ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΠΜΣ: ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΛΕΓΧΟΙ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ ΓΙΑ ΜΟΝΤΕΛΑ ΕΥΠΑΘΕΙΑΣ ΜΕΣΩ ΜΕΤΡΩΝ ΑΠΟΚΛΙΣΗΣ

Βασίλειος Σοφοκλέους

Επιβλέπουσα: Φιλία Βόντα
Καθηγήτρια Ε.Μ.Π.

Αθήνα 2023

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στην επιβλέπουσα της διπλωματικής μου εργασίας καθηγήτρια Φιλία Βόντα για την υποστήριξη, την καθοδήγηση και την εμπιστοσύνη που μου παρείχε κατά τη διάρκεια της εκπόνησης της διπλωματικής μου.

Ευχαριστώ για την ευκαιρία που μου δώσατε να ερευνήσω ένα τόσο σημαντικό θέμα και να αναπτύξω τις ικανότητές μου στην περιοχή αυτή. Η πολύτιμη σας εμπειρία και γνώση με βοήθησαν να αποκτήσω μια βαθύτερη κατανόηση του αντικειμένου και να προχωρήσω σε αναλύσεις και συμπεράσματα που θα με στηρίξουν στην επαγγελματική μου πορεία.

Επίσης, θα ήθελα να ευχαριστήσω θερμά την καθηγήτρια Χρυσηίς Καρώνη και τον καθηγητή Βασίλη Παπανικολάου που δέχτηκαν να είναι μέλη της επιτροπής και επίσης για τον χρόνο και την προσοχή που αφιέρωσαν στην ανάγνωση και αξιολόγηση της διπλωματικής μου.

Τέλος, ευχαριστώ πολύ την οικογένεια και τους φίλους μου για την ψυχολογική υποστήριξη που μου παρείχαν καθώς και την Ανθή Βαμβακά για την ψυχραιμία που την διακατείχε να διορθώσει τυχόν συντακτικά λάθη.

Περίληψη

Η ανάλυση επιβίωσης (survival analysis) ασχολείται με την ανάλυση δεδομένων που αφορούν στο χρόνο που μεσολαβεί μέχρι κάποιο συγκεκριμένο συμβάν. Η ανάλυση επιβίωσης ασχολείται κυρίως με τη μελέτη των χρόνων μέχρι τον θάνατο των ασθενών και για αυτό το λόγο πήρε και το συγκεκριμένο όνομα.

Σε πολλές έρευνες, στις οποίες εξετάζεται ο χρόνος μέχρι την εκδήλωση ενός γεγονότος, παρατηρούνται κάποιες διαφορές μεταξύ των πειραματικών μονάδων, οι οποίες δεν μπορούν να ερμηνευτούν σε ικανοποιητικό βαθμό από τις διαθέσιμες επεξηγηματικές μεταβλητές. Το να αγνοηθούν οι διαφορές αυτές έχει σημαντικές συνέπειες στην επιτυχή ανάλυση του γεγονότος. Επομένως, η ανάπτυξη μεθόδων ενσωμάτωσης της ετερογένειας αποτελεί βασικό στόχο της ανάλυσης επιβίωσης και αξιοπιστίας. Βασικά εργαλεία για την ενσωμάτωση της ετερογένειας στην ανάλυση επιβίωσης αποτελούν τα μοντέλα ευπάθειας.

Η παρούσα μεταπτυχιακή εργασία γράφτηκε στο πλαίσιο της απόκτησης μεταπτυχιακού διπλώματος στην περιοχή των εφαρμοσμένων μαθηματικών και ειδικότερα στην περιοχή των πιθανοτήτων και της στατιστικής. Στην εργασία εξετάζονται έλεγχοι καλής προσαρμογής για μοντέλα ευπάθειας μέσω μέτρων απόκλισης.

Συγκεκριμένα, στο πρώτο κεφάλαιο μελετάται το ημι-παραμετρικό μοντέλο του Cox που ανήκει στα μοντέλα αναλογικών κινδύνων. Το μοντέλο αυτό χρησιμοποιείται για να ερμηνεύσει τη σχέση μεταξύ μιας μεταβλητής που περιγράφει το χρόνο επιβίωσης ενός ατόμου με άλλες συμμεταβλητές, χωρίς να χρειάζεται να οριστεί μια συγκεκριμένη μορφή για την αναφορική συνάρτηση κινδύνου. Στην συνέχεια, στο δεύτερο κεφάλαιο, εισάγεται η έννοια του μοντέλου ευπάθειας, το οποίο αποτελεί μια γενίκευση του μοντέλου του Cox. Με την βοήθεια των μοντέλων ευπάθειας επιτυγχάνεται η μοντελοποίηση των διαφοροποιήσεων που μπορεί να εμφανίζουν οι πληθυσμιακές μονάδες, όσον αφορά τους χρόνους που θα τους συμβεί το

παρατηρούμενο γεγονός. Στο κεφάλαιο αυτό δίνεται ιδιαίτερη έμφαση στα μονομεταβλητά μοντέλα ευπάθειας και ιδίως σε Γάμμα και Inverse Gaussian κατανομές ευπάθειας. Στο τρίτο κεφάλαιο αναφέρονται ορισμένα χρήσιμα μέτρα απόκλισης, τα οποία βοηθούν στην εξέταση του κατά πόσο δυο κατανομές είναι ‘’κοντά’’ μεταξύ τους, όπου αυξάνεται η τιμή του μέτρου τόσο μεγαλύτερες είναι οι διαφοροποιήσεις μεταξύ των κατανομών. Γίνεται ειδική αναφορά στα μέτρα φ-Divergence μέσω των οποίων θα γίνει ο έλεγχος που προτείνουμε και θα ερευνήσουμε στην διπλωματική αυτή. Ο έλεγχος αυτός, εξετάζεται διεξοδικά στο τέταρτο κεφάλαιο με την βοήθεια του στατιστικού πακέτου R studio, διερευνώντας αν ένα σύνολο δεδομένων προέρχονται από συγκεκριμένο μοντέλο ευπάθειας. Η συμπεριφορά του ελέγχου εξετάζεται μέσω του μεγέθους και της ισχύος του για διάφορα μεγέθη δείγματος και δύο κατανομές ευπάθειας.

Abstract

Survival analysis deals with the analysis of data related to the time until a specific event occurs. Survival analysis primarily focuses on studying the times until death of patients, which is why it is named as such.

In many studies examining the time until the occurrence of an event, differences are observed among the experimental units that cannot be adequately explained by the available explanatory variables. Ignoring these differences has significant consequences for the successful analysis of the event. Therefore, developing methods to incorporate heterogeneity is a key goal of survival and reliability analysis. Frailty models are fundamental tools for incorporating heterogeneity in survival analysis.

This thesis was written as part of obtaining a postgraduate degree in the field of applied mathematics, specifically in the area of probability and statistics. The thesis examines goodness-of-fit tests for frailty models using divergence measures.

Specifically, the first chapter investigates the semi-parametric Cox model, which belongs to the class of proportional hazards models. This model is used to interpret the relationship between a variable describing an individual's survival time and other covariates, without the need to specify a particular form for the hazard function. In the second chapter, the concept of frailty models is introduced, which is a generalization of the Cox model. Frailty models are used in modeling the differences that the population units may exhibit regarding the times at which the observed event occurs. This chapter particularly emphasizes univariate frailty models, especially Gamma and Inverse Gaussian frailty distributions.

The third chapter discusses certain divergence measures that assist in examining how "close" two distributions are to each other, where the value of the measure increases with greater differences between the distributions. Special attention is given to φ-Divergence measures, which will be employed and explored in this thesis. The

proposed test is thoroughly examined in the fourth chapter using the statistical software package R Studio, investigating whether a given dataset originates from a specific frailty model. The behavior of the test is examined through its size and power for various sample sizes and two frailty distributions.

Περιεχόμενα

Κεφάλαιο 1
Το μοντέλο του Cox

1.1 Εισαγωγή 10
1.2 Ορισμός 11
1.3 Εκτίμηση των παραμέτρων του μοντέλου Cox 12
1.4 Στρωματοποιημένο μοντέλο του Cox (Stratified Cox model) 13
1.5 Ισόπαλοι χρόνοι διακοπής 15
1.6 Έλεγχος υποθέσεων 16
1.6.1 Έλεγχος του λόγου πιθανοφανειών (Likelihood Ratio Test) 16
1.6.2 Έλεγχος Wald 17
1.6.3 Score test 17
1.7 Έλεγχος υπόθεσης αναλογικών κινδύνων 18
1.7.1 Γραφικός έλεγχος της υπόθεσης αναλογικών κινδύνων 18
1.7.2 Έλεγχος των αναλογικών κινδύνων με τη χρήση εξαρτώμενων από το χρόνο μεταβλητών 19
1.8 Υπόλοιπα του μοντέλου του Cox 21

Κεφάλαιο 2
Μοντέλα ευπάθειας

2.1 Εισαγωγή 23
2.2 Μονομεταβλητά μοντέλα ευπάθειας. 24
2.2.1 Γάμμα μοντέλο ευπάθειας 28
2.2.2 Inverse Gaussian μοντέλο ευπάθειας 32
2.2.3 Positive stable μοντέλο ευπάθειας 33
2.2.4 Power Variance Function («PVF») μοντέλο ευπάθειας 34
2.2.5 Compound Poisson μοντέλο ευπάθειας 35
2.3 Πολυμεταβλητά μοντέλα ευπάθειας 37
2.3.1 Shared frailty model – Από κοινού μοντέλα ευπάθειας. 38
2.3.2 Μοντέλα ευπάθειας συσχετιζόμενων δεδομένων (correlated frailty

models) 40
2.4 Γενικευμένη Συνάρτηση Πιθανοφάνειας μοντέλων ευπάθειας 42

Κεφάλαιο 3
Μέτρα Απόκλισης (Divergence Measures)

3.1 Εισαγωγή 47
3.2 Μέτρα απόκλισης 47
3.2.1 Απόσταση Kolmogorov (Kolmogorov Distance). 48
3.2.2 Απόσταση Lévy (Lévy Distance). 48
3.2.3 Μέτρο Απόκλισης Kullback-Leibler (Kullback-Leibler Divergence Measure) 49
3.2.4 Μέτρο Απόκλισης Jeffreys (Jeffreys Divergence Measure). 49
3.2.5 Μέτρο Απόκλισης Rényi (Rényi Divergence Measure). 49
3.3 $\boldsymbol{\varphi}$-Divergence Measures. 50
3.4 Goodness-of-fit: Simple Null Hypothesis. 52
3.5 Phi-divergences and Goodness-of-fit with Fixed Number of Classes. 54
3.6 Composite null hypothesis - Εκτιμητής Μέγιστης Πιθανοφάνειας (Maximum Likelihood Estimator). 55
3.7 Πίνακας Πληροφορίας του Fisher (Fisher Information Matrix) 57
3.8 Κατανομή αθροίσματος ανεξάρτητων Γάμμα κατανεμόμενων τυχαίων μεταβλητών. 58

Κεφάλαιο 4
Έλεγχος μέσω μέτρων απόκλισης για τα μοντέλα ευπάθειας

4.1 Υποθέσεις και θεωρητικό υπόβαθρο για τις προσομοιώσεις 63
4.1.1 Συνάρτηση αθροιστικού κινδύνου 63
4.1.2 Γάμμα μοντέλο ευπάθειας 63
4.1.3 Inverse Gaussian μοντέλο ευπάθειας 64
4.1.4 Ομαδοποίηση - Διαμερισμός των χρόνων επιβίωσης 65

4.1.5 Ελεγχοσυνάρτηση και μέτρα απόκλισης 68
4.1.6 Ασυμπτωτική κατανομή της ελεγχοσυνάρτησης κάτω από την μηδενική υπόθεση 68
4.1.6.1 Πληροφοριακός αριθμός του Fisher $\mathbf{I}_F(\theta_0)$ για το Γάμμα μοντέλο ευπάθειας 72
4.1.6.2 Πληροφοριακός αριθμός του Fisher $\mathbf{I}_F(\theta_0)$ για το Inverse Gaussian μοντέλο ευπάθειας. 73
4.1.6.3 Πληροφοριακός αριθμός του Fisher $I_F(\theta_0)$ για το Γάμμα μοντέλο ευπάθειας 77
4.1.6.4 Πληροφοριακός αριθμός του Fisher $I_F(\theta_0)$ για το Inverse Gaussian μοντέλο ευπάθειας 78
4.2 Προσομοιώσεις 79
4.2.1 Γάμμα μοντέλο ευπάθειας 80
4.2.2 Inverse Gaussian μοντέλο ευπάθειας 81
4.2.3 Έλεγχος καλής προσαρμογής και για τα δύο μοντέλα 81
4.2.4 Ασυμπτωτική κατανομή της ελεγχοσυνάρτησης 84
4.3 Πίνακες Προσομοιώσεων – Μέγεθος και Ισχύς του ελέγχου 86
4.4 Σχόλια – Συμπεράσματα. 108

Βιβλιογραφία.....110

ΚΕΦΑΛΑΙΟ 1

Μοντέλο του Cox

1.1 Εισαγωγή

Η στατιστική ανάλυση επιβίωσης αποτελεί μια πολύ σημαντική τεχνογνωσία για πολλά επιστημονικά πεδία, όπως η ιατρική, η επιδημιολογία, η βιολογία, η οικονομία, η μηχανική κλπ. Είναι σύνηθες να υποθέτουμε ότι ο πληθυσμός είναι ομοιογενής, δηλαδή ότι κάθε εξαρτώμενη τυχαία μεταβλητή έχει ίσες πιθανότητες να της συμβεί το γεγονός που μας ενδιαφέρει. Ωστόσο, πολλές φορές παρατηρείται μια σημαντική ετερογένεια μεταξύ των πληθυσμιακών μονάδων όσον αφορά τους χρόνους εκδήλωσης του ενδεχομένου. Αυτό μπορεί να οφείλεται στις εγγενείς διαφορές μεταξύ των μονάδων.

Δεδομένων των ως άνω, κρίνεται επιτακτική η χρησιμοποίηση επεξηγηματικών μεταβλητών για την ενσωμάτωση των μεταξύ τους διαφοροποιήσεων.

Στην ανάλυση επιβίωσης μελετώνται οι συνήθεις ανεξάρτητες μεταβλητές όπως η ηλικία, το φύλο, η κληρονομικότητα, το βάρος. Όμως πολλές μεταβλητές που επηρεάζουν την επιβίωση των μονάδων δεν τις γνωρίζουμε ή δεν έχουμε την ικανότητα να τις συμπεριλάβουμε στην στατιστική ανάλυση. Για το λόγο αυτό, κάποιες φορές κρίνεται δύσκολη η υιοθέτηση ενός συγκεκριμένου παραμετρικού μοντέλου.

Στο παρόν κεφάλαιο παρουσιάζεται το μοντέλο του Cox, ένα ημι-παραμετρικό μοντέλο κατά το οποίο δε χρειάζεται να επιλέξουμε ένα βασικό παραμετρικό μοντέλο για τα δεδομένα μας. Το μοντέλο του Cox συναντάται σε ένα ευρύ φάσμα επιστημών, και ιδίως στον τομέα της βιοϊατρικής.

1.2 Ορισμός

Το μοντέλο του Cox είναι ένα μοντέλο αναλογικού κινδύνου, δηλαδή είναι της μορφής $\mathrm{h}\,(\mathrm{t}|\,\mathrm{x}) = g(x) \cdot h_0(t)$. Συγκεκριμένα οι συμμεταβλητές x δρουν στην συνάρτηση κινδύνου μέσω της σχέσης:

$$h\,(t|\,x) = h_0(t) \cdot e^{\beta' x} \tag{1.1}$$

όπου $h_0(t)$ είναι μια ακαθόριστη και μη αρνητική συνάρτηση κινδύνου, η οποία ορίζεται ως βασική συνάρτηση κινδύνου (baseline hazard function) και β είναι ένα διάνυσμα συντελεστών, οι οποίοι εκφράζουν την επίδραση των συμμεταβλητών x.

Οι επεξηγηματικές μεταβλητές $x_i\ ,\ i = 1,2,\ldots,p$ δεν εξαρτώνται από το χρόνο (θεωρούμε ότι έχουν καταγραφεί στην αρχή της μελέτης και παραμένουν αμετάβλητες). Στην περίπτωση που $\beta' = \{0\}^{\mathrm{p}}$, δηλαδή όταν για κάθε συμμεταβλητή $x_i\ , i = 1,2,\ldots,p$ ο συντελεστής συσχέτισης β_i ισούται με 0, τότε οι συμμεταβλητές του μοντέλου δεν επηρεάζουν την εξαρτώμενη μεταβλητή και η συνάρτηση κινδύνου ισούται με τη βασική συνάρτηση κινδύνου.

Γνωρίζουμε ότι η αθροιστική συνάρτηση κινδύνου είναι της μορφής :

$H(t\,|\,x) = \int_0^t \mathrm{h(u|x)\ du}$

Οπότε , $H(t\,|\,x) = \int_0^t h_0(\mathrm{u}) e^{\beta' x}\ \mathrm{du} = H_0(t)\, e^{\beta' x}$

είναι η αθροιστική συνάρτηση κινδύνου του μοντέλου του Cox και

$$S(t\,|\,x) = exp\,(-H(t\,|\,x)) \quad \Leftrightarrow$$

$$S(t\,|\,x) = exp\,\{-H_0(t)\, e^{\beta' x}\} \quad \Leftrightarrow$$

$$S(t\,|\,x) = \{S_0(t)\}^{e^{\beta' x}} \tag{1.2}$$

είναι η συνάρτηση επιβίωσης και $S_0(t)$ η βασική συνάρτηση επιβίωσης του μοντέλου.

Το κύριο χαρακτηριστικό του μοντέλου του Cox είναι ότι οι παραμετρικές μορφές των βασικών συναρτήσεων $h_0(t)$, $S_0(t)$ δεν καθορίζονται και δεν χρειάζεται να

υπολογιστούν για να εξαχθούν συμπεράσματα για την παράμετρο ενδαφέροντος β. Εξ ου και η εναλλακτική ονομασία του μοντέλου του Cox ως ημι-παραμετρικού.

1.3 Εκτίμηση των παραμέτρων του μοντέλου του Cox

Η εκτίμηση των παραμέτρων του Cox πραγματοποιείται με τη μέθοδο της μέγιστης πιθανοφάνειας. Η εν λόγω μέθοδος προσαρμόζεται στις ανάγκες και τις ιδιαιτερότητες του μοντέλου, ήτοι στον μη καθορισμό της βασικής συνάρτησης κινδύνου $h_0(t)$.

Έστω οι χρονικές στιγμές $t_{(1)} < t_{(2)} < t_{(3)} < \ldots < t_{(k)}$. Κατά τη χρονική στιγμή $t_{(i)}$ διακόπτεται η λειτουργία της μονάδας i με συμμεταβλητές $x_{(i)}$. Με $R_{(i)}$ συμβολίζεται ο πληθάριθμος των μονάδων που είναι σε κίνδυνο αμέσως πριν από τη χρονική στιγμή $t_{(i)}$. Επίσης υποθέτουμε ότι οι χρόνοι διακοπής των μονάδων δεν συμπίπτουν, δηλαδή $d_{(j)}= 1 \ , \forall\, j = 1,2,\ldots,k$ με $d_{(j)}$ να είναι ο αριθμός των μονάδων που σταματάει η λειτουργία τους την χρονική στιγμή $t_{(j)}$.

Από την βασική θεωρία πιθανοτήτων γνωρίζουμε ότι η πιθανότητα να διακοπεί η λειτουργία μιας μονάδας j δοθέντος ότι σταματά να λειτουργεί μία οποιαδήποτε μονάδα του συνόλου $R_{(j)}$, είναι:

$$\frac{h(t_{(j)}\,|x_j)}{\sum_{i\in R_{(j)}} h(t_{(j)}\,|x_i)} = \frac{e^{\beta' x_j}}{\sum_{i\in R_{(j)}} e^{\beta' x_i}}$$

Έτσι, ο Cox, το 1972, όρισε την ακόλουθη συνάρτηση μερικής πιθανοφάνειας για την εκτίμηση του β στο μοντέλο αυτό :

$$Lik(\beta) = \prod_{j=1}^{k}\{\frac{e^{\beta' x_j}}{\sum_{i\in R_{(j)}} e^{\beta' x_i}}\} \qquad (1.3)$$

Λογαριθμίζοντας τη συνάρτηση μερικής πιθανοφάνειας βρίσκουμε τη συνάρτηση:

$$l(\beta) = \sum_{j=1}^{k} \beta' x_j - \sum_{j=1}^{k} \ln\{\sum_{i\in R_{(j)}} e^{\beta' x_i}\}$$

όπου οι μερικές παράγωγοι είναι :

$$\frac{\partial l}{\partial \beta_r} = \sum_{j=1}^{k} x_{jr} - \sum_{j=1}^{k} \left[\sum_{i \in R_j} x_{ir} e^{\beta' x_i} \Big/ \sum_{i \in R_j} e^{\beta' x_i} \right]$$

Εξισώνοντας τις συναρτήσεις αυτές με το 0 για κάθε $r = 1,2,\ldots,p$ και λύνοντας ως προς β, βρίσκουμε τις εκτιμήτριες μέγιστης μερικής πιθανοφάνειας $\hat{\beta}$. Οι εκτιμήσεις διασποράς των εκτιμήσεων $\hat{\beta}$ υπολογίζονται από τον αντίστροφο πίνακα της παρατηρούμενης πληροφορίας, με (r, s) στοιχείο να είναι το $-\frac{\partial^2 l}{\partial \beta_r \partial \beta_s} \Big|_{\hat{\beta}}$, όπου:

$$-\frac{\partial^2 l}{\partial \beta_r \partial \beta_s} = \sum_{j=1}^{k} \sum_{i \in R_{(j)}} x_{ir} \left[x_{is} - \frac{\sum_{l \in R_{(j)}} x_{ls} e^{\beta' x_l}}{\sum_{l \in R_{(j)}} e^{\beta' x_l}} \right] \frac{e^{\beta' x_i}}{\sum_{l \in R_{(j)}} e^{\beta' x_l}}$$

1.4 Στρωματοποιημένο μοντέλο του Cox (Stratified Cox model)

Σε ορισμένες αναλύσεις, μια μεταβλητή μπορεί να έχει επίπεδα που απαντώνται σε συναρτήσεις κινδύνου που δεν ικανοποιούν την υπόθεση αναλογικότητας. Σε αυτές τις περιπτώσεις, εφαρμόζουμε την στρωματοποίηση ως προς αυτή τη μεταβλητή. Για παράδειγμα, εισάγοντας την κατηγορική μεταβλητή «φύλο», την χωρίζουμε σε δύο στρώματα, ήτοι άντρες και γυναίκες. Ομοίως π.χ. για την ανάλυση του καρκίνου του εντέρου χωρίζουμε την κατηγορική μεταβλητή σε τρείς κατηγορίες, σε καρκίνο σταδίου I, καρκίνο σταδίου II και καρκίνο σταδίου III, καθώς εντοπίζονται σημαντικές διαφορές πρόγνωσης της επιβίωσης. Για το λόγο αυτό, χρειάζεται να ορίσουμε διαφορετικές βασικές συναρτήσεις κινδύνου στα στάδια του καρκίνου του εντέρου για πιο αξιόπιστη μελέτη της επιβίωσης.

Μία επέκταση του μοντέλου του Cox, η οποία μας βοηθάει να ξεπεράσουμε το πρόβλημα του αναλογικού κινδύνου, είναι η στρωματοποιημένη ανάλυση του Cox. Υποθέτουμε ότι υπάρχει μια μεταβλητή με s επίπεδα. Στην περίπτωση αυτή, η συνάρτηση κινδύνου ενός ατόμου που ανήκει στο i επίπεδο-στρώμα δίνεται από τη σχέση:

$$h_i(t|\,x) = h_{0i}(t)\, e^{\beta' x} \quad , \; i = 1, 2, \ldots, s$$

Η υπόθεση του αναλογικού κινδύνου ισχύει για τις συμμεταβλητές της ίδιας κατηγορίας. Για παράδειγμα, για δύο άτομα που ανήκουν στο ίδιο στρώμα d, $i = 1, \ldots, s$ με μεταβλητές x_1 και x_2 ισχύει:

$$\frac{h_d(t|x_1)}{h_d(t|x_2)} = \frac{h_{0d}(\mathrm{t})\, e^{\beta' x_1}}{h_{0d}(\mathrm{t})\, e^{\beta' x_2}} = e^{\beta'(x_1 - x_2)}$$

Για την εκτίμηση των συντελεστών $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$, χρησιμοποιείται η μέθοδος της μερικής πιθανοφάνειας. Για κάθε στρώμα m, ο λογάριθμος της μερικής πιθανοφάνειας είναι:

$$l_m(\beta) = \sum_{j=1}^{k_m} \beta' x_{mj} - \sum_{j=1}^{k_m} \ln\{ \sum_{i \in R_{mj}} e^{\beta' x_{mi}} \}$$

Παρατηρούμε ότι μόνο οι πληροφορίες του στρώματος m συμμετέχουν στη συνάρτηση l_m. Αθροίζοντας όλα τα στρώματα για $m = 1,2, \ldots, s$ έχουμε:

$$l(\beta) = \sum_{m=1}^{s} l_m(\beta) = \sum_{m=1}^{s} \sum_{j=1}^{k_m} \beta' x_{mj} - \sum_{m=1}^{s} \sum_{j=1}^{k_m} \ln\{ \sum_{i \in R_{mj}} e^{\beta' x_{mi}} \}$$

όπου $l(\beta)$ είναι ο λογάριθμος της συνάρτησης μερικής πιθανοφάνειας. Στη συνέχεια για την εκτίμηση του β προχωράμε με βάση τα οριζόμενα στην παράγραφο 1.3.

1.5 Ισόπαλοι χρόνοι διακοπής

Είναι πιθανόν σε μελέτη ανάλυσης επιβίωσης να συμπίπτουν κάποιοι χρόνοι διακοπής. Σε αυτήν την περίπτωση, πρέπει να διαμορφώσουμε τη συνάρτηση μερικής πιθανοφάνειας, έτσι ώστε να περιλαμβάνονται όλες οι παρατηρήσεις. Εάν οι χρόνοι μέτρησης είναι συνεχείς, εισάγουμε την μεταβλητή d_j, η οποία ορίζεται ως ο αριθμός των μονάδων που διακόπηκαν την χρονική στιγμή t_j (αν η μέτρηση ήταν μεγαλύτερης ή απόλυτης ακρίβειας, η πιθανότητα να συμβεί ισότητα (tie) θα ήταν απειροελάχιστη). Συνηθέστερη παραλλαγή της μερικής πιθανοφάνειας για ισόπαλους χρόνους διακοπής είναι η πιθανοφάνεια του Breslow(1974), ωστόσο χρησιμοποιείται και η πιθανοφάνεια του Efron, καθώς και η διακριτή πιθανοφάνεια για διακριτούς χρόνους. Στη συνέχεια αναλύεται η πιθανοφάνεια του Breslow.

Ορίζουμε $z_j = \sum_{k=1}^{d_j} x_k$, με x_k το διάνυσμα των συμμεταβλητών της k μονάδας, η οποία διακόπτεται την χρονική στιγμή t_j. Το διάνυσμα $s_t = \sum_{m \in d_i} z_m$ είναι το άθροισμα των διανυσμάτων z_m, δηλαδή των ατόμων που αποτυγχάνουν τη χρονική στιγμή t_i. Ο Breslow, χρησιμοποίησε τον όρο $\frac{e^{\beta' z_j}}{\{\sum_{i \in R_{(j)}} e^{\beta' x_i}\}^{d_j}}$ για τη μέθοδο πιθανοφάνειας. Εάν ορίσουμε $d_j = 1$, βλέπουμε ότι εμφανίζεται ο συνήθης όρος μερικής πιθανοφάνειας που αναλύθηκε στο υποκεφάλαιο 1.3 ανωτέρω.

Συνεπώς, η συνάρτηση μερικής πιθανοφάνειας του Breslow είναι:

$$L_{Br}(\beta) = \prod_{j=1}^{k} \{ \frac{e^{\beta' s_j}}{\{\sum_{m \in R_{(j)}} e^{\beta' z_m}\}^{d_j}} \} \qquad (1.4)$$

Η παραλλαγή αυτής της μεθόδου μέγιστης πιθανοφάνειας είναι λιγότερο ακριβής αλλά αρκετά πιο γρήγορη για αυτό και επιλέγεται σε σχέση με της υπόλοιπες μεθόδους για ισόπαλους χρόνους διακοπής.

1.6 Έλεγχος Υποθέσεων

Ο έλεγχος των υποθέσεων $\beta_i = 0$ γίνεται συνήθως με τον έλεγχο του λόγου των πιθανοφανειών, καθώς έχει το πλεονέκτημα σε μικρά δείγματα να αποδίδει πιο αξιόπιστα αποτελέσματα. Όμως χρησιμοποιούνται και άλλοι έλεγχοι, όπως ο έλεγχος WALD ή ο έλεγχος Score test. Στη συνέχεια θα αναφέρουμε μερικούς από αυτούς τους ελέγχους.

1.6.1. Έλεγχος του λόγου πιθανοφανειών (Likelihood Ratio Test)

Έστω ότι σε μία μελέτη κάνουμε την υπόθεση H_0: $\beta_i = 0$, δηλαδή ότι η συμμεταβλητή x_i δεν επηρεάζει την εξαρτημένη μεταβλητή μας. Στον έλεγχο αυτό το μοντέλο προσαρμόζεται με ή χωρίς τη συμμεταβλητή x_i, η αφαίρεση της οποίας ισοδυναμεί με το μοντέλο της υπόθεσής μας. Έστω $\hat{l}_1$ η μεγιστοποιημένη τιμή του λογαρίθμου της μερικής πιθανοφάνειας για $\beta_i \neq 0$, δηλαδή του μοντέλου με τη συμμεταβλητή x_i, και $\hat{l}_0$ για $\beta_i = 0$. Έτσι γίνεται η σύγκριση της τιμής του -2 ($\hat{l}_0 - \hat{l}_1$) με την χ_1^2 κατανομή, κρίνοντας αν η συμμεταβλητή x_i είναι στατιστικά σημαντική για το μοντέλο μας. Σημειώνεται ότι στο μοντέλο αναλογικού κινδύνου του Cox, όταν υπάρχει ένας ικανοποιητικά μεγάλος αριθμός παρατηρήσεων, οι εκτιμήσεις των συντελεστών παλινδρόμησης ακολουθούν προσεγγιστικά κανονική κατανομή.

Παρομοίως πράττουμε και για τον έλεγχο της σημαντικότητας της αφαίρεσης περισσοτέρων από μία συμμεταβλητές. Για τον έλεγχο της μηδενικής υπόθεσης, συγκρίνεται η μεταβολή της τιμής -2 ($\hat{l}_0 - \hat{l}_s$) με την τιμή της χ_s^2 κατανομής, όπου s είναι ο αριθμός των συμμεταβλητών που αφαιρούμε από το μοντέλο.

1.6.2. Έλεγχος Wald

Το Wald test ελέγχει την μηδενική υπόθεση H_0: $\beta_j = 0$ χρησιμοποιώντας την ελεγχοσυνάρτηση :

$$W = (\widehat{\beta_j} - \beta_0)^T I(\widehat{\beta_j})^{-1}(\widehat{\beta_j} - \beta_0)$$

$$= \{\frac{\widehat{\beta_j}}{se(\widehat{\beta_j})}\}^2$$

με την τελευταία ισότητα να ισχύει καθώς η παράμετρος β_j είναι μονοδιάστατη. Η τιμή W συγκρίνεται με την κατανομή χ_1^2, ελέγχοντας έτσι εάν η συμμεταβλητή x_j είναι στατιστικά σημαντική.

1.6.3. Score test

Η στατιστική συνάρτηση του Score test έχει μια ασυμπτωτικά κανονική κατανομή με μέση τιμή 0 και διασπορά $I(\beta)$: $U(\beta) \sim N_p(\,0\,,\ I(\beta)\,)$, όπου $U(\theta) = \frac{\partial LogLik(\theta \mid x)}{\partial \theta}$.

Το score test ελέγχει την μηδενική υπόθεση H_0: $\beta_0 = \beta_1 = \ldots = \beta_p$, χρησιμοποιώντας την ελεγχοσυνάρτηση:

$Q = U^T(\hat{\theta})\, I(\hat{\theta})^{-1} U(\hat{\theta})$

Η τιμή Q συγκρίνεται με τη κατανομή χ_p^2, για τον έλεγχο της σημαντικότητας των p συμμεταβλητών στο μοντέλο.

1.7 Έλεγχος υπόθεσης αναλογικών κινδύνων

Στον κλάδο των βιοϊατρικών επιστημών καθώς και σε άλλους κλάδους, όπου δεν έχουμε προηγούμενη εμπειρία ή δεν έχουν γίνει πολλές στατιστικές αναλύσεις της εξαρτώμενης μεταβλητής, δεν γνωρίζουμε αν η υπόθεση των αναλογικών κινδύνων

ικανοποιείται από τα δεδομένα μας. Έτσι χρειάζεται να προβούμε σε στατιστικούς ελέγχους πριν υιοθετήσουμε το μοντέλο του Cox. Παρακάτω περιγράφονται τρείς χρήσιμοι έλεγχοι αναλογικού κινδύνου.

1.7.1. Γραφικός έλεγχος της υπόθεσης αναλογικών κινδύνων

Στην υπόθεση αναλογικού κινδύνου (Proportional Hazards, PH), για τη συνάρτηση επιβίωσης ενός ατόμου με διάνυσμα συμμεταβλητών $x = (x_1, x_2, \dots, x_p)$ ισχύει ότι:

$$S(t|x) = [S_0(t)]^{e^{\beta' x}}$$

Λογαριθμίζοντας τα δύο μέλη της εξίσωσης, έχουμε :

$$\log\big(S(t|x)\big) = e^{\beta' x}[logS_0(t)]$$

$$\Leftrightarrow \qquad \log[-\log\big(S(t|x)\big)] = \beta' x + log[-logS_0(t)]$$

Έστω x_1, x_2 διανύσματα συμμεταβλητών δυο μονάδων του μοντέλου αναλογικού κινδύνου. Παρατηρούμε ότι η διαφορά των συναρτήσεων $\log[-\log\big(S(t|x_1)\big)]$ και $\log[-\log\big(S(t\,|x_2)\big)]$ είναι σταθερή, καθώς :

$$\log[-\log\big(S(t|x_1)\big)] - \log[-\log\big(S(t\,|x_2)\big)] = \beta'(x_1 - x_2)$$

Σχεδιάζοντας, λοιπόν, τις γραφικές παραστάσεις των συναρτήσεων $\log[-\log\big(S(t|x)\big)]$ για τα δύο διανύσματα x_1 και x_2 στο ίδιο γράφημα πρέπει οι καμπύλες να είναι παράλληλες μεταξύ τους, ήτοι να έχουν περίπου ίση απόσταση σε όλο τους το μήκος. Σε αυτή την περίπτωση, ο πρώτος έλεγχος εκτελείται παρατηρώντας ‘εμπειρικά’ αν οι δυο καμπύλες του γραφήματος είναι ‘σχεδόν’

παράλληλες. Εάν δεν είναι, προβαίνουμε σε μετασχηματισμούς των δεδομένων μας, ελέγχοντας έπειτα πάλι αν ικανοποιείται η συνθήκη αναλογικού κινδύνου, ειδάλλως επιλέγουμε μια διαφορετική μέθοδο ανάλυσης.

1.7.2. Έλεγχος των αναλογικών κινδύνων με τη χρήση εξαρτώμενων από το χρόνο μεταβλητών

Στο μοντέλο αναλογικού κινδύνου, ο λόγος :

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)e^{\beta' x_i}}{h_0(t)e^{\beta' x_j}} = e^{\beta'(x_i - x_j)}$$

μεταξύ των συναρτήσεων κινδύνου δύο μονάδων i και j, είναι ανεξάρτητος του χρόνου.

Έστω ότι θέλουμε να εξετάσουμε αν μία εκ των $x_1, x_2, \dots, x_k$ ικανοποιεί την ιδιότητα του αναλογικού κινδύνου ή είναι εξαρτωμένη του χρόνου. Θέτουμε $x_l(t) = x_l \cdot g(t)$, όπου $g(t)$ η συνάρτηση του χρόνου t. Τότε, η μεταβλητή x_l μετατρέπεται σε μεταβλητή εξαρτώμενη του χρόνου t.

Το μετασχηματισμένο μοντέλο του Cox θα έχει τη μορφή:

$$h(t|x) = h_0(t)\exp(\beta_l x_l + \gamma x_l(t) + \beta^- x^-)$$

όπου με x^- συμβολίζουμε το διάνυσμα των υπολοίπων k -1 μεταβλητών και β^- αντίστοιχα τους συντελεστές τους.

Έτσι τον έλεγχο της μηδενικής υπόθεσης $H_0 : \gamma = 0$ τον πραγματοποιούμε στην παρακάτω εξίσωση:

$$h(t|x) = h_0(t)\exp(\beta_l x_l + \gamma x_l g(t) + \beta^- x^-)$$

Αν η υπόθεση H_0 γίνει δεκτή, τότε η συμμεταβλητή x_l ικανοποιεί την υπόθεση αναλογικών κινδύνων, άλλως συμπεραίνουμε ότι η συμμεταβλητή αυτή είναι εξαρτώμενη του χρόνου t.

Για την εξέταση της υπόθεσης για όλες τις μεταβλητές $x_1, x_2, \dots, x_k$ ταυτοχρόνως, ο έλεγχος της μηδενικής υπόθεσης $H_0 : \gamma_i = 0$, γίνεται στην παρακάτω τροποποιημένη εξίσωση του μοντέλου του Cox:

$$h(t|x) = h_0(t)\exp(\sum_{i=1}^{k}\beta_i x_i + \sum_{i=1}^{k}\gamma_i[x_i g(t)])$$

Όμοια, καταλήγουμε σε αντίστοιχες παρατηρήσεις. Συνήθως για $g(t)$ χρησιμοποιούμε τις συναρτήσεις t και lnt .

1.8 Υπόλοιπα του μοντέλου του Cox

Η γραφική απεικόνιση των υπολοίπων προσφέρει ένα πολύ χρήσιμο έλεγχο για τη διερεύνηση της καταλληλότητας του εκάστοτε εκτιμώμενου μοντέλου. Στο πλαίσιο αυτό, θα αναφέρουμε δυο κατηγορίες υπολοίπων που διερευνώνται στο μοντέλο του Cox. Γνωρίζουμε ότι η σωρευτική συνάρτηση κινδύνου ορίζεται ως:

$$H(t) = -lnS(t) \qquad (1.5)$$

Προσαρμόζοντας τα δεδομένα μας στο μοντέλο του Cox λαμβάνουμε τα υπόλοιπα Cox-Snell:

$$-ln\hat{S}(t_j|x_j) = \hat{H}(t_j|x_j) \qquad \Leftrightarrow$$
$$-ln\hat{S}(t_j|x_j) = \hat{H}_0(t_j)e^{\hat{\beta}'x_j}$$

όπου $\hat{H}_0(t)$ είναι μια μη-παραμετρική εκτιμήτρια και $S(t_j|x_j) = S_0(t_j)^{e^{\beta' x_j}}$

Το διάνυσμα β συνήθως εκτιμάται μεγιστοποιώντας τη συνάρτηση μερικής πιθανοφάνειας $l(\beta)$ (MLE). Η βασική συνάρτηση $H_0(t)$ μπορεί να εκτιμηθεί με τη μέθοδο Breslow.

Εκτιμητής Breslow : $\hat{H}_0(t) = \sum_{t_i \le t} \frac{d_i}{\sum_{j\in R_{(t_i)}} e^{\hat{\beta}' x_j}}$,

όπου $d_{(j)}$ ο αριθμός των μονάδων, των οποίων η λειτουργία σταματάει την χρονική στιγμή $t_{(j)}$.

Ένα άλλο είδος υπολοίπων που χρησιμοποιείται συχνά σε μη παραμετρικά μοντέλα είναι τα υπόλοιπα Schoenfeld (Schoenfeld, 1982).

Τα υπόλοιπα αυτά ορίζονται ως:

$\hat{r}_j = x_j - E(\mathrm{x}|\mathrm{R}_j)$,

με $$E(x|\mathrm{R}_j) = \sum_{k\in R_j} x_k p_k = \frac{\sum_{k\in R_j} x_k e^{\beta\prime x_k}}{\sum_{i\in R_j} e^{\beta\prime x_i}}$$

και p_k είναι η πιθανότητα διακοπής της λειτουργίας της k μονάδας, δοθέντος ότι σταματά να λειτουργεί μία οποιαδήποτε μονάδα του συνόλου $R_{(j)}$:

$$p_k = \frac{e^{\beta' x_k}}{\sum_{i\in R_{(j)}} e^{\beta' x_i}}$$

ΒΙΒΛΙΟΓΡΑΦΙΑ ΚΕΦΑΛΑΙΟΥ 1

Aalen, O. O. (1989). A linear regression model for the analysis of lifetimes. Stat. in Medicine ,8, 907-925

Aalen, O.O. (1988). Heterogeneity in survival analysis. Statistics in Medicine, 7, 1121-37.

Andersen, P. K. and Gill, R. D.(1982). Cox's regression model for counting processes: A large sample study. Annals of Stat., 10:1100-1120

Androulakis, E., Koukouvinos, C. and Vonta, F. (2012). Estimation and variable selection via frailty models with penalized likelihood, Statist. Med. 2012, 31, 2223–2239

Bennett, S. (1983). Analysis of survival data by the proportional odds model. Stat Med, 2: 273–277

Cox, D.R. (1972). Regression models with life tables (with discussion). J Roy Stat Soc B 34: 187–220

Fan, J., Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. Ann Stat 30(1):74–99

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. Ann Stat 30(1): 74–99

Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on EM algorithm. Biometrics, 48, 795-806.

Vonta, F & Karagrigoriou, A. (2007). Variable selection strategies in survival models with multiple imputations, Lifetime Data Analysis, 13, 295–315

ΚΕΦΑΛΑΙΟ 2

Μοντέλα Ευπάθειας

2.1 Εισαγωγή

Τα μοντέλα ευπάθειας (frailty models) ξεκίνησαν να χρησιμοποιούνται με σκοπό τη μοντελοποίηση των διαφοροποιήσεων που παρατηρούνταν στους χρόνους επιβίωσης των μονάδων ενός δείγματος, οι οποίες δε μπορούσαν να επεξηγηθούν με βάση τις μεταβλητές που έχουν συμπεριληφθεί στη μελέτη. Στο πλαίσιο αυτό, προέκυψε η ανάγκη να εισαχθεί ο όρος της «ετερογένειας» του πληθυσμού, προκειμένου να διαφοροποιηθούν οι «ευπαθείς» μονάδες, οι οποίες δεν διατρέχουν τον ίδιο κίνδυνο να τους συμβεί το παρατηρούμενο γεγονός σε σύγκριση με τις υπόλοιπες μονάδες.

Η έννοια αυτή της ευπάθειας εισήχθηκε στα στατιστικά μοντέλα μέσω των συναρτήσεων κινδύνου των ετερογενών μονάδων. Δεδομένου ότι δε γνωρίζουμε τις «κρυφές» μεταβλητές που οδηγούν σε αυτή τη διαφοροποίηση, η τροποποίηση της συνάρτησης κινδύνου γίνεται με την εισαγωγή μιας τυχαίας επίδρασης, μέσω μιας συνάρτησης G, η οποία δρα πολλαπλασιαστικά στη βασική συνάρτηση κινδύνου h_0.

Ο Clayton ήταν ο πρώτος που χρησιμοποίησε, το 1978, τον όρο «ετερογένεια» στην ανάλυση δεδομένων διάρκειας ζωής για να συμπεριλάβει στην ανάλυσή του την κοινή πληροφορία που μοιράζονται πατέρας και γιός ως προς την εμφάνιση χρόνιων παθήσεων. Η ‘’ευπάθεια’’ όμως πρωτοεμφανίστηκε ως ‘’frailty’’, το 1979, από τους Vaupel et al. για την εξήγηση της παρατηρούμενης ετερογένειας που εμφανιζόταν σε μελέτες διάρκειας ζωής για δημογραφικούς σκοπούς. Σύμφωνα με τους Keyfitz και Littman (1979), όταν αγνοείται η ετερογένεια σε μια στατιστική έρευνα, υπερεκτιμάται η διάρκεια ζωής των ατόμων με ευπάθεια, οδηγώντας σε λάθος εκτιμήσεις των παραμέτρων του μοντέλου.

Τα μοντέλα ευπάθειας αποτελούν στην ουσία μία γενίκευση του μοντέλου του Cox, το οποίο αναλύθηκε στο προηγούμενο κεφάλαιο και είναι ένα μοντέλο που έχει την ιδιότητα των αναλογικών κινδύνων για τις μονάδες του πληθυσμού. Τα μοντέλα ευπάθειας χωρίζονται σε δύο κατηγορίες, και πιο συγκεκριμένα α) στην κατηγορία των μοντέλων ευπάθειας όπου η κάθε μονάδα έχει τη δική της ευπάθεια (univariate

frailty model) και δε συναντάται καμία ομοιογένεια μεταξύ των μονάδων, και β) στην κατηγορία των μοντέλων ευπάθειας όπου κάποιες μονάδες μοιράζονται την ίδια ευπάθεια (multivariate frailty model), όπου υπάρχει αμοιβαία εξάρτηση μεταξύ «συγγενικών» μονάδων. Όταν η ευπάθεια που εισάγεται σε ένα μοντέλο εκφράζει διαφοροποιήσεις στους χρόνους επιβίωσης μεταξύ μονάδων, το μοντέλο αυτό αναφέρεται ως «μοντέλο ατομικής ευπάθειας» ή ως «μοντέλο μη κοινής ευπάθειας» (non-shared frailty models), ενώ, όταν οι διαφοροποιήσεις αυτές εμφανίζονται μεταξύ γκρουπ μονάδων, αναφέρεται ως «μοντέλο κοινής ευπάθειας»' (shared frailty models).

2.2 Μονομεταβλητά μοντέλα ευπάθειας

Στα μοντέλα ευπάθειας, οι ανεξάρτητες μεταβλητές χωρίζονται σε δύο κατηγορίες. Στη μία κατηγορία, ανήκουν όλες οι γνωστές μεταβλητές, τις οποίες εισάγουμε στο μοντέλο για να ερευνήσουμε την επίδραση στο παρατηρούμενο γεγονός, ενώ στη δεύτερη κατηγορία εντάσσονται οι άγνωστες μεταβλητές, που δε μπορούν πρακτικά να συμπεριληφθούν στο μοντέλο μας. Οι άγνωστες μεταβλητές εισάγονται στην κατηγορία των ''άγνωστων'' είτε λόγο απουσίας πληροφοριών, είτε λόγω του μεγάλου αριθμού συμμεταβλητών είτε, σε πολλές περιπτώσεις, λόγω του υψηλού οικονομικού κόστους για τη διερεύνησή τους.

Σε ένα μοντέλο αναλογικών κινδύνων, η απουσία κάποιου υποσυνόλου μεταβλητών από το σύστημα οδηγεί σε μεροληπτικές εκτιμήσεις των παραμέτρων παλινδρόμησης και του κινδύνου. Συνήθως, όταν δε λαμβάνουμε υπ' όψιν την ευπάθεια σε ένα μοντέλο, παρατηρούμε ότι η συνάρτηση κινδύνου αυξάνεται όσο «περνά» ο χρόνος , στη συνέχεια φτάνει ένα μέγιστο και τέλος μειώνεται.

Το μονομετάβλητο μοντέλο ευπάθειας είναι ένα εκτεταμένο μοντέλο του Cox, με την προσθήκη μιας επιπρόσθετης τυχαίας μεταβλητής z, η οποία δρα πολλαπλασιαστικά στη συνάρτηση κινδύνου. Συγκεκριμένα, η συνάρτηση κινδύνου παίρνει την μορφή:

$$h(t|z,x) = zh_0(t)e^{\beta' x} \qquad (2.1)$$

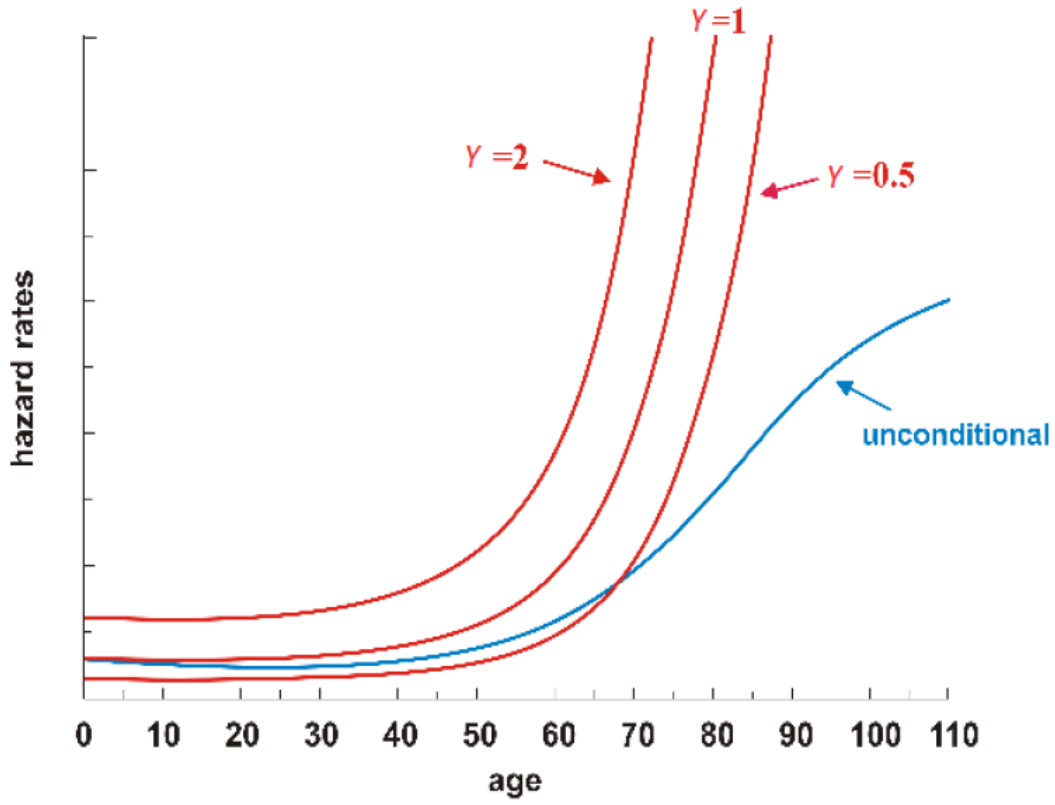
όπου $h_0(t)$, η βασική-αναφορική συνάρτηση κινδύνου, x, το διάνυσμα των συμμεταβλητών και z, η μεταβλητή ευπάθειας. Η μεταβλητή ευπάθειας z μεταβάλλεται ανάλογα με τον πληθυσμό. "Όταν $z < 1$, υπάρχει μειωμένος ατομικός κίνδυνος των μονάδων του πληθυσμού να τους συμβεί το παρατηρήσιμο γενονός , όταν $z > 1$, ο κίνδυνος είναι αυξημένος και όταν $z = 1$, ο πληθυσμός δε θεωρείται ευπαθής. Αυτό παρατηρείται εύκολα στο κάτωθι σχήμα 2.2.1, στο οποίο παρατηρείται ότι, όσο μεγαλώνει η ευπάθεια ενός πληθυσμού, τόσο αυξάνεται ο κίνδυνος θνησιμότητας των μελών-μονάδων του. Επίσης, στο εν λόγω σχήμα διακρίνουμε ότι, αν δε συμπεριλάβουμε όρους ευπάθειας στον πληθυσμό, η συνάρτηση κινδύνου του μοντέλου έχει διαφορετική κατανομή από τις συναρτήσεις κινδύνου των υπό συνθήκη ευπαθών πληθυσμών.

Η συνάρτηση επιβίωσης του συνόλου του πληθυσμού που έχει επιβιώσει έως τη χρονική στιγμή t, ισούται με:

$$S(t|z,x) = \exp(-ze^{\beta' x}H_0(t)) \qquad (2.2)$$

όπου $H_0(\mathrm{t})$ η αθροιστική αναφορική συνάρτηση κινδύνου και $S(t|z,x)$ η πιθανότητα ένα άτομο να έχει επιζήσει έως και τη χρονική στιγμή t.

Conditional and unconditional hazard rates



Σχήμα 2.2.1 : Δεσμευμένες και μη συναρτήσεις κινδύνου σε προσομοιωμένα δεδομένα για την ανθρώπινη θνησιμότητα. Οι κόκκινες γραμμές απεικονίζουν τις υπό συνθήκη συναρτήσεις κινδύνου με αντίστοιχες ευπάθειες 0.5 , 1 και 2, ενώ η μπλε γραμμή απεικονίζει τη συνάρτηση κινδύνου χωρίς ευπάθεια.

Για να εξετάσουμε το μοντέλο στο σύνολο του πληθυσμού, παίρνουμε τη συνάρτηση επιβίωσης του πληθυσμού να ισούται με το μέσο όρο των ατομικών συναρτήσεων επιβίωσης που είδαμε προηγουμένως. Συγκεκριμένα, η μη δεσμευμένη ως προς την ευπάθεια συνάρτηση επιβίωσης στο χρόνο t, λαμβάνεται από τη $S(t|z,x)$, ολοκληρώνοντάς τη ως προς την ευπάθεια:

$$S(t|x) = \int_0^\infty e^{-ze^{\beta^T x} H_0(t)} dF_z(z)$$

όπου F_z η συνάρτηση κατανομής της ευπάθειας.

Ορίζοντας τη συνάρτηση

$$G(w) = -\ln(\int_0^\infty e^{-zw} dF_z(z)) \qquad (2.3)$$

η συνάρτηση επιβίωσης παίρνει την μορφή:

$$S(t|x) = e^{-G(H_0(t)\cdot e^{\beta^T x})} \qquad (2.4)$$

Αν ορίσουμε διαφορετική κατανομή ευπάθειας, παράγεται διαφορετική συνάρτηση G. Υποθέτοντας μια κατανομή για την ευπάθεια και κατά συνέπεια, μια συγκεκριμένη συνάρτηση G, μπορούμε να βρούμε τις εκτιμήσεις των παραμέτρων, ως επί το πλείστων, με τη μέθοδο της μεγιστοποίησης της συνάρτησης πιθανοφάνειας. Σε περίπτωση που κάτι τέτοιο δεν καθίσταται εφικτό, χρησιμοποιούνται άλλες μέθοδοι, με τις πιο διαδεδομένες εξ αυτών η μέθοδος Monte Carlo και η αριθμητική ολοκλήρωση.

Μείζον ζήτημα στην ανάλυση δεδομένων μέσω μοντέλων ευπάθειας συνιστά η επιλογή της κατάλληλης κατανομής ευπάθειας. Συνήθως, λόγω ευχρηστίας, χρησιμοποιείται η Γάμμα κατανομή (Clayton 1978, Vaupel et al. 1979) και η αντίστροφη Gaussian κατανομή. Υπάρχουν, ωστόσο, και αρκετές εφαρμογές με την compound poison κατανομή (Aalen 1988, 1992), με την PVF κατανομή, καθώς και με τη λογαριθμοκανονική κατανομή (McGilchrist and Aisbett, 1991). Οι κατανομές αυτές θα αναλυθούν στη συνέχεια.

Το 1984, ο Hougaard χρησιμοποίησε τους μετασχηματισμούς Laplace για την εύρεση της κατανομής της μεταβλητής ευπάθειας Z. Έκτοτε, οι μετασχηματισμοί Laplace χρησιμοποιούνται συχνά όταν είναι εφικτό, καθώς επιτρέπουν στη συνέχεια τη χρησιμοποίηση της μεθόδου της μέγιστης πιθανοφάνειας για την εκτίμηση των παραμέτρων του μοντέλου. Παρακάτω παρατίθεται ο μετασχηματισμός Laplace συναρτήσει της συνάρτησης επιβίωσης και της συνάρτησης κινδύνου σε μοντέλα ευπάθειας:

$$S(t) = ES(t|Z) = E\left(e^{-ZH_0(t)}\right) = L\left(H_0(t)\right) \qquad (2.4.1)$$

$$f(t) = -h_0(t)L'\big(H_0(t)\big) \quad (2.4.2)$$

$$h(t) = \frac{f(t)}{S(t)} = -h_0(t)\frac{L'\big(H_0(t)\big)}{L\big(H_0(t)\big)} \quad (2.4.3)$$

$$EZ = -L'(0) \quad (2.4.4)$$

$$V(Z) = L''(0) - (L'(0))^2 \quad (2.4.5)$$

2.2.1 Γάμμα μοντέλο ευπάθειας

Η κατανομή Γάμμα είναι μία ευέλικτη κατανομή, η οποία παρουσιάζει τις ιδιότητες οικογένειας κατανομών και ειδικότερα μιας γενικευμένης οικογένειας της εκθετικής κατανομής.

Η ευπάθεια δεν μπορεί να λάβει αρνητικές τιμές, για αυτό και η Γάμμα κατανομή είναι πολύ χρήσιμη σε μοντέλα ευπάθειας, καθώς χρησιμοποιείται για μη αρνητικές τυχαίες μεταβλητές. Η ευρεία χρήση της εν λόγω κατανομής οφείλεται σε δύο ακόμη λόγους. Καταρχάς, η κατανομή ευπάθειας ενός ατόμου του πληθυσμού για δεδομένη χρονική στιγμή t παραμένει Γάμμα κατανομή, με ίδια παράμετρο σχήματος (shape parameter) (αλλά άλλη παράμετρο κλίμακας) με την κατανομή ευπάθειας Γάμμα του ατόμου τη χρονική στιγμή t_0 (στιγμή έναρξης της μελέτης του ατόμου). Επίσης, η κατανομή της ευπάθειας των ατόμων που πεθαίνουν σε δεδομένο χρόνο t, σε σχέση με την κατανομή της ευπάθειας εκείνων που έχουν επιβιώσει μέχρι το χρόνο αυτό, δηλαδή δεδομένου ότι $T > t$, είναι Γάμμα με ίδια παράμετρο κλίμακας και παράμετρο σχήματος αυξημένη κατά μία μονάδα.

Έστω τυχαία μεταβλητή $T > 0$ που ακολουθεί κατανομή Γάμμα, $T \sim Γάμμα(\alpha, \lambda)$, με παράμετρο σχήματος α και παράμετρο rate λ. Η συνάρτηση πυκνότητας πιθανότητάς της T δίνεται από τον τύπο :

$$f_T(t) = \frac{1}{\Gamma(\alpha)}\lambda^{\alpha}t^{\alpha-1}e^{-\lambda t}$$

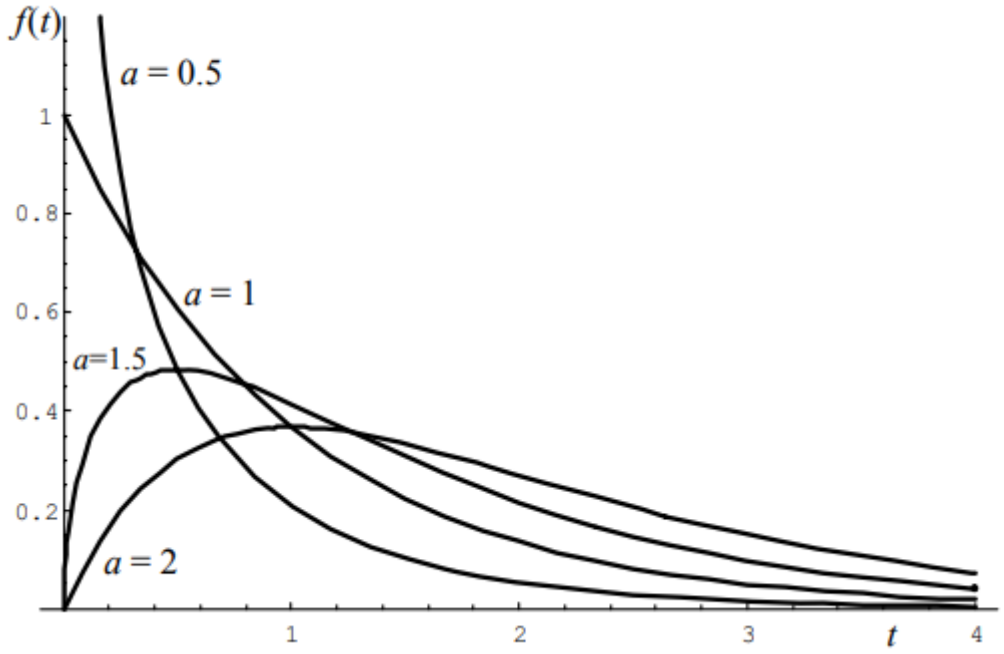
όπου οι παράμετροι $\alpha > 0$ και $\lambda > 0$ παράμετροι, $t \geq 0$, και η συνάρτηση Γάμμα ορίζεται ως: $\Gamma(\kappa) = \int_0^{\infty} x^{k-1}\, e^{-x} dx$

με τις ιδιότητες : $\Gamma(\kappa+1)=\kappa\Gamma(\kappa)$, $\Gamma(n)=(n-1)!$, n ∈ N , $\Gamma(0{,}5)=\sqrt{\pi}$

Η μέση τιμή και η τυπική απόκλιση της κατανομής Γάμμα είναι:

$$E(T)=\frac{\alpha}{\lambda} \quad , \quad Var(T)=\frac{\alpha}{\lambda^2}$$

Παρατηρούμε ότι, καθώς η παράμετρος α μεταβάλλεται, η κατανομή Γάμμα παίρνει διάφορες μορφές κατανομών και γραφημάτων. Παραδείγματος χάρη, όταν $\alpha=1$, η κατανομή Γάμμα μετατρέπεται σε εκθετική κατανομή, ενώ όταν το α παίρνει πολύ ¨μεγάλες¨ τιμές το σχήμα της προσεγγίζει την καμπανοειδή μορφή της κανονικής κατανομής. Αυτό αποτυπώνεται γραφικά στο Σχήμα 2.2.2, όπου οι γραφικές παραστάσεις των κατανομών της Γάμμα λαμβάνουν διαφορετικά σχήματα, όταν η τιμή της παραμέτρου α αλλάζει.



Σχήμα 2.2.2. : Κοινό γράφημα τεσσάρων συναρτήσεων πυκνότητας πιθανότητας της $\Gamma\acute{\alpha}\mu\mu\alpha(\alpha,\lambda)$, με παράμετρο $\alpha=0.5$, $\alpha=1$, $\alpha=1.5$, $\alpha=2$ αντίστοιχα.

Η συνάρτηση κινδύνου για το μοντέλο ευπάθειας χωρίς γνωστές συμμεταβλητές, δίνεται από τον τύπο:

$$h(t|z) = zh_0(t)$$

όπου z η τ.μ. ευπάθεια και $h_0(t)$ η βασική συνάρτηση κινδύνου. Ανάλογα η συνάρτηση επιβίωσης δίνεται από:

$$S(t|z) = e^{-zH_0(t)}$$

όπου $H_0(t)$ η βασική αθροιστική συνάρτηση κινδύνου. Εισάγοντας τον όρο $e^{\beta' X}$ στο μοντέλο όπου X το διάνυσμα των γνωστών συμμεταβλητών και β η παράμετρος παλινδρόμησης ή ενδιαφέροντος παίρνουμε την πιο γενικευμένη συνάρτηση κινδύνου που δίνεται από τον τύπο:

$$h(t|z, X) = zh_0(t)e^{\beta' X}$$

Η βασική μας ιδέα για να χειριστούμε ένα τέτοιο μοντέλο και αφού η ευπάθεια z είναι άγνωστη είναι να ορίσουμε την μη δεσμευμένη ως προς την ευπάθεια συνάρτηση επιβίωσης του πληθυσμού (χωρίς συμμεταβλητές) ως τον σταθμικό μέσο των δεσμευμένων συναρτήσεων επιβίωσης και από την εξίσωση (2.4.1) παίρνουμε:

$$S(t) = E\big(S(t|z)\big) = E\big(e^{-zH_0(t)}\big) = L(H_0(t))$$

όπου L ο μετασχηματισμός Laplace.

Συγκεκριμένα, για την κατανομή Γάμμα:

$$\begin{aligned} L_\Gamma(u) &= \frac{1}{\Gamma(\alpha)}\lambda^\alpha \int e^{-ux} x^{\alpha-1} e^{-\lambda x}\, dx \\ &= \frac{\lambda^\alpha}{(\lambda+u)^\alpha}(\lambda+u)^\alpha \frac{1}{\Gamma(\alpha)} \int x^{\alpha-1}\, e^{-(\lambda+u)x}\, dx \\ &= (1+\frac{u}{\lambda})^{-\alpha} \end{aligned}$$

Οπότε , η συνάρτηση επιβίωσης ισούται με:

$$S(t) = L_\Gamma\big(H_0(t)\big) = (1+\frac{H_0(t)}{\lambda})^{-\alpha} \qquad (2.5)$$

με συνάρτηση πυκνότητας-πιθανότητας να λαμβάνετε από την εξίσωση (2.4.2):

$$f(t) \quad = -h_o(t)L'\big(H_0(t)\big) = -h_o(t)(-\alpha\frac{1}{\lambda})\frac{(1+\frac{H_0(t)}{\lambda})^{-\alpha}}{(1+\frac{H_0(t)}{\lambda})}$$

$$= -h_o(t)(-\alpha\frac{1}{\lambda})\lambda\frac{(1+\frac{H_0(t)}{\lambda})^{-\alpha}}{\lambda+H_0(t)} = h_o(t)\alpha\frac{(1+\frac{H_0(t)}{\lambda})^{-\alpha}}{\lambda+H_0(t)}$$

και συνάρτηση κινδύνου μέσω του (2.4.3) του μετασχηματισμού Laplace:

$$h(t) = \frac{f(t)}{S(t)} = \frac{h_o(t)\alpha\frac{(1+\frac{H_0(t)}{\lambda})^{-\alpha}}{\lambda+H_0(t)}}{(1+\frac{H_0(t)}{\lambda})^{-\alpha}} = \alpha\frac{h_0(t)}{\lambda+H_0(t)}$$

Για λόγους αναγνωρισιμότητας των παραμέτρων στα μοντέλα ευπάθειας γενικά απαιτείται η μέση τιμή της ευπάθειας να είναι 1, άρα εδώ θα πρέπει να υποθέσουμε ότι $\alpha = \lambda$.

ΠΑΡΑΔΕΙΓΜΑ 2.2.1.

Έστω μοντέλο ευπάθειας με την αναφορική συνάρτηση κινδύνου $h_0(t) = \mu e^{bt}$ να ακολουθεί την κατανομή $Gompertz$ και την ευπάθεια Z να ακολουθεί κατανομή Γάμμα, $Z \sim \Gamma(\frac{1}{\sigma^2}, \frac{1}{\sigma^2})$ με συντελεστή σχήματος $\alpha = \frac{1}{\sigma^2}$ και κλίμακας λ να ισούνται με $\frac{1}{\sigma^2}$. Οι συναρτήσεις επιβίωσης και κινδύνου, με τη βοήθεια των μετασχηματισμών Laplace και της εξίσωσης (2.5), δίνονται παρακάτω:

$$H_0(t) \; = \int_0^t \mu e^{bx}dx = [\frac{\mu}{b}e^{bx}]_0^t = \frac{\mu}{b}e^{bt} - \frac{\mu}{b}$$

$$S(t) \quad = \left(1+\frac{H_0(t)}{\lambda}\right)^{-\alpha} = \big(1+\sigma^2 H_0(t)\big)^{-\frac{1}{\sigma^2}}$$

$$= (1+\sigma^2\frac{\mu}{b}(e^{bt}-1))^{-\frac{1}{\sigma^2}}$$

$$h(t) = \alpha \frac{h_0(t)}{\lambda + H_0(t)} = \frac{1}{\sigma^2} \frac{\mu e^{bt}}{\frac{1}{\sigma^2} + \frac{\mu}{b}(e^{bt} - 1)}$$

$$= \frac{\mu e^{bt}}{1 + \sigma^2 \frac{\mu}{b}(e^{bt} - 1)}$$

Πολλοί στατιστικοί ξεκίνησαν να χρησιμοποιούν την κατανομή Γάμμα για να μελετήσουν την ετερογένεια των πληθυσμών από τα τέλη της δεκαετίας του εβδομήντα. Ο Lancaster (1979), βασιζόμενος σε αυτό το μοντέλο, ερεύνησε τη διάρκεια της ανεργίας, ενώ ο Vaupel, το ίδιο έτος, χρησιμοποίησε την κατανομή Γάμμα για τη μελέτη της θνησιμότητας του πληθυσμού. Ένα ακόμη παράδειγμα στην επιστήμη της βιοστατιστικής είναι αυτό του Andersen (1993), ο οποίος μελέτησε τον κίνδυνο που διέτρεχαν τα άτομα με κακοήθες μελάνωμα, ενώ, επίσης, οι Ellerman κ.ά. (1992) εφάρμοσαν την κατανομή Γάμμα-Weibull σε μοντέλο ευπάθειας για να ερευνήσουν την υποτροπή των εγκληματιών σε παράνομες πράξεις.

2.2.2 Inverse Gaussian μοντέλο ευπάθειας

Η αντίστροφη κατανομή *Gaussian* εισήχθη στην επιστήμη της βιοστατιστικής από τον Hougaard (1984) ως μια εναλλακτική της Γάμμα κατανομής για την κατανομή της ευπάθειας.

Η συνάρτηση πυκνότητας πιθανότητας μιας τυχαίας μεταβλητής με αντίστροφη *Γκαουσιανή* κατανομή δίνεται από τον τύπο:

$$f(z) = \frac{\sqrt{\lambda}}{\sqrt{2\pi z^3}} \exp\left(-\frac{\lambda}{2\mu^2 z}(z - \mu)^2\right)$$

όπου $\mu \geq 0$ μέση τιμή και $\lambda > 0$ παράμετρος σχήματος.

Για την αντίστροφη κατανομή Gaussian, ο μετασχηματισμός Laplace δίνεται από τον τύπο:

$$L(u) = E(e^{-uZ})$$
$$= \int \frac{\sqrt{\lambda}}{\sqrt{2\pi z^3}} e^{-uZ} \exp(-\frac{\lambda}{2z\mu^2}(z-\mu)^2) dz$$
$$= \int \frac{\sqrt{\lambda}}{\sqrt{2\pi z^3}} \exp(-\frac{(\lambda+2\mu^2 u)z^2 - 2\mu\lambda z + \lambda\mu^2}{2\mu^2 z}) dz$$

Απλοποιώντας το μετασχηματισμό Laplace η εξίσωση παίρνει τη μορφή:

$$L_{IG}(u) = \exp[\frac{\lambda}{\mu}\left(1 - \sqrt{1 + \frac{2\mu^2 u}{\lambda}}\right)]$$

Οι συναρτήσεις επιβίωσης και κινδύνου μιας τυχαίας συνάρτησης, η οποία ακολουθεί αντίστροφη Γκαουσιανή κατανομή με μέση τιμή $\mu = 1$ και τυπική απόκλιση $\sigma^2 := \frac{1}{\lambda}$, δίνεται από τους τύπους:

$$S(t) = L(H_0(t)) = \exp[\frac{1}{\sigma^2}\left(1 - \sqrt{1 + 2\sigma^2 H_0(t)}\right)]$$

$$h(t) = \frac{h_0(t)}{(1 + 2\sigma^2 H_0(t))^{\frac{1}{2}}}$$

2.2.3 Positive stable μοντέλο ευπάθειας

Μία κατανομή χαρακτηρίζεται ως *positive stable* κατανομή, εάν ο γραμμικός συνδυασμός n ανεξάρτητων μεταβλητών από αυτήν την κατανομή, έχει την ίδια κατανομή με μια γραμμική συνάρτηση μιας μόνο τυχαίας μεταβλητής X από αυτή την κατανομή. Συγκεκριμένα, αν η κατανομή της X είναι positive stable, τότε για κάθε θετικό αριθμό a_i, με $i = 1,2,\dots,n$, υπάρχει θετικός αριθμός b και πραγματικός αριθμός c, ώστε:

$$a_1 X_1 + a_2 X_2 + \cdots + a_n X_n \sim bX + c$$

με X_i να είναι τυχαίες μεταβλητές ανεξάρτητες της X και ο συμβολισμός '$\sim$' να δηλώνει ότι το αριστερό και το δεξί μέλος του συμβολισμού αυτού είναι τ.μ. που ακολουθούν την ίδια κατανομή. Δεδομένου ότι δεν υπάρχει κλειστή μορφή για τον τύπο της συνάρτησης πυκνότητας πιθανότητας ο μετασχηματισμός Laplace, ο οποίος ορίζεται απλά, μας βοηθάει να ορίσουμε τις συναρτήσεις επιβίωσης και κινδύνου. Η συνάρτηση του μετασχηματισμού Laplace για την $positive\ stable$ κατανομή δίνεται από τον τύπο:

$$L(u) = e^{-\frac{ku^{\gamma}}{\gamma}} \qquad , \qquad \mu\varepsilon\ \gamma \in (0,1]$$

Για λόγους ευκολίας, συνήθως περιορίζουμε την κατανομή ευπάθειας με δυο παραμέτρους στην απλή περίπτωση, όπου $k = \gamma$.

Ο μετασχηματισμός Laplace, λοιπόν, παίρνει τη μορφή: $L(u) = e^{-u^{\gamma}}$

Εκ των ανωτέρω προκύπτουν οι συναρτήσεις:

$$S(t) = L\big(H_0(t)\big) = e^{-H_0(t)^{\gamma}}$$

$$f(t) = \gamma h_0(t) H_0(t)^{\gamma-1} e^{-H_0(t)^{\gamma}}$$

$$h(t) = \gamma h_0(t) H_0(t)^{\gamma-1}$$

2.2.4 Power Variance Function («PVF») μοντέλο ευπάθειας

Η οικογένεια κατανομών $power\ variance\ function (PVF)$ εισήχθη από τον Tweedy(1984) και στη συνέχεια μελετήθηκε από τον Hougaard (1986a). Πρόκειται για μια οικογένεια κατανομών μοντέλων ευπάθειας με τρείς παραμέτρους, η οποία ορίζεται ως $PVF(\gamma, k, \lambda)$ και, ανάλογα με τις χρησιμοποιούμενες παραμέτρους, μπορεί να πάρει μορφές διαφόρων κατανομών ευπάθειας, όπως της Γάμμα, της αντίστροφης Gaussian και της Positive stable κατανομής. Ο μετασχηματισμός Laplace για την κατανομή PVF έχει τον τύπο

$$L(u) = e^{-\frac{k}{\gamma}\left((\lambda+u)^{\gamma}-\lambda^{\gamma}\right)}$$

Για μεταβλητή ευπάθειας Z, η οποία ακολουθεί την κατανομή PVF, η μέση τιμή και η διακύμανση ορίζεται ως εξής:

$$E(Z) = k\lambda^{\gamma-1} \quad και \quad V(Z) = k(1-\gamma)\lambda^{\gamma-2}$$

με συνάρτηση επιβίωσης: $S(t) = e^{-\frac{k}{\gamma}\left(\left(\lambda+H_0(t)\right)^{\gamma}-\lambda^{\gamma}\right)}$

και συνάρτηση κινδύνου: $h(t) = kh_0(t)(\lambda + H_0(t))^{\gamma-1}$

Όταν το μοντέλο ορίζεται με τους περιορισμούς $0 \leq \gamma \leq 1, k > 0$ και $\gamma > 0$, τότε αρκεί $\lambda \geq 0$, ενώ για $\gamma \geq 0$, δεχόμαστε $\lambda > 0$. Σε μελέτη επιβίωσης ευπαθών πληθυσμών τη χρονική στιγμή t, η κατανομή μας ορίζεται ως $PVF(\gamma, k, \lambda + H_0(t))$. Παρατηρούμε τη δυνατότητα του μοντέλου να προσαρμόζεται σε διάφορες κατανομές ευπάθειας, αφού για $\gamma = 0$, η κατανομή είναι όμοια με την κατανομή $Γάμμα(\alpha = k, \lambda)$, για $\gamma = 0.5$, παίρνει τη μορφή αντίστροφης κατανομής Gaussian, ενώ για $\lambda = 0$ και $\gamma \neq 0$, προσαρμόζεται σε positive stable κατανομή.

Περιορίζοντας την κατανομή ευπάθειας με $E(Z) = 1$ $και$ $V(Z) = \sigma^2 = \frac{1-\gamma}{\lambda}$ για τους λόγους που έχουμε προαναφέρει η συνάρτηση κινδύνου παίρνει την πιο απλή μορφή:

$$h(t) = \frac{h_0(t)}{(1 + \frac{\sigma^2}{1-\gamma}H_0(t))^{1-\gamma}}$$

2.2.5 Compound Poisson μοντέλο ευπάθειας

Η κατανομή Compound Poisson εισάχθηκε στην ανάλυση μοντέλων ευπάθειας το 1988 από τον Aalen(1988,1992). Η εν λόγω κατανομή είναι ευρέως χρησιμοποιούμενη στη βιοϊατρική και στη δημογραφία. Η κατανομή αυτή μπορεί να οριστεί ως το άθροισμα ενός πλήθους N ανεξάρτητων μεταβλητών X_i, με N να είναι τ.μ. που

ακολουθεί κατανομή Poisson, οι οποίες να ακολουθούν κατανομή Γάμμα. Έστω μοντέλο:

$$Z = \begin{cases} X_1 + X_2 + \ldots + X_N & \text{, αν } N > 0, \\ 0 & \text{, αν } N = 0, \end{cases}$$

με N να ακολουθεί κατανομή Poisson και $X_1, X_2, \ldots, X_N$ να ακολουθούν $\Gamma(k, \lambda)$.

Έστω ο μετασχηματισμός Laplace της κατανομής Γάμμα που αναλύθηκε ανωτέρω L_Γ και L_{Po} ο τύπος μετασχηματισμού Laplace για την κατανομή $Poisson$:

$$L_\Gamma(s) = \left(1 + \frac{s}{\lambda}\right)^{-k}$$

$$L_{Po}(s) = e^{-\rho + \rho e^{-s}}$$

Συνεπώς, ο μετασχηματισμός Laplace για την κατανομή Compound Poisson δίνεται από τον τύπο:

$$L(s) = E(e^{-sZ}) = E\left(e^{-s(X_1 + X_2 + \cdots + X_N)}\right) = E\left(\prod_{i=1}^{N} e^{-sX_i}\right) = E(L_\Gamma(s)^N) =$$

$$= L_{Po}\left(-lnL_\Gamma(s)\right) = L_{Po}\left(-ln\left(1 + \frac{s}{\lambda}\right)^{-k}\right) = \exp\left\{-\rho + \rho e^{ln[\left(1+\frac{s}{\lambda}\right)^{-k}]}\right\}$$

$$= e^{-\rho + \rho\left(1+\frac{s}{\lambda}\right)^{-k}}$$

Χρησιμοποιώντας την παρακάτω παραμετροποίηση :

$$\rho = -\frac{k\lambda^\gamma}{\gamma} \quad , \quad \lambda = \lambda \quad , \quad \kappa = -\gamma \quad ,$$

ο τύπος μετασχηματισμού Laplace της κατανομής compound Poisson παίρνει την εξής μορφή:

$$L(s) = \exp\left\{\frac{k\lambda^\gamma}{\gamma} - \frac{k\lambda^\gamma}{\gamma}\left(1 + \frac{s}{\lambda}\right)^\gamma\right\}$$

$$= \exp\{-\frac{k}{\gamma}[\lambda^\gamma\left(1 + \frac{s}{\lambda}\right)^\gamma - \lambda^\gamma]\}$$

$$= e^{-\frac{k}{\gamma}[(\lambda+s)^\gamma - \lambda^\gamma]}$$

Με τη βοήθεια, λοιπόν, του μετασχηματισμού Laplace, βρίσκουμε τις συναρτήσεις επιβίωσης και κινδύνου:

$$S(t) = e^{-\frac{k}{\gamma}[(\lambda+H_0(t))^{\gamma}-\lambda^{\gamma}]}$$

$$h(t) = kh_0(t)(\lambda + H_0(t))^{\gamma-1}$$

2.3 Πολυμεταβλητά μοντέλα ευπάθειας

Στην ανάλυση επιβίωσης μοντέλων ευπάθειας πολλές φορές παρατηρείται ετερογένεια όσον αφορά στο χρόνο επιβίωσης ορισμένων μελών του πληθυσμού. Η ομαδοποίηση της ετερογένειας αυτής, ανάλογα με τις ιδιαιτερότητες των μελών του πληθυσμού, οδηγεί στην ανάπτυξη των πολυμεταβλητών μοντέλων επιβίωσης. Συχνά τέτοιες ομαδοποιήσεις παρατηρούμε σε δεδομένα χρόνου επιβίωσης όπως π.χ. εμφάνιση νόσου μεταξύ διδύμων, συγγενικών προσώπων, ανθρώπων που έχουν μια κοινή πάθηση ή κοινό ιστορικό λοιμώξεων κλπ.

Για να εξηγηθεί και να αντιμετωπιστεί η ετερογένεια αυτή, η εξάρτηση μεταξύ μελών των ομάδων εισάγεται στα πολυμετάβλητα μοντέλα επιβίωσης, πολλαπλασιάζοντας την αναφορική συνάρτηση κινδύνου κάθε μέλους της ομάδας με μία λανθάνουσα μεταβλητή ευπάθειας (latent). Για δύο συσχετιζόμενα άτομα με διανύσματα παρατηρούμενων συμμεταβλητών x_1, x_2 και χρόνους επιβίωσης t_1, t_2, οι δεσμευμένες συναρτήσεις επιβίωσης συμβολίζονται με $S(\, t_1 \,|z, x_1\,)$ και $S(\, t_2 \,|z, x_2\,)$ αντίστοιχα.

Η συνάρτηση επιβίωσης για το πολυμεταβλητό μοντέλο παρατηρήσεων σε ζεύγη είναι της μορφής:

$$S(t_1, t_2) = \int_0^{\infty} S(\, t_1 \,|z, x_1\,)\, S(\, t_2 \,|z, x_2\,) g(z) dz \qquad (2.6)$$

με g να ορίζει την πυκνότητα πιθανότητας της ευπάθειας Z η οποία ακολουθεί μια δεδομένη κατανομή.

2.3.1 Shared frailty model – Από κοινού μοντέλα ευπάθειας

Σε ένα μοντέλο ευπάθειας με n ομάδες (clusters), έστω ότι η i κατά σειρά ομάδα αποτελείται από n_i πλήθος ατόμων, τα οποία χαρακτηρίζονται από την κοινή μεταβλητή ευπάθειας $Z_i (1 \leq i \leq n)$. Με X_{ij} $(1 \leq i \leq n,\ 1 \leq j \leq n_i)$ συμβολίζεται το διάνυσμα των παρατηρούμενων ανεξάρτητων συμμεταβλητών του j-οστού ατόμου που ανήκει στην i ομάδα ευπάθειας με χρόνο επιβίωσης T_{ij}.

Το shared μοντέλο ευπάθειας προϋποθέτει ότι τα άτομα της ομάδας i μοιράζονται κοινή ευπάθεια Z_i καθ' όλη τη διάρκεια μελέτης τους.

Η έννοια του shared frailty model εισήχθη από τον Clayton (1978), με την υπόθεση ότι οι χρόνοι επιβίωσης των μελών μιας ευπαθούς ομάδας του πληθυσμού είναι ανεξάρτητοι μεταξύ τους, δεδομένης της ευπάθειάς τους (δηλαδή για δεδομένη τιμή της ευπάθειας τους). Με τον τρόπο αυτό, προσέγγισε σε ένα βαθμό την ομογενοποίηση του πληθυσμού.

Η συνάρτηση επιβίωσης στο shared model είναι της μορφής:

$$h(t|x_{ij}, z_i) = z_i h_0(t) e^{\beta^T x_{ij}} \qquad (2.7)$$

με $i = 1, \dots, n$ και □=1, … , n_i. Η $h_0(t)$ ορίζει την αναφορική συνάρτηση κινδύνου και οι ευπάθειες z_i είναι τυχαίες ανεξάρτητες μεταβλητές με συνάρτηση πυκνότητας-πιθανότητας την συνάρτηση της κατανομής που έχει επιλεγεί να ακολουθεί η ευπάθεια του μοντέλου.

Η δεσμευμένη συνάρτηση επιβίωσης των ατόμων της $i-$ ομάδας που μοιράζονται ευπάθεια Z_i είναι:

$$S(t_{i.}|X_{i.}, z_i) = \prod_{j=1}^{n_i} S\left(t_{ij}|x_{ij}, z_i\right) = \prod_{j=1}^{n_i} \exp\left(-H\left(t_{ij}|x_{ij}, z_i\right)\right) =$$

$$= \prod_{j=1}^{n_i} \exp\left(-z_i H_0\left(t_{ij}\right) e^{\beta^T x_{ij}}\right)$$

$$= \exp(-z_i \sum_{j=1}^{n_i} H_0(t_{ij}) e^{\beta^T x_{ij}})$$

όπου $H_0(t) = \int_o^t h_0(t)ds$ ορίζει την αθροιστική αναφορική συνάρτηση κινδύνου, $X_{i.} = (x_{i1}, \dots, x_{in_i})$ ο πίνακας συμμεταβλητών των ατόμων της i ομάδας και $t_{i.} = (t_{i1}, \dots, t_{in_i})$ το διάνυσμα των αντίστοιχων χρόνων επιβίωσης τους.

Η μη δεσμευμένη ως προς την ευπάθεια συνάρτηση επιβίωσης του μοντέλου είναι:

$$S(t_{i.}|X_{i.}) = ES(t_{i.}|X_{i.}, z_i)$$

$$= E \exp\left(-z_i \sum_{j=1}^{n_i} H_0\left(t_{ij}\right) e^{\beta^T x_{ij}}\right)$$

$$= L(\sum_{j=1}^{n_i} H_0\left(t_{ij}\right) e^{\beta^T x_{ij}})$$

όπου L ο μετασχηματισμός Laplace της επιλεγμένης κατανομής ευπάθειας. Ο τύπος της από κοινού συνάρτησης επιβίωσης όλου του πληθυσμού είναι:

$$S(t_{..}|X_{..}) = S\left(t_{11}, t_{12} \dots, t_{21}, \dots, t_{nn_n}|x_{11}, x_{12}, \dots, x_{21}, \dots, x_{nn_n}\right)$$

$$= \prod_{i=1}^{n} L(\sum_{j=1}^{n_i} H_0\left(t_{ij}\right) e^{\beta^T x_{ij}})$$

ως γινόμενο συναρτήσεων επιβίωσης των n ανεξάρτητων μεταξύ τους ομάδων.

Για τη μονομεταβλητή συνάρτηση επιβίωσης, ισχύει το εξής:

$$S(t_{ij}|x_{ij}) = ES(t_{ij}|x_{ij}, z_i) = E\exp\left(-z_i H_0(t_{ij})e^{\beta^T x_{ij}}\right) = L(H_0(t_{ij})e^{\beta^T x_{ij}})$$

$$\Leftrightarrow$$

$$L^{-1}\left(S(t_{ij}|x_{ij})\right) = H_0(t_{ij})e^{\beta^T x_{ij}}$$

όπου L^{-1} η αντίστροφη συνάρτηση του μετασχηματισμού Laplace.

Έτσι, η μη δευσμευμένη συνάρτηση επιβίωσης για κάθε ομάδα i μπορεί να δοθεί ως εξής:

$$S(t_{i\cdot}|x_{i\cdot}) = L(L^{-1}(S(t_{i1}|x_{i1})) + \cdots + L^{-1}\left(S(t_{in_i}|x_{in_i})\right))$$

2.3.2 Μοντέλα ευπάθειας συσχετιζόμενων δεδομένων (correlated frailty models)

Στην ανάλυση δεδομένων για διμεταβλητούς χρόνους αποτυχίας, η ευπάθεια ενός ζεύγους δεδομένων ενδέχεται να χαρακτηρίζεται από δύο τυχαίες μεταβλητές, οι οποίες είναι μεταξύ τους συσχετιζόμενες. Οι ευπάθειες των μελών του ζεύγους σε ένα μοντέλο ευπάθειας συσχετιζόμενων δεδομένων δεν είναι αναγκαστικά οι ίδιες και σε αυτό συνίσταται και η διαφοροποίησή του από τα shared frailty models. Η συσχέτιση μπορεί να έχει είτε θετικό είτε αρνητικό πρόσημο, το τελευταίο, δε, εμφανίζεται στην περίπτωση όπου η συσχετιζόμενη μεταβλητή επιδρά ''ιαματικά'' στην ευπάθεια κάποιου μέλους.

Οι ευπάθειες στο μοντέλο ευπάθειας συσχετιζόμενων δεδομένων δρουν πολλαπλασιαστικά στην αναφορική συνάρτηση κινδύνου, ανεξαρτητοποιώντας τους χρόνους επιβίωσης των μελών του ζεύγους.

Οι συσχετιζόμενες τυχαίες μεταβλητές ακολουθούν την κοινή κατανομή ευπάθειας που έχουμε υποθέσει. Για παράδειγμα, έστω τρείς θετικές σταθερές k_0 , k_1 , k_2 και

$\lambda_0 = k_0$, $\lambda_1 = k_0 + k_1$, $\lambda_2 = k_0 + k_2$. Επίσης έστω Y_0 , Y_1 , Y_2 ανεξάρτητες τυχαίες μεταβλητές κατανομής Γάμμα με $Y_0 \sim \Gamma(k_0, \lambda_0)$, $Y_1 \sim \Gamma(k_1, \lambda_1)$, $Y_2 \sim \Gamma(k_2, \lambda_2)$.

Επομένως ,

$$Z_1 = \frac{\lambda_0}{\lambda_1} Y_0 + Y_1 \sim \Gamma(k_0 + k_1, \lambda_1)$$

$$Z_2 = \frac{\lambda_0}{\lambda_2} Y_0 + Y_2 \sim \Gamma(k_0 + k_1, \lambda_2)$$

με $EZ_1 = EZ_2 = 1$, $V(Z_1) = \frac{1}{\lambda_1} := {\sigma_1}^2$, $V(Z_2) = \frac{1}{\lambda_2} := {\sigma_2}^2$

Οι Yashin and Iachine (1995) για την ανάλυση δεδομένων συσχετιζόμενου μοντέλου ευπάθειας χρησιμοποίησαν τη διμεταβλητή κατανομή επιβίωσης με μορφή:

$$S(t_1, t_2) = \frac{S_1(t_1)^{1-\frac{\sigma_1}{\sigma_2}\rho} \, S_2(t_2)^{1-\frac{\sigma_2}{\sigma_1}\rho}}{(S_1(t_1)^{-{\sigma_1}^2} + S_2(t_2)^{-{\sigma_2}^2} - 1)^{\frac{\rho}{\sigma_1 \sigma_2}}}$$

με ρ να συμβολίζει την συσχέτιση των ατόμων του ζεύγους , όπου:

$$\rho = \frac{cov(Z_1, Z_2)}{\sqrt{V(Z_1)V(Z_2)}} = \frac{k_0}{\sqrt{(k_0 + k_1)(k_0 + k_2)}}$$

και $0 \leq |\rho| \leq \min\{\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_1}\}$

Παρατηρούμε ότι, αν $\sigma_1 \neq \sigma_2$ τότε η συσχέτιση είναι μικρότερη του 1.

Συνεπώς, όσο περισσότερο διαφέρουν οι τυπικές αποκλίσεις μεταξύ τους, τόσο μικρότερος είναι ο βαθμός συσχέτισης της ευπάθειας των δύο παρατηρήσεων.

2.4 Γενικευμένη Συνάρτηση Πιθανοφάνειας μοντέλων ευπάθειας

Η συνάρτηση πιθανοφάνειας στα μοντέλα ευπάθειας χρησιμοποιείται για την εκτίμηση των συντελεστών παλινδρόμησης και των παραμέτρων της εκάστοτε κατανομής.

Όπως έχουμε ήδη αναφέρει, η συνάρτηση κινδύνου ενός μοντέλου ευπάθειας ορίζεται από τον τύπο *(2.1)*:

$$h(t|z,x) = zh_0(t)e^{\beta^T X}$$

όπου z η μεταβλητή ευπάθειας, $h_0(t)$ η αναφορική συνάρτηση κινδύνου και β το διάνυσμα των συντελεστών των ανεξάρτητων συμμεταβλητών του μοντέλου.

Η υπό συνθήκη συνάρτηση επιβίωσης του μοντέλου δίνεται, αντιστοίχως, από τον τύπο:

$$S(t|z,x) = e^{-zH_0(t)e^{\beta^T X}}$$

Για την εύρεση της συνάρτησης επιβίωσης δεδομένου μόνο των συμμεταβλητών X αρκεί να ολοκληρώσουμε την $S(t|z,x)$ ως προς την ευπάθεια. Συγκεκριμένα,

$$S(t|x) = \int_0^\infty e^{-zH_0(t)e^{\beta^T x}} \mathrm{d}F_z(z) = e^{-G(e^{\beta^T x}H_0(t))} \qquad (2.8)$$

με

$$G(x) = -\ln\left(\int_0^\infty e^{-zx}\mathrm{d}F_z(z)\right) \qquad (2.9)$$

και F_z η κατανομή ευπάθειας.

Παρατηρούμε ότι η συνάρτηση G είναι ίση με το μείον του λογαρίθμου του μετασχηματισμού Laplace της εκάστοτε κατανομής ευπάθειας.

Όπως αναφέρθηκε στην ενότητα 2.2.1. και την σχέση (2.5):

$$S(t) = L_\Gamma\left(H_0(t)\right) = (1 + \frac{H_0(t)}{\lambda})^{-\alpha}$$

Έτσι μέσω της σχέσης (2.8) ισχύει η ισότητα:

$$e^{-G(x)} = \left(1 + \frac{x}{\lambda}\right)^{-\alpha} \qquad \leftrightarrow$$

$$G(x) = -\ln\left(1 + \frac{x}{\lambda}\right)^{-\alpha}$$

με α παράμετρος σχήματος (shape) και λ παράμετρος θέσης (rate).

Όταν η κατανομή ευπάθειας ακολουθεί κατανομή $\Gamma\acute{\alpha}\mu\mu\alpha(a,\kappa)$, με a παράμετρος shape και κ παράμετρος scale(scale=1/rate), τότε η συνάρτηση G δίνεται από τον ακόλουθο τύπο:

$$G\big(x, \Gamma(a,\kappa)\big) = -ln(1 + x \cdot \kappa)^{-a} = a \cdot \ln(1 + x \cdot \kappa) \qquad (2.10)$$

Αντιστοίχως, όταν η κατανομή ευπάθειας ακολουθεί κατανομή $Inverse\ Gaussian(\mu, \lambda)$, η συνάρτηση G ορίζεται ως εξής:

$$G\big(x, InvGauss(\mu,\lambda)\big) = -\frac{\lambda}{\mu}\left(1 - \sqrt{1 + \frac{2\mu^2 x}{\lambda}}\right)$$

$$= \sqrt{(\frac{\lambda}{\mu})^2 + 2\lambda x} \quad - \quad \frac{\lambda}{\mu} \qquad (2.11)$$

Για λόγους αναγνωρισιμότητας, η μέση τιμή της ευπάθειας ορίζεται συνήθως ίση με 1 ($EZ = 1$). Στην περίπτωση που η ευπάθεια εκτιμάται ότι ακολουθεί κατανομή $Gamma$ έστω ότι $Z \sim \Gamma\acute{\alpha}\mu\mu\alpha(\frac{1}{k}, \mathrm{k})$, με παράμετρο σχήματος $\frac{1}{k}$ και scale παράμετρο k έχουμε $E(Z) = 1$ και $Var(Z) = k,$ ενώ στην περίπτωση που η ευπάθεια εκτιμάται ότι ακολουθεί κατανομή αντίστροφη Γκαουσιανή με παραμέτρους μ=1 και $\lambda = \frac{1}{\sigma^2}$: shape, θα υποθέσουμε στη συνέχεια το μετασχηματισμό $E(Z) = 1\ και\ Var(Z) = \sigma^2 = 1/2b$. Στην γενική περίπτωση των λογοκριμένων δεδομένων (όταν υπάρχουν παρατηρήσεις για τις οποίες δεν γνωρίζουμε όλο το διάστημα επιβίωσης τους παρά μόνο ένα μέρος του), η συνάρτηση πιθανοφάνειας ορίζεται ως εξής:

$$L = \Pi_i[h(t_i)^{\delta_i} S(t_i)]$$

ή

$$L = \Pi_{i\epsilon U} f(t_i) \Pi_{i\epsilon C} S(t_i) = \Pi_i\big\{f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}\big\} = \Pi_i\big\{h(t_i)^{\delta_i} S(t_i)\big\}$$

όπου $\delta_i = \begin{cases} 1, & \mu\eta\ \lambda o\gamma o\kappa\rho\iota\mu\acute{\epsilon}\nu\epsilon\varsigma \\ 0, & \lambda o\gamma o\kappa\rho\iota\mu\acute{\epsilon}\nu\epsilon\varsigma \end{cases}$

και $U = \{\sigma\acute{\upsilon}\nu o\lambda o\ \mu\eta\ \lambda o\gamma o\kappa\rho\iota\mu\acute{\epsilon}\nu\omega\nu\ \pi\alpha\rho\alpha\tau\eta\rho\acute{\eta}\sigma\epsilon\omega\nu\}$

$C = \{\sigma\acute{\upsilon}\nu o\lambda o\ \lambda o\gamma o\kappa\rho\iota\mu\acute{\epsilon}\nu\omega\nu\ \pi\alpha\rho\alpha\tau\eta\rho\acute{\eta}\sigma\epsilon\omega\nu\}$

Θέτουμε $W_i = \min(T_i, C_i)$ δηλαδή για την i παρατήρηση W_i είναι ο μικρότερος εκ των χρόνων επιβίωσης και λογοκρισίας.

Οπότε κάθε παρατήρηση μπορεί να χαρακτηρισθεί πλήρως από το διάνυσμα πληροφορίας (W_i, X_i, δ_i) όπου X_i το διάνυσμα συμμεταβλητών.

Εισάγοντας την ευπάθεια στη συνάρτηση πιθανοφάνειας, μέσω των μετασχηματισμών Laplace, αυτή παίρνει την ακόλουθη μορφή με βάση τις σχέσεις (1.2), (2.4.3) και (2.3):

$$L = \prod_i [\left\{ \left(-e^{\beta^T X_i} h_0(W_i) \right) \frac{L'\left(e^{\beta^T X_i} H_0(W_i) \right)}{L\left(e^{\beta^T X_i} H_0(W_i) \right)} \right\}^{\delta_i} L(e^{\beta^T X_i} H_0(W_i))]$$

$$L = \prod_i [\left\{ \left(e^{\beta^T X_i} h_0(W_i) \right) G'\left(e^{\beta^T X_i} H_0(W_i) \right) \right\}^{\delta_i} e^{-G(e^{\beta^T X_i} H_0(W_i))}]$$

βάσει του ορισμού της συνάρτησης G στη σχέση που ορίστηκε προηγουμένως στην (2.9) , δηλαδή $G(w) = -\ln(\int_0^\infty e^{-zw} dF_z(z))$.

Λογαριθμίζοντας τη συνάρτηση πιθανοφάνειας προκύπτει ο εξής τύπος:

$$logL = \sum_{i=1}^{n} \delta_i [\log(G'\left(e^{\beta^T X_i} H_0(W_i) \right)) + \log\left(e^{\beta^T X_i} h_0(W_i) \right)] - \sum_{i=1}^{n} G(e^{\beta^T X_i} H_0(W_i))$$

Στη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας, εκτός από το διάνυσμα β, το οποίο θέλουμε να εκτιμήσουμε, έχουμε και τις άγνωστες παραμέτρους $h_0(W_i)$ και $H_0(W_i)$. Για διευκόλυνση, ορίζουμε τη μορφή της αναφορικής συνάρτησης κινδύνου. Θεωρούμε λοιπόν ένα παραμετρικό μοντέλο ευπάθειας. Για παράδειγμα, στην περίπτωση που υποθέτουμε ότι οι παρατηρήσεις μας ακολουθούν την εκθετική κατανομή $Exp(\lambda)$, τότε η αναφορική συνάρτηση κινδύνου θα είναι της μορφής:

$$h_0(t) = h_0(t, \lambda) = \lambda$$

Συνεπώς η αθροιστική αναφορική συνάρτηση κινδύνου είναι :

$$H_0(t) = H_0(t, \lambda) = \lambda t$$

Άρα, η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας παίρνει την κάτωθι μορφή:

$$logL = \sum_{i=1}^{n} \delta_i [\log (G'(e^{\beta^T X_i} \lambda W_i)) + \log(e^{\beta^T X_i} \lambda)] - \sum_{i=1}^{n} G(e^{\beta^T X_i} \lambda W_i)$$

και η μεγιστοποίηση της $logL$ θα γίνει ως προς τις παραμέτρους λ και β.

Αναλόγως, στην περίπτωση που υποθέτουμε ότι οι παρατηρήσεις μας ακολουθούν κατανομή $Gompertz$ και η αναφορική συνάρτηση κινδύνου είναι της μορφής:

$$h_0(t) = h_0(t, \mu, b) = \mu e^{bt}$$

η αθροιστική αναφορική συνάρτηση κινδύνου είναι της απλής μορφής :

$$H_0(t) = H_0(t, \mu, b) = \frac{\mu e^{bt}}{b}$$

Τότε η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας παίρνει την μορφή :

$$logL = \sum_{i=1}^{n} \delta_i [\log (G'\left(\mu \frac{\exp(\beta^T X_i + bW_i)}{b}\right)) + \log(\mu * \exp(\beta^T X_i + bW_i))] - \sum_{i=1}^{n} G\left(\mu \frac{\exp(\beta^T X_i + bW_i)}{b}\right)$$

Στην ως άνω περίπτωση, η μεγιστοποίηση της $logL$ θα έχει σκοπό την εκτίμηση των παραμέτρων β, μ, b.

ΒΙΒΛΙΟΓΡΑΦΙΑ ΚΕΦΑΛΑΙΟΥ 2

Aalen, O.O. (1988.) Heterogeneity in Survival Analysis. Statistics in Medicine 7, 1121 - 1137

Androulakis, E., Koukouvinos, C. and Vonta, F. (2012). Estimation and variable selection via frailty models with penalized likelihood, Statist. Med. 2012, 31, 2223–2239

Duchateau, L., Janssen, P. (2008). The Frailty Model, Springer, New York, Biometrical Journal, 51(3):540-541

Hanagal, D.D. (2007b). Gamma frailty regression models in mixture distributions. Economic Quality Control, 22(2), 295-302.

Henderson, R., Oman, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. Journal of the Royal Statistical Society B 61, 367 – 379

Hougaard, P. (1986b). A class of multivariate failure time distributions. Biometrika 73, 671-78

Hougaard, P. (1991). Modelling Heterogeneity in Survival Data. Journal of Applied Probability 28, 695 – 701

McGilchrist, C.A., Aisbett, C.W. (1991). Regression with Frailty in Survival Analysis. Biometrics 47, 461-466

Vonta, F. (1996). Efficient estimation in a nonproportional hazards model in survival analysis, Scand. J. Statist., 23, No 1, 49-61

ΚΕΦΑΛΑΙΟ 3

Μέτρα Απόκλισης (Divergence Measures)

3.1 Εισαγωγή

Πριν από αρκετές δεκαετίες, ο στατιστικός Mahalanobis (1936) εισήγαγε την έννοια της απόστασης μεταξύ δύο κατανομών πιθανότητας, με σκοπό να μελετηθεί κατά πόσο δύο κατανομές πιθανοτήτων είναι πιθανοτικά «κοντά» μεταξύ τους . Η ιδέα αυτή ερευνήθηκε, ακόμη, από πολλούς άλλους στατιστικούς, όπως ο Rao (1949, 1954), ο οποίος όρισε το μέτρο απόστασης Rao, ο Chernoff (1952), με το μέτρο διακριτών δεδομένων (measures of discriminatory information), ο Kullback (1959), με το Kullback divergence distance, καθώς και ο Kolmogorov (1963), ο οποίος εισήγαγε τα μέτρα measures of variation-distance. Η τιμή των εν λόγω μέτρων αυξάνεται όσο μεγαλύτερη είναι η «απόσταση» μεταξύ δυο κατανομών. Εφεξής τα μέτρα αυτά θα αναφέρονται με τον όρο «μέτρα απόκλισης» ή «divergence measures».

3.2 Μέτρα απόκλισης

Στην παρούσα ενότητα θα αναλύσουμε ορισμένα σημαντικά μέτρα απόκλισης, τα οποία είναι χρήσιμα για τη διερεύνηση πλήθους στατιστικών μελετών.

Έστω X τυχαία μεταβλητή και $(\mathcal{X}, \beta_\chi, P_\theta)$ χώρος πιθανοτήτων, όπου $\mathcal{X}$ ο χώρος τιμών για την τυχαία μεταβλητή X και β_χ σ-άλγεβρα υποσυνόλων Borel. Συμβολίζουμε με $\{P_\theta\}, \theta \in \Theta$ την οικογένεια κατανομών που είναι απολύτως συνεχείς σε σχέση με το μέτρο μ το οποίο είναι ένα σ-πεπερασμένο μέτρο πάνω στο χώρο $(\mathcal{X}, \beta_\chi)$. Για λόγους ευκολίας ορίζουμε το μέτρο μ να είναι είτε το μέτρο Lebesque ($P_\theta(C) = 0$ για ένα ενδεχόμενο C που έχει μέτρο Lebesgue ίσο με 0), είτε ένα μέτρο απαρίθμησης για το

οποίο υπάρχει κάποιο πεπερασμένο ή μετρήσιμο σύνολο S_X για το οποίο ισχύει ότι $P_\theta(\mathcal{X}- S_X)=0$. Παρακάτω παραθέτουμε τη συνάρτηση $f_\theta(x)$ όταν το μ είναι μέτρο Lebesgue και όταν είναι μέτρο απαρίθμησης αντίστοιχα:

$$f_\theta(x) = \frac{dP_\theta}{d\mu}(x) = \begin{cases} f_\theta(x) & \text{αν } \mu \text{ είναι μέτρο Lebesgue} \\ Pr_\theta(X = x) = p_\theta(x) & \text{με } x \epsilon S_X \text{ αν } \mu \text{ μέτρο απαρίθμησης} \end{cases}$$

Έστω δυο μέτρα πιθανότητας P_{θ_1} και P_{θ_2}, τα οποία αντιστοιχούν στις συναρτήσεις κατανομής F_{θ_1} και F_{θ_2} .

Θα μπορούσαμε να ορίσουμε ως F_{θ_1} την κατανομή που ακολουθούν τα δεδομένα μας και με F_{θ_2} την εμπειρική κατανομή των δεδομένων ή την κατανομή που θέλουμε να συγκρίνουμε για να ελέγξουμε αν η τελευταία δύναται να περιγράψει τα δεδομένα αυτά.

3.2.1 Απόσταση Kolmogorov (Kolmogorov Distance)

Η απόσταση Kolmogorov μεταξύ των F_{θ_1} και F_{θ_2}(ή μεταξύ των μέτρων πιθανότητας P_{θ_1} και P_{θ_2}) ορίζεται ως εξής:

$$K_1\left(F_{\theta_1}, F_{\theta_2}\right) = sup_{x\in\mathbb{R}}|F_{\theta_1}(x) - F_{\theta_2}(x)|$$

όπου $sup_{x\in\mathbb{R}}A$ το άνω φράγμα του συνόλου A, για κάθε $x \in \mathbb{R}$.

3.2.2 Απόσταση Lévy (Lévy Distance)

Η απόσταση Lévy μεταξύ των F_{θ_1} και F_{θ_2}(ή μεταξύ των μέτρων πιθανότητας P_{θ_1} και P_{θ_2}) ορίζεται ως εξής:

$$K_2\left(F_{\theta_1}, F_{\theta_2}\right) = \inf\{\varepsilon > 0 : F_{\theta_1}(x-\varepsilon) \le F_{\theta_2}(x) \le F_{\theta_1}(x+\varepsilon)\}$$

3.2.3 Μέτρο Απόκλισης Kullback-Leibler (Kullback-Leibler Divergence Measure)

Ένα πολύ γνωστό μέτρο απόστασης-απόκλισης είναι αυτό των Kullback-Leibler (1951). Ειδικότερα, το εν λόγω μέτρο μεταξύ των κατανομών πιθανοτήτων P_{θ_1} και P_{θ_2} ορίζεται ως εξής:

$$D_{Kull}(\theta_1, \theta_2) = \int_X f_{\theta_1}(x)\, log \frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} d\mu(x)$$

$$= E_{\theta_1}[\log\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right)]$$

(3.1)

Παρόλη τη χρησιμότητα του εν λόγω μέτρου, υπολείπεται ως προς την ιδιότητα της συμμετρίας, καθώς $D_{Kull}(\theta_1, \theta_2) \neq D_{Kull}(\theta_2, \theta_1)$.

3.2.4 Μέτρο Απόκλισης Jeffreys (Jeffreys Divergence Measure)

Ο στατιστικός Jeffreys επιχείρησε να ορίσει ένα συμμετρικό μέτρο απόστασης με βάση το μέτρο των Kullback-Leibler. Το μέτρο αυτό, μεταξύ των κατανομών πιθανοτήτων P_{θ_1} και P_{θ_2} , συμβολίζεται $J(\theta_1, \theta_2)$ και ορίζεται ως εξής:

$$J(\theta_1, \theta_2) = D_{Kull}(\theta_1, \theta_2) + D_{Kull}(\theta_2, \theta_1)$$

3.2.5 Μέτρο Απόκλισης Rényi (Rényi Divergence Measure)

Ο Rényi (1961) παρουσίασε μια γενίκευση της απόστασης μεταξύ δύο κατανομών με παραμέτρους θ_1 *και* θ_2 και κατανομή πιθανότητας P_{θ_1} και P_{θ_2} αντίστοιχα, με το κάτωθι μέτρο απόστασης:

$$D_r^1(\theta_1, \theta_2) = \frac{1}{r-1}\, log \int_X f_{\theta_1}(x)^r\, f_{\theta_2}(x)^{1-r}\; d\mu(x)$$

$$= \frac{1}{r-1} \log(E_{\theta_1}\left[\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right)^{r-1}\right])$$

για $r > 0$, $r \neq 1$.

Το μέτρο αυτό επεκτάθηκε, το 1987, από τους Liese and Vajda ώστε να ορίζεται για κάθε $r \neq 0,\ r \neq 1$. Αυτό επετεύχθη πολλαπλασιάζοντας το μέτρο του Rényi με $\frac{1}{r}$

Συγκεκριμένα:

$$D'^1_r(\theta_1, \theta_2) = \frac{1}{r(r-1)}\ log \int_X f_{\theta_1}(x)^r\, f_{\theta_2}(x)^{1-r}\ d\mu(x)$$

$$= \frac{1}{r(r-1)} \log(E_{\theta_1}\left[\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right)^{r-1}\right])$$

Στην περίπτωση όπου $r = 1\ και\ r = 0$, ισχύει αντίστοιχα το κάτωθι:

$$D'^1_1(\theta_1, \theta_2) = \lim_{r\to 1} D'^1_r(\theta_1, \theta_2) = D_{kull}(\theta_1, \theta_2)$$

$$και$$

$$D'^1_0(\theta_1, \theta_2) = \lim_{r\to 0} D'^1_r(\theta_1, \theta_2) = D_{kull}(\theta_2, \theta_1)$$

3.3 φ-Divergence Measures

Το μέτρο απόκλισης φ (φ divergence measure) μεταξύ των κατανομών πιθανότητας P_{θ_1} και P_{θ_2} ορίζεται ως εξής:

$$\begin{aligned} D_\varphi(P_{\theta_1}, P_{\theta_2}) &= D_\varphi(\theta_1, \theta_2) = \\ &= \int_X f_{\theta_2}(x)\, \varphi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) d\mu(x) \\ &= E_{\theta_2}\left[\varphi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right)\right] \end{aligned} \qquad (3.2)$$

με $\varphi \in \Phi^*$, όπου Φ^* είναι η οικογένεια όλων των κυρτών συναρτήσεων $\varphi(x)$ για τις οποίες ισχύουν οι παρακάτω ιδιότητες:

Για κάθε $x \geq 0$:

- $\varphi(1) = 0$
- $0 * \varphi({}^{0}/_{0}) = 0$
- $0 * \varphi({}^{p}/_{0}) = \lim_{u\to\infty} {}^{\varphi(u)}/_{u}$

Πρόταση 3.3

Έστω $\varphi \in \Phi^*$ συνάρτηση διαφορίσιμη με συνεχή παράγωγο για $x = 1$. Τότε, η συνάρτηση $\psi(x)$, με μορφή:

$$\psi(x) = \varphi(x) - \varphi'(1)(x-1)$$

ανήκει, επίσης, στην Φ^*, όπου $\psi(1) = \varphi(1) - \varphi'(1) * 0 = \varphi(1) = 0$
και έχει παράγωγο $\psi'(x) = \varphi'(x) - \varphi'(1)$,δηλαδή για $x = 1$: $\psi'(1) = 0$.

Επίσης ισχύει ότι:

$$D_\psi(\theta_1, \theta_2) = \int_X f_{\theta_2}(x) \left[\varphi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) - \varphi'(1)\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} - 1\right)\right] d\mu(x) =$$

$$= \int_X f_{\theta_2}(x)\, \varphi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) d\mu(x) = D_\varphi(\theta_1, \theta_2)$$

Ήτοι τα δυο μέτρα απόκλισης είναι ίσα όταν υπάρχει η εξής ισοδυναμία συνόλων:
$\Phi \equiv \Phi^* \cap \{\varphi : \varphi'(1) = 0\}$.

Στον παρακάτω πίνακα παρουσιάζονται.

ϕ-function	Divergence
$x\log x - x + 1$	Kullback-Leibler (1959)
$-\log x + x - 1$	Minimum Discrimination Information
$(x-1)\log x$	J-Divergence
$\frac{1}{2}(x-1)^2$	Pearson (1900), Kagan (1963)
$\frac{(x-1)^2}{(x+1)^2}$	Balakrishnan and Sanghvi (1968)
$\frac{-x^s+s(x-1)+1}{1-s}$, $s \neq 1$,	Rathie and Kannappan (1972)
$\frac{1-x}{2} - \left(\frac{1+x^{-r}}{2}\right)^{-1/r}$, $r > 0$	Harmonic mean (Mathai and Rathie (1975))
$\frac{(1-x)^2}{2(a+(1-a)x)}$, $0 \leq a \leq 1$	Rukhin (1994)
$\frac{ax\log x-(ax+1-a)\log(ax+1-a)}{a(1-a)}$, $a \neq 0, 1$	Lin (1991)
$\frac{x^{\lambda+1}-x-\lambda(x-1)}{\lambda(\lambda+1)}$, $\lambda \neq 0, -1$	Cressie and Read (1984)
$\lvert 1-x^a\rvert^{1/a}$, $0 < a < 1$	Matusita (1964)
$\lvert 1-x\rvert^a$, $a \geq 1$	χ – divergence of order a (Vajda 1973) Total Variation if $a = 1$ (Saks 1937)

3.4 Goodness-of-fit: Simple Null Hypothesis

Προκειμένου να μελετηθεί, με βάση το test Goodness of fit, μια κατανομή πιθανότητας σε πραγματικό χρόνο, ενδείκνυται ο διαχωρισμός των δεδομένων σε M ίσα διαστήματα, ελέγχοντας την υπόθεση $H_0 : p = p^0$ σε έκαστο εξ αυτών.

Έστω $P = \{E_i\}_{i=1,\dots,M}$, ο διαμερισμός του συνόλου των δεδομένων πραγματικού χρόνου σε M διαστήματα. Επίσης, έστω ότι $p = (p_1, \dots, p_M)^T$ και $p^0 = (p_1^0, \dots, p_M^0)^T$ είναι τα διανύσματα πιθανοτήτων των δεδομένων που ανήκουν στα διαστήματα E_i για $i = 1, \dots, M$, για την πραγματική και την υποθετική κατανομή πιθανοτήτων αντίστοιχα. Συγκεκριμένα, $p_i = Pr_F(E_i)$ και $p_i^0 = Pr_{F_0}(E_i) = \int_{E_i} dF_0$, για $i = 1, \dots, M$.

Έστω τυχαίο δείγμα $Y_1, Y_2, \dots, Y_n$ κατανομής F και $N_i = \sum_{j=1}^{n} I_{E_i}(Y_i)$,

με

$$I_{E_i}(Y_i) = \begin{cases} 1\,, & \alpha\nu\ Y_i \in E_i \\ 0, & \alpha\nu\ \notin E_i \end{cases} \tag{3.3}$$

και $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_M)^T$, με $\hat{p}_i = N_i/n$ για $i = 1, \ldots, M$, να είναι η εκτίμηση της πιθανότητας μια τυχαία παρατήρηση να βρίσκεται στο i υποσύνολο.

Ένας συνήθης τρόπος για να ελεγχθεί η μηδενική υπόθεση:

$$H_0 : p = p^0$$

είναι μέσω του Pearson's test statistic ή αλλιώς X^2-test statistic, όπου:

$$X^2 \equiv \sum_{i=1}^{M} \frac{(N_i - np_i^0)^2}{np_i^0} \tag{3.4}$$

Ενώ, ο έλεγχος δύναται να πραγματοποιηθεί και μέσω του likelihood ratio test, όπου:

$$G^2 \equiv 2\sum_{i=1}^{M} N_i \cdot log\,\frac{N_i}{np_i^0}$$

Τα δύο ως άνω test statistics ανήκουν στην οικογένεια power-divergence test statistics, η οποία εισήχθηκε από τους Cressie and Read(1984) και δίνεται από την ελεγχοσυνάρτηση:

$$T_n^{\lambda}(\hat{p}, p^0) = \frac{2n}{\lambda(\lambda+1)}\sum_{i=1}^{M} \hat{p}_\iota \left(\left(\frac{\hat{p}_\iota}{p_i^0}\right)^{\lambda} - 1\right) = \frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{M} \mathrm{N}_i \left(\left(\frac{\mathrm{N}_i}{np_i^0}\right)^{\lambda} - 1\right)$$

όπου $-\infty < \lambda < \infty$.

Τα test $T_n^0(\hat{p}, p^0)$ και $T_n^{-1}(\hat{p}, p^0)$ ορίζονται από την ελεγχοσυνάρτηση $T_n^{\lambda}(\hat{p}, p^0)$ καθώς $\lambda \to 0$ και $\lambda \to -1$ αντιστοίχως. Για διάφορες τιμές της μεταβλητής λ, δημιουργούνται διαφορετικά test μέσω της οικογένειας power-divergence test statistics.

Στον παρακάτω πίνακα αναφέρονται ορισμένα γνωστά test με βάση την τιμή της μεταβλητής λ:

λ	*Test Statistic*	$T_n^{\lambda}(\hat{p}, p^0)$
λ=0	Likelihood ratio test	$2\sum_{i=1}^{M} N_i \log \frac{N_i}{np_i^0}$
λ=1	Chi-square test	$\sum_{i=1}^{M} \frac{(N_i - np_i^0)^2}{np_i^0}$
λ=-1/2	Freeman-Tukey test	$8n\left(1 - \sum_{i=1}^{M} \sqrt{\frac{p_i^0 N_i}{n}}\right)$
λ=-1	Minimum discrimination information test	$2\sum_{i=1}^{M} N_i \log\left(\frac{np_i^0}{N_i}\right)$
λ=-2	Modified Chi-square test	$\sum_{i=1}^{M} \frac{(np_i^0 - N_i)^2}{N_i}$
λ=2/3	Cressie-Read test	$\frac{9}{5}n\left(\sum_{i=1}^{M} \hat{p_i}\left(\frac{\hat{p_i}}{p_i^0}\right)^{2/3} - 1\right)$

Όπως αποδεικνύεται από τα ανωτέρω, η οικογένεια power-divergence test statistics παρουσιάζει αξιοσημείωτη ευελιξία, γεγονός που την καθιστά πολύ σημαντικό εργαλείο για τη σύγκριση δυο κατανομών πιθανοτήτων.

Μια πιο απλή και γενικευμένη οικογένεια test statistics είναι η φ −divergence test statistics η οποία ορίζεται ως εξής:

$$T_n^{\varphi}(\hat{p}, p^0) = \frac{2n}{\varphi''(1)} \sum_{i=1}^{M} p_i^0 \varphi\left(\frac{\hat{p}_i}{p_i^0}\right) \ , \quad \varphi \in \Phi^* \qquad (3.5)$$

3.5 Phi-divergences and Goodness-of-fit with Fixed Number of Classes

Ο Pearson(1900) απέδειξε ότι $X^2 \xrightarrow[n\to\infty]{} X_{M-1}^2$ με $X^2 \equiv \sum_{i=1}^{M} \frac{(N_i - np_i^0)^2}{np_i^0}$, δηλαδή απέδειξε ότι η ελεγχοσυνάρτηση $T_n^{\lambda}(\hat{p}, p^0)$ τείνει στη X_{M-1}^2 για $\lambda = 1$. Εν συνεχεία, οι

Cressie and Read(1984) επέκτειναν την ως άνω σχέση για κάθε $\lambda \in \mathrm{R}$ όπου κάτω από την μηδενική υπόθεση $H_0 : p = p^0$, η ελεγχοσυνάρτηση $T_n^{\lambda}(\hat{p}, p^0) \xrightarrow[n\to\infty]{} X_{M-1}^2$.

Ακολούθως, οι Zografos et. al (1990) έδειξαν ότι $T_n^{\varphi}(\hat{p}, p^0) \xrightarrow[n\to\infty]{} X_{M-1}^2$ κάτω από τη μηδενική υπόθεση $H_0 : p = p^0$ για κάθε $\varphi \in \Phi^*$.

<u>Θεώρημα 3.5.1</u>

Κάτω από τη μηδενική υπόθεση $H_0 : p = p^0 = (p_1^0, \dots, p_M^0)^T$, η ασυμπτωτική κατανομή της συναρτήσεως φ-divergence test statistic, $T_n^{\varphi}(\hat{p}, p^0)$, είναι X^2 με $M-1$ βαθμούς ελευθερίας:

$$T_n^{\varphi}(\hat{p}, p^0) = \frac{2n}{\varphi''(1)} D_{\varphi}(\hat{p}, p^0) \xrightarrow[n\to\infty]{} X_{M-1}^2 \qquad (3.6)$$

Βάσει των προαναφερθέντων, σε μια μελέτη με μεγάλο πλήθος παρατηρήσεων έχουμε τη δυνατότητα, μέσω της chi-square κατανομής X_{M-1}^2 με $M-1$ βαθμούς ελευθερίας, να αποδεχτούμε ή να απορρίψουμε τη μηδενική υπόθεση. Η εν λόγω απόφαση παίρνεται σε επίπεδο σημαντικότητας $\alpha = \Pr\left(X_{M-1}^2 > X_{M-1,\alpha}^2\right)$, όπου απορρίπτουμε τη μηδενική υπόθεση $H_0 : p = p^0 = (p_1^0, \dots, p_M^0)^T$ αν:

$$T_n^{\varphi}(\hat{p}, p^0) > X_{M-1,\alpha}^2$$

3.6 Composite null hypothesis - Εκτιμητής Μέγιστης Πιθανοφάνειας (Maximum Likelihood Estimator)

Έστω $(\mathcal{X}, \beta_{\chi}, P_{\theta})$ ο χώρος πιθανοτήτων τυχαίας μεταβλητής X με β_{χ} σ-άλγεβρα Borel υποσυνόλων $A \subset \mathcal{X}$ και $\{P_{\theta}\}_{\theta\in\Theta}$ οικογένεια κατανομών πιθανοτήτων, οι οποίες ορίζονται από το μετρικό χώρο $(\mathcal{X}, \beta_{\chi})$ με Θ ένα ανοιχτό σύνολο στο $\mathbb{R}^{M_0}$. Θέτοντας $\mathrm{P} = \{E_i\}_{i=1,\dots,M}$ ως μια διαμέριση του $\mathcal{X}$, η πιθανότητα $p_i(\theta) = Pr_{\theta}(E_i)$, $i =$

$1, \dots, M$ προσδιορίζει ένα διακριτό στατιστικό μοντέλο, όπου $Pr_\theta(E_i)$ είναι η πιθανότητα, δεδομένου του θ, να ανήκει ένα τυχαίο δείγμα στο σύνολο E_i .
Έστω $Y_1, \dots, Y_n$ ένα τυχαίο δείγμα του πληθυσμού που περιγράφεται από τη μεταβλητή X και έστω $N_i = \sum_{j=1}^{n} I_{E_i}(Y_j)$, όπου I η δείκτρια συνάρτηση, δηλαδή $I_{E_k}(Y_m) = 1$ εάν $Y_m \in E_k$, αλλιώς $I_{E_k}(Y_m) = 0$. Επιπροσθέτως, $\hat{p}_i = \frac{N_i}{n}$, $i = 1, \ . \ . \ . \ , M$.
Για την εκτίμηση της παραμέτρου θ, μέσω της μεθόδου της μέγιστης πιθανοφάνειας, ενδείκνυται να γίνει μεγιστοποίηση για καθορισμένα $n_1, \dots, n_M$,όπου:

$$Pr_\theta(N_1 = n_1, \dots, N_M = n_M) = \frac{n!}{n_1! \cdots n_M!}(p_1(\theta)^{n_1} \cdots p_M(\theta)^{n_M}) \tag{3.7}$$

Δοθέντων των ως άνω, η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας παίρνει τη μορφή:

$$\begin{aligned} l(\theta) &= \log\big(Pr_\theta(N_1 = n_1, \dots, N_M = n_M)\big) \\ &= \log\left(\frac{n!}{n_1! \cdots n_M!}\right) + n\sum_{i=1}^{M} \hat{p}_i \log(p_i(\theta)) \\ &= \log\left(\frac{n!}{n_1! \cdots n_M!}\right) - n\sum_{i=1}^{M} \hat{p}_i \log\left(\frac{1}{p_i(\theta)}\right) + \\ &\quad + n\sum_{i=1}^{M} \hat{p}_i \log(\hat{p}_i) - n\sum_{i=1}^{M} \hat{p}_i \log(\hat{p}_i) \\ &= -n \cdot D_{Kull}(\hat{p}, p(\theta)) + k \end{aligned} \tag{3.8}$$

με $\hat{p} = (\hat{p}_1, \dots, \hat{p}_M)^T$, $p(\theta) = (p_1(\theta), \dots, p_M(\theta))^T$,
$D_{Kull}\big(\hat{p}, p(\theta)\big) = \sum_{i=1}^{M} \hat{p}_i \log\left(\frac{\hat{p}_i}{p_i(\theta)}\right)$ και $k = \log\left(\frac{n!}{n_1! \cdots n_M!}\right) + n\sum_{i=1}^{M} \hat{p}_i \log(\hat{p}_i)$
ανεξάρτητο του θ.

Κατά συνέπεια, για την εκτίμηση του θ, είναι αναγκαία η μεγιστοποίηση της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας ή ομοίως το minimizing του μέτρου απόκλισης Kullback-Leibler. Η διαδικασία αυτή πραγματώνεται όμοια, μεγιστοποιώντας το $-D_{Kull}\big(\hat{p}, p(\theta)\big)$.

3.7 Πίνακας Πληροφορίας του Fisher (Fisher Information Matrix)

Στο πλαίσιο της στατιστικής, με τον όρο Fisher Information $I(\theta)$ εννοούμε την ποσότητα της πληροφορίας που παρέχει μια τυχαία μεταβλητή Y για μια παράμετρο θ, με βάση τη συνάρτηση πιθανότητας $p(\mathrm{Y}|\theta)$. Το μέτρο αυτό ορίζεται ως εξής:

$$I(\theta^*) = V_{\theta^*}([\frac{\partial}{\partial\theta}\log p(\mathrm{Y}|\theta)]_{\theta=\theta^*}) \qquad (3.9)$$

όπου $V(X)$ η διακύμανση της τυχαίας μεταβλητής X.

Για την τιμή της πληροφορίας του Fisher, ισχύει η σχέση:

$$I(\theta^*) = E((\frac{\partial}{\partial\theta}\log p(\mathrm{Y}|\theta))^2|\theta = \theta^*) \qquad \Leftrightarrow$$

$$I(\theta^*) = [\int_0^\infty \left(\frac{\partial}{\partial\theta}\log f(y,\theta)\right)^2 \cdot f(y,\theta)\, dy]|_{\theta=\theta^*} \qquad (3.10)$$

για μία παρατήρηση του δείγματος.

Εάν η $\log p(\mathrm{Y}|\theta)$ είναι δύο φορές παραγωγίσιμη, τότε το μέτρο Fisher Information ισούται με:

$$I(\theta^*) = E_{\theta^*}(-[\frac{\partial^2}{\partial\theta^2}\log p(\mathrm{Y}|\theta)]_{\theta=\theta^*}) \qquad (3.11)$$

όπου $E(X)$ η μέση τιμή της κατανομής της τυχαίας μεταβλητής X.

Αντίστοιχα Fisher Information Matrix αποκαλούμε μια γενίκευση του Fisher Information, κατά την οποία δημιουργείται ένας πίνακας, στον οποίο το στοιχείο που ανήκει στην i γραμμή και στην j στήλη δίνεται από τον ακόλουθο τύπο:

$$I(\theta^*)_{i,j} = E_{\theta^*}(([\frac{\partial}{\partial\theta_i}\log p(\mathrm{Y}|\theta)]_{\theta=\theta^*})([\frac{\partial}{\partial\theta_j}\log p(\mathrm{Y}|\theta)]_{\theta=\theta^*}))$$

Ομοίως, αποδεικνύεται ότι :

$$I(\theta^*)_{i,j} = -E_{\theta^*}([\frac{\partial^2}{\partial\theta_i\theta_j}\log p(\mathrm{Y}|\theta)]_{\theta=\theta^*})$$

Θεώρημα 3.7.1

Οι Morales et. al (1995) διερεύνησαν την ελεγχοσυνάρτηση $T_n^{\varphi}(\hat{\theta})$, όπου $\hat{\theta}$ η εκτίμηση της παραμέτρου θ μέσω της μεθόδου MLE και απέδειξαν ότι η ασυμπτωτική της κατανομή δίνεται από την σχέση:

$$\frac{2n}{\varphi''(1)} D_{\varphi}\left(\hat{p}, p(\hat{\theta})\right) \xrightarrow[n\to\infty]{} X^2_{M-M_0-1} + \sum_{j=1}^{M_0} (1-\lambda_j) Z_j^2 \qquad (3.12)$$

όπου Z_j με $j = 1,2,\dots,M_0$ είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν τυπική κανονική κατανομή $Z_j \sim N(0,1)$ και λ_j όπου $0 \le \lambda_j \le 1$, είναι οι λύσεις της εξίσωσης:

$$\det\left(\mathbf{I}_F(\theta_0) - \lambda I_F(\theta_0)\right) = 0\,, \qquad (3.13)$$

δεδομένου ότι $\mathbf{I}_F(\theta_0)$ και $I_F(\theta_0)$ είναι πίνακες πληροφορίας Fisher (Fisher Information matrices) του πραγματικού (original) και του διακριτού (discretized) μοντέλου αντίστοιχα και $\theta_0 \in \Theta$ με Θ υποσύνολο του ανοιχτού συνόλου $\mathbb{R}^{M_0}$ (σύνολο όλων των πραγματικών αριθμών διάστασης M_0) και $M_0 < M - 1$.

3.8 Κατανομή αθροίσματος ανεξάρτητων Γάμμα κατανεμόμενων τυχαίων μεταβλητών

Έστω X_i , για $i = 1,\dots,n$, ένα σύνολο ανεξάρτητων τυχαίων μεταβλητών που ακολουθούν Γάμμα κατανομή με παραμέτρους $\alpha_i > 0$ και $\beta_i > 0$. Η συνάρτηση πυκνότητας-πιθανότητας των τ.μ. δίνεται από τον τύπο:

$$f_i(x_i) = \frac{1}{{\beta_i}^{\alpha_i}\Gamma(\alpha_i)} {x_i}^{\alpha_i - 1} e^{-\frac{x_i}{\beta_i}} \quad , \text{για } x_i > 0$$

$$\text{και} \quad f_i(x_i) = 0 \ \text{για} \quad x_i \le 0$$

Ο Moschopoulos (1985) (και πριν από αυτόν ο Mathai (1982)) ερεύνησε την κατανομή της συνάρτησης $Y = X_1 + X_2 + \cdots + X_n$ προσεγγίζοντας την κατανομή του

αθροίσματος Γάμμα κατανομών μέσω της ροπογεννήτριας συνάρτησης m.g.f. (moment generating function). Συγκεκριμένα, λόγω του ότι οι κατανομές X_i είναι ανεξάρτητες, η συνάρτηση ροπογεννήτριας της Y δίνεται από τον τύπο:

$$\boldsymbol{M}(t) = \prod_{i=1}^{n}(1-\beta_i t)^{-\alpha_i}$$

Υποθέτουμε ότι $\beta_1 = \min(\beta_i)$. Η σχέση $(1-\beta_i t)$ λοιπόν παίρνει την μορφή:

$$1-\beta_i t = (1-\beta_1 t)(\frac{\beta_i}{\beta_1})[1-(1-\frac{\beta_1}{\beta_i})/(1-\beta_1 t)]$$

Λογαριθμίζοντας την ροπογεννήτρια $\boldsymbol{M}(t)$ έχουμε ότι:

$$\log \boldsymbol{M}(t) = \log[C\cdot(1-\beta_1 t)^{-\rho}] + \sum_{k=1}^{\infty}\gamma_k\,(1-\beta_1 t)^{-k}$$

όπου, $C = \prod_{i=1}^{n}(\frac{\beta_1}{\beta_i})^{\alpha_i}$ (3.14)

$$\gamma_k = \frac{\sum_{i=1}^{n}\alpha_i\left(1-\frac{\beta_1}{\beta_i}\right)^k}{k},\quad k = 1,2,.. \tag{3.15}$$

$$\rho = \sum_{i=1}^{n}\alpha_i > 0. \tag{3.16}$$

Η σχέση αυτή ισχύει για κάθε t που επαληθεύει την ανίσωση:

$$max_i\left|(1-\frac{\beta_1}{\beta_i})/(1-\beta_1 t)\right| < 1.$$

Έτσι η ροπογεννήτρια $M(t)$ παίρνει την μορφή:

$$M(t) = C\cdot(1-\beta_1 t)^{-\rho}\exp(\sum_{k=1}^{\infty}\gamma_k\,(1-\beta_1 t)^{-k})$$

Θέτοντας $\exp(\sum_{k=1}^{\infty}\gamma_k\,(1-\beta_1 t)^{-k}) = \sum_{k=0}^{\infty}\delta_k\,(1-\beta_1 t)^{-k}$

και παραγωγίζοντας αυτή τη σχέση ως προς $(1-\beta_1 t)^{-1}$ βρίσκουμε ότι αναδρομικά οι συντελεστές δ_k είναι της μορφής:

$$\delta_{k+1} = \frac{1}{1+k}\sum_{i=1}^{k+1} i\gamma_i \delta_{k+1-i} \ , \qquad k = 0,1,2, \ \ldots \tag{3.17}$$

και $\delta_0 = 1$.

ΘΕΩΡΗΜΑ 3.8

Για $\{X_\iota\}$, $i = 1,2,\ldots,n$ ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν Γάμμα κατανομή με παραμέτρους $\alpha_i > 0$ και $\beta_i > 0$ αντίστοιχα, η συνάρτηση πυκνότητας της συνάρτησης $Y = X_1 + X_2 + \cdots + X_n$ μπορεί να εκφραστεί ως εξής:

$$g(y) = C\sum_{k=o}^{\infty} \delta_k \, y^{\rho+k-1} e^{-\frac{y}{\beta_1}} / [\Gamma(\rho+k){\beta_1}^{\rho+k}], \quad y > 0$$

και $g(y) = 0$ για $y \leq 0$

με ρ, δ_k, C όπως ορίστηκαν παραπάνω.

Η συνάρτηση κατανομής, λοιπόν, $\boldsymbol{F}(w) = \Pr(Y \leq w)$ λαμβάνει την παρακάτω μορφή:

$$\boldsymbol{F}(w) = C\sum_{k=o}^{\infty} \delta_k \int_{k=0}^{w} \frac{y^{\rho+k-1} \cdot e^{-\frac{y}{\beta_1}}}{\left[\Gamma(\rho+k)\cdot {\beta_1}^{\rho+k}\right]} dy, \quad w \in R. \tag{3.18}$$

ΒΙΒΛΙΟΓΡΑΦΙΑ ΚΕΦΑΛΑΙΟΥ 3

Agresti, A. (1996). An Introduction to Categorical Data Analysis. John Wiley & Sons, New York, 2,1-53

Balakrishnan, V. and Sanghvi, L. D. (1968). Distance between populations on the basis of attribute. Biometrics, 24, 4, 859-865

Bonett, D. G. (1989). Pearson chi-square estimator and test for log-linear models with expected frequencies subject to linear constraints. Statistics and Probability Letters, 8, 175-177

Broffit, J. D. and Randles, R. H. (1977). A power approximation for the chi-square goodness-of-fit test: Simple hypothesis case. Journal of the American Statistical Association, 72, 604-607

Ferentinos, K., Papaioannou, T. and Zografos (1990). φ-divergence statistics: Sampling properties, multinomial goodness of fit and divergence tests. Communications in Statistics (Theory and Methods), 19, 5, 1785-1802

Mathai (1982). Storage capacity of a dam with gamma type inputs. Annals of the Institute of Statistical Mathematics, 34, 591–597

Menendez, M. L., Morales, D., Pardo, L. and Vajda, I. (2001b). Minimum divergence estimators based on grouped data. Annals of the Institute of Statistical Mathematics, 53,2, 277-288

Menendez, M. L., Morales, D., Pardo, L. and Salicru, M. (1995). Asymptotic behavior and statistical applications of divergence measures in multinomial populations: A unified study. Statistical Papers, 36, 1-29

Moschopoulos, P. G. (1984). The distribution of the sum of independent gamma random variables. Annals of the Institute of Statistical Mathematics, 37, 541–544

Pardo, L.(2005). Statistical Inference Based on Divergence Measures. Statistics: A Series of Textbooks and Monographs Book 185. Mathematics & Statistics, 1st Edition , Chapters 1-6

ΚΕΦΑΛΑΙΟ 4

Έλεγχος μέσω μέτρων απόκλισης για τα μοντέλα ευπάθειας

Σε αυτό το κεφάλαιο προτείνουμε έναν έλεγχο καλής προσαρμογής βάσει των μέτρων απόκλισης για μοντέλα ευπάθειας. Η ελεγχοσυνάρτηση στην οποία θα βασιστεί ο έλεγχος δίνεται στην (3.5) και συγκεκριμένα είναι η φ −divergence test statistic η οποία ορίζεται ως εξής:

$$T_n^{\varphi}(\hat{p}, p^0) = \frac{2n}{\varphi''(1)} \sum_{i=1}^{M} p_i^0 \varphi(\frac{\hat{p}_i}{p_i^0}) \ , \quad \varphi \in \Phi^*$$

Η μηδενική υπόθεση του ελέγχου καλής προσαρμογής είναι

$$\mathcal{H}_0 : \ p^0(t, \theta) = 1 - S_0(t, \theta) = 1 - e^{-G(H_0(t))} \tag{4.1}$$

όπου $S_0(t)$ η συνάρτηση επιβίωσης και $H_0(t)$ η αθροιστική συνάρτηση κινδύνου κάτω από τη μηδενική υπόθεση. Αυτός ο ορισμός των μοντέλων ευπάθειας δίνεται στη σχέση (2.4) για την περίπτωση συμμεταβλητών $x = 0$ και η συνάρτηση G ορίζεται στην (2.3).

Σκοπός μας είναι να εξετάσουμε τη συμπεριφορά του προτεινόμενου ελέγχου μέσω του μεγέθους και της ισχύος του για να δούμε αν ο έλεγχος είναι αποτελεσματικός για τα μοντέλα ευπάθειας. Θα θεωρήσουμε δύο μοντέλα ευπάθειας, το Γάμμα μοντέλο ευπάθειας και το Inverse Gaussian μοντέλο ευπάθειας. Η συνάρτηση φ που θα θεωρήσουμε θα είναι αυτή των Kullback-Leibler.

Η διερεύνηση της αποδοτικότητας του προτεινόμενου ελέγχου θα γίνει μέσω προσομοιώσεων βάσει κώδικα που έχει γραφεί στο στατιστικό πακέτο R και στο περιβάλλον R Studio.

Η ασυμπτωτική κατανομή της ελεγχοσυνάρτησης $T_n^{\varphi}(\hat{p}, p^0)$ θα θεωρηθεί γνωστή και δίνεται από τις σχέσεις (3.12)-(3.13) του θεωρήματος 3.7.1 του Κεφαλαίου 3. Το θεώρημα δίνεται στην εργασία Morales et al. (1995).

4.1 Υποθέσεις και θεωρητικό υπόβαθρο για τις προσομοιώσεις

4.1.1 Συνάρτηση αθροιστικού κινδύνου

Θα υποθέσουμε στη συνέχεια ότι η αθροιστική αναφορική συνάρτηση κινδύνου, όπως είδαμε στο δεύτερο κεφάλαιο, δίνεται απο την εκθετική κατανομή $Exp(\theta)$, οπότε :

$$H_0(t) = H_0(t, \theta) = \theta t \qquad (4.2)$$

και συνεπώς η παράμετρος $\theta \in R$ και $M_0 = 1$, η διάσταση της παραμέτρου θ.

4.1.2 Γάμμα μοντέλο ευπάθειας

Θυμίζουμε εδώ ότι το να προέρχονται τα δεδομένα από κάποιο μοντέλο ευπάθειας σημαίνει ότι στον πληθυσμό παρατηρείται ετερογένεια μεταξύ των ατόμων που τον αποτελούν.

Υποθέτουμε ότι η μεταβλητή ευπάθειας ακολουθεί Γάμμα κατανομή $\Gamma(\alpha, \kappa)$ με α παράμετρο σχήματος (shape) και κ παράμετρο κλίμακας (scale). Θεωρήσαμε αυτή την περίπτωση γιατί η περίπτωση της Γάμμα κατανομής είναι ευρέως χρησιμοποιούμενη στην ανάλυση επιβίωσης λόγω της ευκολίας που παρουσιάζει στους μαθηματικούς υπολογισμούς.

Όπως αναφέρθηκε στο κεφάλαιο (2.2.1) και μέσω της σχέσης (2.10) ισχύει ότι:

$$G\big(H_0(t)\big) = -ln(1 + H_0(t) \cdot \kappa)^{-\alpha} = \alpha \cdot \ln(1 + H_0(t) \cdot \kappa) \qquad (4.3)$$

και

$$S(t) = e^{-G(H_0(t))} \quad \leftrightarrow$$

$$-lnS(t) = G(H_0(t)) \tag{4.4}$$

Έτσι εισάγουμε στο μοντέλο μας κάτω από την μηδενική υπόθεση, ευπάθεια που ακολουθεί Γάμμα κατανομή συνδυάζοντας τις ισότητες (4.2) και (4.3), λαμβάνοντας την παρακάτω ισότητα :

$$G\big(H_0(t)\big) = -ln(1 + \theta \cdot t \cdot \kappa)^{-\alpha}$$

$$\xrightarrow{(4.4)} \quad -lnS(t) = -ln(1 + \theta \cdot t \cdot \kappa)^{-\alpha}$$

$$\longrightarrow \quad S(t) = (1 + \theta \cdot t \cdot \kappa)^{-\frac{1}{\kappa}} \tag{4.5}$$

όπου $\alpha = \frac{1}{\kappa}$ αφού για λόγους αναγνωρισιμότητας των παραμέτρων στα μοντέλα ευπάθειας γενικά απαιτείται η μέση τιμή της ευπάθειας να είναι 1, άρα εδώ θα πρέπει να υποθέσουμε ότι $E(T) = \alpha \cdot \kappa = 1 \quad \Rightarrow \alpha = \frac{1}{\kappa}$. Στο μοντέλο αυτό, όπου α θα αντικαθιστούμε με $\frac{1}{\kappa}$. Στη συνέχεια λοιπόν στις προσομοιώσεις θα υποθέσουμε μοντέλο ευπάθειας Γάμμα με μέση τιμή 1 και διασπορά κ.

4.1.3. Inverse Gaussian μοντέλο ευπάθειας

Στην περίπτωση που η μεταβλητή ευπάθειας ακολουθεί Inverse Gaussian κατανομή $IG(\mu, \lambda)$ με μ να είναι η μέση τιμή και λ η παράμετρος σχήματος (shape), όπως αναφέρθηκε στο κεφάλαιο 2 και μέσω της εξίσωσης (2.11) , η συνάρτηση G ορίζεται ως εξής:

$$G\big(x, InvGauss(\mu, \lambda)\big) = \sqrt{(\frac{\lambda}{\mu})^2 + 2\lambda x} \quad - \quad \frac{\lambda}{\mu}$$

Για λόγους αναγνωρισιμότητας των παραμέτρων στα μοντέλα ευπάθειας γενικά απαιτείται η μέση τιμή της ευπάθειας να είναι 1. Υποθέτοντας ότι η παράμετρος

σχήματος λ ισούται με $2b$, ή αντίστοιχα η διασπορά με $\frac{1}{2b}$, η συνάρτηση ευπάθειας G λαμβάνει την παρακάτω μορφή:

$$G(x) = -2b + \sqrt{4b^2 + 4b \cdot x}$$

$$\xrightarrow{x=H_0(t)}$$

$$G\big(H_0(t)\big) = -2b + \sqrt{4b\big(b + H_0(t)\big)}$$

Έτσι, λαμβάνοντας υπόψιν ότι η αθροιστική συνάρτηση κινδύνου δίνεται από την εκθετική κατανομή $Exp(\theta)$:

$$\xrightarrow{(4.2)} \qquad G\big(H_0(t)\big) = -2b + \sqrt{4b(b + \mathrm{t} \cdot \theta)} \qquad (4.6)$$

Μέσω της εξίσωσης (4.4) η παραπάνω εξίσωση γίνεται:

$$-lnS(t) = -2b + \sqrt{4b(b + \mathrm{t} \cdot \theta)}$$

$$\Rightarrow \qquad S(t) = \exp\{\, 2b - \sqrt{4b(b + \mathrm{t} \cdot \theta)}\,\} \qquad (4.7)$$

Στη συνέχεια λοιπόν στις προσομοιώσεις θα υποθέσουμε μοντέλο ευπάθειας Inverse Gaussian με μέση τιμή 1 και διασπορά $\frac{1}{2b}$.

4.1.4 Ομαδοποίηση - Διαμερισμός των χρόνων επιβίωσης

Όπως είδαμε στο κεφάλαιο 3.4, προκειμένου να μελετηθεί, με βάση το test Goodness of fit, μια κατανομή πιθανότητας, είναι απαραίτητος ο διαχωρισμός των δεδομένων σε M (πιθανώς ίσα) διαστήματα, ελέγχοντας την υπόθεση $H_0 : p = p^0$ σε έκαστο εξ αυτών. Έστω τυχαίο δείγμα $T_1, \dots, T_n$ από χρόνους επιβίωσης. Ταξινομούμε τους χρόνους επιβίωσης T_i σε αύξουσα σειρά, δηλαδή έχουμε $t_1 < t_2 < \cdots < t_n$ και δημιουργούμε M διαστήματα, όπου κάθε διάστημα j, με $j = 1,2, \dots, M$, έχει την μορφή:

$$\left[\ t_1 + (j-1) \cdot \frac{t_n - t_1}{M} \quad ,\ t_1 + j \cdot \frac{t_n - t_1}{M}\ \right) \ , \qquad \mu\varepsilon\ j = 1, \dots, M$$

Ορίζουμε τα άκρα των διαστημάτων αυτών ${t'}_{a_j}$ και ${t'}_{b_j}$ με:

$$t'_{a_j} = t_1 + (j-1) \cdot \frac{t_n - t_1}{M}$$
$$t'_{b_j} = t_1 + j \cdot \frac{t_n - t_1}{M}$$

Κατά τη διάρκεια των προσομοιώσεων παρατηρήσαμε ότι ο τρόπος αυτός δεν ήταν ο ιδανικός διότι κάποια διαστήματα προς το τέλος δεν περιείχαν καθόλου παρατηρήσεις πράγμα που δυσκόλευε την αποδοτική εκτίμηση των παραμέτρων. Εναλλακτικά αποφασίσαμε να χωρίσουμε σε M διαστήματα που περιέχουν τις παρατηρήσεις με ίση πιθανότητα. Συγκεκριμένα, ορίσαμε M διαστήματα που περιέχουν το ίδιο πλήθος παρατηρήσεων, όπου σε κάθε ένα από αυτά περιέχονται n/M παρατηρήσεις, με την προυπόθεση ότι n/M είναι ακέραιος. Αλλιώς θα πάρουμε το ακέραιο μέρος του n/M. Έτσι ορίζουμε κάθε διάστημα j, με $j = 2, \dots, M-1$, να έχει την μορφή:

$$[\ t_{\left(\frac{n}{M}\right)*(j-1)}\ ,\ t_{\left(\frac{n}{M}\right)*j+1}\ )\ ,\quad \text{με } j = 2, \dots, M-1$$

ενώ το πρώτο διάστημα που θα περιέχει τις πρώτες n/M παρατηρήσεις να είναι το εξής : $[\ t_1\ ,\ t_{\left(\frac{n}{M}\right)+1}\ )$

καθώς αντίστοιχα το τελευταίο διάστημα είναι το εξής :

$$[\ t_{\left(n-\frac{n}{M}\right)}\ ,\ t_n + 0{,}0001\ )$$

Για ευκολία συμβολίζουμε τα άκρα των διαστημάτων αυτών t'_{a_j} και t'_{b_j} με j να είναι το j-στο διάστημα το οποίο δημιουργείται από τα άκρα αυτά.

Έστω $P = \{E_j\}_{j=1,\dots,M}$, ο διαμερισμός του συνόλου των χρόνων επιβίωσης σε M διαστήματα. Επίσης, έστω ότι $p = (p_1, \dots, p_M)^T$ και $p^0 = (p_1^0, \dots, p_M^0)^T$ να είναι τα διανύσματα πιθανοτήτων να ανήκουν οι χρόνοι επιβίωσης στα διαστήματα E_j για $j = 1, \dots, M$, για την πραγματική και την υποθετική, κάτω από τη μηδενική υπόθεση, κατανομή πιθανοτήτων αντίστοιχα. Συγκεκριμένα, $p_j = Pr_F(E_j)$ και $p_j^0 = Pr_{F_0}(E_j) = \int_{E_j} dF_0$, για $j = 1, \dots, M$.

Είναι γνωστό ότι η πιθανότητα μία μονάδα-άτομο του πληθυσμού να της συμβεί το γεγονός που μελετάμε μεταξύ των χρόνων t_1 και t_2, με $t_1 \le t_2$ ισούται με:

$$P(t_1 \le T \le t_2) = \int_{t_1}^{t_2} f(u)du = F(t_2) - F(t_1)$$

Η συνάρτηση επιβίωσης μιας μονάδας στον χρόνο t_k έχει μορφή:

$$S(t_k) = P(T > t_k) = 1 - F(t_k)$$

Συνδυάζοντας λοιπόν, τις δύο παραπάνω εξισώσεις λαμβάνουμε την παρακάτω μορφή:

$$P(t_1 \le T \le t_2) = S(t_1) - S(t_2)$$

Η εξίσωση αυτή δίνει την πιθανότητα, κάτω από την μηδενική υπόθεση (4.1), μια τυχαία παρατήρηση για το υποθετικό μοντέλο να βρίσκεται στο j διάστημα, με $j = 1, \dots, M$.

Πιο συγκεκριμένα, για κάθε διάστημα j, η πιθανότητα μια παρατήρηση να ανήκει σε αυτό, ισούται με:

$$p_j^0 = P\left(t'_{a_j} \le T \le t'_{b_j}\right) = S\left(t'_{a_j}\right) - S\left(t'_{b_j}\right) =$$

$$= e^{-G\left(H_0(t'_{a_j})\right)} - e^{-G\left(H_0\left(t'_{b_j}\right)\right)} \qquad (4.8)$$

Για την περίπτωση που η ευπάθεια ακολουθεί Γάμμα κατανομή $\Gamma(\frac{1}{\kappa}, \kappa)$ και η αθροιστική συνάρτηση κινδύνου ορίζεται από την εκθετική κατανομή, η πιο πάνω πιθανότητα, μέσω της ισότητας (4.5), ισούται με:

$$p_j^0 = \left(1 + \theta \cdot t'_{a_j} \cdot \kappa\right)^{-\frac{1}{\kappa}} - \left(1 + \theta \cdot t'_{b_j} \cdot \kappa\right)^{-\frac{1}{\kappa}} \qquad (4.9)$$

Αντίστοιχα στην περίπτωση που η ευπάθεια ακολουθεί αντίστροφη Γκαουσιανή κατανομή $IG(1,2b)$ και η αθροιστική συνάρτηση κινδύνου ορίζεται από την εκθετική κατανομή, η πιο πάνω πιθανότητα, μέσω της ισότητας (4.6), ισούται με:

$$p_j^0 = \exp\left\{2b - \sqrt{4b\left(b + t'_{a_j} \cdot \theta\right)}\right\} - \exp\left\{2b - \sqrt{4b\left(b + t'_{b_j} \cdot \theta\right)}\right\} \qquad (4.10)$$

4.1.5 Ελεγχοσυνάρτηση και μέτρα απόκλισης

Υπενθυμίζουμε όπως είδαμε από τη σχέση (3.5), ότι η ελεγχοσυνάρτηση που θα χρησιμοποιήσουμε ορίζεται ως

$$T_n^{\varphi}(\hat{p}, p^0) = \frac{2n}{\varphi''(1)} \sum_{i=1}^{M} p_i^0 \varphi\left(\frac{\hat{p}_i}{p_i^0}\right) \ , \quad \varphi \in \Phi^*$$

Θα χρειαστεί να επιλέξουμε μια συνάρτηση φ όπως είδαμε στην πρόταση 3.3. Θα επιλέξουμε αυτή των Kullback-Leibler, η οποία είναι της μορφής :

$$\varphi(x) = x \cdot logx - x + 1 \tag{4.11}$$

με $\varphi'(x) = logx$ και $\varphi''(x) = \frac{1}{x}$.

'Εχουμε μιλήσει ήδη για το πως ορίζονται οι πιθανότητες $p_i^0, i = 1, .., M$. Για τον υπολογισμό της ελεγχοσυνάρτησης αναφέρουμε πρώτα πως ορίζονται οι εκτιμήτριες $\hat{p}_i, i = 1, .., M$ από τα δεδομένα $T_1, \dots, T_n$. Οπως ορίστηκαν στην αρχή της παραγράφου 3.4, η εκτίμηση της πιθανότητας μια τυχαία παρατήρηση να βρίσκεται στο i διάστημα, για το διακριτό μοντέλο, ισούται με:

$$\hat{p}_i = N_i/n, \quad \text{για} \quad i = 1, \dots, M, \ \text{με}$$

$$N_i = \sum_{j=1}^{n} I_{E_i}(T_j) \tag{4.12}$$

4.1.6 Ασυμπτωτική κατανομή της ελεγχοσυνάρτησης κάτω απο την μηδενική υπόθεση

Παρακάτω θα αναλύσουμε την διαδικασία για την εύρεση της ασυμπτωτικής κατανομής της ελεγχοσυνάρτησης $T_n^{\varphi}(\hat{p}, p^0)$ που δίνεται στη σχέση (3.12) (σε συνδυασμό με τη σχέση (3.13)) του θεωρήματος 3.7.1 του Κεφαλαίου 3. Το θεώρημα δίνεται στην εργασία Morales et. al (1995) και πιο συγκεκριμένα έχουμε ότι:

$$\frac{2n}{\varphi''(1)} D_{\varphi}\left(\hat{p}, p(\hat{\theta})\right) \xrightarrow[n\to\infty]{} X^2_{M-M_0-1} + \sum_{j=1}^{M_0} (1 - \lambda_j) Z_j^2$$

όπου Z_j ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν τυπική κανονική κατανομή $N(0,1)$ και $0 \leq \lambda_j \leq 1$, με λ_j οι λύσεις της εξίσωσης (3.13):

$$\det\left(\mathbf{I}_F(\theta_0) - \lambda \cdot I_F(\theta_0)\right) = 0$$

όπου $\mathbf{I}_F(\theta_0)$ και $I_F(\theta_0)$ είναι πίνακες πληροφορίας Fisher (Fisher Information matrices) του πραγματικού (original) και του διακριτού (discretized) μοντέλου αντίστοιχα και $\theta_0 \in \Theta$ με Θ υποσύνολο του ανοιχτού συνόλου $\mathbb{R}^{M_0}$ (σύνολο όλων των πραγματικών αριθμών διάστασης M_0) και $M_0 < M - 1$.

Το μοντέλο που ερευνούμε σε αυτή την εργασία έχει $\theta \in \Theta$ με Θ υποσύνολο του ανοιχτού συνόλου $\mathbb{R}$ (σύνολο όλων των πραγματικών αριθμών διάστασης 1) και $M_0 = 1 < M - 1$.

Η σχέση (3.12) λοιπόν, παίρνει την παρακάτω μορφή:

$$\frac{2n}{\varphi''(1)} D_\varphi\left(\hat{p}, p(\hat{\theta})\right) \xrightarrow[n\to\infty]{} X^2_{M-2} + (1-\lambda)Z^2 \qquad (4.13)$$

όπου Z είναι τυχαία μεταβλητή που ακολουθεί τυπική κανονική κατανομή $Z \sim N(0,1)$ και $0 \leq \lambda \leq 1$, με λ η λύση της εξίσωσης (3.13):

$$\mathbf{I}_F(\theta_0) - \lambda \cdot I_F(\theta_0) = 0 \rightarrow \lambda = \frac{\mathbf{I}_F(\theta_0)}{I_F(\theta_0)} \qquad (4.14)$$

Για την εύρεση του λ λοιπόν, θα χρειαστεί να υπολογίσουμε τους τύπους από τους οποίους δίνονται γενικά οι δύο πληροφοριακοί αριθμοί $\mathbf{I}_F(\theta_0)$ και $I_F(\theta_0)$ οι οποίοι στη συνέχεια θα πρέπει να υπολογιστούν στις προσομοιώσεις για την πραγματική τιμή του $\theta = \theta_0$ η οποία θα εκτιμηθεί από τα δεδομένα βάσει της μεθόδου μεγίστης πιθανοφάνειας.

Για τον υπολογισμό του πληροφοριακού αριθμού $\mathbf{I}_F(\theta_0)$, του πραγματικού μοντέλου, θα βασιστούμε στην συνάρτηση πιθανοφάνειας. Για ένα ανεξάρτητα και ισόνομα κατανεμημένο δείγμα, η συνάρτηση πιθανοφάνειας είναι:

$$\mathcal{L} = \prod_{i=1}^{n} f(t_i, \theta) = \prod_{i=1}^{n} f(t_i, \theta)$$

Όπως είδαμε στην ανάλυση επιβίωσης, και μέσω της σχέσης (2.4), ισχύουν τα κάτωθι:

$$F(t) = 1 - S(t) \Rightarrow$$

$$F'(t) = f(t) = -S'(t) \overset{(2.4)}{\Longrightarrow}$$

$$f(t) = -\left(e^{-G\left(H_0(t)\right)}\right)' \Rightarrow$$

$$f(t) = h_0(t) e^{-G\left(H_0(t)\right)} G'(w)|_{w=H_0(t)} \qquad (4.15)$$

Έτσι η γενικευμένη συνάρτηση πιθανοφάνειας, για το παραμετρικό μοντέλο, όπου η αθροιστική συνάρτηση κινδύνου είναι γνωστή εκτός από τη μονοδιάστατη παράμετρο θ, λαμβάνει την μορφή:

$$\mathcal{L} = \prod_{i=1}^{n} f(t_i, \theta) = \prod_{i=1}^{n} h_0(t_i, \theta) e^{-G\left(H_0(t_i, \theta)\right)} G'(w)|_{w=H_0(t_i, \theta)}$$

Λογαριθμίζοντας την, έχουμε ότι:

$$\log(\mathcal{L}) = \sum_{i=1}^{n} [\log\left(h_0(t_i, \theta)\right) - G\left(H_0(t_i, \theta)\right) + \log\left(G'(w)|_{w=H_0(t_i, \theta)}\right)]$$

Παραγωγίζοντας ως προς θ, και συμβολίζοντας την παράγωγο ως προς θ με τελεία (ενώ η παράγωγος ως προς w έχει συμβολιστεί με $'$) έχουμε

$$\frac{\partial \log(\mathcal{L})}{\partial \theta} = \sum_{i=1}^{n} [\ \frac{h^{\cdot}{}_0(t_i,\theta)}{h_0(t_i,\theta)} - G'(w)|_{w=H_0(t_i,\theta)} H_0^{\cdot}(t_i,\theta) + \frac{G''(w)|_{w=H_0(t_i,\theta)} H_0^{\cdot}(t_i,\theta)}{G'(w)|_{w=H_0(t_i,\theta)}}\ ] \quad (4.16)$$

Είναι λοιπόν αναγκαίο να ορίσουμε τις συναρτήσεις $h_0(t_i,\theta)$, $h^{\cdot}{}_0(t_i,\theta)$, $G'\big(H_0(t_i,\theta)\big)$, $G''\big(H_0(t_i,\theta)\big)$ και $H_0^{\cdot}(t_i,\theta)$.

Όπως ορίσαμε στην εξίσωση (4.2), για την αθροιστική συνάρτηση κινδύνου στο μοντέλο μας, ισχύει ότι:

$$H_0(t,\theta) = \theta t \qquad \Rightarrow$$

$$h_0(t,\theta) = \frac{d}{dt} H_0\ (t,\theta) = \theta$$

$$H_0^{\cdot}(t,\theta) = t$$

και

$$h^{\cdot}{}_0(t_i,\theta) = 1.$$

Για την διευκόλυνση μας θα χρειαστεί να κάνουμε το συμβολισμό:

$$B(t_i,\theta) = \left[-G'(w)|_{w=H_0(t_i,\theta)} + \frac{G''(w)|_{w=H_0(t_i,\theta)}}{G'(w)|_{w=H_0(t_i,\theta)}}\right] \cdot H_0^{\cdot}(t_i,\theta) + \frac{h^{\cdot}{}_0(t_i,\theta)}{h_0(t_i,\theta)} \quad (4.17)$$

Για την εκθετική κατανομή η συνάρτηση παίρνει τη μορφή

$$B(t_i,\theta) = \left[-G'(t_i\theta) + \frac{G''(t_i\theta)}{G'(t_i\theta)}\right] \cdot t_i + \frac{1}{\theta} \quad (4.18)$$

Έτσι λοιπόν η συνάρτηση πιθανοφάνειας γράφεται με την παρακάτω μορφή:

$$\frac{\partial \log(\mathcal{L})}{\partial \theta} = \sum_{i=1}^{n} B(t_i, \theta)$$

Σύμφωνα με την παράγραφο (3.7), για τη συνάρτηση πληροφορίας του Fisher για το αρχικό μοντέλο και για μία παρατήρηση, είδαμε ότι ισχύει η σχέση (3.10), η οποία με τον κατάλληλο συμβολισμό γράφεται ως:

$$\mathbf{I}(\theta^*) = \int_0^{\infty} \left(\frac{\partial}{\partial \theta} \log f(y, \theta)\right)^2 \cdot f(y, \theta)\, dy$$

Έτσι για τη δική μας περίπτωση ο πληροφοριακός αριθμός με βάση μία παρατήρηση και υπολογισμένος στο θ_0 γράφεται ως

$$\mathbf{I}_1(\theta_0) = \int_0^{\infty} [\left(\frac{\partial}{\partial \theta} \log f(t_i, \theta)\right)^2 \cdot f(t_i, \theta)]|_{\theta=\theta_0}\, dt_i \qquad \Leftrightarrow$$

$$\mathbf{I}_1(\theta_0) = \int_0^{\infty} [\left(B(t_i, \theta)\right)^2 \cdot f(t_i, \theta)]|_{\theta=\theta_0}\, dt_i \qquad (4.19)$$

Ο πληροφοριακός αριθμός για όλες τις παρατηρήσεις από γνωστά αποτελέσματα γράφεται ως

$$\mathbf{I}_F(\theta_0) = n\mathbf{I}_1(\theta_0)$$

Ετσι λοιπόν υπολογίζουμε τον πληροφοριακό αριθμό του Fisher $\mathbf{I}_F(\theta_0)$ για το πραγματικό-αρχικό μοντέλο.

4.1.6.1 Πληροφοριακός αριθμός του Fisher $\mathbf{I}_F(\theta_0)$ για το Γάμμα μοντέλο ευπάθειας

Για Γάμμα κατανομή ευπάθειας, δηλαδή $Z \sim \Gamma\left(\frac{1}{\kappa}, \kappa\right)$, (παράμετροι shape-scale) με μέση τιμή $E(Z) = 1$ και $Var(Z) = \kappa$ μέσω της εξίσωσης (4.3) βλέπουμε ότι:

$$G(w) = \frac{1}{\kappa}\ln(1 + \mathrm{w} \cdot \kappa) \qquad \Rightarrow$$

$$G'(w)|_{w=H_0(t_i,\theta)} = \frac{\kappa}{\kappa(1+w\cdot\kappa)}|_{w=H_0(t,\theta)} = \frac{1}{1+w\cdot\kappa}|_{w=H_0(t,\theta)} = \frac{1}{(1+H_0(t,\theta)\cdot\kappa)}$$

και:

$$G''(w)|_{w=H_0(t_i,\theta)} = -\frac{\kappa}{(1+w\cdot\kappa)^2}|_{w=H_0(t,\theta)} = -\frac{\kappa}{(1+H_0(t,\theta)\cdot\kappa)^2}$$

Συνεπώς, για τη Γάμμα κατανομή, η (4.17) γίνεται,

$$B(t_i,\theta) = \left\{-\frac{1}{1+\kappa\theta t_i} - \frac{\frac{\kappa}{(1+\kappa\theta t_i)^2}}{\frac{1}{1+\kappa\theta t_i}}\right\} t_i \quad + \frac{1}{\theta} =$$

$$= \left\{-\frac{1}{1+\kappa\theta t_i} - \frac{\kappa}{1+\kappa\theta t_i}\right\} t_i \quad + \frac{1}{\theta} = \left\{-\frac{1+\kappa}{1+\kappa\theta t_i}\right\} t_i \quad + \frac{1}{\theta}$$

Άρα, ο πληροφοριακός αριθμός στην (4.19) ορίζεται ως

$$\mathrm{I}_1(\theta_0) = \int_0^\infty \left[\left(\left\{-\frac{(1+\kappa)t_i}{1+\kappa\theta t_i}\right\} + \frac{1}{\theta}\right)^2 \cdot e^{-\frac{1}{\kappa}\ln(1+\kappa\theta t_i)} \frac{\theta}{(1+\kappa\theta t_i)}\right]|_{\theta=\theta_0}\, dt_i \qquad (4.20)$$

4.1.6.2 Πληροφοριακός αριθμός του Fisher $\mathbf{I}_F(\theta_0)$ για το Inverse Gaussian μοντέλο ευπάθειας

Για την Inverse Gaussian κατανομή, δηλαδή $Z \sim InvGaussian(1,2b)$ έτσι ώστε $E(Z) = 1$ $\kappa\alpha\iota$ $Var(Z) = \frac{1}{2b}$, έχουμε ότι:

$$G(w) \quad = -2b + \sqrt{4b(b+w)}$$

και

$$G'(w) = \frac{\sqrt{b}}{\sqrt{b+w}}$$

και

$$G''(w) = -\frac{\sqrt{b}}{2(b+w)^{3/2}}$$

Άρα,

$$G'(w)|_{w=H_0(t_i,\theta)} = \frac{\sqrt{b}}{\sqrt{b+H_0(t_i,\theta)}}$$

Επίσης,

$$G''(w)|_{w=H_0(t_i,\theta)} = -\frac{\sqrt{b}}{2(b+H_0(t_i,\theta))^{3/2}}$$

Συνεπώς για την Inverse Gaussian κατανομή η (4.17) γίνεται,

$$B(t_i,\theta) = -\frac{\sqrt{b}\ t_i}{\sqrt{(b+\ \theta\, t_i)}} - \frac{\frac{\sqrt{b}\ t_i}{2\left((b+\theta t_i)\right)^{\frac{3}{2}}}}{\frac{\sqrt{b}}{\sqrt{(b+\ \theta\, t_i)}}} + \frac{1}{\theta} =$$

$$= -\frac{\sqrt{b}\ t_i}{\sqrt{(b+\ \theta\, t_i)}} - \frac{t_i}{2(b+\ \theta\, t_i)} + \frac{1}{\theta}$$

Άρα, ο πληροφοριακός αριθμός στην (4.19) ορίζεται ως

$$\mathbf{I}_1(\theta_0) = \int_0^{\infty} \left[\left(-\frac{\sqrt{b}\ t_i}{\sqrt{(b+\ \theta\, t_i)}} - \frac{t_i}{2(b+\ \theta\, t_i)} + \frac{1}{\theta}\right)^2 \cdot e^{2b-\sqrt{4b(b+\theta t_i)}}\ \frac{\sqrt{b}\theta}{\sqrt{b+\theta t_i}}\right]|_{\theta=\theta_0}\, dt_i$$

(4.21)

Στην συνέχεια για τον υπολογισμό του πληροφοριακού αριθμού του Fisher $I_F(\theta_0)$, για το διακριτό μοντέλο θα βασιστούμε στη συνάρτηση πιθανοφάνειας του διακριτού μοντέλου όπως είδαμε στο κεφάλαιο 3.6. Είδαμε ότι η εξίσωση (3.7) μας δίνει τη συνάρτηση πιθανοφάνειας:

$$L = Pr_\theta(N_1 = n_1, \ldots, N_M = n_M) = \frac{n!}{n_1! \cdots n_M!}(p_1(\theta)^{n_1} \cdots p_M(\theta)^{n_M})$$

Έτσι λογαριθμίζοντας έχουμε:

$$logL = \log\left(\frac{n!}{n_1!\cdots n_M!}\right) + \sum_{i=1}^{M} n_i \cdot \log(p_i(\theta)) \qquad (4.22)$$

Λαμβάνοντας υπόψιν την σχέση (4.8) βρίσκουμε ότι οι πιθανότητες p_i^0 κάτω από την μηδενική υπόθεση βρίσκονται από:

$$p_i^0(\theta) = P\left(t'_{\alpha_i} \le T \le t'_{b_i}\right) = P\left(T\epsilon[\,\alpha_i\,, b_i)\right) = S(\alpha_i) - S(b_i) =$$

$$= e^{-G\left(H_0(a_i,\theta)\right)} - e^{-G(H_0(b_i,\theta))}$$

όπου α_i και b_i γνωστά σημεία της διαμέρισης του $[0,\infty)$ με $\alpha_0 = 0$.

Έτσι ο λογάριθμος της πιθανοφάνειας παίρνει την παρακάτω μορφή:

$$logL = \log\left(\frac{n!}{n_1! \cdots n_M!}\right) + \sum_{i=1}^{M} n_i \cdot \log(e^{-G\left(H_0(a_i,\theta)\right)} - e^{-G(H_0(b_i,\theta))})$$

Για τον πληροφοριακό αριθμό του Fisher, στο τρίτο κεφάλαιο, είδαμε ότι ισχύει η σχέση (3.11), για πιθανοφάνεια δυο φορές παραγωγίσιμη ως προς θ:

$$I(\theta^*) = E_{\theta^*}(-[\frac{\partial^2}{\partial\theta^2} \log L(\mathrm{Y}|\theta)]_{\theta=\theta^*})$$

Έτσι, για το διακριτό μοντέλο μας και τη μηδενική υπόθεση ισχύει:

$$I_F(\theta_0) = E_{\theta_0}(-[\frac{\partial^2}{\partial\theta^2} \log L(Y|\theta)]_{\theta=\theta_0})$$

Παραγωγίζοντας τον λογάριθμο της συνάρτησης πιθανοφάνειας ως προς θ (4.16), έχουμε ότι:

$$\frac{\partial}{\partial\theta} \log L = \sum_{i=1}^{M} n_i \frac{-G'(w)|_{w=H_0(a_i,\theta)} H_0^{\cdot}(a_i,\theta) e^{-G\left(H_0(a_i,\theta)\right)} + G'(w)|_{w=H_0(b_i,\theta)} H_0^{\cdot}(b_i,\theta)\, e^{-G\left(H_0(b_i,\theta)\right)}}{e^{-G\left(H_0(a_i,\theta)\right)} - e^{-G(H_0(b_i,\theta))}} \Rightarrow$$

$$\frac{\partial}{\partial\theta} \log L \equiv \sum_{i=1}^{M} n_i\, \Delta_i\,(a_i, b_i, \theta) \qquad (4.23)$$

Κάνοντας το συμβολισμό $g(a_i,\theta) = -G'\left(H_0(a_i,\theta)\right) H_0^{\cdot}(a_i,\theta)\, e^{-G\left(H_0(a_i,\theta)\right)}$,

έχουμε

$$\sum_{i=1}^{M} n_i \frac{g(a_i,\theta) - g(b_i,\theta)}{p_i(\theta)} \qquad (4.24)$$

Η δεύτερη παράγωγος της συνάρτησης πιθανοφάνειας ισούται με

$$\frac{\partial}{\partial\theta}\left[\sum_{i=1}^{M} n_i\, \Delta_i\,(a_i, b_i, \theta)\right] = \sum_{i=1}^{M} n_i\ \frac{\partial}{\partial\theta}\, \Delta_i\,(a_i, b_i, \theta) =$$

$$\sum_{i=1}^{M} n_i \left[\frac{\left(\frac{\partial}{\partial\theta} g(a_i,\theta) - \frac{\partial}{\partial\theta} g(b_i,\theta)\right) p_i(\theta)}{p_i(\theta)^2} - \frac{(g(a_i,\theta) - g(b_i,\theta))\frac{\partial}{\partial\theta} p_i(\theta)}{p_i(\theta)^2}\right] =$$

$$\sum_{i=1}^{M} n_i \left[\frac{\left(\mathcal{K}(a_i,\theta) - \mathcal{K}(b_i,\theta)\right)p_i(\theta)}{p_i(\theta)^2} - \frac{(g(a_i,\theta) - g(b_i,\theta))(g(a_i,\theta) - g(b_i,\theta))}{p_i(\theta)^2}\right] =$$

$$\sum_{i=1}^{M} n_i \left[\frac{\left(\mathcal{K}(a_i,\theta) - \mathcal{K}(b_i,\theta)\right)}{p_i(\theta)} - \frac{(g(a_i,\theta) - g(b_i,\theta))^2}{p_i(\theta)^2}\right] \qquad (4.25)$$

όπου $\mathcal{K}(a_i,\theta) = \frac{\partial}{\partial\theta} g(a_i,\theta) = -G''(w)|_{w=H_0(a_i,\theta)} \left(H_0^{\cdot}(a_i,\theta)\right)^2 \; e^{-G\left(H_0(a_i,\theta)\right)}$

$$-G'(w)|_{w=H_0(a_i,\theta)} \; H_0^{\cdot\cdot}(a_i,\theta)\, e^{-G\left(H_0(a_i,\theta)\right)}$$

$$+G'^2(w)|_{w=H_0(a_i,\theta)} \left(H_0^{\cdot}(a_i,\theta)\right)^2 e^{-G\left(H_0(a_i,\theta)\right)} =$$

$$= -G'(H_0(a_i,\theta))H_0^{\cdot}(a_i,\theta)e^{-G\left(H_0(a_i,\theta)\right)} \left\{\left(+\frac{G''\left(H_0(a_i,\theta)\right)}{G'(H_0(a_i,\theta))} - G'\left(H_0(a_i,\theta)\right)\right) H_0^{\cdot}(a_i,\theta) + \frac{H_0^{\cdot\cdot}(a_i,\theta)}{H_0^{\cdot}(a_i,\theta)}\right\}$$

$$\equiv +g(a_i,\theta)BB(a_i,\theta)$$

με

$$BB(\alpha_i,\theta) = \left[-G'(w)|_{w=H_0(a_i,\theta)} + \frac{G''(w)|_{w=H_0(a_i,\theta)}}{G'(w)|_{w=H_0(a_i,\theta)}}\right] \cdot H_0^{\cdot}(a_i,\theta) + \frac{H_0^{\cdot\cdot}(a_i,\theta)}{H_0^{\cdot}(a_i,\theta)}$$

Συνεπώς η (4.25) γράφεται:

$$\sum_{i=1}^{M} n_i \left[\frac{\left(g(a_i,\theta)BB(a_i,\theta) - g(b_i,\theta)BB(b_i,\theta)\right)}{p_i(\theta)} - \frac{(g(a_i,\theta) - g(b_i,\theta))^2}{p_i(\theta)^2}\right]$$

Ο πληροφοριακός αριθμός για το διακριτό μοντέλο ορίζεται ως

$$I_F(\theta_0) \quad = E_0\left(-\sum_{i=1}^{M} n_i \; \frac{\partial}{\partial\theta} \; \Delta_i\,(a_i,b_i,\theta)|_{\theta_0}\right)$$

$$= -\sum_{i=1}^{M} E_0(n_i) \; \frac{\partial}{\partial\theta} \; \Delta_i\,(a_i,b_i,\theta)|_{\theta_0}$$

$$= -\sum_{i=1}^{M} n\ p_i^0(\theta)\frac{\partial}{\partial\theta}\ \Delta_i\ (a_i, b_i, \theta)|_{\theta_0}$$

$$= n\sum_{i=1}^{M} - p_i(\theta_0)\left\{\frac{g(a_i,\theta)}{p_i(\theta)}BB(a_i,\theta) - \frac{g(b_i,\theta)}{p_i(\theta)}BB(b_i,\theta)\right.$$

$$\left. - \frac{(g(a_i,\theta) - g(b_i,\theta))^2}{p_i(\theta)^2}\right\}|_{\theta_0} \qquad (4.26)$$

Για την περίπτωση της εκθετικής αθροιστικής συνάρτησης κινδύνου έχουμε ότι

$$\dot{H_0}(t,\theta) = t$$

οπότε η πιο πάνω παράγωγος του λογαρίθμου της συνάρτησης πιθανοφάνειας ως προς θ ορίζεται:

$$\sum_{i=1}^{M} n_i \frac{-G'(w)|_{w=a_i\theta}\ a_i\ \ e^{-G(a_i\theta)} + G'(w)|_{w=b_i\theta}\ b_i\ e^{-G(b_i\theta)}}{e^{-G(a_i\theta)} - e^{-G(b_i\theta)}}$$

Για την εκθετική κατανομή, η δεύτερη παράγωγος απλοποιείται λίγο αφού

$$BB(a_i,\theta) = \left(+\frac{G''(a_i\theta)}{G'(a_i\theta)} - G'(a_i\theta)\right) a_i$$

και $$g(a_i,\theta) = -G'(a_i\theta)\ a_i\ e^{-G(a_i\theta)}$$

4.1.6.3 Πληροφοριακός αριθμός του Fisher $I_F(\theta_0)$ για το Γάμμα μοντέλο ευπάθειας

Για το Γάμμα μοντέλο ευπάθειας έχουμε ότι

$$BB(a_i,\theta) = \left(-\frac{\kappa}{1+\kappa\theta a_i} - \frac{1}{(1+\theta a_i\kappa)}\right) a_i = -\frac{a_i\kappa + a_i}{1+\kappa\theta a_i}$$

και $$g(a_i,\theta) = -\frac{a_i}{(1+\theta a_i\kappa)}\, e^{-\frac{1}{\kappa}\ln(1+\theta a_i\kappa)}$$

$$p_i(\theta) = e^{-\frac{1}{\kappa}\ln(1+\theta a_i\kappa)} - e^{-\frac{1}{\kappa}\ln(1+\theta b_i\kappa)}$$

Άρα, ο πληροφοριακός αριθμός $I_F(\theta_0)$ για το Γάμμα μοντέλο ορίζεται ως

$$n\sum_{i=1}^{M} -\{\frac{{a_i}^2(\kappa+1)}{(1+\theta a_i\kappa)^2}\, e^{-\frac{1}{\kappa}\ln(1+\theta a_i\kappa)} - \frac{{b_i}^2(\kappa+1)}{(1+\theta b_i\kappa)^2}\, e^{-\frac{1}{\kappa}\ln(1+\theta b_i\kappa)} -$$

$$\frac{(\frac{b_i}{(1+\theta b_i\kappa)}\, e^{-\frac{1}{\kappa}\ln(1+\theta b_i\kappa)} - \frac{a_i}{(1+\theta a_i\kappa)}\, e^{-\frac{1}{\kappa}\ln(1+\theta a_i\kappa)})^2}{e^{-\frac{1}{\kappa}\ln(1+\theta a_i\kappa)} - e^{-\frac{1}{\kappa}\ln(1+\theta b_i\kappa)}}\}|_{\theta_0} \qquad (4.27)$$

4.1.6.4 Πληροφοριακός αριθμός του Fisher $I_F(\theta_0)$ για το Inverse Gaussian μοντέλο ευπάθειας

Για το Inverse Gaussian μοντέλο ευπάθειας έχουμε ότι:

$$BB(a_i,\theta) = -\frac{a_i}{2(b+a_i\theta)} - \frac{a_i\sqrt{b}}{\sqrt{b+a_i\theta}}$$

και $$g(a_i,\theta) = -\frac{\sqrt{b}}{\sqrt{b+a_i\theta}}\, a_i\, e^{2b-\sqrt{4b(b+a_i\theta)}},$$

$$p_i(\theta) = e^{2b-\sqrt{4b(b+a_i\theta)}} - e^{2b-\sqrt{4b(b+b_i\theta)}}$$

Ο πληροφοριακός αριθμός $I_F(\theta_0)$ για το το Inverse Gaussian μοντέλο ευπάθειας ορίζεται ως:

$$n\sum_{i=1}^{M}-\left\{\begin{array}{c}\frac{\sqrt{b}\,a_i\,e^{2b-\sqrt{4b(b+a_i\theta)}}}{\sqrt{b+a_i\theta}}\left(\frac{\sqrt{b}\;a_i}{\sqrt{(b+\;\theta\,a_i)}}+\frac{a_i}{2(b+\;\theta\,a_i)}\right)-\\ \frac{\sqrt{b}\,a_i\,e^{2b-\sqrt{4b(b+b_i\theta)}}}{\sqrt{b+b_i\theta}}\left(\frac{\sqrt{b}\;b_i}{\sqrt{(b+\;\theta\,b_i)}}+\frac{b_i}{2(b+\;\theta\,b_i)}\right)-\\ \frac{(\frac{\sqrt{b}}{\sqrt{b+b_i\theta}}\;a_i\,e^{2b-\sqrt{4b(b+b_i\theta)}}-\frac{\sqrt{b}}{\sqrt{b+a_i\theta}}\;a_i\,e^{2b-\sqrt{4b(b+a_i\theta)}})^2}{e^{2b-\sqrt{4b(b+a_i\theta)}}-e^{2b-\sqrt{4b(b+b_i\theta)}}}\end{array}\right\}|_{\theta_0}$$

$$\Rightarrow$$

$$n\sum_{i=1}^{M}\left\{\begin{array}{c}\left(\frac{b\cdot\;{b_i}^2e^{2b-\sqrt{4b(b+b_i\theta)}}}{b+b_i\theta}+\frac{\sqrt{b}\;{b_i}^2e^{2b-\sqrt{4b(b+b_i\theta)}}}{2(b+\;\theta\,b_i)^{3/2}}\right)-\\ \left(\frac{b\cdot\;{a_i}^2\cdot e^{2b-\sqrt{4b(b+a_i\theta)}}}{b+a_i\theta}+\frac{\sqrt{b}\cdot\;{a_i}^2\cdot e^{2b-\sqrt{4b(b+a_i\theta)}}}{2(b+\;\theta\,a_i)^{3/2}}\right)+\\ \frac{(\frac{\sqrt{b}}{\sqrt{b+b_i\theta}}\;a_i\;\cdot e^{2b-\sqrt{4b(b+b_i\theta)}}-\frac{\sqrt{b}}{\sqrt{b+a_i\theta}}\;a_i\;\cdot e^{2b-\sqrt{4b(b+a_i\theta)}})^2}{e^{2b-\sqrt{4b(b+a_i\theta)}}-e^{2b-\sqrt{4b(b+b_i\theta)}}}\end{array}\right\}|_{\theta_0}$$

(4.28)

4.2 Προσομοιώσεις

Στόχος της διπλωματικής εργασίας είναι να ελέγξουμε τη συμπεριφορά του προτεινόμενου ελέγχου καλής προσαρμογής, που βασίζεται στην ελεγχοσυνάρτηση T_n^{φ}, για την περίπτωση των μοντέλων ευπάθειας, εξετάζοντας το μέγεθος και την ισχύ του ελέγχου για επίπεδο σημαντικότητας 5%. Η ασυμπτωτική κατανομή της

ελεγχοσυνάρτησης δίνεται από τις σχέσεις (4.13) και (4.14). Το κρίσιμο σημείο για α=5% θα βρεθεί από αυτήν την κατανομή.

Στην συνέχεια θα δώσουμε τις λεπτομέρειες των προσομοιώσεών μας, οι οποίες πραγματοποιήθηκαν μέσω του στατιστικού πακέτου R studio και θα αναφέρουμε τις τιμές των διαφόρων παραμέτρων που χρησιμοποιήσαμε και στηριχτήκαμε στην έρευνα αυτή.

Για αρχή αναφέρουμε ότι προσομοιώσαμε δεδομένα για διάφορα μεγέθη δείγματος και συγκεκριμένα για $n = 60, 120, 240$ έτσι ώστε να εξετάσουμε μικρά, μεσαία και μεγάλα δείγματα. Μέσω της εντολής $runif(n)$, για την ομοιόμορφη κατανομή U(0,1) δημιουργήσαμε πιθανότητες επιβίωσης $S(t_i), i = 1, \dots, n$ για διάφορα μεγέθη δειγμάτων n. Τελικός μας σκοπός είναι να δημιουργήσουμε n χρόνους επιβίωσης οι οποίοι θα ακολουθούν το Γάμμα ή Inverse Gaussian μοντέλο ευπάθειας.

4.2.1. Γάμμα μοντέλο ευπάθειας

Λύνοντας λοιπόν ως προς τον χρόνο t, στον ορισμό του μοντέλου (4.5), βρίσκουμε για το Γάμμα μοντέλο ευπάθειας ότι:

$$t = \frac{S(t)^{-\kappa} - 1}{\theta \cdot \kappa}$$

Έτσι για κάθε πιθανότητα επιβίωσης $S(t_i)$ μιας μονάδας του δείγματος που ακολουθεί το Γάμμα μοντέλο ευπάθειας, βρίσκουμε τον αντίστοιχο χρόνο επιβίωσης t_i, με $i = 1,2, \dots, n$.

Θα θεωρήσουμε στη συνέχεια ότι η βασική συνάρτηση κινδύνου δίνεται από την εκθετική κατανομή $Exp(1)$, δηλαδή η πραγματική τιμή της παραμέτρου $\theta = 1$. Επίσης υποθέτουμε ότι η ευπάθεια ακολουθεί Γάμμα κατανομή $\Gamma(\alpha, \kappa)$ με $\alpha = \frac{1}{\kappa}$ για διάφορες τιμές του $\kappa = 0.25, 0.5, 1$. Εφόσον η παράμετρος κ είναι η διασπορά της ευπάθειας η επιλογή αυτών των τιμών έγινε για να εξετάσουμε την περίπτωση μικρής

αλλά και μεγάλης διασποράς. Η παράμετρος κ θα θεωρηθεί γνωστή σε αυτή την εργασία ενώ η παράμετρος θ θα θεωρηθεί άγνωστη.

Στη συνέχεια ομαδοποιούμε τους χρόνους επιβίωσης που δημιουργήσαμε σε M διαστήματα, όπως συζητήθηκε στην παράγραφο 4.1.3. Ο αριθμός των διαστημάτων M που εξετάσαμε ήταν αρχικά $M = 4, 6, 8, 10$ αλλά κατά τη διάρκεια των προσομοιώσεων παρατηρήσαμε ότι η τιμή του M εξαρτάται απο το μέγεθος του δείγματος n οπότε εξετάστηκαν και μεγαλύτερα M κατά περίπτωση για τη βελτιστοποίηση του ελέγχου.

4.2.2. Inverse Gaussian μοντέλο ευπάθειας

Όσον αφορά το μοντέλο με ενσωματωμένη Inverse Gaussian μοντέλο ευπάθειας $Gaussian(\mu, \lambda)$, με $\mu = 1$ $και$ $\lambda = 2b$, για να βρούμε τους χρόνους επιβίωσης των παρατηρήσεων που προσομοιώσαμε χρειάζεται να λύσουμε την εξίσωση (4.7) ως προς t :

$$t = \frac{lnS(t) \cdot (lnS(t) - 4b\,)}{4 \cdot \mathrm{b} \cdot \theta}$$

Κάτω από τις υποθέσεις για το μοντέλο ευπάθειας που έχουμε κάνει, θα ερευνήσουμε το μοντέλο αυτό για τις ίδιες τιμές της διασποράς ευπάθειας που είδαμε για το Γάμμα μοντέλο για σκοπούς σύγκρισης. Εφόσον λοιπόν ερευνούμε τις τρείς περιπτώσεις με διασποράς $0.25, 0.5, 1$

και αφού $Var(x) = \frac{1}{2b}$, οι τιμές του b που θα επιλέξουμε είναι 2, 1 και 0.5 αντίστοιχα.

4.2.3. Ελεγχος καλής προσαρμογής και για τα δύο μοντέλα

'Εχοντας δημιουργήσει τα δεδομένα μας, στη συνέχεια ομαδοποιούμε τους χρόνους επιβίωσης που δημιουργήσαμε σε M διαστήματα που περιέχουν το ίδιο πλήθος παρατηρήσεων, όπως συζητήθηκε στην παράγραφο 4.1.4.

Στην ελεγχοσυνάρτηση $T_n^{\varphi}(\hat{p}, p^0)$ χρειάζεται να βρούμε το διάνυσμα των πιθανοτήτων p^0 καθώς και το διάνυσμα των πιθανοτήτων $\hat{p}$ για να υπολογίσουμε την τιμή της ελεγχοσυνάρτησης.

Για να βρούμε το διάνυσμα των πιθανοτήτων $\hat{p}$, αυτό επιτυγχάνεται στην R με την χρήση της εντολής *time.freq* μετρώντας το πλήθος των παρατηρήσεων που βρίσκονται σε κάθε ένα από τα διαστήματα E_i και αποθηκεύοντας τα σε διάνυσμα N_i. Έτσι βρίσκουμε τις εκτιμήσεις των πιθανοτήτων p_i μέσω εκτιμητών μεγίστης πιθανοφάνειας που έχουμε ορίσει στο κεφάλαιο 3.4:

$$\hat{p}_i = N_i/n\,, \qquad \text{για } i = 1,\ \ldots, M$$

Καλούμαστε επίσης, και για τα δύο μοντέλα, να βρούμε τις πιθανότητες, μια τυχαία παρατήρηση να βρίσκεται στο i υποσύνολο (E_i), με $i = 1,2, \ldots, M$, για το διακριτό μοντέλο κάτω από την μηδενική υπόθεση, δηλαδή για το υποθετικό $(p_i^0, i = 1, .., M)$ μοντέλο.

Για να υπολογίσουμε τις πιθανότητες για το υποθετικό μοντέλο (p_j^0) θα χρειαστεί να χρησιμοποιήσουμε τους τύπους (4.5) και (4.7) για Γάμμα και για Inverse Gaussian μοντέλο ευπάθειας αντίστοιχα.

Για δεδομένη διαμέριση των χρόνων επιβίωσης στο δείγμα και δεδομένη παράμετρο διασποράς κ για το Gamma μοντέλο ή αντίστοιχα δεδομένη παράμετρο b για το Inverse Gaussian μοντέλο, χρειάζεται να εκτιμήσουμε την παράμετρο θ_0 για να υπολογίσουμε το διάνυσμα των πιθανοτήτων p_j^0. Αυτό επιτυγχάνεται μέσω της μεθόδου μέγιστης πιθανοφάνειας. Θυμίζουμε ότι συνάρτηση πιθανοφάνειας που χρησιμοποιήθηκε στην R έχει την εξής μορφή:

$$\mathcal{L} = \prod_{i=1}^{n} f(t_i, \theta) = \prod_{i=1}^{n} h_0(t_i, \theta) e^{-G(H_0(t_i,\theta))} G'(w)\big|_{w=H_0(t_i,\theta)}$$

με λογάριθμο:

$$\log(\mathcal{L}) = \sum_{i=1}^{n} [\, \log\big(h_0(t_i,\theta)\big) - G\big(H_0(t_i,\theta)\big) + \log\Big(G^{'}(w)|_{w=H_0(t_i,\theta)}\Big)]$$

Έτσι με την χρήση του αλγορίθμου nlm (Non-Linear Minimization) κάνουμε ελαχιστοποίηση της συνάρτησης $-\log(\mathcal{L})$ βρίσκοντας την εκτίμηση της παραμέτρου θ την οποία συμβολίζουμε ως $\hat{\theta}$.

Μπορούμε λοιπόν να υπολογίσουμε σε αυτό το σημείο την τιμή της ελεγχοσυνάρτησης $T_n^{\varphi}(\hat{p}, p^0)$. Ορίζουμε στην R την συνάρτηση $Tnkull$ με βάση τον τύπο (3.5):

$$T_n^{\varphi}(\hat{p}, p^0) = \frac{2n}{\varphi''(1)} \sum_{i=1}^{M} p_i^0 \varphi(\frac{\hat{p}_i}{p_i^0})$$

και για φ παίρνουμε αυτή της Kullback-Leibler. Έτσι βρίσκουμε την τιμή της $T_n^{\varphi_{Kull}}(\hat{p}, p^0)$ με βάση το δείγμα των χρόνων επιβίωσης που προσομοιώσαμε.

Σκοπός μας είναι να εξετάσουμε το μέγεθος και την ισχύ του ελέγχου σε επίπεδο σημαντικότητας 5%. Για να επιτευχθεί αυτό θα χρειαστεί να τρέξουμε τον παραπάνω αλγόριθμο αρκετές φορές, έτσι ώστε να βρούμε την εμπειρική κατανομή της ελεγχοσυνάρτησης $T_n^{\varphi}(\hat{p}, \hat{p}^0)$ και να μπορούμε να την συγκρίνουμε με την ασυμπτωτική κατανομή $X_{M-2}^2 + (1-\lambda)Z^2$ κυρίως μέσω των 95ου ποσοστημορίων τους.

Στο πρόγραμμα, επαναλαμβάνουμε τον παραπάνω αλγόριθμο 10000 φορές, παίρνοντας 10000 τιμές για την ελεγχοσυνάρτηση.

Έτσι είμαστε σε θέση να εκτιμήσουμε, με βάση τα δείγματα αυτά, την κατανομή της ελεγχοσυνάρτησης και συγκεκριμένα να βρούμε το 95ο ποσοστημόριό της, καθώς και την τιμή για το 95ο εμπειρικό ποσοστημόριο της εμπειρικής κατανομής της ελεγχοσυνάρτησης. Το ποσοστημόριο p_a είναι το σημείο της κατανομής για το οποίο

το α% των παρατηρήσεων είναι μικρότερες ή ίσες από αυτό και το υπόλοιπο (1-α)% των παρατηρήσεων είναι μεγαλύτερες ή ίσες από αυτό.

Το 95ο εμπειρικό ποσοστημόριο μπορεί να συγκριθεί με την τιμή p_{95} που θα βρούμε για το 95ο ποσοστημόριο της ασυμπτωτικής κατανομής της ελεγχοσυνάρτησης. Το μέγεθος του ελέγχου είναι ο αριθμός των φορών, στις 10000, που η τιμή της ελεγχοσυνάρτησης υπερβαίνει το κρίσιμο σημείο της ασυμπτωτικής κατανομής για δεδομένο α=5%, δηλαδή το σημείο p_{95}, όταν τα δεδομένα μας προέρχονται πράγματι από την μηδενική κατανομή. Θα θέλαμε σε αυτή την περίπτωση η πιθανότητα απόρριψης της μηδενικής υπόθεσης να μην υπερβαίνει το 5%.

Η ισχύς του ελέγχου είναι ο αριθμός των φορών, στις 10000, που η τιμή της ελεγχοσυνάρτησης υπερβαίνει το κρίσιμο σημείο της ασυμπτωτικής κατανομής για δεδομένο α=5%, δηλαδή το σημείο p_{95}, όταν τα δεδομένα μας δεν προέρχονται από την μηδενική κατανομή. Θα θέλαμε σε αυτή την περίπτωση η πιθανότητα απόρριψης της μηδενικής υπόθεσης να είναι μεγάλη και να πλησιάζει ιδανικά το 1.

4.2.4. Ασυμπτωτική κατανομή της ελεγχοσυνάρτησης

Στην συνέχεια, για να βρούμε την ασυμπτωτική κατανομή της ελεγχοσυνάρτησης για τις διάφορες περιπτώσεις, θα χρειαστεί να υπολογίσουμε τον πληροφοριακό αριθμό του Fisher, για το πραγματικό-αρχικό μοντέλο $\mathbf{I}_F(\theta)$ και για το διακριτό μοντέλο $I_F(\theta)$.

Σε μοντέλο με κατανομή ευπάθειας Γάμμα, ο πληροφοριακός αριθμός του Fisher $I_F(\theta_0)$ δίνεται από τον τύπο (4.27) που είδαμε στο κεφάλαιο 4.1., ενώ ο πληροφοριακός αριθμός $\mathbf{I}_F(\theta_0)$ δίνεται από τον τύπο $\mathbf{I}_F(\theta_0) = n\mathbf{I}_1(\theta_0)$, με $\mathbf{I}_1(\theta_0)$ όπως ορίστηκε στην εξίσωση (4.20).

Αντίστοιχα σε μοντέλο με κατανομή ευπάθειας Inverse Gaussian, ο πληροφοριακός αριθμός του Fisher $I_F(\theta_0)$ δίνεται από τον τύπο (4.28), ενώ ο πληροφοριακός αριθμός $\mathbf{I}_F(\theta_0)$ δίνεται από τον τύπο $\mathbf{I}_F(\theta_0) = n\mathbf{I}_1(\theta_0)$, με $\mathbf{I}_1(\theta_0)$ όπως ορίστηκε στην εξίσωση (4.21).

Η τιμή του λ που χρειάζεται να βρεθεί για την ασυμπτωτική κατανομή της ελεγχοσυνάρτησης που ορίζεται στις (4.13) και (4.14), όπου $0 \leq \lambda \leq 1$, είναι η λύση της εξίσωσης:

$$\mathbf{I}_F(\theta_0) - \lambda \cdot I_F(\theta_0) = 0 \rightarrow \lambda = \frac{\mathbf{I}_F(\theta_0)}{I_F(\theta_0)}.$$

Για να βρούμε την τιμή του ποσοστημορίου p_{95} θα χρειαστεί η συνάρτηση κατανομής της ασυμπτωτικής κατανομής. Την κατανομή αυτή θα την υπολογίσουμε με την βοήθεια της εργασίας Moschopoulos (1985), ο οποίος ερεύνησε την κατανομή αθροίσματος πολλών Γάμμα ανεξάρτητων και τυχαίων μεταβλητών.

Μια κατανομή chi-square $X \sim \chi^2_\kappa$ είναι ειδική περίπτωση της οικογένειας κατανομών Γάμμα. Συγκεκριμένα:

$$\chi^2_\kappa \sim \Gamma\left(\frac{\kappa}{2}, 2\right),$$

με $\Gamma\left(\frac{\kappa}{2}, 2\right)$ να συμβολίζει κατανομή Γάμμα με παραμέτρους shape $\frac{\kappa}{2}$ και scale 2 αντίστοιχα.

Επίσης από ιδιότητες της Γάμμα κατανομής ισχύει ότι για $c > 0$:

$$c \cdot \Gamma(\kappa: shape, \theta: scale) \sim \Gamma(k,\ c \cdot \theta).$$

Έτσι η ασυμπτωτική κατανομή του μοντέλου μας, μπορεί να γραφτεί στη μορφή αθροίσματος δύο Γάμμα κατανεμημένων ανεξάρτητων τυχαίων μεταβητών :

$$X^2_{M-2} + (1-\lambda)Z^2 \quad \sim \quad \Gamma\left(\frac{M-2}{2}, 2\right) \quad + \quad \Gamma\left(\frac{1}{2}, (1-\lambda)\cdot 2\right)$$

καθώς η τ.μ. Z^2 $(Z \sim N(0,1))$, ακολουθεί Chi-square κατανομή με βαθμό ελευθερίας 1 (X^2_1).

Με βάση τη θεωρία που αναπτύξαμε στην παράγραφο (3.8) και συγκεκριμένα με βάση τις σχέσεις (3.14)-(3.18), γράφτηκε κώδικας στην R, με τις παραμέτρους ως εξής:

$$(\alpha_1, \alpha_2) = \left(\frac{1}{2}\ ,\ \frac{M-2}{2}\right)$$

$$(\beta_1, \beta_2) = \left((1-\lambda)\cdot 2\ , 2\right)$$

και ως λ χρησιμοποιήσαμε την μέση τιμή των λ_i που βρήκαμε (για κάθε τιμή της ελεγχοσυνάρτησης T_n που προσομοιώσαμε βρήκαμε ένα λ).

Για δεδομένο M και λ, λοιπόν, βρίσκουμε από την συνάρτηση κατανομής της ασυμπτωτικής κατανομής $\boldsymbol{F}$, την τιμή του ποσοστημορίου p_{95}, δηλαδή

$$\boldsymbol{F}(p_{95}) = \Pr(Y \leq p_{95}) = 0.95$$

Για να υπολογίσουμε την τιμή του εμπειρικού ποσοστημορίου $\boldsymbol{p_{95}}$ ταξινομούμε τις τιμές της ελεγχοσυνάρτησης που έχουμε δημιουργήσει μέσω των προσομοιώσεών μας μας και αποθηκεύουμε την τιμή του $95ου$ ποσοστημορίου. Με αυτόν τον τρόπο έχουμε μια εικόνα για το εάν η εμπειρική κατανομή και η ασυμπτωτική κατανομή της ελεγχοσυνάρτησης είναι κοντά, ελέγχοντας αν οι τιμές p_{95} και $\boldsymbol{p_{95}}$ είναι κοντά.

Στην συνέχεια παραθέτουμε ορισμένους πίνακες με τα αποτελέσματα (σφάλμα τύπου Ι και ισχύς) που λάβαμε από το στατιστικό πακέτο R studio για διάφορες τιμές της μεταβλητής κ και b (σχετίζεται με τη διασπορά της ευπάθειας) , M (αριθμός διαστημάτων που λαμβάνεται στη διαμέριση) και n (αριθμός παρατηρήσεων για κάθε προσομοίωση).

4.3. Πίνακες Προσομοιώσεων - Μέγεθος και Ισχύς του ελέγχου

Στους παρακάτω πίνακες, στη στήλη $''P_t''$ εμφανίζεται το ποσοστό (για κάθε έλεγχο) όπου ο αριθμός των τιμών της ελεγχοσυνάρτησης υπερβαίνει το κρίσιμο σημείο της ασυμπτωτικής κατανομής (p_{95}). Για κάθε περίπτωση τιμών πραγματικών παραμέτρων, κάνουμε 15 ελέγχους, δηλαδή επαναλαμβάνουμε τη διαδικασία προσομοιώσεων 15 φορές, λαμβάνοντας 15 τιμές P_t.

Ως $''mean(\lambda)''$ δηλώνεται η μέση τιμή από τα 10000 λ που έχουν υπολογιστεί. Στην γραμμή ''Mean $(\hat{\theta})$'' δηλώνεται η μέση τιμή των $\hat{\theta}_j$, $j = 1,2, \dots ,15$, όπου $\hat{\theta}_j$ είναι η μέση τιμή των 10000 εκτιμήσεων θ πού έχουμε βρει, για κάθε προσομοίωση j από τις 15.

Στην γραμμή $''p_{95}''$ σημειώνεται το κρίσιμο σημείο της ασυμπτωτικής κατανομής για κάθε M που επιλέξαμε και στην γραμμή $''\boldsymbol{p_{95}}''$ η μέση τιμή, από το σύνολο των 15 τιμών που βρίσκουμε για κάθε προσομοίωση, του $95^{ου}$ ποσοστημορίου της ελεγχοσυνάρτησης .

Όσον αφορά την ισχύ του μοντέλου, θα στηριχθούμε στην ίδια μηδενική υπόθεση που στηριχθήκαμε και ανωτέρω $H_0 : p = p^0$ με την μεταβλητή ευπάθειας να κατανέμεται ως Γάμμα με διασπορά $\kappa = \kappa_1$ ($1/2b_1$ για Inverse Gaussian κατανομή ευπάθειας), ενώ οι χρόνοι που θα προσομοιώσουμε θα κατανέμονται ως Γάμμα με άλλη παράμετρο κ_2 (ή Inverse Gaussian με άλλη παράμετρο b_2) κάτω από την εναλλακτική υπόθεση η οποία είναι αληθινή.

<u>Κατανομή ευπάθειας: Γάμμα, Σφάλμα τύπου I</u>

- Για : $n = 60$, $\kappa = 0.25$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0289	0.0798	0.0383	0.0873	0.0469	0.0895	0.0539	0.0912
2	0.0342	0.0801	0.0431	0.0874	0.0452	0.0895	0.0545	0.0914
3	0.0322	0.0801	0.0429	0.0867	0.0490	0.0900	0.0515	0.0916
4	0.0335	0.0803	0.0393	0.0875	0.0461	0.0901	0.0518	0.0913
5	0.0362	0.0802	0.0435	0.0868	0.0461	0.0892	0.0518	0.0918
6	0.0318	0.0801	0.0430	0.0873	0.0477	0.0898	0.0542	0.0917
7	0.0348	0.0806	0.0377	0.0872	0.0473	0.0896	0.0515	0.0913
8	0.0375	0.0804	0.0403	0.0874	0.0454	0.0899	0.0519	0.0912
9	0.0285	0.0802	0.0410	0.0870	0.0446	0.0897	0.0563	0.0909
10	0.0311	0.0802	0.0416	0.0865	0.0476	0.0894	0.0504	0.0916
11	0.0312	0.0798	0.0401	0.0872	0.0459	0.0895	0.0534	0.0918
12	0.0322	0.0799	0.0441	0.0875	0.0454	0.0898	0.0545	0.0914

13	0.0332	0.0798	0.0395	0.0871	0.0477	0.0901	0.0547	0.0913
14	0.0324	0.0800	0.0458	0.0876	0.0436	0.0897	0.0530	0.0916
15	0.0327	0.0797	0.0413	0.0870	0.0482	0.0898	0.0507	0.0918
mean	0.0327	0.0801	0.0414	0.0872	0.0464	0.0897	0.0529	0.0915

Mean ($\hat{\theta}$)	1.018942	1.016914	1.019562	1.016832
p_{95}	7.609707	10.88095	13.8905	16.75016
$\boldsymbol{p_{95}}$	6.716496	10.36527	13.68994	16.97493

Σχόλια

Για μέγεθος δείγματος $n = 60$ και διασπορά της ευπάθειας $\kappa = 0.25$ το μέγεθος του ελέγχου για M=8 ή 10 είναι περίπου 0.05 όπως θα θέλαμε και αποδεχόμαστε την μηδενική υπόθεση 95% περίπου των φορών. Φαίνεται ότι η κατανομή της ελεγχοσυνάρτησης T_n περιγράφεται καλά από την ασυμπτωτική κατανομή. Επειδή θέλουμε το ποσοστό όπου το πλήθος των τιμών της ελεγχοσυνάρτησης υπερβαίνει το κρίσιμο σημείο της ασυμπτωτικής κατανομής (p_{95}) να προσεγγίζει το 0.05, επιλέγουμε στον έλεγχο αυτό το μέγεθος M να ισούται με 10 με το κρίσιμο σημείο της ασυμπτωτικής κατανομής $''p_{95}''$ να ισούται προσεγγιστικά με 16.75 ενώ το 95ο ποσοστημόριο της ελεγχοσυνάρτησης T_n, $''\boldsymbol{p_{95}}''$, ισούται με 16.97 .

Για τον λόγο αυτό, για να ελέγξουμε την ισχύ του αλγορίθμου, θα επιλέξουμε το βέλτιστο μέγεθος M, δηλαδή για $M = 10$.

Ισχύς Ελέγχου ($\kappa_1 = 0.25\,, M = 10$)

Για $\kappa_2 = 0.5$: $P_t = 0.1191$

Για $\kappa_2 = 1$: $P_t = 0.5854$

Για $\kappa_2 = 1.5$: $P_t = 0.9216$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 120$, $\kappa = 0.25$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0295	0.0968	0.0337	0.105	0.0401	0.108	0.0408	0.109
2	0.0311	0.0968	0.0367	0.104	0.0387	0.108	0.0399	0.110
3	0.0292	0.0968	0.0368	0.105	0.0406	0.108	0.0435	0.110
4	0.0255	0.0968	0.0341	0.104	0.0371	0.108	0.0439	0.109
5	0.0284	0.0968	0.0361	0.105	0.0427	0.108	0.0406	0.109
6	0.0294	0.0969	0.0357	0.105	0.0398	0.108	0.0454	0.110
7	0.0300	0.0968	0.0386	0.105	0.0390	0.108	0.0413	0.110
8	0.0304	0.0968	0.0336	0.105	0.0388	0.108	0.0442	0.110
9	0.0284	0.0965	0.0351	0.105	0.0385	0.108	0.0389	0.109
10	0.0295	0.0967	0.0327	0.105	0.0381	0.108	0.0418	0.110
11	0.0289	0.0965	0.0341	0.105	0.0399	0.108	0.0437	0.110
12	0.0314	0.0968	0.0348	0.105	0.0358	0.108	0.0421	0.109
13	0.0285	0.0968	0.0343	0.105	0.0387	0.108	0.0416	0.109
14	0.0261	0.0966	0.0352	0.105	0.0378	0.108	0.0429	0.110
15	0.0276	0.0969	0.0349	0.104	0.0390	0.108	0.0424	0.110
mean	0.0289	0.0968	0.0351	0.105	0.039	0.108	0.0422	0.11

Mean ($\hat{\theta}$)	1.007231	1.0085	1.009972	1.010041
p_{95}	7.567228	10.84512	13.85573	16.71846
$\boldsymbol{p_{95}}$	6.441991	9.986232	13.15572	16.22043

Εδώ σημειώνουμε ότι με M=16, ο έλεγχος δίνει καλύτερο μέγεθος καθώς $P_t = 0.0493$.

Ισχύς Ελέγχου ($\kappa_1 = 0.25 , M = 16$)

Για $\kappa_2 = 0.5$: $P_t = 0.1637$

Για $\kappa_2 = 1$: $P_t = 0.8695$

Για $\kappa_2 = 1.5$: $P_t = 0.9989$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 240$, $\kappa = 0.25$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0275	0.107	0.0338	0.116	0.0373	0.119	0.0385	0.121
2	0.0255	0.107	0.0340	0.116	0.0357	0.119	0.0384	0.121
3	0.0267	0.107	0.0315	0.116	0.0379	0.119	0.0379	0.121
4	0.0237	0.107	0.0339	0.116	0.0378	0.119	0.0396	0.121
5	0.0245	0.107	0.0315	0.116	0.0382	0.119	0.0375	0.121
6	0.0259	0.107	0.0307	0.116	0.0363	0.119	0.0398	0.121
7	0.0262	0.107	0.0327	0.116	0.0382	0.119	0.0339	0.121
8	0.0248	0.107	0.0321	0.116	0.0388	0.119	0.0394	0.121
9	0.0276	0.107	0.0339	0.116	0.0358	0.119	0.0401	0.121
10	0.0270	0.107	0.0295	0.116	0.0378	0.120	0.0368	0.121
11	0.0274	0.107	0.0306	0.116	0.0369	0.119	0.0378	0.121
12	0.0272	0.107	0.0323	0.116	0.0328	0.119	0.0388	0.121
13	0.0287	0.107	0.0312	0.116	0.0357	0.119	0.0384	0.121
14	0.0274	0.107	0.0302	0.116	0.0366	0.119	0.0376	0.121
15	0.0277	0.107	0.0334	0.116	0.0350	0.119	0.0407	0.121
mean	0.0265	0.107	0.0321	0.116	0.0367	0.119	0.0383	0.121

Mean ($\hat{\theta}$)	1.004945	1.004643	1.005423	1.004179
p_{95}	7.540702	10.82125	13.83419	16.69769
$\boldsymbol{p_{95}}$	6.291468	9.755889	12.95283	15.88819

Εδώ σημειώνουμε ότι με M=14, ο έλεγχος δίνει καλύτερο μέγεθος καθώς $P_t = 0.0496$.

Ισχύς Ελέγχου ($\kappa_1 = 0.25\,, M = 14$)

Για $\kappa_2 = 0.5$: $P_t = 0.2629$

Για $\kappa_2 = 1$: $P_t = 0.9909$

Για $\kappa_2 = 1.5$: $P_t = 1$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 60$, $\kappa = 0.5$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0302	0.0744	0.0404	0.0790	0.0454	0.0806	0.0540	0.0814
2	0.0309	0.0746	0.0422	0.0790	0.0485	0.0804	0.0519	0.0814
3	0.0297	0.0745	0.0416	0.0787	0.0435	0.0803	0.0532	0.0816
4	0.0321	0.0741	0.0432	0.0786	0.0485	0.0805	0.0518	0.0812
5	0.0331	0.0741	0.0441	0.0788	0.0477	0.0802	0.0527	0.0812
6	0.0325	0.0745	0.0423	0.0788	0.0447	0.0805	0.0543	0.0812
7	0.0299	0.0743	0.0398	0.0788	0.0479	0.0804	0.0536	0.0814
8	0.0300	0.0742	0.0403	0.0788	0.0480	0.0802	0.0524	0.0812
9	0.0322	0.0743	0.0382	0.0785	0.0462	0.0807	0.0524	0.0811
10	0.0311	0.0743	0.0389	0.0785	0.0492	0.0805	0.0531	0.0814
11	0.0316	0.0746	0.0441	0.0787	0.0454	0.0805	0.0490	0.0811
12	0.0316	0.0740	0.0365	0.0786	0.0439	0.0804	0.0531	0.0814
13	0.0339	0.0746	0.0432	0.0786	0.0430	0.0800	0.0508	0.0809
14	0.0310	0.0746	0.0381	0.0788	0.0434	0.0803	0.0519	0.0812
15	0.0321	0.0745	0.0380	0.0785	0.0456	0.0798	0.0508	0.0814
mean	0.0315	0.0744	0.0407	0.0787	0.0461	0.0804	0.0523	0.0813

Mean ($\hat{\theta}$)	1.020615	1.021138	1.019278	1.019675
p_{95}	7.623517	10.89986	13.90992	16.76905
$\boldsymbol{p_{95}}$	6.676938	10.35678	13.66558	16.91297

Ισχύς Ελέγχου ($\kappa_1 = 0.5\,, M = 10$)

Για $\kappa_2 = 1$: $P_t = 0.2629$

Για $\kappa_2 = 1.5$: $P_t = 0.7184$

Για $\kappa_2 = 2$: $P_t = 0.9473$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 120$, $\kappa = 0.5$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0294	0.0858	0.0361	0.0908	0.0371	0.0925	0.0429	0.0936
2	0.0265	0.0861	0.0358	0.0908	0.0381	0.0928	0.0435	0.0935
3	0.0263	0.0860	0.0354	0.0906	0.0352	0.0928	0.0401	0.0934
4	0.0253	0.0859	0.0308	0.0909	0.0366	0.0927	0.0421	0.0935
5	0.0249	0.0858	0.0355	0.0909	0.0409	0.0926	0.0421	0.0936
6	0.0256	0.0859	0.0337	0.0908	0.0372	0.0928	0.0399	0.0935
7	0.0271	0.0859	0.0325	0.0908	0.0374	0.0925	0.0390	0.0935
8	0.0256	0.0859	0.0333	0.0908	0.0367	0.0926	0.0401	0.0937
9	0.0261	0.0859	0.0370	0.0911	0.0375	0.0924	0.0457	0.0934
10	0.0259	0.0860	0.0345	0.0906	0.0407	0.0926	0.0415	0.0934
11	0.0311	0.0859	0.0333	0.0905	0.0395	0.0927	0.0446	0.0935
12	0.0287	0.0858	0.0345	0.0908	0.0420	0.0926	0.0418	0.0936
13	0.0275	0.0858	0.0372	0.0908	0.0370	0.0925	0.0416	0.0936

14	0.0279	0.0860	0.0343	0.0906	0.0382	0.0926	0.0412	0.0935
15	0.0282	0.0859	0.0327	0.0908	0.0359	0.0924	0.0436	0.0935
mean	0.0271	0.0859	0.0344	0.0908	0.038	0.0926	0.042	0.0935

Mean ($\hat{\theta}$)	1.007778	1.009426	1.011997	1.012558
p_{95}	7.594871	10.8739	13.88574	16.74716
$\boldsymbol{p_{95}}$	6.335025	9.924608	13.10953	16.19015

Εδώ σημειώνουμε ότι με M=14, ο έλεγχος δίνει καλύτερο μέγεθος καθώς $P_t = 0.0482$.

Ισχύς Ελέγχου ($\kappa_1 = 0.5\,, M = 14$)

Για $\kappa_2 = 1$: $P_t = 0.4456$

Για $\kappa_2 = 1{,}5$: $P_t = 0.9479$

Για $\kappa_2 = 2$: $P_t = 0.9989$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 240$, $\kappa = 0.5$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0246	0.0927	0.0319	0.0978	0.0339	0.0997	0.0401	0.101
2	0.0277	0.0927	0.0312	0.0978	0.0360	0.0997	0.0409	0.101
3	0.0287	0.0927	0.0291	0.0978	0.0327	0.0997	0.0378	0.101
4	0.0259	0.0927	0.0316	0.0977	0.0353	0.0997	0.0392	0.101
5	0.0253	0.0926	0.0319	0.0978	0.0345	0.0998	0.0386	0.101
6	0.0238	0.0926	0.0272	0.0979	0.0316	0.0997	0.0359	0.101
7	0.0271	0.0926	0.0303	0.0978	0.0360	0.0996	0.0394	0.101
8	0.0248	0.0927	0.0299	0.0978	0.0348	0.0997	0.0376	0.101
9	0.0252	0.0926	0.0331	0.0979	0.0359	0.0997	0.0377	0.101

10	0.0257	0.0928	0.0305	0.0977	0.0330	0.0997	0.0375	0.101
11	0.0268	0.0926	0.0330	0.0979	0.0321	0.0996	0.0396	0.101
12	0.0285	0.0927	0.0322	0.0978	0.0347	0.0998	0.0411	0.101
13	0.0272	0.0928	0.0311	0.0977	0.0355	0.0996	0.0374	0.101
14	0.0247	0.0927	0.0328	0.0977	0.0339	0.0998	0.0347	0.101
15	0.0235	0.0927	0.0319	0.0976	0.0378	0.0998	0.0350	0.101
mean	0.026	0.0927	0.0312	0.0978	0.0345	0.0997	0.0382	0.101

Mean ($\hat{\theta}$)	1.005612	1.004671	1.005598	1.004357
p_{95}	7.577525	10.85941	13.87156	16.73452
$\boldsymbol{p_{95}}$	6.249861	9.740504	12.85659	15.91684

Εδώ σημειώνουμε ότι με M=14, ο έλεγχος δίνει καλύτερο μέγεθος καθώς $P_t = 0.0501$.

<u>Ισχύς Ελέγχου</u> $(\kappa_1 = 0.5\,, M = 14)$

Για $\kappa_2 = 1$: $P_t = 0.7311$

Για $\kappa_2 = 1.5$: $P_t = 0.9987$

Για $\kappa_2 = 2$: $P_t = 1$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 60$, $\kappa = 1$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0312	0.0584	0.0400	0.0607	0.0481	0.0618	0.0576	0.0621
2	0.0293	0.0583	0.0395	0.0609	0.0489	0.0618	0.0542	0.0621
3	0.0297	0.0584	0.0423	0.0609	0.0450	0.0618	0.0476	0.0622
4	0.0308	0.0583	0.0400	0.0610	0.0453	0.0618	0.0500	0.0623
5	0.0302	0.0585	0.0366	0.0609	0.0423	0.0619	0.0489	0.0622

6	0.0316	0.0583	0.0410	0.0609	0.0438	0.0617	0.0527	0.0622
7	0.0281	0.0583	0.0409	0.0607	0.0519	0.0618	0.0499	0.0622
8	0.0304	0.0583	0.0426	0.0610	0.0458	0.0617	0.0532	0.0622
9	0.0284	0.0583	0.0398	0.0607	0.0494	0.0618	0.0510	0.0620
10	0.0324	0.0583	0.0410	0.0608	0.0449	0.0618	0.0548	0.0624
11	0.0278	0.0583	0.0439	0.0608	0.0460	0.0618	0.0533	0.0621
12	0.0307	0.0584	0.0380	0.0608	0.0442	0.0617	0.0516	0.0621
13	0.0298	0.0584	0.0399	0.0609	0.0468	0.0617	0.0477	0.0622
14	0.0332	0.0583	0.0400	0.0607	0.0441	0.0618	0.0527	0.0622
15	0.0313	0.0585	0.0401	0.0606	0.0440	0.0618	0.0508	0.0624
mean	0.0303	0.0584	0.0404	0.0608	0.046	0.0618	0.0517	0.0622

Mean ($\hat{\theta}$)	1.023396	1.025466	1.026335	1.025898
p_{95}	7.664169	10.93815	13.94483	16.8034
$\boldsymbol{p_{95}}$	6.622192	10.39418	13.70016	16.90978

<u>Ισχύς Ελέγχου</u> $(\kappa_1 = 1\,, M = 10)$

Για $\kappa_2 = 0.5$: $P_t = 0.0656$

Για $\kappa_2 = 0.25$: $P_t = 0.1206$

Για $\kappa_2 = 0.1$: $P_t = 0.1886$

Για $\kappa_2 = 2$: $P_t = 0.5667$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 120$, $\kappa = 1$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0260	0.0648	0.0334	0.0675	0.0363	0.0684	0.0433	0.0688
2	0.0286	0.0649	0.0343	0.0675	0.0346	0.0684	0.0402	0.0689
3	0.0309	0.0649	0.0343	0.0674	0.0362	0.0684	0.0392	0.0688
4	0.0245	0.0648	0.0351	0.0675	0.0382	0.0683	0.0416	0.0688
5	0.0267	0.0649	0.0322	0.0674	0.0377	0.0684	0.0412	0.0688
6	0.0257	0.0647	0.0323	0.0675	0.0372	0.0683	0.0404	0.0688
7	0.0264	0.0648	0.0316	0.0674	0.0387	0.0683	0.0414	0.0687
8	0.0267	0.0648	0.0299	0.0673	0.0368	0.0684	0.0426	0.0688
9	0.0273	0.0648	0.0341	0.0674	0.0392	0.0685	0.0387	0.0688
10	0.0286	0.0648	0.0314	0.0675	0.0380	0.0682	0.0401	0.0688
11	0.0278	0.0649	0.0315	0.0674	0.0410	0.0685	0.0427	0.0689
12	0.0285	0.0649	0.0322	0.0675	0.0371	0.0683	0.0448	0.0688
13	0.0266	0.0649	0.0345	0.0674	0.0355	0.0683	0.0392	0.0687
14	0.0252	0.0649	0.0334	0.0674	0.0361	0.0684	0.0412	0.0688
15	0.0279	0.0647	0.0325	0.0675	0.0321	0.0684	0.0410	0.0687
mean	0.0272	0.0648	0.0328	0.0674	0.037	0.0684	0.0412	0.0688

Mean ($\hat{\theta}$)	1.013443	1.013654	1.014616	1.01233
p_{95}	7.648352	10.92341	13.93207	16.79197
$\boldsymbol{p_{95}}$	6.346559	9.863945	13.09772	16.19873

Εδώ σημειώνουμε ότι με M=14, ο έλεγχος δίνει καλύτερο μέγεθος καθώς $P_t = 0.0499$.

<u>Ισχύς Ελέγχου</u> ($\kappa_1 = 1$, $M = 14$)

Για $\kappa_2 = 1.5$: $P_t = 0.2957$

Για $\kappa_2 = 2$: $P_t = 0.8296$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 240$, $\kappa = 1$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0242	0.0684	0.0317	0.0710	0.0345	0.0720	0.0324	0.0724
2	0.0250	0.0683	0.0263	0.0711	0.0335	0.0719	0.0391	0.0724
3	0.0268	0.0684	0.0320	0.0710	0.0347	0.0719	0.0360	0.0724
4	0.0245	0.0684	0.0285	0.0710	0.0299	0.0720	0.0371	0.0724
5	0.0241	0.0684	0.0304	0.0710	0.0338	0.0719	0.0320	0.0724
6	0.0244	0.0684	0.0297	0.0710	0.0367	0.0720	0.0345	0.0724
7	0.0242	0.0684	0.0307	0.0710	0.0328	0.0720	0.0370	0.0723
8	0.0240	0.0684	0.0298	0.0710	0.0284	0.0719	0.0361	0.0723
9	0.0237	0.0684	0.0307	0.0710	0.0374	0.0719	0.0370	0.0724
10	0.0269	0.0684	0.0305	0.0710	0.0340	0.0719	0.0398	0.0723
11	0.0261	0.0684	0.0305	0.0710	0.0330	0.0719	0.0353	0.0724
12	0.0228	0.0683	0.0287	0.0709	0.0362	0.0719	0.0357	0.0724
13	0.0237	0.0684	0.0315	0.0710	0.0309	0.0719	0.0375	0.0724
14	0.0246	0.0684	0.0321	0.0710	0.0305	0.0720	0.0378	0.0724
15	0.0229	0.0684	0.0300	0.0710	0.0339	0.0719	0.0383	0.0724
mean	0.0245	0.0684	0.0302	0.071	0.0333	0.0719	0.0364	0.0724

Mean ($\hat{\theta}$)	1.005546	1.006097	1.007556	1.005052
p_{95}	7.638972	10.91602	13.92514	16.78528
$\boldsymbol{p_{95}}$	6.193746	9.689256	12.82308	15.86092

Εδώ σημειώνουμε ότι με M=14, ο έλεγχος δίνει καλύτερο μέγεθος μ καθώς $P_t = 0.0483$.

Ισχύς Ελέγχου ($\kappa_1 = 1 \ , M = 14$)

Για $\kappa_2 = 1.5$: $P_t = 0.521$

Για $\kappa_2 = 2$: $P_t = 0.9854$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

Κατανομή ευπάθειας: Inverse Gaussian, Σφάλμα τύπου I

- Για : $n = 60$, $b = 2$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0355	0.0746	0.0463	0.0821	0.0487	0.0859	0.0529	0.0873
2	0.0323	0.0742	0.0410	0.0822	0.0470	0.0855	0.0549	0.0879
3	0.0337	0.0738	0.0438	0.0818	0.0470	0.0860	0.0519	0.0878
4	0.0301	0.0740	0.0456	0.0821	0.0445	0.0854	0.0565	0.0880
5	0.0323	0.0741	0.0435	0.0821	0.0460	0.0849	0.0519	0.0879
6	0.0315	0.0745	0.0445	0.0823	0.0468	0.0859	0.0558	0.0878
7	0.0346	0.0740	0.0415	0.0817	0.0487	0.0854	0.0554	0.0872
8	0.0329	0.0744	0.0437	0.0824	0.0477	0.0853	0.0536	0.0880
9	0.0286	0.0740	0.0444	0.0823	0.0469	0.0852	0.0504	0.0875
10	0.0352	0.0746	0.0405	0.0822	0.0496	0.0849	0.0530	0.0871
11	0.0353	0.0744	0.0390	0.0820	0.0492	0.0847	0.0563	0.0877
12	0.0343	0.0738	0.0414	0.0827	0.0450	0.0858	0.0489	0.0878
13	0.0375	0.0746	0.0457	0.0831	0.0435	0.0851	0.0559	0.0874
14	0.0390	0.0743	0.0420	0.0823	0.0458	0.0857	0.0526	0.0878
15	0.0274	0.0743	0.0418	0.0822	0.0479	0.0853	0.0528	0.0882
mean	0.0333	0.0742	0.043	0.0822	0.047	0.0854	0.0535	0.0877

Mean ($\hat{\theta}$)	1.017678	1.018444	1.019505	1.019102
p_{95}	7.623935	10.89207	13.89937	16.75679
$\boldsymbol{p_{95}}$	6.791576	10.48712	13.72204	16.96949

Ισχύς Ελέγχου ($b_1 = 2 \ , M = 10$)

Για $b_2 = 1$: $P_t = 0.0767$

Για $b_2 = 0.5$: $P_t = 0.1628$

Για $b_2 = 0.33$: $P_t = 0.2824$

Για $b_2 = 0.25$: $P_t = 0.3800$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 120$, $b = 2$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0310	0.0923	0.0347	0.101	0.0386	0.106	0.0434	0.108
2	0.0322	0.0924	0.0329	0.102	0.0395	0.106	0.0410	0.108
3	0.0295	0.0925	0.0345	0.102	0.0422	0.106	0.0423	0.108
4	0.0306	0.0925	0.0339	0.102	0.0381	0.105	0.0430	0.108
5	0.0306	0.0926	0.0354	0.102	0.0399	0.106	0.0432	0.108
6	0.0290	0.0927	0.0367	0.102	0.0422	0.106	0.0409	0.108
7	0.0311	0.0926	0.0367	0.102	0.0423	0.106	0.0460	0.108
8	0.0277	0.0928	0.0375	0.102	0.0390	0.106	0.0435	0.108
9	0.0299	0.0926	0.0331	0.102	0.0409	0.106	0.0456	0.108
10	0.0309	0.0929	0.0329	0.102	0.0415	0.106	0.0436	0.108
11	0.0273	0.0929	0.0364	0.102	0.0393	0.106	0.0425	0.108
12	0.0294	0.0927	0.0361	0.102	0.0393	0.106	0.0452	0.108
13	0.0278	0.0929	0.0371	0.101	0.0430	0.106	0.0436	0.108
14	0.0274	0.0925	0.0347	0.102	0.0407	0.106	0.0405	0.108

15	0.0279	0.0929	0.0355	0.102	0.0393	0.106	0.0424	0.108
mean	0.0295	0.0927	0.0352	0.102	0.0404	0.106	0.0431	0.108

Mean ($\hat{\theta}$)	1.007012	1.007955	1.007884	1.009858
p_{95}	7.577025	10.85052	13.86007	16.72176
$\boldsymbol{p_{95}}$	6.50061	9.983808	13.25902	16.26987

Εδώ σημειώνουμε ότι με M=14, ο έλεγχος δίνει καλύτερο μέγεθος καθώς $P_t = 0.0487$.

Ισχύς Ελέγχου ($b_1 = 2 \ \ , M = 14$)

Για $b_2 = 1$: $P_t = 0.0776$

Για $b_2 = 0.5$: $P_t = 0.2482$

Για $b_2 = 0.33$: $P_t = 0.4716$

Για $b_2 = 0.25$: $P_t = 0.646$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 240$, $b = 2$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0272	0.105	0.0290	0.115	0.0372	0.120	0.0357	0.122
2	0.0310	0.105	0.0318	0.115	0.0352	0.119	0.0417	0.122
3	0.0239	0.105	0.0354	0.115	0.0338	0.119	0.0382	0.122
4	0.0275	0.105	0.0330	0.115	0.0371	0.119	0.0380	0.122
5	0.0293	0.105	0.0334	0.115	0.0362	0.119	0.0346	0.122
6	0.0278	0.105	0.0342	0.115	0.0347	0.119	0.0388	0.122
7	0.0258	0.105	0.0358	0.115	0.0345	0.120	0.0386	0.122
8	0.0269	0.105	0.0337	0.115	0.0340	0.120	0.0398	0.122
9	0.0277	0.105	0.0360	0.115	0.0438	0.119	0.0369	0.122
10	0.0292	0.105	0.0342	0.115	0.0371	0.120	0.0417	0.122

11	0.0281	0.105	0.0351	0.115	0.0379	0.119	0.0371	0.122
12	0.0253	0.105	0.0341	0.115	0.0369	0.119	0.0387	0.122
13	0.0287	0.105	0.0335	0.115	0.0377	0.119	0.0372	0.122
14	0.0278	0.105	0.0330	0.115	0.0361	0.119	0.0375	0.122
15	0.0297	0.105	0.0364	0.115	0.0392	0.119	0.0387	0.122
mean	0.0277	0.105	0.0339	0.115	0.0368	0.119	0.0382	0.122

Mean ($\hat{\theta}$)	1.004561	1.004369	1.003508	1.004991
p_{95}	7.546917	10.82271	13.83447	16.69616
$\boldsymbol{p_{95}}$	6.37809	9.851113	12.96157	15.88462

Εδώ σημειώνουμε ότι με M=20, ο έλεγχος δίνει καλύτερο μέγεθος καθώς $P_t = 0.0489$.

Ισχύς Ελέγχου ($b_1 = 2 \ , M = 20$)

Για $b_2 = 1$: $P_t = 0.0998$

Για $b_2 = 0.5$: $P_t = 0.4465$

Για $b_2 = 0.33$: $P_t = 0.7662$

Για $b_2 = 0.25$: $P_t = 0.9186$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 60$, $b = 1$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0325	0.0676	0.0435	0.0739	0.0486	0.0766	0.0517	0.0784
2	0.0352	0.0678	0.0403	0.0737	0.0502	0.0767	0.0570	0.0783
3	0.0335	0.0677	0.0421	0.0742	0.0461	0.0766	0.0558	0.0782
4	0.0345	0.0672	0.0396	0.0744	0.0485	0.0764	0.0537	0.0780
5	0.0322	0.0672	0.0406	0.0735	0.0485	0.0759	0.0513	0.0780
6	0.0326	0.0673	0.0403	0.0736	0.0487	0.0760	0.0514	0.0779

7	0.0356	0.0670	0.0419	0.0738	0.0452	0.0760	0.0551	0.0782
8	0.0313	0.0676	0.0439	0.0741	0.0490	0.0764	0.0519	0.0779
9	0.0345	0.0674	0.0429	0.0736	0.0468	0.0766	0.0547	0.0784
10	0.0356	0.0673	0.0378	0.0740	0.0479	0.0763	0.0523	0.0781
11	0.0323	0.0673	0.0410	0.0738	0.0436	0.0768	0.0544	0.0787
12	0.0347	0.0672	0.0401	0.0739	0.0487	0.0764	0.0535	0.0781
13	0.0339	0.0671	0.0398	0.0734	0.0458	0.0763	0.0542	0.0782
14	0.0322	0.0673	0.0444	0.0733	0.0460	0.0764	0.0518	0.0783
15	0.0340	0.0673	0.0388	0.0741	0.0474	0.0760	0.0525	0.0778
mean	0.0336	0.0674	0.0411	0.0738	0.0474	0.0764	0.0534	0.0782

Mean ($\hat{\theta}$)	1.02249	1.022742	1.023162	1.021624
p_{95}	7.641586	10.90936	13.91724	16.77543
$\boldsymbol{p_{95}}$	6.783739	10.39628	13.76054	16.99461

<u>Ισχύς Ελέγχου</u> $(b_1 = 1 \ , M = 10)$

Για $b_2 = 0.5$: $P_t = 0.0921$

Για $b_2 = 0.33$: $P_t = 0.1543$

Για $b_2 = 0.25$: $P_t = 0.2290$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 120$, $b = 1$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0258	0.0829	0.0337	0.0906	0.0361	0.0939	0.0400	0.0960
2	0.0297	0.0830	0.0368	0.0904	0.0415	0.0938	0.0460	0.0953
3	0.0308	0.0833	0.0352	0.0905	0.0405	0.0941	0.0394	0.0955

4	0.0310	0.0829	0.0331	0.0902	0.0415	0.0939	0.0405	0.0954
5	0.0280	0.0830	0.0370	0.0907	0.0401	0.0938	0.0467	0.0958
6	0.0306	0.0830	0.0363	0.0908	0.0367	0.0936	0.0454	0.0953
7	0.0260	0.0831	0.0379	0.0907	0.0384	0.0940	0.0447	0.0956
8	0.0273	0.0829	0.0352	0.0905	0.0395	0.0936	0.0451	0.0951
9	0.0285	0.0829	0.0378	0.0905	0.0396	0.0940	0.0403	0.0953
10	0.0259	0.0831	0.0362	0.0906	0.0373	0.0937	0.0465	0.0953
11	0.0295	0.0830	0.0382	0.0904	0.0390	0.0940	0.0441	0.0955
12	0.0285	0.0832	0.0389	0.0905	0.0363	0.0938	0.0430	0.0956
13	0.0283	0.0831	0.0345	0.0903	0.0400	0.0937	0.0414	0.0958
14	0.0309	0.0831	0.0328	0.0902	0.0386	0.0936	0.0451	0.0954
15	0.0305	0.0832	0.0368	0.0906	0.0362	0.0938	0.0415	0.0956
mean	0.0288	0.083	0.036	0.0905	0.0388	0.0938	0.0433	0.0955

Mean ($\hat{\theta}$)	1.011309	1.010218	1.011364	1.009715
p_{95}	7.601653	10.87425	13.88301	16.7435
$\boldsymbol{p_{95}}$	6.483348	10.06617	13.178	16.31902

Εδώ σημειώνουμε ότι με M=14, ο έλεγχος δίνει καλύτερο μέγεθος καθώς $P_t = 0.0491$.

Ισχύς Ελέγχου ($b_1 = 1 \ , M = 14$)

Για $b_2 = 0.5$: $P_t = 0.1090$

Για $b_2 = 0.33$: $P_t = 0.2271$

Για $b_2 = 0.25$: $P_t = 0.3618$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 240$, $b = 1$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0258	0.0938	0.0300	0.102	0.0345	0.106	0.0380	0.107
2	0.0299	0.0938	0.0349	0.102	0.0325	0.105	0.0365	0.107
3	0.0300	0.0938	0.0318	0.102	0.0356	0.105	0.0406	0.107
4	0.0258	0.0937	0.0306	0.102	0.0341	0.105	0.0395	0.107
5	0.0302	0.0937	0.0338	0.102	0.0366	0.105	0.0375	0.107
6	0.0270	0.0935	0.0338	0.102	0.0380	0.105	0.0396	0.107
7	0.0280	0.0937	0.0333	0.102	0.0364	0.105	0.0403	0.107
8	0.0269	0.0935	0.0311	0.102	0.0330	0.105	0.0419	0.107
9	0.0273	0.0936	0.0349	0.101	0.0326	0.105	0.0416	0.107
10	0.0273	0.0938	0.0323	0.102	0.0383	0.105	0.0376	0.107
11	0.0254	0.0936	0.0333	0.102	0.0365	0.106	0.0362	0.107
12	0.0276	0.0936	0.0314	0.102	0.0372	0.105	0.0369	0.107
13	0.0261	0.0935	0.0329	0.102	0.0348	0.105	0.0361	0.107
14	0.0300	0.0934	0.0329	0.102	0.0342	0.105	0.0390	0.107
15	0.0273	0.0938	0.0344	0.102	0.0354	0.105	0.0385	0.107
mean	0.0276	0.0937	0.0328	0.102	0.0353	0.105	0.0387	0.107

Mean ($\hat{\theta}$)	1.004991	1.00364	1.005397	1.006324
p_{95}	7.574991	10.851	13.86139	16.72221
$\boldsymbol{p_{95}}$	6.343039	9.783176	12.88561	15.91671

Εδώ σημειώνουμε ότι με M=18, ο έλεγχος δίνει καλύτερο μέγεθος καθώς $P_t = 0.0484$.

<u>Ισχύς Ελέγχου</u> ($b_1 = 1$, $M = 18$)

Για $b_2 = 0.5$: $P_t = 0.1489$

Για $b_2 = 0.33$: $P_t = 0.3887$

Για $b_2 = 0.25$: $P_t = 0.6340$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 60$, $b = 0.5$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0340	0.0561	0.0403	0.0617	0.0457	0.0638	0.0537	0.0653
2	0.0293	0.0565	0.0385	0.0616	0.0469	0.0640	0.0558	0.0654
3	0.0315	0.0561	0.0447	0.0618	0.0471	0.0643	0.0505	0.0652
4	0.0330	0.0562	0.0398	0.0619	0.0432	0.0640	0.0556	0.0651
5	0.0328	0.0561	0.0402	0.0617	0.0474	0.0640	0.0538	0.0653
6	0.0342	0.0562	0.0411	0.0618	0.0496	0.0640	0.0560	0.0654
7	0.0325	0.0563	0.0443	0.0617	0.0436	0.0637	0.0517	0.0651
8	0.0316	0.0563	0.0453	0.0620	0.0454	0.0641	0.0533	0.0653
9	0.0329	0.0565	0.0417	0.0617	0.0479	0.0640	0.0488	0.0652
10	0.0319	0.0569	0.0439	0.0617	0.0502	0.0641	0.0523	0.0654
11	0.0304	0.0563	0.0425	0.0614	0.0460	0.0639	0.0550	0.0654
12	0.0343	0.0563	0.0413	0.0621	0.0517	0.0640	0.0540	0.0656
13	0.0347	0.0559	0.0401	0.0618	0.0500	0.0637	0.0544	0.0654
14	0.0316	0.0564	0.0417	0.0614	0.0450	0.0639	0.0516	0.0654
15	0.0346	0.0567	0.0412	0.0617	0.0477	0.0638	0.0556	0.0654
mean	0.0326	0.0563	0.0418	0.0617	0.0472	0.064	0.0535	0.0653

Mean ($\hat{\theta}$)	1.019615	1.02503	1.024904	1.02238
p_{95}	7.668768	10.93596	13.94087	16.798
$\boldsymbol{p_{95}}$	6.791246	10.47509	13.77255	17.01714

Ισχύς Ελέγχου ($b_1 = 0.5$, $M = 10$)

Για $b_2 = 0.33$: $P_t = 0.0754$

Για $b_2 = 0.25$: $P_t = 0.1081$

Για $b_2 = 0.16$: $P_t = 0.1903$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 120$, $b = 0.5$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0290	0.0700	0.0367	0.0760	0.0389	0.0787	0.0462	0.0800
2	0.0268	0.0698	0.0377	0.0757	0.0355	0.0788	0.0417	0.0799
3	0.0261	0.0700	0.0372	0.0759	0.0416	0.0785	0.0424	0.0800
4	0.0288	0.0697	0.0341	0.0757	0.0344	0.0786	0.0417	0.0800
5	0.0270	0.0699	0.0373	0.0757	0.0376	0.0785	0.0410	0.0800
6	0.0274	0.0700	0.0315	0.0760	0.0425	0.0785	0.0406	0.0799
7	0.0297	0.0695	0.0334	0.0759	0.0389	0.0785	0.0436	0.0802
8	0.0273	0.0700	0.0342	0.0759	0.0377	0.0786	0.0431	0.0799
9	0.0304	0.0696	0.0359	0.0759	0.0371	0.0783	0.0432	0.0800
10	0.0291	0.0696	0.0359	0.0760	0.0351	0.0788	0.0443	0.0800
11	0.0300	0.0697	0.0339	0.0759	0.0406	0.0787	0.0430	0.0802
12	0.0292	0.0699	0.0359	0.0759	0.0371	0.0785	0.0422	0.0799
13	0.0267	0.0696	0.0350	0.0759	0.0411	0.0787	0.0440	0.0805
14	0.0283	0.0696	0.0352	0.0757	0.0385	0.0785	0.0438	0.0801
15	0.0272	0.0699	0.0332	0.0758	0.0406	0.0785	0.0441	0.0801
mean	0.0282	0.0698	0.0351	0.0759	0.0385	0.0786	0.043	0.08

Mean ($\hat{\theta}$)	1.011922	1.012522	1.015126	1.011203
p_{95}	7.635211	10.90566	13.91237	16.77128
$\boldsymbol{p_{95}}$	6.475261	10.04215	13.17945	16.30717

Εδώ σημειώνουμε ότι με M=14, ο έλεγχος δίνει καλύτερο μέγεθος καθώς $P_t = 0.0483$.

<u>Ισχύς Ελέγχου</u> ($b_1 = 0.5 \ , M = 14$)

Για $b_2 = 0.33$: $P_t = 0.0740$

Για $b_2 = 0.25$: $P_t = 0.1262$

Για $b_2 = 0.16$: $P_t = 0.2801$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

- Για : $n = 240$, $b = 0.5$ και $theta = 1$

M

	4		**6**		**8**		**10**	
#	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0253	0.0787	0.0317	0.0855	0.0350	0.0883	0.0383	0.0899
2	0.0269	0.0788	0.0325	0.0853	0.0369	0.0884	0.0387	0.0899
3	0.0244	0.0788	0.0332	0.0854	0.0361	0.0884	0.0357	0.0898
4	0.0244	0.0789	0.0332	0.0854	0.0351	0.0884	0.0405	0.0900
5	0.0271	0.0787	0.0304	0.0853	0.0331	0.0884	0.0392	0.0899
6	0.0262	0.0789	0.0297	0.0856	0.0350	0.0884	0.0357	0.0899
7	0.0245	0.0788	0.0302	0.0854	0.0377	0.0881	0.0398	0.0897
8	0.0292	0.0787	0.0305	0.0854	0.0355	0.0883	0.0370	0.0900
9	0.0252	0.0789	0.0293	0.0852	0.0345	0.0884	0.0389	0.0899
10	0.0283	0.0791	0.0321	0.0853	0.0338	0.0884	0.0377	0.0900
11	0.0259	0.0789	0.0321	0.0855	0.0390	0.0884	0.0393	0.0898
12	0.0229	0.0788	0.0318	0.0854	0.0375	0.0882	0.0384	0.0898
13	0.0283	0.0787	0.0297	0.0853	0.0354	0.0882	0.0394	0.0900
14	0.0254	0.0789	0.0311	0.0853	0.0338	0.0883	0.0385	0.0899
15	0.0258	0.0788	0.0322	0.0853	0.0377	0.0883	0.0399	0.0898
mean	0.026	0.0788	0.0313	0.0854	0.0357	0.0883	0.0385	0.0899

Mean ($\hat{\theta}$)	1.006241	1.007911	1.007768	1.006822

p_{95}	7.61265	10.88552	13.89348	16.75377
$\boldsymbol{p_{95}}$	6.321641	9.764375	12.9654	15.94666

Εδώ σημειώνουμε ότι με M=20, ο έλεγχος δίνει καλύτερο μέγεθος καθώς $P_t = 0.0492$.

Ισχύς Ελέγχου ($b_1 = 0.5 \ , M = 14$)

Για $b_2 = 0.33$: $P_t = 0.0865$

Για $b_2 = 0.25$: $P_t = 0.1884$

Για $b_2 = 0.16$: $P_t = 0.4866$

Η ισχύς του ελέγχου αυξάνεται όπως αναμενόταν όσο η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική υπόθεση.

4.4. Σχόλια - Συμπεράσματα

Με βάση τους παραπάνω πίνακες προσομοιώσεων, παρατηρούμε ότι ο προτεινόμενος έλεγχος καλής προσαρμογής συμπεριφέρεται καλά ως προς το μέγεθος αλλά και την ισχύ. Κάτω από την μηδενική υπόθεση υποθέτουμε ότι τα δεδομένα μας ακολουθούν μονομεταβλητά μοντέλα ευπάθειας με γνωστή διασπορά. Στην εργασία αυτή έχουμε υποθέσει ως κατανομές ευπάθειας τη Γάμμα και την Inverse Gaussian.

Η ελεγχοσυνάρτηση του ελέγχου βασίζεται σε μέτρα απόκλισης και πιο συγκεκριμένα στην κλάση των φ- μέτρων απόκλισης (φ- divergence measures) από τα οποία επιλέξαμε να χρησιμοποιήσουμε ένα από τα πιο βασικά το μέτρο Kullback-Leibler. Η ελεγχοσυνάρτηση T_n^{φ} απαιτεί τη διαμέριση των δεδομένων σε M γενικά μη επικαλυπτόμενα διαστήματα, όπως άλλωστε απαιτούν γενικά οι έλεγχοι καλής προσαρμογής. Βρήκαμε ότι για για κάθε μέγεθος δείγματος n υπάρχει ένα M για το οποίο η εμπειρική κατανομή της ελεγχοσυνάρτησης T_n^{φ} είναι πολύ κοντά με την ασυμπτωτική κατανομή της η οποία δίνεται από το άθροισμα των τ.μ. $X_{M-2}^2 + (1 -$

$\lambda)Z^2$. Από την ασυμπτωτική κατανομή πήραμε την κρίσιμη σταθερά για το χωρίο απορρίψεως του ελέγχου.

Ο στατιστικός έλεγχος έγινε σε επίπεδο σημαντικότητας 5%, οπότε το μέγεθος του ελέγχου (το ποσοστό $''P_t''$) είναι κοντά στο επίπεδο σημαντικότητας σε όλες τις περιπτώσεις που εξετάσαμε. Ερευνώντας το μοντέλο μας αρχικά για αριθμό διαστημάτων $M = 4,6,8,10$ παρατηρήσαμε ότι όσο μεγαλώνει το μέγεθος του δείγματος πρέπει να μεγαλώνει και ο αριθμός των διαστημάτων M της διαμέρισης για να επιτυγχάνεται καλό μέγεθος αλλά και ισχύς. Το συμπέρασμα αυτό ισχύει και για τις δύο κατανομές ευπάθειας που χρησιμοποιήσαμε. Αυτό εξηγείται και διαισθητικά καθώς όσο λιγότερα τα διαστήματα στα οποία χωρίζεται και ομαδοποιείται ένα μεγάλο σύνολο δεδομένων τόσο περισσότερες πληροφορίες χάνονται.

Μετά την έρευνά μας, προτείνουμε για μέγεθος δείγματος 60, το M να είναι 10, ενώ για μεγαλύτερα μεγέθη δείγματος 120 ή 240 το M να είναι 14 (ή σε κάποιες περιπτώσεις και λίγο μεγαλύτερο από 14, μέχρι 20, ειδικά για την Inverse Gaussian κατανομή).

Γενικά φαίνεται ότι όσο μεγαλώνει το μέγεθος του δείγματος, η ισχύς μεγαλώνει και ειδικά για τη Γάμμα κατανομή και μέγεθος δείγματος 240 η ισχύς πλησιάζει το 1 όταν απομακρυνόμαστε όλο και πιο πολύ απο τη μηδενική υπόθεση.

Οσο πιο μεγάλη η διασπορά της ευπάθειας, ο έλεγχος δυσκολεύεται να ξεχωρίσει μεταξύ της μηδενικής και εναλλακτικής υπόθεσης και η ισχύς γενικά μειώνεται. Για τη Γάμμα κατανομή όμως γενικά είναι πολύ ικανοποιητική ενώ όχι τόσο για την Inverse Gaussian, κάτι που πρέπει στο μέλλον να διερευνηθεί περισσότερο. Φυσικά και για την Inverse Gaussian όσο μεγαλώνει το μέγεθος του δείγματος η ισχύς βελτιώνεται.

Εν κατακλείδι, ο προτεινόμενος έλεγχος βάσει μέτρων απόκλισης συμπεριφέρεται πολύ καλά για τον έλεγχο του εάν ένα σετ δεδομένων προέρχεται από μοντέλο ευπάθειας με δεδομένη διασπορά. Περαιτέρω διερεύνηση φυσικά χρειάζεται ως προς άλλες κατανομές ευπάθειας, άλλες κατανομές της βασικής συνάρτησης κινδύνου, άλλα μέτρα φ-απόκλισης, επιπρόσθετα μεγέθη δείγματος και επιπρόσθετα επίπεδα σημαντικότητας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Aalen, O.O. (1988). Heterogeneity in survival analysis. Statistics in Medicine, 7, 1121-37.

Aalen, O. O. (1989). A linear regression model for the analysis of lifetimes. Stat. in Medicine ,8, 907-925

Aisbett, C.W., McGilchrist, C.A. (1991). Regression with Frailty in Survival Analysis. Biometrics 47, 461-466

Agresti, A. (1996). An Introduction to Categorical Data Analysis. John Wiley & Sons, New York, 2,1-53

Andersen, P. K. and Gill, R. D.(1982). Cox's regression model for counting processes: A large sample study. Annals of Stat., 10:1100-1120

Androulakis, E., Koukouvinos, C. and Vonta, F. (2012). Estimation and variable selection via frailty models with penalized likelihood, Statist. Med. 2012, 31, 2223–2239

Balakrishnan, V. and Sanghvi, L. D. (1968). Distance between populations on the basis of attribute. Biometrics, 24, 4, 859-865

Bennett, S. (1983). Analysis of survival data by the proportional odds model. Stat Med, 2: 273–277

Bonett, D. G. (1989). Pearson chi-square estimator and test for log-linear models with expected frequencies subject to linear constraints. Statistics and Probability Letters, 8, 175-177

Broffit, J. D. and Randles, R. H. (1977). A power approximation for the chi-square goodness-of-fit test: Simple hypothesis case. Journal of the American Statistical Association, 72, 604-607

Cox, D.R. (1972). Regression models with life tables (with discussion). J Roy Stat Soc B 34: 187–220

Duchateau, L., Janssen, P. (2008). The Frailty Model, Springer, New York, Biometrical Journal, 51(3):540-541

Fan, J., Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. Ann Stat 30(1):74–99

Ferentinos, K., Papaioannou, T. and Zografos (1990). φ-divergence statistics: Sampling properties, multinomial goodness of fit and divergence tests. Communications in Statistics (Theory and Methods), 19, 5, 1785-1802

Hanagal, D.D. (2007b). Gamma frailty regression models in mixture distributions. Economic Quality Control, 22(2), 295-302.

Henderson, R., Oman, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. Journal of the Royal Statistical Society B 61, 367 – 379

Hougaard, P. (1986b). A class of multivariate failure time distributions. Biometrika 73, 671-78

Hougaard, P. (1991). Modelling Heterogeneity in Survival Data. Journal of Applied Probability 28, 695 – 701

Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on EM algorithm. Biometrics, 48, 795-806.

Mathai (1982). Storage capacity of a dam with gamma type inputs. Annals of the Institute of Statistical Mathematics, 34, 591–597

Menendez, M. L., Morales, D., Pardo, L. and Vajda, I. (2001b). Minimum divergence estimators based on grouped data. Annals of the Institute of Statistical Mathematics, 53,2, 277-288

Menendez, M. L., Morales, D., Pardo, L. and Salicru, M. (1995). Asymptotic behavior and statistical applications of divergence measures in multinomial populations: A unified study. Statistical Papers, 36, 1-29

Moschopoulos, P. G. (1984). The distribution of the sum of independent gamma random variables. Annals of the Institute of Statistical Mathematics, 37, 541–544

Pardo, L.(2005). Statistical Inference Based on Divergence Measures. Statistics: A Series of Textbooks and Monographs Book 185. Mathematics & Statistics, 1st Edition , Chapters 1-6

Vonta, F. (1996). Efficient estimation in a nonproportional hazards model in survival analysis, Scand. J. Statist., 23, No 1, 49-61

Vonta, F & Karagrigoriou, A. (2007). Variable selection strategies in survival models with multiple imputations, Lifetime Data Analysis, 13, 295–315