ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
Δ.Π.Μ.Σ ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

Μη Παραμετρικοί Έλεγχοι Υποθέσεων για τα Μοντέλα Ευπάθειας

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΝΙΚΟΛΑΟΥ Α. ΕΛΕΥΘΕΡΙΟΥ

Αθήνα, 2018
Επιβλέπουσα Καθηγήτρια : ΦΙΛΙΑ ΒΟΝΤΑ

Η παρούσα Διπλωματική Εργασία εκπονήθηκε
στα πλαίσια των σπουδών για την απόκτηση του
Μεταπτυχιακού Διπλώματος Ειδίκευσης στις
Εφαρμοσμένες Μαθηματικές Επιστήμες.

Ονοματεπώνυμο	**Βαθμίδα**
Φιλία Βόντα (Επιβλέπουσα)	Αναπληρώτρια Καθηγήτρια Ε.Μ.Π.
Χρυσηίς Καρώνη	Καθηγήτρια Ε.Μ.Π.
Καραγρηγορίου Αλέξανδρος	Καθηγητής Παν. Αιγαίου

Περίληψη

Η παρούσα εργασία αποτελείται από τέσσερα κεφάλαια. Στο πρώτο κεφάλαιο παρουσιάζονται οι εισαγωγικές έννοιες της Ανάλυσης Επιβίωσης και αναλύεται το μοντέλο αναλογικών κινδύνων του Cox και η έννοια της μερικής πιθανοφάνειας. Στην συνέχεια περιγράφουμε τις εισαγωγικές έννοιες της τυχαίας μεταβλητής που περιγράφει την ευπάθεια, τον συνδυασμό της με το μοντέλο του Cox και εισαγάγουμε τα μοντέλα ευπάθειας ή μοντέλα μετασχηματισμού. Στο δεύτερο κεφάλαιο παρουσιάζονται οι ορισμοί και ιδιότητες των μέτρων φ-απόκλισης και αναφέρονται μερικές από τις πιο γνωστές αποκλίσεις. Στο τρίτο κεφάλαιο αναλύουμε κάποιες ελεγχοσυναρτήσεις που βασίζονται στις φ-αποκλίσεις και χρησιμοποιούνται στους ελέγχους καλής προσαρμογής και επίσης, παρουσιάζεται η βάση της μεθοδολογίας που χρησιμοποιείται στο τέταρτο κεφάλαιο. Στο τέταρτο κεφάλαιο ορίζονται καινούργιοι έλεγχοι καλής προσαρμογής με βάση τα μέτρα φ-απόκλισης για λογοκριμένα δεδομένα που ακολουθούν μοντέλα ευπάθειας κάτω από τη μηδενική υπόθεση. Η συμπεριφορά των ελέγχων ως προς το μέγεθος παρουσιάζεται μέσω προσομοιώσεων.

Λέξεις κλειδιά: « Μοντέλα ευπάθειας, φ μέτρα απόκλισης, έλεγχος καλής προσαρμογής »

Abstract

This thesis consists of four chapters. In the first chapter the basic concepts of Survival Analysis are presented and the Cox proportional hazards model as well as the concept of partial likelihood are analysed. Subsequently, we describe the basic concepts of the frailty random variable, we explain its association with the Cox model and we introduce the frailty models or otherwise transformation models. In the second chapter, the definitions and properties of the phi-divergence measures are presented and a few of the most important are described. In the third chapter we analyse some of the test statistics that are based on phi-divergences and are used in goodness-of-fit tests and we present the basis of the methodology that is used in the fourth chapter. In the fourth chapter, new goodness-of-fit tests based on phi-divergence measures for censored data that follow frailty models under the null hypothesis are defined. The performance of the tests with respect to their size is assessed by simulations.

Keywords: « Frailty models, phi-divergence measures, goodness-of-fit-tests**»**

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα καθηγήτρια κυρία Φιλία Βόντα για την εμπιστοσύνη που μου έδειξε αναθέτοντάς μου την εργασία αυτή. Την ευχαριστώ για την υπομονή και την αμέριστη καθοδήγησή της με καίριες υποδείξεις καθ' όλη την διάρκεια της εκπόνησης της μεταπτυχιακής μου εργασίας.

Επιπλέον, θα ήθελα να ευχαριστήσω την καθηγήτρια κυρία Χρυσηίδα Καρώνη και τον καθηγητή κύριο Αλέξανδρο Καραγρηγορίου για την τιμή που μου έκαναν να συμμετάσχουν στην τριμελή εξεταστική επιτροπή.

Πίνακας περιεχομένων

ΚΕΦΑΛΑΙΟ 1

ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ ΚΑΙ ΕΥΠΑΘΕΙΑ

1.1 Βασικές Έννοιες της Ανάλυσης Επιβίωσης

Η ανάλυση επιβίωσης είναι μια σημαντική ερευνητική περιοχή που σχετίζεται με διάφορα πεδία όπως η ιατρική, η βιολογία, η επιδημιολογία, η δημογραφία, η μηχανική. Για τα δεδομένα ανάλυσης επιβίωσης απαιτείται ειδική στατιστική μεθοδολογία για την επεξεργασία και ανάλυσή τους. Τα δεδομένα αυτά συνήθως αφορούν χρόνους μέχρι να συμβεί ένα γεγονός, για παράδειγμα ο χρόνος μέχρι τον θάνατο, ο χρόνος μέχρι να εμφανισθεί κάποια ασθένεια, ο χρόνος μέχρι να χαλάσει κάποια μηχανή. Ένας λόγος για τον οποίο απαιτείται ειδική στατιστική μεθοδολογία είναι το πρόβλημα της λογοκρισίας. Ένα χαρακτηριστικό παράδειγμα του προβλήματος αυτού είναι π.χ. η μελέτη για άτομα από ένα πληθυσμό για τα οποία ο ερευνητής έχει μόνο την πληροφορία ότι το συμβάν προς μελέτη δεν πραγματοποιήθηκε πριν από ένα συγκεκριμένο χρονικό σημείο. Άρα μια λογοκριμένη παρατήρηση περιέχει μόνο μερική πληροφορία για τον τυχαίο χρόνο που μας ενδιαφέρει.

Σύμφωνα με τον Wienke (βλ. [14]) θεωρούμε μια τυχαία μεταβλητή T^*, η οποία είναι μη αρνητική και αντιπροσωπεύει το χρόνο από ένα αρχικό χρονικό σημείο μέχρι να συμβεί ένα συμβάν. Το συμβάν μπορεί να είναι για παράδειγμα, θάνατος ή η εκδήλωση μιας ασθένειας ή επιπλοκές μετά από μια εγχείρηση. Ο χρόνος T^* ονομάζεται γενικώς, χρόνος επιβίωσης. Ο χρόνος T^* μέχρι να συμβεί ένα συμβάν θεωρείται ότι ακολουθεί κάποια συνεχή κατανομή και όλες οι συναρτήσεις κατανομής για τον χρόνο αυτό ορίζονται στο διάστημα $[0, \infty)$. Αν η συνάρτηση πυκνότητας πιθανότητας συμβολίζεται με f τότε η αθροιστική συνάρτηση πιθανότητας ορίζεται από την σχέση:

$$F(t) = \boldsymbol{P}(T^* \leq t) = \int_{-\infty}^{t} f(s)ds$$

Η πιθανότητα ένα άτομο να επιβιώσει πέρα από τον χρόνο t δίνεται από την σχέση:

$$S(t) = 1 - F(t) = P(T^* > t) = \int_{t}^{\infty} f(s)ds$$

Μια άλλη βασική συνάρτηση της ανάλυσης επιβίωσης είναι η συνάρτηση κινδύνου που ορίζεται ως ο στιγμιαίος ρυθμός διακοπής (θανάτου) μιας μονάδας (ή ατόμου) την αμέσως επόμενη χρονική στιγμή t, δεδομένου ότι έχει επιβιώσει μέχρι τη στιγμή t. Η σχέση από την οποία προσδιορίζεται είναι η παρακάτω:

$$h(t) = \lim_{dt \to 0} \frac{\boldsymbol{P}(t < T^* \leq t + dt | T^* > t)}{dt} = \frac{f(t)}{1 - F(t)}$$

το οποίο ισχύει αφού:

$$\boldsymbol{P}(t < T^* \leq t + dt | T^* > t) = \frac{\boldsymbol{P}(t < T^* \leq t + dt)}{\boldsymbol{P}(T^* > t)} = \frac{F(t + dt) - F(t)}{S(t)} \approx \frac{f(t)dt}{S(t)}$$

και με αντικατάσταση ορίζεται η παραπάνω σχέση. Ομοίως προσδιορίζουμε την αθροιστική συνάρτηση κινδύνου ως:

$$H(t) = \int_0^t h(s)ds$$

ή

$$H(t) = \int_0^t h(s)ds = \int_0^t \frac{f(s)}{S(s)}ds = \int_0^t \frac{(-S'(s))}{S(s)}ds = -[\ln S(s)]_0^t = -lnS(t)$$

Οπότε θα ισχύει:

$$S(t) = e^{-H(t)} = e^{-\int_0^t h(s)ds}$$

1.3.2 Λογοκρισία

Μια άλλη βασική έννοια της ανάλυσης επιβίωσης και ένα από τα σημαντικότερα προβλήματα είναι η λογοκρισία. Όπως αναφέρθηκε στην αρχή του κεφαλαίου μια λογοκριμένη παρατήρηση αποτελεί μια ελλιπή παρατήρηση αφού περιέχει μερική πληροφορία σχετικά με τον χρόνο συμβάντος. Για παράδειγμα, στην περίπτωση ενός ασθενή που είναι υπό επίβλεψη μόνο για ένα χρονικό διάστημα και το συμβάν δεν έχει συμβεί μέσα σε αυτό το διάστημα. Το μόνο που είναι γνωστό είναι ότι ο πραγματικός χρόνος συμβάντος ξεπερνάει τον παρατηρούμενο λογοκριμένο χρόνο. Υπάρχουν διάφορα είδη λογοκρισίας, για παράδειγμα λογοκρισία από δεξιά, λογοκρισία από αριστερά, λογοκρισία σε διάστημα καθώς και ταυτόχρονη λογοκρισία και από δεξιά και από αριστερά. Στην εργασία αυτή θα ασχοληθούμε μόνο με την λογοκρισία από δεξιά.

Έστω $T_1^*, T_2^*, \dots, T_n^*$ ανεξάρτητες και όμοια κατανεμημένες μεταβλητές που αποτελούν τους χρόνους επιβίωσης. Οι χρόνοι αυτοί έχουν αθροιστική συνάρτηση κατανομής F και έστω $C_1, C_2, \dots, C_n$ ανεξάρτητες και όμοια κατανεμημένες μεταβλητές που αποτελούν τους χρόνους λογοκρισίας με αθροιστική συνάρτηση κατανομής G. Οι συναρτήσεις F, G θεωρούμε ότι είναι απόλυτα συνεχείς. Επίσης, έστω f και g οι συναρτήσεις πυκνότητας πιθανότητας που σχετίζονται με τις συναρτήσεις F, G. Οι

παρατηρήσεις αντιπροσωπεύονται από τα ζεύγη δεδομένων (T_1, Δ_1), $(T_2, \Delta_2), \dots, (T_n, \Delta_n)$, όπου $T_i = min\{T_i^*, C_i\}$ αποτελεί τον παρατηρούμενο χρόνο και

$$\Delta_i = \begin{cases} 1: \alpha\nu\ T_i^* \le C_i\ \kappa\alpha\iota\ \acute{\alpha}\rho\alpha\ \tau o\ T_i\ \theta\alpha\ \epsilon\acute{\iota}\nu\alpha\iota\ \mu\eta\ \lambda o\gamma o\kappa\rho\iota\mu\acute{\epsilon}\nu o\varsigma\ \chi\rho\acute{o}\nu o\varsigma\ \delta\iota\alpha\kappa o\pi\acute{\eta}\varsigma \\ 0: \alpha\nu\ T_i^* > C_i\ \ \kappa\alpha\iota\ \acute{\alpha}\rho\alpha\ \tau o\ T_i\ \theta\alpha\ \alpha\pi o\tau\epsilon\lambda\epsilon\acute{\iota}\ \lambda o\gamma o\kappa\rho\iota\mu\acute{\epsilon}\nu o\ \chi\rho\acute{o}\nu o\ \delta\iota\alpha\kappa o\pi\acute{\eta}\varsigma \end{cases}$$

Η λογοκρισία είναι ένα κοινό πρόβλημα σε περιπτώσεις όπως οι κλινικές μελέτες, όπου οι ασθενείς μπορεί να εισάγονται στις μελέτες σε διαφορετικούς χρόνους και να αποχωρούν ίσως σε διαφορετικούς χρόνους. Ενδιαφέρον προς μελέτη παρουσιάζουν οι χρόνοι συμβάντων των ασθενών, ενώ η λογοκρισία τους μπορεί να παρουσιάζεται για τους ακόλουθους λόγους:

- Ο ασθενής μπορεί να έχει μετακομίσει και να μην μπορεί να πάρει μέρος στην μελέτη.
- Η θεραπεία μπορεί να έχει πολύ ισχυρές παρενέργειες και να είναι απαραίτητο να διακοπεί η μελέτη.
- Η μελέτη έχει προκαθορισμένη διάρκεια, οπότε και σταματάει η παρακολούθηση των ασθενών.
- Το συμβάν δεν μπορεί να παρατηρηθεί λόγω άλλου συμβάντος του ασθενή, όπως ο θάνατος του λόγω ατυχήματος.

1.3.2 Παραμετρικά Μοντέλα

Στη συνέχεια θεωρούμε ανεξαρτησία μεταξύ του χρόνου επιβίωσης και του χρόνου λογοκρισίας, καθώς επίσης και ότι η κατανομή του χρόνου λογοκρισίας είναι ανεξάρτητη από τις παραμέτρους που υπεισέρχονται στην κατανομή του χρόνου επιβίωσης. Εάν $\boldsymbol{\theta}$ είναι η παράμετρος από την οποία εξαρτάται η κατανομή του χρόνου επιβίωσης με πυκνότητα πιθανότητας $f(t, \boldsymbol{\theta})$, υποθέτουμε ότι η κατανομή του χρόνου λογοκρισίας δεν εξαρτάται από το $\boldsymbol{\theta}$ και άρα λοιπόν δεν περιέχεται σε αυτήν πληροφορία για το άγνωστο $\boldsymbol{\theta}$. Αποτέλεσμα αυτών των υποθέσεων είναι ότι για την περίπτωση της δεξιάς λογοκρισίας για τα δεδομένα επιβίωσης $(T_i, \Delta_i), i = 1, \dots, n$ η συνάρτηση πιθανοφάνειας για ένα άτομο i είναι ανάλογη του

$$L_i(\boldsymbol{\theta}) = f(t_i; \boldsymbol{\theta})^{\delta_i} S(t_i; \boldsymbol{\theta})^{1-\delta_i} = h(t_i; \boldsymbol{\theta})^{\delta_i} S(t_i; \boldsymbol{\theta}) = h(t_i; \boldsymbol{\theta})^{\delta_i} e^{\int_0^{t_i} h(s;\boldsymbol{\theta}) ds}$$

Για ένα δείγμα ανεξάρτητων χρόνων ζωής $(T_i, \Delta_i), i = 1, \dots, n$ η συνάρτηση πιθανοφάνειας των δεδομένων είναι της μορφής:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}) = \prod_{i=1}^{n} h(t_i; \boldsymbol{\theta})^{\delta_i} e^{\int_0^{t_i} h(s;\boldsymbol{\theta}) ds}$$

Παρακάτω θα αναπτύξουμε κάποιες κατανομές πιθανότητας που χρησιμοποιούνται στην ανάλυση επιβίωσης. Όπως γνωρίζουμε οποιαδήποτε κατανομή μη αρνητικών τυχαίων μεταβλητών μπορεί να χρησιμοποιηθεί για την περιγραφή του χρόνου ζωής. Στα παραμετρικά μοντέλα η κατανομή που ακολουθεί η τυχαία μεταβλητή του χρόνου, υποθέτουμε ότι είναι γνωστή εκτός από μία παράμετρο πεπερασμένης διάστασης. Παρακάτω, θα αναφερθούμε σε μερικά κύρια παραμετρικά μοντέλα.

Η Εκθετική κατανομή

Το μοντέλο της εκθετικής κατανομής ($T \sim Exp(\lambda)$) είναι το πιο απλό παραμετρικό μοντέλο ζωής, με μια μόνο παράμετρο $\lambda > 0$. Αντιπροσωπευτικές συναρτήσεις του μοντέλου είναι:

$$Συνάρτηση\ πυκνότητας\ πιθανότητας: f(t) = \lambda e^{-\lambda t}$$

$$Συνάρτηση\ Επιβίωσης: S(t) = e^{-\lambda t}$$

$$Συνάρτηση\ Κινδύνου: h(t) = \lambda$$

$$Αθροιστική\ Συνάρτηση\ Κινδύνου: H(t) = \lambda t$$

$$Μέση\ Τιμή: E(T) = \frac{1}{\lambda}$$

$$Διασπορά: V(T) = \frac{1}{\lambda^2}$$

Το μοντέλο αυτό, λόγω της σταθερής συνάρτησης κινδύνου η οποία είναι ανεξάρτητη του χρόνου δεν αποτελεί πολύ ρεαλιστικό μοντέλο για την ανάλυση δεδομένων ζωής. Παρ'όλα αυτά λόγω της απλότητας των παραπάνω συναρτήσεων χρησιμοποιείται αρκετά στην ανάλυση δεδομένων.

Γάμμα κατανομή

Η Γάμμα κατανομή είναι μια προέκταση της εκθετικής κατανομής. Όπως φαίνεται παρακάτω επειδή οι συναρτήσεις επιβίωσης και κινδύνου της κατανομής δεν έχουν κλειστή μορφή, είναι δύσκολη η χρήση τους στην παραμετρική εκτίμηση. Αν η τυχαία μεταβλητή T ακολουθεί την κατανομή Γάμμα με παράμετρο κλίμακας λ και παράμετρο σχήματος κ, $T \sim G(\lambda, \kappa)$, τότε η συνάρτηση πιθανότητας είναι

$$f(t) = \frac{\lambda^{\kappa} t^{\kappa-1} e^{-\lambda t}}{\Gamma(\kappa)}, \quad \kappa > 0, \lambda > 0$$

Άλλες βασικές συναρτήσεις είναι:

$$S(t) = \frac{\int_{\lambda t}^{\infty} u^{\kappa-1} e^{-u} du}{\Gamma(\kappa)}$$

$$h(t) = \frac{\lambda^{\kappa} t^{\kappa-1} e^{-\lambda t}}{\int_{\lambda t}^{\infty} u^{\kappa-1} e^{-u} du}$$

$$E(T) = \frac{\kappa}{\lambda}$$

$$V(T) = \frac{\kappa}{\lambda^2}$$

Αντίστροφη Γκαουσιανή Κατανομή

Η συνάρτηση πυκνότητας πιθανότητας της κατανομής είναι:

$$f(t) = \sqrt{\frac{\lambda}{2\pi t^3}} e^{\frac{-\lambda(t-\mu)^2}{2t\mu^2}}, \qquad t > 0, \mu > 0, \lambda > 0$$

Η συνάρτηση επιβίωσης και κινδύνου έχουν την μορφή:

$$S(t) = 1 - \Phi\left(\sqrt{\frac{\lambda}{t}}\left(\frac{t}{\mu} - 1\right)\right) - e^{\frac{2\lambda}{\mu}}\Phi\left(-\sqrt{\frac{\lambda}{t}}\left(1 + \frac{t}{\mu}\right)\right)$$

$$h(t) = \frac{\sqrt{\frac{\lambda}{2\pi t^3}} e^{\frac{-\lambda(t-\mu)^2}{2t\mu^2}}}{1 - \Phi\left(\sqrt{\frac{\lambda}{t}}\left(1 - \frac{t}{\mu}\right)\right) - e^{\frac{2\lambda}{\mu}}\Phi\left(-\sqrt{\frac{\lambda}{t}}\left(1 + \frac{t}{\mu}\right)\right)}$$

όπου Φ είναι η συνάρτηση κατανομής της τυποποιημένης Κανονικής κατανομής $N(0,1)$.
Με μέση τιμή και διασπορά:

$$E(T) = \mu, \qquad V(T) = \frac{\mu^3}{\lambda}$$

1.2 Μοντέλα Παλινδρόμησης – Μοντέλο Αναλογικών Κινδύνων

Τα παραπάνω μοντέλα που αναπτύξαμε αφορούν την περίπτωση ανεξάρτητων και όμοια κατανεμημένων μεταβλητών, δηλαδή περιπτώσεις ομοιογενών πληθυσμών. Παρ'όλα αυτά συνήθως ο πληθυσμός που μελετούμε δεν είναι ομοιογενής. Για παράδειγμα, τα άτομα του πληθυσμού μιας επιδημιολογικής μελέτης μπορεί να διαφέρουν στην ηλικία, στο φύλο, στο επίπεδο εκπαίδευσης, στην οικογενειακή κατάσταση, στην γενετική προδιάθεση καθώς και πολλούς άλλους παράγοντες. Μπορεί κάποιες από αυτές τις συμμεταβλητές να είναι ιδιαίτερα σημαντικές, όπως η επίδραση μιας θεραπείας σε μια μελέτη, ή μπορεί να είναι παράγοντες που απαιτείται η ρύθμιση της επίδρασης τους για την ανάλυση. Το μοντέλο αναλογικών κινδύνων του Cox (1972) είναι ένα μοντέλο παλινδρόμησης όπου ο χρόνος συμβάντος είναι εξαρτημένη μεταβλητή. Επιτρέπει να περιληφθούν πληροφορίες για γνωστές, παρατηρούμενες συμμεταβλητές που πιθανώς συνδέονται με το χρόνο επιβίωσης και αποτελεί ένα από τα πιο χρησιμοποιούμενα μοντέλα στην ανάλυση επιβίωσης.

Έστω $h(t|\boldsymbol{X})$ η συνάρτηση κινδύνου ενός ατόμου την χρονική στιγμή t με διάνυσμα συμμεταβλητών $\boldsymbol{X}' = (X_1, \dots, X_k)$. Το $\boldsymbol{X}'$ αποτελεί το ανάστροφο διάνυσμα του διανύσματος $\boldsymbol{X}$.
Η συνάρτηση κινδύνου στο μοντέλο του Cox ορίζεται ως εξής:

$$h(t|\boldsymbol{X}) = h_0(t)g(\boldsymbol{X}), \qquad (1)$$

όπου $h_0(t)$ είναι η βασική συνάρτηση κινδύνου και $g(\cdot)$ είναι μια θετική συνάρτηση. Στο μοντέλο θεωρούμε ότι ο βασικός κίνδυνος αποτελεί ένα κοινό χαρακτηριστικό όλων των ατόμων της μελέτης και αντιστοιχεί σε άτομο με συμμεταβλητές $\boldsymbol{X} = \boldsymbol{0}$. Οι παράμετροι παλινδρόμησης περιέχονται στο $g(\boldsymbol{X}) = g(\boldsymbol{\beta}, \boldsymbol{X})$ και ορίζεται ως εξής:

$$g(\boldsymbol{X}) = e^{\boldsymbol{\beta' X}} \qquad (2)$$

όπου $\boldsymbol{\beta'} = (\beta_1, \dots, \beta_k)$ είναι το διάνυσμα των παραμέτρων παλινδρόμησης. Σε αυτό το μοντέλο οι συμμεταβλητές δρούν πολλαπλασιαστικά στον βασικό κίνδυνο, προσθέτοντας επιπλέον ρίσκο σύμφωνα με τα δεδομένα του κάθε ατόμου. Μπορεί να υποτεθεί ότι κάθε ατομική μεταβολή του κινδύνου μπορεί να χαρακτηριστεί από ένα πεπερασμένο διάνυσμα των παρατηρούμενων συμμεταβλητών (επεξηγηματικές μεταβλητές, παράγοντες κινδύνου, παράγοντες παλινδρόμησης). Η βασική ιδέα αυτής της υπόθεσης είναι ο διαχωρισμός, αφ' ενός της επίδρασης του χρόνου στον βασικό κίνδυνο και αφ' ετέρου της επίδρασης των συμμεταβλητών μέσω ενός εκθετικού όρου. Ουσιαστικά η υπόθεση αυτή ορίζει ότι ο κίνδυνος δύο ατόμων την χρονική στιγμή t σχετίζεται με μια σταθερά αναλογίας που δεν εξαρτάται από τον χρόνο. Η απλή περίπτωση δύο δειγμάτων λαμβάνεται για $k = 1$, όπου έχουμε την περίπτωση του δυαδικού διανύσματος συμμεταβλητών X με $X = 0$ ή $X = 1$ που δηλώνει δύο ομάδες, την ομάδα μηδέν και την ομάδα ένα. Το μοντέλο για τις δύο ομάδες έχει την μορφή:

$$h_i(t|\boldsymbol{X}) = \begin{cases} h_0(t) & \alpha\nu\ X = 0 \\ h_0(t)e^{\beta} & \alpha\nu\ X = 1 \end{cases}, i = 0{,}1$$

Το $h_0(t)$ αποτελεί τον κίνδυνο την χρονική στιγμή t για την ομάδα μηδέν, ενώ το e^{β} αποτελεί τον λόγο κινδύνου για την ομάδα ένα σε σχέση με την ομάδα μηδέν για την χρονική στιγμή t, όπως φαίνεται και από την παρακάτω σχέση:

$$\frac{h_1(t|\boldsymbol{X})}{h_0(t|\boldsymbol{X})} = \frac{h(t|X=1)}{h(t|X=0)} = \frac{h_0(t)e^{1*\beta}}{h_0(t)e^{0*\beta}} = e^{(1-0)\beta} = e^{\beta}, \qquad \forall\, t \geq 0$$

Το μοντέλο (1) διαχωρίζει την επίδραση του χρόνου με την επίδραση των συμμεταβλητών. Αν πάρουμε τον λογάριθμο της σχέσης αυτής παρατηρούμε το παρακάτω γραμμικό μοντέλο:

$$\log h(t|\boldsymbol{X}) = \log h_0(t) + \boldsymbol{\beta' X}$$

Το $h_0(t)$ αποτελεί το μη παραμετρικό κομμάτι του μοντέλου αφού μπορεί να πάρει οποιαδήποτε μορφή ως συνάρτηση του t, με μόνο περιορισμό $h_0(t) > 0$. Το $\boldsymbol{\beta' X}$ αποτελεί το παραμετρικό κομμάτι του μοντέλου, οπότε και το μοντέλο του Cox καλείται ως ημιπαραμετρικό μοντέλο.

Η συνάρτηση επιβίωσης του T είναι:

$$S(t|\boldsymbol{X}) = S_0(t)^{e^{\boldsymbol{\beta' X}}}$$

όπου

$$S_0(t) = e^{-\int_0^t h_0(s)ds} = e^{-H_0(t)}$$

αποτελεί την βασική συνάρτηση επιβίωσης.

Στην περίπτωση του μοντέλου του Cox επειδή ο όρος $e^{\boldsymbol{\beta}'\boldsymbol{X}}$ είναι πάντα θετικός τότε και ο ατομικός κίνδυνος $h(t|\boldsymbol{X})$ είναι και αυτός ένας μη αρνητικός όρος για όλες τις τιμές των t και $\boldsymbol{\beta}$, όπως θα έπρεπε να είναι αφού δηλώνει πιθανότητα.

Στην περίπτωση του ημιπαραμετρικού μοντέλου του Cox, επειδή η πιθανοφάνεια περιέχει την άγνωστη συνάρτηση κινδύνου $h_0(t)$ απαιτείται τροποποίηση της ώστε να μπορεί να εκτιμηθεί το διάνυσμα $\boldsymbol{\beta}$, χωρίς να απαιτείται η εκτίμηση της βασικής συνάρτησης κινδύνου. Ο Cox (1972, 1975) πρότεινε την μέθοδο της μερικής πιθανοφάνειας. Σε αυτή την περίπτωση για δείγμα μεγέθους n για τα παρατηρούμενα δεδομένα $(t_i, \delta_i, \boldsymbol{X_i}),\ i = 1, \dots, n$, όπου $t_i = \min(T_i^*, C_i)$ οι παρατηρούμενοι χρόνοι με $T_i^*, C_i,\ i = 1, \dots, n$ να είναι οι χρόνοι επιβίωσης και λογοκρισίας αντίστοιχα και δ_i δείκτρια συνάρτηση για την οποία ισχύει $I[T_i^* \le C_i], i = 1, \dots, n$ και εφόσον δεν υπάρχουν ισοπαλίες χρόνων τότε από την σχέση:

$$f(t_i) = h(t_i)S(t_i)$$

η σχέση για τον υπολογισμό της πιθανοφάνειας των δεδομένων επιβίωσης μπορεί να πάρει την μορφή:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i) = \prod_{i=1}^{n} \left(h_0(t_i)e^{\boldsymbol{\beta}'\boldsymbol{X_i}}\right)^{\delta_i} S(t_i)$$

$$= \prod_{i=1}^{n} \left(\frac{h_0(t_i)e^{\boldsymbol{\beta}'\boldsymbol{X_i}}}{\sum_{l\in R(t_i)} h_0(t_i)e^{\boldsymbol{\beta}'\boldsymbol{X_l}}}\right)^{\delta_i} \left(\sum_{l\in R(t_i)} h_0(t_i)e^{\boldsymbol{\beta}'\boldsymbol{X_l}}\right)^{\delta_i} S(t_i)$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{\beta}'\boldsymbol{X_i}}}{\sum_{l\in R(t_i)} e^{\boldsymbol{\beta}'\boldsymbol{X_l}}}\right)^{\delta_i} \left(\sum_{l\in R(t_i)} h_0(t_i)e^{\boldsymbol{\beta}'\boldsymbol{X_l}}\right)^{\delta_i} S(t_i)$$

όπου $R(t)$ ορίζεται ως το σύνολο που περιέχει το πλήθος των ατόμων που την χρονική στιγμή t είναι σε κίνδυνο να συμβεί το συμβάν. Η μερική πιθανοφάνεια που απέδειξε ο Cox ότι περιέχει όλη την πληροφορία σχετικά με το $\boldsymbol{\beta}$ ενώ παράλληλα είναι ανεξάρτητη από το $h_0(t)$ δίνεται από την σχέση (και είναι κομμάτι της ολικής πιθανοφάνειας):

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{\beta}'\boldsymbol{X_i}}}{\sum_{l\in R(t_i)} e^{\boldsymbol{\beta}'\boldsymbol{X_l}}}\right)^{\delta_i}$$

Η πιθανοφάνεια αυτή χρησιμοποιείται για την εκτίμηση των συντελεστών παλινδρόμησης στο ημιπαραμετρικό μοντέλο αναλογικών κινδύνων ως εξής:

$$l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \log\left(\prod_{i=1}^{n}\left(\frac{e^{\boldsymbol{\beta}'\boldsymbol{X_i}}}{\sum_{l\in R(t_i)} e^{\boldsymbol{\beta}'\boldsymbol{X_l}}}\right)^{\delta_i}\right) =$$

$$= \sum_{i=1}^{n} \delta_i \left[\boldsymbol{\beta}'\boldsymbol{X_i} - \log\left(\sum_{l\in R(t_i)} e^{\boldsymbol{\beta}'\boldsymbol{X_l}}\right)\right]$$

Οπότε μεγιστοποιώντας τον λογάριθμο της πιθανοφάνειας ως προς $\boldsymbol{\beta}$ μπορούμε να βρούμε εκτιμήσεις για τους συντελεστές παλινδρόμησης. Η συνάρτηση score της μερικής πιθανοφάνειας δίνεται από την σχέση:

$$U(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \delta_i \left[\boldsymbol{X}_i - \frac{\sum_{l\in R(t_i)} \boldsymbol{X_l} e^{\boldsymbol{\beta}'\boldsymbol{X_l}}}{\sum_{l\in R(t_i)} e^{\boldsymbol{\beta}'\boldsymbol{X_l}}}\right]$$

Το μέγιστο υπολογίζεται για $U(\boldsymbol{\beta}) = 0$.

Η μερική πιθανοφάνεια εφαρμόζεται και στον έλεγχο υποθέσεων. Ιδιαίτερα για το μοντέλο του Cox για μονοδιάστατη παράμετρο παλινδρόμησης, ο έλεγχος υποθέσεων με μηδενική υπόθεση $H_0: \boldsymbol{\beta} = \boldsymbol{\beta}_0$ και εναλλακτική υπόθεση $H_1: \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ εξετάζεται παρακάτω με τρείς τρόπους. Ο πρώτος είναι ο έλεγχος του λόγου πιθανοφάνειας που βασίζεται στην μερική πιθανοφάνεια. Οπότε υπό την μηδενική υπόθεση, ο έλεγχος έχει την μορφή:

$$T = -2\left(\log \frac{\boldsymbol{L(\beta_0)}}{\boldsymbol{L(\hat{\beta})}}\right)$$

Η ελεγχοσυνάρτηση αυτή, ασυμπτωτικά έχει κατανομή X^2 με ένα βαθμό ελευθερίας. Αν γίνεται έλεγχος από κοινού των μεταβλητών παλινδρόμησης, τότε ο έλεγχος υποθέσεων, ασυμπτωτικά έχει κατανομή X^2 με βαθμό ελευθερίας ίσο με τον αριθμό των παραμέτρων προς έλεγχο.
Παρόμοια, ένας άλλος έλεγχος υποθέσεων είναι ο Wald, με ελεγχοσυνάρτηση:

$$T = \left(\frac{\hat{\beta} - \beta_0}{se(\hat{\beta})}\right)^2$$

που επίσης έχει κατανομή X^2 με ένα βαθμό ελευθερίας, υπό την μηδενική υπόθεση. Το τυπικό σφάλμα της εκτιμήτριας της παραμέτρου βρίσκεται από τον αντίστροφο της μείον δεύτερης παραγώγου της συνάρτησης της λογαριθμισμένης μερικής πιθανοφάνειας.
Ένας τρίτος τρόπος είναι το score τεστ με ελεγχοσυνάρτηση:

$$T = -\frac{l_1^2}{l_2}$$

όπου

$$l_1 = \frac{\partial \log \boldsymbol{L(\beta)}}{\partial \beta}\Big|_{\beta=\hat{\beta}}$$

και

$$l_2 = \frac{\partial^2 \log \boldsymbol{L}(\boldsymbol{\beta})}{\partial \beta^2}|_{\beta=\hat{\beta}}$$

είναι η πρώτη και η δεύτερη παράγωγος του λογαρίθμου της συνάρτησης της μερικής πιθανοφάνειας. Όμοια όπως και οι δύο προηγούμενες ελεγχοσυναρτήσεις ακολουθεί ασυμπτωτικά, X^2 κατανομή με ένα βαθμό ελευθερίας.

1.3 Ευπάθεια

Σύμφωνα με τον Wienke (βλ. [14]) τα βασικά μοντέλα επιβίωσης ασχολούνται με την απλή περίπτωση δεδομένων που είναι ανεξάρτητα και ισόνομα κατανεμημένα. Αυτό βασίζεται στην υπόθεση ότι ο πληθυσμός που μελετούμε είναι ομοιογενής. Παρ'όλα αυτά όπως είναι κατανοητό τα άτομα του πληθυσμού να διαφέρουν αρκετά μεταξύ τους σε σχέση για παράδειγμα, με τις επιδράσεις ενός φαρμάκου, μιας θεραπείας, ή γενικότερα με την επίδραση διάφορων επεξηγηματικών μεταβλητών. Αυτή η ανομοιογένεια συχνά αναφέρεται ως μεταβλητότητα και αποτελεί μια από τις πιο σημαντικές ποσότητες με εφαρμογές στην επιδημιολογία, στην ιατρική, στην βιολογία. Αυτή η ανομοιογένεια συνήθως είναι δύσκολο να παρατηρηθεί, με αποτέλεσμα να έχουν κατασκευαστεί ιδιαίτερα μοντέλα που να μπορούν να περιγράψουν την μεταβλητότητα αυτή που ονομάζεται ευπάθεια (frailty). Η βασική ιδέα αυτών των μοντέλων βασίζεται στο γεγονός ότι τα άτομα του πληθυσμού έχουν διαφορετικές ευπάθειες, οπότε υποτίθεται ότι στους πιο ευπαθείς θα συμβεί το συμβάν προς ανάλυση, όπως για παράδειγμα ο θάνατος, πιο γρήγορα από τους λιγότερο ευπαθείς. Όταν εκτιμούνται τα ποσοστά θνησιμότητας γενικά ενδιαφερόμαστε για το πώς μεταβάλλονται σε σχέση με τον χρόνο ή την ηλικία. Μερικές φορές παρατηρείται κατά την αρχή της περιόδου της μελέτης ότι υπάρχει αύξηση των ποσοστών θνησιμότητας μέχρι ενός μεγίστου και μετά ύφεση. Αυτή η κατάσταση είναι κάτι σύνηθες, για παράδειγμα στα ποσοστά θανάτου καρκινοπαθών ασθενών. Δηλαδή, όσο περισσότερο ο ασθενής επιζεί αφότου έγινε η διάγνωση και η θεραπεία τόσο καλύτερη η πιθανότητα του να επιβιώσει την ασθένεια.

Πολύ συχνά ο κίνδυνος ενός πληθυσμού αρχίζει να μειώνεται επειδή απλώς τα άτομα υψηλού κινδύνου έχουν πεθάνει, ενώ παρ' όλα αυτά ο ατομικός κίνδυνος συνεχίζει να αυξάνεται. Στην ανάλυση ενός μοντέλου κινδύνου είναι πολύ δύσκολο να συμπεριληφθούν όλοι οι σημαντικοί παράγοντες κινδύνου αφού ο ερευνητής συνήθως δεν έχει την δυνατότητα να έχει όλες τις πληροφορίες σε ατομικό επίπεδο. Αυτό κυρίως ισχύει, για παράδειγμα στις έρευνες πληθυσμού όπου συνήθως οι μόνες γνωστές μεταβλητές είναι το φύλο και η ηλικία. Επίσης μπορεί να μην έχουμε καμία πληροφορία σχετικά με τον παράγοντα κινδύνου ή μπορεί ο υπολογισμός του παράγοντα να απαιτεί μεγάλο οικονομικό κόστος ή και μεγάλη χρονική διάρκεια. Σε τέτοιες περιπτώσεις παρατηρούνται δύο τύποι μεταβλητότητας στα δεδομένα. Ο πρώτος τύπος μεταβλητότητας αφορά τους παρατηρούμενους παράγοντες κινδύνου που περιλαμβάνονται στο μοντέλο και άρα θεωρητικά είναι προβλέψιμος. Ο δεύτερος τύπος μεταβλητότητας αφορά την ετερογένεια που προκαλείται από άγνωστες συμμεταβλητές και που θεωρητικά δεν είναι προβλέψιμος. Σύμφωνα με τον Hougaard (1991) η ετερογένεια μπορεί να εξηγήσει κάποια απροσδόκητα αποτελέσματα, όπως για παράδειγμα τις φθίνουσες συναρτήσεις κινδύνου. Αν κάποια άτομα έχουν υψηλότερο κίνδυνο αποτυχίας ή θανάτου τότε τα υπόλοιπα άτομα θα πρέπει είναι μια

ομάδα με χαμηλότερο κίνδυνο. Μια εκτίμηση του ατομικού κινδύνου χωρίς να συμπεριληφθεί η μη παρατηρούμενη ευπάθεια, θα έχει σαν αποτέλεσμα την υποεκτίμηση της συνάρτησης κινδύνου σε πολύ μεγαλύτερο βαθμό με την πάροδο του χρόνου.
Για αυτό το λόγο χρησιμοποιούνται τα μικτά μοντέλα, όπου ο πληθυσμός θεωρείται ότι είναι μια μίξη ατόμων με μερικώς άγνωστα, διαφορετικά ρίσκα. Τα μη παρατηρούμενα ρίσκα περιγράφονται από την μεταβλητή μίξης που αποτελεί την ευπάθεια στην ανάλυση επιβίωσης. Είναι μια τυχαία μεταβλητή που ακολουθεί κάποια κατανομή. Η ακριβής σχέση μεταξύ της γήρανσης των ατόμων και του πληθυσμού βασίζεται στην κατανομή της ευπάθειας μεταξύ των ατόμων. Με βάση την επιλογή της κατανομής της μεταβλητής της ευπάθειας, η διασπορά στην ευπάθεια καθορίζει τον βαθμό της μη παρατηρούμενης ετερογένειας και αποτελεί έναν δείκτη σημαντικών παραγόντων κινδύνου οι οποίοι λείπουν από το μοντέλο κινδύνου.

Για την αντιμετώπιση του προβλήματος της μη παρατηρούμενης ετερογένειας στους χρόνους συμβάντων που προκαλούνται από άγνωστες συμμεταβλητές, ο Beard (1959) και αργότερα οι Vaupel (1979) και Lancaster (1979) πρότειναν ένα μοντέλο τυχαίων επιδράσεων για τους χρόνους, ώστε να βελτιωθεί η προσαρμοστικότητα των μοντέλων κινδύνου στους πληθυσμούς.

Αναφερόμαστε αρχικά στο μοντέλο ευπάθειας χωρίς παρατηρούμενες συμμεταβλητές. Το πιο συχνά εφαρμοσμένο μοντέλο έχει την δομή ενός μοντέλου αναλογικών κινδύνων το οποίο εξαρτάται από την τυχαία μεταβλητή επίδρασης, την ευπάθεια. Πιο συγκεκριμένα, η συνάρτηση κινδύνου ενός ατόμου βασίζεται στην μη παρατηρούμενη, χρονικά ανεξάρτητη, τυχαία μεταβλητή Z, η οποία ονομάζεται ευπάθεια. Η συνάρτηση κινδύνου δεδομένης της μεταβλητής Z είναι η παρακάτω:

$$h(t|Z) = Zh_0(t) \tag{3}$$

όπου Z είναι μια μη αρνητική τυχαία μεταβλητή και h_0 η βασική συνάρτηση κινδύνου. Συνήθως, γίνεται κάποια κανονικοποίηση ώστε η κατανομή της ευπάθειας να έχει μέση τιμή $E(Z) = 1$. Εάν η διασπορά $\sigma^2 = V(Z)$, η οποία αποτελεί ένα μέγεθος μέτρησης της ετερογένειας στον πληθυσμό όταν είμαστε στην περίπτωση του βασικού κινδύνου, είναι μικρή τότε οι τιμές της Z είναι κοντά στο ένα. Εν αντιθέσει, αν η διασπορά είναι μεγάλη τότε οι τιμές της Z είναι πιο διασκορπισμένες προκαλώντας μεγαλύτερη ετερογένεια στον ατομικό κίνδυνο $Zh_0(t)$.
Το πρόβλημα που παρατηρείται σε μια έρευνα δεν είναι ο υπό συνθήκη κίνδυνος αλλά το συνολικό αποτέλεσμα όλων των ατόμων του πληθυσμού με διαφορετικές τιμές της τυχαίας μεταβλητής Z. Το μοντέλο στην σχέση (3) θεωρείται ένα σχετικά απλό μοντέλο ευπάθειας για την περιγραφή της δράσης της ετερογένειας. Σύμφωνα με το μοντέλο του Cox και την σχέση της μη παρατηρούμενης μεταβλητής Z, ο παρακάτω τύπος περιέχει και τις παρατηρούμενες συμμεταβλητές του μοντέλου:

$$h(t|\boldsymbol{X}, Z) = Zh_0(t)e^{\boldsymbol{\beta}'\boldsymbol{X}} \tag{4}$$

με $\boldsymbol{X} = (X_1, \dots, X_k)$ και $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)$ να αποτελούν τις συμμεταβλητές και τις παραμέτρους παλινδρόμησης αντίστοιχα. Όπως φαίνεται το μοντέλο ευπάθειας είναι μια γενίκευση του μοντέλου αναλογικών κινδύνων, το οποίο λαμβάνεται αν για την κατανομή της ευπάθειας ισχύει $Z = 1$ για όλα τα άτομα. Για το μοντέλο της σχέσης (3)

μπορούμε να υποθέσουμε ότι $S(t|Z)$ είναι η δεσμευμένη συνάρτηση επιβίωσης ενός ατόμου σε σχέση με την μεταβλητή ευπάθειας Z και θα έχει την μορφή:

$$S(t|Z) = e^{-\int_0^t h(S|Z)ds} = e^{-Z\int_0^t h_0(s)ds} = e^{-ZH_0(t)}$$

όπου $H_0(t) = \int_0^t h_0(s)ds$ αποτελεί την αθροιστική βασική συνάρτηση κινδύνου. Οι μορφές των παραπάνω εξισώσεων μελετούνται στο ατομικό επίπεδο, ενώ για το επίπεδο σε σχέση με τον πληθυσμό είναι αναγκαίο να θεωρήσουμε την αφαίρεση του όρου της ευπάθειας ολοκληρώνοντας. Η συνάρτηση επιβίωσης του πληθυσμού είναι ο σταθμισμένος μέσος των δεσμευμένων συναρτήσεων επιβίωσης με βάρη που υπολογίζονται από την συνάρτηση πυκνότητας πιθανότητας της κατανομής της ευπάθειας. Η συνάρτηση επιβίωσης του πληθυσμού λαμβάνεται από την δεσμευμένη συνάρτηση επιβίωσης $S(t|Z)$ αφαιρώντας τον όρο της ευπάθειας με ολοκλήρωση. Μπορεί να θεωρηθεί ως η μη δεσμευμένη συνάρτηση επιβίωσης ενός ατόμου τυχαία επιλεγμένου από τον πληθυσμό που μελετάμε και έχει μαθηματική μορφή όπως φαίνεται παρακάτω:

$$S(t) = \boldsymbol{E}S(t|Z) = \boldsymbol{E}\left[e^{-ZH_0(t)}\right] = \boldsymbol{L}(H_0(t))$$

όπου $\boldsymbol{L}_Z(s) = \boldsymbol{E}[e^{-Zs}] = \int_0^\infty e^{-sZ} f_Z(t)dt$, ο μετασχηματισμός Laplace και s μια μιγαδική μεταβλητή με μη αρνητικό πραγματικό μέρος. Οι παράγωγοι του μετασχηματισμού Laplace χρησιμοποιούνται στους παρακάτω υπολογισμούς για την κατανομή της ευπάθειας:

$$f(t) = -S'(t) = -[\boldsymbol{L}(H_0(t))]' = -h_0(t)\boldsymbol{L}'(H_0(t))$$

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

Από ιδιότητες του μετασχηματισμού Laplace ισχύει:

$$\boldsymbol{L}_{\boldsymbol{Z}}^{(k)}(s) = (-1)^k \boldsymbol{E}[Z^k e^{-sZ}] => \boldsymbol{E}[Z^k] = (-1)^k \boldsymbol{L}_{\boldsymbol{Z}}^{(k)}(0)$$

$$\boldsymbol{E}[Z] = -\boldsymbol{L}'(0)$$

$$\boldsymbol{V}(Z) = \boldsymbol{E}[Z^2] - (\boldsymbol{E}[Z])^2 = \boldsymbol{L}''(0) - (\boldsymbol{L}'(0))^2$$

όπου $\boldsymbol{L}'$ και $\boldsymbol{L}''$ αποτελούν την πρώτη και δεύτερη παράγωγο του μετασχηματισμού Laplace, αντίστοιχα. Σύμφωνα με τον Hougaard (1984,1986a,b), η χρήση του μετασχηματισμού Laplace για την εύρεση κατανομών κατάλληλων για την μεταβλητή ευπάθειας βοηθάει στην απλούστευση της παραμετρικής εκτίμησης. Επειδή συνήθως τα ευπαθή άτομα πεθαίνουν νωρίτερα, η κατανομή της ευπάθειας στον πληθυσμό αλλάζει με τον χρόνο.

Θεώρημα 1.1

Έστω το μοντέλο ευπάθειας της σχέσης (3) . Ο κίνδυνος του πληθυσμού $h(t) = \frac{f(t)}{s(t)}$ συμβολίζεται και ως $h(t) = \boldsymbol{E}[h(t|z)|T > t]$ ή πιο συγκεκριμένα:

$$h(t) = \int_0^\infty h(t|z) f(z|T > t) dz = h_0(t) \int_0^\infty z f(z|T > t) dz$$

όπου $f(z|T > t)$ είναι η πυκνότητα πιθανότητας της ευπάθειας μεταξύ των επιζώντων μέχρι και τον χρόνο t.

Απόδειξη

Σύμφωνα με την σχέση (3) γνωρίζουμε ότι ισχύουν τα παρακάτω:

$$h(t|Z) = \frac{f(t|z)}{S(t|z)} = z h_0(t) => f(t|z) = z h_0(t) S(t|z)$$

$$f(t, z) = z h_0(t) S(t|z) f_Z(z)$$

Άρα η μη δεσμευμένη συνάρτηση κατανομής δίνεται από την παρακάτω σχέση:

$$f(t) = h_0(t) \int_0^\infty z S(t|z) f_Z(z) dz \qquad (5)$$

όπου f_Z είναι η συνάρτηση πυκνότητας πιθανότητας της κατανομής της ευπάθειας. Οπότε χρησιμοποιώντας την σχέση (5) στην σχέση του κινδύνου του πληθυσμού θα ισχύει:

$$h(t) = \frac{f(t)}{S(t)} = \frac{h_0(t) \int_0^\infty z S(t|z) f_Z(z) dz}{S(t)}$$

Επειδή η επιβίωση ως και τον χρόνο t υπονοεί χρόνο θανάτου μεγαλύτερο από t θα ισχύει:

$$f(z, T > t) = \int_t^\infty f(z, s) ds = f_Z(z) \int_t^\infty z h_0(s) S(s|z) ds = f_Z(z) S(t|z)$$

$$f(z|T > t) = \frac{f_Z(z) S(t|z)}{S(t)}$$

Οπότε και ολοκληρώνεται η απόδειξη.

□

Με αποτέλεσμα ο κίνδυνος του πληθυσμού να θεωρείται ότι είναι ο σταθμισμένος μέσος των ατομικών κινδύνων των επιζώντων. Τα βάρη καθορίζονται από την κατανομή της ευπάθειας. Το αποτέλεσμα της σχέσης δείχνει ότι ο κίνδυνος των ατόμων αυξάνει πιο γρήγορα σε σύγκριση με τον κίνδυνο της ομάδας που τα άτομα ανήκουν.

Δηλαδή σύμφωνα με τον Vaupel et al. (1979) και τους Manton και Stallard (1981) ο κίνδυνος του πληθυσμού δεν αντιπροσωπεύει τον κίνδυνο των ατόμων από αυτό τον πληθυσμό. Για αυτό για την εύρεση του κινδύνου των ατόμων χρησιμοποίησαν την μεταβλητή της ευπάθειας.

1.3.1 Μοντέλα Μετασχηματισμού-Γενική κλάση μοντέλων τυχαίων επιδράσεων

Από την συνάρτηση κινδύνου της σχέσης (4) για την περίπτωση της ευπάθειας με συμμεταβλητές μπορούν να υπολογιστούν η αθροιστική συνάρτηση κινδύνου και η δεσμευμένη συνάρτηση επιβίωσης όπως παρουσιάζεται παρακάτω:

$$H(t|\boldsymbol{X},Z)=\int_0^t h(t|\boldsymbol{X},Z)dt=\int_o^t Zh_0(t)e^{\boldsymbol{\beta}'\boldsymbol{X}}dt=ZH_0(t)e^{\boldsymbol{\beta}'\boldsymbol{X}}$$

και

$$S(t|\boldsymbol{X},Z)=e^{-H(t|\boldsymbol{X},Z)}=e^{-ZH_0(t)e^{\boldsymbol{\beta}'\boldsymbol{X}}}$$

Από τις σχέσεις αυτές μπορούν να υπολογιστούν οι ακόλουθες συναρτήσεις $S(t|\boldsymbol{X})$ και $F(t|\boldsymbol{X})$ ως εξής:

$$S(t|\boldsymbol{X})=\int_0^\infty S(t|\boldsymbol{X},y)dF_Z(y)=\int_0^\infty e^{-yH_0(t)e^{\boldsymbol{\beta}'\boldsymbol{X}}}dF_Z(y)=e^{-G(H_0(t)e^{\boldsymbol{\beta}'\boldsymbol{X}})} \qquad (6)$$

και

$$F(t|\boldsymbol{X})=1-S(t|\boldsymbol{X})=1-e^{-G(H_0(t)e^{\boldsymbol{\beta}'\boldsymbol{X}})} \qquad (7)$$

όπου $F_Z(.)$ είναι η αθροιστική συνάρτηση κατανομής (cdf) της θετικής μεταβλητής της ευπάθειας Z (βλ. [11]). Η συνάρτηση G ορίζεται ως εξής:

$$G(w)=-\ln\left(\int_0^\infty e^{-yw}\,dF_Z(y)\right)$$

Παρατηρούμε ότι η συνάρτηση G είναι ίση με $G(x)=-\ln(\boldsymbol{L}(x))$, όπου $\boldsymbol{L}(x)$ είναι ο μετασχηματισμός Laplace της συνάρτησης κατανομής της ευπάθειας. Επειδή η συνάρτηση κατανομής της μεταβλητής της ευπάθειας είναι γνωστή συνεπάγεται ότι και η συνάρτηση G θα είναι γνωστή. Η κλάση που περιέχει μοντέλα όπως αυτό της σχέσης (7) αποτελεί την κλάση των μοντέλων μετασχηματισμού.

Οι διάφορες συναρτήσεις που μπορούν να δημιουργηθούν από την G συνάρτηση μας δίνουν γνωστά μοντέλα όπως αυτό των Clayton-Cuzick (1986) το οποίο λαμβάνεται όταν η μεταβλητή της ευπάθειας ακολουθεί την Γάμμα κατανομή με $\Gamma(1/c,1/c),\ c>0$. Για αυτή την περίπτωση θα ισχύει $G(x,c)=\ln(1+cx)/c$. Άλλες γνωστές μορφές μοντέλων ευπάθειας είναι για $G(x)=x$ που αφορά το μοντέλο αναλογικών κινδύνων του Cox, καθώς και ισχύει $G(x,b)=\sqrt{4b(b+x)}-2b, b>0$ όταν η μεταβλητή της ευπάθειας ακολουθεί την Αντίστροφη Γκαουσιανή κατανομή $IG(b,b)$. Η συνάρτηση G θεωρείται ότι είναι τρεις φορές παραγωγίσιμη, αυστηρώς αύξουσα, κοίλη συνάρτηση στο $(0,\infty)$ για την οποία ισχύει $G(0)=0$ και $G(\infty)=\infty$ (βλ. [5], [9], [11]). Για το

ημιπαραμετρικό μοντέλο του Cox χωρίς ετερογένεια γνωρίζουμε ότι ισχύουν οι σχέσεις:

$$h(t|\boldsymbol{X}) = h_0(t)e^{\boldsymbol{\beta}'\boldsymbol{X}}$$

και

$$S(t|\boldsymbol{X}) = e^{-e^{\boldsymbol{\beta}'\boldsymbol{X}}H_0(t)}$$

Τότε θα ισχύει:

$$\log\left[-\log\left(S(t|\boldsymbol{X})\right)\right] = \log(H_0(t)) + \boldsymbol{\beta}'\boldsymbol{X}$$

όπου $\log(H_0(t))$ είναι μια αυστηρά αύξουσα συνάρτηση. Ομοίως το μοντέλο ευπάθειας της σχέσης (6) για δεξιά λογοκριμένα δεδομένα μπορεί να γραφτεί ως ένα μοντέλο μετασχηματισμού στην παρακάτω μορφή όπως ορίζεται από τους Cheng και συν. (βλ. [3]):

$$g\left(S(t|\boldsymbol{X})\right) = \log\left\{G^{-1}\left(-\log\left(S(t|\boldsymbol{X})\right)\right)\right\} = \log H_0(t) + \boldsymbol{\beta}'\boldsymbol{X}$$

Η $g(.)$ είναι μια γνωστή φθίνουσα συνάρτηση.

Παρακάτω παρουσιάζονται αναλυτικά κάποιες συχνά χρησιμοποιούμενες κατανομές που επιλέγονται ως κατανομές για την μεταβλητή ευπάθειας και είναι επιλεγμένες ώστε στον υπολογισμό της πιθανοφάνειας να είναι σχετικά εύκολη η ολοκλήρωση και να είναι εύκολη η μελέτη του μοντέλου ευπάθειας.

Γάμμα Μοντέλο Ευπάθειας

Η Γάμμα κατανομή εφαρμόζεται ιδιαίτερα σε δεδομένα αποτυχίας ή θανάτου. Είναι σχετικά εύκολο να υπολογιστούν οι κλειστές μορφές των παραστάσεων των μη δεσμευμένων συναρτήσεων επιβίωσης, αθροιστικής κατανομής και κινδύνου με την χρήση του μετασχηματισμού Laplace. Λόγω της ευκολίας στον υπολογισμό του μετασχηματισμού Laplace, αποτελεί μια από τις πιο συχνά χρησιμοποιούμενες κατανομές για την μοντελοποίηση μη αρνητικών μεταβλητών, όπως η ευπάθεια. Η συνάρτηση πυκνότητας πιθανότητας και ο αντίστοιχος μετασχηματισμός Laplace της κατανομής είναι οι παρακάτω:

$$f(z) = \frac{1}{\Gamma(k)}\lambda^k z^{k-1}e^{-\lambda z}, \qquad \lambda, \kappa > 0$$

$$\boldsymbol{L}(u) = \frac{1}{\Gamma(k)}\lambda^k \int_0^\infty z^{k-1}e^{-\lambda z}e^{-uz}dz =$$

$$= \frac{\lambda^k}{(\lambda+u)^k}\frac{1}{\Gamma(k)}(\lambda+u)^k \int_0^\infty z^{k-1}e^{-(\lambda+u)z}dz = \left(1+\frac{u}{\lambda}\right)^{-k}$$

$$\boldsymbol{L}'(u) = -\frac{k}{\lambda}\left(1+\frac{u}{\lambda}\right)^{-k-1}$$

$$\boldsymbol{L}''(u) = \frac{k(k+1)}{\lambda^2}\left(1+\frac{u}{\lambda}\right)^{-k-2}$$

Για $u = 0$ θα ισχύει

$$\boldsymbol{E}[Z] = -\boldsymbol{L}'(0) = \frac{k}{\lambda}$$

$$\boldsymbol{V}(Z) = \boldsymbol{E}[Z^2] - (\boldsymbol{E}[Z])^2 = \boldsymbol{L}''(0) - \big(\boldsymbol{L}'(0)\big)^2 = \frac{k(k+1)}{\lambda^2} - \frac{k^2}{\lambda^2} = \frac{k}{\lambda^2}$$

Λόγω της ευπάθειας θέλουμε να ισχύει $\boldsymbol{E}Z = 1$, άρα χρησιμοποιείται για το μοντέλο της Γάμμα κατανομής ο περιορισμός $k = \lambda$. Ορίζεται, επίσης η διασπορά ως $\sigma^2 = \frac{1}{\lambda}$ και επομένως η συνάρτηση πυκνότητας της μεταβλητής ευπάθειας $Z \sim \Gamma(\frac{1}{\sigma^2}, \frac{1}{\sigma^2})$ θα είναι:

$$f(z) = \frac{1}{\Gamma(\frac{1}{\sigma^2})}\left(\frac{1}{\sigma^2}\right)^{\frac{1}{\sigma^2}} z^{\frac{1}{\sigma^2}-1} e^{-\frac{z}{\sigma^2}}$$

Επίσης η μη δεσμευμένη συνάρτηση επιβίωσης σύμφωνα με τον μετασχηματισμό Laplace θα είναι:

$$S(t) = \boldsymbol{L}\big(H_0(t)\big) = \frac{1}{(1+\sigma^2 H_0(t))^{\frac{1}{\sigma^2}}}$$

και

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

Και η συνάρτηση κινδύνου θα είναι:

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

Αντίστροφο Γκαουσιανό Μοντέλο Ευπάθειας

Η αντίστροφη Γκαουσιανή κατανομή προτάθηκε σαν εναλλακτική κατανομή για την ευπάθεια από τον Hougaard (1984). Παρουσιάζει και αυτή η κατανομή σχετική ευκολία στον υπολογισμό των μη δεσμευμένων συναρτήσεων επιβίωσης και κινδύνου. Η συνάρτηση πυκνότητας πιθανότητας μιας τυχαίας μεταβλητής με παραμέτρους $\mu, \lambda > 0$ δίνεται από την σχέση:

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

Ο μετασχηματισμός Laplace της κατανομής θα είναι:

$$\boldsymbol{L}(u) = \boldsymbol{E}e^{-uz} = \int_0^\infty \frac{\sqrt{\lambda}}{\sqrt{2\pi z^3}} e^{-\frac{\lambda}{2\mu^2 z}(z-\mu)^2} e^{-uz} dz$$

$$= \int_0^\infty \frac{\sqrt{\lambda}}{\sqrt{2\pi z^3}} e^{-\frac{(\lambda+2\mu^2 u)z^2 - 2\mu\lambda z + \lambda\mu^2}{2\mu^2 z}} dz$$

$$= \int_0^\infty \frac{\sqrt{\lambda}}{\sqrt{2\pi z^3}} e^{-\frac{z}{2}\frac{\lambda+2\mu^2 u}{\mu^2} + \frac{\lambda}{\mu} - \frac{\lambda}{2z}} dz$$

$$= e^{-\frac{\lambda\sqrt{1+\frac{2\mu^2 u}{\lambda}}}{\mu} + \frac{\lambda}{\mu}} \int_0^\infty \frac{\sqrt{\lambda}}{\sqrt{2\pi z^3}} e^{-\frac{\lambda z}{2}\frac{1+\frac{2\mu^2 u}{\lambda}}{\mu^2} + \frac{\lambda\sqrt{1+\frac{2\mu^2 u}{\lambda}}}{\mu} - \frac{\lambda}{2z}} dz \qquad (8)$$

Όμως λόγω της παρακάτω σχέσης:

$$-\frac{\lambda z}{2}\frac{1+\frac{2\mu^2 u}{\lambda}}{\mu^2} + \frac{\lambda\sqrt{1+\frac{2\mu^2 u}{\lambda}}}{\mu} - \frac{\lambda}{2z} = -\frac{\lambda}{2\frac{\mu^2}{1+\frac{2\mu^2 u}{\lambda}}z}\left(z - \frac{\mu}{\sqrt{1+\frac{2\mu^2 u}{\lambda}}}\right)^2$$

η συνάρτηση μέσα στο ολοκλήρωμα στη σχέση (8) γράφεται ως εξής:

$$\frac{\sqrt{\lambda}}{\sqrt{2\pi z^3}} e^{-\frac{\lambda z}{2}\frac{1+\frac{2\mu^2 u}{\lambda}}{\mu^2} + \frac{\lambda\sqrt{1+\frac{2\mu^2 u}{\lambda}}}{\mu} - \frac{\lambda}{2z}} = \frac{\sqrt{\lambda}}{\sqrt{2\pi z^3}} e^{-\frac{\lambda}{2\frac{\mu^2}{1+\frac{2\mu^2 u}{\lambda}}z}\left(z-\frac{\mu}{\sqrt{1+\frac{2\mu^2 u}{\lambda}}}\right)^2}$$

και άρα αποτελεί την πυκνότητα πιθανότητας της Αντίστροφης Γκαουσιανής κατανομής με παραμέτρους $\frac{\mu}{\sqrt{1+\frac{2\mu^2 u}{\lambda}}}$ και λ. Οπότε το ολοκλήρωμα στην (8) ισούται με την μονάδα. Άρα ο μετασχηματισμός Laplace για μια τυχαία μεταβλητή που ακολουθεί την Αντίστροφη Γκαουσιανή κατανομή θα είναι:

$$\boldsymbol{L}(u) = e^{-\frac{\lambda\sqrt{1+\frac{2\mu^2 u}{\lambda}}}{\mu} + \frac{\lambda}{\mu}} = e^{\frac{\lambda}{\mu}\left(1-\sqrt{1+\frac{2\mu^2 u}{\lambda}}\right)}$$

Επίσης θα ισχύει τότε:

$$\boldsymbol{L}'(u) = -\frac{\mu}{\sqrt{1+\frac{2\mu^2 u}{\lambda}}} e^{\frac{\lambda}{\mu}\left(1-\sqrt{1+\frac{2\mu^2 u}{\lambda}}\right)}$$

$$\boldsymbol{L}''(u) = \frac{\mu^3}{\lambda\left(1+\frac{2\mu^2 u}{\lambda}\right)^{\frac{3}{2}}} e^{\frac{\lambda}{\mu}\left(1-\sqrt{1+\frac{2\mu^2 u}{\lambda}}\right)} + \frac{\mu^2}{1+\frac{2\mu^2 u}{\lambda}} e^{\frac{\lambda}{\mu}\left(1-\sqrt{1+\frac{2\mu^2 u}{\lambda}}\right)}$$

Άρα η αναμενόμενη τιμή και η διασπορά της κατανομής της ευπάθειας θα είναι για $u = 0$:

$$\boldsymbol{E}Z = -\boldsymbol{L}'(0) = \mu$$

$$\boldsymbol{V}(Z) = \boldsymbol{L}''(0) - \left(\boldsymbol{L}'(0)\right)^2 = \frac{\mu^3}{\lambda}$$

Τότε με τον περιορισμό για την ευπάθεια $\boldsymbol{E}Z = \mu = 1$ ισχύει $\boldsymbol{V}(Z) = \sigma^2 = 1/\lambda$. Άρα τελικά ο μετασχηματισμός Laplace απλοποιείται στην μορφή:

$$\boldsymbol{L}(u) = e^{\frac{1}{\sigma^2}\left(1-\sqrt{1+2\sigma^2 u}\right)}$$

Οι πηγές που χρησιμοποιήθηκαν σε αυτό το κεφάλαιο αφορούν κυρίως τα [5], [9], [11], [14] στην βιβλιογραφία.

ΚΕΦΑΛΑΙΟ 2

ΜΕΤΡΑ ΑΠΟΚΛΙΣΗΣ : ΟΡΙΣΜΟΙ ΚΑΙ ΙΔΙΟΤΗΤΕΣ

2.1 Βασικές Έννοιες Θεωρίας Πιθανοτήτων

Σύμφωνα με τον Pardo (βλ. [8]), έστω $\boldsymbol{X}$ τυχαία μεταβλητή που παίρνει τιμές σε ένα δειγματικό χώρο X που αποτελεί συνήθως, ένα υποσύνολο του R^n. Υποθέτουμε ότι η συνάρτηση κατανομής F του $\boldsymbol{X}$ εξαρτάται από ένα συγκεκριμένο αριθμό παραμέτρων και υποθέτουμε επιπλέον ότι η συναρτησιακή μορφή της F είναι γνωστή εκτός ίσως από ένα πεπερασμένο αριθμό αυτών των παραμέτρων. Έστω $\boldsymbol{\theta}$ το διάνυσμα των άγνωστων παραμέτρων που σχετίζεται με την F. Έστω $(X, \beta_X, P_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta}$ αποτελεί ένα στατιστικό χώρο που σχετίζεται με την τυχαία μεταβλητή $\boldsymbol{X}$, όπου β_X είναι η σ-άλγεβρα των Borel υποσυνόλων A⊂X και $\{P_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\Theta}$ είναι μια οικογένεια από κατανομές πιθανότητας που ορίζονται στο μετρικό χώρο (X, β_X) με Θ ένα ανοιχτό υποσύνολο του $R^{M_0},\ M_0 \geq 1$.

Υποθέτουμε ότι οι πιθανότητες κατανομής P_θ είναι απόλυτα συνεχείς σε σχέση με ένα σ-πεπερασμένο μέτρο μ στον χώρο (X, β_X). Το μ υποθέτουμε ότι είναι είτε το μέτρο Lebesgue ή ένα απαριθμητικό μέτρο (δηλαδή υπάρχει ένα πεπερασμένο ή αριθμήσιμο σύνολο S_X με την ιδιότητα $P_\theta(X - S_X) = 0$). Η παρακάτω σχέση,

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{dP_{\boldsymbol{\theta}}}{d\mu}(\boldsymbol{x}) = \begin{cases} f_\theta(\boldsymbol{x}), & \text{αν } \mu \text{ είναι το μέτρο Lebesgue} \\ \mathrm{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}), & \text{αν } \mu \text{ είναι ένα απαριθμητικό μέτρο } (\boldsymbol{x} \in S_x) \end{cases}$$

αποτελεί την οικογένεια συναρτήσεων πυκνότητας πιθανότητας αν το μ είναι το μέτρο Lebesgue ή αποτελεί την οικογένεια συναρτήσεων μάζας πιθανότητας αν το μ είναι απαριθμητικό μέτρο. Στην πρώτη περίπτωση το $\boldsymbol{X}$ είναι μια τυχαία μεταβλητή με απόλυτα συνεχή κατανομή και στην δεύτερη περίπτωση είναι μια διακριτή τυχαία μεταβλητή με στήριγμα S_X.

Έστω h μετρήσιμη συνάρτηση. Η αναμενόμενη τιμή της $h(\boldsymbol{X})$ δίνεται από την σχέση:

$$E_{\boldsymbol{\theta}}[h(\boldsymbol{X})] = \begin{cases} \int_X h(\boldsymbol{x}) f_{\boldsymbol{\theta}}(\boldsymbol{x}) d\boldsymbol{x}, & \alpha\nu\ \mu\ \epsilon\acute{\iota}\nu\alpha\iota\ \mu\acute{\epsilon}\tau\rho o\ Lebesgue \\ \sum_{x \in S_X} h(\boldsymbol{x}) p_{\boldsymbol{\theta}}(\boldsymbol{x}), & \alpha\nu\ \mu\ \epsilon\acute{\iota}\nu\alpha\iota\ \acute{\epsilon}\nu\alpha\ \alpha\pi\alpha\rho\iota\theta\mu\eta\tau\iota\kappa\acute{o}\ \mu\acute{\epsilon}\tau\rho o \end{cases}$$

Για την μελέτη στατιστικών προβλημάτων, καθώς και της οικογένειας μέτρων απόκλισης που θα αναλυθούν παρακάτω, δίνονται οι ορισμοί δυο σημαντικών αποστάσεων της θεωρίας πιθανοτήτων, αυτές των Kolmogorov και Lévy. Έστω δύο μέτρα πιθανότητας $P_{\boldsymbol{\theta_1}}$ *και* $P_{\boldsymbol{\theta_2}}$ με τις αντίστοιχες τους μονοδιάστατες συναρτήσεις κατανομής $F_{\boldsymbol{\theta_1}}$ *και* $F_{\boldsymbol{\theta_2}}$. Η απόσταση Kolmogorov που εισήγαγε ο Kolmogorov το 1933 μεταξύ των $F_{\boldsymbol{\theta_1}}$ *και* $F_{\boldsymbol{\theta_2}}$(ή μεταξύ των $P_{\boldsymbol{\theta_1}}$ *και* $P_{\boldsymbol{\theta_2}}$) δίνεται από την σχέση:

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

Στην παραπάνω απόσταση Kolmogorov στηρίζεται το θεώρημα Glivenko-Cantelli που ορίζει ότι η συνάρτηση εμπειρικής κατανομής είναι μια ισχυρά συνεπής ομοιόμορφη εκτιμήτρια της πραγματικής συνάρτησης κατανομής. Δηλαδή, δοθέντος ενός τυχαίου δείγματος $X_1, X_2, \dots, X_n$ από ένα πληθυσμό με συνάρτηση κατανομής $F_{\boldsymbol{\theta_0}}$ για $\varepsilon > 0$ θα ισχύει:

$$\lim_{n \to \infty} \mathbf{P}\{K_1(F_n, F_{\boldsymbol{\theta_0}}) > \varepsilon\} = \mathbf{0}$$

όπου F_n είναι η εμπειρική συνάρτηση κατανομής και έχει την μορφή:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(x_i)$$

όπου I_A είναι η δείκτρια συνάρτηση του συνόλου Α.
Ομοίως, η απόσταση Lévy (1925) δίνεται από την σχέση:

$$K_2(F_{\boldsymbol{\theta_1}}, F_{\boldsymbol{\theta_2}}) = inf\{\varepsilon > 0 \colon F_{\boldsymbol{\theta_1}}(x - \varepsilon) \leq F_{\boldsymbol{\theta_2}}(x) \leq F_{\boldsymbol{\theta_1}}(x + \varepsilon), \gamma\iota\alpha\ \acute{o}\lambda\alpha\ \tau\alpha\ x\}$$

με τιμές στο [0,1] διάστημα.

2.2 φ-μέτρα Απόκλισης μεταξύ δύο κατανομών πιθανότητας

Το μέτρο απόκλισης Kullback-Leibler ανάμεσα σε δύο κατανομές πιθανότητας $P_{\boldsymbol{\theta_1}}$ *και* $P_{\boldsymbol{\theta_2}}$ προτάθηκε από τους Kullback και Leibler το 1951 και δίνεται από την σχέση:

$$D_{Kull}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \int_X f_{\boldsymbol{\theta_1}}(\boldsymbol{x}) \log \frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})} d\mu(\boldsymbol{x}) = E_{\boldsymbol{\theta_1}} \left[\log \left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{X})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{X})} \right) \right] \qquad (9)$$

Ο Jeffreys το 1946 χρησιμοποίησε μια συμμετρική μορφή του παραπάνω τύπου ως ένα μέτρο απόκλισης μεταξύ δύο κατανομών πιθανότητας. Αυτό το μέτρο απόκλισης καλείται και ως J-απόκλιση και ορίζεται από την σχέση:

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

Ο Rényi (1961) παρουσίασε την πρώτη παραμετρική γενίκευση της σχέσης (9):

$$D_r^1(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \frac{1}{r-1} \log \int_X f_{\boldsymbol{\theta_1}}(\boldsymbol{x})^r f_{\boldsymbol{\theta_2}}(\boldsymbol{x})^{1-r} d\mu(\boldsymbol{x}) =$$
$$= \frac{1}{r-1} \log E_{\boldsymbol{\theta_1}} \left[\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{X})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{X})} \right)^{r-1} \right], r > 0, r \neq 1$$

Αργότερα οι Liese και Vajda (1987) επέκτειναν την γενίκευση του Rényi για όλα τα $r \neq 1, 0$ ως εξης:

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

$$= \frac{1}{r(r-1)} \log E_{\theta_1} \left[\left(\frac{f_{\theta_1}(\boldsymbol{X})}{f_{\theta_2}(\boldsymbol{X})} \right)^{r-1} \right], \; r \neq 0, 1.$$

Η σχέση αυτή αναφέρεται και ως απόκλιση Rényi. Οι περιπτώσεις $r = 1$ και $r = 0$ ορίζονται ως εξής:

$$D_1^1(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \lim_{r \to 1} D_r^1(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = D_{Kull}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$$

και

$$D_0^1(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \lim_{r \to 0} D_r^1(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = D_{Kull}(\boldsymbol{\theta_2}, \boldsymbol{\theta_1})$$

Το μέτρο απόκλισης $D_{Kull}(\boldsymbol{\theta_2}, \boldsymbol{\theta_1})$ ορίζεται ως η ελάχιστη πληροφορία διάκρισης (Minimum discrimination information) μεταξύ των κατανομών πιθανότητας $P_{\boldsymbol{\theta_1}}$ *και* $P_{\boldsymbol{\theta_2}}$. Άλλες δύο πολύ γνωστές παραμετρικές γενικεύσεις της σχέσης (9) είναι τα r-τάξης και s-βαθμού μέτρα απόκλισης και τα 1ης τάξης και s-βαθμού μέτρα απόκλισης από τους Sharma και Mittal (1977) που δίνονται από την σχέση:

$$D_r^s(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \frac{1}{(s-1)} \left(\left(\int_X f_{\boldsymbol{\theta_1}}(\boldsymbol{x})^r f_{\boldsymbol{\theta_2}}(\boldsymbol{x})^{1-r} d\mu(\boldsymbol{x}) \right)^{\frac{s-1}{r-1}} - 1 \right) =$$

$$= \frac{1}{(s-1)} \left(\left(E_{\boldsymbol{\theta_1}} \left[\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{X})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{X})} \right)^{r-1} \right] \right)^{\frac{s-1}{r-1}} - 1 \right), \qquad r, s \neq 1$$

και

$$D_1^s(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \frac{1}{(s-1)} \left(exp \left((s-1) \int_X f_{\boldsymbol{\theta_1}}(\boldsymbol{x}) \log \frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})} d\mu(\boldsymbol{x}) \right) - 1 \right)$$

$$= \frac{1}{(s-1)}\left(exp\left((s-1)E_{\boldsymbol{\theta_1}}\left[\log\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{X})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{X})}\right)\right]\right)-1\right), \qquad s \neq 1$$

Ορισμός 2.1

Το φ-μέτρο απόκλισης ανάμεσα στις κατανομές πιθανότητας των $P_{\boldsymbol{\theta_1}}$ *και* $P_{\boldsymbol{\theta_2}}$ ορίζεται ως:

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

$$= E_{\boldsymbol{\theta_2}}\left[\varphi\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{X})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{X})}\right)\right], \varphi \in \boldsymbol{\Phi}^* \qquad (10)$$

όπου $\boldsymbol{\Phi}^*$ είναι η κλάση όλων των κυρτών συναρτήσεων $\varphi(x), x \geq 0$ τέτοια ώστε για $x = 1,\ \varphi(x) = 0$ και για $x = 0,\ 0\varphi\,(0/0) = 0$ και $0\varphi(p/0) = \lim_{u\to\infty} \varphi(u)/u$.

Παρατήρηση 2.1

Έστω $\varphi \in \boldsymbol{\Phi}^*$ διαφορίσιμη στο $x = 1$ τότε η συνάρτηση

$$\psi(x) \equiv \varphi(x) - \varphi'(1)(x-1) \qquad (11)$$

ανήκει επίσης στο $\boldsymbol{\Phi}^*$ και έχει την επιπλέον ιδιότητα ότι $\psi'(1) = 0$. Αυτή η ιδιότητα σε συνδυασμό με την κυρτότητα συνεπάγεται ότι $\psi(x) \geq 0\ \gamma\iota\alpha\ \kappa\acute{\alpha}\theta\varepsilon\ x \geq 0$. Επιπλέον, ισχύει:

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

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

Αφού τα δύο μέτρα απόκλισης συμπίπτουν μπορεί να θεωρηθεί ότι το σύνολο $\boldsymbol{\Phi}^*$ είναι ισοδύναμο με το σύνολο

$$\boldsymbol{\Phi} \equiv \boldsymbol{\Phi}^* \cap \{\varphi: \varphi'(1) = 0\}$$

Το μέτρο απόκλισης Kullback-Leibler υπολογίζεται από την σχέση:

$$\psi(x) \equiv x\log x - x + 1 \ \text{ή}\ \ \varphi(x) = x\log x$$

και άρα

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

Θεωρούμε ότι κάθε συνάρτηση φ ανήκει στο $\boldsymbol{\Phi}$ ή $\boldsymbol{\Phi}^*$. Στον παρακάτω πίνακα παρουσιάζονται κάποια σημαντικά μέτρα απόκλισης που αποτελούν περιπτώσεις της $\varphi - \alpha\pi\acute{o}\kappa\lambda\iota\sigma\eta\varsigma$.

$\varphi - συνάρτηση$	Απόκλιση
$x\log x - x + 1$	Kullback-Leibler (1959)
$-\log x + x - 1$	Πληροφορία Ελάχιστης Διάκρισης
$(x-1)\log x$	J-απόκλιση
$\frac{1}{2}(x-1)^2$	Pearson (1900), Kagan (1963)
$\frac{(x-1)^2}{(x+1)^2}$	Balakrishnan και Sanghvi (1968)
$\frac{-x^s + s(x-1) + 1}{1-s}, \quad s \neq 1$	Rathie και Kannappan (1972)
$\frac{1-x}{2} - \left(\frac{1+x^{-r}}{2}\right)^{\frac{-1}{r}}, \quad r > 0$	Αρμονικός Μέσος (Mathai και Rathie (1975))
$\frac{(1-x)^2}{2(a+(1-a)x)}, \quad 0 \leq a \leq 1$	Rukhin (1994)
$\frac{ax\log x - (ax+1-a)\log(ax+1-a)}{a(1-a)}, \quad a \neq 0,1$	Lin (1991)
$\frac{x^{\lambda+1} - x - \lambda(x-1)}{\lambda(\lambda+1)}, \quad \lambda \neq 0, -1$	Cressie και Read (1984)
$\lvert 1 - x^a \rvert^{1/a}, \quad 0 < a < 1$	Matusita (1964)
$\lvert 1 - x^a \rvert^a, \quad a \geq 1$	X- απόκλιση τάξης α (Vajda (1973)) Απόκλιση Ολικής Μεταβολής αν α=1 (Saks 1937)

Στην στατιστική μια πολύ σημαντική οικογένεια φ-αποκλίσεων είναι των Cressie και Read (1984). Ονομάζεται και ως οικογένεια απόκλισης δύναμης και ορίζεται ως εξής:

$$I_\lambda(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \equiv D_{\varphi_{(\lambda)}}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \frac{1}{\lambda(\lambda+1)}\left(\int_X \frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})^{\lambda+1}}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})^{\lambda}} d\mu(\boldsymbol{x}) - 1\right) =$$

$$= \frac{1}{\lambda(\lambda+1)}\left(E_{\boldsymbol{\theta_1}}\left[\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{X})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{X})}\right)^{\lambda}\right] - 1\right) \; για - \infty < \lambda < \infty.$$

Η οικογένεια απόκλισης δύναμης δεν ορίζεται για $\lambda = -1$ ή $\lambda = 0$. Παρ'όλα αυτά αν ορίσουμε τις περιπτώσεις αυτές σύμφωνα με τα συνεχή όρια του $I_\lambda(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ καθώς $\lambda \to -1 \; και \; \lambda \to 0$, τότε το $I_\lambda(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ είναι συνεχές στο λ. Άρα θα ισχύει ότι:

$$\lim_{\lambda \to 0} I_\lambda(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = D_{Kull}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$$

και

$$\lim_{\lambda \to -1} I_\lambda(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = D_{Kull}(\boldsymbol{\theta_2}, \boldsymbol{\theta_1})$$

Η οικογένεια απόκλισης δύναμης δίνεται από την σχέση (10) ως εξής:

$$\varphi(x) = \begin{cases} \varphi_{(\lambda)}(x) = \dfrac{1}{\lambda(\lambda+1)}\left(x^{\lambda+1} - x - \lambda(x-1)\right), & \lambda \neq 0, -1 \\ \varphi_{(0)}(x) = \lim\limits_{\lambda \to 0} \varphi_{(\lambda)}(x) = x\log x - x + 1 \\ \varphi_{(-1)}(x) = \lim\limits_{\lambda \to -1} \varphi_{(\lambda)}(x) = -\log x + x - 1 \end{cases}$$

Η οικογένεια απόκλισης δύναμης προτάθηκε ανεξάρτητα από τους Liese και Vajda (1987) ως μια φ-απόκλιση, την οποία ονόμασαν I_α-απόκλιση. Χρησιμοποιήθηκε ιδιαίτερα από τους Cressie και Read για διακριτές τυχαίες μεταβλητές με πεπερασμένο στήριγμα.

Τα μέτρα απόκλισης των Rényi και Sharma και Mittal καθώς και το μέτρο απόκλισης από τον Bhattacharyya (1943) με σχέση:

$$B(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = -\log \int_X \sqrt{f_{\boldsymbol{\theta_1}}(\boldsymbol{x}) f_{\boldsymbol{\theta_2}}(\boldsymbol{x})} d\mu(\boldsymbol{x})$$

δεν αποτελούν φ-μέτρα απόκλισης. Παρ 'όλα αυτά, τα μέτρα αυτά μπορούν να γραφούν με την ακόλουθη μορφή:

$$D_\varphi^h(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = h(D_\varphi(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}))$$

όπου h είναι μια παραγωγίσιμη αύξουσα πραγματική συνάρτηση που έχει απεικόνιση από το $\left[0, \varphi(0) + \lim\limits_{t\to\infty} \frac{\varphi(t)}{t}\right]$ στο $[0, \infty)$ με $h(0) = 0$ *και* $h'(0) > 0$ *και* $\varphi \in \boldsymbol{\Phi}^*$.

Στον παρακάτω πίνακα παρουσιάζονται οι h και φ συναρτήσεις που δίνουν τις παραπάνω αποκλίσεις:

Απόκλιση	$h(x)$	$\varphi(x)$
Rényi	$\dfrac{1}{r(r-1)}\log(r(r-1)x+1)$ $r \neq 0,1$	$\dfrac{x^r - r(x-1) - 1}{r(r-1)}$ $r \neq 0,1$
Sharma-Mittal	$\dfrac{1}{s-1}\log(r(r-1)x+1)$ $r \neq 0,1$	$\dfrac{x^r - r(x-1) - 1}{r(r-1)}$ $r \neq 0,1$
Bhattacharyya	$-\log(-x+1)$	$-x^{\frac{1}{2}} + 1/2(x+1)$

2.3 Βασικές Ιδιότητες των φ-μέτρων Απόκλισης

Σε αυτή την παράγραφο, παρουσιάζονται μερικές από τις πιο σημαντικές ιδιότητες των $\varphi -$ μέτρων απόκλισης, από στατιστικής άποψης. Είναι λογικό να συμβαίνει μια αύξηση της απόκλισης μεταξύ δύο κατανομών που διαφέρουν ολοένα και περισσότερο. Η επόμενη πρόταση είναι απόρροια της ιδέας αυτής.

Πρόταση 2.2

Έστω $P_{\boldsymbol{\theta_1}}$ και $P_{\boldsymbol{\theta_2}}$ είναι δύο κατανομές πιθανότητας και έστω $\varphi \in \boldsymbol{\Phi}^*$ είναι παραγωγίσιμη στο $t = 1$. Τότε θα ισχύει:

$$0 \le D_\varphi(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \le \varphi(0) + \lim_{r\to\infty} \frac{\varphi(r)}{r}$$

όπου

$$D_\varphi(\boldsymbol{\theta_1}\, \boldsymbol{\theta_2}) = 0 \; \alpha\nu \; P_{\boldsymbol{\theta_1}} = P_{\boldsymbol{\theta_2}} \tag{12}$$

και

$$D_\varphi(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \varphi(0) + \lim_{r\to\infty} \frac{\varphi(r)}{r} \; \alpha\nu \; S_1 \cap S_2 = \emptyset \tag{13}$$

Αν η φ είναι αυστηρά κυρτή στο $t = 1$ τότε η (12) ισχύει αν και μόνο αν $P_{\boldsymbol{\theta_1}} = P_{\boldsymbol{\theta_2}}$. Αν επίσης ισχύει η σχέση:

$$\varphi(0) + \lim_{r\to\infty} \frac{\varphi(r)}{r} < \infty$$

τότε η (13) ισχύει αν και μόνο αν $S_1 \cap S_2 = \emptyset$, όπου $S_i,\ i = 1{,}2$ είναι το στήριγμα της κατανομής πιθανότητας $P_{\boldsymbol{\theta_i}},\ i = 1, 2.$

Απόδειξη

Χρησιμοποιώντας το γεγονός ότι συνάρτηση ψ που δίνεται στην σχέση (11) είναι μη αρνητική θα έχουμε ότι ισχύει $D_\psi(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \ge 0$, αλλά γνωρίζουμε ότι ισχύει η σχέση $D_\varphi(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = D_\psi(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$, επομένως θα έχουμε και ότι $D_\varphi(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \ge 0$.

Γνωρίζουμε ότι για κάθε κυρτή συνάρτηση φ ισχύει η ακόλουθη ανισότητα.

$$\varphi(t) \le \varphi(0) + t \lim_{r\to\infty} \frac{\varphi(r)}{r}, \qquad (t \ge 0) \tag{14}$$

Αν η φ είναι αυστηρά κυρτή σε κάποιο $t_0 \in (0, \infty)$ τότε η ανισότητα στο (14) είναι αυστηρά μικρότερη για κάθε $t > 0$. Χρησιμοποιώντας την (14) έχουμε:

$$D_\varphi(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \le \int_X f_{\boldsymbol{\theta_2}}\ (\boldsymbol{x}) \left(\varphi(0) + \frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})} \lim_{r\to\infty} \frac{\varphi(r)}{r} \right) d\mu(\boldsymbol{x})$$

$$= \varphi(0) + \lim_{r\to\infty} \frac{\varphi(r)}{r}$$

Όπως φαίνεται αν $P_{\boldsymbol{\theta_1}} = P_{\boldsymbol{\theta_2}}$ τότε θα ισχύει ότι $D_\varphi(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = 0$. Ενώ αν $S_1 \cap S_2 = \emptyset$ τότε θα ισχύει:

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

$$= \int_{S_1{}^C \cap S_2} f_{\boldsymbol{\theta_2}}(\boldsymbol{x}) \varphi\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})}\right) d\mu(\boldsymbol{x}) + \int_{S_1 \cap S_2{}^C} f_{\boldsymbol{\theta_2}}(\boldsymbol{x}) \varphi\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})}\right) d\mu(\boldsymbol{x})$$

$$= \varphi(0) + \lim_{r \to \infty} \frac{\varphi(r)}{r}$$

Θα δείξουμε ότι αν η φ είναι αυστηρά κυρτή στο $t = 1$ τότε επειδή $D_{\varphi}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = 0$ συνεπάγεται ότι $P_{\boldsymbol{\theta_1}} = P_{\boldsymbol{\theta_2}}$. Για την ακρίβεια αν η φ είναι αυστηρά κυρτή στο $t = 1$ τότε ισχύει:

$$\psi\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})}\right) > 0$$

Η σχέση αυτή ισχύει για $f_{\boldsymbol{\theta_1}}(\boldsymbol{x})/f_{\boldsymbol{\theta_2}}(\boldsymbol{x}) > 1$ και για $f_{\boldsymbol{\theta_1}}(\boldsymbol{x})/f_{\boldsymbol{\theta_2}}(\boldsymbol{x}) < 1$ και η ψ ορίζεται στην (11). Αν $D_{\psi}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = 0$ τότε $f_{\boldsymbol{\theta_1}}(\boldsymbol{x})/f_{\boldsymbol{\theta_2}}(\boldsymbol{x}) \leq 1$ ή $f_{\boldsymbol{\theta_1}}(\boldsymbol{x})/f_{\boldsymbol{\theta_2}}(\boldsymbol{x}) \geq 1$.

Αρχικά υποθέτουμε ότι $f_{\boldsymbol{\theta_1}}(\boldsymbol{x})/f_{\boldsymbol{\theta_2}}(\boldsymbol{x}) \leq 1$.
Γνωρίζουμε ότι ισχύει:

$$D_{\varphi}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = D_{\psi}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = 0$$

και ισχύει:

$$0 = D_{\psi}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \int_X f_{\boldsymbol{\theta_2}}(\boldsymbol{x}) \psi\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})}\right) d\mu(\boldsymbol{x})$$

$$= \int_X f_{\boldsymbol{\theta_2}}(\boldsymbol{x}) \left(\varphi\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})}\right) - \varphi'(1)\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})} - 1\right)\right) d\mu(\boldsymbol{x})$$

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

$$= 0 - \varphi'(1) \int_X f_{\boldsymbol{\theta_2}}(\boldsymbol{x}) \left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})} - 1\right) d\mu(\boldsymbol{x}) = -\varphi'(1) \int_X \left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})} - 1\right) dP_{\boldsymbol{\theta_2}}$$

Αφού φ είναι αυστηρά κυρτή στο $t = 1$ πρέπει να ισχύει $P_{\boldsymbol{\theta_1}} = P_{\boldsymbol{\theta_2}}$. Παρόμοια αποδεικνύεται για $f_{\theta_1}(x)/f_{\theta_2}(x) \geq 1$.

Η αυστηρή κυρτότητα της φ στο $t = 1$ συνεπάγεται την αυστηρή ανισότητα στην (14) δηλαδή:

$$\varphi(t) < \varphi(0) + t \lim_{r\to\infty} \frac{\varphi(r)}{r}, \qquad \forall\, t > 0$$

Τότε η συνάρτηση

$$l(t) = \varphi(0) - \varphi(t) + t \lim_{r\to\infty} \frac{\varphi(r)}{r}$$

είναι θετική $\forall\, t > 0$.
Αν πάρουμε ένα $\boldsymbol{x} \in S_1$, δηλαδή $\boldsymbol{x}$ τέτοιο ώστε $f_{\boldsymbol{\theta_1}}(x) > 0$, τότε $t = \frac{f_{\boldsymbol{\theta_1}}(x)}{f_{\boldsymbol{\theta_2}}(x)} > 0$ και $l\left(\frac{f_{\boldsymbol{\theta_1}}(x)}{f_{\boldsymbol{\theta_2}}(x)}\right) > 0$. Με αποτέλεσμα να ισχύει:

$$D_l(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \int_X f_{\boldsymbol{\theta_2}}(\boldsymbol{x}) l\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})}\right) d\mu(\boldsymbol{x})$$

$$= \int_X f_{\boldsymbol{\theta_2}}(\boldsymbol{x}) \left(\varphi(0) - \varphi\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})}\right) + \frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})} \lim_{r\to\infty} \frac{\varphi(r)}{r}\right) d\mu(\boldsymbol{x})$$

$$= -D_\varphi(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) + \varphi(0) + \lim_{r\to\infty} \frac{\varphi(r)}{r},$$

Αλλά από την (13) έχουμε:

$$D_\varphi(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \varphi(0) + \lim_{r\to\infty} \frac{\varphi(r)}{r}$$

Άρα θα ισχύει:

$$D_l(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \int_X f_{\boldsymbol{\theta_2}}(\boldsymbol{x}) l\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})}\right) d\mu(\boldsymbol{x}) = 0$$

όπου

$$l\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})}\right) > 0$$

Τότε $f_{\boldsymbol{\theta_2}}(\boldsymbol{x}) = 0$ επειδή $D_l(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = 0$ και $l\left(\frac{f_{\boldsymbol{\theta_1}}(x)}{f_{\boldsymbol{\theta_2}}(x)}\right) > 0$ δηλαδή $\boldsymbol{x} \notin S_2$.

□

Έστω $X_1, \dots, X_n$ είναι ένα δείγμα από το $P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta$. Έστω μ είναι το μέτρο Lebesgue ή ένα απαριθμητικό μέτρο και έστω $f_{\boldsymbol{\theta}}(x) = \frac{dP_{\boldsymbol{\theta}}}{d\mu}(\boldsymbol{x})$, όπου $\boldsymbol{x} = (x_1, \dots, x_n)$. Υποθέτουμε ότι T είναι ένας μετρήσιμος μετασχηματισμός από τον (X^n, β_{χ^n}) στον μετρήσιμο χώρο (Y, β_Y).

Ορίζουμε

$$Q_{\boldsymbol{\theta_i}}(A) = P_{\boldsymbol{\theta_i}}\big(T^{-1}(A)\big), \qquad i = 1,2 \tag{15}$$

όπου $A \in \beta_Y$ και

$$g_{\boldsymbol{\theta}_i}(t) = \frac{dQ_{\boldsymbol{\theta_i}}}{d\mu}(\boldsymbol{t}),\ f_{\boldsymbol{\theta}_i}\left(\frac{\boldsymbol{x}}{\boldsymbol{t}}\right) = \frac{dP_{\boldsymbol{\theta}_i}}{dQ_{\boldsymbol{\theta}_i}}, \quad i = 1,2 \tag{16}$$

όπου με $\boldsymbol{t}$ συμβολίζουμε τις τιμές του T. Με βάση τα παραπάνω έχουμε την ακόλουθη πρόταση ιδιότητας.

Πρόταση 2.3

Έστω $\varphi \in \boldsymbol{\Phi}^*$ και $Q_{\boldsymbol{\theta}_i},\ P_{\boldsymbol{\theta}_i},\ i = 1,2$, είναι δύο μέτρα πιθανότητας που ορίζονται από τις (15) και (16). Τότε έχουμε:

$$D_\varphi\big(Q_{\boldsymbol{\theta_1}}, Q_{\boldsymbol{\theta_2}}\big) \le D_\varphi\big(P_{\boldsymbol{\theta_1}}, P_{\boldsymbol{\theta_2}}\big)$$

Η ισότητα ισχύει αν ο μετασχηματισμός T είναι επαρκής για τις κατανομές πιθανότητας $P_{\boldsymbol{\theta_1}}$ και $P_{\boldsymbol{\theta_2}}$.

Απόδειξη

Ισχύει:

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

$$= \int_X \int_Y f_{\boldsymbol{\theta_2}}\left(\frac{\boldsymbol{x}}{\boldsymbol{t}}\right) g_{\boldsymbol{\theta_2}}(\boldsymbol{t})\varphi\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})}\right)d\mu(\boldsymbol{t})d\mu(\boldsymbol{x})$$

$$= \int_Y g_{\boldsymbol{\theta_2}}(\boldsymbol{t})\left(\int_X f_{\boldsymbol{\theta_2}}\left(\frac{\boldsymbol{x}}{\boldsymbol{t}}\right)\varphi\left(\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})}\right)d\mu(\boldsymbol{x})\right)d\mu(\boldsymbol{t})$$

Εφαρμόζοντας την ανισότητα του Jensen θα έχουμε:

$$D_\varphi\big(P_{\boldsymbol{\theta_1}}, P_{\boldsymbol{\theta_2}}\big) \ge \int_Y g_{\boldsymbol{\theta_2}}(\boldsymbol{t})\left(\varphi\left(\int_X f_{\boldsymbol{\theta_2}}\left(\frac{\boldsymbol{x}}{\boldsymbol{t}}\right)\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})}d\mu(\boldsymbol{x})\right)\right)d\mu(\boldsymbol{t})$$

Αλλά

$$\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})} = \frac{\frac{dP_{\boldsymbol{\theta_1}}}{d\mu}}{\frac{dP_{\boldsymbol{\theta_2}}}{d\mu}} = \frac{\frac{dQ_{\boldsymbol{\theta_1}}}{d\mu}}{\frac{dQ_{\boldsymbol{\theta_2}}}{d\mu}}\frac{\frac{dP_{\boldsymbol{\theta_1}}}{dQ_{\boldsymbol{\theta_1}}}}{\frac{dP_{\boldsymbol{\theta_2}}}{dQ_{\theta_2}}} = \frac{g_{\boldsymbol{\theta_1}}(\boldsymbol{t})}{g_{\boldsymbol{\theta_2}}(\boldsymbol{t})}\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x/t})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x/t})} \tag{17}$$

Τότε

$$D_{\varphi}(P_{\boldsymbol{\theta_1}}, P_{\boldsymbol{\theta_2}}) \geq \int_{Y} g_{\boldsymbol{\theta_2}}(\boldsymbol{t})\varphi\left(\frac{g_{\boldsymbol{\theta_1}}(\boldsymbol{t})}{g_{\boldsymbol{\theta_2}}(\boldsymbol{t})}\right) d\mu(\boldsymbol{t}) = D_{\varphi}(Q_{\boldsymbol{\theta_1}}, Q_{\boldsymbol{\theta_2}}).$$

Αν η φ είναι κυρτή, τότε η ισότητα ισχύει αν και μόνο αν

$$\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})} = \int_{X} f_{\boldsymbol{\theta_2}}(\boldsymbol{x/t})\frac{f_{\boldsymbol{\theta_1}}(\boldsymbol{x})}{f_{\boldsymbol{\theta_2}}(\boldsymbol{x})} d\mu(\boldsymbol{x}), \qquad \gamma\iota\alpha\ \acute{o}\lambda\alpha\ \tau\alpha\ \boldsymbol{x}$$

Ο δεύτερος όρος στην προηγούμενη ανισότητα είναι ίσος με $g_{\boldsymbol{\theta_1}}(\boldsymbol{t})/g_{\boldsymbol{\theta_2}}(\boldsymbol{t})$ σύμφωνα με την (17). Τότε χρησιμοποιώντας το παραγοντικό θεώρημα, η ισότητα ισχύει αν ο μετασχηματισμός T είναι επαρκής για τις κατανομές πιθανότητας $P_{\boldsymbol{\theta_1}}$ και $P_{\boldsymbol{\theta_2}}$. Οπότε και ολοκληρώθηκε η απόδειξη της πρότασης.

Στην επόμενη πρόταση το $\{P_{\theta}\}_{\theta\in\Theta}, \boldsymbol{\Theta} \subset R$ είναι μια οικογένεια από μέτρα πιθανότητας που ορίζονται στην σ-άλγεβρα υποσυνόλων του Borel της ευθείας πραγματικών αριθμών με μονότονο λόγο πιθανοφάνειας στο x. Δηλαδή αν για κάποια θ_1, θ_2 ισχύει $\theta_1 < \theta_2$ τότε οι f_{θ_1}, f_{θ_2} είναι διακριτές και ο λόγος $f_{\theta_1}/f_{\theta_2}$ είναι μια αύξουσα συνάρτηση του x.

Πρόταση 2.4

Υποθέτουμε ότι οι κατανομές πιθανότητας $\{P_{\theta}\}_{\theta\in\boldsymbol{\Theta}}$ ορίζονται στην ευθεία των πραγματικών αριθμών με $\boldsymbol{\theta} \in (a, b) \subset R$ και έστω P_{θ} είναι μια απόλυτα συνεχής κατανομή σε σχέση με ένα σ-πεπερασμένο μέτρο μ. Υποθέτουμε επίσης ότι οι αντίστοιχες συναρτήσεις πυκνότητας ή συναρτήσεις μάζας πιθανότητας έχουν μονότονο λόγο πιθανοφάνειας στο x. Αν $\alpha < \boldsymbol{\theta_1} < \boldsymbol{\theta_2} < \boldsymbol{\theta_3} < b$ και η συνάρτηση w είναι συνεχής τότε ισχύει:

$$D_w(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \leq D_w(\boldsymbol{\theta_1}, \boldsymbol{\theta_3}), w \in \boldsymbol{\Phi}^* \qquad (18)$$

Απόδειξη

Υποθέτουμε ότι μ είναι το μέτρο Lebesque και ορίζουμε την παρακάτω σχέση:

$$\widetilde{D}_{\varphi}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \int_{\mathbb{R}} f_{\boldsymbol{\theta_1}}(x)\varphi\left(\frac{f_{\boldsymbol{\theta_2}}(x)}{f_{\boldsymbol{\theta_1}}(x)}\right) dx$$

και θα αποδείξουμε ότι ισχύει:

$$\widetilde{D}_{\varphi}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \leq \widetilde{D}_{\varphi}(\boldsymbol{\theta_1}, \boldsymbol{\theta_3}), \varphi \in \boldsymbol{\Phi}^* \qquad (19)$$

Αν ισχύει η σχέση (19) τότε και η (18) ισχύει, γιατί αν θεωρήσουμε την συνάρτηση:

$$\varphi(t) = tw\left(\frac{1}{t}\right) \in \boldsymbol{\Phi}^*$$

Θα έχουμε:

$$\widetilde{D}_{\varphi}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \int_{\mathbb{R}} f_{\boldsymbol{\theta_1}}(x) \frac{f_{\boldsymbol{\theta_2}}(x)}{f_{\boldsymbol{\theta_1}}(x)} w\left(\frac{f_{\boldsymbol{\theta_2}}(x)}{f_{\boldsymbol{\theta_1}}(x)}\right) dx = D_w(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}).$$

Αφού από την υπόθεση, η οικογένεια κατανομών $\{P_\theta\}_{\theta \in \boldsymbol{\Theta} \subset \mathbb{R}}$ έχει λόγο πιθανοφάνειας μονότονο και μη φθίνων, τότε οι συναρτήσεις ως προς x:

$$h_2(x) = \frac{f_{\boldsymbol{\theta_2}}(x)}{f_{\boldsymbol{\theta_1}}(x)} \; \kappa\alpha\iota \; h_3(x) = \frac{f_{\boldsymbol{\theta_3}}(x)}{f_{\boldsymbol{\theta_1}}(x)}$$

είναι μη φθίνουσες. Το ίδιο ισχύει και για τον λόγο τους:

$$\frac{h_3(x)}{h_2(x)} = \frac{f_{\boldsymbol{\theta_3}}(x)}{f_{\boldsymbol{\theta_2}}(x)}$$

Από την σχέση αυτή θεωρούμε 3 περιπτώσεις:

a) $h_3(x) < h_2(x)$ για όλα τα x

b) $h_3(x) > h_2(x)$ για όλα τα x

c) Υπάρχει αριθμός α τέτοιος ώστε να ισχύει: $h_3(x) \leq h_2(x)$ για $x < \alpha$ και $h_3(x) \geq h_2(x)$ για $x > \alpha$.

Γνωρίζουμε ότι:

$$E_{\theta_1}[h_3(X)] = \int_{\mathbb{R}} f_{\boldsymbol{\theta_1}}(x) \frac{f_{\boldsymbol{\theta_3}}(x)}{f_{\boldsymbol{\theta_1}}(x)} dx = E_{\theta_1}[h_2(X)] = 1$$

Αν $E_{\theta_1}[h_3(X)] = E_{\theta_1}[h_2(X)] = 1$, τότε οι περιπτώσεις (a) και (b) δεν μπορούν να ισχύουν, οπότε έχουμε την περίπτωση (c). Χρησιμοποιώντας την μονοτονία των $h_2(x)$ και $h_3(x)$ έχουμε:

$$\{x : h_2(X) \leq b\} \subset \{x : h_2(X) \leq b\}, \qquad \alpha\nu \; b < h_2(a)$$

και

$$\{x : h_2(X) \leq b\} \supset \{x : h_3(X) \leq b\}, \qquad \alpha\nu \; b > h_2(a).$$

Αν ορίσουμε τις:

$$F_{h_2(X)}(t) = P_{r_{\theta_1}}(h_2(X) \leq t) = P_{r_{\theta_1}}(x \in \mathbb{R} : h_2(X) \leq t)$$

$$F_{h_3(X)}(t) = P_{r_{\theta_1}}(h_3(X) \leq t) = P_{r_{\theta_1}}(x \in \mathbb{R} : h_3(X) \leq t)$$

Θα έχουμε τότε για $t < h_2(a)$:

$$F_{h_2(X)}(t) = P_{r_{\theta_1}}(x \in \mathbb{R} : h_2(X) \leq t) \leq P_{r_{\theta_1}}(x \in \mathbb{R} : h_3(X) \leq t) = F_{h_3(X)}(t)$$

και για $t > h_2(a)$ θα έχουμε:

$$F_{h_2(X)}(t) \geq F_{h_3(X)}(t)$$

Για την συνέχεια της απόδειξης θα αποδείξουμε ότι οι παρακάτω σχέσεις:

- $E_{\theta_1}[h_3(X)] = E_{\theta_2}[h_3(X)]$
- $F_{h_2(X)}(t) \leq F_{h_3(X)}(t)$

συνεπάγονται την σχέση:

$$E_{\theta_1}[|h_2(X) - k|] \leq E_{\theta_1}[|h_3(X) - k|], \qquad \gamma\iota\alpha\ \acute{o}\lambda\alpha\ \tau\alpha\ k. \tag{20}$$

Είναι γνωστό ότι η αναμενόμενη τιμή μιας μη αρνητικής μεταβλητής X μπορεί να γραφεί ως εξής:

$$E[X] = \int_0^\infty (1 - F_X(x))dx$$

Άρα για την περίπτωση που αναλύουμε θα έχουμε:

$$E_{\theta_1}[h_3(X)] = \int_0^\infty (1 - F_{h_3(X)}(x))dx = \int_0^\infty (1 - F_{h_2(X)}(x))dx = E_{\theta_1}[h_2(X)]$$

Ορίζουμε τις σχέσεις:

$$I_1 \equiv \int_0^{h_2(a)} \Big(\big(1 - F_{h_3(X)}(x)\big) - \big(1 - F_{h_2(X)}(x)\big)\Big)\, dx$$

$$I_2 \equiv \int_{h_2(a)}^\infty \Big(\big(1 - F_{h_3(X)}(x)\big) - \big(1 - F_{h_2(X)}(x)\big)\Big)\, dx$$

Οι σχέσεις αυτές μπορούν να γραφτούν και ως εξής:

$$I_1 \equiv \int_0^{h_2(a)} \Big(F_{h_2(X)}(x) - F_{h_3(X)}(x)\Big)\, dx\,, \quad I_2 \equiv \int_{h_2(a)}^\infty \Big(F_{h_2(X)}(x) - F_{h_3(X)}(x)\Big)\, dx$$

Θα έχουμε τότε:

$$E_{\theta_1}[h_3(X)] - E_{\theta_1}[h_2(X)] =$$

$$= \int_0^{h_2(a)} \left(F_{h_2(X)}(x) - F_{h_3(X)}(x)\right) dx + \int_{h_2(a)}^{\infty} \left(F_{h_2(X)}(x) - F_{h_3(X)}(x)\right) dx = 0$$

Άρα ισχύει:

$$\int_0^{h_2(a)} \left(F_{h_2(X)}(x) - F_{h_3(X)}(x)\right) dx = \\ = \int_{h_2(a)}^{\infty} \left(F_{h_3(X)}(x) - F_{h_2(X)}(x)\right) dx \tag{21}$$

Τώρα αποδεικνύουμε την σχέση (20). Παρατηρούμε ότι ισχύει:

$$E_{\theta_1}[|h_i(X) - k|] = \int_0^{k} F_{h_i(X)}(x)dx + \int_{k}^{\infty} \left(1 - F_{h_i(X)}(x)\right) dx$$

Υποθέτοντας ότι $k \geq h_2(a)$, μια ανάλογη απόδειξη γίνεται και στην περίπτωση $k < h_2(a)$. Θα έχουμε:

$$E_{\theta_1}[|h_i(X) - k|] =$$

$$= \int_0^{h_2(a)} F_{h_i(X)}(x)dx + \int_{h_2(a)}^{k} F_{h_i(X)}(x)dx + \int_{k}^{\infty} \left(1 - F_{h_i(X)}(x)\right) dx \ \text{, για } i = 2,3$$

Ορίζουμε s τέτοιο ώστε:

$$s = E_{\theta_1}[|h_3(X) - k|] - E_{\theta_1}[|h_2(X) - k|]$$

Και ώστε να ισχύει:

$$s = \int_0^{h_2(a)} \left(F_{h_3(X)}(x) - F_{h_2(X)}(x)\right) dx -$$

$$- \int_{h_2(a)}^{k} \left(F_{h_2(X)}(x) - F_{h_3(X)}(x)\right) dx + \int_{k}^{\infty} \left(F_{h_2(X)}(x) - F_{h_3(X)}(x)\right) dx$$

Από (21) θα έχουμε:

$$\int_0^{h_2(a)} \left(F_{h_3(X)}(x) - F_{h_2(X)}(x)\right) dx = \int_{h_2(a)}^{\infty} \left(F_{h_2(X)}(x) - F_{h_3(X)}(x)\right) dx \\ \geq \int_{h_2(a)}^{k} \left(F_{h_2(X)}(x) - F_{h_3(X)}(x)\right) dx$$

Οπότε και θα έχουμε:

$$s \geq \int_{k}^{\infty} \left(F_{h_2(X)}(x) - F_{h_3(X)}(x)\right) dx \geq 0$$

Άρα

$$E_{\theta_1}[|h_3(X) - k|] \geq E_{\theta_1}[|h_2(X) - k|] \tag{22}$$

Τελικά αποδεικνύουμε την (19) ή ισοδύναμα την σχέση:

$$E_{\theta_1}[w(h_3(X))] \geq E_{\theta_1}[w(h_2(X))]$$

Αφού η w είναι συνεχής και κυρτή θα έχουμε:

$$w(z) - w(0) = \int_0^z b(k)dk$$

Όπου b είναι μη φθίνουσα συνάρτηση και φραγμένη στο διάστημα $[0, z]$. Ολοκληρώνοντας κατά μέλη συνεπάγεται:

$$w(z) - w(0) = zb(z) - \int_0^z kdb(k) = \int_0^z (z - k)db(k) + zb(0)$$

Θεωρούμε την συνάρτηση:

$$b^*(k) = \begin{cases} b(k), & \alpha\nu\ k \in [0, z] \\ c, & \alpha\nu\ k > z \end{cases}$$

Τότε θα έχουμε:

$$w(z) - w(0) = \int_0^z (z - k)db^*(k) + zb^*(0) + \int_z^\infty (z - k)db^*(k) =$$
$$= \int_0^\infty (z - k)db^*(k) + zb^*(0)$$

Αφού $\int_z^\infty (z - k)db^*(k) = 0$. Άρα θα ισχύει:

$$E[w(Z)] = E\left[\int_0^\infty (Z - k)db^*(k) + Zb^*(0) + w(0)\right]$$
$$= \int_0^\infty \int_0^\infty (z - k)db^*(k)dF_Z(z) + E[Z]b^*(0) + w(0)$$
$$= \int_0^\infty E[Z - k]db^*(k) + E[Z]b^*(0) + w(0)$$

Όμως,

$$\int_0^\infty E[|Z - k|]db^*(k) = \int_0^\infty \left(\int_0^\infty (z - k)dF_Z(z)\right)db^*(k)$$

$$= \int_0^\infty \left(\int_0^z (z - k)db^*(k) + \int_z^\infty -(z - k)db^*(k)\right)dF_Z(z)$$

Και

$$\int_z^{\infty}(z-k)db^*(k)=0$$

Τότε

$$\int_0^{\infty}E[|Z-k|]db^*(k)=\int_0^{\infty}E[Z-k]db^*(k)$$

Και άρα:

$$E[w(Z)]=\frac{1}{2}\int_0^{\infty}E[(Z-k)+|Z-k|]db^*(k)+E[Z]b^*(0)+w(0)$$

Οπότε αν θεωρήσουμε ότι $Z\equiv h_2(X)$ και επειδή $E_{\theta_1}[h_i(X)]=1,\ \gamma\iota\alpha\ \ i=2,3$ θα έχουμε:

$$E_{\theta_1}[w(h_2(X))]=\frac{1}{2}\int_0^{\infty}(1-k+E[|h_2(X)-k|])db^*(k)+b^*(0)+w(0)$$

Όμοια για $Z\equiv h_3(X)$

$$E_{\theta_1}[w(h_3(X))]=\frac{1}{2}\int_0^{\infty}(1-k+E[|h_3(X)-k|])db^*(k)+b^*(0)+w(0)$$

Όμως από (22) τελικά παρατηρούμε ότι ισχύει:

$$E_{\theta_1}[w(h_3(X))]\geq E_{\theta_1}[w(h_2(X))]$$

Και άρα ολοκληρώθηκε η απόδειξη της σχέσης (19).

Οι πηγές που χρησιμοποιήθηκαν σε αυτό το κεφάλαιο αφορούν κυρίως το [8] στην βιβλιογραφία.

ΚΕΦΑΛΑΙΟ 3

ΕΛΕΓΧΟΣ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ: ΑΠΛΗ ΜΗΔΕΝΙΚΗ ΥΠΟΘΕΣΗ

3.1 Έλεγχος Καλής Προσαρμογής

Σύμφωνα με τον Pardo (βλ. [8]) το πρόβλημα της καλής προσαρμογής μιας κατανομής στην πραγματική ευθεία, εξετάζεται υπό την μηδενική υπόθεση $H_0 : F = F_0$ και αντιμετωπίζεται συχνά διαμερίζοντας το εύρος των δεδομένων σε ξένα μεταξύ τους διαστήματα και ελέγχοντας την υπόθεση $H_0 : \boldsymbol{p} = \boldsymbol{p}^0$ σχετικά με το διάνυσμα παραμέτρων μιας πολυωνυμικής κατανομής.

Έστω $P = \{E_i\}_{i=1,\dots,M}$ είναι μια διαμέριση της ευθείας των πραγματικών αριθμών σε M διαστήματα. Έστω $\boldsymbol{p} = (p_1, \dots, p_M)^T$ και $\boldsymbol{p^0} = (p_1^0, \dots, p_M^0)^T$ είναι οι πραγματικές και υποθετικές πιθανότητες των διαστημάτων $E_i,\ i = 1, \dots, M$, αντίστοιχα με τέτοιο τρόπο ώστε να ισχύει :

$$p_i = P_F(E_i)\ ,\ \ i =, \dots, M, \text{ και } p_i^0 = P_{F_0}(E_i) = \int_{E_i} dF_0\ ,\ i = 1, \dots, M.$$

Έστω $Y_1, \dots, Y_n$ είναι ένα τυχαίο δείγμα από το F και $N_i = \sum_{j=1}^{n} I_{E_i}(Y_j)$, με $I_{E_i}(Y_j) = 1$ αν $Y_j \in E_i$ και 0 διαφορετικά, οι απόλυτες συχνότητες των διαστημάτων. Επίσης, έστω $\hat{\boldsymbol{p}} = (\hat{p}_1, \dots, \hat{p}_M)^T$ με $\hat{p}_i = N_i/n,\ i = 1, \dots, M$ είναι οι σχετικές συχνότητες των διαστημάτων.
Αν θέλουμε να εξετάσουμε την απλή μηδενική υπόθεση:

$$H_0 : \boldsymbol{p} = \boldsymbol{p^0} \tag{23}$$

μερικά από τα πιο χρησιμοποιούμενα στατιστικά τεστ παρουσιάζονται παρακάτω. Όπως είναι η ελεγχοσυνάρτηση του Pearson, X^2 (ή X-τετράγωνο έλεγχος).

$$X^2 \equiv \sum_{i=1}^{M} \frac{(N_i - np_i^0)^2}{np_i^0} \tag{24}$$

Καθώς και η ελεγχοσυνάρτηση του λόγου πιθανοφάνειας G^2:

$$G^2 \equiv 2\sum_{i=1}^{M} N_i \log\frac{N_i}{np_i^0} \tag{25}$$

Αυτές οι δύο ελεγχοσυναρτήσεις είναι ειδικές περιπτώσεις της οικογένειας στατιστικών κριτηρίων απόκλισης-δύναμης των Cressie και Read (1984) με μαθηματικό τύπο:

$$\begin{aligned} T_n^{\lambda}(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) &= \frac{2n}{\lambda(\lambda+1)}\sum_{i=1}^{M} \hat{p}_i\left(\left(\frac{\hat{p}_i}{p_i^0}\right)^{\lambda} - 1\right) \\ &= \frac{2n}{\lambda(\lambda+1)}\sum_{i=1}^{M} N_i\left(\left(\frac{N_i}{np_i^0}\right)^{\lambda} - 1\right), \end{aligned} \tag{26}$$

όπου $-\infty < \lambda < \infty$. Η ελεγχοσυνάρτηση $T_n^0(\hat{\boldsymbol{p}}, \boldsymbol{p^0})$ και $T_n^{-1}(\hat{\boldsymbol{p}}, \boldsymbol{p^0})$ ορίζονται ως τα όρια του $T_n^{\lambda}(\hat{\boldsymbol{p}}, \boldsymbol{p^0})$ καθώς το $\lambda \to 0$ και καθώς το $\lambda \to -1$, αντίστοιχα. Μερικές συγκεκριμένες τιμές που αντιστοιχούν σε γνωστά στατιστικά κριτήρια είναι για παράδειγμα, το στατιστικό κριτήριο X-τετράγωνο X^2 ($για\ \lambda = 1$), το στατιστικό κριτήριο λόγου πιθανοφάνειας G^2 ($\lambda = 0$), η ελεγχοσυνάρτηση Freeman-Tukey ($\lambda = -1/2$), το τροποποιημένο στατιστικό κριτήριο λόγου πιθανοφάνειας ή κριτήριο ελάχιστης διάκρισης πληροφορίας (Gokhale και Kullback, 1978) ($\lambda = -1$), η τροποποιημένη ελεγχοσυνάρτηση Neyman ή αλλιώς τροποποιημένη ελεγχοσυνάρτηση χ-τετράγωνο ($\lambda = -2$) και η ελεγχοσυνάρτηση των Cressie-Read ($\lambda = 2/3$). Οι παραστάσεις των ελεγχοσυναρτήσεων X^2 και G^2 δίνονται από τις (23) και (24) αντίστοιχα. Οι εκφράσεις των υπολοίπων στατιστικών κριτηρίων δίνονται παρακάτω ως εξής :

i) $\lambda = -2$ (τροποποιημένη ελεγχοσυνάρτηση X-τετράγωνο)

$$T_n^{-2}(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) = n\sum_{i=1}^{M} \frac{(p_i^0 - \hat{p}_i)^2}{\hat{p}_i} = \sum_{i=1}^{M} \frac{(np_i^0 - N_i)^2}{N_i}$$

ii) $\lambda = -1$ ($\lambda \to -1$) (Τροποποιημένο στατιστικό κριτήριο λόγου πιθανοφάνειας)

$$T_n^{-1}(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) = 2n\sum_{i=1}^{M} p_i^0 \log\left(\frac{p_i^0}{\hat{p}_i}\right) = 2\sum_{i=1}^{M} N_i \log\left(\frac{np_i^0}{N_i}\right)$$

iii) $\lambda = -1/2$ (ελεγχοσυνάρτηση Freeman-Tukey)

$$T_n^{-\frac{1}{2}}(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) = 8n\left(1 - \sum_{i=1}^{M} \sqrt{p_i^0 \hat{p}_i}\right) = 8n\left(1 - \sum_{i=1}^{M} \sqrt{\frac{N_i p_i^0}{n}}\right)$$

iv) $\lambda = 2/3$ (ελεγχοσυνάρτηση Cressie – Read)

$$T_n^{2/3}(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) = \frac{9}{5} n \left(\sum_{i=1}^{M} \hat{p}_i \left(\frac{\hat{p}_i}{p_i^0} \right)^{2/3} - 1 \right)$$

Αν και τα στατιστικά κριτήρια απόκλισης – δύναμης αποτελούν μια σημαντική και εύχρηστη οικογένεια κριτηρίων, είναι δυνατόν να θεωρήσουμε μια γενικότερη οικογένεια ελεγχοσυναρτήσεων για να εξετάσουμε την (23) η οποία να μπορεί να περιλαμβάνει την (26) ως ιδιαίτερη περίπτωση. Τα στατιστικά κριτήρια φ-απόκλισης ορίζονται ως εξής:

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

Υποθέτουμε για την $\varphi(x)$ στην παραπάνω σχέση ότι είναι δύο φορές παραγωγίσιμη για $x > 0$ και ισχύει $\varphi''(1) \neq 0$.

3.2 φ-αποκλίσεις και Έλεγχος Καλής Προσαρμογής με Σταθερό Αριθμό Κλάσεων

Ο Pearson (1900) απέδειξε για την ασυμπτωτική συμπεριφορά του στατιστικού κριτηρίου X^2 ότι ισχύει:

$$X^2 \xrightarrow[n\to\infty]{L} X_{M-1}^2,$$

με X^2 να δίνεται από την σχέση (24). Αργότερα αυτό το αποτέλεσμα επεκτάθηκε για το στατιστικό κριτήριο του λόγου πιθανοφάνειας και για το τροποποιημένο X-τετράγωνο στατιστικό κριτήριο των Neyman και Pearson (1928) και Neyman (1949). Αργότερα οι Cressie και Read (1984) απέδειξαν ότι:

$$T_n^{\lambda}(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) \xrightarrow[n\to\infty]{L} X_{M-1}^2, \text{υπό την μηδενική υπόθεση } H_0 : \boldsymbol{p} = \boldsymbol{p^0} \text{ για κάθε } \lambda \in \mathbb{R}.$$

Ο Zografos (1990) απέδειξε για την ασυμπτωτική συμπεριφορά των στατιστικών κριτηρίων φ-αποκλίσεων ότι:

$$T_n^{\varphi}(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) \xrightarrow[n\to\infty]{L} X_{M-1}^2, \text{υπό την μηδενική υπόθεση } H_0 : \boldsymbol{p} = \boldsymbol{p^0} \text{ για κάθε } \varphi \in \Phi^*.$$

Παρακάτω θα αναλύσουμε την ασυμπτωτική κατανομή των στατιστικών κριτηρίων φ-αποκλίσεων $T_n^{\varphi}(\hat{\boldsymbol{p}}, \boldsymbol{p^0})$ υπό την μηδενική υπόθεση $H_0 : \boldsymbol{p} = \boldsymbol{p^0}$ με εναλλακτική την υπόθεση $H_1 : \boldsymbol{p} = \boldsymbol{p}^* \neq \boldsymbol{p^0}$. Παρουσιάζονται δύο προτάσεις που θα βοηθήσουν στην απόδειξη για την ανάλυση της ασυμπτωτικής κατανομής.

Πρόταση 3.1

Έστω πείραμα του οποίου τα αποτελέσματα ανήκουν σε μια από τις M $(M \geq 2)$ αμοιβαίως αποκλειόμενες και πλήρεις κατηγορίες, $E_1, \dots, E_M$ και έστω p_i $(0 < p_i < 1)$ είναι η πιθανότητα ότι το αποτέλεσμα ανήκει στην ιοστή κατηγορία $(i = 1, \dots, M)$ και ισχύει $\sum_{i=1}^{M} p_i = 1$. Έστω ότι το πείραμα πραγματοποιείται n φορές, $Y_1, \dots, Y_n$, και τα n αποτελέσματα είναι ανεξάρτητα. Επίσης, έστω N_j ορίζεται ως ο αριθμός των αποτελεσμάτων που ανήκουν στην κατηγορία E_j $(j = 1, \dots, M)$ με $\sum_{j=1}^{M} N_j = n$. Η τυχαία μεταβλητή $(N_1, \dots, N_M)$ ακολουθεί την πολυωνυμική κατανομή με παραμέτρους n και $\boldsymbol{p} = (p_1, \dots, p_M)^T$ και ισχύει επίσης:

$$\sqrt{n}(\widehat{\boldsymbol{p}} - \boldsymbol{p}) \xrightarrow[n\to\infty]{L} N(\boldsymbol{0}, \boldsymbol{\Sigma_p})$$

Αν ορίσουμε

$$T_j^i = \begin{cases} 1, & \alpha\nu\, Y_i \in E_j \\ 0, & \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases} \quad i = 1, \dots, n, \;\; j = 1, \dots, M$$

τότε τα διανύσματα $\left(T_1^i, \dots, T_M^i\right),\ i = 1, \dots, n$ είναι συναρτήσεις των Y_i και έχουν τιμές στο σύνολο

$$\{(1,0, \dots ,0), \dots, (0, \dots ,0,1)\}.$$

Επίσης, αν ορίσουμε ως $M(n; p_1, \dots, p_M)$ την πολυωνυμική κατανομή με παραμέτρους n και $p_1, \dots, p_M$, τότε θεωρώντας ότι ο συμβολισμός της ταύτισης $\equiv$ χρησιμοποιείται για να δηλώσει ότι τα διανύσματα κατανέμονται με αυτό τον τρόπο, θα ισχύουν τα παρακάτω:

$$Y_1 \to (T_1^1, \dots, T_M^1) \equiv M(1; p_1, \dots, p_M)$$

$$Y_2 \to (T_1^2, \dots, T_M^2) \equiv M(1; p_1, \dots, p_M)$$

$$\dots\dots \qquad\qquad \dots$$

$$Y_n \to (T_1^n, \dots, T_M^n) \equiv M(1; p_1, \dots, p_M)$$

$$\underbrace{\left(\sum_{i=1}^{n} T_1^i, \dots, \sum_{i=1}^{n} T_M^i\right)}_{(N_1, \dots, N_M)} \equiv M(n; p_1, \dots, p_M) \qquad (27)$$

Αν θεωρήσουμε ακολουθία τυχαίων μεταβλητών M −διαστάσεων $\{\boldsymbol{U}_n\}_{n\in\mathbb{N}}$ με

$$\boldsymbol{U}_n = (U_{1n}, \dots, U_{Mn})^T$$

Και

$$E[\boldsymbol{U}_n] = \mu \quad \kappa\alpha\iota \quad Cov[\boldsymbol{U}_n] = \boldsymbol{\Sigma}$$

Από το Κεντρικό Οριακό Θεώρημα για ένα τυχαίο διάνυσμα:

$$\bar{U}_n = \left(\frac{1}{n}\sum_{j=1}^{n} U_{1j}\,, \dots, \frac{1}{n}\sum_{j=1}^{n} U_{Mj} \right)$$

θα ισχύει:

$$\sqrt{n}(\bar{U}_n - \mu) \xrightarrow[n\to\infty]{L} N(\mathbf{0}, \boldsymbol{\Sigma})$$

Για το συγκεκριμένο πείραμα ορίζουμε:

$$\boldsymbol{A} = \left(\frac{N_1 - np_1}{\sqrt{n}}, \dots, \frac{N_M - np_M}{\sqrt{n}} \right)$$

Άρα από (27) θα έχουμε:

$$\boldsymbol{A} = \left(\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} T_1^i - np_1 \right), \dots, \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} T_M^i - np_M \right) \right) =$$

$$= \left(\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} T_1^i - p_1 \right), \dots, \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} T_M^i - p_M \right) \right)$$

Άρα από το Κεντρικό Οριακό Θεώρημα έχουμε:

$$\left(\frac{N_1 - np_1}{\sqrt{n}}, \dots, \frac{N_M - np_M}{\sqrt{n}} \right) \xrightarrow[n\to\infty]{L} N(\mathbf{0}, \boldsymbol{\Sigma_p})$$

όπου

$$\boldsymbol{\Sigma_p} = diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^{\boldsymbol{T}}$$

Άρα τελικά:

$$\sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p}) \xrightarrow[n\to\infty]{L} N(\mathbf{0}, \boldsymbol{\Sigma_p})$$

Πρόταση 3.2

Έστω $\boldsymbol{X}$ μια τυχαία, κανονική πολυμεταβλητή k-μεταβλητών με διάνυσμα μέσης τιμής $\mathbf{0}$ και πίνακα διασποράς-συνδιασποράς $\boldsymbol{\Sigma}$. Τότε η κατανομή της τυχαίας μεταβλητής $\boldsymbol{X^T X}$ είναι η X-τετράγωνο με r βαθμούς ελευθερίας αν και μόνο αν ο πίνακας $\boldsymbol{\Sigma}$ είναι μια προβολή τάξης r. Αφού ο $\boldsymbol{\Sigma}$ πίνακας είναι συμμετρικός και τετραγωνικός, τότε θα αποτελεί προβολή αν ισχύει $\boldsymbol{\Sigma^2} = \boldsymbol{\Sigma}$ και άρα $rank(\boldsymbol{\Sigma}) = trace(\boldsymbol{\Sigma})$.

Με την χρήση των δύο αυτών προτάσεων αποδεικνύουμε το παρακάτω θεώρημα.

Θεώρημα 3.1

Υπό την μηδενική υπόθεση $H_0 : \boldsymbol{p} = \boldsymbol{p^0} = (p_1^0, \dots, p_M^0)^{\boldsymbol{T}}$, η ασυμπτωτική κατανομή της ελεγχοσυνάρτησης της φ-απόκλισης, $T_n^{\varphi}(\widehat{\boldsymbol{p}}, \boldsymbol{p^0})$ είναι η X-τετράγωνο με $M-1$ βαθμούς ελευθερίας.

Απόδειξη

Έστω g: $\mathbb{R}^M \to \mathbb{R}^+$ είναι μια συνάρτηση που ορίζεται ως εξής:

$$g(y_1, \dots, y_M) = \sum_{i=1}^{M} p_i^0 \varphi\left(\frac{y_i}{p_i^0}\right) \tag{28}$$

Ένα δεύτερης τάξης ανάπτυγμα Taylor της συνάρτησης g γύρω από το $\boldsymbol{p^0}$ για $\widehat{\boldsymbol{p}} = (\hat{p}_1, \dots, \hat{p}_M)^T$ έχει ως αποτέλεσμα:

$$g(\hat{p}_1, \dots, \hat{p}_M) = g(p_1^0, \dots, p_M^0) + \sum_{i=1}^{M} \left(\frac{\partial g(y_1, \dots, y_M)}{\partial y_i}\right)\Bigg|_{\boldsymbol{p}=\boldsymbol{p^0}} (\hat{p}_i - p_i^0) +$$

$$+\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} \left(\frac{\partial^2 g(y_1, \dots, y_M)}{\partial y_i \partial y_j}\right)\Bigg|_{\boldsymbol{p}=\boldsymbol{p^0}} (\hat{p}_i - p_i^0)(\hat{p}_j - p_j^0) + o\left(\left\|\widehat{\boldsymbol{p}} - \boldsymbol{p^0}\right\|^2\right)$$

‘Ομως,

$$g(\hat{p}_1, \dots, \hat{p}_M) = D_{\varphi}(\widehat{\boldsymbol{p}}, \boldsymbol{p^0}), \qquad g(p_1^0, \dots, p_M^0) = D_{\varphi}(\boldsymbol{p^0}, \boldsymbol{p^0}) = \varphi(1) = 0$$

και ισχύει για $\boldsymbol{y} = (y_1, \dots, y_M)$:

$$\left(\frac{\partial g(\boldsymbol{y})}{\partial y_i}\right)\Bigg|_{\boldsymbol{p}=\boldsymbol{p^0}} = \varphi'(1), \qquad \left(\frac{\partial^2 g(\boldsymbol{y})}{\partial y_i \partial y_j}\right)\Bigg|_{\boldsymbol{p}=\boldsymbol{p^0}} = \begin{cases} \varphi''(1)\dfrac{1}{p_i^0}, & j = i \\ 0, & j \neq i \end{cases}$$

Οπότε τελικά θα ισχύει:

$$D_{\varphi}(\widehat{\boldsymbol{p}}, \boldsymbol{p^0}) = \frac{1}{2}\varphi''(1)\sum_{i=1}^{M}\frac{1}{p_i^0}(\hat{p}_i - p_i^0)^2 + o\left(\left\|\widehat{\boldsymbol{p}} - \boldsymbol{p^0}\right\|^2\right)$$

Όμως,

$$n\, o\left(\left\|\widehat{\boldsymbol{p}} - \boldsymbol{p^0}\right\|^2\right) = o_P(1)$$

Αφού από την πρόταση 3.1 ισχύει $\sqrt{n}(\widehat{\boldsymbol{p}} - \boldsymbol{p^0}) \xrightarrow[n\to\infty]{L} N(\boldsymbol{0}, \boldsymbol{\Sigma_{p^0}})$ όπου

$$\boldsymbol{\Sigma_{p^0}} = diag(\boldsymbol{p^0}) - \boldsymbol{p^0}(\boldsymbol{p^0})^{\boldsymbol{T}}$$

Άρα τελικά οι τυχαίες μεταβλητές:

$$T_n^{\varphi}(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) = \frac{2n}{\varphi''(1)} D_{\varphi}(\hat{\boldsymbol{p}}, \boldsymbol{p^0})$$

και

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

έχουν την ίδια ασυμπτωτική κατανομή. Όμως

$$n \sum_{i=1}^{M} \frac{1}{p_i^0} (\hat{p}_i - p_i^0)^2 = \sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p^0})^T \boldsymbol{C} \sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p^0})$$

όπου $\boldsymbol{C}$ είναι ένας $M \times M$ πίνακας που ορίζεται ως $\boldsymbol{C} = diag((\boldsymbol{p^0})^{-1})$. Οπότε για $\boldsymbol{X} = \sqrt{n}\, diag\big((\boldsymbol{p^0})^{-1/2}\big)(\hat{\boldsymbol{p}} - \boldsymbol{p^0})$ θα έχουμε:

$$\sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p^0})^T \boldsymbol{C} \sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p^0}) = \boldsymbol{X}^T \boldsymbol{X}$$

Η ασυμπτωτική συμπεριφορά της τυχαίας μεταβλητής $\boldsymbol{X}$ είναι η κανονική με διάνυσμα μέσης τιμής το $\boldsymbol{0}$ και πίνακα διασποράς-συνδιασποράς που ορίζεται ως εξής:

$$\boldsymbol{L} = diag\big((\boldsymbol{p^0})^{-1/2}\big)\, \boldsymbol{\Sigma_{p^0}}\, diag\big((\boldsymbol{p^0})^{-1/2}\big)$$

Στην συνέχεια θα αποδείξουμε ότι το $\boldsymbol{L}$ είναι προβολή τάξης $M - 1$. Παρατηρούμε ότι ισχύουν:

$$\boldsymbol{L} = \boldsymbol{I} - diag\big((\boldsymbol{p^0})^{-1/2}\big)\, \boldsymbol{p^0}(\boldsymbol{p^0})^T\, diag\big((\boldsymbol{p^0})^{-1/2}\big)$$

και

$$\boldsymbol{L} \times \boldsymbol{L} = \boldsymbol{I} - diag\left((\boldsymbol{p^0})^{-\frac{1}{2}}\right) \boldsymbol{p^0}(\boldsymbol{p^0})^T\, diag\left((\boldsymbol{p^0})^{-\frac{1}{2}}\right) -$$

$$-diag\left((\boldsymbol{p^0})^{-\frac{1}{2}}\right) \boldsymbol{p^0}(\boldsymbol{p^0})^T\, diag\left((\boldsymbol{p^0})^{-\frac{1}{2}}\right) +$$

$$+diag\left((\boldsymbol{p^0})^{-\frac{1}{2}}\right) \boldsymbol{p^0}(\boldsymbol{p^0})^T\, diag\left((\boldsymbol{p^0})^{-\frac{1}{2}}\right) diag\left((\boldsymbol{p^0})^{-\frac{1}{2}}\right) \boldsymbol{p^0}(\boldsymbol{p^0})^T diag\left((\boldsymbol{p^0})^{-\frac{1}{2}}\right)$$

Όμως ισχύει:

$$(\boldsymbol{p^0})^T\, diag\left((\boldsymbol{p^0})^{-\frac{1}{2}}\right) diag\left((\boldsymbol{p^0})^{-\frac{1}{2}}\right) \boldsymbol{p^0} = \boldsymbol{1}$$

Άρα

$$\boldsymbol{L} \times \boldsymbol{L} = \boldsymbol{I} - diag\left((\boldsymbol{p^0})^{-\frac{1}{2}}\right) \boldsymbol{p^0}(\boldsymbol{p^0})^T\, diag\left((\boldsymbol{p^0})^{-\frac{1}{2}}\right) = \boldsymbol{L}$$

Επίσης,

$$rank(\boldsymbol{L}) = rank\left(diag\left((\boldsymbol{p^0})^{-\frac{1}{2}}\right) \boldsymbol{\Sigma_{p^0}}\right) = rank\big(\boldsymbol{C\Sigma_{p^0}}\big) = trace\big(\boldsymbol{C\Sigma_{p^0}}\big)$$

Όμως,

$$\boldsymbol{C\Sigma_{p^0}} = \left(\delta_{ij} - p_j^0\right)_{i,j=1,\dots,M}$$

Τότε θα ισχύει:

$$trace\left(\boldsymbol{C\Sigma_{p^0}}\right) = \sum_{j=1}^{M}\left(1 - p_j^0\right) = M - 1$$

Οπότε από την πρόταση 3.2 θα ισχύει τελικά ότι:

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

Πόρισμα 3.1

Υπό την μηδενική υπόθεση $H_0 : \boldsymbol{p} = \boldsymbol{p^0}$, η ασυμπτωτική κατανομή της ελεγχοσυνάρτησης της φ-απόκλισης $T_n^{\varphi}(\boldsymbol{p^0}, \hat{\boldsymbol{p}})$, είναι η X-τετράγωνο με $M-1$ βαθμούς ελευθερίας.

Απόδειξη

Θεωρούμε την συνάρτηση $\varphi(x) = x\mathrm{h}(x^{-1})$. Αν $\mathrm{h} \in \boldsymbol{\Phi}^*$ τότε $\varphi \in \boldsymbol{\Phi}^*$ και από το θεώρημα 3.1 θα ισχύει ότι :

$$T_n^{\varphi}(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) \xrightarrow[n\to\infty]{L} X_{M-1}^2$$

Λαμβάνοντας υπόψη ότι $\varphi''(1) = h''(1)$ ισχύει:

$$T_n^{\varphi}(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) = \frac{2n}{\varphi''(1)} D_{\varphi}(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) = \frac{2n}{\varphi''(1)} \sum_{j=1}^{M} p_i^0 \varphi\left(\frac{\hat{p}_i}{p_i^0}\right) = \frac{2n}{h''(1)} \sum_{j=1}^{M} p_i^0 \frac{\hat{p}_i}{p_i^0} h\left(\frac{p_i^0}{\hat{p}_i}\right)$$

$$= T_n^h(\boldsymbol{p^0}, \hat{\boldsymbol{p}})$$

Παρατήρηση 3.1

a) Στην περίπτωση της Kullback-Leibler απόκλισης θα ισχύουν τα παρακάτω:

$$T_n^0(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) = 2nD_{Kull}(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) \xrightarrow[n\to\infty]{L} X_{M-1}^2$$

και

$$T_n^0(\boldsymbol{p^0}, \hat{\boldsymbol{p}}) = 2nD_{Kull}(\boldsymbol{p^0}, \hat{\boldsymbol{p}}) \xrightarrow[n\to\infty]{L} X_{M-1}^2$$

Η πρώτη ελεγχοσυνάρτηση είναι το στατιστικό κριτήριο του λόγου πιθανοφάνειας και η δεύτερη το κριτήριο του τροποποιημένου λόγου πιθανοφάνειας.

b) Στην περίπτωση των (h,φ) −αποκλίσεων, η ασυμπτωτική κατανομή των παρακάτω ελεγχοσυναρτήσεων:

$$T_n^{\varphi,h}(\widehat{\boldsymbol{p}},\boldsymbol{p^0}) = \frac{2n}{h'(0)\varphi''(1)} D_\varphi^h(\widehat{\boldsymbol{p}},\boldsymbol{p^0})$$

και

$$T_n^{\varphi,h}(\boldsymbol{p^0},\widehat{\boldsymbol{p}}) = \frac{2n}{h'(0)\varphi''(1)} D_\varphi^h(\boldsymbol{p^0},\widehat{\boldsymbol{p}})$$

είναι η X-τετράγωνο με $M-1$ βαθμούς ελευθερίας.

Με βάση το θεώρημα 3.1, αν το μέγεθος του δείγματος είναι αρκετά μεγάλο μπορούμε να χρησιμοποιήσουμε το $100(1-\alpha)$ εκατοστημόριο, $X^2_{M-1,\alpha}$, της X-τετράγωνο κατανομής με $M-1$ βαθμούς ελευθερίας που ορίζεται από την εξίσωση $\mathbf{P}\left(X^2_{M-1} \geq X^2_{M-1,\alpha}\right) = a$. Με αποτέλεσμα να μπορούμε να απορρίπτουμε την μηδενική υπόθεση H_0 σε επίπεδο σημαντικότητας α, αν ισχύει ότι:

$$T_n^{\varphi}(\widehat{\boldsymbol{p}},\boldsymbol{p^0}) > X^2_{M-1,\alpha} \ \left(\text{ή } T_n^{\varphi}(\boldsymbol{p^0},\widehat{\boldsymbol{p}}) > X^2_{M-1,\alpha}\right) \tag{29}$$

Η σχέση αυτή αποτελεί ένα κανόνα απόφασης για τον έλεγχο καλής προσαρμογής βάσει του στατιστικού κριτηρίου φ-απόκλισης. Βάσει αυτού του κριτηρίου απόφασης ισχύει το παρακάτω θεώρημα.

Θεώρημα 3.2

Έστω $\boldsymbol{p}^* = (p_1^*, \dots, p_M^*)^T$ είναι μια κατανομή πιθανότητας με $\boldsymbol{p}^* \neq \boldsymbol{p^0}$. Η ισχύς του ελέγχου με βάση τον κανόνα απόφασης της σχέσης (23) για $\boldsymbol{p}^* = (p_1^*, \dots, p_M^*)^T$ ορίζεται από την σχέση:

$$\beta_{n,\varphi}(p_1^*, \dots, p_M^*) = 1 - \Phi_n\left(\frac{1}{\sigma_1(\boldsymbol{p}^*)}\left(\frac{\varphi''(1)}{2\sqrt{n}} X^2_{M-1,\alpha} - \sqrt{n} D_\varphi(\boldsymbol{p}^*,\boldsymbol{p^0})\right)\right)$$

με Φ_n να τείνει ομοιόμορφα στην τυποποιημένη κανονική κατανομή $\Phi(x)$ και

$$\sigma_1^2(\boldsymbol{p}^*) = \sum_{i=1}^{M} p_i^*\left(\varphi'\left(\frac{p_i^*}{p_i^0}\right)\right)^2 - \left(\sum_{i=1}^{M} p_i^*\varphi'\left(\frac{p_i^*}{p_i^0}\right)\right)^2 \tag{30}$$

Απόδειξη

Αποδεικνύουμε αρχικά ότι υπό την υπόθεση $H_1: \boldsymbol{p} = \boldsymbol{p}^* \neq \boldsymbol{p^0}$ ισχύει:

$$\sqrt{n}\left(D_\varphi(\widehat{\boldsymbol{p}},\boldsymbol{p^0}) - D_\varphi(\boldsymbol{p}^*,\boldsymbol{p^0})\right) \xrightarrow[n\to\infty]{L} N\left(0, \sigma_1^2(\boldsymbol{p}^*)\right),$$

με $\sigma_1^2(\boldsymbol{p}^*) > 0$ και $\sigma_1^2(\boldsymbol{p}^*)$ να δίνεται από την σχέση (30). Ένα πρώτης τάξης ανάπτυγμα Taylor της συνάρτησης g της σχέσης (28) γύρω από το $\boldsymbol{p}^* = (p_1^*, \dots, p_M^*)^T$ και για $\hat{\boldsymbol{p}} = (\hat{p}_1, \dots, \hat{p}_M)^T$ δίνει:

$$D_\varphi(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) = D_\varphi(\boldsymbol{p}^*, \boldsymbol{p^0}) + \sum_{i=1}^{M} \left(\frac{\partial D_\varphi(\boldsymbol{p}, \boldsymbol{p^0})}{\partial p_i} \right)\Bigg|_{\boldsymbol{p}=\boldsymbol{p}^*} (\hat{p}_i - p_i^*) + o(\|\hat{\boldsymbol{p}} - \boldsymbol{p}^*\|)$$

όπου

$$\left(\frac{\partial D_\varphi(\boldsymbol{p}, \boldsymbol{p^0})}{\partial p_i} \right)\Bigg|_{\boldsymbol{p}=\boldsymbol{p}^*} = \varphi'\left(\frac{p_i^*}{p_i^0}\right), i = 1, \dots, M$$

Υπό την υπόθεση $H_1: \boldsymbol{p} = \boldsymbol{p}^*$ έχουμε:

$$\sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p}^*) \xrightarrow[n\to\infty]{L} N(\boldsymbol{0}, \boldsymbol{\Sigma_{p^*}})$$

με $\boldsymbol{\Sigma_{p^*}} = diag(\boldsymbol{p}^*) - \boldsymbol{p}^*(\boldsymbol{p}^*)^T$ τότε $\sqrt{n}\, o(\|\hat{\boldsymbol{p}} - \boldsymbol{p}^*\|) = o_P(1)$. Άρα η ασυμπτωτική κατανομή των τυχαίων μεταβλητών:

$$\sqrt{n}\left(D_\varphi(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) - D_\varphi(\boldsymbol{p}^*, \boldsymbol{p^0})\right) \; και \sqrt{n}\sum_{i=1}^{M} t_i(\hat{p}_i - p_i^*)$$

$$με\; t_i = \varphi'\left(\frac{p_i^*}{p_i^0}\right), i = 1, \dots, M$$

είναι η ίδια. Όμως για $\boldsymbol{T} = (\tau = t_1, \dots, t_M)^T$ ισχύει:

$$\sqrt{n}\sum_{i=1}^{M} t_i(\hat{p}_i - p_i^*) = \sqrt{n}\,\boldsymbol{T}^T(\hat{\boldsymbol{p}} - \boldsymbol{p}^*)$$

Συγκλίνει κατά κανόνα στην κανονική κατανομή με μέση τιμή μηδέν και διασπορά $\boldsymbol{T}^T\boldsymbol{\Sigma_{p^*}}\boldsymbol{T} = \sigma_1^2(\boldsymbol{p}^*)$. Τότε θα έχουμε:

$$\begin{aligned}\beta_{n,\varphi}(p_1^*, \dots, p_M^*) &= Pr\left(T_n^\varphi(\hat{\boldsymbol{p}}, \boldsymbol{p^0}) > X_{M-1,\alpha}^2 | H_1: \boldsymbol{p} = \boldsymbol{p}^*\right) \\ &= 1 - \Phi_n\left(\frac{1}{\sigma_1(\boldsymbol{p}^*)}\left(\frac{\varphi''(1)}{2\sqrt{n}}X_{M-1,\alpha}^2 - \sqrt{n}D_\varphi(\boldsymbol{p}^*, \boldsymbol{p^0})\right)\right)\end{aligned}$$

με $\Phi_n(x)$ να τείνει στην τυποποιημένη κανονική κατανομή $\Phi(x)$ και $\sigma_1(\boldsymbol{p}^*)$ δίνεται από την σχέση (30). Με βάση αυτό το αποτέλεσμα η προσέγγιση της συνάρτησης ισχύος του τεστ με βάση τον κανόνα απόφασης της σχέσης (29) για $\boldsymbol{p}^* = (p_1^*, \dots, p_M^*)^T$ είναι:

$$\beta_{n,\varphi}(p_1^*, \dots, p_M^*) \cong 1 - \Phi\left(\frac{1}{\sigma_1(\boldsymbol{p}^*)}\left(\frac{\varphi''(1)}{2\sqrt{n}}X_{M-1,\alpha}^2 - \sqrt{n}D_\varphi(\boldsymbol{p}^*, \boldsymbol{p^0})\right)\right)$$

Ενώ το $\lim_{n\to\infty} \beta_{n,\varphi}(p_1^*, \dots, p_M^*) = 1$ μας δείχνει ότι το τεστ είναι συνεπές.

Πόρισμα 3.2

Με βάση το προηγούμενο θεώρημα ισχύει επίσης:

$$\sqrt{n}\big(D_\varphi(\boldsymbol{p^0}, \widehat{\boldsymbol{p}}) - D_\varphi(\boldsymbol{p^0}, \boldsymbol{p}^*)\big) \xrightarrow[n\to\infty]{L} N\big(0, \sigma_2(\boldsymbol{p}^*)\big),$$

με $\sigma_2(\boldsymbol{p}^*)$ να δίνεται από την σχέση:

$$\sigma_2^2(\boldsymbol{p}^*) = \sum_{i=1}^{M} p_i^* s_i^2 - \left(\sum_{i=1}^{M} p_i^* s_i\right)^2$$

και

$$s_i = \varphi\left(\frac{p_i^0}{p_i^*}\right) - \frac{p_i^0}{p_i^*}\varphi'\left(\frac{p_i^0}{p_i^*}\right), \quad i = 1, \dots, M$$

Πόρισμα 3.3

a) Στην περίπτωση του μέτρου απόκλισης Kullback-Leibler ισχύει:

$$\sigma_1^2(\boldsymbol{p}^*) = \sum_{i=1}^{M} p_i^* \left(\log\frac{p_i^*}{p_i^0}\right)^2 - \left(\sum_{i=1}^{M} p_i^* \log\frac{p_i^*}{p_i^0}\right)^2,$$

$$\sigma_2^2(\boldsymbol{p}^*) = \sum_{i=1}^{M} \frac{(p_i^0)^2}{p_i^*} - 1$$

b) Στην περίπτωση των (h, φ) αποκλίσεων ισχύει:

$$\sigma_1^2(\boldsymbol{p}^*) = \sum_{i=1}^{M} p_i^* \left(h'\big(D_\varphi(\boldsymbol{p}^*, \boldsymbol{p^0})\big)\varphi'\left(\frac{p_i^*}{p_i^0}\right)\right)^2 - \left(\sum_{i=1}^{M} p_i^* \, h'\big(D_\varphi(\boldsymbol{p}^*, \boldsymbol{p^0})\big)\varphi'\left(\frac{p_i^*}{p_i^0}\right)\right)^2,$$

και

$$\sigma_2^2(\boldsymbol{p}^*) = \sum_{i=1}^{M} p_i^* \left(\Big(h'\big(D_\varphi(\boldsymbol{p^0}, \boldsymbol{p}^*)\big)\Big)\left(\varphi\left(\frac{p_i^0}{p_i^*}\right) - \frac{p_i^0}{p_i^*}\varphi'\left(\frac{p_i^0}{p_i^*}\right)\right)\right)^2$$
$$- \left(\sum_{i=1}^{M} p_i \Big(h'\big(D_\varphi(\boldsymbol{p^0}, \boldsymbol{p}^*)\big)\Big)\left(\varphi\left(\frac{p_i^0}{p_i^*}\right) - \varphi'\left(\frac{p_i^0}{p_i^*}\right)\right)\right)^2$$

Προκειμένου να παραχθεί ένας ασυμπτωτικός έλεγχος ισχύος, ο Cochran (1952) πρότεινε ένα σύνολο τοπικών εναλλακτικών κοντινών (contiguous) στην μηδενική

υπόθεση καθώς το n (μέγεθος δείγματος) αυξάνει. Θεωρούμε το πολυωνυμικό διάνυσμα πιθανοτήτων:

$$\boldsymbol{p}_n \equiv \boldsymbol{p}^0 + \frac{\boldsymbol{d}}{\sqrt{n}}$$

με $\boldsymbol{d} = (d_1, \dots, d_M)^{\boldsymbol{T}}$ ένα σταθερό $M \times 1$ διάνυσμα τέτοιο ώστε να ισχύει $\sum_{j=1}^{M} d_j = 0$ και n είναι ο συνολικός αριθμός παραμέτρων στην πολυωνυμική κατανομή. Καθώς $n \to \infty$ η ακολουθία των διανυσμάτων πιθανότητας $\{\boldsymbol{p}_n\}_{n\in N}$ συγκλίνει στο διάνυσμα πιθανότητας $\boldsymbol{p}^0$ ύπο την μηδενική υπόθεση με τάξη $O\left(n^{-\frac{1}{2}}\right)$. Μπορούμε να πούμε ότι η παρακάτω σχέση

$$H_{1,n}: \boldsymbol{p} = \boldsymbol{p}_n \equiv \boldsymbol{p}^0 + \frac{\boldsymbol{d}}{\sqrt{n}}$$

είναι μια ακολουθία συναφών-κοντινών εναλλακτικών υποθέσεων ως προς την μηδενική υπόθεση $\boldsymbol{p}^0$. Αν θεωρήσουμε $\boldsymbol{p}^* \neq \boldsymbol{p}^0$ τότε θα ισχύει ότι $\boldsymbol{p}^* = \boldsymbol{p}^0 + n^{-1/2}\left(\sqrt{n}(\boldsymbol{p}^* - \boldsymbol{p}^0)\right)$ και αν θεωρήσουμε ότι $\boldsymbol{p}_n \equiv \boldsymbol{p}^0 + \boldsymbol{d}/\sqrt{n}$ με $\boldsymbol{d} = \sqrt{n}(\boldsymbol{p}^* - \boldsymbol{p}^0)$ τότε μπορούμε να θεωρήσουμε την παρακάτω έκφραση ως μια προσέγγιση της συνάρτησης ισχύος στο $\boldsymbol{p}^*$.

$$\beta_{n,\varphi}(\boldsymbol{p}_n) = P\left(T_n^{\varphi}(\widehat{\boldsymbol{p}}, \boldsymbol{p}^{\boldsymbol{0}}) > X^2_{M-1,\alpha} | H_{1,n}: \boldsymbol{p} = \boldsymbol{p}_n\right)$$

3.3 Έλεγχος Υποθέσεων

3.3.1 Έλεγχος Υποθέσεων Παραμετρικών Κατανομών

Έστω $(X, \beta_X, P_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}$ είναι ο στατιστικός χώρος που σχετίζεται με την τυχαία μεταβλητή $\boldsymbol{X}$. Με β_X συμβολίζεται η σ-άλγεβρα των Borel υποσυνόλων $A \subset X$ και $\{P_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}$ είναι μια οικογένεια από κατανομές πιθανότητας που ορίζονται στον αριθμήσιμο χώρο (X, β_X) με Θ ένα ανοιχτό υποσύνολο του $\mathbb{R}^{M_0}$, $M_0 \geq 1$. Έστω $P = \{E_i\}_{i=1,\dots,M}$ είναι μια διαμέριση του χώρου X. Ο τύπος $P_{\boldsymbol{\theta}} = p_i(\boldsymbol{\theta}), i = 1, .., M$ ορίζει ένα διακριτό στατιστικό μοντέλο. Έστω $Y_1, \dots, Y_n$ είναι ένα τυχαίο δείγμα από πληθυσμό που περιγράφεται από την τυχαία μεταβλητή $\boldsymbol{X}$, και έστω $N_i = \sum_{j=1}^{n} I_{E_i}(Y_j)$ και $\widehat{p_i} = N_i/n,\ i = 1, .., M$. Η εκτίμηση του $\boldsymbol{\theta}$ με την μέθοδο μέγιστης πιθανοφάνειας για το διακριτό στατιστικό μοντέλο ορίζεται από την μεγιστοποιήση των $n_1, \dots, n_M$,

$$P_{\boldsymbol{\theta}}(N_1 = n_1, \dots, N_M = n_M) = \frac{n!}{n_1! \dots n_M!} p_1(\boldsymbol{\theta})^{n_1} \times \dots \times p_M(\boldsymbol{\theta})^{n_M}$$

Θα ισχύει

$$\log P_{\boldsymbol{\theta}}(N_1 = n_1, \dots, N_M = n_M) = -nD_{Kull}\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\theta})\right) + k \qquad (31)$$

με $\widehat{\boldsymbol{p}} = (\hat{p}_1, .., \hat{p}_M)^T$, $\boldsymbol{p}(\boldsymbol{\theta}) = \left(p_1(\boldsymbol{\theta}), .., p_M(\boldsymbol{\theta})\right)^T$ και k είναι ανεξάρτητο του $\boldsymbol{\theta}$.

Η σχέση (31) ισχύει επειδή αν ορίσουμε $l(\boldsymbol{\theta}) = \log \boldsymbol{P}_{\boldsymbol{\theta}}(N_1 = n_1, \dots, N_M = n_M)$ παρατηρούμε ότι:

$$l(\boldsymbol{\theta}) = \log \frac{n!}{n_1! \dots n_M!} + n \sum_{i=1}^{M} \widehat{p_i} \log p_i(\boldsymbol{\theta})$$

$$= \log \frac{n!}{n_1! \dots n_M!} - n \sum_{i=1}^{M} \widehat{p_i} \log \frac{1}{p_i(\boldsymbol{\theta})} + n \sum_{i=1}^{M} \widehat{p_i} \log \widehat{p_i} - n \sum_{i=1}^{M} \widehat{p_i} \log \widehat{p_i} =$$

$$= \log \frac{n!}{n_1! \dots n_M!} - n \sum_{i=1}^{M} \widehat{p_i} \log \frac{\widehat{p_i}}{p_i(\boldsymbol{\theta})} + n \sum_{i=1}^{M} \widehat{p_i} \log \widehat{p_i}$$

$$= -n \sum_{i=1}^{M} \widehat{p_i} \log \frac{\widehat{p_i}}{p_i(\boldsymbol{\theta})} + k = -n D_{Kull}\big(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\theta})\big) + k$$

Οπότε εκτιμώντας το $\boldsymbol{\theta}$ με τον εκτιμητή μέγιστης πιθανοφάνειας του διακριτού μοντέλου παρατηρούμε ότι η διαδικασία αυτή είναι ισοδύναμη με την ελαχιστοποίηση της Kullback-Leibler απόκλισης για $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^{M_0}$. Αφού η απόκλιση Kullback-Leibler δεν είναι το μοναδικό μέτρο απόκλισης μπορεί να επιλεγεί ως εκτιμήτρια του $\boldsymbol{\theta}$, η τιμή $\overline{\boldsymbol{\theta}}$ για την οποία θα ισχύει:

$$D\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\overline{\boldsymbol{\theta}})\right) = \inf_{\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^{M_0}} D\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\overline{\boldsymbol{\theta}})\right)$$

με D ένα μέτρο απόκλισης.

Παρακάτω θεωρούμε ότι υπάρχει μια συνάρτηση:

$$\boldsymbol{p}(\boldsymbol{\theta}) = \big(p_1(\boldsymbol{\theta}), .., p_M(\boldsymbol{\theta})\big)^T$$

που απεικονίζει κάθε $\boldsymbol{\theta} = (\theta_1, \dots, \theta_M)^T$ σε ένα σημείο στο Δ_M. Το σύνολο Δ_M είναι το κυρτό σύνολο των μέτρων πιθανότητας που ορίζονται στον δειγματικό χώρο X και $\Delta_M = \{\boldsymbol{p} = (p_1, \dots, p_M)^T : p_i \geq 0,\ i = 1, \dots, M,\ \sum_{i=1}^{M} p_i = 1\}$. Το $\boldsymbol{\theta}$ κυμαίνεται στις τιμές του Θ και το $\boldsymbol{p}(\boldsymbol{\theta})$ κυμαίνεται σε ένα υποσύνολο T του Δ_M. Όταν υποθέτουμε ότι ένα μοντέλο είναι το «σωστό», ουσιαστικά υποθέτουμε ότι υπάρχει μια τιμή $\boldsymbol{\theta}_0 \in \Theta$ τέτοια ώστε $\boldsymbol{p}(\boldsymbol{\theta}_0) = \boldsymbol{\pi}$, όπου $\boldsymbol{\pi}$ είναι η πραγματική τιμή της πολυωνυμικής πιθανότητας, δηλαδή $\boldsymbol{\pi} \in T$.

Ορισμός 3.1

Έστω $Y_1, \dots, Y_n$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό που περιγράφεται από την τυχαία μεταβλητή $\boldsymbol{X}$ που συνδέεται με τον στατιστικό χώρο $(X, \beta_X, P_{\boldsymbol{\theta}})_{\boldsymbol{\theta} \in \boldsymbol{\Theta}}$. Ο εκτιμητής ελάχιστης φ-απόκλισης του $\boldsymbol{\theta}_0$ είναι οποιοδήποτε $\widehat{\boldsymbol{\theta}}_{\varphi} \in \Theta$ το οποίο επαληθεύει την παρακάτω σχέση:

$$D_{\varphi}\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}\left(\widehat{\boldsymbol{\theta}}_{\varphi}\right)\right) = \inf_{\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^{M_0}} D_{\varphi}\big(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\theta})\big)$$

Διαφορετικά μπορούμε να υποθέσουμε ότι ο εκτιμητής ικανοποιεί την σχέση:

$$\widehat{\boldsymbol{\theta}}_{\varphi} = arg \inf_{\boldsymbol{\theta}\in\Theta\subseteq\mathbb{R}^{M_0}} D_{\varphi}\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\theta})\right)$$

Παρατήρηση 3.2

Αν θεωρήσουμε την οικογένεια των μέτρων απόκλισης-δύναμης μπορούμε να λάβουμε την εκτιμήτρια ελάχιστης απόκλισης–δύναμης που μελέτησαν οι Cressie και Read (1984). Η σχέση που ορίζει την εκτιμήτρια είναι η παρακάτω:

$$\hat{\theta}_{(\lambda)} = arg \inf_{\boldsymbol{\theta}\in\Theta\subseteq\mathbb{R}^{M_0}} D_{\varphi(\lambda)}\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\theta})\right)$$

Όπου

$$D_{\varphi(\lambda)}\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\theta})\right) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{M} \hat{p}_i \left(\left(\frac{\hat{p}_i}{p_i(\theta)} \right)^{\lambda} - 1 \right)$$

Για $\lambda \to 0$ λαμβάνουμε την εκτιμήτρια μέγιστης πιθανοφάνειας, για $\lambda = 1$ την ελάχιστη εκτιμήτρια X-τετράγωνο, για $\lambda = -2$ την ελάχιστη εκτιμήτρια του τροποποιημένου X-τετράγωνο κριτηρίου (ή ελάχιστη εκτιμήτρια της τροποποιημένης Neyman ελεγχοσυνάρτησης), για $\lambda \to -1$ λαμβάνουμε την ελάχιστη εκτιμήτρια της τροποποιημένης πιθανοφάνειας, για $\lambda = -0.5$ την Freeman-Tukey εκτιμήτρια και για $\lambda = 2/3$ την Cressie-Read εκτιμήτρια.

3.3.2 Έλεγχος Υποθέσεων για Οικογένεια Παραμετρικών Κατανομών

Στην αρχή του κεφαλαίου εξετάσαμε την μηδενική υπόθεση $H_0 : F = F_0$, δηλαδή την περίπτωση που συνάρτηση κατανομής F είναι γνωστή. Παρακάτω αναλύουμε την περίπτωση που η F ανήκει σε μια οικογένεια παραμετρικών κατανομών $\{F_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\boldsymbol{\theta}}$ με Θ ένα ανοικτό υποσύνολο στο $\mathbb{R}^{M_0}$ και η μηδενική υπόθεση σε αυτή την περίπτωση είναι η:

$$H_0 : F = F_{\boldsymbol{\theta}} \tag{32}$$

Μια προσέγγιση σε αυτό το πρόβλημα είναι να θεωρήσουμε ένα διακριτό στατιστικό μοντέλο που να συνδέεται με το αρχικό μοντέλο. Για να το πραγματοποιήσουμε αυτό αρχικά, θεωρούμε μια διαμέριση $P = \{E_i\}_{i=1,\dots,M}$ του αρχικού δειγματικού χώρου. Οι πιθανότητες των στοιχείων της διαμέρισης $E_i,\ i = 1, \dots, M$ βασίζονται στην άγνωστη παράμετρο $\boldsymbol{\theta}$. Δηλαδή:

$$p_i(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(E_i) = \int_{E_i} dF_{\boldsymbol{\theta}}\,, \quad i = 1, \dots, M$$

Η υπόθεση στην σχέση (32) μπορεί να αναλυθεί αν εξετάσουμε τις παρακάτω υποθέσεις:

$$H_0 \colon \boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta}_0) \in T \ για\ \boldsymbol{\theta}_0 \in \Theta \tag{33}$$

Έναντι της εναλλακτικής υπόθεσης,

$$H_1: \boldsymbol{p} \in \Delta_M - T$$

όπου $T = \{\boldsymbol{p}(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), \dots, p_M(\boldsymbol{\theta}))^T \in \Delta_M : \boldsymbol{\theta} \in \Theta\}$, $\Theta \subset R^{M_0}$ ανοικτό υποσύνολο και $M_0 < M - 1$.

Θεώρημα 3.3

Σύμφωνα με τον Pardo (βλ. [8]) υπό την μηδενική υπόθεση της σχέσης (33) και θεωρώντας ότι ισχύουν οι συνθήκες κανονικότητας του Birch τότε θα ισχύει:

$$T_n^{\varphi_1}(\widehat{\boldsymbol{\theta}}_{\varphi_2}) = \frac{2n}{\varphi_1''(1)} D_{\varphi_1}(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\widehat{\boldsymbol{\theta}}_{\varphi_2})) \xrightarrow[n\to\infty]{L} X_M^2 - M_0 - 1, \quad \varphi_1, \varphi_2 \in \Phi^*$$

Οι συνθήκες κανονικότητας του Birch (1964) για $\boldsymbol{\pi} = \boldsymbol{p}(\boldsymbol{\theta}_0)$, $\boldsymbol{\theta}_0$ είναι άγνωστες παράμετροι και $M_0 < M - 1$ είναι:

1. $\boldsymbol{\theta}_0$ είναι ένα εσωτερικό σημείο του Θ.

2. $\pi_i = p_i(\boldsymbol{\theta}_0) > 0 \; για \; i = 1, \dots, M$. Άρα $\boldsymbol{\pi} = (\pi_1, \dots, \pi_M)^T$ είναι ένα εσωτερικό σημείου του συνόλου Δ_M.

3. Η απεικόνιση $\boldsymbol{p}: \Theta \to \Delta_M$ είναι ολικά διαφορίσιμη στο $\boldsymbol{\theta}_0$, ώστε οι μερικές παράγωγοι του $p_i(\boldsymbol{\theta}_0)$ ως προς κάθε θ_j υπάρχουν για $\boldsymbol{\theta}_0$ και $p_i(\boldsymbol{\theta})$ έχει μια γραμμική προσέγγιση στο $\boldsymbol{\theta}_0$ που δίνεται από την σχέση:

$$p_i(\boldsymbol{\theta}) = p_i(\boldsymbol{\theta}_0) + \sum_{j=1}^{M} (\theta_j - \theta_{0j}) \frac{\partial p_i(\boldsymbol{\theta}_0)}{\partial \theta_j} + o(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|)$$

$$όπου \; ισχύει \; \lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}_0} \frac{o(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|)}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|} = 0$$

4. Ο Ιακωβιανός πίνακας είναι μέγιστης τάξης δηλαδή, M_0.

$$\boldsymbol{J}(\boldsymbol{\theta}_0) = \left(\frac{\partial \boldsymbol{p}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \left(\frac{\partial p_i(\boldsymbol{\theta})}{\partial \theta_j}\right)_{\substack{i=1,\dots,M \\ j=1,\dots,M_0}}$$

5. Η αντίστροφη απεικόνιση $\boldsymbol{p}^{-1} : T \to \Theta$ είναι συνεχής για $\boldsymbol{p}(\boldsymbol{\theta}_0) = \boldsymbol{\pi}$.

6. Η απεικόνιση $\boldsymbol{p}: \Theta \to \Delta_M$ είναι συνεχής σε κάθε σημείο $\boldsymbol{\theta} \in \Theta$.

Οι πηγές που χρησιμοποιήθηκαν σε αυτό το κεφάλαιο αφορούν κυρίως το [8] στην βιβλιογραφία.

ΚΕΦΑΛΑΙΟ 4

Έλεγχοι υποθέσεων καλής προσαρμογής για δεδομένα που περιγράφονται από μοντέλα ευπάθειας

4.1 Εισαγωγή

Στο κεφάλαιο αυτό εξετάζεται μέσω προσομοιωμένων δεδομένων, η συμπεριφορά στατιστικών ελεγχοσυναρτήσεων που βασίζονται στην οικογένεια μέτρων απόκλισης του Csiszar. Κάτω από τη μηδενική υπόθεση υποθέτουμε ότι τα δεδομένα μας τα οποία είναι λογοκριμένα από δεξιά, ακολουθούν ένα μοντέλο ευπάθειας όπως αυτό ορίστηκε στο κεφάλαιο 1, στη σχέση (6). Οι ελεγχοσυναρτήσεις που ορίζουμε είναι γενίκευση αυτών που ορίστηκαν από τους Βόντα και Καραγρηγορίου (βλ. [13]) στην περίπτωση των μοντέλων ευπάθειας. Η αποδοτικότητα των ελέγχων εξετάζεται μέσω προσομοιώσεων για μικρά και μεγάλα δείγματα και για δύο κατανομές της μεταβλητής ευπάθειας, τη Γάμμα και Αντίστροφη Κανονική. Πρέπει να τονίσουμε όμως εδώ ότι οι έλεγχοι αυτοί μπορούν να εφαρμοστούν για οποιαδήποτε κατανομή ευπάθειας με μια μικρή αλλαγή στον κώδικα. Αυτό οφείλεται στον γενικό ορισμό (6) των μοντέλων ευπάθειας.

Με την μεθοδολογία που παρουσιάστηκε στα προηγούμενα κεφάλαια και που θα αναλυθεί παρακάτω, ορίζουμε τις ελεγχοσυναρτήσεις και στη συνέχεια βρίσκουμε την εμπειρική τους κατανομή. Ο λόγος είναι ότι η ασυμπτωτική κατανομή των ελεγχοσυναρτήσεων δεν ειναι επακριβώς γνωστή. Οι έλεγχοι λοιπόν ορίζονται με βάση εμπειρικές κρίσιμες τιμές που ορίζουν τα χωρία απορρίψεως της μηδενικής υπόθεσης. Οι προτεινόμενοι έλεγχοι εξετάζονται ως προς το μέγεθός τους. Όλα τα προγράμματα που χρησιμοποιήθηκαν σε αυτό το κεφάλαιο έχουν κατασκευαστεί στην R.

4.1.1 Ορισμός Ελεγχοσυναρτήσεων μέσω φ-μέτρων απόκλισης

Οι έλεγχοι καλής προσαρμογής που θα ορίσουμε αφορούν δεξιά λογοκριμένα δεδομένα. Έστω ότι έχουμε δείγμα μεγέθους n, με χρόνους αποτυχίας (failure time) και λογοκρισίας (censoring time) που συμβολίζονται με X_i και C_i, $i = 1 \dots n$ αντίστοιχα και με την F να αποτελεί την συνάρτηση κατανομής των X_i. Τα δεδομένα που εξετάζουμε αποτελούνται από τα ζευγάρια (T_i, δ_i) με $T_i = \min(X_i, C_i)$ και $\delta_i =$

$I_{[T_i \le C_i]}$, $όπου\ i = 1 \dots n$. Όπως είχαμε αναλύσει στα προηγούμενα κεφάλαια υποθέτουμε ότι το εύρος των δεδομένων που εξετάζουμε ανήκουν σε ένα χρονικό διάστημα $[0, t]$ το οποίο το διαμερίζουμε σε M διαστήματα $\{E_i = (t_{i-1}, t_i]\}_{i=1,\dots,M}$ με $0 = t_0 < t_1 < \cdots < t_M = t$.

Έστω ότι το πλήθος των χρόνων αποτυχίας και λογοκρισίας σε κάθε διάστημα i συμβολίζονται με d_i και c_i, $i = 1, \dots, M$ αντίστοιχα και έστω ότι n_i είναι ο αριθμός των μονάδων του δείγματος που βρίσκονται σε κίνδυνο στην αρχή του ιοστού διαστήματος, με $n_1 = n$. Έστω, επίσης ότι συμβολίζουμε με $h_i^c = 1 - h_i$ και h_i τους ρυθμούς επιβίωσης και κινδύνου του ιοστού διαστήματος. Στην προσομοίωση που θα αναλύσουμε παρακάτω εξετάζουμε την περίπτωση χωρίς συμμεταβλητές, δηλαδή σύμφωνα με την σχέση (6) θα ισχύει, $S(t|\boldsymbol{X} = 0) = e^{-G(H(t))} = e^{-G(H(t,l))}$ $όπου\ l$ είναι η παράμετρος που υπεισέρχεται στην (βασική) αθροιστική συνάρτηση κινδύνου. Θα επικεντρωθούμε δηλαδή εδώ στην περίπτωση των παραμετρικών μοντέλων ευπάθειας όπου η αθροιστική συνάρτηση κινδύνου είναι γνωστής συναρτησιακής μορφής εκτός από την παράμετρο l που είναι πεπερασμένης διαστάσεως. Υπό την υπόθεση ότι η λογοκρισία συμβαίνει την χρονική στιγμή t_i, θεωρούμε ότι:

$$h_i = \frac{(F(t_i) - F(t_{i-1}))}{1 - F(t_{i-1})}, i = 1, \dots, M$$

Εξετάζουμε την παρακάτω μηδενική υπόθεση:

$$H_0: F = F_0 \text{ ή } ισοδύναμα\ H_0: \boldsymbol{h} = \boldsymbol{h}_0 = (h_{10}, \dots, h_{M0})' \tag{34}$$

όπου ισχύει:

$$H_0: h = h_i = h_{i0} = \frac{(F_0(t_i) - F_0(t_{i-1}))}{1 - F_0(t_{i-1})} = \frac{1 - S_0(t_i) - (1 - S_0(t_{i-1}))}{1 - (1 - S_0(t_{i-1}))} =$$

$$= \frac{S_0(t_{i-1}) - S_0(t_i)}{S_0(t_{i-1})} = \frac{e^{-G(H(t_{i-1},l))} - e^{-G(H(t_i,l))}}{e^{-G(H(t_{i-1},l))}} \tag{35}$$

σύμφωνα με τον ορισμό των μοντέλων ευπάθειας.

Για τον έλεγχο υποθέσεων η μεθοδολογία που ακολουθούμε (βλ. [2] και [13]) όπως και αναφέραμε παραπάνω, στηρίζεται σε μια γενική κλάση ελεγχοσυναρτήσεων για λογοκριμένα δεδομένα, η οποία βασίζεται στην οικογένεια μέτρων απόκλισης του Csiszar και ορίζεται γενικά ως εξής:

$$D^{\varphi}(\boldsymbol{d}, \boldsymbol{n}, \boldsymbol{h}) = \frac{2}{\varphi''(1)} \sum_{\iota=1}^{M} n_i \left(h_i \varphi\left(\frac{d_i/n_i}{h_i}\right) + h_i^c \varphi\left(\frac{{d_i}^c/n_i}{h_i^c}\right)\right)$$

όπου $\boldsymbol{d} = (d_1, \dots, d_M)', \boldsymbol{n} = (n_1, \dots, n_M)',\ \boldsymbol{h} = (h_1, \dots, h_M)', d_i^c = n_i - d_i$.

Όταν η άγνωστη κατανομή ανήκει σε παραμετρική οικογένεια κατανομών $\{\boldsymbol{F_\theta}\}$, με $\boldsymbol{\theta}$ άγνωστη παράμετρο μπορούμε να θεωρήσουμε γενική κλάση ελεγχοσυναρτήσεων με την εξής μορφή:

$$D^{\varphi}(\boldsymbol{d},\boldsymbol{n},\boldsymbol{h}(\boldsymbol{\theta})) = \frac{2}{\varphi''(1)}\sum_{\iota=1}^{M} n_i\left(h_i(\boldsymbol{\theta})\varphi\left(\frac{d_i/n_i}{h_i(\boldsymbol{\theta})}\right) + h_i^c\varphi\left(\frac{{d_i}^c/n_i}{h_i^c(\boldsymbol{\theta})}\right)\right) \quad (36)$$

όπου στην γενική περίπτωση ισχύει ότι $\boldsymbol{h}(\boldsymbol{\theta}) = \big(h_{\boldsymbol{1}}(\boldsymbol{\theta}),\dots,h_M(\boldsymbol{\theta})\big)\colon \boldsymbol{\theta}\in\Theta$, όπου Θ ανοικτό υποσύνολο του $\mathbb{R}^m$. Υποθέτουμε ότι $h_i(\boldsymbol{\theta})$ είναι δύο φορές συνεχώς παραγωγίσιμη και έστω ότι $\boldsymbol{\theta}_0$ είναι η πραγματική τιμή της παραμέτρου και ισχύει $\boldsymbol{h}_0 = \boldsymbol{h}(\boldsymbol{\theta}_0)$**.** Παρατηρούμε ότι οι πιθανότητες αποτυχίας στα διάφορα διαστήματα βασίζονται στην άγνωστη m-διαστάσεων παράμετρο $\boldsymbol{\theta}$, με αποτέλεσμα να απαιτείται η χρήση ενός εκτιμητή $\widehat{\boldsymbol{\theta}}$ της $\boldsymbol{\theta}$ παραμέτρου. Στην περίπτωση των διαμερίσεων $\{E_i\}_{i=1,\dots,M}$ που αναφέραμε παραπάνω μπορούμε να χρησιμοποιήσουμε έναν εκτιμητή ελάχιστης φ-απόκλισης για το $\boldsymbol{\theta}$ υπό την μηδενική υπόθεση:

$H_0\colon F = F_0(\boldsymbol{\theta})$ ή *ισοδύναμα* $H_0\colon \boldsymbol{h} = \boldsymbol{h}_0(\boldsymbol{\theta}) = (h_{10}(\boldsymbol{\theta}),\dots,h_{M0}(\boldsymbol{\theta}))'$.

Ο εκτιμητής τότε θα είναι οποιοδήποτε $\widehat{\boldsymbol{\theta}}_\varphi \in \Theta$ που ικανοποιεί την παρακάτω σχέση:

$$\widehat{\boldsymbol{\theta}}_\varphi = arg\min_{\boldsymbol{\theta}\in\Theta}\sum_{\iota=1}^{M} n_i\left(h_{i0}(\boldsymbol{\theta})\varphi\left(\frac{d_i/n_i}{h_{i0}(\boldsymbol{\theta})}\right) + h_{i0}^c(\boldsymbol{\theta})\varphi\left(\frac{{d_i}^c/n_i}{h_{i0}^c(\boldsymbol{\theta})}\right)\right) \quad (37)$$

Όπως παρατηρούμε ο κάθε εκτιμητής που υπολογίζουμε τελικά θα βασίζεται στην φ-συνάρτηση που έχουμε επιλέξει. Η ασυμπτωτική κατανομή της τελευταίας ελεγχοσυνάρτησης σύμφωνα με τις εργασίες [5] και [7] είναι η X^2_{M-m}, δηλαδή η X τετράγωνο με $M-m$ βαθμούς ελευθερίας. Αυτό το σημείο όμως θέλει περαιτέρω διερεύνηση γιατί η ασυμπτωτική κατανομή φαίνεται να απέχει αρκετά από την εμπειρική κατανομή της ελεγχοσυνάρτησης για κάποιες περιπτώσεις.

Παρατήρηση 4.1

Σημαντική επισήμανση είναι ότι η συνάρτηση φ της σχέσης (37) που χρησιμοποιούμε για την εκτίμηση της άγνωστης παραμέτρου δεν απαιτείται να είναι η ίδια με την συνάρτηση φ που χρησιμοποιούμε στην σχέση (36) που υπολογίζεται η ελεγχοσυνάρτηση.

4.1.2 Προσομοιώσεις

Για τον έλεγχο της μηδενικής υπόθεσης της σχέσης (34) θα επικεντρωθούμε όπως είπαμε και πιο πάνω σε παραμετρικά μοντέλα ευπάθειας ή μετασχηματισμού και πιο συγκεκριμένα σε μοντέλα όπου η βασική αθροιστική συνάρτηση κινδύνου δίνεται από την εκθετική κατανομή. Το μοντέλο ευπάθειας που υποθέτουμε δίνεται από την σχέση

$$S(t) = e^{-G(H(t,l))}, \quad (38)$$

όπου $H(t,l) = t * l$ και για τις προσομοιώσεις υποθέσαμε ότι $l = 1$. Η G συνάρτηση που χρησιμοποιούμε αφορά δύο κατανομές της ευπάθειας, την Γάμμα και συγκεκριμένα τη $\Gamma(1/c, 1/c)$ και την Αντίστροφη Γκαουσιανή $IG(b,b)$. Για αυτές τις δύο περιπτώσεις, η G συνάρτηση θα έχει αντίστοιχα την μορφή

$$G(x,c) = \ln(1+cx)/c$$

και $$G(x,b) = \sqrt{4b(b+x)} - 2b, \ με\ b,c > 0.$$

Η διασπορά της ευπάθειας για την περίπτωση της Γάμμα κατανομής είναι c ενώ για την περίπτωση της Αντίστροφης Γκαουσιανής είναι $1/2b$. Θυμίζουμε ότι η μέση τιμή της ευπάθειας είναι 1. Οι παράμετροι c και b που υπεισέρχονται στην συνάρτηση G θα υποτεθούν γνωστές και πιο συγκεκριμένα η τιμή της c παραμέτρου για την Γάμμα κατανομή για την προσομοίωση θα είναι $c = 1/4$ (και αντιστοιχεί σε μικρή διασπορά της ευπάθειας) και αντίστοιχα για την Αντίστροφη Γκαουσιανή κατανομή θα υποθέσουμε τις τιμές $b = 2$ (μικρή διασπορά) καθώς και $b = 1/8$ (μεγάλη διασπορά). Η τιμή $b = 2$ οδηγεί σε ίση διασπορά της ευπάθειας με την Γάμμα κατανομή με $c = 1/4$ και μας ενδιαφέρει αυτό για σκοπούς σύγκρισης. Η περίπτωση της Γάμμα κατανομής με μεγάλη διασπορά παραλείπεται από την εργασία γιατί παρουσιάζει προβλήματα τα οποία πρέπει να εξετασθούν περαιτέρω.

Οι πραγματικοί ρυθμοί επιβίωσης κάτω από την μηδενική υπόθεση δίνονται από την σχέση (35). Θεωρούμε στα πλαίσια των προσομοιώσεων όπως είπαμε και πιο πάνω μονοδιάστατη παράμετρο $\theta = l$ όπου η εκτιμήτρια της παραμέτρου ορίζεται αντίστοιχα από την σχέση (37).

Οι ελεγχοσυναρτήσεις βασίζονται στις φ-συναρτήσεις απόκλισης που παρουσιάστηκαν στο κεφάλαιο 2. Οι συναρτήσεις απόκλισης που χρησιμοποιούμε στην προσομοίωση παρουσιάζονται παρακάτω:

- συνάρτηση Kullback-Leibler: $\varphi_{KL}(x) = x\log x$
- συνάρτηση Pearson: $\varphi_P(x) = (1-x)^2$
- συνάρτηση Cressie- Read με τύπο:

$$\varphi_{CR}(x) = \frac{x^{\lambda+1} - x - \lambda(x-1)}{\lambda(\lambda+1)}, \qquad \lambda \neq 0, -1$$

- συνάρτηση που βασίζεται στο μέτρο απόκλισης BHHJ και δίνεται από την σχέση:

$$\varphi_{BH}(x) = x^{\lambda+1} - \left(1 + \frac{1}{\lambda}\right)x^{\lambda} + \frac{1}{\lambda}, \ \lambda \neq 0$$

Το μέτρο απόκλισης BHHJ σύμφωνα με τους K. Mattheou, S. Lee, A. Karagrigoriou και Mattheou, A. Karagrigoriou (βλ. [6] και [7]) είναι ένα πρόσφατα προτεινόμενο μέτρο απόκλισης των Basu-Harris-Hjort-Jones, που ορίζεται ως εξής:

$$I_X^{\lambda}(g, f_{\boldsymbol{\theta}}) = \int \left\{ f_{\boldsymbol{\theta}}^{\lambda+1}(z) - \left(1 + \frac{1}{\lambda}\right) g(z) f_{\boldsymbol{\theta}}^{\lambda}(z) + \frac{1}{\lambda} g^{\lambda+1}(z) \right\} dz\,, \qquad \lambda > 0$$

όπου αν $\boldsymbol{X} = (X_1, \dots, X_n)$ τυχαίο διάνυσμα με τα X_i να είναι ανεξάρτητα και ισόνομα κατανεμημένα τότε η $g(.\,, \boldsymbol{\theta_0})$ είναι η άγνωστη πραγματική συνάρτηση κατανομής τους και $\boldsymbol{\theta_0} = \left(\theta_{01}, \dots, \theta_{0p}\right)'$ να είναι η πραγματική αλλά άγνωστη τιμή της p-διάστατης παραμέτρου της κατανομής και $f_{\boldsymbol{\theta}}$ η κατανομή που εξετάζουμε την απόκλιση της από την πραγματική κατανομή. Στα πλαίσια της προσομοίωσης για την ελεγχοσυνάρτηση BHHJ θεωρούμε ότι $\lambda = 10/9$ που από προγενέστερη εργασία, (βλ. [13]) αποτελεί βέλτιστη τιμή για την παράμετρο της ελεγχοσυνάρτησης φ_{BH} και παρακάτω στην εργασία για ευκολία συμβολίζεται ως φ_1. Εξετάζεται επίσης η ελεγχοσυνάρτηση για $\lambda = 0.8$ που παρακάτω θα συμβολίζεται με φ_2. Ομοίως για την ελεγχοσυνάρτηση Cressie-Read θεωρούμε την τιμή $\lambda = 2/3$ η οποία αποτελεί βέλτιστη τιμή για την παράμετρο.

Στις προσομοιώσεις κατασκευάζουμε σε κάθε περίπτωση 20000 τυχαία δείγματα και εξετάζουμε μεγέθη δείγματος μικρά και μεγάλα, δηλαδή $n = 20, 50, 100, 200$. Τα δεδομένα που μελετούμε είναι λογοκριμένα από δεξιά με ποσοστά λογοκρισίας $Censor = 10\%, 30\%, 50\%$ έτσι ώστε να δούμε αν το ποσοστό λογοκρισίας (μικρό ή μεγάλο) παίζει ρόλο στη συμπεριφορά των ελέγχων.
Για τα εύρη διαστήματος $\{E_i = (t_{i-1}, t_i]\}_{i=1,\dots,M}$ της παραπάνω μεθοδολογίας έχουμε επιλέξει να πάρουμε ισομήκη διαστήματα που χωρίζουν το διάστημα [0,t] όπου t η μέγιστη σε κάθε παραγόμενο δείγμα παρατήρηση. Φυσικά τα διαστήματα μπορούν να χωριστούν και με άλλο τρόπο αλλά βρήκαμε τον πιο πάνω τρόπο πιο αποδοτικό. Το πόσα διαστήματα θα πάρουμε εξαρτάται από το μέγεθος του δείγματος n. Πιο συγκεκριμένα, για μέγεθος δείγματος $n = 20$ έχουμε επιλέξει αριθμό ισομηκών διαστημάτων $M = 3$ και 5, ενώ για $n = 50$, οι τιμές για το M είναι $M = 3, 5, 7$. Για $n = 100, 200$ το $M = 5, 7, 10$.

4.1.3 Υπολογισμός ελεγχοσυνάρτησης – ποσοστημόρια εμπειρικής κατανομής της ελεγχοσυνάρτησης

Οι χρόνοι αποτυχίας κατασκευάζονται τυχαία από τη σχέση (38) για τις κατανομές της Γάμμα και της Αντίστροφης Γκαουσιανής αφού έχουμε παράγει τυχαία τιμές για τη συνάρτηση επιβίωσης S από την ομοιόμορφη κατανομή στο διάστημα (0,1). Μετά κατασκευάζουμε τυχαία τους χρόνους λογοκρισίας (μικρότερους από τους χρόνους αποτυχίας) και με τη βοήθεια τυχαίων χρόνων από κατάλληλη διωνυμική κατανομή ορίζουμε τελικά όταν υπάρχει επιτυχία ο χρόνος γεγονότος να είναι ο χρόνος αποτυχίας και όταν υπάρχει αποτυχία ο χρόνος γεγονότος να είναι ο χρόνος λογοκρισίας. Με αυτή τη διαδικασία μπορούμε να ελέγχουμε το ποσοστό λογοκρισίας ώστε να λαμβάνουμε όποιο ποσοστό λογοκρισίας θέλουμε. Αφότου έχουν κατασκευαστεί οι χρόνοι στους οποίους έχουμε κάποιο γεγονός βρίσκουμε το μέγιστο τους το οποίο θα αποτελέσει τον χρόνο t, δηλαδή το δεξί άκρο του διαστήματος $[0, t]$ που θα διαμερίσουμε. Ανάλογα με την τιμή M το χρονικό διάστημα διαμερίζεται σε M ισομήκη διαστήματα. Στη συνέχεια υπολογίζουμε τις τιμές των d_i, d_i^c και n_i της σχέσης (36). Για τα h_i, h_i^c

γνωρίζουμε την μορφή της συνάρτησης G καθώς και τις διαμερίσεις του $[0,t]$ διαστήματος, οπότε με χρήση της σχέσης (37) μπορούμε να υπολογίσουμε την εκτιμήτρια $\hat{\theta}_{\varphi}$ με βάση εντολή της R που εντοπίζει το ελάχιστο συναρτήσεων.

Επισημαίνουμε ότι η συνάρτηση φ της σχέσης (37) που χρησιμοποιήσαμε στις προσομοιώσεις είναι το μέτρο απόκλισης BHHJ με $\lambda = 10/9$ για όλες τις ελεγχοσυναρτήσεις. Εδώ εφαρμόσαμε την παρατήρηση 4.1 για διευκόλυνση των υπολογισμών. Για τον υπολογισμό της τιμής της ελεγχοσυνάρτησης αφού υπολογιστεί η εκτιμήτρια ελάχιστης απόκλισης του θ χρησιμοποιήθηκε η σχέση (36) για τις διάφορες συναρτήσεις φ.

Έχοντας βρει τις τιμές της ελεγχοσυνάρτησης για τα 20000 δείγματα, υπολογίζονται τα $90\%, 95\%, 99\%$ εμπειρικά ποσοστημόρια που θα χρησιμοποιήσουμε στο δεύτερο κομμάτι της προσομοίωσης. Τα αποτελέσματα παρουσιάζονται στους παρακάτω πίνακες.

Πίνακας 1

G-INV GAUSSIAN / $b = 0.125$						
n	M	Censor	φ	quant_90	quant_95	quant_99
20	**3**	**50%**	CR	5.990166	7.810775	11.68536
20	3	50%	φ_2	6.190935	8.044235	12.0705
20	3	50%	φ_1	6.518678	8.52253	12.84439
20	3	50%	PR	6.292985	8.199012	12.29596
20	3	50%	KB	6.1704	7.993872	11.9984
20	**5**	50%	CR	9.274828	11.57289	16.11574
20	5	50%	φ_2	9.632056	12.01076	16.77607
20	5	50%	φ_1	10.39374	13.07265	18.26766
20	5	50%	PR	9.937933	12.44141	17.37369
20	5	50%	KB	9.645731	11.86231	16.60833
50	**3**	50%	CR	5.904381	7.6773	11.86766
50	3	50%	φ_2	6.035494	7.873224	12.14403
50	3	50%	φ_1	6.471665	8.562956	13.32036
50	3	50%	PR	6.243617	8.233413	12.76366
50	3	50%	KB	6.03739	7.683403	11.8395
50	**5**	50%	CR	9.334882	11.65326	16.2757
50	5	50%	φ_2	9.631108	12.00397	16.82097
50	5	50%	φ_1	10.61446	13.51125	19.21344
50	5	50%	PR	10.10294	12.72687	18.04398
50	5	50%	KB	9.448284	11.55419	15.9591
50	**7**	50%	CR	12.4895	15.12459	20.73741
50	7	50%	φ_2	12.93653	15.64448	21.3393
50	7	50%	φ_1	14.56134	18.00723	25.03999
50	7	50%	PR	13.7092	16.80177	23.20499
50	7	50%	KB	12.6001	15.01519	20.19398
100	**5**	50%	CR	9.322507	11.67218	16.57544
100	5	50%	φ_2	9.538984	11.92554	16.98473
100	5	50%	φ_1	10.49223	13.65488	19.94996

100	5	50%	PR	10.02407	12.81011	18.66766
100	5	50%	KB	9.165873	11.27583	15.92018
100	**7**	50%	CR	12.35489	15.05869	20.64782
100	7	50%	φ_2	12.69876	15.43998	21.26988
100	7	50%	φ_1	14.4396	18.17914	25.36573
100	7	50%	PR	13.54814	16.83969	23.42602
100	7	50%	KB	12.20865	14.44868	19.65843
100	**10**	50%	CR	16.86823	20.00329	26.43005
100	10	50%	φ_2	17.44452	20.67501	27.18642
100	10	50%	φ_1	20.1951	24.76584	33.36956
100	10	50%	PR	18.75479	22.74809	30.22054
100	10	50%	KB	16.58746	19.1894	25.10712
200	**5**	50%	CR	9.116305	11.32095	17.12388
200	5	50%	φ_2	9.242051	11.46729	17.36569
200	5	50%	φ_1	10.08767	12.97764	20.48263
200	5	50%	PR	9.699331	12.26368	19.08166
200	5	50%	KB	8.872225	10.9284	15.74758
200	**7**	50%	CR	12.24565	14.98547	20.97988
200	7	50%	φ_2	12.45809	15.23888	21.30926
200	7	50%	φ_1	13.95278	17.67227	26.0918
200	7	50%	PR	13.24907	16.57405	24.02995
200	7	50%	KB	11.94284	14.25002	18.89352
200	**10**	50%	CR	16.63244	19.87882	26.59717
200	10	50%	φ_2	17.02725	20.35021	27.04032
200	10	50%	φ_1	19.60111	24.30859	34.28125
200	10	50%	PR	18.33616	22.46445	30.96269
200	10	50%	KB	16.26017	18.77816	24.25389
20	**3**	**30%**	CR	5.429577	7.149811	11.53795
20	3	30%	φ_2	5.565708	7.342685	11.88962
20	3	30%	φ_1	6.134683	8.225659	12.85018
20	3	30%	PR	5.808329	7.777732	12.25396
20	3	30%	KB	5.35388	7.050242	11.66382
20	**5**	30%	CR	8.738519	10.83988	16.07322
20	5	30%	φ_2	9.124093	11.30965	16.85219
20	5	30%	φ_1	10.15642	12.68914	18.28689
20	5	30%	PR	9.576496	11.89655	17.29765
20	5	30%	KB	8.931608	11.22669	16.79486
50	**3**	30%	CR	4.75945	6.622931	11.52596
50	3	30%	φ_2	4.864336	6.770974	11.83492
50	3	30%	φ_1	5.449807	7.819143	13.7913
50	3	30%	PR	5.173196	7.350431	12.86125
50	3	30%	KB	4.752266	6.308493	10.97665
50	**5**	30%	CR	8.329587	10.5592	16.46109
50	5	30%	φ_2	8.664165	10.95298	17.00555

50	5	30%	φ_1	10.00027	12.81327	20.01796
50	5	30%	PR	9.322872	11.84644	18.52615
50	5	30%	KB	8.418038	10.4403	15.71834
50	**7**	30%	CR	11.33118	13.94292	20.59513
50	7	30%	φ_2	11.77481	14.5402	21.42272
50	7	30%	φ_1	14.10468	17.30062	25.60209
50	7	30%	PR	12.92324	15.88402	23.5366
50	7	30%	KB	11.38919	13.92828	19.87873
100	5	30%	CR	7.877476	9.975038	15.96998
100	5	30%	φ_2	8.155152	10.26354	16.34814
100	5	30%	φ_1	8.835773	12.2171	20.23061
100	5	30%	PR	8.432342	11.28202	18.46992
100	5	30%	KB	8.148425	10.12178	14.8442
100	**7**	30%	CR	10.8414	13.45075	19.65582
100	7	30%	φ_2	11.26694	13.90584	20.39886
100	7	30%	φ_1	12.80439	16.89222	25.63301
100	7	30%	PR	12.04041	15.43567	23.12474
100	7	30%	KB	11.1033	13.34202	18.59584
100	**10**	30%	CR	14.81932	18.08683	25.59774
100	10	30%	φ_2	15.53535	18.79962	26.55246
100	10	30%	φ_1	18.07088	23.48971	33.89217
100	10	30%	PR	16.77052	21.11733	30.25643
100	10	30%	KB	15.19461	17.91696	24.55575
200	**5**	30%	CR	7.600804	9.567984	14.16206
200	5	30%	φ_2	7.806701	9.83235	14.4858
200	5	30%	φ_1	8.060016	10.93513	17.57796
200	5	30%	PR	7.883616	10.3314	16.12107
200	5	30%	KB	8.028203	9.798848	13.89786
200	**7**	30%	CR	10.48259	12.83742	18.73802
200	7	30%	φ_2	10.86064	13.17146	19.2602
200	7	30%	φ_1	11.40382	15.45784	24.20267
200	7	30%	PR	11.04259	14.24155	21.81124
200	7	30%	KB	11.14128	13.14971	17.96158
200	**10**	30%	CR	14.58989	17.63584	24.56513
200	10	30%	φ_2	15.20507	18.33317	25.44481
200	10	30%	φ_1	16.50365	22.05371	32.05482
200	10	30%	PR	15.7973	20.01021	28.92319
200	10	30%	KB	15.39456	17.9538	23.3497
20	**3**	**10%**	CR	3.913447	5.887653	10.39658
20	3	10%	φ_2	4.136232	6.148081	10.89563
20	3	10%	φ_1	4.027194	6.351269	11.99116
20	3	10%	PR	3.994979	6.1459	11.32868
20	3	10%	KB	4.324354	6.674673	11.08288
20	**5**	10%	CR	6.352956	9.262042	15.34095

20	5	10%	φ_2	6.668841	9.929404	16.56902
20	5	10%	φ_1	6.605793	9.777286	17.04459
20	5	10%	PR	6.48037	9.571553	16.34936
20	5	10%	KB	7.257414	10.61745	17.91643
50	**3**	10%	CR	3.67867	5.572544	10.64726
50	3	10%	φ_2	3.780779	5.709565	10.96389
50	3	10%	φ_1	4.045586	6.427814	12.57002
50	3	10%	PR	3.89993	6.054719	11.57759
50	3	10%	KB	3.669365	5.384577	10.6356
50	**5**	10%	CR	6.475825	9.087043	16.22762
50	5	10%	φ_2	6.749759	9.471208	17.0768
50	5	10%	φ_1	7.185529	10.47562	20.35293
50	5	10%	PR	6.884499	9.827469	18.5958
50	5	10%	KB	6.785491	9.400413	16.54318
50	**7**	10%	CR	8.582148	11.91994	20.67287
50	7	10%	φ_2	9.026751	12.56835	21.99278
50	7	10%	φ_1	9.555935	13.86879	25.84842
50	7	10%	PR	9.150643	13.03628	23.67497
50	7	10%	KB	9.25603	12.60064	21.77358
100	**5**	10%	CR	5.937934	8.140381	14.61488
100	5	10%	φ_2	6.203084	8.446495	15.12975
100	5	10%	φ_1	6.587577	9.77978	19.0075
100	5	10%	PR	6.332044	9.058326	17.13429
100	5	10%	KB	6.634233	8.257686	13.80542
100	**7**	10%	CR	8.248676	10.96657	18.68211
100	7	10%	φ_2	8.742355	11.50943	19.66526
100	7	10%	φ_1	9.29329	13.44864	25.64957
100	7	10%	PR	8.880777	12.39651	23.0283
100	7	10%	KB	9.21428	11.43938	18.02096
100	**10**	10%	CR	11.31018	14.97404	25.12916
100	10	10%	φ_2	12.00796	15.79511	26.85112
100	10	10%	φ_1	12.86803	18.40368	35.19368
100	10	10%	PR	12.19582	16.90268	30.73735
100	10	10%	KB	12.67973	15.71183	25.04281
200	**5**	10%	CR	5.920002	7.564329	11.89299
200	5	10%	φ_2	6.375925	7.989318	12.34089
200	5	10%	φ_1	6.071323	8.477141	14.77128
200	5	10%	PR	5.978092	8.003928	13.52658
200	5	10%	KB	7.502076	8.969087	12.26324
200	**7**	10%	CR	8.422838	10.49603	16.5759
200	7	10%	φ_2	9.014383	11.11532	17.29614
200	7	10%	φ_1	8.950495	12.2811	21.39608
200	7	10%	PR	8.653009	11.51087	19.36165
200	7	10%	KB	10.22705	12.23299	16.78313

200	**10**	10%	CR	11.59803	14.65485	23.06158
200	10	10%	φ_2	12.43394	15.54967	24.21836
200	10	10%	φ_1	12.56314	17.71422	30.8886
200	10	10%	PR	12.13812	16.45725	27.68206
200	10	10%	KB	13.92581	16.55832	22.59012

Πίνακας 2

G-INV GAUSSIAN / $b = 2$						
n	M	Censor	φ	quant_90	quant_95	quant_99
20	**3**	**50%**	CR	4.66598	5.786858	8.723995
20	3	50%	φ_2	4.942743	6.095835	9.075358
20	3	50%	φ_1	4.483589	5.67788	8.685081
20	3	50%	PR	4.60642	5.724714	8.67288
20	3	50%	KB	5.629743	6.972144	9.875265
20	**5**	50%	CR	7.660473	9.18156	12.73291
20	5	50%	φ_2	8.072183	9.713788	13.35699
20	5	50%	φ_1	7.448709	9.14944	12.71321
20	5	50%	PR	7.555404	9.16733	12.77813
20	5	50%	KB	9.14464	10.85669	14.81828
50	**3**	50%	CR	4.657846	5.980671	8.725693
50	3	50%	φ_2	4.840794	6.262405	9.217035
50	3	50%	φ_1	4.415151	5.616867	8.453647
50	3	50%	PR	4.514183	5.753471	8.51079
50	3	50%	KB	5.366287	7.056272	10.55444
50	**5**	50%	CR	7.725531	9.321537	12.97159
50	5	50%	φ_2	8.080895	9.754848	13.55486
50	5	50%	φ_1	7.43864	9.088949	13.00625
50	5	50%	PR	7.556042	9.128345	13.01081
50	5	50%	KB	9.04746	10.87835	15.05373
50	**7**	50%	CR	10.26856	12.12561	16.33264
50	7	50%	φ_2	10.80308	12.73131	17.13855
50	7	50%	φ_1	10.09035	12.1512	16.95942
50	7	50%	PR	10.19673	12.08631	16.54584
50	7	50%	KB	12.08524	14.17972	18.49094
100	**5**	50%	CR	8.138804	9.824922	13.26228
100	5	50%	φ_2	8.474665	10.23182	13.77963
100	5	50%	φ_1	7.708753	9.351616	13.0387
100	5	50%	PR	7.864626	9.548602	13.10474
100	5	50%	KB	9.49978	11.47385	15.48319
100	**7**	50%	CR	10.72277	12.51354	16.32408
100	7	50%	φ_2	11.17318	12.99713	16.91543
100	7	50%	φ_1	10.32048	12.24789	16.56357
100	7	50%	PR	10.48063	12.33386	16.39418
100	7	50%	KB	12.42471	14.52161	18.86532

100	**10**	50%	CR	14.34369	16.51612	21.49615
100	10	50%	φ_2	15.02623	17.23647	22.39702
100	10	50%	φ_1	14.25156	16.76366	22.41504
100	10	50%	PR	14.23023	16.64651	21.9652
100	10	50%	KB	16.69011	19.08121	24.36199
200	**5**	50%	CR	9.029613	10.81888	15.0071
200	5	50%	φ_2	9.319215	11.13305	15.50948
200	5	50%	φ_1	8.488537	10.14902	14.1445
200	5	50%	PR	8.71371	10.43012	14.46345
200	5	50%	KB	10.31108	12.4314	17.32998
200	**7**	50%	CR	11.65436	13.6699	18.16528
200	7	50%	φ_2	12.076	14.14305	18.70557
200	7	50%	φ_1	11.07472	13.05927	17.72849
200	7	50%	PR	11.32915	13.31261	17.74582
200	7	50%	KB	13.42818	15.78869	20.78775
200	**10**	50%	CR	15.36965	17.58501	22.47911
200	10	50%	φ_2	15.89859	18.17967	23.3151
200	10	50%	φ_1	14.74808	17.36194	22.83955
200	10	50%	PR	15.0212	17.38774	22.60242
200	10	50%	KB	17.65674	20.3291	25.61276
20	**3**	**30%**	CR	3.886245	5.072317	7.953584
20	3	30%	φ_2	4.149783	5.371987	8.347844
20	3	30%	φ_1	3.701596	4.972461	8.026906
20	3	30%	PR	3.75703	4.974798	8.030721
20	3	30%	KB	4.87695	6.025756	9.215404
20	**5**	30%	CR	6.72515	8.315411	11.99562
20	5	30%	φ_2	7.16704	8.855176	12.76302
20	5	30%	φ_1	6.601698	8.278467	12.10212
20	5	30%	PR	6.655635	8.294482	12.06591
20	5	30%	KB	8.358858	10.10748	14.21818
50	**3**	30%	CR	4.182685	5.185942	7.487194
50	3	30%	φ_2	4.486621	5.617104	7.959921
50	3	30%	φ_1	3.752912	4.716739	7.508024
50	3	30%	PR	3.923746	4.929208	7.483765
50	3	30%	KB	5.483447	6.897815	9.429006
50	**5**	30%	CR	6.813567	8.287715	11.89676
50	5	30%	φ_2	7.252002	8.822023	12.56314
50	5	30%	φ_1	6.40992	8.027053	12.15433
50	5	30%	PR	6.595978	8.124072	11.9411
50	5	30%	KB	8.50254	10.3966	14.27104
50	**7**	30%	CR	9.23921	10.97905	15.56921
50	7	30%	φ_2	9.832744	11.68089	16.45344
50	7	30%	φ_1	8.977666	11.03674	16.40746
50	7	30%	PR	9.079123	10.95925	16.0711

50	7	30%	KB	11.45456	13.47605	18.15002
100	**5**	30%	CR	7.541057	9.017631	12.19567
100	5	30%	φ_2	7.988998	9.506459	12.875
100	5	30%	φ_1	6.828095	8.291476	11.76945
100	5	30%	PR	7.114491	8.576738	11.81411
100	5	30%	KB	9.435124	11.29826	15.12823
100	**7**	30%	CR	9.802558	11.56433	15.57072
100	7	30%	φ_2	10.39333	12.25101	16.43166
100	7	30%	φ_1	9.137109	11.04987	15.90317
100	7	30%	PR	9.423681	11.20077	15.66409
100	7	30%	KB	12.1625	14.38428	18.87975
100	**10**	30%	CR	13.01376	15.1808	20.02398
100	10	30%	φ_2	13.82109	16.08583	21.05362
100	10	30%	φ_1	12.5674	15.46198	21.53558
100	10	30%	PR	12.75505	15.17962	20.69797
100	10	30%	KB	16.03994	18.42474	23.70781
200	**5**	30%	CR	9.052906	10.72952	14.28356
200	5	30%	φ_2	9.453849	11.27834	14.99347
200	5	30%	φ_1	8.164452	9.647777	12.90149
200	5	30%	PR	8.545639	10.07708	13.43351
200	5	30%	KB	11.06092	13.28445	17.89039
200	**7**	30%	CR	11.42604	13.39641	17.4491
200	7	30%	φ_2	11.96658	14.09675	18.23102
200	7	30%	φ_1	10.38845	12.2514	16.15365
200	7	30%	PR	10.81184	12.73289	16.58284
200	7	30%	KB	14.03384	16.53302	21.32629
200	**10**	30%	CR	14.62683	16.72988	21.56562
200	10	30%	φ_2	15.40283	17.63248	22.64434
200	10	30%	φ_1	13.49689	15.7661	21.30559
200	10	30%	PR	13.94902	16.09199	21.22633
200	10	30%	KB	18.05977	20.66011	26.28925
20	**3**	**10%**	CR	2.957438	4.430503	7.513498
20	3	10%	φ_2	3.184715	4.710815	7.931842
20	3	10%	φ_1	2.652116	3.964137	7.436331
20	3	10%	PR	2.829954	4.223156	7.507619
20	3	10%	KB	3.869337	5.561951	9.31269
20	**5**	10%	CR	5.127053	6.758814	10.91682
20	5	10%	φ_2	5.581617	7.32876	11.80117
20	5	10%	φ_1	4.693007	6.361185	11.1002
20	5	10%	PR	4.880541	6.521877	11.09173
20	5	10%	KB	6.810296	8.826874	13.6922
50	**3**	10%	CR	3.038666	4.012142	7.272498
50	3	10%	φ_2	3.292623	4.25041	7.556974
50	3	10%	φ_1	2.747103	4.00807	7.419601

50	3	10%	PR	2.848696	3.951987	7.314747
50	3	10%	KB	4.078604	5.015969	8.426772
50	**5**	10%	CR	5.471945	6.98995	11.43313
50	5	10%	φ_2	5.909753	7.5066	12.11622
50	5	10%	φ_1	5.085062	6.827134	12.21266
50	5	10%	PR	5.225974	6.856791	11.85872
50	5	10%	KB	7.184985	8.813499	13.53547
50	**7**	10%	CR	7.469752	9.409828	15.09866
50	7	10%	φ_2	8.081349	10.13278	16.08406
50	7	10%	φ_1	6.987273	9.254583	16.51192
50	7	10%	PR	7.192148	9.320288	15.81237
50	7	10%	KB	9.808599	11.95853	17.93412
100	**5**	10%	CR	5.780819	7.003795	10.49863
100	5	10%	φ_2	6.27698	7.575182	11.13452
100	5	10%	φ_1	5.149936	6.569652	10.99793
100	5	10%	PR	5.393641	6.724679	10.73788
100	5	10%	KB	7.751709	9.332877	12.95574
100	**7**	10%	CR	7.864216	9.541295	13.80741
100	7	10%	φ_2	8.532818	10.26417	14.71299
100	7	10%	φ_1	7.211379	9.212133	14.93043
100	7	10%	PR	7.470224	9.329498	14.36985
100	7	10%	KB	10.3361	12.31763	16.67346
100	**10**	10%	CR	10.51335	12.68333	18.75924
100	10	10%	φ_2	11.40125	13.66713	19.92503
100	10	10%	φ_1	9.843517	12.73383	21.4685
100	10	10%	PR	10.12408	12.72384	20.44861
100	10	10%	KB	13.76361	16.06031	22.00165
200	**5**	10%	CR	6.612358	7.95714	11.02409
200	5	10%	φ_2	7.109726	8.581277	11.96427
200	5	10%	φ_1	5.758074	6.997291	9.986773
200	5	10%	PR	6.114457	7.388292	10.25888
200	5	10%	KB	8.829746	10.74313	14.86023
200	**7**	10%	CR	8.755059	10.30545	13.76375
200	7	10%	φ_2	9.454432	11.12263	14.74488
200	7	10%	φ_1	7.804872	9.470457	13.52435
200	7	10%	PR	8.204279	9.807979	13.39361
200	7	10%	KB	11.625	13.62332	17.92067
200	**10**	10%	CR	11.62664	13.60667	18.42704
200	10	10%	φ_2	12.52414	14.66841	19.72511
200	10	10%	φ_1	10.6489	13.03656	19.14521
200	10	10%	PR	11.04534	13.21806	18.65663
200	10	10%	KB	15.2184	17.70412	22.74673

Πίνακας 3

***G-GAMMA* / $c = 0.25$**						
n	M	Censor	φ	quant_90	quant_95	quant_99
20	**3**	**50%**	CR	4.641548	5.760278	8.70522
20	3	50%	φ_2	4.924843	6.090557	9.044097
20	3	50%	φ_1	4.416882	5.7035	8.703575
20	3	50%	PR	4.529339	5.710336	8.693734
20	3	50%	KB	5.636833	6.942152	9.878424
20	**5**	50%	CR	7.687145	9.171171	12.67974
20	5	50%	φ_2	8.119361	9.699619	13.27012
20	5	50%	φ_1	7.556443	9.228265	12.86814
20	5	50%	PR	7.608408	9.213832	12.82203
20	5	50%	KB	9.115757	10.83516	14.68521
50	**3**	50%	CR	4.727236	5.973536	8.741777
50	3	50%	φ_2	4.936493	6.285207	9.165271
50	3	50%	φ_1	4.477124	5.640438	8.454468
50	3	50%	PR	4.58692	5.793932	8.561622
50	3	50%	KB	5.530987	7.173975	10.53706
50	**5**	50%	CR	7.749361	9.361948	12.93959
50	5	50%	φ_2	8.124787	9.819674	13.60927
50	5	50%	φ_1	7.52566	9.251871	13.02727
50	5	50%	PR	7.619751	9.279718	12.92308
50	5	50%	KB	9.083851	10.88612	15.05498
50	**7**	50%	CR	10.34841	12.13879	16.22182
50	7	50%	φ_2	10.88009	12.75339	17.0326
50	7	50%	φ_1	10.3069	12.44521	17.24178
50	7	50%	PR	10.31531	12.28064	16.72838
50	7	50%	KB	12.10185	14.05958	18.60539
100	**5**	50%	CR	8.343672	9.969151	13.60372
100	5	50%	φ_2	8.69789	10.33229	14.16383
100	5	50%	φ_1	7.881522	9.63102	13.30887
100	5	50%	PR	8.087711	9.760455	13.43032
100	5	50%	KB	9.739045	11.61441	15.94895
100	**7**	50%	CR	10.90102	12.72905	16.5718
100	7	50%	φ_2	11.36906	13.26849	17.16003
100	7	50%	φ_1	10.57635	12.5853	17.12455
100	7	50%	PR	10.6962	12.61871	16.8662
100	7	50%	KB	12.67968	14.78291	19.04678
100	**10**	50%	CR	14.51546	16.79364	21.50286
100	10	50%	φ_2	15.17842	17.55235	22.40332
100	10	50%	φ_1	14.57027	17.37568	23.00546
100	10	50%	PR	14.53402	17.09182	22.46983
100	10	50%	KB	16.82765	19.32365	24.38539
200	**5**	50%	CR	9.305134	11.12419	14.85214

200	5	50%	φ_2	9.603409	11.42787	15.31999
200	5	50%	φ_1	8.748964	10.55268	14.25397
200	5	50%	PR	8.977694	10.78074	14.50138
200	5	50%	KB	10.64652	12.68947	17.22807
200	**7**	50%	CR	12.06545	14.09928	18.61752
200	7	50%	φ_2	12.47309	14.61287	19.2559
200	7	50%	φ_1	11.52905	13.57128	17.99586
200	7	50%	PR	11.7381	13.76619	18.02291
200	7	50%	KB	13.82768	16.23383	21.38906
200	**10**	50%	CR	15.8371	18.15481	23.03856
200	10	50%	φ_2	16.43184	18.79436	23.88273
200	10	50%	φ_1	15.35677	17.85759	23.55159
200	10	50%	PR	15.55241	17.91062	23.17771
200	10	50%	KB	18.22833	20.81408	26.25229
20	**3**	**30%**	CR	3.886734	5.044788	8.019336
20	3	30%	φ_2	4.160841	5.330505	8.441868
20	3	30%	φ_1	3.768182	5.010432	8.165922
20	3	30%	PR	3.792531	5.014713	8.149421
20	3	30%	KB	4.877314	6.030902	9.248544
20	**5**	30%	CR	6.749708	8.26771	11.94522
20	5	30%	φ_2	7.209203	8.799707	12.67254
20	5	30%	φ_1	6.669889	8.355913	12.19788
20	5	30%	PR	6.730742	8.312358	12.01277
20	5	30%	KB	8.315586	9.974303	14.02029
50	**3**	30%	CR	4.178183	5.172551	7.601885
50	3	30%	φ_2	4.509005	5.57596	7.986659
50	3	30%	φ_1	3.722451	4.715877	7.682145
50	3	30%	PR	3.900359	4.914524	7.593252
50	3	30%	KB	5.511029	6.852612	9.252088
50	**5**	30%	CR	6.882168	8.237004	12.03628
50	5	30%	φ_2	7.333614	8.795255	12.73286
50	5	30%	φ_1	6.490466	8.084851	12.60434
50	5	30%	PR	6.643498	8.089205	12.27434
50	5	30%	KB	8.543219	10.39396	14.22928
50	**7**	30%	CR	9.278713	10.95116	15.21673
50	7	30%	φ_2	9.882908	11.6576	16.09825
50	7	30%	φ_1	9.088135	11.1845	16.23499
50	7	30%	PR	9.167382	11.02318	15.71707
50	7	30%	KB	11.46339	13.39329	17.83832
100	**5**	30%	CR	7.562487	9.044843	12.47463
100	5	30%	φ_2	7.997707	9.560847	13.06208
100	5	30%	φ_1	6.883639	8.353895	11.93127
100	5	30%	PR	7.17437	8.585055	12.08152
100	5	30%	KB	9.509573	11.27705	15.23481

100	**7**	30%	CR	9.92427	11.77959	15.98766
100	7	30%	φ_2	10.54115	12.50533	16.89175
100	7	30%	φ_1	9.330762	11.36383	16.47025
100	7	30%	PR	9.605109	11.50615	16.14945
100	7	30%	KB	12.35448	14.57339	19.39946
100	**10**	30%	CR	13.10511	15.30055	20.14852
100	10	30%	φ_2	13.92176	16.22111	21.31918
100	10	30%	φ_1	12.75113	15.61316	21.84269
100	10	30%	PR	12.88349	15.40701	21.05407
100	10	30%	KB	16.22037	18.66724	23.8286
200	**5**	30%	CR	9.159137	10.82517	14.29301
200	5	30%	φ_2	9.596208	11.3713	14.99215
200	5	30%	φ_1	8.236642	9.72135	12.84159
200	5	30%	PR	8.60512	10.16828	13.37829
200	5	30%	KB	11.30438	13.53611	17.98504
200	**7**	30%	CR	11.80768	13.74146	17.65548
200	7	30%	φ_2	12.4303	14.43885	18.6052
200	7	30%	φ_1	10.7443	12.60267	16.60091
200	7	30%	PR	11.18636	13.06494	16.97909
200	7	30%	KB	14.56864	16.97087	21.91864
200	**10**	30%	CR	15.06715	17.21469	22.06766
200	10	30%	φ_2	15.88214	18.18179	23.33568
200	10	30%	φ_1	14.00443	16.33201	21.90102
200	10	30%	PR	14.4561	16.60761	21.72306
200	10	30%	KB	18.66569	21.26368	26.96754
20	**3**	**10%**	CR	2.967732	4.497354	7.76848
20	3	10%	φ_2	3.196684	4.777702	8.147914
20	3	10%	φ_1	2.706839	4.078237	7.628088
20	3	10%	PR	2.85934	4.245707	7.792298
20	3	10%	KB	3.83471	5.564833	9.404414
20	**5**	10%	CR	5.076677	6.846895	11.10013
20	5	10%	φ_2	5.507405	7.391451	12.02496
20	5	10%	φ_1	4.631497	6.458491	11.18672
20	5	10%	PR	4.816016	6.635277	11.19127
20	5	10%	KB	6.725506	8.802214	13.76331
50	**3**	10%	CR	3.007066	4.073756	7.479326
50	3	10%	φ_2	3.269051	4.26847	7.831965
50	3	10%	φ_1	2.784066	4.076591	7.774526
50	3	10%	PR	2.85822	4.063257	7.619837
50	3	10%	KB	3.941289	4.912809	8.418869
50	**5**	10%	CR	5.394265	7.077411	11.62104
50	5	10%	φ_2	5.819921	7.551377	12.29523
50	5	10%	φ_1	5.057904	6.999046	12.59176
50	5	10%	PR	5.179875	6.996933	12.18583

50	5	10%	KB	7.052704	8.755264	13.57178
50	**7**	10%	CR	7.425953	9.412393	15.04862
50	7	10%	φ_2	8.033615	10.1187	16.21622
50	7	10%	φ_1	6.982875	9.333084	16.88035
50	7	10%	PR	7.171843	9.339517	16.1158
50	7	10%	KB	9.679072	11.87724	17.77542
100	**5**	10%	CR	5.74555	6.907204	10.48989
100	5	10%	φ_2	6.228665	7.491707	11.09143
100	5	10%	φ_1	5.099802	6.594114	11.34135
100	5	10%	PR	5.389097	6.698071	11.04623
100	5	10%	KB	7.721179	9.327215	12.62293
100	**7**	10%	CR	7.768221	9.465757	14.13986
100	7	10%	φ_2	8.385939	10.25756	14.91602
100	7	10%	φ_1	7.145024	9.262773	15.90429
100	7	10%	PR	7.3817	9.326944	15.16049
100	7	10%	KB	10.20292	12.24743	16.73286
100	**10**	10%	CR	10.43355	12.5477	18.93993
100	10	10%	φ_2	11.26731	13.54713	20.25644
100	10	10%	φ_1	9.813352	12.68468	22.27071
100	10	10%	PR	10.08305	12.60272	20.58002
100	10	10%	KB	13.54438	15.9783	21.8616
200	**5**	10%	CR	6.57658	7.915511	10.78966
200	5	10%	φ_2	7.100837	8.560181	11.70861
200	5	10%	φ_1	5.709801	6.878715	10.2068
200	5	10%	PR	6.082507	7.307652	10.2597
200	5	10%	KB	8.827203	10.6757	14.72822
200	**7**	10%	CR	8.858165	10.3587	13.83273
200	7	10%	φ_2	9.539803	11.20651	14.93697
200	7	10%	φ_1	7.852112	9.525197	14.10379
200	7	10%	PR	8.279526	9.867448	13.86836
200	7	10%	KB	11.71905	13.78294	17.79711
200	**10**	10%	CR	11.6305	13.56121	18.44349
200	10	10%	φ_2	12.55174	14.56154	19.77493
200	10	10%	φ_1	10.66994	13.06359	19.94123
200	10	10%	PR	11.08156	13.24442	19.28464
200	10	10%	KB	15.18534	17.55641	22.94363

Εδώ παραθέτουμε το 90ο, 95ο και 99ο ποσοστημόριο της X^2 κατανομής με $M-1$ βαθμούς ελευθερίας για τις διάφορες τιμές του M που χρησιμοποιήσαμε για σκοπούς σύγκρισης.

Πίνακας 4

X^2_{M-1}			
M	quant_90	quant_95	quant_99
3	4.60517	5.991465	9.21034
5	7.77944	9.487729	13.2767
7	10.64464	12.59159	16.81189
10	14.68366	16.91898	21.66599

Τα μεγέθη των ελέγχων που προέκυψαν από τη χρήση των ποσοστημορίων της ασυμπτωτικής κατανομής δεν ήταν καλά ακριβώς διότι είπαμε ότι η ασυμπτωτική κατανομή των ελεγχοσυναρτήσεων δεν είναι ακριβώς η X^2 κατανομή με $M-1$ βαθμούς ελευθερίας.

4.2 Μοντέλα παλινδρόμησης για εκτίμηση ποσοστημορίων-κρίσιμων τιμών

Επειδή τα εμπειρικά ποσοστημόρια όπως βλέπουμε και στους πίνακες πιο πάνω εξαρτώνται από το μέγεθος του δείγματος n, από τον αριθμό των διαστημάτων M και από το ποσοστό λογοκρισίας στη συνέχεια θα εξετάσουμε την περίπτωση να βρούμε ποσοστημόρια που δεν θα εξαρτώνται από αυτές τις ποσότητες και θα εφαρμόζονται σε όλες τις περιπτώσεις στο χωρίο απορρίψεως της μηδενικής υπόθεσης για συγκεκριμένη φυσικά συνάρτηση φ.
Σε αυτό το σημείο θα δημιουργήσουμε μοντέλα παλινδρόμησης στα οποία η μεταβλητή απόκρισης θα είναι τα εμπειρικά ποσοστημόρια. Άλλο μοντέλο σχηματίζεται φυσικά για το 90ο, άλλο για το 95ο και άλλο για το 99ο ποσοστημόριο καθώς επίσης άλλα μοντέλα σχηματίζονται για καθε συνάρτηση φ. Οι ανεξάρτητες μεταβλητές θα είναι το n, M και συναρτήσεις αυτών καθώς και το ποσοστό λογοκρισίας censor. Τα δεδομένα για όλα αυτά τα μοντέλα δίνονται στους Πίνακες 1 έως 3. Το μοντέλο πολλαπλής παλινδρόμησης που θεωρούμε γράφεται στη γενική μορφή:

$$y = \beta_0 + \beta_1 n + \beta_2 M + \beta_3 \sqrt{n} + \beta_4 \sqrt{M} + \beta_5 censor + \beta_6 \ln(n) + \beta_7 \ln(M) + \varepsilon \quad (39)$$

Σκοπός μας είναι η εκτίμηση των ποσοστημορίων με βάση τις πιο πάνω ανεξάρτητες μεταβλητές, δηλαδή η εκτίμηση των ανωτέρω συντελεστών παλινδρόμησης σε πρώτη φάση. Επίσης σκοπός μας είναι η εύρεση του 'καλύτερου' και την ίδια στιγμή του λιγότερο πολύπλοκου μοντέλου που θα μας ικανοποιεί για όλες τις περιπτώσεις.

Οι συντελεστές παλινδρόμησης εκτιμούνται και οι σημαντικές ανεξάρτητες μεταβλητές επιλέγονται σύμφωνα με το κριτήριο πληροφορίας του Akaike καθώς και με το κριτήριο του διορθωμένου συντελεστή προσδιορισμού R^2(Adjusted R^2).

Τονίζουμε ότι τα δεδομένα που χρησιμοποιούνται αφορούν κάθε μία από τις πέντε ελεγχοσυναρτήσεις φ-απόκλισης που μελετούμε χωριστά και αντίστοιχα χωριστά και για τις περιπτώσεις που τα δεδομένα προέρχονται από την Αντίστροφη Γκαουσιανή κατανομή με $b = 0.125$ και $b = 2$ ή την Γάμμα κατανομή με $c = 0.25$.

Στους Πίνακες παρακάτω με την χρήση του κριτηρίου πληροφορίας του Akaike επιλέγεται το μοντέλο με την μικρότερη τιμή AIC και καταγράφονται οι p-τιμές του. Όμοια γίνεται και η επιλογή του μοντέλου με βάσει τον προσαρμοσμένο συντελεστή προσδιορισμού, όπου επιλέγεται το μοντέλο με την μεγαλύτερη τιμή $adjusted\ R^2$, χρησιμοποιώντας την εντολή regsubsets του πακέτου leaps της R. Οι εκτιμητές των συντελεστών παλινδρόμησης με τις δύο μεθόδους φαίνονται στους παρακάτω πίνακες για τις διάφορες περιπτώσεις.

Τα αποτελέσματα για την Αντίστροφη Γκαουσιανή κατανομή με $b = 1/8$ είναι τα ακόλουθα:

Πίνακας 5

90% Ποσοστημόριο									
AIC_Selection_Models									
b= 1/8	intercept	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Ln(n)	Ln(M)	$R^2 - adj$
CR	-5.2927		0.7067	-0.0596	3.3464	8.8847			0.9759
Phi0.8	-2.6594		1.1224	-0.0693		8.5672		2.0174	0.9793
Phi 1.1	-13.7600			-0.1422	13.0586	11.2669		-5.4495	0.9564
PR	-12.4280			-0.1076	11.8285	10.2977		-4.8194	0.9648
KB	-1.0049	0.0188	1.0513	-0.3971		6.2416		2.4247	0.9890
P_values									
CR	0.0425		0.1017	0.0801	0.1195	2.68e-15			
Phi 0.8	0.0052		3.80e-06	0.0358		2.26e-15		0.0960	
Phi1.1	2.38e-10			0.0141	0.0004	1.79e-12		0.1736	
PR	3.52e-11			0.0236	0.0001	1.70e-13		0.1472	
KB	0.1852	0.0067	3.53e-08	0.0047		1.95e-15		0.0071	

Πίνακας 6

90% Ποσοστημόριο									
R^2_adj _Selection_Models									
b= 1/8	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log (M)	$R^2 - adj$
CR	-4.8625		0.6849		3.4806	8.8847	-0.2731		0.9760
Phi 0.8	-4.8247		0.7396		3.5994	8.5672	-0.3194		0.9796
Phi 1.1	-14.3101	-0.0073			13.1666	11.2669		-5.6301	0.9571
PR	-12.8398	-0.0054			11.9169	10.2977		-4.9743	0.9649
KB	-1.0049	0.0188	1.0513	-0.3971		6.2416		2.4247	0.9889

Πίνακας 7

95% Ποσοστημόριο									
AIC_Selection_Models									
b= 1/8	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log(M)	$R^2 - adj$
CR	-1.4755		1.2270	-0.1150		8.6892		2.4002	0.9816
Phi 0.8	-1.4324		1.2792	-0.1385		8.2821		2.6367	0.9853
Phi 1.1	-16.7604	-0.0077			18.3055	10.5067		-9.2639	0.9724
PR	-13.8045			-0.1340	15.4659	9.8038		-7.1541	0.9768
KB	0.7154	0.0270	1.0462	-0.6444		5.7119		3.6962	0.9919
P_values									
CR	0.1115		1.28e-06	0.0012		2.68e-15		0.0545	
Phi 0.8	0.0952		1.59e-07	6.83e-05		1.22e-15		0.0240	
Phi 1.1	8.63e-13	0.0056			2.32e-06	2.84e-12		0.0184	
PR	7.93e-13			0.0037	1.48e-06	1.35e-13		0.0262	
KB	0.3035	0.0001	8.32e-09	1.03e-05		2.39e-15		5.10e-05	

Πίνακας 8

95% Ποσοστημόριο									
R^2_adj _Selection_Models									
b= 1/8	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log (M)	$R^2 - adj$
CR	-1.4755		1.2270	-0.1150		8.6892		2.4002	0.9815
Phi 0.8	-1.4324		1.2792	-0.1385		8.2821		2.6366	0.9852
Phi 1.1	-3.6453	-0.0148	2.1105			10.5067	0.7081		0.9730
PR	-14.3251	-0.0069			15.5642	9.8038		-7.3146	0.9775
KB	0.7154	0.0270	1.0462	-0.6444		5.7119		3.6962	0.9918

Πίνακας 9

99% Ποσοστημόριο									
AIC_Selection_Models									
b= 1/8	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log(n)	Log(M)	$R^2 - adj$
CR	-1.7040		1.5934	-0.6331		4.9532	1.9958	3.1605	0.9741
Phi 0.8	-1.3116		1.6218	-0.6704		3.9335	1.9034	3.8193	0.9747
Phi 1.1	-26.2631	-0.0366			26.8677	3.0711	2.6837	-14.825	0.9747
PR	-22.1860			-1.0108	21.1600	3.8805	3.8842	-10.042	0.9754
KB	3.2458		1.2163	-0.3376		2.2097		5.2644	0.9704
P_values									

CR	0.4679		9.73e-06	0.0074		1.81e-06	0.0495	0.0850	
Phi 0.8	0.5863		1.13e-05	0.0060		7.26e-05	0.0672	0.0451	
Phi 1.1	1.55e-09	0.0003			3.97e-07	0.0141	0.0021	0.0050	
PR	5.12e-08			0.0007	1.53e-06	0.0006	0.0029	0.0226	
KB	0.0175		0.0002	5.44e-08		0.0101		0.0046	

Πίνακας 10

99% Ποσοστημόριο									
$\boldsymbol{R^2\_adj}$ **_Selection_Models**									
b= 1/8	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log (M)	$R^2 - adj$
CR	-5.8637		1.0339	-0.6308	5.3904	4.9532	1.9861		0.9741
Phi 0.8	-1.3115		1.6218	-0.6704		3.9335	1.9034	3.8193	0.9747
Phi 1.1	-4.9885	-0.0372	2.9380			3.0711	2.7581		0.9754
PR	-22.1860			-1.0107	21.160	3.8804	3.8842	-10.0423	0.9753
KB	-5.3707			-0.3402	10.3853	2.2096			0.9710

Παρατηρούμε από τους πίνακες ότι υπάρχουν κάποια μοντέλα τα οποία φαίνεται να είναι σημαντικά σύμφωνα με τα δύο κριτήρια και από τις πέντε ελεγχοσυναρτήσεις. Για παράδειγμα μοντέλα με μεταβλητές όπως $n,\ M,\ \log M\,,\ Censor$, καθώς και το μοντέλο με μεταβλητές τα $n,\ Censor,\ \log M\,,\ \sqrt{M}$ και το $M,\ Censor,\ \log M\,,\ \sqrt{n}$ καθώς και μερικοί άλλοι συνδυασμοί που αφορούν μεμονωμένες περιπτώσεις φ-αποκλίσεων.

Ομοίως παρακάτω παρουσιάζονται τα αποτελέσματα για την Αντίστροφη Γκαουσιανή κατανομή με $b = 2$.

Πίνακας 11

90% Ποσοστημόριο									
AIC_Selection_Models									
b=2	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log(M)	$R^2 - adj$
CR	-2.9933	0.0104	0.8491			6.5954		1.8224	0.9792
Phi.8	-2.9468	0.0102	0.8985			6.2177		1.9838	0.9818
Phi 1.1	-3.2364	0.0071	0.8803			7.2471		1.6407	0.9785
PR	-3.1304	0.0085	0.8641			6.9861		1.7235	0.9792
KB	-2.6605	0.0122	1.0179			5.0017		2.3348	0.9806
P_values									

CR	0.0011	8.34e-08	7.30e-05			2.65e-13		0.1050	
Phi 0.8	0.0010	7.41e-08	2.35e-05			5.43e-13		0.0709	
Phi1.1	0.0004	3.18e-05	3.69e-05			1.77e-14		0.1365	
PR	0.0006	2.17e-06	4.46e-05			3.97e-14		0.1171	
KB	0.0083	4.34e-08	3.32e-05			2.04e-09		0.0675	

Πίνακας 12

90% Ποσοστημόριο									
R^2_adj _Selection_Models									
b=2	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log (M)	$R^2 - adj$
CR	-2.9933	0.0104	0.8491			6.5954		1.8224	0.9791
Phi 0.8	-2.9468	0.0102	0.8985			6.2177		1.9838	0.9818
Phi 1.1	-3.2364	0.0071	0.8803			7.2471		1.6407	0.9785
PR	-2.6118	0.0168	0.8475	-0.1698		6.9861		1.9201	0.9791
KB	-2.6604	0.0122	1.0179			5.0017		2.3348	0.9806

Πίνακας 13

95% Ποσοστημόριο									
AIC_Selection_Models									
b=2	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log(M)	$R^2 - adj$
CR	-2.6496	0.0104	0.8918			6.9603		2.3679	0.9795
Phi 0.8	-1.8411	0.0223	0.9058	-0.2456		6.5161		2.9008	0.9825
Phi 1.1	-2.7264	0.0053	1.0912			7.3142		1.7363	0.9824
PR	-2.6984	0.0075	0.9915			7.1484		2.0457	0.9819
KB	-2.3535	0.0138	0.9979			5.3660		3.2098	0.9805
P_values									
CR	0.0065	3.73e-07	0.0001			5.23e-13		0.0547	
Phi 0.8	0.0831	0.0181	6.20e-05	0.1808		1.61e-12		0.0181	
Phi 1.1	0.0028	0.0011	2.16e-06			2.42e-14		0.1236	
PR	0.0033	2.20e-05	1.04e-05			5.07e-14		0.0738	
KB	0.0301	2.65e-08	0.0001			3.43e-09		0.0246	

Πίνακας 14

95% Ποσοστημόριο									
R^2_adj _Selection_Models									
b=2	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log (M)	$R^2 - adj$
CR	-1.9561	0.0216	0.8695	-0.2270		6.9603		2.6308	0.9798
Phi 0.8	-1.8411	0.0223	0.9058	-0.2456		6.5161		2.9008	0.9824
Phi 1.1	-2.7264	0.0053	1.0912			7.3142		1.7363	0.9824
PR	-2.0721	0.0175	0.9715	-0.2050		7.1484		2.2832	0.9821
KB	-8.1936	0.0273		-0.2788	7.7311	5.3660			0.9810

Πίνακας 15

99% Ποσοστημόριο									
AIC_Selection_Models									
b=2	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log(M)	$R^2 - adj$
CR	-0.3831	0.0232	0.9847	-0.3235		6.2968		4.1320	0.9752
Phi 0.8	-0.1835	0.0257	0.9928	-0.3764		5.7427		4.6544	0.9795
Phi 1.1	-0.8246		1.5106			4.9576		2.4733	0.9774
PR	-1.2023		1.2913			5.5326		3.2781	0.9763
KB	0.4262	0.0366	0.9223	-0.4646		4.8122		5.6954	0.9787
P_values									
CR	0.7831	0.0626	0.0007	0.1905		2.19e-09		0.0129	
Phi 0.8	0.8889	0.0315	0.0003	0.1098		4.45e-09		0.0037	
Phi 1.1	0.4880		2.91e-06			1.95e-07		0.1180	
PR	0.3080		2.48e-05			1.90e-08		0.0387	
KB	0.7740	0.0080	0.0023	0.0816		1e-06		0.0019	

Πίνακας 16

99% Ποσοστημόριο									
R^2_adj _Selection_Models									
b=2	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log (M)	$R^2 - adj$
CR	-7.2663	0.0241		-0.3423	8.3071	6.2968			0.9758
Phi 0.8	-13.0686	0.0990		-2.9839	8.8497	5.7427	5.2396		0.9802
Phi 1.1	-4.6353	-0.0021	1.0117		4.6798	4.9576			0.9774
PR	-1.2022		1.2912			5.5326		3.2780	0.9762
KB	-6.6067	0.0365		-0.4623	9.2853	4.8122			0.9793

Παρατηρούμε και εδώ κάποια μοντέλα που έχουν κοινές μεταβλητές και για τις πέντε ελεγχοσυναρτήσεις, όπως για παράδειγμα το μοντέλο με επεξηγηματικές μεταβλητές $n, M, \log M, Censor$, το μοντέλο με μεταβλητές $M, Censor, \log M$ καθώς και με μεταβλητές $n, M, Censor, \log M, \sqrt{n}$.

Τα αποτελέσματα για την Γάμμα κατανομή με $c = 0.25$ είναι τα εξής:

Πίνακας 17

90% Ποσοστημόριο									
AIC_Selection_Models									
C=1/4	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log(M)	$R^2 - adj$
CR	-9.6954	0.0118			8.4485	7.0993		-3.2547	0.9754
Phi 0.8	-10.0315	0.0117			8.9836	6.7814		3.4631	0.9775
Phi 1.1	-10.3888	0.0080			9.0124	7.8392		-3.8252	0.9734
PR	-10.0886	0.0095			8.7781	7.5143		-3.5916	0.9747
KB	-10.6857	0.0142			10.1741	5.6783		-3.8646	0.9759
P_values									
CR	8.30e-12	7.86e-08			0.000196	7.12e-13		0.178694	
Phi 0.8	3.20e-12	7.42e-08			8.60e-05	1.78e-12		0.151	
Phi 1.1	2.43e-12	4.79e-05			0.00011	9.64e-14		0.12190	
PR	3.16e-12	2.72e-06			0.000122	1.74e-13		0.13847	
KB	2.55e-11	3.35e-08			0.000123	2.70e-09		0.16820	

Πίνακας 18

90% Ποσοστημόριο									
R^2_adj _Selection_Models									
C=1/4	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log (M)	$R^2 - adj$
CR	-8.9553	0.0214		-0.1963	8.2714	7.0993		-2.9284	0.9754
Phi 0.8	-9.2662	0.0217		-0.2030	8.8005	6.7814		-3.1257	0.9776
Phi 1.1	-9.6329	0.0178		-0.2005	8.8315	7.8392		-3.4920	0.9734
PR	-9.3150	0.0196		-0.2052	8.5930	7.5143		-3.2505	0.9748
KB	-5.5629	0.0142	0.6864		3.5727	5.6783			0.9759

Πίνακας 19

95% Ποσοστημόριο									
AIC_Selection_Models									
C=1/4	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log(M)	$R^2 - adj$
CR	-7.8372	0.0265		-0.3070	6.5782	7.4118			0.9740
Phi 0.8	-8.1049	0.0269		-0.3162	6.9981	7.0237			0.9778
Phi 1.1	-5.3818	0.0058	0.7980		3.0476	7.9291			0.9774
PR	-10.5875	0.0083			9.9258	7.6997		-3.8614	0.9769
KB	-9.4549	0.0153			7.6404	5.8806			0.9763
P_values									
CR	4.19e-09	0.019		0.162	2e-16	3.91e-12			
Phi 0.8	9.55e-10	0.014		0.136	2e-16	5.64e-12			
Phi 1.1	0.05076	0.0022	0.0801		0.1744	1.69e-13			
PR	3.77e-12	4.66e-05			5.25e-05	3.71e-13		0.132	
KB	3.30e-14	2.77e-08			2e-16	5.20e-09			

Πίνακας 20

95% Ποσοστημόριο									
R^2_adj _Selection_Models									
C=1/4	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log (M)	$R^2 - adj$
CR	-8.7010	0.0253		-0.2779	8.7942	7.4118		-2.6868	0.9740
Phi 0.8	-8.9447	0.0257		-0.2879	9.1526	7.0237		-2.6122	0.9778
Phi 1.1	-3.0201	0.0058	1.1141			7.9291		1.7889	0.9774
PR	-9.7162	0.0196		-0.2311	9.7173	7.6997		-3.4773	0.9771
KB	-7.7414	0.0219			7.7799	5.8806	-0.6263		0.9768

Πίνακας 21

99% Ποσοστημόριο									
AIC_Selection_Models									
C=1/4	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log(M)	$R^2 - adj$
CR	-1.0432	0.0079	1.1145			6.3230		3.2544	0.9699
Phi 0.8	0.0280	0.0250	1.1158	-0.3493		5.7060		4.0828	0.9741
Phi 1.1	-12.8682				15.8409	4.4786		-7.1512	0.9794
PR	-11.3801				13.4342	5.2483		-4.9076	0.9779
KB	0.4964	0.0403	1.0057	-0.5079		5.3162		5.3427	0.9730

P_values									
CR	0.4295	0.0016	0.0005			1.37e-08		0.0668	
Phi 0.8	0.9851	0.0652	0.0004	0.1925		7.06e-08		0.0224	
Phi 1.1	1.50e-12				5.84e-07	9.59e-07		0.0222	
PR	2.00e-11				6.81e-06	4.29e-08		0.1040	
KB	0.7730	0.0112	0.0039	0.0997		2.19e-06		0.0097	

Πίνακας 22

99% Ποσοστημόριο									
R^2_adj **_Selection_Models**									
C=1/4	inter	n	M	$\sqrt{n}$	$\sqrt{M}$	Censor	Log (n)	Log (M)	$R^2 - adj$
CR	-7.4935	0.0227		-0.3047	8.3568	6.3230			0.9704
Phi 0.8	-7.5788	0.0262		-0.3773	8.9189	5.7060			0.9746
Phi 1.1	-12.8682				15.8409	4.4786		-7.1512	0.9793
PR	-11.3801				13.4341	5.2483		-4.9075	0.9778
KB	-6.8941	0.0406		-0.5150	9.4098	5.3162			0.9739

Και εδώ υπάρχουν μερικά σημαντικά μοντέλα που μπορούμε να διακρίνουμε όπως τα $n, M, \log M, Censor$, το μοντέλο με $Censor, \log M, \sqrt{M}$ και εκείνο με μεταβλητές τα $n, M, Censor, \log M, \sqrt{n}$.

Από τα αποτελέσματα των παραπάνω πινάκων επιλέγονται μοντέλα με μεταβλητές που φαίνεται να είναι σημαντικές για όλες τις ελεγχοσυναρτήσεις που εξετάζουμε καθώς και μοντέλα με μεταβλητές που επαναλαμβάνονται στα μοντέλα που έχουμε αναφέρει μέχρι τώρα και πιθανόν να είναι σημαντικές στην εύρεση μιας καλής εκτίμησης ποσοστημορίων. Από αυτά τα μοντέλα επιλέγουμε ένα που είναι σχετικά απλό αλλά παρουσιάζει μεγάλους συντελεστές προσδιορισμού σε όλες τις περιπτώσεις αλλά είναι σημαντικό και σε σχέση με το κριτήριο AIC. Περιέχει τις εξής ανεξάρτητες μεταβλητές παλινδρόμησης και συμβολίζεται ως μοντέλο (I):

$$n, M, Censor, \log M$$

4.3 Υπολογισμός Συντελεστών Παλινδρόμησης

Στο επόμενο κομμάτι των προσομοιώσεων, για κάθε ελεγχοσυνάρτηση και για τις δύο κατανομές που εξετάζουμε, καταγράφουμε τους εκτιμητές των συντελεστών

παλινδρόμησης, τις p-τιμές και τον συντελεστή προσδιορισμού για το παραπάνω γραμμικό μοντέλο. Τα αποτελέσματα φαίνονται στους παρακάτω πίνακες.

Τα αποτελέσματα για την Αντίστροφη Γκαουσιανή κατανομή με $b = 0.125$ παρουσιάζονται παρακάτω:

Πίνακας 23

90% Ποσοστημόριο						
b= 0.125	inter	n	M	Censor	Log(M)	R^2
CR	-2.8663	-0.0027	1.0610	8.8847	1.8730	0.9786
Phi 0.8	-2.8546	-0.0032	1.1303	8.5672	1.9128	0.9815
Phi 1.1	-4.1064	-0.0073	1.3655	11.2669	2.0985	0.9624
PR	-3.5995	-0.0054	1.2372	10.2977	2.0132	0.9693
KB	-2.2178	-0.0006	1.0901	6.2416	1.9649	0.9874
P_Values						
CR	0.00477	0.10296	1.83e-05	3.24e-15	0.13773	
Phi 0.8	0.00363	0.05240	4.02e-06	3.02e-15	0.11583	
Phi 1.1	0.01182	0.01071	0.00035	1.51e-12	0.302799	
PR	0.00841	0.02142	0.00013	1.62e-13	0.236621	
KB	0.00428	0.64354	1.52e-07	2.52e-14	0.04473	

Πίνακας 24

95% Ποσοστημόριο						
b= 0.125	inter	n	M	Censor	Log(M)	R^2
CR	-1.8228	-0.0055	1.2385	8.6892	2.2613	0.9836
Phi 0.8	-1.8444	-0.0066	1.2935	8.2821	2.4598	0.9865
Phi 1.1	-2.5671	-0.0077	1.9004	10.5067	1.4706	0.9758
PR	-2.2556	-0.0069	1.6162	9.8038	1.8099	0.9803
KB	-1.2525	-0.0045	1.1093	5.7119	2.9501	0.9858
P_Values						

CR	0.05665	0.00173	1.28e-06	3.46e-15	0.07036	
Phi 0.8	0.040729	0.000142	2.01e-07	2.29e-15	0.037609	
Phi 1.1	0.09020	0.00555	2.40e-06	2.93e-12	0.44957	
PR	0.06674	0.00231	1.05e-06	9.48e-14	0.25281	
KB	0.13663	0.00357	1.18e-06	5.54e-12	0.01024	

Πίνακας 25

99% Ποσοστημόριο						
b= 0.125	inter	n	M	Censor	Log(M)	R^2
CR	1.5610	-0.0102	1.5544	4.9532	3.5585	0.9766
Phi 0.8	1.5990	-0.0130	1.5912	3.9335	4.1217	0.9777
Phi 1.1	1.5543	-0.0087	2.6535	3.0711	2.4579	0.9694
PR	1.2408	-0.0092	2.0983	3.8804	3.4121	0.9731
KB	2.2288	-0.0163	1.2502	2.2097	4.8530	0.9717
P_Values						
CR	0.257443	0.000183	1.49e-05	2.07e-06	0.054135	
Phi 0.8	0.2535	9.36e-06	1.34e-05	7.11e-05	0.0296	
Phi 1.1	0.4931	0.0345	9.74e-06	0.0341	0.4093	
PR	0.49949	0.00738	1.39e-05	0.00169	0.16235	
KB	0.112210	1.92e-07	0.000261	0.013425	0.010993	

Ομοίως, τα αποτελέσματα για την Αντίστροφη Γκαουσιανή κατανομή με $b = 2$ *είναι* :

Πίνακας 26

90% Ποσοστημόριο						
b=2	inter	n	M	Censor	Log(M)	R^2
CR	-2.9933	0.0104	0.8491	6.5954	1.8224	0.9818
Phi 0.8	-2.9468	0.0102	0.8985	6.2177	1.9838	0.9841
Phi 1.1	-3.2364	0.0071	0.8803	7.2471	1.6407	0.9812

PR	-3.1304	0.0085	0.8641	6.9861	1.7235	0.9818
KB	-2.6605	0.0122	1.0179	5.0017	2.3348	0.9830
P_Values						
CR	0.00119	8.34e-08	7.30e-05	2.65e-13	0.10504	
Phi 0.8	0.00105	7.41e-08	2.35e-05	5.43e-13	0.07094	
Phi 1.1	0.000463	3.18e-05	3.69e-05	1.77e-14	0.136514	
PR	0.000627	2.17e-06	4.46e-05	3.97e-14	0.117113	
KB	0.00831	4.34e-08	3.32e-05	2.04e-09	0.06758	

Πίνακας 27

95% Ποσοστημόριο						
b=2	inter	n	M	Censor	Log(M)	R^2
CR	-2.6496	0.0104	0.8918	6.9603	2.3679	0.9820
Phi 0.8	-2.5912	0.0103	0.9299	6.5161	2.6164	0.9842
Phi 1.1	-2.7264	0.0053	1.0912	7.3142	1.7363	0.9846
PR	-2.6984	0.0075	0.9915	7.1484	2.0457	0.9842
KB	-2.3535	0.0138	0.9979	5.3660	3.2098	0.9829
P_Values						
CR	0.00650	3.73e-07	0.00011	5.23e-13	0.05474	
Phi 0.8	0.00636	2.97e-07	4.71e-05	1.38e-12	0.03099	
Phi 1.1	0.00287	0.00115	2.16e-06	2.42e-14	0.12366	
PR	0.00331	2.20e-05	1.04e-05	5.07e-14	0.07387	
KB	0.030190	2.65e-08	0.000144	3.43e-09	0.024602	

Πίνακας 28

99% Ποσοστημόριο						
b=2	inter	n	M	Censor	Log(M)	R^2
CR	-1.3709	0.0074	1.0164	6.2968	3.7575	0.9777
Phi 0.8	-1.3328	0.0072	1.0296	5.7427	4.2186	0.9810
Phi 1.1	-1.0050	-0.0021	1.4980	4.9576	2.7418	0.9803
PR	-1.0979	0.0012	1.2984	5.5326	3.1228	0.9788
KB	-0.9925	0.01382	0.9678	4.8122	5.1575	0.9799
P_Values						
CR	0.255932	0.001333	0.000543	2.15e-09	0.022026	

Phi 0.8	0.250078	0.001137	0.000305	6.1e-09	0.008400	
Phi 1.1	0.4045	0.3218	3.81e-06	2.32e-07	0.0891	
PR	0.3623	0.5641	2.99e-05	2.99e-08	0.0543	
KB	0.44889	1.51e-06	0.00204	1.62e-06	0.00503	

Τα αποτελέσματα για την Γάμμα κατανομή με $c = 0.25$ είναι:

Πίνακας 29

90% Ποσοστημόριο						
c= 0.25	inter	n	M	Censor	Log(M)	R^2
CR	-3.1416	0.0117	0.8778	7.0993	1.6949	0.9785
Phi 0.8	-3.0627	0.0117	0.9334	6.7814	1.8001	0.9803
Phi 1.1	-3.3971	0.0079	0.9366	7.8392	1.4539	0.9767
PR	-3.2782	0.0095	0.9123	7.5143	1.5497	0.9779
KB	-2.7921	0.0142	1.0575	5.6783	2.0942	0.9789
P_Values						
CR	0.002181	8.04e-08	0.000197	7.13e-13	0.175872	
Phi 0.8	0.00256	7.61e-08	8.63e-05	1.79e-12	0.14865	
Phi 1.1	0.00124	4.90e-05	0.00011	9.64e-14	0.25097	
PR	0.001464	2.78e-06	0.000122	1.74e-13	0.213719	
KB	0.015034	3.41e-08	0.000123	2.69e-09	0.149435	

Πίνακας 30

95% Ποσοστημόριο						
c= 0.25	inter	n	M	Censor	Log(M)	R^2
CR	-2.7308	0.0117	0.9402	7.4118	2.1483	0.9768
Phi 0.8	-2.7290	0.0116	0.9779	7.0237	2.4240	0.9799
Phi 1.1	-3.0201	0.0058	1.1141	7.9291	1.7889	0.9803
PR	-2.8859	0.0083	1.0318	7.6997	1.9509	0.9798
KB	-2.4208	0.0155	1.0606	5.8806	2.8874	0.9793
P_Values						

CR	0.015254	8.27e-07	0.000377	5.08e-12	0.132143	
Phi 0.8	0.012459	5.35e-07	0.000165	8.11e-12	0.081103	
Phi 1.1	0.00456	0.00224	1.76e-05	1.69e-13	0.17441	
PR	0.00654	4.76e-05	5.20e-05	3.69e-13	0.14097	
KB	0.047192	2.87e-08	0.000296	6.54e-09	0.069316	

Πίνακας 31

99% Ποσοστημόριο						
c= 0.25	inter	n	M	Censor	Log(M)	R^2
CR	-1.0432	0.0079	1.1145	6.3230	3.2544	0.9737
Phi 0.8	-1.0387	0.0078	1.1500	5.7060	3.6784	0.9767
Phi 1.1	-0.7447	-0.0017	1.6319	4.4786	2.3745	0.9817
PR	-0.8037	0.0018	1.4063	5.2483	2.7322	0.9805
KB	-1.0545	0.0154	1.0554	5.3162	4.7547	0.9748
P_Values						
CR	0.429596	0.001645	0.000567	1.37e-08	0.066808	
Phi 0.8	0.424731	0.001636	0.000346	7.31e-08	0.037170	
Phi 1.1	0.535	0.398	9.23e-07	1.27e-06	0.137	
PR	0.4958	0.3849	7.41e-06	5.77e-08	0.0843	
KB	0.48480	2.52e-06	0.00326	3.19e-06	0.02152	

Παρατηρούμε ότι ο συντελεστής προσδιορισμού για όλες τις ελεγχοσυναρτήσεις, για τα διαφορετικά ποσοστημόρια που εξετάζουμε και για τις δύο κατανομές είναι αρκετά υψηλός με τιμές που κυμαίνονται από 0.97 έως *και* 0.99 περίπου. Αυτό σημαίνει ότι ένα υψηλό ποσοστό της μεταβλητότητας των μεταβλητών απόκρισης, δηλαδή των ποσοστημορίων, επεξηγείται από αυτό τον συνδυασμό μεταβλητών που επιλέξαμε (μοντέλο (I)).

4.4 Μέγεθος Ελέγχων

Στο τελευταίο μέρος αυτού του κεφαλαίου θα χρησιμοποιήσουμε τις εκτιμήσεις των συντελεστών του μοντέλου (I) που υπολογίστηκαν παραπάνω και θα υπολογίσουμε κρίσιμες τιμές για κάθε κατανομή σύμφωνα με την G συνάρτηση που εξετάζουμε και κάθε φ ελεγχοσυνάρτηση ξεχωριστά. Ο κώδικας είναι όμοια κατασκευασμένος σύμφωνα με την μεθοδολογία που αναπτύχθηκε στην πρώτη προσομοίωση μέχρι και τον υπολογισμό των ελεγχοσυναρτήσεων.

Στην συνέχεια το πρόγραμμα δέχεται τις τιμές των $n, M, Censor$ και αντίστοιχα και το $\log M$, για τις διάφορες περιπτώσεις που εξετάσαμε στο πρώτο κομμάτι της προσομοίωσης και σε συνδυασμό με τους συντελεστές παλινδρόμησης που υπολογίσαμε παραπάνω για το μοντέλο (I) υπολογίζονται οι κρίσιμες τιμές των χωρίων απορρίψεως για επίπεδα σημαντικότητας 1% και 5%. Έχοντας υπολογίσει τις κρίσιμες τιμές, τις συγκρίνουμε με τις τιμές των ελεγχοσυναρτήσεων για τα 20000 τυχαία δείγματα και καταγράφουμε πόσες φορές οι τιμές των ελεγχοσυναρτήσεων είναι μεγαλύτερες από τις τιμές των αντίστοιχων κρίσιμων τιμών. Σκοπός μας είναι τα μεγέθη των προτεινόμενων ελέγχων να είναι περίπου στα επίπεδα σημαντικότητας που έχουμε επιλέξει.

Τα αποτελέσματα για την Αντίστροφη Γκαουσιανή κατανομή με $b = 0.125$ είναι τα παρακάτω:

Πίνακας 32

Inverse Gaussian – $b = 1/8$ – α=1%							
n	M	Censor	CR	φ_2	φ_1	Pearson	Kullback
20	3	**10%**	0.0098	0.00985	0.00735	0.00935	0.0098
20	5	10%	0.00975	0.01085	0.00305	0.00565	0.0124
50	3	10%	0.0116	0.011	0.01105	0.0115	0.0097
50	5	10%	0.01385	0.0135	0.0145	0.01415	0.01325
50	7	10%	0.01365	0.01365	0.0127	0.01285	0.01445
100	5	10%	0.01035	0.0098	0.0109	0.0106	0.0077
100	7	10%	0.0098	0.00965	0.01165	0.01125	0.00775
100	10	10%	0.0107	0.011	0.01275	0.012	0.0111
200	5	10%	0.0064	0.0065	0.006	0.0061	0.0058
200	7	10%	0.00785	0.0074	0.0076	0.008	0.00785
200	10	10%	0.00905	0.0095	0.0093	0.00945	0.00965
20	3	**30%**	0.0074	0.0076	0.0067	0.00715	0.00875
20	5	30%	0.0062	0.0084	0.00235	0.0031	0.01175
50	3	30%	0.0085	0.00845	0.00985	0.00955	0.00795
50	5	30%	0.01095	0.0114	0.01295	0.0116	0.01155
50	7	30%	0.0108	0.01145	0.0115	0.01095	0.01315
100	5	30%	0.00795	0.00835	0.0103	0.0095	0.00675
100	7	30%	0.00765	0.00835	0.01115	0.01015	0.00715
100	10	30%	0.009	0.0099	0.0119	0.01055	0.0102

200	5	30%	0.005	0.00525	0.00575	0.0057	0.00485
200	7	30%	0.0059	0.0066	0.00705	0.00715	0.0069
200	10	30%	0.0075	0.0083	0.00925	0.00895	0.00875
20	3	**50%**	0.0065	0.0068	0.00595	0.0063	0.00815
20	5	50%	0.0028	0.00555	0.00185	0.0021	0.011
50	3	50%	0.00635	0.00705	0.00875	0.00775	0.0071
50	5	50%	0.0077	0.0086	0.01115	0.00985	0.00985
50	7	50%	0.0084	0.0098	0.0098	0.00935	0.0125
100	5	50%	0.00595	0.00655	0.00935	0.0082	0.0054
100	7	50%	0.00615	0.00695	0.01015	0.0085	0.0066
100	10	50%	0.00735	0.00865	0.01115	0.0094	0.0093
200	5	50%	0.00395	0.0045	0.0053	0.005	0.0041
200	7	50%	0.0043	0.00525	0.0067	0.0062	0.00585
200	10	50%	0.0062	0.0075	0.0091	0.00865	0.00795

Γενικά τα αποτελέσματα των ελέγχων είναι καλά για όλες τις ελεγχοσυναρτήσεις. Δηλαδή, ο αριθμός των φορών που οι ελεγχοσυναρτήσεις είναι μεγαλύτερες από την κρίσιμη τιμή που έχουμε υπολογίσει, με την χρήση του γραμμικού μοντέλου παλινδρόμησης (I), είναι ικανοποιητικός σύμφωνα με το επίπεδο σημαντικότητας που έχουμε επιλέξει. Το μέγεθος των ελέγχων δεν υπερβαίνει πουθενά κατά πολύ το 1%.

Παρατηρούμε ότι για όλα τα μεγέθη των δειγμάτων εκτός του 200 και χαμηλό ποσοστό λογοκρισίας 10% τα μεγέθη των ελέγχων είναι πολύ καλά. Αλλά και στην περίπτωση του μεγέθους δείγματος 200 με $M = 10$ τα μεγέθη δεν είναι άσχημα. Όσο το ποσοστό λογοκρισίας αυξάνει οι έλεγχοι συμπεριφέρονται καλά για μεσαίου μεγέθους δείγματα 50 ή 100 ενώ η περίπτωση του μεγέθους 200 χειροτερεύει σταδιακά. Πάντα όμως ο συνδυασμός 200, $M = 10$ δίνει καλύτερα αποτελέσματα για αυτό το μέγεθος δείγματος. Παρατηρούμε επίσης ότι η συνάρτηση φ_1 δίνει σε πολλές περιπτώσεις καλά αποτελέσματα ακόμα για την περίπτωση μεγέθους δείγματος 200. Τα όχι τόσο καλά αποτελέσματα που παίρνουμε για μεγάλο μέγεθος δείγματος οφείλεται ίσως στο γεγονός ότι έπρεπε να θεωρήσουμε και μεγαλύτερες τιμές για το M ή στο ότι οι εκτιμήσεις μέσω του μοντέλου (I) δεν είναι τόσο καλές για μεγάλα δείγματα. Να τονίσουμε επίσης ότι για πολύ μικρό μέγεθος δείγματος (π.χ. 20) τα μεγέθη χειροτερεύουν όσο το ποσοστό λογοκρισίας αυξάνεται αλλά αυτό είναι αναμενόμενο.

Για την περίπτωση που το επίπεδο σημαντικότητας είναι $\alpha = 5\%$ έχουμε τον παρακάτω πίνακα.

Πίνακας 33

Inverse Gaussian – $b = 1/8$ – α=5%							
n	M	Censor	CR	φ_2	φ_1	Pearson	Kullback
20	3	**10%**	0.06655	0.06665	0.0604	0.06105	0.0656
20	5	10%	0.0544	0.0554	0.0468	0.0499	0.0611

50	3	10%	0.0633	0.0598	0.0686	0.06585	0.04555
50	5	10%	0.0566	0.0549	0.05575	0.05615	0.05015
50	7	10%	0.05105	0.0504	0.0464	0.04825	0.05015
100	5	10%	0.047	0.04585	0.05225	0.0502	0.03615
100	7	10%	0.04285	0.0423	0.0458	0.04535	0.0384
100	10	10%	0.0402	0.04125	0.03855	0.03845	0.0417
200	5	10%	0.04515	0.04795	0.0456	0.04595	0.0575
200	7	10%	0.04395	0.04555	0.04315	0.04375	0.05615
200	10	10%	0.0416	0.04275	0.0412	0.0412	0.0529
20	3	**30%**	0.03695	0.0386	0.0349	0.0362	0.04725
20	5	30%	0.0383	0.04065	0.03525	0.03585	0.0495
50	3	30%	0.03365	0.0332	0.03585	0.03495	0.03135
50	5	30%	0.0366	0.0366	0.03565	0.036	0.03785
50	7	30%	0.0372	0.03835	0.03475	0.0359	0.04145
100	5	30%	0.02715	0.02735	0.03195	0.0301	0.02395
100	7	30%	0.028	0.02885	0.0313	0.0299	0.02865
100	10	30%	0.029	0.03015	0.0273	0.028	0.03365
200	5	30%	0.0223	0.024	0.0279	0.0261	0.03245
200	7	30%	0.02605	0.02805	0.02995	0.028	0.0372
200	10	30%	0.0279	0.0303	0.0314	0.02995	0.03905
20	3	**50%**	0.0185	0.0265	0.02095	0.0209	0.03625
20	5	50%	0.024	0.02585	0.02375	0.0239	0.0375
50	3	50%	0.0189	0.01945	0.01915	0.01915	0.021
50	5	50%	0.0256	0.0267	0.02525	0.0254	0.02925
50	7	50%	0.0291	0.0303	0.02895	0.0294	0.03495
100	5	50%	0.0169	0.01725	0.0195	0.0183	0.0167
100	7	50%	0.02095	0.02165	0.0218	0.02165	0.02165
100	10	50%	0.02245	0.0236	0.02255	0.0226	0.0267
200	5	50%	0.0121	0.01315	0.01585	0.0139	0.01805
200	7	50%	0.01655	0.01815	0.0195	0.0183	0.02535
200	10	50%	0.0193	0.0214	0.0227	0.02045	0.02935

Παρατηρούμε εδώ ότι τα μεγέθη ξεκινούν καλά για χαμηλό ποσοστό λογοκρισίας αλλά χειροτερεύουν όσο το ποσοστό λογοκρισίας μεγαλώνει. Η συνάρτηση Kullback φαίνεται να υπερτερεί των άλλων για επίπεδο σημαντικότητας 5%.

Για την περίπτωση της Αντίστροφης Γκαουσιανής κατανομής με $b = 2$ τα αποτελέσματα φαίνονται στον παρακάτω πίνακα:

Πίνακας 34

Inverse Gaussian – $b = 2$ – α=1%							
n	M	Censor	CR	φ_2	φ_1	Pearson	Kullback
20	3	**10%**	0.0177	0.0173	0.01235	0.01325	0.01645

20	5	10%	0.012	0.01175	0.00855	0.0103	0.0119
50	3	10%	0.01295	0.01125	0.0122	0.0124	0.0089
50	5	10%	0.01225	0.01195	0.0125	0.01265	0.0107
50	7	10%	0.01295	0.01295	0.0125	0.01265	0.0127
100	5	10%	0.0079	0.0076	0.00905	0.0089	0.0059
100	7	10%	0.008	0.008	0.00955	0.00905	0.007
100	10	10%	0.0097	0.00985	0.01115	0.0109	0.0091
200	5	10%	0.0065	0.00645	0.00665	0.00665	0.0085
200	7	10%	0.00505	0.0055	0.00645	0.00575	0.0062
200	10	10%	0.0067	0.0075	0.0081	0.0078	0.00655
20	3	**30%**	0.00865	0.00885	0.00785	0.00785	0.01005
20	5	30%	0.0073	0.00825	0.0058	0.00635	0.0095
50	3	30%	0.0072	0.0072	0.00795	0.0077	0.00575
50	5	30%	0.00815	0.0086	0.0099	0.00935	0.0081
50	7	30%	0.0093	0.00995	0.01055	0.0101	0.00995
100	5	30%	0.00465	0.0045	0.0071	0.0063	0.0036
100	7	30%	0.00545	0.00565	0.0078	0.007	0.0056
100	10	30%	0.0071	0.0077	0.00985	0.0093	0.0075
200	5	30%	0.0027	0.00345	0.0049	0.0041	0.0056
200	7	30%	0.00295	0.0031	0.00485	0.00415	0.0043
200	10	30%	0.0045	0.00495	0.0069	0.0064	0.00505
20	3	**50%**	0.00505	0.0055	0.00515	0.00495	0.00675
20	5	50%	0.004	0.0056	0.003	0.00345	0.0074
50	3	50%	0.00385	0.00415	0.0055	0.0048	0.00415
50	5	50%	0.00575	0.0063	0.00765	0.0069	0.00615
50	7	50%	0.00635	0.00735	0.0084	0.0075	0.008
100	5	50%	0.0026	0.00285	0.0053	0.00455	0.0021
100	7	50%	0.00385	0.00415	0.0064	0.00555	0.00405
100	10	50%	0.0051	0.00595	0.00885	0.0072	0.0055
200	5	50%	0.00125	0.00165	0.00305	0.00215	0.00345
200	7	50%	0.0019	0.00215	0.00385	0.00315	0.0029
200	10	50%	0.003	0.00335	0.00605	0.0052	0.0039

Όπως και στον προηγούμενο πίνακα παρατηρούμε ότι τα μεγέθη ξεκινούν καλά για χαμηλό ποσοστό λογοκρισίας αλλά χειροτερεύουν όσο το ποσοστό λογοκρισίας μεγαλώνει. Οι συναρτήσεις φ_1 και Kullback φαίνεται να υπερτερούν των άλλων για επίπεδο σημαντικότητας 1%.

Ομοίως ισχύει και στην περίπτωση για $a = 5\%$:

Πίνακας 35

Inverse Gaussian – $b = 2$ – α=5%							
n	M	Censor	CR	φ_2	φ_1	Pearson	Kullback
20	3	**10%**	0.07685	0.0684	0.07325	0.0782	0.06095
20	5	10%	0.0552	0.0541	0.04995	0.05245	0.0541
50	3	10%	0.0546	0.05025	0.06625	0.06345	0.0379
50	5	10%	0.054	0.05165	0.05535	0.0544	0.0465
50	7	10%	0.04975	0.04925	0.04965	0.0502	0.04715
100	5	10%	0.04015	0.0401	0.04605	0.04345	0.0423
100	7	10%	0.0423	0.04225	0.04605	0.04405	0.0401
100	10	10%	0.04025	0.04055	0.0424	0.0421	0.03985
200	5	10%	0.03915	0.04145	0.04155	0.041	0.04395
200	7	10%	0.03725	0.0389	0.0387	0.03785	0.04125
200	10	10%	0.03695	0.0389	0.0384	0.0373	0.0423
20	3	**30%**	0.03585	0.0368	0.03255	0.03555	0.04345
20	5	30%	0.03235	0.03375	0.0298	0.0311	0.03765
50	3	30%	0.02605	0.025	0.03125	0.0284	0.02265
50	5	30%	0.02945	0.0302	0.03205	0.0316	0.0307
50	7	30%	0.0305	0.0322	0.03195	0.0315	0.0342
100	5	30%	0.0203	0.02225	0.0241	0.0219	0.0265
100	7	30%	0.02375	0.0251	0.0281	0.0266	0.02745
100	10	30%	0.027	0.02895	0.03025	0.02875	0.0289
200	5	30%	0.01855	0.02185	0.0181	0.0179	0.0298
200	7	30%	0.01945	0.02275	0.02135	0.0199	0.02895
200	10	30%	0.0222	0.0253	0.02735	0.02475	0.0315
20	3	**50%**	0.0194	0.02155	0.0168	0.0189	0.0284
20	5	50%	0.01815	0.02055	0.01965	0.0195	0.02635
50	3	50%	0.0137	0.0144	0.01475	0.0142	0.0144
50	5	50%	0.01785	0.01875	0.01925	0.01885	0.02045
50	7	50%	0.0206	0.0223	0.02155	0.02155	0.0246
100	5	50%	0.0118	0.0129	0.01485	0.01385	0.01585
100	7	50%	0.0147	0.0163	0.0177	0.0161	0.0196
100	10	50%	0.01765	0.02	0.02045	0.01965	0.0215
200	5	50%	0.00925	0.0117	0.00915	0.00865	0.0194
200	7	50%	0.01	0.01215	0.0124	0.0102	0.01925
200	10	50%	0.0144	0.0162	0.01855	0.0161	0.0219

Παρακάτω παρουσιάζονται τα μεγέθη ελέγχου για την Γάμμα κατανομή με $c = 1/4 = 0.25$.

Πίνακας 36

Gamma – $c = 1/4$ – α=1%							
n	M	Censor	CR	φ_2	φ_1	Pearson	Kullback
20	3	**10%**	0.0174	0.01825	0.013	0.01345	0.01855

20	5	10%	0.01235	0.01245	0.00825	0.01015	0.01295
50	3	10%	0.0128	0.01175	0.0126	0.01265	0.0093
50	5	10%	0.013	0.0123	0.01285	0.0131	0.01115
50	7	10%	0.0135	0.0133	0.01305	0.0132	0.01275
100	5	10%	0.00815	0.00725	0.00975	0.00945	0.00525
100	7	10%	0.00895	0.0087	0.01065	0.0102	0.00715
100	10	10%	0.00965	0.01	0.01125	0.01065	0.00835
200	5	10%	0.0048	0.00605	0.006	0.0057	0.0072
200	7	10%	0.00535	0.00525	0.0065	0.00635	0.00585
200	10	10%	0.00675	0.007	0.0081	0.0079	0.00735
20	3	**30%**	0.00955	0.00985	0.00855	0.0087	0.0109
20	5	30%	0.0076	0.00855	0.00545	0.0063	0.01
50	3	30%	0.0074	0.00735	0.009	0.00835	0.00635
50	5	30%	0.00855	0.00885	0.01045	0.00965	0.00835
50	7	30%	0.0091	0.0097	0.01095	0.01025	0.0097
100	5	30%	0.00485	0.0049	0.00795	0.0067	0.0034
100	7	30%	0.0061	0.00635	0.00895	0.00835	0.00505
100	10	30%	0.00675	0.00715	0.0101	0.0091	0.0059
200	5	30%	0.00255	0.0032	0.00475	0.0035	0.0047
200	7	30%	0.00305	0.0033	0.00525	0.00475	0.0038
200	10	30%	0.00455	0.0051	0.0071	0.00635	0.005
20	3	**50%**	0.0058	0.00605	0.00555	0.0056	0.00745
20	5	50%	0.00395	0.0056	0.00285	0.0034	0.0075
50	3	50%	0.00425	0.00445	0.00615	0.0055	0.0042
50	5	50%	0.006	0.00645	0.00805	0.0069	0.0061
50	7	50%	0.0061	0.00715	0.00865	0.00755	0.00725
100	5	50%	0.00305	0.00325	0.00625	0.0053	0.00245
100	7	50%	0.00395	0.00435	0.0078	0.0064	0.00365
100	10	50%	0.0049	0.0058	0.0089	0.00745	0.0047
200	5	50%	0.00135	0.00195	0.003	0.00215	0.0032
200	7	50%	0.0018	0.00215	0.00435	0.00315	0.00275
200	10	50%	0.0035	0.00385	0.00645	0.00565	0.00345

Και σε αυτή την περίπτωση που τα δεδομένα προέρχονται από την Γάμμα κατανομή για μικρό μέγεθος δείγματος και χαμηλό ποσοστό λογοκρισίας, παρατηρούμε καλά μεγέθη ελέγχου. Τα μεγέθη χειροτερεύουν όσο το ποσοστό λογοκρισίας αυξάνεται. Οι συναρτήσεις φ_1 και Kullback φαίνεται να υπερτερούν των άλλων για επίπεδο σημαντικότητας 1%.

Όμοια και στην επόμενη περίπτωση:

Πίνακας 37

Gamma – $c = 1/4$ – α=5%							
n	M	Censor	CR	φ_2	φ_1	Pearson	Kullback
20	3	**10%**	0.08035	0.0726	0.08075	0.08115	0.06425
20	5	10%	0.0584	0.0581	0.05205	0.0552	0.0564
50	3	10%	0.05795	0.0538	0.0707	0.06715	0.03835
50	5	10%	0.057	0.0539	0.0592	0.0586	0.0478
50	7	10%	0.05095	0.05005	0.0497	0.05035	0.04835
100	5	10%	0.04015	0.0402	0.04525	0.0424	0.0432
100	7	10%	0.0415	0.0415	0.04475	0.043	0.042
100	10	10%	0.03725	0.0378	0.04015	0.0393	0.03715
200	5	10%	0.03565	0.03845	0.0376	0.0365	0.04095
200	7	10%	0.0354	0.03775	0.0386	0.0362	0.0413
200	10	10%	0.0336	0.0353	0.03825	0.036	0.03845
20	3	**30%**	0.03785	0.03875	0.03475	0.03725	0.04425
20	5	30%	0.03435	0.0357	0.03175	0.03355	0.0385
50	3	30%	0.0261	0.02505	0.0327	0.0292	0.0224
50	5	30%	0.03065	0.03115	0.033	0.03215	0.03065
50	7	30%	0.0317	0.03275	0.03225	0.03185	0.03285
100	5	30%	0.01955	0.02095	0.0244	0.0219	0.0239
100	7	30%	0.0243	0.02535	0.0286	0.02625	0.0277
100	10	30%	0.0246	0.0259	0.0286	0.027	0.0258
200	5	30%	0.0159	0.0192	0.0172	0.01635	0.0263
200	7	30%	0.01775	0.0202	0.02075	0.0185	0.0263
200	10	30%	0.021	0.0232	0.02545	0.0225	0.0277
20	3	**50%**	0.02075	0.0226	0.01725	0.01915	0.0289
20	5	50%	0.0185	0.021	0.0205	0.01965	0.02625
50	3	50%	0.0138	0.0143	0.01595	0.01515	0.013
50	5	50%	0.01855	0.01965	0.01965	0.01935	0.02075
50	7	50%	0.0211	0.02315	0.0231	0.023	0.0243
100	5	50%	0.0108	0.0113	0.0144	0.01355	0.01305
100	7	50%	0.0148	0.0161	0.0178	0.01665	0.0188
100	10	50%	0.01745	0.01905	0.0202	0.01925	0.0191
200	5	50%	0.0066	0.0085	0.0083	0.0074	0.0164
200	7	50%	0.00905	0.0109	0.01255	0.0108	0.0155
200	10	50%	0.01315	0.0153	0.0174	0.01565	0.01945

Συμπεράσματα

- Πρέπει να τονίσουμε ότι οι έλεγχοι που ορίσαμε με βάση τις φ συναρτήσεις για λογοκριμένα δεδομένα μοντέλων ευπάθειας παρουσιάζουν καλά μεγέθη. Τα μεγέθη όπως είναι αναμενόμενο είναι καλύτερα για μικρό ποσοστό λογοκρισίας.

- Μια γενική παρατήρηση είναι επίσης ότι για το μέγεθος δείγματος 20 φαίνεται να είναι καλύτερο να επιλέγονται 5 διαστήματα. Για το μέγεθος δείγματος 50 φαίνεται να πρέπει να επιλέγονται 5 ή 7 διαστήματα, για μέγεθος 100, 7 ή 10 διαστήματα και για μέγεθος 200, 10 ή περισσότερα διαστήματα. Αυτό το σημείο απαιτεί ίσως περισσότερη διερεύνηση.

- Αν και η ασυμπτωτική κατανομή των ελεγχοσυναρτήσεων δεν είναι επακριβώς γνωστή, η μέθοδος που ακολουθήσαμε σε αυτό το κεφάλαιο για να εκτιμήσουμε τις κρίσιμες τιμές των χωρίων απορρίψεως της μηδενικής υπόθεσης μέσω μοντέλων παλινδρόμησης απέβη αποδοτική, αλλά και αυτό το σημείο χρήζει περισσότερης διερεύνησης αφού φαίνεται να υπάρχει περιθώριο βελτίωσης των αποτελεσμάτων.

Οι πηγές που χρησιμοποιήθηκαν σε αυτό το κεφάλαιο αφορούν κυρίως τα [2], [6], [7], [13] στην βιβλιογραφία.

Πηγές-Βιβλιογραφία

[1] Basu, A., Shioya, H., Park, C. (2011), *Statistical Inference: The Minimum Distance Approach*, Chapman and Hall/CRC

[2] Chen, H. S., Lai, K. and Ying, Z. (2004), *Goodness-of-fit tests and minimum power divergence estimators for survival data,* Statistica Sinica Vol. 14 (pp. 231-248)

[3] Cheng, S. C., Wei, L. J. and Ying Z. (1995), *Analysis of transformation models with censored data,* Biometrika vol. 82 pp. 835–845

[4] Gao, F., Manatunga, A. K., Chen, S. (2006), *Non-parametric estimation for baseline hazards function and covariate effects with time-dependent covariates*, Statistics in Medicine John Wiley & Sons Vol. 37 Issue 6, DOI: 10.1002/sim.2574

[5] Huber-Carol, C. and Vonta, F. (2004), *Frailty models for arbitrarily censored and truncated Data*, Lifetime Data Anal., 10, 369–388

[6] Mattheou, K. and Karagrigoriou, A. (2009), *On new developments in divergence statistics*, Journal of Mathematical Sciences 163, 227-237

[7] Mattheou, K., Lee, S. and Karagrigoriou, A. (2009), *A model selection criterion based on the BHHJ measure of divergence*, Journal of Statistical Planning and Inference 139, 228–235

[8] Pardo, L. (2006), *Statistical Inference Based on Divergence Measures,* Chapman and Hall/CRC

[9] Slud, E. V. and Vonta, F. (2004), *Consistency of the NPML Estimator in the Right-Censored Transformation Model*, Scand. J. Statist., 31, 21-41

[10] Vonta, F. (2005), *Efficient estimation in regression frailty or transformation models based on an algorithm*, Australian & New Zealand Journal of Statistics Vol. 47, Issue 4, December 2005, Pages 503–514, DOI: 10.1111/j.1467-842X.2005.00412.x

[11] Vonta, F. and Karagrigoriou, A. (2007), *Variable selection strategies in survival models with multiple imputations*, Lifetime Data Analysis, 13, Issue 3, 295-315

[12] Vonta, F. and Karagrigoriou, A. (2010), *Generalized Measures of Divergence in Survival Analysis and Reliability*, Journal of Applied Probability, Vol. 47, No. 1 (pp. 216-234) published by Applied Probability Trust

[13] Vonta, I. and Karagrigoriou, A. (2014), *Goodness-of-fit tests via ϕ-measures of divergence for censored data*, Journal of Statistical Computation and Simulation, DOI:10.1080/00949655.2012.733396

[14] Wienke, A. (2011), *Frailty Models in Survival Analysis*, Chapman and Hall/CRC Biostatistics Series

[15] Zeng, D. and Lin, D. Y. (2007), *Semiparametric Transformation Models with Random Effects for Recurrent Events*, Journal of the American Statistical Association, Vol. 102, No. 477 (pp. 167-180) published by American Statistical Association

Ιστοσελίδες, pdf

[16] https://www4.stat.ncsu.edu/~dzhang2/st745/chap6.pdf

[17] http://data.princeton.edu/pop509/NonParametricSurvival.pdf

[18] https://www.uni-salzburg.at/fileadmin/oracle_file_imports/246178.PDF

[19] http://www.math.ucsd.edu/~rxu/math284/slect5.pdf

[20] http://www.public.iastate.edu/~kkoehler/stat565/coxph.4page.pdf

[21] https://sundoc.bibliothek.uni-halle.de/habil-online/07/07H056/t3.pdf

[22] http://www2.hawaii.edu/~taylor/z632/Rbestsubsets.pdf