



# ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

## ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

### ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ: ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΜΕΘΟΔΟΛΟΓΙΑΣ

Γεώργιος Β. Αρβανίτης

#### ΕΡΓΑΣΙΑ

Που υποβλήθηκε στο Τμήμα Στατιστικής  
του Οικονομικού Πανεπιστημίου Αθηνών  
ως μέρος των απαιτήσεων για την απόκτηση

Μεταπτυχιακού Διπλώματος Ειδίκευσης  
στη Στατιστική Μερικής Φοίτησης (Part-time)  
με κατεύθυνση «Εφαρμοσμένη Στατιστική για  
Εκπαιδευτικούς και Στελέχη Επιχειρήσεων  
& Οργανισμών»

Αθήνα  
Ιούνιος 2007



ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ  
ΚΑΤΑΛΟΓΟΣ





ΟΙΚΟΝΟΜΙΚΟ  
ΠΑΝΕΠΙΣΤΗΜΙΟ  
ΑΘΗΝΩΝ  
ΒΙΒΛΙΟΘΗΚΗ  
εισ. 81970  
Αρ.  
ταξ.

# ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

## ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

### ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ: ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΜΕΘΟΔΟΛΟΓΙΑΣ

ΓΕΩΡΓΙΟΣ Β. ΑΡΒΑΝΙΤΗΣ

#### ΕΡΓΑΣΙΑ

Που υποβλήθηκε στο Τμήμα Στατιστικής  
του Οικονομικού Πανεπιστημίου Αθηνών  
ως μέρος των απαιτήσεων για την απόκτηση  
Μεταπτυχιακού Διπλώματος  
Ειδίκευσης στη Στατιστική  
Μερικής Παρακολούθησης (Part-time)



Αθήνα  
Ιούνιος 2007





ΟΙΚΟΝΟΜΙΚΟ  
ΠΑΝΕΠΙΣΤΗΜΙΟ  
ΑΘΗΝΩΝ  
ΒΙΒΛΙΟΘΗΚΗ  
εισ. 812 fο  
Αρ.  
παξ.

# ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

## ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Εργασία που υποβλήθηκε ως μέρος των απαιτήσεων για την απόκτηση

Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική  
Μερικής Φοίτησης (Part-Time) με κατεύθυνση «Εφαρμοσμένη Στατιστική  
για Εκπαιδευτικούς και Στελέχη Επιχειρήσεων & Οργανισμών»

## ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ: ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΜΕΘΟΔΟΛΟΓΙΑΣ

Γεώργιος Β. Αρβανίτης

Υπεύθυνος μέλος ΔΕΠ:

Κ. Δημάκη

Αναπληρώτρια Καθηγήτρια

Ο Διευθυντής Μεταπτυχιακών Σπουδών

Επαμεινώνδας Πανάς

Καθηγητής



## **ΑΦΙΕΡΩΣΗ**

Στην οικογένεια μου, που μου συμπαραστάθηκε σε όλη τη διάρκεια του μεταπτυχιακού μου.

Την έδρανη απεριόριζην την επιβλαστική καθηγήσω της Αρι Αλεξανδρίνης Δημητρή, που μεν, εργασταίνει ως μεταβατικός μεν τον ανεπαντίσμαντη βιολογικότερο, αλλά απόλυτη γνώσης της και της αναδυόμενής της καριέρας την διάρκεια της πανεπιστημιακής λεγόμενης. Επίσης μπορεί να γίνεται τον καθηγητή της πανεπιστημιακής τρίτης της Επαναστατικής πατέρας μου, ονόματι Κωνσταντίνου Καραϊσκάκη, από τον οποίο παραγγέλλεται το παρόν έργο.



## ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια Δρ. Αικατερίνη Δημάκη για την εμπιστοσύνη που μου έδειξε αναθέτοντάς μου την συγκεκριμένη διπλωματική, την πολύτιμη βοήθειά της και τις συμβουλές της κατά την διάρκεια της παρούσης εργασίας. Επίσης είμαι ευγνώμων στο σύλλογο των καθηγητών του μεταπτυχιακού τμήματος της Στατιστικής γιατί μου έδωσε την δυνατότητα να συμμετάσχω στο μεταπτυχιακό του πρόγραμμα.





## ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

George B. Arvanitis

Γεννήθηκα στην Θεσσαλονίκη τον Μάρτιο του 1966. Το 1984 πέτυχα στο τμήμα Μαθηματικών του Πανεπιστημίου Αθηνών από όπου αποφοίτησα το 1991. Σήμερα εργάζομαι ως καθηγητής Μαθηματικών στον Ιδιωτικό Τομέα και στην Σχολή Εμπορικού Ναυτικού Ασπροπύργου. Τον Οκτώβριο του 2004 έγινα δεκτός στο Μεταπτυχιακό Πρόγραμμα της Συμπληρωματικής Ειδίκευσης στη Στατιστική. Τα ερευνητικά μου ενδιαφέροντα εστιάζονται στον τομέα της Ανάλυσης Δεδομένων Επιβίωσης και στην εφαρμογή τους μέσω διαφόρων στατιστικών προγραμμάτων.

The present of the author gives a brief of motivation for the presentation. Next all in this note we will focus on some important methods in the statistical analysis for the data in the survival analysis. This paper describes the methodology of survival analysis and its application in the medical statistics at the end of the article a new approach of self-adaptive k-means life, plays decisive role in many areas. Survival analysis and its applications are very useful in life and must be help our effort for the better. In the last section, the author will give a brief of analysis of survival modeling in medical, more focused on the medical field of econometrics and its applications.

In the first part of this note, the author will introduce the basic concepts and methods of survival analysis and its applications in medical survival. For this purpose, the author will start by defining the basic concepts of survival analysis and its applications. Next, the author will introduce the concept of survival function and its properties. Also, the author will introduce the concept of hazard rate and its properties. Then, the author will introduce the concept of survival modeling and its applications. Also, the author will introduce the concept of self-adaptive k-means life and its applications. Finally, the author will give a brief of analysis of survival modeling in medical, more focused on the medical field of econometrics and its applications.





## **ABSTRACT**

George B. Arvanitis

### **SURVIVAL ANALYSIS: SHORT DESCRIPTION OF METHODOLOGY**

June 2007

The duration of life constituted always pole of attraction for the scientists. Methods in the analysis of survival were presented not only in the statistical science but also in the medicine as well as in the social sciences. The repercussion of explanatory variables, which emanates and is used mainly in the medical statistics as well as in the sector of control the industrial and enterprising life, plays decisive role in the comprehension and analysis of quality of our life and thus it helps our effort for improvement or still for her elongation. Also, the applications of analysis of survival extending in very big field, from the field of physics until the field of econometrics. and economy.

The aim of present work is to present characteristically mathematic models and recent methods for the confrontation of problems of data of survival and for the shaping of suitable statistics conclusions with regard to the population of these data. More specifically, diplomatic is constituted by two parts. First is a import in the analysis of survival., with definition of survivor function and hazard rate function, with report in her types, continuous and distinguishable variables. Also becomes extensive report in the censored data and in the form of likelihood function in the case where they exist censored data. In the second part is analyzed the parameter estimate of survivor function, with examination of parameter models, Exponential and Weibull distributions, as well as comparison of two models





# ΠΕΡΙΛΗΨΗ

Γεώργιος Β. Αρβανίτης

## ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ: ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΜΕΘΟΔΟΛΟΓΙΑΣ

Ιούνιος 2007

Η διάρκεια ζωής αποτελούσε ανέκαθεν πόλο έλξης για τους επιστήμονες. Μέθοδοι στην ανάλυση επιβίωσης εμφανίστηκαν όχι μόνο στη στατιστική επιστήμη αλλά και στην ιατρική καθώς και στις κοινωνικές επιστήμες. Η επίπτωση των επεξηγηματικών μεταβλητών, οι οποίες προέρχονται και χρησιμοποιούνται κυρίως στην ιατρική στατιστική καθώς επίσης και στο κλάδο του ελέγχου της βιομηχανικής και επιχειρηματικής ζωής, παίζει καθοριστικό ρόλο στην κατανόηση και ανάλυση της ποιότητας της ζωής μας και έτσι βοηθάει την προσπάθεια μας για βελτίωση ή ακόμη για επιμήκυνσή της. Επίσης, οι εφαρμογές της ανάλυσης επιβίωσης εκτείνονται σε πολύ μεγάλο πεδίο, από το χώρο της φυσικής έως το χώρο της οικονομετρίας, και της οικονομίας..

Ο σκοπός της παρούσας εργασίας είναι να παρουσιάσει χαρακτηριστικά μαθηματικά μοντέλα και πρόσφατες μεθόδους για την αντιμετώπιση προβλημάτων δεδομένων επιβίωσης και για το σχηματισμό κατάλληλης στατιστικής συμπερασματολογίας σχετικά με τον πληθυσμό αυτών των δεδομένων. Ειδικότερα, η διπλωματική αποτελείται από δύο μέρη. Το πρώτο είναι μια εισαγωγή στην ανάλυση επιβίωσης., με ορισμό της συνάρτησης επιβίωσης και της συνάρτησης κινδύνου, με αναφορά στα είδη της, συνεχών και διακριτών μεταβλητών. Επίσης γίνεται εκτενής αναφορά στα λογοκριμένα δεδομένα και στην μορφή της συνάρτησης πιθανοφάνειας στην περίπτωση που υπάρχουν λογοκριμένα δεδομένα. Στο δεύτερο μέρος αναλύεται η παραμετρική εκτίμηση της συνάρτησης επιβίωσης, με ανασκόπηση των παραμετρικών μοντέλων, της εκθετικής και της Weibull κατανομής, καθώς και σύγκριση των δύο μοντέλων





# ΚΑΤΑΛΟΓΟΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

	Σελίδα
<b>1 Ανάλυση Επιβίωσης</b>	1
1.1 Εισαγωγή	1
1.2 Συνάρτηση επιβίωσης και συνάρτηση κινδύνου μίας συνεχούς τυχαίας μεταβλητής	2
1.3 Συνάρτηση επιβίωσης και συνάρτηση κινδύνου μίας διακριτής τυχαίας μεταβλητής	8
1.4 Βασικές μορφές της συνάρτησης κινδύνου	15
1.5 Λογοκριμένα ή περικομμένα δεδομένα	18
1.5.1. Δεξιά λογοκρισία τύπου I (type I right censoring)	22
1.5.2. Δεξιά λογοκρισία τύπου II (type II right censoring)	28
1.5.3. Τυχαία λογοκρισία (random censoring)	29
1.5.4. Αριστερή λογοκρισία, λογοκρισία σε διάστημα, περικομμένα δεδομένα (left censoring, interval censoring, truncated data)	31
1.6 Συνάρτηση πιθανοφάνειας και λογοκριμένα δεδομένα	33
1.6.1. Τυχαία λογοκρισία	33
1.6.2. Δεξιά λογοκρισία τύπου I	35
1.6.3. Δεξιά λογοκρισία τύπου II	37
1.6.4. Αριστερή λογοκρισία, λογοκρισία σε διάστημα, περικοπή	37
<b>2 Παραμετρική Εκτίμηση Της Συνάρτησης Επιβίωσης</b>	47
2.1 Εισαγωγή	47
2.2 Ανασκόπηση παραμετρικών μοντέλων	48
2.2.1. Εκθετική (exponential) κατανομή	48
2.2.2. Κατανομή Weibull	53
2.3 Γενικές Παρατηρήσεις	63

KATAVOLIOΣ ΕΛΛΗΝΩΝ

ΕΛΛΑΣ

X



## **ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ**

<b>Πίνακας</b>	<b>Σελίδα</b>
1.2 Σύνδεση μεταξύ των συναρτήσεων $f(t)$ , $F(t)$ , $S(t)$ , $h(t)$ , $H(t)$	6
1.5.1 Δεδομένα του παραδείγματος 1.5.1	19
1.5.2 Πίνακας ανάλυσης των δεδομένων μέσω SPSS	20
1.5.3 Πίνακας μέσου και διαμέσου για το χρόνο επιβίωσης μέσω SPSS	20
1.5.4 Πληροφορίες για το χρόνο ζωής κάθε εξαρτήματος	25
1.5.5 Δεδομένα σε μορφή πίνακα	25
1.5.6 Πίνακας ανάλυσης των δεδομένων μέσω SPSS	26
1.5.7 Πίνακας μέσου και διαμέσου για το χρόνο επιβίωσης μέσω SPSS	26



## ΚΑΤΑΛΟΓΟΣ ΓΡΑΦΗΜΑΤΩΝ

Γράφημα	Σελίδα
1.2.1 Διαγράμματα για τις συναρτήσεις $f(t)$ , $F(t)$ , $S(t)$ , $h(t)$ , $H(t)$	7-8
1.3.1 Γραφήματα των δεδομένων μέσω των συναρτήσεων $f(t)$ , $F(t)$ , $S(t)$ , $h(t)$ , $H(t)$	13-14
1.4.1 Τα πιο συνηθισμένα είδη μορφών συναρτήσεων κινδύνου	16
1.5.1 Διαγράμματα των συναρτήσεων επιβίωσης και κίνδυνου	21
1.5.2 Απεικόνιση πληροφοριών για τους χρόνους ανάπτυξης του όγκου	23
1.5.3 Διάγραμμα της συνάρτησης επιβίωσης	27
1.5.4 Πληροφορίες για τους χρόνους επανεμφάνισης της οξείας λευχαιμίας	30
1.6.1 Διάγραμμα του εκθετικού μοντέλου $g(y)$	42
2.2.1 Απεικόνιση των συναρτήσεων πυκνότητας $f(t)$ , επιβίωσης $S(t)$ , κινδύνου $h(t)$ της εκθετικής κατανομής με παράμετρο $\lambda$	49
2.2.2 Απεικόνιση της συνάρτησης πυκνότητας $f(t)$ της κατανομής <i>Weibull</i> με παραμέτρο $\lambda$ , για διάφορες τιμές της $\alpha$ (παράμετρος μορφής)	56
2.2.3 Απεικόνιση της συνάρτησης πυκνότητας $S(t)$ της κατανομής <i>Weibull</i> με παραμέτρο $\lambda$ , για διάφορες τιμές της $\alpha$ (παράμετρος μορφής)	57
2.2.4 Απεικόνιση της συνάρτησης πυκνότητας $h(t)$ της κατανομής <i>Weibull</i> με παραμέτρο $\lambda$ , για διάφορες τιμές της $\alpha$ (παράμετρος μορφής)	58
2.3.1 $-\log(\hat{S}(t_i))$ vs $t_i$	66
2.3.2 $\log(-\log(\hat{S}(t_i)))$ vs $\log t_i$	67



# ΚΕΦΑΛΑΙΟ 1<sup>ο</sup>

## **ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ**

### **1.1 Εισαγωγή**

Η ανάλυση επιβίωσης (*survival analysis*) ή ανάλυση χρόνων αποτυχίας (*analysis of failuretime data*) είναι μία συλλογή στατιστικών μεθόδων ανάλυσης δεδομένων τα οποία προκύπτουν ως τιμές μίας μεταβλητής που δηλώνει το χρόνο μέχρις ότου συμβεί κάποιο ενδεχόμενο (αποτυχία). Με τον όρο χρόνο επιβίωσης (*survival time*) ή χρόνο ζωής (*lifetime*) ή χρόνο αποτυχίας (*failuretime*) δηλώνουμε το χρόνο (ημέρες, εβδομάδες, μήνες κτλ.) που μεσολαβεί από τη χρονική στιγμή παρακολούθησης ενός ατόμου (άνθρωπος, αντικείμενο, φαινόμενο κτλ.) μέχρι τη στιγμή που το άτομο θα αντιμετωπίσει το ενδεχόμενο. Χαρακτηριστικά παραδείγματα τέτοιων χρόνων είναι ο χρόνος ζωής ενός ανθρώπου (το ενδεχόμενο είναι ο θάνατος), ο χρόνος επανεμφάνισης των συμπτωμάτων μίας νόσου, ο χρόνος από την έναρξη μίας θεραπείας ως την απόκριση σε αυτή, ο χρόνος ζωής μιας ηλεκτρονικής συσκευής (το ενδεχόμενο είναι η βλάβη της συσκευής), κτλ. Εάν μπορούμε να αναπτύξουμε τα ακριβή αναλυτικά πρότυπα για τη διάρκεια ζωής ατόμων ή αντικειμένων μέσα από προγράμματα, θα έχουμε γρηγορότερη και πιο λεπτομερή εξερεύνηση των τεχνικών διαχείρισης μνήμης.

Για την περιγραφή της κατανομής του χρόνου ζωής  $T$  ενός ατόμου θα χρησιμοποιήσουμε μη αρνητικές τυχαίες μεταβλητές. Αν συμβολίσουμε με  $R_T$  το σύνολο τιμών της  $T(R_T \subseteq [0, \infty))$  και με την αθροιστική συνάρτηση κατανομής της  $F_T$ , τότε  $F(t) = P(T \leq t)$ ,  $t \in R$  και  $F$  αύξουσα συνάρτηση και συνεχής από δεξιά.

- αν η  $T$  είναι μια μη αρνητική συνεχής τυχαία μεταβλητή με συνάρτηση πυκνότητας  $f(t)$  και σύνολο τιμών  $R(t) = [0, \infty)$ , τότε  $F(t) = 0$  για  $t < 0$



$$F(t) = \int_0^t f(u)du, \quad t \geq 0$$

και

$$f(t) = F'(t) = \lim_{\Delta t \rightarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

➤ αν η  $T$  είναι μια μη αρνητική διακριτή τυχαία μεταβλητή με συνάρτηση πιθανότητας  $f(t)$  και σύνολο τιμών  $R(t) = \{t_1, t_2, t_3, \dots\}$  με  $0 < t_1$ , τότε  $F(t) = 0$  με  $t < t_1$ .

$$F(t) = \sum_{j: t_j \leq t} f(t_j), \quad t \geq t_1 \quad \text{και} \quad f(t_j) = F(t_j) - F(t_{j-1}), \quad j = 1, 2, \dots, \quad F(t_0) = 0$$

Η  $f(t)$  και η  $F(t)$  καθορίζει πλήρως τη συμπεριφορά της τυχαίας μεταβλητής  $T$ . Άλλες ποσότητες που καθορίζουν τη συμπεριφορά της  $T$  είναι η συνάρτηση επιβίωσης και η συνάρτηση κινδύνου της τυχαίας μεταβλητής  $T$ .

## 1.2 Συνάρτηση επιβίωσης και συνάρτηση κινδύνου μιας συνεχούς τυχαίας μεταβλητής

Έστω  $T$  μια συνεχής τυχαία μεταβλητή με σύνολο τιμών  $R(t) = [0, \infty)$  η οποία δηλώνει το χρόνο ζωής ενός ατόμου. Η πιθανότητα  $S(t)$  του ενδεχομένου  $\{T > t\}$ ,  $t \geq 0$ , ονομάζεται συνάρτηση επιβίωσης (*survival function*) της τυχαίας μεταβλητής  $T$  και ορίζεται με τον τύπο:  $S(t) = P(T > t)$ ,  $t \geq 0$ . Η συνάρτηση επιβίωσης  $S(t)$  δηλώνει την πιθανότητα να είναι ο χρόνος ζωής ενός ατόμου μεγαλύτερος του χρόνου  $t$  ή πιο απλά η πιθανότητα να φθάσει το άτομο την ηλικία  $t$ . Στα πλαίσια της θεωρίας αξιοπιστίας η συνάρτηση  $S(t)$  είναι γνωστή ως συνάρτηση αξιοπιστίας (*reliability function*). Ο Hayes εισήγαγε την διάκριση μεταξύ «αδύνατου» και «ισχυρού» γενικά για τα δεδομένα μίας έρευνας. Η επεξήγηση ότι η αδύνατη υπόθεση είναι: «τα πρόσφατα δημιουργημένα αντικείμενα έχουν πολύ υψηλότερη θηλυμότητα από τα αντικείμενα που είναι

παλαιότερα». Επίσης διατύπωσε την ισχυρή υπόθεση (που εισάγει στην πραγματικότητα): «*ακόμα κι αν τα αντικείμενα εν λόγω δεν δημιουργούνται πρόσφατα, τα νεώτερα αντικείμενα έχουν μια υψηλότερη θνησιμότητα από τα σχετικά παλαιότερα αντικείμενα*». Η μελλοντική αναμενόμενη διάρκεια ζωής ενός αντικειμένου είναι ανάλογη με την τρέχουσα ηλικία του. Κατά συνέπεια, η προηγούμενη διάρκεια ζωής είναι ισχυρός προάγγελος της μελλοντικής (υπόλοιπης) διάρκειας ζωής.

Η συνάρτηση  $S(t)$  συνδέεται με τη συνάρτηση κατανομής  $F(T)$  της  $T$  σύμφωνα με την σχέση

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

Η συνάρτηση  $S(t)$  είναι φθίνουσα συνάρτηση, συνεχής,  $S(0)=1$  και  $\lim_{t \rightarrow \infty} S(t) = 0$ , με δεδομένο  $F(0) = 0$  και  $\lim_{t \rightarrow \infty} F(t) = 1$ . Επίσης έχουμε

$$S(t) = \int f(u)du$$

όπου  $f(t)$  είναι η συνάρτηση πυκνότητας της τυχαίας μεταβλητής  $T$ . Για συνεχείς τυχαίες μεταβλητές ισχύει ότι  $S(t) = P(T > t) = P(T \geq t)$ ,  $t \geq 0$ , αλλά για διακριτές τυχαίες μεταβλητές δεν ισχύει.

Η συνάρτηση  $f(t)$  συνδέεται με την  $S(t)$  σύμφωνα με τη σχέση  $f(t) = -S'(t)$ , αφού

$$f(t) = \frac{dF(t)}{dt} = \frac{d(1 - S(t))}{dt} = -\frac{dS(t)}{dt} = -S'(t)$$

Μια άλλη βασική ποσότητα στην ανάλυση επιβίωσης είναι η συνάρτηση κινδύνου (*hazard function* ή *hazard rate*) της τυχαίας μεταβλητής  $T$  η οποία συμβολίζεται με  $h(t)$

και ορίζεται με τον τύπο  $h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < T + \Delta t / T \geq t)}{\Delta t}$ ,  $t \geq 0$

Η συνάρτηση κινδύνου  $h(t)$  δηλώνει τη “στιγμαία πιθανότητα θανάτου” ενός ατόμου το χρόνο  $t$  δοθέντος ότι αυτό επέζησε μέχρι τη χρονική στιγμή  $t$ . Η ποσότητα  $h(t)\Delta t$ ,  $\Delta t$

μικρό, είναι προσεγγιστικά η πιθανότητα θανάτου ενός ατόμου στο διάστημα  $[t, t + \Delta t]$  γνωρίζοντας ότι το άτομο έχει επιβιώσει μέχρι τη χρονική στιγμή  $t$ , δηλαδή

$$h(t)\Delta t \approx P(t \leq T < t + \Delta t / T \geq t)$$

Όταν η  $h(t)$  είναι αύξουσα συνάρτηση του χρόνου τότε και η πιθανότητα  $P(t \leq T < t + \Delta t / T \geq t)$  είναι αύξουσα συνάρτηση του χρόνου. Σε αυτή την περίπτωση το άτομο “γερνά” με την πάροδο του χρόνου και λέμε ότι ο χρόνος ζωής του  $T$  έχει την ιδιότητα *IFR* (*increasing failure rate*). Στην αντίθετη περίπτωση ( $h(t)$  φθίνουσα συνάρτηση του χρόνου) το άτομο “βελτιώνεται” με την πάροδο του χρόνου και λέμε ότι ο χρόνος ζωής του  $T$  έχει την ιδιότητα *DFR* (*decreasing failure rate*).

Η συνάρτηση  $h(t)$  είναι γνωστή στην θεωρία αξιοπιστίας ως βαθμίδα αποτυχίας (*failure rate*), στη δημογραφία ως ένταση θνησιμότητας (*force of mortality*), στα οικονομικά ως αντίστροφος λόγος του Mill (*inverse Mill's ratio*), στη θεωρία ακραίων τιμών ως συνάρτηση έντασης (*intensity function*).

Η  $h(t)$  ικανοποιεί τη σχέση  $h(t) = \frac{f(t)}{S(t)}$  αφού

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t / T \geq t)}{\Delta t} = \frac{1}{P(T \geq t)} \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} = \frac{f(t)}{S(t)}$$

και συνεπώς  $h(t) \geq 0$ ,  $t \geq 0$ . Επειδή  $f(t) = -S'(t)$  προκύπτει άμεσα ότι η  $h(t)$  συνδέεται

με την  $S(t)$  σύμφωνα με τη σχέση  $h(t) = -\frac{S'(t)}{S(t)} = -\frac{d \log(S(t))}{dt}$  (1)

Ολοκληρώνοντας τα δύο μέλη της σχέσης (1) ως προς  $t$  και χρησιμοποιώντας τη συνθήκη  $S(0)=1$  προκύπτει ότι

$$\int h(u)du = - \int \frac{d \log(S(u))}{du} du = - \log(S(t))$$

άρα

$$S(t) = \exp(-\int_0^t h(u)du)$$

Επίσης, χρησιμοποιώντας τη σχέση  $f(t) = h(t)S(t)$ , παίρνουμε

$$f(t) = h(t) \exp(-\int_0^t h(u)du)$$

Η συνάρτηση  $H(t)$  που ορίζεται με τον τύπο

$$H(t) = \int_0^t h(u)du$$

ονομάζεται αθροιστική συνάρτηση κινδύνου (*cumulative hazard function*), και άρα έχουμε

$$S(t) = e^{-H(t)} , \quad H(t) = -\log(S(t))$$

Επειδή  $0 \leq S(t) \leq 1$  προκύπτει ότι  $0 \leq H(t) < \infty$ . Επίσης, με δεδομένο ότι  $\lim_{t \rightarrow \infty} S(t) = 0$  προκύπτει ότι  $\lim_{t \rightarrow \infty} H(t) = \infty$ . Έτσι για τη συνάρτηση  $h(t)$  προκύπτει η ιδιότητα

$$\int_0^\infty h(t)dt = \lim_{t \rightarrow \infty} H(t) = \infty$$

Η γνώση μόνο μιας από τις ποσότητες  $f(t)$  ,  $F(t)$  ,  $S(t)$  ,  $h(t)$  ,  $H(t)$  αρκεί για την εύρεση των υπολοίπων τεσσάρων. Κατασκευάζω ένα πίνακα για να απεικονίσω την σύνδεση μεταξύ των ποσοτήτων  $f(t)$  ,  $F(t)$  ,  $S(t)$  ,  $h(t)$  ,  $H(t)$ . Οι ποσότητες στην πρώτη στήλη δίνονται συναρτήσει των ποσοτήτων των άλλων στηλών

	$f(t)$	$F(t)$	$S(t)$	$h(t)$	$H(t)$
$f(t)$	-	$F'(t)$	$-S'(t)$	$h(t) \exp(-\int h(u)du)$	$H'(t) e^{-H(t)}$
$F(t)$	$\int f(u)du$	-	$1-S(t)$	$1 - \exp(-\int h(u)du)$	$1 - e^{-H(t)}$
$S(t)$	$\int f(u)du$	$1-F(t)$	-	$\exp(-\int h(u)du)$	$e^{-H(t)}$
$h(t)$	$\frac{f(t)}{\int f(u)du}$	$\frac{F'(t)}{1-F(t)}$	$-\log(S(t))'$	-	$H(t)$
$H(t)$	$-\log(\int f(u)du)$	$-\log(1-F(t))$	$-\log(S(t))$	$\int h(u)du$	-

**Πίνακας 1.2 Σύνδεση μεταξύ των συναρτήσεων  $f(t)$ ,  $F(t)$ ,  $S(t)$ ,  $h(t)$ ,  $H(t)$**

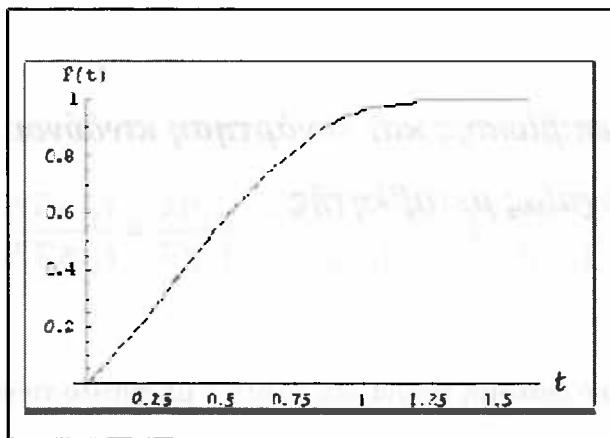
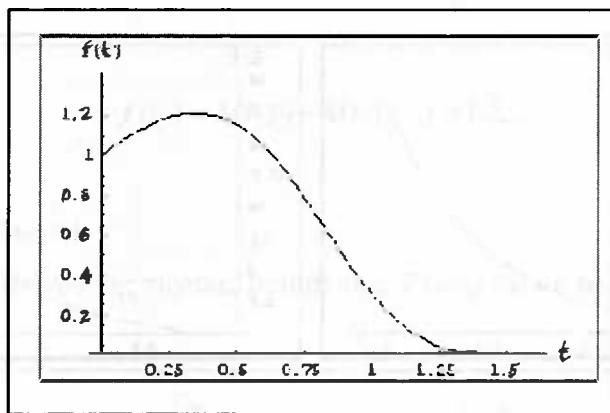
### **ΠΑΡΑΔΕΙΓΜΑ 1.2.1**

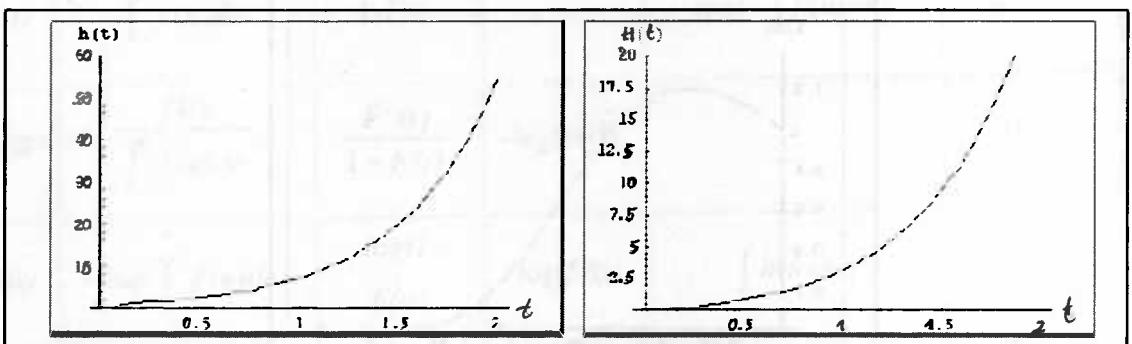
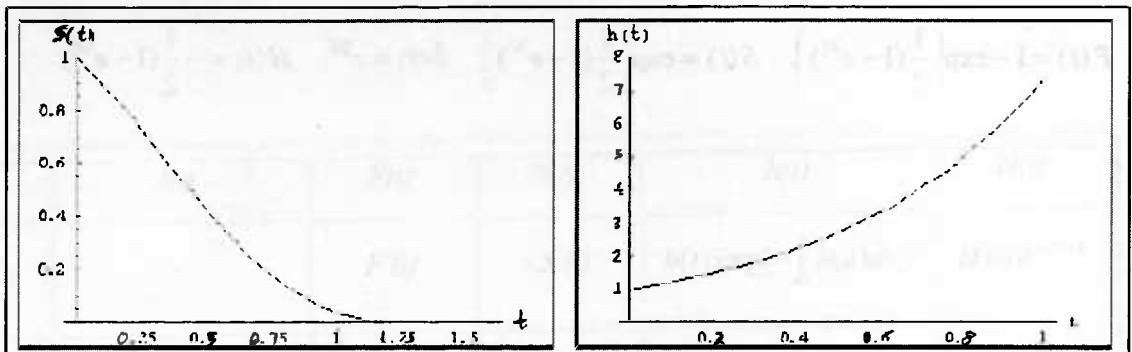
Ας θεωρήσουμε τη συνεχή τυχαία μεταβλητή  $T$  με συνάρτηση πυκνότητας πιθανότητας που δίνεται από τον τύπο

$$f(t) = \exp\left(2t + \frac{1}{2}(1-e^{2t})\right), \quad t \geq 0$$

άρα

$$F(t) = 1 - \exp\left(-\frac{1}{2}(1-e^{2t})\right), \quad S(t) = \exp\left(-\frac{1}{2}(1-e^{2t})\right), \quad h(t) = e^{2t}, \quad H(t) = -\frac{1}{2}(1-e^{2t})$$





**Διάγραμμα 1.2.1** Διαγράμματα για τις συναρτήσεις  $f(t)$ ,  $F(t)$ ,  $S(t)$ ,  $h(t)$ ,  $H(t)$

### 1.3 Συνάρτηση επιβίωσης και συνάρτηση κινδύνου μιας διακριτής τυχαίας μεταβλητής

Έστω μια μη αρνητική διακριτή τυχαία μεταβλητή  $T$  με σύνολο τιμών  $R_T = \{t_1, t_2, t_3, \dots\}$ ,  $0 < t_1 < t_2 < \dots$ , συνάρτηση πιθανότητας  $f(t)$ , και συνάρτηση κατανομής  $F(t)$ . Η συνάρτηση επιβίωσης  $S(t)$  της  $T$  ορίζεται όπως στη συνεχή περίπτωση, δηλαδή από τον τύπο

$$S(t) = P(T > t) = \sum_{t_j > t} f(t_j), \quad t \geq 0.$$

Παρατηρούμε ότι η συνάρτηση επιβίωσης είναι φθίνουσα, συνεχής από δεξιά,  $S(0)=1$  και  $\lim_{t \rightarrow \infty} S(t) = S(\infty) = 0$ . Ισχύει

$$S(t) = 1, \quad 0 \leq t < t_1,$$

$$S(t_j) = f(t_{j-1}) + f(t_{j-2}) + \dots, \quad j = 1, 2, \dots$$

Και

$$f(t_j) = S(t_{j-1}) - S(t_j), \quad j = 1, 2, \dots$$

με δεδομένο ότι  $S(t_0) = 1$ .

Η συνάρτηση κινδύνου της τυχαίας μεταβλητής  $T$  ορίζεται με τον τύπο

$$h(t) = \begin{cases} 0, & t \neq t_j \\ P(T = t_j | T \geq t_j), & t = t_j \end{cases}$$

οπότε

$$h(t_j) = \frac{P(T = t_j)}{P(T \geq t_j)} = \frac{f(t_j)}{S(t_{j-1})} = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})}, \quad j = 1, 2, \dots$$

και

$$1 - h(t_j) = \frac{S(t_j)}{S(t_{j-1})}, \quad j = 1, 2, \dots.$$

Η συνάρτηση επιβίωσης και η συνάρτηση κινδύνου της τυχαίας μεταβλητής  $T$  συνδέονται με τη σχέση

$$S(t) = \prod_{j:t_j \leq t} [1 - h(t_j)], \quad t \geq 0$$

όπου για το παραπάνω γινόμενο ορίζουμε ότι  $S(t) = 1$ ,  $0 \leq t < t_1$

Επίσης ορίζουμε ότι  $h(t_0) = 0$  με  $0 < t_0 < t_1$ , έχουμε ότι για  $0 \leq t < t_1$

$$S(t) = \prod_{j:t_j \leq t} [1 - h(t_j)] = 1 - h(t_0) = 1$$

Για την απόδειξη της παραπάνω σχέσης ας θεωρήσουμε ότι  $t_j \leq t < t_{j+1}$ ,  $j=1,2,\dots$ . Τότε

$$\begin{aligned} S(t) &= P(T > t_j) \\ &= P(T > t_1) \cdot \frac{P(T > t_2)}{P(T > t_1)} \cdot \frac{P(T > t_3)}{P(T > t_2)} \cdots \frac{P(T > t_j)}{P(T > t_{j-1})} \\ &= P(T > t_1) \cdot P(T > t_2 | T > t_1) \cdot P(T > t_3 | T > t_2) \cdots P(T > t_j | T > t_{j-1}) \\ &= S(t_1) \cdot \frac{S(t_2)}{S(t_1)} \cdot \frac{S(t_3)}{S(t_2)} \cdots \frac{S(t_j)}{S(t_{j-1})} \\ &= \prod_{j:t_j \leq t} \frac{S(t_j)}{S(t_{j-1})} = \prod_{j:t_j \leq t} [1 - h(t_j)]. \end{aligned}$$

Επίσης με τη σχέση  $H(t) = -\log S(t)$  που ισχύει στη συνεχή περίπτωση και στη διακριτή περίπτωση με τον ίδιο τύπο

$$\log(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \dots, \quad -1 \leq x < 1.$$

Αρα

$$H(t) = -\log S(t) = -\log(\prod_{j:t_j \leq t} [1-h(t_j)]) = -\sum_{j:t_j \leq t} \log[1-h(t_j)] = \sum_{j:t_j \leq t} \left| h(t_j) + \frac{[h(t_j)]^2}{2} + \frac{[h(t_j)]^3}{3} + \dots \right|$$

και μια προσέγγιση πρώτης τάξης για την  $H(t)$  είναι

$$H(t) = \sum_{j:t_j \leq t} h(t_j)$$

Ο παραπάνω τύπος της  $H(t)$  είναι ένας άλλος ορισμός της αθροιστικής συνάρτησης κινδύνου  $H(t)$  στη διακριτή περίπτωση

$$H(t) = \int_0^t h(u) du$$

και χρησιμοποιείται πιο συχνά στην πράξη.

### ΠΑΡΑΔΕΙΓΜΑ 1.3.1

Ας θεωρήσουμε τη διακριτή τυχαία μεταβλητή  $T$  με συνάρτηση πιθανότητας που δίνεται από τον τύπο

$$f(t_j) = f(j) = 0.25, \quad j = 1, 2, 3, 4.$$

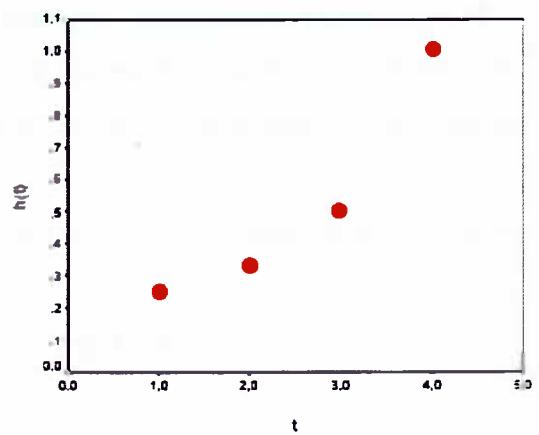
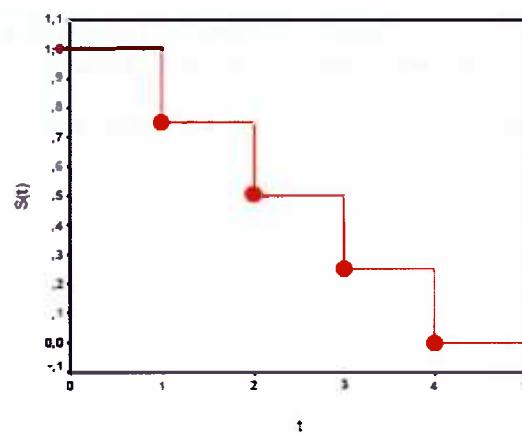
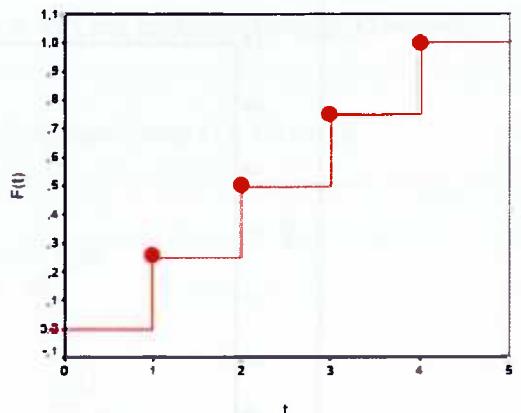
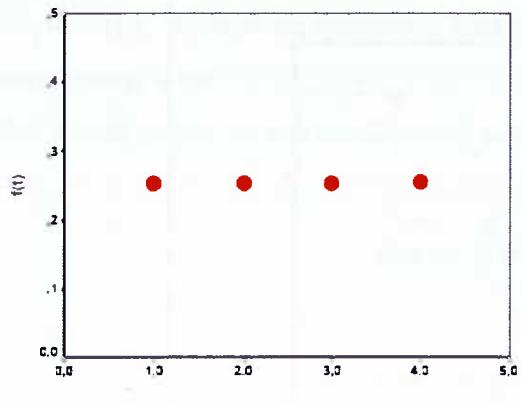
Τότε

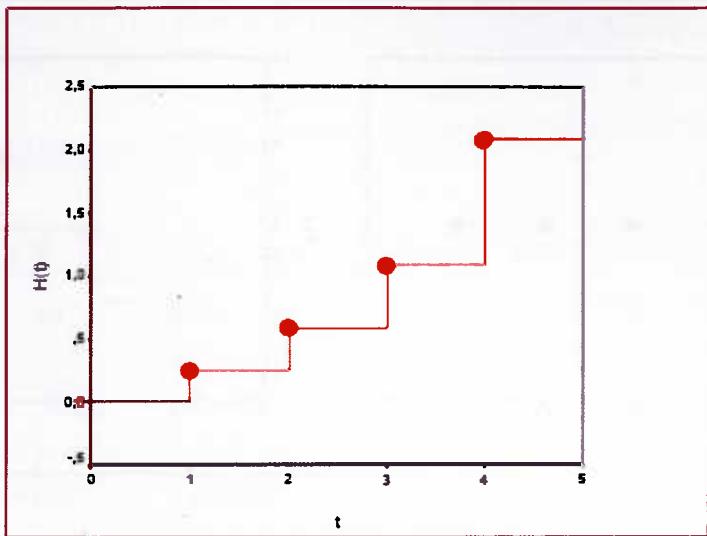
$$F(t) = \begin{cases} 0, & t < 1 \\ 0.25, & 1 \leq t < 2 \\ 0.5, & 2 \leq t < 3 \\ 0.75, & 3 \leq t < 4 \\ 1, & t \geq 4 \end{cases}$$

$$S(t) = \begin{cases} 1, & t < 1 \\ 0.75, & 1 \leq t < 2 \\ 0.5, & 2 \leq t < 3 \\ 0.25, & 3 \leq t < 4 \\ 0, & t \geq 4 \end{cases}$$

$$h(t) = \begin{cases} 0, & t \neq 1, 2, 3, 4 \\ 0.25, & t = 1 \\ 0.3, & t = 2 \\ 0.5, & t = 3 \\ 1, & t = 4 \end{cases}$$

$$H(t) = \begin{cases} 0, & t < 1 \\ 0.25, & 1 \leq t < 2 \\ 0.58\bar{3}, & 2 \leq t < 3 \\ 1.08\bar{3}, & 3 \leq t < 4 \\ 2.08\bar{3}, & t \geq 4 \end{cases}$$





**Διάγραμμα 1.3.1 Γραφήματα των δεδομένων μέσω των συναρτήσεων  $f(t)$ ,  $F(t)$ ,  $S(t)$ ,  $h(t)$ ,  $H(t)$**

Η «μελλοντική αναμενόμενη διάρκεια ζωής  $C(x)$ » ενός αντικειμένου είναι η διαφορά μεταξύ της αναμενόμενης διάρκειας ζωής και της τρέχουσας ηλικία του. Η μελλοντική αναμενόμενη διάρκεια ζωής ενός αντικειμένου είναι ανάλογη στην τρέχουσα ηλικία του (1).

Η αναμενόμενη διάρκεια ζωής με τυχαία μεταβλητή επιβίωσης  $L$  υπολογίζεται ως εξής:

$$\begin{aligned} E[L \mid L \geq x] &= \frac{\int_0^\infty t f(t \mid t \geq x) dt}{\int_x^\infty f(t) dt} \\ &= \frac{\int_x^\infty t f(t) dt}{\int_x^\infty f(t) dt} \\ &= \frac{\int_x^\infty t f(t) dt}{s(x)}. \end{aligned}$$

όπου το  $x$  είναι η τρέχουσα ηλικία. Κατά συνέπεια

$$C(x) = E[L \mid L \geq x] - x$$

άρα η πρόταση (1) παίρνει την μορφή: ( $\exists \psi > 0, \forall x \geq 0$ )  $C(x) = \psi \cdot x$  και  $\psi$  : σταθερά αναλογικότητας μεταξύ της τρέχουσας ηλικίας  $x$  και της αναμενόμενης μελλοντικής διάρκειας ζωής  $C(x)$

Θα βρούμε τη χρήση για μια εναλλακτική μορφή της πρότασης (1): Θέτουμε

$$G(x) = \int_0^x tf(t)dt, \text{ ára}$$

$$E[L | L \geq x] = \frac{G(x)}{s(x)}$$

Επίσης θέτω:

$$\Lambda(x) = \frac{G(x)}{xs(x)} = \frac{E[L | L \geq x]}{x} = \frac{C(x) + x}{x} = \frac{C(x)}{x} + 1$$

και η πόταση (1) παίρνει την μορφή:

$$(\exists \psi' > 1)(\forall x \geq 0)\Lambda(x) = \psi'$$

με δεδομένο:

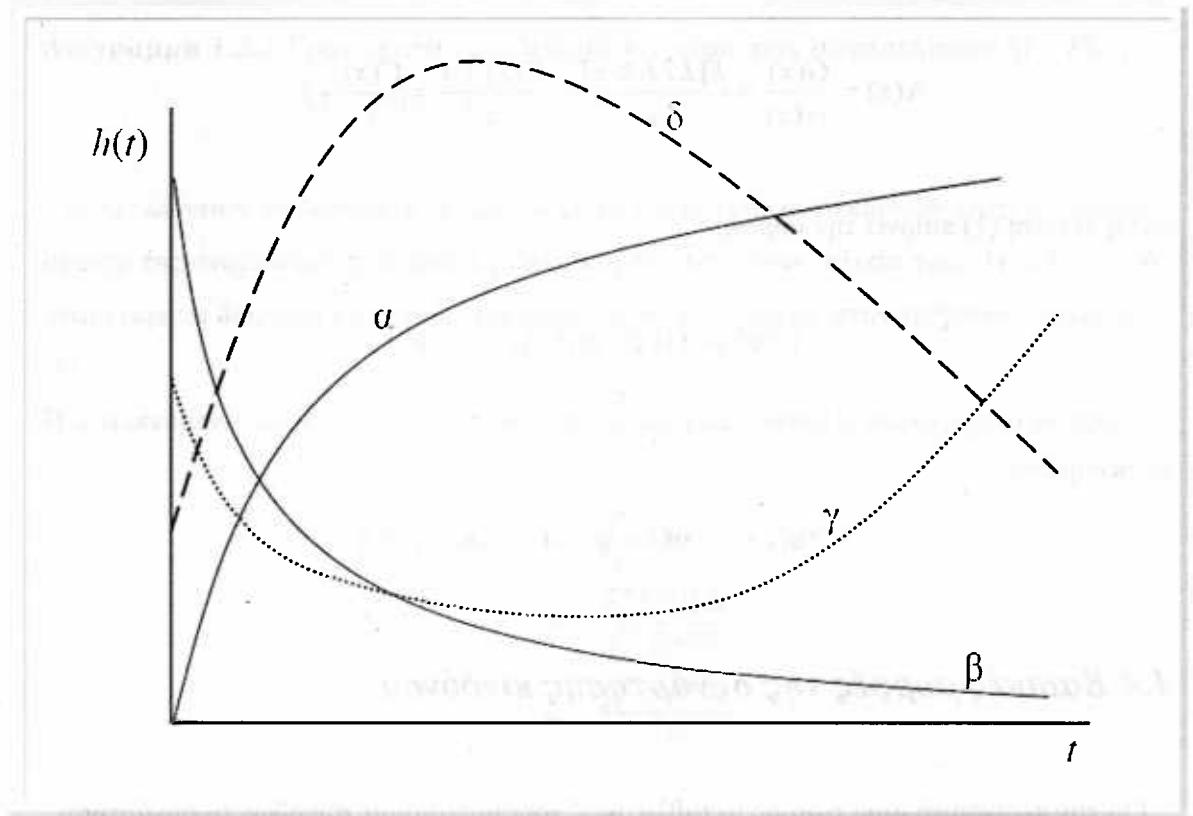
$$\psi' = \psi + 1$$

## 1.4 Βασικές μορφές της συνάρτησης κινδύνου

Για την περιγραφή μιας τυχαίας μεταβλητής  $T$  χρησιμοποιούμε συνήθως τη συνάρτηση πυκνότητάς της ή τη συνάρτηση πιθανότητάς της  $f(t)$  (ανάλογα με το αν είναι συνεχής ή διακριτή) ή ισοδύναμα τη συνάρτηση κατανομής της  $F(t)$ , ή ισοδύναμα τη συνάρτηση επιβίωσής της  $S(t)$ . Η συνάρτηση κινδύνου  $h(t)$  της  $T$ , αν και δεν χρησιμοποιείται συχνά, είναι ιδιαιτέρως χρήσιμη για την περιγραφή κατανομών χρόνου ζωής (*lifetime*

*distributions*) αφού δηλώνει τον τρόπο με τον οποίο μεταβάλλεται η “στιγμιαία πιθανότητα θανάτου” ενός ατόμου συναρτήσει του χρόνου. Σε πολλές εφαρμογές μπορεί να υπάρχουν ποιοτικές πληροφορίες για τη συνάρτηση κινδύνου  $h(t)$  της  $T$  οι οποίες μπορούν να μας βοηθήσουν αρχικά στην αναγνώριση και τελικά στην επιλογή του κατάλληλου παραμετρικού μοντέλου για την περιγραφή της  $T$ .

Στο ακόλουθο διάγραμμα δίνονται τέσσερα βασικά είδη μορφών συναρτήσεων κινδύνου που παρατηρούνται στην πράξη.



**Διάγραμμα 1.4.1** Τα πιο συνηθισμένα είδη μορφών συναρτήσεων κινδύνου

Η κατανομή (α) έχει αύξουσα συνάρτηση κινδύνου (*IFR*), η (β) έχει φθίνουσα συνάρτηση κινδύνου (*DFR*). η (γ) έχει συνάρτηση κινδύνου “λεκανοειδούς” μορφής

(*bathtub-shaped*), ενώ η (δ) έχει συνάρτηση κινδύνου μορφής “καμπούρας” (*hump-shaped*).

Αύξουσες συναρτήσεις κινδύνου (μορφή (α)) εμφανίζονται πολύ συχνά στην πράξη. Τυπικό παράδειγμα αποτελεί η περίπτωση όπου εμφανίζεται φυσική γήρανση (*aging*) ή φθορά (*wear out*) των υπό μελέτη ατόμων με την πάροδο του χρόνου (υπολειπόμενος χρόνος ζωής ενός ατόμου ηλικίας 50 ετών, υπολειπόμενος χρόνος ζωής μιας ηλεκτρονικής συσκευής η οποία ήδη λειτουργεί για ένα χρόνο). Φθίνουσες συναρτήσεις κινδύνου (μορφή (β)) εμφανίζονται σπάνια και αφορούν περιπτώσεις όπου υπάρχει αυξημένη πιθανότητα αποτυχίας σε πρώιμα στάδια δηλαδή υπάρχει βελτίωση με την πάροδο του χρόνου (ο χρόνος λειτουργίας ορισμένου τύπου ηλεκτρονικών συσκευών εμφανίζεται να έχει φθίνουσα συνάρτηση κινδύνου κατά την αρχική περίοδο χρήστης των, όπως και η αντοχή ορισμένων υλικών στο χρόνο (μπετόν, μέταλλα)).

Συναρτήσεις κινδύνου λεκανοειδούς μορφής (μορφή (γ)) προκύπτουν όταν μελετούμε το χρόνο λειτουργίας μιας ολοκαίνουργιας συσκευής ή το χρόνο ζωής ενός ανθρώπου από τη στιγμή της γεννήσεώς του, για μια μεγάλη χρονική περίοδο. Σε τέτοιες περιπτώσεις η συνάρτηση κινδύνου είναι φθίνουσα στην αρχική περίοδο (βρεφική περίοδος, *early life period*) όπου αποτυχίες (βλάβη για τη συσκευή, θάνατος για τον άνθρωπο) μπορούν να αποδοθούν σε αδυναμίες σχεδίασης των εξαρτημάτων για τη συσκευή ή σε βρεφικές ασθένειες για τον άνθρωπο. Στη συνέχεια υπάρχει μια περίοδος (χρήσιμη περίοδος, *useful period*) όπου η συνάρτηση κινδύνου είναι σχεδόν σταθερή και οι αποτυχίες οφείλονται σε τυχαίους λόγους (το τέλος αυτής της περιόδου για τον άνθρωπο είναι συνήθως τα 30 έτη). Τέλος στην τρίτη περίοδο (περίοδος φθοράς, *wear-out period*) η συνάρτηση κινδύνου είναι αύξουσα και απεικονίζει τη φθορά (για τη συσκευή) ή τη γήρανση (για τον άνθρωπο) με την πάροδο του χρόνου.

Συναρτήσεις κινδύνου μορφής καμπούρας (μορφής (δ)), δηλαδή αύξουσα συνάρτηση την αρχική περίοδο και φθίνουσα μετέπειτα, εμφανίζονται όταν μελετούμε το χρόνο αποτυχίας μετά από μια επιτυχή χειρουργική επέμβαση όπου αρχικά υπάρχει αυξημένος κίνδυνος θανάτου λόγω μετεγχειρητικών επιπλοκών (αιμορραγία, μολύνσεις κτλ.), ο οποίος μειώνεται σταθερά με την πάροδο του χρόνου.

Είναι πλέον αντιληπτό ότι η μορφή της συνάρτησης κινδύνου μιας κατανομής χρόνου ζωής έχει σαφή φυσική ερμηνεία. Συνεπώς οποιαδήποτε πληροφορία γύρω από τη φύση

(μορφή) της συνάρτησης κινδύνου είναι χρήσιμη για την αναγνώριση και τελικά την επιλογή ενός κατάλληλου παραμετρικού μοντέλου για την περιγραφή της υπό μελέτη κατανομής.

## 1.5 Λογοκριμένα δεδομένα

Έστω ότι ο χρόνος ζωής των ατόμων ενός πληθυσμού περιγράφεται από μια τυχαία μεταβλητή  $X$  με συνάρτηση επιβίωσης  $S(t)$  και έστω  $X_1, X_2, X_3, \dots, X_n$  ένα τυχαίο δείγμα χρόνων ζωής μεγέθους  $n$  από τον πληθυσμό. Στην περίπτωση που οι παρατηρούμενοι χρόνοι ζωής  $x_1, x_2, \dots, x_n$  είναι γνωστοί μπορούμε να χρησιμοποιήσουμε γνωστές (παραμετρικές και μη παραμετρικές) στατιστικές μεθόδους για τη μελέτη του πληθυσμού (εμπειρική συνάρτηση κατανομής, έλεγχοι  $\chi^2$  καλής προσαρμογής, κριτήριο Kolmogorov-Smirnov, κτλ.). Το ιδιαίτερο χαρακτηριστικό που εμφανίζεται στην ανάλυση επιβίωσης είναι ότι οι χρόνοι ζωής ορισμένων ατόμων του δείγματος είναι λογοκριμένοι (*censored*). Μια λογοκριμένη παρατήρηση παρέχει μόνο μερική πληροφόρηση για το χρόνο ζωής του αντίστοιχου ατόμου και μπορεί να είναι είτε λογοκριμένος από δεξιά (*right censoring*) στην περίπτωση που γνωρίζουμε ότι ο χρόνος ζωής του ατόμου είναι μεγαλύτερος από κάποιο χρόνο  $U$ , είτε λογοκριμένος από αριστερά (*left censoring*) στην περίπτωση που γνωρίζουμε ότι ο χρόνος ζωής του ατόμου είναι μικρότερος από κάποιο χρόνο  $U$ , είτε λογοκριμένος σε διάστημα (*interval censoring*) στην περίπτωση που γνωρίζουμε ότι ο χρόνος ζωής του ατόμου βρίσκεται εντός ενός διαστήματος της μορφής  $(L, R)$  με  $L < R$ .

### ΠΑΡΑΔΕΙΓΜΑ 1.5.1:

Υπάρχουν 8 ασθενείς (I~VIII) σε μια κλινική μελέτη 12-μηνών ( $D$ : θάνατος,  $L$ : χαμένες πληροφορίες).

Σε αυτή την έρευνα, οι ακριβείς χρόνοι αποτυχίας μερικών ασθενών είναι άγνωστοι είτε επειδή οι ασθενείς αποσύρονται από τη μελέτη ή επειδή οι ασθενείς ήταν ακόμα ζωντανοί στο τέλος της μελέτης. Αναφέρουμε τα ανωτέρω κατάσταση ως "λογοκρισία".

Η λογοκρισία είναι τόσο κοινή σε ιατρικά πειράματα που οι στατιστικές μέθοδοι την χρησιμοποιούν στις έρευνές τους.

Για το ανωτέρω παράδειγμα, έχουμε τα ακόλουθα στοιχεία:

6	7*	9.5*	7	10*	7	6*	11
---	----	------	---	-----	---	----	----

όπου "\*" αντιπροσωπεύει τη λογοκρισία. Περαιτέρω, έχουμε τον ακόλουθο πίνακα με πληροφορίες για αυτούς τους ασθενείς:

Subject	Survival time	Censor indicator	Group	# of cigarette	Gender	Age
I	6	1	T	20	1	45
II	7	0	T	30	1	20
III	9.5	0	T	5	0	38
IV	7	1	T	40	1	26
V	10	0	C	3	0	42
VI	7	1	C	40	0	17
VII	6	0	C	60	1	25
VIII	11	1	C	10	0	29

**Πίνακας 1.5.1** Δεδομένα του παραδείγματος 1.5.1

όπου ο δείκτης λογοκριτών (**1: θάνατος, 0: λογοκριμένο δεδομένο**) και ομάδα (**T: ομάδα θεραπείας, C: ομάδα ελέγχου**).

#### Στόχος:

Για τα ανωτέρω χαρακτηριστικά στοιχεία επιβίωσης, ενδιαφερόμαστε για

1. η λειτουργία επιβίωσης
2. η σύγκριση δύο ομάδων στοιχείων επιβίωσης
3. ποιοι παράγοντες (αριθμός των τσιγάρων, του φύλου ή της ηλικίας) είναι σημαντικός για την απόφαση του ποσοστού αποτυχίας



## Kaplan-Meier

**Case Processing Summary**

Total N	N of Events	Censored	
		N	Percent
8	4	4	50,0%

**Survival Table**

	Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
			Estimate	Std. Error		
1	6,000	0	.875	.117	1	7
2	6,000	1	.	.	1	6
3	7,000	0	.729	.165	2	5
4	7,000	1	.	.	2	4
5	7,000	1	.	.	2	3
6	9,500	0	.486	.227	3	2
7	10,000	0	.243	.206	4	1
8	11,000	1	.	.	4	0

**Πίνακας 1.5.2** Πίνακας ανάλυσης των δεδομένων μέσω SPSS

**Means and Medians for Survival Time**

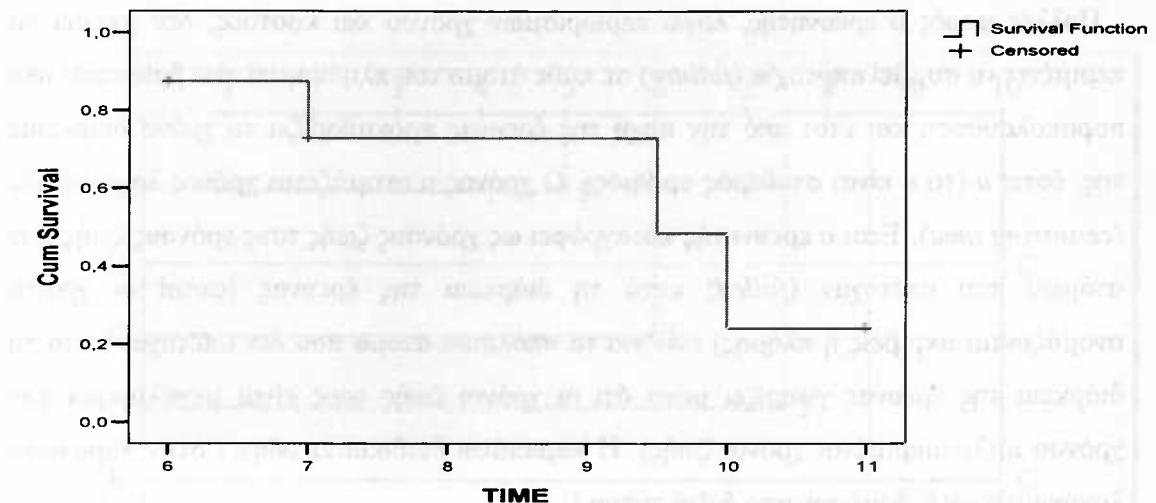
Mean <sup>a</sup>				Median			
Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound			Lower Bound	Upper Bound
9,184	.654	7,902	10,466	9,500	1,400	6,756	12,244

a. Estimation is limited to the largest survival time if it is censored.

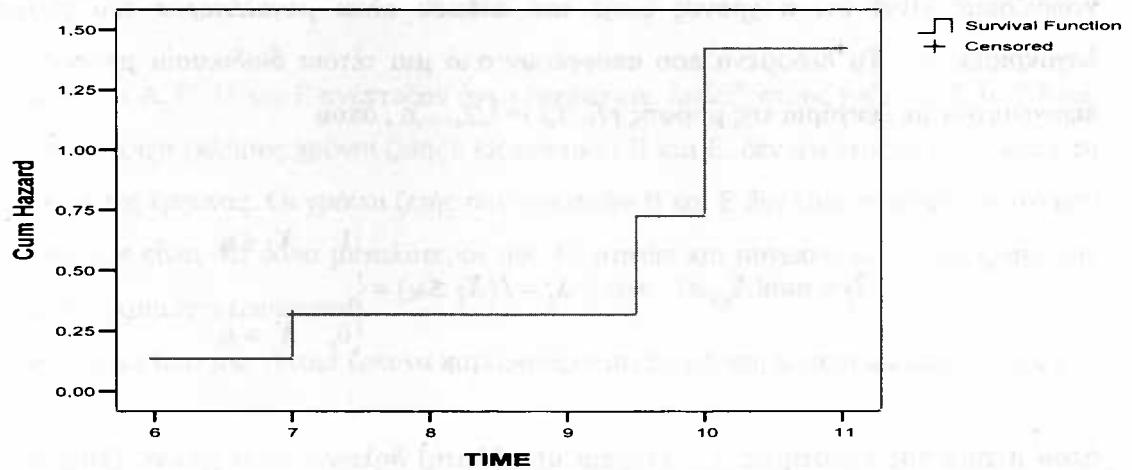
**Πίνακας 1.5.3** Πίνακας μέσου και διαμέσου για το χρόνο επιβίωσης μέσω SPSS

### Τα διαγράμματα που ακολουθούν παρουσιάζουν την επιβίωση και την κίνδυνο για ένα σύνολο δεδομένων.

**Survival Function**



**Hazard Function**



**Διάγραμμα 1.5.1 Διαγράμματα των συναρτήσεων επιβίωσης και κίνδυνον**

### 1.5.1. Δεξιά λογοκρισία τύπου I (type I right censoring)

Πολλές φορές ο ερευνητής, λόγω περιορισμών χρόνου και κόστους, δεν μπορεί να περιμένει να συμβεί αποτυχία (*failure*) σε κάθε άτομο του πληθυσμού που βρίσκεται υπό παρακολούθηση και έτσι από την αρχή της έρευνας προκαθορίζει το χρόνο διάρκειάς της, έστω  $u$  (το  $u$  είναι σταθερός αριθμός). Ο χρόνος  $u$  ονομάζεται χρόνος λογοκρισίας (*censoring time*). Έτσι ο ερευνητής καταγράφει ως χρόνους ζωής τους χρόνους ζωής των ατόμων που απέτυχαν (*failed*) κατά τη διάρκεια της έρευνας (αυτοί οι χρόνοι ονομάζονται ακριβείς ή πλήρης) ενώ για τα υπόλοιπα άτομα που δεν απέτυχαν κατά τη διάρκεια της έρευνας γνωρίζει μόνο ότι οι χρόνοι ζωής τους είναι μεγαλύτεροι του χρόνου  $u$  (λογοκριμένοι χρόνοι ζωής). Η παραπάνω διαδικασία οδηγεί στην παραγωγή λογοκριμένων δεδομένων από δεξιά τύπου I.

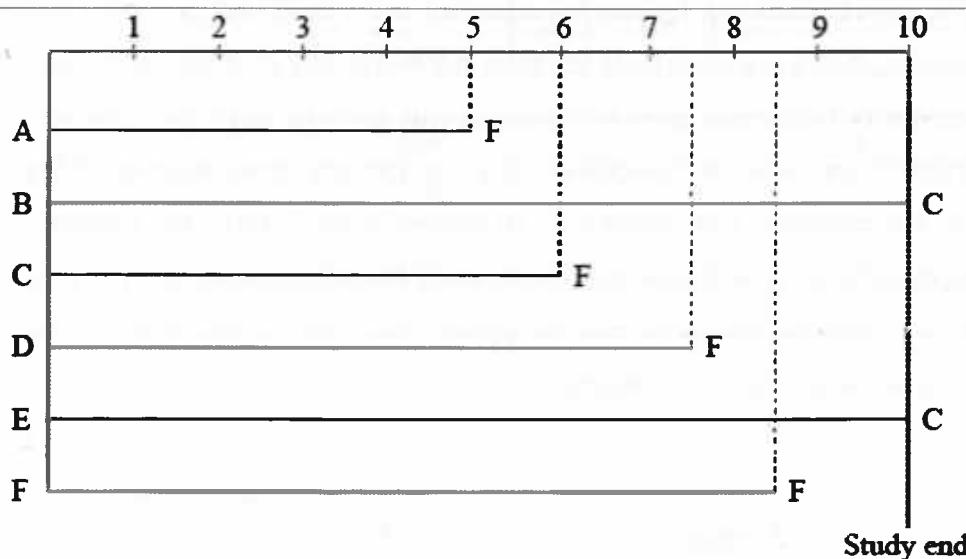
Στη γενική περίπτωση ας συμβολίσουμε με  $X_1, X_2, X_3, \dots, X_n$  τους χρόνους ζωής των  $n$  ατόμων του πληθυσμού από τη χρονική στιγμή αρχής της έρευνας. Στην περίπτωση της δεξιάς λογοκρισίας τύπου I (*type I right censoring*) ο πλήρης χρόνος ζωής  $X_i$  του ατόμου  $i$  παρατηρείται αν και μόνο αν  $X_i \leq u$ , ενώ στην περίπτωση που  $X_i > u$  το μόνο που γνωρίζουμε είναι ότι ο χρόνος ζωής του ατόμου είναι μεγαλύτερος του χρόνου λογοκρισίας  $u$ . Τα δεδομένα που απορρέουν από μια τέτοια διαδικασία μπορούν να περιγραφούν με ζευγάρια της μορφής  $(T_i, \Delta_i)$   $i=1,2,\dots,n$ , όπου

$$T_i = \min(X_i, u) \quad \text{και} \quad \Delta_i = I(X_i \leq u) = \begin{cases} 1, & X_i \leq u \\ 0, & X_i > u \end{cases}$$

όπου η τιμή της ποσότητας  $\Delta_i$  (τυχαία μεταβλητή) δηλώνει αν ο χρόνος ζωής του  $i$  ατόμου είναι λογοκριμένος  $\Delta_i=0$  ή πλήρης  $\Delta_i=1$ .

Για παράδειγμα ας θεωρήσουμε ότι 6 ποντικοί (A, B, C, D, E και F) υποβάλλονται (την ίδια χρονική στιγμή) σε διαδικασία καρκινογένεσης με εμβολιασμό καρκινικών κυττάρων και μας ενδιαφέρει ο χρόνος που απαιτείται για την ανάπτυξη όγκου προκαθορισμένου μεγέθους (αποτυχία). Ο ερευνητής αποφασίζει να τερματίσει το

πείραμά του μετά από 10 μήνες ( $u=10$ ). Το διάγραμμα απεικονίζει τις πληροφορίες που συλλέξαμε για τους χρόνους ανάπτυξης του όγκου



**Διάγραμμα 1.5.2 Απεικόνιση πληροφοριών για τους χρόνους ανάπτυξης του όγκου**

Οι ποντικοί A, C, D και F ανέπτυξαν όγκο (απέτυχαν, *failed*) στους χρόνους 5, 6, 7.5 και 8.5, αντίστοιχα (πλήρης χρόνοι ζωής). Οι ποντικοί B και E, δεν ανέπτυξαν όγκο κατά τη διάρκεια της έρευνας. Οι χρόνοι ζωής των ποντικών B και E δεν είναι γνωστοί. Αυτό που γνωρίζουμε είναι ότι είναι μεγαλύτεροι των 10 μηνών και συνεπώς οι χρόνοι ζωής των είναι λογοκριμένοι (*censored*).

Τα δεδομένα από μια τέτοια έρευνα παρουσιάζονται σε μορφή πίνακα ως ακολούθως

	A	B	C	D	E	F
$(t_i, \delta_i)$	(5, 1)	(10, 0)	(6, 1)	(7.5, 1)	(10, 0)	(8.5, 1)

ή πιο απλά, τα δεδομένα μας καταγράφονται ως 5, 10+, 6, 7.5, 10+, 8.5 όπου το σύμβολο “+” δηλώνει λογοκριμένο χρόνο ζωής (από δεξιά).

Μια πιο γενικευμένη περίπτωση λογοκρισίας τύπου I προκύπτει στην περίπτωση που η διάρκεια παρακολούθησης (χρόνοι λογοκρισίας) των  $n$  ατόμων αν και είναι γνωστή (και προκαθορισμένη) για κάθε άτομο δεν είναι η ίδια για όλα τα άτομα (στο προηγούμενο παράδειγμα οι 6 ποντικοί ήταν στη διάθεσή μας από την αρχή της έρευνας). Τέτοιες καταστάσεις μπορούν να προκύψουν όταν η χρονική αρχή παρακολούθησης κάθε ατόμου δεν συμπίπτει αναγκαστικά με τη χρονική αρχή έναρξης της έρευνας. Σε αυτή την περίπτωση σε κάθε άτομο αντιστοιχεί ένας προκαθορισμένος χρόνος λογοκρισίας, έστω  $u_i$ , και τα δεδομένα που απορρέουν από μια τέτοια έρευνα μπορούν να περιγραφούν με ζευγάρια της μορφής  $(T_i, \Delta_i)$ ,  $i=1,2,\dots,n$

$$T_i = \min(X_i, u_i) \quad \text{και} \quad \Delta_i = I(X_i \leq u_i) = \begin{cases} 1, & X_i \leq u_i \\ 0, & X_i > u_i \end{cases}$$

Ας θεωρήσουμε το ακόλουθο παράδειγμα που οφείλεται στον Bartholomew Στο παράδειγμα αυτό υπάρχει ένα σύστημα στο οποίο εγκαθίστανται εξαρτήματα σε διαφορετικές ημερομηνίες (τα εξαρτήματα παράγονται το ένα κατόπιν του άλλου σύμφωνα με ένα αυστηρό χρονοδιάγραμμα και συνεπώς δεν είναι όλα διαθέσιμα από την αρχή). Μας ενδιαφέρει να μελετήσουμε το χρόνο ζωής των εξαρτημάτων προκαθορίζοντας ως ημερομηνία λήξης της έρευνας την 31η Αυγούστου. Μερικά από τα εξαρτήματα μπορεί να αποτύχουν κατά τη διάρκεια της έρευνας ενώ μερικά άλλα παραμένουν σε χρήση πέραν της 31ης Αυγούστου. Ο παρακάτω πίνακας δίνει πληροφορίες για το χρόνο ζωής κάθε εξαρτήματος:

Εξάρτημα	Ημερομηνία εγκατάστασης	Ημερομηνία αποτυχίας	Χρόνος ζωής $X_i$	Χρόνος λογοκρισίας $u_i$
1	11/06	13/06	2	81
2	21/06	-	-	71
3	22/06	12/08	51	70
4	02/07	-	-	60
5	21/07	23/08	33	41
6	31/07	27/08	27	31
7	31/07	14/08	14	31
8	01/08	25/08	24	30
9	02/08	06/08	4	29
10	10/08	-	-	21

**Πίνακας 1.5.4** Πληροφορίες για το χρόνο ζωής κάθε εξαρτήματος

Παρατηρούμε ότι το εξάρτημα 1 έχει (προκαθορισμένο) χρόνο λογοκρισίας  $u_1=81$  επειδή το χρονικό διάστημα που μεσολαβεί από την ημερομηνία εγκατάστασης του εξαρτήματος έως τη λήξη της έρευνας είναι 81 ημέρες. Συνεπώς ο πραγματικός χρόνος ζωής για το εξάρτημα 1 θα καταγραφεί μόνο αν το εξάρτημα αποτύχει εντός του χρόνου λογοκρισίας. Έτσι το δεδομένο που προκύπτει για το εξάρτημα 1 (σύμφωνα με τον πίνακα) είναι της μορφής

$$(t_1, \delta_1) = (\min(x_1, u_1), \delta_1) = (\min(2, 81), \delta_1) = (2, 1)$$

Τα σύνολο των δεδομένων παρουσιάζονται σε μορφή πίνακα ως ακολούθως

$i$	1	2	3	4	5	6	7	8	9	10
$(t_i, \delta_i)$	(2, 1)	(71, 0)	(51, 1)	(60, 0)	(33, 1)	(27, 1)	(14, 1)	(24, 1)	(4, 1)	(21, 0)

**Πίνακας 1.5.5** Δεδομένα σε μορφή πίνακα

τα δεδομένα μας καταγράφονται και ως  
2, 71+, 51, 60+, 33, 27, 14, 24, 4, 21+.

## Kaplan-Meier

**Survival Table**

	Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
			Estimate	Std. Error		
1	2,000	1	.	.	0	9
2	4,000	1	.	.	0	8
3	14,000	1	.	.	0	7
4	21,000	0	,857	,132	1	6
5	24,000	1	.	.	1	5
6	27,000	1	.	.	1	4
7	33,000	1	.	.	1	3
8	51,000	1	.	.	1	2
9	60,000	0	,429	,310	2	1
10	71,000	0	,000	,000	3	0

**Πίνακας 1.5.6** Πίνακας ανάλυσης των δεδομένων μέσω SPSS

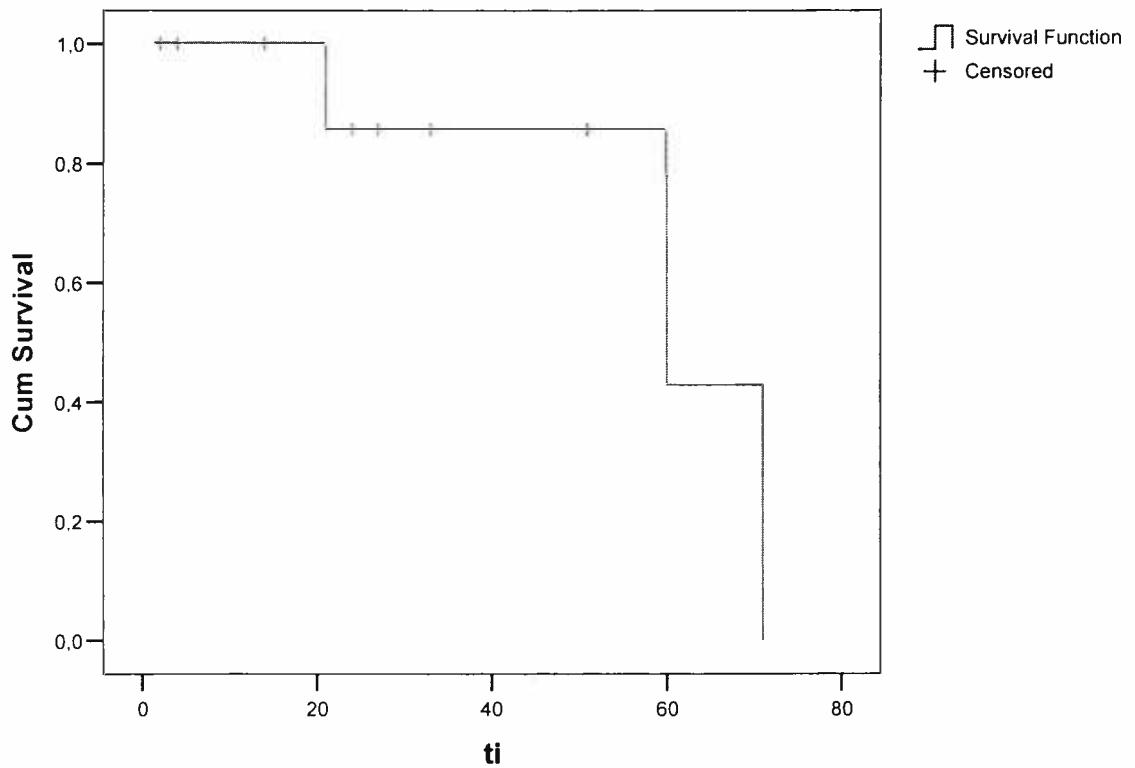
**Means and Medians for Survival Time**

Mean <sup>a</sup>				Median			
Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound			Lower Bound	Upper Bound
59,143	8,284	42,906	75,380	60,000	28,226	4,677	115,323

a. Estimation is limited to the largest survival time if it is censored.

**Πίνακας 1.5.7** Πίνακας μέσου και διαμέσου για το χρόνο επιβίωσης μέσω SPSS

### Survival Function



**Διάγραμμα 1.5.3** Διάγραμμα της συνάρτησης επιβίωσης

### 1.5.2. Δεξιά λογοκρισία τύπου II (type II right censoring)

Στην περίπτωση της δεξιάς λογοκρισίας τύπου II (type II right censoring) αποφασίζουμε από την αρχή της έρευνας ότι αυτή θα τερματιστεί τη χρονική στιγμή (που είναι τυχαία) που θα αποτύχουν συνολικά  $r$  άτομα ,  $r < n$ . Επομένως τα δεδομένα μας αποτελούνται από τους πλήρης χρόνους ζωής των πρώτων  $r$  ατόμων που απέτυχαν (διατεταγμένοι χρόνοι ζωής), ενώ για τα υπόλοιπα  $n-r$  άτομα γνωρίζουμε ότι ο χρόνος ζωής τους είναι μεγαλύτερος από τον μέγιστο χρόνο ζωής των  $r$  ατόμων που απέτυχαν  $X_{(r)}$  . Συνεπώς τα δεδομένα που απορρέουν από μια τέτοια έρευνα μπορούν να περιγραφούν με ζευγάρια της μορφής  $(T_{(i)}, A_{(i)})$  ,  $i = 1, 2, \dots, r$  όπου

$$T_{(i)} = \begin{cases} X_{(i)}, & 1 \leq i \leq r \\ X_{(r)}, & r+1 \leq i \leq n \end{cases} \quad \text{και} \quad A_{(i)} = I(i \leq r) = \begin{cases} 1, & 1 \leq i \leq r \\ 0, & r+1 \leq i \leq n \end{cases}$$

Για παράδειγμα, στο πείραμα με τα 6 ( $6 = n$ ) ποντίκια, αν ο ερευνητής είχε αποφασίσει να τερματίσει την έρευνα όταν τρεις ( $r = 3$ ) από τους ποντικούς εμφανίσουν ογκο, τα δεδομένα που θα κατέγραφε θα ήταν τα ακόλουθα

$$5, 6, 7.5, 7.5+, 7.5+, 7.5+.$$

Η δεξιά λογοκρισία τύπου II είναι χρήσιμη σε περιπτώσεις όπου δεν μπορούμε να προκαθορίσουμε ένα κατάλληλο χρόνο λογοκρισίας όπως στην περίπτωση της δεξιάς λογοκρισίας τύπου II. Επίσης, αξίζει να σημειώσουμε ότι σε ορισμένες έρευνες χρησιμοποιείται συνδυασμός της λογοκρισίας τύπου I και τύπου II (η έρευνα σταματά τη χρονική στιγμή  $u$  ή όταν αποτύχουν συνολικά  $r$  άτομα οτιδήποτε από τα δύο συμβεί πρώτο). Το σημαντικότερο μειονέκτημα της δεξιάς λογοκρισίας τύπου II είναι ότι ο συνολικός χρόνος  $T_{(r)}$  που διαρκεί η έρευνα είναι άγνωστος .

### 1.5.3. Τυχαία λογοκρισία (random censoring)

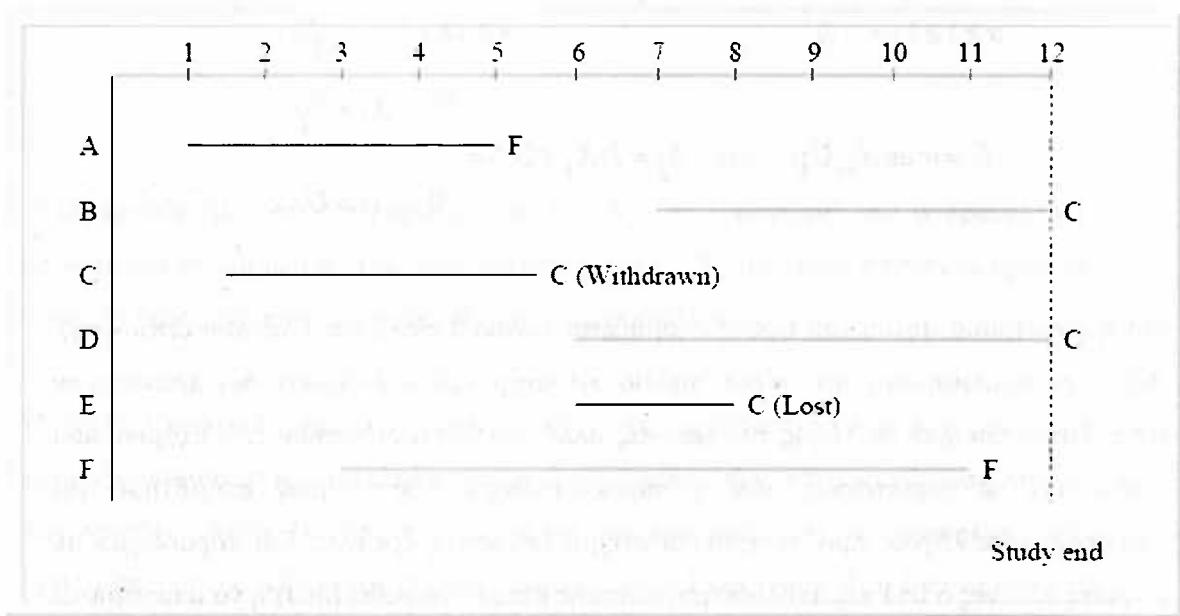
Σε πολλές περιπτώσεις ο χρόνος λογοκρισίας που αντιστοιχεί σε κάθε υπό παρακολούθηση άτομο δεν είναι σταθερός αλλά είναι τυχαίος. Για παράδειγμα σε κλινικές μελέτες, ενώ η χρονική στιγμή αρχής και τέλους της έρευνας είναι προκαθορισμένη, οι ασθενείς εισέρχονται σε αυτή σε διαφορετικές (τυχαίες) χρονικές στιγμές (π.χ. τη στιγμή που θα γίνει διάγνωση της ασθένειας) με αποτέλεσμα οι χρόνοι λογοκρισίας τους να είναι τυχαίοι. Έτσι σε κάθε άτομο αντιστοιχεί μια τυχαία μεταβλητή  $X$  που δηλώνει το χρόνο ζωής του ατόμου (από τη στιγμή που θα εισέλθει στην έρευνα) και μια τυχαία μεταβλητή  $U$  που δηλώνει το χρόνο λογοκρισίας του ατόμου. Επομένως, τα δεδομένα που απορρέουν από μια τέτοια έρευνα μπορούν να περιγραφούν με ζευγάρια της μορφής  $(T_i, \Delta_i)$ ,  $i=1,2,\dots,n$  όπου

$$T_i = \min(X_i, U_i) \quad \text{και} \quad \Delta_i = I(X_i \leq U_i) = \begin{cases} 1, & X_i \leq U_i \\ 0, & X_i > U_i \end{cases}$$

και η λογοκρισία αυτού του είδους ονομάζεται τυχαία λογοκρισία (*random censoring*). Αξίζει να σημειώσουμε στο παρόν σημείο ότι λογοκριμένα δεδομένα δεν προκύπτουν μόνο λόγω του χρόνου λήξης της έρευνας αλλά και λόγω διαφυγών των ατόμων που οφείλονται σε περιπτώσεις που η παρακολούθηση των ατόμων σταμάτησε για οποιουσδήποτε λόγους πριν τη χρονική στιγμή λήξης της έρευνας. Για παράδειγμα σε ιατρικές έρευνες ο υπό παρακολούθηση ασθενής μπορεί να χαθεί (*lost*) ή να αποχωρήσει (*withdraws*) πριν τη λήξη της έρευνας. Είναι συνηθισμένες οι περιπτώσεις που ο ασθενής μπορεί να πεθάνει κατά τη διάρκεια της παρακολούθησης (υποθέτουμε ότι ο θάνατος δεν είναι το ενδεχόμενο που μας ενδιαφέρει), ή να σταματήσει η θεραπευτική του αγωγή λόγω σοβαρών παρενεργειών (αποχωρήσεις). Επίσης μετά από κάποια χρονική περίοδο παρακολούθησης ο ασθενής μπορεί να χαθεί από την έρευνα λόγω αλλαγής κατοικίας

του ή αλλαγή του θεράποντος ιατρού, ή να αρνηθεί να συνεχίσει να λαμβάνει μέρος σε μια τέτοια διαδικασία.

Για να διασαφηνίσουμε τις παραπάνω έννοιες ας υποθέσουμε ότι 6 ασθενείς (A, B, C, D, E και F) με οξεία λευχαιμία εισέρχονται σε μια κλινική μελέτη που διαρκεί 12 εβδομάδες. Υποθέτουμε επίσης ότι και οι 6 ασθενείς ανταποκρίνονται στη θεραπεία που τους χορηγείται με αποτέλεσμα την υποχώρηση των συμπτωμάτων της νόσου μετά από κάποιο χρονικό διάστημα. Μόλις συμβεί αυτό η θεραπευτική αγωγή σταματά και μας ενδιαφέρει να μελετήσουμε τους χρόνους επανεμφάνισης των συμπτωμάτων της νόσου. Είναι προφανές ότι λόγω της τυχαιότητας του χρόνου ανταπόκρισης στην θεραπευτική αγωγή ο χρόνος λογοκρισίας κάθε ατόμου είναι τυχαίος. Το Διάγραμμα 1.5.4 δίνει πληροφορίες για τους χρόνους επανεμφάνισης της νόσου



**Διάγραμμα 1.5.4** Πληροφορίες για τους χρόνους επανεμφάνισης της οξείας λευχαιμίας

Στους ασθενείς A, B, C, D, E και F παρατηρείται υποχώρηση των συμπτωμάτων της νόσου τους χρόνους 1, 7, 1.5, 6, 6 και 3, αντίστοιχα. Συνεπώς οι αντίστοιχοι χρόνοι λογοκρισίας είναι

$$u_1=11, u_2=5, u_3=10.5, u_4=6, u_5=6, u_6=9$$

Στους ασθενείς Α και Φ επανεμφανίστηκαν τα συμπτώματα της ασθένειας στους χρόνους 5 και 11, αντίστοιχα. Στους ασθενείς Β και Δ η υποχώρηση των συμπτωμάτων διαρκεί τουλάχιστον μέχρι τη λήξη της έρευνας. Ο ασθενής Ζ αποχώρησε από τη μελέτη λόγω θανάτου το χρόνο 5.5, ενώ ο ασθενής Ε χάθηκε το χρόνο 8 λόγω αλλαγής της κατοικίας του. Τα δεδομένα που προκύπτουν από αυτή την έρευνα είναι τα ακόλουθα

$$4, 5+, 4+, 6+, 2+, 8.$$

#### **1.5.4. Αριστερή λογοκρισία, λογοκρισία σε διάστημα, περικομμένα δεδομένα (*left censoring, interval censoring, truncated data*)**

Στην περίπτωση της αριστερής λογοκρισίας (*left censoring*) ο πλήρης χρόνος ζωής  $X_i$  του ατόμου  $i$  παρατηρείται αν και μόνο αν  $X_i \leq U_i$ , ενώ στην περίπτωση που  $X_i < U_i$  το μόνο που γνωρίζουμε είναι ότι ο χρόνος ζωής του ατόμου είναι μικρότερος του χρόνου λογοκρισίας  $U_i$ . Τα δεδομένα που απορρέουν από μια τέτοια διαδικασία μπορούν να περιγραφούν με ζευγάρια της μορφής  $(T_i, \Delta_i)$ ,  $i = 1, 2, \dots, n$  όπου

$$T_i = \max(X_i, U_i) \quad \text{και} \quad \Delta_i = I(X_i \geq U_i) = \begin{cases} 1 & X_i \geq U_i \\ 0 & X_i < U_i \end{cases}$$

Για παράδειγμα ας θεωρήσουμε μια έρευνα που αφορά στη μελέτη της ηλικίας κατά την οποία μια συγκεκριμένη ομάδα μαθητών 12 ετών μαθαίνουν να εκτελούν μια ειδική εργασία (ο υπό μελέτη χρόνος ζωής αντιστοιχεί στο χρονικό διάστημα που μεσολαβεί από την ημερομηνία της γέννησης των παιδιών μέχρι την ημερομηνία που μαθαίνουν να εκτελούν την ειδική εργασία). Στην περίπτωση που υπάρχουν μαθητές που ήδη είχαν εκτελέσει την ειδική εργασία στο παρελθόν αλλά δεν γνωρίζουμε πότε αυτό συνέβη τότε προκύπτουν λογοκριμένα δεδομένα από αριστερά. Στην περίπτωση που οι μαθητές μαθαίνουν να εκτελούν την ειδική εργασία κατά τη διάρκεια της μελέτης τότε

προκύπτουν πλήρη δεδομένα. Στην ειδική περίπτωση που η έρευνα έχει προκαθορισμένη ημερομηνία λήξης τότε μπορεί να έχουμε ταυτόχρονα λογοκριμένα δεδομένα και από αριστερά και από δεξιά, οπότε ομιλούμε για διπλή λογοκρισία (*doubly censored*).

Επίσης, ένας άλλος μηχανισμός παραγωγής λογοκριμένων δεδομένων είναι η λογοκρισία σε διάστημα (*interval censoring*). Η λογοκρισία σε διάστημα απαντάται συχνά σε περιπτώσεις όπου άτομα περνούν από περιοδικές εξετάσεις σε προκαθορισμένους χρόνους για την ανίχνευση μιας συγκεκριμένης ασθένειας (ή περιοδική επιθεώρηση καλής λειτουργίας ενός εξαρτήματος μιας μηχανής). Σε τέτοιες περιπτώσεις αν το ενδεχόμενο συμβεί μεταξύ δύο διαδοχικών διαφορετικών χρόνων εξέτασης, έστω  $L_i$  και  $R_i$ , με  $L_i < R_i$ , τότε ο πλήρης χρόνος  $X$  είναι άγνωστος και γνωρίζουμε μόνο ότι  $L_i < X < R_i$ .

Κλείνοντας θα αναφερθούμε στα περικομμένα δεδομένα (*truncated data*). Η περικοπή (*truncation*) ορίζεται ως μια συνθήκη την οποία πρέπει να ικανοποιούν τα άτομα που θα τεθούν υπό την παρακολούθηση μας. Τα άτομα που δεν ικανοποιούν τη συνθήκη δεν τελούν υπό παρακολούθηση και ο ερευνητής αγνοεί την ύπαρξή τους (ο χρόνος ζωής  $X$  καταγράφεται μόνο αν  $X \leq y$  ή  $X \geq y$ , όπου  $y$  είναι ο χρόνος του ενδεχομένου που περικόπτει τα δεδομένα μας. ‘Ενα παράδειγμα περικομμένων δεδομένων (από δεξιά) αναφέρουμε την περίπτωση μιας έρευνας που δημοσιεύθηκε το 1988 και αφορούσε το χρόνο (έως την 30η Ιουνίου 1986) που χρειάστηκε να αναπτύξουν AIDS άτομα που μολύνθηκαν με τον ιό του AIDS (λόγω μολυσμένης μετάγγισης αίματος) την 1η Απριλίου 1978. Τα μολυσμένα άτομα που δεν ανέπτυξαν AIDS έως την 30η Ιουνίου 1986 δεν συμπεριλαμβάνονται στη μελέτη μας. Κατά τον ίδιο τρόπο, ένα άλλο παράδειγμα περικομμένων δεδομένων (από αριστερά) αναφέρουμε την περίπτωση μιας έρευνας που αφορούσε την ηλικία θανάτου ατόμων που διέμεναν σε ένα οίκο ευγηρίας. Επειδή ένα άτομο πρέπει να επιβιώσει μέχρι μια συγκεκριμένη ηλικία για να εισαχθεί στον οίκο ευγηρίας, τα άτομα που δεν επιβιώσαν μέχρι αυτή την ηλικία δεν θα τεθούν υπό παρακολούθηση.

## 1.6 Συνάρτηση πιθανοφάνειας και λογοκριμένα δεδομένα

Στην παρούσα παράγραφο δίνεται η μορφή της συνάρτησης πιθανοφάνειας στην περίπτωση που υπάρχουν λογοκριμένα δεδομένα. Για όλα τα είδη λογοκρισίας από δεξιά η μορφή της συνάρτησης πιθανοφάνειας είναι κοινή

### 1.6.1. Τυχαία λογοκρισία

Στην περίπτωση της τυχαίας λογοκρισίας οι παρατηρήσεις είναι της μορφής  $(T_i, \Delta_i)$   $i = 1, 2, \dots, n$  όπου

$$T_i = \min(X_i, U_i), \quad \Delta_i = \begin{cases} 1, & X_i \leq U_i \\ 0, & X_i > U_i \end{cases}$$

Ας συμβολίσουμε με  $f(\cdot)$ ,  $S(\cdot)$  τη συνάρτηση πυκνότητας και τη συνάρτηση επιβίωσης της τυχαίας μεταβλητής  $X$ , αντίστοιχα, και με  $g(\cdot)$ ,  $G(\cdot)$  τη συνάρτηση πυκνότητας και τη συνάρτηση επιβίωσης της τυχαίας μεταβλητής  $U$ , αντίστοιχα. Ας υποθέσουμε ότι οι τυχαίες μεταβλητές  $X$  και  $U$  είναι ανεξάρτητες. Για την κατανομή του ζεύγους  $(T_i, \Delta_i)$  έχουμε ότι

$$f_{T_i, \Delta_i}(t_i, \delta_i) = f(t_i, \delta_i) = \lim_{h \rightarrow 0} \frac{P(t_i \leq T_i < t_i + h, \Delta_i = \delta_i)}{h}, \quad t_i \geq 0, \quad \delta_i \in \{0, 1\}$$

Παρατηρούμε ότι

$$\begin{aligned}
f(t_i, 1) &= \lim_{h \rightarrow 0} \frac{P(t_i \leq T_i < t_i + h, \Delta_i = 1)}{h} = \lim_{h \rightarrow 0} \frac{P(t_i \leq X_i < t_i + h, U_i \geq X_i)}{h} \\
&= P(U_i \geq t_i) \lim_{h \rightarrow 0} \frac{P(t_i \leq X_i < t_i + h)}{h} = G(t_i) f(t_i), \\
f(t_i, 0) &= \lim_{h \rightarrow 0} \frac{P(t_i \leq T_i < t_i + h, \Delta_i = 0)}{h} = \lim_{h \rightarrow 0} \frac{P(t_i \leq U_i < t_i + h, X_i > U_i)}{h} \\
&= P(X_i > t_i) \lim_{h \rightarrow 0} \frac{P(t_i \leq U_i < t_i + h)}{h} = S(t_i) g(t_i)
\end{aligned}$$

Έτσι μπορούμε να γράψουμε ότι

$$f(t_i, \delta_i) = [f(t_i)G(t_i)]^{\delta_i} [g(t_i)S(t_i)]^{1-\delta_i}, \quad t_i \geq 0, \quad \delta_i \in \{0,1\}$$

Ας υποθέσουμε ότι  $f(\cdot) = f(\cdot, \theta)$  και  $g(\cdot) = g(\cdot, \varphi)$ . Τότε η συνάρτηση πιθανοφάνειας δίνεται από τη σχέση

$$\begin{aligned}
L(\theta, \varphi) &= L(\theta, \varphi | \mathbf{t}, \hat{\mathbf{o}}) \\
&= \prod_{i=1}^n f(t_i, \delta_i; \theta, \varphi) \\
&= \prod_{i=1}^n [f(t_i; \theta)G(t_i; \varphi)]^{\delta_i} [g(t_i; \varphi)S(t_i; \theta)]^{1-\delta_i} \\
&= \left( \prod_{i=1}^n [f(t_i; \theta)]^{\delta_i} [S(t_i; \theta)]^{1-\delta_i} \right) \left( \prod_{i=1}^n [G(t_i; \varphi)]^{\delta_i} [g(t_i; \varphi)]^{1-\delta_i} \right)
\end{aligned}$$

Ο στόχος μιας τέτοιου είδους ανάλυσης είναι να βρεθεί ο ΕΜΠ της παραμέτρου  $\theta$  της συνάρτησης πυκνότητας  $f(\cdot, \theta)$  της τυχαίας μεταβλητής  $X$ . Αφού έχουμε υποθέσει ότι οι τυχαίες μεταβλητές  $X$  και  $U$  είναι ανεξάρτητες είναι λογικό να υποθέσουμε ότι δεν υπάρχει σχέση μεταξύ των ποσοτήτων  $\theta$  και  $\varphi$  και επομένως ο όρος

$$\prod_{i=1}^n [G(t_i; \boldsymbol{\varphi})]^{\delta_i} [g(t_i; \boldsymbol{\varphi})]^{1-\delta_i}$$

που εμφανίζεται στη συνάρτηση πιθανοφάνειας  $L(\theta, \varphi)$  μπορεί να θεωρηθεί ως μια σταθερά που η τιμή της προφανώς δεν επηρεάζει την ανάλυση για την εύρεση του ΕΜΠ της παραμέτρου  $\theta$ . Συνεπώς για την εύρεση ΕΜΠ της παραμέτρου  $\theta$  μπορούμε να χρησιμοποιήσουμε ως “συνάρτηση πιθανοφάνειας” την ποσότητα

$$L(\theta) = \prod_{i=1}^n [f(t_i; \theta)]^{\delta_i} [S(t_i; \theta)]^{1-\delta_i} = \left( \prod_{i \in D} f(t_i; \theta) \right) \left( \prod_{i \in C} S(t_i; \theta) \right)$$

όπου τα σύνολα  $D$  και  $C$  αποτελούν διαμέριση του συνόλου των δεικτών  $\{1, 2, \dots, n\}$ , με το σύνολο  $D$  να περιέχει τους δείκτες που δηλώνουν τα πλήρη δεδομένα και το σύνολο  $C$  να περιέχει τους δείκτες που δηλώνουν τα λογοκριμένα δεδομένα. Επίσης αν συμβολίσουμε με  $h(\cdot) = h(\cdot, \theta)$  και με  $H(\cdot) = H(\cdot, \theta)$  τη συνάρτηση κινδύνου και την αθροιστική συνάρτηση κινδύνου της τυχαίας μεταβλητής  $X$  και χρησιμοποιήσουμε τις σχέσεις

$$f(t_i; \theta) = h(t_i; \theta) S(t_i; \theta), \quad S(t_i; \theta) = \exp(-H(t_i; \theta))$$

και άρα

$$L(\theta) = \prod_{i=1}^n [h(t_i; \theta)]^{\delta_i} S(t_i; \theta) = \prod_{i=1}^n [h(t_i; \theta)]^{\delta_i} \exp[-H(t_i; \theta)]$$

### 1.6.2. Δεξιά λογοκρισία τόπου I

Στην περίπτωση της δεξιάς λογοκρισίας τόπου I οι παρατηρήσεις είναι της μορφής  $(T_i, A_i)$ ,  $i = 1, 2, \dots, n$  όπου

$$T_i = \min(X_i, u_i), \quad \Delta_i = \begin{cases} 1, & X_i \leq u_i \\ 0, & X_i > u_i \end{cases}$$

Στην περίπτωση αυτή έχουμε ότι

$$f_{T_i, \Delta_i}(t_i, 0) = f(t_i, 0) = \begin{cases} 0, & t_i \neq u_i \\ P(X_i > t_i) = S(t_i), & t_i = u_i \end{cases}$$

και

$$f_{T_i, \Delta_i}(t_i, 1) = f(t_i, 1) = \lim_{h \rightarrow 0} \frac{P(t_i \leq T_i < t_i + h, \Delta_i = 1)}{h} = \lim_{h \rightarrow 0} \frac{P(t_i \leq X_i < t_i + h)}{h} = f(t_i), \quad 0 \leq t_i \leq u_i$$

οπότε

$$f_{T_i, \Delta_i}(t_i, \delta_i) = f(t_i, \delta_i) = f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}, \quad 0 \leq t_i \leq u_i, \quad \delta_i \in \{0, 1\}$$

Συνεπώς η συνάρτηση πιθανοφάνειας δίνεται από τον τύπο

$$L(\theta) = \prod_{i=1}^n f(t_i, \delta_i; \theta, \phi) = \prod_{i=1}^n [f(t_i; \theta)]^{\delta_i} [S(t_i; \theta)]^{1-\delta_i} = \left( \prod_{i \in D} f(t_i; \theta) \right) \left( \prod_{i \in C} S(t_i; \theta) \right)$$

Στην ίδια συνάρτηση πιθανοφάνειας καταλήγουμε και στην περίπτωση που

$u_1 = u_2 = u_3 = \dots = u_n = u$

### 1.6.3. Δεξιά λογοκρισία τύπου II

Στην περίπτωση της δεξιάς λογοκρισίας τύπου II οι παρατηρήσεις είναι της μορφής  $(T_{(i)}, \Delta_i), i=1,2,\dots,n$

$$T_{(i)} = \begin{cases} X_{(i)}, & 1 \leq i \leq r \\ X_{(r)}, & r+1 \leq i \leq n \end{cases}, \quad \Delta_i = \begin{cases} 1, & 1 \leq i \leq r \\ 0, & r+1 \leq i \leq n \end{cases}$$

Από τη θεωρία των διατεταγμένων δειγμάτων (*order statistics*) προκύπτει ότι η από κοινού κατανομή του δείγματος είναι ίση με

$$\frac{n!}{(n-r)!} \prod_{i=1}^r [f(t_{(i)}; \theta)]^{\delta_i} [S(t_{(r)}; \theta)]^{1-\delta_i} = \frac{n!}{(n-r)!} [S(t_{(r)}; \theta)]^{n-r} \prod_{i=1}^r [f(t_{(i)}; \theta)]$$

Συνεπώς για την εύρεση ΕΜΠ μπορούμε να χρησιμοποιήσουμε ως συνάρτηση πιθανοφάνειας την ποσότητα

$$L(\theta) = [S(t_{(r)}; \theta)]^{n-r} \prod_{i=1}^r f(t_{(i)}; \theta) = \left( \prod_{i \in D} f(t_i; \theta) \right) \left( \prod_{i \in C} S(t_i; \theta) \right)$$

### 1.6.4. Αριστερή λογοκρισία, λογοκρισία σε διάστημα, περικοπή.

Για όλες τις περιπτώσεις λογοκρισίας που έχουμε αναφέρει η συνάρτηση πιθανοφάνειας υπακούει στο γενικό τύπο

$$L(\theta) \propto \left( \prod_{i \in D} f(t_i; \theta) \right) \left( \prod_{i \in C_r} S(t_i; \theta) \right) \left( \prod_{i \in C_l} [1 - S(t_i; \theta)] \right) \left( \prod_{i \in I} [S(L_i; \theta) - S(R_i; \theta)] \right)$$

όπου τα σύνολα  $D$ ,  $C_r$ ,  $C_l$  και  $I$  αποτελούν διαμέριση του συνόλου των δεικτών {1,2,...,n}, με το σύνολο  $D$  να περιέχει τους δείκτες που δηλώνουν τα πλήρη δεδομένα, το σύνολο  $C_r$  να περιέχει τους δείκτες που δηλώνουν τα λογοκριμένα δεδομένα από δεξιά, το σύνολο  $C_l$  να περιέχει τους δείκτες που δηλώνουν τα λογοκριμένα δεδομένα από αριστερά, και το σύνολο  $I$  να περιέχει τους δείκτες που δηλώνουν τα λογοκριμένα δεδομένα σε διάστημα.

Στην περίπτωση που υπάρχουν περικομμένα δεδομένα από αριστερά ( $X > y$ ) δουλεύουμε με τη συνάρτηση επιβίωσης

$$S(x | X > y) = \frac{S(x)}{S(y)}, \quad x > y$$

και τη συνάρτηση πυκνότητας

$$f(x | X > y) = \frac{d}{dx} F(x | X > y) = \frac{d}{dx} (1 - S(x | X > y)) = \frac{d}{dx} \left( \frac{S(y) - S(x)}{S(y)} \right) = \frac{f(x)}{S(y)}, \quad x > y$$

Για παράδειγμα, όταν σε μια έρευνα έχουμε περικομμένα δεδομένα από αριστερά ( $X > y$ ) και τυχαία λογοκρισία τότε

$$L(\theta) = \prod_{i=1}^n \left[ \frac{f(t_i; \theta)}{S(y; \theta)} \right]^{\delta_i} \left[ \frac{S(t_i; \theta)}{S(y; \theta)} \right]^{1-\delta_i} = \left( \prod_{i \in D} \frac{f(t_i; \theta)}{S(y; \theta)} \right) \left( \prod_{i \in C_r} \frac{S(t_i; \theta)}{S(y; \theta)} \right)$$

Επίσης όταν έχουμε περικομμένα δεδομένα από δεξιά ( $X < y$ ) δουλεύουμε με τη συνάρτηση πυκνότητας

$$f(x | X < y) = \frac{d}{dx} F(x | X < y) = \frac{f(x)}{1 - S(y)}, \quad x < y$$

Τότε

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n \frac{f(t_i; \boldsymbol{\theta})}{1 - S(y; \boldsymbol{\theta})}$$

### **ΠΑΡΑΔΕΙΓΜΑ 1.6.1**

Έστω ότι έχουμε στη διάθεσή μας ένα τυχαίο δείγμα χρόνων ζωής μεγέθους  $n$  από την εκθετική κατανομή με παράμετρο  $\lambda$  ( $X \sim Exp(\lambda)$ ).

- Για πλήρη δεδομένα έχουμε ότι η συνάρτηση πιθανοφάνειας δίνεται από τον τύπο

$$L(\lambda) = \prod_{i=1}^n f(x_i; \lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$$

Παρατηρούμε ότι

$$\ell(\lambda) = \log L(\lambda) = n \log \lambda - \lambda \sum_{i=1}^n x_i$$

$$\frac{d}{d\lambda} \ell(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^n x_i, \quad \frac{d^2}{d\lambda^2} \ell(\lambda) = -\frac{n}{\lambda^2} < 0$$

οπότε ο ΕΜΠ της παραμέτρου  $\lambda$  δίνεται από τον τύπο

$$\hat{\lambda} = \hat{\lambda}(\mathbf{X}) = \frac{n}{\sum_{i=1}^n X_i} = \bar{X}$$

Για ελέγχους υποθέσεων και διαστήματα εμπιστοσύνης για την παράμετρο  $\lambda$  δεν είναι απαραίτητο να καταφύγουμε στην ασυμπτωτική κατανομή του ΕΜΠ. Για την κατασκευή

διαστήματος εμπιστοσύνης (με συντελεστή εμπιστοσύνης  $1 - \alpha$ ) για την παράμετρο  $\lambda$  παρατηρούμε αρχικά ότι η τυχαία μεταβλητή  $\sum_{i=1}^n X_i$ , ακολουθεί την κατανομή  $G(\lambda, n)$ .

Έτσι

$$2\lambda \sum_{i=1}^n X_i \sim G(1/2, n) \quad \text{οπότε} \quad 2\lambda \sum_{i=1}^n X_i \sim \chi^2_{2n} \quad \text{ή ισοδύναμα} \quad \frac{2n\lambda}{\hat{\lambda}} \sim \chi^2_{2n}$$

Αφού

$$P(\chi^2_{2n;1-(\alpha/2)} \leq 2\lambda \sum_{i=1}^n X_i \leq \chi^2_{2n;\alpha/2}) = 1 - \alpha$$

άρα

$$P\left(\frac{\chi^2_{2n;1-(\alpha/2)}}{2 \sum_{i=1}^n X_i} \leq \lambda \leq \frac{\chi^2_{2n;\alpha/2}}{2 \sum_{i=1}^n X_i}\right) = 1 - \alpha$$

οπότε το ζητούμενο διάστημα εμπιστοσύνης (με συντελεστή εμπιστοσύνης  $(1-\alpha)$  για την παράμετρο  $\lambda$  είναι το

$$\left[ \frac{\chi^2_{2n;1-(\alpha/2)}}{2 \sum_{i=1}^n X_i}, \frac{\chi^2_{2n;\alpha/2}}{2 \sum_{i=1}^n X_i} \right] = \left[ \frac{\hat{\lambda} \chi^2_{2n;1-(\alpha/2)}}{2n}, \frac{\hat{\lambda} \chi^2_{2n;\alpha/2}}{2n} \right]$$

Τα μονόπλευρα διαστήματα εμπιστοσύνης (με συντελεστή εμπιστοσύνης  $\alpha$ . 1) για την παράμετρο  $\lambda$  είναι τα

$$\left[ \frac{\chi^2_{2n;\alpha}}{2 \sum_{i=1}^n X_i}, \infty \right], \quad \left[ 0, \frac{\chi^2_{2n;1-\alpha}}{2 \sum_{i=1}^n X_i} \right]$$

Για τὸν ἔλεγχο της υπόθεσης

$$H_0: \lambda = \lambda_0 \quad — \quad H_1: \lambda > \lambda_0$$

γνωρίζουμε ότι η κρίσιμη περιοχή είναι της μορφής  $\sum_{i=1}^n X_i < c$  (προκύπτει από το Λήμμα Neyman-Pearson).

Έτσι η σταθερά  $c$  προσδιορίζεται από τη σχέση

$$P(\sum_{i=1}^n X_i < c | H_0 \text{ αληθής}) = P_{\lambda=\lambda_0}(\sum_{i=1}^n X_i < c) = \alpha$$

ή ισοδύναμα από τη σχέση

$$P(2\lambda_0 \sum_{i=1}^n X_i < 2\lambda_0 c) = \alpha$$

Έτσι,  $2\lambda_0 c = \chi^2_{2n, 1-\alpha}$  και η κρίσιμη περιοχή  $K$  (περιοχή απόρριψης) είναι η

$$K: 2\lambda_0 \sum_{i=1}^n X_i < \chi^2_{2n, 1-\alpha}$$

Αντίστοιχα, για τὸν ἔλεγχο της υπόθεσης

$$H_0: \lambda = \lambda_0 \quad H_1: \lambda < \lambda_0$$

η κρίσιμη περιοχή είναι η

$$K: 2\lambda_0 \sum_{i=1}^n X_i > \chi^2_{2n, \alpha}$$

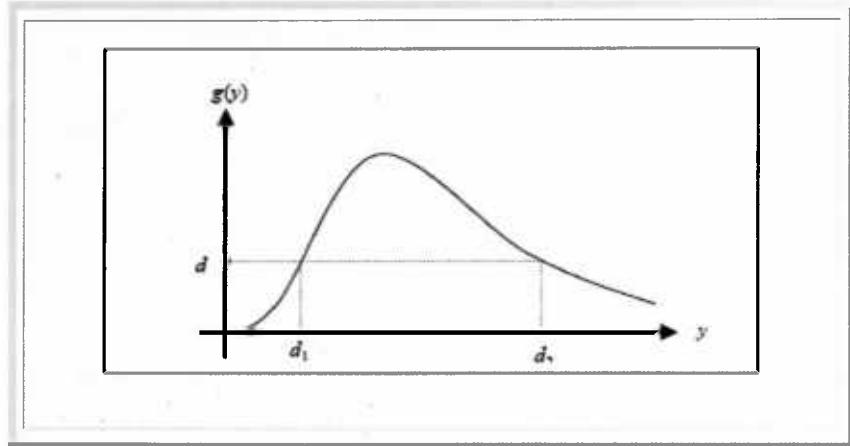
Για τὸν ἔλεγχο της υπόθεσης

$$H_0: \lambda = \lambda_0$$

$$H_1: \lambda \neq \lambda_0$$

χρησιμοποιώντας το τεστ γενικευμένου λόγου πιθανοφανειών προκύπτει ότι η κρίσιμη περιοχή είναι της μορφής  $K: \Lambda \leq c$ , δίνεται από τη σχέση

$$\Lambda = \frac{L(\lambda_0)}{L(\hat{\lambda})} = e^n \left( \frac{\lambda_0 \sum_{i=1}^n X_i}{n} \right)^n e^{-\lambda_0 \sum_{i=1}^n X_i} = e^n y^n e^{-ny} \leq c, \quad y = \frac{\lambda_0 \sum_{i=1}^n X_i}{n}$$



**Διάγραμμα 1.6.1** Διάγραμμα του εκθετικού μοντέλου  $g(y)$

Μπορεί να διαπιστωθεί ότι η ανισότητα  $g(y) = y^n e^{-ny} \leq d$  με  $d \leq e^{-n}$ , ικανοποιείται αν και μόνο αν  $y \leq d_1$  ή  $y \geq d_2$  όπου  $0 < d_1 < d < d_2$  και  $g(d_1) = g(d_2) = d$ . Έτσι η ανισότητα  $\Lambda \leq c$  είναι ισοδύναμη με την:

$$\frac{\lambda_0 \sum_{i=1}^n X_i}{n} \leq c_1 \quad \text{ή} \quad \frac{\lambda_0 \sum_{i=1}^n X_i}{n} \geq c_2$$

δηλαδή

$$2\lambda_0 \sum_{i=1}^n X_i \leq c_1 \quad \text{ή} \quad 2\lambda_0 \sum_{i=1}^n X_i \geq c_2$$

Έτσι η κρίσιμη περιοχή δίνεται από την σχέση

$$\text{Κ: } 2\lambda_0 \sum_{i=1}^n X_i \leq \chi^2_{2n;1-(\alpha/2)} \quad \text{ή} \quad 2\lambda_0 \sum_{i=1}^n X_i \geq \chi^2_{2n;\alpha/2}$$

όταν

$$2\lambda_0 \sum_{i=1}^n X_i \sim \chi^2_{2n}$$

- Για δεξιά λογοκρισία τύπου I και για τυχαία λογοκρισία η συνάρτηση πιθανοφάνειας δίνεται από τον τύπο

$$L(\lambda) = \prod_{i=1}^n [f(t_i; \lambda)]^{\delta_i} [S(t_i; \lambda)]^{1-\delta_i} = \prod_{i=1}^n (\lambda e^{-\lambda t_i})^{\delta_i} (e^{-\lambda t_i})^{1-\delta_i} = \lambda^\delta e^{-\lambda t}$$

όπου  $\delta = \delta_1 + \delta_2 + \dots + \delta_n$  και  $t = t_1 + t_2 + \dots + t_n$ . Παρατηρούμε ότι

$$\ell(\lambda) = \log L(\lambda) = \delta \log \lambda - \lambda t, \quad \frac{d}{d\lambda} \ell(\lambda) = \frac{\delta}{\lambda} - t, \quad \frac{d^2}{d\lambda^2} \ell(\lambda) = -\frac{\delta}{\lambda^2}$$

οπότε ο ΕΜΠ της παραμέτρου  $\lambda$  δίνεται από τη σχέση

$$\hat{\lambda} = \hat{\lambda}(\mathbf{T}) = \frac{\delta}{\sum_{i=1}^n T_i}$$

Για τον ΕΜΠ  $\hat{\lambda}$  έχουμε ότι  $\hat{\lambda} \sim N(\lambda, (I(\lambda))^{-1})$ . Αφού



$$I(\lambda) = -E\left(\frac{d^2}{d\lambda^2} \ell(\lambda)\right) = E\left(\frac{\delta}{\lambda^2}\right) = \frac{\delta}{\lambda^2} = I_0(\lambda)$$

μπορούμε να γράψουμε ότι

$$\hat{\lambda} \stackrel{a}{\sim} N\left(\lambda, \frac{\lambda^2}{\delta}\right)$$

Η παραπάνω μορφή της ασυμπτωτικής κατανομής του  $\hat{\lambda}$  δεν είναι κατάλληλη για την κατασκευή διαστημάτων εμπιστοσύνης για την παράμετρο  $\lambda$  αφού η διακύμανση εξαρτάται από την (άγνωστη) παράμετρο  $\lambda$ . Έτσι χρησιμοποιούμε ως ασυμπτωτική κατανομή του  $\hat{\lambda}$  την

$$\hat{\lambda} \sim N(\lambda, (I(\hat{\lambda}))^{-1})$$

ή

$$\hat{\lambda} \stackrel{a}{\sim} N\left(\lambda, \frac{\hat{\lambda}^2}{\delta}\right)$$

ή

$$\hat{\lambda} \stackrel{a}{\sim} N\left(\lambda, \frac{\delta}{t^2}\right)$$

Ένα (ασυμπτωτικό) διάστημα εμπιστοσύνης με συντελεστή εμπιστοσύνης 1- $\alpha$  για την παράμετρο  $\lambda$  είναι το

$$\hat{\lambda} \pm z_{\alpha/2} \frac{\sqrt{\delta}}{t}$$

- Για δεξιά λογοκρισία τύπου II η συνάρτηση πιθανοφάνειας δίνεται από τον τύπο

$$L(\lambda) = \frac{n!}{(n-r)!} [e^{-\lambda t_{(r)}}]^{n-r} \prod_{i=1}^r \lambda e^{-\lambda t_{(i)}} = \frac{n!}{(n-r)!} \lambda^r \exp \left\{ -\lambda \left( \sum_{i=1}^r t_{(i)} + (n-r)t_{(r)} \right) \right\}$$

Παρατηρούμε ότι

$$\ell(\lambda) = \log L(\lambda) = \log \frac{n!}{(n-r)!} + r \log \lambda - \lambda \left( \sum_{i=1}^r t_{(i)} + (n-r)t_{(r)} \right)$$

και

$$\frac{d}{d\lambda} \ell(\lambda) = \frac{r}{\lambda} - \left( \sum_{i=1}^r t_{(i)} + (n-r)t_{(r)} \right)$$

επίσης

$$\frac{d^2}{d\lambda^2} \ell(\lambda) = -\frac{r}{\lambda^2}$$

οπότε ο ΕΜΠ της παραμέτρου  $\lambda$  είναι

$$\hat{\lambda} = \hat{\lambda}(\mathbf{T}) = \frac{r}{\sum_{i=1}^r T_{(i)} + (n-r)T_{(r)}} = \frac{\delta}{\sum_{i=1}^n T_i}$$

Παρατήρηση: ο ΕΜΠ έχει την ίδια μορφή με την προηγούμενη περίπτωση της δεξιάς λογοκρισίας τύπου I και της τυχαίας λογοκρισίας. Μπορεί να δειχθεί ότι

$$2r\lambda \sum_{i=1}^n T_i \sim \chi^2_{2r}$$

ή

$$\frac{2r\lambda}{\hat{\lambda}} \sim \chi^2_{2r}$$

οπότε ένα ασυμπτωτικό διάστημα εμπιστοσύνης με συντελεστή εμπιστοσύνης  $1-\alpha$  για την παράμετρο  $\lambda$  είναι το

$$\left[ \frac{\chi^2_{2r,1-(\alpha/2)}}{2\sum_{i=1}^n T_i}, \frac{\chi^2_{2r,\alpha/2}}{2\sum_{i=1}^n T_i} \right] = \left[ \frac{\hat{\lambda} \chi^2_{2r,1-(\alpha/2)}}{2r}, \frac{\hat{\lambda} \chi^2_{2r,\alpha/2}}{2r} \right]$$

## **ΚΕΦΑΛΑΙΟ 2<sup>ο</sup>**

### **ΠΑΡΑΜΕΤΡΙΚΗ ΕΚΤΙΜΗΣΗ ΤΗΣ ΣΥΝΑΡΤΗΣΗΣ ΕΠΙΒΙΩΣΗΣ**

#### **2.1 Εισαγωγή**

Βασιζόμενοι σε ένα τυχαίο λογοκριμένο δείγμα χρόνων ζωής μιας τυχαίας μεταβλητής  $T$ , αντιμετωπίζουμε το πρόβλημα της εκτίμησης (μη παραμετρικής και ημι-παραμετρικής) της συνάρτησης επιβίωσης, της συνάρτησης κινδύνου και της αθροιστικής συνάρτησης κινδύνου της  $T$ . Όμως αν έχουμε πληροφορίες για την κατανομή της τυχαίας μεταβλητής  $T$  τότε ένα παραμετρικό μοντέλο θα έδινε πιο ακριβείς εκτιμήσεις των ποσοτήτων που μας ενδιαφέρουν αφού στα παραμετρικά μοντέλα υπάρχει μικρός αριθμός παραμέτρων που πρέπει να εκτιμηθούν και συνήθως έχουν μικρά τυπικά σφάλματα (με την προϋπόθεση ότι το παραμετρικό μοντέλο έχει καλή προσαρμογή στα δεδομένα).

Μελετάμε τις περιπτώσεις όπου η κατανομή της τυχαίας μεταβλητής  $T$  είναι εκθετική, *Weibull*. Για τα παραμετρικά μοντέλα που μελετάμε δείχνουμε ότι μπορούν να θεωρηθούν και ως γραμμικά μοντέλα του λογάριθμου των χρόνων ζωής (*log linear models*). Ειδικότερα, το μοντέλο *Weibull* μπορεί να θεωρηθεί ως μοντέλο αναλογικού κινδύνου.

## 2.2 Ανασκόπηση παραμετρικών μοντέλων

### 2.2.1. Εκθετική (exponential) κατανομή

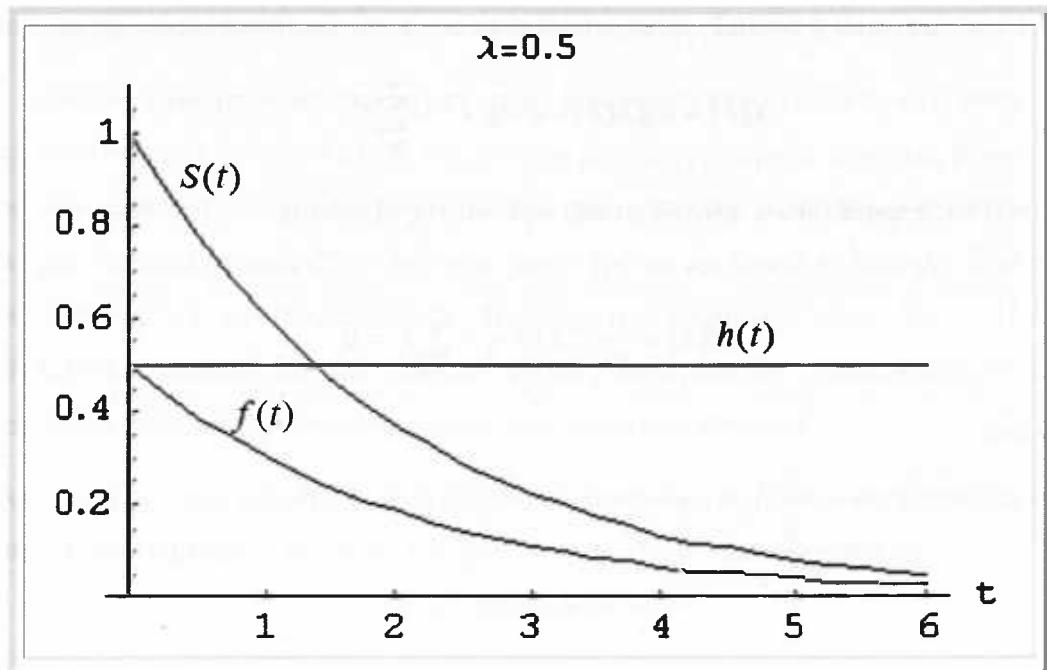
Η συνάρτηση πυκνότητας  $f(t)$ , η συνάρτηση επιβίωσης  $S(t)$  και η συνάρτηση κινδύνου  $h(t)$  της εκθετικής κατανομής με παράμετρο  $\lambda$ , συμβολικά  $Exp(\lambda)$ , έχουν ως ακολούθως

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

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

$$h(t) = \lambda$$

Το κύριο χαρακτηριστικό της εκθετικής κατανομής είναι ότι έχει σταθερή συνάρτηση κινδύνου το οποίο οφείλεται στην ιδιότητα έλλειψης μνήμης της εκθετικής κατανομής. Έτσι μπορούμε να πούμε ότι ένα άτομο με εκθετικό χρόνο ζωής παραμένει “αγέραστο” στο χρόνο αφού η “στιγμαία πιθανότητα θανάτου” είναι σταθερή συνάρτηση του χρόνου, αλλά όχι “αθάνατο” αφού  $h(t) \neq 0$ .



**Διάγραμμα 2.2.1** Απεικόνιση των συναρτήσεων πυκνότητας  $f(t)$ , επιβίωσης  $S(t)$ , κινδύνου  $h(t)$  της εκθετικής κατανομής με παράμετρο  $\lambda$

Ας υποθέσουμε ότι έχουμε στη διάθεσή μας  $n$  παρατηρήσεις της μορφής  $(T_j, \Delta_j)$ ,  $n j$ .  $1 \leq j \leq n$ , από την κατανομή  $Exp(\lambda)$ . Η συνάρτηση πιθανοφάνειας δίνεται από τη σχέση

$$L(\lambda) = \prod_{j=1}^n [f(t_j | \lambda)]^{\delta_j} [S(t_j | \lambda)]^{1-\delta_j} = \prod_{j=1}^n (\lambda e^{-\lambda t_j})^{\delta_j} (e^{-\lambda t_j})^{1-\delta_j} = \prod_{j=1}^n \lambda^{\delta_j} e^{-\lambda t_j}$$

Οπότε

$$\ell(\lambda) = \log L(\lambda) = \log \lambda \sum_{j=1}^n \delta_j - \lambda \sum_{j=1}^n t_j.$$

Θέτοντας  $\sum_{j=1}^n \delta_j = d$  (η ποσότητα  $d$  δηλώνει τον αριθμό των πλήρων χρόνων ζωής του δείγματος), παίρνουμε

$$\ell(\lambda) = \log L(\lambda) = d \log \lambda - \lambda \sum_{j=1}^n t_j.$$

Ο ΕΜΠ της παραμέτρου  $\lambda$  προκύπτει από τη λύση της εξίσωσης

$$U(\lambda) = \frac{d}{d\lambda} \ell(\lambda) = \frac{d}{\lambda} - \sum_{j=1}^n t_j = 0$$

Δηλαδή

$$\hat{\lambda} = \frac{d}{\sum_{j=1}^n t_j} \quad (1)$$

(εναλλακτικά  $\hat{\lambda} = \frac{d}{\sum_{j=1}^n T_j}$ ). Για τον ΕΜΠ  $\hat{\lambda}$  ισχύει ότι

$$\hat{\lambda} \stackrel{a}{=} N(\lambda, (I_0(\lambda))^{-1})$$

αλλά στην πράξη χρησιμοποιείτε η σχέση

$$\hat{\lambda} \stackrel{a}{=} N(\lambda, (I_0(\hat{\lambda}))^{-1}) \quad (2)$$

όπου

$$\frac{d^2}{d\lambda^2} \ell(\lambda) = -\frac{d}{\lambda^2}, \quad I_0(\lambda) = -\frac{d^2}{d\lambda^2} \ell(\lambda) = \frac{d}{\lambda^2}, \quad I_0(\hat{\lambda}) = \frac{d}{\hat{\lambda}^2}.$$

Έτσι η ασυμπτωτική διακύμανση του ΕΜΠ  $\hat{\lambda}$  έχουμε ότι εκτιμάται με τη σχέση

$$\hat{V}(\hat{\lambda}) = \frac{\hat{\lambda}^2}{d}. \quad (3)$$

Θα πρέπει να σημειώσουμε ότι αποφεύγουμε να χρησιμοποιήσουμε τη σχέση (2) για την εύρεση (προσεγγιστικών) διαστημάτων εμπιστοσύνης για την παράμετρο  $\lambda$  αφού

μπορείνα προκύψει αρνητικό κατώτερο όριο διαστήματος. Επίσης η κατανομή του ΕΜΠ

$\hat{\lambda}$  τείνει να είναι ασυμμετρική και συνεπώς η υπόθεση της κανονικής κατανομής δεν μπορεί να δικαιολογηθεί για μικρά και μεσαίου μεγέθους δείγματα. Συνήθως βρίσκουμε διάστημα εμπιστοσύνης για την ποσότητα  $\log \lambda$  που κατανέμεται πιο συμμετρικά και στη συνέχεια “εκθετικοποιούμε” το διάστημα αυτό για να προκύψει κατάλληλο διάστημα εμπιστοσύνης για την παράμετρο  $\lambda$  (διάστημα εμπιστοσύνης τύπου  $\log$ ). Η ίδια μεθοδολογία εφαρμόζεται και για την εύρεση διαστημάτων εμπιστοσύνης για τα ποσοστιαία σημεία της κατανομής η οποία αναλύεται στη συνέχεια.

Από τη στιγμή που ο ΕΜΠ  $\hat{\lambda}$  είναι διαθέσιμος μπορούμε να βρούμε εκτιμήσεις των ποσοστιαίων σημείων  $t_p$ ,  $0 < p < 1$ , των χρόνων ζωής. Λύνοντας την εξίσωση

$$S(t_p) = \exp(-\hat{\lambda} t_p) = 1 - p$$

προκύπτει άμεσα ότι

$$\hat{t}_p = -\frac{\log(1-p)}{\hat{\lambda}}.$$

Για την εκτίμηση της διακύμανσης του  $\hat{t}_p$ , θα εκτιμηθεί πρώτα η διακύμανση του  $\log \hat{t}_p$ , δηλαδή θα εκτιμηθεί η διακύμανση της ποσότητας

$$\log \hat{t}_p = \log[-\log(1-p)] - \log \hat{\lambda} = c_p - \log \hat{\lambda}, \quad c_p = \log[-\log(1-p)]$$

Επειδή η ποσότητα  $\log \hat{t}_p$  είναι συνάρτηση του  $\hat{\lambda}$ , αφού  $\log \hat{t}_p = c_p - \log \hat{\lambda} = g(\hat{\lambda})$ , και ισχύει ότι

$$E(\log \hat{t}_p) = E[g(\hat{\lambda})] \cong g[E(\hat{\lambda})] = g(\lambda) = c_p - \log \lambda = \log t_p$$

$$V(\log \hat{t}_p) = V[g(\hat{\lambda})] \cong V(\hat{\lambda}) \cdot \left( \frac{dg(\hat{\lambda})}{d\hat{\lambda}} \Big|_{\hat{\lambda}=\lambda} \right)^2 = V(\hat{\lambda}) \cdot \left( -\frac{1}{\lambda} \right)^2$$

οπότε

$$\hat{V}(\log \hat{t}_p) = \hat{V}(\hat{\lambda}) \cdot \left(-\frac{1}{\hat{\lambda}}\right)^2 = \frac{\hat{\lambda}^2}{d} \cdot \left(-\frac{1}{\hat{\lambda}}\right)^2 = \frac{1}{d}.$$

Αφού

$$\hat{\lambda} \sim N(\lambda, (I_0(\lambda))^{-1}), \quad \log \hat{t}_p = g(\hat{\lambda}), \quad E(\log \hat{t}_p) \cong \log t_p, \quad \hat{V}(\log \hat{t}_p) = \frac{1}{d}$$

οπότε προκύπτει ότι

$$\log \hat{t}_p \sim N(\log t_p, 1/d)$$

.Συνεπώς ένα προσεγγιστικό διάστημα εμπιστοσύνης για την ποσότητα  $\log t_p$  με

συντελεστή εμπιστοσύνης  $1 - a$  είναι το  $(\hat{se}(\log \hat{t}_p) = \sqrt{\hat{V}(\log \hat{t}_p)} = \frac{1}{\sqrt{d}})$

$$\log \hat{t}_p \pm \frac{Z_{\alpha/2}}{\sqrt{d}}$$

και το αντίστοιχο διάστημα εμπιστοσύνης για την ποσότητα  $t_p$  προκύπτει με εκθετικοποίηση (για να είναι τα άκρα του διαστήματος μη αρνητικοί αριθμοί) και είναι το

$$\hat{t}_p \cdot \exp[\pm z_{\alpha/2} / \sqrt{d}]. \quad (5)$$

Επίσης χρησιμοποιώντας το Παράδειγμα 1.5 προκύπτει άμεσα ότι

$$E(\hat{t}_p) \cong t_p, \quad V(\hat{t}_p) \cong V(\log \hat{t}_p)(t_p)^2$$

Οπότε

$$\hat{V}(\hat{t}_p) = \frac{(\hat{t}_p)^2}{d}.$$

Παρατηρούμε ότι  $\hat{se}(\log \hat{t}_p) = \frac{1}{\sqrt{d}} = (\hat{t}_p)^{-1} \cdot \hat{se}(\hat{t}_p)$  οπότε το προσεγγιστικό διάστημα εμπιστοσύνης για την ποσότητα  $t_p$  με συντελεστή εμπιστοσύνης  $1 - \alpha$  μπορεί να γραφεί στην ισοδύναμη μορφή

$$\hat{t}_p \cdot \exp[\pm z_{\alpha/2}(\hat{t}_p)^{-1} \hat{se}(\hat{t}_p)].$$

### 2.2.2. Κατανομή Weibull

(a) Η εκθετική κατανομή είναι ειδική περίπτωση της κατανομής *Weibull*. Η συνάρτηση πυκνότητας  $f(t)$ , η συνάρτηση επιβίωσης  $S(t)$  και η συνάρτηση κινδύνου  $h(t)$  της κατανομής *Weibull* με παραμέτρους  $\lambda$  (παράμετρος κλίμακας (*scale parameter*)) και  $\alpha$  (παράμετρος μορφής (*shape parameter*)), συμβολισμός  $W(\lambda, \alpha)$ , δίνονται στο ακόλουθο πλαίσιο

$$f(t) = \lambda \alpha t^{\alpha-1} \exp(-\lambda t^\alpha), \quad t \geq 0, \quad \lambda, \alpha > 0$$

$$S(t) = \exp(-\lambda t^\alpha)$$

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

Αν η τυχαία μεταβλητή  $X \sim Exp(\lambda)$  τότε μπορεί εύκολα να δειχθεί ότι η τυχαία μεταβλητή  $T = X^\alpha$  με  $0 > \alpha > 1$  ακολουθεί την κατανομή  $W(\lambda, \alpha)$ . Έτσι, για  $\alpha=1$  η κατανομή  $W(\lambda, \alpha)$

ανάγεται στην εκθετική κατανομή με παράμετρο  $\lambda$ , δηλαδή  $W(\lambda, 1) = \text{Exp}(\lambda)$ . Η κατανομή της τυχαίας μεταβλητής  $Y = \log T$ , όπου  $T \sim W(\lambda, \alpha)$ , έχει συνάρτηση πυκνότητας

$$f_Y(y) = f_T(e^y) \cdot e^y = \lambda \alpha \exp(y\alpha - \lambda e^{y\alpha}), \quad -\infty < y < \infty$$

η οποία μπορεί να γραφεί στη μορφή

$$f_Y(y) = \sigma^{-1} \exp\left[\frac{y-\mu}{\sigma} - \exp\left(\frac{y-\mu}{\sigma}\right)\right], \quad -\infty < y < \infty$$

όπου  $\mu = -(\log \lambda)/\alpha$  και  $\sigma = \alpha^{-1}$ . Γράφοντας την  $Y$  στην μορφή  $Y = \mu + \sigma W$ , οπότε  $W = (Y - \mu)/\sigma$ , προκύπτει ότι η συνάρτηση πυκνότητας της τυχαίας μεταβλητής  $W$  είναι ίση με

$$f_W(w) = \exp(w - e^w), \quad -\infty < w < \infty$$

δηλαδή η τυχαία μεταβλητή  $W$  ακολουθεί την *standard extreme value* κατανομή. Συνεπώς για την κατανομή της τυχαίας μεταβλητής  $T$  μπορούμε να γράψουμε  $T = \exp(\mu + \Sigma w)$

(β) Σε μερικές περιπτώσεις, η κατανομή Weibull εκφράζεται, εκτός της μιας παραμέτρου, με δύο (two-parameter) και με τρεις παραμέτρους (three-parameter). Σε μερικές περιπτώσεις, η κατανομή Weibull με τρεις-παραμέτρους παρέχει μια καλύτερη προσαρμογή δεδομένων από την Weibull κατανομή δύο παραμέτρων.

Η διαφορά στις δύο κατανομές είναι η παράμετρος θέσης  $a$ , η οποία μετατοπίζει τη κατανομή κατά μήκος του  $X$ -άξονα. Εξ ορισμού, υπάρχει μια μηδενική πιθανότητα αποτυχίας για  $x < a$ . Αν και δεν είναι σύνηθες, η θέση μπορεί να είναι αρνητική και αυτό υπονοεί ότι τα στοιχεία ήταν αποτυχημένα πριν από τη δοκιμή.

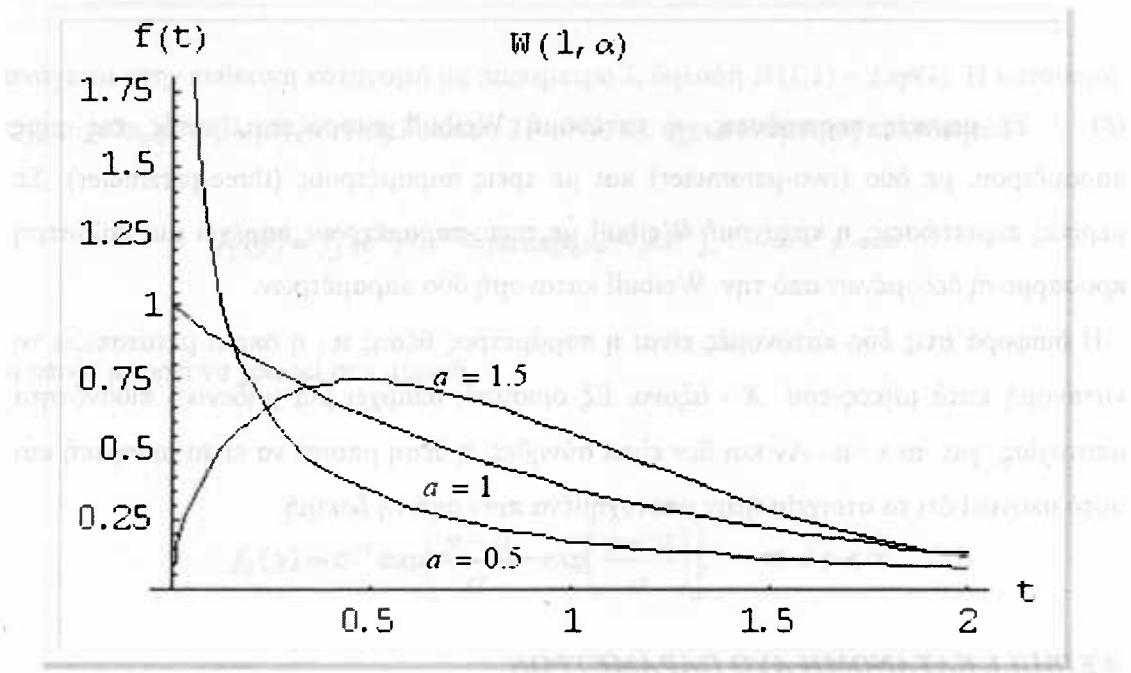
### WEIBULL ΚΑΤΑΝΟΜΗ ΔΥΟ ΠΑΡΑΜΕΤΡΩΝ

$$f(x) = \frac{\beta x^{\beta-1}}{\theta^\beta} \exp\left(-\frac{x}{\theta}\right)^\beta, \quad x \geq 0$$

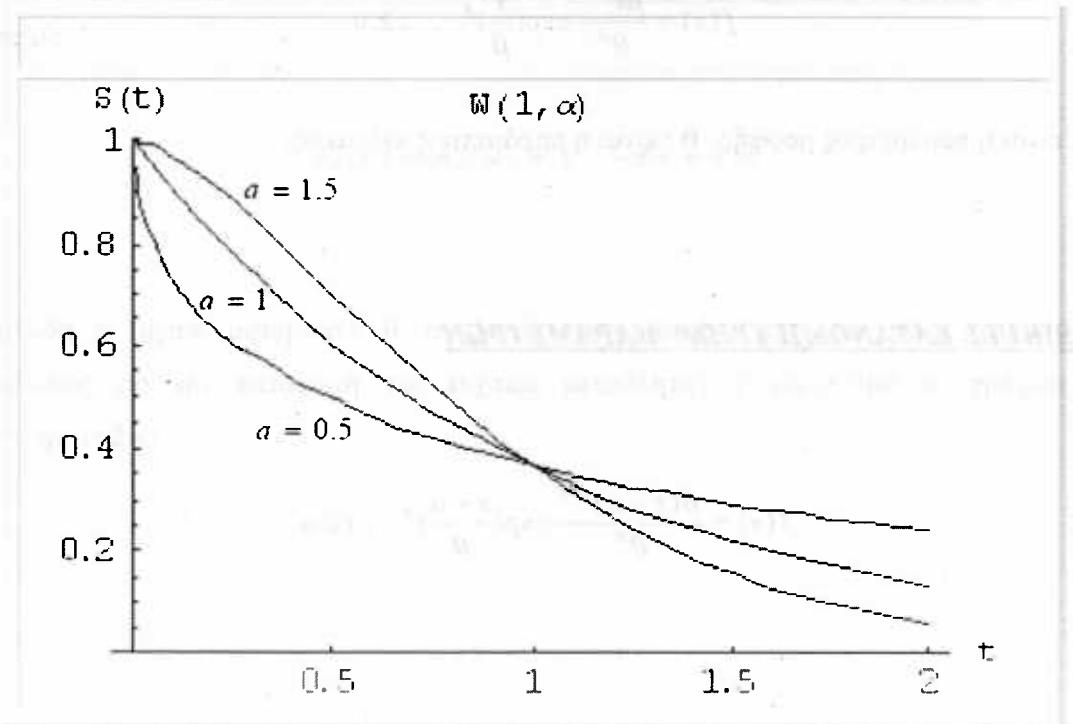
$\beta$  : είναι η παράμετρος μορφής,  $\theta$  : είναι η παράμετρος κλίμακας.

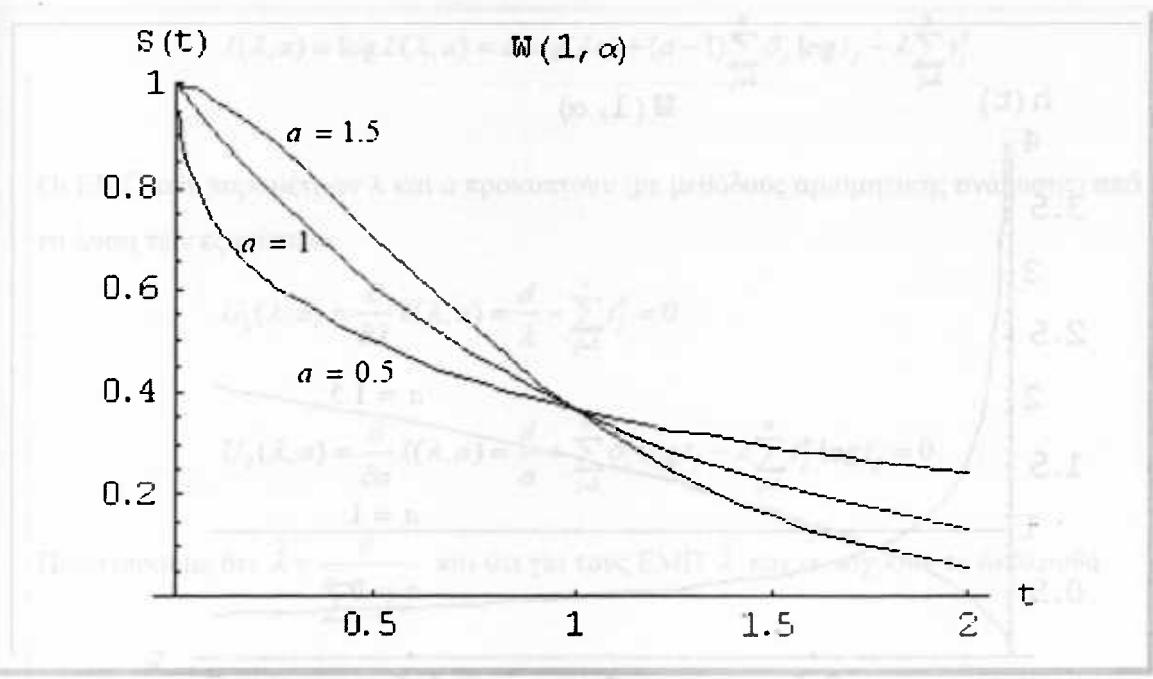
### WEIBULL ΚΑΤΑΝΟΜΗ ΤΡΙΩΝ ΠΑΡΑΜΕΤΡΩΝ

$$f(x) = \frac{\beta(x-\alpha)^{\beta-1}}{\theta^\beta} \exp\left(-\frac{x-\alpha}{\theta}\right)^\beta, \quad x \geq a$$



**Διάγραμμα 2.2.2** Απεικόνιση της συνάρτησης πυκνότητας  $f(t)$  της κατανομής *Weibull* με παραμέτρο  $\lambda$ , για διάφορες τιμές της  $\alpha$  (παράμετρος μορφής)



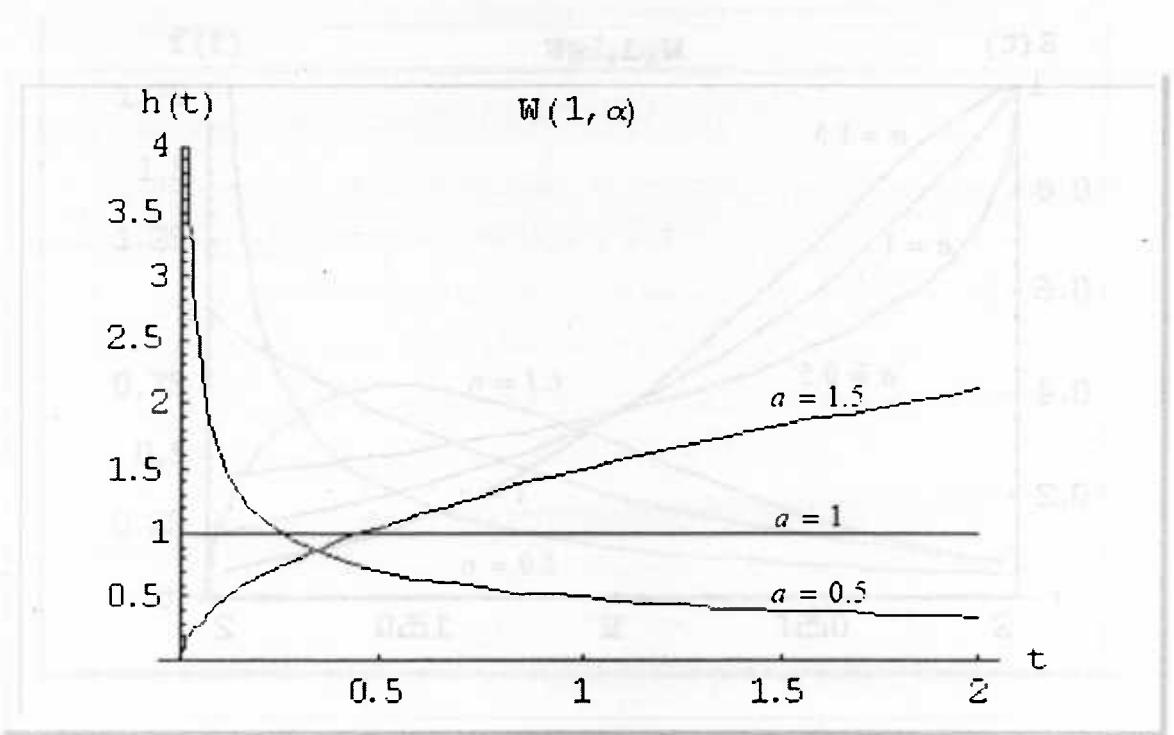


**Διάγραμμα 2.2.3** Απεικόνιση της συνάρτησης πυκνότητας  $S(t)$  της κατανομής *Weibull* με παραμέτρο  $\lambda$ , για διάφορες τιμές της  $\alpha$  (παράμετρος μορφής)

Το μοντέλο της κατανομής  $W(\lambda, \alpha)$  χρησιμοποιείται συχνά στην πράξη λόγω του ευπροσάρμοστου και των ιδιοτήτων της συνάρτησης κινδύνου της. Η συνάρτηση κινδύνου της κατανομής  $W(\lambda, \alpha)$  είναι:

- αύξουσα για  $\alpha > 1$ ,
- φθίνουσα για  $0 < \alpha < 1$ ,
- σταθερή για  $\alpha = 1$ .

Στις εφαρμογές η τιμή της παραμέτρου  $\alpha$  κυμαίνεται συνήθως από 1 έως 3.



**Διάγραμμα 2.2.4** Απεικόνιση της συνάρτησης πυκνότητας  $h(t)$  της κατανομής Weibull με παραμέτρο  $\lambda$ , για διάφορες τιμές της  $\alpha$  (παράμετρος μορφής)

Ας υποθέσουμε ότι έχουμε στη διάθεσή μας  $n$  παρατηρήσεις της μορφής  $(T_j, \Delta_j)$ ,  $1 \leq j \leq n$ , από την κατανομή  $W(\lambda, \alpha)$ . Η συνάρτηση πιθανοφάνειας δίνεται από τη σχέση

$$L(\lambda, \alpha) = \prod_{j=1}^n [f(t_j | \lambda, \alpha)]^{\delta_j} [S(t_j | \lambda, \alpha)]^{1-\delta_j} = \prod_{j=1}^n [\lambda \alpha t_j^{\alpha-1} \exp(-\lambda t_j^\alpha)]^{\delta_j} [\exp(-\lambda t_j^\alpha)]^{1-\delta_j}$$

άρα

$$\ell(\lambda, \alpha) = \log L(\lambda, \alpha) = \log(\lambda \alpha) \sum_{j=1}^n \delta_j + (\alpha - 1) \sum_{j=1}^n \delta_j \log t_j - \lambda \sum_{j=1}^n t_j^\alpha.$$

Θέτοντας  $\sum_{j=1}^n \delta_j = d$ , έχουμε

$$\ell(\lambda, a) = \log L(\lambda, a) = d \log(\lambda a) + (a-1) \sum_{j=1}^n \delta_j \log t_j - \lambda \sum_{j=1}^n t_j^a.$$

Οι ΕΜΠ των παραμέτρων  $\lambda$  και  $a$  προκύπτουν (με μεθόδους αριθμητικής ανάλυσης) από τη λύση των εξισώσεων

$$U_1(\lambda, a) = \frac{\partial}{\partial \lambda} \ell(\lambda, a) = \frac{d}{\lambda} - \sum_{j=1}^n t_j^a = 0$$

$$U_2(\lambda, a) = \frac{\partial}{\partial a} \ell(\lambda, a) = \frac{d}{a} + \sum_{j=1}^n \delta_j \log t_j - \lambda \sum_{j=1}^n t_j^a \log t_j = 0.$$

Παρατηρούμε ότι  $\hat{\lambda} = \frac{d}{\sum_{j=1}^n t_j^a}$  και ότι για τους ΕΜΠ  $\hat{\lambda}$  και  $\hat{a}$  ισχύουν τα ακόλουθα

$$(\hat{\lambda}, \hat{a}) \stackrel{d}{=} N_2((\lambda, a), \mathbf{I}_0^{-1}(\hat{\lambda}, \hat{a})), \quad \hat{\lambda} \stackrel{d}{=} N(\lambda, I_0^{11}(\hat{\lambda}, \hat{a}))$$

$$\hat{a} \stackrel{d}{=} N(a, I_0^{22}(\hat{\lambda}, \hat{a})) \quad (7)$$

όπου ο πίνακας  $I_0(\hat{\lambda}, \hat{a})$ , που αποτελεί εκτίμηση του πίνακα διακυμάνσεων-συνδιακυμάνσεων του διανύσματος  $(\hat{\lambda}, \hat{a})$  και τα στοιχεία  $I_0^{11}(\hat{\lambda}, \hat{a})$  και  $I_0^{22}(\hat{\lambda}, \hat{a})$  προκύπτουν από τις σχέσεις

$$\mathbf{I}_0(\hat{\lambda}, \hat{a}) = \begin{bmatrix} I_{0,11}(\hat{\lambda}, \hat{a}) & I_{0,12}(\hat{\lambda}, \hat{a}) \\ I_{0,21}(\hat{\lambda}, \hat{a}) & I_{0,22}(\hat{\lambda}, \hat{a}) \end{bmatrix}, \quad \mathbf{I}_0^{-1}(\hat{\lambda}, \hat{a}) = \begin{bmatrix} I_0^{11}(\hat{\lambda}, \hat{a}) & I_0^{12}(\hat{\lambda}, \hat{a}) \\ I_0^{21}(\hat{\lambda}, \hat{a}) & I_0^{22}(\hat{\lambda}, \hat{a}) \end{bmatrix}$$

με

$$I_{0,11}(\hat{\lambda}, \hat{a}) = \left( -\frac{\partial^2}{\partial \lambda^2} \ell(\lambda, a) \right) \Big|_{\lambda=\hat{\lambda}, a=\hat{a}}, \quad I_{0,22}(\hat{\lambda}, \hat{a}) = \left( -\frac{\partial^2}{\partial a^2} \ell(\lambda, a) \right) \Big|_{\lambda=\hat{\lambda}, a=\hat{a}},$$

$$I_{0,12}(\hat{\lambda}, \hat{a}) = I_{0,21}(\hat{\lambda}, \hat{a}) = \left( -\frac{\partial^2}{\partial \lambda \partial a} \ell(\lambda, a) \right) \Big|_{\lambda=\hat{\lambda}, a=\hat{a}}$$

$$(\hat{V}(\hat{\lambda}) = I_0^{11}(\hat{\lambda}, \hat{a}), \quad \hat{V}(\hat{a}) = I_0^{22}(\hat{\lambda}, \hat{a}), \quad \text{Cov}(\hat{\lambda}, \hat{a}) = I_0^{12}(\hat{\lambda}, \hat{a}) = I_0^{21}(\hat{\lambda}, \hat{a}))$$

Από τη στιγμή που οι ΕΜΠ  $\hat{\lambda}$  και  $\hat{a}$  είναι διαθέσιμοι μπορούμε να βρούμε εκτιμήσεις των ποσοστιαίων σημείων  $t_p$ ,  $0 < p < 1$ , των χρόνων ζωής. Λύνοντας την εξίσωση

$$S(t_p) = \exp(-\lambda t_p^\alpha) = 1 - p$$

προκύπτει άμεσα ότι

$$\hat{t}_p = \left( -\frac{\log(1-p)}{\hat{\lambda}} \right)^{1/\alpha}$$

Για την εκτίμηση της διακύμανσης του  $\hat{t}_p$ , θα εκτιμηθεί πρώτα η διακύμανση του  $\log \hat{t}_p$ , δηλαδή η διακύμανση της ποσότητας

$$\log \hat{t}_p = \frac{1}{\hat{a}} (\log[-\log(1-p)] - \log \hat{\lambda}) = \frac{1}{\hat{a}} (c_p - \log \hat{\lambda}), \quad c_p = \log[-\log(1-p)]$$

Επειδή η ποσότητα  $\log \hat{t}_p$  είναι μια συνάρτηση της μορφής  $g(\hat{\lambda}, \hat{a})$  και ισχύει ότι το διάνυσμα  $(\hat{\lambda}, \hat{a})$  ακολουθεί ασυμπτωτικά διδιάστατη κανονική κατανομή, προκύπτει ότι η κατανομή του  $\log \hat{t}_p$  είναι ασυμπτωτικά κανονική με μέσο

$$E(\log \hat{t}_p) = E[g(\hat{\lambda}, \hat{a})] \cong g(E(\hat{\lambda}), E(\hat{a})) = g(\lambda, a) = \log t_p$$

Για την διακύμανση του  $\log \hat{t}_p$  έχουμε ότι ασυμπτωτικά ικανοποιεί τη σχέση

$$\begin{aligned} V(\log \hat{t}_p) &= V[g(\hat{\lambda}, \hat{a})] \cong \left( \frac{\partial g(\hat{\lambda}, \hat{a})}{\partial \hat{\lambda}} \Big|_{\hat{\lambda}=\lambda, \hat{a}=a} \right)^2 \cdot V(\hat{\lambda}) + \left( \frac{\partial g(\hat{\lambda}, \hat{a})}{\partial \hat{a}} \Big|_{\hat{\lambda}=\lambda, \hat{a}=a} \right)^2 \cdot V(\hat{a}) \\ &\quad + 2 \cdot \left( \frac{\partial g(\hat{\lambda}, \hat{a})}{\partial \hat{\lambda}} \Big|_{\hat{\lambda}=\lambda, \hat{a}=a} \right) \cdot \left( \frac{\partial g(\hat{\lambda}, \hat{a})}{\partial \hat{a}} \Big|_{\hat{\lambda}=\lambda, \hat{a}=a} \right) \cdot Cov(\hat{\lambda}, \hat{a}) \end{aligned}$$

Επίσης

$$\frac{\partial g(\hat{\lambda}, \hat{a})}{\partial \hat{\lambda}} \Big|_{\hat{\lambda}=\lambda, \hat{a}=a} = -\frac{1}{\lambda a}, \quad \frac{\partial g(\hat{\lambda}, \hat{a})}{\partial \hat{a}} \Big|_{\hat{\lambda}=\lambda, \hat{a}=a} = -\frac{1}{a^2} (c_p - \log \lambda)$$

προκύπτει ότι

$$V(\log \hat{t}_p) = \frac{1}{(\lambda a)^2} \cdot V(\hat{\lambda}) + \frac{(c_p - \log \lambda)^2}{a^4} \cdot V(\hat{a}) + \frac{2(c_p - \log \lambda)}{\lambda a^3} \cdot Cov(\hat{\lambda}, \hat{a})$$

οπότε

$$\begin{aligned}\hat{V}(\log \hat{t}_p) &= \frac{1}{\hat{\lambda}^2 \hat{a}^4} [\hat{a}^2 \hat{V}(\hat{\lambda}) + \hat{\lambda}^2 (c_p - \log \hat{\lambda})^2 \hat{V}(\hat{a}) + 2\hat{\lambda}\hat{a}(c_p - \log \hat{\lambda}) \cdot \text{Cov}(\hat{\lambda}, \hat{a})] \\ &= \frac{1}{\hat{\lambda}^2 \hat{a}^4} [\hat{a}^2 I_0^{11}(\hat{\lambda}, \hat{a}) + \hat{\lambda}^2 (c_p - \log \hat{\lambda})^2 I_0^{22}(\hat{\lambda}, \hat{a}) + 2\hat{\lambda}\hat{a}(c_p - \log \hat{\lambda}) \cdot I_0^{12}(\hat{\lambda}, \hat{a})]\end{aligned}$$

Συνεπώς ένα προσεγγιστικό διάστημα εμπιστοσύνης για την ποσότητα  $\log t_p$  με

$$\text{συντελεστή εμπιστοσύνης } 1-\alpha \text{ είναι το } \hat{se}(\log \hat{t}_p) = \sqrt{\hat{V}(\log \hat{t}_p)}$$

$$\log \hat{t}_p \pm z_{\alpha/2} \hat{se}(\log \hat{t}_p)$$

και το αντίστοιχο διάστημα εμπιστοσύνης για την ποσότητα  $t_p$  προκύπτει με εκθετικοποίηση και είναι το

$$\hat{t}_p \exp[\pm z_{\alpha/2} \hat{se}(\log \hat{t}_p)]$$

Χρησιμοποιώντας τη σχέση  $V(\hat{t}_p) \cong (\hat{t}_p)^2 \cdot V(\log \hat{t}_p)$  παίρνουμε άμεσα ότι μια εκτίμηση της ποσότητας  $V(\hat{t}_p)$  δίνεται από τη σχέση

$$\hat{V}(\hat{t}_p) = \frac{(\hat{t}_p)^2}{\hat{\lambda}^2 \hat{a}^4} [\hat{a}^2 I_0^{11}(\hat{\lambda}, \hat{a}) + \hat{\lambda}^2 (c_p - \log \hat{\lambda})^2 I_0^{22}(\hat{\lambda}, \hat{a}) + 2\hat{\lambda}\hat{a}(c_p - \log \hat{\lambda}) \cdot I_0^{12}(\hat{\lambda}, \hat{a})]. \quad (9)$$

Παρατηρούμε ότι το  $\hat{se}(\log \hat{t}_p) = (\hat{t}_p)^{-1} \hat{se}(\hat{t}_p)$  οπότε το προσεγγιστικό διάστημα εμπιστοσύνης για την ποσότητα  $t_p$  με συντελεστή εμπιστοσύνης  $1-\alpha$  μπορεί να γραφεί στην ισοδύναμη μορφή

$$\hat{\epsilon}_p \cdot \exp[\pm z_{\gamma_2}(\hat{\epsilon}_p)^{-1} se(\hat{\epsilon}_p)]$$

## 2.3 ΓΕΝΙΚΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ

1. Ένα τυχαίο δείγμα των ατόμων ν με διάρκειας ζωής  $T_1, \dots, T_n$  από έναν πληθυσμό με σ.π.π.  $f(t)$  και συνάρτηση επιβίωσης  $S(t)$
2. Συνδεμένος με κάθε άτομο, ένας σταθερός χρόνος λογοκρισίας  $L_i > 0$
3. Για κάθε άτομο η παρατήρηση είναι  $(t_i, \delta_i)$ ,  $i = 1 .. 2. \dots, n$ , όπου

$$t_i = \min(T_i, L_i) \quad \text{and} \quad \delta_i = \begin{cases} 1 & \text{if } t_i = T_i \\ 0 & \text{if } t_i = L_i \end{cases}$$

4. Παρατηρούμε για την κατανομή Weibull

$$S(t) = \exp(-(\frac{t}{\theta})^\alpha)$$

Επομένως,

$$\log S(t) = -(\frac{t}{\theta})^\alpha \Rightarrow \log(-\log S(t)) = -\alpha \log \theta + \alpha \log t$$

5. Η συνάρτηση πιθανοφάνειας, βασισμένη σε αυτό το δείγμα είναι

$$\begin{aligned} l &= \sum_{i \in D} \log h(t_i) + \sum_{i=1}^n \log S(t_i) \\ &= \sum_{i \in D} \log h(t_i) - \sum_{i=1}^n H(t_i) = \\ &= r \log \alpha - \alpha r \log \theta + (\alpha - 1) \sum_{i \in D} \log t_i - \sum_{i=1}^n \left(\frac{t_i}{\theta}\right)^\alpha \end{aligned}$$

όπου το D είναι το σύνολο περιέχει όλες τις παρατηρούμενες αποτυχίες

**6.** Οι εκτιμητές μέγιστης πιθανότητας  $\hat{\theta}$  και  $\hat{\alpha}$  είναι εκείνες οι τιμές των  $\theta$  και  $\alpha$  που μεγιστοποιούν την πιθανότητα της λογαριθμισμένης πιθανοφάνειας

## 7. Score functions

$$u_{\theta} = \frac{\partial l}{\partial \theta} \quad \text{και} \quad u_{\alpha} = \frac{\partial l}{\partial \alpha}$$

**8.** Με δεδομένο  $\alpha$ , υπολογίζουμε το  $\hat{\theta}(\alpha)$ , δηλ.., βρίσκει το  $\theta$  έτσι ώστε  $u_{\theta} = 0$

$$-\frac{\alpha r}{\theta} + \frac{\lambda}{\theta^{\lambda+1}} \sum_{i=1}^n t_i^\alpha \Rightarrow \hat{\theta}(\alpha) = \left( \frac{\sum_{i=1}^n t_i^\alpha}{r} \right)^{\frac{1}{\alpha}}$$

Αντικαθιστούμε  $\hat{\theta}(\alpha)$  στο  $u_{\alpha}$

$$u_{\alpha}(\hat{\lambda}(\alpha), \alpha) = \frac{r}{\alpha} - r \log \hat{\theta}(\alpha) + \sum_{i \in D} \log t_i - \sum_{i=1}^n \left( \frac{t_i}{\hat{\theta}(\alpha)} \right)^{\alpha} \log \left( \frac{t_i}{\hat{\theta}(\alpha)} \right)$$

Θέτουμε την παραπάνω σχέση ίση με μηδέν και λύνοντας ως προς  $\alpha$ , έχω

$$\frac{r}{\alpha} + \sum_{i \in D} \log t_i - \left( \frac{r}{\sum_{i=1}^n t_i^\alpha} \right) \sum_{i=1}^n t_i^\alpha \log(t_i) = 0 \quad (*)$$

- Πρέπει να λυθεί η  $(*)$  αριθμητικά για την εύρεση  $\hat{\alpha}$
- Με δεδομένο την τιμή του  $\hat{\alpha}$ , αντικαθιστούμε στο  $\hat{\lambda}(\alpha)$  και θέτουμε  $\hat{\lambda} =$

$$= \hat{\lambda}(\hat{\alpha})$$

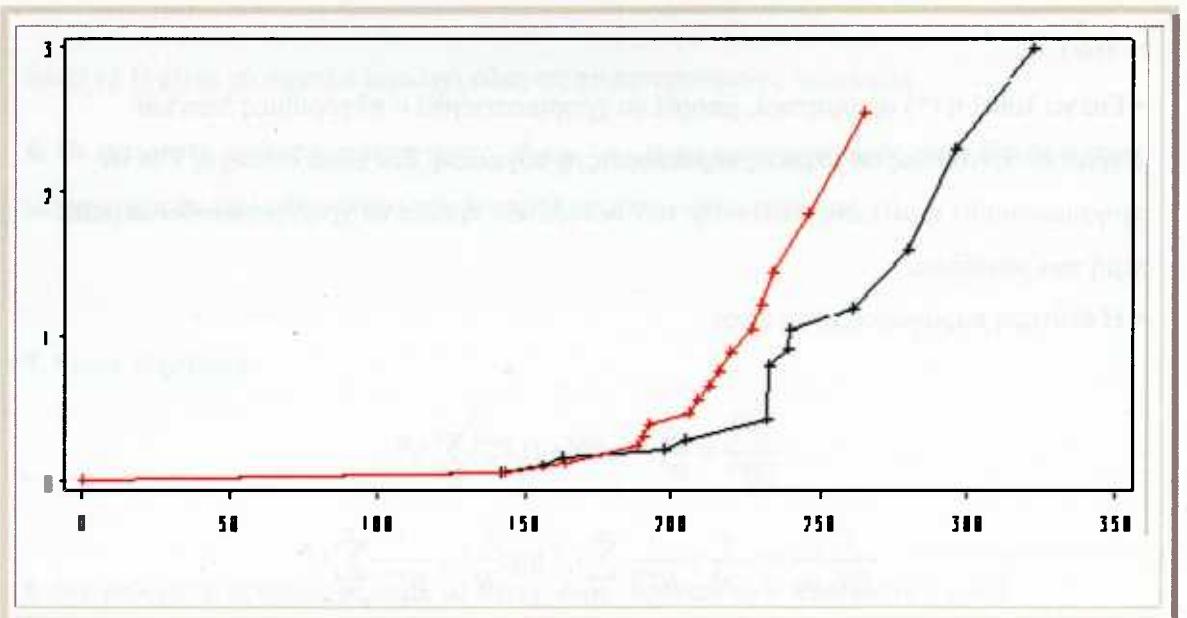
- Για να λυθεί η (\*) αριθμητικά, μπορεί να χρησιμοποιηθεί ο αλγόριθμος Newton-Raphson. Εντούτοις σε μερικές περιπτώσεις, η σύγκλιση δεν είναι σταθερή. Για να χρησιμοποιηθεί η μέθοδος ανάλυσης των δεδομένων πρέπει να χρησιμοποιηθεί ακραία τιμή του μοντέλου.
- Η δεύτερη παράγωγος του  $l$  είναι

$$\begin{aligned}\frac{\partial^2 l}{\partial \theta^2} &= \frac{\alpha r}{\theta^2} - \alpha(\alpha+1)\lambda^{\alpha+2} \sum_{i=1}^n t_i^{-\alpha} \\ \frac{\partial^2 l}{\partial \theta \cdot \partial \alpha} &= -\frac{r}{\theta} + \frac{1}{\theta^{\alpha+1}} \sum_{i=1}^n t_i^{-\alpha} \log\left(\frac{t_i}{\theta}\right) + \frac{1}{\theta^{\alpha+1}} \sum_{i=1}^n t_i^{-\alpha} \\ \frac{\partial^2 l}{\partial \alpha^2} &= -\frac{r}{\alpha^2} - \frac{1}{\theta^\alpha} \sum_{i=1}^n t_i^{-\alpha} (\log\left(\frac{t_i}{\theta}\right))^2\end{aligned}$$

## ΠΑΡΑΔΕΙΓΜΑ ΣΥΓΚΡΙΣΗΣ ΕΚΘΕΤΙΚΟΥ ΜΟΝΤΕΛΟΥ ΚΑΙ ΠΡΟΤΥΠΟΥ WEIBULL 2.3.1

### (a) ΕΚΘΕΤΙΚΟ ΠΡΟΤΥΠΟ

Μπορούμε να σχεδιάσουμε τον αρνητικό λογάριθμο επιβίωσης, δύο ομάδων ποντικών που έχουν εκτεθεί σε μεγάλη ποσότητα ραδιενέργειας, σε σχέση με τον χρόνο αποτυχίας (αριθμός ημερών πριν τον θάνατό τους)



$t_i$ : ημέρες πριν τον θάνατο ποντικών

----- group 1 ----- group 2

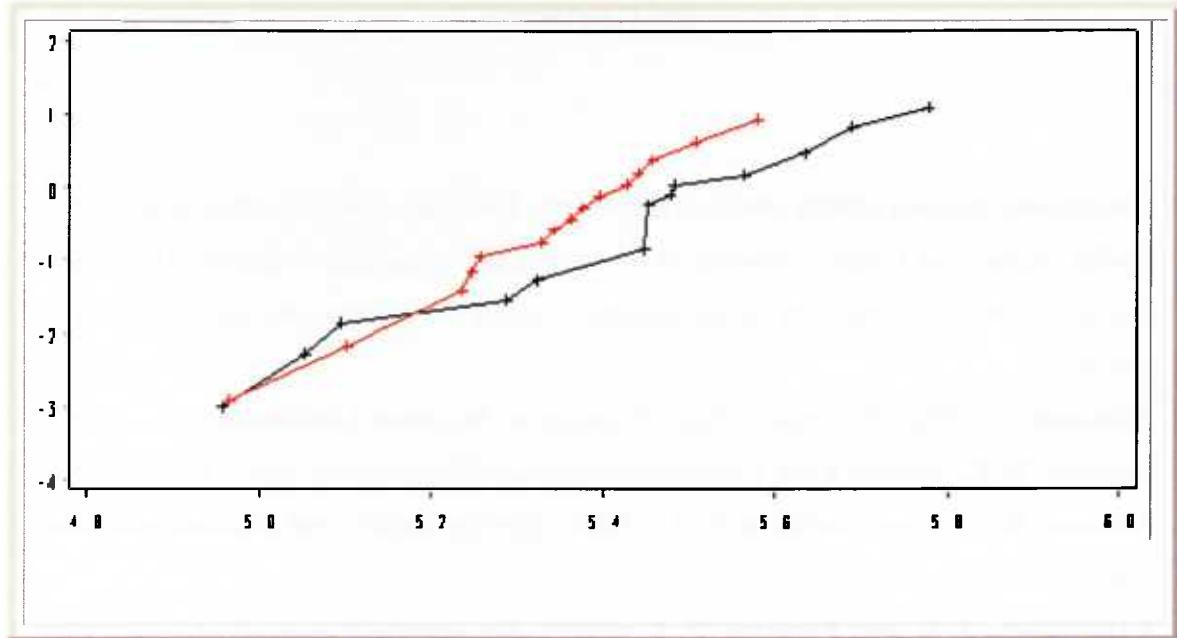
Διάγραμμα 2.3.1  $-\log(\hat{S}(t_i))$  vs  $t_i$

Από την EIKONA 1, φαίνεται ότι το εκθετικό πρότυπο δεν είναι το μόνο κατάλληλο.

Δοκιμάζουμε και το πρότυπο Weibull

### (β) ΠΡΟΤΥΠΟ WEIBULL

Σχεδιάζουμε τον λογάριθμο του αρνητικού λογαρίθμου επιβίωσης των δύο ομάδων ποντικών, σε σχέση με τον λογαριθμισμένο χρόνο αποτυχίας. Το πρότυπο Weibull μπορεί να είναι πιο κατάλληλο



$\log t_i$ : λογαριθμισμένος αριθμός ημερών πριν τον θάνατο ποντικών  
 ----- group 1    ----- group 2

Διάγραμμα 2.3.2  $\log(-\log(\hat{S}(t_i)))$  vs  $\log t_i$

Χρησιμοποιούμε το πρότυπο Weibull για καλήτερη προσαρμογή των δεδομένων

## ΑΝΑΦΟΡΕΣ

- Αικατερίνη Δημάκη (2006). *Ανάλυση Επιβίωσης, Οικονομικό Πανεπιστήμιο Αθηνών*
- Darko Stefanovic (2003). Department of Electrical Engineering Princeton University. *Analytical models of memory object lifetimes. Article found from the site of Princeton University*
- Johansen, S (1978). *The Product Limit Estimator as Maximum Likelihood*
- Johnson, N. L., Samuel Kotz. *Continuous Univariate Distributions, Vol. 2.* 67-110
- Johnson, R. C. E. and Johnson, N. L. (1980). *Survival models and data analysis*, John Wiley, 27-49
- Kalbfleisch, J. D. and Prentice, R. L. (1980). *The statistical analysis of failure time data*. John Wiley, New York
- Klein, J. P. and Moeschberger, M. L. (1997). *Survival analysis: Techniques for censored and truncated data*, Springer Verlag. 56, 123-180
- Lawless, J. F. (1982). *Statistical models & methods for lifetime data*, John Wiley, New York. 55,68-90
- Lee, E. T. (1992). *Statistical Methods for Survival Data Analysis*. New York. 7, 98-125
- London, D. (1997). *Survival models and their estimation*, Actex Publications, Winsted, Connecticut
- T. Menhaj (1999) *Stochastic Systems*, University of Massachusetts Amherst, MA 01003. 19-87
- Miller, R. J., Gong, G. and Munoz, A. (1981). *Survival analysis*, John Wiley, New York.
- .
- .



Δωρεά

