**Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο  Π Ε Ι Ρ Α Ι Ω Σ**

**ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ**
**ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ**

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ**
**ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

# ΣΤΑΤΙΣΤΙΚΗ ΜΕΛΕΤΗ ΠΡΟΒΛΗΜΑΤΩΝ ΜΕ ΜΗ ΠΛΗΡΗ ΔΕΔΟΜΕΝΑ

Νικόλαος Σ. Καποδίστριας

Διπλωματική Εργασία
που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής
Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των
απαιτήσεων για την απόκτηση του Μεταπτυχιακού
Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Οκτώβριος 2007

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ... ... ..
συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική
Τα μέλη της Επιτροπής ήταν:
- Καθηγητής Μ. Κούτρας (Επιβλέπων)
- Αναπληρωτής Καθηγητής Ε. Χατζηκωνσταντινίδης
- Λέκτορας Γ. Τζαβελάς
Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

**U N I V E R S I T Y  O F  P I R A E U S**

**DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE**

**POSTGRADUATE PROGRAM IN APPLIED STATISTICS**

# STATISTICAL ANALYSIS WITH INCOMPLETE DATA SETS

By
Nikolaos S. Kapodistrias

MSc Dissertation
submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
October 2007

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

*Στους γονείς μου*
*Σπύρο και Κρυσταλλία*

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

# Κατάλογος Πινάκων


| | | |
|---|---|---|
| 1.5.1 | Μονότονα-Μη μονότονα δεδομένα | 15 |
| 2.1.1 | Πλήρες δείγμα | 19 |
| 2.1.2 | Εκτιμήσεις παραμέτρων για το πλήρες δείγμα | 19 |
| 2.1.3 | Ελλιπές δείγμα | 20 |
| 2.1.4 | Εκτιμήσεις παραμέτρων για το ελλιπές δείγμα | 20 |
| 2.3.1 | Μέθοδοι απλής αντικατάστασης | 32 |
| 2.3.2 | Ιστόγραμμα συχνοτήτων ύστερα από την εφαρμογή της μεθόδου αντικατάστασης μέσης τιμής | 32 |
| 5.2.1 | Δεδομένα παραδείγματος *Freireich* 1963 | 96 |
| 5.2.2 | Εκτιμήτρια Kaplan-Meier για την ομάδα *Treatment* | 97 |
| 5.2.3 | Διαστήματα εμπιστοσύνης 95% τύπου *plain* με τον τύπο του Greenwood για τη συνάρτηση επιβίωσης | 102 |
| 5.4.1 | Πίνακες επιβίωσης με χρήση του SPSS | 107 |


ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

# Κατάλογος Σχημάτων


5.4.1 Εκτιμήτρια Kaplan-Meier 109
5.4.2 95% διάστημα εμπιστοσύνης τύπου *plain* για την ομάδα *Treatment* 110
5.4.3 95% διάστημα εμπιστοσύνης τύπου *log-log* για την ομάδα *Treatment* 110
5.4.4 95% διάστημα εμπιστοσύνης τύπου *plain* με τον τύπο του *Greenwood* 112


ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

# Κατάλογος Συντομογραφιών

| | |
|---|---|
| τ.μ. | τυχαία μεταβλητή |
| σ.π.π. | συνάρτηση πυκνότητας πιθανότητας |
| σ.κ. | συνάρτηση κατανομής |
| ΔΕΤ | Δείκτης Ελλιπών Τιμών |
| ε.σ. | επίπεδο σημαντικότητας |
| δ.ε. | διάστημα εμπιστοσύνης |
| AC | Available Case Analysis |
| CC | Complete Case Analysis |
| MCAR | Missing Completely At Random |
| MAR | Missing At Random |
| MNAR | Missing Not At Random |
| NINR | Non Ignorable Non Response |

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

# Περιεχόμενα

**Κατάλογος Πινάκων** **vii**
**Κατάλογος Σχημάτων** **ix**
**Κατάλογος Συντομογραφιών** **xi**

**Εισαγωγή** 1

**1. Ελλιπείς τιμές, ΔΕΤ και Μηχανισμοί ελλιπών τιμών**
1.1 Ελλιπής τιμή και λογοκριμένα δεδομένα 5
1.2 Μηχανισμοί που οδηγούν σε ελλιπείς τιμές 6
1.3 Δείκτης ελλιπών τιμών 9
1.4 Ορισμός βασικών μηχανισμών ελλιπών τιμών 10

**2. Απλές μέθοδοι αντιμετώπισης ελλιπών τιμών**
2.1 Απλές τεχνικές χειρισμού ελλιπών τιμών 17
2.1.1 Ανάλυση με χρήση μόνο των πλήρων περιπτώσεων 17
2.1.2 Ανάλυση με χρήση των απαραίτητων κάθε φορά πλήρων δεδομένων 21
2.1.3 Ανάλυση με τη χρήση βαρών 22
2.2 Διαδικασία επαναληπτικών επισκέψεων 26
2.3 Μέθοδοι Αντικατάστασης ελλιπών τιμών 27
2.3.1 Η μέθοδος του Buck 28
2.3.2 Ad-hoc μέθοδοι 30
2.3.3 Αντικατάσταση με χρήση τυχαίου παράγοντα 30
2.3.4 Μέθοδοι απλής αντικατάστασης 31
2.3.5 Εκτίμηση της διακύμανσης των εκτιμητριών απλής αντικατάστασης 34

**3. ΕΜΠ, Πολλαπλή αντικατάσταση, Αλγόριθμος ΕΜ**
3.1 Εκτιμήτριες μέγιστης πιθανοφάνειας 39
3.2 Η συνάρτηση μέγιστης πιθανοφάνειας για δεδομένα με ελλιπείς τιμές 40
3.3 Η συνάρτηση πιθανοφάνειας στην περίπτωση *MAR* 41
3.4 Η συνάρτηση πιθανοφάνειας στην περίπτωση *MNAR* 43
3.5 Πολλαπλή Αντικατάσταση 44
3.6 *EM* αλγόριθμος 46
3.7 Μέθοδος Gauss, Αλυσιδωτές Ισότητες, Μέθοδοι για μονότονα δεδομένα 48
3.8 Μπευζιανές μέθοδοι 49
3.9 Στατιστική ανάλυση με χρήση στατιστικών πακέτων 51

**4. Λογοκριμένα δεδομένα**
4.1 Λογοκριμένα δεδομένα 57

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

4.1.1 Τύποι λογοκρισίας 59
4.2 Συνάρτηση επιβίωσης και βαθμίδα αποτυχίας 69
4.3 Εκτιμήτριες μέγιστης πιθανοφάνειας 70
4.4 Η μέθοδος Newton 71
4.5 Διακύμανση και συνδιακύμανση της *ΕΜΠ* 72

**5. ΕΜΠ και Απαραμετρικές μέθοδοι για λογοκριμένα δεδομένα**
5.1 ΕΜΠ και μη παραμετρικές εκτιμήτριες για λογοκριμένα δεδομένα 73
5.1.1 Συνάρτηση πιθανοφάνειας για λογοκριμένα δεδομένα 73
5.1.2 Είδη λογοκρισίας και συνάρτηση πιθανοφάνειας 74
5.1.3 Παράδειγμα εκτίμησης παραμέτρων για δεξιά λογοκριμένα δεδομένα 78
5.1.4 Έλεγχοι υποθέσεων και διαστήματα εμπιστοσύνης σε δεξιά λογοκρισία τύπου I και II 84
5.2 Μη παραμετρική εκτίμηση για τη συνάρτηση επιβίωσης 87
5.2.1 Πίνακες επιβίωσης 88
5.2.2 Εκτίμηση της διακύμανσης της συνάρτησης επιβίωσης 92
5.2.3 Εμπειρική συνάρτηση επιβίωσης 93
5.2.4 Εκτιμήτρια Kaplan-Meier 94
5.2.5 Εκτιμήτρια Flemming-Harrington ή Nelson Aalen 100
5.2.6 Αλγόριθμος Turnbull 102
5.3 Απαραμετρικά μοντέλα 105
5.4 Στατιστική ανάλυση με χρήση στατιστικών πακέτων 107

**Περίληψη** 115
**Abstract** 117
**Βιβλιογραφία** 119

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

# ΕΙΣΑΓΩΓΗ

Δυο από τα σημαντικότερα προβλήματα που αντιμετωπίζουν οι στατιστικοί σχεδόν σε κάθε εργασία την οποία πραγματοποιούν είναι οι ελλιπείς τιμές (*missing values*) και τα λογοκριμένα δεδομένα (*censored values*). Στις στατιστικές αναλύσεις, τα δεδομένα που συλλέγονται, οργανώνονται συνήθως σε πίνακες δεδομένων όπου κάθε γραμμή αντιπροσωπεύει μία μονάδα και κάθε στήλη μια μεταβλητή η οποία μετράται για την μονάδα αυτή. Οι πίνακες αυτοί μπορούν να περιλαμβάνουν ποιοτικές μεταβλητές (φύλο, πρόθεση ψήφου) διατάξιμες (μορφωτικό επίπεδο, κατηγορία bonus malus), και ποσοτικές, διακριτές ή συνεχείς (βάρος, ηλικία, ύψος).

Σε πολλές περιπτώσεις λόγω μιας σειράς γνωστών ή άγνωστων παραγόντων, μερικές μονάδες του πίνακα παρουσιάζουν ελλιπείς τιμές. Επίσης, επειδή κάποιες έρευνες δεν ολοκληρώνονται (για διάφορους λόγους) μέχρις ότου ο αναλυτής λάβει αποτελέσματα για όλο το δείγμα που εξετάζει, είναι δύσκολο να συγκεντρωθούν πλήρη δεδομένα. Μία τέτοια έρευνα τερματίζεται συνήθως προτού όλες οι μονάδες που βρίσκονται υπό εξέταση αποβιώσουν, και έχουμε την πληροφορία ότι μερικές από αυτές είναι εν ζωή στο τέλος της έρευνας, χωρίς να υπάρχει δυνατότητα καταγραφής της χρονικής στιγμής που θα τερματίσει η λειτουργία τους. Αυτό αποτελεί και το κίνητρο της θεωρίας των λογοκριμένων δεδομένων. Σχετικά με το πρώτο πρόβλημα κάποια στατιστική μέθοδος, σχεδιασμένη να αναλύσει πίνακες με ελλιπείς τιμές, θεωρητικά αναμένει σε κάθε μονάδα του πίνακα και τον αντίστοιχο πραγματικό αριθμό, συνεπώς δεν μπορεί να εφαρμοστεί σε μονάδες που παρουσιάζουν ελλιπείς τιμές. Τα διάφορα στατιστικά πακέτα, προκειμένου να αντιμετωπίσουν τα παραπάνω προβλήματα, είναι προγραμματισμένα να αγνοούν κάθε μονάδα στην οποία εμφανίζονται ελλιπείς τιμές και να αναλύουν ένα πλήρες υποσύνολο των πραγματικών μονάδων. Η μέθοδος αυτή είναι γνωστή ως **Ανάλυση των πλήρων παρατηρήσεων** (*Complete-Case Analysis(CC)*.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Έχουν προταθεί κατά καιρούς ποικίλες μέθοδοι για την αντιμετώπιση και το χειρισμό δεδομένων με ελλιπείς τιμές. Οι μέθοδοι αυτοί μπορεί είτε να κάνουν απλή διαγραφή των μονάδων με ελλιπείς τιμές, είτε, με διάφορες τεχνικές, να προβαίνουν στην αντικατάστασή τους με άλλες. Σκοπός του στατιστικού αναλυτή είναι η δημιουργία ενός πλήρου δείγματος, ώστε τα αποτελέσματα που θα προκύψουν της έρευνας να είναι έγκυρα ενώ συνάμα να μη χαθεί καμία πληροφορία η οποία περιέχεται στο δείγμα.

Στη συγκεκριμένη εργασία θα παρουσιαστούν διάφορες τεχνικές χειρισμού δειγμάτων με ελλιπείς τιμές. Αρχικά θα δοθεί η μέθοδος της διαγραφής των ελλιπών μονάδων από το δείγμα, που σκοπό έχει μία πρώτη προσέγγιση του προβλήματος από τον αναλυτή. Θα παρουσιαστούν μέθοδοι απλής αντικατάστασης των ελλιπών τιμών από ένα απλό σύνολο περιγραφικών στατιστικών στοιχείων (πχ. μέση τιμή, διάμεσος, επικρατούσα τιμή). Στη συνέχεια θα γίνει μία εκτενής αναφορά σε πιό σύνθετες μεθόδους (πχ. στον ΕΜ αλγόριθμο, πολλαπλή αντικατάσταση κτλ) οι οποίες και δίνουν πιό αντιπροσωπευτικές τιμές αντικατάστασης. Αυτές στηρίζονται στην ιδέα των επαναληπτικών διαδικασιών. Τέλος, θα παρουσιαστεί και η μέθοδος των εκτιμητριών μέγιστης πιθανοφάνειας που αποτελεί το πιό σημαντικό κομμάτι της εκτιμητικής, η οποία δίνει αξιόπιστα αποτελέσματα ακόμη και όταν θέλουμε να αναλύσουμε δείγματα με ελλιπείς τιμές.

Τα λογοκριμένα καθώς και τα περικομμένα δεδομένα ήταν ανέκαθεν σημαντικό θέμα έρευνας για τους στατιστικούς και τους αναλογιστές. Απαντώνται στη βιομηχανία ασφαλίσεων ζωής, σε δημογραφικές μελέτες, στη βιοστατιστική, στον ποιοτικό έλεγχο αλλά και στη θεωρία αξιοπιστίας. Λόγω της ανάγκης ανάλυσης τους σε πολλές εφαρμοσμένες περιοχές απαιτούν και τον ανάλογο σοβαρό χειρισμό ώστε τα αποτελέσματα που θα προκύψουν να είναι σωστά και τα συμπεράσματα έγκυρα.

Διακρίνουμε διάφορες μορφές λογοκρισίας. Οι βασικές μορφές είναι η δεξιά λογοκρισία τύπου Ι και ΙΙ, η αριστερή λογοκρισία, η λογοκρισία σε διάστημα και η τυχαία λογοκρισία. Φυσικά στην πράξη

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

εμφανίζεται συνήθως κάποιος συνδιασμός των παραπάνω μορφών τόσο μεταξύ τους όσο και με τα περικομμένα δεδομένα.

Στην εργασία αυτή θα παρουσιαστούν μερικές απλές τεχνικές αντιμετώπισης δειγμάτων με λογοκριμένα δεδομένα. Οι τεχνικές αυτές περιλαμβάνουν τρείς κατηγορίες: την εμπειρική κατανομή, τις εκτιμήτριες μέγιστης πιθανοφάνειας και την απαραμετρική εκτίμηση. Η εμπειρική συνάρτηση κατανομής του δείγματος είναι η πιό απλή μορφή προσέγγισης λογοκριμένων δεδομένων. Βασικό μειονέκτημά της είναι ο αποκλισμός από την εκτίμηση της συνάρτησης, όσων δεδομένων είναι λογοκριμένα. Οι εκτιμήτριες μέγιστης πιθανοφάνειας αποτελούν το βασικό εργαλείο για την επίλυση προβλημάτων με λογοκριμένα δεδομένα, διότι δεν αποκλείονται οι λογοκριμένες τιμές και συνεπώς δεν χάνεται καθόλου πληροφορία. Τέλος στην εργασία θα παρουσιαστούν και μερικές απαραμετρικές μέθοδοι που περιλαμβάνουν την εκτιμήτρια Kaplan-Meier και Flemming-Harrington για λογοκρισία από δεξιά και αριστερά και τον αλγόριθμο Turnbull για λογοκριμένες τιμές σε διάστημα.

Η εργασία ολοκληρώνεται με αναφορά σε διάφορα στατιστικά πακέτα. Τα πακέτα που κυκλοφορούν στην αγορά έχουν συμπεριλάβει στις βιβλιοθήκες τους ειδικές ρουτίνες για την επεξεργασία τέτοιων δειγμάτων. Συνεπώς είναι απαραίτητο ο στατιστικός να γνωρίζει πως λειτουργούν τα προγράμματα αυτά. Για το λόγο αυτό, στα πλαίσια αυτής της διπλωματικής παρουσιάζονται οι διαδικασίες που πρέπει να ακολουθήσει κάποιος σε καθένα από τα πακέτα αυτά για το χειρισμό λογοκριμένων δεδομένων καθώς και δειγμάτων με ελλιπείς τιμές

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

# ΚΕΦΑΛΑΙΟ 1

## 1.1 ΕΛΛΙΠΗΣ ΤΙΜΗ ΚΑΙ ΛΟΓΟΚΡΙΜΕΝΑ ΔΕΔΟΜΕΝΑ

Ένα σημείο το οποίο οφείλουμε να ξεκαθαρίσουμε από την αρχή είναι η διαφορά μεταξύ της ελλιπούς τιμής (***missing value***) και των λογοκριμένων δεδομένων (***censored data***). Για το σκοπό αυτό παρατίθενται τα παρακάτω παράδειγματα.

Σε μία έρευνα, που σκοπό έχει να εξετάσει αν η παχυσαρκία σχετίζεται με το φύλο του ατόμου, συμμετέχουν 1000 άτομα εκ των οποίων τα 600 είναι αγόρια και τα 400 είναι κορίτσια. Στο σύνολο των αγοριών έχει καταγραφεί το βάρος των 560 και από το σύνολο των κοριτσιών μόλις των 380. Είναι εμφανές ότι η διακριτή μεταβλητή φύλο είναι πλήρως καθορισμένη, αντιθέτως η συνεχής μεταβλητή βάρος δεν εμφανίζει παντού τιμές. Η απάντηση στο ερώτημα του βάρους του κάθε ατόμου υπάρχει, αλλά ο ερευνητής δεν τη γνωρίζει και συνεπώς στον πίνακα των δεδομένων πρέπει να εμφανιστεί ως ελλιπής τιμή.

Με βάση τα παραπάνω η ελλιπής τιμή μπορεί να οριστεί γενικά ως *μια* ***παρατήρηση*** *της οποίας η τιμή είναι* ***σαφώς καθορισμένη*** *αλλά για κάποιο λόγο* ***δεν είναι γνωστή*** *στον ερευνητή.*

Όταν όλοι οι χρόνοι ζωής που αφορούν μία σειρά ατόμων είναι γνωστοί, τα δεδομένα είναι **πλήρη** για την έρευνα που πραγματοποιείται. Για να επιτευχθεί η συγκέντρωση πλήρων δεδομένων ο ερευνητής θα πρέπει να συνεχίσει το πείραμα μέχρι να αποβιώσουν όλα τα άτομα που τέθηκαν σε έλεγχο.

Στην πράξη, η αναμονή μέχρι να αποβιώσουν όλα τα άτομα είναι συνήθως ανέφικτη ή οικονομικά ασύμφορη. Για το λόγο αυτό, στις πιο πολλές περιπτώσεις, ο ερευνητής διακόπτει το πείραμα πριν ολοκληρωθεί η διαδικασία, με αποτέλεσμα να μην είναι γνωστές όλες οι τιμές των χρόνων ζωής των ατόμων. Στην περίπτωση αυτή τα δεδομένα είναι **λογοκριμένα** (***censored data***).

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

## 1.2 ΜΗΧΑΝΙΣΜΟΙ ΠΟΥ ΟΔΗΓΟΥΝ ΣΕ ΕΛΛΙΠΕΙΣ ΤΙΜΕΣ

Το πιο καίριο ερώτημα που πρέπει να απαντηθεί προκειμένου να επιλεγεί κατάλληλη στατιστική μέθοδος για την ανάλυση μονάδων με ελλιπείς τιμές είναι για ποιο λόγο εμφανίζονται αυτές. Αν κάποιος γνωρίζει το μηχανισμό που τις δημιουργεί, τότε είναι σε θέση να δημιουργήσει την κατάλληλη μέθοδο για να τις αντιμετωπίσει, Little (1995).

Υπάρχουν δύο βασικές κατηγορίες μηχανισμών που δημιουργούν ελλιπείς τιμές. Η πρώτη κατηγορία περιλαμβάνει ελλιπείς τιμές που οφείλονται σε τυχαίους παράγοντες και η δεύτερη ελλιπείς τιμές οι οποίες οφείλονται στο σχεδιασμό της έρευνας. Μερικά παραδείγματα από τις δύο αυτές κατηγορίες είναι τα ακόλουθα:

1) Ελλιπείς τιμές μπορεί να οφείλονται σε **τυχαίους παράγοντες**, όπως
   - α) Ανθρώπινα σφάλματα κατά την καταχώρηση, κωδικοποίηση ή υπολογισμό δεδομένων (*human errors*).
   - β) Ασθενής αρνούμενος να υποβληθεί σε τεστ ή θεραπεία (*non-compliance*).
   - γ) Άτομο αρνούμενο να συμμετάσχει σε έρευνα.
   - δ) Ασθενής που συμμετέχει σε μια κλινική δοκιμή και χάνεται από την παρακολούθηση λόγω αλλαγής της διεύθυνσης διαμονής (*mover*).
   - ε) Ασθενής ο οποίος χάνεται από μία σειρά ιατρικών επισκέψεων (*follow up*).
   - στ) Άτομο το οποίο συμμετέχει σε έρευνα και αρνείται να απαντήσει σε ερώτηση είτε για προσωπικούς λόγους είτε γιατί δεν είναι σίγουρος για το απόρρητο της έρευνας.
   - ζ) Ασθενής που λαμβάνει μέρος σε κλινική δοκιμή για μία ασθένεια και δεν προσέρχεται σε προγραμματισμένη επίσκεψη επειδή δεν αισθανόταν καλά.
   - η) Αμέλεια ή απειρία των ερευνητών (πχ. αδυνατούν να εντοπίσουν τους ερευνώμενους, παραλείπουν καταχωρήσεις

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

απαντήσεων κλπ), μη κατανόηση των ερωτήσεων ή των απαντήσεων. Συμβαίνει ακόμα ορισμένοι ερευνητές να μη γίνονται συμπαθείς στους ερευνώμενους, οπότε εισπράττουν την άρνησή τους.

θ) Μη δυνατότητα πρόσβασης σε δεδομένα/ εμπιστευτικότητα.

2) Ελλιπείς τιμές μπορεί να οφείλονται στο **σχεδιασμό της έρευνας**. Ακολουθούν δύο παραδείγματα που παρουσιάζουν τέτοιες περιπτώσεις.

α) Σε ένα νοσοκομείο πραγματοποιείται μία μελέτη για την παχυσαρκία. Συγκεκριμένα, ερευνάται το ανθρώπινο λίπος (μεταβλητή $Y_1$) διαφόρων ατόμων από ένα πληθυσμό και η συσχέτισή του με άλλους παράγοντες. Υποθέτουμε ότι η μέτρηση της μεταβλητής $Y_1$ είναι ιδιαίτερα δαπανηρή, όμως μια εξίσου καλή εναλλακτική επιλογή είναι ο δείκτης ανθρώπινης μάζας $Y_2$ (που είναι μία συνάρτηση του ύψους και του βάρους). Ο υπολογισμός του τελευταίου δεν είναι καθόλου δαπανηρός, καθώς εκφράζεται με μία απλή μαθηματική σχέση η οποία δίνεται από τον τύπο $\left(\frac{\text{βάρος ατόμου}}{\text{ύψος στο τετράγωνο}}\right)$. Για κάθε άτομο καθορίζεται η μεταβλητή $Y_2$ και οι διάφοροι άλλοι παράγοντες με τους οποίους πρόκειται να μελετηθεί η συσχέτισή της. Στη συνέχεια επιλέγεται ένα υποσύνολο του αρχικού μας δείγματος και υποβάλλεται σε ένα test για τον υπολογισμό της $Y_1$.

Είναι φανερό ότι, στην υπό μελέτη περίπτωση η μεταβλητή $Y_2$ έχει παρατηρήσεις με καθορισμένες τιμές, ενώ η $Y_1$ αποτελείται από παρατηρήσεις οι οποίες σε ορισμένες περιπτώσεις είναι ελλιπείς (*partial non-response*).

β) Διενεργείται μία έρευνα προκειμένου να διερευνηθεί αν υπάρχει κάποια σχέση μεταξύ μίας νόσου ($Y$) και κάποιου

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

παράγοντα ($X$). Έστω ότι η ανίχνευση του παράγοντα είναι ακριβή, και για το λόγο αυτό αποφασίζεται να χρησιμοποιηθεί μία φθηνότερη μέθοδος της οποίας τα αποτελέσματα $Z$ είναι ενδεικτικά της έκθεσης ή μη του ατόμου στον παράγοντα. (1.3.1)

Για κάθε άτομο από το δείγμα καθορίζεται η μεταβλητή $Y$ (νόσος ή μη) και η $Z$ (φθηνή μέθοδος έκθεσης ή μη στον παράγοντα $X$). Στη συνέχεια, επιλέγεται ένα υποσύνολο του αρχικού δείγματός μας και υποβάλλεται στο ακριβό test $X$.

Με τον τρόπο αυτό υπάρχουν δύο μεταβλητές $Y$ και $Z$ πάντα καθορισμένες, ενώ αντιθέτως η μεταβλητή $X$ παρουσιάζει σε ορισμένες περιπτώσεις παρατηρήσεις με ελλιπείς τιμές (***partial non-response***). Στο συγκεκριμένο παράδειγμα ο ερευνητής γνωρίζει πλήρως το μηχανισμό που δημιουργεί τις ελλιπείς τιμές, αφού αυτός τον καθορίζει.

ΠΑΡΑΤΗΡΗΣΕΙΣ

- Ο μηχανισμός ελλιπών τιμών δεν είναι δυνατόν να βρίσκεται πάντα υπό τον πλήρη έλεγχο του ερευνητή.
- Ο ρόλος του μηχανισμού ελλιπών τιμών είναι καθοριστικός στην επιλογή του μοντέλου που θα χρησιμοποιηθεί για τη στατιστική ανάλυση των δεδομένων.
- Μία μεταβλητή μπορεί να παρουσιάζει ελλιπείς τιμές σε κάποιες περιπτώσεις (***partial non-response***) ή στο σύνολό της (***total non-response***).
- Οι αιτίες που οδήγησαν σε μια ελλιπή τιμή παίζουν σημαντικό ρόλο στο κόστος που θα είχε το να αγνοηθεί η ελλιπής τιμή και να γίνει στατιστική ανάλυση χρησιμοποιώντας μόνο τις μονάδες με πλήρεις παρατηρήσεις.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

## 1.3 ΔΕΙΚΤΗΣ ΕΛΛΙΠΩΝ ΤΙΜΩΝ

Ο Δείκτης Ελλιπών Τιμών (*Missing Data Indicator*) ο οποίος προτάθηκε από τους Little and Rubin (2002) είναι η πλέον βασική έννοια για τη μελέτη μη πλήρων δεδομένων αφού μέσω αυτής γίνεται η περαιτέρω ανάλυσή τους.

Έστω $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_k)$ μία $k$-διάστατη τυχαία μεταβλητή από ένα τυχαίο δείγμα μεγέθους $n$, στο οποίο κάποιες παρατηρήσεις μπορεί να είναι ελλιπείς.

Ο Δείκτης Ελλιπών Τιμών (ΔΕΤ) παριστάνεται με το αγγλικό γράμμα $\boldsymbol{R} = (R_{ij}) \quad i=1,\ldots,n \;\; j=1,\ldots,k$ και ορίζεται ως εξής:

$$R_{ij} = \begin{array}{l} 1 \text{ αν η } j \text{ συνιστώσα της } i \text{ μονάδας έχει καταγραφεί} \\ 0 \text{ αν η } j \text{ συνιστώσα ττη } i \text{ μονάδας είναι ελλιπής.} \end{array}$$

Όταν το δείγμα είναι μονοδιάστατο τότε ο $R$ αναφέρεται κάθε φορά σε μία μονάδα του δείγματος, συνεπώς,

$$\boldsymbol{R} = (R_j) \quad j=1,2,\ldots,k,$$

όπου

$$R_j = \begin{cases} 0 \text{ αν η } j \text{ συνιστώσα έχει καταγραφεί} \\ 1 \text{ αν η } j \text{ συνιστώσα δεν υπάρχει.} \end{cases}$$

Ο Δείκτης Ελλιπών Τιμών είναι μία $nxk$-διάστατη τυχαία μεταβλητή (ή $k$-διάστατη αν ληφθεί υπ' όψιν ο δεύτερος ορισμός). Οι τιμές της δεν είναι γνωστές πριν την επιλογή του τυχαίου δείγματος (δηλαδή δεν είναι γνωστό εκ των προτέρων που θα εμφανιστούν οι ελλιπείς τιμές).

Οι γραμμές του πίνακα $R$, στον πρώτο ορισμό, είναι στατιστικά ανεξάρτητες, δεδομένου ότι και οι δύο ορισμοί του Δείκτη αναφέρονται σε ανεξάρτητες μονάδες και αντιμετωπίζουν με τον ίδιο τρόπο το πρόβλημα.

Η τυχαία μεταβλητή $R$ εξαρτάται μόνο από τις τιμές των $Y_j$ της μονάδας στην οποία αναφέρεται. Στην περίπτωση που είναι γνωστό ότι κάποιες μεταβλητές δεν είναι ποτέ ελλιπείς (πχ γένος, ηλικία), μπορεί στις εφαρμογές η διάσταση της τυχαίας μεταβλητής $R$ να ελαττωθεί.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Με τη χρήση του Δείκτη Ελλιπών Τιμών μπορεί να περιγραφεί κάθε σύνολο δεδομένων με ελλιπείς τιμές.

Στη μονοδιάστατη περίπτωση, η τυχαία μεταβλητή $Y$ χωρίζεται συνήθως σε δύο συνιστώσες. Το σύνολο των μονάδων που έχουν παρατηρήσεις με συγκεκριμένες τιμές και επομένως είναι γνωστές, αποτελεί το **παρατηρηθέν τμήμα** της, και συμβολίζεται με $\boldsymbol{Y}_{obs}$. Προφανώς

$$\boldsymbol{Y}_{\boldsymbol{obs}} = (Y_j : R_j = 1)$$

Το σύνολο των μονάδων που έχουν ελλιπείς τιμές αποτελεί το **ελλιπές τμήμα** της, και συμβολίζεται με $\boldsymbol{Y}_{mis}$, δηλαδή

$$\boldsymbol{Y}_{\boldsymbol{mis}} = (Y_j : R_j = 0)$$

Με κατάλληλη αναδιάταξη του δείγματος, μπορούμε να γράψουμε

$$\boldsymbol{Y} = (\boldsymbol{Y}_{\boldsymbol{obs}}, \boldsymbol{Y}_{\boldsymbol{mis}}).$$

## 1.4 ΟΡΙΣΜΟΣ ΒΑΣΙΚΩΝ ΜΗΧΑΝΙΣΜΩΝ ΕΛΛΙΠΩΝ ΤΙΜΩΝ

Θεωρούμε ένα τυχαίο δείγμα μεγέθους $n$ από την $k$-διάστατη τυχαία μεταβλητή $\boldsymbol{Y}$ και συμβολίζουμε με $\boldsymbol{R}$ τον αντίστοιχο δείκτη ελλιπών τιμών. Ο μηχανισμός ελλιπών τιμών που θα αναλυθεί παρακάτω είναι ουσιαστικά η κατανομή $f$ της τυχαίας μεταβλητής $\boldsymbol{R}$ η οποία παρέχει την πιθανότητα σε μία μεταβλητή $Y$ κάποια $Y_i$ (μονάδες) να παρουσιάζουν ελλιπείς τιμές, Little and Rubin (2002).

Στη συνέχεια θα χρησιμοποιήσουμε το συμβολισμό

$$f(r_1,...,r_k) = P(\boldsymbol{R} = (r_1,...,r_k)),\ r_j \in \{0,1\}$$

για την από κοινού κατανομή των τυχαίων μεταβλητών $R_1, R_2, ..., R_k$.

Η γνώση του μηχανισμού ελλιπών τιμών ισοδυναμεί με γνώση της παραπάνω κατανομής.

Ο μηχανισμός ελλιπών τιμών καθορίζει και τους λόγους για τους οποίους οι τιμές αυτές είναι ελλιπείς. Ο Rubin (1976), εξηγώντας τους λόγους που εμφανίζονται ελλιπείς παρατηρήσεις και τον τρόπο με τον οποίο αυτές σχετίζονται με τις παρατηρούμενες τιμές μίας μεταβλητής,

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

παρουσίασε και περιέγραψε τρεις διαφορετικούς μηχανισμούς, που οδηγούν σε ελλιπή σύνολα δεδομένων: το μηχανισμό των **πλήρως τυχαίων ελλιπών δεδομένων** (*Missing Completely At Random*), το μηχανισμό των **τυχαίων ελλιπών δεδομένων** (*Missing At Random*) και το μηχανισμό των **μη τυχαίων ελλιπών δεδομένων** (*Missing Not At Random*). Σύμφωνα με τον πρώτο από αυτούς, τα δεδομένα ονομάζονται **Πλήρως Τυχαία Ελλιπή** όταν ο μηχανισμός που δημιουργεί την έλλειψη είναι τελείως ανεξάρτητος τόσο από την φύση όσο και από τιμές των υπό μελέτη μεταβλητών. Στην περίπτωση αυτή μπορούμε να γράψουμε

$$f_{R|Y}(\boldsymbol{r} \mid \boldsymbol{y}) = f(\boldsymbol{r})$$

αφού η κατανομή του δείκτη ελλιπών τιμών $\boldsymbol{R}$ είναι ανεξάρτητη του $\boldsymbol{Y}$. Για το λόγο αυτό, λέμε ότι οι ελλιπείς τιμές οφείλονται εντελώς στην τύχη. Ο μηχανισμός αυτός θα αναφέρεται με τα αρχικά **MCAR** και δηλώνει ότι η εμφάνιση ελλιπών τιμών στο δείγμα δε σχετίζεται με κανένα παράγοντα, είτε γνωστό είτε άγνωστο που αφορά την έρευνα (π.χ. σχεδιασμός της έρευνας, τυχαίοι παράγοντες, ανθρώπινα λάθη κ.τ.λ.).

Στην περίπτωση που οι ελλιπείς τιμές παρουσιάζονται με βάση το μηχανισμό MCAR, οι πλήρεις παρατηρήσεις είναι ένα απλό τυχαίο δείγμα του αρχικού δείγματος.

Ένα χαρακτηριστικό παράδειγμα μηχανισμού *MCAR* είναι το ακόλουθο.

- Ένας σωλήνας που περιέχει δείγμα αίματος από ένα ασθενή σπάει από απροσεξία του γιατρού, με αποτέλεσμα οι παράμετροι του αίματος να μη μπορούν να μετρηθούν. Στο παράδειγμα αυτό η τυχαία μεταβλητή $\boldsymbol{Y}$ είναι οι παραμέτροι του αίματος του ασθενούς που πρέπει να εκτιμηθούν (π.χ. σάκχαρο) και ο ΔΕΤ δηλώνει με την τιμή μηδέν ότι ο σωλήνας έσπασε και δεν υπάρχει μέρτηση του σακχάρου (ελλιπής τιμή), ενώ με την τιμή 1 ότι υπάρχει συγκεκριμένη τιμή για τον ασθενή (εάν ο σωλήνας δε σπάσει). Επειδή το αν θα σπάσει η όχι ο σωλήνας θεωρείται τυχαίο γεγονός, οι τιμές που θα πάρει ο ΔΕΤ είναι εντελώς

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

τυχαίες και η κατανομή του ανεξάρτητη από τις παρατηρήσεις του δείγματος που θα προκύψει.

Κάποια άλλα παραδείγματα μηχανισμού *MCAR* είναι τα εξής:

- Πραγματοποιείται μία έρευνα και χάνεται ένα συμπληρωμένο ερωτηματολόγιο για άγνωστους λόγους.
- Δημιουργούνται ελλιπείς τιμές που οφείλονται σε λάθος πληκτρολόγηση.
- Κάποιος ασθενής δεν προσέρχεται σε προγραμματισμένη εξέταση για άγνωστους λόγους.
- Κάποια δεδομένα μπορεί να χαθούν από λάθη ενός υπαλλήλου.

Στην περίπτωση *MCAR*, η πλέον απλή τεχνική αντιμετώπισης δεδομένων με ελλιπείς τιμές, δηλαδή η ανάλυση με χρήση μόνο των πλήρων περιπτώσεων δίνει αμερόληπτα αποτελέσματα (Greenland και Finkle (1995)).

Σύμφωνα με το δεύτερο μηχανισμό που περιέγραψε ο Rubin (1976), τα δεδομένα χαρακτηρίζονται ως **Τυχαία Ελλιπή**, όταν ο μηχανισμός της έλλειψης εξαρτάται από τις παρατηρηθείσες τιμές, χωρίς όμως να εξαρτάται και από τις ελλιπείς παρατηρήσεις. Ο μηχανισμός αυτός αποκαλείται **MAR** (*Missing At Random*), και μπορούμε να γράψουμε

$$f_{R|Y}(\boldsymbol{r} \mid \boldsymbol{y}) \equiv f(\boldsymbol{r} \mid \boldsymbol{y}_{mis}, \boldsymbol{y}_{obs}) = f(\boldsymbol{r} \mid \boldsymbol{y}_{obs}).$$

Η παραπάνω ισότητα δηλώνει ότι η εμφάνιση ελλιπών τιμών εξαρτάται μόνο από τις γνωστές τιμές. Στην πράξη *MAR* μπορεί να έχουμε σε σύνολα δεδομένων στα οποία κάποιες μεταβλητές δεν είναι ποτέ ελλιπείς (πχ φύλο, ηλικία) και η πιθανότητα οι υπόλοιπες μεταβλητές να είναι ελλιπείς εξαρτάται μόνο από τις πρώτες.

Ο μηχανισμός *MAR* προϋποθέτει ότι υπάρχει κάποια συνάρτηση *g* των μονάδων με πλήρεις παρατηρήσεις η οποία καθορίζει την πιθανότητα οι υπόλοιπες μονάδες να παρουσιάζουν ελλιπείς παρατηρήσεις (Collins,Schafer and Kam (2001)).

Η σημασία του μηχανισμού *MAR* είναι ιδιαίτερα σημαντική αφού στην περίπτωση που αυτός ισχύει, δεν απαιτείται ακριβής καθορισμός

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

της συνάρτησης $g(\boldsymbol{y}_{obs}) = f(\boldsymbol{r} \mid \boldsymbol{y}_{obs})$, κάτι το οποίο θα διαπιστωθεί στη συνέχεια με τη χρήση της εκτιμήτριας μέγιστης πιθανοφάνειας.

Τυπικά παραδείγματα *MAR* είναι τα ακόλουθα:

- Πραγματοποιείται μία έρευνα και ζητείται να εκτιμηθεί η προβλεπόμενη τιμή για ένα συγκεκριμένο διαγνωστικό τεστ, όμως τα αποτελέσματα είναι γνωστά για όλα τα άτομα που έχουν νοσήσει και άγνωστα για ένα τυχαίο δείγμα όσων δεν έχουν νοσήσει. Σε αυτό το παράδειγμα, ο μηχανισμός ελλιπών τιμών είναι *MAR*, καθώς η πιθανότητα να εμφανιστούν ελλιπείς τιμές στο δείγμα εξαρτάται από το αν το άτομο έχει εκδηλώσει τη νόσο ή όχι (Oostenbrink, Moons, Bleeker, Moll and Grobbee (2003)).
- Αν το φύλο σε ένα δείγμα είναι γνωστό σε κάθε μονάδα τότε η πιθανότητα να εμφανιστεί ελλιπής τιμή σε οποιαδήποτε άλλη μεταβλητή του δείγματος μπορεί να εξαρτάται μόνο από το φύλο.

Και στην περίπτωση μηχανισμού ελλιπών τιμών MAR, η πλέον απλή τεχνική αντιμετώπισης των δεδομένων, είναι η ανάλυση των πλήρων παρατηρήσεων.

Όταν ο μηχανισμός ελλιπών τιμών δεν μπορεί να απλοποιηθεί (δηλαδή η κατανομή $f_{R|Y}(\boldsymbol{r} \mid \boldsymbol{y}_{mis}, \boldsymbol{y}_{obs})$ του ΔΕΤ εξαρτάται τόσο από τις παρατηρηθείσες όσο και από τις ελλιπείς τιμές) ο Rubin (1976)πρότεινε τη χρήση του όρου **μη αμελητέος**, **NINR** (*Non Ignorable Non Response*), ή **μη τυχαίος MNAR** (*Missing Not At Random*). Στην περίπτωση *NINR* ο ερευνητής είναι υποχρεωμένος να καθορίσει πλήρως την πιθανότητα ελλιπών τιμών (δηλαδή την κατανομή του ΔΕΤ), ώστε να βρεθούν συνεπείς εκτιμήτριες των παραμέτρων ενός μοντέλου παλινδρόμησης με εξαρτημένη μεταβλητή $Y$ (η οποία είναι πάντοτε γνωστή) και ανεξάρτητη $X$ (στην οποία πιθανόν να εμφανιστούν ελλιπείς τιμές). Ουσιαστικά στο συγκεκριμένο μηχανισμό, η πιθανότητα μία μονάδα να εμφανίζει ελλιπή τιμή εξαρτάται από την τιμή της και δε μπορεί να αποδοθεί εξ ολοκλήρου σε γνωστούς παράγοντες (Carpenter, Kenward and Vansteelandt (2006)).
Ένα τυπικό παράδειγμα *NINR* είναι το ακόλουθο:

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

- Ζητείται από ερωτώμενους να δηλώσουν το επίπεδο του μηνιαίου μισθού τους. Είναι πολύ πιθανό να εμφανιστούν ελλιπείς τιμές όταν το επίπεδο των μηνιαίων απολαβών είναι σχετικά υψηλό. Η εμφάνιση ελλιπών τιμών δεν οφείλεται εντελώς στην τύχη, αλλά σχετίζεται με άγνωστους παράγοντες (χαρακτηριστικά) του ερωτώμενου.
- Σημειώνουμε ότι και τα λογοκριμένα δεδομένα που θα εξετασθούν στις επόμενες παραγράφους αποτελούν ειδική περίπτωση μηχανισμού *NINR* (*MNAR*) .

Όταν ο μηχανισμός είναι *NINR*, χάνονται χρήσιμες πληροφορίες από τα δεδομένα ενώ δεν υπάρχει μία γενική μέθοδος για την αποτελεσματική διαχείριση των δεδομένων αυτών.

Ένα σημαντικό πρόβλημα στο πεδίο της μελέτης ελλιπών δεδομένων προκύπτει, συνήθως, όταν σε μία μελέτη υπάρχουν πολλαπλές μεταβλητές με ελλιπείς τιμές. Στην περίπτωση αυτή, το δείγμα θα καλείται **μονότονο** *(monotone*), αν ο πίνακας δεδομένων δύναται να διαμορφωθεί με τέτοιο τρόπο ώστε να υπάρχει μία ιεραρχία στις ελλιπείς τιμές. Πιο συγκεκριμένα, σε ένα μονότνο δείγμα, ζητάμε, όταν διαθέτουμε για μία μονάδα του δείγματος την τιμή της μεταβλητής $X_b$ να συνεπάγεται ότι η παρατήρηση της μεταβλητής $X_a$, για $a<b$ για την ίδια μονάδα είναι επίσης γνωστή. Ο πίνακας 1.5.1 παρουσιάζει δύο υποθετικά σύνολα δεδομένων. Το αριστερό έχει διαμορφωθεί ώστε να σχηματίσει ένα μονότονο δείγμα (stairstep), κάτι το οποίο δεν γίνεται να πραγματοποιηθεί με το δεύτερο (δεξιό) σύνολο δεδομένων. Όταν το δείγμα είναι **μη-μονότονο** (*Non-Monotone*) τα μοντέλα που θα χρησιμοποιηθούν για την εκτίμηση των ελλιπών τιμών κάθε μιας από τις μεταβλητές μπορεί να συμπεριλάβουν συνδιακυμάνσεις, των οποίων οι τιμές θα είναι επίσης ελλιπείς. Αντίθετα, όταν το δείγμα είναι μονότονο, απλοποιούνται οι σχετικές διαδικασίες. Ωστόσο ένα μονότονο δείγμα είναι ασύνηθες να παρουσιαστεί σε μία έρευνα.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

**Πίνακας 1.5.1**

Μονότονα-μη μονότονα δεδομένα

| Monotone | | | | Non-Monotone | | |
|---|---|---|---|---|---|---|
| **Δείγμα** | **X1** | **X2** | **X3** | **X1** | **X2** | **X3** |
| **1** | Obs | Obs | Obs | Obs | Obs | Obs |
| **2** | Obs | Obs | M | Obs | Obs | M |
| **3** | Obs | M | M | Obs | M | M |
| **4** | M | M | M | Obs | M | Obs |
| **5** | | | | M | Obs | Obs |
| **6** | | | | M | M | Obs |
| | | | | | | |

ΠΑΡΑΤΗΡΗΣΕΙΣ

- Μία συνέπεια από την παρουσία ελλιπών τιμών είναι η μεροληψία των εκτιμητριών. Η φύση της μεροληψίας εξαρτάται από τον εκάστοτε μηχανισμό ελλιπών τιμών. Όταν ο μηχανισμός είναι *MCAR*, προκύπτουν συνήθως αμερόληπτες εκτιμήτριες, ενώ όταν ο μηχανισμός είναι *MAR* ή *NINR* (*MNAR*) εμφανίζεται αρκετά συχνά μεροληψία στις χρησιμοποιούμενες εκτιμήτριες ( Greenland and Finkle (1995), Vach (1994), Schafer (1997).
- Γενικά η παρουσία ελλιπών τιμών έχει ως αποτέλεσμα την υποεκτίμηση των διακυμάνσεων των εκάστοτε δειγμάτων.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

# ΚΕΦΑΛΑΙΟ 2

Στο δεύτερο μέρος της παρούσας διπλωματικής εργασίας θα παρουσιαστούν κάποιες μέθοδοι για την αντιμετώπιση προβλημάτων με ελλιπείς τιμές. Θα δοθούν μέθοδοι οι οποίες στηρίζονται στην ιδέα της συμπλήρωσης των ελλιπών τιμών αφού είναι οι πλέον συνήθεις μέθοδοι και αργότερα μπορούν να χρησιμοποιηθούν για να δώσουν αρχικές τιμές σε επαναληπτικές διαδικασίες.

## 2.1 ΑΠΛΕΣ ΤΕΧΝΙΚΕΣ ΧΕΙΡΙΣΜΟΥ ΕΛΛΙΠΩΝ ΤΙΜΩΝ

Η πλέον απλή και αρκετά κοινή μέθοδος ανάλυσης δεδομένων με ελλιπείς τιμές είναι η **διαγραφή** (*Deletion*) εκείνων των μονάδων που παρουσιάζουν ελλιπείς τιμές και η ανάλυση του συγκεκριμένου δείγματος χρησιμοποιώντας μονάδες με πλήρεις παρατηρήσεις.

### 2.1.1 Ανάλυση με χρήση μόνο των πλήρων περιπτώσεων

Έστω ένα τυχαίο δείγμα $\boldsymbol{Y} = (Y_1,...,Y_n) = (Y_1,...,Y_m,Y_{m+1},...,Y_n)$ όπου οι $m$ πρώτες παρατηρήσεις είναι πλήρεις ενώ οι $n$-$m$ τελευταίες είναι ελλιπείς. Η πιο απλή προσέγγιση για την ανάλυση του προαναφερθέντος δείγματος είναι η **ανάλυση με χρήση μόνο των πλήρων περιπτώσεων** όπου το μέγεθος του δείγματος θεωρείται ίσο με $m$, (Little and Rubin (2002), Karla Nobrega (2002). Η μέθοδος αυτή είναι επίσης γνωστή ως **Listwise Deletion**(***LD***) ή **Casewise Deletion**(***CD***). Όλες οι μονάδες με ελλιπείς παρατηρήσεις διαγράφονται και η στατιστική ανάλυση πραγματοποιείται στις πλήρεις μονάδες. Είναι προφανές ότι αν το τυχαίο δείγμα προέρχεται από μία κατανομή $f(\boldsymbol{y})$ το δείγμα μεγέθους $m$ προέρχεται από την $f_{\boldsymbol{Y}|\boldsymbol{R}=1}$ όπου $\boldsymbol{R}$ είναι ο ΔΕΤ για κάθε μονάδα. Τα συμπεράσματα που προκύπτουν είναι πιο ασφαλή όταν ο μηχανισμός ελλιπών τιμών είναι *MCAR*, καθώς τότε το παρατηρηθέν δείγμα είναι απλό τυχαίο δείγμα του πλήρους (δηλαδή $f = f_{\boldsymbol{Y}|\boldsymbol{R}=1}$) και η CC ανάλυση δίνει αμερόληπτες εκτιμήτριες για τις παραμέτρους του πληθυσμού. Το πλεονέκτημα της μεθόδου εντοπίζεται στο ότι όλες οι μονάδες που χρησιμοποιεί είναι πλήρεις και επομένως

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

μπορούν να εφαρμοστούν όλες οι γνωστές πολυδιάστατες μέθοδοι (πχ. να εκτιμηθούν συντελεστές παλινδρόμησης, συνδιασπορές κλπ).

Τα μειονεκτήματα της μεθόδου είναι τα ακόλουθα.

- Η συγκεκριμένη μέθοδος παρουσιάζει απώλεια σημαντικού μεγέθους πληροφορίας η οποία είναι διαθέσιμη στις μονάδες με ελλιπείς παρατηρήσεις. Σε περιπτώσεις μάλιστα που η διάσταση των μονάδων είναι αρκετά μεγάλη, αυξάνει και η πιθανότητα κάποια μονάδα να περιέχει ελλιπείς τιμές και ως εκ τούτου να αγνοηθεί τελείως από την CC ανάλυση.
- Στην περίπτωση που ο μηχανισμός ελλιπών τιμών δεν είναι *MCAR* ενδέχεται τα αποτελέσματα να μην είναι αμερόληπτα. Οι Rosenbaum and Rubin (1983) αναφέρουν ότι η συγκεκριμένη μέθοδος είναι καλή για την εξάλειψη της μεροληψίας μόνο όταν ο μηχανισμός ελλιπών τιμών είναι *MAR*.
- Σε μία μελέτη με μεγάλο αριθμό μεταβλητών και μικρό αριθμό ελλιπών τιμών ανά μεταβλητή, παίρνουμε συνήθως σημαντικό ποσοστό ελλιπών τιμών στο σύνολο. Αυτό έχει ως άμεσο αποτέλεσμα ένα μεγάλο μέρος από τις μονάδες να απορριφθεί από το δείγμα.

Θα δώσουμε στη συνέχεια ένα παράδειγμα μέσα από το οποίο θα διευκρινιστεί η μεθοδολογία ανάλυσης δεδομένων με χρήση της μεθόδου CC.

Θεωρούμε έναν πληθυσμό με τα παρακάτω χαρακτηριστικά. Μεταβλητή $X_1$(φύλο)=1 αν πρόκειται για άνδρα και 0 εάν πρόκειται για γυναίκα. Η μεταβλητή $X_2$(ηλικία)~ $N(50,10^2)$ ακολουθεί την κανονική κατανομή με τις ανωτέρω παραμέτρους. Τέλος η μεταβλητή βάρος $Y$ θεωρείται ότι εκφράζεται μέσω των $X_1$, $X_2$ από τη σχέση $Y = \beta_0 + X_1\beta_1 + (X_2 - 20)\beta_2 + \varepsilon$, όπου $\varepsilon \sim N(0,\sigma^2)$. Από έρευνα που πραγματοποιήθηκε παλαιότερα είναι γνωστό για τον πληθυσμό ότι $\beta_0 = 125, \beta_1 = 40, \beta_2 = 1$ και $\sigma = 15$. Λαμβάνεται από τον πληθυσμό αυτό ένα δείγμα $n = 20$ ατόμων για το οποίο είναι γνωστό ότι η μεταβλητή

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

$X_1$ είναι πλήρης και οι μεταβλητές $X_2$ και $Y$ παρουσιάζουν ελλιπείς τιμές. Ο Πίνακας 2.1.1 των δεδομένων είναι ο ακόλουθος, όπου με τον αριθμό 1 συμβολίζονται οι άντρες και με 0 απεικονίζονται οι γυναίκες.

**Πίνακας 2.1.1**

Πλήρες δείγμα

| περίπτωση | φύλο | ηλικία | βάρος |
|---|---|---|---|
| 1 | 0 | 28 | 160 |
| 2 | 1 | 19 | 218 |
| 3 | 1 | 37 | 235 |
| 4 | 0 | 24 | 150 |
| 5 | 1 | 18 | 170 |
| 6 | 1 | 20 | 176 |
| 7 | 1 | 36 | 225 |
| 8 | 0 | 42 | 200 |
| 9 | 0 | 28 | 170 |
| 10 | 1 | 46 | 195 |
| 11 | 0 | 23 | 148 |
| 12 | 0 | 29 | 200 |
| 13 | 1 | 44 | 221 |
| 14 | 0 | 37 | 179 |
| 15 | 0 | 21 | 160 |
| 16 | 1 | 41 | 204 |
| 17 | 0 | 40 | 190 |
| 18 | 1 | 37 | 208 |
| 19 | 0 | 50 | 204 |
| 20 | 1 | 43 | 235 |

Στην περίπτωση που τα δεδομένα είναι πλήρη, η στατιστική ανάλυση των δεδομένων μέσα από το SPSS δίνει τα παρακάτω αποτελέσματα:

**Πίνακας 2.1.2**

Εκτιμήσεις παραμέτρων για το πλήρες δείγμα

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 140,342 | 13,622 | | 9,515 | ,000 |
| | ΦΥΛΟ | 40,127 | 7,604 | ,562 | 3,926 | ,001 |
| | ΗΛΙΚΙΑ | 1,175 | ,389 | ,531 | 3,711 | ,002 |

a. Dependent Variable: ΒΑΡΟΣ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Παρατηρείται ότι οι εκτιμίσεις για τα $b_1$ και $b_2$ είναι πολύ κοντά στις πραγματικές. Μεγάλη διαφορά εμφανίζεται μόνο στη σταθερά $b_0$ στην οποία όμως συνήθως δε δίνεται και ιδιαίτερη σημασία.

Όταν στο δείγμα υπάρχουν μονάδες με ελλιπείς τιμές, οι μονάδες στις οποίες αντιστοιχούν θα διαγραφούν και στην ανάλυση θα συμμετέχουν μόνο οι πλήρεις. Συνεπώς ένα μέρος της πληροφορίας την οποία περιέχουν χάνεται. Έτσι, αν θεωρήσουμε ως ελλιπείς τις τιμές που σημειώνονται με ? στον Πίνακα 2.1.3 θα πάρουμε μέσω του SPSS τα αποτελέσματα του Πίνακα 2.1.4.

**Πίνακας 2.1.3**

Ελλιπές δείγμα

| περίπτωση | φύλο | ηλικία | βάρος |
|---|---|---|---|
| 1 | 0 | 28 | 160 |
| 2 | 1 | ? | 218 |
| 3 | 1 | ? | ? |
| 4 | 0 | 24 | 150 |
| 5 | 1 | ? | 170 |
| 6 | 1 | ? | 176 |
| 7 | 1 | 36 | 225 |
| 8 | 0 | 42 | ? |
| 9 | 0 | 28 | 170 |
| 10 | 1 | ? | 195 |
| 11 | 0 | ? | 148 |
| 12 | 0 | 29 | ? |
| 13 | 1 | 44 | ? |
| 14 | 0 | 37 | 179 |
| 15 | 0 | 21 | ? |
| 16 | 1 | 41 | ? |
| 17 | 0 | 40 | 190 |
| 18 | 1 | ? | 208 |
| 19 | 0 | ? | ? |

**Πίνακας 2.1.4**

Εκτιμήσεις παραμέτρων για το ελλιπές δείγμα

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 151,986 | 47,270 | | -,617 | ,571 |
| | ΦΥΛΟ | 70,119 | 16,354 | -,865 | -,915 | ,412 |
| | ΗΛΙΚΙΑ | -,496 | ,283 | 1,197 | 1,267 | ,274 |

a. Dependent Variable: ΒΑΡΟΣ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Παρατηρούμε ότι τώρα οι παράμετροι $b_1$ και $b_2$ διαφέρουν κατά πολύ από τις πραγματικές παραμέτρου του πληθυσμού, κάτι το οποίο προφανώς οφείλεται στη διαγραφή των μονάδων με ελλιπείς παρατηρήσεις. Αυτή είναι και η βασική αδυναμία της μεθόδου. Στην στατιστική ανάλυση που προηγήθηκε αξιοποιήθηκαν 7 μονάδες ενώ 13 παρέμειναν εκτός.

Οι εκτιμήσεις των παραμέτρων του μοντέλου παλινδρόμησης διαφέρουν από τις πραγματικές τιμές. Οι διαπιστώσεις των Little and Rubin (2002) για την *CC* ανάλυση είναι τα παρακάτω:

1. Η *CC* ανάλυση δεν είναι πάντοτε αποτελεσματική.
2. Η *CC* ανάλυση είναι αποτελεσματική μόνο στην περίπτωση που ελλιπείς τιμές εμφανίζονται στην μεταβλητή *Y* (εξαρτημένη) του μοντέλου παλινδρόμησης.

### 2.1.2 Ανάλυση με χρήση των απαραίτητων κάθε φορά πλήρων δεδομένων

Η **ανάλυση με χρήση των απαραίτητων κάθε φορά πλήρων δεδομένων** (*Available Case Analysis* (*AC*)) είναι μία απλή μέθοδος η οποία αξιοποιεί δείγματα με μονάδες που εμφανίζουν ελλιπείς τιμές. Καλείται και **Pairwise Deletion** (***PD***) καθώς διαγράφει τις μονάδες εκείνες με ελλιπείς τιμές από κάθε στατιστική ανάλυση, βασιζόμενη στις μεταβλητές οι οποίες χρησιμοποιούνται για τη συγκεκριμένη ανάλυση. Είναι ιδιαίτερα χρήσιμη σε πολυδιάστατα δεδομένα.

Έστω ότι $\boldsymbol{Y} = (\boldsymbol{Y_1},...,\boldsymbol{Y_n})$ είναι ένα τυχαίο δείγμα με $\boldsymbol{Y_i} = (Y_{i1},...,Y_{ik})$ και $\boldsymbol{R} = (\boldsymbol{R_1},...,\boldsymbol{R_n})$ ο ΔΕΤ για κάθε μονάδα. Σύμφωνα με τους Little and Rubin (1987) στην *AC* για κάθε μονάδα $\boldsymbol{Y}_i$ χρησιμοποιούνται όλες οι παρατηρήσεις με $R_j = 1$ για κάθε $j = 1,2,...,k$. Για παράδειγμα όταν σκοπός είναι να εκτιμηθεί ο πίνακας συσχετίσεων από το ανωτέρω πολυδιάστατο τυχαίο δείγμα, τότε η *CC* ανάλυση θα χρησιμοποιήσει το πλήρες δείγμα, διαγράφοντας όλες τις μονάδες με ελλιπείς τιμές. Αντιθέτως η *AC* ανάλυση θα χρησιμοποιήσει όλες τις παρατηρήσεις οι οποίες είναι γνωστές από κοινού για τις μονάδες $Y_j$ και $Y_k$ ώστε να

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

εκτιμήσει την συσχέτιση της $Y_j$ και $Y_k$. Στην *AC* ανάλυση, για τον υπολογισμό της εκάστοτε συσχέτισης, υπάρχει ένα πλήρες δείγμα (πχ πλήρεις παρατηρήσεις για τον υπολογισμό της συσχέτισης της μονάδας $Y_j$ και $Y_k$). Συνεπώς, η μέθοδος κάνει καλύτερη χρήση της διαθέσιμης πληροφορίας από την *CC* ανάλυση. Το μέγεθος του πλήρους δείγματος διαφέρει από συσχέτιση σε συσχέτιση, δημιουργώντας πιθανά προβλήματα στην περίπτωση που ο μηχανισμός ελλιπών τιμών δεν είναι *MCAR* ή οι μεταβλητές είναι υψηλά συσχετισμένες. Στην παρουσία υψηλής συσχέτισης μεταξύ των μεταβλητών, η μέθοδος δεν εγγυάται ότι ο πίνακας συσχετίσεων θα είναι θετικά ορισμένος, κάτι το οποίο διαπίστωσε με παραδείγματα ο Haitovsky (1968). Αν και η *AC* ανάλυση είναι εύχρηστη, τα τυπικά σφάλματα που προκύπτουν από τους υπολογισμούς είναι πολύπλοκα και σύνθετα. Σε γενικές γραμμές η μέθοδος δεν συνιστάται, ακόμη και όταν ο μηχανισμός είναι *MCAR*.

### 2.1.3 Ανάλυση με την χρήση βαρών

Μία διαφορετική προσέγγιση για την εκτίμηση ελλιπών τιμών είναι η χρήση της **μεθόδου βαρών** (*weighting method*). Σκοπός της μεθόδου είναι ο μετασχηματισμός της κατανομής των μονάδων του δείγματος (αφού στο δείγμα υπάρχουν μονάδες με πλήρεις και ελλιπείς παρατηρήσεις) ώστε να ανταποκρίνεται στην κατανομή του πληθυσμού από τον οποίο έχουν ληφθεί. Η χρήση της μεθόδου είναι διαδεδομένη στην τοπογράφηση, χαρτογράφηση και γενικά σε οποιαδήποτε μελέτη πραγματοποιείται σε πληθυσμούς και για οποιοδήποτε σκοπό. Πολλές έρευνες χρησιμοποιούν στρωματοποιημένη δειγματοληψία (*stratified*) ή σύνολα ομοειδών ατόμων (*cluster*). Όταν το δείγμα που επιλέγεται δεν αποτελεί απλό τυχαίο δείγμα, πρέπει να ληφθεί υπ'όψιν η πιθανότητα κάποια μονάδα να συμπεριληφθεί σε αυτό. Σε αυτό βασίζεται και η συγκεκριμένη μέθοδος.

Αποτελεί μία διαφοροποίηση της *CC* ανάλυσης διότι χειρίζεται με ιδιαίτερο τρόπο μονάδες (άτομα) που παρουσιάζουν ελλιπείς τιμές (*non response*). Η μέθοδος διαγράφει τις μονάδες με ελλιπείς τιμές και ρυθμίζει τα δειγματικά βάρη των πλήρων μονάδων στην έρευνα ώστε



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

να αντισταθμίσουν εκείνες τις μονάδες που είναι ελλιπείς. Σύμφωνα με τους Little and Rubin (2002), Robins, Rotnitzky, Zao (1995), Horton and Lipsitz (1999), η ακριβής διαδικασία ανάλυσης μέσω της μεθόδου των βαρών είναι η επόμενη.

Θεωρείται ένας πληθυσμός $P$, μεγέθους $N$, ο οποίος είναι διαμερισμένος σε $C$ ομάδες μεγέθους $N_1, N_2, ..., N_C$. Στον πληθυσμό αυτό υπάρχουν μονάδες με πλήρεις αλλά και μονάδες με ελλιπείς παρατηρήσεις. Είναι επιθυμητή η λήψη δείγματος μεγέθους $n$ και η διαμέρισή του σε ομάδες μεγέθους $n_1, n_2, ..., n_c$. Επιλέγεται ένα τυχαίο δείγμα σύμφωνα με καποιο δειγματικό σχέδιο (στρωματοποιημένη δειγματοληψία ή σύνολα ομοειδών ατόμων και υποθέτουμε ότι οι μονάδες είναι ανεξάρτητες η μία από την άλλη. Προφανώς στο δείγμα θα συμπεριλαμβάνονται και μονάδες με ελλιπείς παρατηρήσεις. Σκοπός της μεθόδου είναι η αντικατάσταση των μονάδων αυτών με άλλες από τον ίδιο πληθυσμό οι οποίες θα είναι πλήρεις ώστε να εκτιμηθεί το συνολικό μέγεθος του πληθυσμού. Η μέθοδος αναλύεται παρακάτω.

Η πιθανότητα μία μονάδα με πλήρεις παρατηρήσεις από τον πληθυσμό να αντικαταστήσει στο δείγμα μία μονάδα με ελλιπείς παρατηρήσεις είναι

$$p_i = P(R_i = 1 | s) = P(i \in s_r | i \in s), \; i \in s.$$

Η πιθανότητα αυτή αρχικά είναι άγνωστη και πρέπει να εκτιμηθεί. Το $s$ είναι το σύνολο των μονάδων (με πλήρεις και ελλιπείς παρατηρήσεις) στο αρχικό δείγμα, $s_r$ είναι το σύνολο των μονάδων με πλήρεις παρατηρήσεις (υποσύνολο του $s$). Για την αντιμετώπιση του προβλήματος καθορίζεται ένα μοντέλο για την πιθανότητα $p_i$ και εκτιμάται από το $\hat{p}_i$.

Υπάρχουν απλές μέθοδοι για την εκτίμηση του $p_i$. Τα δύο συνηθέστερα μοντέλα που χρησιμοποιούνται για την εκτίμηση του $p_i$ στην πράξη είναι τα παρακάτω.

- Το λογιστικό μοντέλο $\hat{p}_i = \left(1 + e^{-z_i b}\right)^{-1}$, όπου $z_i$ είναι μία βοηθητική μεταβλητή.



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

- Οι μονάδες να είναι ομοιόμορφα κατανομημένες μέσα στις ομάδες. Δηλαδή κάθε μονάδα (με πλήρεις παρατηρήσεις) που προέρχεται από τα σύνολα $N_1, N_2, ..., N_C$ έχει την ίδια πιθανότητα να συμπεριληφθεί στην αντίστοιχη ομάδα $n_1, n_2, ..., n_c$ του δείγματος $n$, δηλαδή $p_i = p_c$ για όλα τα $i$.

Η πιθανότητα $p_c$ είναι το ποσοστό των μονάδων με πλήρεις παρατηρήσεις της ομάδας $n_c$, προτού γίνει η αντικατάσταση των ελλιπών τιμών που περιλαμβάνονται σε αυτή.

Εν συνεχεία, καθορίζονται τα νέα δειγματικά βάρη, εφόσον πλέον έχουν διαγραφεί οι μονάδες με ελλιπείς τιμές από κάθε ομάδα, με τη βοήθεια του τύπου $w_i^* = \frac{w_i}{\hat{p}_i}$, όπου $w_i$ είναι τα προγενέστερα βάρη. Οι Eltinge and Yansaneh (1997) έχουν προτείνει μία συγκεκριμένη τεχνική για τον καθορισμό των βαρών. Στην μέθοδο της ανάλυσης με τη χρήση βαρών η εκτιμήτρια του συνόλου του πληθυσμούείναι η εξής (Haziza and Nobrega (2002)):

$$\hat{Y}^* = \sum_{i \in s_r} w_i^* y_i = \sum_{i \in s_r} \frac{w_i}{\hat{p}_i} y_i ,$$

όπου $y_i$ είναι η τιμή που έχει καταγραφεί για την $i$ μονάδα.

Αποδεικνύεται ότι η $\hat{Y}^*$ είναι αμερόληπτη εκτιμήτρια της $Y$ αν και μόνο αν $\hat{p}_i = p_i$.

Αν τώρα ο πληθυσμός $P$ δεν ήταν αρχικά διαμερισμένος σε C ομάδες, η εκτιμήτρια της $Y$ δίνεται από τον τύπου

$$\hat{Y}_1^* = \sum_{i \in s_r} w_i^* y_i = \sum_{i \in s_r} \frac{w_i}{\hat{p}} y_i = \sum_{i \in s_r} w_i y_i \left( \frac{\sum_{i \in s_r} w_i}{\sum_{i \in s} w_i} \right),$$

όπου $\hat{p} = \left( \frac{\sum_{i \in s_r} w_i}{\sum_{i \in s} w_i} \right)$.

Η εκτιμήτρια $\hat{Y}_1^*$ δεν είναι αμερόληπτη. Πιο συγκεκριμένα ισχύει

$$Bias(\hat{Y}_1^*) = E_p E_r (\hat{Y}_1^* - Y) = \frac{1}{P} \sum_P (p_i - \bar{P})(y_i - \bar{Y}),$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

όπου $\bar{P} = \frac{1}{N}\sum_{P} p_i$ , (Haziza and Nobrega (2002)).

Η μεροληψία της εκτιμήτριας $\hat{Y}_1^*$ είναι μηδέν στις επόμενες δύο περιπτώσεις:

1. $p_i = p \; \forall i \in P$ (ομοιόμορφη κατανομή του πληθυσμού).
2. $y_i$ σταθερό για κάθε $i \square P$E

Οι δύο προαναφερθείσες συνθήκες όμως δεν εμφανίζονται στην πράξη.. Όταν ο πληθυσμός $P$ χωρίζεται σε $C$ ομάδες $P_1,...,P_C$ μεγέθους $N_1,...,N_C$ γίνεται πάντα με τέτοιο τρόπο ώστε

$$P = \sum_{n=1}^{C} P_n \quad \text{όπου} \quad P_n \cap P_l = \varnothing \quad \forall v \neq l .$$

Με την οργάνωση αυτή προκύπτουν $s_1,...,s_c$ ομάδες με μέγεθος $n_1,...,n_c$ έτσι ώστε

$$s_n = s \cap P_n , \; v = 1,...,C .$$

Τη χρονική στιγμή κατά την οποία θα έχουν δημιουργηθεί οι ομάδες, υπολογίζονται τα δειγματικά βάρη και ο ερευνητής μπορεί να συνεχίσει τη στατιστική ανάλυση έχοντας παράλληλα περιορίσει τις ελλιπείς τιμές τις οποίες είχε εξ αρχής.

Η μέθοδος παρουσιάζει τους επόμενους πολύ σημαντικούς περιορισμούς.

- Μεγάλο μέρος της πληροφορίας που περιέχεται στις ελλιπείς τιμές χάνεται και για το λόγο αυτό η μέθοδος κρίνεται ως αναποτελεσματική.
- Οι εκτιμήτριες που προκύπτουν έχουν μη αποδεκτή υψηλή διακύμανση.
- Απαιτεί πάντοτε τον υπολογισμό βαρών.
- Απαιτεί πολύπλοκους υπολογισμούς διακυμάνσεων.

Τα πλεονεκτήματα της μεθόδου είναι τα παρακάτω:

- Είναι ιδιαίτερα εύχρηστη μέθοδο όταν η δειγματοληψία που εφαρμόζεται είναι είτε στρωματοποιημένη είτε βασίζεται σε σύνολα ομοειδών ατόμων.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

- Μπορεί να διατηρήσει την πολυμεταβλητή δομή των δεδομένων .
- Τα βάρη μπορεί να ληφθούν υπ'όψιν στον υπολογισμό διακυμάνσεων.

## 2.2 ΔΙΑΔΙΚΑΣΙΑ ΕΠΑΝΑΛΗΠΤΙΚΩΝ ΕΠΙΣΚΕΨΕΩΝ

Μία από τις πιο συνηθισμένες περιπτώσεις εμφάνισης ελλιπών τιμών είναι οι μη ανταποκρινόμενοι σε μία έρευνα, Haziza and Nobrega (2002). Στη συγκεκριμένη κατηγορία περιλαμβάνονται οι αρνούμενοι να συνεργασθούν (***refusals***), οι απόντες κατά την επίσκεψη του ερευνητή (***non conducts***), αυτοί που έκαναν αλλαγή διεύθυνσης διαμονής ή εργασίας, αυτοί που έχουν ανικανότητα (***incapacity***) παροχής πληροφοριών και αυτοί που δεν εντοπίζονται (***not found***). Το πρόβλημα της μη ανταπόκρισης έχει ως συνέπεια είτε τη μείωση του μεγέθους του δείγματος είτε την εμφάνιση ελλιπών τιμών. Το πλέον σοβαρό πρόβλημα για τη μελέτη δεδομένων τέτοιας μορφής είναι η διαχείριση περιπτώσεων όπου το ερευνώμενο χαρακτηριστικό έχει διαφορετική κατανομή σε ανταποκρινόμενους και μη ανταποκρινόμενους.

Το πρόβλημα αντιμετωπίζεται με **επαναληπτικές επισκέψεις** (***follow up***) στους μη ανταποκρινόμενους, με σκοπό την ελαχιστοποίησή τους, ή με αντικαταστάσεις των μη ανταποκρινόμενων μελών με άλλα μέλη που θεωρούνται ότι έχουν συναφείς ιδιότητες.

Τα κύρια πλεονεκτήματα της μεθόδου είναι ότι

- συμπληρώνει τα δεδομένα του δείγματος με πραγματικά στοιχεία, διότι αντικαθιστά άτομα που δεν ανταποκρίθηκαν στην έρευνα με άλλα που έχουν συναφείς ιδιότητες και
- διατηρεί την πολυμεταβλητή δομή των δεδομένων,

ενώ τα βασικά της μειονεκτήματα εστιάζονται στο ότι

- έχει αυξημένο αυξημένο κόστος,
- είναι ιδιαίτερα χρονοβόρα, και
- συνήθως είναι αδύνατο να ολοκληρωθεί επιτυχώς η διαδικασία μέσα στους χρονικούς περιορισμούς που θέτει ο ερευνητής.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

## 2.3 ΜΕΘΟΔΟΙ ΑΝΤΙΚΑΤΑΣΤΑΣΗΣ ΕΛΛΙΠΩΝ ΤΙΜΩΝ

Το επόμενο βήμα, μετά τις *CC*, *AC*, την ανάλυση με χρήση βαρών και τη μέθοδο επαναληπτικών επισκέψεων, είναι η προσπάθεια συμπλήρωσης των ελλιπών τιμών. Στην πράξη προϋποθέτει ότι ο ερευνητής πρέπει να μαντέψει με κάποιο τρόπο τις ελλιπείς τιμές που παρουσιάζουν κάποιες μονάδες, κάνοντας χρήση των παρατηρηθέντων (γνωστών) τιμών των υπολοίπων μονάδων.

Σύμφωνα με την κλασσική στατιστική, τα αποτελέσματα που προκύπτουν από μια έρευνα δεν πρέπει να βασίζονται εξ ολοκλήρου στο δείγμα από το οποίο προήλθαν. Είναι απαραίτητο η μελέτη να επαναληφθεί με διαφορετικό δείγμα και τα αποτελέσματά της να είναι παραπλήσια των αρχικών. Τα συμπεράσματα της έρευνας δεν θα πρέπει να στηρίζονται αποκλειστικά στις μονάδες του συγκεκριμένου δείγματος. Κάθε μονάδα ενός τυχαία επιλεγμένου δείγματος μπορεί να αναπληρωθεί από μία νέα, τυχαία επιλεγμένη μονάδα από τον ίδιο πληθυσμό, χωρίς να αλλοιώσει τα αποτελέσματα. Οι μέθοδοι της **αντικατάστασης** (*imputation*) στηρίζονται στη βασική αυτή αρχή.

Για παράδειγμα, σε ένα διαγνωστικό test η τιμή του οποίου είναι ενδεικτική του αν έχει νοσήσει ένα άτομο ή όχι είναι δυνατό να αναπληρώσουμε οποιονδήποτε ασθενή δεν έχει νοσήσει αλλά εμφανίζει ελλιπή τιμή στο αποτέλεσμα του test, με ένα νέο, από τον ίδιο πληθυσμό που να μην έχει νοσήσει όμως η τιμή στο test να είναι γνωστή. Ας σημειωθεί ότι τα γενικά χαρακτηριστικά του νέου ασθενή δεν απαιτείται να είναι τα ίδια με αυτά που έχει ο αρχικά επιλεγμένος. Το αποτέλεσμα θα είναι ένα νέο τυχαίο δείγμα από τον πληθυσμό και τα συμπεράσματα που θα ακολουθήσουν της ανάλυσης θα είναι εξίσου έγκυρα. Φυσικά αν το άτομο που δεν έχει νοσήσει και εμφανίζει ελλιπή τιμή στο test είναι άντρας, η αντικατάστασή του θα πρέπει να γίνει από το σύνολο των αντρών που δεν έχουν εμφανίσει την νόσο αλλά η τιμή του test είναι γνωστή.

Θα δούμε στη συνέχεια μερικές από τις μεθόδους αντικατάστασης ελλιπών τιμών.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

### 2.3.1 Η μέθοδος του Buck

Η πλέον γνωστή ίσως μέθοδος συμπλήρωσης είναι η **συμπλήρωση δεσμευμένων μέσων τιμών του Buck** (*Buck's method*). Επειδή η συγκεκριμένη μέθοδος στηρίζεται στη χρήση γραμμικού μοντέλου παλινδρόμησης είναι εξίσου γνωστή ως **αντικατάσταση με τη χρήση μοντέλου παλινδρόμησης** (*Regression Imputation*).

Έστω το τυχαίο δείγμα

$$(\boldsymbol{L_1},...,\boldsymbol{L_n})$$

με $\boldsymbol{L}=(Y,X_1,...,X_p)$ στο οποίο η μεταβλητή $Y$ είναι ελλιπής σε κάποια σημεία ενώ οι μεταβλητές $X_i$ είναι πλήρεις. Έστω επίσης το πολλαπλό γραμμικό μοντέλο με ανεξάρτητες μεταβλητές τις $X_1,...,X_p$ και εξαρτημένη την $Y$, δηλαδή

$$E(Y \mid X_1,...,X_p)=\beta_0+\beta_1 X_1+...+\beta_p X_p .$$

Μία λογική διαδικασία εκτίμησης των ελλιπών τιμών θα ήταν, αφού εκτιμηθούν οι συντελεστές $b_i$ με τη μέθοδο ελαχίστων τετραγώνων (χρησιμοποιώντας τα άτομα για τα οποία διαθέτουμε πλήρεις παρατηρήσεις- μεθόδους *CC*), στη συνέχεια η ελλιπής τιμή της $i$ μονάδας της μεταβλητής $Y$, να συμπληρωθεί από την

$$\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1 X_{i1}+...+\hat{\beta}_p X_{ip}$$

δηλαδή από την εκτίμηση της μέσης τιμής απόκρισης που προκύπτει από το μοντέλο με τα πλήρη δεδομένα.

Τα δεδομένα $(Y_1,...,Y_k,\hat{Y}_{k+1},...,\hat{Y}_n)$ (οι $n-k$ τελευταίες μονάδες είχαν ελλιπείς τιμές) θεωρούνται ως πλήρη δεδομένα και είναι δυνατός ο υπολογισμός των παραμέτρων της τυχαίας μεταβλητής $Y$.

Στην περίπτωση που ο μηχανισμός ελλιπών τιμών είναι *MCAR* ή ακόμα υπό την ασθενέστερη συνθήκη *MAR*, η παραπάνω μέθοδος δίνει γενικά καλές εκτιμήτριες της πραγματικής μέσης τιμής $m_Y$ της τμ $Y$. Από τη δειγματική μέση τιμή του συμπληρωμένου δείγματος έχουμε

$$\hat{\mu}_Y=\frac{1}{n}\left(\sum_{i=1}^{k} Y_i+\sum_{i=k+1}^{n}\hat{Y}_i\right).$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Θα πρέπει να τονιστεί ότι, όταν ο μηχανισμός ελλιπών τιμών είναι *MAR*, για να είναι αμερόληπτη η παραπάνω εκτιμήτρια, θα πρέπει να ισχύει η υπόθεση της γραμμικής εξάρτησης της δεσμευμένης μέσης τιμής από τις μεταβλητές $X_i$, δηλαδή ότι $E(Y \mid X_1,...,X_p) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$.

Ιδιαίτερη προσοχή απαιτείται όταν η μέθοδος εφαρμόζεται σε μονάδες που παρουσιάζουν ελλιπή τιμή στη μεταβλητή $Y$, αλλά με ακραίες τιμές σε κάποιο ή κάποια από τα $X_i$.

Παρ'όλα τα καλά αποτελέσματα της μεθόδου για την εκτίμηση της μέσης τιμής, δεν προκύπτουν εξίσου ικανοποιητικά αποτελέσματα αν τα δεδομένα που προκύπτουν από αυτή χρησιμοποιηθούν για την εκτίμηση άλλων παραμέτρων όπως π.χ. της διασποράς $s_Y^2$. Αυτό οφείλεται στο γεγονός ότι, ενώ οι ελλιπείς τιμές αναμένεται να βρίσκονται γύρω από την ευθεία παλινδρόμησης, στην πραγματικότητα αντικαθίστανται με τιμές επάνω στην εκτιμηθείσα ευθεία (ουσιαστικά με τη μέση τιμή τους). Έτσι ένα μέρος της διασποράς των ελλιπών τιμών χάνεται με αποτέλεσμα η δειγματική διασπορά να υποεκτιμά την πραγματική διασπορά της τμ $Y$, η οποία δίνεται από τον τύπο

$$S_Y^2 = \frac{1}{n-1}\left( \sum_{i=1}^{k} (Y_i - \hat{\mu}_Y)^2 + \sum_{\iota=\kappa+1}^{n} \left(\hat{Y}_i - \hat{\mu}_Y\right)^2 \right).$$

ΠΑΡΑΤΗΡΗΣΕΙΣ

- Το φαινόμενο της υποεκτίμησης είναι συνέπεια του ότι τα δεδομένα που προκύπτουν από τη μέθοδο αυτή δεν προέρχονται των πραγματικά δεδομένα.
- Στην πραγματικότητα η κατανομή των δεδομένων θεωρείται γνωστή (εκτός κάποιων παραμέτρων της) από την στιγμή που υιοθετείται το γραμμικό μοντέλο.
- Μια τετριμμένη περίπτωση της μεθόδου του Buck είναι η μονοδιάστατη περίπτωση (δηλαδή $p = 0$) στην οποία κάθε ελλιπής τιμή αντικαθίσταται από τη μέση τιμή των παρατηρηθέντων. Είναι προφανές ότι με την αντικατάσταση αυτή, η μεν δειγματική μέση τιμή δεν αλλάζει, η δε δειγματική διασπορά ελαττώνεται,

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

και μάλιστα η ελάττωση αυτή μεγαλώνει όσο περισσότερες είναι οι ελλιπείς τιμές. Το παράδειγμα αυτό φανερώνει και την βασική αδυναμία της μεθόδου.

### 2.3.2 Ad-hoc μέθοδοι

Από τους Horton and Kleinman (2007) έχει προταθεί μία σειρά ad-hoc μεθόδων με σκοπό το χειρισμό και την αντιμετώπιση των ελλιπών τιμών σύμφωνα με του. Μια προσέγγιση για συνεχείς τυχαίες μεταβλητές περιλαμβάνει τη δημιουργία μίας νέας μεταβλητής η οποία θα είναι ο ΔΕΤ (εδώ προτιμάται ο δεύτερος ορισμός). Για όσες μονάδες εμφανίζουν πλήρεις παρατηρήσεις σε όλες τις μεταβλητές, ο ΔΕΤ θα πάρει την τιμή ένα. Εκείνες όμως οι μονάδες που παρουσιάζουν έστω και μία ελλιπή τιμή σε κάποια από τις μεταβλητές τότε ο ΔΕΤ θα πάρει την τιμή μηδέν. Έπειτα οι συνεχείς τυχαίες μεταβλητές, ο ΔΕΤ καθώς και η αλληλεπίδρασή τους χρησιμοποιούνται από τον αναλυτή σε οποιοδήποτε μοντέλο παλινδρόμησης έχει υιοθετήσει για την ανάλυσή του. Η τεχνική είναι επίσης γνωστή με την ονομασία **Imputation with dummy**.

Μία παρόμοια μέθοδος για διακριτές τυχαίες μεταβλητές περιλαμβάνει τη δημιουργία μιας νέας κατηγορίας που θα είναι ο δείκτης ελλιπών τιμών. Και οι δύο τεχνικές παρουσιάζουν μεροληψία αφού οδηγούν σε υποεκτίμηση της διακύμανσης. Οι μέθοδοι προτιμώνται όταν ο μηχανισμός ελλιπών τιμών είναι *MCAR*, σύμφωνα με τους Jones (1996) και Greenland and Finkle (1995).

Μία δεύτερη προσέγγιση περιλαμβάνει τον αποκλεισμό από την ανάλυση εκείνων των μεταβλητών που παρουσιάζουν ελλιπείς τιμές σε μεγάλο ποσοστό. Η τεχνική δεν είναι καθόλου ελκυστική αφού μπορεί να οδηγήσει στην απόρριψη σημαντικής πληροφορίας από οποιοδήποτε μοντέλο παλινδρόμησης με πιθανές συνέπειες τη μεροληψία και τα μεγάλα τυπικά σφάλματα, (Carpenter (2004), Cook, Zeng, Yi (2004), Jansen (2006)).

### 2.3.3 Αντικατάσταση με χρήση τυχαίου παράγοντα

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Οι Haziza and Nobrega (2002) παρουσιάζουν τη μέθοδο **αντικατάστασης με χρήση τυχαίου παράγοντα** (*random regression imputation*) η οποία χρησιμοποιεί τον ίδιο μηχανισμό αντικατάστασης ελλιπών τιμών με την μέθοδο του Buck (δηλαδή χρήση κάποιου γραμμικού μοντέλου παλινδρόμησης). Η μόνη διαφορά είναι ένας τυχαίος παράγοντας ο οποίος προστίθεται στην τιμή που έχει εκτιμηθεί για τις προβλεπόμενες τιμές ($\hat{Y}_i$) της τμ $Y$ που αρχικά ήταν ελλιπείς. Ο τυχαίος αυτός παράγοντας διαφέρει σε κάθε εκτιμηθείσα παρατήρηση και συνήθως βασίζεται στα σφάλματα (*residuals*) που προκύπτουν από το μοντέλο της παλινδρόμησης. Η χρήση του παράγοντα αυτού αποβλέπει στην καλύτερη εκτίμηση της δειγματικής διασποράς.

### 2.3.4 Μέθοδος απλής αντικατάστασης

Μία αρκετά διαδεδομένη και εύχρηστη μέθοδος αντικατάστασης ελλιπών τιμών είναι η **απλή αντικατάσταση** (*Single Imputation*). Η μέθοδος παράγει μία τεχνητή τιμή ώστε να αντικαταστήσει την ελλιπή. Οι συνήθεις τεχνικές αναπλήρωσης περιλαμβάνουν

1. τη **μέση τιμή** (*mean*)
2. τη **διάμεσο** (*median*)
3. την **επικρατούσα τιμή** (*mode*)
4. την **αναλογία** ή **λογο** (*ratio*)
5. τη **hot-deck** και
6. τον **πλησιέστερο γείτονα** (*nearest neighbour*).

Έστω $P$ ένας πληθυσμός μεγέθους $N$ για τον οποίο επιχειρείται να εκτιμηθεί η πληθυσμιακή μέση τιμή ενός χαρακτηριστικού, $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. Για το σκοπό αυτό επιλέγεται ένα απλό τυχαίο δείγμα μεγέθους $n$. Είναι γνωστό ότι ο δειγματικός μέσος $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ αποτελεί αμερόληπτη εκτιμήτρια της μέσης τιμής του πληθυσμού $\bar{Y}$, στην περίπτωση που όλες οι παρατηρήσεις είναι γνωστές. Όταν στο δείγμα εμφανιστούν ελλιπείς τιμές, ο δειγματικός μέσος δεν αποτελεί

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

αμερόληπτη εκτιμήτρια της μέσης τιμής του πληθυσμού. Ως εκτιμήτρια για τον $\bar{Y}$ θα μπορούσε να χρησιμοποιηθεί η ποσότητα

$$\bar{y} = \frac{1}{n}\left( \sum_{i \in r} y_i + \sum_{i \in m} y_i^* \right),$$

όπου $r$ είναι το σύνολο των μονάδων της τμ $Y$ που οι τιμές τους είναι πλήρεις, $m$ είναι το σύνολο των μονάδων που οι τιμές τους είναι ελλιπείς και $y_i^*$ είναι μία τιμή που δημιουργείται, με σκοπό να αντικαταστήσει μία ελλιπή τιμή του δείγματος.

Ο Πίνακας 2.3.1 συνοψίζει τις διαδικασίες κατασκευής των τιμών αυτών.

**Πίνακας 2.3.1**

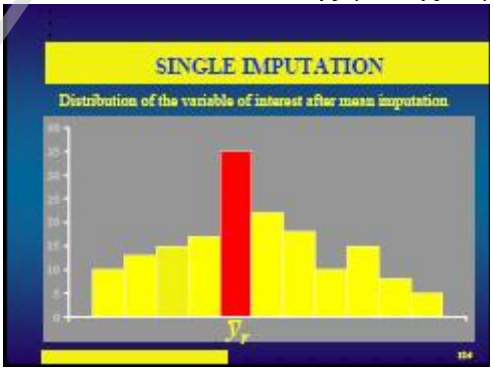
Μέθοδοι απλής αντικατάστασης

| Imputation Method | Mean Imputation | Ratio Imputation | Regression Imputation | Random Hot-Deck Imputation | Nearest-Neighbour Imputation |
|---|---|---|---|---|---|
| Imputed Value | $y_i^* = \frac{1}{r_{i,m}} y_r = \bar{y}_r$ | $y_i^* = \frac{\bar{y}_r}{\bar{x}_r} x_i$ | $y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}$ | $y_i^* = y_j$ for some $j \in s_r$ such that $d(x_i^*, x_j) = \min_{k \in r} d(x_i, x_k)$ | $y_i^* = y_j$ for some $j \in s_r$ such that $dist(x_i, x_j)$ is minimum |
| Auxiliary Variable(s)? | NO | YES (one) | YES (two or more) | NO | YES (one or more) |

Στη μέθοδο αντικατάστασης μέσων τιμών (*mean imputation*) κάθε ελλιπής τιμή αντικαθίσταται από τη μέση τιμή που προκύπτει από τις πλήρεις μονάδες, Haziza and Nobrega (2002). Η ίδια ακριβώς διαδικασία

**Πίνακας 2.3.2**

Ιστόγραμμα συχνοτήτων ύστερα από την εφαρμογή της μεθόδου αντικατάστασης μέσης τιμής




ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ακολουθείται με τη διάμεσο αλλά και με την επικρατούσα τιμή. Φυσικά ο ερευνητής θα κρίνει ποια θα επιλέξει, ανάλογα με την περίπτωση που αντιμετωπίζει κάθε φορά.

Όταν αντικατασταθούν όλες οι ελλιπείς τιμές από τη μέση τιμή των πλήρων μονάδων, τότε κατασκευάζοντας ένα ιστόγραμμα διαπιστώνουμε μία έντονη αύξηση στη στήλη της μέσης τιμής. Μάλιστα σε πολλές περιπτώσεις η μέση τιμή γίνεται πλέον και επικρατούσα τιμή του δείγματος (Πίνακας 2.3.2).

Στην αντικατάσταση λόγου (***ratio imputation***), είναι απαραίτητη η παρουσία δύο τυχαίων μεταβλητών $Y$ και $X$ όπου η πρώτη εμφανίζει ελλιπείς τιμές, ενώ η δεύτερη είναι πάντα πλήρης. Η μέθοδος εφαρμόζεται σε μοντέλα παλινδρόμησης με δύο τυχαίες μεταβλητές. Κάθε μονάδα $y_i$ με ελλιπή τιμή αντικαθίσταται από την προβλεπόμενη τιμή $y_i^*$ που λαμβάνεται από τον λόγο $\frac{\bar{y}_r}{\bar{x}_r}x_i$, όπου $\bar{y}_r$ είναι η δειγματική μέση τιμή των μονάδων της τμ $Y$ που είναι πλήρεις, $\bar{x}_r$ η δειγματική μέση τιμή της τμ $X$, και $x_i$ η τιμή που αντιστοιχεί στην μονάδα $i$ για τη μεταβλητή $X$, Haziza and Nobrega (2002).

Η hot-deck αντικατάσταση είναι ένας διαφορετικός τρόπος αντικατάστασης. Οι μονάδες με ελλιπείς τιμές αντικαθίστανται από πλήρεις μονάδες οι οποίες επιλέγονται από το δείγμα με πιθανότητα $P\left(y_i^* = y_i\right) = \frac{1}{r}$. Η μέθοδος παρουσιάζει υπερεκτίμηση της διασποράς και δε βελτιώνει τη δειγματική μέση τιμή σε σχέση με τη μέση τιμή που θα προκύψει εφαρμόζοντας τη *CC* ανάλυση.

Στη μέθοδο του πλησιέστερου γείτονα (***nearest neighbour***) η ελλιπής τιμή αντικαθίσταται από την τιμή του πλησιέστερου γείτονα (πλήρης μονάδα), Haziza and Nobrega (2002), ο οποίος καθορίζεται με βάση την απόσταση δύο ή περισσοτέρων μεταβλητών. Το είδος της απόστασης πρέπει να είναι σαφώς καθορισμένο (πχ Ευκλείδεια). Στην περίπτωση πολλών μεταβλητών είναι προτιμότερο, πριν εφαρμοστεί η τεχνική αυτή, να τυποποιηθούν οι παρατηρήσεις των μονάδων (πχ αφαιρώντας τη μέση τιμή και διαιρώντας με την τυπική απόκλιση) .

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Η μέθοδος απλής αντικατάστασης παρουσιάζει πλεονεκτήματα τα οποία συνοψίζονται ακολούθως.

- Δημιουργεί πλήρη δεδομένα (απλά για κάθε ερευνητή).
- Αντίθετα με τις μεθόδους βαρών, επιτρέπει τη χρήση ενός βάρους (*single weight*).
- Επιχειρεί να χρησιμοποιήσει όλα τα δεδομένα ώστε να μη χαθεί πληροφορία.
- Είναι διαθέσιμη σε πολλά στατιστικά πακέτα διαθέτουν την επιλογή της απλής αντικατάστασης.

Τα μειονεκτήματα της μεθόδου είναι τα εξής:

- Οι τιμές που χρησιμοποιούνται για αντικατάσταση είναι τεχνητές οπότε τα αποτελέσματα που προκύπτουν στην ανάλυση δεν είναι ακριβή.
- Το γεγονός ότι οι τιμές που δημιουργούνται θεωρούνται ως παρατηρηθείσες τιμές οδηγεί σε πιθανή υποεκτίμηση της διακύμανσης των εκτιμητριών, ιδιαίτερα όταν οι ελλιπείς τιμές είναι πάρα πολλές.

### 2.3.5 Εκτίμηση της διακύμανσης των εκτιμητριών απλής αντικατάστασης

Η εκτίμηση της διακύμανσης των εκτιμητριών απλής αντικατάστασης θεωρείται αναγκαία αφού αποτελεί ένα είδος αξιολόγησης της ποιότητας των τιμών που θα δημιουργηθούν για να αντικαταστήσουν τις ελλιπείς τιμές. Συμβάλει στην εξαγωγή σωστών συμπερασμάτων ενώ παράλληλα ο ερευνητής πληροφορείται για τυχόν λανθασμένες επιλογές (εάν η διακύμανση είναι πολύ υψηλή). Υπάρχουν τρείς βασικές τεχνικές για την εκτίμηση της διακύμανσης, οι οποίες παρουσιάζονται παρακάτω.

Η πρώτη τεχνική είναι γνωστή ως **προσέγγιση με τη βοήθεια μοντέλου** (*model-assisted approach*) και προτάθηκε από τον Sarndal (1992). Έστω ότι η εκτιμήτρια για τη μέση τιμή του πληθυσμού $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ είναι η



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

$$\bar{y}_I = \frac{1}{n}\left(\sum_{i\in s_r} w_i y_i + \sum_{i\in s_m} w_i y_i^*\right),$$

όπου $w_i$ είναι τα δειγματικά βάρη.

Θεωρούμε ότι

$$\bar{y}_I - \bar{Y} = (\bar{y} - \bar{Y}) + (\bar{y}_I - \bar{y}),$$

όπου $\bar{y}$ είναι η μέση τιμή των πλήρων μονάδων του δείγματος, και

$\bar{y} - \bar{Y}$ =σφάλμα οφειλόμενο στη δειγματοληψία,

$\bar{y}_I - \bar{y}$ =σφάλμα οφειλόμενο στις ελλιπείς τιμές.

Η διακύμανση της εκτιμήτριας μπορεί να αναλυθεί ως εξής:

$$V_{tot} = V\left(\bar{y}_I - \bar{Y}\right) = E\left(\bar{y}_I - \bar{Y}\right)^2 = V_{sam} + V_{imp} + V_{mix}$$

Η συνιστώσα $V_{mix}$ είναι μηδέν για τις μεθόδους απλής αντικατάστασης ή είναι πολύ μικρή συγκρινόμενη με τις υπόλοιπες δύο διακυμάνσεις. $V_{sam}$ είναι η διακύμανση που οφείλεται στη δειγματοληψία και $V_{imp}$ είναι η διακύμανση που οφείλεται στις νέες τιμές που αντικατέστησαν τις ελλιπείς.

Δεύτερη στη σειρά μέθοδος είναι η μέθοδος **Jackknife** την οποία πρότειναν ο Rao and Shao (1992). Αρχικά λαμβάνεται ένα δείγμα μεγέθους $n$ από έναν πληθυσμό $P$ μεγέθους $N$ σύμφωνα με κάποιο δειγματικό σχέδιο. Έστω $Y$ τυχαία μεταβλητή που αντιστοιχεί στο συνολικό αριθμό του πληθυσμού ο οποίος πρέπει να εκτιμηθεί και $\hat{Y}_{HT} = \sum_{i\in s} w_i y_i$ η εκτιμήτρια του $Y$. Η διακύμανση με την τεχνική Jackknife λαμβάνεται ως εξής (θεωρείται ότι στο δείγμα δεν έχουμε παρουσία ελλιπών τιμών):

1. Αφαιρείται μία παρατήρηση από το δείγμα
2. Προσαρμόζονται τα βάρη του δείγματος
3. Υπολογίζεται η $\hat{Y}_{HT}$ με τα βάρη αυτά
4. Επαναφέρεται στο δείγμα η παρατήρηση που αφαιρέθηκε, αφαιρείται η επόμενη και υπολογίζεται ξανά η $\hat{Y}_{HT}$.
5. Επαναλαμβάνονται τα βήματα 1-4 μέχρις ότου να αφαιρεθούν όλες οι παρατηρήσεις.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Η εκτιμήτρια *Jackknife* της διακύμανσης δίνεται από τον τύπο

$$V_{Jackknife}\left(\hat{Y}_{HT}\right)=\frac{n-1}{n}\sum_{j=1}^{n}\left(\hat{Y}_{HT_{(j)}}-\hat{Y}_{HT}\right)^2,$$

όπου

$$\hat{Y}_{HT_{(j)}}=\sum_{i\neq j\in s} w_{i(j)}y_i \ ,$$

και

$$w_{i(j)}=\begin{cases}\dfrac{n}{n-1}w_i, & i\neq j\\[2ex] 0, & i=j\end{cases}$$

ΠΑΡΑΤΗΡΗΣΗ

- Η εκτιμήτρια διακύμανσης με την τεχνική Jackknife δίνει πολύ καλά αποτελέσματα, όταν το δείγμα είναι μεγάλο και όταν η μέθοδος αντικατάστασης που εφαρμόζεται είναι εκείνη του λόγου (*ratio*). Σε διαφορετική περίπτωση, τα αποτελέσματα είναι φτωχά.

Μία τρίτη τεχνική αποτελεί η μέθοδος **Bootstrap.** Προτάθηκε από τους Shao and Sitter (1996) και περιλαμβάνει τα επόμενα βήματα.

1. Επιλέγεται ένα δείγμα bootstrap $s^*$ (απλό τυχαίο δείγμα με επανάθεση) μεγέθους $n^*=n-1$ από το δείγμα $s$ (το οποίο περιλαμβάνει τις μονάδες με πλήρεις παρατηρήσεις αλλά και τις μονάδες με ελλιπείς τιμές), δηλαδή $s=\{y_i : i\in s_r\}\cup\{y_j : j\in s_m\}$. Σημειώνεται ότι $s^*=s_r^*\cup s_m^*$ (όπου $s_r^*$ και $s_m^*$ αντιστοιχούν στα δείγματα των μονάδων με πλήρεις και ελλιπείς τιμές αντίστοιχα).
2. Έστω $R_i$ ο δείκτης ελλιπών τιμών ο οποίος σχετίζεται με το $y_i\in s^*$ όπως φαίνεται παρακάτω.

$$s_m^*=\{i\in s^* : R_i=0\} \text{ και } s_r^*=\{i\in s^* : R_i=1\}$$

Στη συνέχεια χρησιμοποιείται οποιαδήποτε τεχνική απλής αντικατάστασης για να αντικατασταθούν οι ελλιπείς τιμές στο $s^*$.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

3. Τέλος υπολογίζεται η εκτιμήτρια

$$\hat{y}_I^* = \frac{1}{N}\left[\sum_{i\in s_r^*} w_i^* y_i + \sum_{i\in s_m^*} w_i^* y_i\right],$$

όπου

$$w_i^* = w_i \frac{n}{n-1}$$

Επαναλαμβάνονται τα βήματα 1-3 $B$ φορές (το $B$ καθορίζεται από τον ερευνητή, ανάλογα με το πόσες φορές επιθυμεί να γίνει η επανάληψη).

Η διακύμανση με την τεχνική bootstrap δίνεται από τον παρακάτω τύπο

$$V_{Bootstrap} = \frac{1}{B-1}\left[\sum_{b=1}^{B}\left(\bar{y}_{I(b)}^* - \bar{\bar{y}}_I^*\right)^2\right],$$

όπου

$$\bar{\bar{y}}_I^* = \frac{1}{B}\sum_{b=1}^{B}\bar{y}_{I(b)}^* .$$

Η ποσότητα $\bar{y}_{I(b)}^*$ δηλώνει την εκτιμήτρια που υπολογίστηκε στο $b$-οστό βήμα του αλγορίθμου.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

# ΚΕΦΑΛΑΙΟ ΤΡΙΤΟ

## 3.1 ΕΚΤΙΜΗΤΡΙΕΣ ΜΕΓΙΣΤΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ

Οι εκτιμήτριες πιθανοφάνειας είναι μία εναλλακτική προσέγγιση για την επίλυση προβλημάτων με ελλιπείς τιμές.

Θεωρείται το πλήρες τυχαίο δείγμα μεγέθους $n$

$$\boldsymbol{Y} = (Y_1,...,Y_n),$$

του οποίου οι παρατηρήσεις προέρχονται από κάποια τυχαία μεταβλητή με κατανομή $f(\boldsymbol{y};\theta)$. Η τυχαία μεταβλητή $\boldsymbol{Y}$ μπορεί να είναι είτε μονοδιάστατη είτε πολυδιάστατη. Η από κοινού συνάρτηση κατανομής του τυχαίου δείγματος θα δίνεται από την έκφραση

$$f(\boldsymbol{y};\theta) = f(y_1,...,y_n;\theta) = \prod_{i=1}^{n} f(y_i;\theta).$$

Ως εκτιμήτρια μέγιστης πιθανοφάνειας της παραμέτρου $q$, ορίζεται η τιμή που μεγιστοποιεί την ακόλουθη συνάρτηση, για δεδομένες τιμές του $Y$.

$$L(\theta) = f(\theta \mid \boldsymbol{y}) = f(\boldsymbol{y};\theta).$$

Η $L(q)$ ονομάζεται συνάρτηση πιθανοφάνειας και είναι η από κοινού κατανομή του τυχαίου δείγματος με τη διαφορά ότι εδώ θεωρείται σαν συνάρτηση του $q$ ενώ τα $Y_i$ θεωρούνται σταθερά (έχουν λάβει τις τιμές που παρατηρήθηκαν στο δείγμα).

Συνήθως είναι ευκολότερο, από άποψη πράξεων, να μεγιστοποιείται, αντί του $L(\theta)$ ο πιθανολογάριθμος

$$l(\theta) = \ln(L(\theta))$$

Αν η παράμετρος $q$ είναι $p$-διάστατη $\boldsymbol{\theta} = (\theta_1,...,\theta_p)$ και $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1,...,\hat{\theta}_p)$ είναι η εκτιμήτρια μέγιστης πιθανοφάνειας, τότε η $\hat{\theta}$ θα ικανοποιεί την εξίσωση μέγιστης πιθανοφάνειας

$$\frac{dL(\boldsymbol{\theta})}{d\theta} = \left( \frac{\partial L(\theta_1)}{\partial \theta_1},...,\frac{\partial L(\theta_p)}{\partial \theta_p} \right) = 0$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ή ισοδύναμα την $\frac{dl(\boldsymbol{\theta})}{d\theta}=\left(\frac{\partial l(\theta_1)}{\partial\theta_1},...,\frac{\partial l(\theta_p)}{\partial\theta_p}\right)=0$.

Σημειώνεται ότι στην περίπτωση αυτή έχουμε στην πραγματικότητα ένα σύστημα $p$ εξισώσεων με $p$ αγνώστους. Στις περισσότερες γνωστές κατανομές οι λύσεις των εξισώσεων μέγιστης πιθανοφάνειας είναι και οι αντίστοιχες εκτιμήτριες, συνεπώς το ενδιαφέρον εστιάζεται στην επίλυση των παραπάνω εξισώσεων. Στις πλέον συνηθισμένες περιπτώσεις, η λύση των εξισώσεων μέγιστης πιθανοφάνειας είναι μοναδικές και παρουσιάζουν πολύ καλές ιδιότητες όπως η αμεροληψία, συνέπεια κτλ.

## 3.2 Η ΣΥΝΑΡΤΗΣΗ ΜΕΓΙΣΤΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ ΓΙΑ ΔΕΔΟΜΕΝΑ ΜΕ ΕΛΛΙΠΕΙΣ ΤΙΜΕΣ

Η μέθοδος εκτιμητριών μέγιστης πιθανοφάνειας εφαρμόζεται, τόσο σε πλήρη δεδομένα όσο και σε δεδομένα με ελλιπείς τιμές, αρκεί να κατασκευαστούν οι κατάλληλες συναρτήσεις πιθανοφάνειας. Η κατασκευή της συνάρτησης πιθανοφάνειας έχει ως εξής:

Σύμφωνα με τους Little and Rubin (2002) αρχικά επιλέγεται μια τυχαία μεταβλητή $Y$ μεγέθους $n$, η οποία περιέχει ελλιπείς τιμές και έχει συνάρτηση πυκνότητας πιθανότητας $f(\boldsymbol{y},\boldsymbol{\theta})$ (η δειγματοληπτική διαδικασία που ακολουθήθηκε δεν παίζει ουσιαστικό ρόλο). Παράλληλα με το δείγμα που επιλέχθηκε επιλέγεται και ένα τυχαίο δείγμα μεγέθους $n$ από την τυχαία μεταβλητή $R$, όπου $R$ είναι ο δείκτης ελλιπών τιμών. Συνεπώς όλη η πληροφορία περιέχεται στο τυχαίο διάνυσμα

$$(\boldsymbol{Y},\boldsymbol{R})=(Y_1,R_1,...,Y_n,R_n)=(Y_1,...,Y_n,R_1,...,R_n).$$

Δεδομένου όμως ότι $\boldsymbol{Y}=(\boldsymbol{Y_{obs}},\boldsymbol{Y_{mis}})$ και ότι $R_i=0$ υποδηλώνει τις μονάδες με ελλιπείς τιμές, το δείγμα που περιέχει όλη την πληροφορία είναι το

$$(\boldsymbol{Y_{obs}},\boldsymbol{R}).$$

Με βάση όσα αναφέρθηκαν ανωτέρω η από κοινού συνάρτηση πυκνότητας πιθανότητας των δεδομένων με ελλιπείς τιμές είναι

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

$$f(\boldsymbol{y}_{obs}, R; \theta, \phi).$$

Η παράμετρος $q$ είναι η παράμετρος της κατανομής του αρχικού πληθυσμού η οποία πρέπει να εκτιμηθεί, ενώ η $f$ περιέχει πιθανές νέες παραμέτρους, τις οποίες ίσως να έχει εισάγει ο μηχανισμός ελλιπών τιμών στο μοντέλο. Άρα η συνάρτηση πιθανοφάνειας για τα δεδομένα με ελλιπείς τιμές είναι η παρακάτω:

$$L(\theta, \phi) = f(\theta, \phi \mid \boldsymbol{y}_{obs}, \boldsymbol{r}) = f(\boldsymbol{y}_{obs}, \boldsymbol{r}; \theta, \phi).$$

Όμως η $f(Y_{obs}, R; \theta, \phi)$ αποτελεί την περιθώρια κατανομή της $f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \boldsymbol{r}; \theta, \phi)$, οπότε

$$f(\boldsymbol{y}_{obs}, r; \theta, \phi) = \int f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, r; \theta, \varphi) d\boldsymbol{y}_{mis}. \quad (3.2.1)$$

Λαμβάνοντας υπ' όψιν την έκφραση

$$f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \boldsymbol{r}; \theta, \phi) = f(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}; \phi) f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}; \theta),$$

μπορούμε να γράψουμε την (3.2.1) στη μορφή

$$f(\boldsymbol{y}_{obs}, \boldsymbol{r}; \theta, \phi) = \int f(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}; \varphi) f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}; \theta) d\boldsymbol{y}_{mis}.$$

Συνεπώς η συνάρτηση πιθανοφάνειας είναι

$$L(\theta, \phi) = \int f(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}; \varphi) f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}; \theta) d\boldsymbol{y}_{mis}. \quad (3.2.2)$$

Η (3.2.2) είναι πολύ χρήσιμη καθώς χωρίζει τη συνάρτηση πιθανοφάνειας σε δύο παράγοντες. Ο πρώτος είναι ο μηχανισμός ελλιπών τιμών και ο δεύτερος η κατανομή του πλήρους δείγματος. Είναι προφανές ότι η γνώση του μηχανισμού ελλιπών τιμών κάνει δυνατό τον υπολογισμό της συνάρτησης πιθανοφάνειας όταν η κατανομή του πλήρους δείγματος θεωρείται ορισμένη. Πρέπει να αναφερθεί στο σημείο αυτό ότι οι παράμετροι $q$ και $f$ δεν είναι απαραιτήτως ξεχωριστές με την έννοια ότι η $f$ μπορεί να περιέχει τμήμα ή ολόκληρη την $q$. Βέβαια αυτό που είναι επιθυμητό στην πράξη είναι η ανεξαρτησία, διότι έτσι διευκολύνονται κατά πολύ οι υπολογισμοί.

## 3.3 Η ΣΥΝΑΡΤΗΣΗ ΠΙΘΑΝΟΦΑΝΕΙΑΣ ΣΤΗΝ ΠΕΡΙΠΤΩΣΗ MAR

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Η συνάρτηση πιθανοφάνειας λαμβάνει πολύ απλή μορφή στην περίπτωση που ο μηχανισμός ελλιπών τιμών είναι *MAR*. Όπως έχει ήδη αναφερθεί ο μηχανισμός *MAR* προϋποθέτει ότι η πιθανότητα εμφάνισης ελλιπών τιμών εξαρτάται μόνο από το παρατηρηθέν τμήμα των δεδομένων. Σύμφωνα με τον Rubin (1976), στην περίπτωση *MAR* έχουμε

$$f(\boldsymbol{r} \mid \boldsymbol{y}) = f(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}) = f(\boldsymbol{r} \mid \boldsymbol{y}_{obs}),$$

και βάση την σχέση (3.2.2) προκύπτει

$$\begin{aligned} L(\theta,\phi) &= \int f(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}; \varphi) f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}; \theta) d\boldsymbol{y}_{mis} = \\ &= \int f(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}; \theta) f(\boldsymbol{r} \mid \boldsymbol{y}_{obs}; \varphi) d\boldsymbol{y}_{mis} = \\ &= f(\boldsymbol{r} \mid \boldsymbol{y}_{obs}; \phi) \int f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}; \theta) d\boldsymbol{y}_{mis} = \\ &= f(\boldsymbol{r} \mid \boldsymbol{y}_{obs}; \phi)\ f(\boldsymbol{y}_{obs}; \theta), \end{aligned} \tag{3.3.1}$$

όπου $f(\boldsymbol{y}_{obs}; \theta) = \int f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}; \theta) d\boldsymbol{y}_{mis}$ είναι η περιθώρια κατανομή της κατανομής του πλήρους δείγματος.

Όπως φαίνετα από τη σχέση (3.3.1), στην περίπτωση που οι παράμετροι $q$ και $f$ είναι ξεχωριστές, η παράμετρος $q$ επηρεάζει μόνο τον παράγοντα

$$f(\boldsymbol{y}_{obs}; \theta) = \int f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}; \theta) d\boldsymbol{y}_{mis}, \tag{3.3.2}$$

οπότε προκειμένου να βρεθεί η εκτιμήτρια μέγιστης πιθανοφάνειας αρκεί να βρεθεί το $q$ που μεγιστοποιεί την (3.3.2). Σε αυτό το σημείο γίνεται αντιληπτό αυτό που ειπώθηκε και στην Παράγραφο 1.5, ότι υπό την υπόθεση *MAR* δεν απαιτείται ο ακριβής καθορισμός της συνάρτησης, $g(\boldsymbol{y}_{obs}) = f(\boldsymbol{r} \mid \boldsymbol{y}_{obs})$ προκειμένου να βρεθούν οι άγνωστες παράμετροι της κατανομής του πληθυσμού. Συμπερασματικά, όσον αφορά την ανάλυση δεδομένων με ελλιπείς τιμές, η συνθήκη *MAR* καθίσταται ικανή ώστε ο μηχανισμός ελλιπών τιμών να είναι αγνοήσιμος (*ignorable*), κάτι το οποίο δεν είναι ξεκάθαρο από την αρχή, όπως στο *MCAR*.

Όσον αφορά στην ανάλυση δεδομένων *MAR*, προτεραιότητα δίνεται στην εύρεση μεθόδων μεγιστοποίησης της συνάρτησης $f(\boldsymbol{y}_{obs}; \theta)$.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

## 3.4 Η ΣΥΝΑΡΤΗΣΗ ΠΙΘΑΝΟΦΑΝΕΙΑΣ ΣΤΗΝ ΠΕΡΙΠΤΩΣΗ MNAR

Στο σημείο αυτό παρουσιάζεται ένας τρόπος αντιμετώπισης προβλημάτων δεδομένων με ελλιπείς τιμές, ο μηχανισμός των οποίων είναι μη αγνοήσιμος *MNAR* ή *NINR*. Όταν στο δείγμα υπάρχουν δεδομένα με ελλιπείς τιμές, η συνάρτηση πιθανοφάνειας θα είναι, κατά το Little (1993)

$$f(\boldsymbol{y}_{obs},\boldsymbol{r};\theta,\varphi)=\int f(\boldsymbol{y}_{obs},\boldsymbol{y}_{mis},\boldsymbol{r};\theta,\varphi)d\boldsymbol{y}_{mis},$$

από την οποία προκύπτει η

$$L(\theta,\varphi)=\int f(\boldsymbol{r}|\boldsymbol{y}_{obs},\boldsymbol{y}_{mis};\varphi)f(\boldsymbol{y}_{obs},\boldsymbol{y}_{mis};\theta)d\boldsymbol{y}_{mis}.$$

Σε αντίθεση με το μηχανισμό *MAR* στον οποίο δεν απαιτείται ο ακριβής καθορισμός της συνάρτησης $g(\boldsymbol{y}_{obs})=f(\boldsymbol{r}\,|\,\boldsymbol{y}_{obs})$, στην περίπτωση *NINR* πρέπει να καθοριστεί το μοντέλο του μηχανισμού ελλιπών τιμών. Δηλαδή η δεσμευμένη συνάρτηση πυκνότητας πιθανότητας

$$f(\boldsymbol{r}\,|\,\boldsymbol{y}_{obs},\boldsymbol{y}_{mis};\phi).$$

Στην περίπτωση που τα δεδομένα περιέχουν και πλήρεις μεταβλητές, τότε η σχέση (3.2.2) μετασχηματίζεται ως εξής (Little (1993)

$$L(\theta,\varphi)=\int f(\boldsymbol{r}|\boldsymbol{y}_{obs},\boldsymbol{y}_{mis},\boldsymbol{z};\varphi)f(\boldsymbol{y}_{obs},\boldsymbol{y}_{mis}\,|\,\boldsymbol{z};\theta)d\boldsymbol{y}_{mis},$$

όπου το *Z* αντιστοιχεί στις πλήρεις μεταβλητές του μοντέλου.

Αν η παράμετρος $f$ είναι γνωστή, τότε ο μηχανισμός ελλιπών τιμών είναι καθορισμένος.

Το μοντέλο το οποίο παράγεται σε κάθε περίπτωση (δηλαδή σε κάθε διαφορετικό δείγμα που λαμβάνεται) είναι αρκετά εξειδικευμένο αφού απαιτεί καθορισμό της κατανομής τόσο των πλήρων παρατηρήσεων όσο και του μηχανισμού ελλιπών τιμών. Η εκτίμηση των παραμέτρων για το υπό εξέταση μοντέλο είναι γενικά ιδιαίτερα επίπονη.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

## 3.5 ΠΟΛΛΑΠΛΗ ΑΝΤΙΚΑΤΑΣΤΑΣΗ

Μία διαφοροποίηση της μεθόδου της αντικατάστασης αποτελεί η **πολλαπλή αντικατάσταση** (*multiple imputaion*) (Rubin (1987)). Έστω μία τυχαία μεταβλητή $\boldsymbol{Y} = (Y_1,...,Y_n) = (Y_1,...,Y_m, Y_{m+1},...,Y_n)$ μεγέθους $n$. Οι $m$ πρώτες μονάδες είναι πλήρεις και οι υπόλοιπες $n-m$ είναι ελλιπείς. Σε αντίθεση με τη μέθοδο της απλής αντικατάστασης κατά την οποία αντικαθίστανται οι $n-m$ ελλιπείς αυτές τιμές από ένα απλό σετ τιμών, στην πολλαπλή αντικατάσταση δημιουργείται ένας αλγόριθμος με σκοπό την πιό ασφαλή εκτίμηση των τιμών που θα αντικαταστήσουν τις ελλιπείς. Μια απλή μορφή πολλαπλής αντικατάστασης είναι η παρακάτω διαδικασία που προτάθηκε από τους Haziza and Nobrega (2002).

Έστω το διάνυσμα

$$\boldsymbol{y}_{sam} = \{y_i : i \in s\} = \{\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}\},$$

όπου το $s_r$ είναι υποσύνολο του συνόλου των δεδομένων και περιέχει τις μονάδες με πλήρεις παρατηρήσεις, το $s_m$ είναι το υποσύνολο του συνόλου των δεδομένων που περιλαμβάνει τις μονάδες με ελλιπείς τιμές, και $\boldsymbol{y}_{obs} = \{y_i : i \in s_r\}$, $\boldsymbol{y}_{mis} = \{y_i : i \in s_m\}$.

Έστω ακόμη $Q$ μία βοηθητική παράμετρος (διαβαθμήσιμη) . Η μέθοδος της πολλαπλής αντικατάστασης υποθέτει ότι

$$\frac{Q - \overline{Q}_{IM}}{\sqrt{T_M}} \sim N(0,1),$$

όπου $\overline{Q}_{IM}$ είναι η εκτιμήτρια της παραμέτρου $Q$ και δηλώνει την παράμετρο που θα χρησιμοποιείτο εάν δεν υπήρχαν ελλιπείς τιμές στο αρχικό δείγμα, και $T_M$ είναι η διακύμανση της εκτιμήτριας.

Δημιουργούνται $M = 5$ διαφορετικά δείγματα ίδιου μεγέθους ($n$) με το αρχικό. Σε αυτά οι $m$ μονάδες με πλήρεις τιμές του αρχικού δείγματος δεν έχουν αλλάξει (έχουν παραμείνει οι ίδιες) ενώ οι $n-m$ υπόλοιπες που ήταν ελλιπείς έχουν αντικατασταθεί με κάποιες άλλες διαφορετικές σε καθένα από τα 5 νέα δείγματα (η αντικατάσταση γίνεται με μέθοδο απλής αντικατάστασης).

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Σκοπός της παραπάνω διαδικασίας είναι να εκτιμηθεί η μέση τιμή του πληθυσμού από το οποίο προέρχεται το δείγμα. Για το λόγο αυτό υπολογίζεται ο δειγματικός μέσος $\overline{Q}_I^{(i)}$ σε καθένα από τα 5 πλήρη σετ δεδομένων και στη συνέχεια υπολογίζουμε τον ολικό μέσο $\overline{Q}_{IM} = \frac{1}{M}\sum_{i=1}^{M}\overline{Q}_I^{(i)}$ .

Για τον υπολογισμό της διακύμανσης της εκτιμήτριας $\overline{Q}_{IM}$ εκτιμάται πρώτα η διακύμανση εντός των ομάδων ($M=5$) με τη χρήση του τύπου

$$\overline{U}_M = \frac{1}{M}\sum_{i=1}^{M}\frac{U^{(i)}}{n},$$

όπου $U^{(i)}$ είναι η διακύμανση της ομάδας (δείγματος) $i$. Έπειτα εκτιμάται η διακύμανση μεταξύ των ομάδων, με τη χρήση του τύπου

$$B_M = \frac{1}{M-1}\sum_{i=1}^{M}\left(\overline{Q}_I^{(i)} - \overline{Q}_{IM}\right)^2,$$

ενώ τέλος η ολική διακύμανση της εκτιμήτριας δίνεται από τον τύπο

$$T_M = \overline{U}_M + \left(1+\frac{1}{M}\right)B_M .$$

Αποδεικνύεταο ότι η τυποποιημένη μεταβλητή

$$\frac{Q-\overline{Q}_{IM}}{\sqrt{T_M}} .$$

ακολουθεί την κατανομή κατανομή $t$ με $U_M$ βαθμούς ελευθερίας οι οποίοι δίνονται από τον τύπο

$$U_M = (M-1)\left(1+\frac{1}{r_M}\right)^2 \text{ ενώ } r_M = \frac{\left(1+\frac{1}{M}\right)B_M}{\overline{U}_M}.$$

ΠΑΡΑΤΗΡΗΣΕΙΣ

1. Όταν στο δείγμα δεν υπάρχουν ελλιπείς τιμές τότε $r_M = 0$ και $U_M = \infty$.
2. Ένα βασικό ερώτημα που προκύπτει για την εφαρμογή της μεθόδου είναι το πόσα δείγματα $M$ θα πρέπει να δημιουργηθούν.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Ο Rubin (1987) καθόρισε ένα φράγμα για το σκοπό αυτό, το οποίο δίνεται από τη σχέση

$$g_M = \frac{r_M + \frac{2}{(U_M + 3)}}{r_M + 1}.$$

3. Ένας δείκτης ο οποίος φανερώνει πόσο ακριβή είναι τα αποτελέσματα που προκύπτουν από την μέθοδο πολλαπλής αντικατάστασης σε σχέση με τον αριθμό των δειγμάτων που δημιούργησε ο στατιστικός αναλυτής είναι ο $RE = \left(1 + \frac{g_M}{M}\right)^{-1}$.

   Χρησιμοποιώντας σε αυτόν την τιμή του $\gamma_M$ τα αποτελέσματα που δίνει είναι ενδεικτικά για το $M$ που θα πρέπει να επιλεγεί. Όσο μεγαλύτερη είναι η τιμή του δείκτη τόσο το $M$ που θα επιλεγεί θα δώσει αντιπροσωπευτικότερη τελική τιμή που θα αντικαταστήσει την ελλιπή τιμή.

| | $\gamma_M$ | | | | |
|---|---|---|---|---|---|
| $M$ | 10% | 20% | 30% | 50% | 70% |
| 3 | 0.9677 | 0.9375 | 0.9091 | 0.8571 | 0.8108 |
| 5 | 0.9804 | 0.9615 | 0.9434 | 0.9091 | 0.8772 |
| 10 | 0.9901 | 0.9804 | 0.9709 | 0.9524 | 0.9346 |
| 20 | 0.9950 | 0.9901 | 0.9852 | 0.9756 | 0.9662 |

4. Ένα $(1-a)\%$ διάστημα εμπιστοσύνης για το $Q$ είναι το

$$\overline{Q}_{IM} \pm t_{U_M, a/2}\sqrt{T_M}$$

Για περισσότερες πληροφορίες ο ενδιαφερόμενος αναγνώστης μπορεί να απευθυνθεί στην εργασία των Haziza and Nobrega (2002).

## 3.6 Ο ΕΜ ΑΛΓΟΡΙΘΜΟΣ

Η κατασκευή αλγορίθμων επαναληπτικής προσέγγισης των εκτιμητριών μέγιστης πιθανοφάνειας αποτελεί την βασική ιδέα του ΕΜ αλγορίθμου ο οποίος χρονολογείται από από τις αρχές του αιώνα

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

(Dempster, Laird and Rubin (1977), Little and Rubin (1987) και οι McLachlan and Krishnan (1997)). Στην περίπτωση πλήρων δεδομένων θα ήταν δυνατή η χρήση κλασικών επαναληπτικών μεθόδων για την επίλυση των εξισώσεων αυτών όπως η μέθοδος Newton-Raphson.

Ο αλγόριθμος EM (***Expectation-Maximization***) βασίζεται σε μία απλή ιδέα και είχε χρησιμοποιηθεί αρκετά νωρίς στην επίλυση διαφόρων προβλημάτων. Η ad hoc αντικατάσταση των ελλιπών τιμών αποτελεί και την ιδέα αυτή.

Αρχικά δίνεται μία πρώτη τιμή σε κάθε ελλιπή τιμή. Οι μέθοδοι που θα δώσουν την πρώτη αυτή τιμή αναφέρθηκαν σε προηγούμενη παράγραφο (πχ. απλή αντικατάσταση, μέθοδος του Buck).

Εφόσον το δείγμα θεωρείται πλέον πλήρες εκτιμώνται οι άγνωστες παράμετροι της κατανομής του πλήρους δείγματος χρησιμοποιώντας κλασικές στατιστικές μεθόδους.

Θεωρώντας ότι οι εκτιμήτριες που προκύπτουν αποτελούν τις άγνωστες παραμέτρους της κατανομής του δείγματος, γίνεται νέα εκτίμηση των ελλιπών τιμών με αποτέλεσμα να δημιουργηθεί μία επαναληπτική διαδικασία η οποία συνεχίζεται έως ότου ικανοποιηθεί ένα κριτήριο σύγκλισης.

Η χρήση της ονομασίας μεγιστοποίηση (***Maximization***) οφείλεται στη χρήση των εκτιμητριών μέγιστης πιθανοφάνειας για το πλήρες δείγμα. Η ονομασία εκτίμηση (***Expectation***) οφείλεται στον τρόπο με τον οποίο εκτιμούνται οι ελλιπείς τιμές. Σε κάθε ελλιπή τιμή δίδεται μία αναμενόμενη τιμή κάτω από την υπόθεση ότι αυτές ακολουθούν την κατανομή που προκύπτει από τις παραμέτρους που εκτιμήθηκαν.

Αναλυτικά, αφού η από κοινού κατανομή του τυχαίου δείγματος $\boldsymbol{Y}$ γράφεται στη μορφή

$$f(\boldsymbol{y};\theta) = f(\boldsymbol{y}_{obs},\boldsymbol{y}_{mis};\theta) = f(\boldsymbol{y}_{obs};\theta) f(\boldsymbol{y}_{mis} \mid \boldsymbol{y}_{obs};\theta),$$

για τον πιθανολογάριθμο του δείγματος θα ισχύει

$$\lambda(\theta;\boldsymbol{Y}) = \lambda(\theta;\boldsymbol{y}_{obs}) + \lambda(\boldsymbol{y}_{mis} \mid \boldsymbol{y}_{obs};\theta),$$

από όπου παίρνουμε

$$\lambda(\theta;\boldsymbol{y}_{obs}) = \lambda(\theta;\boldsymbol{y}) - \lambda(\boldsymbol{y}_{mis} \mid \boldsymbol{y}_{obs};\theta). \qquad (3.6.1)$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Στην περίπτωση που ο μηχανισμός ελλιπών τιμών είναι *MAR* για την εκτίμηση της $q$ απαιτείται μεγιστοποίηση της (3.6.1)

## 3.7 ΜΕΘΟΔΟΣ GAUSS, ΑΛΥΣΙΔΩΤΕΣ ΙΣΟΤΗΤΕΣ, ΜΕΘΟΔΟΙ ΓΙΑ ΜΟΝΟΤΟΝΑ ΔΕΔΟΜΕΝΑ

Η **μέθοδος του Gauss** (*Conditional Gaussian*) (Schafer (1997)) εφαρμόζεται σε πίνακες δεδομένων, οι οποίοι περιλαμβάνουν τόσο μεταβλητές με συνεχή όσο και μεταβλητές με διακριτά δεδομένα όπου όμως και στις δύο κατηγορίες μεταβλητών εμφανίζονται ελλιπείς τιμές. Συνοπτικά, η μέθοδος χρησιμοποιεί ένα μοντέλο λογιστικής παλινδρόμησης (*log-linear*) για τις διακριτές τυχαίες μεταβλητές (Bishop, Fienberg and Holland (1975)) ενώ για τις συνεχείς τυχαίες μεταβλητές θεωρεί ότι ακολουθούν την πολυδιάστατη κανονική κατανομή. Βασικό μειονέκτημα της μεθόδου είναι η παρουσία πολλών παραμέτρων (Olkin and Tate (1961)).

Μία εναλλακτική μέθοδος πολλαπλής αντικατάστασης περιλαμβάνει την προσέγγιση μεταβλητή-παρά-μεταβλητή χρησιμοποιώντας **αλυσιδωτές ισότητες** (*chained equations*). Το μοντέλο αντικατάστασης των ελλιπών τιμών καθορίζεται ξεχωριστά για κάθε μεταβλητή. Στο πρώτο βήμα του αλγόριθμου δημιουργείται μία τιμή που θα αντικαταστήσει την πρώτη στη σειρά ελλιπή τιμή και στη συνέχεια η νέα αυτή τιμή θα χρησιμοποιηθεί στη διαδικασία αντικατάστασης της επόμενης (Horton and Kleinman (2007), van Buuren (1999), van Buuren (2006), Raghunathan (2001)). Η επαναληπτική διαδικασία στηρίζεται στην **μέθοδο δειγματοληψίας του Gibbs** (*Gibbs Sampling*).

Σύμφωνα με τη δειγματοληψία του Gibbs (Haziza and Nobrega (2002)) θεωρείται η τυχαία μεταβλητή

$$Z \in \mathbf{Z} = (Z_1,...,Z_p),$$

για την οποία οι δεσμευμένες συναρτήσεις πυκνότητας πιθανότητας είναι οι ακόλουθες

$$f_1,...,f_p,$$

δηλαδή



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

$$Z_i \mid z_1, z_2, \ldots, z_{i-1}, z_{i+1}, \ldots, z_p \sim f_i\left(z_i \mid z_1, z_2, \ldots, z_{i-1}, z_{i+1}, \ldots, z_p\right).$$

Έστω ότι η αρχική τιμή που έχει δοθεί είναι η

$$\boldsymbol{z}^{(1)} = \left(z_1^{(1)}, \ldots, z_p^{(1)}\right),$$

ο αλγόριθμος χρησιμοποιεί, στο δεύτερο βήμα, την τιμή αυτή για να υπολογίσει την αμέσως επόμενη ελλιπή τιμή σε κάθε τυχαία μεταβλητή

$$Z_1^{(2)} \sim f_1\left(z_1 \mid z_2^{(1)}, \ldots, z_p^{(1)}\right),$$

$$Z_2^{(2)} \sim f_2\left(z_2 \mid z_1^{(2)}, z_3^{(1)}, \ldots, z_p^{(1)}\right),$$

και τέλος

$$Z_p^{(2)} \sim f_p\left(z_p \mid z_1^{(2)}, \ldots, z_{p-1}^{(2)}\right).$$

Η διαδικασία συνεχίζεται έως ότου ικανοποιηθεί ένα κριτήριο σύγκλισης. Το βασικότερο πρόβλημα της μεθόδου είναι ότι η διαδικασία μπορεί να μην συγκλίνει σε κάποια σταθερή κατανομή.

Όταν το σύνολο των δεδομένων είναι μονότονο, μία άλλη μέθοδος που χρησιμοποιείται πιο συχνά είναι εκείνη του **predictive mean matching** (Horton and Kleiman (2007)).

## 3.8 ΜΠΕΥΖΙΑΝΕΣ ΜΕΘΟΔΟΙ

Έναν εναλλακτικό τρόπο χειρισμού προβλημάτων με ελλιπείς τιμές όταν μάλιστα το μέγεθος του δείγματος είναι σχετικά μικρό, αποτελούν οι Μπεϋζιανές προσεγγίσεις. Οι συγκεκριμένες μέθοδοι είναι αρκετά ευέλικτες και εύχρηστες.

Έστω $q$ η άγνωστη παράμετρος της κατανομής του πλήρους δείγματος, $\boldsymbol{R}$ ο δείκτης ελλιπών τιμών και τίθεται ως $\boldsymbol{X}$ ένας $nxq$ πίνακας μεταβλητών οι οποίες είναι πάντα πλήρεις (π.χ. για την $i$ γραμμή του πίνακα ισχύει $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iq})$, όπου $x_{ij}$ είναι η τιμή της μεταβλητής $X_j$ για την $i$ μονάδα ).

Αν ο μηχανισμός ελλιπών τιμών είναι αγνοήσιμος, η κατανομή της παραμέτρου $q$ είναι η ακόλουθη κατά τους Little and Rubin (2002)

$$p(\theta \mid \boldsymbol{Y}, \boldsymbol{R}, \boldsymbol{X}) = p(q \mid \boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \boldsymbol{r}, \boldsymbol{x}) = p(\theta \mid \boldsymbol{y}_{obs}, \boldsymbol{r}, \boldsymbol{x}).$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Με τη βοήθεια του θεωρήματος του Bayes προκύπτει ότι

$$p(\theta \mid \boldsymbol{y}_{obs}, \boldsymbol{r}, \boldsymbol{x}) = \frac{p(\theta, \boldsymbol{y}_{obs}, \boldsymbol{r}, \boldsymbol{x})}{p(\boldsymbol{y}_{obs}, \boldsymbol{r}, \boldsymbol{x})},$$

όμως ο δείκτης ελλιπών τιμών μπορεί να παραληφθεί καθώς η παράμετρος $q$ εξαρτάται μόνο από τις μονάδες με πλήρεις παρατηρήσεις (οι οποίες έχουν $R_{ij} = 1$). Συνεπώς

$$p(\theta \mid \boldsymbol{y}_{obs}, \boldsymbol{x}) = \frac{p(\theta, \boldsymbol{y}_{obs}, \boldsymbol{x})}{p(\boldsymbol{y}_{obs}, \boldsymbol{x})},$$

και με την βοήθεια του θεωρήματος ολικής πιθανότητας λαμβάνεται η επόμενη τελική μορφή:

$$p(\theta \mid \boldsymbol{y}_{obs}, \boldsymbol{x}) = \frac{p(\boldsymbol{y}_{obs} \mid \boldsymbol{x}, \theta) p(\boldsymbol{x}, \theta)}{p(\boldsymbol{y}_{obs}, \boldsymbol{x})} =$$

$$= \frac{p(\boldsymbol{y}_{obs} \mid \boldsymbol{x}, \theta) p(\theta \mid \boldsymbol{x}) p(\boldsymbol{x})}{p(\boldsymbol{y}_{obs} \mid \boldsymbol{x}) p(\boldsymbol{x})} =$$

$$= \frac{p(\boldsymbol{y}_{obs} \mid \boldsymbol{x}, \theta) p(\theta \mid \boldsymbol{x})}{p(\boldsymbol{y}_{obs} \mid \boldsymbol{x})}.$$

Εκείνο το οποίο ενδιαφέρει είναι το γινόμενο $p(\boldsymbol{y}_{obs} \mid \boldsymbol{x}, \theta) p(\theta \mid \boldsymbol{x})$ στο οποίο η $p(\theta \mid \boldsymbol{x})$ αποδίδει τη συνεισφορά της μπεϋζιανής μεθόδου στην απλούστευση της εκτίμησης της παραμέτρου $\theta$, ενώ η $p(\boldsymbol{y}_{obs} \mid \boldsymbol{x}, \theta)$ είναι η συνάρτηση πυκνότητας πιθανότητας των πλήρων παρατηρήσεων του δείγματος. Αντιθέτως η αρχική συνάρτηση $p(\theta \mid \boldsymbol{y}_{obs}, \boldsymbol{r}, \boldsymbol{x})$ σπάνια έχει αναλυτική μορφή που να είναι απλή για τα μη πλήρη δεδομένα. Αρκετά συχνά χρησιμοποιούνται μέθοδοι προσομοίωσης για να παράγουν τιμές για την παράμετρο $q$. Μία από τις μεθόδους αυτές είναι η αύξηση των δεδομένων (Tanner and Wong (1987)) μίας επαναληπτικής διαδικασίας που εφαρμόζεται στην $p(\theta \mid \boldsymbol{y}_{obs}, \boldsymbol{r}, \boldsymbol{x})$ για την εκτίμηση του $q$. Η συγκεκριμένη τεχνική συνδυάζει χαρακτηριστικά του ΕΜ αλγόριθμου και της πολλαπλής αντικατάστασης με $M$ αντικαταστάσεις για κάθε μονάδα με ελλιπή τιμή σε κάθε επανάληψη. Μπορεί να θεωρηθεί ως μία τεχνική τελειοποίησης του ΕΜ αλγορίθμου χρησιμοποιώντας ταυτόχρονα και διαδικασίες προσομοίωσης, με το βήμα της αντικατάστασης να είναι αντίστοιχο με το βήμα $E$ του ΕΜ.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Υπάρχουν πολλές τεχνικές παρακολούθησης και ελέγχου σύγκλισης αλγορίθμων που χρησιμοποιούν επαναληπτική μπεϋζιανή προσομοίωση για την πολλαπλή αντικατάσταση ορθογώνιων σύνολο δεδομένων (Gelman (1995)). Αφορούν κυρίως, όμως, αγνοήσιμους μηχανισμούς ελλιπών τιμών.

## 3.9 ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΜΕ ΧΡΗΣΗ ΣΤΑΤΙΣΤΙΚΩΝ ΠΑΚΕΤΩΝ

Η χρήση στατιστικών πακέτων για την ανάλυση δεδομένων με ελλιπείς τιμές είναι αρκετά διαδεδομένη. Στη συνέχεια θα γίνει μία παρουσίαση του τρόπου που χειρίζονται τα ελλιπή δεδομένα ορισμένα από τα πιό δημοφιλή στατιστικά πακέτα.

**Στατιστικό πακέτο SPSS**

Στο SPSS, όταν ο μηχανισμός ελλιπών τιμών είναι *MCAR*, γίνεται διαχωρισμός των μονάδων του δείγματος σε εκείνες που εμφανίζουν ελλιπείς τιμές και σε εκείνες που δεν εμφανίζουν ελλιπείς τιμές. Στη συνέχεια επιλέγοντας *t*-test για τους μέσους σε μεταβλητές όπως εισόδημα, ηλικία και φύλο ή άλλες κύριες μεταβλητές επιβεβαιώνεται ή όχι ο μηχανισμός *MCAR*. Ένας εναλλακτικός τρόπος είναι η επιλογή *Missing Values Analysis* (MVA), η οποία περιλαμβάνει το *Little's MCAR test*. Το τελευταίο είναι ένα $x^2$-τεστ για ελλιπείς τιμές που οφείλονται εντελώς στην τύχη. Όταν το *p*-value του ελέγχου είναι μεγαλύτερο από το επίπεδο σημαντικότητας που έχει αποφασίσει να χρησιμοποιήσει ο αναλυτής τότε γίνεται αποδεκτή η υπόθεση της ύπαρξης του μηχανισμού *MCAR*. Μία ακόμη επιλογή μέσα από το SPSS είναι η ακόλουθη. Απο το Analyse, Missing Value Analysis γίνεται επιλογή του αλγόριθμου *EM*. Το τεστ του *Little* θα εκτυπώσει τον πίνακα μέσων *EM* ( *EM Means*), τον πίνακα συνδιακυμάνσεων *EM* ( *EM Covariances*) και τον πίνακα συσχετίσεων *EM* ( *EM Correlations*).

Αν ο μηχανισμός ελλιπών τιμών επιβεβαιωθεί ότι είναι MCAR τότε ο στατιστικός μπορεί να επιλέξει την ανάλυση με χρήση μόνο των



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

πλήρων παρατηρήσεων (*CC*) για να ολοκληρώσει την επεξεργασία των δεδομένων του..

Όταν είναι επιθυμητό να ελεγχθεί αν ο μηχανισμός ελλιπών τιμών είναι ***MAR*** χρησιμοποιείται και πάλι η επιλογή MVA. Το πρόγραμμα θα παράγει έναν πίνακα με την ονομασία ΄Separate Variance t test'. Κάθε γραμμή του πίνακα αυτού περιέχει μεταβλητές οι οποίες έχουν 1% ή παραπάνω ελλιπείς τιμές, ενώ κάθε στήλη του πίνακα αποτελείται από μεταβλητές. Σε κάθε κελί (*cell*), αν η τιμή του *p*-value του αμφίπλευρου ελέγχου είναι ≤0.05, συμπεραίνουμε ότι οι μεταβλητές με τις ελλιπείς τιμές (που βρίσκονται στις γραμμές) σχετίζονται με τις αντίστοιχες μεταβλητές (που είναι πλήρεις) και βρίσκονται στις στήλες, οπότε ο μηχανισμός ελλιπών τιμών είναι ***MAR***.

Το SPSS μέσα απο το MVA προσφέρει τρεις επιλογές εκτίμησης: την μέθοδο CC, την επιλογή ενός μοντέλου παλινδρόμησης για την εκτίμηση ελλιπών τιμών και τον EM αλγόριθμο. Το ***Transform, Replace Missing Cases*** (***RMC***) παρέχει πέντε μεθόδους εκτίμησης: αντικατάσταση ελλιπών τιμών με τη χρήση του μέσου, αντικατάσταση με τη χρήση μέσου ή διαμέσου από τις πλησιέστερες μονάδες (δηλαδή μονάδες με παραπλήσια χαρακτηριστικά), γραμμική παρεμβολή (***linear interpolation***) μεταξύ προηγούμενων και επόμενων πλήρων μονάδων από εκείνη που εμφανίζει ελλιπείς τιμές ή τέλος παρεμβολή μεταξύ της αμέσως προηγούμενης και της αμέσως επόμενης πλήρους μονάδας από εκείνη που εμφανίζει ελλιπείς τιμές. Προσφέρει επίσης και την επιλογή ενός μοντέλου παλινδρόμησης για την αντικατάσταση των ελλιπών τιμών με εκείνες που θα εκτιμηθούν.

Το SPSS προσφέρει τη δυνατότητα μέσα από το ***Missing Values Analysis*** τον καθορισμό της φύσης (***pattern***) των ελλιπών δεδομένων που υπάρχουν στο δείγμα. Οι ελλιπείς μονάδες του δείγματος μπορεί να μην παρουσιάζουν κάποια τιμή για ένα χαρακτηριστικό, μπορούν επίσης να δηλωθούν με κάποια πολύ υψηλή τιμή (***extremely high value***) σε σχέση με τις πλήρεις μονάδες του δείγματος ή να δηλωθούν με κάποια πολύ χαμηλή τιμή (***extremely low value***). Συνεπώς, δημιουργούνται τρείς πίνακες που απεικονίζουν τη συχνότητα κάθε είδους ελλιπούς τιμής που εμφανίζεται στο δείγμα (είναι γνωστό ότι

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

στο πακέτο SPSS μία ελλιπής τιμή μπορεί να δηλωθεί είτε με μία πολύ υψηλή τιμή είτε με μία πολύ χαμηλή τιμή). Οι πίνακες αυτοί είναι

- Ένας πίνακας συνάφειας (***crosstabulation***). Για κάθε κατηγορική μεταβλητή, δημιουργείται ένας πίνακας που παρουσιάζει για κάθε κατηγορία της κατηγορικής μεταβλητής τη συχνότητα και το ποσοστό των πλήρων μονάδων για τις υπόλοιπες μεταβλητές
- Ένας πίνακας ποσοστού αποτυχίας (***percent mismatch***). Για κάθε ζευγάρι μεταβλητών, απεικονίζεται το ποσοστό των περιπτώσεων όπου η μία μεταβλητή έχει ελλιπή τιμή και η άλλη όχι. Τα διαγώνια στοιχεία του πίνακα αποτελούν το ποσοστό των ελλιπών τιμών για κάθε μεταβλητή.
- Ένας πίνακας με *t*-τεστ. Σε αυτόν δίνεται μία σειρά από *t*-τεστ για να ληφθεί απόφαση σε διάφορους στατιστικούς ελέγχους.

Στο SPSS, η έλλειψη της απάντησης αντιμετωπίζεται με τον ορισμό τιμής/τιμών για τη μεταβλητή που θα δηλώνει ακριβώς αυτό: τη μη-ύπαρξη απάντησης. Προφανώς οι τιμές αυτές πρέπει να είναι εκτός των ορίων των καταγεγραμμένων (ή απλώς πιθανών) τιμών της συγκεκριμένης μεταβλητής. Το SPSS υλοποιεί χωριστά το κάθε τεστ που ζητήθηκε χρησιμοποιώντας όλες τις περιπτώσεις που είναι έγκυρες για τις μεταβλητές που συμμετέχουν σε αυτό (***exclude cases test by test***), δηλαδή χρησιμοποιεί την μέθοδο *AC*. Είναι δυνατή η επιλογή στο πρόγραμμα της χρησιμοποίησης σε όλα τα τεστ μόνο των περιπτώσεων που είναι ταυτόχρονα έγκυρες για όλες τις μεταβλητές του καταλόγου( ***test variable list, exclude cases listwise***). Μία τέτοια επιλογή οδηγεί στη χρήση της μεθόδου *CC*. Τέλος, σε ορισμένες διαδικασίες, το SPSS μπορεί να χρησιμοποιήσει όλες τις καταγραμμένες περιπτώσεις αντικαθιστώντας τις ελλείπουσες τιμές, με τη μέση τιμή της αντίστοιχης μεταβλητής (***replace with mean***).

Στο SPSS ο χρήστης μπορεί να καθορίσει την κατανομή των δεδομένων την οποία θα χρησιμοποιήσει ο αλγόριθμος *EM*. Για το σύνολο αυτό δίνονται διάφορες επιλογές, όπως κανονική κατανομή, κατανομή Student *t*.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

**Ρουτίνα AmeliaII / Στατιστικό πακέτο *R***

Το AmeliaII που κατασκεύασαν οι Hanker, King and Blackwell (2006) χρησιμοποιεί τη μέθοδο bootstrap βασισμένη στον αλγόριθμο *EM*. Η μέθοδος αυτή είναι αρκετά γρήγορη. Το πακέτο επιτρέπει στον χρήστη να εισάγει αρχικές τιμές για κάθε ελλιπή τιμή. Το άρθρο των Honaker and King(2006) προσφέρει επιπλέον πληροφορίες για το συγκεκριμένο πακέτο.

Το AmeliaII πραγματοποιεί μόνο το βήμα της αντικατάστασης. Επιπλέον ανάλυση και συνδυασμός αποτελεσμάτων μπορούν να ληφθούν από το πακέτο *R* χρησιμοποιώντας το λογισμικό Zelig των Imai, King και Lau (2006) ή από ξεχωριστά στατιστικά πακέτα (πχ SAS ή STATA).

**Ρουτίνα Hmisc / Στατιστικό πακέτο *R* και *S*-Plus**

Η ρουτίνα Hmisc που πρότεινε ο Harrell (2006) συνδυάζεται με τα πακέτα *R* και *S*-Plus και μπορεί να εφαρμοστεί και σε δείγματα με μονότονα δεδομένα. Χρησιμοποιεί τη μέθοδο predictive mean matching (για μονότονα δεδομένα) ή τη μέθοδο βαρών. Η μέθοδος του predictive mean matching λειτουργεί τόσο για κατηγορικές όσο και για συνεχείς μεταβλητές, χωρίς να είναι απαραίτητος ο υπολογισμός σφαλμάτων. Αυτή η προσέγγιση λαμβάνει υπόψη κάθε μορφή αβεβαιότητας. Υπάρχει παράλληλα και η δυνατότητα αντικατάστασης με τη χρήση μοντέλου παλινδρόμησης.

**Ρουτίνα ICE / Στατιστικό πακέτο STATA**

Η ρουτίνα ICE του Patrick Royston κάνει χρήση της μεθόδου των αλυσιδωτών ισοτήτων (*imputation using chained equation ICE*), και εκτελείται μέσω του πακέτου STATA. Παρέχει υποστήριξη για κατηγορικές μονάδες που παρουσιάζουν ελλιπείς τιμές. Για την αντικατάσταση των ελλιπών τιμών, χρησιμοποιείται είτε ένα μοντέλο λογιστικής παλινδρόμησης, είτε πολυωνυμική κατανομή. Επίσης είναι δυνατη η εφαρμογή της μεθόδου των βαρών μέσα από την εντολή *pweights*.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

**Ρουτίνα IVEWARE / Στατιστικό πακέτο SAS**

Η ρουτίνα IVEWARE χρησιμοποιείται στο πακέτο SAS. Για την αντικατάσταση των ελλιπών τιμών που παρουσιάζονται στις μονάδες του δείγματος βασίζεται στην μέθοδο αλυσιδωτών ισοτήτων αλλά και στην πολλαπλή αντικατάσταση.

**Στατιστικό πακέτο LogXact / LogXact7**

Το πακέτο LogXact χρησιμοποιεί τη μέθοδο των εκτιμητριών μέγιστης πιθανοφάνειας που πρότεινε ο Ibrahim (1990) και μπορεί να λειτουργήσει ακόμη και με δέκα κατηγορικές μεταβλητές (μέχρι πέντε επίπεδα η καθεμιά) που παρουσιάζουν ελλιπείς τιμές.

**Ρουτίνα MICE / Στατιστικό πακέτο *R* και *S*-Plus**

Η ρουτίνα MICE (***multiple imputation by cained equation***), δηλαδή πολλαπλή αντικατάσταση με τη χρήση αλυσιδωτών ισοτήτων υποστηρίζεται από τα πακέτα *R* και *S*-Plus. Παρέχει τη δυνατότητα αντικατάστασης ελλιπών τιμών με την βοήθεια πληθώρας μεθόδων όπως πολλαπλή αντικατάσταση, αλυσιδωτές ισοτήτες, μέθοδο του Buck, μοντέλα λογιστικής παλινδρόμησης και ανάλυση συνιστωσών. Επίσης δίνει το πλεονέκτημα στο χρήστη να προγραμματίσει δικές του μεθόδους αντικατάστασης με αποτέλεσμα να είναι δυνατή η αντιμετώπιση προβλημάτων στα οποία ο μηχανισμός ελλιπών τιμών είναι *MNAR*.

**Στατιστικό πακέτο SAS**

Η ανάλυση βάσει πολλαπλής αντικατάστασης από το πακέτο SAS πραγματοποιείται σε τρία βήματα. Αρχικά, η αντικατάσταση ολοκληρώνεται από τη ρουτίνα SAS PROC. Έπειτα, το πακέτο SAS χρησιμοποιεί για την ανάλυση κάθε γνωστή μέθοδο εφόσον το δείγμα είναι πλέον πλήρες. Τέλος τα αποτελέσματα συνδυάζονται χρησιμοποιώντας τη ρουτίνα PROC MIANALYZE.

Η ρουτίνα SAS PROC βασίζεται σε μία πληθώρα μεθόδων για την αντικατάσταση των ελλιπών τιμών όπως η μέθοδος Gauss και διάφορα μοντέλα παλινδρόμησης.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

**Στατιστικό πακέτο *S*-Plus**

Το στατιστικό πακέτο S-Plus περιλαμβάνει μία εκτεταμένη βιβλιοθήκη για ελλιπείς τιμές που βασίζεται στη μεθοδολογία του Schafer (1997) με τη χρήση του EM αλγόριθμου του Dempster (1977).

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

# ΚΕΦΑΛΑΙΟ ΤΕΤΑΡΤΟ

## 4.1 ΛΟΓΟΚΡΙΜΕΝΑ ΔΕΔΟΜΕΝΑ

Οι στατιστικές μέθοδοι που σχετίζονται με τα λογοκριμένα δεδομένα ήταν ανέκαθεν ένα ενδιαφέρον ερευνητικό θέμα. Εμφανίστηκαν τον $17^{ο}$ αιώνα με την ανάπτυξη των αναλογιστικών μαθηματικών στη βιομηχανία των ασφαλίσεων ζωής και των δημογραφικών μελετών. Ο αστρονόμος Edmund Halley θεωρείται μεταξύ άλλων ένας από τους θεμελιωτές της θεωρίας των λογοκριμένων δεδομένων. Στα μέσα του $20^{ου}$ αιώνα, οι μέθοδοι αυτές εξελίχθηκαν, παράλληλα με τις βιοιατρικές και κλινικές μελέτες σε μεγάλο πεδίο της βιομετρίας. Παρόμοιες (και συχνά πανομοιότυπες) στατιστικές μέθοδοι αναπτύσσονταν και σε δύο άλλα πεδία: εκείνο της θεωρίας της αξιοπιστίας , με βιομηχανικές και μηχανολογικές εφαρμογές και στην οικονομετρία. Στην τελευταία μάλιστα γίνονταν προσπάθειες να κατανοηθούν οι σχέσεις μεταξύ οικονομικών δυνάμεων σε ομάδες ανθρώπων. Τελικά, συναντούμε τα ίδια μαθηματικά προβλήματα και μερικές παρόμοιες μεθόδους επίλυσης ακόμη και στη σύγχρονη αστρονομία.

Έστω ότι ο χρόνος ζωής μιας μονάδας (ατόμου) περιγράφεται από μία μη αρνητική τυχαία μεταβλητή $T$ η οποία ακολουθεί μία συνεχή κατανομή με συνάρτηση πυκνότητας $f(t)$, αθροιστική συνάρτηση κατανομής $F(t)$ και συνάρτηση επιβίωσης $S(t)=1-F(t)$, $t\geq 0$. Θεωρείται ένα τυχαίο δείγμα $T_1,T_2,...,T_n$ από την κατανομή αυτή (δηλαδή $T_i, i=1,2,...,n$ είναι οι χρόνοι ζωής $n$ ανεξαρτήτων μονάδων που χρησιμοποιήθηκαν στη μελέτη) και συμβολίζονται με $t_1,t_2,...,t_n$ οι χρόνοι ζωής που παρατηρήθηκαν (πειραματικά/στατιστικά δεδομένα). Αυτό το οποίο έχει βασική σημασία είναι η εξαγωγή συμπερασμάτων για την άγνωστη κατανομή της τυχαίας μεταβλητής $T$ όταν στο δείγμα παρουσιάζονται και λογοκριμένα δεδομένα.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Στην περίπτωση που οι παρατηρούμενοι χρόνοι ζωής $t_1, t_2, ..., t_n$ είναι γνωστοί, ο αναλυτής μπορεί να χρησιμοποιήσει γνωστές στατιστικές μεθόδους για τη μελέτη του πληθυσμού (παραμετρικές και μη παραμετρικές). Τόσο στη θεωρία αξιοπιστίας και τον αναλογισμό όσο και στην ανάλυση επιβίωσης, ιδιαίτερο ενδιαφέρον παρουσιάζουν οι περιπτώσεις όπου όταν οι χρόνοι ζωής ορισμένων μονάδων (ατόμων) του δείγματος είναι λογοκριμένοι (*censored*). Συνεπώς ένα τέτοιο δείγμα δεν παρέχει πλήρη πληροφόρηση για τους χρόνους ζωής των ατόμων καθώς οι παρατηρήσεις μπορεί να είναι, είτε **λογοκριμένες από δεξιά** (*right censored*)στην περίπτωση που ο αναλυτής γνωρίζει ότι οι χρόνοι ζωής των μονάδων είναι μεγαλύτεροι από κάποιο χρόνο $t_0$, είτε **λογοκριμένες από αριστερά** (*left censored*), όταν γνωρίζει ότι οι χρόνοι ζωής των μονάδων είναι μικρότεροι από κάποιο χρόνο $t_0$, είτε **λογοκριμένοι σε διάστημα** (*interval censored*), στην περίπτωση που γνωρίζει ότι οι χρόνοι ζωής των μονάδων βρίσκονται εντός ενός διαστήματος $(L, R)$ με $L < R$. Στην πράξη συναντώνται κατά πλειοψηφία λογοκριμένοι χρόνοι ζωής από δεξιά.

Μερικές μελέτες απαιτούν ένα αρκετά μεγάλο διάστημα για να ολοκληρωθούν, ενώ σε άλλες το κόστος είναι υπέρογκο για τον ερευνητή. Δεν είναι δυνατή η αναμονή μέχρι να συμβεί η αποτυχία σε κάθε μονάδα του δείγματος που βρίσκεται υπό παρακολούθηση, και για τον λόγο αυτό προκαθορίζεται η χρονική διάρκειά της. Η συγκεκριμένη αυτή χρονική στιγμή λήξης της έρευνας καλείται χρόνος **λογοκρισίας** (*censoring time*). Καταγράφονται, συνεπώς, ως πραγματικοί χρόνοι ζωής μόνο οι χρόνοι των μονάδων που απέτυχαν κατά την διάρκεια της έρευνας. Αυτοί αποτελούν και τις πλήρεις παρατηρήσεις. Οι υπόλοιπες μονάδες (άτομα) που συνέχισαν να λειτουργούν έχουν χρόνους ζωής που υπερβαίνουν το $t_0$. Για αυτούς είναι γνωστό μόνο ότι ξεπερνούν την χρονική στιγμή $t_0$ (λογοκριμένοι χρόνοι ζωής), ενώ η πραγματική τους τιμή είναι άγνωστη στον ερευνητή.

Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι ακόλουθες περιπτώσεις λογοκρισίας.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

### 4.1.1 Τύποι λογοκρισίας

### α. Λογοκρισία τύπου I (Type I censoring)

Στην περίπτωση λογοκρισίας τύπου I (*type I right censoring*) ο χρόνος ζωής $T_i$ για την μονάδα $i$ καταγράφεται μόνο όταν $T_i \le t_0$. Σε οποιαδήποτε άλλη περίπτωση, ο χρόνος ζωής θεωρείται λογοκριμένος, δηλαδή μεγαλύτερος του χρόνου λογοκρισίας $t_0$. Τα δεδομένα που απορρέουν από μία τέτοια διαδικασία οδηγούν σε μία δίτιμη μεταβλητή $C_i$ (τυχαία μεταβλητή) που δηλώνει αν ο χρόνος ζωής της $i$ μονάδας είναι λογοκριμένος $(C_i = 0)$ ή πλήρης $(C_i = 1)$.

Όλες οι μονάδες τίθενται σε λειτουργία στην αρχή του πειράματος ($t$=0) και το πείραμα συνεχίζεται μέχρι μία συγκεκριμένη χρονική στιγμή $t_0$ (*censoring time*) οπότε και σταματά η εκτέλεσή του ή η συλλογή στατιστικών στοιχείων. Στην περίπτωση αυτή γνωρίζουμε μόνο

1. τους ακριβείς χρόνους ζωής των μονάδων που καταστράφηκαν πριν τη χρονική στιγμή $t_0$.
2. το πλήθος των μονάδων που καταστράφηκαν μετά τη χρονική στιγμή $t_0$.

Αν συμβολιστούν με $t_{(1)}, t_{(2)}, ..., t_{(n)}$ οι διατεταγμένοι χρόνοι ζωής των μονάδων ή ατόμων (παρατηρηθείσες τιμές), και υποτεθεί ότι $m$ ακριβώς μονάδες (άτομα)καταστράφηκαν στο χρονικό διάστημα $[0, t_0)$ προκύπτει ότι

$$t_{(1)} \le t_{(2)} \le ... \le t_{(m)} \le t_0 \le t_{(m+1)} \le ... \le t_{(n)}.$$

Έτσι τα διαθέσιμα στοιχεία κατά τον τερματισμό του πειράματος σχετίζονται με τις τυχαίες μεταβλητές (διατεταγμένοι χρόνοι ζωής)

$$T_{(1)} \le T_{(2)} \le ... \le T_{(M)}$$

και $M$ (πλήθος μονάδων που έχουν καταστραφεί μέχρι τον τερματισμό του πειράματος)

Ως παράδειγμα, ας θεωρήσουμε μία μελέτη στην οποία ένας επιδημιολόγος επιθυμεί να προσδιορίσει πώς ο χρόνος ζωής ενός

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ατόμου (εξαρτημένη μεταβλητή) σχετίζεται με τον αριθμό των τσιγάρων που καταναλώνει κάθε μέρα (ανεξάρτητη μεταβλητή). Το πείραμα έχει διάρκεια δέκα χρόνια και μέσα σε αυτά μερικά από τα άτομα που συμμετέχουν στην έρευνα πεθαίνουν ενώ άλλα όχι. Για τον χρόνο ζωής των τελευταίων είναι γνωστό ότι ξεπερνάει τη χρονική διάρκεια του πειράματος, δηλαδή το πείραμα διακόπτεται πριν ολοκληρωθεί η καταστροφή όλων των μονάδων που συμμετέχουν σε αυτό. Για το λόγο αυτό οι τιμές αυτές αποκαλούνται **δεξιά λογοκριμένα δεδομένα** (*right censored values*).

Λογοκρισία τύπου I ή δεξιά λογοκριμένα δεδομένα τύπου I συναντώνται συχνά

1. στην φαρμακευτική/βιοστατιστική (επίπεδα επιβίωσης ύστερα από χειρουργική επέμβαση ή φαρμακευτική αγωγή).
2. θεωρία αξιοπιστίας (χρόνοι αποτυχίας μονάδων που αποτελούν ένα σύστημα).
3. αναλογισμό (πίνακες επιβίωσης).

## β. Λογοκρισία τύπου II (Type II censoring)

Όλες οι μονάδες τίθενται σε λειτουργία στην αρχή του πειράματος ($t=0$) και το πείραμα συνεχίζεται μέχρι την χρονική στιγμή που καταστρέφεται η $r$ μονάδα ($r$ δεδομένος ακέραιος, $1 \le r \le n$). Τα δεδομένα που είναι διαθέσιμα στην περίπτωση αυτή είναι τα

$$t_{(1)} \le t_{(2)} \le \ldots \le t_{(r)}$$

και σχετίζονται με τις τυχαίες μεταβλητές

$$T_{(1)} \le T_{(2)} \le \ldots \le T_{(r)}$$

ενώ ο χρόνος διάρκειας του πειράματος είναι τυχαίος (ισούται με $T_{(r)}$).

Ως παράδειγμα, ας θεωρήσουμε μία εταιρεία κατασκευής κινητήρων η οποία επιθυμεί να διαπιστώσει το μέσο χρόνο βλάβης τους ως συνάρτηση της ταχύτητας λειτουργίας τους. Το πείραμα πραγματοποιείται έως ότου το 20% των κινητήρων να καταστραφεί και συνεπώς το 80% θα είναι λογοκριμένες τιμές. Ένα ακόμη παράδειγμα

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

αποτελεί ένα πείραμα με 8 λαμπτήρες, όπου ο ερευνητής έχει αποφασίσει να τερματίσει την έρευνα όταν 4 λαμπτήρες ($r=4$) αποτύχουν.

### γ. Λογοκρισία τύπου III (Type III censoring)

Αποτελεί συνδυασμό των δύο προηγούμενων περιπτώσεων. Το πείραμα τερματίζεται όταν συμβεί ένα από τα δύο κριτήρια που τέθηκαν στις λογοκρισίες τύπου I και II, δηλαδή η συλλογή στοιχείων σταματάει τη χρονική στιγμή $t_0$ ή όταν χαλάσουν $r$ συνολικά μονάδες (ότι από τα δύο συμβεί πρώτο).

### δ. Τυχαία λογοκρισία (Random Censoring or Type IV censoring)

Όλες οι μονάδες τίθενται σε λειτουργία τη χρονική στιγμή $t=0$ και σε καθεμία αντιστοιχεί ένας χρόνος λογοκρισίας που είναι τυχαία μεταβλητή. Έτσι με την $i$ μονάδα, εκτός του χρόνου $T_i$ που είναι τυχαία μεταβλητή, σχετίζεται και μία άλλη τυχαία μεταβλητή $S_i$ που περιγράφει πότε θα σταματήσει η παρακολούθηση της μονάδας (αν αυτή έχει επιβιώσει μέχρι τότε). Ο χρόνος $T_i$ καλείται δυνητικός χρόνος ζωής (*potential lifetime*) της $i$ μονάδας. Τα $(T_i, S_i)$, $i=1,2,...,n$ θεωρούνται ανεξάρτητες και ισόνομες τυχαίες μεταβλητές ενώ τα στατιστικά στοιχεία που συλλέγονται αφορούν τις τυχαίες μεταβλητές

$$Y_i = \min(T_i, S_i), i=1,2,...,n$$

και το κατά πόσον για την $i$ μονάδα έχουμε αποτυχία ή λογοκρισία. Το τελευταίο περιγράφεται από την δίτιμη τυχαία μεταβλητή

$$d_i = \begin{cases} 1, \text{ αν } T_i \text{ ή } S_i \\ \qquad\qquad i=1,2,...,n. \\ 0, \text{ αν } T_i \text{ ή } S_i \end{cases}$$

Αυτού του είδους η λογοκρισία ονομάζεται **τυχαία λογοκρισία** ή **λογοκρισία τύπου IV**.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Μερικά χρήσιμα πρακτικά συμπεράσματα στα οποία εμφανίζεται η παραπάνω μορφή λογοκρισίας είναι τα εξής:

**Παράδειγμα 1.** Έστω ότι στη θεωρία αξιοπιστίας είναι επιθυμητή η συλλογή στατιστικών στοιχείων για $n$ μονάδες τα οποία ξεκινούν να παρακολουθούνται την χρονική στιγμή $t_1 = 0$ και μέχρι την $t_2 = t_0$. Όμως στην αρχή του πειράματος δεν είναι διαθέσιμες όλες οι μονάδες αφού ο χρόνος κατασκευής τους είναι τυχαία μεταβλητή και έτσι κάθε μια τίθεται για πρώτη φορά σε λειτουργία κάποια τυχαία χρονική στιγμή.

Αν τώρα μεταφερθούν όλοι οι χρόνοι ζωής στην αρχή των χρόνων λαμβάνεται ακριβώς ότι περιγράφηκε στη λογοκρισία τύπου IV. Προφανώς ο χρόνος λογοκρισίας της $i$ μονάδας θα περιγράφεται από την τυχαία μεταβλητή $S_i = t_0 - X_i$ όπου $X_i$ είναι η καθυστέρηση (τυχαία μεταβλητή) στην κατασκευή της $i$ μονάδας σε σχέση με την έναρξη του πειράματος.

**Παράδειγμα 2.** Στην ανάλυση επιβίωσης, όταν παρακολουθούνται χρόνοι επιβίωσης ασθενών οι οποίοι πάσχουν από συγκεκριμένη ασθένεια, μπορεί να εμφανιστεί τυχαία λογοκρισία εξαιτίας αποχώρησης ατόμων πριν την ημερομηνία ολοκλήρωσης της έρευνας (λόγω αλλαγής τόπου κατοικίας ή θεράποντος ιατρού ή ακόμα και λόγω θανάτου από άλλη αιτία).

**Παράδειγμα 3.** Έστω ότι ενδιαφέρει να μελετηθεί ο χρόνος ζωής $T$ ενός εξαρτήματος $A$. Το εξάρτημα αυτό βρίσκεται συνδεδεμένο σε σειρά με ένα άλλο εξάρτημα $B$ με αποτέλεσμα, όταν το $B$ καταστραφεί να μη συνεχίζει να λειτουργεί το σύστημα, οπότε παύει η παρατήρηση του πραγματικού χρόνου ζωής του $A$. Έτσι ο χρόνος ζωής $S$ του $B$ (τυχαία μεταβλητή) λειτουργεί ως τυχαία λογοκρισία για τον χρόνο ζωής του $T$ του $A$.

**Παράδειγμα 4.** Σε κλινικές μελέτες, ενώ η χρονική στιγμή αρχής και τέλους της έρευνας είναι προκαθορισμένη, οι ασθενείς εισέρχονται σε

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

αυτή σε διαφορετικές (τυχαίες) χρονικές στιγμές (πχ. τη στιγμή που θα γίνει διάγνωση της ασθένειας) με αποτέλεσμα, οι χρόνοι λογοκρισίας να είναι τυχαίοι. Συνεπώς, σε κάθε άτομο αντιστοιχεί μία τυχαία μεταβλητή $T$ που δηλώνει το χρόνο ζωής του ατόμου (από τη στιγμή που θα εισέλθει στην έρευνα)και μία τυχαία μεταβλητή $S$ που δηλώνει το χρόνο λογοκρισίας του ατόμου.

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

Σε περίπτωση **αριστερής λογοκρισίας** (*left censoring*) ο πλήρης χρόνος ζωής $T_i$ της μονάδας $i$ παρατηρείται αν και μόνο αν $T_i \geq t_0$, ενώ στην περίπτωση που $T_i < t_0$, το μόνο που γνωρίζει ο αναλυτής είναι ότι ο χρόνος ζωής της μονάδας είναι μικρότερος του χρόνου λογοκρισίας $t_0$. Δίνονται στη συνέχεια δύο παραδείγματα που αφορούν τον παραπάνω τύπο λογοκρισίας.

**Παράδειγμα 1**. Μία ομάδα φοιτητών αναλαμβάνουν να ολοκληρώσουν τη διπλωματική τους μέσα σε ένα εξάμηνο. Εάν κάποιοι από αυτούς την ξεκινήσουν και την τελειώσουν σε διάστημα πριν από το εξάμηνο που τους έχει ανατεθεί, τότε δεν είναι γνωστό πότε αυτό συνέβη. Τα δεδομένα που θα προκύψουν θα είναι λογοκριμένα από τα αριστερά. Στην περίπτωση που οι μαθητές την ολοκληρώσουν μέσα στο διάστημα του εξαμήνου τότε προκύπτουν πλήρη δεδομένα. Επειδή όμως υπάρχει προκαθορισμένη ημερομηνία λήξης (τέλος του εξαμήνου) μπορεί να έχουμε ταυτόχρονα λογοκριμένα δεδομένα και από αριστερά και από δεξιά, οπότε και γίνεται λόγος για **διπλή λογοκρισία** (*doubly censored*).

**Παράδειγμα 2.** Ένας αστρονόμος παρατηρεί ένα δείγμα από γειτονικούς γαλαξίες ώστε να καθορίσει τη σχέση ανάμεσα σε σκόνη



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

και μοριακά αέρια. Στους μισούς εντοπίζεται ερυθρή φωτοβολία αλλά μόνο λίγοι από αυτούς παρουσιάζουν φωτοβολία μέσα στο ορατό από το ανθρώπινο μάτι φάσμα της ακτινοβολίας (δηλαδή τιμή ανώτερη του κάτω ορίου της ανθρώπινης όρασης). Οι υπόλοιποι με ερυθρή φωτοβολία δίνουν τιμή κατώτερη του κάτω ορίου και συνεπώς μη ορατή από τον άνθρωπο (λογοκριμένο δεδομένο).

Ένας διαφορετικός μηχανισμός παραγωγής λογοκριμένων δεδομένων είναι η **λογοκρισία σε διάστημα** (*interval censoring*). Απαντάται όταν ένας ασθενής εξετάζεται σε τακτά χρονικά διαστήματα και κάποια στιγμή εντοπίζεται υποτροπή. Σε αυτή την περίπτωση η μοναδική πληροφορία που υπάρχει είναι ότι η υποτροπή εμφανίζεται μεταξύ δύο διαφορετικών εξετάσεων.

Σημαντική είναι και η αναφορά στα **περικομμένα δεδομένα** (*truncated data*). Η περικοπή ορίζεται ως μία συνθήκη την οποία πρέπει να ικανοποιούν τα άτομα που θα τεθούν υπό παρακολούθηση. Όσα από τα άτομα δεν ικανοποιούν τη συνθήκη δεν τίθενται υπό παρακολούθηση και απλά αγνοούνται.

Είναι άξιο προσοχής ότι, όπως έγινε φανερό από τα παραδείγματα που δόθηκαν προηγουμένως, τα λογοκριμένα δεδομένα δεν οφείλονται αποκλειστικά και μόνο στο χρόνο λήξης της έρευνας αλλά και στη διαφυγή ατόμων, των οποίων που η παρακολούθηση σταμάτησε για οποιουσδήποτε λόγους πριν το χρόνο λήξης της έρευνας. Για παράδειγμα, σε ιατρικές έρευνες ο υπό παρακολούθηση ασθενής μπορεί να χαθεί ή να αποχωρήσει πριν τη λήξη της έρευνας. Δεν είναι λίγες οι περιπτώσεις που ο ασθενής μπορεί να πεθάνει κατά τη διάρκεια της παρακολούθησης, ή να σταματήσει την θεραπευτική αγωγή λόγω σοβαρών παρενεργειών. Όσον αφορά την επιστήμη του αναλογισμού, περιπτώσεις λογοκρισίας εμφανίζονται στις περιπτώσεις ασφάλισης όπου έχουν τεθεί όρια ή αφαιρετέα ποσά.

Για να διασαφηνιστούν οι παραπάνω έννοιες, ας υποτεθεί ότι 7 ασθενείς με γρίπη εισέρχονται σε κλινική μελέτη που διαρκεί 1 εβδομάδα. Υποτίθεται ότι και οι 7 ασθενείς ανταποκρίνονται στη θεραπεία που τους χορηγείται με αποτέλεσμα την υποχώρηση των συμπτωμάτων μετά από κάποιο χρονικό διάστημα. Μόλις συμβεί αυτό,



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

η θεραπευτική αγωγή σταματά και ενδιαφέρει πλέον ο χρόνος επανεμφάνισης της νόσου. Προφανώς λόγω της τυχαιότητας του χρόνου ανταπόκρισης στη θεραπευτική αγωγή, ο χρόνος λογοκρισίας κάθε ατόμου είναι τυχαίος.

Οι ασφαλιστικές εταιρείες παρέχουν τη δυνατότητα σε ασφαλισμένους να κρατήσουν ένα ποσό του κινδύνου, συνήθως της επιλογής τους, και να ασφαλιστούν για περιπτώσεις κινδύνου που ξεπερνά το ποσό αυτό. Η πολιτική που ακολουθεί η εταιρεία έχει ως σκοπό να αποφύγει μικροζημιές και μικροαποζημιώσεις, οι οποίες συχνά είναι πολλές. Από την άλλη πλευρά ο ίδιος ο ασφαλισμένος αναλαμβάνει προσωπικά ένα μικρό ποσοστό του κινδύνου με αποτέλεσμα να είναι πιο προσεκτικός, να λαμβάνει μέτρα πρόληψης και αποφυγής του κινδύνου, ενώ παράλληλα καταβάλλει και λιγότερα ασφάλιστρα.

Για παράδειγμα, ένας ασφαλισμένος ασφαλίζεται για το ποσό των 30000 ευρώ με αφαιρετέο ποσό $d=2000$ ευρώ για πιθανή ατομική ζημιά $X$. Πρακτικά, αν σε κάποια πραγματοποίηση του κινδύνου προκληθεί ζημιά $X \le d$, θα πρέπει ο ασφαλισμένος να πληρώσει από μόνος του ενώ η εταιρία δεν καταβάλλει κανένα ποσό ως αποζημίωση. Αν η πραγματοποίηση του κινδύνου επιφέρει ζημιά $X > d$=2000 (πχ έστω $X$=20000), ο ασφαλισμένος θα πληρώσει για την αποκατάσταση της ζημιάς 2000 ευρώ και η εταιρία το υπόλοιπο (δηλ. $20000-2000=18000$ ευρώ). Το αφαιρετέο αυτό ποσό είναι μία μορφή περικομμένων δεδομένων από αριστερά, διότι οποιαδήποτε πραγματοποίηση κινδύνου έχει ως αποτέλεσμα ύψος ζημιάς $X \le d$, απλά δεν ενδιαφέρει την εταιρεία. (Η τυχαία μεταβλητή $Z$ που περιγράφει την αποζημίωση που πληρώνει η εταιρία, μπορεί να εκφραστεί μέσω του τύπου):

$$Z = \begin{cases} 0\,, \text{ αν } X \le d \\ \\ X-d\,, \text{ αν } X > d\,, \end{cases}$$

ή ισοδύναμα

$$Z = (X-d)_+ = \max(0, X-d)\,.$$



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Μία μορφή λογοκρισίας σε ασφαλιστικές καλύψεις είναι και το όριο ιδίας κράτησης, το οποίο συμβολίζεται με $U$. Οποιοσδήποτε κίνδυνος επιφέρει ζημιά μέχρι εκείνο το ποσό, η ασφαλιστική εταιρεία τον αποζημιώνει. Ζημιά μεγαλύτερη του ποσού αυτού αποζημιώνεται μόνο μέχρι το όριο.

Η αποζημίωση $Z$ που καταβάλλει η εταιρία, μπορεί να περιγραφεί τώρα:

$$Z = \begin{cases} X, \text{ αν } X \le U \\ U, \text{ αν } X > U, \end{cases}$$

ή ισοδύναμα

$$Z = \min(X, U).$$

Το ποσό ιδίας κράτησης είναι μία μορφή λογοκρισίας από δεξιά.

Οι ασφαλιστικές εταιρείες, για να χειριστούν κινδύνους που πιθανότατα να αποφέρουν μεγάλες ζημιές, κρατούν ένα μέρος του κινδύνου και το υπόλοιπο μέρος του κινδύνου το μεταφέρουν σε μία άλλη ασφαλιστική εταιρεία. Η πρώτη εταιρεία ονομάζεται πρωτασφαλίστρια ενώ η δεύτερη αντασφαλίστρια.

Για την πρωτασφαλίστρια εταιρία ισχύει

$$Z = \begin{cases} X, \text{ αν } X \le U \\ U, \text{ αν } X > U \end{cases}$$

ενώ για την αντασφαλίστρια

$$Y = \begin{cases} 0, \text{ αν } X \le U \\ X - U, \text{ αν } X > U. \end{cases} = \max(0, X - U)$$

Τέλος αξίζει να σημειωθεί ότι, στην πλειοψηφία των ασφαλιστηρίων συμβολαίων περιλαμβάνονται όρια και αφαιρεταία ποσά ταυτόχρονα σε ένα συμβόλαιο.

Αν συμβεί αυτό, τότε η αποζημίωση για την πρωτασφαλίστρια εταιρεία μπορεί να εκφραστεί ως εξής:

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

$$Z=\begin{cases}0\,,\ \text{αν}\ \ X\le d\\ X-d\,,\ \text{αν}\ \ d<X<U\\ U-d\,,\ \text{αν}\ \ X\le U\,.\end{cases}$$

**Παράδειγμα 1.** Σε ένα χαρτοφυλάκιο $N$=8 συμβολαίων παρατηρούνται οι εξής κίνδυνοι:

403, 1490, 1948, 443, 1866, 1704, 1221, 823

και ορίζεται το $d=1600$.

Τα ποσά που θα καταβάλλει ο ασφαλιστής είναι 1948-1600=348, 1866-1600=266, 1704-1600=104.

Οι υπόλοιπες ζημιές δεν αποζημιώνονται διότι είναι μικρότερες από το αφαιρεταίο ποσό. Το άθροισμα των τριών αυτών ζημιών θα είναι $S_R$.

Η παρατηρηθείσα διαδικασία αποζημιώσεων θα είναι:

$$S_R=0+0+348+0+266+106+0+0$$

Η συνάρτηση κατανομής των ατομικών ζημιών δίνεται από τον τύπο

$$F_W(w)=P(W\le w)=P(X-d\le w\,|\,X>d)=P(X\le w+d\,|\,X>d)=$$
$$=\frac{P(d\le X\le w+d)}{P(X>d)}=\frac{F_X(w+d)-F_X(d)}{1-F_X(d)}$$

και η συνάρτηση πυκνότητας πιθανότητας παίρνει τη μορφή

$$f_W(w)=F_W'(w)=\frac{f_x(w+d)}{1-F_X(d)}\,.$$

**Παράδειγμα 2.** Έστω ότι η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής $X$ του μεγέθους της ατομικής ζημιάς είναι της μορφής

$$f_X(x)=0.02\left(1-\frac{x}{100}\right),\ \ 0\ \text{˝ά}x\ \text{˝ά}100$$

Ας υποθέσουμε ακόμη ότι το ποσό που καταβάλλεται ως αποζημίωση για κάθε ζημιά είναι το 80% του ποσού της ζημιάς που υπερβαίνει ένα ποσό ίσο με 15 νομισματικές μονάδες και ότι ο αριθμός των ζημιών είναι τυχαία μεταβλητή που ακολουθεί την κατανομή Poisson με παράμετρο $l=4$. Ζητείται το αναμενόμενο συνολικό κόστος που αντιστοιχεί στον ασφαλισμένο και το αναμενόμενο συνολικό κόστος που αντιστοιχεί στον ασφαλιστή.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Η τυχαία μεταβλητή $Y$ που θα δίνεται από τον παρακάτω τύπο εκφράζει το μέγεθος της αποζημίωσης από

$$Y = \begin{cases} 0, & X \le 15 \\ (X-15)0.8, & X > 15, \end{cases}$$

ενώ η τυχαία μεταβλητή $S_I$ αντιστοιχεί στην συνολική αποζημίωση που καταβάλλεται στους ασφαλισμένους της εταιρίας μπορεί να γραφεί στη μορφή:

$$S_I = Y_1 + Y_2 + \ldots + Y_N, \; N \sim P(\lambda), \; l = 4.$$

Έστω $Z$ η τυχαία μεταβλητή που περιγράφει να είναι το ατομικό κόστος που καλύπτει ο ασφαλισμένος κατά την εμφάνιση μιας ζημιάς. Τότε η τυχαία μεταβλητή $S_A$ που δίνει το συνολικό κόστος ασφάλισης θα δίνεται από το άθροισμα

$$S_A = Z_1 + Z_2 + \ldots + Z_N \;,$$

όπου

$$Z_i = \begin{cases} X_i, & X_i \text{ ´ά} 15 \\ 0.2(X_i \quad 15), & X_i > 15. \end{cases}$$

Τελικά, το μέγεθος της συνολικής ζημιάς θα έχει την μορφή

$$S = X_1 + X_2 + \ldots + X_N, \; S = S_I + S_A,$$

με μέση τιμή

$$E(S) = E(S_I) + E(S_A).$$

Επίσης θα ισχύει

$$E(S_A) = E(N)E(X) - E(N)E(Y),$$

$$E(S_A) = 4(E(X) - E(Y)),$$

και αφού

$$E(X) = \int_0^{100} x f_X(x)dx = \int_0^{100} 0.02x\left(1 - \frac{x}{100}\right)dx = 33.3,$$

$$E(Y) = \int_{15}^{100} 0.8(x-15) f_X(x)dx = \int_{15}^{100} 0.8(x-15)0.02\left(1 - \frac{x}{100}\right)dx = 24.03,$$

βρίσκουμε τελικά

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

$$E(S_A) = 4(33.3 - 24.03) = 37.08 \text{ νομισματικές μονάδες.}$$

**Παράδειγμα 3.** Εάν μία ασφαλιστική εταιρία έχει θέσει ένα όριο $U$ στην ανάληψη κινδύνου ενώ υπάρχει και ένα αφαιρεταίο ποσό $d$, τότε η μέση αποζημίωση που θα καταβάλλει θα δίνεται από τον τύπο:
Έστω

$$E(S) = \int_d^U (x-d) f_X(x)dx + U \cdot P(X > U) = \int_d^U (x-d) f_X(x)dx + U[1 - F_X(U)].$$

Η τυχαία μεταβλητή $S$ δηλώνει την συνολική αποζημίωση που καταβάλλεται στον ασφαλισμένο.
Αυτή είναι μία χαρακτηριστική περίπτωση δείγματος με λογοκριμένα δεδομένα από δεξιά και περικομμένα από αριστερά.

## 4.2 ΣΥΝΑΡΤΗΣΗ ΕΠΙΒΙΩΣΗΣ ΚΑΙ ΒΑΘΜΙΔΑ ΑΠΟΤΥΧΙΑΣ

Γίνεται χρήση δύο διαφορετικών τρόπων για να προσδιοριστεί μια κατανομή, είτε μέσω της συνάρτησης κατανομής είτε μέσω της συνάρτησης πυκνότητας πιθανότητας. Στους χρόνους επιβίωσης, προτιμάται η **συνάρτηση επιβίωσης** η οποία σχετίζεται με εκείνη της κατανομής μέσα από τη σχέση

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

Για τους χρόνους επιβίωσης ισχύει πάντα $t>0$. Η συνάρτηση αυτή παριστάνει την πιθανότητα ένα άτομο να επιβιώσει τουλάχιστον $t$ χρόνια (για δεδομένο $t$). Μπορεί επίσης να χρησιμοποιηθεί και η σχέση

$$S_{T_x}(t) = P(T_x > t)$$

η οποία δηλώνει την πιθανότητα ένα άτομο ηλικίας $x$ να ζήσει τουλάχιστον t χρόνια.

Μία χρήσιμη συνάρτηση, διαδεδομένη στην ανάλυση επιβίωσης, είναι η βαθμίδα αποτυχίας (Hosmer&Lemeshow (1999))

$$l(t) = \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{S(t)},$$

στην οποία το $T$ είναι μία μη αρνητική συνεχής τυχαία μεταβλητή και η συνάρτηση εκφράζει την πιθανότητα ένα άτομο να πεθάνει σύντομα, εφόσον έχει ζήσει $t$ έτη.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Παράδειγμα: Στην εκθετική κατανομή η βαθμίδα αποτυχίας είναι ίση με

$$l(t) = \frac{f(t)}{S(t)} = \frac{l e^{-lt}}{e^{-lt}} = l\ ,\ t > 0\ .$$

Το αποτέλεσμα είναι αναμενόμενο (Sheldon (1994)), καθώς είναι γνωστό ότι η εκθετική κατανομή έχει την **αμνήμονα ιδιότητα** (*memoryless*), δηλαδή η πιθανότητα να πεθάνει ένα άτομο ή να καταστραφεί μία μονάδα που βρίσκεται σε λειτουργία δεν εξαρτάται από τον χρόνο. Με την βοήθεια την βαθμίδας αποτυχίας είναι δυνατός ο υπολογισμός της συνάρτησης επιβίωσης διότι συνδέονται ακολούθως:

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

Το εκθετικό μοντέλο (1) έχει βαθμίδα αποτυχίας:

$$l_i = \frac{1}{\exp(b_0 + b_1 x_i)},\ i = 1,2,...n\ ,$$

και το μοντέλο της Weibull

$$l_i(t) = \frac{c\frac{1}{b}(\frac{t}{b})^c \exp-(\frac{t}{b})^c}{\exp-\left(\frac{t}{b}\right)^c} = ct^{c-1}(\frac{1}{b})^c = (\frac{1}{\exp(b_0 + b_1 x_i)})^{\frac{1}{s}}\frac{1}{s}t^{\frac{1}{s}}\ .$$

## 4.3 ΕΚΤΙΜΗΤΡΙΕΣ ΜΕΓΙΣΤΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ

Η μέθοδος των εκτιμητών μέγιστης πιθανοφάνειας χρησιμοποιείται για την εύρεση των βέλτιστων εκτιμητών των παραμέτρων του μοντέλου που έχει προσαρμοστεί στα δεδομένα (Morgenthaler (1997)). Οι παράμετροι που πρέπει να εκτιμηθούν στο μοντέλο (1) είναι οι $b = (b_0, b_1) \in R^2$. Αρχικά είναι αναγκαίος ο υπολογισμός του πιθανολογάριθμου:

$$f_T(t) = \frac{1}{\exp(b_0 + b_1 x_i)}\exp(\frac{-t}{\exp(b_0 + b_1 x_i)}),\ t > 0, i = 1,2,...,n\ ,$$

έτσι

$$l(b_0, b_1) = \log(\prod_{i=1}^{n} f_T(t)) = \sum_{i=1}^{n}\left[-(b_0 + b_1 x_i) - \frac{t_i}{\exp(b_0 + b_1 x_i)}\right].$$

**Ειδική περίπτωση**: $b_1 = 0$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Εάν υποτεθεί ότι η παράμετρος $b_1 = 0$, τότε η εκτίμηση του $b_0$ είναι αρκετά απλή:

$$\frac{\partial l}{\partial b_0}(\beta_0) = 0 \Leftrightarrow \sum_{i=1}^{n} \frac{t_i}{\exp(b_0)} = n \Leftrightarrow \hat{b}_0 = \log(\bar{t}).$$

**Γενική περίπτωση**: $b_1 \neq 0$

Για να μεγιστοποιηθεί η συνάρτηση του πιθανολογάριθμου θα τεθούν οι μερικές παράγωγοι ίσες με το μηδέν

$$\left\{ \begin{array}{l} \frac{\partial l}{\partial b_0}(b) = 0 \\ \frac{\partial l}{\partial b_1}(b) = 0 \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{l} -n + \sum_{i=1}^{n} \frac{t_i}{\exp(b_0 + b_1 x_i)} = 0 \\ \sum_{i=1}^{n} \frac{x_i t_i}{\exp(b_0 + b_1 x_i)} - \sum_{i=1}^{n} x_i = 0 \end{array} \right\}.$$

Σημειώνεται ότι δεν υφίσταται κάποια συγκεκριμένη κλειστή μορφή επίλυσης του παραπάνω συστήματος και για το λόγο αυτό προτιμάται η μέθοδος Newton.

## 4.4 Η ΜΕΘΟΔΟΣ NEWTON

Η μέθοδος Newton (Quarteroni (2000))λειτουργεί με την αρχή ότι είναι εύκολη η μεγιστοποίηση μιας τετραγωνικής προσέγγισης της συνάρτησης $b = (b_0, b_1) \in R^2$. Τίθεται ως $l_Q(b)$ η τετραγωνική προσέγγιση του $l(b)$ για το $b_0$, όπου το $b_0$ είναι ένα διάνυσμα $1xq$.

$\nabla l(b^0)_j = \frac{\partial l}{\partial b_j}(b^0)$ ένα διάνυσμα 1xp και $H(b^0)_{ij} = \frac{\partial^2 l}{\partial b_i b_j}(b^0)$ πίνακας $pxp$

$$l_Q(b) = l(b^0) + \nabla l(b^0)(b - b^0) + \frac{1}{2}(b - b^0)H(b^0)(b - b^0).$$

Το μέγιστο της τετραγωνικής προσέγγισης:

$$\max l_Q(b) \Leftrightarrow \nabla l_Q(b) = 0 \Leftrightarrow \nabla l(b^0) + H(b^0)(b - b^0) = 0 \Leftrightarrow b = b^0 - H^{-1}(b^0)\nabla l(b^0).$$

Ξεκινώντας από την τιμή $b^0$ για παράμετρο $b$, χρησιμοποιείται η φόρμουλα

$$b^{k+1} = b^k - H^{-1}(b^k)\nabla l(b^k),$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

και για πρακτικούς λόγους θα τεθεί μία υπόθεση, ώστε ο αλγόριθμος να λάβει τέλος ύστερα από κάποιες επαναλήψεις. Η υπόθεση αυτή μπορεί για παράδειγμα να είναι η απόλυτη σύγκλιση $\left|x^{k+1}-x^{k}\right|<c$ όπου $c>0$.

**Εφαρμογή της μεθόδου στο εκθετικό μοντέλο**

Για να εφαρμοστεί η μέθοδος Newton στο εκθετικό μοντέλο (1), πρέπει να υπολογιστεί η ποσότητα

$$\nabla l(b)=\left(-n+\sum_{i=1}^{n}\frac{t_i}{\exp(b_0+b_1x_i)},\sum_{i=1}^{n}\frac{x_it_i}{\exp(b_0+b_1x_i)}-\sum_{i=1}^{n}x_i\right),$$

$$H(b)=\begin{pmatrix}-\sum_{i=1}^{n}\frac{t_i}{\exp(b_0+b_1x_i)} & -\sum_{i=1}^{n}\frac{x_it_i}{\exp(b_0+b_1x_i)}\\ -\sum_{i=1}^{n}\frac{x_it_i}{\exp(b_0+b_1x_i)} & -\sum_{i=1}^{n}\frac{x_i^2t_i}{\exp(b_0+b_1x_i)}\end{pmatrix}.$$

Φυσικά η μέθοδος λειτουργεί και για την κατανομή Weibull, για την οποία όμως η παράμετρος σ θα πρέπει να περιέχεται στο διάνυσμα $b$.

## 4.5 ΔΙΑΚΥΜΑΝΣΗ ΚΑΙ ΣΥΝΔΙΑΚΥΜΑΝΣΗ ΤΟΥ ΕΜΠ

Είναι γνωστό ότι ο ΕΜΠ είναι ασυμπτωτικά αμερόληπτος (Morgenthaler (1997)). Παρ'όλα αυτά ενδιαφέρει η εκτίμηση της διακύμανσης του ΕΜΠ ώστε να καθοριστούν στη συνέχεια διαστήματα εμπιστοσύνης και διάφορα τεστ στο μοντέλο. Εκτιμάται ασυμπτωτικά η διακύμανση του ΕΜΠ (Cox (1972))με την χρήση του πίνακα πληροφορίας του Fisher:

$$I_{ij}=E\left[-\frac{\partial^2 l(b)}{\partial b_i\partial b_j}\right].$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

# ΚΕΦΑΛΑΙΟ ΠΕΜΠΤΟ

## 5.1 ΕΜΠ ΚΑΙ ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ ΕΚΤΙΜΗΤΡΙΕΣ ΓΙΑ ΛΟΓΟΚΡΙΜΕΝΑ ΔΕΔΟΜΕΝΑ

Το προηγούμενο κεφάλαιο ήταν μια εκτενής αναφορά σε δείγμα πλήρων δεδομένων που ακολουθούσαν ένα εκθετικό μοντέλο. Οι διαδικασίες που ακολουθήθηκαν για την εκτίμηση των παραμέτρων του μοντέλου είναι οι ίδιες και για δείγμα με λογοκριμένα δεδομένα. Οι μέθοδοι που θα παρουσιαστούν βρίσκουν εφαρμογή τόσο στη θεωρία αξιοπιστίας και την ανάλυση επιβίωσης όσο και στον αναλογισμό.

### 5.1.1 Συνάρτηση πιθανοφάνειας για λογοκριμένα δεδομένα

Θα ξεκινήσουμε με την εύρεση της συνάρτησης πιθανοφάνειας, αφού αυτή θα δώσει τη δυνατότητα υπολογισμού της ΕΜΠ ώστε να γίνει προσαρμογή ενός μοντέλου με λογοκριμένα δεδομένα. Για να αναλυθούν τέτοια δεδομένα χρησιμοποιείται η δίτιμη μεταβλητή $C_i$ (η οποία ενδέχεται να είναι και απλή σταθερά) που ορίζεται ως εξής:

$$C_i = \begin{cases} 1, \text{ αν η } i \text{ παρατήρηση δεν είναι λογοκριμένη} \\ \\ 0, \text{ αν η } i \text{ παρατήρηση είναι λογοκριμένη.} \end{cases}$$

Ας θεωρήσουμε ότι η συνάρτηση πυκνότητας $f_b(t_i)$ του χρόνου ζωής $T_i$ της $i$ μονάδας εξαρτάται από την παράμετρο $\beta$ και ας συμβολίσουμε με $F_\beta(t)$ και $S_\beta(t)$ την αντίστοιχη συνάρτηση κατανομής και συνάρτηση επιβίωσης.

Η συνάρτηση πιθανοφάνειας για τα λογοκριμένα δεδομένα είναι παρόμοια εκείνης που χρησιμοποιήθηκε για τα πλήρη, με τη διαφορά ότι πρέπει να εισαχθεί σε αυτή και η πληροφορία που προκύπτει από τα λογοκριμένα δεδομένα. Για δεξιά λογοκριμένα δεδομένα έχουμε

$$L(b) = \prod_{i=1}^{n} f_b^{c_i}(t_i) S^{(1-c_i)}(t_i),$$



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

όπου $S_\beta(t) = P(T > t) = 1 - F_\beta(t) = 1 - F(t)$ είναι η συνάρτηση επιβίωσης. Για τα αριστερά λογοκριμένα δεδομένα έχουμε

$$L(b) = \prod_{i=1}^{n} f_b^{c_i}(t_i) F^{(1-c_i)}(t_i),$$

και τέλος για τη λογοκρισία σε διάστημα λαμβάνεται ο τύπος

$$L(b) = \prod_{i=1}^{n} f_b^{c_i}(t_i) [S(L_i) - S(R_i)]^{(1-c_i)},$$

Σημειώνεται ότι οι στη συνέχεια θα χρησιμοποιούνται εναλλακτικά οι συμβολισμοί $f_b(t_i)$ και $f(t_i \mid b)$ για να δηλώσουν τη συνάρτηση πυκνότητας του χρόνου ζωής $T$. Αντίστοιχη παρατήρηση ισχύει και για τη συνάρτηση κατανομής και τη συνάρτηση επιβίωσης οι οποίες θα σημειώνονται ως $F_\beta(t)$ ή $F(t \mid \beta)$ και $S_\beta(t)$ ή $S(t \mid \beta)$ αντίστοιχα.

Ο γενικός τύπος της συνάρτησης πιθανοφάνειας για όλες τις περιπτώσεις λογοκρισίας είναι ο ακόλουθος:

$$L(\beta) = \left( \prod_{i \in P} f(t_i \mid \beta) \right) \left( \prod_{i \in C_r} S(t_i \mid \beta) \right) \left( \prod_{i \in C_l} [1 - S(t_i \mid \beta)] \right) \left( \prod_{i \in I} [S(L_i \mid \beta) - S(R_i \mid \beta)] \right).$$

Τα σύνολα $P, C_r, C_l, I$ αποτελούν διαμέριση του συνόλου των δεικτών $\{1,2,...,n\}$. Το σύνολο $P$ περιέχει τους δείκτες που δηλώνουν τα πλήρη δεδομένα, το σύνολο $C_r$ περιέχει τους δείκτες που δηλώνουν τα λογοκριμένα δεδομένα από δεξιά, το σύνολο $C_l$ περιέχει τους δείκτες που δηλώνουν τα λογοκριμένα δεδομένα από αριστερά ενώ τέλος, το σύνολο $I$ περιέχει τους δείκτες που δηλώνουν τα λογοκριμένα δεδομένα σε διάστημα. Ανάλογα με την περίπτωση που αντιμετωπίζει κάθε φορά ο στατιστικός αναλυτής χρησιμοποιεί και την κατάλληλη συνάρτηση πιθανοφάνειας με τις ανάλογες μετατροπές.

### 5.1.2 Είδη λογοκρισίας και συνάρτηση πιθανοφάνειας

**Λογοκρισία σε διάστημα, αριστερή λογοκρισία**

Η αριστερή λογοκρισία και η λογοκρισία σε διάστημα (Long (1979)) δεν απαντώνται αρκετά συχνά όμως η εξέτασή τους απαιτεί ιδιαίτερη προσοχή. Στην περίπτωση της αριστερής λογοκρισίας οι

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

παρατηρήσεις θα είναι της μορφής $(Z_i, C_i)$, $i = 1,2,...,n$ όπου $Z_i = \max(T_i, t_0)$ . Συνεπώς προκύπτει ότι

$$f_{Z_i,C_i}(t_i,0) = f(t_i,0) = \begin{cases} 0, & t_i > t_0 \\ P(T_i < t_i) = F(t_i), & t_i = t_0, \end{cases}$$

και

$$f_{Z_i,C_i}(t_1,1) = f(t_i,1) = \lim_{h\to 0} \frac{P(t_i \le Z_i \le t_i + h, C_i = 1)}{h} = \lim_{h\to 0} \frac{P(t_i \le T_i \le t_i + h)}{h} = f(t_i),$$

Επομένως

$$f_{Z_i,C_i}(t_i,c_i) = f(t_i,c_i) = f(t_i)^{c_i} F(t_i)^{1-c_i}, \; c_i \in \{0,1\}.$$

Συμπερασματικά, η συνάρτηση πιθανοφάνειας δίνεται από τον τύπο

$$L(b) = \prod_{i=1}^{n} f(t_i, c_i \mid b, f) = \prod_{i=1}^{n} [f(t_i \mid b)]^{c_i} [F(t_i \mid b)]^{1-c_i} = \left( \prod_{i\in P}^{n} f(t_i \mid b) \right) \left( \prod_{i\in C_l} F(t_i \mid b) \right)$$

**Δεξιά λογοκρισία τύπου Ι**

Όταν συναντάται η περίπτωση της δεξιάς λογοκρισίας τύπου Ι οι παρατηρήσεις θα είναι της μορφής $(Z_i, C_i)$, $i = 1,2,...,n$ όπου $Z_i = \min(T_i, t_0)$ .Συνεπώς προκύπτει ότι:

$$f_{Z_i,C_i}(t_i,0) = f(t_i,0) = \begin{cases} 0, & t_i > t_0 \\ P(T_i > t_i) = S(t_i), & t_i = t_0, \end{cases}$$

και

$$f_{Z_i,C_i}(t_1,1) = f(t_i,1) = \lim_{h\to 0} \frac{P(t_i \le Z_i < t_i + h, C_i = 1)}{h} = \lim_{h\to 0} \frac{P(t_i \le T_i < t_i + h)}{h} = f(t_i),$$

$0 \le t_i \le t_0$.

Επομένως

$$f_{Z_i,C_i}(t_i,c_i) = f(t_i,c_i) = f(t_i)^{c_i} S(t_i)^{1-c_i}, \; 0 \le t_i \le t_0, \; c_i \in \{0,1\},$$

και η συνάρτηση πιθανοφάνειας θα δίνεται από τον τύπο

$$L(b) = \prod_{i=1}^{n} f(t_i, c_i \mid b, f) = \prod_{i=1}^{n} [f(t_i \mid b)]^{c_i} [S(t_i \mid b)]^{1-c_i} = \left( \prod_{i\in P}^{n} f(t_i \mid b) \right) \left( \prod_{i\in C_r} S(t_i \mid b) \right)$$

**Δεξιά λογοκρισία τύπου ΙΙ**

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Όταν η μορφή της λογοκρισίας είναι τύπου ΙΙ, τότε οι παρατηρήσεις είναι της μορφής $(Z_{(i)}, C_i)$, $i = 1,2,...,n$ όπου

$$T_{(i)} = \begin{cases} T_{(i)}, 1 \text{ για } i \text{ για } r \\ \\ T_{(r)}, \ r+1 \text{ για } i \text{ για } n \ . \end{cases}$$

Η από κοινού κατανομή του δείγματος θα εκφράζεται ως

$$\frac{n!}{(n-r)!}\prod_{i=1}^{n}\left[f(t_i \mid \theta)\right]^{c_i}\left[S(t_{(r)} \mid \theta)\right]^{1-c_i} = \frac{n!}{(n-r)!}\left[S(t_{(r)} \mid \theta)\right]^{n-r}\prod_{i=1}^{r}\left[f(t_{(i)} \mid \theta)\right],$$

και η συνάρτηση πιθανοφάνειας δίνεται από τον τύπο

$$L(b) = \left[S(t_{(r)} \mid b)\right]^{n-r}\prod_{i=1}^{r} f(t_i \mid b) = \left(\prod_{i \in P} f(t_i \mid b)\right)\left(\prod_{i \in C_r} S(t_i \mid b)\right),$$

**Τυχαία λογοκρισία ή λογοκρισία τύπου IV**

Στην περίπτωση της τυχαίας λογοκρισίας οι παρατηρήσεις είναι της μορφής $(Z_i, C_i)$, $i = 1,2,...,n$ όπου

$$Z_i = \min(T_i, S_i).$$

Για ευκολία στους υπολογισμούς θα χρησιμοποιηθούν και πλαι οι συμβολισμοί $f(\cdot), S(\cdot)$ για τη συνάρτηση πυκνότητας και την συνάρτηση επιβίωσης της τυχαίας μεταβλητής $T$, ενώ με $g(\cdot), G(\cdot)$ θα συμβολίσουμε τη συνάρτηση πυκνότητας και την συνάρτηση επιβίωσης της τυχαίας μεταβλητής $S_i$. Οι υπολογισμοί θα γίνουν υπό την υπόθεση της ανεξαρτησίας των $S$ και $T$.

Η από κοινού συνάρτηση πυκνότητας πιθανότητας είναι η παρακάτω:

$$f_{Z_i, C_i}(t_i, c_i) = f(t_i, c_i) = \lim_{h \to 0} \frac{P(t_i \text{ για } Z_i < t_i + h, C_i = c_i)}{h}, \ t_i \text{ για } 0, c_i \in \{0,1\},$$

και

$$f_{Z_i, C_i}(t_i, 0) = f(t_i, 0) = \lim_{h \to 0} \frac{P(t_i \text{ για } Z_i < t_i + h, C_i = 0)}{h} = \lim_{h \to 0} \frac{P(t_i \text{ για } S_i < t_i + h, T_i > S_i)}{h} =$$

$$= P(T_i > S_i) \lim_{h \to 0} \frac{P(t_i \text{ για } S_i < t_i + h)}{h} = S(t_i) g(t_i),$$

$$f_{Z_i, C_i}(t_i, 1) = f(t_i, 1) = \lim_{h \to 0} \frac{P(t_i \le Z_i < t_i + h, C_i = 1)}{h} = \lim_{h \to 0} \frac{P(t_i \le T_i < t_i + h, T_i \le S_i)}{h} =$$

$$= P(S_i \ge T_i) \lim_{h \to 0} \frac{P(t_i \le T_i < t_i + h)}{h} = G(t_i) f(t_i).$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Άρα

$$f_{Z_i,C_i}(t_i,c_i) = f(t_i,c_i) = [f(t_i)G(t_i)]^{c_i}[g(t_i)S(t_i)]^{1-c_i},\ t_i \geq 0,\ c_i \in \{0,1\}.$$

Η συνάρτηση πιθανοφάνειας θα δίνεται από τον τύπο

$$L(\beta,\phi) = \prod_{i=1}^{n} f(t_i,c_i \mid \beta,\phi) = \prod_{i=1}^{n}[f(t_i \mid \beta)G(t_i \mid \phi)]^{c_i}[g(t_i \mid \phi)S(t_i \mid \beta)]^{1-c_i} =$$

$$= \left(\prod_{i=1}^{n}[f(t_i \mid \beta)]^{c_i}[S(t_i \mid \beta)]^{1-c_i}\right)\left(\prod_{i=1}^{n}[G(t_i \mid \phi)]^{c_i}[g(t_i \mid \phi)]^{1-c_i}\right)$$

Λόγω της υπόθεσης της ανεξαρτησίας που διατυπώθηκε στην αρχή δεν υπάρχει κάποια σχέση μεταξύ των παραμέτρων $\beta$ και $\varphi$ και συνεπώς ο όρος

$$\left(\prod_{i=1}^{n}[G(t_i \mid \phi)]^{c_i}[g(t_i \mid \phi)]^{1-c_i}\right),$$

ο οποίος εμφανίζεται στη συνάρτηση πιθανοφάνειας μπορεί να θεωρηθεί ως μία σταθερά που δεν επηρεάζει την εύρεση της εκτιμήτριας μέγιστης πιθανοφάνειας της παραμέτρου $b$. Επομένως ως συνάρτηση πιθανοφάνειας προς μεγιστοποίηση μπορεί να χρησιμοποιηθεί η ποσότητα

$$L(\beta) = \left(\prod_{i=1}^{n}[f(t_i \mid \beta)]^{c_i}[S(t_i \mid \beta)]^{1-c_i}\right) = \prod_{i\in P} f(t_i \mid \theta)\prod_{\iota\in C} S(t_i \mid \theta),$$

όπου τα σύνολα $P$ και $C$ έχουν τον ίδιο ακριβώς τρόπο ορισμού όπως και προηγούμενα.

**Περικομμένα δεδομένα**

Όταν στο δείγμα που έχει επιλεγεί περιέχονται περικομμένα δεδομένα από δεξιά, δηλαδή ο χρόνος ζωής $T$ καταγράφεται μόνο όταν ισχύει $T < y$ (όπου $y$ είναι ο χρόνος περικοπής), τότε

$$f(t \mid T < y)\frac{d}{dt}F(t \mid T < y) = \frac{f(t)}{F(y)} = \frac{f(t)}{[1-S(y)]} \text{ για } t < y,$$

και η συνάρτηση πιθανοφάνειας θα είναι η ακόλουθη

$$L(\beta) = \prod_{i=1}^{n}\frac{f(t_i \mid \beta)}{1-S(y \mid \beta)}.$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Όταν στο δείγμα που έχει επιλεγεί περιέχονται περικομμένα δεδομένα από αριστερά, δηλαδή ο χρόνος ζωής $T$ καταγράφεται μόνο αν ισχύει $T \geq y$, τότε

$$f(t \mid T > y) = \frac{d}{dt} F(t \mid T > y) = \frac{d}{dt}\left[1 - S(t \mid T > y)\right] = \frac{d}{dt}\left[1 - \frac{S(t)}{S(y)}\right] = \frac{d}{dt}\left(\frac{S(y) - S(t)}{S(y)}\right) = \frac{f(t)}{S(y)}.$$

Δηλαδή

$$S(t \mid T \geq y) = \frac{S(t)}{S(y)} \text{ για } t > y,$$

και η συνάρτηση πιθανοφάνειας παίρνει τη μορφή

$$L(\beta) = \prod_{i=1}^{n} \frac{f(t_i \mid \beta)}{S(y \mid \beta)}.$$

Τέλος υπάρχουν και περιπτώσεις στις οποίες στο δείγμα να περιέχονται περικομμένα και λογοκριμένα δεδομένα. Η διαδικασία εύρεσης της συνάρτησης πιθανοφάνειας γίνεται πολυπλοκότερη αλλά όχι και αδύνατη.

### 5.1.3 Παράδειγμα εκτίμησης παραμέτρων για δεξιά λογοκριμένα δεδομένα

**α. Δεξιά λογοκρισία τύπου Ι**

Στο σημείο αυτό θα παρουσιαστεί η προσαρμογή ενός παραμετρικού μοντέλου σε δείγμα με λογοκριμένα δεδομένα. Για το σκοπό αυτό θα υιοθετηθεί το μοντέλο της εκθετικής

$$T_i = \exp(b_0 + b_1 x_i) Z_i, \; Z_i \sim Exp(\frac{1}{s}), \; b_0, b_1 \in R, s > 0, i = 1,2,...,n.$$

Αρχικά και για λόγους ευκολίας θα γίνει μία παρουσίαση του μοντέλου στην απλή περίπτωση που $Z_i \sim \exp(1)$.

Στο συγκεκριμένο μοντέλο, η $x_i$ είναι μια επεξηγηματική μεταβλητή, η οποία ερμηνεύεται ως η ηλικία του $i$ ασθενή. Η χρησιμότητα του μοντέλου έγκειται στο γεγονός ότι η $x_i$ λειτουργεί πολλαπλασιαστικά στο χρόνο, διότι επιταχύνει ή επιβραδύνει τον χρόνο ζωής, εξαρτώμενη πάντα από το πρόσημο της παραμέτρου $b_1$ (Hosmer & Lemeshow (1999)). Το μοντέλο έχει έννοια, καθώς παίρνει μόνο θετικές τιμές για οποιαδήποτε τιμή των παραμέτρων $b_0, b_1 \in R$. Ένα επιπλέον πλεονέκτημα αφορά τον αναμενόμενο χρόνο ζωής:



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

$$\mathrm{E}(T_i \mid X_i) = \mathrm{E}[\exp(\beta_0 + \beta_1 x_i) Z_i \mid X_i] = \exp(\beta_0 + \beta_1 x_i)\mathrm{E}[Z_i] = \exp(\beta_0 + \beta_1 x_i).$$

Η αθροιστική κατανομή της τυχαίας μεταβλητής $T_i$ είναι:

$$F_T(t) = P(T_i \le t) = P(\exp(b_0 + b_1 x_i) Z_i \le t) = P\left( Z_i \le \frac{t}{\exp(b_0 + b_1 x_i)} \right) =$$

$$F_{\exp(1)}\left( \frac{t}{\exp(b_0 + b_1 x_i)} \right) = 1 - e^{-\frac{t}{\exp(b_0 + b_1 x_i)}}, t \ge 0$$

Με άλλα λόγια

$$T_i \sim \exp\left( \frac{t}{\exp(b_0 + b_1 x_i)} \right) b_0, b_1 \in R, i = 1,2,...,n.$$

Η μελέτη του $\log(T_i)$ αντί της $T_i$ έχει το πλεονέκτημα της δημιουργίας γραμμικού μοντέλου. Συνεπώς:

$$\log(T_i) = b_0 + b_1 x_i + \log(Z_i) \ , b_0, b_1 \in R, i = 1,2,...,n$$

Είναι δυνατή η εκτίμηση των παραμέτρων $b_0$ και $b_1$ χρησιμοποιώντας την μέθοδο των ελαχίστων τετραγώνων.

Έστω $\varepsilon_i = \log(Z_i)$, τότε η συνάρτηση κατανομής του $e_i$ είναι η εξής:

$$F_{e_i}(t) = P(\log(Z_i) \le t) = P(Z_i \le e^t) = F_z(e^t) = 1 - e^{-e^t}, \ t \in R.$$

ενώ η συνάρτηση πυκνότητας πιθανότητας

$$f_{e_i}(t) = e^t \exp(-e^t), \ t \in R.$$

Για να γίνει μια απλή γενικευση του εκθετικού μοντέλου είναι δυνατή η προσθήκη της παραμέτρου $s \ge 0$ στο $e_i$. Το μοντέλο τότε παίρνει την παρακάτω μορφή $\log(T_i) = b_0 + b_1 x_i + s e_i$. Συνεπώς η συνάρτηση κατανομής του $e_i$ είναι η

$$F_{se}(t) = P(se \le t) = P(e \le \frac{t}{s}) = 1 - \exp(-e^{\frac{t}{s}}) \ \ t \in R, s > 0.$$

Προκύπτει ότι

$$F_{\exp(se)}(t) = P(\exp(se) \le t) = P(se \le \log t) = 1 - \exp(-t^{\frac{1}{s}}) \ \ t > 0, s > 0,$$

το οποίο είναι

$$\mathrm{T}_i = \exp(b_0 + b_1 x_i) Z_i \ \text{και} \ \ Z_i \sim W(1, \frac{1}{s}), b_0, b_1 \in R, s > 0, i = 1,2,...,n.$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Γνωστό είναι επίσης ότι η γενική μορφή της συνάρτησης κατανομής της Weibull (Nelson (2004))είναι η ακόλουθη

$$F_W(t) = 1 - \exp\left[-\left(\frac{t}{b}\right)^c\right] \quad t > 0, b, c > 0.$$

Για το συγκεκριμένο παράδειγμα , παρατηρείται ότι $T_i \sim W(\exp(b_0 + b_1 x_i), \frac{1}{s})$,

$$P(T_i \le t) = P(Z_i \le \frac{t}{\exp(b_0 + b_1 x_i)}) = 1 - \exp(-(\frac{t}{\exp(b_0 + b_1 x_i)})^{\frac{1}{s}}).$$

Θα επανέλθουμε τώρα στο αρχικό μας μοντέλο όπου $Z_i \sim Exp(\frac{1}{\sigma})$ για να εκτιμήσουμε τις παραμέτρους $b_0, b_1$ και $s$ για δεξιά λογοκριμένα δεδομένα με την βοήθεια του ΕΜΠ.

$$L(b_0, b_1, s) = \prod_{i=1}^{n}\left[\frac{(\frac{1}{s})t_i^{(\frac{1}{s}-1)}}{\exp(b_0 + b_1 x_i)^{\frac{1}{s}}}\exp-(\frac{t_i}{\exp(b_0 + b_1 x_i)})\right]^{c_i}\left[\exp-\left(\frac{t_i}{\exp(b_0 + b_1 x_i)}\right)^{\frac{1}{s}}\right]^{(1-c_i)}$$

$$= \prod_{i=1}^{n}\left\{\left[\frac{\left(\frac{1}{s}\right)t_i^{(\frac{1}{s}-1)}}{\exp(b_0 + b_1 x_i)^{\frac{1}{s}}}\right]^{c_i}\exp-\left[\left(\frac{t_i}{\exp(b_0 + b_1 x_i)}\right)^{\frac{1}{s}}\right]\right\}.$$

Λογαριθμίζοντας προκύπτει ο ακόλουθος πιθανολογάριθμος

$$\sum_{i=1}^{n}\left\{c_i\left[-\log(\sigma) + \left(\frac{1}{\sigma} - 1\right)\log(t_i) - \frac{1}{\sigma}(\beta_0 + \beta_1 x_i)\right] - \frac{t_i}{\sigma\exp(\beta_0 + \beta_1 x_i)}\right\}.$$

Οι μερικές παράγωγοι που θα προκύψουν με την παραπάνω διαδικασία είναι δύσκολο να υπολογιστούν. Για το σκοπό αυτό θα χρησιμοποιηθεί η μέθοδος Newton όπως παρουσιάστηκε στα προηγούμενα. Τα ίδια βήματα ακολουθούνται για αριστερά λογοκριμένα δεδομένα αλλά και για λογοκρισία σε διάστημα. Ωστόσο, στην περίπτωση της λογοκρισίας σε διάστημα είναι ιδιαίτερα δύσκολη η μεγιστοποίηση του πιθανολογάριθμου (Polak (1997)).

**Ειδική περίπτωση**: $b_1 = 0$ και $s = 1$

Στο πολύ απλό μοντέλο για δεξιά λογοκριμένα δεδομένα

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

$$T_i = \exp(\beta_0)Z_i,\ Z_i \sim \exp(1),\ b_0 \in R, i = 1,2,...,n,$$

η εκτίμηση του $b_0$ είναι εύκολη και προκύπτει απο την επίλυση της εξίσωσης

$$\frac{\partial l}{\partial \beta_0}(\beta_0) = 0.$$

Εισάγωντας τους παρακάτω συμβολισμούς

$n_n$ :το πλήθος των πλήρων δεδομένων

$n_c$ :το πλήθος των λογοκριμένων δεδομένων

$n = n_n + n_c$: συνολικό πλήθος δεδομένων

μπορούμε να γράψουμε

$$l(\beta_0) = \sum_{i=1}^{n} c_i \beta_0 - \frac{t_i}{\exp(\beta_0)},$$

$$l(\beta_0) = 0 \Leftrightarrow \sum_{i=1}^{n} c_i \beta_0 - \frac{t_i}{\exp(\beta_0)} = 0$$

$$\Leftrightarrow \exp(b_0) = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} c_i} = \frac{\sum_{i=1}^{n} t_i}{n_n} \Leftrightarrow \hat{b}_0 = \log\left(\frac{\sum_{i=1}^{n} t_i}{n_n}\right).$$

Το $\sum_{i=1}^{n} t_i$ στις παραπάνω εκφράσεις εμπεριέχει την τιμή $t_i$ για τα πλήρη και την τιμή με βάση την οποία έγινε η λογοκρισία, για τα λογοκριμένα. Ο αντίστοιχος εκτιμητής για τα πλήρη δεδομένα ήταν

$$\hat{b}_0 = \log \bar{t}.$$

**β. Δεξιά λογοκρισία τύπου ΙΙ**

Εδώ τα δεδομένα αποτελούνται από τις παρατηρηθείσες τιμές $t_{(1)}, t_{(2)}, ..., t_{(r)}$ των $r$ τυχαίων μεταβλητών

$$T_{(1)} \le T_{(2)} \le ... \le T_{(r)},$$

($r$ δεδομένος ακέραιος, $1 \le r \le n$). Οι υπόλοιπες μονάδες (άτομα) απέτυχαν (απεβίωσαν) μετά την χρονική στιγμή $t_{(r)}$. Εάν βρεθεί η βαθμίδα αποτυχίας, είναι εύκολος και ο υπολογισμός της συνάρτησης πυκνότητας πιθανότητας και το αντίστροφο. Αυτό διευκολύνει κατά

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

πολύ την εύρεση της εκτιμήτριας μέγιστης πιθανοφάνειας των παραμέτρων της κατανομής που ακολουθούν τα δεδομένα.

Αρχικά πρέπει να υπολογίστεί η από κοινού συνάρτηση πυκνότητας των διατεταγμένων χρόνων ζωής $T_{(i)}$. Είναι γνωστό ότι αν $f, F$ είναι η σ.π. και η σ.κ. αντίστοιχα των τυχαίων μεταβλητών $T_1, T_2, ..., T_n$ και $1 \le r \le n$, τότε η απο κοινού σ.π. των διατεταγμένων παρατηρήσεων $T_{(1)}, T_{(2)}, ..., T_{(r)}$, σύμφωνα με την θεωρία διατεταγμένων δειγμάτων δίνεται από τον τύπο

$$f_{T_{(1)},T_{(2)},...,T_{(r)}}(\tau_1, \tau_2, ..., \tau_r) = \frac{n!}{(n-r)!}[1 - F(\tau_r)]^{n-r} \prod_{i=1}^{r} f(\tau_i), \text{ για } 0 \le t_1 \le t_2 \le ... \le t_r.$$

Έστω περίπτωση που οι χρόνοι ζωής $T_i$ ακολουθούν την εκθετική κατανομή( $f(t) = \lambda e^{-\lambda t}$, $F(t) = 1 - e^{-\lambda t}$), συνεπώς προκύπτει ότι

$$f_{T_{(1)},T_{(2)},...,T_{(r)}}(t_1, t_2, ..., t_r) = \frac{n!}{(n-r)!}[1 - e^{-l t_r}]^{r} \prod_{i=1} l e^{-l t_r},$$

και η συνάρτηση πιθανοφάνειας είναι η

$$L(l; t_1, t_2, ..., t_r) = f_{T_{(1)},T_{(2)},...,T_{(r)}}(t_1, t_2, ..., t_r) =$$

$$\frac{n!}{(n-r)!} \lambda^r \exp( -\lambda ( \sum_{i=1}^{r} \tau_i + (n-r)\tau_r)) \quad .$$

Ο πιθανολογάριθμος αντίστοιχα είναι

$$l(l; t_1, t_2, ..., t_r) = \log L(l; t_1, t_2, ..., t_r) = r \ln l - l \sum_{i=1}^{r} t_i - l(n-r)t_r,$$

και ελαχιστοποιείται για

$$\lambda = \frac{r}{\sum_{i=1}^{r} \tau_i + (n-r)\tau_r}.$$

Άρα η εκτιμήτρια μέγιστης πιαθανοφάνειας της βαθμίδας αποτυχίας $l$ δίνεται απο τον τύπο

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^{r} T_{(i)} + (n-r)T_{(r)}}.$$

Θα εξετάσουμε στη συνέχεια ένα παράδειγμα όπου οι χρόνοι ζωής ακολουθούν την κατανομή Γάμμα. Ένας ερευνητής έχει στη διάθεσή του $n = 45$ όμοιες μονάδες, οι οποίες έχουν λειτουργήσει

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ανεξάρτητα η μία από την άλλη και υποθέτουμε ότι, η βαθμίδα αποτυχίας των μονάδων είναι της μορφής

$\lambda(t)=\frac{\lambda^{\alpha}}{\Gamma(\alpha,\lambda t)}t^{a-1}e^{-\lambda t}, t\geq 0$, όπου $a, l$ άγνωστες θετικές παράμετροι ( το σύμβολο $\Gamma(\alpha,\lambda t)$ δίνει τη μη πλήρη συνάρτηση Γάμμα). Έστω επίσης ότι δεν είναι διαθέσιμο όλο το δείγμα αλλά υπάρχει λογοκρισία στις παρατηρήσεις μετά την $25^{\eta}$ διατεταγμένη (λογοκρισία τύπου II, με $r=25$). Θεωρώντας το $\alpha$ γνωστό και ίσο με 4 ζητάμε να εκτιμήσουμε το $\lambda$..

Εφόσον ο ερευνητής δεν έχει στη διάθεσή του όλο το δείγμα αλλά υφίσταται λογοκρισία μετά την $25^{\eta}$ διατεταγμένη παρατήρηση, θα χρησιμοποιηθεί η συνάρτηση πιθανοφάνειας όπως ακριβώς αναπτύχθηκε στα προηγούμενα, δηλαδή

$$L(\lambda;\tau_1,\tau_2,...,\tau_r)=$$

$$f_{T_{(1)},T_{(2)},...,T_{(r)}}(t_1,t_2,...,t_r)=\frac{45!}{(45-25)!}\left[\frac{G(4,l\,t_r)}{G(4)}\right]^{45-25}\prod_{i=1}^{25}\frac{l^4}{G(4)}t^{4-1}e^{-lt}.$$

Λογαριθμίζοντας και παραγωγίζοντας ως προς $l$ παίρνουμε την εξίσωση

$$-\sum_{i=1}^{25}t_i+\frac{4\cdot 25}{l}+(45-25)\log\left[\frac{G(4,l\,t_r)}{G(4)}\right]=0.$$

Το $t_r$ βρίσκεται από τα δεδομένα, καθώς το $r=25$ και από προηγούμενη εξίσωση μπορούμε πλέον να υπολογίσουμε το $\lambda$ (με αριθμητικές μεθόδους). Έτσι αν $t_r$=2.1 θα προκύψει ότι $\hat{l}=1.37$. Στη συνέχεια η εκτίμηση της συνάρτησης επιβίωσης είναι πλέον απλή και γίνεται μέσω του τύπου

$$S(t)=1-F(t)=\frac{\Gamma(\alpha,\hat{\lambda}t)}{\Gamma(\alpha)}=\frac{\Gamma(4,1.37t)}{\Gamma(4)}.$$

Με την ίδια διαδικασία υπολογίζει κανείς εκτιμήτριες μέγιστης πιθανοφάνειας και για την κατανομή Weibull, η οποία είναι πολύ διαδεδομένη στον τομέα της θεωρίας αξιοπιστίας.

Η μέθοδος μέγιστης πιθανοφάνειας δίνει αμερόληπτες εκτιμήτριες για τις παραμέτρους της κατανομής των δεδομένων του

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

δείγματος, και τα αποτελέσματα που προκύπτουν είναι αρκετά κοντά στα πραγματικά. Όταν οι παρατηρήσεις ακολουθούν κατανομές με παραπάνω από μία παραμέτρους, απαιτείται συνήθως η εφαρμογή μεθόδων αριθμητικής ανάλυσης, για την επίλυση των εξισώσεων που προκύπτουν.

### 5.1.4 Έλεγχοι υποθέσεων και διαστήματα εμπιστοσύνης σε δεξιά λογοκρισία τύπου I και II

Ας ξεκινήσουμε με ένα τυχαίο δείγμα χρόνων ζωής μεγέθους $n$ από την εκθετική κατανομή, δηλαδή $T \sim \exp(\lambda)$. Το δείγμα αποτελείται από μονάδες που τίθενται σε λειτουργία και κάθε μία από λειτουργεί ανεξάρτητα από την άλλη. Θα πραγματοποιηθεί έλεγχος υποθέσεων και θα κατασκευαστούν διαστήματα εμπιστοσύνης για τις περιπτώσεις δεξιάς λογοκρισίας τύπου I και II.

Αρχικά εξετάζεται η δεξιά λογοκρισία τύπου II, της οποίας η συνάρτηση πιθανοφάνειας, όπως είδαμε προηγουμένως, δίνεται από τον τύπο

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

Ο πιθανολογάριθμος είναι ίσος με

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

Παραγωγίζοντας και εξισώνοντάς το αποτέλεσμα με το μηδέν παίρνουμε

$$\frac{d}{dl} l(l) = 0 \Leftrightarrow \frac{r}{l} - \left(\sum_{i=1}^{r} t_{(i)} + (n-r)t_{(r)}\right) = 0.$$

οπότε λαμβάνεται η εκτιμήτρια μέγιστης πιθανοφάνειας της παραμέτρου $l$

$$\hat{l} = \frac{r}{\sum_{i=1}^{r} T_{(i)} + (n-r)T_{(r)}} = \frac{r}{\sum_{i=1}^{n} T_i}.$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Είναι γνωστό από την εκτιμητική ότι όταν $T_i \sim \exp(\lambda)$ τότε η ποσότητα $2\lambda\sum_{i=1}^{n} T_i$ ακολουθεί την κατανομή $X^2_{2r}$. Είναι δυνατή πλέον η κατασκευή ασυμπτωτικού διαστήματος εμπιστοσύνης, με συντελεστή $(1-a)\%$ για την παράμετρο $l$ :

$$\left[\frac{X^2_{2r,1-\left(\frac{a}{2}\right)}}{2\sum_{i=1}^{n} T_i}, \frac{X^2_{2r,\frac{a}{2}}}{2\sum_{i=1}^{n} T_i}\right].$$

Στη συνέχεια εξετάζεται η δεξιά λογοκρισία τύπου I, της οποίας η συνάρτηση πιθανοφάνειας είναι

$$L(\lambda) = \prod_{i=1}^{n} f(t_i, c_i \mid \lambda, \phi) = \prod_{i=1}^{n} \left[f(t_i \mid \lambda)\right]^{c_i} \left[S(t_i \mid \lambda)\right]^{1-c_i} = \left(\prod_{i\in P} f(t_i \mid \lambda)\right)\left(\prod_{i\in C} S(t_i \mid \lambda)\right)$$

$$= \prod_{i=1}^{n} \left(l e^{-l t_i}\right)^{c_i} \left(e^{-l t_i}\right)^{1-c_i} = \prod_{i=1}^{n} l^{c_i} e^{-l t_i} .$$

Εάν τεθεί $c = c_1 + c_2 + ... + c_n$ και $t = t_1 + t_2 + ... + t_n$ τότε η πιθανοφάνεια παίρνει μορφή:

$$L(\lambda) = \lambda^c e^{-\lambda t} .$$

Ενώ πιθανολογάριθμος θα είναι ίσος με

$$l(\lambda) = c\log(\lambda) - \lambda t .$$

Μεγιστοποιώντας τον πιθανολογάριθμο, λαμβάνεται

$$\frac{d}{d\lambda} l(\lambda) = 0 \Leftrightarrow \frac{c}{\lambda} - t = 0 \Leftrightarrow \hat{\lambda} = \frac{c}{t} = \frac{c}{\sum_{i=1}^{n} T_i} .$$

Η εκτιμήτρια μέγιστης πιθανοφάνειας της παραμέτρου $l$ είναι γνωστό ότι ακολουθεί ασυμπτωτικά τη μονοδιάστατη κανονική κατανομή με μέση τιμή $l$ και διακύμανση τον αντίστροφο του πίνακα πληροφορίας του Fisher. Ο τελευταίος είναι ίσος με

$$I(\lambda) = -\frac{d}{d\lambda^2} l(\lambda) = -\left(-\frac{c}{\lambda^2}\right) = \frac{c}{\lambda^2},$$

οπότε

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

$$\hat{l} \sim N\left( l, \frac{l^2}{c} \right).$$

Η τελευταία ποσότητα αν χρησιμοποιηθεί για την κατασκευή διαστημάτων εμπιστοσύνης θα οδηγήσει σε δύχρηστες εκφράσεις, διότι περιέχει την άγνωστη παράμετρο $l$ και στον παρονομαστή. Αν στον παρονομαστή χρησιμοποιηθεί στη θέση του $l$ το $\hat{l}$,
Θα προκύψει το επόμενο ασυμπτωτικό διάστημα εμπιστοσύνης για την παράμετρο $l$, με συντελεστή εμπιστοσύνης $(1-a)\%$:

$$\left( \hat{l} - z_{\frac{a}{2}} \sqrt{\frac{c}{T^2}}, \hat{l} + z_{\frac{a}{2}} \sqrt{\frac{c}{T^2}} \right)$$

Με την ίδια διαδικασία μπορεί να πραγματοποιηθεί στατιστική συμπερασματολογία και για την κατανομή Weibull, η οποία είναι διαδεδομένη στους τομείς της θεωρίας αξιοπιστίας και ανάλυσης επιβίωσης. Η δυσκολία που προκύπτει στην περίπτωση αυτή είναι ότι η κατανομή Weibull έγκειται στο γεγονός ότι δεν είναι μονοπαραμετρική, και επομένως θα πρέπει να αναζητηθούν ΕΜΠ και για τις δύο παραμέτρους της.

Θεωρείται ένα τυχαίο δείγμα χρόνων ζωής μεγέθους $n$ από την κατανομή Weibull, δηλαδή $T \sim W(\lambda, \alpha)$. Το δείγμα αποτελείται από μονάδες που τίθενται σε λειτουργία και κάθε μία λειτουργεί ανεξάρτητα από την άλλη. Θα εξεταστεί η περίπτωση της λογοκρισίας απο δεξιά τύπου I.
Η συνάρτηση πυκνότητας της κατανομής είναι ίση με:

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

και οδηγεί στην επόμενη συνάρτηση πιθανοφάνειας:

$$L(\lambda, a) = \prod_{i=1}^{n} f(t_i, c_i \mid \lambda, \phi) = \prod_{i=1}^{n} [f(t_i \mid \lambda)]^{c_i} [S(t_i \mid \lambda)]^{1-c_i} = \left( \prod_{i \in P}^{n} f(t_i \mid \lambda) \right) \left( \prod_{i \in C} S(t_i \mid \lambda) \right) =$$

$$= \prod_{i=1}^{n} \left[ \lambda a {t_i}^{a-1} \exp\left(-\lambda {t_i}^a\right) \right]^{c_i} \left[ \exp\left(-\lambda {t_i}^a\right) \right]^{1-c_i} = \prod_{i=1}^{n} \lambda^{c_i} a^{c_i} {t_i}^{a-1} \exp\left(-\lambda {t_i}^a\right) =$$

$$= \lambda^c a^c \prod_{i=1}^{n} \left( {t_i}^{a-1} \right) \exp\left( -\lambda \sum_{i=1}^{n} {t_i}^a \right)$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Η τελευταία σχέση προέκυψε θέτοντας $c = c_1 + c_2 + ... + c_n$, ενώ ο πιθανολογάριθμος παίρνει τη μορφή

$$l(\lambda, a) = c\log\lambda + c\log a + (a-1)\sum_{i=1}^{n}\log t_i - \lambda\sum_{i=1}^{n} t_i^a \ .$$

Η εύρεση της ΕΜΠ για την παράμετρο $l$ στην περίπτωση που η παραμέτρος $\alpha$ είναι γνωστή απαιτεί την επίλυση της εξίσωσης

$$\frac{d}{d\lambda} l(\lambda, \alpha) = 0 \Leftrightarrow \frac{c}{\lambda} - \sum_{i=1}^{n} t_i^a = 0 \Leftrightarrow \hat{\lambda} = \frac{c}{\sum_{i=1}^{n} T_i^a} \ .$$

Αφού $T \sim W(l, a)$ η τυχαία μεταβλητή $T^a$ θα ακολουθεί την εκθετική κατανομή όπως φαίνεται από την επόμενη ισότητα

$$P\left(T^a \le t\right) = P\left(T \le t^{\frac{1}{a}}\right) = F\left(t^{\frac{1}{a}}\right) = 1 - \exp\left(-l\left(t^{\frac{1}{a}}\right)^a\right) = 1 - \exp(-lt).$$

Το τυχαίο άθροισμα ($\sum_{i=1}^{n} T_i^a$) $n$ ανεξαρτήτων και ισόνομων τυχαίων μεταβλητών που ακολουθούν την εκθετική κατανομή είναι γνωστό ότι θα ακολουθεί την κατανομή Γάμμα με παραμέτρους $c, l$ (δηλαδή $\sum_{i=1}^{n} T_i^a \sim Gamma(c, \lambda)$ .

Η κατασκευή ελέγχου υποθέσεων και η δημιουργία διαστημάτων εμπιστοσύνης είναι όμοιοι με εκείνους της εκθετικής κατανομής.

Όσον αφορά την παράμετρο $a$ ο υπολογισμός του ΕΜΠ δεν είναι τόσο απλός καθώς απαιτεί αριθμητικές μεθόδους για την επίλυσή του.

## 5.2 ΜΗ ΠΑΡΑΜΕΤΡΙΚΗ ΕΚΤΙΜΗΣΗ ΓΙΑ ΤΗ ΣΥΝΑΡΤΗΣΗ ΕΠΙΒΙΩΣΗΣ

Η μη παραμετρική εκτίμηση της συνάρτησης επιβίωσης ενός συνόλου δεδομένων με λογοκριμένες παρατηρήσεις είναι αρκετά διαδεδομένη. Η χρήση της εντοπίζεται σε δείγματα για τα οποία η κατανομή είναι άγνωστη ή γίνεται μία πρώτη προσπάθεια εκτίμησης διαφόρων χαρακτηριστικών (μέτρων θέσης και διασποράς). Περιλαμβάνει ένα σύνολο μεθόδων, οι οποίες διαφέρουν τόσο σε δομή

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

όσο και στην εφαρμογή τους. Τα αποτελέσματα και τα συμπεράσματα που εξάγονται δεν συμπίπτουν πάντοτε καθώς κάθε μία έχει τα πλεονεκτήματα και τα μειονεκτήματά της. Κεντρική ιδέα και εδώ είναι η εκτίμηση της συνάρτησης επιβίωσης (αξιοπιστίας). Αυτό συμβαίνει διότι, εάν υπολογιστεί η συνάρτηση επιβίωσης, ο στατιστικός αναλυτής μπορεί να εκτιμήσει τη σ.π.π. και τη σ.κ, να πραγματοποιήσει ελέγχους και να κατασκευάσει διαστήματα εμπιστοσύνης. Τέλος, με διάφορους ελέγχους (*Q-Q Plot, P-P Plot, Kolmogorov- Smirnov test*) επιβεβαιώνει τα αποτελέσματά του. Για την εφαρμογή των μη παραμετρικών αυτών μεθόδων είναι απαραίτητη μια αναφορά στους πίνακες επιβίωσης.

### 5.2.1 Πίνακες επιβίωσης

Οι πίνακες επιβίωσης (*life tables*) είναι ιδιαίτερα χρήσιμοι σε διάφορους τομείς της αναλογιστικής επιστήμης, όπως στις μεθόδους συνταξιοδότησης και στις μεθόδους τιμολόγησης ασφαλίστρων καθώς και στο πεδίο της δημογραφίας. Η παρουσία λογοκριμένων δεδομένων τους καθιστά βασικό εργαλείο για την αντιμετώπιση τους.

Όπως και στα προηγούμενα, θεωρείται ένα τυχαίο δείγμα μεγέθους $n$ από κάποιο πληθυσμό. Το δείγμα αυτό περιέχει πλήρη και λογοκριμένα δεδομένα. Τα δεδομένα είναι συγκεντρωμένα σε ένα πίνακα επιβίωσης (ο οποίος παρουσιάζει πολλά κοινά χαρακτηριστικά με τον πίνακα συχνοτήτων) και σε $k+1$ διαστήματα της μορφής $I_j = \lfloor a_{j-1}, a_j \rangle$ με $j = 1,2,...,k+1$ για τα οποία ισχύει $a_0 = 0$ και $a_{k+1} = \infty$. Για κάθε διάστημα υπάρχουν οι εξής πληροφορίες (Berkson, Gage (1950)):

- ο αριθμός $d_j$ των θανάτων στο διάστημα $I_j$ (πλήρεις χρόνοι ζωής).
- ο αριθμός $c_j$ των απομακρύνσεων στο διάστημα $I_j$ (λογοκριμένοι χρόνοι ζωής).
- ο αριθμός $r_j$ των ατόμων που βρίσκονται σε κίνδυνο στην αρχή του διαστήματος $I_j$.

Με την φράση “ατόμα που βρίσκονται σε κίνδυνο” ($r_j$) εννοούμε εκείνα τα οποία δύνανται να αποτύχουν (θάνατοι) ή να λογοκριθούν οι

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

χρόνοι ζωής τους από τη χρονική στιγμή $a_{j-1}$ ή αργότερα. Ο αριθμός $r_j$ είναι διαφορετικός από τον πραγματικό αριθμό των ατόμων που βρίσκονται εν ζωή πέραν του διαστήματος $I_{j-1}$ διότι σε αυτόν περιλαμβάνονται και χρόνοι ζωής οι οποίοι έχουν λογοκριθεί στο παρελθόν, δηλαδή στα διαστήματα $I_1, I_2, ..., I_{j-1}$. Το $j$ έχει οριστεί από 1 έως $k+1$, παρ'όλα αυτά το διάστημα $I_{k+1}$ δε χρησιμοποιείται πάντοτε. Στην περίπτωση που χρησιμοποιηθεί, θέτουμε $r_{k+1} = d_{k+1}$ και $c_{k+1} = 0$. Στο διάστημα $I_1$ ο αριθμός των ατόμων που βρίσκονται σε κίνδυνο είναι ίσος με το συνολικό πλήθος του δείγματος, δηλαδή $r_1 = n$, ενώ για τα υπόλοιπα διαστήματα ισχύει ότι $r_j = r_{j-1} - d_{j-1} - c_{j-1}$.

Πέρα από τις παραπάνω ποσότητες είναι απαραίτητος ο ορισμός των επομένων:

- $P_j = P(T \geq a_j) = S(a_j)$, είναι η πιθανότητα ένα άτομο να επιζήσει πέραν του διαστήματος $I_j$.
- $p_j = P(T \geq a_j \mid T \geq a_{j-1}) = \frac{P_j}{P_{j-1}}$, είναι η πιθανότητα ένα άτομο να επιζήσει πέραν του διαστήματος $I_j$ δεδομένου ότι επέζησε πέραν του διαστήματος $I_{j-1}$.
- $q_j = 1 - p_j = 1 - \frac{P_j}{P_{j-1}}$, είναι η πιθανότητα ένα άτομο να πεθάνει στο διάστημα $I_j$ δεδομένου ότι επέζησε πέραν του διαστήματος $I_{j-1}$.

Οι τυχαίες μεταβλητές $T$ και $S(t)$ ορίζονται όπως και στην αρχή και παράλληλα ισχύει ότι $P_{k+1} = 0, p_{k+1} = 0, q_{k+1} = 1$, ενώ $P_0 = S(0) = 1$, $p_0 = 1, q_0 = 0$.

Βασικός σκοπός στην ανάλυση των πινάκων επιβίωσης είναι η εκτίμηση της ποσότητας $P_j = S(a_j)$. Στην ιδανική περίπτωση δείγματος με πλήρεις χρόνους ζωής, προφανής εκτιμήτρια της $P_j$ θα ήταν η ποσότητα $\frac{r_j}{n}$ ή αλλιώς το ποσοστό των ατόμων που βρίσκονται στη ζωή

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

τη χρονική στιγμή $a_j$. Όμως κάτι τέτοιο είναι αρκετά δύσκολο αφού συχνά υπάρχουν απομακρύνσεις με συνέπεια την εμφάνιση λογοκριμένων χρόνων ζωής. Συνεπώς η ποσότητα $r_j$ δε συμπίπτει με τον πραγματικό αριθμό ατόμων που βρίσκονται στη ζωή τη χρονική στιγμή $a_j$. Επομένως η ποσότητα $\frac{r_j}{n}$ υποεκτιμά την ποσότητα $P_j$.

Για το λόγο αυτό θα δοθούν αρχικά κάποιες εκτιμήσεις για τα $p_j$ και $q_j$ και εν συνεχεία θα εκτιμηθεί η ποσότητα $P_j$.

Αν στο διάστημα $I_j$ δεν υπάρχουν απομακρύνσεις τότε η εκτίμηση της ποσότητας $q_j$ είναι ίση με

$$\hat{q}_j = \frac{d_j}{r_j}.$$

Δηλώνει το ποσοστό εκείνων που πέθαναν σε σχέση με τον αριθμό των ατόμων που βρίσκονται σε κίνδυνο.

Αν στο διάστημα $I_j$ υπάρχουν απομακρύνσεις τότε η εκτίμηση του $q_j$ δε μπορεί να είναι η ίδια, διότι το $r_j$ μεταβάλλεται με διαφορετικό ρυθμό λόγο των απομακρύνσεων. Μία λύση στο ανωτέρω πρόβλημα είναι η χρησιμοποίηση της αναλογιστικής υπόθεσης (***actuarial assumption***). Η υπόθεση θεωρεί ότι οι απομακρύνσεις $c_j$ που γίνονται εντός του διαστήματος $I_j$ κατανέμονται ομοιόμορφα σε αυτό και κατά μέσο όρο μόνο οι μισές από αυτές είναι ενεργές. Συνεπώς θεωρείται ότι ο αριθμός των ατόμων που βρίσκονται σε κίνδυνο στην αρχή του διαστήματος $I_j$ είναι ίσος με

$$r_j^{'} = r_j - \frac{c_j}{2}.$$

Τελικά προκύπτει ότι

$$\hat{q}_j = \frac{d_j}{r_j^{'}} = \frac{d_j}{r_j - \frac{c_j}{2}}.$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Αναφέρεται ότι το $r_j > 0$, και στην περίπτωση που συμβαίνει ισχύει $r_j = 0$, ορίζεται αξιωματικά $\hat{q}_j = 1$.

Είναι γνωστό ότι $p_j = 1 - q_j$. Επομένως

$$\hat{p}_j = 1 - \hat{q}_j = 1 - \frac{d_j}{r_j'} = \frac{r_j' - d_j}{r_j'}.$$

Παρατηρείται όμως ότι

$$P_j = S(a_j) = P(T \tilde{\acute{\varepsilon}} a_j) = \frac{P(T \tilde{\acute{\varepsilon}} a_1)}{P(T \tilde{\acute{\varepsilon}} a_0)} \cdot \frac{P(T \tilde{\acute{\varepsilon}} a_2)}{P(T \tilde{\acute{\varepsilon}} a_1)} \cdots \frac{P(T \tilde{\acute{\varepsilon}} a_j)}{P(T \tilde{\acute{\varepsilon}} a_{j-1})} =$$
$$= P(T \tilde{\acute{\varepsilon}} a_1 \mid T \tilde{\acute{\varepsilon}} a_0) \cdot P(T \tilde{\acute{\varepsilon}} a_2 \mid T \tilde{\acute{\varepsilon}} a_1) \cdots P(T \tilde{\acute{\varepsilon}} a_j \mid T \tilde{\acute{\varepsilon}} a_{j-1}) =$$
$$= p_1 p_2 \ldots p_j,$$

οπότε

$$\hat{P}_j = \hat{p}_1 \hat{p}_2 \ldots \hat{p}_j = \prod_{i=1}^{j} \left(1 - \frac{d_i}{r_i'}\right), \quad j = 1,2,\ldots,k+1$$

Κάνοντας χρήση της εκτίμησης της συνάρτησης επιβίωσης, είναι δυνατή η εκτίμηση της σ.π.π. και της βαθμίδας αποτυχίας. των δεδομένων του δείγματος. Πιο συγκεκριμένα, αφού $f(t) = -S'(t)$ και $\lambda(t) = \frac{f(t)}{S(t)}$, θα έχουμε

για το διάστημα $t \in (a_{j-1}, a_j)$,

$$f(t) = -S'(t) = \frac{\hat{S}(a_{j-1}) - \hat{S}(a_j)}{a_j - a_{j-1}} = \frac{\hat{P}_{j-1} - \hat{P}_j}{a_j - a_{j-1}},$$

$$\hat{\lambda}(t) = \frac{f(t)}{\hat{S}(t)} = \frac{\dfrac{\hat{S}(a_{j-1}) - \hat{S}(a_j)}{a_j - a_{j-1}}}{\hat{S}(a_j)} = \frac{\dfrac{\hat{P}_{j-1} - \hat{P}_j}{a_j - a_{j-1}}}{\hat{P}_j}.$$

Οι πίνακες επιβίωσης προτιμώνται, ιδίως στην περίπτωση που ο όγκος των δεδομένων είναι μεγάλος. Όσον αφορά την επιλογή των διαστημάτων $I_j$ ισχύουν οι αρχές που χρησιμοποιούνται και στους πίνακες συχνοτήτων. Έτσι το πλήθος των τάξεων δεν πρέπει να είναι πολύ μικρό ή πολύ μεγάλο και τα πλάτη των διαστημάτων δεν είναι απαραίτητο να είναι ίσα.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

### 5.2.2 Εκτίμηση της διακύμανσης της συνάρτησης επιβίωσης

Για την εκτίμηση της διακύμανσης, θα εξετασθούν ξεχωριστά δύο περιπτώσεις. Η περίπτωση δείγματος με πλήρη δεδομένα και εκείνη δείγματος με την παρουσία λογοκριμένων δεδομένων.

- **Περίπτωση στην οποία δεν υπάρχουν απομακρύνσεις**

Θεωρείται ένα πλήρες τυχαίο δείγμα μεγέθους $n$ από ένα πληθυσμό και συμβολίζεται με $D_j \quad j=1,2,...,k+1$ η τυχαία μεταβλητή που δηλώνει τον αριθμό των θανάτων στο διάστημα $I_j$ και $\Pr_j$ η πιθανότητα ένα άτομο να πεθάνει σε αυτό το διάστημα. Θα έχουμε

$$\Pr_j = P(a_{j-1} \le T \le a_j) = S(a_{j-1}) - S(a_j) = P_{j-1} - P_j = p_1 p_2 ... p_{j-1} q_j ,$$

και

$$D_1 + ... + D_{k+1} = n .$$

Το τυχαίο διάνυσμα $(D_1 + ... + D_{k+1})$ ακολουθεί την πολυωνυμική κατανομή με συνάρτηση πιθανοφάνειας

$$L(q) = \prod_{i=1}^{n} \frac{n!}{D_1!...D_{k+1}!} p_1^{D_1} p_2^{D_2} ... p_{k+1}^{D_{k+1}} = \frac{n!}{D_1!...D_{k+1}!} \prod_{i=1}^{n} \left(p_1 p_2 ... p_{j-1} q_j\right)^{D_j} =$$

$$= \frac{n!}{D_1!...D_{k+1}!} \prod_{i=1}^{n} q_j^{D_j} p_j^{R_j - D_j} = \frac{n!}{D_1!...D_{k+1}!} \prod_{i=1}^{n} q_j^{D_j} \left(1 - q_j\right)^{R_j - D_j} .$$

όπου $R_j = D_j + D_{j+1} + ... + D_{k+1}$ είναι ο αριθμός των ατόμων που βρίσκονται εν ζωή πέραν του διαστήματος $I_{j-1}$.

Η συνάρτηση πιθανοφάνειας $L(q)$ μεγιστοποιείται όταν

$$\hat{q}_j = \frac{D_j}{R_j} ,$$

οπότε η εκτιμήτρια μέγιστης πιθανοφάνειας του $\hat{P}_j$ είναι

$$\hat{P}_j = \frac{R_{j+1}}{n} .$$

Η τυχαία μεταβλητή $R_{j+1}$ ακολουθεί τη διωνυμική κατανομή με παραμέτρους παραμέτρους $n$ και $P_j$ διότι δηλώνει τον αριθμό των

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ατόμων που βρίσκονται εν ζωή την χρονική στιγμή $a_j$ (είναι δηλαδή ένα τυχαίο άθροισμα από ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την κατανομή Bernouli). Επομένως

$$E\left(\hat{P}_j\right)=E\left(\frac{R_{j+1}}{n}\right)=\frac{1}{n}E(R_{j+1})=\frac{R_{j+1}}{n}=P_j,$$

$$V(\hat{P}_j)=V\left(\frac{R_{j+1}}{n}\right)=\frac{1}{n^2}V(R_{j+1})=\frac{P_j(1-P_j)}{n}.$$

Η εκτίμηση της διακύμανσης της εκιμήτριας $\hat{P}_j$ είναι η

$$\hat{V}(\hat{P}_j)=\hat{V}\left(\frac{R_{j+1}}{n}\right)=\frac{1}{n^2}\hat{V}(R_{j+1})=\frac{\hat{P}_j(1-\hat{P}_j)}{n}.$$

- **Περίπτωση στην οποία υπάρχουν απομακρύνσεις**

Σε αυτή την περίπτωση ισχύουν κάποια αποτελέσματα τα οποία ισχύουν και στην περίπτωση που δεν υπάρχουν διαφυγές όπως

$$E(\hat{q}_j)=q_j,\quad V(\hat{q}_j)=p_jq_j.$$

Άρα

$$E(\hat{p}_j)=E(1-\hat{q}_j)=1-q_j=p_j \text{ και } V(\hat{p}_j)=V(1-\hat{q}_j)=\frac{p_jq_j}{r_j'}.$$

Συνεπώς μία εκτίμηση της ποσότητας $V(\hat{P}_j)$ είναι η παρακάτω

$$\hat{V}(\hat{P}_j)=\left(\hat{P}_j\right)^2\sum_{i=1}^{j}\frac{\hat{q}_i}{r_i'\hat{p}_i}=\left(\hat{P}_j\right)^2\sum_{i=1}^{j}\frac{d_i}{r_i'\left(r_i'-d_i\right)}.$$

## 5.2.3 Εμπειρική συνάρτηση επιβίωσης

Ας υποτεθεί ότι $n$ μονάδες τίθενται σε λειτουργία. Καταγράφονται οι χρονικές στιγμές $t_i$, $i=1,2,...,n$ κατά τις οποίες αποτυγχάνουν. Η πλέον γνωστή εκτιμήτρια για τη συνάρτηση κατανομής είναι η εμπειρική συνάρτηση κατανομής $\hat{F}(t)$ , η οποία εκφράζει το ποσοστό των μονάδων με χρόνο ζωής μικρότερο ή ίσο του $t$. Άρα η εκτιμήτρια της αντίστοιχης συνάρτησης επιβίωσης θα είναι η

$$\hat{S}(t)=1-\hat{F}(t),$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ή ισοδύναμα

$$\hat{S}(t) = \begin{cases} 1, \text{ αν } t \le t_{(1)} \\ \dfrac{n-i+1}{n} \text{ αν } t_{(i-1)} \le t \le t_{(i)} \\ 0, \text{ αν } t \ge t_{(n)}. \end{cases}$$

Το ζήτημα που προκύπτει είναι, τι γίνεται αν θέλουμε να εκτιμήσουμε τη συνάρτηση επιβίωσης σε δείγμα με δεξιά λογοκριμένα δεδομένα. Η απάντηση είναι εξής: Είναι αδύνατος ο υπολογισμός της για χρόνους ζωής $t$ μεγαλύτερους από το δεξί όριο λογοκρισίας $R_i$ διότι αν υπάρχει παρατήρηση με χρόνο ζωής $t > R_i$, δεν είναι γνωστό το που πρέπει να τοποθετηθεί. Μία πιθανή αντιμετώπιση του προβλήματος είναι να αγνοηθούν τα λογοκριμένα δεδομένα και η εμπειρική κατανομή να υπολογιστεί για τις πλήρεις παρατηρήσεις. Φυσικά το αποτέλεσμα δεν θα είναι αμερόληπτο διότι χάνεται μέρος της πληροφορίας. Ο εκτιμητής που προκύπτει με την παραπάνω διαδικασία καλείται trivial estimator.

### 5.2.4 Εκτιμήτρια Kaplan-Meier

Ένα μεγάλο πλεονέκτημα που παρουσιάζει η εκτιμήτρια Kaplan-Meier (KM) (Kaplan and Meier (1958)) είναι η δυνατότητα αντιμετώπισης προβλημάτων με δεδομένα στα οποία υπάρχει λογοκρισία τύπου IV καθώς και σε εκείνα με δεξιά λογοκρισία (λογοκρισία τύπου II). Στην πράξη αυτό το οποίο παρατηρείται είναι οι ακριβείς χρόνικές στιγμές αποτυχίας ενώ είναι γνωστό και το πλήθος των ατόμων ή μονάδων $n_j(r_j)$ που βρίσκονται υπό παρακολούθηση (δεν έχουν λογοκριθεί ακόμα). Με βάση τα παραπάνω ενδιαφέρει η εκτίμηση της συνάρτησης επιβίωσης $S(t)$ για τα διάφορα $t \ge 0$.

Η ιδέα για τον εκτιμητή Kaplan-Meier προέκυψε ως εξής: Αν $t_j \le t_{j+1}$, $j=0,1,\ldots$ τότε μπορούμε να γράψουμε

$$S(t_j) = P(T > t_j) = P(T > t_j, T > t_{j-1}) = P(T > t_j \mid T > t_{j-1})P(T > t_{j-1}) =$$
$$= P(T > t_j \mid T > t_{j-1})P(T > t_{j-1})P(T > t_{j-1} \mid T > t_{j-2})...P(T > t_0 = 0).$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Υποτίθεται ότι στην αρχή της μελέτης όλες οι μονάδες (άτομα) βρίσκονται εν ζωή, συνεπώς $P(T > t_0 = 0) = 1$. Επίσης, με βάση όσα ανφέρθηκαν στην Παράγραφο 5.2.1, μπορούμε να γράψουμε

$$P(T > t_j \mid T > t_{j-1}) = \frac{r_j - d_j}{r_j},$$

και η εκτιμήτρια Kaplan-Meier παίρνει τη μορφή

$$\hat{S}(t) = \prod_{j:t_j<t} \frac{r_j - d_j}{r_j} = \prod_{j:t_j<t} \left(1 - \frac{d_j}{r_j}\right), \; t \geq 0.$$

Για την τελευταία χρησιμοποιείται και η εναλλακτική ονομασία **εκτιμήτρια γινομένου-ορίου** (*Product Limit Estimate PLE*). Σημειώνεται ότι οι Kaplan-Meier κατασκεύασαν αρχικά την παραπάνω εκτιμήτρια ως μη παραμετρική εκτιμήτρια μέγιστης πιθανοφάνειας της ποσότητας $S(t)$.

Είναι φανερό ότι η εκτιμήτρια ΚΜ

- Είναι συνεχής από δεξιά.
- Είναι σκαλωτή συνάρτηση (όπως η εμπειρική συνάρτηση επιβίωσης).
- Ισούται με 1 για $0 \leq t \leq t_{(1)}$.
- Ελαττώνεται κατά ένα παράγοντα ίσο με

  $$\frac{r_j - d_j}{r_j} = 1 - \frac{d_j}{r_j},$$

  στις θέσεις $t_j$ (οι χρόνοι που παρατηρούνται αποτυχίες).
- Δε μεταβάλλεται τις χρονικές στιγμές όπου γίνεται λογοκρισία (και οι οποίες είναι άγνωστες), τα συμβάντα όμως επηρεάζουν το ύψος του σκαλοπατιού της γραφικής παράστασης. Το μειονέκτημα της ΚΜ είναι ότι ποτέ δεν θα φτάσει στην τιμή 0 αν υπάρχει λογοκρισία στην τελευταία μονάδα που έχει μείνει για παρακολούθηση. Γι' αυτό θεωρείται ως μη καθορισμένη μετά το χρόνο της τελευταίας παρατηρηθείσας αποτυχίας.

Στην περίπτωση που κάποια χρονική στιγμή λογοκρισίας συμπίπτει με το χρόνο ζωής του ατόμου, θεωρείται ότι η πρώτη είναι απειροστά μικρότερη από τη δεύτερη αφού διαισθητικά είναι

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

κατανοητό ότι για να λογοκριθεί ένα άτομο θα πρέπει να έχει επιβιώσει μέχρι εκείνη τη στιγμή.

Θα πρέπει να σημειωθεί ότι το πλήθος των ατόμων που πέθαναν είναι στην πραγματικότητα τυχαία μεταβλητή $D_j$, δηλαδή η εκτιμήτρια Kaplan-Meier έχει τη μορφή

$$\hat{S}_{KM}(t) = \prod_{j:t_j<t} \frac{r_j - D_j}{r_j} = \prod_{j:t_j<t} \left(1 - \frac{D_j}{r_j}\right) = \prod_{j:t_j<t} \left(1 - \hat{q}_j\right),$$

όπου

$$\hat{q}_j = \frac{D_j}{r_j},\ j = 1,2,\ldots\ .$$

*Παράδειγμα 5.2.3.1*: Ο Freireich το 1963 συγκέντρωσε δεδομένα από ασθενείς με οξεία λευχαιμία οι οποίοι ανταποκρίθηκαν στη θεραπευτική αγωγή που ακολούθησαν (***prednisone***) με αποτέλεσμα την υποχώρηση των συμπτωμάτων της νόσου. Στη συνέχεια οι ασθενείς χωρίστηκαν σε δύο ομάδες των 21 ασθενών. Η πρώτη ομάδα ασθενών (***Treatment***) ακολούθησε θεραπευτική αγωγή με το φάρμακο 6-MP (***6-mercaptopurine***) ενώ στη δεύτερη ομάδα ασθενών χορηγήθηκε ψευδοφάρμακο (***Placebo***). Σκοπός της έρευνας ήταν να αξιολογηθεί η επίδραση του φαρμάκου 6-MP ως προς την ικανότητα του να συντηρήσει την υποχώρηση (***remission***) των συμπτωμάτων της νόσου. Τα δεδομένα της έρευνας δίνονται στον Πίνακα 5.2.1.

**Πίνακας 5.2.1**

Δεδομένα παραδείγματος *Freireich* (1963)

| Treatment Group, $n = 21$ | Placebo Group, $n = 21$ |
|---|---|
| 6, 6, 6, 7, 10, 13, 16, 22, 23,<br>6+, 9+, 10+, 11+, 17+, 19+<br>20+, 25+, 32+, 32+, 34+, 35+ | 1, 1, 2, 2, 3, 4, 4,<br>5, 5, 8, 8, 8, 8, 11, 11<br>12, 12, 15, 17, 22, 23 |
| # Failed: 9<br># Censored: 12<br>Total: 21 | # Failed: 21<br># Censored: 0<br>Total: 21 |

Στον πίνακα αυτό με την ονομασία ***Failed*** είναι τα άτομα τα οποία πέθαναν κατά τη διάρκεια της θεραπευτικής αγωγής. Με την ονομασία ***Censored*** είναι τα άτομα που για κάποιο λόγο δεν είναι γνωστό πότε πέθαναν (είτε αποχώρησαν από τη θεραπευτική αγωγή, είτε υποχώρησαν τα συμπτώματα της νόσου).

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Για τον υπολογισμό της εκτιμήτριας Kaplan-Meier για την ομάδα *Treatment* κατασκευάζεται ο ακόλουθος Πίνακας.

**Πίνακας 5.2.2**

Εκτιμήτρια Kaplan-Meier για την ομάδα *Treatment*

***PLE* για την ομάδα *Treatment***

| $t_j$ | $r_j$ | $d_j$ | $c_j$ | $1-(d_j/r_j)$ | $t$ | $\hat{S}(t)$ |
|---|---|---|---|---|---|---|
| $t_0=0$ | $r_0=21$ | $d_0=0$ | $c_0=0$ | 1 | $t\le 6$ | 1 |
| $t_1=6$ | $r_1=21$ | $d_1=3$ | $c_1=1$ | 0.857 | $6<t\le 7$ | 0.857 |
| $t_2=7$ | $r_2=17$ | $d_2=1$ | $c_2=1$ | 0.941 | $7<t\le 10$ | 0.807 |
| $t_3=10$ | $r_3=15$ | $d_3=1$ | $c_3=2$ | 0.933 | $10<t\le 13$ | 0.753 |
| $t_4=13$ | $r_4=12$ | $d_4=1$ | $c_4=0$ | 0.917 | $13<t\le 16$ | 0.690 |
| $t_5=16$ | $r_5=11$ | $d_5=1$ | $c_5=3$ | 0.910 | $16<t\le 22$ | 0.627 |
| $t_6=22$ | $r_6=7$ | $d_6=1$ | $c_6=0$ | 0.857 | $22<t\le 23$ | 0.538 |
| $t_7=23$ | $r_7=6$ | $d_7=1$ | $c_7=5$ | 0.833 | $23<t\le 35$ | 0.448 |

Όπως αναφέρθηκε και προηγουμένως στην περίπτωση που η μέγιστη παρατήρηση είναι λογοκριμένος χρόνος (στο παράδειγμα η παρατήρηση 35+ για την ομάδα *Treatment*), η εκτιμήτρια ορίζεται μέχρι αυτή την παρατήρηση.

Σύμφωνα με τον τύπο της εκτιμήτριας Kaplan-Meier ο υπολογισμός μπορεί να δίνει αναδρομικά με χρήση των σχέσεων

$$\hat{S}(t_1+)=\frac{r_1-d_1}{r_1},$$

$$\hat{S}(t_j+)=\hat{S}(t_{j-1}+)\frac{r_j-d_j}{r_j},\ j=2,3,...,k.$$

Για να εκτιμηθεί η διακύμανση της εκτιμήτριας ΚΜ, είναι αναγκαία η παρουσίαση της μεθόδου Delta. Αποτελεί μία τεχνική υπολογισμού της διακύμανσης της τυχαίας μεταβλητής $Y=f(X)$ όπου $f(\cdot)$ είναι γνωστή συνάρτηση και $X$ είναι τυχαία μεταβλητή με μέση τιμή $\mu=E(X)$. Η μέθοδος κάνει χρήση των δύο πρώτων όρων του αναπτύγματος της σειράς Taylor γύρω από το $m$, για την εκτίμηση της διακύμανσης (Samartzis (2006)). Πιο συγκεκριμένα, χρησιμοποιώντας την προσεγγιστική έκφραση

$$\begin{aligned}&f(X)=f(m)+f'(m)(X-m),\\&V(f(X))=V\big(f(m)+f'(m)(X-m)\big)=\\&=f''(m)V(X-m)=f''(m)V(X)=f''(m)s^2,\end{aligned}$$



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

όπου $\sigma^2 = V(X)$.

Για παράδειγμα, αν εφαρμόσουμε τη μεθόδο Delta για τις παρακάτω περιπτώσεις $f(x) = \log x$ και $f(x) = \exp(x)$ θα πάρουμε τα εξής:

$$\log(X) = \log(m) + (X - m)\frac{1}{m} \Rightarrow V(\ln X) = \left(\frac{1}{m}\right)^2 s^2.$$

$$\exp(X) = \exp(m) + (X - m)\exp(X) \Rightarrow V\left(\exp(X)\right) = \exp^2(m)s^2.$$

Επανερχόμαστε στην εκτιμήτρια ΚΜ

.

$$\log \hat{S}_{KM}(t) = \log \prod_{j:t_j<t}\left(1-\frac{d_j}{r_j}\right) = \sum_{j:t_j<t}\log\left(1-\frac{d_j}{r_j}\right).$$

και θεωρώντας ότι τα $d_j \sim B(r_j, 1-p_j)$ και ότι οι τυχαίες μεταβλητές $1-\frac{D_j}{r_j}$, είναι ανεξάρτητες προκύπτει

$$\hat{V}\left\{\sum_{j:t_j<t}\log\left(1-\frac{d_j}{r_j}\right)\right\} = \hat{V}\left(\sum_{j:t_j<t}\log\left(1-\hat{q}_j\right)\right) = \hat{V}\left(\sum_{j:t_j<t}\log\left(\hat{p}_j\right)\right) =$$
$$= \sum_{j:t_j<t}\hat{V}\left(\log\left(\hat{p}_j\right)\right).$$

Γνωρίζοντας ότι

$$\hat{V}\left(\log \hat{S}(t)\right) = \left(\frac{1}{E\left(\hat{S}(t)\right)}\right)^2 \hat{V}\left(\hat{S}(t)\right),$$

ανάλογα προκύπτει ότι

$$\hat{V}\left(\log\left(\hat{p}_j\right)\right) = \left(\frac{1}{p_j}\right)^2 \frac{p_j(1-p_j)}{r_j} =$$
$$= \frac{d_j}{r_j(r_j - d_j)}$$
$$\Rightarrow \log\left[\hat{V}\left(\hat{S}(t)\right)\right] = \sum_{j:t_j<t}\frac{d_j}{r_j(r_j-d_j)}.$$

Το αποτέλεσμα που λαμβάνεται τελικά είναι

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

$$\hat{V}\left(\hat{S}(t)\right) = \hat{V}\left\{\exp\left[\log\left(\hat{S}(t)\right)\right]\right\} =$$
$$= \exp^2\left[\log\left(\hat{S}(t)\right)\right] \sum_{j:t_j<t} \frac{d_j}{r_j(r_j - d_j)} =$$
$$= \hat{S}^2(t) \sum_{j:t_j<t} \frac{d_j}{r_j(r_j - d_j)}$$

ο οποίος είναι γνωστός ως ο **τύπος του Greenwood** και παρουσιάστηκε το 1926 από τον Greenwood (Greenwood (1926))**.**

Η εκτιμήτρια Kaplan-Meier, ως εκτιμήτρια μέγιστης πιθανοφάνειας της $S(t)$, αποδεικνύεται ότι ασυπτωτικά ακολουθεί την κανονική κατανομή. Επίσης είναι συνεπής και αμερόληπτη εκτιμήτρια. Ισχύει ότι

$$\hat{S}(t) \sim N\left(S(t), \hat{V}\left(\hat{S}(t)\right)\right),$$

όπου $\hat{V}\left(\hat{S}(t)\right)$ είναι οποιαδήποτε συνεπής εκτίμηση του $V\left(\hat{S}(t)\right)$. Χρησιμοποιώντας για την εκτίμηση της $V\left(\hat{S}(t)\right)$ τον τύπο του Greenwoodπροκύπτει το επόμενο ασυμπτωτικό διάστημα εμπιστοσύνης $(1-a)\%$ για την $S(t)$

$$\left[\hat{S}(t) - z_{a/2}\sqrt{\hat{V}\left(\hat{S}(t)\right)}, \hat{S}(t) + z_{a/2}\sqrt{\hat{V}\left(\hat{S}(t)\right)}\right].$$

Το τελευταίο αναφέρεται συνήθως ως **απλό**(*plain*) διάστημα εμπιστοσύνης για τη συνάρτηση επιβίωσης.

Στην ειδική περίπτωση όπου, είτε το κάτω όριο βρεθεί μικρότερο του μηδέν είτε το άνω όριο μεγαλύτερο της μονάδας, αντικαθίστανται με μηδέν και ένα αντίστοιχα.

Ένα άλλο διάστημα εμπιστοσύνης $(1-a)\%$ που δίνεται συνήθως έτοιμο από τα στατιστικά πακέτα αντί του απλού είναι το διάστημα εμπιστοσύνης για την ποσότητα $\log[-\log S[t]]$, δηλαδή το

$$\left[\hat{S}(t)^{\exp\left(\pm z_{a/2}\hat{S}_{\log(-\log\hat{S}(t)))}\right)}\right],$$

Το τελευταίο είναι γνωστό ως **διάστημα εμπιστοσύνης log-log.**

Τέλος υπάρχει και το διάστημα εμπιστοσύνης $(1-a)\%$ γνωστό ως **log,** το οποίο δίνεται από τον τύπο

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

$$\left[\hat{S}(t)\exp(\pm z_{a/2}\hat{S}_{\log\hat{S}(t)})\right].$$

Κλείνοντας την παράγραφο αυτή αναφέρουμε ότι, αν στο δείγμα περιέχονται δεδομένα λογοκριμένα από αριστερά, τότε είναι ευκολότερο να εκτιμηθεί η συνάρτηση κατανομής αντί της συνάρτησης επιβίωσης. Έτσι, χρησιμοποιώντας μία τεχνική αντίστοιχη με αυτήν που δημιούργησε την εκτιμήτρια Kaplan-Meier μπορούμε να γράψουμε

$$F(t_j) = P(T \le t_j) = P(T \le t_j, T \le t_{j+1}) = P(T \le t_j \mid T \le t_{j-1})P(T \le t_{j+1}) =$$
$$= P(T \le t_j \mid T \le t_{j+1})P(T \le t_{j+1})P(T \le t_{j+1} \mid T \le t_{j+2})...P(T \le t_n)$$

Ισχύει ότι $P(T \le t_n) = 1$, διότι η χρονική στιγμή $t_n$ είναι η μεγαλύτερη τιμή που καταγράφηκε. Η εκτιμήτρια του $F(t)$ παίρνει τη μορφή

$$\hat{F}(t) = \prod_{j:t_j>t} \frac{r_j - d_j}{r_j} = \prod_{j:t_j>t}\left[1 - \frac{d_j}{r_j}\right],$$

όπου τώρα $d_j$ είναι ο αριθμός των ατόμων που εισέρχονται στην έρευνα (αντίστοιχα στη λογοκρισία από δεξιά ήταν ο αριθμός των ατόμων που πέθαιναν) και $r_j$ είναι ο αριθμός των ατόμων στην έρευνα τη χρονική στιγμή $t_j$. Όταν ένα δεδομένο είναι λογοκριμένο από αριστερά, θα εισέλθει στην έρευνα τη χρονική στιγμή του αριστερού ορίου λογοκρισίας. Αντίστοιχα όταν ένα δεδομένο είναι πλήρες, τότε θα εισέλθει στην έρευνα τη χρονική στιγμή $t_j$ και θα καταμετρηθεί ως $d_j$.

Για τη διακύμανση του εκτιμητή, δύναται να τροποποιηθεί ο τύπος του Greenwood ως εξής (Greenwood (1926)):

$$\hat{V}\left(\hat{F}(t)\right) = \hat{F}^2(t)\sum_{j:t_j>t}\frac{d_j}{r_j(r_j - d_j)}.$$

### 5.2.5 Εκτιμήτρια Flemming-Harrington ή Nelson-Aalen

Με βάση τύπο των Kaplan-Meier μπορούμε να γράψουμε

$$\hat{S}(t) = \prod_{j:t_j<t}\frac{r_j - d_j}{r_j} = \prod_{j:t_j<t}\left(1 - \frac{d_j}{r_j}\right) \Rightarrow \log\hat{S}(t) = \sum_{j:t_j<t}\log\left(1 - \frac{d_j}{r_j}\right).$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Στην περίπτωση που

- ο αριθμός των ατόμων που βρίσκονται σε κίνδυνο (δηλαδή ο $r_j$) είναι μεγάλος, και
- ο αριθμός των διαστημάτων $I_j$ είναι σχετικά μεγάλος, με αποτέλεσμα ο αριθμός των ατόμων που πεθαίνουν σε καθένα από αυτά να είναι πολύ μικρός,

η παραπάνω σχέση διαφοροποιείται ακολούθως

$$-\log\left(1-\frac{d_j}{r_j}\right)\cong\frac{d_j}{r_j}\Rightarrow-\log\hat{S}(t)\cong\sum_{j:t_j<t}\frac{d_j}{r_j}$$

(έγινε χρήση της προσεγγιστικής σχέσης $-\log(1-x)\approx x$ που προκύπτει από το ανάπτυγμα $\log(1-x)=-x-\frac{x^2}{2}-\frac{x^3}{3}-\ldots$).

Συνεπώς μία εκτίμηση της συνάρτησης επιβίωσης δίνεται από τον τύπο

$$\hat{S}_{NA}(t)=\exp\left(\sum_{j:t_j<t}\frac{d_j}{r_j}\right),$$

ο οποίος είναι γνωστός ως **εκτιμήτρια Flemming-Harrington ή Nelson-Aalen** (Flemming and Harrington (1984)).

Η εκτιμήτρια Flemming-Harrington ακολουθεί ασυμπτωτικά την κανονική κατανομή

$$\hat{S}_{NA}(t)\sim N\left(S(t),\hat{V}\left(\hat{S}_{NA}(t)\right)\right)$$

Για την εκτίμηση της διακύμανσης $V\left(\hat{S}_{NA}(t)\right)$ χρησιμοποιούνται συνήθως οι παρακάτω τύποι

- $\hat{V}\left(\hat{S}_{NA}(t)\right)=\left[\hat{S}_{NA}(t)\right]^2\sum_{j:t_j<t}\frac{d_j}{r_j(r_j-d_j)}$, (τύπος του Greenwood).
- $\hat{V}\left(\hat{S}_{NA}(t)\right)=\left[\hat{S}_{NA}(t)\right]^2\sum_{j:t_j<t}\frac{d_j(r_j-d_j)}{r_j^3}$, (τύπος του Klein).

Ένα ασυμπτωτικό διάστημα εμπιστοσύνης $(1-a)\%$ για την ποσότητα $S_{NA}(t)$ θα είναι το

$$\left[\hat{S}_{NA}(t)\pm z_{a/2}\sqrt{\hat{V}\left(\hat{S}_{NA}(t)\right)}\right],$$

το οποίο καλείται **απλό** (*plain*).

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Χρησιμοποιώντας το Παράδειγμα της παραγράφου 5.2.3, υπολογίζεται, με τη βοήθεια και του τύπου του *Greenwood*, ένα 95% διαστήμα εμπιστοσύνης τύπου *log* για τη συνάρτηση επιβίωσης $S_{NA}(t)$ της ομάδας *Treatment*. Τα αποτελέσματα παρατίθενται στον Πίνακα 5.2.3

**Πίνακας 5.2.3**

**Διάστημα εμπιστοσύνης 95% τύπου *plain* με τον τύπο του *Greenwood* για τη συνάρτηση επιβίωσης**

| $t$ | $\hat{S}_{NA}(t)$ | $\sqrt{\hat{V}[\hat{S}_{NA}(t)]}$ | Lower limit | Upper limit |
|---|---|---|---|---|
| (6,7] | 0.8669 | 0.0772 | 0.7155 | 1.0182 |
| (7,10] | 0.8174 | 0.0881 | 0.6447 | 0.9900 |
| (10,13] | 0.7646 | 0.0978 | 0.5729 | 0.9564 |
| (13,16] | 0.7035 | 0.1089 | 0.4901 | 0.9169 |
| (16,22] | 0.6424 | 0.1168 | 0.4135 | 0.8712 |
| (22,23] | 0.5569 | 0.1328 | 0.2966 | 0.8171 |
| (23,35] | 0.4714 | 0.1416 | 0.1939 | 0.7488 |

## 5.2.6 Αλγόριθμος Turnbull

Μέχρι αυτό το σημείο έχουν αναφερθεί μέθοδοι υπολογισμού μη παραμετρικής εκτιμήτριας για πλήρη δεδομένα, δεξιά λογοκριμένα δεδομένα και αριστερά λογοκριμένα δεδομένα. Ιδιαίτερο ενδιαφέρον παρουσιάζει και η περίπτωση λογοκρισίας σε διάστημα (*interval-censored data*). Ο Turnbull δίνει ένα αλγόριθμο (Turnbull (1976)) που χρησιμοποιείται για τον υπολογισμό μη παραμετρικής εκτιμήτριας σε αυτή τη μορφή λογοκρισίας. Ο αλγόριθμος λειτουργεί με βάση τις αρχές του αλγόριθμου EM (A.P.Dempster (1977)). Υπενθυμίζουμε ότι, αν υποτεθεί ότι υπάρχουν μη πλήρη δεδομένα, και είναι επιθυμητή η εκτίμηση μιας παραμέτρου της κατανομής των δεδομένων, ο EM αλγόριθμος ξεκινά από μια αρχική τιμή για την παράμετρο και εκτιμά τις ελλιπείς τιμές με βάση αυτή. Εν συνεχεία υπολογίζει τις παραμέτρους που μεγιστοποιούν την πιθανοφάνεια, χρησιμοποιώντας στη θέση των ελλιπών τιμών τις εκτιμήτριες που υπολόγισε στο προηγούμενο βήμα. Τέλος εναλλάσσει αυτές τις δύο διαδικασίες (δηλαδή *expectation-maximization*) μέχρις ότου επιτευχθεί σύγκλιση σε κάποια υπόθεση.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

**Θα περιγράψουμε στη συνέχεια πως λειτουργεί ο αλγόριθμος Turnbull. Θεωρούμε**

αρχικά μία διαμέριση του χώρου $R_+$, μέσω των αριθμών $k_1, k_2, ..., k_{m-1}$ δημιουργώντας τα παρακάτω διαστήματα

$$J_1 = (o, k_1], \ J_2 = (k_1, k_2], \ldots, \ J_{m-1} = (k_{m-2}, k_{m-1}], \ J_m = (k_{m-1}, \infty).$$

Έστω $s = (s_1, s_2, ..., s_m)$ ένα $m$-διάνυσμα με $s_i > 0$, $i = 1,2,...,n$ και $\sum_{i=1}^{m} s_i = 1$.

Το $s$ αντιπροσωπεύει το βήμα της συνάρτησης κατανομής $\hat{F}$ μέσα σε κάθε διαστήμα. Θα πρέπει να εκτιμηθεί το $s$ ώστε να ληφθεί στη συνέχεια μία μη παραμετρική εκτιμήτρια της συνάρτησης επιβίωσης. Η τελευταία δίνεται μέσα από την σχέση:

$$\hat{S}(t) = 1 - \prod_{\{i \in K\}} \sum_{\{i \le t\}} s_i \ , t > 0.$$

Χρησιμοποιώντας τα $n$ δεδομένα που έχουμε στη διάθεσή μας, κατασκευάζουμε έναν πίνακα Α $a$ διαστάσεων $n \times m$ ως εξής:

$$a_{ij} = \begin{cases} 1 \text{ αν το γεγονός } i \text{ μπορούσε να συμβεί κατά την διάρκεια του διαστήματος } J_j \\ 0, \text{ αν το γεγονός } i \text{ δεν μπορούσε να συμβεί κατά την διάρκεια του διαστήματος } J_j, \end{cases}$$

όπου $i = 1,2,...,n$ και $j = 1,2,...,m$.

Παράλληλα ορίζεται και η δείκτρια τυχαία μεταβλητή

$$I_{ij} = \begin{cases} 1, \text{ αν το γεγονός } i \text{ συμβαίνει κατά την διάρκεια του διαστήματος } J_j \\ 0, \text{ αν το γεγονός } i \text{ δεν συμβαίνει κατά την διάρκεια του διαστήματος } J_j \end{cases}$$

όπου $i = 1,2,...,n$ και $j = 1,2,...,m$.

Εξαιτίας της παρουσίας των λογοκριμένων δεδομένων στο δείγμα, είναι αδύνατο να είναι γνωστός ο πίνακας $I$ για όλες τις τιμές του $i$ και $j$.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Συνεπώς χρησιμοποιείται η αναμενόμενη τιμή του με την υπόθεση ότι το $s$ είναι γνωστό, πιο συγκεκριμένα

$$m_{ij} = E\left[I_{ij}\right]\frac{a_{ij}s_j}{\sum_{k=1}^{m} a_{ik}s_k},\ i=1,2,...,n\ ,\ j=1,2,...,m\,.$$

Εδώ το $m_{ij}$ αντιπροσωπεύει την πιθανότητα το γεγονός $i$ να συμβεί στο διάστημα $J_j$ υποθέτοντας ότι ο χρόνος ζωής έχει την κατανομή που καθορίζει το $s$.

Αν υποτεθεί ότι τα $m_{ij}$ είναι παρατηρηθείσες τιμές (πλήρεις), τότε το ποσοστό των δεδομένων στο διάστημα $J_j$ δίνεται από τον τύπο

$$p_j(s) = \left(\frac{1}{n}\right)\sum_{i=1}^{n} m_{ij},\ j=1,2,...,m.$$

Με βάση όλα τα παραπάνω μπορεί ο αλγόριθμος Turnbull να παράγει εκτιμήσεις για το $s$ μέχρις ότου συγκλίνει σε κάποιο κριτήριο.

**Βήμα 1ο**

Από ένα σύνολο δεδομένων, δημιουργείται η διαμέριση του χώρου $R_+$ και κατασκευάζεται και ένα πίνακας $a$ διαστάσεων $n\times m$ με στοιχεία $a_{ij}$ όπως περιγράφηκε.

**Βήμα 2ο**

Λαμβάνεται το διάνυσμα $s$ διάστασης $m$ που αντιπροσωπεύει το βήμα της συνάρτησης επιβίωσης στο διάστημα $J_j$ για $j=1,2,...,m.$

**Βήμα 3ο**

Για $k=0$, θέτουμε $s^0 = \left(\frac{1}{m},\frac{1}{m},...,\frac{1}{m}\right).$

**Βήμα 4ο**

Υπολογίζονται αναδρομικά οι τιμές του $s_j^k$ με χρήση του τύπου

$$s_j^{k+1} = p_j\left(s^k\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{a_{ij}s_j^k}{\sum_{l=1}^{m} a_i s_l^k},\ k=k+1 \text{ και } j=1,2,...,m$$

και η διαδικασία ολοκληρώνεται όταν ικανοποιείται τουλάχιστον ένα από τα παρακάτω κριτήρια

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

$$\sum_{j=1}^{m}\left(s_j^{k+1}-s_j^{k}\right)^2 < c \text{ ή } \max_{j=1,2,\ldots,n}\left|s_j^{k+1}-s_j^{k}\right| < c \text{ για } c>0.$$

## 5.3 ΑΠΑΡΑΜΕΤΡΙΚΑ ΜΟΝΤΕΛΑ

Όπως ήδη έχει αναφερθεί, τα μη παραμετρικά μοντέλα δεν κάνουν καμία υπόθεση για τον πληθυσμό από τον οποίο προέρχεται το δείγμα και σπάνια στηρίζονται σε τεχνικές μέγιστης πιθανοφάνειας. Τα μονοδιάστατα μοντέλα κατασκευάστηκαν για να καθορίζουν τα διάφορα χαρακτηριστικά του πληθυσμού από τον οποίο προέκυψε το δείγμα με λογοκριμένα δεδομένα και να συγκρίνουν αυτά τα χαρακτηριστικά για δύο ή περισσότερα λογοκριμένα δείγματα. Τα δισδιάστατα μοντέλα κατασκευάστηκαν για να μετρούν, εάν μία μεταβλητή που παρουσιάζει ελλιπείς τιμές σχετίζεται με κάποια άλλη μεταβλητή του πληθυσμού. Στην περίπτωση που συμβαίνει κάτι τέτοιο προσδιορίζουν ένα μοντέλο παλινδρόμησης που να εκφράζει τη σχέση μεταξύ των δύο μεταβλητών. Σημειώνουμε ότι υπάρχουν επίσης και αντίστοιχες πολυμεταβλητές μέθοδοι για λογοκριμένα δεδομένα, ωστόσο είναι ελάχιστες και δεν είναι λειτουργικές τέτοιες τεχνικές.

Δύο μέθοδοι είναι από τη φύση τους μονοδιάστατες. Η πρώτη, η οποία αναπτύχθηκε εκτενώς, είναι η εκτιμήτρια Kaplan-Meier η οποία έκανε την εμφάνισή της το 1958. Η δεύτερη μέθοδος παρέχει μία πληθώρα τρόπων ελέγχου για το αν δύο λογοκριμένα δείγματα προέρχονται από τον ίδιο πληθυσμό. Αποτελούν κυρίως γενίκευση απλών τεστ για λογοκριμένα δείγματα, όπως το τεστ του Wilcoxon και Logrank για δύο δείγματα. Η διαφορά τους έγκειται στα βάρη που χρησιμοποιούνται (*weights*) για τα λογοκριμένα δεδομένα, όταν προχωράμε στον υπολογισμ της στατιστικής συνάρτησης. Έτσι, το καθένα παρουσιάζει διαφορετική ευαισθησία στο ένα ή στο άλλο άκρο της κατανομής. Για παράδειγμα τα τεστ των Gehan και Logrank χρησιμοποιούνται ευρέως στην βιομετρία, ενώ κάποια άλλα τεστ δεν χρησιμοποιούνται.

Υπάρχουν πολλές διαφορετικές δισδιάστατες μέθοδοι, που περιλαμβάνουν δύο τεστ συσχέτισης και τρία μοντέλα γραμμικής

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

παλινδρόμησης. Το μοντέλο κινδύνου του Cox (*Cox hazard model*) το οποίο είναι τεστ συσχέτισης, Ο ΕΜ αλγόριθμος και η μέθοδος Buckley-James (μοντέλο γραμμικής παλινδρόμησης) μπορούν να επεξεργαστούν πολλές ανεξάρτητες μεταβλητές εφόσον η εξαρτημένη μεταβλητή αποτελείται από ένα είδος λογοκριμένων δεδομένων (λογοκρισία από δεξιά ή αριστερά). Το γενικευμένο Kendall's tau (τεστ συσχέτισης) και η γραμμική παλινδρόμηση του Schmitt μπορούν να επεξεργαστούν εξαρτημένες μεταβλητές με ταυτόχρονη παρουσία διαφορετικών ειδών λογοκριμένων δεδομένων (συμπεριλαμβανομένων λογοκριμένων δεδομένων και στην ανεξάρτητη μεταβλητή), υπό την προυπόθεση ότι η ανεξάρτητη τυχαία μεταβλητή είναι μόνο μία.

Οι περισσότερες από τις παραπάνω μέθοδους αναλύονται εκτενώς από τους Schmitt (1985), Feigelson and Nelson (1985) και Isobe (1986).

Εντούτοις, πέρα από τα πλεονεκτήματα που παρουσιάζουν όταν χρησιμοποιούνται για την επίλυση προβλημάτων με λογοκριμένα δεδομένα, εμφανίζουν και διάφορα μειονεκτήματα τα οποία συνοψίζονται στη συνέχεια

- Η εκτιμήτρια Kaplan-Meier λειτουργεί με οποιαδήποτε κατανομή, εφόσον όμως η λογοκρισία είναι τυχαία.
- Το τεστ ελέγχου για δύο λογοκριμένα δείγματα απλά συγκρίνει τα δύο αυτά δείγματα χωρίς να αναζητά και να καθορίζει την σωστή κατανομή του πληθυσμού από τον οποίο προέρχονται.
- Το τεστ του Gehan υποθέτει ότι τα χαρακτηριστικά της λογοκρισίας των δύο δείγματων είναι τα ίδια αλλά δεν απαιτεί να είναι τυχαία.
- Το γενικευμένο Kendall's tau και το μοντέλο κινδύνου του Cox μερικές φορές δε δίνουν ακριβή αποτελέσματα.
- Υπάρχουν περιορισμοί στα μοντέλα παλινδρόμησης (*Emmehod* και *Schmitt*). Πιο συγκεκριμένα το μοντέλο παλινδρόμησης του Schmitt λειτουργεί όταν υπάρχει λογοκρισία και στις δύο μεταβλητές (εξαρτημένη και ανεξάρτητη), ενώ το μοντέλο του

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Emmehod απαιτεί τα σφάλματα να ακολουθούν την κανονική κατανομή.

- Η χρήση του μοντέλου του Schmitt είναι προβληματική όταν τα λογοκριμένα δεδομένα στο δείγμα είναι πάρα πολλά.

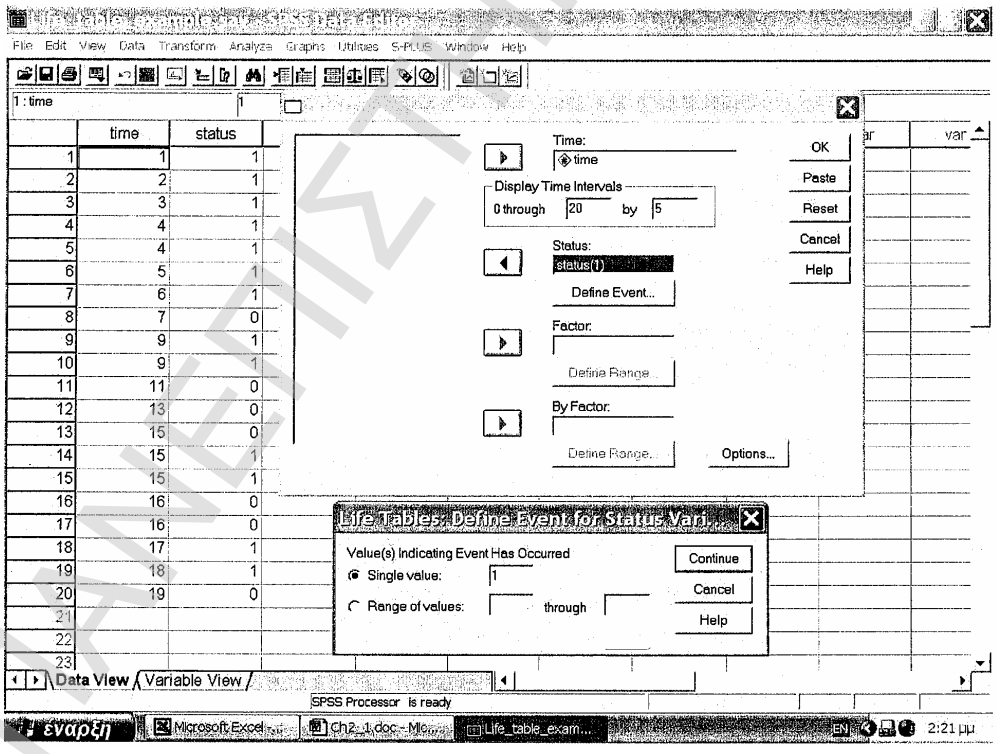
## 5.4 ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΜΕ ΧΡΗΣΗ ΣΤΑΤΙΣΤΙΚΩΝ ΠΑΚΕΤΩΝ

**Στατιστικό πακέτο SPSS**

Το στατιστικό πακέτο SPSS μπορεί να χρησιμοποιηθεί για την περιληπτική παρουσίαση δεδομένων χρόνων ζωής σε μορφή πινάκων επιβίωσης. Πιο συγκεκριμένα μετά την εισαγωγή των δεδομένων επιλέγεται *Analyse>Survival>Life Tables* και στο μενού που εμφανίζεται, δίνονται διάφορες επιλογές όπως φαίνεται στο κάτωθι σχήμα

**Πίνακας 5.4.1**

Πίνακες επιβίωσης με χρήση SPSS




ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Παράλληλα παρέχει τη δυνατότητα εφαρμογής της μεθόδου Kaplan-Meier και της παλινδρόμησης του Cox, με πολύ καλό εύρος αποτελεσμάτων. Αντιθέτως δεν υποστηρίζει παραμετρική παλινδρόμηση.

**Στατιστικό πακέτο SPLUS**

Αποτελεί την καλύτερη επιλογή για στατιστική ανάλυση λογοκριμένων δεδομένων, καθώς συνδυάζει τις δυνατότητες του SPSS και του Minitab. Παρόλο που οι επιλογές από το μενού είναι περιορισμένες, όταν εργάζεται κανείς στο *command mode* πρακτικά δεν υπαρχουν περιορισμοί.

Στο παράδειγμα της Παραγράφου 5.2.3, με τη βοήθεια του στατιστικού πακέτου SPLUS, μπορεί να υπολογιστεί η εκτιμήτρια Kaplan-Meier τρέχοντας το ακόλουθο *script file*.

```
attach(Remission)
rem <-survfit(Surv(TIME,STATUS)~GROUP,conf.type="none")
summary(rem)
plot(rem,xlab="Time in weeks",ylab="Survival", main="Remission Data",lty=2:1)
legend(25,0.9,c("Treatment","Placebo"),lty=2:1)
```

προκύπτει η ανάλυση και για τα 42 άτομα του παραδείγματος, το οποίο περιέχει τις μεταβλητές *TIME* (που υποδηλώνει το χρόνο ζωής), *STATUS* (που φανερώνει αν ο χρόνος είναι λογοκριμένος (*STATUS*=0) ή όχι (*STATUS*=1) καθώς και τη μεταβλητή *GROUP* η οποία για την ομάδα *TREATMENT* είναι 0 και για τη *PLACEBO* είναι 1.

GROUP=0

| time | n.risk | n.event | survival | std.err |
|---|---|---|---|---|
| 6 | 21 | 3 | 0.857 | 0.0764 |
| 7 | 17 | 1 | 0.807 | 0.0869 |
| 10 | 15 | 1 | 0.753 | 0.0963 |
| 13 | 12 | 1 | 0.690 | 0.1068 |
| 16 | 11 | 1 | 0.627 | 0.1141 |
| 22 | 7 | 1 | 0.538 | 0.1282 |
| 23 | 6 | 1 | 0.448 | 0.1346 |

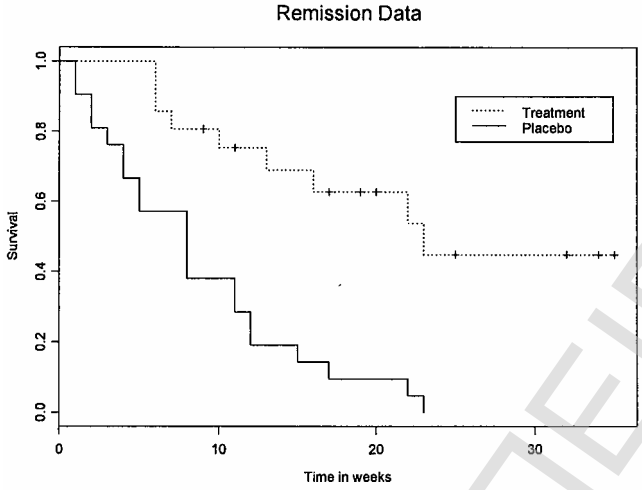
GROUP=1

| time | n.risk | n.event | survival | std.err |
|---|---|---|---|---|
| 1 | 21 | 2 | 0.9048 | 0.0641 |
| 2 | 19 | 2 | 0.8095 | 0.0857 |
| 3 | 17 | 1 | 0.7619 | 0.0929 |
| 4 | 16 | 2 | 0.6667 | 0.1029 |
| 5 | 14 | 2 | 0.5714 | 0.1080 |
| 8 | 12 | 4 | 0.3810 | 0.1060 |
| 11 | 8 | 2 | 0.2857 | 0.0986 |
| 12 | 6 | 2 | 0.1905 | 0.0857 |
| 15 | 4 | 1 | 0.1429 | 0.0764 |
| 17 | 3 | 1 | 0.0952 | 0.0641 |
| 22 | 2 | 1 | 0.0476 | 0.0465 |
| 23 | 1 | 1 | 0.0000 | NA |

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

**Σχήμα 5.4.1**

Εκτιμήτρια Kaplan-Meier




Στη γραφική παράσταση της εκτιμήτριας Kaplan-Meier για την ομάδα *Treatment* με + σημειώνονται οι λογοκριμένοι χρόνοι ζωής. Συμπεραίνεται ότι το φάρμακο 6-MP συγκρινόμενο με το *Placebo* έχει την ιδιότητα να καθυστερεί την επανεμφάνιση της νόσου.

Όσον αφορά τα διαστήματα εμπιστοσύνης, συνηθίζεται η παρουσίαση διαστημάτων τύπου *log-log* από το πακέτο STATA, τύπου *plain* από το πακέτο SAS και τύπου *log* από το SPLUS (το οποίο όμως δίνει τη δυνατότητα δημιουργίας και των υπολοίπων διαστημάτων), όπως θα παρουσιαστεί παρακάτω. Σημειώνεται ότι το πακέτο SPSS δε δίνει διαστήματα εμπιστοσύνης.

Με χρήση του SPLUS, τρέχοντας τα ακόλουθα *script files* προκύπτουν τα αποτελέσματα που ακολουθούν.

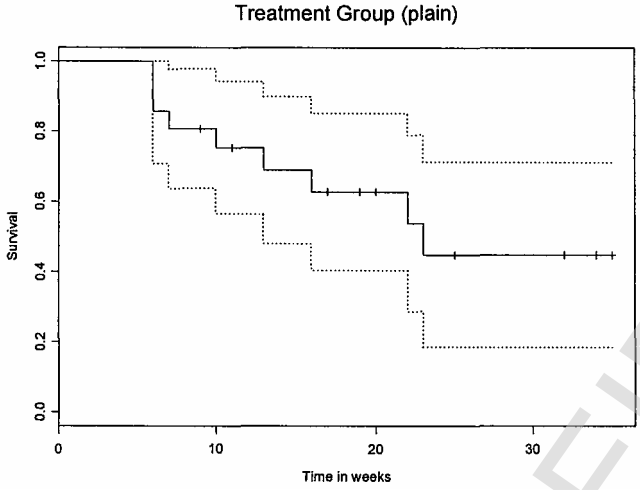
**Διάστημα εμπιστοσύνης τύπου *plain* για την ομάδα *Treatment***

```
attach(Rem0)
rem0 <-survfit(Surv(TIME,STATUS),conf.int=0.95, conf.type="plain")
summary(rem0)
plot(rem0,xlab="Time in weeks",ylab="Survival", main="Treatment Group (plain)")
```

| time | n.risk | n.event | survival | std.err | lower 95% CI | upper 95% CI |
|---|---|---|---|---|---|---|
| 6 | 21 | 3 | 0.857 | 0.0764 | 0.707 | 1.000 |
| 7 | 17 | 1 | 0.807 | 0.0869 | 0.636 | 0.977 |
| 10 | 15 | 1 | 0.753 | 0.0963 | 0.564 | 0.942 |
| 13 | 12 | 1 | 0.690 | 0.1068 | 0.481 | 0.900 |
| 16 | 11 | 1 | 0.627 | 0.1141 | 0.404 | 0.851 |
| 22 | 7 | 1 | 0.538 | 0.1282 | 0.286 | 0.789 |
| 23 | 6 | 1 | 0.448 | 0.1346 | 0.184 | 0.712 |

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

## Σχήμα 5.4.2

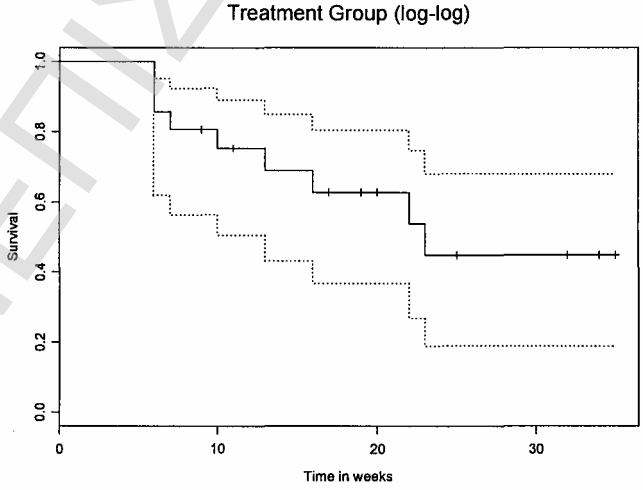



**Διάστημα εμπιστοσύνης τύπου *log-log* για την ομάδα *Treatment***

```
attach(Rem0)
rem0 <-survfit(Surv(TIME,STATUS),conf.int=0.95, conf.type="log-log")
summary(rem0)
plot(rem0,xlab="Time in weeks",ylab="Survival", main="Treatment Group (log-log)")
```

```
 time n.risk n.event survival std.err lower 95% CI   upper 95% CI
    6     21       3    0.857  0.0764        0.620         0.952
    7     17       1    0.807  0.0869        0.563         0.923
   10     15       1    0.753  0.0963        0.503         0.889
   13     12       1    0.690  0.1068        0.432         0.849
   16     11       1    0.627  0.1141        0.368         0.805
   22      7       1    0.538  0.1282        0.268         0.747
   23      6       1    0.448  0.1346        0.188         0.680
```

## Σχήμα 5.4.3




ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

**Διάστημα εμπιστοσύνης τύπου *log* για την ομάδα *Treatment***

```
attach(Rem0)
rem0 <-survfit(Surv(TIME,STATUS),conf.int=0.95, conf.type="log")
summary(rem0)
plot(rem0,xlab="Time in weeks",ylab="Survival", main="Treatment Group (log)")
```

| time | n.risk | n.event | survival | std.err | lower 95% CI | upper 95% CI |
|---|---|---|---|---|---|---|
| 6 | 21 | 3 | 0.857 | 0.0764 | 0.720 | 1.000 |
| 7 | 17 | 1 | 0.807 | 0.0869 | 0.653 | 0.996 |
| 10 | 15 | 1 | 0.753 | 0.0963 | 0.586 | 0.968 |
| 13 | 12 | 1 | 0.690 | 0.1068 | 0.510 | 0.935 |
| 16 | 11 | 1 | 0.627 | 0.1141 | 0.439 | 0.896 |
| 22 | 7 | 1 | 0.538 | 0.1282 | 0.337 | 0.858 |
| 23 | 6 | 1 | 0.448 | 0.1346 | 0.249 | 0.807 |

Διαφορετικά θα μπορούσε να χρησιμοποιήσει κανείς τη διαδικασία *Statistics>Survival>Nonparametric Survival*, στο πεδίο *Model* να σημειώσει ως *Curve type*:Kaplan-Meier, και στο πεδίο *Formula* να συμπληρώσει *Surv*(*TIME, STATUS*). Τέλος στο πεδίο *Options* θα επιλεγεί ο συντελεστής εμπιστοσύνης (*Confidence Interval Level*), ως*Standard Error Formula* ο τύπος του *Greenwood* και *Confidence Interval type*: *plain, log, lol-log*.

Η εκτιμήτρια Flemming-Harrington καθώς και τα διατήματα εμπιστοσύνης τύπου *plain* και *log*, μπορούν επίσης να κατασκευαστούν από το στατιστικό πακέτο SPLUS, με τη βοήθεια του τύπου του *Greenwood*, χρησιμοποιώντας το ακόλουθο *script file*:

- **95% διάστημα εμπιστοσύνης τύπου *plain* με τον τύπο του *Greenwood***

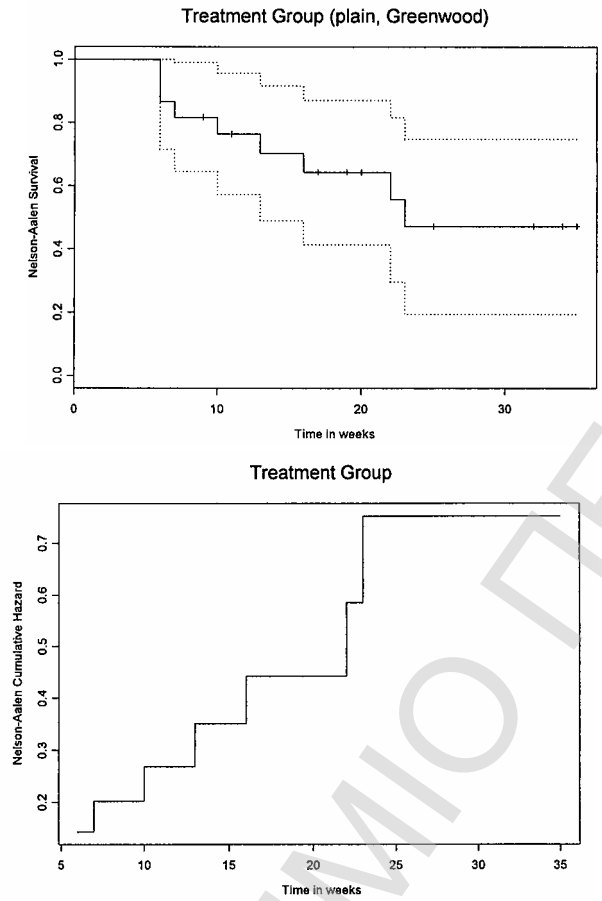
```
attach(Rem0)
nelson <-survfit(Surv(TIME,STATUS),conf.int=0.95, conf.type="plain",
type="fleming-harrington", error="greenwood")
summary(nelson)
plot(nelson,xlab="Time in weeks",ylab="Nelson-Aalen Survival", main="Treatment Group
    (plain, Greenwood)")
H <- -log(nelson$surv)
t <- nelson$time
plot(t,H,type="s",xlab="Time in weeks",ylab="Nelson-Aalen Cumulative Hazard",
   main="Treatment Group")
```

Έτσι προκύπτουν τα επόμενα αποτελέσματα

| time | n.risk | n.event | survival | std.err | lower 95% CI | upper 95% CI |
|---|---|---|---|---|---|---|
| 6 | 21 | 3 | 0.867 | 0.0772 | 0.716 | 1.000 |
| 7 | 17 | 1 | 0.817 | 0.0881 | 0.645 | 0.990 |
| 10 | 15 | 1 | 0.765 | 0.0978 | 0.573 | 0.956 |
| 13 | 12 | 1 | 0.704 | 0.1089 | 0.490 | 0.917 |
| 16 | 11 | 1 | 0.642 | 0.1168 | 0.414 | 0.871 |
| 22 | 7 | 1 | 0.557 | 0.1328 | 0.297 | 0.817 |
| 23 | 6 | 1 | 0.471 | 0.1416 | 0.194 | 0.749 |

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

**Σχήμα 5.4.4**




Τέλος όσον αφορά την παλινδρόμηση με παραμετρικά μοντέλα, μπορεί να υλοποιηθεί μέσα από την επιλογή *Statistics>Survival>Parametric Survival*.

**Στατιστικό πακέτο SAS**

Το στατιστικό πακέτο SAS αν και δεν είναι εύχρηστο είναι αρκετά σταθερό και προσφέρει εξαιρετικές δυνατότητες ώστε ο χρήστης μπορεί να ασχοληθεί με οποιαδήποτε πτυχή της ανάλυσης επιβίωσης.

**Στατιστικό πακέτο MINITAB**

Το πακέτο MINITAB έχει πολλές δυνατότητες για παραμετρική ανάλυση παλινδρόμησης με λογοκριμένα δεδομένα για την κατανομή Weibull, την εκθετική, την κανονική και την λογαριθμοκανονική. Παρ'όλα αυτά δεν υποστηρίζει την παλινδρόμηση του Cox κάτι το οποίο αποτελεί βασικό μειονέκτημα του πακέτου.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Το στατιστικό πακέτο STSDAS είναι από τα πλέον ισχυρά στην αντιμετώπιση προβλημάτων με λογοκριμένα δεδομένα και προτάθηκε από τους Feigelson και Isobe στο Πανεπιστήμιο της Πενσυλβάνια. Διάφορες άλλες ρουτίνες για την αντιμετώπιση λογοκριμένων δεδομένων έχουν προταθεί από τους Wagner και Meeker (1985). Πολλές από αυτές έχουν ενσωματωθεί στα στατιστικά πακέτα όπως το SAS, SPSS, BMDP και GLIM. Τέλος ενδιαφέροντα θεωρούνται και τα πακέτα CENSOR, DASH, LIMDEP, STATPAC, SURVAN, SURVREG.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

# Περίληψη

Οι ελλιπείς τιμές αποτελούν ένα σημαντικό πρόβλημα κατά την ανάλυση πραγματικών δεδομένων, διότι συχνά οδηγούν σε εκτιμήτριες οι οποίες δεν έχουν την ιδιότητα της αμεροληψίας ενώ και τα αντίστοιχα συμπεράσματα δεν είναι τόσο ακριβή όσο στην περίπτωση πλήρων δεδομένων. Οι στατιστικές μέθοδοι χειρισμού ελλιπών δεδομένων περιλαμβάνουν την αντικατάσταση, την χρήση εκτιμητριών μέγιστης πιθανοφάνειας και τη μέθοδο βαρών. Κάθε μέθοδος λειτουργεί διαφορετικά λαμβάνοντας υπόψη τον εκάστοτε μηχανισμό ελλιπών τιμών και το είδος των μεταβλητών (συνεχείς, διακριτές). Τα τελευταία χρόνια είναι ευρέως διαδεδομένα μοντέλα παλινδρόμησης για την εκτίμηση των ελλιπών τιμών, ενώ και τα σύγχρονα στατιστικά πακέτα έχουν συμπεριλάβει στις βιβλιοθήκες τους ρουτίνες για το χειρισμό τους.

Η ανάλυση επιβίωσης είναι ιδιαίτερα σημαντική στη βιοστατιστική, αναλογισμό και σε πολλά άλλα πεδία. Όταν χρησιμοποιούμε μεθόδους ανάλυσης επιβίωσης αρκετά συχνά εμφανίζονται λογοκριμένα δεδομένα. Τότε είναι απαραίτητο να αναγνωρίζουμε τη μορφή λογοκρισίας που εμφανίζεται στο δείγμα μας, και να εφαρμόζουμε την κατάλληλη μέθοδο για την εκτίμηση της συνάρτησης επιβίωσης και της βαθμίδας αποτυχίας. Στην εργασία αυτή παρουσιάζουμε τις σε μεθόδους εκτίμησης των Kaplan-Meier, Flemming-Harrington, τον αλγόριθμο Turnbull και τις εκτιμήτριες μέγιστης πιθανοφάνειας. Τέλος έχουμε καταγράψει πως μπορούν να χρησιμοποιηθούν τα συνήθη στατιστικά πακέτα για την ανάλυση τόσο ελλιπών όσο και λογοκριμένων δεδομένων.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

# Abstract

Missing data are a major problem when real data are to be analyzed, because they may cause bias to the estimators used or lead to inefficient analysis. Statistical methods addressing missingness have been actively pursued in the recent years, and include imputation, likelihood and weighting approaches. The approach becomes more complicated when many patterns of missing values are present, or when both categorical and continuous random variables are involved. Implementations of routines to incorporate observations with incomplete variables in regression models are nowdays widely available. The most popular statistical packages have by now incorporated several routines to handle missing values.

Survival Analysis is becoming increasingly important in biostatistics, actuarial science and many other scientific fields. When resorting to survival analysis methods, it is important to identify the type of censoring, so that the appropriate method can be used in estimation of the survival and hazard function. In the present MSc, we present in some detail Kaplan-Meier estimator, the Flemming-Harrington estimator, the Turnbull algorithm and maximum likelihood estimators. Finally we will provide details on how the most popular statistical packages can be exploited for the analysis of missing or truncated data.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

# ΒΙΒΛΙΟΓΡΑΦΙΑ


Berkson, J., and Gage, R. P. (1950). Calculation of Survival Rates for Canser, *Proc Staff Meet Mayo* Clin, **24**, (11), 270-286.

Bishop, Y. M. M., Fienberg, S.E., and Holland, P. W. (1975). *Discrete Multivariate Analyses:Theory and Practice*,Cambridge, MA:MIT Press.

Carpenter, J. R., Kenward, M. G. and Vansteelandt, S. (2006). A Comparison of Multiple Imputation and Doubly Robust Estimation for Analyses with Missing Data, *Journal of the Royal Statistic Society*, Series A, **169**, 571-584.

Collins, L. M., Schafer, J. L., and Kam, C. M. (2001). A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures, *Psychological Methods*, **6**, 330-351.

Cox, D. R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society*, **B 34**, 1-34.

Dempster, A. P., Laird, N. M., and Rubin, D. B., (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society Series B*, **39**,1–38.

Diggle, P., and Kenward, M. G. (1994). Informative Drop-Out in Longitudinal Data Analysis, *Applied Statistics*, **43**, 49-73.

Eltinge, and Yansaneh, (1997). Diagnostics for Formation of Nonresponse Adjustment Cells With an Application to Income Nonresponse in U.S. Consumer Expenditure Survey, *Canada's National Statistics Agency*, **23**, 54-68.

Flemming, T. R., and Harrington, D. P. (1984). Nonparametric Estimation of the Survival Distribution in Censored Data, *Communication in Statistics-Theory and Methods*, **13**, 2469-2456.

Freireich, E. O. (1963). The effect of 6-mercaptopmine on the duration of steroid included remission in acute leukemia, *Blood*, **21**, 699-716.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. B. (1995). *Bayesian Data Analysis*, Chapman & Hall, New York .

Greenland, S. and Finkle, W. D. (1995). A Critical look at Methods for Handling Missing Covariates in Epidemiologic regression Analyses, *American Journal of Epidemiology*, **142**, 1255-1264.

Greenwood, M. (1926). The natural duration of cancer, *Reports on Public health and Medical Subjects*, **33**, 45-60.

Haitovsky, Y. (1968). Missing data in regression analysis, *Journal of the Royal Statistical Society B*, **30**, 67–81.

Haziza, D., and Nobrega, K. SSC-2002, Hamilton.

Hedeker, D., Gibbons, D. (1997). Application of Random-Effects Pattern-Mixture, Models for Missing Data in Logitudinal Studies, *The American Psychological Association Inc*, **2**, 64-78.

Horton, N. J., and Kleinman, K. P. (2007). A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models, *The American Statistician*, **61**, 79-90.

Horton, N. J., and Lipsitz, S. R. (1997). Review of Software to Fit Generalized Estimating Equation (GEE) Regression Models, *The American Statistician*, **53**, 160-169.


ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Hosmer, D. W., and Lemeshow, S. (1999). *Applied Survival Analysis*, Wiley Interscience.

Jones, M. P. (1996). Indicatior and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression, *Journal of the American Statistical Association*, **91**, 222-230.

Kaplan, E. L., Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**, 190-221.

Klein, J. P., Moeschberger, M. L. (1997). *Survival Analysis*, Springer.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.

Little, R. J. A.(1993). Pattern-mixture models for multivariate incomplete data, *Journal of the American Statistical Association*, **88**, 125–134.

Little, R. J. A. (1995). Modeling the drop-out mechanism in longitudinal studies, *Journal of the American Statistical Association*, **90**, 1112–1121.

Little, R. J. A., and Rubin, D.B. (2002). *Statistical Analysis With Missing Data*, 2nd ed..Wiley, NewYork.

McLachlan, G. J., and Krishnan, T. (1997). *The EM Algorithm and Extensions*, Wiley, New York.

Morgenthaler, S. (1997). *Introduction a la statistique*.

Nelson, W. B. (2004). *Applied Life Data Analysis*, Wiley, New York.

Olkin, I., and Tate, R. F. (1961). Multivariate Correlation Models With Mixed Discrete and Continuous Variables, *The Annuals of Mathematical Statistics*, **32**, 448-465.

Oostenbrink, R., Moons, K.G.M., Bleeker, S.E., Moll, H.A., and Grobbee, D.E. (2003). Diagnostic research on routine care data: prospects and problems, *J Clin. Epidemiol*., **56**, 501–506.

Polak, E. (1997). *Optimization*, Springer.

Quarteroni, A. (2000). *Methode numeriques pour le calcul scienifique*, Springer.

Raghunathan, T. E., Lepkowski, J. M., van Hocwyk, J., and Solenberger, P. (2001). A multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, *Survey Methodology*, **27**, 85-95.

Rao, and Shao, J. (1992), Jackknife Variance Estimation With Survey Data Under Hot Deck Imputation, *Biometrika*, **60**, 34-40.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data, *Journal of the American Statistical Association*, **90**, 106-121.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, **70**, 41 –55.

Rubin, D. B. (1976). Inference and Missing Data, *Biometrika*, **63**, 581-590.

Samartzis, L. (2006). *Survival and censored data*, Ecole polytechnique Federale de Lausanne.

Sarndal (1992). Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, Vol **87**, 376-382.

Schafer, J.L. (1997). *Analysis of incomplete multivariate data*, Chapman & Hall/CRC Press, London.

Shao, J., Sitter, R. R. (1996). Bootstrap for Imputed Survey Data, *Journal of the American Statistical Association*, **91**, 150-163.

Shao, J., Steel, P. (1999). Variance Estmation for Survey Data With Composite Imputation and Nonnegligible Sampling Fractions, *Journal of the American Statistical Association*, **94**, 198-206.

PANEPISTIMIO PIERAIA

Sheldon M. Ross. (1994). *Initiation aux propabilites.*

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association*, **82**, 528–550.

Tanner, M. A., (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*, Springer-Verlag, New York.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of theRoyal Statistical Society*, **60**, 123-140.

Vach, W. (1994). *Logistic regression with missing values in the covariates*, Springer, New York.

van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis, *Statistics in Medicine*, **18**, 681-694.

van Buuren, S., Brand, J. P. L., Groothius-Oudshoom, C. G. M., and Rubin, D. B. (2006). Fully Conditional Specification in Multivariate Imputation, *Journal of Statistical Computation and Simulation*, **76**, 1049-1064.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ