



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ
ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΠΑΡΑΜΕΤΡΙΚΕΣ ΚΑΙ ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ
ΜΕΘΟΔΟΙ ΑΝΑΛΥΣΗΣ ΣΧΕΔΙΑΣΜΩΝ ΜΕ
ΑΠΟΚΟΜΜΕΝΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ

Διπλωματική εργασία

ΒΟΝΤΑ ΒΑΣΙΛΙΚΗ

Επιβλέπων: Κουκουβίνος Χρήστος, Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2013

ΠΕΡΙΛΗΨΗ

Η Στατιστική μπορεί να οριστεί ως η επιστήμη που ασχολείται με τη συγκέντρωση, παρουσίαση, αξιολόγηση και την εν συνεχεία επεξεργασία της πληροφορίας. Είναι φανερό ότι ο κλάδος της στατιστικής, ήδη από τα τέλη του 18^{ου} αιώνα, εξαπλώνεται και βρίσκει εφαρμογές και σε άλλους επιστημονικούς κλάδους όπως η αστρονομία, η βιολογία, οι κοινωνικές επιστήμες κ.α. Στη βιομηχανία μέσω της στατιστικής ο πειραματιστής μπορεί να εξετάσει τους παράγοντες εκείνους που κάνουν το παραγόμενο προϊόν ανθεκτικότερο. Επίσης, στη βιοϊατρική μέσω ερευνητικών διαδικασιών εξετάζεται η εξέλιξη μιας ασθένειας, καθώς και ποιοι παράγοντες την επηρεάζουν. Η ανάλυση των ερευνητικών δεδομένων μπορεί να οδηγήσει τόσο στην πρόληψη της, όσο και στην καταπολέμηση της.

Η παρούσα εργασία εστιάζει, κυρίως, στην ανάλυση επιβίωσης ή αξιοπιστίας, τον κλάδο, εκείνο, της στατιστικής που μελετά προβλήματα των οποίων τα πειραματικά ή ερευνητικά δεδομένα μπορεί να αφορούν τη χρονική στιγμή που προκαλείται βλάβη σε ένα μηχανικό σύστημα ή θάνατος σε ένα βιολογικό οργανισμό.

Πιο συγκεκριμένα, στο πρώτο κεφάλαιο γίνεται μια εισαγωγική αναφορά σε σχετικές με την ανάλυση αξιοπιστίας έννοιες. Επίσης, περιγράφονται τα είδη πειραματικών δεδομένων και οι κατανομές που συχνά προσαρμόζονται σε αυτά. Στο δεύτερο κεφάλαιο παρατίθενται τα βασικά είδη παραγοντικών σχεδιασμών, οι οποίοι χρησιμεύουν στην οργάνωση πειραματικών δεδομένων. Με τον όρο παραγοντικό σχεδιασμό, εννοείται ότι σε κάθε πλήρη δοκιμή ή επανάληψη του πειράματος, εξετάζονται όλοι οι δυνατοί συνδυασμοί των επιπέδων των παραγόντων. Εν συνεχεία, στο τρίτο κεφάλαιο περιγράφονται πλήρως παραμετρικές μέθοδοι ανάλυσης δεδομένων που εμπεριέχουν εκτός από πλήρεις και αποκομμένες παρατηρήσεις, ενώ δίνονται σχετικά παραδείγματα. Τέλος, στο τέταρτο και τελευταίο κεφάλαιο της εργασίας αυτής, παρουσιάζονται μη παραμετρικές μέθοδοι ανάλυσης δεδομένων που εμπεριέχουν επίσης αποκομμένες παρατηρήσεις. Οι μέθοδοι, αυτοί, καλούνται μη παραμετρικές, καθώς δεν απαιτείται καμία υπόθεση για την ακριβή μορφή της κατανομής του δείγματος.

ABSTRACT

Statistics can be defined as the science that deals with the collection, presentation, evaluation and the further process of information. It is obvious that the principle of statistics, since the 18th century, has been spread and applied in other principles such as astronomy, biology, social sciences etc. In industry through statistics the experimenter can examine those factors that make the final product more sustainable. Also, in biomedical the statistics through a lot of experiments can monitor the causes of a disease, and the real factors that affect it. The analysis of the found data can lead to the prevention of the disease, as well as to the therapy of it.

The present paper, focuses mainly on survival analysis or reliability, the branch of statistics which considers problems whose experimental or research data may relate to the exact moment of damaging a mechanical system or causing the death of a biological organism.

Specifically, in the first chapter a reference introduction is mentioned on the aspect of survival analysis. Furthermore, someone can find a description of the types of experimental data and the distributions which can be adjusted to them. The second chapter deals with the basic types of factorial designs, which can be used in order the experimental data to be organized. The term of factorial design, means that in every test or a full repetition of the experiment, are examined all the possible combinations of different levels of factors. The third chapter mentions the most important parametric methods of data analysis which include full observations and censored. Finally, in the last chapter of this dissertation, non-parametric methods of data analysis, which include, also, censored data are presented. These methods, are called non-parametric, because no assumption for the precise form of the distribution sample is needed.

ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά θα ήθελα να ευχαριστήσω θερμά τον Καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου κ. Χρήστο Κουκουβίνο για την επίβλεψη και την καθοδήγηση του κατά τη διάρκεια εκπόνησης αυτής της διπλωματικής εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα Ανδρουλάκη Εμμανουήλ για την πολύτιμη βοήθεια, στήριξη και το συνεχές ενδιαφέρον που έδειξε.

Επιπλέον, θα ήταν παράλειψη μου να μην ευχαριστήσω την φίλη μου και συμφοιτήτρια μου Αράπη Κλεοπάτρα για τη συνεχή παρότρυνση της και τις εύστοχες παρατηρήσεις της.

Τέλος, οφείλω ένα μεγάλο ευχαριστώ στην οικογένεια μου για τα εφόδια που μου προσέφερε όλα αυτά τα χρόνια, την αμέριστη συμπαράσταση τους και ενθάρρυνση τους σε κάθε βήμα μου.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ.....	i
ABSTRACT.....	ii
ΕΥΧΑΡΙΣΤΙΕΣ.....	iii

ΚΕΦΑΛΑΙΟ 1 : ΕΙΣΑΓΩΓΗ **1**

1.1 Εισαγωγικές έννοιες.....	1
1.2 Είδη αποκομμένων δεδομένων.....	2
1.3 Είδη αποκοπής παρατηρήσεων.....	5
1.4 Βασικά στοιχεία.....	6
1.4.1 Συνάρτηση κατανομής (distribution function).....	6
1.4.2 Συνάρτηση αξιοπιστίας ή επιβίωσης (survivor function).....	6
1.4.3 Διάμεσος (median).....	7
1.4.4 Συνάρτηση πυκνότητας πιθανότητας (probability density function).....	7
1.4.5 Συνάρτηση διακινδύνευσης (hazard function).....	9
1.4.6 Σωρευτική συνάρτηση διακινδύνευσης (integrated or cumulative hazard).....	9
1.5 Μοντέλα διάρκειας ζωής	10
1.5.1 Εκθετική κατανομή (Exponential distribution).....	10
1.5.2 Η κατανομή Weibull (Weibull distribution).....	11
1.5.3 Η κατανομή Γάμμα (Gamma distribution).....	13
1.5.4 Η Λογαριθμο-κανονική κατανομή (Lognormal distribution).....	15
1.5.5 Η λογαριθμο-λογιστική κατανομή (Log-logistic distribution).....	18
1.6 Αριθμητικό παράδειγμα.....	20

ΚΕΦΑΛΑΙΟ 2: ΠΑΡΑΓΟΝΤΙΚΟΙ ΣΧΕΔΙΑΣΜΟΙ 23

2.1 Εισαγωγή.....	23
2.1.1 Βασικές έννοιες και ορισμοί.....	23
2.1.2 Σχεδιασμός και υλοποίηση πειραμάτων.....	25
2.1.3 Ιστορική Αναδρομή.....	27
2.2 Είδη παραγοντικών σχεδιασμών.....	29
2.2.1 Ο 2^k παραγοντικός σχεδιασμός.....	29
2.2.1.1 Ο 2^2 παραγοντικός σχεδιασμός.....	30
2.2.1.2 Ένα παράδειγμα 2^4 παραγοντικού σχεδιασμού.....	31
2.2.2 Ο 2^k κλασματικός παραγοντικός σχεδιασμός.....	35
2.2.2.1 Το ένα δεύτερο κλάσμα του 2^k παραγοντικού σχεδιασμού...36	
2.2.3 Αναλυτική τάξη σχεδιασμού (design resolution).....	38
2.2.4 Ο 3^k παραγοντικός σχεδιασμός (full factorial design).....	40
2.2.4.1 Ένα παράδειγμα 3^4 παραγοντικού σχεδιασμού.....	41
2.2.5 Σχεδιασμός επαναλαμβανόμενων μετρήσεων (repeated measures design).....	43

ΚΕΦΑΛΑΙΟ 3: ΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΠΑΡΑΓΟΝΤΙΚΩΝ ΣΧΕΔΙΑΣΜΩΝ 45

3.1 Η μέθοδος ελαχίστων τετραγώνων.....	45
3.2 Η μέθοδος μέγιστης πιθανοφάνειας.....	48
3.2.1 Η μέθοδος μέγιστης πιθανοφάνειας για δεδομένα που περιέχουν αποκομμένες παρατηρήσεις.....	49
3.2.2 Κριτήρια σύγκρισης μοντέλων.....	55
3.3 Μέθοδος ανάλυσης παραγοντικών σχεδιασμών με αποκομμένες παρατηρήσεις.....	58
3.3.1 Αριθμητικό παράδειγμα.....	61
3.3.1.1 Περιγραφή του προβλήματος.....	61
3.3.1.2 Ανάλυση δεδομένων.....	63
3.3.1.3 Σύγκριση του αποτελέσματος της ανάλυσης του πειράματος με πλήρη δεδομένα.....	66

ΚΕΦΑΛΑΙΟ 4: ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ ΑΝΑΛΥΣΗΣ ΠΑΡΑΓΟΝΤΙΚΩΝ ΣΧΕΔΙΑΣΜΩΝ

67

4.1 Η εκτιμήτρια Kaplan-Meier.....	67
4.1.1 Αριθμητικό παράδειγμα.....	68
4.2 Ο μη παραμετρικός έλεγχος log-rank.....	71
4.3 Η εκτιμήτρια Nelson-Aalen.....	73
4.4 Μη παραμετρικές μέθοδοι για την ανάλυση παραγοντικών σχεδιασμών με αποκομμένες παρατηρήσεις.....	74
4.4.1 Μη παραμετρική μέθοδος ανάλυσης παραγοντικών σχεδιασμών με αποκομμένες παρατηρήσεις των Lee – Ing Tong και Chao – Ton Su.....	74
4.4.1.1 Αριθμητικό παράδειγμα.....	77
4.4.1.2 Συμπεράσματα.....	82
4.4.2 Μη παραμετρική μέθοδος ανάλυσης παραγοντικών σχεδιασμών με αποκομμένες παρατηρήσεις των Michael G. Akritas και Edgar Brunner...83	83
4.4.2.1 Ο στατιστικός έλεγχος.....	85
4.4.2.2 Προσέγγιση μικρού δείγματος.....	86
4.4.2.3 Αριθμητικό παράδειγμα και ανάλυση δεδομένων.....	89
4.4.3 Μη παραμετρική μέθοδος ανάλυσης παραγοντικών σχεδιασμών με αποκομμένες παρατηρήσεις των John T. O’Gorman και Michael G. Akritas.....	90
4.4.3.1 Αριθμητικό παράδειγμα.....	93
4.4.3.2 Αριθμητικό παράδειγμα.....	95
4.4.3.3 Συμπεράσματα.....	96
BIBΛΙΟΓΡΑΦΙΑ.....	I

ΚΕΦΑΛΑΙΟ 1:

ΕΙΣΑΓΩΓΗ

1.1 Εισαγωγικές έννοιες

Τόσο στην βιομηχανία, όσο και στην βιοϊατρική συχνά πραγματοποιούνται πειράματα στα οποία καταγράφεται ο χρόνος μέχρι ότου συμβεί ένα γεγονός. Ένα γεγονός μπορεί να είναι τελικό, όπως για παράδειγμα ο θάνατος, που μπορεί να συμβεί μόνο μια φορά ή μη τελικό όπως η εμφάνιση μιας αρρώστιας, μια μηχανολογική βλάβη ή ακόμα η επιτυχής ανάρρωση ενός ασθενούς, που μπορεί να συμβεί πολλές φορές. Πρακτικά το γεγονός σχετίζεται με την ποιοτική αλλαγή της κατάστασης ενός ατομικού προσώπου ή αντικειμένου. Ο χρόνος από την αρχή της έρευνας μέχρι να συμβεί το γεγονός που καταγράφουμε μπορεί να μετρείται σε ώρες, μέρες, μήνες ή ακόμα και σε χρόνια. Εναλλακτικά ο χρόνος μπορεί να σχετίζεται με την ηλικία του εμπλεκόμενου όταν συμβεί το γεγονός. Επίσης σε πολλές περιπτώσεις, κυρίως σε βιομηχανικά πειράματα, η διάρκεια λειτουργίας της μονάδας μπορεί να μην μετρείται σε μονάδες χρόνου αλλά σε χιλιόμετρα, στροφές, κ.α.. Ακόμα και σε αυτή την περίπτωση, όμως, η απόκριση θεωρείται συμβατικά χρόνος ζωής. Όλοι αυτοί οι χρόνοι μέχρι να συμβεί ένα γεγονός αναφέρονται ως *δεδομένα επιβίωσης ή δεδομένα διάρκειας ζωής*.

Ο κλάδος της στατιστικής που σαν σκοπό έχει την ανάλυση δεδομένων επιβίωσης ονομάζεται *ανάλυση αξιοπιστίας (reliability analysis)*, όταν πρόκειται για εφαρμογές θετικών επιστημών, ή *ανάλυση επιβίωσης (survival analysis)*, όταν πρόκειται για βιοϊατρικές εφαρμογές. Κάνοντας χρήση των μεθόδων της ανάλυσης επιβίωσης, ο πειραματιστής αποσκοπεί στην πρόβλεψη της επιβίωσης ή της μέσης διάρκειας ζωής των πληθυσμών που ερευνώνται, καθώς και στην σύγκριση τους με άλλους πειραματικούς πληθυσμούς με σκοπό την εύρεση κινδύνων ή/και χρήσιμων προγνωστικών παραγόντων.

Ο λόγος που αναπτύχθηκε ο συγκεκριμένος κλάδος της στατιστικής έγκειται στο ότι τα δεδομένα διάρκειας ζωής δεν μπορούν να αναλυθούν μέσω των τυπικών στατιστικών μεθόδων. Μία από τις ιδιαιτερότητες τους είναι ότι γενικότερα τα δεδομένα επιβίωσης δεν είναι συμμετρικά κατανομημένα. Συνήθως,

ένα ιστόγραμμα που κατασκευάζεται από τους χρόνους επιβίωσης μιας ομάδας παρόμοιων μονάδων τείνει να είναι *θετικά λοξό*, δηλαδή, παρατηρείται μια μακρύτερη «ουρά» στα δεξιά του διαστήματος που περιλαμβάνει τον κύριο όγκο των παρατηρήσεων. Συνεπώς, τα δεδομένα διάρκειας ζωής δεν μπορούν να θεωρηθούν κανονικά κατανομημένα. Το πρόβλημα, αυτό, μπορεί να αντιμετωπιστεί μετασχηματίζοντας τα δεδομένα, παραδείγματος χάριν λογαριθμίζοντας τα, έτσι ώστε να προκύψει μια πιο συμμετρική κατανομή. Ωστόσο, μια πιο ικανοποιητική προσέγγιση είναι η υιοθέτηση ενός εναλλακτικού μοντέλου για τα αρχικά δεδομένα.

Το κύριο χαρακτηριστικό των δεδομένων διάρκειας ζωής που καθιστά ακατάλληλη την ανάλυση τους μέσω τυπικών διαδικασιών είναι ότι συχνά συναντάται το φαινόμενο των *αποκομμένων παρατηρήσεων (censored data)*. Αποκομμένα ονομάζονται τα δεδομένα για τα οποία δεν έχει παρατηρηθεί πειραματικά ο χρόνος επιβίωσης, η στιγμή, δηλαδή, που συνέβη το γεγονός. Τα δεδομένα που δεν είναι αποκομμένα ονομάζονται *πλήρη*. Η έννοια των αποκομμένων δεδομένων χρησιμοποιήθηκε για πρώτη φορά από τον Hald (1949). Στην πλειονότητα των περιπτώσεων τα αποκομμένα δεδομένα είναι αποτέλεσμα ατελούς παρατήρησης της εξέλιξης κάποιων ασθενών ή ακόμα και βίαιης διακοπή του πειράματος σε ανύποπτη χρονική στιγμή.

1.2 Είδη αποκομμένων δεδομένων

Τα είδη των αποκομμένων παρατηρήσεων είναι τα εξής:
(Collett, 2003)

- **Δεξιά αποκομμένες παρατηρήσεις (*right-censored data*) :**

Σε αυτή την κατηγορία ανήκουν οι παρατηρήσεις που η ακριβής χρονική στιγμή που συνέβη το γεγονός είναι άγνωστη, είναι γνωστό όμως ότι είναι μεγαλύτερη ή ίση από την χρονική στιγμή t_0 . Η χρονική στιγμή t_0 ονομάζεται αποκομμένος χρόνος επιβίωσης και είναι μικρότερος από τον πραγματικό χρόνο επιβίωσης.

Παράδειγμα 1.2.1

Σε ένα πείραμα εξετάζεται η διάρκεια ζωής κάποιων αμορτισέρ οχημάτων. Όπως είναι φυσικό οι μονάδες μέτρησης της διάρκειας ζωής στη συγκεκριμένη περίπτωση είναι χιλιόμετρα λειτουργίας και όχι χρόνος λειτουργίας. Έστω ότι όταν πραγματοποιήθηκε η συλλογή των δεδομένων κάποια αμορτισέρ συνέχιζαν να λειτουργούν, τότε δεν είναι γνωστός ο ακριβής χρόνος επιβίωσης αλλά είναι γνωστό ότι είναι μεγαλύτερος από το χρόνο της τελευταίας καταγραφής.

- ***Αριστερά αποκομμένες παρατηρήσεις (left-censored data) :***

Σε αυτή την κατηγορία ανήκουν οι παρατηρήσεις που ο πραγματικός χρόνος επιβίωσης είναι άγνωστος και μικρότερος από τον παρατηρούμενο. Η από αριστερά αποκοπή παρουσιάζεται αρκετά σπάνια σε εφαρμογές.

Παράδειγμα 1.2.2

Για την καλύτερη κατανόηση της παραπάνω έννοιας ας θεωρήσουμε μία έρευνα κατά την οποία συμμετέχουν καρκινοπαθείς που έχουν υποβληθεί σε χειρουργική επέμβαση αφαίρεσης του αρχικού όγκου. Μετά από τρεις μήνες οι ασθενείς εξετάζονται για τυχόν επανεμφάνιση του καρκίνου. Τη συγκεκριμένη χρονική στιγμή διαπιστώνεται ότι μερικοί ασθενείς έχουν υποτροπιάσει. Για τους συγκεκριμένους ασθενείς ο πραγματικός χρόνος υποτροπιασμού είναι μικρότερος από τρεις μήνες και ανήκει στην κατηγορία των αριστερά αποκομμένων παρατηρήσεων.

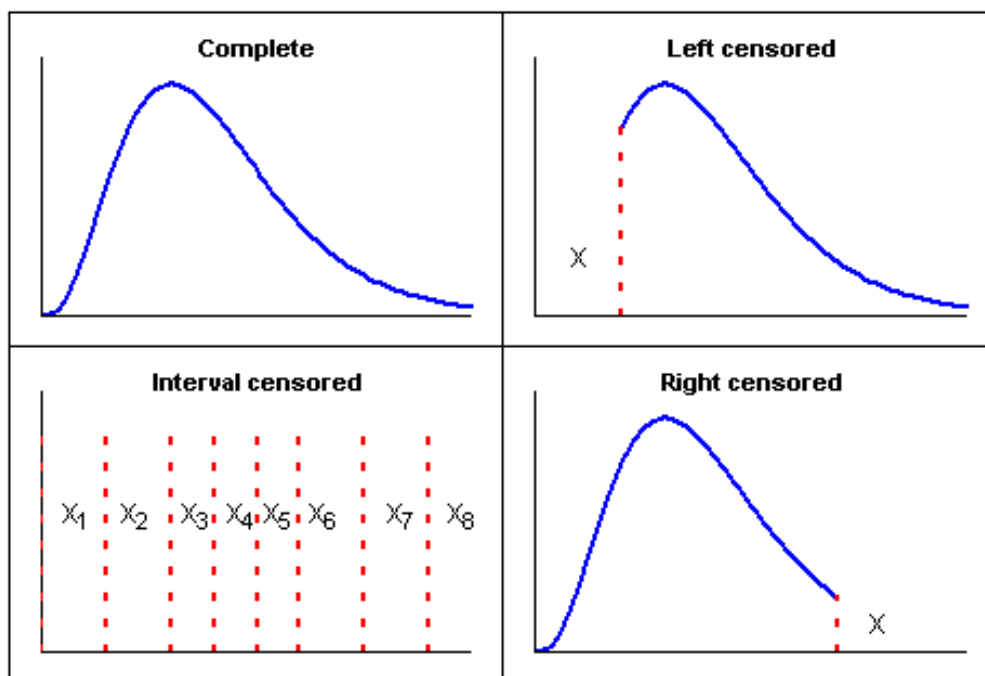
- ***Αποκομμένες παρατηρήσεις σε διάστημα (interval censored):***

Στην συγκεκριμένη περίπτωση η ακριβής χρονική στιγμή που συνέβη το γεγονός δεν είναι καταγεγραμμένη, αλλά είναι γνωστό ότι είναι μεγαλύτερη από τη χρονική στιγμή t_1 και μικρότερη από τη χρονική στιγμή t_2 .

Παράδειγμα 1.2.3

Ας θεωρήσουμε πάλι την περίπτωση όπου μελετάται η πορεία της υγείας καρκινοπαθών που έχουν υποβληθεί σε χειρουργική επέμβαση. Εάν κατά την εξέταση που πραγματοποιήθηκε ένα τρίμηνο μετά την εγχείρηση δεν παρατηρηθεί ο υποτροπιάσμος κάποιου ασθενή, αλλά έξι μήνες μετά σε μια δεύτερη εξέταση

διαπιστωθεί επανεμφάνιση του καρκίνου, η ακριβής χρονική στιγμή υποτροπιασμού είναι μεταξύ τριών και έξι μηνών. Τότε, η χρονική στιγμή που συνέβη το γεγονός λέγεται ότι είναι αποκομμένη σε διάστημα.



Γράφημα 1.1

Τα παραπάνω τρία είδη αποκομμένων δεδομένων προϋποθέτουν η αποκοπή να πραγματοποιείται τυχαία και ανήκουν στο είδος της μη πληροφοριακής αποκοπής (*uninformative censoring*). Γενικά, εμφανίζονται σε πλήρως ελεγχόμενα πειράματα, όπως σε εφαρμογές θετικών επιστημών, όπου η αποκοπή δε σχετίζεται με την μετέπειτα διάρκεια ζωής της μονάδας.

- **Πληροφοριακή αποκοπή (*informative censoring*):**

Στο παράδειγμα 1.2.1 όπου εξετάζεται η διάρκεια ζωής κάποιων αμορτισέρ οχημάτων, έστω ότι αποσύρεται μια μονάδα όταν αρχίζει να παρουσιάζει κάποιες δυσλειτουργίες. Έστω t_0 η χρονική στιγμή απόσυρσης. Τότε, η πραγματική χρονική στιγμή βλάβης του αμορτισέρ (T) θα είναι μεγαλύτερη από τη χρονική στιγμή t_0 και επιπλέον θα ισχύει $T = t_0 + \delta$, όπου δ πολύ μικρό. Η συγκεκριμένη περίπτωση καλείται πληροφοριακή αποκοπή (*informative censoring*).

Συγκεκριμένα, σε εφαρμογές που στηρίζονται στον ανθρώπινο παράγοντα, η αποκοπή μπορεί να προκληθεί από την αυτόβουλη αποχώρηση του ασθενή από το πείραμα, η οποία μπορεί φαινομενικά να θεωρείται τυχαία, αλλά στην πραγματικότητα συχνά σχετίζεται με την επιδείνωση της υγείας του. Τέτοια μορφή αποκοπής αποτελεί η πληροφοριακή αποκοπή. Στην συγκεκριμένη περίπτωση οι συνήθεις μέθοδοι ανάλυσης μπορούν να οδηγήσουν σε λανθασμένα αποτελέσματα.

1.3 Είδη αποκοπής παρατηρήσεων

Παρακάτω δίνονται οι τρεις βασικοί μηχανισμοί αποκοπής παρατηρήσεων (Καρώνη 2009):

- **Αποκοπή τύπου I:**

Το χρονικό διάστημα c παρακολούθησης των μονάδων είναι προκαθορισμένο και γνωστό. Γνωρίζουμε τον ακριβή χρόνο επιβίωσης αν $T_i < c$, σε διαφορετική περίπτωση γνωρίζουμε ότι η διάρκεια ζωής έχει υπερβεί το χρόνο c .

Είναι δυνατόν, επίσης, να έχει προκαθοριστεί ο χρόνος παρακολούθησης για κάθε μία μονάδα ξεχωριστά. Δηλαδή να γνωρίζουμε τα c_1, c_2, \dots, c_n . Βέβαια ενδέχεται κάποια c_i να είναι ίσα μεταξύ τους.

Στην συγκεκριμένη περίπτωση, οι χρόνοι c_i είναι δεδομένοι ενώ ο αριθμός των μονάδων που καταστρέφονται είναι τυχαίος.

- **Αποκοπή τύπου II:**

Διακόπτεται η παρακολούθηση του πειράματος όταν καταστραφούν k μονάδες. Εδώ το πλήθος k είναι προκαθορισμένο, ενώ η διάρκεια παρακολούθησης είναι τυχαία.

- **Αποκοπή τύπου III:**

Η συγκεκριμένη περίπτωση αποκοπής παρατηρήσεων εμφανίζεται κυρίως σε μελέτες ιατρικής φύσεως. Στην αποκοπή τύπου III η χρονική διάρκεια της μελέτης είναι προκαθορισμένη αλλά η στιγμή εισαγωγής ενός ασθενή στην μελέτη διαφέρει.

1.4 Βασικά στοιχεία

Για τη μελέτη των κυριότερων κατανομών που μπορούν να προσαρμοστούν στα δεδομένα επιβίωσης είναι χρήσιμο να ορίσουμε κάποιες βασικές έννοιες.

1.4.1 Συνάρτηση κατανομής (distribution function)

Έστω T η διάρκεια ζωής συνεχής τυχαίας μεταβλητής. Η συνάρτηση κατανομής ορίζεται ως εξής:

$$F(t) = P[T \leq t]$$

Δηλαδή εκφράζει την πιθανότητα η τιμή της παρατήρησης που μελετάται να είναι μικρότερη από κάποια δοσμένη τιμή t .

Εξ' ορισμού η $F(t)$ είναι αύξουσα και ισχύει

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

1.4.2 Συνάρτηση αξιοπιστίας ή επιβίωσης (survivor function)

Η συνάρτηση επιβίωσης ορίζεται ως εξής :

$$S(t) = P[T \geq t]$$

Εκφράζει δηλαδή την πιθανότητα η τιμή της παρατήρησης που μελετάται να είναι μεγαλύτερη ή ίση από κάποια δοσμένη τιμή t .

Από τον ορισμό της συνάρτησης αξιοπιστίας φαίνεται ότι συνδέεται με την συνάρτηση κατανομής από τον τύπο $S(t) = 1 - F(t)$.

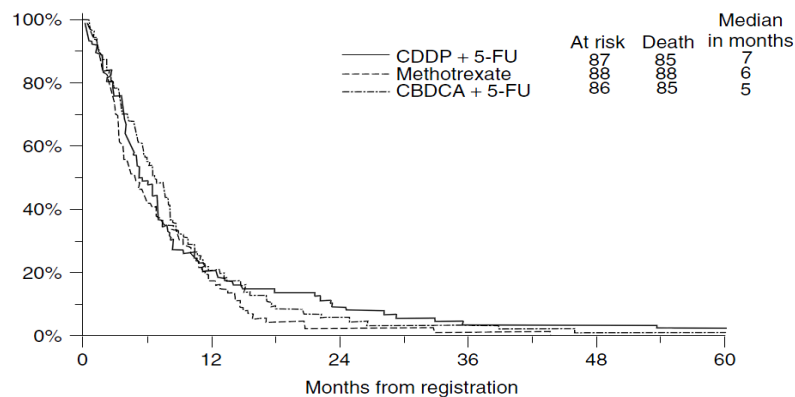
Έπεται, λοιπόν, για μία φθίνουσα συνάρτηση για την οποία ισχύει

$$\lim_{t \rightarrow 0} S(t) = 1 \text{ και } \lim_{t \rightarrow \infty} S(t) = 0.$$

Η γραφική παράσταση της $S(t)$ συναρτήσεως του χρόνου t , ονομάζεται καμπύλη επιβίωσης (survival curve), ορισμός που δόθηκε για πρώτη φορά από τον Joseph Berkson (1899 – 1982) το 1942.

Στο γράφημα 1.2 βλέπουμε ένα παράδειγμα καμπυλών επιβίωσης ατόμων που πάσχουν από καρκίνο και ακολουθούν τρεις διαφορετικές θεραπείες.

(Crowley J. and Ankerst D.P., 2006)



Γράφημα 1.2

1.4.3 Διάμεσος (median)

Η καμπύλη επιβίωσης τέμνεται με την καμπύλη της συνάρτησης κατανομής όταν $F(t_{0.5})=S(t_{0.5})=0.5$. Το σημείο $t_{0.5}$ καλείται διάμεσος και είναι σύνηθες μέτρο περιγραφής της μέσης διάρκειας ζωής.

1.4.4 Συνάρτηση πυκνότητας πιθανότητας (probability density function)

Η συνάρτηση πυκνότητας πιθανότητας (σ.π.π) , $f(t)$, της συνεχής μεταβλητής T ορίζεται από την σχέση:

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}.$$

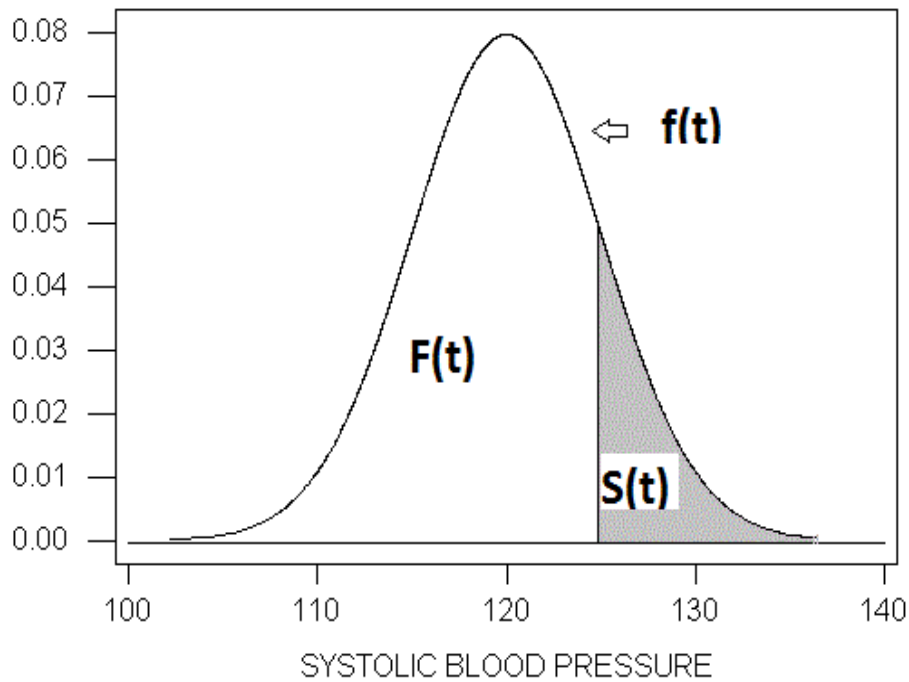
Επισημαίνεται ότι τόσο η συνάρτηση κατανομής, όσο και η συνάρτηση αξιοπιστίας μπορούν να οριστούν μέσω της σ.π.π ως εξής:

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

και $S(t) = P[t \geq T] = \int_t^\infty f(u)du.$

Συνεπώς, ισχύει ότι $\int_0^\infty f(u)du = 1.$

Στο Γράφημα 1.3 βλέπουμε τη συνάρτηση πυκνότητας πιθανότητας μιας θεωρητικής κατανομής που περιγράφει την αρτηριακή πίεση μιας ομάδας φοιτητών.



Γράφημα 1.3

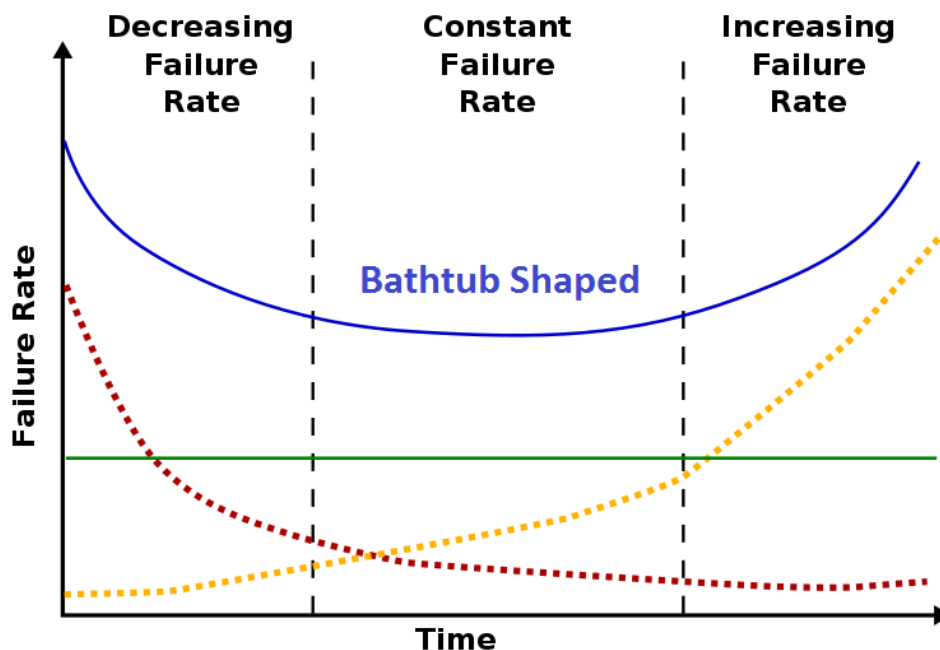
1.4.5 Συνάρτηση διακινδύνευσης (hazard function)

Η συνάρτηση διακινδύνευσης, $h(t)$, εκφράζει την τάση προς διακοπή ενός αντικειμένου στο χρονικό διάστημα $(t, t+\delta t]$ με δεδομένη την επιβίωση του μέχρι τη χρονική στιγμή t .

Συνεπώς ορίζεται μέσω του ακόλουθου τύπου:

$$h(t) = \lim_{\delta t \rightarrow 0} P[t < T \leq t + \delta t | T > t] = \lim_{\delta t \rightarrow 0} \frac{(S(t) - S(t + \delta t))/S(t)}{\delta t} = \frac{f(t)}{S(t)}.$$

Κατ' επέκταση η συνάρτηση διακινδύνευσης $h(t)$ εκφράζει το στιγμιαίο ρυθμό διακοπής (*failure rate*), ενώ η $h(t)\delta t$ είναι η υπο συνθήκη πιθανότητα της επικείμενης διακοπής μιας παρατήρησης δοθέντος ότι επέζησε μέχρι τη συγκεκριμένη χρονική στιγμή t .



Γράφημα 1.4

Όπως παρατηρείται στο Γράφημα 1.4 η συνάρτηση διακινδύνευσης μπορεί να είναι σταθερή, αύξουσα ή φθίνουσα. Επιπλέον, υπάρχει η δυνατότητα να συνδυάζει και τα τρία αυτά χαρακτηριστικά. Στην συγκεκριμένη περίπτωση αρχικά μειώνεται ο στιγμιαίος ρυθμός διακοπής καθώς αυξάνεται η ηλικία (*decreasing failure rate*), στη συνέχεια ακολουθεί μια φάση σταθερότητας (*constant failure rate*) και τέλος, παρατηρείται μία συνεχής αύξηση του ρυθμού διακοπής (*increasing failure rate*). Το σχήμα αυτής της καμπύλης ονομάζεται «διακινδύνευση της μπανιέρας» (*bathtub shaped*). Αν και σε αρκετές περιπτώσεις έχει την πιο ρεαλιστική συμπεριφορά, εντούτοις παρουσιάζονται δυσκολίες στην μοντελοποίηση του.

1.4.6 Σωρευτική συνάρτηση διακινδύνευσης (integrated or cumulative hazard)

Η σωρευτική συνάρτηση διακινδύνευσης (σ.σ.δ.) ορίζεται ως

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

Η σ.σ.δ. είναι ιδιαίτερα χρήσιμη για την επιλογή ενός κατάλληλου στατιστικού μοντέλου κατά την ανάλυση δεδομένων.

Από τους παραπάνω ορισμούς προκύπτει ότι:

$$H(t) = \int_0^t \frac{f(u)}{S(u)} du = \int_0^t \frac{-\frac{dS(u)}{du}}{S(u)} du = [-\ln S(u)]_0^t = -\ln S(t).$$

Συνεπώς, $S(t) = \exp\{-H(t)\}$.

Παρατηρούμε ότι από τους παραπάνω ορισμούς και σχέσεις οι συναρτήσεις $h(t), f(t), S(t), F(t), H(t)$ είναι μαθηματικά ισοδύναμες καθώς αν γνωρίζουμε την μία, μπορούμε να βρούμε και τις υπόλοιπες.

1.5 Μοντέλα διάρκειας ζωής

Στην παρούσα παράγραφο παρουσιάζονται τα κυριότερα μοντέλα διάρκειας ζωής, δηλαδή, κατανομές που είναι ιδιαίτερα χρήσιμες σε πρακτικές εφαρμογές καθώς συχνά προσαρμόζονται ικανοποιητικά στα δεδομένα επιβίωσης.

1.5.1 Εκθετική κατανομή (exponential distribution)

Η συνάρτηση πυκνότητας πιθανότητας της Εκθετικής κατανομής μια τυχαίας μεταβλητής T με παράμετρο ρυθμού λ είναι

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

Σύμφωνα με τις σχέσεις που αναφέρθηκαν στην προηγούμενη παράγραφο η συνάρτηση αξιοπιστίας δίνεται από τη σχέση:

$$S(t) = \int_t^{\infty} f(u) du = e^{-\lambda t},$$

ενώ η συνάρτηση διακινδύνευσης δίνεται από τη σχέση:

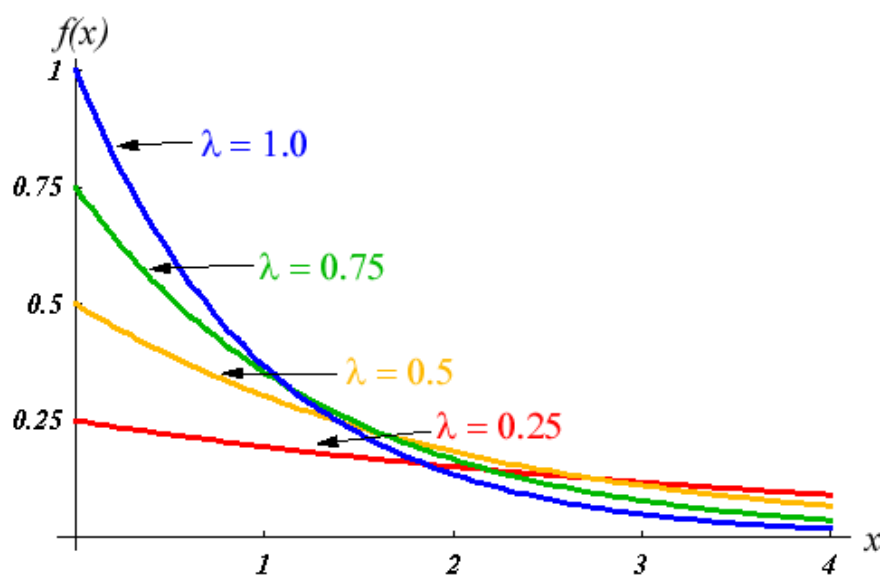
$$h(t) = \frac{f(t)}{S(t)} = \lambda.$$

Η εκθετική κατανομή είναι το πιο απλό μοντέλο διάρκειας ζωής, όμως η συνάρτηση διακινδύνευσης της είναι ανεξάρτητη του χρόνου κάτι που την καθιστά μη ρεαλιστικό μοντέλο.

Αναφέρατε, επίσης, η αναμενόμενη τιμή και η διασπορά οι οποίες δίνονται από τους εξής τύπους:

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

$$V(T) = 1/\lambda^2$$



Γράφημα 1.5

Στο παραπάνω γράφημα φαίνεται η σ.π.π της εκθετικής κατανομής για διάφορες τιμές της παραμέτρου λ .

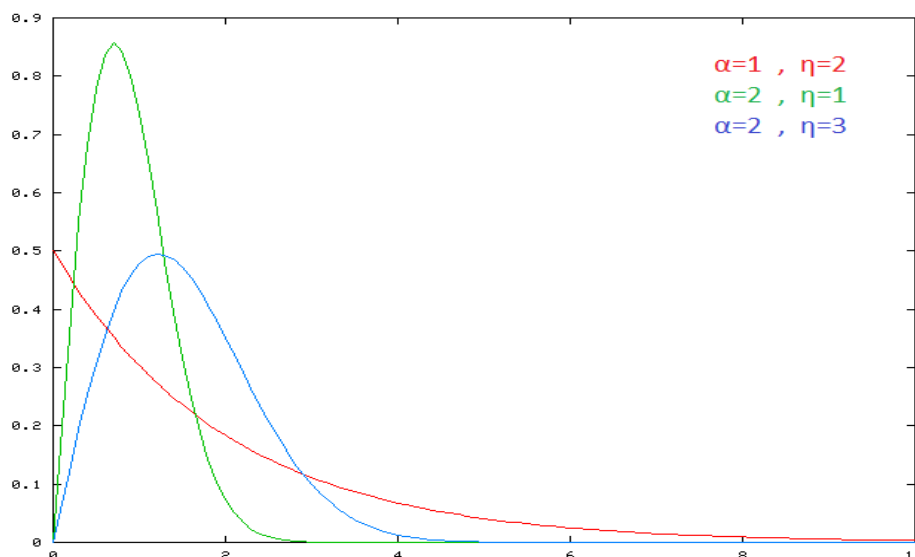
1.5.2 Η κατανομή Weibull (Weibull distribution)

Η κατανομή Weibull είναι η πλέον διαδεδομένη κατανομή στην ανάλυση αξιοπιστίας, κάτι που αποδεικνύεται τόσο εμπειρικά όσο και θεωρητικά. Κατασκευάστηκε από τον Waloddi Weibull το 1937. Το 1951 σε δημοσίευση του, ισχυρίστηκε ότι η κατανομή του μπορεί να εφαρμοστεί σε ποικίλλα προβλήματα, το οποίο με το πέρασμα του χρόνου έχει αποδειχθεί.

Η σ.π.π. με παράμετρο κλίμακας $a > 0$ και παράμετρο σχήματος $\eta > 0$ ορίζεται ως:

$$f(t) = \eta a^{-\eta} t^{\eta-1} \exp\{-(t/a)^\eta\}, t > 0.$$

Στο παρακάτω γράφημα φαίνονται οι γραφικές παραστάσεις της σ.π.π. της κατανομής Weibull για διάφορες τιμές των παραμέτρων a, η . Παρατηρείται ότι για $\eta=1$ προκύπτει η εκθετική κατανομή. Ενώ, για $\eta=2$ έχουμε την περίπτωση της κατανομής Rayleigh, η οποία είναι γνωστή για τη χρήση της στη μοντελοποίηση του ύψους κύματος σχετικά με τους θαλάσσιους κυματισμούς.



Γράφημα 1.6

Η συνάρτηση αξιοπιστίας δίνεται από τον τύπο :

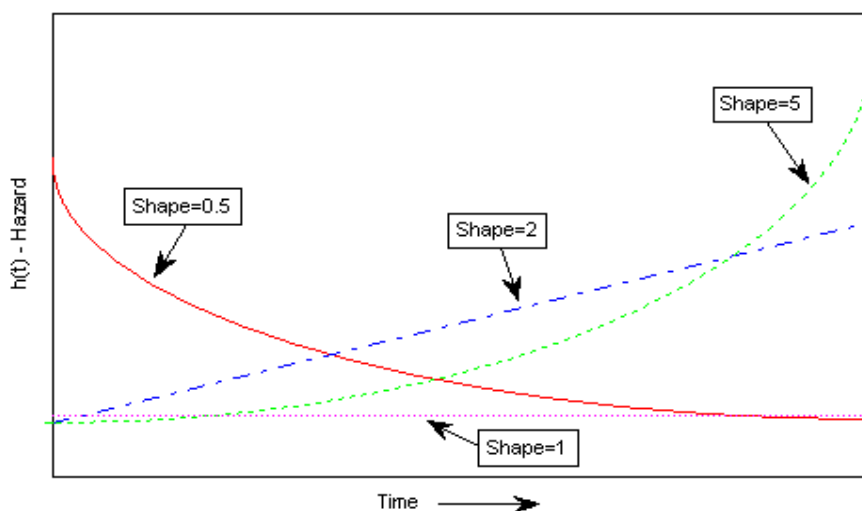
$$S(t) = \int_t^{\infty} \eta \alpha^{-\eta} (\tau/\alpha)^{\eta-1} e^{-(\tau/\alpha)^{\eta}} d\tau = \exp\{-(t/a)^{\eta}\},$$

ενώ η συνάρτηση διακινδύνευσης από τη σχέση:

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

Από την τελευταία σχέση, προκύπτει ότι η συνάρτηση διακινδύνευσης είναι αύξουσα όταν η παράμετρος σχήματος είναι ίση με τη μονάδα, φθίνουσα όταν $\eta < 1$ και σταθερή όταν $\eta = 1$ (γράφημα 1.7). Η συγκεκριμένη παρατήρηση καθιστά την κατανομή Weibull ως ένα αρκετά ευέλικτο μοντέλο αξιοπιστίας

The Weibull Hazard Function



Γράφημα 1.7

Τέλος η αναμενόμενη διάρκεια ζωής (μέση τιμή) ισούται με :

$$E(T) = \alpha\Gamma(1 + \eta^{-1}),$$

όπου όπου Γ η συνάρτηση Γάμμα για την οποία ισχύει:

$$\Gamma(u) = \int_0^{\infty} t^{u-1} e^{-t} dt.$$

Συνεπώς, η διασπορά της διάρκειας ζωής ισούται με

$$V(T) = \alpha^2[\Gamma(1 + 2\eta^{-1}) - [\Gamma(1 + \eta^{-1})]^2].$$

Μέσω της κατανομής Weibull μπορούν να εκφραστούν αρκετές άλλες κατανομές όπως η εκθετική, η κανονική, η κατανομή Rayleigh και μερικές φορές η Poisson και η διωνυμική κατανομή. Η παραπάνω ιδιότητα της Weibull κατανομής είναι ένας από τους λόγους για τους οποίους είναι συχνή η χρήση της στην ανάλυση επιβίωσης. Επιπλέον, η συγκεκριμένη κατανομή μπορεί να εφαρμοστεί και σε πολύ μικρό αριθμό δεδομένων, κάτι που την καθιστά ιδιαίτερα χρήσιμη στον τομέα της μηχανικής. (The New Weibull Handbook, 2004)

1.5.3 Η κατανομή Γάμμα (Gamma distribution)

Η κατανομή Γάμμα με παραμέτρους κλίμακας και σχήματος $\lambda > 0$ και $\alpha > 0$, αντίστοιχα έχει συνάρτηση πυκνότητας πιθανότητας:

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

όπου η συνάρτηση Γ για ακέραιους αριθμούς ορίζεται ως : $\Gamma(\alpha) = (\alpha - 1)!$.

Συνεπώς, αν θεωρήσουμε την παράμετρο σχήματος ίση με την μονάδα ($\alpha=1$) προκύπτει η σ.π.π. της εκθετικής κατανομής (γράφημα 1.8). Επιπλέον η κατανομή γάμμα είναι πιο ευέλικτη από την εκθετική, καθώς μπορεί να συμπεριλάβει περισσότερες πληροφορίες. Παραδείγματος χάριν, ας θεωρήσουμε ότι ένα προϊόν έχει επιβιώσει μέχρι τη χρονική στιγμή t_0 , και η πιθανότητα να επιβιώσει για χρονικό διάστημα μεγαλύτερο της συγκεκριμένης χρονικής στιγμής εξαρτάται και από την ηλικία του προϊόντος. Τότε η κατανομή γάμμα προσαρμόζεται καλύτερα από την εκθετική κατανομή.

Αν και όπως είδαμε παραπάνω η κατανομή γάμμα αποτελεί ένα ευέλικτο μοντέλο, η συνάρτηση αξιοπιστίας είναι σχετικά δύσκολη μορφολογικά και προκύπτει ως εξής:

$$S(t) = \int_t^{\infty} f(\tau) d\tau = \int_t^{\infty} \frac{\lambda^{\alpha} \tau^{\alpha-1} e^{-\lambda\tau}}{\Gamma(\alpha)} d\tau = \int_u^{\infty} \frac{u^{\alpha-1} e^{-u}}{\Gamma(\alpha)} du = \frac{\Gamma(\alpha, \lambda t)}{\Gamma(\alpha)}$$

Στον υπολογισμό του παραπάνω ολοκληρώματος χρησιμοποιήθηκε ο μετασχηματισμός $u = \lambda t$, ενώ η συνάρτηση $\Gamma(\alpha, \lambda t)$ που εμφανίζεται στην τελευταία ισότητα αποτελεί την άνω ατελή συνάρτηση Γάμμα (*upper incomplete Gamma function*) και δίνεται από την σχέση:

$$\Gamma(\alpha, \lambda t) = \int_{\lambda t}^{\infty} t^{\alpha-1} e^{-t} dt .$$

Η συνάρτηση διακυνδύευσης ισούται με:

$$h(t) = \frac{\lambda^{\alpha} t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha, \lambda t)}$$

και συμπεριφέρεται ως εξής:

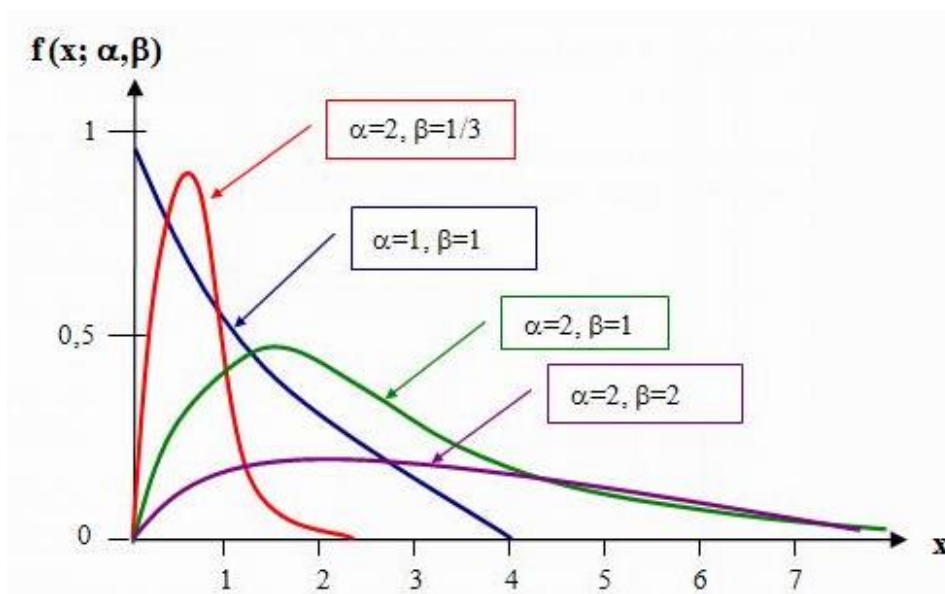
$$h(t) = \begin{cases} \text{φθίνουσα} & , 0 < \alpha < 1 \\ \text{σταθερή} & , \alpha = 1 \\ \text{αύξουσα} & , \alpha > 1 \end{cases}$$

Τέλος, η αναμενόμενη τιμή και η διασπορά της τυχαίας μεταβλητής T δίνονται από τους τύπους:

$$E(T) = \frac{\Gamma(\alpha + 1)}{\lambda \Gamma(\alpha)} = \frac{\alpha}{\lambda}$$

και

$$V(T) = \frac{\alpha(\alpha + 1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2} .$$



Γράφημα 1.8

1.5.4 Η Λογαριθμο-κανονική κατανομή (lognormal distribution)

Η κανονική κατανομή με σ.π.π. $f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\}$, $-\infty < t < \infty$, όπου $\mu \in (-\infty, +\infty)$ η παράμετρος θέσης και $\sigma > 0$ η παράμετρος κλίμακας, είναι συμμετρική. Το παραπάνω χαρακτηριστικό της κανονικής κατανομής αντιτίθεται στην λοξότητα που παρουσιάζουν οι περισσότερες εφαρμογές της ανάλυσης επιβίωσης, η οποία προκύπτει από το γεγονός ότι τα δεδομένα δεν μπορούν να πάρουν αρνητικές τιμές. Για να υπερβούμε αυτό το εμπόδιο μπορούμε να χρησιμοποιήσουμε το μετασχηματισμό $Y = \ln T$. Η τ.μ. Y θα ακολουθεί λογαριθμο-κανονική κατανομή, η οποία αποτελεί χρησιμότερο μοντέλο διάρκειας ζωής.

Συνεπώς, προκύπτει ο παρακάτω ορισμός:

$$T \sim N(\mu, \sigma^2) \Leftrightarrow Y = \ln T \sim \text{log-normal}(\mu, \sigma^2)$$

Για την εύρεση της σ.π.π. της λογαριθμο-κανονικής κατανομής εργαζόμαστε ως εξής:

Έστω $T \sim N(\mu, \sigma^2)$ τότε:

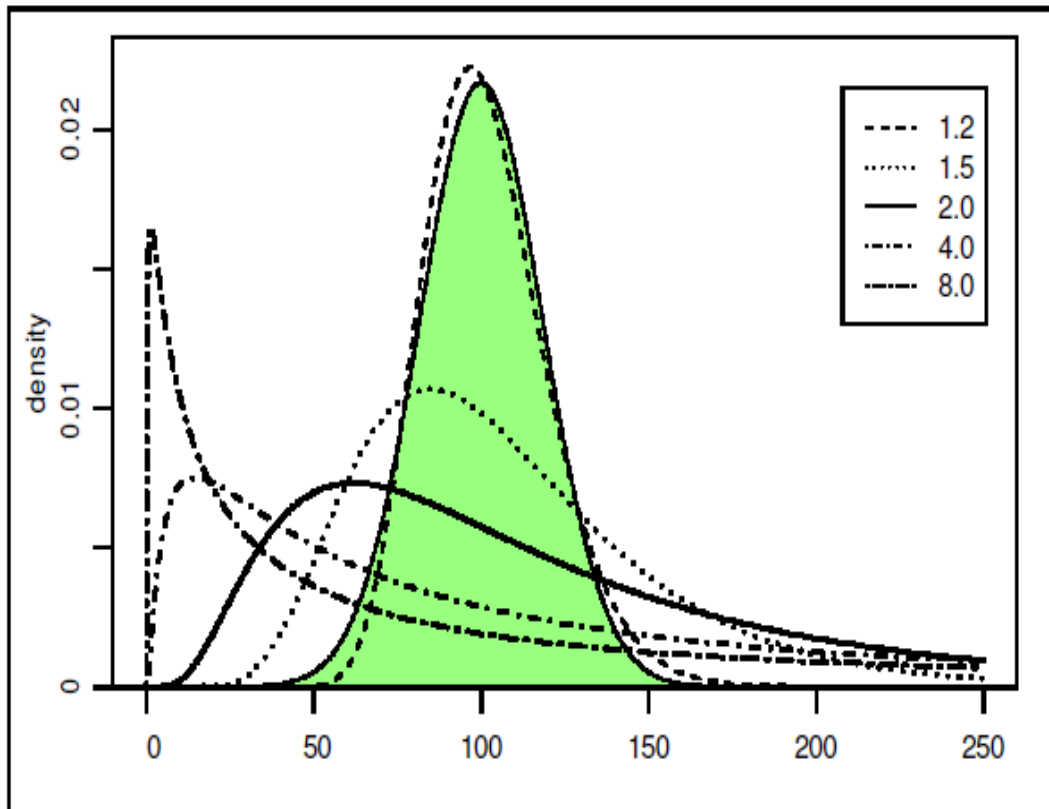
$$\begin{aligned} F(t) &= P(T \leq t) = P(e^Y \leq t) = P(Y \leq \ln t) = \Phi\left(\frac{\ln t - \mu}{\sigma}\right) = \\ &= \int_{-\infty}^{\ln t} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\} dx = \int_0^t \frac{1}{\sigma u\sqrt{2\pi}} \exp\left\{-\frac{(\ln u - \mu)^2}{2\sigma^2}\right\} du \end{aligned}$$

όπου $u = e^x$.

Άρα,

$$f(t) = \frac{dF(t)}{dt} = \frac{1}{\sigma t\sqrt{2\pi}} \exp\left\{-\frac{(\ln t - \mu)^2}{2\sigma^2}\right\}$$

η σ.π.π. της λογαριθμο-κανονικής κατανομής με παράμετρο κλίμακας e^μ και παράμετρο σχήματος $\sigma > 0$.



Γράφημα 1.9

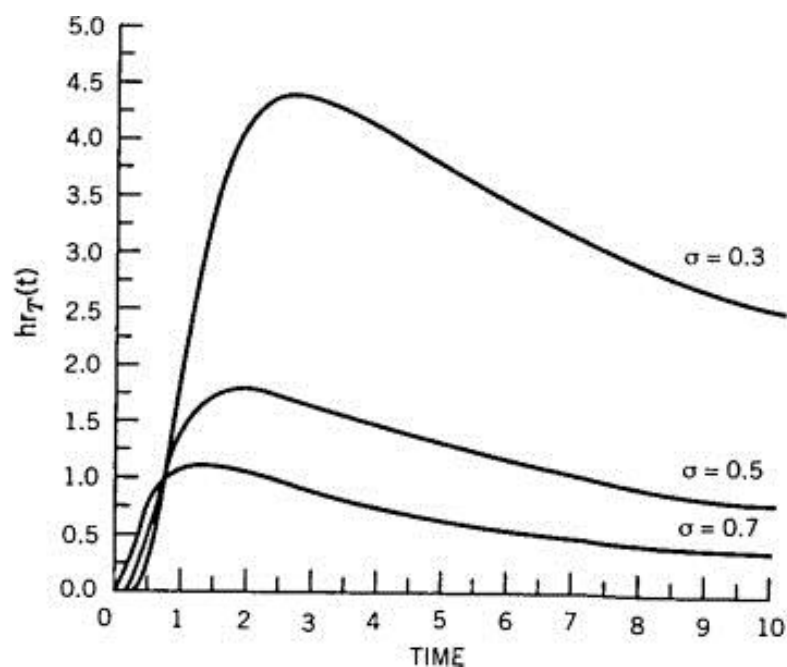
Στο παραπάνω γράφημα παρατηρούμε πέντε διαφορετικές γραφικές παραστάσεις σ.π.π. λογαριθμο-κανονικής κατανομής, συγκρινόμενες με την κανονική κατανομή $N(100, 20^2)$. Στις συγκεκριμένες γραφικές παραστάσεις η παράμετρος θέσης μ παραμένει ίδια, ενώ η παράμετρος σχήματος σ μεταβάλλεται. Διαπιστώνεται ότι αυξάνοντας την παράμετρο σχήματος σ , η κατανομή μας, τείνει πιο γρήγορα στο μηδέν. Τέλος, μια αλλαγή της παραμέτρου μ θα επηρέαζε την κλιμάκωση τόσο στην οριζόντια όσο και στην κατακόρυφη κατεύθυνση, στην ουσία όμως το σχήμα θα παρέμενε ίδιο.

Πίνακας 1.1: Μια σύντομη σύγκριση μεταξύ της κανονικής και της λογαριθμο-κανονικής κατανομής.

	Κανονική Κατανομή (<i>normal distribution</i>)	Λογαριθμο-κανονική Κατανομή (<i>lognormal distribution</i>)
Γράφικη παράσταση σ.π.π.	Συμμετρική	Λοξή
	Χαρακτηριστικά	
Διάμεσος	\bar{x} , αριθμητική	\bar{x}^* , γεωμετρική
Τυπική Απόκλιση	s , προσθετική	s^* , πολλαπλασιαστική
Μέτρο Διασποράς	s/\bar{x}	s^*
Διαστήματα Εμπιστοσύνης		
68.3%	$\bar{x} \pm s$	$\bar{x}^* \times / s^*$
95.5%	$\bar{x} \pm 2s$	$\bar{x}^* \times / (s^*)^2$
99.7%	$\bar{x} \pm 3s$	$\bar{x}^* \times / (s^*)^3$
Υποσημείωση:	$\alpha \times / \beta = [\frac{\alpha}{\beta}, \alpha \cdot \beta]$	

(Limpert et al. 2001)

Επίσης, μεγάλο ενδιαφέρον παρουσιάζει η συνάρτηση διακινδύνευσης της λογαριθμο-κανονικής κατανομής. Όπως φαίνεται στο διάγραμμα 1.10, αρχικά και μέχρι μια χρονική στιγμή αυξάνεται και στη συνέχεια τείνει στο μηδέν ($h(t) \rightarrow 0$) καθώς $t \rightarrow \infty$.



Γράφημα 1.10

Αυτό σημαίνει ότι αντικείμενα που ακολουθούν λογαριθμο-κανονική κατανομή μέχρι κάποια ηλικία διατρέχουν πολύ υψηλό κίνδυνο, όμως καθώς μεγαλώνουν ο κίνδυνος μειώνεται μέχρι που τείνει να γίνει μηδενικός.

Η αναμενόμενη τιμή και η διασπορά δίνονται από τους τύπους:

$$E(T) = \exp(\mu + \sigma^2) , \quad V(T) = [\exp(2\mu + \sigma^2)][\exp(\sigma^2) - 1].$$

1.5.5 Η λογαριθμο-λογιστική κατανομή (log-logistic distribution)

Ορίζουμε την λογαριθμο-λογιστική κατανομή ως εξής:

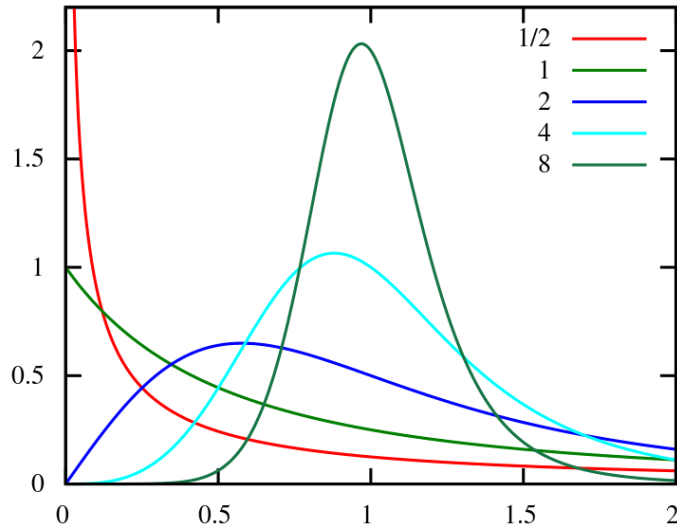
$$T \sim \text{log-logistic} \Leftrightarrow Y = \ln T \sim \text{logistic} .$$

Μέσω της σ.π.π. της λογιστικής κατανομής, $f(t) = \frac{\tau^{-1} \exp\{(t-\nu)/\tau\}}{[1 + \exp\{(t-\nu)/\tau\}]^2}$, όπου $\nu \in (-\infty, \infty)$ η παράμετρο θέσης και $\tau > 0$ η παράμετρος κλίμακας, αποδεικνύεται η σ.π.π. της λογαριθμο-λογιστικής κατανομής (η απόδειξη είναι παρόμοια με αυτήν της σ.π.π. της λογαριθμο-κανονικής κατανομής), η οποία δίνεται από τη σχέση:

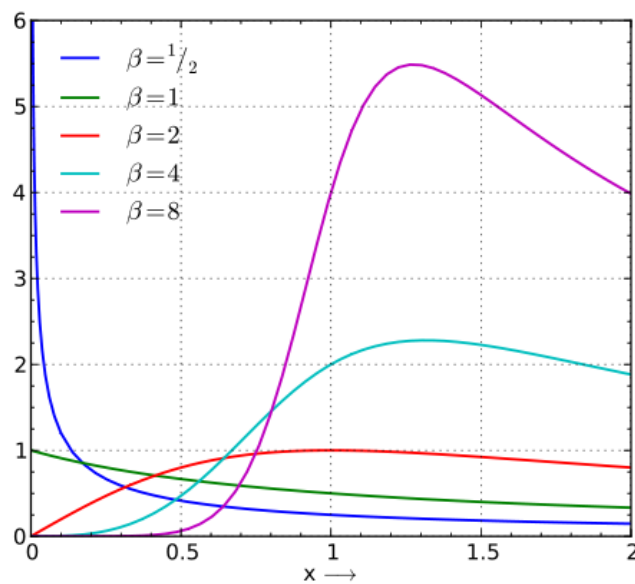
$$f(t) = \frac{\alpha \beta t^{\beta-1}}{(1 + at^\beta)^2}, t \geq 0$$

όπου $\beta = \tau^{-1}$ και $\alpha = e^{-\nu/\tau}$.

Η συνάρτηση αξιοπιστίας ισούται με: $S(t) = (1 + at^\beta)^{-1}$ και η συνάρτηση διακινδύνευσης με: $h(t) = \frac{\alpha\beta t^{\beta-1}}{1+at^\beta}$.



Γράφημα 1.11: Συνάρτηση πυκνότητας πιθανότητας



Γράφημα 1.12: Συνάρτηση διακινδύνευσης

(Wikipedia, the log-logistic distribution)

Στα γραφήματα 1.11 και 1.12 φαίνονται οι γραφικές παραστάσεις τόσο της σ.π.π. της λογαριθμο-λογιστικής κατανομής, όσο και της συνάρτησης διακινδύνευσης, όπου η παράμετρος θέσης $\alpha=1$ και είναι σταθερή σε όλες τις κατανομές, ενώ η παράμετρος σχήματος β παίρνει τις τιμές που αναγράφονται στα διαγράμματα.

Η λογαριθμο-λογιστική κατανομή χρησιμοποιείται στις εξής περιπτώσεις:

- Στην ανάλυση επιβίωσης σαν παραμετρικό μοντέλο για γεγονότα που η τιμή τους αυξάνεται μέχρι κάποια χρονική στιγμή και στη συνέχεια μειώνεται.

Σε αρκετές περιπτώσεις προτιμάται της κατανομής Weibull, καθώς όταν η παράμετρος σχήματος είναι μεγαλύτερη της μονάδας ($\beta > 1$) η καμπύλη διακινδύνευσης δεν είναι μονότονη. Επιπλέον, η συνάρτηση διακινδύνευσης είναι σε κλειστή μορφή κάτι που την καθιστά αρκετά εύχρηστο μοντέλο για την επεξεργασία δεδομένων με αποκομμένες παρατηρήσεις.

Χαρακτηριστικό παράδειγμα αποτελεί η θνησιμότητα ανθρώπων που πάσχουν από καρκίνο και ακολουθούν κάποια συγκεκριμένη θεραπεία.

- Στην υδρολογία για τη μοντελοποίηση της ροής του ρεύματος και της βροχόπτωσης.
- Στις οικονομικές επιστήμες ως μοντέλο κατανομής του πλούτου ή των εισόδων.

1.6 Αριθμητικό παράδειγμα

Στην ανάλυση επιβίωσης, αρκετά σημαντικό είναι να ελεγχθεί η καταλληλότητα του προσαρμοσμένου μοντέλου. Δηλαδή, να διαπιστωθεί αν η προτεινόμενη κατανομή περιγράφει το σύνολο των δεδομένων. Υπάρχουν αρκετοί στατιστικοί και γραφικοί έλεγχοι υποθέσεων που συμβάλλουν στην επιλογή του καταλληλότερου μοντέλου. Στο παρακάτω παράδειγμα χρησιμοποιείται ο έλεγχος

Anderson-Darling, ο οποίος χρησιμεύει, κυρίως, για τη σύγκριση διαφορετικών προσαρμοσμένων μοντέλων.

Αρχικά, για το κάθε μοντέλο ξεχωριστά υπολογίζεται το στατιστικό ελέγχου:

$$A^2 = -N - S ,$$

όπου $S = \sum_{i=1}^N \frac{(2i-1)}{N} [\ln F(Y_{(i)}) + \ln(1 - F(Y_{(N+1-i)}))]$

και $Y_{(i)}$ τα ταξινομημένα δεδομένα. Το μοντέλο που αντιστοιχεί στην μικρότερη τιμή είναι το καταλληλότερο (*Καρώνη, 2009*).

Ας θεωρήσουμε τα παρακάτω δεδομένα διάρκειας ζωής (σε χιλιόμετρα), τα οποία αφορούν 38 αμορτισέρ οχημάτων. Στα δεδομένα αυτά εμπεριέχονται συνολικά 27 από δεξιά αποκομμένες παρατηρήσεις, ενώ στην πραγματικότητα συνέβησαν μόλις 11 γεγονότα. Σκοπός της παρούσας ανάλυσης είναι να επιλεγεί το μοντέλο διάρκειας ζωής που περιγράφει καλύτερα τα δεδομένα.

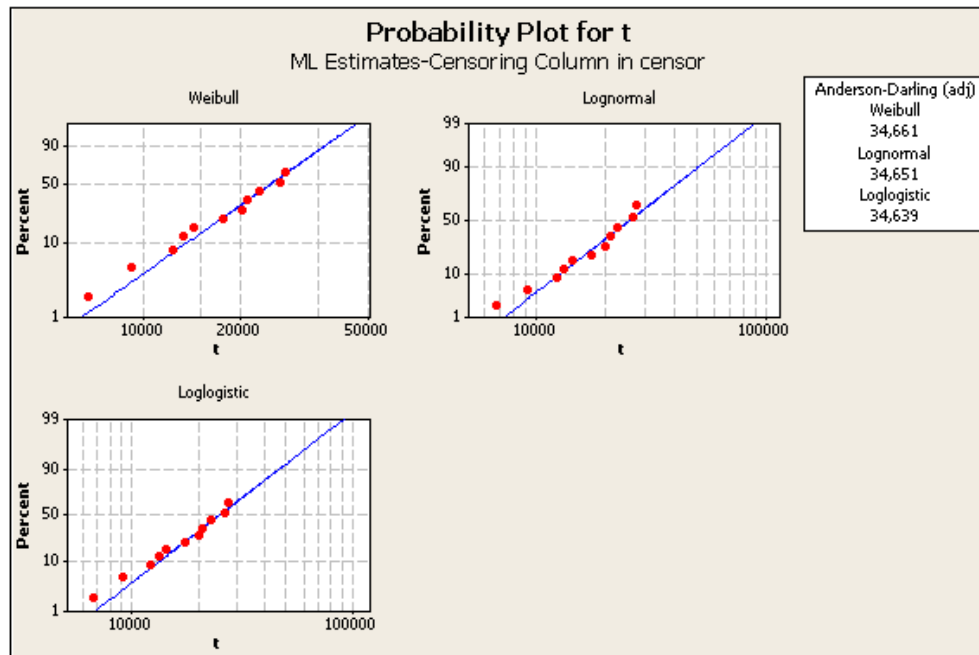
6700	6950 *	7820 *	8790 *	9120	9660 *
9820 *	11310 *	11690 *	11850 *	11880 *	12140 *
12200	12870 *	13150	13330 *	13470 *	14040 *
14300	17520	17540 *	17890 *	18450 *	18960 *
18980 *	19410 *	20100 *	20100	20150 *	20320 *
20900	22700	23490 *	26510	27410 *	27490
27890 *	28100 *				

(με αστερίσκο σημειώνονται οι αποκομμένες παρατηρήσεις)

Η επιλογή του καταλληλότερου μοντέλου θα γίνει ανάμεσα στις τρεις κατανομές, οι οποίες έχει αποδειχθεί ότι αποτελούν τα δημοφιλέστερα μοντέλα διάρκειας ζωής: την κατανομή Weibull, τη Λογαριθμο-κανονική και τη Λογαριθμο-λογιστική κατανομή.

Όπως παρατηρείται στα παρακάτω γραφήματα πιθανότητας (Γράφημα 1.12) τα σημεία προσεγγίζουν ικανοποιητικά τις ευθείες, κάτι που υποδεικνύει την καλή προσαρμογή και των τριών μοντέλων στα δεδομένα μας. Σύμφωνα με τον έλεγχο *Anderson-Darling*, ελάχιστα καλύτερη προσαρμογή στα δεδομένα έχει η

Λογαριθμο-λογιστική κατανομή καθώς αντιστοιχεί στην μικρότερη τιμή του στατιστικού, με πολύ μικρή διαφορά από τις άλλες δύο.



Γράφημα 1.12

Μέσω του συγκεκριμένου παραδείγματος διαπιστώνεται ότι τα τρία δημοφιλέστερα μοντέλα διάρκειας ζωής (Weibull, Lognormal, Loglogistic) σε πολλές εφαρμογές της ανάλυσης επιβίωσης προσαρμόζονται εξίσου καλά.

ΚΕΦΑΛΑΙΟ 2:

ΠΑΡΑΓΟΝΤΙΚΟΙ ΣΧΕΔΙΑΣΜΟΙ

2.1 Εισαγωγή

2.1.1 Βασικές έννοιες και ορισμοί

(Prvan et al., 2002)

Ορισμός 2.1.1:

Πείραμα (experiment) ονομάζεται μια δοκιμή ή μια σειρά δοκιμών όπου ο πειραματιστής πραγματοποιεί σκόπιμες αλλαγές στις μεταβλητές εισόδου μιας διαδικασίας ή ενός συστήματος έτσι ώστε να μπορούν να παρατηρηθούν ή να εντοπιστούν οι παράγοντες που επηρεάζουν τις μεταβλητές εξόδου.

Ορισμός 2.1.2:

Η *πειραματική σχεδίαση (design of experiments)* ασχολείται με τη σχεδίαση και τη διεξαγωγή πειραμάτων, τα οποία μέσω της ανάλυσης των αποτελεσμάτων, θα οδηγήσουν τους ερευνητές στην απόκτηση έγκυρων και αντικειμενικών συμπερασμάτων.

Ορισμός 2.1.3:

Παράγοντας (factor) ονομάζεται οποιοδήποτε χαρακτηριστικό των πειραματικών μονάδων μπορεί να επηρεάσει την παρατηρούμενη απόκριση του αποτελέσματος. Οι παράγοντες που εφαρμόζονται από τον πειραματιστή ονομάζονται *παράγοντες ελέγχου (treatment/ control factors)*, ενώ οι τυχόν παράγοντες που είναι εγγενείς ιδιότητες των πειραματικών μονάδων ονομάζονται *παράγοντες θορύβου (block/noise factors)*. Τέλος, οι πιθανές τιμές που μπορεί να πάρει ένας παράγοντας ονομάζονται *επίπεδα (levels)* του παράγοντα.

Ορισμός 2.1.4:

Με τον όρο *παραγοντικό σχεδιασμό (factorial design)*, εννοούμε ότι σε κάθε πλήρη δοκιμή ή επανάληψη του πειράματος, εξετάζονται όλοι οι δυνατοί συνδυασμοί των επιπέδων (σταθμών) των παραγόντων.

Ορισμός 2.1.5:

Με τον όρο *αγωγή (treatment combination)* ονομάζουμε κάθε πιθανό συνδυασμό των επιπέδων των παραγόντων ελέγχου.

Παρακάτω παρατίθενται κάποιοι από τους σημαντικότερους παραγοντικούς σχεδιασμούς, μερικοί από τους οποίους θα παρουσιαστούν αναλυτικότερα σε επόμενες παραγράφους.

Ορισμός 2.1.6:

Ο σχεδιασμός, στον οποίο, κάθε δυνατός συνδυασμός επιπέδων παραγόντων εμφανίζεται ίσο αριθμό φορών ονομάζεται *πλήρης παραγοντικός σχεδιασμός (complete factorial design)* και συμβολίζεται ως FD.

Ορισμός 2.1.7:

Στον *πλήρη 2^k σχεδιασμό* υπεισέρχονται k παράγοντες, οι οποίοι λαμβάνουν τιμές σε δύο στάθμες. Ενώ, *κλασματικός παραγοντικός σχεδιασμός (fractional factorial design, FFD)* ονομάζεται ο σχεδιασμός στον οποίο εμφανίζεται μόνο ένα υποσύνολο των δυνατών συνδυασμών επιπέδων παραγόντων. Συνηθίζεται ο συμβολισμός 2^{k-p} για να περιγραφεί το $1/2^p$ κλάσμα ενός 2^k σχεδιασμού.

Ορισμός 2.1.8:

Ο BIB-σχεδιασμός (*balanced incomplete block design*) με παραμέτρους (v, b, r, k, λ) αποτελείται από b ομάδες (*blocks*) μεγέθους k . Ο αριθμός των αγωγών στον συγκεκριμένο σχεδιασμό είναι v , ενώ η καθεμία ανήκει σε r μπλοκ. Τέλος, κάθε ζευγάρι αγωγών ανήκει σε ακριβώς λ μπλοκ.

2.1.2 Σχεδιασμός και υλοποίηση πειραμάτων

(Jeff Wu et al., 2000)

Στη συγκεκριμένη παράγραφο παρουσιάζονται συνοπτικά τα σημαντικά βήματα που ένας πειραματιστής πρέπει να ακολουθήσει τόσο για το σχεδιασμό όσο και για την υλοποίηση του πειράματος.

Ο στόχος του πειράματος πρέπει να είναι σαφώς καθορισμένος, κάτι που σημαίνει ότι όλα τα ενδιαφερόμενα μέρη πρέπει να συμβάλλουν στον καθορισμό ενός κοινού στόχου. Παραδείγματος χάριν, για την κατασκευή ενός προϊόντος θα πρέπει να ληφθεί υπ' όψιν η άποψη επιστημόνων ποικίλων ειδικοτήτων, όπως είναι οι μηχανικοί ειδικευμένοι στο σχεδιασμό προϊόντων, καθώς και οι υπεύθυνοι για το σχεδιασμό και την επίβλεψη της κατασκευαστικής διαδικασίας. Επιπλέον, ήδη από το στάδιο του σχεδιασμού της παραγωγικής διαδικασίας θα πρέπει να είναι σαφείς οι ανάγκες και οι προσδοκίες της αγοράς.

Η επιλογή της επιθυμητής απόκρισης (*response*) αποτελεί εξίσου σημαντικό μέρος της πειραματικής διαδικασίας. Γενικά, τα αποτελέσματα ενός πειράματος μπορεί να είναι διακριτά (*discrete*) ή συνεχή (*continuous*). Τα διακριτά αποτελέσματα εκτός από αριθμούς, μπορεί να αντιστοιχούν και σε κατηγορίες, π.χ. καλό/κακό ή εύκολο/μέτριο/δύσκολο. Στην πλειονότητα των περιπτώσεων, όμως, προτιμάται το αποτέλεσμα του πειράματος να εκφράζεται σε συνεχή μορφή. Ανεξάρτητα από τη μορφή της απόκρισης θα πρέπει να είναι σαφής η επιθυμητή βελτιστοποίηση της. Οι τρεις πιθανές κατηγορίες βελτιστοποίησης της απόκρισης είναι οι εξής:

- *Smaller-the-better*: Στην περίπτωση αυτή, η επιθυμητή απόκριση είναι η ελάχιστη δυνατή. Για παράδειγμα, σε μια μελέτη για την εξέλιξη ενός όγκου σε ασθενείς που υποβάλλονται σε μια συγκεκριμένη θεραπεία, επιθυμητή απόκριση είναι η όσο το δυνατόν μεγαλύτερη συρρίκνωση του όγκου.
- *Larger-the-better*: Στην περίπτωση αυτή, η επιθυμητή απόκριση είναι η μέγιστη δυνατή. Για παράδειγμα, ας θεωρήσουμε ένα πείραμα στο οποίο μελετάται η διάρκεια ζωής ενός συνόλου ρουλεμάν κάτω από στρεσογόνες

συνθήκες. Προφανώς, στόχο του πειράματος αποτελεί ο καθορισμός των συνθηκών κάτω από τις οποίες μεγιστοποιείται η απόκριση.

- ο Nominal-the-best: Σπανιότερα, η επιθυμητή απόκριση λαμβάνει συγκεκριμένη τιμή.

Στο τρίτο βήμα του σχεδιασμού του πειράματος θα πρέπει να επιλεγούν οι παράγοντες που πιθανώς επηρεάζουν την απόκριση, καθώς και τα επίπεδα αυτών. Για την επιλογή των πιθανών παραγόντων, χρησιμοποιείται συχνά ανάλυση διασποράς, ενώ συνηθίζεται και η χρήση γραφικών μεθόδων. Στην επιλογή των παραγόντων και των επιπέδων τους θα πρέπει να ληφθούν σοβαρά υπ' όψιν πρακτικοί περιορισμοί, καθώς και περιορισμοί που αφορούν το κόστος. Σε αρκετές περιπτώσεις, κάποιος συνδυασμός παραγόντων μπορεί να οδηγήσει σε καταστροφικά αποτελέσματα, γεγονός που αποδεικνύει τη σημαντικότητα της ορθής επιλογής των παραγόντων και των επιπέδων τους.

Η επιλογή του πειραματικού σχεδιασμού είναι κρίσιμη. Μία κακή επιλογή σχεδιασμού μπορεί συχνά να οδηγήσει σε εξόρυξη μερικής πληροφορίας, με αποτέλεσμα ο ερευνητής να μην καταλήγει σε στατιστικά σημαντικά συμπεράσματα. Από την άλλη πλευρά, αν το πείραμα είναι καλά σχεδιασμένο, η διαδικασία ανάλυσης των πειραματικών παρατηρήσεων καθίσταται απλούστερη.

Στο στάδιο εκτέλεσης του πειράματος, η χρήση ενός πίνακα σχεδιασμού (*planning matrix*) είναι καθοριστική. Ο πίνακας αυτός περιλαμβάνει το σύνολο των εκτελέσεων του πειράματος. Επίσης, σε αρκετές περιπτώσεις συστήνεται μια δοκιμαστική εκτέλεση του πειράματος έτσι ώστε να διαπιστωθούν τυχόν δυσκολίες, δηλαδή, να εντοπισθούν τυχόν προβλήματα σχετικά με τον καθορισμό των παραγόντων και τη μέτρηση της απόκρισης.

Στη συνέχεια ο πειραματιστής πρέπει να προχωρήσει σε προσεκτική *ανάλυση των δεδομένων*. Η ανάλυση που θα διεξαχθεί θα πρέπει να είναι κατάλληλη για το σχεδιασμό που χρησιμοποιήθηκε για τη συλλογή των δεδομένων. Αυτή περιλαμβάνει συχνά την προσαρμογή ενός στατιστικού μοντέλου, καθώς και τον έλεγχο συγκεκριμένων υποθέσεων.

Τελευταίο βήμα αποτελεί η *εξαγωγή συμπερασμάτων*. Με βάση την ανάλυση των δεδομένων, προκύπτουν συμπεράσματα σε σχέση με τους παράγοντες που επηρεάζουν σημαντικά την απόκριση και τις βέλτιστες στάθμες αυτών, ούτως ώστε να ικανοποιούνται οι αρχικοί στόχοι. Παράλληλα, κατασκευάζεται ένα

μοντέλο που εξηγεί την απόκριση μέσω των στατιστικά σημαντικών παραγόντων. Τέλος, η ανάλυση μπορεί ενίοτε να αναδείξει την ανάγκη για περαιτέρω πειραματισμό. Για παράδειγμα, στην περίπτωση που δύο μοντέλα εξηγούν εξίσου καλά την απόκριση, ένα περαιτέρω πείραμα παρακολούθησης θα ήταν αναγκαίο, έτσι ώστε να επιλεγεί το καταλληλότερο μοντέλο, με τελικό στόχο την επίτευξη βελτιστοποίησης.

2.1.3 Ιστορική Αναδρομή

Η έννοια του παραγοντικού σχεδιασμού εισήχθη για πρώτη φορά από τον R. A. Fisher το 1926. Η δουλειά του Fisher, την οποία εξέλιξαν αργότερα οι F. Yates και D.J. Finney, είχε ως αφετηρία προβλήματα του τομέα της γεωργίας και της βιολογίας.

Γενικά, στα πειράματα που σχετίζονται με τον κλάδο της γεωργίας ο όγκος των δεδομένων είναι συχνά μεγάλος, ενώ χρειάζεται αρκετός χρόνος για την ολοκλήρωσή τους. Καθίσταται επίσης σαφές το γεγονός ότι τέτοιου είδους πειράματα επηρεάζονται άμεσα από τις μεταβολές του χώρου στον οποίο εφαρμόζονται. Η ιδιαίτερη αυτή φύση των γεωργικών και βιολογικών πειραμάτων οδήγησε ώστε ήδη οι πρώτες μελέτες σχετικά με τη σχεδίαση πειραμάτων να οδηγήσουν στην κατασκευή και τη χρήση κλασματικών παραγοντικών σχεδιασμών (*fractional factorial designs*), κυρίως για λόγους οικονομίας. Παράλληλα, η ανάμειξη στους πλήρεις και κλασματικούς σχεδιασμούς (*blocking*) άρχισε σταδιακά να αποτελεί όλο και συχνότερη πρακτική, ώστε να καθίσταται δυνατή η εκτέλεση του συνολικού πλήθους των επαναλήψεων σε ομάδες (*blocks*). Τέλος, οι έρευνες σε σχέση με το σχεδιασμό των γεωργικών πειραμάτων ανέδειξαν τη σημαντικότητα της χρήσης επαναληπτικών, καθώς και ορθογώνιων σχεδιασμών.

Στη συνέχεια, η εξέλιξη της θεωρίας των συνδυαστικών σχεδιασμών (*combinatorial designs*), στην οποία η συμβολή του Ινδο-αμερικανού μαθηματικού R.C. Bose (1939) υπήρξε θεμελιώδης, υποκινήθηκε από προβλήματα στην ανάμειξη σχεδιασμών. Στο πλαίσιο αυτό, κατασκευάστηκαν οι BIB-σχεδιασμοί (*balanced incomplete block designs*), καθώς και οι SBIB-σχεδιασμοί (*symmetric balanced incomplete block designs*). Η πρόοδος που

σημειώθηκε τη συγκεκριμένη περίοδο βρήκε, επίσης, εφαρμογές στις κοινωνικές επιστήμες και σε βιομηχανίες κλωστοϋφαντουργίας και ξυλείας.

Η επόμενη εποχή ταχείας ανάπτυξης ήρθε αμέσως μετά τον δεύτερο Παγκόσμιο Πόλεμο. Στην προσπάθεια τους να εφαρμόσουν παλιές τεχνικές για να λύσουν προβλήματα σε χημικές βιομηχανίες, ο G.E.P. Box και οι συνεργάτες του στις Imperial Chemical Industries ανακάλυψαν ότι καινούριες τεχνικές και έννοιες πρέπει να αναπτυχθούν έτσι ώστε να αντιμετωπιστούν τα μοναδικά χαρακτηριστικά που προκύπτουν σε προβλήματα βιομηχανιών μεταποίησης. Οι νέες τεχνικές στόχευαν περισσότερο στην μοντελοποίηση και την βελτιστοποίηση της διαδικασίας, σε αντίθεση με τον πρωταρχικό στόχο των γεωργικών πειραμάτων, που ήταν η σύγκριση διαφορετικών συνδυασμών παραγόντων. Τα πειράματα τα οποία μελέτησε ο G.E.P. Box και οι συνεργάτες του είναι μικρότερης διάρκειας, το οποίο σε συνδυασμό με το κόστος τους ευνοεί το διαδοχικό πειραματισμό.

Με βάση τις συγκεκριμένες παρατηρήσεις, προέκυψε η ανάπτυξη νέων τεχνικών όσον αφορά τον πειραματικό σχεδιασμό και κυρίως τον βέλτιστο σχεδιασμό (*optimal design*), καθώς και η δημιουργία των κεντρικών σύνθετων σχεδιασμών (*central composite designs*). Η ανάλυση τους βασίζεται σε μεγαλύτερο βαθμό στο μοντέλο παλινδρόμησης, καθώς και στη γραφική ανάλυση. Την περίοδο αυτή, δόθηκε επίσης έμφαση στην εύρεση μιας ευέλικτης στρατηγικής με τη βοήθεια της οποίας να βρίσκονται σχεδιασμοί, οι οποίοι θα ταιριάζουν σε ένα συγκεκριμένο μοντέλο ή πειραματική περιοχή. Με τη χρήση γρήγορων υπολογιστικών αλγορίθμων, οι βέλτιστοι σχεδιασμοί (των οποίων πρωτοπόρος ήταν ο J. Kiefer) έγιναν ένα σημαντικό κομμάτι αυτής της στρατηγικής.

Σχετικά πρόσφατα δόθηκε έμφαση στην προσπάθεια μείωσης της διασποράς του σφάλματος κατά τη διάρκεια ενός πειράματος. Η συγκεκριμένη προσπάθεια υπήρξε πηγή έμπνευσης, έτσι ώστε να δημιουργηθούν καινούριες τεχνικές στην σχεδίαση πειραμάτων. Στον κατασκευαστικό τομέα η ικανότητα να φτιαχτούν αρκετά κομμάτια με ελάχιστα μειονεκτήματα αποτελεί ένα πλέον ανταγωνιστικό πλεονέκτημα. Ως εκ τούτου, η μείωση της διασποράς στα ποιοτικά χαρακτηριστικά αυτών των εξαρτημάτων έχει οδηγήσει στη βελτίωση τόσο της ποιότητας, όσο και της παραγωγικότητας. Ο G. Taguchi υποστήριξε τη χρήση των εύρωστων παραμετρικών σχεδιασμών (*robust parameter designs*) για τη βελτίωση

ενός συστήματος (προϊόν ή διαδικασία), καθιστώντας το λιγότερο ευαίσθητο στις αλλαγές, οι οποίες είναι δύσκολο να ελεγχθούν κατά τη διάρκεια κανονικής λειτουργίας ή χρήσης του προϊόντος ή της διεργασίας. Οι παράγοντες του συστήματος μπορούν να κατηγοριοποιηθούν σε παράγοντες θορύβου και παράγοντες ελέγχου. Έτσι, με την αξιοποίηση των αλληλεπιδράσεων μεταξύ των δύο διαφορετικών τύπων παραγόντων, μπορεί κανείς να πετύχει ευρωστία, επιλέγοντας να ρυθμίσει με τέτοιο τρόπο τους παράγοντες ελέγχου έτσι ώστε το σύστημα να γίνει λιγότερο ευαίσθητο στους παράγοντες θορύβου.

Παραδοσιακά, όταν λαμβάνονται υπ' όψιν ο δειγματικός μέσος (*mean*) και η διασπορά (*variation*), η διασπορά χρησιμοποιείται για να εκτιμηθεί η διακύμανση του δειγματικού μέσου, όπως στο t τεστ ή στην ανάλυση διασποράς. Η έμφαση, λοιπόν, στη διασπορά και την ταξινόμηση των παραγόντων στους δύο παραπάνω τύπους οδήγησε στην ανάπτυξη νέων εννοιών και μεθόδων στο σχεδιασμό και την ανάλυση των εύρωστων παραμετρικών σχεδιασμών. Η αρχική διατύπωση του προβλήματος και κάποιες βασικές έννοιες αναπτύχθηκαν από τον G.Taguchi. Άλλες βασικές αρχές και στατιστικές τεχνικές έχουν αναπτυχθεί από στατιστικούς από τα μέσα της δεκαετίας του 1980, ενώ οι έρευνες στη συγκεκριμένη περιοχή συνεχίζονται μέχρι και σήμερα.

2.2 Είδη παραγοντικοί σχεδιασμοί

2.2.1 Ο 2^k παραγοντικός σχεδιασμός

Στον 2^k παραγοντικό σχεδιασμό παρατηρούνται k παράγοντες, ο καθένας από τους οποίους δέχεται τιμές σε δύο στάθμες. Μια πλήρης επανάληψη ενός τέτοιου σχεδιασμού απαιτεί $\underbrace{2 \times 2 \times \dots \times 2}_{k\text{-φορές}} = 2^k$ παρατηρήσεις.

Ο 2^k σχεδιασμός είναι ιδιαίτερα χρήσιμος στα αρχικά στάδια της πειραματικής διεργασίας, όταν υπάρχουν πολλοί παράγοντες που πρέπει να εξεταστούν. Ο συγκεκριμένος σχεδιασμός, μας πληροφορεί για το μικρότερο αριθμό εκτελέσεων (*runs*) που πρέπει να πραγματοποιηθεί, έτσι ώστε να μελετηθούν και οι k παράγοντες. Δεδομένου ότι υπάρχουν μόνο δύο στάθμες για κάθε παράγοντα, συνηθίζεται να μελετάται η γραμμική επίδραση της απόκρισης πάνω στο εύρος των επιλεγμένων επιπέδων των παραγόντων.

2.2.1.1 Ο 2² παραγοντικός σχεδιασμός

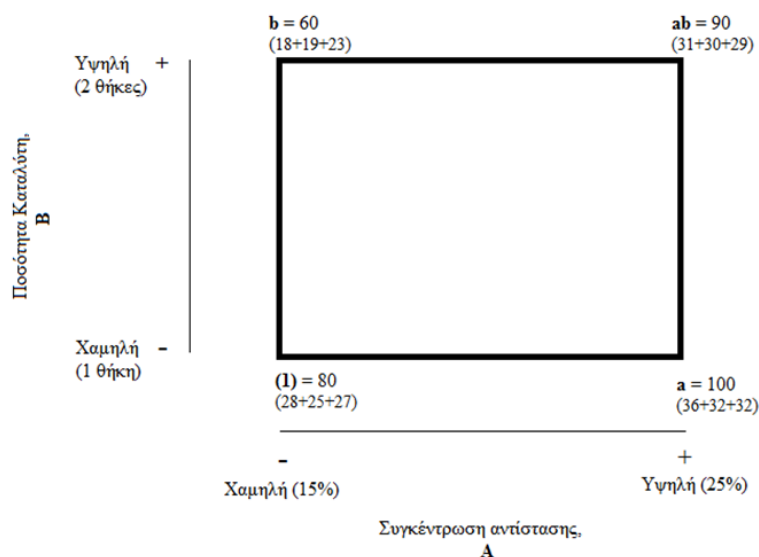
Στην απλούστερη περίπτωση που έχουμε μόνο δύο παράγοντες, έστω Α και Β και καθέναν σε δύο στάθμες, αναφερόμαστε σε ένα 2² παραγοντικό σχεδιασμό. Για ευκολία τα επίπεδα των παραγόντων αντιστοιχίζονται σε «υψηλή» και «χαμηλή» στάθμη.

Για καλύτερη κατανόηση, θα μελετήσουμε τον 2² παραγοντικό σχεδιασμό μέσω ενός παραδείγματος, το οποίο εξετάζει την επίδραση της συγκέντρωσης της αντίστασης και της ποιότητας του καταλύτη στον χρόνο αντίδρασης μιας χημικής διαδικασίας. Έστω ότι η συγκέντρωση της αντίστασης είναι ο παράγοντας Α και οι υπό μελέτη στάθμες του είναι οι 15% και 25%. Επιπλέον, ως παράγοντα Β θεωρείται η ποιότητα του καταλύτη, όπου η «υψηλή» στάθμη αντιστοιχεί στη χρήση δύο θηκών του καταλύτη, και η «χαμηλή» στάθμη συμβολίζει τη χρήση μίας θήκης. Το πείραμα επαναλαμβάνεται τρεις φορές και τα δεδομένα δίνονται στον παρακάτω πίνακα:

Παράγοντας (Factor)		Συνδυασμός Αγωγής (Treatment combination)	Επανάληψη (n) (Run)			Σύνολο (Total)
			I	II	III	
A	B		_____			
-	-	A χαμηλή, B χαμηλή	28	25	27	80
+	-	A υψηλή, B χαμηλή	36	32	32	100
-	+	A χαμηλή, B υψηλή	18	19	23	60
+	+	A υψηλή, B υψηλή	31	30	29	90

Πίνακας 2.1

Συνήθως, η επίδραση ενός παράγοντα συμβολίζεται με ένα κεφαλαίο λατινικό γράμμα. Για παράδειγμα, η επίδραση του παράγοντα A θα συμβολίζεται ως «A», ενώ η αλληλεπίδραση των παραγόντων A και B ως «AB». Επιπλέον, στον 2^2 σχεδιασμό συνηθίζεται ο συμβολισμός «+», «-» για την υψηλή και την χαμηλή στάθμη, αντίστοιχα, ενός παράγοντα.



Σχήμα 2.1 : Συνδυασμοί αγωγών στον 2^2 σχεδιασμό

Στο σχήμα 2.1 παρατηρούμε τους τέσσερις συνδυασμούς των αγωγών. Η υψηλή στάθμη ενός παράγοντα συμβολίζεται με το αντίστοιχο μικρό γράμμα, ενώ η χαμηλή στάθμη με την απουσία γράμματος. Επίσης, το σύμβολο (1) αντιστοιχίζεται στην περίπτωση που και οι δύο παράγοντες βρίσκονται στη χαμηλή στάθμη. Ο παραπάνω συμβολισμός χρησιμοποιείται σε όλους τους σχεδιασμούς 2^k .

2.2.1.2 Ένα παράδειγμα 2^4 παραγοντικού σχεδιασμού

(Design & Analysis of Experiments 2009 Montgomery)

Σε ένα παραγοντικό πείραμα εξετάζεται το κατά πόσο επηρεάζεται ο ρυθμός διήθησης (*filtration rate*) της ρητίνης από τους εξής τέσσερις παράγοντες: A = θερμοκρασία, B = πίεση, C = συγκέντρωση της formaldehyde και D = ρυθμός ανάδευσης. Ο κάθε παράγοντας παίρνει τιμές σε δύο στάθμες μία χαμηλή και μία

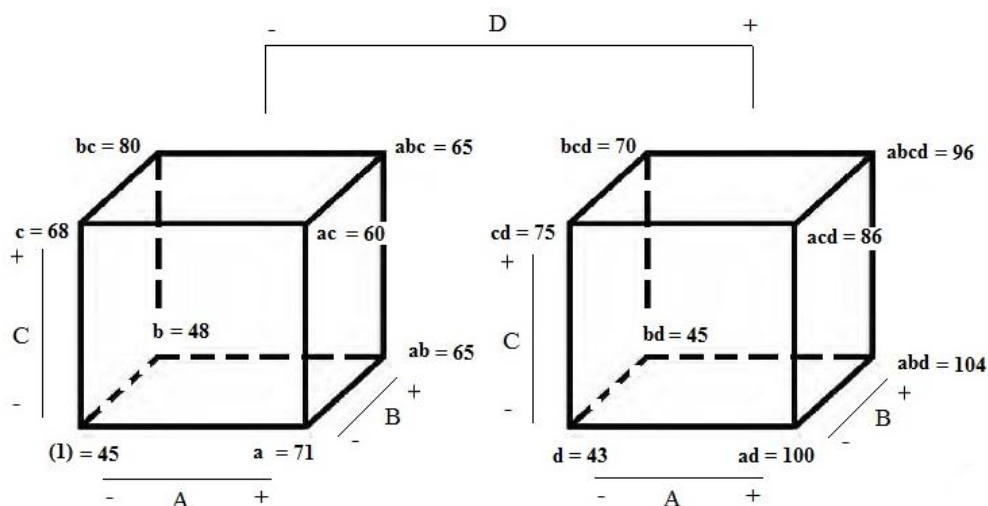
υψηλή. Συνεπώς, για τον εντοπισμό των παραγόντων ή των συνδυασμών παραγόντων που επηρεάζουν περισσότερο το πείραμα θα χρησιμοποιηθεί ένας 2^4 παραγοντικός σχεδιασμός.

Σκοπός του μηχανικού της διαδικασίας αποτελεί η μεγιστοποίηση του ρυθμού διήθησης της ρητίνης, περιορίζοντας, όμως, όσο το δυνατόν περισσότερο τη συγκέντρωση της formaldehyde.

Στον Πίνακα 2.2 παρουσιάζονται τα δεδομένα που προκύπτουν από μία επανάληψη του 2^4 παραγοντικού σχεδιασμού. Γενικότερα, σε μία επανάληψη ενός 2^k παραγοντικού σχεδιασμού πρέπει να πραγματοποιηθούν 2^k εκτελέσεις. Συγκεκριμένα, θα γίνουν k εκτελέσεις που θα μετράνε την κύρια επίδραση των k παραγόντων, μία εκτέλεση στην οποία και οι k παράγοντες θα λαμβάνουν τιμές που αντιστοιχούν στην χαμηλή τους στάθμη, $\binom{k}{2} = \frac{k!}{(k-2)! 2!}$ εκτελέσεις που εξετάζουν την αλληλεπίδραση δύο παραγόντων, $\binom{k}{3}$ που αφορούν την αλληλεπίδραση τριών παραγόντων, κ.ο. Στο παράδειγμα που μελετάται οι 16 εκτελέσεις έγιναν με τυχαία σειρά.

Pilot Plant Filtration Rate Experiment						
Run Number	Factor				Run Label	Filtration Rate (gal/h)
	A	B	C	D		
1	-	-	-	-	(1)	45
2	+	-	-	-	a	71
3	-	+	-	-	b	48
4	+	+	-	-	ab	65
5	-	-	+	-	c	68
6	+	-	+	-	ac	60
7	-	+	+	-	bc	80
8	+	+	+	-	abc	65
9	-	-	-	+	d	43
10	+	-	-	+	ad	100
11	-	+	-	+	bd	45
12	+	+	-	+	abd	104
13	-	-	+	+	cd	75
14	+	-	+	+	acd	86
15	-	+	+	+	bcd	70
16	+	+	+	+	abcd	96

Πίνακας 2.2



Σχήμα 2.2

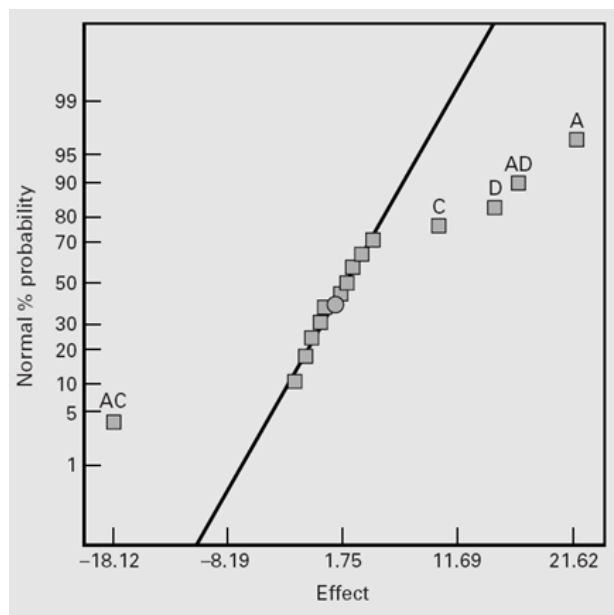
Στο σχήμα 2.2 φαίνεται η γεωμετρική ερμηνεία των συνδυασμών των παραγόντων ενώ ο πίνακας με τα θετικά και τα αρνητικά πρόσημα για τις σταθερές των αντιθέσεων για τον 2^4 παραγοντικό σχεδιασμό παρουσιάζονται στον Πίνακα 2.3.

Contrast Constants for the 2^4 Design															
	A	B	AB	C	AC	BC	ABC	D	AD	BD	ABD	CD	ACD	BCD	ABCD
(1)	-	-	+	-	+	+	-	-	+	+	-	+	-	-	+
a	+	-	-	-	-	+	+	-	-	+	+	+	+	-	-
b	-	+	-	-	+	-	+	-	+	-	+	+	-	+	-
ab	+	+	+	-	-	-	-	-	-	-	-	+	+	+	+
c	-	-	+	+	-	-	+	-	+	+	-	-	+	+	-
ac	+	-	-	+	+	-	-	-	-	+	+	-	-	+	+
bc	-	+	-	+	-	+	-	-	+	-	+	-	+	-	+
abc	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-
d	-	-	+	-	+	+	-	+	-	-	+	-	+	+	-
ad	+	-	-	-	+	+	+	+	+	-	-	-	+	+	+
bd	-	+	-	-	+	-	+	+	-	+	-	-	+	-	+
abd	+	+	+	-	-	-	-	+	+	+	+	-	-	-	-
cd	-	-	+	+	-	-	+	+	-	-	+	+	-	-	+
acd	+	-	-	+	+	-	-	+	+	-	-	+	+	-	-
bcd	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
abcd	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Πίνακας 2.3

Τέλος, κάνοντας χρήση του κανονικού διαγράμματος (Σχήμα 2.3) μπορούμε να αποφανθούμε για τη σημαντικότητα των επιδράσεων και των αλληλεπιδράσεων. Συγκεκριμένα, όλες οι επιδράσεις και αλληλεπιδράσεις που

βρίσκονται πάνω στην ευθεία γραμμή είναι αμελητέες, ενώ οι επιδράσεις και οι αλληλεπιδράσεις που βρίσκονται μακριά από την γραμμή είναι οι πιο σημαντικές. Σύμφωνα με την παραπάνω ανάλυση στο συγκεκριμένο παράδειγμα οι κύριες επιδράσεις που επηρεάζουν περισσότερο την απόκριση είναι οι A, D και C, ενώ οι πιο σημαντικές αλληλεπιδράσεις είναι οι AC και AD.



Σχήμα 2.3

Παρατηρείται ότι η επίδραση των παραγόντων A, C, D είναι θετική αυτό σημαίνει ότι αν αυξήσουμε τους παράγοντες από την χαμηλή στην υψηλή στάθμη θα αυξηθεί η απόδοση. Συνεπώς, ο ρυθμός διήθησης μεγιστοποιείται αν τεθούν οι παραπάνω τρεις παράγοντες στην υψηλή στάθμη. Εντούτοις, πρέπει να ληφθούν υπ' όψιν και οι σημαντικές αλληλεπιδράσεις. Δεδομένου ότι η αλληλεπίδραση AC είναι αρνητική συμπεραίνεται ότι η επίδραση της θερμοκρασίας (A) είναι πολύ μικρή όταν η συγκέντρωση (C) είναι στην υψηλή στάθμη και πολύ μεγάλη όταν η συγκέντρωση είναι στην χαμηλή στάθμη, στην οποία περίπτωση έχουμε καλύτερα αποτελέσματα. Τέλος, η αλληλεπίδραση AD δείχνει ότι ο ρυθμός ανάδευσης (D) έχει μικρή επίδραση σε μικρή θερμοκρασία και μεγάλη θετική επίδραση σε υψηλή θερμοκρασία. Συνεπώς, η καλύτερη απόκριση φαίνεται να εμφανίζεται όταν οι παράγοντες A, D παίρνουν τιμές στην υψηλή στάθμη και ο παράγοντας C στην χαμηλή. Με αυτό τον συνδυασμό ο πειραματιστής μπορεί να πετύχει τον

καλύτερο ρυθμό διήθησης της ρητίνης περιορίζοντας, επιπλέον, τη συγκέντρωση της formaldehyde.

2.2.2 Ο 2^k κλασματικός παραγοντικός σχεδιασμός

Σε πολλές περιπτώσεις για μια πλήρη επανάληψη ενός 2^k παραγοντικού σχεδιασμού χρειάζεται να γίνουν αρκετές εκτελέσεις, κάτι που αυξάνει αρκετά τόσο τη χρονική διάρκεια του πειράματος, όσο και το κόστος του. Παραδείγματος χάριν, ας θεωρήσουμε ένα πείραμα γεωπονικής φύσεως στόχος του οποίου είναι η μελέτη της επίδρασης δέκα θρεπτικών συστατικών στην παραγωγή ρυζιού. Αν ο πειραματιστής χρησιμοποιήσει έναν πλήρη παραγοντικό σχεδιασμό θα χρειαστούν $2^8 = 1024$ εκτελέσεις. Σε αυτόν το σχεδιασμό 10 βαθμοί ελευθερίας αντιστοιχούν σε κύριες επιδράσεις, 45 σε αλληλεπιδράσεις δύο παραγόντων και 120 σε αλληλεπιδράσεις τριών παραγόντων. Οι υπόλοιποι βαθμοί ελευθερίας αντιστοιχούν σε αλληλεπιδράσεις περισσότερων παραγόντων, ενώ οι συνολικοί βαθμοί ελευθερίας είναι 1023. Σε ένα τόσο μεγάλο πείραμα είναι εύγλωττο ο πειραματιστής να αντιμετωπίσει δυσκολίες τόσο με το κόστος του πειράματος όσο και με τη χρονική διάρκεια εκτέλεσης του.

Καταλήγοντας, οι πλήρεις παραγοντικοί σχεδιασμοί είναι εύχρηστοι μόνο σε περιπτώσεις όπου το πείραμα περιλαμβάνει λίγους παράγοντες. Στις περισσότερες περιπτώσεις ο πειραματιστής αναγκάζεται να συμπεριλάβει μόνο μερικούς από τους συνδυασμούς παραγόντων, να εκτελέσει δηλαδή μόνο ένα κλάσμα του πλήρους παραγοντικού σχεδιασμού. Το 1945 ο Finney εισήγαγε την έννοια της κλασματικής επαναληψής (*fractional replication*) η οποία ορίζεται ως η τεχνική ανάκτησης χρήσιμων πληροφοριών σε λογικό βαθμό ακρίβειας παρατηρώντας μόνο ένα μέρος του πλήρους παραγοντικού σχεδιασμού. Ο κλασματικός παραγοντικός σχεδιασμός συμβολίζεται ως 2^{k-p} σχεδιασμός με $p < k$, όπου λαμβάνονται υπόψη k παράγοντες και ο αριθμός επαναλήψεων είναι 2^{k-p} .

Οι κλασματικοί παραγοντικοί σχεδιασμοί εφαρμόζονται κυρίως σε πειράματα κρησαρίσματος (*screening experiments*). Στα πειράματα αυτά θεωρείται ότι μετέχουν αρκετοί παράγοντες και σκοπός είναι η αναγνώριση εκείνων των παραγόντων που έχουν μεγάλες επιδράσεις. Τα πειράματα κρησαρίσματος συνήθως εκτελούνται στα αρχικά στάδια μιας έρευνας, όταν είναι πιθανόν να

έχουν θεωρηθεί αρκετοί παράγοντες με μικρή ή μηδενική επίδραση στην απόκριση. Οι υπόλοιποι παράγοντες που έχουν θεωρηθεί σημαντικοί ερευνώνται περαιτέρω στα επόμενα πειράματα.

2.2.2.1 Το ένα δεύτερο κλάσμα του 2^k παραγοντικού σχεδιασμού

Ας υποθέσουμε μία κατάσταση στην οποία πρέπει να εξεταστούν τρεις παράγοντες οι οποίοι παίρνουν τιμές σε δύο στάθμες. Ο πειραματιστής θα πρέπει να εξετάσει συνολικά και τους $2^3 = 8$ συνδυασμούς αγωγών αλλά έχει τη δυνατότητα μόνο για τέσσερις εκτελέσεις. Συνεπώς θα οδηγηθεί στη χρήση ενός κλασματικού παραγοντικού σχεδιασμού και συγκεκριμένα στον 2^{3-1} παραγοντικό σχεδιασμό, το οποίο αποτελεί διαφορετική ονομασία του $\frac{1}{2}$ κλάσματος του 2^3 παραγοντικού σχεδιασμού.

Συνδυασμός Αγωγών	Παραγοντική επίδραση							
	I	A	B	C	AB	AC	BC	ABC
a	+	+	-	-	-	-	+	+
b	+	-	+	-	-	+	-	+
c	+	-	-	+	+	-	-	+
abc	+	+	+	+	+	+	+	+
ab	+	+	+	-	+	-	-	-
ac	+	+	-	+	-	+	-	-
bc	+	-	+	+	-	-	+	-
(1)	+	-	-	-	+	+	+	-

Πίνακας 2.4

Στον παραπάνω πίνακα φαίνονται οι εκτελέσεις του 2^3 παραγοντικού σχεδιασμού. Οι στήλες αλληλεπίδρασης των μεταβλητών προκύπτουν με πολλαπλασιασμό των αντίστοιχων στηλών και ουσιαστικά χρειάζονται για την ανάλυση των παρατηρήσεων και όχι για το σχεδιασμό αυτών.

Ο παραπάνω πίνακας είναι χωρισμένος σύμφωνα με τα δύο κύρια κλάσματα που προκύπτουν. Ο 2^{3-1} παραγοντικός σχεδιασμός σχηματίζεται επιλέγοντας εκείνους τους συνδυασμούς αγωγών που έχουν θετικό πρόσημο στην στήλη ABC, η οποία ονομάζεται γεννήτορας (*generator*) του συγκεκριμένου κλάσματος. Γενικότερα, γεννήτορας ονομάζεται η στήλη εκείνη που υποδεικνύει πιο $\frac{1}{2}$ κλάσμα θα χρησιμοποιηθεί. Μερικές φορές γίνεται αναφορά σε ένα γεννήτορα σαν μία λέξη.

Ένας 2^{k-p} κλασματικός παραγοντικός σχεδιασμός ορίζεται μέσω ενός συνόλου p ανεξάρτητων μεταβλητών. Οι λέξεις που αντιπροσωπεύουν αυτές τις p μεταβλητές ονομάζονται ανεξάρτητες ορίζουσες λέξεις (*independent defining words*). Τα γινόμενα των ανεξάρτητων ορίζουσών λέξεων ονομάζονται ορίζουσες λέξεις (*defining words*). Σημειώνεται ότι τα γινόμενα προκύπτουν από απαλοιφή των γραμμών που εμφανίζονται δύο φορές. Συνολικά υπάρχουν $2^p - 1$ ορίζουσες λέξεις, ενώ οι αντίστοιχες παραγοντικές επιδράσεις καλούνται ορίζουσες επιδράσεις (*defining effects*) του κλάσματος. Οι ορίζουσες επιδράσεις αποτελούν την ορίζουσα σχέση (*defining relation*). Γενικότερα, η ορίζουσα σχέση για έναν κλασματικό παραγοντικό σχεδιασμό θα είναι πάντοτε το σύνολο όλων των στηλών που είναι ίσες με τη μοναδιαία στήλη I. Στο σχεδιασμό μας η σχέση $I=ABC$ αποτελεί την ορίζουσα σχέση. Τέλος, οι υπόλοιπες $2^k - 2^p$ παραγοντικές επιδράσεις χωρίζονται σε $2^{k-p} - 1$ ομάδες ταυτόσημων επιδράσεων (*alias sets*) μεγέθους 2^p . Για την εύρεση των ταυτόσημων επιδράσεων αρκεί να πολλαπλασιαστεί οποιαδήποτε στήλη με την ορίζουσα σχέση. Στο παράδειγμα που μελετάται ισχύει:

$A \cdot I = A \cdot ABC = A^2BC \Leftrightarrow A = BC$, διότι το τετράγωνο οποιασδήποτε στήλης είναι ίσο με τη μοναδιαία I.

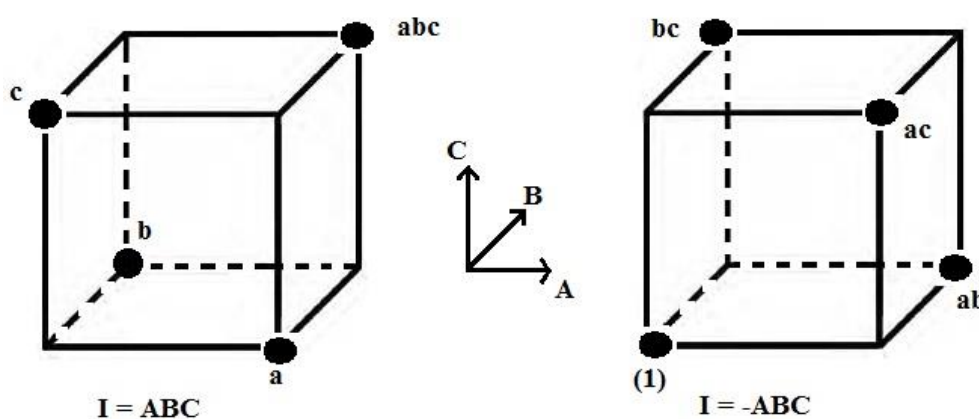
$$B \cdot I = B \cdot ABC = AB^2C \Leftrightarrow B = AC$$

$$C \cdot I = C \cdot ABC = ABC^2 \Leftrightarrow C = AB$$

Το $\frac{1}{2}$ κλάσμα του 2^3 παραγοντικού σχεδιασμού με $I = ABC$ ονομάζεται κύριο (*principle*) κλάσμα. Θα μπορούσε να είχε χρησιμοποιηθεί και το εναλλακτικό ή

συμπληρωματικό $\frac{1}{2}$ κλάσμα, το οποίο φαίνεται στο δεύτερο μισό του Πίνακα 2.4. Η ορίζουσα σχέση του συγκεκριμένου σχεδιασμού δίνεται από την ισότητα: $I = -ABC$, ενώ οι ταυτόσημες επιδράσεις είναι οι εξής : $BC = -A$, $AC = -B$ και $AB = -C$.

Στην πράξη, τον πειραματιστή δεν τον ενδιαφέρει πιο από τα δύο κλάσματα θα χρησιμοποιηθεί. Και τα δύο κλάσματα ανήκουν στην ίδια οικογένεια, καθώς και τα δύο $\frac{1}{2}$ κλάσματα αποτελούν έναν πλήρη 2^3 σχεδιασμό. Η γεωμετρική ερμηνεία των δύο κλασμάτων του 2^3 σχεδιασμού φαίνονται στο Σχήμα 2. 4.



Σχήμα 2.4

2.2.3 Αναλυτική τάξη σχεδιασμού (design resolution)

Ένας σχεδιασμός είναι αναλυτικής τάξης R, αν καμία επίδραση p παραγόντων δεν είναι ταυτόσημη με άλλη επίδραση που περιέχει R-p παράγοντες. Ο παραπάνω ορισμός εισήχθη από τους Box και Hunter το 1961. Η χρήση ενός συμβόλου Ρωμαϊκής επίδρασης, συνήθως, υποδηλώνει την αναλυτική τάξη του σχεδιασμού. Επομένως, το $\frac{1}{2}$ κλάσμα του 2^3 σχεδιασμού με ορίζουσα σχέση την $I = ABC$ (ή την $I = -ABC$) αποτελεί ένα σχεδιασμό III αναλυτικής τάξης και συμβολίζεται ως 2_{III}^{3-1} . Ειδικότερα αναφέρουμε (*Fractional Factorial Plans, Alope Dey*):

Σχεδιασμοί αναλυτικής τάξης III:

Αυτοί είναι σχεδιασμοί στους οποίους καμία κύρια επίδραση δεν είναι ταυτόσημη με κάποια άλλη κύρια επίδραση, αλλά είναι ταυτόσημες με αλληλεπιδράσεις δύο παραγόντων. Επίσης, οι αλληλεπιδράσεις δύο παραγόντων μπορεί να είναι ταυτόσημες μεταξύ τους.

Όπως αναφέρθηκε και προηγουμένως, ο 2^{3-1} σχεδιασμός του Πίνακα 2.4 είναι αναλυτικής τάξης III (2_{III}^{3-1}).

Σχεδιασμοί αναλυτικής τάξης IV:

Στους συγκεκριμένους σχεδιασμούς καμία κύρια επίδραση δεν είναι ταυτόσημη με άλλη κύρια επίδραση ή με αλληλεπίδραση δύο παραγόντων. Οι αλληλεπιδράσεις δύο παραγόντων είναι ταυτόσημες με άλλες αλληλεπιδράσεις δύο παραγόντων.

Χαρακτηριστικό παράδειγμα αποτελεί ο 2^{4-1} παραγοντικός σχεδιασμός με ορίζουσα σχέση $I = ABCD$ (2_{IV}^{4-1}).

Σχεδιασμοί αναλυτικής τάξης V:

Καμία κύρια επίδραση ή αλληλεπίδραση δύο παραγόντων δεν είναι ταυτόσημη με οποιαδήποτε άλλη κύρια επίδραση ή αλληλεπίδραση δύο παραγόντων. Όμως, οι τελευταίες είναι ταυτόσημες με αλληλεπιδράσεις τριών παραγόντων. Παράδειγμα σχεδιασμού αναλυτικής τάξης V αποτελεί ένας 2^{5-1} σχεδιασμός με $I=ABCDE$ και συμβολίζεται 2_V^{5-1} .

Γενικά, η αναλυτική τάξη ενός κλασματικού παραγοντικού σχεδιασμού με δύο στάθμες είναι ίση με το μικρότερο μήκος οποιασδήποτε λέξης στην ορίζουσα σχέση, δηλαδή με το μικρότερο αριθμό γραμμάτων που παρουσιάζεται στις συγκεκριμένες λέξεις. Συνεπώς, οι παραπάνω ορισμοί μπορούν να ονομαστούν ως σχεδιασμοί τριών, τεσσάρων και πέντε γραμμάτων αντίστοιχα. Συνήθως, προτιμάται η χρήση κλασματικών παραγοντικών σχεδιασμών με την υψηλότερη δυνατή αναλυτική τάξη. Με αυτό τον τρόπο, οι περιορισμοί τους οποίους πρέπει να λάβει υπ' όψιν του ο πειραματιστής για το ποιες αλληλεπιδράσεις θα

θεωρηθούν αμελητέες έτσι ώστε να λάβει μια μοναδική ερμηνεία των δεδομένων, ελαχιστοποιούνται.

2.2.4 Ο 3^k παραγοντικός σχεδιασμός (full factorial design)

Σε αντιστοιχία με τον 2^k παραγοντικό σχεδιασμό, ο 3^k είναι ένας παραγοντικός σχεδιασμός που λαμβάνει υπ' όψιν k παράγοντες καθένας από τους οποίους παίρνει τιμές σε τρεις στάθμες και απαιτεί 3^k επαναλήψεις. Συνήθως, γίνεται αναφορά στις τρεις στάθμες των παραγόντων ως χαμηλή, μέτρια και υψηλή. Υπάρχουν αρκετοί τρόποι για να συμβολιστούν οι τρεις στάθμες των k παραγόντων, ο συνηθέστερος είναι με την χρήση των ψηφίων 0 (χαμηλή), 1 (μέτρια) και 2 (υψηλή). Κάθε συνδυασμός αγωγών στον 3^k παραγοντικό σχεδιασμό θα συμβολίζεται με k ψηφία, όπου το πρώτο ψηφίο θα υποδεικνύει τη στάθμη του παράγοντα A, το δεύτερο θα δείχνει τη στάθμη του παράγοντα B,..., και το k ψηφίο θα υποδεικνύει τη στάθμη του παράγοντα K.

Για παράδειγμα, σε έναν 3^2 παραγοντικό σχεδιασμό έχουμε τους συνδυασμούς αγωγών που φαίνονται στο Πίνακα 2.5.

		B		
		0	1	2
A	0	00	01	02
	1	10	11	12
	2	20	21	22

Πίνακας 2.5

Στον παραπάνω πίνακα ο συμβολισμός 00 αναφέρεται στην αλληλεπίδραση των παραγόντων A και B όταν και οι δύο λαμβάνουν τιμές που αντιστοιχούν στη χαμηλή στάθμη, ενώ το 01 συμβολίζει τον συνδυασμό αγωγών που αντιστοιχεί στον A στην χαμηλή στάθμη και στον B στην μεσαία.

Το σύστημα συμβολισμού που μόλις προαναφέρθηκε, θα μπορούσε να χρησιμοποιηθεί και για την περιγραφή των 2^k σχεδιασμών που παρουσιάστηκαν στην προηγούμενη παράγραφο, με το 0 να συμβολίζει την χαμηλή στάθμη και το

1 την υψηλή στάθμη των παραγόντων. Εντούτοις, προτιμάται ο συμβολισμός \pm ή ± 1 , καθώς διευκολύνει τη γεωμετρική ερμηνεία του σχεδιασμού και επιπλέον, μπορεί να εφαρμοστεί άμεσα στο μοντέλο παλινδρόμησης, στην ομαδοποίηση και στην κατασκευή των κλασματικών παραγοντικών σχεδιασμών.

2.2.4.1 Ένα παράδειγμα 3^4 παραγοντικού σχεδιασμού

Ας θεωρήσουμε ένα πείραμα κατά το οποίο εξετάζεται η επίδραση τεσσάρων παραγόντων στην αντοχή των παραγόμενων ζωνών ασφαλείας φορτηγών. Οι τέσσερις παράγοντες είναι οι εξής:

A: η ξαφνική πίεση που ασκείται στη ζώνη

B: το νεκρό μήκος της ζώνης

C: το μήκος του καλωδίου που συνδέει τη ζώνη στον υποδοχέα

D: ο τύπος του υποδοχέα

Κάθε παράγοντας παίρνει τιμές σε τρεις στάθμες, οι οποίες συμβολίζονται με τα ψηφία 0, 1 και 2 και παρουσιάζονται στον Πίνακα 2.6. Συνεπώς, ο σχεδιασμός μας είναι ένας 3^4 παραγοντικός σχεδιασμός και απαιτούνται $3^4 = 27$ εκτελέσεις για την ολοκλήρωση του πειράματος. Όπως παρατηρείται και στον Πίνακα 2.7 η εκτέλεση κάθε συνδυασμού παραγόντων επαναλαμβάνεται τρεις φορές.

Παράγοντας (Factor)	Στάθμη (level)		
	0	1	2
Πίεση (psi)	1100	1400	1700
Νεκρό μήκος (mm)	10	10,2	10,4
Μήκος καλωδίου (mm)	18	23	27
Τύπος υποδοχέα (αριθμός)	P74	P75	P76

Πίνακας 2.6

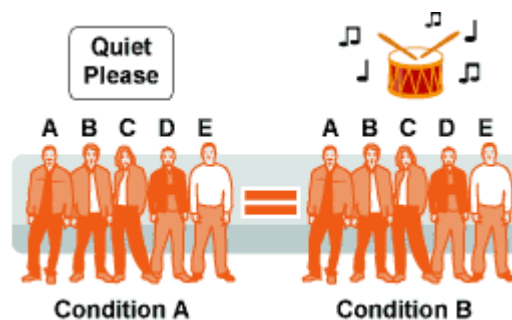
Εκτέ- λεση (run)	Παράγοντας (factor)				Δύναμη			Flash		
	A	B	C	D						
1	0	0	0	0	5164	6615	5959	12.89	12.70	12.74
2	0	0	1	1	5356	6117	5224	12.83	12.73	13.07
3	0	0	2	2	3070	3773	4257	12.37	12.47	12.44
4	0	1	0	1	5547	6566	6320	13.29	12.86	12.70
5	0	1	1	2	4754	4401	5436	12.64	12.50	12.61
6	0	1	2	0	5524	4050	4526	12.76	12.72	12.94
7	0	2	0	2	5684	6251	6214	13.17	13.33	13.98
8	0	2	1	0	5735	6271	5843	13.02	13.11	12.67
9	0	2	2	1	5744	4797	5416	12.37	12.67	12.54
10	1	0	0	1	6843	6895	6957	13.28	13.65	13.58
11	1	0	0	1	6843	6895	6957	13.28	13.65	13.58
12	1	0	2	0	6152	5819	5963	13.19	12.94	13.15
13	1	1	0	2	6854	6804	6907	14.65	14.98	14.40
14	1	1	1	0	6799	6703	6792	13.00	13.35	12.87
15	1	1	2	1	6513	6503	6568	13.13	13.40	13.80
16	1	2	0	0	6473	6974	6712	13.55	14.10	14.41
17	1	2	1	1	6832	7034	5057	14.86	13.27	13.64
18	1	2	2	2	4968	5684	5761	13	13.58	13.45
19	2	0	0	2	7148	6920	6220	16.70	15.85	14.90
20	2	0	1	0	6905	7068	7156	14.70	13.97	13.66
21	2	0	2	1	6933	7194	6667	13.51	13.64	13.92
22	2	1	0	0	7227	7170	7015	15.54	16.16	16.14
23	2	1	1	1	7014	7040	7200	13.97	14.09	14.52
24	2	1	2	2	6215	6260	6488	14.35	13.56	13
25	2	2	0	1	7145	6868	6964	15.70	16.45	15.85
26	2	2	1	2	7161	1263	6937	15.21	13.77	14.34
27	2	2	2	0	7060	7050	6950	13.51	13.42	13.07

Στο πείραμα υπάρχουν δύο μεταβλητές απόκρισης. Η πρώτη είναι η δύναμη αντοχής του καλωδίου που συνδέει τον ιμάντα με τον υποδοχέα. Μονάδες μέτρησης της δύναμης είναι το lb ($1lb \approx 4.448222 N$) και η κατώτατη τιμή που έχει οριστεί από τον πειραματιστή είναι 4000 lb. Η δεύτερη μεταβλητή απόκρισης είναι η ανάφλεξη (mm), η οποία ορίζεται ως το παραπανίσιο μέταλλο που προκύπτει από την διαδικασία στρεσαρίσματος της ζώνης. Η τιμή της ανάφλεξης δεν μπορεί να υπερβεί τα 14mm. Η δύναμη πρέπει να μεγιστοποιηθεί έτσι ώστε να είναι σίγουρο ότι η ζώνη δεν θα διαλυθεί σε περίπτωση ατυχήματος. Αντιθέτως η ανάφλεξη θα πρέπει να ελαχιστοποιηθεί καθώς προκαλεί ρωγμές στο πλαστικό κάλυμμα του υποδοχέα. Αρχικά, οι προτεινόμενες ρυθμίσεις ήταν η πίεση να παίρνει τιμές στην υψηλή στάθμη και το μήκος του καλωδίου στη μεσαία. Συνεπώς, στόχος του πειράματος είναι να βρεθεί ο κατάλληλος συνδυασμός παραγόντων έτσι ώστε να επιτευχθεί η μέγιστη δυνατή δύναμη και η ελάχιστη ανάφλεξη λαμβάνοντας υπ' όψιν τους παραπάνω περιορισμούς. Ειδικότερα, η επιθυμητή μέση τιμή της δύναμης ισούται με 7000 lb με τυπική απόκλιση 250 lb και η μέση τιμή της ανάφλεξης με 13,50 mm με τυπική απόκλιση 0,2 mm.

2.2.5 Σχεδιασμός επαναλαμβανόμενων μετρήσεων (repeated measurements design)

Ο σχεδιασμός επαναλαμβανόμενων μετρήσεων είναι ένας πειραματικός σχεδιασμός που περιλαμβάνει τον έλεγχο ενός συνόλου μονάδων δύο ή περισσότερες φορές.

Παραδείγματος χάριν, σε ένα πείραμα οι συμμετέχοντες καλούνται να συμπληρώσουν κάποια test μνήμης κάτω από διαφορετικές συνθήκες θορύβου.



Το βασικό πλεονέκτημα των σχεδιασμών επαναλαμβανόμενων μετρήσεων είναι ότι μειώνει την επίδραση των παραγόντων θορύβου. Οι παράγοντες θορύβου (*noise factors*) αποτελούν μη ελεγχόμενους παράγοντες οι οποίοι συνδέονται συνήθως με το περιβάλλον της παραγωγής ή της λειτουργίας. Οι παράγοντες αυτοί διακρίνονται στους εξωτερικούς παράγοντες θορύβου, όπως η υγρασία, η θερμοκρασία περιβάλλοντος κ.α. και στους εσωτερικούς παράγοντες, όπως οι αποκλίσεις των μηχανημάτων από τις προδιαγραφές της παραγωγικής διαδικασίας. Επίσης, οι σχεδιασμοί επαναλαμβανόμενων μετρήσεων απαιτούν μικρότερο αριθμό παρατηρούμενων μονάδων, δεδομένου ότι όλες οι επαναλήψεις του πειράματος εξετάζουν το ίδιο σύνολο μονάδων. Οι σχεδιασμοί αυτοί χρησιμοποιούνται, αρκετά, στις κλινικές δοκιμές ή στη φαρμακευτική βιομηχανία, όταν ο σκοπός είναι η βελτίωση χρόνιων παθήσεων και όχι η θεραπεία.

Οι σχεδιασμοί επαναλαμβανόμενων μετρήσεων παρουσιάζουν, επίσης, και ορισμένα μειονεκτήματα. Αρχικά, τα πειράματα διαρκούν περισσότερο χρόνο και υπάρχει ο κίνδυνος αποχώρησης ατόμων που συμμετέχουν στο πείραμα. Στη περίπτωση που κάτι τέτοιο συμβεί το δείγμα που βρίσκεται υπό εξέταση είναι αρκετά μικρό και δεν μπορεί να θεωρηθεί στατιστικά σημαντικό. Επίσης, υπάρχουν αρκετές απειλές για την εσωτερική εγκυρότητα του σχεδιασμού. Πιο συγκεκριμένα, όταν οι ίδιες μονάδες εξετάζονται αρκετές φορές οι τιμές τους τείνουν να παλινδρομούν προς τη μέση τιμή. Επιπλέον, εξωτερικές αλλαγές που λαμβάνουν χώρα μεταξύ των διαφορετικών μετρήσεων μπορούν να επηρεάσουν την απόκριση του πειράματος.

ΚΕΦΑΛΑΙΟ 3:

ΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ ΑΝΑΛΥΣΗΣ ΠΑΡΑΓΟΝΤΙΚΩΝ ΣΧΕΔΙΑΣΜΩΝ

3.1 Η μέθοδος ελαχίστων τετραγώνων

Σε πολλά πειράματα υπάρχει μία γραμμική σχέση ανάμεσα στα μετρούμενα μεγέθη. Για παράδειγμα, η ταχύτητα ενός σώματος το οποίο εκτελεί ελεύθερη πτώση, μεταβάλλεται γραμμικά με το χρόνο, εφόσον αγνοηθεί η αντίσταση του αέρα. Τοποθετώντας τα σημεία σε ένα διάγραμμα, παρατηρείται ότι αυτά προσεγγίζουν μία ευθεία γραμμή. Το επόμενο βήμα είναι να βρεθεί η κλίση της ευθείας η οποία προσεγγίζει περισσότερο αυτά τα σημεία, και το σημείο στο οποίο αυτή τέμνει τον άξονα y (τεταγμένη). Σε κάθε περίπτωση, δεν αναμένεται η ευθεία να διέρχεται από όλα τα σημεία, λόγω της παρουσίας τυχαίων σφαλμάτων. Εντούτοις, υπάρχει η δυνατότητα εύρεσης προσεγγιστικών τιμών τόσο για την κλίση όσο και για την τεταγμένη, εάν σχεδιαστεί μία ευθεία η οποία να διέρχεται ανάμεσα από τα διεσπαρμένα σημεία.

Έστω ότι ένα γραμμικό μοντέλο της μορφής

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (3.1)$$

προσαρμόζεται στα δεδομένα που μελετούνται. Τα β_0 , β_1 και ε είναι άγνωστα και το ε είναι δύσκολο να βρεθεί καθώς μεταβάλλεται για κάθε παρατήρηση Y . Ωστόσο, τα β_0 και β_1 παραμένουν σταθερά και μολονότι δεν μπορούν να υπολογισθούν ακριβώς, χρησιμοποιώντας τη διαθέσιμη πληροφορία των δεδομένων μπορούν να βρεθούν οι εκτιμήσεις $\hat{\beta}_0$ και $\hat{\beta}_1$. Επομένως το γραμμικό μοντέλο μπορεί να γραφεί ως

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \quad (3.2)$$

όπου \hat{Y} η προβλεπόμενη τιμή της Y για μια δοθείσα τιμή της X , όταν οι τιμές των $\hat{\beta}_0$ και $\hat{\beta}_1$ έχουν προσδιοριστεί.

Η ακριβέστερη μέθοδος για την εκτίμηση των β_0 και β_1 είναι η *μέθοδος των ελαχίστων τετραγώνων*.

Ας υποθέσουμε ότι διαθέτουμε n σύνολα παρατηρήσεων $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ κανονικά καταναμημένα με μέση τιμή μ και διασπορά σ^2 . Τότε η αθροιστική συνάρτηση κατανομής των X_i με $i = 1, 2, \dots, n$ εκφράζεται ως εξής:

$$F(X_i) = \Phi\left(\frac{X_i - \mu}{\sigma}\right)$$

Όπου η συνάρτηση $\Phi()$ αποτελεί την τυπική αθροιστική συνάρτηση της κανονικής κατανομής. Από την παραπάνω εξίσωση προκύπτει η εκτιμήτρια:

$$\hat{F}(X_i) = \Phi\left(\frac{X_i - \hat{\mu}}{\hat{\sigma}}\right)$$

για την οποία είναι γνωστό ότι ισχύει:

$$\hat{F}(X_i) = \frac{i}{n+1} \quad (\text{D'Agostino et al., 1986})$$

Η τελευταία εξίσωση γράφεται αλλιώς:

$$\Phi^{-1}(\hat{F}(X_i)) = \frac{X_i - \hat{\mu}}{\hat{\sigma}} = -\frac{\hat{\mu}}{\hat{\sigma}} + \frac{1}{\hat{\sigma}} X_i, \quad (3.3)$$

Η (3.3) μπορεί να θεωρηθεί γραμμικό μοντέλο παλινδρόμησης της μορφής

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

με $\hat{Y}_i = \Phi^{-1}(\hat{F}(X_i))$, $\hat{\beta}_0 = -\frac{\hat{\mu}}{\hat{\sigma}}$ και $\hat{\beta}_1 = \frac{1}{\hat{\sigma}}$.

Επομένως τα δεδομένα σύμφωνα με την εξίσωση (3.1) μπορούν να γραφούν ως

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \text{ για } i = 1, 2, \dots, n.$$

Σύμφωνα με την μέθοδο ελαχίστων τετραγώνων οι εκτιμήτριες των β_0 και β_1 είναι εκείνες οι τιμές που θα μας δώσουν την ελάχιστη δυνατή τιμή του αθροίσματος των τετραγώνων των αποκλίσεων από την πραγματική γραμμή. Δηλαδή οι $\hat{\beta}_0$ και $\hat{\beta}_1$ θα πρέπει να ικανοποιούν την εξίσωση:

$$\min \sum_{i=1}^n \varepsilon_i^2 = \min \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2. \quad (3.4)$$

Με μερική παραγωγή της 3.4 ως προς β_0 και β_1 έχουμε τις εκτιμήτριες:

$$\hat{\beta}_1 = \frac{\sum X_i Y_i - [\sum X_i][\sum Y_i]/n}{\sum X_i^2 - (\sum X_i)^2/n} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \quad (3.5)$$

και

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \quad (3.6)$$

Για την περιγραφή της εκτιμήτριας $\hat{\beta}_1$ αρκετά συχνά χρησιμοποιείται ένας διαφορετικός συμβολισμός. Συγκεκριμένα το μέγεθος S_{XY} ορίζεται ως:

$$\begin{aligned} S_{XY} &= \sum (X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \sum (X_i - \bar{X})Y_i \\ &= \sum X_i(Y_i - \bar{Y}) \\ &= \sum X_i Y_i - \frac{[(\sum X_i)(\sum Y_i)]}{n} \\ &= \sum X_i Y_i - n\bar{X}\bar{Y} \end{aligned}$$

Αντίστοιχα, το μέγεθος S_{XX} ορίζεται ως:

$$\begin{aligned} S_{XX} &= \sum (X_i - \bar{X})^2 \\ &= \sum (X_i - \bar{X})X_i \\ &= \sum X_i^2 - \frac{(\sum X_i)^2}{n} \\ &= \sum X_i^2 - n\bar{X}^2 \end{aligned}$$

και το μέγεθος S_{YY} ορίζεται ως:

$$\begin{aligned} S_{YY} &= \sum (Y_i - \bar{Y})^2 \\ &= \sum (Y_i - \bar{Y})Y_i \\ &= \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} \\ &= \sum Y_i^2 - n\bar{Y}^2 \end{aligned}$$

Συνεπώς σύμφωνα με τους παραπάνω συμβολισμούς έχουμε ότι $\hat{\beta}_1 = S_{XY}/S_{XX}$. Ο συγκεκριμένος τύπος είναι πιο βολικός καθώς είναι πιο εύκολος να τον θυμάται κανείς.

Τέλος, αντικαθιστώντας την σχέση 3.6 στην σχέση 3.2 παίρνουμε την εκτίμηση της εξίσωσης παλινδρόμησης που είναι η εξής :

$$\hat{Y} = \bar{Y} + \hat{\beta}_1(X - \bar{X}).$$

Αρκετό ενδιαφέρον παρουσιάζει το γεγονός ότι υπάρχει μια διαφωνία σχετικά με το ποιος πρώτος ανακάλυψε τη μέθοδο ελαχίστων τετραγώνων. Φαίνεται ότι η μέθοδος ανακαλύφθηκε ανεξάρτητα από τους Carl Friedrich Gauss (1777-1855) και Adrien Marie Legendre (1752-1833). Ο Gauss φαίνεται ότι άρχισε να τη χρησιμοποιεί πριν από το 1803, ο ίδιος βέβαια υποστηρίζει ότι άρχισε να τη χρησιμοποιεί το 1795 αλλά κάτι τέτοιο δεν είναι επιβεβαιωμένο, ενώ η πρώτη ανακοίνωση της δημοσιεύθηκε το 1805 από το Legendre. Όταν ο Gauss το 1809 έγραψε ότι είχε χρησιμοποιήσει τη μέθοδο νωρίτερα από την ημερομηνία δημοσίευσης της, άρχισε μια αμφισβήτηση για την προτεραιότητα. (Norman Draper et al., 1997)

3.2 Η μέθοδος μέγιστης πιθανοφάνειας

Ας θεωρήσουμε ένα τυχαίο δείγμα X_1, X_2, \dots, X_n από έναν πληθυσμό με άγνωστη παράμετρο $\vec{\theta} = (\theta_1, \theta_2, \dots, \theta_n) \in \theta$ και συνάρτηση πυκνότητας πιθανότητας $f(\vec{x}|\vec{\theta})$. Σκοπός είναι η εκτίμηση της παραμέτρου $\vec{\theta}$. Αν $f(x_i|\vec{\theta})$ είναι η συνάρτηση πυκνότητας πιθανότητας κάθε τιμής του τυχαίου δείγματος, τότε η από κοινού συνάρτηση πυκνότητας πιθανότητας των μεταβλητών X_1, X_2, \dots, X_n είναι

$$f(x_1, x_2, \dots, x_n|\theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdot \dots \cdot f(x_n|\theta)$$

Για συγκεκριμένες τιμές x_1, x_2, \dots, x_n τυχαίου δείγματος η παραπάνω σχέση είναι συνάρτηση μόνο της παραμέτρου θ και συμβολίζεται ως:

$$L(\theta|x_1, x_2, \dots, x_n) = f(x_1|\theta) \cdot f(x_2|\theta) \cdot \dots \cdot f(x_n|\theta) = \prod_{i=1}^n f(x_i|\theta) \quad (3.2.1)$$

Η σχέση (3.2.1) ονομάζεται συνάρτηση πιθανοφάνειας (*likelihood function*) του τυχαίου δείγματος X_1, X_2, \dots, X_n και εκφράζει το πόσο πιθανοφανείς ή διαφορετικά το πόσο σύμφωνες είναι με το συγκεκριμένο δείγμα είναι οι διάφορες τιμές της παραμέτρου θ .

Η μέθοδος της μέγιστης πιθανοφάνειας η οποία προτάθηκε από τον R.A. Fisher, συνίσταται στην επιλογή της τιμής $\hat{\theta}$ η οποία μεγιστοποιεί την (3.2.1).

$$L(\hat{\theta}|x_1, x_2, \dots, x_n) = \sup L(\theta|x_1, x_2, \dots, x_n).$$

Η τιμή $\hat{\theta}$ καλείται εκτιμήτρια μέγιστης πιθανοφάνειας.

Η εκτιμήτρια μέγιστης πιθανοφάνειας μεγιστοποιεί επίσης και την συνάρτηση $\log L(\theta|x_1, x_2, \dots, x_n)$, κάτι που ευκολύνει αρκετά τις πράξεις. Συνεπώς η εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n)$ προκύπτει από την λύση των εξισώσεων:

$$\frac{d \log L(\theta|x_1, x_2, \dots, x_n)}{d\theta} = 0, \quad i = 1, 2, \dots, n$$

Φυσικά, για να είναι η $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n)$ λύση μεγίστου θα πρέπει οι αντίστοιχοι πίνακες των δευτέρων μερικών παραγώγων της $\log L(\theta|x_1, x_2, \dots, x_n)$ να είναι γνήσια αρνητικοί για $\theta = \hat{\theta}$.

3.2.1 Η μέθοδος μέγιστης πιθανοφάνειας για δεδομένα που περιέχουν αποκομμένες παρατηρήσεις.

Η μέθοδος μέγιστης πιθανοφάνειας μπορεί να χρησιμοποιηθεί και στην περίπτωση που το δείγμα μας περιλαμβάνει αποκομμένες παρατηρήσεις. Στην συγκεκριμένη περίπτωση οι παρατηρήσεις $\{x_i: i = 1, \dots, n\}$ χωρίζονται στα ακόλουθα δύο υποσύνολα

$$U = \{\text{μη-αποκομμένες παρατηρήσεις}\}$$

$$C = \{\text{αποκομμένες παρατηρήσεις}\}$$

και η συνάρτηση πιθανοφάνειας δίνεται από τον εξής τύπο:

$$L(\theta|x_i: i = 1, 2, \dots, n) = L(\theta|x_i: i \in U)L(\theta|x_i: i \in C) = L_U L_C,$$

όπου $L_U = \prod_{i \in U} f(t_i)$ ενώ ο όρος L_C εξαρτάται από τη μορφή αποκοπής των δεδομένων.

Συνάρτηση πιθανοφάνειας για δεξιά αποκομμένες παρατηρήσεις

Στην περίπτωση της από δεξιά αποκοπής, η τιμή x_i αντιστοιχεί στο γεγονός $X > x_i$, συνεπώς ισχύει:

$$L_C = \prod_{i \in C} P[X > x_i] = \prod_{i \in C} S(x_i).$$

Οπότε, η συνάρτηση πιθανοφάνειας διαμορφώνεται ως εξής:

$$L = \prod_{i \in U} f(x_i) \prod_{i \in C} S(x_i) = \prod_i \{f(x_i)^{\delta_i} S(x_i)^{1-\delta_i}\} = \prod_i h(t_i)^{\delta_i} S(t_i)$$

όπου $\delta_i = \begin{cases} 1, & i \in U \\ 0, & i \in C \end{cases}$.

Συνάρτηση πιθανοφάνειας για αριστερά αποκομμένες παρατηρήσεις

Στην περίπτωση της από αριστερά αποκοπής, η τιμή x_i αντιστοιχεί στο γεγονός $X \leq x_i$, συνεπώς ισχύει:

$$L_C = \prod_{i \in C} P[X \leq x_i] = \prod_{i \in C} [1 - S(x_i)]$$

Οπότε, η συνάρτηση πιθανοφάνειας διαμορφώνεται ως εξής:

$$L = \prod_{i \in U} f(x_i) \prod_{i \in C} [1 - S(x_i)] = \prod_i \{f(x_i)^{\delta_i} [1 - S(x_i)^{1-\delta_i}]\}.$$

Συνάρτηση πιθανοφάνειας για αποκομμένες παρατηρήσεις σε διάστημα

Στην περίπτωση της αποκοπής εντός διαστήματος, $x_{a_i} < x_i \leq x_{\beta_i}$, ισχύει:

$$L_C = \prod_{i \in C} P[x_{a_i} < X \leq x_{\beta_i}] = \prod_{i \in C} [S(x_{a_i}) - S(x_{\beta_i})]$$

Οπότε, η συνάρτηση πιθανοφάνειας διαμορφώνεται ως εξής:

$$L = \prod_{i \in U} f(x_i) \prod_{i \in C} [S(x_{a_i}) - S(x_{\beta_i})].$$

Γενικότερα, υπάρχει η δυνατότητα και οι τρεις μορφές να περιληφθούν ταυτόχρονα στην συνάρτηση πιθανοφάνειας. Όπως και στην περίπτωση που τα δεδομένα μας δεν περιέχουν αποκομμένες παρατηρήσεις, έτσι και εδώ η εκτιμήτρια μέγιστης πιθανοφάνειας προκύπτει από την μεγιστοποίηση της συνάρτησης πιθανοφάνειας.

Το παρακάτω παράδειγμα θα βοηθήσει στην καλύτερη κατανόηση της μεθόδου. Ζητούνται οι εκτιμήτριες μέγιστης πιθανοφάνειας για το μοντέλο της λογαριθμο-λογιστικής κατανομής που εκτός από χρόνους διακοπής περιλαμβάνει από αριστερά και από δεξιά αποκομμένες παρατηρήσεις.

Η σ.π.π για τη Λογαριθμο-λογιστική κατανομή είναι:

$$f(t) = \frac{\alpha \gamma t^{\gamma-1}}{(1+at^\gamma)^2}, t > 0, a > 0, \gamma > 0$$

Αντίστοιχα, για τη συνάρτηση επιβίωσης ισχύει ότι:

$$S(t) = (1+at^\gamma)^{-1}, t > 0, a > 0, \gamma > 0,$$

καθώς και για τη συνάρτηση διακινδύνευσης ότι:

$$h(t) = \frac{\alpha \gamma t^{\gamma-1}}{(1+at^\gamma)}, t > 0, a > 0, \gamma > 0.$$

Η συνάρτηση πιθανοφάνειας, όταν έχουμε από δεξιά και από αριστερά αποκομμένες παρατηρήσεις, δίνεται από τη σχέση:

$$L = L_u L_{c_L} L_{c_R} = \prod_{i \in U} f(t_i) \prod_{i \in c_L} (1 - S(t_i)) \prod_{i \in c_R} S(t_i),$$

όπου

$$U = \{i \in \{1, 2, \dots, n\} : t_i \text{ μη αποκομμένη παρατήρηση}\}$$

$$C_L = \{i \in \{1, 2, \dots, n\} : t_i \text{ από αριστερά αποκομμένη παρατήρηση}\}$$

$$C_R = \{i \in \{1, 2, \dots, n\} : t_i \text{ από δεξιά αποκομμένη παρατήρηση}\}$$

Αντικαθιστώντας στο μοντέλο της Λογαριθμο-λογιστικής κατανομής θα έχουμε:

$$L = L_u L_{c_L} L_{c_R} = \prod_{i \in U} \frac{\alpha \gamma t_i^{\gamma-1}}{(1+at_i^\gamma)^2} \prod_{i \in c_L} (1 - (1+at_i^\gamma)^{-1}) \prod_{i \in c_R} (1+at_i^\gamma)^{-1} \Rightarrow$$

$$L = \prod_{i \in U} \frac{\alpha \gamma t_i^{\gamma-1}}{(1+at_i^\gamma)^2} \prod_{i \in c_L} \left(1 - \frac{1}{(1+at_i^\gamma)}\right) \prod_{i \in c_R} (1+at_i^\gamma)^{-1} \Rightarrow$$

$$L = \prod_{i \in U} \frac{\alpha \gamma t_i^{\gamma-1}}{(1 + at_i^\gamma)^2} \prod_{i \in c_L} \left(\frac{(1 + at_i^\gamma)}{(1 + at_i^\gamma)} - \frac{1}{(1 + at_i^\gamma)} \right) \prod_{i \in c_R} (1 + at_i^\gamma)^{-1} \Rightarrow$$

$$L = \prod_{i \in U} \frac{\alpha \gamma t_i^{\gamma-1}}{(1 + at_i^\gamma)^2} \prod_{i \in c_L} \frac{at_i^\gamma}{(1 + at_i^\gamma)} \prod_{i \in c_R} (1 + at_i^\gamma)^{-1} \Rightarrow$$

$$L = \prod_{i \in U} \frac{\alpha \gamma t_i^{\gamma-1}}{(1 + at_i^\gamma)^2} \prod_{i \in c_L} \frac{at_i^\gamma}{1 + at_i^\gamma} \prod_{i \in c_R} \frac{1}{1 + at_i^\gamma} \quad (3.2.2)$$

Ορίζουμε τις δείκτριες συναρτήσεις:

$$\delta_{ui} = \begin{cases} 1, & \text{αν } t_i \text{ χρόνος διακοπής} \\ 0, & \text{διαφορετικά} \end{cases}$$

$$\delta_{Li} = \begin{cases} 1, & \text{αν } t_i \text{ χρόνος αποκοπής από αριστερά} \\ 0, & \text{διαφορετικά} \end{cases}$$

$$\delta_{Ri} = \begin{cases} 1, & \text{αν } t_i \text{ χρόνος αποκοπής από δεξιά} \\ 0, & \text{διαφορετικά} \end{cases}$$

Λογαριθμίζοντας την (3.2.2) και χρησιμοποιώντας τις δείκτριες συναρτήσεις έχουμε:

$$l = \ln L = \sum_{i=1}^n \left\{ \delta_{ui} \ln \left(\frac{\alpha \gamma t_i^{\gamma-1}}{(1 + at_i^\gamma)^2} \right) + \delta_{Li} \ln \left(\frac{at_i^\gamma}{1 + at_i^\gamma} \right) + \delta_{Ri} \ln \left(\frac{1}{1 + at_i^\gamma} \right) \right\} \Rightarrow$$

$$l = \sum_{i=1}^n \left\{ \delta_{ui} [\ln(\alpha \gamma t_i^{\gamma-1}) - \ln(1 + at_i^\gamma)^2] + \delta_{Li} [\ln(at_i^\gamma) - \ln(1 + at_i^\gamma)] - \delta_{Ri} \ln(1 + at_i^\gamma) \right\}$$

$$l = \sum_{i=1}^n \{ \delta_{ui} [\ln \alpha + \ln \gamma + (\gamma - 1) \ln t_i - 2 \ln(1 + at_i^\gamma)] + \delta_{Li} [\ln a + \gamma \ln t_i - \ln(1 + at_i^\gamma)] - \delta_{Ri} \ln(1 + at_i^\gamma) \}$$

$$\Rightarrow l = \sum_{i=1}^n \{ \delta_{ui} \ln \alpha + \delta_{ui} \ln \gamma + \delta_{ui} (\gamma - 1) \ln t_i - 2 \delta_{ui} \ln(1 + at_i^\gamma) + \delta_{Li} \ln a + \delta_{Li} \gamma \ln t_i - \delta_{Li} \ln(1 + at_i^\gamma) - \delta_{Ri} \ln(1 + at_i^\gamma) \} \Rightarrow$$

$$l = \sum_{i=1}^n \{ (\delta_{ui} + \delta_{Li}) \ln \alpha + \delta_{ui} \ln \gamma + \delta_{ui} \gamma \ln t_i - \delta_{ui} \ln t_i + \delta_{Li} \gamma \ln t_i - \ln(1 + at_i^\gamma) [2 \delta_{ui} + \delta_{Li} + \delta_{Ri}] \}$$

Τελικά, η λογαριθμοποιημενη-πιθανοφάνεια για το μοντέλο της Λογαριθμο-λογιστικής κατανομής με από αριστερά και δεξιά αποκομμένες παρατηρήσεις δίνεται από τη σχέση:

$$l = \sum_{i=1}^n \{ (\delta_{ui} + \delta_{Li}) \ln \alpha + \delta_{ui} \ln \gamma + \gamma \ln t_i (\delta_{ui} + \delta_{Li}) - \delta_{ui} \ln t_i - \ln(1 + at_i^\gamma) [2 \delta_{ui} + \delta_{Li} + \delta_{Ri}] \}$$

Ισοδύναμα,

$$l = \ln L = \sum_{i=1}^n (\delta_{ui} + \delta_{Li}) \ln \alpha + \sum_{i=1}^n \delta_{ui} \ln \gamma + \sum_{i=1}^n \gamma \ln t_i (\delta_{ui} + \delta_{Li}) - \sum_{i=1}^n \delta_{ui} \ln t_i - \sum_{i=1}^n \ln(1 + at_i^\gamma) [2 \delta_{ui} + \delta_{Li} + \delta_{Ri}] \Rightarrow$$

Επομένως, αν k ο συνολικός αριθμός των διακοπών, μ ο συνολικός αριθμός των από αριστερά αποκομμένων παρατηρήσεων και ν ο συνολικός αριθμός των από δεξιά αποκομμένων παρατηρήσεων, θα ισχύει:

$$l = (k + \mu) \ln \alpha + k \ln \gamma + \gamma (k + \mu) \sum_{i=1}^n \ln t_i - k \sum_{i=1}^n \ln t_i - (2k + \mu + \nu) \sum_{i=1}^n \ln(1 + at_i^\gamma) \quad (3.2.3)$$

Για να βρούμε τις εκτιμήτριες μέγιστης πιθανοφάνειας για το μοντέλο της Λογαριθμο-λογιστικής κατανομής που εκτός από χρόνους διακοπής περιλαμβάνει από αριστερά και από δεξιά αποκομμένες παρατηρήσεις, αρχικά θα παραγωγίσουμε μερικώς τη συνάρτηση της λογαριθμημένης πιθανοφάνειας ως προς τις δύο παραμέτρους α και γ . Στη συνέχεια, θέτοντας τις μερικές

παραγώγους ίσες με το μηδέν και λύνοντας το σύστημα που προκύπτει, λαμβάνουμε τις ζητούμενες εκτιμήτριες μέγιστης πιθανοφάνειας \hat{a} και $\hat{\gamma}$.

Παραγωγίζοντας μερικώς την (3.2.3) ως προς a έχουμε:

$$\frac{\partial l}{\partial a} = (k + \mu) \frac{1}{a} - (2k + \mu + \nu) \frac{\partial \sum_{i=1}^n \ln(1 + at_i^\gamma)}{\partial a} \Rightarrow$$

$$\frac{\partial l}{\partial a} = (k + \mu) \frac{1}{a} - (2k + \mu + \nu) \frac{\partial \{\ln(1 + at_1^\gamma) + \ln(1 + at_2^\gamma) + \dots + \ln(1 + at_n^\gamma)\}}{\partial a}$$

Όμως,

$$\frac{\partial \ln(1 + at^\gamma)}{\partial a} = \frac{\partial \ln(1 + at^\gamma)}{\partial (1 + at^\gamma)} \frac{\partial (1 + at^\gamma)}{\partial a} \Rightarrow$$

$$\frac{\partial \ln(1 + at^\gamma)}{\partial a} = \frac{1}{(1 + at^\gamma)} t^\gamma$$

Άρα,

$$\frac{\partial l}{\partial a} = (k + \mu) \frac{1}{a} - (2k + \mu + \nu) \sum_{i=1}^n \frac{t_i^\gamma}{(1 + at_i^\gamma)} \quad (3.2.4)$$

Αντίστοιχα,

$$\frac{\partial l}{\partial \gamma} = k \frac{1}{\gamma} + (k + \mu) \sum_{i=1}^n \ln t_i - (2k + \mu + \nu) \frac{\partial \sum_{i=1}^n \ln(1 + at_i^\gamma)}{\partial \gamma}$$

Όμως όπως και στην προηγούμενη περίπτωση θα ισχύει αντίστοιχα:

$$\frac{\partial \ln(1 + at^\gamma)}{\partial \gamma} = \frac{\partial \ln(1 + at^\gamma)}{\partial (1 + at^\gamma)} \frac{\partial (1 + at^\gamma)}{\partial \gamma} \Rightarrow$$

$$\frac{\partial \ln(1 + at^\gamma)}{\partial \gamma} = \frac{1}{(1 + at^\gamma)} a t^\gamma \ln t$$

Συνεπώς,

$$\frac{\partial l}{\partial \gamma} = k \frac{1}{\gamma} + (k + \mu) \sum_{i=1}^n \ln t_i - (2k + \mu + \nu) \sum_{i=1}^n \frac{at_i^\gamma \ln t_i}{(1 + at_i^\gamma)} \quad (3.2.5)$$

Εξισώνοντας τις (3.2.4) και (3.2.5) με το μηδέν έχουμε:

$$\frac{\partial l}{\partial \alpha} = (k + \mu) \frac{1}{\alpha} - (2k + \mu + \nu) \sum_{i=1}^n \frac{t_i^\gamma}{(1 + at_i^\gamma)} = 0 \quad (3.2.6) \text{ και}$$

$$\frac{\partial l}{\partial \gamma} = k \frac{1}{\gamma} + (k + \mu) \sum_{i=1}^n \ln t_i - (2k + \mu + \nu) \sum_{i=1}^n \frac{at_i^\gamma \ln t_i}{(1 + at_i^\gamma)} = 0 \quad (3.2.7)$$

Λύνοντας, τέλος, το σύστημα των (3.2.6) και (3.2.7) ως προς α και γ , υπολογίζουμε τις ζητούμενες εκτιμήτριες μέγιστης πιθανοφάνειας $\hat{\alpha}$ και $\hat{\gamma}$.

3.2.2 Κριτήρια σύγκρισης προσαρμοσμένων μοντέλων

Ένας αρκετά αξιόπιστος τρόπος για να αποφανθούμε για το καταλληλότερο μοντέλο είναι να κάνουμε χρήση κάποιων μέτρων καταλληλότητας μοντέλου, όπως είναι τα κριτήρια AIC και BIC.

Το κριτήριο AIC χρησιμοποιεί την τιμή της συνάρτησης:

$$AIC = 2p - 2\hat{l}$$

Αντίστοιχα, το κριτήριο BIC χρησιμοποιεί την τιμή της συνάρτησης:

$$BIC = p \cdot \ln n - 2\hat{l},$$

όπου p το πλήθος των παραμέτρων του μοντέλου, n το μέγεθος του δείγματος και \hat{l} η μεγιστοποιημένη τιμή της λογαριθμημένης πιθανοφάνειας για το μοντέλο αυτό. Βέλτιστο θεωρείται το μοντέλο με την μικρότερη τιμή των AIC και BIC.

Για τα δεδομένα του παραδείγματος 1.6.1, με τη βοήθεια των κριτηρίων AIC και BIC θα επιλεγεί το μοντέλο διάρκειας ζωής που περιγράφει καλύτερα τα δεδομένα. Η επιλογή του καταλληλότερου μοντέλου θα γίνει ανάμεσα στις τρεις κατανομές, οι οποίες έχει αποδειχθεί ότι αποτελούν τα δημοφιλέστερα μοντέλα διάρκειας ζωής: την κατανομή Weibull, τη Λογαριθμο-κανονική και τη Λογαριθμο-λογιστική κατανομή.

Για να υπολογίσουμε τις τιμές των παραπάνω κριτηρίων χρειαζόμαστε τις εκτιμήσεις της μέγιστης πιθανοφάνειας. Από την παραμετρική ανάλυση της κατανομής Weibull προκύπτουν τα εξής αποτελέσματα:

Distribution Analysis: t

Variable: t

Censoring Information	Count
Uncensored value	11
Right censored value	27

Censoring value: censor = 0

Estimation Method: Maximum Likelihood

Distribution: Weibull

Parameter Estimates

Parameter	Estimate	Standard Error	95,0% Normal CI	
			Lower	Upper
Shape	3,16047	0,730818	2,00873	4,97257
Scale	27718,7	3046,02	22347,8	34380,5

Log-Likelihood = -123,995

Goodness-of-Fit
Anderson-Darling (adjusted) = 34,661

Επομένως, για την κατανομή Weibull ισχύει ότι:

$$\hat{l} = -123,995$$

Αντίστοιχα, για τη Λογαριθμο-κανονική κατανομή η ανάλυση προέκυψε ως εξής:

Distribution Analysis: t

Variable: t

Censoring Information	Count
Uncensored value	11
Right censored value	27

Censoring value: censor = 0

Estimation Method: Maximum Likelihood

Distribution: Lognormal

Parameter Estimates

Parameter	Estimate	Standard Error	95,0% Normal CI	
			Lower	Upper
Location	10,1448	0,144175	9,86219	10,4273
Scale	0,530068	0,112683	0,349447	0,804047

Log-Likelihood = -124,609

Goodness-of-Fit
Anderson-Darling (adjusted) = 34,651

Επομένως, για την Λογαριθμο-κανονική κατανομή ισχύει ότι:

$$\hat{I} = -124,609$$

Τέλος, για τη Λογαριθμο-λογιστική κατανομή έχουμε:

Distribution Analysis: t				
Variable: t				
Censoring Information		Count		
Uncensored value		11		
Right censored value		27		
Censoring value: censor = 0				
Estimation Method: Maximum Likelihood				
Distribution: <u>Loglogistic</u>				
Parameter Estimates				
		Standard	95,0% Normal CI	
Parameter	Estimate	Error	Lower	Upper
Location	10,1291	0,122225	9,88958	10,3687
Scale	0,280982	0,0663879	0,176833	0,446470
<u>Log-Likelihood = -124,365</u>				
Goodness-of-Fit				
Anderson-Darling (adjusted) = 34,639				

Επομένως, η λογαριθμοποιημενη-πιθανοφάνεια για τη Λογαριθμο-λογιστική κατανομή υπολογίσθηκε ως εξής:

$$\hat{I} = -124,365$$

Για τον υπολογισμό των κριτηρίων AIC και BIC για τις τρεις κατανομές έχουμε:

$$AIC - Weibull = -2 * (-123,995) + 2 * p = 247,99 + 2p$$

$$AIC - Lognormal = -2 * (-124,609) + 2 * p = 249,218 + 2p$$

$$AIC - Loglogistic = -2 * (-124,365) + 2 * p = 248,73 + 2p$$

Αντίστοιχα,

$$BIC - Weibull = -2 * (-123,995) + p \ln(n) = 247,99 + p \ln(n)$$

$$BIC - Lognormal = -2 * (-124,609) + p \ln(n) = 249,218 + p \ln(n)$$

$$BIC - \text{Loglogistic} = -2 * (-124,365) + \text{pln}(n) = 248,73 + \text{pln}(n)$$

Επομένως, οι μικρότερες τιμές για τα κριτήρια AIC και BIC εμφανίζονται, με μικρή διαφορά, στην κατανομή Weibull.

Συμπεραίνουμε, λοιπόν, ότι η κατανομή Weibull είναι εκείνη που με μικρή διαφορά προσαρμόζεται καλύτερα στα δεδομένα του πειράματος κι έτσι αποτελεί το καταλληλότερο μοντέλο για την παρούσα ανάλυση.

3.3 Μέθοδος ανάλυσης παραγοντικών σχεδιασμών με αποκομμένες παρατηρήσεις

(Chien-Hui Yang, 2006)

Ο κύριος σκοπός της παραγοντικής σχεδίασης και του παραμετρικού σχεδιασμού του Taguchi είναι ο προσδιορισμός των κύριων παραγόντων που επηρεάζουν τα ποιοτικά χαρακτηριστικά μιας διαδικασίας ή ενός προϊόντος καθώς και η επιλογή του συνδυασμού των σταθμών των παραγόντων που θα έχει σαν αποτέλεσμα τη βέλτιστη απόκριση. Στη συγκεκριμένη παράγραφο προτείνεται μια μέθοδος για την ανάλυση πειραματικών σχεδιασμών στα δεδομένα των οποίων περιέχονται εκτός από πλήρεις και αποκομμένες παρατηρήσεις τύπου II. Υπενθυμίζεται ότι οι αποκομμένες παρατηρήσεις τύπου II προκύπτουν όταν ο τερματισμός παρακολούθησης του πειράματος συμπίπτει με τη χρονική στιγμή καταστροφής της k-οστής μονάδας. Στη συγκεκριμένη περίπτωση, η διάρκεια παρακολούθησης του πειράματος είναι τυχαία.

Συνήθως, όταν χρησιμοποιείται η ανάλυση διασποράς (ANOVA) για τον προσδιορισμό του βέλτιστου συνδυασμού επιπέδων των παραγόντων, απαιτείται η υπόθεση κανονικότητας της απόκρισης. Ως εκ τούτου, στη συγκεκριμένη παράγραφο τα αποκομμένα δεδομένα τύπου II που θα μελετηθούν, θεωρούνται ότι ακολουθούν κανονική κατανομή. Στην περίπτωση μη κανονικότητας των μεταβλητών απόκρισης μπορούν να χρησιμοποιηθούν αρκετοί μετασχηματισμοί των δεδομένων, έτσι ώστε η κατανομή τους να πλησιάσει την κανονική. Μερικοί από τους μετασχηματισμούς που χρησιμοποιούνται είναι οι εξής: $Y = \ln X$, $Y = \arcsin X$, $Y = \sqrt{X}$. Για τον έλεγχο της κανονικότητας των δεδομένων

μπορούν να εφαρμοστούν τόσο γραφικοί έλεγχοι, όσο και στατιστικοί όπως το χ^2 -test ή το Kolmogorov-Smirnov test.

Παρακάτω παρατίθενται τα βήματα της διαδικασίας ανάλυσης των παραγοντικών σχεδιασμών που περιέχουν αποκομμένες παρατηρήσεις τύπου II.

- Εκτίμηση των αποκομμένων δεδομένων κάθε συνδυασμού ενός παραγοντικού σχεδιασμού χρησιμοποιώντας τη συνάρτηση MMLP.
- Κατασκευή ενός ψευδο-πλήρους συνόλου μέσω του συνδυασμού των εκτιμητριών των αποκομμένων παρατηρήσεων και των μη αποκομμένων παρατηρήσεων. Στη συνέχεια, το ψευδο-πλήρες σύνολο δεδομένων αναλύεται με τη βοήθεια της μεθόδου ανάλυσης διασποράς.
- Προσδιορισμός του βέλτιστου συνδυασμού σταθμών παραγόντων.

Σημειώνεται ότι, η συνάρτηση MMLP που χρησιμοποιείται στο πρώτο βήμα της μεθόδου αποτελεί την τροποποιημένη εκτιμήτρια μέγιστης πιθανοφάνειας (*modified maximum likelihood predictor*). Η συγκεκριμένη συνάρτηση χρησιμοποιείται για την εκτίμηση αποκομμένων παρατηρήσεων, που προκύπτουν στην περίπτωση τερματισμού του πειράματος προτού καταστραφούν όλες οι υπό μελέτη μονάδες. Έστω ότι οι παρατηρήσεις $X = (X_{(1)}, X_{(2)}, \dots, X_{(r)})$ αποτελούν τις πλήρεις παρατηρήσεις και ο σκοπός είναι να εκτιμηθούν οι αποκομμένες παρατηρήσεις $X_{(s)}$, με $1 \leq r < s \leq n$ και n το μέγεθος του δείγματος. Τα δεδομένα ικανοποιούν την υπόθεση της κανονικότητας, ενώ η μέση τιμή μ και η διασπορά σ^2 είναι άγνωστα μεγέθη.

Η εκτιμήτρια μέγιστης πιθανοφάνειας της μέσης τιμής και της διασποράς δίνονται από τις σχέσεις:

$$\hat{\mu} = \bar{X}_r + \frac{\hat{\sigma}}{r} \sum_{i=r+1}^n E(Z_{i:n})$$

και

$$\hat{\sigma} = \left\{ \alpha(X_{r:n} - \bar{X}_r) + \sqrt{[a(X_{r:n} - \bar{X}_r)]^2 + 4r[\sum_{i=1}^r (X_{i:n} - \bar{X}_r)^2]} \right\} / 2r,$$

όπου

$$Z_{i:n} = (X_{i:n} - \mu) / \sigma \quad \text{για } i = 1, 2, \dots, n,$$

$$\bar{X}_r = (1/r) \sum_{i=1}^r X_{i:n}$$

$$\text{και} \quad a = \sum_{i=r+1}^n E(Z_{i:n})$$

Τέλος η συνάρτηση MMLP δίνεται από τη σχέση :

$$\hat{X}_{s:n} = \begin{cases} \hat{\mu} + \hat{\sigma}E(Z_{s:n}) , & \text{αν } \hat{X}_{s:n} > \hat{X}_{r:n} \text{ και } r+1 < s \leq n \\ X_{r:n} , & \text{αν } \hat{X}_{s:n} \leq \hat{X}_{r:n} \text{ ή } s = r+1 \end{cases} \quad (3.3.1)$$

Η παραπάνω διαδικασία ανάλυσης μπορεί να εφαρμοστεί και στην περίπτωση του πειραματικού σχεδιασμού του Taguchi με αποκομμένα δεδομένα. Η διαδικασία έχει ως εξής:

- Εκτίμηση των αποκομμένων δεδομένων κάθε συνδυασμού ενός παραγοντικού σχεδιασμού χρησιμοποιώντας τη συνάρτηση MMLP.
- Κατασκευή ενός ψευδο-πλήρους συνόλου μέσω του συνδυασμού των εκτιμητριών των αποκομμένων παρατηρήσεων και των μη αποκομμένων παρατηρήσεων και υπολογισμός του δείκτη αναλογίας θορύβου (SN) κάθε αγωγής. Εν συνεχεία εφαρμόζοντας τη μέθοδο ανάλυσης διασποράς και αναλύοντας τα γραφήματα των επιδράσεων των παραγόντων, πραγματοποιείται ο καθορισμός των σημαντικών παραγόντων.
- Προσδιορισμός του βέλτιστου συνδυασμού των σταθμών των παραγόντων.

Υπενθυμίζεται ότι ο δείκτης αναλογίας θορύβου κάθε αγωγής υπολογίζεται από τη σχέση:

$$SN = -10 \log(s^2 + \bar{X}_{i:n}^2),$$

όπου s , $\bar{X}_{i:n}$ οι εκτιμήτριες της διασποράς και της μέσης τιμής κάθε αγωγής, αντίστοιχα, που έχουν υπολογισθεί με τη μέθοδο ελαχίστων τετραγώνων.

3.3.1 Αριθμητικό παράδειγμα παραμετρικής ανάλυσης δεδομένων στα οποία εμπεριέχονται αποκομμένες παρατηρήσεις

3.3.1.1 Περιγραφή του προβλήματος

Ένα πείραμα με 16 εκτελέσεις έλαβε χώρα σε ένα εργοστάσιο κατασκευής ημιαγωγών με σκοπό τη διερεύνηση έξι παραγόντων που επηρεάζουν την καμπυλότητα και την κυρτότητα των υπόγειων συσκευών που κατασκευάζονται στο εργοστάσιο. Στον πίνακα 3.1 που ακολουθεί, παρουσιάζονται οι έξι μεταβλητές (παράγοντες) και τα επίπεδα τους. Οι παράγοντες είναι οι εξής:

A: θερμοκρασία πλαστικοποίησης ($^{\circ}\text{C}$)

B: χρονική διάρκεια πλαστικοποίησης (h)

C: πίεση πλαστικοποίησης (Ton)

D: θερμοκρασία πυρακτώσεως ($^{\circ}\text{C}$)

E: χρονική διάρκεια πυρακτώσεως (h)

F: σημείο πυρακτώσεως ($^{\circ}\text{C}$)

Κάθε εκτέλεση επαναλήφθηκε τέσσερις φορές κατά τις οποίες λήφθηκαν οι τιμές της κυρτότητας των συσκευών. Στον πίνακα 3.2 παρουσιάζονται τα σχετικά δεδομένα, όπου με τα σύμβολα +1 και -1 χρησιμοποιήθηκαν για να εκφράσουν την υψηλή και χαμηλή στάθμη των παραγόντων.

Εάν η μεγαλύτερη τιμή που λαμβάνει κάθε συνδυασμός παραγόντων θεωρηθεί αποκομμένη, τότε προκύπτουν οι από δεξιά αποκομμένες παρατηρήσεις τύπου II με $n = 4, r = 3$ και $s = 4$. Το Kolmogorov-Smirnov test χρησιμοποιήθηκε για να επαληθευθεί η υπόθεση της κανονικότητας των πειραματικών δεδομένων. Τα αποτελέσματα του test επιβεβαίωσαν την υπόθεση κανονικότητας.

Label	Factor	Low level	High level
A	Lamination Temperature(°C)	55	75
B	Lamination Time (s)	10	25
C	Lamination Pressure (Ton)	5	10
D	Firing Temperature (°C)	1,580	1,620
E	Firing Cycle Time (h)	17.5	29
F	Firing Dew Point (°C)	20	26

Πίνακας 3.1

Run	A	B	C	D	E	F	Camber for replicate			
							X_1	X_2	X_3	X_4
1	-1*	-1	-1	-1	-1	-1	167	128	149	185
2	1	-1	-1	-1	1	-1	62	66	44	20
3	-1	1	-1	-1	1	1	41	42	43	50
4	1	1	-1	-1	-1	1	73	81	39	30
5	-1	-1	1	-1	1	1	47	47	40	89
6	1	-1	1	-1	-1	1	219	258	147	296
7	-1	1	1	-1	-1	-1	121	90	92	86
8	1	1	1	-1	1	-1	191	186	162	106
9	-1	-1	-1	1	-1	1	32	23	77	69
10	1	-1	-1	1	1	1	78	158	60	45
11	-1	1	-1	1	1	-1	43	27	28	28
12	1	1	-1	1	-1	-1	186	137	159	158
13	-1	-1	1	1	1	-1	110	86	101	158
14	1	-1	1	1	-1	-1	65	109	126	71
15	-1	1	1	1	-1	1	155	158	145	145
16	1	1	1	1	1	1	93	124	110	133

Πίνακας 3.2

3.3.1.2 Ανάλυση δεδομένων

Για την ανάλυση των δεδομένων, τα οποία περιέχουν από δεξιά αποκομμένα δεδομένα τύπου II θα χρησιμοποιηθεί η διαδικασία που περιγράφηκε στην παράγραφο 3.3. Η συνάρτηση 3.3.1 που αποτελεί την MMLP χρησιμοποιείται για την πρόβλεψη των αποκομμένων παρατηρήσεων σε κάθε συνδυασμό παραγόντων. Επειδή $s = r + 1$, η μεταβλητή κάθε συνδυασμού παραγόντων $X_{s:n}$ αντικαθίσταται από την ποσότητα $X_{r:n}$. Ο πίνακας 3.3 παρουσιάζει τα ψευδο-συμπληρωμένα δεδομένα.

No.	$X_{1:4}$	$X_{2:4}$	$X_{3:4}$	$\hat{X}_{4:4}$
1	128	149	167	167
2	20	44	62	62
3	41	42	43	43
4	30	39	73	73
5	40	47	47	47
6	147	219	258	258
7	86	90	92	92
8	106	162	186	186
9	23	32	69	69
10	45	60	78	78
11	27	28	28	28
12	137	158	159	159
13	86	101	110	110
14	65	71	109	109
15	145	145	155	155
16	93	110	124	124

Πίνακας 3.3

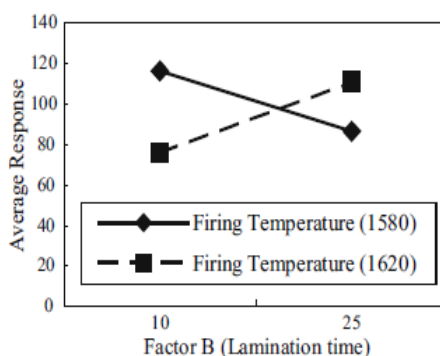
Στη συνέχεια χρησιμοποιείται η ανάλυση διασποράς με σκοπό τον καθορισμό του συνδυασμού επιπέδων παραγόντων που οδηγεί στη βέλτιστη απόκριση. Τα αποτελέσματα της ανάλυσης διασποράς παρουσιάζονται στον πίνακα 3.4. Από την ανάλυση παρατηρείται μια σαφής αλληλεπίδραση των παραγόντων B και D, καθώς η p -τιμή του BD είναι πολύ μικρότερη του $\alpha = 0.05$. Οι υπόλοιπες αλληλεπιδράσεις παραγόντων είναι μη σημαντικές αφού η p -τιμή τους είναι μεγαλύτερη από $\alpha = 0.05$, για αυτό το λόγο στον πίνακα 3.4 αναφέρεται μόνο η αλληλεπίδραση BD. Επιπλέον, από την ανάλυση διασποράς προκύπτει, ότι οι κύριες επιδράσεις A, C και E είναι σημαντικές καθώς η p -τιμή τους είναι, επίσης, μικρότερη του $\alpha = 0.05$.

Source of variation	Sum of squares	Degree of freedom	Mean square	F_0	P -value
A	14762.2500	1	14762.2500	7.66	0.0076
B	105.0625	1	105.0625	0.05	0.8162
C	35815.5625	1	35815.5625	18.59	<.0001
D	1024.0000	1	1024.0000	0.53	0.4691
E	31506.2500	1	31506.2500	16.35	0.0002
F	1722.2500	1	1722.2500	0.89	0.3485
BD	16835.0625	1	16835.0625	8.74	0.0046
Error	107909.3125	56	1926.9520		
Total	209679.7500	63			

Πίνακας 3.4

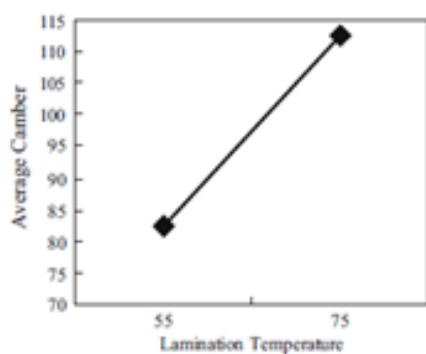
Ακολουθούν κάποια γραφήματα που αφορούν τις σημαντικές αλληλεπιδράσεις και τις σημαντικές κύριες επιδράσεις που προέκυψαν από την ανάλυση διασποράς.

Το γράφημα 3.1 παρουσιάζει τις μέσες αποκρίσεις σε συνάρτηση με τις τιμές που λαμβάνουν οι παράγοντες B και D. Το πόσο σημαντική είναι η αλληλεπίδραση εκφράζεται από το γεγονός ότι οι δύο ευθείες του γραφήματος είναι τέμνουσες και όχι παράλληλες. Το γράφημα υποδεικνύει ότι η αλλαγή της τιμής του παράγοντα B από την χαμηλή στάθμη στην υψηλή επηρεάζει τη μέση τιμή της κυρτότητας. Συγκεκριμένα, αν ταυτόχρονα η τιμή του παράγοντα D αντιστοιχεί στην υψηλή στάθμη, τότε η μέση απόκριση μειώνεται. Αντιθέτως, αν ο παράγοντας D βρίσκεται στη χαμηλή του στάθμη και η τιμή του παράγοντα B αυξάνεται, τότε η μέση τιμή της κυρτότητας αυξάνεται. Η απόκριση του πειράματος ανήκει στην κατηγορία smaller-the-better, συνεπώς ο συνδυασμός που φαίνεται να έχει το καλύτερο αποτέλεσμα είναι όταν η τιμή του παράγοντα B αντιστοιχεί στη χαμηλή στάθμη και ο παράγοντας D αντιστοιχεί στην υψηλή του στάθμη.

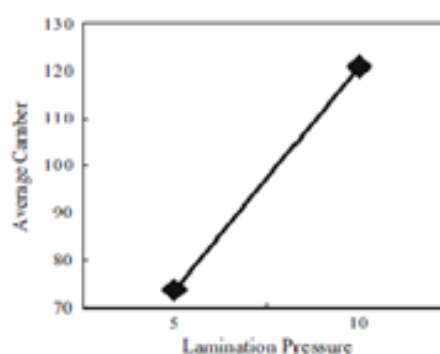


Γράφημα 3.1

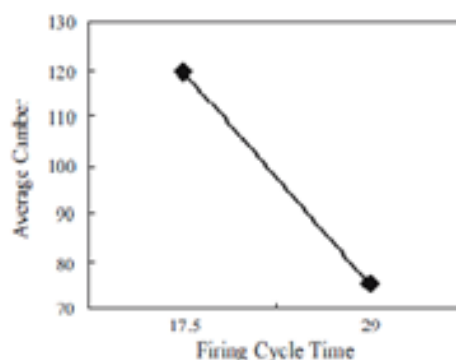
Τα γραφήματα 3.2, 3.3 και 3.4 απεικονίζουν τα διαγράμματα των σημαντικών κύριων επιδράσεων A, C και E αντίστοιχα. Στα συγκεκριμένα γραφήματα παρουσιάζεται η οριακή μέση απόκριση συναρτήσει των επιπέδων των τριών παραγόντων. Δεδομένου ότι η μικρότερη απόκριση είναι η επιθυμητή, μέσω των διαγραμμάτων συμπεραίνεται ότι ο καλύτερος συνδυασμός παραγόντων είναι όταν οι παράγοντες A και C λαμβάνουν τιμές στην χαμηλή τους στάθμη και ο E στην υψηλή του. Ως εκ τούτου ο βέλτιστος συνδυασμός παραγόντων – επιπέδων μπορεί να οριστεί ως $A_{-}B_{-}C_{-}D_{+}E_{+}F_{+}$. Τα σύμβολα (+) και (-) χρησιμοποιούνται για να εκφράσουν τη χαμηλή και την υψηλή στάθμη κάθε παράγοντα, αντίστοιχα.



Γράφημα 3.2



Γράφημα 3.3



Γράφημα 3.4

3.3.1.3 Σύγκριση του αποτελέσματος της ανάλυσης του πειράματος με πλήρη δεδομένα

Στην περίπτωση που δε θεωρηθούν αποκομμένες παρατηρήσεις, η ανάλυση των πλήρων πειραματικών δεδομένων υποδεικνύει ως σημαντικούς παράγοντες τους A, B, C, D και E. Ο βέλτιστος συνδυασμός των επιπέδων των παραγόντων που προκύπτει αποτελείται από τους παράγοντες A, B και C στη χαμηλή τους στάθμη και τους παράγοντες D και E στην υψηλή τους. Το συγκεκριμένο αποτέλεσμα συμπίπτει με το βέλτιστο συνδυασμό που προέκυψε από την προηγούμενη μέθοδο, κάτι που υποδεικνύει την αποτελεσματικότητα της μεθόδου.

ΚΕΦΑΛΑΙΟ 4:

ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ ΑΝΑΛΥΣΗΣ ΠΑΡΑΓΟΝΤΙΚΩΝ ΣΧΕΔΙΑΣΜΩΝ

4.1 Η εκτιμήτρια Kaplan-Meier

(Bewick et al. 2004)

Η εκτιμήτρια Kaplan-Meier είναι μία από τις πιο δημοφιλείς μεθόδους εκτίμησης της καμπύλης επιβίωσης δεδομένων που συμπεριλαμβάνουν αποκομμένες παρατηρήσεις. Πρόκειται για μία μη παραμετρική μέθοδο που δεν απαιτεί καμία υπόθεση για την ακριβή μορφή της κατανομής του δείγματος. Η μέθοδος βασίζεται στην ιδέα ότι η πιθανότητα να επιβιώσει μία μονάδα k ή περισσότερα χρονικά διαστήματα από τη στιγμή εισόδου στο πείραμα ισούται με το γινόμενο των k παρατηρούμενων ποσοστών επιβίωσης των χρονικών διαστημάτων, δηλαδή το συσσωρευτικό ποσοστό επιβίωσης.

Πιο συγκεκριμένα, έστω τυχαίο δείγμα n μονάδων, μερικές από τις οποίες καταστρέφονται κατά τις διακεκριμένες χρονικές στιγμές

$$t_{(1)} < t_{(2)} < \dots < t_{(k)}, \quad k \leq n$$

Η συνάρτηση επιβίωσης τη χρονική στιγμή $j \in (1, k)$ δίνεται από τη σχέση:

$$\begin{aligned} S(t_{(j)}) &= P(T > t_{(j)}) \\ &= P(T > t_{(1)}) \cdot P(T > t_{(2)} | T > t_{(1)}) \cdot \dots \cdot P(T > t_{(j)} | T > t_{(j-1)}) \end{aligned}$$

Η εκτιμήτρια της $P(T > t_{(1)})$ είναι η

$$\hat{P}(T > t_{(1)}) = 1 - p_1 = 1 - \frac{d_1}{r_1} = \frac{r_1 - d_1}{r_1}$$

Όπου η $p_1 = \frac{d_1}{r_1}$ είναι η σχετική συχνότητα των κατεστραμμένων μονάδων στο χρονικό διάστημα $(0, t_{(1)})$, με d_1 το πλήθος των μονάδων που καταστρέφονται τη χρονική στιγμή $t_{(1)}$ και r_1 το πλήθος των μονάδων που λειτουργούσαν ακριβώς πριν από τη συγκεκριμένη χρονική στιγμή. Ομοίως, υπολογίζονται οι εκτιμήσεις $\hat{P}(T > t_{(j)})$ για κάθε χρονική στιγμή $t_{(j)}$, παραδείγματος χάριν

$$\hat{P}(T > t_{(1)}) = 1 - \frac{d_2}{r_2} = \frac{r_2 - d_2}{r_2}.$$

Συνεπώς, η εκτιμήτρια Kaplan-Meier δίνεται από τη σχέση:

$$\hat{S}(t) = \frac{r_1 - d_1}{r_1} \times \frac{r_2 - d_2}{r_2} \times \dots \times \frac{r_i - d_i}{r_i}, i: t_{(i)} \leq t < t_{(i+1)}$$

$$= \begin{cases} \prod_{j:t_{(j)} \leq t} \frac{r_j - d_j}{r_j}, & t \geq t_{(1)} \\ 1, & t < t_{(1)} \end{cases}$$

Το τυπικό σφάλμα της $\hat{S}(t)$ δίνεται μέσω του τύπου του Greenwood και εκφράζει μια εκτίμηση της ακρίβειας της εκτιμήτριας Kaplan-Meier:

$$se(\hat{S}(t)) = \hat{S}(t) \left\{ \sum_{t_{(j)} \leq t} \frac{d_j}{r_j(r_j - d_j)} \right\}^{1/2}.$$

4.1.1 Αριθμητικό παράδειγμα

Στον πίνακα 4.1 παρουσιάζονται δεδομένα διάρκειας ζωής ασθενών που ακολουθούν μία από τις δύο διαφορετικές θεραπείες 1 και 2.

Patient number	Survival time (days)	Outcome	Treatment	Age (years)
1	1	Died	2	75
2	1	Died	2	79
3	4	Died	2	85
4	5	Died	2	76
5	6	Unknown	2	66
6	8	Died	1	75
7	9	Survived	2	72
8	9	Died	2	70
9	12	Died	1	71
10	15	Unknown	1	73
11	22	Died	2	66
12	25	Survived	1	73
13	37	Died	1	68
14	55	Died	1	59
15	72	Survived	1	61

Πίνακας 4.1

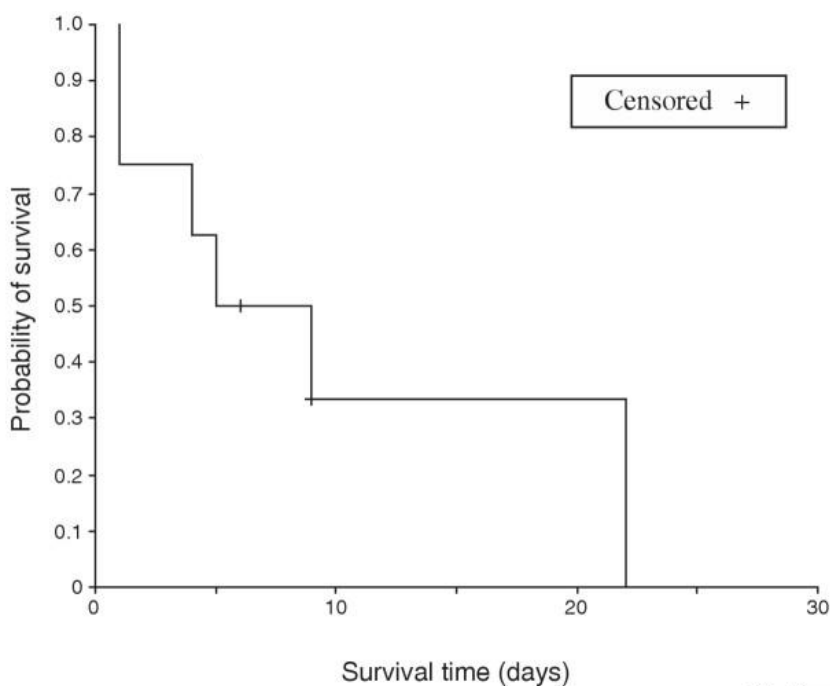
Για τον υπολογισμό της εκτιμήτριας Kaplan-Meier τα δεδομένα διάρκειας ζωής, συμπεριλαμβανομένων και των αποκομμένων παρατηρήσεων, θα πρέπει να είναι διατεταγμένα κατά αύξουσα σειρά. Στην περίπτωση που μία αποκομμένη τιμή συμπίπτει με μία πλήρη, τότε η πλήρη θα προηγείται της αποκομμένης. Στον

πίνακα 4.2 παρουσιάζονται οι υπολογισμοί της εκτιμήτριας Kaplan-Meier των ασθενών που ακολουθούν τη θεραπεία 2.

Patient number	Survival time (days)	Number known to be alive (r_i)	Deaths (d_i)	Proportion surviving (p_i)	Cumulative proportion surviving ($S(t)$)
	0				1
1	1	8			
2	1	8	2	$(8 - 2)/8 = 0.750$	$1 \times 0.750 = 0.750$
3	4	6	1	$(6 - 1)/6 = 0.833$	$0.750 \times 0.833 = 0.625$
4	5	5	1	$(5 - 1)/5 = 0.800$	$0.625 \times 0.800 = 0.500$
5	6+				
7	9	3	1	$(3 - 1)/3 = 0.667$	$0.500 \times 0.667 = 0.333$
8	9+				
11	22	1	1	$(1 - 1)/1 = 0.00$	$0.333 \times 0.00 = 0.000$

Πίνακας 4.2

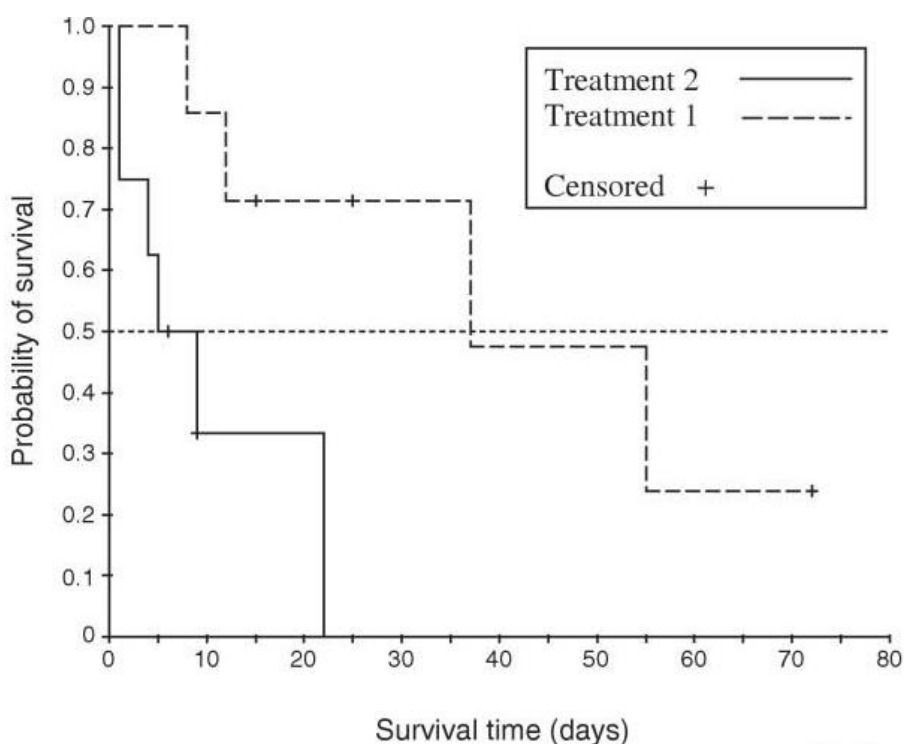
Στην περίπτωση αποκομμένης τιμής το ποσοστό επιβίωσης p_i ισούται με 1. Το συγκεκριμένο γεγονός δεν επηρεάζει το σωρευτικό ποσοστό επιβίωσης, $\hat{S}(t)$, συνεπώς οι συγκεκριμένες τιμές παραλείπονται από τον πίνακα 4.2.



Γράφημα 4.1

Το γράφημα 4.1 μας δίνει την καμπύλη επιβίωσης των ασθενών που ακολουθούν την θεραπεία 2. Με + συμβολίζονται στο γράφημα οι αποκομμένες τιμές.

Μέσω γραφημάτων που αναπαριστούν τις εκτιμήτριες των συναρτήσεων επιβίωσης διαφορετικών πληθυσμών η μέθοδος Kaplan-Meier δίνει τη δυνατότητα σύγκρισης τους. Το γράφημα 4.2 παρουσιάζει με διακεκομμένη γραμμή την καμπύλη επιβίωσης των ασθενών που ακολουθούν την 1^η θεραπεία και με συνεχή την καμπύλη επιβίωσης εκείνων που ακολουθούν την 2^η. Από το γράφημα διαπιστώνεται ότι οι ασθενείς που ακολουθούν τη 1^η θεραπεία έχουν μεγαλύτερο ποσοστό επιβίωσης από εκείνους που ακολουθούν την 2^η. Το γράφημα της καμπύλης επιβίωσης μπορεί να χρησιμοποιηθεί, επίσης, για την εκτίμηση του μέσου χρόνου επιβίωσης. Ο χρόνος, αυτός, αντιστοιχεί στη χρονική στιγμή της οποίας η πιθανότητα επιβίωσης ισούται με 0,5. Ο μέσος χρόνος επιβίωσης για εκείνους τους ασθενείς που υπόκεινται στη θεραπεία 2 φαίνεται να είναι, περίπου, 5 ημέρες έναντι 37 ημερών για τους ασθενείς που τους χορηγείται η θεραπεία 1.



Γράφημα 4.2

4.2 Ο μη παραμετρικός έλεγχος log-rank

Ο μη παραμετρικός έλεγχος log-rank δίνει τη δυνατότητα σύγκρισης δύο συναρτήσεων επιβίωσης χωρίς να προσδιορίζονται μαθηματικά οι συναρτήσεις αυτές. Χρησιμοποιείται, δηλαδή, για τον έλεγχο της μηδενικής υπόθεσης $H_0: S_1(t) = S_2(t)$. Προτάθηκε από τους Mantel και Haenszel το 1959.

Έστω $t_{(1)} < t_{(2)} < \dots < t_{(n)}$ διακεκριμένες χρονικές στιγμές κατά τις οποίες παύουν να λειτουργούν μονάδες που προέρχονται από δύο διαφορετικές ομάδες. Η ελεγχοσυνάρτηση του ελέγχου log-rank δίνεται από τον τύπο:

$$\chi^2(\log - rank) = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \quad (4.2.1)$$

όπου O_1 και O_2 ο συνολικός αριθμός των παρατηρούμενων γεγονότων στους πληθυσμούς 1 και 2, αντίστοιχα, και E_1 και E_2 ο συνολικός αριθμός των αναμενόμενων γεγονότων.

Ο συνολικός αριθμός των αναμενόμενων γεγονότων μίας ομάδας ($E_j: j = 1, 2$) προκύπτει αθροίζοντας τον αριθμό των αναμενόμενων γεγονότων ($E_{ij}: i = 1, \dots, n$) κάθε χρονικής στιγμής $t_{(i)}$ κατά την οποία συμβαίνει το γεγονός. Όπου ο αναμενόμενος αριθμός γεγονότων της χρονικής στιγμής $t_{(i)}$ υπολογίζεται ως το ρίσκο θανάτου τη συγκεκριμένη χρονική στιγμή πολλαπλασιασμένο με τον αριθμό των μονάδων που είναι 'ζωντανές' αμέσως πριν από την συγκεκριμένη στιγμή (r_i). Υπό τη μηδενική υπόθεση, το ρίσκο θανάτου που ισούται με το πηλίκο του πλήθους των 'θανάτων' προς το πλήθος των 'ζωντανών' μονάδων (d_i/r_i), μπορεί να υπολογιστεί από το συνδυασμό των δεδομένων των δύο ομάδων. Στον πίνακα 4.3 παρουσιάζονται οι υπολογισμοί για την εύρεση του αναμενόμενου αριθμού 'θανάτου', E_2 , για τους ασθενείς που ακολουθούν την 2^η θεραπεία του αριθμητικού παραδείγματος 4.1.1.

Survival time (days)	Treatment group	Number known to be alive (r_i)	Deaths (d_i)	Risk for death (d_i/r_i)	Number known to be alive from treatment group 2 (r_{2i})	Expected number of events in treatment group 2 (E_{2i})
0						
1	2	15	2	$2/15 = 0.133$	8	$8 \times 0.133 = 1.07$
1	2					
4	2	13	1	$1/13 = 0.077$	6	$6 \times 0.077 = 0.46$
5	2	12	1	$1/12 = 0.083$	5	$5 \times 0.083 = 0.42$
6+	2	11	0	$0/11 = 0$	4	$4 \times 0 = 0.00$
8	1	10	1	$1/10 = 0.100$	3	$3 \times 0.100 = 0.30$
9	2	9	1	$1/9 = 0.111$	3	$3 \times 0.111 = 0.33$
9+	2	8	0	$0/8 = 0$	2	$2 \times 0 = 0.00$
12	1	7	1	$1/7 = 0.143$	1	$1 \times 0.143 = 0.14$
15+	1	6	0	$0/6 = 0$	1	$1 \times 0 = 0.00$
22	2	5	1	$1/5 = 0.200$	1	$1 \times 0.200 = 0.20$
25+	1	4	0	$0/4 = 0$	0	$0 \times 0 = 0.00$
37	1	3	1	$1/3 = 0.333$	0	$0 \times 0 = 0.00$
55	1	2	1	$1/2 = 0.500$	0	$0 \times 0 = 0.00$
72+	1					
						$E_2 = 2.92$

Πίνακας 4.3

Από τον πίνακα παρατηρείται ότι στο ξεκίνημα της 4^{ης} μέρας, κατά την οποία πραγματοποιείται ο τρίτος θάνατος (3^ο γεγονός), υπήρχαν ακόμα 13 ζωντανοί ασθενείς. Ο ένας απεβίωσε οπότε το ρίσκο θανάτου ανέρχεται στο $1/13 = 0,077$. Οι έξι από τους δεκατρείς ασθενείς άνηκαν στην 2^η ομάδα, οπότε ο αναμενόμενος αριθμός θανάτων στο 3^ο γεγονός ισούται με $6 \times 0,077 = 0,46$. Με την ίδια λογική πραγματοποιήθηκαν και οι υπόλοιποι υπολογισμοί που φαίνονται στον πίνακα 4.3. Συνεπώς, ο συνολικός αριθμός αναμενόμενων θανάτων για την 2^η ομάδα ασθενών προκύπτει από τη σχέση:

$$E_2 = \sum_{i=1}^k \frac{d_i}{r_i} r_{2i}$$

όπου r_{2i} είναι ο αριθμός των εν ζωή ασθενών της 2^{ης} ομάδας τη χρονική στιγμή $t_{(i)}$. Η ποσότητα E_1 μπορεί να υπολογιστεί ως $n - E_2$, όπου n ο συνολικός αριθμός των γεγονότων.

Στην περίπτωση που ο αριθμός των γεγονότων δεν είναι πολύ μικρός η ελεγχοσυνάρτηση log-rank ακολουθεί ασυμπτωτικά την χ_1^2 , δηλαδή την ακολουθεί την κατανομή χ^2 με ένα βαθμό ελευθερίας.

Για τα δεδομένα του παραδείγματος 4.1.1, ο συνολικός αριθμός των αναμενόμενων θανάτων για τη 2^η ομάδα υπολογίστηκε 2,92 ενώ ο συνολικός αριθμός θανάτων είναι 10. Συνεπώς, $E_1 = 10 - 2,92 = 7,08$. Αντικαθιστώντας, λοιπόν, τις συγκεκριμένες τιμές στην (1), η ελεγχοσυνάρτηση log-rank ισούται με:

$$(4.2.1) \Leftrightarrow \frac{(4 - 7,08)^2}{7,08} + \frac{(6 - 2,92)^2}{2,92} = 4,59$$

Από τους πίνακες της κατανομής χ_1^2 η p -τιμή ισούται με 0,032 , κάτι που υποδεικνύει μια σημαντική διαφορά μεταξύ των καμπυλών επιβίωσης των δύο πληθυσμών.

4.3 Η εκτιμήτρια Nelson-Aalen

Η εκτιμήτρια Nelson-Aalen εισήχθη για πρώτη φορά από τον Nelson (1972) στο πλαίσιο των μοντέλων αξιοπιστίας καθώς επίσης και από τον Aalen (1978) ο οποίος χρησιμοποίησε σύγχρονες τεχνικές καταμέτρησης. Η εκτιμήτρια έχει ένα σύνολο ιδιοτήτων που την καθιστούν αρκετά δημοφιλή. Μία εκ των οποίων είναι το γεγονός ότι παρέχει καλύτερη απόδοση σε δείγματα μικρού μεγέθους από άλλες μεθόδους. Η εκτιμήτρια Nelson-Aalen χρησιμοποιείται, επίσης, αρκετά στη βιοστατιστική. Κυρίως εφαρμόζεται με δύο τρόπους στην ανάλυση επιβίωσης. Αρχικά, χρησιμοποιείται για την σύγκριση μεταξύ διαφορετικών παραμετρικών μοντέλων και την επιλογή που εκφράζει καλύτερα τα δεδομένα του πειράματος. Δεύτερον, παρέχει εκτιμήσεις του ποσοστού κινδύνου που μπορεί στη συνέχεια να εξομαλυνθεί.

Η εκτιμήτρια Nelson-Aalen χρησιμοποιείται για την εκτίμηση της σωρευτικής συνάρτησης διακινδύνευσης του δείγματος. Πρόκειται για μία μη παραμετρική μέθοδο.

Η σωρευτική συνάρτηση διακινδύνευσης συνδέεται με τη συνάρτηση επιβίωσης μέσω της σχέσης: $H(t) = -\ln S(t)$. Συνεπώς, η εκτιμήτρια Nelson-Aalen, $\hat{H}(t)$, υπολογίζεται μέσω της εκτιμήτριας Kaplan- Meier, $\hat{S}(t)$.

$$\begin{aligned}\hat{H}(t) &= -\ln \hat{S}(t) \\ &= -\ln \prod_{j:t_{(j)} \leq t} \frac{r_j - d_j}{r_j} \\ &= -\sum_{j:t_{(j)} \leq t} \ln \left(1 - \frac{d_j}{r_j} \right) \\ &\cong \sum_{j:t_{(j)} \leq t} \frac{d_j}{r_j}\end{aligned}$$

Συνεπώς,

$$\hat{H}(t) = \begin{cases} \sum_{j:t_{(j)} \leq t} \frac{d_j}{r_j}, & \text{όταν } t \geq t_{(1)} \\ 0, & \text{όταν } t < t_{(1)} \end{cases}$$

Η εκτιμήτρια της διασποράς της εκτιμήτριας Nelson-Aalen υπολογίζεται από τον τύπο:

$$\hat{V}(\hat{H}) = \sum_{j:t_{(j)} \leq t} \frac{d_j}{r_j^2}$$

4.4 Μη παραμετρικές μέθοδοι για την ανάλυση παραγοντικών σχεδιασμών με αποκομμένες παρατηρήσεις

4.4.1 Μη παραμετρική μέθοδος ανάλυσης παραγοντικών σχεδιασμών με αποκομμένες παρατηρήσεις των Lee – Ing Tong και Chao – Ton Su

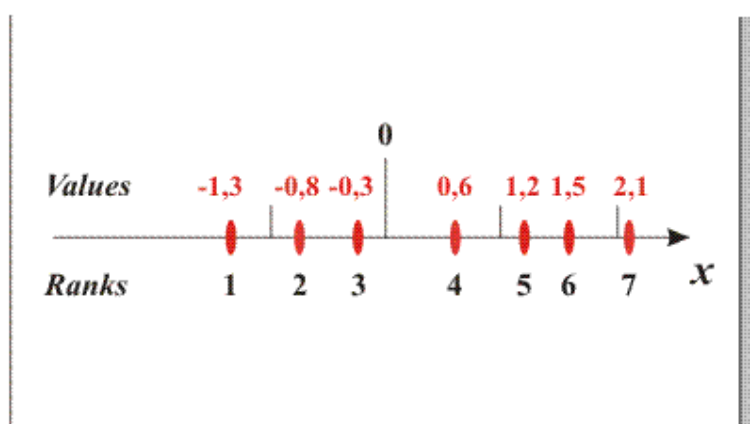
Η μη παραμετρική μέθοδος που θα αναλυθεί στη συνέχεια, σε συνδυασμό με την ανάλυση παλινδρόμησης όχι μόνο λαμβάνει υπόψη την τυπική απόκλιση της απόκρισης, αλλά μπορεί ακόμα να φανεί χρήσιμη σε επαναλαμβανόμενα ή μη πειράματα που εμπεριέχουν από δεξιά ή από αριστερά αποκομμένες παρατηρήσεις. Επιπλέον, η συγκεκριμένη μέθοδος κάνει χρήση των τάξεων των παρατηρήσεων.

Γενικότερα, η χρήση της τάξης των παρατηρήσεων, αντί για τις ακριβείς τιμές των δεδομένων, είναι μία τυπική προσέγγιση στον τομέα της μη παραμετρικής στατιστικής, καθώς παρουσιάζει αρκετά πλεονεκτήματα και δίνει τη δυνατότητα απλοποίησης της διαδικασίας ανάλυσης των δεδομένων. Αρχικά, οι τάξεις των αποκομμένων παρατηρήσεων μπορούν να βρεθούν ευκολότερα από τις αναμενόμενες αποκρίσεις τους. Επιπλέον, η κατανομή των τάξεων παραμένει η ίδια με την κατανομή του αρχικού συνόλου δεδομένων και δεν επηρεάζεται από μονότονες μεταθέσεις των δεδομένων, καθώς επίσης και από οποιαδήποτε αλλαγή κλίμακας στην περίπτωση ποσοτικών μεταβλητών. Τέλος, η χρήση των τάξεων των παρατηρήσεων είναι μια ισχυρή μέθοδος ανάλυσης ενός προβλήματος για το

οποίο δεν υπάρχουν αρκετές πληροφορίες για την κατανομή που ακολουθούν τα δεδομένα του.

Παρακάτω, βλέπουμε ένα παράδειγμα ενός ταξινομημένου δείγματος 7 παρατηρήσεων. Στην πάνω γραμμή του γραφήματος φαίνονται οι πραγματικές τιμές των παρατηρήσεων για τη μεταβλητή x σε αύξουσα σειρά, ενώ στην τελευταία γραμμή φαίνεται η τάξη κάθε παρατήρησης.

Γενικά σε ένα δείγμα n παρατηρήσεων η μικρότερη παρατήρηση θεωρείται πως είναι τάξης 1 (*rank 1*) και η μεγαλύτερη τάξης n (*rank n*).



Γράφημα 4.3

Στο σημείο αυτό, θα δούμε αναλυτικά τα βήματα που ακολουθούνται από τη μη παραμετρική μέθοδο ανάλυσης δεδομένων (Tong et al., 1996):

1. Σαν πρώτο βήμα, απαραίτητος είναι ο διαχωρισμός των πειραματικών δεδομένων σε πλήρη (*uncensored*) και αποκομμένα (*censored*). Με Y_U θα συμβολίζεται το σύνολο των πλήρων και με Y_C το σύνολο των αποκομμένων.

Έστω N το μέγεθος του δείγματος, δηλαδή θεωρείται ότι υπάρχουν N αποκρίσεις. Εάν το σημείο αποκοπής είναι το C και υπάρχουν n πλήρη δεδομένα ισχύει: $Y_U = [y_1, y_2, \dots, y_n]^T$ και $Y_C = [y_{n+1}, y_{n+2}, \dots, y_N]^T$, όπου με y_i συμβολίζονται οι παρατηρούμενες τιμές. Τα πλήρη δεδομένα ταξινομούνται κατά αύξουσα σειρά μεγέθους και συμβολίζονται με $R_U = [r_1, r_2, \dots, r_n]^T$, όπου με r_i συμβολίζεται η τάξη της παρατήρησης y_i .

Στην περίπτωση που κάποιες παρατηρούμενες τιμές είναι ίσες, τότε στην κάθε μία αποδίδεται ο μέσος όρος των αντίστοιχων τάξεων.

2. Στη συνέχεια αν ορίσουμε Z_U τον πίνακα των σταθμών των παραγόντων για τα πλήρη δεδομένα και $\hat{\beta}_U$ τον πίνακα των εκτιμητριών των συντελεστών παλινδρόμησης μπορούμε να βρούμε, πάντα με χρήση της ανάλυσης παλινδρόμησης, τη σχέση ανάμεσα στο διάνυσμα Y_U και στον πίνακα Z_U όπου:

$$\hat{\mu}_{Y_U|Z_U} = Z_U \hat{\beta}_U \quad (4.4.1)$$

με $\hat{Y}_U = \hat{\mu}_{Y_U|Z_U}$.

3. Σαν τρίτο βήμα απαιτείται η εύρεση της εκτιμήτριας \hat{Y}_C τοποθετώντας στη σχέση (4.4.1) αντί για τον πίνακα Z_U , τον πίνακα Z_C , δηλαδή τον πίνακα σταθμών παραγόντων για τα αποκομμένα δεδομένα.

4. Επόμενο βήμα αποτελεί η ταξινόμηση της εκτιμήτριας αποκομμένων δεδομένων \hat{Y}_C , με τα δεδομένα να τοποθετούνται σε αύξουσα σειρά. Το ταξινομημένο διάνυσμα συμβολίζεται ως $\hat{R}_C = [r_{n+1}, r_{n+2}, \dots, r_N]^T$. Έτσι, αν σε δείγμα μεγέθους N με n το πλήθος πλήρεις παρατηρήσεις, έχουμε μια παρατήρηση r_i αποκομμένη από δεξιά θα πρέπει να βρίσκεται ανάμεσα στις παρατηρήσεις r_{n+1} και r_N , ενώ αν έχουμε μια παρατήρηση r_i αποκομμένη από αριστερά, θα βρίσκεται ανάμεσα στις παρατηρήσεις 1 και $N-n$.

5. Στο πέμπτο βήμα βρίσκουμε το μοντέλο παλινδρόμησης για τη μέση απόκριση και την τυπική απόκλιση για κάθε μια από τις επαναλήψεις.

Αν ορίσουμε τους πίνακες $R = [R_U | \hat{R}_C]^T$ και $Z = [Z_U | Z_C]^T$ τότε μπορούμε να υπολογίσουμε για τη j -οστή επανάληψη τη μέση τιμή των τάξεων R_j και την τυπική απόκλιση των τάξεων S_j . Στη συνέχεια, βρίσκουμε τη σχέση μεταξύ των R_j και Z , καθώς και των S_j και Z από τις παρακάτω σχέσεις:

$$\hat{\mu}_{R|Z} = Z \hat{\beta}_R \quad (4.4.2)$$

$$\text{και } \hat{\mu}_{S|Z} = Z\hat{\beta}_S \quad (4.4.3)$$

όπου $\hat{\beta}_R$ και $\hat{\beta}_S$ είναι οι πίνακες των εκτιμητριών των συντελεστών παλινδρόμησης.

6. Εντοπισμός των παραγόντων που επηρεάζουν σημαντικά τη μέση απόκριση, καθώς και την τυπική απόκλιση. Σ' αυτό μπορούν να φανούν χρήσιμα τα γραφήματα των εκτιμητριών $\hat{\beta}_R$ και $\hat{\beta}_S$.
7. Εντοπισμός του βέλτιστου συνδυασμού σταθμών παραγόντων.

Η παραπάνω διαδικασία εφαρμόζεται, κυρίως, σε επαναλαμβανόμενα πειράματα που περιέχουν αποκομμένες παρατηρήσεις. Από την άλλη πλευρά, στην περίπτωση των μη επαναλαμβανόμενων πειραμάτων με αποκομμένες παρατηρήσεις, η διαδικασία ανάλυσης είναι η ίδια με τη μόνη διαφορά ότι λαμβάνεται υπ' όψιν μόνο η βελτιστοποίηση της απόκρισης.

4.4.1.1 Αριθμητικό παράδειγμα

Ας δούμε ένα αριθμητικό παράδειγμα που θα βοηθήσει στην καλύτερη κατανόηση της μη παραμετρικής διαδικασίας ανάλυσης δεδομένων με αποκομμένες παρατηρήσεις, η οποία περιγράφηκε στην προηγούμενη παράγραφο. Το σύνολο των δεδομένων που θα μελετηθεί αποτελείται από 16 παρατηρήσεις ενός πειράματος πέντε παραγόντων, καθένας εκ των οποίων λαμβάνει τιμές σε 2 στάθμες. Στον πίνακα 4.4 παρουσιάζονται τα δεδομένα του προβλήματος. Υποτίθεται ότι η μικρότερη ποιότητα των χαρακτηριστικών είναι η επιθυμητή, ενώ δεν προσδιορίζεται η ακριβής τιμή της απόκρισης πέντε παρατηρήσεων. Τα συγκεκριμένα δεδομένα είναι από δεξιά αποκομμένα και στον πίνακα 4.4 συμβολίζονται με (*). Το σημείο αποκοπής είναι η τιμή 67.

Trial	Factors					Responses		R_U	
	A	B	C	D	E				
1	1	1	1	1	1	66	66	10.5	10.5
2	1	1	2	2	2	*(68)	63	*	7.5
3	1	2	1	2	2	*(80)	*(88)	*	*
4	1	2	2	1	1	63	65	7.5	9
5	2	1	1	1	2	*(73)	*(71)	*	*
6	2	1	2	2	1	37	42	1	4
7	2	2	1	2	1	38	39	2	3
8	2	2	2	1	2	57	48	6	5

Note:

* censored data; the entries in parentheses are the values of the censored data

Πίνακας 4.4

Για την εφαρμογή της μεθόδου ακολουθούνται τα βήματα που παρουσιάστηκαν στην παράγραφο 4.4.

1. Αρχικά, ορίζεται το διάνυσμα Y_U που αποτελείται από τις πλήρεις παρατηρήσεις και στη συνέχεια, οι τιμές αυτές διατάσσονται κατά αύξουσα σειρά μεγέθους στο διάνυσμα R_U . Συνεπώς από τον πίνακα 4.4 έχουμε:

$$Y_U = [66,66,63,63,65,37,42,38,39,57,48]^T$$

και

$$R_U = [10.5,10.5,7.5,7.5,9,1,4,2,3,6,5]^T$$

2. Εν συνεχεία, ορίζουμε τον πίνακα, Z_U , των σταθμών των παραγόντων για τα πλήρη δεδομένα.

A	B	C	D	E	AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
-----	-----	-----	-----	-----	------	------	------	------	------	------	------	------	------	------

$$Z_U = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 & 1 & 1 & 2 & 2 & 1 & 1 & 4 & 2 & 2 & 2 & 1 \\ 1 & 2 & 1 & 2 & 2 & 1 & 2 & 4 & 4 & 2 & 2 & 2 & 1 & 4 & 2 \\ 1 & 2 & 2 & 1 & 2 & 1 & 4 & 2 & 4 & 2 & 2 & 4 & 2 & 2 & 1 \\ 1 & 2 & 2 & 2 & 1 & 2 & 4 & 4 & 2 & 4 & 4 & 2 & 4 & 2 & 4 \end{pmatrix}_{5 \times 16}$$

Με βάση τον πίνακα Z_U και τη σχέση (4.4.1) υπολογίζεται η εκτιμήτρια \hat{Y}_U .

$$\hat{\mu}_{Y_U|Z_U} = Z_U \hat{\beta}_U$$

$$= 88 - 17,5A - 1,5B - 0,5C - 8,5D + 6E \quad (4.4.4)$$

3. Στο τρίτο βήμα πραγματοποιείται η εκτίμηση των αποκομμένων παρατηρήσεων, δηλαδή υπολογίζεται η εκτιμήτρια \hat{Y}_C . Οι εκτιμήσεις των αποκομμένων παρατηρήσεων βρίσκονται αντικαθιστώντας στην εξίσωση (4.4.4) τα επίπεδα όπου παρατηρούνται οι αποκομμένες παρατηρήσεις. Συνεπώς, οι εκτιμήτριες των αποκρίσεων για τις εκτελέσεις 2, 3 και 5 ισούνται με 63, 62 και 54,5 αντίστοιχα.
4. Από τη στιγμή που τα δεδομένα είναι αποκομμένα από τα δεξιά και υπάρχουν πέντε αποκομμένες τιμές, οι τάξεις των εκτιμητριών αυτών των παρατηρήσεων κυμαίνονται μεταξύ των αριθμών 12 και 16. Συνεπώς, $\hat{R}_C = [16, 14, 15, 12, 13]^T$.
5. Στη συνέχεια, ορίζονται οι πίνακες $R = [R_U | \hat{R}_C]^T$ και $Z = [Z_U | Z_C]^T$, μέσω των οποίων υπολογίζεται για τη j -οστή επανάληψη η μέση τιμή των τάξεων R_j και η τυπική απόκλιση των τάξεων S_j . Τα αποτελέσματα αυτής της διαδικασίας συνοψίζονται στον πίνακα 4.5.

Trial	1	2	3	4	5	6	7	8
R_j	10.5	11.75	14.5	8.25	12.5	3	2.5	5.5
S_j	0	3.324	0.5	0.75	0.5	1.5	0.5	0.5

Πίνακας 4.5

Το μοντέλο παλινδρόμησης για τα R_j , $j = 1, 2, \dots, 8$ και Z καθώς και για τα S_j και Z δίνονται από τις σχέσεις (4.4.5) και (4.4.6), αντίστοιχα.

$$\begin{aligned} \hat{\mu}_{R|Z} &= Z\hat{\beta}_R \\ &= 7,25 + 1,75A + 4,25B - 1,75C - 1,25D + 5E - 4AB - 0,75AC \end{aligned} \quad (4.4.5)$$

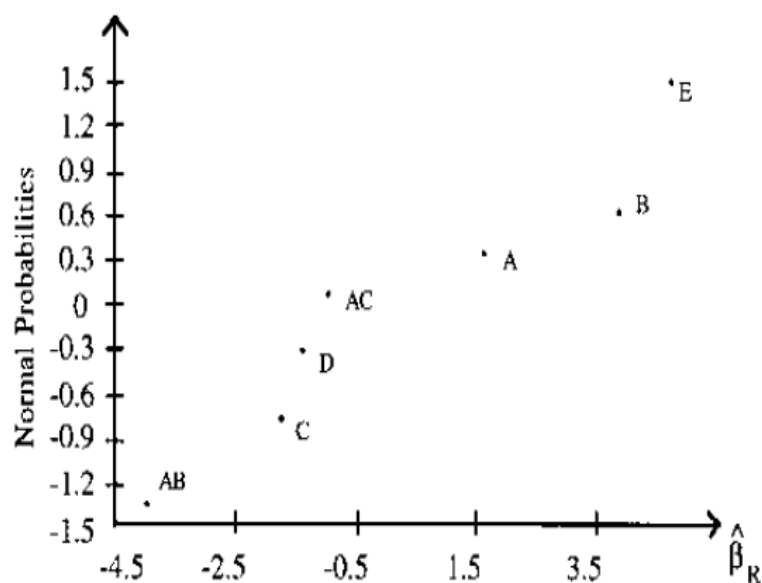
$$\hat{\mu}_{S|Z} = Z\hat{\beta}_S$$

$$= -3,0185 + 0,7315A - 1,574B + 3,074C + 1,0185D + 0,5185E + 0,537AB - 1,287AC \quad (4.4.46)$$

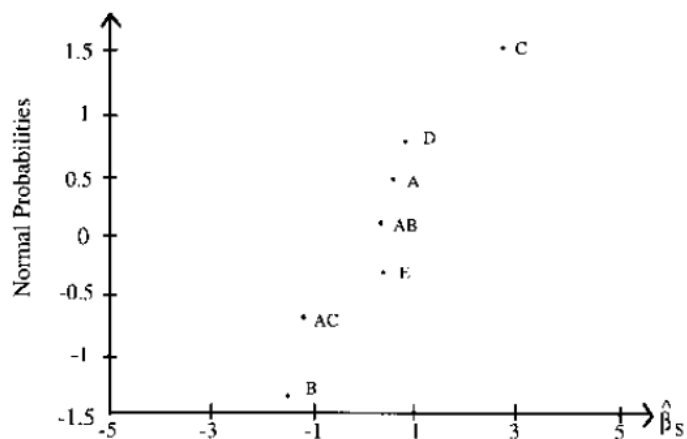
6. Οι κανονικές πιθανότητες για τους συντελεστές των εξισώσεων υπολογίζονται στον πίνακα 4.6. Ενώ, στα γραφήματα 4.4 και 4.5, αντίστοιχα, παρουσιάζονται τα διαγράμματα κανονικής πιθανότητας. Από τα συγκεκριμένα διαγράμματα συμπεραίνεται ότι οι παράγοντες AB, A, B και E επηρεάζουν σημαντικά τη μέση απόκριση, ενώ οι παράγοντες AC, B και C έχουν σημαντικές επιπτώσεις στη τυπική απόκλιση.

Trial	1	2	3	4	5	6	7	8
R_j	10.5	11.75	14.5	8.25	12.5	3	2.5	5.5
S_j	0	3.324	0.5	0.75	0.5	1.5	0.5	0.5

Πίνακας 4.6



Γράφημα 4.4



Γράφημα 4.5

7. Τέλος, αναζητάται ο βέλτιστος συνδυασμός των παραγόντων, κάτι που αποτελεί και τον κύριο στόχο της ανάλυσης που εφαρμόζεται.

Από την εξίσωση (4.4.5) συμπεραίνεται ότι η τιμή του παράγοντα E θα πρέπει να αντιστοιχεί στη χαμηλή του στάθμη, καθώς η επιθυμητή απόκριση είναι η μικρότερη, δηλαδή το πείραμα μας ανήκει στην κατηγορία smaller-the-better. Στον πίνακα 4.7 παρουσιάζονται οι τιμές της εκτιμήτριας $\hat{\mu}_{R|Z}$ για τις διάφορες στάθμες των παραγόντων.

Factors (regression coefficients)	A (1.75)	B (4.25)	AB (-4)	The difference in $\hat{\mu}_{R Z}$
Level setting 1	1	1	1	2
Level setting 2	1	2	2	2.25
Level setting 3	2	1	2	-0.35
Level setting 4	2	2	4	-4(min.)

Πίνακας 4.7

Από τη σύγκριση των διάφορων τιμών της $\hat{\mu}_{R|Z}$ και δεδομένου ότι η μικρότερη τιμή είναι η επιθυμητή, οι παράγοντες A και B θα πρέπει να αντιστοιχούν στη χαμηλή τους στάθμη. Συνεπώς, ο ενδεικτικός βέλτιστος συνδυασμός των επιπέδων των παραγόντων που προκύπτει είναι ο $A_2B_2E_1$, όπου με (2) συμβολίζεται η υψηλή στάθμη κάθε παράγοντα και με (1) η χαμηλή. Ομοίως, στον πίνακα 4.8 παρουσιάζονται οι διάφορες τιμές της εκτιμήτριας $\hat{\mu}_{S|Z}$. Από τη σύγκριση των τιμών του πίνακα 4.8 και από την εξίσωση (4.4.6) προκύπτει ότι ο ενδεικτικός βέλτιστος

συνδυασμός των σταθμών των παραγόντων για την επίδραση στη διασπορά, είναι ο εξής: $A_2B_2C_1$. Δεδομένου ότι οι δύο παραπάνω περιπτώσεις αντιστοιχούν η μία στην άλλη, προκύπτει ο βέλτιστος συνδυασμός των παραγόντων επιπέδων του πειράματος, $A_2B_2C_1E_1$.

Factors (regression coefficients)	A (0.7315)	B (3.074)	AB (-1.287)	The difference in $\hat{\mu}_{S Z}$
Level setting 1	1	1	1	2.5185
Level setting 2	1	2	2	4.3055
Level setting 3	2	1	2	1.963(min.)
Level setting 4	2	2	4	2.463

Πίνακας 4.8

Εάν, γινόταν η ανάλυση των 16 δεδομένων που παρουσιάζονται στον πίνακα 4.4 με τη μέθοδο ANOVA, ο βέλτιστος συνδυασμός παραγόντων που θα προέκυπτε θα ήταν ο $A_2B_2C_2E_1$. Τα αποτελέσματα για τις δύο τεχνικές είναι αρκετά παρόμοια. Κατά συνέπεια, αποδεικνύεται η ισχύς της προτεινόμενης μεθόδου.

4.4.1.2 Συμπεράσματα

Μια αποτελεσματική διαδικασία βασισμένη στην τάξη των παρατηρήσεων και στην ανάλυση παλινδρόμησης αναλύθηκε στη παράγραφο 4.4 για την ανάλυση ενός πειράματος που περιέχει κάποιες αποκομμένες παρατηρήσεις. Η προτεινόμενη διαδικασία είναι αρκετά απλούστερη από κάποιες συμβατικές μεθόδους, όπως η εκτίμηση μέγιστης πιθανοφάνειας. Για να εξασφαλιστεί η αποτελεσματικότητα της συγκεκριμένης μεθόδου προτείνεται τουλάχιστον τα δύο τρίτα των δεδομένων που εμπλέκονται σε ένα πείραμα να είναι πλήρη, διαφορετικά δεν υπάρχουν αρκετά στοιχεία για την ανάλυση. Από την άλλη πλευρά, αν ο συντελεστής προσδιορισμού του προβλεπόμενου μοντέλου παλινδρόμησης είναι πολύ χαμηλός, συνίσταται η επανεξέταση των παραγόντων του σχεδιασμού, διαφορετικά η προτεινόμενη μέθοδος θα πρέπει να απορριφθεί και να χρησιμοποιηθεί κάποια εναλλακτική. Η προτεινόμενη διαδικασία, αν και υστερεί αυστηρής θεωρητικής απόδειξης, μπορεί εύκολα να υλοποιηθεί σε βιομηχανικά πειράματα.

4.4.2 Μη παραμετρική μέθοδος ανάλυσης παραγοντικών σχεδιασμών με αποκομμένες παρατηρήσεις των Michael G. Akritas και Edgar Brunner

Τα τελευταία χρόνια υπάρχει μεγάλο ενδιαφέρον γύρω από μη παραμετρικές μεθόδους ανάλυσης γραμμικών μοντέλων παλινδρόμησης με αποκομμένα δεδομένα. Με εξαίρεση τη δουλειά των Akritas και La Valley (1996), οι οποίοι θεώρησαν ένα γραμμικό μοντέλο για παραγοντικούς σχεδιασμούς και ανέπτυξαν διαδικασίες ελέγχου για παραμετρικές υποθέσεις των μη κύριων επιδράσεων, των μη αλληλεπιδράσεων κτλ., φαίνεται ότι οι παραγοντικοί σχεδιασμοί δεν έχουν συμπεριληφθεί στη πρόσφατη επιστημονική δραστηριότητα.

Η παρούσα παράγραφος αφορά την ανάλυση παραγοντικών σχεδιασμών με ανεξάρτητες παρατηρήσεις στο πλαίσιο ενός μη παραμετρικού μοντέλου και την ανάπτυξη διαδικασιών ελέγχου μη παραμετρικών υποθέσεων. Αντίστοιχες μη παραμετρικές υποθέσεις παρουσιάστηκαν από τους Akritas και Arnold (1994) στο πλαίσιο των σχεδιασμών επαναλαμβανόμενων μετρήσεων και επεκτάθηκαν στους παραγοντικούς σχεδιασμούς με ανεξάρτητες παρατηρήσεις από τους Akritas, Arnold και Brunner (1997). Συνεπώς, οι παρούσες μέθοδοι διαφέρουν από αυτές των Akritas και La Valley.

Οι διαδικασίες ελέγχου που αναφέρονται στην παράγραφο είναι μη παραμετρικές και ως εκ τούτου μπορούν να θεωρηθούν ως προεκτάσεις σε παραγοντικούς σχεδιασμούς διαδικασιών ελέγχου ισότητας k δειγμάτων (Andersen et al. 1982). Αυτά τα στατιστικά προκύπτουν από τα επιχειρήματα του Mantel (1966), τα σταθμισμένα log-rank στατιστικά των Tarone και Ware (1977) και τη γενική τάξη K ελέγχων του Gill (1980). Συγκεκριμένα, ο Mantel θεωρούσε τα δεδομένα σαν μια σειρά $k \times 2$ πινάκων για κάθε μία από τις διαφορετικές στιγμές αποκοπής, εφάρμοζε τον έλεγχο των Mantel και Haenszel (1959) και συνδύαζε τους πίνακες σαν να ήταν ανεξάρτητοι. Το τελικό στατιστικό των Mantel και Haenszel (log-rank statistic) βασίζεται στο άθροισμα των διανυσμάτων των συχνοτήτων των σπανιότερων εμφανιζόμενων παρατηρήσεων για κάθε ένα από τους $k \times 2$ πίνακες:

$$LR = \sum_{l=1}^L (D_{1l} - E(D_{1l}), \dots, D_{kl} - E(D_{kl}))$$

όπου L είναι ο συνολικός αριθμός των διαφορετικών χρόνων αποκοπής, D_{il} το πλήθος των αποτυχιών του δείγματος i την στιγμή αποκοπής l και $E(D_{il})$ η αναμενόμενη τιμή κάτω από τη μηδενική υπόθεση.

Έστω ένας $a \times b$ παραγοντικός σχεδιασμός, ο οποίος περιέχει $k = ab$ πληθυσμούς οι οποίοι προκύπτουν από όλους τους δυνατούς συνδυασμούς των επιπέδων παραγόντων. Η προηγούμενη τάξη ελέγχων η οποία εξετάζει την ομοιότητα των k πληθυσμών δεν έχει επεκταθεί προκειμένου να ελέγχει πιο εξειδικευμένες υποθέσεις οι οποίες ενδιαφέρουν τους πειραματιστές στην περίπτωση ανάλυσης δύο παραγόντων. Μερικές από αυτές τις υποθέσεις αφορούν την αλληλεπίδραση των παραγόντων. Για την καλύτερη κατανόηση της δυσκολίας που προκύπτει είναι απαραίτητο να ληφθεί υπ' όψιν το συμπέρασμα του Mantel για τον έλεγχο log-rank: Κάτω από μία εξειδικευμένη μηδενική υπόθεση είναι αδύνατον να υπολογιστεί με μη παραμετρικό τρόπο η ποσότητα $E(D_{(rc)}, l)$.

Στην παρούσα παράγραφο παρουσιάζεται ένας τρόπος για να ξεπεραστεί η δυσκολία υπολογισμού της $E(D_{(rc)}, l)$. Οι μη παραμετρικές υποθέσεις των Akritas και Arnold (1994) αποτελούν τη βάση της μεθόδου. Οι υποθέσεις αυτές γίνονται στο πλαίσιο ενός γενικού μη παραμετρικού μοντέλου, το οποίο ορίζει μόνο ότι οι ποσότητες T_{rci} έχουν μια συνάρτηση κατανομής F_{rc} , όπου T_{rci} είναι η τυχαία μεταβλητή που αντιστοιχεί στο αντικείμενο i του κελιού (r, c) .

Έστω $\mathbf{F} = (F_{11}, \dots, F_{1b}, \dots, F_{ab})'$, τότε οι μη παραμετρικές υποθέσεις των μη κύριων επιδράσεων, της μη αλληλεπίδρασης και των μη απλών επιδράσεων μπορούν να γραφτούν στη μορφή:

$$H_0(A): \mathbf{C}_A \mathbf{F} = \mathbf{0}, \quad H_0(AB): \mathbf{C}_{AB} \mathbf{F} = \mathbf{0}, \quad H_0(A|B): \mathbf{C}_{A|B} \mathbf{F} = \mathbf{0}$$

όπου \mathbf{C} κατάλληλοι πίνακες αντιθέσεων. Σε προηγούμενες εργασίες (Akritas and Arnold 1994, Akritas et al. 1997) το στατιστικό έλεγχο που χρησιμοποιήθηκε για τις παραπάνω υποθέσεις είναι σε τετραγωνική μορφή βασισμένο στο διάνυσμα $\mathbf{C} \int (1 - \hat{H}(x)) d\hat{\mathbf{F}}(x)$, όπου $\hat{\mathbf{F}} = (\hat{F}_{11}, \dots, \hat{F}_{1b}, \dots, \hat{F}_{ab})'$ είναι το διάνυσμα των εμπειρικών συναρτήσεων κατανομών για κάθε κελί και \hat{H} είναι η εμπειρική συνάρτηση κατανομής για το συνδυασμένο δείγμα.

Στη συνέχεια εφαρμόζεται η ίδια προσέγγιση με τη διαφορά ότι χρησιμοποιείται η εκτιμήτρια Kaplan-Meier στη θέση της εμπειρικής συνάρτησης κατανομής.

4.4.2.1 Ο στατιστικός έλεγχος

Για την ανάλυση της μεθόδου χρησιμοποιείται ένας $a \times b$ παραγοντικός σχεδιασμός. Ο σχεδιασμός λαμβάνει υπ' όψιν του τον παράγοντα A ο οποίος λαμβάνει τιμές σε a στάθμες και τον παράγοντα B που λαμβάνει τιμές σε b στάθμες. Για κάθε συνδυασμό επιπέδων των παραγόντων έχουν ληφθεί αρκετές παρατηρήσεις οι οποίες έχουν κατανεμηθεί σε ένα πίνακα. Αρχικά, από κάθε κελί (r,c) του πίνακα λαμβάνεται η πληροφορία (X_{rck}, Δ_{rck}) , $k = 1, \dots, n_{rc}$. Η ποσότητα X_{rck} εκφράζει την ελάχιστη τιμή μεταξύ του ζητούμενου χρόνου αποτυχίας T_{rck} και της αποκομμένης παρατήρησης C_{rck} . Επίσης, $\Delta_{rck} = I(X_{rck} = T_{rck})$, όπου με τον όρο $I(E)$ υποδηλώνεται ο δείκτης του γεγονότος E. Οι δείκτες r και c εκφράζουν τα επίπεδα των παραγόντων A και B, αντίστοιχα, ενώ ο όρος k χρησιμοποιείται για την αρίθμηση των παρατηρήσεων κάθε κελιού (r,c) .

Για την υλοποίηση της μεθόδου θεωρείται ότι οι παρατηρήσεις που βρίσκονται σε διαφορετικά κελιά είναι ανεξάρτητες μεταξύ τους και ότι οι αποκομμένες παρατηρήσεις είναι ανεξάρτητες σε σχέση με τις μη αποκομμένες. Επίσης, θεωρείται ότι οι παρατηρήσεις T_{rck} είναι ανεξάρτητες και ακολουθούν την κατανομή F_{rc} , ενώ οι αποκομμένες παρατηρήσεις C_{rck} είναι ανεξάρτητες και ακολουθούν την κατανομή G_{rc} . Τέλος, οι κατανομές F_{rc} και G_{rc} πρέπει να ικανοποιούν την σχέση:

$$\int_0^{\tau_{rc}} \frac{dF_{rc}(s)}{1 - G_{rc}(s-)} < \infty, \quad \forall(r, c)$$

όπου $\tau_{rc} = \sup\{t: (1 - F_{rc}(t-))(1 - G_{rc}(t-)) > 0\}$.

Έστω $\tau_L = \tau_{11} \wedge \dots \wedge \tau_{ab}$ και $\tau_U = \tau_{11} \vee \dots \vee \tau_{ab}$. Επειδή οι μη παραμετρικές υποθέσεις περιλαμβάνουν τις συναρτήσεις κατανομών όλων των κελιών, δεν μπορούν να ελεγχθούν πέρα της τιμής τ_L . Έστω $T_L = T_{11} \wedge \dots \wedge T_{ab}$ και $T_U = T_{11} \vee \dots \vee T_{ab}$, όπου με T_{rc} συμβολίζεται η μεγαλύτερη παρατήρηση που περιλαμβάνεται στο κελί (r,c) . Με βάση τα παραπάνω, οι στατιστικοί έλεγχοι που αναπτύσσονται είναι βασισμένοι σε σταθμισμένες συγκρίσεις των εκτιμητριών Kaplan-Meier που τερματίζουν στην τιμή T_L . Επιπλέον, αν ικανοποιείται τουλάχιστον μία από τις υποθέσεις

$$\tau_L = \tau_U \quad \text{ή} \quad \tau_{F_{rc}} < \tau_{G_{rc}}, \quad \forall(r, c),$$

τότε οι συγκρίσεις μπορούν να επεκταθούν μέχρι την τιμή T_U . Σημειώνεται ότι $\tau_F = \sup\{t: 1 - F(t-) > 0\}$ για κάθε συνάρτηση κατανομής F . Η πρώτη από τις παραπάνω υποθέσεις είναι πιο εύκολο να συναντηθεί στην πράξη, παραδείγματος χάριν όταν σε μία μελέτη έχει οριστεί μια κοινή χρονική στιγμή τερματισμού για όλα τα κελιά, ενώ η δεύτερη δίνει τη δυνατότητα να συμπεριληφθούν οι αποκομμένες παρατηρήσεις. Στη συνέχεια, συμβολίζουμε $T = T_L$ και $\tau = \tau_L$, ενώ οι διαδικασίες ελέγχου που αναπτύσσονται χρησιμοποιούν συγκρίσεις που τερματίζουν στην τιμή T_L .

Έστω $\mathbf{F} = (F_{11}, \dots, F_{1b}, \dots, F_{ab})'$ και $\hat{\mathbf{F}} = (\hat{F}_{11}, \dots, \hat{F}_{1b}, \dots, \hat{F}_{ab})'$ τα διανύσματα των συναρτήσεων κατανομής και των εκτιμητριών Kaplan-Meier, αντιστοίχως. Σύμφωνα με την προηγούμενη παράγραφο, το στατιστικό ελέγχου για τη μηδενική υπόθεση $H_0: \mathbf{C}\mathbf{F} = 0$ βασίζεται στο διάνυσμα:

$$\mathbf{C} \int_0^T \hat{S}_H d\hat{\mathbf{F}}$$

όπου $\hat{S}_H(s) = 0.5n^{-1}(Y_{..}(s) + Y_{..}(s-))$ και \mathbf{C} ένας πίνακας αντιθέσεων.

Συγκεκριμένα το στατιστικό ελέγχου δίνεται από την σχέση:

$$Q(\mathbf{C}) = n(\mathbf{C} \int_0^T \hat{S}_H d\hat{\mathbf{F}})'(\mathbf{C}\hat{\mathbf{V}}\mathbf{C}')^{-1} \times (\mathbf{C} \int_0^T \hat{S}_H d\hat{\mathbf{F}})$$

όπου $\hat{\mathbf{V}}$ ένας διαγώνιος πίνακας $ab \times ab$ με στοιχεία $(n/n_{rc})\hat{\sigma}_{rc}(T)$ και \mathbf{C} ένας $k \times ab$ ($k < ab$) πίνακας αντιθέσεων. Το στατιστικό ελέγχου ακολουθεί την χ-τετράγωνο κατανομή με k βαθμούς ελευθερίας (χ_k^2).

4.4.2.2 Προσέγγιση μικρού δείγματος

Μελέτες προσομοίωσης δείχνουν ότι όταν ο αριθμός των επιπέδων παραγόντων είναι μεγάλος, η προσέγγιση της ασυμπτωτικής κατανομής χ^2 των στατιστικών $Q(\mathbf{C}_A)$, $Q(\mathbf{C}_B)$, $Q(\mathbf{C}_{AB})$ είναι μάλλον αργή. Συγκεκριμένα, για μικρά δείγματα και πολυάριθμα επίπεδα παραγόντων, οι διαδικασίες ελέγχου είναι περισσότερο ή λιγότερο ελεύθερες. Συνεπώς απαιτείται μια κατάλληλη προσέγγιση για μικρά δείγματα. Εδώ χρησιμοποιείται η ιδέα της προσέγγισης του Box (1954), η οποία είναι υπολογιστικά απλή και η προσαρμογή της σε μικρά δείγματα είναι αρκετά καλή.

Έστω $C_1 = P_a \otimes b^{-1} \mathbf{1}'_b$, $C_2 = P_a \otimes P_b$ και $C_3 = P_a \otimes I_b$, όπου για κάθε ακέραιο d ισχύει $P_d = I_d - d^{-1} \mathbf{1}_d \mathbf{1}'_d$. Οι πίνακες, αυτοί, είναι πίνακες αντιθέσεων οι οποίοι μπορούν να χρησιμοποιηθούν με σκοπό την εναλλακτική περιγραφή της μηδενικής υπόθεσης. Συγκεκριμένα, η μηδενική υπόθεση έχει ως εξής:

$$H_0(A): C_1 F = \mathbf{0}, H_0(B): C_2 F = \mathbf{0} \text{ και } H_0(AB): C_3 F = \mathbf{0}.$$

Έστω, τώρα, C ένας από τους προαναφερθέντες πίνακες αντιθέσεων και έστω $W = C'(CC')^{-1}C$, όπου $(CC')^{-1}$ ένας γενικός αντίστροφος του CC' . Σημειώνεται ότι ο πίνακας W είναι ένας πίνακας προβολών, του οποίου τα διαγώνια στοιχεία ισούνται με m . Εάν $\tilde{Q}(W) = \hat{p}'(T)W\hat{p}(T)$, όπου $\hat{p}(T) = \int_0^T \hat{S}_H d\hat{F}$, το στατιστικό ελέγχου σύμφωνα με την προσέγγιση του Box δίνεται από τη σχέση:

$$B_n(W) = n\tilde{Q}(W)/[m \cdot tr(\hat{V})]$$

και ακολουθεί ασυμπτωτικά μια $\chi^2_{\hat{f}}/\hat{f}$ κατανομή με βαθμούς ελευθερίας

$\hat{f} = m^2 [tr(\hat{V})]^2 / tr(W\hat{V}W\hat{V})$, όπου $tr(\cdot)$ συμβολίζει το ίχνος ενός τετραγωνικού πίνακα.

Μια ένδειξη της συμπεριφοράς των στατιστικών καθώς και της ποιότητας των προσεγγίσεων στην περίπτωση των αποκομμένων δεδομένων φαίνεται στον Πίνακα 4.9, στον οποίο παρουσιάζονται τα αποτελέσματα μιας μελέτης προσομοίωσης.

$n = 15$ Level	Main effects				Interaction	
	A(a = 2)		B(b = 3)		AB(2 × 3)	
	χ^2_1	Box approximation	χ^2_2	Box approximation	χ^2_2	Box approximation
.10	.103	.102	.125	.108	.125	.104
.05	.057	.055	.071	.055	.067	.053
.01	.013	.011	.015	.011	.016	.012

Πίνακας 4.9

Στις μελέτες χρησιμοποιήθηκε ένας 2×3 σχεδιασμός με ισομεγέθη δείγματα $n_{ij} = n = 15$, όπου η κατανομή των χρόνων επιβίωσης είναι η λογαριθμοκανονική (*lognormal*), δηλαδή $T_{ijk} \sim \exp(Z_{ij})$, όπου οι ποσότητες Z_{ij}

είναι κανονικά κατανομημένες με διακύμανση ίση με 1 και μέση τιμή $\mu_{ij} = 0, i = 1,2, j = 1,2,3$. Η αποκομμένη μεταβλητή προκύπτει σύμφωνα με την ίδια κατανομή βήτα για κάθε κελί, με αποτέλεσμα 18% αποκοπή. Για κάθε περίπτωση έγιναν 5.000 προσομοιώσεις. Τα αποτελέσματα των προσομοιώσεων με τα περισσότερα επίπεδα ανά παράγοντα, έδειξαν ότι το στατιστικό $B_n(W)$ έχει καλή εφαρμογή σε τέτοιες περιπτώσεις. Τα αποτελέσματα για τα προσομοιωμένα 10%, 5% και 1% επίπεδα των στατιστικών $Q(C_A)$, $Q(C_B)$ και $Q(C_{AB})$ παρουσιάζονται στον Πίνακα 4.9.

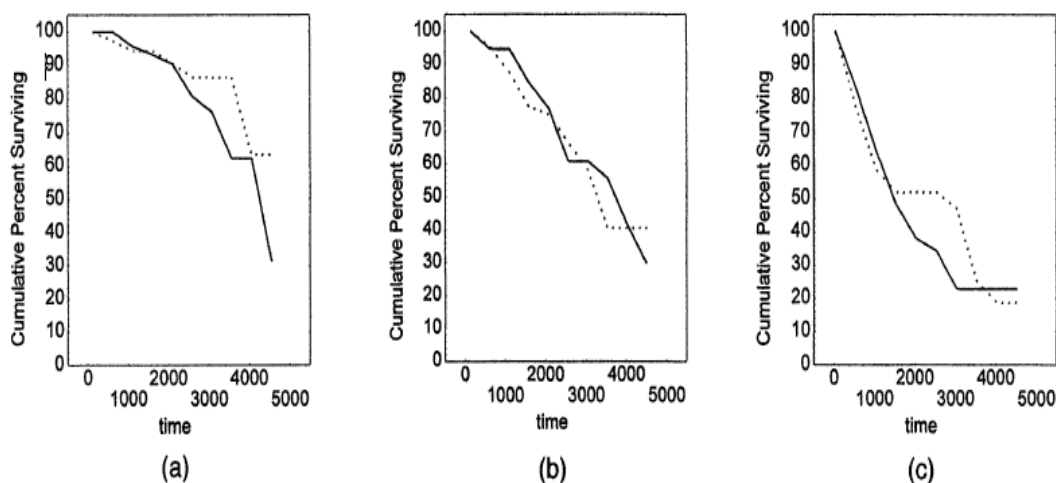
Στη συνέχεια γίνεται σύγκριση μεταξύ του προτεινόμενου στατιστικού και του ελέγχου log-rank. Η σύγκριση γίνεται μέσω $\alpha=3$ αγωγών, χρησιμοποιώντας τυχαίες μεταβλητές που ακολουθούν εκθετική κατανομή και έγιναν $n \equiv n_i = 15$ επαναλήψεις. Στην περίπτωση που $X_{ij} \sim \exp(1 + \delta(i - 1))$ φαίνεται να είναι βέλτιστος ο έλεγχος log-rank, ενώ στην περίπτωση που $X_{ij} \sim \exp(1) + \delta(i - 1)$, $i = 1,2,3, j = 1, \dots, 15$, φαίνεται να εφαρμόζεται καλύτερα η προτεινόμενη μέθοδος. Στον Πίνακα 4.10 παρουσιάζονται τα αποτελέσματα της σύγκρισης.

$X_{ij} \sim \exp(1 + \delta(i - 1))$			$X_{ij} \sim \exp(1) + \delta(i - 1)$		
δ	Log-Rank	$B_n(W)$	δ	Log-Rank	$B_n(W)$
0	.071	.053	0	.067	.057
.5	.400	.284	.15	.207	.228
1	.722	.575	.2	.329	.365
1.5	.872	.759	.3	.618	.677
2	.942	.863	.4	.864	.896

Πίνακας 4.10

4.4.2.3 Αριθμητικό παράδειγμα και ανάλυση δεδομένων

Στο παράδειγμα που ακολουθεί προσαρμόζονται τα δεδομένα που προέκυψαν κατά τη διάρκεια μια τυχαιοποιημένης κλινικής δοκιμής για την μελέτη της πρωτοπαθούς χολικής κίρρωσης (Fleming and Harrington, 1991). Το ενδιαφέρον της έρευνας εστιάζεται στην επίδραση των δύο θεραπειών, D πενικιλλαμίνη και P εικονικό φάρμακο, στον χρόνο μεταξύ της έναρξης της έρευνας και του προγενέστερου θανάτου, της μεταμόσχευσης ήπατος ή της ανάλυσης της μελέτης. Ο παράγοντας A αντιστοιχεί στην θεραπεία (δύο επίπεδα), ενώ ο παράγοντας B αντιστοιχεί στο ιστολογικό στάδιο της νόσου (τρία επίπεδα). Τα μεγέθη των δειγμάτων είναι $n_{11} = 47$ (11), $n_{12} = 36$ (6), $n_{21} = 56$ (20), $n_{22} = 64$ (23), $n_{31} = 55$ (34) και $n_{32} = 54$ (31), για ένα σύνολο 312 παρατηρήσεων, όπου οι αριθμοί στις παρενθέσεις αντιστοιχούν στις αποκομμένες παρατηρήσεις. Στο γράφημα 4.6 παρουσιάζονται οι εκτιμήτριες Kaplan – Meier για τις δύο διαφορετικές θεραπείες συναρτήσει των διαφόρων σταδίων πάθησης.



Γράφημα 4.6

Το γεγονός της διασταύρωσης των συναρτήσεων επιβίωσης δικαιολογεί την χρήση της προτεινόμενης μεθόδου. Τα αποτελέσματα των ελέγχων της μηδενικής υπόθεσης είναι τα εξής: $Q(C_A) = 1,22$ ($p = 0.27$), $Q(C_B) = 58,78$ ($p < 0.00005$) και $Q(C_{AB}) = 1,11$ ($p = 0.58$). Συνεπώς, φαίνεται ότι η επίδραση τόσο της αλληλεπίδρασης και της θεραπείας δεν είναι σημαντική, ενώ η επίδραση του σταδίου της νόσου είναι αρκετά σημαντικός.

4.4.2.4 Συμπεράσματα

Στην παραπάνω ενότητα παρουσιάστηκε η μέθοδος των Akritas και Brunner (1997) για τον έλεγχο της υπόθεσης της μη ύπαρξης κύριων επιδράσεων, αλληλεπιδράσεων και απλών επιδράσεων σε παραγοντικούς σχεδιασμούς με αποκομμένες παρατηρήσεις. Οι διαδικασίες αναπτύχθηκαν στο πλαίσιο ενός μη παραμετρικού μοντέλου το οποίο δεν υπαγορεύει κάποια δομή για τις κατανομές που παράγουν τα δεδομένα για κάθε συνδυασμό παραγόντων επιπέδων. Οι προσομοιώσεις έδειξαν ότι για μικρά δείγματα με μετρίως μεγάλο αριθμό επιπέδων παραγόντων η ασυμπτωτική θεωρία δεν παρέχει καλή προσέγγιση της κατανομής του στατιστικού. Για την αντιμετώπιση της συγκεκριμένης δυσκολίας αναπτύχθηκε μια τροποποίηση της μεθόδου βασισμένη στην ιδέα του Box (1954). Οι διαδικασίες ελέγχου που αναφέρθηκαν είναι εύκολα υλοποιήσιμες και οι προσομοιώσεις δείχνουν ότι έχουν καλές ιδιότητες ισχύος.

4.4.3 Μη παραμετρική μέθοδος ανάλυσης παραγοντικών σχεδιασμών με αποκομμένες παρατηρήσεις των John T. O' Gorman και Michael G. Akritas.

Σκοπός της παρούσας παραγράφου είναι η μη παραμετρική προσέγγιση της ανάλυσης των σχεδιασμών επαναλαμβανόμενων μετρήσεων που περιλαμβάνουν αποκομμένες παρατηρήσεις.

Προκειμένου να παρουσιαστεί μια συνοπτική περιγραφή του μη παραμετρικού μοντέλου και των στατιστικών ελέγχων, που θα εφαρμόζονται σε όλους τους σχεδιασμούς επαναλαμβανόμενων μετρήσεων και σε όλες τις υποθέσεις που τυχόν να ενδιαφέρουν τέτοιους σχεδιασμούς, είναι αναγκαίο να γίνουν ορισμένες παραδοχές.

Αρχικά, γίνεται ο διαχωρισμός των παραγόντων σε παράγοντες στήλης (*column factors*) και παράγοντες γραμμής (*row factors*). Οι παράγοντες στήλης είναι αυτοί των οποίων τα επίπεδα διασταυρώνονται με τις μονάδες που είναι υπό εξέταση, ενώ παράγοντες γραμμής ονομάζονται εκείνοι των οποίων τα επίπεδα είναι εμφωλευμένα στις μονάδες. Παραδείγματος χάριν, σε μία μελέτη που αφορούσε την διαβητική αμφιβληστροειδοπάθεια, ο τύπος του διαβήτη (νεανικός ή ενήλικος) αντιστοιχεί στον παράγοντα γραμμή, ενώ ο τύπος της θεραπείας (laser ή καμία θεραπεία) αποτελεί τον παράγοντα στήλη.

Επιπλέον, είναι χρήσιμη η υιοθέτηση του συμβολισμού $MM(x; y)$ για έναν σχεδιασμό επαναλαμβανόμενων μετρήσεων ο οποίος περιλαμβάνει x παράγοντες γραμμής και y παράγοντες στήλης. Παραδείγματος χάρη, ένας σχεδιασμός επαναλαμβανόμενων μετρήσεων συμβολίζεται ως $MM(0; 1)$.

Στη συνέχεια, παρουσιάζεται η μέθοδος για την περίπτωση του $MM(1; 1)$ σχεδιασμού. Η συγκεκριμένη μέθοδος εφαρμόζεται, επίσης, σε όλους τους σχεδιασμούς επαναλαμβανόμενων μετρήσεων.

Συνεπώς, οι σημαντικές αποκρίσεις εκφράζονται μέσω των τυχαίων και ανεξάρτητων διανυσμάτων $T_{ik} = (T_{i1k}, \dots, T_{ick})'$, $i = 1, \dots, r$. Το κάθε διάνυσμα αντιστοιχεί στις c παρατηρήσεις της k μονάδας όταν η τιμή του παράγοντα γραμμής αντιστοιχεί στο i επίπεδο του. Επιπλέον, θεωρείται η ύπαρξη των ανεξάρτητων και τυχαίων διανυσμάτων $C_{ik} = (C_{i1k}, \dots, C_{ick})'$, $i = 1, \dots, r$ και $k = 1, \dots, n_i$, καθένα από τα οποία παριστάνει τις αποκομμένες μεταβλητές. Για $i = 1, \dots, r$ και $k = 1, \dots, n_i$ το διάνυσμα T_{ik} πρέπει να είναι ανεξάρτητο του C_{ik} .

Εν συνεχεία, οι παρατηρούμενες ποσότητες ορίζονται μέσω των διανυσμάτων:

$$X_{ik} = (X_{i1k}, \dots, X_{ick})' \text{ και } \Delta_{ik} = (\Delta_{i1k}, \dots, \Delta_{ick})' \quad (1)$$

$$i = 1, \dots, r \text{ και } k = 1, \dots, n_i$$

όπου $X_{ijk} = \min(T_{ijk}, C_{ijk})$ και $\Delta_{ijk} = I(X_{ijk} = T_{ijk})$ για $j = 1, \dots, c$. Η συνάρτηση $I(E)$ υποδηλώνει το δείκτη του γεγονότος E . Το οριακό μη παραμετρικό μοντέλο για παραγοντικούς σχεδιασμούς με αποκομμένες παρατηρήσεις καθορίζει μόνο ότι $T_{ijk} \sim F_{ij}$ και $C_{ijk} \sim G_{ij}$, $i = 1, \dots, r$, $j = 1, \dots, c$. Δηλαδή, ότι οι σημαντικές αποκρίσεις T_{ijk} και οι αποκομμένες παρατηρήσεις C_{ijk} ακολουθούν τις συναρτήσεις κατανομών F_{ij} και G_{ij} , αντιστοίχως, για τις οποίες δεν υπάρχει καμία πληροφορία.

Σύμφωνα με τους Akritas και Arnold (1994) οι μη παραμετρικές επιδράσεις ορίζονται μέσω της σχέσης:

$$F_{ij}(t) = \bar{F}_{..}(t) + A_i(t) + B_j(t) + (AB)_{ij}(t)$$

η οποία είναι μοναδική και υπόκειται στους περιορισμούς:

- $\sum_i A_i(t) = 0$
- $\sum_j B_j(t) = 0$
- $\sum_i (AB)_{ij}(t) = 0$ για όλα τα i

- $\sum_j (AB)_{ij}(t) = 0$ για όλα τα j

Συνεπώς, ισχύουν οι σχέσεις:

- $A_i = \bar{F}_{i.} - \bar{F}_{..}$
- $B_j = \bar{F}_{.j} - \bar{F}_{..}$
- $(AB)_{ij} = F_{ij} - \bar{F}_{i.} - \bar{F}_{.j} + \bar{F}_{..}$

όπου $\bar{F}_{i.} = \sum_j F_{ij}/c$, $\bar{F}_{.j} = \sum_i F_{ij}/r$ και $\bar{F}_{..} = \sum_i \sum_j F_{ij}/r c$.

Οι συναρτήσεις A_i , B_j και $(AB)_{ij}$ αποτελούν τις μη παραμετρικές κύριες επιδράσεις των παραγόντων A , B και τη μη παραμετρική επίδραση της αλληλεπίδρασης AB , αντίστοιχα.

Οι μη παραμετρικές υποθέσεις παρουσιάζουν αρκετά πλεονεκτήματα, το κυριότερο από τα οποία είναι ότι παραμένουν αναλλοίωτες στην περίπτωση μετασχηματισμού των δεδομένων. Συγκεκριμένα, οι υποθέσεις αυτές ορίζουν ότι οι αντίστοιχες μη παραμετρικές επιδράσεις είναι μηδέν. Σημειώνεται ότι η μη παραμετρική υπόθεση των μη κύριων επιδράσεων για τον παράγοντα A , $H_0^F(A): A_i = 0$ για κάθε $i = 1, \dots, r$, συνεπάγεται ότι $\bar{F}_{1.} = \bar{F}_{2.} = \dots = \bar{F}_{r.}$. Ομοίως, η μη παραμετρική υπόθεση $H_0^F(A|B): A_i + (AB)_{ij} = 0$ για κάθε $i = 1, \dots, r$ και $j = 1, \dots, c$ είναι ισοδύναμη με τις σχέσεις: $F_{11} = \dots = F_{1c}, F_{21} = \dots = F_{2c}, \dots, F_{r1} = \dots = F_{rc}$.

Η βασική ιδέα για την ανάπτυξη μιας γενικής θεωρίας ελέγχων είναι να γραφούν οι παραπάνω υποθέσεις στη μορφή: $H_0^F: \mathbf{M}\mathbf{F} = \mathbf{0}$, όπου $\mathbf{F} = (F_{11}, \dots, F_{1c}, \dots, F_{r1}, \dots, F_{rc})'$ και \mathbf{M} ένας πίνακας αντιθέσεων. Μια οικογένεια σταθμισμένων στατιστικών Kaplan-Meier για τον έλεγχο τέτοιων υποθέσεων προκύπτει εκτιμώντας το διάνυσμα $T_M = \mathbf{M} \int_0^T \varphi d\mathbf{F}$, όπου φ σταθμισμένη συνάρτηση. Στη συγκεκριμένη μέθοδο επιλέγεται η σταθμισμένη συνάρτηση $\varphi = 1 - H$, όπου H είναι το όριο της εμπειρικής συνάρτησης κατανομής \hat{H} του δείγματος. Συνεπώς, το διάνυσμα T_M εκτιμάται από τη σχέση

$$\hat{T}_M = \mathbf{M} \int_0^T (1 - \hat{H}) d\hat{\mathbf{F}} = \mathbf{M} \int_0^T \hat{S}_H d\hat{\mathbf{F}},$$

όπου $\hat{\mathbf{F}} = (\hat{F}_{11}, \dots, \hat{F}_{1c}, \dots, \hat{F}_{r1}, \dots, \hat{F}_{rc})'$ και \hat{F}_{ij} είναι η εκτιμήτρια Kaplan-Meier για τα δεδομένα του κελιού (i,j) . Συνεπώς, ο στατιστικός έλεγχος είναι σε τετραγωνική μορφή βασισμένος στο \hat{T}_M .

Στα πραγματικά παραδείγματα που παρουσιάζονται στις επόμενες παραγράφους χρησιμοποιείται το στατιστικό έλεγχου που περιλαμβάνει το

παρακάτω θεώρημα. Πριν την παρουσίαση του, όμως, είναι αναγκαίος ο ορισμός μερικών ποσοτήτων. Συνεπώς, έστω $N = \sum_{i=1}^r \sum_{j=1}^c n_{ij} = c \sum_{i=1}^r n_i$ και $\tau_{ij} = \sup\{t: (1 - F_{ij}(t-))(1 - G_{ij}(t-)) > 0\}$, όπου για οποιαδήποτε δεξιά συνεχή συνάρτηση A , με $A(s-)$ συμβολίζεται η δεξιά συνεχή πλευρά της.

Θεώρημα 4.1:

Έστω M ένας πίνακας αντιθέσεων πλήρης διάταξης και έστω \hat{V} ο πίνακας συνδιακύμανσης. Κάτω από την υπόθεση $H_0^F: MF = \mathbf{0}$ και δεδομένου ότι ισχύουν οι συνθήκες:

1. $N \rightarrow \infty$ και $\frac{n_i}{N} \rightarrow c_i > 0$, όπου $\sum c_i = 1$ και

2. $\frac{\int_0^{\tau_{ij}} dF_{ij}(s)}{(1 - G_{ij}(s-))} < \infty$ για κάθε (i, j) .

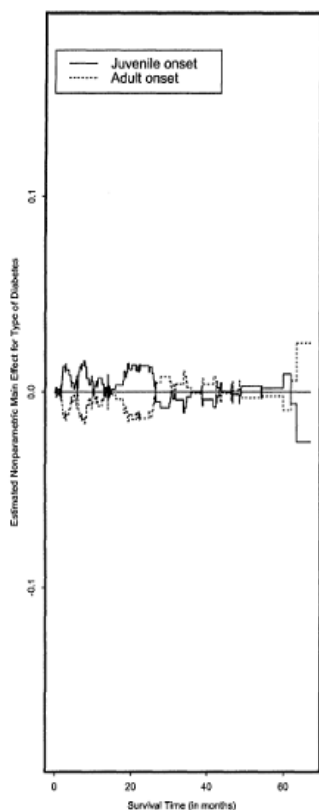
το στατιστικό $Q_N(M) = N(M \int_0^T \hat{S}_H d\hat{F})'(M\hat{V}M')^{-1} \times (M \int_0^T \hat{S}_H d\hat{F})$ ακολουθεί ασυμπτωτικά την χ_v^2 - κατανομή, όπου v ο βαθμός του πίνακα M .

4.4.3.1 Αριθμητικό παράδειγμα

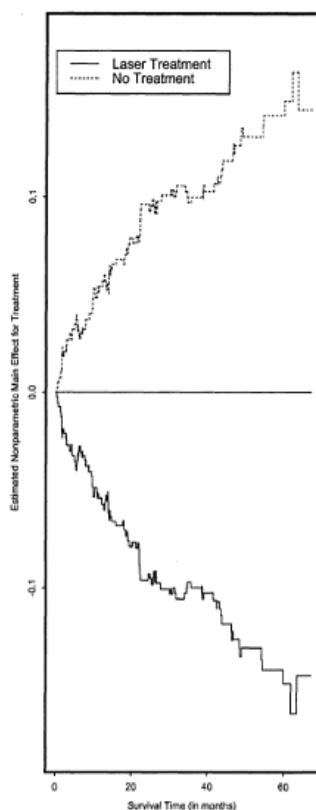
Η μελέτη διαβητικής αμφιβληστροειδοπάθειας (DRS) εξετάζει την αποτελεσματικότητα της φωτοπηξίας με laser στην καθυστέρηση της εκδήλωσης της τύφλωσης σε άτομα που πάσχουν από διαβητική αμφιβληστροειδοπάθεια. Το σύνολο των δεδομένων για πρώτη φορά αναλύθηκε από τους Huster, Brookmeyer και Self το 1989. Για τις ανάγκες της μελέτης το ένα μάτι των ασθενών που συμμετείχαν υποβλήθηκε σε θεραπεία ενώ το άλλο όχι. Η επιλογή του ματιού που υποβλήθηκε σε θεραπεία έγινε τυχαία. Ο σχεδιασμός του πειράματος είναι ένας $MM(1,1)$ στον οποίο ο παράγοντας στήλη (παράγοντας B) αντιστοιχεί στη θεραπεία, ενώ ο παράγοντας γραμμή (παράγοντας A) είναι ο τύπος του διαβήτη (ενήλικος ή νεανικός). Η μεταβλητή απόκρισης είναι ο χρόνος μέχρι η οπτική οξύτητα σε ένα μάτι να είναι λιγότερο από 5/200. Στο πείραμα συμμετείχαν 114 ασθενείς που έπασχαν από νεανικό διαβήτη και 83 ασθενείς που έπασχαν από διαβήτη τύπου II (ενήλικο), ενώ το 61% των παρατηρήσεων είναι αποκομμένες.

Τα γραφήματα 4.7 και 4.8 παρουσιάζουν τις εκτιμώμενες μη παραμετρικές κύριες επιδράσεις για τον τύπο του διαβήτη και της θεραπείας,

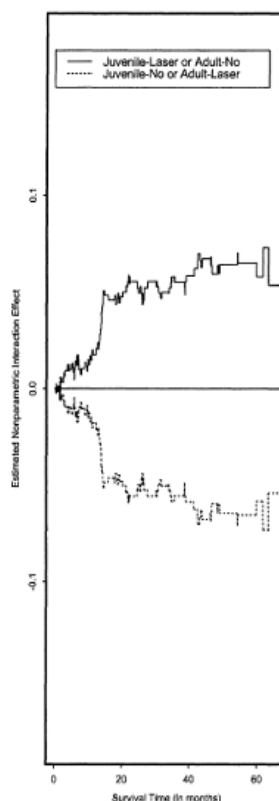
αντίστοιχα. Ενώ στο γράφημα 4.9 παρουσιάζονται οι μη παραμετρικές αλληλεπιδράσεις.



Γράφημα 4.7



Γράφημα 4.8



Γράφημα 4.9

Το γεγονός ότι στο γράφημα 4.8 η εκτιμώμενη μη παραμετρική επίδραση της θεραπείας είναι πάντα μικρότερη από τη μη θεραπεία συνεπάγεται ότι $\bar{F}_{11}(t) > \bar{F}_{21}(t)$, δηλαδή η μέση εκτιμώμενη πιθανότητα η μεταβλητή απόκρισης να είναι μικρότερη από t είναι μεγαλύτερη για το σύνολο που δεν υπόκειται σε θεραπεία για κάθε $t \leq 66.93$. Επίσης, στο γράφημα 4.7 παρατηρείται εκτεταμένη διασταύρωση των εκτιμώμενων μη παραμετρικών επιδράσεων που αφορούν τον τύπο του διαβήτη κάτι που αποδεικνύει τη μη σημαντικότητα της επίδρασης του τύπου του διαβήτη. Τέλος, με βάση το γράφημα 4.9 θα μπορούσε να υποστηριχτεί ότι $(AB)_{11}(t) > (AB)_{12}(t)$, $t \leq 66.93$, βέβαια οι μη παραμετρικές επιδράσεις των αλληλεπιδράσεων είναι λιγότερο σαφή.

Για το σύνολο των δεδομένων του συγκεκριμένου προβλήματος το στατιστικό Q_N για τον παράγοντα A ισούται με $Q_N(A) = 0.002$ με p-τιμή = 0,961, για τον παράγοντα B ισούται με $Q_N(B) = 38,744$ με p-τιμή < 0,001 και για την

αλληλεπίδραση AB ισούται με $Q_N(AB) = 8,970$ με p-τιμή = 0,003. Συμπεραίνεται, λοιπόν, ότι η μη παραμετρική αλληλεπίδραση των παραγόντων, AB, είναι σημαντική. Για την καλύτερη ανάλυση πραγματοποιούνται δύο συμπληρωματικοί έλεγχοι, οι οποίοι εξετάζουν την επίδραση της θεραπείας σε κάθε μία από τις δύο ομάδες ασθενών. Οι έλεγχοι για την επίδραση της θεραπείας με laser στην ομάδα ασθενών που πάσχουν από διαβήτη τύπου I (νεανικός διαβήτης) και διαβήτη τύπου II (ενήλικο διαβήτη) δίνουν p-τιμές 0,0141 και <0,0001, αντίστοιχα. Συνεπώς, η θεραπεία με laser είναι πιο αποτελεσματική για την ομάδα των ασθενών που πάσχουν από διαβήτη τύπου II.

Είναι επίσης δύνατη η ανάλυση του συγκεκριμένου συνόλου δεδομένων χρησιμοποιώντας άλλες μεθόδους. Μία από αυτές είναι η μη παραμετρική μεθοδολογία των Wei και Lachin (1984), η οποία ελέγχει την ισότητα των κατανομών $F_1(t_1, t_2)$ και $F_2(t_1, t_2)$ οι οποίες αντιστοιχούν στις ομάδες που πάσχουν από νεανικό και ενήλικο διαβήτη, αντίστοιχα. Η υπόθεση αυτή συνεπάγεται ισότητα των αντίστοιχων οριακών συναρτήσεων κατανομής το οποίο είναι ισοδύναμο με την υπόθεση $H_0^F(A|B)$. Το στατιστικό έλεγχο των Wei και Lachin για την υπόθεση $H_0^F(A|B)$ υπολογισμένο για το συγκεκριμένο σύνολο δεδομένων ισούται με 8,09 το οποίο αντιστοιχεί στην p-τιμή 0,0175 σύμφωνα με την χ_2^2 κατανομή. Σημειώνεται ότι η μεθοδολογία των Wei και Lachin (1984) δεν μπορεί να χρησιμοποιηθεί για τον έλεγχο της επίδρασης της θεραπείας.

4.4.3.2 Αριθμητικό παράδειγμα

Η εθνική συνεργατική μελέτη Χολολιθίασης (National Cooperative Gallstone) ήταν μια κλινική δοκιμή, σκοπός της οποίας ήταν ο έλεγχος της ασφάλειας και της αποτελεσματικότητας της χρήσης χολικού άλατος chenodiol (χηνοδεσοξυχολικό οξύ) για τη διάλυση χολολίθων. Τα δεδομένα της έρευνας και μία πρώτη ανάλυση τους δημοσιεύθηκαν από τους Wei και Lachin το 1984. Κατά τη διάρκεια του πειράματος μετρήθηκαν τα επίπεδα της χοληστερόλης του ορού των υποκειμένων σε πέντε διαφορετικές χρονικές περιόδους. Η πρώτη μέτρηση έγινε κατά την έναρξη του πειράματος και ακολούθησαν άλλες τέσσερις μετρήσεις, 6, 12, 20 και 24 μήνες αργότερα. Οι συμμετέχοντες άνηκαν σε δύο ομάδες. Στην πρώτη ομάδα ($n_1 = 64$) χορηγήθηκαν εικονικά φάρμακα (placebo), ενώ στη δεύτερη ($n_2 = 48$) χορηγήθηκαν φάρμακα υψηλής δόσης χηνοδεσοξυχολικού οξέος. Λόγω της μη πραγματοποίησης

ορισμένων επισκέψεων, της διακοπής συμμετοχής στη μελέτη καθώς και της απώλειας δειγμάτων υπήρξαν χαμένες παρατηρήσεις. Οι συγκεκριμένες παρατηρήσεις θεωρείται ότι χάθηκαν τυχαία και δεν εξαρτώνται από τη χρονική περίοδο ή από την ομάδα στην οποία άνηκε ο ασθενής.

Όπως και στην εργασία των Wei και Lachin (1984), ως μεταβλητή απόκρισης, χρησιμοποιείται η αύξηση των επιπέδων χοληστερόλης πάνω από τα αρχικά κατά την εξέταση των 6, 12, 20 και 24 μηνών. Για τον καλύτερο χειρισμό των χαμένων δεδομένων, θεωρείται ότι τα επίπεδα χοληστερόλης οριοθετούνται κάτω από 142. Συνεπώς, η αύξηση του επιπέδου της χοληστερόλης μιας χαμένης παρατήρησης θεωρείται αποκομμένη παρατήρηση στο διάστημα κάτω της τιμής 142. Ο παράγοντας στήλη (παράγοντας B) αντιστοιχεί στο χρόνο (6, 12, 20, 24 μήνες) και ο παράγοντας γραμμή (παράγοντας A) αντιστοιχεί στη θεραπεία (παράγοντα A).

Χρησιμοποιώντας τα στατιστικά ελέγχου για μεγάλα δείγματα ισχύουν τα εξής αποτελέσματα: $Q_N(A) = 4,461$ με p -τιμή = 0,035, $Q_N(B) = 18,915$ με p -τιμή = 0,0003 και $Q_N(AB) = 3,773$ με p -τιμή = 0,287.

Το συγκεκριμένο σύνολο δεδομένων αναλύθηκε από τους Wei και Lachin (1984) σαν ένα πρόβλημα πολλών μεταβλητών με έναν παράγοντα. Ο έλεγχος της υπόθεσης $H_0^F(A|B)$, που πραγματοποιήσαν, απέδωσε p -τιμή = 0,121. Συνεπώς, φαίνεται ότι η συγκεκριμένη μέθοδος των Akritas και O' Gorman παρουσιάζει ένα σαφές πλεονέκτημα καθώς δίνεται η δυνατότητα ελέγχου συγκεκριμένων υποθέσεων που αφορούν τις μη κύριες επιδράσεις και αλληλεπιδράσεις.

4.4.3.3 Συμπεράσματα

Στην παραπάνω παράγραφο παρουσιάστηκε το μη παραμετρικό μοντέλο των John T. O' Gorman και Michael G. Akritas για σχεδιασμούς επαναλαμβανόμενων μετρήσεων με μία επανάληψη και αποκομμένα δεδομένα. Το συγκεκριμένο μοντέλο δεν θέτει κανέναν περιορισμό στις συναρτήσεις κατανομών κάθε κελιού, ενώ εφαρμόζεται για όλα τα είδη δεδομένων (συνεχή, διατεταγμένα κ.α.). Τα στατιστικά που παρουσιάζονται είναι εύκολα υπολογίσιμα και αποτελούνται από τετραγωνικές μορφές των στατιστικών Kaplan-Meier. Τέλος, παρουσιάστηκαν κάποια γραφήματα τα οποία εξετάζουν τις μη παραμετρικές επιδράσεις και λειτουργούν ως συμπλήρωμα των τυπικών διαδικασιών ελέγχου.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Aitchison J. and Brown J.A.C. (1957), “*The Log-normal Distribution*”, Cambridge (UK): Cambridge University Press.
- Akritas, M. G., and Arnold, S. F. (1994), "Fully Nonparametric Hypotheses for Factorial Designs I: Multivariate Repeated Measures Designs," *Journal of the American Statistical Association*, **89**, pp.336-343
- Akritas M. G., Arnold S. F. and Brunner E. (1997), "Nonparametric Hypotheses and Rank Statistics for Unbalanced Factorial Designs," *Journal of the American Statistical Association*, **92**, pp.258-265
- Akritas, M. G. and Brunner E. (1997), “Nonparametric Methods for Factorial Designs With Censored Data”, *Journal of the American Statistical Association*, **92**, pp.568-576
- Akritas M. G. and LaValley M. P. (1996), "Nonparametric Inference in Factorial Designs With Censored Data," *Biometrics*, **52**, pp.913-924
- Andersen P. K., Borgan O., Gill R. D., and Keiding N. (1982), "Linear Nonparametric Tests for Comparison of Counting Processes With Applications to Censored Survival Data," *International Statistical Review*, **50**, pp.219-258
- Ανδρουλάκης Ε. (2008) Μεταπτυχιακή διπλωματική εργασία: “Μέθοδοι επιλογής μεταβλητών στο μοντέλο αναλογικής διακινδύνευσης του Cox και εφαρμογές σε πραγματικά ιατρικά δεδομένα με αποκομμένες παρατηρήσεις”
- Αράπη Κ. (2011) Διπλωματική εργασία: “Ανάλυση πειραματικών σχεδιασμών με αποκομμένες παρατηρήσεις”
- Bain L.J., Hardt M.E. (1991), “*Statistical analysis of reliability and life-testing model: theory and methods*”, 2nd edn. Marcel Dekker, New York
- Barton R.R. (1997), “Pre-Experiment Planning for Designed Experiments: Graphical Methods”, *Journal of Quality Technology*, 29, pp.307-316
- Bennett (1983), "Log-Logistic Regression Models for Survival Data", *Journal of the Royal Statistical Society*

- Bewick V., Ball J. and Cheek L., “Survival analysis”, *US National Library of Medicine, National Institutes of Health* (www.ncbi.nlm.nih.gov)
- Bisgaard S. (1992), “Industrial use of statistically designed experiments: case study references and some historical anecdotes”, *Quality Eng.* 4, pp.547–562
- Boag J.W. (1949), “Maximum likelihood estimates of the proportion of patients cured by cancer therapy”, *Journal of the Royal Statistical Society B*, **11**, pp.15–53
- Box G.E.P. and Hunter J.S. (1961), “The 2^{k-p} Fractional Factorial Designs”, *Technometrics*, **3**, pp.311-351
- Chien-Hui Yang and Lee-Ing Tong (2006) “Predicting type II censored data from factorial experiments using modified maximum likelihood predictor”, *Int J Adv Manuf Technol*, **30** pp.887–896
- Collett David (2003), “*Modelling Survival Data in Medical Research*”, 2nd edn., Chapman & Hall/CRC, p.1-54
- Cox D.R. (1958), “*Planning of Experiments*”, Wiley New York
- Crow E.L., Shimizu K, eds. (1988), “*Log-normal Distributions: Theory and Application*”, New York: Dekker
- Crowley J. and Ankerst D.P. (2006), “*Handbook of Statistics in Clinical Oncology*”, 2nd edn., Chapman & Hall/CRC
- C.F. Jeff Wu and Michael Hamada (2000), “Experiments: Planning, Analysis, and Parameter Design Optimization”, *Wiley Series in Probability and Statistics*, pp.1-255
- Δεσπούδης Θ. και Φυτίλης Χ. (2009), Έρευνα και εφαρμογή σε βιογραφικά δεδομένα μουσικών “Ανάλυση Επιβίωσης”
- Dey A. and Mukerjee R. (1999), “*Fractional Factorial Plans*”, John Wiley & Sons
- Durie B.G.M., Jacobson J., Barlogie B., and Crowley J.(2004), “Magnitude of response with myeloma frontline therapy does not predict outcome: importance of time to progression in Southwest Oncology Group chemotherapy trials”, *J. Clin. Onc.*, **22**, pp.1857–1863
- Finney D.J. (1945), “*The fractional replication of the factorial arrangements*”, *Ann.Eugen.* pp. 291-301

- Hinkelmann K. and Kempthorne O. (1994), “*Design and Analysis of Experiments*”, New York: John Wiley & Sons
- Hirano S.S., Nordheim E.V., Arny D.C. and Upper C.D. (1982), “Log-normal distribution of epiphytic bacterial populations on leaf surfaces”, *Applied and Environmental Microbiology*, **44**, pp.695-700
- Huster W. J., Brookmeyer R. and Self S. G. (1989), “Modelling paired survival data with covariates”, *Biometrics*, **45**, pp.145-156.
- Kempthorne O. (1952), “*The Design and Analysis of Experiments*”, Wiley, New York.
- Klein J.P. and Moeschberger M.L. (1997), “*Survival Analysis: Techniques for Censored and Truncated Data*”, Springer, New York
- Kaplan E.L. and Meier P. (1958), “Nonparametric estimation from incomplete observations”, *J.Amer. Statist. Assoc.*, pp.457-481
- Καρώνη X. (2009), “Μοντέλα Αξιοπιστίας και Επιβίωσης”, Συμμεών
- Κουκουβίνος X. (2005), “Γραμμικά μοντέλα και σχεδιασμοί”
- Κουκουβίνος X. (2005), “Στατιστικοί σχεδιασμοί”, pp.1-19
- Lawless JF (1982), “*Statistical models and methods for lifetime data*”, Wiley, New York
- Lee-Ing Tong and Chao-Ton Su (1997), “A non-parametric method for experimental analysis with censored data”, *International Journal of Quality & Reliability Management*, **14**, pp. 456-463
- Limpert E., Stahel W. and Abbt M. (2001) “Log-normal distributions across the Sciences: Keys and Clues”, *BioScience*, **51**, pp.341-352
- Mantel N., and Haenszel W. (1959), "Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease," *Journal of the National Cancer Institute*, **22**, pp.719-748
- Miller RG (1981), “*Survival analysis*”, Wiley, New York
- Montgomery DC (2001), “*Design and analysis of experiments*”, 5th edn. Wiley, New York
- Nelson W.B. (1972), “Theory and applications of hazard plotting for censored failure data”, *Technometrics*, pp.945-966

- O'Gorman J.T. and Akritas M.G. (2001), “Nonparametric Models and Methods for Designs with Dependent Censored Data: Part I”, *Biometrics*, **57**, pp. 88-95
- Prvan T. and Street D.J. (2002), “An annotated bibliography of application papers using certain classes of fractional factorial and related designs”, *Journal of Statistical Planning and Inference*, **106**, pp. 245 – 269
- Raqab M.Z. (1997) “Modified maximum likelihood predictors of future order statistics from normal samples”, *Comput Stat Data Anal* , **25**, pp.91–106
- Shoukri M.M., Mian I.U.M. and Tracy D.S. (1988), "Sampling Properties of Estimators of the Log-Logistic Distribution with Application to Canadian Precipitation Data", *The Canadian Journal of Statistics*, **16**, pp.223-236
- Su CT and Miao CL (1998) “Neural network procedures for experimental analysis with censored data”, *Int J Qual Sci* ,**3**, pp.239–253
- Taguchi, G. (1987), “*Introduction to Quality Engineering, 5-day Seminar Course Manual*”, American Supplier Institute, Inc.
- Wei L. J. and Lachin, J. M. (1984), “Two-sample asymptotically distribution-free tests for incomplete multivariate observations”, *Journal of the American Statistical Association*, **79**, pp.653-661.
- Wikipedia the free encyclopedia, “the log-logistic distribution”, www.wikipedia.com