ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΑΝΑΛΥΣΗ ΠΕΙΡΑΜΑΤΙΚΩΝ ΣΧΕΔΙΑΣΜΩΝ ΜΕ ΑΠΟΚΟΜΜΕΝΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ

Διπλωματική Εργασία

ΑΡΑΠΗ ΚΛΕΟΠΑΤΡΑ

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

Αθήνα, Οκτώβριος 2011

ΠΕΡΙΛΗΨΗ

Με την εξέλιξη της κοινωνίας και της βιομηχανίας κατά τη διάρκεια των τελευταίων δεκαετιών, οι μαθηματικοί κλάδοι της περιγραφικής στατιστικής, καθώς και της στατιστικής συμπερασματολογίας αποτέλεσαν εξαιρετικά σημαντικές περιοχές έρευνας. Οι μέθοδοι της περιγραφικής στατιστικής αποτελούν το επιστημονικό εργαλείο για τη συγκέντρωση, οργάνωση και παρουσίαση πειραματικών δεδομένων σε εύληπτη μορφή, ενώ οι μέθοδοι της στατιστικής συμπερασματολογίας, καθιστούν δυνατή την προσέγγιση τόσο ποσοτικών όσο και ποιοτικών χαρακτηριστικών ενός ευρύτερου συνόλου, όπως αυτά απορρέουν από τη μελέτη ενός σχετικά μικρού υποσυνόλου του.

Στα πλαίσια της εργασίας αυτής, δίνεται ιδιαίτερο βάρος στη στατιστική μελέτη προβλημάτων της ανάλυσης αξιοπιστίας ή επιβίωσης, τα πειραματικά ή ερευνητικά δεδομένα της οποίας μπορεί να αφορούν τη χρονική στιγμή που προκαλείται βλάβη σε ένα μηχανικό σύστημα ή θάνατος σε ένα βιολογικό οργανισμό. Οι μέθοδοι της στατιστικής συμπερασματολογίας σε τέτοιου είδους εφαρμογές, στοχεύουν στον προσδιορισμό των παραγόντων που επιδρούν σημαντικά στο υπό μελέτη γεγονός, καθώς και στον προσδιορισμό του βέλτιστου δυνατού συνδυασμού συνθηκών, ώστε να επιτυγχάνεται το επιθυμητό αποτέλεσμα σε κάθε περίπτωση. Το αποτέλεσμα αυτό μπορεί να είναι η επίτευξη της βέλτιστης ποιότητας ενός παραγόμενου βιομηχανικού προϊόντος, όπως επίσης και η ποσοστιαία αύξηση της αναμενόμενης διάρκειας ζωής ενός βιολογικού οργανισμού.

Στο πρώτο κεφάλαιο της παρούσας εργασίας 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 περιγράφεται το γραμμικό μοντέλο και οι μέθοδοι εκτίμησης των συντελεστών αυτού και στη συνέχεια, αναλύεται το μοντέλο αναλογικής διακινδύνευσης του Cox, ενώ παράλληλα γίνεται μια εισαγωγή στις μεθόδους ελέγχου καταλληλότητας μοντέλων και στους ελέγχους υποθέσεων. Στο τρίτο κεφάλαιο παρατίθενται τα βασικά είδη παραγοντικών σχεδιασμών που χρησιμεύουν στην οργάνωση των πειραματικών δεδομένων και τέλος, στο τέταρτο κεφάλαιο περιγράφονται αναλυτικά μέθοδοι ανάλυσης δεδομένων που εμπεριέχουν εκτός από πλήρεις και αποκομμένες παρατηρήσεις, ενώ δίνονται σχετικά παραδείγματα.

ABSTRACT

With the evolution of society and industry over the past decades, the mathematical fields of descriptive statistics and statistical inference have proved to be extremely important research areas. The methods of descriptive statistics form a scientific tool for collecting, organizing and presenting experimental data in a quite easily understandable form and the methods of statistical inference make it possible to approach both quantitative and qualitative characteristics of a larger whole, by studying those characteristics arising from a relatively small sample.

As long as this work is related, focus is given in particular on the statistical study of reliability problems and survival analysis. In those cases, experimental or research data may concern the time when failure is observed in a mechanical system or death in a biological organism. The methods of statistical inference, in such applications, aim to identify the factors that impact significantly on the studied event, and also to identify the optimum combination of conditions in order to achieve the desired result in each case. This result might be to achieve the highest quality of an industrial product, as well as to increase the life expectancy of a biological organism.

In the first chapter of this work, an introductory reference is made to concepts related to reliability analysis. The different types of experimental data are also discussed, as well as the distributions that are often adjusted to it. In the second chapter, the linear model is described and also the methods of estimation of the model coefficients. Furthermore, the proportional hazard model of Cox is analyzed, while an introduction is made on the methods of validation of chosen models and hypothesis testing. In the third chapter, the main types of factorial designs used in organizing experimental data are presented and finally, in the fourth chapter methods of data analysis are thoroughly described, involving other than full and censored observations, while illustrative examples are given.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τον Καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου κ. Χρήστο Κουκουβίνο, για τη συνεχή ενθάρρυνση, καθοδήγηση και εμπιστοσύνη που έδειξε καθ' όλη τη διάρκεια εκπόνησης αυτής της διπλωματικής. Θεωρώ επίσης υποχρέωσή μου να ευχαριστήσω τον υποψήφιο διδάκτορα Ανδρουλάκη Εμμανουήλ για την πολύτιμη βοήθεια, στήριξη και το συνεχές ενδιαφέρον που έδειξε. Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου για τα εφόδια που μου προσέφεραν, τη φροντίδα, τη συμπαράσταση και την υπομονή τους.

.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

ΚΕΦΑΛΑΙΟ 2

ΤΟ ΗΜΙ-ΠΑΡΑΜΕΤΡΙΚΟ ΜΟΝΤΕΛΟ ΑΝΑΛΟΓΙΚΗΣ ΔΙΑΚΙΝΔΥΝΕΥΣΗΣ ΤΟΥ COX

ΚΕΦΑΛΑΙΟ 3

ΠΕΙΡΑΜΑΤΙΚΟΙ ΣΧΕΔΙΑΣΜΟΙ

ΚΕΦΑΛΑΙΟ 4

ΜΕΘΟΔΟΙ ΑΝΑΛΥΣΗΣ ΣΧΕΔΙΑΣΜΩΝ ΜΕ ΑΠΟΚΟΜΜΕΝΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Εισαγωγικές έννοιες

Σε διάφορους τομείς της επιστήμης πραγματοποιούνται πειράματα και έρευνες που έχουν σαν στόχο είτε τη βελτίωση των ανθρώπινων συνθηκών ζωής, είτε την αναβάθμιση της παραγωγικής διαδικασίας. Σε τομείς όπως η βιομηχανία, ένα πείραμα έχει συνήθως σαν στόχο την εξαγωγή συμπερασμάτων με γνώμονα τη βελτίωση της ποιότητας του παραγόμενου προϊόντος, ενώ στον τομέα της υγείας μια έρευνα συχνά αποσκοπεί στη βαθύτερη κατανόηση των λόγων που οδηγούν στην εκδήλωση ασθενειών και κατά προέκταση στην αμεσότερη διάγνωση και αντιμετώπισή τους.

Ο κλάδος της στατιστικής που εξετάζει σε ποιά χρονική στιγμή αναμένεται να προκληθεί θάνατος σε κάποιο βιολογικό οργανισμό, καθώς και την πιθανότητα αποτυχίας σε μηχανικά συστήματα ή μέρη αυτών, ονομάζεται ανάλυση αξιοπιστίας ή επιβίωσης (*reliability or survival analysis*). Κάνοντας χρήση των μεθόδων της ανάλυσης αξιοπιστίας, ένας ερευνητής δύναται να αποφανθεί για τους παράγοντες που επιδρούν σημαντικά στην εμφάνιση του υπό μελέτη γεγονότος, όπως επίσης και για το ποσοστό επίδρασής τους σε αυτό.

Στην πλειονότητα των περιπτώσεων, τα πειραματικά δεδομένα της ανάλυσης επιβίωσης αφορούν, το λεγόμενο χρόνο αποτυχίας *(failure time)*, το χρόνο δηλαδή κατά τον οποίο προκαλείται θάνατος ή βλάβη στο υπό μελέτη άτομο ή αντικείμενο αντίστοιχα. Στις περιπτώσεις όπου ο χρόνος αποτυχίας δεν είναι δυνατό να παρατηρηθεί πειραματικά, (για παράδειγμα λόγω βίαιης διακοπής του πειράματος), τότε η παρατήρηση θεωρείται αποκομμένη *(censored)*. Η έννοια των αποκομμένων δεδομένων χρησιμοποιήθηκε για πρώτη φορά από τον Hald (1949). Τα δεδομένα που δεν είναι αποκομμένα ονομάζονται πλήρη.

Στα στατιστικά δεδομένα της ανάλυσης επιβίωσης είναι αρκετά συχνό το φαινόμενο των αποκομμένων παρατηρήσεων, οι οποίες στην πλειονότητα των περιπτώσεων είναι αποτέλεσμα ατελούς παρατήρησης της εξέλιξης κάποιων ασθενών ή ακόμα μπορούν να προκύψουν από κάποια βίαιη διακοπή του πειράματος σε ανύποπτη χρονική στιγμή. Ένα βασικό χαρακτηριστικό τους είναι ότι λόγω της ασυμμετρίας που παρουσιάζουν, δεν ακολουθούν κανονική κατανομή και συνεπώς δεν επιτρέπουν τη χρήση των συνηθισμένων στατιστικών τεχνικών για την εξαγωγή συμπερασμάτων.

Σε κάποιες περιπτώσεις, όπως θα δούμε και στη συνέχεια, είναι δυνατή η μετατροπή των δεδομένων ώστε να προσαρμόζονται στο μη συμμετρικό μοντέλο της λογαριθμο-κανονικής κατανομής (Lee and Wang, 2003). Στην πλειονότητα των περιπτώσεων, όμως, προτιμάται η υιοθέτηση άλλων μη συμμετρικών μοντέλων, ούτως ώστε να μην κρίνεται απαραίτητη η μετατροπή των αρχικών δεδομένων.
Συγκεκριμένα, οι κατανομές που συχνότερα προσαρμόζονται σε τέτοιου είδους εφαρμογές είναι η κατανομή Weibull και Γάμμα (Καρώνη, 2005), λόγω κυρίως της ευελιξίας που παρουσιάζουν ως μοντέλα αξιοπιστίας.

Ένα παράδειγμα εφαρμογής κατά τη διάρκεια μελέτης της οποίας είναι πιθανό να προκύψουν αποκομμένες παρατηρήσεις, αποτελεί η διαδικασία ανεύρεσης του αναμενόμενου χρόνου ζωής T μιας μπαταρίας αυτοκινήτου κάτω από στρεσογόνες συνθήκες, καθώς και του χρόνου ζωής *(survival time)* T ενός ασθενούς που πάσχει από καρκίνο του παγκρέατος.

1.2 Είδη αποκομμένων δεδομένων

Τα είδη αποκομμένων δεδομένων (Collett, 2003) είναι τρία:

- **Τα από δεξιά αποκομμένα δεδομένα** *(right-censored data)*

Σε αυτή την κατηγορία ανήκουν οι παρατηρήσεις των οποίων η πραγματική τιμή είναι άγνωστη, είναι όμως γνωστό ότι είναι μεγαλύτερη ή ίση με κάποια δεδομένη τιμή L. ($T \geq L$)

- **Τα από αριστερά αποκομμένα δεδομένα** *(left-censored data)*

Σε αυτή την κατηγορία ανήκουν οι παρατηρήσεις των οποίων η πραγματική τιμή είναι άγνωστη, είναι όμως γνωστό ότι είναι μικρότερη ή ίση με κάποια δεδομένη τιμή L. ($T \leq L$)

- **Τα δεδομένα αποκομμένα σε διάστημα** *(doubly censored)*

Η τελευταία κατηγορία αφορά ουσιαστικά τα δεδομένα που είναι ταυτόχρονα αποκομμένα από δεξιά και από αριστερά. Αυτό σημαίνει ότι και στην περίπτωση αυτή η πραγματική τιμή είναι άγνωστη, είναι όμως γνωστό ότι βρίσκεται εντός κάποιου γνωστού διαστήματος τιμών ($m \leq T \leq M$).

Το είδος αποκομμένων παρατηρήσεων που συναντάται συχνότερα σε στατιστικά δεδομένα είναι οι από δεξιά αποκομμένες παρατηρήσεις.

1.3 Είδη αποκοπής δεδομένων

Στη συνέχεια δίνονται οι τρεις βασικοί μηχανισμοί με τους οποίους καταλήγουμε σε τέτοιου είδους δεδομένα.

- **Η αποκοπή τύπου I** *(type I censoring)*

Στην αποκοπή τύπου I η διάρκεια της μελέτης του δείγματος είναι καθορισμένη εκ των προτέρων και ίση με κάποιο $t > 0$. Αυτό πρακτικά σημαίνει ότι αν η μεταβλητή T παίρνει τιμή μικρότερη ή ίση με τη διάρκεια της έρευνας, τότε η τιμή της T θεωρείται γνωστή. Αν, όμως, ισχύει $T > t$ τότε η μοναδική πληροφορία που έχουμε για τη μεταβλητή T είναι ότι δέχεται τιμή μεγαλύτερη από τη γνωστή t.

<u>Παράδειγμα 1.1</u>

Ας θεωρήσουμε ότι 100 μπαταρίες μελετώνται ως προς τη διάρκεια ζωής τους για διάστημα επτά μηνών ($t = 7$). Τότε οι μπαταρίες που μετά το πέρας των επτά μηνών θα εξακολουθούν να είναι σε λειτουργία αποτελούν αποκομμένες παρατηρήσεις ($T > t=7$). Σε αυτή την περίπτωση το πλήθος των ολοκληρωμένων παρατηρήσεων στη διάρκεια διεξαγωγής του πειράματος είναι τυχαίο.

- **Η αποκοπή τύπου II** *(type II censoring)*

Στην αποκοπή τύπου II η διάρκεια της μελέτης του δείγματος δεν είναι ορισμένη εξ'αρχής. Έτσι, η διαδικασία ολοκληρώνεται όταν το γεγονός που μελετάται παρατηρηθεί σε έναν προκαθορισμένο αριθμό k το πλήθος ατόμων, με $k \in N$.

Παράδειγμα 1.2

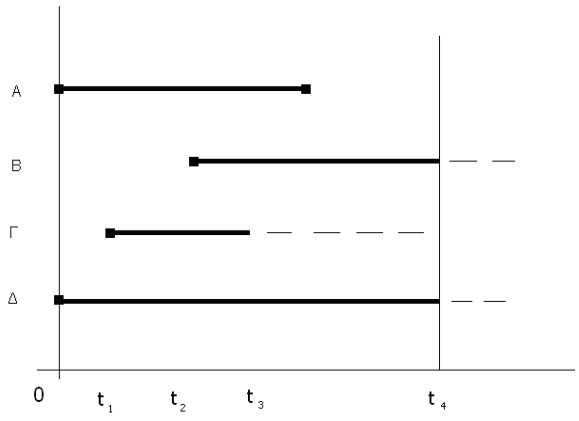
Στην περίπτωση του πειράματος των 100 μπαταριών το πείραμα θα ολοκληρωθεί όταν συμπληρωθεί ένας συγκεκριμένος αριθμός παρατηρήσεων. Για παράδειγμα, έστω ότι ορίζουμε το πείραμα να ολοκληρώνεται μόλις σταματήσει να λειτουργεί η κατά σειρά 80ή μπαταρία. Έτσι, η διάρκεια του πειράματος είναι άγνωστη, εξασφαλίζεται όμως η εξασφάλιση επαρκούς πληροφορίας για την εξαγωγή αξιόπιστων στατιστικών συμπερασμάτων. Οι τελευταίες 20 παρατηρήσεις θεωρούνται αποκομμένες.

- **Η αποκοπή τύπου ΙΙΙ** *(type III / random censoring)*

Στην αποκοπή τύπου ΙΙΙ εμφανίζεται κατά κόρον στις μελέτες ιατρικής φύσεως. Το κύριο χαρακτηριστικό αυτού του μηχανισμού αποκοπής είναι ότι η διάρκεια της μελέτης είναι προκαθορισμένη, ενώ οι ασθενείς δεν εισέρχονται την ίδια χρονική στιγμή σε αυτήν.

Παράδειγμα 1.3

Έστω ότι η Ιατρική Σχολή του πανεπιστημίου Ιωαννίνων με πρωτοβουλία ομάδας καθηγητών αποφασίζει να μελετήσει για πέντε χρόνια την επίδραση του καπνίσματος στη διάρκεια ζωής ασθενών που πάσχουν από χρόνια αποφρακτική πνευμονοπάθεια. Ένας σεβαστός αριθμός ατόμων που πάσχουν από την ασθένεια αρκεί για να τεθεί η έρευνα σε λειτουργία. Αν όμως κατά τη διάρκεια διεξαγωγής της μελέτης αυτής προκύψουν και άλλοι πάσχοντες που ενδιαφέρονται να συμμετέχουν τότε είναι πιθανό να προστεθούν στο δείγμα παρά το γεγονός ότι η έρευνα είναι ήδη σε εξέλιξη. Αυτό έχει σαν αποτέλεσμα τη δημιουργία παρατηρήσεων των οποίων η μορφή φαίνεται παρακάτω:



Γράφημα 1.1

- Στην κατηγορία παρατηρήσεων Α ανήκει το σύνολο των ατόμων που εισήλθαν στην έρευνα τη χρονική στιγμή $t_0 = 0$ και το προς μελέτη γεγονός προκλήθηκε όσο ακόμα η μελέτη ήταν σε εξέλιξη.
- Στην κατηγορία Β ανήκουν άτομα που εισήλθαν στο πείραμα τη χρονική στιγμή $t_2 > t_0$ και μέχρι και τη χρονική στιγμή t_4 όπου έληξε το πείραμα δεν είχε προκληθεί το υπό μελέτη γεγονός.
- Στην κατηγορία Γ ανήκουν άτομα που εισήλθαν στο πείραμα τη χρονική στιγμή $t_1 > t_0$ και τη χρονική στιγμή t_3 αποχώρησαν από το πείραμα χωρίς ακόμα να έχει προκληθεί το υπό μελέτη γεγονός.
- Στην κατηγορία Δ ανήκουν τα άτομα που εισήλθαν στο πείραμα στην αρχή του και μέχρι τη λήξη του δεν είχε ακόμα παρατηρηθεί το υπό μελέτη γεγονός.

Ενδεικτικά αναφέραμε κάποιες περιπτώσεις. Είναι προφανές ότι οι κατηγορίες παρατηρήσεων Β, Γ, Δ όπως ορίστηκαν αποτελούν από δεξιά αποκομμένα δεδομένα.

1.4 Προϋποθέσεις για τη δυνατότητα ανάλυσης των χρόνων επιβίωσης

Γενικά για τη μελέτη δεδομένων επιβίωσης στα οποία περιλαμβάνονται και αποκομμένες παρατηρήσεις, εκτός από την προϋπόθεση της ανεξαρτησίας των παρατηρήσεων μεταξύ τους, αποτελεί εξίσου βασική προϋπόθεση η ανεξαρτησία των χρόνων αποκοπής από τον χρόνο επιβίωσης. Έτσι, στην κατηγορία των παρατηρήσεων Β τα αποκομμένα δεδομένα προέκυψαν λόγω διαφορετικής χρονικής στιγμής εισόδου στη μελέτη κι επομένως ισχύει η υπόθεση της ανεξαρτησίας. Στην κατηγορία των περιπτώσεων Γ όμως, η βίαιη απομάκρυνση από το πεδίο μελέτης μπορεί να μεταφράζεται σε μερική ή ολική εξάρτηση της αποκοπής, από το χρόνο επιβίωσης.

Για παράδειγμα αν ένας ασθενής αρνείται να συμμετάσχει πλέον σε ένα πρόγραμμα παρακολούθησης της υγείας του, λόγω ξαφνικής αδυναμίας που οφείλεται σε επιδείνωση της κατάστασής του, τότε ο χρόνος αποκοπής έχει άμεση εξάρτηση από τον τελικό χρόνο επιβίωσης.

Ας δούμε τώρα ένα αριθμητικό παράδειγμα διάρκειας ζωής μπαταριών λιθίου (Canadian Electronics Magazine, 2009) ελαφρά τροποποιημένο ώστε να καθιστά σαφή την έννοια των αποκομμένων δεδομένων. Στο παράδειγμα αυτό, όπως και σε πολυάριθμες πειραματικές εφαρμογές εμπεριέχονται εκτός από πλήρεις και αποκομμένες παρατηρήσεις.

Παράδειγμα 1.4

Σε ένα δείγμα 20 μπαταριών ιόντων λιθίου τις οποίες φορτίζουμε καθημερινά την ίδια ώρα. Μελετούμε μετά από πόσες μέρες η μπαταρία θα φτάσει να αποδίδει λιγότερο από το 1% της αρχικής της απόδοσης. Το πείραμα έχει οριστεί από την αρχή ότι θα διαρκέσει ακριβώς 1 χρόνο (365 ημέρες). Τα αποτελέσματα φαίνονται στον πίνακα που ακολουθεί:

ΠΙΝΑΚΑΣ 1.1

ΠΕΙΡΑΜΑ ΔΙΑΡΚΕΙΑΣ ΖΩΗΣ ΜΠΑΤΑΡΙΩΝ ΛΙΘΙΟΥ			
Α/Α	**Διάρκεια ζωής**	**Α/Α**	**Διάρκεια ζωής**
1.	307	11.	354
2.	298	12.	339
3.	349	13.	288
4.	301	14.	361
5.	365*	15.	365*
6.	365*	16.	331
7.	268	17.	357
8.	328	18.	282
9.	340	19.	345
10.	365*	20.	322

Στις 16 από τις 20 παρατηρήσεις είχαμε πλήρη χρόνο διάρκειας ζωής. Στις παρατηρήσεις όμως με αστερίσκο, δηλαδή σε εκείνες που έχουν αύξοντα αριθμό 5,6,10 και 15 δεν είχε παρατηρηθεί μείωση της απόδοσης των μπαταριών κάτω από 1% μέχρι και τη χρονική στιγμή λήξης του πειράματος. Αυτό σημαίνει ότι οι τέσσερις αυτές παρατηρήσεις αποτελούν από δεξιά αποκομμένα δεδομένα.

Έχει ενδιαφέρον απλώς να αναφέρουμε ότι οι μπαταρίες ιόντων λιθίου χρησιμοποιούνται σήμερα σε ένα τεράστιο εύρος συσκευών, από τα κινητά τηλέφωνα μέχρι τα υβριδικά οχήματα, αφού μπορούν να αποθηκεύουν μεγάλα ποσά ενέργειας ανά μονάδα βάρους. Έτσι παρατείνεται η διάρκεια ζωής της μπαταρίας, ενώ είναι πιο ελαφριά αφού το λίθιο είναι το πιο ελαφρύ μέταλλο.

1.5 Κατανομές

1.5.1 Βασικά στοιχεία

Για τη μελέτη των κυριότερων κατανομών που προσαρμόζονται στα δεδομένα επιβίωσης είναι απαραίτητος ο ορισμός κάποιων βασικών εννοιών:

1.5.1.1 Συνάρτηση Κατανομής

Η συνάρτηση κατανομής F(t) *(distribution function)* της μεταβλητής T προκύπτει ως εξής:

$$F(t) = P(T \leq t),$$

δηλαδή η συνάρτηση κατανομής μας δίνει την πιθανότητα η τιμή της μεταβλητής που μελετάται να είναι μικρότερη ή ίση από κάποια δοσμένη τιμή t. Επομένως η F(t) είναι αύξουσα συνάρτηση, με $\lim_{t\to 0} F(t) = 0$ και $\lim_{t\to\infty} F(t) = 1$.

1.5.1.2 Συνάρτηση αξιοπιστίας ή συνάρτηση επιβίωσης *(survival function)*

Η συνάρτηση αξιοπιστίας ή αλλιώς συνάρτηση επιβίωσης ορίζεται ως εξής:

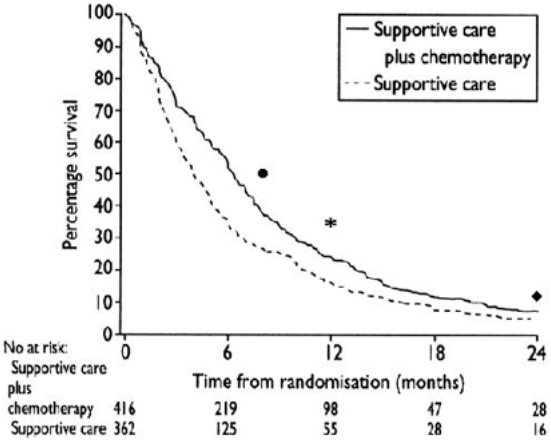
$$S(t) = P(T>t) = 1-F(t)$$

και αναφέρεται στην πιθανότητα ένα άτομο να επιβιώσει για χρόνο μεγαλύτερο από χρόνο t (Lee-Wang,2003).

Έπεται, λοιπόν, άμεσα ότι πρόκειται για μια φθίνουσα συνάρτηση για την οποία ισχύει ότι για τη χρονική στιγμή μηδέν (t=0), η πιθανότητα επιβίωσης είναι ίση με τη μονάδα (S(t)=1), ενώ για άπειρο χρόνο ($t \to \infty$) η πιθανότητα επιβίωσης είναι μηδενική (S(t)=0).

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

Στο παρακάτω σχήμα βλέπουμε ένα παράδειγμα καμπυλών επιβίωσης μετά από έρευνα σε άτομα που πάσχουν από καρκίνο του παγκρέατος (British Medical Journal, 1995):



Γράφημα 1.2

1.5.1.3 Συνάρτηση πυκνότητας πιθανότητας

Η συνάρτηση πυκνότητας πιθανότητας f(t) *(probability density function,)* η οποία συναντάται στη βιβλιογραφία ως σ.π.π της μεταβλητής T, προκύπτει από τη σχέση:

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

Σημειώνεται ότι η συνάρτηση Κατανομής συνδέεται με τη συνάρτηση πυκνότητας πιθανότητας με τη σχέση:

$$F(t) = P(T \le t) = \int_0^t f(u)du$$

Ενώ η συνάρτηση αξιοπιστίας συνδέεται με τη συνάρτηση πυκνότητας πιθανότητας με τη σχέση:

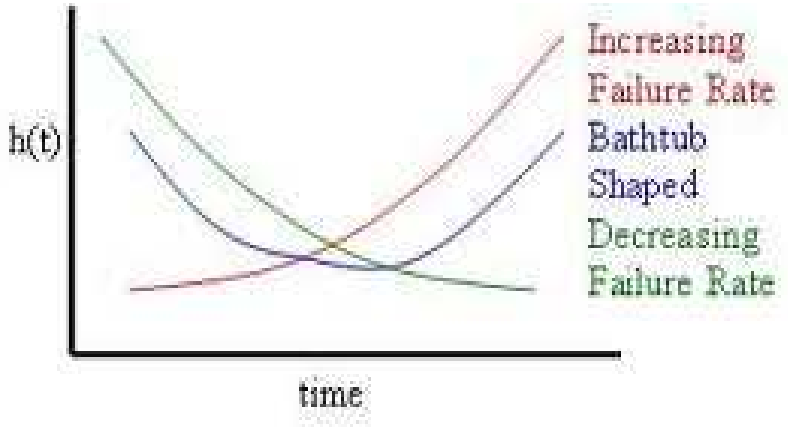
$$S(t) = P(T > t) = \int_{t}^{\infty} f(u)du$$

1.5.1.4 Συνάρτηση διακινδύνευσης

Η συνάρτηση διακινδύνευσης ή συνάρτηση βαθμού κινδύνου *(hazard function)*, συμβολίζεται με h(t) και εκφράζει την τάση του προς μελέτη ατόμου ή αντικειμένου να αποτύχει στο χρονικό διάστημα (t,t+δt] . Ορίζεται ως:

$$h(t) = \lim_{\delta t \to 0} \left[\frac{[S(t) - S(t+\delta t)] / S(t)}{\delta t} \right] = \frac{f(t)}{S(t)}$$

Στο γράφημα που ακολουθεί φαίνεται με κόκκινο χρώμα η συνάρτηση αύξουσας διακινδύνευσης, με πράσινο η συνάρτηση φθίνουσας διακινδύνευσης και με μπλε η "διακινδύνευση μπανιέρας" *(bath-tub hazard)* που αποτελεί το πιο ρεαλιστικό μοντέλο διακινδύνευσης, αφού ξεκινά με πτωτική τάση, συνεχίζει με μια παροδική σταθεροποίηση της διακινδύνευσης και με το πέρασμα του χρόνου αποκτά έντονα αυξητική τάση.



Γράφημα 1.3

Όταν η συνάρτηση διακινδύνευσης είναι φθίνουσα τότε η στιγμιαία πιθανότητα θανάτου μειώνεται με το πέρασμα του χρόνου, κάτι που γενικά δεν αναμένεται να συμβεί. Αντίθετα, όταν είναι αύξουσα τότε η στιγμιαία πιθανότητα θανάτου αυξάνεται με την πάροδο του χρόνου που είναι και το αναμενόμενο.

1.5.1.5 Σωρευτική συνάρτηση διακινδύνευσης

Η σωρευτική συνάρτηση διακινδύνευσης *(cumulative hazard function)*, συμβολίζεται με H(t) και ορίζεται ως :

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

Αποδεικνύεται ιδιαίτερα χρήσιμη για την επιλογή του καταλληλότερου μοντέλου επιβίωσης.

Από όλα τα παραπάνω προκύπτει άμεσα ότι οι συναρτήσεις S(t), h(t), f(t), F(t) είναι ισοδύναμες, αφού:

$$H(t) = \int_0^t h(u)du = \int_0^t \frac{f(u)}{S(u)}du = \int_0^t \frac{-S'(u)}{S(u)}du = \left[-\ln(S(u))\right]_0^t = -\ln S(t)$$

,

και επομένως:

$$S(t) = \exp\{-H(t)\}$$

.

Με άλλα λόγια, αυτό σημαίνει πως η γνώση μιας από τις παραπάνω συναρτήσεις αρκεί για τον υπολογισμό των υπολοίπων.

1.5.2 Εκθετική κατανομή

Η εκθετική κατανομή *(exponential)* συχνά ασχολείται με το χρονικό διάστημα μέχρι κάποιο συγκεκριμένο συμβάν. Για παράδειγμα, το μήκος που αποκτά, μέσα σε λίγα λεπτά, η γραμμή υπεραστικών τηλεφωνικών κλήσεων μιας επιχείρησης, καθώς και το χρονικό διάστημα, σε μήνες, που διαρκεί μια μπαταρία αυτοκινήτου. Μπορεί να αποδειχθεί, επίσης, ότι το ποσό των χρημάτων που έχουμε στην τσέπη μας ακολουθεί μια εκθετική κατανομή.

Η κατανομή αυτή αποτελεί το απλούστερο, ίσως, μοντέλο διάρκειας ζωής. Δεν τυγχάνει όμως μεγάλης χρηστικής αξίας στον τομέα της αξιοπιστίας, αφού σε

ελάχιστες περιπτώσεις κρίνεται ως το καταλληλότερο μοντέλο σε εφαρμογές που εμπεριέχουν αποκομμένες παρατηρήσεις.

Συνάρτηση πυκνότητας πιθανότητας

Η συνάρτηση πυκνότητας πιθανότητας μιας εκθετικής κατανομής με παράμετρο ρυθμού λ είναι:

$$f(t;\lambda) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0, \\ 0, & t < 0. \end{cases} \qquad \lambda>0$$

Μέση τιμή, διασπορά και διάμεσος

Η μέση τιμή της αναμενόμενης τιμής μιας εκθετικά κατανεμημένης μεταβλητής T δίνεται από τον τύπο:

$$\mathrm{E}[T] = \frac{1}{\lambda}$$

Η διασπορά της μεταβλητής T δίνεται από τον τύπο:

$$\mathrm{Var}[T] = \frac{1}{\lambda^2}$$

Η διάμεσος της μεταβλητής T δίνεται από τον τύπο:

$$\mathrm{m}[T] = \frac{\ln 2}{\lambda} < \mathrm{E}[T]$$

Συνάρτηση αξιοπιστίας

Η συνάρτηση αξιοπιστίας της εκθετικής κατανομής δίνεται από τον τύπο:

$$\mathrm{S}(\mathrm{t}) = \int_t^\infty f(\tau)d\tau = e^{-\lambda t}$$

Συνάρτηση διακινδύνευσης

Η συνάρτηση διακινδύνευσης είναι:

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

Βλέπουμε ότι η συνάρτηση διακινδύνευσης είναι σταθερή και ανεξάρτητη του χρόνου t, δηλαδή είναι ανεξάρτητη από την ηλικία της υπό μελέτη μονάδας. Το γεγονός αυτό είναι που μειώνει τη χρηστικότητα του μοντέλου στο ελάχιστο, αφού είναι ελάχιστες οι ρεαλιστικές εφαρμογές στις οποίες η γήρανση δεν επηρεάζει την πιθανότητα θνησιμότητας.

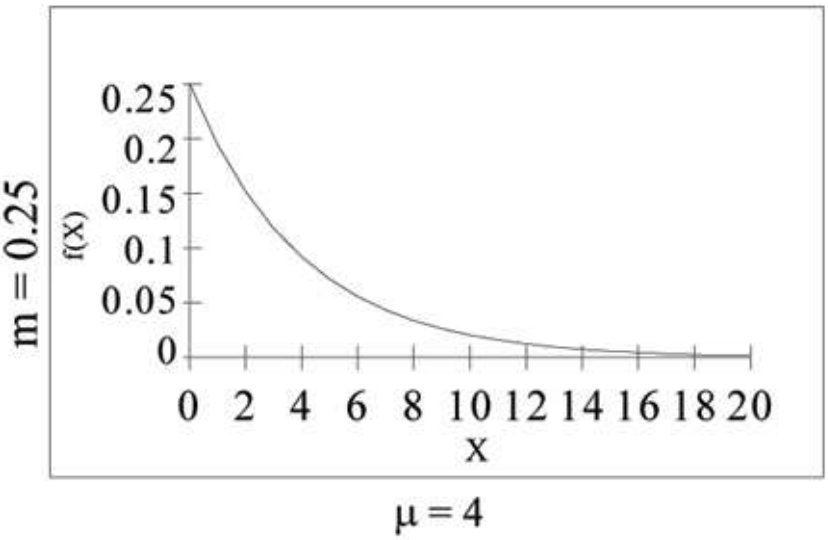
Παράδειγμα 1.5

Χαρακτηριστικό παράδειγμα εκθετικής κατανομής (Lopez, 2009)

Έστω η συνεχής τυχαία μεταβλητή X που εκφράζει το χρόνο εξυπηρέτησης ενός πελάτη σε κάποιο τηλεφωνικό κέντρο με μέση τιμή $\mu = \overline{X} = 4\,\text{min}$ και παράμετρο m=0.25. Η τυπική απόκλιση ταυτίζεται με τη μέση τιμή, άρα η κατανομή περιγράφεται ως εξής X~Exp(m). Δηλαδή, X~Exp(0.25). Η συνάρτηση πυκνότητας πιθανότητας ορίζεται ως:

$$f(x) = me^{-mx},$$

και τελικά καταλήγουμε στο γράφημα που φαίνεται παρακάτω:



Γράφημα 1.4

1.5.3 Κατανομή Weibull

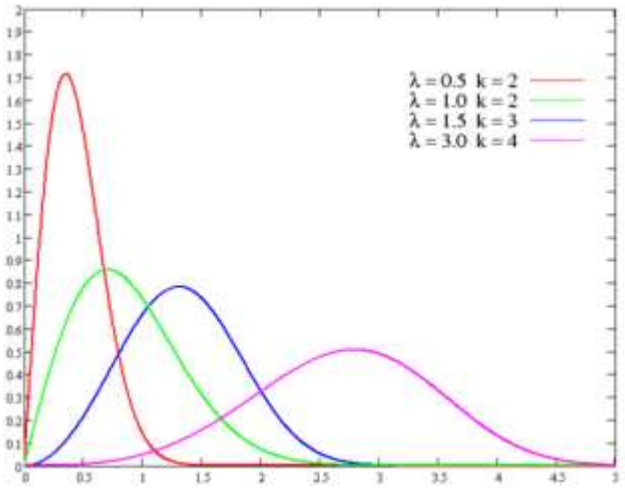
Αποτελεί την πλέον διαδεδομένη κατανομή για τις μεταβλητές διάρκειας ζωής, χάρη στη μεγάλη ευελιξία που τη χαρακτηρίζει ως μοντέλο αξιοπιστίας.

Συνάρτηση πυκνότητας πιθανότητας

Η συνάρτηση πυκνότητας πιθανότητας δίνεται από τη σχέση:

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

όπου η μεταβλητή λ>0 ονομάζεται παράμετρος κλίμακας και η μεταβλητή k>0 καλείται παράμετρος σχήματος. (Βούρος, 2007) Παρατηρούμε ότι για k=1 έχουμε την περίπτωση της Εκθετικής κατανομής, ενώ για διάφορες τιμές των παραμέτρων η μορφή της σ.π.π φαίνεται παρακάτω:



Γράφημα 1.5

Για k=2 σημειώνεται ότι έχουμε την περίπτωση της κατανομής Rayleigh, η οποία είναι γνωστή για τη χρήση της στη μοντελοποίηση του ύψους κύματος σχετικά με τους θαλάσσιους κυματισμούς.

Συνάρτηση αξιοπιστίας

Η συνάρτηση αξιοπιστίας είναι:

$$\mathrm{S(t)} = \int_t^{\infty} k\lambda^{-1}(\tau/\alpha)^{\eta-1} e^{-(\tau/\lambda)^k} d\tau = \int_{(t/\lambda)^{\eta}}^{\infty} e^{-u} du \text{, όπου } u = (t/\lambda)^k \text{ και τελικά,}$$

$$S(t) = \exp\{-(t/\lambda)^k\}$$

Συνάρτηση διακινδύνευσης

Η συνάρτηση διακινδύνευσης είναι:

$$h(t) = \frac{f(t)}{S(t)} = k\lambda^{-k} t^{k-1}$$

Μέση τιμή και διασπορά

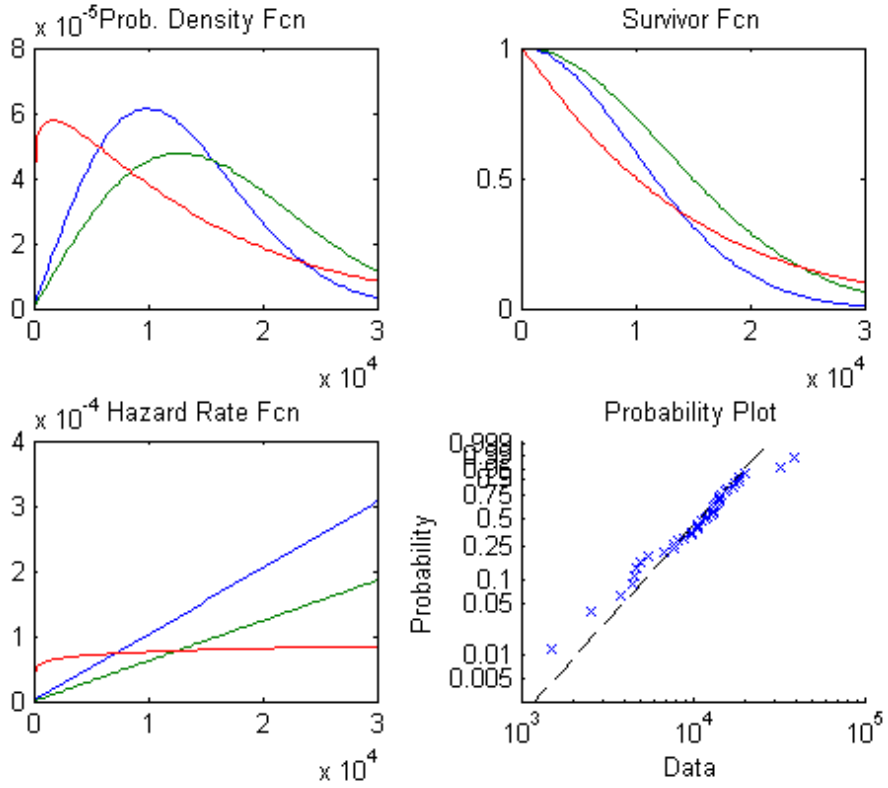
Για τη μέση τιμή και τη διασπορά στο μοντέλο της κατανομής Weibull έχουμε αντίστοιχα:

$$\mathrm{E}(\mathrm{T}) = \lambda\Gamma(1+k^{-1}) \text{ και } \mathrm{V}(\mathrm{T}) = \lambda^2[\Gamma(1+2k^{-1}) - \{\Gamma(1+k^{-1})\}^2],$$

όπου Γ η συνάρτηση Γάμμα για την οποία ισχύει:

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

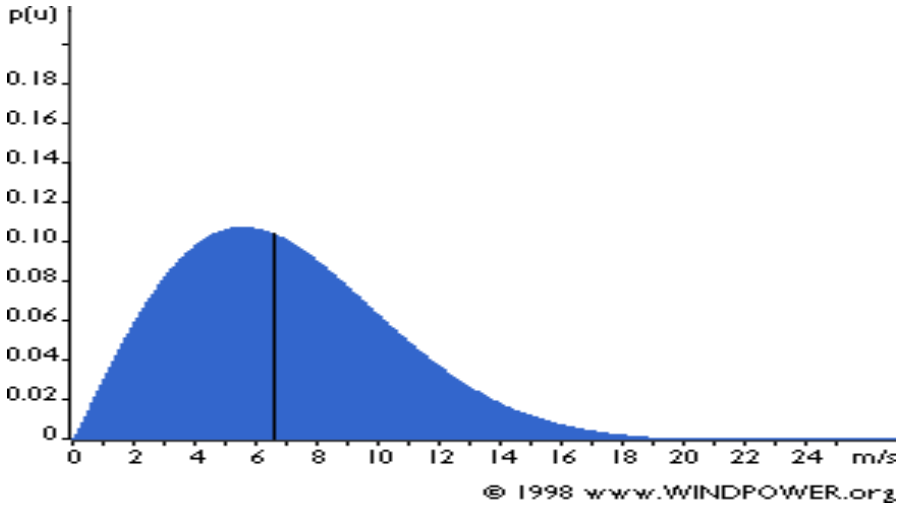
Στη συνέχεια βλέπουμε για διάφορες τιμές των παραμέτρων κλίμακας και σχήματος το γράφημα της συνάρτησης πυκνότητας πιθανότητας της κατανομής Weibull, καθώς και της συνάρτησης επιβίωσής της. Ενδιαφέρον παρουσιάζει η καμπύλη της συνάρτησης διακινδύνευσης, καθώς παρατηρείται αυξητική τάση όσο μεγαλώνει η ηλικία. Είναι επόμενο, λοιπόν, η κατανομή Weibull να αποτελεί χρήσιμο μοντέλο διάρκειας ζωής.



Γράφημα 1.6

Παρόλο που η κατανομή Weibull είναι ίσως το πιο διαδεδομένο μοντέλο στην ανάλυση αξιοπιστίας, οι πρακτικές της εφαρμογές δε σταματούν εκεί. Αρκετά καθημερινά φαινόμενα περιγράφονται ικανοποιητικά με χρήση της κατανομής αυτής, όπως είναι για παράδειγμα τα ανεμολογικά χαρακτηριστικά στις περιοχές της εύκρατης ζώνης για ύψος μέχρι 100m από το έδαφος.

Στο παρακάτω γράφημα φαίνεται η κατανομή πυκνότητας πιθανότητας p(v) της τυχαίας μεταβλητής v που εκφράζει την ταχύτητα του ανέμου σε μια περιοχή ενός τετραγωνικού μέτρου εύκρατης ζώνης.



Γράφημα 1.7

Διάμεσος=6.6m/sec

Αυτό σημαίνει ότι το μισό χρόνο η ταχύτητα του ανέμου έχει τιμή μικρότερη από 6.6m/sec και τον άλλο μισό μεγαλύτερη από 6.6m/sec

Μέση τιμή ανέμου: 7m/sec

1.5.4 Κατανομή Γάμμα *(Gamma distribution)*

Συνάρτηση πυκνότητας πιθανότητας

Η δι-παραμετρική κατανομή Γάμμα με παράμετρο κλίμακας $\lambda>0$ και παράμετρο θέσης $\alpha>0$ έχει σ.π.π

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

Συνάρτηση αξιοπιστίας

Η συνάρτηση αξιοπιστίας της κατανομής Γάμμα είναι σχετικά δύσχρηστη μορφολογικά και προκύπτει ως εξής:

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

και τελικά,

$S(t) = \frac{\Gamma(\alpha, \lambda t)}{\Gamma(\alpha)}$, όπου Γ(α,λt) είναι η άνω ατελής συνάρτηση Γάμμα *(upper incomplete Gamma function)* που δίνεται από τη σχέση:

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

Μέση τιμή και διασπορά

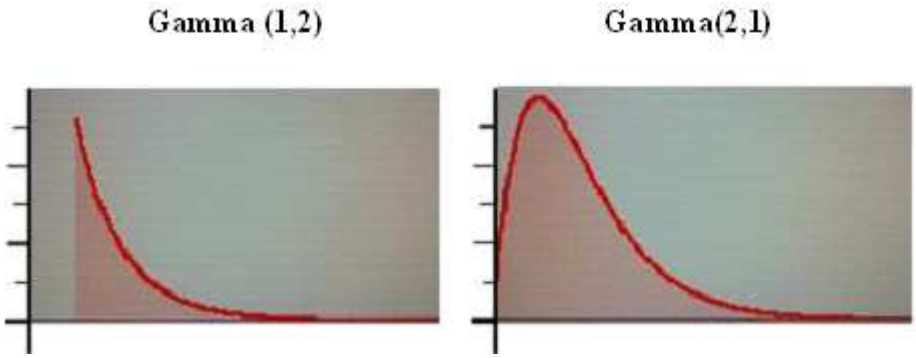
Η αναμενόμενη τιμή της τυχαίας μεταβλητής Τ δίνεται από τη σχέση:

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

και η διασπορά είναι:

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

Το γράφημα της συνάρτησης πυκνότητας πιθανότητας της δι-παραμετρικής κατανομής Γάμμα για κάποιες τιμές των παραμέτρων της φαίνεται παρακάτω:



Γράφημα 1.8

1.5.5 Κατανομή Gumbel

Η κατανομή Gumbel αλλιώς αναφέρεται στη βιβλιογραφία ως κατανομή ακραίων τιμών *(Smallest Extreme Value)* και επίσης αποτελεί μοντέλο διάρκειας ζωής (Καρώνη, 2005). Μεγάλο μέρος της πρακτικής της αξίας προκύπτει από τη σχέση:

T~Weibull $\Leftrightarrow$ lnT~Gumbel

Συνάρτηση Πυκνότητας Πιθανότητας

Η συνάρτηση πυκνότητας πιθανότητας δίνεται από τη σχέση:

$$f(t) = \sigma^{-1} e^{\{(\frac{t-\mu}{\sigma})S(t)\}}, -\infty < t < \infty$$

όπου μ η παράμετρος θέσης και σ η παράμετρος κλίμακας.

Συνάρτηση Επιβίωσης

Η συνάρτηση επιβίωσης της κατανομής Gumbel είναι:

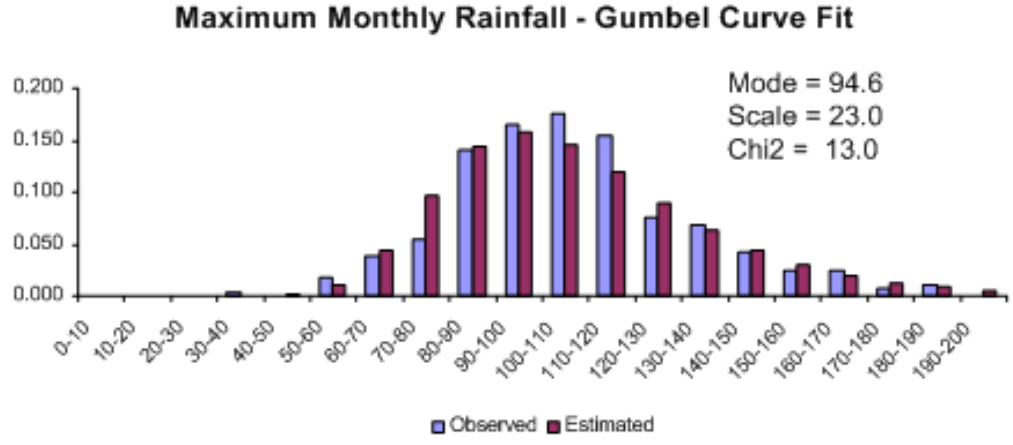
$$S(t) = e^{-e^{\{\frac{t-\mu}{\sigma}\}}}$$

Συνάρτηση Διακινδύνευσης

Τέλος, η συνάρτηση διακινδύνευσης δίνεται από τη σχέση:

$$h(t) = \sigma^{-1} e^{\frac{t-\mu}{\sigma}}.$$

Η κατανομή Gumbel μοντελοποιεί το πρόβλημα της διάρκειας ζωής κάποιων προϊόντων, των οποίων η καταναλωτική αξία μειώνεται κατακόρυφα μόλις συμπληρώσουν κάποια ηλικία. Είναι επίσης κατάλληλη για τη στατιστική μελέτη των βροχοπτώσεων μιας περιοχής σε μηνιαία βάση (Κουτσογιάννης, 2008). Στη συνέχεια παρατίθεται σχετικό γράφημα:



Γράφημα 1.9

1.5.6 Λογαριθμοκανονική κατανομή *(lognormal distribution)*

Σε πολλές περιπτώσεις η τυχαία μεταβλητή Τ που μας απασχολεί δεν ακολουθεί κανονική κατανομή, με έναν απλό μετασχηματισμό όμως συχνά

καταλήγουμε σε κανονική τυχαία μεταβλητή (E.Limpert et al, 2001). Ένας μετασχηματισμός που μπορεί να προσαρμόζει τα πειραματικά δεδομένα σε κανονική κατανομή είναι ο logT. Θεωρώντας γνωστή την κανονική κατανομή με τυπική απόκλιση σ>0 , μέση τιμή μ και σ.π.π :

$$f(t)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(t-\mu)^2/(2\sigma^2)}, t>0 \quad , \ t>0$$

Ορίζουμε τη λογαριθμο-κανονική κατανομή ως εξής:

Αν T ~ log-normal(μ, σ^2)

Τότε Y = logT ~ N (μ, σ^2)

Συνάρτηση πυκνότητας πιθανότητας

Η συνάρτηση πυκνότητας πιθανότητας δίνεται από τη σχέση:

$$f(t)=\frac{1}{\sigma t\sqrt{2\pi}}e^{-(\log t-\mu)^2/(2\sigma^2)}, t>0 \quad , \ t>0$$

Μέση τιμή και διασπορά

Η μέση τιμή και η διασπορά δίνονται από τις σχέσεις:

$$\mathrm{E(T)}=e^{\mu+\frac{1}{2}\sigma^2}$$

και

$$Var(T)=E(T^2)-(E(T))^2=e^{2\mu+\sigma^2}(e^{\sigma^2}-1).$$

<u>Παράδειγμα 1.6</u>

(Δαμιανού, 2003)

- Έστω T η χρονική διάρκεια επώασης μιας μεταδοτικής νόσου, τότε έχει διαπιστωθεί εμπειρικά ότι η μεταβλητή logT ακολουθεί κανονική κατανομή.
- Ακόμα, η αντοχή X ενός υλικού σε συγκεκριμένες καταπονήσεις δεν ακολουθεί κανονική κατανομή. Η μεταβλητή logX όμως ακολουθεί.
- Τέλος, αν P η ποσότητα ενός φαρμάκου που παραμένει στον οργανισμό μας μετά από συγκεκριμένο χρονικό διάστημα από τη στιγμή χορήγησής του έχει παρατηρηθεί ότι η logP ακολουθεί κατά προσέγγιση κανονική κατανομή.

Το επόμενο παράδειγμα είναι αρκετά διαφωτιστικό όσον αφορά την πρακτική αξία της λογαριθμο-κανονικής κατανομής.

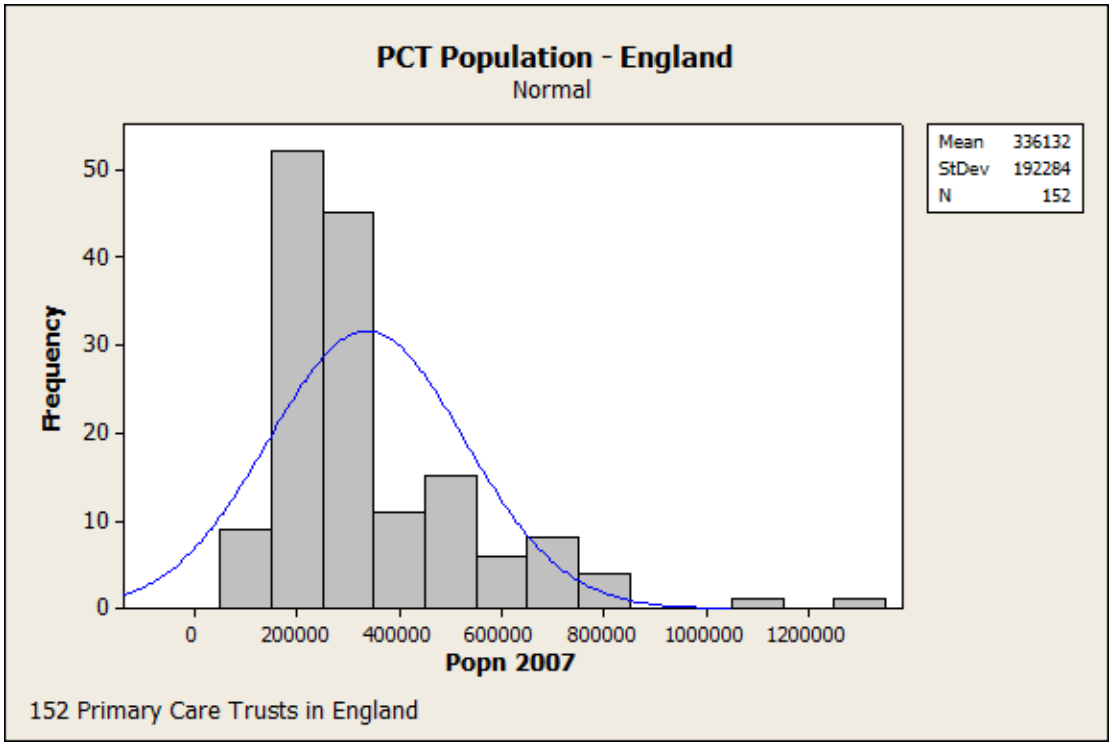
Παράδειγμα 1.7

(Health Service Journal, 2009)

Μετά την ολοκλήρωση μιας βρετανικής έρευνας που αφορά στο πλήθος των ατόμων που εξυπηρετήθηκαν από ένα δείγμα 152 εργαζομένων του εθνικού συστήματος υγείας της Αγγλίας κατά το ημερολογιακό έτος 2007, τα αποτελέσματα δόθηκαν στη δημοσιότητα.

Στο γράφημα 1.10 θα δούμε το ιστόγραμμα των παρατηρήσεων, καθώς και μια προσπάθεια προσέγγισης του με την κανονική κατανομή. Στη συνέχεια, στο γράφημα 1.11 θα δούμε το ίδιο ιστόγραμμα παρατηρήσεων, όπως επίσης και την προσέγγισή του μέσω της λογαριθμο-κανονικής κατανομής.

Ιστόγραμμα - Γράφημα συνάρτησης πυκνότητας πιθανότητας κανονικής Κατανομής

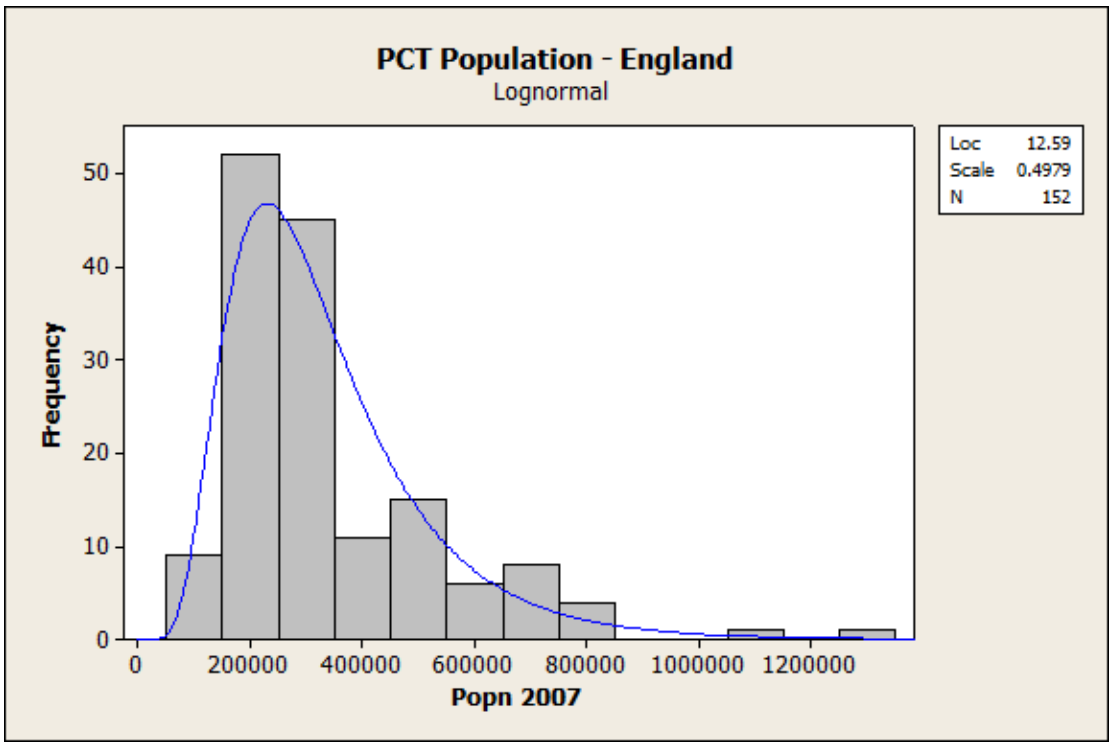


Γράφημα 1.10

Είναι προφανές ότι η κανονική κατανομή δεν είναι η καταλληλότερη για την προσέγγιση των στατιστικών δεδομένων της έρευνας, αφού μάλιστα δεν εμπεριέχει καθόλου τις τιμές που ξεπερνούν το 1.000.000 σε πλήθος ατόμων.

Αντίθετα, από το επόμενο ιστόγραμμα καθίσταται σαφές ότι η λογαριθμο-κανονική κατανομή προσεγγίζει με πολύ μεγαλύτερη ακρίβεια τα δεδομένα.

Ιστόγραμμα - Γράφημα συνάρτησης πυκνότητας πιθανότητας λογαριθμο-κανονικής Κατανομής



Γράφημα 1.11

Βλέπουμε πως το ιστόγραμμα των λογαριθμημένων παρατηρήσεων, περιγράφεται με αρκετά μεγάλη ακρίβεια από τη συνάρτηση πυκνότητας πιθανότητας της λογαριθμο-κανονικής κατανομής.

Από το παραπάνω παράδειγμα φαίνεται η χρηστική αξία της λογαριθμο-κανονικής κατανομής σε πραγματικά δεδομένα, καθώς και η προσαρμοστικότητα αυτής στην προσέγγιση των πειραματικών παρατηρήσεων.

Σημειώνεται ότι η λογαριθμο-κανονική κατανομή έχει καλή προσαρμογή σε πραγματικά προβλήματα, όπου η τυχαία μεταβλητή T παίρνει σχετικά μικρές τιμές.

ΚΕΦΑΛΑΙΟ 2

ΤΟ ΗΜΙ-ΠΑΡΑΜΕΤΡΙΚΟ ΜΟΝΤΕΛΟ ΑΝΑΛΟΓΙΚΗΣ ΔΙΑΚΙΝΔΥΝΕΥΣΗΣ ΤΟΥ COX

2.1 Εισαγωγικά στοιχεία

Γενικά, βασικό στόχο κάθε έρευνας ή επιστημονικής ανάλυσης αποτελεί η μοντελοποίηση των σχέσεων μεταξύ μεταβλητών. Οι μεταβλητές κατηγοριοποιούνται σε εξαρτημένες και ανεξάρτητες. Οι ανεξάρτητες ή επεξηγηματικές μεταβλητές είναι εκείνες των οποίων οι τιμές καταγράφονται κατά τη διάρκεια του πειράματος, ώστε τελικά να καταλήξει ο ερευνητής σε ένα συμπέρασμα για την εξαρτώμενη μεταβλητή. Επίσης, οι δύο βασικές κατηγορίες μεταβλητών είναι οι ποσοτικές και οι ποιοτικές. Οι ποσοτικές μεταβλητές χωρίζονται σε συνεχείς και διακριτές και οι ποιοτικές σε κατηγορικές και μεταβλητές διάταξης..

Το τελικό συμπέρασμα για την εξαρτώμενη μεταβλητή προκύπτει ύστερα από ανάλυση, κατά τη διάρκεια της οποίας καθορίζεται μεταξύ άλλων το επίπεδο σημαντικότητας κάθε επεξηγηματικής μεταβλητής. Για να συμβεί αυτό, είναι απαραίτητη η εκτίμηση των συντελεστών παλινδρόμησης, η οποία επιτυγχάνεται μέσω της συνάρτησης μερικής πιθανοφάνειας, όταν όλες οι παρατηρήσεις είναι πλήρεις, ή μέσω παραλλαγών αυτής, όταν στα δεδομένα εμπεριέχονται και αποκομμένες παρατηρήσεις.

Για τα προβλήματα της ανάλυσης επιβίωσης και όχι μόνο, ιδιαίτερα δημοφιλές είναι το μοντέλο παλινδρόμησης του Cox ή αλλιώς μοντέλο αναλογικής διακινδύνευσης του Cox, το οποίο επιτρέπει τη σύγκριση των επεξηγηματικών μεταβλητών ακόμα και σε προβλήματα που περιέχουν αποκομμένες παρατηρήσεις. Τη χρήση του μοντέλου αυτού θα εξετάσουμε αναλυτικά στη συνέχεια.

Στο σημείο αυτό αξίζει να αναφερθούμε στο γεγονός ότι ακόμα και οι λογοκριμένες (αποκομμένες) παρατηρήσεις είναι πολύτιμες για την εξαγωγή στατιστικών συμπερασμάτων. Αν για παράδειγμα έχει οριστεί η διάρκεια

παρακολούθησης γυναικών με καρκίνο του στήθους στα πέντε χρόνια, με ενδεχόμενο το θάνατο και εναλλακτικό ενδεχόμενο την ίαση, τότε ενδέχεται μέχρι και το τέλος αυτής σε κάποιες περιπτώσεις γυναικών να μην έχει προκληθεί τίποτε από τα δύο. Αυτό, όμως δε σημαίνει ότι δεν επιβίωσαν τουλάχιστον μέχρι αυτή τη χρονική στιγμή. Επομένως, το να μη λάβει κάποιος υπόψη τα δεδομένα των γυναικών αυτών θα διαστρέβλωνε σημαντικά τα αποτελέσματα της έρευνας.

Στη συνέχεια θα δοθεί ο ορισμός του πολλαπλού γραμμικού μοντέλου παλινδρόμησης, καθώς και μια σύντομη περιγραφή των μεθόδων που χρησιμοποιούνται για την εκτίμηση των συντελεστών αυτού.

2.1.1 Πολλαπλό γραμμικό μοντέλο παλινδρόμησης

Το πολλαπλό γραμμικό μοντέλο παλινδρόμησης (Κουκουβίνος, 2005) χρησιμοποιείται για να μελετήσει τη σχέση μεταξύ μιας εξαρτώμενης μεταβλητής και διάφορων ανεξάρτητων μεταβλητών.

Η γενική μορφή του μοντέλου είναι η εξής:

$\tilde{y} = X\tilde{\beta} + \tilde{\varepsilon}$, ή ισοδύναμα:

$$y_i = \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i, i = 1,2,....,n \quad (1)$$

όπου Χ ο πίνακας των εξαρτημένων (επεξηγηματικών) μεταβλητών $\tilde{x}_1,\ldots,\tilde{x}_k$ και $\tilde{y}$ η εξαρτώμενη ή αποκριτική μεταβλητή. Οι δείκτες i είναι οι n το πλήθος παρατηρήσεις του δείγματος και ε_i το σφάλμα κάθε παρατήρησης, με $\tilde{\varepsilon} = \{\varepsilon_1\varepsilon_2.....\varepsilon_n\}$.

2.1.2 Μέθοδοι για την εκτίμηση των συντελεστών ενός γραμμικού μοντέλου παλινδρόμησης

Οι πιο συνηθισμένες μέθοδοι για την εκτίμηση των συντελεστών ενός μοντέλου παλινδρόμησης είναι η μέθοδος ελαχίστων τετραγώνων και η μέθοδος μέγιστης πιθανοφάνειας (Smith, 1997), οι οποίες περιγράφονται εν συντομία στη συνέχεια.

2.1.2.1 Μέθοδος Ελαχίστων Τετραγώνων *(Least Square Method)*

Η εκτιμήτρια $\hat{\beta}$ του διανύσματος των συντελεστών του γραμμικού μοντέλου δίνεται από τη σχέση:

$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}y = (\frac{1}{n}\sum x_i x_i')^{-1}(\frac{1}{n}\sum x_i y_i)$$

Αντικαθιστώντας το διάνυσμα των συντελεστών στο μοντέλο (1) έχουμε την εκτίμηση της μεταβλητής απόκρισης $\hat{y}$.

2.1.2.2 Μέθοδος Μέγιστης Πιθανοφάνειας *(Maximum Likelihood Estimation)*

Θεωρούμε τη συνάρτηση πυκνότητας πιθανότητας $f(X \mid \tilde{\beta})$ και ορίζουμε την πιθανοφάνεια ως:

$L(\tilde{\beta}) = \prod_{i=1}^{k} f(x_i \mid \tilde{\beta}) = f(x_1 \mid \tilde{\beta}) f(x_2 \mid \tilde{\beta}) \ldots f(x_k \mid \tilde{\beta})$, ενώ λογαριθμώντας έχουμε:

$$\ln\mathrm{L}(\tilde{\beta}) = \ln f(x_1 \mid \tilde{\beta}) + \ldots\ldots + \ln f(x_k \mid \tilde{\beta})$$

και τελικά παραγωγίζοντας ως προς $\tilde{\beta}$ έχουμε την εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\beta}$ που προκύπτει από τη σχέση:

$$\frac{1}{L(\hat{\beta})}\frac{\partial L(\hat{\beta})}{\partial \hat{\beta}} = 0$$

2.2 Το μοντέλο αναλογικής διακινδύνευσης του Cox
(Cox proportional hazard model)

Το μοντέλο αναλογικής διακινδύνευσης του Cox, είναι ένα μοντέλο παλινδρόμησης που μοντελοποιεί τη συνάρτηση διακινδύνευσης $h(t)$ και επιτρέπει τη σύγκριση των επεξηγηματικών μεταβλητών, ώστε να γίνει δυνατή η επιλογή των στατιστικά σημαντικότερων εξ αυτών. Συγκεκριμένα, εξετάζει την επίδρασή τους στην εκτίμηση της μεταβλητής απόκρισης και χρησιμοποιείται, όπως ήδη αναφέραμε, σε δεδομένα που περιλαμβάνουν εκτός από πλήρεις και αποκομμένες παρατηρήσεις.

Είναι αναμφισβήτητα ένα από τα δημοφιλέστερα μοντέλα που χρησιμοποιούνται στην ανάλυση επιβίωσης, λόγω της απλής εφαρμογής του (Molinero, 2001).

2.2.1 Ορισμός βασικών συναρτήσεων

Για την κατανόηση του μοντέλου του Cox πρέπει πρώτα να ορίσουμε τις συναρτήσεις $S(t), h(t), H(t)$. Για τις οποίες έχουμε:

Συνάρτηση διακινδύνευσης

Η συνάρτηση διακινδύνευσης για το μοντέλο αναλογικού κινδύνου του Cox, με μεταβλητή απόκρισης t και διάνυσμα επεξηγηματικών μεταβλητών $\tilde{x}$ φαίνεται παρακάτω:

$$h(t,\tilde{x}) = h_0(t)e^{\beta_1 x_1 + ... + \beta_k x_k}$$

Σημειώνεται ότι η συνάρτηση διακινδύνευσης $h_0(t)$ εκφράζει τον κίνδυνο θανάτου ή αποτυχίας, όταν όλες οι επεξηγηματικές μεταβλητές x_j, για j=1,2,….,k είναι ίσες με μηδέν.

Θεωρούμε τώρα το λόγο:

$$\frac{h(t,\tilde{x}_i)}{h(t,\tilde{x}_{i+1})} = \frac{h_0(t)e^{\tilde{x}_i'\tilde{\beta}}}{h_0(t)e^{\tilde{x}_{i+1}'\tilde{\beta}}} = e^{(\tilde{x}_i - \tilde{x}_{i+1})'\tilde{\beta}},$$

που είναι γνωστός ως αναλογία κινδύνου *(hazard ratio)*.

Το διάνυσμα $\tilde{x}_i$ αντιστοιχεί στο διάνυσμα των k επεξηγηματικών μεταβλητών για το i-οστό άτομο που συμμετέχει στο πείραμα.

Εφόσον η τιμή των επεξηγηματικών μεταβλητών δεν εξαρτάται από το χρόνο t, η ποσότητα $e^{(\tilde{x}_i - \tilde{x}_{i+1})'\tilde{\beta}}$ είναι σταθερή και για αυτό το μοντέλο του Cox είναι γνωστό ως μοντέλο αναλογικής διακινδύνευσης.

Ισοδύναμη έκφραση του μοντέλου είναι η εξής:

$$\ln[\frac{\mathrm{h}(\mathrm{t},\tilde{x})}{h_0(t)}] = \beta_1 x_1 + ... + \beta_k x_k$$

Η τελευταία μορφή απλοποιεί σημαντικά την ερμηνεία των συντελεστών. Συγκεκριμένα, το β_i ισοδυναμεί με το λογάριθμο του σχετικού κινδύνου όταν έχουμε αύξηση μιας μονάδας στη μεταβλητή x_i, ενώ οι άλλες παραμένουν σταθερές. Με άλλα λόγια, η σχέση e^{β_i} εκφράζει το σχετικό ρίσκο που λαμβάνεται όταν η επεξηγηματική μεταβλητή x_i αυξάνεται κατά μια μονάδα, ενώ οι υπόλοιπες μεταβλητές παραμένουν σταθερές.

Γενικά, η συνάρτηση διακινδύνευσης $h(t,\tilde{x})$ εξαρτάται από δύο παράγοντες: τη συνάρτηση $h_0(t)$ και το διάνυσμα $\tilde{\beta}=(\beta_1,\beta_2,...,\beta_k)$. Η συνάρτηση $h_0(t)$, η οποία εξαρτάται μόνο από το χρόνο, στο μοντέλο του Cox αφήνεται αυθαίρετη και θεωρείται ίδια για το σύνολο των n ατόμων της μελέτης. Για το λόγο αυτό, το μοντέλο αναλογικού κινδύνου του Cox θεωρείται ημι-παραμετρικό *(semi-parametric),* αφού δεν καθορίζει τη μορφή της $h_0(t)$, αλλά υποθέτει ότι οι επιδράσεις των μεταβλητών παραμένουν σταθερές στο χρόνο και είναι προσθετικές σε μια συγκεκριμένη κλίμακα.

Σωρευτική συνάρτηση διακινδύνευσης

Η σωρευτική συνάρτηση διακινδύνευσης για το μοντέλο του Cox ορίζεται ως εξής:

$$H(t,\tilde{x})=\int_0^t h_0(u)e^{\tilde{x}'\tilde{\beta}}du=H_0(t)e^{\tilde{x}'\tilde{\beta}}$$

Συνάρτηση αξιοπιστίας

Ενώ η συνάρτηση επιβίωσης ή αξιοπιστίας δίνεται από τη σχέση:

$$S(t)=\exp\{-H(t)\}\Rightarrow$$

$$S(t)=\exp\left\{-H_0(t)e^{x'\beta}\right\}=\left[S_0(t)\right]^{e^{x'\beta}}$$

2.2.2 Εκτίμηση των συντελεστών $\tilde{\beta}_i$

Για την εκτίμηση των συντελεστών με τη μέθοδο μέγιστης πιθανοφάνειας είναι απαραίτητο να οριστεί πρώτα η συνάρτηση πιθανοφάνειας $L(\tilde{\beta})$, για την εύρεση της οποίας υπάρχουν αρκετές μέθοδοι. Μεταξύ αυτών η μέθοδος που δόθηκε

πρώτη και χρησιμοποιείται ακόμα και σήμερα, οφείλεται στον Cox και βασίζεται στη συνάρτηση μερικής πιθανοφάνειας.

Συνάρτηση μερικής πιθανοφάνειας

Ας θεωρήσουμε ότι μετά τη λήξη ενός πειράματος έχουμε n αριθμό παρατηρήσεων, εκ των οποίων οι k παρατηρήσεις είναι μη αποκομμένες (πλήρεις), ενώ οι n-k είναι αποκομμένες. Θεωρούμε τους αντίστοιχους χρόνους αποτυχίας $t_{(1)}, t_{(2)}, ..., t_{(k)}$. Τότε αν R_j το σύνολο των ατόμων που βρίσκονται σε κίνδυνο τη χρονική στιγμή $t_{(j)}$ η συνάρτηση μερικής πιθανοφάνειας ορίζεται ως εξής:

$$L(\underset{\sim}{\beta}) = \prod_{j=1}^{k} \frac{e^{\underset{\sim}{x}_{(j)}'\underset{\sim}{\beta}}}{\sum_{i \in R_j} e^{\underset{\sim}{x}_{(i)}'\underset{\sim}{\beta}}}$$

Το γεγονός ότι στην παραπάνω διαδικασία δεν προσδιορίζεται η $h_0(t)$, εξηγεί τον όρο "μερική πιθανοφάνεια". Στο σημείο αυτό, αξίζει να σημειωθεί ότι η μοναδική συμβολή των πειραματικών δεδομένων στην πιθανοφάνεια, είναι οι χρονικές στιγμές στις οποίες παρατηρούνται τα μη αποκομμένα γεγονότα.

Όπως απέδειξε ο Cox, η μερική πιθανοφάνεια μπορεί να χρησιμοποιηθεί ως μια συνηθισμένη συνάρτηση πιθανοφάνειας για τον υπολογισμό της εκτίμησης $\hat{\beta}$ των $\tilde{\beta}_i$. Είναι αποτελεσματική, όμως, μόνο στην περίπτωση που κάθε χρόνος επιβίωσης εμφανίζεται μόνο μια φορά στα πειραματικά μας δεδομένα.

Σε αντίθετη περίπτωση οι απαραίτητοι υπολογισμοί για την εκτίμηση των συντελεστών είναι αρκετά πολύπλοκοι. Τα στατιστικά πακέτα που χρησιμοποιούνται για την εκτίμηση αυτών, χρησιμοποιούν συχνά μια προσέγγιση που δόθηκε το 1990 από τον βρετανό μαθηματικό R.Peto, η οποία αποδεικνύεται αποτελεσματική υπό την προϋπόθεση ο αριθμός των ταυτόχρονων γεγονότων να είναι μικρός σε σύγκριση με τα άτομα που βρίσκονται σε κίνδυνο για κάθε χρονική στιγμή. Επίσης οι προσεγγίσεις της μερικής πιθανοφάνειας των Breslow και Efron αποδεικνύονται αποτελεσματικές στην περίπτωση πολλαπλών ταυτόχρονων θανάτων.

Το μοντέλο του Cox ισχύει μόνο κάτω από την υπόθεση της αναλογικής διακινδύνευσης. Αυτό σημαίνει ότι είναι αναγκαίο να γίνεται έλεγχος της αναλογικότητας των κινδύνων.

2.2.3 Το μοντέλο διακινδύνευσης του Cox στην περίπτωση χρονο-εξαρτώμενων επεξηγηματικών μεταβλητών

Στο μοντέλο αναλογικής διακινδύνευσης του Cox οι μεταβλητές είναι ποσοτικές και σταθερές ως προς το χρόνο. Υπάρχει όμως περίπτωση αυτές να μεταβάλλονται, με αποτέλεσμα να έχουμε δύο κατηγορίες χρονο-εξαρτώμενων *(time-dependent)* μεταβλητών: τις εσωτερικές και τις εξωτερικές.

Ας θεωρήσουμε ένα πρόβλημα επιβίωσης που αφορά την εξέλιξη ασθενών με όγκο στον ήπαρ. Τότε στην πρώτη κατηγορία ανήκουν οι μεταβλητές των οποίων οι τιμές μπορούν να ληφθούν μόνο όσο ο ασθενής βρίσκεται εν ζωή, ενώ στη δεύτερη κατηγορία ανήκουν οι μεταβλητές που μπορούν να λάβουν τιμή ανεξαρτήτως από το αν ο ασθενής βρίσκεται εν ζωή ή όχι. Ένα παράδειγμα εσωτερικής χρονο-εξαρτώμενης μεταβλητής αποτελεί το μέγεθος του όγκου του ασθενούς, ενώ ένα παράδειγμα εξωτερικής χρονο-εξαρτώμενης μεταβλητής αποτελεί η θερμοκρασία του αέρα την κάθε χρονική στιγμή.

Συνάρτηση διακινδύνευσης για το μοντέλο του Cox με χρονο-εξαρτώμενες μεταβλητές

Έστω ότι στο προηγούμενο παράδειγμα ο αριθμός των ασθενών που συμμετέχουν στην έρευνα είναι n και το σύνολο των συμμεταβλητών, δηλαδή το σύνολο των επεξηγηματικών μεταβλητών που εισάγονται στο πρόβλημα, είναι p. Τότε, η συνάρτηση διακινδύνευσης για το i-οστό άτομο που βρίσκεται υπό μελέτη δίνεται από την παρακάτω σχέση:

$$h_i(t) = h_0(t)e^{\sum_{j=1}^{p}\beta_j x_{ji}},$$

όπου x_{ji} η τιμή της j-οστής επεξηγηματικής μεταβλητής για το i-οστό άτομο.

Η σχέση αυτή αποτελεί ισοδύναμη μορφή της σχέσης:

$$h(t,\tilde{x}) = h_0(t)e^{\beta_1 x_1 + ... + \beta_k x_k}.$$

Αν τώρα θεωρήσουμε τις συμμεταβλητές x_{ji} μη σταθερές, αλλά εξαρτώμενες από το χρόνο τότε έχουμε την τελική μορφή:

$$\tilde{h}(t) = h_0(t)e^{\sum_{j=1}^{p}\beta_j \tilde{x}_j(t)}\tilde{x}_j(t)$$

Εκτιμήτρια μερικής πιθανοφάνειας

Ο λογάριθμος της μερικής πιθανοφάνειας στην περίπτωση των χρονο-εξαρτώμενων μεταβλητών δίνεται από τη σχέση:

$\ln L(\tilde{\beta}) = \sum_{i=1}^{n} \delta_i (\sum_{j=1}^{p} x_{ji}(t_i)\beta_j - \log(\sum_{k \in R(t_i)} e^{\beta_j x_{jk}(t_i)}))$, όπου:

t_i η χρονική στιγμή θανάτου του i-οστού ατόμου

R(t_i): το σύνολο των ατόμων σε κίνδυνο τη χρονική στιγμή t_i

δ_i: η χαρακτηριστική συνάρτηση αποκοπής για την οποία αν Α το σύνολο των αποκομμένων παρατηρήσεων και Β το σύνολο των μη αποκομμένων έχουμε:

$$\delta_i = \begin{cases} 0, & i \in A \\ 1, & i \in B \end{cases}.$$

Είναι προφανές ότι στην περίπτωση των χρονο-εξαρτώμενων συμμεταβλητών και η σχετική διακινδύνευση $d(t) = \frac{h_i(t)}{h_0(t)}$ εξαρτάται από το χρόνο, επομένως στην περίπτωση αυτή δεν ισχύει η υπόθεση της αναλογικής διακινδύνευσης.

2.2.4 Το στρωματοποιημένο μοντέλο του Cox

Στην περίπτωση που σε μια στατιστική μελέτη έχουμε μια ή παραπάνω κατηγορικές μεταβλητές, η υπόθεση της αναλογικότητας δύναται να μην ισχύει και για το λόγο αυτό πρέπει να ελεγχθεί.

Αν επιπλέον, σε μια πειραματική διαδικασία μια από τις ανεξάρτητες μεταβλητές είναι για παράδειγμα το φύλο, η συνάρτηση διακινδύνευσης μπορεί να διαμορφώνεται διαφορετικά για τις δύο τιμές της μεταβλητής, έτσι ώστε να ισχύει η υπόθεση της αναλογικότητας που δε θα ίσχυε διαφορετικά. Η παραπάνω διαδικασία αποτελεί στρωματοποίηση του μοντέλου ως προς τη συγκεκριμένη κατηγορική μεταβλητή.

Στην περίπτωση αυτή, τα άτομα για τα οποία η κατηγορική μεταβλητή λαμβάνει την ίδια τιμή (π.χ άνδρας) θεωρούμε ότι ανήκουν στο ίδιο στρώμα. Αν το πλήθος των στρωμάτων είναι s, τότε για κάθε στρώμα λαμβάνουμε διαφορετική βασική συνάρτηση διακινδύνευσης h_{0k} (t), όπου k={1,2,…,s}.

Συνάρτηση διακινδύνευσης

Η συνάρτηση διακινδύνευσης του i-οστού ατόμου είναι η εξής:

$$h_{ik}(t) = h_{0k}(t)e^{\tilde{\beta}x_{ik}},$$

όπου k το στρώμα στο οποίο ανήκει το άτομο και $\tilde{x}_{ik}$ το διάνυσμα τιμών των επεξηγηματικών μεταβλητών για το άτομο αυτό.

Η εύρεση της εκτιμήτριας μέγιστης πιθανοφάνειας δεδομένης της συνάρτησης διακινδύνευσης, γίνεται κατά τα γνωστά όπως διατυπώθηκε στην ενότητα 2.2.2 ή με χρήση ειδικών στατιστικών πακέτων.

2.3 Έλεγχοι καταλληλότητας μοντέλου

Έχοντας περιγράψει τον τρόπο εύρεσης των εκτιμητριών μέγιστης πιθανοφάνειας, κρίνεται απαραίτητος πλέον ο έλεγχος της καταλληλότητας του μοντέλου για κάθε συντελεστή, ο οποίος πραγματοποιείται κατά βάση με δύο τρόπους:

- **Με χρήση της γραφικής μεθόδου**
- **Με χρήση ελέγχου υπολοίπων**

2.3.1 Γραφική μέθοδος ελέγχου καταλληλότητας μοντέλου

Ο γραφικός έλεγχος καταλληλότητας πραγματοποιείται ως εξής:
Λογαριθμώντας τη συνάρτηση επιβίωσης

$$S(t,\tilde{x}) = \exp\{-H(t,\tilde{x})\} = \exp\{-H_0(t)e^{\tilde{x}'\tilde{\beta}}\},$$

έχουμε:

$$\ln\{-\ln S(t,\tilde{x})\} - \ln H_0(t) = \tilde{x}'\tilde{\beta}.$$

Αυτό γραφικά σημαίνει ότι οι καμπύλες $\ln\{-\ln S(t,\tilde{x})\}$ και $\ln H_0(t)$ πρέπει να είναι παράλληλες για κάθε $\tilde{x}$ και t>0. Κατά συνέπεια όλες οι καμπύλες $\ln\{-\ln S(t,\tilde{x}_i)\}$ για διάφορα $\tilde{x}_i$ θα πρέπει να είναι παράλληλες αν ισχύει η αναλογικότητα κινδύνου.

Βάσει της παραπάνω διαπίστωσης μπορούμε να αποφανθούμε γραφικά, σχετικά με την καταλληλότητα του μοντέλου, πραγματοποιώντας έλεγχο για κάθε

μεταβλητή που εμπεριέχεται στο αρχικό πρόβλημα. Για να είναι η μέθοδος αξιόπιστη πρέπει ο αριθμός των συμμεταβλητών να είναι περιορισμένος. Το κύριο πλεονέκτημα, όμως, της μεθόδου, είναι ότι μπορεί να χρησιμοποιηθεί για οποιοδήποτε μοντέλο αναλογικής διακινδύνευσης και όχι μόνο για το μοντέλο του Cox.

Σημειώνεται ότι στην προηγούμενη μέθοδο για να γίνει η γραφική αναπαράσταση της καμπύλης $\ln\{-\ln S(t,\tilde{x})\}$ γίνεται χρήση της εκτιμήτριας Kaplan-Meier της συνάρτησης επιβίωσης, η οποία ορίζεται στη συνέχεια.

2.3.1.1 Εκτιμήτρια Kaplan-Meier

Η εκτιμήτρια Kaplan-Meier της συνάρτησης επιβίωσης η οποία πήρε το όνομά της από τους Edward L. Kaplan και Paul Meier (Kaplan et al., 1958), χρησιμοποιείται σε εφαρμογές που εμπεριέχουν από δεξιά αποκομμένες παρατηρήσεις γι' αυτό και η χρηστική της αξία, ιδιαίτερα στα προβλήματα της ανάλυσης επιβίωσης, είναι πολύ μεγάλη.

Έστω η χρονική στιγμή t_j, τότε ορίζουμε ως d_j τον αριθμό των υπό παρατήρηση γεγονότων που πραγματοποιούνται αυτή τη χρονική στιγμή.

Για παράδειγμα, ας υποθέσουμε ότι εκτελείται ένα πείραμα σχετικά με το χρόνο περιστροφής ενός τύπου ρουλεμάν σε κάποια συγκεκριμένη συχνότητα έως ότου αυτά υποστούν βλάβη. Έστω ότι τη χρονική στιγμή t_j αποτυγχάνουν 3 ρουλεμάν, τη χρονική στιγμή t_j' αποτυγχάνει ένα και τη χρονική στιγμή t_j'' δεν αποτυγχάνει κανένα. Τότε, σε αντιστοιχία έχουμε:

$$d_j = 3, d_j' = 1, d_j'' = 0.$$

Η εκτιμήτρια Kaplan-Meier ορίζεται ως εξής:

$$\widehat{S}(t) = \frac{n_1 - d_1}{n_1}\frac{n_2 - d_2}{n_2}.....\frac{n_i - d_i}{n_i},$$

όπου n_i ο αριθμός των μονάδων που ήταν σε λειτουργία ακριβώς πριν από τη χρονική στιγμή t_i και $i : t_i \le t < t_{i+1}$.

Τελικά,

$$\widehat{S}(t) = \prod_{j:t_j \le t} \frac{n_j - d_j}{n_j}$$

2.3.2 Έλεγχος καταλληλότητας μοντέλου μέσω υπολοίπων

Ένας άλλος τρόπος ελέγχου της καταλληλότητας του μοντέλου είναι μέσω υπολοίπων τα οποία αποτελούν ένα μέτρο συμφωνίας των προβλέψεων της στατιστικής ανάλυσης με τις πραγματικές τιμές.

Στη συνέχεια υποθέτουμε το μοντέλο του Cox με k επεξηγηματικές μεταβλητές τέτοιες ώστε να ισχύει:

$$\hat{\underset{\sim}{\beta}}'\tilde{x}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki}$$

και θεωρούμε ένα δείγμα n ατόμων όπου οι n-r χρόνοι επιβίωσης είναι δεξιά αποκομμένοι. Η συνάρτηση διακινδύνευσης για το i άτομο, με i=1,2,…,n, είναι η εξής:

$$\hat{h}_i(t) = \exp(\hat{\underset{\sim}{\beta}}'\tilde{x}_i)\hat{h}_0(t),$$

με $\hat{h}_0(t)$ να είναι η εκτίμηση της βασικής συνάρτησης διακινδύνευσης. Στη συνέχεια θα ορίσουμε ενδεικτικά κάποιες κατηγορίες υπολοίπων.

Υπόλοιπα Cox-Snell

Το Cox-Snell υπόλοιπο για το i άτομο, δίδεται από τον τύπο

$$r_{CSi} = \exp(\hat{\underset{\sim}{\beta}}'\tilde{x}_i)\hat{H}_0(t_i),$$

όπου t_i ο παρατηρούμενος χρόνος επιβίωσης του i-οστού ατόμου.

Τα υπόλοιπα αυτά χρησιμοποιούνται ευρέως στα παραμετρικά μοντέλα και όχι τόσο στο ημι-παραμετρικό μοντέλο του Cox.

Υπόλοιπα Martingale

Τα υπόλοιπα Martingale ορίζονται βάσει των Cox-Snell υπολοίπων ως εξής:

$r_{Mi} = \delta_i - r_{CSi}$, όπου:

$$\delta_i = \begin{cases} 0, \text{ για αποκομμένη παρατήρηση} \\ 1, \text{ για μη αποκομμένη παρατήρηση} \end{cases}.$$

Τα υπόλοιπα Martingale παίρνουν τιμές μεταξύ του $-\infty$ και της μονάδας και χρησιμεύουν για την εύρεση της συναρτησιακής μορφής μιας μεταβλητής που πρόκειται να εισαχθεί στο μοντέλο του Cox.

Υπόλοιπα Deviance

Ορίζονται ως:

$$r_{Di} = \text{sgn}(r_{Mi})\left[-2\{r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\}\right]^{1/2}, \text{ με}$$

$$\text{sgn}(x) = \begin{cases} 1, \text{ για } x > 0 \\ -1, \text{ για } x < 0 \end{cases}$$

και χρησιμοποιούνται αντί των Martingale λόγω μεγαλύτερης ευχρηστίας στη γραφική τους ερμηνεία.

Υπόλοιπα Schoenfeld

Σε αντίθεση με όσα είδαμε μέχρι τώρα, τα υπόλοιπα Schoenfeld δεν ορίζονται για όλα τα άτομα του δείγματος αλλά μόνο για τις μη αποκομμένες παρατηρήσεις, ενώ δε δίνουν ακριβή τιμή για κάθε μονάδα αλλά ένα σύνολο τιμών. Ακόμα, για τον υπολογισμό τους δεν απαιτείται εκτίμηση της σωρευτικής συνάρτησης διακινδύνευσης.

Προτού προχωρήσουμε στον ορισμό των υπολοίπων υπενθυμίζουμε τα εξής: η πιθανότητα αποτυχίας της j-οστής υπό μελέτη μονάδας είναι η εξής:

$$p_j = \frac{e^{\tilde{\beta}'\tilde{x}_j}}{\sum_{i \in R_j} e^{\tilde{\beta}'\tilde{x}_i}}.$$

όπου R_j το σύνολο των ατόμων που είναι σε κίνδυνο τη χρονική στιγμή t_j

Το διάνυσμα των επεξηγηματικών μεταβλητών έχει αναμενόμενη τιμή:

$$E(\tilde{x} \mid R_j) = \sum_{k \in R_j} \tilde{x}_k p_k = \frac{\sum_{k \in R_j} \tilde{x}_k e^{\tilde{\beta}'\tilde{x}_k}}{\sum_{i \in R_j} e^{\tilde{\beta}'\tilde{x}_i}}.$$

Τελικά, τα υπόλοιπα Schoenfeld ορίζονται ως:

$\hat{r}_j = \tilde{x}_j - \hat{E}(\tilde{x} \mid R_j)$, όπου η εκτιμήτρια της αναμενόμενης μέσης τιμής προκύπτει με αντικατάσταση των εκτιμητριών $\hat{\tilde{\beta}}$.

2.4 Έλεγχοι υποθέσεων

Στο σημείο αυτό θα αναφερθούμε εν συντομία σε δύο δημοφιλείς τρόπους ελέγχου υποθέσεων. Σε πολλές περιπτώσεις οι έλεγχοι υποθέσεων μπορούν να χρησιμοποιηθούν για την ενίσχυση ή απόρριψη ενός ισχυρισμού, σχετικά με το επίπεδο σημαντικότητας μιας ή περισσοτέρων επεξηγηματικών μεταβλητών (Παύλου, 2006).

2.4.1 Έλεγχοι λόγου πιθανοφάνειας (Likelihood Ratio tests)

- Θεωρούμε τη μηδενική υπόθεση:

$H_0 : \beta_j = \beta_0 \neq 0$

με εναλλακτική την:

$H_1 : \beta_j = 0$

- Θεωρούμε τη στατιστική συνάρτηση:

$$L(\beta_0) = 2l(\hat{\beta}) - 2l(\beta_0)$$

όπου l($\hat{\beta}$) ο λογάριθμος πιθανοφάνειας για το εναλλακτικό μοντέλο, δηλαδή για το μοντέλο χωρίς τη συμμεταβλητή β_j και l(β_0) ο λογάριθμος της πιθανοφάνειας για το μοντέλο της μηδενικής υπόθεσης, δηλαδή για το μοντέλο όπου $\beta_j = \beta_0$.

- Στη συνέχεια ελέγχουμε αν η L(β_0) ακολουθεί κατανομή X_p^2 με βαθμό ελευθερίας p, ίσο με το σύνολο των συμμεταβλητών του προβλήματος που μελετάται.

2.4.2 Έλεγχος Wald

- Θεωρούμε την ελεγχοσυνάρτηση Wald για τον έλεγχο της μηδενικής υπόθεσης:

$H_0 : \beta = \beta_0$,

η οποία ορίζεται για κάθε μεταβλητή j ως εξής:

$$W = \{\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}\}^2$$

- Συγκρίνουμε με την κατανομή X_1^2 .

Τέλος, για την καλύτερη κατανόηση του αντικειμένου της ανάλυσης επιβίωσης παραθέτουμε ένα αριθμητικό παράδειγμα.

2.5 Παράδειγμα ανάλυσης επιβίωσης

Πρόκειται για μια έρευνα του πανεπιστημιακού νοσοκομείου της Μαδρίτης (Hospital Universitario Ramon y Cajal (Kleinbaum, 2006)) η οποία αφορά ένα τυπικό πρόβλημα επιβίωσης, όπου μελετάται ο χρόνος επιβίωσης των ατόμων του δείγματος με χρήση τριών συμμεταβλητών: το επίπεδο χοληστερίνης στο αίμα, το φύλο και το αν το υπό παρατήρηση άτομο είναι καπνιστής ή όχι.

Για κάθε επεξηγηματική μεταβλητή βρίσκουμε αρχικά μια εκτίμηση του συντελεστή (στήλη Coef), υπολογίζεται το τυπικό σφάλμα (standard error $(se(\hat{\beta}))$ και η ποσότητα $(\frac{\hat{\beta}}{se(\hat{\beta})})^2$ (ελεγχοσυνάρτηση Wald) που μέσω σύγκρισής της με την κατανομή X^2 δίνει την p-τιμή της προτελευταίας στήλης του πίνακά μας. Με τον τρόπο αυτό, η ελεγχοσυνάρτηση Wald επιτρέπει μια πρώτη εκτίμηση για τη στάθμη σημαντικότητας της εν λόγω επεξηγηματικής μεταβλητής.

ΠΙΝΑΚΑΣ 2.1

Μεταβλητή	Coef.	st.err.	$(\frac{\hat{\beta}}{se(\hat{\beta})})^2$	P	σ.σ
Χοληστερίνη	0.1091	0.0333	10.738	0.0017	p < 0.01
Φύλο	-0.0488	0.4716	0.011	0.9180	NO
Κάπνισμα	1.0638	0.3946	7.268	0.0091	p < 0.01

Από τον παραπάνω πίνακα προκύπτει ότι δεν υπάρχει άμεση σχέση ανάμεσα στο φύλο και στην επιβίωση του υπό μελέτη ατόμου, ενώ οι τιμές της πιθανότητας (στήλη p) δείχνουν ότι υπάρχει άμεση εξάρτηση της επιβίωσης από τα επίπεδα χοληστερίνης και το κάπνισμα, αφού οι δύο μεταβλητές είναι στατιστικά σημαντικές με στάθμη σημαντικότητας 99%.

Η παραπάνω διαπίστωση ενισχύεται αν παρουσιάσουμε τον πίνακα του σχετικού κινδύνου με ένα 95% διάστημα εμπιστοσύνης (δ.ε).

ΠΙΝΑΚΑΣ 2.3

Μεταβλητή	Σχετικός κίνδυνος	Inf δ.ε.	Sup δ.ε.
Χοληστερίνη	1.12	1.04	1.19
Φύλο	0.95	0.38	2.40
Κάπνισμα	2.90	1.34	6.28

Ο σχετικός κίνδυνος δίνεται από τη σχέση e^{β_i} .

Στο σημείο αυτό έχει ενδιαφέρον να παρατηρήσουμε πως αν ο σχετικός κίνδυνος παίρνει την τιμή 1, τότε η εκτίμηση του συντελεστή της αντίστοιχης επεξηγηματικής μεταβλητής είναι ίση με μηδέν και αυτό σημαίνει ότι αν πρόκειται για κάποια δίτιμη μεταβλητή όπως είναι στο παράδειγμά μας το κάπνισμα (καπνίζει: ναι=1 όχι=0), τότε ο κίνδυνος είναι ο ίδιος ανεξάρτητα από την τιμή της μεταβλητής.

Αντίστοιχα, ένας σχετικός κίνδυνος μεγαλύτερος από τη μονάδα υποδεικνύει μεγαλύτερο κίνδυνο για τα άτομα με το συγκεκριμένο χαρακτηριστικό. Έτσι, στην περίπτωσή μας ένας καπνιστής βρίσκεται σε 2.9 φορές μεγαλύτερο κίνδυνο από έναν μη καπνιστή όταν όλα τα άλλα στοιχεία είναι ίδια και για τα δύο άτομα.

ΚΕΦΑΛΑΙΟ 3

ΠΕΙΡΑΜΑΤΙΚΟΙ ΣΧΕΔΙΑΣΜΟΙ

3.1 Εισαγωγή

Με τη ραγδαία εξέλιξη των ερευνητικών και βιομηχανικών διαδικασιών, κυρίως κατά τη διάρκεια των τελευταίων δεκαετιών, η διεξαγωγή πειραμάτων, καθώς και η μελέτη των πειραματικών δεδομένων έχει μετατραπεί σε ένα έργο αρκετά σύνθετο. Σκοπός κάθε ερευνητικής διαδικασίας είναι η εξαγωγή του μεγαλύτερου δυνατού όγκου πληροφορίας, ούτως ώστε τα συμπεράσματα να είναι αξιόπιστα και να περιορίζεται στο ελάχιστο το κόστος και η διάρκεια διεξαγωγής του εκάστοτε πειράματος. Για να γίνει αυτό απαιτείται η χρήση διαδικασιών για την οργάνωση και στη συνέχεια την ανάλυση των πειραματικών δεδομένων.

Για την οργάνωση των πειραματικών δεδομένων με χρήση παραγοντικών σχεδιασμών ο εκάστοτε ερευνητής επιλέγει εκ των προτέρων κάποιους παράγοντες που θεωρεί εν δυνάμει σημαντικούς. Σημειώνεται ότι με τον όρο παράγοντας ορίζουμε μια ποιοτική μεταβλητή η οποία δέχεται τιμές από ένα πεπερασμένο σύνολο. Οι τιμές αυτές ονομάζονται στάθμες *(levels)*, ενώ κάθε δυνατός συνδυασμός παραγόντων λέγεται αγωγή *(treatment)*. Προφανώς, όταν μελετάται η επίδραση μόνο ενός παράγοντα οι αγωγές ταυτίζονται με τις στάθμες.

Ενδεικτικά αναφέρουμε κάποια είδη παραγοντικών σχεδιασμών με δύο στάθμες , όπως είναι ο 2^k σχεδιασμός που λαμβάνει υπόψη k παράγοντες και απαιτεί 2^k επαναλήψεις, ο 2^k κλασματικός σχεδιασμός που λαμβάνει υπόψη τους ίδιους παράγοντες και απαιτεί μικρότερο αριθμό επαναλήψεων, ο 2^k σχεδιασμός αναμεμειγμένος σε blocks, καθώς και ο σχεδιασμός ορθογώνιων αντιθέσεων.

Στη συνέχεια με την ανάλυση δεδομένων είναι δυνατόν να ερευνηθεί η συμβολή μιας πηγής μεταβολής (παράγοντας) στην ολική μεταβολή των δεδομένων (απόκριση).

3.1.1 Χρησιμότητα πειραματικών σχεδιασμών

Ο στατιστικός κλάδος του παραγοντικού σχεδιασμού πειραμάτων προσφέρει, σήμερα, μια αρκετά μεγάλη ποικιλία διαδικασιών για το σχεδιασμό πειραμάτων, με τελικό στόχο τον έλεγχο της ποιότητας των προϊόντων που διατίθενται στην αγορά, αλλά και την εξαγωγή στατιστικών συμπερασμάτων στον τομέα της υγείας, της γεωργίας και της βιομηχανίας. Με τη χρήση των διαδικασιών αυτών, ο ερευνητής εντοπίζει τελικά τους στατιστικά σημαντικότερους παράγοντες που επηρεάζουν το πρόβλημα με το οποίο καταπιάνεται κι έτσι μειώνεται σημαντικά η υποκειμενικότητα που περικλείει ο ανθρώπινος παράγοντας στη λήψη αποφάσεων, αφού οι σχέσεις μεταξύ των μεταβλητών ενός πειράματος μοντελοποιούνται και ερμηνεύονται πλήρως με μαθηματικούς όρους.

3.1.2 Ο πειραματικός σχεδιασμός του Taguchi

Τα τελευταία χρόνια, ο σχεδιασμός του ιάπωνα G.Taguchi *(robust parameter design)* έχει, επίσης, αποδειχθεί ιδιαίτερα χρήσιμος για τη βελτίωση της ποιότητας των παραγόμενων τεχνολογικών προϊόντων (Sreenivas et al., 2008). Αυτό συμβαίνει διότι σε τέτοιου είδους εφαρμογές η απόκριση *(response)* εξαρτάται συχνά από παράγοντες σήματος, για τους οποίους ο θόρυβος *(noise)* παίζει καθοριστικό ρόλο. Ο σχεδιασμός του Taguchi αποδεικνύεται αποτελεσματικός στη περίπτωση αυτή, καθώς μέσω αυτού η παραγωγική διαδικασία γίνεται λιγότερο ευαίσθητη στο θόρυβο κι έτσι βελτιώνεται ραγδαία τόσο η ποιότητα του προϊόντος, όσο και η αποτελεσματικότητα της παραγωγικής διαδικασίας.

3.2 Είδη πειραματικών σχεδιασμών

Όπως ήδη έχουμε αναφέρει τα είδη των πειραματικών σχεδιασμών ποικίλουν. Στην ενότητα αυτή θα αναφερθούμε εκτενέστερα σε κάποια από αυτά, ενώ για την πλήρη κατανόησή τους θα παρατεθούν σχετικά παραδείγματα.

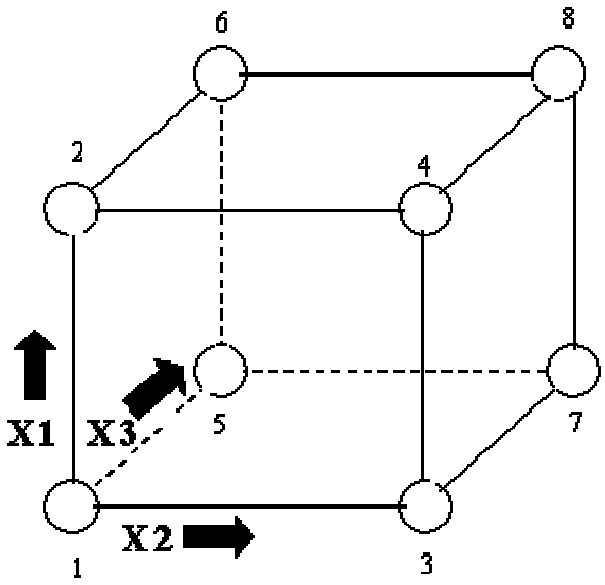
Σημειώνεται ότι στην περίπτωση αποκομμένων παρατηρήσεων δε διαφοροποιείται η δομή των πειραματικών σχεδιασμών. Αυτό που διαφοροποιείται είναι η μέθοδος ανάλυσής τους. Στις μεθόδους αυτές θα αναφερθούμε εκτενώς σε επόμενο κεφάλαιο.

3.2.1 Ο 2^k παραγοντικός σχεδιασμός *(full factorial design)*

Ο 2^k πειραματικός σχεδιασμός αποτελείται από τους 2^k δυνατούς συνδυασμούς των k παραγόντων, οι οποίοι δέχονται τιμές σε δύο στάθμες. Κάθε συνδυασμός τιμών παραγόντων ονομάζεται αγωγή *(treatment)* ή επανάληψη *(run)*. Αυτό σημαίνει πως ένας 2^k πειραματικός σχεδιασμός αποτελείται από 2^k επαναλήψεις. Συχνά οι δύο στάθμες συμβολίζονται με (-) και (+) ή με -1 και 1 αντίστοιχα.

Ας θεωρήσουμε ένα 2^3 παραγοντικό σχεδιασμό με παράγοντες X1, X2, X3. Τότε μπορούμε να παραστήσουμε γραφικά το σχεδιασμό αυτό, όπως φαίνεται στο γράφημα 3.1. Οι ακμές του κύβου δείχνουν την αύξηση των παραγόντων και οι κορυφές αντιστοιχούν στις $2^3 = 8$ επαναλήψεις.

Σημειώνεται ότι, με τον όρο παραγοντικό σχεδιασμό εννοούμε ότι σε κάθε πλήρη δοκιμή του πειράματος εξετάζονται όλοι οι δυνατοί συνδυασμοί των σταθμών των παραγόντων.



Γράφημα 3.1

Σε μορφή πίνακα ο σχεδιασμός δίνεται ως εξής:

ΠΙΝΑΚΑΣ 3.1

Α/Α Επανάληψης	Παράγοντες		
	X1	**X2**	**X3**
1	-1	-1	-1
2	1	-1	-1
3	-1	1	-1
4	1	1	-1
5	-1	-1	1
6	1	-1	1
7	-1	1	1
8	1	1	1

Η σειρά αυτή των επαναλήψεων δεν είναι τυχαία, αλλά αποτελεί την τυπική μορφή επαναλήψεων. Για παράδειγμα η κορυφή 1 αναπαριστά την κατάσταση όπου και οι τρεις παράγοντες βρίσκονται στη χαμηλότερη στάθμη τους, ενώ η κορυφή 2 την κατάσταση όπου ο παράγοντας X1 βρίσκεται στην υψηλότερη στάθμη του, ενώ οι παράγοντες X2, X3 στη χαμηλότερη.

3.2.1.1 Παράδειγμα ενός 2^4 παραγοντικού σχεδιασμού (Hamada-Wu, 2000)

Στη διαδικασία παραγωγής ηλεκτρονικών συσκευών ένα πρωταρχικό βήμα αποτελεί η δημιουργία λεπτών κρυσταλλικών στρωμάτων στις πλάκες πυριτίου. Για να γίνει αυτό οι πλάκες τοποθετούνται σε έναν περιστρεφόμενο κύλινδρο και ψεκάζονται με χημικές ουσίες, κάτω από ορισμένη θερμοκρασία και για ορισμένο χρονικό διάστημα. Για τη βελτιστοποίηση της διαδικασίας πραγματοποιήθηκε πείραμα στο οποίο συμπεριλήφθηκαν οι εξής τέσσερις παράγοντες: η μέθοδος περιστροφής του κυλίνδρου, η θέση του ψεκαστήρα, η θερμοκρασία και ο χρόνος.

Ο σχεδιαστικός πίνακας φαίνεται στον πίνακα που ακολουθεί:

ΠΙΝΑΚΑΣ 3.2

Α/Α Αγωγής	Παράγοντες			
	Μέθοδος περιστροφής	**Θέση ψεκαστήρα**	**Θερμοκρασία (^oF)**	**Χρονική διάρκεια**
1	εναλλασσόμενη	6	1210	Μικρή
2	εναλλασσόμενη	2	1210	Μικρή
3	εναλλασσόμενη	2	1220	Μεγάλη
4	εναλλασσόμενη	6	1220	Μικρή
5	συνεχής	2	1220	Μεγάλη
6	συνεχής	2	1210	Μικρή
7	εναλλασσόμενη	2	1210	Μεγάλη
8	συνεχής	6	1220	Μικρή
9	συνεχής	2	1210	Μεγάλη
10	συνεχής	6	1220	Μεγάλη
11	εναλλασσόμενη	2	1220	Μικρή
12	συνεχής	2	1220	Μικρή
13	εναλλασσόμενη	6	1210	Μεγάλη
14	εναλλασσόμενη	6	1220	Μεγάλη
15	συνεχής	6	1210	Μεγάλη
16	συνεχής	6	1210	Μικρή

Η σειρά των επαναλήψεων στο σχεδιασμό είναι τυχαία. Σημειώνεται επίσης ότι η επιλογή των τεσσάρων παραγόντων έγινε σύμφωνα με την κρίση του ερευνητή. Κάλλιστα θα μπορούσε να έχει επιλέξει κάποιους από αυτούς ή ακόμα να είχε προσθέσει άλλους, όπως η υγρασία ή η διαθεσιμότητα οξυγόνου στο δωμάτιο.

Στη συνέχεια, δίνεται μια κωδικοποίηση των παραγόντων και των σταθμών, καθώς και ο σχεδιαστικός πίνακας που προκύπτει μετά από την κωδικοποίηση αυτή.

ΠΙΝΑΚΑΣ 3.3

Παράγοντας	Κωδικοποίηση
Μέθοδος περιστροφής	A
Θέση ψεκαστήρα	B
Θερμοκρασία (^{o}F)	C
Χρονική διάρκεια	D

ΠΙΝΑΚΑΣ 3.4

Παράγοντας	(+)	(-)
Μέθοδος περιστροφής	συνεχής	εναλλασσόμενη
Θέση ψεκαστήρα	6	2
Θερμοκρασία (^{o}F)	1220	1210
Χρονική διάρκεια	Μεγάλη	Μικρή

Ο καθορισμός των σταθμών ως υψηλή (+) και χαμηλή (-) είναι προφανές ότι πραγματοποιήθηκε εμπειρικά. Η μεγάλη χρονική διάρκεια ψεκασμού, δηλαδή, θεωρήθηκε υψηλή στάθμη, ενώ η μικρή θεωρήθηκε χαμηλή.

Ο τελικός πίνακας σχεδιασμού για τα πειραματικά δεδομένα του παραδείγματος 3.2.1.1 είναι ο εξής:

ΠΙΝΑΚΑΣ 3.5

Α/Α Επανάληψης	**Παράγοντες**			
	A	**B**	**C**	**D**
1	-	+	-	-
2	-	-	-	-
3	-	-	+	+
4	-	+	+	-

5	+	-	+	+
6	+	-	-	-
7	-	-	-	+
8	+	+	+	-
9	+	-	-	+
10	+	+	+	+
11	-	-	+	-
12	+	-	+	-
13	-	+	-	+
14	-	+	+	+
15	+	+	-	+
16	+	+	-	-

Από το παραπάνω παράδειγμα γίνεται εμφανής η χρησιμότητα της κωδικοποίησης των παραγόντων για την απλούστευση του σχεδιασμού, καθώς και ο τρόπος με τον οποίο ο παραγοντικός σχεδιασμός οργανώνει τα πειραματικά δεδομένα.

Γενικά, για τη διεξαγωγή του εκάστοτε πειράματος δημιουργείται αρχικά ένας πίνακας σχεδιασμού στον οποίο φαίνονται τα πειραματικά δεδομένα στη φυσική τους μορφή, για να αποφεύγονται τυχόν συγχύσεις. Στη συνέχεια κωδικοποιούνται, για να διευκολυνθεί η στατιστική ανάλυση αυτών.

3.2.2 Ο 2^k κλασματικός παραγοντικός σχεδιασμός *(fractional factorial design)*

Όπως είδαμε στην ενότητα 3.2.1, ο 2^k παραγοντικός σχεδιασμός απαιτεί 2^k επαναλήψεις. Κατ' αντιστοιχία ο 3^k σχεδιασμός, δηλαδή ο σχεδιασμός όπου οι k παράγοντες δέχονται τιμές σε τρεις στάθμες, απαιτεί ακριβώς 3^k επαναλήψεις. Επομένως για μεγάλο αριθμό παραγόντων, ο αριθμός των επαναλήψεων που απαιτείται είναι τεράστιος. (Box, 2005)

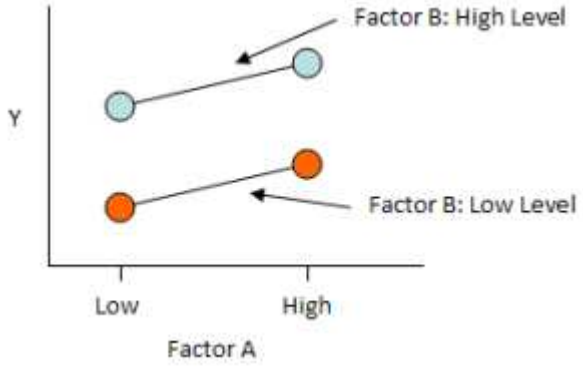
Για παράδειγμα, σε ένα σχεδιασμό με 8 παράγοντες, δύο σταθμών ο καθένας, οι επαναλήψεις που απαιτούνται είναι 2^8=256. Ενώ, αν ο ένας παράγοντας ήταν δύο σταθμών και οι υπόλοιποι επτά τριών θα απαιτούνταν $2^1 * 3^7 = 4374$ επαναλήψεις.

Από τα παραπάνω γίνεται εμφανής η επιτακτική ανάγκη για περιορισμό των επαναλήψεων σε προβλήματα με μεγάλο αριθμό παραγόντων, με σκοπό των υπολογισμό των σημαντικότερων κυρίων επιδράσεων και αλληλεπιδράσεων αυτών.

Ως επίδραση ενός παράγοντα ορίζεται η μεταβολή που προκαλείται στην απόκριση *(response)* από τη μεταβολή στο επίπεδο του παράγοντα. Πρόκειται για τη λεγόμενη κύρια επίδραση *(main effect)* των βασικών παραγόντων, ενώ ως αλληλεπίδραση παραγόντων ορίζουμε τη μεταβολή που προκαλείται στην απόκριση από την ταυτόχρονη μεταβολή στη στάθμη των παραγόντων.

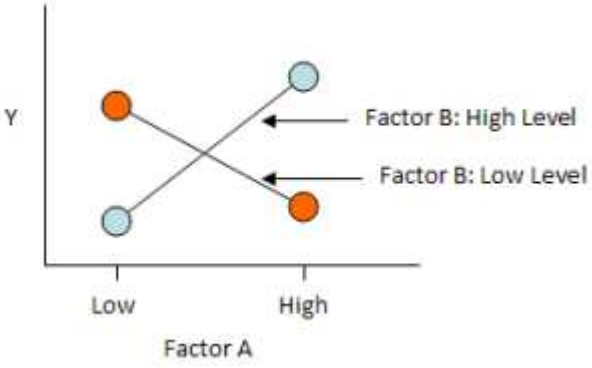
Όταν οι παράγοντες είναι ταξινομημένοι σε ένα παραγοντικό σχεδιασμό τότε χωρίζονται σε εγκάρσιους και διασταυρωμένους *(crossed)*.

Στο Γράφημα 3.2 βλέπουμε ένα παράδειγμα παραγόντων που δεν αλληλεπιδρούν μεταξύ τους:



Γράφημα 3.2

Ενώ στο Γράφημα 3.3 βλέπουμε ένα παράδειγμα πειράματος με αλληλεπίδραση των παραγόντων Α και Β:



Γράφημα 3.3

Η αρχική ιδέα για τον περιορισμό των επαναλήψεων ενός πειράματος ήταν, λοιπόν, ο εκάστοτε ερευνητής να επιλέγει με προσοχή τους συνδυασμούς παραγόντων των οποίων οι αλληλεπιδράσεις τον ενδιαφέρουν περισσότερο και να αγνοεί ορισμένες αλληλεπιδράσεις υψηλής τάξης τις οποίες θεωρεί αμελητέες. Με τον τρόπο αυτό εκτελείται μόνο ένα κλάσμα του πλήρους παραγοντικού σχεδιασμού.

Η διαδικασία αυτή ονομάστηκε κλασματικός παραγοντικός σχεδιασμός πειραμάτων και χρησιμοποιείται σήμερα ευρέως στις εφαρμογές λόγω της υψηλής απόδοσης και οικονομίας που προσφέρει. Συμβολίζεται ως 2^{k-p} σχεδιασμός με $p < k$, όπου λαμβάνονται υπόψη k παράγοντες και ο αριθμός επαναλήψεων είναι 2^{k-p}.

Για παράδειγμα, στον 2^3 σχεδιασμό υπάρχουν δύο κύρια $\frac{1}{2}$-κλάσματα:

- Το κύριο κλάσμα με γεννήτορα I = ABC
- Το κύριο κλάσμα με γεννήτορα I = - ABC

Αυτό σημαίνει ότι η στήλη της αλληλεπίδρασης ABC λαμβάνει μόνο την υψηλή ή τη χαμηλή στάθμη. Έτσι, I = ABC αν λαμβάνει μόνο την υψηλή και I = - ABC αν λαμβάνει αντίστοιχα μόνο τη χαμηλή.

Στο παράδειγμα που ακολουθεί βλέπουμε ένα παράδειγμα $\frac{1}{2}$-κλάσματος ενός 2^3 παραγοντικού σχεδιασμού με γεννήτορα I = ABC, με $2^2 = 4$ επαναλήψεις (2^{3-1} κλασματικός σχεδιασμός).

Παράδειγμα 3.1

Α/Α	**Συνδυασμός αγωγών**	**Παράγοντες**				**Αλληλεπιδράσεις**			
		I	**A**	**B**	**C**	**AB**	**AC**	**BC**	**ABC**
1	A	1	1	-1	-1	-1	-1	1	1
2	B	1	-1	1	-1	-1	1	-1	1
3	C	1	-1	-1	1	1	-1	-1	1
4	ABC	1	1	1	1	1	1	1	1

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.

Στο παράδειγμα αυτό θα μπορούσε κατά την ανάλυση να δημιουργηθεί σύγχυση ανάμεσα στην κύρια επίδραση του παράγοντα Α και την αλληλεπίδραση των ΒC αφού οι δύο στήλες ταυτίζονται, καθώς επίσης και του Β με την ΑC αλλά και του C με την ΑΒ, αντίστοιχα. Δύο ή περισσότερες επιδράσεις που έχουν αυτή την ιδιότητα ονομάζονται ταυτόσημες επιδράσεις *(aliases)*. Αυτό δε σημαίνει ότι οι επιδράσεις αυτές ταυτίζονται, αλλά ότι ο σχεδιασμός αδυνατεί να τις διαχωρίσει. Επίσης, η αλληλεπίδραση των ABC είναι ίδια σε κάθε επανάληψη. Αυτό σημαίνει ότι δεν προσθέτει καμία παραπάνω πληροφορία για τον ερευνητή.

3.2.2.1 Αναλυτική τάξη σχεδιασμών *(resolution)*

Για την αναλυτική τάξη ενός σχεδιασμού χρησιμοποιείται συνήθως ως υποσημείωση κάποιο σύμβολο της ρωμαϊκής αρίθμησης. Στο Παράδειγμα 3.1 έχουμε έναν 2^{3-1}_{III} σχεδιασμό, δηλαδή έναν σχεδιασμό αναλυτικής τάξης ΙΙΙ. Ειδικότερα αναφέρουμε (Κουκουβίνος, 2005) :

- **Σχεδιασμοί αναλυτικής τάξης ΙΙΙ**

Καμία κύρια επίδραση δεν είναι ταυτόσημη με άλλη, αλλά οι κύριες επιδράσεις είναι ταυτόσημες με αλληλεπιδράσεις δύο παραγόντων. Επίσης, αλληλεπιδράσεις δύο παραγόντων, μπορεί να είναι ταυτόσημες μεταξύ τους.

Παράδειγμα σχεδιασμού αναλυτικής τάξης ΙΙΙ αποτελεί ένας 2^{3-1} σχεδιασμός με I=ABC και συμβολίζεται 2^{3-1}_{III}

- **Σχεδιασμοί αναλυτικής τάξης IV**

Καμία κύρια επίδραση δεν είναι ταυτόσημη με άλλη ή με οποιαδήποτε αλληλεπίδραση δύο παραγόντων, αλλά αλληλεπιδράσεις δύο παραγόντων είναι ταυτόσημες με άλλες αλληλεπιδράσεις.

Παράδειγμα σχεδιασμού αναλυτικής τάξης IV αποτελεί ένας 2^{4-1} σχεδιασμός με I=ABCD και συμβολίζεται 2^{4-1}_{IV}.

- **Σχεδιασμοί αναλυτικής τάξης V**

Καμία κύρια επίδραση ή αλληλεπίδραση δύο παραγόντων δεν είναι ταυτόσημη με άλλη κύρια επίδραση ή αλληλεπίδραση δύο παραγόντων, αλλά αλληλεπιδράσεις δύο παραγόντων είναι ταυτόσημες με αλληλεπιδράσεις τριών.

Παράδειγμα σχεδιασμού αναλυτικής τάξης V αποτελεί ένας 2^{5-1} σχεδιασμός με I=ABCDE και συμβολίζεται 2_V^{5-1}.

Σημειώνεται ότι, η τυπική διάταξη ενός σχεδιασμού αποτελεί έναν ορθογώνιο πίνακα. Οι ιδιότητες ενός ορθογώνιου πίνακα σχεδιασμού, όταν η χαμηλή και η υψηλή στάθμη των παραγόντων έχουν κωδικοποιηθεί με -1 και 1, αντίστοιχα, είναι οι εξής:

- Οι στήλες είναι ανά δύο ορθογώνιες,. Αυτό σημαίνει ότι το συνολικό άθροισμα των γινομένων των στοιχείων κάθε στήλης για κάθε επανάληψη είναι μηδέν.
- Το άθροισμα των στοιχείων κάθε στήλης είναι μηδέν, εκτός από τη στήλη που συνδυάζει όλους τους παράγοντες.

Η ιδιότητα της ορθογωνιότητας είναι σημαντική, καθώς μειώνει την πιθανότητας εμφάνισης ταυτόσημων επιδράσεων *(aliases)*.

3.2.3 Σχεδιασμοί ορθογώνιων πινάκων *(orthogonal arrays)*

Οι σχεδιασμοί ορθογώνιων πινάκων αποτελούν κλασματικούς σχεδιασμούς *(highly fractionated factorial designs)* και αποδεικνύονται αρκετά χρήσιμοι σε εφαρμογές που εμπεριέχουν αποκομμένα δεδομένα και για τη στατιστική τους ανάλυση χρησιμοποιούνται οι μέθοδοι Taguchi που θα δούμε αναλυτικά στο επόμενο κεφάλαιο. Για τον ορισμό ενός ορθογώνιου πίνακα σχεδιασμού πρέπει πρώτα να ληφθούν υπόψη τα εξής:

- Ο αριθμός των παραγόντων που μελετώνται
- Οι στάθμες κάθε παράγοντα
- Οι αλληλεπιδράσεις παραγόντων για τις οποίες απαιτείται εκτίμηση
- Οι δυσκολίες που θα προκύψουν στη διεξαγωγή του πειράματος

Είδαμε ότι στον 2^k παραγοντικό ή κλασματικό παραγοντικό σχεδιασμό μπορούμε να εκτιμήσουμε την αλληλεπίδραση παραγόντων. Σε κάποιες, όμως, περιπτώσεις δύναται να προκληθεί σύγχυση. Στο σχεδιασμό ορθογώνιων πινάκων ο οποίος αποτελεί εξαιρετικά τμηματικό σχεδιασμό η πληροφορία που παράγεται είναι συνάρτηση της φύσης της σύγχυσης που προκαλούν άλλου τύπου σχεδιασμοί, όπως επίσης και ανθρώπινων υποθέσεων σχετικά με την αναμενόμενη λειτουργία του υπό μελέτη συστήματος. Αυτό, με απλά λόγια, σημαίνει πως η αλληλεπίδραση δύο παραγόντων θεωρείται αμελητέα εκτός και αν ο ερευνητής τη θεωρήσει a-priori σημαντική και άξια εκτίμησης. Για να αρχίσει να γίνεται πιο κατανοητή η χρήση του σχεδιασμού παραθέτουμε το επόμενο εισαγωγικό παράδειγμα:

Παράδειγμα 3.2

Σε ένα πείραμα με σκοπό τη βέλτιστη κατασκευή ενός τούνελ στην Ιαπωνία του 1959 που πραγματοποιήθηκε από την εθνική εταιρία σιδηροδρόμων μελετήθηκαν οι εξής 9 παράγοντες δύο σταθμών (Blunt et al., 2002):

Παράγοντες	**Χαμηλή στάθμη (-)**	**Υψηλή στάθμη (+)**
A-Είδος εργαλείου συγκόλλησης	J100	B17
B-Διάρκεια ψησίματος ράβδων	Χωρίς ψήσιμο	Μια ημέρα
C-Υλικό προς συγκόλληση	SS41	SB35
D-Πάχος υλικού προς συγκόλληση	8mm	12mm
E-Γωνία συγκόλλησης	70^o	60^o
F-Άνοιγμα συγκολλητικής συσκευής	1.5mm	3mm
G-Ένταση ρεύματος	150A	130A
H-Μέθοδος συγκόλλησης	εναλλασσόμενη	σταθερή
I-Προθέρμανση	Χωρίς προθέρμανση	Προθέρμανση στους $150^o C$

Κατά την κρίση του ερευνητή εκτός από τους 9 παράγοντες του προβλήματος, μπορεί να επιλέξει ένα πλήθος αλληλεπιδράσεων που θεωρεί μη αμελητέες. Στη συνέχεια, βλέπουμε τον πίνακα σχεδιασμού για τους 9 παράγοντες και τις 4 αλληλεπιδράσεις που επιλέχθηκαν.

ΠΙΝΑΚΑΣ 3.6

A/A	A	G	AG	H	AH	GH	B	D	E	F	I	e	e	C	AC
1	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	+	-	+	+	+	+	+	+	+	+
3	-	-	-	+	+	+	+	-	-	-	-	+	+	+	+
4	-	-	-	+	+	+	+	+	+	+	+	-	-	-	-
5	-	+	+	-	-	+	+	-	-	+	+	-	+	+	+
6	-	+	+	-	-	+	+	+	+	-	-	+	-	-	-
7	-	+	+	+	+	-	-	-	-	+	+	+	-	-	-
8	-	+	+	+	+	-	-	+	+	-	-	-	+	+	+
9	+	-	+	-	+	-	+	-	+	-	+	-	-	-	+
10	+	-	+	-	+	-	+	+	-	+	-	+	+	+	-
11	+	-	+	+	-	+	-	-	+	-	+	+	+	+	-
12	+	-	+	+	-	+	-	+	-	+	-	-	-	-	+
13	+	+	-	-	+	+	-	-	+	+	-	-	+	+	-
14	+	+	-	-	+	+	-	+	-	-	+	+	-	-	+
15	+	+	-	+	-	-	+	-	+	+	-	+	-	-	+
16	+	+	-	+	-	-	+	+	-	-	+	-	+	+	-

Οι στήλες e πιθανόν να χρησιμοποιηθούν κατά την ανάλυση για την εκτίμηση των σφαλμάτων.

Ο παραπάνω σχεδιασμός συμβολίζεται ως $L_{16}(2^{15})$ ή εναλλακτικά συναντάται στη βιβλιογραφία ως $\mathrm{OA}(16, 2^{15})$ και μπορεί ισοδύναμα να μελετηθεί ως ένας 2^{9-5} κλασματικός παραγοντικός σχεδιασμός, αφού είναι προφανές ότι δεν πρόκειται για πλήρη σχεδιασμό.

Σημειώνεται ότι πλήρης παραγοντικός σχεδιασμός ονομάζεται ο σχεδιασμός όπου σε κάθε συνδυασμό των σταθμών του λαμβάνουμε μη μηδενικό αριθμό παρατηρήσεων. Γενικά σπάνια αρκεί μια παρατήρηση για κάθε συνδυασμό σταθμών διότι έτσι καθίσταται αδύνατος ο υπολογισμός σφαλμάτων. Οι πλήρεις σχεδιασμοί είναι ορθογώνιοι. Οι ορθογώνιοι όμως δεν είναι κατ' ανάγκην και πλήρεις όπως είδαμε στο παραπάνω παράδειγμα.

3.2.4 Σχεδιασμοί λατινικών τετραγώνων

(latin square designs)

Ένα Λατινικό Τετράγωνο (Κουκουβίνος, 2005) τάξης n είναι ένας n x n πίνακας με ακριβώς n διαφορετικά σύμβολα, όπου κάθε σύμβολο εμφανίζεται μία φορά σε κάθε γραμμή και μία φορά σε κάθε στήλη. Τα Λατινικά Τετράγωνα είναι σχεδιασμοί 3 παραγόντων με n στάθμες ο καθένας, όπου δε λαμβάνουμε παρατηρήσεις για κάθε συνδυασμό των σταθμών των τριών παραγόντων, αλλά για κάθε συνδυασμό σταθμών οποιονδήποτε 2 εξ' αυτών.

Λέγεται ότι είναι σε κανονική μορφή ή κανονικοποιημένο αν η πρώτη του γραμμή και η πρώτη του στήλη είναι σε φυσική σειρά. Μπορούμε εύκολα να κανονικοποιήσουμε ένα Λατινικό Τετράγωνο με αντιμεταθέσεις των γραμμών και των στηλών του. Στη συνέχεια βλέπουμε ένα Λατινικό Τετράγωνο 5 x 5 πρώτα σε τυχαία μορφή και στη συνέχεια κανονικοποιημένο.

Παράδειγμα 3.3

Μη κανονικοποιημένο λατινικό τετράγωνο 5 x 5

A	**B**	**C**	**D**	**E**
D	**E**	**A**	**B**	**C**
B	**C**	**D**	**E**	**A**
E	**A**	**B**	**C**	**D**
C	**D**	**E**	**A**	**B**

Λατινικό τετράγωνο 5 x 5 σε κανονική μορφή

A	**B**	**C**	**D**	**E**
B	**C**	**D**	**E**	**A**
C	**D**	**E**	**A**	**B**
D	**E**	**A**	**B**	**C**
E	**A**	**B**	**C**	**D**

Για κάθε Λατινικό Τετράγωνο n τάξης υπάρχει τουλάχιστον ένα Λατινικό Τετράγωνο που σχηματίζεται με κυκλική εναλλαγή των γραμμάτων. Δυο Λατινικά Τετράγωνα τέτοια ώστε σε όλες τις θέσεις όπου στο πρώτο εμφανίζεται ένα συγκεκριμένο σύμβολο, στο άλλο να εμφανίζονται όλα τα άλλα σύμβολα εκτός αυτού ονομάζονται ορθογώνια. Γενικά υπάρχουν το πολύ n-1 ανά δύο ορθογώνια Λατινικά Τετράγωνα τάξης $n > 1$. Υπάρχουν όμως Λατινικά Τετράγωνα για τα οποία δεν υπάρχουν ορθογώνια όπως το παρακάτω:

Παράδειγμα 3.4

5	**6**	**1**	**4**	**3**	**2**
6	**5**	**2**	**1**	**4**	**3**
1	**2**	**3**	**5**	**6**	**4**
2	**3**	**4**	**6**	**1**	**5**
4	**1**	**5**	**3**	**2**	**6**
3	**4**	**6**	**2**	**5**	**1**

Στο Λατινικό Τετράγωνο του παραδείγματος 3.4 αντί για λατινικά γράμματα από το A έως το F, χρησιμοποιήθηκαν ισοδύναμα αριθμοί από το 1 έως το 6.

Γενικά στο σχεδιασμό Λατινικών Τετραγώνων οι γραμμές εκφράζουν τον ένα παράγοντα, οι στήλες τον δεύτερο και τα γράμματα τις αγωγές.

Παράδειγμα 3.5

Παράδειγμα εφαρμογής σχεδιασμού Λατινικών Τετραγώνων
(Styan et al., 2008)

Ο Palluel το 1778 χρησιμοποίησε το σχεδιασμό Λατινικών Τετραγώνων για να επιλέγει κάθε μέρα 4 πρόβατα προς σφαγή τα οποία όμως να είναι διαφορετικής ράτσας και να ακολουθούν διαφορετική δίαιτα το καθένα.

Στη συνέχεια παρατίθεται το περί ού ο λόγος Λατινικό Τετράγωνο, οι γραμμές του οποίου αντιστοιχούν σε διαφορετικές ράτσες και οι στήλες σε διαφορετικές δίαιτες.

Λατινικό Τετράγωνο του Palluel

A	**B**	**C**	**D**
D	**A**	**B**	**C**
C	**D**	**A**	**B**
B	**C**	**D**	**A**

3.2.5 Παραγοντικοί σχεδιασμοί αναμεμειγμένοι σε ομάδες *(blocks)*

Η ανάμειξη *(confounding)* είναι μια τεχνική σχεδιασμού για να τακτοποιήσουμε ένα πλήρες ή κλασματικό παραγοντικό πείραμα σε blocks όπου το μέγεθος κάθε block είναι μικρότερο από τον αριθμό των συνδυασμών αγωγών σε μια επανάληψη. Με την τεχνική αυτή αποκτάται πληροφορία σχετικά με ορισμένες επιδράσεις αγωγών (κυρίως υψηλής τάξης).

Για να γίνει κατανοητή η τεχνική της ανάμειξης σχεδιασμών σε blocks, ας δούμε ένα παράδειγμα ενός 2^4 πλήρους παραγοντικού σχεδιασμού αναμεμειγμένου σε δύο blocks:

Παράδειγμα 3.6

Όπως ήδη έχουμε δει στον πλήρη 2^4 παραγοντικό σχεδιασμό απαιτούνται $2^4=16$ επαναλήψεις. Αν λοιπόν χωρίσουμε τις επαναλήψεις σε δύο ομάδες κάθε ομάδα θα αποτελείται από 8 επαναλήψεις. Έστω οι παράγοντες A,B,C,D. Τότε, ένας πιθανός προσδιορισμός των δύο blocks με αναμεμειγμένη την αλληλεπίδραση υψηλότερης τάξης φαίνεται στον Πίνακα 3.7:

ΠΙΝΑΚΑΣ 3.7

Block 1		Block 2
I		A
AB		B
AC		C
BC		D
AD		ABC
BD		BCD
CD		ACD
ABCD		ABD

Παρατηρούμε ότι στο πρώτο block η αλληλεπίδραση ABCD όλων των παραγόντων είναι πάντα θετική, ενώ στο δεύτερο πάντα αρνητική. Αυτό προκύπτει από το γεγονός ότι η αλληλεπίδραση ABCD υψηλότερης τάξης είναι αναμεμειγμένη στα δύο blocks.

Ένα βασικό πλεονέκτημα του σχεδιασμού αναμεμειγμένου σε blocks είναι ότι ο ερευνητής μπορεί να επιλέξει τη χρονική στιγμή που θα εκτελέσει τις επαναλήψεις κάθε ομάδας, χωρίς κατ' ανάγκην να πρέπει να πραγματοποιηθούν οι επαναλήψεις των δύο ομάδων ταυτόχρονα. Αυτό είναι συχνά ιδιαίτερα χρήσιμο ειδικά όταν το πείραμα απαιτεί μεγάλο χρονικό διάστημα για να διεξαχθεί ολόκληρο.

Με την ίδια ακριβώς λογική λειτουργεί η ανάμειξη και στους κλασματικούς παραγοντικούς σχεδιασμούς.

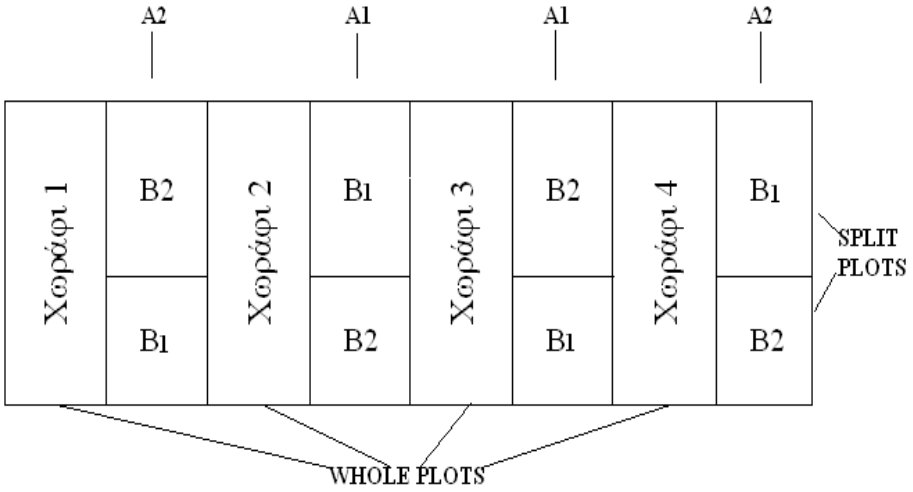
3.2.6 Ο παραγοντικός σχεδιασμός split-plot

Ο παραγοντικός σχεδιασμός διαιρεμένων τεμαχίων *(split-plot factorial design)* αναπτύχθηκε από τον Fischer το 1925 και αρχικά χρησιμοποιήθηκε κατά κόρον σε πειράματα που αφορούσαν γεωργικές καλλιέργειες (Fischer, 1925). Ουσιαστικά πρόκειται για σχεδιασμό αναμεμειγμένο σε ομάδες στον οποίο τα ίδια τα blocks λειτουργούν σαν πειραματικές μονάδες για κάποιους από τους παράγοντες.

Αυτό σημαίνει ότι τελικά δημιουργούνται δύο επιπέδων πειραματικές μονάδες. Οι κύριες μονάδες *(whole plots)* αποτελούνται από ολόκληρα τα blocks, ενώ οι μονάδες που εμπεριέχονται στα blocks και επίσης συμπεριφέρονται ως πειραματικά αντικείμενα, θεωρούνται δευτερεύουσες *(subplots ή split plots)*. Σε

αντιστοιχία με τα δύο επίπεδα αντικειμένων υπό μελέτη, υπάρχουν και δύο επίπεδα τυχαιοποίησης για τη σειρά των αγωγών.

Ας θεωρήσουμε μια μελέτη για την επίδραση δύο μεθόδων άρδευσης (Παράγοντας Α) και δύο διαφορετικών ειδών λιπασμάτων (Παράγοντας Β) στην παραγωγική διαδικασία τεσσάρων μικρών χωραφιών. Στην περίπτωση αυτή, δεν είναι δυνατό να εφαρμοστεί διαφορετική μέθοδος άρδευσης μέσα στο ίδιο χωράφι, αντίθετα όμως είναι εφικτό να χρησιμοποιηθεί διαφορετικό λίπασμα ανά λίγα τετραγωνικά μέτρα γης. Ας φανταστούμε, λοιπόν κάθε χωράφι να χωρίζεται σε δύο μέρη. Το γράφημα που ακολουθεί είναι αρκετά διαφωτιστικό όσο αφορά τη χρήση πειραματικού σχεδιασμού split-plot (Jones et al., 2009) :



Γράφημα 3.4

Με Α1 και Α2 συμβολίζονται οι δύο διαφορετικοί τρόποι άρδευσης και με Β1 και Β2 τα δύο είδη λιπασμάτων. Κάθε χωράφι αρδεύεται με ενιαίο τρόπο, αλλά σε κάθε τεμάχιο αυτού χρησιμοποιείται διαφορετικό λίπασμα. Ο παράγοντας Α θεωρείται κύριος παράγοντας *(whole-plot factor)* και ο Β διαιρεμένος *(split-plot factor).*

3.2.7 Ο παραγοντικός σχεδιασμός Plackett-Burman

Οι σχεδιασμοί Plackett-Burman αποτελούν μια ειδική κατηγορία ορθογώνιων κλασματικών σχεδιασμών, οι οποίοι κατασκευάζονται με κυκλικές μεταθέσεις μιας δοσμένης εκτέλεσης ή ομάδων εκτελέσεων. Δόθηκαν στη δημοσιότητα το 1946 από τους Plackett και Burman (Plackett et al., 1946) και αποτελούν ιδιαίτερα χρήσιμο

στατιστικό εργαλείο, αφού επιτρέπουν την εκτίμηση k κυρίων επιδράσεων εκτελώντας μόνο k+1 επαναλήψεις.

Στους σχεδιασμούς αυτούς οι επαναλήψεις είναι ακέραιο πολλαπλάσιο του 4. Στην περίπτωση όπου ο αριθμός των επαναλήψεων αποτελεί δύναμη του 2, τότε ο σχεδιασμός Plackett-Burman ταυτίζεται με τον κλασματικό παραγοντικό σχεδιασμό δύο παραγόντων, αναλυτικής τάξης III.

Στη συνέχεια θα δούμε ένα παράδειγμα τέτοιου σχεδιασμού (Lu, 1992) ο οποίος μάλιστα φαίνεται να εμπεριέχει και αποκομμένες παρατηρήσεις.

Παράδειγμα 3.7

Σε μια μελέτη με σκοπό τη βελτίωση της αξιοπιστίας μιας συσκευής θέρμανσης/ψύξης ο Specht (1985) πραγματοποίησε ένα πείραμα διάρκειας ζωής κάνοντας χρήση ενός σχεδιασμού Pluckett-Burnman 12 επαναλήψεων, που λάμβανε υπόψη 10 παράγοντες δύο σταθμών.

Από αυτούς οι τέσσερις αφορούσαν την κατεργασία των υλικών:

- A (εκτεταμένη-1,ειδική-2)
- F (χωρίς κατεργασία-1, με κατεργασία-2)
- J (παλαιού τύπου-1, νέου τύπου-2)
- K (ενισχυμένη-1, τυπική-2)

Δύο αφορούσαν μεθόδους επιλογής υλικών:

- C (ιδιαίτερη-1, τυπική-2),
- E (ειδική ποιότητα-1, κλασική-2)

και τέλος, τέσσερις αφορούσαν το σχεδιασμό του προϊόντος:

- B (στυλ1-1, στυλ2-2),
- D (βαρύ-1,ελαφρύ-2),
- G (με κοφτερά μέρη-1, χωρίς-2), και
- H (σε φυσική μορφή-1, σε μη φυσική μορφή-2)

Ο πίνακας σχεδιασμού και τα δεδομένα διάρκειας ζωής φαίνονται στη συνέχεια. Κάθε μονάδα χρόνου ζωής αντιστοιχεί σε 100 κύκλους λειτουργίας της συσκευής. Η διάρκεια ζωής μετράται είτε έως ότου επέλθει θραύση των γωνιακού σωλήνα (*TCC: tube corner cracks)* είτε έως ότου αποτύχει ο ενδιάμεσος σωλήνας (*DAG: duct angle*

cracks). Όπως θα δούμε στα πειραματικά δεδομένα εμπεριέχονται πλήρεις αλλά και αποκομμένες παρατηρήσεις.

ΠΙΝΑΚΑΣ 3.8

Α/α	ΠΑΡΑΓΟΝΤΕΣ										Χρόνοι αποτυχίας	
	F	B	A	C	D	E	G	H	J	K	TCC	DAC
1	1	1	1	1	1	1	1	1	1	1	$(164,\infty)$	(128,140)
2	1	1	1	1	1	2	2	2	2	2	$(164,\infty)$	$(164,\infty)$
3	1	1	2	2	2	1	1	2	2	2	(0,42)	(116, 128)
4	1	2	1	2	2	2	2	1	1	2	(93.5, 105)	(56.5, 71)
5	1	2	2	1	2	1	2	1	2	1	(82,93.5)	(71, 82)
6	1	2	2	2	1	2	1	2	1	1	(93.5, 105)	(71,82)
7	2	1	2	2	1	2	2	1	2	1	$(164,\infty)$	$(164,\infty)$
8	2	1	2	1	2	2	1	1	1	2	(56.5, 71)	(56.5, 71)
9	2	1	1	2	2	1	2	2	1	1	$(164,\infty)$	$(164,\infty)$
10	2	2	2	1	1	1	2	2	1	2	(56.5 , 71)	(0,42)
11	2	2	1	2	1	1	1	1	2	2	(128,140)	$(164,\infty)$
12	2	2	1	1	2	2	1	2	2	1	$(164,\infty)$	$(164,\infty)$

Παρατηρούμε ότι και στις δύο αποκρίσεις, οι 5 από το σύνολο των 12 παρατηρήσεων, δηλαδή ένα ποσοστό ύψους πάνω από 40% έχουν χρόνο αποκοπής τους 164 (x100) κύκλους ζωής.

3.2.8 Ο 3^k παραγοντικός σχεδιασμός

Με την ίδια ακριβώς λογική που ορίσαμε σε προηγούμενες ενότητες τον 2^k πειραματικό σχεδιασμό, στην ενότητα αυτή θα ορίσουμε, καθώς επίσης και θα δούμε κάποια παραδείγματα σχετικά με τον 3^k πλήρη και κλασματικό παραγοντικό σχεδιασμό.

3.2.8.1 Ο 3^k πλήρης παραγοντικός σχεδιασμός *(full factorial design)*

Όπως είναι λογικό ο 3^k πλήρης παραγοντικός σχεδιασμός λαμβάνει υπόψη k αριθμό παραγόντων που δέχονται τιμές σε τρεις στάθμες και απαιτεί 3^k επαναλήψεις. Αυτό πρακτικά σημαίνει ότι ένας 3^5 πλήρης παραγοντικός σχεδιασμός απαιτεί 3^5=243 επαναλήψεις, διαδικασία εξαιρετικά χρονοβόρα και επίπονη. Λόγω, λοιπόν, του μεγάλου αριθμού αγωγών που διαθέτουν τέτοιου είδους σχεδιασμοί, συνήθως δεν πραγματοποιούνται πλήρως αλλά κλασματικά.

3.2.8.2 Ο 3^k κλασματικός παραγοντικός σχεδιασμός *(fractional factorial design)*

Ήδη έχουμε εξηγήσει πως λειτουργεί ένας κλασματικός πειραματικός σχεδιασμός. Ας δούμε, λοιπόν, ένα παράδειγμα ενός 3^{3-1} σχεδιασμού.

Ο σχεδιασμός αυτός μελετά 3 παράγοντες, έστω A, B, C οι οποίοι δέχονται τιμές σε τρεις στάθμες ο καθένας. Ας θεωρήσουμε ότι η χαμηλή στάθμη κάθε παράγοντα συμβολίζεται με τον αριθμό 1, η μεσαία με τον αριθμό 2 και η υψηλή με τον αριθμό 3. Ο σχεδιασμός απαιτεί 3^{3-1}=9 επαναλήψεις και ον σχεδιαστικός πίνακας φαίνεται στη συνέχεια:

ΠΙΝΑΚΑΣ 3.9

Α/Α	**Παράγοντες**		
	A	**B**	**C**
1	1	1	1
2	1	2	3
3	1	3	2
4	2	1	3
5	2	2	2
6	2	3	1
7	3	1	2
8	3	2	1
9	3	3	3

ΚΕΦΑΛΑΙΟ 4

ΜΕΘΟΔΟΙ ΑΝΑΛΥΣΗΣ ΣΧΕΔΙΑΣΜΩΝ ΜΕ ΑΠΟΚΟΜΜΕΝΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ

4.1 Εισαγωγικά στοιχεία

Όπως ήδη έχουμε αναφέρει ο παραγοντικός σχεδιασμός καθώς και πιο συγκεκριμένα ο πειραματικός σχεδιασμός του Taguchi χρησιμοποιούνται κατά κόρον από τις βιομηχανίες και όχι μόνο, για τη βελτίωση της παραγωγικής διαδικασίας. Στις εφαρμογές της ανάλυσης αξιοπιστίας ή επιβίωσης, όμως, πολύ συχνά προκύπτουν ελλιπή (αποκομμένα) δεδομένα , κυρίως τύπου ΙΙ.

Αυτό σημαίνει ότι πριν από την έναρξη του πειράματος ορίζεται ο συνολικός αριθμός αντικειμένων ή ατόμων πάνω στα οποία θα πραγματοποιηθεί η μελέτη. Η λήξη, όμως, της πειραματικής διαδικασίας δεν ορίζεται ως η χρονική στιγμή που θα αποτύχει και το τελευταίο αντικείμενο, αλλά ως η χρονική στιγμή όπου ένα ορισμένο ποσοστό αυτών θα έχει αποτύχει.

Στις περιπτώσεις αυτές, απαιτείται από την πλευρά του αναλυτή ιδιαίτερη επιδεξιότητα ώστε να γίνει η προσαρμογή του καταλληλότερου μοντέλου στα πειραματικά δεδομένα, αφού, όπως θα δούμε αναλυτικά στη συνέχεια, η μέθοδος ανάλυσης διασποράς *(analysis of variance (ANOVA) method)* η οποία αποδεικνύεται ιδιαίτερα αποτελεσματική σε εφαρμογές με πλήρη δεδομένα, δεν αποτελεί τόσο χρήσιμο εργαλείο όταν στα πειραματικά δεδομένα εμπεριέχονται εκτός από πλήρεις και αποκομμένες παρατηρήσεις.

4.2 Σύντομη ιστορική αναδρομή

Η προσπάθεια στατιστικής ανάλυσης αποκομμένων παρατηρήσεων μετρά πάνω από τέσσερις δεκαετίες ζωής. Οι αμερικανοί Wayne Nelson και Gerald J.Hahn ήταν από τους πρώτους που ασχολήθηκαν εκτενώς με το θέμα στα τέλη της δεκαετίας του ’60, ενώ αξιοσημείωτη βιβλιογραφία αρχίζει να υπάρχει από τη δεκαετία του ’70 και μετά.

Στους δύο αμερικανούς οφείλεται και η χρήση της μεθόδου ελαχίστων τετραγώνων *(LS: least square method),* καθώς και της επαναληπτικής μεθόδου ελαχίστων τετραγώνων *(ILS: iterative least square method)* στην ανάλυση αξιοπιστίας (Hahn et al., 1981). Άλλη μέθοδος στατιστικής ανάλυσης πειραματικών δεδομένων που χρησιμοποιείται και σε εφαρμογές με ελλιπείς παρατηρήσεις είναι η μέθοδος εκτίμησης μέγιστης πιθανοφάνειας *(MLE: maximum likelihood estimation method)* που χρησιμοποιείται ευρέως σε γραμμικά μοντέλα. Η μέθοδος αυτή χρονολογείται από τις αρχές του 20ου αιώνα, αλλά η χρήση της στην ανάλυση επιβίωσης καθιερώθηκε από τους Hamada και Wu (Hamada et al., 1991) αρκετά χρόνια μετά.

Στις αρχές της δεκαετίας του ’80 ο κινέζος Julong Deng διατύπωσε τη θεωρία “grey system” (Deng, 1989), η οποία συμπεριλαμβάνει τη λεγόμενη μέθοδο “grey prediction” *(grey prediction method)* που αποδεικνύεται ιδιαίτερα χρήσιμη, αφού απαιτεί γενικά μικρό δείγμα ατόμων για τη διεξαγωγή της πειραματικής διαδικασίας, ενώ η εφαρμογή της είναι σχετικά απλή.

Επίσης, η μέθοδος MAA *(minute accumulating analysis)* του ιάπωνα Taguchi (Taguchi et al., 1987), αποδεικνύεται ιδιαίτερα χρήσιμη σε δεδομένα αποκομμένα σε διάστημα, ενώ η μέθοδος ανάλυσης του Torres χρησιμοποιείται ευρέως σε μη επαναλαμβανόμενους και κλασματικούς παραγοντικούς σχεδιασμούς (Torres, 1993).

Το 1994 οι Lu και Unal πρότειναν τον αλγόριθμο EMM *(expectation-modeling-maximization)* που χρησιμοποιεί ψευδο-πλήρη δεδομένα και κάνει εφαρμογή της ανάλυσης παλινδρόμησης για τη σύγκριση των κύριων επιδράσεων και αλληλεπιδράσεων δύο παραγόντων (Lu et al., 1994), ενώ το 1995 οι Hamada και Wu χρησιμοποίησαν μια πολυπλοκότερη μέθοδο υιοθετώντας μια Μπεϋζιανή *(Bayesian)* προσέγγιση με σκοπό να εκτιμήσουν τα ελλιπή δεδομένα που προέκυψαν σε ένα κλασματικό πειραματικό σχεδιασμό, δεδομένου ότι η εκτιμήτρια μέγιστης πιθανοφάνειας δεν υπήρχε ή απειριζόταν (Hamada et al., 1995).

Τέλος, ιδιαίτερα αποτελεσματική αποδεικνύεται, επίσης, η χρήση μη παραμετρικών *(non-parametrical)* μεθόδων και ανάλυσης παλινδρόμησης για την ανάλυση αποκομμένων δεδομένων που προέκυψαν σε επαναλαμβανόμενους ή μη πειραματικούς σχεδιασμούς, αφού με τον τρόπο αυτό λαμβάνεται υπόψη και η μεταβλητότητα των παραγόντων.

4.3 Μέθοδοι ανάλυσης δεδομένων

Στη συνέχεια θα δούμε κάποιες προτεινόμενες μεθόδους για την ανάλυση πειραματικών δεδομένων που εμπεριέχουν αποκομμένες παρατηρήσεις:

4.3.1 Μέθοδος εκτίμησης ελαχίστων τετραγώνων για αποκομμένα πειραματικά δεδομένα

Σε πειράματα που εμπεριέχουν αποκομμένες παρατηρήσεις, η μέθοδος εκτίμησης ελαχίστων τετραγώνων μπορεί να χρησιμοποιηθεί για την εκτίμηση της μέσης τιμής και της διασποράς μόνο κάτω από την υπόθεση της κανονικότητας των πειραματικών δεδομένων.

Έστω ότι το σύνολο των παρατηρήσεων $X_{r+1} < X_{r+2} < ... < X_{n-s}$ αποτελεί ένα κανονικά κατανεμημένο δείγμα δεδομένων τύπου II, αποκομμένων σε διάστημα με μέση τιμή μ και διασπορά σ^2.

Τότε η αθροιστική συνάρτηση κατανομής των X_i με i = r+1 , r+2 ,…., n-s εκφράζεται ως εξής:

$$F(X_i) = \Phi(\frac{\mathrm{X}_i - \mu}{\sigma}), \quad r+1 \le i \le n-s$$

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

$$\hat{F}(X_i) = \Phi(\frac{\mathrm{X}_i - \hat{\mu}}{\hat{\sigma}}), \quad r+1 \le i \le n-s$$ για την οποία είναι γνωστό ότι ισχύει:

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

Η τελευταία εξίσωση γράφεται αλλιώς:

$$\Phi^{-1}(\hat{F}(X_i)) = \frac{X_i - \hat{\mu}}{\hat{\sigma}} = -\frac{\hat{\mu}}{\hat{\sigma}} + \frac{1}{\hat{\sigma}}\mathrm{X}_{i,} \quad r+1 \le i \le n-s \quad (1)$$

Η (1) μπορεί να θεωρηθεί γραμμικό μοντέλο παλινδρόμησης της μορφής:

$$Y_i = \beta_0 + \beta_1 \mathrm{X}_i \ ,$$

με $Y_i = \Phi^{-1}(\hat{F}(X_i))$, $\beta_0 = -\frac{\hat{\mu}}{\hat{\sigma}}$ και $\beta_1 = \frac{1}{\hat{\sigma}}$.

Επομένως, οι εκτιμήτριες μέσης τιμής και διασποράς που συμβολίζονται με $\hat{\mu}$ και $\hat{\sigma}$ αντίστοιχα, μπορούν να υπολογισθούν από την ελαχιστοποίηση του αθροίσματος τετραγωνικών σφαλμάτων ως εξής:

$$\min \sum_{i=r+1}^{n-s} \varepsilon_i^2 = \min \sum_{i=r+1}^{n-s} (Y_i - \beta_0 - \beta_1 \mathrm{X}_i)^2 \quad (2)$$

Με μερική παραγώγιση της (2) ως προς β_0 και β_1 έχουμε τις εκτιμήτριες:

$$\hat{\beta}_1 = \frac{\sum_{i=r+1}^{n-s} (X_i - \bar{X}) Y_i}{\sum_{i=r+1}^{n-s} (X_i - \bar{X})^2} = \frac{1}{\hat{\sigma}} \quad \text{και} \quad \hat{\beta}_0 = \bar{\mathrm{Y}} - \hat{\beta}_1 \bar{\mathrm{X}} \quad (3)$$

Από την (3) προκύπτουν οι ζητούμενες εκτιμήτριες ως:

$$\hat{\sigma} = \frac{\sum_{i=r+1}^{n-s} (X_i - \bar{X})^2}{\sum_{i=r+1}^{n-s} (X_i - \bar{X}) Y_i} \quad \text{και} \quad \hat{\mu} = \bar{\mathrm{X}} - \hat{\sigma} \bar{\mathrm{Y}} \ .$$

Σημειώνεται ότι για r=0 έχουμε την περίπτωση δεξιά αποκομμένων δεδομένων τύπου II, ενώ για s=0 έχουμε την περίπτωση αριστερά αποκομμένων δεδομένων τύπου II αντίστοιχα.

4.3.2 Επαναληπτική μέθοδος ελαχίστων τετραγώνων για αποκομμένα πειραματικά δεδομένα

Η μέθοδος αυτή αναφέρεται στην ξένη βιβλιογραφία ως *(ILS: iterative least squares)* και αποτελεί μια εξίσου απλή μέθοδο ανάλυσης δεδομένων στην οποία αρχικά οι αποκομμένοι χρόνοι αποτυχίας αντιμετωπίζονται σαν να ήταν πλήρεις.

Ας δούμε τα βήματα της μεθόδου:

i. Σ' αυτό το βήμα και μόνο, τα ελλιπή δεδομένα θεωρούνται πλήρη και σ' αυτά γίνεται μια πρώτη εφαρμογή της γνωστής μεθόδου ελαχίστων τετραγώνων.
ii. Στη συνέχεια και με χρήση των αποτελεσμάτων αυτής της πρώτης εφαρμογής, εκτιμάται ο αναμενόμενος χρόνος αποτυχίας για κάθε αποκομμένη παρατήρηση.

iii. Η μέθοδος των ελαχίστων τετραγώνων επαναλαμβάνεται και αυτή τη φορά οι ελλιπείς παρατηρήσεις που στο βήμα i θεωρήθηκαν πλήρεις λαμβάνουν την τιμή που εκτιμήθηκε στο βήμα ii.

iv. Τα βήματα ii και iii επαναλαμβάνονται, μέχρις ότου μετά από πολλαπλές επαναλήψεις της μεθόδου, οι εκτιμήσεις φτάσουν να συγκλίνουν σε μια τιμή.

Ο βασικός περιορισμός της επαναληπτικής μεθόδου ελαχίστων τετραγώνων έγκειται στο γεγονός ότι το τελικό μοντέλο επηρεάζεται σημαντικά από το αρχικά επιλεγμένο, καθώς και το ότι δε λαμβάνονται υπόψη οι αλληλεπιδράσεις και η απόκλιση των παραγόντων.

4.3.3 Μέθοδος ΜΑΑ για την ανάλυση πειραματικών δεδομένων αποκομμένων σε διάστημα

Στη μέθοδο ΜΑΑ *(minute accumulating analysis)* του Taguchi τα δεδομένα κωδικοποιούνται με 0 και 1. Σε κάθε επανάληψη, αν το υπό μελέτη αντικείμενο είναι σε λειτουργία εκφράζεται με το σύμβολο 1, ενώ αν έχει αποτύχει εκφράζεται με 0. Στη συνέχεια, τα δεδομένα αντιμετωπίζονται σαν να προήλθαν από σχεδιασμό split-plot. Σαν "whole-plot factors*"* θεωρούμε όλους τους παράγοντες ελέγχου, ενώ σαν "split-plot factor" τον παράγοντα χρόνο (Tong et al., 1995).

Στα πλαίσια της εργασίας αυτής δε θα ασχοληθούμε περαιτέρω με τη συγκεκριμένη μέθοδο, μιας και το γεγονός ότι κατά την ανάλυση των δεδομένων οι αποκομμένοι χρόνοι αντιμετωπίζονται σαν να ήταν πλήρεις χρόνοι αποτυχίας, οδηγεί συχνά σε ελλιπή συμπεράσματα, αφού οι μη παρατηρούμενοι χρόνοι αποτυχίας μπορεί να απέχουν πολύ από τους χρόνους αποκοπής.

4.3.4 Ανάλυση πειραματικών σχεδιασμών κάνοντας χρήση της ανάλυσης του Torres

Στη μέθοδο αυτή θα χρησιμοποιήσουμε αρχικά τη μέθοδο ελαχίστων τετραγώνων με σκοπό να εκτιμήσουμε τη μέση τιμή και την τυπική απόκλιση κάθε αγωγής.

Στη συνέχεια θα εφαρμόσουμε την ανάλυση του Torres με σκοπό την εκτίμηση της επίδρασης κάθε παράγοντα και κάποιων συνδυασμών παραγόντων στην ολική απόκριση (Tong et al., 2006).

Τα βήματα αυτής περιγράφονται εν συντομία παρακάτω:

a. Θεωρούμε ένα διάνυσμα **L** που εμπεριέχει τις εκτιμήσεις ελαχίστων τετραγώνων του συνόλου των αγωγών, καθώς και το διάνυσμα ταξινόμησης αυτού **R**.
b. Ορίζουμε έναν πίνακα X που να αποτελείται από 1 και -1 και του οποίου η πρώτη στήλη είναι ένα μοναδιαίο διάνυσμα και οι υπόλοιπες αναφέρονται στους κατά τον ερευνητή σημαντικούς παράγοντες και συνδυασμούς παραγόντων.
c. Ορίζουμε το **R** ως διάνυσμα απόκρισης, οι συντελεστές παλινδρόμησης του οποίου υπολογίζονται από την παρακάτω σχέση:

$$a_R = [X^T X]^{-1} X^T R$$

d. Οι εκτιμημένοι στο προηγούμενο βήμα συντελεστές και επιδράσεις, χρησιμοποιούνται για την κατασκευή ενός διαγράμματος κανονικής πιθανότητας, από τη μελέτη του οποίου ο ερευνητής τελικά αποφαίνεται για το βέλτιστο συνδυασμό σταθμών παραγόντων.

4.3.5 Μέθοδος για την ανάλυση αποκομμένων δεδομένων που προκύπτουν από τον παραμετρικό σχεδιασμό του Taguchi

Στα περισσότερα πειράματα η συλλογή πειραματικών δεδομένων αποσκοπεί στην ανάλυση της μέσης απόκρισης. Ο Taguchi, όμως, όπως ήδη έχουμε δει, δίνει γενικά μεγάλη έμφαση στη μελέτη της απόκλισης της απόκρισης, χρησιμοποιώντας ένα δείκτη αναλογίας σήματος/θορύβου *(signal-to-noise (S/N) ratio)*. Αρκετές είναι οι μεθοδολογίες ανάλυσης που λαμβάνουν υπόψη την αναλογία αυτή, κάποιες όμως από αυτές έχουν ευρεία εφαρμογή και χωρίζονται ανάλογα με το είδος της απόκρισης στην οποία αποσκοπούν.

Πιο συγκεκριμένα, χρησιμοποιείται ελαφρώς διαφορετική μεθοδολογία όταν ο ερευνητής ζητά την ελαχιστοποίηση της τιμής απόκρισης *(smaller-the-better),* διαφορετική όταν ζητά τη μεγιστοποίηση αυτής *(larger-the-better)* και τέλος διαφορετική όταν αποσκοπεί σε συγκεκριμένη τιμή *(nominal-the-best)*. (Byrne et al.,

1987). Στην τελευταία περίπτωση, προηγείται της ανάλυσης μια διαδικασία που ακολουθείται με σκοπό η μέση τιμή της απόκρισης να πλησιάσει τη ζητούμενη τιμή, όσο αυτό είναι δυνατό (Peace, 1993).

i. Το πρώτο βήμα για την ενίσχυση της αξιοπιστίας του παραγόμενου προϊόντος είναι φυσικά η μείωση της απόκλισης. Επομένως, αρχικά ο ερευνητής πρέπει να εντοπίσει τους παράγοντες που την επηρεάζουν σημαντικά και στη συνέχεια να καθορίσει το συνδυασμό των επιπέδων των παραγόντων που επηρεάζεται λιγότερο από το θόρυβο. Με τον τρόπο αυτό επιτυγχάνεται και η τελική μείωση της απόκλισης.
ii. Αφού πραγματοποιηθεί η μείωση της απόκλισης των δεδομένων, το επόμενο βήμα είναι να μετατοπισθεί η μέση τιμή ώστε να ταυτίζεται με τη ζητούμενη. Ο ερευνητής σ' αυτό το σημείο επικεντρώνεται στους παράγοντες που επηρεάζουν σημαντικά την απόκριση, αλλά όχι την απόκλιση αυτής κι έτσι γίνεται δυνατή η μετακίνηση της κατανομής των παρατηρήσεων ώστε η μέση τιμή αυτών να ταυτίζεται με τη ζητούμενη, αλλά να μην επηρεαστεί η ίδια η κατανομή τους.

Με τον τρόπο, δηλαδή, αυτό γίνεται δυνατή η μετακίνηση της μέσης τιμής απόκρισης προς την επιθυμητή τιμή *(target value)*.

Αν, όμως, σκοπός της πειραματικής διαδικασίας είναι απλώς η ελαχιστοποίηση της απόκρισης τότε ο τύπος για το δείκτη αναλογίας σήματος θορύβου δίνεται ως εξής:

$$SN_{stb} = -10\log(\frac{1}{n}\sum_{i=1}^{n} y_i^2)$$

Δηλαδή ,

$$SN_{stb} = -10\log\{\frac{1}{n}\sum_{i=1}^{n}[(y_i - \overline{y}) + \overline{y}]^2\}$$

$$= -10\log(s^2 + \overline{y}^2) \quad (4) ,$$

όπου με s^2 συμβολίζεται η σχέση $s^2 = \dfrac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n}$ και

stb: smaller-the-better.

Αντίστοιχα, αν ζητείται η μεγιστοποίηση της απόκρισης *(ltb: larger-the-better),* τότε ο ερευνητής καλείται να πραγματοποιήσει τις κατάλληλες μετατροπές ώστε τελικά το πρόβλημα να αναχθεί στην προηγούμενη περίπτωση: smaller-the-better κι έτσι να χρησιμοποιηθεί ο δείκτης αναλογίας της σχέσης (4).

Τέλος, στην περίπτωση που η ζητούμενη απόκριση θέλουμε να παίρνει συγκεκριμένη τιμή *(ntb: nominal-the-best)* ο δείκτης αναλογίας δίνεται από την παρακάτω σχέση:

$$SN_{ntb} = 10\log(\frac{\overline{y}^2}{s^2}) \quad (5).$$

Ας δούμε τώρα αναλυτικά τα βήματα που ακολουθούνται τελικά για την ανάλυση αποκομμένων δεδομένων που προκύπτουν από τον πειραματικό σχεδιασμό του Taguchi:

i. Αν η απόκριση είναι τύπου larger-the-better, τότε πρέπει αρχικά να μετατραπεί σε smaller-the-better, ενώ αν είναι τύπου smaller-the-better ή nominal-the-best, τότε στο πρώτο βήμα δεν πραγματοποιείται καμία απολύτως ενέργεια .

ii. Μέσω της ήδη γνωστής μας μεθόδου εκτίμησης ελαχίστων τετραγώνων *(least square estimation method)* υπολογίζεται η εκτιμήτρια μέσης τιμής και τυπικής απόκλισης για κάθε μια από τις αγωγές *(treatments).*

iii. Η μέση τιμή και η τυπική απόκλιση που εκτιμήθηκαν στο βήμα ii χρησιμοποιούνται για να υπολογισθούν οι δείκτες αναλογίας σήματος/θορύβου *(S/N ratios)* για κάθε μια αγωγή ξεχωριστά. Σε κάθε τύπο απόκρισης χρησιμοποιείται η αντίστοιχη σχέση για το δείκτη σήματος/θορύβου (4) ή (5).

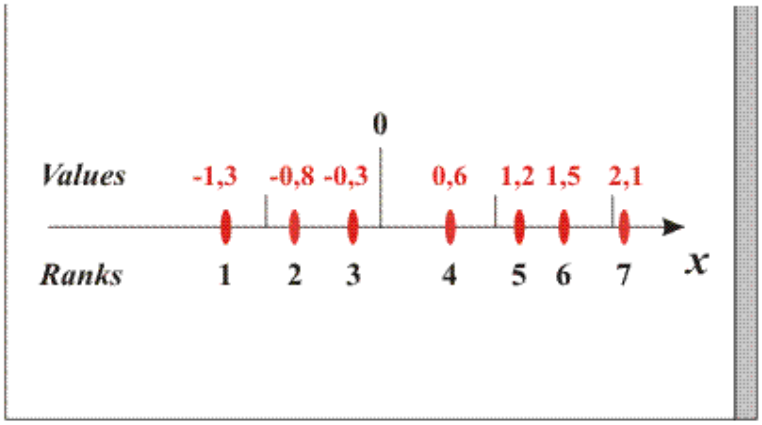
iv. Το επίπεδο ενός παράγοντα επιλέγεται βάσει της τιμής του δείκτη αναλογίας σήματος/θορύβου. Όταν η απόκριση είναι τύπου smaller-the-better, τότε ο συνδυασμός σταθμών παραγόντων που αντιστοιχεί στον υψηλότερο μέσο δείκτη αναλογίας επιλέγεται ως βέλτιστος, ενώ όταν η απόκριση είναι τύπου nominal-the-best ακολουθείται η διαδικασία που περιγράψαμε νωρίτερα για την τελική επιλογή του βέλτιστου συνδυασμού επιπέδων παραγόντων.

4.3.6 Μη παραμετρική μέθοδος για την ανάλυση πειραματικών δεδομένων με αποκομμένες παρατηρήσεις

Συχνά στην ανάλυση δεδομένων με αποκομμένες παρατηρήσεις χρησιμοποιούνται μη παραμετρικές μέθοδοι, καθώς αυτές γίνονται εύκολα κατανοητές από τον εκάστοτε ερευνητή, έχουν απλή εφαρμογή και δεν απαιτούν προϋπάρχουσα γνώση για την κατανομή των μεταβλητών. Η μη παραμετρική μέθοδος που θα αναλυθεί στη συνέχεια, σε συνδυασμό με την ανάλυση παλινδρόμησης όχι μόνο λαμβάνει υπόψη την απόκλιση των παραγόντων, αλλά μπορεί ακόμα να φανεί χρήσιμη σε επαναλαμβανόμενα ή μη πειράματα που εμπεριέχουν από δεξιά ή από αριστερά αποκομμένες παρατηρήσεις.

Ακόμα, στα πλαίσια της μεθόδου αυτής γίνεται χρήση των τάξεων των παρατηρήσεων. Στη συνέχεια, 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 4.1

Η χρήση των τάξεων των παρατηρήσεων, αντί για τις ακριβείς τιμές των δεδομένων, προσφέρει αρκετά πλεονεκτήματα, αφού με τον τρόπο αυτό απλοποιείται σημαντικά η διαδικασία ανάλυσης των πειραματικών δεδομένων. Επιπλέον, η κατανομή των τάξεων παραμένει αναλλοίωτη σε μονότονες μεταθέσεις των

δεδομένων, καθώς επίσης και σε οποιαδήποτε αλλαγή κλίμακας στην περίπτωση ποσοτικών μεταβλητών. Τέλος, αποδεικνύεται ιδιαίτερα χρήσιμη όταν δεν υπάρχει αρκετή πληροφορία για το είδος της κατανομής που ακολουθούν τα πειραματικά δεδομένα.

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

I. Σαν πρώτο βήμα, απαραίτητος είναι ο διαχωρισμός των πειραματικών δεδομένων σε πλήρη *(uncensored)* και αποκομμένα *(censored)*. Με Y_u θα συμβολίζουμε το σύνολο των πλήρων και με Y_c το σύνολο των αποκομμένων.

II. 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:

$$\hat{\mu}_{Y_u|Z_u} = Z_u \hat{\beta}_u \quad (6), \text{ με } \hat{Y}_u = \hat{\mu}_{Y_u|Z_u}$$

III. Σαν τρίτο βήμα απαιτείται η εύρεση της εκτιμήτριας $\hat{Y}_c$ τοποθετώντας στη σχέση (6) αντί για τον πίνακα Z_u, τον πίνακα Z_c, δηλαδή τον πίνακα σταθμών παραγόντων για τα αποκομμένα δεδομένα.

IV. Επόμενο βήμα αποτελεί η ταξινόμηση της εκτιμήτριας αποκομμένων δεδομένων $\hat{Y}_c$, με τα δεδομένα να τοποθετούνται σε αύξουσα σειρά. Το ταξινομημένο διάνυσμα συμβολίζεται $\hat{R}_c = [r_{n+1}, r_{n+2}, ..., r_N]$, όπου N είναι το μέγεθος του δείγματος και n το πλήθος των παρατηρήσεων που δεν είναι αποκομμένες. Έτσι, αν 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.

V. 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 $R = [R_u \mid \hat{R}_c]^T$ και $Z = [Z_u \mid Z_c]^T$ 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:

$$\hat{\mu}_{R|Z} = Z\hat{\beta}_R \quad \text{και} \quad \hat{\mu}_{S|Z} = Z\hat{\beta}_S \ ,$$

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

VI. 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 $\hat{\beta}_S$.

VII. Εντοπισμός του βέλτιστου συνδυασμού σταθμών παραγόντων.

Γενικά, για την επιτυχή εφαρμογή της παραπάνω μεθόδου συνίσταται η χρήση της σε πειράματα όπου τουλάχιστον τα $\frac{2}{3}$ των παρατηρήσεων αποτελούν πλήρη πειραματικά δεδομένα.

4.3.7 Η μέθοδος ανάλυσης αποκομμένων δεδομένων "grey prediction"

Η μέθοδος αυτή δεν απαιτεί ιδιαίτερες προϋποθέσεις και η εφαρμογή της είναι σχετικά απλή, γι' αυτό συχνά χρησιμοποιείται και από ερευνητές που δε διαθέτουν ισχυρό στατιστικό υπόβαθρο.

Αρχικά ο ερευνητής επεξεργάζεται τα πειραματικά δεδομένα ως εξής:

i. Δημιουργεί μια ακολουθία στοιχείων $x^{(0)}$, όπου τοποθετεί τα πειραματικά δεδομένα με την ίδια σειρά με την οποία τα εξέλαβε από το πείραμα.

ii. Δημιουργεί από τη $x^{(0)}$ μια νέα ακολουθία δεδομένων $x^{(1)}$ ως εξής:

$$\text{Αν} \quad x^{(0)} = (x^{(0)}(1), x^{(0)}(2), ..., x^{(0)}(n))$$
$$= (x^{(0)}(k) : k = 1, 2, ..., n) ,$$

Τότε η ακολουθία $x^{(1)}$ συμβολίζεται ως 1-AGO *(accumulated generating operation)* της $x^{(0)}$ και δίνεται ως εξής:

$$x^{(1)} = (\sum_{k=1}^{1} x^{(0)}(k), \sum_{k=1}^{2} x^{(0)}(k), ..., \sum_{k=1}^{n} x^{(0)}(k)) .$$

Έτσι, η ακολουθία $x^{(0)}$ μπορεί εύκολα να βρεθεί εκτελώντας την αντίστροφη διαδικασία *(IAGO: inverse accumulated generating operation).* Η διαδικασία με σύμβολα δίνεται ως: $\text{IAGO}\, x^{(1)} = x^{(0)}$ και για τη $x^{(0)}(\kappa)$ ισχύει η αναδρομική σχέση:

$$x^{(0)}(k) = x^{(1)}(k) - x^{(1)}(k+1) .$$

Στη συνέχεια, ο ερευνητής καλείται να κατασκευάσει ένα μοντέλο, όπως είναι για παράδειγμα τα GM(1,1), GM(1,N) και GM(0,N). Ο συμβολισμός GM *(grey modeling)* αναφέρεται στο γεγονός ότι πρόκειται για μοντέλο της μεθόδου που μελετάμε. Η πρώτη παράμετρος μέσα στην παρένθεση συμβολίζει την τάξη της παραγώγισης του μοντέλου και η δεύτερη τον αριθμό των παραγόντων που μελετώνται.

Το μοντέλο που κατασκευάζεται είναι ένα μοντέλο διαφορικών εξισώσεων, κάτι που σημαίνει ότι η κατασκευή του στηρίζεται σε ένα σύνολο γραμμικών διαφορικών εξισώσεων. Το σύνηθες διαφορικό μοντέλο πρώτης τάξης δίνεται ως εξής:

$$\frac{dx}{dt} + \alpha x = b .$$

Οι διαφορικές εξισώσεις χρησιμοποιούνται κατά κύριο λόγο για τη διαφόριση συνεχών (μη διακριτών) δεδομένων. Η ακολουθία δεδομένων $x^{(0)}$, όμως, αποτελείται από διακριτές, μη διαφορίσιμες παρατηρήσεις. Για το λόγο αυτό ο Deng χρησιμοποίησε το σύνηθες διαφορικό μοντέλο για να κατασκευάσει ένα μοντέλο διαφορικών εξισώσεων για την ακολουθία $x^{(1)}$ και το αποτέλεσμα ήταν το εξής:

$$\frac{dx^{(1)}}{dt} + \alpha x^{(1)} = b \text{ (Deng, 1982),}$$

Βάσει αυτού ας παρακολουθήσουμε την κατασκευή του μοντέλου GM(1,1) που αποτελεί και το δημοφιλέστερο μοντέλο της μεθόδου "grey prediction".

$x^{(0)}(k)+\alpha Z^{(1)}(k)=b$ (6), όπου $Z^{(1)}(k)=\frac{1}{2}x^{(1)}(k)+\frac{1}{2}x^{(1)}(k-1)$

και τα α,b είναι παράμετροι του μοντέλου, των οποίων οι τιμές προκύπτουν με χρήση της μεθόδου ελαχίστων τετραγώνων.

Από την εξίσωση (6), τώρα, για k=2,3,…,n έχουμε:

$$x^{(0)}(2)+\alpha Z^{(1)}(2)=b$$
$$x^{(0)}(3)+\alpha Z^{(1)}(3)=b$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$x^{(0)}(n)+\alpha Z^{(1)}(n)=b$$

Έστω, λοιπόν, $Y_N=[x^{(0)}(2),x^{(0)}(3),....,x^{(0)}(n)]^T$, $B=\begin{matrix}-Z^{(1)}(2) & 1\\ -Z^{(1)}(3) & 1\\ ..... & ...\\ -Z^{(1)}(n) & 1\end{matrix}$ και $\hat{a}=[a,b]^T$.

Τότε το μοντέλο γίνεται $Y_N=B\hat{a}$ και με πολλαπλασιασμό από αριστερά του ανάστροφου B^T έχουμε:

$B^TY_N-B^TB\hat{a}=0$.

Τέλος, με εφαρμογή της μεθόδου ελαχίστων τετραγώνων, προκύπτει ο πίνακας των παραμέτρων του μοντέλου ως εξής:

$\hat{a}=(B^TB)^{-1}B^TY_N$.

Έτσι, αντικαθιστώντας τις τιμές των παραμέτρων, η εξίσωση απόκρισης δίνεται από τη σχέση:

$x^{(1)}(k+1)=[x^{(0)}(1)-\frac{b}{a}]e^{ak}+\frac{b}{a}$ (7)

Από την τελευταία, προκύπτει η $x^{(0)}$, αφού $x^{(0)}=\text{IAGO}\,x^{(1)}$. Έχουμε, λοιπόν:

$$x^{(0)}(k+1)=x^{(1)}(k+1)-x^{(1)}(k)$$
$$=(1-e^{a})[x^{(0)}(1)-\frac{b}{a}]e^{-ak}.$$

Έστω, τώρα, η ακολουθία $x^{(0)} = (x^{(0)}(1), x^{(0)}(2), ..., x^{(0)}(n))$ όπως ορίστηκε στην αρχή. Τότε ορίζουμε το δείκτη αναλογίας κλάσης, που συμβολίζεται με $\sigma^{(0)}(k)$ ως εξής:

$$\sigma^{(0)}(k) = \frac{x^{(0)}(k+1)}{x^{(0)}(k)}, k = 2, 3, ..., n.$$

Βασικό προαπαιτούμενο για τη χρήση της μεθόδου είναι ο δείκτης αναλογίας κλάσης *(class ratio)* να βρίσκεται ανάμεσα στις τιμές 0.1353 και 7.389 κατά την κατασκευή του GM(1,1) (Deng, 1993).

Περιληπτικά τα βήματα για την κατασκευή του μοντέλου φαίνονται στη συνέχεια:

I. Έχοντας την ακολουθία δεδομένων $x^{(0)} = (x^{(0)}(1), x^{(0)}(2), ..., x^{(0)}(n))$, υπολογίζουμε το δείκτη αναλογίας κλάσης *(class ratio)*

II. Εφαρμόζουμε τη διαδικασία AGO *(accumulated generating operation)* για τη δημιουργία της ακολουθίας $x^{(1)}(k)$

III. Υπολογίζουμε την ακολουθία $Z^{(1)}(k) = \frac{1}{2}x^{(1)}(k) + \frac{1}{2}x^{(1)}(k-1)$.

IV. Υπολογίζουμε τις παραμέτρους α και b του μοντέλου χρησιμοποιώντας τη μέθοδο ελαχίστων τετραγώνων

V. Αφού υπολογίσουμε τις τιμές των α και b τις αντικαθιστούμε στην εξίσωση απόκρισης (εξίσωση (7)) κι έτσι έχουμε το μοντέλο GM(1,1).

Μετά την ολοκλήρωση και της διαδικασίας κατασκευής του μοντέλου, ο ερευνητής είναι πλέον έτοιμος να ακολουθήσει τα δύο απλά στάδια από τα οποία απαρτίζεται η μέθοδος ανάλυσης δεδομένων "grey prediction".
Ας δούμε αναλυτικά τα βήματα των δύο αυτών σταδίων ανάλυσης:

<u>Στάδιο 1ο: Υπολογιστική προσέγγιση των αποκομμένων δεδομένων κάθε επανάληψης του πειράματος</u>

i. Διαχωρισμός του συνόλου Y_c των αποκομμένων παρατηρήσεων από το σύνολο Y_u των πειραματικών δεδομένων που είναι πλήρη.

ii. Κατασκευή της ακολουθίας $x^{(0)}$ βάσει του συνόλου Y_u, υπολογισμός του δείκτη αναλογίας κλάσης *(class ratio)* και έλεγχος για το αν αυτός βρίσκεται εντός των επιτρεπτών ορίων για κάθε αγωγή.

iii. Εφαρμογή της διαδικασίας AGO με σκοπό την κατασκευή μοντέλων GM(1,1) που βασίζονται στη $x^{(0)}$ του βήματος ii, όπως αυτή προκύπτει σε κάθε επανάληψη του πειράματος.

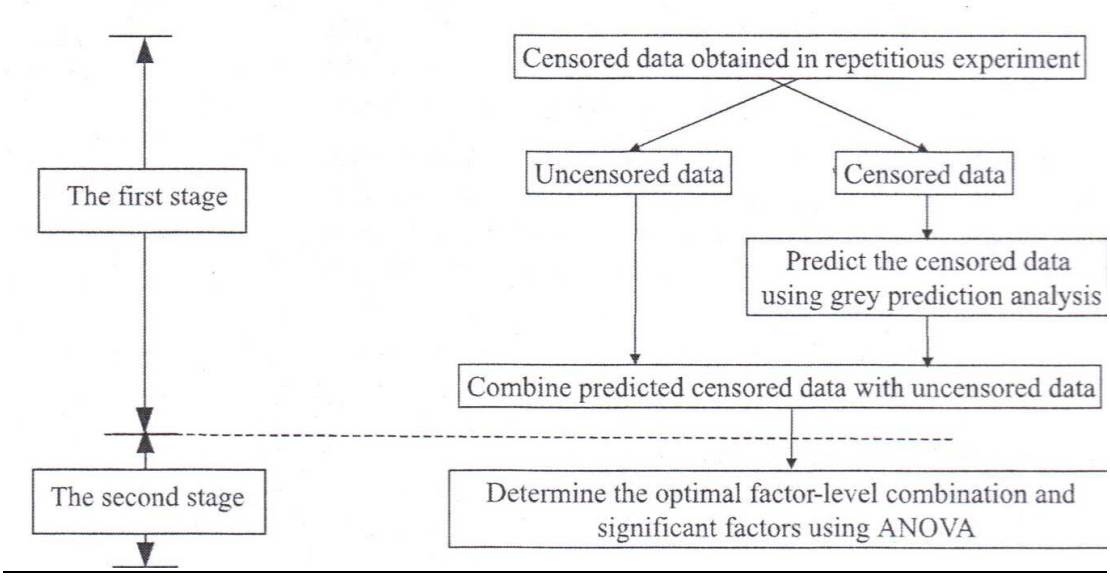
iv. Υπολογισμός κατά προσέγγιση των αποκομμένων παρατηρήσεων κάθε επανάληψης, με χρήση των μοντέλων GM(1,1) του βήματος iii.

v. Συνδυασμός των κατά προσέγγιση υπολογισμένων παρατηρήσεων με τα πλήρη δεδομένα, με σκοπό την τελική δημιουργία ψευδο-πλήρων δεδομένων.

Στάδιο 2^{ο}

Χρήση της γνωστής ανάλυσης διασποράς *(ANOVA method),* για την ανάλυση των πλήρων και ψευδο-πλήρων δεδομένων του πειράματος, με σκοπό την επιλογή του βέλτιστου συνδυασμού σταθμών παραγόντων.

Στη συνέχεια βλέπουμε σχηματικά τα δύο αυτά στάδια της μεθόδου:



Γράφημα 4.2

4.4 Παραδείγματα ανάλυσης πειραματικών σχεδιασμών με αποκομμένες παρατηρήσεις

Στην ενότητα αυτή θα δούμε κάποια αριθμητικά παραδείγματα ανάλυσης παραγοντικών σχεδιασμών στα οποία εμπεριέχονται αποκομμένες παρατηρήσεις. Με τον τρόπο αυτό, στοχεύουμε στην βαθύτερη και πληρέστερη κατανόηση των μεθόδων ανάλυσης που περιγράφθηκαν στην προηγούμενη ενότητα.

4.4.1 Ένα αριθμητικό παράδειγμα πειραματικού σχεδιασμού

4.4.1.1 Περιγραφή του προβλήματος

Ας δούμε την περιγραφή ενός άκρως ρεαλιστικού πειράματος 16 επαναλήψεων (Montgomery, 2001), καθώς και την ανάλυση των δεδομένων αυτού με χρήση της μεθόδου που περιγράφθηκε στην ενότητα 4.3.4.

Το πείραμα λαμβάνει χώρα σε ένα εργοστάσιο παραγωγής ημι-αγωγών και πραγματοποιείται με σκοπό τη μελέτη των επιδράσεων 6 παραγόντων δύο σταθμών ο καθένας, στην καμπυλότητα των παραγόμενων συσκευών. Στον πίνακα 4.1 βλέπουμε τις μεταβλητές του πειράματος με τα επίπεδά τους.

ΠΙΝΑΚΑΣ 4.1

Κωδικοποίηση παράγοντα	Πλήρης ονομασία παράγοντα	Χαμηλή στάθμη	Υψηλή στάθμη
A	Θερμοκρασία σφυρηλάτησης (^{0}C)	55	75
B	Διάρκεια σφυρηλάτησης (sec)	10	25
C	Πίεση σφυρηλάτησης (Ton)	5	10
D	Θερμοκρασία πύρωσης (^{0}C)	1580	1620
E	Διάρκεια κύκλου πυρακτώσεως (hours)	17.5	29
F	Τελική θερμοκρασία δρόσου πυράκτωσης (^{0}C)	20	26

Κάθε συνδυασμός επιπέδων παραγόντων επαναλήφθηκε 4 φορές, κάτι που σημαίνει ότι πρόκειται για έναν επαναλαμβανόμενο *(replicated)* κλασματικό 2^{6-2} πειραματικό σχεδιασμό, σε κάθε επανάληψη του οποίου πραγματοποιήθηκε μέτρηση της καμπυλότητας του παραγόμενου προϊόντος.

Όλη η σχετική πληροφορία για το πείραμα φαίνεται στον πίνακα που ακολουθεί:

ΠΙΝΑΚΑΣ 4.2

Α/Α Αγωγής	**Παράγοντες**						**Παρατήρηση μετά από κάθε επανάληψη**			
	A	**B**	**C**	**D**	**E**	**F**	**X1**	**X2**	**X3**	**X4**
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

4.4.1.2 Ανάλυση των δεδομένων του πειράματος

Στην ενότητα αυτή θα θεωρήσουμε, για την ανάλυση των πειραματικών δεδομένων, τις μεγαλύτερες παρατηρήσεις του πίνακα 4.2, τύπου II από δεξιά αποκομμένα δεδομένα και τη μεταβλητή απόκρισης κανονικά κατανεμημένη. Έτσι, οι παράμετροι n, r και s της μεθόδου ελαχίστων τετραγώνων παίρνουν τιμές n=4, s=1 και r=0 αφού πρόκειται για δεδομένα αποκομμένα από δεξιά. Στη συνέχεια, σε επόμενη ενότητα θα συγκρίνουμε το αποτέλεσμα της ανάλυσης με αποκομμένες παρατηρήσεις με το αποτέλεσμα της ανάλυσης με πλήρεις, με σκοπό να αποδείξουμε την εγκυρότητα των αποτελεσμάτων της μεθόδου.

Στον πίνακα 4.3 φαίνονται τα αποτελέσματα της μεθόδου ελαχίστων τετραγώνων *(LSE method)* και στην τελευταία στήλη φαίνεται η τάξη *(rank)* αυτών.

ΠΙΝΑΚΑΣ 4.3

A/A	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	LSE	Rank
1	128	149	167	157.993	13
2	20	44	62	52.796	4
3	41	42	43	42.512	2
4	30	39	73	59.771	6
5	40	47	47	47.000	3
6	147	219	258	237.062	16
7	86	90	92	90.914	8
8	106	162	186	172.711	15
9	23	32	69	54.778	5
10	45	60	78	69.497	7
11	27	28	28	28.000	1
12	137	158	159	158.364	14
13	86	101	110	105.238	10
14	65	71	109	95.171	9
15	145	145	155	151.836	12
16	93	110	124	116.948	11

Επομένως όπως προκύπτει από τον πίνακα 4.3:

$L=[157.99,52.8,42.51,59.77,47,237.06,90.91,172.71,54.78,69.5,28,158.36,105.24,95.17,151.84,116.95]^T$
και $R=[13,4,2,6,3,16,8,15,5,7,1,14,10,9,12,11]^T$

Σκοπός μας είναι τελικά να επιτύχουμε την κατασκευή ενός διαγράμματος κανονικής πιθανότητας και μέσω αυτού να αποφανθούμε για το βέλτιστο συνδυασμό σταθμών παραγόντων. Έχοντας αυτό κατά νου, υπολογίζουμε το διάνυσμα εκτίμησης των συντελεστών παλινδρόμησης a_R για τους σημαντικότερους παράγοντες του και συνδυασμούς παραγόντων του προβλήματος, δηλαδή στην περίπτωση αυτή για τους: A, B, C, D, E, F, AB, AC, AD, AE, AF, BD, BF, ABD και ACD.

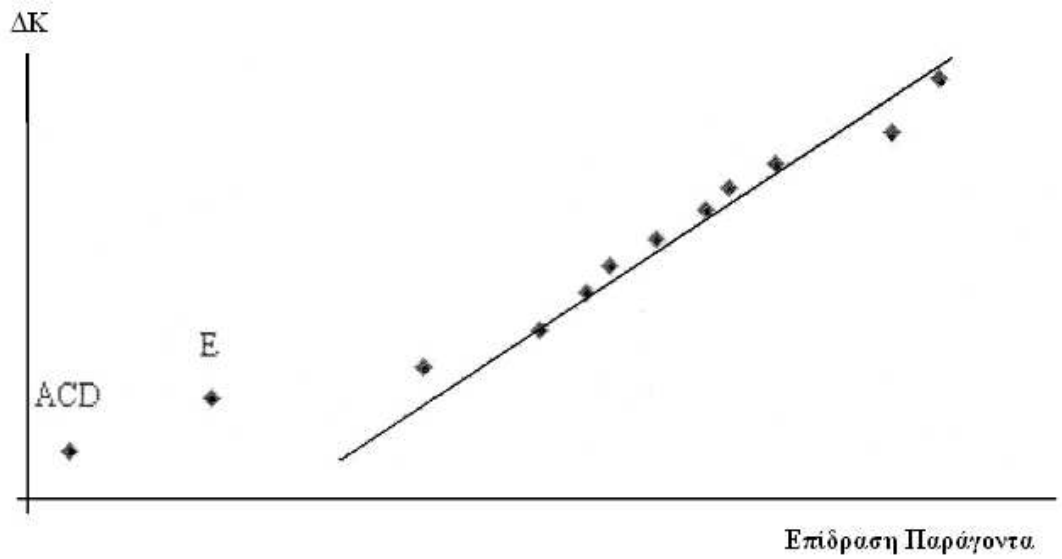
Στον πίνακα 4.4 που ακολουθεί φαίνεται το διάνυσμα a_R, καθώς και ο δείκτης κανονικότητας για κάθε παράγοντα.

ΠΙΝΑΚΑΣ 4.4

Παράγοντες	a_R	**Δείκτης κανονικότητας *(n.p)**	**Τάξη**	**Αθροιστική πιθανότητα**
A	1.750	1.2450	14	0.8934
B	0.125	-0.2491	6.5	0.4016
C	2.000	1.7394	15	0.9590
D	0.125	-0.2491	6.5	0.4016
E	-1.875	-1.2450	2	0.1066
F	-0.750	-0.9458	3	0.1721
AB	1.125	0.9458	13	0.8279
AC	0.500	0.2491	9.5	0.5984
AD	-0.125	-0.6113	4.5	0.2705
AE	0.875	0.7137	12	0.7623
AF	0.500	0.2491	9.5	0.5984
BD	0.750	0.5150	11	0.6967
BF	-0.125	-0.6113	4.5	0.2705
ABD	0.250	0.0000	8	0.5000
ACD	-2.625	-0.7394	1	0.0410

**n.p: normal probability*

Εκμεταλλευόμενοι τα στοιχεία του πίνακα 4.4 μπορούμε πλέον να κατασκευάσουμε το διάγραμμα κανονικής πιθανότητας μαζί με την προσαρμοσμένη ευθεία παλινδρόμησης:



Γράφημα 4.3

Από το διάγραμμα αυτό φαίνεται ότι η επίδραση του παράγοντα Ε και η αλληλεπίδραση των παραγόντων ACD είναι σημαντική. Επειδή η ζητούμενη απόκριση πρέπει να είναι όσο το δυνατό μικρότερη, ορίζουμε τον παράγοντα Ε στην υψηλή του στάθμη, αφού η εκτιμήτρια του συντελεστή παλινδρόμησής του δέχεται αρνητική τιμή (-1.875).

Επίσης, αφού η αλληλεπίδραση ACD φαίνεται να είναι σημαντική πρέπει να συνεκτιμηθούν και οι εκτιμητές των συντελεστών για τους παράγοντες και συνδυασμούς παραγόντων: A,C,D,AC,AD και ACD.

Στον επόμενο πίνακα φαίνεται το πώς επηρεάζεται η τιμή της απόκρισης *(rank),* από διαφορετικούς συνδυασμούς των επιδράσεων και αλληλεπιδράσεων A,C,D,AC,AD και ACD.

ΠΙΝΑΚΑΣ 4.5

Παράγοντες	A	C	D	AC	AD	ACD	Απόκριση (rank)
Συντελεστής Παλινδρόμησης	1.750	2.000	0.125	0.500	-0.125	-2.625	-
Περίπτωση 1	1	1	1	1	1	1	1.625
>> 2	1	1	-1	1	-1	-1	6.875
>> 3	1	-1	1	-1	1	-1	1.875
>> 4	1	-1	-1	-1	-1	1	-3.375
>> 5	-1	1	1	-1	-1	-1	2.625
>> 6	-1	1	-1	-1	1	1	-3.125
>> 7	-1	-1	1	1	-1	1	-5.625
>> 8	-1	-1	-1	1	1	-1	-0.875

Από τον πίνακα 4.5 καθίσταται σαφές ότι τη μικρότερη απόκριση την έχουμε στην κατάσταση 7, όπου οι παράγοντας A και C βρίσκονται στη χαμηλή στάθμη τους και ο παράγοντας D στην υψηλή. Έτσι, έχουμε καταλήξει ότι η βέλτιστη στάθμη για τους παράγοντες A και C είναι η χαμηλή, ενώ για τους παράγοντες D και E η υψηλή. Για τους παράγοντες E και F αντίστοιχα πάλι μπορούμε να αποφανθούμε από τους συντελεστές παλινδρόμησής τους κι έτσι έχουμε τον παράγοντα E να έχει βέλτιστο επίπεδο το χαμηλό και τον F το υψηλό.

Συνοψίζοντας, ο βέλτιστος συνδυασμός σταθμών παραγόντων είναι ο εξής: $A_-B_-C_-D_+E_+F_+$.

4.4.1.3 Σύγκριση του αποτελέσματος της ανάλυσης του πειράματος με αποκομμένα και πλήρη δεδομένα

Αν τα δεδομένα ήταν πλήρη, τότε οι παράγοντες A, B, C, D και E θα θεωρούνταν σημαντικοί. Με το βέλτιστο επίπεδο των παραγόντων A, B και C να είναι το χαμηλό και των παραγόντων D και E το υψηλό. Η διαφορά του αποτελέσματος αυτού από το αποτέλεσμα της ενότητας 4.4.1.2 είναι ότι στην

προηγούμενη ανάλυση ο παράγοντας Β δε θεωρήθηκε σημαντικός. Παρόλα αυτά, το βέλτιστο επίπεδο και για τον παράγοντα αυτόν βρέθηκε να είναι το χαμηλό.

4.4.2 Αριθμητικό παράδειγμα ανάλυσης παραμετρικού σχεδιασμού του Taguchi

4.4.2.1 Περιγραφή του προβλήματος

Το ακόλουθο πείραμα πραγματοποιήθηκε με σκοπό την εξεύρεση του βέλτιστου τρόπου προσάρτησης μιας υποδοχής, πάνω σε ένα συνθετικό σωλήνα, ούτως ώστε να εξασφαλίζεται η ασφαλής και αξιόπιστη χρήση αυτού στη σύνδεση εξαρτημάτων κινητήρων αυτοκινήτων. Τελικός σκοπός, με άλλα λόγια, του πειράματος είναι η δυνατότητα μεγιστοποίησης της εξωτερικής δύναμης που ασκείται στα εξαρτήματα, χωρίς να υπάρχει κίνδυνος βλάβης του κινητήρα και αποκόλλησης των εξαρτημάτων αυτού (Byrne et al., 1987).

Οι παράγοντες που μελετήθηκαν ήταν στο σύνολό τους 7, τέσσερις από αυτούς ήταν παράγοντες ελέγχου τριών σταθμών ο καθένας και τρεις ήταν παράγοντες θορύβου. Οι τρεις αυτοί παράγοντες θορύβου, πιο συγκεκριμένα ο χρόνος, η θερμοκρασία και η σχετική υγρασία τη στιγμή του πειράματος δε θα συμπεριληφθούν στην παρούσα ανάλυση. Αντίθετα, για τους 4 παράγοντες ελέγχου έχουμε τον πίνακα που ακολουθεί:

ΠΙΝΑΚΑΣ 4.6

Κωδικοποίηση	Παράγοντας	Στάθμες		
A	Παρεμβολές	Λίγες-1	Αρκετές-2	Πολλές-3
B	Πάχος τοιχώματος σύνδεσης	Λεπτό-1	Μέτριο-2	Παχύ-3
C	Βάθος διείσδυσης	Επιφανειακό-1	Μέτριο-2	Βαθύ-3
D	Ποσοστό βύθισης στην υποδοχή	Μικρό-1	Μεσαίο-2	Υψηλό-3

Η δύναμη που χρειάζεται για την αποκόλληση των εξαρτημάτων, μετράται 8 φορές για κάθε συνδυασμό σταθμών παραγόντων. Συνολικά οι αγωγές που μελετώνται είναι 9. Άρα πρόκειται για έναν επαναλαμβανόμενο (replicated) κλασματικό 3^{4-2} παραγοντικό σχεδιασμό, η ανάλυση του οποίου θα πραγματοποιηθεί με τη βοήθεια της μεθόδου της ενότητας 4.3.5.

Στον πίνακα 4.7 που ακολουθεί, προσαρτάται όλη η σχετική με το πείραμα πληροφορία. Σημειώνεται ότι, οι μεγαλύτερες παρατηρήσεις θεωρούνται από δεξιά αποκομμένα δεδομένα τύπου II και αυτό σημαίνει ότι οι τιμές των παραμέτρων για τη μετέπειτα εφαρμογή της μεθόδου ελαχίστων τετραγώνων, με σκοπό την εκτίμηση της μέσης τιμής και διασποράς, είναι οι εξής:

$$n=8\ ,\quad r=0\quad \text{και}\quad s=1.$$

ΠΙΝΑΚΑΣ 4.7

A/A	Παράγοντας				Παρατήρηση σε κάθε επανάληψη								S/N ratio
	A	B	C	D	X1	X2	X3	X4	X5	X6	X7	X8	
1	1	1	1	1	15.6	9.5	16.9	19.9	19.6	19.6	20.0	19.1	24.045
2	1	2	2	2	15.0	16.2	19.4	19.6	19.7	19.8	24.2	21.9	25.522
3	1	3	3	3	16.3	16.7	19.1	15.6	22.6	18.2	23.3	20.4	25.335
4	2	1	2	3	18.3	17.4	18.9	18.6	21.0	18.9	23.2	24.7	25.904
5	2	2	3	1	19.7	18.6	19.4	25.1	25.6	21.4	27.5	25.3	26.908
6	2	3	1	2	16.2	16.3	20.0	19.8	14.7	19.6	22.5	24.7	25.326
7	3	1	3	2	16.4	19.1	18.4	23.6	16.8	18.6	24.3	21.6	25.711
8	3	2	1	3	14.2	15.6	15.1	16.8	17.8	19.6	23.2	24.4	24.832
9	3	3	2	1	16.1	19.9	19.3	17.3	23.1	22.7	22.6	28.6	26.152

4.4.2.2 Ανάλυση των δεδομένων του πειράματος

Στο πρόβλημα αυτό, η ζητούμενη απόκριση είναι τύπου larger-the-better, επομένως σύμφωνα με τα βήματα της μεθόδου που περιγράψαμε αναλυτικά στην ενότητα 4.3.5, σαν πρώτο βήμα πρέπει να πραγματοποιήσουμε τις κατάλληλες μετατροπές ώστε η απόκριση να γίνει τύπου smaller-the-better. Στα πλαίσια της διαδικασίας αυτής, τα τύπου II από δεξιά αποκομμένα δεδομένα *(right censored data)* του πειράματος θα μετατραπούν σε τύπου II από αριστερά αποκομμένα δεδομένα *(left censored data).* Αυτό σημαίνει αυτόματα πως οι τιμές των παραμέτρων n,r και s διαμορφώνονται ως εξής:

n=8 , r=1 και s=0.

Στον πίνακα 4.8 βλέπουμε τα από αριστερά πλέον αποκομμένα δεδομένα του πειράματος:

ΠΙΝΑΚΑΣ 4.8

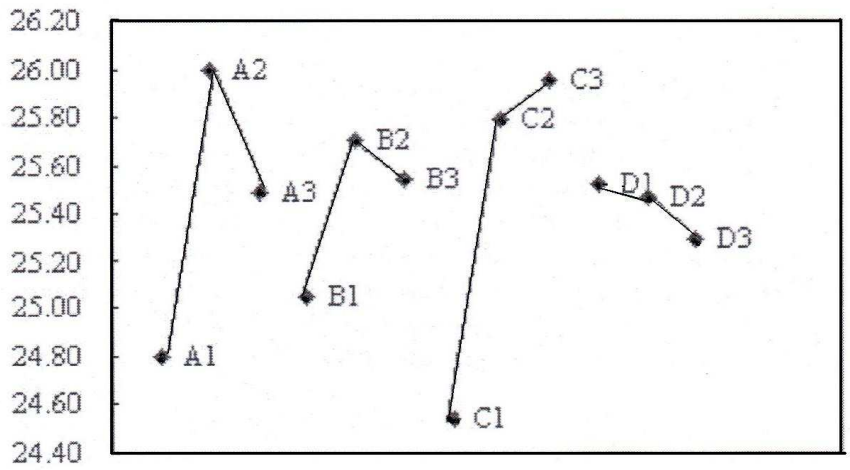
Α/Α	**Παρατηρούμενη δύναμη αποκόλλησης**							
	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$	$X_{(8)}$
1	-	0.05025	0.05102	0.05102	0.05236	0.05917	0.06410	0.10526
2	-	0.04566	0.05076	0.05076	0.05102	0.05155	0.06173	0.06667
3	-	0.04425	0.05236	0.05236	0.05495	0.05988	0.06135	0.06410
4	-	0.04310	0.05291	0.05291	0.05291	0.05376	0.05464	0.05747
5	-	0.03906	0.03984	0.03984	0.04673	0.05076	0.05155	0.05376
6	-	0.04444	0.05051	0.05051	0.05102	0.06135	0.06173	0.06803
7	-	0.04237	0.05236	0.05236	0.05376	0.05435	0.05952	0.06098
8	-	0.04310	0.05618	0.05618	0.05952	0.06410	0.06623	0.07042
9	-	0.04329	0.04425	0.04425	0.05025	0.05181	0.05780	0.06211

Στη συνέχεια, με την εφαρμογή της μεθόδου ελαχίστων τετραγώνων βρίσκουμε την εκτιμήτρια μέσης τιμής και τυπικής απόκλισης κάθε αγωγής, ενώ από τον τύπο $SN_{stb} = -10\log(s^2 + \bar{y}^2)$ υπολογίζουμε το δείκτη αναλογίας θορύβου για κάθε μια από αυτές. Στον πίνακα 4.9 που ακολουθεί φαίνονται τα σχετικά αποτελέσματα.

ΠΙΝΑΚΑΣ 4.9

A/A αγωγής	$\hat{\mu}$	$\hat{\sigma}$	SN ratio
1	0.05577	0.00123	23.627
2	0.05196	0.00013	25.475
3	0.05330	0.00011	25.300
4	0.05045	0.00006	25.845
5	0.04420	0.00009	26.888
6	0.05309	0.00016	25.260
7	0.05107	0.00010	25.675
8	0.05622	0.00019	24.742
9	0.04860	0.00012	26.052

Υπάρχουν παραπάνω από ένας τρόποι ανάλυσης δεδομένων που μπορούν να χρησιμοποιηθούν από το σημείο αυτό και μετά. Στην παρούσα ανάλυση θα κατασκευάσουμε το διάγραμμα του μέσου δείκτη θορύβου *(average S/N ratio)* για κάθε επίπεδο των τεσσάρων παραγόντων ελέγχου. Έτσι, γνωρίζοντας ότι ζητούμενο είναι η κατά το δυνατό μεγαλύτερη απόκριση *(larger-the-better)* μπορούμε να αποφανθούμε μέσω του διαγράμματος για τη σημαντικότητα και το βέλτιστο επίπεδο κάθε παράγοντα.



Γράφημα 4.4

Πράγματι, από το γράφημα 4.4 διαφαίνεται ότι το βέλτιστο επίπεδο για τους παράγοντες Α και Β είναι το επίπεδο 2, για τον C το 3 και για τον D το 1. Επίσης, σημαντικότεροι φαίνεται να είναι οι παράγοντες Α και C, ενώ λιγότερο σημαντικοί οι Β και D. Συνεπώς, βέλτιστος συνδυασμός σταθμών παραγόντων είναι ο: $A_2B_2C_3D_1$.

4.4.2.3 Σύγκριση του αποτελέσματος της ανάλυσης του πειράματος με αποκομμένα και πλήρη δεδομένα

Στην παρούσα πειραματική εφαρμογή, η ανάλυση με πλήρη δεδομένα δε θα είχε επιφέρει διαφορετικά αποτελέσματα. Οι τέσσερις παράγοντες θα θεωρούνταν σημαντικοί, με βέλτιστα επίπεδα τα ίδια με αυτά που προέκυψαν στην ανάλυση της ενότητας 4.4.2.2.

Μετά από αυτή τη σύντομη εφαρμογή της μεθόδου καθίσταται πλέον σαφής η μεγάλη χρηστική αξία της, αφού εύκολα μπορεί να εφαρμοστεί ακόμα και από άτομα που δεν έχουν ευρεία γνώση στατιστικών μεθόδων. Η χρήση της είναι απλή, όχι ιδιαίτερα χρονοβόρα και μπορεί να εφαρμοστεί σε ευρεία γκάμα βιομηχανικών πειραμάτων.

4.4.3 Αριθμητικό παράδειγμα μη παραμετρικής ανάλυσης δεδομένων στα οποία εμπεριέχονται αποκομμένες παρατηρήσεις

4.4.3.1 Περιγραφή του προβλήματος

Στο παρόν παράδειγμα θα μελετήσουμε έναν L_8 ορθογώνιο πειραματικό σχεδιασμό με πέντε παράγοντες, δύο σταθμών ο καθένας. Το σύνολο των εκτελούμενων συνδυασμών επιπέδων παραγόντων είναι 8 και κάθε συνδυασμός εκτελείται δύο φορές. Το είδος της ζητούμενης απόκρισης είναι smaller-the-better.

Στον πίνακα 4.10 βλέπουμε τα πειραματικά δεδομένα. Οι παρατηρήσεις που δίνονται με το σύμβολο * αποτελούν από δεξιά αποκομμένα δεδομένα με σημείο αποκοπής το 67. Οι τελευταίες δύο στήλες (R_u) υποδεικνύουν την τάξη κάθε μη αποκομμένης παρατήρησης και θα φανούν χρήσιμες κατά την ανάλυση των δεδομένων με βάση τη μέθοδο της ενότητας 4.3.6.

ΠΙΝΑΚΑΣ 4.10

Α/Α Αγωγής	ΠΑΡΑΓΟΝΤΕΣ					ΑΠΟΚΡΙΣΗ		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

4.4.3.2 Ανάλυση των δεδομένων του πειράματος

I. Από τον πίνακα 4.10 λαμβάνουμε το διάνυσμα πλήρων *(uncensored)* δεδομένων Y_u, καθώς και το αντίστοιχο διάνυσμα τάξεων R_u:

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

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

II. Στη συνέχεια με τη βοήθεια της ανάλυσης παλινδρόμησης, βρίσκουμε τη σχέση ανάμεσα στο διάνυσμα Y_u και τον πίνακα Z_u των σταθμών παραγόντων των 5 αποκομμένων δεδομένων του πειράματος. Έχουμε:

$Z_u =$

	A	**B**	**C**	**D**	**E**	**AB**	**AC**	**AD**	**AE**	**BC**	**BD**	**BE**	**CD**	**CE**	**DE**
1	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	2	1
1	2	1	2	2	1	2	4	4	2	2	2	1	4	2	2
1	2	1	1	2	1	4	2	4	2	2	4	2	2	1	2
1	2	2	2	1	2	4	4	2	4	4	2	4	2	4	2

(5x16)

και $\hat{\mu}_{Y_u|Z_u} = 88 - 17.5A - 1.5B - 0.5C - 8.5D + 6E$ (8)

III. Εκτιμούμε τα αποκομμένα δεδομένα εισάγοντας το επίπεδο των παραγόντων κάθε αγωγής στην εξίσωση (8). Έτσι για την αγωγή 2 η αποκομμένη παρατήρηση εκτιμάται να έχει την τιμή 63, για την αγωγή 3 την τιμή 62 και τέλος για την αγωγή 5 την τιμή 54.5.

IV. Αφού το πείραμα εμπεριέχει 5 από δεξιά αποκομμένες παρατηρήσεις σε ένα σύνολο 16 παρατηρήσεων, οι τάξεις αυτών θα είναι από 12 μέχρι 16. Το διάνυσμα ταξινόμησης των αποκομμένων *(censored)* δεδομένων του πειράματος προκύπτει ως εξής:

$$\hat{R}_c = [16,14,15,12,13]^T$$

V. Υπολογίζουμε τη μέση τιμή των τάξεων R_j , καθώς και την τυπική απόκλιση αυτών S_j χρησιμοποιώντας το διάνυσμα $R = [R_u \mid \hat{R}_c]^T$. Τα αποτελέσματα φαίνονται στον πίνακα 4.11:

ΠΙΝΑΚΑΣ 4.11

Αγωγή	**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	3.32	0.5	0.75	0.5	1.5	0.5	0.5

Τελικά κατασκευάζουμε όπως και πριν τα μοντέλα παλινδρόμησης:

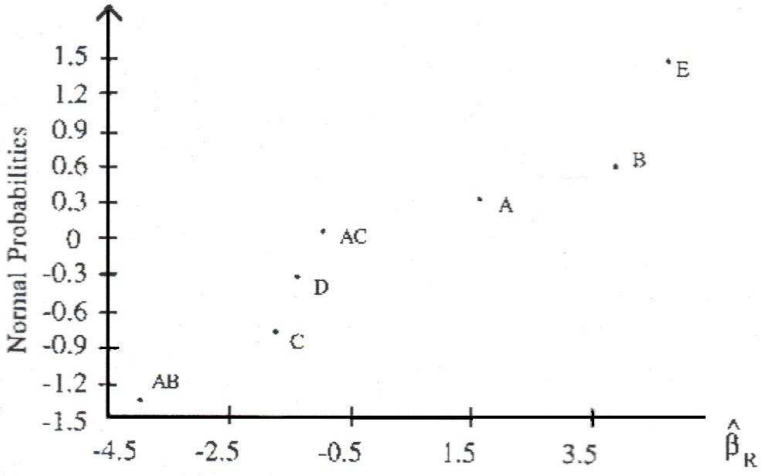
$\hat{\mu}_{R|Z} = 7.25 + 1.75A + 4.25B - 1.75C - 1.25D + 5E - 4AB - 0.75AC$ (9) και

$\hat{\mu}_{S|Z} = -3.02 + 0.73A - 1.57B + 3.07C + 1.02D + 0.52E + 0.54AB - 1.29AC$ (10)

VI. Κατασκευάζουμε το διάγραμμα κανονικής πιθανότητας για το μοντέλο (9) (Γράφημα 4.5), με σκοπό τον έλεγχο της υπόθεσης της κανονικότητας των παρατηρήσεων. Στη συνέχεια, πραγματοποιούμε την ίδια ακριβώς διαδικασία για το μοντέλο (10) (Γράφημα 4.6).

ΠΙΝΑΚΑΣ 4.12

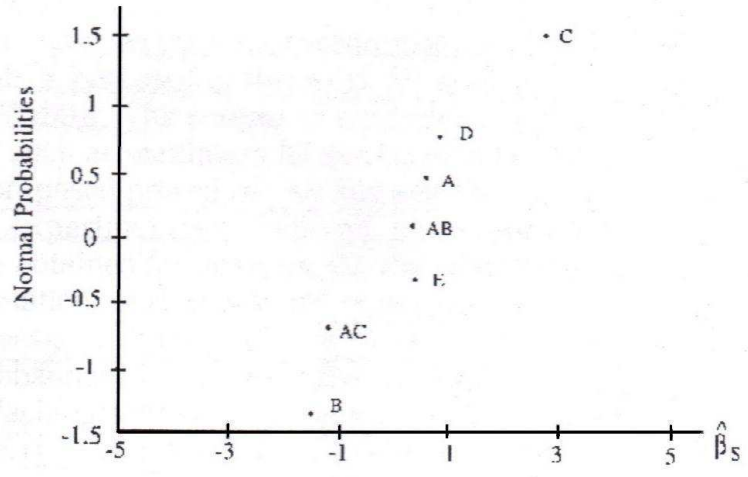
Επίδραση	A	B	C	D	E	AB	AC
$\hat{\beta}_R$	1.75	4.25	-1.75	-1.25	5	-4	-0.75
n.p	0.35147	0.75613	-0.75613	-0.35147	1.36459	-1.36459	0



Γράφημα 4.5

ΠΙΝΑΚΑΣ 4.13

Επίδραση	A	B	C	D	E	AB	AC
$\hat{\beta}_s$	0.7315	-1.574	3.074	1.0185	0.5185	0.537	-1.287
n.p	0.35147	-1.36459	1.36459	0.75613	-0.35147	0	-0.75613



Γράφημα 4.6

VII. Αφού πραγματοποιηθεί ο έλεγχος κανονικότητας, το τελικό βήμα είναι ο εντοπισμός του βέλτιστου συνδυασμού επιπέδων παραγόντων. Από το γράφημα 4.5 προκύπτει ότι οι παράγοντες AB, E, A και B επιδρούν σημαντικά στη μέση τιμή, ενώ από το γράφημα 4.6 προκύπτει, αντίστοιχα, πως οι παράγοντες AC, A, B και C επιδρούν σημαντικά στην τυπική απόκλιση. Από την εξίσωση (9) καθίσταται σαφές ότι βέλτιστο επίπεδο για τον παράγοντα E είναι το επίπεδο 1, αφού ο τύπος της απόκρισης είναι smaller-the-better. Για τους παράγοντες A και B έχουμε τα εξής:

ΠΙΝΑΚΑΣ 4.14

Παράγοντες	**A**	**B**	**AB**	$\hat{\mu}_{R\|Z}$
Συντελεστής Παλινδρόμησης	1.75	4.25	-4	-
Περίπτωση 1	1	1	1	2
>> 2	1	2	2	2.25
>> 3	2	1	2	-0.35
>> 4	2	2	4	-4

Από τον πίνακα 4.14 προκύπτει ότι το βέλτιστο επίπεδο για τους παράγοντες A και B είναι το επίπεδο 2, αφού στην Περίπτωση 4 ελαχιστοποιείται η μέση απόκριση. Αντίστοιχα, από τον πίνακα 4.15, επιβεβαιώνεται ότι το βέλτιστο επίπεδο για τον παράγοντα A είναι το 2, ενώ για τον παράγοντα C βέλτιστο φαίνεται να είναι το επίπεδο 1, αφού στην Περίπτωση 3 ελαχιστοποιείται η μέση τυπική απόκλιση.

ΠΙΝΑΚΑΣ 4.15

Παράγοντες	**A**	**C**	**AC**	$\hat{\mu}_{S\|Z}$
Συντελεστής Παλινδρόμησης	0.7315	3.074	-1.287	-
Περίπτωση 1	1	1	1	2.5185
>> 2	1	2	2	4.3055
>> 3	2	1	2	1.963
>> 4	2	2	4	2.463

Τέλος, από τη σχέση (10) επιβεβαιώνεται ότι το βέλτιστο επίπεδο για τον παράγοντα Β είναι το 2. Τελικά, από την εφαρμογή της μεθόδου, βέλτιστος συνδυασμός επιπέδων είναι ο εξής: $A_2B_2C_1E_1$.

4.4.3.3 Σύγκριση του αποτελέσματος της ανάλυσης του πειράματος με αποκομμένα και πλήρη δεδομένα

Στην περίπτωση ανάλυσης των δεδομένων του πίνακα 4.10 χωρίς να θεωρηθούν αποκομμένες οι 5 από τις 16 παρατηρήσεις, ο βέλτιστος συνδυασμός σταθμών παραγόντων που θα προέκυπτε από την ανάλυση του πίνακα ANOVA είναι ο εξής: $A_2B_2C_2E_1$. Συνεπώς, το τελικό αποτέλεσμα της μη παραμετρικής μεθόδου που ακολουθήθηκε, δε διαφέρει σημαντικά από αυτό της παραδοσιακής ανάλυσης διασποράς.

4.4.4 Παράδειγμα ανάλυσης δεδομένων με χρήση της μεθόδου "grey prediction"

Στα πλαίσια της ενότητας αυτής θα μελετήσουμε έναν πλήρη παραγοντικό σχεδιασμό δύο μεταβλητών, κάθε μια από τις οποίες λαμβάνει τιμές σε δύο στάθμες. Στην πειραματική διαδικασία, πραγματοποιήθηκαν 8 επαναλήψεις του πειράματος για κάθε συνδυασμό σταθμών παραγόντων, στις οποίες δεν προέκυψαν αποκομμένες παρατηρήσεις. Στην παρούσα ανάλυση όμως, όπως κάναμε και σε προηγούμενες ενότητες, θα θεωρήσουμε τις δύο τελευταίες παρατηρήσεις μεγαλύτερης τάξης αποκομμένες, δημιουργώντας έτσι από δεξιά αποκομμένα δεδομένα τύπου ΙΙ. Στη συνέχεια, αφού πραγματοποιήσουμε την ανάλυση αυτών με τη μέθοδο "grey prediction" και καταλήξουμε σε συμπεράσματα, θα τα συγκρίνουμε με το αποτέλεσμα της ανάλυσης με πλήρη πειραματικά δεδομένα.

4.4.4.1 Περιγραφή του προβλήματος

Σκοπός του πειράματος είναι ο καθορισμός των σημαντικών παραγόντων, καθώς και του βέλτιστου συνδυασμού επιπέδων αυτών, ώστε να μεγιστοποιείται ο αναμενόμενος χρόνος ζωής των παραγόμενων πυκνωτών επιφάνειας (Condra, 1993).

Στην πραγματικότητα, μελετάται ο χρόνος ζωής 8 πυκνωτών σε κάθε αγωγή. Παρόλο που το πείραμα ολοκληρώνεται όταν αποτύχουν και οι 8, στην παρούσα ανάλυση θεωρείται πως η πειραματική διαδικασία περατώνεται μόλις αποτύχουν οι 6 πρώτοι από αυτούς. Ο πίνακας των πλήρων πειραματικών δεδομένων είναι ο εξής:

ΠΙΝΑΚΑΣ 4.16

Α/Α Αγωγής	**Παράγοντες**		**Παρατηρούμενη απόκριση**							
	Α	**Β**	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
1	1	1	430	950	560	210	310	230	250	230
2	1	2	1080	1060	890	450	430	320	340	430
3	2	1	890	1060	680	310	310	310	250	230
4	2	2	1100	1080	1080	460	620	370	580	430

Στον πίνακα 4.17 που ακολουθεί, βλέπουμε τα δεδομένα του πειράματος κατ' αύξουσα σειρά:

ΠΙΝΑΚΑΣ 4.17

Α/Α Αγωγής	**Παράγοντες**		**Παρατηρούμενη απόκριση**							
	Α	**Β**	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
1	1	1	210	230	230	250	310	430	560	950
2	1	2	320	340	430	430	450	890	1060	1080
3	2	1	230	250	310	310	310	680	890	1060
4	2	2	370	430	460	580	620	1080	1080	1100

4.4.4.2 Ανάλυση των δεδομένων του πειράματος

Κάνοντας χρήση των βημάτων της μεθόδου που περιγράφθηκαν αναλυτικά στην ενότητα 4.3.7, πραγματοποιούμε την παρακάτω ανάλυση για τα δεδομένα του πίνακα 4.18.

Στάδιο 1^{ο} :

i. Με Y_u συμβολίζουμε το σύνολο των δεδομένων που θεωρούνται πλήρη και με Y_c το σύνολο των αποκομμένων.

ΠΙΝΑΚΑΣ 4.18

Α/Α Αγωγής	Παράγοντες		Y_u						Y_c	
	A	**B**	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
1	1	2	210	230	230	250	310	430	-	-
2	1	2	320	340	430	430	450	890	-	-
3	2	1	230	250	310	310	310	680	-	-
4	2	2	370	430	460	580	620	1080	-	-

ii. Κατασκευάζουμε το διάνυσμα $x^{(0)}$ για κάθε αγωγή και υπολογίζουμε το δείκτη αναλογίας κλάσης (πίνακας 4.19):

ΠΙΝΑΚΑΣ 4.19

Α/Α	$x^{(0)}$	**Class ratio**
1	(210,230,230,250,310,430)	(-,0.91,1.00,0.92,0.81,0.72)
2	(320,340,430,430,450,890)	(-,0.94,0.79,1.00,0.96,0.51)
3	(230,250,310,310,310,680)	(-,0.92,0.81,1.00,1.00,0.46)
4	(370,430,460,580,620,1080)	(-,0.86,0.93,0.79,0.95,0.57)

Σημειώνεται ότι, όλοι οι δείκτες της τρίτης στήλης του πίνακα βρίσκονται μέσα στα επιτρεπτά όρια από 0.1353 μέχρι 7.389.

iii. Εφαρμόζουμε τη διαδικασία AGO με σκοπό την κατασκευή μοντέλων GM(1,1) για κάθε αγωγή.

ΠΙΝΑΚΑΣ 4.20

Α/Α	$x^{(0)}$	**GM(1,1)** ($x^{(0)}(k+1)$ = **…**)
1	(210,230,230,250,310,430)	$974.5835(1-e^{-0.18085})e^{0.18085k}$
2	(320,340,430,430,450,890)	$968.482(1-e^{-0.25217})e^{0.25217k}$
3	(230,250,310,310,310,680)	$613.023(1-e^{-0.27133})e^{0.27133k}$
4	(370,430,460,580,620,1080)	$1172.946(1-e^{-0.2571})e^{0.2571k}$

iv. Υπολογίζουμε κατά προσέγγιση τις αποκομμένες παρατηρήσεις κάθε επανάληψης, με χρήση των μοντέλων GM(1,1) του πίνακα 4.20.

ΠΙΝΑΚΑΣ 4.21

Α/Α	**Εκτιμήσεις αποκομμένων παρατηρήσεων**	
	$x^{(0)}(7)$	$x^{(0)}(8)$
1	477.185	571.777
2	980.143	1261.268
3	742.012	973.303
4	1243.65	1608.261

Ο τελικός πίνακας σχεδιασμού περιλαμβάνει τα πλήρη πειραματικά δεδομένα του πίνακα 4.18, καθώς και τα δεδομένα του πίνακα 4.21.

ΠΙΝΑΚΑΣ 4.22

Α/Α Αγωγής	**Παράγοντες**		**Ψευδο-πλήρη πειραματικά δεδομένα**							
	Α	**Β**	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
1	1	1	210	230	230	250	310	430	*477.185	*571.777
2	1	2	320	340	430	430	450	890	*980.143	*1261.268
3	2	1	230	250	310	310	310	680	*742.012	*973.303
4	2	2	370	430	460	580	620	1080	*1243.65	*1608.261

Στάδιο $2^{ο}$:

Για τα ψευδο-πλήρη δεδομένα του πίνακα 4.22 χρησιμοποιούμε τη γνωστή από την κλασική στατιστική ανάλυση διασποράς *(analysis of variance)*, με σκοπό να εντοπίσουμε τον παράγοντα ή τους παράγοντες που επιδρούν σημαντικά στο χρόνο ζωής των πυκνωτών. Είναι γενικά γνωστό, ότι ένας στατιστικός αναλυτής δύναται με χρήση ανάλυσης διασποράς να εξετάσει αν η μεταβλητότητα των τιμών της εξαρτημένης μεταβλητής εξηγείται από τις ανεξάρτητες μεταβλητές (Οικονόμου, 2010). Στη συνέχεια, βλέπουμε την κατασκευή του πίνακα ANOVA για το παράδειγμα αυτό.

ΠΙΝΑΚΑΣ 4.23

ANOVA

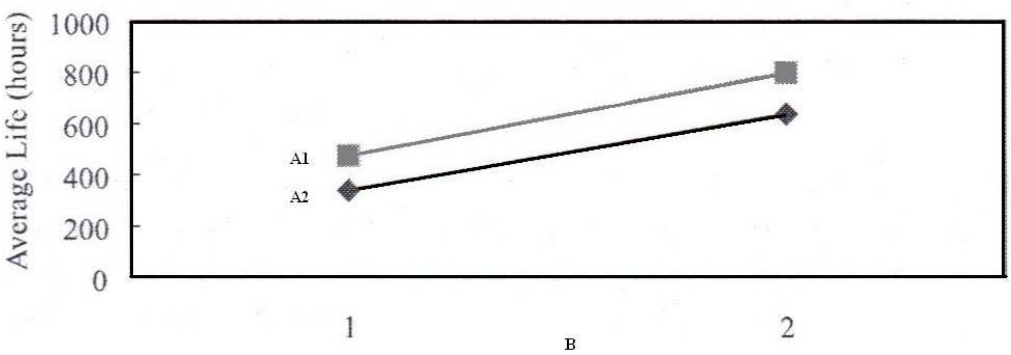
Πηγή	SS	df	MS	F	p-value
A	178033.3	1	178033.3	1.654911	0.208827
B	774715	1	774715	7.201373	0.012091
AB	1177.909	1	1177.909	0.010949	0.917408
Σφάλμα	3012206	28	107578.8	-	-
Ολική	3966132	31	-	-	-

SS: sum of squares, df: degrees of freedom, MS: mean square

όπου, $\mathrm{MS} = \frac{SS}{df}$ και $\mathrm{F} = \frac{\mathrm{MS}}{df}$.

Η p-τιμή για τον παράγοντα B είναι μικρότερη από τη στάθμη σημαντικότητας α=0.05, άρα ο παράγοντας B θεωρείται ότι επιδρά σημαντικά στην απόκριση.

Για να αποφανθούμε για το βέλτιστο συνδυασμό παραγόντων και εφόσον από τον πίνακα ANOVA υποδεικνύεται πως η αλληλεπίδραση AB δεν είναι σημαντική, σχεδιάζουμε όπως και στο γράφημα 3.2 το διάγραμμα των κύριων επιδράσεων για τους παράγοντες A και B.



Γράφημα 4.7

Από το γράφημα 4.7 προκύπτει ότι ο βέλτιστος συνδυασμός σταθμών παραγόντων, δηλαδή ο συνδυασμός που μεγιστοποιεί τη μέση διάρκεια ζωής *(average life)* είναι ο εξής: A_1B_2 .

4.4.4.3 Σύγκριση του αποτελέσματος της ανάλυσης του πειράματος με αποκομμένα και πλήρη δεδομένα

Για να αποφανθούμε σχετικά με το αν η μέθοδος της προηγούμενης ενότητας είναι αξιόπιστη, ας δούμε ποιο θα ήταν το αποτέλεσμα για το βέλτιστο συνδυασμό επιπέδων παραγόντων μετά από την ανάλυση των πλήρων δεδομένων του πειράματος. Ο πίνακας ANOVA για τα πλήρη αποκομμένα δεδομένα θα ήταν ο εξής:

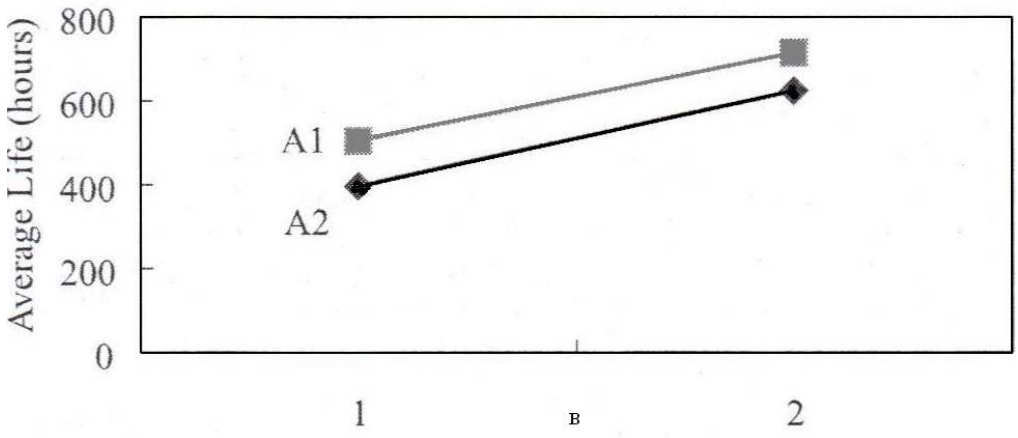
ΠΙΝΑΚΑΣ 4.23

ANOVA

Πηγή	SS	df	MS	F	p-value
A	79003.13	1	79003.13	0.83488	0.368667
B	385003.1	1	385003.1	4.068591	0.053367
AB	703.125	1	703.125	0.00743	0.931921
Σφάλμα	2649588	28	94628.13	-	-
Ολική	3114297	31	-	-	-

Στην περίπτωση αυτή, η p-τιμή για τον παράγοντα Β είναι σχεδόν ίση με τη στάθμη σημαντικότητας α=0.05, κάτι που υποδεικνύει πως ο παράγοντας Β είναι σημαντικός. Η αλληλεπίδραση των παραγόντων Α και Β έχει p-τιμή 0.93 >> 0.05, άρα δεν επιδρά σημαντικά στο χρόνο ζωής των πυκνωτών.

Ας δούμε, λοιπόν, και στην περίπτωση των πλήρων πειραματικών δεδομένων το διάγραμμα των κύριων επιδράσεων των παραγόντων (γράφημα 4.8):



Γράφημα 4.8

Όπως φαίνεται ξεκάθαρα από το διάγραμμα αυτό, ο βέλτιστος συνδυασμός επιπέδων παραγόντων προκύπτει και σε αυτή την περίπτωση ο A_1B_2. Επομένως, τα συμπεράσματα της ανάλυσης με πλήρη και με αποκομμένα δεδομένα στο παρόν παράδειγμα ταυτίζονται κι έτσι καταλήγουμε να θεωρήσουμε αξιόπιστη τη μέθοδο ανάλυσης που χρησιμοποιήθηκε στην ενότητα 4.4.4.2.

4.5 Συμπεράσματα

Η ποικιλία των μεθόδων ανάλυσης πειραματικών σχεδιασμών στους οποίους εμπεριέχονται αποκομμένες παρατηρήσεις αποδεικνύεται πως είναι αρκετά μεγάλη. Όπως προέκυψε από την προηγούμενη ανάλυση, κάποιες από τις μεθόδους αυτές, όπως για παράδειγμα η μέθοδος “grey prediction”, έχουν αρκετά εύκολη εφαρμογή χωρίς να απαιτείται ιδιαίτερα επιστημονικό υπόβαθρο από την πλευρά του ερευνητή. Το γεγονός αυτό αυξάνει κατά πολύ τη χρηστική τους αξία.

Επίσης, εξαιρετικά χρηστικές φαίνεται πως είναι οι μέθοδοι ανάλυσης αποκομμένων δεδομένων που προκύπτουν από τον παραμετρικό σχεδιασμό του Taguchi, λόγω κυρίως της πολύ μεγάλης γκάμας βιομηχανικών πειραμάτων στα όποια δίνουν επαρκή και αξιόπιστα συμπεράσματα.

Γενικά, οι μέθοδοι ανάλυσης σχεδιασμών με αποκομμένες παρατηρήσεις αποτελούν σημαντικό και αναπόσπαστο κομμάτι του συνόλου των μεθόδων στατιστικής συμπερασματολογίας, ενώ η ευρεία χρήση τους συμβάλλει σημαντικά στη βελτίωση της παραγωγικής διαδικασίας και κατά συνέπεια στη βελτίωση των σύγχρονων συνθηκών ζωής.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Ανδρουλάκης Ε., (2008) Μεταπτυχιακή διπλωματική εργασία: "Μέθοδοι επιλογής μεταβλητών στο μοντέλο αναλογικής διακινδύνευσης του Cox και εφαρμογές σε πραγματικά ιατρικά δεδομένα με αποκομμένες παρατηρήσεις", pp.75-83.
- Βασδέκης Β., "Σημειώσεις στα εφαρμοσμένα γραμμικά μοντέλα" , pp.67-81.
- Berkson J., (2005) "Encyclopedia of Biostatistics".
- Bier V., Cox L. Jr, (2002) "Probabilistic Risk Analysis for Engineered Systems".
- Blunt J., Balchin N.C., (2002) "Health and Safety in Welding and Allied Processes".
- Βούρος Δ., (2007) Διπλωματική εργασία: "Αποτίμηση ρίσκου με εφαρμογή στον ελληνικό θαλάσσιο χώρο".
- Box G.E., Hunter J.S, Hunter W.G, (2005) "Statistics for Experimenters: Design, Innovation, and Discovery".
- Breslow N., (1974) "Covariance analysis of censored survival data. *Biometrics*", pp.89-100 .
- Collet D., (2003) "Modelling Survival Data in Medical Research", pp.1-54.
- Condra L.W., (1993) "Reliability improvement with design of experiments".
- Cox D.R., (1972) "Regression models and life-tables", pp. 187-220.
- Cox D.R., (1975) "Partial likelihood", pp. 269-276.
- D'Agostino R.B, Stephens M.A, (1986) "Goodness-of-fit Techniques".
- Δαμιανού Χ., Παπαδάτος Ν., Χαραλαμπίδης Χ., (2003) "Εισαγωγή στις πιθανότητες και τη στατιστική", σελ.133-135 .
- Deng J.L., (1982) "Control problems of grey system", pp.288-294.
- Deng J.L., (1989) "Introduction to grey system theory", pp.1-24.

- Deng J.L., (1993) “On judging the admissibility of grey modeling via class ratio”, pp.249-252.
- Fisher R.A., (1925) “Statistical Methods for Research Workers”.
- Hahn G.J., Morgon C.B., Schmee J., (1981) “The analysis of a fractional factorial experiment with censored data using iterative least square”, pp.33-36.
- Hamada M., Tse S.K, (1988), “A Note on the Existence of Maximum Likelihood, Estimates in Linear Regression Models Using Interval-Censored Data”, pp.293-296.
- Hamada M., Wu C.F.G, (1991) “Analysis of censored data from highly fractionated experiments”, pp.25-38.
- Hamada M., Wu C.F.G., (1995) “Analysis of censored data from fractionated experiments: a Bayesian approach”, pp.467-477.
- Hamada M., Wu C.F.G, (2000) “Experiments: Planning, Analysis and Parameter Design Optimization”.
- Jones B., Nachtsheim C.J., (2009) “Split-Plot Designs :What, Why, and How”, pp.340-360.
- Kaplan, E.L., Meier, P., (1958) “Non-parametric estimation from incomplete observations.” pp. 457-481.
- Καρώνη Χ., (2005) “Μοντέλα αξιοπιστίας και Επιβίωσης” .
- Kleinbaum DG., (2006) “Statistics in the health sciences: Survival analysis”.
- Κοκολάκης Γ., Σπηλιώτης Ι., (2002) “Εισαγωγή στη Θεωρία Πιθανοτήτων και Στατιστική”.
- Κουκουβίνος Χ., (2005) “Γραμμικά μοντέλα και σχεδιασμοί”, pp.241-260.
- Κουτσογιάννης Δ., (2008) “Σύγχρονες τάσεις στην εκτίμηση των βροχοπτώσεων”.
- Kuehl R.O., (2000) “Design of Experiments: Statistical Principles of Research Design and Analysis”, pp.469-472.
- Lee E.T., Wang J.W., (2003) “Statistical methods for survival data analysis”, pp.134-196.
- Limpert E., Stahel W., Abbt M., (2001) “Log-normal Distributions across the Sciences: Keys and Clues”, pp.341-352.

- Lopez A.M., (2009) “Analisis estadisticos de datos de vida”, pp.7-16.
- Montgomery D.C., (2001) “Design and Analysis of experiments”.
- Lu J.C., (1992) “Analysis of location and dispersion effects based on censored data from unreplicated experiments”, pp.4-21.
- Lu J.C., Unal C., (1994) “Process characterization and optimization based on censored data from highly fractionated experiments”, pp.145-155.
- Molinero L.M., (2001) “Modelos de regresión de Cox para el tiempo de supervivencia”.
- Nelson W., (1982) “Applied life data analysis”, pp.319-326.
- Οικονόμου Π., Καρώνη Χ., “Στατιστικά μοντέλα παλινδρόμησης”, pp.25-45.
- Ott R.L, Longnecker M., (2010) “An introduction to Statistical Methods and Data Analysis”, pp.1095-1101.
- Παύλου Ε., (2006) Μεταπτυχιακή Διπλωματική εργασία: “Το μοντέλο αναλογικού κινδύνου του Cox στην ανάλυση επιβίωσης”, pp.76-77.
- Peace G.S, (1993) “Taguchi methods: A Hands-On approach”.
- Plackett R.L., Burman J.P., (1946) "The Design of Optimum Multifactorial Experiments”, pp.305-325.
- Rossi M., Zhang J.Y., Steenaart W., (2001) “Iterative least squares: Design of perfect reconstruction”.
- Saville D.J., Wood G.R., “Split -plot design”, pp.7-34.
- Schmee J., Hahn G.J, (1979) “A simple method for regression analysis with censored data”, pp.417-434.
- Smith D.N., (1997), “Εφαρμοσμένη Ανάλυση Παλινδρόμησης”.
- Specht N., (1985), “Heat Exchanger Product Design via Taguchi Methods”, pp.302-318.
- Sreenivas R., Kumar C.G, Prakasham R.S., Hobbs P.J., (March 2008) “The Taguchi methodology as a statistical tool for biotechnological applications: A critical appraisal” Biotechnology Journal, pp.510-523.
- Styan G. P. H., Boyer C., Chu K.L., (2008) “Some comments on Latin squares and on Greco-Latin squares, illustrated with postage stamps and old playing cards”.

- Taguchi S., Byrne D.M., (1987) “The Taguchi approach to parameter design”, pp.19-26.
- Tinsson W., (2010) “Plans d’experience: constructions et analyses statistiques”, pp.1-8.
- Tong L.I., Su C.T., (1996) “A non-parametric method for experimental analysis with censored data”, pp.456-463.
- Tong L.I., Yang C.H., (2006) “Analyzing Type II Censored Data Obtained from Repetitious Experiments”, pp.50-62.
- Tong L.I., Wang C.H., Hsiao L.C., (2005) “Optimizing processes based on censored data obtained in repetitious experiments using grey prediction”, pp. 991-998.
- Torres V.A., (1993) “Simple analysis of unreplicated factorials with possible abnormalities”, pp.183-187.
- Τσάντας Ν., Μωυσιάδης Χ., Μπαγιάτης Ν., Χατζηπαντελής Θ., (1999) “Ανάλυση δεδομένων με τη βοήθεια στατιστικών πακέτων”, pp.14-48.
- Vial J., Jardy A., (1998) “Utilisation des plans d’experiences pour evaluer la robustesse d’une methode d’analyse quantitative”, pp.15-24.

- British Medical Journal, (1995)
- Canadian Electronics Magazine, (2009)
- Health Service Journal, (2009)

- http://www.statsoft.fr
- http://www.statgraphics.fr
- http://www.slideshare.net/