ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

Τμημα Μαθηματικων

Φωτεινή Θ. Ζόγκα

Ελεγχοι καλης προσαρμογης για την κατανομη Poisson

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

Ιωάννινα, 2023

UNIVERSITY OF IOANNINA

Department of Mathematics

Foteini Th. Zogka

Goodness-of-fit tests for the Poisson distribution

Master's Thesis

Ioannina, 2023

Αφιερώνεται στην οικογένειά μου.

Η παρούσα Μεταπτυχιακή Διατριβή εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική και Επιχειρησιακή Έρευνα που απονέμει το Τμήμα Μαθηματικών του Πανεπιστημίου Ιωαννίνων.

Εγκρίθηκε την 27/01/2023 από την Τριμελή Εξεταστική Επιτροπή:

Ονοματεπώνυμο	**Βαθμίδα**
Ιωάννης Δημητρίου	Αν. Καθηγητής
Δημήτριος Μπάγκαβος	Επ. Καθηγητής
Απόστολος Μπατσίδης	Αν. Καθηγητής (**Επιβλέπων**)

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

"Δηλώνω υπεύθυνα ότι η παρούσα διατριβή εκπονήθηκε κάτω από τους διεθνείς ηθικούς και ακαδημαϊκούς κανόνες δεοντολογίας και προστασίας της πνευματικής ιδιοκτησίας. Σύμφωνα με τους κανόνες αυτούς, δεν έχω προβεί σε ιδιοποίηση ξένου επιστημονικού έργου και έχω πλήρως αναφέρει τις πηγές που χρησιμοποίησα στην εργασία αυτή."

Φωτεινή Θ. Ζόγκα

Ευχαριστιες

Η παρούσα Μεταπτυχιακή Διατριβή εκπονήθηκε στον Τομέα Πιθανοτήτων, Στατιστικής και Επιχειρησιακής Έρευνας του Τμήματος Μαθηματικών της Σχολής Θετικών Επιστημών του Πανεπιστημίου Ιωαννίνων, με Επιβλέποντα τον Αναπληρωτή Καθηγητή κ. Απόστολο Μπατσίδη και μέλη της Τριμελούς Εξεταστικής Επιτροπής τους κ.κ. Ιωάννη Δημητρίου και Δημήτριο Μπάγκαβο.

Αρχικά θα ήθελα να ευχαριστήσω τον κ. Απόστολο Μπατσίδη που δέχτηκε να αναλάβει την επίβλεψη αυτής της μεταπτυχιακής διατριβής. Τον ευχαριστώ για την καθοδήγησή του, την υπομονή και τον πολύτιμο χρόνο που μου αφιέρωσε για να βγει εις πέρας αυτή η μεταπτυχιακή διατριβή.

Ακόμη θα ήθελα να ευχαριστήσω τους κκ. Ιωάννη Δημητρίου, Αναπληρωτή Καθηγητή του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων και Δημήτριο Μπάγκαβο, Επίκουρο Καθηγητή του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων, για τη συμμετοχή τους στην Τριμελή Εξεταστική Επιτροπή και για τις παρατηρήσεις και υποδείξεις τους. Ευχαριστώ και όλα τα άλλα μέλη του Τομέα Πιθανοτήτων, Στατιστικής και Επιχειρησιακής Έρευνας, για τις γνώσεις που μου παρείχαν τόσο σε προπτυχιακό, όσο και σε μεταπτυχιακό πρόγραμμα σπουδών.

Επίσης ευχαριστώ τους συμφοιτητές μου, που ήταν συνοδοιπόροι σε αυτό το μεταπτυχιακό πρόγραμμα, αλλά και τους φίλους μου, που πίστεψαν σε μένα.

Τέλος από καρδιάς ευχαριστώ τους γονείς μου Θεόδωρο και Αθηνά και τον αδερφό μου Δημήτρη, για την ψυχολογική και συναισθηματική τους στήριξη καθόλη την πορεία των προπτυχιακών και μεταπτυχιακών μου σπουδών.

Ιωάννινα, Ιανουάριος 2023

Φωτεινή Θ. Ζόγκα

Περιληψη

Η εξέχουσα θέση της κατανομής Poisson, μεταξύ των διακριτών κατανομών, τόσο από θεωρητικής όσο και από πρακτικής πλευράς, έχει ως συνέπεια την αναγκαιότητα ύπαρξης στατιστικών μεθοδολογιών ελέγχου ότι όντως ένα διαθέσιμο σύνολο δεδομένων προέρχεται από έναν πληθυσμό που περιγράφεται ικανοποιητικά από μία κατανομή Poisson. Έλεγχοι αυτής της μορφής καλούνται «έλεγχοι καλής προσαρμογής» των δεδομένων σε μια συγκεκριμένη κατανομή και έχουν προταθεί αρκετοί στο πλαίσιο της κατανομής Poisson. Η παρουσίαση και η συγκριτική μελέτη των ελέγχων καλής προσαρμογής της κατανομής Poisson, αποτελεί το βασικό αντικείμενο μελέτης αυτής της μεταπτυχιακής διατριβής. Ειδικότερα, στο Κεφάλαιο 1 (Εισαγωγή) δίνεται μία σύντομη ιστορική αναδρομή στον τρόπο εισαγωγής αυτής της κατανομής στη βιβλιογραφία, ενώ το κεφάλαιο ολοκληρώνεται με μία ενότητα ο στόχος της οποίας είναι διτός. Από τη μία μεριά να γίνει υπενθύμιση βασικών εννοιών και αποτελεσμάτων, και από την άλλη μεριά να παρουσιαστούν χρήσιμες ιδιότητες της κατανομής Poisson, που θα χρησιμοποιηθούν στο υπόλοιπο τμήμα αυτής της μεταπτυχιακής διατριβής. Στο Κεφάλαιο 2 (Ανασκόπηση) παρουσιάζονται οι πιο συχνά χρησιμοποιούμενοι σε πρακτικές εφαρμογές έλεγχοι καλής προσαρμογής που έχουν παρουσιαστεί στη σχετική βιβλιογραφία. Ταυτόχρονα, γίνεται προσπάθεια οι προαναφερθέντες έλεγχοι να

ταξινομηθούν ανάλογα με τα χαρακτηριστικά τους. Στο Κεφάλαιο 3 (Συγκριτική μελέτη), το ενδιαφέρον επικεντρώνεται στη συγκριτική μελέτη της απόδοσης των ελέγχων που παρουσιάστηκαν στο προηγούμενο κεφάλαιο, τόσο ως προς τη διατήρηση του επιπέδου σημαντικότητας, όσο και ως προς την ισχύ. Ειδικότερα, συνοπτικά αναφέρονται τα διαθέσιμα στη βιβλιογραφία συμπεράσματα, ενώ παρατίθενται και τα αποτελέσματα της συγκριτικής μελέτης που πραγματοποιήθηκε στα πλαίσια αυτής της μεταπτυχιακής διατριβής. Στο Κεφάλαιο 4 (Επεκτάσεις - Επίλογος) αναφέρονται συνοπτικά έλεγχοι καλής προσαρμογής για τη διδιάστατη και την m-διάστατη, με $m \geq 3$, κατανομή Poisson. Στον επίλογο του κεφαλαίου και της διατριβής αναδεικνύεται η αναγκαιότητα ύπαρξης εναλλακτικών στην Poisson κατανομών, καθώς και η επακόλουθη συνέπεια της ύπαρξης ελέγχων καλής προσαρμογής για αυτές τις εναλλακτικές κατανομές, που αποτελούν μελλοντικό πιθανό πεδίο έρευνας,δηλαδή κατανομών που γενικεύουν την κατανομή Poisson. Τέλος, η μεταπτυχιακή διατριβή ολοκληρώνεται με τη Βιβλιογραφία.

Abstract

The importance of the Poisson distribution among the discrete distributions has led to the development of several hypothesis tests, for testing whether the available count data set can be considered as it comes from the population which can be described by a Poisson distribution. Tests of this form are well-known as goodness-of-fit tests for the Poisson distribution and the main aim of this Master's thesis is the presentation and the comparative study of these goodness-of-fit tests. In particular, Chapter 1 (Introduction) gives the historical background of the genesis of the Poisson distribution and some of its useful properties. Moreover, a review of fundamental concepts and results, which will be used in the rest of this thesis, is also provided. Chapter 2 (Review) presents the most widely used goodness-of-fit tests that have been presented in the statistical literature. Simultaneously, these tests are classified into categories according to their characteristics. Chapter 3 (Comparative study), the interest is focused on comparing and evaluating the performance of the tests presented in the previous chapter based on Monte Carlo simulation studies. Since the prominent role of the Poisson distribution resulted in its extension to two or more dimensions, in Chapter 4 (Extensions- Epilogue) we briefly describe some goodness-of-fit tests for the bivariate and multivariate Poisson distributions.

In the last section of this chapter we discuss the weakness of Poisson distribution to model several random phenomena, which resulted in generalizations of the Poisson distribution and the necessity of goodness-of-fit techniques for the new models. Finally, the thesis is completed with the Bibliography.

Περιεχομενα

1 Εισαγωγή **5**

1.1 Ιστορική αναδρομή και ορισμός της κατανομής Poisson 5

1.2 Χρήσιμα αποτελέσματα και ιδιότητες της κατανομής Poisson . . . 11

1.2.1 Ορισμοί και αποτελέσματα ασυμπτωτικής στατιστικής . . . 11

1.2.2 Αθροιστική συνάρτηση κατανομής και εκτίμησή της 13

1.2.3 Ολοκληρώσιμη συνάρτηση κατανομής και εκτίμησή της . . 16

1.2.4 Πιθανογεννήτρια συνάρτηση και εκτίμησή της 17

1.2.5 Ροπογεννήτρια, ροπές, ημιαναλλοίωτες και άλλες ιδιότητες 23

1.2.6 Μέση απόσταση 29

1.2.7 Παραμετρικό bootstrap στον έλεγχο καλής προσαρμογής 32

2 Έλεγχοι καλής προσαρμογής της κατανομής Poisson **35**

2.1 Χι-τετράγωνο έλεγχος 36

2.1.1 Τροποποιήσεις και γενικεύσεις 41

2.2 Έλεγχοι με την ε.α.σ.κ. 45

2.2.1 Kolmogorov-Smirnov 46

2.2.2 Cramér-von Mises, Watson, Anderson-Darling και Klar 49

2.3 Έλεγχος με την ε.ο.σ.κ. 56

2.4 Έλεγχοι με την ε.π.σ. 58

2.5 Έλεγχοι που στηρίζονται στις ροπές 69

2.5.1 Έλεγχοι με τον δειγματικό δείκτη διασποράς 69

2.5.2 Έλεγχος των Gupta et al. (1994) 80

2.5.3 Έλεγχος των Kyriakoussis et al. (1998) 82

2.5.4 Έλεγχος των Pettigrew and Mohler (1967) 85

2.6 Έλεγχοι βασισμένοι στην επαρκή στατιστική συνάρτηση 88

2.7 Smooth tests 92

2.8 Γραφικοί έλεγχοι 97

3 Συγκριτική μελέτη 103

3.1 Ανασκόπηση 104

3.2 Συγκριτική μελέτη-συμπεράσματα 111

3.2.1 Σχεδιασμός μελέτης προσομοίωσης 112

3.2.2 Αποτελέσματα συγκριτικής μελέτης 114

3.2.3 Συμπεράσματα συγκριτικής μελέτης 121

4 Επεκτάσεις - Επίλογος 125

4.1 Εισαγωγή 126

4.2 Έλεγχοι για τη διδιάστατη κατανομή Poisson 131

4.2.1 Γενικεύσεις των δεικτών διασποράς του Fisher 131

4.2.2 Έλεγχοι με την ε.π.σ. 134

4.3 Έλεγχοι για την πολυδιάστατη κατανομή Poisson 140

4.4 Επίλογος 142

Βιβλιογραφία **145**

ΚΕΦΑΛΑΙΟ 1

Εισαγωγη

Στο πρώτο κεφάλαιο, αρχικά, θα δοθεί μία σύντομη ιστορική αναδρομή στην κατανομή Poisson, ενώ στη συνέχεια ο σκοπός του κεφαλαίου είναι διτός. Από τη μια μεριά θα παρουσιαστούν χρήσιμες ιδιότητες της κατανομής Poisson και από την άλλη μεριά θα γίνει υπενθύμιση βασικών εννοιών και αποτελεσμάτων, τα οποία είναι ιδιαίτερα χρήσιμα για την κατανόηση όσων έπονται στο επόμενο κεφάλαιο και αφορούν τους ελέγχους καλής προσαρμογής της κατανομής Poisson.

1.1 Ιστορική αναδρομή και ορισμός της κατανομής Poisson

Η κατανομή Poisson είναι, ίσως, η σημαντικότερη συνήθης διακριτή κατανομή με πλήθος εφαρμογών, όπως αναλυτικά θα αναφέρουμε στη συνέχεια. Ονομάζεται έτσι προς τιμήν του Simeon Denis Poisson (1781 -1840), ο οποίος παρουσίασε την κατανομή αυτή σε εργασία του το 1837 (βλέπε Poisson (1837)). Ειδικότερα, ο Poisson αρχικά διαπίστωσε ότι η χρήση της συνάρτησης πιθανότητας της διωνυμικής κατανομής παρουσιάζει αρκετές δυσκολίες για τον υπολογισμό πιθανοτήτων, όταν το πλήθος, n, των ανεξάρτητων επαναλήψεων μιας δοκιμής Bernoulli παίρνει μεγάλες τιμές, η πιθανότητα επιτυχίας είναι αρκετά μικρή ($p \longrightarrow 0$) και η τιμή της τυχαίας μεταβλητής για την οποία θέλουμε να υπολογιστεί η πιθανότητα δεν είναι κοντά είτε στο 0 είτε στο n. Σε όσα ακολουθούν με e συμβολίζεται ο αριθμός του Euler ($e = 2.71828.....$) και με ! το σύμβολο του παραγοντικού.Στη συνέχεια, θέλοντας να προσδιορίσει έναν εναλλακτικό τρόπο υπολογισμού των πιθανοτήτων της διωνυμικής κατανομής, οδηγήθηκε, ως όριο της διωνυμικής κατανομής και υπό τις υποθέσεις της πρότασης που ακολουθεί, στην κατανομή που σήμερα μας είναι γνωστή ως κατανομή Poisson.

Πρόταση 1.1.1. *Έστω ότι η τ.μ. X ακολουθεί τη διωνυμική κατανομή με παραμέτρους n, $p \in (0,1)$, δηλαδή $X \sim Bin(n,p)$. Αν για $n \to +\infty$, η πιθανότητα επιτυχίας p συγκλίνει στο 0, $p \to 0$, έτσι ώστε η αναμενόμενη τιμή της τ.μ. να συγκλίνει σε έναν σταθερό αριθμό $\lambda > 0$, δηλαδή έτσι ώστε $E(X) = np \to \lambda$, με $\lambda > 0$, τότε:*

$$\lim_{n\to+\infty} p_X(x) = \lim_{n\to+\infty} \left(\binom{n}{x} p^x \ (1-p)^{n-x} \right) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots .$$

Εύκολα αποδεικνύεται ότι η συνάρτηση που προέκυψε στην παραπάνω πρόταση ως όριο της διωνυμικής κατανομής είναι όντως συνάρτηση πιθανότητας, μιας και είναι μη αρνητική συνάρτηση και επιπρόσθετα $\sum_{x=0}^{\infty} e^{-\lambda} \dfrac{\lambda^x}{x!} = 1$, καθώς ως γνωστόν $\sum_{x=0}^{\infty} \dfrac{\lambda^x}{x!} = e^{\lambda}$. Επομένως,η παραπάνω πρόταση, οδήγησε στον παρακάτω ορισμό.

Ορισμός 1.1.1. *Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την **κατανομή Poisson** με παράμετρο λ, $\lambda > 0$, αν οι δυνατές της τιμές x είναι $x \in \{0, 1, 2, \ldots, \}$ και η συνάρτηση πιθανότητάς της δίνεται από τη σχέση:*

$$f_0(x,\lambda) = p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots \tag{1.1}$$

Στην περίπτωση αυτή θα συμβολίζεται $X \sim \mathcal{P}(\lambda)$.

Στο σημείο αυτό επισημαίνεται ότι ο ρόλος της παραμέτρου λ θα διευκρινιστεί στην επόμενη ενότητα, όπου παρατίθενται χρήσιμες ιδιότητες της κατανομής.

Από όσα προηγήθηκαν προκύπτει ότι η κατανομή Poisson μπορεί να χρησιμοποιηθεί προσεγγιστικά, για τη μοντελοποίηση του αριθμού των επιτυχιών σε ένα πάρα πολύ μεγάλο πλήθος από ανεξάρτητες επαναλήψεις μίας δοκιμής Bernoulli με σταθερή και πάρα πολύ μικρή πιθανότητα επιτυχίας σε κάθε δοκιμή. Για τον λόγο αυτόν αναφέρεται και ως **κατανομή των σπάνιων ενδεχομένων** (distribution of rare events). Ωστόσο, η χρήση της δεν περιορίζεται μόνο σε αυτές τις περιπτώσεις, καθώς, υπό κάποιους περιορισμούς-υποθέσεις που σχετίζονται με αυτές της πρότασης 1.1.1, αποδεικνύεται ότι η κατανομή αυτή μπορεί να χρησιμοποιηθεί για τη μοντελοποίηση τυχαίων πειραμάτων που εξελίσσονται στον χρόνο και το ενδιαφέρον μας επικεντρώνεται στον αριθμό των εμφανίσεων ενός ενδεχομένου σε αυτό το χρονικό διάστημα. Ενδεικτικά παραπέμπουμε, μεταξύ άλλων, στους Feller (1968) και Ζωγράφος (2008). Ειδικότερα, η κατανομή Poisson βρίσκει εφαρμογή και σε περιπτώσεις όπου σε ένα τυχαίο πείραμα μας ενδιαφέρει πόσες φορές πραγματοποιείται ένα ενδεχόμενο σε ένα χρονικό διάστημα t ή σε ένα διάστημα μήκους t ή σε μία επιφάνεια εμβαδού t. Σε τέτοιες περιπτώσεις δεν έχουμε πλέον μία τ.μ. αλλά σύμφωνα με τον επόμενο ορισμό μία στοχαστική διαδικασία.

Ορισμός 1.1.2. *Μία* ***στοχαστική διαδικασία*** *είναι μία συλλογή, μία οικογένεια τυχαίων μεταβλητών* $\{X(t) : t \in T\}$, *όπου* t *είναι μία παράμετρος που παίρνει τιμές σε ένα κατάλληλα ορισμένο σύνολο* T, *που καλείται παραμετρικός χώρος. Δηλαδή, για κάθε* t *η* $X(t)$ *είναι τ.μ. Η ειδική περίπτωση της στοχαστικής διαδικασίας που παριστάνει τον συνολικό αριθμό των συμβάντων που* *έχουν πραγματοποιηθεί στον χρόνο t, με $X(t) \geq 0$ και σύνολο δυνατών τιμών $\{0, 1, 2, \ldots\}$, καλείται **διαδικασία καταμέτρησης ή απαριθμήτρια διαδικασία** όταν επιπλέον ισχύει:*

α) $X(0) = 0$

β) $X(t) \geq X(s)$ *για* $s < t$

γ) Το πλήθος των συμβάντων στο $(s, t]$ *είναι* $N(t) - N(s)$.

Μία διαδικασία καταμέτρησης λέμε ότι είναι **διαδικασία Poisson** με μέσο ρυθμό λ στη μονάδα του χρόνου (όγκου, μήκους, ανάλογα) αν $X(0) = 0$ και επιπλέον πληροί τις ακόλουθες υποθέσεις:

1. Υπόθεση 1 (*Ιδιότητα Στατικότητας*). Η πιθανότητα πραγματοποίησης ακριβώς ενός γεγονότος σε ένα μικρό χρονικό διάστημα μήκους dt είναι κατά προσέγγιση ανάλογη του μήκους του διαστήματος, δηλαδή ισχύει ότι:

$$P(X(dt) = 1) = \lambda \cdot dt + o(dt),$$

   όπου ο συμβολισμός $o(dt)$ χρησιμοποιείται για να δηλώσει μια συνάρτηση που είναι τέτοια ώστε $\lim\limits_{dt \to 0} \frac{o(dt)}{dt} = 0$.

2. Υπόθεση 2. Η πιθανότητα να εμφανιστεί το γεγονός δύο ή περισσότερες φορές σε ένα μικρό χρονικό διάστημα dt είναι αμελητέα, δηλαδή ισχύει ότι:

$$P(X(dt) \geq 2) = o(dt).$$

3. Υπόθεση 3 (*Ιδιότητα Ανεξαρτησίας*). Ο αριθμός των γεγονότων που εμφανίζονται σε ένα χρονικό διάστημα είναι ανεξάρτητος από τον αριθμό των

γεγονότων που πραγματοποιούνται σε ένα οποιοδήποτε άλλο μη επικαλυπτόμενο χρονικό διάστημα ή διαφορετικά οι αριθμοί των συμβάντων που λαμβάνουν χώρα σε μη επικαλυπτόμενα χρονικά διαστήματα είναι ανεξάρτητοι μεταξύ τους.

Οι τρεις παραπάνω υποθέσεις ουσιαστικά οδηγούν στην ακόλουθη ιδιότητα: για οποιοδήποτε χρονικό διάστημα $(s, s+t]$, με $s \geq 0$ και $t > 0$, ο αριθμός των γεγονότων σε αυτό, $X(s+t) - X(s)$, ακολουθεί την κατανομή Poisson με παράμετρο λt, δηλαδή ισχύει ότι:

$$P\left(X(s+t) - X(s) = x\right) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, \quad x = 0, 1, 2, \ldots$$

Πρόταση 1.1.2. *Αν $X(t)$ είναι διαδικασία Poisson που περιγράφει τον αριθμό των συμβάντων σε ένα χρονικό διάστημα t (ή αντίστοιχα σε μία επιφάνεια εμβαδού t ή σε μία απόσταση μήκους t), τότε*

$$P\left(X(t) = x\right) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, \quad x = 0, 1, 2, \ldots \tag{1.2}$$

όπου η παράμετρος λ εκφράζει τον μέσο αριθμό των συμβάντων που πραγματοποιούνται στη μονάδα του χρόνου (ή μήκους ή επιφάνειας) ή αλλιώς τον ρυθμό εμφάνισης.

Οι παραπάνω ιδιότητες έχουν κάνει την κατανομή και τη διαδικασία Poisson να εφαρμόζονται σε πλήθος επιστημονικών πεδίων. Ενδεικτικά παραδείγματα όπου η κατανομή αυτή βρίσκει εφαρμογή αποτελούν (βλέπε Οικονόμου κ.ά. (2022) και τις εκεί αναφορές) η μοντελοποίηση του αριθμού:

- των εναέριων βομβών που έπληξαν το Λονδίνο κατά τη διάρκεια του Δεύτερου Παγκοσμίου Πολέμου,

- των άστρων σε συγκεκριμένο τμήμα του διαστήματος,
- των ασθενών που καταφθάνουν σε ένα νοσοκομείο κατά τη διάρκεια μιας ώρας,
- των φωτονίων λέιζερ που χτυπούν έναν ανιχνευτή σε ένα χρονικό διάστημα,
- των τηλεφωνικών κλήσεων ή των αφίξεων πελατών σε ένα σύστημα εξυπηρέτησης, σε ένα χρονικό διάστημα,
- των μεταλλάξεων σε ένα τμήμα DNA ανά μονάδα μήκους,
- των βακτηρίων σε πλάκα άγαρ συγκεκριμένης επιφάνειας, και
- των τερμάτων σε έναν ποδοσφαιρικό αγώνα.

Προφανώς, ο παραπάνω κατάλογος είναι ανεξάντλητος, ενώ δεν πρέπει να ξεχνάμε και τις εφαρμογές της κατανομής Poisson ως όριο της διωνυμικής κατανομής.

Η σπουδαιότητα της κατανομής Poisson στη μοντελοποίηση πραγματικών τυχαίων φαινομένων, έχει ως συνέπεια την αναγκαιότητα ύπαρξης τρόπων ελέγχου ότι όντως ένα διαθέσιμο σύνολο δεδομένων προέρχεται από έναν πληθυσμό που περιγράφεται ικανοποιητικά από μία κατανομή Poisson. Έλεγχοι αυτής της μορφής καλούνται «έλεγχοι καλής προσαρμογής» των δεδομένων σε μια συγκεκριμένη κατανομή και έχουν προταθεί αρκετοί στο πλαίσιο της κατανομής Poisson. Αντικείμενο μελέτης των επόμενων δύο κεφαλαίων είναι η παρουσίαση και η συγκριτική μελέτη των κυριότερων τέτοιων ελέγχων που έχουν εμφανιστεί στη βιβλιογραφία. Για την καλύτερη κατανόηση όσων θα ακολουθήσουν, στην ενότητα που ακολουθεί υπενθυμίζονται βασικές έννοιες και αποτελέσματα, ενώ παρατίθενται βασικές ιδιότητες της κατανομής Poisson.

1.2 Χρήσιμα αποτελέσματα και ιδιότητες της κατανομής Poisson

Στην ενότητα αυτή θα παρατεθούν χρήσιμα αποτελέσματα που αφορούν τόσο έννοιες που θα χρησιμοποιηθούν στο επόμενο κεφάλαιο της διατριβής όσο και ιδιότητες της κατανομής Poisson. Η παράθεση αυτών γίνεται με απώτερο στόχο τη διευκόλυνση της μελέτης των υπόλοιπων κεφαλαίων αυτής της διατριβής.

1.2.1 Ορισμοί και αποτελέσματα ασυμπτωτικής στατιστικής

Στην ενότητα αυτή θα παρατεθούν ορισμοί των ειδών σύγκλισης ακολουθιών τυχαίων μεταβλητών, καθώς και θεμελιώδη θεωρήματα της ασυμπτωτικής στατιστικής που θα χρησιμοποιηθούν εκτενώς στο δεύτερο κεφάλαιο της μεταπτυχιακής διατριβής. Για περισσότερες λεπτομέρειες παραπέμπουμε στο σύγγραμμα του Van der Vaart (2000).

Ορισμός 1.2.1. *Η ακολουθία των τυχαίων μεταβλητών* X_n, $n = 1, 2, \ldots$, *λέμε ότι συγκλίνει κατά πιθανότητα (converge in probability) στην τυχαία μεταβλητή* X *και γράφουμε ότι* $X_n \xrightarrow{\pi} X$, *αν για κάθε* $\varepsilon > 0$

$$\lim_{n\to\infty} P\{|X_n - X| > \varepsilon\} = 0.$$

Ορισμός 1.2.2. *Έστω* X_n, $n = 1, 2, \ldots$ *μια ακολουθία τ.μ. και* X *μία άλλη τ.μ., με αθροιστικές συναρτήσεις κατανομής* $F_n(\cdot)$ *και* $F(\cdot)$, *αντίστοιχα. Η ακολουθία των τυχαίων μεταβλητών* X_n, $n = 1, 2, \ldots$, *λέμε ότι συγκλίνει κατά κατανομή (converge in distribution) στην τυχαία μεταβλητή* X *και γράφουμε ότι*

$X_n \xrightarrow{d} X$, *αν*

$$\lim_{n\to\infty} F_n(x) = F(x)$$

για κάθε $x \in \mathbb{R}$, *σημείο συνέχειας της* $F(x)$.

Ορισμός 1.2.3. *Η ακολουθία των τυχαίων μεταβλητών* X_n, $n = 1, 2, \ldots$, *λέμε ότι συγκλίνει σχεδόν βέβαια (converge almost surely) στην τυχαία μεταβλητή* X *και γράφουμε ότι* $X_n \xrightarrow{\sigma.\beta.} X$ *αν*

$$P(\{\omega \in \Omega : X_n(\omega) \to X(\omega)\}) = 1.$$

Θεώρημα 1.2.1. ***(Slutsky)*** *Έστω* $X_n \xrightarrow{d} X$ *και* $Y_n \xrightarrow{\pi} c$, *όπου* $c \in \mathbb{R}$ *μία σταθερά. Τότε,* $X_n + Y_n \xrightarrow{d} X + c$, $X_n Y_n \xrightarrow{d} cX$ *και* $\dfrac{X_n}{Y_n} \xrightarrow{d} \dfrac{X}{c}$, *εφόσον* $c \neq 0$ *(βλέπε Van der Vaart (2000)).*

Κεντρικό Οριακό Θεώρημα

Στη Θεωρία Πιθανοτήτων το Κεντρικό Οριακό Θεώρημα θεμελιώνει ότι, σε πολλές περιπτώσεις, το άθροισμα πολλών ανεξάρτητων τυχαίων μεταβλητών τείνει να ακολουθεί κατάλληλα ορισμένη κανονική κατανομή ακόμα κι αν οι αρχικές μεταβλητές δεν ακολουθούν κανονική κατανομή. Στη συνέχεια διατυπώνεται το Κεντρικό Οριακό Θεώρημα στην πολυδιάστατη περίπτωση.
Σε όσα ακολουθούν με a^t συμβολίζεται το αντίστροφο του διανύσματος a.

Θεώρημα 1.2.2. ***(Πολυδιάστατο Κεντρικό Οριακό Θεώρημα)*** *Έστω ανεξάρτητα και ισόνομα τυχαία διανύσματα* $X^j = (X_1^j, ..., X_k^j)^t$, $j = 1, 2, ..., n$, *από έναν πληθυσμό με μέσο διάνυσμα* $\mu = (\mu_1, \mu_2, ..., \mu_k)^t$ *και πίνακα διακυμάνσεων συνδιακυμάνσεων* $\Sigma = [\sigma_{ij}] \in \mathbb{R}^{k\times k}$, *όπου* $E(X_i) = \mu_i$ *και* $Cov(X_i, X_j) = \sigma_{ij}$. *Τότε ισχύει ότι:*

$$\sqrt{n}\left(\bar{X} - \mu\right) \xrightarrow{d} N_k(0, \Sigma),$$

όπου $\bar{X} = \frac{1}{n}\left(\sum_{j=1}^{n} X_1^j, ..., \sum_{j=1}^{n} X_k^j\right)^t$.

Δέλτα μέθοδος

Στη Στατιστική η δέλτα μέθοδος αφορά την προσεγγιστική κατανομή μίας συνάρτησης ενός ασυμπτωτικά κανονικού εκτιμητή. Ειδικότερα, αν T_n είναι ένας εκτιμητής της παραμέτρου θ με $\sqrt{n}(T_n - \theta) \xrightarrow{d} T$ και $\phi : R^k \to R^m$ μία δοθείσα συνάρτηση ορισμένη τουλάχιστον σε μία περιοχή του θ, με ϕ να είναι διαφορίσιμη στο θ, τότε $\sqrt{n}\left(\phi(T_n) - \theta\right) \xrightarrow{d} \phi_\theta' T$, όπου $\phi(x_1, ..., x_k) = (\phi_1(x_1, ..., x_k), ..., \phi_m(x_1, ..., x_k))$, ενώ ϕ_θ' ο $m \times k$ πίνακας με στοιχεία $\frac{\partial \phi_i}{\partial x_j}$, για $i = 1, ..., m$, $j = 1, ..., k$, υπολογισμένα στο θ.

1.2.2 Αθροιστική συνάρτηση κατανομής και εκτίμησή της

Η αθροιστική συνάρτηση κατανομής $F_X(\cdot)$ μιας τ.μ. X επιτρέπει τον υπολογισμό πιθανοτήτων της μορφής $P\left(X \in B\right)$, για κάθε B, με $B \subseteq \mathbb{R}$, και ορίζεται ως εξής:

Ορισμός 1.2.4. *Έστω* $(\Omega, \mathcal{A}, P)$ *είναι χώρος πιθανότητας και* $X : \Omega \to S_X \subseteq \mathbb{R}$ *μια τυχαία μεταβλητή. Η συνάρτηση κατανομής ή αθροιστική συνάρτηση κατανομής (α.σ.κ.) της τυχαίας μεταβλητής* X *συμβολίζεται με* $F_X(\cdot)$ *και είναι* $F_X : \mathbb{R} \to [0, 1]$ *μια πραγματική συνάρτηση που ορίζεται από τη σχέση:*

$$F_X(x) = P(X \leq x) = P\left(\{\omega \in \Omega : X(\omega) \leq x\}\right), \quad x \in \mathbb{R}. \qquad (1.3)$$

Χρησιμοποιώντας τη σχέση (1.1) και τον προηγούμενο ορισμό προκύπτει το αποτέλεσμα που δίνεται στην πρόταση που ακολουθεί.

Πρόταση 1.2.3. *Έστω* $X \sim \mathcal{P}(\lambda)$, $\lambda > 0$. *Τότε η α.σ.κ. της τ.μ.* X

προσδιορίζεται από τη σχέση:

$$F_0(x,\lambda) = P(X \le x) = \begin{cases} 0, & x < 0, \\ \displaystyle\sum_{y=0}^{\lfloor x \rfloor} \frac{e^{-\lambda}\lambda^y}{y!}, & x \ge 0, \end{cases} \tag{1.4}$$

όπου το $\lfloor x \rfloor$ *συμβολίζει το ακέραιο μέρος του* x.

Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $F(\cdot)$. Εκτιμούμε την F με την εμπειρική αθροιστική συνάρτηση κατανομής (ε.α.σ.κ.), ο ορισμός της οποίας ακολουθεί.

Ορισμός 1.2.5 (ε.α.σ.κ.). *Έστω* $X_1, X_2, \ldots, X_n$ *ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής (α.σ.κ.)* $F(\cdot)$. *Η εμπειρική αθροιστική συνάρτηση κατανομής (ε.α.σ.κ.) συμβολίζεται με* $F_n(x)$ *και ορίζεται ως εξής:*

$$F_n(x) = \frac{\textit{πλήθος των } X_i \leqslant x}{n}, \;\; x \in \mathbb{R}, \tag{1.5}$$

ή ισοδύναμα

$$F_n(x) = \frac{\sum_{i=1}^{n} I_{(-\infty,x]}(X_i)}{n}, x \in \mathbb{R}, \tag{1.6}$$

όπου

$$I_C(X_i) = \begin{cases} 1\,, & X_i \in C, \\ 0\,, & X_i \notin C, \end{cases}$$

η συνάρτηση που συχνά αναφέρεται ως δείκτρια συνάρτηση.

Από τον ορισμό της εμπειρικής αθροιστικής συνάρτησης κατανομής (ή, αλλιώς, εμπειρική συνάρτηση κατανομής) γίνεται άμεσα αντιληπτό ότι πρόκειται για μία στατιστική συνάρτηση, καθώς είναι συνάρτηση των $X_1, \ldots, X_n$.

Ακολούθως, παρουσιάζονται κάποιες χρήσιμες ιδιότητες της εμπειρικής αθροιστικής συνάρτησης κατανομής (Gibbons and Chakraborti (2020)).

Θεώρημα 1.2.4. *Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(x)$.*

α) Για σταθερό x, η κατανομή της στατιστικής συνάρτησης $nF_n(x)$ είναι η διωνυμική με παραμέτρους n και $F(x)$.

β) Η ε.α.σ.κ. $F_n(x)$ είναι συνεπής εκτιμητής της αθροιστικής συνάρτησης κατανομής $F(x)$, και ακολουθεί προσεγγιστικά κανονική κατανομή με μέση τιμή $F(x)$ και διακύμανση $\{F(x)\,(1 - F(x))\}\,/n$.

Επιπλέον, στο θεώρημα που ακολουθεί και αποδείχτηκε από τον Glivenko (1933) για συνεχή συνάρτηση κατανομής F και από τον Cantelli (1933) για γενική συνάρτηση κατανομής F (για την απόδειξή του παραπέμπουμε, μεταξύ άλλων, στο σύγγραμμα του Loève (1977)) θεμελιώνεται ότι με πιθανότητα 1 η σύγκλιση της $F_n(x)$ στην $F\,(x)$ είναι ομοιόμορφη στο x ή, διαφορετικά, για μεγάλο μέγεθος δείγματος η προσέγγιση της $F\,(x)$ από την $F_n(x)$ είναι αρκετά ακριβής.

Θεώρημα 1.2.5. *Αν $F_n(x)$ είναι η εμπειρική αθροιστική συνάρτηση κατανομής ενός τυχαίου δείγματος $X_1, \ldots, X_n$, από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F\,(x)$, τότε:*

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{\sigma.\beta.} 0,$$

δηλαδή $P\left(\sup\limits_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0\right) = 1.$

Δηλαδή, σύμφωνα με το παραπάνω θεώρημα, η διαφορά της ε.α.σ.κ. από την πραγματική α.σ.κ. γίνεται μικρότερη, καθώς αυξάνει το μέγεθος του δείγματος.

Το αποτέλεσμα αυτό και η ερμηνεία του έχει οδηγήσει να προταθούν στη βιβλιογραφία διάφοροι έλεγχοι καλής προσαρμογής που στηρίζονται σε μέτρα εγγύτητας μεταξύ της εμπειρικής αθροιστικής συνάρτησης κατανομής και της αθροιστικής συνάρτησης κατανομής υπό τη μηδενική υπόθεση (βλέπε Ενότητα 2.2).

1.2.3 Ολοκληρώσιμη συνάρτηση κατανομής και εκτίμησή της

Αρχικά θα δοθεί ο ορισμός της ολοκληρώσιμης συνάρτησης κατανομής και ιδιότητές της, ενώ η ενότητα θα ολοκληρωθεί με τον ορισμό του δειγματικού ανάλογού της.

Ορισμός 1.2.6. *Έστω X μία διακριτή τ.μ. με συνάρτηση πιθανότητας $f(\cdot)$ και πεπερασμένη μέση τιμή. Η ολοκληρώσιμη συνάρτηση κατανομής ορίζεται ως ο μετασχηματισμός:*

$$\Psi_X(t) := E(X-t)^+ = \sum_{k=\lfloor t \rfloor+1}^{\infty} (k-t)P(X=k), \tag{1.7}$$

όπου $\lfloor t \rfloor$ το ακέραιο μέρος του t, ενώ $y^+ = max(0, y)$.

Επισημαίνεται σε αυτό το σημείο ότι η ολοκληρώσιμη συνάρτηση κατανομής έχει εμφανιστεί στη βιβλιογραφία του γνωστικού αντικειμένου της θεωρίας κινδύνου (risk theory) υπό τον όρο stop-loss tranform, ενώ έχει χρησιμοποιηθεί σε πλήθος εργασιών που αφορούν τη στοχαστική διάταξη (βλέπε Goovaerts et al. (1990) και τις εκεί αναφορές). Η ιδιότητα που κάνει την ολοκληρώσιμη συνάρτηση κατανομής να μπορεί να χρησιμοποιηθεί στη στατιστική και να είναι χρήσιμη στην πράξη είναι αυτή που διατυπώνεται στην πρόταση που ακολουθεί, για την ειδική περίπτωση διακριτών τυχαίων μεταβλητών, και ουσιαστικά είναι το μονοσήμαντο της ολοκληρώσιμης συνάρτησης κατανομής.

Πρόταση 1.2.6 (θεώρημα μονοσήμαντου της ολοκληρώσιμης συνάρτησης)**.** *Έστω X και Y δύο διακριτές τυχαίες μεταβλητές με τιμές στους μη αρνητικούς ακεραίους και πεπερασμένες μέσες τιμές. Τότε οι X και Y ακολουθούν την ίδια κατανομή αν και μόνο αν $\Psi_X(t) = \Psi_Y(t)$, για κάθε $t \in R$.*

Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη ολοκληρώσιμη αθροιστική συνάρτηση κατανομής. Εκτιμούμε την άγνωστη ολοκληρώσιμη συνάρτηση κατανομής με την εμπειρική ολοκληρώσιμη συνάρτηση κατανομής ο ορισμός της οποίας ακολουθεί.

Ορισμός 1.2.7. *Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής (α.σ.κ.) $F(\cdot)$ και ολοκληρώσιμη συνάρτηση κατανομής $\Psi(\cdot)$. Η εμπειρική ολοκληρώσιμη συνάρτηση κατανομής (ε.ο.σ.κ.) συμβολίζεται με $\Psi_n(t)$ και ορίζεται ως εξής:*

$$\Psi_n(t) = \frac{1}{n}\sum_{i=1}^{n}(X_i - t)I\{X_i > t\}(t), \tag{1.8}$$

όπου

$$I\{X_i > t\}\,(t) = \begin{cases} 1\,, & X_i > t, \\ 0\,, & X_i \le t. \end{cases}$$

1.2.4 Πιθανογεννήτρια συνάρτηση και εκτίμησή της

Αρχικά στην ενότητα αυτή θα δοθεί ο ορισμός της πιθανογεννήτριας συνάρτησης (probability generating function) και του δειγματικού ανάλογού της, ενώ θα εξηγηθεί και η ευρεία χρήση τους στη στατιστική. Τέλος, η ενότητα θα ολοκληρωθεί με την παράθεση ιδιοτήτων που ικανοποιεί η πιθανογεννήτρια συνάρτησης της κατανομής Poisson.

Ορισμός 1.2.8. *Για μία διακριτή τυχαία μεταβλητή X με τιμές στο σύνολο $\{0, 1, ...\}$ και συνάρτηση πιθανότητας $p_X(x)$, η πιθανογεννήτρια συνάρτηση $G_X(t)$ ορίζεται μέσω του ακόλουθου μετασχηματισμού:*

$$G_X(t) = E\left(t^X\right) = \sum_{x=0}^{\infty} t^x P(X = x),$$

με την παράμετρο t να είναι τέτοια ώστε να εξασφαλίζεται η σύγκλιση της σειράς (γενικά $|t| \leq 1$ με $t \in \mathbb{C}$, καθώς η πιθανογεννήτρια ορίζεται τουλάχιστον εντός του μοναδιαίου κύκλου).

Η χρήση της πιθανογεννήτριας συνάρτησης στην περίπτωση διακριτών δεδομένων για τη διεξαγωγή στατιστικής συμπερασματολογίας (εκτιμητική, έλεγχος υποθέσεων), καθώς και σε πλήθος άλλων προβλημάτων (επιλογής μοντέλου, ελέγχου καλής προσαρμογής, εντοπισμού σημείου αλλαγής κ.ο.κ.) είναι πλήρως αιτιολογημένη, καθώς, σύμφωνα με τους Nakamura and Pérez-Abreu (1993) η πιθανογεννήτρια συνάρτηση έχει τις περισσότερες φορές πιο απλή μορφή από τη συνάρτηση πιθανότητας ή την αθροιστική συνάρτηση κατανομής, υπάρχει πάντοτε για $t \in [0, 1]$ (σε αντίθεση με τη ροπογεννήτρια συνάρτηση, η οποία μπορεί να μην υπάρχει), ενώ χαρακτηρίζει πλήρως την κατανομή (μονοσήμαντο πιθανογεννητριών) λόγω της επόμενης πρότασης.

Πρόταση 1.2.7. *Έστω X και Y δύο διακριτές τυχαίες μεταβλητές με τιμές στους μη αρνητικούς ακεραίους. Οι τυχαίες μεταβλητές X και Y ακολουθούν την ίδια κατανομή αν και μόνο αν $G_X(t) = G_Y(t)$, για κάθε $t \in [-1, 1]$, όπου $G_X(t) = E(t^X)$ και $G_Y(t) = E(t^Y)$.*

Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη πιθανογεννήτρια συνάρτηση (probability generating function).Εκτιμούμε την άγνωστη πιθανογεννήτρια συνάρτηση με την εμπειρική πιθανογεννήτρια συνάρτηση, ο ορισμός της οποίας και κάποιες βασικές ιδιότητές της που αιτιολογούν τη χρήση της για ελέγχους καλής προσαρμογής.

Ορισμός 1.2.9. *Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής (α.σ.κ.) $F(\cdot)$ και πιθανογεννήτρια συνάρτηση $g(\cdot)$. Η εμπειρική πιθανογεννήτρια συνάρτηση (ε.π.σ.),$g_n(x)$, ορίζεται ως εξής:*

$$g_n(t) = \frac{1}{n}\sum_{i=1}^{n} t^{X_i},\ t \in [0,1]. \tag{1.9}$$

Παρατηρούμε ότι σε αντίθεση με την ε.α.σ.κ., η ε.π.σ. είναι συνεχής πραγματική συνάρτηση στο διάστημα $[0,1]$. Επιπλέον, η ε.π.σ. πληροί κάποιες πολύ χρήσιμες ιδιότητες που εξηγούν τη χρησιμότητά της στη στατιστική συμπερασματολογία και στον έλεγχο καλής προσαρμογής. Κάποιες από αυτές τις ιδιότητες παρατίθενται στη συνέχεια (βλέπε, μεταξύ άλλων, Nakamura and Pérez-Abreu (1993)).

Θεώρημα 1.2.8. *Αν $g_n(t)$ είναι η εμπειρική πιθανογεννήτρια συνάρτηση ενός τυχαίου δείγματος $X_1, \ldots, X_n$, από έναν πληθυσμό με πιθανογεννήτρια συνάρτηση $g(\cdot)$, τότε:*

α) η ε.π.σ. $g_n(t)$ είναι αμερόληπτος εκτιμητής της πιθανογεννήτριας συνάρτησης κατανομής,

β) η ε.π.σ. $g_n(t)$ είναι συνεπής εκτιμητής της πιθανογεννήτριας συνάρτησης κατανομής, και ισχύει ότι η ποσότητα $\sqrt{n}\{g_n(t) - g(t)\}$ συγκλίνει κατά κατανομή σε μία κανονική τυχαία μεταβλητή με μέση τιμή μηδέν και διακύμανση $\sigma^2(t) = g(t^2) - g^2(t)$.

Επιπλέον, στο θεώρημα που ακολουθεί (βλέπε, για παράδειγμα, Feuerverger (1988) και Nakamura and Pérez-Abreu (1993)) θεμελιώνεται ότι, για μεγάλο μέγεθος δείγματος, η προσέγγιση της $g(\cdot)$, του λογαρίθμου της και των παραγώγων της από την $g_n(t)$, τον λογάριθμό της και τις παραγώγους της είναι αρκετά ακριβής.

Θεώρημα 1.2.9. *Αν $g_n(\cdot)$ είναι η εμπειρική πιθανογεννήτρια συνάρτηση ενός τυχαίου δείγματος $X_1, \ldots, X_n$, από έναν πληθυσμό με πιθανογεννήτρια συνάρτηση $g(\cdot)$, τότε:*

$$\sup_{0\le t\le 1} |g_n(t) - g(t)| \xrightarrow{\sigma.\beta.} 0,$$

$$\sup_{0\le t\le 1} \left|g_n^{(k)}(t) - g^{(k)}(t)\right| \xrightarrow{\sigma.\beta.} 0,$$

ενώ

$$\sup_{0\le \epsilon\le t\le 1} |Y_n(t) - Y(t))| \xrightarrow{\sigma.\beta.} 0,$$

$$\sup_{0\le t\le 1} \left|Y_n^{(k)}(t) - Y^{(k)}(t)\right| \xrightarrow{\sigma.\beta.} 0,$$

όπου $Y_n(t) = \log(g_n(t))$, $Y(t) = \log(g(t))$, ενώ γενικά $h^{(k)}$ συμβολίζει την k-οστή παράγωγο της συνάρτησης h, με $k = 1, 2, ...$

Σύμφωνα με το παραπάνω θεώρημα, η διαφορά της ε.π.σ., του λογαρίθμου της αλλά και της παραγώγου k-οστής τάξης αυτής από τις αντίστοιχες πραγματικές γίνεται μικρότερη, καθώς αυξάνει το μέγεθος του δείγματος. Το αποτέλεσμα αυτό και η ερμηνεία του έχει οδηγήσει να προταθούν στη βιβλιογραφία διάφοροι έλεγχοι καλής προσαρμογής που στηρίζονται σε μέτρα εγγύτητας μεταξύ της εκτιμώμενης υπό τη μηδενική υπόθεση πιθανογεννήτριας συνάρτησης και της εμπειρικής πιθανογεννήτριας συνάρτησης.

Στην πρόταση που ακολουθεί προσδιορίζεται η πιθανογεννήτρια συνάρτηση της κατανομής Poisson. Ο προσδιορισμός της επιτυγχάνεται εύκολα ύστερα από λίγη άλγεβρα ανακαλώντας ότι εξ ορισμού ισούται με την αναμενόμενη τιμή $E\left(t^X\right)$.

Πρόταση 1.2.10. *Έστω $X \sim \mathcal{P}(\lambda)$, $\lambda > 0$. Τότε η πιθανογεννήτρια συνάρτησή της δίνεται από τη σχέση:*

$$g(t, \lambda) = e^{\lambda(t-1)}, t \in [0, 1]. \tag{1.10}$$

Απόδειξη. Με χρήση του ορισμού της πιθανογεννήτριας συνάρτησης και σ.π.π. της κατανομής Poisson, ισχύει ότι:

$$\begin{aligned} g(t, \lambda) &= \sum_{x=0}^{\infty} t^x \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} t^x \frac{(\lambda t)^x}{x!} \\ &= e^{-\lambda} e^{\lambda t} \\ &= e^{\lambda(t-1)}. \end{aligned}$$

κάνοντας χρήση της σειράς Taylor της εκθετικής συνάρτησης. □

Επιπλέον προκύπτει ότι η πιθανογεννήτρια της κατανομής Poisson ικανοποιεί την ιδιότητα που δίνεται στην επόμενη πρόταση.

Πρόταση 1.2.11. *Έστω $\mathcal{G}$ το σύνολο των πιθανογεννητριών συναρτήσεων που αντιστοιχούν σε τυχαίες μεταβλητές με τιμές στο $\mathbb{N}_0 = \{0, 1, ...\}$ και πεπερασμένη μέση τιμή. Τότε η πιθανογεννήτρια συνάρτηση της κατανομής Poisson είναι η μοναδική πιθανογεννήτρια συνάρτηση που ικανοποιεί τη διαφορική εξίσωση:*

$$\frac{\partial}{\partial t} g(t) - \lambda g(t) = 0.$$

Απόδειξη. Εύκολα μπορούμε να επιβεβαιώσουμε ότι η πιθανογεννήτρια συνάρτηση της κατανομής Poisson πληροί την εν λόγω διαφορική εξίσωση. Στη συνέχεια, θα αποδείξουμε ότι αυτή είναι η μοναδική πιθανογεννήτρια συνάρτηση. Είναι γνωστό ότι η γραμμική διαφορική εξίσωση πρώτης τάξης της μορφής $y' + p(t)y = 0$, όπου $y = y(t)$, $y' = (\partial/\partial t)y(t)$ και $p(t)$ μια συνεχής συνάρτηση στο t, δίνεται από τη σχέση $y = C\exp(-\int p(t)dt)$, όπου C είναι μια αυθαίρετη σταθερά. Εφαρμόζοντας αυτή τη γενική λύση στην παραπάνω διαφορική εξίσωση έχουμε ότι:

$$g(t) = C\exp\left(\lambda \int 1dt\right) = C\exp\left(\lambda t\right).$$

Λαμβάνοντας επιπρόσθετα υπόψη ότι εξ ορισμού $g(1) = E(1) = 1$ έχουμε ότι $C = \exp\left(-\lambda\right)$. Έτσι $g(t) = \exp\left(\lambda(t-1)\right)$, που ταυτίζεται με την πιθανογεννήτρια συνάρτηση της Poisson με παράμετρο λ. □

Από το αποτέλεσμα της προηγούμενης πρότασης, προκύπτει το ακόλουθο πόρισμα.

Πόρισμα 1.2.12. *Έστω $\mathcal{G}$ το σύνολο των πιθανογεννητριών συναρτήσεων που αντιστοιχούν σε τυχαίες μεταβλητές με τιμές στο $\mathbb{N}_0 = \{0, 1, ...\}$ και πεπερασμένη μέση τιμή. Τότε η πιθανογεννήτρια συνάρτηση της κατανομής Poisson είναι η μοναδική πιθανογεννήτρια συνάρτηση για την οποία ισχύει ότι:*

$$\frac{\partial^2}{\partial t^2}\log(g(t)) = 0.$$

Απόδειξη. Από την Πρόταση 1.2.11 έχουμε ότι η πιθανογεννήτρια της κατανομής Poisson είναι η μόνη πιθανογεννήτρια στο σύνολο των $\mathcal{G}$ η οποία πληροί τη σχέση:

$$\frac{\frac{\partial}{\partial t}g(t)}{g(t)} = \lambda,$$

ή ισοδύναμα η μόνη πιθανογεννήτρια στο σύνολο των $\mathcal{G}$ για την οποία ισχύει ότι:

$$\frac{\partial}{\partial t}\log(g(t)) = \lambda.$$

Παίρνοντας τη δεύτερη παράγωγο προκύπτει άμεσα το ζητούμενο. □

1.2.5 Ροπογεννήτρια, ροπές, ημιαναλλοίωτες και άλλες ιδιότητες

Αρχικά στην ενότητα αυτή θα δοθεί ο ορισμός της ροπογεννήτριας και της ημιαναλλοίωτης γεννήτριας συνάρτησης (moment generating, cumulant generating function, αντίστοιχα), θα προσδιοριστεί η ροπογεννήτρια συνάρτηση της κατανομής Poisson και θα χρησιμοποιηθεί για τον προσδιορισμό σχέσεων που αφορούν τις ροπές και τις ημιαναλλοίωτες αυτής της κατανομής. Τέλος, η ενότητα θα ολοκληρωθεί με την παράθεση ιδιοτήτων που συνδέονται με τις ροπές της κατανομής Poisson.

Ορισμός 1.2.10. *Έστω X μία τυχαία μεταβλητή με α.σ.κ. $F_X(\cdot)$. Η ροπογεννήτρια συνάρτηση της τυχαίας μεταβλητής X συμβολίζεται με $M_X(t)$ και ορίζεται από τη σχέση:*

$$M_X(t) = E\left(e^{tX}\right)$$

υπό την προϋπόθεση ότι η αναμενόμενη τιμή υπάρχει για t στη γειτονιά του 0. Αυτό σημαίνει ότι υπάρχει ένα $h > 0$ τέτοιο ώστε για όλα τα $t \in (-h, h)$ η αναμενόμενη τιμή να υπάρχει. Αν η αναμενόμενη τιμή δεν υπάρχει στη γειτονιά του 0, λέμε ότι η ροπογεννήτρια συνάρτηση δεν υπάρχει. Τέλος, η ημιαναλλοίωτη γεννήτρια συνάρτηση ορίζεται από τη σχέση $K_X(t) = log(M_X(t))$.

Η ροπογεννήτρια συνάρτηση καλείται έτσι καθώς αν υπάρχει σε ένα ανοικτό διάστημα γύρω από το μηδέν, επιτρέπει τον υπολογισμό των απλών ροπών τάξης k μέσω της σχέσης:

$$E(X^k) = \frac{d^k M_X(t)}{dt^k}\Big|_{t=0},$$

δηλαδή η k-οστή τάξης απλή ροπή ισούται με την k-οστή παράγωγο της ροπογεννήτριας ως προς t υπολογισμένη στο $t = 0$, με $k = 1, 2, \ldots$. Επιπλέον, οι ημιαναλλοίωτες τάξης r, έστω k_r, αποκτώνται γράφοντας την ημιαναλλοίωτη συνάρτηση ως δυναμοσειρά και τότε είναι:

$$K_X(t) = \sum_{r=1}^{\infty} \frac{k_r t^r}{r!}.$$

Παρατήρηση 1.2.13. *Η σχέση μεταξύ των πρώτων τεσσάρων ημιανναλοίωτων και των ροπών $E(X^k)$ είναι η ακόλουθη:*

$$k_1 = E(X),\ k_2 = E(X^2) - (E(X))^2,$$

$$k_3 = E(X^3) - 3E(X^2)E(X) + 2(E(X))^3,$$

και

$$k_4 = E(X^4) - 4E(X^3)E(X) - 3(E(X^2))^2 + 12E(X^2)(E(X))^2 - 6(E(X))^4.$$

Χρησιμοποιώντας τη σχέση (1.1) και τον ορισμό της ροπογεννήτριας και της ημιαναλλοίωτης γεννήτριας συνάρτησης προκύπτει το αποτέλεσμα που δίνεται στην πρόταση που ακολουθεί.

Πρόταση 1.2.14. *Έστω $X \sim \mathcal{P}(\lambda)$, $\lambda > 0$. Τότε η ροπογεννήτρια και η ημιαναλλοίωτη γεννήτρια συνάρτησή της δίνονται από τις σχέσεις:*

$$M_X(t) = \exp[\lambda(e^t - 1)], t \in \mathbb{R}, \tag{1.11}$$

και

$$K_X(t) = \lambda(e^t - 1) = \lambda \sum_{r=1}^{\infty} \frac{t^r}{r!}, t \in \mathbb{R},$$

αντίστοιχα.

Λαμβάνοντας υπόψη τα παραπάνω εύκολα προκύπτουν τα αποτελέσματα των δύο προτάσεων που ακολουθούν.

Πρόταση 1.2.15. *Έστω X η τυχαία μεταβλητή που ακολουθεί $\mathcal{P}(\lambda)$, με σ.π. που προσδιορίζεται στη σχέση (1.1). Τότε οι τέσσερις πρώτες μη κεντρικές ροπές της, $\mu_i^{'} = E(X^i)$, δίνονται από τις σχέσεις:*

$$E(X) = \lambda, \tag{1.12}$$

$$E(X^2) = \lambda^2 + \lambda, \tag{1.13}$$

$$E(X^3) = \lambda^3 + 3\lambda^2 + \lambda, \tag{1.14}$$

ενώ

$$E(X^4) = \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda. \tag{1.15}$$

Πρόταση 1.2.16. *Έστω $X \sim \mathcal{P}(\lambda)$, με σ.π. που προσδιορίζεται στη σχέση (1.1). Τότε οι ημιαναλλοίωτές της είναι ίσες με την παράμετρο λ, δηλαδή $k_r = \lambda$, για $r = 1, 2, \ldots$.*

Από την Πρόταση 1.2.15 προκύπτει ότι η παράμετρος λ είναι η μέση τιμή της κατανομής και επομένως παριστάνει τον μέσο αριθμό εμφανίσεων γεγονότων στη μονάδα του χρόνου. Επιπρόσθετα, χρησιμοποιώντας τα αποτελέσματα της Πρότασης 1.2.15 προκύπτει ότι η διακύμανση της τ.μ. $X \sim \mathcal{P}(\lambda)$, $\lambda > 0$, είναι

$$Var(X) = \lambda. \tag{1.16}$$

Επομένως, προκύπτει το ακόλουθο συμπέρασμα.

Πόρισμα 1.2.17. *Η κατανομή Poisson έχει τη χαρακτηριστική ιδιότητα ότι η μέση τιμή της είναι ίση με τη διακύμανση. Εναλλακτικά ο λεγόμενος δείκτης διασποράς (index of dispersion), έστω δ, που ορίζεται ως το πηλίκο της διακύμανσης προς τη μέση τιμή, δηλαδή $\delta = \dfrac{\sigma^2}{\mu}$, και αποτελεί ένα κανονικοποιημένο μέτρο της διασποράς μίας κατανομής, είναι ίσος με 1 στην περίπτωση της κατανομής Poisson.*

Παρατήρηση 1.2.18. *Οι υψηλότερης τάξης μη κεντρικές ροπές της κατανομής Poisson προσδιορίζονται από την ακόλουθη σχέση:*

$$E(X^k) = \sum_{i=0}^{k} \lambda^i \left\{ \begin{matrix} k \\ i \end{matrix} \right\},$$

όπου

$$\left\{ \begin{matrix} n \\ k \end{matrix} \right\} = \frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} (k-i)^n,$$

οι λεγόμενοι αριθμοί Stirling δεύτερου είδους.

Επιπρόσθετα, χρησιμοποιώντας τα αποτελέσματα της Πρότασης 1.2.15 προσδιορίζονται εύκολα, μετά από λίγη άλγεβρα και λαμβάνοντας υπόψη την σχέση που συνδέει απλές και κεντρικές ροπές, δηλαδή ότι

$$\mu_i = \sum_{r=0}^{i} (-1)^r \binom{i}{r} (E(X))^r E(X^{i-r}),$$

όπου $\mu_i = E\left(X - E(X)\right)^i$, για $i = 1, ..., 4$, καθώς και οι συντελεστές λοξότητας και κύρτωσης της κατανομής αυτής. Τα αποτελέσματα αυτά συνοψίζονται στο πόρισμα που ακολουθεί.

Πόρισμα 1.2.19. *Έστω X η τυχαία μεταβλητή που ακολουθεί $\mathcal{P}(\lambda)$, με σ.π. που προσδιορίζεται στη σχέση (1.1). Τότε οι τέσσερις πρώτες κεντρικές ροπές της δίνονται από τις σχέσεις:*

$$\mu_1 = 0,\ \mu_2 = \lambda,\ \mu_3 = \lambda, \mu_4 = \lambda(1 + 3\lambda)$$

ενώ οι συντελεστές λοξότητας και κύρτωσης δίνονται από τις σχέσεις:

$$Skew(X) = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{1}{\sqrt{\lambda}}, \tag{1.17}$$

και

$$Kurt(X) = \frac{\mu_4}{\mu_2^2} - 3 = \frac{1}{\lambda}, \tag{1.18}$$

αντίστοιχα.

Λαμβάνοντας υπόψη τις σχέσεις (1.17) και (1.18) προκύπτει άμεσα μετά από λίγη άλγεβρα η ακόλουθη ιδιότητα της κατανομής Poisson, η οποία θα χρησιμοποιηθεί για την κατασκευή ενός στατιστικού ελέγχου καλής προσαρμογής της κατανομής Poisson (βλέπε Gupta et al. (1994)).

Πρόταση 1.2.20. *Έστω X η τυχαία μεταβλητή που ακολουθεί $\mathcal{P}(\lambda)$, με σ.π.*

που προσδιορίζεται στη σχέση (1.1). Τότε

$$\mu_2\mu_4 - 3\mu_2^3 - \mu_3^2 = 0 \tag{1.19}$$

όπου $\mu_i = E\left(X - E(X)\right)^i$.

Απόδειξη. Στην ειδική περίπτωση της Poisson με παράμετρο λ έχουμε ότι $\mu_1 = 0$, $\mu_2 = \lambda$, $\mu_3 = \lambda$ και $\mu_4 = 3\lambda^2 + \lambda$. Άρα:

$$\mu_2\mu_4 - 3\mu_2^3 - \mu_3^2 = \lambda(3\lambda^2 + \lambda) - 3(\lambda)^3 - \lambda^2 = 3\lambda^3 + \lambda^2 - 3\lambda^3 - \lambda^2 = 0, \tag{1.20}$$

που αποδεικνύει το ζητούμενο. □

Μία άλλη ιδιότητα της κατανομής Poisson δίνεται στην πρόταση που ακολουθεί (βλέπε Kyriakoussis et al. (1998)).

Πρόταση 1.2.21. *Έστω X η τυχαία μεταβλητή που ακολουθεί $\mathcal{P}(\lambda)$, με σ.π. που προσδιορίζεται στη σχέση (1.1). Τότε*

$$\frac{E(X^2) - E(X)}{(E(X))^2} - 1 = 0 \tag{1.21}$$

ή ισοδύναμα

$$\frac{\mu_2^{'} - \mu_1^{'}}{\left(\mu_1^{'}\right)^2} = 1, \tag{1.22}$$

όπου $\mu_j^{'} = E(X^j)$.

Απόδειξη. Λαμβάνοντας υπόψιν ότι $\mu_1^{'} = \lambda$ και $\mu_2^{'} = \lambda^2 + \lambda$ προκύπτει ότι:

$$\frac{\mu_2^{'} - \mu_1^{'}}{\left(\mu_1^{'}\right)^2} = \frac{\lambda + \lambda^2 - \lambda}{(\lambda)^2} = 1.$$

□

Δεσμευμένη κατανομή δοθέντος της επαρκούς στατιστικής συνάρτησης

Έστω $X_1, X_2, ..., X_n$ είναι διακριτές τυχαίες μεταβλητές που ακολουθούν κατανομή Poisson με παράμετρο λ. Τότε, με χρήση του παραγοντικού θεωρήματος, αποδεικνύεται ότι η στατιστική συνάρτηση (σ.σ.) $T_n = \sum_{i=1}^{n} X_i$ είναι επαρκής σ.σ. για την παράμετρο λ. Το αποτέλεσμα αυτό σημαίνει ότι η T_n περιέχει όλη την πληροφορία του τ.δ. για την άγνωστη παράμετρο λ, ενώ ταυτόχρονα από τον ορισμό της επάρκειας έχουμε ότι η δεσμευμένη κατανομή των $X_1, X_2, ..., X_n$ δοθέντος ότι $T_n = t$ είναι ανεξάρτητη της παραμέτρου λ. Ειδικότερα, αξιοποιώντας το γεγονός ότι η κατανομή της σ.σ. T_n είναι Poisson με παράμετρο $n\lambda$ (προκύπτει με τη μέθοδο της ροπογεννήτριας ως άθροισμα ανεξάρτητων και ισόνομων τ.μ., βλέπε Παπαιωάννου (1993)), έχουμε ότι:

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n | T_n = t) = \frac{t!}{x_1! \cdots x_n!} \left(\frac{1}{n}\right)^t. \tag{1.23}$$

Από την τελευταία σχέση συμπεραίνουμε ότι η δεσμευμένη κατανομή των X_1, X_2,...,X_n δοθέντος της επαρκούς σ.σ. $T_n = \sum_{i=1}^{n} X_i$ είναι πολυωνυμική με πιθανότητες $p_1 = p_2 = ... = p_n = \frac{1}{n}$.

1.2.6 Μέση απόσταση

Θα παραθέσουμε αρχικά τον ορισμό της έννοιας της μέσης απόστασης και στη συνέχεια με βάση αυτόν τον ορισμό θα δοθούν κάποιες ιδιότητες.

Ορισμός 1.2.11. *Έστω X τυχαία μεταβλητή με τιμές στο $\mathbb{N}_0 = \{0, 1, ...\}$ και συνάρτηση πιθανότητας $f_X(\cdot)$. Η μέση απόσταση m_k ενός σταθεροποιημένου* *πραγματικού αριθμού $k \in \mathbb{R}$ από όλες τις πιθανές τιμές 0,1,2,... της τ.μ. X ορίζεται από τη σχέση:*

$$m_k^X = E(|k - X|), k \in \mathbb{R}$$

ή ισοδύναμα

$$m_k^X = \sum_{j=0}^{\infty} |k - j| P(X = j), \text{ για } k \in \mathbb{R}.$$

Σε περιπτώσεις που δεν θα δημιουργείται σύγχυση αντί του συμβολισμού m_k^X θα χρησιμοποιείται ο συμβολισμός m_k.

Η ιδιότητα που κάνει την παραπάνω συνάρτηση να μπορεί να χρησιμοποιηθεί στον έλεγχο καλής προσαρμογής είναι αυτή που διατυπώνεται στην πρόταση που ακολουθεί (βλέπε Székely and Rizzo (2004)).

Πρόταση 1.2.22. *Έστω X και Y δύο διακριτές τυχαίες μεταβλητές με τιμές στους μη αρνητικούς ακεραίους, με πεπερασμένες μέσες τιμές. Οι τυχαίες μεταβλητές X και Y ακολουθούν την ίδια κατανομή αν και μόνο αν $m_k^X = m_k^Y$, για κάθε μη αρνητικό ακέραιο k.*

Επομένως, στην παραπάνω πρόταση διατυπώνεται το μονοσήμαντο αυτού του μετασχηματισμού και αξιοποιώντας το αποτέλεσμά της προκύπτει ότι ο έλεγχος καλής προσαρμογής της κατανομής Poisson ανάγεται στον έλεγχο ότι η άγνωστη μέση απόσταση m_k^X του πληθυσμού ταυτίζεται με τη μέση απόσταση m_{ok} υπό τη μηδενική υπόθεση ότι τα δεδομένα προέρχονται από την Poisson με παράμετρο λ, για κάθε τιμή k. Στην πρόταση που ακολουθεί προσδιορίζεται η μέση απόσταση m_k^X, όταν $X \sim \mathcal{P}(\lambda)$, $\lambda > 0$.

Πρόταση 1.2.23. *Έστω $X \sim \mathcal{P}(\lambda)$, $\lambda > 0$. Τότε η μέση απόσταση m_k^X ενός σταθεροποιημένου πραγματικού αριθμού $k \in R$ από όλες τις πιθανές τιμές 0,1,2,...*

της τ.μ. X προσδιορίζεται από τη σχέση:

$$E(|k-X|) = 2(k-\lambda)F_0(k-1,\lambda) + 2\lambda f_0(k-1,\lambda) - (k-\lambda), \tag{1.24}$$

όπου $F_0(x,\lambda)$ και $f_0(x,\lambda)$ η α.σ.κ. και η σ.π. της $X \sim \mathcal{P}(\lambda)$, $\lambda > 0$.

Απόδειξη. Από τον ορισμό της μέσης απόστασης m_k^X, λαμβάνοντας υπόψη τη σ.π. της κατανομής Poisson με παράμετρο λ προκύπτει ότι:

$$E(|k-X|) = 2\sum_{i=0}^{k-i} \frac{e^{-\lambda}\lambda^i}{i!} + \lambda - k.$$

Στη συνέχεια, λαμβάνοντας υπόψη τις σχέσεις:

$$\lambda f_0(k-1,\lambda) = kf_0(k,,\lambda) \text{ και } \sum_{i=0}^{k} f_0(i,\lambda) = \lambda F_0(k-1,\lambda),$$

παίρνουμε το ζητούμενο αποτέλεσμα. □

Παρατήρηση 1.2.24. *Στην περίπτωση της Poisson κατανομής μια αναδρομική σχέση για τον προσδιορισμό των m_k προκύπτει ακολουθώντας το εξής σκεπτικό (βλέπε Székely and Rizzo (2004)). Για $k = 0$ εξ ορισμού είναι $m_0 = E|X| = \lambda$. Επιπλέον για $k = 1$ έχουμε ότι:*

$$\begin{aligned} m_1 &= E\left(|1-X|\right) = \sum_{x=0}^{\infty} |1-x| P(X=x) \\ &= P(X=0) + \sum_{x=2}^{\infty} (x-1)P(X=x), \end{aligned}$$

ή ύστερα από λίγη άλγεβρα

$$m_1 = 2P(X=0) + E(X) - 1 = 2f_0(0,\lambda) - (1-\lambda).$$

Επομένως προκύπτει ότι

$$P(X=0)=\frac{m_1+1-\lambda}{2},$$

ή ισοδύναμα, καθώς $P(X=0)=f_0(0,\lambda)=F_0(0,\lambda)$, *έχουμε ότι:*

$$f_0(0,\lambda)=F_0(0,\lambda)=\frac{m_1+1-\lambda}{2}. \tag{1.25}$$

Λύνοντας τη σχέση (1.24) *ως προς* $f_0(k,\lambda)$, *προκύπτει ότι:*

$$f_0(k,\lambda)=\frac{m_{k+1}-(k+1-\lambda)(2F_0(k-1,\lambda)-1)}{2(k+1)}. \tag{1.26}$$

1.2.7 Παραμετρικό bootstrap στον έλεγχο καλής προσαρμογής

Όπως θα δούμε στο επόμενο κεφάλαιο πολλές φορές η ασυμπτωτική κατανομή της στατιστικής συνάρτησης που χρησιμοποιείται για τον έλεγχο καλής προσαρμογής για την κατανομή Poisson εξαρτάται από την άγνωστη παράμετρο λ ή ακόμη είναι και δύσκολο να προσδιοριστεί. Για την αντιμετώπιση αυτού του προβλήματος, τις περισσότερες φορές οι συγγραφείς καταφεύγουν στη χρησιμοποίηση παραμετρικού bootstrap. Στην ενότητα αυτή θα δοθούν τα βήματα υλοποίησης αυτής της μεθόδου στην περίπτωση που το διαθέσιμο σύνολο δεδομένων είναι το X_1,...,X_n, η στατιστική συνάρτηση που χρησιμοποιείται είναι η $D_n(\hat{\lambda})$ (βλέπε 2.2.1)και η μηδενική υπόθεση απορρίπτεται για μεγάλες τιμές της στατιστικής συνάρτησης ελέγχου.

Στο πλαίσιο που περιγράφηκε παραπάνω, τα βήματα είναι τα ακόλουθα.

1. Από τις διαθέσιμες παρατηρήσεις εκτιμούμε την άγνωστη παράμετρο λ, έστω με την εκτιμήτρια $\hat{\lambda}$.

2. Υπολογίζουμε την τιμή της στατιστικής συνάρτησης $D_n(\hat{\lambda})$.

3. Δημιουργούμε B το πλήθος (B μεγάλος αριθμός) τυχαία δείγματα μεγέθους n από την κατανομή με αθροιστική συνάρτηση κατανομής $F_0(x; \hat{\lambda})$, δηλαδή από την κατανομή Poisson$(\widehat{\lambda})$ και για καθένα από αυτά υπολογίζουμε την τιμή της στατιστικής συνάρτησης $D_n(\hat{\lambda}^{(j)})$, όπου $\hat{\lambda}^{(j)}$, $j = 1, ..., B$ ο εκτιμητής της παραμέτρου λ που προκύπτει στο $j-$οστό δείγμα.

4. Εκτιμούμε την p-τιμή του ελέγχου ως το ποσοστό των φορών που η τιμή της στατιστικής συνάρτησης $D_n(\hat{\lambda}^{(j)})$ είναι μεγαλύτερη από την τιμή $D_n(\hat{\lambda})$.

5. Αν η p-τιμή είναι μεγαλύτερη ή ίση από το επίπεδο σημαντικότητας απορρίπτεται η μηδενική υπόθεση.

ΚΕΦΑΛΑΙΟ 2

Ελεγχοι καλησ προσαρμογησ τησ κατανομησ Poisson

Έστω $X_1, X_2, ..., X_n$ ένα τυχαίο δείγμα μεγέθους n από έναν πληθυσμό που λαμβάνει τιμές στο $\mathbb{N}_0 = \mathbb{N} \cup 0 = \{0, 1, 2, ...\}$, με αθροιστική συνάρτηση κατανομής (α.σ.κ.) $F(x)$ (αντίστοιχα, πιθανογεννήτρια κ.ο.κ.), η οποία είναι άγνωστη. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση $H_0 : F(x) = F_0(x, \lambda)$, για κάθε $x \in \mathbb{R}$, όπου $F_0(x, \lambda)$ είναι η αθροιστική συνάρτηση κατανομής της κατανομής Poisson. Το πρόβλημα ελέγχου της παραπάνω υπόθεσης καλείται **έλεγχος καλής προσαρμογής της κατανομής Poisson** και διακρίνεται στον απλό έλεγχο καλής προσαρμογής, αν η παράμετρος λ είναι γνωστή, ενώ διαφορετικά λέμε ότι έχουμε έναν σύνθετο έλεγχο καλής προσαρμογής. Είναι προφανές ότι συνηθέστερα σε πρακτικές εφαρμογές το ενδιαφέρον επικεντρώνεται στον σύνθετο έλεγχο καλής προσαρμογής. Επισημαίνεται ότι οι απλοί έλεγχοι δεν μπορούν μα χρησιμοποιηθούν σε περιπτώσεις σύνθετων ελέγχων, καθώς τα αποτελέσματά τους παύουν να ισχύουν σε αυτή την περίπτωση.

Στη στατιστική βιβλιογραφία έχουν αναπτυχθεί διάφορες στατιστικές μεθοδολογίες για τον έλεγχο καλής προσαρμογής στην κατανομή Poisson. Η έρευνα σε αυτό το αντικείμενο είναι πάρα πολύ μεγάλη και έχουν εμφανιστεί τόσο γραφικοί όσο και στατιστικοί τρόποι ελέγχου καλής προσαρμογής. Ο αριθμός των ελέγχων που έχουν προταθεί είναι ιδιαίτερα μεγάλος, καθώς οι τρόποι απόκλισης από την κατανομή Poisson είναι αρκετοί και αυτό έχει ως αποτέλεσμα να μην είναι εφικτό να υπάρξει ένας έλεγχος που θα είναι για όλες τις εναλλακτικές κατανομές ο πλέον ισχυρός. Για παράδειγμα, έχουν εμφανιστεί στατιστικοί τρόποι ελέγχου που στηρίζονται σε μέτρα αποκλίσεων μεταξύ της εμπειρικής αθροιστικής συνάρτησης και της εκτιμώμενης αθροιστικής συνάρτησης κατανομής υπό την υπόθεση ότι τα δεδομένα προέρχονται από την κατανομή Poisson, άλλοι που στηρίζονται σε μέτρα αποκλίσεων μεταξύ της εμπειρικής πιθανογεννήτριας συνάρτησης και της εκτιμώμενης πιθανογεννήτριας συνάρτησης της υπό έλεγχο κατανομής. Επιπρόσθετα, έχουν εμφανιστεί στη βιβλιογραφία έλεγχοι που βασίζονται σε χαρακτηρισμούς και ιδιότητες της κατανομής Poisson κ.ο.κ. Στο κεφάλαιο αυτό, αρχικά, παρουσιάζεται, για ιστορικούς κυρίως λόγους, ο έλεγχος χι-τετράγωνο καλής προσαρμογής, που αποτελεί τον αρχαιότερο τέτοιον έλεγχο. Στη συνέχεια, παρουσιάζονται στατιστικοί έλεγχοι καλής προσαρμογής ταξινομημένοι σε διακριτές κατηγορίες (π.χ. έλεγχοι με βάση την εμπειρική αθροιστική συνάρτηση ή την εμπειρική πιθανογεννήτρια συνάρτηση), ενώ το κεφάλαιο ολοκληρώνεται με την παρουσίαση κάποιων γραφικών τρόπων ελέγχου καλής προσαρμογής της κατανομής Poisson.

2.1 Χι-τετράγωνο έλεγχος

Ο αρχαιότερος και περισσότερο γνωστός έλεγχος καλής προσαρμογής είναι ο X^2 (χι-τετράγωνο), ο οποίος προτάθηκε από τον Karl Pearson (1857-1936) στην εργασία του Pearson (1900). Ο έλεγχος αυτός απαιτεί τα δεδομένα να δίνονται σε k το πλήθος ομάδες ή να δύναται να χωριστούν σε k το πλήθος ομάδες και αξιοποιεί, στη συνέχεια, τις συχνότητες εμφάνισης κάθε ομάδας. Αρχικά υποθέτουμε ότι η κατανομή Poisson μας είναι πλήρως γνωστή υπό τη μηδενική υπόθεση (απλός έλεγχος καλής προσαρμογής).

Απλός έλεγχος χι-τετράγωνο καλής προσαρμογής

Σε όσα ακολουθούν συμβολίζουμε με p_{0i} την πιθανότητα, υπό τη μηδενική υπόθεση, μία τυχαία παρατήρηση από την τυχαία μεταβλητή X να ανήκει στην i-οστή κατηγορία, $i = 1, ..., k$, ενώ ο έλεγχος παρουσιάζεται για τη γενική περίπτωση που οι διαθέσιμες ομάδες είναι οι $(a, x_1]$,...,$(x_{k-1}, \beta]$. Επομένως,

$$p_{01} = F_0\left(x_1\right) - F_0\left(\alpha\right), p_{02} = F_0\left(x_2\right) - F_0\left(x_1\right), ..., p_{0k} = F_0\left(\beta\right) - F_0\left(x_{k-1}\right),$$

με $\sum_{i=1}^{k} p_{0i} = 1$ και F_0 την α.σ.κ. της Poisson υπό τη μηδενική υπόθεση. Τότε, αν $N_1, N_2, \ldots, N_k$ είναι οι τυχαίες μεταβλητές που παριστάνουν το πλήθος των παρατηρήσεων που ανήκουν σε κάθε μία από τις k το πλήθος ομάδες, στις n που έχουν επιλεχθεί, τότε μέσω ορισμού ισχύει ότι το τυχαίο διάνυσμα $(N_1, ..., N_k)$ ακολουθεί, υπό τη μηδενική υπόθεση, πολυωνυμική κατανομή με παραμέτρους n, p_{01},..., p_{0k}. Στο παραπάνω πλαίσιο, έστω n_i (e_i) ο παρατηρούμενος (αναμενόμενος υπό τη μηδενική υπόθεση) αριθμός των παρατηρήσεων που ανήκουν σε κάθε μία από τις k το πλήθος ομάδες. Τότε, καθώς η περιθώρια κατανομή της τυχαίας μεταβλητής N_i ακολουθεί, υπό τη μηδενική υπόθεση H_0, διωνυμική κατανομή με παραμέτρους n και p_{0i}, έχουμε ότι:

$$e_i = E(N_i) = np_{0i}, i = 1, ..., k,$$

με $\sum_{i=1}^{k} e_i = n$. Ο X^2 έλεγχος καλής προσαρμογής αντιπαραβάλλει τον αριθμό των παρατηρούμενων παρατηρήσεων που ανήκουν στο i-οστό υποδιάστημα με τον αναμενόμενο αριθμό αυτών υπό την H_0 και βασίζεται στη στατιστική συνάρτηση:

$$
\begin{aligned}
X^2 &= \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \\
&= \sum_{i=1}^{k} \frac{n_i^2}{e_i} - n. \qquad (2.1)
\end{aligned}
$$

Όταν δεν ισχύει η μηδενική υπόθεση, τότε κάθε παρατηρούμενος αριθμός n_i θα διαφέρει αρκετά από τον αντίστοιχο αναμενόμενο e_i. Άρα, καθώς, η στατιστική συνάρτηση είναι το άθροισμα των τετραγώνων αυτών των διαφορών διαιρεμένων με τον αντίστοιχο αναμενόμενο αριθμό αυτών, συνεπάγεται ότι η μηδενική υπόθεση H_0 θα απορρίπτεται για μεγάλες τιμές της στατιστικής συνάρτησης. Το εύλογο ερώτημα που προκύπτει είναι ποιες τιμές της στατιστικής συνάρτησης μπορούν να θεωρηθούν μεγάλες. Η απάντηση αυτή μπορεί να δοθεί, αν προσδιοριστεί η κατανομή της στατιστικής συνάρτησης υπό τη μηδενική υπόθεση. Ο προσδιορισμός αυτός επιτυγχάνεται στην πρόταση που ακολουθεί (βλέπε, μεταξύ άλλων, DasGupta (2008)).

Πρόταση 2.1.1. *Έστω $N_1, N_2, \ldots, N_k$ είναι οι τυχαίες μεταβλητές που παριστάνουν το πλήθος των παρατηρήσεων που ανήκουν σε κάθε μία από τις k το πλήθος ομάδες, στις n τυχαία επιλεγμένες παρατηρήσεις από την X. Τότε, υπό την υπόθεση ότι η αθροιστική συνάρτηση κατανομής της τυχαίας μεταβλητής X είναι η $F_0(\cdot)$, η τυχαία μεταβλητή*

$$
X^2 = \sum_{i=1}^{k} \frac{(N_i - e_i)^2}{e_i}, \qquad (2.2)
$$

όπου $e_i = np_{0i}$, ακολουθεί ασυμπτωτικά χι-τετράγωνο κατανομή με $k-1$ βαθμούς ελευθερίας (χ^2_{k-1}).

Από την παραπάνω πρόταση προκύπτει ότι απορρίπτουμε σε επίπεδο σημαντικότητας α την H_0, αν για την παρατηρούμενη τιμή της στατιστικής συνάρτησης ελέγχου, έστω X^2_{obs}, ισχύει ότι:

$$X^2_{obs} = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \geqslant \chi^2_{k-1,a}, \tag{2.3}$$

όπου $\chi^2_{k-1,a}$ είναι η τιμή εκείνη για την οποία ισχύει ότι $P\left(\chi^2_{k-1} \geqslant \chi^2_{k-1,a}\right) = a$, όπου a το ποσοστιαίο σημείο της χι-τερτάγωνο. Τιμές των $\chi^2_{k-1,a}$ είναι διαθέσιμες στη βιβλιογραφία για διάφορους βαθμούς ελευθερίας και επίπεδα σημαντικότητας. Άμεσα προκύπτει ότι η p-τιμή του ελέγχου προσδιορίζεται από τη σχέση:

$$P\left(\chi^2_{k-1} \geqslant X^2_{obs}\right) = 1 - F_{\chi^2_{k-1}}\left(X^2_{obs}\right),$$

όπου $F_{\chi^2_{k-1}}$ η αθροιστική συνάρτηση κατανομής της χ^2_{k-1} κατανομής.

Παρατήρηση 2.1.2. *Ένα πρώτο μειονέκτημα του ελέγχου χι-τετράγωνο καλής προσαρμογής είναι ότι πρόκειται για ασυμπτωτικό έλεγχο και, επομένως, πρέπει να έχουμε διαθέσιμο μεγάλο σε μέγεθος δείγμα. Επιπλέον, απαιτεί τα δεδομένα να διαχωριστούν σε ομάδες και το ερώτημα που προκύπτει είναι αν ο αριθμός των ομάδων και ο τρόπος καθορισμού τους επηρεάζει την ισχύ του ελέγχου. Δυστυχώς, η απάντηση είναι καταφατική και έχουν προταθεί διάφορες επιλογές για τον προσδιορισμό του πλήθους των ομάδων k. Ειδικότερα, ο Moore (1986) πρότεινε ο αριθμός των ομάδων k να είναι $k = 2n^{\frac{2}{5}}$, οι Kvam and Vidakovic (2007) πρότειναν να ισχύει $k \geq 3$, $\frac{n^2}{k} \geq 10$ και $e_i \geq 0.25$, ενώ οι Gibbons and Chakraborti (2014) πρότειναν ο αριθμός των ομάδων να είναι τέτοιος ώστε κάθε ομάδα να έχει, υπό τη μηδενική υπόθεση, αναμενόμενο αριθμό παρατηρήσεων πάνω από πέντε. Ο τελευταίος κανόνας είναι και ο πιο γνωστός και σε περίπτωση που παραβιάζεται ενοποιούμε ομάδες έτσι ώστε να ισχύει.*

Σύνθετος έλεγχος χι-τετράγωνο καλής προσαρμογής

Η περίπτωση μίας πλήρους ορισμένης κατανομής Poisson είναι σχεδόν σπάνια στις πρακτικές εφαρμογές. Δηλαδή, συνήθως, έχουμε ένα τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(x)$, η οποία είναι άγνωστη, και θέλουμε να ελέγξουμε τη μηδενική υπόθεση $H_0 : F(x) = F_0(x; \lambda)$, για κάθε $x \in \mathbb{R}$ και για κάποιο $\lambda > 0$. Επομένως, σε αυτήν την περίπτωση, οι πιθανότητες η τ.μ. να ανήκει στην i-οστή ομάδα I_i, $P(X \in I_i | H_0) = p_{i0}(\lambda)$ δεν είναι άμεσα υπολογίσιμες, καθώς είναι συνάρτηση της άγνωστης παραμέτρου λ. Σε αυτήν τη συνήθη περίπτωση, εκτιμούμε την άγνωστη παράμετρο από την εκτιμήτρια στατιστική συνάρτηση έστω $\hat{\lambda}$. Επομένως, για τον έλεγχο της σύνθετης υπόθεσης θα χρησιμοποιήσουμε τη στατιστική συνάρτηση:

$$X^2(\hat{\lambda}) = \sum_{i=1}^{k} \frac{(n_i - np_{i0}(\hat{\lambda}))^2}{np_{i0}(\hat{\lambda})}. \tag{2.4}$$

Ο Fisher (1924) ήταν ο πρώτος που παρατήρησε ότι, στη γενική περίπτωση μίας άγνωστης, υπό τη μηδενική υπόθεση, α.σ.κ. η ασυμπτωτική κατανομή της στατιστικής συνάρτησης δεν είναι απαραίτητο να ακολουθεί χι-τετράγωνο κατανομή με $k-1$ βαθμούς ελευθερίας και ότι διαφορετικές μέθοδοι εκτίμησης των άγνωστων παραμέτρων, αντανακλούν στις ιδιότητες της δειγματικής κατανομής του $X^2(\hat{\lambda})$. Επίσης, επιχειρηματολόγησε ότι η κατάλληλη μέθοδος εκτίμησης είναι η εκτίμηση με τη μέθοδο της μέγιστης πιθανοφάνειας που στηρίζεται στον παρατηρούμενο αριθμό n_i κάθε ομάδας. Τότε προκύπτει ότι για την εύρεση του εκτιμητή έχουμε να επιλύσουμε το σύστημα των εξισώσεων (βλέπε Moore (1986)):

$$\sum_{i=1}^{k} \frac{n_i}{p_{i0}(\lambda)} \frac{\partial p_{i0}(\lambda)}{\partial \lambda} = 0. \tag{2.5}$$

Για την ειδική περίπτωση της κατανομής Poisson ισχύει ότι ο εκτιμητής μέγιστης πιθανοφάνειας, χρησιμοποιώντας τον παρατηρούμενο αριθμό των ομάδων, είναι το ίδιο αποδοτικός με τον κλασικό εκτιμητή μέγιστης πιθανοφάνειας. Επιπρόσθετα, παρατήρησε ότι ένας ασυμπτωτικά ισοδύναμος εκτιμητής του παραπάνω προκύπτει προσδιορίζοντας τις τιμές των παραμέτρων έτσι ώστε το X^2 της σχέσης (2.4) να γίνεται όσον το δυνατό μικρότερο. Αυτή είναι η γνωστή στη στατιστική βιβλιογραφία ως ελάχιστη χι-τετράγωνο μέθοδος εκτίμησης (minimum chi-square method of estimation). Η οριακή κατανομή του $X^2(\hat{\lambda})$ για αυτήν τη μέθοδο εκτίμησης προσδιορίστηκε από τους Fisher (1924) και Neyman and Pearson (1928). Ειδικότερα, αν το μέγεθος του δείγματος n είναι μεγάλο και η μηδενική υπόθεση είναι αληθής, ασυμπτωτικά κατανέμεται ως χ^2_{k-2}. Έτσι, απορρίπτουμε τη μηδενική υπόθεση, σε επίπεδο σημαντικότητας α, αν και μόνο αν $X^2 \geqslant \chi^2_{k-2,a}$. Επισημαίνεται ότι η εκτίμηση με τη μέθοδο της μέγιστης πιθανοφάνειας είναι στην περίπτωση της κατανομής Poisson ασυμπτωτικά ισοδύναμη με αυτήν της μεθόδου εκτίμησης X^2 και για τον λόγο αυτόν τα παραπάνω αποτελέσματα συνεχίζουν να ισχύουν ακόμα και αν χρησιμοποιηθεί ο εκτιμητής μέγιστης πιθανοφάνειας, της παραμέτρου λ.

2.1.1 Τροποποιήσεις και γενικεύσεις

Στη σχετική βιβλιογραφία έχουν εμφανιστεί διάφορες στατιστικές μεθοδολογίες που αποτελούν τροποποιήσεις, επεκτάσεις και γενικεύσεις του χι-τετράγωνο ελέγχου καλής προσαρμογής. Στην ενότητα αυτή, θα παρουσιαστούν κάποιες μεθοδολογίες που ανήκουν σε αυτήν την κατηγορία, χωρίς όμως η παράθεση να είναι εξαντλητική και λεπτομερής.

Ειδικότερα, κάποιες φορές, χρησιμοποιείται ως εναλλακτικό στον χι-τετράγωνο έλεγχο καλής προσαρμογής το τεστ πηλίκου πιθανοφανειών που εισήχθει στη βιβλιογραφία από τους Neyman and Pearson (1928). Το τεστ πηλίκου πιθανοφανειών, όταν χρησιμοποιείται στο πλαίσιο του ελέγχου καλής προσαρμογής, είναι ουσιαστικά ο λόγος της συνάρτησης πιθανοφάνειας όταν αυτή μεγιστοποιείται υπό την υπόθεση ότι τα δεδομένα προέρχονται από την υπό έλεγχο κατανομή προς το μέγιστο της συνάρτησης πιθανοφάνειας χρησιμοποιώντας τα παρατηρούμενα δεδομένα. Κατά αυτόν τον τρόπο προκύπτει ότι ο φυσικός λογάριθμος αυτού του πηλίκου πολλαπλασιασμένος με το -2 οδηγεί στην ακόλουθη στατιστική συνάρτηση:

$$X^2_{LR} = 2\sum_{i=1}^{k} n_i \ln\left(\frac{n_i}{e_i}\right). \tag{2.6}$$

Η ασυμπτωτική κατανομή της X^2_{LR}, υπό τη μηδενική υπόθεση ότι τα δεδομένα προέρχονται από κατανομή Poisson με άγνωστη παράμετρο λ, είναι χι-τετράγωνο με $k-2$ βαθμούς ελευθερίας. Επομένως, απορρίπτεται η μηδενική υπόθεση αν η παρατηρούμενη τιμή της στατιστικής συνάρτησης είναι μεγαλύτερη από την τιμή $\chi^2_{k-2,a}$ (βλέπε Titterington et al. (1985)).

Παρατήρηση 2.1.3. *Με μελέτες προσομοίωσης (Young and Young (1998), Larntz (1978)) έχει εξαχθεί το συμπέρασμα ότι για μεγέθη δείγματος μικρότερα από 100, ο στατιστικός έλεγχος καλής προσαρμογής με το X^2_{LR} τείνει να έχει εμπειρική πιθανότητα σφάλματος τύπου I μεγαλύτερη από την αντίστοιχη του κλασικού χι-τετράγωνο ελέγχου καλής προσαρμογής, ειδικά για μεγάλες τιμές της παραμέτρου λ. Αυτό έχει ως συνέπεια να μην προτιμάται ο έλεγχος αυτός συγκριτικά με τον χι-τετράγωνο. Για τον λόγο αυτό ο έλεγχος αυτός δεν θα μας απασχολήσει στα υπόλοιπα κεφάλαια αυτής της μεταπτυχιακής διατριβής.*

Όπως αναφέρθηκε στην προηγούμενη ενότητα, ένα πρόβλημα που προκύπτει κατά την εφαρμογή του χι-τετράγωνο ελέγχου καλής προσαρμογής είναι εκείνο της ύπαρξης διαστημάτων με μικρές αναμενόμενες συχνότητες. Για την αντιμετώπιση αυτού του προβλήματος, ο Nass (1959) πρότεινε τη σ.σ.:

$$N = \frac{\sum_{i=0}^{m} \left(\frac{n_i^2}{e_i} \right) - n - (m-1)}{\sqrt{\frac{m-1}{m} \left[2m - \frac{(m+1)^2 + 2m}{n} + \sum_{i=0}^{m} \frac{1}{e_i} \right]}} \tag{2.7}$$

με m να είναι η μέγιστη παρατηρούμενη συχνότητα και n το μέγεθος του δείγματος. Αποδεικνύεται ότι η σ.σ. N ακολουθεί, υπό τη μηδενική υπόθεση, ασυμπτωτικά τυπική κανονική κατανομή. Προφανώς, λόγω του τρόπου κατασκευής, η μηδενική υπόθεση απορρίπτεται για μεγάλες τιμές της σ.σ.ε. N, δηλαδή σε ασυμπτωτικό επίπεδο σημαντικότητας α απορρίπτεται όταν $N \geq z_\alpha$.

Παρατήρηση 2.1.4. *Ένα μειονέκτημα της στατιστικής συνάρτησης N είναι ότι η αύξηση της διακύμανσης καθιστά τον έλεγχο συντηρητικό και αυτό διότι προκύπτουν μεγάλες αποκλίσεις από την υψηλή μεταβλητότητα των δεδομένων. Τέλος, μέσω προσομοίωσης φαίνεται να μην είναι ικανοποιητική η κανονική προσέγγιση και αυτός είναι ο λόγος που ο έλεγχος αυτός δεν θα ληφθεί υπόψιν περαιτέρω.*

Ένας άλλος έλεγχος που ανήκει σε αυτήν την κατηγορία είναι αυτός των Freeman and Tukey (1950), ο οποίος βασίζεται στη στατιστική συνάρτηση

$$\mathrm{FT}^2 = 4 \sum_{i=1}^{k} (\sqrt{n_i} - \sqrt{e_i})^2. \tag{2.8}$$

Επισημαίνεται ότι η παρακάτω σ.σ. προκύπτει ως ειδική περίπτωση μίας μεγαλύτερης οικογένειας σ.σ. που προτάθηκαν από τους Read and Cressie (1988) και δίνεται από τη σχέση:

$$I(a) = \frac{2}{a(a+1)} \sum_{i=1}^{k} e_i \left[\left(\frac{n_i}{e_i} \right)^a - 1 \right], \tag{2.9}$$

όπου a παράμετρος που καθορίζει ποιο μέλος αυτής της οικογένειας στατιστικών συναρτήσεων χρησιμοποιείται. Ειδικότερα, για $a = 1$ το τεστ είναι ισοδύναμο με το χι-τετράγωνο του Pearson. για $a \to 0$ το στατιστικό τεστ τείνει στο τεστ πηλίκου πιθανοφανειών, ενώ για $a = -0.5$ προκύπτει το τεστ των Freeman-Tukey. Οι Cressie and Read (1984) μελέτησαν τις ιδιότητες αυτής της οικογένειας στατιστικών συναρτήσεων και συμπέραναν ότι υπό τη μηδενική υπόθεση κάθε μέλος της οικογένειας έχει την ίδια ασυμπτωτική κατανομή. Επίσης συνέστησαν τη χρήση του $a = \frac{2}{3}$ διότι για αυτήν την επιλογή η προσέγγιση χι-τετράγωνο υπό τη μηδενική υπόθεση είναι καλύτερη.

Τέλος, μία γενίκευση του παραπάνω ελέγχου προκύπτει με χρήση της σ.σ. (βλέπε Κεφάλαιο 6 της μονογραφίας του Pardo (2006) και τις εκεί αναφορές):

$$T_n^{\phi_1}(\hat{\lambda}_{\phi_2}) = \frac{2n}{\phi_1''(1)} D_{\phi_1}\left(\hat{p}, p(\hat{\lambda}_{\phi_2})\right),$$

όπου $\hat{p} = \left(\frac{n_1}{n}, ..., \frac{n_k}{n}\right)$,

$$\hat{\lambda}_\phi = \arg inf_{\lambda>0} D_\phi(\hat{p}, p(\lambda)),$$

με

$$D_\phi(\hat{p}, p(\lambda)) = \sum_{i=1}^{k} \hat{p}_i(\lambda) \phi\left(\frac{\hat{p}}{\hat{p}_i(\lambda)}\right),$$

ενώ ϕ_1 και ϕ_2 είναι δύο φορές συνεχώς παραγωγίσιμες για $x > 0$ με δεύτερες παραγώγους τέτοιες ώστε $\phi_1''(1) \neq 0$, $\phi_2''(1) \neq 0$ και επιπλέον ανήκουν στο σύνολο όλων των κυρτών συναρτήσεων $\Phi^* = \{\phi(x), x \geq 0\}$ που είναι τέτοιες ώστε $\phi(1) = 0$, $0\phi(0/0) = 0$ και $0\phi(p/0) = \lim_{u\to\infty}\phi(u)/u$. Η παραπάνω στατιστική συνάρτηση ακολουθεί ασυμπτωτικά χι-τετράγωνο κατανομή με $k-2$ βαθμούς ελευθερίας.

Επισημαίνεται ότι όλοι οι έλεγχοι που παρουσιάστηκαν σε αυτήν την ενότητα προϋποθέτουν τη διαμέριση των δεδομένων, ενώ και η επιλογή του αριθμού των ομάδων χρήζει ιδιαίτερης προσοχής, ειδικά όταν δεν υποδεικνύεται από τα ίδια τα δεδομένα. Το παραπάνω γεγονός θεωρείται αδυναμία αυτών των ελέγχων και οδηγεί να μην χρησιμοποιούνται πολύ συχνά σε πρακτικές εφαρμογές και σε μελέτες προσομοίωσης. Αυτός είναι και ο λόγος που στην παρούσα μεταπτυχιακή διατριβή δεν θα γίνει περαιτέρω αναφορά αυτών των στατιστικών τεστ και δεν θα χρησιμοποιηθούν στη συγκριτική μελέτη του επόμενου κεφαλαίου.

2.2 Έλεγχοι με την ε.α.σ.κ.

Στην ενότητα αυτήν παρουσιάζονται έλεγχοι καλής προσαρμογής της Poisson κατανομής που στηρίζονται σε μέτρα εγγύτητας μεταξύ μίας εκτιμήτριας της άγνωστης αθροιστικής συνάρτησης κατανομής, που είναι η εμπειρική αθροιστική συνάρτηση κατανομής, και της (εκτιμώμενης) αθροιστικής συνάρτησης κατανομής υπό τη μηδενική υπόθεση. Η αιτιολόγηση της εισαγωγής στη βιβλιογραφία τέτοιων ελέγχων προκύπτει εύκολα ανακαλώντας το Θεώρημα 1.2.5, σύμφωνα με το οποίο η διαφορά της ε.α.σ.κ. από την πραγματική α.σ.κ. γίνεται μικρότερη, καθώς αυξάνει το μέγεθος του δείγματος. Στη συνέχεια αυτής της ενότητας το ενδιαφέρον επικεντρώνεται στην περίπτωση του σύνθετου ελέγχου καλής προσαρμογής.

2.2.1 Kolmogorov-Smirnov

Μεταξύ των ελέγχων καλής προσαρμογής που έχουν παρουσιασθεί στη βιβλιογραφία και βασίζονται στην εμπειρική αθροιστική συνάρτηση κατανομής, ο πιο δημοφιλής είναι ο λεγόμενος έλεγχος των Kolmogorov-Smirnov (K-S test) ή, απλώς, έλεγχος του Kolmogorov. Ο έλεγχος των Kolmogorov-Smirnov προτάθηκε αρχικά για συνεχείς κατανομές, αλλά αργότερα επεκτάθηκε και σε διακριτές κατανομές (βλέπε Conover (1972)). Ειδικότερα, στη γενική περίπτωση, ο Kolmogorov (1933) πρότεινε να χρησιμοποιείται η στατιστική συνάρτηση

$$D_n = \sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)|\,, \tag{2.10}$$

η οποία, στην ουσία, μετρά πόσο αποκλίνει η ε.α.σ.κ. $F_n(\cdot)$ από την α.σ.κ. $F_0(\cdot)$ και αναζητά τη μέγιστη κατακόρυφη απόσταση μεταξύ των γραφημάτων των $F_n(x)$ και $F_0(x)$.

Για τον σύνθετο έλεγχο καλής προσαρμογής της κατανομής Poisson, προτείνεται η στατιστική συνάρτηση που προκύπτει με αντικατάσταση της άγνωστης παραμέτρου λ με την εκτιμήτριά της, έστω $\hat{\lambda}$, δηλαδή προτείνεται η σ.σ.:

$$D_n\left(\hat{\lambda}\right) = \sup_{x\geq 0}\left|F_n(x) - F_0(x,\hat{\lambda})\right|. \tag{2.11}$$

Αποδεικνύεται (βλέπε για παράδειγμα Henze (1996)) ότι η στατιστική συνάρτηση $D_n\left(\hat{\lambda}\right)$ δεν έχει την κατανομή του D_n, κλειστή μορφή για την κατανομή της στατιστικής συνάρτησης $D_n\left(\hat{\lambda}\right)$ είναι άγνωστη, ενώ η μέθοδος εκτίμησης που χρησιμοποιείται (μέγιστης πιθανοφάνειας, ροπών ή κάποια άλλη) επιδρά στην απόδοση του ελέγχου (βλέπε Weber et al. (2006)).

Για την ειδική περίπτωση της κατανομής Poisson, ο Henze (1996) απέδειξε ότι η κρίσιμη τιμή του ελέγχου εξαρτάται από το επιθυμητό επίπεδο σημαντικότητας, το μέγεθος του διαθέσιμου δείγματος, αλλά και από την άγνωστη παράμετρο λ. Ένας τρόπος για να ξεπεραστεί το παραπάνω πρόβλημα είναι να χρησιμοποιηθούν τεχνικές Monte Carlo για την εκτίμηση της p-τιμής του ελέγχου μέσω παραμετρικού ελέγχου bootstrap. Ο έλεγχος που προκύπτει κατά αυτόν τον τρόπο έχει ασυμπτωτικό επίπεδο σημαντικότητας α, καθώς τα n και B (μέγεθος δείγματος και πλήθος bootstrap δειγμάτων, αντίστοιχα) τείνουν στο άπειρο (Henze (1996)).

Παρατήρηση 2.2.1. *Ένας άλλος τρόπος προσδιορισμού των κρίσιμων τιμών για την ειδική περίπτωση της Poisson προκύπτει διεξάγοντας μελέτη προσομοίωσης και αποκτώντας κατάλληλους πίνακες. Τέτοιοι πίνακες παρατίθενται, για παράδειγμα, στην εργασία των Campbell and Oprian (1979) για διάφορα μεγέθη δείγματος. Επισημαίνεται ότι αυτοί οι πίνακες των κρίσιμων τιμών υποθέτουν ότι η άγνωστη παράμετρος αντικαθίσταται από τη δειγματική μέση τιμή, άρα ότι χρησιμοποιείται η εκτίμηση μέγιστης πιθανοφάνειας. Τέλος, επισημαίνεται ότι οι πίνακες των κρίσιμων τιμών διαφοροποιούνται ανάλογα με την τιμή της δειγματικής μέσης τιμής. Πιο συγκεκριμένα, οι κρίσιμες τιμές διαφοροποιούνται για τιμές της δειγματικής τιμής στα ακόλουθα μη επικαλυπτόμενα διαστήματα* $(0,1]$, $(1,2]$, $(2,3]$, $(3,5]$ *και* $(5,10]$.

Πριν ολοκληρωθεί αυτή η υποενότητα, αξίζει να επισημανθεί ότι για τον υπολογισμό της σ.σ. $D_n\left(\hat{\lambda}\right)$, αν $M = X_{(n)}$ και $k \in \{0, 1, ..., M\}$, έχουμε ότι (βλέπε Henze (1996)):

$$D_n\left(\hat{\lambda}\right) = \max_{0 \le k \le M} \left| F_n(k) - F_0(k; \hat{\lambda}) \right|. \tag{2.12}$$

Παρατήρηση 2.2.2. *Χρησιμοποιώντας την ιδιότητα που δόθηκε στη σχέση*

(1.23) ότι η δεσμευμένη κατανομή των $(X_1, X_2, ..., X_n)$ *δοθέντος της επαρκούς στατιστικής συνάρτησης* $T_n = \sum_{i=1}^{n} X_i = t$ *είναι πολυωνυμική* $M(t, 1/n, ..., 1/n)$*, ο Frey (2012) πρότεινε έναν τροποποιημένο Kolmogorov-Smirnov τύπου έλεγχο. Ειδικότερα, η προτεινόμενη στατιστική συνάρτηση βασίζεται στη μεγαλύτερη απόσταση μεταξύ της εμπειρικής αθροιστικής συνάρτησης κατανομής και της αναμενόμενης τιμής της εμπειρικής αθροιστικής συνάρτησης κατανομής, δεδομένης της τιμής του επαρκούς στατιστικού* T *και δίνεται από τη σχέση:*

$$D = \sup_{x \in \mathcal{R}} |F_n(x) - E[F_n(x)|T]|.$$

Μετά από πράξεις προκύπτει ότι η σ.σ. δίνεται από τη σχέση:

$$D = \max_{x \in \{0,...,t\}} \left| F_n(x) - \sum_{i=0}^{x} \binom{t}{i} \left(\frac{1}{n}\right)^i \left(1 - \frac{1}{n}\right)^{t-i} \right|. \tag{2.13}$$

Είναι προφανές ότι η σ.σ. D *απορρίπτει τη μηδενική υπόθεση για μεγάλες τιμές και ο υπολογισμός της p-τιμής του ελέγχου ανάγεται στον υπολογισμό της πιθανότητας* $P(D \geq v|T = t)$*, η οποία μπορεί να προσδιοριστεί με βάση έναν αναλυτικό αλγόριθμο που έχει δοθεί από τους Frey (2012). Ωστόσο, καθώς με μελέτες προσομοίωσης που έχουν πραγματοποιηθεί τόσο από άλλους συγγραφείς όσο και από τους ίδιους, η απόδοση του ελέγχου δεν είναι ικανοποιητική, δεν παρατίθενται περισσότερες λεπτομέρειες σχετικές με τον αλγόριθμο υπολογισμού των p-τιμών ή των κρίσιμων τιμών του ελέγχου. Επιπρόσθετα, για τους παραπάνω λόγους, ο έλεγχος αυτός δεν θα συμπεριληφθεί στην συγκριτική μελέτη του επόμενου κεφαλαίου.*

2.2.2 Cramér-von Mises, Watson, Anderson-Darling και Klar

Εκτός από το Kolmogorov-Smirnov τεστ υπάρχει ένας μεγάλος αριθμός από ελέγχους καλής προσαρμογής που στηρίζονται στην εμπειρική αθροιστική συνάρτηση κατανομής. Στην ενότητα αυτή, θα παρουσιαστούν, πολύ σύντομα, κάποιοι από τους πιο δημοφιλείς τέτοιους ελέγχους. Ωστόσο, αρχικά, για λόγους πληρότητας, θα παραθέσουμε τις σ.σ. που προτάθηκαν για τη γενική περίπτωση του απλού ελέγχου καλής προσαρμογής, όταν η α.σ.κ. είναι συνεχής.

Ο Cramér (1928) και ο von Mises (1928), ανεξάρτητα, πρότειναν για τη γενική περίπτωση του απλού ελέγχου καλής προσαρμογής τη στατιστική συνάρτηση:

$$W_n^2 = n \int_{-\infty}^{+\infty} \{F_n(x) - F_0(x)\}^2 dF_0(x),$$

ενώ ο Watson (1961) πρότεινε μία τροποποίηση του ελέγχου των Cramér-von Mises για τη γενική περίπτωση του απλού ελέγχου καλής προσαρμογής θεωρώντας τη στατιστική συνάρτηση:

$$U_n^2 = n \int_{-\infty}^{+\infty} \left\{F_n(x) - F_0(x) - \int_{-\infty}^{+\infty} (F_n(x) - F_0(x)) dF_0(x)\right\}^2 dF_0(x).$$

Τέλος, οι Anderson and Darling (1952) τροποποίησαν τον έλεγχο των Cramér-von Mises, έτσι ώστε να δίνει μεγαλύτερη προσοχή και βαρύτητα στις ουρές της κατανομής και πρότειναν τη στατιστική συνάρτηση:

$$A_n^2 = n \int_{-\infty}^{+\infty} \{F_n(x) - F_0(x)\}^2 \ \psi(F_0(x)) dF_0(x),$$

όπου $\psi(t) \geqslant 0$, $0 \leqslant t \leqslant 1$, είναι μια προκαθορισμένη συνάρτηση βάρους που επιλέγεται, έτσι ώστε να εξετάζονται συγκεκριμένες περιοχές της κατανομής. Για επιπλέον ιδιότητες που πρέπει να πληροί η συνάρτηση βάρους παραπέμπουμε στους

Anderson and Darling (1952). Στην ειδική περίπτωση όπου $\psi(t) = 1$, $0 \leqslant t \leqslant 1$, προκύπτει η στατιστική συνάρτηση των Cramér-von Mises, ενώ οι Anderson and Darling (1952) θεώρησαν την ειδική περίπτωση όπου $\psi(t) = \dfrac{1}{t(1-t)}$, για $0 \leqslant t \leqslant 1$, η οποία οδηγεί σε ένα τεστ που είναι ευαίσθητο στις ουρές της κατανομής. Είναι προφανές από τον τρόπο ορισμού των σ.σ. W_n^2, U_n^2 και A_n^2 ότι απορρίπτεται η μηδενική υπόθεση για μεγάλες τιμές της.

Στην περίπτωση του σύνθετου ελέγχου καλής προσαρμογής της κατανομής Poisson οι στατιστικές συναρτήσεις που προκύπτουν από τις παραπάνω με αντικατάσταση της άγνωστης παραμέτρου από την εκτιμήτρια $\hat{\lambda}$, αλλά και τροποποιημένες εκδοχές αυτών, έχουν παρουσιαστεί και μελετηθεί στις εργασίες των Henze (1996) και Spinelli and Stephens (1997). Ειδικότερα, ο Henze (1996) θεώρησε τη στατιστική συνάρτηση των Cramér-von Mises που ορίζεται από τη σχέση:

$$C_n = n \sum_{k=0}^{\infty} \{F_n(k) - F_0(k, \widehat{\lambda})\}^2 \{F(k, \widehat{\lambda}) - F(k-1, \widehat{\lambda})\}, \tag{2.14}$$

καθώς και την ακόλουθη τροποποίησή του

$$C_n^* = n \sum_{k=0}^{\infty} \{F_n(k) - F_0(k, \widehat{\lambda})\}^2 \{F_n(k) - F_n(k-1)\}. \tag{2.15}$$

Όπως επισημαίνει ο Henze (1996) η σ.σ. C_n^* περιέχει το άθροισμα το πολύ $M+1$ το πλήθος μη μηδενικών όρων, καθώς οι όροι για $k = M+2$ και μεγαλύτερες τιμές του k είναι πάντοτε μηδέν. Τέλος, είναι προφανές ότι οι στατιστικές συναρτήσεις C_n^* και C_n, έχουν την ίδια οριακή κατανομή, υπό τη μηδενική υπόθεση. Ειδικότερα, ο Henze (1996) απέδειξε ότι η ασυμπτωτική κατανομή των δύο σ.σ. υπό τη μηδενική υπόθεση είναι η ίδια, με τις κρίσιμές τιμές τους, οι οποίες ταυτίζονται, να εξαρτώνται από το επιθυμητό επίπεδο σημαντικότητας, το μέγεθος του διαθέσιμου δείγματος, αλλά και από την άγνωστη παράμετρο λ. Ένας τρόπος για να ξεπεραστεί το παραπάνω πρόβλημα είναι να χρησιμοποιηθούν και πάλι τεχνικές Monte Carlo για την εκτίμηση της p-τιμής του ελέγχου μέσω παραμετρικού ελέγχου bootstrap. Ο έλεγχος που προκύπτει κατά αυτόν τον τρόπο έχει ασυμπτωτικό επίπεδο σημαντικότητας α καθώς τα n και B (μέγεθος δείγματος και πλήθος bootstrap δειγμάτων, αντίστοιχα) τείνουν στο άπειρο.

Παρατήρηση 2.2.3. *Για τον υπολογισμό της σ.σ. C_n στην πράξη, αν $M = X_{(n)}$, έχουμε ότι (βλέπε Henze (1996)):*

$$C_n = n \sum_{k=0}^{l} \{F_n(k) - F_0(k, \widehat{\lambda})\}^2 f(k, \widehat{\lambda}) + R,$$

όπου, για $l \geq M$, με R συμβολίζονται οι όροι εκείνοι που ικανοποιούν τη σχέση: $0 \leq R \leq n\{1 - F_0(l, \widehat{\lambda})\}^3$. Η επιλογή του l που συνήθως συνίσταται και η οποία παρέχει επαρκή αριθμητική ακρίβεια, ορίζεται να είναι η ακόλουθη

$$l := min\{m \geq M : n\{1 - F_0(m, \widehat{\lambda})\}^3 \leq 10^{-4}\}.$$

Από την άλλη μεριά, οι Spinelli and Stephens (1997) πρότειναν και μελέτησαν τρεις διαφορετικές εκδοχές στατιστικών συναρτήσεων που ανήκουν σε αυτήν την κατηγορία. Σε όσα ακολουθούν με $\hat{p}_k$ συμβολίζεται η εκτιμώμενη πιθανότητα που δίνεται από τη σχέση:

$$\hat{p}_k = P(X = k | X \sim \mathcal{P}(\hat{\lambda})),$$

ενώ με $o_k = n\{F_n(k) - F_n(k-1)\}$ και με $e_k = n\hat{p}_k$ συμβολίζεται ο παρατηρούμενος και αναμενόμενος αριθμός παρατηρήσεων, αντίστοιχα, με τιμή ίση με k, για $k = 0, 1, ...,$. Τότε οι Spinelli and Stephens (1997) ορίζουν τις σ.σ. που δίνονται από τι σχέσεις:

$$W^2 = n^{-1} \sum_{k=0}^{\infty} Z_k^2 \hat{p}_k, \tag{2.16}$$

$$A^2 = n^{-1} \sum_{k=0}^{\infty} \frac{Z_k^2 \hat{p}_k}{H_k(1-H_k)}, \tag{2.17}$$

και

$$W_m^2 = n^{-1} \sum_{k=0}^{\infty} Z_k^2, \tag{2.18}$$

όπου

$$Z_k = \sum_{i=0}^{k} (o_i - e_i) \text{ και } H_k = \sum_{i=0}^{k} \hat{p}_i.$$

Παρατηρούμε ότι εξ' ορισμού η σ.σ. Z_k παριστάνει τη διαφορά του αναμενόμενου αριθμού των παρατηρήσεων με τιμή ίση ή μικρότερη από k ($\sum_{i=0}^{k} e_i$) από τον αντίστοιχο παρατηρούμενο αριθμό παρατηρήσεων ($\sum_{i=0}^{k} o_i$). Επομένως, από τον ορισμό της ε.α.σ.κ. και της α.σ.κ., έχουμε ότι:

$$Z_k = nF_n(k) - nF_0(k, \hat{\lambda}) = n \left\{ F_n(k) - F_0(k, \hat{\lambda}) \right\}.$$

Με παρόμοιο σκεπτικό προκύπτει ότι $H_k = F_0(k, \hat{\lambda})$. Επομένως, λαμβάνοντας επιπλέον υπόψη ότι:

$$\hat{p}_k = P(X = k | X \sim \mathcal{P}(\hat{\lambda})) = F_0(k, \hat{\lambda}) - F_0(k-1, \hat{\lambda}),$$

οι εκφράσεις των σ.σ. W^2, A^2 και W_m^2 ισοδύναμα γράφονται

$$\begin{aligned} W^2 &= n \sum_{k=0}^{\infty} \left\{ F_n(k) - F_0(k, \hat{\lambda}) \right\}^2 \left\{ F_0(k, \hat{\lambda}) - F_0(k-1, \hat{\lambda}) \right\} \\ &= n \sum_{k=0}^{\infty} \left\{ F_n(k) - F_0(k, \hat{\lambda}) \right\}^2 f_0(k, \hat{\lambda}), \end{aligned}$$

$$\begin{aligned} A^2 &= n \sum_{k=0}^{\infty} \frac{\left\{F_n(k) - F_0(k,\hat{\lambda})\right\}^2}{F_0(k,\hat{\lambda})\left(1 - F_0(k,\hat{\lambda})\right)} \left\{F_0(k,\hat{\lambda}) - F_0(k-1,\hat{\lambda})\right\} \\ &= n \sum_{k=0}^{\infty} \frac{\left\{F_n(k) - F_0(k,\hat{\lambda})\right\}^2}{F_0(k,\hat{\lambda})\left(1 - F_0(k,\hat{\lambda})\right)} f_0(k,\hat{\lambda}), \end{aligned}$$

και

$$W_m^2 = n \sum_{k=0}^{\infty} \left\{F_n(k) - F_0(k,\hat{\lambda})\right\}^2 \left\{F_0(k,\hat{\lambda}) - F_0(k-1,\hat{\lambda})\right\},$$

αντίστοιχα. Παρατηρήστε ότι η σ.σ. W^2 ταυτίζεται με τη σ.σ. C_n που δόθηκε στη σχέση (2.14), ενώ σε όρους του παραπάνω συμβολισμού η σ.σ. C_n^* γράφεται στη μορφή $C_n^* = n^{-2} \sum_{k=0}^{\infty} Z_k^2 o_k$.

Παρατήρηση 2.2.4. *Οι παραπάνω στατιστικές συναρτήσεις ουσιαστικά διαφοροποιούνται με την εισαγωγή διαφορετικής στάθμισης στα τετράγωνα των διαφορών $F_n(k) - F_0(k,\hat{\lambda})$. Στην πραγματικότητα για να μπορούν να υπολογιστούν, μιας και ο υπολογισμός ενός άπειρου αθροίσματος στην πράξη δεν είναι εφικτός, οι αθροίσεις τους σταματούν σε μία τιμή M που ορίζεται ως το ανώτερο όριο άθροισης του πεπερασμένου αθροίσματος. Διάφοροι τρόποι επιλογής της τιμής M έχουν προταθεί στη βιβλιογραφία. Για παράδειγμα, οι Karlis and Xekalaki (2000) πρότειναν να επιλέγεται η τιμή M κατά τέτοιον τρόπο ώστε η πιθανότητα παρατήρησής της υπό την κατανομή Poisson με παράμετρο $\hat{\lambda}$ να μην υπερβαίνει την τιμή 10^{-4}. Από την άλλη μεριά, οι Mijburgh and Visagie (2020) πρότειναν τη χρήση της τιμής $M = 100$, που ισχυρίζονται ότι, σε σχέση με αυτήν που προτάθηκε από τους Karlis and Xekalaki (2000) οδηγεί στο να συμπεριλαμβάνονται περισσότεροι όροι.*

Άλλος ένας στατιστικός έλεγχος που στηρίζεται στην εμπειρική αθροιστική συνάρτηση κατανομής $F_n(\cdot)$, είναι αυτός που προτάθηκε από τον Klar (1999). Η σ.σ. του ελέγχου δίνεται από τη σχέση:

$$\tilde{T}_n = \sum_{k\geq 0} |Z_{n,k}|, \tag{2.19}$$

όπου

$$Z_{n,k} = \sqrt{n}(F_n(k) - F_0(k, \hat{\lambda}))$$

Όπως επισημαίνεται από τον Klar (1999), πλην του όρου $\sqrt{n}$ η παραπάνω σ.σ. δεν είναι παρά ειδική περίπτωση της μετρικής Mallow μεταξύ της ε.α.σ.κ. και της α.σ.κ. για $r = 1$. Είναι προφανές ότι η μηδενική υπόθεση απορρίπτεται για μεγάλες τιμές της ελεγχοσυνάρτησης. Παρότι υπάρχουν διαθέσιμα θεωρητικά αποτελέσματα στην εργασία του Klar (1999), που αφορούν την ασυμπτωτική κατανομή αυτής της σ.σ., αυτή είναι πολύπλοκη και εξαρτάται από την άγνωστη παράμετρο λ. Για τον λόγο αυτό χρησιμοποιείται παραμετρικό bootstrap για την υλοποίηση του ελέγχου. Ο έλεγχος που προκύπτει κατά αυτόν τον τρόπο έχει, όπως απέδειξε ο Klar (1999), ασυμπτωτικό επίπεδο σημαντικότητας a καθώς τα n και B (μέγεθος δείγματος και πλήθος bootstrap δειγμάτων, αντίστοιχα) τείνουν στο άπειρο.

Επισημαίνεται ότι στην πράξη, για τον υπολογισμό της σ.σ. περιορίζουμε το άπειρο άθροισμα σε πεπερασμένο και χρησιμοποιείται η σχέση:

$$\tilde{T}_n = \sum_{k=0}^{X_{(n)}} |Z_{n,k}| + \sqrt{n} - \sqrt{n}\bar{X} \sum_{k=0}^{X_{(n)}} (1 - F_0(k, \widehat{\lambda})).$$

Παρατήρηση 2.2.5. *Οι Székely and Rizzo (2004) πρότειναν έναν Cramér-von Mises τύπου έλεγχο με τη διαφοροποίηση ότι για την εκτίμηση της άγνωστης* *αθροιστικής συνάρτησης κατανομής δεν χρησιμοποιούν την εμπειρική αθροιστική συνάρτηση κατανομής, αλλά την εκτίμηση που προκύπτει από τις ιδιότητες της μέσης απόστασης (βλέπε Ενότητα 1.2.6). Ειδικότερα, χρησιμοποιώντας τις σχέσεις (1.25) και (1.26) προτείνουν την ακόλουθη εκτίμηση:*

$$\hat{f}(0) = \hat{F}(0) = \frac{\hat{m}_1 + 1 - \hat{\lambda}}{2}, \tag{2.20}$$

και

$$\hat{f}(k) = \frac{\hat{m}_{k+1} - (k+1-\hat{\lambda})(2\hat{F}(k-1) - 1)}{2(k+1)}, \text{ για } k = 1, 2... \tag{2.21}$$

με

$$\hat{\lambda} = \bar{X} \text{ και } \hat{m}_k = \frac{1}{n}\sum_{i=1}^{n} |k - X_i|.$$

Στο παραπάνω πλαίσιο, πρότειναν τη στατιστική συνάρτηση που προσδιορίζεται από τη σχέση:

$$SR = n\sum_{j=0}^{\infty}(\hat{F}(j) - F_0(j,\hat{\lambda}))^2 f_0(j,\hat{\lambda}), \tag{2.22}$$

όπου

$$\hat{F}(0) = \frac{\hat{m}_1 + 1 - \hat{\lambda}}{2} \text{ και } \hat{F}(k) = \sum_{j=0}^{k} \hat{f}(j),\ k = 1, 2, ...$$

με τις ποσότητες $\hat{f}(j)$ να έχουν προσδιοριστεί παραπάνω.

Η κατανομή της σ.σ.ε. υπό τη μηδενική υπόθεση δεν είναι απαλλαγμένη της άγνωστης παραμέτρου λ και για αυτό χρησιμοποιείται παραμετρικός έλεγχος bootstrap.

2.3 Έλεγχος με την ε.ο.σ.κ.

Στην προηγούμενη ενότητα παρουσιάστηκαν έλεγχοι που βασίζονται σε μέτρα απόκλισης μεταξύ της εκτιμώμενης α.σ.κ. υπό τη μηδενική υπόθεση, ήτοι της α.σ.κ. της κατανομής Poisson που προκύπτει με αντικατάσταση της άγνωστης παραμέτρου λ από έναν κατάλληλο εκτιμητή της (Εμπειρικής Μέγιστης Πιθανοφάνειας (Ε.Μ.Π.), μέθοδος ροπών) και μιας εκτίμησης της άγνωστης α.σ.κ. με βάση τα διαθέσιμα δεδομένα, με την ε.α.σ.κ. να είναι η συνηθέστερη επιλογή. Ωστόσο, στη βιβλιογραφία έχουν εμφανιστεί διάφορες μεθοδολογίες ελέγχου καλής προσαρμογής της κατανομής Poisson οι οποίες στηρίζονται σε μέτρα απόκλισης που χρησιμοποιούν στον ορισμό τους αντί της α.σ.κ. και του δειγματικού ανάλογού της κάποιον άλλον μετασχηματισμό της τυχαίας μεταβλητής X. Στην ενότητα αυτή, θα παρουσιαστεί ο έλεγχος που προτάθηκε από τον Klar (1999) και ο οποίος βασίζεται σε ένα μέτρο απόκλισης μεταξύ της εκτιμώμενης, υπό τη μηδενική υπόθεση, ολοκληρώσιμης συνάρτησης κατανομής (integrated distribution function) και του δειγματικού ανάλογού της (βλέπε Ενότητα 1.2.3), ενώ έλεγχοι που βασίζονται σε μέτρα απόκλισης μεταξύ της εκτιμώμενης, υπό τη μηδενική υπόθεση, πιθανογεννήτριας συνάρτησης και του δειγματικού ανάλογού της θα αποτελέσουν αντικείμενο μελέτης της επόμενης ενότητας.

Ειδικότερα, βασιζόμενος στο μονοσήμαντο της ολοκληρώσιμης συνάρτησης κατανομής (βλέπε Πρόταση 1.2.6), ο Klar (1999) πρότεινε έναν έλεγχο καλής προσαρμογής τύπου Kolmogorov-Smirnov, με τη διαφοροποίηση ότι η εκτιμώμενη α.σ.κ. υπό τη μηδενική υπόθεση θα αντικατασταθεί από την εκτιμώμενη ολοκληρώσιμη συνάρτηση κατανομής και η εμπειρική συνάρτηση κατανομής από το δειγματικό ανάλογο της ολοκληρώσιμης συνάρτησης κατανομής, ήτοι την εμπειρική ολοκληρώσιμη συνάρτηση κατανομή (βλέπε Ορισμό 1.2.7). Επομένως, η σ.σ.

που προτάθηκε από τον Klar (1999) είναι η

$$I_n = \sup_{t \geq 0} \sqrt{n} |\Psi_n(t) - \Psi(t, \hat{\lambda})|, \tag{2.23}$$

με $\Psi(t, \lambda)$ να είναι η ολοκληρώσιμη συνάρτηση κατανομής υπό την υπόθεση της κατανομής Poisson με παράμετρο λ. Παρατηρήστε ότι από τον ορισμό της ολοκληρώσιμης συνάρτησης κατανομής άμεσα προκύπτει ότι $\Psi(0, \lambda) = E(X) = \lambda$.

Χρησιμοποιώντας το γεγονός ότι οι συναρτήσεις $\Psi_n(t)$ και $\Psi(t, \hat{\lambda})$ είναι γραμμικές στο διάστημα $(k, k+1]$, $k \in \mathbb{N}_0$, ο Klar (1999) οδηγήθηκε στις ακόλουθες ισοδύναμες εκφράσεις

$$I_n = \sup_{k \in \mathbb{N}_0} \sqrt{n} |\Psi_n(k) - \Psi(k, \hat{\lambda})| = \max_{0 \leq k \leq X_{(n)}} \sqrt{n} |\Psi_n(k) - \Psi(k, \hat{\lambda})|.$$

Επιπρόσθετα, καθώς από τη σχέση (1.8) έχουμε άμεσα ότι $\Psi_n(0) = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$, για τον υπολογισμό της σ.σ. I_n στην πράξη μπορεί εναλλακτικά να χρησιμοποιηθεί και η σχέση:

$$I_n = \sqrt{n} \sup_{1 \leq k \leq M} \left| \sum_{i=0}^{k-1} (F_n(i) - F_0(i, \hat{\lambda})) \right|,$$

όπου F_0 η α.σ.κ. της κατανομής Poisson.

Από τον τρόπο ορισμού της σ.σ. I_n είναι προφανές ότι η μηδενική υπόθεσης του σύνθετου ελέγχου καλής προσαρμογής απορρίπτεται για μεγάλες τιμές της σ.σ.ε. Ωστόσο κρίσιμες τιμές του ελέγχου δεν είναι διαθέσιμες, καθώς η ασυμπτωτική κατανομή της σ.σ. είναι, όπως αποδείχθηκε από τον Klar (1999), πολύπλοκη και επιπλέον εξαρτάται από την άγνωστη παράμετρο λ. Για τον λόγο αυτό για την υλοποίηση του ελέγχου χρησιμοποιείται παραμετρικός έλεγχος bootstrap. Ο έλεγχος που προκύπτει κατά αυτόν τον τρόπο έχει, όπως απέδειξε ο Klar (1999), ασυμπτωτικό επίπεδο σημαντικότητας α καθώς τα n και B (μέγεθος δείγματος και πλήθος bootstrap δειγμάτων, αντίστοιχα) τείνουν στο άπειρο.

2.4 Έλεγχοι με την ε.π.σ.

Στις προηγούμενες δύο ενότητες παρουσιάστηκαν έλεγχοι καλής προσαρμογής της κατανομής Poisson που στηρίζονται σε μέτρα εγγύτητας μεταξύ μίας εκτιμήτριας της άγνωστης αθροιστικής συνάρτησης κατανομής, που είναι η εμπειρική αθροιστική συνάρτηση κατανομής, και της εκτιμώμενης αθροιστικής συνάρτησης κατανομής υπό τη μηδενική υπόθεση, καθώς και με παρόμοιο σκεπτικό σε ένα μέτρο εγγύτητας μεταξύ της εκτιμώμενης ολοκληρώσιμης συνάρτησης κατανομής και της εμπειρικής ολοκληρώσιμης συνάρτησης κατανομής. Στη συνέχεια αυτής της ενότητας το ενδιαφέρον επικεντρώνεται στην περίπτωση του σύνθετου ελέγχου καλής προσαρμογής που διεξάγεται με στατιστικές συναρτήσεις που βασίζονται σε μέτρα εγγύτητας, αλλά και σε χαρακτηρισμούς της κατανομής Poisson που εκφράζονται με τη βοήθεια της πιθανογεννήτριας συνάρτησης και της εμπειρικής πιθανογεννήτριας συνάρτησης (βλέπε Ενότητα 1.2.4).

Τεστ των Kocherlakota and Kocherlakota (1986).
Οι Kocherlakota and Kocherlakota (1986) πρότειναν να χρησιμοποιείται για τον έλεγχο καλής προσαρμογής της Poisson η σ.σ. που δίνεται από τη σχέση:

$$K = \sqrt{n}\frac{g_n(t) - g(t, \bar{X})}{\widehat{\sigma}_n},$$

όπου $\widehat{\sigma}_n = e^{\bar{X}(t^2-1)} - e^{2\bar{X}(t-1)}(1 + \bar{X}(t-1)^2)$ και $g(t, \lambda)$ η πιθανογεννήτρια της Poisson με παράμετρο λ.

Η ιδέα των Kocherlakota and Kocherlakota (1986) στηρίζεται στο αποτέλεσμα που διατυπώθηκε στο Θεώρημα 1.2.8 και ήταν να θεωρηθεί ως σ.σ.ε., για συγκεκριμένη τιμή του t, η διαφορά μεταξύ της εμπειρικής πιθανογεννήτριας συνάρτησης και της εκτιμώμενης πιθανογεννήτριας συνάρτησης κατανομής υπό τη μηδενική υπόθεση. Τότε, αν $g(t, \lambda)$ η πιθανογεννήτρια συνάρτηση υπό τη μηδενική υπόθεση, εφαρμόζοντας ανάπτυγμα Taylor γύρω από την παράμετρο λ ισχύει ότι:

$$g(t, \hat{\lambda}) = g(t, \lambda) + (\hat{\lambda} - \lambda)\frac{d}{d\lambda}g(t, \lambda),$$

και επομένως

$$\left(g_n(t) - g(t, \hat{\lambda})\right) = (g_n(t) - g(t, \lambda)) - (\hat{\lambda} - \lambda)\frac{d}{d\lambda}g(t, \lambda).$$

Ανακαλώντας τη σχέση (1.10) προκύπτει ότι:

$$\frac{d}{d\lambda}g(t, \lambda) = (t-1)e^{\lambda(t-1)}.$$

Επομένως, έχουμε ότι:

$$\left(g_n(t) - g(t, \hat{\lambda})\right) = (g_n(t) - g(t, \lambda)) - (\hat{\lambda} - \lambda)(t-1)e^{\lambda(t-1)}.$$

Τότε, λαμβάνοντας υπόψη το Θεώρημα 1.2.9, την πιθανογεννήτρια της κατανομής Poisson (βλέπε σχέση (1.9)), το γεγονός ότι $\hat{\lambda} = \bar{X}$ και ότι

$$Cov(t^X, X) = E(t^X X) - E(X)E(t^X) = t\frac{d}{dt}g(t, \lambda) - \lambda g(t),$$

και ύστερα από αρκετή άλγεβρα (βλέπε Kocherlakota and Kocherlakota (1986) προκύπτει ότι η σ.σ. $\sqrt{n}\left(g_n(t) - g(t, \hat{\lambda})\right)$ συγκλίνει κατά κατανομή στην κανονική κατανομή με μέση τιμή 0 και διακύμανση ίση με

$$\sigma^2(t,\lambda) = \exp(\lambda(t^2-1)) - \exp(2\lambda(t-1))(1+\lambda(t-1)^2).$$

Με αντικατάσταση στην έκφραση της διακύμανσης της άγνωστης τιμής λ από έναν συνεπή εκτιμητή της (π.χ. την $\hat{\lambda} = \bar{X}$) προκύπτει η σ.σ. που προτάθηκε από τους Kocherlakota and Kocherlakota (1986). Δηλαδή κατά αυτόν τον τρόπο προκύπτει η σ.σ. της επόμενης σχέσης

$$K = \sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n} t^{X_i} - e^{\bar{X}(t-1)}}{e^{\bar{X}(t^2-1)} - e^{2\bar{X}(t-1)}(1+\bar{X}(t-1)^2)}. \tag{2.24}$$

Επομένως, σε επίπεδο σημαντικότητας α η μηδενική υπόθεση απορρίπτεται αν $|K| > z_{\alpha/2}$.

Ωστόσο, όπως επισημαίνουν οι Karlis and Xekalaki (2000) , ο παραπάνω έλεγχος έχει το μειονέκτημα ότι εξαρτάται από την επιλογή του t. Ειδικότερα, παρότι οι Kocherlakota and Kocherlakota (1986) έδειξαν ότι ο έλεγχος δεν είναι πολύ ευαίσθητος στην επιλογή της παραμέτρου t, μπορεί η απόδοσή του να βελτιωθεί χρησιμοποιώντας περισσότερες από μία τιμές της παραμέτρου t, έχοντας όμως ως συνέπεια το υπολογιστικό κόστος. Σε περίπτωση που χρησιμοποιείται μία τιμή, τότε προτείνεται να επιλέγεται μία μικρή θετική τιμή κοντά στο 0. Τέλος, ένας άλλος τρόπος αντιμετώπισης του προβλήματος της επιλογής της τιμής του t είναι να χρησιμοποιηθεί ο έλεγχος που προτάθηκε από τους Rueda et al. (1991). Ο έλεγχος αυτός, καθώς και γενικεύσεις του, παρουσιάζονται στη συνέχεια.

Ο έλεγχος των Rueda et al. (1991) και η γενίκευσή του

Οι Rueda et al. (1991) για να αποφευχθεί το πρόβλημα επιλογής του t και παρακινούμενοι από τη λογική του ελέγχου των Cramér-von Mises πρότειναν να βασίζεται ο έλεγχος καλής προσαρμογής της Poisson κατανομής στη σ.σ. που προκύπτει από το ολοκλήρωμα στο διάστημα $[0,1]$ των τετραγώνων των διαφορών της εμπειρικής πιθανογεννήτριας συνάρτησης από την εκτιμώμενη, υπό τη μηδενική υπόθεση, πιθανογεννήτρια συνάρτηση. Δηλαδή, ο έλεγχος τους βασίζεται στη στατιστική συνάρτηση:

$$R = n\int_0^1 \left(g_n(t) - g(t, \hat{\lambda} = \bar{X})\right)^2 dt.$$

Λαμβάνοντας υπόψη τον ορισμό της ε.π.σ. και την έκφραση της πιθανογεννήτριας συνάρτησης της κατανομής Poisson έχουμε με αντικατάσταση ότι:

$$\begin{aligned} R &= n\int_0^1 \left(\frac{1}{n}\sum_{i=1}^n t^{X_i} - e^{\bar{X}(t-1)}\right)^2 dt \\ &= n\int_0^1 \left(\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n t^{X_i+X_j} - 2\frac{1}{n}\sum_{i=1}^n t^{X_i}e^{\bar{X}t}e^{-\bar{X}} + e^{2\bar{X}t}e^{-2\bar{X}}\right) dt \\ &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \int_0^1 t^{X_i+X_j}dt - 2e^{-\bar{X}}\sum_{i=1}^n\int_0^1 t^{X_i}e^{\bar{X}t}dt + ne^{-2\bar{X}}\int_0^1 e^{2\bar{X}t}dt \\ &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \frac{1}{X_i+X_j+1} - 2e^{-\bar{X}}\sum_{i=1}^n T(X_i,\bar{X}) + n\left(\frac{1-e^{-2\bar{X}}}{2\bar{X}}\right), \end{aligned} \tag{2.25}$$

όπου

$$T(X_i,\bar{X}) = \int_0^1 t^{X_i}e^{\bar{X}t}dt.$$

Προφανώς, η μηδενική υπόθεση απορρίπτεται για μεγάλες τιμές της στατιστικής συνάρτησης. Η ασυμπτωτική κατανομή της σ.σ. υπό τη μηδενική υπόθεση έχει προσδιοριστεί από τους Rueda et al. (1991), όντας το ολοκλήρωμα του τετραγώνου μιας γκαουσιανής διαδικασίας με μέση τιμή μηδέν και συνάρτηση συνδιακύμανσης που εξαρτάται από την άγνωστη παράμετρο λ. Για τον λόγο αυτό στην πράξη χρησιμοποιείται παραμετρικός έλεγχος bootstrap.

Ο έλεγχος των Rueda et al. (1991) γενικεύθηκε από τους Baringhaus et al. (2000), οι οποίοι εισήγαγαν στο ολοκλήρωμα μια συνάρτηση βάρους στην προσπάθειά τους να γίνει ο έλεγχος πιο ισχυρός στον εντοπισμό αποκλίσεων από την κατανομή Poisson. Ειδικότερα, εισάγοντας ως συνάρτηση βάρους την $w(t) = t^a$, με $a \geq 0$ να είναι μια σταθερά, πρότειναν τη σ.σ.:

$$R_a = n \int_0^1 \left(g_n(t) - g(t, \hat{\lambda} = \bar{X}) \right)^2 t^a dt.$$

Η επιλογή μίας μεγάλης τιμής του a στην σ.σ. σημαίνει ότι μπαίνει περισσότερο βάρος κοντά στο τελικό σημείο $t = 1$. Επειδή οι ροπές του X σχετίζονται με τις μονόπλευρες παραγώγους της πιθανογεννήτριας συνάρτησης $g(t)$, ισχύει ότι για μεγάλες τιμές του a, ο έλεγχος της μηδενικής υπόθεσης με βάση τη σ.σ.ε. R_a είναι ιδιαίτερα ευαίσθητο ως προς τις αποκλίσεις των ροπών του X, από τις αντίστοιχες τιμές υπό την κατανομή Poisson. Προφανώς, η μηδενική υπόθεση και πάλι απορρίπτεται για μεγάλες τιμές της σ.σ. R_a. Στο σημείο αυτό επισημαίνεται ότι η ασυμπτωτική κατανομή αυτής της σ.σ., υπό τη μηδενική υπόθεση, μελετήθηκε από τους Gürtler and Henze (2000), οι οποίοι απέδειξαν ότι εξαρτάται τόσο από την παράμετρο a όσο και από την άγνωστη παράμετρο λ. Επιπλέον απέδειξαν ότι η χρήση παραμετρικού bootstrap οδηγεί σε έλεγχο με ασυμπτωτικό επίπεδο σημαντικότητας α, καθώς το μέγεθος δείγματος και το πλήθος των bootstrap δειγμάτων τείνει στο άπειρο. Τέλος, μία εναλλακτική μορφή της σ.σ. που καθιστά πιο εύκολο τον υπολογισμό της έχει δοθεί από τους Gürtler and Henze (2000) και είναι η ακόλουθη:

$$R_a = n \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} \left[\frac{(f_n(i) - f_0(i, \bar{X}))(f_n(k) - f_0(k, \bar{X}))}{i + k + a + 1} \right] \tag{2.26}$$

όπου $f_n(x) = \dfrac{1}{n}\sum_{i=1}^{n} I(X_i = x)$ και $f_0(x,\lambda)$ η συνάρτηση πιθανότητας της Poisson με παράμετρο λ.

Έλεγχοι των Baringhaus and Henze (1992) , Treutler (1995) και Meintanis and Nikitin (2008)

Οι τρεις αυτοί έλεγχοι βασίζονται στη χαρακτηριστική ιδιότητα της κατανομής Poisson που δόθηκε στην Πρόταση 1.2.11, σύμφωνα με την οποία η πιθανογεννήτρια συνάρτηση της κατανομής Poisson είναι η μοναδική πιθανογεννήτρια συνάρτηση στο σύνολο των πιθανογεννητριών συναρτήσεων που αντιστοιχούν σε τ.μ. με τιμές στο σύνολο $\{0, 1, ...\}$ και με πεπερασμένη μέση τιμή που ικανοποιεί τη διαφορική εξίσωση $\dfrac{\partial}{\partial t} g(t) - \lambda g(t) = 0$.

Χρησιμοποιώντας τον παραπάνω χαρακτηρισμό, η ιδέα των Baringhaus and Henze (1992) είναι ότι το δειγματικό ανάλογο της παραπάνω διαφορικής εξίσωσης, μπορεί να χρησιμεύσει ως ένας τρόπος ελέγχου αποκλίσεων από αυτήν την κατανομή. Στο πλαίσιο αυτό, πρότειναν τη στατιστική συνάρτηση:

$$T = n\int_0^1 (\bar{X} g_n(t) - g_n'(t))^2 dt.$$

Λαμβάνοντας υπόψη ότι

$$g_n(t) = \frac{1}{n}\sum_{i=1}^{n} t^{X_i} \text{ και } g_n'(t) = \frac{1}{n}\sum_{i=1}^{n} X_i t^{X_i - 1},$$

ύστερα από λίγη άλγεβρα προκύπτει η ακόλουθη ισοδύναμη μορφή:

$$T = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\bar{X}^2}{X_i + X_j + 1} + \frac{X_i X_j}{X_i + X_j - 1}\right) - (n - f(0))\bar{X}, \qquad (2.27)$$

όπου είναι $f(0) = \frac{1}{n}(\sum_{i=1}^{n} I(X_i = 0))^2$ (βλέπε Gürtler and Henze (2000)).

Μία γενίκευση του παραπάνω ελέγχου προτάθηκε στη διδακτορική διατριβή Treutler (1995), με την εισαγωγή της συνάρτησης βάρους t^a στο ολοκλήρωμα. Ειδικότερα, προτάθηκε η σ.σ. που δίνεται από τη σχέση:

$$T_a = n \int_0^1 (g_n'(t) - \bar{X} g_n(t))^2 t^a dt$$

ή ισοδύναμα, όπως προκύπτει μετά από λίγη άλγεβρα, από τη σχέση:

$$T_a = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{\bar{X}^2}{X_i + X_j + a + 1} - \frac{\bar{X}(X_i + X_j)}{X_i + X_j + a} + \frac{X_i X_j}{X_i + X_j + a - 1} \right) \tag{2.28}$$

όπου $a \geq 0$ είναι μια σταθερά. Με την εισαγωγή της συνάρτησης του βάρους, t^a, να επιτυγχάνεται, ο έλεγχος να είναι πιο ευαίσθητος σε σχέση με επιλεγμένες εναλλακτικές (βλέπε Gürtler and Henze (2000)).

Τέλος, ο έλεγχος που προτάθηκε από τους Meintanis and Nikitin (2008) διαφοροποιείται από τον έλεγχο που προτάθηκε από τον Treutler (1995) (άρα και από αυτόν των Baringhaus and Henze (1992)), καθώς παρότι στον ορισμό της περιέχει συνάρτηση βάρους δεν υψώνεται στο τετράγωνο το δειγματικό ανάλογο της διαφορικής εξίσωσης που πληροί η κατανομή Poisson. Ειδικότερα, έχοντας ως συνάρτηση βάρους την t^α, με α μία θετική παράμετρο, προτείνεται η ακόλουθη σ.σ.:

$$MN_a^* = \sqrt{n} \int_0^1 (g_n'(t) - \bar{X} g_n(t)) t^a dt. \tag{2.29}$$

Ισοδύναμα έχουμε ότι:

$$\begin{aligned} MN_a^* &= \sqrt{n}\int_0^1 \left(\frac{1}{n}\sum_{i=1}^n X_i t^{X_i-1} - \bar{X}\frac{1}{n}\sum_{i=1}^n t^{X_i}\right) t^a dt \\ &= \sqrt{n}\frac{1}{n}X_i\sum_{i=1}^n \int_0^1 t^{X_i-1}t^a dt - \sqrt{n}\bar{X}\frac{1}{n}\sum_{i=1}^n\int_0^1 t^{X_i}t^a dt \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{X_i}{X_i+a} - \frac{\bar{X}}{X_i+a+1}\right). \end{aligned}$$

Προφανώς, η μηδενική υπόθεση και πάλι απορρίπτεται για μεγάλες τιμές των σ.σ. T και T_a, ενώ απορρίπτεται για μεγάλες κατά απόλυτο τιμές της σ.σ. MN_a^*. Στο σημείο αυτό επισημαίνεται ότι η ασυμπτωτική κατανομή της σ.σ. T υπό τη μηδενική υπόθεση μελετήθηκε από τους Baringhaus and Henze (1992), οι οποίοι απέδειξαν ότι εξαρτάται από την άγνωστη παράμετρο λ. Επιπλέον απέδειξαν ότι η χρήση του παραμετρικού bootstrap οδηγεί σε έλεγχο με ασυμπτωτικό επίπεδο σημαντικότητας α, καθώς το μέγεθος δείγματος και το πλήθος των bootstrap δειγμάτων τείνει στο άπειρο και με ισχύ που τείνει στη μονάδα για οποιαδήποτε εναλλακτική κατανομή με πεπερασμένη μέση τιμή.

Από την άλλη μεριά, η σ.σ. που προτάθηκε από τους Meintanis and Nikitin (2008) ασυμπτωτικά υπό τη μηδενική υπόθεση ακολουθεί κανονική κατανομή με μέση τιμή 0 και διακύμανση που προσδιορίζεται από τη σχέση:

$$V_a(\lambda) = \lambda^2\bar{\epsilon}_{a+2} + \lambda(\lambda+1)\bar{\epsilon}_{a+1} + 2\lambda^2(\epsilon_{a+2} - \epsilon_{a+1}) - \lambda\epsilon_{a+1}^2,$$

όπου

$$\epsilon_a = E[(X+a)^{-1})] \text{ και } \bar{\epsilon}_a = E[(X+a)^{-2})],$$

τα οποία υπολογίζονται μέσω των σχέσεων:

$$\lambda\epsilon_{a+1} = 1 - a\epsilon_a \text{ και } \lambda\bar{\epsilon}_{a+1} = \epsilon_a - a\bar{\epsilon}_a.$$

Καθώς η διακύμανση της ασυμπτωτικής κατανομής, υπό τη μηδενική υπόθεση, εξαρτάται από την άγνωστη παράμετρο λ, σε πρακτικές εφαρμογές χρησιμοποιείται η σ.σ. που προκύπτει με αντικατάσταση στην έκφραση της διακύμανσης της άγνωστης παραμέτρου από έναν συνεπή εκτιμητή. Επομένως, σε πρακτικές εφαρμογές χρησιμοποιείται η σ.σ.

$$MN_a(\hat{\lambda}) = \frac{MN_a^*}{\sqrt{V_a(\hat{\lambda})}}. \tag{2.30}$$

Η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας α όταν $|MN_a(\hat{\lambda})| > z_{\alpha/2}$.

Τεστ Nakamura and Pérez-Abreu (1993)

Ο έλεγχος αυτός στηρίζεται στην ιδιότητα της κατανομής Poisson που δόθηκε στο Πόρισμα 1.2.12, από όπου έχουμε ότι η πιθανογεννήτρια συνάρτησή της είναι η μόνη για την οποία ισχύει ότι

$$\frac{\partial^2}{\partial t^2}\log(g(t)) = 0.$$

Άρα, υπό την υπόθεση ότι τα δειγματικά δεδομένα προέρχονται από την κατανομή Poisson και όταν το μέγεθος δείγματος n είναι μεγάλο, τότε ο λογάριθμος της εμπειρικής πιθανογεννήτριας συνάρτησης, $\log(g_n(t))$, είναι σχεδόν ευθεία γραμμή. Παίρνοντας τη δεύτερη παράγωγο του λογαρίθμου της εμπειρικής πιθανογεννήτριας συνάρτησης και αναλύοντάς την:

$$\frac{\partial^2}{\partial t^2}\log(g_n(t)) = \frac{g_n(t)\frac{\partial^2}{\partial t^2}g_n(t) - \left(\frac{\partial}{\partial t}g_n(t)\right)^2}{g_n^2(t)}$$

μπορούμε να ποσοτικοποιήσουμε την απόκλιση της $\log(g_n(t))$ από την ευθεία γραμμή. Ο αριθμητής του παραπάνω πηλίκου εύκολα μπορεί να γραφεί στη μορφή

$$g_n(t)\frac{\partial^2}{\partial t^2}g_n(t) - \left(\frac{\partial}{\partial t}g_n(t)\right)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} t^{X_i+X_j-2}(X_i(X_i - X_j - 1)).$$

Επίσης καθώς $g_n^2(t) > 0$ για όλα τα t η δεύτερη παράγωγος του λογαρίθμου της εμπειρικής πιθανογεννήτριας συνάρτησης είναι μηδέν αν και μόνο αν η $N_n(t)$ που ορίζεται από τη σχέση:

$$N_n(t) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} t^{X_i+X_j-2}(X_i(X_i - X_j - 1)),$$

είναι ίση με μηδέν. Παρατηρήστε ότι

$$N_n(1) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i(X_i - X_j - 1)) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 - \frac{1}{n}\sum_{j=1}^{n}X_j = S^2 - \bar{X},$$

Η ιδιότητα αυτή χρησιμοποιήθηκε από τους Nakamura and Pérez-Abreu (1993) στο ακόλουθο πλαίσιο. Αν $X_{(n)} = \max\{X_1, ..., X_n\}$, τότε παρατήρησαν αρχικά ότι το $N_n(t)$ είναι ένα τυχαίο πολυώνυμο ως προς t βαθμού $2X_{(n)} - 2$. Αυτό ερμηνεύεται ότι μπορεί να γραφεί στην ισοδύναμη μορφή:

$$N_n(t) = \sum_{k=0}^{2X_{(n)}-2} a_k t^k,$$

με τους πολυωνυμικούς συντελεστές a_k να προσδιορίζονται από τη σχέση:

$$a_k = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i(X_i - X_j - 1))I(X_i + X_j - 2 = k)$$

με $I(\cdot)$ τη συνήθη δείκτρια συνάρτηση. Καθώς αυτό το πολυώνυμο πρέπει να είναι ίσο με το μηδέν, πρότειναν ως στατιστική συνάρτηση το άθροισμα των τετραγώνων των πολυωνυμικών συντελεστών a_k και μεγάλες τιμές αυτού του αθροίσματος οδηγούν σε απόρριψη της μηδενικής υπόθεσης. Συνοψίζοντας, η σ.σ. που προτάθηκε από τους Nakamura and Pérez-Abreu (1993) είναι η:

$$V := \sum_{k=0}^{2X_{(n)}-2} a_k^2,$$

ισοδύναμα

$$V = \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}(X_i(X_i - X_j - 1)X_k(X_k - X_l - 1)I(X_i + X_j = X_k + X_l). \tag{2.31}$$

Επιπλέον, οι Nakamura and Pérez-Abreu (1993) απέδειξαν την ασθενή σύγκλιση του V υπό τη μηδενική υπόθεση σε μια σειρά σταθμισμένων ανεξάρτητων χ_1^2 τυχαίων μεταβλητών. Οι κρίσιμες τιμές αυτού του ορίου προσεγγίστηκαν με μια περικοπή της άπειρης δυναμοσειράς και είναι διαθέσιμες, για συγκεκριμένες τιμές της παραμέτρου λ ή της εκτίμησή της $\hat{\lambda}$, σε πίνακες που παρατίθενται στην εργασία των Nakamura and Pérez-Abreu (1993) και στη σχετική προδημοσίευση αυτής.

Οι Nakamura and Pérez-Abreu (1993) θέλοντας να προτείνουν μία σ.σ. που θα οδηγεί σε κρίσιμες τιμές που είναι ανεξάρτητες από την παράμετρο λ της κατανομής Poisson οδηγήθηκαν, μετά από προσομοιώσεις, στην ακόλουθη τροποποίηση της στατιστικής συνάρτησης:

$$V^* = \frac{nV}{\bar{X}^{1.45}}. \tag{2.32}$$

Προφανώς και αυτήν τη φορά η μηδενική υπόθεση απορρίπτεται για μεγάλες τιμές της στατιστικής συνάρτησης ελέγχου.

2.5 Έλεγχοι που στηρίζονται στις ροπές

Το ενδιαφέρον στην ενότητα αυτή επικεντρώνεται στην παρουσίαση ελέγχων που αξιοποιούν ιδιότητες της κατανομής Poisson που σχετίζονται με τις πληθυσμιακές ροπές και συναρτήσεις αυτών.

2.5.1 Έλεγχοι με τον δειγματικό δείκτη διασποράς

Στο Πόρισμα 1.2.17 διατυπώθηκε το συμπέρασμα ότι στην περίπτωση της Poisson κατανομής με παράμετρο $\lambda > 0$, καθώς η μέση τιμή και η διακύμανση είναι ίσες με το λ, προκύπτει ότι η τιμή του δείκτη διασποράς δ είναι ίση με 1. Από την άλλη μεριά, σε περιπτώσεις κατανομών όπου η διακύμανση είναι μεγαλύτερη (μικρότερη, αντίστοιχα) από τη μέση τιμή, δηλαδή σε περιπτώσεις υπερδιασκορπισμένων (υποδιασκορπισμένων, αντίστοιχα) κατανομών, η τιμή του πληθυσμιακού δείκτη διασποράς λαμβάνει τιμές μεγαλύτερες (μικρότερες, αντίστοιχα) από τη μονάδα.

Με βάση τα παραπάνω, ένας τρόπος ελέγχου αποκλίσεων από την κατανομή Poisson μπορεί να αναχθεί στον έλεγχο αν ο πληθυσμιακός δείκτης διασποράς είναι ίσος με τη μονάδα ή ένα πολλαπλάσιό του με την αντίστοιχη πολλαπλάσια τιμή. Το πρώτο βήμα για τη διεξαγωγή τέτοιων ελέγχων είναι ο προσδιορισμός ενός εκτιμητή της αντίστοιχης πληθυσμιακής ποσότητας. Στο πλαίσιο αυτό, έχουν οριστεί ο Poisson δείκτης διασποράς και ο δείκτης διασποράς του Fisher. Ειδικότερα, ο Poisson δείκτης διασποράς ορίζεται από τη σχέση:

$$D_n = \frac{(n-1)S^2}{\bar{X}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\bar{X}}, \tag{2.33}$$

όπου $\bar{X}$ και S^2 είναι η δειγματική μέση τιμή και διακύμανση, αντίστοιχα, ενώ ο δείκτης διασποράς του Fisher ορίζεται από τη σχέση:

$$FI = \frac{S^2}{\bar{X}} = \frac{D_n}{n-1}. \tag{2.34}$$

Επομένως, άμεσα, προκύπτει ότι ο Poisson δείκτης διασποράς είναι απλά το πηλίκο του δείκτη διασποράς του Fisher με το $(n-1)$. Στη βιβλιογραφία έχει εμφανιστεί ως Poisson δείκτης διασποράς και το πηλίκο $\frac{D_n}{n}$, το οποίο είναι ασυμπτωτικά ισοδύναμο με τον δείκτη FI.

Το επόμενο βήμα για τη διεξαγωγή ενός στατιστικού ελέγχου υποθέσεων είναι να προσδιοριστεί, αρχικά, η κατανομή της σ.σ.ε. υπό τη μηδενική υπόθεση και στη συνέχεια η κρίσιμη περιοχή του ελέγχου. Ακολουθώντας αυτήν την πορεία, στην ενότητα αυτή θα παρουσιαστούν έλεγχοι που έχουν εμφανιστεί στη βιβλιογραφία και βασίζονται στις σ.σ. D_n και FI.

Στο θεώρημα που ακολουθεί προσδιορίζεται η ασυμπτωτική κατανομή της σ.σ. D_n όταν το τυχαίο δείγμα προέρχεται από την κατανομή Poisson (βλέπε, μεταξύ άλλων, Fisher and Thornton (1922), Selby (1965), Hoel (1943), Avelino (1978) και τις εκεί αναφορές).

Θεώρημα 2.5.1. *Έστω $X_1, X_2, ..., X_n$ ένα τυχαίο δείγμα μεγέθους n από την κατανομή Poisson με παράμετρο λ. Τότε η σ.σ. D_n ακολουθεί προσεγγιστικά* *χι-τετράγωνο κατανομή με $n-1$ βαθμούς ελευθερίας, δηλαδή*

$$D_n \xrightarrow{d} \chi^2_{n-1}. \tag{2.35}$$

Απόδειξη. Για την απόδειξη του θεωρήματος παραπέμπουμε ενδεικτικά στους Rao and Chakravarti (1956). □

Σύμφωνα με το αποτέλεσμα που δόθηκε στο Θεώρημα 2.5.1, η μηδενική υπόθεση απορρίπτεται για μεγάλες ή μικρές τιμές της σ.σ.ε. D_n. Προφανώς, οι κρίσιμες τιμές του ελέγχου, δηλαδή ποιες τιμές θεωρούνται μικρές ή μεγάλες προσδιορίζονται με τη βοήθεια της χι-τετράγωνο κατανομής με $n-1$ βαθμούς ελευθερίας, έτσι ώστε να επιτυγχάνεται (ασυμπτωτικό) επίπεδο σημαντικότητας ίσο με α. Στο σημείο αυτό, πρέπει να επισημανθεί ότι η προσέγγιση της κατανομής της σ.σ.ε. D_n από τη χι-τετράγωνο κατανομή με $n-1$ βαθμούς ελευθερίας που δόθηκε στο Θεώρημα 2.5.1 είναι, όπως διαπίστωσαν οι Anderson and Siddiqui (1994), ικανοποιητική ακόμα και για μικρά σε μέγεθος δείγματα (π.χ. για δείγματα μεγέθους ίσα με 5), υπό την προϋπόθεση ότι η μέση τιμή της κατανομής είναι μεγαλύτερη από 3 ή 5 (βλέπε Selby (1965)), ενώ όταν η μέση τιμή είναι μικρότερη από 1.5 (ή από 5, βλέπε Selby (1965)) χρειάζεται για να είναι ικανοποιητική η παραπάνω προσέγγιση πολύ μεγάλο μέγεθος δείγματος. Αυτές οι παρατηρήσεις οδήγησαν τους Anderson and Siddiqui (1994) να προσπαθήσουν να βελτιώσουν την προσέγγιση της κατανομής της σ.σ. D_n. Με αυτόν τον στόχο, αρχικά, προσέγγισαν τις τέσσερις πρώτες ροπές της σ.σ. D_n. Έπειτα, πρότειναν να προσεγγιστεί η κατανομή της σ.σ. D_n από τη γάμμα κατανομή με παραμέτρους τέτοιες ώστε οι δύο πρώτες ροπές της γάμμα κατανομής να ταυτίζονται με τις δύο πρώτες ροπές που προηγούμενα είχαν προσδιορίσει για την D_n. Επιπλέον, σε μία προσπάθεια να βελτιωθεί περισσότερο η προσέγγιση που προτάθηκε χρησιμοποίησαν τη μέθοδο προσέγγισης με ορθογώνια πολυώνυμα. Πιο συγκεκριμένα, χρησιμοποιώντας την τρίτη και τέταρτη τάξης ροπή της D_n, πρότειναν πολυώνυμα Hermite και Laguerre τρίτης και τέταρτης τάξης, καθώς και σειρές Edgeworth τέταρτης τάξης. Μέσω μελέτης προσομοίωσης κατέληξαν στο συμπέρασμα ότι καμία από τις παραπάνω μεθόδους δεν παρέχει βελτίωση στην προσέγγιση της κατανομής της σ.σ. D_n. Τα παραπάνω συμπεράσματα έχουν ως συνέπεια να μην χρησιμοποιείται στην πράξη ο έλεγχος που βασίζεται στη σ.σ. D_n, καθώς τα αποτελέσματά του με χρήση της ασυμπτωτικής κατανομής δεν είναι έγκυρα παρά μόνο για πολύ μεγάλο μέγεθος δείγματος (Karlis and Xekalaki (2000)). Για τον λόγο αυτό, δεν παρουσιάζονται στη συνέχεια περισσότερες λεπτομέρειες και αποτελέσματα σχετικά με αυτόν τον έλεγχο.

Στο θεώρημα που ακολουθεί (βλέπε, μεταξύ άλλων, Henze and Klar (1996)) προσδιορίζεται η ασυμπτωτική κατανομή μίας σ.σ. που αποτελεί συνάρτηση του D_n.

Θεώρημα 2.5.2. *Έστω $X_1, X_2, ..., X_n$ ένα τυχαίο δείγμα μεγέθους n από μια κατανομή που ανήκει στην οικογένεια των κατανομών για τις οποίες η μέση τιμή, μ, είναι ίση με τη διακύμανση, δηλαδή είναι τέτοια ώστε $\mu_2 = \mu$, με $\mu_2 = E(X-\mu)^2$. Τότε, για μεγάλο μέγεθος δείγματος, ισχύει ότι:*

$$U_1 = \frac{1}{\sqrt{2n}}(D_n - n) \xrightarrow{d} N(0, \sigma^2), \tag{2.36}$$

όπου

$$\sigma^2 = \frac{1}{2\mu^2}(\mu_4 - 2\mu_3 - 2\mu_2\mu + \mu_2 + \mu^2),$$

με $\mu_i = E(X-\mu)^i$, για $i = 2, 3, 4$.

Απόδειξη. Αρχικά, λαμβάνοντας υπόψη τη σχέση (2.33), παρατηρούμε ότι:

$$\begin{aligned}\frac{1}{\sqrt{2n}}(D_n - n) &= \frac{1}{\sqrt{2n\bar{X}}}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 - n\bar{X}\right] \\ &= \frac{1}{\sqrt{2n\bar{X}}}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 - n\bar{X}\right],\end{aligned}$$

και λαμβάνοντας υπόψιν ότι

$$\bar{X}^2 = (\bar{X} - \mu)^2 + 2\bar{X}\mu - \mu^2,$$

έχουμε ότι:

$$\begin{aligned}\frac{1}{\sqrt{2n}}(D_n - n) &= \frac{1}{\sqrt{2n\bar{X}}}\left[\sum_{i=1}^{n} X_i^2 - n(\bar{X} - \mu)^2 - 2\sum_{i=1}^{n} X_i\mu + n\mu^2 - n\bar{X}\right] \\ &= \frac{1}{\sqrt{2n\bar{X}}}\sum_{i=1}^{n}\left[X_i^2 - (1+2\mu)X_i + \mu^2\right] - \frac{1}{\sqrt{2n\bar{X}}}n(\bar{X} - \mu)^2 \\ &= \frac{1}{\sqrt{2n\bar{X}}}\sum_{i=1}^{n}\left[X_i^2 - (1+2\mu)(X_i - \mu) + \mu^2 - (1+2\mu)\mu\right] \\ &- \frac{1}{\sqrt{2n\bar{X}}}n(\bar{X} - \mu)^2,\end{aligned}$$

ή ισοδύναμα

$$\begin{aligned}\frac{1}{\sqrt{2n}}(D_n - n) &= \frac{1}{\sqrt{2n\bar{X}}}\sum_{i=1}^{n}\left[X_i^2 - \mu^2 - \mu - (1+2\mu)(X_i - \mu)\right] \\ &- \frac{1}{\sqrt{2n\bar{X}}}n(\bar{X} - \mu)^2.\end{aligned}$$

Ο δεύτερος από τους παραπάνω όρους συγκλίνει υπό τη μηδενική υπόθεση κατά πιθανότητα στο μηδέν. Όσον αφορά τον πρώτο όρο προσδιορίζεται η ασυμπτωτική του κατανομή υπό τη μηδενική υπόθεση με εφαρμογή του Κεντρικού Οριακού Θεωρήματος και της δέλτα μεθόδου. Ειδικότερα ορίζοντας την τυχαία μεταβλητή

$$Y_i = X_i^2 - \mu - \mu^2 - (1+2\mu)(X_i - \mu),$$

ο προσδιορισμός της ασυμπτωτικής, υπό τη μηδενική υπόθεση, κατανομής του πρώτου όρου ανάγεται στον προσδιορισμό της ασυμπτωτικής κατανομής υπό τη μηδενική υπόθεση της

$$\sqrt{n}\phi\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \frac{1}{n}\sum_{i=1}^{n} X_i\right),$$

όπου $\phi(x,y) = \dfrac{x}{\sqrt{2y}}$, με $\phi' = \left(\frac{d}{dx}\frac{x}{\sqrt{2y}}, \frac{d}{dy}\frac{x}{\sqrt{2y}}\right) = \left(\frac{1}{\sqrt{2y}}, -\frac{x}{\sqrt{2y^2}}\right)$.

Από το πολυδιάστατο κεντρικό οριακό θεώρημα (βλέπε Van der Vaart (2000))

$$\sqrt{n}\left[\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n} Y_i\\ \frac{1}{n}\sum_{i=1}^{n} X_i\end{pmatrix} - \begin{pmatrix}0\\ \mu\end{pmatrix}\right] \xrightarrow{d} N_2\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \Sigma\right),$$

καθώς, υπό τη μηδενική υπόθεση $Var(X) = \mu$ και $E(X_i - \mu) = 0$, και επομένως:

$$\begin{aligned} E(Y_i) &= E(X_i^2) - \mu - \mu^2 - (1+2\mu)E(X_i - \mu) \\ &= Var(X_i) + (EX_i)^2 - \mu - \mu^2 - (1+2\mu)(E(X_i) - \mu) \\ &= \mu + \mu^2 - \mu - \mu^2 = 0, \end{aligned}$$

ενώ, καθώς $E(Y_i) = 0$, έχουμε ότι:

$$\Sigma = \begin{pmatrix} Var(Y_i) & Cov(X_i, Y_i) \\ Cov(X_i, Y_i) & Var(X_i) \end{pmatrix} = \begin{pmatrix} Var(Y_i) & E(X_iY_i) \\ E(X_iY_i) & Var(X_i) \end{pmatrix}.$$

Ωστόσο, για την εφαρμογή της δέλτα μεθόδου καθώς είναι $\phi'(x,y)\big|_{(0,\mu)} = \left(\frac{1}{\mu}, 0\right)$ απαιτείται μόνο ο προσδιορισμός της $Var(Y_i)$, η οποία μετά από λίγη άλγεβρα προκύπτει ότι δίνεται από τη σχέση:

$$Var(Y_i) = \mu_4 - 2\mu_3 - 2\mu_2\mu + \mu_2 + \mu^2.$$

Το επιθυμητό αποτέλεσμα ότι $U_1 \xrightarrow{d} N(0, \sigma^2)$ προκύπτει με εφαρμογή της δέλτα μεθόδου, καθώς ύστερα από λίγη άλγεβρα προκύπτει ότι $\sigma^2 = \frac{1}{2\mu^2} Var(Y_i)$. □

Ως ειδική περίπτωση του παραπάνω θεωρήματος προκύπτει το ακόλουθο αποτέλεσμα.

Πόρισμα 2.5.3. *Έστω $X_1, X_2, ..., X_n$ ένα τυχαίο δείγμα μεγέθους n από την κατανομή Poisson με παράμετρο λ. Τότε, για μεγάλο μέγεθος δείγματος, ισχύει ότι:*

$$U = \frac{1}{\sqrt{2n}}(D_n - n) \xrightarrow{d} N(0,1), \tag{2.37}$$

ή ισοδύναμα

$$U^2 \xrightarrow{d} \chi_1^2. \tag{2.38}$$

Απόδειξη. Το επιθυμητό αποτέλεσμα προκύπτει άμεσα από το αποτέλεσμα της Πρότασης 2.5.2, καθώς η κατανομή Poisson πληροί τις υποθέσεις εφαρμογής της, με: $\mu = \lambda$, $\mu_2 = Var(X) = \lambda$, $\mu_3 = \lambda$ και $\mu_4 = \lambda + 3\lambda^2$. Άρα σε αυτήν την περίπτωση, μετά από λίγη άλγεβρα, έχουμε ότι

$$\sigma^2 = \frac{\mu + 3\mu - 2\mu - 2\mu^2 + \mu + \mu^2}{2\mu^2} = 1.$$

Άρα η στατιστική συνάρτηση ελέγχου U ακολουθεί ασυμπτωτικά την τυπική κανονική κατανομή και επομένως $U^2 \xrightarrow{d} \chi_1^2$. □

Παρατήρηση 2.5.4. *Η στατιστική συνάρτηση U ταυτίζεται με αυτήν που προτάθηκε από τους Zelterman and Chen (1988) στα πλαίσια του ελέγχου της υπόθεσης ότι τα δεδομένα προέρχονται από την κατανομή Poisson έναντι της εναλλακτικής υπόθεσης ότι τα δεδομένα προέρχονται από μείξη Poisson.*

Με βάση το Πόρισμα 2.5.3, η μηδενική υπόθεση ότι τα δεδομένα προέρχονται από την Poisson κατανομή απορρίπτεται τόσο για μικρές όσο και για μεγάλες τιμές της σ.σ.ε. U ή ισοδύναμα για μεγάλες τιμές της σ.σ. U^2. Ειδικότερα, σε επίπεδο σημαντικότητας α απορρίπτουμε τη μηδενική υπόθεση αν $|U| \geq z_{\alpha/2}$ ή αν $U^2 \geq \chi^2_{1,\alpha}$. Στο σημείο αυτό επισημαίνεται ότι, παρότι η δειγματική κατανομή της σ.σ. D_n δεν μπορεί να προσεγγισθεί καλά από τη χι-τετράγωνο κατανομή με $n-1$ βαθμούς ελευθερίας, η προσέγγιση που δίνεται στο Πόρισμα 2.5.3 είναι ικανοποιητική (βλέπε, μεταξύ άλλων, Karlis and Xekalaki (2000)).

Παρατήρηση 2.5.5. *Σε επόμενη ενότητα θα δούμε ότι η σ.σ. U^2 συνδέεται με τους ελέγχους που προτάθηκαν από τους Rayner and Best (1990), όντας στην πραγματικότητα το τετράγωνο της πρώτης μη μηδενικής συνιστώσας του Neyman smooth ελέγχου της Poisson κατανομής.*

Εκτός από τη μη ικανοποιητική προσέγγιση της σ.σ. D_n από τη χι-τετράγωνο κατανομή με $n-1$ βαθμούς, ο έλεγχος με χρήση της σ.σ. D_n έχει οδηγήσει πολλές φορές, όπως αναφέρουν οι Henze and Klar (1996), στην εσφαλμένη αντίληψη ότι στην περίπτωση απόρριψης της μηδενικής υπόθεσης, η σ.σ. D_n παρέχει μία άμεση διάγνωση όσον αφορά τον τύπο της απόκλισης της από τη μηδενική υπόθεση H_0 με την έννοια ότι μεγάλες (μικρές, αντίστοιχα) τιμές της σ.σ. D_n υποδεικνύουν ότι η κατανομή του πληθυσμού έχει διακύμανση μεγαλύτερη (μικρότερη, αντίστοιχα) από τη μέση τιμή. Ωστόσο, από τη μία μεριά οι μεγάλες ή μικρές τιμές αξιολογούνται από την κατανομή της σ.σ. υπό την υπόθεση ότι η άγνωστη α.σ.κ. F είναι αυτή της κατανομής Poisson, ενώ από την άλλη μεριά ο πληθυσμιακός δείκτης διασποράς μπορεί να είναι ίσος με 1 ή η διαφορά $\delta - 1$ ίση με το μηδέν, ακόμα και αν η α.σ.κ. F δεν ειναι αυτή της Poisson. Επίσης, διαπίστωσαν ότι μία μεγάλη απόλυτη τιμή της σ.σ. U_1 μπορεί να προκύψει όχι μόνο αν ο πληθυσμός είναι υπερδιασκορπισμένος ή υποδιασκορπισμένος, αλλά και για μεγάλες τιμές της διακύμανσης σ^2. Χρησιμοποιώντας το αποτέλεσμα του Θεωρήματος 2.5.2 οι Henze and Klar (1996) ασχολήθηκαν με τον δίπλευρο έλεγχο της μηδενικής υπόθεσης ότι τα δεδομένα προέρχονται από κατανομή με μέση τιμή ίση με τη διακύμανση χρησιμοποιώντας το αποτέλεσμα που διατυπώνεται στο θεώρημα που ακολουθεί.

Θεώρημα 2.5.6. *Έστω $X_1, X_2, ..., X_n$ ένα τυχαίο δείγμα μεγέθους n από μια κατανομή που ανήκει στην οικογένεια των κατανομών για τις οποίες η μέση τιμή είναι ίση με τη διακύμανση. Τότε, για μεγάλο μέγεθος δείγματος, ισχύει ότι:*

$$S^* = \frac{U_1}{\hat{\sigma}} \xrightarrow{d} N(0,1), \tag{2.39}$$

όπου

$$\hat{\sigma}^2 = \frac{1}{2n\bar{X}^2} \sum_{i=1}^{n} [(X_i - \bar{X})^2 - X_i]^2.$$

Ισοδύναμα

$$S^* = \frac{n\bar{X}(FI-1)}{\sum_{i=1}^{n}[(X_i - \bar{X})^2 - X_i]} = \frac{\bar{X}(D_n - n)}{\sum_{i=1}^{n}[(X_i - \bar{X})^2 - X_i]} \xrightarrow{d} N(0,1). \tag{2.40}$$

Απόδειξη. Από το Θεώρημα 2.5.2 και χρησιμοποιώντας το Θεώρημα του Slutsky έχουμε ότι $\dfrac{U_1}{\hat{\sigma}} \xrightarrow{d} N(0,1)$, με $\hat{\sigma}$ έναν συνεπή εκτιμητή του σ. Ένας τέτοιος εκτιμητής προκύπτει με αντικατάσταση των πληθυσμιακών ροπών με τις αντίστοιχες δειγματικές ροπές, δηλαδή με αντικατάσταση των μ_i, $i = 2, 3, 4$, από τις ποσότητες $\widehat{\mu}_i = \dfrac{1}{n} \sum_{i=1}^{n}(X_i - \bar{X})^i$, για $i = 2, 3, 4$, και της μέσης τιμής μ με τη δειγματική μέση τιμή $\bar{X}$. Τότε, προκύπτει ότι:

$$\hat{\sigma}^2 = \frac{1}{2n\bar{X}^2} \left(\sum_{i=1}^{n}(X_i - \bar{X})^4 - 2\sum_{i=1}^{n} X_i^3 + 4\bar{X}\sum_{i=1}^{n} X_i^2 - 2n\bar{X}^3 + \sum_{i=1}^{n} X_i^2 \right),$$

ή έπειτα από λίγη άλγεβρα ότι:

$$\hat{\sigma}^2 = \frac{1}{2n\bar{X}^2} \sum_{i=1}^{n} \left[(X_i - \bar{X})^2 - X_i\right]^2.$$

και άρα καταλήξαμε στο επιθυμητό αποτέλεσμα.

□

Χρησιμοποιώντας το αποτέλεσμα του Θεωρήματος 2.5.6, η μηδενική προς έλεγχο υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας α όταν $|S^*| \geq z_{\alpha/2}$.

Παρατήρηση 2.5.7. *Εναλλακτικά, από το Θεώρημα 2.5.6 μπορεί να χρησιμοποιηθεί η σ.σ.*

$$S^{*2} = \frac{\bar{X}^2(D_n - n)^2}{\left(\sum_{i=1}^{n}\left[(X_i - \bar{X})^2 - X_i\right]\right)^2} = \frac{n^2\bar{X}^2(FI - 1)^2}{\left(\sum_{i=1}^{n}\left[(X_i - \bar{X})^2 - X_i\right]\right)^2},$$

η οποία ακολουθεί, υπό τη μηδενική υπόθεση, ασυμπτωτικά χ_1^2 κατανομή.

Χρησιμοποιώντας το αποτέλεσμα που διατυπώθηκε στο Θεώρημα 2.5.6, προκύπτει το ακόλουθο αποτέλεσμα.

Θεώρημα 2.5.8. *Έστω $X_1, X_2, ..., X_n$ ένα τυχαίο δείγμα μεγέθους n από την κατανομή Poisson με παράμετρο λ. Τότε, για μεγάλο μέγεθος δείγματος, ισχύει ότι:*

$$W = \frac{n}{2}(FI - 1)^2 = \frac{n}{2}\left(\frac{S^2}{\bar{X}} - 1\right)^2 \xrightarrow{d} \chi_1^2. \tag{2.41}$$

Επιπλέον

$$O_2 = \sqrt{\frac{n-1}{2}}(FI - 1) = \sqrt{\frac{n-1}{2}}\left(\frac{S^2}{\bar{X}} - 1\right) \xrightarrow{d} N(0,1). \tag{2.42}$$

Απόδειξη. Λαμβάνοντας υπόψη ότι:

$$\frac{1}{2n}(D_n - n)^2 \xrightarrow{d} \chi_1^2$$

, έχουμε, άμεσα, ότι

$$\frac{1}{2n}\left((n-1)FI - n\right)^2 \xrightarrow{d} \chi_1^2.$$

Επιπρόσθετα, μετά από λίγη άλγεβρα, προκύπτει ότι:

$$\frac{n-1}{2}\left(FI - \frac{n}{n-1}\right)^2 \xrightarrow{d} \chi_1^2$$

ή ισοδύναμα λόγω του Θεωρήματος του Slutsky ότι

$$\frac{n}{2}(FI-1)^2 \xrightarrow{d} \chi_1^2.$$

Επιπλέον, καθώς ισχύει ότι:

$$\sqrt{\frac{n}{2}}(FI-1) = \sqrt{\frac{n}{2}}\left(\frac{S^2}{\bar{X}} - 1\right) \xrightarrow{d} N(0,1), \tag{2.43}$$

το αποτέλεσμα για τη σ.σ. O_2 προκύπτει άμεσα με αντικατάσταση του n από το $n-1$. □

Παρατήρηση 2.5.9. *Η σ.σ. O_2 προτάθηκε από τους Potthoff and Whittinghill (1966) και Böhning (1994), στα πλαίσια του ελέγχου της Poisson κατανομής έναντι της εναλλακτικής υπόθεσης ότι τα δεδομένα προέρχονται από μια μείξη Poisson κατανομών. Μία ισοδύναμη έκφραση για τη σ.σ. O_2 είναι η εξής:*

$$O_2 = D_n \frac{1}{\sqrt{2n-1}} - \frac{\sqrt{(n-1)}}{\sqrt{2}}.$$

Επιπλέον, αξίζει να επισημανθεί ότι η στατιστική συνάρτηση W προτάθηκε από τους Rayner and McIntyre (1985) για τον έλεγχο της υπόθεσης ότι τα δεδομένα προέρχονται από την Poisson έναντι της εναλλακτικής ότι προέρχονται από την κατανομή με συνάρτηση πιθανότητας:

$$P_x(\lambda,\vartheta) = \frac{\lambda(\lambda + x\vartheta)^{x-1} e^{-(\lambda+x\vartheta)}}{x!},$$

για $x = 0, 1, 2, ..., \lambda > 0$ *και* $|\vartheta| < 1$ *και* $P_x(\lambda, \vartheta) = 0$ *αν* $x \geq m$ *και* $\lambda + m\vartheta \leq 0$.

Η προσέγγιση που δίνεται στο Θεώρημα 2.5.8 για τη σ.σ. W είναι ικανοποιητική για μεγάλα μεγέθη δείγματος και μεγάλες τιμές της παραμέτρου λ. Χρησιμοποιώντας το αποτέλεσμα του θεωρήματος έχουμε ότι σε ασυμπτωτικό επίπεδο σημαντικότητας α η κρίσιμη περιοχή είναι $W > \chi^2_{1,\alpha}$ και $|O_2| > z_{\alpha/2}$.

Παρατήρηση 2.5.10. *Η σ.σ.* U *και η σ.σ* O_2 *είναι αρκετά παρόμοιες με μόνο μικρές διαφορές στην περίπτωση των πρακτικών εφαρμογών.*

Παρατήρηση 2.5.11. *Η σ.σ.* R_a *των Baringhaus et al. (2000) συνδέεται με τον δειγματικό δείκτη διασποράς μέσω της ακόλουθης σχέσης:*

$$\lim_{a \to \infty} a^5 R_a = 6\bar{X}^2 \frac{(D_n - n)^2}{n}, \tag{2.44}$$

ενώ αποδεικνύεται (βλέπε Baringhaus et al. (2000))) ότι η σ.σ. T_a *συνδέεται με τον δειγματικό δείκτη διασποράς μέσω της ακόλουθης σχέσης:*

$$\lim_{a \to \infty} a^3 T_a = 2\bar{X}^2 \frac{(D_n - n)^2}{n}. \tag{2.45}$$

Τέλος, ένας ακόμη έλεγχος που στηρίζεται σε μία παραλλαγή της σ.σ D_n είναι αυτός που προτάθηκε στην εργασία των Kyriakoussis et al. (1998) και ο οποίος θα παρουσιαστεί σε επόμενη ενότητα.

2.5.2 Έλεγχος των Gupta et al. (1994)

Οι Gupta et al. (1994) πρότειναν έναν έλεγχο καλής προσαρμογής της κατανομής Poisson ο οποίος στηρίζεται στην ιδιότητα που δόθηκε στην Πρόταση 1.2.20. Σύμφωνα με αυτήν, αν η τ.μ. X ακολουθεί κατανομή Poisson τότε

$\mu_2\mu_4 - 3\mu_2^3 - \mu_3^2 = 0$, όπου $\mu_i = E\left(X - E(X)\right)^i$, με $i = 2, 3, 4$. Προφανώς, το εύλογο ερώτημα που προκύπτει είναι αν ισχύει το αντίστροφο. Δυστυχώς, το αντίστροφο δεν είναι πάντοτε αληθές, καθώς μπορούν να βρεθούν κι άλλες κατανομές για τις οποίες να ισχύει η παραπάνω σχέση. Για παράδειγμα, μπορούμε εύκολα να δείξουμε ότι η παραπάνω σχέση ικανοποιείται και για την κανονική κατανομή με μέση τιμή 0 και διακύμανση σ^2. Ωστόσο, οι Gupta et al. (1994) απέδειξαν ότι η κατανομή Poisson και η κανονική κατανομή με μέση τιμή 0 είναι οι μόνες κατανομές στην κλάση των λεγόμενων infinitely divisible distribution που ικανοποιούν την παραπάνω σχέση. Με τον όρο infinitely divisible distributions περιγράφονται εκείνες οι τ.μ. για τις οποίες για κάθε $n \in N$ υπάρχει μία ακολουθία ανεξάρτητων και ισόνομων τυχαίων μεταβλήτων X_1,...,X_n τέτοιες ώστε $X_1 + X_2 + ... + X_n$ και X να έχουν την ίδια κατανομή, δηλαδή $X \overset{\mathrm{d}}{=} X_1 + X_2 + \cdots + X_n$.

Στηριζόμενοι στην παραπάνω ιδιότητα, η ιδέα των Gupta et al. (1994) είναι να προταθεί ένας έλεγχος που βασίζεται στο δειγματικό ανάλογο της προαναφερθείσας διαφοράς και στο αποτέλεσμα που δίνεται στην πρόταση που ακολουθεί.

Πρόταση 2.5.12. *Έστω $X_1, X_2, ..., X_n$ ένα τυχαίο δείγμα μεγέθους n από την κατανομή Poisson με παράμετρο λ. Τότε, για μεγάλο μέγεθος δείγματος, ισχύει ότι:*

$$T = \frac{\sqrt{n}}{2} \frac{\left(m_2 m_4 - 3m_2^3 - m_3^2\right)}{\bar{X}^2\sqrt{(1 + 24\bar{X} + 6\bar{X}^2)}} \xrightarrow{d} N(0, 1), \tag{2.46}$$

όπου $m_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$, $k = 2, 3, 4$.

Απόδειξη. Το αποτέλεσμα προκύπτει με εφαρμογή αρχικά του πολυδιάστατου κεντρικού οριακού θεωρήματος για το τυχαίο διάνυσμα $\sqrt{n}(N_1, N_2, N_3, N_4)^t$, όπου $N_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$, $j = 1, ..., 4$ και έπειτα της μεθόδου δέλτα και του Θεωρήματος του Slutsky. Ειδικότερα, θεωρώντας τη σ.σ. $m_2m_4 - 3m_2^3 - m_3^2$, όπου

$m_k = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$, προκύπτει ότι, υπό τη μηδενική υπόθεση της κατανομής Poisson με παράμετρο λ, είναι:

$$\sqrt{n}\left(m_2 m_4 - 3m_2^3 - m_3^2\right) \xrightarrow{d} N\left(0, 4\lambda^4(1+24\lambda+6\lambda^2)\right).$$

Έπειτα, εφαρμόζοντας το θεώρημα του Slutsky άμεσα προκύπτει το ζητούμενο.

$\square$

Συνοψίζοντας, ο έλεγχος των Gupta et al. (1994) στηρίζεται στη σ.σ. T και απορρίπτει τη μηδενική υπόθεση για μεγάλες ή μικρές τιμές της σ.σ.ε. ή διαφορετικά με ασυμπτωτικό επίπεδο σημαντικότητας α, η κρίσιμη περιοχή απόρριψης της μηδενικής υπόθεσης θα είναι $|T| \geq z_\alpha$.

2.5.3 Έλεγχος των Kyriakoussis et al. (1998)

Οι Kyriakoussis et al. (1998) πρότειναν έναν έλεγχο καλής προσαρμογής της κατανομής Poisson ο οποίος στηρίζεται στην ιδιότητα που δόθηκε στην Πρόταση 1.2.21. Σύμφωνα με αυτήν την ιδιότητα, αν η τ.μ. X ακολουθεί κατανομή Poisson τότε $\frac{\mu_2' - \mu_1'}{(\mu_1')^2} = 1$, όπου $\mu_i' = E(X^i)$, $i = 1, 2$. Προφανώς, το εύλογο ερώτημα που προκύπτει είναι αν ισχύει το αντίστροφο. Δυστυχώς, το αντίστροφο δεν είναι πάντοτε αληθές, καθώς μπορούν να βρεθούν κι άλλες κατανομές για τις οποίες ισχύει η παραπάνω σχέση. Για παράδειγμα, μπορούμε εύκολα να δείξουμε ότι η παραπάνω σχέση ικανοποιείται και για την κανονική κατανομή με μέση τιμή 1 και διακύμανση 1, καθώς τότε $\mu_1' = 1$ και $\mu_2' = 2$. Ωστόσο, οι Kyriakoussis et al. (1998) απέδειξαν ότι η κατανομή Poisson είναι η μόνη κατανομή που ικανοποιεί την παραπάνω ιδιότητα στην κλάση των κατανομών με συνάρτηση πιθανότητας της μορφής

$$P(X=x)=\frac{a(\theta)\theta^x}{\sum_{j=0}^{m} a(j)\theta^j},\ x=0,1,...,m,$$

όπου m θετικός ακέραιος ή άπειρο και $a(x)\theta^x > 0$.

Στηριζόμενοι στην παραπάνω ιδιότητα, η ιδέα των Kyriakoussis et al. (1998) ήταν να προταθεί ένας έλεγχος που βασίζεται στο δειγματικό ανάλογο της προαναφερθείσας διαφοράς και στο αποτέλεσμα που δίνεται στην πρόταση που ακολουθεί.

Πρόταση 2.5.13. *Έστω $X_1, X_2, ..., X_n$ ένα τυχαίο δείγμα μεγέθους n από την κατανομή Poisson με παράμετρο λ. Τότε, για μεγάλο μέγεθος δείγματος, ισχύει ότι:*

$$T_p=\sqrt{\frac{n}{2}}\bar{X}\,(\hat{c}-1)\xrightarrow{d} N(0,1),$$

όπου

$$\hat{c}=\frac{\frac{1}{n}\sum_{i=1}^{n}X_i^2-\frac{1}{n}\sum_{i=1}^{n}X_i}{\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2}.$$

Απόδειξη. Το αποτέλεσμα προκύπτει με εφαρμογή του πολυδιάστατου κεντρικού οριακού θεωρήματος για το τυχαίο διάνυσμα $\sqrt{n}(M_1, M_2)^t$, όπου $M_j = \frac{1}{n}\sum_{i=1}^{n}X_i^j$, $j=1,2$ και έπειτα της μεθόδου δέλτα για $\phi(x,y)=\frac{y-x}{x^2}$ και τέλος του θεωρήματος του Slutsky. Ειδικότερα, από το πολυδιάστατο κεντρικό οριακό θεώρημα έχουμε:

$$\sqrt{n}\left[\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}X_i\\ \frac{1}{n}\sum_{i=1}^{n}X_i^2\end{pmatrix}-\begin{pmatrix}E(X)\\ E(X^2)\end{pmatrix}\right]\xrightarrow{d} N_2\left(\begin{pmatrix}0\\0\end{pmatrix},\Sigma\right),$$

όπου, υπό τη μηδενική υπόθεση $E(X)=\lambda$ και $E(X^2)=Var(X)+(E(X))^2=$

$\lambda+\lambda^2$ και έπειτα από λίγη άλγεβρα

$$\Sigma = \begin{pmatrix} Var(X) & Cov(X,X^2) \\ Cov(X^2,X) & Var(X^2) \end{pmatrix} = \begin{pmatrix} \lambda & \lambda+2\lambda^2 \\ \lambda+2\lambda^2 & 4\lambda^3+6\lambda^2+\lambda \end{pmatrix}.$$

Για την εφαρμογή της δέλτα μεθόδου με $\phi^{'}(x,y) = (\frac{x-2y}{x^3}, \frac{1}{x^2})$ έχουμε

$$\sqrt{n}\left[\phi\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i \\ \frac{1}{n}\sum_{i=1}^n X_i^2 \end{pmatrix}) - \phi\begin{pmatrix} E(X) \\ E(X^2) \end{pmatrix}\right] \xrightarrow{d} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2\right),$$

όπου έπειτα από λίγη άλγεβρα

$$\sigma^2 = \phi^{'}\big|_{(\lambda,\lambda+\lambda^2)}\Sigma\phi^{'T}\big|_{(\lambda,\lambda+\lambda^2)} = \frac{2}{\lambda^2}.$$

Συνεπώς

$$\sqrt{n}\left[\phi\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i \\ \frac{1}{n}\sum_{i=1}^n X_i^2 \end{pmatrix}) - \phi\begin{pmatrix} E(X) \\ E(X^2) \end{pmatrix}\right] \xrightarrow{d} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{2}{\lambda^2}\right),$$

ή ισοδύναμα

$$\sqrt{n}\,(\hat{c}-1) \xrightarrow{d} N\left(0, \frac{2}{\lambda^2}\right).$$

Συνδυάζοντας αυτό το αποτέλεσμα με το θεώρημα του Slutsky και το γεγονός ότι $\bar{X} \xrightarrow{p} \lambda$, συνεπάγεται ότι $T_p \xrightarrow{d} N(0,1)$. □

Συνοψίζοντας, ο έλεγχος των Kyriakoussis et al. (1998) στηρίζεται στη σ.σ. T_p και απορρίπτει τη μηδενική υπόθεση για μεγάλες ή μικρές τιμές της σ.σ.ε. ή διαφορετικά για μεγάλες κατά απόλυτο τιμές της σ.σ.ε. Δηλαδή, με επίπεδο σημαντικότητας α, η κρίσιμη περιοχή απόρριψης της μηδενικής υπόθεσης θα είναι $|T_p| \geq z_{\alpha/2}$.

Παρατήρηση 2.5.14. *Χρησιμοποιώντας ότι:*

$$\frac{1}{n}\sum_{i=0}^{n} X_i^2 = \frac{1}{n}\sum_{i=0}^{n}\left(X_i - \bar{X}\right)^2 + \bar{X}^2 = S^2 + \bar{X}^2,$$

προκύπτει ότι

$$\hat{\lambda}(\hat{c}-1) = \frac{S^2 - \bar{X}}{\bar{X}}$$

και έπειτα από λίγη άλγεβρα ότι

$$T_p = \frac{\sqrt{n}}{2}(FI - 1).$$

Άρα ο έλεγχος αυτός είναι ασυμπτωτικά ισοδύναμος με τον έλεγχο που στηρίζεται είτε στο δείκτη FI ή στον δείκτη D_n.

2.5.4 Έλεγχος των Pettigrew and Mohler (1967)

Οι Pettigrew and Mohler (1967) πρότειναν έναν έλεγχο καλής προσαρμογής της κατανομής Poisson ο οποίος στηρίζεται στην ιδιότητα που διατυπώθηκε στην Πρόταση 1.2.16, σύμφωνα με την οποία οι ημιαναλλοίωτες κάθε τάξης της κατανομής Poisson είναι ίσες με την παράμετρο λ. Στο πλαίσιο αυτό, πρότειναν τη χρήση της σ.σ.

$$Z_p = \frac{\hat{k}_p - \bar{X}}{\sqrt{Var(\hat{k}_p|\bar{X})}}, \tag{2.47}$$

για $p = 2, 3, 4$, όπου $\hat{k}_p$ είναι η δειγματική τάξης p ημιαναλλοίωτη και $Var(\hat{k}_p|\bar{X})$ η διακύμανσή της δοθέντος της δειγματικής μέσης τιμής $\bar{X}$.

Η πιο συχνά χρησιμοποιούμενη στατιστική συνάρτηση Z_p, είναι για $p = 2$, με

$\hat{k}_2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$. Τότε προκύπτει η σ.σ.:

$$Z_2 = \frac{S^2 - \bar{X}}{\sqrt{2\bar{X}(n\bar{X}-1)}}\sqrt{n(n-1)}. \tag{2.48}$$

Η ασυμπτωτική κατανομή, υπό τη μηδενική υπόθεση, της σ.σ. Z_2 δίνεται στην πρόταση που ακολουθεί (βλέπε Khamkong (2010) και Gart and Pettigrew (1970)).

Θεώρημα 2.5.15. *Έστω $X_1, X_2, ..., X_n$ ένα τυχαίο δείγμα μεγέθους n από την κατανομή Poisson με παράμετρο λ. Τότε η σ.σ.ε. Z_2 υπό τη μηδενική υπόθεση H_0 ακολουθεί τυπική κανονική κατανομή, δηλαδή*

$$Z_2 \xrightarrow{d} N(0,1).$$

Απόδειξη. Η στατιστική συνάρτηση ελέγχου μπορεί να γραφεί στην ακόλουθη μορφή

$$Z_2 = \frac{Z_2'}{\sqrt{T_{Z_2}}},$$

όπου $Z_2' = S^2 - \bar{X}$ και $T_{Z_2} = \dfrac{2\bar{X}(n\bar{X}-1)}{n(n-1)}$.

Υπό τη μηδενική υπόθεση και εφαρμόζοντας το πολυδιάστατο κεντρικό οριακό θεώρημα έχουμε ότι (βλέπε απόδειξη Πρότασης 2.5.13):

$$\sqrt{n}\left[\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n} X_i \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2\end{pmatrix} - \begin{pmatrix}E(X) \\ E(X^2)\end{pmatrix}\right] \xrightarrow{d} N_2\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \Sigma\right),$$

όπου, υπό τη μηδενική υπόθεση $E(X) = \lambda$ και $E(X^2) = \lambda + \lambda^2$ και έπειτα από λίγη άλγεβρα

$$\Sigma = \begin{pmatrix} \lambda & \lambda + 2\lambda^2 \\ \lambda + 2\lambda^2 & 4\lambda^3 + 6\lambda^2 + \lambda \end{pmatrix}.$$

Εφαρμόζοντας τη δέλτα μέθοδο με $\phi(x, y) = y - x^2 - x$ και $\phi^{'}(x, y) = (-2x - 1, y)$ έχουμε ότι:

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 - \bar{X}\right] \xrightarrow{d} N\left(0, 6\lambda^2\right),$$

ή ισοδύναμα ότι

$$Z_2' \xrightarrow{d} N\left(0, \frac{2\lambda^2}{n-1}\right).$$

Επιπλέον, ισχύει ότι:

$$E\left[\bar{X}^2 - \frac{\bar{X}}{n}\right] = \lambda^2.$$

Λαμβάνοντας υπόψη τα παραπάνω και κάνοντας χρήση του θεωρήματος Slutsky για σταθερή παράμετρο λ και $n \to \infty$, προκύπτει το ζητούμενο. $\square$

Από το αποτέλεσμα του προηγούμενου θεωρήματος προκύπτει ότι σε ασυμπτωτικό επίπεδο σημαντικότητας α η περιοχή απόρριψης της μηδενικής υπόθεσης είναι $|Z_2| > z_{\alpha/2}$.

Από συγκριτικές μελέτες προσομοίωσης έχει προκύψει ότι οι έλεγχοι καλής προσαρμογής που έχουν προταθεί στις εργασίες των Gupta et al. (1994), Kyriakoussis et al. (1998) και Pettigrew and Mohler (1967), δεν είναι ανταγωνιστικοί σε σχέση με άλλους διαθέσιμους ελέγχους, καθώς παρουσιάζουν τη μικρότερη ισχύ σε όλες τις περιπτώσεις εναλλακτικών κατανομών που έχουν θεωρηθεί. Αυτός είναι και ο λόγος που δεν θα συμπεριληφθούν στη συγκριτική μελέτη του επόμενου κεφαλαίου.

2.6 Έλεγχοι βασισμένοι στην επαρκή στατιστική συνάρτηση

Πλην του τροποποιημένου Kolmogorov-Smirnov τύπου έλεγχο που παρουσιάστηκε στην Παρατήρηση 2.2.2 και ο οποίος αξιοποιεί την ιδιότητα που δόθηκε στη σχέση (1.23), σύμφωνα με την οποία η δεσμευμένη κατανομή του τυχαίου διανύσματος $(X_1, X_2, ..., X_n)$ δοθέντος της επαρκούς στατιστικής συνάρτησης $\sum_{i=1}^{n} X_i = t$ είναι πολυωνυμική $M(t, 1/n, ..., 1/n)$, στη βιβλιογραφία έχει εμφανισθεί ένας ακόμη έλεγχος. Αντικείμενο μελέτης αυτής της ενότητας είναι η παρουσίαση και επεξήγηση αυτού του ελέγχου. Ειδικότερα, στην ενότητα αυτή θα παρουσιασθεί ο έλεγχος καλής προσαρμογής της Poisson κατανομής που προτάθηκε από τους González-Barrios et al. (2006). Στο σημείο αυτό αξίζει να επισημανθεί ότι μία γενική μέθοδος κατασκευής ελέγχων καλής προσαρμογής με τη δεσμευμένη σ.π.π. ή σ.π. των $X_1, X_2, ..., X_n$ δοθέντος της τιμής της επαρκούς σ.σ. έχει προταθεί από τους Lockhart (2012), ενώ ο πρώτος που χρησιμοποίησε αυτήν την ιδέα ήταν ο Fisher (1950).

Στο παραπάνω πλαίσιο, σύμφωνα με τους González-Barrios et al. (2006) αρχικά δοθέντος των τιμών (t, n) βρίσκουμε όλες τις l το πλήθος διαφορετικές διατάξεις (με πλήθος στοιχείων n) που είναι τέτοιες $(y_n, y_{n-1}, ..., y_1)$, με $y_n \geq y_{n-1} \geq ... \geq y_1$ και $\sum_{i=1}^{n} y_i = t$. Έπειτα, για κάθε μία υπολογίζουμε την πιθανότητα $p = P(X_1 = y_1, X_2 = y_2, \cdots, X_n = y_n | T = t)$ από τη σχέση:

$$\begin{aligned} p &= \binom{t}{y_n}\binom{t-y_n}{y_{n-1}}\cdots\binom{t-y_n-...-y_2}{y_1}\left(\frac{1}{n}\right)^t \\ &= \frac{t!}{y_1!...y_n!}\left(\frac{1}{n}\right)^t. \end{aligned}$$

Στο επόμενο βήμα οι πιθανότητες αυτές τοποθετούνται σε φθίνουσα διάταξη και αθροίζονται μέχρι να βρούμε την τιμή εκείνη που είναι πιο κοντινή στην τιμή $1-\alpha$, όπου α είναι το επίπεδο σημαντικότητας του ελέγχου. Όσες διατάξεις δεν έχουν συμπεριληφθεί στο προαναφερθέν άθροισμα αποτελούν την κρίσιμη περιοχή του ελέγχου.

Για την κατανόηση των παραπάνω ακολουθεί το αριθμητικό παράδειγμα που παρουσιάσθηκε στην εργασία των González-Barrios et al. (2006).

Παράδειγμα κατανόησης.

Έστω $n = 10$ και $t = 8$, δηλαδή ότι έχουμε 10 το πλήθος παρατηρήσεις με το άθροισμα των τιμών τους να είναι ίσο με 8. Τότε προκύπτει ότι υπάρχουν $l = 22$ το πλήθος διατάξεις της μορφής $(y_n, y_{n-1}, ..., y_1)$, με $y_n \geq y_{n-1} \geq ... \geq y_1$ και $\sum_{i=1}^{n} y_i = 8$, οι οποίες δίνονται στην πρώτη στήλη του Πίνακα 2.1. Για κάθε μία τέτοια διάταξη υπολογίζουμε την πιθανότητα εμφάνισής της χρησιμοποιώντας τη σχέση που δόθηκε παραπάνω για την πιθανότητα $P(X_1 = y_1, X_2 = y_2, \cdots, X_n = y_n|T = t)$ για $n = 10$ και $t = 8$. Οι τιμές αυτών των πιθανοτήτων δίνονται στη δεύτερη στήλη του Πίνακα 2.1. Ένας ενδεικτικός υπολογισμός είναι ο ακόλουθος:

$$P(5, 2, 1, 0, 0, 0, 0, 0, 0, 0) = \binom{8}{5}\binom{8-5}{2}\binom{8-5-2}{1}\left(\frac{1}{10}\right)^8.$$

Το επόμενο βήμα για την υλοποίηση του ελέγχου είναι να διαταχθούν οι πιθανότητες σε φθίνουσα σειρά, ενώ ταυτόχρονα υπολογίζεται σε μία ξεχωριστή στήλη το άθροισμα αυτών των πιθανοτήτων. Κατά αυτόν τον τρόπο προκύπτει ο Πίνακας 2.6.

Επομένως, παρατηρώντας ότι το άθροισμα των πιθανοτήτων για τη διάταξη $(3, 2, 2, 1, 0, 0, 0, 0, 0, 0)$ είναι 0.910224, προκύπτει ότι η κρίσιμη περιοχή επιπέδου σημαντικότητας 1-0.910224 =0.089776 είναι το συμπλήρωμα του ακόλουθου συνόλου διατάξεων:

$$R = \{(2,2,1,1,1,1,0,0,0,0),(3,2,1,1,1,0,0,0,0,0),$$
$$(2,1,1,1,1,1,1,0,0,0),(2,2,2,1,1,0,0,0,0,0),$$
$$(3,1,1,1,1,1,0,0,0,0),(3,2,2,1,0,0,0,0,0,0)\}.$$

Παρατήρηση 2.6.1. *Από τα παραπάνω παράδειγμα είναι αντιληπτό ότι ο έλεγχος είναι ακριβής και όχι ασυμπτωτικός. Ωστόσο, καθώς στηριζόμαστε στην πολυωνυμική κατανομή, είναι προφανές ότι μπορεί να εφαρμοστεί, όπως όλοι οι έλεγχοι των οποίων οι σ.σ. ακολουθούν διακριτή κατανομή, για συγκεκριμένες τιμές του επιπέδου σημαντικότητας* α. *Επίσης, είναι προφανές ότι για μεγάλες τιμές των* n *και* t *ο έλεγχος απαιτεί εξαντλητική απαρίθμηση με συνέπεια να έχουμε μεγάλο υπολογιστικό χρόνο. Για τον λόγο αυτόν προτείνεται να χρησιμοποιείται για μικρά σε μέγεθος δείγματος, λαμβάνοντας υπόψη και ότι είναι ακριβής έλεγχος σε αντίθεση με άλλους που είναι ασυμπτωτικοί. Τέλος, για μία τροποποίηση του ελέγχου παραπέμπουμε στους Beltrán-Beltrán and O'Reilly (2019). Ωστόσο, ο έλεγχος αυτός δεν παρουσιάζεται καθώς οι ίδιοι ερευνητές που τον πρότειναν κατέληξαν στο συμπέρασμα ότι, παρότι είναι καλύτερος από αυτών των González-Barrios et al. (2006), δεν είναι ανταγωνιστικός σε σχέση με τους κλασικούς ελέγχους που παρουσιάστηκαν σε προηγούμενες ενότητες ή θα παρουσιαστούν σε επόμενες.*

Διατάξεις	Πιθανότητα υπό την H_0
(8,0,0,0,0,0,0,0,0,0)	0.00000001
(7,1,0,0,0,0,0,0,0,0)	0.0000072
(6,2,0,0,0,0,0,0,0,0)	0.0000252
(6,1,1,0,0,0,0,0,0,0)	0.0002016
(5,3,0,0,0,0,0,0,0,0)	0.0000504
(5,2,1,0,0,0,0,0,0,0)	0.0012096
(5,1,1,1,0,0,0,0,0,0)	0.0028224
(4,4,0,0,0,0,0,0,0,0)	0.0000315
(4,3,1,0,0,0,0,0,0,0)	0.002016
(4,2,2,0,0,0,0,0,0,0)	0.001512
(4,2,1,1,0,0,0,0,0,0)	0.021168
(4,1,1,1,1,0,0,0,0,0)	0.021168
(3,3,2,0,0,0,0,0,0,0)	0.002016
(3,3,1,1,0,0,0,0,0,0)	0.014112
(3,2,2,1,0,0,0,0,0,0)	0.042336
(3,2,1,1,1,0,0,0,0,0)	0.169344
(3,1,1,1,1,1,0,0,0,0)	0.084672
(2,2,2,2,0,0,0,0,0,0)	0.005292
(2,2,2,1,1,0,0,0,0,0)	0.127008
(2,2,1,1,1,1,0,0,0,0)	0.31752
(2,1,1,1,1,1,1,0,0,0)	0.169344
(1,1,1,1,1,1,1,1,0,0)	0.018144

Πίνακας 2.1: Διατάξεις της μορφής $(y_{10}, ..., y_1)$, με $y_{10} \geq y_9 \geq ... \geq y_1$ και $\sum_{i=1}^{10} y_i = 8$ και πιθανότητες εμφάνισής τους υπό την H_0.

Διατάξεις	Πιθανότητες	Άθροισμα πιθανοτήτων
(2,2,1,1,1,1,0,0,0,0)	0.31752	0.31752
(3,2,1,1,1,0,0,0,0,0)	0.169344	0.486864
(2,1,1,1,1,1,1,0,0,0)	0.169344	0.656208
(2,2,2,1,1,0,0,0,0,0)	0.127008	0.783216
(3,1,1,1,1,1,0,0,0,0)	0.084672	0.867888
(3,2,2,1,0,0,0,0,0,0)	0.042336	0.910224
(4,1,1,1,1,0,0,0,0,0)	0.021168	0.931392
(4,2,1,1,0,0,0,0,0,0)	0.021168	0.952560
(1,1,1,1,1,1,1,1,0,0)	0.018144	0.970704
(3,3,1,1,0,0,0,0,0,0)	0.014112	0.984816
(2,2,2,2,0,0,0,0,0,0)	0.005292	0.990108
(5,1,1,1,0,0,0,0,0,0)	0.0028224	0.9929304
(3,3,2,0,0,0,0,0,0,0)	0.002016	0.9949464
(4,3,1,0,0,0,0,0,0,0)	0.002016	0.9969624
(4,2,2,0,0,0,0,0,0,0)	0.001512	0.9984744
(5,2,1,0,0,0,0,0,0,0)	0.0012096	0.9996840
(6,1,1,0,0,0,0,0,0,0)	0.0002016	0.9998856
(5,3,0,0,0,0,0,0,0,0)	0.0000504	0.999936
(4,4,0,0,0,0,0,0,0,0)	0.0000315	0.9999675
(6,2,0,0,0,0,0,0,0,0)	0.0000252	0.9999927
(7,1,0,0,0,0,0,0,0,0)	0.0000072	0.9999999
(8,0,0,0,0,0,0,0,0,0)	0.00000001	1.0000000

Πίνακας 2.2: Πίνακας διατάξεων της μορφής $(y_{10}, ..., y_1)$, με $y_{10} \geq y_9 \geq ... \geq y_1$ και $\sum_{i=1}^{10} y_i = 8$, πιθανοτήτων εμφάνισής τους σε φθίνουσα διάταξη και αθροισμάτων αυτών.

2.7 Smooth tests

Στη γενική περίπτωση, οι έλεγχοι αυτής της κατηγορίας είναι έλεγχοι καλής προσαρμογής που κατασκευάζονται ορίζοντας αρχικά μία τάξης k εναλλακτική στην υπό τη μηδενική υπόθεση σ.π. ή σ.π.π. της μορφής:

$$f(x;\theta,\lambda) = C(\theta,\lambda)exp\left(\sum_{i=1}^{k}\theta_i h_i(x;\lambda)\right)f(x,\lambda), \tag{2.49}$$

όπου $\theta = (\theta_1, ..., \theta_k)$, $f(x,\lambda)$ είναι η σ.π.π. ή η σ.π. υπό τη μηδενική υπόθεση, $C(\theta,\lambda)$ μία σταθερά και $h_i(x;\lambda)$ είναι ένα σύνολο ορθοκανονικών συναρτήσεων στην $f(x,\lambda)$ που είναι τέτοιες ώστε για οποιαδήποτε τιμή των x και λ να ισχύει ότι:

$$E_0 \{h_i(x;\lambda)h_j(x;\lambda)\} = \delta_{ij}, \ 0 \leq i, j \leq k,$$

με $\delta_{ij} = 1$ όταν $i = j$ και μηδέν διαφορετικά, ενώ E_0 συμβολίζει τη μέση τιμή υπό την κατανομή με σ.π.π. ή σ.π. $f_0(x, \lambda)$.

Στη συνέχεια, ο έλεγχος καλής προσαρμογής ανάγεται στον έλεγχο της υπόθεσης ότι το k-διάστατο διάνυσμα θ είναι ίσο με το μηδέν έναντι της εναλλακτικής ότι η υπόθεση αυτή απορρίπτεται. Δηλαδή ανάγεται στον έλεγχο της υπόθεσης ότι τα δεδομένα περιγράφονται ικανοποιητικά από την $f(x, \lambda)$ έναντι της εναλλακτικής ότι τα δεδομένα περιγράφονται καλύτερα από μία κατανομή που είναι μέλος της $f(x; \theta, \lambda)$. Η σ.σ.ε. για αυτήν την υπόθεση προσδιορίζεται ως ένα score test. Υπό αυτό το σκεπτικό, οι έλεγχοι αυτοί είναι βέλτιστοι για μεγάλα μεγέθη δείγματος, ενώ καλούνται smooth tests καθώς κατασκευάζονται για να έχουν καλή ισχύ έναντι εναλλακτικών των οποίων οι σ.π. ή οι σ.π.π. διαφέρουν smoothly από την κατανομή υπό τη μηδενική υπόθεση, όπως αυτό ορίζεται στην σχέση 2.49. Ο πρώτος τέτοιος έλεγχος παρουσιάστηκε από τον Neyman (1937) για τον έλεγχο καλής προσαρμογής της ομοιόμορφης κατανομής, ενώ καθοριστική συμβολή στην απόκτηση μεγαλύτερης δημοφιλίας είχε το έργο των Rayner and Best (1990) με μία σειρά σχετικών εργασιών, αλλά και με μία σχετική μονογραφία (βλέπε, Rayner and Best (1988), Rayner et al. (2009) και τις εκεί αναφορές). Στην ενότητα αυτή το ενδιαφέρον θα επικεντρωθεί στη συνοπτική περιγραφή του ελέγχου στην ειδική περίπτωση της κατανομής Poisson.

Στην περίπτωση της Poisson κατανομής η εναλλακτική συνάρτηση πιθανότητας τάξης k που θεωρούν οι Rayner and Best (1988) είναι αυτή που παρουσιάστηκε από τον Charlier (1905), και η οποία δίνεται από τη σχέση:

$$f(x;\theta,\lambda) = C(\theta,\lambda)exp\left(\sum_{i=1}^{k}\theta_i h_i(x;\lambda)\right) f_0(x,\lambda), \qquad (2.50)$$

όπου $C(\theta,\lambda)$ σταθερά και $h_i(x;\lambda)$ είναι ο i-οστός όρος του λεγόμενου πολυωνύμου Charlier ή πολυωνύμου Poisson-Charlier που ορίζεται από τη σχέση:

$$h_i(x;\lambda) = \sqrt{\left(\frac{\lambda^i}{i!}\right)}\sum_{u=0}^{i}(-1)^{i-u}\binom{i}{u}u!\frac{1}{\lambda^u}\binom{x}{u}.$$

Αποδεικνύεται εύκολα ότι για $i=0$ είναι $h_0(x;\lambda)=1$ για οποιαδήποτε τιμή των x και λ, ενώ

$$h_1(x;\lambda) = \sqrt{\lambda}\left(\frac{x}{\lambda}-1\right). \qquad (2.51)$$

Επιπλέον (βλέπε Karlis and Xekalaki (2000)) μια χρήσιμη αναδρομική σχέση για τον προσδιορισμό των $h_i(x,\lambda)$ είναι η ακόλουθη:

$$h_i(x;\lambda)\sqrt{(i\lambda)} = (i+x-\lambda-1)h_{i-1}(x-1;\lambda) - \sqrt{\left(\frac{i-1}{\lambda}\right)}(x-1)h_{i-2}(x-2;\lambda).$$

Επιπρόσθετες σχέσεις προσδιορισμού των h_j, για $j = 1,...,4$, δίνονται στην εργασία των Best and Rayner (1999).

Με το παραπάνω σκεπτικό, οι Rayner and Best (1988) προτείνουν ένα smooth test για τον έλεγχο καλής προσαρμογής της Poisson χρησιμοποιώντας τη σ.σ.:

$$S_k = \sum_{i=2}^{k} V_i^2, \qquad (2.52)$$

όπου οι συνιστώσες V_i, $i = 2, ..., k$, προσδιορίζονται από τη σχέση:

$$V_i = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} h_i(X_j; \hat{\lambda} = \bar{X}).$$

Μία πρώτη παρατήρηση είναι ότι στην ελεγχοσυνάρτηση δεν περιλαμβάνεται η συνιστώσα V_1, καθώς, λαμβάνοντας υπόψη τη σχέση (2.51), είναι

$$V_1 = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \sqrt{\bar{X}} \left(\frac{X_j}{\bar{X}} - 1 \right) = \sqrt{\frac{1}{n\bar{X}}} \sum_{j=1}^{n} (X_j - \bar{X}) = 0.$$

Παρατήρηση 2.7.1. *Για $k = 2$, μετά από λίγη άλγεβρα προκύπτει ότι*

$$V_2 = \frac{1}{\sqrt{2n}} \sum_{j=1}^{n} \frac{(X_j - \bar{X})^2 - X_j}{\bar{X}} = \frac{1}{\sqrt{2n}} \frac{\sum_{j=1}^{n} (X_j - \bar{X})^2 - n\bar{X}}{\bar{X}} = \frac{D_n - n}{\sqrt{2n}},$$

όπου D_n ο Poisson δείκτης διασποράς. Άρα ο έλεγχος που στηρίζεται στη σ.σ. V_2 είναι ασυμπτωτικά ισοδύναμος με αυτόν που στηρίζεται στη σ.σ. D_n.

Παρατήρηση 2.7.2. *Αξίζει να επισημανθεί ότι οι Ledwina and Wyłupek (2017) θεώρησαν αντί της εναλλακτικής οικογένειας κατανομών της σχέσης (2.50) την οικογένεια κατανομών με σ.π. που δίνεται από τη σχέση:*

$$f(x; \theta, \lambda) = \frac{\left(1 + \sum_{i=1}^{k} \theta_i h_i(x; \lambda)\right)^2}{1 + \sum_{i=1}^{k} \theta_i^2} f(x, \lambda). \tag{2.53}$$

Παρόλα αυτά κατέληξαν στην ίδια σ.σ. με αυτήν των Rayner and Best (1988).

Η ασυμπτωτική κατανομή των V_j υπό τη μηδενική υπόθεση είναι τυπική κανονική (βλέπε Best and Rayner (1999) και τις εκεί αναφορές), ενώ η ασυμπτωτική κατανομή της σ.σ. S_k είναι χι-τετράγωνο με $k - 1$ βαθμούς ελευθερίας, με τη μηδενική υπόθεση να απορρίπτεται για μεγάλες τιμές της σ.σ. S_k.

Οι Rayner et al. (2009) μελέτησαν την ειδική περίπτωση της σ.σ.ε. για $k = 5$ μέσω προσομοιώσεων και κατέληξαν στο συμπέρασμα ότι η δειγματική κατανομή του S_5 εξαρτάται από το n και την παράμετρο λ και πρότειναν για $\alpha = 0.05$ την ακόλουθη διορθωμένη κρίσιμη τιμή

$$\chi^2_{4,\alpha}\left(1 - \frac{1.14}{\sqrt{n\bar{X}}}\right)$$

για τις περιπτώσεις που $n\bar{X} > 10$. Η μη ικανοποιητική προσέγγιση της χι-τετράγωνο κατανομής επιβεβαιώθηκε, για διάφορες τιμές του k, και από την εκτενή μελέτη προσομοίωσης των Ledwina and Wyłupek (2017). Ειδικότερα, από τη μελέτη αυτή προέκυψε ότι για μικρά σε μέγεθος δείγματα, για παράδειγμα $n = 50$, η προσέγγιση της χι-τετράγωνο κατανομής δεν είναι καθόλου ικανοποιητική. Επιπλεόν, οι ίδιοι συγγραφείς κατέληξαν στο συμπέρασμα ότι η επιλογή του k επιδρά στην απόδοση του ελέγχου ως προς την ισχύ του ελέγχου. Επομένως, το εύλογο ερώτημα που χρήζει απάντησης είναι ποιος είναι ο αριθμός k που πρέπει να επιλέγεται. Στη βιβλιογραφία έχουν εμφανισθεί διάφορες απαντήσεις σε αυτό το ερώτημα. Ενδεικτικά, οι Rayner et al. (2009) προτείνουν να αγνοούμε τις συνιστώσες V_j οι οποίες είναι μικρότερες κατά απόλυτο τιμή από την τιμή 2, ενώ οι Ledwina and Wyłupek (2017) προτείνουν να επιλέγεται το k μέσω του ακόλουθου κανόνα:

$$S = min\{1 \leq k \leq d(n), W_k - k\log n \geq W_j - j\log n, 1 \leq j \leq d(n)\},$$

όπου $\{d(n)\}_{n=1}^{\infty}$ μία φθίνουσα ακολουθία φυσικών αριθμών.

2.8 Γραφικοί έλεγχοι

Στις προηγούμενες ενότητες το ενδιαφέρον επικεντρώθηκε στην παρουσίαση ελέγχων καλής προσαρμογής της κατανομής Poisson μέσω κατάλληλων στατιστικών ελέγχων. Ωστόσο, στη στατιστική βιβλιογραφία έχουν παρουσιαστεί και μεθοδολογίες που μας προσφέρουν έναν απλό και αποτελεσματικό τρόπο αξιολόγησης της προσαρμογής του μοντέλου της Poisson κατανομής μέσω μιας γραφικής παράστασης (βλέπε Gan et al. (1991)). Αν και οι γραφικοί τρόποι ελέγχου καλής προσαρμογής θεωρούνται λιγότερο επίσημοι και περισσότερο υποκειμενικοί από τους τρόπους που συζητήθηκαν στις προηγούμενες ενότητες αυτές του κεφαλαίου, έχει προταθεί (βλέπε, μεταξύ άλλων, D'Agostino (1986)) οι μεθοδολογίες των προηγούμενων ενοτήτων να συμπληρώνονται με γραφικούς τρόπους ελέγχου καλής προσαρμογής. Η πρόταση αυτή αιτιολογείται καθώς οι γραφικοί έλεγχοι αποτελούν διερευνητικά εργαλεία για την κατανόηση σχέσεων που μπορούν να υπάρχουν στα δεδομένα του δείγματος και οι οποίες σχέσεις δεν αναδεικνύονται από τους στατιστικούς ελέγχους. Για τον λόγο αυτόν, στην ενότητα αυτή παρουσιάζονται δημοφιλείς γραφικοί τρόποι ελέγχου καλής προσαρμογής της Poisson κατανομής.

Γράφημα εκτιμώμενης α.σ.κ. και ε.α.σ.κ.

Ο πιο απλός και εύκολος γραφικός τρόπος ελέγχου της καλής προσαρμογής της Poisson κατανομής στηρίζεται στην εμπειρική αθροιστική συνάρτηση κατανομής και την ιδιότητά της ότι συγκλίνει στην πραγματική αθροιστική συνάρτηση κατανομής. Ειδικότερα, στο ίδιο ορθογώνιο σύστημα αξόνων αναπαρίστανται η εμπειρική αθροιστική συνάρτηση κατανομής και η εκτιμώμενη αθροιστική συνάρτηση κατανομής υπό τη μηδενική υπόθεση. Δηλαδή, τα βήματα που ακολουθούνται είναι:

Βήμα 1. Διατάσσονται σε αύξουσα τάξη μεγέθους οι διαθέσιμες παρατηρήσεις. Έστω $x_{(1)}$,...,$x_{(n)}$ οι διατεταγμένες παρατηρήσεις.

Βήμα 2. Υπολογίζονται οι τιμές $F_n(x_{(i)})$, για $i = 1, ..., n$.

Βήμα 3. Υπολογίζεται ο εκτιμητής $\bar{X}$ και οι τιμές $F_0(x_{(i)}, \lambda = \bar{X})$, για $i = 1, ..., n$.

Βήμα 4. Στο ίδιο ορθογώνιο σύστημα αξόνων αναπαρίστανται τα ζεύγη παρατηρήσεων $\big(x_{(i)}, F_n(x_{(i)})\big)$ και $\big(x_{(i)}, F_0(x_{(i)}, \lambda = \bar{X})\big)$, για $i = 1, ..., n$.

Όσο πιο πολύ απέχουν τα δύο γραφήματα που προκύπτουν από την παραπάνω μέθοδο τόσο περισσότερο ενισχύεται η υπόθεση ότι τα δεδομένα αποκλίνουν από την κατανομή Poisson.

Q-Q (quantile-quantile) γράφημα

Ένας άλλος γραφικός τρόπος ελέγχου είναι το Q-Q (quantile-quantile) γράφημα το οποίο συγκρίνει τα ποσοστιαία σημεία (quantile) του δείγματος έναντι των πληθυσμιακών ποσοστιαίων σημείων της κατανομής Poisson. Αν τα σημεία είναι κοντά σε ευθεία γραμμή δεν υπάρχει ένδειξη για απόκλιση από την κατανομή Poisson. Η κατασκευή ενός Q-Q γραφήματος διέπεται από τα ακόλουθα βήματα:

Βήμα 1. Διατάσσονται σε αύξουσα τάξη μεγέθους οι διαθέσιμες παρατηρήσεις. Έστω $x_{(1)}$,...,$x_{(n)}$ οι διατεταγμένες παρατηρήσεις.

Βήμα 2. Στο σημείο $x_{(i)}$ ανατίθεται, ως διόρθωση συνέχειας, το δειγματικό ποσοστιαίο σημείο $p_i = \dfrac{i - 0.5}{n}$ (αντί του i/n).

Βήμα 3. Έστω q_1,...,q_n τα σημεία για τα οποία ισχύει ότι $F_0(q_i, \hat{\lambda}) = p_i$.

Βήμα 4. Σε ένα σύστημα ορθογωνίων αξόνων αναπαρίστανται τα ζεύγη παρατηρήσεων $(q_i, x_{(i)})$, για $i = 1, ..., n$.

Αξίζει να επισημανθεί ότι στη βιβλιογραφία έχουν αναφερθεί κι άλλες επιλογές των p_i πέραν αυτής που αναφέρθηκε παραπάνω, όπως είναι οι $p_i = \frac{i}{n+1}$ ή $p_i = \frac{i-0.375}{n+0.25}$ ή $p_i = \frac{i-0.3}{n+0.4}$ κ.ά.

Παρατήρηση 2.8.1. *Πολλές φορές συμπληρώνουμε και επιβεβαιώνουμε τις πληροφορίες που παρέχονται στο Q-Q γράφημα, μέσω του υπολογισμού του συντελεστή προσδιορισμού R^2 που ορίζεται από τη σχέση:*

$$R^2 = \frac{(\sum_{i=1}^{n}(x_{(i)} - \bar{x})(q_i - \bar{q}))^2}{\sum_{i=1}^{n}(x_{(i)} - \bar{x})^2 \sum_{i=1}^{n}(q_i - \bar{q})^2},$$

όπου $x_{(i)}$ οι διατεταγμένες παρατηρήσεις και q_i τα αναμενόμενα ποσοστιαία σημεία από την Poisson κατανομή με εκτιμώμενη παράμετρο $\widehat{\lambda} = \bar{x}$. Το R^2 δεν είναι τίποτε άλλο παρά το τετράγωνο του δειγματικού συντελεστή συσχέτισης των ζευγών $(x_{(i)}, q_i)$, $i = 1, ..., n$, με πιθανές τιμές που κυμαίνονται από 0 έως 1, δηλαδή $0 \leq R^2 \leq 1$. Το συμπέρασμα που μπορεί να προκύψει είναι ότι, όσο πιο κοντά στο 1 είναι η τιμή του R^2, τόσο ισχυρότερος είναι ο βαθμός γραμμικής συσχέτισης μεταξύ των $x_{(i)}$ και q_i. Προφανώς, η μηδενική υπόθεση απορρίπτεται για μικρές τιμές της σ.σ. και οι κρίσιμες τιμές προσδιορίζονται μέσω παραμετρικού bootstrap.

Poissonness γράφημα

Ένα παρόμοιο διάγραμμα ελέγχου με το Q-Q γράφημα προτάθηκε από τον Hoaglin (1980). Το Poissonness γράφημα, όπως λέγεται, βασίζεται στην υπόθεση ότι για κάποια σταθερή τιμή της παραμέτρου λ, κάθε παρατηρούμενη συχνότητα, έστω n_k, με $\sum_k n_k = n$, ισούται με την αναμενόμενη συχνότητα $nf_0(k, \lambda) = n\frac{\lambda^k e^{-\lambda}}{k!}$, $k = 0, 1, 2, ...$

Ισοδύναμα θα πρέπει ο λογάριθμος κάθε παρατηρούμενης συχνότητας να ισούται με τον λογάριθμο της αναμενόμενης συχνότητας ή ισοδύναμα παίρνοντας λογάριθμο και στα δύο μέλη της παραπάνω σχέσης προκύπτει, έπειτα από λίγη άλγεβρα, ότι:

$$\log\left(\frac{n_k k!}{n}\right) = k\log(\lambda) - \lambda.$$

Επομένως, προκύπτει ότι η γραφική παράσταση του πρώτου μέλους $\log\left(\frac{n_k k!}{n}\right)$ συναρτήσει του k καταλήγει να είναι μία ευθεία γραμμή με κλίση $\log(\lambda)$ και σταθερό όρο $-\lambda$.

Η ιδιότητα αυτή χρησιμοποιείται με παρόμοιο τρόπο, όπως στην περίπτωση του Q-Q γραφήματος, και η ευθεία γραμμή της Poisson που σχηματίζεται, χρησιμοποιείται για τον έλεγχο της μηδενικής υπόθεσης αν τα δεδομένα προέρχονται ή όχι από την Poisson κατανομή. Στην περίπτωση που για ένα σύνολο δεδομένων, το Poissonness γράφημα είναι ευθεία γραμμή, χρησιμοποιώντας είτε τον εκτιμητή μέγιστης πιθανοφάνειας $\widehat{\lambda}$, είτε τον εκτιμητή που προέρχεται από την κλίση της Poissonness γραμμής, μπορούμε να προσδιορίσουμε αν τα δεδομένα μας προέρχονται από την κατανομή Poisson. Για περαιτέρω πληροφορίες σχετικά με το Poissonness διάγραμμα, παραπέμπουμε τον\την αναγνώστη\στρια στις εργασίες των Hoaglin (1980), Hoaglin et al. (2011) και τις εκεί αναφορές.

Γράφημα της ε.π.σ.

Ένας άλλος γραφικός τρόπος διερεύνησης αν τα δεδομένα προέρχονται από την κατανομή Poisson ή κάποια άλλη διακριτή κατανομή είναι αυτός που περιγράφεται στη συνέχεια (βλέπε Nakamura and Pérez-Abreu (1993)). Ειδικότερα, καθώς ο λογάριθμος της πιθανογεννήτριας συνάρτησης της κατανομής Poisson με παράμετρο λ είναι ίσος με $\lambda(t-1)$, έχουμε ότι πρόκειται για ευθεία γραμμή. Από την άλλη μεριά, όπως επισημαίνεται από τους Nakamura and Pérez-Abreu (1993), ο λογάριθμος της πιθανογεννήτριας της διωνυμικής κατανομής είναι μία κοίλη συνάρτηση, ενώ ο λογάριθμος της πιθανογεννήτριας της αρνητικής διωνυμικής κατανομής ή της μείξης κατανομών Poisson έχει πάντοτε σχήμα κυρτής συνάρτησης. Επιπλέον, στην περίπτωση της περικομμένης στο 0 κατανομής Poisson η συμπεριφορά του λογαρίθμου της πιθανογεννήτριας κοντά στο 0 είναι όπως αυτή της συνάρτησης $\log(t)$, ενώ για τιμές κοντά στο 1 μοιάζει με ευθεία γραμμή. Καθώς η πιθανογεννήτρια συνάρτηση είναι άγνωστη, καταφεύγουμε στη γραφική αναπαράσταση του λογαρίθμου της εμπειρικής πιθανογεννήτριας συνάρτησης και την εξαγωγή συμπερασμάτων από αυτήν σύμφωνα με τις παραπάνω κατευθυντήριες οδηγίες.

Άλλοι γραφικοί έλεγχοι

Άλλα διαγράμματα για τον ελέγχο της Poisson κατανομής που έχουν προταθεί κατά καιρούς, βασίζονται στον λόγο διαδοχικών παρατηρούμενων συχνοτήτων. Συγκεκριμένα, ο Dubey (1966) πρότεινε την απεικόνιση του λόγου $\dfrac{f_x}{f_{x+1}}$ συναρτήσει του x, για $x = 0, 1, 2, \ldots$. Παρατηρούμε ότι στην περίπτωση της κατανομής Poisson έχουμε ότι η γραφική παράσταση που προκύπτει είναι μία ευθεία γραμμή με κλίση και σταθερό όρο ίσο με $\dfrac{1}{\lambda}$, καθώς

$$\frac{f_0(x,\lambda)}{f_0(x+1,\lambda)} = \frac{e^{-\lambda}\lambda^x}{x!}\frac{(x+1)!}{e^{-\lambda}\lambda^{x+1}} = \frac{x+1}{\lambda}.$$

Αργότερα, ο Ord (1967) διευρύνοντας τη σκέψη του Dubey (1966) υπέδειξε ότι μία καλύτερη διαγνωστική γραφική παράσταση είναι αυτή που προέρχεται από την $u_x = \dfrac{xf_x}{f_{x-1}}$ και η οποία στην ειδική περίπτωση της κατανομής Poisson δίνεται από τη σχέση $u_x = \lambda + 0x$, όπου λ η παράμετρος της κατανομής Poisson. Παρότι οι αντίστοιχες δειγματικές ποσότητες δεν ικανοποιούν πλήρως και με μεγάλη ακρίβεια τις παραπάνω σχέσεις, ο Ord (1967) υποστήριξε, ότι οι δειγματοληπτικές γραφικές παραστάσεις αυτού του τύπου, μπορούν να δώσουν μία ικανοποιητική ένδειξη για το αν υπάρχουν ενδείξεις αποκλίσεων από την κατανομή Poisson.

ΚΕΦΑΛΑΙΟ 3

Συγκριτική μελέτη

Στο προηγούμενο κεφάλαιο παρουσιάστηκαν διάφοροι τρόποι ελέγχου καλής προσαρμογής για την Poisson κατανομή. Η ύπαρξη τόσων πολλών ελέγχων είναι πλήρως αιτιολογημένη από το γεγονός ότι υπάρχουν ποικίλοι τρόποι απόκλισης από αυτήν την κατανομή. Το προφανές ερώτημα που μπορεί ήδη να έχει προκύψει είναι ποιος ή ποιοι από αυτούς τους ελέγχους είναι προτιμότερο να χρησιμοποιούνται ή ακόμη και ποιοι δεν δίνουν αξιόπιστα αποτελέσματα. Η απάντηση σε αυτό το ερώτημα προφανώς απαιτεί εκτενές συγκριτικές μελέτες για την αξιολόγηση των ελέγχων ως προς τη διατήρηση της πιθανότητας σφάλματος τύπου Ι, αλλά και ως προς την ισχύ. Δηλαδή, απαιτείται να αξιολογηθεί ότι η πραγματική πιθανότητα απόρριψης της μηδενικής υπόθεσης, ενώ αυτή είναι αληθής, είναι κοντά στο ονομαστικό επίπεδο σημαντικότητας α, ενώ από την άλλη μεριά η πιθανότητα απόρριψής της όταν τα δεδομένα προέρχονται από άλλες, εναλλακτικές κατανομές είναι υψηλή. Στο παραπάνω πλαίσιο, το κεφάλαιο αυτό διαρθρώνεται σε δύο μέρη. Στο πρώτο μέρος, παρατίθενται συνοπτικά συμπεράσματα εκτενών συγκριτικών μελετών που έχουν εμφανιστεί στη βιβλιογραφία, ενώ στο δεύτερο μέρος του κεφαλαίου παρουσιάζονται τα αποτελέσματα της συγκριτικής μελέτης που διεξήχθη στα πλαίσια αυτής της μεταπτυχιακής διατριβής. Τα αποτελέσματα της συγκριτικής μελέτης προέκυψαν μέσω του κώδικα της R, που είναι διαθέσιμος στον παρακάτω ιστότοπο: `https://github.com/fotinizogka/Myproject/tree/master`.

3.1 Ανασκόπηση

Όσο είμαστε σε θέση να γνωρίζουμε, οι τρεις πιο εκτενείς συγκριτικές μελέτες των ελέγχων καλής προσαρμογής της Poisson κατανομής έχουν διενεργηθεί από τους Karlis and Xekalaki (2000), Gürtler and Henze (2000) και πιο πρόσφατα από τους Mijburgh and Visagie (2020). Ωστόσο, συγκριτικές μελέτες περιορισμένης κλίμακας έχουν διενεργηθεί και από άλλους συγγραφείς-ερευνητές, στο πλαίσιο της εισαγωγής από τους ίδιους, ενός νέου στατιστικού ελέγχου καλής προσαρμογής. Στην ενότητα αυτή παρατίθενται τα συμπεράσματα των εκτενών συγκριτικών μελετών, καθώς τα όποια επιμέρους συμπεράσματα συμπεριλαμβάνονται σ΄ αυτά.

Μελέτη των Karlis and Xekalaki (2000)

Οι Karlis and Xekalaki (2000) συνέκριναν ως προς την ισχύ και την απόδοση τους ελέγχους που παρατίθενται στον Table 1 της εργασίας τους και τμήμα του οποίου αναπαράγεται στον Πίνακα 3.1 αυτής της μεταπτυχιακής διατριβής.

Τεστ	Αναφορά	Ενότητα	Σχέση
X^2_{LR}	Titterington et al. (1985)	2.1.1	(2.6)
$D_n(\widehat{\lambda})$	Kolmogorov (1933)	2.2.1	(2.11)
C_n	Henze (1996)	2.2.2	(2.14)
V^*	Nakamura and Pérez-Abreu (1993)	2.4	(2.32)
D_n	Fisher and Thornton (1922)	2.5.1	(2.33)
O_2	Böhning (1994)	2.5.1	(2.42)
T	Gupta et al. (1994)	2.5.2	(2.46)
$Z_{p=4}$	Gart and Pettigrew (1970)	2.5.4	(2.47)

Πίνακας 3.1: Τμήμα των στατιστικών συναρτήσεων της συγκριτικής μελέτης των Karlis and Xekalaki (2000).

Στο πλαίσιο αυτό, η συγκριτική μελέτη των ελέγχων σε επίπεδο σημαντικότητας $\alpha = 0.05$ αποτελείται από δύο βήματα. Στο πρώτο βήμα προσδιορίζονται μέσω προσομοίωσης bootstrap οι κρίσιμες τιμές για τους στατιστικούς ελέγχους και για όλες τις υπό τη μηδενική υπόθεση κατανομές Poisson με $\lambda = 1, 3, 5$ και δειγματικά μεγέθη $n = 50, 100, 500$. Στο επόμενο βήμα, δοθέντος του δειγματικού μεγέθους, 10.000 δείγματα δημιουργούνται από διάφορες εναλλακτικές κατανομές. Σημειώνεται ότι οι εναλλακτικές κατανομές επιλέχθηκαν κατά τέτοιον τρόπο ώστε να ταιριάζουν ως προς τις πρώτες ροπές, αλλά και να παριστάνουν κάποια εναλλακτικά μοτίβα, κάποιους δυνατούς τρόπους απόκλισης από την Poisson κατανομή. Επιπλέον, θεωρήθηκαν και εναλλακτικές κατανομές με μέση τιμή ίση με τη διακύμανση. Συνοψίζοντας, οι Karlis and Xekalaki (2000) θεώρησαν τόσο υπερδιασκορπισμένες και υποδιασκορπισμένες κατανομές όσο και κατανομές που είναι ισοδιασκορπισμένες.

Συμπεράσματα των Karlis and Xekalaki (2000) για υπερδιασκορπισμένες και υποδιασκορπισμένες εναλλακτικές κατανομές:

- Ο έλεγχος των Gupta et al. (1994) που παρουσιάστηκε στην Ενότητα 2.5.2 έχει κατώτερη απόδοση σε όλες τις περιπτώσεις και η ισχύς του είναι μικρή. Παρόμοια είναι η απόδοση του ελέγχου με την 4^{η} ημιαναλλοίωτη των Gart and Pettigrew (1970) (βλέπε Ενότητα 2.5.4).

- Η απόδοση όλων των εναλλακτικών μορφών του δείκτη διασποράς D_n είναι παρόμοια.

- Στις περιπτώσεις που ο πληθυσμιακός δείκτης διασποράς μειώνεται, τα στατιστικά τεστ που βασίζονται στην εμπειρική συνάρτηση κατανομής έχουν καλύτερη απόδοση.

- Προτείνουν στην περίπτωση ελέγχου της Poisson κατανομής έναντι υπερδιασκορπισμένων ή υποδιασκορπισμένων κατανομών να προτιμάται ένας έλεγχος που βασίζεται στο πηλίκο διακύμανσης προς μέση τιμή. Η στατιστική συνάρτηση ελέγχου της σχέσης (2.6) και το Hellinger deviance test έχουν καλή απόδοση για την περίπτωση υπερδιασκορπισμένων εναλλακτικών κατανομών, αλλά απαιτούν περισσότερο υπολογιστικό χρόνο. Επιπρόσθετα, προτείνεται η χρήση της σ.σ. του Böhning (1994) που παρουσιάστηκε στην Ενότητα 2.5, καθώς η κατανομή της, με βάση προσομοιώσεις, είναι κοντά στην τυπική κανονική.

Συμπεράσματα των Karlis and Xekalaki (2000) για εναλλακτικές κατανομές που είναι ισοδιασκορπισμένες:

- Σε αυτήν την περίπτωση οι έλεγχοι που βασίζονται στο πηλίκο διακύμανσης προς μέση τιμή, όπως αναμενόταν, αποτυγχάνουν να ανιχνεύσουν αποκλίσεις από την Poisson.
- Για τα μικρότερα σε μέγεθος δείγματα, ο έλεγχος που βασίζεται στην 4^η ημιαναλλοίωτη και αυτός των Gupta et al. (1994) είναι προτιμότερο. Αυτό έρχεται σε αντίθεση με το συμπέρασμα που έχει διατυπωθεί για την απόδοση αυτών των ελέγχων στην περίπτωση υπερδιασκορπισμένων και υποδιασκορπισμένων κατανομών.
- Ο έλεγχος που προτάθηκε από τους Nakamura and Pérez-Abreu (1993) (βλέπε Ενότητα 2.4, σχέση (2.32)) φαίνεται να υπερέχει των ελέγχων που στηρίζονται στην ε.π.σ., ενώ ο έλεγχος που βασίζεται στη σ.σ.ε. C_n των

Cramér-von Mises υπερέχει αυτού που βασίζεται στη σ.σ.ε. D_n των Kolmogorov Smirnov .

Συνολικό συμπέρασμα: Από τα τεστ που βασίζονται στην εμπειρική πιθανογεννήτρια συνάρτηση, προτιμότερο είναι αυτό των Nakamura and Pérez-Abreu (1993). Από όλους τους ελέγχους που στηρίζονται στον λόγο διακύμανσης προς αναμενόμενη τιμή, προτιμότερος είναι αυτός των Böhning (1994) που βασίζεται στη σ.σ.ε O_2.

Μελέτη των Gürtler and Henze (2000)

Οι Gürtler and Henze (2000) συνέκριναν ως προς την απόδοση τους ελέγχους που παρατίθενται στον Πίνακα 3.2.

Τεστ	Αναφορά	Ενότητα	Σχέση
$D_n(\widehat{\lambda})$	Kolmogorov (1933)	2.2.1	(2.11)
C_n	Henze (1996)	2.2.2	(2.14)
C_n^*	Henze (1996)	2.2.2	(2.15)
$\tilde{T}_n$	Klar (1999)	2.2.2	(2.19)
I_n	Klar (1999)	2.3	(2.23)
R_a	Baringhaus et al. (2000)	2.4	(2.26)
T_a	Treutler (1995)	2.4	(2.28)
V	Nakamura and Pérez-Abreu (1993)	2.4	(2.31)
V^*	Nakamura and Pérez-Abreu (1993)	2.4	(2.32)
U^2	Rayner and McIntyre (1985)	2.5.1	(2.38)

Πίνακας 3.2: Στατιστικές συναρτήσεις της συγκριτικής μελέτης των Gürtler and Henze (2000).

Στη μελέτη προσομοίωσής τους θεώρησαν 10.000 δείγματα μεγέθους $n = 50$ με επίπεδο σημαντικότητας $\alpha = 0.1$.

Συμπεράσματα των Gürtler and Henze (2000):

- Από τα τεστ που βασίζονται στην εμπειρική συνάρτηση, την καλύτερη ισχύ, για τις περισσότερες εναλλακτικές κατανομές που εξετάστηκαν, την έχει η σ.σ.ε του Klar ($\tilde{T}_n$), έναντι των C_n, C_n^* και $D_n(\widehat{\lambda})$.

- Η ισχύς των R_a και T_a για σταθεροποιημένο a, είναι σχεδόν η ίδια. Συγκεκριμένα, όταν οι εναλλακτικές κατανομές είναι είτε διακριτές ομοιόμορφες κατανομές ή μείξη Poisson κατανομών, με την αύξηση της παραμέτρου a φαίνεται η ισχύς να παραμένει σχεδόν η ίδια, εν αντιθέσει με άλλες εναλλακτικές κατανομές όπου αυξάνοντας την τιμή της παραμέτρου a, αυξάνεται και η ισχύ.

- Ο έλεγχος U^2, θεωρείται ισχυρός, εκτός από περιπτώσεις εναλλακτικών κατανομών για τις οποίες ισχύει ότι η αναμενόμενη τιμή και η διακύμανση είναι σχεδόν ίσες. Ο συγκεκριμένος έλεγχος χρήζει μεγάλης προσοχής, μιας και το πραγματικό επίπεδο σημαντικότητάς του μπορεί να πέσει πολύ πιο χαμηλά από το επίπεδο σημαντικότητας που έχει οριστεί για τη διεξαγωγή του ελέγχου.

- Οι σ.σ.ε V και V^* παρουσιάζουν είτε χαμηλή είτε υψηλή ισχύ, ανάλογα με το αν οι εναλλακτικές κατανομές είναι υποδιασκορπισμένες (για παράδειγμα διωνυμική) ή σχεδόν ισοδιασκορπισμένες, όπως στην περίπτωση της διακριτής ομοιόμορφης κατανομής $DU(0, 4)$.

Συνολικά, τα τεστ που βασίζονται στις σ.σ. T_5, R_5, $\tilde{T}_n$, I_n και V^* έχουν καλύτερη απόδοση.

Μελέτη των Mijburgh and Visagie (2020)

Πρόσφατα, οι Mijburgh and Visagie (2020), ήταν αυτοί που πραγματοποίησαν μία συγκριτική ανασκόπηση και αξιολόγηση των ελέγχων καλής προσαρμογής της Poisson κατανομής. Ειδικότερα, συνέκριναν ως προς την απόδοσή τους, τους ελέγχους του Πίνακα 3.3 αυτής της μεταπτυχιακής διατριβής.

Τεστ	Αναφορά	Ενότητα	Σχέση
$D_n(\widehat{\lambda})$	Kolmogorov (1933)	2.2.1	(2.11)
C_n	Henze (1996)	2.2.2	(2.14)
$\tilde{T}_n$	Klar (1999)	2.2.2	(2.19)
I_n	Klar (1999)	2.3	(2.23)
R_a	Baringhaus et al. (2000)	2.4	(2.26)
T_a	Treutler (1995)	2.4	(2.28)
MN_a	Meintanis and Nikitin (2008)	2.4	(2.30)
V^*	Nakamura and Pérez-Abreu (1993)	2.4	(2.32)
U^2	Rayner and McIntyre (1985)	2.5.1	(2.38)
SR	Székely and Rizzo (2004)	2.2.2	(2.22)

Πίνακας 3.3: Στατιστικές συναρτήσεις της συγκριτικής μελέτης των Mijburgh and Visagie (2020).

Στο σημείο αυτό επισημαίνεται ότι μία συλλογή των εναλλακτικών κατανομών που θεώρησαν ήταν οι λεγόμενες σταθμισμένες Poisson κατανομές $WP(\lambda, a, b)$ με σ.π. που δίνεται από τη σχέση

$$f(x) = \frac{w(x)f_\lambda(x)}{E(w(\tilde{X}))}, \; x = 0, 1, ..., \tag{3.1}$$

όπου

$$E(w(\tilde{X})) = \sum_{j=0}^{\infty} w(j)f_\lambda(x) < \infty,$$

με τη συνάρτηση βάρους να είναι το ακόλουθο πολυώνυμο δευτέρου βαθμού

$$w(x) = ax^2 + bx + 1.$$

Άλλες κατανομές που θεώρησαν ήταν οι: διωνυμική $Bin(m, p)$, αρνητική διωνυμική $NB(r, p)$, μείξη δύο Poisson $PM(\lambda_1, \lambda_2)$, γενικευμένη Poisson $GP(\lambda_1, \lambda_2)$, Zero Inflated Poisson $ZIP(p, \lambda)$ και διακριτή ομοιόμορφη $DU(a, b)$. Επομένως, θεωρήθηκαν εναλλακτικές κατανομές, οι οποίες είναι τόσο υπερδιασκορπισμένες όσο και υποδιασκορπισμένες ή ισοδιασκορπισμένες. Παραδείγματα της πρώτης κατηγορίας κατανομών είναι η $PM(0.01, 1, 5)$ και η $NB(1, 0.5)$, της δεύτερης κατηγορίας η $DU(0, 2)$ και η $Bin(50, 0.01)$, ενώ στην κατηγορία των ισοδιασκορπισμένων κατανομών ανήκουν για παράδειγμα η $DU(0, 4)$ και η $WP(1, 1, -1)$.

Στο πλαίσιο αυτό, οι Mijburgh and Visagie (2020) στη μελέτη προσομοίωσής τους θεώρησαν 50.000 τυχαία δείγματα μεγέθους $n = 30, 50, 100, 200$ από τις προαναφερθείσες εναλλακτικές κατανομές.

Συμπεράσματα που εξήγαγαν οι Mijburgh and Visagie (2020), είναι τα ακόλουθα:

- Τα περισσότερα στατιστικά τεστ ελέγχου της κατανομής Poisson που μελετήθηκαν έναντι εναλλακτικών κατανομών ισοδιασποράς (όπως Poisson, διακριτή ομοιόμορφη και σταθμισμένη Poisson) δεν παρέχουν υψηλή ισχύ. Τα μόνα τεστ που παρέχουν την υψηλότερη ισχύ έναντι των κατανομών αυτών είναι αυτά που βασίζονται στις σ.σ.ε. $\tilde{T}_n$, V^* και SR και έχουν προταθεί από τον Klar (1999), τους Nakamura and Pérez-Abreu (1993) και τους Székely and Rizzo (2004), αντίστοιχα.

- Το πιο ισχυρό στατιστικό τεστ έναντι των υποδιασκορπισμένων εναλλακτικών κατανομών (όπως είναι η διωνυμική κατανομή) είναι αυτό που βασίζεται στη σ.σ.ε. MN_a και το οποίο προτάθηκε από τους Meintanis and Nikitin (2008) με τα τεστ που βασίζονται στη σ.σ.ε. T_a και SR των Treutler (1995) και Székely and Rizzo (2004) , αντίστοιχα, να ακολουθούν.

- Στην περίπτωση των υπερδιασκορπισμένων εναλλακτικών κατανομών (όπως είναι η αρνητική διωνυμική και η μείξη Poisson), η υψηλότερη ισχύς παρέχεται από την σ.σ.ε U^2 των Rayner and McIntyre (1985), με τα τεστ που στηρίζονται στις σ.σ. $D_n(\widehat{\lambda})$, C_n και V^* των Kolmogorov Smirnov, Cramér-von Mises και των Nakamura and Pérez-Abreu (1993), αντίστοιχα, να ακολουθούν.

- Το τεστ που στηρίζεται στη σ.σ. MN_a και το οποίο προτάθηκε από τους Meintanis and Nikitin (2008), παρότι δεν υπερτερεί, έναντι όλων των άλλων τεστ, έναντι οποιασδήποτε υπερδιασκορπισμένης εναλλακτικής κατανομής, παρέχει σχετικά υψηλές τιμές για όλες τις εναλλακτικές κατανομές. Συνεπώς, ο έλεγχος αυτής προτείνεται σε κάθε περίπτωση εναλλακτικής κατανομής.

3.2 Συγκριτική μελέτη-συμπεράσματα

Στην ενότητα αυτή θα παρουσιασθούν και θα σχολιασθούν τα αποτελέσματα της συγκριτικής μελέτης που διεξήχθη στα πλαίσια αυτής της μεταπτυχιακής διατριβής. Ειδικότερα, συγκρίθηκαν ως προς την απόδοση οι έλεγχοι καλής προσαρμογής που δίνονται στον Πίνακα 3.4, ήτοι θεωρήθηκαν δεκαπέντε (15) διαφορετικοί έλεγχοι καλής προσαρμογής. Ωστόσο, πριν από την παράθεση αυτών των αποτελεσμάτων θα περιγραφεί ο σχεδιασμός της μελέτης προσομοίωσης. Τέλος, επισημαίνεται ότι για τους ελέγχους που εξαρτώνται από μία πρόσθετη παράμετρο a, όπως αυτοί που στηρίζονται στις σ.σ. T_a, R_a και MN_a επιλέχθηκε, ακολουθώντας τα συμπεράσματα προηγούμενων συγκριτικών μελετών, η τιμή $a = 5$.

Τεστ	Αναφορά	Ενότητα	Σχέση
U^2	Rayner and McIntyre (1985)	2.5.1	(2.38)
$D_n(\widehat{\lambda})$	Kolmogorov (1933)	2.2.1	(2.11)
C_n	Henze (1996)	2.2.2	(2.14)
C_n^*	Henze (1996)	2.2.2	(2.15)
$\tilde{T}_n$	Klar (1999)	2.2.2	(2.19)
W^2	Spinelli and Stephens (1997)	2.2.2	(2.16)
A^2	Spinelli and Stephens (1997)	2.2.2	(2.17)
SR	Székely and Rizzo (2004)	2.2.2	(2.22)
I_n	Klar (1999)	2.3	(2.23)
R	Rueda et al. (1991)	2.4	(2.25)
R_a	Baringhaus et al. (2000)	2.4	(2.26)
T	Baringhaus and Henze (1992)	2.4	(2.27)
T_a	Treutler (1995)	2.4	(2.28)
MN_a	Meintanis and Nikitin (2008)	2.4	(2.30)
V^*	Nakamura and Pérez-Abreu (1993)	2.4	(2.32)

Πίνακας 3.4: Στατιστικές συναρτήσεις της συγκριτικής μελέτης της Ενότητας 3.2.

3.2.1 Σχεδιασμός μελέτης προσομοίωσης

Αρχικά επισημαίνεται ότι η διεξαγωγή των ελέγχων προσαρμογής θα γίνει σε όλες τις περιπτώσεις με τη βοήθεια παραμετρικού bootstrap, για να εξασφαλιστεί, ας μας επιτραπεί η έκφραση, η υποκειμενικότητα, στα αποτελέσματα. Η επιλογή της χρήσης παραμετρικού bootstrap δικαιολογείται, καθώς, αν όχι όλοι, οι περισσότεροι έλεγχοι καλής προσαρμογής έχουν είτε άγνωστες κρίσιμες τιμές ή πίνακες κρίσιμων τιμών για αυτούς έχουν δοθεί για περιορισμένο αριθμό επιλογών μεγέθους δείγματος και επιπέδου σημαντικότητας. Επιπλέον, η ασυμπτωτική κατανομή πολλών εξ αυτών είτε εξαρτάται από την άγνωστη παράμετρο της κατανομής Poisson ή είναι πολύπλοκη ή η προσέγγιση που υποδεικνύεται δεν είναι ικανοποιητική.

Στο παραπάνω πλαίσιο, ακολουθώντας τον σχεδιασμό των συγκριτικών μελετών της προηγούμενης ενότητας, για την αξιολόγηση των ελέγχων ως προς τη διατήρηση του επιπέδου σημαντικότητας παράγονται 10.000 τυχαία δείγματα μέγεθους $n = 50, 100, 200$ από την κατανομή Poisson , με παράμετρο $\lambda = 0.5, 1, 5, 10, 30$. Για κάθε συνδυασμό μεγέθους δείγματος και τιμής της παραμέτρου ο έλεγχος διεξάγεται σε επίπεδο σημαντικότητας $\alpha = 0.05$, με χρήση $B = 500$ bootstrap δειγμάτων. Σε κάθε περίπτωση το εμπειρικό επίπεδο σημαντικότητας καταγράφεται, δηλαδή υπολογίζεται το ποσοστό των φορών που απορρίπτεται η μηδενική υπόθεση, ενώ αυτή είναι αληθής.

Στη συνέχεια για την αξιολόγηση της ισχύς των ελέγχων καλής προσαρμογής ακολουθείται εκ νέου η παραπάνω διαδικασία με την τροποποίηση αυτή τη φορά ότι τα δεδομένα δεν προέχονται από την κατανομή Poisson , αλλά από εναλλακτικές κατανομές. Ακολουθώντας τις συγκριτικές μελέτες της προηγούμενης ενότητας, θεωρούνται τόσο υπερδιασκορπισμένες όσο και υποδιασκορπισμένες κατανομές. Ειδικότερα, θεωρήθηκαν οι ακόλουθες υποδιασκορπισμένες κατανομές:

- διωνυμικές $Bin(2, 0.5), Bin(4, 0.25), Bin(10, 0.1), Bin(20, 0.25)$, και η
- διακριτή ομοιόμορφη $DU(0, 2)$,

ενώ θεωρήθηκαν οι ακόλουθες υπερδιασκορπισμένες κατανομές:

- αρνητικές διωνυμικές $NB(1, 0.5)$, $NB(3, 0.75)$, $NB(15, 0.75)$, $NB(45, 0.9)$,
- μείξη Poisson $PM(0.25, 1, 5)$, $PM(0.05, 1, 5)$, $PM(0.5, 2, 5)$, $PM(0.01, 1, 5)$,
- διακριτές ομοιόμορφες $DU(0, 5)$, $DU(0, 6)$, και
- γενικευμένες Poisson $GPD(4, 0.1)$, $GPD(3.75, 0.25)$, $GPD(4.5, 0.1)$.

Τέλος, θεωρήθηκε και μία ισοδιασκορπισμένη κατανομή, η διακριτή ομοιόμορφη $DU(0, 4)$.

3.2.2 Αποτελέσματα συγκριτικής μελέτης

Τα αποτελέσματα που αφορούν το εμπειρικό επίπεδο σημαντικότητας παρατίθενται στους Πίνακες 3.5-3.7. Από την άλλη μεριά, τα αποτελέσματα που αφορούν την ισχύ των στατιστικών ελέγχων καλής προσαρμογής για υποδιασκορπισμένες εναλλακτικές κατανομές δίνονται στους Πίνακες 3.8-3.10, ενώ τα αντίστοιχα αποτελέσματα που αφορούν τις υπερδιασκορπισμένες εναλλακτικές κατανομές δίνονται στους Πίνακες 3.11-3.16. Τέλος, τα αποτελέσματα που αφορούν την ισοδιασκορπισμένη εναλλακτική κατανομή που θεωρήθηκε παρατίθενται στον Πίνακα 3.17.

	$\mathcal{P}(0.5)$	$\mathcal{P}(1)$	$\mathcal{P}(5)$	$\mathcal{P}(10)$	$\mathcal{P}(30)$
U^2	0.0531	0.051	0.0518	0.0513	0.0502
$D_n(\widehat{\lambda})$	0.0526	0.0484	0.0536	0.0469	0.0496
C_n	0.0552	0.0497	0.0496	0.0491	0.0503
C_n^*	0.0558	0.0505	0.0504	0.0470	0.0498
$\tilde{T}_n$	0.0502	0.0483	0.0522	0.0473	0.0474
W^2	0.0552	0.0497	0.0496	0.0503	0.0475
A^2	0.0558	0.0513	0.0506	0.0475	0.0493
SR	0.0537	0.0483	0.0541	0.0506	0.0519
I_n	0.0497	0.0487	0.0518	0.0471	0.0472
R	0.0522	0.0496	0.0502	0.0491	0.0495
R_a	0.0465	0.0495	0.0544	0.0549	0.0503
T	0.0508	0.0483	0.0528	0.0517	0.0524
T_a	0.0494	0.049	0.0546	0.0503	0.053
MN_a	0.0482	0.0494	0.0531	0.0533	0.0551
V^*	0.0528	0.0534	0.0499	0.0517	0.0518

Πίνακας 3.5: Εμπειρικό επίπεδο σημαντικότητας για α=0.05 και $n = 50$.

	$\mathcal{P}(0.5)$	$\mathcal{P}(1)$	$\mathcal{P}(5)$	$\mathcal{P}(10)$	$\mathcal{P}(30)$
U^2	0.0476	0.0493	0.0505	0.0527	0.0506
$D_n(\widehat{\lambda})$	0.0499	0.0478	0.05	0.0551	0.0529
C_n	0.0498	0.0483	0.0514	0.0553	0.0513
C_n^*	0.0489	0.0489	0.0524	0.054	0.0508
$\tilde{T}_n$	0.0497	0.046	0.0518	0.0517	0.05
W^2	0.0498	0.0483	0.0514	0.0533	0.0513
A^2	0.05	0.0473	0.052	0.0527	0.0516
SR	0.0481	0.0485	0.0502	0.0543	0.0515
I_n	0.0487	0.0469	0.053	0.0522	0.052
R	0.0487	0.0502	0.0484	0.0507	0.0514
R_a	0.0472	0.0491	0.0523	0.0518	0.052
T	0.0466	0.0487	0.0488	0.0517	0.0523
T_a	0.0471	0.0491	0.0529	0.0527	0.052
MN_a	0.0468	0.0477	0.0521	0.0526	0.0522
V^*	0.0496	0.0484	0.0516	0.0523	0.0516

Πίνακας 3.6: Εμπειρικό επίπεδο σημαντικότητας για α=0.05 και $n = 100$.

	$\mathcal{P}(0.5)$	$\mathcal{P}(1)$	$\mathcal{P}(5)$	$\mathcal{P}(10)$	$\mathcal{P}(30)$
U^2	0.0525	0.0557	0.0517	0.0505	0.0542
$D_n(\widehat{\lambda})$	0.0536	0.0498	0.0555	0.05	0.0533
C_n	0.0494	0.0556	0.0507	0.0498	0.0495
C_n^*	0.0507	0.0553	0.0525	0.0522	0.0517
$\tilde{T}_n$	0.0507	0.0544	0.0487	0.0517	0.0529
W^2	0.0494	0.0556	0.0507	0.0498	0.0495
A^2	0.049	0.057	0.05	0.0507	0.0524
SR	0.0508	0.0563	0.0517	0.0537	0.0515
I_n	0.0503	0.0552	0.0486	0.049	0.0528
R	0.0517	0.0568	0.0507	0.0513	0.0525
R_a	0.0495	0.0567	0.0522	0.0516	0.0527
T	0.0496	0.0577	0.0513	0.0502	0.0535
T_a	0.049	0.0567	0.0519	0.055	0.053
MN_a	0.0497	0.0572	0.0496	0.0476	0.0464
V^*	0.0514	0.0539	0.0545	0.0519	0.0499

Πίνακας 3.7: Εμπειρικό επίπεδο σημαντικότητας για α=0.05 και $n = 200$.

	Bin(2,0.5)	Bin(4,0.25)	Bin(10,0.1)	Bin(20,0.25)	DU(0,2)
U^2	0.9213	0.1833	0.0533	0.2212	0.3733
$D_n(\widehat{\lambda})$	0.6314	0.2194	0.0723	0.2048	0.2937
C_n	0.8261	0.2305	0.0719	0.1932	0.3119
C_n^*	0.8099	0.2413	0.0598	0.1799	0.3124
$\tilde{T}_n$	0.7456	0.1813	0.0749	0.1708	0.5238
W^2	0.8261	0.2305	0.0719	0.1932	0.3119
A^2	0.8954	0.2493	0.0672	0.167	0.4125
SR	0.8907	0.1839	0.073	0.1673	0.5257
I_n	0.2384	0.0501	0.0354	0.0779	0.5298
R	0.8676	0.235	0.0752	0.1654	0.5088
R_a	0.9178	0.1835	0.0648	0.2175	0.593
T	0.8097	0.2186	0.0753	0.1727	0.5456
T_a	0.9153	0.2208	0.0634	0.2491	0.6741
MN_a	**0.9733**	**0.2582**	**0.0759**	**0.2702**	0.338
V^*	0.890	0.0526	0.0201	0.0819	**0.7486**
Δείκτης FI	0.5	0.75	0.9	0.75	0.67

Πίνακας 3.8: Ισχύς των ελέγχων καλής προσαρμογής για υποδιασκορπισμένες εναλλακτικές κατανομές για α=0.05 και $n = 50$.

	Bin(2,0.5)	Bin(4,0.25)	Bin(10,0.1)	Bin(20,0.25)	DU(0,2)
U^2	0.9992	0.4546	0.0813	0.4547	0.8229
$D_n(\widehat{\lambda})$	0.9924	0.4607	0.0993	0.4277	0.601
C_n	0.9995	0.4594	**0.0995**	0.3964	0.6217
C_n^*	0.9958	0.3114	0.0741	0.3767	0.6341
$\tilde{T}_n$	0.995	0.3495	0.0838	0.3158	0.9327
W^2	0.995	0.4594	**0.0995**	0.3964	0.6217
A^2	**0.9996**	0.3816	0.0804	0.3442	0.632
SR	**0.9996**	0.3603	0.0891	0.2977	0.917
I_n	0.5224	0.1236	0.0314	0.1498	0.968
R	0.9932	0.3365	0.088	0.2382	0.9694
R_a	0.9982	0.3617	0.0861	0.3728	0.9699
T	0.9848	0.392	0.0943	0.2337	0.945
T_a	0.998	0.428	0.096	0.4471	0.9831
MN_a	0.997	**0.501**	0.0994	**0.4695**	0.6915
V^*	0.9994	0.1794	0.0321	0.3178	**0.9999**
Δείκτης FI	0.5	0.75	0.9	0.75	0.67

Πίνακας 3.9: Ισχύς των ελέγχων καλής προσαρμογής για υποδιασκορπισμένες εναλλακτικές κατανομές για α=0.05 και $n = 100$.

0.8ςμ

	Bin(2,0.5)	Bin(4,0.25)	Bin(10,0.1)	Bin(20,0.25)	DU(0,2)
U^2	**0.988**	**0.817**	0.1551	**0.7955**	**1**
$D_n(\widehat{\lambda})$	0.95	0.7998	0.1673	0.7378	0.9438
C_n	0.949	0.7938	0.1664	0.731	0.9524
C_n^*	0.956	0.7937	0.1228	0.7346	0.9527
$\tilde{T}_n$	0.982	0.6477	0.1244	0.5518	**1**
W^2	0.949	0.7938	0.1664	0.731	0.9524
A^2	0.968	0.6975	0.1447	0.6554	0.9718
SR	0.983	0.654	0.1305	0.5525	**1**
I_n	0.978	0.2664	0.0311	0.2741	0.9999
R	0.976	0.6824	0.127	0.5193	0.9998
R_a	0.982	0.6941	0.1317	0.682	**1**
T	0.975	0.6768	0.155	0.5294	**1**
T_a	0.979	0.7595	0.1567	0.7673	**1**
MN_a	0.980	0.808	**0.1767**	0.7685	0.9614
V^*	0.9644	0.4745	0.0601	0.5923	**1**
Δείκτης FI	0.5	0.75	0.9	0.75	0.67

Πίνακας 3.10: Ισχύς των ελέγχων καλής προσαρμογής για υποδιασκορπισμένες εναλλακτικές κατανομές για α=0.05 και $n = 200$.

	NB(1,0.5)	NB(3,0.75)	NB(15,0.75)	NB(45,0.9)	PM(0.25,1,5)	PM(0.05,1,5)	PM(0.5,2,5)	PM(0.01,1,5)
U^2	**0.8618**	**0.357**	**0.3653**	0.112	0.8668	0.1551	**0.7732**	0.063
$D_n(\widehat{\lambda})$	0.8099	0.2961	0.3192	0.1092	0.7092	**0.2530**	0.6204	0.071
C_n	0.8138	0.3091	0.3196	0.0989	0.7254	0.2528	0.5711	0.0712
C_n^*	0.8251	0.3241	0.3338	0.0708	0.7356	0.2529	0.5836	**0.0887**
$\tilde{T}_n$	0.7426	0.1996	0.1781	0.1043	0.8912	0.0816	0.5616	0.0599
W^2	0.8138	0.3091	0.3196	0.0989	0.7254	0.2528	0.5711	0.0712
A^2	0.8223	0.3155	0.3206	0.0917	0.8786	0.2516	0.6992	0.0879
SR	0.7314	0.1955	0.1774	0.0624	0.5355	0.0813	0.5512	0.0539
I_n	0.8145	0.2943	0.2189	0.1081	0.8972	0.0829	0.653	0.0584
R	0.763	0.248	0.2861	0.097	0.95	0.1453	0.6604	0.0608
R_a	0.7736	0.2502	0.2881	0.1041	**0.9559**	0.1405	0.7268	0.0703
T	0.7804	0.2692	0.2906	0.1013	0.9456	0.1233	0.6521	0.0583
T_a	0.7862	0.2774	0.3012	0.1045	0.9496	0.1328	0.7689	0.0592
MN_a	0.836	0.3124	0.3088	0.0905	0.6255	0.2235	0.7509	0.0692
V^*	0.8228	0.3273	0.2827	**0.1362**	0.8599	0.1386	0.7092	0.0689
Δείκτης FI	2	1.33	1.33	1.11	1.75	1.16	1.64	1.03

Πίνακας 3.11: Ισχύς των ελέγχων καλής προσαρμογής για υπερδιασκορπισμένες εναλλακτικές κατανομές για α=0.05 και $n = 50$.

	NB(1,0.5)	NB(3,0.75)	NB(15,0.75)	NB(45,0.9)	PM(0.25,1,5)	PM(0.05,1,5)	PM(0.5,2,5)	PM(0.01,1,5)
U^2	**0.9934**	**0.5455**	**0.5667**	**0.1363**	0.9879	0.2191	**0.9587**	0.0673
$D_n(\widehat{\lambda})$	0.9869	0.4815	0.5456	0.1345	0.9794	0.357	0.8785	0.0818
C_n	0.9891	0.5099	0.5337	0.1383	0.9586	**0.3612**	0.8579	0.0822
C_n^*	0.9822	0.5862	0.5227	0.1384	0.9657	0.3544	0.8694	**0.0919**
$\tilde{T}_n$	0.9672	0.3516	0.3082	0.0864	0.9917	0.115	0.9433	0.0671
W^2	0.9891	0.5099	0.5337	0.1383	0.9586	**0.3612**	0.8579	0.0822
A^2	0.9649	0.5207	0.4462	0.1091	0.9911	0.3599	0.9339	0.0918
SR	0.9611	0.3478	0.3043	0.0827	0.9485	0.1185	0.9223	0.0507
I_n	0.9868	0.4639	0.3592	0.1193	0.9922	0.1102	0.9471	0.0618
R	0.9686	0.4053	0.4175	0.117	0.9983	0.19846	0.8978	0.055
R_a	0.9722	0.4135	0.47	0.1244	**0.9986**	0.2036	0.9583	0.0687
T	0.9628	0.4613	0.4308	0.1192	0.9874	0.152	0.9305	0.513
T_a	0.9721	0.4694	0.5035	0.1255	0.9981	0.1995	0.9583	0.0693
MN_a	0.9889	0.5097	0.5277	0.1251	0.9864	0.3176	0.9	0.0713
V^*	0.9701	0.4742	0.4531	0.0856	0.9739	0.1792	0.9298	0.0635
Δείκτης FI	2	1.33	1.33	1.11	1.75	1.16	1.64	1.03

Πίνακας 3.12: Ισχύς των ελέγχων καλής προσαρμογής για υπερδιασκορπισμένες εναλλακτικές κατανομές για α=0.05 και $n = 100$.

	NB(1,0.5)	NB(3,0.75)	NB(15,0.75)	NB(45,0.9)	PM(0.25,1,5)	PM(0.05,1,5)	PM(0.5,2,5)	PM(0.01,1,5)
U^2	0.9999	**0.7982**	**0.8348**	**0.2099**	**1**	0.3485	**1**	0.069
$D_n(\widehat{\lambda})$	0.9989	0.7715	0.7976	0.1975	**1**	0.5643	**1**	0.0927
C_n	**1**	0.7816	0.7906	0.1924	0.9999	**0.5682**	**1**	0.0929
C_n^*	**1**	0.7846	0.7916	0.1965	**1**	0.5596	**1**	**0.0951**
$\tilde{T}_n$	**1**	0.6199	0.549	0.1168	**1**	0.1862	0.9999	0.0613
W^2	**1**	0.7817	0.7906	0.1924	0.9999	**0.5682**	**1**	0.0929
A^2	**1**	0.769	0.7618	0.1827	**1**	0.5318	**1**	0.0926
SR	0.9999	0.6158	0.5352	0.1121	**1**	0.1827	0.9999	0.0514
I_n	**1**	0.7317	0.5727	0.1327	**1**	0.1986	0.9999	0.0526
R	**1**	0.6225	0.7343	0.1604	**1**	0.3346	**1**	0.054
R_a	**1**	0.6915	0.7402	0.1657	**1**	0.3464	**1**	0.0616
T	**1**	0.7211	0.7637	0.1841	**1**	0.3299	**1**	0.0662
T_a	**1**	0.7387	0.7748	0.1879	1	0.3364	**1**	0.0698
MN_a	**1**	0.7784	0.7984	0.1719	**1**	0.5342	**1**	0.0853
V^*	**1**	0.6916	0.6883	0.158	**1**	0.29015	**1**	0.0675
Δείκτης FI	2	1.33	1.33	1.11	1.75	1.16	1.64	1.03

Πίνακας 3.13: Ισχύς των ελέγχων καλής προσαρμογής για υπερδιασκορπισμένες εναλλακτικές κατανομές για α=0.05 και $n = 200$.

	DU(0,5)	DU(0,6)	GPD(4,0.1)	GPD(3.75,0.25)	GPD(4.5,0.1)
U^2	0.1061	0.3687	**0.2295**	**0.8246**	**0.2352**
$D_n(\widehat{\lambda})$	0.4311	0.7516	0.2102	0.5841	0.1491
C_n	0.4256	0.7304	0.2022	0.5334	0.1176
C_n^*	0.4223	0.7331	0.2186	0.5472	0.1267
$\tilde{T}_n$	0.6788	0.785	0.1156	0.7713	0.2086
W^2	0.4256	0.7304	0.2022	0.5334	0.1176
A^2	0.5438	0.7318	0.1707	0.7244	0.1828
SR	0.6599	0.6827	0.1133	0.2967	0.1073
I_n	0.4765	0.6431	0.1372	0.785	0.2204
R	0.7692	0.8808	0.1794	0.7891	0.1783
R_a	0.7409	0.8714	0.1838	0.7897	0.2081
T	0.5445	0.8414	0.1865	0.7076	0.1822
T_a	0.4229	0.6781	0.2038	0.7946	0.2111
MN_a	0.3382	0.669	0.1912	0.3115	0.2072
V^*	**0.8455**	**0.9082**	0.1949	0.7661	0.2217
Δείκτης FI	1.17	1.33	1.24	1.78	1.24

Πίνακας 3.14: Ισχύς των ελέγχων καλής προσαρμογής για υπερδιασκορπισμένες εναλλακτικές κατανομές για α=0.05 και $n = 50$.

	DU(0,5)	DU(0,6)	GPD(4,0.1)	GPD(3.75,0.25)	GPD(4.5,0.1)
U^2	0.1939	0.6367	**0.3576**	**0.9752**	**0.3644**
$D_n(\widehat{\lambda})$	0.7193	0.9518	0.3253	0.8515	0.2222
C_n	0.708	0.9436	0.3274	0.8276	0.1832
C_n^*	0.7527	0.9531	0.3285	0.8429	0.1939
$\tilde{T}_n$	0.956	0.9827	0.1756	0.954	0.3024
W^2	0.708	0.9436	0.3274	0.8276	0.1832
A^2	0.708	0.9942	0.3356	0.9384	0.275
SR	0.9588	0.9831	0.1875	0.8245	0.1286
I_n	0.8183	0.9303	0.2272	0.9573	0.3149
R	0.8875	0.9837	0.2457	0.914	0.2672
R_a	0.9866	0.9998	0.3045	0.9638	0.3293
T	0.901	0.9839	0.2515	0.9258	0.2763
T_a	0.7351	0.9557	0.3167	0.9655	0.3323
MN_a	0.6197	0.9278	0.3194	0.8724	0.1454
V^*	**0.9991**	**0.9999**	0.2698	0.9576	0.3152
Δείκτης FI	1.17	1.33	1.24	1.78	1.24

Πίνακας 3.15: Ισχύς των ελέγχων καλής προσαρμογής για υπερδιασκορπισμένες εναλλακτικές κατανομές για α=0.05 και $n = 100$.

	DU(0,5)	DU(0,6)	GPD(4,0.1)	GPD(3.75,0.25)	GPD(4.5,0.1)
U^2	0.3708	0.9158	**0.5928**	**0.9999**	**0.5923**
$D_n(\widehat{\lambda})$	0.9526	**1**	0.5335	0.988	0.3572
C_n	0.9496	0.9999	0.5319	0.9871	0.3244
C_n^*	0.9497	**1**	0.5423	0.9883	0.3348
$\tilde{T}_n$	0.9999	**1**	0.321	0.9989	0.4963
W^2	0.9496	0.9999	0.5319	0.9871	0.3244
A^2	**1**	**1**	0.5428	0.9985	0.4575
SR	0.9993	0.9999	0.3131	0.9822	0.3207
I_n	0.9982	0.9999	0.3612	0.9989	0.5052
R	0.9928	0.989	0.411	0.9962	0.4166
R_a	0.9999	0.9999	0.4746	**0.9999**	0.5332
T	0.9942	0.999	0.4143	0.9978	0.429
T_a	0.9813	0.9999	0.5316	**0.9999**	0.5372
MN_a	0.8968	0.9999	0.5383	0.9987	0.4157
V^*	**1**	**1**	0.4346	0.9988	0.5137
Δείκτης FI	1.17	1.33	1.24	1.78	1.24

Πίνακας 3.16: Ισχύς των ελέγχων καλής προσαρμογής για υπερδιασκορπισμένες εναλλακτικές κατανομές για α=0.05 και $n = 200$.

	$n = 50$	$n = 100$	$n = 200$
U^2	0.0119	0.0132	0.0138
$D_n(\widehat{\lambda})$	0.1031	0.1829	0.3876
C_n	0.0943	0.159	0.3408
C_n^*	0.106	0.1613	0.3341
$\tilde{T}_n$	0.5358	0.879	0.9964
W^2	0.0943	0.1829	0.3408
A^2	0.1231	0.1608	**1**
SR	0.5292	0.8782	0.9984
I_n	0.3191	0.6723	0.9888
R	0.6157	0.957	0.7962
R_a	0.6127	0.9497	0.9996
T	0.1719	0.5013	0.882
T_a	0.1766	0.4491	0.9671
MN_a	0.062	0.1163	0.1996
V^*	**0.7338**	**0.9989**	**1**

Πίνακας 3.17: Ισχύς των ελέγχων καλής προσαρμογής για την ισοδιασκορπισμένη εναλλακτική κατανομή $DU(0, 4)$ για α=0.05.

3.2.3 Συμπεράσματα συγκριτικής μελέτης

Από τα αποτελέσματατα που δίνονται στους Πίνακες 3.5-3.7 και αφορούν το εμπειρικό επίπεδο σημαντικότητας προκύπτει, όπως αναμενόταν αφού χρησιμοποιείται η μέθοδος bootstrap, ότι όλοι οι έλεγχοι διατηρούν το ονομαστικό επίπεδο σημαντικότητας.

Από τη σύγκριση ως προς την απόδοση των στατιστικών συναρτήσεων ελέγχου για υποδιασκορπισμένες εναλλακτικές κατανομές (βλέπε Πίνακες 3.8-3.10) προκύπτουν τα συμπεράσματα που ακολουθούν.

- Ο έλεγχος MN_a των Meintanis and Nikitin (2008) μπορεί να προταθεί ως εκείνος με ανταγωνιστική ισχύ για όλες τις υποδιασκορπισμένες ενναλακτικές κατανομές που θεωρήθηκαν.
- Ο έλεγχος C_n μπορεί να προταθεί μεταξύ των ελέγχων που παρουσιάστηκαν στην Ενότητα 2.2.2 , ήτοι που χρησιμοποιούν αποστάσεις μεταξύ της α.σ.κ. και της ε.α.σ.κ.
- Οι έλεγχοι που στηρίζονται στην ε.π.σ. υπερισχύουν γενικότερα των ελέγχων που στηρίζονται στην ε.α.σ.κ.
- Ο έλεγχος I_n που στηρίζεται στην ε.ο.σ.κ. δεν είναι ανταγωνιστικός με τους υπόλοιπους ελέγχους για μικρά σε μέγεθος δείγματα.
- Η απόδοση των ελέγχων είναι καλύτερη για τιμές του πληθυσμιακού δείκτη FI που είναι πιο απομακρυσμένες από την τιμή 1.

Από τη σύγκριση ως προς την απόδοση των στατιστικών συναρτήσεων ελέγχου για υπερδιασκορπισμένες εναλλακτικές κατανομές (βλέπε Πίνακες 3.11-3.16) προκύπτουν τα συμπεράσματα που ακολουθούν.

- Ο έλεγχος U^2 των Rayner and McIntyre (1985) παρουσιάζει την υψηλότερη ισχύ, στην πλειοψηφία των υπερδιασκορπισμένων εναλλακτικών κατανομών και για τα τρία μεγέθη δείγματος $n = 50, 100, 200$.

- Από τους ελέγχους που στηρίζονται στην ε.π.σ. καλή απόδοση έχει ο έλεγχος V^* στην περίπτωση των διακριτών ομοιόμορφων εναλλακτικών κατανομών, με τους ελέγχους που βασίζονται στις σ.σ. T_a και R_a, να ακολουθούν με, επίσης, υψηλή ισχύ.

- Η απόδοση των ελέγχων που στηρίζονται στην ε.α.σ.κ., κυρίως για μεγάλο μέγεθος δείγματος ($n = 200$), είναι επίσης ικανοποιητική.

- Η σ.σ.ε. MN_a, των Meintanis and Nikitin (2008), παρότι δεν υπερισχύει έναντι όλων των άλλων σ.σ.ε., παρατηρείται, ότι έχει σχετικά υψηλή ισχύ, για όλες τις υπερδιασκορπισμένες εναλλακτικές κατανομές που θεωρήθηκαν.

- Η σ.σ.ε SR, των Székely and Rizzo (2004), παρουσιάζει τη χαμηλότερη ισχύ, στην πλειοψηφία των υπερδιασκορπισμένων εναλλακτικών κατανομων, με εξαίρεση αυτών της διακριτής ομοιόμορφης.

Τέλος, από τη σύγκριση ως προς την απόδοση των στατιστικών συναρτήσεων ελέγχου για την ισοδιασκορπισμένη εναλλακτική κατανομή που θεωρήθηκε (βλέπε Πίνακα 3.17) προκύπτουν τα συμπεράσματα που ακολουθούν.

- Η πλειοψηφία των ελέγχων δεν παρουσιάζει καλή απόδοση, ειδικά για δείγματα μεγέθους $n = 50$ και $n = 100$.

- Από τους ελέγχους που στηρίζονται στην ε.α.σ.κ. καλύτερη απόδοση, γενικά, έχουν οι $\tilde{T}_n$ και SR.

- Από τους ελέγχους που στηρίζονται στην ε.π.σ. καλύτερη απόδοση έχει ο έλεγχος V^*. Ο έλεγχος αυτός είναι και ο πλέον ισχυρότερος ανάμεσα στους δεκαπέντε που θεωρήθηκαν για την υπό εξέταση ισοδιασκορπισμένη εναλλακτική κατανομή.

ΚΕΦΑΛΑΙΟ 4

Επεκτασεισ - Επιλογοσ

Διδιάστατα ή πολυδιάστατα απαριθμητά δεδομένα εμφανίζονται σε πολλά διαφορετικά ερευνητικά πεδία. Η εξέχουσα θέση της Poisson κατανομής είχε ως αποτέλεσμα την επέκτασή της σε δύο ή περισσότερες διαστάσεις και την εισαγωγή στη βιβλιογραφία των αντίστοιχων κατανομών που διαδραματίζουν σημαντικό ρόλο στη μοντελεποίηση τέτοιων δεδομένων. Η σπουδαιότητα αυτή αιτιολογεί πλήρως και την εισαγωγή στη βιβλιογραφία μεθόδων ελέγχου καλής προσαρμογής ότι για αυτές τις διδιάστατες και πολυδιάστατες γενικεύσεις. Για λόγους πληρότητας, στο κεφάλαιο αυτό αναφέρονται, συνοπτικά, έλεγχοι καλής προσαρμογής για τη διδιάστατη και m-διάστατη, με $m \geq 3$, κατανομή Poisson. Στον επίλογο του κεφαλαίου και της διατριβής αναδεικνυέται η αναγκαιότητα ύπαρξης εναλλακτικών στην Poisson κατανομών. Η αναγκαιότητα αυτή έχει ως επακόλουθη συνέπεια την αναγκαιότητα ύπαρξης ελέγχων καλής προσαρμογής για αυτές τις εναλλακτικές κατανομές και κάποιες βιβλιογραφικές αναφορές σε αυτό το πλαίσιο παρατίθενται για τον/την ενδιαφερόμενο/μενη αναγνώστη/στρια. Πριν προχωρήσουμε στην παρουσίαση όσων αναφέρθηκαν, στην επόμενη εισαγωγική ενότητα, θα παραθέσουμε κάποιες χρήσιμες έννοιες, αποτελέσματα και ορισμούς που θα διευκολύνουν την κατανόηση του κεφαλαίου.

4.1 Εισαγωγή

Παρότι στη βιβλιογραφία έχουν εμφανιστεί διάφοροι ορισμοί της διδιάστατης Poisson κατανομής (παραπέμπουμε στις εργασίες των Kocherlakota and Kocherlakota (1992), Novoa-Muñoz and Jiménez-Gamero (2014), καθώς και στις εκεί αναφορές), σε αυτήν τη διατριβή υιοθετείται ο ορισμός της διδιάστατης και πολυδιάστατης Poisson που συνηθέστερα χρησιμοποιείται στη βιβλιογραφία (Balakrishnan et al. (1998)) και οι οποίοι ορισμοί παρατίθενται στη συνέχεια.

Ορισμός 4.1.1. *Έστω $X_1 = Y_1 + Y_3$ και $X_2 = Y_2 + Y_3$, με Y_1, Y_2, Y_3 ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν Poisson κατανομή με παραμέτρους $\lambda_1' = \lambda_1 - \lambda_3 > 0$, $\lambda_2' = \lambda_2 - \lambda_3 > 0$ και $\lambda_3 > 0$, αντίστοιχα. Τότε λέμε ότι η από κοινού κατανομή των X_1 και X_2 είναι η διμεταβλητή Poisson με παράμετρο $\lambda = (\lambda_1, \lambda_2, \lambda_3)$, με συνάρτηση πιθανότητας που δίνεται από τη σχέση:*

$$P(X_1 = x_1, X_2 = x_2) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1^{x_1}}{x_1!} \frac{\lambda_2^{x_2}}{x_2!} \sum_{k=0}^{min(x_1,x_2)} \binom{x_1}{k} \binom{x_2}{k} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2} \right)^k .$$

Σε αυτήν την περίπτωση γράφουμε ότι $(X_1, X_2) \sim BP(\lambda)$, όπου $\lambda = (\lambda_1, \lambda_2, \lambda_3) \in \Theta$, με $\Theta = \{(\lambda_1, \lambda_2, \lambda_3) \in \mathbb{R}^3 : \lambda_1 > \lambda_3, \lambda_2 > \lambda_3, \lambda_3 > 0\}$.

Παρατήρηση 4.1.1. *Εύκολα προκύπτει ότι $E(X_1) = Var(X_1) = \lambda_1$, $E(X_2) = Var(X_2) = \lambda_2$, ενώ $Cov(X_1, X_2) = \lambda_3$. Επομένως, οι δειγματικές μέσες τιμές και η δειγματική συνδιακύμανση είναι οι εκτιμήσεις που προκύπτουν με τη μέθοδο των ροπών για τις παραμέτρους λ_1, λ_2 και λ_3, αντίστοιχα. Επιπλέον, καθώς $Cov(X_1, X_2) = \lambda_3$, με $\lambda_3 > 0$, έχουμε ότι η διμεταβλητή Poisson κατανομή μπορεί να χρησιμοποιηθεί για τη μοντελοποίηση θετικά συσχετισμένων δεδομένων.*

Ο Ορισμός 4.1.1 γενικεύεται στις m διαστάσεις ως εξής (Balakrishnan et al. (1998)).

Ορισμός 4.1.2. *Έστω $X_1 = Y_1 + Y_{m+1}$,...,$X_m = Y_m + Y_{m+1}$, με Y_1,...Y_{m+1} ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν κατανομή Poisson με παραμέτρους $\lambda_1' = \lambda_1 - \lambda_{m+1} > 0$,...,$\lambda_m' = \lambda_m - \lambda_{m+1} > 0$ και $\lambda_{m+1} > 0$, αντίστοιχα. Τότε λέμε ότι η από κοινού κατανομή των X_1, X_2,...,X_m είναι η m-διάστατη Poisson κατανομή με παράμετρο $\lambda = (\lambda_1, \lambda_2, ..., \lambda_{m+1})$. Σε αυτήν την περίπτωση γράφουμε ότι $(X_1, X_2, ...X_m) \sim MP(\lambda)$, όπου $\lambda = (\lambda_1, ..., \lambda_{m+1}) \in \Theta$, με $\Theta = \{(\lambda_1, ..., \lambda_{m+1}) \in \mathbb{R}^{m+1} : \lambda_1 > \lambda_{m+1}, ..., \lambda_m > \lambda_{m+1}, \lambda_{m+1} > 0\}$.*

Στην πρόταση που ακολουθεί διατυπώνεται μία χαρακτηριστική ιδιότητα της διδιάστατης Poisson. Για τη διατύπωση και την απόδειξή της στη γενική περίπτωση των m-διαστάσεων παραπέμπουμε στον Teicher (1954).

Πρόταση 4.1.2. *Αν $(X_1, X_2) \sim BP(\lambda)$ και $\lambda_i \to \infty$, για $i = 1, 2, 3$, έτσι ώστε $\frac{\lambda_3}{\sqrt{\lambda_1 \lambda_2}} \to \rho$, όπου ρ μία σταθερά, τότε η οριακή κατανομή του τυχαίου διανύσματος (Y_1, Y_2), με $Y_i = \frac{X - \lambda_i}{\sqrt{\lambda_i}}$, $i = 1, 2$, είναι διδιάστατη κανονική κατανομή με μηδενικό μέσο διάνυσμα και πίνακα διακυμάνσεων συνδιακυμάνσεων με διαγώνια στοιχεία ίσα με τη μονάδα και μη διαγώνια στοιχεία ίσα με $\rho \geq 0$. Επιπρόσθετα, ισχύει ότι η τυχαία μεταβλητή:*

$$\frac{Y_1^2 - 2\rho Y_1 Y_2 + Y_2^2}{1 - \rho^2},$$

ακολουθεί προσεγγιστικά χι-τετράγωνο κατανομή με 2 βαθμούς ελευθερίας, ήτοι την χ_2^2.

Με παρόμοιο σκεπτικό με εκείνο που διατυπώθηκε για τις μονοδιάστατες διακριτές κατανομές, στις διδιάστατες και πολυδιάστατες διακριτές κατανομές σημαντικό ρόλο διαδραματίζει η πιθανογεννήτρια συνάρτηση κατανομής, που ορίζεται ως $E(t_1^{X_1} t_2^{X_2})$ και $E(t_1^{X_1} t_2^{X_2} ... t_m^{X_m})$, για $m \geq 3$. Στην πρόταση που ακολουθεί προσδιορίζεται, χρησιμοποιώντας τον ορισμό και μετά από λίγη άλγεβρα, η πιθανογεννήτρια συνάρτηση της διδιάστατης και της m-διάστατης κατανομής Poisson.

Πρόταση 4.1.3. *Έστω $(X_1, X_2) \sim BP(\lambda)$ με $\lambda = (\lambda_1, \lambda_2, \lambda_3) \in \Theta$, τότε η πιθανογεννήτρια συνάρτηση δίνεται από τη σχέση:*

$$g(t_1, t_2; \lambda) = e^{\lambda_1(t_1-1)+\lambda_2(t_2-1)+\lambda_3(t_1-1)(t_2-1)}, \text{ με } t = (t_1, t_2) \in [0,1]^2.$$

Αν $(X_1, X_2, ..., X_m) \sim MP(\lambda)$ τότε:

$$g(t_1, ..., t_m; \lambda) = e^{\sum_{i=1}^m \lambda_i(t_i-1)+\lambda_{m+1}(\prod_{i=1}^m t_i - \sum_{i=1}^m t_i + m - 1)}.$$

Στην Πρόταση 1.2.11 είχε διατυπωθεί ότι η πιθανογεννήτρια συνάρτηση της κατανομής Poisson είναι η μόνη πιθανογεννήτρια συνάρτηση στο σύνολο των πιθανογεννητριών συναρτήσεων που αντιστοιχούν σε τυχαίες μεταβλητές με τιμές στο $\mathbb{N}_0 = \{0, 1, ...\}$ και πεπερασμένη μέση τιμή που ικανοποιεί τη διαφορική εξίσωση $\dfrac{\partial}{\partial t} g(t) - \lambda g(t) = 0$. Επισημαίνεται ότι η ιδιότητα αυτή αποτέλεσε την κεντρική ιδέα του ελέγχου των Baringhaus and Henze (1992).Στη συνέχεια παρατίθενται δύο ανάλογες ιδιότητες που ικανοποιούν η πιθανογεννήτρια συνάρτηση της διδιάστατης και m-διάστατης Poisson και οι οποίες ιδιότητες έχουν αποτελέσει το θεμέλιο για την κατασκευή ελέγχων καλής προσαρμογής. Για την απόδειξη των αποτελέσματων της πρότασης παραπέμπουμε στις εργασίες Novoa-Muñoz and Jiménez-Gamero (2014) και Novoa-Muñoz and Jiménez-Gamero (2016).

Πρόταση 4.1.4. *Έστω G_2 είναι το σύνολο των συναρτήσεων $g : [0,1]^2 \to R$, που είναι τέτοιες ώστε g να είναι μία πιθανογεννήτρια συνάρτηση, με τις μερικές παραγώγους $\frac{\partial}{\partial t_i} g(t_1, t_2)$, $i = 1, 2$, να υπάρχουν για κάθε $(t_1, t_2) \in [0,1]^2$. Τότε η πιθανογεννήτρια συνάρτηση της διδιάστατης Poisson είναι η μόνη πιθανογεννήτρια συνάρτηση στο παραπάνω σύνολο πιθανογεννητριών συναρτήσεων* *που ικανοποιεί το σύστημα εξισώσεων* $D_i(t;\lambda) = 0$, $i = 1, 2$, *όπου*

$$D_1(t;\lambda) = \frac{\partial}{\partial t_1} g(t_1, t_2; \lambda) - (\lambda_1 + \lambda_3(t_2 - 1))\, g(t_1, t_2; \lambda),$$

και

$$D_2(t;\lambda) = \frac{\partial}{\partial t_2} g(t_1, t_2; \lambda) - \{\lambda_2 + \lambda_3(t_1 - 1)\} g(t_1, t_2; \lambda).$$

Η γενίκευση της παραπάνω πρότασης στις m-διαστάσεις διατυπώνεται στην πρόταση που ακολουθεί.

Πρόταση 4.1.5. *Έστω* G_m *είναι το σύνολο των συναρτήσεων* $g : [0,1]^m \to R$, *που είναι τέτοιες ώστε* g *να είναι μία πιθανογεννήτρια συνάρτηση, με τις μερικές παραγώγους* $\frac{\partial}{\partial t_i} g(t_1, ..., t_m)$, $i = 1, 2, .., m$ *να υπάρχουν για κάθε* $(t_1, ..., t_m) \in [0,1]^m$. *Τότε η πιθανογεννήτρια συνάρτηση της* m*-διάστατης Poisson είναι η μόνη πιθανογεννήτρια συνάρτηση στο παραπάνω σύνολο πιθανογεννητριών συναρτήσεων που ικανοποιεί το σύστημα εξισώσεων* $D_i(t;\lambda) = 0$, $i = 1, ..., m$, *όπου για* $i = 1, ..., m$ *είναι:*

$$D_i(t;\lambda) = \frac{\partial}{\partial t_i} g(t) - \lambda_i + \lambda_{m+1} \left(\prod\nolimits_{j \neq i} t_j - 1 \right) g(t).$$

Τέλος, ένας νέος χαρακτηρισμός, ο οποίος βασίζεται στην πιθανογεννήτρια συνάρτηση, δόθηκε πρόσφατα για τη διδιάστατη περίπτωση από τον Novoa-Muñoz (2021) και παρατίθεται στην πρόταση που ακολουθεί.

Πρόταση 4.1.6. *Έστω* G_2 *είναι το σύνολο των συναρτήσεων* $g : [0,1]^2 \to R$, *που είναι τέτοιες ώστε* g *να είναι μία πιθανογεννήτρια συνάρτηση, με τις μερικές παραγώγους* $\frac{\partial}{\partial t_i} g(t_1, t_2)$, $i = 1, 2$ *να υπάρχουν για κάθε* $(t_1, t_2) \in [0,1]^2$. *Τότε η πιθανογεννήτρια συνάρτηση της διδιάστατης Poisson είναι η μόνη πιθανογεννήτρια συνάρτηση στο παραπάνω σύνολο πιθανογεννητριών συναρτήσεων* *που ικανοποιεί το σύστημα των εξισώσεων* $E_i(t;\lambda)=0$, $i=1,2,3$, *όπου*

$$E_1(t;\lambda)=\frac{\partial g(t_1,1)}{\partial t_1}-\lambda_1 g(t_1,1),$$

$$E_2(t;\lambda)=\frac{\partial g(1,t_2)}{\partial t_2}-\lambda_2 g(1,t_2),$$

και

$$E_3(t;\lambda)=\frac{\partial^2 g(t_1,t_2)}{\partial t_1\partial t_2}-\left(\lambda_3+\{\lambda_2+\lambda_3(t_1-1)\}\{\lambda_1+\lambda_3(t_2-1)\}\right)g(t_1,t_2).$$

Έστω $\boldsymbol{X}_1=(X_{11},...,X_{1d})$, $\boldsymbol{X}_2=(X_{21},...,X_{2d})$,..., $\boldsymbol{X}_n=(X_{n1},...,X_{nd})$ ένα τυχαίο δείγμα μεγέθους n από έναν d-διαστατο πληθυσμό με τιμές στο N_0^d, όπου $N_0=\{0,1,...,\}$. Τότε η εμπειρική πιθανογεννήτρια συνάρτηση συμβολίζεται με $g_n(t)$ και ορίζεται ως εξής:

$$g_n(t)=\frac{1}{n}\sum_{i=1}^{n}t_1^{X_{i1}}t_2^{X_{i2}}\dots t_d^{X_{id}}\quad ,t\in W, \tag{4.1}$$

για κάποιο κατάλληλο σύνολο W που είναι υποσύνολο του R^d.

Η ε.π.σ. που ορίζεται για m-διάστατα δεδομένα πληροί, κατά αναλογία με την κλασική ε.π.σ., κάποιες πολύ χρήσιμες ιδιότητες που εξηγούν τη χρησιμότητά της στη στατιστική συμπερασματολογία και στον έλεγχο καλής προσαρμογής. Κάποιες από αυτές τις ιδιότητες παρατίθενται στη συνέχεια (βλέπε Novoa-Muñoz and Jiménez-Gamero (2014)). Σε όσα ακολουθούν για οποιαδήποτε συνάρτηση $h:S\subset\mathbb{R}^d\to\mathbb{R}$ είναι

$$D^{a_1\dots a_d}h(u)=\frac{\partial^k}{\partial u_1^{a_1}\dots\partial u_d^{a_d}}\ h(u),$$

$\forall a_1,\dots,a_d\in\mathbb{N}_0$, τέτοια ώστε $k=a_1+\dots+a_d$.

Θεώρημα 4.1.7. *Αν $g_n(t)$ είναι η εμπειρική πιθανογεννήτρια συνάρτηση ενός τυχαίου δείγματος $\boldsymbol{X}_1 = (X_{11}, ..., X_{1d})$, $\boldsymbol{X}_2 = (X_{21}, ..., X_{2d})$,..., $\boldsymbol{X}_n = (X_{n1}, ..., X_{nd})$, από έναν m-διάστατο πληθυσμό με τιμές στον $\mathbb{N}_0^d$ με πιθανογεννήτρια συνάρτηση $g(t)$, ορισμένη στο σύνολο W που είναι υποσύνολο του $\mathbb{R}^d$ τότε:*

α) Ισχύει

$$\sup_{u \in R^d} |g_n(u) - g(u)| \xrightarrow{\sigma.\beta.} 0. \tag{4.2}$$

β) Επιπλέον, αν $D^{a_1 \cdots a_d} g(u)$ υπάρχει στο R, τότε:

$$\sup_{u \in R^d} |D^{a_1 \ldots a_d} g_n(u) - D^{a_1 \cdots a_d} g(u)| \xrightarrow{\sigma.\beta.} 0. \tag{4.3}$$

4.2 Έλεγχοι για τη διδιάστατη κατανομή Poisson

Στην ενότητα αυτή θα παρουσιαστούν εν συντομία οι κυριότεροι, και όχι όλοι, έλεγχοι καλής προσαρμογής για τη διδιάστατη Poisson κατανομή.

4.2.1 Γενικεύσεις των δεικτών διασποράς του Fisher

Παρακινούμενοι από την ιδέα ότι ο δείκτης διασποράς του Fisher χρησιμοποιείται συχνά για τον έλεγχο της υπόθεσης ότι τα δεδομένα προέρχονται από την κατανομή Poisson, οι Loukas and Kemp (1986) προτείνουν έναν έλεγχο που στηρίζεται σε μία στατιστική συνάρτηση την οποία καλούν διδιάστατο δείκτη διασποράς. Η σ.σ. που προτάθηκε από τους Loukas and Kemp (1986) αξιοποιεί την ιδιότητα της διδιάστατης Poisson που διατυπώθηκε στην Πρόταση 4.1.2.

Ειδικότερα, αν $\boldsymbol{X}_i = (X_{i1}, X_{i2})$, $i = 1, ..., n$, είναι n το πλήθος διδιάστατα τυχαία διανύσματα από τη διδιάστατη κατανομή Poisson με γνωστό διάνυσμα παραμέτρων λ, τότε με βάση το αποτέλεσμα της Πρότασης 4.1.2, έχουμε ότι

$$\frac{1}{1-\rho^2}\sum_{i=1}^{n}(W_{i1}^2 - 2\rho W_{i1}W_{i2} + W_{i2}^2) \xrightarrow{d} \chi^2_{2n},$$

όπου $\rho = \dfrac{\lambda_3}{\sqrt{\lambda_1\lambda_2}}$ και $W_{ik} = \dfrac{X_{ik}-\lambda_k}{\sqrt{\lambda_k}}$, για $k = 1, 2$ και $i = 1, 2, ..., n$.

Καθώς στην πράξη οι παράμετροι λ_i είναι άγνωστοι, οι Loukas and Kemp (1986) προτείνουν να αντικαθίστανται με τους εκτιμητές τους με τη μέθοδο των ροπών. Δηλαδή, προτείνουν τη χρήση των

$$\hat{\lambda}_j = \bar{X}_j = \frac{\sum_{i=1}^n X_{ij}}{n},\ j = 1, 2$$

και

$$\hat{\lambda}_3 = S_{X_1X_2} = \frac{1}{n}\sum_{i=1}^{n}(X_{i1}-\bar{X}_1)(X_{i2}-\bar{X}_2).$$

Τότε προκύπτει η σ.σ.

$$I_B = \frac{\sum_{i=1}^n \left[\dfrac{(X_{i1}-\bar{X}_1)^2}{\bar{X}_1} - \dfrac{2S_{X_1X_2}(X_{i1}-\bar{X}_1)(X_{i2}-\bar{X}_2)}{\bar{X}_1\bar{X}_2} + \dfrac{(X_{i2}-\bar{X}_2)^2}{\bar{X}_2}\right]}{1-\dfrac{S^2_{X_1X_2}}{\bar{X}_1\bar{X}_2}},$$

η οποία ισοδύναμα μπορεί να γραφεί

$$I_B = n\frac{\bar{X}_2 S^2_{X_1} - 2S^2_{X_1X_2} + \bar{X}_1 S^2_{X_2}}{\bar{X}_1\bar{X}_2 - S^2_{X_1X_2}}, \tag{4.4}$$

με $S^2_{X_j}$, για $j = 1, 2$, να είναι οι δειγματικές διακυμάνσεις, ενώ όπως πρωτύτερα ορίστηκε $S_{X_1X_2}$ είναι η δειγματική συνδιακύμανση. Τότε καθώς εκτιμήθηκαν τρεις πληθυσμιακές παράμετροι σε επίπεδο σημαντικότητας α η μηδενική υπόθεση απορρίπτεται αν $I_B \geq \chi^2_{2n-3,\alpha}$.

Όπως προαναφέρθηκε, η σ.σ. I_B προέκυψε χρησιμοποιώντας τους εκτιμητές που προκύπτουν με χρήση της μεθόδου των ροπών. Οι Rayner and Best (1997) παρατήρησαν με μελέτες προσομοίωσης ότι όσο μεγαλώνει η τιμή της παραμέτρου $\rho = \frac{\lambda_3}{\sqrt{\lambda_1\lambda_2}}$ αυξάνει η πιθανότητα η τιμή της σ.σ. I_B να είναι αρνητική και αυτό έχει ως συνέπεια η κατανομή της να μην προσεγγίζεται ικανοποιητικά από τη χι-τετράγωνο κατανομή. Επιπρόσθετα, όταν $\widehat{\rho}^2 > \frac{1}{2}\left(\frac{S^2_{X_1}}{\bar{X}_1 +}\frac{S^2_{X_2}}{\bar{X}_2}\right)$, προκύπτει ότι $I_B < 0$.

Για την αντιμετώπιση τον παραπάνω προβλημάτων, οι Rayner and Best (1997), πρότειναν την εκτίμηση της άγνωστης παραμέτρου από τον συντελεστή συσχέτισης του Pearson. Κατά αυτόν τον τρόπο, οι Rayner and Best (1997) πρότειναν τη σ.σ. που ορίζεται από τη σχέση:

$$NI_B = \frac{n}{1-r^2}\left(\frac{S^2_{X_1}}{\bar{X}_1} - 2r^2\sqrt{\frac{S^2_{X_1}S^2_{X_2}}{\bar{X}_1\bar{X}_2}} + \frac{S^2_{X_2}}{\bar{X}_2}\right), \tag{4.5}$$

όπου r ο συντελεστής συσχέτισης του δείγματος.

Για τη σ.σ. NI_B ισχύει ότι για μεγάλο μέγεθος δείγματος n ακολουθεί προσεγγιστικά χ^2_{2n-3} και η μηδενική υπόθεση απορρίπτεται, σε ασυμπτωτικό επίπεδο σημαντικότητας α, αν $NI_B \geq \chi^2_{2n-3,\alpha}$.

Ένας ακόμη έλεγχος που μπορεί να θεωρηθεί επέκταση του ελέγχου που στηρίζεται στον δείκτη διασποράς είναι ο έλεγχος που προτάθηκε από τους Rayner and Best (1997) και αποτελεί ουσιαστικά μία αναθεωρημένη έκδοση του ελέγχου που προτάθηκε από τον Crockett (1979). Ο έλεγχος αυτός βασίζεται σε μία σ.σ. που δεν είναι τίποτε άλλο παρά μία τυποποιημένη μορφή του Poisson δείκτη. Πιο συγκεκριμένα, η σ.σ. που προτάθηκε δίνεται από τη σχέση:

$$T = \frac{n}{2}\frac{\bar{X_2}^2(S^2_{X_1} - \bar{X_1})^2 - 2S^2_{X_1X_2}(S^2_{X_1} - \bar{X_1})(S^2_{X_2} - \bar{X_2}) + \bar{X_1}^2(S^2_{X_2} - \bar{X_2})^2}{\bar{X_1}^2\bar{X_2}^2 - S^4_{X_1X_2}}. \tag{4.6}$$

Ουσιαστικά, ο έλεγχος αυτός αξιολογεί πώς οι διακυμάνσεις των δεδομένων διαφέρουν από αυτές που αναμένονται υπό τη μηδενική υπόθεση. Αποδεικνύεται ότι η ασυμπτωτική κατανομή της σ.σ. T, υπό τη μηδενική υπόθεση, είναι χι-τετράγωνο με 2 βαθμούς ελευθερίας και επομένως προκύπτει ότι, σε ασυμπτωτικό επίπεδο σημαντικότητας α, απορρίπτεται η μηδενική υπόθεση αν $T \geq \chi^2_{2,\alpha}$.

4.2.2 Έλεγχοι με την ε.π.σ.

Στην ενότητα αυτή θα παρουσιαστούν έλεγχοι καλής προσαρμογής που έχουν προταθεί στις εργασίες των Novoa-Muñoz and Jiménez-Gamero (2014), Novoa-Muñoz and Jiménez-Gamero (2016), Novoa-Muñoz (2021) και στηρίζονται σε ιδιότητες της ε.π.σ. και της πιθανογεννήτριας συνάρτησης της διδιάστατης Poisson.

Ειδικότερα, οι Novoa-Muñoz and Jiménez-Gamero (2014) παρατηρούν αρχικά ότι από το Θεώρημα 4.1.7 ένας συνεπής εκτιμητής της πιθανογεννήτριας συνάρτησης είναι η ε.π.σ., ενώ αν $\hat{\lambda}_n$ είναι συνεπής εκτιμητής της άγνωστης παραμέτρου λ, τότε η εκτιμώμενη, υπό τη μηδενική υπόθεση, πιθανογεννήτρια συνάρτηση $g(t, \hat{\lambda}_n)$ είναι συνεπής εκτιμητής της πιθανογεννήτριας συνάρτησης $g(t, \lambda)$. Λαμβάνοντας επιπλέον υπόψιν ότι η κατανομή ενός τυχαίου διανύσματος προσδιορίζεται μοναδικά από την πιθανογεννήτρια συνάρτηση προτείνουν την ακόλουθη σ.σ.:

$$R_{n,w}(\hat{\lambda}_n) = n\int_0^1\int_0^1 (\{g_n(t) - g(t; \hat{\lambda}_n)\})^2 w(t)dt, \tag{4.7}$$

με $w(t)$ μια συνάρτηση βάρους που είναι τέτοια ώστε $w(t) \geq 0$, $\forall t \in [0,1]^2$ και

$$\int_0^1 \int_0^1 w(t)dt < \infty,$$

έτσι ώστε να εξασφαλίζεται ότι είναι πεπερασμένο για κάθε n το διπλό ολοκλήρωμα που εμφανίζεται στην έκφραση της σ.σ. $R_{n,w}(\hat{\lambda}_n)$. Μία τέτοια συνάρτηση που προτάθηκε από τους Novoa-Muñoz and Jiménez-Gamero (2014) είναι η $w(t) = t_1^{a_1} t_2^{a_2}$, με $a_1, a_2 \in (-1, \infty)$.

Από τον τρόπο κατασκευής της σ.σ.ε. προκύπτει άμεσα ότι η μηδενική υπόθεση απορρίπτεται για μεγάλες τιμές της σ.σ. $R_{n,w}(\hat{\lambda}_n)$. Παρότι έχουν προσδιοριστεί οι ασυμπτωτικές ιδιότητες της σ.σ. $R_{n,w}(\hat{\lambda}_n)$ στην πράξη χρησιμοποιείται παραμετρικό bootstrap, το οποίο εκτιμά με συνέπεια την κατανομή της στατιστικής συνάρτησης υπό τη μηδενική υπόθεση. Δηλαδή, ο έλεγχος που προκύπτει με παραμετρικό bootstrap έχει ασυμπτωτικό επίπεδο σημαντικότητας α, καθώς το μέγεθος δείγματος και το πλήθος bootstrap δειγμάτων τείνουν στο άπειρο (βλέπε Novoa-Muñoz and Jiménez-Gamero (2014)).

Παρατήρηση 4.2.1. *Ο παραπάνω έλεγχος ουσιαστικά αποτελεί μια γενίκευση στις δύο διαστάσεις των ελέγχων που προτάθηκαν από τους Rueda et al. (1991) και Baringhaus et al. (2000) για τον στατιστικό έλεγχο καλής προσαρμογής στην περίπτωση της μονοδιάστατης κατανομής Poisson. Οι έλεγχοι αυτοί παρουσιάστηκαν στην Ενότητα 2.4.*

Δύο ακόμη έλεγχοι καλής προσαρμογής για τη διδιάστατη κατανομή Poisson που στηρίζονται στην ε.π.σ. έχουν προταθεί στη βιβλιογραφία στις εργασίες Novoa-Muñoz and Jiménez-Gamero (2014) και Novoa-Muñoz (2021). Οι έλεγχοι αυτοί στηρίζονται αυτήν τη φορά στην ιδιότητα που πληροί η διδιάστατη Poisson και έχει διατυπωθεί στην Πρόταση 4.1.4, ήτοι ότι η πιθανογεννήτρια της διδιάστατης Poisson είναι η μοναδική πιθανογεννήτρια συνάρτηση στο σύνολο G_2 που ικανοποιεί το ακόλουθο σύστημα εξισώσεων $D_i(t;\lambda)=0$, $i=1,2$, όπου τα $D_i(t;\lambda)$ ορίστηκαν στην Ενότητα 4.1. Αξιοποιώντας αυτήν την ιδιότητα και το γεγονός ότι (βλέπε Θεώρημα 4.1.7) η εμπειρική πιθανογεννήτρια συνάρτηση και οι παράγωγοί της είναι συνεπείς εκτιμητές της πιθανογεννήτριας συνάρτησης και των παραγώγων της, αντίστοιχα, συμπεραίνουν ότι υπό τη μηδενική υπόθεση τα δειγματικά ανάλογα των $D_i(t,\lambda)$ που ορίζονται από τις σχέσεις:

$$D_{1n}(t;\hat{\lambda}_n)=\frac{\partial}{\partial t_1}g_n(t_1,t_2)-\{\widehat{\lambda}_{1n}+\widehat{\lambda}_{3n}(t_2-1)\}g_n(t_1,t_2),$$

και

$$D_{2n}(t;\hat{\lambda}_n)=\frac{\partial}{\partial t_2}g_n(t_1,t_2)-\{\widehat{\lambda}_{2n}+\widehat{\lambda}_{3n}(t_1-1)\}g_n(t_1,t_2),$$

θα πρέπει να λαμβάνουν και αυτά τιμές κοντά στο 0. Η παραπάνω σκέψη ανάλογα με τον τρόπο που θα ερμηνευθεί οδηγεί σε δύο διαφορετικούς ελέγχους.

Ειδικότερα, οι Novoa-Muñoz and Jiménez-Gamero (2014), παρακινούμενοι από το σκεπτικό του ελέγχου που προτάθηκε από τους Baringhaus and Henze (1992) στην περίπτωση της μονοδιάστατης Poisson, προτείνουν τη σ.σ. που δίνεται από τη σχέση:

$$S_{n,w}(\hat{\lambda}_n)=n\int_0^1\int_0^1\{D_{1n}^2(t;\widehat{\lambda}_n)+D_{2n}^2(t;\widehat{\lambda}_n)\}w(t)dt, \tag{4.8}$$

με $w(t)$ μία συνάρτηση βάρους που ικανοποιεί τις ιδιότητες που προηγούμενα αναφέρθηκαν. Προφανώς, από τον τρόπο κατασκευής της σ.σ.ε. προκύπτει άμεσα ότι η μηδενική υπόθεση απορρίπτεται για μεγάλες τιμές της σ.σ. $S_{n,w}(\hat{\lambda}_n)$. Παρότι έχουν προσδιοριστεί οι ασυμπτωτικές ιδιότητες της σ.σ. $S_{n,w}(\hat{\lambda}_n)$ στην πράξη χρησιμοποιείται παραμετρικό bootstrap, και ο έλεγχος που προκύπτει κατά αυτόν τον τρόπο έχει ασυμπτωτικό επίπεδο σημαντικότητας α, καθώς το μέγεθος δείγματος και το πλήθος bootstrap δειγμάτων τείνουν στο άπειρο (βλέπε Novoa-Muñoz and Jiménez-Gamero (2014)).

Οι Novoa-Muñoz and Jiménez-Gamero (2016), παρακινούμενοι από το σκεπτικό που αναπτύχθηκε από τους Nakamura and Pérez-Abreu (1993) στην περίπτωση του ελέγχου της μονοδιάστατης Poisson, πρότειναν και έναν άλλο τρόπο ερμηνείας της ιδιότητας ότι υπό τη μηδενική υπόθεση τα δειγματικά ανάλογα $D_{in}(t, \hat{\lambda}_n)$, $i = 1, 2$, είναι κοντά στο μηδέν. Ειδικότερα, αρχικά, παρατήρησαν ότι:

$$D_{in}(t;\widehat{\lambda}) = \sum_{r_1 \geq 0} \sum_{r_2 \geq 0} d_i(r_1, r_2; \widehat{\lambda}) t_1^{r_1} t_2^{r_2}, \text{ για } i = 1, 2,$$

με

$$d_1(r_1, r_2; \widehat{\lambda}) = (r_1 + 1) p_n(r_1 + 1, r_2) - (\widehat{\lambda}_1 - \widehat{\lambda}_3) p_n(r_1, r_2) - \widehat{\lambda}_3 p_n(r_1, r_2 - 1),$$

$$d_2(r_1, r_2; \widehat{\lambda}) = (r_2 + 1) p_n(r_1, r_2 + 1) - (\widehat{\lambda}_2 - \widehat{\lambda}_3) p_n(r_1, r_2) - \widehat{\lambda}_3 p_n(r_1 - 1, r_2),$$

ενώ

$$p_n(r_1, r_2) = \frac{1}{n} \sum_{k=1}^{n} I(X_{k1} = r_1, X_{k2} = r_2),$$

η σχετική συχνότητα εμφάνισης του ζεύγους (r_1, r_2). Τότε, άμεσα, προκύπτει ότι $D_{in}(t;\widehat{\lambda}) = 0$ αν και μόνο αν οι πολυωνυμικοί συντελεστές του $t_1^{r_1} t_2^{r_2}$ είναι ίσοι με το μηδέν $\forall r_1, r_2 \geq 0$. Η σκέψη αυτή τους οδήγησε να προτείνουν τη σ.σ.:

$$W_n(\widehat{\lambda}) = \sum_{r_1 \geq 0} \sum_{r_2 \geq 0} \{d_1(r_1, r_2; \widehat{\lambda})^2 + d_2(r_1, r_2; \widehat{\lambda})^2\},$$

η οποία ισοδύναμα μπορεί να γραφεί στη μορφή

$$W_n(\widehat{\lambda}) = \sum_{r_1,r_2=0}^{M} \{d_1(r_1,r_2;\widehat{\lambda})^2 + d_2(r_1,r_2;\widehat{\lambda})^2\},$$

όπου $M = \max\{X_{(n)1}, X_{(n)1}\}$, με $X_{(n)k} = \max_{1\le i\le n} X_{ik}$, $k = 1, 2$.

Προφανώς, από τον τρόπο κατασκευής της σ.σ.ε. προκύπτει άμεσα ότι η μηδενική υπόθεση απορρίπτεται για μεγάλες τιμές της σ.σ. $W_n(\widehat{\lambda})$. Παρότι έχουν προσδιοριστεί οι ασυμπτωτικές ιδιότητες της σ.σ. $W_n(\widehat{\lambda})$ στην πράξη χρησιμοποιείται παραμετρικό bootstrap, και ο έλεγχος που προκύπτει κατά αυτόν τον τρόπο έχει ασυμπτωτικό επίπεδο σημαντικότητας α, καθώς το μέγεθος δείγματος και το πλήθος bootstrap δειγμάτων τείνουν στο άπειρο (βλέπε Novoa-Muñoz and Jiménez-Gamero (2016)).

Τέλος, με ανάλογο σκεπτικό και χρησιμοποιώντας την ιδιότητα που δόθηκε στην Πρόταση 4.1.6, ο Novoa-Muñoz (2021) πρότεινε τη σ.σ.:

$$T_{n,w}(\widehat{\lambda}_n) = n\int_0^1\int_0^1 \{E_{1n}^2(t;\widehat{\lambda}_n) + E_{2n}^2(t;\widehat{\lambda}_n) + E_{3n}^2(t;\widehat{\lambda}_n)\}w(t)dt, \tag{4.9}$$

με $E_{in}(t;\widehat{\lambda}_n)$ τα δειγματικά ανάλογα των $E_i(t;\lambda)$ που προκύπτουν με αντικατάσταση της πιθανογεννητριας συνάρτησης από την εμπειρική πιθανογεννήτρια συνάρτηση και της άγνωστης παραμέτρου από έναν συνεπή εκτιμητή της. Δηλαδή, είναι

$$E_{1n}(t;\widehat{\lambda}_n) = \frac{\partial g_n(t_1,1)}{\partial t_1} - \widehat{\lambda}_{1n} g_n(t_1,1),$$

$$E_{2n}(t;\widehat{\lambda}_n) = \frac{\partial g_n(1,t_2)}{\partial t_2} - \widehat{\lambda}_{2n} g_n(1,t_2),$$

και

$$E_{3n}(t;\widehat{\lambda}_n) = \frac{\partial^2 g(t_1,t_2)}{\partial t_1 \partial t_2} - \widehat{\lambda}_{3n} - \{\widehat{\lambda}_{2n} + \widehat{\lambda}_{3n}(t_1 - 1)\}\{\widehat{\lambda}_{1n} + \widehat{\lambda}_{3n}(t_2 - 1)\}g_n(t_1,t_2).$$

Προφανώς, από τον τρόπο κατασκευής της σ.σ.ε. προκύπτει άμεσα ότι η μηδενική υπόθεση απορρίπτεται για μεγάλες τιμές της σ.σ. $T_{n,w}(\widehat{\lambda}_n)$. Παρότι έχουν προσδιοριστεί οι ασυμπτωτικές ιδιότητες της σ.σ. στην πράξη χρησιμοποιείται παραμετρικό bootstrap, και ο έλεγχος που προκύπτει κατά αυτόν τον τρόπο έχει ασυμπτωτικό επίπεδο σημαντικότητας α, καθώς το μέγεθος δείγματος και το πλήθος bootstrap δειγμάτων τείνουν στο άπειρο (βλέπε Novoa-Muñoz (2021)).

Παρατήρηση 4.2.2. *Από μελέτη προσομοίωσης που έχει διεξαχθεί από τους Novoa-Muñoz and Jiménez-Gamero (2016) και από τον Novoa-Muñoz (2021) έχει εξαχθεί το συμπέρασμα ότι οι έλεγχοι που στηρίζονται στην ε.π.σ. μπορούν να ανιχνεύσουν όλες τις αποκλίσεις που έχουν θεωρηθεί στις εναλλακτικές κατανομές που θεωρήθηκαν. Κάτι τέτοιο δεν επιτυγχάνεται από τους ελέγχους που παρουσιάστηκαν στην πρώτη υποενότητα, γεγονός που ήταν αναμενόμενο, αφού οι έλεγχοι αυτοί είναι μη συνεπείς διότι βασίζονται σε ιδιότητες των πληθυσμιακών ροπών. Επιπρόσθετα, οι Novoa-Muñoz and Jiménez-Gamero (2016) εξάγουν το συμπέρασμα ότι για τις μισές εκ των εναλλακτικών η απόδοση των ελέγχων που στηρίζονται στην ε.π.σ. είναι περίπου ίδια, ενώ στις υπόλοιπες ο έλεγχος που προτάθηκε στην εργασία τους έχει καλύτερη απόδοση. Τέλος, από τη μελέτη προσομοίωσης που διεξήχθη από τον Novoa-Muñoz (2021) εξάγεται το συμπέρασμα ότι ο έλεγχος του είναι ανταγωνιστικός στους υπόλοιπους που βασίζονται στην ε.π.σ.*

4.3 Έλεγχοι για την πολυδιάστατη κατανομή Poisson

Στην προηγούμενη ενότητα παρουσιάστηκαν έλεγχοι που αφορούν τη διδιάστατη Poisson. Ωστόσο, αρκετά συχνά τα διαθέσιμα απαριθμητά δεδομένα είναι m-διάστατα και υπάρχει αναγκαιότητα για τον έλεγχο καλής προσαρμογής της m-διάστατης κατανομής Poisson. Η αναγκαιότητα αυτή έχει αντιμετωπιστεί στις εργασίες των Novoa-Muñoz and Jiménez-Gamero (2014) και Novoa-Muñoz and Jiménez-Gamero (2016). Ειδικότερα, κατά αναλογία με τους ελέγχους που παρουσιάστηκαν στις ίδιες εργασίες για τη διδιάστατη περίπτωση, ο πρώτος έλεγχος από αυτούς στηρίζεται στο γεγονός ότι η εμπειρική πιθανογεννήτρια συνάρτηση είναι συνεπής εκτιμητής της άγνωστης πιθανογεννήτριας συνάρτησης, ενώ οι άλλοι δύο έλεγχοι στηρίζονται στην ιδιότητα που πληροί η m-διάστατη Poisson και έχει διατυπωθεί στην Πρόταση 4.1.5. Δηλαδή, οι δύο άλλοι έλεγχοι στηρίζονται στο γεγονός ότι η πιθανογεννητρια της m-διάστατης Poisson είναι η μοναδική πιθανογεννήτρια συνάρτηση στο σύνολο G_m που ικανοποιεί το ακόλουθο σύστημα εξισώσεων $D_i(t;\lambda)=0$, $i=1,...,m$, όπου τα $D_i(t;\lambda)$ ορίστηκαν στην Ενότητα 4.1.

Στη συνέχεια παρατίθενται οι σχέσεις προσδιορισμού αυτών των στατιστικών συναρτήσεων. Σε όσα ακολουθούν με $\hat{\lambda}_n=(\hat{\lambda}_{1,n},...,\hat{\lambda}_{m+1,n})$ συμβολίζουμε έναν συνεπή εκτιμητή της παραμέτρου $\lambda=(\lambda_1,...,\lambda_{m+1})$.

Πιο συγκεκριμένα, ο πρώτος τρόπος ελέγχου στηρίζεται στη σ.σ.:

$$R_{m,n,w}(\hat{\lambda}_n)=\int_{[0,1]^m}(\sqrt{n}\{g_n(t)-g_{\widehat{\lambda}}(t)\})^2 w(t)dt, \tag{4.10}$$

με $w(t)$ μια συνάρτηση βάρους που είναι τέτοια ώστε $w(t) \geq 0$, $\forall t \in [0,1]^m$ και

$$\int_{[0,1]^m} w(t)dt < \infty,$$

έτσι ώστε να εξασφαλίζεται ότι το ολοκλήρωμα που εμφανίζεται στην έκφραση της σ.σ. $R_{m,n,w}(\hat{\lambda}_n)$ είναι πεπερασμένο για κάθε n. Μία τέτοια συνάρτηση που προτάθηκε από τους Novoa-Muñoz and Jiménez-Gamero (2014) είναι η $w(t) = t_1^{a_1} t_2^{a_2} ... t_m^{a_m}$, με $a_i \in (-1, \infty)$, για $i = 1, ..., m$.

Ο δεύτερος τρόπος ελέγχου στηρίζεται στη σ.σ.:

$$S_{m,n,w}(\hat{\lambda}_n) = n \int_{[0,1]^m} \{D_{1n}^2(t; \widehat{\lambda}_n) + ... + D_{mn}^2(t; \widehat{\lambda}_n)\} w(t)dt, \qquad (4.11)$$

με $w(t)$ μία συνάρτηση βάρους που ικανοποιεί τις ιδιότητες που προηγούμενα αναφέρθηκαν, ενώ για $i = 1, ..., m$ είναι:

$$D_{in}(t; \hat{\lambda}_n) = \frac{\partial}{\partial t_i} g_n(t) - \{\hat{\lambda}_{i,n} + \hat{\lambda}_{m+1,n} \left(\prod\nolimits_{j \neq i} t_j - 1\right)\} g_n(t).$$

Τέλος, ο τρίτος τρόπος ελέγχου στηρίζεται στη σ.σ.:

$$W_{m,n}(\widehat{\lambda}) = \sum_{r_1 ... r_m \geq 0} \left(\sum_{i=1}^{m} d_i(r_1, ..., r_m; \widehat{\lambda})^2 \right) = \sum_{r_1 ... r_m = 0}^{M} \left(\sum_{i=1}^{m} d_i(r_1, ..., r_m; \widehat{\lambda})^2 \right),$$

όπου

$$\begin{aligned} d_i(r_1, ..., r_m; \widehat{\lambda}) &= (r_i + 1) p_n(r_1 ... r_{i-1}, r_i + 1, r_{i+1}, ... r_m) \\ &- (\widehat{\lambda}_i - \widehat{\lambda}_{m+1}) p_n(r_1, ..., r_m) \\ &- \widehat{\lambda}_{m+1} p_n(r_{i-1}, ... r_{i-1} - 1, r_i, r_{i+1} - 1, ..., r_m - 1) \end{aligned}$$

με $p_n(r_1,...,r_m) = \dfrac{1}{n}\sum_{k=1}^{n} I(X_{k1} = r_1,...,X_{km} = r_m)$ η σχετική συχνότητα εμφάνισης της m-άδας $(r_1,...,r_m)$, ενώ $M = \max\{X_{(n)1},...,X_{(n)m}\}$, με $X_{(n)k} = \max\limits_{\{1\leq i\leq n\}} X_{ik}$, $1 \leq k \leq m$.

Προφανώς, από τον τρόπο κατασκευής των σ.σ.ε. προκύπτει άμεσα ότι η μηδενική υπόθεση απορρίπτεται για μεγάλες τιμές αυτών. Παρότι έχουν προσδιοριστεί οι ασυμπτωτικές ιδιότητες των σ.σ. στην πράξη χρησιμοποιείται παραμετρικό bootstrap και οι έλεγχοι που προκύπτουν κατά αυτόν τον τρόπο έχουν ασυμπτωτικό επίπεδο σημαντικότητας α, καθώς το μέγεθος δείγματος και το πλήθος bootstrap δειγμάτων τείνουν στο άπειρο (βλέπε Novoa-Muñoz and Jiménez-Gamero (2014), Novoa-Muñoz and Jiménez-Gamero (2016)).

Παρατήρηση 4.3.1. *Ο έλεγχος που προτάθηκε από τους Novoa-Muñoz and Jiménez-Gamero (2016) έχει επεκταθεί στις τρεις διαστάσεις στην ίδια εργασία, ωστόσο είναι ιδιαίτερα πολύπλοκος και η γενίκευσή του σε περισσότερες διαστάσεις μπορεί να θεωρηθεί ανέφικτη.*

4.4 Επίλογος

Η Poisson αλλά και οι γενικεύσεις της σε περισσότερες διαστάσεις, βρίσκουν, όπως ήδη έχουμε αναφέρει, πλήθος εφαρμογών. Για τον λόγο αυτό έχουν προταθεί ποικίλοι τρόποι ελέγχου καλής προσαρμογής και ο σκοπός αυτής της διατριβής ήταν να παρουσιάσει τους σημαντικότερους εξ αυτών. Ωστόσο, θα πρέπει να επισημάνουμε ότι παρότι η κατανομή Poisson μοντελοποιεί τον αριθμό των συμβάντων στη μονάδα του χρόνου (όγκου, μήκους, εμβαδού) θα πρέπει να είμαστε ιδιαίτερα προσεκτικοί στη χρήση της, γιατί θα πρέπει να ικανοποιούνται οι Υποθέσεις 1-3 που παρατέθηκαν στο πρώτο κεφάλαιο. Επιπλέον, με βάση τις ιδιότητες της κατανομής Poisson, θα πρέπει να χρησιμοποιείται στη μοντελοποίηση τυχαίων φαινομένων όπου η μέση τιμή ταυτίζεται πρακτικά με τη διακύμανση, καθώς στην κατανομή Poisson, με παράμετρο λ είναι $E(X) = Var(X) = \lambda$. Από την άλλη καθώς από τη σ.π. της κατανομής Poisson προκύπτει ότι η πιθανότητα να μην πραγματοποιηθεί ένα γεγονός στη μονάδα του χρόνου είναι ίση με $P(X = 0) = e^{-\lambda}$ και επομένως $P(X \geq 1) = 1 - e^{-\lambda}$, η κατανομή Poisson δεν μπορεί να χρησιμοποιηθεί για τη μοντελοποίηση τυχαίων φαινομένων που είναι βέβαιη η πραγματοποίηση, εμφάνιση ενός γεγονότος. Τέλος, με την ίδια αιτιολόγηση δεν μπορεί να χρησιμοποιηθεί για τη μοντελοποίηση τυχαίων φαινομένων με μεγάλη πιθανότητα μη εμφάνισης γεγονότων σε συγκεκριμένο χρονικό διάστημα. Για όλους τους παραπάνω λόγους, έχουν εμφανιστεί γενικεύσεις της κατανομής Poisson με στόχο την μοντελοποίηση τέτοιων πραγματικών φαινομένων. Όλα τα παραπάνω έχουν οδηγήσει να εισαχθούν στη βιβλιογραφία εναλλακτικές κατανομές για τη μοντελοποίηση πραγματικών τυχαίων φαινομένων που αδυνατεί να μοντελοποιήσει η κατανομή Poisson. Ενδεικτικά, αναφέρονται η Neyman τύπου A, η κατανομή Bell, η Bell Touchard και η οικογένεια των γενικευμένων Poisson. Προφανώς, η εισαγωγή αυτών των κατανομών είχε ως συνέπεια και την εισαγωγή αντίστοιχων ελέγχων καλής προσαρμογής (βλέπε Batsidis and Lemonte (2021), Batsidis et al. (2020), Meintanis (2008) και στις εκεί αναφορές). Κατά αντίστοιχο τρόπο, στην περίπτωση των περισσότερων διαστάσεων ενδεικτικά παραπέμπουμε στην εργασία του Meintanis (2007) και στις εκεί αναφορές που έχουν προταθεί. Η συγκριτική μελέτη των ελέγχων προσαρμογής για αυτές τις κατανομές και η εισαγωγή νέων χρησιμοποιώντας ιδιότητες και χαρακτηριστικά τους είναι ένα θέμα που χρήζει περαιτέρω έρευνας.

Βιβλιογραφία

Anderson, J. C. and Siddiqui, M. M. (1994). The sampling distribution of the index of dispersion, *Communications in Statistics-Theory and Methods* **23**(3): 897–911.

Anderson, T. and Darling, D. (1952). Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes, *Annals of Mathematical Statistics* **23**: 193–212.

Avelino, E. G. (1978). *The index of dispersion*, B.Sc. Thesis, University of British Columbia.

Balakrishnan, N., Johnson, N. L. and Kotz, S. (1998). A note on relationships between moments, central moments and cumulants from multivariate distributions, *Statistics & Probability Letters* **39**(1): 49–54.

Baringhaus, L., Gürtler, N. and Henze, N. (2000). Theory & methods: weighted integral test statistics and components of smooth tests of fit, *Australian & New Zealand Journal of Statistics* **42**(2): 179–192.

Baringhaus, L. and Henze, N. (1992). A goodness of fit test for the Poisson distribution based on the empirical generating function, *Statistics & Probability Letters* **13**(4): 269–274.

Batsidis, A., Jiménez-Gamero, M. D. and Lemonte, A. J. (2020). On goodness-of-fit tests for the Bell distribution, *Metrika* **83**(3): 297–319.

Batsidis, A. and Lemonte, A. J. (2021). On Goodness-of-Fit Tests for the Neyman Type A Distribution: Accepted- November 2021, *REVSTAT-Statistical Journal* .

Beltrán-Beltrán, J. I. and O'Reilly, F. J. (2019). On goodness of fit tests for the Poisson, negative binomial and binomial distributions, *Statistical Papers* **60**(1): 1–18.

Best, D. J. and Rayner, J. C. W. (1999). Goodness of fit for the Poisson distribution, *Statistics & Probability Letters* **44**(3): 259–265.

Böhning, D. (1994). A note on a test for Poisson overdispersion, *Biometrika* **81**(2): 418–419.

Campbell, D. B. and Oprian, C. A. (1979). On the Kolmogorov-Smirnov test for the Poisson distribution with unknown mean, *Biometrical Journal* **21**(1): 17–24.

Cantelli, F. P. (1933). Sulla determinazione empirica delle leggi di probabilità, *Giorn. Ist. Ital. Attuari* **4**: 421–424.

Charlier, C. V. L. (1905). Uber die Darstellung willkurlicher Functione, *Arkiv for Matematik, Astronomi och Fysik* **2**: 1–35.

Conover, W. J. (1972). A Kolmogorov goodness-of-fit test for discontinuous distributions, *Journal of the American Statistical Association* **67**(339): 591–596.

Cramér, H. (1928). On the composition of elementary errors, *Scandinavian Actuarial Journal* **1**: 13–74.

Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society: Series B (Methodological)* **46**(3): 440–464.

Crockett, N. G. (1979). *Interactive Statistics*, D. McNeil(ed.), chapter A quick test of fit of a bivariate distribution.

D'Agostino, R. B. (1986). *Goodness-of-fit-techniques*, Routledge.

DasGupta, A. (2008). *Asymptotic theory of statistics and probability*, Springer.

Dubey, S. D. (1966). The Teacher's Corner: Graphical Tests for Discrete Distributions, *The American Statistician* **20**(3): 23–24.

Feller, W. (1968). An Introduction to Probability Theory and its applications.

Feuerverger, A. (1988). A bound for estimation in nonlinear time series models by independence testing methods, *Statistics & Probability Letters* **6**(4): 237–241.

Fisher, R. A. (1924). The Conditions Under Which χ^2 Measures the Discrepancey Between Observation and Hypothesis, *Journal of the Royal Statistical Society* **87**(3): 442–450.

Fisher, R. A. (1950). The significance of deviations from expectation in a Poisson series, *Biometrics* **6**(1): 17–24.

Fisher, R. A. and Thornton, H. G. (1922). The accuracy of the plating method of estimating the density of bacterial populations, *Annals of Applied Biology* **9**: 325–359.

Freeman, M. F. and Tukey, J. W. (1950). Transformations related to the angular and the square root, *The Annals of Mathematical Statistics* **21**(4): 607–611.

Frey, J. (2012). An exact Kolmogorov–Smirnov test for the Poisson distribution with unknown mean, *Journal of Statistical Computation and Simulation* **82**(7): 1023–1033.

Gan, F. F., Koehler, K. J. and Thompson, J. C. (1991). Probability plots and distribution curves for assessing the fit of probability models, *The American Statistician* **45**(1): 14–21.

Gart, J. J. and Pettigrew, H. M. (1970). On the conditional moments of the k-statistics for the Poisson distribution, *Biometrika* **57**(3): 661–664.

Gibbons, J. D. and Chakraborti, S. (2014). *Nonparametric Statistical Inference*, CRC Press.

Gibbons, J. D. and Chakraborti, S. (2020). *Nonparametric Statistical Inference, Fourth Edition Revised and Expanded*, Chapman and Hall/CRC.

Glivenko, V. (1933). Sulla determinazione empirica delle leggi di probabilità, *Giorn. Ist. Ital. Attuari* **4**: 92–99.

González-Barrios, J. M., O'Reilly, F. and Rueda, R. (2006). Goodness of fit for discrete random variables using the conditional density, *Metrika* **64**(1): 77–94.

Goovaerts, M. J., Kaas, R., Van Heerwaarden, A. E., Bauwelinckx, T. et al. (1990). *Effective actuarial methods*, Amsterdam North-Holland.

Gupta, A. K., Móri, T. F. and Székely, G. J. (1994). Testing for Poissonity-normality vs. other infinite divisibility, *Statistics & Probability Letters* **19**(3): 245–248.

Gürtler, N. and Henze, N. (2000). Recent and classical goodness-of-fit tests for the Poisson distribution, *Journal of Statistical Planning and Inference* **90**(2): 207–225.

Henze, N. (1996). Empirical-distribution-function goodness-of-fit tests for discrete models, *Canadian Journal of Statistics* **24**(1): 81–93.

Henze, N. and Klar, B. (1996). Properly rescaled components of smooth tests of fit are diagnostic, *Australian Journal of Statistics* **38**(1): 61–74.

Hoaglin, D. C. (1980). A Poissonness plot, *The American Statistician* **34**(3): 146–149.

Hoaglin, D. C., Mosteller, F. and Tukey, J. W. (2011). *Exploring data tables, trends, and shapes*, John Wiley & Sons.

Hoel, P. G. (1943). On indices of dispersion, *The Annals of Mathematical Statistics* **14**(2): 155–162.

Karlis, D. and Xekalaki, E. (2000). A simulation comparison of several procedures for testing the Poisson assumption, *Journal of the Royal Statistical Society: Series D (The Statistician)* **49**(3): 355–382.

Khamkong, M. M. (2010). *Asymptotic Tests for Poisson Distribution*, PhD thesis, School of Applied Statistics National Institute of Development Administration.

Klar, B. (1999). Goodness-of-fit tests for discrete models based on the integrated distribution function, *Metrika* **49**(1): 53–69.

Kocherlakota, S. and Kocherlakota, K. (1986). Goodness of fit tests for discrete distributions, *Communications in Statistics-Theory and Methods* **15**(3): 815–829.

Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate discrete distributions*, CRC Press.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione, *G. Ist. Ital. Attuari* **4**: 83–91.

Kvam, P. H. and Vidakovic, B. (2007). *Nonparametric statistics with applications to science and engineering*, John Wiley & Sons.

Kyriakoussis, A., Li, G. and Papadopoulos, A. (1998). On characterization and goodness-of-fit test of some discrete distribution families, *Journal of Statistical Planning and Inference* **74**(2): 215–228.

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics, *Journal of the American Statistical Association* **73**(362): 253–263.

Ledwina, T. and Wyłupek, G. (2017). On Charlier polynomials in testing Poissonity, *Communications in Statistics-Simulation and Computation* **46**(3): 1918–1932.

Lockhart, R. A. (2012). Conditional limit laws for goodness-of-fit tests, *Bernoulli* **18**(3): 857–882.

Loève, M. (1977). *Probability Theory I*, Springer, New York, NY.

Loukas, S. and Kemp, C. D. (1986). The index of dispersion test for the bivariate Poisson distribution, *Biometrics* **42**(4): 941–948.

Meintanis, S. G. (2007). A new goodness-of-fit test for certain bivariate distributions applicable to traffic accidents, *Statistical Methodology* **4**(1): 22–34.

Meintanis, S. G. (2008). New inference procedures for generalized Poisson distributions, *Journal of Applied Statistics* **35**(7): 751–762.

Meintanis, S. G. and Nikitin, Y. Y. (2008). A class of count models and a new consistent test for the Poisson distribution, *Journal of Statistical Planning and Inference* **138**(12): 3722–3732.

Mijburgh, P. A. and Visagie, I. J. H. (2020). An overview of goodness-of-fit tests for the Poisson distribution, *South African Statistical Journal* **54**(2): 207–230.

Moore, D. S. (1986). Tests of Chi-Squared Type Goodness of Fit Techniques.

Nakamura, M. and Pérez-Abreu, V. (1993). Use of an empirical probability generating function for testing a Poisson model, *Canadian Journal of Statistics* **21**(2): 149–156.

Nass, C. A. G. (1959). The χ^2 test for small expectations in contingency tables, with special reference to accidents and absenteeism, *Biometrika* **46**(3/4): 365–385.

Neyman, J. (1937). Smooth test for goodness of fit, *Scandinavian Actuarial Journal* **1937**(3-4): 149–199.

Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I, *Biometrika* **20**(1/2): 175–240.

Novoa-Muñoz, F. (2021). Goodness-of-fit tests for the bivariate Poisson distribution, *Communications in Statistics-Simulation and Computation* **50**(7): 1998–2014.

Novoa-Muñoz, F. and Jiménez-Gamero, M. D. (2014). Testing for the bivariate Poisson distribution, *Metrika* **77**(6): 771–793.

Novoa-Muñoz, F. and Jiménez-Gamero, M. D. (2016). A goodness-of-fit test for the multivariate Poisson distribution, *SORT (Statistics and Operations Research Transactions)* **40**(1): 113–138.

Ord, J. K. (1967). Graphical methods for a class of discrete distributions, *Journal of the Royal Statistical Society: Series A (General)* **130**(2): 232–238.

Pardo, L. (2006). *Statistical inference based on divergence measures*, Chapman and Hall/CRC.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **50**(302): 157–175.

Pettigrew, H. M. and Mohler, W. C. (1967). A rapid test for the Poisson distribution using the range, *Biometrics* pp. 685–692.

Poisson, S. D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile: précédées des règles générales du calcul des probabilités*, Bachelier.

Potthoff, R. F. and Whittinghill, M. (1966). Testing for homogeneity: II. the Poisson distribution, *Biometrika* **53**(1/2): 183–190.

Rao, C. R. and Chakravarti, I. M. (1956). Some small sample tests of significance for a Poisson distribution, *Biometrics* **12**(3): 264–282.

Rayner, J. C. W. and Best, D. J. (1988). Smooth tests of goodness of fit for regular distributions, *Communications in Statistics-Theory and Methods* **17**(10): 3235–3267.

Rayner, J. C. W. and Best, D. J. (1990). Smooth tests of goodness of fit: An Overview, *International Statistical Review/Revue Internationale de Statistique* **58**(1): 9–17.

Rayner, J. C. W. and Best, D. J. (1997). Crockett's test of fit for the bivariate Poisson, *Biometrical Journal* **39**(4): 423–430.

Rayner, J. C. W. and McIntyre, R. I. (1985). Use of the Score Statistic for Testing Goodness-of-Fit of Some Generalised Distributions, *Biometrical Journal* **27**(2): 159–166.

Rayner, J. C. W., Thas, O. and Best, D. J. (2009). *Smooth tests of goodness of fit: using R*, John Wiley & Sons.

Read, T. R. C. and Cressie, N. A. C. (1988). Testing the Models: Large-Sample Results, *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer, pp. 44–63.

Rueda, R., O'Reilly, F. and Pérez-Abreu, V. (1991). Goodness of fit for the Poisson distribution based on the probability generating function, *Communications in Statistics-Theory and Methods* **20**(10): 3093–3110.

Selby, B. (1965). The index of dispersion as a test statistic, *Biometrika* **52**(3/4): 627–629.

Spinelli, J. J. and Stephens, M. A. (1997). Cramér-von Mises tests of fit for the Poisson distribution, *Canadian Journal of Statistics* **35**(1): 257–268.

Székely, G. J. and Rizzo, M. L. (2004). Mean distance test of Poisson distribution, *Statistics & Probability Letters* **67**(3): 241–247.

Teicher, H. (1954). On the multivariate Poisson distribution, *Scandinavian Actuarial Journal* **1954**(1): 1–9.

Titterington, D. M., Afm, S., Smith, A. F. M., Makov, U. E. et al. (1985). *Statistical analysis of finite mixture distributions*, Vol. 198, John Wiley & Sons Incorporated.

Treutler, B. (1995). *Test for the Poisson distribution (in German)*, PhD thesis, University of Karlsruhe.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Vol. 3, Cambridge University Press.

von Mises, R. E. (1928). *Wahrscheinlichkeit, Statistik und Wahrheit*, Vol. 20, Julius Springer.

Watson, G. S. (1961). Goodness-of-fit tests on a circle, *Biometrika* **48**: 109–114.

Weber, M. D., Leemis, L. M. and Kincaid, R. K. (2006). Minimum Kolmogorov-Smirnov test statistic parameter estimates, *Journal of Statistical Computation and Simulation* **76**(3): 195–206.

Young, L. J. and Young, J. H. (1998). Goodness-of-fit tests, *Statistical Ecology*, Springer, pp. 42–74.

Zelterman, D. and Chen, C.-F. (1988). Homogeneity tests against central-mixture alternatives, *Journal of the American Statistical Association* **83**(401): 179–182.

Ζωγράφος, Κ. (2008). *Πιθανότητες*, Πανεπιστήμιο Ιωαννίνων, Ιωάννινα.

Οικονόμου, Π. Μαλεφάκη, Σ. και . Μπατσίδης. Α. (2022). *Πιθανότητες-Στατιστική*, Ηλεκτρονικά Ακαδημαϊκά Συγγράμματα Κάλλιπος, Υπό Έκδοση.

Παπαιωάννου, Τ. (1993). *Εισαγωγή στις Πιθανότητες και τη Στατιστική, Μέρος Ι: Πιθανότητες*, Ιωάννινα.