

**ΠΑΝΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΟΙΝΩΝΙΚΩΝ ΚΑΙ ΠΟΛΙΤΙΚΩΝ
ΕΠΙΣΤΗΜΩΝ**

PANTEION UNIVERSITY OF SOCIAL AND POLITICAL SCIENCES



**ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΟΙΚΟΝΟΜΙΑΣ ΚΑΙ ΔΗΜΟΣΙΑΣ
ΔΙΟΙΚΗΣΗΣ**

ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΗΣ & ΠΕΡΙΦΕΡΕΙΑΚΗΣ ΑΝΑΠΤΥΞΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

**ΚΑΤΕΥΘΥΝΣΗ: ΕΦΑΡΜΟΣΜΕΝΑ ΟΙΚΟΝΟΜΙΚΑ ΚΑΙ
ΔΙΟΙΚΗΣΗ**

Η εφαρμογή μεθόδων της ανάλυσης επιβίωσης στη διάρκεια λειτουργίας επιχειρήσεων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Νικόλαος Καπετανέας

Αθήνα, 2019

Τριμελής Επιτροπή

Ριτσάρντσον Κλάιβ, Καθηγητής Παντείου Πανεπιστημίου (Επιβλέπων)

Ντεγιαννάκης Σταύρος, Αναπληρωτής Καθηγητής Παντείου Πανεπιστημίου

Στοφόρος Χρυσόστομος, Αναπληρωτής Καθηγητής Παντείου Πανεπιστημίου



Copyright©Νικόλαος Καπετανέας, 2019

All rights reserved. Με την επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας μεταπτυχιακής εργασίας εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της μεταπτυχιακής εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Η έγκριση μεταπτυχιακής εργασίας από το Πάντειον Πανεπιστήμιο Κοινωνικών και Πολιτικών Επιστημών δεν δηλώνει αποδοχή των γνώμων του συγγραφέα.

*Στην Άννα
και στα παιδιά μας
Γιώργο και Σωτήρη*

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον καθηγητή μου, κ. Ριτσάρντσον Κλάιβ, για την ανάθεση της εργασίας, την καθοδήγηση, τις υποδείξεις και την έμπρακτη υποστήριξή του, για την ολοκλήρωση αυτής της διπλωματικής εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω την οικογένεια μου για την υποστήριξη τους, την υπομονή τους και την αγάπη τους.

Περιεχόμενα

Περίληψη.....	7
Abstract	8
Κεφάλαιο πρώτο	9
Εισαγωγή στην Ανάλυση Επιβίωσης.....	9
1.1 Γενικά για την Ανάλυση Επιβίωσης.....	9
1.2 Αποκοπή δεδομένων (Censoring)	10
1.3 Συνάρτηση κατανομής	11
1.4 Ορολογία και συμβολισμοί	12
1.5 Κατανομές «Διάρκειας Ζωής»	12
Κεφάλαιο δεύτερο	13
Συναρτήσεις «Διάρκειας Ζωής».....	13
2.1 Η Αθροιστική Συνάρτηση κατανομής διάρκειας ζωής.....	13
2.2 Η συνάρτηση πυκνότητας/πιθανότητας διάρκειας ζωής	14
2.3 Η Συνάρτηση Επιβίωσης.....	15
2.4 Η Συνάρτηση Κινδύνου	17
2.5 Η αθροιστική συνάρτηση κινδύνου	19
2.6 Η συνάρτηση μέσης υπολειπόμενης ζωής.....	20
Κεφάλαιο τρίτο	21
Ειδικά παραμετρικά μοντέλα διάρκειας ζωής	21
3.1 Μέθοδοι ανάλυσης δεδομένων διάρκειας ζωής.....	21
3.2 Παραμετρικές Μέθοδοι	21
3.2.1 Η Εκθετική κατανομή	21
3.2.2 Η κατανομή Weibull	23
3.2.3 Η λογαριθμοκανονική κατανομή	24
3.2.4 Η γάμμα κατανομή	25
3.3 Μη-Παραμετρικές Μέθοδοι.....	25
3.3.1 Η μέθοδος Kaplan-Meier.....	25
3.3.2 Η Καμπύλη επιβίωσης.....	30
3.3.3 Υπολογισμός της εκτιμήτριας Kaplan – Meier	31
3.3.4 Διαστήματα Εμπιστοσύνης της Εκτιμήτριας Kaplan – Meier.....	33
3.4 Πίνακες Επιβίωσης (Life Tables).....	33
3.4.1 Πίνακες Επιβίωσης Πληθυσμού (Population Life Tables)	34

3.4.2 Κλινικοί Πίνακες Επιβίωσης – Η Αναλογιστική μέθοδος (Clinical Life Tables – The Actuarial method).....	34
Κεφάλαιο τέταρτο	38
Μη – παραμετρικές μέθοδοι για τη σύγκριση καμπυλών επιβίωσης	38
4.1 Το Logrank Test	38
4.1.1 Ο λόγος διακινδύνευσης	41
4.2 Το Cox - Mantel Test.....	41
Κεφάλαιο πέμπτο	44
Το μοντέλο αναλογικών κινδύνων του COX.....	44
5.1 Ορισμός του Μοντέλου	44
Κεφάλαιο 6	47
Μελέτη της «διάρκειας ζωής» εταιρειών στην Ιταλία	47
6.1 Δεδομένα μελέτης.....	47
6.2 Αποτελέσματα	48
6.2.1 Σύγκριση ανάμεσα στις περιοχές.....	48
6.2.2 Σύγκριση ανάμεσα σε διαφορετικές δραστηριότητες.....	50
6.2.3 Η χρήση του μοντέλου του Cox.....	53
6.4 Συμπεράσματα	61
Βιβλιογραφία	62

Περίληψη

Η ανάλυση επιβίωσης περιλαμβάνει την εξέταση της χρονικής διάρκειας μεταξύ ενός σταθερού χρονικού σημείου έναρξης (π.χ. διάγνωση καρκίνου, έναρξη λειτουργίας μιας επιχείρησης) και τερματισμού (π.χ. θάνατος, τερματισμού των εργασιών μιας επιχείρησης). Το βασικό χαρακτηριστικό που διακρίνει αυτά τα δεδομένα από άλλους τύπους είναι ότι το γεγονός (event) είναι πιθανόν να μη συμβεί αναγκαστικά σε όλες τις μεμονωμένες περιπτώσεις μέχρι τη λήξη της μελέτης και γι' αυτές τις περιπτώσεις οι πλήρεις χρόνοι επιβίωσης τους είναι άγνωστοι. Επομένως, απαιτούνται ειδικές μέθοδοι για αυτούς τους τύπους δεδομένων. Στη μελέτη μας, εστιάζουμε αρχικά στη μέθοδο Kaplan-Meier, και η οποία χρησιμοποιείται περισσότερο από άλλες ως μέθοδο ανάλυσης επιβίωσης. Επιτρέπει τον υπολογισμό της πιθανότητας να συμβεί ένα γεγονός (event) όπως είναι ο θάνατος ενός ασθενή ή μια εταιρεία να διακόψει οριστικά τη λειτουργία της. Με την Kaplan-Meier ανάλυση εξηγείται πώς λειτουργεί η μέθοδος, πώς υπολογίζονται οι πιθανότητες επιβίωσης και πώς τα δεδομένα επιβίωσης μπορούν να συνοψιστούν σε ομάδες και να συγκριθούν χρησιμοποιώντας έλεγχο υποθέσεων με το logrank test. Επιπλέον, παρέχει σημαντικές πληροφορίες σχετικά με τις καμπύλες επιβίωσης. Δεύτερον, πραγματοποιήσαμε μια ανάλυση παλινδρόμησης του Cox και πιο συγκεκριμένα για τη μεταβλητή ανάλογων κινδύνων. Συζητήσαμε μεθόδους για τον έλεγχο της σωστής εφαρμογής των αποτελεσμάτων τόσο με τις γραφικές παραστάσεις όσο και με την προσαρμογή των μοντέλων στα αποτελέσματα. Στην ανάλυσή μας μελετήσαμε την ανάλυση επιβίωσης, τόσο με μη παραμετρικά μοντέλα (μέθοδο Kaplan-Meier) όσο και με ημι-παραμετρικά (μοντέλο παλινδρόμησης Cox) για τη διάρκεια λειτουργίας των επιχειρήσεων σε δύο περιοχές της Ιταλίας και παρουσιάσαμε τις βασικές έννοιες της ανάλυσης επιβίωσης, συμπεριλαμβανομένου και της ερμηνείας των καμπυλών επιβίωσης μεταξύ των επιχειρήσεων των δύο περιοχών. Τα αποτελέσματά μας έδειξαν ότι υπάρχει πλεονέκτημα στη διάρκεια επιβίωσης για τις επιχειρήσεις που δραστηριοποιούνται στην περιοχή PG σε σχέση με την περιοχή TR, καθώς και πλεονέκτημα στη διάρκεια επιβίωσης για ορισμένες δραστηριότητες εταιρειών έναντι άλλων. Τέλος, και λόγω του ότι η διακοπή λειτουργίας των εταιριών (event) μπορεί να οφείλεται σε διαφορετικούς λόγους, εξετάζουμε επίσης τις διαφορετικές αιτίες κλεισίματος σε μια ανταγωνιστική ανάλυση κινδύνων.

Λέξεις κλειδιά: Ανάλυση επιβίωσης, Kaplan-Meier ανάλυση, μοντέλο παλινδρόμησης του Cox, μοντέλο αναλογικών κινδύνων

Abstract

Survival analysis involves the consideration of the time between a fixed starting point of time (e.g. diagnosis of cancer, start of a firm's operations) and a terminating event (e.g. death, termination of a firm's operations). The key feature that distinguishes such data from other types is that the event will not necessarily have occurred in all individual cases by the time the study ends, and for these cases their exact survival times are unknown. Thus, special methods are required for these types of data. In our study, first we focus on the Kaplan–Meier method, the most popular method used for survival analysis. It makes it possible to calculate the incidence rate of events like deaths or closure of operation, by using information from all subjects at risk for these events. We explain how the K-M method works, and how survival probabilities are calculated, survival data can be summarized and survival in groups can be compared using the logrank test for hypothesis testing. In addition, it provides important information regarding the survival curve over time. Secondly, we describe Cox's proportional hazards model for survival analysis. We discuss methods for checking the overall fit of the results with both graphical and formal goodness of fit tests. In our application of survival analysis, both non-parametric (Kaplan-Meier method) and semi-parametric (Cox regression model) analyses for firms' length of operation in two districts of Italy is carried out. We present and interpret survival curves and quantify and test survival difference between the firms of the two districts. Our results showed that there appears to be a survival advantage for the firms which operate in PG district compared to TR and also survival advantage for some branches of economic activities over others. Furthermore, we examine different causes of closure in a competing risks analysis.

Keywords: Survival analysis, Kaplan-Meier method, Cox's regression analysis, Proportional hazard analysis

Κεφάλαιο πρώτο

Εισαγωγή στην Ανάλυση Επιβίωσης

1.1 Γενικά για την Ανάλυση Επιβίωσης

Η μελέτη των περιπτώσεων όπου η χρονική διάρκεια μέχρι να συμβεί ένα γεγονός ή αποτέλεσμα με θετικές ή αρνητικές συνέπειες έχει ιδιαίτερη σημασία τόσο σε προσωπικό επίπεδο όσο και στο επίπεδο των επιστημών (υγείας, βιολογίας, μηχανολογίας, οικονομικών κ.α.).Ως παραδείγματα θα μπορούσαν να αναφερθούν ενδεικτικά η ανθρώπινη διάρκεια ζωής, η διάρκεια ζωής από τη διάγνωση μιας θανατηφόρου ασθένειας, η διάρκεια ενός γάμου, η διάρκεια παραμονής ενός εργαζόμενου στην εργασία ή επανασύλληψη ενός αποφυλακισμένου, κτλ.

Για την ανάλυση αυτής της μορφής των δεδομένων έχει αναπτυχθεί εκτεταμένη στατιστική θεωρία, η οποία ονομάζεται ανάλυση αξιοπιστίας (reliability analysis), όταν πρόκειται για εφαρμογές θετικών επιστημών, ή ανάλυση επιβίωσης (survival analysis), όταν πρόκειται για εφαρμογές βιοϊατρικής.

Ακριβέστερα στην ανάλυση επιβίωσης η μελέτη αφορά τις μεταβλητές οι οποίες αντιπροσωπεύουν χρονικές διάρκειες των οποίων ο τερματισμός χαρακτηρίζεται από την εμφάνιση κάποιου ενδεχόμενου, το οποίο ανεξάρτητα της θετικής ή αρνητικής του συνέπειας, ονομάζεται αποτυχία.(Καλαματιανού, 2001).

Η αποτυχία δηλαδή, μπορεί να αντιστοιχεί στο ενδεχόμενο θανάτου, της καταστροφής, της εμφάνισης μιας ασθένειας, της έκδοσης ενός διαζυγίου, στο ενδεχόμενο της ανάρρωσης, της εύρεσης απασχόλησης κ.α. Σε κάθε περίπτωση η αποτυχία δηλώνει τον τερματισμό μιας διάρκειας και η διάρκεια αυτή στο χώρο της ανάλυσης επιβίωσης ονομάζεται γενικώς διάρκεια ζωής ή επιβίωσης.

Αν και θα μπορούσε να υποστηριχθεί ότι η ανάλυση δεδομένων διάρκειας ζωής μπορεί να πραγματοποιηθεί στα βήματα της ανάλυσης δεδομένων σύμφωνα με την κλασική στατιστική εν τούτοις υπάρχουν δύο αντικειμενικές δυσκολίες.

Τα δεδομένα -χρονικές διάρκειες- έχουν πάντα θετικές τιμές και συχνά εμφανίζουν δεξιά (θετική) Λοξότητα Αυτό σημαίνει ότι για την εκτίμηση της κατανομής πρέπει να χρησιμοποιηθούν μετασχηματισμοί ή μη παραμετρικές διαδικασίες.

Το δεύτερο σημείο που διαφοροποιεί τις μεθόδους ανάλυσης επιβίωσης είναι η παρουσία της αποκοπής των δεδομένων, δηλαδή όταν με τη λήξη της περιόδου παρατήρησης κάποιες μονάδες συνεχίζουν να λειτουργούν και επομένως δεν είναι ακριβής ο χρόνος διάρκειας της λειτουργίας τους ή της επιβίωσής τους.

1.2 Αποκοπή δεδομένων (Censoring)

Στην ανάλυση επιβίωσης, όπως ήδη έχουμε αναφέρει παραπάνω, έχουμε αποκομμένα δεδομένα όταν υπάρχουν παρατηρήσεις, των οποίων οι χρόνοι διάρκειας ζωής δεν είναι ακριβείς. Έχουμε δηλαδή μερική πληροφορία όταν ένα πείραμα -κατά το οποίο καταγράφεται ο χρόνος λειτουργίας ενός δείγματος μονάδων-τερματίζεται, πολύ πιθανό μερικές μονάδες να λειτουργούν ακόμη. Αν και δεν γνωρίζουμε τη διάρκεια ζωής των μονάδων του δείγματος, διαθέτουμε την πληροφορία ότι η διάρκεια λειτουργίας της μονάδας έχει ξεπεράσει τη χρονική διάρκεια κατά την οποία η μονάδα ήταν στο πείραμα.

Συμβολίζουμε με (T) την τυχαία μεταβλητή διάρκειας ζωής μιας μονάδας. Λόγω του ότι (T) συμβολίζει χρόνο, σε όλες τις περιπτώσεις οι πιθανές τιμές θα είναι ίσες ή μεγαλύτερες του μηδενός. Επίσης, συμβολίζουμε με (t) κάθε τιμή ενδιαφέροντος της τυχαίας μεταβλητής (T) . Για παράδειγμα, αν θέλουμε να εξακριβώσουμε αν ένας ασθενής επιβιώνει περισσότερο από πέντε έτη μετά από μια θεραπεία, τότε $t=5$ και το ερώτημα που τίθεται είναι εάν T είναι μεγαλύτερο του 5.

Υπάρχουν τρεις τρόποι αποκοπής :

I.Αποκοπή από δεξιά (Right censoring)

Αποκοπή από δεξιά έχουμε όταν το γεγονός που ενδιαφέρει δεν λαμβάνει χώρα πριν τον τερματισμό της μελέτης. Είναι ο πιο συνηθισμένος τύπος αποκοπής και διακρίνεται σε:

Συστηματική (ή σταθερή), όπου καθορίζεται από τη διαδικασία της μελέτης.

Τυχαία όπου καθορίζεται από ένα μηχανισμό εκτός ελέγχου του ερευνητή.

Έστω ένα πείραμα το οποίο γίνεται με 25 ρουλεμάν ίδιου τύπου όπου δοκιμάζονται με το ίδιο φορτίο (Caroni, 2002). Καταγράφεται ο αριθμός των περιστροφών μέχρι να υποστούν βλάβη. Ο ερευνητής αποφασίζει να τερματίσει το πείραμα μετά από 100.000 περιστροφές για παράδειγμα. Ο χρόνος παρατήρησης είναι κοινός για όλα τα

υποκείμενα της έρευνας (αριθμός περιστροφών). Σε αυτή την περίπτωση παρατηρείται αποκοπή τύπου I.

Αν στη ίδια έρευνα ο ερευνητής αποφάσιζε τη διακοπή της έρευνας για συγκεκριμένο αριθμό αποτυχιών, για παράδειγμα αν τα 4/6 των ρουλεμάν παρουσίασαν βλάβη τότε η αποκοπή θα ήταν τύπου II (Caroni, 2002).

II Αποκοπή από αριστερά (Left censoring)

Έχουμε αποκοπή από αριστερά όταν έχουμε καθυστερημένη εισαγωγή των υποκείμενων στη μελέτη.

III Αποκοπή Διαστήματος (Interval censoring)

Έχουμε αποκοπή διαστήματος όταν υπάρχουν διαστήματα κατά τη διάρκεια της μελέτης κατά τα οποία οι υποκείμενοι δεν παρακολουθούνται συνεχώς αλλά μόνο (π.χ) μια φορά την ημέρα ή εβδομάδα.

Πέραν του παραπάνω διαχωρισμού η αποκοπή ανάλογα με τα χαρακτηριστικά της μπορεί να υποστεί και άλλους διαχωρισμούς. Σε κάθε περίπτωση η υπό εξέταση μονάδα εισέρχεται στη μελέτη σε χρόνο t_0 και παρακολουθείται μέχρι και την προκαθορισμένη χρονική στιγμή, έστω t_c . Η διάρκεια παρακολούθησης της μελέτης είναι γνωστή, έστω $c = t_c - t_0$. Με τη λήξη της περιόδου παρατήρησης μπορούμε να έχουμε πειραματικές μονάδες για τις οποίες καταγράφηκαν αποτυχίες, π.χ. $t_1 < c$, $t_2 < c$, $t_3 < c$. Επίσης, κάποιες πειραματικές μονάδες δεν εκδήλωσαν την αποτυχία κατά τη διάρκεια της περιόδου παρακολούθησης, π.χ. $t_4 > c$, $t_5 > c$, $t_6 > c$. Με βάση τα παραπάνω γίνεται αποκοπή στον χρόνο $t_0 + c$, αυτή λοιπόν η περίπτωση κατά την οποία γνωρίζουμε ότι ο χρόνος ζωής της υπό εξέταση μονάδας είναι μεγαλύτερος από κάποιο χρόνο t_c αντιστοιχεί στη λεγόμενη δεξιά αποκοπή. Ονομάζεται δεξιά αποκοπή επειδή το γεγονός που δεν καταγράφηκε βρίσκεται στα δεξιά του χρόνου δηλαδή γνωρίζουμε ότι το γεγονός δεν έχει συμβεί στο τέλος της έρευνας. Δηλαδή:

$$Y_i = T_i \text{ αν } T_i \leq c \text{ or } c \text{ αν } c < T_i$$

Σημειώνεται ότι η συνάρτηση κατανομής των Y έχει θετική μάζα στο σημείο $y=c$.

1.3 Συνάρτηση κατανομής

Έστω T η συνάρτηση διάρκειας ζωής με συνάρτηση πυκνότητας πιθανότητας $f(t)$, $t \geq 0$.

Η συνάρτηση κατανομής ορίζεται ως:

$$F(t) = P[T \leq t] = \int_0^t f(u) du$$

Από τα παραπάνω προκύπτει ότι η $f(t)$ είναι αύξουσα,

$$\lim_{t \rightarrow 0} F(t) = 0, \lim_{t \rightarrow \infty} F(t) = 1.$$

1.4 Ορολογία και συμβολισμοί

Πλήρη Δεδομένα

Πλήρη στοιχεία ή δεδομένα είναι τα στοιχεία που η διάρκεια της κάθε μονάδας του δείγματος που παρατηρείται είναι γνωστή. Τα πλήρη δεδομένα είναι πολύ πιο εύκολο να επεξεργαστούν από ότι τα αποκομμένα δεδομένα.

1.5 Κατανομές «Διάρκειας Ζωής»

Η τυχαία μεταβλητή της διάρκειας ζωής, όπως και κάθε μεταβλητή, περιγράφεται πλήρως από την κατανομή της και από μερικές ακόμα παραμέτρους που περιγράφουν περισσότερο περιγραφικά την κατανομή. Η κατανομή αυτή προσδιορίζεται στη βάση συναρτήσεων συγκεκριμένης μορφής με τις οποίες περιγράφεται η κατανομή της διάρκειας ζωής και η μορφή αυτών των συναρτήσεων θα έχουν παραμετρικό ή μη παραμετρικό χαρακτήρα.

Κεφάλαιο δεύτερο

Συναρτήσεις «Διάρκειας Ζωής»

Η κατανομή του χρόνου επιβίωσης, χαρακτηρίζεται από έξι συναρτήσεις, τη Συνάρτηση Κατανομής μιας τυχαίας μεταβλητής καθώς και την Αθροιστική Συνάρτηση Κατανομής της, τη Συνάρτηση Πυκνότητας Πιθανότητας (probability density function), τη Συνάρτηση Επιβίωσης (survivor function), τη Συνάρτηση Κινδύνου (hazard function), την Αθροιστική Συνάρτηση Κινδύνου (cumulative hazard function) και τη Συνάρτηση Μέσης Υπολειπόμενης Ζωής (mean residual lifetime).

2.1 Η Αθροιστική Συνάρτηση κατανομής διάρκειας ζωής

Η αθροιστική συνάρτηση κατανομής της τυχαίας μεταβλητής διάρκειας ζωής, T , συμβολίζεται με F και ορίζεται η συνάρτηση με πεδίο ορισμού το σύνολο των μη αρνητικών πραγματικών αριθμών και πεδίο τιμών το διάστημα $[0,1]$ και η οποία ικανοποιεί τη σχέση $F(t) = P(T \leq t)$ για κάθε $t \geq 0$, όπου P συμβολίζει την πιθανότητα να συμβεί το ενδεχόμενο. Δηλαδή, είναι μια συνάρτηση κατανομής μιας τυχαίας μεταβλητής που όμως στην περίπτωση που εξετάζουμε η τυχαία μεταβλητή αντιστοιχεί στη διάρκεια ζωής και συμβολίζεται ως T .

Σύμφωνα με τους Mood, Graybill & Boes (1974) η εν λόγω συνάρτηση είναι μια συνάρτηση κατανομής, εφόσον μας εξηγεί πως οι τιμές μιας τυχαίας μεταβλητής κατανέμονται και επίσης είναι αθροιστική εφόσον δίνει την κατανομή σε αθροιστική μορφή.

Για παράδειγμα αν η τυχαία μεταβλητή T εκφράζει ανθρώπινη ηλικία σε έτη, μέχρι το ενδεχόμενο του θανάτου και έστω $t=50$, τότε η $F(50)=P(T \leq 50)$, εκφράζει τη πιθανότητα ένα άτομο τυχαία επιλεγμένο από τον πληθυσμό να επιβιώσει μέχρι την ηλικία των πενήντα ετών. Επίσης εκφράζει και την αναμενόμενη αναλογία των μελών του πληθυσμού που επιβιώνουν μέχρι την ηλικία των πενήντα ετών.

Κάθε αθροιστική συνάρτηση κατανομής διάρκειας ζωής ικανοποιεί τις σχέσεις

$F(0) = 0$ και $F(\infty) = 1$ και επίσης είναι μια μονότονη μη φθίνουσα συνάρτηση του t , δηλαδή αν,

$$t_1 < t_2 \text{ τότε ισχύει } F(t_1) \leq F(t_2).$$

Από τα παραπάνω προκύπτει, ότι αυξανόμενου του χρόνου, η πιθανότητα η τυχαία

μεταβλητή διάρκειας ζωής να πάρει μεγαλύτερη τιμή αυξάνεται ή παραμένει η ίδια. Αν για παράδειγμα υποθέσουμε δύο διάρκειες ανθρώπινης ζωής $t=50$ και $t=60$ όπου ισχύει $t_1 < t_2$, τότε σύμφωνα με τα παραπάνω ισχύει ότι $F(t_1) \leq F(t_2)$ δηλαδή, η πιθανότητα ένα άτομο να πεθάνει πριν τα εξήντα έτη είναι μεγαλύτερη ή ίση με την πιθανότητα το άτομο να πεθάνει πριν τα πενήντα έτη.

Επίσης, αν $F(t_0)=0$ τότε επίσης ισχύει ότι $F(t)=0$ για κάθε $t \leq t_0$

Για παράδειγμα, αν θεωρήσουμε ότι η τυχαία μεταβλητή δηλώνει τι διάρκεια έχει την επανασύλληψη ενός κακοποιού, τότε αν γνωρίζουμε ότι $F(30) = P(T \leq 30) = 0$, δηλαδή ότι ένα άτομο δεν έχει συλληφθεί μέσα σε διάστημα 30 ετών τότε επίσης ισχύει ότι το άτομο δεν έχει συλληφθεί και σε όλο το διάστημα που προηγείται αυτού, δηλαδή $F(20) = P(T \leq 20) = 0$.

2.2 Η συνάρτηση πυκνότητας/πιθανότητας διάρκειας ζωής

Όταν η διάρκεια ζωής T είναι μια διακριτή τυχαία μεταβλητή με τιμές $t_1, t_2, \dots, t_n, \dots \in [0, \infty)$, τότε η συνάρτηση πιθανότητας διάρκειας ζωής που συμβολίζεται με f ικανοποιεί τη σχέση:

$$f(t) = \begin{cases} P(T = t_i) & \text{αν } t = t_i, i = 1, 2, \dots, n, \dots \\ 0 & \text{αν } t \neq t_i \end{cases}$$

Η συνάρτηση πιθανότητας διάρκειας ζωής είναι μια συνάρτηση με πεδίο ορισμού το σύνολο των μη αρνητικών πραγματικών αριθμών με πεδίο τιμών το διάστημα $[0,1]$ και οι τιμές της εκφράζουν τις πιθανότητες με τις οποίες η τυχαία μεταβλητή διάρκειας ζωής T μπορεί να πάρει τις συγκεκριμένες διακριτές τιμές.

Οι ιδιότητες της συνάρτησης πιθανότητας διάρκειας ζωής είναι οι εξής:

α. $f(t_i) \geq 0$ για κάθε $i=1,2,\dots$ όπου $t_i \geq 0$

β. $\sum_i f(t_i) = 1$ όπου $t_i \geq 0$, και το άθροισμα καλύπτει το σύνολο των σημείων $t_i, i=1,2,\dots,n,\dots$

Η ιδιότητα αυτή εκφράζει το γεγονός ότι τα ενδεχόμενα η τυχαία μεταβλητή διάρκεια ζωής να πάρει κάθε μια από τις διακριτές τιμές $t_1, t_2, \dots, t_n, \dots$ είναι αμοιβαίως αποκλειόμενα ενδεχόμενα και το σύνολο τους αποτελεί το βέβαιο ενδεχόμενο.

Η σχέση συνάρτησης πιθανότητας διάρκειας ζωής και αθροιστικής συνάρτησης κατανομής διάρκειας ζωής ορίζεται από τη σχέση ότι η

$$F(t_r) = \sum_{i=1}^{i=r} f(t_i)$$

Επίσης, αν υποθέσουμε τώρα ότι η F είναι γνωστή, τότε η

$$f(t_i) = F(t_i) - \lim_{h \rightarrow 0} F(t_i - h)$$

με h μεγαλύτερο από το μηδέν και να τείνει στο μηδέν. ($0 < h \rightarrow 0$).

Όταν η διάρκεια ζωής είναι μια συνεχής τυχαία μεταβλητή, τότε η συνάρτηση που συμβολίζεται με f είναι παράγωγος της αθροιστικής συνάρτησης κατανομής διάρκειας ζωής ως προς t , δηλαδή,

$$f(t) = dF(t)/dt (=F'(t))$$

για όλες τις τιμές t της T που υπάρχει παράγωγος.

2.3 Η Συνάρτηση Επιβίωσης

Συμβολίζοντας το χρόνο επιβίωσης με T , η συνάρτηση επιβίωσης (survival function) $S(t)$ ορίζεται ως η πιθανότητα επιβίωσης ενός ατόμου πέραν τη χρονική στιγμή t και δίνεται από τη σχέση :

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

Η συνάρτηση επιβίωσης είναι μη αρνητική και μη αύξουσα συνάρτηση του t με $S(0) = 1$ και $S(\infty) = 0$. Η γραφική παράσταση της $S(t)$ συναρτήσει του t είναι γνωστή ως καμπύλη επιβίωσης και είναι πολύ σημαντική στην ανάλυση δεδομένων χρόνου επιβίωσης.

Η σχέση της S με την F και την f

Η σχέση της συνάρτησης επιβίωσης με την αθροιστική συνάρτηση κατανομής διάρκειας ζωής είναι ότι η συνάρτηση επιβίωσης είναι η συμπληρωματική συνάρτηση της αθροιστικής συνάρτησης κατανομής.

Οι παραπάνω σχέσεις ορισμού των εν λόγω συναρτήσεων υποδεικνύουν ότι η συνάρτηση επιβίωσης μπορεί να προσδιοριστεί ως

$$f(t) = -dS(t)/dt$$

δηλαδή, από την αθροιστική συνάρτηση κατανομής διάρκειας ζωής ή από τη συνάρτηση πιθανότητας /πυκνότητας διάρκειας ζωής και αντιστρόφως.

Γενικά η συνάρτηση επιβίωσης εκφράζει την πιθανότητα «επιβίωσης» πέραν μιας συγκεκριμένης χρονικής στιγμής t , δηλαδή εκφράζει την πιθανότητα η τυχαία μεταβλητή διάρκεια ζωής T , να παίρνει τιμές μεγαλύτερες μιας συγκεκριμένης τιμής t . Για παράδειγμα, αν υποθέσουμε ότι η T συμβολίζει την ηλικία ενός ατόμου μέχρι το ενδεχόμενο του θανάτου, τότε η τιμή της συνάρτησης επιβίωσης $S(50) = P(T > 50)$ μπορεί να ερμηνευθεί είτε ως πιθανότητα ώστε ένα άτομο να επιβιώσει πέραν της ηλικίας των 50 ετών, είτε ως η αναμενόμενη αναλογία των μελών ενός πληθυσμού που επιβιώνουν πέραν της ηλικίας των 50 ετών.

Οι ιδιότητες της συνάρτησης επιβίωσης οριοθετούν τις τιμές της συνάρτησης και εξυπηρετούν στον υπολογισμό και στην ερμηνεία των τιμών της.

Δηλαδή,

Κάθε συνάρτηση επιβίωσης ικανοποιεί τις σχέσεις:

$$S(0)=1, \text{ και } S(\infty)=0$$

Η πρώτη σχέση δείχνει ότι το ενδεχόμενο επιβίωσης μιας πειραματικής μονάδας πέραν της στιγμής μηδέν είναι ένα βέβαιο ενδεχόμενο και συνεπώς η πιθανότητα επιβίωσης πέραν της αρχής αυτής είναι ίση με 1, ενώ από την δεύτερη σχέση προκύπτει ότι το ενδεχόμενο ώστε μια μονάδα να επιβιώσει πέραν του απείρου (δηλαδή να επιβιώνει για πάντα) είναι αδύνατον και η πιθανότητα είναι ίση με μηδέν.

Επίσης, αυξανόμενου του χρόνου η πιθανότητα ώστε μια μονάδα να επιβιώσει πέραν ενός χρονικού σημείου δεν αυξάνεται, αντίθετα μειώνεται ή παραμένει ίδια. Για παράδειγμα αν η τυχαία μεταβλητή T δηλώνει διάρκεια έγγαμου βίου σε έτη, και $t_1=10 < t_2=15$ είναι δύο τιμές αυτής της μεταβλητής, τότε ισχύει $S(10) \geq S(15)$. Δηλαδή, η πιθανότητα ο έγγαμος βίος ενός ατόμου να διαρκέσει πέραν των 15 ετών είναι μικρότερη ή το πολύ ίση της πιθανότητας να διαρκέσει περισσότερο από 10 έτη.

Θεωρητικά, το διάγραμμα της συνάρτησης επιβίωσης ως προς το χρόνο T , είναι μια συνεχής καμπύλη που ξεκινά από το σημείο $(0,1)$, έχει καθοδική πορεία και καταλήγει στο σημείο $(\infty, 0)$, όπου και συναντά τον άξονα του χρόνου.

2.4 Η Συνάρτηση Κινδύνου

Η συνάρτηση κινδύνου, $h(t)$, ορίζεται ως η πιθανότητα απεβίωσης (ή πραγμάτωσης του γεγονότος που εξετάζεται) τη χρονική στιγμή t , δεδομένου ότι το άτομο έχει επιβιώσει μέχρι τη χρονική στιγμή t .

Δηλαδή,

$$h(t) = \lim_{dt \rightarrow 0} P(t < T \leq t + dt \mid T > t)/dt$$

όπου $t \geq 0$

Η συνάρτηση κινδύνου δίνει ένα μέτρο του πόσο πιθανό είναι ένα άτομο να αποβιώσει ως συνάρτηση της ηλικίας του ατόμου, για παράδειγμα ο κίνδυνος θανάτου ανάμεσα σε αυτούς που είναι ζωντανοί τη συγκεκριμένη στιγμή.

Οι Cox & Oakes (1982), αναφέρουν ότι η συνάρτηση κινδύνου είναι ο πιο διαδεδομένος τρόπος μελέτης του μοντέλου της κατανομής της διάρκειας ζωής στις βιοϊατρικές επιστήμες. Δηλαδή, αν έχουμε μια τυχαία μεταβλητή της διάρκειας ζωής T την οποία θεωρούμε ως μια συνεχή τυχαία μεταβλητή, t μια τιμή της T και dt ένα πολύ μικρό διάστημα τιμών της T , τότε η πιθανότητα ώστε η T να πάρει τιμές στο διάστημα $(t, t+dt]$ ή η πιθανότητα το ενδεχόμενο της αποτυχίας να εμφανιστεί στο διάστημα $(t, t+dt]$ είναι $P(t < T \leq t + dt)$ και μπορεί να υπολογιστεί όταν είναι γνωστή η συνάρτηση πυκνότητας διάρκειας ζωής ή η συνάρτηση επιβίωσης.

Δηλαδή,

$$P(t < T \leq t + dt) = \int_t^{t+dt} f(u) du = F(t+dt) - F(t) = S(t) - S(t+dt)$$

Αν υποθέσουμε ότι το ενδεχόμενο της αποτυχίας δεν έχει εμφανιστεί, δηλαδή $T > t$, τότε η πιθανότητα το ενδεχόμενο της αποτυχίας να εμφανιστεί στο διάστημα $(t, t + dt]$ θα καθορίζεται από τη σχέση,

$$P(t < T \leq t + dt \mid T > t)/dt = P(t < T \leq t + dt)/P(T > t),$$

Διαιρώντας την παραπάνω πιθανότητα με το μήκος του διαστήματος $(t, t+dt)$, δηλαδή με το dt έχουμε,

$$P(t < T \leq t + dt \mid T > t)/dt$$

με την οποία σχέση παίρνουμε ένα μέσο ρυθμό αποτυχίας, δηλαδή τις αποτυχίες που αντιστοιχούν στη μονάδα χρόνου.

Κλείνουμε τη θεωρητική προσέγγιση αν πάρουμε το όριο του παραπάνω ρυθμού καθώς και το μήκος του διαστήματος dt τείνει στο μηδέν, τότε ορίζουμε τον στιγμιαίο ρυθμό αποτυχίας που ταυτίζεται με τη συνάρτηση κινδύνου.

Η σημασία των τιμών της συνάρτησης κινδύνου

Ο Kleinbaum (1995), ερμηνεύει τη συνάρτηση κινδύνου $h(t)$, ως τη στιγμιαία πιθανότητα ανά μονάδα χρόνου ώστε μία πειραματική μονάδα να εμφανίσει το ενδεχόμενο «αποτυχία», δοθέντος ότι έχει επιβιώσει μέχρι και τη χρονική στιγμή t . Επίσης, η σημασία της συνάρτησης κινδύνου μπορεί να αποδοθεί και ως δεσμευμένη πιθανότητα με τον αριθμητή του κλάσματος που βρίσκεται μέσα στο όριο να είναι η δεσμευμένη πιθανότητα ώστε η αποτυχία να εμφανιστεί κατά τη διάρκεια ενός διαστήματος μήκους dt χρονικών μονάδων, το οποίο έπεται της χρονικής στιγμής t , λαμβάνοντας υπόψη ότι η αποτυχία δεν έχει εμφανιστεί μέχρι και τη χρονική στιγμή t .

Συνεπώς, μπορούμε να πούμε ότι η συνάρτηση κινδύνου, βασίζεται στην εμφάνιση του ενδεχόμενου που αποκαλούμε «αποτυχία», σε αντίθεση με τη συνάρτηση επιβίωσης που στηρίζεται στη μη εμφάνιση αυτού του ενδεχόμενου. Με άλλα λόγια, η συνάρτηση κινδύνου δίνει την αρνητική πλευρά της πληροφορίας που δίνει η συνάρτηση επιβίωσης.

Επίσης, η συνάρτηση κινδύνου γενικότερα κυρίως είναι ρυθμός και λιγότερο πιθανότητα και ουσιαστικά είναι ο λόγος δύο ποσοτήτων. Δηλαδή, είναι ο λόγος μιας δεσμευμένης πιθανότητας προς ένα χρονικό διάστημα και το αποτέλεσμα είναι δεσμευμένη πιθανότητα στη μονάδα χρόνου. Όταν διαιρέσουμε την πιθανότητα με το μήκος χρονικού διαστήματος δεν παίρνουμε πιθανότητα, αλλά ρυθμό. Η κλίμακα μέτρησης του ρυθμού κυμαίνεται από το μηδέν μέχρι το άπειρο, ενώ της πιθανότητας από το μηδέν μέχρι το ένα και εξαρτάται από την κλίμακα μέτρησης του χρόνου. Για παράδειγμα, αν η δεσμευμένη πιθανότητα είναι 0,40 και το χρονικό διάστημα είναι 15 ημέρες (μισός μήνας), τότε ο ρυθμός είναι $(0,40)/1/2=0,80/\text{μήνα}$. Ουσιαστικά, η δεσμευμένη αυτή πιθανότητα εκτιμάται με βάση τη σχετική πυκνότητα των πειραματικών μονάδων που εμφανίζουν το ενδεχόμενο της αποτυχίας στη διάρκεια ενός χρονικού διαστήματος. Για παράδειγμα, αν η T συμβολίζει διάρκεια ανθρώπινης

ζωής μέχρι το ενδεχόμενο του θανάτου, τότε μια τιμή της συνάρτησης κινδύνου έστω $h(90) = 0,80/\mu\eta\gamma\alpha$, σημαίνει ότι τα άτομα ηλικίας 90 ετών αναμένεται να πεθαίνουν με ρυθμό 0,80 το μήνα.

Από τα παραπάνω προκύπτει ότι οι τιμές της συνάρτησης κινδύνου είναι πάντα μεγαλύτερες ή ίσες του μηδενός και ότι οι τιμές της $h(t)$ δεν έχουν ανώτατο όριο.

Η ανάλυση της συνάρτησης κινδύνου έχει σημαντική συνεισφορά στη μελέτη της ανάλυσης επιβίωσης γιατί:

α) μπορεί να ερμηνευτεί ως το «ποσό του κινδύνου» που εμπεριέχει μια δεδομένης ηλικίας πειραματική μονάδα να εμφανίσει το ενδεχόμενο της αποτυχίας στο αμέσως επόμενο χρονικό διάστημα.

β) Η συνάρτηση κινδύνου είναι χρήσιμη όταν θέλουμε να συγκρίνουμε τον τρόπο αλλαγής του προαναφερόμενου κινδύνου στη διάρκεια του χρόνου, μεταξύ πληθυσμών.

γ) Η συνάρτηση κινδύνου εμβαθύνει περισσότερο στον υπό-συνθήκη ρυθμό αποτυχίας ή στον υπό- συνθήκη τερματισμό της διάρκειας ζωής ή λειτουργίας.

2.5 Η αθροιστική συνάρτηση κινδύνου

Ακόμη μια συνάρτηση η οποία περιγράφει την κατανομή των δεδομένων της διάρκειας ζωής είναι η Αθροιστική Συνάρτηση Κινδύνου (cumulative hazard function) και στην περίπτωση που έχουμε συνεχή τυχαία μεταβλητή T , συμβολίζεται ως H και ορίζεται ως:

$$H(t) = \int_0^t h(u) du \text{ όπου } t \geq 0$$

Στην περίπτωση όπου η τυχαία μεταβλητή T είναι διακριτή, τότε η αθροιστική συνάρτηση διακινδύνευσης της T συμβολίζεται ως

$$H(t) = \sum_{t_i < t} h(t_i) \text{ όπου } t \geq 0$$

Από τα παραπάνω συμπεραίνουμε ότι οι αθροιστική συνάρτηση κινδύνου έχει τις παρακάτω ιδιότητες:

α) κάθε αθροιστική συνάρτηση κινδύνου μπορεί να πάρει τιμές από το μηδέν μέχρι το άπειρο

$$H(0)=0 \text{ και } H(\infty) = \infty$$

β) είναι μη-φθίνουσα συνάρτηση του t , δηλαδή:

$$\text{αν } t_1 < t_2 \text{ τότε } H(t_1) \leq H(t_2)$$

γ) αν T είναι συνεχής τότε η αθροιστική συνάρτηση κινδύνου διάρκειας ζωής είναι συνεχής από δεξιά, δηλαδή:

$$\lim_{h \rightarrow 0^+} H(t+h) = H(t)$$

2.6 Η συνάρτηση μέσης υπολειπόμενης ζωής

Τελευταία μορφής συνάρτησης που περιγράφει την κατανομή δεδομένων διάρκειας ζωής είναι η συνάρτηση μέσης υπολειπόμενης ζωής (mean residual life function) και ορίζεται για συνεχής και διακριτές τυχαίες μεταβλητές T ως:

$$L(t) = E[T-t \mid T > t] \text{ για } t \geq 0$$

Με τον παραπάνω ορισμό αποτυπώνεται ότι η αναμενόμενη τιμή της τυχαίας μεταβλητής της διάρκειας ζωής T , που υπολογίζεται για εκείνες τις τιμές της διάρκειας ζωής που είναι μεγαλύτερες μιας δοθείσας τιμής. Με άλλα λόγια, η μέση υπολειπόμενη ζωή εκφράζει το πόσο ακόμα αναμένεται να επιβιώσει μια πειραματική μονάδα λαμβάνοντας υπόψη το γεγονός ότι αυτή η μονάδα έχει ήδη διανύσει μια συγκεκριμένη διάρκεια ζωής. Για παράδειγμα ας υποθέσουμε ότι η διάρκεια ζωής εκφράζει ανθρώπινη ζωή, μετρούμενη σε έτη, τότε η τιμή $L(80) = E(T-80 \mid T > 80) = 5$, εκφράζει τα πόσα χρόνια αναμένεται να ζήσει ένα άτομο που είναι ήδη 80 ετών, στην άνω περίπτωση αναμένεται να ζήσει πέντε ακόμη έτη.

Κεφάλαιο τρίτο

Ειδικά παραμετρικά μοντέλα διάρκειας ζωής

3.1 Μέθοδοι ανάλυσης δεδομένων διάρκειας ζωής

Με τον όρο «Μέθοδοι ανάλυσης δεδομένων διάρκειας ζωής», εννοούμε τις μεθόδους που μας οδηγούν στον προσδιορισμό της συγκεκριμένης μορφής της κατανομής που εκθέτει ένας συγκεκριμένος πληθυσμός δεδομένων διάρκειας ζωής. Οι μέθοδοι κατηγοριοποιούνται περαιτέρω σε παραμετρικές και μη παραμετρικές.

Οι παραμετρικές μέθοδοι χρησιμοποιούνται :

1. όταν υποθέτουμε ότι ο πληθυσμός του ενδιαφέροντός μας από τον οποίο προέρχεται το δείγμα τιμών διάρκειας ζωής που έχουμε στη διάθεσή μας, ακολουθεί μια συγκεκριμένη κατανομή, όπως για παράδειγμα, η εκθετική κατανομή, η κατανομή Weibull, η Λογαριθμική Κατανομή και η Γάμμα κατανομή.
2. Όταν ένα κατάλληλο γνωστό μοντέλο προσαρμόζεται στα δεδομένα.

Και στις δύο περιπτώσεις, όταν βρεθεί ότι ένα μοντέλο προσαρμόζεται κατάλληλα στα δεδομένα που έχουμε, τότε η κατανομή του πληθυσμού που μας ενδιαφέρει μπορεί να περιγραφεί από τις εκτιμηθείσες τιμές των παραμέτρων του μοντέλου.

Οι μη παραμετρικές μέθοδοι χρησιμοποιούνται όταν ο πληθυσμός του ενδιαφέροντός μας δεν ακολουθεί κανένα θεωρητικό μοντέλο ή δεν μπορούμε να υποθέσουμε ότι συμβαίνει κάτι τέτοιο διότι, για παράδειγμα, ο σχετικός έλεγχος είναι μια δαπανηρή ή χρονοβόρα διαδικασία. Τότε λέμε ότι αφήνουμε τα δεδομένα να δείξουν αυτά το πρότυπό τους.

3.2 Παραμετρικές Μέθοδοι

3.2.1 Η Εκθετική κατανομή

Η εκθετική κατανομή χρησιμοποιείται ευρέως και κατέχει έναν πολύ σημαντικό ρόλο στην ανάλυση επιβίωσης. Ένα ιδιαίτερο χαρακτηριστικό της είναι το γεγονός ότι είναι μια συνεχής κατανομή που χαρακτηρίζεται από έλλειψης μνήμης, δηλαδή η ηλικία των πειραματικών μονάδων δεν επηρεάζει την μελλοντική τους επιβίωση.

Επίσης, ένα ακόμη χαρακτηριστικό της εκθετικής κατανομής είναι ότι έχει μόνο θετική παράμετρο, που τη συμβολίζουμε με λ , και την οποία συνήθως ονομάζεται ρυθμός αποτυχίας.

Αν για οποιαδήποτε τυχαία μεταβλητή T , που ακολουθεί την εκθετική κατανομή με παράμετρο λ η συνάρτηση πυκνότητας της T είναι:

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{για } t \geq 0 \\ 0 & \text{για } t < 0 \end{cases}$$

Αν, όπου $T \geq 0$, η T συμβολίζει την τυχαία μεταβλητή της διάρκειας ζωής και t μια οποιοδήποτε τιμή της T και αν υποθέσουμε ότι η T ακολουθεί την εκθετική κατανομή με παράμετρο λ , τότε οι σχέσεις της κατανομής της διάρκειας ζωής δίνονται από τις παρακάτω συναρτήσεις.

$$F(t) = 1 - e^{-\lambda t}$$

$$f(t) = \lambda e^{-\lambda t}$$

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

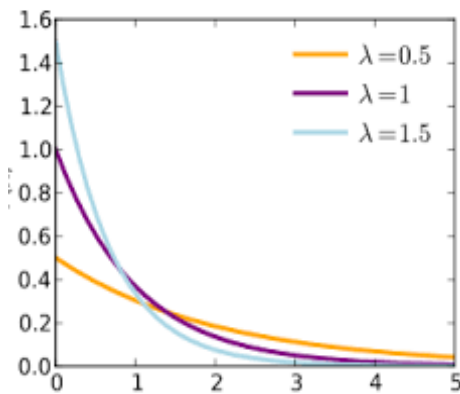
$$h(t) = \lambda$$

$$H(t) = \lambda t$$

$$L(t) = 1/\lambda$$

$$\text{Για όλα τα } t \geq 0.$$

Από τις παραπάνω σχέσεις προκύπτει ότι η συνάρτηση κινδύνου $h(t) = \lambda$, της εκθετικής κατανομής είναι σταθερή για οποιαδήποτε τιμή της παραμέτρου λ , που σημαίνει ότι αν η διάρκεια ζωής μιας πειραματικής μονάδας ακολουθεί την εκθετική κατανομή, ο κίνδυνος αποτυχίας παραμένει ίδιος καθ' όλη τη διάρκεια ζωής, με άλλα λόγια είναι ανεξάρτητος της ηλικίας της πειραματικής μονάδας. Παρακάτω παραθέτουμε διαγραμματική απεικόνιση της εκθετικής κατανομής για την $f(t)$.



3.2.2 Η κατανομή Weibull

Η κατανομή Weibull χρησιμοποιείται ευρέως στην ιατρική αλλά ιδιαίτερα στη μηχανική. Οι περιορισμοί της εκθετικής κατανομής που περιγράψαμε στην προηγούμενη παράγραφο, δηλαδή η ιδιότητα έλλειψη μνήμης και η σταθερή συνάρτηση κινδύνου, αίρονται με την Weibull κατανομή, η οποία υποδηλώνει ποικιλία δεδομένων συμπεριλαμβανομένου της φθοράς που επέρχεται λόγω του χρόνου, όπως συμβαίνει στην πραγματικότητα, είτε πρόκειται για μηχανήματα, είτε για ανθρώπινη ζωή. Η Weibull κατανομή είναι γενίκευση της εκθετικής κατανομής και είναι κατάλληλη για την περιγραφή μεταβλητών διάρκειας ζωής που έχουν σταθερή, αύξουσα ή φθίνουσα συνάρτηση κινδύνου. Αν υποθέσουμε ότι η διάρκεια ζωής T ακολουθεί την κατανομή Weibull με παραμέτρους έστω κ και λ , οι παραστάσεις της κατανομής δίνονται από τις σχέσεις:

$$F(t) = 1 - e^{-(\lambda t)^\kappa}$$

$$f(t) = \kappa \lambda^\kappa t^{\kappa-1} e^{-(\lambda t)^\kappa}$$

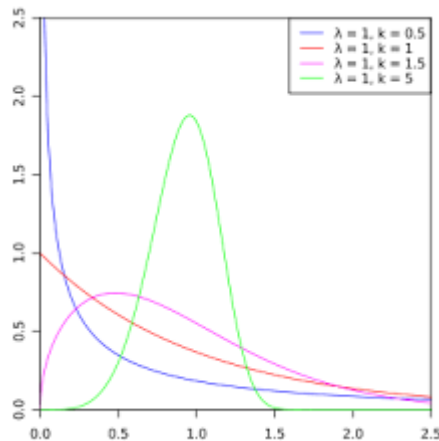
$$S(t) = e^{-(\lambda t)^\kappa}$$

$$h(t) = \kappa \lambda^\kappa t^{\kappa-1}$$

$$H(t) = (\lambda t)^\kappa$$

Για όλα τα $t \geq 0$.

Στη συνέχεια παραθέτουμε διαγραμματική απεικόνιση της κατανομής Weibull για την $f(t)$.



3.2.3 Η λογαριθμοκανονική κατανομή

Η λογαριθμοκανονική κατανομή υποδηλώνει ότι ο λογάριθμος του αθροίσματος n ανεξάρτητων τυχαίων μεταβλητών ισούται με το άθροισμα των λογαρίθμων των τυχαίων μεταβλητών. Δηλαδή, αν $Y_1, Y_2, \dots, Y_n$, οι n ανεξάρτητες τυχαίες μεταβλητές και $T_n = \sum_{i=1}^n Y_i$, τότε $\log T_n = \sum_{i=1}^n \log Y_i$.

Χαρακτηριστικό αυτής της κατανομής είναι ότι στην πιο απλή μορφή της μπορεί να οριστεί ως μια κατανομή μιας μεταβλητής της οποίας ο λογάριθμος ακολουθεί την κανονική κατανομή. Αν πούμε, ότι η τυχαία μεταβλητή της διάρκειας ζωής T είναι τέτοια ώστε ο $\log_e T$, ακολουθεί την κανονική κατανομή με μέση τιμή μ και διασπορά σ^2 , τότε λέμε ότι η διάρκεια ζωής T ακολουθεί τη λογαριθμοκανονική κατανομή και τη συμβολίζουμε ως εξής:

$T \sim \Lambda(\mu, \sigma^2)$, με τα μ, σ^2 να είναι αντιστοίχως παράμετροι κλίμακας και σχήματος της λογαριθμοκανονικής κατανομής.

Η συνάρτηση πυκνότητας της αυτής της κατανομής είναι λοξή προς τα δεξιά και όσο μεγαλύτερη είναι η τιμή του σ^2 τόσο μεγαλύτερη είναι η λοξότητα της. Αυτό συνεπάγεται, ότι η συνάρτηση κινδύνου αυξάνεται ως ένα μέγιστο αρχικά με το χρόνο και στη συνέχεια φθίνει, θεωρητικά μέχρι το μηδέν, γεγονός το οποίο μας

επιτρέπει ότι αυτό το μοντέλο είναι κατάλληλοι στις περιπτώσεις όπου το μοτίβο επιβίωσης έχει αρχικά αυξανόμενο ρυθμό αποτυχιών και στη συνέχεια ο ρυθμός αυτός μειώνεται.

3.2.4 Η γάμμα κατανομή

Η γάμμα κατανομή που χρησιμοποιείται αρκετά στο χώρο της ανάλυσης δεδομένων διάρκειας ζωής, αποτελεί γενίκευση της εκθετικής κατανομής ενώ η ίδια η κατανομή περιλαμβάνει και τη χ^2 κατανομή. Με άλλα λόγια, η γάμμα κατανομή προκύπτει ως η κατανομή που θεωρητικά σχετίζεται με τυχαίες μεταβλητές που ακολουθούν την κανονική κατανομή, ως η κατανομή αθροίσματος των τετραγώνων των ανεξάρτητων τυχαίων μεταβλητών που ακολουθούν την τυποποιημένη κανονική κατανομή.

3.3 Μη-Παραμετρικές Μέθοδοι

Οι μη-παραμετρικές μέθοδοι χρησιμοποιούνται για την εκτίμηση μιας κατανομής δεδομένων διάρκειας ζωής, όταν δεν υπάρχει παραμετρικό μοντέλο που να προσαρμόζεται στα δεδομένα της υπό μελέτης περίπτωσης. Επίσης, μερικές φορές έχουμε μικρό ή τέτοιο δείγμα δεδομένων, γεγονός που να εμποδίζει την επιβεβαίωση της ορθότητας των παραμετρικών μοντέλων.

Στην ανάλυση επιβίωσης χρησιμοποιούνται δυο βασικές μη-παραμετρικές μέθοδοι ανάλυσης δεδομένων διάρκεια ζωής. Η μέθοδος Πινάκων Ζωής και η μέθοδος Kaplan-Meier (1958), ή μέθοδος ορίου γινομένων. Ουσιαστικά, θα προσπαθήσουμε να εκτιμήσουμε τη μία εκ των βασικών συναρτήσεων που περιγράφει την κατανομή διάρκειας ζωής που είναι η συνάρτηση επιβίωσης.

3.3.1 Η μέθοδος Kaplan-Meier

Ας υποθέσουμε ότι θέλουμε να εκτιμήσουμε την πιθανότητα επιβίωσης κάποιων ασθενών για διάστημα ενός έτους οι οποίοι σε κάποιο χρονικό σημείο διαγνώστηκαν να έχουν μια ασθένεια. Αρχικά με την υπόθεση ότι έχουμε πλήρη δεδομένα η μέθοδος Kaplan-Meier βασίζεται στην παραδοχή πως για να επιβιώσουν οι ασθενείς ένα χρόνο μετά από την διάγνωση, θα πρέπει να επιβιώνουν καθημερινά από την πρώτη μέχρι και την τελευταία ημέρα του υπό μελέτη χρονικού διαστήματος (στην προκειμένη περίπτωση του ενός έτους). Είναι λοιπόν εύλογο για τον υπολογισμό της πιθανότητας $\Pr(T > t)$, θα πρέπει να υπολογισθεί για συγκεκριμένη ημέρα η αναλογία των ασθενών, ανάμεσα σε αυτούς που εισέρχονται στη μελέτη και σε αυτούς που

καταφέρνουν να επιβιώσουν μέχρι την επόμενη ημέρα.

Αρχικά λοιπόν η πιθανότητα υπολογίζεται από το παρακάτω κλάσμα

$p(1) = \text{Αριθμός ασθενών που εισέρχονται στη μελέτη} \text{ μείον αριθμός θανάτων την πρώτη μέρα} / \text{Αριθμός ασθενών που εισέρχονται στη μελέτη}$

Στη συνέχεια όμως ο αριθμός των εν ζωή ασθενών οι οποίοι και απεικονίζονται στον παρονομαστή είναι πολύ πιθανό να είναι αρκετά μικρότερος από τον αρχικό. Για παράδειγμα την 300η ημέρα από την αρχή της έρευνας κανείς από τους ασθενείς που είχαν αποτυχία (θάνατος) ή πιθανόν έχουν αποκοπεί σε κάποια από τις προηγούμενες ημέρες δεν μεταβιβάζει κάποια πληροφορία στην πιθανότητα επιβίωσης της 300ης ημέρας. Μόνο ασθενείς που είναι ζωντανοί και υπό παρακολούθηση ακριβώς πριν την 300η ημέρα συμπεριλαμβανομένων και αυτών των οποίων η ο χρόνος αποτυχίας είναι η 300η ημέρα, αυτή η ομάδα ασθενών ονομάζεται ασθενείς σε κίνδυνο και η πιθανότητα $P(300) = \text{Αριθμός ασθενών σε κίνδυνο την 300η ημέρα} \text{ μείον αριθμός θανάτων την 300η ημέρα} / \text{Αριθμός ασθενών σε κίνδυνο την 300η ημέρα}$. Αυτό το οποίο μας αφορά είναι πως θα συνδυάσουμε τις παραπάνω πιθανότητες έτσι ώστε να φθάσουμε στην $P(365)$. Ο υπολογισμός της παραπάνω πιθανότητας γίνεται βήμα-βήμα υπολογίζοντας αρχικά τις ημερήσιες πιθανότητες επιβίωσης οι οποίες προκύπτουν από κάποια για παράδειγμα $P(2) = p(1) * p(2)$. Στη συνέχεια υπολογίζεται η πιθανότητα επιβίωσης των τριών ημερών $P(3) = P(2) * p(3)$ και ούτω καθεξής μέχρι και την $P(365)$.

Επομένως, μια εκτίμηση για την συνάρτηση επιβίωσης όταν έχουμε πλήρη δεδομένα είναι $\hat{S}(t_{(i)}) = (n-i)/n$.

Στην πράξη η συνάρτηση επιβίωσης εκτιμάται σε κάθε διακριτό σημείο της μεταβλητής διάρκειας ζωής. Η σχέση $\hat{S}(t)$ είναι μια κλιμακωτή συνάρτηση που ξεκινά από το σημείο μηδέν και φθίνει με βήμα $1/n$.

Όταν υπάρχουν αποκομμένα δεδομένα, τότε η εκτίμηση κατά Kaplan-Meier της συνάρτησης επιβίωσης με βάση τη συνάρτηση κινδύνου έχει ως εξής:

$$S(t_i) = \prod_{j \in R(t_i)} [1 - h(t_j)] \quad i=1,2,\dots \text{ και } t_i \geq 0$$

Όπου $R(t_i)$ είναι το συμπλήρωμα του συνόλου κινδύνου κατά τη χρονική στιγμή t_i .

Για καλύτερη κατανόηση, ας θεωρήσουμε ότι έχουμε να μελετήσουμε το εξής ερώτημα:

«Ποια η πιθανότητα οι ασθενείς να επιβιώσουν για μία συγκεκριμένη διάρκεια χρόνου».

Χρησιμοποιώντας τη μέθοδο των Kaplan – Meier, έχουμε:

p_1 : η πιθανότητα επιβίωσης για έναν τουλάχιστο χρόνο.

p_2 : η πιθανότητα επιβίωσης το δεύτερο χρόνο, δεδομένου ότι οι ασθενείς επιβίωσαν τον πρώτο χρόνο.

...

...

p_j : η πιθανότητα επιβίωσης τον χρόνο j , δεδομένου ότι επιβίωσαν τα προηγούμενα $j-1$ χρόνια.

Υποθέτουμε ότι είναι γνωστοί οι πλήρεις και οι αποκομμένοι χρόνοι n ατόμων, σε ένα τυχαίο δείγμα $\{t_1, t_2, \dots, t_n\}$. Έστω $t_{(1)} < t_{(2)} < \dots < t_{(k)}$, $k \in \mathbb{N}$, οι διακεκριμένοι, ταξινομημένοι χρόνοι αποτυχίας (δηλαδή οι χρόνοι στους οποίους συμβαίνουν θάνατοι (αποτυχίες) και έτσι, αντιστοιχούν σε πλήρεις χρόνους ζωής. Όταν σε ένα χρόνο $t_{(j)}$, συμβαίνουν πολλές αποτυχίες, τότε οι παρατηρήσεις αυτές ονομάζονται ισότιμες (tied) παρατηρήσεις ή δεσμοί.

Για $j = 1, 2, \dots, k$ έστω:

n_j : ο αριθμός των ατόμων που είναι σε κίνδυνο στο χρόνο $t_{(j)}$, δηλαδή είναι ο αριθμός των ζωντανών ατόμων στην αρχή του διαστήματος $[t_{(j)}, t_{(j+1)})$ ή διαφορετικά, είναι ο αριθμός των ατόμων που έχουν χρόνους αποτυχίας ή αποκοπής μεγαλύτερους ή ίσους με $t_{(j)}$.

d_j : ο αριθμός των θανάτων στο $t_{(j)}$ και

c_j : ο αριθμός των αποκομμένων τιμών (αντιστοιχούν σε αποκομμένους χρόνους ζωής) στο διάστημα $[t_{(j)}, t_{(j+1)})$.

Έτσι, η ποσότητα $n_j - d_j$ συμβολίζει τον αριθμό των ασθενών που επιβιώνουν στο χρόνο $t_{(j)}$ και η ποσότητα $n_j - d_j - c_j$, συμβολίζει τον αριθμό των ασθενών που είναι σε κίνδυνο στο χρόνο $t_{(j+1)}$:

$$n_{(j+1)} = n_j - d_j - c_j$$

Βάση των πιο πάνω και του ορισμού της p_j έχουμε ότι:

$$p_j = (n_j - d_j) / n_j = 1 - d_j / n_j$$

όπου $p_j = (p(t_{(j)})) = P(T > t_{(j)})$.

Σύμφωνα με το θεμελιώδη τύπο των πιθανοτήτων έχουμε

$$P(A \cap B) = P(A)P(B|A)$$

με $\{T > t_{(j-1)}\}$ με ενδεχόμενο A και ενδεχόμενο B το $\{T > t_{(j)}\}$ υπολογίζουμε την συνάρτηση επιβίωσης $S(t)$ ως

$$S(t_{(j)}) = P(T > t_{(j)}) = P(T > t_{(1)})P(T > t_{(2)}|T > t_{(1)}) \dots P(T > t_{(j)}|T > t_{(j-1)})$$

Όμως αφού

$$P(T > t_{(j)}) = (n_j - d_j) / n_j \text{ και } P(T > t_{(j)}|T > t_{(j-1)})$$

τότε η εκτιμήτρια Kaplan - Meier είναι:

$$\hat{S}(t) = (n_1 - d_1) / n_1 * (n_2 - d_2) / n_2 * \dots * (n_i - d_i) / n_i, i: t_{(i)} \leq t \leq t_{(i+1)}$$

$$\prod (n_i - d_i) / n_i, \text{ όταν } t \geq t_{(1)} \text{ και } 1 \text{ όταν } t < t_{(1)}$$

Διαφορετικά, η εκτιμήτρια $\hat{S}(t)$ μπορεί να υπολογιστεί βάση της αναδρομικής σχέσης:

$$\hat{S}(t_{(j)}) = \hat{S}(t_{(j-1)})p_j$$

Θεωρούμε ότι $\hat{S}(0) = 1$, δηλαδή ότι όλοι οι ασθενείς είναι ζωντανοί στο χρόνο 0.

Παρατηρούμε ότι όταν δεν πεθαίνει κανένα άτομο, δηλαδή $d_j = 0$ ή $p_j = 1$. Έτσι, η εκτιμήτρια Kaplan – Meier της πιθανότητας επιβίωσης στο χρόνο t , αλλάζει μόνο σε χρόνους στους οποίους πεθαίνει τουλάχιστον ένα άτομο. Ως συνέπεια, μπορούμε να παραλείψουμε τους χρόνους στους οποίους δεν παρατηρούνται θάνατοι, στον υπολογισμό της $\hat{S}(t)$.

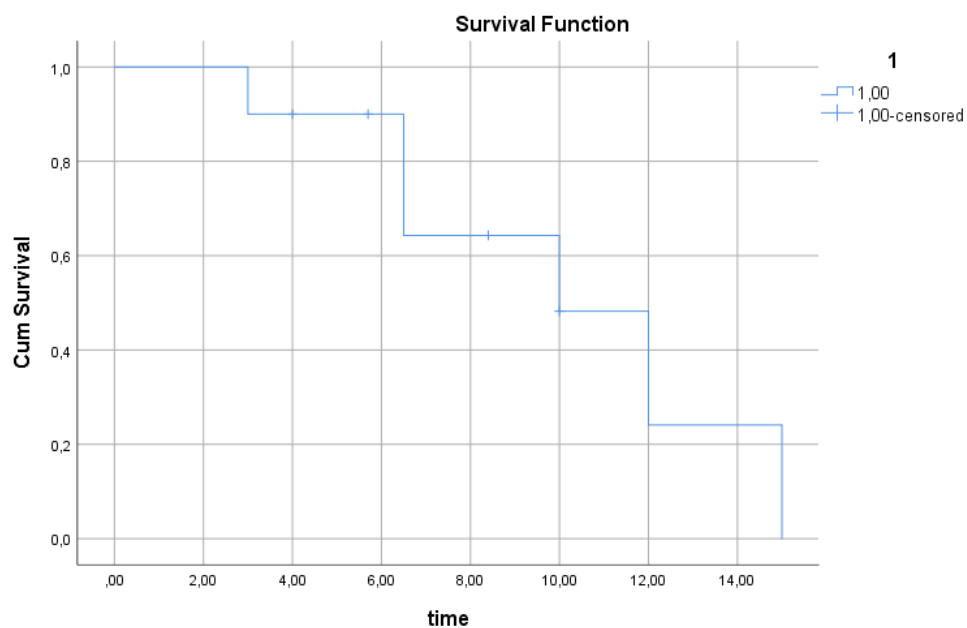
Στην πράξη, ο εκτιμητής γινομένου-ορίου, μπορεί να εκτιμηθεί κατασκευάζοντας έναν πίνακα, ο οποίος θα έχει 5 στήλες.

1. Η πρώτη στήλη θα περιέχει τους διακεκριμένους πλήρεις χρόνους επιβίωσης, αποκομμένους και μη, σε αύξουσα σειρά.
2. Η δεύτερη στήλη θα έχει τον αριθμό των ατόμων που βρίσκονται σε κίνδυνο (n_j).
3. Η τρίτη στήλη θα έχει τον αριθμό των θανάτων (d_j) στους διάφορους χρόνους επιβίωσης (αποτυχίας).
4. Στην τέταρτη στήλη θα είναι οι τιμές των p_j .
5. Η τελευταία στήλη θα είναι οι εκτιμητές Kaplan - Meier

Στα παρακάτω παραδείγματα θα δούμε εφαρμογές της μεθόδου Kaplan –Meier

Παράδειγμα 3.1

Έστω ότι παρατηρούνται οι παρακάτω χρόνοι (διάρκεια) ύφεσης δέκα ασθενών με κακοήθεις όγκους. Έξι (6) ασθενείς υποτροπιάζουν στους 3.0, 6.5, 6.5, 10, 12 και 15 μήνες αντίστοιχα. Επιπλέον ένας (1) ασθενής παύει να παρακολουθείται στους 8.4 μήνες, ενώ τρεις (3) εξακολουθούν να παρακολουθούνται στο τέλος της μελέτης, μετά από 4.0, 5.7 και 10 μήνες. Στο διάγραμμα που ακολουθεί απεικονίζεται η $S(t)$.



Ακολουθεί ο πίνακας υπολογισμών

time	rank	r	$(n-r)/(n-r-1)$	$\hat{S}(t)$
3.0	1	1	9/10	0.900
4.0+	2	-	-	-
5.7+	3	-	-	-
6.5	4	4	6/7	$9/10 * 6/7 = 0.771$
6.5	5	5	5/6	$9/10 * 6/7 * 5/6 = 0.643$
8.4+	6	-	-	-
10.0	7	7	$3/4$	$9/10 * 6/7 * 5/6 * 3/4 = 0.482$
10.0+	8	-	-	-
12.0	9	9	$1/2$	$9/10 * 6/7 * 5/6 * 3/4 * 1/2 = 0.241$
15.0	10	10	0	0

E. Lee pg.71

3.3.2 Η Καμπύλη επιβίωσης

Από τη γραφική παράσταση της $\hat{S}(t)$ συναρτήσεως του t , παίρνουμε τον εκτιμητή Kaplan – Meier της καμπύλης επιβίωσης και επομένως έχουμε μια καλή περιγραφή των δεδομένων. Η $\hat{S}(t)$ είναι φθίνουσα συνάρτηση, συνεχής από αριστερά της

οποίας η τιμή δεν αλλάζει, παρά μόνο στα σημεία όπου παρατηρούνται θάνατοι ή γενικότερα αποτυχίες. Έτσι σε κάθε επόμενο βήμα η τιμή της μειώνεται κατά

$$\frac{n_j - d_j}{n_j}$$

αμέσως μετά το j πλήρη χρόνο, $t_{(j)}$

Επειδή η αποκοπή του ασθενή μειώνει τον αριθμό των ασθενών που συνεισφέρουν στην καμπύλη, (μειώνεται το n_j), κάθε θάνατος μετά από αυτό το σημείο παριστάνει μια μεγαλύτερη αναλογία του υπόλοιπου πληθυσμού από την αναλογία που θα είχαμε αν γνωρίζαμε τους πλήρεις χρόνους. Έτσι, η αποκοπή επηρεάζει την καμπύλη επιβίωσης.

Όταν η μεγαλύτερη παρατήρηση είναι μη-αποκομμένη, τότε ο Kaplan-Meier εκτιμητής στο σημείο αυτό είναι 0, επειδή $n_k = d_k$ και $c_k = 0$. Άρα θα έχουμε ότι $p_k = 0$ και έτσι, $\hat{S}(k) = 0$. Στην καμπύλη επιβίωσης τότε, θα έχουμε μια κάθετη γραμμή στο t_k , από τον προτελευταίο πλήρη χρόνο που θα κατεβαίνει κάθετα στο t_k . Δηλαδή φαίνεται ότι $S(t) = 0$, για κάθε $t > t_k$. Αυτό το αποτέλεσμα βασίζεται μόνο σε ένα ασθενή και είναι λάθος να συμπεράνουμε ότι η πιθανότητα ένας ασθενής (με την ίδια ασθένεια) δεν θα επιβιώσει περισσότερο από το χρόνο t_k . Αντίθετα, όταν η μεγαλύτερη παρατήρηση είναι αποκομμένη, τότε η καμπύλη επιβίωσης δεν είναι 0 μετά το $t_{\max}$, αφού στην περίπτωση αυτή, $n_{\max} \neq d_{\max}$ και συνεπώς $\hat{S}(\max) \neq 0$. Έτσι η καμπύλη μετά το $t_{\max}$, θα συνεχίζει παράλληλα με τον άξονα των t χωρίς να κατεβαίνει προς τα κάτω.

3.3.3 Υπολογισμός της εκτιμήτριας Kaplan – Meier

Για παρακάτω παράδειγμα ας υπολογίσουμε τη $\hat{S}(t)$ για 11 ασθενείς.

Έστω οι παρακάτω χρόνοι επιβίωσης ασθενών μετά από συγκεκριμένη θεραπεία

33,7+	3,9	10,5	5,4	19,5	23,8+	7,9	16,9+	16,6+	33,7+	17,1+
-------	-----	------	-----	------	-------	-----	-------	-------	-------	-------

Πηγή: Elisa T. Lee pg.20

Πίνακας Υπολογισμών

Death time (t)	Number of patients alive at beginning of interval t_i	Number of patients dying during interval t_i	Proportion surviving (p_j)	Cumulative proportion surviving from beginning of study to end of interval $\hat{S}(t)$
3.9	11	1	0.909	0.909
5.4	10	1	0.9	0.818
7.9	9	1	0.889	0.727
10.5	8	1	0.875	0.636
19.5	4	1	0.75	0.477

Η $\hat{S}(t)$ εκτιμάται μόνο για τους χρόνους που παρατηρούνται θάνατοι (πλήρεις χρόνοι).

Στο χρόνο 0 υπάρχουν $n_0 = 11$ ασθενείς που βρίσκονται σε κίνδυνο.

Στο χρόνο 1 υπάρχουν επίσης $n_1 = 11$. Στο διάστημα $[t_{(1)}, t_{(2)}) = [3,9, 5,4)$ έχουμε ένα θάνατο στους 3,9 μήνες, δηλαδή $d_1=1$. Άρα, $p_1 = 1 - \frac{1}{11} = 0,909$ και $\hat{S}(t) = 0,909$ για $t \in [t_{(1)}, t_{(2)})$.

Στην αρχή του επόμενου διαστήματος θα έχουμε $n_2 = 11 - 1 = 10$ ασθενείς που βρίσκονται σε κίνδυνο στο διάστημα $[t_{(2)}, t_{(3)})$ και $d_2=1$ εφόσον έχουμε έναν θάνατο στους 5,4 μήνες, ενώ δεν έχουμε αποκομμένες παρατηρήσεις στο πρώτο διάστημα.

$$\text{Άρα } p_2 = 1 - \frac{1}{10} = 0,9$$

και $\hat{S}(t) = 0,909 * 0,9 = 0,818$ για $t \in [t_{(2)}, t_{(3)})$.

Συνεχίζουμε με τον ίδιο τρόπο για να υπολογίσουμε τις τιμές των υπολοίπων p_j και $\hat{S}(t)$.

$$p_3 = 1 - \frac{1}{9} = 0,889 \text{ και } \hat{S}(t) = 0,818 * 0,889 = 0,727 \text{ για } t \in [t_{(3)}, t_{(4)}).$$

Στο διάστημα $[t_{(4)}, t_{(5)}) = [10,5, 19,5)$ έχουμε $n_4 = 8$ άτομα σε κίνδυνο και έχουμε ένα θάνατο στους 10,5 μήνες. Επιπλέον έχουμε και 3 άτομα τα οποία χάθηκαν κατά την περίοδο της έρευνας, δηλαδή 3 αποκομμένες παρατηρήσεις, δηλαδή $c_4=3$.

Άρα: $p_4 = 1 - \frac{1}{8} = 0,875$ και $\hat{S}(t) = 0,727 * 0,875 = 0,636$ για $t \in [t_{(4)}, t_{(5)})$.

Στο τελευταίο διάστημα, έχουμε $n_5 = 8 - 1 - 3 = 4$ ασθενείς σε κίνδυνο εφόσον σε αυτό το διάστημα έχουμε ένα θάνατο και έχουν χαθεί από την μελέτη τρία άτομα. Άρα, $p_5 = 1 - \frac{1}{4} = 0,75$ και $\hat{S}(t) = 0,636 * 0,75 = 0,477$ για $t_{(5)}$.

3.3.4 Διαστήματα Εμπιστοσύνης της Εκτιμήτριας Kaplan – Meier

Με την παραπάνω μέθοδο προκύπτει η δειγματική εκτίμηση της $\hat{S}(t)$, απαιτείται όμως και η εκτίμηση της ακρίβειας αυτής ώστε να κατασκευαστούν διαστήματα εμπιστοσύνης.

Ένα διάστημα εμπιστοσύνης μιας παραμέτρου έστω η αντιστοιχεί σε ένα διάστημα τιμών της παραμέτρου που προσδιορίζεται στη βάση ενός δείγματος, μεγέθους n , τιμών $t_1, t_2, \dots, t_n$ της τυχαίας μεταβλητής T , όπου τα όρια του διαστήματος είναι συναρτήσεις του n , των $t_1, t_2, \dots, t_n$, καθώς και του συντελεστή του διαστήματος $(1 - \alpha)$, η τιμή του οποίου καθορίζεται από τον ερευνητή.

Ένα $(1 - \alpha)100\%$ διάστημα εμπιστοσύνης στο χρόνο t για την εκτιμήτρια $\hat{S}(t)$, θεωρώντας ότι οι εκτιμητές Kaplan – Meier ακολουθούν κανονική κατανομή, είναι :

$$\hat{S}(t) \pm z_{\alpha/2} \text{ se}[\hat{S}(t)]$$

3.4 Πίνακες Επιβίωσης (Life Tables)

Οι πίνακες επιβίωσης (life tables) είναι επίσης μία μη παραμετρική μέθοδος, πολύ διαδεδομένη, για την περιγραφή δεδομένων που αφορούν χρόνους επιβίωσης. Η μέθοδος αυτή χρησιμοποιήθηκε αρχικά ως τεχνική για τη μέτρηση της θνησιμότητας και περιγράφει την κλίμακα επιβίωσης ενός πληθυσμού. Χρησιμοποιείτε από κυβερνητικές υπηρεσίες και από ιατρικούς ερευνητές σε μελέτες της επιβίωσης όπως αύξηση του πληθυσμού, γονιμότητα, μετανάστευση κι ούτω καθεξής.

Οι πίνακες επιβίωσης είναι μια μορφή των συνηθισμένων πινάκων συχνοτήτων (frequency tables) στην περίπτωση που υπάρχουν αποκομμένα δεδομένα.

3.4.1 Πίνακες Επιβίωσης Πληθυσμού (Population Life Tables)

Για την αποτύπωση των δεδομένων επιβίωσης ζωής ενός πληθυσμού, χρησιμοποιούνται κυρίως ο πίνακας της ζωής της ομάδας και ο τρέχον πινάκων επιβίωσης.

Ο πίνακας επιβίωσης της ομάδας (ή κοόρτης – cohort) περιγράφει την επιβίωση ή την εξέλιξη της θνησιμότητας από τη γέννηση έως το θάνατο μιας συγκεκριμένης ομάδας ατόμων που είχαν γεννηθεί περίπου στο ίδιο χρονικό διάστημα. Για παράδειγμα, ο πίνακας επιβίωσης όλων των ατόμων που έχουν γεννηθεί το 1950. Η ομάδα πρέπει να ακολουθείται από το 1950 έως ότου όλα αυτά τα άτομα πεθάνουν. Το ποσοστό θανάτου (ή επιβίωσης) χρησιμοποιείται στη συνέχεια για να κατασκευαστούν πίνακες επιβίωσης για διαδοχικά ημερολογιακά έτη. Αυτός ο τύπος πίνακα, είναι χρήσιμος στην προβολή της επιβίωσης του πληθυσμού αλλά δεν κατασκευάζεται συχνά γιατί απαιτεί μακρά περίοδο παρακολούθησης.

Ο τρέχων πίνακας επιβίωσης κατασκευάζεται με την εφαρμογή ενός λόγου θνησιμότητας του πληθυσμού μιας συγκεκριμένης ηλικίας, σε μια δεδομένη χρονική περίοδο σε μια υποθετική ομάδα των 100.000 ατόμων. Το σημείο εκκίνησης θεωρείται η γέννηση κατά το έτος 0.

Δύο πηγές των δεδομένων απαιτούνται για την κατασκευή ενός πίνακα επιβίωσης του πληθυσμού:

1. Δεδομένα απογραφής όσον αφορά τον αριθμό των ατόμων κάθε ηλικίας που ζουν για ένα δεδομένο έτος, στα τέλη του πρώτου εξαμήνου και
2. Δημογραφικές στατιστικές για τον αριθμό των θανάτων κατά το συγκεκριμένο έτος για κάθε ηλικία.

3.4.2 Κλινικοί Πίνακες Επιβίωσης – Η Αναλογιστική μέθοδος (Clinical Life Tables – The Actuarial method)

Η αναλογιστική μέθοδος των πινάκων επιβίωσης έχει εφαρμοστεί σε κλινικά δεδομένα για πολλές δεκαετίες. Οι Beale et al. (1976) και οι Cutler και Ederer (1958) δίνουν μια μέθοδο πινάκων επιβίωσης για την εκτίμηση της συνάρτησης επιβίωσης, ενώ ο Gehan (1969) παρείχε μεθόδους για την εκτίμηση και των τριών συναρτήσεων – επιβίωσης, κινδύνου και πυκνότητας πιθανότητας.

Η χρήση της αναλογιστικής μεθόδου (actuarial method) δεν είναι δυνατή όταν ο αριθμός των συμμετεχόντων σε μια μελέτη είναι μικρός, όταν τα επιμέρους χρονικά διαστήματα στα οποία χωρίζεται ο συνολικός χρόνος διεξαγωγής της μελέτης είναι σχετικά μεγάλα, όταν οι απώλειες είναι πολλές και όταν δεν είναι λογικό να θεωρηθεί ότι οι απώλειες κατανέμονται ομοιόμορφα σε ένα δεδομένο χρονικό διάστημα. Στις περιπτώσεις αυτές είναι απαραίτητη η εφαρμογή της μεθόδου Kaplan-Meier που περιγράψαμε στην προηγούμενη ενότητα.

Η μέθοδος Kaplan-Meier είναι παρόμοια με την αναλογιστική, αλλά διαφέρει στο ότι δεν απαιτείται να χωριστεί ο συνολικός χρόνος σε ενδιάμεσα χρονικά διαστήματα. Για το λόγο αυτόν, η μέθοδος Kaplan-Meier είναι κατάλληλη, κυρίως, για μελέτες με μικρό αριθμό μονάδων. Η μέθοδος Kaplan-Meier απαιτεί λιγότερο πολύπλοκους υπολογισμούς από την αναλογιστική, κυρίως επειδή η εκτίμηση της επιβίωσης πραγματοποιείται κάθε φορά που υπό μελέτη μονάδα εμφανίζει την έκβαση και όχι σε συγκεκριμένα χρονικά διαστήματα, όπως συμβαίνει στην αναλογιστική. Η μέθοδος Kaplan-Meier, σε αντίθεση με την αναλογιστική, δεν λαμβάνει υπόψη της κατά τους υπολογισμούς τις απώλειες της μελέτης. Επίσης, η αναλογιστική μέθοδος ενσωματώνει όλες τις πληροφορίες επιβίωσης μέχρι το πέρας της μελέτης, δηλαδή χρησιμοποιεί ελλιπή στοιχεία όπως είναι οι απώλειες κατά την παρακολούθηση διαφορετικό της μελέτης, πρόσωπα που αποσύρθηκαν από τη μελέτη χωρίς λόγο καθώς και πλήρη στοιχεία του θανάτου.

Έστω ένα τυχαίο δείγμα n ατόμων από ένα συγκεκριμένο πληθυσμό το οποίο περιέχει πλήρη και αποκομμένα δεδομένα. Τα δεδομένα αυτά χωρίζονται σε m διαστήματα της μορφής $I_j = [t_j, t_{j+1})$. Το t_0 ορίζεται να είναι 0.

Για κάθε διάστημα I_j , έχουμε τις ακόλουθες πληροφορίες:

n_j : ο αριθμός των ατόμων που είναι σε κίνδυνο στην αρχή του διαστήματος I_j , δηλαδή είναι ο αριθμός των ζωντανών ατόμων στην αρχή του διαστήματος.

d_j : ο αριθμός των θανάτων στο διάστημα c_j .

c_j : ο αριθμός των αποκομμένων τιμών (αντιστοιχούν σε αποκομμένους χρόνους ζωής) στο διάστημα I_j .

Προκύπτει ότι $n_0 = n$, και $n_{j+1} = n_j - c_j - d_j$, $j = 0, 1, \dots, m-1$. Όταν δεν υπάρχουν αποκομμένες παρατηρήσεις, όταν δηλαδή $c_j = 0$ για $j = 0, 1, \dots, m-1$, τότε η εκτίμηση του

$q_j = P(\text{θάνατος στο } I_j | \text{μονάδες σε κίνδυνο στην αρχή του διαστήματος})$
είναι d_j/n_j .

Για να ορίσουμε τον αριθμό των μονάδων που είναι σε κίνδυνο στο διάστημα I_j , θεωρούμε ότι, κατά μέσο όρο, η αποκοπή συμβαίνει στο μέσο του διαστήματος I_j . Έτσι, ο μέσος αριθμός ατόμων που είναι σε κίνδυνο στο διάστημα I_j είναι:

$$n'_j = n_j - \frac{c_j}{2}$$

Ο αναλογικός εκτιμητής (actuarial estimator) του q_j είναι:

$$\hat{q}_j = \frac{d_j}{n_j - \frac{c_j}{2}} = \frac{d_j}{n'_j}$$

Έτσι,

$$\hat{p}_j = 1 - \frac{d_j}{n'_j} = \frac{n'_j - d_j}{n'_j}$$

Ορίζουμε το αθροιστικό ποσοστό επιβίωσης $[\hat{S}(t_i)]$ σαν μια εκτίμηση της συνάρτησης επιβίωσης για τον χρόνο t_i . Συχνά αναφέρεται και σαν αθροιστικός ρυθμός επιβίωσης. Για $i = 1$, $\hat{S}(t_1) = 1$ ενώ για $i = 2, \dots, n$ $\hat{S}(t_i) = \hat{p}_{i-1} \hat{S}(t_{i-1})$.

Είναι η συνήθης εκτίμηση πίνακα επιβίωσης και βασίζεται στο γεγονός ότι η επιβίωση στην αρχή του i διαστήματος, σημαίνει και επιβίωση στην αρχή και καθ' όλο το $(i-1)$ διάστημα.

Η εκτιμώμενη συνάρτηση πυκνότητας πιθανότητας $[\hat{f}(t_m)]$ ορίζεται σαν τη πιθανότητα θανάτου στο i διάστημα, ανά μονάδα πλάτους του διαστήματος. Δηλαδή,

$$\hat{f}(t_m) = \frac{\hat{S}(t_i) - \hat{S}(t_{i-1})}{t_i - t_{i-1}}$$

Επίσης, η συνάρτηση κινδύνου $\hat{h}(t_{mi})$ ορίζεται σαν το πηλίκο του αριθμού των

θανάτων ανά μονάδα χρόνου σε ένα διάστημα με το μέσο αριθμό των επιζώντων στο μέσο του διαστήματος, δηλαδή:

$$\hat{h}(t_{mi}) = \frac{\hat{f}(t_{mi})}{\hat{s}(t_{mi})}$$

Κεφάλαιο τέταρτο

Μη – παραμετρικές μέθοδοι για τη σύγκριση καμπυλών επιβίωσης

Στην ανάλυση δεδομένων επιβίωσης, εκτός από τη συνάρτηση επιβίωσης, με την οποία μπορούμε να έχουμε εικόνα για τα δεδομένα διάρκειας ζωής, ενδιαφερόμαστε και για τη σύγκριση του χρόνου επιβίωσης δύο ή περισσότερων υπό μελέτη ομάδων. Συνήθως οι χρόνοι επιβίωσης διαφορετικών ομάδων διαφέρουν και λόγω του γεγονότος ότι δεν κατανέμονται κανονικά πρέπει να εφαρμοστούν επίσης μη-παραμετρικοί έλεγχοι, μερικούς από τους οποίους περιγράφουμε παρακάτω.

4.1 To Logrank Test

Έστω $t_{(1)} < t_{(2)} < \dots < t_{(k)}$ διακεκριμένες χρονικές στιγμές κατά τις οποίες παύουν να λειτουργούν μονάδες που προέρχονται από δύο ομάδες. Αμέσως πριν από τη χρονική στιγμή $t_{(j)}$, θεωρούμε ότι στην ομάδα i ($=1,2$) υπάρχουν n_{ij} μονάδες «σε κίνδυνο», εκ των οποίων d_{ij} μονάδες παύουν να λειτουργούν τη στιγμή $t_{(j)}$. Ορίζουμε:

$$n_j = n_{1j} + n_{2j}$$

και

$$d_j = d_{1j} + d_{2j}.$$

Τα γεγονότα της χρονικής στιγμής $t_{(j)}$ περιγράφονται περιληπτικά από τον παρακάτω πίνακα συνάφειας:

Διακοπή λειτουργίας		A	B	Σ
	Ναι	d_{1j}	d_{2j}	d_j
	Όχι	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
	Σ	n_{1j}	n_{2j}	n_j

Πηγή: (Χ. Καρώνη, 2009)

Αναλύοντας τον παραπάνω πίνακα συνάφειας με το γνωστό χ^2 έλεγχο, υπολογίζουμε τις αναμενόμενες συχνότητες υπό την υπόθεση ανεξαρτησίας του γεγονότος «διακοπή λειτουργίας των μονάδων» από την «ομάδα στην οποία ανήκει». Δηλαδή, οι πιθανότητες να διακοπή η λειτουργία των μονάδων είναι ίδιες και στις δύο ομάδες και συνεπώς και οι συναρτήσεις επιβίωσης, το οποίο είναι και το ζητούμενο. Η αναμενόμενη συχνότητα του κελιού $[d_{1j}]$ που αφορά τις μονάδες της ομάδας A των οποίων η λειτουργία διακόπηκε είναι η παρακάτω:

$$E(d_{1j}) = n_{1j}d_j/n_j = \hat{d}_{1j}$$

και η απόκλιση από την παρατηρούμενη d_{1j} είναι

$$u_j = d_{1j} - (n_{1j}d_j/n_j)$$

Ο έλεγχος της υπόθεσης της ανεξαρτησίας προκύπτει διαιρώντας την ποσότητα u_j με μια εκτιμήτρια του τυπικού σφάλματος ή διαιρώντας το τετράγωνο της ποσότητας αυτής με τη διασπορά της d_{1j}

$$v_j = V(d_{1j}) = n_{1j}n_{2j}d_j(n_j - d_j) / n_j^2(n_j - 1),$$

δηλαδή

$$\frac{\{d_{1j} - (n_{1j}d_j / n_j)\}^2}{n_{1j}n_{2j}d_j(n_j - d_j) / n_j^2(n_j - 1)} = \frac{u_j^2}{v_j}$$

Η τελική μορφή της ελεγχοσυνάρτησης του ελέγχου log-rank προσδιορίζεται αθροίζοντας ως προς όλες τις χρονικές στιγμές (t_j) δηλαδή,

$$u = \sum_j u_j = \sum_j \{d_{1j} - (n_{1j}d_j / n_j)\}$$

και θεωρώντας ότι οι πίνακες συνάφειας για κάθε $t_{(j)}$, $j = 1, \dots, k$, είναι ανεξάρτητοι, τότε οι διασπορά u του αθροίσματος u είναι

$$u = \sum_j u_j = \sum_j n_{1j}n_{2j}d_j(n_j - d_j) / n_j^2(n_j - 1)$$

Υπό την υπόθεση $H_0 : S_1(t) = S_2(t)$, η $u / \sqrt{v}$ ακολουθεί την $N(0,1)$ ασυμπτωτικά όταν ο αριθμός των διακοπών δεν είναι πολύ μικρό, κατά συνέπεια και η ελεγχοσυνάρτηση log-rank u^2 / v ακολουθεί την χ_1^2 ασυμπτωτικά. Πρόκειται για ένα μη-παραμετρικό έλεγχο διότι η μηδενική υπόθεση αφορά την ισότητα δύο συναρτήσεων επιβίωσης, χωρίς να προσδιορίζονται οι συναρτήσεις αυτές.

Μια μεγάλη τιμή του $C_{logrank}^2$ θα οδηγήσει στην απόρριψη της H_0 . Γενικά, έστω k ο αριθμός των υπό μελέτη ομάδων, το $C_{logrank}^2$ συγκρίνεται με το στατιστικό X^2 με $k-1$ βαθμούς ελευθερίας.

Ως παράδειγμα δίνεται ο παρακάτω πίνακας που δίνει το χρόνο (μέρες) μέχρι το θάνατο ή εμφάνισης επιπλοκής μετά από θεραπεία σε 27 ασθενείς με λέμφωμα του Hodgkin (ομάδα HL) ή λέμφωμα άλλου τύπου (NHL)

NHL	42, 53, 57, 63, 81, 140, 176, 210*, 252, 476*, 524, 1037*
HL	30, 36, 41, 52, 52, 108, 132, 180*, 307*, 406*, 446*, 484*, 748*, 1290*, 1345*

Για τον υπολογισμό των u_j και v_j , ταξινομούμε όλους τους (πλήρεις) χρόνους αποτυχίας (και των δύο ομάδων μαζί) σε αύξουσα σειρά και υπολογίζουμε τους αναμενόμενους χρόνους θανάτου σε κάθε μη-αποκομμένο χρόνο.

$t_{(j)}$	n_{1j}	n_{2j}	d_{1j}	d_{2j}	u_j	v_j
30	12	15	0	1	-0,44444	0,2564
36	12	14	0	1	-0,46154	0,2585
41	12	13	0	1	-0,48	0,2496
42	12	12	1	0	0,5	0,2500
52	11	12	0	1	-0,47826	0,2727
53	11	11	1	0	0,5	0,2743
57	10	11	1	0	0,52381	0,2494
62	9	11	0	1	-0,45	0,2475
63	9	10	1	0	0,526316	0,2493
81	8	10	1	0	0,555556	0,2469
108	7	10	0	1	-0,41176	0,2422
132	7	9	0	1	-0,4375	0,2460
140	7	8	1	0	0,533333	0,2488
176	6	8	1	0	0,571429	0,2448
252	2	0	1	0	0	0,0000
ΣΥΝΟΛΟ					0,546934	3,536372

Πίνακας υπολογισμού του ελέγχου log-rank

Όσον αφορά τα αποτελέσματα πέραν της χρονικής στιγμής 180 δε συμβάλλουν στη σύγκριση, διότι δεν υπάρχουν πλέον μονάδες στην ομάδα (HL) που να λειτουργούν.

Άρα:

$$u^2 / v = (0.5469)^2 / 3,5363 = 0,084$$

Με p-τιμή μεγαλύτερη του .2 της χ^2 κατανομής με ένα βαθμό ελευθερίας, συμπεραίνεται ότι η κατανομή της διάρκειας ζωής δε διαφέρει μεταξύ των δύο

ομάδων.

Ο έλεγχος στην περίπτωση αυτή είναι αποτελεσματικός με :

Μηδενική υπόθεση H_0 : $S_1(t) = S_2(t)$ δεν υπάρχει διαφορά ανάμεσα στις καμπύλες επιβίωσης,

έναντι της εναλλακτικής H_1 : $S_1(t) \neq S_2(t)$ υπάρχει διαφορά ανάμεσα στις καμπύλες επιβίωσης,

4.1.1 Ο λόγος διακινδύνευσης

Ο λόγος O_A/E_A είναι μια εκτίμηση του κινδύνου στην ομάδα A. Μια τιμή μεγαλύτερη της μονάδας, υποδεικνύει ότι συμβαίνουν περισσότεροι θάνατοι στην ομάδα A από ότι αναμενόταν. Μπορούμε να συνδυάσουμε τους λόγους των δύο ομάδων και να πάρουμε έτσι ένα στατιστικό που δίνει χρήσιμα αποτελέσματα για μια συγκεκριμένη μελέτη. Συγκεκριμένα, μπορούμε να υπολογίσουμε ένα συντελεστή κινδύνου (Hazard Ratio - HR) που ορίζεται ως εξής:

$$HR = \frac{O_A/E_A}{O_B/E_B}$$

Στο παράδειγμα 4.1 ο συντελεστής κινδύνου είναι $HR = \frac{\frac{1}{\frac{3.75}{5}}}{\frac{1}{2.25}} = 0.12$

Δηλαδή, ο κίνδυνος με τη θεραπεία B (καθόλου θεραπεία) είναι περίπου 8(!) φορές μεγαλύτερος από τον κίνδυνο της θεραπείας A (CFM).

Άρα, όταν η τιμή του HR είναι μεγαλύτερη του 1, υποδεικνύει ότι οι ασθενείς που λαμβάνουν τη θεραπεία A πεθαίνουν με πιο γρήγορο ρυθμό από αυτούς που λαμβάνουν τη θεραπεία B, ενώ όταν η τιμή του HR είναι μικρότερη του 1 το αντίστροφο. Παρόλα αυτά, όταν εφαρμόσαμε το Logrank τεστ στα δεδομένα αυτά δεν βρήκαμε ότι η διαφορά αυτή είναι σημαντική.

4.2 To Cox - Mantel Test

Έστω $t(1) < \dots < t(k)$ οι διακεκριμένοι χρόνοι αποτυχίας στις δύο ομάδες και έστω $m_{(i)}$ ο αριθμός των χρόνων αποτυχίας που ισούνται με t_i , έτσι ώστε

$$\sum_{i=1}^{\kappa} m_{(i)} = r_1 + r_2$$

Επιπλέον, έστω $R(t)$ το σύνολο των ατόμων που βρίσκονται ακόμη σε κίνδυνο στο χρόνο t , με την υπόθεση ότι το συμβάν ή η αποκοπή είναι τουλάχιστον ίση με t .

Αν, n_{1t} και n_{2t} το πλήθος των ατόμων στο $R(t)$ που ανήκουν στις ομάδες 1 και 2 αντίστοιχα, τότε ο συνολικός αριθμός των παρατηρήσεων στο $R(t_i)$ είναι $r(i) = n_{1t} + n_{2t}$

Ορίζουμε,

$$U = r_2 - \sum_{i=1}^k m_{(i)} A_{(i)}$$

$$I = \sum_{i=1}^{\kappa} m_{(i)} \frac{m_{(i)} (r_{(i)} - m_{(i)})}{r_{(i)} - 1} A_{(i)} (1 - A_{(i)})$$

όπου:

$r(i)$: το πλήθος των παρατηρήσεων στο $R(t_i)$ και

$A_{(i)}$: η αναλογία των $r(i)$ που ανήκουν στη δεύτερη ομάδα.

Έτσι θεωρούμε το στατιστικό $C = \frac{U}{\sqrt{I}}$ ως μια τυπική κανονική μεταβλητή κάτω από τη μηδενική υπόθεση (Cox 1972). Για την καλύτερη κατανόηση παραθέτουμε το παράδειγμα με τα δεδομένα της προηγούμενης ενότητας (4.1)

CMF(Group 1)	16+	18+	20+	23	24+
Control (Group 2)	15	18	19	19	20

Με μηδενική υπόθεση

$H_0: S_1 = S_2$ (οι δύο θεραπείες είναι το ίδιο αποτελεσματικές) έναντι της υπόθεσης

$H_1: S_1(t) > S_2$ (η θεραπεία CMF πιο αποτελεσματική από την καθόλου θεραπεία).

Υπάρχουν $k = 5$ διακριτοί χρόνοι αποτυχίας στις δύο ομάδες, $r_1 = 1$, $r_2 = 5$

Στον παρακάτω πίνακα δίνονται οι υπολογισμοί του παραδείγματος

Χρόνος υποτροπίασης t_i	m_i	n_{1t}	n_{2t}	r_i	A_i
15	1	5	5	10	0,5
18	1	4	4	8	0,5
19	2	3	3	6	0,5
20	1	3	1	4	0,25
23	1	2	0	2	0

$$U = 5 - (0.5 + 0.5 + 1 + 0.25 + 0) = 2.25$$

$$I = \frac{1 \times 9}{9} (0.5 \times 0.5) + \frac{1 \times 7}{7} (0.5 \times 0.5) + \frac{2 \times 4}{5} (0.5 \times 0.5) + \frac{1 \times 3}{3} (0.25 \times 0.75) = 0.25 + 0.25 + 0.4 + 0.1875 = 1.0875$$

$$\text{Συνεπώς, } C = 2.75 / \frac{1}{\sqrt{1.0875}} = 2.637 > Z_{0.05} = 1.64 \text{ και τελικά απορρίπτουμε την } H_0,$$

δηλαδή οι δύο θεραπείες δεν είναι το ίδιο αποτελεσματικές.

Κεφάλαιο πέμπτο

Το μοντέλο αναλογικών κινδύνων του COX

Εισαγωγή

Το μοντέλο αναλογικών κινδύνων που προτάθηκε από τον Cox (1972) έχει χρησιμοποιηθεί ευρέως σε δεδομένα επιβίωσης. Είναι εύκολο να εφαρμοστεί και εύκολο να ερμηνευτεί, προτρέποντας συνήθως τον ερευνητή να είναι η πρώτη του επιλογή όταν αντιμετωπίζει δεδομένα διάρκειας ζωής. Όσον αφορά την αξιοπιστία του συγκεκριμένου μοντέλου υπάρχει η άποψη ότι το μοντέλο εμπεριέχει αρκετές ελλείψεις και προτείνεται η περιορισμένη χρήση του για την ανάλυση αναλογικών κινδύνων (Lawless, 1983). Παρόλα αυτά, και όσον αφορά την αξιοπιστία του μοντέλου, αρκετοί ερευνητές δεν συμμερίζονται τις απόψεις του Lawless, και χρησιμοποιούν το μοντέλο στην επεξεργασία των δεδομένων όπως οι Bendell et al., (1986), Bendell et al., (1991), Dale, (1985), Krivtsov et al, (2002), Madeiras et al. , (2011).

Στο προηγούμενο κεφάλαιο αναπτύχθηκαν διάφοροι έλεγχοι, για τον έλεγχο της αξιοπιστίας των συναρτήσεων επιβίωσης διαφόρων ομάδων, λαμβάνοντας υπόψη τις τιμές ενός άλλου χαρακτηριστικού που μπορεί να επηρεάζει το χρόνο ζωής των υπό εξέταση μονάδων. Όταν έχουμε περισσότερες από μία μεταβλητή, τότε οι μέθοδοι γίνονται περίπλοκες και είναι δύσκολο να εφαρμοστούν. Το Logrank test, όπως και οι άλλοι έλεγχοι, δε μπορούν να προσαρμοστούν για τη διερεύνηση της επίδρασης διαφόρων μεταβλητών ταυτόχρονα που είναι γνωστό ότι επηρεάζουν την διάρκεια ζωής.

Η εύρεση της σχέσης μεταξύ μιας μεταβλητής που δηλώνει το χρόνο επιβίωσης μιας μονάδας (ατόμου ή μηχανήματος) και άλλων συμμεταβλητών, επιτυγχάνεται συνήθως μέσω ενός μοντέλου παλινδρόμησης. Όταν έχουμε αποκομμένα δεδομένα επιβίωσης, χρησιμοποιείται συνήθως το μοντέλο παλινδρόμησης του Cox (Cox regression model) ή διαφορετικά το μοντέλο αναλογικού κινδύνου του Cox (Cox proportional hazard model).

5.1 Ορισμός του Μοντέλου

Θεωρούμε ότι έχουμε ένα συγκεκριμένο αριθμό μονάδων προς μελέτη π.χ. n και ότι το $x_1, x_2, \dots, x_p$ με p να ορίζουμε των αριθμό των μεταβλητών που πιστεύουμε ότι

επηρεάζουν το χρόνο ζωής αυτών των μονάδων. Οι μεταβλητές αυτές μπορεί να παριστάνουν διάφορα χαρακτηριστικά, αν υποθέσουμε ότι οι μονάδες αυτές είναι άτομα, τότε μερικές από τις μεταβλητές μπορεί να είναι:

Θεραπείες ασθενών

Χαρακτηριστικά ασθενών

Επιπλέον ορίζουμε το $x_i = (x_{1i}, x_{2i}, \dots, x_{pi})$, $i = 1, 2, \dots, n$, να είναι οι τιμές των συμμεταβλητών που αντιστοιχούν στο i άτομο.

Η γενική μορφή ενός μοντέλου αναλογικού κινδύνου είναι:

$$h(t|x) = h_0(t)g(x)$$

όπου $g(x)$ είναι μια συνάρτηση του x και $h_0(t)$ μπορεί να θεωρηθεί η βάση της γραμμής κινδύνου για κάθε μονάδα κατά την οποία $g(x) = 1$. Το μοντέλο αναλογικού κινδύνου του Cox υποθέτει ότι η συνάρτηση κινδύνου είναι μια εξαρτημένη μεταβλητή που δίδεται ο τύπος ως παρακάτω

$$h(t|x) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p) = h_0(t) \exp\left(\sum_{j=1}^p \beta_j x_j\right)$$

όπου $h_0(t)$ είναι η συνάρτηση κινδύνου της κατανομής των δεδομένων διάρκειας ζωής όταν οι μεταβλητές x θεωρούνται ότι ισούνται με το μηδέν και τα βήτα (β) είναι οι παράμετροι του μοντέλου.

Αρχικά υποθέτουμε ότι οι συμμεταβλητές δεν εξαρτώνται από το χρόνο, ότι δηλαδή οι τιμές των συμμεταβλητών x_i καταγράφηκαν στην αρχή της μελέτης, στο $t = 0$, και ότι οι τιμές αυτές είναι σταθερές καθ'όλη τη διάρκεια της μελέτης. Άρα σύμφωνα με τα παραπάνω, στο μοντέλο αναλογικής διακινδύνευσης Cox, οι συμμεταβλητές x δρουν στη συνάρτηση διακινδύνευσης μέσω της σχέσης

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

όπου $h_0(t)$ είναι μια βασική συνάρτηση κινδύνου και β' ένα διάνυσμα p συντελεστών, οι οποίοι εκφράζουν ποσοτικά την επίδραση της καθεμιάς των συμμεταβλητών x . Η ανεξαρτησία της διακινδύνευσης και κατά συνέπεια και της επιβίωσης από τη συμμεταβλητή x_i σημαίνει ότι $\beta'_i = 0$.

Τελικά, η διακινδύνευση $h(t|x)$, εξαρτάται από το χρόνο και τις συµµεταβλητές. Όσον αφορά το χρόνο, $h_0(t)$, είναι µια συνάρτηση του χρόνου που θεωρείται η ίδια και για τα n άτοµα και όσον αφορά τις συµµεταβλητές θεωρούνται µια ποσότητα που εξαρτάται από τις συµµεταβλητές µόνο µέσω του διανύσµατος β' .

Ας θεωρήσουµε την κινδυνότητα (Hazard Ratio – HR) ως το λόγο των συναρτήσεων διακινδύνευσης δύο ατόµων. Στο µοντέλο αναλογικής διακινδύνευσης του Cox, παρατηρείται η εξής ιδιότητα: οι συναρτήσεις διακινδύνευσης των εκάστοτε ατόµων είναι ανάλογες µεταξύ τους. Δηλαδή, έστω ότι $[h(t|x_1)/h(t|x_2)]$ ο λόγος των συναρτήσεων διακινδύνευσης δύο ατόµων και $x_1 = (x_{11}, x_{12}, \dots, x_{1p})$, $x_2 = (x_{21}, x_{22}, \dots, x_{2p})$ να είναι τα διανύσµατα των συµµεταβλητών. Τότε ο λόγος αυτός τότε είναι σταθερός (ανεξάρτητος του χρόνου):

$$HR(t) = \frac{h(t|x_1)}{h(t|x_2)} = \frac{h_0(t)e^{\beta'x_1}}{h_0(t)e^{\beta'x_2}} = e^{\beta'(x_1 - x_2)}$$

Η κύρια παραδοχή σύµφωνα µε το µοντέλο το Cox είναι ότι η αναλογία κινδύνου, που είναι η αναλογία της συνάρτησης κινδύνου ως προς τον βασικό κίνδυνο, είναι συνεχής µε την πάροδο του χρόνου και ερµηνεύεται ως σχετική και όχι ως απόλυτη. Η χρήση της εκθετικής συνάρτησης εξασφαλίζει ότι ο κίνδυνος είναι θετικός.

Κεφάλαιο 6

Μελέτη της «διάρκειας ζωής» εταιρειών στην Ιταλία

6.1 Δεδομένα μελέτης

Firms data

Για την ολοκλήρωση της μελέτης δίδονται στοιχεία για τη “διάρκεια ζωής” 969 εταιρειών σε δύο περιοχές της Ιταλίας.

Οι μεταβλητές τις οποίες θα χρησιμοποιήσουμε ορίζονται ως παρακάτω:

Case_number A/A από 1 έως 969

Location Η γεωγραφική περιοχή: 0 = PG / 1 = TR

Activity Κάθε ξεχωριστός κωδικός αντιστοιχεί σε μια δραστηριότητα για την κάθε εταιρεία. Αυτή η πληροφορία δεν καταγράφεται για όλες τις εταιρείες.

Start Το έτος ίδρυσης της εταιρείας

Outcome Τι απέγινε: 1 = ακόμα υπάρχει (το Φεβ 2015 οπότε τελείωσε η μελέτη), 2 / 3 / 4 = έχει κλείσει, με διάφορους τρόπους

Status 0 = ακόμα υπάρχει = right-censored στην ορολογία της ανάλυσης επιβίωσης
1 = έχει κλείσει (δηλ., οι κωδικοί 2, 3 ,4 της Outcome)

Duration Η “διάρκεια ζωής” από την έναρξη μέχρι το κλείσιμο (ή στο τέλος της μελέτης αν δεν έχει κλείσει)

Κατά την αξιολόγηση των δεδομένων παρατηρήθηκε ότι η ημερομηνία έναρξης ορισμένων εταιρειών δεν ήταν ορισμένη και συνεπώς δεν μπορούσε να εκτιμηθεί η διάρκεια ζωής των συγκεκριμένων εταιρειών και τελικά αφαιρέθηκαν τα σχετικά δεδομένα. Έμειναν 959 εταιρείες στην ανάλυση.

Σκοπός της μελέτης είναι η ανάλυση επιβίωσης της διάρκειας ζωής των εταιρειών στις δύο αναφερόμενες περιοχές της Ιταλίας και τα χαρακτηριστικά τους, μέσω των Καμπυλών Kaplan-Meier, τους ελέγχους log-rank και της παλινδρόμηση του Cox.

Πιο συγκεκριμένα η μελέτη αφορά στο αν η διάρκεια ζωής των επιχειρήσεων διαφέρει μεταξύ των περιοχών ή ανάλογα με την δραστηριότητα ή ανάλογα με το έτος έναρξης εργασιών.

Επίσης, σε αυτά τα δεδομένα έχουμε “competing risks” – ο “θάνατος” δεν έρχεται μόνο με έναν τρόπο, αλλά υπάρχουν 3 τρόποι (οι κωδικοί 2, 3, 4 του Outcome). Στη συνέχεια θα δούμε τον καθένα από αυτούς. Δηλαδή θα κάνουμε τις παραπάνω αναλύσεις 4 φορές

(α) Status ως έχει παραπάνω (= κλείσιμο από οποιαδήποτε αιτία)

(β) Status = 1 αν έκλεισε για τον 2ο λόγο, 0 αλλιώς (δηλαδή, υπάρχει ακόμα, ή έκλεισε αλλά όχι με τον 2ο τρόπο)

(γ) Status = 1 αν έκλεισε για τον 3ο λόγο, 0 αλλιώς

(δ) Status = 1 αν έκλεισε για τον 4ο λόγο, 0 αλλιώς

6.2 Αποτελέσματα

6.2.1 Σύγκριση ανάμεσα στις περιοχές

Όσον αφορά τη σύγκριση ανάμεσα στις περιοχές έχουμε:

Case Processing Summary				
Location	Total N	N of Events	Censored	
			N	Percent
PG	157	101	56	35,7%
TR	801	578	223	27,8%
Overall	958	679	279	29,1%

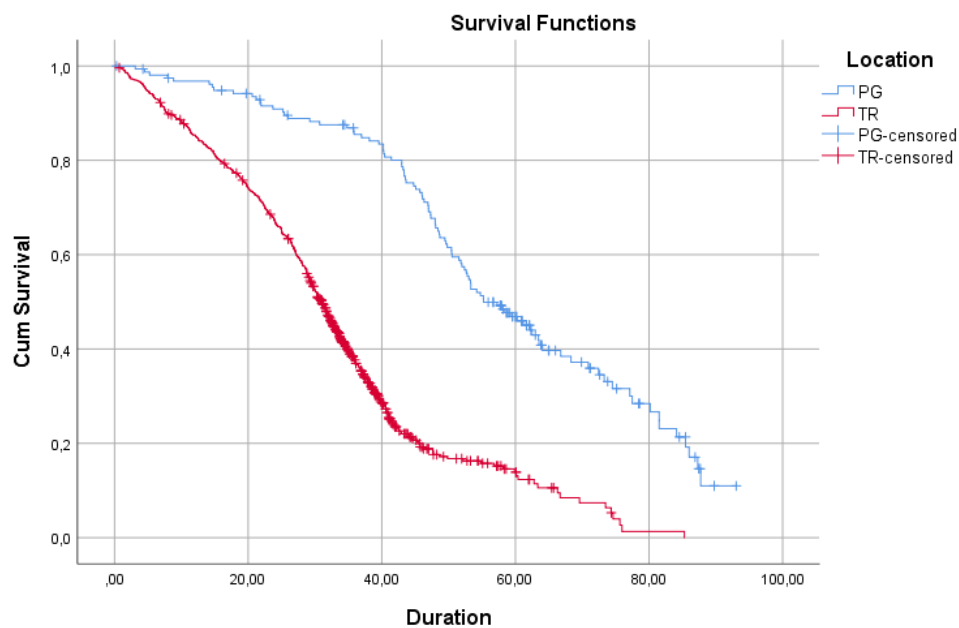
Πίνακας 6.1 Δεδομένα της K-M ανά κατηγορία της μεταβλητής «Location»

Medians for Survival Time				
Location	Median			
	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
PG	55,222	3,452	48,456	61,987
TR	31,025	0,627	29,795	32,254
Overall	33,715	0,808	32,131	35,299

Πίνακας 6.2 Αποτελέσματα της K-M ανά κατηγορία της μεταβλητής «Location»

Overall Comparisons			
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	118,137	1	,000

Πίνακας 6.3 Αποτελέσματα Log Rank test



Εικόνα 6.1 Γραφική αναπαράσταση ανά κατηγορία της μεταβλητής «Location»

Στον Πίνακα 6.1 « Case Processing Summary» βλέπουμε τα συγκεντρωτικά στοιχεία από την Kaplan-Meier ανάλυση δεδομένων. Έχουμε το σύνολο των μονάδων ανά περιοχή, τον αριθμό των συμβάντων “event” και τον αριθμό των αποκομμένων μονάδων καθώς και το ποσοστό αυτών. Στον Πίνακα 6.2 «Medians for Survival time» βλέπουμε το δείκτη « median», όπου για την περιοχή PG είναι 55,222 χρόνια, ενώ για

την περιοχή TR είναι 31,025, επίσης πολύ μικρότερος. Δηλαδή, βλέπουμε αρκετά σημαντική διαφορά στα χρόνια λειτουργίας των εταιρειών ανάμεσα στις δύο περιοχές της Ιταλίας. Στη συνέχεια θέλουμε να εξακριβώσουμε αν αυτή η διαφορά είναι και στατιστικά σημαντική, όπου σύμφωνα με τον Πίνακα 6.3 «Overall Comparisons» παρατηρούμε ότι η τιμή-P του ελέγχου στατιστικής σημαντικότητας, significant test, είναι 0.001, δηλαδή στατιστικά σημαντικό άρα ισχύει η υπόθεση ότι οι διαφορές στη διάρκεια λειτουργίας των εταιρειών είναι εξαρτώμενες από την περιοχή. Τέλος, συγκρίνουμε τις δύο καμπύλες επιβίωσης που αφορούν τις δύο περιοχές Εικόνα 6.1. Η μπλε καμπύλη αφορά την PG περιοχή και η κόκκινη καμπύλη αφορά την TR, πάνω στις οποίες είναι επίσης αποτυπωμένες και οι περιπτώσεις αποκοπής δεδομένων (censored). Σε συμφωνία με τα προηγούμενα, η προβολή της καμπύλης της περιοχής PG στον άξονα της χρονικής διάρκειας δείχνει ότι η διάρκεια λειτουργίας- επιβίωσης για τις εταιρείες που λειτουργούν σ' αυτή, είναι πολύ μεγαλύτερη. Ως συμπέρασμα, μπορούμε να πούμε ότι η περιοχή λειτουργίας των εταιρειών επηρεάζει και τη διάρκεια της λειτουργίας τους.

6.2.2 Σύγκριση ανάμεσα σε διαφορετικές δραστηριότητες

Όσον αφορά τη σύγκριση ανάμεσα στις διαφορετικές δραστηριότητες των εταιρειών έχουμε:

Case Processing Summary				
Activity	Total N	N of Events	Censored	
			N	Percent
0	54	26	28	51,9%
1	72	47	25	34,7%
2	93	60	33	35,5%
3	24	18	6	25,0%
4	292	182	110	37,7%
5	21	9	12	57,1%
6	74	39	35	47,3%
7	18	10	8	44,4%
8	27	8	19	70,4%
9	6	4	2	33,3%
Overall	681	403	278	40,8%

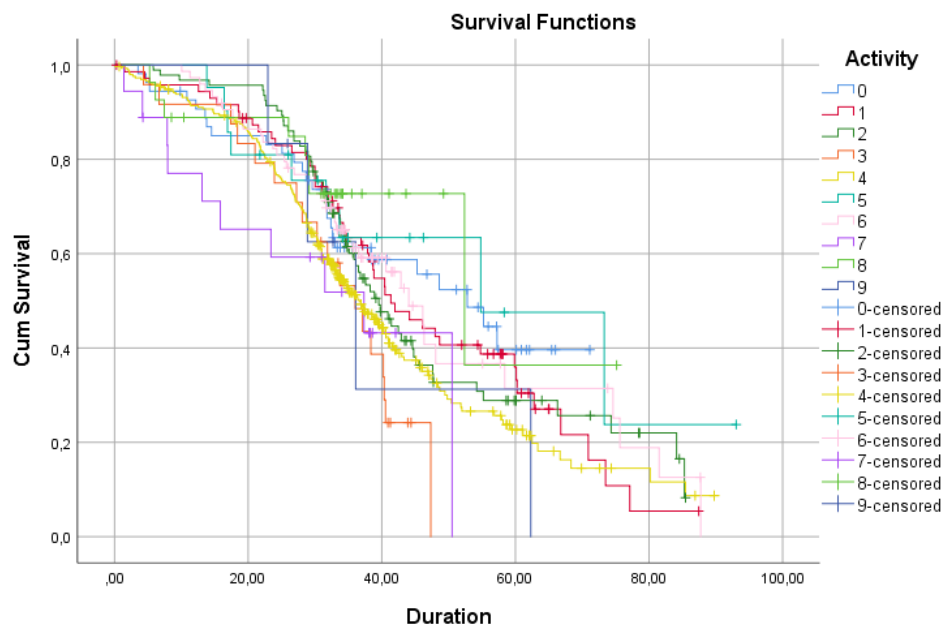
Πίνακας 6.4 Δεδομένα της K-M ανά κατηγορία της μεταβλητής «Activity»

Medians for Survival Time				
Activity	Median			
	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
0	52,781	7,091	38,882	66,680
1	41,367	3,769	33,980	48,754
2	39,562	2,376	34,905	44,219
3	36,008	3,788	28,584	43,433
4	36,625	1,844	33,010	40,239
5	54,808	16,121	23,211	86,405
6	44,047	3,592	37,007	51,086
7	37,332	11,330	15,124	59,539
8	52,386	16,688	19,678	85,095
9	36,099	5,667	24,991	47,206
Overall	39,573	1,102	37,413	41,732

Πίνακας 6.5 Αποτελέσματα της K-M ανά κατηγορία της μεταβλητής «Activity»

Overall Comparisons			
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	17,860	9	,037

Πίνακας 6.6 Αποτελέσματα Log Rank test



Εικόνα 6.2 Γραφική αναπαράσταση ανά κατηγορία της μεταβλητής «Activity»

Στον Πίνακα 6.4 « Case Processing Summary» βλέπουμε τα συγκεντρωτικά στοιχεία από την Kaplan-Meier ανάλυση δεδομένων. Έχουμε το σύνολο των μονάδων ανά

δραστηριότητα , τον αριθμό των συμβάντων “event” και τον αριθμό των αποκομμένων μονάδων καθώς και το ποσοστό αυτών. Στον Πίνακα 6.5 « Medians for Survival time» βλέπουμε το δείκτη « median», όπου για την δραστηριότητα 7 η τιμή είναι 37,332 χρόνια, ενώ για την δραστηριότητα 5 είναι 54,808, τιμή πολύ μεγαλύτερη από τις υπόλοιπες δραστηριότητες. Δηλαδή, βλέπουμε αρκετά σημαντική διαφορά στα χρόνια λειτουργίας των εταιρειών ανάμεσα στις δραστηριότητες και για τις δύο περιοχές της Ιταλίας. Στη συνέχεια θέλουμε να εξακριβώσουμε αν αυτή η διαφορά είναι και στατιστικά σημαντική, όπου σύμφωνα με τον Πίνακα 6.6 «Overall Comparisons» παρατηρούμε ότι η τιμή-P του ελέγχου στατιστικής σημαντικότητας, significant test, είναι 0.037, δηλαδή στατιστικά σημαντικό άρα ισχύει η υπόθεση ότι οι διαφορές στη διάρκεια λειτουργίας των εταιρειών είναι εξαρτώμενες από την δραστηριότητα. Τέλος, συγκρίνουμε τις καμπύλες επιβίωσης που αφορούν τις δραστηριότητες (Εικόνα 6.2). Η μπλε καμπύλη αφορά την δραστηριότητα 5 και η μωβ καμπύλη αφορά την δραστηριότητα 7, πάνω στις οποίες είναι επίσης αποτυπωμένες και οι περιπτώσεις αποκοπής δεδομένων (censored). Σε συμφωνία με τα προηγούμενα, η προβολή της καμπύλης της δραστηριότητας 5 στον άξονα της χρονικής διάρκειας δείχνει ότι η διάρκεια λειτουργίας- επιβίωσης για τις εταιρείες που έχουν ως αντικείμενο αυτή τη δραστηριότητα, είναι πολύ μεγαλύτερη από τις υπόλοιπες. Ως συμπέρασμα, μπορούμε να πούμε ότι η δραστηριότητα των εταιρειών επηρεάζει και τη διάρκεια της λειτουργίας τους.

6.2.3 Η χρήση του μοντέλου του Cox

Όσον αφορά την εφαρμογή του μοντέλου των συμμεταβλητών του Cox έχουμε:

Case Processing Summary

		N	Percent
Cases available in analysis	Event ^a	679	70,9%
	Censored	278	29,0%
	Total	957	99,9%
Cases dropped	Cases with missing values	0	0,0%
	Cases with negative time	0	0,0%
	Censored cases before the earliest event in a stratum	1	0,1%
	Total	1	0,1%
Total		958	100,0%

a. Dependent Variable: Duration

Πίνακας 6.7 Δεδομένα της Cox Ανάλυση με εξαρτημένη μεταβλητή «Duration»

Categorical Variable Codings^a

		Frequency	(1)	(2)	(3)	(4)
New_activity ^b	1,00	72	1	0	0	0
	2,00	93	0	1	0	0
	3,00	427	0	0	1	0
	4,00	292	0	0	0	1
	6,00	74	0	0	0	0

a. Category variable: New_activity

b. Indicator Parameter Coding

Πίνακας 6.8 Δεδομένα παραμετροποίησης νέας μεταβλητής «New Activity»

Omnibus Tests of Model Coefficients^a

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi- square	df	Sig.	Chi- square	df	Sig.	Chi- square	df	Sig.
	7785,981	343,635	6	0,000	452,098	6	0,000	452,098	6

a. Beginning Block Number 1. Method = Enter

Πίνακας 6.9 Πίνακας Παραμέτρων Μοντέλου

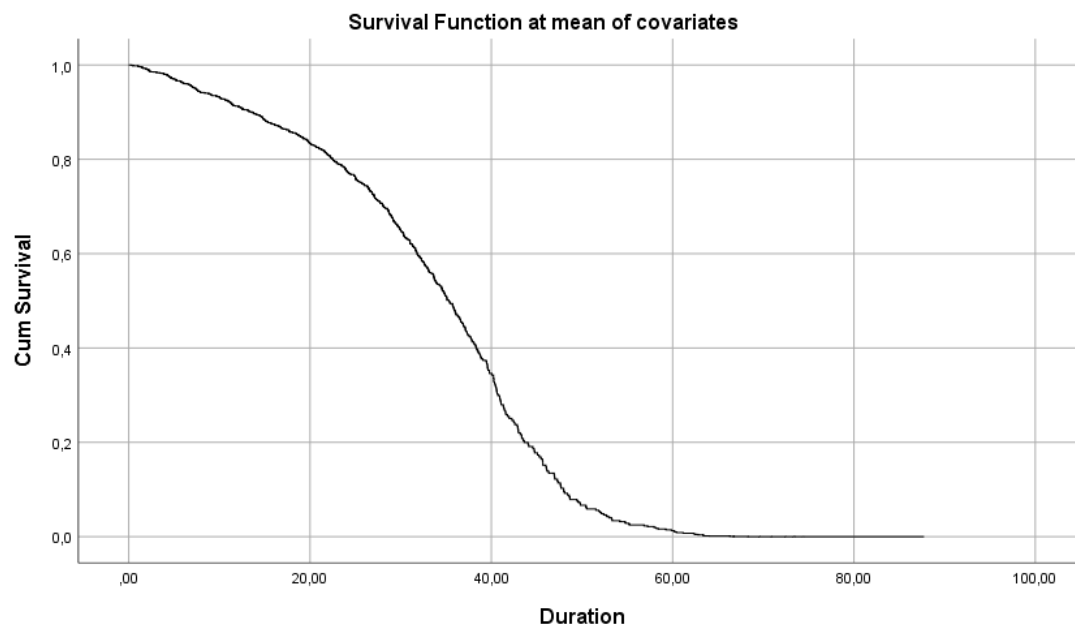
Variables in the Equation								
	B	SE	Wald	df	Sig.	Exp(B)	95,0% CI for Exp(B)	
							Lower	Upper
Location	,382	,144	7,008	1	,008	1,465	1,104	1,944
Start	,070	,005	191,963	1	,000	1,073	1,062	1,083
New_activity			85,189	4	,000			
New_activity(1)	,293	,219	1,795	1	,180	1,340	,873	2,057
New_activity(2)	,195	,206	,896	1	,344	1,216	,811	1,822
New_activity(3)	,933	,170	30,277	1	,000	2,542	1,823	3,544
New_activity(4)	,275	,177	2,424	1	,119	1,317	,931	1,863

Πίνακας 6.10 Έλεγχος υποθέσεων Μοντέλου

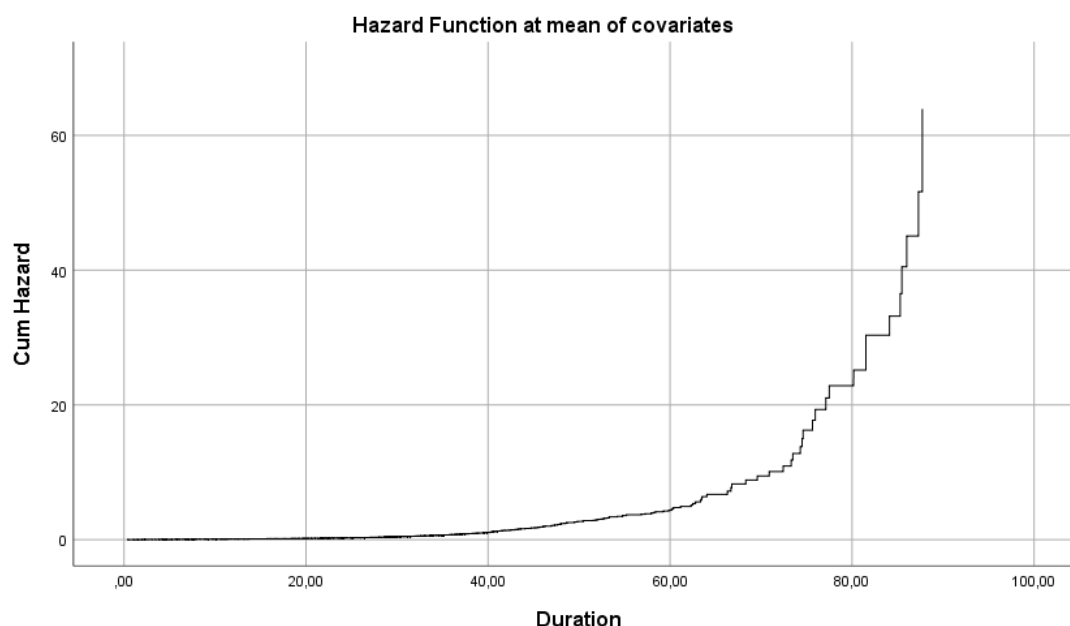
Covariate Means

	Mean
Location	,837
Start	1971,980
New_activity(1)	,074
New_activity(2)	,097
New_activity(3)	,446
New_activity(4)	,305

Πίνακας 6.11 Πίνακας συσχετίσεων μεταβλητής «New Activity»



Εικόνα 6.3 Γραφική αναπαράσταση της $S(t)$



Εικόνα 6.4 Γραφική αναπαράσταση της $H(t)$

Στον Πίνακα 6.7 « Case Processing Summary» βλέπουμε τα συγκεντρωτικά στοιχεία από την Cox ανάλυση δεδομένων. Έχουμε το σύνολο των μονάδων στην μελέτη, τον αριθμό των συμβάντων “event” και τον αριθμό των αποκομμένων μονάδων καθώς και το ποσοστό αυτών. Επίσης, βλέπουμε και την κατηγοριοποίηση των μεταβλητών «Δραστηριότητα» μετά τη δημιουργία της νέας μεταβλητής «New Activity» (Πίνακας 6.8)

Λόγω του ότι μερικές κατηγορίες της μεταβλητής «Activity» έχουν χαμηλές συχνότητες, δημιουργήσαμε με νέα μεταβλητή «New_activity» που κρατάει τους αρχικούς κωδικούς 1, 2, 4 και 6 ως έχουν, ενώ οι υπόλοιποι κωδικοί ενώνονται στο καινούργιο κωδικό 3. Αυτό συμβαίνει επειδή οι χαμηλές συχνότητες εμποδίζουν την προσαρμογή του μοντέλου στα δεδομένα.

Στον Πίνακα 6.9 «Omnibus Tests of Model Coefficients» βλέπουμε το χ^2 value, τους βαθμούς ελευθερίας και η τιμή-P του ελέγχου στατιστικής σημαντικότητας, significant test, είναι 0.001. Συμπεραίνουμε από τα στοιχεία αυτού του πίνακα ότι υπάρχει τουλάχιστον μια μεταβλητή πρόβλεψης των δεδομένων για τα δεδομένα διάρκειας επιβίωσης. Στον επόμενο Πίνακα 6.10 «Variables in the Equation» αρχικά παρατηρούμε τη μεταβλητή «New_activity». Εφόσον έχουμε ότι η τιμή-P του

ελέγχου στατιστικής σημαντικότητας, significant test, είναι (sign.=0.001), τότε υπάρχει διαφοροποίηση της επιβίωσης μεταξύ των κατηγοριών της μεταβλητής αυτής. Στη συνέχεια, κοιτάζουμε τους επιμέρους ελέγχους για κάθε κατηγορία σε σύγκριση με την κατηγορία αναφοράς – η οποία είναι η τελευταία (New_Activity = 6). Βλέπουμε ότι είναι η κατηγορία 3 που διαφέρει (με μεγαλύτερο hazard επειδή $B > 0$).

Στην περίπτωση που ο ολικός έλεγχος ήταν μη σημαντικός, τότε δεν θα προχωρούσαμε στην αξιολόγηση των επιμέρους κατηγοριών.

Όσον αφορά την αναλογία κινδύνου σύμφωνα με το μοντέλο του Cox παρατηρούμε στη στήλη $\exp(B)$ τις παρακάτω τιμές: Για την μεταβλητή «Location» η τιμή είναι 1.465, για την μεταβλητή «Start» είναι 1.073, για τη μεταβλητή «New Activity 1» η τιμή είναι 1.34, για τη μεταβλητή «New Activity 2» η τιμή είναι 1.216, για τη μεταβλητή «New Activity 3» η τιμή είναι 2.542 και για τη μεταβλητή «New Activity 4» η τιμή είναι 1.317.

Μια τιμή κοντά στο 1, σημαίνει ότι δεν υπάρχει συσχέτιση ανάμεσα στη μεταβλητή και στην πρόβλεψη της διάρκειας ζωής των υπό εξέταση μονάδων (εταιρειών), όπως είναι το αποτέλεσμα για το σύνολο των υπό εξέταση μεταβλητών, εκτός από την μεταβλητή «New Activity 3», όπου η τιμή είναι 2.542. Συνεπώς και σε συνέχεια των προηγούμενων η συγκεκριμένη μεταβλητή επηρεάζει τη διάρκεια λειτουργίας των εταιρειών. Με άλλα λόγια, η αναλογικότητα συνεπάγεται ότι οι ποσότητες $\exp(b_i)$ ονομάζονται λόγοι κινδύνου. Μια τιμή b_i μεγαλύτερη από μηδέν, ή ισοδύναμα ένας λόγος επικινδυνότητας μεγαλύτερος από ένα, δείχνει ότι καθώς η τιμή της i -οστής μεταβλητής αυξάνεται, ο κίνδυνος συμβάντος αυξάνεται και συνεπώς το μήκος της επιβίωσης μειώνεται. Επίσης, ένας λόγος επικινδυνότητας άνω του 1 υποδεικνύει μια συμμεταβλητή που συνδέεται θετικά με την πιθανότητα συμβάντος και έτσι σχετίζεται αρνητικά με το μήκος της επιβίωσης.

6.2.4 Competing risks

Στη συνέχεια μια ακόμη ανάλυση κρίθηκε αναγκαία για τη διερεύνηση σε βάθος στο κατά πόσο ο λόγος διακοπής των εργασιών των εταιρειών θα μπορούσε να συσχετιστεί με τη διάρκεια επιβίωσης λειτουργίας τους. Στην προηγούμενη ενότητα, η ανάλυση εστίαζε στη συσχέτιση της διάρκειας επιβίωσης με τη διακοπή λειτουργίας της εταιρίας για οποιοδήποτε λόγο. Σε αυτή την ενότητα, θα διερευνήσουμε στο αν οι

διαφορετικοί λόγοι (competing risks) διακοπής συσχετίζονται με τη διάρκεια επιβίωσης των εταιρειών. Συγκεκριμένα, από τα δεδομένα έχουμε στην μεταβλητή «Outcome» τις παρακάτω περιπτώσεις:

Κωδικός «2» για «closed»

Κωδικός «3» για «Dissolution_Liquidation»

Κωδικός «4» για «Legal Procedure»

Ακολουθούν πίνακες των σχετικών αποτελεσμάτων.

Variables in the Equation							95,0% CI for Exp(B)	
	B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
Location	,138	,256	,290	1	,590	1,148	,695	1,896
Start	,150	,012	164,943	1	,000	1,162	1,136	1,189
New_activity			12,032	4	,017			
New_activity(1)	,341	,310	1,208	1	,272	1,406	,766	2,582
New_activity(2)	,278	,281	,975	1	,323	1,320	,761	2,293
New_activity(3)	,589	,240	6,010	1	,014	1,803	1,125	2,889
New_activity(4)	,146	,247	,351	1	,553	1,158	,714	1,878

Πίνακας 6.12 με χρήση της μεταβλητής «outcome» και της εναλλακτικής [2]

Στον Πίνακα 6.12 «Variables in the Equation» αρχικά παρατηρούμε τη μεταβλητή «New_activity». Εφόσον έχουμε την τιμή-P του ελέγχου στατιστικής σημαντικότητας, significant test, να είναι (sign.=0.017), τότε υπάρχει διαφοροποίηση της επιβίωσης μεταξύ των κατηγοριών της μεταβλητής αυτής. Στη συνέχεια, κοιτάζουμε τους επιμέρους ελέγχους για κάθε κατηγορία σε σύγκριση με την κατηγορία αναφοράς – η οποία είναι η τελευταία (New_Activity = 6). Βλέπουμε ότι είναι η κατηγορία 3 που διαφέρει (με μεγαλύτερο hazard επειδή $B > 0$).

Στην περίπτωση που ο ολικός έλεγχος ήταν μη σημαντικός, τότε δεν θα προχωρούσαμε στην αξιολόγηση των επιμέρους κατηγοριών.

Όσον αφορά την αναλογία κινδύνου σύμφωνα με το μοντέλο του Cox παρατηρούμε στη στήλη exp (B) τις παρακάτω τιμές: Για την μεταβλητή «Location» η τιμή είναι 1.148, για την μεταβλητή «Start» είναι 1.162, για τη μεταβλητή «New Activity 1» η τιμή είναι 1.406, για τη μεταβλητή «New Activity 2» η τιμή είναι 1.320, για τη

μεταβλητή «New Activity 3» η τιμή είναι 1.803 και για τη μεταβλητή «New Activity 4» η τιμή είναι 1.158. Μεγαλύτερη συσχέτιση έχουμε με την μεταβλητή «New Activity 3» με την τιμή-P του ελέγχου στατιστικής σημαντικότητας, significant test, να είναι.=0.014».

Variables in the Equation							95,0% CI for Exp(B)	
	B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
Location	,779	,221	12,378	1	,000	2,179	1,412	3,362
Start	,034	,006	30,314	1	,000	1,035	1,022	1,048
New_activity			93,959	4	,000			
New_activity(1)	,340	,370	,847	1	,357	1,405	,681	2,902
New_activity(2)	-,140	,379	,135	1	,713	,870	,413	1,830
New_activity(3)	1,401	,286	23,975	1	,000	4,059	2,317	7,111
New_activity(4)	,492	,299	2,703	1	,100	1,636	,910	2,940

Πίνακας 6.13 με χρήση της μεταβλητής «outcome» και της εναλλακτικής [3]

Στον Πίνακα 6.13 «Variables in the Equation» αρχικά παρατηρούμε τη μεταβλητή «New_activity». Εφόσον έχουμε την τιμή-P του ελέγχου στατιστικής σημαντικότητας, significant test, να είναι (sign.=0.001), τότε υπάρχει διαφοροποίηση της επιβίωσης μεταξύ των κατηγοριών της μεταβλητής αυτής. Στη συνέχεια, κοιτάζουμε τους επιμέρους ελέγχους για κάθε κατηγορία σε σύγκριση με την κατηγορία αναφοράς – η οποία είναι η τελευταία (New_Activity = 6). Βλέπουμε ότι είναι η κατηγορία 3 που διαφέρει (με μεγαλύτερο hazard επειδή B>0).

Στην περίπτωση που ο ολικός έλεγχος ήταν μη σημαντικός, τότε δεν θα προχωρούσαμε στην αξιολόγηση των επιμέρους κατηγοριών.

Όσον αφορά την αναλογία κινδύνου σύμφωνα με το μοντέλο του Cox παρατηρούμε στη στήλη exp (B) τις παρακάτω τιμές: Για την μεταβλητή «Location» η τιμή είναι 2.179, για την μεταβλητή «Start» είναι 1.035, για τη μεταβλητή «New Activity 1» η τιμή είναι 1.405, για τη μεταβλητή «New Activity 2» η τιμή είναι 0.870, για τη μεταβλητή «New Activity 3» η τιμή είναι 4.059 και για τη μεταβλητή «New Activity 4» η τιμή είναι 1.636. Μεγαλύτερη συσχέτιση έχουμε με την μεταβλητή «New Activity 3» με την τιμή-P του ελέγχου στατιστικής σημαντικότητας, significant test, να είναι «sign.=0.001».

Variables in the Equation								
	B	SE	Wald	df	Sig.	Exp(B)	95,0% CI for Exp(B)	
							Lower	Upper
Location	-,536	,343	2,434	1	,119	,585	,299	1,147
Start	,078	,014	32,277	1	,000	1,081	1,052	1,110
New_activity			1,761	4	,780			
New_activity(1)	,312	,580	,290	1	,590	1,367	,439	4,257
New_activity(2)	,622	,538	1,336	1	,248	1,863	,649	5,351
New_activity(3)	,437	,489	,796	1	,372	1,547	,593	4,037
New_activity(4)	,274	,495	,306	1	,580	1,315	,498	3,473

Πίνακας 6.14 με χρήση της μεταβλητής «outcome» και της εναλλακτικής [4]

Στον Πίνακα 6.14 «Variables in the Equation» αρχικά παρατηρούμε τη μεταβλητή «New_activity». Εφόσον έχουμε την τιμή-P του ελέγχου στατιστικής σημαντικότητας, significant test, να είναι (sign.=0.780), δηλαδή μη σημαντικό, τότε δεν υπάρχει διαφοροποίηση της επιβίωσης μεταξύ των κατηγοριών της μεταβλητής αυτής.

Ακολουθεί συγκεντρωτικός πίνακας με τα αποτελέσματα όσον αφορά τους δείκτες exp(b) και CI

Var.	competing risk	sign.	Exp(b)	95% Confidence Interval	
				Lower Bound	Upper Bound
Location	competing risk 2	.590	1.145	.695	1.896
Location	competing risk 3	.000	2.179	1.412	3.362
Location	competing risk 4	.119	.585	.299	1.142
Location	Όλες οι αιτίες	.008	1.465	1.104	1.944
Start	competing risk 2	.000	1.162	1.136	1.189
Start	competing risk 3	.000	1.035	1.022	1.048
Start	competing risk 4	.000	1.081	1.052	1.110
Start	Όλες οι αιτίες	.000	1.073	1.062	1.083
Activity 1	competing risk 2	.272	1.406	0.766	2.582
Activity 1	competing risk 3	.357	1.405	0.681	2.902
Activity 1	competing risk 4	.59	1.367	0.439	4.257
Activity 1	Όλες οι αιτίες	.180	1.340	0.873	2.057
Activity 2	competing risk 2	.323	1.32	0.761	2.293
Activity 2	competing risk 3	.713	0.87	0.413	1.83
Activity 2	competing risk 4	.248	1.863	0.649	5.351
Activity 2	Όλες οι αιτίες	.344	1.216	.811	1.822
Activity 3	competing risk 2	.014	1.803	1.125	2.889
Activity 3	competing risk 3	.000	4.059	2.317	7.111
Activity 3	competing risk 4	.372	1.547	0.593	4.037
Activity 3	Όλες οι αιτίες	.000	2.542	1.823	3.544
Activity 4	competing risk 2	.553	1.158	0.714	1.878
Activity 4	competing risk 3	.1	1.636	0.91	2.94
Activity 4	competing risk 4	.58	1.315	0.498	3.473
Activity 4	Όλες οι αιτίες	.119	1.317	0.931	1.863

Πίνακας 6.15 Συγκεντρωτικός πίνακας αποτελεσμάτων

Σύμφωνα με τα παραπάνω, συσχέτιση της μεταβλητής «Location» ανιχνεύεται μόνο στην περίπτωση που έχουμε competing risk 3, η μεταβλητή «Start» συσχετίζεται και με τις τρεις περιπτώσεις του competing risk, ενώ από το σύνολο των μεταβλητών «Activity» συσχέτιση υπάρχει στην περίπτωση της «Activity 3» με τις περιπτώσεις του «competing risk2» και «competing risk 3». Επίσης, το αποτέλεσμα «για όλες τις αιτίες μαζί» μπορεί να θεωρηθεί κατά κάποιο τρόπο ο μέσος όρος των επιμέρους εκτιμήσεων, που ελαχιστοποιεί τις διαφορές που υπάρχουν μεταξύ τους.

Η παραπάνω προσέγγιση στην ανάλυση των competing risks ονομάζεται “cause-specific hazards”, επειδή ουσιαστικά εξετάζουμε κάθε αιτία «θανάτου» ξεχωριστά. Δηλαδή, σε κάθε επιμέρους ανάλυση, οι θάνατοι από αιτίες άλλες από την υπό εξέταση, δε μετράνε ως θάνατοι αλλά ως αποκομμένες παρατηρήσεις (με χρόνο αποκοπής το χρόνο θανάτου από άλλη αιτία). Αυτή δεν αποτελεί τη μοναδική προσέγγιση που διαθέτει η στατιστική μεθοδολογία. Βλέπετε, για παράδειγμα, Putter et al. (2007).

6.4 Συμπεράσματα

Τα παραμετρικά μοντέλα PH είναι παρόμοια στην ερμηνεία με το μοντέλο Cox (PH). Η βασική διαφορά μεταξύ των δύο είναι ότι ο κίνδυνος υποτίθεται ότι ακολουθεί μια συγκεκριμένη στατιστική κατανομή όταν ένα πλήρως παραμετρικό μοντέλο PH είναι προσαρμοσμένο στα δεδομένα, ενώ το μοντέλο Cox δεν επιβάλλει κανένα τέτοιο περιορισμό. Εκτός από αυτό, οι δύο τύποι μοντέλων είναι ισοδύναμοι. Οι αναλογίες κινδύνου έχουν την ίδια ερμηνεία, είτε προέρχονται από ένα μοντέλο Cox είτε από ένα πλήρως παραμετρικό μοντέλο, ενώ η αναλογικότητα των κινδύνων εξακολουθεί να θεωρείται δεδομένη και στις δύο περιπτώσεις.

Μπορεί να προκύψει ένας αριθμός διαφορετικών παραμετρικών μοντέλων PH επιλέγοντας διαφορετικές κατανομές κινδύνου. Όπως έχει αναφερθεί σε προηγούμενες ενότητες, υπάρχει άμεση σχέση μεταξύ της συνάρτησης επιβίωσης και του κινδύνου και η επιλογή της κατανομής της συνάρτησης κινδύνου καθορίζει εκείνη της επιβίωσης. Στην πραγματικότητα, τα μοντέλα που χρησιμοποιούνται συνήθως, όπως τα μοντέλα Exponential, Weibull ή Gamma, παίρνουν τα ονόματά τους από την κατανομή που υποτίθεται ότι ακολουθούν για τους χρόνους επιβίωσης, αλλά και τα χαρακτηριστικά όσον αφορά τη συνάρτηση κινδύνου. Σύμφωνα με τα παραπάνω αποτελέσματα είτε ακολουθώντας παραμετρικό μοντέλο (KM) είτε μη παραμετρικό μοντέλο (Cox), υπάρχει συνέπεια τους πίνακες αποτελεσμάτων, εφόσον και στις δύο περιπτώσεις η δραστηριότητα της κάθε επιχείρησης είναι μεταβλητή προσδιορισμού της διάρκειας της. Στην περίπτωση του μοντέλου (KM), υπήρχαν διαφορές ανάμεσα στις δραστηριότητες στο μέσο όρο διάρκειας της λειτουργίας των επιχειρήσεων, αλλά όχι σε βαθμό όπου η δραστηριότητα που συγκέντρωνε τη μεγαλύτερη βαθμολογία, να επηρεάζει άμεσα και τη διάρκεια.

Βιβλιογραφία

- Καλαματιανού, Α. (2001). Δεδομένα διάρκειας ζωής, κατανομές & μέθοδοι ανάλυσης. Αθήνα: Εκδόσεις Παπαζήζη ΑΕΒΕ.
- Καρώνη, Χ. (2009). Μοντέλα αξιοπιστίας και επιβίωσης. Αθήνα: Εκδόσεις Συμεών.
- Beale, E. M. L., Kendall, M.G., & Mann, D. W. (1976). The Discarding of Variables in Multivariate Analysis. *Biometrika*, 54, 357-366.
- Bendell A., Walley M., Wightman D. W., Wood L. M., (1986). Proportional hazards modelling in reliability analysis - an application to brake discs on high speed trains. *Quality and Reliability Engineering International*, vol. 2, num. 1, p. 45-52.
- Bendell A., Wightman D. W., Walker E. V., (1991). Applying proportional hazards modelling in reliability. *Reliability Engineering and System Safety*, vol. 34, num. 1, p. 35-53.
- Caroni, C. (2002) The Correct “Ball Bearings” Data. *Lifetime Data Analysis*, 8, 395-399.
- Cutler, S.J., & Ederer, F. (1958). Maximum Utilization of the Life Table Method in Analyzing Survival. *Journal of Chronic Diseases*, 8, 699-712.
- Cox, D. R. (1972). Regression Model and Life Tables. *Journal of the Royal Statistical Society*. 34, 187-220.
- Cox, D. R. & Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman & Hall.
- Dale C. J., (1985). Application of the proportional hazards model in the reliability field. *Reliability Engineering*, vol. 10, num. 1, p. 1-14.
- Gehan, E. A. (1969). Estimating Survival Function from the Life Table. *Journal of Chronic Diseases*, 21, 629-644.
- Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457-481.
- Kleinbaum, G.D. (1995). *Survival analysis, A self- learning text*. NY: Springer
- Krivtsov V. V., Tananko D. E., Davis T. P., (2002). Regression approach to tire reliability analysis. *Reliability Engineering and System Safety*, vol. 78, num. 3, p. 267-273.
- Lawless, J. F. (1983). Statistical methods in reliability (with discussion), *Technometrics*, 25, 305-335.
- Lee, T.E.(1992). *Statistical methods for survival data analysis*. NY: John Wiley & Sons.
- Madeira S., Infante P., Didelet F., (2011). Use of survival models in a refinery. *Revstat*, vol. 11, num. 1, p. 45-65.

Parmar, M.K.B. & Machin, D. (1996). *Survival Analysis, A practical approach*. NY: John Wiley & Sons.

Putter, M. Fiocco & R.B. Geskus (2007) Tutorial in biostatistics: Competing risks and mult-state models. *Statistics in Medicine* 26, 2389-2430.

Mood, A.M., Graybill, F.N. & Boes, D.C. (1974). *Introduction to the theory of statistics*. McGraw-Hill.