ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ
ΕΠΙΣΤΗΜΩΝ

Μεταπτυχιακή Εργασία

ΣΤΑΤΙΣΤΙΚΗ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ ΓΙΑ ΜΟΝΤΕΛΑ ΕΥΠΑΘΕΙΑΣ

ΓΕΩΡΓΑΛΗ ΑΡΓΥΡΩ ΣΟΦΙΑ

Τριμελής Επιτροπή: Βόντα Φιλία, Καθηγήτρια, Επιβλέπουσα
Καρώνη Χρυσής, Ομότιμη Καθηγήτρια
Παπανικολάου Βασίλειος, Ομότιμος Καθηγητής

Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών «Εφαρμοσμένες Μαθηματικές Επιστήμες με Ειδίκευση στη Στατιστική»

25 Ιουνίου 2025

Περίληψη

Η ανάλυση των δεδομένων διάρκειας ζωής παίζει σημαντικό ρόλο στην ιατρική, την επιδημιολογία, τη βιολογία, τη δημογραφία, την οικονομία, την μηχανική, την αναλογιστική επιστήμη και άλλους τομείς. Η Ανάλυση Επιβίωσης είναι μια περιοχή της Στατιστικής επιστήμης η οποία ασχολείται με την ανάλυση και την ερμηνεία δεδομένων σχετικά με τη διάρκεια ζωής. Η στατιστική μεθοδολογία που χρησιμοποιείται σε αυτή την περιοχή είναι ζωτικής σημασίας σε όλους αυτούς τους τομείς, όπου η κατανόηση του χρόνου μέχρι να συμβεί ένα γεγονός ενδιαφέροντος είναι απαραίτητη. Για την ανάλυση τέτοιου είδους δεδομένων έχουν αναπτυχθεί και εφαρμοστεί πολλές στατιστικές μέθοδοι.

Το κύριο πρόβλημα στην Ανάλυση Επιβίωσης είναι η λογοκρισία, μία έννοια η οποία συνήθως σημαίνει ότι σε μία μελέτη ο ερευνητής έχει πληροφορίες μόνο για το ότι το συμβάν ενδιαφέροντος δεν συνέβη πριν από μία συγκεκριμένη χρονική στιγμή. Αυτό υπονοεί πως μια λογοκριμένη παρατήρηση περιέχει μόνο μέρος της πληροφορίας για μια τυχαία μεταβλητή που μας ενδιαφέρει. Λόγω αυτών των ελλιπών παρατηρήσεων αναπτύχθηκαν διάφορες ειδικές στατιστικές μέθοδοι.

Ο εκτιμητής Kaplan–Meier (1958) της συνάρτησης επιβίωσης είναι ένα σημαντικό βήμα στην μη παραμετρική εκτίμηση της συνάρτησης επιβίωσης διάφορων μοντέλων που υποθέτονται για τέτοιου είδους δεδομένα. Οι περισσότερες εκτιμήσεις γίνονται υπό όρους με βάση το τι είναι γνωστό τη στιγμή της ανάλυσης, αλλά αυτό αλλάζει με την πάροδο του χρόνου, το οποίο δείχνει ότι πολλές τυπικές στατιστικές προσεγγίσεις δεν μπορούν να εφαρμοστούν στην Ανάλυση Επιβίωσης. Τα μοντέλα που βασίζονται στη συνάρτηση κινδύνου κυριαρχούν στην Ανάλυση Επιβίωσης από τότε που προτάθηκε το μοντέλο αναλογικών κινδύνων από τον Cox (1972). Ο λόγος που αυτό το μοντέλο είναι τόσο δημοφιλές βασίζεται στο πόσο εύκολα διαχειρίζεται δύσκολες έννοιες όπως η λογοκρισία (censoring) και η περικοπή (truncation).

Αυτή η διπλωματική εργασία εστιάζει σε μοντέλα ευπάθειας, μια πολύ συγκεκριμένη περιοχή της Ανάλυσης Επιβίωσης. Η έννοια της ευπάθειας εισάγει την έννοια της πιθανής ετερογένειας στα δεδομένα και στις συσχετίσεις που υπάρχουν μέσα σε μοντέλα για δεδομένα επιβίωσης. Ένα μοντέλο ευπάθειας είναι ένα πολλαπλασιαστικό μοντέλο κινδύνου που αποτελείται από τρεις συνιστώσες: την ευπάθεια (τυχαία επίδραση), τη βασική συνάρτηση κινδύνου (παραμετρική ή μη παραμετρική), και έναν όρο που μοντελοποιεί την επίδραση των παρατηρούμενων συμμεταβλητών (σταθερές επιρροές).

Η εργασία αυτή επικεντρώνεται στη μελέτη παραμετρικών μοντέλων

ευπάθειας στην Ανάλυση Επιβίωσης, με στόχο την αποτελεσματική διαχείριση της μη παρατηρήσιμης ετερογένειας μεταξύ των ατόμων μιας μελέτης. Εξετάζονται αναλυτικά τα Γάμμα και Inverse Gaussian μοντέλα ευπάθειας, στη μονομεταβλητή περίπτωση και ταυτόχρονα στην περίπτωση λογοκριμένων δεδομένων από δεξιά, με παραμετροποίηση της τυχαίας επίδρασης μέσω της παραμέτρου ευπάθειας. Κατασκευάζεται ένας στατιστικός έλεγχος καλής προσαρμογής για τέτοιου είδους μοντέλα, βασισμένος σε ϕ-μέτρα απόκλισης μεταξύ της εμπειρικής και της θεωρητικής κατανομής. Ο βασικός μηχανισμός του ελέγχου στηρίζεται στην ελεγχοσυνάρτηση T_n^{ϕ} (Leandro Pardo (2006)), της οποίας μελετάται η ασυμπτωτική κατανομή κάτω από τη μηδενική υπόθεση την οποία την υποθέτουμε σύνθετη. Βάσει θεωρημάτων που δίνονται στο βιβλίο Leandro Pardo (2006) και στις εργασίες Chen, Lai and Ying (2004) και Vonta and Karagrigoriou (2014), η ασυμπτωτική κατανομή της ελεγχοσυνάρτησης είναι η χ^2 με βαθμούς ελευθερίας $M-k$ (ίσους με τον αριθμό των διαστημάτων στη διαμέριση των δεδομένων μείον τον αριθμό των παραμέτρων που εκτιμώνται) όπως συνηθίζεται να γίνεται στους ελέγχους καλής προσαρμογής. Βρέθηκε όμως σε αυτή την εργασία ότι η ασυμπτωτική κατανομή της ελεγχοσυνάρτησης προσεγγίζεται καλύτερα μέσω ενός γραμμικού συνδυασμού από δύο ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν χ^2_{M-k} και χ^2_1 κατανομές (Morales et al. (1995)). Αυτός ο γραμμικός συνδυασμός χρησιμοποιήθηκε για τον προσδιορισμό των κρίσιμων τιμών του ελέγχου για επίπεδο σημαντικότητας 5% και για $k=1$.

Η εγκυρότητα και η απόδοση του ελέγχου αξιολογούνται μέσω εκτεταμένων προσομοιώσεων, σε σενάρια με δεξιά λογοκρισία, για διάφορα μεγέθη δείγματος n, διαφορετικές τιμές της διασποράς της παραμέτρου ευπάθειας, και μεταβαλλόμενα επίπεδα λογοκρισίας. Σε κάθε πείραμα μελετάται η συμπεριφορά του σφάλματος τύπου Ι, η εκτίμηση της μονοδιάστατης παραμέτρου θ που υπεισέρχεται στο μοντέλο μέσω της βασικής συνάρτησης κινδύνου κάτω από την μηδενική υπόθεση, καθώς και η συμφωνία μεταξύ του 95ου εμπειρικού και θεωρητικού ποσοστημορίου της κατανομής του T_n^{ϕ}.

Abstract

Survival data analysis plays a significant role in medicine, epidemiology, biology, demography, economics, engineering, actuarial science, and other fields. Survival analysis is a branch of statistical science concerned with the analysis and interpretation of time-to-event data. The statistical methodology used in this area is crucial in all these disciplines, where understanding the time until an event of interest occurs is essential. Numerous statistical methods have been developed and applied for analyzing such data.

The main issue in survival analysis is censoring, which typically means that in a study, the researcher knows only that the event of interest has not occurred before a specific time. This implies that a censored observation contains only partial information about the random variable of interest. Various specialized statistical methods have been developed to handle these incomplete observations.

The Kaplan–Meier (1958) estimator of the survival function is an important step in the non-parametric estimation of survival functions for such data. Most estimations are conditional on what is known at the time of analysis, which changes over time. This shows that many typical statistical approaches cannot be directly applied in survival analysis. Hazard-based models have dominated survival analysis since the introduction of the proportional hazards model by Cox (1972). The popularity of this model lies in its ability to conveniently handle complex concepts like censoring and truncation.

This thesis focuses on frailty models, a specific topic within survival analysis. The concept of frailty introduces the idea of potential heterogeneity in the data and correlations that exist within survival models. A frailty model is a multiplicative hazard model consisting of three components: the frailty term (random effect), the baseline hazard function (parametric or non-parametric), and a term modeling the effect of observed covariates (fixed effects).

This study examines parametric frailty models in survival analysis, aiming to efficiently manage unobservable heterogeneity among study subjects. The Gamma and Inverse Gaussian frailty models are analyzed in detail, in the univariate case and under right-censored data, with the random effect parameterized through the frailty term. A statistical goodness-of-fit test is constructed for such models, based on ϕ-divergence measures between empirical and theoretical distributions. The core mechanism of the test relies on the test statistic T_n^{ϕ} (Statistical Inference Based on Divergence Measures, Leandro Pardo, 2006),

whose asymptotic distribution under the null hypothesis (assumed to be composite) is studied. Based on the theorems presented in Pardo's book and the works of Chen, Lai and Ying (2004), and Vonta and Karagrigoriou (2014), the asymptotic distribution of the test statistic is shown to follow a chi-squared χ^2 distribution with $M-k$ degrees of freedom (equal to the number of intervals in the data partition minus the number of estimated parameters), as is common in goodness-of-fit testing. However, this thesis finds that the asymptotic distribution is better approximated by a linear combination of two independent chi-squared random variables with χ^2_{M-k} and χ^2_1 distributions (Morales et al. (1995)). This linear combination is used to determine critical values for the test at a level of significance 5% and for $k = 1$.

The validity and performance of the test are evaluated through extensive simulations, under scenarios with right censoring, various sample sizes n, different values of frailty variance, and varying censoring levels. In each experiment, the behavior of the Type I error, the estimation of the unidimensional parameter θ involved in the baseline hazard under the null hypothesis, and the agreement between the 95th empirical and theoretical quantiles of the T_n^{ϕ} distribution are studied.

ΕΥΧΑΡΙΣΤΙΕΣ

Με την παρούσα διπλωματική εργασία ολοκληρώνονται οι σπουδές μου στο διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών «Εφαρμοσμένες Μαθηματικές Επιστήμες με Ειδίκευση στη Στατιστική» του Εθνικού Μετσόβιου Πολυτεχνείου. Στις σπουδές μου ήταν καθοριστική η συμβολή των καθηγητών μου στα γνωστικά αντικείμενα που παρακολούθησα, στους οποίους οφείλω να εκφράσω τις ειλικρινείς μου ευχαριστίες για τη συμβολή τους στην ολοκλήρωση των σπουδών μου.

Ιδιαίτερα επιθυμώ να ευχαριστήσω την Καθηγήτρια μου και επιβλέπουσα στην παρούσα διπλωματική εργασία, κυρία Βόντα Φιλία, για την επιστημονική και συμβουλευτική καθοδήγηση που μου προσέφερε σε όλα τα στάδια εκπόνησης της εργασίας με τις εύστοχες και πολύ εποικοδομητικές παρατηρήσεις της.

Οφείλω επίσης να εκφράσω τις ευχαριστίες μου προς την επιβλέπουσα επιτροπή, που αποτελούνταν από τον Ομότιμο Καθηγητή Παπανικολάου Βασίλειο και την Ομότιμη Καθηγήτρια Καρώνη Χρυσηίδα, για το χρόνο που αφιέρωσαν.

Περιεχόμενα

1 Εισαγωγή

Σε αυτή την ενότητα θα γίνει μία επισκόπηση των βασικών αρχών που διέπουν το πεδίο της Ανάλυσης Επιβίωσης. Είναι σημαντικό να επισημανθεί ότι η ανάλυση των δεδομένων επιβίωσης έχει απαιτήσει την ανάπτυξη ενός ξεχωριστού στατιστικού πλαισίου, που θα αποκλίνει σημαντικά από τις μεθοδολογίες που χρησιμοποιούνται σε άλλους στατιστικούς τομείς.

Ο πυρήνας της είναι η ανάλυση δεδομένων διάρκειας ζωής (survival data) τα οποία αντικατοπτρίζουν τη χρονική διάρκεια έως ότου συμβεί ένα γεγονός. Το "συμβάν" αναφέρεται συνήθως σε ένα ή περισσότερα περιστατικά ενδιαφέροντος, όπως είναι ο θάνατος σε βιολογικούς οργανισμούς, η ίαση μίας ασθένειας ή η αστοχία σε μηχανικά συστήματα. Ο πρωταρχικός στόχος της ανάλυσης επιβίωσης δεν είναι απλώς να προβλέψει εάν ένα γεγονός θα συμβεί, αλλά κυρίως, πότε θα συμβεί αυτό το γεγονός.

1.1 Εισαγωγικές Έννοιες

Βασικό χαρακτηριστικό της Ανάλυσης Επιβίωσης αποτελεί η αντιμετώπιση λογοκριμένων (censored) ή περικομμένων (truncated) παρατηρήσεων, όπου για ορισμένες από αυτές, το συμβάν ενδιαφέροντος (όπως ο θάνατος, η αποτυχία ή η υποτροπή) δεν έχει συμβεί μέχρι το τέλος της περιόδου της μελέτης και είναι διαθέσιμες μόνο μερικές πληροφορίες. Αυτές οι πληροφορίες απαιτούν την ανάπτυξη εξειδικευμένων στατιστικών τεχνικών για την ακριβή εκτίμηση των χρόνων επιβίωσης και τη σύγκριση των ποσοστών επιβίωσης.

Σε αυτό το σημείο θα αναφερθούν κάποιες βασικές έννοιες, όπως η Συνάρτηση Επιβίωσης που εκτιμά την πιθανότητα ο χρόνος μέχρι το γεγονός ενδιαφέροντος να είναι μεγαλύτερος από κάποιο καθορισμένο χρόνο. Η Συνάρτηση Επιβίωσης (Survival function) είναι ένα βασικό εργαλείο στην Ανάλυση Επιβίωσης, που συχνά συμβολίζεται με $S(t)$ και δείχνει την πιθανότητα επιβίωσης πέραν του χρόνου t. Στη συνέχεια έχουμε τη Συνάρτηση Κινδύνου (hazard function), η οποία αντιπροσωπεύει τη στιγμιαία πιθανότητα εμφάνισης του γεγονότος που μας ενδιαφέρει τη στιγμή t, δεδομένου ότι το συμβάν δεν έχει

ακόμη συμβεί μέχρι το t. Έχουμε επίσης τη Λογοκρισία, μια θεμελιώδη πτυχή των δεδομένων επιβίωσης, όπου η ακριβής ώρα του συμβάντος είναι άγνωστη για ορισμένες παρατηρήσεις. Τεχνικές όπως η εκτιμήτρια Kaplan-Meier (Kaplan, E. L., & Meier, P. (1958)) και το μοντέλο αναλογικών κινδύνων Cox (Cox, D. R. (1972)) έχουν αναπτυχθεί για τον χειρισμό λογοκριμένων δεδομένων.

Η Ανάλυση Επιβίωσης παρέχει εργαλεία για να απαντηθούν ερωτήματα όπως η διάμεσος ή ο μέσος χρόνος επιβίωσης, η πιθανότητα επιβίωσης μετά από έναν ορισμένο χρόνο και πώς οι συμμεταβλητές επηρεάζουν τον κίνδυνο εμφάνισης του γεγονότος. Μέσω τεχνικών όπως η Kaplan-Meier και το μοντέλο Cox, οι ερευνητές μπορούν να εκτιμήσουν τις συναρτήσεις επιβίωσης, να συγκρίνουν χρόνους επιβίωσης μεταξύ των ομάδων και να αξιολογήσουν το αντίκτυπο των επεξηγηματικών μεταβλητών στην επιβίωση, ακόμη και παρουσία λογοκριμένων δεδομένων.

1.1.1 Αθροιστική Συνάρτηση Κατανομής, Συνάρτηση Επιβίωσης και Συνάρτηση Κινδύνου

Θεωρούμε μια τυχαία μη αρνητική μεταβλητή $\mathbf{T}^*$, η οποία αντιπροσωπεύει το χρόνο από ένα καλά καθορισμένο σημείο έναρξης μέχρι την εμφάνιση ενός γεγονότος. Εάν το γεγονός αυτό είναι ο θάνατος, το $\mathbf{T}^*$ είναι ο χρόνος επιβίωσης. Αυτή η μεταβλητή συνήθως είναι συνεχής, αλλά λόγω της δυσκολίας πολλές φορές να μετρηθεί σε όλη τη χρονική διάρκεια του πειράματος μπορεί να εμφανισθεί και σε διακριτή μορφή. Όλες οι συναρτήσεις κατανομής του χρόνου ενός συμβάντος ορίζονται στο διάστημα $[0, \infty)$.

Η συνάρτηση πυκνότητας πιθανότητας (p.d.f.) συμβολίζεται με $\mathbf{f}$. Η κατανομή μιας τυχαίας μεταβλητής καθορίζεται πλήρως και μοναδικά από την συνάρτηση πυκνότητας πιθανότητας. Υπάρχουν και άλλες χρήσιμες συναρτήσεις που μπορούν να ληφθούν από τη συνάρτηση πυκνότητας πιθανότητας, όπως αυτή της Αθροιστικής Συνάρτησης Κατανομής της $\mathbf{T}^*$,

$$F(t) = P(T^* \leq t) = \int_0^t f(s)\, ds, \tag{1}$$

όπου $\mathbf{P}(\mathbf{A})$ είναι η πιθανότητα να συμβεί το γεγονός $\mathbf{A}$.

Στην ανάλυση επιβίωσης, ενδιαφέρον αποτελεί η πιθανότητα ενός ατόμου να επιβιώσει πέραν του χρόνου $\mathbf{t}$, η οποία δίνεται από τη συνάρτηση επιβίωσης

$$S(t) = 1 - F(t) = P(T^* > t) = \int_t^\infty f(s)\, ds. \tag{2}$$

Κύρια έννοια όμως της ανάλυσης επιβίωσης αποτελεί η συνάρτηση κινδύνου και ορίζεται από τη σχέση,

$$\mu(t) = \lim_{\varepsilon \to 0} \frac{P(t < T^* \leq t + \varepsilon | T^* > t)}{\varepsilon} = \frac{f(t)}{1 - F(t)} \tag{3}$$

Η συνάρτηση αυτή χαρακτηρίζει τον κίνδυνο θανάτου που μεταβάλλεται με την πάροδο του χρόνου. Καθορίζει τη στιγμιαία πιθανότητα αποτυχίας τη χρονική στιγμή $\mathbf{t}$, δεδομένου ότι το άτομο επιβιώνει μέχρι τη χρονική στιγμή $\mathbf{t}$. Ονομάζεται επίσης και ρυθμός εξόδου, καθώς δείχνει την έξοδο της παρατήρησης από την κατάσταση του ενδιαφέροντος.

Μερικές φορές, είναι χρήσιμο να ασχοληθούμε με τη σωρευτική συνάρτηση κινδύνου,

$$M(t) = \int_0^t \mu(s)\, ds. \tag{4}$$

Σημαντικό είναι να αναφερθεί η έννοια του μετασχηματισμού **Laplace L** μιας τυχαίας μεταβλητής:

$$\mathbf{L}(u) = \mathbb{E}e^{-uT^*} = \int_0^\infty e^{-ut} f(t)\, dt. \tag{5}$$

Όλες οι παραπάνω συναρτήσεις προσδιορίζουν με ισοδύναμο τρόπο την κατανομή της τυχαίας και μη αρνητικής μεταβλητής $\mathbf{T}^*$.

Είναι εύκολο να προκύψουν οι σχέσεις μεταξύ των παραπάνω εννοιών για παράδειγμα από την (1) συνεπάγεται ότι,

$$S(t) = 1 - F(t) = e^{-\int_0^t \mu(s)\, ds} = e^{-M(t)}. \tag{6}$$

Η εξίσωση αυτή είναι η κύρια εκθετική συνάρτηση της Ανάλυσης Επιβίωσης, καθώς παρουσιάζει τη συνάρτηση κατανομής και τη συνάρτηση επιβίωσης μέσω της συνάρτησης κινδύνου ή σωρευτικού κινδύνου. Αποδεικνύεται έτσι ότι η συνάρτηση κινδύνου είναι πιο εύχρηστη σε σύγκριση με την συνάρτηση πυκνότητας πιθανότητας ή τη συνάρτηση κατανομής και επιβίωσης, λόγο της ουσιαστικής πιθανολογικής ερμηνείας της και της απλότητας της σε πιθανοτικές εκφράσεις.

1.2 Μη Λογοκριμένα, Λογοκριμένα ή Περικομμένα Δεδομένα Επιβίωσης

Στην Ανάλυση Επιβίωσης, επειδή πρωταρχικός στόχος της είναι να εξεταστεί ο χρόνος μέχρις ότου συμβεί ένα γεγονός, τα δεδομένα μπορούν γενικά να ταξινομηθούν σε δύο βασικές κατηγορίες, τα μη λογοκριμένα δεδομένα (Observed-Uncensored) και τα περικομμένα (truncated) ή λογοκριμένα (censored) δεδομένα ή και λογοκριμένα και περικομμένα ταυτόχρονα.

1.2.1 Μη Λογοκριμένα Δεδομένα Επιβίωσης

Τα μη λογοκριμένα δεδομένα αναφέρονται σε καταστάσεις όπου παρατηρείται πλήρως το πότε συμβαίνει ένα γεγονός ενδιαφέροντος (π.χ. θάνατος, αποτυχία, ανατροπή). Ο ακριβής χρόνος μέχρι να συμβεί το γεγονός αυτό είναι γνωστός. Αυτός ο τύπος δεδομένων παρέχει πλήρεις πληροφορίες σχετικά με τους χρόνους επιβίωσης ατόμων ή αντικειμένων υπό μελέτη. Για παράδειγμα, εάν μια μελέτη διερευνά το χρόνο μέχρι να αποτύχει ένα συγκεκριμένο στοιχείο, κάθε στοιχείο που αποτυγχάνει συνεισφέρει δεδομένα στην ανάλυση χωρίς λογοκρισία. Τα μη λογοκριμένα δεδομένα είναι ζωτικής σημασίας για την Ανάλυση Επιβίωσης, καθώς παρέχουν άμεσες, ποσοτικοποιήσιμες πληροφορίες για το χρόνο των γεγονότων.

Στα δεδομένα με παρατήρηση χωρίς λογοκρισία, κάθε περίπτωση έχει σαφώς καθορισμένο «χρόνο συμβάντος», ο οποίος είναι ο ακριβής χρόνος από την έναρξη της παρατήρησης μέχρι την εμφάνιση του συμβάντος. Αυτός ο τύπος δεδομένων είναι ιδανικός για την ανάλυση επιβίωσης, καθώς παρέχει τις πιο λεπτομερείς και αδιαμφισβήτητες πληροφορίες σχετικά με το χρόνο μέχρι το συμβάν, επιτρέποντας αξιόπιστες εκτιμήσεις παραμέτρων, ποσοτήτων και πιθανοτήτων επιβίωσης.

Η ανάλυση των παρατηρημένων-μη λογοκριμένων δεδομένων μπορεί να χρησιμοποιήσει διάφορες στατιστικές μεθόδους όπως αυτή της εμπειρικής εκτιμήτριας της συνάρτησης επιβίωσης από παρατηρηθέντα-μη λογοκριμένα δεδομένα. Μπορεί επίσης να γίνει χρήση του μοντέλου αναλογικών κινδύνων Cox, το οποίο αποτελεί ένα μοντέλο παλινδρόμησης που χρησιμοποιείται για την ταυτόχρονη εξέταση της επίδρασης πολλών μεταβλητών στο χρόνο επιβίωσης, ιδανικό για το χειρισμό παρατηρημένων-μη λογοκριμένων δεδομένων. Παραμετρικά μοντέλα επιβίωσης, επίσης μπορούν να χρησιμοποιηθούν, όπως το εκθετικό, το Weibull, το λογαριθμοκανονικό και άλλα, τα οποία υποθέτουν μια συγκεκριμένη κατανομή για τους χρόνους επιβίωσης και μπορούν να προσαρμοστούν

σε παρατηρούμενα-μη λογοκριμένα δεδομένα για να εκτιμηθούν οι πιθανότητες επιβίωσης και οι διάμεσοι χρόνοι επιβίωσης.

Παράδειγμα 1.1. Έστω μια κλινική δοκιμή που εξετάζει την αποτελεσματικότητα ενός νέου φαρμάκου που πρόκειται να αυξήσει τον χρόνο επιβίωσης των ασθενών με έναν συγκεκριμένο τύπο καρκίνου. Οι ασθενείς εγγράφονται στη μελέτη κατά τη διάγνωση και παρακολουθούνται για περίοδο 5 ετών.

- **Γεγονός ενδιαφέροντος:** Θάνατος λόγω καρκίνου.
- **Παρατήρηση:** Καταγράφεται ο χρόνος κάθε ασθενούς από την εγγραφή έως το θάνατο.
- **Συλλογή δεδομένων:** Έστω ότι υπάρχουν 100 ασθενείς στη δοκιμή. Κατά τη διάρκεια της περιόδου των 5 ετών, 80 από αυτούς τους ασθενείς πεθαίνουν από καρκίνο και οι ακριβείς χρόνοι θανάτου τους καταγράφονται με ακρίβεια. Αυτοί οι 80 ασθενείς αντιπροσωπεύουν τα παρατηρούμενα-μη λογοκριμένα δεδομένα, καθώς οι χρόνοι εκδήλωσής τους είναι γνωστοί και δεν υπάρχει λογοκρισία.
- **Ανάλυση:** Χρησιμοποιώντας την εκτιμήτρια Kaplan-Meier, η οποία στην περίπτωση των μη λογοκριμένων δεδομένων ταυτίζεται με την εμπειρική εκτιμήτρια της συνάρτησης επιβίωσης, θα μπορούσε κανείς να υπολογίσει την πιθανότητα επιβίωσης σε διάφορα χρονικά σημεία (π.χ. 1 έτος, 3 έτη, 5 έτη) με βάση τους παρατηρηθέντες χρόνους θανάτου. Εάν κάποιος επιθυμεί να διερευνήσει την επίδραση παραγόντων όπως η ηλικία, το φύλο ή προηγούμενες καταστάσεις υγείας, θα μπορούσε να εφαρμόσει ένα μοντέλο Cox χρησιμοποιώντας τους πλήρεις, μη λογοκριμένους χρόνους συμβάντων αυτών των 80 ασθενών.

Σε αυτό το παράδειγμα, τα παρατηρηθέντα-μη λογοκριμένα δεδομένα παρέχουν μια σαφή και πλήρη εικόνα των χρόνων συμβάντων, γεγονός που είναι πλεονεκτικό για την ανάλυση. Ωστόσο, οποιαδήποτε ανάλυση στον πραγματικό κόσμο θα πρέπει επίσης να λάβει υπόψη τυχόν λογοκριμένα δεδομένα από τους υπόλοιπους 20 ασθενείς που δεν εμφάνισαν το συμβάν εντός της περιόδου μελέτης, ώστε να αποφευχθεί η μεροληψία στην εκτίμηση της συνάρτησης επιβίωσης. □

1.2.2 Λογοκριμένα ή Περικομμένα Δεδομένα Επιβίωσης

Τα λογοκριμένα και περικομμένα δεδομένα είναι έννοιες που εμφανίζονται συχνά στην ανάλυση επιβίωσης, όπου η καθεμία αντιπροσωπεύει διαφορετικούς τύπους ελλιπών πληροφοριών σχετικά με τη μεταβλητή ενδιαφέροντος, συνήθως χρόνο μέχρι να συμβεί ένα συμβάν.

Η λογοκρισία συμβαίνει όταν οι πληροφορίες σχετικά με την εκδήλωση ενός γεγονότος στη μελέτη είναι μόνο εν μέρει γνωστές. Υπάρχουν διάφοροι τύποι λογοκρισίας, όπως η δεξιά λογοκρισία (Right-censoring), η οποία αποτελεί τον πιο συνηθισμένο τύπο λογοκρισίας, όπου το γεγονός ενδιαφέροντος δεν έχει συμβεί μέχρι το τέλος της μελέτης ή το άτομο εγκαταλείπει τη μελέτη νωρίς. Γνωρίζουμε μόνο ότι ο χρόνος του συμβάντος είναι μεγαλύτερος από μια ορισμένη τιμή. Υπάρχει επίσης η αριστερή λογοκρισία (Left-censoring), όπου το γεγονός έχει ήδη συμβεί πριν από την έναρξη της μελέτης, επομένως γνωρίζουμε ότι ο χρόνος του συμβάντος είναι μικρότερος από μια ορισμένη τιμή. Τέλος έχουμε τη λογοκρισία διαστήματος (Interval-censoring), κατά την οποία γνωρίζουμε ότι το γεγονός συνέβη μέσα σε ένα συγκεκριμένο χρονικό διάστημα, αλλά δεν γνωρίζουμε τον ακριβή χρόνο.

Σε αυτή την εργασία εξετάζουμε κυρίως τον τύπο της δεξιάς λογοκρισίας, επειδή αυτός ο τύπος λογοκρισίας είναι ο πιο συνηθισμένος σε πραγματικά δεδομένα και εφαρμογές της ανάλυσης επιβίωσης.

Έστω $T_1^*, T_2^*, \ldots, T_n^*$, n ανεξάρτητοι και ισόνομα κατανεμημένοι χρόνοι επιβίωσης με αθροιστική συνάρτηση κατανομής F και έστω $C_1, C_2, ..., C_n$ είναι επίσης ανεξάρτητοι χρόνοι λογοκρισίας με αθροιστική συνάρτηση κατανομής G. Γενικά υποθέτουμε ότι οι F και G είναι απολύτως συνεχείς. Επιπλέον, έστω f και g, συναρτήσεις πυκνότητας πιθανότητας σε σχέση με τις F και G. Είμαστε σε θέση να παρατηρήσουμε μόνο τα δεδομένα $(T_1, \Delta_1), (T_2, \Delta_2), \ldots, (T_n, \Delta_n)$, όπου $T_i = \min\{T_i^*, C_i\}$ συμβολίζει το χρόνο της παρατήρησης και

$$\Delta_i = \begin{cases} 1 & : \text{εάν } T_i^* \leq C_i, \ , T_i \text{ δεν είναι λογοκριμένο} \\ 0 & : \text{εάν } T_i^* > C_i, \ , T_i \text{ είναι λογοκριμένο} \end{cases}$$

Χρησιμοποιούμε το T^* και C, χωρίς δείκτες, ως συντομογραφία για όλες τις T_i^* και C_i^* μεταβλητές διάρκειας ζωής και λογοκρισίας, αντίστοιχα, και συμβολίζουμε με H τη συνάρτηση κατανομής του χρόνου παρατήρησης $T = \min\{T^*, C\}$. Τότε μπορεί εύκολα να προκύψει η ακόλουθη σχέση:

$$
\begin{aligned}
H(t) &= P(\min\{T^*, C\} \leq t) \\
&= 1 - P(\min\{T^*, C\} > t) \\
&= 1 - P(T^* > t, C > t).
\end{aligned} \tag{7}
$$

Υποθέτοντας ανεξαρτησία μεταξύ του χρόνου γεγονότος T^* και του χρόνου λογοκρισίας C συνεπάγεται η απλοποίηση,

$$
\begin{aligned}
H(t) &= 1 - P(T^* > t)P(C > t) \\
&= 1 - (1 - P(T^* \leq t))(1 - P(C \leq t)) \\
&= 1 - (1 - F(t))(1 - G(t)).
\end{aligned} \tag{8}
$$

Αυτό υπογραμμίζει τη σημασία της παραδοχής της ανεξαρτησίας όσον αφορά τα γεγονότα και τους χρόνους λογοκρισίας που συνήθως γίνονται στην ανάλυση επιβίωσης

Πρέπει να προσέξουμε πως η αθροιστική συνάρτηση κατανομής των μη λογοκριμένων δεδομένων (αν επικεντρωθούμε μόνο στις λογοκριμένες παρατηρήσεις) δεν είναι F.

$$
\begin{aligned}
P(T \leq t, \Delta = 1) &= P(T^* \leq t, T^* \leq C) \\
&= \int_{t^* \leq t} \int_{t^* \leq c} f(t^*) g(c)\, dt^*\, dc \\
&= \int_{t^* \leq t} f(t^*) \left(\int_{t^* \leq c} g(c)\, dc \right) dt^* \\
&= \int_{t^* \leq t} f(t^*)\, (1 - G(t^*))\, dt^* \\
&= F(t)
\end{aligned} \tag{9}
$$

Συνεπώς, η αγνόηση των λογοκριμένων χρόνων των γεγονότων συνεπάγεται μεροληπτικές εκτιμήσεις.

Θεώρημα 1. Η συνάρτηση πυκνότητας πιθανότητας των δεδομένων επιβίωσης (T, Δ) είναι,

$$
f(t, \delta) = \left(f(t)(1 - G(t))\right)^{\delta} \left(g(t)(1 - F(t))\right)^{1-\delta} \tag{10}
$$

Παράδειγμα 1.2. Δεξιά Λογοκριμένα Δεδομένα

Ας θεωρήσουμε ένα σενάριο όπου παράγουμε ένα σύνολο 10 παρατηρήσεων που δοκιμάζονται για αποτυχία κατά τη διάρκεια μιας χρονικής περιόδου, και ορισμένες μονάδες αποτυγχάνουν κατά τη διάρκεια της περιόδου παρατήρησης, ενώ

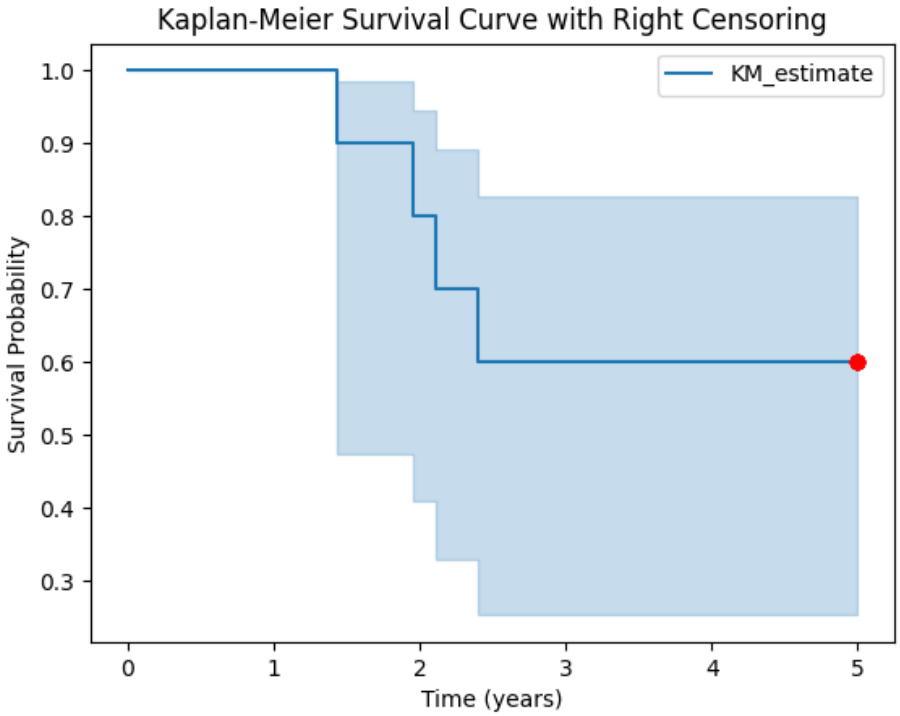
άλλες εξακολουθούν να λειτουργούν μέχρι το τέλος της μελέτης και συνεπώς είναι δεξιά λογοκριμένες. Μπορούμε να χρησιμοποιήσουμε το πακέτο lifelines της Python, το οποίο έχει σχεδιαστεί για ανάλυση επιβίωσης, ή αντίστοιχα τη βιβλιοθήκη survival στην R. Παρακάτω παρατίθενται τα αποτελέσματα που παράγονται από τα δεδομένα επιβίωσης με δεξιά λογοκρισία και προσαρμόζονται με μια καμπύλη Kaplan-Meier.

Δημιουργούμε έναν δείκτη λογοκρισίας όπου περίπου το 50% των παρατηρήσεων λογοκρίνονται. Επίσης χρησιμοποιούμε την εκτιμήτρια Kaplan-Meier για να προσαρμόσουμε το μοντέλο επιβίωσης και να σχεδιάσουμε την καμπύλη επιβίωσης. Οι λογοκριμένες παρατηρήσεις των δεδομένων φαίνονται στο γράφημα με κόκκινους κύκλους.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Σχήμα 1.

```
import numpy as np
import pandas as pd
from lifelines import KaplanMeierFitter
import matplotlib.pyplot as plt

# Set the seed for reproducibility
np.random.seed(123)

# Simulate some data
n = 10  # number of units
study_duration = 5  # total study time in years
failure_times = np.random.uniform(0, study_duration, n)
   # random failure times
censoring_indicator = np.random.uniform(0, 1, n) > 0.5  #
    random censoring, with about a 50% chance
observed_times = np.where(censoring_indicator,
   failure_times, study_duration)
status = np.where(censoring_indicator, 1, 0)  # 1 for
   event occurred, 0 for censored

data = pd.DataFrame({
    'id': range(1, n+1),
    'time': observed_times,
    'status': status
})
```

```
# Initialize the KaplanMeierFitter object
kmf = KaplanMeierFitter()

# Fit the model to the data
kmf.fit(durations=data['time'], event_observed=data['
    status'])

# Plot the survival curve
kmf.plot_survival_function()
plt.title('Kaplan-Meier Survival Curve with Right
    Censoring')
plt.xlabel('Time (years)')
plt.ylabel('Survival Probability')

# Annotate censored points on the plot
for _, row in data[data['status'] == 0].iterrows():
    plt.plot(row['time'], kmf.survival_function_.loc[row[
        'time']], 'ro')

plt.show()
print(data)
```



Σχήμα 1: Καμπύλη Kaplan-Meier για δεξιά λογοκριμένα δεδομένα.

Η καμπύλη επιβίωσης Kaplan-Meier δείχνει τις πιθανότητες επιβίωσης των ατόμων/στοιχείων με την πάροδο του χρόνου, δεδομένης της παρουσίας δεξιάς λογοκρισίας στα δεδομένα. Η βηματική συνάρτηση μειώνεται στους χρόνους των συμβάντων, δείχνοντας μια πτώση της πιθανότητας επιβίωσης κάθε φορά που συμβαίνει ένα γεγονός. Υπάρχουν τέσσερα τέτοια γεγονότα, που συμβαίνουν περίπου στα 5.0, 1.43, 5.0, 5.0, 5.0, 2.11, 5.0, 5.0, 2.40 και 1.96 έτη.

Οι κόκκινες κουκκίδες στο διάγραμμα, που είναι οι λογοκριμένες παρατηρήσεις, παρουσιάζονται στο ίδιο σημείο στο γράφημα φυσικά, δείχνουν ότι αυτά τα άτομα/στοιχεία δεν βίωσαν το γεγονός κατά τη διάρκεια της περιόδου πα-

ρατήρησης. Σύμφωνα με το σύνολο δεδομένων, τα 6 από τα 10 στοιχεία ήταν λογοκριμένα (τα 1, 3, 4, 5, 7 και 8). Αυτά τα στοιχεία λογοκρίνονται δεξιά τη χρονική στιγμή 5, που σημαίνει ότι εξακολουθούσαν να λειτουργούν στο τέλος της μελέτης, οπότε οι πραγματικοί χρόνοι επιβίωσής τους είναι μεγαλύτεροι από 5. Τα συμβάντα (αποτυχίες) αντιπροσωπεύονται από τις πτώσεις στην καμπύλη επιβίωσης. Τα στοιχεία 2, 6, 9 και 10 βίωσαν το συμβάν στις αντίστοιχες χρονικές στιγμές τους και οι χρόνοι επιβίωσής τους είναι ακριβώς γνωστοί.

Αρχικά, η πιθανότητα επιβίωσης είναι 1 (ή 100%), καθώς δεν έχει συμβεί κανένα γεγονός. Το πρώτο γεγονός οδηγεί σε μείωση της πιθανότητας επιβίωσης, η οποία συνεχίζεται με κάθε επόμενο συμβάν. Τα επίπεδα τμήματα της καμπύλης αντιστοιχούν σε διαστήματα όπου δεν παρατηρούνται γεγονότα. Το πλάτος του διαστήματος εμπιστοσύνης (σκιασμένη περιοχή) αυξάνεται με την πάροδο του χρόνου, υποδεικνύοντας αυξανόμενη αβεβαιότητα σχετικά με την εκτίμηση της πιθανότητας επιβίωσης, ιδίως καθώς μειώνεται ο αριθμός των ατόμων που διατρέχουν κίνδυνο. Η περίοδος μελέτης είναι 5 έτη, η οποία είναι και ο μέγιστος χρόνος παρατήρησης για τις λογοκριμένες μονάδες.

□

Στα περικομμένα δεδομένα, τα άτομα αποκλείονται εντελώς από τη μελέτη εάν ο χρόνος εκδήλωσης τους δεν εμπίπτει σε ένα συγκεκριμένο εύρος. Υπάρχουν δύο κύριοι τύποι περικοπής, η αριστερή περικοπή (Left-truncation), όπου τα άτομα περιλαμβάνονται στη μελέτη μόνο εάν ο χρόνος συμβάντος τους υπερβαίνει μια ορισμένη τιμή. Για παράδειγμα, σε μια μελέτη όπου ενδιαφέρον έχει η διάρκεια εργασίας, όσοι εργάζονταν πριν από την ημερομηνία έναρξης της μελέτης μπορούν να συμπεριληφθούν μόνο εάν συνεχίσουν να εργάζονται και μετά την έναρξη της μελέτης. Έχουμε και τη δεξιά περικοπή (Right-truncation), όπου τα άτομα περιλαμβάνονται στη μελέτη μόνο εάν ο χρόνος εκδήλωσης τους είναι κάτω από μια συγκεκριμένη τιμή και είναι λιγότερο συνηθισμένη από την αριστερή περικοπή.

Η βασική διαφορά μεταξύ λογοκριμένων και περικομμένων δεδομένων είναι ότι με λογοκριμένα δεδομένα, όλα τα άτομα αποτελούν μέρος της μελέτης, αλλά ορισμένα έχουν ελλιπείς πληροφορίες, ενώ με περικομμένα δεδομένα, ορισμένα άτομα δεν περιλαμβάνονται καθόλου στη μελέτη με βάση τους χρόνους συμβάντων τους. Ο σωστός χειρισμός αυτών των τύπων ημιτελών δεδομένων είναι

απαραίτητος για την ακριβή ανάλυση επιβίωσης, καθώς επηρεάζουν σημαντικά την εκτίμηση των συναρτήσεων επιβίωσης και την ανάλυση του χρόνου μέχρι τα δεδομένα του συμβάντος.

1.3 Πιθανοφάνεια

Σε αυτό το σημείο θα αναλυθεί η συνάρτηση πιθανοφάνειας, η οποία προκύπτει με βάση τα αποτελέσματα του Θεωρήματος 1 που δόθηκε προηγουμένως. Εδώ εξετάζεται μια παραμετρική κατάσταση, η οποία σημαίνει ότι η κατανομή του χρόνου επιβίωσης θεωρείται γνωστή μέχρι ένα πεπερασμένης διαστάσεως άγνωστο διάνυσμα παραμέτρων θ. Ένα παραμετρικό μοντέλο για την επιβίωση είναι η εκθετική κατανομή. Είναι πολύ σπάνιο τα δεδομένα επιβίωσης να ακολουθούν αυτή την κατανομή επειδή δεν είναι πολύ ευέλικτη. Αλλά για την κατανόηση όλων των άλλων κατανομών επιβίωσης, το εκθετικό μοντέλο έχει μεγάλη σημασία. Άλλες παραμετρικές προσεγγίσεις, όπως το μοντέλο Weibull (που χρησιμοποιείται συχνά στην αξιοπιστία) και το μοντέλο Gompertz (που προτιμάται στην αναλογιστική και στις δημογραφικές ρυθμίσεις) περιέχουν την εκθετική κατανομή ως ειδική περίπτωση. Στις βιοστατιστικές εφαρμογές η εκθετική κατανομή γίνεται όλο και πιο δημοφιλής. Οι κατανομές μελετώνται στην πιο απλή περίπτωση ανεξάρτητων και ισόνομα κατανεμημένων τυχαίων μεταβλητών.

Εδώ εκτός από την ανεξαρτησία μεταξύ των γεγονότων και των χρόνων λογοκρισίας, υποθέτουμε μη πληροφοριακή λογοκρισία (non-informative censoring). Αυτό σημαίνει ότι η κατανομή λογοκρισίας δεν πρέπει να εξαρτάται από το διάνυσμα παραμέτρων θ της κατανομής επιβίωσης. Διαφορετικά, η κατανομή λογοκρισίας θα περιείχε πληροφορίες σχετικά με τις παραμέτρους που μας ενδιαφέρουν. Επιπλέον, δεν στοχεύουμε στην εκτίμηση των παραμέτρων της κατανομής λογοκρισίας. Ως εκ τούτου, οι όροι $g(t)$ και $G(t)$ στη συνάρτηση πυκνότητας γίνονται προσθετικές σταθερές, ανεξάρτητες από το θ στη συνάρτηση λογαριθμικής πιθανοφάνειας. Αυτές οι προσθετικές σταθερές δεν παίζουν ρόλο στην παράγωγο της συνάρτησης λογαριθμικής πιθανοφάνειας ως προς θ και επομένως μπορούν να απαλειφθούν από την πιθανοφάνεια των δεδομένων. Ως συνέπεια αυτού και του τύπου (10), η συμβολή των από δεξιά λογοκριμένων δεδομένων επιβίωσης (t_i, δ_i), όπου $i = 1, ..., n$ στην συνάρτηση πιθανότητας απλοποιείται σε,

$$L_i(\theta) = f(t_i; \theta)^{\delta_i} S(t_i; \theta)^{1-\delta_i} \tag{11}$$

Βλέπουμε πως όπως είναι σύνηθες η πιθανότητα είναι η πυκνότητά τους, και οι λογοκριμένες παρατηρήσεις παρέχουν την πληροφορία ότι ο άγνωστος χρόνος επιβίωσης υπερβαίνει τον παρατηρούμενο λογοκριμένο χρόνο. Κατά συνέπεια, ο δεύτερος όρος στην εξίσωση (11) αντικατοπτρίζει την πιθανότητα επιβίωσης τουλάχιστον μέχρι το t_i. Ο τύπος (11) μπορεί εύκολα να εκφραστεί ως συνάρτηση της συνάρτησης κινδύνου,

$$L_i(\theta) = \mu(t_i;\theta)^{\delta_i} e^{-\int_0^{t_i} \mu(s;\theta)\, ds}. \tag{12}$$

Θεωρούμε ένα δείγμα ανεξάρτητων δισδιάστατων τ.μ. $(t_1, \delta_1), \ldots, (t_n, \delta_n)$. Η συνάρτηση πιθανοφάνειας των δεδομένων ορίζεται ως,

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta) = \prod_{i=1}^{n} \left(\mu(t_i;\theta)^{\delta_i} e^{-\int_0^{t_i} \mu(s;\theta)\, ds} \right) \tag{13}$$

λόγω της ανεξαρτησίας μεταξύ των τυχαίων διανυσμάτων (μονομεταβλητό μοντέλο).

Για να υπογραμμιστεί η σημασία της κοινής παραδοχής σχετικά με τις ανεξαρτησία της λογοκρισίας στην Ανάλυση Επιβίωσης, παρατίθεται το ακόλουθο παράδειγμα, το οποίο αφορά την περίπτωση της εξαρτημένης λογοκρισίας για να μελετήσουμε τις διαφορές. Αποδεικνύεται ότι η συνάρτηση πιθανοφάνειας στην περίπτωση εξαρτημένης λογοκρισίας είναι το γινόμενο των παραγώγων της από κοινού συνάρτησης επιβίωσης των χρόνων ζωής και των χρόνων λογοκρισίας.

Παράδειγμα 1.3. Θεωρούμε ότι έχουμε δεδομένα επιβίωσης (T, Δ), όπου $T = \min(T^*, C)$ και $\Delta_i = 1$ εάν $T^* \leq C$ (παρατηρούμενο γεγονός). Θέτουμε επίσης, $S(t^*, c)$ και $f(t^*, c)$ να είναι η κοινή συνάρτηση επιβίωσης και η πυκνότητα πιθανότητας της T^* και C, αντίστοιχα. Κατά συνέπεια, οι συναρτήσεις κατανομής που απαιτούνται για την κατασκευή της συνάρτησης πιθανότητας μπορούν να προκύψουν ως εξής,

- **Συναρτήσεις Κατανομής**

– Για μη λογοκριμένες παρατηρήσεις ($\Delta = 1$):

$$\begin{aligned} H_1(t) &= \mathbf{P}(T \leq t, \Delta = 1) \\ &= \mathbf{P}(T^* \leq t, T^* \leq C) \\ &= \iint_{\{t^* \leq t, t^* \leq c\}} f(t^*, c)\, dc\, dt^* \\ &= -\int_0^t S_1(t^*, t^*)\, dt^* \end{aligned}$$

με $S_1(t,c) = \frac{\partial S(t,c)}{\partial t}$. Αυτό σημαίνει ότι η πυκνότητα μιας μη λογοκριμένης ($\delta = 1$) παρατήρησης είναι μια παράγωγος της συνάρτησης κατανομής,

$$h_1(t) = \frac{dH_1(t)}{dt} = -S_1(t,t).$$

– Για τις λογοκριμένες παρατηρήσεις ($\Delta = 0$):

$$\begin{aligned} H_0(t) &= \mathbf{P}(T \leq t, \Delta = 0) \\ &= \mathbf{P}(C \leq t, C < T^*) \\ &= \iint_{\{c \leq t, c < t^*\}} f(t^*, c)\, dt^*\, dc \\ &= -\int_0^t S_2(c, c)\, dc \end{aligned}$$

με $S_2(t^*, c) = \frac{\partial S(t^*,c)}{\partial c}$ και $h_0(t) = \frac{dH_0(t)}{dt} = -S_2(t,t)$.

- **Συνάρτηση Πιθανοφάνειας:**

– Κατά συνέπεια, η πιθανότητα στην περίπτωση ενός παραμετρικού μοντέλου με $\boldsymbol{\theta}$, ως το διάνυσμα των άγνωστων παραμέτρων, είναι μια σύνθεση των συναρτήσεων πυκνότητας, $S_1(t,t;\boldsymbol{\theta}) = S_1(t,t)$ και $S_2(t,t;\boldsymbol{\theta}) = S_2(t,t)$. Για ένα δείγμα $(t_1,\delta_1),...,(t_n,\delta_n)$, η συνάρτηση πιθανοφάνειας είναι :

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(-S_1(t_i,t_i;\boldsymbol{\theta})\right)^{\delta_i} \left(-S_2(t_i,t_i;\boldsymbol{\theta})\right)^{1-\delta_i}.$$

Στην περίπτωση ανεξαρτησίας της λογοκρισίας με $S(t, c; \boldsymbol{\theta}) = (1 - F(t; \boldsymbol{\theta})(1 - G(c)))$, απλοποιείται σε,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}) = \prod_{i=1}^{n} \mu(t_i; \boldsymbol{\theta})^{\delta_i} \exp\left(- \int_0^{t_i} \mu(s; \boldsymbol{\theta})\, ds \right)$$

.

□

Μέχρι τώρα, είχε εξεταστεί μόνο η περίπτωση των δεδομένων με δεξιά λογοκρισία. Ωστόσο, σε ορισμένες περιπτώσεις, οι χρόνοι των γεγονότων είναι γνωστοί μόνο όταν βρίσκονται σε ένα συγκεκριμένο διάστημα. Η κατάσταση αυτή προκύπτει ιδίως όταν τα υποκείμενα της μελέτης δεν βρίσκονται υπό συνεχή παρακολούθηση, όπως, για παράδειγμα, οι ασθενείς που επισκέπτονται το γιατρό τους σε προκαθορισμένες ώρες (ή ώρες που τους βολεύουν), όπου η εμφάνιση του συμβάντος μπορεί να διαγνωστεί γνωρίζοντας ότι το συμβάν είχε ή δεν είχε συμβεί κατά τη στιγμή της προηγούμενης επίσκεψης. Αυτό το είδος λογοκρισίας ονομάζεται λογοκρισία διαστήματος και εξετάστηκε λεπτομερώς από την Sun (2006). Γενικά, η δεξιά λογοκρισία είναι μια ειδική περίπτωση λογοκρισίας διαστήματος και ορισμένες από τις μεθόδους για την δεξιά λογοκρισία μπορούν να εφαρμοστούν άμεσα, ή με μικρές αλλαγές, σε λογοκρισία διαστήματος. Ωστόσο, οι περισσότερες από τις προσεγγίσεις για δεδομένα με δεξιά λογοκρισία δεν είναι κατάλληλες για δεδομένα με λογοκρισία διαστήματος, επειδή ο μηχανισμός λογοκρισίας πίσω από αυτή είναι πολύ πιο περίπλοκος από ότι στην περίπτωση της δεξιάς λογοκρισίας.

Μια σημαντική ειδική περίπτωση δεδομένων με λογοκρισία διαστήματος είναι τα λεγόμενα δεδομένα τρέχουσας κατάστασης (current status data). Ο όρος δεδομένα τρέχουσας κατάστασης προέρχεται από εφαρμογές στον τομέα της δημογραφίας (Diamond et al. 1986). Σημαίνει ότι η πληροφορία για την επιβίωση κάθε ατόμου περιλαμβάνει είτε το μηδέν είτε το άπειρο. Τέτοιου είδους δεδομένα εμφανίζονται συνήθως όταν κάθε υποκείμενο της μελέτης παρατηρείται μόνο μία φορά και η μόνη διαθέσιμη πληροφορία για το υπό μελέτη γεγονός είναι το αν το γεγονός έχει συνέβη πριν από τη λήψη της παρατήρησης. Κατά συνέπεια, τα δεδομένα τρέχουσας κατάστασης δίνονται με τη μορφή (T, Δ),

όπου T δηλώνει το χρόνο παρακολούθησης (ο οποίος δεν είναι ο χρόνος κατά τον οποίο συμβαίνει το συμβάν) και Δ είναι ο δείκτης που δείχνει κατά πόσον το γεγονός έχει ήδη συμβεί πριν από την παρακολούθηση ή όχι. Στην παραμετρική περίπτωση, η συνάρτηση πιθανότητας ενός δείγματος $(t_1, \delta_1), ..., (t_n, \delta_n)$ με άγνωστο διάνυσμα παραμέτρων θ προς εκτίμηση μπορεί να γραφεί στη μορφή (Sun 2006),

$$L(\theta) = \prod_{i=1}^{n} \left((1 - S(t_i; \theta))^{\delta_i} S(t_i; \theta)^{1-\delta_i} \right) \tag{14}$$

με στοιχεία ανάλογα με το αν το συμβάν έχει ήδη συμβεί πριν από τους χρόνους παρακολούθησης ($\delta_i = 1$) ή όχι ($\delta_i = 0$).

1.4 Παραμετρικά Μοντέλα

Στη συνέχεια θα εξετασθούν ορισμένες κατανομές πιθανοτήτων που είναι χρήσιμες στον τομέα της ανάλυσης επιβίωσης. Φυσικά, οποιαδήποτε κατανομή μη αρνητικών τυχαίων μεταβλητών μπορεί να χρησιμοποιηθεί για να περιγράψει τη διάρκεια ζωής. Στην παραμετρική περίπτωση, στα μοντέλα που υποθέτονται, ο βασικός κίνδυνος προσδιορίζεται από έναν μικρό αριθμό μονοδιάστατων παραμέτρων ή ένα πεπερασμένο διάνυσμα παραμέτρων. Οι κατανομές επιβίωσης είναι συνεχείς με μία εξαίρεση - την σε σημεία σταθερή συνάρτηση κινδύνου. Σε όλη τη βιβλιογραφία για την ανάλυση επιβίωσης, ορισμένα παραμετρικά μοντέλα έχουν χρησιμοποιηθεί επανειλημμένα, όπως το εκθετικό, το Weibull και το Gompertz. Αυτές οι κατανομές έχουν συγκεκριμένες εκφράσεις για την επιβίωση, την πυκνότητα και τις συναρτήσεις κινδύνου. Οι κατανομές Γάμμα και η Λογαριθμοκανονική είναι υπολογιστικά λιγότερο βολικές, αλλά εφαρμόζονται συχνά. Για την αποφυγή ζητημάτων εγκυρότητας του μοντέλου, η μη παραμετρική προσέγγιση, βασισμένη στον εκτιμήτρια Kaplan-Meier, είναι συνήθως η προτιμώμενη πορεία. Ωστόσο, αυτή η εναλλακτική λύση είναι συχνά αναποτελεσματική, όπως σημειώνει ο Miller (1983). Ειδικότερα, τα τυπικά σφάλματα των εκτιμήσεων των παραμέτρων σε παραμετρικά μοντέλα τείνουν να είναι μικρότερα από ότι σε μη παραμετρικά μοντέλα. Ωστόσο, η επάρκεια της επιλεγμένης κατανομής πρέπει να ελέγχεται.

Στην παρούσα και στις επόμενες ενότητες η τυχαία μεταβλητή T υποδηλώνει το χρόνο μέχρι την εμφάνιση ενός γεγονότος, για το οποίο μας ενδιαφέρει να βγάλουμε συμπεράσματα. Μόνο η απλούστερη περίπτωση ανεξάρτητων και πανομοιότυπα κατανεμημένων χρόνων γεγονότων $T_1, T_2, ..., T_n$ εξετάζεται εδώ.

Γενικά, θα χρησιμοποιούμε κεφαλαία γράμματα για τις τυχαίες μεταβλητές και πεζά γράμματα για τις υλοποιήσεις τους.

Θα ξεκινήσουμε με την εκθετική κατανομή, καθώς είναι θεμελιώδης στην ανάλυση επιβίωσης, ακόμη και αν είναι σχετικά σπάνιο το γεγονός τα δεδομένα του χρόνου συμβάντος να ακολουθούν αυτή τη μονοπαραμετρική κατανομή. Ωστόσο, η εκθετική κατανομή βοηθά στην κατανόηση των βασικών ζητημάτων πολλών άλλων τυπικών κατανομών επιβίωσης.

1.4.1 Εκθετική Κατανομή

Το εκθετικό μοντέλο $T \sim Exp(\lambda)$ είναι το απλούστερο παραμετρικό μοντέλο διάρκειας ζωής. Έχει μόνο μία παράμετρο λ. Το μοντέλο υποθέτει σταθερό κίνδυνο με την πάροδο του χρόνου, κάτι το οποίο αντανακλά την ιδιότητα της κατανομής κατάλληλα, η οποία ιδιότητα ονομάζεται έλλειψη μνήμης. Η πιθανότητα αποτυχίας εντός ενός συγκεκριμένου χρονικού διαστήματος εξαρτάται μόνο από το μήκος αλλά όχι από τη θέση αυτού του διαστήματος. Αυτό σημαίνει ότι η κατανομή $T - t$ υπό την προϋπόθεση ότι $T > t$ είναι η ίδια με με την αρχική κατανομή. Με άλλα λόγια, ισχύει ότι

$$P(t < T \leq t + \epsilon \mid T > t) = P(T \leq \epsilon)$$

για κάθε θετικό ε.

Η παραπάνω σχέση αποδεικνύεται εύκολα ως εξής:

$$\begin{aligned} P(T > t + t_0 \mid T > t) &= \frac{P(T > t + t_0, T > t)}{P(T > t)} \\ &= \frac{P(T > t + t_0)}{P(T > t)} \\ &= \frac{S(t + t_0)}{S(t)} \\ &= \frac{e^{-\lambda(t+t_0)}}{e^{-\lambda t}} \\ &= e^{-\lambda t_0} \\ &= P(T > t_0). \end{aligned}$$

Παράδειγμα 1.4. Ας υποθέσουμε ότι η τυχαία μεταβλητή **T** ακολουθεί κατανομή με συνάρτηση πυκνότητας πιθανότητας,

$$f(t) = \lambda e^{-\lambda t}, \quad t \geq 0, \tag{15}$$

όπου λ όπου είναι η παράμετρος ρυθμού της κατανομής.

Οι σχετικές συναρτήσεις είναι:

$$\begin{aligned} \text{Συνάρτηση Επιβίωσης} \quad & S(t) = e^{-\lambda t} \\ \text{Συνάρτηση Κινδύνου} \quad & \mu(t) = \lambda \\ \text{Αθροιστική συνάρτηση κινδύνου} \quad & M(t) = \lambda t. \end{aligned}$$

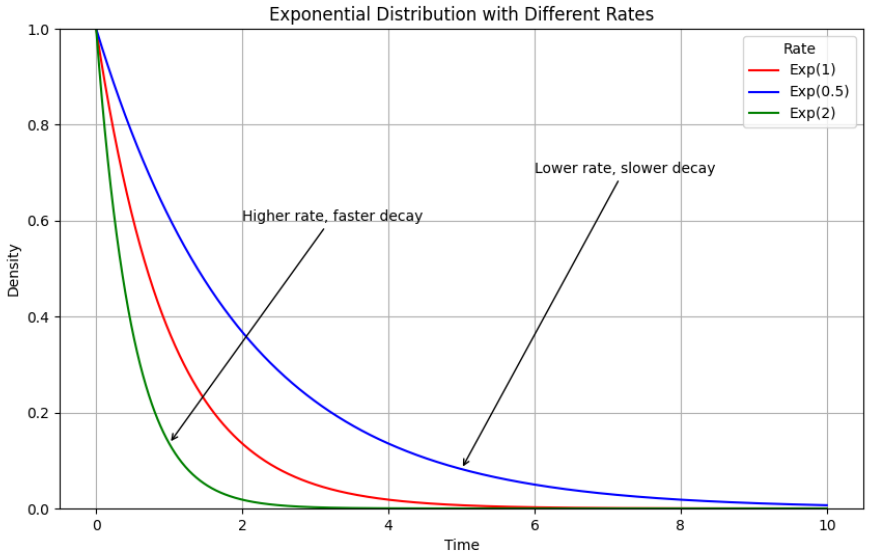
□

Η εκθετική κατανομή χρησιμοποιείται για τη μοντελοποίηση του χρόνου μέχρι να συμβεί ένα γεγονός, όπως η αστοχία ενός ηλεκτρονικού εξαρτήματος, και έχει σταθερή συνάρτηση κινδύνου, πράγμα που σημαίνει ότι το γεγονός είναι εξίσου πιθανό να συμβεί σε οποιαδήποτε χρονική στιγμή.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Σχήμα 2.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Generating the x values
x = np.linspace(0, 10, 1000)

# Calculating the exponential distributions for different
    rates
y1 = np.exp(-1 * x)  # rate = 1
y2 = np.exp(-0.5 * x)  # rate = 0.5
y3 = np.exp(-2 * x)  # rate = 2
```

```
# Creating data frames
df1 = pd.DataFrame({'x': x, 'Density': y1, 'Rate': 'Exp
    (1)'})
df2 = pd.DataFrame({'x': x, 'Density': y2, 'Rate': 'Exp
    (0.5)'})
df3 = pd.DataFrame({'x': x, 'Density': y3, 'Rate': 'Exp
    (2)'})

# Combining the data frames
df = pd.concat([df1, df2, df3])

# Plotting
plt.figure(figsize=(10, 6))
sns.lineplot(data=df, x='x', y='Density', hue='Rate',
    palette=['red', 'blue', 'green'])

# Annotations to explain the effect of different rates
plt.annotate('Higher rate, faster decay', xy=(1, np.exp
    (-2)), xytext=(2, 0.6),
             arrowprops=dict(facecolor='black',
                arrowstyle='->'))
plt.annotate('Lower rate, slower decay', xy=(5, np.exp
    (-0.5 * 5)), xytext=(6, 0.7),
             arrowprops=dict(facecolor='black',
                arrowstyle='->'))

plt.title('Exponential Distribution with Different Rates'
    )
plt.xlabel('Time')
plt.ylabel('Density')
plt.ylim(0, 1)
plt.legend(title='Rate')
plt.grid(True)
plt.savefig("Exponential_density_rates.png")
plt.show()
```



Σχήμα 2: Εκθετική Κατανομή με διαφορετικούς ρυθμούς πυκνότητας.

Στο γράφημα 2 παραπάνω ο χρόνος t αντιστοιχίζεται στον άξονα x και οι τιμές που παίρνει η συνάρτηση πυκνότητας πιθανότητας στον άξονα y. Παρατηρούμε πως, έχουμε υψηλότερο ρυθμό σε ένα τμήμα της πράσινης γραμμής, το οποίο δείχνει πού η πυκνότητα έχει μειωθεί σημαντικά μέχρι το $x = 1$. Χαμηλότερο ρυθμό δείχνει ένα τμήμα της μπλε γραμμής, υποδεικνύοντας μια πιο αργή μείωση της πυκνότητας γύρω στο $x = 5$.

Το διάγραμμα απεικονίζει οπτικά πώς ο ρυθμός μιας εκθετικής κατανομής επηρεάζει τον ρυθμό αποτυχίας. Οι ταχύτεροι ρυθμοί οδηγούν σε πιο απότομη μείωση της πυκνότητας, που παρατηρείται σε μικρότερο εύρος κατά μήκος του άξονα x. Αντίθετα, οι βραδύτεροι ρυθμοί παρουσιάζουν μια πιο σταδιακή μείωση, εξαπλώνοντας την πτώση.

1.4.2 Κατανομή Weibull

Το μοντέλο Weibull εισήχθη από τον Waloddi Weibull (1939) και είναι μία δημοφιλής γενίκευση του εκθετικού μοντέλου με δύο θετικές παραμέτρους. Η δεύτερη παράμετρος ν επιτρέπει μεγάλη ευελιξία στο μοντέλο και διαφορετικά σχήματα της συνάρτησης κινδύνου. Η ευκολία του μοντέλου αυτού οφείλεται, αφενός, σε αυτή την ευελιξία και, αφετέρου, στην απλότητα της συνάρτησης κινδύνου και των συναρτήσεων επιβίωσης.

Παράδειγμα 1.5. Ας υποθέσουμε ότι η τυχαία μεταβλητή $\mathbf{T}$ ακολουθεί κατανομή με συνάρτηση πυκνότητας πιθανότητας,

$$f(t) = \lambda t^{\nu-1} e^{-\lambda t^{\nu}}, \quad t \geq 0, \tag{16}$$

όπου λ, ν είναι μονοδιάστατες μη αρνητικές παράμετροι. Ισχύουν οι ακόλουθες σχέσεις,

$$\begin{gathered}
\text{Συνάρτηση Επιβίωσης} \quad S(t) = e^{-\lambda t^{\nu}} \\
\text{Συνάρτηση Κινδύνου} \quad \mu(t) = \lambda \nu t^{\nu-1} \\
\text{Αθροιστική συνάρτηση κινδύνου} \quad M(t) = \lambda t^{\nu} \\
\text{Μέση Τιμή} \quad ET = \lambda^{-\frac{1}{\nu}} \Gamma\left(1 + \tfrac{1}{\nu}\right) \\
\text{Διασπορά} \quad V(T) = \lambda^{-\frac{2}{\nu}} \left(\Gamma\left(1 + \tfrac{2}{\nu}\right) - \Gamma\left(1 + \tfrac{1}{\nu}\right)^2\right), \\
\text{όπου η Γάμμα συνάρτηση είναι: } \Gamma(k) = \int_0^{\infty} s^{k-1} e^{-s}\, ds \quad (k > 0).
\end{gathered}$$

□

Όπως αναφέρθηκε προηγουμένως, η έννοια της συνάρτησης κινδύνου είναι ιδιαίτερα χρήσιμη στην Ανάλυση Επιβίωσης. Δηλαδή, οι ιδιότητες της κατανομής των χρόνων των συμβάντων γενικά χαρακτηρίζονται με βάση τις ιδιότητες της συνάρτησης κινδύνου, η οποία περιγράφει τον τρόπο με τον οποίο η στιγμιαία πιθανότητα αποτυχίας για ένα άτομο μεταβάλλεται με το χρόνο. Οι εφαρμογές συχνά περιέχουν ποιοτικές πληροφορίες σχετικά με τη συνάρτηση κινδύνου, οι οποίες είναι χρήσιμες για την επιλογή ενός μοντέλου με καλή προσαρμογή.

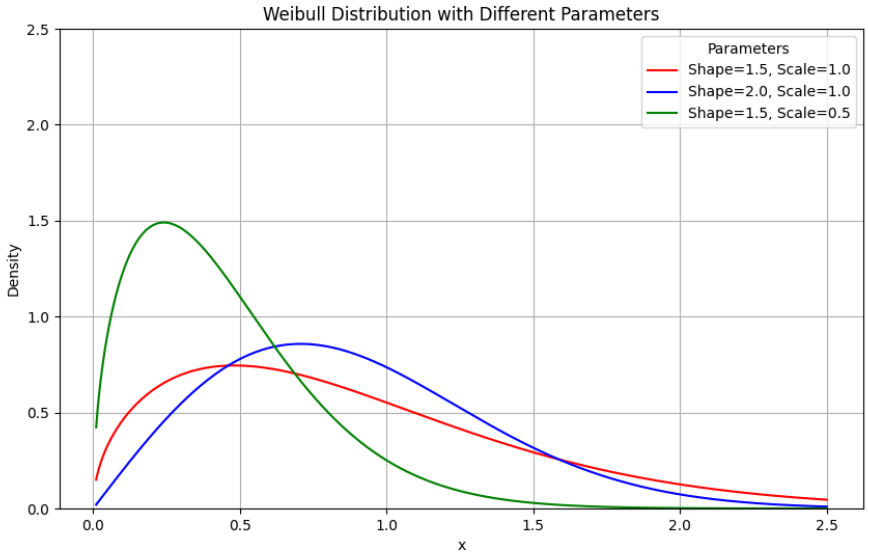
Η συνάρτηση κινδύνου μπορεί να τροποποιηθεί με διάφορους τρόπους. Μπορεί να είναι αυξανόμενη (για παράδειγμα, μια κατανομή Weibull με παράμετρο σχήματος $\nu > 1$), φθίνουσα (κατανομή Weibull με $\nu < 1$), σταθερή (κατανομή Weibull με $\nu = 1$), μπορεί να παίρνει σχηματική μορφή $J-$, $U-$ ή καμπανόμορφη (κατανομή log-normal).

Τα κύρια σημεία που πρέπει να θυμόμαστε εδώ είναι ότι η συνάρτηση κινδύνου αντιπροσωπεύει μια πτυχή της κατανομής της πιθανότητας μιας μη αρνητικής τυχαίας μεταβλητής που έχει άμεση φυσική σημασία, και ότι οι ποιοτικές πληροφορίες σχετικά με την μορφή της συνάρτησης κινδύνου είναι χρήσιμες για την επιλογή μιας ενός κατάλληλου (παραμετρικού) μοντέλου για την υπό εξέταση κατάσταση. Επιπλέον, από μια πιο πρακτική άποψη, τα μοντέλα που βασίζονται στην συνάρτηση κινδύνου μπορούν εύκολα να χειριστούν τη λογοκρισία και την αποκοπή που εμφανίζονται συχνά σε δεδομένα επιβίωσης.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Σχήμα 3.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import weibull_min

# Generating the x values
x = np.linspace(0.01, 2.5, 1000)  # Avoiding zero because
    the Weibull function can't handle it at x = 0

# Parameters for the Weibull distributions
params = [
    (1.5, 1.0),  # shape = 1.5, scale = 1.0
    (2.0, 1.0),  # shape = 2.0, scale = 1.0
    (1.5, 0.5)   # shape = 1.5, scale = 0.5
]

# Calculating the Weibull distributions for different
   parameters
dfs = []
colors = ['red', 'blue', 'green']
labels = ['Shape=1.5, Scale=1.0', 'Shape=2.0, Scale=1.0',
    'Shape=1.5, Scale=0.5']
for (shape, scale), color, label in zip(params, colors,
   labels):
    y = weibull_min.pdf(x, shape, scale=scale)
    df = pd.DataFrame({'x': x, 'Density': y, '
       Distribution': label})
```

```
    dfs.append(df)

# Combining the data frames
df = pd.concat(dfs)

# Plotting
plt.figure(figsize=(10, 6))
sns.lineplot(data=df, x='x', y='Density', hue='
   Distribution', palette=colors)
plt.title('Weibull Distribution with Different Parameters
   ')
plt.xlabel('x')
plt.ylabel('Density')
plt.ylim(0, 2.5)
plt.legend(title='Parameters')
plt.grid(True)
plt.savefig("Weibull_distribution.png")
plt.show()
```



Σχήμα 3: Οι συναρτήσεις πυκνότητας πιθανότητας της Weibull Κατανομής.

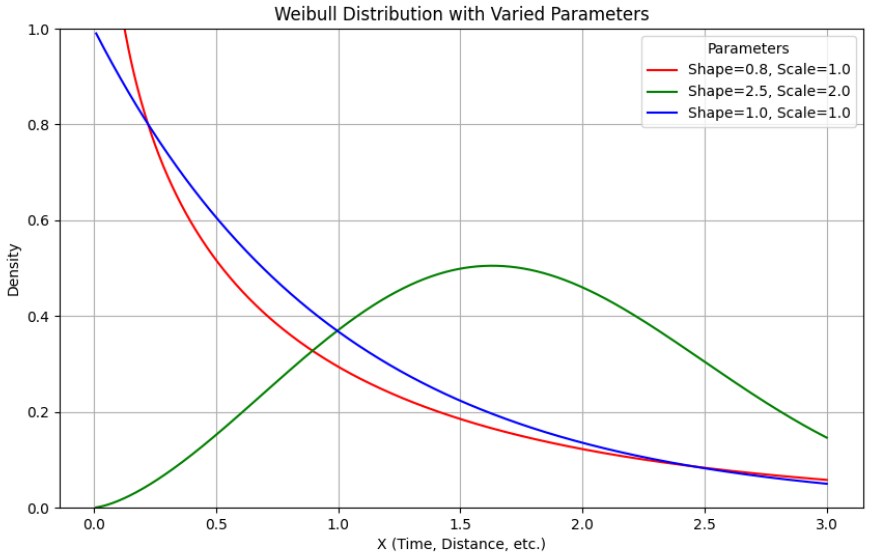
Στο Διάγραμμα 3 απεικονίζονται οι συναρτήσεις πυκνότητας πιθανότητας των κατανομών Weibull (λ=1,ν=1.5), Weibull (λ=1,ν=2) και Weibull (λ=0.5,ν=1.5).

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Σχήμα 4.

```
import numpy as np
import pandas as pd
```

```
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import weibull_min

# Generating the x values
x = np.linspace(0.01, 3, 1000)  # Using a larger range to
    better display differences

# Parameters for the Weibull distributions
params = [
    (0.8, 1.0),  # shape < 1, indicating a decreasing
       failure rate
    (2.5, 2.0),  # shape > 1, indicating an increasing
       failure rate
    (1.0, 1.0)   # shape = 1, which resembles the
       exponential distribution
]

# Calculating the Weibull distributions for different
   parameters
dfs = []
colors = ['red', 'green', 'blue']
labels = ['Shape=0.8, Scale=1.0', 'Shape=2.5, Scale=2.0',
    'Shape=1.0, Scale=1.0']
for (shape, scale), color, label in zip(params, colors,
   labels):
    y = weibull_min.pdf(x, shape, scale=scale)
    df = pd.DataFrame({'x': x, 'Density': y, '
       Distribution': label})
    dfs.append(df)

# Combining the data frames
df = pd.concat(dfs)

# Plotting
plt.figure(figsize=(10, 6))
sns.lineplot(data=df, x='x', y='Density', hue='
   Distribution', palette=colors)
plt.title('Weibull Distribution with Varied Parameters')
plt.xlabel('X (Time, Distance, etc.)')
```

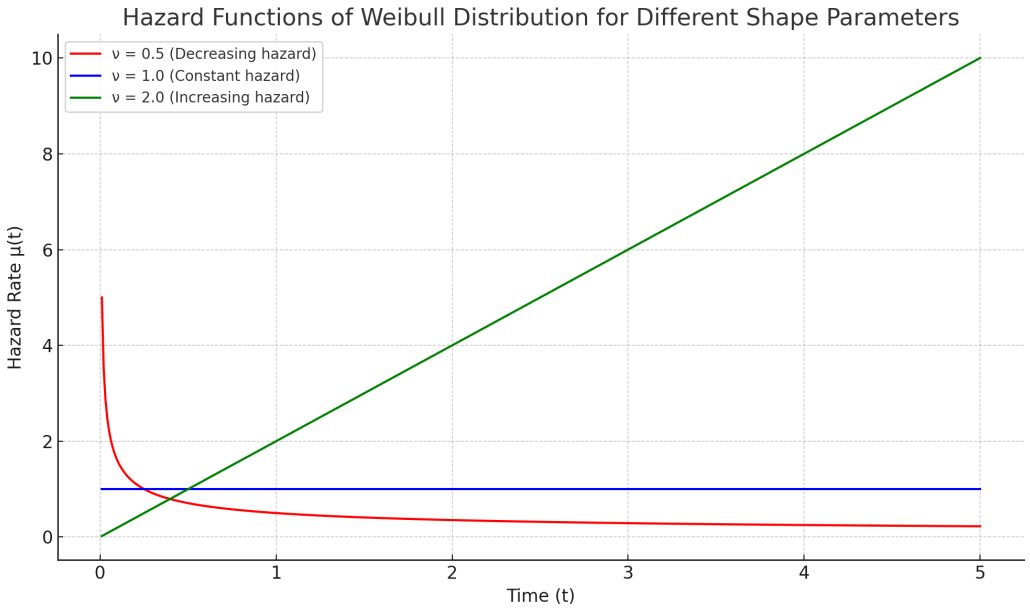
```
plt.ylabel('Density')
plt.ylim(0, 1)
plt.legend(title='Parameters')
plt.grid(True)
plt.savefig("Weibull_varied_parameters.png")
plt.show()
```



Σχήμα 4: Οι συναρτήσεις πυκνότητας πιθανότητας της Weibull Κατανομής με διαφορετική επιλογή παραμέτρων.

Στο Σχήμα 4 απεικονίζονται οι συναρτήσεις πυκνότητας πιθανότητας των

κατανομών Weibull (λ=1,ν=0.8), Weibull (λ=2,ν=2.5) και Weibull (λ=1,ν=1).



Σχήμα 5: Οι συναρτήσεις πυκνότητας πιθανότητας της Weibull Κατανομής με διαφορετική επιλογή των παραμέτρων ν για τη συνάρτηση κινδύνου.

Στο Σχήμα 5 απεικονίζονται οι συναρτήσεις πυκνότητας πιθανότητας των κατανομών Weibull (λ=1,ν=0.5), Weibull (λ=1,ν=1) και Weibull (λ=1,ν=2).

1.4.3 Λογαριθμολογιστική Κατανομή

Μια εναλλακτική πρόταση στην παραπάνω κατανομή Weibull είναι η λογαριθμολογική κατανομή. Είναι αρκετά ευέλικτη με δύο παραμέτρους, που συμβολίζονται με $\log L(\nu, \kappa)$. Είναι ένα από τα παραμετρικά μοντέλα επιβίωσης-χρόνου

στα οποία για $\kappa \leq 1$ ο ρυθμός αποτυχίας μειώνεται με την πάροδο του χρόνου και για $\kappa > 1$ ο ρυθμός αποτυχίας έχει κυρτή συμπεριφορά, δηλαδή αρχικά αυξάνεται και έπειτα μειώνεται με την πάροδο του χρόνου.

Παράδειγμα 1.6. Ας υποθέσουμε ότι η τυχαία μεταβλητή **T** ακολουθεί κατανομή με συνάρτηση πυκνότητας πιθανότητας

$$f(t) = \frac{\nu\kappa(\upsilon t)^{\kappa-1}}{(1+(\upsilon t)^{\kappa})^2} \quad (\nu > 0, \kappa > 0) \tag{17}$$

Ισχύουν οι ακόλουθες σχέσεις,

$$\text{Συνάρτηση Επιβίωσης} \quad S(t) = \frac{1}{1+(\upsilon t)^{\kappa}}$$

$$\text{Συνάρτηση Κινδύνου} \quad \mu(t) = \frac{\nu(\upsilon t)^{\kappa-1}}{1+(\upsilon t)^{\kappa}}$$

$$\text{Αθροιστική συνάρτηση κινδύνου} \quad M(t) = \ln(1+(\upsilon t)^{\kappa})$$

$$\text{Μέση Τιμή} \quad ET = \lambda^{-\frac{1}{\nu}}\Gamma\left(1+\tfrac{1}{\nu}\right)$$

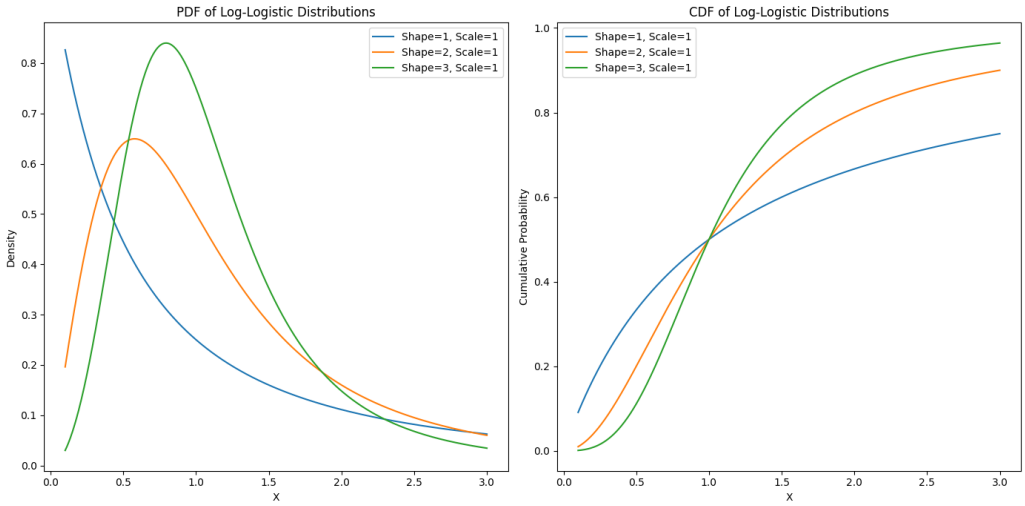
□

Για την απεικόνιση λογαριθμολογιστικών κατανομών θα δημιουργήσουμε ένα παράδειγμα στην Python όπου θα σχεδιάσουμε τη συνάρτηση πυκνότητας πιθανότητας (PDF) και τη συνάρτηση αθροιστικής κατανομής (CDF) της κατανομής log-logistic χρησιμοποιώντας διαφορετικές παραμέτρους. Η λογαριθμολογαριθμική κατανομή έχει το όνομα fisk.

Με τον παρακάτω κώδικα δημιουργούμε το Σχήμα 5.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
from scipy.stats import fisk

# Parameters for the log-logistic distribution
params = [
    (1, 1),   # shape = 1, scale = 1
    (2, 1),   # shape = 2, scale = 1
    (3, 1)    # shape = 3, scale = 1
]

# Generating the x values
x = np.linspace(0.1, 3, 1000)

# Prepare the plot
plt.figure(figsize=(14, 7))

# Plotting PDF and CDF for each set of parameters
for c, scale in params:
    pdf = fisk.pdf(x, c, scale=scale)
    cdf = fisk.cdf(x, c, scale=scale)

    plt.subplot(1, 2, 1)
    plt.plot(x, pdf, label=f'Shape={c}, Scale={scale}')
    plt.title('PDF of Log-Logistic Distributions')
    plt.xlabel('X')
    plt.ylabel('Density')
    plt.legend()

    plt.subplot(1, 2, 2)
    plt.plot(x, cdf, label=f'Shape={c}, Scale={scale}')
    plt.title('CDF of Log-Logistic Distributions')
    plt.xlabel('X')
    plt.ylabel('Cumulative Probability')
    plt.legend()

plt.tight_layout()
plt.show()
```



Σχήμα 6: Συνάρτηση πυκνότητας πιθανότητας (αριστερά) και αθροιστική συνάρτηση κατανομής (δεξιά) της λογαριθμολογιστικής κατανομής.

Χρησιμοποιούμε τρία σύνολα παραμέτρων για να απεικονίσουμε τις διαφορετικές συμπεριφορές της λογαριθμολογιστικής κατανομής. Όλες χρησιμοποιούν παράμετρο κλίμακας 1, αλλά μεταβάλλουν την παράμετρο κ.

Η συνάρτηση πυκνότητας πιθανότητας ορίζει την πιθανότητα εμφάνισης κάθε τιμής της κατανομής. Καθώς αυξάνεται η παράμετρος κ, η κορυφή της κατανομής γίνεται πιο έντονη και μετακινείται ελαφρώς προς τα δεξιά, υποδεικνύοντας υψηλότερο μέγιστο ποσοστό αποτυχίας αλλά σε μεταγενέστερο χρόνο.

Η αθροιστική συνάρτηση κατανομής δείχνει την αθροιστική πιθανότητα μέχρι μια ορισμένη τιμή. Καθώς αυξάνεται ο χρόνος, η καμπύλη γίνεται πιο απότομη στην άνοδό της, υποδεικνύοντας ταχύτερη προσέγγιση στη βεβαιότητα (ή υψηλότερη αξιοπιστία μέχρι ένα ορισμένο σημείο).

Αυτό το παράδειγμα δείχνει πώς η παράμετρος κ επηρεάζει τόσο την πυκνότητα όσο και τη σωρευτική πιθανότητα των αποτελεσμάτων, καθιστώντας τη λογαριθμολογιστική κατανομή ιδιαίτερα χρήσιμη για τη μοντελοποίηση χρόνων επιβίωσης, όπου ο ρυθμός εμφάνισης του γεγονότος μεταβάλλεται κατά τη διάρκεια της ζωής των ατόμων/στοιχείων που μελετώνται.

1.4.4 Κατανομή Gompertz

Στις πιθανότητες και τη στατιστική, η κατανομή Gompertz είναι μια συνεχής κατανομή πιθανότητας, που πήρε το όνομά της από τον Βρετανό Benjamin Gompertz. Η κατανομή αυτή εφαρμόζεται συχνά για την περιγραφή της κατανομής της διάρκειας ζωής των ενηλίκων. Συναφείς επιστημονικοί τομείς όπως η βιολογία και η γεροντολογία χρησιμοποιούν επίσης την κατανομή Gompertz για την ανάλυση της επιβίωσης. Πιο πρόσφατα, οι επιστήμονες πληροφορικής άρχισαν επίσης να μοντελοποιούν τα ποσοστά αποτυχίας του κώδικα υπολογιστών με την κατανομή αυτή.

Παράδειγμα 1.7. Ας υποθέσουμε ότι η τυχαία μεταβλητή $\mathbf{T}$ ακολουθεί την κατανομή Gompertz (λ, ϕ) με συνάρτηση πυκνότητας πιθανότητας,

$$f(t) = \lambda e^{\phi t} e^{-\lambda\left(\frac{e^{\phi t}-1}{\phi}\right)} \tag{18}$$

όπου λ, ϕ, t μη αρνητικές παράμετροι. Ισχύουν οι ακόλουθες σχέσεις,

$$\begin{gathered}\text{Συνάρτηση Επιβίωσης} \quad S(t) = e^{-\lambda\left(\frac{e^{\phi t}-1}{\phi}\right)} \\ \text{Συνάρτηση Κινδύνου} \quad \mu(t) = \lambda e^{\phi t} \\ \text{Αθροιστική συνάρτηση κινδύνου} \quad M(t) = \tfrac{\lambda}{\phi}(e^{\phi t} - 1)\end{gathered}$$

Η συνάρτηση κινδύνου είναι αυξανόμενη ξεκινώντας από το λ στο χρόνο μηδέν. Για τιμές της παραμέτρου $\phi < 0$, η συνάρτηση κινδύνου είναι φθίνουσα και η αθροιστική συνάρτηση κινδύνου συγκλίνει στη σταθερά $-\frac{\lambda}{\phi}$ για $t \to \infty$, έτσι ώστε να μην βιώνουν όλα τα άτομα, του πληθυσμού της μελέτης, το γεγονός. Προφανώς, η εκθετική κατανομή αποτελεί ειδική περίπτωση της κατανομής Gompertz στην περίπτωση της $\phi = 0$. Το μοντέλο Gompertz γενικεύτηκε στην κατανομή Gompertz-Makeham (Makeham 1860) με την προσθήκη μιας σταθεράς c στη συνάρτηση κινδύνου,

$$\mu(t) = \lambda e^{\phi t} + c.$$

Εδώ η πρόσθετη παράμετρος c περιγράφει μια μη-γηραντική πτυχή της μελέτης του πληθυσμού που είναι ανεξάρτητη από το χρόνο t, ενώ το τμήμα Gompertz εξακολουθεί να αντιπροσωπεύει την εξαρτώμενη από την ηλικία πτυχή με εκθετική μορφή. Οι παράμετροι λ και c δεν μπορούν να προσδιοριστούν στην περίπτωση $\phi = 0$, μόνο το άθροισμά τους μπορεί να εκτιμηθεί.

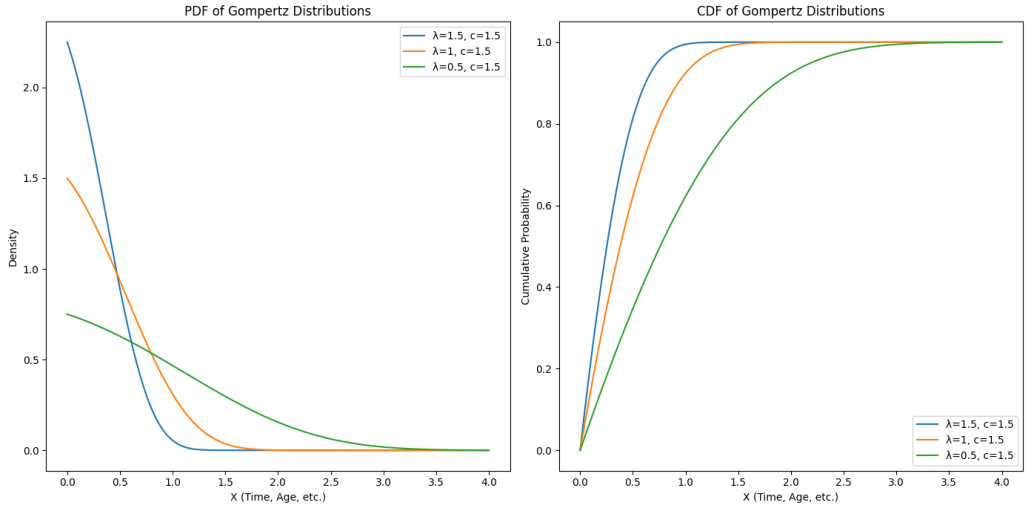
Η κατανομή Gompertz-Makeham περιγράφει την δυναμική της ηλικίας στην ανθρώπινη θνησιμότητα, με μεγάλη ακρίβεια στο ηλικιακό εύρος περίπου 30-80 ετών. Σε πιο προχωρημένες ηλικίες τα ποσοστά θνησιμότητας δεν αυξάνονται τόσο γρήγορα όσο προβλέπει ο νόμος της θνησιμότητας - ένα φαινόμενο γνωστό ως επιβράδυνση της θνησιμότητας στα τέλη της ζωής. Το φαινόμενο αυτό αποτέλεσε ένα από τα σημεία εκκίνησης για την ανάπτυξη μονομεταβλητών μοντέλων ευπάθειας.

□

Θα σχεδιάσουμε το PDF και το CDF της κατανομής Gompertz, που βλέπουμε στο σχήμα 6.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import gompertz

# Define the parameters for the Gompertz distribution
lambda_values = [1.5, 1, 0.5]
c = 1.5

# Generate x values
x = np.linspace(0, 4, 1000)

# Prepare the plot
plt.figure(figsize=(14, 7))

# Plotting PDF and CDF for each value of lambda
for lambda_val in lambda_values:
```

```
    pdf = gompertz.pdf(x, c=c, scale=1/lambda_val)
    cdf = gompertz.cdf(x, c=c, scale=1/lambda_val)

    plt.subplot(1, 2, 1)
    plt.plot(x, pdf, label=f'l={lambda_val}, c={c}')
    plt.title('PDF of Gompertz Distributions')
    plt.xlabel('X (Time, Age, etc.)')
    plt.ylabel('Density')
    plt.legend()

    plt.subplot(1, 2, 2)
    plt.plot(x, cdf, label=f'l={lambda_val}, c={c}')
    plt.title('CDF of Gompertz Distributions')
    plt.xlabel('X (Time, Age, etc.)')
    plt.ylabel('Cumulative Probability')
    plt.legend()

plt.tight_layout()
plt.show()
```



Σχήμα 7: Συνάρτηση πυκνότητας πιθανότητας (αριστερά) και αθροιστική συνάρτηση κατανομής (δεξιά) της κατανομής Gompertz.

Βλέπουμε πώς για διαφορετικές τιμές του λ έχουμε διαφορετικές επιρροές στη κατανομή, προσομοιώνοντας διαφορετικούς βαθμούς «τεντώματος» κατά μήκος του άξονα x. Χαμηλές τιμές του λ έχουν ως αποτέλεσμα μια πιο τεντωμένη κατανομή (βραδύτερη συσσώρευση του ποσοστού αποτυχίας), ενώ υψηλότερες τιμές συμπιέζουν την κατανομή (ταχύτερη συσσώρευση).

Διατηρώντας το c σταθερό, ο ρυθμός αύξησης της αποτυχίας με την πάροδο του χρόνου παραμένει ο ίδιος σε όλα τα διαγράμματα, αλλά η συνολική επίδραση στην εξάπλωση και την κορυφή της κατανομής διαμορφώνεται κατά λ.

Αυτή η τροποποίηση παρέχει μια σαφή απεικόνιση του τρόπου με τον οποίο διαφορετικές παράμετροι κλίμακας (λ) αλληλεπιδρούν με μια σταθερή παράμετρο σχήματος (c) για τη διαμόρφωση της κατανομής Gompertz, παρέχοντας πολύτιμες πληροφορίες για τη μοντελοποίηση δεδομένων.

1.4.5 Λογαριθμοκανονική Κατανομή

Στη θεωρία πιθανοτήτων, η λογαριθμοκανονική κατανομή είναι μια συνεχής κατανομή πιθανότητας μιας τυχαίας μεταβλητής της οποίας ο λογάριθμος ακολουθεί κανονική κατανομή. Έτσι, αν η τυχαία μεταβλητή X ακολουθεί λογαριθμοκανονική κατανομή, τότε η $Y = \ln(X)$ έχει κανονική κατανομή. Ισοδύναμα, αν η Y ακολουθεί κανονική κατανομή, τότε η εκθετική συνάρτηση της Y, $X = \exp(Y)$, έχει λογαριθμοκανονική κατανομή. Μια τυχαία μεταβλητή που ακολουθεί λογαριθμοκανονική κατανομή παίρνει μόνο θετικές πραγματικές τιμές.

Παράδειγμα 1.8. Όταν μία τυχαία μεταβλητή X ακολουθεί την κατανομή $N(m, s^2)$, τότε η τυχαία μεταβλητή $T = \exp(X)$ ακολουθεί τη Λογαριθμοκανονική κατανομή $\log\mathrm{N}(m, s^2)$ με παραμέτρους m, s^2. Ισχύουν οι ακόλουθες σχέσεις,

$$\text{Συνάρτηση Πυκνότητας Πιθανότητας} \quad f(t) = \frac{1}{\sqrt{2\pi}st} e^{-\frac{(\log t - m)^2}{2s^2}}$$

$$\text{Συνάρτηση Επιβίωσης} \quad S(t) = 1 - \Phi\left(\frac{\log t - m}{s}\right)$$

$$\text{Συνάρτηση Κινδύνου} \quad \mu(t) = \frac{\frac{1}{st}\phi\left(\frac{\log t - m}{s}\right)}{1-\Phi\left(\frac{\log t - m}{s}\right)}$$

$$\text{Μέση Τιμή} \quad ET = e^{m+\frac{s^2}{2}}$$

$$\text{Διασπορά} \quad V(T) = e^{2m+s^2}(e^{s^2}-1)$$

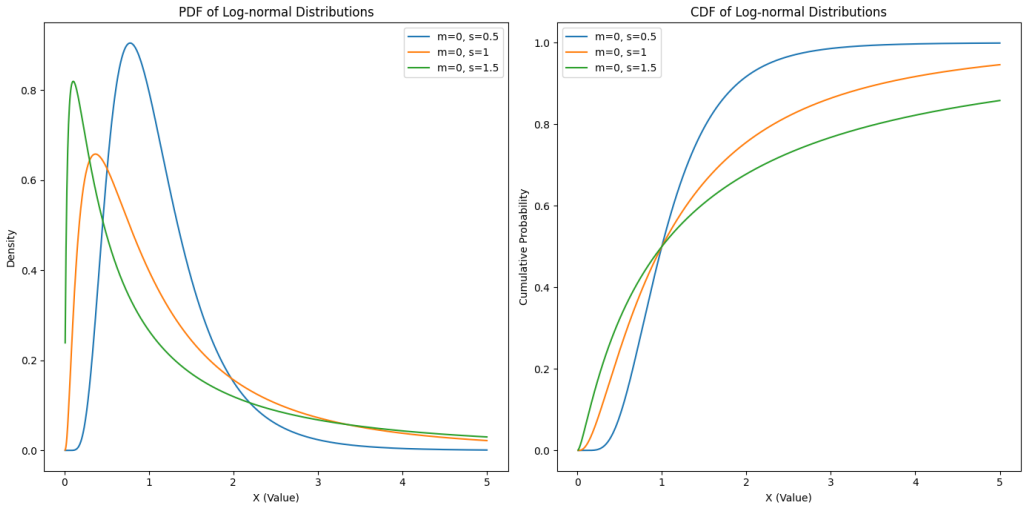
Στα παραπάνω η συνάρτηση Φ αποτελεί την συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής. Στη συνάρτηση κινδύνου για αυτήν την κατανομή βλέπουμε πως για $t = 0$ λαμβάνει τιμή 0, εν συνεχεία αυξάνεται έως ένα ολικό μέγιστο και μειώνεται τείνοντας προς το μηδέν, καθώς ο χρόνος τείνει στο άπειρο. Αυτό καθιστά την εν λόγω κατανομή ακατάλληλη για δεδομένα επιβίωσης ανθρώπων ιδιαίτερα για ανθρώπους μεγάλης ηλικίας. Ωστόσο, αν εξετάζουμε δεδομένα επιβίωσης νεότερων ανθρώπων ή βρεφών φαίνεται να λειτουργεί καλύτερα.

□

Θα σχεδιάσουμε το PDF και το CDF της Λογαριθμοκανονικής κατανομής, σχήμα 7.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import lognorm

# Define the parameters for the log-normal distribution
sigmas = [0.5, 1, 1.5]
mu = 0

# Generate x values
x = np.linspace(0.01, 5, 1000)

# Prepare the plot
plt.figure(figsize=(14, 7))
```

```
# Plotting PDF and CDF for each sigma
for sigma in sigmas:
    pdf = lognorm.pdf(x, s=sigma, scale=np.exp(mu))
    cdf = lognorm.cdf(x, s=sigma, scale=np.exp(mu))

    plt.subplot(1, 2, 1)
    plt.plot(x, pdf, label=f'm={mu}, s={sigma}')
    plt.title('PDF of Log-normal Distributions')
    plt.xlabel('X (Value)')
    plt.ylabel('Density')
    plt.legend()

    plt.subplot(1, 2, 2)
    plt.plot(x, cdf, label=f'm={mu}, s={sigma}')
    plt.title('CDF of Log-normal Distributions')
    plt.xlabel('X (Value)')
    plt.ylabel('Cumulative Probability')
    plt.legend()

plt.tight_layout()
plt.show()
```



Σχήμα 8: Συνάρτηση πυκνότητας πιθανότητας (αριστερά) και αθροιστική συνάρτηση κατανομής (δεξιά) της Λογαριθμοκανονικής κατανομής.

Η εντολή *lognorm* προσαρμόζει τη θέση της κατανομής με βάση το μέσο όρο του λογαρίθμου της μεταβλητής. Μεταβάλλοντας την τυπική απόκλιση, τα διαγράμματα απεικονίζουν διαφορετικούς βαθμούς διασποράς γύρω από τη μέση τιμή. Μια μεγαλύτερη διασπορά έχει ως αποτέλεσμα μια πιο διασκορπισμένη κατανομή, επηρεάζοντας τόσο την κορυφή όσο και τις ουρές της κατανομής. Αυτή η ρύθμιση βοηθά στην οπτικοποίηση του τρόπου με τον οποίο η λογαριθμοκανονική κατανομή συμπεριφέρεται με διαφορετικά επίπεδα μεταβλητότητας (ελέγχεται από τη διασπορά), καθιστώντας την ένα χρήσιμο εργαλείο.

1.4.6 Κατανομή Gamma

Η κατανομή Gamma είναι μια άλλη επέκταση της εκθετικής κατανομής. Είναι πιο δύσκολη η χρήση τους στην ανάλυση επιβίωσης επειδή τα μοντέλα αυτά δεν

έχουν κλειστής μορφής εκφράσεις για τις συναρτήσεις επιβίωσης και κινδύνου. Και τα δύο περιλαμβάνουν το ατελές ολοκλήρωμα,

$$I_k(x) = \frac{\int_0^x s^{k-1}e^{-s}\,ds}{\Gamma(k)}$$

Παράδειγμα 1.9. Κατά συνέπεια, η παραδοσιακή εκτίμηση μέγιστης πιθανοφάνειας δεν είναι απλή και απαιτεί τον υπολογισμό ατελών ολοκληρωμάτων, επιβάλλοντας αριθμητικά προβλήματα στην εκτίμηση των παραμέτρων. Εάν η μεταβλητή T ακολουθεί κατανομή Gamma με παράμετρο σχήματος k και αντίστροφη παράμετρο κλίμακας $\lambda(T \sim \Gamma(k,\lambda))$, τότε ισχύουν οι ακόλουθες σχέσεις,

$$\text{Συνάρτηση Πυκνότητας Πιθανότητας} \quad f(t) = \frac{\lambda^k t^{k-1}e^{-\lambda t}}{\Gamma(k)} \quad (k>0, \lambda>0)$$

$$\text{Συνάρτηση Επιβίωσης} \quad S(t) = 1 - I_k(\lambda t)$$

$$\text{Συνάρτηση Κινδύνου} \quad \mu(t) = \frac{\lambda^k t^{k-1}e^{-\lambda t}}{(1-I_k(\lambda t))\Gamma(k)}$$

$$\text{Μέση Τιμή} \quad ET = \frac{k}{\lambda}$$

$$\text{Διασπορά} \quad V(T) = \frac{k}{\lambda^2}$$

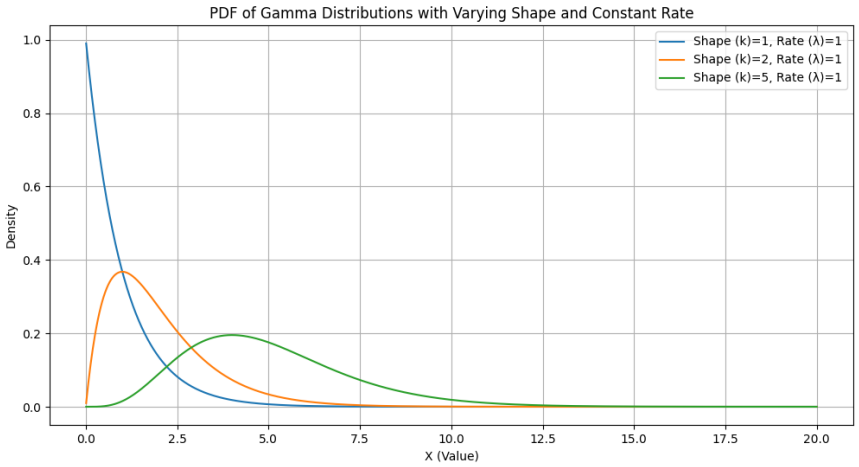
$$\text{Συνάρτηση Laplace} \quad L(u) = Ee^{-Tu} = (1+\tfrac{u}{\lambda})^{-k}$$

Εάν $k = 1$, η κατανομή Gamma μετατρέπεται σε εκθετική κατανομή. Με ακέραιο k, η κατανομή Gamma συχνά ονομάζεται ειδική κατανομή Erlangian. □

Θα σχεδιάσουμε το PDF της Gamma κατανομής, σχήμα 8.

```
import numpy as np
import matplotlib.pyplot as plt
```

```
from scipy.stats import gamma

# Define the shape and rate parameters
k_values = [1, 2, 5]  # Different shape parameters
lambda_value = 1  # Constant rate parameter

# Generate x values
x = np.linspace(0.01, 20, 1000)

# Prepare the plot
plt.figure(figsize=(12, 6))

# Plotting PDF for each shape parameter with constant
   rate
for k in k_values:
    pdf = gamma.pdf(x, a=k, scale=1/lambda_value)
    plt.plot(x, pdf, label=f'Shape (k)={k}, Rate (l)={
       lambda_value}')

plt.title('PDF of Gamma Distributions with Varying Shape
   and Constant Rate')
plt.xlabel('X (Value)')
plt.ylabel('Density')
plt.legend()
plt.grid(True)
plt.show()
```



Σχήμα 9: Απεικόνιση της κατανομής Gamma.

Διαφορετικές τιμές για την μεταβλητή k καταδεικνύουν πώς η παράμετρος λ επηρεάζει την κατανομή, με τη λ να διατηρείται σταθερή για να απομονωθεί η επίδραση της k. Παραπάνω έχουμε μια γραφική παράσταση που εμφανίζει την επίδραση της αλλαγής της παραμέτρου λ στη μορφή της κατανομής. Με μια υψηλότερη τιμή της k, η κατανομή γίνεται πιο συμμετρική και σε σχήμα καμπάνας, σε αντίθεση από το λοξό σχήμα που είναι τυπικό για τις χαμηλότερες k. Αυτό βοηθά στην κατανόηση της συμπεριφοράς διαδικασιών όπου ο ρυθμός εμφάνισης ή γεγονότων διέπεται από μια κατανομή Gamma.

1.4.7 Κατανομή Pareto

Η κατανομή Pareto εισήχθη με στόχο να εξηγήσει την κατανομή του εισοδήματος ενός συγκεκριμένου πληθυσμού στα τέλη του 19ου αιώνα. Η κατανομή είναι λοξή και με βαριά ουρά με δύο παραμέτρους $\omega > 0$ και $\zeta > 0$. Η k-οστή

ροπή της κατανομής Pareto είναι πεπερασμένη εάν ισχύει ο περιορισμός $\omega > k$. Η επικινδυνότητα είναι μονότονα φθίνουσα.

Παράδειγμα 1.10. Ισχύουν οι ακόλουθες σχέσεις,

$$\text{Συνάρτηση Πυκνότητας Πιθανότητας} \quad f(t) = \frac{\zeta}{\omega}\left(\frac{\omega}{\omega+t}\right)^{\zeta+1} (\omega > 0, \zeta > 0)$$

$$\text{Συνάρτηση Επιβίωσης} \quad S(t) = \left(\frac{\omega}{\omega+t}\right)^{\zeta}$$

$$\text{Συνάρτηση Κινδύνου} \quad \mu(t) = \frac{\zeta}{\omega+t}$$

$$\text{Αθροιστική Συνάρτηση Κινδύνου} \quad M(t) = -\zeta \log\left(\frac{\omega}{\omega+t}\right)$$

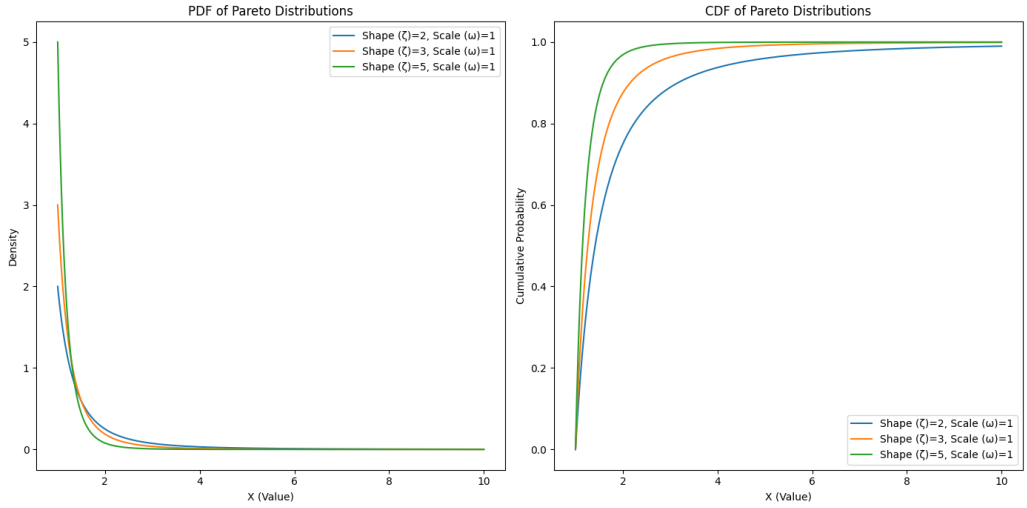
$$\text{Μέση Τιμή} \quad ET = \frac{\omega}{\zeta-1} \quad (\zeta > 1)$$

□

Θα σχεδιάσουμε το PDF και το CDF της κατανομής Pareto, σχήμα 9.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import pareto

# Define the scale and shape parameters
omega = 1  # Scale parameter (minimum value)
zeta_values = [2, 3, 5]  # Different shape parameters

# Generate x values
x = np.linspace(omega, 10, 1000)  # x must be >= omega

# Prepare the plot
plt.figure(figsize=(14, 7))
```

```
# Plotting PDF and CDF for each shape parameter
for zeta in zeta_values:
    pdf = pareto.pdf(x, b=zeta, scale=omega)
    cdf = pareto.cdf(x, b=zeta, scale=omega)

    plt.subplot(1, 2, 1)
    plt.plot(x, pdf, label=f'Shape (z)={zeta}, Scale (w)
       ={omega}')
    plt.title('PDF of Pareto Distributions')
    plt.xlabel('X (Value)')
    plt.ylabel('Density')
    plt.legend()

    plt.subplot(1, 2, 2)
    plt.plot(x, cdf, label=f'Shape (z)={zeta}, Scale (w)
       ={omega}')
    plt.title('CDF of Pareto Distributions')
    plt.xlabel('X (Value)')
    plt.ylabel('Cumulative Probability')
    plt.legend()

plt.tight_layout()
plt.show()
```



Σχήμα 10: Συνάρτηση πυκνότητας πιθανότητας (αριστερά) και αθροιστική συνάρτηση κατανομής (δεξιά) της κατανομής Pareto.

Διαφορετικές τιμές για το ζ δείχνουν πώς αλλάζει η ουρά της κατανομής. Τα γραφήματα δείχνουν πώς αλλάζουν το PDF και το CDF για διαφορετικές ζ τιμές. Υψηλότερες τιμές για το ζ οδηγούν σε ταχύτερη πτώση της πυκνότητας και ταχύτερη αύξηση της αθροιστικής συνάρτησης, γεγονός που αντανακλά σε λιγότερες ακραίες τιμές.

1.5 Μη Παραμετρική Εκτίμηση των συναρτήσεων Επιβίωσης και Κινδύνου

Στην περίπτωση της παραμετρικής εκτίμησης, απαιτούνται υποθέσεις σχετικά με την κατανομή των χρόνων αποτυχίας. Αυτή η προσέγγιση μπορεί να είναι λογική, ωστόσο, εάν οι πληροφορίες αυτές δεν είναι διαθέσιμες, χρησιμοποιούνται συνήθως μη παραμετρικά μοντέλα. Η απλούστερη μη παραμετρική εκτίμηση

μιας συνάρτησης κατανομής είναι η εμπειρική συνάρτηση κατανομής, η οποία οδηγεί σε μια διακριτή εκτίμηση ακόμη και για μια συνεχή κατανομή.

Σημαντικά μη παραμετρικά και ημιπαραμετρικά μοντέλα για την ανάλυση επιβίωσης, με λογοκριμένες παρατηρήσεις, χρησιμοποιούν την εκτιμήτρια Kaplan-Meier (1958) και την επέκτασή της σε περικομμένους/αποκομμένους χρόνους ζωής (Turnbull 1976, Tsai et al. 1987). Άλλη μία επίσης σημαντική τεχνική αποτελεί το ημιπαραμετρικό μοντέλο παλινδρόμησης αναλογικών κινδύνων (Cox 1972).

Η επιλογή μεταξύ ενός παραμετρικού ή ενός μη παραμετρικού μοντέλου είναι ζωτικής σημασίας. Τα μη παραμετρικά μοντέλα είναι ευέλικτα και μπορούν να χειριστούν οποιαδήποτε κατανομή πιθανότητας, αλλά απαιτούν πολύ περισσότερα δεδομένα για να δώσουν αξιόπιστα αποτελέσματα. Επιπλέον, η εκτίμηση της συνάρτησης κινδύνου είναι δύσκολη με τις μη παραμετρικές μεθόδους, αντίθετα, τα παραμετρικά μοντέλα μπορούν να αποδώσουν καλά αποτελέσματα ακόμη και με μικρά μεγέθη δείγματος. Καθώς αν οι υποθέσεις του μοντέλου είναι σωστές, η εκτίμηση είναι πιο αποτελεσματική από ό,τι με τις μη παραμετρικές μεθόδους.

1.5.1 Εκτιμήτρια Kaplan–Meier

Ένα χρήσιμο μέσο για τον χαρακτηρισμό της επιβίωσης σε μια ομάδα ατόμων είναι ο υπολογισμός και η γραφική απεικόνιση της εμπειρικής συνάρτησης επιβίωσης (empirical estimator).

$$S(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(t_i > t),$$

όπου $\mathbf{1}$ είναι η δείκτρια συνάρτηση, η οποία παίρνει τιμή 1 όταν ισχύει $t_i > t$ και μηδέν οπουδήποτε αλλού. Εάν δεν υπάρχουν λογοκριμένες παρατηρήσεις στο δείγμα, η εμπειρική συνάρτηση επιβίωσης τη χρονική στιγμή t είναι ο λόγος των επιζώντων τη χρονική στιγμή t ως προς το μέγεθος του δείγματος n. Αυτή η βηματική συνάρτηση μειώνεται κατά $1/n$ αμέσως μετά από κάθε παρατηρούμενη αποτυχία. Συνεπώς, ο εκτιμητής είναι ουσιαστικά το ποσοστό αυτών που έχουν επιβιώσει μέχρι και την χρονική στιγμή t.

Στη περίπτωση όμως που υπάρχουν και λογοκριμένες παρατηρήσεις, αναπτύχθηκε η εκτιμήτρια Kaplan-Meier, η οποία είναι ουσιαστικά μία επέκταση του εμπειρικού εκτιμητή για δεδομένα που περιέχουν και λογοκριμένες τιμές. Η εκτιμήτρια Kaplan-Meier στην ουσία δίνει εκτιμήσεις για τις τιμές της συνάρτησης επιβίωσης για μη λογοκριμένους χρόνους και για τον υπολογισμό της

εκτιμήτριας λαμβάνονται υπόψιν και οι ακριβείς χρόνοι επιβίωσης και οι λογοκριμένοι χρόνοι. Θεωρούμε τους παρακάτω διατεταγμένους χρόνους επιβίωσης,

$$t_{(1)} < t_{(2)} < t_{(3)} < \cdots < t_{(n)},$$

έχοντας έτσι τότε την εκτίμηση Kaplan-Meier της συνάρτησης επιβίωσης για τους μη λογοκριμένους χρόνους,

$$S(t) = \begin{cases} 1 & \text{εάν } t_{(1)} > t, \\ \prod_{i:t_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right) & \text{εάν } t_{(1)} \leq t. \end{cases}$$

όπου d_i είναι ο αριθμός των ατόμων που απεβίωσαν τη χρονική στιγμή t_i και το n_i αποτελεί το σύνολο των ατόμων που επέζησαν ακριβώς πριν αυτή τη χρονική στιγμή t_i.

Παράδειγμα 1.11. Ας υποθέσουμε ότι εξετάζουμε την αποτελεσματικότητα ενός νέου φαρμάκου στην παράταση της ζωής των ασθενών που έχουν διαγνωστεί με έναν συγκεκριμένο τύπο καρκίνου. Θέλουμε να χρησιμοποιήσουμε την εκτιμήτρια Kaplan-Meier για να απεικονίσουμε τις πιθανότητες επιβίωσης με την πάροδο του χρόνου.

Στα δεδομένα μας έχουμε δύο ομάδες, όπου η κάθε μία αποτελείται απο 10 ασθενείς,

Ομάδα Α: Ασθενείς που λαμβάνουν το νέο φάρμακο.
Ομάδα Β: Ασθενείς που λαμβάνουν μια τυπική θεραπεία.

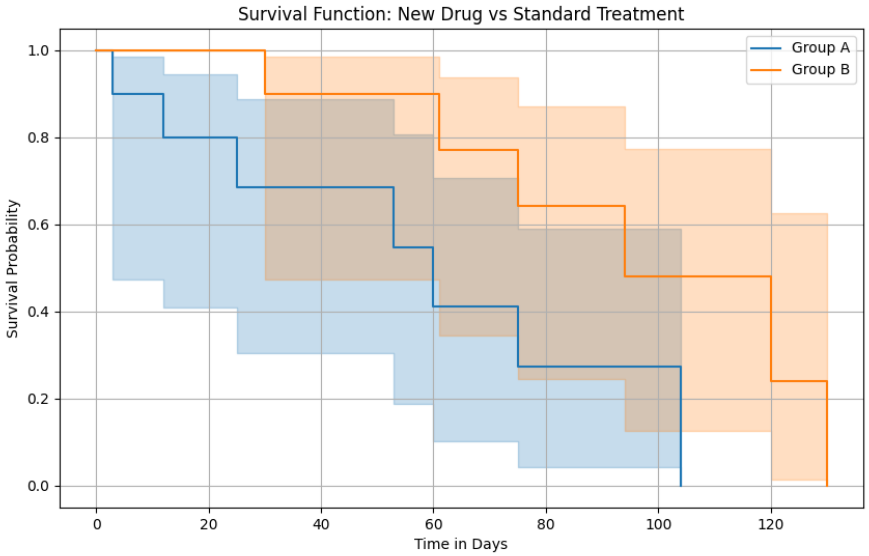
Τα δεδομένα για κάθε ομάδα μπορεί να περιλαμβάνουν, το χρόνο μέχρι το θάνατο ή την τελευταία παρακολούθηση (σε ημέρες). Επίσης είναι γνωστό και το γεγονός του αν στην ουσία είναι γνωστός ο χρόνος μέχρι τη λήξη της ζωής, το οποίο δείχνει εάν επήλθε ο θάνατος (1 εάν επήλθε το συμβάν (θάνατος), 0 εάν λογοκρίθηκε). Οι τιμές του χρόνου που παρέχονται στο σύνολο δεδομένων κυμαίνονται από τον μικρότερο παρατηρούμενο χρόνο (3 ημέρες) έως τον μεγαλύτερο (130 ημέρες), καλύπτοντας ένα εύρος περίπου 4 μηνών. Το εύρος αυτό υποδεικνύει την περίοδο κατά την οποία παρακολουθήθηκε η επιβίωση των ασθενών στη μελέτη, δίνοντας πληροφορίες για τις πιθανότητες επιβίωσης σε διαφορετικές χρονικές στιγμές εντός αυτών των 130 ημερών.

Χρησιμοποιούμε τη βιβλιοθήκη lifelines της Python,

```
import pandas as pd
import numpy as np
from lifelines import KaplanMeierFitter
import matplotlib.pyplot as plt

# Sample Data
data = {
    'time': [3, 12, 19, 25, 44, 53, 60, 75, 90, 104, 30,
       42, 56, 61, 75, 85, 94, 110, 120, 130],
    'event': [1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1,
       1, 0, 1, 0, 1, 1],
    'group': ['A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A'
       , 'A', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B'
       , 'B']
}

df = pd.DataFrame(data)

# Create a KaplanMeierFitter object
kmf = KaplanMeierFitter()

# Plot settings
plt.figure(figsize=(10, 6))

for group in df['group'].unique():
    group_data = df[df['group'] == group]
    kmf.fit(group_data['time'], event_observed=group_data
       ['event'], label=f'Group {group}')
    kmf.plot_survival_function()

plt.title('Survival Function: New Drug vs Standard
   Treatment')
plt.xlabel('Time in Days')
```

```
plt.ylabel('Survival Probability')
plt.grid(True)
plt.legend()
plt.show()
```

Δημιουργούμε αρχικά ένα DataFrame με δειγματικούς χρόνους επιβίωσης, γεγονότα και ομάδες. Στη συνέχεια γίνεται μια προσαρμογή του μοντέλου, το οποίο γίνεται για κάθε ομάδα για τον υπολογισμό της συνάρτησης επιβίωσης. Οι συναρτήσεις επιβίωσης σχεδιάζονται για κάθε ομάδα, δείχνοντας την πιθανότητα επιβίωσης με την πάροδο του χρόνου.



Σχήμα 11: Καμπύλη επιβίωσης Kaplan-Meier, η οποία απεικονίζει τις πιθανότητες επιβίωσης των ασθενών με την πάροδο του χρόνου για δύο διαφορετικές ομάδες - την ομάδα A και την ομάδα B.

Η γραφική παράσταση απεικονίζει τις εκτιμώμενες καμπύλες επιβίωσης για τους ασθενείς που λαμβάνουν το νέο φάρμακο σε σχέση με εκείνους που λαμβάνουν την τυπική θεραπεία. Αυτή η απεικόνιση βοηθά στη σύγκριση της αποτελεσματικότητας των θεραπειών. Εάν η καμπύλη για την ομάδα Α (νέο φάρμακο) παραμένει υψηλότερη από την ομάδα Β (τυπική θεραπεία), αυτό υποδηλώνει ότι οι ασθενείς που λαμβάνουν το νέο φάρμακο τείνουν να επιβιώνουν περισσότερο, υποδεικνύοντας την πιθανή αποτελεσματικότητά του.

Και οι δύο ομάδες ξεκινούν με πιθανότητα επιβίωσης 1 (όλοι οι ασθενείς είναι ζωντανοί) και βιώνουν το πρώτο γεγονός (θάνατος ή εγκατάλειψη μελέτης) σε διαφορετικούς χρόνους. Κάθε βήμα προς τα κάτω σε μια γραμμή αντιπροσωπεύει την εμφάνιση ενός συμβάντος. Μια πιο απότομη κλίση στην καμπύλη επιβίωσης υποδηλώνει υψηλότερη πιθανότητα θανάτου στο συγκεκριμένο χρονικό διάστημα. Για παράδειγμα, εάν η γραμμή της ομάδας Β πέφτει ταχύτερα από εκείνη της ομάδας Α, αυτό υποδηλώνει ότι οι ασθενείς της ομάδας Β εμφανίζουν συμβάντα με υψηλότερο ρυθμό κατά τη διάρκεια εκείνης της περιόδου. Τα σημεία στα οποία η γραμμή κατεβαίνει προς τα κάτω υποδεικνύουν χρονικές στιγμές κατά τις οποίες παρατηρείται ένα συμβάν (όπως ο θάνατος). Τα οριζόντια τμήματα των γραμμών υποδηλώνουν περιόδους χωρίς κανένα συμβάν ή όπου τα δεδομένα ενδέχεται να λογοκρίνονται (π.χ. ένας ασθενής που χάθηκε για παρακολούθηση ή είναι ακόμη ζωντανός χωρίς εμφάνιση συμβάντος στο τέλος της μελέτης).

Συγκρίνοντας τις καμπύλες επιβίωσης, μπορεί κανείς να εκτιμήσει ποια ομάδα τα πηγαίνει καλύτερα με την πάροδο του χρόνου όσον αφορά την επιβίωση. Σε αυτό το διάγραμμα, εάν η γραμμή της ομάδας Α είναι σταθερά πάνω από τη γραμμή της ομάδας Β, αυτό υποδηλώνει ότι η ομάδα Α έχει μεγαλύτερη πιθανότητα επιβίωσης τις περισσότερες φορές κατά τη διάρκεια της μελέτης. Αντίθετα, αν η γραμμή της ομάδας Β είναι πάνω από αυτήν, αυτό υποδηλώνει καλύτερη επιβίωση για την ομάδα αυτή. Απλά κοιτάζοντας τις καμπύλες, εάν διαχωρίζονται σημαντικά και δεν επικαλύπτονται πολύ (όπως τα διαστήματα εμπιστοσύνης που εμφανίζονται ως σκιασμένες περιοχές γύρω από τις καμπύλες), αυτό μπορεί να υποδηλώνει μια στατιστικά σημαντική διαφορά στην επιβίωση μεταξύ των δύο ομάδων. Ωστόσο, η οπτική επιθεώρηση από μόνη της δεν αρκεί για οριστικά συμπεράσματα. Συνήθως, πρέπει να χρησιμοποιηθεί ένας έλεγ-

χος log-rank ή άλλοι στατιστικοί έλεγχοι για να προσδιοριστεί εάν οι διαφορές μεταξύ των καμπύλων είναι στατιστικά σημαντικές, λαμβάνοντας υπόψη τον αριθμό των ασθενών και των συμβάντων.

□

Η εκτιμήτρια Kaplan-Meier αποτελεί δηλαδή μία βηματική συνάρτηση και αυτό εκφράζεται και από τη γραφική της αναπαράσταση. Οι εκτιμήσεις απεικονίζονται στην καμπύλη επιβίωσης όπου οι μη λογοκριμένοι χρόνοι είναι χωρισμένοι σε διαστήματα και σε κάθε ένα από αυτά αντιστοιχεί μία τιμή για την συνάρτηση επιβίωσης. Δηλαδή η Kaplan-Meier μας δίνει μία εκτίμηση για την συνάρτηση επιβίωσης.

Σημαντικό είναι να αναφέρουμε πως θα πρέπει να διερευνήσουμε την ακρίβεια της εκτιμήτριας. Ένα μέτρο ακρίβειας μίας εκτιμήτριας αντικατοπτρίζεται στο τυπικό της σφάλμα. Ουσιαστικά, μικρές τιμές του τυπικού σφάλματος υποδηλώνουν μικρές διακυμάνσεις της εκτιμήτριας από δείγμα σε δείγμα. Επιπλέον, με τη βοήθεια του τυπικού σφάλματος μπορούν να κατασκευαστούν και κατάλληλα διαστήματα εμπιστοσύνης.

Η εκτίμηση Kaplan-Meier για τη συνάρτηση επιβίωσης για μία τυχαία τιμή t στο χρονικό διάστημα $[t_{(k)}, t_{(k+1)})$ μπορεί να γραφεί στη μορφή,

$$S(t) = \prod_{j=1}^{k} p_j$$

όπου $\kappa = 1, \ldots, n,\ \ p_j = (n_j - d_j)/n_j$ είναι η εκτίμηση της πιθανότητας ότι μία μονάδα του πληθυσμού επέζησε στο χρονικό διάστημα $[t_{(k)}, t_{(k+1)}]$, όπου $j = \kappa = 1, \ldots, n$.

Αφού λογαριθμίσουμε την παραπάνω σχέση, έχουμε τη διακύμανση της Kaplan-Meier και ο τύπος Greenwood (Greenwood 1926) υπολογίζεται από την έκφραση,

$$Var(\log(S(t))) = \sum_{j=1}^{k} Var(\log(p_j)).$$

1.5.2 Μέθοδος Nelson-Aalen

Η μέθοδος Nelson-Aalen (Nelson, W. (1972) & Aalen, O. O. (1978)) αποτελεί μία μη παραμετρική μέθοδο για την εκτίμηση της αθροιστικής συνάρτησης κινδύνου $\hat{M}_T(\cdot)$, η οποία προτάθηκε από τον Nelson (1972) ως ένας δείκτης

αξιοπιστίας και επαναχρησιμοποιήθηκε από τον Aalen (1978). Η εκτιμήτρια Nelson-Aalen για την αθροιστική συνάρτηση κινδύνου δίνεται από τη σχέση,

$$\hat{M}_T(t) = \sum_{t_j \leq t} \frac{d_j}{n_j}$$

όπου n_j είναι τα άτομα που βρίσκονται σε κίνδυνο ακριβώς πριν την χρονική στιγμή t_j. Συνεπώς, η εκτιμήτρια Nelson-Aalen αποτελεί μία βηματική μέθοδο, συνεχή από δεξιά, η οποία δίνει εκτιμήσεις σημειακά για τους παρατηρούμενους πραγματικούς χρόνους επιβίωσης. Η διασπορά της εκτιμήτριας δίνεται από τη σχέση,

$$Var(\hat{M}_T(t)) = \sum_{t_j \leq t} \frac{d_j}{n_j^2}.$$

Μπορεί να αποδειχθεί ότι η εκτιμήτρια καθώς και η διασπορά της εκτιμήτριας Nelson-Aalen είναι σχεδόν αμερόληπτες. Σε μεγάλα δείγματα η εκτιμήτρια Nelson-Aalen, που αξιολογείται σε δεδομένη χρονική στιγμή t, είναι περίπου κανονικά κατανεμημένη, οπότε ένα $(100(1-\alpha)\%)$ διάστημα εμπιστοσύνης για την $M_T(t)$ έχει τη μορφή,

$$\left(\hat{M}_T(t) - Z_{1-\frac{\alpha}{2}} \cdot se(\hat{M}_T(t)), \hat{M}_T(t) + Z_{1-\frac{\alpha}{2}} \cdot se(\hat{M}_T(t))\right),$$

όπου $se(\hat{M}_T(t))$ είναι η τετραγωνική ρίζα του $Var(\hat{M}_T(t))$ και $z_{\frac{1-\alpha}{2}}$ το $(1-\alpha)/2$ ποσοστιαίο σημείο της τυποποιημένης Κανονικής κατανομής.

Η προσέγγιση της κανονικής κατανομής βελτιώνεται με τη χρήση ενός λογαριθμικού μετασχηματισμού που δίνει το διάστημα εμπιστοσύνης,

$$\hat{M}_T(t) \exp\left[\pm z_{\frac{1-\alpha}{2}} \frac{se(\hat{M}_T(t))}{\hat{M}_T(t)}\right].$$

1.6 Μοντέλο αναλόγων κινδύνων του Cox

Τα μοντέλα που έχουν παρουσιαστεί μέχρι στιγμής αφορούν απλές περιπτώσεις ανεξάρτητων μεταβλητών, όπου συνεπάγεται μια ομοιογένεια στον υπό μελέτη πληθυσμό. Ωστόσο, στις περισσότερες πραγματικές εφαρμογές, ο υπό μελέτη πληθυσμός δεν είναι ομοιογενής. Για παράδειγμα, τα άτομα σε επιδημιολογικές μελέτες μπορεί να διαφέρουν ως προς την ηλικία, το φύλο, την κοινωνικοοικονομική κατάσταση, την εκπαίδευση και άλλους παράγοντες. Ο ερευνητής

λοιπόν στις περισσότερες περιπτώσεις ενδιαφέρεται να εξετάσει κατά πόσο και πώς επηρεάζουν οι μεταβλητές αυτές τον χρόνο επιβίωσης. Το μοντέλο αναλογικών κινδύνων που εισήγαγε ο Cox (1972) είναι ένα μοντέλο παλινδρόμησης με το χρόνο του συμβάντος να εμπλέκεται ως εξαρτημένη μεταβλητή μέσω της συνάρτησης κινδύνου. Αυτό στην ουσία επιτρέπει τη συμπερίληψη πληροφοριών σχετικά με γνωστές (παρατηρούμενες) συμμεταβλητές σε μοντέλα δεδομένων επιβίωσης με εύκολο τρόπο και είναι το το πιο ευρέως χρησιμοποιούμενο μοντέλο στον τομέα αυτό. Το μοντέλο του Cox λοιπόν αποτελεί μία στατιστική τεχνική, ώστε να προβλέψουμε π.χ. την επίδραση της θεραπείας στην επιβίωση των ασθενών, δεδομένων των τιμών των επεξηγηματικών μεταβλητών που επηρεάζουν τον χρόνο επιβίωσης του ασθενούς και κατ' επέκταση τον κίνδυνο θανάτου του.

Έστω $\mu(t \mid \mathbf{X})$ ο κίνδυνος ενός ατόμου τη χρονική στιγμή t με διάνυσμα συμμεταβλητών $\mathbf{X}' = (X_1, \ldots, X_k)$. Εδώ το $\mathbf{X}'$ δηλώνει τον ανάστροφο του διανύσματος στηλών $\mathbf{X}$. Το μοντέλο αναλογικών κινδύνων ορίζεται από τη σχέση,

$$\mu(t \mid X) = \mu_0(t)h(X) \tag{19}$$

όπου $\mu_0(t)$ είναι η βασική συνάρτηση κινδύνου και $h(\cdot)$ κάποια θετική συνάρτηση. Το μοντέλο υποθέτει ένα κοινό χαρακτηριστικό, όπως έναν κίνδυνο (π.χ. κίνδυνος θανάτου), που θα έχουν τα άτομα στον υπό μελέτη πληθυσμό. Οι παράμετροι πρωταρχικού ενδιαφέροντος περιέχονται στο $h(\mathbf{X}) = h(\beta, \mathbf{X})$, συχνά

$$h(X) = e^{\beta'\mathbf{X}}$$

όπου $\beta' = (\beta_1, \beta_2, \ldots, \beta_k)$ το διάνυσμα των παραμέτρων παλινδρόμησης. Σε αυτό το μοντέλο, οι συμμεταβλητές δρουν πολλαπλασιαστικά στον αρχικό κίνδυνο, καθώς βάση των προγνωστικών πληροφοριών ενός ατόμου, προσθέτουν περαιτέρω κινδύνους. Στο μοντέλο αυτό γίνεται η υπόθεση πως η ατομική διακύμανση κινδύνου χαρακτηρίζεται από ένα πεπερασμένης διάστασης διάνυσμα, παρατηρούμενων συμμεταβλητών, οι οποίες ονομάζονται επεξηγηματικές μεταβλητές. Η βασική ιδέα πίσω από αυτό είναι ο διαχωρισμός του χρόνου της συνάρτησης κινδύνου, αφενός, και της επίδρασης των συμμεταβλητών σε έναν εκθετικό όρο από την άλλη. Στην ουσία, η υπόθεση αυτή δείχνει ότι οι κίνδυνοι δύο ατόμων τη χρονική στιγμή t συνδέονται με μια σταθερά που δεν εξαρτάται από τη χρονική στιγμή t. Η απλή περίπτωση δύο δειγμάτων επιτυγχάνεται με τον περιορισμό σε μία μόνο ($k = 1$) δυαδική τυχαία μεταβλητή X στο μοντέλο με $X = 0$ ή $X = 1$, ανάλογα με τη συμμετοχή στην ομάδα. Ο όρος e^{β} δηλώνει

τον λόγο κινδύνου μεταξύ των δύο ομάδων. Ανάλογα με το αν η βασική συνάρτηση κινδύνου $\mu_0(t)$ είναι άγνωστη ή άγνωστη μόνο μέσω μίας πεπερασμένης παραμέτρου, το μοντέλο ονομάζεται μη παραμετρικό ή ημιπαραμετρικό.

Η συνάρτηση επιβίωσης δεδομένων των συμμεταβλητών $\mathbf{X}$ είναι,

$$S(t \mid \mathbf{X}) = S_0(t)^{e^{\beta' \mathbf{x}}},$$

όπου, $S_0(t) = e^{-\int_0^t \mu_0(s)\, ds}$ αποτελεί τη βασική συνάρτηση επιβίωσης και οι συνιστώσες του διανύσματος β είναι άγνωστες παράμετροι παλινδρόμησης. Αυτό σημαίνει ότι η συνάρτηση επιβίωσης ενός ατόμου με διάνυσμα συμμεταβλητών $\mathbf{X}$ είναι μια δύναμη της βασικής συνάρτησης επιβίωσης.

Δύο διαφορετικές προσεγγίσεις είναι δυνατές στο μοντέλο αναλογικών κινδύνων. Ορισμένες φορές οι συμμεταβλητές έχουν λοξές κατανομές, για παράδειγμα, όταν μόνο ένα μικρό κλάσμα των ατόμων εκτίθεται στον παράγοντα κινδύνου που μας ενδιαφέρει. Όμως είναι επίσης πολύ συχνό ένα μεγάλο μέρος των παρατηρήσεων να είναι λογοκριμένο. Ειδικά σε μεγάλες μελέτες που αναλύονται οι επιδράσεις μιας σπάνιας έκθεσης σε ένα συμβάν, ο αριθμός των εκτεθειμένων περιπτώσεων μπορεί να είναι πολύ μικρός. Κάποιος μπορεί τότε να αμφισβητήσει την εγκυρότητα των συμπερασμάτων που βασίζονται σε ασυμπτωτικά αποτελέσματα. Στην παραμετρική περίπτωση, ο κίνδυνος επιλέγεται από την κατηγορία των παραμετρικών κατανομών διάρκειας ζωής. Για παράδειγμα, η συνάρτηση πιθανότητας στο μοντέλο Weibull με $\theta = (\lambda, \nu)$ έχει τη μορφή,

$$L(\beta, \theta) = \prod_{i=1}^{n} \left(\lambda \nu t_i^{\nu-1} e^{\beta' X_i} \right)^{\delta_i} e^{-\lambda {t_i}^{\nu} e^{\beta' X_i}} \tag{20}$$

και με $\theta = (\lambda, \phi)$ έχουμε το μοντέλο Gompertz,

$$L(\beta, \theta) = \prod_{i=1}^{n} \left(\lambda e^{\phi t_i} e^{\beta' X_i} \right)^{\delta_i} e^{-\lambda \frac{(e^{\phi t_i} - 1)}{\phi} e^{\beta' X_i}}. \tag{21}$$

Οι εκτιμήσεις των παραμέτρων β εξαρτώνται από την παραμετρική υπόθεση σχετικά με την συνάρτηση κινδύνου. Αυτό περιορίζει τη δυνατότητα εφαρμογής του λόγου αναλογικών κινδύνων, διότι λεπτομερέστερες υποθέσεις σχετικά με τη μορφή της συνάρτησης κινδύνου είναι απαραίτητες. Το μοντέλο Weibull είναι το μόνο μοντέλο αναλογικών κινδύνων που είναι και ταυτόχρονα ένα μοντέλο επιταχυνόμενου χρόνου αποτυχίας (AFT). Οι παράμετροι μπορούν εύκολα να μετασχηματιστούν από το ένα μοντέλο στο άλλο και αντίστροφα. Παρόμοια με το μοντέλο αναλογικών κινδύνων, το μοντέλο (AFT) επιτρέπει μια εύκολη

και φυσική ερμηνεία των εκτιμήσεων των παραμέτρων. Η παραμετρική προσέγγιση είναι πολύ πιο σημαντική στα μοντέλα ευπάθειας από ότι στα μοντέλα αναλογικών κινδύνων.

Στην ημιπαραμετρική περίπτωση (χωρίς παραμετρικό προσδιορισμό της βασικής συνάρτησης κινδύνου), το μοντέλο είναι φυσικό και αρκετά ευέλικτο. Δεδομένου ότι το $e^{\beta' X}$ είναι πάντα θετικό, ο ατομικός κίνδυνος $\mu(t|X)$ είναι αυτομάτως μη αρνητικός για όλα τα t και όλες τις τιμές του β. Ένας δεύτερος τρόπος ενίσχυσης αυτού του μοντέλου είναι το γεγονός ότι η λογοκρισία και οι αναλογικοί κίνδυνοι προσαρμόζονται εύκολα σε αυτή την περίπτωση. Εφόσον η πιθανότητα περιέχει την απροσδιόριστη συνάρτηση κινδύνου, η χρήση της παραδοσιακής μέγιστης πιθανοφάνειας καθίσταται προβληματική. Ως εκ τούτου, χρειαζόμαστε μια προσαρμοσμένη έκδοση της πιθανοφάνειας που περιέχει επαρκείς πληροφορίες για το διάνυσμα παραμέτρων β αλλά όχι την απροσδιόριστη βασική συνάρτηση κινδύνου. Ο Cox πρότεινε τη χρήση της προσέγγισης της μερικής πιθανοφάνειας. Συμβολίζουμε τα παρατηρούμενα δεδομένα διάστασης n με $(t_i, \delta_i, X_i)(i = 1, ..., n)$. Η πιθανοφάνεια των δεδομένων επιβίωσης που δίνεται στην σχέση (21), μπορεί τώρα να γραφεί στη μορφή,

$$\prod_{i=1}^{n} \left(\mu_0(t_i)e^{\beta' X_i}\right)^{\delta_i} \exp\left(-M_0(t_i)e^{\beta' X_i}\right) \tag{22}$$

Αυτή η πιθανοφάνεια περιέχει την άγνωστη συνάρτηση βασικού κινδύνου $\mu_0(\cdot)$ αλλά και την άγνωστη σωρευτική βασική συνάρτηση κινδύνου. Θα θεωρήσουμε ότι ο βασικός κίνδυνος $\mu_0(\cdot)$, είναι μηδενικός εκτός από τις χρονικές στιγμές t_i, στις οποίες λαμβάνει χώρα ένα συμβάν. Αυτό οδηγεί σε έναν ορισμό της σωρευτικής βασικής συνάρτησης κινδύνου, ο οποίος αποτελεί διακριτή εκδοχή του αθροιστικού βασικού κινδύνου,

$$M_0^d(t) = \sum_{t_j \leq t} \mu_0(t_j).$$

Αντικαθιστώντας αυτή τη συνάρτηση βασικού αθροιστικού κινδύνου στην πιθανοφάνεια (22) προκύπτει,

$$\prod_{i=1}^{n} \left(\left(\mu_0(t_i)e^{\beta' X_i}\right)^{\delta_i} \exp\left(-(\sum_{j \in R(t_i)} \mu_0(t_j))e^{\beta' X_i} \right) \right),$$

$$\prod_{i=1}^{n}\left(\left(\mu_0(t_i)e^{\beta' X_i}\right)\exp\left(-\mu_0(t_i)\sum_{j\in R(t_i)}e^{\beta' X_j}\right)\right)^{\delta_i}, \tag{23}$$

όπου $R(t)$ αποτελεί το σύνολο κινδύνου τη χρονική στιγμή t που περιλαμβάνει όλα τα άτομα που εξακολουθούν να διατρέχουν τον κίνδυνο να βιώσουν το γεγονός που μας ενδιαφέρει τη χρονική στιγμή t. Πρόκειται για μια τροποποιημένη συνάρτηση πιθανοφάνειας. Λαμβάνοντας τον λογάριθμο και τη μερική παράγωγο ως προς $\mu_0(t_i)$ της σχέσης (23) έχουμε,

$$\frac{1}{\mu_0(t_i)} - \sum_{j\in R(t_i)} e^{\beta' X_j}$$

Η εξίσωση αυτών των μερικών παραγώγων με το μηδέν συνεπάγεται,

$$\hat{\mu}_0(t_i) = \left(\frac{\delta_i}{\sum_{j\in R(t_i)} e^{\beta' X_j}}\right). \tag{24}$$

Εισάγοντας αυτή την σχέση στην (23), προκύπτει μια συνάρτηση πιθανοφάνειας που περιέχει ως αγνώστους μόνο τις παραμέτρους παλινδρόμησης:

$$L(\beta) = \prod_{i=1}^{n}\left(\frac{e^{\beta' X_i}}{\sum_{j\in R(t_i)} e^{\beta' X_j}}\right)^{\delta_i} \tag{25}$$

όπου $R(t_i)$ είναι το σύνολο των μονάδων του πληθυσμού που βρίσκονται σε κίνδυνο λίγο πριν τη χρονική στιγμή t_i. Αυτή η συνάρτηση πιθανοφάνειας ορίζεται ως συνάρτηση μερικής πιθανοφάνειας.

Ο λογάριθμος της μερικής συνάρτησης πιθανοφάνειας εκφράζεται ως εξής:

$$l(\beta) = \ln(L(\beta)) = \sum_{i=1}^{n}\delta_i\left(\beta' Q_i - \ln\left(\sum_{j\in R(t_i)}\exp(\beta' Q_j)\right)\right). \tag{26}$$

Οι εκτιμήσεις των συντελεστών προκύπτουν μεγιστοποιώντας την παραπάνω συνάρτηση χρησιμοποιώντας επαναληπτικές – αριθμητικές μεθόδους, όπως την επαναληπτική μέθοδο Newton – Raphson.

Σε αυτό το σημείο θα αναφέρουμε πως, στην περίπτωση που θέλουμε να ελέγξουμε τη μηδενική, το οποίο σημαίνει πως ένας συντελεστής του μοντέλου είναι ίσος με μηδέν, δηλαδή $H_0 : \beta_i = 0$, με εναλλακτική ότι $H_1 : \beta_i \neq 0$

για κάθε συντελεστή χωριστά, χρησιμοποιούμε το Wald test. Το στατιστικό ελέγχου δίνεται για αυτή την μηδενική υπόθεση από την παρακάτω σχέση:

$$T = \left(\frac{\hat{\beta}_j - \beta_0}{se(\hat{\beta}_j)} \right)^2 \tag{27}$$

όπου, $se(\hat{\beta}_j)$ αποτελεί το τυπικό σφάλμα της εκτίμησης του συντελεστή β_j. Το στατιστικό αυτό ακολουθεί X^2 κατανομή με 1 βαθμό ελευθερίας.

Το Wald test μπορεί να γενικευθεί στην περίπτωση που η παράμετρος β που μας ενδιαφέρει είναι διάνυσμα διαστάσεως p.

Για δύο διανύσματα τιμών των επεξηγηματικών μεταβλητών $\mathbf{X}$ και $\mathbf{X}^*$ έχουμε:

$$\frac{h(t \mid \mathbf{X})}{h(t \mid \mathbf{X}^*)} = \frac{h_0(t) \exp\left(\sum_{i=1}^{p} \beta_i X_i\right)}{h_0(t) \exp\left(\sum_{i=1}^{p} \beta_i X_i^*\right)} = \exp\left(\sum_{i=1}^{p} \beta_i (X_i - X_i^*) \right) = \gamma. \tag{28}$$

Η σχέση αυτή αποτελεί μία σταθερά αναλογίας, έστω γ και παρατηρείται ότι είναι μία σχέση ανεξάρτητη από τον χρόνο t. Συνεπώς, ο λόγος των ρυθμών κινδύνου δύο διαφορετικών ατόμων είναι σταθερός για γνωστές τιμές των επεξηγηματικών μεταβλητών και εκφράζει το σχετικό κίνδυνο θανάτου ενός ατόμου με παράγοντες κινδύνου $\mathbf{X}$ συγκριτικά με ένα άτομο με παράγοντες κινδύνου $\mathbf{X}^*$. Υπάρχει λοιπόν μία σχέση αναλογίας μεταξύ των δύο συναρτήσεων κινδύνου για τα δύο διανύσματα τιμών των επεξηγηματικών μεταβλητών και γι' αυτό τον λόγο το μοντέλο παλινδρόμησης του Cox ονομάζεται και μοντέλο αναλόγων κινδύνων.

Το μοντέλο έχει στόχο να εξετάσει την μορφή της συνάρτησης κινδύνου, καθώς και τις επιδράσεις των επεξηγηματικών μεταβλητών στην συνάρτηση κινδύνου, άρα και στον χρόνο επιβίωσης. Η συνάρτηση κινδύνου $\ln\{h(t \mid \mathbf{X})\}$ συνδέεται γραμμικά με τις επεξηγηματικές μεταβλητές $X_1, ...X_p$. Συνεπώς, οι συντελεστές του μοντέλου, στην περίπτωση ποσοτικών επεξηγηματικών μεταβλητών, μπορούν να ερμηνευθούν ως,

$$\begin{aligned} \ln h(t \mid \mathbf{X}) &= \ln\left(h_0(t) g(\mathbf{X})\right) \\ &= \ln\left(h_0(t) \exp(\beta_1 X_1 + \cdots + \beta_p X_p)\right) \\ &= \ln h_0(t) + \beta_1 X_1 + \cdots + \beta_p X_p \end{aligned} \tag{29}$$

$$\ln\left(\frac{h(t \mid \mathbf{X})}{h_0(t)}\right) = \beta_1 X_1 + \cdots + \beta_p X_p \tag{30}$$

Υποθέτουμε ότι για το μοντέλο του Cox, οι επεξηγηματικές μεταβλητές επιδρούν προσθετικά στην συνάρτηση κινδύνου, το οποίο σημαίνει πως δεν υπάρχουν αλληλεπιδράσεις μεταξύ των μεταβλητών και ότι η συνάρτηση κινδύνου συνδέεται γραμμικά με τους συντελεστές του μοντέλου $\beta_1, ..., \beta_p$. Το μοντέλο του Cox δεν υποθέτει μία ιδιαίτερη κατανομή για τους χρόνους επιβίωσης, αλλά υποθέτει ότι οι επιδράσεις διαφορετικών μεταβλητών στην επιβίωση είναι σταθερές ως προς τον χρόνο.

Σε αυτό το σημείο θα μπορούσαμε να δώσουμε μία ερμηνεία για τους συντελεστές του μοντέλου. Αρχικά θα υποθέσουμε ότι το μοντέλο αναλογικών κινδύνων περιλαμβάνει μία συνεχή επεξηγηματική μεταβλητή, έστω $\mathbf{X}$. Τότε, το μοντέλο δίνεται από τη σχέση,

$$h(t \mid \mathbf{X}) = h_0(t)e^{\beta \mathbf{X}}, \tag{31}$$

όπου ο λογάριθμος του λόγου συναρτήσεων κινδύνου, μπορεί να γραφεί γραμμικά,

$$\ln\left(\frac{h(t \mid \mathbf{X})}{h_0(t)}\right) = \beta \mathbf{X} \tag{32}$$

Έστω ότι θέλουμε να συγκρίνουμε τις συναρτήσεις κινδύνου για δύο μονάδες του πληθυσμού, όπου η μία λαμβάνει την τιμή x για την επεξηγηματική μεταβλητή X και η άλλη την τιμή $x + 1$. Τότε,

$$\frac{h(t \mid x+1)}{h(t \mid x)} = \frac{h_0(t)e^{\beta(x+1)}}{h_0(t)e^{\beta x}} = \frac{e^{\beta(x+1)}}{e^{\beta x}} = e^{\beta} \tag{33}$$

Συνεπώς, η εκτίμηση της παραμέτρου β είναι η μεταβολή του λογαρίθμου του λόγου της συνάρτησης κινδύνου, όταν η επεξηγηματική μεταβλητή αυξηθεί κατά μία μονάδα. Αν υποθέσουμε ότι το μοντέλο περιλαμβάνει μία επεξηγηματική μεταβλητή, η οποία είναι κατηγορική, έχουμε δύο περιπτώσεις $X = 0$ ή $X = 1$. Τότε η συνάρτηση κινδύνου αναφέρεται στους ασθενείς, που λαμβάνουν την τιμή 1 για την επεξηγηματική μεταβλητή X. Επίσης, το $exp(\beta)$ είναι η μεταβολή στην τιμή της συνάρτησης κινδύνου ενός ασθενή με τιμή 1 για την επεξηγηματική μεταβλητή X σε σχέση με έναν ασθενή με τιμή 0 για την επεξηγηματική μεταβλητή X.

Τώρα σημαντικό είναι να αναφερθεί, πως για να ελέγξουμε αν το μοντέλο προσαρμόζεται στα δεδομένα μας, δηλαδή ότι περιγράφει στον μεγαλύτερο δυνατό βαθμό τα δεδομένα μας, ένας τρόπος είναι η χρήση της στατιστική συνάρτηση Deviance. Η συνάρτηση αυτή στηρίζεται στον έλεγχο του λόγου των πιθανοφανειών δύο επιλεγμένων μοντέλων. Δηλαδή, ελέγχεται η μηδενική υπόθεση, $H_0 : \beta_h = 0$ και εναλλακτική την $H_1 : \beta_h \neq 0$, το οποίο στην ουσία αποτελεί μια σύγκριση μεταξύ του μοντέλου που προσαρμόζουμε και του ιδανικού μοντέλου.

Στην ουσία γίνεται σύγκριση της μεγιστοποιημένης πιθανοφάνειας υπό το μοντέλο της επιλογής μας με την μεγιστοποιημένη πιθανοφάνεια για το καλύτερο μοντέλο. Το saturated μοντέλο είναι ουσιαστικά το καλύτερο μοντέλο που θα μπορούσαμε να πάρουμε από τα δεδομένα μας. Είναι ένα μοντέλο με αριθμό επεξηγηματικών μεταβλητών ίσο με το μέγεθος του δείγματος. Άρα, το μοντέλο αυτό αποτελεί το μοντέλο με την καλύτερη προσαρμογή στα δεδομένα και συγκρίνοντας το με το μοντέλο της επιλογής μας λαμβάνουμε μία εικόνα του πόσο καλά περιγράφει το μοντέλο μας τα δεδομένα μας. Για το μοντέλο του Cox χρησιμοποιώντας τον λόγο των μερικών συναρτήσεων Πιθανοφάνειας λαμβάνουμε ότι,

$$Deviance = -2 \ln \left(\frac{L_{fitted}}{L_{saturated}} \right).$$

Βλέπουμε πως η Deviance ακολουθεί τη κατανομή X^2 και έχει βαθμούς ελευθερίας ίσους με τη διαφορά των παραμέτρων των δύο μοντέλων, $d_{satured} - d_{fitted}$, όπου το $d_{satured}$ είναι ο αριθμός των παραμέτρων του μοντέλου και d_{fitted} ο αριθμός των παραμέτρων του υπό μελέτη μοντέλου. Με τη χρήση της ελεγχοσυνάρτησης μπορεί να γίνει επίσης σύγκριση δύο εμφωλευμένων μοντέλων, όπου η ελεγχοσυνάρτηση $Deviance_1 - Deviance_2$ ακολουθεί την κατανομή X^2 με βαθμούς ελευθερίας ίσους με τον αριθμό της διαφοράς των παραμέτρων των δύο μοντέλων.

1.7 Κριτήρια επιλογής μοντέλου

Έχουν αναπτυχθεί αριθμητικές ποσότητες, με στόχο την αξιολόγηση μοντέλων, ώστε να γίνεται η βέλτιστη επιλογή, ανάμεσα σε άλλα μοντέλα. Για την επιλογή του κατάλληλου μοντέλου, αλλά και για την σύγκρισή τους χρησιμοποιούνται τα μέτρα καταλληλότητας. Κάποια από αυτά αποτελούν τα κριτήρια AIC και BIC.

1.7.1 Akaike's information criterion (AIC)

Το μοντέλο AIC (Akaike, H. (1974)) είναι ένα μέτρο που χρησιμοποιείται για σύγκριση και επιλογή μοντέλων. Βασίζεται στην έννοια της εντροπίας, με στόχο να βρεθεί το μοντέλο που εξηγεί καλύτερα τα δεδομένα με τον ελάχιστο αριθμό παραμέτρων. Ορίζεται από τη σχέση,

$$AIC = 2k - 2logL,$$

όπου k είναι ο αριθμός των παραμέτρων στο μοντέλο και L είναι η μέγιστη τιμή της συνάρτησης πιθανοφάνειας για δεδομένο μοντέλο.

Το AIC χρησιμοποιείται κυρίως για τη σύγκριση διαφορετικών μοντέλων που εφαρμόζονται στο ίδιο σύνολο δεδομένων. Το μοντέλο με το χαμηλότερο AIC θεωρείται το καλύτερο μεταξύ των συγκριτικών μοντέλων. Το AIC εξισορροπεί την αντιστάθμιση μεταξύ της καλής προσαρμογής του μοντέλου και της πολυπλοκότητας του μοντέλου. Απορρίπτονται τα μοντέλα με περισσότερες παραμέτρους για να αποφευχθεί η υπερβολική προσαρμογή και αυτό γίνεται με τη ποσότητα $2d$ που καλείται ποινή (penalty). Αυτό σημαίνει ότι η προσθήκη παραμέτρων σε ένα μοντέλο θα ευνοηθεί μόνο εάν βελτιώνει σημαντικά την προσαρμογή του μοντέλου. Οι τιμές AIC είναι σχετικές, που σημαίνει ότι είναι χρήσιμες μόνο όταν γίνεται σύγκριση μεταξύ μοντέλων. Η ίδια η απόλυτη τιμή του AIC δεν παρέχει ουσιαστικές πληροφορίες.

1.7.2 Bayesian information criterion (BIC)

Το BIC (Schwarz, G. (1978)) χρησιμοποιείται για την επιλογή μοντέλου μεταξύ ενός πεπερασμένου συνόλου μοντέλων, όπου προτιμάται το μοντέλο με το χαμηλότερο BIC. Εισάγει έναν όρο ποινής για τον αριθμό των παραμέτρων, ο οποίος είναι μεγαλύτερος από αυτόν του AIC, καθιστώντας το πιο συντηρητικό. Ορίζεται από τη σχέση,

$$BIC = kln(n) - 2ln(L),$$

όπου k είναι ο αριθμός των παραμέτρων του μοντέλου, n είναι ο αριθμός των παρατηρήσεων του συνόλου δεδομένων και το L αποτελεί, όπως και πριν, τη μέγιστη τιμή της συνάρτησης πιθανοφάνειας για το μοντέλο. Η ποινή για τον αριθμό των παραμέτρων στο BIC είναι $ln(n)$, η οποία αυξάνεται με το μέγεθος του δείγματος, καθιστώντας το BIC πιο αυστηρό κριτήριο ως προς την προσθήκη παραμέτρων από το AIC.

1.8 Μοντέλα επιταχυνόμενου χρόνου αποτυχίας

Το μοντέλο επιταχυνόμενου χρόνου αποτυχίας (AFT) (Buckley, J., & James, I. (1979)) είναι ένα παραμετρικό μοντέλο επιβίωσης που χρησιμοποιείται για την ανάλυση δεδομένων που αφορούν το χρόνο μέχρι το συμβάν. Σε αντίθεση με το μοντέλο αναλογικών κινδύνων Cox, το οποίο υποθέτει ότι η συνάρτηση αναφοράς της συνάρτησης κινδύνου $h_0(\cdot)$ είναι ανεξάρτητη των τιμών των επεξηγηματικών μεταβλητών, το μοντέλο (AFT) υποθέτει ότι οι επεξηγηματικές μεταβλητές επηρεάζουν πολλαπλασιαστικά το χρόνο συμβάντος. Η συνάρτηση κινδύνου του μοντέλου, ορίζεται ως,

$$\mu(t \mid X) = \mu_0(te^{\beta' X})e^{\beta' X}, \tag{34}$$

όπου $\mu_0(\cdot)$ η βασική συνάρτηση κινδύνου.

Τα μοντέλα (AFT) χρησιμοποιούνται κυρίως με βάση παραμετρικές προσεγγίσεις με Λογαριθμοκανονική, Γάμμα και αντίστροφους γκαουσιανούς βασικούς κινδύνους. Σε αντίθεση με το μοντέλο Cox, μπορεί να χαρακτηριστεί και να ερμηνευτεί καλύτερα από την άποψη της συνάρτηση επιβίωσης. Υποθέτοντας ένα μοντέλο με μία μόνο δυαδική μεταβλητή X (για παράδειγμα, που υποδεικνύει τη θεραπεία ($X = 1$) και την ομάδα ελέγχου ($X = 0$), σε μία τυχαιοποιημένη κλινική δοκιμή) για τη συνάρτηση επιβίωσης, η σχέση

$$S(t \mid X = 1) = S(e^{\beta}t \mid X = 0) \tag{35}$$

ισχύει. Η συνάρτηση κινδύνου (34) ενός ατόμου στην ομάδα ελέγχου είναι τότε ο βασικός κίνδυνος $\mu(t|0) = \mu(t)$. Ο παράγοντας e^{β}, ονομάζεται παράγοντας επιτάχυνσης, που σημαίνει ότι η πιθανότητα επιβίωσης στο χρονικό σημείο t, στην ομάδα θεραπείας, είναι παρόμοια με την πιθανότητα επιβίωσης στο χρονικό σημείο te^{β} στην ομάδα ελέγχου. Αυτό σημαίνει, για παράδειγμα, ότι ο διάμεσος χρόνος επιβίωσης ενός ασθενούς στην ομάδα θεραπείας είναι e^{β} φορές μεγαλύτερος από εκείνον ενός ασθενούς στην ομάδα ελέγχου. Συνεπώς, κάποια άτομα μεγαλώνουν e^{β} φορές γρηγορότερα, με κάποια βιολογική έννοια, στην ομάδα ελέγχου σε σύγκριση με την ομάδα θεραπείας, γεγονός που επιτρέπει μια πολύ απλή και φυσική ερμηνεία των παραμέτρων του μοντέλου.

2 Μοντέλα Ευπάθειας

Ο χρόνος επιβίωσης, διαφορετικών ατόμων, χρησιμοποιείται ως μεταβλητή απόκρισης, για τα μοντέλα παλινδρόμησης που αφορούν δεδομένα διάρκειας ζωής, με σκοπό την εκτίμηση της συνάρτησης κινδύνου, και κατά συνέπεια της συνάρτησης επιβίωσης. Ως επεξηγηματικές μεταβλητές χρησιμοποιούνται διάφοροι παράγοντες που θεωρούμε ότι επηρεάζουν το χρόνο επιβίωσης. Παρόλα αυτά όμως, υπάρχουν και άλλες μεταβλητές που μπορεί να επηρεάζουν το μοντέλο μας, ως προς το χρόνο επιβίωσης, οι οποίες μπορεί να είναι δύσκολο να συμπεριληφθούν στην ανάλυσή μας, καθώς μπορεί να αφορούν παράγοντες που μπορεί να μην γνωρίζουμε. Για το λόγο αυτό έχουν αναπτυχθεί δύο πηγές μεταβλητότητας στα δεδομένα μας, η υπολογίσιμη μεταβλητότητα από παρατηρούμενους παράγοντες και η μεταβλητότητα, η οποία δεν μπορεί να υπολογιστεί αφού υπάρχουν αστάθμητοι παράγοντες.

Οι άγνωστοι παράγοντες μπορούν να συμπεριληφθούν στο μοντέλο μέσω μίας λανθάνουσας μεταβλητής (latent variable), η οποία αποτελεί τον παράγοντα τυχαίων επιδράσεων και καλείται ευπάθεια (frailty) στην ανάλυση επιβίωσης και συνεπώς τα μοντέλα αυτά ονομάζονται μοντέλα ευπάθειας (frailty models). Τον όρο ευπάθεια εισήγαγαν οι Vaupel et al. (1979) προκειμένου να χαρακτηρίσουν αυτά τα μοντέλα τυχαίων επιδράσεων. Ο όρος αυτός χρησιμοποιήθηκε για να προσδιορίσει ότι διαφορετικές μονάδες του πληθυσμού μπορεί να βρίσκονται σε κίνδυνο όταν εκτίθενται σε διάφορους παράγοντες ακόμα και αν φαινομενικά εμφανίζουν ίδια χαρακτηριστικά, όπως ηλικία, φύλο, βάρος κ.λ.π..

Για παράδειγμα, σε μία έρευνα που εξετάζει δύο θεραπείες για την αντιμετώπιση του καρκίνου συλλέγουμε δεδομένα για κάποια χαρακτηριστικά των ασθενών, όπως φύλο, ηλικία κ.λ.π. και προσαρμόζουμε το μοντέλο αναλόγων κινδύνων του Cox. Έτσι, για τους ασθενείς ίδιας ηλικίας και φύλου, που λαμβάνουν την ίδια θεραπεία, η κατανομή του χρόνου επιβίωσης θα είναι ίδια. Ωστόσο, η υπόθεση αυτή για μελέτες σε ζωντανούς οργανισμούς δεν είναι απαραίτητα σωστή, καθώς σίγουρα θα υπάρχουν πολλοί παράγοντες που επηρεάζουν τον χρόνο επιβίωσης αυτών και τις περισσότερες φορές δεν είναι γνωστοί για να συμπεριληφθούν στο μοντέλο.

Η γενίκευση του μοντέλου αναλόγων κινδύνων του Cox είναι το πιο ευρέως διαδεδομένο μοντέλο παλινδρόμησης ευπάθειας, εισάγοντας τον παράγοντα των τυχαίων επιδράσεων (random effects) ως μία τυχαία μεταβλητή για τον χειρισμό της αλληλεπίδρασης μεταξύ των μεταβλητών, που δεν έχει ληφθεί υπόψη στο μοντέλο, αλλά και της μη παρατηρούμενης ανομοιογένειας που χαρακτηρίζει τον πληθυσμό που εξετάζουμε. Στα μονοδιάστατα μοντέλα επιβίωσης (univariate survival models), ένα μοντέλο ευπάθειας μπορεί να χρησιμοποιηθεί για την εκτίμηση της ανομοιογένειας μεταξύ των διαφορετικών μονάδων του πληθυσμού. Στα πολυδιάστατα (multivariate survival models), χρησιμοποιούνται τα από κοινού μοντέλα ευπάθειας (shared frailty models). Το από κοινού μοντέλο ευπάθειας είναι ένα μοντέλο τυχαίων επιδράσεων, το οποίο υποθέτει μεταβλητότητα μεταξύ διαφορετικών ομάδων (frailty) και μεταβλητότητα μεταξύ των μονάδων του πληθυσμού, η οποία εξηγείται από την συνάρτηση κινδύνου (hazard function).

Για την μοντελοποίηση του χρόνου επιβίωσης μπορούν να χρησιμοποιηθούν και παραμετρικά, αλλά και ημιπαραμετρικά μοντέλα παλινδρόμησης. Στα παραμετρικά μοντέλα γίνεται η υπόθεση ότι έχουμε μία γνωστή κατανομή για την βασική συνάρτηση κινδύνου (baseline hazard function) και εκτιμούμε τις παραμέτρους αυτής χρησιμοποιώντας τα δεδομένα, ενώ στην περίπτωση των ημιπαραμετρικών μοντέλων δεν γίνεται καμία υπόθεση για την κατανομή της βασικής συνάρτησης κινδύνου και η εκτίμηση αυτής γίνεται με άλλες τεχνικές, όπως ο αλγόριθμος EM.

2.1 Παραμετροποίηση μοντέλου ευπάθειας

Θεωρούμε T μία τυχαία μεταβλητή ως το χρόνο επιβίωσης, η οποία ακολουθεί μία συνεχή κατανομή. Στη συνέχεια ορίζουμε επίσης μία μη αρνητική τυχαία μεταβλητή Z, που θα καλείται ευπάθεια (frailty) (Vaupel et al., 1979) αν η δεσμευμένη συνάρτηση κινδύνου δεδομένου του Z έχει την μορφή,

$$h(t \mid Z) = Zh_0(t), \tag{36}$$

όπου $h_0(t)$, η βασική συνάρτηση κινδύνου.

Στη συνέχεια, έχουμε τη δεσμευμένη συνάρτηση επιβίωσης,

$$S(t \mid Z) = e^{-ZH_0(t)} \tag{37}$$

με $H_0(t) = \int_0^t h_0(s)\, ds$ να είναι η αθροιστική βασική συνάρτηση κινδύνου για το επίπεδο αναφοράς. Μπορούμε επίσης να βρούμε τη συνάρτηση επιβίωσης

$S(t)$, η οποία προκύπτει ολοκληρώνοντας τη δεσμευμένη συνάρτηση επιβίωσης $S(t \mid Z)$ ως προς Z, υπολογίζοντας δηλαδή την αναμενόμενη τιμή,

$$S(t) = \int_0^\infty \exp(-zH_0(t)) f_Z(z)\, dz = \mathbb{E}[S(t \mid Z)] = \mathbb{E}[e^{-ZH_0(t)}], \tag{38}$$

όπου $f_Z(z)$ η συνάρτηση πυκνότητας πιθανότητας της ευπάθειας Z.

Σημαντικό είναι να αναφέρουμε πως ιδιαίτερη έμφαση δίνεται στον μετασχηματισμό Laplace της ευπάθειας, επειδή η συνάρτηση επιβίωσης και η συνάρτηση κινδύνου μπορούν εύκολα να εκφραστούν με τη χρήση αυτού του μετασχηματισμού. Ως εκ τούτου, η συνάρτηση πιθανότητας μπορεί επίσης να εκφραστεί μέσω του μετασχηματισμού Laplace.

Η κ-οστή παράγωγος του μετασχηματισμού Laplace της ευπάθειας δίνεται από την σχέση,

$$L^{(k)}(s) = (-1)^k \mathbb{E}\left[Z^k \exp(-Zs)\right] \tag{39}$$

όπου $s = H_0(t)$.

Επιπλέον,

$$\begin{aligned} f(t) &= (-1)\frac{dS(t)}{dt} \\ &= \int_0^\infty \frac{d}{dt}\left[\exp(-zH_0(t)) f_Z(z)\right] dz \\ &= -\int_0^\infty z \cdot h_0(t) \exp(-zH_0(t)) f_Z(z) dz. \end{aligned}$$

το οποίο έχει ως αποτέλεσμα τη συνάρτηση πυκνότητας πιθανότητας,

$$f(t) = -h_0(t) L'(H_0(t)), \quad \text{όπου } L' = L^{(1)} \tag{40}$$

Αφού η συνάρτηση κινδύνου δίνεται από τη σχέση $h(t) = f(t)/S(t)$ και επειδή έχουμε σχέσεις (39) και (40) έχουμε ότι,

$$h(t) = \frac{f(t)}{S(t)} = -\frac{h_0(t) L'(H_0(t))}{L(H_0(t))} = -h_0(t)\frac{L'(H_0(t))}{L(H_0(t))}. \tag{41}$$

Η μέση τιμή και η διασπορά της τυχαίας μεταβλητής ευπάθειας Z,

$$E(Z) = -L'(0), \tag{42}$$

και επειδή,

$$L''(0) = (-1)^2 E(Z^2) = E(Z^2), \tag{43}$$

Τότε έχουμε ότι,

$$V(Z) = E(Z^2) - (E(Z))^2, \tag{44}$$

και

$$V(Z) = L''(0) - (L'(0))^2. \tag{45}$$

2.2 Μονομεταβλητά Μοντέλα Ευπάθειας

Στη συνέχεια θα εισαχθούν παρατηρούμενες συμμεταβλητές στο μοντέλο παλινδρόμησης ευπάθειας ή μοντέλο τυχαίων επιδράσεων για δεδομένα επιβίωσης παρόμοιο με το Cox μοντέλο, το οποίο ορίζεται ως,

$$h(t \mid X, Z) = Z h_0(t) e^{\beta' x} \tag{46}$$

όπου $h(t \mid X, Z)$ είναι ο κίνδυνος θανάτου ενός ασθενούς την χρονική στιγμή t, εξαιτίας των επιδράσεων του χρόνου επιβίωσής του. Επιπλέον, $h_0(t)$ είναι το επίπεδο αναφοράς της συνάρτησης κινδύνου για έναν ασθενή και αντιστοιχεί στην πιθανότητα θανάτου ενός ασθενή τη χρονική στιγμή t όταν όλες οι επεξηγηματικές μεταβλητές λαμβάνουν την τιμή 0. Να αναφέρουμε πως η ευπάθεια λειτουργεί πολλαπλασιαστικά στην συνάρτηση $h(t \mid X, Z)$ και η τυχαία μεταβλητή ευπάθειας Z είναι μη αρνητική. Η Z στην ουσία εκφράζει τη μη παρατηρούμενη πληροφορία, που επηρεάζει τη συνάρτηση κινδύνου και λαμβάνει διαφορετική τιμή για κάθε μονάδα του πληθυσμού. Αν $0 < Z < 1$ τότε ο ασθενής είναι λιγότερο ευπαθής, δηλαδή η πιθανότητα να επέλθει το υπό μελέτη γεγονός για αυτόν τον ασθενή μειώνεται, ενώ για $Z > 1$ ο ασθενής είναι περισσότερο ευπαθής από το μέσο ασθενή, δηλαδή το επίπεδο αναφοράς.

Σημαντικό είναι επίσης να σημειωθεί ότι $X = (X_1, ..., X_k)$ και $\beta = (\beta_1, ..., \beta_k)$ αποτελούν τις συμμεταβλητές και τις παραμέτρους παλινδρόμησης, αντίστοιχα. Κατά συνέπεια, ένα μοντέλο ευπάθειας είναι μια γενίκευση του μοντέλου αναλογικών κινδύνων. Το μοντέλο αναλογικών κινδύνων προκύπτει εάν η $Z = 1$ για όλα τα άτομα.

Συνεπώς η συνάρτηση πιθανοφάνειας για το μοντέλο ευπάθειας μπορεί να γραφεί λόγω της (46) ως,

$$L(\beta) = \prod_{i=1}^{n} \left[Z_i h_0(t_i) \exp(\beta' X_i) \right]^{\delta_i} \exp\left(-Z_i H_0(t_i) \exp(\beta' X_i) \right) \tag{47}$$

Η παραπάνω συνάρτηση πιθανοφάνειας αποτελεί τη βάση για την εκτίμηση των παραμέτρων του μοντέλου.

Η ευπάθεια σαν τυχαία μεταβλητή μπορεί να ακολουθεί διάφορες κατανομές, οι οποίες χρησιμοποιούνται ευρέως σε πολλές εφαρμογές και θα αναλύσουμε παρακάτω κάποιες πρακτικές εφαρμογές.

2.2.1 Γάμμα μοντέλο ευπάθειας

Η κατανομή Γάμμα $\Gamma(\kappa, \lambda)$, ως μια γενίκευση της εκθετικής κατανομής ταιριάζει πολύ καλά σε δεδομένα ευπάθειας. Αποτελεί μια ευέλικτη κατανομή καθώς δέχεται διαφορετικές τιμές για τη παράμετρο k, όπου για $k = 1$, είναι πανομοιότυπη με την εκθετική, για μεγάλο k, παίρνει μια μορφή καμπάνας που θυμίζει την κανονική κατανομή.

Οι συναρτήσεις που εξετάζονται στην ανάλυση επιβίωσης είναι η συνάρτηση πυκνότητας πιθανότητας,

$$f(z) = \frac{1}{\Gamma(k)} \lambda^k z^{k-1} e^{-\lambda z} \tag{48}$$

και ύστερα από τη χρήση του μετασχηματισμού Laplace λαμβάνουμε ότι:

$$L(s) = \frac{\lambda^k}{\Gamma(k)} \int_0^\infty e^{-uz} z^{k-1} e^{-\lambda z} \, dz = \left(1 + \frac{u}{\lambda}\right)^{-k} \tag{49}$$

Η πρώτη και η δεύτερη παράγωγος του μετασχηματισμού Laplace είναι,

$$L'(u) = -\frac{k}{\lambda} \left(1 + \frac{u}{\lambda}\right)^{-k-1} \tag{50}$$

$$L''(u) = \frac{k(k+1)}{\lambda^2} \left(1 + \frac{u}{\lambda}\right)^{-k-2} \tag{51}$$

Για τις παραγώγους, για $u = 0$ έχουμε,

$$E(Z) = \frac{k}{\lambda}$$

$$V(Z) = \frac{k(k+1)}{\lambda^2} - \frac{k^2}{\lambda^2} = \frac{k}{\lambda^2}$$

Για να είναι το μοντέλο καλώς ορισμένο, τότε θα πρέπει να κάνουμε την παραδοχή για την Γάμμα κατανομή, ότι $\kappa = \lambda$, το οποίο οδηγεί στο συμπέρασμα ότι $E(Z) = 1$ και η διασπορά $V(Z) = \frac{1}{\lambda}$.

Αν θεωρήσουμε ότι η διασπορά είναι η άγνωστη παράμετρος για την κατανομή Γάμμα με $\frac{1}{\sigma^2} = \lambda$, λαμβάνουμε για την συνάρτηση επιβίωσης,

$$S(t) = L(M_0(t)) = \frac{1}{(1 + \sigma^2 M_0(t))^{1/\sigma^2}} \tag{52}$$

την συνάρτηση κινδύνου,

$$f(t) = \frac{\mu_0(t)}{(1 + \sigma^2 M_0(t))^{1/\sigma^2+1}} \tag{53}$$

καθώς και την συνάρτηση πυκνότητας πιθανότητας,

$$\mu(t) = \frac{\mu_0(t)}{1 + \sigma^2 M_0(t)} \tag{54}$$

Για λόγους ευκολίας, η ευπάθεια θεωρείται σταθερή με την πάροδο του χρόνου για κάθε άτομο. Όμως, η κατανομή της ευπάθειας, στον πληθυσμό που εξακολουθεί να διατρέχει κίνδυνο, αλλάζει με την πάροδο του χρόνου. Στην συνέχεια επεκτείνουμε το μοντέλο συμπεριλαμβάνοντας όρους από το μοντέλο του Cox, με αντικατάσταση της αθροιστικής συνάρτησης κινδύνου $M_0(t)$ με $M_0(t)e^{\beta' X}$. Έτσι λαμβάνουμε κάποια αποτελέσματα χρήσιμα για την εκτίμηση των παραμέτρων του μοντέλου, ιδιαίτερα στην περίπτωση ενός ημιπαραμετρικού Γάμμα μοντέλου παλινδρόμησης ευπάθειας.

Η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής Z δεδομένου των επεξηγηματικών μεταβλητών του μοντέλου για αυτούς που έχουν επιβιώσει μέχρι τη χρονική στιγμή t για $T > t$,

$$\begin{aligned} f(z \mid \mathbf{X}, T > t) &= \frac{S(t \mid \mathbf{X}, z) f(z)}{S(t \mid \mathbf{X})} \\ &= \frac{\exp\left(-z M_0(t) e^{\beta' \mathbf{X}}\right) z^{\frac{1}{\sigma^2}-1} \exp\left(-\frac{z}{\sigma^2}\right)}{\Gamma\left(\frac{1}{\sigma^2}\right) \sigma^{\frac{2}{\sigma^2}} \left(1 + \sigma^2 M_0(t) e^{\beta' \mathbf{X}}\right)^{-\frac{1}{\sigma^2}}} \\ &= \frac{\left(\frac{1}{\sigma^2} + M_0(t) e^{\beta' \mathbf{X}}\right)^{\frac{1}{\sigma^2}} z^{\frac{1}{\sigma^2}-1} \exp\left(-z\left(\frac{1}{\sigma^2} + M_0(t) e^{\beta' \mathbf{X}}\right)\right)}{\Gamma\left(\frac{1}{\sigma^2}\right)}. \end{aligned} \tag{55}$$

με παραμέτρους, $\frac{1}{\sigma^2} = \lambda$ και $\frac{1}{\sigma^2} + M_0(t) e^{\beta' X} = k$.

Για $T = t$, δηλαδή για αυτούς που έχουν χρόνο επιβίωσης t, η συνάρτηση πυκνότητας πιθανότητας της Z δεδομένου των επεξηγηματικών μεταβλητών δίνεται από τη σχέση,

$$
\begin{aligned}
f(z \mid \mathbf{X}, T = t) &= \frac{f(t \mid \mathbf{X}, z) f(z)}{f(t \mid \mathbf{X})} \\
&= z\mu_0(t) \exp\left(-zM_0(t)e^{\beta' \mathbf{X}}\right) z^{\frac{1}{\sigma^2}-1} \exp\left(-\frac{z}{\sigma^2}\right) \\
&= \frac{\left(\frac{1}{\sigma^2}\right)^2 \mu_0(t)(1+\sigma^2 M_0(t)e^{\beta' \mathbf{X}}) z^{\frac{1}{\sigma^2}-1}}{\Gamma\left(\frac{1}{\sigma^2}\right)\sigma^2} \\
&= \frac{(1+M_0(t)e^{\beta' \mathbf{X}})^{\frac{1}{\sigma^2}+1} z^{\frac{1}{\sigma^2}-1} \exp\left(-z\left(\frac{1}{\sigma^2}+M_0(t)e^{\beta' \mathbf{X}}\right)\right)}{\Gamma\left(\frac{1}{\sigma^2}+1\right)}.
\end{aligned} \tag{56}
$$

η οποία αποτελεί Γάμμα κατανομή με παραμέτρους $\frac{1}{\sigma^2}+1=\lambda$ και $\frac{1}{\sigma^2}+M_0(t)e^{\beta' X}=k$.

Ειδικότερα, προκύπτει ότι η μέση ευπάθεια μεταξύ των θανάτων στην ηλικία t είναι,

$$
\mathbb{E}(Z \mid \mathbf{X}, T = t) = \frac{1+\sigma^2}{1+\sigma^2 M_0(t)e^{\beta' \mathbf{X}}}, \tag{57}
$$

σε σχέση με,

$$
\mathbb{E}(Z \mid \mathbf{X}, T > t) = \frac{1}{1+\sigma^2 M_0(t)e^{\beta' \mathbf{X}}}. \tag{58}
$$

μεταξύ των επιζώντων στην ίδια ηλικία.

Τα άτομα που πεθαίνουν τη χρονική στιγμή t έχουν υψηλότερη μέση ευπάθεια σε σύγκριση με τους επιζώντες αυτής της χρονικής στιγμής. Επιπλέον, ισχύει ότι η διακύμανση της ευπάθειας μεταξύ των ατόμων που πεθαίνουν τη χρονική στιγμή t είναι,

$$
V(Z \mid \mathbf{X}, T = t) = \frac{\sigma^2(1+\sigma^2)}{(1+\sigma^2 M_0(t)e^{\beta' \mathbf{X}})^2} \tag{59}
$$

και

$$
V(Z \mid \mathbf{X}, T > t) = \frac{\sigma^2}{(1+\sigma^2 M_0(t)e^{\beta' \mathbf{X}})^2}. \tag{60}
$$

μεταξύ των επιζώντων.

Κατά συνέπεια, η διακύμανση της ευπάθειας μειώνεται επίσης κατά τη διάρκεια, έτσι ο πληθυσμός της μελέτης γίνεται πιο ομοιογενής. Ωστόσο, ο

συντελεστής μεταβλητότητας παραμένει σταθερός με την πάροδο του χρόνου, οπότε ο πληθυσμός δεν γίνεται πιο ομοιογενής σε σχέση με τον μέσο όρο.

2.2.2 Γάμμα παραμετρικά μοντέλα ευπάθειας

Όπως και στο μοντέλο αναλογικών κινδύνων μπορούμε να χρησιμοποιήσουμε είτε παραμετρική προσέγγιση, είτε ημιπαραμετρική, ώστε να εκτιμήσουμε τις παραμέτρους του μοντέλου. Θα εξετάσουμε αρχικά την παραμετρική περίπτωση, όπου με βάση την σχέση (47), η συνάρτηση πιθανοφάνειας για το μοντέλο ευπάθειας στη γενική του μορφή δίνεται από τη σχέση,

$$L(\beta \mid Z_1, Z_2, \ldots, Z_n) = \prod_{i=1}^{n} \left[Z_i h_0(t_i, \theta) e^{\beta' X_i} \right]^{\delta_i} \exp\left(-Z_i H_0(t_i; \theta) e^{\beta' X_i} \right) \quad (61)$$

όπου θ είναι το άγνωστο διάνυσμα παραμέτρων που υπεισέρχεται στην βασική συνάρτηση κινδύνου. Υπό την υπόθεση ότι η ευπάθεια ακολουθεί Γάμμα κατανομή, μπορούμε να ολοκληρώσουμε ως προς την ευπάθεια και χρησιμοποιώντας τις σχέσεις (48) και (49), παίρνουμε την συνάρτηση πιθανοφάνειας,

$$L(\beta, \theta, \sigma^2) = \prod_{i=1}^{n} \left[\frac{h_0(t_i; \theta) \exp(\beta' X_i)}{(1 + \sigma^2 H_0(t_i; \theta) \exp(\beta' X_i))^{\frac{1}{\sigma^2}}} \right]^{\delta_i} \left(1 + \sigma^2 H_0(t_i; \theta) \exp(\beta' x_i)\right)^{-\frac{1}{\sigma^2}} \quad (62)$$

Από τις σχέσεις (57) και (58) η ευπάθεια για κάθε ασθενή $Z_i, i = 1, ..., n$ εκτιμάται από την αναμενόμενη τιμή,

$$\hat{Z}_i = \frac{\frac{1}{\hat{\sigma}^2} + \delta_i}{\frac{1}{\hat{\sigma}^2} + H_0(t_i; \hat{\theta}) \exp(\hat{\beta}' X_i)}, \quad (63)$$

όπου $\hat{\theta}$, αποτελεί την εκτίμηση του διανύσματος των παραμέτρων που εισέρχεται στην συνάρτηση κινδύνου και $\hat{\beta}$ είναι η εκτίμηση των παραμέτρων του μοντέλου. Επίσης, έχουμε ότι δ_i είναι ένας δείκτης που λαμβάνει την τιμή 1 αν έχουμε χρόνο επιβίωσης και την τιμή 0 αν έχουμε λογοκριμένο χρόνο, καθώς και $\hat{\sigma}^2$ είναι η εκτίμηση της διασποράς της μεταβλητής της ευπάθειας.

2.2.3 Γάμμα ημιπαραμετρικά μοντέλα ευπάθειας

Στο ημιπαραμετρικό μοντέλο ευπάθειας, δεν γίνεται καμία υπόθεση σχετικά με τη μορφή της βασικής συνάρτησης κινδύνου. Αυτό απαιτεί νέες στρατηγικές

εκτίμησης σε σύγκριση με το παραμετρικό μοντέλο. Προηγουμένως εξετάσαμε την επέκταση του παραμετρικού μοντέλου αναλογικών κινδύνων στο παραμετρικό μοντέλο Γάμμα της ευπάθειας. Στη συνέχεια εξετάζεται η επέκταση του ημιπαραμετρικού μοντέλου αναλογικών κινδύνων (Cox) στο Γάμμα ημιπαραμετρικό μοντέλο ευπάθειας.

Τώρα ο κίνδυνος $h_0(t)$ αντιμετωπίζεται ως οχληρά παράμετρος (nuisamce parameter) και θα χρησιμοποιηθεί ο αλγόριθμος EM (Expectation Maximization algorithm), ο οποίος επιτρέπει την εκτίμηση παραμέτρων στο Γάμμα ημιπαραμετρικό μοντέλο ευπάθειας. Ο αλγόριθμος αυτός αποτελείται από δύο βήματα, Expectation step και Maximization step. Στο πρώτο, εκτιμώνται οι αναμενόμενες τιμές για τις μη παρατηρούμενες μεταβλητές ευπάθειας δεδομένων των παρατηρήσεων και των εκτιμημένων παραμέτρων. Οι εκτιμήσεις αυτές των αναμενόμενων τιμών χρησιμοποιούνται στο δεύτερο βήμα μεγιστοποίησης ως πραγματική πληροφορία και υπολογίζονται νέες εκτιμήσεις για τις παραμέτρους μεγιστοποιώντας την συνάρτηση πιθανοφάνειας. Παρακάτω θα παρουσιάσουμε τον αλγόριθμο EM για το Γάμμα μοντέλο ευπάθειας.

Αρχικά εξετάζουμε την συνάρτηση πιθανοφάνειας όπου οι μεταβλητές ευπάθειας αποτελούν τυχαίες μεταβλητές Z_i. Η από κοινού συνάρτηση πιθανοφάνειας για το $(t_i, \delta_i, Z_i)(i = 1, ..., n)$ και έχει τη μορφή,

$$\begin{aligned} L(\beta, \sigma^2|\mathbf{Z}) = \\ \prod_{i=1}^{n} f(t_i, \delta_i, Z_i; \beta, \sigma^2) = \\ \prod_{i=1}^{n} f(t_i, \delta_i, \beta|Z_i) \prod_{i=1}^{n} f(Z_i; \sigma^2) = \\ L_1(\beta|\mathbf{Z}) L_2(\sigma^2|\mathbf{Z}), \end{aligned} \tag{64}$$

με $\mathbf{Z} = (Z_1, ..., Z_n)$ να αποτελεί το τυχαίο δείγμα των ευπαθειών και το $\beta = \beta_1, ..., \beta_p$ να είναι το διάνυσμα των αγνώστων παραμέτρων του μοντέλου ευπάθειας.

Έτσι έχουμε ότι ο πρώτος όρος είναι η συνάρτηση πιθανοφάνειας των παρατηρούμενων γεγονότων δεδομένων των ευπαθειών,

$$L_1(\beta|\mathbf{Z}) = \prod_{i=1}^{n} \left[Z_i h_0(t_i) \exp(\beta' X_i) \right]^{\delta_i} \exp\left(-Z_i H_0(t_i) \exp(\beta' X_i) \right), \tag{65}$$

ενώ ο δεύτερος όρος της σχέσης (64) προκύπτει από τη συνάρτηση πυκνότητας πιθανότητας της μεταβλητής ευπάθειας,

$$L_2(\sigma^2|\mathbf{Z}) = \prod_{i=1}^{n} f_Z(Z_i;\sigma^2) \tag{66}$$

Εάν οι μεταβλητές ευπάθειας Z_i ήταν γνωστές, οι παράμετροι παλινδρόμησης β θα μπορούσαν να εκτιμηθούν με τη μέθοδο της μερικής πιθανοφάνειας αντικαθιστώντας τους όρους $Z_i exp(\beta' X_i)$ της σχέσης L_1 με $exp(\beta' X_i)+log(Z_i)$ χρησιμοποιώντας το $log(Z_i)$ ως σταθερές τιμές αντιστάθμισης fixed offset values. Κατά συνέπεια, απαιτείται το βήμα της αναμενόμενης τιμής για να ληφθούν οι εκτιμήσεις των τιμών των μεταβλητών ευπάθειας. Οι εκτιμήσεις αυτές χρησιμοποιούνται για την μεγιστοποίηση της συνάρτησης Πιθανοφάνειας, ώστε τελικά να υπολογίσουμε τις άγνωστες παραμέτρους του μοντέλου.

Στο Maximization step, με βάση τη Μερική συνάρτηση Πιθανοφάνειας που παρουσιάστηκε στο μοντέλο του Cox μπορεί να γραφεί η αντίστοιχη συνάρτηση για το μοντέλο ευπάθειας.

$$L(\beta|Z) = \prod_{i=1}^{n} \left(\frac{e^{\beta' X_i + \log(Z_i)}}{\sum_{j \in R(t_i)} Z_j e^{\beta' X_j}} \right)^{\delta_i} \tag{67}$$

Οι άγνωστες τυχαίες μεταβλητές Z_i και $log(Z_i)$ αντικαθίστανται τώρα από τις αναμενόμενες τιμές τους, $E_{(\kappa)}(Z_i)$ και $E_{(\kappa)}(logZ_i)$, στο κ βήμα του αλγορίθμου.

$$\log L(\beta, \sigma^2) = \sum_{i=1}^{n} \delta_i \left[\beta' X_i + E_{(\kappa)}(\log(Z_i)) - \log \left(\sum_{j \in R(t_i)} E_{(\kappa)}(Z_j) e^{\beta' X_j} \right) \right] \tag{68}$$

Μεγιστοποιώντας την παραπάνω σχέση μπορούμε να λάβουμε εκτιμήσεις για τις παραμέτρους β με διάφορες αριθμητικές επαναληπτικές μεθόδους.

Στη συνέχεια για το Expectation step, όπως και με την παραμετρική προσέγγιση η μη παρατηρούμενη ευπάθεια για κάθε ασθενή Z_i εκτιμάται από την αναμενόμενη τιμή που τελικά δίνεται από τη σχέση,

$$E_{(\kappa+1)}(Z_i) = \frac{1}{\sigma^2_{(\kappa)} + \delta_i} \cdot \frac{1}{\sigma^2_{(\kappa)} + H_{0\kappa}(t_i;\theta) \exp(\beta'_{(\kappa)} X_i)}, \tag{69}$$

όπου $H_{0\kappa}(\cdot)$ να αποτελεί ένα μη παραμετρικό εκτιμητή της αθροιστικής βασικής συνάρτησης κινδύνου με βάση τις εκτιμήσεις από το κ βήμα. Για παράδειγμα χρησιμοποιώντας τον εκτιμητή Nelson-Aalen μία εκτίμηση θα μπορούσε να είναι η εξής,

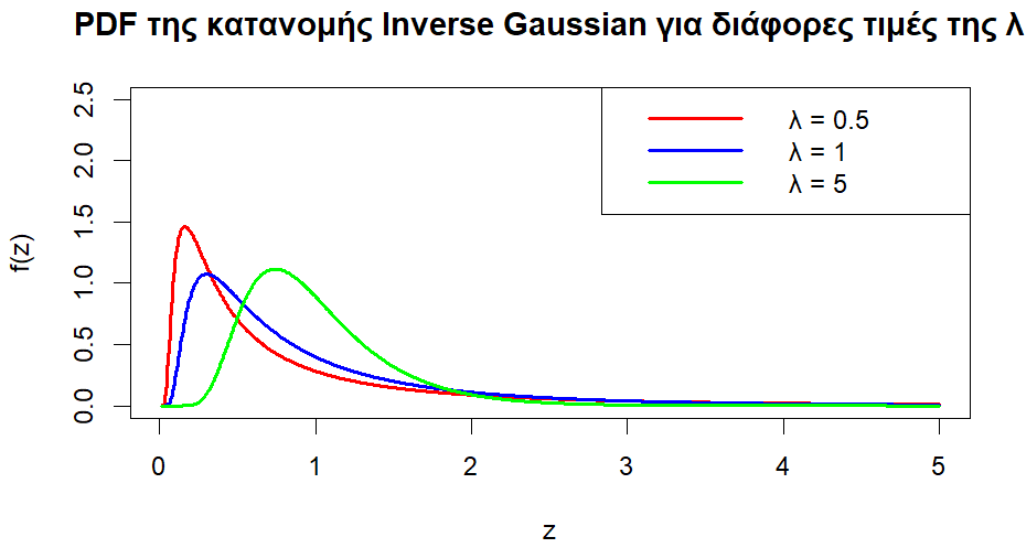
$$H_{0\kappa}(t) = \sum_{i:t_i \leq t} \frac{\delta_i}{\sum_{j \in R(t_i)} E_{(\kappa)}(Z_j) e^{\beta'_{(\kappa)} X_j}}. \tag{70}$$

2.2.4 Inverse Gaussian μοντέλο ευπάθειας

Η αντίστροφη Inverse Gaussian (Inverse Normal), αποτελεί μια εναλλακτική κατανομή ευπάθειας της κατανομής Γάμμα. Η πυκνότητα πιθανότητας μιας τυχαίας μεταβλητής που ακολουθεί Inverse Gaussian κατανομή, με παραμέτρους $\mu > 0$, $\lambda > 0$, $z > 0$,

$$f(z) = \frac{\sqrt{\lambda}}{\sqrt{2\pi} z^{3/2}} \exp\left(-\frac{\lambda}{2\mu^2 z} (z - \mu)^2 \right). \tag{71}$$

Παρακάτω έχουμε ένα διάγραμμα για διαφορετικές τιμές παραμέτρων.



Σχήμα 12: Συναρτήσεις πυκνότητας πιθανότητας της κατανομής Inverse Gaussian με μέση αναμενόμενη τιμή 1 και διακυμάνσεις 0.5, 1 και 5.

Παρατηρούμε ότι η θετική παράμετρος λ ελέγχει τη διασπορά της κατανομής και όταν δέχεται μεγαλύτερες τιμές, έχουμε αυστηρότερη συγκέντρωση της κατανομής γύρω από τον μέσο όρο.

Ο μετασχηματισμός Laplace για μία τυχαία μεταβλητή που ακολουθεί Inverse Gaussian κατανομή,

$$\mathscr{L}(s) = E(e^{-sZ}) = \exp\left(\frac{\lambda}{\mu}\left(1 - \sqrt{1 + \frac{2\mu^2 s}{\lambda}}\right)\right) \tag{72}$$

Η πρώτη και η δεύτερη παράγωγος του μετασχηματισμού Laplace δίνονται ως εξής,

$$L'(s) = -\frac{\mu}{\sqrt{1 + \frac{2\mu^2 s}{\lambda}}} \exp\left(\frac{\lambda}{\mu}\left(1 - \sqrt{1 + \frac{2\mu^2 s}{\lambda}}\right)\right) \tag{73}$$

και

$$\begin{aligned} L''(s) = \frac{\mu^3}{\lambda\left(1 + \frac{2\mu^2 s}{\lambda}\right)^{3/2}} \exp\left(\frac{\lambda}{\mu}\left(1 - \sqrt{1 + \frac{2\mu^2 s}{\lambda}}\right)\right) + \\ \frac{\mu^2}{1 + \frac{2\mu^2 s}{\lambda}} \exp\left(\frac{\lambda}{\mu}\left(1 - \sqrt{1 + \frac{2\mu^2 s}{\lambda}}\right)\right) \end{aligned} \tag{74}$$

Θέτοντας $s = 0$ λαμβάνουμε τις εκφράσεις για τη μέση τιμή και τη διασπορά της τυχαίας μεταβλητής ευπάθειας Z,

$$E(Z) = -L'(0) = \mu \tag{75}$$

και

$$V(Z) = L''(0) - (L'(0))^2 = \frac{\mu^3}{\lambda} \tag{76}$$

Υπό την υπόθεση $E(Z) = \mu = 1$, έχουμε ότι $V(Z) = 1/\lambda = \sigma^2$ και έτσι ο μετασχηματισμός Laplace μπορεί να γραφεί,

$$L(s) = e^{\frac{1}{\sigma^2}\left(1 - \sqrt{1 + 2\sigma^2 s}\right)} \tag{77}$$

Ως εκ τούτου, η συνάρτηση επιβίωσης και η συνάρτηση κινδύνου μπορεί να γραφεί,

$$S(t) = e^{\frac{1}{\sigma^2}\left(1 - \sqrt{1 + 2\sigma^2 H_0(t)}\right)} \tag{78}$$

και

$$h(t) = \frac{h_0(t)}{(1 + 2\sigma^2 H_0(t))^{1/2}} \tag{79}$$

Η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής Z, μεταξύ των επιζώντων μέχρι τη χρονική στιγμή t μπορεί να γραφεί στη μορφή,

$$f(z \mid X, T > t) = \frac{S(t \mid X, z) f(z)}{S(t \mid X)}$$

$$= \frac{1}{\sqrt{2\pi z^3}} \exp\left(- \frac{\left(z - \left(1 + 2\sigma^2 H_0(t) e^{\beta' X}\right)^{-\frac{1}{2}}\right)^2}{2\sigma^2 z} \right) \frac{1}{1 + 2\sigma^2 H_0(t) e^{\beta' X}}$$

Η παραπάνω σχέση αποτελεί τη συνάρτηση πυκνότητας πιθανότητας της Inverse Gaussian κατανομής με μέση τιμή διασπορά που δίνεται από τις σχέσεις (Wienke, 2011),

$$E(Z \mid X, T > t) = \frac{1}{\sqrt{1 + \sigma^2 H_0(t) e^{\beta' X}}} \tag{80}$$

$$V(Z \mid X, T > t) = \frac{\sigma^2}{(1 + \sigma^2 H_0(t) e^{\beta' X})^2} \tag{81}$$

2.2.5 Positive Stable μοντέλο ευπάθειας

Μια κατανομή ονομάζεται σταθερή (stable) εάν το κανονικοποιημένο άθροισμα n ανεξάρτητων τυχαίων μεταβλητών, από την κατανομή αυτή, έχει την ίδια κατανομή με έναν παράγοντα κλίμακας, πολλαπλασιασμένο με μια μόνο τυχαία μεταβλητή. Η κανονικοποίηση δίνεται από τη σχέση, $n^{1/\gamma}$, όπου η παράμετρος γ προέρχεται από το διάστημα $(0, 2]$ και ονομάζεται χαρακτηριστικός εκθέτης characteristic exponent. Για να εξασφαλίσουμε μια κατανομή στους θετικούς αριθμούς, περιοριζόμαστε στην περίπτωση των θετικών σταθερών κατανομών, που χαρακτηρίζονται από $\gamma \in (0, 1]$. Η συνάρτηση πυκνότητας πιθανότητας μιας τέτοιας θετικής μονοπαραμετρικής σταθερής τυχαίας κατανομής δίνεται από τη σχέση (Feller 1971),

$$f(z) = \frac{1}{\pi} \sum_{k=1}^{\infty} (-1)^{k+1} \frac{\Gamma(k\gamma + 1)}{k!} z^{-k\gamma - 1} \sin(k\gamma\pi) \tag{82}$$

όπου, $z \geq 0$ και $0 < \gamma \leq 1$.

Αυτή η έκφραση είναι μια δυναμοσειρά, που συγκλίνει γρήγορα για μεγάλες τιμές του z, και αργά για μικρές τιμές του z. Στην ειδική περίπτωση $\gamma = 1$, η κατανομή ευπάθειας εκφυλίζεται σε $Z = 1$. Αν και η συνάρτηση πυκνότητας

πιθανότητας μιας τυχαίας μεταβλητής με θετική σταθερή κατανομή μπορεί να αναπαρασταθεί μόνο με άπειρες σειρές, ο μετασχηματισμός Laplace έχει πολύ απλή μορφή,

$$L(u) = e^{-u^{\gamma}}. \tag{83}$$

Η πρώτη παράγωγος του μετασχηματισμού Laplace τείνει στο άπειρο, άρα και η αναμενόμενη τιμή της μεταβλητής ευπάθειας απειρίζεται και έτσι η διασπορά της Z δεν ορίζεται.

$$\lim_{u \to 0^{+}} L'(u) = -\gamma \lim_{u \to 0^{+}} \frac{e^{-u^{\gamma}}}{u^{1-\gamma}} = -\infty \tag{84}$$

Η σύγκριση της συνάρτησης κινδύνου ενός πληθυσμού με τη συνάρτηση κινδύνου υπό συνθήκη για ένα άτομο, με τιμή ευπάθειας ένα, δεν έχει νόημα επειδή η αναμενόμενη τιμή της μεταβλητής ευπάθειας δεν δεν υπάρχει. Κατά συνέπεια, ένα άτομο με ευπάθεια ίση με ένα δεν μπορεί να χρησιμεύσει, με τον ίδιο τρόπο όπως σε άλλα μοντέλα. Ο κύριος λόγος που η εν λόγω κατανομή χρησιμοποιήθηκε στα μοντέλα ευπάθειας, παρ΄ όλο που δεν ορίζεται η μέση της τιμή, είναι ότι πεπερασμένη μέση τιμή της κατανομής της ευπάθειας είναι μία μόνο απαίτηση για την αναγνωρισιμότητα των παραμέτρων σε μονομεταβλητά μοντέλα ευπάθειας (Wienke, 2011).

Χρησιμοποιώντας τον μετασχηματισμό Laplace μιας θετικής σταθερής μεταβλητής ευπάθειας, που δόθηκε προηγουμένως, η συνάρτηση πυκνότητας πιθανότητας, η συνάρτηση επιβίωσης και η συνάρτηση κινδύνου είναι,

$$S(t) = e^{-M_0(t)^{\gamma}} \tag{85}$$

$$f(t) = \gamma \mu_0(t) M_0(t)^{\gamma - 1} e^{-M_0(t)^{\gamma}} \tag{86}$$

$$\mu(t) = \gamma \mu_0(t) M_0(t)^{\gamma - 1} \tag{87}$$

3 Μέτρα απόκλισης

Τα μέτρα απόκλισης είναι κρίσιμα εργαλεία στη στατιστική για την ποσοτικοποίηση της διαφοράς μεταξύ δύο κατανομών πιθανότητας. Η κατανόηση των μέτρων απόκλισης είναι απαραίτητη για τη σύγκριση θεωρητικών κατανομών με εμπειρικά δεδομένα, την αξιολόγηση της απόδοσης μοντέλων ή την εφαρμογή αλγορίθμων μηχανικής μάθησης.

3.1 Εισαγωγή στα Μέτρα Απόκλισης

Έστω X μια τυχαία μεταβλητή που παίρνει τιμές σε ένα δειγματικό χώρο $\mathscr{X}$, ο οποίος συνήθως αποτελεί υποσύνολο του $\mathbb{R}^n$. Υποθέτουμε ότι η συνάρτηση κατανομής F της X εξαρτάται από έναν ορισμένο αριθμό παραμέτρων και ότι η συνάρτηση της F είναι γνωστή, εκτός ίσως από ένα πεπερασμένο σύνολο άγνωστων παραμέτρων θ, που σχετίζονται με την F. Έστω $(\mathscr{X}, \beta_{\mathscr{X}}, P_{\mathscr{X}})_{\theta\in\Theta}$ ο χώρος πιθανότητας που σχετίζεται με την την τυχαία μεταβλητή X, όπου $\beta_{\mathscr{X}}$ είναι το σ-πεδίο Borel υποσυνόλων $A \subset \mathscr{X}$ και $\{P_\theta\}_{\theta\in\Theta}$ μια οικογένεια κατανομών πιθανότητας που ορίζονται στον μετρήσιμο χώρο $(\mathscr{X}, \beta_{\mathscr{X}})$ με Θ ένα ανοικτό υποσύνολο του R^{M_0}, $M_0 \geq 1$. Στη συνέχεια το στήριγμα της κατανομής πιθανότητας P_θ συμβολίζεται με $S_{\mathscr{X}}$.

Υποθέτουμε ότι οι κατανομές πιθανοτήτων P_θ είναι απολύτως συνεχείς ως προς ένα σ-πεπερασμένο μέτρο μ στο $(\mathscr{X}, \beta_{\mathscr{X}})$. Για λόγους απλότητας το μ είναι είτε το μέτρο Lebesgue, δηλαδή ικανοποιεί τη συνθήκη $P_\theta(C) = 0$, όποτε το C έχει μέτρο ίσο με μηδέν, είτε ένα μέτρο απαρίθμησης, δηλαδή υπάρχει ένα πεπερασμένο ή μετρήσιμο σύνολο $S_{\mathscr{X}}$ με την ιδιότητα $P_\theta(\mathscr{X} - S_{\mathscr{X}}) = 0$.

$$f_\theta(x) = \frac{dP_\theta}{d\mu}(x) = \begin{cases} f_\theta(x) & \text{εάν } \mu \text{ μέτρο Lebesgue} \\ \Pr_\theta(X = x) = p_\theta(x) & \text{εάν } \mu \text{ μέτρο απαρίθμησης} \end{cases} \quad (x \in S_{\mathscr{X}}) \tag{88}$$

Στην πρώτη περίπτωση η X είναι τυχαία μεταβλητή με απολύτως συνεχή κατανομή και στην δεύτερη περίπτωση είναι μια διακριτή τυχαία μεταβλητή με στήριγμα $S_{\mathscr{X}}$.

Έστω h μια μετρήσιμη συνάρτηση. Η αναμενόμενη τιμή της $h(X)$ συμβολίζεται με,

$$E_\theta[h(X)] = \begin{cases} \int h(x) f_\theta(x)\, dx & \text{εάν } \mu \text{ μέτρο Lebesgue} \\ \sum_{x \in S_X} h(x) p_\theta(x) & \text{εάν } \mu \text{ μέτρο απαρίθμησης} \end{cases} \tag{89}$$

Μετά την εισαγωγή της απόστασης μεταξύ δύο κατανομών από τον Mahalanobis (1936), έχουν προταθεί διάφοροι συντελεστές για να αντικατοπτρίζουν το γεγονός ότι ορισμένες κατανομές πιθανοτήτων είναι πιο κοντά μεταξύ τους από άλλες και κατά συνέπεια ότι μπορεί να είναι ευκολότερο να διακρίνει κανείς μεταξύ ενός ζεύγους κατανομών που είναι μακριά το ένα από το άλλο, από ότι μεταξύ εκείνων που είναι πιο κοντά. Πολλά στατιστικά τέστ, όπως ο λόγος πιθανοφάνειας, το chi-square, το Wald test, ορίζονται με βάση τα μέτρα απόκλισης.

Οι αναφερόμενοι συντελεστές έχουν την κοινή ιδιότητα να αυξάνονται καθώς οι δύο κατανομές που εμπλέκονται «απομακρύνονται η μία από την άλλη». Στη συνέχεια, ένας συντελεστής με αυτή την ιδιότητα θα ονομάζεται μέτρο απόκλισης μεταξύ δύο κατανομών πιθανότητας.

3.1.1 Απόκλιση Kolmogorov

Έστω δύο μέτρα πιθανότητας P_{θ_1} και P_{θ_2} με τις μονοδιάστατες συναρτήσεις κατανομής F_{θ_1} και F_{θ_2}, αντίστοιχα. Η απόκλιση Kolmogorov (1933), μεταξύ των F_{θ_1} και F_{θ_2} (ή μεταξύ των P_{θ_1} και P_{θ_2}) είναι,

$$K_1(F_{\theta_1}, F_{\theta_2}) = \sup_{x \in \mathbb{R}} |F_{\theta_1}(x) - F_{\theta_2}(x)| \tag{90}$$

Σύμφωνα με το θεώρημα Glivenko-Cantelli, που βασίζεται στην προηγούμενη απόσταση, έχουμε ότι η εμπειρική συνάρτηση κατανομής είναι μια ομοιόμορφα ισχυρά συνεπής εκτιμήτρια της πραγματικής συνάρτησης κατανομής, δηλαδή, δεδομένου ενός τυχαίου δείγματος $X_1, ..., X_n$ ενός πληθυσμού με συνάρτηση κατανομής F_{θ_0}, για κάθε $\epsilon > 0$, ισχύει

$$\lim_{n \to \infty} \Pr\left(K_1(F_n, F_{\theta_0}) > \epsilon\right) = 0, \tag{91}$$

όπου F_n είναι η εμπειρική συνάρτηση κατανομής,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(x_i), \tag{92}$$

και I_A είναι η δείκτρια συνάρτηση του συνόλου Α.

3.1.2 Απόκλιση Levy

Σε αυτή τη περίπτωση έχουμε,

$$K_2(F_{\theta_1}, F_{\theta_2}) = \inf \{\varepsilon > 0 : F_{\theta_1}(x-\varepsilon) \leq F_{\theta_2}(x) \leq F_{\theta_1}(x+\varepsilon), \text{ για όλα τα } x\} \tag{93}$$

Η παραπάνω σχέση παίρνει τιμές στο [0, 1] και δεν είναι εύκολο να υπολογιστεί. Είναι ενδιαφέρον να σημειωθεί ότι η σύγκλιση κατά Levy συνεπάγεται ασθενή σύγκλιση για την συνάρτηση κατανομής στο R (Lukacs, 1975). Είναι αναλλοίωτη σε μετατόπιση, αλλά όχι σε κλίμακα.

3.2 Φ - Μέτρα Απόκλισης μεταξύ δύο κατανομών

Το μέτρο απόκλισης Kullback-Leibler (Kullback, S., & Leibler, R. A. (1951)), μεταξύ των κατανομών πιθανότητας P_{θ_1} και P_{θ_2} είναι,

$$D_{Kull}(\theta_1, \theta_2) = \int f_{\theta_1}(x) \log\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) d\mu(x) = E_{\theta_1}\left[\log\left(\frac{f_{\theta_1}(X)}{f_{\theta_2}(X)}\right)\right] \tag{94}$$

Ο Jeffreys (1946) χρησιμοποίησε μια συμμετρική εκδοχή της (95),

$$J(\theta_1, \theta_2) = D_{Kull}(\theta_1, \theta_2) + D_{Kull}(\theta_2, \theta_1), \tag{95}$$

ως μέτρο απόκλισης μεταξύ δύο κατανομών πιθανότητας. Αυτή η απόκλιση ονομάζεται επίσης απόκλιση J.

Ο Renyi (1961) παρουσίασε την πρώτη παραμετρική γενίκευση της (95),

$$D_r^1(\theta_1, \theta_2) = \frac{1}{r-1} \log \int_x f_{\theta_1}(x)^r f_{\theta_2}(x)^{1-r}\, d\mu(x) = \frac{1}{r-1} \log E_{\theta_1}\left[\left(\frac{f_{\theta_1}(X)}{f_{\theta_2}(X)}\right)^{r-1}\right], \tag{96}$$

όπου $\quad r > 0, \quad r \neq 1.$

Αργότερα, οι Liese και Vajda (1987) την επέκτειναν για όλα τα $\quad r \neq 0, 1,$

$$D_r^1(\theta_1, \theta_2) = \frac{1}{r(r-1)} \log \int f_{\theta_1}(x)^r f_{\theta_2}(x)^{1-r} d\mu(x) = \frac{1}{r(r-1)} \log E_{\theta_1}\left[\left(\frac{f_{\theta_1}(X)}{f_{\theta_2}(X)}\right)^{r-1}\right]. \tag{97}$$

Στη συνέχεια, η (98) θα αναφέρεται ως απόκλιση Renyi. Οι περιπτώσεις $r = 1$ και $r = 0$ ορίζονται ως εξής,

$$D'_1(\theta_1, \theta_2) = \lim_{r\to 1} D'_r(\theta_1, \theta_2) = D_{Kull}(\theta_1, \theta_2) \tag{98}$$

και

$$D'_0(\theta_1, \theta_2) = \lim_{r\to 0} D'_r(\theta_1, \theta_2) = D_{Kull}(\theta_2, \theta_1) \tag{99}$$

αντίστοιχα. Το μέτρο απόκλισης $D_{Kull}(\theta_2, \theta_1)$ ονομάζεται ελάχιστο πληροφοριακό μέτρο διάκρισης μεταξύ των κατανομών P_{θ_1} και P_{θ_2}. Άλλες δύο γνωστές παραμετρικές γενικεύσεις της (95) είναι το r-order και s-βαθμού μέτρο, και το μέτρο απόκλισης 1-τάξης και s-βαθμού, Sharma και Mittal (1977),

$$\begin{aligned} D^r_s(\theta_1, \theta_2) &= \frac{1}{s-1}\left(\left(\int f_{\theta_1}(x)^r f_{\theta_2}(x)^{1-r} d\mu(x)\right)^{\frac{s-1}{r-1}} - 1\right) \\ &= \frac{1}{s-1}\left(E_{\theta_1}\left[\left(\frac{f_{\theta_1}(X)}{f_{\theta_2}(X)}\right)^{r-1}\right]^{\frac{s-1}{r-1}} - 1\right) \end{aligned} \tag{100}$$

για $r, s \neq 1$ και

$$\begin{aligned} D^1_s(\theta_1, \theta_2) &= \frac{1}{s-1}\left(\exp\left((s-1)\int f_{\theta_1}(x) \log\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} d\mu(x)\right) - 1\right) \\ &= \frac{1}{s-1}\left(\exp\left((s-1) E_{\theta_1}\left[\log\frac{f_{\theta_1}(X)}{f_{\theta_2}(X)}\right]\right) - 1\right), \end{aligned} \tag{101}$$

για $s \neq 1$.
Επίσης ισχύει ότι,

$$\begin{aligned} &i) \quad \lim_{s\to 1} D^s_r(\theta_1, \theta_2) = r D^1_r(\theta_1, \theta_2) \\ &ii) \quad \lim_{r\to 1} D^s_r(\theta_1, \theta_2) = D^s_1(\theta_1, \theta_2) \\ &iii) \quad \lim_{s\to 1} D^s_1(\theta_1, \theta_2) = D_{Kull}(\theta_1, \theta_2), \quad \lim_{r\to -1} D^1_r(\theta_1, \theta_2) = D_{Kull}(\theta_1, \theta_2). \end{aligned} \tag{102}$$

Το μέτρο απόκλισης Kullback-Leibler είναι το πιο διάσημο μέτρο της οικογένειας των Φ-μέτρων απόκλισης.

Ορισμός 3.1. Το Φ-μέτρο απόκλισης μεταξύ των κατανομών πιθανότητας P_{θ_1} και P_{θ_2}, ορίζεται από τη σχέση,

$$\begin{aligned} D_\phi(P_{\theta_1}, P_{\theta_2}) &= D_\phi(\theta_1, \theta_2) \\ &= \int f_{\theta_2}(x)\phi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) d\mu(x) \\ &= E_{\theta_2}\left[\phi\left(\frac{f_{\theta_1}(X)}{f_{\theta_2}(X)}\right)\right] \end{aligned} \tag{103}$$

για $\phi \in \Phi^*$ όπου Φ^*, είναι η κλάση όλων των κυρτών συναρτήσεων $\phi(x)$, $x \geq 0$, τέτοιων ώστε στο $x = 1$, $\phi(1) = 0$, στο $x = 0$, $0\phi(0/0) = 0$ και $0\phi(p/0) = \lim_{u\to\infty} \frac{\phi(u)}{u}$.

Παρατήρηση 3.1. Έστω, $\phi \in \Phi^*$ διαφορίσιμη στο $x = 1$, τότε η συνάρτηση,

$$\psi(x) = \phi(x) - \phi'(1)(x - 1) \tag{104}$$

ανήκει επίσης στην κλάση συναρτήσεων Φ^* και έχει την ιδιότητα ότι $\psi'(1) = 0$. Αυτή η ιδιότητα σε συνδυασμό με την κυρτότητα, συνεπάγεται ότι $\psi(x) \geq 0$, για κάθε $x \geq 0$. Επίσης,

$$\begin{aligned} D_\psi(\theta_1, \theta_2) &= \int_x f_{\theta_2}(x)\left(\phi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) - \phi'(1)\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} - 1\right)\right) d\mu(x) \\ &= \int_x f_{\theta_2}(x)\phi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right) d\mu(x) \\ &= D_\phi(\theta_1, \theta_2). \end{aligned}$$

Δεδομένου ότι τα δύο μέτρα απόκλισης συμπίπτουν, μπορούμε να θεωρήσουμε ότι το σύνολο Φ^* είναι ισοδύναμο με το σύνολο,

$$\Phi = \Phi^* \cap \{\phi : \phi'(1) = 0\}.$$

Το μέτρο απόκλισης Kullback-Leibler προκύπτει για $\psi(x) = xlog(x) - x + 1$ ή $\phi(x) = xlog(x)$. Μπορούμε να παρατηρήσουμε ότι $\psi(x) = \phi(x) - \phi^{'}(1)(x-1)$. Θα συμβολίζουμε με ϕ κάθε συνάρτηση που ανήκει στη Φ ή στη Φ^*. Στον παρακάτω πίνακα παρουσιάζουμε μερικά σημαντικά μέτρα απόκλισης.

Πίνακας 1: Συναρτήσεις απόκλισης και οι αντίστοιχες Φ-συναρτήσεις

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

Από στατιστική άποψη, η σημαντικότερη οικογένεια Φ-αποκλίσεων είναι ίσως η οικογένεια που μελέτησαν οι Cressie και Read (1984), δηλαδή η οικογένεια δύναμης-απόκλισης (power divergence).

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

$$= \frac{1}{\lambda(\lambda+1)} \left(E_{\theta_1} \left[\left(\frac{f_{\theta_1}(X)}{f_{\theta_2}(X)} \right)^\lambda \right] - 1 \right),$$

για $-\infty < \lambda < \infty$.

Η οικογένεια της δύναμης-απόκλισης, δεν ορίζεται για $\lambda = -1$ ή $\lambda = 0$. Ωστόσο, τα συνεχή όρια του $I_\lambda(\theta_1, \theta_2)$, όταν το λ συγκλίνει στο -1 και το 0, έχουν ως αποτέλεσμα η $I_\lambda(\theta_1, \theta_2)$ να είναι συνεχής στο λ. Δεν είναι δύσκολο να διαπιστωθεί ότι,

$$\lim_{\lambda \to 0} I_\lambda(\theta_1, \theta_2) = D_{Kull}(\theta_1, \theta_2) \tag{105}$$

$$\lim_{\lambda \to -1} I_\lambda(\theta_1, \theta_2) = D_{Kull}(\theta_2, \theta_1) \tag{106}$$

Μπορούμε να παρατηρήσουμε έτσι ότι η οικογένεια της δύναμης-απόκλισης προκύπτει από την (104) με

$$\phi(x) = \begin{cases} \phi_{(\lambda)}(x) = \frac{1}{\lambda(\lambda+1)}(x^{\lambda+1} - x - \lambda(x-1)), & \lambda \neq 0, \lambda \neq -1, \\ \phi_{(0)}(x) = \lim_{\lambda \to 0} \phi_\lambda(x) = x \log x - x + 1, \\ \phi_{(-1)}(x) = \lim_{\lambda \to -1} \phi_\lambda(x) = -\log x + x - 1. \end{cases} \tag{107}$$

Τα μέτρα απόκλισης των Renyi, Sharma και Mittal που δίνονται στις (98) και (94), καθώς και το μέτρο που δίνεται από τον Bhattacharyya (1943),

$$B(\theta_1, \theta_2) = -\log \left(\int_{\mathscr{X}} \sqrt{f_{\theta_1}(x) f_{\theta_2}(x)}\, d\mu(x) \right), \tag{108}$$

δεν είναι μέτρα Φ-αποκλίσεων. Ωστόσο, τα μέτρα αυτά μπορούν να γραφούν στη μορφή,

$$D^h_\phi(\theta_1, \theta_2) = h(D_\phi(\theta_1, \theta_2)), \tag{109}$$

όπου h είναι μια διαφορίσιμη αύξουσα πραγματική συνάρτηση που απεικονίζεται από,

$$\left[0, \phi(0) + \lim_{t \to \infty} \frac{\phi(t)}{t}\right]$$

στο $[0, \infty)$. Η συνθήκη αυτή θα δειχθεί στην Πρόταση (3.1) παρακάτω.

3.3 Βασικές ιδιότητες των Φ-Μέτρων Απόκλισης

Είναι λογικό να απαιτούμε από μια απόκλιση την ιδιότητα να αυξάνεται όταν δύο κατανομές αποκλίνουν. Η Πρόταση (3.1) αποτελεί άμεση συνέπεια αυτής της ιδέας. Στη συνέχεια υποθέτουμε την ύπαρξη της πρώτης παραγώγου της φ για $x = 1$. Η υπόθεση αυτή δεν είναι απαραίτητη αλλά κάνει τη προσέγγιση αποδείξεων ευκολότερη.

Πρόταση 3.1. Έστω P_{θ_1} και P_{θ_2} δύο κατανομές πιθανοτήτων και έστω $\phi \in \Phi^*$ διαφορίσιμη για $t = 1$. Τότε,

$$0 \leq D_\phi(\theta_1, \theta_2) \leq \phi(0) + \lim_{r\to\infty} \frac{\phi(r)}{r}$$

όπου,

$$D_\phi(\theta_1, \theta_2) = 0, \quad \text{εάν} \;\; P_{\theta_1} = P_{\theta_2}, \tag{110}$$

και

$$D_\phi(\theta_1, \theta_2) = \phi(0) + \lim_{r\to\infty} \frac{\phi(r)}{r}, \quad \text{εάν } S_1 \cap S_2 = \emptyset. \tag{111}$$

Εάν η ϕ είναι επίσης αυστηρά κυρτή για $t = 1$, τότε η (111) ισχύει εάν και μόνο εάν $P_{\theta_1} = P_{\theta_2}$. Εάν επιπλέον,

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

τότε η (112) ισχύει αν και μόνο αν $S_1 \cap S_2$ είναι ίση με το κενό, όπου S_i, $i = 1, 2$, είναι το στήριγμα της κατανομής πιθανότητας P_{θ_i} , $i = 1, 2$.

Έστω τώρα $X_1, ..., X_2$ ένα δείγμα από το P_θ, $\theta \in \Theta$. Για μ που είναι το μέτρο Lebesgue ή ένα μέτρο απαρίθμησης, έστω $f_\theta(x) = \frac{dP_\theta}{d\mu}(x)$ όπου $x = (x_1, ..., x_n)$. Ας υποθέσουμε ότι T είναι ένας μετρήσιμος μετασχηματισμός από τον $(\mathscr{X}^n, \beta_{\mathscr{X}^n})$ σε έναν μετρήσιμο χώρο $(\mathscr{Y}, \beta_{\mathscr{Y}})$. Συμβολίζουμε,

$$Q_{\theta_i}(A) = P_{\theta_i}(T^{-1}(A)), \quad i = 1, 2, \tag{112}$$

όπου, $A \in \beta_{\mathscr{Y}}$ και,

$$g_{\theta_i}(t) = \frac{dQ_{\theta_i}(t)}{d\mu}, \quad f_{\theta_i}(x \mid t) = \frac{dP_{\theta_i}}{dQ_{\theta_i}}, \quad i = 1, 2; \tag{113}$$

με t συμβολίζουμε τις τιμές του T.

Πρόταση 3.2. Έστω $\phi \in \Phi^*$ και Q_{θ_i} , P_{θ_i}, $i = 1, 2$, δύο μέτρα πιθανότητας που ορίζονται στο (113) και (114). Τότε έχουμε

$$D_\phi(Q_{\theta_1}, Q_{\theta_2}) \leq D_\phi(P_{\theta_1}, P_{\theta_2}). \tag{114}$$

Η ισότητα ισχύει εάν η T είναι επαρκής για τις κατανομές πιθανότητας P_{θ_1} και P_{θ_2}.

Στην ακόλουθη πρόταση $\{P_\theta\}_{\theta\in\Theta}, \Theta \subseteq \mathbb{R},$, είναι μια οικογένεια μέτρων πιθανότητας που ορίζονται στο σ-πεδίο των Borel υποσυνόλων, με μονότονο λόγο πιθανοφάνειας ως προς x, δηλαδή, αν για κάθε $\theta_1 < \theta_2$, τα $f_{\theta_1}(x)$ και $f_{\theta_2}(x)$ διαφέρουν και ο λόγος $f_{\theta_1}(x)/f_{\theta_2}(x)$ είναι μη φθίνουσα συνάρτηση του x. Είναι επίσης δυνατό να οριστούν οικογένειες πυκνοτήτων με μη αύξοντα λόγο πιθανοφάνειας ως προς x, αλλά τέτοιες οικογένειες μπορούν να αντιμετωπιστούν με συμμετρία.

Πρόταση 3.3. Ας υποθέσουμε ότι οι κατανομές πιθανοτήτων $\{P_\theta\}_{\theta\in\Theta}$, παίρνουν τιμές στην πραγματική ευθεία, $\theta \in (a, b) \subseteq \mathbb{R}$ και έστω ότι η P_θ είναι απολύτως συνεχής ως προς ένα σ-πεπερασμένο μέτρο μ (μέτρο Lebesgue ή μέτρο απαρίθμησης). Έστω επίσης ότι οι αντίστοιχες συναρτήσεις πυκνότητας πιθανότητας ή συναρτήσεις μάζας πιθανότητας έχουν μονότονο λόγο πιθανοφάνειας ως προς x. Αν $a < \theta_1 < \theta_2 < \theta_3 < b$ και η συνάρτηση ϕ είναι συνεχής, ισχύει,

$$D_\phi(\theta_1, \theta_2) \leq D_\phi(\theta_1, \theta_3), \quad \phi \in \Phi^*. \tag{115}$$

Παρατήρηση 3.2. Είναι προφανές ότι αν η h είναι μια διαφορίσιμη αύξουσα πραγματική απεικόνιση, τα (h, ϕ)-μέτρα απόκλισης ικανοποιούν επίσης τις προτάσεις 3.1, 3.2 και 3.3.

Παρατήρηση 3.3. Αν θεωρήσουμε μια συνάρτηση $\phi \in \Phi^*$ η οποία είναι αυστηρά κυρτή στο $x = 1$, η αντίστοιχη ϕ-απόκλιση είναι ένα μέτρο στο χώρο $P = P_{\theta\in\Theta}$. Είναι δυνατό να ορίσουμε ένα νέο μέτρο απόκλισης, με βάση μια δεδομένη ϕ-απόκλιση, με τέτοιο τρόπο ώστε το νέο μέτρο απόκλισης να είναι και συμμετρικό. Αυτό είναι δυνατό αν θεωρήσουμε το μέτρο απόκλισης που σχετίζεται με τη συνάρτηση $\varphi(t) = \phi(t) + t\phi(1/t)$ Vajda (1995).

3.4 Έλεγχοι Καλής Προσαρμογής με βάση τα ϕ-μέτρα απόκλισης

Οι στατιστικοί έλεγχοι με βάση τις Φ-αποκλίσεις χρησιμοποιούνται για τον έλεγχο υποθέσεων, ιδίως για τους ελέγχους καλής προσαρμογής, ειδικότερα για να ελέγξουμε αν ένα δείγμα προέρχεται από μια συγκεκριμένη κατανομή.

3.4.1 Εκτιμητής Μεγίστης πιθανοφάνειας με βάση ελεγχοσυναρτήσεις Φ-αποκλίσεων

Η μέθοδος μεγίστης πιθανοφάνειας (MLE) και οι στατιστικοί έλεγχοι που βασίζονται στις Φ-αποκλίσεις είναι δύο ισχυρά στατιστικά εργαλεία που χρησιμοποιούνται για την εκτίμηση παραμέτρων και τον έλεγχο υποθέσεων. Η MLE επιλέγει τις παραμέτρους που μεγιστοποιούν τη συνάρτηση πιθανοφάνειας, η οποία μετρά πόσο πιθανά είναι τα παρατηρούμενα δεδομένα, δεδομένου διαφορετικών τιμών των παραμέτρων. Στην ουσία επιλέγουμε μια κατανομή πιθανότητας που μοντελοποιεί τα δεδομένα με άγνωστες παραμέτρους και κατασκευάζουμε τη συνάρτηση πιθανοφάνειας, που είναι η από κοινού πιθανότητα των παρατηρούμενων δεδομένων ως συνάρτηση των παραμέτρων. Επίσης γίνεται εύρεση των τιμών των παραμέτρων που μεγιστοποιούν τη συνάρτηση πιθανότητας. Αυτό γίνεται συχνά με τη χρήση λογισμού (μηδενισμός της παραγώγου) ή αριθμητικών μεθόδων.

Για να αναλύσουμε την περίπτωση στην οποία έχουμε μόνο ομαδοποιημένα δεδομένα σε αντίθεση με πρόσβαση σε μη ομαδοποιημένες παρατηρήσεις, είναι σημαντικό να εξετάσουμε τις επιπτώσεις όσον αφορά τη στατιστική εκτίμηση και τον έλεγχο. Η συζήτηση θα περιστραφεί γύρω από τη χρήση των ελάχιστων Φ-αποκλίσεων στην περίπτωση των ομαδοποιημένων δεδομένων και των εκτιμητών μέγιστης πιθανοφάνειας MLE όταν είναι διαθέσιμα τα αρχικά δεδομένα.

Ένα πρώτο σενάριο αφορά τα ομαδοποιημένα δεδομένα και την εκτίμηση της ελάχιστης Φ-απόκλισης. Όταν τα δεδομένα είναι ομαδοποιημένα σε M διαστήματα, με N_i, όπου $i = 1, ..., M$ να αντιπροσωπεύουν τον αριθμό των παρατηρήσεων σε κάθε διάστημα, τα μεμονωμένα σημεία των δεδομένων $Y_1, ...Y_n$ δεν είναι άμεσα παρατηρήσιμα. Αυτό περιορίζει τις στατιστικές τεχνικές που μπορούν να χρησιμοποιηθούν για την εκτίμηση των παραμέτρων.

Η εκτίμηση της ελάχιστης Φ-απόκλισης, ελαχιστοποιεί την Φ-απόκλιση μεταξύ της εμπειρικής κατανομής των παρατηρούμενων ομαδοποιημένων δεδομένων και του θεωρητικού μοντέλου που εξαρτάται από την παράμετρο θ. Ο εκτιμητής $\hat{\theta}_\phi$ ορίζεται ελαχιστοποιώντας την,

$$D_\varphi(P, Q_\theta) = \sum_{i=1}^{M} \varphi\left(\frac{p_i}{q_i(\theta)}\right) q_i(\theta) \tag{116}$$

όπου $p_i = \frac{N_i}{n}$ και $q_i(\theta)$ είναι η θεωρητική πιθανότητα του i-οστού διαστήματος, που εξαρτάται από την παράμετρο θ.

Η ομαδοποίηση των δεδομένων έχει ως αποτέλεσμα την απώλεια πληροφοριών, καθώς υπολογίζει τον μέσο όρο. Αυτή η ομαδοποίηση δεδομένων μπορεί να οδηγήσει σε λιγότερο αποτελεσματικές εκτιμήσεις.

Ένα δεύτερο σενάριο τώρα, αφορά τα αρχικά δεδομένα και τον εκτιμητή μεγίστης πιθανοφάνειας. Έχουμε ότι, όταν οι αρχικές παρατηρήσεις $Y_1, ...Y_n$ είναι διαθέσιμες, μπορούν να χρησιμοποιηθούν πιο εξελιγμένες και δυνητικά πιο ισχυρές στατιστικές μέθοδοι. Ο MLE συνήθως προτιμάται σε τέτοια σενάρια. Ο MLE βρίσκει την παράμετρο θ που μεγιστοποιεί τη συνάρτηση πιθανοφάνειας, η οποία είναι η πιθανότητα της παρατήρησης του συγκεκριμένου δείγματος, δεδομένου του θ. Το $\hat{\theta}_{MLE}$ προκύπτει από τη μεγιστοποίηση,

$$L(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} f_\theta(y_i), \tag{117}$$

όπου $f_\theta(y_i)$ είναι η συνάρτηση πυκνότητας πιθανότητας του Y_i υπό το θ.

Το πλεονέκτημα στο παραπάνω, αποτελεί ότι ο MLE αξιοποιεί πλήρως τα δεδομένα, καταγράφοντας όλες τις διαθέσιμες πληροφορίες από κάθε παρατήρηση Y_i. Αυτό συνήθως οδηγεί σε πιο αποτελεσματικούς εκτιμητές, δηλαδή έχουν μικρότερες διασπορές μεταξύ των αμερόληπτων εκτιμητών.

Ο MLE γενικά παρέχει πιο αποτελεσματικούς εκτιμητές από τους εκτιμητές της ελάχιστης Φ-απόκλισης, όταν τα δεδομένα είναι ομαδοποιημένα. Η πλήρης αξιοποίηση των μεμονωμένων σημείων των δεδομένων στον MLE αξιοποιεί περισσότερες πληροφορίες από ότι οι συγκεντρωτικές ομάδες. Παρόλο που οι MLE είναι στατιστικά προτιμότεροι, ενδέχεται να είναι πιο πολύπλοκοι στην εφαρμογή τους, ιδίως σε περιπτώσεις που αφορούν μεγάλα σύνολα δεδομένων ή πολύπλοκα μοντέλα. Οι έλεγχοι που βασίζονται στον MLE είναι συνήθως πιο ισχυρές λόγω της υψηλότερης αποτελεσματικότητας των εκτιμητών. Είναι καλύτεροι στον εντοπισμό μικρών αλλά σημαντικών αποκλίσεων από μηδενικές υποθέσεις σε σύγκριση με τους ελέγχους που βασίζονται σε ομαδοποιημένα δεδομένα.

Τώρα θα παρουσιάσουμε το θεώρημα που δηλώνει την ασυμπτωτική κατανομή της Φ-απόκλισης, η οποία συμβολίζεται με $T_n^\phi(\hat{\theta})$, όπου $\hat{\theta}$ είναι ένας

εκτιμητής της παραμέτρου. Η απόδειξη βρίσκεται στο Θεώρημα Morales et al. (1995), για τα μέτρα Φ-απόκλισης και στους Chernoff και Lehman (1954) για την ειδική περίπτωση όπου, $\phi_1(x) = \frac{1}{2}(x-1)^2$. Ξέρουμε πως οι Φ-αποκλίσεις χρησιμοποιούνται για τη μέτρηση της απόστασης μεταξύ δύο κατανομών πιθανότητας, παρόμοια με την απόκλιση Kullback-Leibler και βασίζονται στη κυρτή συνάρτηση $\phi(x) = (x-1)^2$, που αντιστοιχεί στη τετραγωνική απόκλιση. Το $T_n^{\phi}(\hat{\theta})$ κατασκευάζεται χρησιμοποιώντας το μέτρο Φ-απόκλισης. Χρησιμοποιείται για τον έλεγχο υποθέσεων σχετικά με την παράμετρο θ ενός στατιστικού μοντέλου. Το στατιστικό αυτό αξιολογεί γενικά την προσαρμογή ενός μοντέλου συγκρίνοντας τα παρατηρούμενα δεδομένα με αυτό που προβλέπει το μοντέλο με τις εκτιμώμενες παραμέτρους $\hat{\theta}$.

Θεώρημα 3.1. Σύμφωνα με τις συνθήκες που δίνονται στους Morales et al. (1995) η ασυμπτωτική κατανομή του στατιστικού ελέγχου Φ-απόκλισης $T_n^{\phi}(\hat{\theta})$, όπου $\hat{\theta}$ είναι ο εκτιμητής μεγίστης πιθανοφάνειας με βάση τα αρχικά δεδομένα, δίνονται από

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

όπου Z_j είναι ανεξάρτητες και κανονικά κατανεμημένες τυχαίες μεταβλητές με μέση τιμή μηδέν και μοναδιαία διακύμανση, και οι λ_j, $0 \leq \lambda_j \leq 1$, είναι οι ρίζες της εξίσωσης,

$$det(I_F(\theta_0) - \lambda \mathscr{I}_F(\theta_0)) = 0,$$

όπου $\mathscr{I}_F(\theta_0)$ και $I_F(\theta_0)$ είναι οι πίνακες πληροφοριών Fisher από το αρχικό και το διακριτοποιημένο μοντέλο, αντίστοιχα.

Το θεώρημα στην ουσία δηλώνει ότι η ασυμπτωτική κατανομή της $T_n^{\phi}(\hat{\theta})$ συγκλίνει σε μια μη κεντρική X^2 κατανομή καθώς το μέγεθος του δείγματος n πλησιάζει στο άπειρο. Η δεύτερη παράγωγος της συνάρτησης ϕ, υπολογίζεται στο 1 και η επιλογή της επηρεάζει τον τύπο της απόκλισης, με συνήθεις επιλογές που περιλαμβάνουν το τετραγωνικό σφάλμα (που οδηγεί στο τεστ του Pearson) και τη λογαριθμική απόκλιση (που οδηγεί στην απόκλιση Kullback-Leibler). Το ϕ- μέτρο απόκλισης $D_{\phi}(\hat{p}, p(\hat{\theta}))$ μετρά την απόσταση μεταξύ της κατανομής των εμπειρικών δεδομένων, που εδώ συμβολίζονται με $\hat{p}$, και της κατανομής του μοντέλου $p(\hat{\theta})$, κάτω από την μηδενική υπόθεση όπου η άγνωστη παράμετρος εκτιμάται από τον εκτιμητή μεγίστης πιθανοφάνειας (MLE) $\hat{\theta}$. Το

ϕ- μέτρο απόκλισης ποσοτικοποιεί την ασυμφωνία μεταξύ των παρατηρούμενων δεδομένων και του μοντέλου κάτω από την μηδενική υπόθεση. Η ασυμπτωτική κατανομή της ελεγχοσυνάρτησης αποτελείται από μία X^2 κατανομή με $M - M_0 - 1$ βαθμούς ελευθερίας και ένα γραμμικό συνδυασμό κατανομών Q^2. Οι βαθμοί ελευθερίας $M - M_0 - 1$ προκύπτουν από τον αριθμό των ανεξάρτητων κατηγοριών-διαστημάτων μειωμένων από τον αριθμό των εκτιμώμενων παραμέτρων θ.

3.5 Πίνακας Πληροφορίας του Fisher (Fisher Information Matrix)

Με τον πληροφοριακό αριθμό του Fisher (Fisher Information) $I(\theta)$ αντιπροσωπεύουμε την ποσότητα της πληροφορίας που παρέχει μια τυχαία μεταβλητή Y για μια παράμετρο θ, με βάση τη συνάρτηση πυκνότητας πιθανότητας $p(Y|\theta)$. Ο πληροφοριακός αριθμός ορίζεται σε μία δεδομένη τιμή της παραμέτρου $\theta = \theta^*$ η οποία συχνά είναι η πραγματική τιμή που συμβολίζεται με θ_0 ως,

$$I(\theta^*) = V_{\theta^*}\left(\frac{\partial}{\partial\theta}\log p(Y \mid \theta)\Big|_{\theta=\theta^*}\right) \tag{118}$$

όπου $V(x)$ να είναι η διακύμανση της τυχαίας μεταβλητής X. Για την τιμή της πληροφορίας του Fisher, για μία παρατήρηση του δείγματος, ισχύει η σχέση,

$$\begin{aligned} I(\theta^*) &= E\left(\left(\frac{\partial}{\partial\theta}\log p(Y \mid \theta)\right)^2\Big|_{\theta=\theta^*}\right) \\ \iff I(\theta^*) &= \int_0^\infty \left(\frac{\partial}{\partial\theta}\log f(y,\theta)\right)^2 f(y,\theta)\Big|_{\theta=\theta^*} dy \end{aligned} \tag{119}$$

Εάν η $logp(Y|\theta)$ είναι δύο φορές παραγωγίσιμη, τότε ο πληροφοριακός αριθμός ισούται με,

$$I(\theta^*) = E_{\theta^*}\left(-\frac{\partial^2}{\partial\theta^2}\log p(Y \mid \theta)\Big|_{\theta=\theta^*}\right) \tag{120}$$

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

Επίσης αποκαλούμε Fisher Information Matrix μια γενίκευση του Fisher Information για την περίπτωση που η παράμετρος θ είναι διάνυσμα, κατά την οποία δημιουργείται ένας πίνακας, στον οποίο το στοιχείο που ανήκει στην i γραμμή και στην j στήλη,

$$I(\theta^*)_{i,j} = E_{\theta^*}\left(\left(\frac{\partial}{\partial\theta_i}\log p(Y \mid \theta)\Big|_{\theta=\theta^*}\right)\left(\frac{\partial}{\partial\theta_j}\log p(Y \mid \theta)\Big|_{\theta=\theta^*}\right)\right) \quad (121)$$

Ομοίως, αποδεικνύεται ότι,

$$I(\theta^*)_{i,j} = -E_{\theta^*}\left(\frac{\partial^2 \log p(Y \mid \theta)}{\partial\theta_i\partial\theta_j}\Big|_{\theta=\theta^*}\right) \quad (122)$$

4 Στατιστικός έλεγχος βάσει $\phi-$μέτρων απόκλισης για τα μοντέλα ευπάθειας – Θεωρία και Προσομοιώσεις

Στατιστικός έλεγχος βάσει $\phi-$μέτρων απόκλισης για τα μοντέλα ευπάθειας έχει προταθεί και εξεταστεί ήδη στην διπλωματική εργασία Σοφοκλέους και Βόντα (2023) για δεδομένα χωρίς λογοκρισία. Λόγω του ότι είναι πολύ σύνηθες στην ανάλυση επιβίωσης να υπάρχουν ημιτελείς παρατηρήσεις π.χ. με τη μορφή της λογοκρισίας γενικεύουμε σε αυτή την εργασία τον έλεγχο έτσι ώστε να χειρίζεται επίσης λογοκριμένες παρατηρήσεις από δεξιά.

4.1 Ελεγχοσυνάρτηση

Για τον έλεγχο καλής προσαρμογής με σύνθετη μηδενική υπόθεση

$$H_0 : \mathbf{p}^0(\theta) = (p_1^0(\theta), ..., p_M^0(\theta))^T,$$

η γενική μορφή της ελεγχοσυνάρτησης (Σοφοκλέους και Βόντα 2023) η οποία βασίζεται στο βιβλίο του Pardo(2006) είναι

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

όπου $\hat{\mathbf{p}}$ μια μη παραμετρική εκτίμηση της κατανομής των παρατηρήσεων και $\mathbf{p}^0(\theta)$, η παραμετρική κατανομή που υποθέτουμε κάτω από την μηδενική υπόθεση. Όπως συνηθίζεται σε ελέγχους καλής προσαρμογής τα δεδομένα χωρίζονται σε M διαστήματα. Η ελεγχοσυνάρτηση μετρά την απόσταση μεταξύ μη παραμετρικής και παραμετρικής κατανομής και η μηδενική απορρίπτεται αν η απόσταση αυτή είναι μεγάλη.

Λόγω του ότι υποθέτουμε την ύπαρξη λογοκριμένων παρατηρήσεων από δεξιά, η μηδενική υπόθεση του προτεινόμενου ελέγχου γράφεται σε σχέση με την συνάρτηση κινδύνου h ως

$$H_0 : \mathbf{h}^0(\theta) = (h_1^0(\theta), ..., h_M^0(\theta))^T.$$

Παρόμοια και η ελεγχοσυνάρτηση του ελέγχου γενικεύεται ως εξής

$$T_n^{\phi}(\hat{\mathbf{h}}, \mathbf{h}^0(\theta)) = \frac{2}{\phi''(1)} \sum_{i=1}^{M} r_i \left\{ h_i^0(\theta) \phi\left(\frac{\hat{h}_i}{h_i^0(\theta)}\right) + (1 - h_i^0(\theta)) \phi\left(\frac{1-\hat{h}_i}{1-h_i^0(\theta)}\right) \right\} \tag{123}$$

όπου r_i ο αριθμός των ατόμων που είναι σε κίνδυνο λίγο πριν το i διάστημα και

$$\hat{h}_i = d_i/r_i, \quad i = 1, ..., M$$

η εκτιμήτρια Nelson-Aalen της συνάρτησης κινδύνου στο i διάστημα, με d_i να είναι ο αριθμός των αποτυχιών στο i διάστημα.

Παρατηρούμε ότι η ελεγχοσυνάρτηση εξαρτάται από την πραγματική τιμή της παραμέτρου $\theta = \theta_0$ η οποία είναι στην πραγματικότητα άγνωστη. Για τον υπολογισμό της ελεγχοσυνάρτησης η παράμετρος πρέπει να εκτιμηθεί όπως συμβαίνει σε κάθε σύνθετη μηδενική υπόθεση. Η εκτιμήτρια που προτείνεται για την εκτίμηση της παραμέτρου ορίζεται ως

$$\hat{\theta_{\phi}} = arg \min_{\theta \in \Theta} T_n^{\phi}(\hat{\mathbf{h}}, \mathbf{h}^0(\theta)) = \tag{124}$$

$$arg \min_{\theta \in \Theta} \frac{2}{\phi''(1)} \sum_{i=1}^{M} r_i \left\{ h_i^0(\theta) \phi\left(\frac{\hat{h}_i}{h_i^0(\theta)}\right) + (1 - h_i^0(\theta)) \phi\left(\frac{1-\hat{h}_i}{1-h_i^0(\theta)}\right) \right\}.$$

Όταν σε αυτή την ελαχιστοποίηση χρησιμοποιείται η συνάρτηση ϕ των Kullback-Leibler η εκτιμήτρια που προκύπτει είναι η εκτιμήτρια μεγίστης πιθανοφάνειας του θ.

Το χωρίο απορρίψεως της μηδενικής υπόθεσης είναι της μορφής

$$T_n^{\phi}(\hat{\mathbf{h}}, \mathbf{h}(\hat{\theta}_{\phi})) \geq c_a \tag{125}$$

όπου η σταθερά c_a θα υπολογιστεί κατάλληλα από την κατανομή της ελεγχοσυνάρτησης κάτω από την μηδενική υπόθεση για δεδομένο επίπεδο σημαντικότητας a.

Πιο συγκεκριμένα σχετικά με το διαμερισμό των δεδομένων σε M διαστήματα, έστω ο διαμερισμός της κατανομής των γεγονότων σε M διαστήματα τα οποία συμβολίζουμε με $E_j, j = 1, .., M$. Πιο συγκεκριμένα έστω η διαμέριση του διαστήματος παρατήρησης $[0, \tau]$ ως εξής $0 = a_0 < a_1 < a_2 < ... < a_M = \tau$ έτσι ώστε $E_j = (a_{j-1}, a_j], j = 1, ..., M$. Έστω h_j ο ρυθμός κινδύνου στο j διάστημα και αντίστοιχα $1 - h_j$ ο ρυθμός επιβίωσης στο ίδιο διάστημα.

Λόγω του ότι υποθέτουμε την ύπαρξη λογοκριμένων παρατηρήσεων από δεξιά, τα δεδομένα μας είναι έρχονται από ζεύγη τ.μ. $(T_1, \delta_1), ..., (T_n, \delta_n)$ με $T_i = \min(X_i, C_i)$ όπου T_i οι χρόνοι επιβίωσης, C_i οι χρόνοι λογοκρισίας και δ_i η δείκτρια συνάρτηση της λογοκρισίας η οποία είναι ίση με 1 αν έχουμε μη λογοκριμένη παρατήρηση και 0 αλλιώς.

Η συνάρτηση κινδύνου h_j^0 στο διάστημα $(a_{j-1}, a_j]$ ισούται εξ ορισμού με

$$h_j^0(\theta) = \frac{S_0(a_{j-1}, \theta) - S_0(a_j, \theta)}{S_0(a_{j-1}, \theta)}$$

και για το μοντέλο ευπάθειας γράφεται ως

$$\frac{e^{-G(H_0(a_{j-1},\theta))} - e^{-G(H_0(a_j,\theta))}}{e^{-G(H_0(a_{j-1},\theta))}}$$

βάσει του ορισμού της συνάρτησης επιβίωσης των μοντέλων ευπάθειας που σε αυτή την εργασία θα είναι

$$S_0(t, \theta) = e^{-G(H_0(t,\theta))} \tag{126}$$

όπου $H_0(t)$ η αθροιστική συνάρτηση κινδύνου κάτω από τη μηδενική υπόθεση και

$$G(x) = -ln(\int_0^\infty e^{-xz} dF_Z(z))$$

ο πλην λογάριθμος του μετασχηματισμού Laplace της τ.μ. ευπάθειας Z.

Στόχος αυτού του κεφαλαίου είναι η εξέταση της συμπεριφοράς του προτεινόμενου ελέγχου μέσω του μεγέθους και της ισχύος του για να δούμε αν ο έλεγχος είναι αποτελεσματικός για τα μοντέλα ευπάθειας. Συγκεκριμένα θα εξεταστούν τα μοντέλα ευπάθειας, Γάμμα και Inverse Gaussian. Η συνάρτηση ϕ που θα θεωρήσουμε είναι $\phi(x) = xlogx - x + 1$, με $\phi^{'}(x) = logx + 1$ και $\phi^{''}(x) = \frac{1}{x}$ δηλαδή αυτή των Kullback-Leibler . Επίσης θα θεωρήσουμε την εκθετική κατανομή για να μοντελοποιήσουμε την συνάρτηση κινδύνου κάτω από την μηδενική υπόθεση. Λεπτομέρειες θα δοθούν πιο κάτω σε αυτό το κεφάλαιο. Το μέγεθος και η ισχύς του ελέγχου θα μελετηθούν μέσω προσομοιώσεων δεδομένων με τη χρήση του στατιστικού πακέτου R.

4.2 Ασυμπτωτική Κατανομή της ελεγχοσυνάρτησης κάτω από την μηδενική υπόθεση

Παρακάτω θα αναλύσουμε την διαδικασία για την εύρεση της ασυμπτωτικής κατανομής της ελεγχοσυνάρτησης $T_n^\phi(\hat{\mathbf{h}}, \mathbf{h}(\hat{\theta}_\phi))$ κάτω από την μηδενική υπόθεση

με βάση το Θεώρημα Morales όπως δόθηκε και παραπάνω, οπότε

$$T_n^{\phi}(\hat{\mathbf{h}}, \mathbf{h}(\hat{\theta}_{\phi})) \xrightarrow[n\to\infty]{} X^2_{M-M_0} + \sum_{j=1}^{M_0}(1-\lambda_j)Z_j^2,$$

αφού είναι γνωστό ότι

$$T_n^{\phi}(\hat{\mathbf{h}}, \mathbf{h}(\hat{\theta}_{\phi})) \xrightarrow[n\to\infty]{} X^2_{M-M_0}$$

Vonta and Karagrigoriou (2014). Σε αυτή την εργασία θα υποθέσουμε $\theta \in \Theta$ με $\Theta \subseteq R$ και άρα $M_0 = 1$ και έτσι το θεώρημα του Morales γίνεται,

$$T_n^{\phi}(\hat{\mathbf{h}}, \mathbf{h}(\hat{\theta}_{\phi})) \xrightarrow[n\to\infty]{} X^2_{M-1} + \sum_{j=1}^{1}(1-\lambda_j)Z_j^2,$$

ή τελικά

$$T_n^{\phi}(\hat{\mathbf{h}}, \mathbf{h}(\hat{\theta}_{\phi})) \xrightarrow[n\to\infty]{} X^2_{M-1} + (1-\lambda)Z^2, \tag{127}$$

όπου Z είναι τυχαία μεταβλητή που ακολουθεί τυπική κανονική κατανομή και $0 \leq \lambda \leq 1$, να είναι η λύση της εξίσωσης,

$$I_F(\theta_0) - \lambda \cdot \mathscr{I}_F(\theta_0) = 0 \implies \lambda = \frac{I_F(\theta_0)}{\mathscr{I}_F(\theta_0)}. \tag{128}$$

όπου θ_0 η πραγματική τιμή της παραμέτρου θ, $\mathscr{I}_F(\theta_0)$, και $I_F(\theta_0)$ οι πληροφοριακοί αριθμοί του αρχικού και του διακριτοποιημένου μοντέλου αντίστοιχα.

Για την εύρεση του λ, θα χρειαστεί να υπολογίσουμε τους τύπους από τους οποίους δίνονται γενικά οι δύο πληροφοριακοί αριθμοί $\mathscr{I}_F(\theta_0)$ και $I_F(\theta_0)$ για την περίπτωση των μοντέλων ευπάθειας.

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

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

Η παραπάνω σχέση αντιπροσωπεύει την πιθανοφάνεια για κάθε μεμονωμένη παρατήρηση t_i, υποθέτοντας ότι όλες οι παρατηρήσεις είναι χωρίς λογοκρισία και παρέχουν πλήρεις πληροφορίες. Για λογοκριμένες παρατηρήσεις, εμπλέκεται

η συνάρτηση επιβίωσης $S(t_i, \theta) = P(T > t_i)$, η οποία δίνει την πιθανότητα ότι το γεγονός ενδιαφέροντος θα συμβεί μετά τη στιγμή t_i.

Σε αυτή την εργασία θα θεωρήσουμε παρατηρήσεις οι οποίες είναι δυνητικά λογοκριμένες από δεξιά. Έστω δ_i ο δείκτης λογοκρισίας, όπου $\delta_i = 1$ εάν το συμβάν παρατηρηθεί (χωρίς λογοκρισία) και $\delta_i = 0$ εάν η παρατήρηση λογοκρίνεται. Έτσι η συνάρτηση πιθανοφάνειας ορίζεται ως,

$$\mathscr{L}(t_1, \delta_1, ..., t_n, \delta_n) = \prod_{i=1}^{n} \left[f(t_i, \theta)^{\delta_i} \cdot S(t_i, \theta)^{1-\delta_i} \right].$$

Ισχύει εξ ορισμού ότι,

$$F(t, \theta) = 1 - S(t, \theta)$$

όπου η συνάρτηση επιβίωσης για τα μοντέλα ευπάθειας ορίζεται γενικά σε αυτή την εργασία ως,

$$S(t, \theta) = e^{-G(H(t,\theta))}$$

με

$$G(x) = -ln(\int_0^\infty e^{-xz} dF_Z(z)).$$

Δηλαδή η συνάρτηση G είναι ο αρνητικός λογάριθμος του μετασχηματισμού Laplace της τ.μ. ευπάθειας Z.

Επίσης,

$$\frac{d}{dt} F(t, \theta) = f(t, \theta) = -\frac{d}{dt} S(t, \theta),$$

και

$$f(t, \theta) = -\frac{d}{dt} \left(e^{-G(H(t,\theta))} \right).$$

Η μορφή της συνάρτησης πυκνότητας για το μοντέλο ευπάθειας στο θ_0 είναι,

$$f(t, \theta_0) = h(t, \theta_0) e^{-G(H(t,\theta_0))} G'(w) \Big|_{w=H(t,\theta_0)}$$

όπου η παράγωγος ως προς w θα συμβολίζεται με $'$.

Στη συνέχεια θα χρησιμοποιήσουμε τον ακόλουθο συμβολισμό για να τονίσουμε ότι μας ενδιαφέρει η ασυμπτωτική κατανομή της ελεγχοσυνάρτησης κάτω από την μηδενική υπόθεση

$$f_0(t, \theta) = h_0(t, \theta) e^{-G(H_0(t,\theta))} G'(w) \Big|_{w=H_0(t,\theta)}.$$

Η συνάρτηση πιθανοφάνειας βάση αυτού του ορισμού και του ορισμού της συνάρτησης επιβίωσης γίνεται,

$$\mathscr{L} = \prod_{i=1}^{n} \left[\left(h_0(t_i,\theta) e^{-G(H_0(t_i,\theta))} G'(w)\Big|_{w=H_0(t_i,\theta)} \right)^{\delta_i} \cdot \left(e^{-G(H_0(t_i,\theta))} \right)^{1-\delta_i} \right]$$

Μπορούμε να απλοποιήσουμε τους όρους ως εξής,

$$\mathscr{L} = \prod_{i=1}^{n} \left[(h_0(t_i,\theta))^{\delta_i} \left(G'(w)\Big|_{w=H_0(t_i,\theta)} \right)^{\delta_i} \right] e^{-G(H_0(t_i,\theta))} \tag{129}$$

Λογαριθμίζοντας την, έχουμε ότι

$$\log \mathscr{L} = \sum_{i=1}^{n} \left[\delta_i \log h_0(t_i,\theta) + \delta_i \log G'(w)\Big|_{w=H_0(t_i,\theta)} - G(H_0(t_i,\theta)) \right]. \tag{130}$$

Στη συνέχεια παραγωγίζουμε ως προς θ, και συμβολίζουμε την παράγωγο ως προς θ με τελεία, οπότε

$$\frac{\partial}{\partial\theta}(\delta_i \log h_0(t_i,\theta)) = \delta_i \frac{1}{h_0(t_i,\theta)} \frac{\partial h_0(t_i,\theta)}{\partial\theta} = \delta_i \frac{\dot{h_0}(t_i,\theta)}{h_0(t_i,\theta)},$$

και,

$$\frac{\partial}{\partial\theta} \left(\delta_i \log G'(w)\Big|_{w=H_0(t_i,\theta)} \right) = \delta_i \frac{G''(w)}{G'(w)}\Big|_{w=H_0(t_i,\theta)} \dot{H_0}(t_i,\theta).$$

Συνεπώς,

$$\begin{aligned} \frac{\partial \log \mathscr{L}}{\partial\theta} = \sum_{i=1}^{n} \Bigg[& \delta_i \frac{\dot{h_0}(t_i,\theta)}{h_0(t_i,\theta)} \\ & + \delta_i \frac{G''(w)}{G'(w)}\Big|_{w=H_0(t_i,\theta)} \dot{H_0}(t_i,\theta) - G'(w)\Big|_{w=H_0(t_i,\theta)} \dot{H_0}(t_i,\theta) \Bigg]. \end{aligned} \tag{131}$$

Στη συνέχεια και εφόσον έχουμε υποθέσει σε αυτή την εργασία ότι η παράμετρος θ είναι μονοδιάστατη, θα υποθέσουμε ότι η συνάρτηση αθροιστικού κινδύνου (και συνεπώς η συνάρτηση κινδύνου) ορίζονται από την εκθετική κατανομή, δηλαδή,

$$H_0(t,\theta) = \theta_0 t$$

ή

$$H_0(t_i,\theta) = \theta t_i,$$
$$h_0(t_i,\theta) = H_0^{'}(t_i,\theta) = \theta,$$
$$\dot{H_0}(t_i,\theta) = t_i$$

και

$$\dot{h_0}(t,\theta) = 1.$$

Για διευκόλυνση, θα χρησιμοποιήσουμε το συμβολισμό,

$$B(t_i,\theta) = \frac{\dot{h_0}(t_i,\theta)}{h_0(t_i,\theta)} + \{ -G'(w) + \frac{G''(w)}{G'(w)} \}\Big|_{w=H_0(t_i,\theta)} \dot{H_0}(t_i,\theta).$$

Για την εκθετική κατανομή η συνάρτηση παίρνει τη μορφή,

$$B(t_i,\theta) = \frac{1}{\theta} + \{ -G'(t_i\theta) + \frac{G''(t_i\theta)}{G'(t_i\theta)} \} \cdot t_i. \tag{132}$$

Έτσι έχουμε γενικά ότι ,

$$\frac{\partial \log \mathscr{L}}{\partial \theta} = \sum_{i=1}^{n} \delta_i B(t_i,\theta) - (1-\delta_i) G'(w)\Big|_{w=H_0(t_i,\theta)} \dot{H_0}(t_i,\theta)$$

και για την εκθετική κατανομή

$$\frac{\partial \log \mathscr{L}}{\partial \theta} = \sum_{i=1}^{n} \delta_i B(t_i,\theta) - (1-\delta_i) G'(w)\Big|_{w=t_i\theta} t_i.$$

4.3 Πληροφοριακός αριθμός Fisher $\mathscr{I}_F(\theta_0)$ για το αρχικό μοντέλο ευπάθειας

Για τον πληροφοριακό αριθμό του Fisher για το αρχικό μοντέλο και για μία παρατήρηση, είδαμε ότι ισχύει η σχέση (120), η οποία με τον κατάλληλο συμβολισμό γράφεται ως,

$$\mathscr{I}_1(\theta_0) = \sum_{\delta_i=0}^{1} \int_0^{\infty} \left(\frac{\partial \log \mathscr{L}_1(t_i, \delta_i, \theta)}{\partial \theta} \right)^2 f(t_i, \theta) f(\delta_i) \Big|_{\theta=\theta_0} dt_i = \tag{133}$$

$$\sum_{\delta_i=0}^{1} \int_0^{\infty} \Big(\delta_i B(t_i, \theta) - (1-\delta_i) G^{'}(H_0(t_i, \theta)) \dot{H_0}(t_i, \theta)) \Big)^2 \cdot f(t_i, \theta) \Big|_{\theta=\theta_0} f(\delta_i)\, dt_i = \tag{134}$$

$$\int_0^{\infty} \left\{ (B(t_i, \theta))^2 p + (-G^{'}(H_0(t_i, \theta)) \dot{H_0}(t_i, \theta))^2 (1-p) \right\} \cdot f(t_i, \theta) \Big|_{\theta=\theta_0} dt_i \tag{135}$$

όπου ορίζουμε την πιθανότητα μη λογοκριμένης παρατήρησης $p = P(\delta_i = 1)$ και αντίστοιχα την πιθανότητα παρατήρησης μιας λογοκριμένης παρατήρησης ως $1 - p = P(\delta_i = 0)$.

Ο πληροφοριακός αριθμός για το αρχικό μοντέλο ευπάθειας για όλες τις παρατηρήσεις είναι,

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

4.3.1 Πληροφοριακός αριθμός του Fisher $\mathscr{I}_F(\theta_0)$ για το αρχικό Γάμμα μοντέλο ευπάθειας

Σύμφωνα με τη Γάμμα κατανομή ευπάθειας $Z \sim \Gamma\left(\frac{1}{k}, k\right)$, όπου η μέση τιμή είναι $E(Z) = 1$ και $Var(Z) = \kappa$, έχουμε ότι,

$$G(w) = \frac{1}{\kappa} \ln(1 + w \cdot \kappa),$$

με παράγωγο,

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

και δεύτερη παράγωγο,

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

Συνεπώς, στην δική μας περίπτωση αρχικά, για την σχέση, $\frac{G''(w)}{G'(w)}\Big|_{w=H_0(t,\theta)}$, έχουμε ότι,

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

και έτσι έχουμε ότι για την περίπτωση του εκθετικού μοντέλου,

$$B(t_i,\theta) = \frac{1}{\theta} + \{-\frac{1}{1+\theta t_i\kappa} - \kappa\cdot\frac{1}{1+\theta t_i\kappa}\}\cdot t_i, \tag{136}$$

ή

$$B(t_i,\theta) = \frac{1}{\theta} - \{\frac{1+\kappa}{1+\theta t_i\kappa}\}\cdot t_i. \tag{137}$$

Άρα, ο πληροφοριακός αριθμός για το Γάμμα μοντέλο ευπάθειας ορίζεται ως,

$$\mathscr{I}_F(\theta_0) = n\mathscr{I}_1(\theta_0) = n\int_0^\infty \Bigg\{ \Bigg(\left(\frac{1}{\theta} - \frac{(1+\kappa)t_i}{1+\kappa\theta t_i}\right)^2 p + \left(\frac{-t_i}{1+\kappa\theta t_i}\right)^2 (1-p)\Bigg) \cdot e^{-\frac{1}{\kappa}\ln(1+\kappa\theta t_i)}\cdot\frac{\theta}{1+\kappa\theta t_i}\Bigg\}\Big|_{\theta=\theta_0} dt_i.$$

4.3.2 Πληροφοριακός αριθμός του Fisher $\mathscr{I}_F(\theta_0)$ για το αρχικό Inverse Gaussian μοντέλο ευπάθειας

Για την Inverse Gaussian κατανομή, δηλαδή $Z \sim \text{InvGaussian}(1, 2b)$, όπου $E(Z) = 1$ και $Var(Z) = \frac{1}{2b}$ έχουμε ότι,

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

και

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

και

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

Άρα,

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

Επίσης,

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

Συνεπώς στην περίπτωσή μας έχουμε,

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

$$\frac{1}{\theta} - \Big(\frac{\sqrt{b}}{\sqrt{b+t_i\theta}} + \frac{\sqrt{b+t_i\theta}}{2(b+t_i\theta)^{3/2}}\Big)t_i =$$

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

Άρα, ο πληροφοριακός αριθμός, για το Inverse Gaussian μοντέλο ευπάθειας ορίζεται ως,

$$\mathscr{I}_F(\theta_0) = n\mathscr{I}_1(\theta_0) = n\int_0^\infty \Bigg\{\Bigg(\Bigg(\frac{1}{\theta} - \Big(\frac{2\sqrt{b(b+t_i\theta)}+1}{2(b+t_i\theta)}\Big)t_i\Bigg)^2 p + \left(-\frac{\sqrt{b}t_i}{\sqrt{b+t_i\theta}}\right)^2 (1-p)\Bigg)e^{-2b+\sqrt{4b(b+t_i\theta)}}\frac{\sqrt{b}\theta}{\sqrt{b+t_i\theta}}\Bigg\}\Big|_{\theta=\theta_0} dt_i.$$

4.4 Πληροφοριακός αριθμός Fisher $I_F(\theta_0)$ για το διακριτοποιημένο μοντέλο ευπάθειας

Στην συνέχεια, για τον υπολογισμό του πληροφοριακού αριθμού του Fisher $I_F(\theta_0)$, για το διακριτό μοντέλο, θα βασιστούμε στη συνάρτηση πιθανοφάνειας του διακριτού μοντέλου. Πιο κάτω περιγράφουμε τη μοντελοποίηση του προβλήματος στην περίπτωση των λογοκριμένων δεδομένων.

Έστω ο διαμερισμός της κατανομής των χρόνων επιβίωσης σε M διαστήματα τα οποία συμβολίζουμε με $E_j, j = 1, .., M$ και πιο συγκεκριμένα έστω η διαμέριση του διαστήματος παρατήρησης $[0, \tau]$ ως εξής $0 = a_0 < a_1 < a_2 < ... < a_M = \tau$ έτσι ώστε $E_j = (a_{j-1}, a_j], j = 1, ..., M$. Έστω r_j τα άτομα τα οποία είναι σε κίνδυνο ακριβώς πριν το διάστημα E_j ή ακριβώς πριν το χρόνο a_{j-1}. Προφανώς $r_1 = n$. Έστω d_j ο αριθμός των ατόμων που αποτυγχάνουν στο διάστημα E_j. Έστω h_j ο ρυθμός κινδύνου στο j διάστημα και αντίστοιχα $1 - h_j$ ο ρυθμός επιβίωσης στο ίδιο διάστημα. Δεδομένης της ιστορίας του τι έχει συμβεί πριν το διάστημα E_j, η τ.μ. d_j ακολουθεί διωνυμική κατανομή στο διάστημα $(a_{j-1}, a_j]$ και πιο συγκεκριμένα

$$P(d_j = x) = \binom{r_j}{x} h_j(\theta)^x (1 - h_j(\theta))^{r_j - x}, \quad x = 1, ..., r_j$$

Λόγω του ότι υποθέτουμε την ύπαρξη λογοκριμένων παρατηρήσεων από δεξιά, η μηδενική υπόθεση του προτεινόμενου ελέγχου γράφεται σε σχέση με την συνάρτηση κινδύνου h στα διαστήματα E_j ως

$$H_0 : \mathbf{h}^0(\theta) = (h_1^0(\theta), ..., h_M^0(\theta))^T.$$

Η συνάρτηση κινδύνου h_j^0 στο διάστημα $(a_{j-1}, a_j]$ για το μοντέλο ευπάθειας είναι

$$h_j^0(\theta) = \frac{S_0(a_{j-1}, \theta) - S_0(a_j, \theta)}{S_0(a_{j-1}, \theta)} = \frac{e^{-G(H_0(a_{j-1}, \theta))} - e^{-G(H_0(a_j, \theta))}}{e^{-G(H_0(a_{j-1}, \theta))}}$$

ενώ ο ρυθμός επιβίωσης είναι $1 - h_j^0(\theta)$.

Η συνάρτηση πιθανοφάνειας για όλα τα διαστήματα, υποθέτοντας ανεξαρτησία ανάμεσα στα διαστήματα δοθέντος του παρελθόντος πριν το εκάστοτε διάστημα, ορίζεται ως

$$L(d_1, d_2, ..., d_M, \theta) = \prod_{i=1}^{M} \binom{r_i}{d_i} (h_i^0(\theta))^{d_i} (1 - (h_i^0(\theta))^{r_i - d_i}$$

Έτσι ο λογάριθμος της πιθανοφάνειας παίρνει την μορφή,

$$\log L(d_1, ..., d_M, \theta) = \sum_{i=1}^{M} \left[\log \binom{r_i}{d_i} + d_i \log(h_i^0(\theta)) + (r_i - d_i) \log(1 - (h_i^0(\theta)) \right]$$

Ο αριθμός Fisher ποσοτικοποιεί τον όγκο των πληροφοριών που παρέχει ένα δείγμα σχετικά με την παράμετρο θ. Είναι γνωστό ότι ο πληροφοριακός

αριθμός Fisher για το θ, ορίζεται ως η μέση τιμή της πλην δεύτερης παραγώγου του λογαρίθμου της συνάρτησης πιθανοφάνειας ως προς θ.

Για τον πληροφοριακό αριθμό του Fisher, για το διακριτό μοντέλο ευπάθειας κάτω από τη μηδενική υπόθεση ισχύει,

$$I_{F_0}(\theta)=E_\theta\left(-\frac{\partial^2}{\partial\theta^2}\log L(d_1,...,d_M,\theta)|_{\theta=\theta_0}\right)=E_\theta\left(\left(\frac{\partial}{\partial\theta}\log L(d_1,...,d_M,\theta)\right)^2|_{\theta=\theta_0}\right)$$

Για τον υπολογισμό του πληροφοριακού αριθμού, έχουμε ότι η μερική παράγωγος,

$$\frac{\partial h_i^0(\theta)}{\partial\theta}=\frac{\partial}{\partial\theta}\left(\frac{S_0(a_{i-1},\theta)-S_0(a_i,\theta)}{S_0(a_{i-1},\theta)}\right)$$

$$=\frac{\partial}{\partial\theta}\left(S_0(a_{i-1},\theta)-S_0(a_i,\theta)\right)\frac{S_0(a_{i-1},\theta)}{S_0(a_{i-1},\theta)^2}-(S_0(a_{i-1},\theta)-S_0(a_i,\theta))\frac{\dot{S}_0(a_{i-1},\theta)}{S_0(a_{i-1},\theta)^2}$$

$$=[\dot{S}_0(a_{i-1},\theta)-\dot{S}_0(a_i,\theta)]\frac{S_0(a_{i-1},\theta)}{S_0(a_{i-1},\theta)^2}-\left(S_0(a_{i-1},\theta)-S_0(a_i,\theta)\right)\frac{\dot{S}_0(a_{i-1},\theta)}{S_0(a_{i-1},\theta)^2}$$

$$=\frac{S_0(a_i,\theta)\dot{S}_0(a_{i-1},\theta)-\dot{S}_0(a_i,\theta)S_0(a_{i-1},\theta)}{S_0(a_{i-1},\theta)^2}$$

$$=\frac{e^{-G(H_0(a_i,\theta))}}{e^{-G(H_0(a_{i-1},\theta))}}(G'(H_0(a_{i-1},\theta))\dot{H}_0(a_{i-1},\theta)-G'(H_0(a_i,\theta))\dot{H}_0(a_i,\theta)) \quad (138)$$

Επίσης,

$$\frac{\partial(1-h_i^0(\theta))}{\partial\theta}=\frac{\partial(h_i^0(\theta))}{\partial\theta}$$

η οποία δίνεται από την σχέση (140).

Παραγωγίζοντας τον λογάριθμο της συνάρτησης πιθανοφάνειας ως προς θ, έχουμε

$$\frac{\partial\log L}{\partial\theta}=\sum_{i=1}^{M}\left[d_i log(\dot{h_i^0}(\theta))+(r_i-d_i)log(1\dot{-}h_i^0(\theta))\right]$$

$$=\sum_{i=1}^{M}\left[d_i\frac{1}{h_i^0(\theta)}\dot{h_i^0}(\theta)-(r_i-d_i)\frac{1}{1-h_i^0(\theta)}\dot{h_i^0}(\theta)\right]$$

$$=\sum_{i=1}^{M}\dot{h_i^0}(\theta)\left[\frac{d_i(1-h_i^0(\theta))-(r_i-d_i)h_i^0(\theta)}{h_i^0(\theta)(1-h_i^0(\theta))}\right]$$

$$= \sum_{i=1}^{M} \dot{h_i^0}(\theta) \frac{d_i - r_i h_i^0(\theta)}{h_i^0(\theta)(1-h_i^0(\theta))}$$

Συνεπώς για να βρεθεί ο πληροφοριακός αριθμός για το διακριτό μοντέλο, αρχικά έχουμε,

$$I_{F_0}(\theta) = E_0 \left\{ \left(\sum_{i=1}^{M} \dot{h_i^0}(\theta) \frac{d_i - r_i h_i^0(\theta)}{h_i^0(\theta)(1-h_i^0(\theta))} \right)^2 \right\} = \sum_{i=1}^{M} E_0 \left(\dot{h_i^0}(\theta) \frac{d_i - r_i h_i^0(\theta)}{h_i^0(\theta)(1-h_i^0(\theta))} \right)^2$$

$$= \sum_{i=1}^{M} \left(\frac{\dot{h_i^0}(\theta)}{h_i^0(\theta)(1-h_i^0(\theta))} \right)^2 E_0 \left(d_i - r_i h_i^0(\theta) \right)^2 =$$

$$= \sum_{i=1}^{M} \left(\frac{\dot{h_i^0}(\theta)}{h_i^0(\theta)(1-h_i^0(\theta))} \right)^2 Var_0(d_i)$$

λόγω ανεξαρτησίας ανάμεσα στα διαστήματα και διότι $E_0(d_i - r_i h_i^0(\theta)) = 0$.

Η αναμενόμενη τιμή E_0 εξαρτάται από την κατανομή των τ.μ. d_i κάτω από την μηδενική υπόθεση. Αφού τα d_i ακολουθούν Διωνυμική κατανομή $b(r_i, h_i^0(\theta))$ τότε $E_0[d_i] = r_i h_i^0(\theta)$ και $Var_0(d_i) = r_i(h_i^0(\theta)(1-h_i^0(\theta))$ και την αντικαθιστούμε πιο πάνω για τον υπολογισμό του πληροφοριακού αριθμού Fisher ο οποίος γίνεται

$$\sum_{i=1}^{M} \left(\frac{\dot{h_i^0}(\theta)}{h_i^0(\theta)(1-h_i^0(\theta))} \right)^2 r_i h_i^0(\theta)(1-h_i^0(\theta))$$

και τελικά ο πληροφοριακός αριθμός Fisher για το διακριτό μοντέλο ευπάθειας ισούται με

$$I_{F_0}(\theta) = I_F(\theta_0) = \sum_{i=1}^{M} \frac{r_i(\dot{h_i^0}(\theta))^2)}{h_i^0(\theta)(1-h_i^0(\theta))} \tag{139}$$

Σε σχέση με τη συνάρτηση G ο πληροφοριακός αριθμός γράφεται από την σχέση (140)

$$\sum_{i=1}^{M} r_i \frac{(\frac{e^{-2G(H_0(a_i,\theta))}}{e^{-2G(H_0(a_{i-1},\theta))}}(G'(H_0(a_{i-1},\theta))\dot{H}_0(a_{i-1},\theta) - G'(H_0(a_i,\theta))\dot{H}_0(a_i,\theta))^2)}{\left(1 - \frac{e^{-G(H_0(a_i,\theta))}}{e^{-G(H_0(a_{i-1},\theta))}}\right) \frac{e^{-G(H_0(a_i,\theta))}}{e^{-G(H_0(a_{i-1},\theta))}}} =$$

$$\sum_{i=1}^{M} r_i \frac{\frac{e^{-G(H_0(a_i,\theta))}}{e^{-G(H_0(a_{i-1},\theta))}}(G'(H_0(a_{i-1},\theta)\dot{H}_0(a_{i-1},\theta) - G'(H_0(a_i,\theta)\dot{H}_0(a_i,\theta))^2}{\left(1 - \frac{e^{-G(H_0(a_i,\theta))}}{e^{-G(H_0(a_{i-1},\theta))}}\right)} \tag{140}$$

Για το εκθετικό μοντέλο ο παραπάνω πληροφοριακός αριθμός απλοποιείται σε

$$\sum_{i=1}^{M} r_i \frac{\frac{e^{-G(a_i\theta)}}{e^{-G(a_{i-1}\theta)}}(G'(a_{i-1}\theta)a_{i-1} - G'(a_i\theta)a_i)^2}{\left(1 - \frac{e^{-G(a_i\theta)}}{e^{-G(a_{i-1}\theta)}}\right)}\Big|_{\theta=\theta_0}$$

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

Για το Γάμμα μοντέλο ευπάθειας έχουμε ήδη δει ότι

$$G(w) = \frac{1}{\kappa}\ln(1 + w \cdot \kappa),$$

με παράγωγο,

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

και

$$S(t) = [1 + \kappa H_0(t,\theta)]^{-1/\kappa}$$

οπότε η σχέση (142) γίνεται

$$\sum_{i=1}^{M} r_i \frac{\frac{\{[1+\kappa H_0(a_i,\theta)]^{-1/\kappa}}{[1+\kappa H_0(a_{i-1},\theta)]^{-1/\kappa}}\left(\frac{\dot{H}_0(a_{i-1},\theta)}{1+H_0(a_{i-1},\theta)\kappa} - \frac{\dot{H}_0(a_i,\theta)}{1+H_0(a_i,\theta)\kappa}\right)^2}{\left(1 - \frac{[1+\kappa H_0(a_i,\theta)]^{-1/\kappa}}{[1+\kappa H_0(a_{i-1},\theta)]^{-1/\kappa}}\right)} \tag{141}$$

ενώ για το εκθετικό μοντέλο έχουμε

$$\sum_{i=1}^{M} r_i \frac{\frac{\{[1+\kappa a_i\theta)]^{-1/\kappa}}{[1+\kappa a_{i-1}\theta)]^{-1/\kappa}}\left(\frac{a_{i-1}}{1+a_{i-1}\theta\kappa} - \frac{a_i}{1+a_i\theta\kappa}\right)^2}{\left(1 - \frac{[1+\kappa a_i\theta)]^{-1/\kappa}}{[1+\kappa a_{i-1}\theta)]^{-1/\kappa}}\right)}\Big|_{\theta=\theta_0} \tag{142}$$

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

Για το Inverse Gaussian μοντέλο ξέρουμε ήδη ότι

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

άρα,

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

και η συνάρτηση επιβίωσης είναι,

$$S(t) = e^{2b-\sqrt{4b(b+H_0(t_i,\theta))}}.$$

Σε σχέση με την συνάρτηση G η σχέση (142) γίνεται

$$\sum_{i=1}^{M} r_i \frac{\frac{e^{2b-\sqrt{4b(b+H_0(a_i,\theta))}}}{e^{2b-\sqrt{4b(b+H_0(a_{i-1},\theta))}}} \left(\frac{\dot{H}_0(a_{i-1},\theta)\sqrt{b}}{\sqrt{b+H_0(a_{i-1},\theta)}} - \frac{\dot{H}_0(a_i,\theta))\sqrt{b}}{\sqrt{b+H_0(a_i,\theta)}} \right)^2}{\left(1 - \frac{e^{2b-\sqrt{4b(b+H_0(a_i,\theta))}}}{e^{2b-\sqrt{4b(b+H_0(a_{i-1},\theta))}}}\right)} \tag{143}$$

Για το εκθετικό μοντέλο ο παραπάνω πληροφοριακός αριθμός απλοποιείται σε

$$\sum_{i=1}^{M} r_i \frac{\frac{e^{2b-\sqrt{4b(b+a_i\theta)}}}{e^{2b-\sqrt{4b(b+a_{i-1}\theta)}}} \left(\frac{(a_{i-1}\sqrt{b}}{\sqrt{b+a_{i-1}\theta}} - \frac{a_i\sqrt{b}}{\sqrt{b+a_i\theta}} \right)^2}{\left(1 - \frac{e^{2b-\sqrt{4b(b+a_i,\theta)}}}{e^{2b-\sqrt{4b(b+a_{i-1}\theta)}}}\right)}\Bigg|_{\theta=\theta_0}. \tag{144}$$

4.5 Υπολογισμός Κρίσιμης τιμής του χωρίου απορρίψεως του ελέγχου

Σε αυτή την παράγραφο θα ασχοληθούμε με τον υπολογισμό της κρίσιμης τιμής c_a που εμπλέκεται στο χωρίο απορρίψεως του ελέγχου που ορίστηκε στη σχέση (126) και ορίζεται ως

$$T_n^{\phi}(\hat{\mathbf{h}}, \mathbf{h}(\hat{\theta}_\phi)) \geq c_a.$$

Η σταθερά θα υπολογιστεί από την ασυμπτωτική κατανομή της ελεγχοσυνάρτησης κάτω από την μηδενική υπόθεση για δεδομένο επίπεδο σημαντικότητας

α. Για α=5% π.χ., η κρίσιμη τιμή $c_{0.05}$ θα είναι το 95ο ποσοστημόριο της ασυμπτωτικής κατανομής της ελεγχοσυνάρτησης.

Η ασυμπτωτική κατανομή της ελεγχοσυνάρτησης που δίνεται στη σχέση (128) είναι

$$X^2_{M-1} + (1-\lambda)Z^2$$

δηλαδή

$$X^2_{M-1} + (1-\lambda)X^2_1. \tag{145}$$

Βλέπουμε λοιπόν ότι η ασυμπτωτική της κατανομή αποτελείται από ένα γραμμικό συνδυασμό X^2 τυχαίων μεταβλητών ο οποίος έχει εξεταστεί και χρησιμοποιηθεί στην εργασία Σοφοκλέους και Βόντα (2023). Παραθέτουμε τις λεπτομέρειες πιο κάτω.

4.5.1 Κατανομή αθροίσματος ανεξάρτητων Γάμμα κατανεμημένων τυχαίων μεταβλητών

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

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

Ο Moschopoulos (1985), και πριν από αυτόν ο Mathai (1982), ερεύνησε την κατανομή της συνάρτησης $Y = X_1 + X_2 + \ldots + X_n$ προσεγγίζοντας την κατανομή του αθροίσματος Γάμμα κατανομών μέσω της ροπογεννήτριας συνάρτησης m.g.f. (moment generating function). Συγκεκριμένα, λόγω του ότι οι τ.μ. X_i είναι ανεξάρτητες, η συνάρτηση ροπογεννήτριας της τ.μ. Y δίνεται από τον τύπο:

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

Υποθέτουμε χωρίς βλάβη της γενικότητας ότι $\beta_1 = \min(\beta_i)$. Ο όρος $(1-\beta_i t)$ λοιπόν παίρνει την μορφή,

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

Λογαριθμίζοντας την ροπογεννήτρια $M(t)$ έχουμε ότι:

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

όπου,

$$C = \prod_{i=1}^{n} \left(\frac{\beta_1}{\beta_i} \right)^{a_i}$$

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

$$\rho = \sum_{i=1}^{n} a_i > 0.$$

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

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

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

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

Θέτοντας

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

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

$$\delta_{k+1} = \frac{1}{1+k} \sum_{i=1}^{k+1} i \gamma_i \delta_{k+1-i}, \quad k = 0, 1, 2, \ldots \tag{146}$$

και $\delta_0 = 1$.

Θεώρημα 4.1. Για $X_i,\ i = 1, 2, \ldots, n$ ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν Γάμμα κατανομή με παραμέτρους $a_i > 0$ και $\beta_i > 0$ αντίστοιχα, η συνάρτηση πυκνότητας της τυχαίας μεταβλητής $Y = X_1 + X_2 + \ldots + X_n$ μπορεί να εκφραστεί ως εξής:

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

με ρ, δ_k, C όπως ορίστηκαν παραπάνω.

Η συνάρτηση κατανομής λοιπόν της τ.μ. Y είναι

$$F(w) = P(Y \leq w)$$

λαμβάνει την παρακάτω μορφή:

$$F(w) = C \sum_{k=0}^{\infty} \delta_k \int_{k=0}^{w} \frac{y^{\rho+k-1} \cdot e^{-\frac{y}{\beta_1}}}{\Gamma(\rho+k) \cdot \beta_1^{\rho+k}} dy, \quad w \in R. \tag{147}$$

4.5.2 Εφαρμογή του Θεωρήματος για την ασυμπτωτική κατανομή της ελεγχοσυνάρτησης του προτεινόμενου ελέγχου

Για την κατανομή που ορίζεται στη σχέση (146) έχουμε $n = 2$ αφού εμπλέκονται δύο X^2 τ.μ. Η κατανομή X^2_{M-1} είναι $\Gamma(\frac{M-1}{2}, 2)$ ενώ η X^2_1 είναι $\Gamma(\frac{1}{2}, 2)$.

Επίσης από ιδιότητες της Γάμμα κατανομής ισχύει ότι για

$$c' > 0 : c' \cdot \Gamma(\kappa_1 : shape, \kappa_2 : scale) \equiv \Gamma(\kappa_1, c' \cdot \kappa_2).$$

Έτσι η ασυμπτωτική κατανομή (146), μπορεί να γραφτεί στη μορφή αθροίσματος δύο Γάμμα κατανεμημένων ανεξάρτητων τυχαίων μεταβλητών :

$$X^2_{M-1} + (1-\lambda)Z^2 \equiv \Gamma(\frac{M-1}{2}, 2) + \Gamma(\frac{1}{2}, (1-\lambda) \cdot 2).$$

Με βάση τη θεωρία που παραθέτουμε στην προηγούμενη παράγραφο και το θεώρημα 4.1, οι παράμετροι του θεωρήματος διαμορφώνονται ως εξής, θεωρώντας $\beta_1 = (1-\lambda) \cdot 2 < \beta_2 = 2$:

$$(a_1, a_2) = (1/2, (M-1)/2)$$

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

Η σταθερά C ισούται με

$$C = (1-\lambda)^{\frac{M-1}{2}}.$$

Η σταθερά ρ ισούται με

$$\rho = \frac{M}{2}.$$

Οι σταθερές γ_k ισούνται με

$$\gamma_k = \frac{M-1}{2}\frac{\lambda^k}{k}, \quad k = 1, 2...$$

Οι σταθερές δ_k που υπολογίζονται από τον αναδρομικό τύπο (147) ισούνται με

$$\delta_0 = 1, \delta_1 = \gamma_1\delta_0, \delta_2 = \frac{1}{2}(\gamma_1\delta_1 + \gamma_2\delta_0), ...$$

Από τη σχέση (148), έχοντας γνωστές όλες τις σταθερές, λύνουμε ως προς το 95ο ποσοστημόριο της κατανομής $w_{0.95}$ από τη σχέση

$$F(w_{0.95}) = C\sum_{k=0}^{\infty}\delta_k \int_{k=0}^{w_{0.95}} \frac{y^{\rho+k-1}\cdot e^{-\frac{y}{\beta_1}}}{\Gamma(\rho+k)\cdot\beta_1^{\rho+k}}dy = 0.95. \tag{148}$$

Η κρίσιμη τιμή του ελέγχου

$$c_{0.05} = w_{0.95}.$$

Με βάση αυτές τις παραμέτρους και θεωρία διαμορφώθηκε και ο κώδικας στην R που αναπτύχθηκε αρχικά στην εργασία Σοφοκλέους και Βόντα (2023) και τροποποιήθηκε κατάλληλα στην παρούσα εργασία για να γενικευθεί η θεωρία στην περίπτωση των λογοκριμένων παρατηρήσεων. Ως λ στον κώδικα χρησιμοποιήσαμε την μέση τιμή των λ_i που βρέθηκαν μέσω των προσομοιώσεων (για κάθε τιμή της ελεγχοσυνάρτησης T_n που προσομοιώσαμε προφανώς προκύπτει ένα λ_i). Για δεδομένο M και λ, λοιπόν, βρίσκουμε από την συνάρτηση κατανομής της ασυμπτωτικής κατανομής της ελεγχοσυνάρτησης που δίνεται από την (149), την τιμή του 95ου ποσοστημορίου που αποτελεί και την κρίσιμη τιμή του χωρίου απορρίψεως. Πρόσθετες λεπτομέρειες θα δοθούν στην επόμενη παράγραφο.

4.6 Προσομοιώσεις

Στόχο μας πλέον αποτελεί ο έλεγχος της συμπεριφοράς του προτεινόμενου ελέγχου καλής προσαρμογής, που βασίζεται στην ελεγχοσυνάρτηση $T_n{}^{\phi}$, στην

περίπτωση που υπάρχουν λογοκριμένα και μη δεδομένα. Συγκεκριμένα εστιάζουμε στην περίπτωση των μοντέλων ευπάθειας, εξετάζοντας το μέγεθος του ελέγχου για επίπεδο σημαντικότητας 5%. Η ασυμπτωτική κατανομή της ελεγχοσυνάρτησης δίνεται από τις σχέσεις Morales. Το κρίσιμο σημείο για $a = 5\%$ θα βρεθεί από αυτήν την κατανομή.

Στην συνέχεια θα δώσουμε τις λεπτομέρειες των προσομοιώσεων μας. Για αρχή αναφέρουμε ότι προσομοιώσαμε δεδομένα για διάφορα μεγέθη δείγματος και συγκεκριμένα για $n = 60, 120, 240$ έτσι ώστε να εξετάσουμε μικρά, μεσαία και μεγάλα δείγματα. Θα γίνει χρήση της εντολής $runif(n)$, για την ομοιόμορφη κατανομή $U(0, 1)$, ώστε να δημιουργήσουμε πιθανότητες επιβίωσης $S(t_i), i = 1, ..., n$ για διάφορα μεγέθη δειγμάτων n. Σκοπός είναι να δημιουργήσουμε n χρόνους επιβίωσης οι οποίοι θα ακολουθούν το Γάμμα ή Inverse Gaussian μοντέλο ευπάθειας.

4.6.1 Γάμμα μοντέλο ευπάθειας

Όπως αναφέρθηκε στο Κεφάλαιο 2 έχουμε ότι ισχύει για την Γάμμα κατανομή, $\Gamma(a, \kappa)$,

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

και

$$S(t) = e^{-G(H_0(t))} \iff -lnS(t) = G(H_0(t)),$$

άρα,

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

$$\implies -lnS(t) = -ln(1 + \theta \cdot t \cdot \kappa)^{-a}$$

$$\implies S(t) = (1 + \theta \cdot t \cdot \kappa)^{-1/\kappa}.$$

όπου $a = 1/\kappa$.

Λύνοντας λοιπόν ως προς τον χρόνο t, βρίσκουμε για το Γάμμα μοντέλο ευπάθειας ότι,

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

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

Στη συνέχεια θα θεωρήσουμε ότι η βασική συνάρτηση κινδύνου δίνεται από την εκθετική κατανομή $Exp(1)$, δηλαδή η πραγματική τιμή της παραμέτρου $\theta = 1$. Θα γίνει επίσης η υπόθεση ότι η ευπάθεια ακολουθεί Γάμμα κατανομή με $a = 1/\kappa$ για τις τιμές του $\kappa = 0.5, 1.5$. Εφόσον η παράμετρος κ είναι η διασπορά της ευπάθειας η επιλογή αυτών των τιμών έγινε για να εξετάσουμε την περίπτωση μικρής αλλά και μεγάλης διασποράς. Η παράμετρος κ θα θεωρηθεί γνωστή και η παράμετρος $\theta = 1$.

Στη συνέχεια ομαδοποιούμε τους χρόνους επιβίωσης που δημιουργήσαμε σε M διαστήματα. Εξετάστηκε μεγάλος αριθμός διαστημάτων, αλλά επειδή η τιμή του M εξαρτάται από το μέγεθος του δείγματος n, για τη βελτιστοποίηση του ελέγχου χρησιμοποιήθηκε διαφορετική τιμή για το M, ανάλογα με την κάθε περίπτωση που εξετάζαμε.

Επίσης σημαντικό είναι να αναφερθεί, επειδή μελετάμε δεδομένα με λογοκρισία, ορίστηκε και μια αντίστοιχη μεταβλητή, censoring rate $= 10\%, 30\%, 50\%$, με στόχο τη μελέτη διαφορετικών ποσοστών λογοκρισίας για την κάθε περίπτωση.

4.6.2 Inverse Gaussian μοντέλο ευπάθειας

Για την κατανομή Inverse Gaussian, η συνάρτηση G ορίζεται ως,

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

Η μέση τιμή της ευπάθειας είναι $\mu = 1$, $\lambda = 2b$ και η διασπορά της είναι ίση με $\frac{1}{2b}$, η συνάρτηση ευπάθειας λαμβάνει την μορφή,

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

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

Λαμβάνοντας υπόψιν την αθροιστική συνάρτηση κινδύνου, έχουμε

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

$$\implies S(t) = \exp\left(2b - \sqrt{4b(b + t \cdot \theta)}\right).$$

Έτσι λύνοντας ως προς t, έχουμε,

$$t = \frac{\ln S(t) \cdot (\ln S(t) - 4b)}{4 \cdot b \cdot \theta}.$$

Θα ερευνήσουμε το μοντέλο αυτό για δύο τιμές της διασποράς ευπάθειας όπως και στο Γάμμα μοντέλο. Για σκοπούς σύγκρισης, αφού λοιπόν ερευνούμε την περίπτωση όπου η διασπορά της Γάμμα ευπάθειας είναι ίση με 0.5 και αφού $Var(x) = \frac{1}{2b}$ η πρώτη τιμή του b που θα επιλέξουμε είναι 1. Επίσης επιλέξαμε και τις τιμές $b = 0.5$ και $b = 3$.

4.6.3 Έλεγχος καλής προσαρμογής και για τα δύο μοντέλα

Έχοντας δημιουργήσει τα δεδομένα μας, στη συνέχεια ομαδοποιούμε τους χρόνους επιβίωσης που δημιουργήσαμε σε M διαστήματα που περιέχουν το ίδιο πλήθος παρατηρήσεων.

Στην ελεγχοσυνάρτηση $T_n^{\phi}(\hat{h}, h^0)$, πρέπει να βρούμε το διάνυσμα της συνάρτησης κινδύνου h^0 και $\hat{h}$ για να υπολογίσουμε την τιμή της ελεγχοσυνάρτησης. Για να βρεθεί το $\hat{h}$ θα χρησιμοποιήσουμε την Nelson-Aalen εκτιμήτρια της συνάρτησης κινδύνου η οποία δίνεται από την σχέση,

$$\hat{h}_i = d_i/r_i, \quad i = 1, ..., M$$

όπου r_i ο αριθμός των ατόμων που είναι σε κίνδυνο λίγο πριν το i διάστημα και d_i ο αριθμός των αποτυχιών στο i διάστημα.

Για να υπολογίσουμε το υποθετικό μοντέλο h_j^0 θα χρειαστεί να χρησιμοποιήσουμε τον τύπο, όπως ορίστηκε και προηγουμένως,

$$h_j^0(\theta) = \frac{S_0(a_{j-1}, \theta) - S_0(a_j, \theta)}{S_0(a_{j-1}, \theta)},$$

για το Γάμμα και για το Inverse Gaussian μοντέλο ευπάθειας.

Για δεδομένη διαμέριση των χρόνων επιβίωσης στο δείγμα σε M διαστήματα και δεδομένη παράμετρο διασποράς κ για το Γάμμα μοντέλο ή αντίστοιχα δεδομένη παράμετρο b για το Inverse Gaussian μοντέλο, χρειάζεται να εκτιμήσουμε την παράμετρο θ_0 για να υπολογίσουμε το διάνυσμα των συναρτήσεων κινδύνου h_j^0. Αυτό επιτυγχάνεται μέσω της μεθόδου μέγιστης πιθανοφάνειας. Θυμίζουμε ότι συνάρτηση πιθανοφάνειας που χρησιμοποιήθηκε έχει την μορφή,

$$\mathscr{L} = \prod_{i=1}^{n} \left[\left(h_0(t_i, \theta) e^{-G(H_0(t_i,\theta))} G'(w) \Big|_{w=H_0(t_i,\theta)} \right)^{\delta_i} \cdot \left(e^{-G(H_0(t_i,\theta))} \right)^{1-\delta_i} \right]$$

με λογάριθμο,

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

Με τη χρήση του αλγορίθμου nlm (Non-Linear Minimization) κάνουμε ελαχιστοποίηση της συνάρτησης $-log(\mathscr{L})$ για να βρούμε την εκτίμηση της παραμέτρου θ την οποία συμβολίζουμε ως $\hat{\theta}$.

Έτσι θα μπορέσουμε να υπολογίσουμε την τιμή της ελεγχοσυνάρτησης $T_n^{\phi}(\hat{h}, h^0)$. Ορίζουμε στην R την συνάρτηση $Tnkull$ με βάση τον τύπο,

$$T_n^{\phi}(\hat{\mathbf{h}}, \mathbf{h}^0(\theta)) = \frac{2}{\phi''(1)} \sum_{i=1}^{M} r_i \left\{ h_i^0(\theta) \phi \left(\frac{\hat{h}_i}{h_i^0(\theta)} \right) + (1 - h_i^0(\theta)) \phi \left(\frac{1 - \hat{h}_i}{1 - h_i^0(\theta)} \right) \right\}$$

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

Σκοπός μας είναι να εξετάσουμε το μέγεθος του ελέγχου σε επίπεδο σημαντικότητας 5%. Για να επιτευχθεί αυτό θα χρειαστεί να τρέξουμε τον παραπάνω αλγόριθμο αρκετές φορές, έτσι ώστε να βρούμε την εμπειρική κατανομή της ελεγχοσυνάρτησης $T_n^{\phi}(\hat{h}, h^0)$ και να μπορούμε να την συγκρίνουμε με την ασυμπτωτική κατανομή $X_{M-1}^2 + (1-\lambda)Z^2$ κυρίως μέσω των 95ου ποσοστημορίου της. Στο πρόγραμμα, επαναλαμβάνουμε τον παραπάνω αλγόριθμο 10000 φορές, παίρνοντας 10000 τιμές για την ελεγχοσυνάρτηση.

Έτσι μπορούμε να εκτιμήσουμε, με βάση τα δείγματα αυτά, την κατανομή της ελεγχοσυνάρτησης και συγκεκριμένα να βρούμε το 95ο ποσοστημόριό της, καθώς και την τιμή για το 95ο εμπειρικό ποσοστημόριο της εμπειρικής κατανομής της ελεγχοσυνάρτησης. Το ποσοστημόριο p_a είναι το σημείο της κατανομής για το οποίο το $a\%$ των παρατηρήσεων είναι μικρότερες ή ίσες από αυτό και το υπόλοιπο $(1-a)\%$ των παρατηρήσεων είναι μεγαλύτερες ή ίσες από αυτό.

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

4.7 Ασυμπτωτική κατανομή της ελεγχοσυνάρτησης

Σε αυτό το σημείο, για να βρούμε την ασυμπτωτική κατανομή της ελεγχοσυνάρτησης για τις διάφορες περιπτώσεις, θα χρειαστεί να υπολογίσουμε τον πληροφοριακό αριθμό του Fisher, για το πραγματικό-αρχικό μοντέλο $I_F(\theta_0)$ και για το διακριτό μοντέλο $\mathscr{I}_F(\theta_0)$. Σε μοντέλο με κατανομή ευπάθειας Γάμμα, οι πληροφοριακοί αριθμοί δίνονται από τους παρακάτω τύπους,

$$\mathscr{I}_F(\theta_0) = n\mathscr{I}_1(\theta_0) = n\int_0^\infty \left\{ \left(\left(\frac{1}{\theta} - \frac{(1+\kappa)t_i}{1+\kappa\theta t_i}\right)^2 p + \left(\frac{-t_i}{1+\kappa\theta t_i}\right)^2 (1-p)\right) \cdot e^{-\frac{1}{\kappa}\ln(1+\kappa\theta t_i)} \cdot \frac{\theta}{1+\kappa\theta t_i} \right\}\Big|_{\theta=\theta_0} dt_i.$$

και

$$\sum_{i=1}^{M} r_i \frac{\frac{\{[1+\kappa H_0(a_i,\theta)]^{-1/\kappa}}{[1+\kappa H_0(a_{i-1},\theta)]^{-1/\kappa}} \left(\frac{\dot{H}_0(a_{i-1},\theta)}{1+H_0(a_{i-1},\theta)\kappa} - \frac{\dot{H}_0(a_i,\theta)}{1+H_0(a_i,\theta)\kappa}\right)^2}{\left(1 - \frac{[1+\kappa H_0(a_i,\theta)]^{-1/\kappa}}{[1+\kappa H_0(a_{i-1},\theta)]^{-1/\kappa}}\right)}.$$

Αντίστοιχα σε μοντέλο με κατανομή ευπάθειας Inverse Gaussian, οι πληροφοριακοί αριθμοί δίνονται από,

$$\mathscr{I}_F(\theta_0) = n\mathscr{I}_1(\theta_0) = n\int_0^\infty \left\{ \left(\left(\frac{1}{\theta} - \Big(\frac{2\sqrt{b(b+t_i\theta)}+1}{2(b+t_i\theta)}\Big)t_i\right)^2 p + \left(-\frac{\sqrt{b}t_i}{\sqrt{b+t_i\theta}}\right)^2 (1-p)\right) e^{-2b+\sqrt{4b(b+t_i\theta)}} \frac{\sqrt{b}\theta}{\sqrt{b+t_i\theta}} \right\}\Big|_{\theta=\theta_0} dt_i$$

και

$$\sum_{i=1}^{M} r_i \frac{\frac{e^{2b-\sqrt{4b(b+H_0(a_i,\theta))}}}{e^{2b-\sqrt{4b(b+H_0(a_{i-1},\theta))}}} \left(\frac{\dot{H}_0(a_{i-1},\theta)\sqrt{b}}{\sqrt{b+H_0(a_{i-1},\theta)}} - \frac{\dot{H}_0(a_i,\theta))\sqrt{b}}{\sqrt{b+H_0(a_i,\theta)}}\right)^2}{\left(1 - \frac{e^{2b-\sqrt{4b(b+H_0(a_i,\theta))}}}{e^{2b-\sqrt{4b(b+H_0(a_{i-1},\theta))}}}\right)}.$$

Η τιμή του λ, $0 \leq \lambda \leq 1$, χρειάζεται για να βρεθεί η ασυμπτωτική κατανομή της ελεγχοσυνάρτησης, είναι η λύση της εξίσωσης,

$$\lambda = \frac{\mathscr{I}_F(\theta_0)}{I_F(\theta_0)}.$$

Για να βρούμε την τιμή του ποσοστημορίου p_{95} θα χρειαστεί η συνάρτηση κατανομής της ασυμπτωτικής κατανομής. Την κατανομή αυτή θα την υπολογίσουμε με την βοήθεια της εργασίας Moschopoulos (1985), ο οποίος ερεύνησε την κατανομή γραμμικού συνδυασμού πολλών Γάμμα ανεξάρτητων τυχαίων μεταβλητών. Μια κατανομή chi-square $X \sim \chi^2_k$ είναι ειδική περίπτωση της οικογένειας κατανομών Γάμμα.

$$\chi^2_k \sim \Gamma\Big(\frac{k}{2}, 2\Big),$$

όπου $\Gamma(\frac{\kappa}{2}, 2)$, συμβολίζει τη κατανομή Γάμμα με παραμέτρους shape $\frac{\kappa}{2}$ και scale 2 αντίστοιχα. Επίσης από ιδιότητες της Γάμμα κατανομής ισχύει ότι για $c > 0$,

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

Έτσι η ασυμπτωτική κατανομή του μοντέλου μας, μπορεί να γραφτεί στη μορφή γραμμικού συνδυασμού δύο Γάμμα κατανεμημένων ανεξάρτητων τυχαίων μεταβλητών,

$$X^2_{M-1} + (1-\lambda)Z^2 \ \sim \ \Gamma\Big(\tfrac{M-1}{2}, 2\Big) \ + \ \Gamma\Big(\tfrac{1}{2}, (1-\lambda)\cdot 2\Big).$$

Η τυχαία μεταβλητή $Z^2 (Z \sim N(0,1))$, ακολουθεί Chi-square κατανομή με βαθμό ελευθερίας 1 (Q^2_1).

Βάση όλων αυτών γράφτηκε αντίστοιχος κώδικας στην R, με τις παραμέτρους,

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

και

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

και ως λ χρησιμοποιήσαμε την μέση τιμή των λ_i που βρήκαμε (για κάθε τιμή της ελεγχοσυνάρτησης T_n που προσομοιώσαμε βρήκαμε ένα λ). Για δεδομένο M και λ, λοιπόν, βρίσκουμε από την συνάρτηση κατανομής της ασυμπτωτικής κατανομής $\mathbf{F}$, την τιμή του ποσοστημορίου p_{95}, δηλαδή

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

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

Στην συνέχεια παραθέτουμε ορισμένους πίνακες με τα αποτελέσματα (σφάλμα τύπου Ι) που λάβαμε για διάφορες τιμές των παραμέτρων κ και b (οι οποίες σχετίζονται με τη διασπορά της ευπάθειας) , M (αριθμός διαστημάτων που λαμβάνεται στη διαμέριση) και n (αριθμός παρατηρήσεων για κάθε προσομοίωση).

4.8 Πίνακες Προσομοιώσεων

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

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

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

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

Κατανομή ευπάθειας: Γάμμα, Σφάλμα τύπου Ι

Για : censoring rate $= 0.1$, $n = 60$, $\kappa = 0.5$ και $\theta = 1$.

κ	**8**		**10**	
	P_t	mean(λ)	P_t	mean(λ)
1	0.0455	0.003814147	0.0498	0.003242072
2	0.0463	0.003765788	0.0454	0.003251255
3	0.0491	0.003768781	0.0494	0.003251209
4	0.0434	0.003796493	0.0463	0.003254323
5	0.0445	0.003772314	0.0483	0.003253128
6	0.0497	0.003821227	0.04610461	0.003262096
7	0.0457	0.003811730	0.0483	0.003264559
8	0.0479	0.003815094	0.0459	0.003257431
9	0.0472	0.003821007	0.0496	0.003220024
10	0.0474	0.003801434	0.04900490	0.003294788
11	0.0478	0.003813399	0.04920492	0.003274849
12	0.0482	0.003813057	0.0491	0.003268075
13	0.0436	0.003772099	0.0475	0.003294603
14	0.0492	0.003803644	0.0465	0.003240844
15	0.0492	0.003795663	0.0487	0.003260833
Mean	0.04698	0.003799059	0.04794096	0.003259339

Mean ($\hat{\theta}$)	0.9314766	0.9320591
p_{95}	15.49996	18.30108
$\mathbf{p_{95}}$	15.29811	18.15492

Για μέγεθος δείγματος $M = 60$, ποσοστό λογοκρισίας 10% και διασπορά της ευπάθειας $\kappa = 0.5$, το μέγεθος του ελέγχου για $M = \sqrt{n} = 8$ ή $M = \sqrt{2n} = 10$ είναι περίπου 0.05 όπως θα θέλαμε και αποδεχόμαστε την μηδενική υπόθεση 95% περίπου των φορών. Φαίνεται ότι η κατανομή της ελεγχοσυνάρτησης T_n περιγράφεται καλά από την ασυμπτωτική κατανομή. Επειδή θέλουμε το ποσοστό όπου το πλήθος των τιμών της ελεγχοσυνάρτησης υπερβαίνει το κρίσιμο σημείο της ασυμπτωτικής κατανομής να προσεγγίζει το 0.05, επιλέγουμε στον έλεγχο αυτό το μέγεθος M να ισούται με 8 με το κρίσιμο σημείο της ασυμπτωτικής κατανομής p_{95} να ισούται προσεγγιστικά με 15.49996 ενώ το 95ο ποσοστημόριο

της ελεγχοσυνάρτησης T_n, $\mathbf{p_{95}}$, ισούται με 15.29811. Βλέπουμε πως οι δύο τιμές είναι σχεδόν πανομοιότυπες, με την εμπειρική τιμή της T_n να υποδηλώνει ότι η ασυμπτωτική κατανομή προσεγγίζει καλά την πραγματική κατανομή της ελεγχοσυνάρτησης T_n.

Για : censoring rate $= 0.1$, $n = 120$, $\kappa = 0.5$ και $\theta = 1$.

κ	**10**		**12**	
	P_t	mean(λ)	P_t	mean(λ)
1	0.0424	0.002617523	0.0455	0.002332015
2	0.0484	0.002581351	0.0432	0.002350669
3	0.0394	0.002596944	0.0476	0.002338110
4	0.0464	0.002593595	0.0440	0.002345793
5	0.0432	0.002606240	0.0391	0.002326390
6	0.0466	0.002591587	0.0465	0.002353561
7	0.0419	0.002590753	0.0430	0.002363233
8	0.0439	0.002597697	0.0467	0.002377168
9	0.0448	0.002621077	0.0452	0.002355124
10	0.0414	0.002594417	0.0444	0.002343463
11	0.0431	0.002578976	0.0464	0.002350680
12	0.0447	0.002598258	0.0440	0.002370521
13	0.0419	0.002604297	0.0450	0.002352158
14	0.0446	0.002597684	0.0425	0.002347520
15	0.0457	0.002600383	0.0501	0.002330417
Mean	0.04389333	0.002598052	0.04488	0.002349122

Mean ($\hat{\theta}$)	0.9387483	0.9389376
p_{95}	18.30228	21.02195
$\mathbf{p_{95}}$	17.85875	20.64801

Εδώ σημειώνουμε ότι για $M = 12$, ο έλεγχος δίνει μέγεθος $P_t = 0.04488$.

Για : censoring rate $= 0.1$, $n = 240$, $\kappa = 0.5$ και $\theta = 1$.

κ	**M = 10**		**M = 12**		**M = 20**	
	P_t	mean(λ)	P_t	mean(λ)	P_t	mean(λ)
1	0.0471	0.001173969	0.0465	0.001053144	0.0480	0.0007584
2	0.0471	0.001175554	0.0428	0.001053246	0.0498	0.0007541
3	0.0471	0.001173775	0.0449	0.001050554	0.0503	0.0007559
4	0.0426	0.001172100	0.0428	0.001055851	0.0525	0.0007576
5	0.0437	0.001175223	0.0418	0.001052647	0.0477	0.0007555
6	0.0427	0.001180703	0.0463	0.001051489	0.0512	0.0007545
7	0.0453	0.001171043	0.0459	0.001052209	0.0481	0.0007536
8	0.0440	0.001170670	0.0461	0.001053140	0.0466	0.0007582
9	0.0499	0.001177104	0.0429	0.001051109	0.0518	0.0007552
10	0.0466	0.001171770	0.0473	0.001054327	0.0501	0.0007531
11	0.0457	0.001173148	0.0428	0.001058686	0.0478	0.0007528
12	0.0449	0.001176718	0.0467	0.001044621	0.0485	0.0007571
13	0.0432	0.001173380	0.0478	0.001055971	0.0488	0.0007582
14	0.0461	0.001172532	0.0453	0.001049278	0.0475	0.0007556
15	0.0454	0.001176775	0.0485	0.001053089	0.0543	0.0007548
Mean	0.0454	0.0011743	0.0452	0.0010526	0.0495	0.0007556

Mean ($\hat{\theta}$)	0.934127	0.9343081	0.934101
p_{95}	18.30489	21.02422	31.36734
$\mathbf{p_{95}}$	17.97201	20.66881	31.40925

Σε αυτό το σημείο, βλέπουμε πως για $M = 20$ παίρνουμε μέγεθος $P_t = 0.0495$ πιο κοντά δηλαδή στο 0.05 αλλά η τιμή του M έχει ανέβει.

Για : censoring rate $= 0.1$, $n = 60$, $\kappa = 1.5$ και $\theta = 1$.

κ	**12**	
	P_t	mean(λ)
1	0.05020000	0.001839919
2	0.04860486	0.001821504
3	0.05050000	0.001835821
4	0.05220522	0.001819741
5	0.05580558	0.001826682
6	0.05521656	0.001816957
7	0.05032516	0.001846453
8	0.05241048	0.001814116
9	0.04601380	0.001818994
10	0.05140000	0.001836187
11	0.05040504	0.001816520
12	0.05001000	0.001838230
13	0.05181554	0.001825445
14	0.05391078	0.001839564
15	0.05141028	0.001829786
Mean	0.05134889	0.001828395

Mean ($\hat{\theta}$)	0.92524
p_{95}	21.02286
$\mathbf{p_{95}}$	21.11072

Για αυτή τη περίπτωση, το καλύτερο δυνατό αποτέλεσμα το παίρνουμε για $M = 12$.

Για : censoring rate $= 0.1$, $n = 120$, $\kappa = 1.5$ και $\theta = 1$.

κ	**10**		**12**	
	P_t	mean(λ)	P_t	mean(λ)
1	0.0497	0.0009285902	0.0488	0.0007649840
2	0.0500	0.0009295289	0.0534	0.0007692498
3	0.0538	0.0009292562	0.0466	0.0007692133
4	0.0514	0.0009314174	0.0499	0.0007709660
5	0.0501	0.0009256450	0.0522	0.0007698400
6	0.0510	0.0009291358	0.0447	0.0007670932
7	0.0487	0.0009249967	0.0493	0.0007732574
8	0.0502	0.0009328487	0.0471	0.0007689747
9	0.0471	0.0009280693	0.0471	0.0007686613
10	0.0491	0.0009250719	0.0515	0.0007675949
11	0.0495	0.0009268337	0.0479	0.0007676362
12	0.0489	0.0009241215	0.0521	0.0007647586
13	0.0475	0.0009351721	0.0491	0.0007645118
14	0.0483	0.0009328924	0.0474	0.0007675925
15	0.0517	0.0009332458	0.0509	0.0007716141
Mean	0.0498	0.0009291217	0.0492	0.0007683965

Mean ($\hat{\theta}$)	0.9092283	0.9095788
p_{95}	18.30534	21.02472
$\mathbf{p_{95}}$	18.29589	20.95346

Εδώ σημειώνουμε ότι για $M = 10$ ή 12, ο έλεγχος δίνει μέγεθος περίπου 0.05 όπως αναμέναμε.

Για : censoring rate $= 0.1$, $n = 240$, $\kappa = 1.5$ και $\theta = 1$.

ι	**18**	
	P_t	mean(λ)
1	0.0550	0.0002288581
2	0.0553	0.0002280680
3	0.0559	0.0002280595
4	0.0534	0.0002288596
5	0.0547	0.0002296238
6	0.0572	0.0002292595
7	0.0543	0.0002292529
8	0.0525	0.0002283868
9	0.0520	0.0002299031
10	0.0535	0.0002288034
11	0.0551	0.0002285830
12	0.0549	0.0002282449
13	0.0545	0.0002281531
14	0.0519	0.0002283381
15	0.0602	0.0002285845
Mean	0.05469333	0.000228

Mean ($\hat{\theta}$)	0.9026081
p_{95}	28.86895
$\mathbf{p_{95}}$	29.26625

Για αυτή τη περίπτωση, το καλύτερο δυνατό αποτέλεσμα το παίρνουμε για $M = 18$.

Για : censoring rate $= 0.3$, $n = 60$, $\kappa = 0.5$ και $\theta = 1$.

ι	**22**	
	P_t	mean(λ)
1	0.05651704	0.002450220
2	0.05183991	0.002486875
3	0.05944914	0.002489761
4	0.06189821	0.002431951
5	0.06189016	0.002461640
6	0.05882353	0.002490212
7	0.05825776	0.002448265
8	0.06668481	0.002478367
9	0.06203678	0.002459660
10	0.05339806	0.002456589
11	0.05438402	0.002482035
12	0.05834242	0.002509032
13	0.05960632	0.002434949
14	0.05534247	0.002462547
15	0.06244842	0.002469775
Mean	0.05872794	0.000246

Mean ($\hat{\theta}$)	0.7290374
p_{95}	33.92065
$\mathbf{p_{95}}$	34.6072

Σε αυτό το σημείο πρέπει να αναφέρουμε, πως με την αύξηση του ποσοστού λογοκρισίας σε 30%, για να προσαρμοστούν καλύτερα τα δεδομένα και ταυτοχρόνως το μέγεθος του ελέγχου να είναι όσο πιο κοντά στο 0.05, χρειάζεται να γίνει και αύξηση του πλήθους των διαστημάτων. Σε αυτή την περίπτωση έχουμε ότι ο έλεγχος δίνει καλύτερο μέγεθος για $M = 22$.

Για : censoring rate $= 0.3$, $n = 120$, $\kappa = 0.5$ και $\theta = 1$.

ι	**34**	
	P_t	mean(λ)
1	0.01754386	0.0006714274
2	0.03389831	0.0007677446
3	0.00000000	0.0007469049
4	0.03571429	0.0006747564
5	0.07407407	0.0006898196
6	0.03225806	0.0008062941
7	0.07407407	0.0007483271
8	0.06976744	0.0007199819
9	0.09836066	0.0007859981
10	0.03448276	0.0006929075
11	0.07843137	0.0007102623
12	0.03636364	0.0007493115
13	0.05454545	0.0007389458
14	0.08333333	0.0006955439
15	0.03571429	0.0006772079
Mean	0.05	0.0007250289

Mean ($\hat{\theta}$)	0.7215134
p_{95}	48.60131
$\mathbf{p_{95}}$	48.35589

Εδώ σημειώνουμε ότι με $M = 34$, ο έλεγχος δίνει το απαιτούμενο μέγεθος.

Για : censoring rate $= 0.3$, $n = 240$, $\kappa = 0.5$ και $\theta = 1$.

ι	**49**	
	P_t	mean(λ)
1	0.04395604	0.2065380
2	0.04819277	0.2053986
3	0.02352941	0.2045851
4	0.04651163	0.2055795
5	0.03488372	0.2030328
6	0.07954545	0.2060805
7	0.01190476	0.2069866
8	0.06741573	0.2048094
9	0.03370787	0.2052660
10	0.04761905	0.2068303
11	0.11235955	0.2059624
12	0.03448276	0.2064945
13	0.04705882	0.2056887
14	0.04651163	0.2064333
15	0.06741573	0.2067781
Mean	0.049673	0.2057

Mean ($\hat{\theta}$)	0.675865
p_{95}	73.90
$\mathbf{p_{95}}$	73.16

Εδώ σημειώνουμε ότι με $M = 49$, ο έλεγχος δίνει το απαιτούμενο μέγεθος.

Για : censoring rate $= 0.3$, $n = 60$, $\kappa = 1.5$ και $\theta = 1$.

ι	**28**	
	P_t	mean(λ)
1	0.06418219	0.0004736857
2	0.06907895	0.0004808629
3	0.05474860	0.0004750328
4	0.05162738	0.0004770231
5	0.05633803	0.0004811916
6	0.05214724	0.0004814387
7	0.05888651	0.0004867260
8	0.05971770	0.0004793684
9	0.05241521	0.0004757945
10	0.05615551	0.0004789276
11	0.06696935	0.0004804795
12	0.07004310	0.0004820620
13	0.06854839	0.0004778747
14	0.06347555	0.0004822638
15	0.05603448	0.0004756186
Mean	0.06	0.0004788

Mean ($\hat{\theta}$)	0.6746
p_{95}	41.33
$\mathbf{p_{95}}$	42.19

Εδώ σημειώνουμε ότι με $M = 28$, ο έλεγχος δίνει καλό μέγεθος, που παρόλα αυτά είναι μεγαλύτερο από 0.05. Έπειτα από δοκιμές παρατηρήθηκε πως για τόσο μικρό αριθμό παρατηρήσεων, με σχετικά μεγάλη διασπορά ευπάθειας, σε συνδυασμό με ποσοστό λογοκρισίας ίσο με 30%, έπρεπε να αυξηθεί πολύ ο αριθμός τον διαστημάτων που χωρίζουμε τις παρατηρήσεις, κάτι που δεν θα θέλαμε να συμβαίνει.

Για : censoring rate $= 0.3$, $n = 120$, $\kappa = 1.5$ και $\theta = 1$.

ι	**50**	
	P_t	mean(λ)
1	0.07761194	0.0001156003
2	0.10882353	0.0001143269
3	0.06629834	0.0001149638
4	0.08734940	0.0001142324
5	0.08333333	0.0001147165
6	0.09677419	0.0001154221
7	0.10991957	0.0001158176
8	0.09090909	0.0001147350
9	0.09230769	0.0001149230
10	0.11684783	0.0001152454
11	0.10192837	0.0001150631
12	0.08139535	0.0001141205
13	0.10561056	0.0001147556
14	0.10280374	0.0001141245
15	0.08450704	0.0001152791
Mean	0.09376133	0.0001

Mean ($\hat{\theta}$)	0.6538917
p_{95}	66.33
$\mathbf{p_{95}}$	70.85

Εδώ σημειώνουμε ότι με $M = 50$, ο έλεγχος δίνει σχετικά καλό μέγεθος, παρόλα αυτά είναι πολύ μεγαλύτερο από 0.05. Ένας πιθανός λόγος για αυτό είναι η μεγάλη τιμή για $\kappa = 1.5$ για τη διασπορά της ευπάθειας όταν έχουμε σε συνδυασμό μεγαλύτερο ποσοστό λογοκρισίας, όπως σημειώθηκε και πριν, ενώ για $\kappa = 0.5$ έχουμε καλύτερα αποτελέσματα. Επιπρόσθετα, η μεροληψία της εκτίμησης της άγνωστης παραμέτρου θ αυξάνει.

Για : censoring rate $= 0.5$, $n = 60$, $\kappa = 0.5$ και $\theta = 1$.

ι	**18**	
	P_t	mean(λ)
1	0.03925926	0.0009515226
2	0.03918613	0.0009583039
3	0.03678161	0.0009534370
4	0.04154519	0.0009548589
5	0.04824561	0.0009641203
6	0.05371597	0.0009565026
7	0.03876525	0.0009733890
8	0.04964539	0.0009592733
9	0.05247813	0.0009562135
10	0.04571843	0.0009749416
11	0.04851557	0.0009734858
12	0.04842260	0.0009558246
13	0.04867257	0.0009466892
14	0.04194757	0.0009648664
15	0.04888889	0.0009621812
Mean	0.04545254	0.00096

Mean ($\hat{\theta}$)	0.65
p_{95}	28.86
$\mathbf{p_{95}}$	28.49

Τώρα θα αυξηθεί και άλλο το ποσοστό λογοκρισίας σε 50%, άρα σε αυτή τη περίπτωση έχουμε ότι ο έλεγχος δίνει καλύτερο μέγεθος για $M = 18$.

Για : censoring rate $= 0.5$, $n = 120$, $\kappa = 0.5$ και $\theta = 1$.

ι	**30**	
	P_t	mean(λ)
1	0.06259314	0.0003837503
2	0.04827089	0.0003843782
3	0.05110603	0.0003824221
4	0.04672245	0.0003845813
5	0.05191060	0.0003804284
6	0.05379310	0.0003803889
7	0.04719101	0.0003846233
8	0.05828003	0.0003818351
9	0.04269175	0.0003850631
10	0.05232130	0.0003864411
11	0.05747126	0.0003882281
12	0.05468750	0.0003804562
13	0.04687500	0.0003813322
14	0.04175365	0.0003858778
15	0.06341463	0.0003819828
Mean	0.05	0.00038

Mean ($\hat{\theta}$)	0.6
p_{95}	43.77
$\mathbf{p_{95}}$	43.96

Εδώ σημειώνουμε ότι με $M = 30$, ο έλεγχος δίνει καλύτερο μέγεθος.

Για : censoring rate $= 0.5$, $n = 240$, $\kappa = 0.5$ και $\theta = 1$.

λ	**54**	
	P_t	mean(λ)
1	0.04237288	0.0001087958
2	0.04628331	0.0001080386
3	0.05546995	0.0001084574
4	0.05726872	0.0001079240
5	0.06433566	0.0001080252
6	0.05390071	0.0001080046
7	0.04746835	0.0001078595
8	0.04329609	0.0001081274
9	0.05701754	0.0001088608
10	0.05417277	0.0001082632
11	0.04826546	0.0001088421
12	0.04617605	0.0001081954
13	0.07417582	0.0001084259
14	0.04041916	0.0001088117
15	0.04587156	0.0001075233
Mean	0.05	0.0001

Mean ($\hat{\theta}$)	0.52
p_{95}	72.15
$\mathbf{p_{95}}$	72.27

Εδώ σημειώνουμε ότι με $M = 54$, ο έλεγχος δίνει το επιθυμητό μέγεθος.

Για τη Γάμμα κατανομή παρατηρούμε ότι για $\kappa = 0.5$ και στις τρεις περιπτώσεις με ποσοστό λογοκρισίας $10\%, 30\%, 50\%$, μπορεί να βρεθεί οτι ο κατάλληλος αριθμός διαστημάτων για να έχουμε P_t σχεδόν ίσο με 0.05 δεν είναι υπερβολικά μεγάλος. Όσο αυξάνουμε όμως τη διασπορά της μεταβλητής ευπάθειας, τόσο πιο πολύ αυξάνει το πλήθος των διαστημάτων M κάτι το οποίο δεν είναι επιθυμητό. Εξαίρεση αποτελεί η περίπτωση που το ποσοστό της λογοκρισίας είναι μικρό, δηλαδή 0.1. Οσο αυξάνεται η διασπορά της ευπάθειας σε

συνδυασμό με το ποσοστό της λογοκρισίας τόσο μεγαλώνει και η μεροληψία της εκτίμησης της παραμέτρου θ.

Κατανομή ευπάθειας: Inverse Gaussian, Σφάλμα τύπου I.

Για : censoring rate $= 0.1$, $n = 60$, $b = 1$ και $\theta = 1$.

κ	**22**	
	P_t	mean(λ)
1	0.03418896	0.3571885
2	0.03414371	0.3562843
3	0.04230686	0.3558620
4	0.04387194	0.3555340
5	0.04394657	0.3545149
6	0.03896384	0.3557218
7	0.03894432	0.3559208
8	0.03938541	0.3558376
9	0.03946373	0.3566092
10	0.03739523	0.3560809
11	0.03809936	0.3564046
12	0.03623814	0.3562873
13	0.03659197	0.3563154
14	0.04111807	0.3551704
15	0.04748482	0.3541214
Mean	0.04	0.35585

Mean ($\hat{\theta}$)	0.92
p_{95}	39.21299
$\mathbf{p_{95}}$	38.03621

Εδώ σημειώνουμε ότι με $M = 22$, ο έλεγχος δίνει καλό μέγεθος. Επίσης η διασπορά της ευπάθειας είναι ίση εδώ με $1/2$ δηλαδή σχετικά μικρή.

Για : censoring rate $= 0.1$, $n = 120$, $b = 1$ και $\theta = 1$.

ι	**22**	
	P_t	mean(λ)
1	0.03641092	0.3546307
2	0.03730373	0.3545645
3	0.03930393	0.3543792
4	0.03860772	0.3543467
5	0.04070000	0.3543753
6	0.04010000	0.3543243
7	0.03970794	0.3544810
8	0.03960792	0.3534326
9	0.04291287	0.3539501
10	0.04170834	0.3540647
11	0.03950790	0.3545961
12	0.03501401	0.3550088
13	0.04180418	0.3541805
14	0.03770754	0.3548977
15	0.04190838	0.3543370
Mean	0.043	0.3543713

Mean ($\hat{\theta}$)	0.9110263
p_{95}	38.61077
$\mathbf{p_{95}}$	37.42506

Εδώ σημειώνουμε ότι με $M = 22$, ο έλεγχος δίνει καλό μέγεθος.

Για : censoring rate $= 0.1$, $n = 240$, $b = 1$ και $\theta = 1$.

ι	**24**	
	P_t	mean(λ)
1	0.0503	0.3562252
2	0.0486	0.3564017
3	0.0493	0.3566393
4	0.0514	0.3556755
5	0.0469	0.3562738
6	0.0510	0.3564228
7	0.0442	0.3564803
8	0.0454	0.3567565
9	0.0535	0.3559679
10	0.0472	0.3567464
11	0.0492	0.3563856
12	0.0483	0.3565844
13	0.0523	0.3563115
14	0.0449	0.3569496
15	0.0501	0.3557325
Mean	0.04884	0.3563702

Mean ($\hat{\theta}$)	0.9059221
p_{95}	39.43463
$\mathbf{p_{95}}$	39.2928

Εδώ σημειώνουμε ότι με $M = 24$, ο έλεγχος δίνει καλύτερο μέγεθος, πιο κοντά στο επιθυμητό 0.05.

Για : censoring rate $= 0.1$, $n = 120$, $b = 0.5$ και $\theta = 1$.

λ	**32**	
	P_t	mean(λ)
1	0.04858300	0.2739022
2	0.05050505	0.2729669
3	0.06578947	0.2747470
4	0.04833837	0.2724677
5	0.04439960	0.2745377
6	0.04158215	0.2719569
7	0.06054490	0.2725316
8	0.05894309	0.2730658
9	0.05231388	0.2727018
10	0.05668016	0.2733857
11	0.06344411	0.2729498
12	0.05177665	0.2733936
13	0.04439960	0.2745271
14	0.05645161	0.2731010
15	0.04939516	0.2714239
Mean	0.0527	0.2731

Mean ($\hat{\theta}$)	0.9017
p_{95}	50.05517
$\mathbf{p_{95}}$	50.25838

Εδώ σημειώνουμε ότι με $M = 32$, ο έλεγχος δίνει το καλύτερο μέγεθος. Έχουμε καλή απόδοση ακόμα και με $n = 120$. Κάποια μικρή υποεκτίμηση της θ αναμένεται λόγω της λογοκρισίας. Επίσης η διασπορά της ευπάθειας είναι ίση εδώ με 1 δηλαδή μεγαλύτερη από αυτή που θεωρήσαμε στους προηγούμενους πίνακες γι΄ αυτό και ανεβαίνει η τιμή του M.

Για : censoring rate $= 0.1$, $n = 240$, $b = 0.5$ και $\theta = 1$.

ι	**32**	
	P_t	mean(λ)
1	0.05	0.2683578
2	0.06	0.2730067
3	0.05	0.2725437
4	0.05	0.2704933
5	0.08	0.2696848
6	0.04	0.2739718
7	0.07	0.2762614
8	0.04	0.2739845
9	0.02	0.2682839
10	0.08	0.2670887
11	0.09	0.2706815
12	0.06	0.2706927
13	0.05	0.2697592
14	0.02	0.2732671
15	0.04	0.2710817
Mean	0.05333333	0.2712773

Mean ($\hat{\theta}$)	0.8955983
p_{95}	49.6785
$\mathbf{p_{95}}$	49.77933

Εδώ σημειώνουμε ότι με $M = 32$, ο έλεγχος δίνει το καλύτερο μέγεθος.

Για : censoring rate = 0.3, $n = 120$, $b = 0.5$ και $\theta = 1$.

ι	**49**	
	P_t	mean(λ)
1	0.04878049	0.2088366
2	0.03571429	0.2084107
3	0.08641975	0.2042294
4	0.03448276	0.2051800
5	0.05882353	0.2066392
6	0.04597701	0.2070860
7	0.04878049	0.2054284
8	0.02222222	0.2072149
9	0.06593407	0.2074808
10	0.08045977	0.2081745
11	0.08139535	0.2056940
12	0.05376344	0.2069347
13	0.02325581	0.2045553
14	0.02150538	0.2049425
15	0.07865169	0.2061280
Mean	0.052	0.2064623

Mean ($\hat{\theta}$)	0.6795869
p_{95}	74.58467
$\mathbf{p_{95}}$	74.14805

Εδώ σημειώνουμε ότι με $M = 49$, ο έλεγχος δίνει το επιθυμητό μέγεθος. Παρά τη μεγαλύτερη λογοκρισία, ο έλεγχος δίνει καλά αποτελέσματα αλλά με M πολύ μεγάλο. Αναμένεται η μικρότερη τιμή για τη θ, καθώς τα μη παρατηρούμενα δεδομένα μειώνουν τις πληροφορίες.

Για : censoring rate = 0.1, $n = 60$, $b = 3$ και $\theta = 1$.

ι	**26**	
	P_t	mean(λ)
1	0.05183846	0.3165579
2	0.05469318	0.3154629
3	0.05324581	0.3153370
4	0.04787043	0.3180014
5	0.05525868	0.3157102
6	0.06563289	0.3090415
7	0.05530840	0.3152693
8	0.06003866	0.3110297
9	0.06820377	0.3111054
10	0.06361108	0.3123161
11	0.07296869	0.3096153
12	0.05068676	0.3166393
13	0.06377551	0.3114350
14	0.05301292	0.3140164
15	0.06358801	0.3124059
Mean	0.05864888	0.3135962

Mean ($\hat{\theta}$)	0.9295361
p_{95}	42.3789
$\mathbf{p_{95}}$	43.1708

Εδώ σημειώνουμε ότι με M =26, ο έλεγχος δίνει καλό μέγεθος κοντά στο 0.05, καθώς με την παράμετρο ευπάθειας ορισμένη ίση σε 3 πρέπει να γίνει μεγάλη αύξηση των διαστημάτων. Αυτό οφείλεται στο ότι η διασπορά της ευπάθειας έχει μεγαλώσει κι άλλο και έχει γίνει ίση με 1.5.

Για : censoring rate $= 0.1$, $n = 120$, $b = 3$ και $\theta = 1$.

i	**28**	
	P_t	mean(λ)
1	0.05009518	0.2978305
2	0.04817708	0.2990786
3	0.05138736	0.2982935
4	0.05573376	0.2966871
5	0.05226795	0.2980667
6	0.04888800	0.2977123
7	0.05689672	0.2950411
8	0.05211465	0.2976255
9	0.05340681	0.2970955
10	0.05586200	0.2976605
11	0.05129233	0.2973055
12	0.05073699	0.2977758
13	0.04759042	0.2986691
14	0.05359110	0.2983912
15	0.05528292	0.2957660
Mean	0.05222155	0.2975333

Mean ($\hat{\theta}$)	0.9207111
p_{95}	44.65693
$\mathbf{p_{95}}$	44.88769

Εδώ σημειώνουμε ότι με $M = 28$, ο έλεγχος δίνει το καλύτερο δυνατό μέγεθος. Συνεπώς γίνεται αντιληπτό πως με την αύξηση των παρατηρήσεων, έχουμε καλύτερα αποτελέσματα στην περίπτωση που $b = 3$.

Για : censoring rate $= 0.1$, $n = 240$, $b = 3$ και $\theta = 1$.

ι	**30**	
	P_t	mean(λ)
1	0.0260	0.2944564
2	0.0260	0.2949015
3	0.0263	0.2945035
4	0.0265	0.2949345
5	0.0288	0.2944608
6	0.0219	0.2952459
7	0.0310	0.2935432
8	0.0297	0.2937978
9	0.0280	0.2945622
10	0.0272	0.2948539
11	0.0280	0.2944958
12	0.0312	0.2937734
13	0.0227	0.2951686
14	0.0260	0.2945093
15	0.0303	0.2944017
Mean	0.02730667	0.2945072

Mean ($\hat{\theta}$)	0.9174002
p_{95}	51.2189
$\mathbf{p_{95}}$	47.71453

Το εμπειρικό ποσοστό απόρριψης (περίπου 2,73%) είναι κάτω από το επίπεδο 5%. Αυτό υποδεικνύει ότι ο έλεγχος είναι συντηρητικός ενδεχομένως λόγω της τιμής $b = 3$ και της αύξησης διασποράς. Επίσης έχουμε γενικά και ένα μικρό χάσμα ανάμεσα στην ασυμπτωτική κατανομή και την πραγματική κατανομή της ελεγχοσυνάρτησης, συμβάλλοντας στη συντηρητικότητα.

Για : censoring rate = 0.3, $n = 60$, $b = 1$ και $\theta = 1$.

ι	**28**	
	P_t	mean(λ)
1	0.07407407	0.2129671
2	0.04687500	0.2133484
3	0.06329114	0.2139651
4	0.07777778	0.2123473
5	0.09090909	0.2130628
6	0.06250000	0.2148123
7	0.09782609	0.2145845
8	0.06779661	0.2118244
9	0.05882353	0.2137803
10	0.09756098	0.2149332
11	0.03529412	0.2143298
12	0.10144928	0.2140215
13	0.07500000	0.2141643
14	0.02816901	0.2148136
15	0.06250000	0.2137826
Mean	0.066	0.213

Mean ($\hat{\theta}$)	0.6932
p_{95}	42.39
$\mathbf{p_{95}}$	43.76

Η αύξηση του ποσοστού λογοκρισίας σε 0.3 έχει ως αποτέλεσμα την υποεκτίμηση της ϑ σε 0.69 και παρόλη την αύξηση των διαμερίσεων σε 28 η καλύτερη δυνατή τιμή επιτυγχάνεται ίση με $P_t = 0.066$. Αυτό δείχνει ότι ο έλεγχος απορρίπτει τη μηδενική υπόθεση αρκετά συχνά. Αυτό το βλέπουμε και από το γεγονός ότι η κρίσιμη τιμή που χρησιμοποιείται στο τεστ έχει μια μικρή διαφορά σε σύγκριση με την πραγματική του στατιστικού ελέγχου.

Για : censoring rate $= 0.3$, $n = 120$, $b = 1$ και $\theta = 1$.

λ	**42**	
	P_t	mean(λ)
1	0.04748732	0.2291722
2	0.04558271	0.2295551
3	0.04277829	0.2295221
4	0.03875598	0.2296148
5	0.04525965	0.2292364
6	0.04759704	0.2292671
7	0.03924419	0.2292149
8	0.03252834	0.2293202
9	0.04519505	0.2294918
10	0.03851574	0.2293094
11	0.04753363	0.2293353
12	0.04147251	0.2293352
13	0.04263382	0.2292793
14	0.04750923	0.2293708
15	0.04449761	0.2297659
Mean	0.04310607	0.229386

Mean ($\hat{\theta}$)	0.7292481
p_{95}	64.26372
$\mathbf{p_{95}}$	63.41157

Σε αυτό το σημείο παρατηρούμε, πως παρόλο την αυξημένη τιμή της λογοκρισίας, για περισσότερες παρατηρήσεις και σχετικά μεγάλο αριθμό διαμερίσεων έχουμε, καλύτερα αποτελέσματα. Τώρα έχουμε ότι το μέγεθος του ελέγχου είναι πιο κοντά στο 5%, παρόλο που το θ εξακολουθεί να επηρεάζεται από την απώλεια πληροφοριών.

Για : censoring rate $= 0.3$, $n = 240$, $b = 1$ και $\theta = 1$.

ι	**44**	
	P_t	mean(λ)
1	0.06863485	0.2238375
2	0.06573984	0.2241175
3	0.06105308	0.2241279
4	0.06833494	0.2239283
5	0.06222508	0.2241122
6	0.06938468	0.2240776
7	0.06683222	0.2239710
8	0.06268112	0.2240798
9	0.06973727	0.2236746
10	0.06648678	0.2238184
11	0.06530172	0.2241252
12	0.06552279	0.2239794
13	0.06805212	0.2239988
14	0.06474125	0.2240660
15	0.06579798	0.2238350
Mean	0.06603505	0.2239833

Mean ($\hat{\theta}$)	0.7044786
p_{95}	68.27268
$\mathbf{p_{95}}$	70.28986

Για αυτή τη περίπτωση, το καλύτερο δυνατό αποτέλεσμα το παίρνουμε για $M = 44$.

Για : censoring rate $= 0.3$, $n = 60$, $b = 3$ και $\theta = 1$.

ι	**30**	
	P_t	mean(λ)
1	0.08713693	0.1574575
2	0.08514851	0.1572991
3	0.07444668	0.1567497
4	0.10123967	0.1580798
5	0.08661417	0.1573182
6	0.07234043	0.1576336
7	0.08453608	0.1571866
8	0.06263499	0.1581465
9	0.07014028	0.1577115
10	0.06843267	0.1578091
11	0.07283465	0.1574687
12	0.06916996	0.1576891
13	0.07520325	0.1574159
14	0.08651911	0.1569752
15	0.05263158	0.1578993
Mean	0.07660193	0.1575227

Mean ($\hat{\theta}$)	0.7447223
p_{95}	43.5574
$\mathbf{p_{95}}$	45.88848

Για αυτή τη περίπτωση, το καλύτερο δυνατό αποτέλεσμα το παίρνουμε για $M = 30$, το οποίο παραμένει υψηλό και αυτό θεωρείται πως συμβαίνει λόγω ίσως της τιμής $b = 3$, σε συνδυασμό με το ποσοστό λογοκρισίας 30%.

Για : censoring rate $= 0.3$, $n = 120$, $b = 3$ και $\theta = 1$.

ι	**52**	
	P_t	mean(λ)
1	0.05154639	0.1617738
2	0.07627119	0.1622931
3	0.05286344	0.1615449
4	0.09090909	0.1619568
5	0.10593220	0.1614248
6	0.06578947	0.1617094
7	0.07619048	0.1610486
8	0.06008584	0.1608365
9	0.04433498	0.1618006
10	0.07000000	0.1620217
11	0.04867257	0.1616349
12	0.07441860	0.1621205
13	0.07239819	0.1616754
14	0.06410256	0.1616895
15	0.08212560	0.1613592
Mean	0.06904271	0.1616593

Mean ($\hat{\theta}$)	0.7377939
p_{95}	70.51756
$\mathbf{p_{95}}$	72.79363

Για αυτή τη περίπτωση, το καλύτερο δυνατό αποτέλεσμα το παίρνουμε για $M = 52$.

Για : censoring rate $= 0.3$, $n = 240$, $b = 3$ και $\theta = 1$.

λ	**64**	
	P_t	mean(λ)
1	0.05672124	0.1632478
2	0.04603756	0.1633365
3	0.05423494	0.1632431
4	0.04795615	0.1637560
5	0.05187386	0.1635429
6	0.05117026	0.1632081
7	0.05354691	0.1629864
8	0.05765271	0.1631460
9	0.05199629	0.1631730
10	0.05336952	0.1634357
11	0.05767926	0.1629522
12	0.06382979	0.1630733
13	0.05414827	0.1629747
14	0.05669509	0.1633575
15	0.05050274	0.1634670
Mean	0.05382764	0.16326

Mean ($\hat{\theta}$)	0.7343526
p_{95}	89.6127
$\mathbf{p_{95}}$	90.12334

Για αυτή τη περίπτωση, το καλύτερο δυνατό αποτέλεσμα το παίρνουμε για $M = 64$. Βλέπουμε πως έχουμε καλύτερα αποτελέσματα από ότι στις περιπτώσεις που οι παρατηρήσεις είναι $n = 60$ και $n = 120$, καθώς το P_t είναι περίπου 5%.

Για παράδειγμα αν οι παρατηρήσεις αυξηθούν για $n = 300$, για $M = 64$, έχουμε,

λ	**64**	
	P_t	mean(λ)
1	0.04915730	0.1626903
2	0.05179283	0.1632495
3	0.05220884	0.1633472
4	0.06344828	0.1633129
5	0.05722071	0.1636119
6	0.06308725	0.1623618
7	0.05874499	0.1635383
8	0.06693440	0.1624828
9	0.05945946	0.1632796
10	0.05061560	0.1632345
11	0.06873315	0.1626954
12	0.04768392	0.1629959
13	0.05198358	0.1631153
14	0.05802969	0.1641663
15	0.05019815	0.1635401
Mean	0.056	0.1631

Mean ($\hat{\theta}$)	0.7338286
p_{95}	91.99
$\mathbf{p_{95}}$	92.9

Για $M = 64$ βρέθηκε η βέλτιστη τιμή του P_t.

Για : censoring rate $= 0.5$, $n = 60$, $b = 1$ και $\theta = 1$.

κ	**28**	
	P_t	mean(λ)
1	0.0000000	0.1896986
2	0.0000000	0.1898308
3	0.1428571	0.1904947
4	0.1666667	0.1894688
5	0.0000000	0.1893031
6	0.0000000	0.1899909
7	0.0000000	0.1905548
8	0.2000000	0.1902576
9	0.0000000	0.1908271
10	0.0000000	0.1896762
11	0.0000000	0.1893533
12	0.0000000	0.1901925
13	0.0000000	0.1886036
14	0.2500000	0.1894883
15	0.0000000	0.1891606
Mean	0.05	0.189

Mean ($\hat{\theta}$)	0.547
p_{95}	37.38
$\mathbf{p_{95}}$	33.45

Για αυτή τη περίπτωση, το καλύτερο δυνατό αποτέλεσμα το παίρνουμε για $M = 28$. Το εκτιμώμενο θ είναι 0.547, το οποίο απέχει πολύ από την πραγματική τιμή $\theta = 1$. Επίσης παρατηρούμε διαφορά στη κρίσιμη τιμή με την εμπειρική το οποίο σημαίνει ότι το τεστ είναι συντηρητικό (δηλαδή, απορρίπτει λιγότερο συχνά), εξηγώντας γιατί υπάρχουν πολλά $P_t = 0$. Αυτό πολύ πιθανό να συμβαίνει λόγο λίγο παρατηρήσεων σε συνδυασμό με μεγάλο ποσοστό λογοκρισίας.

Για : censoring rate $= 0.5$, $n = 120$, $b = 1$ και $\theta = 1$.

ι	**38**	
	P_t	mean(λ)
1	0.03030303	0.1903678
2	0.07228916	0.1903316
3	0.05747126	0.1901562
4	0.08988764	0.1900967
5	0.02941176	0.1902150
6	0.04494382	0.1901807
7	0.06250000	0.1902363
8	0.05128205	0.1904475
9	0.04901961	0.1901235
10	0.03488372	0.1901475
11	0.03000000	0.1901507
12	0.03960396	0.1902761
13	0.03571429	0.1901484
14	0.01408451	0.1903750
15	0.02500000	0.1901515
Mean	0.04442632	0.190227

Mean ($\hat{\theta}$)	0.5313823
p_{95}	53.48106
$\mathbf{p_{95}}$	52.40626

Για αυτή τη περίπτωση, το καλύτερο δυνατό αποτέλεσμα το παίρνουμε για $M = 38$. Με την αύξηση των δεδομένων αρχίζουμε και έχουμε λίγο καλύτερα αποτελέσματα, όχι όμως και τα ιδανικότερα, καθώς το θ υποεκτιμάται ακόμα.

Για : censoring rate $= 0.5$, $n = 60$, $b = 3$ και $\theta = 1$.

ι	**22**	
	P_t	mean(λ)
1	0.04477612	0.1484353
2	0.03960396	0.1487891
3	0.03333333	0.1492149
4	0.05069124	0.1490630
5	0.03743316	0.1491440
6	0.03669725	0.1493722
7	0.06250000	0.1488865
8	0.04000000	0.1481737
9	0.05741627	0.1486850
10	0.04265403	0.1486504
11	0.02631579	0.1491597
12	0.02116402	0.1491051
13	0.04926108	0.1489640
14	0.02525253	0.1487074
15	0.07352941	0.1485443
Mean	0.04270855	0.1488597

Mean ($\hat{\theta}$)	0.5681059
p_{95}	33.70189
$\mathbf{p_{95}}$	32.81597

Για αυτή τη περίπτωση, το καλύτερο δυνατό αποτέλεσμα το παίρνουμε για $M = 22$. Σε αυτή τη περίπτωση έχουμε καλή συμπεριφορά του ελέγχου ακόμη και για μικρές τιμές για το n, αλλά έχουμε μεροληψία στην εκτίμηση.

Για : censoring rate $= 0.5$, $n = 120$, $b = 3$ και $\theta = 1$.

ι	**34**	
	P_t	mean(λ)
1	0.03714286	0.1512525
2	0.03588517	0.1518181
3	0.06194690	0.1515627
4	0.05185185	0.1517553
5	0.04644809	0.1517713
6	0.06015038	0.1515125
7	0.06417112	0.1514504
8	0.06406685	0.1516565
9	0.06233062	0.1514921
10	0.03519062	0.1518201
11	0.05555556	0.1516995
12	0.05507246	0.1514372
13	0.04878049	0.1512788
14	0.06353591	0.1517264
15	0.05649718	0.1519406
Mean	0.05324174	0.1516116

Mean ($\hat{\theta}$)	0.5541337
p_{95}	48.40663
$\mathbf{p_{95}}$	48.7038

Για αυτή τη περίπτωση, το καλύτερο δυνατό αποτέλεσμα το παίρνουμε για $M = 34$.

Για : censoring rate $= 0.5$, $n = 240$, $b = 3$ και $\theta = 1$.

ι	**58**	
	P_t	mean(λ)
1	0.05660377	0.1535283
2	0.05555556	0.1533648
3	0.05844156	0.1531238
4	0.05280528	0.1535119
5	0.07167235	0.1534314
6	0.05382436	0.1534994
7	0.03797468	0.1532828
8	0.04154303	0.1532673
9	0.04062500	0.1533697
10	0.04587156	0.1531303
11	0.06622517	0.1532844
12	0.05245902	0.1533414
13	0.08000000	0.1533911
14	0.05775076	0.1533042
15	0.05047319	0.1533231
Mean	0.05478835	0.1533436

Mean ($\hat{\theta}$)	0.5455748
p_{95}	77.88988
$\mathbf{p_{95}}$	78.61703

Για αυτή τη περίπτωση, το καλύτερο δυνατό αποτέλεσμα το παίρνουμε για $M = 58$.

Στην περίπτωση της Inverse Gaussian κατανομής, τα καλύτερα αποτελέσματα τα λαμβάνουμε στην περίπτωση όπου το ποσοστό λογοκρισίας είναι 0.1. Με την αύξηση του ποσοστού λογοκρισίας μειώνεται η εκτίμηση του θ, παρόλο που σε κάποιες περιπτώσεις το P_t είναι σχεδόν ίσο με 0.05.

4.8.1 Σχόλια - Συμπεράσματα

Η παρούσα μεταπτυχιακή εργασία ασχολήθηκε με την ανάπτυξη, ανάλυση και αξιολόγηση στατιστικών μεθόδων για δεδομένα επιβίωσης, με επίκεντρο τα μοντέλα ευπάθειας και την εφαρμογή ϕ-μέτρων απόκλισης σε ελέγχους καλής προσαρμογής, με παρουσία δεξιάς λογοκρισίας. Βασικά χαρακτηριστικά της αποτέλεσαν η ανάπτυξη ενός νέου στατιστικού ελέγχου καλής προσαρμογής για παραμετρικά μοντέλα ευπάθειας, ο οποίος βασίζεται σε ϕ-μέτρα απόκλισης μεταξύ της θεωρητικής και της παρατηρούμενης-εμπειρικής κατανομής. Η μελέτη της ελεγχοσυνάρτησης έγινε κάτω από τη μηδενική υπόθεση και θεωρήσαμε ότι τα δεδομένα ακολουθούν μονομεταβλητά μοντέλα ευπάθειας με γνωστή διασπορά. Καθώς επίσης υποθέσαμε ως κατανομές ευπάθειας τη Γάμμα και την Inverse Gaussian. Έγινε ανάλυση της ελεγχοσυνάρτησης T_n^ϕ, η οποία βασίζεται σε μέτρα απόκλισης και συγκεκριμένα στην κλάση των ϕ-μέτρων απόκλισης (ϕ-divergence measures) από τα οποία επιλέξαμε να χρησιμοποιήσουμε το μέτρο Kullback-Leibler. Η ελεγχοσυνάρτηση T_n^ϕ απαιτεί τη διαμέριση των δεδομένων σε M γενικά μη επικαλυπτόμενα διαστήματα. Βρήκαμε ότι για για κάθε μέγεθος δείγματος n υπάρχει ένα M για το οποίο η εμπειρική κατανομή της ελεγχοσυνάρτησης T_n^ϕ είναι πολύ κοντά με την ασυμπτωτική κατανομή της. Από την ασυμπτωτική κατανομή πήραμε την κρίσιμη σταθερά για το χωρίο απορρίψεως του ελέγχου.

Η ενότητα των προσομοιώσεων είχε στόχο την εμπειρική αξιολόγηση της απόδοσης του προτεινόμενου ελέγχου καλής προσαρμογής για μοντέλα ευπάθειας, σε περιπτώσεις παρουσίας λογοκριμένων και μη δεδομένων. Οι πίνακες που παρουσιάστηκαν στην υποενότητα 4.8 συγκεντρώνουν στατιστικά αποτελέσματα όπως το μέσο σφάλμα τύπου I (P_t), η μέση τιμή του εκτιμητή της παραμέτρου θ που υπεισέρχεται στη συνάρτηση κινδύνου του μοντέλου για την οποία υποθέσαμε εκθετική κατανομή, καθώς και οι τιμές του κρίσιμου ποσοστημορίου $p95$ από την ασυμπτωτική και εμπειρική κατανομή, βάσει δεδομένων που προσομοιώθηκαν υπό διαφορετικά μοντέλα ευπάθειας (Γάμμα, Inverse Gaussian), διαφορετικές τιμές διασποράς της ευπάθειας και ποικίλα μεγέθη δείγματος n.

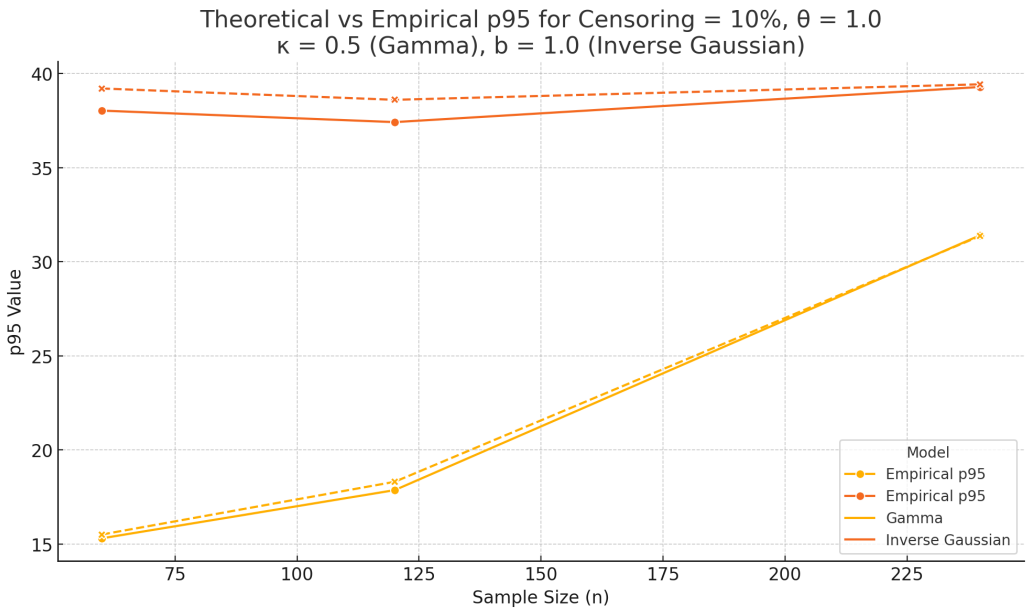
Από την ανάλυση των πινάκων εξάγονται συμπεράσματα όπως, ότι οι τιμές του σφάλματος τύπου I παραμένουν κοντά στο επιθυμητό επίπεδο προσαρμογής 5% για κατάλληλες τιμές της παραμέτρου διαμέρισης M, επιβεβαιώνοντας την ορθή βαθμονόμηση του ελέγχου και την εγκυρότητα της ασυμπτωτικής προσέγγισης. Αυτό όμως συμβαίνει κυρίως στις περιπτώσεις που το ποσοστό λογοκρισίας είναι χαμηλό, γύρω στο $10\% - 30\%$ και πολύ συχνά σε συνδυασμό με μικρή τιμή για την διασπορά της μεταβλητής ευπάθειας. Όσο αυξάνεται το

ποσοστό λογοκρισίας, τόσο λιγότερα «πραγματικά συμβάντα» έχουμε, γεγονός που αποδυναμώνει τον έλεγχο και την ακρίβεια των εκτιμήσεων. Επίσης σε υψηλά ποσοστά λογοκρισίας, το εμπειρικό $\mathbf{p_{95}}$ αποκλίνει συχνότερα από το θεωρητικό p_{95}, και το σφάλμα τύπου I μπορεί να απορρυθμιστεί ελαφρώς. Το οποίο σφάλμα παραμένει συνήθως σταθερό όταν το μοντέλο είναι σωστά προσαρμοσμένο, αλλά τείνει να αποκλίνει όταν η λογοκρισία συνδυάζεται με μικρό αριθμό παρατηρήσεων n.

Επίσης είδαμε πως ο εκτιμητής της παραμέτρου ευπάθειας $\hat{\theta}$ παρουσιάζει μικρή μεροληψία και καλή σύγκλιση προς την πραγματική τιμή θ, ιδίως για μεγαλύτερα δείγματα, γεγονός που ενισχύει τη συνέπεια της μεθόδου εκτίμησης. Η παράμετρος ευπάθειας (διασπορά) ελέγχει το επίπεδο μη παρατηρήσιμης ετερογένειας μεταξύ των παρατηρήσεων έτσι όσο μεγαλύτερη η τιμή της, τόσο εντονότερη η τυχαία επίδραση ευπάθειας και τόσο πιο εξαρτημένα είναι τα άτομα εντός των ομάδων. Αυξημένη ευπάθεια οδηγεί σε μεγαλύτερη διασπορά στους χρόνους επιβίωσης και αύξηση της μεταβλητότητας των εκτιμήσεων. Έτσι σε προσομοιώσεις όπου αυξάνεται η διασπορά της ευπάθειας η εμπειρική $\mathbf{p_{95}}$ τείνει να αποκλίνει περισσότερο από τη θεωρητική, και οι εκτιμήσεις γίνονται λιγότερο σταθερές.

Η εμπειρική $\mathbf{p_{95}}$ τείνει να είναι ελαφρώς υψηλότερη από τη θεωρητική, ιδίως σε μικρά δείγματα, κάτι που δηλώνει ότι ο έλεγχος τείνει να είναι συντηρητικός, απορρίπτοντας λιγότερο συχνά από το αναμενόμενο. Καθώς όμως αυξάνεται το μέγεθος δείγματος, η εμπειρική τιμή συγκλίνει προς τη θεωρητική, επιβεβαι-

ώνοντας την ασυμπτωτική σύγκλιση του ελέγχου T_n^{ϕ}.

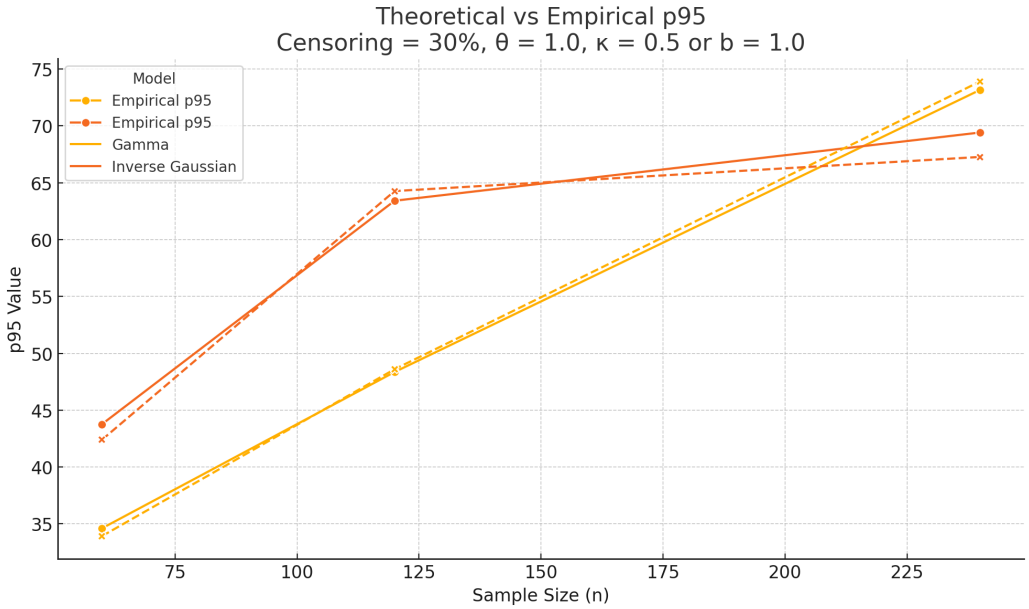


Σχήμα 13: Γράφημα σύγκρισης θεωρητικών και εμπειρικών τιμών του ποσοστημορίου p_{95} σε διάφορα μεγέθη δείγματος, για 10% λογοκρισία.

Το παραπάνω γράφημα δημιουργήθηκε για τη σύγκριση των θεωρητικών και εμπειρικών τιμών του ποσοστημορίου p_{95} σε διάφορα μεγέθη δείγματος. Συγκεκριμένα από τις περιπτώσεις προσομοιώσεων όπου το ποσοστό λογοκρισίας είναι 10% και η διασπορά ευπάθειας είναι ίση με 1 και για τις δυο περιπτώσεις, δηλαδή $\kappa = 0.5$ και $b = 1$. Παρατηρούμε πως το Γάμμα μοντέλο έχει χαμηλότερες τιμές p_{95} και παρουσιάζει καλή σύγκλιση μεταξύ των θεωρητικών και εμπειρικών τιμών. Στο Inverse Gaussian, οι τιμές p_{95} είναι υψηλότερες, αλλά και πάλι η εμπειρική τιμή πλησιάζει πολύ τη θεωρητική, ιδίως στα μεγάλα δείγματα. Βλέπουμε πως και στα δύο μοντέλα, η σύγκλιση της εμπειρικής τιμής προς τη θεωρητική βελτιώνεται με την αύξηση του δείγματος n. Επίσης το Γάμ-

μα μοντέλο είναι ίσως πιο σταθερό για χρήση σε δοκιμές καλής προσαρμογής λόγω χαμηλότερης διακύμανσης στις τιμές του ελέγχου, καθώς παρουσιάζει λιγότερες αποκλίσεις και ομαλότερη συμπεριφορά.

Στη συνέχεια στο παρακάτω διάγραμμα απεικονίζεται η σύγκριση μεταξύ της θεωρητικής και της εμπειρικής τιμής του ποσοστού p_{95}, για ποσοστό λογοκρισίας 30%, για τις δύο κατανομές ευπάθειας Γάμμα (με $\kappa = 0.5$) και Inverse Gaussian (με $b = 1$) και παράμετρο $\theta = 1$.

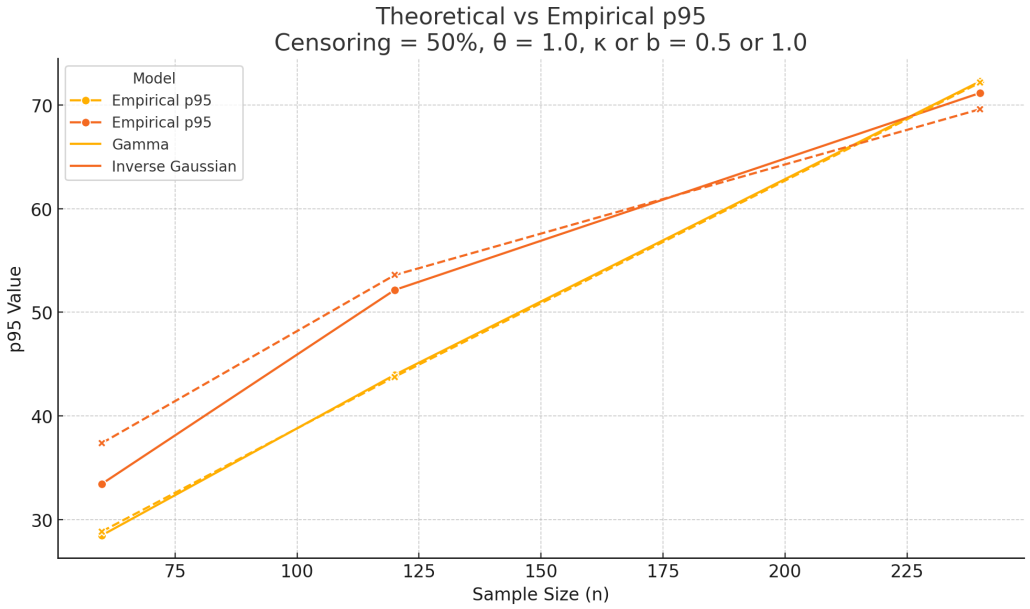


Σχήμα 14: Γράφημα σύγκρισης θεωρητικών και εμπειρικών τιμών του ποσοστημορίου p_{95} σε διάφορα μεγέθη δείγματος, για 30% λογοκρισία.

Στο Γάμμα μοντέλο παρατηρείται σταδιακή και γραμμική αύξηση της τιμής p_{95} με το μέγεθος δείγματος και η εμπειρική τιμή συγκλίνει εξαιρετικά καλά προς τη θεωρητική, ακόμα και σε μικρό δείγμα (π.χ. $n = 60$). Για $n = 240$, η τιμή $\mathbf{p_{95}}$ ξεπερνά την τιμή 73, κάτι που επιβεβαιώνει την επίδραση της

λογοκρισίας στη διασπορά του ελέγχου. Στο Inverse Gaussian μοντέλο οι τιμές $\mathbf{p_{95}}$ είναι γενικά υψηλότερες, με πιο απότομη αύξηση από $n = 60$ σε $n = 120$, αλλά σχετική σταθεροποίηση στη συνέχεια. Η εμπειρική $\mathbf{p_{95}}$ παραμένει κοντά στη θεωρητική, αν και με ελαφρώς μεγαλύτερες αποκλίσεις από το Γάμμα. Η συμπεριφορά υποδηλώνει ότι η κατανομή αυτή ίσως είναι πιο ευαίσθητη στο μέγεθος δείγματος.

Τέλος παρουσιάζεται το διάγραμμα σύγκρισης της θεωρητικής και της εμπειρικής τιμής του ποσοστού p_{95}, για ποσοστό λογοκρισίας 50%, για τις δύο κατανομές ευπάθειας Γάμμα (με $\kappa = 0.5$) και Inverse Gaussian (με $b = 1$) και παράμετρο $\theta = 1$.



Σχήμα 15: Γράφημα σύγκρισης θεωρητικών και εμπειρικών τιμών του ποσοστημορίου p_{95} σε διάφορα μεγέθη δείγματος, για 50% λογοκρισία.

Βλέπουμε όπως και πριν ότι το Γάμμα μοντέλο παρουσιάζει γραμμική και

σταθερή αύξηση της $\mathbf{p_{95}}$ με το μέγεθος δείγματος. Η διαφορά μεταξύ θεωρητικής και εμπειρικής $\mathbf{p_{95}}$ είναι ελάχιστη, επιβεβαιώνοντας την αξιοπιστία του μοντέλου ακόμη και υπό υψηλή λογοκρισία. Στο Inverse Gaussian μοντέλο οι τιμές $\mathbf{p_{95}}$ είναι συστηματικά υψηλότερες από το Γάμμα μοντέλο και σε μικρό δείγμα ($n = 60$), η εμπειρική $\mathbf{p_{95}}$ είναι αισθητά μικρότερη από τη θεωρητική και λαμβάνουμε καλύτερα αποτελέσματα για $n = 120$ και άνω, όπου οι τιμές συγκλίνουν ικανοποιητικά.

Συνεπώς συμπεραίνουμε ότι η αύξηση του ποσοστού λογοκρισίας οδηγεί σε αισθητή αύξηση της τιμής $\mathbf{p_{95}}$ τόσο στο Γάμμα όσο και στο Inverse Gaussian μοντέλο. Το Γάμμα μοντέλο παρουσιάζει πιο ομαλή και γραμμική αύξηση της $\mathbf{p_{95}}$ με το μέγεθος δείγματος και δείχνει πολύ καλή σύγκλιση μεταξύ θεωρίας και πράξης. Ενώ το Inverse Gaussian τείνει να έχει υψηλότερες τιμές $\mathbf{p_{95}}$, ιδιαίτερα σε μικρά δείγματα ή υπό ισχυρή λογοκρισία. Παρ' όλα αυτά, η συμφωνία θεωρητικής και εμπειρικής τιμής βελτιώνεται σημαντικά για n μεγαλύτερο του 120. Η αύξηση του μεγέθους δείγματος οδηγεί σε καλύτερη προσέγγιση της εμπειρικής $\mathbf{p_{95}}$ προς τη θεωρητική, και στα δύο μοντέλα. Για $n = 240$, σχεδόν σε όλες τις περιπτώσεις, η εμπειρική $\mathbf{p_{95}}$ είναι πολύ κοντά στην αντίστοιχη θεωρητική, επιβεβαιώνοντας την εγκυρότητα της ασυμπτωτικής προσέγγισης.

Οι μικρές αποκλίσεις στην εκτίμηση της παραμέτρου θ είναι αναμενόμενες λόγω διαφορών στη διακύμανση της ευπάθειας και του βαθμού λογοκρισίας. Πρέπει επίσης να αναφέρουμε πως και η επιλογή της παραμέτρου M αποδείχθηκε κρίσιμη για την επίτευξη του επιθυμητού μεγέθους του ελέγχου 0.05. Σε κάθε σενάριο εντοπίστηκε η καλύτερη δυνατή τιμή διαμέρισης που προσφέρει το επιθυμητό αποτέλεσμα μεγέθους. Αξίζει να σημειώσουμε εδώ ότι όσο το ποσοστό λογοκρισίας αυξάνει παρατηρούμε ότι ο αριθμός M των διαστημάτων που απαιτούνται για να επιτευχθεί το επιθυμητό μέγεθος του ελέγχου αυξάνει σε βαθμό μεγαλύτερο από ότι θα θέλαμε. Αυτό μας οδηγεί στο συμπέρασμα ότι η ασυμπτωτική κατανομή που χρησιμοποιήθηκε για την ελεγχοσυνάρτηση θέλει περισσότερη διερεύνηση και βελτίωση.

Με βάση τα θεωρητικά αποτελέσματα και τις εκτεταμένες προσομοιώσεις που πραγματοποιήθηκαν, ο προτεινόμενος έλεγχος καλής προσαρμογής αποδείχθηκε αποτελεσματικός, σταθερός και κατάλληλος για την αξιολόγηση μοντέλων ευπάθειας σε δεδομένα με δεξιά λογοκρισία. Οι υποθέσεις της μεθόδου επαληθεύτηκαν στην πράξη, και ο συνδυασμός ϕ-μέτρων απόκλισης με την ελεγχοσυνάρτηση οδήγησε σε καλούς ελέγχους, με σφάλμα τύπου I κοντά στο θεωρητικό επίπεδο. Επομένως, μπορεί να ειπωθεί ότι η μεθοδολογία που αναπτύχθηκε είναι κατάλληλη για πρακτική εφαρμογή, ιδίως όταν απαιτείται έλεγχος καλής προσαρμογής σε στοχαστικά μοντέλα με μη παρατηρούμενη

ετερογένεια. Επιπλέον, η σταθερότητα των εκτιμητών και η συνέπεια της προσέγγισης ακόμα και σε σχετικά μικρά μεγέθη δείγματος καθιστούν το μοντέλο εύρωστο και αξιόπιστο εργαλείο στην ανάλυση επιβίωσης.

5 Βιβλιογραφία (Αγγλικά)

Αναφορές

[1] Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. Annals of Statistics 6, 701–726.

[2] Akaike, H. (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723.

[3] Balakrishnan, N., Peng, Y. (2006) Generalized gamma frailty model. Statistics in Medicine 25, 2797–2816.

[4] Barker, P., Henderson, R. (2004) Modelling converging hazards in survival analysis. Lifetime Data Analysis 10, 263–281.

[5] Barker, P., Henderson, R. (2005) Small sample bias in the gamma frailty model for univariate survival. Lifetime Data Analysis 11, 265–284.

[6] Caroni, C., Crowder, M., Kimber, A. (2010) Proportional hazards models with discrete frailty. Lifetime Data Analysis 16, 374–384.

[7] Chen, H.S. Lai, K. and Ying, Z. (2004). Goodness of fit tests and minimum power divergence estimators for survival data, Statist. Sinica 14, 231–248.

[8] Chernoff, H., & Lehmann, E. L. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. The Annals of Mathematical Statistics.

[9] Collett, D. (2003) Modelling Survival Data in Medical Research. Chapman & Hall/CRC, London.

[10] Commenges, D., Andersen P.K. (1995) Score test of homogeneity for survival data. Lifetime Data Analysis 1, 145–160.

[11] Cox, D.R. (1959) Analysis of exponentially distributed life-times with two types of failure. Journal of the Royal Statistical Society (B) 21, 411–421.

[12] Cox, D.R. (1972) Regression models and life-tables. Journal of the Royal Statistical Society (B) 34, 187–220.

[13] Cox, D.R. (1975) Partial likelihood. Biometrika 62, 269–276. Cox, D.R., Oakes, D.(1984) Analysis of Survival Data. Chapman & Hall, London.

[14] Cressie, N., & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. Journal of the Royal Statistical Society: Series B (Methodological).

[15] Duchateau, L., Janssen, P. (2008) The Frailty Model. Springer, New York. Dunson, D.B., Chen, Z. (2004) Selecting factors predictive of heterogeneity in multivariate event time data. Biometrics 60, 352–358.

[16] Economou, P., Caroni, C. (2005) Graphical tests for the assumption of gamma and inverse Gaussian frailty distributions. Lifetime Data Analysis 11, 565–582.

[17] Economou, P., Caroni, C. (2008) Graphical tests for the frailty distribution in the shared frailty model. Communications in Statistics - Simulations and Computation 37, 978–992.

[18] Efron, B. (1977) The efficiency of Cox's likelihood function for censored data. Journal of the American Statistical Association 72, 557–565.

[19] Farewell, V.T. (1977) A model for a binary variable with time-censored observations. Biometrika 64, 43–46.

[20] Feller, W. (1971) An Introduction to Probability Theory and its Applications. John Wiley and Sons, New York.

[21] Fisher, R. A. (1925). Theory of statistical estimation. Mathematical Proceedings of the Cambridge Philosophical Society.

[22] Fleming, T., Harrington, D. (1991) Counting Processes and Survival Analysis. Wiley & Sons, Chichester.

[23] Geerdens, C., Claeskens, G., & Janssen, P. (2012). Goodness-of-fit tests for the frailty distribution in proportional hazards models with shared frailty.

[24] Gjessing, H.K., Aalen, O.O., Hjort, N.L. (2003) Frailty models based on Levy processes. Advances in Applied Probability 35, 532–550.

[25] Hougaard, P. (1986b) A class of multivariate failure time distributions. Biometrika 73, 671–678.

[26] Hougaard, P. (1995) Frailty models for survival data. Lifetime Data Analysis 1, 255–273.

[27] Hougaard, P. (1999) Fundamentals of survival data. Biometrics 55, 13–22.

[28] Huang, X., Wolfe, R.A. (2002) A frailty model for informative censoring. Biometrics 58, 510–520.

[29] Huber-Carol, C., Vonta, I. (2004) Frailty models for arbitrarily censored and truncated data. Lifetime Data Analysis 10, 369–388.

[30] Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London. Series A.

[31] Kalbfleisch, J., Prentice, R. (2002) The Statistical Analysis of Failure Time Data. Wiley, New York.

[32] Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457–48.

[33] Klein, J.P., Moeschberger, M.L. (2003) Survival Analysis - Techniques for Censored and Truncated Data. Springer, New York.

[34] Kosorok, M.R., Lee, B.L., Fine, J.P. (2004) Semiparametric inference for proportional hazards frailty regression models. Annals of Statistics 32, 1448–1491.

[35] Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. The Annals of Mathematical Statistics.

[36] Mahalanobis, P. C. (1936). On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India.

[37] Miller, R.G. (1981) Survival Analysis. Wiley & Sons, New York.

[38] Morales, D., Pardo, L., & Vajda, I. (1995). Asymptotic divergences of estimates of discrete distributions. Journal of Statistical Planning and Inference.

[39] Morley, E., Perry, H. M., Miller, D. K. (2002) Something about frailty. Journal of Gerontology: Medical Sciences 57A, M698–M704.

[40] Pardo, L. (2006). Statistical Inference Based on Divergence Measures. Chapman & Hall/CRC.

[41] Pareto, V. (1897) Cours d'Economie Politique. Rouge, Paris. Parmar, M., Machin, D. (1995) Survival Analysis: A Practical Approach. Wiley & Sons, New York.

[42] Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics.

[43] Smith, P.J. (2002) Analysis of Failure and Survival Data. Chapman & Hall/CRC, London.

[44] Sun, J. (2006) The Statistical Analysis of Interval-censored Failure Time Data. Springer, New York.

[45] Tableman, M., Kim, J.S. (2004) Survival Analysis using S: Analysis of Timeto-event Data. Chapman & Hall/CRC, London.

[46] Therneau, T.M., Grambsch, P.M. (2000) Modeling Survival Data. Springer, New York.

[47] Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16, 439–54.

[48] Vaupel, J.W., Carey, J.R., Christensen, K., Johnson, T.E., Yashin, A.I., Holm, N.V., Iachine, I.A., Kannisto, V., Khazaeli, A.A., Liedo, P., Longo, V.D., Zeng, Y., Manton, K.G., Curtsinger, J.W. (1998). Biodemographic trajectories of longevity. Science 280, 855–860.

[49] Vaupel, J.W., Yashin, A.I. (1985) Heterogeneity's ruses: some surprising effects of selection on population dynamics. The American Statistician 39, 176–185.

[50] Vonta, F., & Karagrigoriou, A. (2014). Goodness-of-fit tests via ϕ-measures of divergence for censored data. . *Journal of Statistical Computation and Simulation*, Vol. 84, No. 5, 946–963.

[51] Wienke, A. (2010). *Frailty Models in Survival Analysis*. Chapman & Hall/CRC Biostatistics Series.

[52] Zahl, P. (1997) Frailty modelling for the excess hazard. Statistics in Medicine 16, 1573–1585.

6 Βιβλιογραφία (Ελληνικά)

Αναφορές

[1] Σοφοκλέους και Βόντα (2023). Έλεγχοι Καλής Προσαρμογής για Μοντέλα Ευπάθειας μέσω Μέτρων Απόκλισης, Διατριβή Master, ΕΜΠ.

[2] Χαβιατζή Άννα (2015). Μέθοδοι Ποινικοποιημένης Πιθανοφάνειας στα Μοντέλα Ευπάθειας με Ομαδοποιημένα Δεδομένα, Διατριβή Master, ΕΜΠ.

[3] Παναγιωτοπούλου Μαρία - Ελευθερία (2017). Μοντέλα Ευπάθειας με Εφαρμογή σε Δεδομένα Σχετικά με τον Καρκίνο, Διατριβή Master, ΕΜΠ.