πραγματικό αριθμό συστάδων που είναι παρούσες στο σύνολο των δεδομένων και στην παρένθεση τα κεντροειδή των συγκεκριμένων συστάδων που χρησιμοποιήθηκαν για την προσομοίωση. Στις στήλες με τίτλο ετικέτα "*N*", έχουμε καταχωρήσει τον αριθμό επιτυχών ευρέσεων (από τις 100 προσομοιωμένες επαναλήψεις) του σωστού αριθμού συστάδων με τη μέθοδο RUNS-GRAPH και GRAPH αντίστοιχα. Επιπλέον, στις στήλες που χαρακτηρίζονται ως "*N*-1" και "*N*+1", έχουμε καταγράψει τον αριθμό των "σχεδόν επιτυχών" ευρέσεων. Τέλος, οι στήλες που χαρακτηρίζονται ως "Άθροισμα" (Sum) περιέχουν το σύνολο των επιτυχών και σχεδόν επιτυχών ευρέσεων του σωστού αριθμού ομάδων.

Οι υπόλοιπες γραμμές των πινάκων αυτών δίνουν τα αντίστοιχα αποτελέσματα για δεδομένα μεγαλύτερων διαστάσεων. Οι πρώτες δύο συντεταγμένες των συστάδων κρατήθηκαν αμετάβλητες και προστέθηκε μια τρίτη και μια τέταρτη συντεταγμένη (Var3 και Var4 του Πίνακα 7.4).

Η απόδοση των αλγορίθμων συγκριτικά εξετάστηκε και στην περίπτωση των άνισα διεσπαρμένων πληθυσμών (η διασπορά δεν είναι η ίδια για όλες τις κατευθύνσεις). Και πάλι έγινε χρήση συστάδων των οποίων τα μέλη κατανέμονται σύμφωνα με την πολυδιάστατη κανονική διανομή $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_p)$ με $\boldsymbol{\mu}_i$, $i = 1, 2, ..., 9$ του Πίνακα 7.4 και όπου $\boldsymbol{\Sigma}_p$ είναι ένας διαγώνιος πίνακας της μορφής $\begin{bmatrix} s_1^2 & 0 & ... & 0 \\ 0 & s_2^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & s_p^2 \end{bmatrix}$ με $s_i = 0.50, 0.75, 1.00, 1.25$. Η επιλογή αυτή είναι σε συμφωνία με τη μεθοδολογία που προτάθηκε από τους Milligan and Cooper

(1985), δηλαδή οι προκύπτουσες συστάδες να είναι εσωτερικά συνεκτικές και καλά χωρισμένες στην πρώτη διάσταση, ο διαχωρισμός τους να μειώνεται συνεχώς καθώς προχωράμε στη δεύτερη, την τρίτη κλπ διαστάσεις. Παρήχθησαν δέκα παρατηρήσεις για κάθε συστάδα και ολόκληρη η διαδικασία επαναλήφθηκε 50 φορές για κάθε μεμονωμένο σύνολο στοιχείων. Για τα δεδομένα αυτά, δίνεται ο αριθμός επιτυχών και σχεδόν επιτυχών ευρέσεων του αριθμού συστάδων που υπάρχουν στο σύνολο στοιχείων, στον Πίνακα 7.7.

Από τα αριθμητικά αποτελέσματα που δίνονται στους Πίνακες 7.5, 7.6, 7.7 καταλήγουμε στα εξής συμπεράσματα:

Η γενική απόδοση του αλγόριθμου RUNS-GRAPH είναι καλύτερη από αυτή του αλγόριθμου GRAPH και η αποδοτικότητα και των δυο αλγόριθμων μειώνεται όσο αυξάνεται το p. Παρόλα αυτά, ο αλγόριθμος RUNS-GRAPH διατηρεί την ικανότητα να αποκαλύπτει την ακριβή δομή των συστάδων ή να παρέχει ικανοποιητική προσέγγιση αυτής ακόμα και στην περίπτωση που έχουμε $p > N + 2$.

Ένα άλλο σημαντικό στοιχείο είναι ότι, αν έχουμε κάποια εκ των προτέρων πληροφορία για τον αριθμό (κατά προσέγγιση) των συστάδων που υπάρχουν στο δείγμα, τότε μπορούμε να βελτιώσουμε την ταχύτητα και τη γενική απόδοση του αλγόριθμου RUNS-GRAPH προσαρμόζοντας κατάλληλα το εύρος της παραμέτρου c. Ειδικότερα, αν υπάρχει υποψία ότι ο αριθμός των συστάδων ξεπερνά πάρα πολύ τον αριθμό των διαστάσεων, τότε το εύρος των τιμών της παραμέτρου c μπορεί να περιοριστεί στο διάστημα $[0.5\bar{c}, \bar{c}]$. Αν είναι πιθανό ο αριθμός των συστάδων να ξεπερνά κατά πολύ τον αριθμό των διαστάσεων, τότε το εύρος των τιμών μπορεί να περιοριστεί στο διάστημα $[0.9\bar{c}, 1.1\bar{c}]$. Τέλος, αν ο αριθμός των συστάδων είναι περίπου ίσος με τον αριθμό των διαστάσεων, τότε η επιλογή $c = \bar{c}$ δίνει τα καλύτερα αποτελέσματα.

Συμπερασματικά, μπορούμε να πούμε ότι, η ομαλή λειτουργία της προτεινόμενης διαδικασίας έγκειται από τη μια στη σημαντική ευαισθησία του στατιστικού ροών $M_{n,k}$ για την ανίχνευση συστάδων παρόμοιων αποτελεσμάτων και από την άλλη στη σημαντική περιγραφική δύναμη των διαγραμμάτων Andrews. Φαίνεται λογικό στο μέλλον να προχωρήσουμε σε μια βαθύτερη μελέτη των δυο αυτών παραγόντων και να ερευνήσουμε αν μπορούν να βελτιωθούν κατάλληλα ώστε η τεχνική ομαδοποίησης να δίνει καλύτερα αποτελέσματα. Μια πρώτη κατεύθυνση είναι να θεωρήσουμε διαφορετικά είδη ροών ή ακόμα καλύτερα των scans. Μια άλλη πολλά υποσχόμενη κατεύθυνση είναι να αντικαταστήσουμε

τις κλασικές συναρτήσεις Andrews με τις αντίστοιχες τροποποιημένες των Khatree and Naik (2000) Embrechts and Herzberg (1991), Embrechts at al (1995) και άλλων.

Πίνακας 7.5: Αποτελέσματα μελέτης στην περίπτωση τυπικής απόκλισης ίσης με 0.5

p	N	*RUNS-GRAPH*				*GRAPH*			
		N-1	*N*	*N*+1	Sum	*N*-1	*N*	*N*+1	Sum
2	2 (μ_1, μ_2)	0	67	33	100	0	100	0	100
	2 (μ_4, μ_5)	0	100	0	100	0	100	0	100
	2 (μ_4, μ_6)	0	56	43	99	0	89	11	100
	3 (μ_1, μ_2, μ_3)	0	100	0	100	0	100	0	100
	3 (μ_2, μ_4, μ_6)	73	27	0	100	100	0	0	100
	3 (μ_3, μ_4, μ_5)	0	100	0	100	100	0	0	100
	4 (μ_1, μ_2, μ_3, μ_4)	0	100	0	100	5	95	0	100
	5	0	100	0	100	75	9	0	84
3	3	0	100	0	100	13	77	0	90
	4	0	100	0	100	17	30	0	47
	5	0	100	0	100	29	4	0	33
	6	31	69	0	100	0	0	0	0
	7	43	56	0	99	0	0	0	0
	8	81	0	0	81	0	0	0	0
	9	78	0	0	78	0	0	0	0
4	4	0	100	0	100	0	9	0	9
	5	0	100	0	100	0	0	0	0
	6	57	43	0	100	0	0	0	0
	7	86	1	0	87	0	0	0	0
	8	76	0	0	76	0	0	0	0
	9	83	4	0	87	0	0	0	0

Πίνακας 7.6: Αποτελέσματα μελέτης στην περίπτωση τυπικής απόκλισης ίσης με 1.0

p	*N*	*RUN-GRAPH*				*GRAPH*			
		N-1	*N*	*N*+1	Sum	*N*-1	*N*	*N*+1	Sum
2	2	0	37	62	99	0	97	3	100
	3	24	56	20	100	0	100	0	100
	4	9	57	13	79	2	98	0	100
	5	5	64	4	73	48	48	0	96
	6	3	0	0	3	64	18	2	84
3	3	2	53	44	99	0	100	0	100
	4	2	65	31	98	20	70	0	90
	5	4	68	26	98	19	47	0	66
	6	65	14	0	79	14	0	0	14
	7	35	4	0	39	0	0	0	0
4	4	2	35	50	87	17	59	0	76
	5	3	36	15	54	10	11	0	21
	6	51	13	0	64	2	0	0	2
	7	57	9	0	66	0	0	0	0

Πίνακας 7.7: Αποτελέσματα μελέτης στην περίπτωση στην περίπτωση άνισων διακυμάνσεων

p	N	*RUNS-GRAPH*				*GRAPH*			
		N-1	N	N+1	Sum	N-1	N	N+1	Sum
2	2	0	26	26	52	0	49	49	98
	3	1	46	47	94	0	50	50	100
	4	0	45	45	90	11	39	50	100
	5	0	41	41	82	28	11	39	78
	6	39	1	40	80	32	0	32	64
3	3	0	41	41	82	3	47	50	100
	4	0	45	45	90	11	34	45	90
	5	0	41	41	82	28	11	39	78
	6	37	9	46	92	4	0	4	8
	7	39	5	44	88	0	0	0	0
4	4	0	15	15	30	12	19	31	62
	5	0	25	25	50	22	3	25	50
	6	32	10	42	84	1	0	1	2
	7	27	10	37	74	0	0	0	0

7.5. Βιολογικές εφαρμογές

Το γενετικό υλικό DNA είναι ένα μεγάλου μήκους μόριο με μορφή νήματος, που αποτελείται από "νουκλεοτίδια" ή "βάσεις", οι οποίες συνδέονται σε σειρά. Υπάρχουν 4 είδη βάσεων, το σύνολο των οποίων συμβολίζεται ως {A,C,G,T}. Το DNA αποτελείται, συνήθως, από δύο ελικοειδείς αλυσίδες αντίθετης κατεύθυνσης.

Ένας μεγάλος αριθμός ερευνητών ασχολείται με τον προσδιορισμό του είδους και των ιδιοτήτων διαφόρων οργανισμών, αξιοποιώντας τις γενετικές πληροφορίες που περιέχονται στο DNA. Για το λόγο αυτό δημιουργήθηκαν βάσεις δεδομένων, οι οποίες αποτελούνται από καταλόγους μεγάλων τέτοιων ακολουθιών. Για να γίνει πιο εύκολα η μελέτη τους, οι ακολουθίες αυτές μπορούν να θεωρηθούν γραμμικές.

Συνήθως, ενδιαφερόμαστε για τη σύγκριση ακολουθιών DNA, με σκοπό την ομαδοποίηση ή μη των οργανισμών από τους οποίους προέρχονται. Αν οι ακολουθίες αυτές έχουν κάποιο αριθμό τμημάτων που ταυτίζονται, τότε οι οργανισμοί που μελετώνται έχουν παρόμοιες ιδιότητες.

Από στατιστικής απόψεως, ας θεωρήσουμε ως επιτυχία την ταύτιση των βάσεων (ανάμεσα στις δύο ακολουθίες που συγκρίνονται) και ως αποτυχία τη διαφοροποίηση τους. Η απαρίθμηση των τμημάτων μήκους k που περιέχουν τουλάχιστον r επιτυχίες παρουσιάζει

ιδιαίτερο ενδιαφέρον. Σε αυτή την περίπτωση, ο υπολογισμός των αντίστοιχων πιθανοτήτων βασίζεται στη μελέτη των στατιστικών συναρτήσεων σάρωσης. Οι Arratia et al. (1989), Karlin and Cardon (1994) και Karlin and Macken (1991) ανέπτυξαν προσεγγίσεις για τις κατανομές αυτές με τη βοήθεια της μεθόδου Chen-Stein.

Από την άλλη μεριά, κατά τη μελέτη ακολουθιών αμινοξέων, οι μοριακοί βιολόγοι θεωρούν διάφορα σχέδια ταξινόμησης τους. Παραδείγματος χάριν, χημικό αλφάβητο (οκτώ γράμματα), αλφάβητο σύνδεσης (τέσσερα γράμματα-βάσεις) ή αλφάβητο φορτίου {-1,0,+1}. Τέτοια προβλήματα μελετήθηκαν από τους Karlin and Ghandour (1985). Οι ερευνητές συγκρίνουν ακολουθίες από διαφορετικά είδη οργανισμών και εξετάζουν τμήματα μεγάλου μήκους, στα οποία παρατηρούνται ομοιότητες στις περισσότερες θέσεις. Σκοπός τους είναι η ανάπτυξη κριτηρίων με τη βοήθεια των οποίων θα αποφασίζουν πότε μια τέτοια ταύτιση είναι ασυνήθιστα μεγάλη (Glaz and Naus (1991)). Στην περίπτωση που εξετάζουμε τα φορτία $\{-1,0,+1\}$ που περιλαμβάνονται σε ακολουθίες από αμινοξέα είναι σαν να θεωρούμε ακολουθίες με τρίτιμες δοκιμές.

Η τυχαία μεταβλητή $S_{n,k}$ που μελετήθηκε στο κεφάλαιο 3 αλλά και η διδιάστατη κατανομή $(S_{n,k}^{(S)}, N_{n,r}^{(F)})$ που μελετήθηκε στο κεφάλαιο 4 μπορούν να χρησιμοποιηθούν ως κατάλληλα μοντέλα για την ανάπτυξη κριτηρίων με τη βοήθεια των οποίων οι ειδικοί επιστήμονες θα αποφασίζουν πότε μια τέτοια ταύτιση είναι ασυνήθιστα μεγάλη.

7.6. Κριτήρια εκμάθησης στη Ψυχολογία

Η χρήση κριτηρίων επίδοσης είναι ένα συνηθισμένο εργαλείο στις πειραματικές μελέτες μνήμης και εκμάθησης. Τέτοιοι έλεγχοι εφαρμόζονται από ψυχολόγους για να αποφασίσουν τον τερματισμό μιας θεραπευτικής διαδικασίας. Ένα από τα πιο γνωστά κριτήρια είναι το "κριτήριο ροής" (runs criterion) του Grand (1946), το οποίο απορρίπτει την υπόθεση της μη εκμάθησης, αν το αντικείμενο μελέτης απαντά σωστά σε συγκεκριμένο αριθμό διαδοχικών ερωτήσεων. Υποθέτοντας ότι υπάρχουν δύο είδη απαντήσεων (σωστή ή λανθασμένη), ο Grand (1946, 1947) έδωσε πίνακες για τον προσδιορισμό του μέγιστου αριθμού ερωτήσεων, για τον οποίο η πιθανότητα τυχαίας επίτευξης του κριτηρίου είναι μικρότερη από μια συγκεκριμένη τιμή (γνωστός από τον Bogartz (1965) ως "κρίσιμη τιμή του ελέγχου" (criterion risk)). Οι πιθανότητες που σχετίζονται με το κριτήριο αυτό μπορούν να

υπολογισθούν εύκολα με τη βοήθεια της θεωρίας του Κεφαλαίου 3 που αφορά ροές επιτυχιών προκαθορισμένου μήκους. Μια γενίκευση των παραπάνω αποτελεί το ακόλουθο κριτήριο: Ας υποθέσουμε ότι το υπό εξέταση άτομο υποβάλλεται ταυτόχρονα σε δύο διαφορετικές ασκήσεις (δοκιμασίες). Κάθε δοκιμασία αποτελείται από μια σειρά ερωτήσεων. Ένα άτομο αξιολογείται ως ικανό, αν δώσει τουλάχιστον k διαδοχικές σωστές απαντήσεις στην πρώτη δοκιμασία ή τουλάχιστον r διαδοχικές σωστές απαντήσεις στη δεύτερη. Επιπλέον, ενδιαφέρον παρουσιάζει η μελέτη μοντέλων με εξάρτηση μεταξύ διαδοχικών απαντήσεων, με την έννοια ότι μια σωστή απάντηση μπορεί να προκαλέσει με μεγαλύτερη πιθανότητα μια επόμενη σωστή απάντηση και αντίστροφα περίπτωση Μαρκοβιανής εξάρτησης.

Μια ιδιαίτερα χρήσιμη τροποποίηση των άνω κριτηρίων εκμάθησης περιγράφεται στο υπόλοιπο της παραγράφου (βλέπε επίσης Antzoulakos et al. (2005)). Ένα από τα παλαιότερα και πιο γνωστά κριτήρια στη ψυχολογία είναι όπως προαναφέρθηκε το κριτήριο των ροών του Grand (1946) που απορρίπτει την μηδενική υπόθεση της μη εκμάθησης εάν το άτομο υπό μελέτη δίνει τη σωστή απάντηση σε έναν προδιευκρινισμένο αριθμό διαδοχικών δοκιμών.

Ας υποθέσουμε ότι η πιθανότητα της σωστής απάντησης σε μια δοκιμή είναι ίση με το p και ότι η εμφάνιση r διαδοχικών σωστών απαντήσεων θεωρείται ένδειξη ότι το άτομο έχει εκπαιδευτεί. Η δεσμευμένη κατανομή της τ.μ. $X_{n,1}^{(2)}$ δεδομένου ότι $X_{n,r}^{(1)} = 0$ (πρόκειται για τη δεσμευμένη κατανομή της διδιάστατης τυχαίας μεταβλητής $(S_{n,k}, N_{n,r})$, με $S_{n,k} = X_{n,k}^{(1)}, N_{n,r} = X_{n,r}^{(2)}$, που μελετήθηκε στο Κεφάλαιο 4), παρέχει τις πληροφορίες για το συνολικό αριθμό των λανθασμένων απαντήσεων στις n δοκιμές δεδομένου ότι, μέχρι εκείνη την δοκιμή, το άτομο δεν έχει εκπληρώσει το κριτήριο των ροών του Grand (1946).

Εάν επιθυμήσουμε να αγνοήσουμε τις απομονωμένες λανθασμένες απαντήσεις, ο μέσος όρος των ανεπιτυχών απαντήσεων στις n δοκιμές δεδομένου ότι το άτομο δεν κατάφερε να εκπληρώσει το κριτήριο των ροών, είναι ίσος με $E[X_{n,k}^{(2)} = x \mid X_{n,r}^{(1)} = 0]$.

Τέλος, η αναμενόμενη τιμή $E[X_{n,k}^{(2)} = x \mid X_{n,r}^{(1)} = 0]$ δίνει το μέσο όρο των ανεπιτυχών απαντήσεων όταν αγνοούνται οι φραγμοί των λανθασμένων απαντήσεων του μήκους $1,2\ldots,k\text{-}1$.

7.7. Άλλες Εφαρμογές

Πολλές είναι οι εφαρμογές της θεωρίας των ροών και σχηματισμών και του χρόνου αναμονής που συνδέονται με αυτά σε διάφορους τομείς, όπως, π.χ. κοινωνιολογία, οικολογία, ραδιοαστρονομία, κλπ.

Πιο συγκεκριμένα, ένας κοινωνιολόγος μελετά τη συμπεριφορά μιας ομάδας ανθρώπων που είναι διατεταγμένοι σε μια σειρά, διακρίνοντάς τους σε δύο κατηγορίες ως προς το φύλο τους. Συνήθως, μελετάται η εμφάνιση ομαδοποιήσεων του ίδιου φύλου, η οποία ισοδυναμεί με την εμφάνιση ροών του ενός ή του άλλου φύλου (είδους). Από την άλλη μεριά, η θεωρία των ροών μπορεί να εφαρμοσθεί στη μελέτη της κινητικότητας του πληθυσμού ως προς την κοινωνική τάξη (mobility theory). Οι Kemeny and Snell (1976) μελέτησαν το πρόβλημα αυτό θεωρώντας ότι το κοινωνικό σύνολο απαρτίζεται από τρία βασικά στρώματα-τάξεις : κατώτερη, μέση και ανώτερη.

Στον κλάδο της οικολογίας, ιδιαίτερο ενδιαφέρον παρουσιάζει η μελέτη στατιστικών συναρτήσεων οι οποίες σχετίζονται με την εξάπλωση κάποιων τύπων βλάστησης μέσα σε συγκεκριμένες γεωγραφικές περιοχές (για παράδειγμα δασικές εκτάσεις). Επίσης, ενδιαφέρον παρουσιάζει η μελέτη της κατανομής στατιστικών που σχετίζονται με την εξάπλωση κάποιας ασθένειας που τυχόν προσβάλλει κάποιο είδος δέντρου (Pielou (1969)). Για την υλοποίηση της μελέτης οι δασικές εκτάσεις χωρίζονται σε λωρίδες, δημιουργώντας με αυτό τον τρόπο μεγάλες "γραμμικές" ακολουθίες δένδρων. Η εξαγωγή των συμπερασμάτων βασίζεται στη μελέτη της εμφάνισης ροών συγκεκριμένου μήκους από δένδρα ίδιου είδους ή στην εμφάνιση ροών με ασυνήθιστα μεγάλο μήκος. Επιπλέον, είναι σημαντική η μελέτη ομάδων διαδοχικών δένδρων (k-άδων) στις οποίες εμφανίζεται συγκεκριμένος αριθμός από το ίδιο είδος δένδρων. Παρόμοια προβλήματα εκφράζονται στη μελέτη της εμφάνισης συγκεκριμένων λέξεων σε προτάσεις ή κείμενα (Karlin and Ost (1987)). Με τα θέματα αυτά ασχολούνται γλωσσολόγοι, αλλά και ερευνητές της πληροφορικής.

Μια άλλη εφαρμογή της θεωρίας ροών επιτυχιών εμφανίζεται στον κλάδο της πληροφορικής, ο οποίος ασχολείται με την ανάπτυξη αλγορίθμων που συμπιέζουν και αποσυμπιέζουν αρχεία μεγάλου μεγέθους. Υπάρχουν δύο είδη αλγορίθμων. Το πρώτο είδος εφαρμόζεται σε κείμενα που έχουν μεγάλο αριθμό από επαναλαμβανόμενες λέξεις. Οι αλγόριθμοι αυτοί βασίζονται στον εντοπισμό των λέξεων που επαναλαμβάνονται συχνότερα. Στη συνέχεια, οι λέξεις και τα κενά αντικαθίστανται από αριθμούς. Αυτό έχει ως αποτέλεσμα

να μειώνεται σημαντικά ο όγκος που καταλαμβάνουν τα αρχεία. Στην αποσυμπίεση των αρχείων, για την αξιολόγηση του αλγορίθμου, ενδιαφέρον παρουσιάζει η μελέτη του αριθμού των προτάσεων ή των γραμμών (k-άδων) που περιλαμβάνουν περισσότερες από r συμπιεσμένες λέξεις. Αυτό αποτελεί μια απλή εφαρμογή των τυχαίων μεταβλητών τμημάτων σάρωσης. Το δεύτερο είδος αλγορίθμων, αφορά κυρίως δυαδικά αρχεία (δηλαδή, αρχεία που περιέχουν συνδυασμούς των 0 και 1). Στην περίπτωση αυτή, μια ακολουθία, π.χ. 0001100100010000 αντικαθίσταται από την ακολουθία 03 12 02 11 03 11 04, δηλαδή κάθε ροή από 1 ή 0 αντικαθίσταται από μια δυάδα που αποτελείται από το σύμβολο και το μήκος της ροής του συμβόλου. Αν οι ροές είναι μικρού μήκους, τότε η συμπίεση είναι προφανώς μικρότερη.

7.8. Ανακεφαλαίωση

Στο Κεφάλαιο αυτό αναφερθήκαμε διεξοδικά στις εφαρμογές της θεωρίας ροών επιτυχιών και γενικά των σχηματισμών σε διάφορα επιστημονικά πεδία. Ιδιαίτερη προσοχή δόθηκε σε νέα αποτελέσματα που αφορούν τους ελέγχους τυχαιότητας, τη θεωρία αξιοπιστίας συστημάτων, την πολυμεταβλητή στατιστική τεχνική της ανάλυσης συστάδων (cluster analysis), στη ψυχολογία καθώς και περιληπτικά σε μία σειρά παλαιότερων εφαρμογών στις αλυσίδες DNA, στην οικονομική ανάπτυξη και ανταγωνιστικότητα, στην οικολογία και τέλος στη μετεωρολογία.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Agin, M.A. and Godbole, A.P. (1992). A new exact runs test for randomness, In *Computing Science and Statistics* (Eds., C. Page and R. Le Page), Proceedings of the 22nd Symposium on the Interface, pp. 281-285, Springer-Verlag, New York.

Aki, S. (1992). Waiting time problems for a sequence of discrete random variables, *Annals of the Institute of Statistical Mathematics,* **44,** 363-378.

Aki, S. (1997). On sooner and later problems between success and failure runs, In *Advances in Combinatorial Methods and Applications to Probability and Statistics* (Ed., N. Balakrishnan), pp. 385-400, Birkhauser, Boston.

Aki, S. and Hirano, K. (1988). Some characteristics of the binomial distribution of order *k* and related distributions, In *Statistical Theory and Data Analysis* II (Ed., K. Matusita), 211-222, Elsevier Science Publishers, B.V., North Holland.

Aki, S. and Hirano, K. (1993). Discrete distributions related to succession events in a two-state Markov chain, In *Statistical Science and Data Analysis,* (Eds., K. Matusita, M.L. Puri and T. Hayakawa), pp. 467-474, VSP International Science Publishers, Zeist.

Aki, S. and Hirano, K. (2004). Waiting time problems for a two-dimensional pattern, *Annals of the Institute of Statistical Mathematics,* **56,** 169-182.

Aki, S., Balakrishnan, N. and Mohanty, S.G. (1996). Sooner and later waiting time problems for success and failure runs in higher order Markov dependent trials, *Annals of the Institute of Statistical Mathematics,* **48,** 773-787.

Aki, S., Kuboki, H. and Hirano, K. (1984). On discrete distributions of order *k*, *Annals of the Institute of Statistical Mathematics,* **36,** 431-440.

Alexandrou, V. (1997). A study of enumerating random variables in sequences of trials by the aid of Markov chains, and applications, *Ph.D. Thesis,* University of Athens, Greece (in Greek).

Alt, F.B, and Smith, N.D. (1988). Multivariate Process Control. *Handbook of Statistics,* P.R. Krishnaiah and C.R. Rao (eds), Elsevier Science Publishers: North-Holland V.7, 333-351.

Alwan, L.C. (1986). Cusum quality control - Multivariate approach. *Communications in Statistics - Theory and Methods*, **15**, 3531-3543.

Alwan, L.C, Champ, C.W., and Maragah H.D. (1994). Study of the Average Run Lengths for Supplementary Runs Rules in the Presence of Autocorrelation, *Communications in Statistics – Theory and Methods*, **23**, 2, 373-391.

Andrews, D.F. (1972). Plots of high dimensional data, *Biometrics*, **28**, 125-136.

Antzoulakos, D.L. (1999). On waiting time problems associated with runs in Markov dependent trials, *Annals of the Institute of Statistical Mathematics*, **51,** 323-330.

Antzoulakos, D.L. (2001). Waiting times for patterns in a sequence of multistate trials, *Journal of Applied Probability*, **38**, 508-518.

Antzoulakos, D.L. (2003a). Waiting times and number of appearances of runs: A unified approach, *Communications in Statistics – Theory and Methods*, **32**, 7, 1289-1315.

Antzoulakos, D.L. (2003b). *Academic Notes on Statistical Quality Control*, Department of Statistics and Insurance Science, University of Piraeus, Piraeus.

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**, 4, 865-884.

Antzoulakos, D.L., Bersimis, S. and Koutras, M.V. (2004). Waiting Times Associated with the Total Number of Run Lengths, In *Mathematical and Statistical Methods in Reliability*, Edited by B. Lindqvist & K. Doksum, World Scientific.

Antzoulakos, D.L and Chadjiconstantinidis, S. (2001). Distributions of numbers of success runs of fixed length in Markov dependent trials, *Annals of the Institute of Mathematical Statistics*, 5**3**, 599-619.

Aparisi, F., Champ, C.W. and Diaz, J.C.G. (2004). A performance analysis of Hotelling's χ^2 control chart with supplementary runs rules, *Quality Engineering*, **16**, 13-22.

Arratia,R., Goldstein, L. and Gordon, L. (1989). Two moments suffice for Poisson approximations: The Chen-Stein method, *Annals of Probability*, **17,** 9-25.

Balakrishnan N. and Koutras M.V. (2002). *Runs and Scan with Applications,* New York: John Wiley.

Balakrishnan, N., Balasubramanian, K. and Viveros, R. (1993). On sampling inspection plans based on the theory of runs, *The Mathematical Scientist,* **18,** 113-126.

Balakrishnan, N., Balasubramanian, K. and Viveros, R. (1995). Start-up demonstration tests under correlation and corrective action, *Naval Research Logistics,* **42,** 1271-1276.

Balakrishnan, N., Mohanty, S.G. and Aki, S. (1997). Start-up demonstration tests under Markov dependence model with corrective actions, *Annals of the Institute of Statistical Mathematics,* **49,** 155-169.

Balasudramanian, K., Viveros, R and Balakrishnan, N (1993). Sooner and later waiting time problems for Markovian Bernoulli trials, *Statistics and Probability Letters,* **18,** 153-161.

Bersimis, S. (2000). Multivariate Statistical Process Control. MSc Thesis, Athens University of Economic and Business, Department of Statistics, Greece.

Bersimis, S., Psarakis, S. and Panaretos, J. (2005). Multivariate Statistical Process Control Charts: An Overview, *Quality and Reliability Engineering International*, *Accepted Paper*.

Binswanger, K. and Embrechts, P. (1994). Longest runs in coin tossing, *Insurance: Mathematics and Economics,* **15**, 139-149.

Bissel, A.F. (1978). An Attempt to Unify the Theory of Quality Control Procedures, *Bulletin in Applied Statistics,* 5, 113-128.

Bogartz, R. (1965). The criterion method: some analyses and remarks, *Psychological Bulletin*, **64**, 1-14.

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., Ch. A. Charalambides, M.V. Koutras and N. Balakrishnan), pp. 63-86, Chapman and Hall, Boca Raton, Florida.

Chadjiconstantinidis, S., Antzoulakos, D.L. and Koutras M.V. (2000). Joint distributions of successes, failures and patterns in enumeration problems, *Advances in Applied Probability,* **32**, 866-884.

Champ, C.W. (1992). Steady-State Run Length Analysis of a Shewhart Quality Control Chart with Supplementary Runs Rules, *Communications in Statistics – Theory and Methods,* **21**, 3, 765-777.

Champ C.W. and Woodall, W.H. (1987). Exact results for Shewhart control charts with supplementary runs rules, *Technometrics,* **29**, 393-399.

Champ, C.W., and Woodall W.H. (1990). A Program to Evaluate the Run Length Distribution of a Shewhart Control Chart with Supplementary Runs Rules, *Journal of Quality Technology,* **22**, 1, 68-73.

Champ, C.W. and Woodall, W.H. (1997). Signal Probabilities of Runs Rules Supplementing a Shewhart, *Communication in Statistics: Theory and Practice*, **26**, 1347-1360.

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.

Chao, M.T., Fu, J.C. and Koutras, M. V. (1995). A survey of the reliability studies of consecutive -*k*-out -of-*n*: F systems and its related systems, *IEEE Transactions on Reliability,* **44**, 120-127.

Chen, J. and Glaz, J. (1996). Two dimensional scan statistics, *Statistics & Probability Letters*, **31**, 59-68.

Child, I.L. (1946). A note on Grant's "New statistical criteria for learning and problem solution", *Psychological Bulletin,* **43**, 558-561.

Chryssaphinou, O. and Papastavridis, S. (1988). A limit-theorem on the number of overlapping appearances of a pattern in a sequence of independent trials, *Probability Theory and Related Fields,* **79**, 129-143.

Chryssaphinou, O. and Papastavridis, S. (1990). Limit distribution for a consecutive-*k*-out-of-*n*: F system, *Advances in Applied Probability,* **22**, 491-493.

Chryssaphinou, O., Papastavridis, S. and Tsapelas, T. (1993). On the number of overlapping success runs in a sequence of independent Bernoulli trials, *Applications of Fibonacci Numbers* (Eds., G. E. Bergum et al.), **5**, 103-112, Kluwer Academic Publishers, Amsterdam, The Netherlands.

Chryssaphinou, O., Papastavridis, S. and Tsapelas, T. (1994). On the waiting time of appearance of given patterns, *In Runs and Patterns in Probability* (Eds., A.P. Godbole and S.G. Papastavridis), pp. 231-241, Kluwer Academic Publishers, Amsterdam, The Netherlands.

Cochran, W.G. (1938). An extension of Gold's method for examining the apparent persistence of one type of weather, *Royal Meteorological Society Quarterly Journal,* **64**, 631-634.

Coleman, D. (1986). The Power Function of X-bar charts: Symbolic Manipulation of Recurrence Relations and Markov Chains, *Technical Report,* RCA, Princeton, NJ.

Crosier, R.B. (1988). Multivariate generalizations of cumulative sum quality-control schemes, *Technometrics*, **30**, 291-303.

David, F.N. (1947). A power function for tests of randomness in a sequence of alternatives, *Biometrika*, **34**, 335-339.

Dembo, A. and Karlin, S. (1992). Poisson approximations for *r*-scan processes, *Annals of Applied Probability*, **2**, 329-357.

De Moivre, A. (1738). *The Doctrine of Chance*, Chelsea Publishing Co., Third edition, New York.

Derman, C. and Ross, S.M. (1997). *Statistical Aspects of Quality Control*, Academic Press, San Diego, CA.

Divoky J.J., Taylor E.W. (1995). Detecting Process Drift with Combinations of Trend and Zonal Supplementary Runs Rules, *International Journal of Quality and Reliability Management*, **12**, 2, 1995, 60-71.

Doi, M. and Yamamoto, E. (1998). On the joint distribution of runs in a sequence of multi-state trials, *Statistics and Probability Letters*, 3**9**, 133-141.

Dudding, B.P. and Jannet, W.J. (1942). *Quality Control Charts*, B.S. 600 R. London: B.S.I.

Ebneshahrashoob, M. and Sobel, M. (1990). Sooner and later waiting time problems for Bernoulli trials: frequency and run quotas, *Statistics and Probability Letters*, **9**, 5-11.

Embrechts. P. and Herzberg, A.M. (1991). Variations of Andrews plots, *International Statistical Review*, **59**, 175-194.

Embrechts, P., Kluppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, New York.

Embrechts, P., Herzberg, A.M., Kalbfleisch, H.K., Traves, W.N. and Whitla, J.R. (1995). An introduction to wavelets with applications to Andrews' plots, *Journal of Computational and Applied Mathematics*, **64**, 41-56.

Everitt, B. (1978). *Graphical Techniques for Multivariate Data*, Heinemann Educational Books, London.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications*, Vol I, Third Edition, John Wiley & Sons, New York.

Freedman, D. (1965). Bernard Friedman's urn, *Annals of Mathematical Statistics*, **36**, 956-970.

Friedman, B.(1949). A simple urn model, *Communications on Pure and Applied Mathematics*, **2**, 59-70.

Fu, J.C. (1996). Distribution theory of runs and patterns associated with a sequence of multi-state trials, *Statistica Sinica*, **6**, 957-974.

Fu, J.C. and Chang, Y.M. (2003). On ordered series and later waiting time distributions in a sequence of Markov dependent multistate trials, *Journal of Applied Probability*, **40** (3), 623-642.

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.

Fu, J.C. and Lou, W.Y.W. (2003). *Distribution Theory of Runs and Patterns and Its Applications*, World Scientific, New Jersey.

Fu, J.C., Lou, W.Y.W., Z.D. Bai and G. Lai (2002). The exact and limiting distributions for the number of successes in success runs within a sequence of Markov-dependent two-state trials, *Annals of the Institute of Statistical Mathematics*, **54**, 719-730.

Fu, J.C., Shmueli, G. and Chang, Y.M. (2003). A unified Markov chain approach for computing the run length distribution in control charts with simple or compound rules, *Statistics and Probability Letters*, **65**, 457-466.

Fu, J.C., Spiring F.A., and Xie, H.S. (2002). On the average run lengths of quality control schemes using a Markov chain approach, *Statistics and Probability Letters*, **56** (4), 369-380.

Gabriel, K.R. and Neumann, J. (1962). A Markov chain model for daily rainfall occurrence at Tel Aviv, *Quarterly Journal of the Royal Meteorological Society*, **88,** 90-95.

Gibbons, J.D. (1971). *Nonparametric Statistical Inference*, McGraw-Hill, New York.

Glasbey, C.A. (1987). Complete linkage as a multiple stopping rule for single linkage clustering, *Journal of Classification*, **4**, 103-109.

Glaz, J. (1983). Moving window detection for discrete-data, *IEEE Transactions on Information Theory*, **29**, 457-462.

Glaz, J. (1989). Approximations and bounds for the distributions of the scan statistic, *Journal of the American Statistical Association*, **84**, 560-566.

Glaz, J., and Naus, J.I. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data, *Annals of Applied Probability*, **1**, 306-318.

Godbole, A.P. (1990a). Specific formulae for some success run distributions, *Statistics and Probability Letters*, **10**, 119-124.

Godbole, A.P. (1990b). On hypergeometric and related distributions of order *k*, *Communications in Statistics-Theory and Methods*, **19**, 1291-1301.

Godbole, A.P. (1991). Poisson approximations for runs and patterns of rare events, *Advances in Applied Probability*, **23**, 851-865.

Godbole, A.P. (1992). The exact and asymptotic distribution of overlapping success runs, *Communications in Statistics-Theory and Methods*, **21**, 953-996.

Godbole, A.P., Papastavridis S.G. and Weishaar, R.S. (1997). Formulae, recursions and approximations for the joint distribution of success runs of several lengths, *Annals of the Institute of Statistical Mathematics*, **49**, 141-153.

Goldstein, L. (1990). Poisson approximation in DNA sequence matching, *Communications in Statistics-Theory and Methods*, **19**, 4167-4179.

Gower, J.C., Ross, G.J.S. (1969). Minimum spanning trees and single linkage cluster analysis, *Journal of the Royal Statistical Society (Series C)*, **18**, 54-64.

Grant, D. (1946). New statistical criteria for learning and problem solution in experiments involving repeated trials, *Psychological Bulletin*, **43**, 272-282.

Grant, D. (1947). Additional tables of the probability of "run" of correct responses in learning and problem solving, *Psychological Bulletin*, **44**, 276-279.

Guibas, L.J. and Odlyzko, A.M. (1978). Maximal Prefix-Synchronized codes, *SIAM Journal of Applied Mathematics*, **35**, 401-418.

Guibas, L.J. and Odlyzko, A.M. (1980). Long repetitive patterns in random sequences, *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, **53**, 241-262.

Guibas, L.J. and Odlyzko, A.M. (1981). Periods in strings, *Journal of Combinatorial Theory*, **30A**, 19-42.

Hahn, G.J. and Gage, J.B. (1983). Evaluation of a start-up demonstration test, *Journal of Quality Technology*, **15**, 103-105.

Han, Q (2001). Joint distributions of success runs in a sequence of bivariate trials, *unpublished manuscript*.

Han. Q. and Aki, S. (1999). Joint distributions of runs in a sequence of multi-state trials, *Annals of the Institute of Statistical Mathematics*, **51**, 419-447.

Han, Q. and Aki, S. (2000). Waiting time problems in a two-state Markov chain, *Annals of the Institute of Statistical Mathematics*, **52**, 778-789.

Hardy, A. (1996). On the number of clusters, *Computational Statistical Data Analysis*, 23, 83-96.

Hawkins, D.M. (1991). Multivariate quality control based on regression-adjusted variables, *Technometrics*, **33**, 61-75.

Hirano, K. (1986). Some properties of the distributions of order *k*, In *Fibonacci Numbers and their Applications* (Eds., A.N. Philippou, A.F. Horadam and G.E. Bergum), pp. 43-53, D. Reidel Publishing Company, Dordrecht, The Netherlands.

Hirano, K. and Aki, S. (1993). On number of occurrences of success runs of specified length in a two-state Markov chain, *Statistica Sinica*, **3**, 313-320.

Hirano, K., Aki, S., Kashiwagi, N., and Kuboki, H. (1991). On Ling's binomial and negative binomial distributions of order *k*, *Statistics and Probability Letters*, **11**, 503-509.

Hotelling, H. (1947). Multivariate quality control, illustrated by the air testing of sample bombsights, *Techniques of Statistical Analysis* (eds., C. Eisenhart, M.W. Hastay, and W.A. Wallis), 111-184, McGraw-Hill, New York.

Huntington, R. (1978). Distribution of the minimum number of points in a scanning interval on the line, *Stochastic Processes and their Applications*, **7**, 73-77.

Jardine, N. and Sibson, R. (1971). *Mathematical Taxonomy*, Wiley, New York.

Johnson, B.C. and Fu, J.C. (1999). The distribution of increasing ℓ-sequences in random permutations: a Markov chain approach, *Statistics and Probability Letters*, **49** (4), 337-344

Johnson, N., Kotz, S. and Balakrishnan, N. (1995) *Continuous Univariate Distributions, Vol. 1 (Wiley Series in Probability and Statistics)* , Wiley, New York.

Karlin, S. and Cardon, L.R. (1994). Computational DNA-sequence analysis, *Annual Review of Microbiology*, **48**, 619-654.

Karlin, S. and Ghandour, G. (1985). Multiple-alphabet amino acid sequence comparisons of the immunoglobulin *k* constant domain, *Proccedings of the National Academy of Science* USA, **82**, 8597-8601.

Karlin, S. and Macken, C (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data, *Journal of the American Statistical Association*, **86**, 27-35.

Karlin, S. and Ost, F. (1987). Counts of long aligned word matches among random letter sequences, *Advances in Applied Probability*, **19**, 293-351.

Khattree, R. and Naik, D.N. (2002). Andrews Plots for Multivariate Data: Some New Suggestions and Applications, *Journal of Statistical Planning and Inference*, **100**, 411-425.

Kemeny, J.G. and Snell, J.L. (1976). *Finite Markov Chains*, Springer Verlag, New York.

Klein, M. (2000). Two alternatives to the Shewhart X-bar Control Chart, *Journal of Quality Technology*, **32**, 427-431.

Knight, W. (1974). A run-like statistic for ecological transects, *Biometrics*, **30**, 553-555.

Koutras, M.V. (1996a). On a Markov chain approach for the study of reliability structures, *Journal of Applied Probability*, **33**, 357-367.

Koutras, M.V (1996b). On a waiting time distribution in a sequence of Bernoulli trials, *Annals of the Institute of Statistical Mathematics*, **48**,789-806.

Koutras, M.V. (1997a). Waiting time distributions associated with runs of fixed length in two-state Markov chains, *Annals of the Institute of Statistical Mathematics*, **49**, 123-139.

Koutras, M.V. (1997b). Waiting times and number of appearances of events in a sequence of discrete random variables, In *Advances in Combinatorial Methods and Applications to Probability and Statistics* (Ed., N. Balakrishnan), 363-384, Birkhauser, Boston.

Koutras, M.V. (2003). Applications of Markov Chains to the Distribution Theory of Runs and Patterns, In *Handbook of Statistics* (Eds., D.N. Shanbhag and C.R. Rao), 431-472, Elsevier Science.

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

Koutras, M.V and Alexandrou, V.A. (1997a). Non-parametric randomness tests based on success runs of fixed length, *Statistics and Probability Letters*, **32**, 393-404.

Koutras, M.V. and Alexandrou, V.A. (1997b). Sooner waiting time problems in a sequence of trinary trials, *Journal of Applied Probability*, **34**, 593-609.

Koutras, M.V., Bersimis, S. and Antzoulakos, D.L. (2005a). Improving the Performance of the Chi-Square Control Chart via Runs Rules, *Methodology and Computing in Applied Probability, Accepted Paper*.

Koutras, M.V., Bersimis, S. and Antzoulakos, D.L. (2005b). Bivariate Markov Chain Embeddable Variables of Polynomial Type, *Submitted Paper*.

Krolak-Schwerdt, S., and Eckes, T. (1992). A graph theoretic criterion for determining the number of cluster in a data set, *Multivariate Behavioral Research*, 27, 4, 541-565.

Laplace, P.S de (1812). *Theorie Analytique des Probabilites*, Paris. Second edition, 1814;Third edition, 1820; Reprinted in *Oeuvres*, **7**, 1886.

Ling, K.D. (1988). On binomial distributions of order *k*, *Statistics and Probability Letters*, **6**, 247-250

Ling, K.D. and Low, T. (1993). On the soonest and latest waiting time distributions: succession quotas, *Communications in Statistics-Theory and Methods*, **22**, 2207-2221.

Ling, K.D. and Tai, T.H. (1990). On bivariate binomial distributions of order *k*, *Soochow Journal of Mathematics*, **16**, 211-220.

Lowry, C.A. and Montgomery, D. C. (1995). A review of multivariate control charts, *IIE Transactions*, **27**, 800-810.

Lowry, A.C., Champ, C.W., and Woodall W.H. (1995). The Performance of Control Charts for Monitoring Process Variation, *Communications in Statistics-Simulation*, **24**, 2, 409-437.

Makri and Philippou (1996). Exact reliability formulas for linear and circular *m*-consecutive-*k*-out-of-*n*: F systems, *Microelectronics and Reliability*, **36**, 657-660.

Maravelakis, P.E., Bersimis, S., Panaretos, J. and Psarakis, S. (2002). Identifying the Out of Control Variable in a Multivariate Control Chart, *Communications in Statistics- Theory and Methods*, **31**, 12, 2391-2408.

Milligan, G.W. (1985). An algorithm for generating artificial test clusters, *Psychometrika*, **50**, 123-127.

Milligan, G.W. and Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, **50**, 2, 159-179.

Montgomery, D.C. (2000). *Introduction to Statistical Quality Control*, New York: John Wiley.

Mood, A.M. (1940). The distribution theory of runs, *Annals of Mathematical Statistics*, **11**, 367-392.

Moore, P.T. (1958). Some properties of runs in quality control procedures, *Biometrika*, **45**, 89-95.

Mosteller, F. (1941). Note on an application of runs to quality control charts, *Annals of Mathematical Statistics*, **12**, 228-232.

Nair, A.N.K. (1942). On the probability of obtaining *k* sets of consecutive successes in *n* trials, *Mathematics Student*, **10**, 83-84.

Naus, J.I. (1982). Approximations for distributions of scan statistics, *Journal of the American Statistical Association*, **77**, 177-183.

Nelson, L.S. (1984). The Shewhart Control Chart-Test for Special Causes, *Journal of Quality Technology*, **16**, 4, 337-239.

Nelson, L.S. (1985). Interpreting Shewharts X-bar Chart, *Journal of Quality Technology*, **17**, 114-116.

O' Brien, P.C. (1976). A test for randomness, *Biometrics*, **32**, 391-401.

O'Brien, P.C. and Dyck, P.J. (1985). A runs test based on run lengths, *Biometrics*, **41**, 237-244.

Page, E.S. (1954). Continuous Inspections Schemes, *Biometrica*, **40**, 112-123.

Page, E.S. (1955). Control Charts with Warning Lines, *Biometrics*, **42**, 243-257.

Palm, A.C. (1990). Tables of Run Length Percentiles for Determining the Sensitivity of Shewhart Control Charts for Averages with Supplementary Runs Rules, *Journal of Quality Technology*, **22**, 289–298.

Panaretos, J. and Xekalaki, E. (1986). On generalized binomial and multinomial distributions and their relation to generalized Poisson distributions, *Annals of the Institute of Statistical Mathematics*, **38**, 223-231.

Philippou, A.N. (1986). Distributions and Fibonacci polynomials of order *k*, longest runs, and reliability of consecutive *k*-out-of-*n*: *F* systems, In *Fibonacci Numbers and Their Applications* (Eds., A.N. Philippou, G.E. Bergum and A.F. Horadam), 203-227, Reidel Dordrecht.

Philippou, A.N. and Antzoulakos, D.L. (1990). Multivariate Fibonacci polynomials of order *k* and the multiparameter negative binomial distribution of the same order, In *Applications of Fibonacci Numbers* (Eds., G.E. Bergum et al.), 273-279, Kluwer Academic Publishers, Amsterdam, The Netherlands.

Philippou, A.N. and Makri, F.S. (1986). Successes runs and longest runs, *Statistics and Probability Letters*, **4**, 211-215.

Philippou, A.N. and Muwafi, A.A. (1982). Waiting for the *k-th* consecutive success and the Fibonacci sequence of order *k*, *The Fibonacci quarterly*, **20**, 28-32.

Philippou, A.N., Georgiou, C. and Philippou, G.N. (1983). A generalized geometric distribution and some of its properties, *Statistics and Probability Letters*, **1**, 171-175.

Philippou, A.N., Antzoulakos, D.L. and Tripsiannis, G.A. (1990). Multivariate Distributions of order *k*, Part II, *Statistics and Probability Letters*, **32**, 393-404.

Philippou, A.N., Tripsiannis, G.A. and Antzoulakos, D.L. (1989). New Polya and inverse Polya distributions of order *k*, *Communications in Statistics-Theory and Methods*, **18**, 2125-2137.

Pielou, E.C. (1962). Runs of one species with respect to another in transects through plant populations, *Biometric*, **18**, 579-593.

Pielou, E.C. (1963a). Runs of healthy and diseased trees in transects through an infected forest, *Biometrics*, **19**, 603-614.

Pielou, E.C. (1963b). The distribution of the diseased trees with respect to healthy ones in a patchily infected forest, *Biometrics*, **19**, 450-459.

Pielou, E.C. (1969). *An Introduction to Mathematical Ecology*, Wiley-Interscience, New York.

Pielou, E.C. (1977). *Mathematical Ecology* ,John Wiley & Son, New York.

Pignatiello, J.J., Jr. and Runger, G.C. (1990). Comparisons of multivariate CUSUM charts, *Journal of Quality Technology*, 22, 173-186.

Prairie, R.R., Zimmer, W.J. and Brookhouse, J.K. (1962). Some acceptance sampling plans based on the theory of runs, *Technometrics*, **4**, 177-185.

Rakitzis A. (2004). Shewhart Control Charts with Stopping Rules Based on Runs. Unpublished MSc Thesis, University of Piraeus, Department of Statistics and Insurance Science, Greece.

Roberts, S.W. (1958). Properties of control chart zone tests, *Bell System Technical Journal*, **37**, 83-114.

Ryan, T.P. (2000). *Statistical Methods for Quality Improvement (2nd ed.)*, New York: John Wiley.

SAS Institute (1986). SAS / QC User's Guide, Version 5 Edition, SAS Institute, Cary, NC.Shewhart, W. A. (1931).

Schuster, E.F. (1994). Exchangeability and recursion in the conditional distribution theory of number and length of runs, In *Runs and Patterns in Probability*, (Eds., A.P. Godbole and S.P. Papastavridis), 91-118, Kluwer Academic Publishers, Amsterdam, The Netherlands.

Sen, Z. (1980). Statistical analysis of hydrologic critical draughts, *Journal of the Hydraulics Division*, **106**, 99-115.

Shmueli, G. and Cohen, A. (2000). Run-related probability functions applied to sampling inspection, *Technometrics*, **42**, 2, 188-202.

Shmueli G., and Cohen A. (2003). Run-Length Distribution for Control Charts with Runs and Scans Rules, *Communication in Statistics, Part A - Theory and Methods*, 32, 475-495.

Simpson, T. (1740). *The nature and Laws of Chance. The Whole after a new, general, and conspicuous Manner, and illustrated with a great Variety of Examples*, Cave Publishers, London. Reprinted, 1792.

Todhunter, I. (1865). *A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace*, Macmillan, London. Reprintend by Chelsea Publishing Company, New York, 1949.

Tripsiannis, G.A. (1993). Modified multivariate Polya and inverse Polya distributions of order *k*, *Journal of the Indian Society of Statistics and Operations Research*, **14**, 1-14.

Tripsiannis, G.A. and Philippou, A.N. (1997a). A multivariate negative binomial distribution of order *k* arising when success runs are allowed to overlap, In *Advances in Combinatorial Methods and Applications to Probability and Statistics* (Ed., N. Balakrishnan), pp. 427-438, Birkhauser, Boston.

Tripsiannis, G.A. and Philippou, A.N. (1997b). A new multivariate inverse Polya distribution of order *k*, *Communications in Statistics – Theory and Methods*, **26**, 149-158.

Uchida, M. and Aki, S. (1995). Sooner and later waiting time problems in a two-state Markov chain, *Annals of the Institute of Statistical Mathematics*, **47**, 415-433.

Uppuluri, V.R.R. and Patil, S.A. (1983). Waiting times and generalized Fibonacci sequences, *The Fibonacci Quarterly*, **21**, 242-249.

Vassiliou A., Tambouratzis, D., Koutras, M.V. and Bersimis, S. (2004). A New Similarity Measure and its Use in Determining the Number of Clusters in a Data Set, *Communications in Statistics – Theory and Methods*, **33**, 7, 1643-1666.

Viveros R. and Balakrishnan, N. (1993). Statistical inference from start-up demonstration test data, *Journal of Quality Technology*, **22**, 119-130.

Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population, *Annals of Mathematical Statistics*, **11**, 147-162.

Walker, E., Philpot, J.W. and Clement, J. (1991). False signal rates for the Shewhart control chart with supplementary runs tests, *Journal of Quality Technology*, **23**, 247-252.

Wallenstein, S., Naus, J.I. and Glaz, J. (1994). Power of the scan statistic in detecting a changed segment in a Bernoulli sequence, *Biometrika*, **81**, 595-601.

Wang (2001), Convergence theorems for the lengths of consecutive successes of Markov-Bernoulli sequences, *unpublished manuscript*.

Weiler, H. (1953). The use of runs to control the mean in quality control, *Journal of the American Statistical Association*, **48**, 816-825.

Western Electric Company AT&T (1956). *Statistical Quality Control Handbook*, Indianapolis, IN.

Westgard, J.A. and Groth, T. (1977). Power Functions for Statistical Control Rules, *Clinical Chemistry*, **25**, 863-869.

Westgard, J.A., Barr, P.L., Hunt, M.R. and Groth, T. (1979). A Multi-Run Shewhart Chart for Quality Control in Clinical Chemistry, *Clinical Chemistry*, **27**, 493-501.

Wolfowitz, J. (1943). On the theory of runs with some applications to quality control, *Annals of Mathematical Statistics*, **14**, 280-288.

Woodall,W.H. (2000). Controversies and Contradictions in Statistical Process Control, *Journal of Quality Technology*, **32**, 341-378.