



Εισαγωγή στη Στατιστική

Τάκη Παπαϊωάννου Σωτήρη Λουκί



ΒΒΛΙΟΘΗΚΗ
ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΙΩΑΝΝΙΝΩΝ



026000010088

102/95

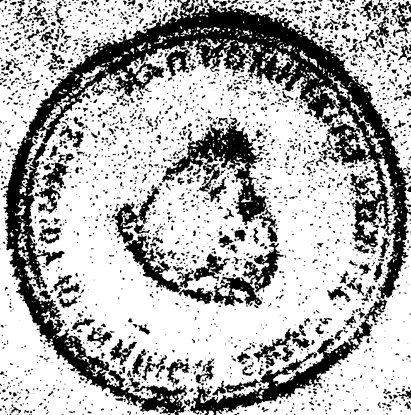


140



ΕΙΣΟΔΥΜΑΤΑ ΠΡΩΤΟ ΕΠΙΣΤΗΜΟΝ





Τμήμα Παιδαγωγικών

Εκπαιδευτική Επιστήμη

Πανεπιστήμιο Κρήτης

Εκδόθηκε το 1990

Αριθμός Έκδοσης: 1

Εκδόθηκε στην Αθήνα

ΕΙΣΑΓΩΓΗ ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ

ΕΙΣΑΓΩΓΗ ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ

1990

Αθήνα



ΠΡΩΤΟΤΥΠΟ ΠΤΟ ΠΥΘΥΡΡΙΣ

Τάκη Παπαιωάννου

Καθηγητή Στατιστικής

Πανεπιστημίου Ιωαννίνων

Σωτήρη Β. Λουκά

Αν. Καθηγητή Στατιστικής

Πανεπιστημίου Ιωαννίνων

Εισαγωγή στη Στατιστική



Ιωάννινα 1990

Δεύτερη Ανατύπωση Μάρτιος 1994



Κάθε γνήσιο αντίτυπο φέρει την υπογραφή των συγγραφέων

102/95

Απαγορεύεται η ολική ή μερική ανατύπωση του παρόντος
με οποιοδήποτε τρόπο χωρίς την έγγραφη άδεια των συγγραφέων

Copyright 1990 by : Takis Papaioannou

Sotiris Loukas

This book or any part thereof must not be reproduced in any
form without the written permission of the authors.

PRINTED IN IOANNINA GREECE

6 5 4 3



ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1

Πιθανότητες-Τυχαιες Μεταβλητες-Κατανομες

1.1	Γενικότητες	1
1.2	Πιθανότητα	3
1.3	Τυχαιές Μεταβλητές - Κατανομές Πιθανότητας - Μετρήσεις.	7
1.4	Πληθυσμός - Τυχαίο Δείγμα - Κατανομή Δείγματος	11
1.5	Ειδικές Κατανομές	14
1.6	Χαρακτηριστικά των Τυχαίων Μεταβλητών και Κατανομών	19
1.7	Γεννήτριες Συναρτήσεις	22

ΚΕΦΑΛΑΙΟ 2

Περιγραφική Στατιστική

2.1	Συχνότητα -Σχετική συχνότητα-Αθροιστική συχνότητα	27
2.2	Περιγραφή και σύμπτυξη των αριθμητικών δεδομένων	29
2.2.1	Πίνακες συχνοτήτων	30
2.2.2	Γραφικές μέθοδοι σύμπτυξης	35
2.2.3	Αριθμητικά μεγέθη - Χαρακτηριστικά των μετρήσεων	40
ΑΣΚΗΣΕΙΣ		54

ΚΕΦΑΛΑΙΟ 3

Στατιστικά-Δειγματικές Κατανομες

3.1	Δειγματοληψία και Συμπερασματολογία Δειγματικές κατανομές	57
3.2	Μέσες τιμές και διακυμάνσεις των $\bar{X}$ και S^2	59
3.3	Κατανομές που απορρέουν από την κανονική	62
3.4	Δειγματοληψία από κανονικούς πληθυσμούς	69
3.5	Κεντρικό Οριακό Θεώρημα	74
3.6	Κανονική προσέγγιση της διωνυμικής κατανομής	81
ΑΣΚΗΣΕΙΣ		86



ΚΕΦΑΛΑΙΟ 4**Στατιστική Συμπερασματολογία**

4.1	Εκτίμηση Παραμέτρων	94
4.1.1	Σημειοεκτιμητική	95
4.1.2	Διαστήματα εμπιστοσύνης	98
4.2	Έλεγχος Στατιστικών Υποθέσεων Στατιστικά τεστ	103
4.2.1	Αξιολόγηση των τεστ-Σφάλματα τύπου I και τύπου II	105
4.2.2	Τα στοιχεία του τεστ	109
4.2.3	Γενικές παρατηρήσεις	113
4.2.4	Τιμή P	114
4.3	Απλή Στατιστική Συμπερασματολογία	115
4.3.1	Συμπερασματολογία για τη μέση τιμή	115
4.3.2	Συμπερασματολογία για το διωνυμικό p	121
4.3.3	Συμπερασματολογία για τη διαφορά δύο μέσων τιμών	125
4.3.4	Συγκρίσεις κατά ζεύγη	135
4.3.5	Συμπερασματολογία για τις διακυμάνσεις	139
4.3.6	Συμπερασματολογία για τη διαφορά δύο ποσοστών (πιθανοτήτων επιτυχίας)	144
4.3.7	Άλλα θέματα	151
	ΑΣΚΗΣΕΙΣ	152

ΚΕΦΑΛΑΙΟ 5**Γραμμική Παλινδρόμηση - Συσχετισή**

5.1	Γενικότητες	161
5.2	Σχέσεις μεταξύ μεταβλητών	162
5.2.1	Συναρτησιακή σχέση μεταξύ δύο μεταβλητών	162
5.2.2	Στατιστική σχέση μεταξύ δύο μεταβλητών	163
5.3	Μοντέλα παλινδρόμησης και χρησιμότητα τους	166
5.4	Απλή γραμμική παλινδρόμηση	169
5.5	Εκτίμηση της συνάρτησης παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων	171



5.6	Ιδιότητες των εκτιμητών ελαχίστων τετραγώνων	178
5.7	Στατιστική συμπερασματολογία	180
5.8	Διάστημα εμπιστοσύνης για την συνάρτηση παλινδρόμησης	182
5.9	Ερμηνεία της στατιστικής συμπερασματολογίας	183
5.10	Η ανάλυση της διακύμανσης στη παλινδρόμηση	184
5.11	Έλεγχος για την καταλληλότητα του μοντέλου παλινδρόμησης- Ανάλυση των Υπολοίπων	189
5.11.1	Έλεγχος για την γραμμικότητα της συνάρτησης παλινδρόμησης	191
5.11.2	Έλεγχος για την σταθερότητα της διακύμανσης των σφαλμάτων	192
5.11.3	Έλεγχος για την ανεξαρτησία των σφαλμάτων	193
5.11.4	Έλεγχος για την κανονικότητα των σφαλμάτων	195
5.11.5	Έλεγχος για την παρουσία ακραίων παρατηρήσεων	196
5.12	Μέτρα επαναφοράς στην καταλληλότητα του μοντέλου	197
5.12.1	Λογαριθμικός μετασχηματισμός	197
5.12.2	Αντίστροφος μετασχηματισμός	198
5.13	Πολλαπλή γραμμική παλινδρόμηση	200
5.14	Συσχέτιση	201
5.15	Οι ποιοτικές μεταβλητές στην παλινδρόμηση	205
5.16	Στατιστικά πακέτα και παλινδρόμηση	206
	ΑΣΚΗΣΕΙΣ	210

ΚΕΦΑΛΑΙΟ 6

Ανάλυση της Διακύμανσης

6.1	Γενικότητες	215
6.2	Σχέση μεταξύ παλινδρόμησης και ανάλυσης της διακύμανσης	216
6.3	Μοντέλα Ανάλυσης της Διακύμανσης και χρησιμότητά τους	219



6.4	Ανάλυση της Διακύμανσης κατά ένα παράγοντα	221
6.4.1	Το στατιστικό μοντέλο	222
6.4.2	Εκτίμηση των παραμέτρων με τη μέθοδο των ελαχίστων τετραγώνων	223
6.4.3	Η Ανάλυση της Διακύμανσης	228
6.4.4	Ανάλυση των μέσων τιμών	234
6.4.5	Έλεγχος για την καταλληλότητα του μοντέλου	239
6.4.6	Στατιστικά πακέττα	242
6.5	Αρχές σχεδιασμού πειραμάτων κατά ένα και δύο παράγοντες	244
6.6.	Ανάλυση της Διακύμανσης κατά δύο παράγοντες	249
6.6.1	Το στατιστικό μοντέλο	250
6.6.2	Εκτίμηση των παραμέτρων με τη μέθοδο των ελαχίστων τετραγώνων	251
6.6.3	Η Ανάλυση της Διακύμανσης	254
6.6.4.	Ανάλυση των μέσων τιμών	259
6.6.5	Έλεγχος για την καταλληλότητα του μοντέλου	260
6.6.6	Στατιστικά πακέττα	261
ΑΣΚΗΣΕΙΣ		262

ΚΕΦΑΛΑΙΟ 7

Μη Παραμετρική Στατιστική

7.1	Γενικότητες	267
7.2	Κλίμακες μετρήσεων	268
7.3	Ένας πληθυσμός - Ένα δείγμα	270
7.3.1	Το χ^2 τεστ καλής προσαρμογής	271
7.3.2	Το τεστ των Kolmogorov - Smirnov	272
7.3.3	Το προσημικό τεστ	274
7.4	Δύο πληθυσμοί - Δύο δείγματα (συσχετισμένα)	277
7.4.1	Προσημικό τεστ για συγκρίσεις κατά ζεύγη	277
7.4.2	Προσημικό τεστ των τάξεων του Wilcoxon	279



7.5 Δύο πληθυσμοί - Δύο δείγματα (ανεξάρτητα)	281
7.5.1 χ^2 τεστ ομογένειας ποσοστών - ανεξαρτησίας	281
7.5.2 Το τεστ των Kolmogorov - Smirnov	284
7.5.3 Το τεστ των Wilcoxon - Mann - Whitney	286
7.6 Μέτρα συσχέτισης	289
7.6.1 Συντελεστής συσχέτισης του Spearman	289
ΑΣΚΗΣΕΙΣ	293
 ΠΑΡΑΡΤΗΜΑ	 295
 ΒΙΒΛΙΟΓΡΑΦΙΑ	 309
 ΑΠΑΝΤΗΣΕΙΣ ΣΤΙΣ ΑΣΚΗΣΕΙΣ	 313
 ΕΥΡΕΤΗΡΙΟ	 321



[illegible]

ΠΡΟΛΟΓΟΣ

Οι πιθανότητες και η Στατιστική είναι ένας ενδιαφέρον και γοητευτικός κλάδος των Εφαρμοσμένων Μαθηματικών με μεγάλη ανάπτυξη τόσο στη θεωρία όσο και τις εφαρμογές. Ένας δείκτης της ανάπτυξης αυτής είναι οι πολυάριθμες Σχολές και Τμήματα ή Ινστιτούτα Στατιστικής τα οποία ξεπήδησαν τα τελευταία τριάντα χρόνια από τα Μαθηματικά κυρίως Τμήματα σε πολλά Πανεπιστήμια ανά τον κόσμο. Στη χώρα μας οι Πιθανότητες και η Στατιστική διδάσκονται στα Μαθηματικά Τμήματα των Πανεπιστημίων από τις αρχές της δεκαετίας του 1960. Διδάσκονται επίσης και σε άλλα Τμήματα των Σχολών Θετικών Επιστημών, στα Πολυτεχνεία, στις Νομικές, Οικονομικές, Βιομηχανικές, Γεωπονικές, Ιατρικές και Φιλοσοφικές Σχολές των Πανεπιστημίων.

Οι Πιθανότητες και η Στατιστική αποτελούν ένα απαραίτητο εργαλείο στα χέρια κάθε ερευνητή. Δημοσκοπήσεις, έρευνες αγοράς, ιατρικές μελέτες, ποιοτικοί έλεγχοι, προβλέψεις και γενικά αναλύσεις μετρήσεων (δεδομένων) χρησιμοποιούν μεθόδους των Πιθανοτήτων και της Στατιστικής. Υπάρχουν πολλά βιβλία Πιθανοτήτων και Στατιστικής, κυρίως στην ξένη βιβλιογραφία. Τα βιβλία αυτά καλύπτουν πολλά θέματα και απευθύνονται σε διάφορα ακροατήρια ανάλογα με το αντικείμενο που ασχολούνται και το επίπεδο της μαθηματικής ωριμότητας του αναγνώστη.

Το βιβλίο αυτό αποτελεί μια εισαγωγή στη Στατιστική η οποία συνήθως διδάσκεται ύστερα από ένα εισαγωγικό μάθημα στις Πιθανότητες. Με εξαίρεση τις Πιθανότητες δεν απαιτεί προηγούμενη γνώση του θέματος. Απευθύνεται κυρίως στους φοιτητές των Μαθηματικών Τμημάτων των Πανεπιστημίων. Απευθύνεται επίσης και στους φοιτητές Πολυτεχνείων και άλλων Σχολών που έχουν το απαιτούμενο υπόβαθρο. Το μαθηματικό υπόβαθρο που χρειάζεται για την κατανόηση των εννοιών και μεθόδων του είναι οι λυκειακές μαθηματικές γνώσεις. Κατά το μεγαλύτερο μέρος είναι αυτάρκες. Οι έννοιες και τα θέματα του εισάγονται περισσότερο διεξοδικά και με παραδείγματα παρά αυστηρά, αυθαίρετα και αυθεντικά. Τα επιχειρήματα και οι αποδείξεις στηρίζονται στην μαθηματική και την πιθανοθεωρητική λογική, τις προτάσεις και τα θεωρήματα που σταδιακά παρουσιάζονται και τη διαίσθηση. Η μαθηματική αυστηρότητα διατηρείται όπου αυτό είναι δυνατό και σκόπιμο. Οι



συγγραφείς πιστεύουν ότι η μελέτη ενός μαθηματικού κλάδου γίνεται καλύτερη όταν προηγείται η κατανόηση των εννοιών και των μεθόδων και μετά ακολουθεί η μαθηματική αυστηρότητα. Αυτό ιδιαίτερα ισχύει για τη Στατιστική η οποία είναι η κατ'εξοχήν επιστήμη που ασχολείται με το να δίνει νόημα και σημασία σε αριθμητικά κυρίως δεδομένα σε αντίθεση με τα Μαθηματικά που ασχολούνται με το να αποδεικνύουν θεωρήματα και λήμματα. Το βιβλίο αυτό γραμμένο κάτω από τη φιλοσοφία αυτή δίνει έμφαση τόσο στη θεωρία όσο και στις εφαρμογές.

Η ύλη του βιβλίου αναπτύσσεται σε επτά κεφάλαια:

Το πρώτο κεφάλαιο είναι εισαγωγικό. Περιγράφει σε γενικές γραμμές τον κλάδο της Στατιστικής, ανακεφαλαιώνει τις έννοιες της πιθανότητας της τυχαίας μεταβλητής και της κατανομής πιθανότητας, παρουσιάζει τις έννοιες του πληθυσμού, του τυχαίου δείγματος, της κατανομής του δείγματος και δίνει συνοπτικά αρκετές χρήσιμες ειδικές κατανομές μαζί με τα βασικά χαρακτηριστικά τους.

Η Περιγραφική Στατιστική παρουσιάζεται στο δεύτερο κεφάλαιο.

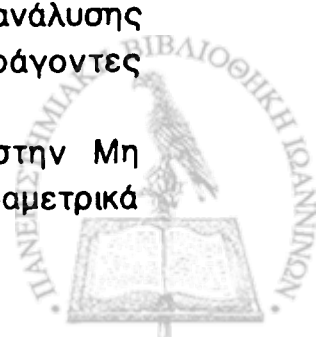
Στο τρίτο κεφάλαιο θέτονται τα θεωρητικά θεμέλια της Στατιστικής. Εισάγονται οι έννοιες του στατιστικού, της δειγματικής κατανομής, οι κατανομές χ^2 , t και F , η δειγματοληψία από κανονικούς πληθυσμούς και το κεντρικό οριακό θεώρημα.

Το τέταρτο κεφάλαιο αποτελεί μια εισαγωγή στη Στατιστική Συμπερασματολογία. Αναπτύσσονται πρώτα οι αρχές, οι μέθοδοι εκτίμησης στατιστικών παραμέτρων και στη συνέχεια δίνονται η φιλοσοφία και η μεθοδολογία των στατιστικών τεστ. Τέλος δίνεται μια σειρά εφαρμογών, η λεγόμενη απλή Στατιστική Συμπερασματολογία για ένα ή δύο πληθυσμούς.

Η Γραμμική Παλινδρόμηση και η Συσχέτιση παρουσιάζονται στο πέμπτο κεφάλαιο. Αναπτύσσονται η απλή γραμμική παλινδρόμηση, η μέθοδος των ελαχίστων τετραγώνων, η εκτίμηση και τα τεστ για τη συνάρτηση παλινδρόμησης, ο έλεγχος καταλληλότητας του μοντέλου, η ανάλυση των υπολοίπων και άλλες χρήσιμες μεθοδολογίες.

Το έκτο κεφάλαιο αποτελεί μία εισαγωγή στην Ανάλυση της Διακύμανσης. Παρουσιάζεται εκεί η μεθοδολογία της ανάλυσης της διακύμανσης κατά ένα παράγοντα και κατά δύο παράγοντες χωρίς αλληλοεπίδραση.

Το έβδομο κεφάλαιο αποτελεί μια εισαγωγή στην Μη Παραμετρική Στατιστική παρουσιάζοντας τα βασικά μη παραμετρικά



τεστ όπως το τεστ καλής προσαρμογής, το τεστ των Kolmogorov-Smirnov, το προσημικό τεστ, το τεστ των τάξεων και άλλα.

Στο τέλος του βιβλίου δίδονται οι στατιστικοί πίνακες, η σχετική βιβλιογραφία, οι απαντήσεις στις ασκήσεις και το ευρετήριο. Οι ορισμοί, τα θεωρήματα, τα πορίσματα, τα παραδείγματα, τα σχήματα και οι ασκήσεις έχουν χωριστή αρίθμηση σε κάθε κεφάλαιο. Έτσι, θεώρημα 4.3 σημαίνει το τρίτο θεώρημα στο τέταρτο Κεφάλαιο. Το βιβλίο περιέχει 24 παραδείγματα που αναπτύσσονται με κάθε λεπτομέρεια και 78 ασκήσεις, θεωρητικές και εφαρμοσμένες, που εμφανίζονται στο τέλος κάθε κεφαλαίου.

Το βιβλίο αυτό αποτελεί εξέλιξη και συμπλήρωση του δεύτερου μέρους (με υπότιτλο Στατιστική) του βιβλίου του Τ. Παπαϊωάννου με τίτλο "Εισαγωγή στις Πιθανότητες και τη Στατιστική" που εκδόθηκε για πρώτη φορά το 1981 και σε δεύτερη έκδοση το 1982. Θα ακολουθήσει ξεχωριστός τόμος με τίτλο "Εισαγωγή στις Πιθανότητες". Είναι προϊόν των παραδόσεων των συγγραφέων σε φοιτητές ΑΕΙ για πολλά χρόνια. Ξεκίνησε με μορφή σημειώσεων για τους πρωτοετείς και δευτεροετείς φοιτητές του Μαθηματικού Τμήματος του Πανεπιστημίου Ιωαννίνων. Οι σημειώσεις αυτές δουλεύτηκαν και ζυμώθηκαν με τη βοήθεια ευστόχων παρατηρήσεων φοιτητών και συναδέλφων των συγγραφέων. Σε όλους αυτούς απευθύνουμε τις ευχαριστίες μας. Η ευθύνη για αβλεψίες ή λάθη είναι δική μας. Κάθε καλοπροαίρετη υπόδειξη και κριτική είναι ευπρόσδεκτη.

Ευχαριστίες απευθύνονται στους Ρόη Λαμπρίδη, Ράνια Πραμαντιώτη και Ιωάννη Ρίζο για την επιμέλεια της έκδοσης. Επίσης ευχαριστίες απευθύνονται στον Μαθηματικό Νικόλαο Αντωνίου που φιλοτέχνησε το εξώφυλλο και στο Τυπογραφείο του Πανεπιστημίου Ιωαννίνων για την επιμέλεια της εκτύπωσης.

Τάκης Παπαϊωάννου

Σωτήρης Λουκάς

Ιωάννινα, Ιανουάριος 1990



1. The first part of the report is a general statement of the purpose of the study. The purpose of the study is to determine the effect of the new teaching method on the learning of the subject.

[illegible][illegible]

14-00000

УЧЕБНО-МЕТОДИЧЕСКОЕ ПОСОБИЕ

THE UNITED STATES OF AMERICA

20 AUG 1961
The following information is available for the above mentioned

SECRET

The following paragraphs are related to the analysis

1. The following are the names of the persons who have been identified as having been in contact with the subject of this investigation:

To: 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026 2027 2028 2029 2030 2031 2032 2033 2034 2035 2036 2037 2038 2039 2040 2041 2042 2043 2044 2045 2046 2047 2048 2049 2050 2051 2052 2053 2054 2055 2056 2057 2058 2059 2060 2061 2062 2063 2064 2065 2066 2067 2068 2069 2070 2071 2072 2073 2074 2075 2076 2077 2078 2079 2080 2081 2082 2083 2084 2085 2086 2087 2088 2089 2090 2091 2092 2093 2094 2095 2096 2097 2098 2099 2100 2101 2102 2103 2104 2105 2106 2107 2108 2109 2110 2111 2112 2113 2114 2115 2116 2117 2118 2119 2120 2121 2122 2123 2124 2125 2126 2127 2128 2129 2130 2131 2132 2133 2134 2135 2136 2137 2138 2139 2140 2141 2142 2143 2144 2145 2146 2147 2148 2149 2150 2151 2152 2153 2154 2155 2156 2157 2158 2159 2160 2161 2162 2163 2164 2165 2166 2167 2168 2169 2170 2171 2172 2173 2174 2175 2176 2177 2178 2179 2180 2181 2182 2183 2184 2185 2186 2187 2188 2189 2190 2191 2192 2193 2194 2195 2196 2197 2198 2199 2200 2201 2202 2203 2204 2205 2206 2207 2208 2209 2210 2211 2212 2213 2214 2215 2216 2217 2218 2219 2220 2221 2222 2223 2224 2225 2226 2227 2228 2229 2230 2231 2232 2233 2234 2235 2236 2237 2238 2239 2240 2241 2242 2243 2244 2245 2246 2247 2248 2249 2250 2251 2252 2253 2254 2255 2256 2257 2258 2259 2260 2261 2262 2263 2264 2265 2266 2267 2268 2269 2270 2271 2272 2273 2274 2275 2276 2277 2278 2279 2280 2281 2282 2283 2284 2285 2286 2287 2288 2289 2290 2291 2292 2293 2294 2295 2296 2297 2298 2299 2300 2301 2302 2303 2304 2305 2306 2307 2308 2309 2310 2311 2312 2313 2314 2315 2316 2317 2318 2319 2320 2321 2322 2323 2324 2325 2326 2327 2328 2329 2330 2331 2332 2333 2334 2335 2336 2337 2338 2339 2340 2341 2342 2343 2344 2345 2346 2347 2348 2349 2350 2351 2352 2353 2354 2355 2356 2357 2358

100

ΚΕΦΑΛΑΙΟ 1

πιθανότητες τυχαιες μεταβλητες κατανομες

1.1 Γενικότητες

Στατιστική είναι η επιστήμη η οποία ασχολείται με τον σχεδιασμό πειραμάτων, τη συλλογή και ανάλυση αριθμητικών δεδομένων και την εξαγωγή συμπερασμάτων. Τα συμπεράσματα αναφέρονται σε άγνωστα χαρακτηριστικά ή ιδιότητες πληθυσμών και εξαγονται με την βοήθεια των πληροφοριών που περιέχονται σε δείγματα από τους πληθυσμούς αυτούς. Στα πειράματα περιλαμβάνονται και οι δειγματοληπτικές έρευνες, οι σφιγμομετρήσεις της κοινής γνώμης καθώς επίσης και οι έρευνες αγοράς. Το θεωρητικό υπόβαθρο της Στατιστικής είναι η Θεωρία των Πιθανοτήτων.

Η Στατιστική μαζί με την Θεωρία των Πιθανοτήτων χρησιμοποιείται σε όλους τους κλάδους της ανθρώπινης δραστηριότητας και γνώσης όπως π.χ. οι επιχειρήσεις, η διοίκηση, η εκπαίδευση, η ψυχολογία, η ιατρική, η γεωπονική, τα οικονομικά, η βιολογία, η γενετική, η λογοτεχνία κ.λ.π. Με την πάροδο του χρόνου και την εξάπλωση των ηλεκτρονικών υπολογιστών η εφαρμογή και η ανάπτυξη της Στατιστικής γίνεται ολοένα και μεγαλύτερη. Ιδιαίτεροι κλάδοι έχουν δημιουργηθεί από την χρησιμοποίηση των στατιστικών

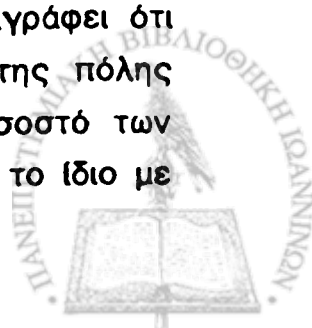


μεθόδων και εννοιών στην Οικονομία, την Βιολογία και την Ιατρική, την Ψυχολογία κ.λ.π. όπως η Οικονομετρία, η Βιομετρία, η Βιοστατιστική, η Ψυχομετρία κ.λ.π.

Συχνά η Στατιστική θεωρείται συνώνυμη με πίνακες στατιστικών στοιχείων, με ιστογράμματα ή άλλα γραφήματα και με εικόνες, με ποσοστά, με μέσους όρους κ.λ.π. Η αντίληψη αυτή είναι πεπαλαιωμένη και αγνοεί τις μεγάλες κατακτήσεις του κλάδου τα τελευταία χρόνια. Σ' αυτές περιλαμβάνονται έννοιες και θέματα όπως διαστήματα εμπιστοσύνης, τέστ σημαντικότητας, παλινδρόμηση, προβλέψεις, συντελεστές συσχέτισης, στοχαστικά μοντέλα κ.λ.π. Τα θέματα που αναφέρονται σε πινακοποίηση δεδομένων, παραστάσεις των δεδομένων με γραφήματα ή εικόνες και υπολογισμό περιγραφικών μέτρων αποτελούν αντικείμενο της Περιγραφικής Στατιστικής και καλύπτονται στο επόμενο κεφάλαιο.

Όταν τα άτομα ή υλικά που διερευνώνται έχουν μεγάλη μεταβλητότητα ή όταν ο ελεγχόμενος πειραματισμός είναι αδύνατος, (π.χ. σφυγμομέτρηση κοινής γνώμης, έρευνες αγοράς, κοινωνιολογία, οικονομία κ.λ.π.) η στατιστική σχεδίαση των δειγματοληπτικών ερευνών παρέχει στους ερευνητές μεθόδους χωρίς τις οποίες θα ήταν αδύνατον να βγάλουν οποιοδήποτε συμπέρασμα.

Η σύγχρονη Στατιστική ασχολείται κυρίως με τη λεγόμενη Στατιστική Συμπερασματολογία. Εφ' όσον δεν χρησιμοποιούμε τα μέτρα που υπολογίζουμε για να κάνουμε γενικεύσεις, τότε απλώς περιγράφουμε ότι παρατηρούμε. Αλλά ευθύς ως κάνουμε μια επαγωγική γενίκευση τότε αφήνουμε την περιγραφή και μπαίνουμε στο χώρο της συμπερασματολογίας (επαγωγής). Ένας δημοσιογράφος, για παράδειγμα, κάνει μια σφυγμομέτρηση της κοινής γνώμης σε μια πόλη, και ερωτά εκατό ανθρώπους αν υποστηρίζουν μία αλλαγή σε κάποιο νόμο. Εβδομήντα πέντε, από αυτούς, αποκρίνονται ότι υποστηρίζουν την αλλαγή. Αν ο δημοσιογράφος γράψει ότι εβδομήντα πέντε (75) από εκατό (100) άτομα, που ερωτήθηκαν, υποστηρίζουν την αλλαγή, τότε απλώς περιγράφει ότι παρατήρησε. Αν όμως γράψει ότι 75% των κατοίκων της πόλης υποστηρίζουν την αλλαγή, τότε συμπεραίνει ότι, το ποσοστό των κατοίκων της πόλης, που υποστηρίζουν την αλλαγή, είναι το ίδιο με



εκείνο του μικρού δείγματός του. Ξεκινώντας από την πληροφορία που του δίνει το δείγμα διατυπώνει μια πρόταση για ένα μεγαλύτερο σύνολο από το οποίο πάρθηκε το δείγμα. Στην πραγματικότητα είναι ίσως δύσκολο να μη γενικεύσει έστω και υποσυνείδητα.

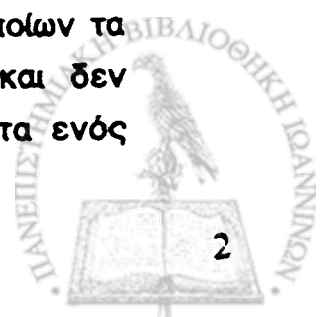
Η παραπάνω στατιστική συμπερασματολογία έχει ισχύ, αν σωστές στατιστικές αρχές εφαρμόστηκαν στην επιλογή του δείγματος και του εκτιμητή. Αν, για παράδειγμα, το δείγμα του δημοσιογράφου ήταν τυχαίο (λεπτομέρειες για τυχαία δείγματα δίνονται στη § 1.4), τότε ορθώς μπορεί να υποθέσει ότι το αληθινό ποσοστό για όλη την κοινότητα που υποστηρίζει την αλλαγή πρέπει να ευρίσκεται γύρω στο 75%. Μήπως είναι 65% ή 85%; Είναι δύσκολο να κρίνει. Η απάντηση χρειάζεται Θεωρία Πιθανοτήτων. Με τη βοήθειά της μπορούμε να κατασκευάσουμε ένα αξιόπιστο διάστημα ποσοστών εκείνων που υποστηρίζουν την αλλαγή στο νόμο.

Η Στατιστική Συμπερασματολογία περιλαμβάνει την Εκτιμητική και τα Στατιστικά Τεστ. Μέσα στο πλαίσιο αυτό αγκαλιάζει ολόκληρη τη στατιστική μεθοδολογία όπως την Παλινδρόμηση, την Ανάλυση της Διακύμανσης, την Πολυμεταβλητή Ανάλυση κ.λ.π. Μερικά από τα θέματα αυτά παρουσιάζονται στα επόμενα κεφάλαια. Διάφορες ειδικές κατανομές της Θεωρίας Πιθανοτήτων θεωρούνται επίσης γνωστές και ανακεφαλαιώνονται στο εδάφιο 1.5. Το ίδιο ισχύει και για τα χαρακτηριστικά των τυχαίων μεταβλητών και κατανομών (μέσες τιμές, διακυμάνσεις, γεννήτριες συναρτήσεις) που παρουσιάζονται περιληπτικά στα εδάφια 1.6 και 1.7.

Η Στατιστική στηρίζεται στη Θεωρία Πιθανοτήτων. Στοιχεία της θεωρίας αυτής θεωρούνται γνωστά. Για λόγους ευκολίας οι έννοιες της πιθανότητας, της τυχαίας μεταβλητής και της κατανομής πιθανότητας ανακεφαλαιώνονται στο εδάφιο 1.2 και στο πρώτο μέρος του εδαφίου 1.3.

1.2 Πιθανότητα

Τυχαία πειράματα είναι πειράματα των οποίων τα αποτελέσματα κυβερνώνται από τους νόμους της τύχης και δεν μπορούν να καθοριστούν εκ των προτέρων. Τα αποτελέσματα ενός



4 Κεφ. 1 Πιθανότητες - Τυχαιες Μεταβλητες - Κατανομες

τυχαίου πειράματος είναι γνωστά ως ενδεχόμενα ή γεγονότα. Τα ατομικά αδιαίρετα αποτελέσματα ενός τυχαίου πειράματος λέγονται στοιχειώδη ή απλά ενδεχόμενα ή δειγματικά σημεία. Το σύνολο των δειγματικών σημείων ενός τυχαίου πειράματος λέγεται δειγματικός χώρος και συμβολίζεται με το S ή το Ω . Τα ενδεχόμενα είναι υποσύνολα του S .

Αν A και B είναι δύο ενδεχόμενα τότε $A + B$ ή $A \cup B$ είναι το ενδεχόμενο που συνίσταται στην πραγματοποίηση τουλάχιστον ενός από τα A, B . $A \cap B$ ή AB είναι το ενδεχόμενο που συνίσταται στην ταυτόχρονη πραγματοποίηση των A και B . A^c ή A^c ή $\bar{A}$ είναι το αντίθετο ενδεχόμενο του A . Ασυμβίβαστα είναι δύο ενδεχόμενα για τα οποία ισχύει $AB = \emptyset$.

Ο κλασικός ορισμός της πιθανότητας έχει ως ακολούθως: Έστω A ένα ενδεχόμενο που αποτελείται από n στοιχειώδη ενδεχόμενα σε ένα δειγματικό χώρο N ισοπίθανων ενδεχομένων. Η πιθανότητα $P(A)$ του A είναι

$$P(A) = \frac{n}{N}. \quad (1.1)$$

Ο εμπειρικός ή στατιστικός ορισμός της πιθανότητας είναι: Έστω n/N η σχετική συχνότητα ενός ενδεχομένου A σε N επαναλήψεις ενός πειράματος. Η πιθανότητα $P(A)$ του A είναι

$$P(A) = \lim_{N \rightarrow \infty} \frac{n}{N}$$

Ο μαθηματικός ορισμός της πιθανότητας στηρίζεται στα ακόλουθα αξιώματα του Kolmogorov:

α. Η πιθανότητα ενός ενδεχομένου είναι ένας μη αρνητικός πραγματικός αριθμός, δηλαδή $P(A) \geq 0$ για κάθε ενδεχόμενο A .



β. $P(S) = 1.$

γ. Αν $A_1, A_2, \dots$ είναι μία πεπερασμένη ή άπειρη ακολουθία ασυμβίβαστων ανά δύο ενδεχόμενων τότε

$$P(\cup_i A_i) = \sum_i P(A_i), \quad (1.2)$$

όπου ο δείκτης i διατρέχει τους αριθμούς $1, 2, \dots, n$ ή $1, 2, \dots$

Οι κυριότερες ιδιότητες των πιθανοτήτων έχουν ως ακολούθως:-

α. Εάν A είναι ένα υποσύνολο ενός διακεκριμένου δειγματικού χώρου και $E_1, E_2, \dots$ μία πεπερασμένη ή άπειρη ακολουθία στοιχειωδών γεγονότων που αντιπροσωπεύουν το A , δηλαδή $A = \cup E_i$ τότε

$$P(A) = \sum_i P(E_i) \quad \text{ή} \quad P(A) = \sum_{s \in A} P(s),$$

όπου $P(s)$ είναι η πιθανότητα ενός δειγματικού σημείου s .

β. $0 \leq P(A) \leq 1$ για κάθε ενδεχόμενο A .

γ. $P(\bar{A}) = 1 - P(A).$

δ. Αν A και B είναι δύο ενδεχόμενα τότε

$$P(A \cup B) = P(A) + P(B) - P(AB). \quad (1.3)$$

Η δεσμευμένη ή υπό συνθήκες πιθανότητα ενός ενδεχόμενου A δοθέντος ενός άλλου ενδεχόμενου B με $P(B) > 0$ συμβολίζεται με $P(A|B)$ και ορίζεται ως



$$P(A|B) = P(AB)/P(B). \quad (1.4)$$

Άμεση απόρροια της δεσμευμένης πιθανότητας είναι ο πολλαπλασιαστικός νόμος

$$P(A_1 A_2 \dots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 A_2) \dots P(A_n|A_1 A_2 \dots A_{n-1}). \quad (1.5)$$

Το Θεώρημα Ολικής πιθανότητας και ο κανόνας του Bayes έχουν ως ακολούθως:

α. Έστω $H_1, H_2, \dots, H_k$ k ασυμβίβαστα μεταξύ τους ενδεχόμενα το άθροισμα των οποίων καλύπτει όλο το δειγματικό χώρο S (δηλ. $S = H_1 \cup H_2 \cup \dots \cup H_k$) και $P(H_i) > 0$ για κάθε $i=1, 2, \dots, k$. Τότε για κάθε ενδεχόμενο A έχουμε

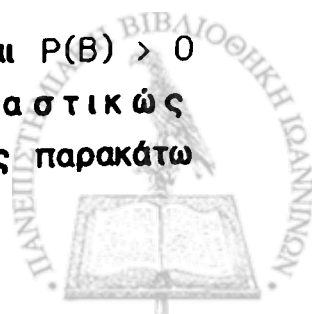
$$P(A) = \sum_{i=1}^k P(AH_i) = \sum_{i=1}^k P(A|H_i)P(H_i). \quad (1.6)$$

β. Έστω $H_1, \dots, H_k$ k ασυμβίβαστα μεταξύ τους γεγονότα καλύπτοντα όλο το δειγματικό χώρο S με $P(H_i) > 0$ για κάθε i . Τότε για κάθε ενδεχόμενο A με $P(A) > 0$ έχουμε

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{\sum_{i=1}^k P(A|H_i)P(H_i)}. \quad (1.7)$$

Οι πιθανότητες $P(H_i)$ λέγονται εκ των προτέρων (a priori) πιθανότητες και οι πιθανότητες $P(H_i|A)$ λέγονται εκ των υστέρων (a posteriori).

Δύο ενδεχόμενα A και B με $P(A) > 0$ και $P(B) > 0$ λέγονται στατιστικώς ή στοχαστικώς ανεξάρτητα αν και μόνο αν ισχύει μία από τις παρακάτω σχέσεις



$$\begin{aligned}
 \text{(i)} \quad P(A|B) &= P(A) \\
 \text{(ii)} \quad P(B|A) &= P(B) \\
 \text{(iii)} \quad P(AB) &= P(A)P(B).
 \end{aligned}
 \tag{1.8}$$

η ενδεχόμενα $A_1, \dots, A_n$ λέγονται στατιστικώς ή στοχαστικώς ανεξάρτητα αν ισχύει η σχέση

$$P(A_{i_1} A_{i_2} \dots A_{i_k}) = P(A_{i_1}) P(A_{i_2}) \dots P(A_{i_k}) \tag{1.9}$$

για κάθε υποσύνολο $(i_1 i_2 \dots i_k)$ του συνόλου $\{1, 2, \dots, n\}$.

1.3 Τυχάλες Μεταβλητές - Κατανομές Πιθανότητας - Μετρήσεις

Οι μονοδιάστατες τυχαίες μεταβλητές (τ.μ.) είναι μονοσήμαντες συναρτήσεις με πεδίο ορισμού ένα δειγματικό χώρο S και τιμές ένα υποσύνολο της ευθείας των πραγματικών αριθμών. Συμβολίζονται συνήθως με κεφαλαία γράμματα X, Y, Z κ.λ.π. ή X_1, X_2, Y_1 κ.λ.π. Αν το σύνολο των τιμών μιας τυχαίας μεταβλητής είναι το πολύ αριθμήσιμο τότε η τ.μ. λέγεται διακριτή ή απαριθμητή ή ασυνεχής. Αν το πεδίο τιμών μιας τυχαίας μεταβλητής είναι ένα πεπερασμένο ή άπειρο διάστημα της ευθείας των πραγματικών αριθμών (ή ολόκληρη η ευθεία) και $P(X=x) = 0$ για κάθε τιμή της τ.μ. τότε η τ.μ. λέγεται (απολύτως) συνεχής.

Κάθε τυχαία μεταβλητή έχει και μία κατανομή πιθανότητας. Αν X διακριτή τ.μ. η συνάρτηση

$$p_X(x) = P(X=x) \tag{1.10}$$

λέγεται συνάρτηση πιθανότητας (σ.π.) ή κατανομή πιθανότητας της X . Αν X συνεχής τ.μ. και υπάρχει μία μη αρνητική συνάρτηση με πεδίο ορισμού το πεδίο τιμών



της X τέτοια ώστε

$$P(X \in \Delta) = \int_{\Delta} f_X(x) dx \quad (1.11)$$

για κάθε διάστημα Δ του πεδίου τιμών του X , τότε η $f_X(x)$ λέγεται **συνάρτηση πυκνότητας πιθανότητας (σ.π.π.)** ή απλώς **κατανομή πιθανότητας**.

Οι συναρτήσεις πιθανότητας έχουν τις ιδιότητες

α. $p_X(x) \geq 0$ για κάθε $x \in R$.

β. $\sum_x p_X(x) = 1.$ (1.12)

γ. $P(X \in B) = \sum_{x \in B} p_X(x), B \in A.$

Οι συναρτήσεις πυκνότητας πιθανότητας έχουν τις ιδιότητες

α. $f_X(x) \geq 0, x \in R.$

β. $\int_{-\infty}^{+\infty} f_X(x) dx = 1.$ (1.13)

γ. $P(x < X \leq x+dx) = f(x)dx.$

Οι κατανομές πιθανότητας δίνονται και με τις **αθροιστικές συναρτήσεις κατανομής (α.σ.κ.)** $F_X(x)$ που ορίζονται από τη σχέση

$$F_X(x) = P(X \leq x), x \in R \quad (1.14)$$



και έχουν τις ακόλουθες ιδιότητες:

α. $P(a < X \leq b) = F_X(b) - F_X(a).$

β. Αν η τ.μ. X είναι συνεχής τότε

$$F_X(x) = \int_{-\infty}^x f(y) dy \text{ και } f(x) = F'(x) \text{ για κάθε } x \in \mathbb{R}.$$

γ. $F(x)$, μη φθίνουσα: αν $x_1 < x_2$ τότε $F(x_1) \leq F(x_2).$

δ. $F(x)$ συνεχής από τα δεξιά για κάθε $x \in \mathbb{R}.$

ε. $P(X < x) = \lim_{y \rightarrow x^-} F(y) \equiv F(x^-).$

στ. $P(X = x) = F(x) - F(x^-).$

Η αθροιστική συνάρτηση κατανομής μιας διακριτής τυχάλες μεταβλητής είναι κλιμακωτή συνάρτηση (step function) με αριθμήσιμο πλήθος ασυνεχειών (πηδημάτων). Αντίθετα, η αθροιστική συνάρτηση κατανομής μιας (απολύτως) συνεχούς τ.μ. είναι συνεχής παντού. Υπάρχουν τυχάλες μεταβλητές οι οποίες δεν είναι ούτε διακριτές ούτε συνεχείς. Είναι μικτού τύπου και οι αθροιστικές συναρτήσεις κατανομής τους έχουν συνεχή και κλιμακωτά τμήματα.

Η αθροιστική συνάρτηση κατανομής μιας τ.μ. είναι παραγωγίσιμη παντού εκτός από ένα το πολύ αριθμήσιμο πλήθος σημείων. Υπάρχουν συνεχείς α.σ.κ. των οποίων η παράγωγος είναι μηδέν σχεδόν παντού. Κατανομές με τέτοιες α.σ.κ. λέγονται ιδιάζουσες.

Τα αριθμητικά δεδομένα που επεξεργάζεται η Στατιστική προέρχονται από μία ή περισσότερες μεταβλητές οι οποίες αναφέρονται σε διάφορα φαινόμενα ή πειράματα (φυσικά, κοινωνικά, ιατρικά κ.λ.π.).



Ιδιαίτερα οι μεταβλητές της Στατιστικής είναι τυχαίες μεταβλητές. Τα φαινόμενα ή πειράματα κυβερνώνται από τους νόμους της τύχης και οι τιμές των μεταβλητών είναι αποτέλεσμα παραγόντων τύχης. Συχνά τα αριθμητικά δεδομένα ή οι τιμές των μεταβλητών λέγονται και παρατηρήσεις ή απλώς μετρήσεις, ιδιαίτερα όταν προκύπτουν από διαδικασίες μετρήσεων ή παρατηρήσεων. Στην αρχική τους φάση οι μετρήσεις ή παρατηρήσεις δεν είναι κατ'ανάγκην αριθμητικού χαρακτήρος. Μπορεί να είναι σύμβολα, εικόνες, καμπύλες κ.λ.π. Εύκολα όμως με κατάλληλες συναρτήσεις (μετασχηματισμούς) μετατρέπονται σε αριθμητικά δεδομένα. Τα δεδομένα αυτά συμβολίζονται με:

$x_1, x_2, \dots, x_n$ όταν προέρχονται από την μεταβλητή X

$y_1, y_2, \dots, y_n$ όταν προέρχονται από την μεταβλητή Y

$(x_1, y_1), \dots, (x_n, y_n)$ όταν προέρχονται από την μεταβλητή (X, Y) .

$x_{ij}, i=1, 2, \dots, k, j=1, 2, \dots, n_i$ με $n_1 + n_2 + \dots + n_k = n$ κ.λ.π. όπου n παριστάνει το πλήθος ή αριθμό των μετρήσεων.

Οι μεταβλητές διακρίνονται σε δύο γενικές κατηγορίες: ποσοτικές και ποιοτικές ή κατηγορικές μεταβλητές. Οι πρώτες είναι μεταβλητές που μπορούν να μετρηθούν π.χ. το μήκος, το ύψος, το πλάτος, το βάρος, η ηλικία, ο χρόνος μεταξύ διαδοχικών εμφανίσεων ενός ενδεχομένου, ο αριθμός τηλεφωνικών κλήσεων σε μία ώρα κ.λ.π. Αντίθετα οι ποιοτικές μεταβλητές δεν μπορούν να μετρηθούν και αναφέρονται σε καταστάσεις, κατηγορίες, επίπεδα, είδη, παρουσία ή απουσία ενός ή περισσότερων γνωρισμάτων κ.λ.π. Για παράδειγμα το φύλο των ανθρώπων, το χρώμα, η οικονομικοκοινωνική κατάσταση (φτωχός, πλούσιος κ.λ.π.), η επιτυχία, η αποτυχία κ.λ.π.

Ανάλογη είναι και η διάκριση των μεταβλητών σε συνεχείς και διακριτές ή απαριθμητές μεταβλητές.

1.4 Πληθυσμός - Τυχαίο Δείγμα - Κατανομή Δείγματος

Οι πρώτες βασικές έννοιες της Στατιστικής είναι ο πληθυσμός και το δείγμα. Η έννοια του πληθυσμού ταυτίζεται με την ομώνυμη έννοια στη Θεωρία Πιθανοτήτων. Το σύνολο των τιμών μίας τυχαίας μεταβλητής (μονοδιάστατης ή πολυδιάστατης) είναι ένας (θεωρητικός) πληθυσμός. Έτσι μερικές φορές ο πληθυσμός συμβολίζεται όπως και οι τυχαίες μεταβλητές και λέμε έστω ο πληθυσμός X, Y, Z κ.λ.π.

Το μέγεθος του πληθυσμού μπορεί να είναι πεπερασμένο, άπειρο ή υπεράριθμο. Το μέγεθος επίσης μπορεί να αυξάνεται ή ελαττούται ανάλογα με το ιδιαίτερο ενδιαφέρον μας. Αν, για παράδειγμα, ενδιαφερόμαστε για την γενική βαθμολογία των εισαγωγικών εξετάσεων στα Ανώτατα Εκπαιδευτικά ιδρύματα μιας χρονιάς, ο πληθυσμός μας αποτελείται από όλους τους υποψήφιους που έδωσαν εξετάσεις και ιδιαίτερα από όλες τις

Αν ενδιαφερόμαστε για τη γενική βαθμολογία του Πολυτεχνικού-Φυσικομαθηματικού Κύκλου τότε έχουμε ένα διαφορετικό (μικρότερο) πληθυσμό. Έτσι οι πληθυσμοί στην Στατιστική διαμορφώνονται ή καθορίζονται ανάλογα με τη σφαίρα του ενδιαφέροντός μας.

Ο πληθυσμός αποτελεί το άγνωστο μέρος της Στατιστικής. Σκοπός της Στατιστικής είναι η εξαγωγή συμπερασμάτων για τον πληθυσμό βάσει του δείγματος.

Το δείγμα ορίζεται απλώς σαν ένα μέρος του πληθυσμού. Αναφερόμενοι στο παραπάνω παράδειγμα οι πρωτοετείς φοιτητές της Ιατρικής Σχολής του Παν/μίου Ιωαννίνων αποτελούν ένα δείγμα και οι γενικές βαθμολογίες τους στις εισαγωγικές εξετάσεις αποτελούν ένα δείγμα γενικών βαθμολογιών. Οι τιμές των μεταβλητών που προκύπτουν από τα μέλη του δείγματος λέγονται και τιμές του δείγματος. Υπάρχουν διάφορα είδη δειγμάτων. Βασικός στόχος είναι το δείγμα να αντιπροσωπεύει τον πληθυσμό. Στην αρχή θα περιοριστούμε στο τυχαίο δείγμα.



Τυχαιό δείγμα είναι το δείγμα που εκλέγεται κατά τέτοιο τρόπο ώστε όλα τα μέλη του πληθυσμού να έχουν ίση και ανεξάρτητη πιθανότητα να συμπεριληφθούν στο δείγμα.

Είναι προφανές ότι οι τιμές του τυχαίου δείγματος είναι τιμές της τυχαίας μεταβλητής (ή των τυχαίων μεταβλητών) του πληθυσμού. Έτσι οι γενικές βαθμολογίες του δείγματος είναι τιμές της τυχαίας μεταβλητής "γενική βαθμολογία εισαγωγικών εξετάσεων". Αν δε φανταστούμε ότι η όλη δειγματοληψία μπορεί να επαναληφθεί νοερώς ένα μεγάλο (άπειρο) αριθμό φορές, βλέπουμε ότι οι τιμές του δείγματος μεταβάλλονται σε όλο το φάσμα των τιμών της τυχαίας μεταβλητής. Με τις νοερές αυτές επαναλήψεις κάθε τιμή του δείγματος γίνεται τυχαία μεταβλητή ισόνομη με την X και ανεξάρτητη από τις άλλες. Έτσι έχουμε τον ακόλουθο ορισμό:

Ορισμός 1.1: Τυχαιό δείγμα μεγέθους n από ένα πληθυσμό X ή μία κατανομή είναι n ανεξάρτητες και ισόνομες τυχαίες μεταβλητές $X_1, X_2, \dots, X_n$.

Από τα παραπάνω βλέπουμε ότι οι τιμές ενός τυχαίου δείγματος παίζουν ένα διπλό ρόλο: αφ'ενός είναι συγκεκριμένες αριθμητικές τιμές, αφ'ετέρου είναι και τυχαίες μεταβλητές με την έννοια των νοερών επαναλήψεων που αναφέραμε πιο πάνω. Έτσι το τυχαίο δείγμα άλλοτε θα συμβολίζεται με $x_1, x_2, \dots, x_n$ και άλλοτε με $X_1, X_2, \dots, X_n$.

Υπάρχουν πολλοί τρόποι με τους οποίους μπορούμε να εκλέξουμε ένα τυχαίο δείγμα. Ο πλέον συνηθισμένος τρόπος είναι με τη βοήθεια πινάκων τυχαίων αριθμών. Δείγμα πίνακα τυχαίων αριθμών παρατίθεται πιο κάτω.

Οι πίνακες αυτοί κατασκευάζονται έτσι ώστε κάθε αριθμός (μονοψήφιος, διψήφιος κ.λ.π.) να έχει την ίδια πιθανότητα ($1/10$, $1/100$, κ.λ.π. αντίστοιχα) να εμφανιστεί στον πίνακα.

99 60 50 50 60 12 48 08 01 88 60 51 73 38 54 47 41 67 43 84 21 35 12 90 43
 88 61 29 18 05 31 29 56 94 33 58 89 68 42 10 08 11 15 96 06 15 51 48 39 96
 19 28 81 63 24 30 90 40 11 59 30 10 01 02 60 30 16 26 83 53 41 09 50 65 01
 53 61 62 34 47 04 37 74 97 09 77 36 92 80 45 99 26 28 24 24 54 63 43 63 54
 40 03 44 30 11 42 25 70 19 79 90 12 36 16 80 28 70 24 80 07 76 17 01 50 80

65 15 18 13 54 05 13 69 91 51 84 57 52 89 88 12 52 03 39 71 19 48 20 94 10
 95 79 58 84 86 00 04 73 69 94 89 12 93 84 29 72 62 79 66 98 65 17 54 09 50
 75 26 86 16 42 65 03 22 43 68 87 68 70 09 18 92 85 94 60 32 97 44 95 82 72
 92 45 48 29 84 56 60 50 64 07 71 46 35 31 52 21 80 61 25 30 31 99 58 07 04
 43 00 97 26 90 99 85 55 75 16 09 55 34 16 16 94 32 12 12 07 32 90 97 02 47

Πίνακας 1.1 Τυχαίοι Αριθμοί

Η τυχαία δειγματοληψία γίνεται ως εξής: Παίρνουμε (ή κατασκευάζουμε) ένα κατάλογο (frame) του πληθυσμού και απαριθμούμε τα στοιχεία αυτού. Έτσι κάθε στοιχείο αντιστοιχεί σ'ένα μονοψήφιο, διψήφιο κ.λ.π. αριθμό. Ξεκινώντας από ένα τυχαίο σημείο του πίνακα τυχαίων αριθμών παίρνουμε ένα αριθμό μονοψήφιο, διψήφιο κ.λ.π. ανάλογα με το αν τα στοιχεία του πληθυσμού είναι μονοψήφια ή διψήφια κ.λ.π. Αν ο αριθμός αυτός αντιστοιχεί σε στοιχεία του πληθυσμού, το στοιχείο αυτό εκλέγεται στο δείγμα. Αν όχι ο αριθμός αυτός απορρίπτεται και παίρνουμε τον γειτονικό αριθμό. Προχωρούμε από γειτονικό σε γειτονικό αριθμό για την εκλογή των μελών του δείγματος.

Αντί για πίνακες τυχαίων αριθμών μπορούμε να χρησιμοποιήσουμε και ένα τηλεφωνικό κατάλογο. Έχει αποδειχθεί ότι τα κεντρικά ψηφία επταψήφιων αριθμών τηλεφώνου συμπεριφέρονται σαν τυχαίοι αριθμοί.

Τυχαία δείγματα δεν εμφανίζονται πάντοτε στην πράξη. Άλλοτε οι πιθανότητες εκλογής των στοιχείων του πληθυσμού είναι άνισες και άλλοτε η ανεξαρτησία δεν ισχύει. Η Στατιστική στηρίζεται σε δείγματα πιθανότητας, δηλαδή δείγματα που έχουν ωρισμένη πιθανότητα να εκλεγούν. Αργότερα θα ονομάσουμε την πιθανότητα αυτή πιθανοφάνεια (likelihood).

Ο πληθυσμός αποτελεί το άγνωστο μέρος της



Στατιστικής. Το δείγμα αποτελεί το γνωστό μέρος. Η Στατιστική είναι επαγωγική επιστήμη, η οποία προσπαθεί να βγάλει συμπεράσματα από το επί μέρους, γνωστό τμήμα του πληθυσμού, το δείγμα, για το καθολικό μέρος, ολόκληρο τον πληθυσμό.

Με κάθε μέτρηση ο νους του ανθρώπου μπορεί να συσχετίσει ένα μέτρο κανονικότητας της μέτρησης (ή έλλειψης αυτής) και πιο συγκεκριμένα ένα μέτρο πιθανότητας πραγματοποίησης της μέτρησης. Αυτό οδηγεί στην έννοια της κατανομής η οποία κατά κάποιο τρόπο δείχνει την θέση της μετρήσεως ως προς τις άλλες μετρήσεις γνωστές ή άγνωστες. Έτσι έχουμε δύο κατανομές: την κατανομή του πληθυσμού (της τ.μ.Χ) και την κατανομή του δείγματος. Η κατανομή του πληθυσμού είναι γνωστή από την Θεωρία Πιθανοτήτων. Η κατανομή του δείγματος στηρίζεται στην έννοια της σχετικής συχνότητας και αναπτύσσεται στο Κεφάλαιο 2.

1.5 **Ειδικές Κατανομές**

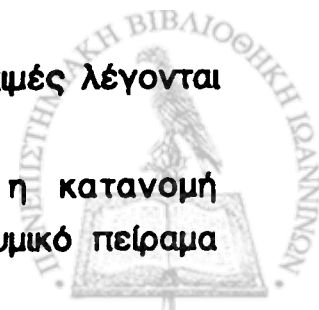
Οι κυριώτερες από τις διακριτές κατανομές πιθανότητας είναι η διωνυμική, η υπεργεωμετρική, η γεωμετρική, η αρνητική διωνυμική και η κατανομή του Poisson.

Ένα πείραμα που αποτελείται από δοκιμές η κάθε μία από τις οποίες έχει μόνο δύο δυνατά αποτελέσματα: επιτυχία (Ε) ή αποτυχία (Α) και για το οποίο ισχύουν

- (i) η πιθανότητα επιτυχίας ή αποτυχίας παραμένει σταθερή σε όλες τις δοκιμές του πειράματος και
- (ii) οι δοκιμές είναι στατιστικώς ανεξάρτητες

λέγεται διωνυμικό πείραμα. Οι επί μέρους δοκιμές λέγονται δοκιμές του Bernoulli.

Η διωνυμική κατανομή είναι η κατανομή πιθανότητας του αριθμού X των επιτυχιών σε ένα διωνυμικό πείραμα



με n δοκιμές και πιθανότητα επιτυχίας p . Η συνάρτηση πιθανότητας της διωνυμικής κατανομής είναι

$$p_X(x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \dots, n \quad (1.15)$$

Συμβολικά γράφουμε $X \sim B(n, p)$.

Η υπεργεωμετρική κατανομή εμφανίζεται όταν κάνουμε δειγματοληψία χωρίς επανάθεση από ένα πεπερασμένο πληθυσμό. Έστω N το μέγεθος του πληθυσμού, Np το πλήθος των στοιχείων μίας κατηγορίας C , Nq το πλήθος των στοιχείων που δεν ανήκουν στην κατηγορία C , p το ποσοστό των στοιχείων της κατηγορίας C , $q=1-p$ και n το μέγεθος του δείγματος. Αν X είναι ο αριθμός των στοιχείων της κατηγορίας C που εμφανίζονται στο δείγμα, τότε η X είναι μία υπεργεωμετρική τ.μ. με κατανομή

$$p_X(x) = \frac{\binom{Np}{x} \binom{Nq}{n-x}}{\binom{N}{n}} \quad (1.16)$$

$$x = \max(0, n - Nq), \dots, \min(n, Np).$$

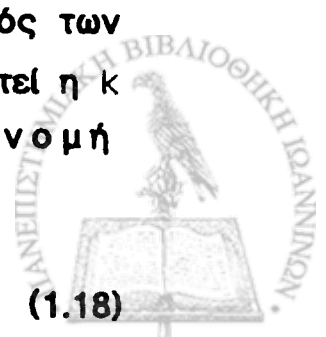
Συμβολικά γράφουμε $X \sim Hg(N, n, p)$.

Ο χρόνος αναμονής (μετρούμενος σε ακέραιες μονάδες) ή ο αριθμός των δοκιμών μέχρις ότου εμφανιστεί η πρώτη επιτυχία σε ένα διωνυμικό πείραμα ακολουθεί τη γεωμετρική κατανομή με σ.π.

$$p_X(x) = pq^{x-1}, \quad x = 1, 2, 3, \dots \quad (1.17)$$

Συμβολικά γράφουμε $X \sim \text{Geo}(p)$. Στο ίδιο πλαίσιο, ο αριθμός των δοκιμών αναμονής ("χρόνος αναμονής") μέχρις ότου εμφανιστεί η k επιτυχία ακολουθεί την αρνητική διωνυμική κατανομή με σ.π.

$$p_X(x) = \binom{x-1}{k-1} p^k q^{x-k}, \quad x = k, k+1, \dots \quad (1.18)$$



Η αρνητική διωνυμική τ.μ. $NB(k, p)$ είναι άθροισμα k ανεξάρτητων γεωμετρικών τ.μ. $Geo(p)$.

Η κατανομή του Poisson είναι μία από τις πιο βασικές κατανομές της Θεωρίας Πιθανοτήτων και της Στατιστικής. Εφαρμόζεται όταν έχουμε μία σειρά τυχαίων ενδεχομένων τα οποία γίνονται με την πάροδο του χρόνου έτσι ώστε (i) η πιθανότητα να συμβεί ένα ενδεχόμενο σε ένα μικρό διάστημα να είναι ανάλογη του μήκους του διαστήματος, (ii) η πιθανότητα να συμβούν δύο ή περισσότερα ενδεχόμενα σε ένα πολύ μικρό διάστημα να είναι αμελητέα και (iii) τα ενδεχόμενα που πραγματοποιούνται σε μη επικαλυπτόμενα χρονικά διαστήματα να είναι στατιστικώς ανεξάρτητα. Έστω X ο αριθμός των ενδεχομένων ('αφίξεων', 'εκπομπών') που εμφανίζονται σε ένα χρονικό διάστημα και που υπακούουν στις παραπάνω προϋποθέσεις. Η τ.μ. X έχει την κατανομή του Poisson με σ.π.

$$p_X(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots \quad (1.19)$$

όπου λ είναι ο μέσος αριθμός 'αφίξεων' στο ίδιο διάστημα. Συμβολικά $X \sim P(\lambda)$.

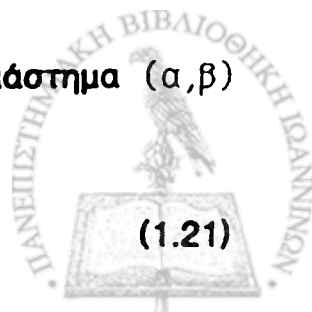
Αν $X(t)$ είναι ο αριθμός των 'αφίξεων' στο διάστημα $(0, t)$, όπου το t είναι χρόνος μεταβλητός, η οικογένεια των τ.μ. $X(t)$ λέγεται στοχαστική διαδικασία του Poisson και έχει σ.π. την

$$p(x, t) = P[X(t)=x] = \frac{e^{-\lambda t} (\lambda t)^x}{x!}, \quad x = 0, 1, 2, \dots \quad (1.20)$$

Οι κυριώτερες συνεχείς κατανομές πιθανότητες είναι η ομοιόμορφη, η εκθετική, οι κατανομές Γάμμα και Βήτα και η κανονική κατανομή.

Η ομοιόμορφη κατανομή στο διάστημα (α, β) έχει σ.π.π.

$$f_X(x) = \frac{1}{\beta - \alpha} I_{[\alpha, \beta]}(x), \quad (1.21)$$



όπου $I_{[\alpha, \beta]}(x)$ είναι η δείκτρια συνάρτηση του συνόλου $[\alpha, \beta]$, δηλαδή έχει τιμή 1 αν $x \in [\alpha, \beta]$ και 0 αν $x \notin [\alpha, \beta]$. Συμβολικά γράφουμε $X \sim U(\alpha, \beta)$.

Ο χρόνος αναμονής μεταξύ δύο διαδοχικών αφίξεων μίας διαδικασίας του Poisson με παράμετρο λ ακολουθεί εκθετική κατανομή με σ.π.π.

$$f_X(x) = \lambda e^{-\lambda x} I_{(0, +\infty)}(x). \quad (1.22)$$

Συμβολικά γράφουμε $X \sim \text{Εκθ}(\lambda)$.

Ο χρόνος αναμονής μέχρι της εμφάνισης του α στη σειρά (α φυσικός αριθμός) τυχαίου ενδεχόμενου σε μία διαδικασία του Poisson με παράμετρο $\lambda = 1/\beta$ ακολουθεί την Γάμμα κατανομή με σ.π.π.

$$f_X(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} I_{(0, +\infty)}(x). \quad (1.23)$$

Συμβολικά γράφουμε $X \sim G(\alpha, \beta)$. Για $\alpha = \nu/2$ και $\beta = 2$ η κατανομή Γάμμα γίνεται η χ^2 τετράγωνο με ν βαθμούς ελευθερίας, δηλαδή

$$\chi_\nu^2 = G(\alpha = \nu/2, \beta = 2) \quad (1.24)$$

Η Βήτα κατανομή έχει σ.π.π.

$$f_X(x) = \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)} I_{(0, 1)}(x) \quad (1.25)$$

Συμβολικά $X \sim \text{Be}(\alpha, \beta)$.



Η τυπική κανονική κατανομή $N(0,1)$ έχει σ.π.π.

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R} \quad (1.26)$$

και η (γενική) κανονική κατανομή $N(\mu, \sigma^2)$ έχει σ.π.π.

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbb{R}. \quad (1.27)$$

Οι παράμετροι μ και σ^2 είναι η μέση τιμή και η διακύμανση της κατανομής αντίστοιχα. Η κανονική κατανομή είναι γνωστή και ως η κατανομή του Gauss ή η κατανομή των σφαλμάτων των μετρήσεων.

Αν $X \sim N(0,1)$, τότε $Y = \sigma X + \mu \sim N(\mu, \sigma^2)$ και αν $X \sim N(\mu, \sigma^2)$, τότε $Y = (X - \mu)/\sigma \sim N(0,1)$. Η α.σ.κ. της τυπικής κανονικής κατανομής συμβολίζεται με $\Phi(x)$ και δίνεται από τη σχέση

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt, \quad x \in \mathbb{R}. \quad (1.28)$$

Ισχύουν οι ακόλουθες σχέσεις:

$$\Phi(x) = 1 - \Phi(-x), \quad x \in \mathbb{R}$$

$$\Phi(x) = \frac{1}{2} + \int_0^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt, \quad x > 0.$$

Η τ.μ. $Z = (X - \mu)/\sigma$ λέγεται τυποποιημένη τ.μ.



1.6 Χαρακτηριστικά των Τυχαίων Μεταβλητών και Κατανομών

Τα χαρακτηριστικά των τυχαίων μεταβλητών ή κατανομών είναι:

η μέση τιμή

$$\mu = E(X) = \begin{cases} \sum x p(x), & \text{αν } X \text{ διακριτή τ.μ.} \\ \int_{-\infty}^{+\infty} x f(x) dx, & \text{αν } X \text{ συνεχής τ.μ.} \end{cases} \quad (1.29)$$

η διακύμανση ή διασπορά

$$\sigma^2 = \text{Var}(X) = E(X - \mu)^2 = EX^2 - (EX)^2 \quad (1.30)$$

ο συντελεστής λοξότητας (συμμετρίας-ασυμμετρίας της κατανομής)

$$\alpha_3 = E[(X - \mu)/\sigma]^3 \quad (1.31)$$

και ο συντελεστής κύρτωσης

$$\alpha_4 = E[(X - \mu)/\sigma]^4. \quad (1.32)$$

Η μαθηματική ελπίδα ή αναμενόμενη τιμή ή μέση τιμή μιας συνάρτησης $h(X)$ της τ.μ. X ορίζεται με τη σχέση

$$Eh(X) = \begin{cases} \sum h(x)p(x) & , \text{αν } X \text{ διακριτή τ.μ.} \\ \int_{-\infty}^{+\infty} h(x)f(x)dx & \text{αν } X \text{ συνεχής τ.μ.} \end{cases} \quad (1.33)$$



Και στις δύο περιπτώσεις η μαθηματική ελπίδα ορίζεται (υπάρχει) υπό τον όρο ότι το άθροισμα ή το ολοκλήρωμα στην (1.29) και την (1.33) συγκλίνουν απόλυτα.

Τυπική απόκλιση μίας τ.μ. X είναι

$$\sigma = \sqrt{\text{Var}(X)}. \quad (1.34)$$

Η διακύμανση μετρά τη μεταβλητότητα της τ.μ. Για c - σταθερά, $\text{Var}(c) = 0$. Επιπλέον $\text{Var}(aX+b) = a^2 \text{Var}(X)$

Η σημασία της διακύμανσης καταφαίνεται και με την ανισότητα του Chebyshev:

$$P(|X-\mu| \geq \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2} \quad (1.35)$$

ή

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - 1/k^2, \quad k > 0. \quad (1.36)$$

Αν η κατανομή είναι συμμετρική περί το μ , $\alpha_3 = 0$ και αν είναι $N(\mu, \sigma^2)$ τότε $\alpha_4 = 3$. Αν $\alpha_3 > 0$ η κατανομή είναι λοξή προς τα δεξιά ενώ αν $\alpha_3 < 0$ είναι λοξή προς τα αριστερά. Αν $\alpha_4 > 3$ η κατανομή είναι περισσότερο κυρτή από την κανονική ενώ αν $\alpha_4 < 3$ είναι λιγότερο κυρτή.

Οι μέσες τιμές και διακυμάνσεις ειδικών κατανομών έχουν ως ακολούθως:

α. Διωνυμική κατανομή $B(n, p)$

$$\mu = np \qquad \sigma^2 = npq$$

β. Υπεργεωμετρική κατανομή $Hg(N, n, p)$

$$\mu = np \qquad \sigma^2 = npq \frac{N-n}{N-1}$$



γ. Γεωμετρική κατανομή $\text{Geo}(p)$

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{q}{p^2}$$

δ. Αρνητική διωνυμική κατανομή $\text{NB}(k, p)$

$$\mu = \frac{k}{p} \qquad \sigma^2 = \frac{kq}{p^2}$$

ε. Κατανομή του Poisson $P(\lambda)$

$$\mu = \lambda \qquad \sigma^2 = \lambda$$

στ. Ομοιόμορφη κατανομή $U(\alpha, \beta)$

$$\mu = \frac{\alpha + \beta}{2} \qquad \sigma^2 = \frac{(\beta - \alpha)^2}{12}$$

ζ. Εκθετική κατανομή $\text{Exp}(\lambda)$

$$\mu = \frac{1}{\lambda} \qquad \sigma^2 = \frac{1}{\lambda^2}$$

η. Κατανομή Γάμμα $G(\alpha, \beta)$

$$\mu = \alpha\beta \qquad \sigma^2 = \alpha\beta^2$$

θ. Κατανομή Βήτα $\text{Be}(\alpha, \beta)$

$$\mu = \frac{\alpha}{\alpha + \beta} \qquad \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)(\alpha + \beta + 1)}$$

ι. Κανονική κατανομή $N(\mu, \sigma^2)$

$$\mu = \mu \qquad \sigma^2 = \sigma^2$$



Εκτός από τη μέση τιμή, τη διακύμανση και τις ροπές διαφόρων τάξεων μίας κατανομής υπάρχουν και άλλα χαρακτηριστικά όπως η κορυφή, η διάμεσος, τα ποσοστιαία σημεία τα οποία περιγράφουν τη θέση μιας κατανομής.

Κορυφή ή επικρατούσα τιμή μιας κατανομής είναι το σημείο ή σημεία στα οποία εμφανίζεται το μέγιστο της $p(x)$ ή $f(x)$. Τα σημεία x_p για τα οποία ισχύει

$$P(X \leq x_p) = F(x_p) = p \quad \text{αν } X \text{ συνεχής}$$

ή (1.37)

$$P(X \leq x_p) \geq p \text{ και } P(X \geq x_p) \geq 1 - p \text{ αν } X \text{ διακριτή}$$

λέγονται p ποσοστιαία σημεία (quantiles) της τ.μ. Τα ποσοστιαία σημεία δεν ορίζονται κατ' ανάγκην μονοσήμαντα. Αν $p=0.01, 0.02, \dots, 0.99$ τα σημεία $x_{0.01}, x_{0.02}, \dots, x_{0.99}$ λέγονται και πρώτο, δεύτερο, ..., εννενηκοστό ένατο εκατοστιαία σημεία ή εκατοστημόρια (percentiles). Αν $p=0.5$ το σημείο $x_{0.5}$ λέγεται διάμεσος της κατανομής και συνήθως συμβολίζεται με το m . Τα σημεία $x_{0.25}, x_{0.50}, x_{0.75}$ λέγονται σημεία του πρώτου, δεύτερου και τρίτου τεταρτημορίου. Η διαφορά $x_{0.75} - x_{0.25}$ λέγεται ενδοτεταρτημοριακό πλάτος (interquantile range).

Μέτρα θέσης μιας κατανομής είναι η μέση τιμή, η διάμεσος, η επικρατούσα τιμή, τα ποσοστιαία σημεία, το $(x_{0.75} - x_{0.25})/2$. Μέτρα διασποράς είναι η διακύμανση, η τυπική απόκλιση, ο συντελεστής μεταβλητότητας σ/μ , το ενδοτεταρτημοριακό πλάτος. Μέτρα λοξότητας είναι ο συντελεστής α_3 . Μέτρα κύρτωσης είναι ο συντελεστής α_4 .

1.7 Γεννήτριες Συναρτήσεις

Οι γεννήτριες συναρτήσεις τυχαίων μεταβλητών ή κατανομών αποτελούν πολύτιμα εργαλεία για τη μελέτη των



κατανομών. Από αυτές η πιο σημαντική είναι η ροπογεννήτρια συνάρτηση μίας τ.μ. X η οποία ορίζεται με τη σχέση

$$m_X(t) = E(e^{tX}) \quad (1.38)$$

Η $m_X(t)$ υπάρχει πάντοτε στο σημείο $t=0$. Υπάρχουν περιπτώσεις που η $m_X(t)$ δεν υπάρχει για καμιά άλλη τιμή του t εκτός της $t=0$. Αν η $m_X(t)$ υπάρχει για μια περιοχή του $t=0$ τότε έχει παραγώγους κάθε τάξεως στην περιοχή αυτή. Αν είναι δυνατή η παραγωγή ως προς t εντός των συμβόλων αθροίσης ή ολοκλήρωσης στην έκφραση της $E(e^{tX})$ τότε ισχύει

$$\mu_r = E(X^r) = \lim_{t \rightarrow 0} \frac{d^r m_X(t)}{dt^r} \equiv \left. \frac{d^r m_X(t)}{dt^r} \right|_{t=0} \quad (1.39)$$

Επίσης ισχύουν και οι σχέσεις

$$m_X(t) = 1 + \sum_{r=1}^{\infty} \mu_r \frac{t^r}{r!} \text{ για κάθε } t \in (-c, c) \quad (1.40)$$

$$m_{aX+b}(t) = e^{bt} m_X(at)$$

Είναι δυνατό μία τ.μ. X να μην έχει καμμία ροπή αλλά να έχει ροπογεννήτρια συνάρτηση. Επίσης να έχει όλες ή μερικές ροπές και να μην έχει ροπογεννήτρια συνάρτηση. Είναι ακόμη δυνατό μία τ.μ. να έχει όλες ή μερικές ροπές, να έχει ροπογεννήτρια συνάρτηση και η συνάρτηση αυτή να μην παράγει τις ροπές.

Η πιο σημαντική ιδιότητα των ροπογεννητριών συναρτήσεων είναι το **Θεώρημα του μονοσήμαντου**: αν οι ροπογεννήτριες συναρτήσεις δύο τ.μ. X και Y είναι ίσες, δηλαδή $m_X(t) = m_Y(t)$ για κάθε t , τότε οι τ.μ. X και Y είναι **ισόνομες**.



Το θεώρημα αυτό μας επιτρέπει τον προσδιορισμό των κατανομών τ.μ. με τη βοήθεια των ροπογεννήτριων συναρτήσεων.

Οι ροπογεννήτριες συναρτήσεις γνωστών κατανομών έχουν ως ακολούθως:

α. Διωνυμική $B(n, p)$

$$m(t) = (pe^t + q)^n, \quad t \in \mathbb{R}.$$

β. Γεωμετρική $Geo(p)$

$$m(t) = \frac{pe^t}{1 - qe^t}, \quad t < -\ln q.$$

γ. Αρνητική διωνυμική $NB(k, p)$

$$m(t) = \frac{p^k e^{tk}}{(1 - qe^t)^k}, \quad t < -\ln q.$$

δ. Poisson $P(\lambda)$

$$m(t) = e^{\lambda(e^t - 1)}, \quad t \in \mathbb{R}.$$

ε. Ομοιόμορφη $U(\alpha, \beta)$

$$m(t) = \frac{e^{\beta t} - e^{\alpha t}}{t(\beta - \alpha)}, \quad t \in \mathbb{R}.$$



στ. Εκθετική $Ek9(a)$

$$m(t) = \frac{a}{a-t}, \quad t < a.$$

ΥΠΟΦΙΛΗ ΣΤΟΙΧΟΣΤΑΤΗ

ζ. Γάμμα $G(\alpha, \beta)$

$$m(t) = (1-\beta t)^{-\alpha}, \quad t < 1/\beta.$$

η.δδ. Κανονική $N(\mu, \sigma^2)$

$$m(t) = e^{\mu t + (\sigma^2 t^2)/2}, \quad t \in \mathbb{R}.$$

2.1 Συχνότητα - Ιχθυική σχέση - Αθροιστική συχνότητα

Εστω $x_1, x_2, \dots, x_n$ μια δείγμα n μεγεθών. Οι τιμές x_i είναι ένας αριθμός n φορές ο ίδιος. Ο αριθμός των φορές που η τιμή x_i εμφανίζεται στο δείγμα λέγεται συχνότητα της τιμής και συμβολίζεται με f_i .

Ιχθυική συχνότητα της τιμής x_i είναι ο αριθμός f_i . Αν k είναι ο αριθμός των διαφορετικών τιμών του δείγματος τότε

$$\sum_{i=1}^k f_i = n$$



$$\frac{(a-b)^2}{a^2+b^2} = \frac{1}{2}$$

$$(a-b)^2 = a^2 + b^2$$

$$(a-b)^2 = a^2 + b^2$$

$$(a-b)^2 = a^2 + b^2$$

$$\frac{(a-b)^2}{a^2+b^2} = \frac{1}{2}$$

$$(a-b)^2 = a^2 + b^2$$

$$w(t) = \frac{1}{2} (a-b)^2$$

$$w(t) = \frac{1}{2} (a-b)^2$$

$$w(t) = \frac{1}{2} (a-b)^2$$

$$L(t) = \frac{1}{2} (a-b)^2$$

$$w(t) = \frac{1}{2} (a-b)^2$$

$$w(t) = \frac{1}{2} (a-b)^2$$

$$w(t) = \frac{1}{2} (a-b)^2$$

ΚΕΦΑΛΑΙΟ 2

ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Αν και η Περιγραφική Στατιστική δεν αποτελεί πια το κύριο ή το πιο σημαντικό θέμα της Στατιστικής, χρησιμοποιείται ευρύτατα σε πρωταρχική φάση κάθε ανάλυσης δεδομένων και τα τελευταία χρόνια γνωρίζει μεγάλη άνθηση χάρις στην αλματώδη ανάπτυξη των γραφικών μεθόδων των ηλεκτρονικών υπολογιστών.

Η Περιγραφική Στατιστική περιλαμβάνει τους πίνακες στατιστικών δεδομένων, τα γραφήματα διαφόρων μορφών και τύπων και ένα σύνολο αριθμητικών χαρακτηριστικών του δείγματος. Πολλά από αυτά στηρίζονται στην έννοια της συχνότητας.

2.1 Συχνότητα - Σχετική συχνότητα - Αθροιστική συχνότητα

Έστω $x_1, x_2, \dots, x_n$ ένα δείγμα από κάποιο πληθυσμό X . Οι τιμές x_i δεν είναι κατ'ανάγκη διάφορες μεταξύ τους. Ο αριθμός των φορών που η τιμή x_i εμφανίζεται στο δείγμα λέγεται **συχνότητα** της τιμής και συμβολίζεται με f_i .

Σχετική συχνότητα της τιμής x_i είναι ο αριθμός f_i/n . Αν k είναι ο αριθμός των διακεκριμένων τιμών του δείγματος τότε

$$\sum_{i=1}^k f_i = n \text{ και } \sum_{i=1}^k \frac{f_i}{n} = 1.$$



Η αθροιστική συχνότητα F_i της τιμής x_i ορίζεται ως εξής:

$$F_i = \begin{cases} \text{άθροισμα των συχνοτήτων } f_j \text{ όλων των} \\ \text{τιμών } x_j \text{ με την ιδιότητα } x_j \leq x_i. \end{cases}$$

Είναι προφανές ότι $x_i < x_j$ αν και μόνο αν $F_i < F_j$.

Η συχνότητα, σχετική συχνότητα κ.λ.π. δεν αναφέρονται μόνο σε μονοδιάστατες μετρήσεις x_i αλλά και σε πολυδιάστατες (x_i, y_i) , (x_i, y_i, z_i) κ.λ.π.

Παραδείγματα :

2.1: Υψη 750 ανδρών (με προσέγγιση του πλησιέστερου εκατοστού).

Ύψος Συχν.	Ύψος συχν.	Ύψος Συχν.	Ύψος Συχν.	Ύψος Συχν.
X (cm) f	X (cm) f	X (cm) f	X (cm) f	X (cm) f
155 3	162 10	169 34	176 40	183 25
156 4	163 11	170 35	177 36	184 23
157 5	164 13	171 36	178 31	185 23
158 5	165 20	172 37	179 30	186 1
159 7	166 24	173 39	180 29	187 5
160 7	167 27	174 42	181 28	188 4
161 9	168 30	175 45	182 26	189 2

2.2: Αριθμός εργατικών ατυχημάτων ενός πυριτιδοποιείου σε περίοδο 5 εβδομάδων.

Αριθμός ατυχημάτων x_i :	0	1	2	3	4	5
Συχνότητα f_i :	447	132	42	21	3	2



2.3: Βάρος και διάμετρος πορτοκαλιών

Βάρος (Προσέγγιση 2 gr)	Διάμετρος (προσέγγιση μισού cm)									Σύνολο
	23.0	23.5	24.0	24.5	25.0	25.5	26.0	26.5	27.0	
190		1								1
192			1							1
194	1	2	4	3	2	1				13
196	1	1	6	7	5		1			21
198		2	8	12	13	5	3	1		44
200	1	3	7	13	20	14	5	2		65
202		1	2	9	11	18	10	4	1	56
204			1	1	1	2	3	2	1	11
206					2	4	5	1	1	13
208						1	2	2		5
210							2			2
Σύνολο	3	10	29	45	54	45	31	12	3	232

2.2 Περιγραφή και σύμπτυξη των αριθμητικών δεδομένων

Η ανάγκη περιγραφής ή περιληπτικής παρουσίασης των μετρήσεων είναι προφανής. Ο νους του ανθρώπου μπορεί να συγκρατήσει μόνο ένα ορισμένο αριθμό στοιχείων, παραστάσεων ή γνώσεων. Μία από τις λειτουργίες του εγκεφάλου είναι η συστηματική ταξινόμηση των γνώσεων. Σήμερα με την πρόοδο της τεχνολογίας ο αριθμός των αριθμητικών δεδομένων τα οποία συλλέγονται για ένα ερευνητικό σκοπό μπορεί να είναι τεράστιος. Γίνεται λοιπόν αναγκαία η σύμπτυξη ή η περιληπτική παρουσίαση των αριθμητικών δεδομένων. Τα αριθμητικά δεδομένα μπορούν να συμπτυχθούν με τρεις τρόπους:

- α) Στατιστικοί πίνακες (Πίνακες συχνοτήτων),
- β) Γραφικές παραστάσεις (ιστογράμματα) και γ)
- Αριθμητικά μεγέθη. Ιστορικά οι κλάδοι της Στατιστικής και



της Ανάλυσης των Μετρήσεων ξεκίνησαν με τους δύο πρώτους τρόπους. Οι σύγχρονες θεωρίες χρησιμοποιούν τον τρίτο τρόπο σχεδόν αποκλειστικά. Παρά το γεγονός αυτό η χρήση και των τριών τρόπων είναι επιβεβλημένη γιατί ο κάθε ένας παρουσιάζει πλεονεκτήματα και αποκαλύπτει πτυχές τού πληθυσμού οι οποίες ενδεχομένως να αποκρύπτονται από τους άλλους.

2.2.1 Πίνακες συχνοτήτων

Υπάρχουν διάφοροι πίνακες συχνοτήτων. Από αυτούς αναφέρουμε τους εξής:

α. Απλός (μη διατεταγμένος) πίνακας συχνοτήτων:

Οι αρχικές μετρήσεις $x_1, x_2, \dots, x_n$ μπορούν να συμπυκθούν με την βοήθεια των συχνοτήτων ως εξής:

Μέτρηση	x_1	x_2	$\dots$	x_k
Συχνότητα	f_1	f_2	$\dots$	f_k

όπου όλες οι τιμές $x_1, x_2, \dots, x_k$ είναι διάφορες μεταξύ τους και k ο αριθμός των μετρήσεων αυτών.

β. Διατεταγμένος πίνακας συχνοτήτων:

Εάν τα $x_1, x_2, \dots, x_k$ διαταχθούν κατά σειρά μεγέθους προκύπτει ο πίνακας:



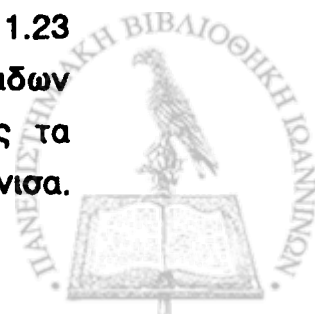
Μέτρηση	Συχνότητα	Αθροιστική συχνότητα
$x(1)$	f_1	$F_1 = f_1$
$x(2)$	f_2	$F_2 = f_1 + f_2$
.....		
$x(i)$	f_i	$F_i = f_1 + \dots + f_i$
.....		
$x(k)$	f_k	$F_k = n$
Σύνολο	n	

Πίνακας 2.1 Διατεταγμένος πίνακας συχνοτήτων όπου
 $x(1) < x(2) < \dots < x(k)$

Προφανώς καμμία απώλεια μετρήσεων δεν γίνεται με τους δύο παραπάνω τρόπους. Εάν όμως ο αριθμός k είναι μεγάλος, τέτοιες συμπτώξεις είναι μικρής σημασίας, διότι είναι σχεδόν ισοδύναμοι με τα αρχικά δεδομένα.

Τα παραδείγματα του εδαφίου 2.1 είναι διατεταγμένοι πίνακες συχνοτήτων.

γ. Ομαδοποιημένες μετρήσεις - Πίνακες συχνοτήτων: Μία από τις πιο συνήθεις μεθόδους σύμπτυξης μετρήσεων είναι η ομαδοποίηση αυτών. Ο αριθμός k των ομάδων, συνήθως 5 ως 30, εκλέγεται αυθαίρετα και το μήκος κάθε κατηγορίας ή ομάδας καθορίζεται από το k και το εύρος (range), δηλ. την διαφορά μεταξύ μέγιστης και ελάχιστης μέτρησης. Καταβάλλεται φροντίδα ώστε τα όρια των ομάδων να είναι διάφορα των μετρήσεων. Αυτό εξασφαλίζεται καθορίζοντας τα όρια των ομάδων να έχουν περισσότερα σημαντικά ψηφία από εκείνα των μετρήσεων. Έτσι αν οι μετρήσεις είναι 1.2, 1.8, 10.6, 4.9 κ.λ.π. δηλαδή έχουν ακρίβεια πρώτου δεκαδικού ψηφίου (γίνεται αυτομάτως στρογγύλεμα στο πλησιέστερο πρώτο δεκαδικό ψηφίο π.χ. το 1.23 καταχωρείται ως 1.2 και 1.27 ως 1.3) τότε τα όρια των ομάδων μπορεί να είναι 0.25-0.75, 0.75-1.25, 1.25-1.75 κ.λ.π. Συνήθως τα μήκη των ομάδων εκλέγονται ίσα, μπορούν όμως να είναι και άνισα.



Κάθε ομάδα αντιπροσωπεύεται από το μέσο σημείο της που λέγεται σημείο ή τιμή (class mark) της ομάδας, τη συχνότητα της ομάδας και την αθροιστική της συχνότητα. Συχνότητα ομάδας είναι ο αριθμός των μετρήσεων που πέφτουν στην ομάδα. Προφανώς η σύμπτυξη αυτή δεν παύει να έχει πρακτική χρησιμότητα, δεδομένου ότι αποτελεί προϋπόθεση κατασκευής ιστογραμμάτων και πολυγώνων, τα οποία κατά προσέγγιση απεικονίζουν την κατανομή του πληθυσμού. Οι ομαδοποιημένες μετρήσεις συνήθως παρουσιάζονται με τη μορφή του Πίνακα 2.2, όπου L_i , U_i είναι το κατώτερο και το ανώτερο όριο αντίστοιχα της ομάδας i και X_i το σημείο ή η τιμή αυτής. Για λόγους συνέχειας παίρνουμε $U_i = L_{i+1}$, $i = 1, 2, \dots, k-1$.

Ο πίνακας των ομαδοποιημένων μετρήσεων μας δίνει την λεγόμενη κατανομή συχνοτήτων των μετρήσεων.

Ομάδα	Όρια	Τιμή ομάδας	Συχνότητα	Αθροιστική συχνότητα
1	L_1, U_1	X_1	f_1	$F_1 = f_1$
2	L_2, U_2	X_2	f_2	$F_2 = f_1 + f_2$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
i	L_i, U_i	X_i	f_i	$F_i = f_1 + \dots + f_i$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
k	L_k, U_k	X_k	f_k	n
Σύνολο			n	

Πίνακας 2.2 Ομαδοποιημένες μετρήσεις

Μερικές φορές τα δεδομένα περιέχουν ακρότατες μετρήσεις, δηλαδή μετρήσεις που είναι ή πολύ μικρότερες ή πολύ μεγαλύτερες σε σχέση με τις άλλες. Τότε αν κάνουμε ομαδοποίηση



με τον συνήθη τρόπο θα προκύψουν ομάδες με μηδενική συχνότητα. Το πρόβλημα αυτό αποφεύγεται με δύο τρόπους: είτε διατηρώντας τον ίδιο αριθμό ομάδων και χρησιμοποιώντας "ανοικτά", δηλαδή απείρου μήκους διαστήματα, για την πρώτη ή την τελευταία ομάδα, είτε μεταβάλλοντας τον αριθμό των ομάδων. Έτσι έχουμε μεν καλύτερη εικόνα της κατανομής των μετρήσεων, χάνουμε όμως τις πραγματικές τιμές των ακρότατων μετρήσεων. Επί πλέον προκύπτουν δυσκολίες στον υπολογισμό των αριθμητικών χαρακτηριστικών του δείγματος.

Παράδειγμα 2.4: Εξήντα άτομα λαμβάνουν μέρος σε μία κλινική δοκιμή. Οι μετρήσεις χοληστερίνης και αίματος με προσέγγιση μονάδος έδωσαν τα ακόλουθα αποτελέσματα:

239	212	249	227	218	310	281	330	226	233
223	161	195	233	249	284	284	174	170	256
169	299	210	301	199	258	258	195	227	244
355	234	195	196	354	282	282	286	286	176
195	163	297	211	288	309	309	225	223	195
248	284	173	256	169	209	209	200	258	284

Ο διατεταγμένος πίνακας συχνοτήτων είναι της μορφής:

X_i	f_i
161	1
163	2
169	2
.	.
355	1
Σύνολο	60



34 Κεφ. 2 Περιγραφική Στατιστική

Μεγαλύτερη σύμπτυξη των δεδομένων επιτυγχάνεται με ομαδοποίηση. Έτσι διαλέγοντας ομαδοποίηση με 8 ομάδες και παρατηρώντας ότι η μέγιστη μείον η ελάχιστη τιμή είναι 355-161=194, διαιρούμε το 194 με το 8: $194/8=24.25$. Άρα παίρνουμε για μήκος της κάθε ομάδας 25 μονάδες με όρια ομάδων τα 160.5-185.5, 185.5-210.5 κ.λ.π.

Οι ομάδες, τα όρια τους, οι τιμές των ομάδων, η καταμέτρηση των συχνοτήτων και οι συχνότητες παρουσιάζονται στον παρακάτω πίνακα.

Αν μία από τις παραπάνω μετρήσεις ήταν 450 τότε θα προέκυπταν ομάδες μηδενικής συχνότητας. Στην περίπτωση αυτή η τελευταία ομάδα θα μπορούσε να γίνει 'ανοικτή' δηλαδή 335.5 και πάνω.

Ομάδα	Όρια	Τιμή	Συχνότητες	Αθροιστ. συχνότ.	Σχετικές συχνότητ.
1	160.5-185.5	173	X III = 8	8	0.1333
2	185.6-210.5	198	X X I = 11	19	0.1833
3	210.5-235.5	223	X X III = 13	32	0.2167
4	235.5-260.5	248	X X = 10	42	0.1667
5	260.5-285.5	273	X I = 6	48	0.1000
6	285.5-310.5	298	X II = 7	55	0.1167
7	310.5-335.5	323	III = 3	58	0.0500
8	335.5-360.5	348	II = 2	60	0.0333
Σύνολο			60		1.000

Πίνακας 2.3 Ομαδοποιημένες μετρήσεις χοληστερίνης

δ. Άλλοι στατιστικοί πίνακες: Γενικά πίνακες που περιέχουν διάφορα αριθμητικά στοιχεία λέγονται στατιστικοί πίνακες.

Παραδείγματα:

Πίνακες με εμβαδόν ή μήκος των ακτών διαφόρων γεωγραφικών περιοχών μιας χώρας.

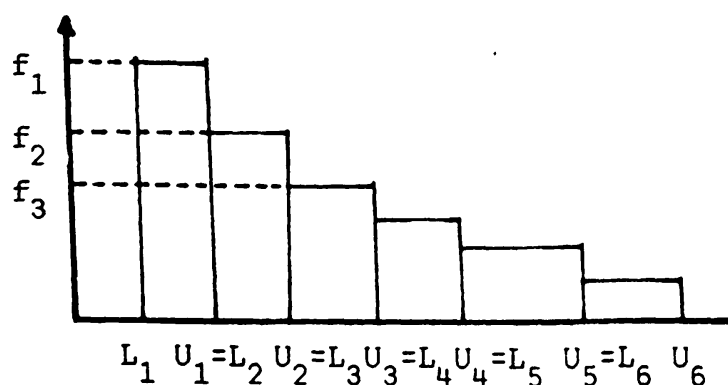


- Πίνακες με την ανά μήνα μέγιστη, ελαχίστη και μέση θερμοκρασία του αέρος σε διάφορες πόλεις.
- Πίνακες με τον ανά μήνα αριθμό ημερών βροχής για διάφορες πόλεις.
- Πίνακες θνησιμότητας ανά ηλικία και αιτία θανάτου για ένα ορισμένο έτος.
- Πίνακες ανεργίας κατά ηλικία, φύλλο, περιοχή κ.λ.π.

Ο αναγνώστης παραπέμπεται για περισσότερη ενημέρωση στις επετηρίδες και στατιστικές εκδόσεις της Εθνικής Στατιστικής Υπηρεσίας της Ελλάδας.

2.2.2 Γραφικές μέθοδοι σύμπτυξης

α. Ιστογράμματα (Histogram): Το ιστόγραμμα είναι ο πλέον γνωστός γραφικός τρόπος σύμπτυξης και παρουσίασης των μετρήσεων. Το διάγραμμα αυτό γίνεται σ'ένα σύστημα αξόνων x και y και αποτελείται από ένα σύνολο ορθογωνίων τοποθετημένων στον άξονα των x , όπως φαίνεται στο παρακάτω σχήμα. Η βάση των ορθογωνίων παριστάνει το μήκος των ομάδων και το ύψος την συχνότητα των ομάδων (ή ποσότητα ανάλογη της συχνότητας).

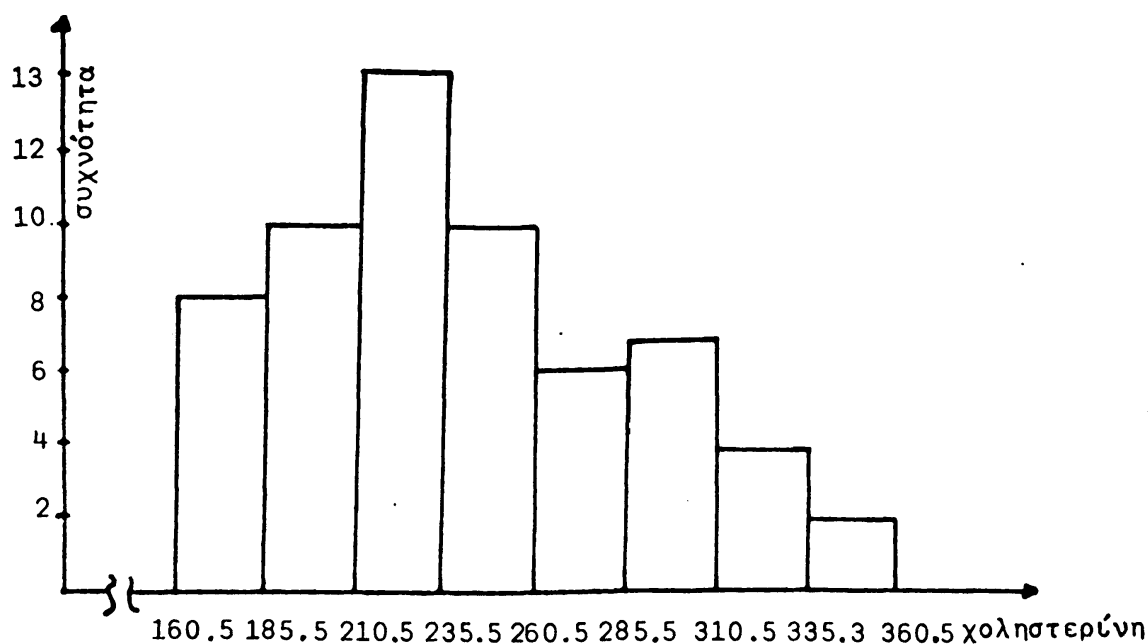


Σχήμα 2.1 Ιστόγραμμα

Το βασικό κριτήριο για την κατασκευή ιστογράμματος είναι ότι το ποσοστό του ιστογραμματικού εμβαδού κάθε ομάδας πρέπει να είναι ίσο προς την σχετική συχνότητα της ομάδας. Έτσι ο σχεδιασμός του ιστογράμματος γίνεται θέτοντας στον άξονα των y τις τιμές f_i/d_i , $i = 1, 2, \dots, k$, όπου $d_i = U_i - L_i$. Εάν οι ομάδες είναι ίσες, δηλαδή έχουν το ίδιο μήκος ($d_1 = d_2 = \dots = d_k$) τότε το d μπορεί να απορροφηθεί στην κλίμακα του άξονα των y . Έτσι οι τιμές στο y θα παριστάνουν τις απλές συχνότητες f_i . Εάν στον άξονα των y θέσουμε τις τιμές f_i/nd_i τότε το συνολικό εμβαδόν του ιστογράμματος θα ισούται με την μονάδα.

Το ιστόγραμμα κατασκευάζεται με βάση τον πίνακα συχνοτήτων ομαδοποιημένων μετρήσεων και με τη συσχέτιση εμβαδού και συχνότητας (ή και σχετικής συχνότητας) επιτυγχάνεται μία γραφική και οπτική αξιολόγηση της κατανομής των μετρήσεων. Αν η πρώτη ή η τελευταία ομάδα είναι "ανοικτή", παραλείπεται στην κατασκευή του ιστογράμματος.

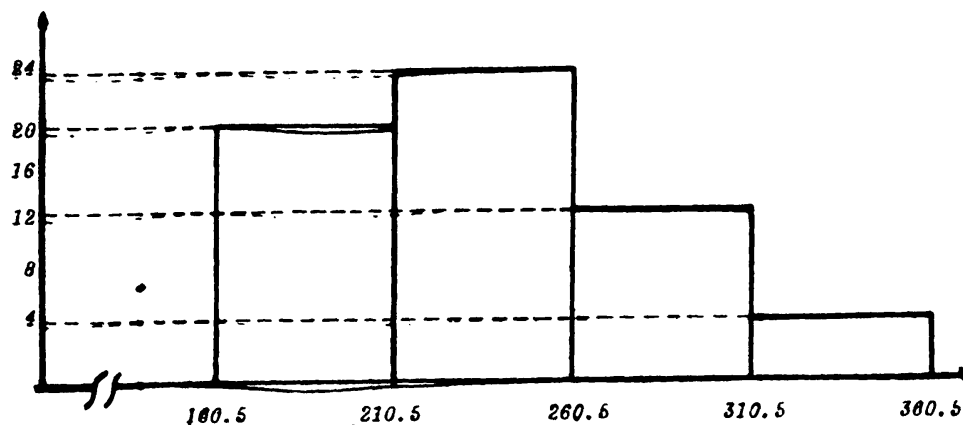
Το ιστόγραμμα των δεδομένων του Παραδείγματος 2.4 (βλ. Πίνακα 2.3) είναι:



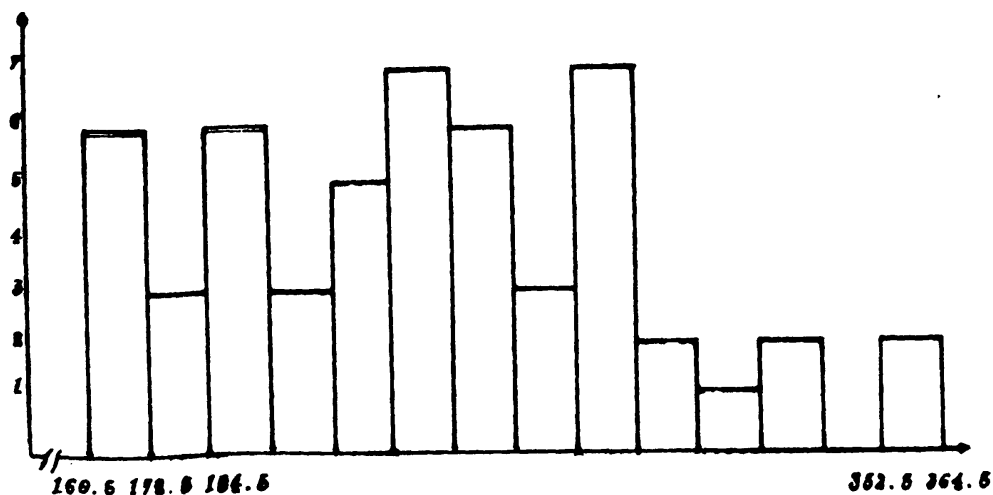
Σχήμα 2.2 Ιστόγραμμα χοληστερίνης με μήκος ομάδων 25



Όταν το μήκος των ομάδων είναι 50 και 12 τότε προκύπτουν τα ακόλουθα ιστογράμματα.



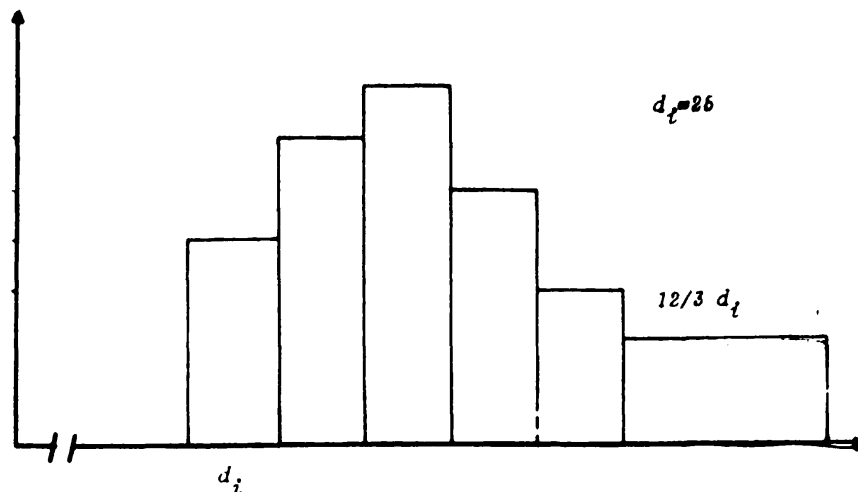
Σχήμα 2.3 Ιστόγραμμα χοληστερίνης με μήκος ομάδων 50



Σχήμα 2.4 Ιστόγραμμα χοληστερίνης με μήκος ομάδων 12



Το μήκος των ομάδων επηρεάζει την εικόνα του ιστογράμματος. Όσο πιο μεγάλο είναι το μήκος των ομάδων τόσο πιο ασαφές και ακαθόριστο είναι το σχήμα της κατανομής των μετρήσεων (βλ. σχήμα 2.3). Αν οι ομάδες είναι πολλές και έχουν μικρό μήκος, το ιστόγραμμα παρουσιάζει ανωμαλίες που αντανακλούν την μεταβλητικότητα της δειγματοληψίας (βλ. σχήμα 2.4).



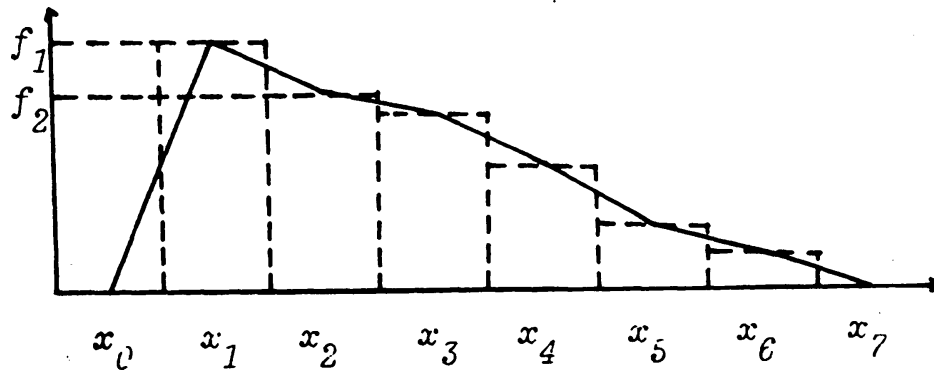
Σχήμα 2.5 Ιστόγραμμα χοληστερίνης με άνισες ομάδες

Αν ενοποιήσουμε τις τρεις τελευταίες ομάδες του πίνακα 2.3 προκύπτει το ιστόγραμμα με άνισες ομάδες του σχήματος 2.5.

β. Πολύγωνο συχνοτήτων: Το πολύγωνο συχνοτήτων αποτελεί το δεύτερο τρόπο γραφικής σύμπτυξης και παρουσίασης των μετρήσεων. Κατασκευάζεται ενώνοντας με ευθείες γραμμές τα μέσα των πάνω πλευρών των ορθογωνίων του ιστογράμματος, όπως φαίνεται στο Σχήμα 2.6.

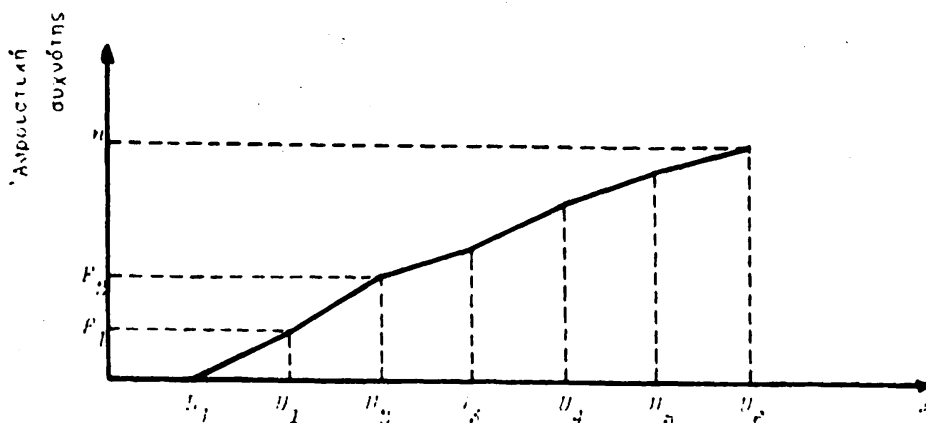
Από πλευράς χρησιμότητας το πολύγωνο συχνοτήτων είναι ισοδύναμο με το ιστόγραμμα.





Σχήμα 2.6 Πολύγωνο συχνοτήτων

γ. Πολύγωνο αθροιστικών συχνοτήτων: Το πολύγωνο αυτό κατασκευάζεται χρησιμοποιώντας τις αθροιστικές συχνότητες και ενώνοντας με ευθείες γραμμές τα ακόλουθα σημεία (x, y) του επιπέδου: $(L_1, 0), (U_1, F_1), (U_2, F_2), \dots, (U_k, n)$ ή ισοδύναμα $(L_1, 0), (U_1, F_1/n), (U_2, F_2/n), \dots, (U_k, 1)$. Χρησιμοποιώντας το ιστόγραμμα του Σχήματος 2.1 και το $1/n$ της κλίμακας έχουμε:



Σχήμα 2.7 Πολύγωνο αθροιστικών συχνοτήτων



Όπως και στην περίπτωση πινάκων συχνότητας έτσι και τώρα ο κάθε ένας από τους γραφικούς τρόπους σύμπτυξης παρουσιάζει ιδιαίτερα πλεονεκτήματα. Συνεπώς είναι αδύνατη η υπόδειξη του πλέον ενδεδειγμένου τρόπου γραφικής παρουσίασης των μετρήσεων. Ο σκοπός και η προβλεπόμενη χρήση της γραφικής παράστασης αποτελούν το καλύτερο κριτήριο επιλογής. Εάν για παράδειγμα σκοπός της σύμπτυξης είναι η εύρεση και αξιολόγηση της συγκριτικής θέσης μελλοντικών μετρήσεων τότε το πολύγωνο αθροιστικών συναρτήσεων είναι η πλέον ενδεδειγμένη γραφική σύμπτυξη.

Εκτός από τους παραπάνω γραφικούς τρόπους παρουσίασης των δεδομένων υπάρχουν και άλλοι, όπως τα ραβδογράμματα (απλά ή σύνθετα), τα ειδολοδιαγράμματα, τα κυκλικά διαγράμματα, οι κύλινδροι, οι έγχρωμοι άτλαντες κ.λ.π. Όλα αυτά αποτελούν παραλλαγές των παραπάνω τριών γραφικών τρόπων. Ο αναγνώστης παραπέμπεται στη σχετική βιβλιογραφία.

2.2.3 Αριθμητικά μεγέθη - Χαρακτηριστικά των μετρήσεων

Οι αριθμητικές μέθοδοι περιγραφής και σύμπτυξης των μετρήσεων χρησιμοποιούν αριθμητικές ποσότητες ή χαρακτηριστικά τα οποία κατά κάποιο τρόπο περιγράφουν τις μετρήσεις. Ο βασικός σκοπός είναι η εξακρίβωση της κατανομής του πληθυσμού και τα χαρακτηριστικά αυτά αποσκοπούν στην απόδοση λεπτομερειών της κατανομής όπως η θέση, η έκταση (εύρος), η συμμετρία, η λοξότητα, η κυρτότητα, η αιχμή κ.λ.π. της κατανομής. Διακρίνουμε τρεις κατηγορίες περιγραφικών μέτρων:

α. Μέτρα θέσης: Συχνά τα αριθμητικά δεδομένα παρουσιάζουν το χαρακτηριστικό γνώρισμα να συγκεντρώνονται γύρω από κάποια "κεντρική" τιμή. Αυτό το φαινόμενο είναι γνωστό σαν κεντρική τάση των μετρήσεων. Η κεντρική τιμή καθορίζει την θέση



του κέντρου ή μέσου της κατανομής των μετρήσεων και ανάλογα με το τι εννοούμε με τον όρο κέντρο προκύπτουν τα διάφορα μέτρα θέσης. Τα μέτρα αυτά είναι αντίστοιχα των μέτρων θέσης της (θεωρητικής) κατανομής του πληθυσμού και είναι η μέση τιμή, η διάμεσος, η κορυφή κ.λ.π.

Ο αριθμητικός μέσος όρος των μετρήσεων ή η δειγματική μέση τιμή ή η μέση τιμή του δείγματος $x_1, \dots, x_n$ συμβολίζεται με $\bar{x}$ και ορίζεται από τη σχέση

$$\bar{x} = \left(\sum_{i=1}^n x_i \right) / n. \quad (2.1)$$

Μετρείται με τις ίδιες μονάδες όπως και οι αρχικές μετρήσεις (cm, gr, δραχμές κ.λ.π.).

Στην περίπτωση μετρήσεων με συχνότητες ή ομαδοποιημένων μετρήσεων η μέση τιμή δίδεται από τη σχέση

$$\bar{x} = \left(\sum_{i=1}^K f_i x_i \right) / n \quad (2.2)$$

όπου x_i η τιμή της ομάδας i και f_i η συχνότητα της.

Για τα δεδομένα του Παραδείγματος 2.4 έχουμε

$$\bar{x} = \frac{239+212+\dots+284}{60} = \frac{14427}{60} = 240.45.$$

Αν όμως χρησιμοποιήσουμε την ομαδοποίηση του Πίνακα 2.3 έχουμε

$$\begin{aligned} \bar{x} &= \frac{173 \times 8 + 198 \times 11 + \dots + 348 \times 2}{60} \\ &= \frac{1384 + 2178 + \dots + 696}{60} = \frac{14330}{60} = 238.85. \end{aligned}$$



Η διαφορά από την προηγούμενη μέση τιμή οφείλεται σε απώλεια πληροφορίας λόγω ομαδοποιήσεως.

Η μέση τιμή του δείγματος με βάρη ή συντελεστές $w_1, \dots, w_n$ δίνεται από τη σχέση

$$\bar{X} = \left(\sum_{i=1}^n w_i x_i \right) / \left(\sum_{i=1}^n w_i \right) \quad (2.3)$$

Τα βάρη εκφράζουν την σημασία που αποδίδουμε σε κάθε τιμή. Αν έχουμε τρία δείγματα με μεγέθη n_1, n_2 και n_3 και μέσες τιμές $\bar{X}_1, \bar{X}_2, \bar{X}_3$ αντίστοιχα, η μέση τιμή $\bar{X}$ όλων των μετρήσεων είναι η μέση τιμή των $\bar{X}_1, \bar{X}_2, \bar{X}_3$ με βάρη n_1, n_2, n_3 δηλαδή

$$\bar{X} = \bar{X}_w = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + n_3 \bar{X}_3}{n_1 + n_2 + n_3}$$

Η τιμή αυτή είναι διαφορετική από την $(\bar{X}_1 + \bar{X}_2 + \bar{X}_3)/3$ εκτός εάν $\bar{X}_1 = \bar{X}_2 = \bar{X}_3$ ή $n_1 = n_2 = n_3$.

Η διάμεσος των μετρήσεων ή η δειγματική διάμεσος ή η διάμεσος του δείγματος είναι η μέση μέτρηση στην κατά σειρά μεγέθους διάταξη τους.

Έτσι αν $X(1), X(2), \dots, X(n)$ είναι ένα διατεταγμένο δείγμα και n περιττό με διακεκριμένη τη μέτρηση $X((n+1)/2)$ (δηλαδή με συχνότητα ίση με ένα) τότε η διάμεσος του δείγματος είναι η $X((n+1)/2)$ μέτρηση. Αν το n είναι άρτιο και οι μετρήσεις $X(n/2), X(n/2+1)$ έχουν συχνότητα ένα, τότε η δειγματική διάμεσος είναι οποιοδήποτε σημείο μεταξύ των μετρήσεων $X(n/2)$ και $X(n/2+1)$.

Συνήθως παίρνουμε το σημείο

$$M = \frac{1}{2} [X(n/2) + X(n/2+1)].$$



Αν οι παραπάνω προϋποθέσεις δεν πληρούνται τότε ως διάμεσο παίρνουμε το σημείο με την ιδιότητα τουλάχιστον 50% των μετρήσεων να είναι μικρότερες ή ίσες από αυτό και τουλάχιστον 50% να είναι μεγαλύτερες ή ίσες. Έτσι για τα δεδομένα του παραδείγματος 2.2. η διάμεσος είναι το σημείο 0.

Αν οι μετρήσεις είναι ομαδοποιημένες (βλ. Πίνακα 2.2) τότε ως διάμεσος του δείγματος λαμβάνεται το σημείο M του άξονα των x με την ιδιότητα η από το M παράλληλος προς τον άξονα των y να χωρίζει το εμβαδόν του ιστογράμματος σε δύο ίσα μέρη. Ο ορισμός αυτός της διαμέσου στηρίζεται στην παραδοχή ότι οι μετρήσεις είναι ομοιόμορφα κατανεμημένες σε κάθε ομάδα. Με τον συμβολισμό της § 2.2.1γ και υπό την προϋπόθεση ότι τα μήκη των ομάδων είναι ίσα προς d , η διάμεσος M ευρίσκεται εντός της ομάδας i για την οποία

$$F_{i-1} < n/2 \leq F_i, \quad i=1,2, \dots, k, \quad F_0 = 0$$

και

$$dF_{i-1} + (M-L_i)f_i = (1/2) \, dn$$

ή

$$M = L_i + \frac{d}{f_i} \left(\frac{n}{2} - F_{i-1} \right). \quad (2.4)$$

Για τα ομαδοποιημένα δεδομένα του Πίνακα 2.3 η διάμεσος M ευρίσκεται στην $i=3$ η ομάδα ($F_2 = 19 < 60/2 < F_3=32$). Έτσι:

$$M = L_3 + \frac{d}{f_3} \left(\frac{n}{2} - F_2 \right) = 210.5 + \frac{25}{13} \left(\frac{60}{2} - 19 \right) = 231.66.$$

Αν ο αριθμός των μετρήσεων είναι μεγάλος (π.χ. μεγαλύτερος του 25 ή 30) η ιδέα της διαμέσου επεκτείνεται με τη χρήση των τεταρτημορίων και των εκατοσπαιών σημείων..

Έστω p ένα από τα κλάσματα 0.01, 0.02, ..., 0.99. Προφανώς το p εκφράζεται και ως $100.p\%$. Το $100.p$ -στό



εκατοστιαίο σημείο του δείγματος είναι η τιμή εκείνη με την ιδιότητα αν οι μετρήσεις διαταχθούν από τη μικρότερη προς τη μεγαλύτερη, τουλάχιστον $100p\%$ των μετρήσεων ευρίσκονται προς τα αριστερά (μικρότερες ή ίσες) της τιμής αυτής και τουλάχιστον $100(1-p)\%$ ευρίσκονται προς τα δεξιά (μεγαλύτερες ή ίσες).

Το πρώτο (κάτω) τεταρτημόριο είναι το εικοστό πέμπτο (25) εκατοστιαίο σημείο, το δεύτερο τεταρτημόριο (διάμεσος) είναι το πενήτηκοστό (50) εκατοστιαίο σημείο και το τρίτο (άνω) τεταρτημόριο είναι το εβδομηκοστό πέμπτο (75) εκατοστιαίο σημείο.

Αν δύο τιμές του δείγματος ικανοποιούν τον παραπάνω ορισμό, τότε συνήθως χρησιμοποιούμε τον μέσο όρο τους ως εκατοστιαίο σημείο.

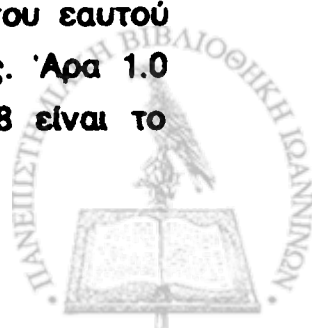
Παράδειγμα 2.5: Δίδονται οι εντάσεις, σε κλίμακα Richter, 35 σεισμικών δονήσεων σε μία σειсмоγενή περιοχή διατεταγμένες κατά σειρά μεγέθους:

0.1	0.5	1.4	2.0	3.1	4.6	5.7
0.2	0.6	1.4	2.1	3.6	4.8	5.7
0.3	0.8	1.7	2.6	3.8	4.8	6.8
0.3	1.0	1.7	2.7	4.0	4.9	6.8
0.3	1.4	2.0	2.7	4.0	4.9	8.2

Πίνακας 2.4 Εντάσεις σεισμών

Να βρεθούν τα τεταρτημόρια και η 20-στή εκατοστιαία μέτρηση.

Για το πρώτο τεταρτημόριο πρέπει να μετρήσουμε τουλάχιστον $0.25 \times 35 = 8.75$ μετρήσεις από τη μικρότερη μέτρηση και τουλάχιστον $0.75 \times 35 = 26.25$ μετρήσεις από τη μεγαλύτερη. Η 9-η μέτρηση 1.0 έχει 9 μετρήσεις μικρότερες ή ίσες του εαυτού της και 27 μετρήσεις μεγαλύτερες ή ίσες του εαυτού της. Άρα 1.0 είναι το πρώτο τεταρτημόριο. Ομοίως ευρίσκουμε ότι 3.8 είναι το



τρίτο τεταρτημόριο και 2.6 η διάμεσος. Για το 20-στό εκατοστιαίο τουλάχιστον $0.20 \times 35 = 7$ μετρήσεις πρέπει να είναι ίσες ή μικρότερες και τουλάχιστον 28 ίσες ή μεγαλύτερες. Τα σημεία 0.6 και 0.8 ικανοποιούν τις συνθήκες αυτές. Άρα η 20-στή εκατοστιαία μέτρηση είναι ο μέσος όρος $(0.6 + 0.8)/2 = 0.7$. Αυτό σημαίνει ότι 20% των 35 σεισμών είχαν έκταση μικρότερη του 0.7 της κλίμακας του Richter.

Όπως είδαμε η μέση τιμή είναι ευαίσθητη στην ύπαρξη λίγων ασυνήθιστα μικρών ή μεγάλων μετρήσεων. Η διάμεσος αγνοεί όλες εκτός από τις κεντρικές τιμές του διατεταγμένου δείγματος. Οι επόμενες δύο μέσες τιμές θεωρούνται σαν μέση λύση μεταξύ της μέσης τιμής και της διαμέσου.

(i) Κλαδεμένη μέση τιμή (Trimmed sample mean). Εδώ αφαιρούμε όλες τις μετρήσεις που είναι μικρότερες από το πρώτο τεταρτημόριο και όλες τις μετρήσεις που είναι μεγαλύτερες από το τρίτο τεταρτημόριο. Στη συνέχεια υπολογίζουμε τη μέση τιμή των μετρήσεων που έχουν απομείνει.

(ii) Μέση τιμή του Winsor (Winsorized sample mean). Εδώ αντικαθιστούμε κάθε μέτρηση που είναι μικρότερη από το πρώτο τεταρτημόριο με την τιμή του πρώτου τεταρτημορίου. Αντικαθιστούμε επίσης κάθε μέτρηση που είναι μεγαλύτερη από το τρίτο τεταρτημόριο με την τιμή του τρίτου τεταρτημορίου. Διατηρούμε αμετάβλητες όλες τις άλλες μετρήσεις. Στη συνέχεια υπολογίζουμε τη μέση τιμή όλων των μετρήσεων, τροποποιηθέντων και μη.

Οι δύο τελευταίες μέσες τιμές δεν επηρεάζονται από την παρουσία ενός μικρού ποσοστού ασυνήθων ή λανθασμένων μετρήσεων που είναι είτε πάρα πολύ μεγάλες είτε πάρα πολύ μικρές. Οι μέσες αυτές τιμές είναι σχεδόν τόσο καλές όσο και η απλή μέση τιμή στην περίπτωση συμμετρικών κατανομών όταν δεν υπάρχουν ασυνήθεις παρατηρήσεις. Συνήθως υπερέχουν όταν υπάρχουν "άγριες" τιμές.

Η κορυφή ή επικρατούσα τιμή των μετρήσεων ή τού δείγματος ή η δειγματική κορυφή είναι, όπως και στην περίπτωση του πληθυσμού, η τιμή με την μεγαλύτερη



εκατοστιαίο σημείο του δείγματος είναι η τιμή εκείνη με την ιδιότητα αν οι μετρήσεις διαταχθούν από τη μικρότερη προς τη μεγαλύτερη, τουλάχιστον $100p\%$ των μετρήσεων ευρίσκονται προς τα αριστερά (μικρότερες ή ίσες) της τιμής αυτής και τουλάχιστον $100(1-p)\%$ ευρίσκονται προς τα δεξιά (μεγαλύτερες ή ίσες).

Το πρώτο (κάτω) τεταρτημόριο είναι το εικοστό πέμπτο (25) εκατοστιαίο σημείο, το δεύτερο τεταρτημόριο (διάμεσος) είναι το πενήτηκοστό (50) εκατοστιαίο σημείο και το τρίτο (άνω) τεταρτημόριο είναι το εβδομηκοστό πέμπτο (75) εκατοστιαίο σημείο.

Αν δύο τιμές του δείγματος ικανοποιούν τον παραπάνω ορισμό, τότε συνήθως χρησιμοποιούμε τον μέσο όρο τους ως εκατοστιαίο σημείο.

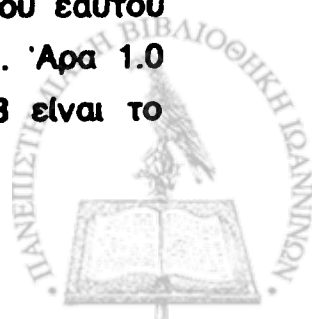
Παράδειγμα 2.5: Δίδονται οι εντάσεις, σε κλίμακα Richter, 35 σεισμικών δονήσεων σε μία σειсмоγενή περιοχή διατεταγμένες κατά σειρά μεγέθους:

0.1	0.5	1.4	2.0	3.1	4.6	5.7
0.2	0.6	1.4	2.1	3.6	4.8	5.7
0.3	0.8	1.7	2.6	3.8	4.8	6.8
0.3	1.0	1.7	2.7	4.0	4.9	6.8
0.3	1.4	2.0	2.7	4.0	4.9	8.2

Πίνακας 2.4 Εντάσεις σεισμών

Να βρεθούν τα τεταρτημόρια και η 20-στή εκατοστιαία μέτρηση.

Για το πρώτο τεταρτημόριο πρέπει να μετρήσουμε τουλάχιστον $0.25 \times 35 = 8.75$ μετρήσεις από τη μικρότερη μέτρηση και τουλάχιστον $0.75 \times 35 = 26.25$ μετρήσεις από τη μεγαλύτερη. Η 9-η μέτρηση 1.0 έχει 9 μετρήσεις μικρότερες ή ίσες του εαυτού της και 27 μετρήσεις μεγαλύτερες ή ίσες του εαυτού της. Άρα 1.0 είναι το πρώτο τεταρτημόριο. Ομοίως ευρίσκουμε ότι 3.8 είναι το



τρίτο τεταρτημόριο και 2.6 η διάμεσος. Για το 20-στό εκατοστιαίο τουλάχιστον $0.20 \times 35 = 7$ μετρήσεις πρέπει να είναι ίσες ή μικρότερες και τουλάχιστον 28 ίσες ή μεγαλύτερες. Τα σημεία 0.6 και 0.8 ικανοποιούν τις συνθήκες αυτές. Άρα η 20-στή εκατοστιαία μέτρηση είναι ο μέσος όρος $(0.6 + 0.8)/2 = 0.7$. Αυτό σημαίνει ότι 20% των 35 σεισμών είχαν έκταση μικρότερη του 0.7 της κλίμακας του Richter.

Όπως είδαμε η μέση τιμή είναι ευαίσθητη στην ύπαρξη λίγων ασυνήθιστα μικρών ή μεγάλων μετρήσεων. Η διάμεσος αγνοεί όλες εκτός από τις κεντρικές τιμές του διατεταγμένου δείγματος. Οι επόμενες δύο μέσες τιμές θεωρούνται σαν μέση λύση μεταξύ της μέσης τιμής και της διαμέσου.

(i) Κλαδεμένη μέση τιμή (Trimmed sample mean). Εδώ αφαιρούμε όλες τις μετρήσεις που είναι μικρότερες από το πρώτο τεταρτημόριο και όλες τις μετρήσεις που είναι μεγαλύτερες από το τρίτο τεταρτημόριο. Στη συνέχεια υπολογίζουμε τη μέση τιμή των μετρήσεων που έχουν απομείνει.

(ii) Μέση τιμή του Winsor (Winsorized sample mean). Εδώ αντικαθιστούμε κάθε μέτρηση που είναι μικρότερη από το πρώτο τεταρτημόριο με την τιμή του πρώτου τεταρτημορίου. Αντικαθιστούμε επίσης κάθε μέτρηση που είναι μεγαλύτερη από το τρίτο τεταρτημόριο με την τιμή του τρίτου τεταρτημορίου. Διατηρούμε αμετάβλητες όλες τις άλλες μετρήσεις. Στη συνέχεια υπολογίζουμε τη μέση τιμή όλων των μετρήσεων, τροποποιηθέντων και μη.

Οι δύο τελευταίες μέσες τιμές δεν επηρεάζονται από την παρουσία ενός μικρού ποσοστού ασυνήθων ή λανθασμένων μετρήσεων που είναι είτε πάρα πολύ μεγάλες είτε πάρα πολύ μικρές. Οι μέσες αυτές τιμές είναι σχεδόν τόσο καλές όσο και η απλή μέση τιμή στην περίπτωση συμμετρικών κατανομών όταν δεν υπάρχουν ασυνήθεις παρατηρήσεις. Συνήθως υπερέρχουν όταν υπάρχουν "άγριες" τιμές.

Η κορυφή ή επικρατούσα τιμή των μετρήσεων ή τού δείγματος ή η δειγματική κορυφή είναι, όπως και στην περίπτωση του πληθυσμού, η τιμή με την μεγαλύτερη



συχνότητα. Έτσι στο παράδειγμα 2.1 το ύψος 175cm είναι η κορυφή του δείγματος. Για τα ομαδοποιημένα δεδομένα του Πίνακα 2.3 η 3^η ομάδα έχει την μεγαλύτερη συχνότητα και ως κορυφή λαμβάνεται η τιμή της ομάδας αυτής 223 mgr. Είναι φανερό ότι οι κατανομές των μετρήσεων μπορεί να έχουν μία μόνο κορυφή ή καμμία κορυφή ή να είναι πολυκόρυφες.

Το μέσο εύρος (midrange) των μετρήσεων είναι

$$1/2 (\min X_i + \max X_i).$$

Είναι και αυτό άλλο ένα μέτρο θέσης της κατανομής των μετρήσεων, που προφανώς επηρεάζεται από τις άκρες τιμές του δείγματος.

Από τα διάφορα μέτρα θέσης ποιό είναι το καλύτερο ή προτιμώτερο; Η απάντηση στο ερώτημα αυτό εξαρτάται από τον σκοπό και τον τρόπο χρησιμοποίησης των μέτρων αυτών. Έτσι χρειάζεται να ξέρουμε τα πλεονεκτήματα και τα μειονεκτήματα των διαφόρων μέτρων.

Αν η κατανομή των μετρήσεων είναι συμμετρική και μονοκόρυφη, τότε η μέση τιμή $\bar{X}$, η διάμεσος M και η κορυφή K συμπίπτουν. Όσο η κατανομή γίνεται πιο ασυμμετρική τόσο οι διαφορές μεταξύ των μέτρων αυτών γίνονται μεγαλύτερες.

Η μέση τιμή $\bar{X}$ είναι ευαίσθητη στις ακρότατες τιμές του δείγματος ενώ το αντίθετο συμβαίνει για τη διάμεσο M και κορυφή K . Αν η κατανομή των μετρήσεων είναι λοξή τότε η διάμεσος και η κορυφή αποτελούν καλύτερα περιγραφικά μέτρα της θέσης της κατανομής των μετρήσεων.

Η μέση τιμή $\bar{X}$ υπερέχει των άλλων μέτρων θέσης όσον αφορά τις μαθηματικές και στατιστικές ιδιότητες. Έχει μεγαλύτερη ευστάθεια (μικρότερη μεταβλητότητα) και επιδέχεται θεωρητική επεξεργασία πλουσιώτερη και ευκολότερη σε σχέση με τα άλλα μέτρα. Τα τελευταία αυτά πλεονεκτήματα κάνουν την μέση τιμή μοναδική σχεδόν εκλογή για όλους εκτός από τους καθαρά περιγραφικούς σκοπούς.



β. Μέτρα μεταβλητότητας (διασποράς ή διακύμανσης) των μετρήσεων: Οι αριθμητικές τιμές των στατιστικών δεδομένων είναι κατά κανόνα διαφορετικές μεταξύ τους. Παρουσιάζουν δηλαδή την λεγόμενη διασπορά ή διακύμανση άλλες φορές σε μικρό και άλλες φορές σε μεγάλο βαθμό. Η μεταβλητότητα αυτή μετρείται με τα λεγόμενα μέτρα διασποράς ή διακύμανσης. Τα μέτρα αυτά μαζί με τα μέτρα θέσης δίνουν μια πληρέστερη περιγραφή των μετρήσεων, δεδομένου ότι μπορεί να έχουμε ομάδες δεδομένων με την ίδια μέση τιμή ή διάμεσο ή κορυφή αλλά τελείως διαφορετικές διακυμάνσεις.

Το πιο απλό μέτρο διασποράς των μετρήσεων είναι το **δειγματικό εύρος (range)** ή το **εύρος του δείγματος** το οποίο ορίζεται από την σχέση

$$R = \max X_i - \min X_i. \quad (2.5)$$

Παρά την ευκολία υπολογισμού, το εύρος δεν είναι και τόσο ικανοποιητικό μέτρο μεταβλητότητας διότι: (i) στηρίζεται μόνο στις δύο άκρες μετρήσεις και αγνοεί τις άλλες, (ii) καθώς ο αριθμός των μετρήσεων μεγαλώνει το εύρος τείνει να γίνει μεγαλύτερο (έτσι μπορεί να χρησιμοποιηθεί για την σύγκριση της μεταβλητότητας δύο ομάδων δεδομένων μόνο όταν οι ομάδες περιέχουν τον ίδιο αριθμό μετρήσεων) και (iii) παρουσιάζει την μικρότερη ευστάθεια (μεγαλύτερη μεταβλητότητα) σε επαναληπτικές δειγματοληψίες σε σχέση με τα άλλα μέτρα διασποράς εκτός των περιπτώσεων που το μέγεθος του δείγματος είναι πολύ μικρό. Λόγω της ευκολίας υπολογισμού και της ευστάθειας του σε μικρά δείγματα, το εύρος χρησιμοποιείται συχνά στον στατιστικό ποιοτικό έλεγχο όπου τέσσερεις ή πέντε μετρήσεις συνήθως αρκούν.

Η **διακύμανση** (ή **διασπορά**) του δείγματος ή **δειγματική διασπορά** δίδεται από την σχέση

$$S'^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2. \quad (2.6)$$



Αναπτύσσοντας το τετράγωνο και χρησιμοποιώντας ιδιότητες του αθροίσματος Σ έχουμε

$$\begin{aligned}\Sigma(x_i - \bar{x})^2 &= \Sigma(x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \Sigma x_i^2 - 2\bar{x} \Sigma x_i + n\bar{x}^2 \\ &= \Sigma x_i^2 - 2 \left(\frac{\Sigma x_i}{n} \right) \Sigma x_i + n \left(\frac{\Sigma x_i}{n} \right)^2 \\ &= \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}.\end{aligned}$$

Έτσι

$$S'^2 = \frac{1}{n} \left(\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} \right). \quad (2.7)$$

Για τον υπολογισμό του S'^2 αρκεί να υπολογίσουμε το άθροισμα Σx_i και το άθροισμα τετραγώνων Σx_i^2 των μετρήσεων. Ο τύπος (2.7) λέγεται τύπος μηχανής.

Για λόγους που θα φανούν όταν μελετήσουμε την εκτίμηση των στατιστικών παραμέτρων, ως διακύμανση του δείγματος συχνά παίρνεται η ποσότητα

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2.8)$$

η οποία λέγεται και μέση τετραγωνική απόκλιση. Ο τύπος μηχανής για την απόκλιση S^2 είναι

$$S^2 = \frac{1}{n-1} \left(\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} \right) \quad (2.9)$$

Οι διακυμάνσεις S'^2 και S^2 εκφράζουν την ίδια στατιστική έννοια: την μεταβλητότητα του δείγματος. Επίσης η διαφορά μεταξύ των πμών των S'^2 και S^2 είναι ασήμαντη, ιδίως όταν το μέγεθος του δείγματος είναι μεγάλο. Δεν παύουν όμως να έχουν διαφορετικές στατιστικές ιδιότητες όπως θα δούμε πιο κάτω (το S'^2 τείνει να

υποεκτιμά τη διακύμανση σ^2).

Η διακύμανση στηρίζεται στις ποσότητες $x_1 - \bar{x}, \dots, x_n - \bar{x}$ που λέγονται αποκλίσεις από τη μέση τιμή. Ο μέσος όρος των αποκλίσεων αυτών δεν προσφέρεται για μέτρο διακύμανσης διότι

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = 0 \quad (2.10)$$

για οποιοδήποτε σύνολο μετρήσεων $x_1, \dots, x_n$.

Αντ' του μέσου όρου των αποκλίσεων μπορούμε να χρησιμοποιήσουμε τον μέσο όρο των απολύτων αποκλίσεων

$$\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|.$$

Το μέτρο αυτό λέγεται μέση απόλυτη απόκλιση. Παρουσιάζει περιορισμένο ενδιαφέρον λόγω των τεχνικών δυσκολιών της απολύτου τιμής.

Η ποσότης

$$\sum_{i=1}^n (x_i - \bar{x})^2 \quad (2.11)$$

λέγεται άθροισμα τετραγώνων των αποκλίσεων ή λόγω απλότητας άθροισμα τετραγώνων και είναι ίση με το μηδέν εάν και μόνον εάν οι τιμές των μετρήσεων είναι ίσες μεταξύ τους. Ομοίως το άθροισμα τετραγώνων παίρνει μεγάλες τιμές εάν και μόνον εάν οι τιμές των μετρήσεων διαφέρουν (έχουν μεγάλη διασπορά) μεταξύ τους. Το ίδιο ισχύει και για τις ποσότητες S'^2 και S^2 .

Το άθροισμα τετραγώνων συνδέεται με έναν αριθμό, ο οποίος είναι γνωστός ως βαθμός ελευθερίας του αθροίσματος. Ο αριθμός αυτός ορίζεται ως ο αριθμός των τετραγώνων μείον τον αριθμό των ανεξαρτήτων γραμμικών



περιορισμών που υπάρχουν επί των όρων του αθροίσματος. Έτσι το άθροισμα τετραγώνων (2.11) έχει $n-1$ βαθμούς ελευθερίας διότι υπάρχει μόνον ένας γραμμικός περιορισμός, ο (2.10) με την βοήθεια του οποίου μπορεί να προσδιορισθεί η n -οστή απόκλιση όταν δίδονται οι άλλες $n-1$ αποκλίσεις. Οι βαθμοί ελευθερίας συνοδεύουν και την ποσότητα S^2 .

Η τυπική απόκλιση του δείγματος ή δειγματική τυπική απόκλιση είναι η τετραγωνική ρίζα της διακύμανσης δηλαδή

$$S' = \sqrt{S^2} \quad \text{ή} \quad S = \sqrt{S^2}.$$

Η τυπική απόκλιση εκφράζει την μεταβλητότητα των μετρήσεων κατά περισσότερο άμεσο τρόπο δεδομένου ότι μετρείται με τις ίδιες μονάδες μετρήσεως όπως και τα αρχικά δεδομένα σε αντίθεση με την διακύμανση που μετρείται με το τετράγωνο των μονάδων.

Ως συντελεστής μεταβλητότητας των μετρήσεων ορίζεται η ποσότης

$$C.V. = \frac{S}{\bar{X}} \quad \text{ή} \quad C.V. = \frac{S}{\bar{X}} \times 100\% \quad (2.12)$$

(C.V. - coefficient of variation), η οποία είναι καθαρός αριθμός και εκφράζει την μεταβλητότητα των μετρήσεων απαλλαγμένη από την επίδραση της μέσης τιμής.

Η χρησιμότητα του συντελεστή μεταβλητότητας φαίνεται από το ακόλουθο παράδειγμα: Ένα σύστημα ή μέθοδος κατασκευάζει σωλήνες (π.χ. κάλυκες) μίας ορισμένης διαμέτρου. Έστω ότι η τυπική απόκλιση της διαμέτρου είναι 0.05 cm. Πότε το σύστημα (μέθοδος) έχει περισσότερη ακρίβεια όταν η μέση διάμετρος είναι 10 cm ή 1/2 cm;

Ας υπολογίσουμε τους συντελεστές μεταβλητότητας:



$$\text{με μέση διάμετρο } 10\text{cm} , \text{C.V.} = \frac{0.05}{10} \times 100\% = 0.5\%$$

$$\text{με μέση διάμετρο } 0.5\text{cm} , \text{C.V.} = \frac{0.05}{0.5} \times 100\% = 10\%.$$

Άρα όταν η μέση διάμετρος είναι 10cm ο συντελεστής μεταβλητότητας είναι μικρότερος και το σύστημα έχει μεγαλύτερη ακρίβεια.

Παράδειγμα 2.6: Για τα δεδομένα του παραδείγματος 2.4 έχουμε:

$$\sum x_i = 14.427 , \sum x_i^2 = 3615497$$

$$S'^2 = \frac{1}{60} \left(3615497 - \frac{(14.427)^2}{60} \right) = 2442.08$$

$$S' = 49.42$$

$$S^2 = \frac{1}{59} \left(3615497 - \frac{(14.427)^2}{60} \right) = 2483.42$$

$$S = 49.83.$$

Είναι φανερό ότι η δειγματική διακύμανση και η τυπική απόκλιση S'^2 και S' αντιστοιχούν εννοιολογικά με την διακύμανση και την τυπική απόκλιση σ^2 και σ του πληθυσμού. Η στατιστική σημασία της διακύμανσης και κατά συνέπεια της τυπικής απόκλισης καταφαίνεται από το Θεώρημα του Chebyshev: για οποιοδήποτε τυχαίο δείγμα $x_1, x_2, \dots, x_n$ τουλάχιστον $(1 - 1/k^2)$ των μετρήσεων ($k = 2, \dots$) θα ευρίσκεται σε απόσταση k τυπικών αποκλίσεων από την μέση τιμή, δηλαδή εντός του διαστήματος $[(\bar{X} - ks, \bar{X} + ks)]$.

Στην περίπτωση μετρήσεων με συχνότητα ή



ομαδοποιημένων μετρήσεων η διακύμανση δίδεται από τις σχέσεις

$$s^2 = \frac{1}{n-1} \sum_{i=1}^k f_i (x_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum f_i x_i^2 - \frac{(\sum f_i x_i)^2}{n} \right).$$

Παράδειγμα 2.7: Για τα ομαδοποιημένα δεδομένα του παραδείγματος 2.4 έχουμε:

$$\sum f_i x_i = 14427, \quad \sum f_i x_i^2 = 3556190$$

$$s^2 = \frac{1}{59} [3556190 - 3468972.15] = 1478.27$$

$$s = 38.45.$$

γ. Ροπές του δείγματος: Οι ροπές του δείγματος ορίζονται όπως οι ροπές του πληθυσμού και έχουν την ίδια εννοιολογική σημασία. Έτσι έχουμε:

Η δειγματική ροπή k τάξης περί το μηδέν είναι:

$$m_k = \frac{1}{n} \sum_{i=1}^n x_i^k, \quad k=1,2,\dots$$

Η δειγματική κεντρική ροπή k τάξης είναι:

$$u_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k, \quad k=1,2,\dots$$



Η δειγματική τυπική ή κανονικοποιημένη ροπή k τάξης είναι:

$$a_k = \frac{1}{n} \sum_{i=1}^n \frac{(x_i - \bar{x})^k}{s^k}, \quad k=1,2,\dots$$

Από τις παραπάνω ροπές πιο χρήσιμες είναι η $m_1 = \bar{x}$, $s_2 = s^2$, a_3 ο δειγματικός συντελεστής λοξότητας της κατανομής των μετρήσεων και a_4 ο δειγματικός συντελεστής κυρτότητας της κατανομής. Η ερμηνεία των a_3 και a_4 είναι η ίδια με την ερμηνεία των αντίστοιχων παραμέτρων α_3 και α_4 του πληθυσμού.

ΑΣΚΗΣΕΙΣ

- 2.1 Για κάθε ομάδα μετρήσεων να βρεθούν (α) η μέση τιμή, (β) η διάμεσος, (γ) η κορυφή, (δ) το εύρος, (ε) η διακύμανση S^2 και (στ) η τυπική απόκλιση S .

(i)	1,	3,	3,	5,	6,	6			
(ii)	-4,	2,	0,	-6,	-4,	6,	2,	0,	4
(iii)	16,	22,	2,	8,	6,	20,	14,	24.	

- 2.2 Δίδονται οι χρόνοι ζωής 60 λυχνιών ενός ορισμένου τύπου. Ο χρόνος ζωής μετρείται σε ώρες με ακρίβεια μετρήσεων την πλησιέστερη ώρα.

365	510	381	354	301	497
321	250	445	462	373	442
331	364	501	286	451	444
364	382	460	404	342	473
271	388	337	438	552	351
342	301	265	309	448	573
455	373	457	399	395	273
332	554	308	443	461	412
270	324	251	299	605	341
539	370	306	541	398	366

- (i) Να γίνει ομαδοποίηση των δεδομένων σε δέκα ομάδες.
- (ii) Να κατασκευαστεί το ιστόγραμμα και το πολύγωνο αθροιστικών συχνοτήτων.
- (iii) Για τις ομαδοποιημένες μετρήσεις να βρεθεί η μέση τιμή, η διάμεσος και η τυπική απόκλιση S .
- (iv) Να βρεθούν τα τεταρτημόρια και το δέκατο εκατοστιαίο σημείο.



2.3 Να δειχθούν οι σχέσεις

$$v_2 = m_2 - m_1^2$$

$$v_3 = m_3 - 3m_1m_2 + 2m_1^3$$

$$v_4 = m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4 .$$

2.4 Να δειχθεί ότι το άθροισμα των τετραγώνων των αποκλίσεων η μετρήσεων από μία τιμή α γίνεται ελάχιστο αν και μόνο αν $\alpha = \bar{X}$.

2.5 Να δειχθεί ότι

$$E(m_k) = \mu_k.$$

2.6 Να δειχθεί ότι

$$(i) \quad S^2 = \frac{1}{2n(n-1)} \sum_{i=1}^n \sum_{j=1}^n (X_i - X_j)^2$$

$$(ii) \quad \text{Var}(S^2) = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right),$$

$$\mu_k = E(X - \mu)^k .$$



1950

10-10-68

1. The following information is being furnished to you for your information:

SECRET, (S) TO ADJUTANT (a) n

$$f(m) = \frac{1}{2}m^2 - \frac{1}{2}m + \frac{1}{2} = \frac{1}{2}(m-1)^2 + \frac{1}{2}$$

13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 1033 1034 1035 1036 1037 1038 1039 1040 1041 1042 1043 1044 1045 104

1. The Commission has received information that the following individuals have been identified as being involved in the activities of the Communist Party, U.S.A., and its front organizations, in the State of New York:

[illegible]

$$(A) \frac{1}{1.5} = (52) \frac{1}{2} = (52) 76V \quad (11)$$

CONFIDENTIAL

14-00000-15

TO: THE DIRECTOR, FBI, WASHINGTON, D. C. FROM: SAC, NEW YORK (100-100000) SUBJECT: JAMES EARL RAY, AKA; ALLEGED ATTEMPT TO OBTAIN PASSPORT FOR VIETNAM; RE: NEW YORK TELETYPE TO BUREAU, 1/10/68.

1. The following information was obtained from the report of the
1. The following information was obtained from the report of the

SECRET

ΚΕΦΑΛΑΙΟ 3

ΣΤΑΤΙΣΤΙΚΑ ΔΕΙΓΜΑΤΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

3.1 Δειγματοληψία και Συμπερασματολογία Δειγματικές κατανομές

Η εύρεση των άγνωστων ποσοτικών ή ποιοτικών στοιχείων ενός πληθυσμού γίνεται με τη βοήθεια δειγμάτων που εκλέγονται απ'αυτόν. Έστω ότι στις δημοτικές εκλογές, ο πληθυσμός μιας πόλης είναι διχασμένος σε εκείνους που υποστηρίζουν τον υποψήφιο Α και εκείνους που υποστηρίζουν τον υποψήφιο Β. Μία σφυγμομέτρηση της κοινής γνώμης επιδιώκει να προβλέψει το αποτέλεσμα της εκλογής. Έτσι εκλέγεται ένα τυχαίο δείγμα 100 ψηφοφόρων (συνήθως χωρίς επανατοποθέτηση) και έστω X ο αριθμός εκείνων που υποστηρίζουν τον υποψήφιο Α. Έστω επίσης p το ποσοστό των ψηφοφόρων της πόλης που υποστηρίζουν τον Α. Αν το p ήταν γνωστό δεν θα υπήρχε ανάγκη πρόβλεψης ή δειγματοληψίας. Η σφυγμομέτρηση προσπαθεί να "μαντέψει" το p βάσει του X . Αν π.χ. $X=70$ ξέρουμε, και χωρίς τη χρήση της Στατιστικής Ανάλυσης, ότι το p θα είναι γύρω στο 70%. Είναι απίθανο το p να είναι 25% και να έχουμε ένα $X=70$. Επίσης ξέρουμε ότι αν το μέγεθος του δείγματος ήταν 1000 αντί 100 θα μπορούσαμε να κάνουμε ακριβέστερη "εκτίμηση" του p . Σκοπός μας είναι να πάμε πιο κάτω από αυτές τις απλές και αρχικές ιδέες και να καταλάβουμε λεπτομερέστερα πως γίνεται η "εκτίμηση" του p και τι είδος ακρίβειας μπορούμε να έχουμε.

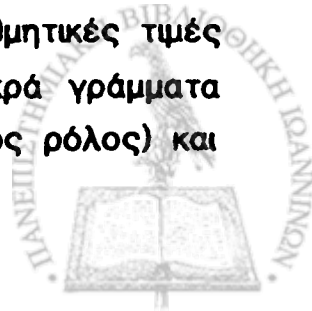


Τα άγνωστα στοιχεία του πληθυσμού συνήθως εκφράζονται συναρτήσει αριθμών που λέγονται (στατιστικά) παράμετροι. Π.χ. το ρ του παραπάνω δείγματος, ο μέσος χρόνος μεταξύ δύο διαδοχικών κλήσεων σένα τηλεφωνικό κέντρο, το μ της κανονικής κατανομής, η $P(X > 1)$ κ.λ.π. Τα συμπεράσματα για μια παράμετρο εξάγονται με τη βοήθεια των στατιστικών (ή στατιστικών συναρτήσεων), τα οποία είναι αριθμοί που μπορούν να υπολογιστούν από το δείγμα. Π.χ. το $\bar{X}$ του παραπάνω παραδείγματος, το $\bar{X}$, το S^2 κ.λ.π.

Ορισμός 3.1: Μία συνάρτηση των μεταβλητών $X_1, X_2, \dots, X_n$ ενός τυχαίου δείγματος που δεν περιέχει άγνωστες παραμέτρους λέγεται στατιστικό ή στατιστική συνάρτηση (σ.σ.).

Τα στατιστικά συμβολίζονται συνήθως με $T = T(X_1, X_2, \dots, X_n)$, $S = S(X_1, X_2, \dots, X_n)$ κ.λ.π.

Από τα προηγούμενα είναι φανερό ότι τόσο οι τιμές X_i ενός τυχαίου δείγματος όσο και τα στατιστικά T παίζουν ένα διπλό ρόλο: αφ'ενός μεν είναι γνωστοί αριθμοί (έχουν γνωστές αριθμητικές τιμές), αφ'ετέρου είναι και τυχαίες μεταβλητές. Στο προηγούμενο παράδειγμα $X=70$. Το X έχει την τιμή 70 αλλά είναι και μία διωνυμική τ.μ. με παραμέτρους (n, p) . Οι στατιστικές συναρτήσεις ως συναρτήσεις τυχαίων μεταβλητών είναι επίσης τυχαίες μεταβλητές. Παίρνουν τις διάφορες τιμές τους έπειτα από n ο.ε.ρ. επαναλήψεις του πειράματος που γεννά τις τιμές του τυχαίου δείγματος. Συνήθως θα συμβολίζουμε τις αριθμητικές τιμές του τυχαίου δείγματος ή των στατιστικών με μικρά γράμματα $x_1, x_2, \dots$ ή $y_1, y_2, \dots$, t , $\bar{x}$, s^2 κ.λ.π. (πρώτος ρόλος) και



τιστυχαίες μεταβλητές με κεφαλαία γράμματα $X_1, X_2, \dots, Y_1, Y_2, \dots, T, \bar{X}, S^2$ κ.λ.π. (δεύτερος ρόλος). Οι κατανομές πιθανότητας των στατιστικών λέγονται **δειγματικές κατανομές**. Έτσι ο $\bar{X}$, τα S^2, S κ.λ.π. έχουν από μία δειγματική κατανομή το κάθε ένα. Τα στατιστικά συμπεράσματα εξάγονται χρησιμοποιώντας τις κατανομές αυτές.

Αν ενδιαφερόμαστε για το μέσο βάρος μ (σε κιλά) του πληθυσμού των ψαριών στη λίμνη Ιωαννίνων, το μ είναι μία άγνωστη παράμετρος και προφανώς θα μείνει για πάντα άγνωστη. Για να "εκτιμήσουμε" το μ παίρνουμε ένα δείγμα ψαριών και υπολογίζουμε το μέσο βάρος τους $\bar{x}$. Με βάσει το $\bar{x}$ βγάζουμε συμπέρασμα για το μ . Χωρίς στατιστική θεωρία, ξέρουμε ότι το $\bar{x}$ "εκτιμά" κάπως το μ , αλλά για να μάθουμε πως να βγάζουμε συμπεράσματα και πόση ακρίβεια θα έχουν τα συμπεράσματα αυτά, χρειαζόμαστε να ξέρουμε την κατανομή του $\bar{X}$, τη δειγματική κατανομή της τ.μ. $\bar{X}$.

Είναι φανερό ότι η εύρεση των δειγματικών κατανομών στηρίζεται στη μεθοδολογία της αλλαγής μεταβλητών. Πριν προχωρήσουμε στη μελέτη των δειγματικών κατανομών θα δώσουμε μερικά γενικά αποτελέσματα που αφορούν τις μέσες τιμές και διακυμάνσεις των χαρακτηριστικών του δείγματος και τα οποία ισχύουν χωρίς συγκεκριμένη υπόθεση για το είδος της κατανομής του πληθυσμού.

3.2 Μέσες τιμές και διακυμάνσεις των $\bar{X}$ και S^2

Εφ'όσον τα στατιστικά $\bar{X}, S^2$ κ.λ.π. είναι τυχαίες μεταβλητές είναι σωστό να μιλάμε για τις μέσες τιμές, διακυμάνσεις τους κ.λ.π.



Θεώρημα 3.1: Έστω $X_1, \dots, X_n$ ένα τυχαίο δείγμα από ένα πληθυσμό με μέση τιμή μ και διακύμανση σ^2 . Τότε

$$(i) \quad E(\bar{X}) = \mu \quad (3.1)$$

$$(ii) \quad \text{Var}(\bar{X}) = \sigma^2/n \quad (3.2)$$

$$(iii) \quad E(S^2) = \sigma^2 \quad (3.3)$$

Απόδειξη:

$$(i) \quad E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \mu.$$

$$(ii) \quad \text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var} X_i = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}.$$

$$\begin{aligned} (iii) \quad E(S^2) &= E\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right] \\ &= \frac{1}{n-1} \sum_{i=1}^n E[(X_i - \mu + \mu - \bar{X})^2] \\ &= \frac{1}{n-1} E\left[\sum_{i=1}^n (X_i - \mu)^2 + \sum_{i=1}^n (\bar{X} - \mu)^2 - 2 \sum_{i=1}^n (\bar{X} - \mu)(X_i - \mu)\right] \\ &= \frac{1}{n-1} E\left[\sum_{i=1}^n (X_i - \mu)^2 + n(\bar{X} - \mu)^2 - 2(\bar{X} - \mu)(n\bar{X} - n\mu)\right] \end{aligned}$$



$$\begin{aligned}
 &= \frac{1}{n-1} E\left[\sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2\right] \\
 &= \frac{1}{n-1} \left[\sum_{i=1}^n E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2\right].
 \end{aligned}$$

Βάσει του (ii) έχουμε $E[(\bar{X} - \mu)^2] = \text{Var}(\bar{X}) = \sigma^2/n$.

Έτσι

$$E(S^2) = \frac{1}{n-1} \left[\sum_{i=1}^n \sigma^2 - n\sigma^2/n\right] = \sigma^2.$$

Πόρισμα 3.1: Έστω $X_{11}, X_{12}, \dots, X_{1n_1}$ ένα τυχαίο δείγμα μεγέθους n_1 από ένα πληθυσμό με μέση τιμή μ_1 και διακύμανση σ_1^2 και $X_{21}, X_{22}, \dots, X_{2n_2}$ ένα τυχαίο δείγμα μεγέθους n_2 από ένα άλλο ανεξάρτητο πληθυσμό με μέση τιμή μ_2 και διακύμανση σ_2^2 . Τότε

$$(i) \quad E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 \quad (3.4)$$

$$(ii) \quad \text{Var}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}, \quad (3.5)$$

όπου $\bar{X}_1 = (\sum_{j=1}^{n_1} X_{1j})/n_1$ και $\bar{X}_2 = (\sum_{j=1}^{n_2} X_{2j})/n_2$.



3.3 Κατανομές που απορρέουν από την κανονική

Οι μονοδιάστατες κατανομές που δίδονται στο εδάφιο αυτό απορρέουν από την κανονική κατανομή και παίζουν όπως θα δούμε πρωταρχικό ρόλο στη στατιστική συμπερασματολογία. Το ίδιο ισχύει και για τα αντίστροφα εκατοστιαία σημεία αυτών.

Ορισμός 3.2: Κατανομή χ^2

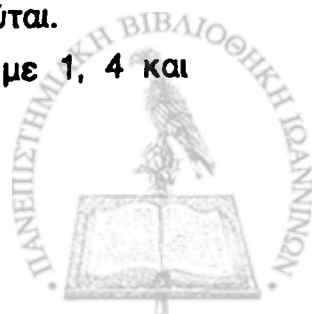
Εστω $X_1, X_2, \dots, X_n$ n ανεξάρτητες και ισόνομες τ.μ. με τυπική κανονική κατανομή. Η κατανομή της τ.μ. $Y = X_1^2 + \dots + X_n^2$ είναι η χ_n^2 , χ^2 - τετράγωνο με n βαθμούς ελευθερίας (β.ε.). Συμβολικά

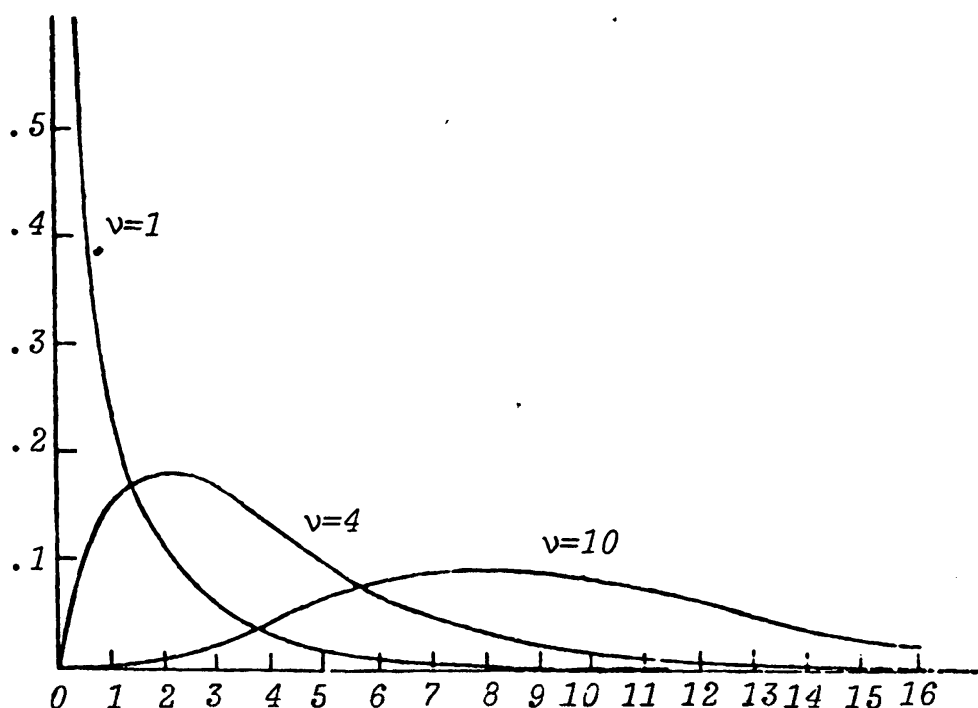
$$\chi_n^2 = \sum_{i=1}^n N(0,1)^2. \quad (3.6)$$

Η χ^2 τετράγωνο κατανομή με n βαθμούς ελευθερίας ορίσθηκε στο εδάφιο 1.5 ως ειδική περίπτωση της Γάμμα κατανομής με $\alpha = n/2$ και $\beta = 2$. Ο ορισμός 3.2 αποτελεί ένα ισοδύναμο τρόπο παρουσίασης της κατανομής χ_n^2 . Πράγματι με τη μέθοδο των ροπογεννητριών μπορούμε να αποδείξουμε ότι η τ.μ. Y του Ορισμού 3.2 έχει κατανομή $G(\alpha = n/2, \beta = 2)$.

Η χ_n^2 κατανομή είναι λοξή προς τα δεξιά και η πυκνότητα της εξαρτάται μόνο από τους βαθμούς ελευθερίας n . Μάλιστα δε καθώς οι βαθμοί ελευθερίας αυξάνουν η λοξότητα ελαττούται.

Το Σχήμα 3.1 μας δίνει τρεις χ^2 κατανομές με 1, 4 και 10 βαθμούς ελευθερίας.





Σχήμα 3.1 χ^2 κατανομές με 1, 4 και 10 βαθμούς ελευθερίας

Μπορεί να αποδειχθεί ότι όταν $v \rightarrow \infty$ η χ^2 κατανομή προσεγγίζεται από την κανονική κατανομή.

Οι συνήθεις πίνακες της χ^2 κατανομής (βλέπε Πίνακα II στο Παράρτημα) δίνουν για διάφορους β.ε. ($v \leq 30$) αντίστροφα (από πάνω) εκατοσπιαία σημεία $\chi_{\alpha, v}^2$ τα οποία ορίζονται από τη σχέση

$$P(\chi_v^2 > \chi_{\alpha, v}^2) = \alpha,$$

(3.7)



π.χ. $\chi_{0.05,20}^2 = 31.41$. Όταν το v είναι μεγαλύτερο του 30 οι πιθανότητες της κατανομής χ^2 συνήθως προσεγγίζονται με τις πιθανότητες της κανονικής κατανομής (βλ. Κεντρικό Οριακό Θεώρημα § 3.5).

Επίσης έχουμε

$$E(\chi_v^2) = v \quad \text{και} \quad \text{Var}(\chi_v^2) = 2v. \quad (3.8)$$

Ορισμός 3.3: Κατανομή t

Εστω $X \sim N(0,1)$, $Y \sim \chi_v^2$ και X, Y ανεξάρτητες τ.μ. Η κατανομή της τ.μ. $Z = \frac{X}{\sqrt{Y/v}}$ λέγεται t_v (τι με v βαθμούς ελευθερίας). Συμβολικά

$$t_v = \frac{N(0,1)}{\sqrt{\chi_v^2/v}}. \quad (3.9)$$

Η κατανομή t_v είναι γνωστή και ως η κατανομή του Student, ψευδώνυμο του W.S. Gosset ο οποίος πρώτος την ανακάλυψε το 1908.

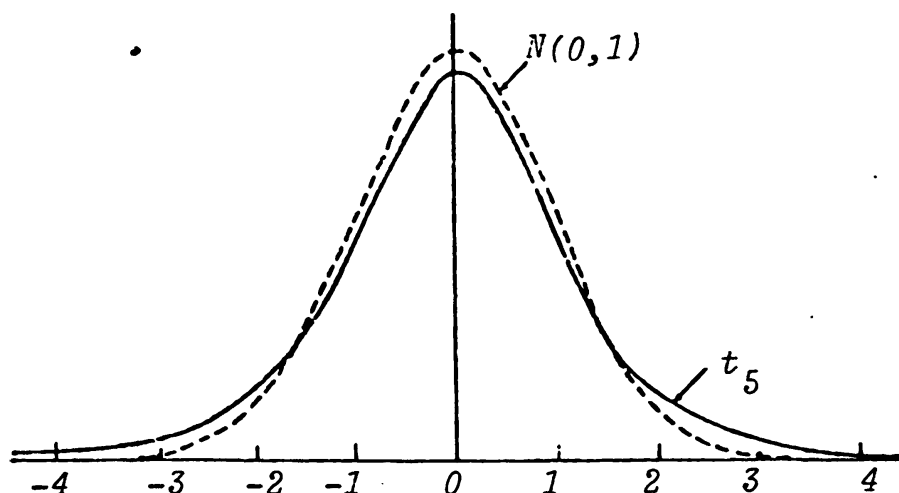
Μπορεί ν'αποδειχθεί ότι η σ.π.π. της t κατανομής δίνεται από τη σχέση

$$f_{t_v}(t) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(v/2)\sqrt{v\pi}} \left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}}, \quad -\infty < t < +\infty.$$



Η t_v κατανομή είναι συμμετρική περί τον άξονα $x = 0$ και έχει μέση τιμή και διακύμανση:

$$E(t_v) = 0 \quad , \quad \text{Var}(t_v) = \frac{v}{v-2} \quad , \quad v > 2 \quad (3.10)$$



Σχήμα 3.2 Κατανομές t_5 και $t_\infty \equiv N(0,1)$

Οι ουρές της κατανομής t έχουν περισσότερη πιθανότητα από τις αντίστοιχες ουρές της τυπικής κανονικής κατανομής η οποία είναι ψηλότερη στο μέσον. Όσο μεγαλώνουν οι βαθμοί ελευθερίας τόσο περισσότερο η κατανομή προσεγγίζει την τυπική κανονική. Στην πραγματικότητα

$$t_\infty \equiv N(0,1) \quad . \quad (3.11)$$

Με ανάλογο τρόπο ορίζονται τα αντίστροφα εκατοστιαία σημεία $t_{\alpha,v}$ της κατανομής t_v , τα οποία δίνονται στους στατιστικούς πίνακες για $t_{\alpha,v} > 0$ και διάφορες τιμές των βαθμών



66 Κεφ. 3. Στατιστικά - Δειγματικές Κατανομές

ελευθερίας ν (Πίνακας III στο Παράρτημα)

$$P(t_{\nu} \geq t_{\alpha, \nu}) = \alpha . \quad (3.12)$$

Λόγω συμμετρίας έχουμε

$$P(t_{\nu} \leq -t_{\alpha, \nu}) = P(t_{\nu} \geq t_{\alpha, \nu}) = \alpha ,$$

$$t_{1-\alpha, \nu} = -t_{\alpha, \nu} .$$

π.χ. $t_{0.01, 15} = 2.602 , \quad t_{0.05, \infty} = 1.645 .$

Ορισμός 3.4: Κατανομή F

Εστω $X_1 \sim \chi_{\nu_1}^2$, $X_2 \sim \chi_{\nu_2}^2$ και X_1, X_2 ανεξάρτητες τ.μ. Η κατανομή της τ.μ. $(X_1/\nu_1)/(X_2/\nu_2)$ λέγεται F_{ν_1, ν_2} (F με ν_1, ν_2 βαθμούς ελευθερίας). Συμβολικά

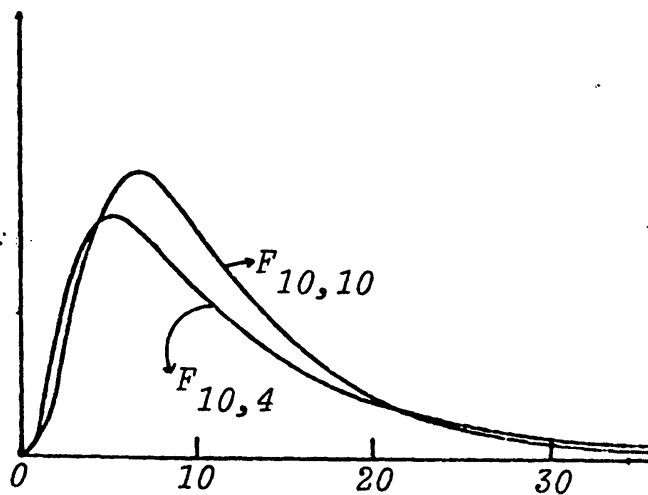
$$F_{\nu_1, \nu_2} = \frac{\chi_{\nu_1}^2 / \nu_1}{\chi_{\nu_2}^2 / \nu_2} . \quad (3.13)$$

Μπορεί να αποδειχθεί ότι η σ.π.π. της F κατανομής δίνεται από τη σχέση



$$F_{v_1, v_2}(x) = \frac{\Gamma\left(\frac{v_1 + v_2}{2}\right) \left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}}}{\Gamma\left(\frac{v_1}{2}\right) \Gamma\left(\frac{v_2}{2}\right)} \frac{x^{\frac{v_1}{2} - 1}}{\left(1 + \frac{v_1 x}{v_2}\right)^{\frac{v_1 + v_2}{2}}}, \quad 0 < x < \infty$$

Η κατανομή F είναι ασυμμετρική. Δύο κατανομές F δίδονται στο Σχήμα 3.3.



Σχήμα 3.3 Κατανομές $F_{10,4}$ και $F_{10,10}$

Η μέση τιμή και η διακύμανση της F_{v_1, v_2} είναι

$$E(F_{v_1, v_2}) = \frac{v_2}{v_2 - 2}, \quad v_2 > 2$$

(3.14a)



$$\text{Var}(F_{v_1, v_2}) = \frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)}, \quad v_2 > 4. \quad (3.14b)$$

Τα αντίστροφα (από πάνω) εκατοστιαία σημεία της F κατανομής ορίζονται από τη σχέση (βλ. Σχήμα 3.3)

$$P(F_{v_1, v_2} \geq F_{\alpha, v_1, v_2}) = \alpha \quad (3.15)$$

και δίνονται σε στατιστικούς πίνακες. (Βλέπε Πίνακα IV στο Παράρτημα για $\alpha=0.05$ και $\alpha=0.01$) Όταν $\alpha > 0.5$ χρησιμοποιούμε τη σχέση

$$F_{1-\alpha, v_1, v_2} = \frac{1}{F_{\alpha, v_2, v_1}}. \quad (3.16)$$

Από τους προηγούμενους ορισμούς εύκολα βλέπουμε ότι ισχύει η σχέση

$$t_v^2 = F_{1, v}. \quad (3.17)$$

Ορισμός 3.5: Έστω $Z \sim N(0, 1)$. Τα αντίστροφα (από πάνω) εκατοστιαία σημεία της τυπικής κανονικής κατανομής ορίζονται από τη σχέση

$$P(Z \geq z_\alpha) = \alpha. \quad (3.18)$$



Ισοδύναμα έχουμε

$$\Phi(z_\alpha) = 1 - \alpha \quad , \quad z_{1-\alpha} = -z_\alpha \quad .$$

Για $\alpha=5\%$, 2.5% και 0.5% έχουμε

$$z_{0.05} = 1.645$$

$$z_{0.025} = 1.96 \quad (3.19)$$

$$z_{0.005} = 2.58 \quad .$$

3.4 Δειγματοληψία από κανονικούς πληθυσμούς

Εστω ότι οι πληθυσμοί από τους οποίους παίρνουμε τα τυχαία δείγματα είναι κανονικοί. Ποίες είναι οι δειγματικές κατανομές των διαφόρων στατιστικών συναρτήσεων; Θα δώσουμε την απάντηση για τις στατιστικές συναρτήσεις $\bar{X}$, S^2 , $(\bar{X}-\mu)\sqrt{n}/S$, S_1^2/S_2^2 , όπου μ είναι η μέση τιμή του πληθυσμού και S_1^2 , S_2^2 οι διακυμάνσεις δύο τυχαίων δειγμάτων από δύο ανεξάρτητους κανονικούς πληθυσμούς.

Θ ε ώ ρ η μ α 3.2 : Εστω $X_1, \dots, X_n$ τυχαίο δείγμα από ένα κανονικό πληθυσμό με μέση τιμή μ και διακύμανση σ^2 . Τότε

(1) η δειγματική κατανομή του $\bar{X}$ είναι κανονική με μέση τιμή μ και διακύμανση σ^2/n , συμβολικά

$$\bar{X} \sim N(\mu, \sigma^2/n) \quad .$$

(3.20)



(ii) η δειγματική κατανομή του $(n-1)S^2/\sigma^2$ είναι χι-τετράγωνο με $n-1$ βαθμούς ελευθερίας, συμβολικά

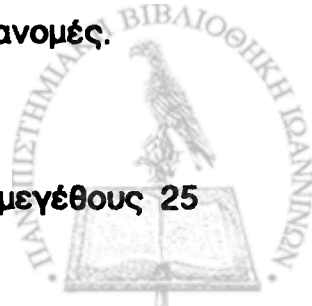
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2 \quad (3.21)$$

(iii) Τα στατιστικά $\bar{X}$ και S^2 είναι ανεξάρτητα.

Απόδειξη : Θα αποδείξουμε μόνο την πρόταση (i). Οι προτάσεις (ii) και (iii) απαιτούν πολυπλοκότερη ανάλυση και ξεφεύγουν από το σκοπό του βιβλίου αυτού. Επειδή οι X_i είναι ανεξάρτητες $N(\mu, \sigma^2)$ τ.μ., είναι εύκολο να αποδειχτεί με τη μέθοδο των ροπογεννητριών, ότι η τ.μ. $\sum X_i$ είναι κανονική με μέση τιμή $\sum \mu_i = n\mu$ και διακύμανση $\sum \sigma_i^2 = n\sigma^2$. Έτσι η ροπογεννήτρια συνάρτηση της $\bar{X} = \sum X_i/n$ είναι

$$\begin{aligned} m_{\bar{X}}(t) &= m_{\sum X_i}(t/n) = e^{n\mu(t/n) + \frac{1}{2}n\sigma^2(t^2/n^2)} \\ &= e^{\mu t} + \frac{\sigma^2}{2n} t^2 \\ &= e^{\mu t} + \frac{1}{2}\left(\frac{\sigma^2}{n}\right)t^2. \end{aligned}$$

Η πρόταση βέβαια μπορεί ν'αποδειχθεί με την βοήθεια του Θεωρήματος 3.1 (i), (ii) και του γεγονότος ότι γραμμικοί συνδιασμοί κανονικών τυχαίων μεταβλητών οδηγούν σε κανονικές κατανομές.



λαμβάνεται από ένα κανονικό πληθυσμό με διακύμανση 625. Να βρεθούν οι πιθανότητες (i) η μέση τιμή του δείγματος να διαφέρει κατά 4 ή και περισσότερο από τη μέση τιμή του πληθυσμού και (ii) η διακύμανση του δείγματος να είναι μεγαλύτερη του 300.

Το Θεώρημα 3.2 μας δίνει

$$\bar{X} \sim N(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} = \frac{625}{25} = 25)$$

και

$$\frac{(n-1)S^2}{\sigma^2} = \frac{24S^2}{625} \sim \chi_{24}^2$$

Έτσι για το πρώτο ερώτημα έχουμε

$$\begin{aligned} P(|\bar{X} - \mu| \geq 4) &= 1 - P(\mu - 4 < \bar{X} < \mu + 4) \\ &= 1 - P\left(-\frac{4}{5} < \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} < \frac{4}{5}\right) \\ &= 1 - P(-0.8 < z < 0.8) \\ &= 1 - (0.2881 + 0.2881) = 0.4238. \end{aligned}$$

Για το δεύτερο

$$\begin{aligned} P(S^2 > 300) &= P\left(\frac{24S^2}{625} > \frac{24 \times 300}{625} = 11.52\right) \\ &= P(\chi_{24}^2 > 11.52). \end{aligned}$$



Από τους συνήθεις πίνακες της κατανομής χ^2 έχουμε

$$97.5\% < P(\chi_{24}^2 > 11.52) \leq 99\%$$

διότι $10.856 < 11.52 < 12.401$.

Θεώρημα 3.3: Έστω $x_1, \dots, x_n$ τυχαίο δείγμα μεγέθους n από ένα κανονικό πληθυσμό με μέση τιμή μ και διακύμανση σ^2 . Τότε η στατιστική συνάρτηση $(\bar{x} - \mu)\sqrt{n}/S$ έχει t -κατανομή με $n-1$ βαθμούς ελευθερίας. Συμβολικά

$$\frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t_{n-1} \quad (3.22)$$

Απόδειξη: Βάσει του Θεωρήματος 3.2 έχουμε:

$$\bar{x} \sim N(\mu, \sigma^2/n) \quad , \quad \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad ,$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

και οι μεταβλητές $\bar{x}$ και S^2 είναι στατιστικώς ανεξάρτητες. Άρα εφαρμόζοντας τον Ορισμό 3.3 παίρνουμε:

$$\frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}} / \sqrt{\frac{(n-1)S^2}{\sigma^2} / (n-1)} = \frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t_{n-1} \quad .$$



Πόρισμα 3.2: Εστω $(\bar{X}_1, S_1^2)$ και $(\bar{X}_2, S_2^2)$ οι μέσες τιμές και διακυμάνσεις δύο τυχαίων δειγμάτων μεγέθους n_1 και n_2 αντίστοιχα από δύο ανεξάρτητους κανονικούς πληθυσμούς με μέσες τιμές μ_1 και μ_2 αντίστοιχα και κοινή διακύμανση σ^2 . Τότε η στατιστική συνάρτηση

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \quad (3.23)$$

όπου
$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \quad (3.24)$$

δηλαδή έχει την t κατανομή με $n_1 + n_2 - 2$ βαθμούς ελευθερίας.

Θεώρημα 3.4: Εστω S_1^2 και S_2^2 οι διακυμάνσεις δύο τυχαίων δειγμάτων μεγέθους n_1 και n_2 από δύο ανεξάρτητους κανονικούς πληθυσμούς με μέσες τιμές και διακυμάνσεις (μ_1, σ_1^2) , (μ_2, σ_2^2) αντίστοιχα. Τότε η στατιστική συνάρτηση $S_2^2 S_1^2 / \sigma_1^2 S_2^2$ έχει F - κατανομή με n_1-1 και n_2-1 βαθμούς ελευθερίας, συμβολικά

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$$



Απόδειξη : Από το Θεώρημα 3.2 έχουμε:

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi_{n_1-1}^2 \quad \text{και} \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi_{n_2-1}^2 .$$

Εφαρμόζοντας τον Ορισμό 3.4 και την ανεξαρτησία των S_1^2 και S_2^2 έχουμε

$$\begin{aligned} & \left(\frac{(n_1-1)S_1^2}{\sigma_1^2} / (n_1-1) \right) / \left(\frac{(n_2-1)S_2^2}{\sigma_2^2} / (n_2-1) \right) \\ &= \frac{S_1^2 \sigma_2^2}{S_2^2 \sigma_1^2} \sim F_{n_1-1, n_2-1} . \end{aligned}$$

3.5 Κεντρικό Οριακό Θεώρημα

Όταν ο πληθυσμός από τον οποίο παίρνουμε το τυχαίο δείγμα δεν είναι κανονικός ποιές είναι οι δειγματικές κατανομές των διαφόρων στατιστικών; Όπως έχουμε αναφέρει η εύρεση των δειγματικών κατανομών είναι από τα σπουδαιότερα θέματα της Μαθηματικής Στατιστικής. Η Στατιστική Συμπερασματολογία στηρίζεται στις δειγματικές κατανομές. Η απάντηση στο παραπάνω ερώτημα απαιτεί μεγαλύτερο μαθηματικό υπόβαθρο και ξεφεύγει από το σκοπό του βιβλίου αυτού. Έχουμε όμως ένα σημαντικό θεώρημα το οποίο μας δίδει την κατά προσέγγιση κατανομή μεγάλων (με πολλούς όρους) αθροισμάτων ανεξάρτητων και ισόνομων τ.μ. και το οποίο

καταδεικνύει τη σπουδαιότητα της κανονικής κατανομής. Από το άθροισμα εύκολα προσδιορίζεται η κατανομή της μέσης τιμής. Το θεώρημα αυτό είναι το περίφημο Κεντρικό Οριακό Θεώρημα (ΚΟΘ) και εμφανίζεται σε δύο μορφές.

**Θεώρημα 3.5 : Κεντρικό Οριακό
Θεώρημα (Μορφή I)**

Εστω $X_1, \dots, X_n$ η ισόνομες και ανεξάρτητες τυχαίες μεταβλητές (τυχαίο δείγμα) η κάθε μία με πεπερασμένη μέση τιμή μ και διακύμανση σ^2 . Τότε αν το n είναι μεγάλο, η τυχαία μεταβλητή $\frac{\sum X_i - n\mu}{\sigma\sqrt{n}}$ είναι κατά προσέγγιση $N(0,1)$ ή ισοδύναμα, το άθροισμα $\sum X_i$ έχει κατά προσέγγιση κανονική κατανομή με μέση τιμή $n\mu$ και διακύμανση $n\sigma^2$.

**Θεώρημα 3.6 : Κεντρικό Οριακό
Θεώρημα (Μορφή II)**

Εστω $X_1, \dots, X_n$ ένα τυχαίο δείγμα από κάποιο πληθυσμό με πεπερασμένη μέση τιμή μ και διακύμανση σ^2 . Αν το n είναι μεγάλο, τότε η δειγματική μέση τιμή $\bar{X}$ έχει κατά προσέγγιση κανονική κατανομή με μέση τιμή μ και διακύμανση σ^2/n ή ισοδύναμα, η τυχαία μεταβλητή $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ είναι κατά προσέγγιση $N(0,1)$.

Το αξιοσημείωτο του Κεντρικού Οριακού Θεωρήματος είναι ότι και αν ακόμη η κατανομή του πληθυσμού δεν είναι κανονική, η μέση τιμή του δείγματος έχει κατά προσέγγιση κανονική κατανομή αν το n είναι μεγάλο. Αυτό μας επιτρέπει να χρησιμοποιούμε τη στατιστική θεωρία που στηρίζεται στην κανονική κατανομή αρκεί το δείγμα να είναι μεγάλο.



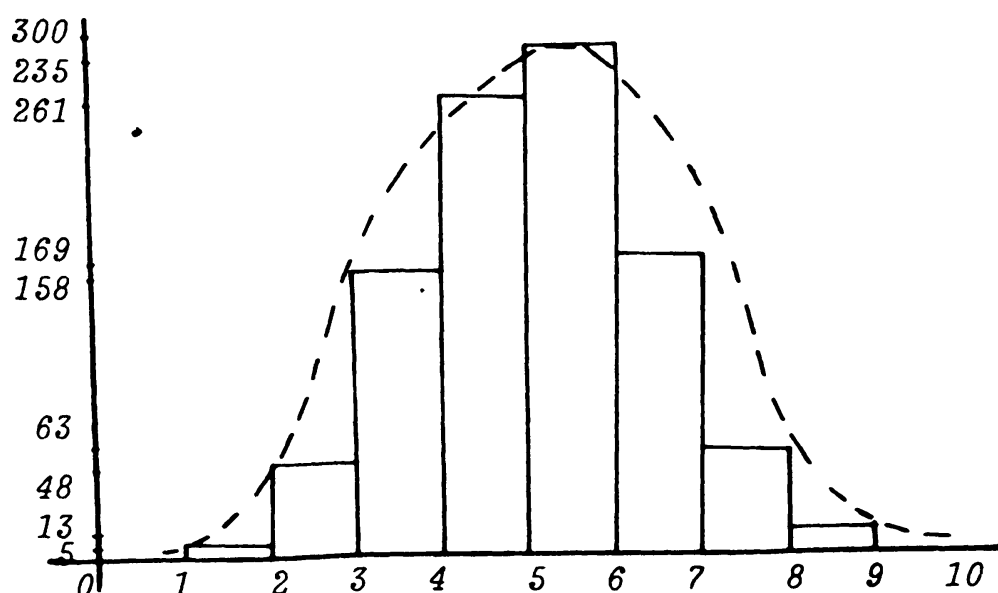
Το Κεντρικό Οριακό Θεώρημα (ΚΟΘ) συχνά ερμηνεύεται ότι συνεπάγεται σύγκλιση της κατανομής του $\bar{X}$ προς μία κανονική κατανομή καθώς το $n \rightarrow \infty$. Αυτό δεν είναι σωστό διότι $\text{Var}(\bar{X}) \rightarrow 0$ καθώς το $n \rightarrow \infty$. Το ΚΟΘ δικαιολογεί την προσέγγιση της κατανομής του $\bar{X}$ με την κατανομή $N(\mu, \sigma^2/n)$ όταν το n είναι μεγάλο.

Η απόδειξη του Κεντρικού Οριακού Θεωρήματος απαιτεί βαθύτερες έννοιες και προτάσεις της Μαθηματικής Στατιστικής και ξεφεύγει από το σκοπό του βιβλίου αυτού. Εμπειρικά το θεώρημα μπορεί να επιβεβαιωθεί (i) παίρνοντας ένα μεγάλο αριθμό τυχαίων δειγμάτων μεγέθους n από μία κατανομή με μέση τιμή μ και διακύμανση σ^2 , (ii) υπολογίζοντας για κάθε δείγμα τη μέση του τιμή $\bar{X}$ και (iii) χαράσσοντας το ιστόγραμμα των τιμών $\bar{X}$. Το ιστόγραμμα αυτό θα πρέπει να προσεγγίζει μία καμπύλη ανάλογη της συνάρτησης πυκνότητας πιθανότητας της κανονικής κατανομής με μέση τιμή μ και διακύμανση σ^2/n . Πράγματι, με τη βοήθεια του minicomputer Hewlett-Packard 45, 1002 τυχαία δείγματα μεγέθους 5 παίρνονται από την ομοιόμορφο κατανομή $U(0,10)$ με βάση το διάστημα $(0, 10)$ και με προσέγγιση δευτέρου δεκαδικού ψηφίου. Για κάθε δείγμα υπολογίζουμε το $\bar{X}$. Κατόπιν οι 1002 δειγματικές μέσες τιμές ομαδοποιούνται σε 10 διαστήματα $[0,1)$, $[1,2)$, ..., $[9,10)$ με τα ακόλουθα αποτελέσματα:

Διάστημα	Συχνότητα
$[0,1)$	0
$[1,2)$	5
$[2,3)$	48
$[3,4)$	158
$[4,5)$	261
$[5,6)$	285
$[6,7)$	169
$[7,8)$	63
$[8,9)$	13
$[9,10)$	0



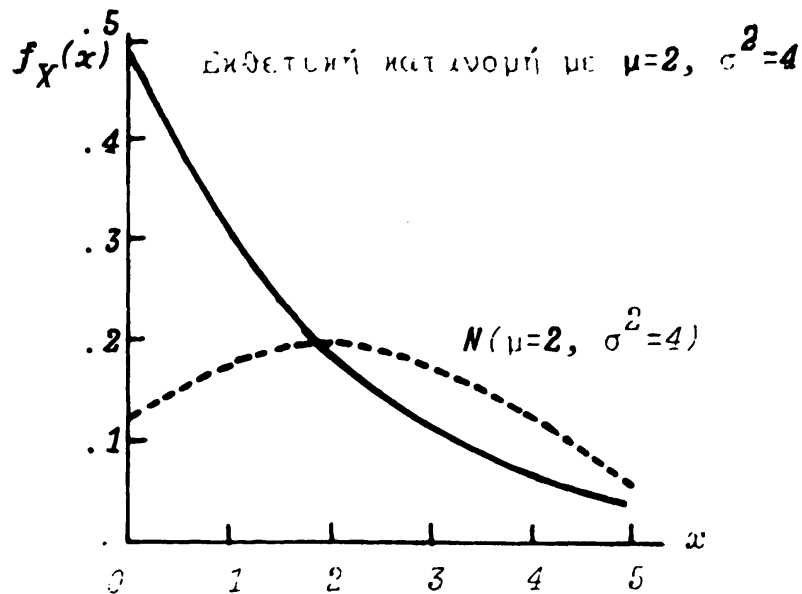
Το ιστόγραμμα των τιμών αυτών δίδεται στο Σχήμα 3.4 από το οποίο φαίνεται η προσέγγιση της κατανομής του $\bar{X}$ με μία κατανομή σχήματος καμπάνας.



Σχήμα 3.4 Κατανομή του $\bar{X}$ με δειγματοληψία από ομοιόμορφη κατανομή

Είναι φανερό ότι όσο μεγαλύτερο το n τόσο καλύτερη η προσέγγιση της κανονικής κατανομής από την κατανομή του $\bar{X}$. Το Σχήμα 3.5 επεξηγεί όχι μόνο το Κεντρικό Οριακό Θεώρημα αλλά και τη βελτίωση της προσέγγισης καθώς το n μεγαλώνει. Έτσι το Σχήμα 3.5 (α) δίνει την πυκνότητα (συνεχής γραμμή) της εκθετικής κατανομής με $\lambda = 1/2$ ($\mu=2$, $\sigma^2=4$) και την πυκνότητα (διακεκομμένη γραμμή) της κανονικής κατανομής με την ίδια μέση τιμή και διακύμανση.



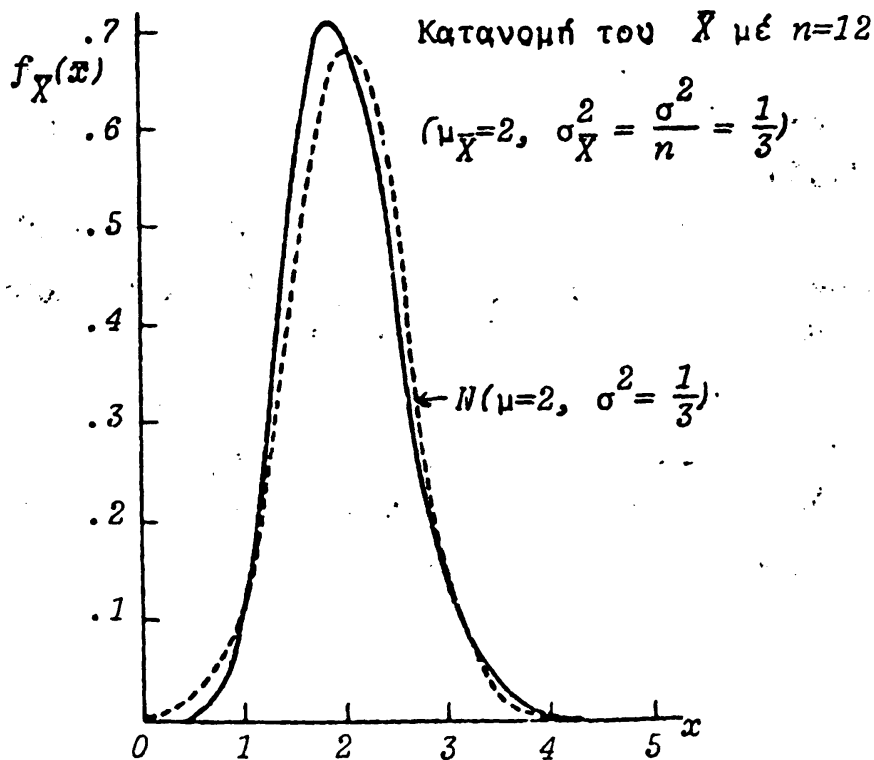
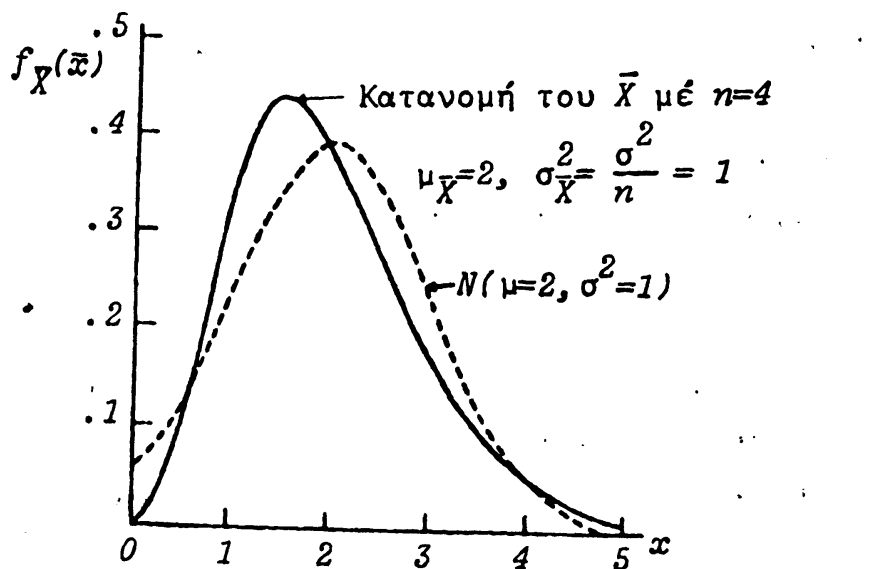


Σχήμα 3.5 (α) Κεντρικό Οριακό Θεώρημα με δειγματοληψία από εκθετική κατανομή

Η μέση τιμή $\bar{X}$ για δείγματα μεγέθους $n = 4$ από τον εκθετικό πληθυσμό του Σχήματος 3.5 (α) έχει μέση τιμή 2 και διακύμανση $\sigma^2/n = 1$. Το Σχήμα 3.5 (β) μας δείχνει την ακριβή κατανομή (συνεχής γραμμή) του $\bar{X}$ και την αντίστοιχη κανονική κατανομή με $\mu=2$ και $\sigma^2 = 1$. Οι δύο καμπύλες είναι αξιοσημείωτα όμοιες. Τα ίδια στοιχεία δίδονται στο Σχήμα 3.5 (γ) αλλά για $n=12$. Η προσέγγιση είναι αξιοθαύμαστη ακόμη και παρά το γεγονός ότι το n δεν είναι υπερβολικά μεγάλο.

Απομένει να καθοριστεί το μέγεθος του n ώστε η προσέγγιση του Κεντρικού Οριακού Θεωρήματος να είναι ικανοποιητική. Δυστυχώς δεν υπάρχει ξεκάθαρη και γενική απάντηση στο ερώτημα αυτό. Η κατάλληλη τιμή του n εξαρτάται από την κατανομή του πληθυσμού και τη χρήση που θα κάνουμε στα αποτελέσματα τα οποία θα πάρουμε. Κατά κανόνα το Κεντρικό Οριακό Θεώρημα δίδει ικανοποιητικά αποτελέσματα ακόμη και για

μικρά δείγματα ($n \geq 25$ ή 30), αρκεί η κατανομή του πληθυσμού να είναι μονοκόρυφη και να μην έχει ουρές με μεγάλη μάζα πιθανότητας.



Σχήμα 3.5 (β,γ) Κεντρικό Οριακό Θεώρημα με δειγματοληψία από εκθετική κατανομή



Οι εφαρμογές του Κεντρικού Οριακού Θεωρήματος είναι πολύπλευρες: Σε πολύπλοκες φυσικές μετρήσεις, το σφάλμα μετρήσεως είναι το άθροισμα $\sum X_i$ πολλών ανεξάρτητων τυχαίων σφαλμάτων X_i . Το ύψος ενός φυτού είναι το άθροισμα $\sum X_i$ πολλών μικρών ανεξάρτητων αυξήσεων X_i . Το ίδιο ισχύει για το ύψος, το βάρος κ.λ.π. ενός ατόμου, την επίδοση στο σχολείο κ.λ.π. Στις περιπτώσεις αυτές η κανονική κατανομή προσφέρει ικανοποιητική προσέγγιση της κατανομής του αθροίσματος.

Παράδειγμα 3.2: Εστω ότι σε ένα οποιοδήποτε μικροδευτερόλεπτο (10^{-6}sec) η πιθανότητα k ηλεκτρόνια να φτάσουν στην άνοδο μιας λυχνίας κενού είναι

$$(0.16)^k e^{-0.16}/k! .$$

Να βρεθεί η πιθανότητα ο αριθμός των ηλεκτρονίων που θα φθάσουν σε ένα δευτερόλεπτο να είναι μεταξύ 159000 και 161000.

Εστω $X_1, X_2, X_3, \dots$ ο αριθμός των ηλεκτρονίων που φτάνουν στην άνοδο το $1^0, 2^0, 3^0$, κ.λ.π. μικροδευτερόλεπτο. Προφανώς $X_i \sim P(\lambda=0.16)$, $i=1, 2, \dots$ και $\mu = EX_i = 0.16$, $\sigma^2 = \text{Var}(X_i) = 0.16$. Το δευτερόλεπτο έχει 10^6 μικροδευτερόλεπτα. Ο αριθμός των ηλεκτρονίων που φτάνουν στην άνοδο σε 1 δευτερόλεπτο είναι $X_1 + X_2 + \dots + X_{10^6}$, $n = 10^6$ και η ζητούμενη πιθανότητα είναι

$$P(159000 \leq X_1 + X_2 + \dots + X_{10^6} \leq 161000).$$

Θα υπολογίσουμε την πιθανότητα αυτή κατά προσέγγιση χρησιμοποιώντας το Κεντρικό Οριακό Θεώρημα και τη διόρθωση συνέχειας.

Θέτουμε $X = X_1 + X_2 + \dots + X_{10^6}$ και έχουμε



$$\mu = E(X_i) = \pi = 0.16, \sigma^2 = \text{Var}(X_i) = \pi = 0.16,$$

$$\mu_X = n\mu = 10^6 \times 0.16 = 160000,$$

$$n\sigma^2 = 10^6 \times 0.16 = 160000.$$

Έτσι

$$\begin{aligned} & P(159000 \leq X_1 + X_2 + \dots + X_{10^6} \leq 161000) \\ &= P(158999.5 < X < 161000.5 | X \sim N(\mu = 160000, \sigma^2 = 160000)) \\ &\approx P\left(\frac{158999.5 - 160000}{\sqrt{160000}} < \frac{X - \mu_X}{\sigma_X} < \frac{161000.5 - 160000}{\sqrt{160000}}\right) \\ &\approx P\left(\frac{-1000.5}{400} < Z < \frac{1000.5}{400}\right) \\ &\approx P(-2.5 < Z < 2.5) = 0.9876. \end{aligned}$$

3.6 Κανονική προσέγγιση της διωνυμικής κατανομής

Το Κεντρικό Οριακό Θεώρημα μπορεί να εφαρμοστεί στην περίπτωση μίας διωνυμικής τυχάας μεταβλητής X με μεγάλο αριθμό n δοκιμών και πιθανότητα επιτυχίας p . Η εφαρμογή αυτή ονομάζεται κανονική προσέγγιση της διωνυμικής κατανομής.

Έστω X_i το αποτέλεσμα της i δοκιμής του διωνυμικού

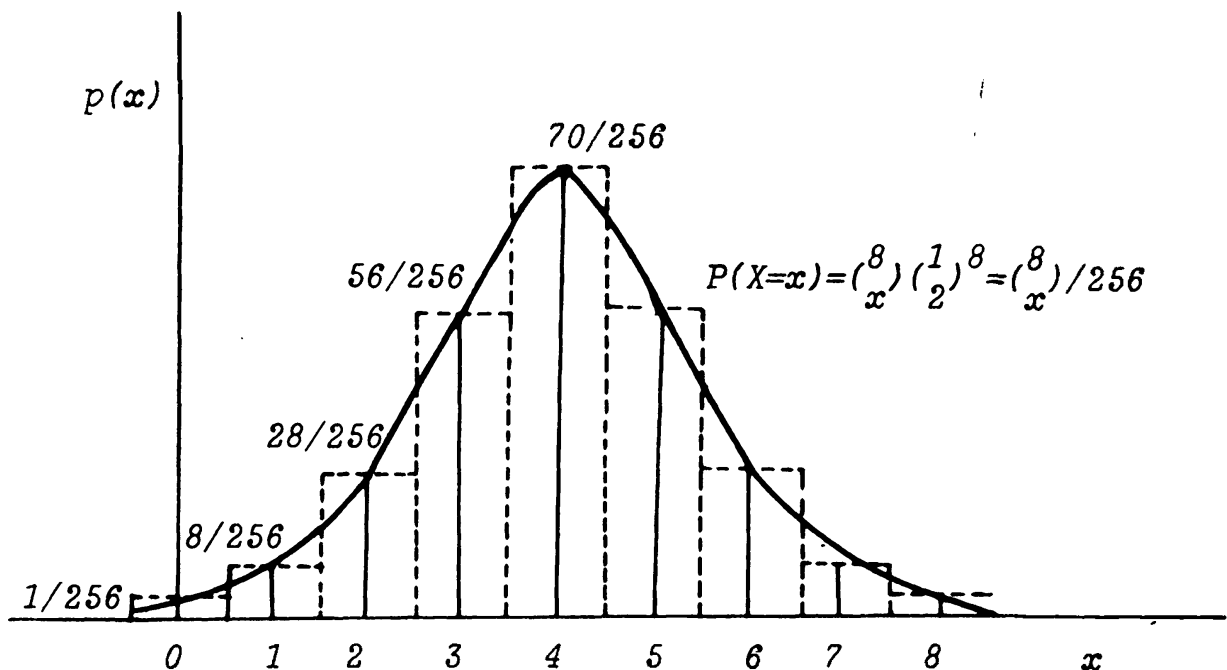


πειράματος, όπου $X_1=1$ αν το αποτέλεσμα είναι επιτυχία E και $X_1=0$ αν το αποτέλεσμα είναι αποτυχία A τότε έχουμε

$$X = X_1 + X_2 + \dots + X_n.$$

Οι τυχαίες μεταβλητές $X_1, \dots, X_n$ είναι ανεξάρτητες και ισόνομες με μέση τιμή p και διακύμανση pq . Έτσι για μεγάλο n η τυχαία μεταβλητή $X = \sum X_i$ έχει κατά προσέγγιση κανονική κατανομή με μέση τιμή np και διακύμανση npq .

Για παράδειγμα ας θεωρήσουμε την περίπτωση $X \sim B(n=8, p=0.5)$. Η συνάρτηση πιθανότητας του X δίδεται στο Σχήμα 3.6.



Σχήμα 3.6 Κανονική προσέγγιση διωνυμικής κατανομής

Αν πάρουμε τα ενδιάμεσα σημεία $-0.5, 0.5, 1.5, \dots, 7.5$ και 8.5 και κατασκευάσουμε ορθογώνια με ύψη τις πιθανότητες $1/256, \dots, 70/256, \dots, 1/256$, τότε τα εμβαδά των ορθογώνιων ισούνται με τις αντίστοιχες πιθανότητες $p(0), p(1), \dots, p(8)$. Έτσι

$$P(3 \leq X \leq 6) = \begin{cases} \text{άθροισμα των εμβαδών των ορθογωνίων με βάσεις} \\ (2.5, 3.5), (3.5, 4.5), (4.5, 5.5), (5.5, 6.5) \end{cases}$$

$$= \frac{56}{256} + \frac{70}{256} + \frac{56}{256} + \frac{28}{256} = \frac{210}{256} = 0.8200.$$

Εστω τώρα μία κανονική κατανομή με μέση τιμή και διακύμανση τις ίδιες όπως η διωνυμική, δηλαδή μέση τιμή

$$\mu = np = 8 \times (1/2) = 4 \quad \text{και} \quad \sigma^2 = npq = 8 \times (1/2) \times (1/2) = 2.$$

Χαράσουμε τη συνάρτηση πυκνότητας πιθανότητας της κανονικής πάνω στο ίδιο σχήμα και βλέπουμε ότι τα εμβαδά των ορθογωνίων προσεγγίζονται από τα αντίστοιχα εμβαδά κάτω από την κανονική συνάρτηση πυκνότητας. Έτσι

$$P(2.5 < X < 6.5 | X \sim N(\mu=4, \sigma^2=2)) = P\left(\frac{2.5-4}{1.414} < \frac{X-\mu}{\sigma} < \frac{6.5-4}{1.414}\right)$$

$$= P(-1.06 < Z < 1.77)$$

$$= 0.3554 + 0.4616$$

$$= 0.8170$$

και

$$P(3 \leq X \leq 6) \approx P(2.5 < X < 6.5 | X \sim N(\mu=np, \sigma^2=npq)) .$$

Η προσέγγισή μας έχει ακρίβεια περίπου 0.4% της αληθινής τιμής.

Η αύξηση του διαστήματος $[3, 6)$ κατά $1/2$ δεξιά και αριστερά λέγεται διόρθωση συνεχείας δεδομένου ότι διακριτές πιθανότητες υπολογίζονται σαν εμβαδά με την βοήθεια συναρτήσεων πυκνοτήτων πιθανοτήτων. Η διόρθωση συνεχείας έχει



σαν αποτέλεσμα την βελτίωση της προσέγγισης.

Γενικά, αν το διωνυμικό ενδεχόμενο είναι

$$a < X \leq b$$

τότε χρησιμοποιούμε το ενδεχόμενο

$$a - \frac{1}{2} < X < b + \frac{1}{2}$$

για την προσέγγιση με την κανονική κατανομή. Η προσέγγιση είναι καλύτερη όσο το n είναι μεγαλύτερο και όσο το p είναι πλησιέστερο στο $1/2$. Από εμπειρικούς υπολογισμούς προκύπτει ότι η προσέγγιση της διωνυμικής κατανομής από τη κανονική είναι ικανοποιητική όταν τα n και p ικανοποιούν τον ακόλουθο πρακτικό κανόνα

$$0 \leq np - 2\sqrt{npq} < np + 2\sqrt{npq} \leq n.$$

Παράδειγμα 3.3: Ένας μεγάλος αριθμός σπόρων μίας ποικιλίας λουλουδιών αναμειγνύεται με τις ακόλουθες αναλογίες ως προς το χρώμα του λουλουδιού που θα παραχθεί: 2 κόκκινα, 2 άσπρα, 1 μπλε. Οι σπόροι αναμειγνύονται και συσκευάζονται τυχαία σε σακκούλες των 100 σπόρων. Να βρεθεί η πιθανότητα μία σακούλα να περιέχει το πολύ 50 "άσπρους" σπόρους.

Έστω X ο αριθμός των "άσπρων" σπόρων μίας σακούλας. Προφανώς το διωνυμικό πείραμα μπορεί να υιοθετηθεί και αν X είναι ο αριθμός των άσπρων σπόρων μίας σακούλας τότε $X \sim B(n=100, p=40\%)$, όπου p = πιθανότητα "άσπρου" σπόρου στο μείγμα $= 2/(2+2+1) = 2/5$. Τότε η πιθανότητα που ζητάμε είναι



$$\sum_{x=0}^{50} \binom{100}{x} (0.4)^x (0.6)^{100-x}$$

Η πιθανότητα αυτή μπορεί να υπολογιστεί κατά προσέγγιση ως ακολούθως:

$$P(X \leq 50) \approx P(X < 50.5 | X \sim N(\mu = 100 \times 0.4, \sigma^2 = 100 \times 0.4 \times 0.6))$$

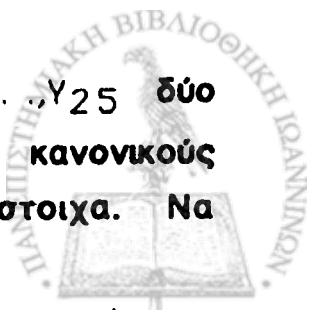
$$\approx P\left(\frac{X - \mu}{\sigma} < \frac{50.5 - 40}{\sqrt{24}} \mid X \sim N(\mu = 40, \sigma^2 = 24)\right)$$

$$\approx P(Z < 2.14) = 0.9842.$$



ΑΣΚΗΣΕΙΣ

- 3.1 Δίδεται ο πληθυσμός των μετρήσεων: 3, 4, 5, 6. (i) Να υπολογιστούν τα μ και σ^2 για τον πληθυσμό αυτό. (ii) Να καταγραφούν όλα τα δυνατά τυχαία δείγματα (με επανάθεση) μεγέθους $n = 2$. (iii) Για κάθε δείγμα να βρεθούν τα $\bar{X}$ και S^2 . (iv) Να βρεθούν τα $E(\bar{X})$ και $Var(\bar{X})$ και να συγκριθούν με τα αποτελέσματα που δίνουν γνωστές σχέσεις. (v) Να βρεθεί η $E(S^2)$ και να συγκριθεί με το σ^2 .
- 3.2 Αν $X_1, \dots, X_n$ είναι τυχαίο δείγμα από μια κανονική κατανομή $N(\mu, \sigma^2)$ να βρεθούν (i) η κατανομή του $X_i - \bar{X}$ για i σταθερό, (ii) η κατανομή του $n(\bar{X} - \mu)^2 / \sigma^2$, (iii) η μέση τιμή του S^2 , όπου $S^2 = [\sum (X_i - \bar{X})^2] / n$ και (iv) η κατανομή του $\sum (X_i - \mu)^2 / \sigma^2$.
- 3.3 Έστω S^2 η διακύμανση ενός τυχαίου δείγματος μεγέθους 10 από μια κανονική κατανομή $N(\mu = 0, \sigma^2 = 4)$. Να βρεθεί η $Var(S^2)$.
- 3.4 Να βρεθεί η κατανομή της τ.μ. $(\sum X_i^2) / \sigma^2$, όπου $X_1, \dots, X_n$ τυχαίο δείγμα από ένα πληθυσμό με κανονική κατανομή $N(0, \sigma^2)$.
- 3.5 Έστω $X_1, X_2, \dots, X_{25}$ και $Y_1, Y_2, \dots, Y_{25}$ δύο τυχαία δείγματα από δύο ανεξάρτητους κανονικούς πληθυσμούς $N(0, 16)$ και $N(1, 9)$ αντίστοιχα. Να



υπολογιστεί η πιθανότητα $P(\bar{X} - \bar{Y} > 0)$.

- 3.6 Μία βιομηχανία κατασκευάζει παιδικές τροφές των οποίων το βάρος ακολουθεί την κανονική κατανομή με μέση τιμή 900 gr. Σε ένα αγορανομικό έλεγχο ο ελεγκτής διαλέγει στην τύχη 40 κουπά από την παραγωγή μίας μέρας και τα ζυγίζει όλα μαζί. Ποιά είναι η πιθανότητα το συνολικό βάρος των 40 κουπών να είναι μικρότερο των 35.500 gr;

- 3.7 Έστω $X_i, i=1, \dots, 5$ ανεξάρτητες τυπικές κανονικές τυχαίες μεταβλητές.

(i) Να βρεθεί η μέση τιμή και η τυπική απόκλιση της $X_1^2 + X_2^2 + \dots + X_5^2$

(ii). Να βρεθεί η σταθερή C έτσι ώστε

$$P(-C \leq 2X_5 / \sqrt{X_1^2 + \dots + X_4^2} \leq C) = 0.90.$$

- 3.8 Αν X_1, X_2, X_3 είναι ανεξάρτητες τυχαίες μεταβλητές η κάθε μία με τυπική κανονική κατανομή

(i) να υπολογιστεί η $P(X_1^2 + X_2^2 + X_3^2 \leq 1)$ και

(ii) να δοθεί η κατανομή της $2X_1^2 / (X_2^2 + X_3^2)$.

- 3.9 Αν X_1, X_2 είναι τυχαίο δείγμα από μια τυπική κανονική κατανομή $N(0,1)$ να βρεθούν οι κατανομές των στατιστικών συναρτήσεων

(i) $(X_2 - X_1) / \sqrt{2},$ (ii) $(X_1 + X_2)^2 / (X_2 - X_1)^2,$



$$(111) (X_1 + X_2) / \sqrt{(X_2 - X_1)^2} \text{ και } X_2^2 / X_1^2 .$$

3.10 Έστω $X_1, \dots, X_n$ τυχαίο δείγμα από μια τυπική κανονική κατανομή $N(0, 1)$. Έστω επίσης

$$\bar{X}_k = \frac{1}{k} \sum_{i=1}^k X_i \text{ και } \bar{X}_{n-k} = \frac{1}{n-k} \sum_{i=k+1}^n X_i .$$

Να βρεθούν οι κατανομές των ακόλουθων στατιστικών συναρτήσεων (i) $(\bar{X}_k + \bar{X}_{n-k})/2$

$$(ii) k\bar{X}_k^2 + (n-k)\bar{X}_{n-k}^2, \quad (iii) X_1^2/X_2^2$$

3.11 Αν $\bar{X}$ είναι η μέση τιμή ενός τυχαίου δείγματος μεγέθους 100 από ένα πληθυσμό με κατανομή χ_{50}^2 να υπολογιστεί κατά προσέγγιση η πιθανότητα $P(49 < \bar{X} < 51)$.

3.12 Τα αποτελέσματα ενός μεγάλου αριθμού ανεξάρτητων μετρήσεων καταγράφονται. Κάθε ένα από τα αποτελέσματα στρογγυλεύεται στον πλησιέστερο ακέραιο. Αν τα σφάλματα που γίνονται λόγω στρογγύλευσης ακολουθούν την κατανομή

$$f(x) = \begin{cases} 2(1-2|x|) & , \quad -1/2 < x < 1/2 \\ 0 & , \quad \text{αλλού} \end{cases}$$



να βρεθεί η πιθανότητα το μέσο σφάλμα στρογγύλευσης να είναι απολύτως μικρότερο του $1/\sqrt{3n}$.

3.13 Έστω τυχαίο δείγμα μεγέθους 48 από την κατανομή

$$f(x) = \begin{cases} 1/2 & , \quad 0 < x < 2 \\ 0 & , \quad \text{αλλού} \end{cases}$$

Να υπολογιστεί κατά προσέγγιση η $P(0.95 < \bar{X} < 1.05)$.

3.14 Η μέση τιμή και η τυπική απόκλιση του χρόνου ζωής μιας λυχνίας είναι 1285 και 150 ώρες αντίστοιχα. Να βρεθεί η πιθανότητα η μέση τιμή ενός τυχαίου δείγματος 100 λυχνιών να είναι μεγαλύτερη από 1300 ώρες.

3.15 Ένα ρολοί κάνει σφάλμα στην ένδειξη της ώρας το πολύ μισό (± 0.5) λεπτό την ημέρα, αλλά γενικά (κατά μέσον όρο) πηγαίνει καλά. Έτσι για το μοντέλο "ακριβείας" του ρολογιού ορίζουμε με X_i το πραγματικό σφάλμα του ρολογιού (σε λεπτά της ώρας) στις 12 το μεσημέρι της i μέρας, $i=1, 2, \dots$ και υποθέτουμε ότι οι τ.μ. X_i , $i=1, 2, \dots$ είναι ανεξάρτητες ομοιόμορφες τ.μ. $U(-1/2, 1/2)$. Να βρεθεί κατά προσέγγιση η πιθανότητα μετά από 100 μέρες το ρολοί να δείχνει λάθος το πολύ ίσο με $3.46 = \sqrt{12}$ λεπτά στις 12 το μεσημέρι.



$$(111) (x_1 + x_2) / \sqrt{(x_2 - x_1)^2} \text{ και } x_2^2 / x_1^2 .$$

3.10 Έστω $x_1, \dots, x_n$ τυχαίο δείγμα από μια τυπική κανονική κατανομή $N(0,1)$. Έστω επίσης

$$\bar{x}_k = \frac{1}{k} \sum_{i=1}^k x_i \text{ και } \bar{x}_{n-k} = \frac{1}{n-k} \sum_{i=k+1}^n x_i .$$

Να βρεθούν οι κατανομές των ακόλουθων στατιστικών συναρτήσεων (i) $(\bar{x}_k + \bar{x}_{n-k})/2$

$$(11) k\bar{x}_k^2 + (n-k)\bar{x}_{n-k}^2, \quad (111) x_1^2/x_2^2$$

3.11 Αν $\bar{X}$ είναι η μέση τιμή ενός τυχαίου δείγματος μεγέθους 100 από ένα πληθυσμό με κατανομή χ_{50}^2 να υπολογιστεί κατά προσέγγιση η πιθανότητα $P(49 < \bar{X} < 51)$.

3.12 Τα αποτελέσματα ενός μεγάλου αριθμού ανεξάρτητων μετρήσεων καταγράφονται. Κάθε ένα από τα αποτελέσματα στρογγυλεύεται στον πλησιέστερο ακέραιο. Αν τα σφάλματα που γίνονται λόγω στρογγύλευσης ακολουθούν την κατανομή

$$f(x) = \begin{cases} 2(1-2|x|) & , \quad -1/2 < x < 1/2 \\ 0 & , \quad \text{αλλού} \end{cases}$$



να βρεθεί η πιθανότητα το μέσο σφάλμα στρογγύλευσης να είναι απολύτως μικρότερο του $1/\sqrt{3n}$.

3.13 Έστω τυχαίο δείγμα μεγέθους 48 από την κατανομή

$$f(x) = \begin{cases} 1/2 & , \quad 0 < x < 2 \\ 0 & , \quad \text{αλλού} \end{cases}$$

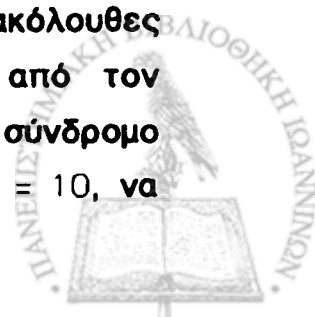
Να υπολογιστεί κατά προσέγγιση η $P(0.95 < \bar{X} < 1.05)$.

3.14 Η μέση τιμή και η τυπική απόκλιση του χρόνου ζωής μιας λυχνίας είναι 1285 και 150 ώρες αντίστοιχα. Να βρεθεί η πιθανότητα η μέση τιμή ενός τυχαίου δείγματος 100 λυχνιών να είναι μεγαλύτερη από 1300 ώρες.

3.15 Ένα ρολοί κάνει σφάλμα στην ένδειξη της ώρας το πολύ μισό (± 0.5) λεπτό την ημέρα, αλλά γενικά (κατά μέσον όρο) πηγαίνει καλά. Έτσι για το μοντέλο "ακριβείας" του ρολογιού ορίζουμε με X_i το πραγματικό σφάλμα του ρολογιού (σε λεπτά της ώρας) στις 12 το μεσημέρι της i μέρας, $i=1, 2, \dots$ και υποθέτουμε ότι οι τ.μ. X_i , $i=1, 2, \dots$ είναι ανεξάρτητες ομοιόμορφες τ.μ. $U(-1/2, 1/2)$. Να βρεθεί κατά προσέγγιση η πιθανότητα μετά από 100 μέρες το ρολοί να δείχνει λάθος το πολύ ίσο με $3.46 = \sqrt{12}$ λεπτά στις 12 το μεσημέρι.



- 3.16 Εργοστάσιο συσκευάζει τα τρανζίστορς που κατασκευάζει σε κιβώτια των 150. Τα βάρη των τρανζίστορς είναι τυχαίες μεταβλητές με μέση τιμή 0.8 gr. και διακύμανση 0.49 gr. Να βρεθεί η πιθανότητα το καθαρό βάρος 30 κιβωτίων να είναι μεγαλύτερο από 3.7 Kgr.
- 3.17 Έστω ότι ο αριθμός των αεροπλάνων που φτάνουν σε κάποιο αεροδρόμιο σε οποιαδήποτε εικοσάλεπτο περίοδο ακολουθεί την κατανομή του Poisson με μέση τιμή 100. (i) Να βρεθεί ένα κάτω φράγμα της πιθανότητας ο αριθμός των αεροπλάνων που φτάνουν σε μια δεδομένη εικοσάλεπτο περίοδο να είναι μεταξύ 80 και 120 χρησιμοποιώντας την ανισότητα του Chebyshev. (ii) Να υπολογιστεί η πιθανότητα στο (i) χρησιμοποιώντας το Κεντρικό Οριακό Θεώρημα.
- 3.18 Έστω X_i ο αριθμός των μετεωροειδών που συγκρούονται με ένα δοκιμαστικό δορυφόρο κατά τη διάρκεια της i τροχιάς και S_n ο συνολικός αριθμός των μετεωροειδών που συγκρούονται με τον δορυφόρο κατά τη διάρκεια n τροχιών. Αν οι τυχαίες μεταβλητές $X_1, X_2, \dots$ είναι ανεξάρτητες, η κάθε μια Poisson με παράμετρο λ (i) να βρεθούν η $E(S_n)$ και η $Var(S_n)$ και (ii) αν $n = 100$ και $\lambda = 4$ να βρεθεί κατά προσέγγιση η $P(S_{100} > 440)$.
- 3.19 Η πιθανότητα θεραπείας ενός ατόμου με ένα σπάνιο σύνδρομο είναι 1%. Σε κάθε μια από τις ακόλουθες περιπτώσεις εκλέγεται ένα τυχαίο δείγμα από τον πληθυσμό των ατόμων που πάσχουν από το σύνδρομο αυτό. Αν το μέγεθος του δείγματος είναι (i) $n = 10$, να



βρεθεί η πιθανότητα θεραπείας ενός ακριβώς ατόμου, (ii) $n = 300$, να βρεθεί η πιθανότητα θεραπείας τριών το πολύ ατόμων χρησιμοποιώντας την προσέγγιση του Poisson και (iii) $n = 600$, να βρεθεί η πιθανότητα θεραπείας 4 μέχρι και 12 ατόμων χρησιμοποιώντας την κανονική προσέγγιση.

3.20 Έστω

$$f(x) = \begin{cases} 1/x^2 & , \quad x > 1 \\ 0 & , \quad \text{αλλού} \end{cases}$$

η σ.π.π. μιας τ.μ. X και ένα τυχαίο δείγμα μεγέθους 72 από την κατανομή αυτή. Να υπολογιστεί κατά προσέγγιση η πιθανότητα περισσότερες από 50 από τις τιμές του τυχαίου δείγματος να είναι μικρότερες του 3.

3.21 Η πιθανότητα να γιαντρευτεί ένα άτομο το οποίο πάσχει από μια ασθένεια και υποβάλλεται σε μια πειραματική θεραπεία είναι 5%. (i) 10 τέτοια άτομα εκλέγονται τυχαία και υποβάλλονται στην θεραπεία. Ποια η πιθανότητα να γιαντρευθούν 3 από αυτά; (ii) Ποια η πιθανότητα να χρειαστούν 15 τέτοια άτομα μέχρις ότου 2 από αυτά γιαντρευτούν. (iii) 600 άτομα με αυτή την ασθένεια εκλέγονται στην τύχη. Ποια η πιθανότητα μετά την θεραπεία να γιαντρευτούν 25 μέχρι 35 άτομα;



- 3.22 Η Ολυμπιακή Αεροπορία (ΟΑ) γνωρίζει ότι 5% των ατόμων που κλείνουν θέση σε μια πτήση δεν εμφανίζονται τελικά για να ταξιδέψουν. Αν η ΟΑ κάνει κρατήσεις για 160 άτομα σε μια πτήση με 150 μόνο θέσεις να βρεθεί η πιθανότητα να υπάρχει θέση διαθέσιμη για όλα τα άτομα που έχουν κλείσει.



ΚΕΦΑΛΑΙΟ 4

ΣΤΑΤΙΣΤΙΚΗ

ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ

Συμπερασματολογία είναι ο κλάδος της Λογικής, ο οποίος ασχολείται με την εξαγωγή συμπερασμάτων. Τα συμπεράσματα μπορεί να έχουν τη μορφή απόφασης, πρόβλεψης, ή απόκτησης καινούργιας γνώσης, διαμόρφωσης γνώμης κ.λ.π. Καθημερινά αντιμετωπίζουμε καταστάσεις που απαιτούν λήψη προσωπικής απόφασης ή πρόβλεψης για το μέλλον. Η κυβέρνηση ενδιαφέρεται να προβλέψει τον τμάρημο το επόμενο έτος. Ο ερευνητής θέλει να αποδείξει πειραματικά τη θεωρία του. Ο ιατρός ενδιαφέρεται να συγκρίνει δύο διαφορετικές θεραπείες. Η νοικοκυρά θέλει να ξέρει αν το απορρυπαντικό Α είναι πιο αποτελεσματικό από το απορρυπαντικό Β. Για όλα αυτά εξάγονται συμπεράσματα με τη βοήθεια όλων των σχετικών πληροφοριών που λέγονται παρατηρήσεις ή δεδομένα.

Ο Στατιστικός συλλέγει τα δεδομένα με πειράματα ή δειγματοληψίες και επιδιώκει να βγάλει συμπεράσματα για τα φαινόμενα που εξετάζει. Χρησιμοποιεί τις μετρήσεις του για εκτίμηση των τιμών των αγνώστων παραμέτρων ή για έλεγχο (τεστ) υποθέσεων περί αυτών. Έτσι η Στατιστική Συμπερασματολογία χωρίζεται σε δύο κλάδους: την Εκτιμητική και τον Έλεγχο των Στατιστικών Υποθέσεων.

Στην § 3.1 παρουσιάσαμε μία σφυγμομέτρηση της κοινής γνώμης κατά την οποία πήραμε ένα δείγμα από τους ψηφοφόρους μίας πόλης οι οποίοι θεωρούνται διχασμένοι σε εκείνους που υποστηρίζουν τον υποψήφιο Α και εκείνους που υποστηρίζουν τον υποψήφιο Β. Υποθέσαμε ότι p είναι το άγνωστο ποσοστό ψηφοφόρων



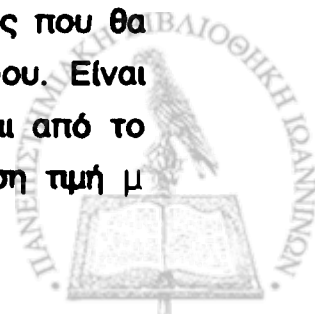
που υποστηρίζουν τον υποψήφιο Α. Σκοποί της σφυγμομέτρησης μπορεί να είναι η εκτίμηση ή πρόβλεψη του p ή ο έλεγχος κάποιας υπόθεσης για το p , π.χ. να ελέγξουμε ότι $p > 1/2$, δηλαδή να προβλέψουμε, βάσει του δείγματος, αν ο Α θα κερδίσει τις εκλογές ή όχι. Και στις δύο περιπτώσεις βγάζουμε στατιστικά συμπεράσματα για το p .

Η στατιστική συμπερασματολογία έχει επαγωγικό χαρακτήρα, δηλαδή γενικεύει από ένα συγκεκριμένο πείραμα σε όλα τα παρόμοια πειράματα. Με άλλα λόγια γενικεύει από το δείγμα στον πληθυσμό. Εξ αιτίας του επαγωγικού της χαρακτήρα, κάθε στατιστική συμπερασματολογία περιέχει αβεβαιότητα. Η αβεβαιότητα αυτή μπορεί να μετρηθεί με τη βοήθεια της έννοιας της πιθανότητας (βλ. Κεφάλαιο 1). Σκοπός της Στατιστικής είναι η μελέτη μεθόδων συμπερασματολογίας και τρόπων μέτρησης της αβεβαιότητας αυτής. Έτσι σε κάθε πρακτική εφαρμογή η στατιστική συμπερασματολογία περιλαμβάνει δύο στοιχεία: (α) το συμπέρασμα και (β) το μέτρο της ορθότητας ή καταλληλότητας του.

Στο Κεφάλαιο αυτό αναπτύσσουμε πρώτα τις αρχές και μεθόδους εκτίμησης στατιστικών παραμέτρων και στη συνέχεια δίνουμε τη φιλοσοφία και μεθοδολογία του ελέγχου των στατιστικών υποθέσεων. Στόχος μας είναι να δώσουμε μία απλή και σύντομη εισαγωγή στις ιδέες της Στατιστικής Συμπερασματολογίας. Για πλήρη ανάπτυξη των θεμάτων αυτών ο αναγνώστης παραπέμπεται σε προχωρημένα συγγράμματα της Στατιστικής.

4.1 Εκτίμηση Παραμέτρων

Έστω ότι έχουμε στα χέρια μας ένα τυχαίο δείγμα $X_1, X_2, \dots, X_n$ από κάποιο πληθυσμό με μία άγνωστη παράμετρο. Το πρόβλημα μας είναι να βρούμε μία ή περισσότερες ποσότητες που θα χρησιμοποιηθούν για την εκτίμηση της άγνωστης παραμέτρου. Είναι προφανές ότι οι ποσότητες αυτές θα πρέπει να προέρχονται από το δείγμα. Για παράδειγμα, όταν θέλουμε να εκτιμήσουμε τη μέση τιμή μ



ενός πληθυσμού X , μπορούμε να χρησιμοποιήσουμε την ποσότητα $\bar{X}$. Όταν χρησιμοποιούμε μια τέτοια ποσότητα, η ποσότητα αυτή λέγεται **εκτιμήτρια συνάρτηση** και η αριθμητική τιμή της που προκύπτει από το δείγμα **εκτιμητής** της παραμέτρου. Οι εκτιμήτριες συναρτήσεις είναι στατιστικές συναρτήσεις και εξαρτώνται μόνο από το δείγμα και όχι την παράμετρο που ζητάμε να εκτιμήσουμε. Στην περίπτωση αυτή εκτιμούμε ένα άγνωστο σημείο με ένα γνωστό σημείο. Η σχετική μεθοδολογία λέγεται **σημειοεκτιμητική**.

Την άγνωστη παράμετρο μπορούμε να εκτιμήσουμε με ένα διάστημα αντί με ένα σημείο. Χρειαζόμαστε δηλαδή κατά κανόνα δύο ποσότητες που προέρχονται από το δείγμα, τα όρια του διαστήματος. Η σχετική μεθοδολογία λέγεται **εκτιμητική με διαστήματα εμπιστοσύνης**.

Σε πολλά προβλήματα της Στατιστικής η κατανομή του πληθυσμού είναι γνωστής μορφής εκτός από το γεγονός ότι περιέχει μία ή περισσότερες άγνωστες παραμέτρους. Η περίπτωση αυτή ονομάζεται (παραμετρικό) **στατιστικό μοντέλο**. Για παράδειγμα η κατανομή του πληθυσμού μπορεί να είναι διωνυμική $B(n, p)$ με n γνωστό και $p \in (0, 1)$ άγνωστο ή κανονική $N(\mu, \sigma^2)$ με $-\infty < \mu < +\infty$, $\sigma^2 > 0$ άγνωστα κ.λ.π. Γενικά οι άγνωστοι παράμετροι θα συμβολίζονται με θ ή θ_1, θ_2 , κ.λ.π. και οι κατανομές του πληθυσμού X με $p(x, \theta)$ ή $f(x_i, \theta)$. Το πρόβλημα μας είναι να εκτιμήσουμε το θ ή συναρτήσεις $g(\theta)$ του θ . Οι εκτιμήτριες συναρτήσεις του θ θα συμβολίζονται με $\hat{\theta}$ ή γενικά $T(X_1, \dots, X_n)$.

4.1.1 Σημειοεκτιμητική

Το πρόβλημα της σημειοεκτιμητικής είναι διπλό: Ποιά είναι τα κριτήρια ή αρχές με τις οποίες αξιολογούμε τους εκτιμητές και ορίζουμε τον "καλύτερο" εκτιμητή μεταξύ των διαφόρων υποψηφίων και ποιές είναι οι μέθοδοι εύρεσεως εκτιμητών;

Τα κριτήρια αξιολογήσεως εκτιμητών είναι πολλά. Από



αυτά αναφέρουμε δύο βασικά:

Αμεροληψία: Ένας εκτιμητής $\hat{\theta}$ μίας παραμέτρου θ λέγεται αμερόληπτος αν έχει αναμενόμενη τιμή θ για κάθε τιμή της παραμέτρου θ , δηλαδή

$$E(\hat{\theta}) = \theta \quad (4.1)$$

για κάθε θ .

Η αμεροληψία ενός εκτιμητή μετρά την ορθότητα (accuracy) της εκτίμησης και εκφράζει την ιδέα ότι σε πολλές νοερές ή πραγματικές επαναλήψεις της δειγματοληψίας μας η μέθοδος εκτίμησης που χρησιμοποιούμε θα μας δώσει κατά μέσο όρο την άγνωστη τιμή του θ .

Είναι εύκολο να δει κανείς ότι υπάρχουν πολλοί αμερόληπτοι εκτιμητές για κάθε παράμετρο θ , π.χ. αν $n=50$ οι εκτιμητές

$$\bar{x}_{50} = \frac{1}{50} \sum_{i=1}^{50} x_i \text{ και } \bar{x}_{25} = \frac{1}{25} \sum_{i=1}^{25} x_i$$

είναι και οι δύο αμερόληπτοι εκτιμητές του μ . Διαισθανόμαστε ότι ο $\bar{x}_{50}$ είναι καλύτερος από τον $\bar{x}_{25}$ διότι προκύπτει από περισσότερες μετρήσεις, δηλ. χρησιμοποιεί περισσότερη πληροφορία από τον $\bar{x}_{25}$. Εκείνο που παίζει ρόλο είναι η μεταβλητότητα του εκτιμητή. Η μεταβλητότητα του εκτιμητή μετρείται με τη διακύμανση του. Η διακύμανση του $\bar{x}_{50}$ είναι $\sigma^2/50$ και η διακύμανση του $\bar{x}_{25}$ είναι $\sigma^2/25$, μεγαλύτερη από την πρώτη. Όσο μικρότερη η διακύμανση τόσο καλύτερη. Αυτές οι ιδέες μας οδηγούν στο δεύτερο κριτήριο.

Ελάχιστη Διακύμανση: Μεταξύ των αμερόληπτων εκτιμητών προτιμητέος είναι εκείνος που έχει τη μικρότερη διακύμανση.



Η ακρίβεια (precision) ενός αμερόληπτου εκτιμητή μετρείται συνήθως με την διακύμανση του

$$\sigma_{\hat{\theta}}^2 = \text{Var}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \quad (4.2)$$

ή την τυπική απόκλιση του

$$\sigma_{\hat{\theta}} = \sqrt{\text{Var}(\hat{\theta})} = \sqrt{E(\hat{\theta} - \theta)^2}.$$

Η τυπική απόκλιση $\sigma_{\hat{\theta}}$ λέγεται μερικές φορές και τυπικό σφάλμα του εκτιμητή. Για το δειγματικό μέσο $\bar{X}$ το τυπικό σφάλμα είναι $\sigma/\sqrt{n}$, δηλαδή

$$\sigma_{\bar{X}} = \sigma/\sqrt{n}. \quad (4.3)$$

Ένας εκτιμητής λέγεται αμερόληπτος ομοιόμορφα ελάχιστης διακύμανσης (Α.Ο.Ε.Δ.) όταν είναι αμερόληπτος και έχει μικρότερη διακύμανση από κάθε άλλο αμερόληπτο εκτιμητή.

Έχουμε δει ότι αν η άγνωστη παράμετρος θ είναι η μέση τιμή του πληθυσμού μ , το $\bar{X}$ έχει την ιδιότητα της αμεροληψίας δηλαδή

$$E(\bar{X}) = \mu. \quad (4.4)$$

Αν επί πλέον ο πληθυσμός είναι κανονικός, το $\bar{X}$ είναι Α.Ο.Ε.Δ. του μ .

Ομοίως αν $X_1, \dots, X_n$ είναι ένα τυχαίο δείγμα από κάποιο πληθυσμό με άγνωστη παράμετρο σ^2 , το στατιστικό S^2 είναι αμερόληπτος εκτιμητής του σ^2 δηλαδή

$$E(S^2) = \sigma^2. \quad (4.5)$$



Το σημαντικό σημείο των σχέσεων (4.4) και (4.5) είναι ότι ισχύουν πάντα ασχέτως των ιδιαιτέρων τιμών που τα μ και σ^2 ενδέχεται να έχουν. Η σχέση (4.5) δικαιολογεί τη χρήση του S^2 και όχι του $S \cdot^2$ για τη διακύμανση του δείγματος. Για να γίνει η διακύμανση του δείγματος αμερόληπτη εκτιμήτρια του σ^2 διαιρέσαμε το άθροισμα των τετραγώνων των αποκλίσεων (2.15) με το $n-1$ και όχι με το n . Η ποσότης $S \cdot^2$ τείνει να υποεκτιμά το σ^2 . Η τυπική απόκλιση του δείγματος

$$S = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

συνήθως χρησιμοποιείται ως εκτιμητής του σ παρά το γεγονός ότι δεν είναι ακριβώς αμερόληπτος.

Το τυπικό σφάλμα του $\bar{x}$ είναι

$$\sigma_{\bar{x}} = \sigma / \sqrt{n}$$

και εκτιμάται με

$$\hat{\sigma}_{\bar{x}} = S / \sqrt{n}.$$

Εκτός από τα παραπάνω κριτήρια αξιολόγησης εκτιμητών υπάρχουν και άλλα όπως η επάρκεια, η συνέπεια κ.λ.π. Υπάρχουν επίσης διάφοροι μέθοδοι ευρέσεως εκτιμητών. Η πλέον σημαντική από αυτές είναι η μέθοδος της μεγίστης πιθανοφάνειας. Τα θέματα αυτά ξεφεύγουν από το σκοπό του βιβλίου αυτού και δεν θα αναπτυχθούν. Αποτελούν αντικείμενο της Εκτιμητικής η οποία για κάθε στατιστικό μοντέλο και περίπτωση μας δίνει τον καλύτερο εκτιμητή ή υποδεικνύει τρόπους εύρεσης αυτού.

4.1.2 Διαστήματα εμπιστοσύνης

Έστω $x_1, x_2, \dots, x_n$ ένα τυχαίο δείγμα από κάποιο



πληθυσμό με μία άγνωστη παράμετρο θ . Έστω $L=L(X_1, X_2, \dots, X_n)$ και $U=U(X_1, X_2, \dots, X_n)$ τα κάτω και άνω όρια ενός διαστήματος αντίστοιχα. Τα όρια αυτά είναι συναρτήσεις των τυχαίων μεταβλητών του δείγματος. Οι τιμές τους είναι $l=l(x_1, x_2, \dots, x_n)$ και $u=u(x_1, x_2, \dots, x_n)$. Μπορούμε να χρησιμοποιήσουμε το διάστημα (L, U) με τιμή (l, u) για την εκτίμηση του θ . Δύο πράγματα πρέπει να ζητάμε: (i) το διάστημα (L, U) να περιέχει την αληθινή τιμή του θ ένα μεγάλο "ποσοστό φορών" και (ii) το διάστημα να έχει όσο το δυνατόν μικρότερο μήκος. Το "ποσοστό φορών" που ένα διάστημα (L, U) περιέχει το θ λέγεται **επίπεδο ή βαθμός εμπιστοσύνης**, συμβολίζεται με $100(1-\alpha)\%$ και φανερώνει την πιθανότητα το (L, U) να περιέχει το θ , δηλαδή

$$\text{βαθμός εμπιστοσύνης} = P(L < \theta < U) = 1 - \alpha. \quad (4.6)$$

Το διάστημα (L, U) λέγεται **διάστημα εμπιστοσύνης (Δ.Ε)**.

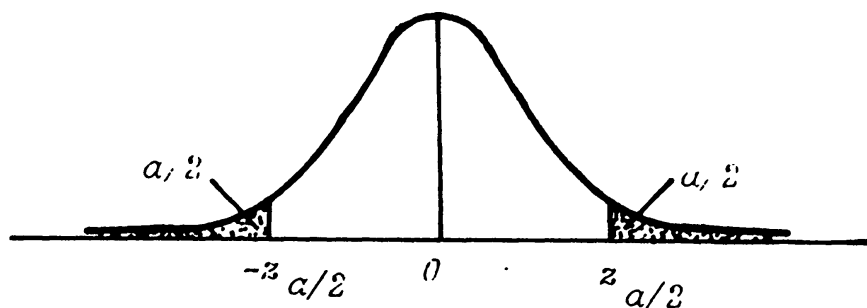
Για να διευκρινήσουμε τις παραπάνω έννοιες θα κατασκευάσουμε ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για τη μέση τιμή μ ενός κανονικού πληθυσμού με βάση ένα τυχαίο δείγμα μεγέθους n και με την υπόθεση ότι η διακύμανση σ^2 του πληθυσμού είναι γνωστή. Η τελευταία αυτή υπόθεση είναι εξωπραγματική. Διευκολύνει όμως στην αρχή την κατανόηση της βασικής διαδικασίας. Διάστημα εμπιστοσύνης στην περίπτωση που το σ^2 είναι άγνωστο δίδεται αργότερα.

Έστω $\bar{X}$ η μέση τιμή του δείγματος. Γνωρίζουμε ότι το $\bar{X}$ έχει κανονική κατανομή $N(\mu, \sigma^2/n)$. Έτσι για κάθε μ

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1). \quad (4.7)$$

Έστω $z_{\alpha/2}$ το αντίστροφο εκατοστιαίο σημείο της τυπικής κανονικής κατανομής που ορίστηκε στη § 3.3, σχέση (3.18). Τότε έχουμε (βλ. Σχήμα 4.1)





Σχήμα 4.1 Τυπική κανονική κατανομή

$$P\{-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\} = 1 - \alpha \quad (4.8)$$

ή

$$P\{-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\} = 1 - \alpha$$

ή

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \quad (4.9)$$

διότι τα ενδεχόμενα στις δύο τελευταίες σχέσεις είναι ισοδύναμα με το ενδεχόμενο $|\bar{X} - \mu| < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. Έτσι το διάστημα (L,U) με όρια

$$L = \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \text{ και } U = \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (4.10)$$

έχει την ιδιότητα



$$P\{L < \mu < U\} = 1 - \alpha \quad (4.11)$$

για κάθε μ . Το διάστημα αυτό είναι διάστημα εμπιστοσύνης (με όρια εμπιστοσύνης τα L και U) για την παράμετρο μ με βαθμό εμπιστοσύνης $100(1-\alpha)\%$.

Παρατηρούμε ότι τα όρια (4.10) είναι τυχαίες μεταβλητές και το διάστημα (L, U) , με την ιδιότητα (4.11) για κάθε μ , είναι τυχαίο διάστημα. Παρατηρούμε επίσης ότι τα σημεία $-z_{\alpha/2}$ και $z_{\alpha/2}$ της (4.8) δεν είναι τα μόνα που ικανοποιούν τη σχέση (4.8). Υπάρχουν άπειρα σημεία z_1, z_2 τέτοια ώστε

$$P\{z_1 < Z < z_2\} = 1 - \alpha.$$

Άρα τα Δ.Ε. σταθερού βαθμού εμπιστοσύνης δεν είναι μονοσήμαντα. Με την απαίτηση του ελάχιστου μήκους γίνονται μονοσήμαντα όπως στην παραπάνω περίπτωση με $z_1 = -z_{\alpha/2}$ και $z_2 = z_{\alpha/2}$.

Παράδειγμα 4.1: Ένα διεγερτικό φάρμακο ελέγχεται για την επίδραση του στην πίεση του αίματος. Οι πιέσεις αίματος είκοσι ατόμων μετρούνται πριν από τη λήψη και μισή ώρα μετά τη λήψη του φαρμάκου και λαμβάνονται οι ακόλουθες διαφορές:

7,	6,	0,	8,	-9,	-4,	0,	1,	-9,	1
2,	7,	0,	6,	-6,	-5,	-1,	6,	-2,	4.

Από προηγούμενες μελέτες είναι γνωστό ότι η προ και μετά τη λήψη του φαρμάκου διαφορά πιέσεων ακολουθεί κανονική κατανομή με $\sigma^2 = 25$. Να κατασκευασθεί ένα 95% διάστημα εμπιστοσύνης της μέσης διαφοράς μ πιέσεων αίματος.

Πράγματι έχουμε: $\bar{X} = 0.6, z_{0.025} = 1.96, \sigma = 5, n = 20$.

Έτσι τα όρια εμπιστοσύνης είναι



$$L = \bar{X} - 1.96 \frac{5}{\sqrt{20}} = 0.6 - 2.19 = -1.59$$

$$U = \bar{X} + 1.96 \frac{5}{\sqrt{20}} = 0.6 + 2.19 = 2.79.$$

Άρα ένα διάστημα εμπιστοσύνης της μέσης διαφοράς πιέσεων πριν και μετά το φάρμακο με βαθμό εμπιστοσύνης 95% είναι το διάστημα $(-1.59, 2.79)$.

Ποτέ δεν γράφουμε

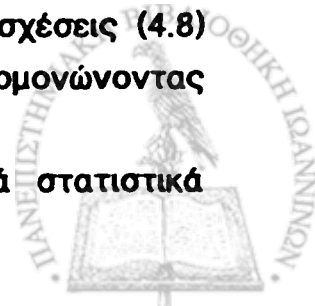
$$P(-1.59 < \mu < 2.79) = 0.95$$

διότι μετά το πείραμα η αληθινή τιμή του μ είτε ευρίσκεται εντός του διαστήματος $(-1.59, 2.79)$ είτε εκτός και η παραπάνω πιθανότητα είτε είναι 1 είτε 0. Λέμε όμως ότι έχουμε 95% εμπιστοσύνη ότι $-1.59 < \mu < 2.79$. Αυτό είναι ένας τρόπος έκφρασης της σχέσης (4.11).

Η εμπιστοσύνη που έχουμε στα όρια -1.59 και 2.79 πηγάζει στην πραγματικότητα από την εμπιστοσύνη μας στη στατιστική διαδικασία που έδωσε τα όρια αυτά. Αυτή η διαδικασία έδωσε τυχαίες μεταβλητές L και U που έχουν 95% πιθανότητα σε επαναληπτική διεγματοληψία να περικλείουν το αληθινό μ . Αν οι ειδικές τιμές τους -1.59 και 2.79 στη πραγματικότητα περικλείουν το αληθινό μ δεν έχουμε τρόπο να το γνωρίζουμε.

Μιλήσαμε για την έννοια του διαστήματος εμπιστοσύνης και δώσαμε ένα τέτοιο διάστημα για το απλούστερο από τα στατιστικά μοντέλα, την κανονική κατανομή με μ άγνωστο και σ^2 γνωστό. Από τον τρόπο εύρεσης του διαστήματος αυτού διαφαίνεται η ακόλουθη γενική ιδέα εύρεσης διαστημάτων εμπιστοσύνης: αρκεί να εύρουμε μία τυχαία ποσότητα όπως η (4.7) με γνωστή κατανομή το διάστημα εμπιστοσύνης ευρίσκεται τότε ανάλογα με τις σχέσεις (4.8) και (4.9) δηλαδή αντιστρέφοντας τις ανισότητες και απομονώνοντας την άγνωστη παράμετρο.

Πράγματι η ιδέα αυτή αποδίδει για πολλά στατιστικά



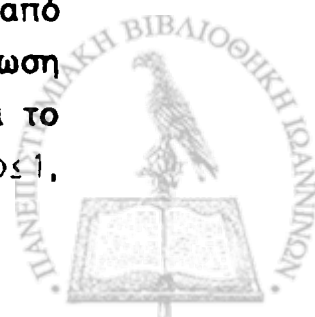
μοντέλα. Οι λεπτομέρειες του θέματος αυτού μαζί με τη γενική θεωρία των διαστημάτων εμπιστοσύνης και τον τρόπο ευρέσεως διαστημάτων ελαχίστου μήκους αποτελούν αντικείμενο παραπέρα μελέτης σε προχωρημένα βιβλία της Στατιστικής.

4.2 Έλεγχος Στατιστικών Υποθέσεων Στατιστικά Τεστ

Θα ξεκινήσουμε με ένα παράδειγμα που περιέχει όλα τα στοιχεία του θέματος.

Το λεγόμενο "τριπλό τεστ" μας δίνει ένα τρόπο ελέγχου των ικανοτήτων ενός υποψήφιου "δοκιμαστή κρασιών" να διακρίνει λεπτές διαφορές στη γεύση, άρωμα και υφή διαφόρων κρασιών. Το τεστ εφαρμόζεται παρουσιάζοντας στον υποψήφιο τρία δείγματα κρασιών, δύο από τα οποία προέρχονται από το ίδιο κρασί, και ζητώντας του αφού τα δοκιμάσει να βρει το διαφορετικό κρασί. Τα δείγματα είναι όσο το δυνατό όμοια και η σειρά παρουσίασης τυχαία. Αν ο υποψήφιος δεν έχει πράγματι την ικανότητα να ξεχωρίζει τα κρασιά, τότε έχει πιθανότητα $1/3$ να βρει το διαφορετικό κρασί και το ερώτημα είναι αν το άτομο μπορεί να φέρει καλύτερα αποτελέσματα.

Για να ελέγξουμε την ικανότητα του υποψηφίου δεν αρκεί ένα μόνο "τριπλό τεστ". Του δίνουμε μία δειρά τέτοιων τεστ με δείγματα παρουσιάζόμενα κάθε φορά στην τύχη για να εξαλείψουμε την περίπτωση συστηματικής μεροληψίας και εξασφαλίσουμε ανεξαρτησία από τεστ σε τεστ. Έστω p η πιθανότητα ο υποψήφιος να βρει σωστά το διαφορετικό κρασί σε μία απλή εφαρμογή του τριπλού τεστ (δοκιμή). Αν ο υποψήφιος έχει την ικανότητα διάκρισης κρασιών τότε $p > 1/3$, αν δεν έχει καμμία ικανότητα τότε $p = 1/3$. Δεχόμαστε ότι το άτομο δεν μπορεί να έχει επίδοση χειρότερη από την τύχη, δηλαδή αποκλείουμε εκ των προτέρων την περίπτωση $p < 1/3$. Δεχόμαστε επίσης ότι η πιθανότητα p είναι σταθερή για το άτομο και δεν μεταβάλλεται με την πάροδο των τεστ. Το p , $0 \leq p \leq 1$,



είναι η άγνωστος στατιστική παράμετρος και κατά κάποιο τρόπο είναι το κριτήριο της διακριτικής ικανότητας του υποψηφίου.

Ορίζουμε τις δύο προηγούμενες υποθέσεις ως ακολούθως

Μηδενική υπόθεση $H_0: p = 1/3$

Εναλλακτική υπόθεση $H_a: p > 1/3$.

Έτσι αντιμετωπίζουμε το πρόβλημα ελέγχου της μηδενικής υπόθεσης H_0 έναντι της εναλλακτικής H_a . Υποθέτουμε επί πλέον εκ των προτέρων ότι δύο είναι οι δυνατές περιπτώσεις: είτε αληθεύει η H_0 είτε αληθεύει η H_a και ότι η εκλογή μας είναι μεταξύ αυτών.

Ο έλεγχος στατιστικών υποθέσεων μας δίδει τον τρόπο για να κρίνουμε ή να αποφασίσουμε για την ορθότητα ή αποδοχή της H_0 . Η όλη διαδικασία ονομάζεται στατιστικό τεστ.

Για να γίνει ο στατιστικός έλεγχος των H_0 και H_a πρέπει να χρησιμοποιήσουμε τις πληροφορίες για το p που μας δίνουν τα στατιστικά δεδομένα, δηλαδή το τυχαίο δείγμα. Αν κάνουμε n δοκιμές με τον υποψήφιο δοκιμαστή και x είναι ο αριθμός των σωστών αναγνωρίσεων που κάνει, τότε η τ.μ. X έχει διωνυμική κατανομή με παραμέτρους το n και κάποιο άγνωστο p :

$$P(X=x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \dots, n.$$

Θέλουμε να χρησιμοποιήσουμε το στατιστικό X για να αποφασίσουμε αν η εναλλακτική υπόθεση H_a είναι πιο λογική ή παραδεκτή από τη μηδενική υπόθεση H_0 , δηλαδή αν ο υποψήφιος έχει την ικανότητα διακρίσεως κρασιών ή όχι. Με άλλα λόγια αν η H_0 αληθεύει ή όχι. Θέλουμε να κατασκευάσουμε, πριν γίνει ο πειραματισμός, ένα κανόνα για να πάρουμε απόφαση. Ο κανόνας αυτός είναι το στατιστικό τεστ. Από άλλη σκοπιά θέλουμε να δούμε με τη βοήθεια ενός στατιστικού τεστ αν τα αποτελέσματα του πειράματος στηρίζουν την H_0 ή την H_a . Αυτά γίνονται κατασκευάζοντας μια περιοχή απόρριψης ή κρίσιμη περιοχή (κ.π.) C , δηλαδή ένα σύνολο τιμών του X που θα μας

οδηγήσει στο να απορρίψουμε την H_0 και να προτιμήσουμε την H_a από την H_0 . Αν το πείραμα δώσει στο X μία αριθμητική τιμή που βρίσκεται εντός του C , τότε απορρίπτουμε την H_0 , διαφορετικά δεν απορρίπτουμε την H_0 .

Για επεξήγηση έστω ότι $n=10$. Θεωρούμε δύο κρίσιμες περιοχές:

$$C_5 = \{5, 6, 7, 8, 9, 10\}$$

και

$$C_8 = \{8, 9, 10\}.$$

Αν χρησιμοποιήσουμε τη C_5 το στατιστικό τεστ είναι: Απορρίπτουμε την H_0 και δεχόμαστε την H_a (δηλαδή υπάρχουν ενδείξεις ότι ο υποψήφιος έχει την ικανότητα διακρίσεως) αν $X \geq 5$. Δεχόμαστε την H_0 (ο υποψήφιος δεν έχει την ικανότητα) αν $X < 5$. Ομοίως το τεστ με κρίσιμο περιοχή την C_8 είναι: Απορρίπτουμε την H_0 αν $X \geq 8$ και δεχόμαστε την H_0 αν $X < 8$.

Οι παραπάνω κρίσιμες περιοχές είναι εύλογες διότι προφανώς όσο πιο μεγάλο είναι το X τόσο μεγαλύτερη ένδειξη έχουμε ότι $p > 1/3$ και επομένως πρέπει να απορρίψουμε την H_0 . Αντίθετα η κρίσιμη περιοχή $C = \{1, 3, 5, 7, 9\}$ δεν είναι εύλογη. Έτσι έχουμε δύο προβλήματα: πως κατασκευάζουμε κρίσιμες περιοχές και πως τις αξιολογούμε. Θα αναπτύξουμε πρώτα το δεύτερο πρόβλημα.

4.2.1 Αξιολόγηση των τεστ - Σφάλματα τύπου I και τύπου II

Ποιά από τις κρίσιμες περιοχές C_5 και C_8 είναι καλύτερη; Όταν κάνουμε ένα στατιστικό τεστ μπορεί να καταλήξουμε σε λανθασμένη απόφαση. Έτσι έχουμε δύο είδη δυνατών σφαλμάτων, τα σφάλματα Τύπου I και Τύπου II:

Σφάλμα Τύπου I: Απορρίπτουμε την H_0 όταν η H_0 αληθεύει.



Σφάλμα Τύπου II: Δεχόμαστε την H_0 όταν η H_a αληθεύει

Διαγραμματικά τα σφάλματα μπορούν να παρασταθούν με τον ακόλουθο πίνακα:

		Στατιστικός (συμπέρασμα)	
		αποδέχεται την H_0	απορρίπτει την H_0
αληθεύει η H_0 πραγματικότητα αληθεύει η H_a	αληθεύει η H_0	0 πιθανότητα = $1-\alpha$	Σφάλμα Τύπου I πιθανότητα = α
	αληθεύει η H_a	Σφάλμα Τύπου II πιθανότητα = β	0 πιθανότητα = $1-\beta$

Σφάλματα Στατιστικών Τεστ

Αρχίζοντας με το άνω αριστερά τετραγωνίδιο και πηγαίνοντας προς τη φορά των δεικτών του ωρολογίου έχουμε: όταν η H_0 αληθεύει και ο Στατιστικός αποδεχθεί την H_0 δεν γίνεται σφάλμα, ενώ αν ο Στατιστικός απορρίψει την H_0 τότε γίνεται σφάλμα Τύπου I. Ομοίως όταν αληθεύει η H_a και ο Στατιστικός αποδεχθεί την H_0 γίνεται σφάλμα Τύπου II, ενώ στην αντίθετη περίπτωση δεν γίνεται σφάλμα. Τα δύο σφάλματά είναι προφανώς τυχαία ενδεχόμενα.

Η επίδοση ενός τεστ ή κρίσιμης περιοχής αξιολογείται με τις πιθανότητες των σφαλμάτων των παραπάνω τύπων:

$$\begin{aligned}\alpha &= P(\text{Σφάλμα Τύπου I}) \\ &= P(\text{Απορρίπτεται η } H_0 | H_0 \text{ αληθινή})\end{aligned}$$

(4.12)

$$\begin{aligned}\beta &= P(\text{Σφάλμα Τύπου II}) \\ &= P(\text{Δεκτή η } H_0 | H_0 \text{ λανθασμένη}).\end{aligned}$$



Η πιθανότητα α του Σφάλματος Τύπου I λέγεται και μέγεθος της κρίσιμης περιοχής ή επίπεδο σημαντικότητας του στατιστικού τεστ. Το $\gamma = 1 - \beta$ λέγεται ισχύς ή δύναμη του τεστ. Είναι φανερό ότι θέλουμε τα α και β να είναι όσο το δυνατό μικρότερα.

Για την κρίσιμη περιοχή C_5 , σφάλμα τύπου I γίνεται όταν ο δοκιμαστής δεν έχει την διακριτική ικανότητα αλλά επιτυχάνει 5 ή περισσότερες επιτυχίες αναγνωρίσεις απλώς στην τύχη. Έτσι απορρίπτομε την H_0 λανθασμένα. Η πιθανότητα να γίνει αυτό είναι:

$$\begin{aligned}\alpha(C_5) &= P(\text{απορρίπτεται η } H_0 | H_0 \text{ αληθινή}) \\ &= P(X \geq 5 | p = 1/3) = \sum_{x=5}^{10} \binom{10}{x} \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{10-x} \\ &= \frac{1}{3^{10}} \left[\binom{10}{5} 2^4 + \binom{10}{6} 2^5 + \dots + \binom{10}{9} 2 + 1 \right] \\ &= 0.2131.\end{aligned}$$

Για την κρίσιμη περιοχή C_8 η πιθανότητα του σφάλματος τύπου I είναι:

$$\alpha(C_8) = P(X \geq 8 | p = 1/3) = \sum_{x=8}^{10} \binom{10}{x} \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{10-x} = 0.0034.$$

Έτσι το $\alpha(C_8)$ είναι μικρότερο του $\alpha(C_5)$. Όσον αφορά το σφάλμα τύπου I το τεστ C_8 είναι καλύτερο από το τεστ C_5 .

Ας υπολογίσουμε τώρα τις πιθανότητες του σφάλματος τύπου II. Όταν η H_0 είναι λανθασμένη, δηλ. $p > 1/3$, το β εξαρτάται από τη συγκεκριμένη τιμή του p που θα θεωρήσουμε. Έτσι έστω $p = 0.70$. Τότε για τις κρίσιμες περιοχές C_5 και C_8 έχουμε



$$\beta(C_5, p=0.70) = P(\text{Δεκτή} \mid H_0 | p=0.7)$$

$$= P(X < 5 | p=0.7) = \sum_{x=0}^4 \binom{10}{x} (0.7)^x (0.3)^{10-x}$$

$$= 0.0473$$

$$\beta(C_8, p=0.70) = P(X < 8 | p=0.7) = \sum_{x=0}^7 \binom{10}{x} (0.7)^x (0.3)^{10-x}$$

$$= 0.6171,$$

δηλαδή $\beta(C_8, p=0.70) > \beta(C_5, p=0.70)$. Έτσι ως προς το σφάλμα τύπου II το τεστ C_8 είναι χειρότερο του C_5 .

Βλέπουμε δηλαδή ότι με γνώμονα το σφάλμα τύπου I η κρίσιμη περιοχή C_8 είναι προτιμώτερη από την C_5 και αντίστροφα η C_5 είναι προτιμώτερη από την C_8 όταν ο γνώμονας είναι το σφάλμα τύπου II. Αυτό είναι αναπόφευκτο: καθώς το α μικραίνει το β μεγαλώνει και αντίστροφα όταν το α μεγαλώνει το β μικραίνει. Στην ακραία περίπτωση που ποτέ δεν κάνουμε σφάλμα Τύπου I, δηλαδή ποτέ δεν απορρίπτομε την H_0 , η κρίσιμη περιοχή C είναι το κενό σύνολο και το $\alpha=0$. Αλλά τότε πάντοτε δεχόμαστε την H_0 και $\beta=1$. Στην άλλη ακραία περίπτωση που ποτέ δεν κάνουμε σφάλμα Τύπου II, δηλαδή ποτέ δεν δεχόμαστε την H_0 , $\beta=0$. Αλλά τότε πάντοτε απορρίπτομε την H_0 και $\alpha=1$. Είναι αδύνατον τα α και β να γίνουν και τα δύο 0 ή και τα δύο 1 για μια συγκεκριμένη κρίσιμη περιοχή.

Με άλλα λόγια είναι αδύνατη η ταυτόχρονη, δηλαδή για το ίδιο C , ελαχιστοποίηση των $\alpha(C)$ και $\beta(C)$. Το πρόβλημα αυτό παρακάμπτεται κρατώντας το α σταθερό, καθορισμένο εκ των προτέρων και επιδιώκοντας να βρούμε εκείνη την κρίσιμη περιοχή C με μέγεθος μικρότερο ή ίσο του α και ελάχιστο β ή ισοδύναμα, τη μεγίστη ισχύ $\gamma = 1 - \beta$. Το σταθερό προκαθορισμένο α λέγεται επίπεδο

σημαντικότητας του τεστ και συνήθως παίρνει τιμές 1%, 5%, ή 10%. Το τεστ λέγεται και τεστ επιπέδου ή μεγέθους α .

Έτσι στο παράδειγμα του "τριπλού τεστ" αν πάρουμε το $\alpha=5\%$ είναι εύκολο να δούμε ότι η κρίσιμη περιοχή

$$C_7 = \{7, 8, 9, 10\}$$

είναι η μεγαλύτερη κρίσιμη περιοχή με

$$\alpha(C_7) = 0.0196 < 0.05.$$

Μπορεί επί πλέον να αποδειχθεί ότι η C_7 είναι η ισχυρότερη (άριστη) κρίσιμη περιοχή επιπέδου σημαντικότητας 5% για τον έλεγχο της H_0 έναντι της H_a . Δηλαδή

$$\beta(C_7, \rho) \leq \beta(C, \rho)$$

για κάθε C με $\alpha(C) \leq 0.05$ και κάθε $\rho > 1/3$.

4.2.2 Τα στοιχεία του τεστ

Από τα παραπάνω είναι φανερό ότι κάθε στατιστικό τεστ αποτελείται από τα ακόλουθα στοιχεία:

1. Την μηδενική H_0 και την εναλλακτική H_a υπόθεση.
2. Το προκαθορισμένο επίπεδο σημαντικότητας α .
3. Την στατιστική συνάρτηση του τεστ και την κρίσιμη περιοχή (κ.π.) C .
4. Την τιμή της στατιστικής συνάρτησης του τεστ που προκύπτει από τα δεδομένα (μετρήσεις).



5. Το συμπέρασμα: αν η τιμή της στατιστικής συνάρτησης του τεστ πέσει μέσα στην κρίσιμη περιοχή C ή H_0 απορρίπτεται υπέρ της H_a · αν πέσει εκτός ή H_0 δεν απορρίπτεται.

Για κάθε ένα από τα παραπάνω στοιχεία έχουμε τις ακόλουθες παρατηρήσεις:

Στοιχείο 1. Οι υποθέσεις H_0 και H_a διατυπώνονται με την βοήθεια των στατιστικών παραμέτρων του μοντέλου που υιοθετείται για την ανάλυση των μετρήσεων. Διακρίνονται σε απλές (π.χ. $p=1/3$) και σύνθετες (π.χ. $p>1/3$). Διακρίνονται επίσης και σε μονόπλευρες (π.χ. $p>1/3$) και δίπλευρες (π.χ. $p\neq 1/3$). Υποθέσεις τύπου $\theta>\theta_0$, όπου θ_0 γνωστό λέγονται δεξιόπλευρες και υποθέσεις τύπου $\theta<\theta_0$ λέγονται αριστερόπλευρες.

Η μηδενική υπόθεση είναι συνήθως το "αντίθετο" του ισχυρισμού που θέλουμε να ελέγξουμε. Μηδενοποιεί τον ισχυρισμό αυτό. Έτσι στο παράδειγμα του "τριπλού τεστ" αν ο ισχυρισμός του υποψηφίου είναι ότι έχει την ικανότητα διακρίσεως (δηλαδή $p > 1/3$) τότε η H_0 είναι $p = 1/3$ και $H_a: p = 1/3$. Αν κάνουμε σφυγμομέτρηση της κοινής γνώμης με σκοπό να αποδείξουμε ότι λιγότεροι του 30% των πολιτών υποστηρίζουν κάποιο θέμα τότε η H_0 είναι $p\geq 0.3$ και η H_a είναι $p < 0.3$, όπου p είναι το ποσοστό των πολιτών που υποστηρίζουν το θέμα. Όταν θέλουμε να ελέγξουμε την υπόθεση ότι η μακρόχρονη αποθήκευση του αίματος ή πλάσματος σε καταψύκτη δεν αλλάζει την περιεκτικότητα του σε χοληστερίνη, τριγλυκερίδια και άλλες ουσίες, η H_0 είναι $\mu=0$ και η H_a είναι $\mu\neq 0$, όπου μ είναι η μέση διαφορά μετρήσεων περιεκτικότητας φρέσκου αίματος και του ίδιου αίματος μετά από ορισμένους μήνες αποθήκευσης.

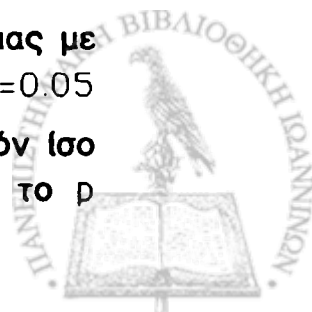
Στοιχείο 2. Τα α εκφράζει το ποσοστό (πιθανότητα) σφάλματος Τύπου I που είμαστε διατεθειμένοι να ανεχτούμε.



Στοιχείο 3. Η στατιστική συνάρτηση του τεστ και η κρίσιμη περιοχή είναι τα κύρια στοιχεία του τεστ. Στην πιο γενική περίπτωση η κ.π. του τεστ είναι ένα υποσύνολο του δειγματικού χώρου S , ο οποίος συνήθως είναι πολυδιάστατος. Επειδή όμως συχνά χρησιμοποιούνται μετασχηματισμοί από το S στην ευθεία γραμμή η κ.π. δίδεται με τη βοήθεια ισοδύναμων μονοδιάστατων κρίσιμων περιοχών. Οι μετασχηματισμοί αυτοί είναι οι στατιστικές συναρτήσεις των τεστ. Έτσι η σ.σ. του τεστ είναι μία τυχαία ποσότητα με γνωστή κατανομή όταν η H_0 αληθεύει. Η μορφή της H_a καθορίζει την θέση της κρίσιμης περιοχής στον άξονα των x . Αν η H_a είναι δεξιόπλευρη η κρίσιμη περιοχή του τεστ είναι συνήθως της μορφής $[c_0, +\infty)$, αν H_a είναι αριστερόπλευρη η κ.π. είναι της μορφής $(-\infty, c_0]$, και τέλος αν η H_a δίπλευρη η κ.π. είναι της μορφής $(-\infty, c_1] \cup [c_2, +\infty)$, όπου c_0, c_1, c_2 γνωστοί αριθμοί. Τα c_0, c_1, c_2 υπολογίζονται από την κατανομή της σ.σ. του τεστ όταν αληθεύει η H_0 . Ανάλογα τα τεστ λέγονται δεξιόπλευρα, αριστερόπλευρα και δίπλευρα ή τεστ δεξιάς ουράς, αριστεράς ουράς ή δύο ουρών της κατανομής.

Στοιχείο 4. Οι μετρήσεις (το τυχαίο δείγμα) πρέπει να "ταιριάζουν" στο στατιστικό μοντέλο που υιοθετήθηκε. Στο παράδειγμά μας, έχουμε πράγματι μετρήσεις από διωνυμικό πείραμα.

Στοιχείο 5. Όταν απορρίπτουμε την H_0 , τότε είτε την απορρίπτουμε σωστά είτε λανθασμένα. Στη δεύτερη περίπτωση το σφάλμα μας ελέγχεται από το α , έχει δηλαδή πιθανότητα μικρότερη ή ίση του α , το οποίο είναι γνωστό προκαθορισμένο με μικρή τιμή. Όταν δεν απορρίπτουμε την H_0 πάλι είτε ενεργούμε σωστά είτε λανθασμένα. Στη δεύτερη περίπτωση το σφάλμα μας ελέγχεται από το β , το οποίο όμως είναι άγνωστο διότι εξαρτάται από τις τιμές του θ που ανήκουν στο H_a . Έτσι δεν ξέρομε αν το σφάλμα μας έχει μικρή ή μεγάλη πιθανότητα. Στο παράδειγμά μας με το "τριπλό τεστ" και με κρίσιμη περιοχή το C_7 και $\alpha=0.05$ αν το p είναι λίγο μεγαλύτερο από το $1/3$, το β είναι σχεδόν ίσο με το $1 - \alpha(C_7) = 1 - 0.01916 = 0.9804 \approx 95\%$, ενώ αν το p

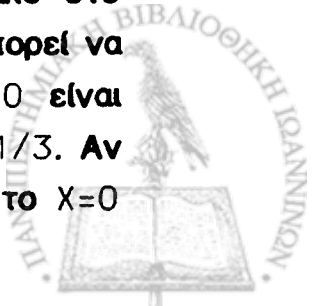


είναι 0.95 τότε $\beta=0.0011$. Γενικά οι τιμές του β είναι μικρές για μεγάλο p αλλά σχεδόν ίσες με $1-\alpha$ για p κοντά στο $1/3$. Αν η τιμή της σ.σ. του τεστ δεν πέσει μέσα στην κ.π. δεν μπορούμε με εμπιστοσύνη να αποδεχθούμε την H_0 διότι δεν μπορούμε με εμπιστοσύνη ν'απορρίψουμε "ολόκληρη" την H_a . Έτσι υπάρχει μία ασυμμετρία στον τρόπο μεταχειρήσεως των H_0 και H_a .

Από τα παραπάνω βγαίνει το ακόλουθο συμπέρασμα: Είναι προτιμώτερο να απορρίπτουμε την H_0 παρά να την αποδεχόμαστε. Όταν απορρίπτουμε την H_0 το ποσοστό σφάλματος είναι εκ των προτέρων γνωστό ή είναι μικρότερο ή ίσο ενός γνωστού ορίου, του α . Για τους ίδιους λόγους αποφεύγουμε να λέμε ότι "αποδεχόμαστε" την H_0 και χρησιμοποιούμε τον όρο "δεν απορρίπτεται η H_0 ", ή λέμε ότι τα στατιστικά δεδομένα δεν μας δίνουν αρκετές ενδείξεις για να απορρίψουμε την μηδενική υπόθεση.

Μερικές φορές τα στατιστικά τεστ λέγονται και τεστ σημαντικότητας. Η διαφορά με τα τεστ στατιστικών υποθέσεων είναι περισσότερο φιλοσοφικής μορφής παρά ουσιαστικής. Τα τεστ στατιστικών υποθέσεων είναι μία μεθοδολογία λήψεως αποφάσεως τύπου "απορρίπτω - δέχομαι". Τα τεστ σημαντικότητας αντίθετα κρίνουν αν ένα πειραματικό αποτέλεσμα είναι στατιστικώς σημαντικό. Ένα αποτέλεσμα είναι στατιστικώς σημαντικό αν η πιθανότητα να συμβεί το αποτέλεσμα αυτό κάτω από ορισμένες προϋποθέσεις H_0 είναι μικρή. Το γεγονός ότι το αποτέλεσμα πραγματοποιείται, δίνει σημαντικές ενδείξεις εναντίον της H_0 και φανερώνει ότι τα πειραματικά δεδομένα δεν την υποστηρίζουν.

Στο παράδειγμα του δοκιμαστού κρασιών ένα X με τιμές 7, 8, 9 ή 10 είναι στατιστικώς σημαντικό στο επίπεδο 1.97% διότι όταν αληθεύει η H_0 , δηλ. $p = 1/3$, η $P(X \geq 7)$ ισούται με 0.0197. Αλλά και η $P(X=0)$ είναι μικρότερη του 1.97%. Είναι λοιπόν ένα X ίσο με το 0 στατιστικώς σημαντικό στο επίπεδο 1.97%; Όχι διότι υποθέσαμε ότι ο δοκιμαστής δεν μπορεί να έχει επίδοση χειρότερη από την τύχη. Αν λοιπόν το $X=0$ είναι ασύνηθες όταν $p=1/3$, είναι περισσότερο ασύνηθες όταν $p>1/3$. Αν όμως το p μπορεί να είναι και μικρότερο του $1/3$ τότε και το $X=0$



είναι στατιστικώς σημαντικό δηλαδή μας δίνει σημαντική ένδειξη εναντίον της $H_0: p=1/3$.

4.2.3 Γενικές παρατηρήσεις

Είναι φανερό ότι ποτέ δεν μπορούμε να αποδείξουμε την ορθότητα μίας υπόθεσης. Πρέπει βέβαια να εξαιρέσουμε τετρωμένες ή εξεζητημένες περιπτώσεις που με την εκτέλεση ενός ή περισσοτέρων πειραμάτων αποκαλύπτεται ο άγνωστος χαρακτήρας του πληθυσμού. Ομοίως και περιπτώσεις που δεν κυβερνώνται από τους νόμους της τύχης. Η θέση της φύσης είναι άγνοση και οιοσδήποτε ή οσοσδήποτε πειραματισμός είναι αδύνατο να την αποκαλύψει.

Η αποδοχή ή απόρριψη της H_0 εξαρτάται από τους κινδύνους (risks) που είμαστε διατεθειμένοι να δεχθούμε. Εξαρτάται από τα α και β . Αλλάζοντας το α μπορεί να αλλάξει η απόφαση ή το συμπέρασμα. Αν η ισχύς του τεστ είναι γνωστή, τότε αποδεχόμενοι την H_0 έχουμε πλήρη γνώση των κινδύνων της απόφασης μας. Συχνά αντί να δεχτούμε μία μηδενική υπόθεση προτιμάμε να εκτιμήσουμε την άγνωστη παράμετρο με ένα σημείο ή ένα διάστημα εμπιστοσύνης.

Η διατύπωση των υποθέσεων πρέπει να γίνεται πριν γίνει το πείραμα και ληφθούν οι μετρήσεις. Διαφορετικά τα συμπεράσματα δεν είναι αμερόληπτα.

Μετά την παρουσίαση της φιλοσοφίας των λογικών αρχών των στατιστικών τεστ απομένει να παρουσιάσουμε την θεωρία εύρεσης και τις ιδιότητες των στατιστικών συναρτήσεων των τεστ. Αυτό αποτελεί αντικείμενο των κλάδων Έλεγχος Στατιστικών Υποθέσεων και Θεωρία Στατιστικών Αποφάσεων και είναι εκτός του σκοπού του παρόντος βιβλίου. Εδώ θα περιοριστούμε στην ακόλουθη αρχή κατασκευής στατιστικής συνάρτησης του τεστ που απορρέει από την ορισμό τους: στατιστική συνάρτηση του τεστ είναι μία τυχαία ποσότητα με γνωστή κατανομή όταν αληθεύει η H_0 .

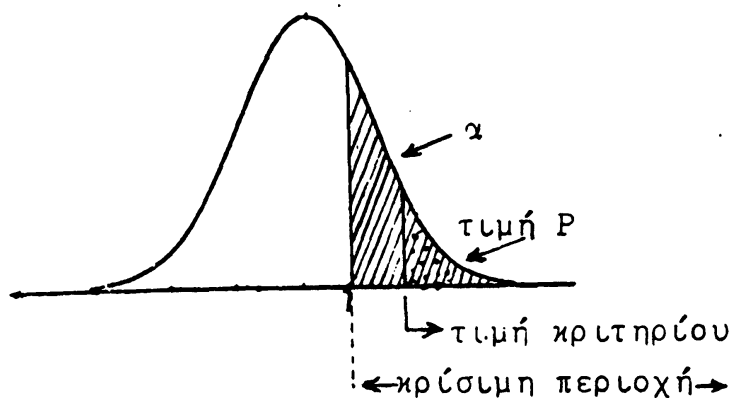


4.2.4 Τιμή P

Τελευταία σε πολλούς κλάδους και ιδίως στο χώρο των Βιοϊατρικών Επιστημών τα στατιστικά τεστ γίνονται με τη βοήθεια της λεγόμενης τιμής P (P value).

Ορισμός: Τιμή P ενός στατιστικού τεστ είναι η πιθανότητα το στατιστικό κριτήριο να πάρει τιμές πιο ακραίες από εκείνη που παρατηρήθηκε, όταν αληθεύει η H_0 .

Όπως έχει αναπτυχθεί πιο πάνω τα στατιστικά τεστ γίνονται με προκαθορισμένο το α . Το προκαθορισμένο α καθορίζει και την κρίσιμη περιοχή. Έτσι αν η τιμή της στατιστικής συνάρτησης του τεστ (του κριτηρίου) που χρησιμοποιούμε είναι μέσα στην κρίσιμη περιοχή, απορρίπτουμε την μηδενική υπόθεση H_0 και το τεστ είναι στατιστικά σημαντικό. Έτσι μεταβάλλοντας το α ένα τεστ μπορεί από σημαντικό να γίνει μη σημαντικό. Η τιμή του α για την οποία γίνεται η αλλαγή αυτή είναι η τιμή P. Η τιμή P λέγεται και "παρατηρηθέν" επίπεδο σημαντικότητας και συχνά είναι η πιθανότητα η τιμή του κριτηρίου του τεστ να είναι μεγαλύτερη εκείνης που παρατηρήθηκε. Μερικές τιμές P υποδηλώνουν απόρριψη της μηδενικής υπόθεσης. Έτσι αν $\alpha=5\%$ και $P<5\%$ η H_0 απορρίπτεται. Όσο πιο μικρή είναι η τιμή P τόσο ισχυρότερη ένδειξη έχουμε εναντίον της H_0 . Γραφικά η τιμή P, το α και η κρίσιμη περιοχή με $P<\alpha$ έχουν ως εξής:



4.3 Απλή Στατιστική Συμπερασματολογία

Στο εδάφιο αυτό παρουσιάζουμε μερικές εφαρμογές της στατιστικής συμπερασματολογίας που θα εμπεδώσουν τόσο τα διαστήματα εμπιστοσύνης όσο και τα στατιστικά τεστ.

4.3.1 Συμπερασματολογία για τη μέση τιμή

Έστω $X_1, X_2, \dots, X_n$ ένα τυχαίο δείγμα μεγέθους n από ένα πληθυσμό με μέση τιμή μ και διακύμανση σ^2 . Έστω $\bar{X}$ η μέση τιμή του δείγματος και S^2 η διακύμανση του. Θέλουμε να εκτιμήσουμε το μ και να ελέγξουμε υποθέσεις για το μ . Υποθέτουμε στην αρχή ότι ο πληθυσμός είναι κανονικός. Αργότερα θα χαλαρώσουμε την υπόθεση αυτή.

Ο καλύτερος εκτιμητής του μ είναι ο

$$\hat{\mu} = \bar{X}$$

με εκτιμώμενο τυπικό σφάλμα

$$\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}}$$

(βλέπε § 4.1.1). Το καλύτερο διάστημα εμπιστοσύνης για το μ είναι

$$L = \bar{X} - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}, \quad U = \bar{X} + t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}. \quad (4.13)$$

Το διάστημα αυτό απορρέει από το γεγονός ότι η ποσότητα

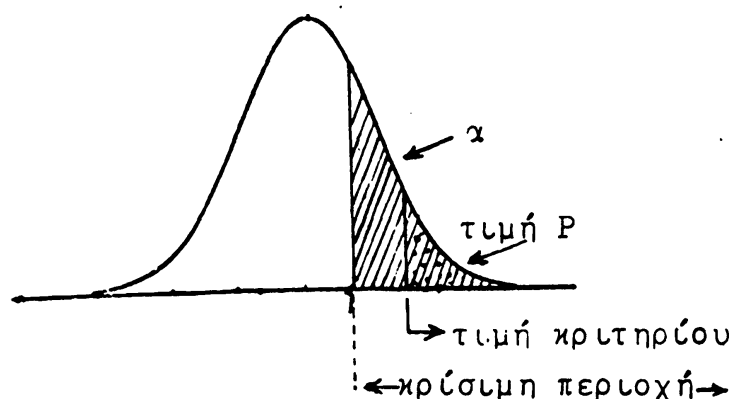
$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

4.2.4 Τιμή P

Τελευταία σε πολλούς κλάδους και ιδίως στο χώρο των Βιοϊατρικών Επιστημών τα στατιστικά τεστ γίνονται με τη βοήθεια της λεγόμενης τιμής P (P value).

Ορισμός: Τιμή P ενός στατιστικού τεστ είναι η πιθανότητα το στατιστικό κριτήριο να πάρει τιμές πιο ακραίες από εκείνη που παρατηρήθηκε, όταν αληθεύει η H_0 .

Όπως έχει αναπτυχθεί πιο πάνω τα στατιστικά τεστ γίνονται με προκαθορισμένο το α . Το προκαθορισμένο α καθορίζει και την κρίσιμη περιοχή. Έτσι αν η τιμή της στατιστικής συνάρτησης του τεστ (του κριτηρίου) που χρησιμοποιούμε είναι μέσα στην κρίσιμη περιοχή, απορρίπτουμε την μηδενική υπόθεση H_0 και το τεστ είναι στατιστικά σημαντικό. Έτσι μεταβάλλοντας το α ένα τεστ μπορεί από σημαντικό να γίνει μη σημαντικό. Η τιμή του α για την οποία γίνεται η αλλαγή αυτή είναι η τιμή P. Η τιμή P λέγεται και "παρατηρηθέν" επίπεδο σημαντικότητας και συχνά είναι η πιθανότητα η τιμή του κριτηρίου του τεστ να είναι μεγαλύτερη εκείνης που παρατηρήθηκε. Μερικές τιμές P υποδηλώνουν απόρριψη της μηδενικής υπόθεσης. Έτσι αν $\alpha=5\%$ και $P<5\%$ η H_0 απορρίπτεται. Όσο πιο μικρή είναι η τιμή P τόσο ισχυρότερη ένδειξη έχουμε εναντίον της H_0 . Γραφικά η τιμή P, το α και η κρίσιμη περιοχή με $P<\alpha$ έχουν ως εξής:



4.3 Απλή Στατιστική Συμπερασματολογία

Στο εδάφιο αυτό παρουσιάζουμε μερικές εφαρμογές της στατιστικής συμπερασματολογίας που θα εμπεδώσουν τόσο τα διαστήματα εμπιστοσύνης όσο και τα στατιστικά τεστ.

4.3.1 Συμπερασματολογία για τη μέση τιμή

Έστω $X_1, X_2, \dots, X_n$ ένα τυχαίο δείγμα μεγέθους n από ένα πληθυσμό με μέση τιμή μ και διακύμανση σ^2 . Έστω $\bar{X}$ η μέση τιμή του δείγματος και S^2 η διακύμανση του. Θέλουμε να εκτιμήσουμε το μ και να ελέγξουμε υποθέσεις για το μ . Υποθέτουμε στην αρχή ότι ο πληθυσμός είναι κανονικός. Αργότερα θα χαλαρώσουμε την υπόθεση αυτή.

Ο καλύτερος εκτιμητής του μ είναι ο

$$\hat{\mu} = \bar{X}$$

με εκτιμώμενο τυπικό σφάλμα

$$\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}}$$

(βλέπε § 4.1.1). Το καλύτερο διάστημα εμπιστοσύνης για το μ είναι

$$L = \bar{X} - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}, \quad U = \bar{X} + t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}. \quad (4.13)$$

Το διάστημα αυτό απορρέει από το γεγονός ότι η ποσότητα

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$



(βλέπε Θεώρημα 3.3).

Για να ελέγξουμε υποθέσεις της μορφής

$$(a) \quad H_0: \mu \leq \mu_0 \quad \text{έναντι} \quad H_a: \mu > \mu_0$$

$$(b) \quad H_0: \mu \geq \mu_0 \quad \text{έναντι} \quad H_a: \mu < \mu_0$$

$$(c) \quad H_0: \mu = \mu_0 \quad \text{έναντι} \quad H_a: \mu \neq \mu_0$$

με επίπεδο σημαντικότητας α χρησιμοποιούμε το στατιστικό

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \quad \text{το οποίο έχει κατανομή } t_{n-1} \quad \text{όταν } \mu = \mu_0 \quad (4.14)$$

και τις ακόλουθες αντίστοιχες κρίσιμες περιοχές:

$$\text{για την (a)} \quad C = (t_{\alpha, n-1}, +\infty), \quad \text{δηλ. } t > t_{\alpha, n-1}$$

$$\text{για την (b)} \quad C = (-\infty, -t_{\alpha, n-1}), \quad \text{δηλ. } t < -t_{\alpha, n-1}$$

$$\text{για την (c)} \quad C = (-\infty, -t_{\alpha/2, n-1}) \cup (t_{\alpha/2, n-1}, +\infty),$$

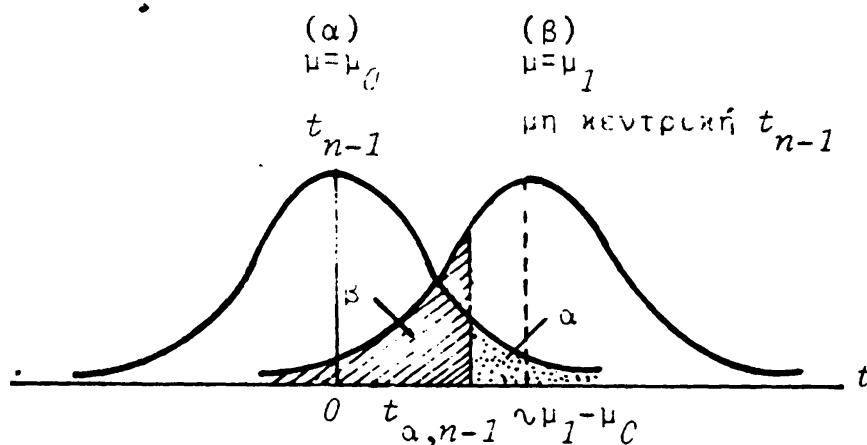
$$\text{δηλ. } |t| > t_{\alpha/2, n-1}$$

Το τεστ αυτό λέγεται t τεστ.

Οι παραπάνω κρίσιμες περιοχές είναι οι καλύτερες δυνατές. Αυτό αποδεικνύεται στη Θεωρία Ελέγχου Στατιστικών Υποθέσεων. Μπορεί να φανεί όμως και διαισθητικά. Πράγματι, για τις υποθέσεις (a) μεγάλες θετικές τιμές του $\bar{X}$ υποδηλώνουν απόρριψη της H_0 και αποδοχή της H_a . Αλλά μεγάλες θετικές τιμές του $\bar{X}$ σημαίνουν μεγάλες θετικές τιμές του t .

Το Σχήμα 4.2 μας δίνει μια γραφική παράσταση της

κρίσιμης περιοχής του t -τεστ, του α και του β (για $\mu = \mu_1 > \mu_0$) για την περίπτωση των υποθέσεων (α). Όταν ισχύει η H_0 η σ.σ. (4.14) του t -τεστ έχει κατανομή t_{n-1} . Αν $\mu = \mu_1 > \mu_0$ η σ.σ. (4.14) έχει μη κεντρική t_{n-1} κατανομή η οποία έχει ανάλογο σχήμα με την t_{n-1} αλλά είναι μετατοπισμένη προς τα δεξιά με μέση τιμή ανάλογη του $\mu_1 - \mu_0$. Η κ.π. καθορίζεται με το σημείο t_α . Δεξιά του και κάτω από την καμπύλη (α) έχουμε το α , ενώ αριστερά του και κάτω από την καμπύλη (β) έχουμε το β .



Σχήμα 4.2 Κρίσιμη περιοχή, α και β για t -τεστ

Παράδειγμα 4.2: Ένας υγειονομικός σταθμός θέλει να ελέγξει αν ο μέσος αριθμός βακτηριδίων ανά μονάδα όγκου θαλασσινού νερού σε μία παραλία υπερβαίνει το επίπεδο ασφαλείας των 200. Δώδεκα δείγματα νερού συλλέγονται και ευρίσκονται οι ακόλουθοι αριθμοί βακτηριδίων ανά μονάδα όγκου:

170,	175,	190,	198,	215,	185
184,	207,	210,	193,	196,	180.

Υπάρχει λόγος ανησυχίας;

Έστω μ ο μέσος αριθμός βακτηριδίων ανά μονάδα όγκου και σ^2 η διακύμανση. Δεδομένου ότι η αποτυχία διαπίστωσης



παραβίασης του ορίου ασφαλείας μπορεί να έχει σοβαρές συνέπειες η μηδενική και η εναλλακτική υπόθεση είναι

$$H_0: \mu \geq 200 \quad \text{και} \quad H_a: \mu < 200.$$

Υποθέτουμε ότι οι μετρήσεις αποτελούν ένα τυχαίο δείγμα από μία κανονική κατανομή και ότι το επίπεδο σημαντικότητας είναι $\alpha=1\%$. Το στατιστικό τεστ

$$t = \frac{\bar{X} - 200}{S/\sqrt{12}}$$

έχει t κατανομή με 11 βαθμούς ελευθερίας και κρίσιμη περιοχή

$$C = (-\infty, -t_{0.01, 11} = -2.718).$$

Οι υπολογισμοί με τα παραπάνω δεδομένα μας δίνουν

$$\bar{X} = 191.9$$

$$S = 14.03$$

$$t = \frac{191.1 - 200}{14.03/\sqrt{12}} = \frac{-8.9}{4.05} = -2.197.$$

Επειδή η παρατηρηθείσα τιμή $t=-2.197$ είναι εκτός της κ.π. C , η H_0 δεν απορρίπτεται στο επίπεδο του 1%. Δεν υπάρχει ένδειξη ότι το μέσο επίπεδο των βακτηριδίων είναι εντός των ορίων ασφαλείας.

Στο σημείο αυτό ίσως θελήσουμε να εκτιμήσουμε το μ . Πράγματι ο εκτιμητής του μ είναι ο $\hat{\mu} = \bar{X} = 191.9$ βακτηρίδια ανά μονάδα όγκου και έχει τυπικό σφάλμα $\hat{\sigma}_{\bar{X}} = S/\sqrt{n} = 4.05$.

Ένα 95% διάστημα εμπιστοσύνης για το μ είναι το



$$\bar{X} \pm t_{0.025, 11} \frac{S}{\sqrt{n}} = 191.9 \pm 2.201 \times 4.05 = 191.9 \pm 8.91.$$

Παράδειγμα 4.3: Σε μια προσπάθεια να προσδιοριστεί αν η ειδική εκπαίδευση αυξάνει ή όχι τον δείκτη ευφυΐας, είκοσι πέντε παιδιά εξετάζονται με ένα βασικό - τυποποιημένο τεστ ευφυΐας. Κατόπιν τα παιδιά αυτά παρακολουθούν ένα ειδικό μάθημα, ο σκοπός του οποίου είναι η αύξηση του δείκτη ευφυΐας. Στο τέλος του μαθήματος εξετάζονται για δεύτερη φορά. Η διαφορά μεταξύ των βαθμών της δευτέρας και πρώτης εξέτασης καταγράφεται για κάθε παιδί. Έστω ότι η μέση τιμή των διαφορών αυτών είναι $\bar{X} = 6$ μονάδες (σε κλίμακα 0-100) και η διακύμανση $S^2 = 64$. Έχει η ειδική εκπαίδευση αυξήσει τον δείκτη ευφυΐας;

Έστω μ η μέση τιμή των διαφορών. Η μηδενική υπόθεση είναι ότι η ειδική εκπαίδευση δεν είναι αποτελεσματική, δηλαδή $H_0: \mu = 0$. Η εναλλακτική είναι $H_a: \mu > 0$, δηλαδή υπάρχει θετική αύξηση. Θα χρησιμοποιήσουμε δεξιόπλευρο t -τεστ και θέτουμε $\alpha = 5\%$. Έτσι $t_{0.05, 24} = 1.711$.

Η τιμή της σ.σ. του t -τεστ είναι

$$t = \frac{\bar{X} - 0}{S/\sqrt{25}} = \frac{6 - 0}{8/5} = 3.75 > 1.711 = t_{\alpha, n-1}.$$

Έτσι απορρίπτεται η H_0 και βγάζουμε το συμπέρασμα ότι πράγματι η ειδική εκπαίδευση αυξάνει το δείκτη ευφυΐας. Το ποσοστό σφάλματος μας είναι 5%.

Όσες φορές χρησιμοποιούμε την κατανομή t_v για υπολογισμό διαστημάτων εμπιστοσύνης ή τεστ στατιστικών υποθέσεων υποθέτουμε ότι τα δεδομένα μας είναι ένα τυχαίο δείγμα από ένα κανονικό πληθυσμό. Η πρώτη υπόθεση του τυχαίου δείγματος είναι ουσιαστική αν θέλουμε να διατυπώσουμε πιθανοθεωρητικά συμπεράσματα για τα αποτελέσματα μας. Δεν μπορεί να χαλαρωθεί.

Η δεύτερη υπόθεση του κανονικού πληθυσμού δεν είναι τόσο ισχυρή και μπορεί να χαλαρωθεί αρκεί να έχουμε ένα αρκετά



μεγάλο δείγμα. Αυτό είναι άμεση συνέπεια του Κεντρικού Οριακού Θεωρήματος.

Μέγεθος του δείγματος: Η στατιστική θεωρία προσφέρει πολύτιμη βοήθεια κατά τη σχεδίαση μελετών, ερευνών ή πειραμάτων. Μία τέτοια περίπτωση είναι ο προσδιορισμός του μεγέθους του δείγματος.

Η ακρίβεια (αξιοπιστία) ενός εκτιμητή $\hat{\theta}$ μετρείται με το τυπικό του σφάλμα $\sigma_{\hat{\theta}}$ το οποίο συνήθως είναι

$$\sigma_{\hat{\theta}} = \sigma / \sqrt{n},$$

όπου σ η τυπική απόκλιση του πληθυσμού και n το μέγεθος του δείγματος. Έτσι αν η μεταβλητότητα (δηλ. το σ) του πληθυσμού είναι γνωστή (π.χ. από προηγούμενες μελέτες) μπορούμε να προσδιορίσουμε το n ώστε να πετύχουμε μία δεδομένη ακρίβεια για το $\hat{\theta}$. Έστω τώρα ότι θέλουμε να εκτιμήσουμε τη μέση τιμή μ ενός πληθυσμού από ένα μεγάλο δείγμα κατά τέτοιο τρόπο ώστε η μέση τιμή $\bar{X}$ του δείγματος να ευρίσκεται (με εμπιστοσύνη $100(1-\alpha)\%$) εντός ω μονάδων από το μ . Με άλλα λόγια θέλουμε το μήκος του Δ.Ε. (4.13) να είναι 2ω . Έτσι

$$U-L = 2t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} = 2\omega$$

ή

$$n = \frac{t_{\alpha/2, n-1}^2 S^2}{\omega^2} \quad (4.15)$$

Η λύση ως προς n της εξίσωσης (4.15) δεν είναι εύκολη. Στην πράξη όμως μπορούμε να προσεγγίσουμε το $t_{\alpha/2, n-1}$ με το $z_{\alpha/2}$ και το S με κάποιο σ που προκύπτει από προσωπική γνώση, παλαιά

δεδομένα, ή πληροφορία από τη βιβλιογραφία. Έτσι έχουμε τη σχέση

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{\omega^2} \quad (4.16)$$

Βλέπουμε δηλαδή ότι καθώς η διακύμανση αυξάνει, το n , επίσης μεγαλώνει. Πληθυσμοί με μεγάλη μεταβλητότητα απαιτούν περισσότερες μετρήσεις (σε σύγκριση με πληθυσμούς με μικρή μεταβλητότητα) για να επιτευχθεί η επιθυμητή ακρίβεια. Επίσης όσο η εμπιστοσύνη της εκτίμησης αυξάνει (δηλ. το α μικραίνει) το n μεγαλώνει. Και καθώς το ω μικραίνει το n μεγαλώνει.

4.3.2 Συμπερασματολογία για το διωνυμικό p

Έστω ότι δίνονται n ανεξάρτητες δοκιμές τύπου Bernoulli με πιθανότητα επιτυχίας p και έστω X ο αριθμός των επιτυχιών. Με άλλα λόγια δίνεται ένα διωνυμικό πείραμα με τιμή X όπου η τυχαία μεταβλητή X έχει διωνυμική κατανομή $B(n, p)$.

Ο φυσιολογικός εκτιμητής $\hat{p}$ του p είναι η παρατηρηθείσα συχνότητα επιτυχίας, δηλαδή

$$\hat{p} = \frac{X}{n} \quad (4.17)$$

Ο εκτιμητής αυτός είναι αμερόληπτος διότι

$$E(\hat{p}) = E(X)/n = np/n = p$$

και έχει τυπικό σφάλμα

$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n} \quad (4.18)$$



που εκτιμάται με

$$\hat{\sigma}_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n} . \quad (4.19)$$

Το Κεντρικό Οριακό Θεώρημα μας δίνει ότι η κατανομή της τ.μ. χ είναι κατά προσέγγιση κανονική με μέση τιμή np και διακύμανση npq (βλ. εδάφιο 3.6). Η προσέγγιση βελτιώνεται όσο το n αυξάνει. Ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για το p κατασκευάζεται με τη βοήθεια της προσέγγισης αυτής:

$$L = \hat{p} - z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n} , \quad U = \hat{p} + z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n} . \quad (4.20)$$

Ο έλεγχος στατιστικών υποθέσεων για το p γίνεται με τη βοήθεια της στατιστικής συνάρτησης χ όπως αναπτύχθηκε στις § 4.2, 4.2.1 και 4.2.3. Το τεστ αυτό είναι ακριβές και γίνεται με τη χρήση λεπτομερών διωνυμικών πινάκων ή σημαντικής υπολογιστικής προσπάθειας. Ένα κατά προσέγγιση (για μεγάλα n) τεστ μπορεί να γίνει χρησιμοποιώντας το στατιστικό

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{\chi - np_0}{\sqrt{np_0(1-p_0)}} \quad (4.21)$$

το οποίο έχει τυπική κανονική κατανομή όταν $p = p_0$.

Έτσι για να ελέγξουμε υποθέσεις της μορφής

$$(a) \quad H_0: p \leq p_0 \quad \text{έναντι} \quad H_A: p > p_0$$

$$(b) \quad H_0: p \geq p_0 \quad \text{έναντι} \quad H_A: p < p_0$$

$$(c) \quad H_0: p = p_0 \quad \text{έναντι} \quad H_A: p \neq p_0$$



χρησιμοποιούμε το Z της (4.21) και τις ακόλουθες κρίσιμες περιοχές

για την (a) $C = (z_\alpha, +\infty)$, δηλ. $Z > z_\alpha$

για την (b) $C = (-\infty, -z_\alpha)$, δηλ. $Z < -z_\alpha$

για την (c) $C = (-\infty, -z_\alpha/2) \cup (z_\alpha/2, +\infty)$, δηλ. $|Z| > z_\alpha/2$

.

Παράδειγμα 4.4: Μια Ιατρική Ερευνητική Ομάδα (Coronary Drug Project, J. Chron, Dis., vol. 27, 267-285) ανέφερε το 1974 ότι από 2789 άνδρες ηλικίας 30-64, οι οποίοι είχαν μία ή περισσότερες καρδιακές προσβολές, υποβλήθηκαν σε θεραπεία και έγιναν καλά, 358 (δηλαδή 12.8%) πέθαναν κατά τη διάρκεια των τριών ετών που ακολούθησαν την θεραπεία. Να εκτιμηθεί η πιθανότητα θανάτου εντός των τριών ετών μετά τη θεραπεία.

Έστω ότι για κάθε δοκιμή, δηλαδή για κάθε καρδιοπαθή, επιτυχία είναι το άτομο να μην πεθάνει κατά τη διάρκεια των τριών ετών μετά τη θεραπεία. Έστω p η πιθανότητα αυτή. Τότε έχουμε $n=2789$ δοκιμές και

$$\hat{p} = \frac{2431}{2789} = 0.872, \quad 1 - \hat{p} = \frac{358}{2789} = 0.128.$$

Το $\hat{p}$ εκτιμά την πιθανότητα ζωής κατά την διάρκεια των τριών ετών και το $1 - \hat{p}$ εκτιμά την πιθανότητα θανάτου εντός των τριών ετών μετά τη θεραπεία ετών. Το τυπικό σφάλμα του $\hat{p}$ είναι

$$\hat{\sigma}_{\hat{p}} = \sqrt{(0.128)(0.872)/2789} = 0.006.$$

Ομοίως



$$\hat{\sigma}_{(1-\hat{p})} = 0.006.$$

Ένα κατά προσέγγιση 95% Δ.Ε. για το p δίδεται από την (4.20):

$$0.872 \pm 1.96(0.006) = 0.872 \pm 0.012,$$

δηλαδή $l = 86\%$ και $u = 88,4\%$.

Με όμοιο τρόπο ένα 95% Δ.Ε. για το $1-p$ είναι

$$0.128 \pm 1.96(0.006) = 0.128 \pm 0.012,$$

δηλαδή $l = 11.6\%$ και $u = 14\%$. Με άλλα λόγια, ο τριετής ρυθμός θανάτου είναι μεταξύ 11.6% και 14%.

Με βάση τα προηγούμενα νομίζετε ότι είναι λογικό να συμπεράνομε ότι ο ετήσιος ρυθμός θανάτου για την κατηγορία των παραπάνω ατόμων είναι μεταξύ 3.9% και 4.7%;

Η παραπάνω ανάλυση ισχύει υπό την προϋπόθεση ότι τα 2789 άτομα αποτελούν ένα τυχαίο δείγμα από τον (μεγάλο) πληθυσμό καρδιοπαθών μίας χώρας. Στην πραγματικότητα τα 2789 άτομα δεν ελήφθησαν με δειγματοληψία. Ελήφθησαν από μια παρεμφερή κλινική δοκιμή της οποίας ο σκοπός ήταν η σύγκριση διαφόρων φαρμάκων ελαττώσεως της χοληστερίνης με το αδρανές φάρμακο placebo ως προληπτικά φάρμακα καρδιακού νοσήματος. Τα 2789 άτομα πήραν το placebo και η μελέτη έγινε σε 53 νοσοκομεία των Ηνωμένων Πολιτειών της Αμερικής που έλαβαν μέρος στην έρευνα. Βέβαια πρέπει να γνωρίζουμε ότι τα άτομα αυτά είχαν συστηματική ιατρική παρακολούθηση με αποτέλεσμα να υπάρχει ένα μικρό ποσοστό μεροληψίας έναντι των τυχαίως εκλεγμένων καρδιοπαθών.

Έστω τώρα ότι θέλουμε να ελέγξουμε την υπόθεση $H_0: p=0.90$ έναντι $H_a: p \neq 0.90$. Η τιμή της σ.σ. του τεστ (4.21) είναι



$$z = \frac{0.872 - 0.90}{\sqrt{(0.90)(0.10)/2789}} = \frac{-0.028}{0.0057} = -4.91.$$

Η τιμή αυτή είναι μέσα στην κρίσιμη περιοχή μεγέθους 5%

$$C = (-\infty, -1.96) \cup (1.96, +\infty)$$

και έτσι με επίπεδο σημαντικότητας $\alpha=5\%$ η H_0 απορρίπτεται. Μαλιστα δε επειδή η πιθανότητα $P(|Z| > 4.91)$ είναι σχεδόν μηδέν η τιμή $z = 4.91$ είναι στατιστικώς πολύ σημαντική και έτσι υπάρχουν ισχυρές ενδείξεις για την απόρριψη της H_0 . Το αποτέλεσμα αυτό δεν μας ξενίζει διότι το $n=2789$ είναι αρκετά μεγάλο και το p εκτιμάται με το $\hat{p}$ με μεγάλη ακρίβεια.

Μέγεθος του δείγματος: Αν θέλουμε να εκτιμήσουμε το p με βαθμό εμπιστοσύνης $100(1-\alpha)\%$ και σφάλμα εκτίμησης εντός $\pm \omega$ από το Δ.Ε. (4.20) έχουμε

$$2\omega = U - L = 2z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n} \quad \text{ή} \quad n = \hat{p}(1-\hat{p}) \left(\frac{z_{\alpha/2}}{\omega} \right)^2$$

και επειδή $\hat{p}(1-\hat{p}) \leq 1/4$ για κάθε $\hat{p} \in (0,1)$

$$n \leq \frac{1}{4} \left(\frac{z_{\alpha/2}}{\omega} \right)^2.$$

4.3.3 Συμπερασματολογία για τη διαφορά δύο μέσων τιμών

Εδώ αναλύουμε δεδομένα που αποτελούνται από δύο ανεξάρτητα τυχαία δείγματα, δείγμα 1 από πληθυσμό 1 και δείγμα 2 από πληθυσμό 2. Συχνά το δείγμα 1 είναι δείγμα ελέγχου και το δείγμα 2 είναι δείγμα "θεραπείας" ή



“δοκιμασίας”. Άλλες φορές τα δείγματα προκύπτουν από πειράματα σύγκρισης των αποτελεσμάτων δύο διαφορετικών θεραπειών.

Στη Στατιστική θεραπείες είναι οποιεσδήποτε διαδικασίες ή μέθοδοι ή ουσίες τα αποτελέσματα των οποίων θέλουμε να εκτιμήσουμε ή συγκρίνουμε. Θεραπείες μπορεί να είναι διαφορετικά φάρμακα ή διαφορετικές δόσεις ενός φαρμάκου διαφορετικές θεραπευτικές αγωγές στην ιατρική, διαφορετικές χημικές ουσίες ή διαφορετικά ποσοστά εφαρμογής μίας χημικής ουσίας, διαφορετικές μηχανές ή διαφορετικοί χειριστές κ.λ.π.

Η σύγκριση δύο θεραπειών γίνεται εφαρμόζοντας τις σε ομοειδές πειραματικό υλικό. Διαιρούμε το πειραματικό υλικό σε πειραματικές μονάδες και κατόπιν εφαρμόζουμε στην τύχη μία από τις δύο θεραπείες σε κάθε πειραματική μονάδα. Αυτό γίνεται είτε με ένα νόμισμα (κορώνα-γράμματα) είτε με τη βοήθεια πινάκων τυχαίων αριθμών. Μερικές φορές επιβάλλουμε και τον δευτερεύοντα περιορισμό κάθε θεραπεία να εφαρμόζεται στον ίδιο αριθμό πειραματικών μονάδων. Σχέδια πειραμάτων της παραπάνω μορφής λέγονται πλήρως τυχαιοποιημένα σχέδια.

Αφού εφαρμοστούν οι θεραπείες παίρνουμε ένα αντικειμενικό μέτρο του αποτελέσματος κάθε θεραπείας από κάθε μονάδα και έτσι καταλήγουμε σε δυο ανεξάρτητα δείγματα από δύο πληθυσμούς, τις δύο θεραπείες.

Στόχος είναι να εκτιμήσουμε τη διαφορά $\mu_1 - \mu_2$ ή να ελέγξουμε την υπόθεση $H_0: \mu_1 = \mu_2$, όπου μ_1 και μ_2 είναι οι μέσες τιμές των πληθυσμών (θεραπειών) 1 και 2 αντίστοιχα.

Έστω $X_{11}, \dots, X_{1n_1}$ τα δεδομένα (τυχαίο δείγμα μεγέθους n_1) από τον πληθυσμό 1 και $X_{21}, \dots, X_{2n_2}$ τα δεδομένα (τυχαίο δείγμα μεγέθους n_2) από τον πληθυσμό 2. Οι πληροφορίες που περιέχονται στα δεδομένα αυτά συνοψίζονται περιληπτικά στον ακόλουθο πίνακα:



Δείγμα	Μέγεθος	Βαθμ. ελευθ. (β.ε.)	Μέση τιμή	Άθροισμ. τετρ. (SS)
1	n_1	$n_1 - 1$	$\bar{X}_1$	$\sum (X_{1i} - \bar{X}_1)^2$
2	n_2	$n_2 - 1$	$\bar{X}_2$	$\sum (X_{2i} - \bar{X}_2)^2$
Σύνολο		$n_1 + n_2 - 2$		Συνολικό SS

Πίνακας 4.1 Υπολογισμοί δύο δειγμάτων

Υπενθυμίζουμε τους τύπους μηχανής για τα αθροίσματα τετραγώνων και τις διακυμάνσεις:

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^n (X_{1i} - \bar{X}_1)^2 = \frac{1}{n_1 - 1} \left(\sum X_{1i}^2 - \frac{(\sum X_{1i})^2}{n_1} \right) \quad (4.22)$$

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^n (X_{2i} - \bar{X}_2)^2 = \frac{1}{n_2 - 1} \left(\sum X_{2i}^2 - \frac{(\sum X_{2i})^2}{n_2} \right)$$

Ο καλύτερος εκτιμητής της διαφοράς $\mu_1 - \mu_2$ είναι ο

$$\hat{\mu}_1 - \hat{\mu}_2 = \bar{X}_1 - \bar{X}_2. \quad (4.23)$$

Επειδή τα τυχαία δείγματα είναι ανεξάρτητα, οι στατιστικές συναρτήσεις $\bar{X}_1$ και $\bar{X}_2$ είναι ανεξάρτητες. Έτσι η διακύμανση του $\bar{X}_1 - \bar{X}_2$ είναι

$$\text{Var}(\bar{X}_1 - \bar{X}_2) = \text{Var}\bar{X}_1 + \text{Var}\bar{X}_2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \quad (4.24)$$

Από τη σχέση αυτή βρίσκουμε ότι το τυπικό σφάλμα του εκτιμητή $\bar{X}_1 - \bar{X}_2$ της διαφοράς $\mu_1 - \mu_2$ είναι



$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} . \quad (4.25)$$

Έτσι το εκτιμώμενο τυπικό σφάλμα του $\bar{X}_1 - \bar{X}_2$ είναι

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} . \quad (4.26)$$

Για να προχωρήσουμε τώρα στα διαστήματα εμπιστοσύνης και τα τεστ για τη διαφορά $\mu_1 - \mu_2$ χρειάζεται να κάνουμε μερικές υποθέσεις για τις κατανομές των $\bar{X}_1$ και $\bar{X}_2$. Αν λοιπόν τα δείγματα 1 και 2 είναι μεγάλα τότε ισχύουν κατά προσέγγιση οι σχέσεις

$$\bar{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1}) \quad \text{και} \quad \bar{X}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2}) . \quad (4.27)$$

Έτσι στην περίπτωση αυτή ένα $100(1-\alpha)\%$ κατά προσέγγιση διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_2$ δίδεται από τις σχέσεις

$$L = \bar{X}_1 - \bar{X}_2 - z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (4.28)$$

$$U = \bar{X}_1 - \bar{X}_2 + z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} .$$

Ομοίως, στην περίπτωση αυτή, υποθέσεις της μορφής



- (a) $H_0: \mu_1 - \mu_2 \leq \delta_0$ έναντι $H_a: \mu_1 - \mu_2 > \delta_0$
- (b) $H_0: \mu_1 - \mu_2 \geq \delta_0$ έναντι $H_a: \mu_1 - \mu_2 < \delta_0$ (4.29)
- (c) $H_0: \mu_1 - \mu_2 = \delta_0$ έναντι $H_a: \mu_1 - \mu_2 \neq \delta_0$

ελέγχονται κατά προσέγγιση με το στατιστικό

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (4.30)$$

το οποίο έχει κατά προσέγγιση τυπική κανονική κατανομή όταν ισχύει η H_0 .

Οι αντίστοιχες κρίσιμες περιοχές είναι οι ακόλουθες

για την (a) $C = (z_\alpha, +\infty)$, δηλ. $Z > z_\alpha$

για την (b) $C = (-\infty, -z_\alpha)$, δηλ. $Z < -z_\alpha$

για την (c) $C = (-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, +\infty)$, δηλ. $|Z| > z_{\alpha/2}$.

Όταν τα μεγέθη n_1 και n_2 των δειγμάτων 1 και 2 είναι μικρά ή προηγούμενη συμπερασματολογία όσον αφορά τα διαστήματα εμπιστοσύνης και τεστ δεν ισχύει. Τα διαστήματα (4.28) και τεστ (4.30) δεν είναι αξιόπιστα. Για να προχωρήσουμε χρειάζονται περισσότερες υποθέσεις για τις κατανομές των πληθυσμών 1 και 2 και τις διακυμάνσεις τους. Χρειάζεται να υποθέσουμε ότι οι πληθυσμοί είναι κανονικοί και έχουν ίσες διακυμάνσεις.

Κανονικοί πληθυσμοί - ίσες διακυμάνσεις: Έστω λοιπόν ότι οι δύο πληθυσμοί είναι κανονικοί και ότι οι διακυμάνσεις τους σ_1^2 και σ_2^2 είναι ίσες



και άγνωστες, δηλαδή $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Όταν το πειραματικό υλικό (πειραματικές μοναδες) είναι ομοειδές και ελέγχονται όλοι οι παράγοντες που είναι δυνατόν να επηρεάζουν τις μετρήσεις (εκτός από τις θεραπείες) τότε η υπόθεση της ισότητας των διακυμάνσεων είναι ρεαλιστική. Βέβαια η υπόθεση αυτή μπορεί να ελεγχθεί και στατιστικώς (βλέπε τεστ ισότητας των διακυμάνσεων).

Ο καλύτερος εκτιμητής της κοινής διακύμανσης σ^2 είναι ο

$$s^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{\text{συνολικό SS}}{\text{συνολικοί β.ε.}} \quad (4.31)$$

έχει n_1+n_2-2 βαθμούς ελευθερίας και λέγεται αθροιστικός εκτιμητής της διακύμανσης σ^2 . Έτσι το εκτιμώμενο τυπικό σφάλμα του εκτιμητή $\bar{X}_1 - \bar{X}_2$ της διαφοράς είναι

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \quad (4.32)$$

Τα εκτιμώμενα τυπικά σφάλματα των εκτιμητών $\hat{\mu}_1 = \bar{X}_1$ και $\hat{\mu}_2 = \bar{X}_2$ είναι

$$\hat{\sigma}_{\bar{X}_1} = s/\sqrt{n_1}, \quad \hat{\sigma}_{\bar{X}_2} = s/\sqrt{n_2}. \quad (4.33)$$

Με τις παραπάνω υποθέσεις το στατιστικό

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (4.34)$$

έχει κατανομή $t_{n_1+n_2-2}$ για κάθε μ_1, μ_2 και σ^2 .

Έτσι το καλύτερο $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_2$ δίνεται από τις σχέσεις



$$L = \bar{X}_1 - \bar{X}_2 - t_{\alpha/2, n_1+n_2-2} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (4.35)$$

$$U = \bar{X}_1 - \bar{X}_2 + t_{\alpha/2, n_1+n_2-2} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} .$$

Οι υποθέσεις (a), (b), (c) της (4.29) για τη διαφορά $\mu_1 - \mu_2$ ελέγχονται με τη στατιστική συνάρτηση

$$t = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (4.36)$$

η οποία έχει κατανομή $t_{n_1+n_2-2}$ όταν ισχύει η H_0 .

Οι αντίστοιχες κρίσιμες περιοχές είναι οι ακόλουθες:

για την (a) $C = (t_{\alpha}, +\infty)$, δηλ. $t > t_{\alpha}$

για την (b) $C = (-\infty, -t_{\alpha})$, δηλ. $t < -t_{\alpha}$

για την (c) $C = (-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, +\infty)$, δηλ. $|t| > t_{\alpha/2}$.

Τα διαστήματα εμπιστοσύνης και τα τεστ για μία από τις μέσες τιμές μ_i χρησιμοποιούν τον αθροιστικό εκτιμητή S του σ και οι ποσότητες $t_{\alpha/2}$ και t έχουν n_1+n_2-2 βαθμούς ελευθερίας.

Παράδειγμα 4.5: Δύο ανεξάρτητες ομάδες παιδιών 1 και 2 αποτελούνται από 10 παιδιά η κάθε μια. Τα παιδιά της ομάδας 1 προέρχονται από υπερτασικούς γονείς ενώ τα παιδιά της ομάδας 2 προέρχονται από γονείς με κανονική πίεση. Η συστολική πίεση του αίματος μετريέται στα παιδιά των δύο ομάδων με τα ακόλουθα αποτελέσματα:



132 Κεφ. 4 Στατιστική Συμπερασματολογία

Ομάδα 1: 100, 102, 96, 106, 110, 110, 120, 112, 112, 90
 Ομάδα 2: 104, 88, 100, 98, 102, 92, 96, 100, 96, 96.

Με επίπεδο σημαντικότητας $\alpha=10\%$ να ελεγχθεί αν υπάρχει διαφορά μεταξύ της μέσης πίεσης μ_1 παιδιών από υπερτασικούς γονείς και της μέσης φυσιολογικής πίεσης μ_2 . Υποθέτουμε ότι η μεταβλητότητα των πιέσεων είναι η ίδια και για τους δύο πληθυσμούς, παιδιά υπερτασικών γονέων και παιδιά φυσιολογικών γονέων.

Για τα παραπάνω δεδομένα έχουμε

Δείγμα 1

$$\sum x_{1i} = 1058$$

$$\bar{x}_1 = 105.8$$

$$\sum x_{1i}^2 = 112644$$

Δείγμα 2

$$\sum x_{2i} = 972$$

$$\bar{x}_2 = 97.2$$

$$\sum x_{2i}^2 = 94680$$

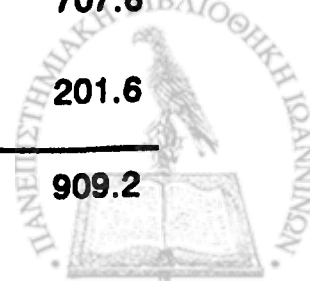
Έτσι

$$\sum (x_{1i} - \bar{x}_1)^2 = \sum x_{1i}^2 - \frac{(\sum x_{1i})^2}{n_1} = 112644 - \frac{(1058)^2}{10} = 707.6$$

$$\sum (x_{2i} - \bar{x}_2)^2 = 94680 - \frac{(972)^2}{10} = 201.6$$

και ο Πίνακας 4.1. είναι

Δείγμα	n	β.ε.	μέση τιμή	SS
1	10	9	105.8	707.6
2	10	9	97.2	201.6
Σύνολο		18		909.2



Η μέση διαφορά πιέσεων $\mu_1 - \mu_2$ εκτιμάται με την ποσότητα

$$\hat{\mu}_1 - \hat{\mu}_2 = \bar{X}_1 - \bar{X}_2 = 8.6 .$$

Επειδή οι πληθυσμοί είναι κανονικοί και έχουν την ίδια μεταβλητότητα $\sigma_1^2 = \sigma_2^2 = \sigma^2$, ο αθροιστικός εκτιμητής του σ^2 είναι (βλ. (4.31))

$$s^2 = \frac{909.2}{18} = 50.51 , \quad (S=7.11) .$$

Η αξιοπιστία του εκτιμητή $\bar{X}_1 - \bar{X}_2$ ζυγίζεται με το τυπικό του σφάλμα (βλ. (4.32)).

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{50.51} \sqrt{\frac{1}{10} + \frac{1}{10}} = 3.18 ,$$

το οποίο είναι σχετικά μικρό.

Η υπόθεση που θέλουμε να ελέγξουμε με επίπεδο σημαντικότητας 10% είναι (βλ. (4.29) (c) με $\delta_0=0$)

$$H_0: \mu_1 = \mu_2 \quad \text{έναντι} \quad H_a: \mu_1 \neq \mu_2$$

Έτσι το στατιστικό t με $\delta_0=0$ (βλ. (4.36)) έχει 18 β.ε. και είναι

$$t = \frac{105.8 - 97.2}{7.11 \sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{8.6}{3.18} = 2.70 .$$

Η κρίσιμη περιοχή για $\alpha=10\%$ είναι

$$|t| > t_{0.05, 18} = 1.734 .$$

Επειδή $2.70 > 1.734$ απορρίπτεται η H_0 και εξάγεται το



συμπέρασμα ότι υπάρχει στατιστικώς σημαντική διαφορά μεταξύ της μέσης πίεσης αίματος παιδιών από υπερτασικούς γονείς και της μέσης φυσιολογικής πίεσης. Το ποσοστό σφάλματος είναι 10%.

Το μέγεθος των δειγμάτων: Η συζήτηση ακολουθεί το σκεπτικό του προσδιορισμού του μεγέθους του δείγματος που αναπτύχθηκε στην περίπτωση του ενός πληθυσμού και ενός δείγματος (βλ. § 4.3.1). Το ενδιαφέρον μας είναι για τη παράμετρο $\delta = \mu_1 - \mu_2$. Υποθέτουμε ότι οι διακυμάνσεις των δύο πληθυσμών είναι ίσες, δηλαδή $\sigma_1^2 = \sigma_2^2 = \sigma^2$, και ότι ο ερευνητής γνωρίζει ή μπορεί να καθορίσει μία λογική τιμή του σ . Υπάρχουν πολλές πειραματικές περιπτώσεις που οι παραπάνω προϋποθέσεις ικανοποιούνται.

Έστω ότι θέλουμε να εκτιμήσουμε τη διαφορά $\delta = \mu_1 - \mu_2$ κατά τέτοιο τρόπο ώστε η δειγματική διαφορά $\bar{X}_1 - \bar{X}_2$ να ευρίσκεται εντός ω μονάδων από το δ με γνωστή εμπιστοσύνη $100(1-\alpha)\%$. Με άλλα λόγια θέλουμε το μήκος του Δ.Ε. (4.35) να είναι 2ω . Έτσι

$$U - L = 2t_{\alpha/2, n_1+n_2-2} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 2\omega.$$

Υπάρχουν πολλοί συνδιασμοί των n_1 και n_2 που ικανοποιούν την παραπάνω σχέση. Θέτοντας $S = \sigma$ και υποθέτοντες ότι τα 2 δείγματα είναι μεγάλα και ίσα, δηλαδή $n_1 = n_2 = n$, και $t_{\alpha/2, n_1+n_2-2} = z_{\alpha/2}$ η προηγούμενη σχέση γίνεται

$$z_{\alpha/2} \sigma \sqrt{\frac{1}{n} + \frac{1}{n}} = \omega$$

και τελικά λύνοντας την ως προς n έχουμε:



$$n = \frac{2z_{\alpha/2}^2 \sigma^2}{\omega^2}$$

n μετρήσεις εκλέγονται από τον πληθυσμό 1 και n από τον πληθυσμό 2.

4.3.4 Συγκρίσεις κατά ζεύγη

Οι συγκρίσεις κατά ζεύγη αναφέρονται σε δεδομένα που λαμβάνονται κατά ζεύγη. Το Παράδειγμα 4.3. είναι μία τέτοια περίπτωση. Τα δεδομένα των συγκρίσεων κατά ζεύγη εμφανίζονται συνήθως σε πειράματα των οποίων ο σκοπός είναι η διερεύνηση της αποτελεσματικότητας μίας θεραπείας. Ζεύγη ατόμων ή πειραματοζώων ή πιο γενικά πειραματικών μονάδων σχηματίζονται με βάση διαφόρους παράγοντες όπως π.χ. η ηλικία, το βάρος, το φύλο, η οξύτητα της ασθένειας, η χημική σύνθεση, η κληρονομικότητα (δίδυμα), το περιβάλλον κ.λ.π., δηλ. παράγοντες που επηρεάζουν τις μετρήσεις. Σε ένα μέλος κάθε ζεύγους εφαρμόζεται η θεραπεία ενώ το άλλο μέλος χρησιμεύει για έλεγχο (control). Μετά λαμβάνονται οι μετρήσεις. Πιο φυσιολογικά, συγκρίσεις κατά ζεύγη εφαρμόζονται όταν η θεραπεία εφαρμόζεται σε κάθε ένα από n άτομα ή πειραματικές μονάδες και λαμβάνονται δύο μετρήσεις, μία πριν από τη θεραπεία και μία μετά τη θεραπεία. Το μέλος του ζεύγους στο οποίο εφαρμόζεται η θεραπεία εκλέγεται στην τύχη. Σχέδια πειραμάτων με τα παραπάνω χαρακτηριστικά λέγονται και σχέδια τυχαιοποιημένων τετραγώνων (randomized block designs). Σε κάθε τετράγωνο έχουμε δύο πειραματικές μονάδες.

Η παράμετρος που μας ενδιαφέρει εδώ είναι η $\delta = \mu_1 - \mu_2$ όπου μ_1 είναι η μέση τιμή της θεραπείας και μ_2 η μέση φυσιολογική τιμή (το control). Η συμπερασματολογία γίνεται με τον ίδιο τρόπο όπως στην § 4.3.1, περίπτωση μίας μέσης τιμής ενός πληθυσμού και ενός δείγματος. Το δείγμα είναι οι διαφορές $D_i = X_{1i} - X_{2i}$, $i = 1, \dots, n$ όπου $X_{11}, \dots, X_{1n}$ είναι οι μετρήσεις θεραπείας και



$X_{21}, \dots, X_{2n}$ οι μετρήσεις control. Βλέπε Παράδειγμα 4.3.
Η διαφορά $\delta = \mu_1 - \mu_2$ εκτιμάται αμερόληπτα με το

$$\bar{D} = \bar{X}_1 - \bar{X}_2 . \quad (4.38)$$

Το εκτιμώμενο τυπικό σφάλμα του $\bar{D}$ είναι

$$\hat{\sigma}_{\bar{D}} = S_D / \sqrt{n} , \quad (4.39)$$

όπου

$$S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2 .$$

Υποθέτοντας ότι οι διαφορές $D_1, \dots, D_n$ αποτελούν ένα τυχαίο δείγμα από ένα κανονικό πληθυσμό, ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για τη διαφορά δ είναι το

$$L = \bar{D} - t_{\alpha/2, n-1} S_D / \sqrt{n} \quad (4.40)$$

$$U = \bar{D} + t_{\alpha/2, n-1} S_D / \sqrt{n} .$$

Υποθέσεις για το δ π.χ., $H_0: \delta = \delta_0$ ελέγχονται με το στατιστικό

$$t = \frac{\bar{D} - \delta_0}{S_D / \sqrt{n}} \quad (4.41)$$

το οποίο έχει κατανομή t_{n-1} όταν αληθεύει η H_0 .

Με συγκρίσεις κατά ζεύγη μπορούμε να συγκρίνουμε και δύο διαφορετικές θεραπείες. Στα πλήρως τυχαιοποιημένα πειράματα σύγκρισης δύο θεραπειών (βλ. § 4.3.1), οι θεραπείες εφαρμόζονται στην τύχη πάνω σε όλες τις πειραματικές μονάδες. Συνήθως η τυχαιοποίηση περιορίζεται έτσι ώστε κάθε θεραπεία να εφαρμόζεται στον ίδιο αριθμό πειραματικών μονάδων (ίσες επαναλήψεις κάθε

θεραπείας). Τα δεδομένα πειραμάτων της μορφής αυτής αναλύονται με τη μεθοδολογία των δύο δειγμάτων - δύο πληθυσμών (βλ. § 4.3.1).

Για τα παραπάνω πειράματα η αποτελεσματικότητα και ακρίβεια της εκτίμησης είναι αντιστρόφως ανάλογες προς τον αθροιστικό εκτιμητή S^2 της κοινής διακύμανσης σ^2 . Όσο μικρότερο το S^2 τόσο μεγαλύτερη η ακρίβεια. Για να αυξήσουμε την ακρίβεια ενός πειράματος, με ένα συγκεκριμένο ποσό πειραματικού υλικού, πρέπει να αλλάξουμε το σχέδιο του πειράματος για να ελαττώσουμε την ανεξήγητη μεταβλητότητα των δεδομένων καθώς αυτή μετριέται με το S^2 . Επειδή το S^2 προκύπτει από τη μεταβλητότητα μεταξύ πειραματικών μονάδων που δέχονται την ίδια θεραπεία, είτε πρέπει να ελαττώσουμε τη μεταβλητότητα αυτή χρησιμοποιώντας περισσότερο ομογενές υλικό, είτε πρέπει να απαλείψουμε μερική από τη μεταβλητότητα αυτή χρησιμοποιώντας εκ των προτέρων πληροφορίες για τις μετρήσεις (αποκρίσεις) που θα προκύψουν από τις πειραματικές μονάδες. Οι ιδέες αυτές υλοποιούνται με τα σχέδια τυχαιοποιημένων τετραγώνων.

Αν οι πειραματικές μονάδες εμφανίζονται κατά ζεύγη ή μπορούν να ομαδοποιηθούν σε ζεύγη έτσι ώστε η μεταβλητότητα των αποκρίσεων μεταξύ των μελών κάθε ζεύγους να είναι μικρότερη από τη μεταβλητότητα μεταξύ μελών διαφόρων ζευγών, μπορούμε να βελτιώσουμε την αποτελεσματικότητα του πειράματος μας εφαρμόζοντας τις δύο θεραπείες τυχαία στα δύο μέλη κάθε ζεύγους. Κάθε μία από τις δύο θεραπείες εφαρμόζεται σε ένα μέλος κάθε ζεύγους. Έτσι εκτιμάται η διαφορά στην επίδοση των δύο θεραπειών για κάθε ζεύγος, η μεταβλητότητα μεταξύ των ζευγών δεν υπεισέρχεται. Αν η τελευταία μεταβλητότητα είναι μεγαλύτερη από τη μεταβλητότητα εκτός των ζευγών, το πείραμα τυχαιοποιημένων τετραγώνων θα είναι καλύτερο (αποτελεσματικότερο για τη σύγκριση των δύο θεραπειών) από το πλήρως τυχαιοποιημένο πείραμα.

Παράδειγμα 4.6: Για να προσδιοριστεί αν μία επεξεργασία θερμότητας είναι αποτελεσματική στο να ελαττώνει ή όχι



τον αριθμό των βακτηριδίων του γάλακτος παίρνονται 10 δείγματα γάλακτος και για κάθε δείγμα γίνεται μικροσκοπική καταμέτρηση των βακτηριδίων πριν και μετά την επεξεργασία. Τα δεδομένα είναι οι λογάριθμοι των αριθμών των βακτηριδίων.

Λογάριθμοι αριθμών βακτηριδίων

Δείγμα	Πριν από την επεξεργασία (X_1)	Μετά την επεξεργασία (X_2)	Διαφορές $D_1 = X_{11} - X_{21}$
1	6.88	6.85	0.03
2	7.05	6.84	0.21
3	8.24	7.07	1.17
4	5.20	5.05	0.15
5	6.16	6.18	-0.02
6	6.67	6.71	-0.04
7	7.01	6.49	0.52
8	5.46	5.24	0.22
9	5.87	5.88	-0.01
10	6.54	6.41	0.13

Για τα παραπάνω δεδομένα έχουμε

$$\sum D_1 = 2.36, \quad \bar{D} = 0.236$$

$$\sum D_1^2 = 1.7742, \quad (\sum D_1)^2/n = 0.5570$$

$$S_D^2 = (1.7742 - 0.5570)/9 = 0.1352$$

$$.1352/10 = .01352$$

Θέλουμε να ελέγξουμε την υπόθεση μη αποτελεσματικότητας της επεξεργασίας $H_0: \delta = \mu_1 - \mu_2 = 0$ έναντι της



υποθέσεως ελαττώσεως του αριθμού των βακτηριδίων $H_a: \delta > 0$.

Η σ.σ. του τεστ

$$t = \frac{\bar{D} - 0}{S_D / \sqrt{n}}$$

έχει τιμή

$$t = \frac{0.236}{\sqrt{0.0135}} = 2.0312$$

και επειδή $t = 2.0312 > t_{0.05,9} = 1.833$ απορρίπτουμε την H_0 με επίπεδο σημαντικότητας 5% και βγάζουμε το συμπέρασμα ότι η επεξεργασία θερμάνσεως ελαττώνει τον αριθμό των βακτηριδίων.

4.3.5 Συμπερασματολογία για τις διακυμάνσεις

α) Μία διακύμανση

Έστω $X_1, X_2, \dots, X_n$ ένα τυχαίο δείγμα από κάποιο πληθυσμό με μέση τιμή μ και διακύμανση σ^2 . Έχουμε δει στα προηγούμενα ότι ο καλύτερος (από πλευράς αμεροληψίας) εκτιμητής του σ^2 είναι η δειγματική διακύμανση

$$S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2 = \frac{1}{n-1} \left(\sum X_i^2 - \frac{(\sum X_i)^2}{n} \right).$$

Ομοίως ο εκτιμητής του σ είναι το S .

Έστω τώρα ότι ο πληθυσμός είναι κανονικός. Το καλύτερο $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για το σ^2 είναι το



$$L = \frac{(n-1)S^2}{\chi_{\alpha/2, n-1}^2}, \quad U = \frac{(n-1)S^2}{\chi_{1-\alpha/2, n-1}^2} \quad (4.42)$$

Το αντίστοιχο διάστημα για την τυπική απόκλιση σ είναι το

$$\left(\sqrt{(n-1)S^2 / \chi_{\alpha/2, n-1}^2}, \sqrt{(n-1)S^2 / \chi_{1-\alpha/2, n-1}^2} \right).$$

Τα διαστήματα αυτά απορρέουν από το γεγονός ότι η ποσότητα

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

για κάθε μ και σ (βλ. Θεώρημα 3.2).

Για να ελέγξουμε υποθέσεις της μορφής

$$(a) \quad H_0: \sigma^2 \leq \sigma_0^2 \quad (\text{ή } \sigma \leq \sigma_0) \quad \text{έναντι} \quad H_a: \sigma^2 > \sigma_0^2 \quad (\text{ή } \sigma > \sigma_0)$$

$$(b) \quad H_0: \sigma^2 \geq \sigma_0^2 \quad (\text{ή } \sigma \geq \sigma_0) \quad \text{έναντι} \quad H_a: \sigma^2 < \sigma_0^2 \quad (\text{ή } \sigma < \sigma_0)$$

$$(c) \quad H_0: \sigma^2 = \sigma_0^2 \quad (\text{ή } \sigma = \sigma_0) \quad \text{έναντι} \quad H_a: \sigma^2 \neq \sigma_0^2 \quad (\text{ή } \sigma \neq \sigma_0)$$

με επίπεδο σημαντικότητας α χρησιμοποιούμε το στατιστικό

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \quad (4.43)$$

το οποίο έχει χ_{n-1}^2 κατανομή όταν $\sigma^2 = \sigma_0^2$.



και τις ακόλουθες κρίσιμες περιοχές:

για την (a) $C = (\chi^2_{\alpha, n-1}, +\infty)$, δηλ. $\chi^2 > \chi^2_{\alpha, n-1}$

για την (b) $C = (0, \chi^2_{1-\alpha, n-1})$, δηλ. $\chi^2 < \chi^2_{1-\alpha, n-1}$

για την (c) $C = (0, \chi^2_{1-\alpha/2, n-1}) \cup (\chi^2_{\alpha/2, n-1}, +\infty)$.

Το τεστ αυτό λέγεται χ^2 (χι - τετράγωνο).

Παράδειγμα 4.7: Ο κατασκευαστής ενός οργάνου ακριβείας ισχυρίζεται ότι η τυπική απόκλιση των μετρήσεων που γίνονται με το όργανο είναι $\sigma=2$. Αν σε ένα πείραμα πάρουμε τις μετρήσεις 4.2, 5.3, 10.3 αληθεύει ο ισχυρισμός του κατασκευαστή; ($\alpha=5\%$).

Για τα δεδομένα του παραδείγματος έχουμε

$$\sum x_i = 19.8 \quad , \quad \sum x_i^2 = 151.82 \quad , \quad n=3$$

$$\sum (x_i - \bar{x})^2 = 151.82 - (19.8)^2/3 = 21.14$$

$$s^2 = \sum (x_i - \bar{x})^2 / (n-1) = 21.14 / 2 = 10.57.$$

Θέλουμε να ελέγξουμε την

$$H_0: \sigma = 2 \quad \text{έναντι} \quad H_a: \sigma \neq 2 .$$

Η σ.σ. του τεστ είναι



$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{2 \times 10.57}{4} = 5.285$$

και

$$\chi_{0.025,2}^2 = 7.378, \quad \chi_{0.975,2}^2 = 0.0506.$$

Έτσι επειδή

$$\chi_{0.975,2}^2 = 0.0506 < \chi^2 = 5.285 < \chi_{0.025,2}^2 = 7.378$$

τα δεδομένα δεν μας δίνουν αρκετές ενδείξεις για να απορρίψουμε τον ισχυρισμό του κατασκευαστή με $\alpha=5\%$. Ένα 95% διάστημα εμπιστοσύνης για το σ είναι

$$L = \sqrt{\frac{(n-1)S^2}{\chi_{0.025,2}^2}} = \sqrt{\frac{21.14}{7.378}} = \sqrt{2.87} = 1.69$$

$$U = \sqrt{\frac{(n-1)S^2}{\chi_{0.975,2}^2}} = \sqrt{\frac{21.14}{0.0506}} = \sqrt{417.79} = 20.44$$

β) Σύγκριση δύο διακυμάνσεων

Παράδειγμα 4.8 : Δύο διαφορετικά ηλεκτρονικά όργανα Α και Β, χρησιμοποιούνται για τη μέτρηση της πίεσης του ματιού. Ενδιαφερόμαστε να συγκρίνουμε τα όργανα αυτά ως προς την ποιότητα τους. Η ποιότητα ενός οργάνου είναι μεγάλη όταν επανειλημμένες μετρήσεις στο ίδιο αντικείμενο έχουν μικρή μεταβλητότητα. Έτσι επανειλημμένες μετρήσεις στο ίδιο μάτι και την ίδια περίπου χρονική στιγμή δίνουν τα ακόλουθα αποτελέσματα.



Με το όργανο	A	B
Αριθμός μετρήσεων	$n_1 = 10$	$n_2 = 8$
Διακύμανση	$S_1^2 = 1.25$	$S_2^2 = 0.28$

Γενικά: Έστω δύο ανεξάρτητα τυχαία δείγματα μεγέθους n_1 και n_2 από δύο κανονικούς πληθυσμούς με μέσες τιμές μ_1, μ_2 και διακυμάνσεις σ_1^2, σ_2^2 αντίστοιχα. Έστω $\bar{X}_1, S_1^2$ και $\bar{X}_2, S_2^2$ οι μέσες τιμές και διακυμάνσεις των δύο δειγμάτων. Ενδιαφερόμαστε να συγκρίνουμε τις δύο διακυμάνσεις. Η μηδενική υπόθεση είναι $H_0: \sigma_1^2 = \sigma_2^2$. Σε άλλες εφαρμογές η υπόθεση αυτή δίδει ίση μεταβλητότητα σε δύο μεθόδους μετρήσεων κ.λ.π.

Η στατιστική συνάρτηση του τεστ είναι

$$F = S_1^2 / S_2^2 \sim F_{n_1-1, n_2-1} \quad \text{όταν ισχύει η } H_0. \quad (4.44)$$

(Βλέπε Θεώρημα 3.4). Το τεστ αυτό λέγεται **F τεστ** και εφαρμόζεται για τις υποθέσεις της ακόλουθης μορφής

$$(a) \quad H_0: \sigma_1^2 \leq \sigma_2^2 \quad \text{έναντι} \quad H_a: \sigma_1^2 > \sigma_2^2$$

$$(b) \quad H_0: \sigma_1^2 \geq \sigma_2^2 \quad \text{έναντι} \quad H_a: \sigma_1^2 < \sigma_2^2$$

$$(c) \quad H_0: \sigma_1^2 = \sigma_2^2 \quad \text{έναντι} \quad H_a: \sigma_1^2 \neq \sigma_2^2$$

με τις ακόλουθες κρίσιμες περιοχές:

$$\text{για την (a)} \quad C = (F_{\alpha, n_1-1, n_2-1}, +\infty)$$

$$\text{για την (b)} \quad C = (0, F_{1-\alpha, n_1-1, n_2-1})$$

$$\text{για την (c)} \quad C = (0, F_{1-\alpha/2, n_1-1, n_2-1}) \cup (F_{\alpha/2, n_1-1, n_2-1}, +\infty).$$



Λύση του Παραδείγματος: Εδώ έχουμε να ελέγξουμε την $H_0: \sigma_1^2 = \sigma_2^2$ έναντι $H_a: \sigma_1^2 \neq \sigma_2^2$. Εστω $\alpha = 10\%$. Η σ.σ. του τεστ

$$F = S_1^2 / S_2^2$$

έχει τιμή $F = 1.25 / 0.28 = 4.46$. Η κρίσιμη περιοχή μεγέθους 10% είναι $C = (0, F_{0.95,9,7}) \cup (F_{0.05,9,7}, +\infty)$. Επειδή

$$F_{1-\alpha, n_1, n_2} = 1 / F_{\alpha, n_2, n_1}$$

$$F_{0.95,9,7} = 1 / F_{0.05,7,9} = 1 / 3.29 = 0.30 \text{ και } F_{0.05,9,7} = 3.68.$$

Επειδή τώρα $F = 4.46 > 3.68$ απορρίπτουμε την H_0 με επίπεδο σημαντικότητας 10%. Δεν υπάρχει αμφιβολία ότι το όργανο Β είναι καλύτερο από το Α. Αυτό είναι φανερό και από τις τιμές των S_1^2 και S_2^2 που είναι διαφορετικές. Δεν αρκεί όμως αυτό, διότι αν π.χ. $S_2^2 = 0.35$ το αντικείμενο δεν είναι στατιστικώς σημαντικό και η διαφορά μεταξύ S_1^2 και S_2^2 οφείλεται στην τύχη. Το ποσοστό σφάλματος είναι 10%.

4.3.6 Συμπερασματολογία για τη διαφορά δύο ποσοστών (πιθανοτήτων επιτυχίας)

α) Σύγκριση δυο ποσοστών κατά ζεύγη

Παράδειγμα 4.9: Το κάθε ένα από 60 δείγματα ρινικών εκκρινμάτων καλλιεργείται σε δύο διαφορετικά μέσα Α και Β με στόχο τη σύγκριση της ικανότητας των δύο μέσων να ανιχνεύσουν ορισμένους ιούς γρίπης. Τα αποτελέσματα δίνονται στον ακόλουθο πίνακα όπου + σημαίνει ότι γίνεται ανίχνευση του ιού της γρίπης και - ότι δεν γίνεται.

Κατανομή 60 ρινικών εκκρίσεων βάσει αποτελεσμάτων
καλλιέργεια σε δύο μέσα

Μέσον Α	+	+	-	-	
Μέσον Β	+	-	+	-	
Αριθμ. πτυέλων	25	14	4	17	Σύνολο 60

Τα αποτελέσματα αυτά μπορούν να περιγραφούν ισοδύναμα με τον ακόλουθο τρόπο:

		Μέσον Β		Σύνολο
		+	-	
Μέσον Α	+	25	14	39
	-	4	17	21
Σύνολο		29	31	60

Η μηδενική υπόθεση είναι ότι τα δύο μέσα είναι εξ ίσου αποτελεσματικά στην ανίχνευση ιών γρίπης.

Γενικά: Συχνά έχουμε ένα πληθυσμό τα μέλη του οποίου υποβάλλονται σε δύο "δοκιμασίες" ή "θεραπείες" 1 και 2 με σκοπό τη σύγκριση της επίδοσης ή αποτελεσματικότητας των δύο δοκιμασιών. Έστω p_1 και p_2 οι άγνωστες πιθανότητες-ποσοστά "επιτυχίας" των δοκιμασιών 1 και 2 αντίστοιχα. Στόχος μας είναι ο έλεγχος της μηδενικής υπόθεσης $H_0: p_1 = p_2$. N μέλη του πληθυσμού εκλέγονται τυχαία και το κάθε ένα υποβάλλεται στις δύο δοκιμασίες με τα ακόλουθα N ζεύγη αποτελεσμάτων:

Δοκιμασία		Αριθμός ζευγών
1	2	
E	E	x
E	A	y
A	E	z
A	A	w
Σύνολο		N



όπου, E - επιτυχία και A - αποτυχία.

Τα ίδια αποτελέσματα μπορούν να παρασταθούν ως ακολούθως

		Δοκιμασία 2		
		E	A	
Δοκιμασία 1	E	x	y	x+y
	A	z	w	z+w
		x+z	y+w	N

Το ποσοστό επιτυχίας της δοκιμασίας 1 στο δείγμα είναι $\hat{p}_1 = \frac{x+y}{N}$ και της δοκιμασίας 2, $\hat{p}_2 = \frac{x+z}{N}$. Είναι φανερό ότι η τυχαία μεταβλητή $X + Y$ είναι διωνυμική $B(N, p_1)$ και η τυχαία μεταβλητή $X + Z$ είναι $B(N, p_2)$. Επί πλέον οι δύο αυτές μεταβλητές είναι εξαρτημένες και όταν ισχύει η H_0

$$E[X+Y-(X+Z)] = Np_1 - Np_2 = 0 \quad \text{ή} \quad E(Y) = E(Z).$$

Με άλλα λόγια, όταν ισχύει η μηδενική υπόθεση, οι αναμενόμενοι αριθμοί των y και z ζευγών είναι ίσοι. Έστω τώρα n ο αριθμός των μη ομοειδών ζευγών, δηλ. $n = y + z$. Τότε δ ο θ έ ν τ ο ς ό τ ι υ π ά ρ χ ο υ ν η τέτοια ζεύγη η τυχαία μεταβλητή Y έχει κατανομή $B(n, 1/2)$ όταν ισχύει η H_0 . Άρα ένα τεστ μεγάλου δείγματος (κατά προσέγγιση) λαμβάνεται χρησιμοποιώντας τη στατιστική συνάρτηση .

$$\frac{Y - n/2}{\sqrt{n/4}} \sim \text{κατά προσέγγιση } N(0, 1) \text{ όταν ισχύει } p_1 = p_2. \quad (4.45)$$

Η προσέγγιση βελτιώνεται όταν χρησιμοποιήσουμε τη λεγόμενη "διόρθωση συνεχείας" μειώνοντας την απόλυτη τιμή του $Y - n/2$ κατά $1/2$. Το παραπάνω τεστ λέγεται και τεστ του $M c N e m a n$.

Ένα κατά προσέγγιση $100(1-\alpha)\%$ διάστημα εμπιστοσύνης της διαφοράς $p_1 - p_2$ των δύο ποσοστών δίδεται από τη σχέση



$$\frac{y-z}{N} \pm z_{\alpha/2} \frac{\sqrt{y+z}}{N} \quad (4.46)$$

Λύση του Παραδείγματος: Η τιμή της στατιστικής συνάρτησης του τεστ (4.45) είναι

$$\frac{14-18/2}{\sqrt{18/4}} = \frac{5}{2.12} = 2.36$$

και η τιμή P του τεστ είναι

$$P=P(|Z|>2.36) = 0.0182 < 0.05.$$

Άρα τα πειραματικά δεδομένα μαρτυρούν ότι οι ικανότητες ανίχνευσης των δύο μέσων είναι διάφορες με επίπεδο σημαντικότητας μικρότερο του 5%. Συγκεκριμένα το μέσον Α είναι περισσότερο αποτελεσματικό του Β. Η διόρθωση συνεχείας δίδει σαν τιμή της στατιστικής συνάρτησης (4.45) την $4.5/2.12 = 2.12$, αποτέλεσμα επίσης στατιστικώς σημαντικό.

Ένα 95% διάστημα εμπιστοσύνης της διαφοράς μεταξύ των δύο ποσοστών των θετικών ανιχνεύσεων είναι

$$\begin{aligned} \frac{14-4}{60} \pm 1.96 \frac{\sqrt{18}}{60} \\ = 0.17 \pm 0.14 \\ = 0.03 \text{ έως } 0.31. \end{aligned}$$

Παρατηρούμε ότι αν και το παραπάνω τεστ σημαντικότητας στηρίζεται μόνο στις δύο συχνότητες y και z , η εκτιμώμενη διαφορά μεταξύ των ποσοστών θετικών ανιχνεύσεων και το τυπικό της σφάλμα εξαρτώνται και από το N . Με άλλα λόγια η ένδειξη για την ύπαρξη διαφοράς στηρίζεται μόνο στα μη ομοειδή



ζεύγη· το μέγεθος της διαφοράς εκτιμάται από όλα τα αριθμητικά δεδομένα.

β) Σύγκριση δύο ποσοστών χωρίς ζεύγη

Παράδειγμα 4.10: Σε μία διπλότυφλη κλινική δοκιμή που αφορά ένα ενδεχόμενα υποτασικό φάρμακο χρησιμοποιούνται 100 ασθενείς. 50 ασθενείς εκλέγονται τυχαία και τους δίδεται το ενεργό φάρμακο και στους υπόλοιπους δίδεται το placebo. Ισοδύναμα το φάρμακο και το placebo μπορούν να δοθούν στους 100 ασθενείς στην τύχη έτσι ώστε 50 από αυτούς να πάρουν το φάρμακο και 50 το placebo. Το σχέδιο του πειράματος αυτού λέγεται πλήρως τυχαιοποιημένο σχέδιο (completely randomized design). Η αντίδραση του ασθενούς στην θεραπεία χαρακτηρίζεται σαν ευνοϊκή και μη ευνοϊκή βάσει του βαθμού και της διάρκειας της πύεσης του αίματος. Μας ενδιαφέρει να διαπιστώσουμε αν το φάρμακο έχει την ικανότητα να ελαττώσει την πίεση του αίματος. Η αντίδραση σε 34 από τους 50 ασθενείς που πήραν το φάρμακο ήταν ευνοϊκή και σε 9 από τους 50 που πήραν το placebo ήταν επίσης ευνοϊκή. Η μηδενική υπόθεση είναι ότι το φάρμακο δεν έχει κανένα αποτέλεσμα.

Γενικά: Έστω δύο πληθυσμοί τα μέλη των οποίων μπορούν να καταταχθούν σε δύο κατηγορίες A και B, π.χ. A είναι η κατηγορία των μελών που έχουν ένα χαρακτηριστικό γνώρισμα a και B η κατηγορία εκείνων που δεν έχουν το a. Έστω p_1 και p_2 οι πιθανότητες των μελών των δύο πληθυσμών να ανήκουν στη κατηγορία A αντίστοιχα.

Ένα τυχαίο δείγμα μεγέθους n_1 εκλέγεται από τον πρώτο πληθυσμό με επανατοποθέτηση ή χωρίς επανατοποθέτηση αν ο πληθυσμός είναι άπειρος. Το δείγμα περιέχει x_1 μέλη της κατηγορίας A. Ομοίως ένα ανεξάρτητο τυχαίο δείγμα μεγέθους n_2 εκλέγεται από τον δεύτερο πληθυσμό και περιέχει x_2 μέλη της κατηγορίας A. Το ποσοστό p_1 εκτιμάται με $\hat{p}_1 = x_1/n_1$ και το p_2 με $\hat{p}_2 = x_2/n_2$

Προφανώς

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

και

$$\text{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}. \quad (4.47)$$

Η διακύμανση της διαφοράς $\hat{p}_1 - \hat{p}_2$ εκτιμάται από την παραπάνω σχέση αντικαθιστώντας τα p_1 και p_2 με $\hat{p}_1$ και $\hat{p}_2$ αντίστοιχα, δηλαδή

$$\widehat{\text{var}}(\hat{p}_1 - \hat{p}_2) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}.$$

Ένα κατά προσέγγιση $100(1-\alpha)\%$ διάστημα εμπιστοσύνης της διαφοράς $p_1 - p_2$ των δύο ποσοστών δίδεται από τη γνωστή σχέση

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}. \quad (4.49)$$

Έστω τώρα η μηδενική υπόθεση $H_0: p_1 = p_2$. Έστω επίσης p η κοινή τιμή των ποσοστών. Τα $\hat{p}_1$ και $\hat{p}_2$ είναι και τα δύο εκτιμητές του p και δεν έχει νόημα να εκτιμούμε το p με δύο διαφορετικές ποσότητες. Όταν αληθεύει η H_0 τα δύο δείγματα προέρχονται στην πραγματικότητα από τον ίδιο πληθυσμό και ο άριστος εκτιμητής του p ευρίσκεται ενώνοντας τα δείγματα, δηλαδή

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}. \quad (4.50)$$

Έτσι ο εκτιμητής της διακύμανσης της $\hat{p}_1 - \hat{p}_2$ είναι



$$\widehat{\text{Var}}(\hat{p}_1 - \hat{p}_2) = \hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \quad (4.51)$$

Η μηδενική υπόθεση ελέγχεται κατά προσέγγιση με τη στατιστική συνάρτηση

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \text{ όταν ισχύει η } H_0. \quad (4.52)$$

Λύση του παραδείγματος: Το ποσοστό των ασθενών που αντιδρούν ευνοϊκά στο φάρμακο είναι $\hat{p}_1 = 34/50 = 0.68$ δεδομένου ότι $n_1 = 50$ και $x_1 = 34$. Ομοίως το ποσοστό των ασθενών που αντιδρούν ευνοϊκά με το placebo είναι $\hat{p}_2 = 9/50 = 0.18$ δεδομένου ότι $n_2 = 50$ και $x_2 = 9$. Το κοινό ποσοστό ευνοϊκής αντίδρασης είναι

$$\hat{p} = \frac{39+4}{50+50} = 0.43.$$

Η μηδενική υπόθεση είναι ότι το φάρμακο δεν έχει κανένα αποτέλεσμα. Ακόμη πιο αρνητικά θα λέγαμε ότι το φάρμακο έχει κανένα ή έχει χειρότερο αποτέλεσμα από το placebo. Έτσι η H_0 είναι $p_1 \leq p_2$. Η στατιστική συνάρτηση του τεστ (4.52) έχει τιμή

$$\begin{aligned} Z &= \frac{0.68 - 0.18}{\sqrt{0.43(1 - 0.43)\left(\frac{1}{50} + \frac{1}{50}\right)}} = \frac{0.50}{\sqrt{0.43 \times 0.57 \times \frac{2}{50}}} \\ &= \frac{0.50}{0.099} = 5.05. \end{aligned}$$

Μεγάλες τιμές του Z υποδεικνύουν απόρριψη της H_0 . Πράγματι η τιμή P του τεστ είναι

$$P = P(Z > 5.05) \approx 0.0 < 0.05.$$



Άρα με επίπεδο σημαντικότητας $\alpha=5\%$ τα πειραματικά δεδομένα οδηγούν στην απόρριψη της υπόθεσης ότι το φάρμακο δεν είναι αποτελεσματικό. Υπάρχουν αρκετές ενδείξεις ότι το φάρμακο έχει την ικανότητα να ελαττώσει την πίεση.

Η διαφορά μεταξύ των δύο ποσοστών αποτελεσματικότητας του φαρμάκου και του placebo είναι $0.68-0.18=0.5$. Ένα 95% διάστημα εμπιστοσύνης της διαφοράς p_1-p_2 έχει όρια

$$0.5 \pm 1.96 \sqrt{\frac{0.68(1-0.68)}{50} + \frac{0.18(1-0.18)}{50}} = 0.5 \pm 0.17.$$

4.3.7 Άλλα θέματα

Τα θέματα της Στατιστικής Συμπερασματολογίας είναι πολλά και πολύπλευρα. Αναφέρουμε χαρακτηριστικά μερικούς τίτλους: τεστ καλής προσαρμογής, πίνακες συνάφειας, ανάλυση k δειγμάτων, μη παραμετρική συμπερασματολογία, ανάλυση των τάξεων των μετρήσεων, γραμμική παλινδρόμηση και συσχέτιση, ανάλυση της διακύμανσης, πολυδιάστατη ανάλυση, κ.λ.π. Σε όλα αυτά έχουμε εκτιμητική και έλεγχο υποθέσεων.

Τα περισσότερα από τα παραπάνω θέματα παρουσιάζονται στα Κεφάλαια που ακολουθούν.



ΑΣΚΗΣΕΙΣ

4.1 Έστω $\underline{X} = (X_1, X_2, \dots, X_{n_1})$ και $\underline{Y} = (Y_1, Y_2, \dots, Y_{n_2})$

δύο ανεξάρτητα τυχαία δείγματα από δύο κανονικές κατανομές με μέσες τιμές και διακυμάνσεις (μ_X, σ^2) και (μ_Y, σ^2) αντίστοιχα. Ναδειχθεί ότι για κάθε $w \in [0, 1]$ η στατιστική συνάρτηση $U = U(X, Y) = wS_X^2 + (1-w)S_Y^2$ είναι ένας αμερόληπτος εκτιμητής του σ^2 .

4.2 Έστω $X_1, \dots, X_n$ τυχαίο δείγμα από μια Poisson κατανομή με παράμετρο λ . Να βρεθούν δύο αμερόληπτοι εκτιμητές του λ .

4.3 Δύο εργαστηριακές μέθοδοι A και B διατίθενται για τη μέτρηση ενός μεγέθους και είναι αμερόληπτοι. Η διακύμανση των μετρήσεων A είναι $\sigma_A^2 = 4$ και των μετρήσεων B $\sigma_B^2 = 3$. Οι μετρήσεις A στοιχίζουν 20 δραχμές η μία και οι μετρήσεις B 50 δραχμές η μία. Διατίθενται 500 δραχμές. Ποιά μέθοδος είναι προτιμητέα αν χρησιμοποιηθεί τελικά η μέση τιμή του δείγματος;

4.4 Ένας πληθυσμός αποτελείται από τις ακόλουθες τρεις τιμές 4, 5, 6. Να βρεθεί ο βαθμός εμπιστοσύνης του διαστήματος εμπιστοσύνης (L, U) για τη μέση τιμή του πληθυσμού μ , όπου $L = \bar{X} - \sigma \frac{2}{\sqrt{X}} / 2$, $U = \bar{X} + \sigma \frac{2}{\sqrt{X}} / 2$ και $\bar{X}$

η μέση τιμή δείγματος με επανάθεση μεγέθους $n=2$.

4.5 Δώδεκα φυτά εκλέγονται στη τύχη και από κάθε τέτοιο φυτό εκλέγεται ένα φύλλο στην τύχη. Για κάθε φύλλο προσδιορίζεται το περιεχόμενο του σε ασκορβικό οξύ με τα ακόλουθα αποτελέσματα:



9.35	8.65	11.68	12.77	8.81	9.52
8.68	9.82	10.29	10.99	10.76	10.55.

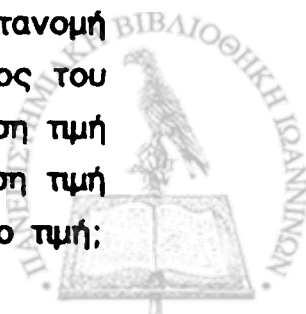
(i) Να εκτιμηθεί το μέσο περιεχόμενο ασκορβικού οξέως και να βρεθεί το τυπικό σφάλμα του εκτιμητή. (ii) Να βρεθεί ένα 95% διάστημα εμπιστοσύνης του μέσου περιεχομένου. (iii) Να βρεθεί ένα διάστημα 95% εμπιστοσύνης του σ .

4.6 Να κατασκευαστεί ένα διάστημα εμπιστοσύνης με επίπεδο σημαντικότητας 95% για την παράμετρο $\theta = 3\mu + 7$ βάσει παρατηρηθέντος δείγματος με τιμές 8, 14, 11 όπου μ είναι η μέση τιμή μιας κανονικής κατανομής $N(\mu, \sigma^2)$ με σ^2 άγνωστο.

4.7 Δίνεται τυχαίο δείγμα $X_1, X_2, \dots, X_9$ από ένα πληθυσμό με κανονική κατανομή $N(\mu, 1)$ και θέλουμε να ελέγξουμε την υπόθεση $H_0: \mu = 20$ ως προς την εναλλακτική $H_a: \mu = 21.5$, χρησιμοποιώντας το ακόλουθο κριτήριο: Απορρίπτουμε την H_0 αν $\bar{X} \geq 21$ και την δεχόμαστε αν $\bar{X} < 21$. Υπολογίστε τις πιθανότητες σφαλμάτων τύπου I και II καθώς και την δύναμη (ισχύ) του κριτηρίου (τεστ).

4.8 Δίνεται μια παρατήρηση X από την κανονική κατανομή $N(\mu, 1)$ και θέλουμε να ελέγξουμε την υπόθεση $H_0: \mu = 10$ ως προς την εναλλακτική $H_a: \mu = 14$ βασιζόμενοι στο ακόλουθο κριτήριο: Απορρίπτουμε την H_0 , εάν $X > 13$ και δεχόμαστε την H_0 οπουδήποτε αλλού. Υπολογίστε τις πιθανότητες των σφαλμάτων τύπου I και II.

4.9 Αν το ύψος του ανθρώπου ακολουθεί κανονική κατανομή με $\sigma = 6.35 \text{ cm}$, ποιο θα πρέπει να είναι το μέγεθος του δείγματος ώστε, με βαθμό εμπιστοσύνης 95%, η μέση τιμή του δείγματος να μη διαφέρει από την αληθινή μέση τιμή του πληθυσμού περισσότερο από 1.27 cm σε απόλυτο τιμή;



- 4.10 10 άτομα εκλέγονται στην τύχη από ένα πληθυσμό και η συστολική πίεση του αίματός τους βρίσκεται 100, 102, 96, 106, 110, 110, 120, 112, 112, 90. Είναι γνωστό ότι η συστολική πίεση ακολουθεί την κανονική κατανομή με μέση τιμή μ και τυπική απόκλιση σ . (i) Κατασκευάστε ένα 95% διάστημα εμπιστοσύνης για το μ . (ii) Να ελεγχθεί η μηδενική υπόθεση $H_0: \mu = 100$ έναντι $H_a: \mu \neq 100$ με βαθμό εμπιστοσύνης 95%.
- 4.11 Πιστεύεται ότι τα νεογέννητα μωρά τα οποία ασκούνται και υποβοηθούνται στο περπάτημα περπατούν τελικά κατά μέσο όρο στους 9.75 μήνες. Έξι μωρά ασκούνται εντατικά και παρατηρούνται οι ακόλουθες ηλικίες περπατήματος (σε μήνες): 9, 9.5, 9.75, 10, 13, 19.5. Να ελεγχθεί στατιστικώς ο ισχυρισμός των 9.75 μηνών με $\alpha = 5\%$ υποθέτοντες ότι η ηλικίες περπατήματος ακολουθούν την κανονική κατανομή.
- 4.12 Ένας χημικός παρασκεύασε μία διάλυση που σκοτώνει το 70% των εντόμων κάποιου είδους. Παιό θα πρέπει να είναι το μέγεθος του δείγματος ώστε, με βαθμό εμπιστοσύνης 95%, τα αποτελέσματα της δειγματοληψίας να είναι εντός 0.02 του αληθινού κλάσματος των εντόμων που σκοτώνονται;
- 4.13 Ένας νέος τύπος λαμπτήρων, οι οποίοι χρησιμοποιούνται σε φλας φωτογραφικών μηχανών, πρόκειται να δοκιμαστεί προκειμένου να εκτιμηθεί η πιθανότητα p ο νέος τύπος λαμπτήρων να δίδει το απαιτούμενο φως στην κατάλληλη στιγμή. Για το σκοπό αυτό ελήφθη ένα δείγμα από 1000 λαμπτήρες του τύπου αυτού και ευρέθη ότι 920 έδιναν το απαιτούμενο φως στην κατάλληλη στιγμή. Σε ποιά όρια θα κείται το p με πιθανότητα 99%;



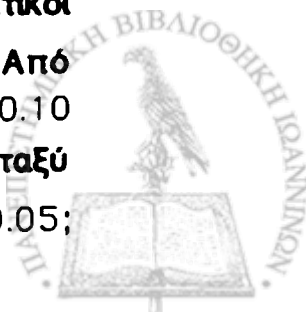
- 4.14 Ένα πείραμα σύγκρισης των χρόνων αντίδρασης δυο φαρμάκων A, B έδωσε τα κάτωθι αποτελέσματα (σε δευτερόλεπτα) για ένα τυχαίο δείγμα 16 ατόμων.

Φάρμακο A	1	3	2	1	2	1	3	2
Φάρμακο B	4	2	3	3	1	2	3	3

Εάν οι κατανομές των χρόνων αντίδρασης των φαρμάκων A και B είναι $N(\mu_A, \sigma^2)$ και $N(\mu_B, \sigma^2)$ αντίστοιχα να βρεθεί ένα 90% διάστημα εμπιστοσύνης για τη διαφορά $\mu_A - \mu_B$.

- 4.15 Για να ελέγξουμε την αποτελεσματικότητα ενός νέου λιπάσματος για την παραγωγή σίτου, χρησιμοποιούμε ένα αγρό ο οποίος διαιρείται σε 60 τετράγωνα με τα ίδια ακριβώς πλεονεκτήματα το καθένα. Το νέο λίπασμα εφαρμόζεται τυχαίως σε 30 τετράγωνα το δε παλιό επίσης τυχαίως στα υπόλοιπα. Η μέση απόδοση των 30 τετραγώνων στα οποία είχε χρησιμοποιηθεί το νέο λίπασμα είναι 18 τόννοι και η τυπική απόκλιση 0.6. Τα αντίστοιχα μεγέθη για τα υπόλοιπα τετράγωνα στα οποία χρησιμοποιήθηκε το παλιό λίπασμα είναι 17 τόννοι και 0.5 αντίστοιχα. Με επίπεδο σημαντικότητας 0.05 να ελεγχθεί η υπόθεση ότι δεν υπάρχει διαφορά μεταξύ των δυό λιπασμάτων.

- 4.16 Δυο μηχανές M_1 και M_2 αυτομάτου συσκευασίας έχουν ρυθμιστεί έτσι ώστε να γεμίζουν πακέτα με βάρος 1Kgr. Υπάρχουν υπόνοιες ότι οι δυο μηχανές δεν λειτουργούν ομοιόμορφα ως προς το βάρος του περιεχομένου. Για το λόγο αυτό λαμβάνονται 100 πακέτα από την παραγωγή κάθε μηχανής M_1, M_2 και ευρίσκεται ότι οι δειγματικοί μέσοι είναι αντίστοιχα 1.07 και 1.18 Kgr. Από προηγούμενες παρατηρήσεις γνωρίζουμε ότι $\sigma_1 = 0.10$ και $\sigma_2 = 0.12$ Kgr. (i) Υπάρχει πράγματι διαφορά μεταξύ των δυο μηχανών με επίπεδο σημαντικότητας $\alpha=0.05$;

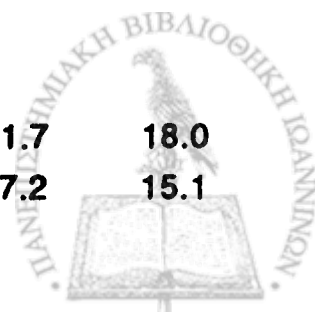


(ii) Να βρεθεί η ισχύς του τεστ που χρησιμοποιείται στο (i) όταν το μέσο βάρος του πακέτου από τη μηχανή M_1 είναι κατά 5 Kgr μεγαλύτερο από το μέσο βάρος του πακέτου για τη μηχανή M_2 .

4.17 Για τη σύγκριση δύο μεθόδων φυσιοθεραπείας ένα δείγμα από 18 ασθενείς εκλέγεται τυχαία. Στους 9 ασθενείς εφαρμόζεται η παλαιά μέθοδος και μετράται ο χρόνος θεραπείας μέχρι να πετύχουμε ένα επιθυμητό αποτέλεσμα. Οι μετρήσεις έδωσαν (σε μέρες θεραπείας) 32, 37, 35, 28, 41, 44, 35, 31, 34. Στους υπόλοιπους ασθενείς εφαρμόζεται η καινούργια μέθοδος και οι χρόνοι (μέρες) θεραπείας μέχρι να πετύχουμε το ίδιο αποτέλεσμα είναι 35, 31, 29, 25, 34, 40, 27, 32, 31. Εάν υποτεθεί ότι οι μετρήσεις ακολουθούν την κανονική κατανομή με μέσους μ_1 και μ_2 αντίστοιχα και με την ίδια διασπορά σ^2 να ελεγχθεί η υπόθεση ότι οι δυο μέθοδοι δίνουν τα ίδια αποτελέσματα με επίπεδο σημαντικότητας 5%.

4.18 Ο Pauling το 1971 ισχυρίστηκε ότι μεγάλες ημερήσιες δόσεις ασκορβικού οξέως μπορούν να εμποδίσουν ή να θεραπεύσουν το κοινό κρυολόγημα. Έτσι παρατηρήθηκε αύξηση της χρήσης βιταμίνης C. Ο Keith το 1974 μελέτησε την επίδραση μαζικών δόσεων ασκορβικού οξέως στην κατανάλωση τροφής, ανάπτυξη, βάρος νεφρών, μορφολογία και σωματική σύνθεση νεαρών αρσενικών χοιριδίων. Τους έδωσε ημερήσιες συμπληρωματικές δόσεις 250, 500 και 1000 mg ασκορβικού οξέως για 9 εβδομάδες. Παρόμοια χοιρίδια χρησιμοποιήθηκαν για control. Ο παρακάτω πίνακας δίνει την περιεκτικότητα σε λιπίδια (σε ποσοστό βάρους σώματος) 8 χοιριδίων (4 control και 4 της ομάδας των 1000 mg).

		Περιεκτικότητα λιπιδίων (%)			
Control	(X)	23.8	15.4	21.7	18.0
1000 mg	(Y)	13.8	9.3	17.2	15.1



Υποθέτοντας ότι η περιεκτικότητα ακολουθεί κανονική κατανομή να ελεγχθεί η υπόθεση ότι το ασκορβικό οξύ δεν μεταβάλλει το ποσοστό λιπιδίων, με επίπεδο σημαντικότητας $\alpha=5\%$.

- 4.19 Σε 9 γυναίκες εφαρμόζεται ορισμένη διαίτα για την ελάττωση του βάρους τους. Ο παρακάτω πίνακας δίνει τα βάρη σε Kgr προ και μετά τη διαίτα.

Προ	67	70	56	62	63	68	60	58	59
Μετά	62	60	58	56	54	56	59	52	52

Με επίπεδο σημαντικότητας 0.01 να ελεγχθεί εάν η διαίτα επέφερε ελάττωση του βάρους.

- 4.20 Τα παρακάτω δεδομένα δείχνουν την πίεση του αίματος εννέα ασθενών που μετρήθηκε με δυο διαφορετικά όργανα A και B.

		Ασθενείς								
		1	2	3	4	5	6	7	8	9
όργανα	A	144	165	125	149	141	118	131	126	147
	B	147	167	124	152	146	120	135	126	149

Με επίπεδο σημαντικότητας 0.01 εξετάστε τον ισχυρισμό ότι κατά μέσο όρο το όργανο B δίνει μεγαλύτερες ενδείξεις από το A.

- 4.21 Δυο ηλεκτρονικά όργανα A και B χρησιμοποιούνται για τη μέτρηση της πίεσης ενός συστήματος. Η ποιότητα των οργάνων εξαρτάται από τη μεταβλητότητα επανειλημμένων μετρήσεων στο ίδιο σύστημα. Μία σειρά από ανεξάρτητες τέτοιες μετρήσεις την ίδια περίπου χρονική στιγμή έδωσε τα ακόλουθα αποτελέσματα.



όργανο	A	B
αριθμ. μετρήσεων	$n_A = 10$	$n_B = 8$
διακύμανση	$S_A^2 = 1.25$	$S_B^2 = 0.28$

Με $\alpha=10\%$ υπάρχει ένδειξη ποιοτικής διαφοράς των οργάνων;

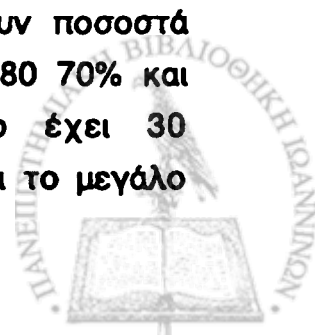
- 4.22 Οι τιμές κλεισίματος των μετοχών δυο ανωνύμων εταιρειών παρατηρήθηκαν για μια περίοδο 16 ημερών. Οι μέσες τιμές και διακυμάνσεις ήταν

$$\bar{y}_1 = 40.33, \bar{y}_2 = 42.54, S_1^2 = 1.54, S_2^2 = 2.96.$$

Εξετάστε εάν τα παραπάνω αριθμητικά δεδομένα δίνουν αρκετές ενδείξεις που να μαρτυρούν διαφορά στην μεταβλητότητα των τιμών των μετοχών των δυο πληθυσμών που συσχετίζονται με τα παραπάνω δείγματα. Υποθέστε ότι οι κατανομές των πληθυσμών είναι οι κανονικές $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$ αντίστοιχα ($\alpha=10\%$).

- 4.23 Από τα στοιχεία ενός εργαστηρίου φαίνονται τα ακόλουθα: 60 transistors τύπου A σε ένα δείγμα χιλίων transistors βρέθηκαν ελαττωματικά και 40 transistors τύπου B σε ένα άλλο δείγμα επίσης χιλίων transistors βρέθηκαν ελαττωματικά. Τα παραπάνω δεδομένα δίνουν αρκετές ενδείξεις που να φαίνεται ότι υπάρχει διαφορά στην ποιότητα μεταξύ transistors τύπου A και τύπου B με βαθμό εμπιστοσύνης 95%;

- 4.24 Ένα μικρό και ένα μεγάλο φροντιστήριο έχουν ποσοστά επιτυχίας στις πανελλήνιες εξετάσεις του 1980 70% και 80% αντίστοιχα. Το μικρό φροντιστήριο έχει 30 υποψήφιους και το μεγάλο 100. Είναι πράγματι το μεγάλο φροντιστήριο καλύτερο από το μικρό; ($\alpha=5\%$).



- 4.25 Δυο διαφορετικά κράματα μετάλλων I, II χρησιμοποιούνται για την κατασκευή εξαρτημάτων μίας οικιακής συσκευής. Εκατό δείγματα από το κάθε ένα υπόκεινται σε ένα τεστ αντοχής σε κάμψεις. Ελαττώματα παρατηρήθηκαν σε 18 από εκείνα που γίνονται με το κράμα I και σε 26 από εκείνα που γίνονται με το κράμα II. Μπορούμε να συμπεράνουμε ότι τα κράματα έχουν την ίδια ανθεκτικότητα ή ότι το κράμα τύπου I είναι ανθεκτικότερο του κράματος τύπου II, $\alpha=5\%$.
- 4.26 Από τα στοιχεία ενός νοσοκομείου φαίνεται ότι 60 άνδρες σε ένα δείγμα από 1000 άνδρες έναντι 40 γυναικών σε ένα άλλο δείγμα από 1000 γυναίκες νοσηλεύονται στο νοσοκομείο αυτό εξ αιτίας καρδιακού νοσήματος. Τα παραπάνω δεδομένα δίνουν αρκετές ενδείξεις που να φαίνεται ότι υπάρχει διαφορά στο ποσοστό καρδιακών παθήσεων μεταξύ των ανδρών και γυναικών που μπήκαν στο νοσοκομείο; ($\alpha=5\%$).



1. The first part of the document is a letter from the President of the United States to the Congress, dated January 1, 1863. It is a very long letter, and it contains a great deal of information about the state of the country at that time.

[illegible]

1. The following is a list of the names of the persons who are known to have been in contact with the subject during the period from 1940 to 1945:

For 1980, the payroll expenditures shown indicate an increase in payroll expenditures of 1980 70% over 1979. To meet expenditures for 80 expenditures for payroll for 1980 100. This amount to payroll expenditures for 1980 and to 1980, to 5X1.

ΚΕΦΑΛΑΙΟ 5

γραμμική παλινδρόμηση

συσχετιση

5.1 Γενικότητες

Συχνά, δύο ή περισσότερες ποσοτικές μεταβλητές εξετάζονται μαζί, με την ελπίδα να προσδιοριστεί η σχέση που υποτίθεται ότι υπάρχει μεταξύ τους. Για παράδειγμα, θα μπορούσε να εξεταστεί η επίδραση που έχει η θερμοκρασία στα πειραματικά αποτελέσματα μιας χημικής διαδικασίας.

Πολλές φορές επίσης, δύο ή περισσότερες ποσοτικές μεταβλητές εξετάζονται με στόχο την πρόβλεψη της μίας από τις άλλες. Για παράδειγμα, γνωρίζοντας τη σχέση μεταξύ εξόδων που δαπανώνται για διαφήμιση και εσόδων από πωλήσεις κάποιου προϊόντος, θα ήταν χρήσιμο να προβλεφθούν, κατά πιθανότητα, τα έσοδα όταν για τη διαφήμιση του προϊόντος δαπανηθεί κάποιο γνωστό ποσό.

Η Στατιστική μεθοδολογία που χρησιμοποιεί τη σχέση μεταξύ δύο ή περισσότερων ποσοτικών μεταβλητών έτσι ώστε η μία να μπορεί να προβλεφθεί από την άλλη ή τις άλλες καλείται Ανάλυση της Παλινδρόμησης (Regression Analysis). Τον όρο "Παλινδρόμηση" χρησιμοποίησε για πρώτη φορά ο F. Galton το 1900 όταν μελετώντας τη σχέση μεταξύ του ύψους των γονέων και αυτού των παιδιών τους, παρατήρησε ένα είδος επαναφοράς (παλινδρόμησης) του ύψους των παιδιών στο ύψος των γονέων τους.

Η Ανάλυση της Παλινδρόμησης αποτελεί ένα από τους σημαντικότερους κλάδους της Στατιστικής με ευρείες εφαρμογές σε



όλες τις σύγχρονες επιστήμες από τις Θεωρητικές μέχρι τις Επιχειρηματικές, Οικονομικές, Βιολογικές, Κοινωνικές κ.λ.π. Επιπλέον περιέχει πολλά και ενδιαφέροντα θεωρητικά θέματα. Έτσι, μία εκτεταμένη μελέτη της είναι πέρα από τους στόχους ενός εισαγωγικού βιβλίου.

Εδώ θα αναπτύξουμε τη μεθοδολογία που εφαρμόζεται στην απλή περίπτωση δύο μόνο μεταβλητών με εξάρτηση γραμμικής μορφής. Θα αναφέρουμε όμως και τις βασικές έννοιες για το γενικώτερο πρόβλημα μη γραμμικών σχέσεων ή περισσότερων από δύο μεταβλητών.

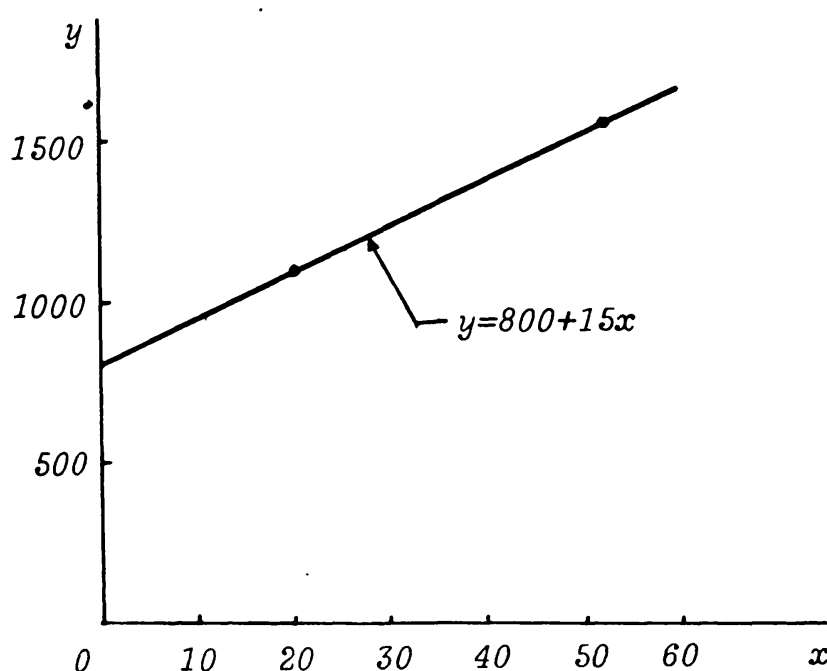
5.2 Σχέσεις μεταξύ μεταβλητών

Η έννοια της σχέσης μεταξύ δύο μεταβλητών, όπως π.χ. μεταξύ εσόδων και εξόδων μιας οικογένειας ή ηλικίας και βάρους ενός παιδιού θεωρείται γνωστή. Οι σχέσεις μεταξύ μεταβλητών διακρίνονται σε συναρτησιακές και στατιστικές ανάλογα με το αν οι μεταβλητές είναι μη στοχαστικές (deterministic) ή στοχαστικές (τυχαίες).

5.2.1. Συναρτησιακή σχέση μεταξύ δύο μεταβλητών

Η συναρτησιακή σχέση μεταξύ δύο μεταβλητών εκφράζεται με κάποιο μαθηματικό τύπο. Αν x είναι η ανεξάρτητη μεταβλητή και y η εξαρτημένη, μία συναρτησιακή σχέση είναι του τύπου $y=f(x)$. Για δεδομένη τιμή του x η συνάρτηση f μας δίνει την αντίστοιχη τιμή του y . Για παράδειγμα, μία εταιρεία ενοικιάζει τα αυτοκίνητά της με βάση τη σχέση $y = 800 + 15x$ όπου, y είναι το ποσό (σε δρχ) που ο πελάτης θα πληρώσει για να διανύσει μια απόσταση x (σε Km). Δύο άτομα που πρόσφατα νοίκιασαν αυτοκίνητα

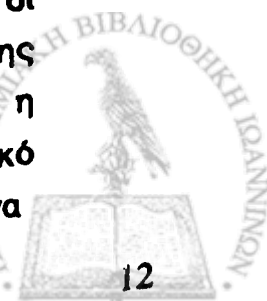
από την εν λόγω εταιρεία, πλήρωσαν 1100 δρχ. και 1565 δρχ. για αποστάσεις 20 Km και 51 Km αντίστοιχα. Η γραφική παράσταση της συναρτησιακής σχέσης που χρησιμοποιεί η εταιρεία όπως και τα δύο ζεύγη (x_i, y_i) των παρατηρήσεων (δεδομένων) που αναφέραμε δίνονται στο Σχήμα 5.1. Είναι φανερό ότι τα σημεία που αντιστοιχούν στις παρατηρήσεις είναι σημεία της καμπύλης, γραφικής παράστασης της συναρτησιακής σχέσης, που εδώ είναι ευθεία γραμμή.



Σχήμα 5.1 Παράδειγμα συναρτησιακής σχέσης

5.2.2 Στατιστική σχέση μεταξύ δύο μεταβλητών

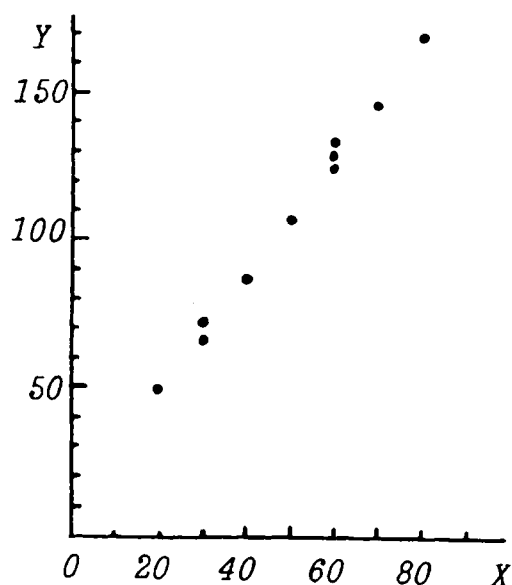
Η στατιστική σχέση διαφέρει από τη συναρτησιακή στο ότι δεν συνιστά μία τέλεια σχέση. Γενικά, οι παρατηρήσεις σε μία στατιστική σχέση δεν αποτελούν σημεία της καμπύλης που δίνεται από τη σχέση. Έστω, για παράδειγμα, ότι η Ελληνική βιομηχανία όπλων (ΕΒΟ) απασχολεί το εργατικό δυναμικό της κατασκευάζοντας κάνες ενός τύπου όπλων μια φορά κάθε μήνα



Μήνας παραγωγής	Μέγεθος παραγγελίας	Ώρες εργασίας
1	X_i	Y_i
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132

Πίνακας 5.1 Δεδομένα μεγέθους παραγγελίας και ωρών εργασίας, παράδειγμα ΕΒΟ

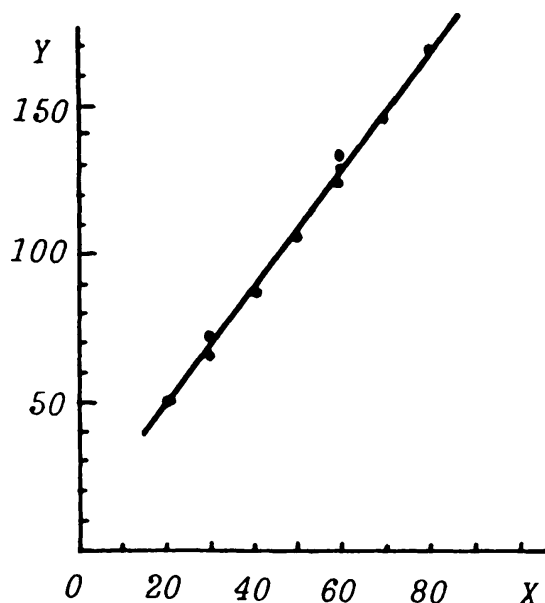
εκτελώντας όλες τις μέχρι τότε παραγγελίες. Στον Πίνακα 5.1 δίνονται οι παραγγελίες προς την ΕΒΟ των δέκα τελευταίων μηνών μαζί με τις αντίστοιχες ώρες εργασίας που απαιτήθηκαν για την εκτέλεσή των, κάτω από παρόμοιες συνθήκες παραγωγής (υποθετικά δεδομένα). Τα δεδομένα του πίνακα που αποτελούν παρατηρήσεις δύο μεταβλητών παριστάνονται γραφικά στο Σχήμα 5.2α. Σαν εξαρτημένη μεταβλητή (Y) θεωρούμε τις ώρες εργασίας και σαν ανεξάρτητη (X) το μέγεθος της παραγγελίας. Το Σχήμα 5.2β προκύπτει από το 5.2α χαράσσοντας την "πλησιέστερη" ευθεία προς τα δεδομένα σημεία. Εδώ φαίνεται καθαρά ότι κάποια σχέση πρέπει να υπάρχει μεταξύ μεγέθους παραγγελίας και των ωρών εργασίας που



Σχήμα 5.2α Διάγραμμα διασποράς, παράδειγμα ΕΒΟ

απαιτούνται για την εκτέλεση της από το γεγονός ότι για μεγαλύτερο αριθμό παραγγελιών υπάρχει η ένδειξη (τάση) ότι απαιτούνται περισσότερες ώρες εργασίας. Όμως η σχέση δεν είναι απόλυτη. Βλέπουμε να υπάρχει μια διασπορά των σημείων που δείχνει ότι μέρος της μεταβλητότητας των ωρών εργασίας δεν συνδέεται με το μέγεθος της παραγγελίας. Για παράδειγμα, το μέγεθος παραγγελίας του 1^{ου} και του 8^{ου} μήνα ήταν το ίδιο (30) όμως για την εκτέλεση απαιτήθηκε διαφορετικός αριθμός ωρών εργασίας. Λόγω της διασποράς των σημείων που παρατηρείται σε μία στατιστική σχέση, ένα σχήμα σαν το 5.2α λέγεται **διάγραμμα διασποράς**. Στη στατιστική ορολογία κάθε σημείο ενός διαγράμματος αποτελεί μία παρατήρηση (δοκιμή). Η ευθεία γραμμή που δίνεται στο Σχήμα 5.2β εκφράζει τη στατιστική σχέση μεταξύ ωρών εργασίας και μεγέθους παραγγελίας. Η ευθεία αυτή, για την οποία θα αναφερθούμε εκτενέστερα πιο κάτω, παριστά γραφικά τη γενική τάση με την οποία οι ώρες εργασίας επηρεάζονται από το μέγεθος της παραγγελίας. Παρατηρούμε ότι τα περισσότερα σημεία δεν αποτελούν σημεία της ευθείας γραμμής. Αυτή η διασπορά των σημείων γύρω από τη γραμμή αντιπροσωπεύει το μέρος της μεταβλητότητας των ωρών εργασίας που δεν συνδέεται με το μέγεθος παραγγελίας και που συνήθως



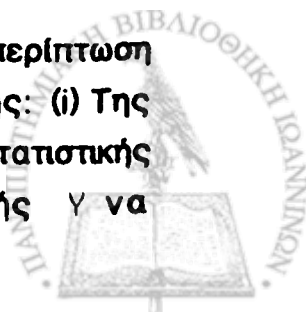


Σχήμα 5.2β Στατιστική σχέση μεταξύ μεγέθους παραγγελίας και ωρών εργασίας, παράδειγμα EBO

θεωρείται ότι είναι τυχαίας προέλευσης. Οι στατιστικές σχέσεις, αν και δεν παρέχουν την ακρίβεια μίας συναρτησιακής, αποτελούν ένα καλύτερο τρόπο μελέτης ενός φαινομένου ή προβλήματος δεδομένου ότι τα φαινόμενα διέπονται από νόμους τύχης, περιέχουν αβεβαιότητες κλπ. που δεν μπορούν να εκφραστούν σε μία συναρτησιακή σχέση που πάντα εξιδανικεύει την πραγματικότητα. Γι'αυτό οι στατιστικές σχέσεις έχουν εφαρμογές σε όλες σχεδόν τις επιστήμες.

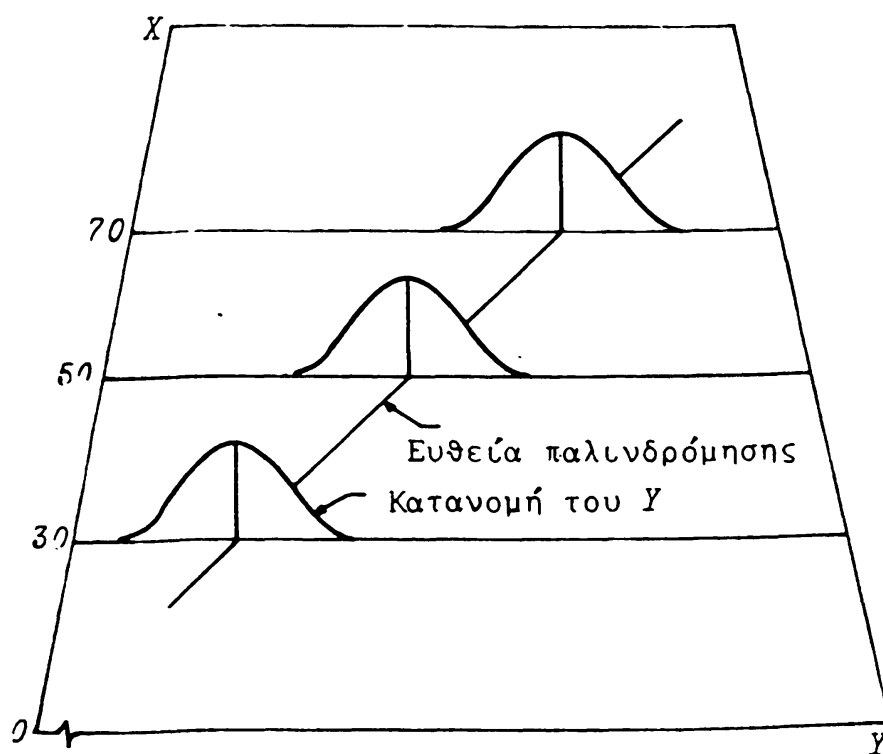
5.3 Μοντέλα παλινδρόμησης και χρησιμότητά τους

Ένα μοντέλο παλινδρόμησης είναι η τυπική περίπτωση έκφρασης των δύο βασικών αρχών μιας στατιστικής σχέσης: (i) Της διασποράς των παρατηρήσεων γύρω από τη καμπύλη της στατιστικής σχέσης. (ii) Της τάσης της εξαρτημένης μεταβλητής Y να



μεταβάλλεται με την ανεξάρτητη ή τις ανεξάρτητες μεταβλητές σύμφωνα με κάποιο συστηματικό τρόπο. Τα στοιχεία αυτά περιέχονται σε ένα μοντέλο παλινδρόμησης αν υποτεθεί ότι:

(i) Οι τιμές της εξαρτημένης μεταβλητής Y που περιέχονται στα δεδομένα αποτελούν τυχαίο δείγμα από ένα πληθυσμό με κάποια κατανομή για κάθε διακεκριμένη τιμή της ανεξάρτητης μεταβλητής X των δεδομένων. (ii) Οι μέσες τιμές των κατανομών αυτών μεταβάλλονται με κάποιο συστηματικό τρόπο. Ας δούμε ξανά το παράδειγμα της ΕΒΟ. Οι ώρες εργασίας Y αποτελούν σε ένα μοντέλο παλινδρόμησης μία τυχαία μεταβλητή. Για κάθε αριθμό παραγγελίας X , θεωρούμε ότι υπάρχει μία κατανομή της Y . Μία τέτοια κατανομή για $X=30$, που αποτελεί την πρώτη παραγγελία του παραδείγματος της ΕΒΟ, φαίνεται στο Σχήμα 5.3α. Οι 73 ώρες εργασίας που απαιτήθηκαν για την εκτέλεση της παραπάνω παραγγελίας θεωρούμε ότι αποτελούν μία τυχαία επιλογή από αυτή τη κατανομή.

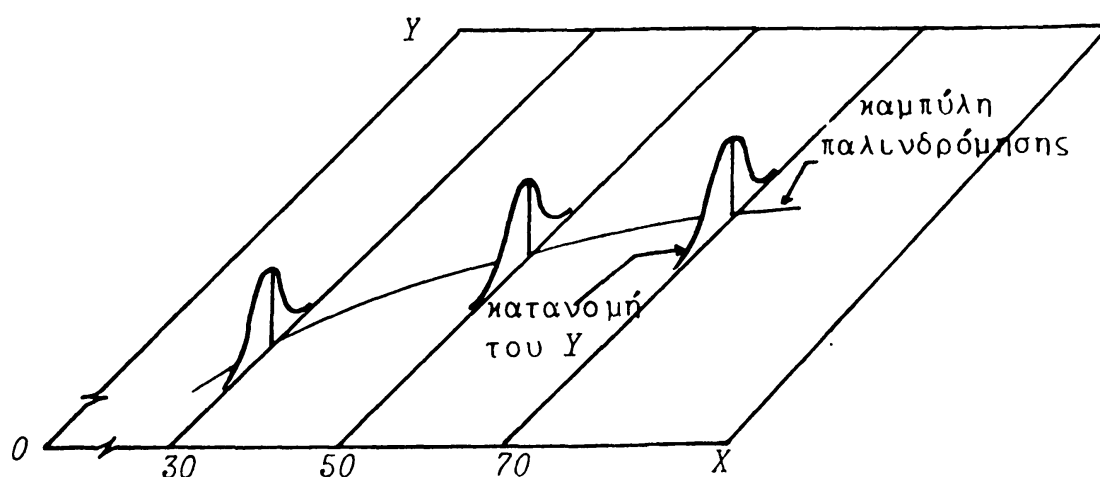


Σχήμα 5.3α Γραφική παράσταση μοντέλου γραμμικής παλινδρόμησης



Στο σχήμα φαίνονται ακόμα οι κατανομές για $X=50$ και $X=70$ όπως και η γραμμή που συνδέει τις μέσες τιμές των κατανομών. Βλέπουμε ότι οι μέσες τιμές μεταβάλλονται με κάποιο συστηματικό τρόπο με τις τιμές της X .

Αυτή η συστηματική σχέση λέγεται συνάρτηση παλινδρόμησης του Y στο X και το γράφημά της καμπύλη παλινδρόμησης. Το Σχήμα 5.3α μας δίνει την περίπτωση γραμμικής συνάρτησης παλινδρόμησης. Για το παράδειγμα της ΕΒΟ αυτό μας δείχνει ότι κατά μέσο όρο οι ώρες εργασίας μεταβάλλονται (σχετίζονται) γραμμικά με το μέγεθος της παραγγελίας. Η γραμμικότητα σαν μορφή εξάρτησης μεταξύ ωρών εργασίας και μεγέθους παραγγελίας δεν είναι από την αρχή γνωστή. Ένα διαφορετικό μοντέλο παλινδρόμησης που θα μπορούσε να ισχύει φαίνεται στο Σχήμα 5.3β όπου η συνάρτηση παλινδρόμησης είναι καμπυλόγραμμη.



Σχήμα 5.3β Γραφική παράσταση μοντέλου
καμπυλόγραμμης παλινδρόμησης

Οι βασικές διαφορές μεταξύ των μοντέλων παλινδρόμησης προκύπτουν από τη μορφή της συνάρτησης παλινδρόμησης και της κατανομής της τ.μ. Y . Στη παλινδρόμηση η μεταβλητή που μας ενδιαφέρει να μελετήσουμε αποτελεί της εξαρτημένη και θεωρείται

τυχαία μεταβλητή. Η ανεξάρτητη δεν είναι τυχαία μεταβλητή και οι παρατηρήσεις που αντιστοιχούν σ' αυτή θεωρούνται ελεγχόμενες ακριβείς μετρήσεις (τιμές) της μεταβλητής, χωρίς σφάλμα. Υπάρχουν πολλές περιπτώσεις που η θεώρηση αυτή είναι ρεαλιστική όπως π.χ. όταν το X παριστάνει δόση φαρμάκου, λιπάσματος, θερμοκρασίας που ο ερευνητής ελέγχει στο πείραμα του ή το εργαστήριό του. Το ότι οι έννοιες "ανεξάρτητη μεταβλητή" και "εξαρτημένη μεταβλητή" είναι περισσότερο τυπικές παρά πραγματικές φαίνεται στην παρακάτω περίπτωση. Έστω ότι θα μας ενδιέφερε να μελετήσουμε τη σχέση μεταξύ του ύψους (Y) και του βάρους (X) νεογέννητων παιδιών. Θεωρώντας την Y σαν εξαρτημένη μεταβλητή θα μπορούσαμε να χρησιμοποιήσουμε την Ανάλυση της Παλινδρόμησης για να κάνουμε προβλέψεις για το ύψος ενός μωρού για το οποίο γνωρίζουμε το βάρος του. Αν μας ενδιέφερε όμως να μπορούμε να προβλέπουμε το βάρος ενός μωρού με δεδομένο ύψος, τότε θα έπρεπε να θεωρήσουμε το X σαν της εξαρτημένη μεταβλητή. Τα μοντέλα παλινδρόμησης χρησιμοποιούνται βασικά για προβλέψεις αλλά και για να περιγράψουν ή να ελέγξουν κάποια κατάσταση. Τη διεύθυνση της ΕΒΟ, για παράδειγμα, θα ενδιέφερε κύρια με βάση παρελθόντα δεδομένα να μπορεί να προβλέπει τις ώρες εργασίας που θα απαιτούσε η εκτέλεση της επόμενης παραγγελίας ώστε να είναι σε θέση να εκτιμήσει τα έξοδα και να οργανώσει καλύτερα τη παραγωγικότητά της. Μετά την εκτέλεση της παραγγελίας κάποια σύγκριση μεταξύ προβλεπόμενων και ωρών που τελικά απαιτήθηκαν θα ήταν ίσως χρήσιμη για κάποιο διοικητικό έλεγχο ή παραπέρα βελτιώσεις.

5.4 Απλή γραμμική παλινδρόμηση

Το μοντέλο της απλής γραμμικής παλινδρόμησης που θα μελετήσουμε με λεπτομέρεια περιλαμβάνει μία ανεξάρτητη μεταβλητή και τη γραμμική συνάρτηση παλινδρόμησης. Έχει τη μορφή



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (5.1)$$

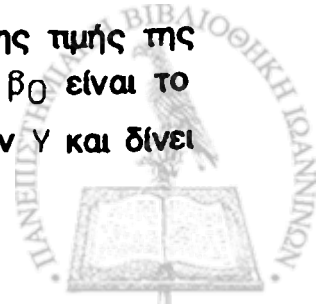
Τα β_0 και β_1 αποτελούν τις άγνωστες παραμέτρους του μοντέλου, τα Y_i , $i=1,2,\dots,n$, είναι οι τιμές (συνήθως λέγονται και αποκρίσεις) της εξαρτημένης μεταβλητής, τα X_i είναι οι τιμές της ανεξάρτητης (ελεγχόμενης) μεταβλητής που θεωρούνται γνωστές σταθερές και n το πλήθος των δεδομένων (X_i, Y_i) που πρόκειται να επεξεργαστούμε. Τα ε_i θεωρούνται ασυσχέτιστες ανά δύο τυχαίες μεταβλητές (βλέπε § 5.14) με $E(\varepsilon_i)=0$ και $\text{Var}(\varepsilon_i)=\sigma^2$. Το ε_i λέγεται τυχαίο σφάλμα και παριστά, για δοθείσα τιμή X_i του X , την απόκλιση της αντίστοιχης τιμής Y_i του Y από την άγνωστη γραμμή παλινδρόμησης.

Το μοντέλο (5.1) λέγεται απλό, γιατί περιέχει μία μόνο ανεξάρτητη μεταβλητή, γραμμικό ως προς τις παραμέτρους, γιατί καμία παράμετρος δεν εμφανίζεται σε εκθέτη ή πολλαπλασιάζεται ή διαφείται με άλλη παράμετρο και γραμμικό ως προς την ανεξάρτητη μεταβλητή, γιατί το X εμφανίζεται μόνο στην πρώτη δύναμη. Ένα μοντέλο παλινδρόμησης με τις παραπάνω ιδιότητες λέγεται και μοντέλο πρώτης τάξης. Σύμφωνα με τις υποθέσεις του μοντέλου (5.1) τα Y_i αποτελούνται από τον σταθερό όρο $\beta_0 + \beta_1 X_i$ και τον τυχαίο όρο ε_i . Είναι δηλαδή τα Y_i ασυσχέτιστες τυχαίες μεταβλητές με $E(Y_i) = \beta_0 + \beta_1 X_i$ και $\text{Var}(Y_i) = \sigma^2$. Για κάποια τιμή του X , έστω X_i , η τ.μ. Y_i έχει κατανομή με μέση τιμή

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

που αποτελεί και τη συνάρτηση παλινδρόμησης του μοντέλου (5.1). Το τυχαίο σφάλμα ε_i , για δοθείσα τιμή X_i , ισούται με την απόκλιση της συγκεκριμένης τιμής Y_i από την αντίστοιχη τιμή της συνάρτησης παλινδρόμησης.

Οι παράμετροι β_0 και β_1 του μοντέλου λέγονται και συντελεστές παλινδρόμησης. Το β_1 είναι η κλίση της γραμμής παλινδρόμησης και δείχνει τη μεταβολή της μέσης τιμής της κατανομής του Y για κάθε μοναδιαία αύξηση στο X . Το β_0 είναι το σημείο που η γραμμή παλινδρόμησης τέμνει τον άξονα των Y και δίνει τη μέση τιμή της κατανομής του Y για $X=0$.



Βασική υπόθεση του μοντέλου είναι ότι η διακύμανση των $\hat{Y}_i$ είναι σταθερή για διάφορες τιμές της ανεξάρτητης μεταβλητής, είναι δηλαδή ανεξάρτητη του X .

5.5 Εκτίμηση της συνάρτησης παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων

Ας υποθέσουμε ότι το μοντέλο της απλής γραμμικής παλινδρόμησης (5.1) είναι κατ'αρχή κατάλληλο για να εκφράσει τη σχέση μεταξύ της εξαρτημένης μεταβλητής Y και της ανεξάρτητης μεταβλητής X που πρόκειται να μελετήσουμε και έστω ότι έχουμε στη διάθεση μας τις παρατηρήσεις (X_i, Y_i) , $i=1, 2, \dots, n$, που πάρθηκαν μετά από κάποιο πείραμα. Οι τιμές των παραμέτρων β_0 και β_1 είναι άγνωστες και κατά συνέπεια θα πρέπει να τις εκτιμήσουμε από τα δεδομένα. Για να βρούμε "καλούς" εκτιμητές των β_0 και β_1 θα χρησιμοποιήσουμε μια τεχνική γνωστή σαν μέθοδος των ελαχίστων τετραγώνων. Για κάθε παρατήρηση (X_i, Y_i) , το σφάλμα όπως φαίνεται στο Σχήμα 5.4 μας δίνει την απόκλιση της Y_i από την (άγνωστη) αναμενόμενη τιμή της αφού $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$:

Έστω

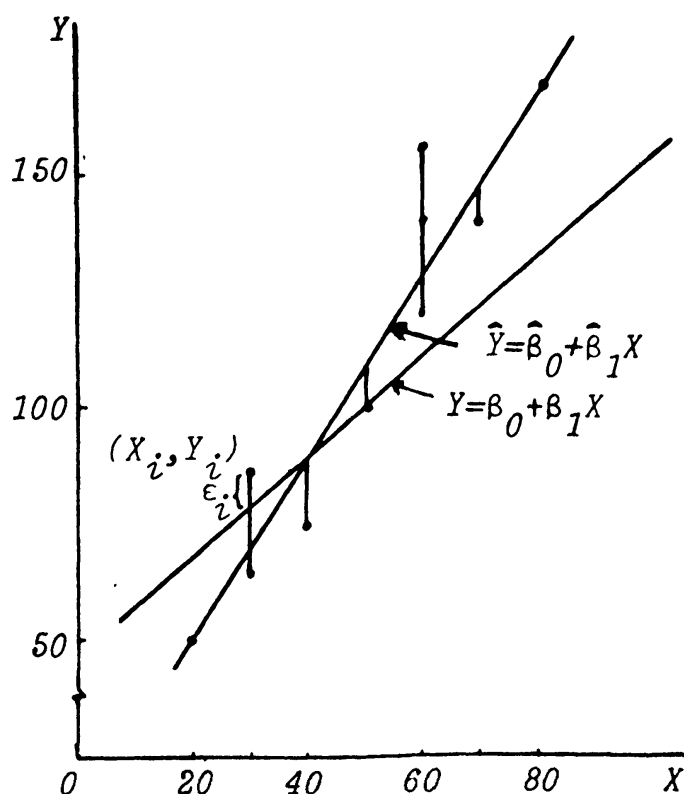
$$Q(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2,$$

το άθροισμα τετραγώνων των n αποκλίσεων. Κατά τη μέθοδο των ελαχίστων τετραγώνων, οι εκτιμητές των β_0 και β_1 είναι αντίστοιχα εκείνες οι τιμές $\hat{\beta}_0$ και $\hat{\beta}_1$ που ελαχιστοποιούν τη συνάρτηση $Q(\beta_0, \beta_1)$. Οι τιμές των β_0 και β_1 που ελαχιστοποιούν τη συνάρτηση $Q(\beta_0, \beta_1)$ προκύπτουν παραγωγίζοντας την $Q(\beta_0, \beta_1)$ ως προς β_0 και β_1 . Είναι



$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_1^n (y_1 - \beta_0 - \beta_1 x_1)$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_1^n x_1 (y_1 - \beta_0 - \beta_1 x_1).$$

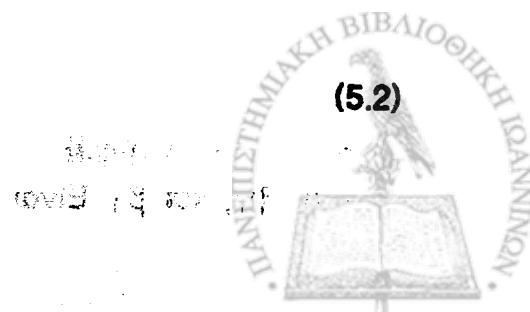


Σχήμα 5.4 Διάγραμμα παλινδρόμησης, παράδειγμα ΕΒΟ

Αν θέσουμε τις μερικές παραγώγους ίσες με το μηδέν και συμβολίσουμε με $\hat{\beta}_0$ και $\hat{\beta}_1$ τις τιμές των β_0 και β_1 αντίστοιχα που ελαχιστοποιούν τη συνάρτηση $Q(\beta_0, \beta_1)$, παίρνουμε τις λεγόμενες κανονικές εξισώσεις:

$$\sum Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i$$

$$\sum X_i Y_i = \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2.$$



Η λύση του συστήματος (5.2) ως προς $\hat{\beta}_0$ και $\hat{\beta}_1$ δίνει

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

(5.3)

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}.$$

Τα $\hat{\beta}_0$ και $\hat{\beta}_1$ λέγονται και εκτιμητές ελαχίστων τετραγώνων των β_0 και β_1 αντίστοιχα.

Για να βρούμε τους εκτιμητές $\hat{\beta}_0$ και $\hat{\beta}_1$ στην περίπτωση της ΕΒΟ, στον Πίνακα 5.2 έχουμε υπολογίσει τις ποσότητες που εμφανίζονται στις σχέσεις (5.3). Αν τώρα αντικαταστήσουμε στις σχέσεις (5.3) τις αντίστοιχες ποσότητες παίρνουμε

$$\hat{\beta}_1 = \frac{61800 - \frac{(500)(1100)}{10}}{28400 - \frac{(500)^2}{10}} = 2.0$$

και

$$\hat{\beta}_0 = \frac{1100}{10} - (2.0) \frac{500}{10} = 10.0$$

Εκτιμάμε δηλαδή ότι θα έχουμε, κατά μέσο όρο, μία αύξηση από δύο ώρες εργασίας για κάθε μοναδιαία αύξηση στον αριθμό της παραγγελίας.



	Y_1	X_1	$Y_1 X_1$	X_1^2
	73	30	2190	900
	50	20	1000	400
	128	60	7680	3600
	170	80	13600	6400
	87	40	3480	1600
	108	50	5400	2500
	135	60	8100	3600
	69	30	2070	900
	148	70	10360	4900
	132	60	7920	3600
Σύνολο	1100	500	61800	28400

Πίνακας 5.2 Βασικοί υπολογισμοί που απαιτούνται για την εύρεση των εκτιμητών $\hat{\beta}_0$ και $\hat{\beta}_1$, παράδειγμα ΕΒΟ

Γνωρίζοντας τώρα τους εκτιμητές των παραμέτρων της συνάρτησης παλινδρόμησης

$$E(Y) = \beta_0 + \beta_1 X$$

είναι φυσικό να την εκτιμήσουμε με την

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \quad (5.4)$$

όπου, $\hat{Y}$ είναι η τιμή της εκτιμώμενης συνάρτησης παλινδρόμησης. Είναι δηλαδή το $\hat{Y}$ ένας εκτιμητής της μέσης τιμής της εξαρτημένης μεταβλητής Y όταν δίνεται ότι η ανεξάρτητη μεταβλητή παίρνει τη τιμή X . Χρησιμοποιώντας τη σχέση (5.4) είμαστε σε θέση να προβλέψουμε, δοθείσης μιας τιμής X_0 της X , την αντίστοιχη μέση τιμή της Y . Έτσι, $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$ είναι ο

εκτιμητής της μέσης τιμής της Y , δηλαδή το $\hat{Y}_0$ είναι η προβλεπόμενη τιμή που αντιστοιχεί στο X_0 . Επίσης το $\hat{Y}_0$ είναι ο εκτιμητής μίας μεμονωμένης τιμής του Y όταν $X=X_0$.

Για τα δεδομένα του παραδείγματος της ΕΒΟ αφού $\hat{\beta}_0=10.0$ και $\hat{\beta}_1=2.0$, η εκτιμώμενη συνάρτηση παλινδρόμησης είναι, $\hat{Y}=10.0+2.0X$. Έτσι, για παράδειγμα, μπορούμε να προβλέψουμε ότι, κατά μέσο όρο, για μέγεθος παραγγελίας $X=55$, οι απαιτούμενες για την εκτέλεση της παραγγελίας ώρες εργασίας θα ήταν $\hat{Y}=10.0+2.0(55)=120$. Για τη τιμή που βρήκαμε μπορούμε να πούμε ότι αν εκτελεστούν πολλές παραγγελίες μεγέθους 55 κάτω από παρόμοιες συνθήκες εργασίας με αυτές των δεδομένων μας τότε, κατά μέσο όρο, οι ώρες εργασίας που θα απαιτηθούν θα είναι περίπου 120. Βέβαια για μεμονωμένη παραγγελία μεγέθους 55 είναι πολύ πιθανόν ότι θα απαιτηθούν περισσότερες ή λιγώτερες από 120 ώρες λόγω της μεταβλητότητας που υπάρχει στο σύστημα και εκφράζεται στο μοντέλο με τα σφάλματα ε_i .

Χρησιμοποιώντας την εκτιμώμενη συνάρτηση παλινδρόμησης (5.4) μπορούμε να υπολογίσουμε για κάθε μία από τις τιμές της ανεξάρτητης μεταβλητής X των δεδομένων, την εκτιμώμενη τιμή της εξαρτημένης μεταβλητής Y . Για παράδειγμα, για τη τιμή $X_1=30$ των δεδομένων της ΕΒΟ παίρνουμε $\hat{Y}_1=10.0+2.0(30)=70$, που διαφέρει από την παρατηρηθείσα τιμή $Y_1=73$. Την διαφορά αυτή την ονομάζουμε υπόλοιπο. Για τα δεδομένα που έχουμε στη διάθεσή μας τα υπόλοιπα, που θα τα συμβολίζουμε με e , υπολογίζονται από τις διαφορές $e_i=Y_i-\hat{Y}_i$, $i=1,2,\dots,n$, και αποτελούν τις κατακόρυφες αποκλίσεις των παρατηρήσεων Y_i από την ευθεία της εκτιμώμενης συνάρτησης παλινδρόμησης. Τα υπόλοιπα e_i δεν πρέπει να συγχέονται με τα τυχαία σφάλματα $\varepsilon_i=Y_i-E(Y_i)$ τα οποία ως γνωστό αποτελούν τις (άγνωστες) κατακόρυφες αποκλίσεις των Y_i από την ευθεία της συνάρτησης παλινδρόμησης που περιλαμβάνει τις άγνωστες παραμέτρους β_0 και β_1 και είναι φυσικά άγνωστη. Τα υπόλοιπα e_i είναι πολύ χρήσιμα στη μελέτη της καταλληλότητας ενός μοντέλου παλινδρόμησης όπως θα δούμε σε επόμενη παράγραφο.



i	X_i	Y_i	$\hat{Y}_i$	$(Y_i - \hat{Y}_i) = e_i$
1	30	73	70	+3
2	20	50	50	0
3	60	128	130	-2
4	80	170	170	0
5	40	87	90	-3
6	50	108	110	-2
7	60	135	130	+5
8	30	69	70	-1
9	70	148	150	-2
10	60	132	130	+2
Σύνολο	500	1100	1100	0

Πίνακας 5.3 Έκτιμώμενες τιμές και υπόλοιπα, παράδειγμα ΕΒΟ

Ο Πίνακας 5.3 δίνει τις παραπάνω ποσότητες για τα δεδομένα της ΕΒΟ. Τα δεδομένα του παραδείγματός μας, οι εκτιμώμενες τιμές $\hat{Y}_i$, τα υπόλοιπα e_i όπως και η εκτιμώμενη γραμμή παλινδρόμησης $Y = 10.0 + 2.0X$ φαίνονται στο Σχήμα 5.5.

Η εκτιμώμενη με τη μέθοδο ελαχίστων τετραγώνων συνάρτηση παλινδρόμησης έχει αρκετές ιδιότητες που αξίζει να τις αναφέρουμε:

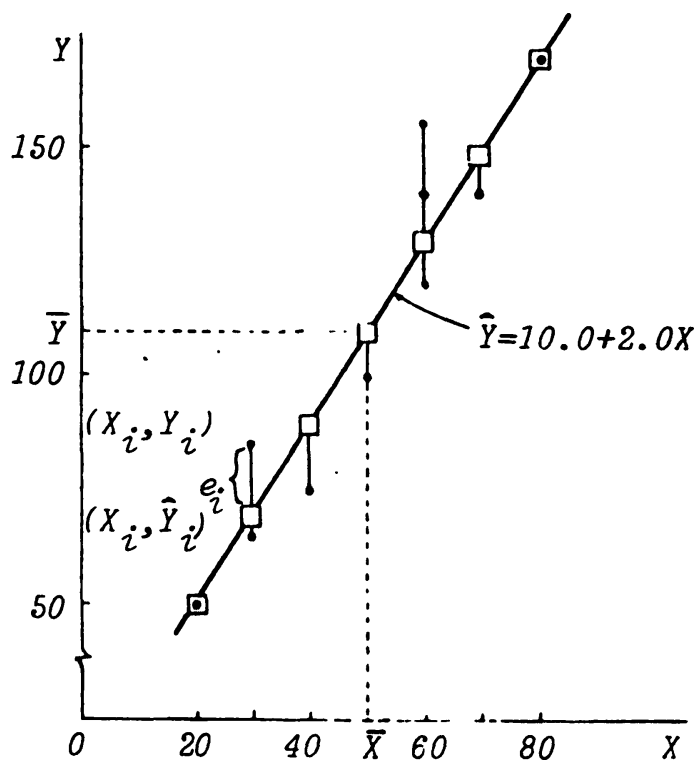
(i) Το κέντρο βάρους των παρατηρήσεων $(\bar{X}, \bar{Y})$, είναι σημείο της εκτιμώμενης γραμμής παλινδρόμησης γιατί, αν $X = \bar{X}$, τότε $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} = \bar{Y} + \hat{\beta}_1 (\bar{X} - \bar{X}) = \bar{Y}$. Η ιδιότητα αυτή φαίνεται στο Σχήμα 5.5.

(ii) Τα υπόλοιπα αθροίζουν στο μηδέν γιατί, χρησιμοποιώντας τη πρώτη από τις κανονικές εξισώσεις (5.2), έχουμε

$$\sum e_i = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \sum Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum X_i = 0.$$



- (iii) Τα υπόλοιπα ικανοποιούν επίσης τη σχέση $\sum x_i e_i = 0$, όπως εύκολα προκύπτει από την δεύτερη από τις κανονικές εξισώσεις (5.2).
- (iv) Το άθροισμα των παρατηρήσεων Y_i ισούται με το άθροισμα των εκτιμώμενων τιμών $\hat{Y}_i$, $\sum Y_i = \sum \hat{Y}_i$, όπως προκύπτει από την πρώτη από τις κανονικές εξισώσεις (5.2).



Σχήμα 5.5 Εκτιμώμενη γραμμή παλινδρόμησης και υπόλοιπα, παράδειγμα ΕΒΟ



5.6 Ιδιότητες των εκτιμητών ελαχίστων τετραγώνων

Ο εκτιμητής ελαχίστων τετραγώνων της παραμέτρου β_1 του μοντέλου (5.1), επειδή $\sum (x_i - \bar{x})\bar{y} / \sum (x_i - \bar{x})^2 = 0$, δίνεται και από τη σχέση

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}.$$

Είναι δηλαδή το $\hat{\beta}_1$ ένας γραμμικός συνδιασμός των ασυσχέπτων τ.μ. y_i . Στηριζόμενοι στις υποθέσεις του μοντέλου (5.1) έχουμε,

$$E(\hat{\beta}_1) = \frac{\sum (x_i - \bar{x}) E(y_i)}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) (\beta_0 + \beta_1 x_i)}{\sum (x_i - \bar{x})^2} = \beta_1 \quad (5.5)$$

είναι δηλαδή ο $\hat{\beta}_1$ αμερόληπτος εκτιμητής του β_1 . Μπορεί ναδειχτεί ότι η διακύμανση του $\hat{\beta}_1$ ισούται με

$$\text{Var}(\hat{\beta}_1) = \sum \left(\frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right)^2 \text{Var}(y_i) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \quad (5.6)$$

Ο εκτιμητής ελαχίστων τετραγώνων της παραμέτρου β_0 του μοντέλου (5.1) δίνεται από τη σχέση

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \sum \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right) y_i$$

Εύκολα βλέπουμε ότι και ο $\hat{\beta}_0$ είναι αμερόληπτος εκτιμητής του β_0 γιατί,



$$\begin{aligned}
E(\hat{\beta}_0) &= \sum \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right) E(y_i) \\
&= \sum \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right) (\beta_0 + \beta_1 x_i) = \beta_0.
\end{aligned} \tag{5.7}$$

Μπορεί επίσης ναδειχτεί ότι η διακύμανση του $\hat{\beta}_0$ ισούται με

$$\begin{aligned}
\text{var}(\hat{\beta}_0) &= \sum \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right)^2 \text{var}(y_i) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right).
\end{aligned} \tag{5.8}$$

Σχετικά με τους εκτιμητές ελαχίστων τετραγώνων $\hat{\beta}_0$ και $\hat{\beta}_1$ ένα σημαντικό θεώρημα, γνωστό σαν θεώρημα των Gauss-Markov, λέει ότι αν ισχύουν οι υποθέσεις του μοντέλου (5.1), οι $\hat{\beta}_0$ και $\hat{\beta}_1$ είναι οι καλύτεροι, με την έννοια της μικρότερης διακύμανσης, γραμμικοί αμερόληπτοι εκτιμητές των β_0 και β_1 αντίστοιχα (Mood, Graybill and Boes, 1974). Έτσι, το $\hat{Y}$ είναι αμερόληπτος εκτιμητής της $E(Y)$ με διακύμανση μικρότερη από οποιονδήποτε άλλο αμερόληπτο γραμμικό εκτιμητή της $E(Y)$.

Για να μπορούμε να εκτιμήσουμε τη διακύμανση των $\hat{\beta}_0$ και $\hat{\beta}_1$ χρειαζόμαστε έναν εκτιμητή της άγνωστης διακύμανσης των σφαλμάτων e_i . Μπορεί να αποδειχθεί ότι ένας αμερόληπτος εκτιμητής του σ^2 δίνεται από τη στατιστική συνάρτηση

$$s^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

το λεγόμενο μέσο τετραγωνικό σφάλμα.



5.7 Στατιστική συμπερασματολογία

Η εύρεση της εκτιμώμενης συνάρτησης παλινδρόμησης αποτελεί μόνο την αρχή στη μελέτη της σχέσης μεταξύ των μεταβλητών X και Y ακόμα και όταν δεν υπάρχουν αποκλίσεις από τις υποθέσεις του μοντέλου (5.1). Εκείνο που βασικά ενδιαφέρει είναι να καταλήξουμε σε διάφορα συμπεράσματα σχετικά με τις παραμέτρους β_0 και β_1 . Ειδικότερα, όταν για παράδειγμα συμπεράνουμε ότι $\beta_1 = 0$, τότε οι μεταβλητές X και Y δεν σχετίζονται γραμμικά. Για το σκοπό αυτό είναι απαραίτητο να προσδιοριστεί κάποια κατανομή των τ.μ. ε_i .

Έστω ότι, πέρα από τις υποθέσεις του μοντέλου (5.1), τα ε_i είναι ανεξάρτητες τ.μ. με κανονική κατανομή, δηλαδή τα Y_i , για κάθε X_i , είναι ανεξάρτητες τ.μ. με κατανομή την κανονική με $E(Y_i) = \beta_0 + \beta_1 X_i$ και $\text{Var}(Y_i) = \sigma^2$.

Η υπόθεση της κανονικότητας των τυχαίων σφαλμάτων δικαιολογείται σε πολλές περιπτώσεις (από το κεντρικό οριακό θεώρημα) καθώς τα ε_i αντιπροσωπεύουν συχνά πλήθος παραγόντων που δεν υπάρχουν στο μοντέλο (5.1) αλλά επηρεάζουν την ανεξάρτητη μεταβλητή Y .

Ο προσδιορισμός της κατανομής των τ.μ. Y_i ταυτόχρονα μας προσδιορίζει και την κατανομή των $\hat{\beta}_0$ και $\hat{\beta}_1$. Τα $\hat{\beta}_0$ και $\hat{\beta}_1$ σαν γραμμικοί συνδιασμοί των τ.μ. Y_i έχουν κανονικές κατανομές με μέσες τιμές και διακυμάνσεις που δίνονται από τις σχέσεις (5.5)-(5.8). Έχουμε δηλαδή,

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right)\right) \quad (5.9)$$

και

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(X_i - \bar{X})^2}\right) \quad (5.10)$$



Επί πλέον, μπορεί να αποδειχθεί (βλέπε π.χ. Mood, Graybill and Boes, 1974) ότι, αν το μοντέλο (5.1) με την πρόσθετη υπόθεση της κανονικότητας των σφαλμάτων ισχύει, η κατανομή της στατιστικής συνάρτησης $(n-2)S^2/\sigma^2$ είναι η χ^2_{n-2} τα δε $\hat{\beta}_0$ και $\hat{\beta}_1$ είναι ανεξάρτητα του S^2 .

Εύκολα τώρα μπορούμε να δείξουμε ότι

$$B_0 = \frac{\hat{\beta}_0 - \beta_0}{S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}} \sim t_{n-2} \quad (5.11)$$

και

$$B_1 = \frac{\hat{\beta}_1 - \beta_1}{S / \sqrt{\sum (x_i - \bar{x})^2}} \sim t_{n-2}. \quad (5.12)$$

Χρησιμοποιώντας τις σχέσεις (5.11) και (5.12) είμαστε σε θέση να ελέγξουμε διάφορες υποθέσεις για τα β_0 και β_1 αντίστοιχα.

Έτσι, με επίπεδο σημαντικότητας α , η μηδενική υπόθεση $H_0: \beta_0 = 0$ με εναλλακτική $H_a: \beta_0 \neq 0$, απορρίπτεται αν για $\beta_0 = 0$, $|B_0| \geq t_{\alpha/2, n-2}$ ή πάλι η μηδενική υπόθεση $H_0: \beta_0 \leq 0$ με εναλλακτική $H_a: \beta > 0$ απορρίπτεται όταν, για $\beta_0 = 0$, $B_0 \geq t_{\alpha, n-2}$.

Όμοια, η μηδενική υπόθεση $H_0: \beta_1 = 0$ με εναλλακτική $H_a: \beta_1 \neq 0$, απορρίπτεται αν, για $\beta_1 = 0$, $|B_1| \geq t_{\alpha/2, n-2}$.

Για τα δεδομένα της ΕΒΟ και με επίπεδο σημαντικότητας $\alpha = 0.05$ έχουμε $S^2 = 60/8 = 7.5$, $\sum (x_i - \bar{x})^2 = 3400$ και $\hat{\beta}_1 = 2.0$. Άρα, $B_1 = 2.0 / 0.0497 = 42.6$. Επειδή $B_1 = 42.6 > t_{0.025, 8} = 2.306$, συμπεραίνουμε ότι $\beta_1 \neq 0$ δηλαδή ότι υπάρχει γραμμική εξάρτηση του Y από το X .

Ένα $(1-\alpha)100\%$ διάστημα εμπιστοσύνης για το β_1 είναι το $\hat{\beta}_1 \pm t_{\alpha/2, n-2} S / \sqrt{\sum (x_i - \bar{x})^2}$.

Έστω πάλι ότι θα μας ενδιέφερε, για τα δεδομένα της ΕΒΟ, να εκτιμήσουμε το β_1 με ένα 95% διάστημα εμπιστοσύνης. Με



βάση τις παραπάνω ποσότητες, ένα 95% διάστημα εμπιστοσύνης για το β_1 είναι το $2.0 \pm 2.306 \sqrt{7.5 / \sqrt{3400}}$ ή το $[1.89, 2.11]$. Έτσι, με συντελεστή εμπιστοσύνης 0.95, εκτιμάμε ότι κατά μέσο όρο οι ώρες εργασίας αυξάνουν κάπου μεταξύ των 1.89 και 2.11 για κάθε μοναδιαία αύξηση στις παραγγελίες. Διαφορετικά, μπορούμε να πούμε ότι το παραπάνω 95% διάστημα εμπιστοσύνης σημαίνει ότι, αν παίρνουμε πολλά ανεξάρτητα δείγματα στα οποία οι τιμές της ελεγχόμενης μεταβλητής X είναι οι ίδιες με αυτές των δεδομένων του παραδείγματός μας και για κάθε δείγμα φτιάξουμε ένα 95% διάστημα εμπιστοσύνης για το β_1 , περιμένουμε σε 95% από τα διαστήματα εμπιστοσύνης να περιέχεται η πραγματική τιμή του β_1 .

5.8 Διάστημα εμπιστοσύνης για την συνάρτηση παλινδρόμησης

Είδαμε στα προηγούμενα ότι η

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

αποτελεί την εκτιμώμενη συνάρτηση παλινδρόμησης για το μοντέλο (5.1) και μπορεί να χρησιμοποιηθεί για να προβλέψει τη μέση τιμή του Y , δοθείσης μίας τιμής X_0 του X .

Για να βγάλουμε διάφορα χρήσιμα συμπεράσματα για το $\hat{Y}_0$ χρειάζεται να προσδιορίσουμε την κατανομή του. Εύκολα παίρνουμε,

$$E(\hat{Y}_0) = E(\hat{\beta}_0 + \hat{\beta}_1 X_0) = \beta_0 + \beta_1 X_0. \quad (5.13)$$

Επίσης μπορεί να δειχθεί ότι

$$\text{var}(\hat{Y}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right) \quad (5.14)$$



Παρατηρούμε ότι η διακύμανση του $\hat{Y}_0$ είναι μικρότερη για τιμές X_0 του X κοντά στο $\bar{X}$. Επειδή τα $\hat{\beta}_0$ και $\hat{\beta}_1$ είναι γραμμικοί μετασχηματισμοί των κανονικών τ.μ. Y_i , το $\hat{Y}_0$ έχει κανονική κατανομή με μέση τιμή και διακύμανση όπως δίνονται στις σχέσεις (5.13) και (5.14) αντίστοιχα. Επί πλέον το $\hat{Y}_0$ σαν συνάρτηση των $\hat{\beta}_0$ και $\hat{\beta}_1$ είναι ανεξάρτητο του S^2 . Έτσι, είναι τώρα εύκολο να δείξουμε ότι

$$\frac{\hat{Y}_0 - E(Y_0)}{\sqrt{\hat{\text{var}}(\hat{Y}_0)}} \sim t_{n-2}. \quad (5.15)$$

Ένα $(1-\alpha)100\%$ διάστημα εμπιστοσύνης για την αληθή τιμή της $E(Y_0)$ είναι το

$$\hat{Y}_0 \pm t_{\alpha/2, n-2} \sqrt{\hat{\text{var}}(\hat{Y}_0)}. \quad (5.16)$$

Ένα 90% διάστημα εμπιστοσύνης για τη μέση τιμή του Y_0 όταν $X_0=55$ για το παράδειγμα της ΕΒΟ είναι το [118.3, 121.7] γιατί, $\hat{Y}_0 = 10.0 + 2.0(55) = 120$, $t_{0.05, 8} = 1.860$ και

$$\hat{\text{var}}(\hat{Y}_0) = 7.5 \left(\frac{1}{10} + \frac{(55-50)^2}{3400} \right) = 0.80515.$$

5.9 Ερμηνεία της στατιστικής συμπερασματολογίας

Όταν, στην απλή γραμμική παλινδρόμηση, η υπόθεση $H_0: \beta_1 = 0$ απορρίπτεται και υπό την προϋπόθεση ότι δεν υπάρχουν αποκλίσεις από τις υποθέσεις του μοντέλου (5.1), μπορούμε να πούμε



ότι η γραμμική παλινδρόμηση εξηγεί κατά ένα ωρισμένο βαθμό την μεταβλητότητα του Y . Το Y μεταβάλλεται με το X και αυξάνει ή μειώνεται αντίστοιχα με το αν το $\hat{\beta}_1$ είναι θετικό ή αρνητικό. Έχουμε δηλαδή ενδείξεις ότι για λόγους πρόβλεψης είναι καλύτερα να χρησιμοποιήσουμε και το X μέσω της εκτιμώμενης συνάρτησης παλινδρόμησης παρά μόνο το Y . Οπωσδήποτε υπάρχει η πιθανότητα τα X και Y να συνδέονται και με άλλου είδους σχέσεις όπως είναι για παράδειγμα η καμπυλόγραμμη. Επίσης δεν μπορούμε να συμπεράνουμε αν το X προκαλεί τις μεταβολές στο Y γιατί είναι πιθανόν τα X και Y να εξαρτώνται από μία τρίτη μη μετρηθείσα μεταβλητή που προκαλεί τις μεταβολές της Y . Όταν τώρα η $H_0: \beta_1 = 0$ δεν απορρίπτεται, υπάρχει ένδειξη ότι τα X και Y δεν συνδέονται γραμμικά (μπορεί βέβαια να συνδέονται με κάποια σχέση καμπυλόγραμμης μορφής). Έτσι, αν περιοριστούμε στο απλό γραμμικό μοντέλο (5.1), δεν χρειάζεται το X για προβλέψεις του Y που σε μία τέτοια περίπτωση γίνονται με βάση την $\hat{Y} = \bar{Y}$.

5.10 Η ανάλυση της διακύμανσης στη παλινδρόμηση

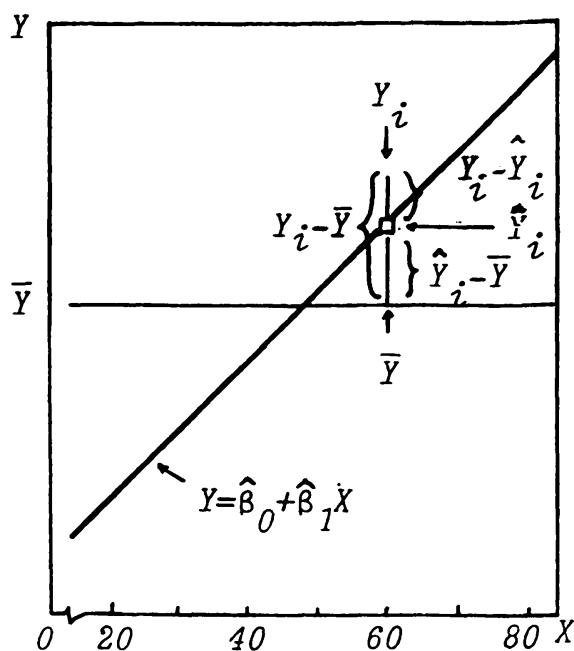
Μία διαφορετική και ισοδύναμη προσέγγιση στη μελέτη της σχέσης μεταξύ των μεταβλητών ενός μοντέλου παλινδρόμησης προσφέρει η μεθοδολογία της Ανάλυσης της Διακύμανσης. Η Ανάλυση της Διακύμανσης, που δεν θα προσθέσει τίποτε σ'αυτά που ήδη γνωρίζουμε για την απλή γραμμική παλινδρόμηση, βασίζεται στη διάσπαση ενός αθροίσματος τετραγώνων και των αντίστοιχων βαθμών ελευθερίας της εξαρτημένης μεταβλητής Y και αποτελεί το βασικό εργαλείο για την μελέτη περισσότερο πολύπλοκων μοντέλων παλινδρόμησης. Επί πλέον η Ανάλυση της Διακύμανσης παρουσιάζει και υπολογιστικά πλεονεκτήματα.

Ας θεωρήσουμε την απόκλιση $Y_i - \bar{Y}$ δηλαδή την βασική ποσότητα που μετρά την διακύμανση (μεταβλητότητα) των παρατηρήσεων Y_i . Η απόκλιση αυτή γράφεται,



$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i \quad (5.17)$$

μπορεί δηλαδή να θεωρηθεί ότι αποτελείται από δύο μέρη. Την απόκλιση της προβλεπόμενης τιμής $\hat{Y}_i$ από την μέση τιμή $\bar{Y} = \bar{\hat{Y}}$ (βλέπε § 5.5) και την απόκλιση της Y_i από την εκτιμώμενη γραμμή παλινδρόμησης. Η σχέση (5.17) δίνεται γραφικά στο Σχήμα 5.6 για μία από τις παρατηρήσεις του παραδείγματος της ΕΒΟ.



Σχήμα 5.6 Ενδεικτική διάσπαση για μία από τις αποκλίσεις $Y_i - \bar{Y}$, παράδειγμα ΕΒΟ

Αν υψώσουμε την (5.17) στο τετράγωνο και αθροίσουμε, παίρνουμε

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 + 2 \sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}).$$

Αλλά



$$\begin{aligned}
\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) &= \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y}) \\
&= \sum (Y_i - \bar{Y} - \hat{\beta}_1 (X_i - \bar{X}))\hat{\beta}_1 (X_i - \bar{X}) \\
&= \hat{\beta}_1 \sum (\hat{Y}_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}_1^2 \sum (X_i - \bar{X})^2 \\
&= 0.
\end{aligned}$$

Βλέπουμε δηλαδή ότι και για τα αθροίσματα τετραγώνων των αποκλίσεων της (5.17) ισχύει ανάλογη σχέση:

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \quad (5.18)$$

ή συμβολικά,

$$SS_{tot} = SS_{reg} + SS_{res}.$$

Η αξιοσημείωτη σχέση (5.18) μας λέει ότι το άθροισμα τετραγώνων των παρατηρήσεων Y_i από την μέση τιμή (ολικό άθροισμα τετραγώνων) ισούται με το άθροισμα τετραγώνων των αποκλίσεων των προβλεπόμενων τιμών $\hat{Y}_i$ από τη μέση τους τιμή (άθροισμα τετραγώνων λόγω παλινδρόμησης) και το άθροισμα τετραγώνων των υπολοίπων. Βλέπουμε δηλαδή ότι ένα μέρος από την μεταβλητότητα της Y οφείλεται στην σχέση της με την X όπως αυτή εκφράζεται στο απλό γραμμικό μοντέλο παλινδρόμησης και ένα άλλο μέρος στο γεγονός ότι οι παρατηρήσεις Y_i δεν ανήκουν όλες στην εκτιμώμενη ευθεία. Από την σχέση (5.18) γίνεται φανερό ότι όσο μεγαλύτερο είναι το SS_{reg} τόσο ισχυρότερη είναι η σχέση εξάρτησης της Y από την X . Έτσι, ένας άλλος τρόπος για να εκτιμήσουμε την αξία της παλινδρόμησης είναι ο λόγος

$$R^2 = \frac{SS_{reg}}{SS_{tot}}$$

να μην είναι πολύ μικρότερος από την μονάδα. Το R^2 λέγεται και συντελεστής προσδιορισμού και μας δίνει το ποσοστό



μεταβλητότητας της τ.μ. Y που οφείλεται στην παλινδρόμηση.

Σε κάθε άθροισμα τετραγώνων αντιστοιχούν ορισμένοι βαθμοί ελευθερίας (β.ε.) που ισοδυναμούν με το πλήθος των ανεξάρτητων συναρτήσεων των Y_i που απαιτούνται για τον υπολογισμό του εν λόγω αθροίσματος. Έτσι, το SS_{tot} έχει $n-1$ β.ε. γιατί $\sum(Y_i - \bar{Y}) = 0$, το $SS_{reg} = \sum(\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1^2 \sum(X_i - \bar{X})^2$ έχει 1

β.ε. γιατί μπορεί να υπολογισθεί από μία συνάρτηση των Y_i την $\hat{\beta}_1$ το δε SS_{res} έχει $n-2$ β.ε. γιατί $\sum(Y_i - \hat{Y}_i) = 0$ και $\sum X_i(Y_i - \hat{Y}_i) = 0$ όπως φαίνεται από τις κανονικές εξισώσεις (5.2). Ισχύει δηλαδή και για τους βαθμούς ελευθερίας που αντιστοιχούν στα αθροίσματα τετραγώνων της (5.18) ανάλογη διασπαστική σχέση.

Η σχέση (5.18) και οι αντίστοιχοι των αθροισμάτων τετραγώνων β.ε. οδηγούν στον γνωστό σαν πίνακα της Ανάλυσης της Διακύμανσης (ΑΝΑΔΙΑ).

ΑΝΑΔΙΑ			
Προέλευση Μεταβλητότητας	Άθροισμα Τετραγώνων	Βαθμοί Ελευθερίας	Μέσο Τετράγωνο
Παλινδρόμηση	$SS_{reg} = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MS_{reg} = \frac{SS_{reg}}{1}$
Υπόλοιπο ή Απόκλιση από την Παλινδρόμηση	$SS_{res} = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MS_{res} = S^2 = \frac{SS_{res}}{n-2}$
Ολική Μεταβλητότητα	$SS_{tot} = \sum(Y_i - \bar{Y})^2$	$n-1$	

Η στήλη "Μέσο Τετράγωνο" προκύπτει διαιρώντας τα αθροίσματα τετραγώνων με τους αντίστοιχους βαθμούς ελευθερίας.

Για τα δεδομένα του παραδείγματος της ΕΒΟ μετά τους σχετικούς υπολογισμούς προκύπτει ο παρακάτω πίνακας:



ΑΝΑΔΙΑ			
Προέλευση Μεταβλητότητας	Άθροισμα Τετραγώνων	Βαθμοί Ελευθερίας	Μέσο Τετράγωνο
Παλινδρόμηση	13600	1	13600
Υπόλοιπο	60	8	7.5
Ολική Μεταβλητότητα	13660	9	

Αν υποθέσουμε ότι οι υποθέσεις του μοντέλου (5.1) με την πρόσθετη υπόθεση της κανονικότητας των σφαλμάτων ισχύουν, είναι δυνατόν να κατασκευάσουμε ένα στατιστικό τεστ αξιολόγησης της παλινδρόμησης βασισμένο στην F κατανομή.

Μπορεί να αποδειχθεί (βλέπε π.χ. Wilks, 1962) ότι οι στατιστικές συναρτήσεις SS_{reg} και SS_{res} είναι ανεξάρτητες και επιπλέον

$$\frac{SS_{reg}}{\sigma^2} \sim \chi^2_1$$

και

$$\frac{SS_{res}}{\sigma^2} \sim \chi^2_{n-2}$$

Έτσι η στατιστική συνάρτηση

$$F = \frac{MS_{reg}}{MS_{res}} \sim F_{1,n-2}. \quad (5.19)$$

Η σχέση (5.19) μας δίνει την δυνατότητα να ελέγξουμε την σημαντικότητα της παλινδρόμησης. Η υπόθεση $H_0: \beta_1=0$ έναντι της $H_a: \beta_1 \neq 0$, για επίπεδο σημαντικότητας α , απορρίπτεται όταν $F > F_{\alpha,1,n-2}$. Σε μια τέτοια κρίση περιοχή μας οδηγούν οι σχέσεις $E(MS_{res}) = \sigma^2$ και $E(MS_{reg}) = \sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2$ από τις οποίες προκύπτει ότι αν $\beta_1=0$, $E(MS_{reg}) = E(MS_{res})$ ενώ αν $\beta_1 \neq 0$, $E(MS_{reg}) > E(MS_{res})$.

Μπορεί εύκολα να αποδειχθεί (βλέπε Neter and Wasserman, 1975) ότι το παραπάνω F -τεστ είναι ισοδύναμο με



το t -τεστ της § 5.7 που χρησιμοποιείται για τον αντίστοιχο έλεγχο.

Για τα δεδομένα του παραδείγματος της ΕΒΟ έχουμε, $F = 13600/7.5 = 1813.3$. Επίσης για $\alpha=0.05$, $F_{0.05,1,8}=5.32$. Επειδή $1813.3 > 5.32$, η $H_0: \beta_1=0$ απορρίπτεται, συμπέρασμα στο οποίο έχουμε ήδη καταλήξει χρησιμοποιώντας το αντίστοιχο t -τεστ.

Για τα ίδια δεδομένα μπορούμε εύκολα να υπολογίσουμε τον συντελεστή προσδιορισμού R^2 . Έχουμε $R^2 = 13600/13660 = 0.9956$, δηλαδή το 99.56% της μεταβλητότητας της Y οφείλεται στην παλινδρόμηση.

5.11 Έλεγχος για την καταλληλότητα του μοντέλου παλινδρόμησης-Ανάλυση των Υπολοίπων

Όταν ένα μοντέλο Παλινδρόμησης, όπως το απλό γραμμικό μοντέλο (5.1), επιλέγεται για κάποια εφαρμογή, δεν είναι συνήθως γνωστό από πριν αν αυτό είναι το καταλληλότερο. Είναι δυνατόν μία ή περισσότερες από τις υποθέσεις του μοντέλου, όπως η γραμμικότητα της συνάρτησης παλινδρόμησης ή η κανονικότητα των σφαλμάτων, να μην ισχύουν για τα δεδομένα που εξετάζονται. Έτσι είναι σημαντικό, προτού προχωρήσουμε σε λεπτομερή ανάλυση των δεδομένων, να ελέγξουμε την καταλληλότητα του μοντέλου που έχουμε επιλέξει. Για τον σκοπό αυτό έχουν διαμορφωθεί διάφορα στατιστικά τεστ και γραφικές μέθοδοι.

Τέτοια στατιστικά τεστ είναι: (i) το τεστ των Durbin - Watson που χρησιμοποιείται για τον έλεγχο της τυχαιότητας των σφαλμάτων, (ii) το τεστ συσχέτισης των τάξεων (rank correlation) που χρησιμοποιείται για τον έλεγχο της σταθερότητας της διακύμανσης, (iii) το τεστ καλής προσαρμογής που χρησιμοποιείται για τον έλεγχο της κανονικότητας των σφαλμάτων κλπ. Ο έλεγχος της καταλληλότητας που γίνεται με τα στατιστικά τεστ είναι συγκεκριμένος, επιστημονικά αυστηρός και αποφεύγει την υποκειμενικότητα που υπεισέρχεται στην ερμηνεία μίας γραφικής παράστασης. Λεπτομέρειες



γί'αυτά τα στατιστικά τεστ που ξεφεύγουν από τον σκοπό ενός εισαγωγικού βιβλίου δίνονται, μεταξύ άλλων στα βιβλία των Neter and Wasserman (1975) και Draper and Smith (1981).

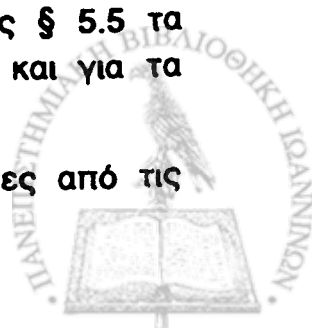
Οι γραφικές μέθοδοι εξ'άλλου συνίστανται σε διάφορες γραφικές παραστάσεις των υπολοίπων και μπορούν να δώσουν μόνον ενδείξεις για τυχόν αποκλίσεις από τις υποθέσεις ενός μοντέλου παλινδρόμησης. Παρ'όλα αυτά, οι γραφικές μέθοδοι επειδή είναι απλές, πρακτικές και συνήθως επαρκείς, έχουν επικρατήσει. Αποτελούνται από γραφικές παραστάσεις των υπολοίπων γί'αυτό και η όλη μεθοδολογία είναι γνωστή σαν ανάλυση των υπολοίπων.

Τα υπόλοιπα (residuals), $e_i = Y_i - \hat{Y}_i$, όπως ορίστηκαν στην § 5.5 είναι οι διαφορές μεταξύ παρατηρηθείσας και εκτιμώμενης τιμής του Y . Μπορεί δηλαδή να θεωρηθούν σαν τα παρατηρηθέντα σφάλματα, σε αντιδιαστολή με τα άγνωστα πραγματικά σφάλματα, $\varepsilon_i = Y_i - E(Y_i)$, του μοντέλου (5.1). Στις υποθέσεις του μοντέλου (5.1), τα ε_i θεωρήθηκαν ανεξάρτητες $N(0, \sigma^2)$ τ.μ. Αν το μοντέλο (5.1) είναι κατάλληλο για τα δεδομένα που εξετάζουμε, τα e_i θα έπρεπε να αντανakλούν τις ιδιότητες που υποθέσαμε ότι έχουν τα ε_i . Σημειώνουμε όμως εδώ ότι επειδή για την μέση τιμή των υπολοίπων έχουμε πάντα $\bar{e} = 0$ (βλ. § 5.5), δεν μπορούμε να ελέγξουμε κατά πόσο τα σφάλματα ικανοποιούν την υπόθεση $E(\varepsilon_i) = 0$. Η διακύμανση των n υπολοίπων e_i για το μοντέλο (5.1) ορίζεται από τη σχέση

$$\frac{\sum (e_i - \bar{e})^2}{n-2} = S^2 = MS_{res},$$

και αποτελεί όπως έχουμε ήδη αναφέρει αμερόληπτο εκτιμητή της διακύμανσης των σφαλμάτων σ^2 . Συχνά για ευκολία στην ανάλυση των υπολοίπων χρησιμοποιούνται στην θέση των υπολοίπων e_i , τα κανονικοποιημένα υπόλοιπα $e_i / \sqrt{MS_{res}}$. Όπως εύκολα διαπιστώνουμε από τις ιδιότητες (ii) και (iii) της § 5.5 τα υπόλοιπα e_i δεν είναι ανεξάρτητες τ.μ. Το ίδιο ισχύει και για τα κανονικοποιημένα υπόλοιπα.

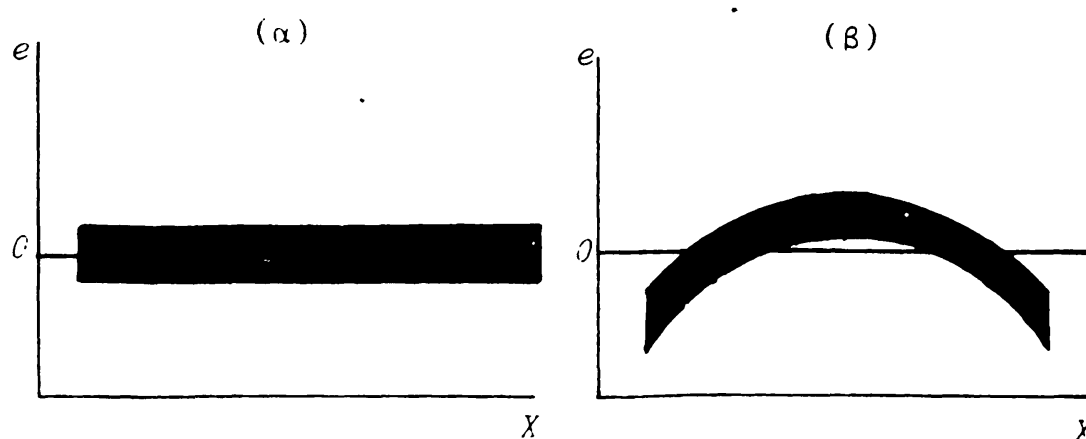
Στην συνέχεια θα περιγράψουμε τις κυριότερες από τις



ἑνδογενῆς μεθόδους τῆς ἀνάλυσης τῶν υπολοίπων γιὰ τὴν περίπτωση τοῦ ἀπλοῦ γραμμικοῦ μοντέλου (5.1), ἀν καὶ τὰ βασικὰ στοιχεῖα ἐφαρμόζονται σὲ ὅλα τὰ στατιστικὰ μοντέλα.

5.11.1 Ἐλεγχος γιὰ τὴν γραμμικότητα τῆς συνάρτησης παλινδρόμησης

Ἡ γραφικὴ παράσταση τῶν υπολοίπων e_i ὡς πρὸς τὴν ἀντίστοιχες τιμὲς X_i τῆς ανεξάρτητης μεταβλητῆς γιὰ τὰ δεδομένα ποὺ ἐξετάζουμε μπορεῖ νὰ μας δώσει ἐνδείξεις γιὰ τυχόν αποκλίσεις ἀπὸ τὴν ὑπόθεση τῆς γραμμικότητας τῆς συνάρτησης παλινδρόμησης. Τὸ Σχῆμα 5.7α δείχνει τὴν μορφή τῆς γραφικῆς παράστασης ὅταν τὸ γραμμικὸ μοντέλο εἶναι κατάλληλο. Τὰ υπολοίπα θὰ πρέπει σ'αυτὴ τὴν περίπτωση νὰ σχηματίζουν μίαν ὀριζόντιαν ζώνη γύρω ἀπὸ τὸ μηδέν χωρὶς νὰ δείχνουν κάποια συστηματικὴ τάση ὅταν εἶναι θετικὰ ἢ ἀρνητικὰ. Ἡ γραφικὴ παράσταση τῶν υπολοίπων γιὰ τὰ δεδομένα τῆς ΕΒΟ (Πίνακας 5.3), ὅπως φαίνεται ἀπὸ τὸ Σχῆμα 5.8 ἐμπίπτει σ' αὐτὴν τὴν περίπτωση.



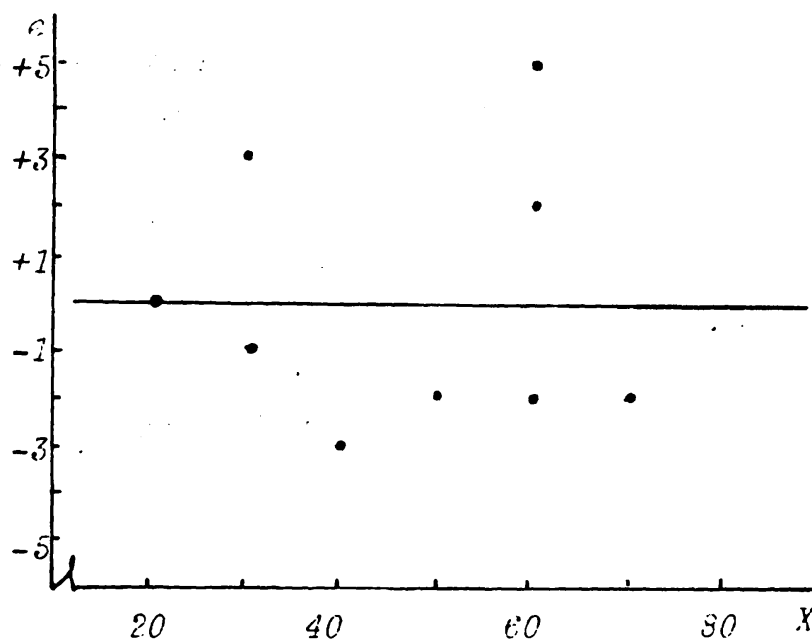
Σχῆμα 5.7 Δυνατές μορφές γραφικῆς παράστασης υπολοίπων ὡς πρὸς τὴν ανεξάρτητη μεταβλητὴ

Τὸ Σχῆμα 5.7β δείχνει τὴν μορφή ποὺ εμφανίζει ἡ γραφικὴ παράσταση τῶν υπολοίπων ὡς πρὸς X ὅταν ὑπάρχουν αποκλίσεις ἀπὸ τὴν γραμμικότητα τοῦ μοντέλου δίνοντας ἐνδείξεις γιὰ



καμπυλόγραμμη συνάρτηση παλινδρόμησης. Σε μία τέτοια περίπτωση τα υπόλοιπα τείνουν να γίνονται θετικά και αρνητικά με κάποιο συστηματικό τρόπο.

Όταν οι αποκλίσεις από την γραμμικότητα της συνάρτησης παλινδρόμησης είναι άλλης μορφής από την καμπυλόγραμμη, η γραφική παράσταση δημιουργεί διαφορετική οπωσδήποτε εικόνα από αυτή του Σχήματος 5.7β.

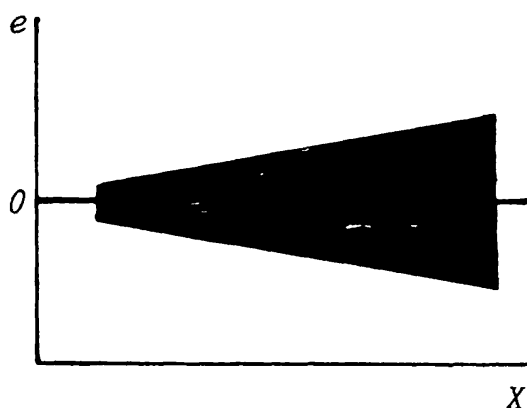


Σχήμα 5.8 Γραφική παράσταση υπολοίπων ως προς την ανεξάρτητη μεταβλητή, παράδειγμα ΕΒΟ

5.11.2 Έλεγχος για την σταθερότητα της διακύμανσης των σφαλμάτων

Η γραφική παράσταση των υπολοίπων ως προς την ανεξάρτητη μεταβλητή X βοηθάει επίσης στο να διαπιστώσουμε κατά πόσον υπάρχουν αποκλίσεις από την υπόθεση της σταθερότητας της διακύμανσης των σφαλμάτων. Το Σχήμα 5.9 δείχνει την μορφή που εμφανίζει η γραφική παράσταση όταν η διακύμανση των σφαλμάτων

αυξάνεται με το X . Σε πολλές εφαρμογές των Οικονομικών Επιστημών εμφανίζεται η εικόνα του Σχήματος 5.9. Μπορούμε επίσης να συναντήσουμε και περιπτώσεις όπου η διακύμανση ελαττώνεται όταν το X αυξάνει ή διάφορες άλλες μορφές μεταβολής. Αν η γραφική παράσταση των υπολοίπων ως προς τις τιμές της X ή ως προς τις τιμές της Y παρουσιάζει την εικόνα του Σχήματος 5.7α τότε μπορούμε να πούμε ότι η υπόθεση ότι η διακύμανση των σφαλμάτων του μοντέλου (5.1) είναι σταθερή, δεν παραβιάζεται από τα συγκεκριμένα δεδομένα που εξετάζουμε. Τέτοια είναι η περίπτωση των δεδομένων της ΕΒΟ όπως φαίνεται στο Σχήμα 5.8.



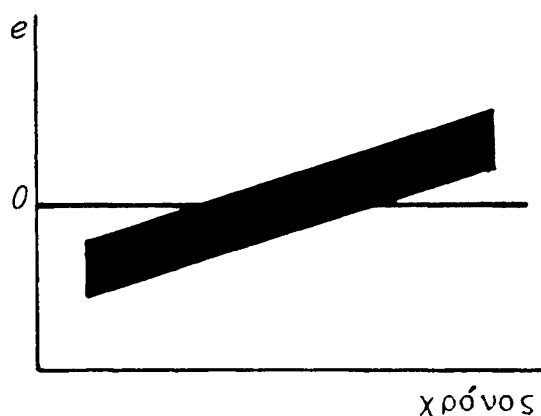
Σχήμα 5.9 Δυνατή μορφή γραφικής παράστασης υπολοίπων ως προς την ανεξάρτητη μεταβλητή

5.11.3 Έλεγχος για την ανεξαρτησία των σφαλμάτων

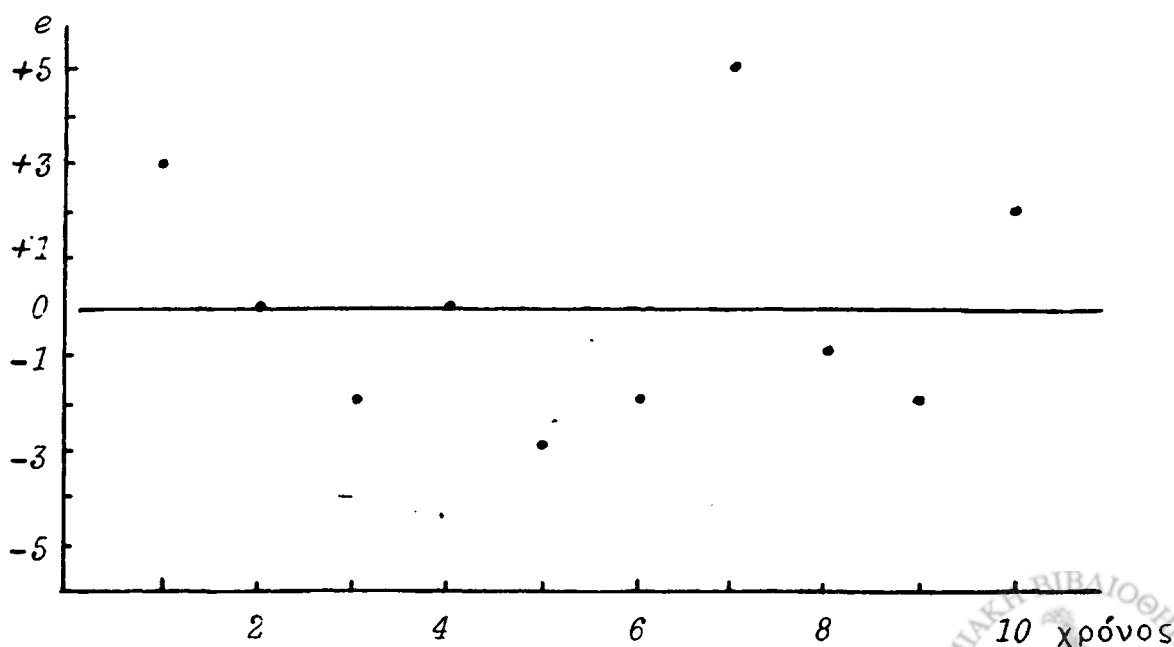
Όταν τα δεδομένα έχουν παρθεί κατά χρονολογική σειρά, είναι χρήσιμο να γίνει η γραφική παράσταση των υπολοίπων ως προς τον χρόνο παρ'ότι ο χρόνος δεν εκφράζεται στο μοντέλο (5.1). Πρόθεση μας είναι να διαπιστώσουμε αν υπάρχει καμιά μορφή σχέσης μεταξύ σφαλμάτων και χρόνου. Μία γραφική παράσταση των υπολοίπων ως προς τον χρόνο με μορφή όπως στο Σχήμα 5.10 δείχνει ότι υπάρχει επίδραση γραμμικής μορφής του χρόνου και



Ισωςθα έπρεπε να συμπεριληφθεί στο μοντέλο μία επιπλέον ανεξάρτητη μεταβλητή, ο χρόνος. Όταν τα σφάλματα είναι ανεξάρτητα τα υπόλοιπα παρουσιάζουν μία λίγο - πολύ τυχαία διασπορά γύρω από το μηδέν όπως στο Σχήμα 5.7α.



Σχήμα 5.10 Δυνατή μορφή γραφικής παράστασης των υπολοίπων ως προς τον χρόνο



Σχήμα 5.11 Γραφική παράσταση των υπολοίπων ως προς τον χρόνο, παράδειγμα ΕΒΟ

Η γραφική παράσταση των υπολοίπων ως προς τον χρόνο για τα δεδομένα του παραδείγματος της EBO (Πίνακας 5.3) δίνεται στο Σχήμα 5.11. Καμιά μορφή συστηματικής εξάρτησης μεταξύ των υπολοίπων και του χρόνου δεν φαίνεται να υπάρχει. Έτσι, για τα δεδομένα της EBO, δεν έχουμε ενδείξεις για απόκλιση από την υπόθεση της ανεξαρτησίας των σφαλμάτων.

5.11.4 Έλεγχος για την κανονικότητα των σφαλμάτων

Γενικά, μικρές αποκλίσεις από την υπόθεση της κανονικότητας των σφαλμάτων δεν προκαλούν ιδιαίτερα προβλήματα. Μεγάλες αποκλίσεις όμως θα έπρεπε να μας απασχολήσουν σοβαρά. Αποκλίσεις από την κανονικότητα των σφαλμάτων μπορούν να φανούν εξετάζοντας διάφορες γραφικές παραστάσεις των υπολοίπων. Το ιστόγραμμα των υπολοίπων ή η γραφική παράσταση των κανονικοποιημένων υπολοίπων σε χαρτί τυπικής κανονικής κατανομής μπορούν να δείξουν, αν υπάρχουν, τις μεγάλες αποκλίσεις από την κανονικότητα. Μια ακόμα δυνατότητα δίνεται από τα κανονικοποιημένα υπόλοιπα. Η υπόθεση της κανονικότητας των σφαλμάτων για το μοντέλο (5.1), $\varepsilon_i \sim N(0, \sigma^2)$, είναι ισοδύναμη με την υπόθεση ότι το 95% των ε_i / σ βρίσκεται μεταξύ του $-z_{0.025} = -1.96$ και $z_{0.025} = 1.96$. Έτσι, αν περίπου το 95% των $\varepsilon_i / \sqrt{MS_{res}}$ για τα δεδομένα που εξετάζουμε βρίσκονται στο διάστημα $(-2, 2)$ δεν έχουμε ενδείξεις για αποκλίσεις από την υπόθεση της κανονικότητας.

Για τα δεδομένα του παραδείγματος της EBO έχουμε βρει $MS_{res} = 7.5$ τα δε υπόλοιπα δίνονται στον Πίνακα 5.3. Τα κανονικοποιημένα υπόλοιπα που προκύπτουν,

$$1.095, 0, -0.730, 0, -1.095, -0.730, 1.826, -0.365, -0.730, 0.730$$

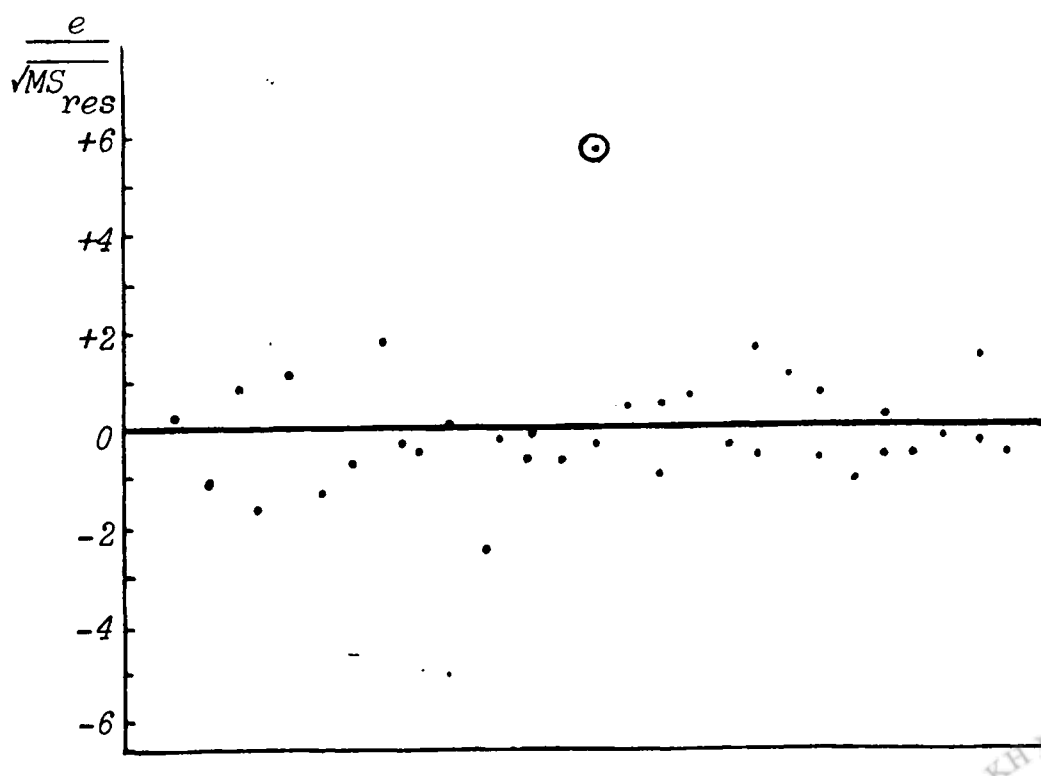
είναι όλα στο διάστημα $(-2, 2)$. Άρα, η υπόθεση της κανονικότητας των σφαλμάτων δεν παραβιάζεται.



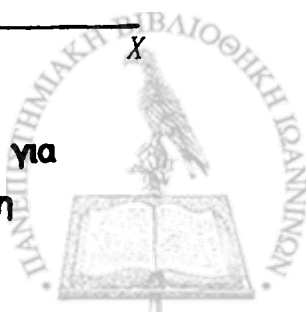
5.11.5 Έλεγχος για την παρουσία ακραίων παρατηρήσεων

Πολλές φορές είναι δυνατόν κάποια από τις τιμές Y_i των δεδομένων να είναι κατά πολύ μεγαλύτερη από τις άλλες με άμεση συνέπεια να επηρεάζει σε δυσανάλογα με τις άλλες μεγάλο βαθμό την μορφή της εκτιμώμενης γραμμής παλινδρόμησης. Η παρουσία ακραίων τιμών (παρατηρήσεων) μπορεί να φανεί αν γίνει η γραφική παράσταση των κανονικοποιημένων υπολοίπων ως προς την ανεξάρτητη μεταβλητή X . Παράδειγμα δεδομένων με μία ακραία τιμή (αυτήν στον κύκλο) δίνεται στο Σχήμα 5.12. Σημειώστε ότι το υπόλοιπο αυτό αντιστοιχεί σε παρατήρηση Y_i περίπου 6 τυπικές αποκλίσεις από την εκτιμώμενη τιμή $\hat{Y}_i$.

Η πρώτη εκτίμηση που συνήθως γίνεται όταν εντοπίζεται μία ακραία τιμή είναι ότι αυτή ίσως οφείλεται είτε σε λάθος, είτε σε απρόσμενο γεγονός και συνεπώς θα έπρεπε να αγνοηθεί.



Σχήμα 5.12 Γραφική παράσταση υπολοίπων για δεδομένα με ακραία παρατήρηση



Από την άλλη μεριά όμως η ακραία αυτή παρατήρηση θα μπορούσε να περιέχει σημαντική πληροφορία, αν δεν αληθεύουν οι παραπάνω εκδοχές, όπως π.χ. όταν η ακραία τιμή οφείλεται σε κάποια αλληλεπίδραση με κάποια άλλη ανεξάρτητη μεταβλητή που δεν συμπεριλαμβάνεται στο μοντέλο. Έχει επικρατήσει, σαν κανόνας, να αγνοείται η ακραία παρατήρηση μόνον όταν αυτή αποδειγμένα αντιπροσωπεύει λάθος αντιγραφής, υπολογιστικό λάθος, κακολειτουργία κάποιου εξαρτήματος ή παρόμοιες καταστάσεις.

5.12 Μέτρα επαναφοράς στην καταλληλότητα του μοντέλου

Αποκλίσεις από τις υποθέσεις του απλού γραμμικού μοντέλου μπορούν να αντιμετωπιστούν είτε με την αντικατάσταση του μοντέλου (5.1) με κάποιο καταλληλότερο αν και πολλές φορές πιο πολύπλοκο μοντέλο, είτε με τη χρησιμοποίηση κάποιου μετασχηματισμού στα δεδομένα έτσι ώστε το μοντέλο (5.1) να καθίσταται κατάλληλο για τα προκύπτοντα δεδομένα. Εδώ θα περιγράψουμε περιληπτικά δύο από τους πιο βασικούς μετασχηματισμούς που χρησιμοποιούνται όταν διαπιστωθούν αποκλίσεις από την γραμμικότητα της συνάρτησης παλινδρόμησης. Τον Λογαριθμικό και τον Αντίστροφο μετασχηματισμό. Λεπτομέρειες γι'αυτήν την περίπτωση όπως και μεθοδολογίες για την αντιμετώπιση αποκλίσεων από τις υποθέσεις της σταθερότητας της διακύμανσης ή της κανονικότητας και ανεξαρτησίας των σφαλμάτων δίνονται, μεταξύ άλλων, στα βιβλία των Neter and Wasserman (1975) και Draper and Smith (1981).

5.12.1 Λογαριθμικός μετασχηματισμός

Ο Λογαριθμικός μετασχηματισμός είναι χρήσιμος όταν το πραγματικό μοντέλο είναι πολλαπλασιαστικό, για



παράδειγμα της μορφής

$$Y = \gamma_0 \gamma_1^X \varepsilon \quad (5.20)$$

όπου, Y εξαρτημένη μεταβλητή, X ανεξάρτητη (ελεγχόμενη) μεταβλητή, γ_0, γ_1 οι παράμετροι του μοντέλου και ε μία τυχαία μεταβλητή, το σφάλμα, με $E(\varepsilon)=1$ και $\text{Var}(\varepsilon)=\sigma^2$. Το μοντέλο (5.20) δεν είναι γραμμικό ούτε ως προς τις παραμέτρους ούτε ως προς X . Παρόλα αυτά μπορεί να μελετηθεί με βάση το γραμμικό μοντέλο (5.1) αν χρησιμοποιηθεί ο κατάλληλος μετασχηματισμός. Εφαρμόζοντας τον μετασχηματισμό $Y' = \log_{10} Y$ στο μοντέλο (5.20) παίρνουμε

$$Y' = \beta_0 + \beta_1 X + \varepsilon' \quad (5.21)$$

όπου, $\beta_0 = \log_{10} \gamma_0$, $\beta_1 = \log_{10} \gamma_1$ και $\varepsilon' = \log_{10} \varepsilon$.

Έτσι, η ανάλυση δεδομένων (X_i, Y_i) με το μοντέλο (5.20) είναι ισοδύναμη με την ανάλυση των μετασχηματισμένων δεδομένων $(X_i, \log_{10} Y_i)$ με το μοντέλο (5.21). Τέτοια πολλαπλασιαστικά μοντέλα βρίσκουν πολλές εφαρμογές, μεταξύ άλλων στις Επιχειρηματικές και Οικονομικές επιστήμες.

5.12.2 Αντίστροφος μετασχηματισμός

Ο αντίστροφος μετασχηματισμός χρησιμοποιείται είτε για να μετασχηματίσει κάποιο μη γραμμικό μοντέλο σε γραμμικό, είτε γιατί ενώ κανένα μοντέλο δεν έχει επιλεγεί, το διάγραμμα διασποράς των δεδομένων παρέχει ενδείξεις καμπυλόγραμμης σχέσης.

Για παράδειγμα το μοντέλο

$$Y = \beta_0 + \beta_1 / X + \varepsilon$$

που δεν είναι γραμμικό ως προς X , με τον μετασχηματισμό $X' = 1/X$



παίρνει την μορφή του απλού γραμμικού μοντέλου

$$Y = \beta_0 + \beta_1 X' + \varepsilon'$$

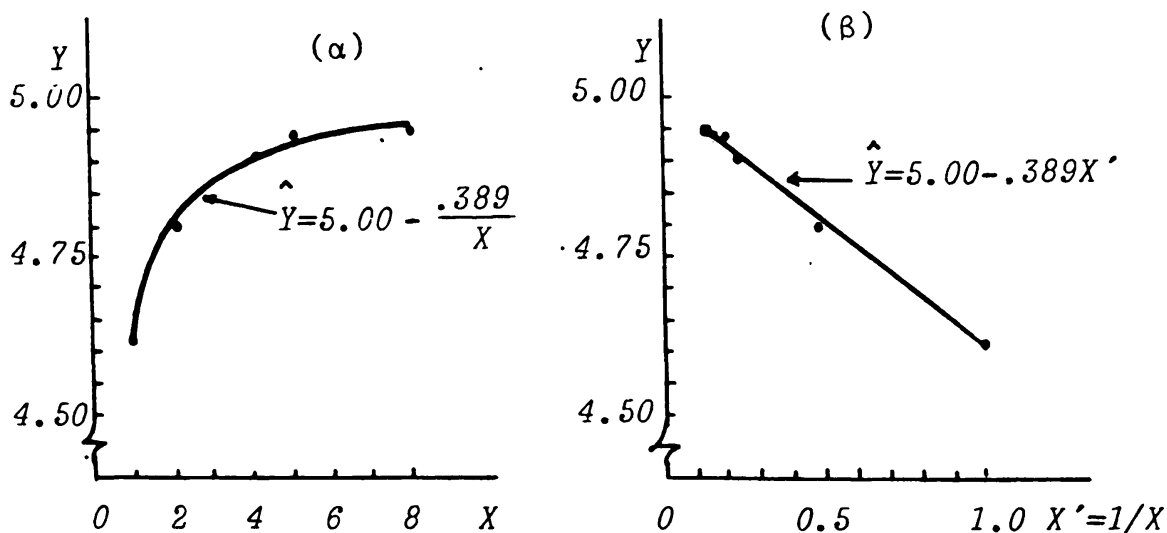
Άλλα απλά γραμμικά μοντέλα που περιέχουν αντίστροφους μετασχηματισμούς είναι τα

$$Y' = \beta_0 + \beta_1 X + \varepsilon' \quad \text{όπου} \quad Y' = 1/Y$$

και

$$Y' = \beta_0 + \beta_1 X' + \varepsilon' \quad \text{όπου} \quad Y' = 1/Y, \quad X' = 1/X.$$

Ένα παράδειγμα αντίστροφου μετασχηματισμού δίνεται στο Σχήμα 5.13. Το Σχήμα 5.13α αποτελεί το διάγραμμα διασποράς των δεδομένων (X_i, Y_i) όπου X χρόνια εμπειρίας και Y κέρδος ανά ώρα πέντε υπαλλήλων κάποιου καταστήματος που κατασκευάζει περρούκες κατά παραγγελία. Τα δεδομένα παρέχουν ενδείξεις ότι η σχέση είναι καμπυλόγραμμη. Το Σχήμα 5.13β δίνει το διάγραμμα διασποράς των δεδομένων $(X_i' = 1/X_i, Y_i)$ με τα σημεία να κείνται σχεδόν επί ευθείας. Η εκτιμώμενη συνάρτηση παλινδρόμησης, για τα μετασχηματισμένα δεδομένα, είναι $\hat{Y} = 5.00 - 0.389X'$ και δίνεται στο Σχήμα 5.13β.



Σχήμα 5.13 Γραφικές παραστάσεις αντίστροφου μετασχηματισμού

Έτσι, η εκτιμώμενη συνάρτηση των αρχικών δεδομένων είναι $\hat{Y} = 5.00 - 0.389/X$ και δίνεται στο Σχήμα 5.13α.



5.13 Πολλαπλή γραμμική παλινδρόμηση

Συχνά υπάρχει η ένδειξη ότι οι τιμές μίας μεταβλητής επηρεάζονται από περισσότερες από μία ελεγχόμενες μεταβλητές. Για παράδειγμα, η ετήσια παραγωγή πορτοκαλιών μίας περιοχής θα μπορούσε να επηρεάζεται τόσο από την μέση ανοιξιόατη βροχή της περιοχής όσο και από την μέση της θερμοκρασία. Η γενίκευση του απλού γραμμικού μοντέλου που μελετήσαμε στα προηγούμενα επιτρέπει την ταυτόχρονη ύπαρξη $k \geq 2$ ανεξάρτητων μεταβλητών στο μοντέλο και αποτελεί το γενικό γραμμικό μοντέλο που δίνεται από την σχέση

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + \varepsilon_i \quad (5.22)$$

όπου, $Y_i, i=1,2,\dots,n$, οι τιμές της εξαρτημένης μεταβλητής, $X_{ji}, j=1,2,\dots,k, i=1,2,\dots,n$ οι τιμές των ανεξάρτητων (ελεγχόμενων) μεταβλητών για τις οποίες υποθέτουμε ότι δεν υπόκεινται σε σφάλματα, $\beta_0, \beta_1, \dots, \beta_k$ οι άγνωστες παράμετροι του μοντέλου και, $\varepsilon_i, i=1,2,\dots,n$, τα σφάλματα για τα οποία υποθέτουμε ότι είναι ανά δύο ασυσχέτιστες τ.μ με $E(\varepsilon_i)=0$.

Η σχετική με το γενικό γραμμικό μοντέλο στατιστική μεθοδολογία καλείται ανάλυση πολλαπλής γραμμικής Παλινδρόμησης και δίνει απαντήσεις σε ερωτήματα όπως το ποιές από τις X_j είναι οι πλέον απαραίτητες στο μοντέλο όταν ενδιαφερόμαστε για προβλέψεις της Y που γενικά, για δεδομένες τιμές X_{j0} των μεταβλητών $X_j, j=1,2,\dots,k$, δίνονται από την

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_{10} + \dots + \hat{\beta}_k X_{k0}$$

όπου, $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ οι εκτιμητές ελαχίστων τετραγώνων των $\beta_0, \beta_1, \dots, \beta_k$ αντίστοιχα.

Όλα τα μοντέλα που είναι γραμμικά ως προς τις παραμέτρους μπορούν να πάρουν την μορφή του μοντέλου (5.22). Για παράδειγμα, το γνωστό σαν πολυωνυμικό μοντέλο

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k + \varepsilon$$



είναι ισοδύναμο με το μοντέλο (5.22) αρκεί να θέσουμε $X_1 = X$, $X_2 = X^2, \dots, X^k = X_k$.

5.14 Συσχέτιση

Στα προηγούμενα μελετήσαμε την εξάρτηση μίας τ.μ. Y από μία ανεξάρτητη μεταβλητή X την οποία θεωρήσαμε ελεγχόμενη. Είναι όμως δυνατόν να είναι τ.μ. τόσο το Y όσο και το X . Εκτός αυτού, συχνά χρειαζόμαστε, για παράδειγμα στην περίπτωση της πολλαπλής παλινδρόμησης, κάποια απλή μεθοδολογία η οποία να δίνει τον βαθμό γραμμικής εξάρτησης δύο ποσοτικών τ.μ. χωρίς να υπεισέρχεται σε λεπτομέρειες της Παλινδρόμησης.

Ένα μέτρο της συμμεταβλητότητας δύο τ.μ. X και Y είναι η συνδιακύμανση ή συνδιασπορά που συμβολίζεται με $\text{Cov}(X, Y)$ και ορίζεται από τη σχέση

$$\text{Cov}(X, Y) = E[(X - EX)(Y - EY)] = EXY - EXEY.$$

Η συνδιακύμανση έχει σαν μονάδες μέτρησης το γινόμενο των μονάδων μέτρησης των X και Y και είναι θετική όταν τα $X - EX$ και $Y - EY$ τείνουν να έχουν το ίδιο σημείο με μεγάλη πιθανότητα και αρνητική όταν τα $X - EX$ και $Y - EY$ τείνουν να έχουν αντίθετο σημείο με μεγάλη πιθανότητα.

Για δύο τ.μ. X και Y ο "θεωρητικός" συντελεστής συσχέτισης συμβολίζεται με $\rho(X, Y)$ ή απλά ρ και ορίζεται από τη σχέση

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}X \text{ Var}Y}}.$$

Ο συντελεστής συσχέτισης αφαιρεί, κατά κάποιο τρόπο, την ατομική μεταβλητότητα των X και Y διαιρώντας τη συνδιακύμανση με το γινόμενο των τυπικών αποκλίσεων των X και Y . Είναι καθαρός



αριθμός μεταξύ του -1 και 1 και επιπλέον είναι ένα μέτρο του βαθμού της γραμμικής συσχέτισης μεταξύ των τ.μ. X και Y . Όταν $\rho=0$ τότε δεν υπάρχει γραμμική σχέση μεταξύ των X και Y (είναι δυνατόν όμως τα X και Y να συσχετίζονται διαφορετικά, π.χ. καμπυλόγραμμα) ενώ όταν $\rho=\pm 1$ οι τ.μ. X και Y συσχετίζονται γραμμικά με τέλεια σχέση. Σημειώστε ότι όταν $\rho(X,Y)=0$, οι τ.μ. X και Y λέγονται ασυσχέτιστες. Επίσης αν οι ασυσχέτιστες τ.μ. X , Y είναι κανονικές, τότε είναι και ανεξάρτητες (βλέπε π.χ. Mood, Graybill and Boes, 1974).

Ο "θεωρητικός" συντελεστής συσχέτισης ρ είναι δύσκολο να υπολογιστεί, έτσι στην πράξη εκτιμάται από τον δειγματικό συντελεστή συσχέτισης του Pearson, που συμβολίζεται με $r(X,Y)$ ή απλά r και για n ζεύγη (X_i, Y_i) αριθμητικών τιμών των τ.μ. X και Y ορίζεται από τη σχέση

$$r(X,Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} .$$

Ο δειγματικός συντελεστής συσχέτισης r είναι ένας καθαρός αριθμός μεταξύ του -1 και 1 και το τετράγωνο του μας δίνει τον συντελεστή προσδιορισμού γιατί,

$$\begin{aligned} r^2 &= \frac{(\sum (X_i - \bar{X})(Y_i - \bar{Y}))^2}{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2} = \frac{\hat{\beta}_1 \sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} \\ &= \frac{SS_{reg}}{SS_{tot}} = R^2 . \end{aligned}$$

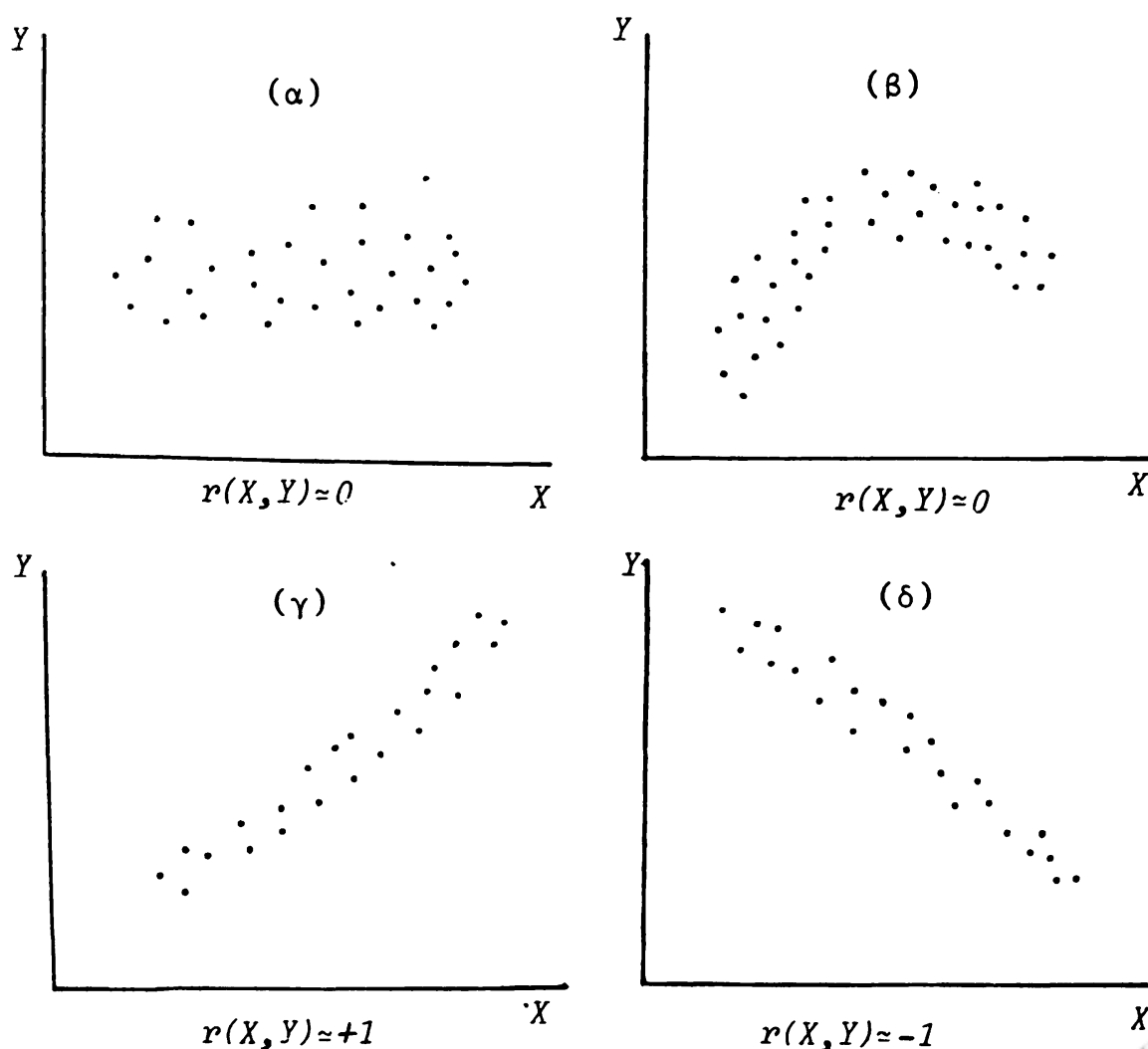
Συνήθως στην πράξη, και για να αποκτήσουμε μία εποπτική εικόνα του προβλήματος κάνουμε πρώτα την γραφική παράσταση των δεδομένων που πρόκειται να αναλύσουμε, δηλαδή το διάγραμμα διασποράς. Στο Σχήμα 5.14 δίνονται μερικές από τις δυνατές μορφές που μπορεί να παρουσιάσει ένα διάγραμμα διασποράς σε σχέση με τον δειγματικό συντελεστή συσχέτισης για κάποιες

ενδεικτικές τιμές του $r(X,Y)$.

Όταν $r(X,Y) \approx 0$ δεν υπάρχει γραμμική σχέση μεταξύ των X και Y (Σχήμα 5.14α), χωρίς βέβαια αυτό να αποκλείει την ύπαρξη κάποιας σχέσης άλλης μορφής, για παράδειγμα καμπυλόγραμμης (Σχήμα 5.14β).

Όταν $r(X,Y) \approx +1$ τότε υπάρχει θετική (Σχήμα 5.16γ) ενώ όταν $r(X,Y) \approx -1$ αρνητική (Σχήμα 5.14δ) γραμμική εξάρτηση μεταξύ των τ.μ. X και Y .

Αν υποθέσουμε ότι οι τ.μ. X και Y έχουν κανονικές κατανομές τότε μπορεί να αποδειχθεί (βλέπε πχ. Mood, Graybill and Boes, 1974) ότι, όταν $\rho=0$, η στατιστική συνάρτηση $t=r\sqrt{n-2}/\sqrt{1-r^2}$ έχει την t_{n-2} κατανομή.



Σχήμα 5.14 Παραδείγματα διαγράμματος διασποράς και αντίστοιχου δειγματικού συντελεστή συσχέτισης



Έτσι, για τον έλεγχο της $H_0: \rho=0$ έναντι της $H_a: \rho \neq 0$, με επίπεδο σημαντικότητας α , η αρχική υπόθεση απορρίπτεται όταν $|t| > t_{\alpha/2, n-2}$.

Για ελέγχους υποθέσεων της μορφής $H_0: \rho = \rho_0$ έναντι $H_a: \rho \neq \rho_0$, χρησιμοποιούμε τον γνωστό σαν μετασχηματισμό του Fisher, $Z = \frac{1}{2} \ln \frac{1+r}{1-r}$. Αν υποθέσουμε ότι οι τ.μ. X και Y έχουν κανονικές κατανομές, μπορεί να αποδειχθεί (βλέπε π.χ. Kendall and Stuart, 1977) ότι, όταν $\rho = \rho_0$, η τ.μ. Z είναι ασυμπτωτικά κανονική με

$$E(Z) = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} \text{ και } \text{Var}(Z) = \frac{1}{n-3}.$$

Έτσι, η στατιστική συνάρτηση $W = \frac{Z - E(Z)}{\sqrt{\text{Var}(Z)}}$ έχει την $N(0, 1)$ κατανομή και με επίπεδο σημαντικότητας α , η $H_0: \rho = \rho_0$ απορρίπτεται όταν $|W| > z_{\alpha/2}$. Έχει βρεθεί ότι η προσέγγιση που επιτυγχάνεται είναι ικανοποιητική όταν $n > 50$.

Για τα δεδομένα της EBO βρίσκουμε

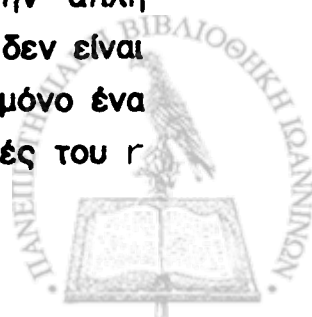
$$r = 6800 / \sqrt{3400 \times 13660} = 0.9978$$

και

$$t = 0.9978 \sqrt{8} / \sqrt{1-0.9956} = 42.6$$

Επειδή $t = 42.6 > t_{0.005, 8} = 3.355$ η αρχική υπόθεση $\rho = 0$ απορρίπτεται σε επίπεδο σημαντικότητας $\alpha = 0.01$. Την ίδια υπόθεση μπορούμε να ελέγξουμε χρησιμοποιώντας το ασυμπτωτικό W -τεστ με $\rho_0 = 0$. Τα δεδομένα της EBO δίνουν $Z = \frac{1}{2} \ln \frac{1.9978}{0.0022} = 3.4057$, $E(Z) = 0$ και $\text{Var}(Z) = 1/7$. Έτσι $W = 9.01$ και επειδή $W = 9.01 > z_{0.005} = 2.58$, η αρχική υπόθεση $\rho = 0$ απορρίπτεται σε επίπεδο σημαντικότητας $\alpha = 0.01$.

Η συσχέτιση είναι στενά συνδεδεμένη με την απλή γραμμική παλινδρόμηση από την σχέση $r^2 = R^2$. Όταν το X δεν είναι τ.μ. δεν μπορούμε να μιλάμε για συσχέτιση και το r^2 είναι μόνο ένα μέτρο καταλληλότητας της προσαρμογής. Σημειώστε ότι τμές του r



κοντά στο 0.5 δεν είναι τίποτα το σπουδαίο. Όταν $r=0.5$ τότε $r^2=R^2=0.25$ δηλαδή μόνο το 25% της μεταβλητότητας της Y οφείλεται στην παλινδρόμηση.

5.15 Οι ποιοτικές μεταβλητές στην παλινδρόμηση

Στα προηγούμενα μελετήσαμε σχέσεις μεταξύ ποσοτικών μεταβλητών, δηλαδή μεταβλητών που παίρνουν συνεχείς τιμές σε κάποιο διάστημα όπως είναι το εισόδημα, η ηλικία, η θερμοκρασία κλπ.

Συχνά μεταβλητές που εμφανίζονται στους χώρους των επιχειρηματικών και οικονομικών επιστημών είναι ποιοτικές, δηλαδή μεταβλητές που δεν μπορούν να μετρηθούν και αναφέρονται σε καταστάσεις, είδη κλπ. όπως είναι π.χ. το φύλο των ανθρώπων (αρσενικό, θηλυκό), η οικονομικοκοινωνική κατάσταση (φτωχός, πλούσιος) ή η σωματική ικανότητα (ικανός, σχετικά ικανός, ανίκανος). Οι ποιοτικές μεταβλητές μπορούν να συμπεριληφθούν για μελέτη σε ένα μοντέλο παλινδρόμησης με την βοήθεια των λεγόμενων δείκτριων (indicator, dummy) μεταβλητών. Σε κάθε ποιοτική μεταβλητή αντιστοιχούμε μία δείκτρια μεταβλητή, δηλαδή μεταβλητή η οποία παίρνει κάποια ενδεικτική τιμή για κάθε μία από τις κατηγορίες της. Ας υποθέσουμε, για παράδειγμα, ότι τα δεδομένα που εξετάζονται προέρχονται από δύο διαφορετικούς τύπους μηχανών, έστω Α και Β. Η ποιοτική μεταβλητή "τύπος μηχανών" μπορεί να συμπεριληφθεί σε ένα μοντέλο παλινδρόμησης με σκοπό να εξακριβωθεί τυχόν επίδραση της στα αποτελέσματα με την βοήθεια μίας δείκτριας μεταβλητής. Για τον σκοπό αυτό εισάγουμε την δείκτρια μεταβλητή X και κάποιον συντελεστή παλινδρόμησης β έτσι ώστε ο πρόσθετος όρος βX εμφανίζεται στο μοντέλο. Στην δείκτρια μεταβλητή X θα μπορούσαν να δοθούν οι παρακάτω τιμές:

$X=0$ αν τα δεδομένα προέρχονται από την μηχανή Α,

$X=1$ αν τα δεδομένα προέρχονται από την μηχανή Β.



Οποιοσδήποτε δύο διακεκριμένες τιμές για την X θα ήταν κατάλληλες αν και οι παραπάνω είναι συνήθως, αλλά όχι πάντα, οι καλύτερες.

Μοντέλα στα οποία όλες οι ανεξάρτητες μεταβλητές είναι ποιοτικές ονομάζονται μοντέλα ανάλυσης της διακύμανσης. Ειδικότερα δε, η ανάλυση της απλής γραμμικής παλινδρόμησης με ποιοτική ανεξάρτητη μεταβλητή ταυτίζεται με την κατά ένα παράγοντα ανάλυση της διακύμανσης που μελετάται στο επόμενο κεφάλαιο.

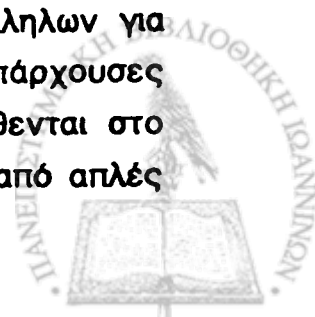
Μοντέλα στα οποία υπάρχουν ποσοτικές και ποιοτικές ανεξάρτητες μεταβλητές ονομάζονται μοντέλα ανάλυσης της συνδιακύμανσης και η μελέτη τους βασίζεται στη μεθοδολογία της πολλαπλής γραμμικής παλινδρόμησης.

Η μελέτη των παραπάνω μοντέλων όπως και μοντέλων στα οποία η εξαρτημένη μεταβλητή είναι ποιοτική ξεφεύγουν από τον σκοπό ενός εισαγωγικού βιβλίου. Λεπτομέρειες δίνονται, μεταξύ άλλων, στα βιβλία των Neter and Wasserman (1975) και Draper and Smith (1981).

5.16 Στατιστικά πακέττα και παλινδρόμηση

Οι υπολογισμοί που υπεισέρχονται στην Ανάλυση της παλινδρόμησης είναι συνήθως επίπονοι όταν ο αριθμός των παρατηρήσεων είναι μεγάλος ή όταν υπάρχουν πολλές ανεξάρτητες μεταβλητές (πολλαπλή γραμμική παλινδρόμηση). Σήμερα, με την ύπαρξη των ηλεκτρονικών υπολογιστών, τέτοιοι υπολογισμοί μπορούν να γίνουν εύκολα και γρήγορα μετά την δημιουργία κατάλληλων προγραμμάτων ηλεκτρονικού υπολογιστή.

Η ανάγκη για ευρεία χρήση των Στατιστικών μεθοδολογιών σε ανάλυση δεδομένων, ακόμα και από επιστήμονες χωρίς ιδιαίτερες γνώσεις στη Στατιστική ή στον προγραμματισμό, οδήγησε διάφορες ερευνητικές ομάδες στη δημιουργία προγραμμάτων κατάλληλων για ανάλυση δεδομένων με τις σπουδαιότερες από τις υπάρχουσες Στατιστικές μεθοδολογίες. Τέτοια προγράμματα που διατίθενται στο εμπόριο, καλύπτουν ένα πολύ ευρύ χώρο της Στατιστικής, από απλές



γραφικές παραστάσεις μέχρι προχωρημένες και πολύπλοκες Στατιστικές μεθοδολογίες (ανάλυση μοντέλων απλής και πολλαπλής παλινδρόμησης, διακύμανσης, συνδιακύμανσης κλπ.) και περιέχονται, μεταξύ άλλων, στα στατιστικά πακέττα BMDP (Biomedical Programs) και SPSS (Statistical Package for the Social Sciences) που έχουν στην διάθεσή τους οι χρήστες του ηλεκτρονικού υπολογιστή του Πανεπιστημίου Ιωαννίνων. Η χρήση των προγραμμάτων που περιέχονται στα στατιστικά πακέττα BMDP και SPSS είναι απλή και δεν απαιτεί ιδιαίτερες γνώσεις Στατιστικής ή προγραμματισμού. Στο ενημερωτικό - βοηθητικό εγχειρίδιο κάθε στατιστικού πακέττου περιγράφονται οι Στατιστικές μεθοδολογίες στις οποίες βασίζεται κάθε ένα από τα περιεχόμενα προγράμματα. Έτσι, μετά την εισαγωγή στον ηλεκτρονικό υπολογιστή τόσο των δεδομένων όσο και του προγράμματος του αντίστοιχου της Στατιστικής μεθοδολογίας που επιλέχτηκε, τα αποτελέσματα των υπολογισμών δίνονται ταχύτατα από τον ηλεκτρονικό υπολογιστή.

Στις προηγούμενες παραγράφους οι διάφοροι υπολογισμοί που υπεισέρχονται στην μεθοδολογία της απλής γραμμικής παλινδρόμησης χρησιμοποιήθηκαν για την Στατιστική ανάλυση των δεδομένων της ΕΒΟ. Εδώ θα δώσουμε τα αποτελέσματα για τα δεδομένα της ΕΒΟ όπως πάρθηκαν από τον ηλεκτρονικό υπολογιστή Cyber 171 του Πανεπιστημίου Ιωαννίνων χρησιμοποιώντας το πρόγραμμα BMDP1R που αντιστοιχεί στη μεθοδολογία της απλής και πολλαπλής γραμμικής παλινδρόμησης.

Τα αριθμητικά αποτελέσματα που ακολουθούν έχουν παραχθεί με χρήση ηλεκτρονικού υπολογιστή και περιέχουν:

- (i) Τα δεδομένα (X_i, Y_i) , τις προβλεπόμενες τιμές $\hat{Y}_i$ και τα υπόλοιπα e_i (§ 5.5).
- (ii) Τα $\hat{\beta}_0$, $\hat{\beta}_1$ και το t-τεστ για τον έλεγχο της σημαντικότητας του β_1 (§ 5.5, § 5.7).
- (iii) Τον συντελεστή προσδιορισμού R^2 (§ 5.10).
- (iv) Τον πίνακα ANADIA και το F-τεστ για τον έλεγχο της σημαντικότητας της παλινδρόμησης (§ 5.10).



(v) Την γραφική παράσταση των υπολοίπων e_i ως προς την ανεξάρτητη μεταβλητή X (§ 5.11).

(vi) Τον δειγματικό συντελεστή συσχέτισης $r(X,Y)$ (§ 5.14).

ΠΑΡΑΔΕΙΓΜΑ ΕΒΟ

(i)

LIST OF PREDICTED VALUES, RESIDUALS, AND VARIABLES

CASE		RESIDUAL	PREDICTED VALUE	VARIABLES	
LABEL	NO.			1 Y	2 X
	1	3.000	70.0	73.0	30.00
	2	.6821E-12	60.0	50.0	20.00
	3	-2.000	130.0	128.0	60.00
	4	0.	170.0	170.0	80.00
	5	-3.000	90.0	87.0	40.00
	6	-2.000	110.0	108.0	50.00
	7	5.000	130.0	135.0	60.00
	8	-1.000	70.0	69.0	30.00
	9	-2.000	150.0	148.0	70.00
	10	2.000	130.0	132.0	60.00

(ii)

VARIABLE	COEFFICIENT	STD. ERROR	STD.REG. COEFF	T	P(2 TAIL)
INTERCEPT	10.00000				
X	2.00000	.04697	.998	42.583	0.0000

(iii)

REGRESSION TITLE IS

EXAMPLE EBO

DEPENDENT VARIABLE. 1 Y

TOLERANCE.0100

ALL DATA CONSIDERED AS A SINGLE GROUP

MULTIPLE R .9978 STD.ERROR OF EST. 2.7386

MULTIPLE R-SQUARE .9956

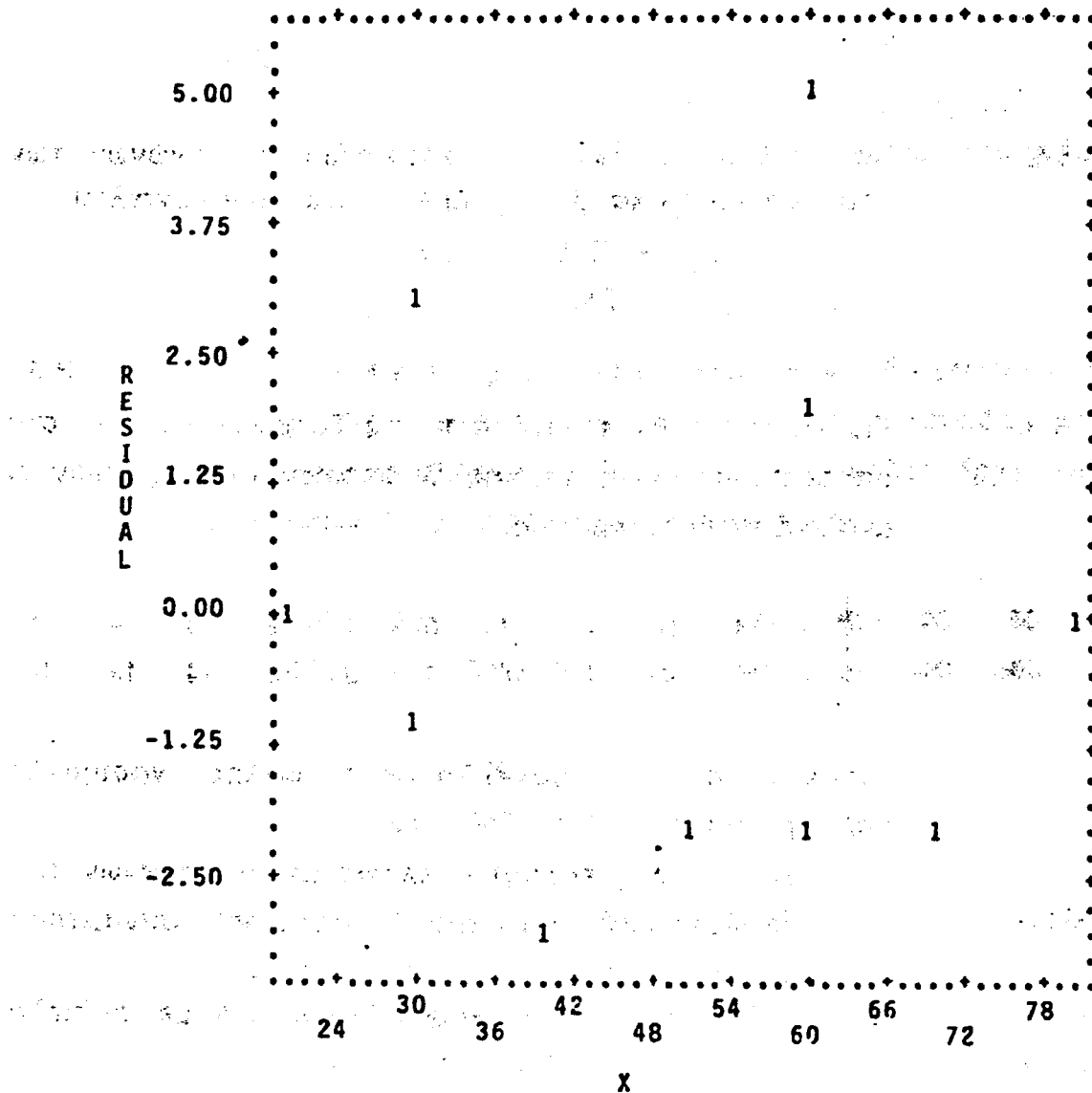
(iv)

ANALYSIS OF VARIANCE

	SUM OF SQUARES	DF	MEAN SQUARE	F RATIO
REGRESSION	13600.0000	1	13600.0000	1813.333
RESIDUAL	60.0000	8	7.5000	



(v)



(vi)

CORRELATION MATRIX

	Y	X
Y	1	1.0000
X	2	.9978



ΑΣΚΗΣΕΙΣ

5.1 Για το απλό γραμμικό μοντέλο να βρεθεί η $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$ και η $\text{Cov}(\bar{Y}, \hat{\beta}_1)$.

5.2 Να βρεθούν οι εκτιμητές ελαχίστων τετραγώνων των παραμέτρων β_0 και β_1 του απλού γραμμικού μοντέλου

$$Y_i = \beta_0 + \beta_1(X_i - \bar{X}) + \varepsilon_i, \quad i=1, 2, \dots, n.$$

5.3 Για να μελετηθεί η επίδραση της θερμοκρασίας (X) στα αποτελέσματα μίας χημικής διαδικασίας (Y) πάρθηκαν οι ακόλουθες παρατηρήσεις:

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

Υιοθετώντας το μοντέλο της απλής γραμμικής παλινδρόμησης,

- Βρείτε τους εκτιμητές ελαχίστων τετραγώνων των παραμέτρων και την εκτιμώμενη συνάρτηση παλινδρόμησης.
- Ελέγξατε την υπόθεση ότι $\beta_1 = 0$ με επίπεδο σημαντικότητας $\alpha = 0.05$.
- Βρείτε ένα 95% διάστημα εμπιστοσύνης (Δ.Ε.) για το β_1 .
- Βρείτε ένα 95% Δ.Ε. για την μέση τιμή του Y όταν $X = 3$.
- Υπάρχουν ενδείξεις για την υιοθέτηση άλλου καταλληλότερου μοντέλου;

5.4 Τα παρακάτω δεδομένα αποτελούν παρατηρήσεις της ποσότητας ύδατος (X) και της αντίστοιχης παραγόμενης χλόης (Y) σε Kgr κάποιου πειραματικού σταθμού.



X	12	18	24	30	36	42	48
Y	5.27	5.68	6.25	7.21	8.02	8.71	8.02

Υιοθετώντας το απλό γραμμικό μοντέλο,

- (i) Βρείτε την εκτιμώμενη συνάρτηση παλινδρόμησης.
- (ii) Ελέγξατε την υπόθεση ότι $\beta_1 = 0$ με επίπεδο σημαντικότητας $\alpha = 0.05$.
- (iii) Βρείτε ένα 95% Δ.Ε. για το β_1 .

5.5

Για να βρεθεί η σχέση μεταξύ εξόδων για διαφήμιση (X) και εσόδων από πωλήσεις (Y) ανά βδομάδα, τα παρακάτω αποτελούν τα δεδομένα (σε χιλιάδες δρχ) των τελευταίων 12 βδομάδων μίας επιχείρησης.

X	40	20	25	20	30	50	40	20	50
Y	365	400	395	365	475	440	490	420	560

X	40	25	50
Y	525	480	510

Υποθέτοντας κατ'αρχήν ότι το απλό γραμμικό μοντέλο είναι κατάλληλο,

- (i) Να βρεθεί η εκτιμώμενη συνάρτηση παλινδρόμησης που θα δίνει τις ανα βδομάδα πωλήσεις όταν θα δίνονται τα έξοδα που δαπανήθηκαν για διαφήμιση.
- (ii) Να γίνει έλεγχος της σημαντικότητας της παλινδρόμησης, $\alpha = 0.05$.
- (iii) Να εξεταστεί η καταλληλότητα του μοντέλου.

5.6

Σε 8 ομάδες των 25 ποντικιών παρατηρήθηκαν οι παρακάτω θάνατοι (Y) όταν τους δόθηκαν οι δόσεις (X) κάποιου δηλητηρίου (σε mg).



Ομάδα		1	2	3	4	5	6	7	8
Δόση	X	4	6	8	10	12	14	16	18
Θάνατοι	Y	1	3	6	8	14	16	20	21

Υιοθετώντας το μοντέλο της απλής παλινδρόμησης,

- Βρείτε την εκτιμώμενη γραμμή παλινδρόμησης.
- Εκτιμήσατε τον αριθμό των ποντικών που θα ψοφήσουν αν σε ομάδα 25 ποντικών δοθεί μία δόση 7 mg από το ίδιο δηλητήριο.

5.7 Σε 10 παιδιά ηλικίας 10-15 ετών μετρήθηκε το βάρος Y σε Kg και το ύψος X σε cm.

X	130	148	154	145	162	158	152	144	138	164
Y	34	42	45	41	46	49	43	46	36	50

Υιοθετώντας το απλό γραμμικό μοντέλο,

- Να βρεθεί η εκτιμώμενη συνάρτηση παλινδρόμησης του βάρους σε σχέση με το ύψος.
- Να εξεταστεί αν το βάρος είναι ανεξάρτητο του ύψους.
- Να δοθεί ένα 95% Δ.Ε. για το βάρος ενός παιδιού ύψους 115 cm.

5.8 Για να προσδιοριστεί η επίδραση του ναρκωτικού εφεδρίνη στους παλμούς της καρδιάς, χορηγείται το φάρμακο αυτό σε έναν ασθενή για 6 μέρες και σε διαφορετικές δόσεις. Τα αποτελέσματα συνοψίζονται παρακάτω όπου X ημερήσια δόση εφεδρίνης σε χάπια και Y αριθμός παλμών ανά λεπτό:

X	3	2	1	3	5
Y	70	60	50	80	100



Υιοθετώντας το μοντέλο της απλής γραμμικής παλινδρόμησης,

- (i) Να βρεθεί η εκτιμώμενη συνάρτηση παλινδρόμησης.
- (ii) Να γίνει το διάγραμμα διασποράς και να σχεδιαστεί η εκτιμώμενη γραμμή παλινδρόμησης.
- (iii) Να υπολογιστούν η διακύμανση των υπολοίπων και ο συντελεστής συσχέτισης.

Τι συμπεράσματα βγάξετε;

5.9

Τα παρακάτω δεδομένα αποτελούν μετρήσεις του ποσού λιπάσματος X (σε Kgr / στρέμμα) που ένας αγρότης χρησιμοποίησε σε 33 χωράφια της ίδιας περιοχής και της απόδοσης Y σε σιτάρι (τόνοι / στρέμμα).

X	112	92	72	66	112	88	42	126	72	52	28
Y	33	28	38	17	35	31	8	37	32	20	17

X	132	23	57	111	69	19	103	141	77	37	23
Y	36	14	25	40	29	12	27	40	26	27	9

X	142	37	127	88	48	61	71	113	88	44	77
Y	38	13	23	31	37	25	14	26	24	17	32

Αν υποθέσουμε ότι τα δεδομένα προέρχονται από κανονικές κατανομές ελέγξατε αν το ρ είναι σημαντικά διάφορο του μηδενός για $\alpha=0.01$. Δικαιολογεί τους ισχυρισμούς του συμπεράσματός σας το διάγραμμα διασποράς;

5.10

Σε 12 ράβδους σιδήρου μετρήθηκαν το μέτρο αντοχής (X) και το μέτρο ελαστικότητας (Y) και από τις μετρήσεις βρέθηκε $r(X,Y)=0.8$. Υπάρχει γραμμική συσχέτιση μεταξύ των μέτρων ελαστικότητας και αντοχής αν υποθέσουμε ότι τα X και Y είναι κανονικές τυχαίες μεταβλητές;



Y0176T3801Y ?

1964-1965

[illegible]

(b) No part of this document is to be released without the written approval of the Director, FBI.

ЭПОСЫ И ПОВЕСТИ НАДЫН

U.S. AIR FORCE - WASHINGTON, D.C. 20330

1964-1965

47-38861-1000

IT INFORMATION SYSTEMS SECURITY OFFICIALS

10/10/2010 10:10:10 AM

[illegible]

146 150 158 152 144 138

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100

100-443887-100

08 16 71 JAN 09 02 00 11 72 CS SER Y

04 25 41 36

SECRET

TT AA BB CCH EE FF GG HH II JJ KK LL MM NN OO PP QQ RR SS TT UU VV WW XX YY ZZ

26 11 45 25 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

000 0000000000 000000 00 00 0000000000 00

AMTVOYUO DAVIS Q 01 YD 3189446 14000000 10000000 00000000

[illegible]

DATE: 11/11/2015 TIME: 11:11:11 AM

100-443887-100

1950年10月1日

[illegible]

РП ОЛБ ЛОА (У) ЗАПИСНИКОВИ ОДЈЕД ОД ЛОА (У)
ЗАПИСНИКОВИ ОДЈЕД ОД ЛОА (У)

100-443886-100

...указавши, що в разі виникнення конфлікту між інтересами держави та інтересами громадян, держава повинна захищати інтереси громадян.

237140734

ΚΕΦΑΛΑΙΟ 6

αναλυση της διακυμανσης

6.1 Γενικότητες

Η Ανάλυση της Διακύμανσης είναι μία ευέλικτη και δυναμική μεθοδολογία, από τις σπουδαιότερες στη Στατιστική, που χρησιμοποιείται για τη μελέτη της σχέσης μεταξύ μίας εξαρτημένης μεταβλητής και μίας ή και περισσότερων ανεξάρτητων μεταβλητών. Δεν απαιτεί προηγούμενη γνώση της μορφής της σχέσης μεταξύ της εξαρτημένης και των ανεξάρτητων μεταβλητών ούτε περιορίζεται σε ποσοτικές μεταβλητές. Στη συνηθισμένη μορφή η εξαρτημένη μεταβλητή είναι ποσοτική και οι ανεξάρτητες μεταβλητές είναι ποιοτικές.

Στην αρχή θα αναφερθούμε στη σχέση (διαφορά) που υπάρχει μεταξύ της Ανάλυσης της Διακύμανσης και της Παλινδρόμησης. Στη συνέχεια θα αναπτύξουμε τη μεθοδολογία που εφαρμόζεται στην Ανάλυση της Διακύμανσης κατά ένα παράγοντα, όταν δηλαδή το μοντέλο της Ανάλυσης της Διακύμανσης περιέχει δύο ανεξάρτητες μεταβλητές. Θα ακολουθήσει η Ανάλυση της Διακύμανσης κατά δύο παράγοντες όπου το μοντέλο της Ανάλυσης της Διακύμανσης περιέχει δύο ανεξάρτητες μεταβλητές. Επίσης θα αναφέρουμε και τις βασικές έννοιες γενικώτερων μοντέλων Ανάλυσης της Διακύμανσης καθώς και ορισμένα στοιχεία από τον Σχεδιασμό Πειραμάτων, η Στατιστική ανάλυση του οποίου δίνεται από την Ανάλυση της Διακύμανσης.



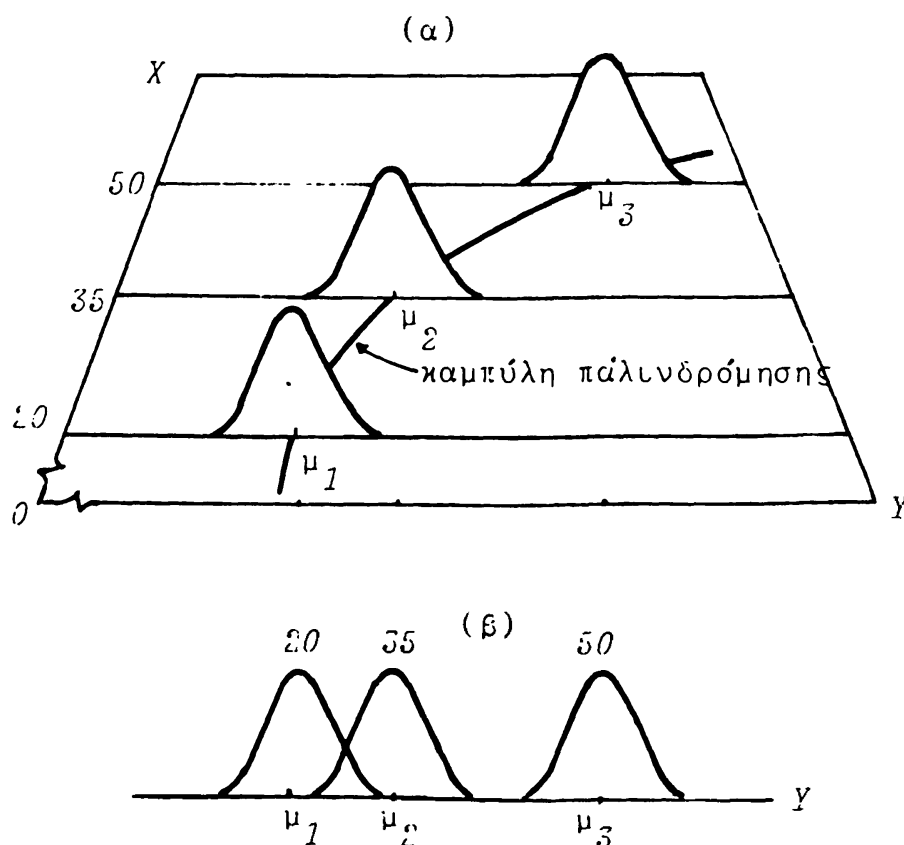
6.2 Σχέση μεταξύ παλινδρόμησης και ανάλυσης της διακύμανσης

Η Ανάλυση της Παλινδρόμησης, όπως είδαμε στο προηγούμενο Κεφάλαιο, μελετά τη στατιστική σχέση μεταξύ μίας ή περισσοτέρων ανεξάρτητων μεταβλητών και μίας εξαρτημένης μεταβλητής. Στα συνήθη μοντέλα Παλινδρόμησης, στα μοντέλα δηλαδή που δεν περιλαμβάνονται οι περιπτώσεις που αναφέραμε στην § 5.15, όλες οι μεταβλητές είναι ποσοτικές. Η συνάρτηση παλινδρόμησης περιγράφει την μορφή της στατιστικής σχέσης μεταξύ της μέσης τιμής της εξαρτημένης μεταβλητής και των ανεξάρτητων (ελεγχόμενων) μεταβλητών.

Η Ανάλυση της Διακύμανσης που οφείλεται κυρίως στην ερευνητική δουλειά του R. A. Fisher στην περίοδο 1912-1962 (Fisher R. A., 1966) μελετά την επίδραση ενός ή περισσοτέρων ανεξάρτητων μεταβλητών (παραγόντων) πάνω σε μία εξαρτημένη μεταβλητή. Μελετά επίσης και την αλληλοεπίδραση των ανεξάρτητων παραγόντων μεταξύ τους σε σχέση με την εξαρτημένη μεταβλητή. Τα μοντέλα της Ανάλυσης της Διακύμανσης μελετούν και αυτά, όπως και τα μοντέλα Παλινδρόμησης, τη στατιστική σχέση μεταξύ μίας εξαρτημένης μεταβλητής και μίας ή και περισσοτέρων ανεξάρτητων μεταβλητών. Έχουν όμως διαφορετική μορφή από τα μοντέλα Παλινδρόμησης. Η μέση τιμή της εξαρτημένης μεταβλητής εκφράζεται σαν άθροισμα των επιδράσεων και αλληλοεπιδράσεων των ανεξάρτητων μεταβλητών. Όπως και στην Παλινδρόμηση, η εξαρτημένη μεταβλητή είναι ποσοτική. Όμως, τα μοντέλα της Ανάλυσης της Διακύμανσης, διαφέρουν από τα συνήθη μοντέλα Παλινδρόμησης σε δύο σημαντικά σημεία: (i) Στα μοντέλα της Ανάλυσης της Διακύμανσης, οι ανεξάρτητες μεταβλητές μπορούν να είναι ποιοτικές. (ii) Αν οι ανεξάρτητες μεταβλητές είναι ποσοτικές, καμία υπόθεση δεν απαιτείται για την μορφή της στατιστικής σχέσης μεταξύ αυτών και της εξαρτημένης μεταβλητής. Το πρόβλημα δηλαδή του προσδιορισμού της μορφής της συνάρτησης Παλινδρόμησης που εμφανίζεται στην Παλινδρόμηση δεν υπάρχει στην Ανάλυση της Διακύμανσης. Στο Σχήμα 6.1 φαίνονται οι βασικές διαφορές μεταξύ



Παλινδρόμησης και Ανάλυσης της Διακύμανσης όταν οι ανεξάρτητες μεταβλητές είναι ποσοτικές. Στο Σχήμα 6.1α δίνεται γραφικά το μοντέλο Παλινδρόμησης που χρησιμοποιήθηκε για να μελετηθεί η επίδραση της ηλικίας του ανθρώπου στην πίεση του αίματος. Σε κάθε μία από τις τιμές (20, 35, 50) της ανεξάρτητης μεταβλητής X (ηλικία σε χρόνια) αντιστοιχεί μία κατανομή πιθανότητας της εξαρτημένης μεταβλητής Y (πίεση του αίματος). Οι μέσες τιμές μ_1, μ_2, μ_3 των κατανομών αυτών είναι σημεία της καμπύλης παλινδρόμησης που περιγράφει τη στατιστική σχέση μεταξύ ηλικίας και πίεσης αίματος. Για την ίδια μελέτη η Ανάλυση της Διακύμανσης ενδιαφέρεται μόνο για την μέση επίδραση των τριών ηλικιών στην πίεση του αίματος. Οι τρεις ηλικίες οδηγούν σε τρεις πληθυσμούς όπου ο καθένας έχει κάποια κατανομή της πίεσης του αίματος όπως φαίνεται στο Σχήμα 6.1β.

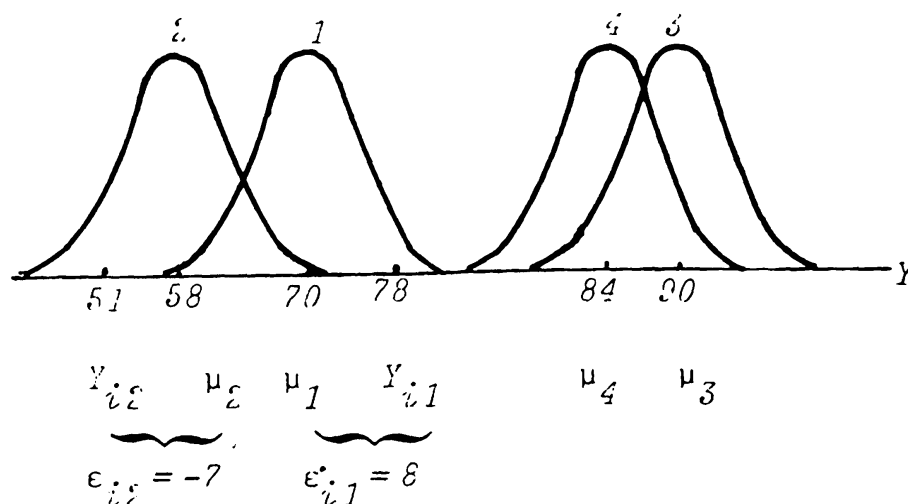


Σχήμα 6.1 Σχέση μεταξύ Παλινδρόμησης και Ανάλυσης της Διακύμανσης (X : ηλικία, Y : πίεση αίματος)



Συχνά στην Ανάλυση της Διακύμανσης οι ανεξάρτητες μεταβλητές είναι ποιοτικές όπως π.χ. όταν μελετάται η επίδραση τεσσάρων διαφορετικών ειδών ενός λιπάσματος (X) στην παραγωγή καλαμποκιού (Y). Εδώ, κάθε τύπος λιπάσματος οδηγεί σε διαφορετικό πληθυσμό με κάποια κατανομή της παραγωγής καλαμποκιού. Επειδή η μεταβλητή "τύπος λιπάσματος" είναι ποιοτική, δεν δίνεται στο Σχήμα το αντίστοιχο μοντέλο Παλινδρόμησης.

Όταν οι ανεξάρτητες μεταβλητές είναι ποιοτικές, δεν υπάρχει διαφορά μεταξύ των μοντέλων της Παλινδρόμησης και της Ανάλυσης της Διακύμανσης.



Σχήμα 6.2 Η Ανάλυση της Διακύμανσης με ποιοτική ανεξάρτητη μεταβλητή (X: είδη λιπάσματος, Y: παραγωγή καλαμποκιού)

Η κατάσταση διαφέρει όταν οι ανεξάρτητες μεταβλητές είναι ποσοτικές. Τότε, από τη μία μεριά υπάρχει η επιλογή μίας μεθοδολογίας που δεν απαιτεί τον προσδιορισμό της στατιστικής σχέσης (Ανάλυση της Διακύμανσης) και από την άλλη μια μεθοδολογία που απαιτεί αυτόν τον προσδιορισμό (Παλινδρόμηση). Αν υπάρχει αμφιβολία για την μορφή της στατιστικής σχέσης τότε, η στρατηγική που συνήθως ακολουθείται είναι, να αρχίσει κανένας με την Ανάλυση της Διακύμανσης για να μελετήσει την επίδραση των ανεξάρτητων μεταβλητών στην εξαρτημένη και μετά να χρησιμοποιήσει την Ανάλυση της Παλινδρόμησης για να προσδιορίσει

την μορφή της στατιστικής σχέσης μεταξύ εξαρτημένης και ανεξάρτητων μεταβλητών έτσι ώστε να μπορεί να προχωρήσει σε πρόβλεψη.

6.3 Μοντέλα Ανάλυσης της Διακύμανσης και χρησιμότητά τους

Η Ανάλυση της Διακύμανσης έχει την δική της ορολογία. Έτσι θεωρούμε χρήσιμο πριν αναφερθούμε σε περιπτώσεις όπου επιβάλλεται χρήση της μεθοδολογίας της Ανάλυσης της Διακύμανσης, να εισάγουμε μερικές από τις σχετικές έννοιες.

Για την Ανάλυση της Διακύμανσης ο όρος παράγοντας είναι μία ανεξάρτητη (ελεγχόμενη) μεταβλητή την οποία θέλουμε να μελετήσουμε. Για παράδειγμα, σε μία έρευνα που αφορά την επίδραση της τιμής κάποιου προϊόντος στις πωλήσεις του, ο παράγοντας που μελετάται είναι η τιμή. Όμοια, σε κάποια μελέτη που σκοπό έχει την απήχηση τεσσάρων τύπων διαφήμισης ενός προϊόντος, ο παράγοντας που ερευνάται είναι ο τύπος διαφήμισης. Μία συγκεκριμένη τιμή ή κάποιος συγκεκριμένος τύπος ενός παράγοντα αποτελεί και ένα επίπεδο του παράγοντα αυτού. Στη μελέτη της επίδρασης της ηλικίας στην πίεση του αίματος (Σχήμα 6.1) χρησιμοποιήθηκαν οι τιμές 20, 35, 50 της μεταβλητής "ηλικία". Κάθε μία από αυτές αποτελεί ένα επίπεδο του παράγοντα "ηλικία". Οι μελέτες διαφέρουν ανάλογα με τον αριθμό των παραγόντων που εξετάζονται. Όταν μόνο ένας παράγοντας εξετάζεται, όπως για παράδειγμα συμβαίνει στη μελέτη της απήχησης των τεσσάρων τύπων διαφήμισης, έχουμε τις μελέτες κατά ένα παράγοντα που αποτελούν το αντικείμενο της Ανάλυσης της Διακύμανσης κατά ένα παράγοντα. Όταν $k \geq 2$ παράγοντες εξετάζονται ταυτόχρονα έχουμε τις μελέτες κατά πολλούς παράγοντες. Ειδικότερα, η περίπτωση δύο παραγόντων όπως π.χ. είναι ο προσδιορισμός της επίδρασης της θερμοκρασίας και της συγκέντρωσης στα αποτελέσματα μιας χημικής διαδικασίας, οδηγεί σε

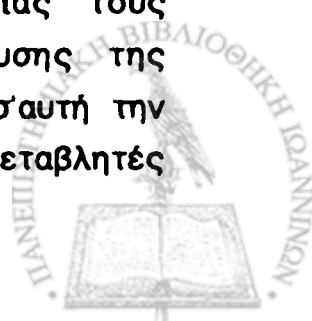


μελέτες κατά δύο παράγοντες που αποτελούν το αντικείμενο της Ανάλυσης της Διακύμανσης κατά δύο παράγοντες.

Σε αντιστοιχία με τις μεταβλητές, οι παράγοντες διακρίνονται σε ποιοτικούς, όταν τα επίπεδα διαφέρουν κατά κάποια ποιοτική ιδιότητα (τύποι διαφημίσεων, τύποι αντισκωριακού κλπ.) και ποσοτικούς, όταν κάθε επίπεδο περιγράφεται με κάποια αριθμητική ποσότητα σε κάποια κλίμακα (θερμοκρασία σε βαθμούς Κελσίου, ηλικία σε χρόνια, τιμή σε δρχ. κλπ.).

Στις μελέτες κατά ένα παράγοντα, μία δοκιμασία αντιστοιχεί σε ένα επίπεδο του παράγοντα. Έτσι, στο παράδειγμα των τεσσάρων τύπων διαφήμισης, κάθε τύπος αποτελεί μία δοκιμασία. Σε μελέτες πολλών παραγόντων, μία δοκιμασία αντιστοιχεί σε ένα συνδυασμό των επιπέδων των παραγόντων. Για παράδειγμα, σε μία μελέτη για τον προσδιορισμό της επίδρασης στις πωλήσεις της τιμής (25 δρχ., 29 δρχ.) και του χρώματος των πακέτων (κόκκινο, πράσινο) ενός προϊόντος, κάθε συνδυασμός, όπως "κόκκινο πακέτο-25 δρχ." είναι μία δοκιμασία. Συνολικά περιέχονται τέσσερες δοκιμασίες όσοι δηλαδή και οι συνδυασμοί "τιμή-χρώμα πακέτων".

Τα μοντέλα της Ανάλυσης της Διακύμανσης χρησιμοποιούνται για να προσδιορίσουν και αναλύσουν τις επιδράσεις μιας ή περισσότερων ανεξάρτητων μεταβλητών στην εξαρτημένη μεταβλητή. Έστω, για παράδειγμα, ότι μία εταιρεία χρησιμοποιεί κάποιο τύπο διαφήμισης των προϊόντων της που φαίνεται να την ικανοποιεί. Σε μία μελέτη βέβαια, θα μπορούσε να ερευνήσει τον τύπο διαφήμισης που χρησιμοποιεί μαζί με άλλους τρεις τύπους. Το μοντέλο της Ανάλυσης της Διακύμανσης κατά ένα παράγοντα θα ήταν σε θέση να διαπιστώσει, ποιός από τους τέσσερεις τύπους διαφήμισης είναι ο πλέον κατάλληλος για την εν λόγω εταιρεία. Έστω πάλι ότι σε μία μελέτη της απόδοσης των υπαλλήλων ενός οργανισμού εξετάζονται οι απουσίες των υπαλλήλων (εξαρτημένη μεταβλητή) σε σχέση με τα χρόνια υπηρεσίας, το φύλο, την οικογενειακή κατάσταση και το είδος της δουλειάς τους (ανεξάρτητες μεταβλητές). Το μοντέλο της Ανάλυσης της Διακύμανσης κατά πολλούς παράγοντες, θα μπορούσε, σ' αυτή την έρευνα, να διαπιστώσει κατά πόσον οι ανεξάρτητες μεταβλητές



αλληλοεπιδρούν σημαντικά και ποιές από αυτές είναι πιο στενά συσχετισμένες με την εξαρτημένη μεταβλητή. Το πρώτο από τα παραδείγματα δείχνει ότι τα μοντέλα της Ανάλυσης της Διακύμανσης σε μελέτες κατά ένα παράγοντα χρησιμοποιούνται για να συγκρίνουν τις διαφορετικές επιδράσεις των επιπέδων του παράγοντα και να διαπιστώσουν τον "καλύτερο". Το άλλο, ότι σε μελέτες με πολλούς παράγοντες τα μοντέλα της Ανάλυσης της Διακύμανσης χρησιμοποιούνται για να διακρίνουν πότε οι διαφορετικοί παράγοντες αλληλοεπιδρούν, ποιοί παράγοντες είναι οι "σπουδαιότεροι", ποιοί συνδιασμοί παραγόντων είναι οι "καλύτεροι" κλπ.

6.4 Ανάλυση της Διακύμανσης κατά ένα παράγοντα

Η μεθοδολογία που υπεισέρχεται στην Ανάλυση της Διακύμανσης κατά ένα παράγοντα αποτελεί γενίκευση της μεθοδολογίας που χρησιμοποιείται για συγκρίσεις των μέσων τιμών δύο ανεξάρτητων πληθυσμών (t-τεστ) έτσι ώστε να επιτυγχάνεται η ταυτόχρονη σύγκριση των μέσων τιμών περισσότερων από δύο ανεξάρτητων πληθυσμών.

Τα βασικά στοιχεία του μοντέλου της Ανάλυσης της Διακύμανσης κατά ένα παράγοντα είναι απλά. Σε κάθε δοκιμασία (επίπεδο παράγοντα) υποθέτουμε ότι αντιστοιχεί μία κατανομή πιθανοτήτων της εξαρτημένης μεταβλητής. Για παράδειγμα, στην περίπτωση της μελέτης των τεσσάρων ειδών του λιπάσματος (§ 6.2), υπάρχει μία κατανομή της παραγωγής καλαμποκιού για κάθε είδος λιπάσματος. Στο μοντέλο της Ανάλυσης της Διακύμανσης υποθέτουμε ότι κάθε μία από τις κατανομές πιθανότητας είναι κανονική με την ίδια διακύμανση και ότι οι παρατηρήσεις που παίρνονται από κάποιες πειραματικές μονάδες, για κάθε δοκιμασία, είναι τυχαίες παρατηρήσεις από την αντίστοιχη κατανομή και ανεξάρτητες από τις παρατηρήσεις κάθε άλλης δοκιμασίας. Το Σχήμα 6.2 δείχνει αυτές τις συνθήκες. Οι κατανομές διαφέρουν μόνο στις μέσες τιμές. Έτσι, τυχόν διαφορές



στις μέσες τιμές, θα έδειχναν τις διαφορές στις επιδράσεις των δοκιμασιών στην εξαρτημένη μεταβλητή. Αυτός είναι βασικά ο λόγος που η Ανάλυση της Διακύμανσης επικεντρώνεται στις μέσες αποκρίσεις (τιμές της εξαρτημένης μεταβλητής) των δοκιμασιών που εξετάζονται. Έτσι, η ανάλυση των δεδομένων βασίζεται στο: να προσδιορισθεί αν οι μέσες τιμές που αντιστοιχούν σε κάθε δοκιμασία είναι οι ίδιες και αν δεν είναι, να εξετασθεί πως διαφέρουν και ποιές είναι οι επιπτώσεις των διαφορών.

6.4.1 Το στατιστικό μοντέλο

Έστω ότι υπάρχουν r δοκιμασίες (επίπεδα του παράγοντα) που εξετάζονται (π.χ. 4 τύποι ενός λιπάσματος) οι οποίες θα συμβολίζονται με τον δείκτη j , $j=1, \dots, r$. Ο αριθμός των παρατηρήσεων για την j δοκιμασία θα συμβολίζεται με n_j και ο συνολικός αριθμός των παρατηρήσεων με n δηλαδή $n=\sum n_j$. Η i -οστή παρατήρηση της j δοκιμασίας, $i=1, \dots, n_j$, θα συμβολίζεται με Y_{ij} .

Το μοντέλο έχει τη μορφή

$$Y_{ij} = \mu_j + \varepsilon_{ij}, \quad (6.1)$$

όπου Y_{ij} είναι η i -οστή τιμή (απόκριση) της εξαρτημένης μεταβλητής για την j δοκιμασία. Διαφορετικά, το Y_{ij} μπορεί να θεωρηθεί ως η i -οστή παρατήρηση του δείγματος από τον j πληθυσμό. Τα μ_j είναι παράμετροι (οι μέσες τιμές των r ανεξάρτητων πληθυσμών), τα δε ε_{ij} είναι ανεξάρτητες $N(0, \sigma^2)$ τυχαίες μεταβλητές και αποτελούν τα τυχαία σφάλματα.

Ας υποθέσουμε ότι το μοντέλο (6.1) εφαρμόζεται στο παράδειγμα των 4^{ων} ειδών λιπάσματος (§ 6.2) και ότι οι παράμετροι είναι: $\mu_1=70$, $\mu_2=58$, $\mu_3=84$ και $\sigma^2=16$. Το Σχήμα 6.2 παριστάνει γραφικά το μοντέλο αυτό, σύμφωνα με το οποίο η παραγωγικότητα σε καλαμπόκι που αντιστοιχεί στον 1^ο τύπο λιπάσματος κατανέμεται

κανονικά με $\mu_1=70$ και $\sigma^2=16$. Αν υποθέσουμε ότι η i -οστή παρατηρούμενη παραγωγικότητα για τον 1^ο τύπο λιπάσματος είναι $Y_{i1}=78$, η τιμή του αντίστοιχου τυχαίου σφάλματος είναι $\varepsilon_{i1}=Y_{i1}-\mu_1=78-70=8$. Η παρατήρηση Y_{i1} όπως και η παρατήρηση $Y_{i2}=51$ για την οποία η τιμή του τυχαίου σφάλματος είναι $\varepsilon_{i2}=-7$ φαίνονται στο Σχήμα 6.2.

Το μοντέλο με την μορφή που δόθηκε στη σχέση (6.1) είναι επαρκές για μελέτες κατά ένα παράγοντα. Μία διαφορετική αλλά ισοδύναμη μορφή, χρήσιμη σε μελέτες με πολλούς παράγοντες, είναι

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij}, \quad (6.2)$$

όπου η παράμετρος $\mu = \sum n_j \mu_j / n$ (σταθερά, κοινή για όλες τις παρατηρήσεις) λέγεται μεγάλη μέση τιμή, η παράμετρος $\tau_j = \mu_j - \mu$ (σταθερά για κάθε δοκιμασία) συμβολίζει την επίδραση (effect) της j δοκιμασίας και τα τυχαία σφάλματα ε_{ij} που είναι ανεξάρτητες $N(0, \sigma^2)$ τ.μ. Είναι προφανές ότι $\sum n_j \tau_j = 0$.

Το αρχικό ενδιαφέρον στις μελέτες κατά ένα παράγοντα, όπως και προηγούμενα αναφέραμε, έγκειται στο κατά πόσον οι μέσοι μ_j είναι ίσοι δηλαδή για παράδειγμα, σε μία μελέτη 5 διαφορετικών τύπων διαφήμισης ενός προϊόντος αρχικά θα ενδιέφερε αν οι 5 τύποι διαφήμισης έχουν ίσες μέσες επιδράσεις στις πωλήσεις του προϊόντος. Έτσι, για το μοντέλο (6.1) οι υποθέσεις που ενδιαφέρουν είναι η $H_0: \mu_1 = \mu_2 = \dots = \mu_r$ έναντι της H_a : όχι όλα τα μ_j ίσα μεταξύ τους. Ισοδύναμα, οι ίδιες υποθέσεις για το μοντέλο (6.2) είναι η $H_0: \tau_1 = \tau_2 = \dots = \tau_r = 0$ έναντι της H_a : όχι όλα τα τ_j ίσα με μηδέν.

6.4.2 Εκτίμηση των παραμέτρων με τη μέθοδο των ελαχίστων τετραγώνων

Ας υποθέσουμε ότι το μοντέλο της Ανάλυσης της Διακύμανσης κατά ένα παράγοντα είναι κατ'αρχήν κατάλληλο για να



μελετήσει την επίδραση r δοκιμασιών στην εξαρτημένη μεταβλητή Y και έστω ότι έχουμε στη διάθεση μας τις παρατηρήσεις Y_{ij} $i=1, \dots, n_j, j=1, \dots, r$ που πάρθηκαν μετά από κάποιο πείραμα. Οι τιμές των παραμέτρων του μοντέλου είναι άγνωστες και κατά συνέπεια θα πρέπει να τις εκτιμήσουμε από τα δεδομένα. Θα χρησιμοποιήσουμε, όπως και στην Παλινδρόμηση, τη μέθοδο των ελαχίστων τετραγώνων για να βρούμε τους εκτιμητές των παραμέτρων. Πριν περιγράψουμε τη μέθοδο θα δώσουμε ορισμένους συμβολισμούς που θα χρειασθούν στην ανάπτυξη της μεθοδολογίας σε συνδιασμό με ένα ενδεικτικό παράδειγμα: Η εταιρεία ζυμαρικών ΑΒΕΖ ενδιαφέρεται να διαπιστώσει κατά πόσο οι κάτοικοι μίας περιοχής δείχνουν ιδιαίτερη προτίμηση ανάμεσα σε τέσσερις τύπους ζυμαρικών που παράγει. Για το σκοπό αυτό επιλέγονται τυχαία δέκα καταστήματα τροφίμων, με σχεδόν ίσες πωλήσεις, σε κάθε ένα από τα οποία δίνεται για πώληση ένας από τους τέσσερις τύπους. Τελικά δύο από τους τέσσερις τύπους ζυμαρικών δίνονται σε τρία από τα δέκα καταστήματα και οι υπόλοιποι δύο σε δύο από τα καταστήματα ο κάθε τύπος. Έγινε προσπάθεια ώστε, εκτός από τον τύπο, άλλες συνθήκες που υπήρχε περίπτωση να επιδράσουν στις πωλήσεις όπως π.χ. η ποσότητα, η τιμή, η τοποθέτηση στα ράφια των καταστημάτων να διατηρηθούν στα ίδια επίπεδα για όλα τα καταστήματα. Οι ποσότητες ζυμαρικών που πωλήθηκαν (αριθμός πακέτων του ενός κιλού), ανά κατάστημα και τύπο, σε διάστημα μίας βδομάδας δίνονται στον Πίνακα 6.1α (υποθετικά δεδομένα). Στον Πίνακα 6.1β δίνονται οι συμβολισμοί που αντιστοιχούν στα δεδομένα της ΑΒΕΖ του Πίνακα 6.1α. Όπως προαναφέραμε το Y_{ij} παριστά την παρατήρηση της i -οστής πειραματικής μονάδας για την j δοκιμασία. Εδώ Y_{ij} είναι ο αριθμός πακέτων που πωλήθηκαν στο i -οστό από τα καταστήματα που διέθετε τον j τύπο ζυμαρικών. Για παράδειγμα, το Y_{11} παριστά τις πωλήσεις του πρώτου από τα καταστήματα που πουλούσαν τον 1ο τύπο ζυμαρικών, εδώ $Y_{11}=12$.



Κατάστημα τροφίμων	(α) Δεδομένα Τύπος ζυμαρικών				Σύνολο
	1	2	3	4	
1	12	14	19	24	
2	18	12	17	30	
3		13	21		
Σύνολο	30	39	57	54	180
Δειγμ. μέσοι	15	13	19	27	18
Αρ. καταστημ.	2	3	3	2	10

Δειγμ. μονάδα (i)	(β) Συμβολισμοί Δοκιμασίες (j) (επίπεδα παράγοντα)				Σύνολο
	1	2	3	4	
1	Y_{11}	Y_{12}	Y_{13}	Y_{14}	
2	Y_{21}	Y_{22}	Y_{23}	Y_{24}	
3		Y_{32}	Y_{33}		
Σύνολο	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	$Y_{.4}$	$Y_{..}$
Δειγμ. μέσοι	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	$\bar{Y}_{.3}$	$\bar{Y}_{.4}$	$\bar{Y}_{..}$
Μέγεθ. δείγματος	n_1	n_2	n_3	n_4	n

Πίνακας 6.1 Δεδομένα μεγέθους πωλήσεων (αριθμός πακέττων του κιλού) ανά κατάστημα για κάθε ένα από τους τέσσερις τύπους ζυμαρικών, παράδειγμα ABEZ

Όμοια, ο αριθμός των πωλήσεων του δεύτερου από τα καταστήματα που διέθεταν τον 3^ο τύπο ήταν $Y_{23} = 17$. Το άθροισμα των παρατηρήσεων που αντιστοιχεί στην j δοκιμασία συμβολίζεται με $Y_{.j}$: $Y_{.j} = \sum_i Y_{ij}$. Για παράδειγμα, από τον 1^ο τύπο συνολικά πωλήθηκαν $Y_{.1} = 30$ πακέττα. Ο δειγματικός μέσος της j δοκιμασίας συμβολίζεται με $\bar{Y}_{.j}$: $\bar{Y}_{.j} = Y_{.j} / n_j$. Το συνολικό άθροισμα των παρατηρήσεων



συμβολίζεται με $Y_{..}$: $Y_{..} = \sum \sum Y_{ij}$. Τέλος η μέση τιμή όλων των παρατηρήσεων (δειγματική γενική μέση τιμή) συμβολίζεται με $\bar{Y}_{..}$: $\bar{Y}_{..} = Y_{..}/n$. Για το παράδειγμα μας, $\bar{Y}_{..} = 180/10 = 18$.

Η μέθοδος των ελαχίστων τετραγώνων χρησιμοποιεί, όπως είδαμε στο προηγούμενο Κεφάλαιο (§ 5.5), το άθροισμα τετραγώνων των αποκλίσεων των Y_{ij} από τις αναμενόμενες τιμές των.

Για το μοντέλο (6.1) έχουμε $E(Y_{ij}) = \mu_j$. Οι εκτιμητές των ελαχίστων τετραγώνων, $\hat{\mu}_j$, των παραμέτρων μ_j , προκύπτουν αν ελαχιστοποιήσουμε ως προς μ_j την συνάρτηση

$$Q(\mu_1, \dots, \mu_r) = \sum_i \sum_j (Y_{ij} - \mu_j)^2$$

$$= \sum_i (Y_{i1} - \mu_1)^2 + \dots + \sum_i (Y_{ir} - \mu_r)^2.$$

Κάθε μία από τις παραμέτρους εμφανίζεται μόνο σε ένα από τα επί μέρους αθροίσματα του δεύτερου μέλους. Έτσι, αν ελαχιστοποιήσουμε τα επί μέρους αθροίσματα, η $Q(\mu_1, \dots, \mu_r)$ γίνεται ελάχιστη. Έστω το j επί μέρους άθροισμα

$$Q_j(\mu_j) = \sum_i (Y_{ij} - \mu_j)^2.$$

Αν παραγωγίσουμε ως προς μ_j την προηγούμενη σχέση και εξισώσουμε με το μηδέν, παίρνουμε τον εκτιμητή ελαχίστων τετραγώνων $\hat{\mu}_j$ της παραμέτρου μ_j :

$$-2 \sum_i (Y_{ij} - \hat{\mu}_j) = 0$$

ή

$$\hat{\mu}_j = \bar{Y}_{.j}, \quad j=1, \dots, r.$$

Για το παράδειγμα της ΑΒΕΖ οι εκτιμητές ελαχίστων τετραγώνων είναι:



$$\hat{\mu}_1 = \bar{Y}_{.1} = 15, \hat{\mu}_2 = \bar{Y}_{.2} = 13, \hat{\mu}_3 = \bar{Y}_{.3} = 19, \hat{\mu}_4 = \bar{Y}_{.4} = 27.$$

Εκτιμάται δηλαδή ότι ο μέσος αριθμός πωλήσεων ανά κατάσταση για τον 1^ο τύπο είναι 15 πακέτα (για τον πληθυσμό των καταστημάτων αυτής της μελέτης) για τον 2^ο τύπο 13 πακέτα κλπ.

Για το μοντέλο (6.2) η συνάρτηση που πρέπει να ελαχιστοποιηθεί είναι η

$$Q(\mu, \tau_1, \dots, \tau_r) = \sum_i \sum_j (Y_{ij} - \mu - \tau_j)^2.$$

Μπορεί εύκολα ναδειχτεί ότι οι εκτιμητές ελαχίστων τετραγώνων δίνονται από τις σχέσεις

$$\hat{\mu} = \bar{Y}_{..}$$

και

(6.4)

$$\hat{\tau}_j = \bar{Y}_{.j} - \bar{Y}_{..}, \quad j=1, \dots, r.$$

Για το παράδειγμα της ABEZ έχουμε

$$\hat{\mu} = 18, \hat{\tau}_1 = \bar{Y}_{.1} - \bar{Y}_{..} = -3, \hat{\tau}_2 = -5, \hat{\tau}_3 = 1, \hat{\tau}_4 = 9.$$

Από τα δεδομένα δηλαδή προκύπτει ότι οι μέσες πωλήσεις ανά κατάσταση για τον 1^ο τύπο ζυμαρικών είναι 3 πακέτα λιγότερα από τις μέσες πωλήσεις που αντιστοιχούν και στους τέσσερις τύπους, ότι για τον 4^ο τύπο οι μέσες πωλήσεις είναι 9 πακέτα περισσότερα κλπ.

Μπορεί να αποδειχθεί ότι οι εκτιμητές ελαχίστων τετραγώνων που δίνονται από τις σχέσεις (6.3) και (6.4) είναι αμερόληπτοι ταυτίζονται δε με τους εκτιμητές μεγίστης πιθανοφάνειας για τα μοντέλα (6.1) και (6.2) αντίστοιχα.

Τα υπόλοιπα e_{ij} είναι, όπως και στην παλινδρόμηση, πολύ χρήσιμα στη μελέτη της καταλληλότητας του μοντέλου όπως θα δούμε σε επόμενη παράγραφο. Ορίζονται από την σχέση



$$e_{ij} = Y_{ij} - \hat{\mu}_j = Y_{ij} - \bar{Y}_{.j}$$

αποτελούν δηλαδή, για κάθε δοκιμασία, τις αποκλίσεις των παρατηρήσεων από την δειγματική μέση τιμή τους. Ο ορισμός αυτός ισχύει και για το μοντέλο (6.2) γιατί $\hat{\mu} + \hat{\tau}_j = \bar{Y}_{..} + (\bar{Y}_{.j} - \bar{Y}_{..}) = \bar{Y}_{.j}$. Τα υπόλοιπα e_{ij} για τα δεδομένα του Πίνακα 6.1 για το παράδειγμα της ΑΒΕΖ δίνονται στον Πίνακα 6.2. Για παράδειγμα, από τον Πίνακα 6.1 βρίσκουμε

$$e_{12} = Y_{12} - \bar{Y}_{.2} = 14 - 13 = +1.$$

Όπως διαπιστώνουμε και από τον Πίνακα, τα υπόλοιπα αθροίζουν στο μηδέν για κάθε δοκιμασία δηλαδή ικανοποιούν τις παρακάτω r συνθήκες:

$$\sum_i e_{ij} = 0, \quad j=1, \dots, r. \quad (6.5)$$

Κατάστημα τροφίμων	Τύπος ζυμαρικών				Σύνολο
	1	2	3	4	
1	-3	+1	0	-3	
2	+3	-1	-2	+3	
3		0	+2		
Σύνολο	0	0	0	0	0

Πίνακας 6.2 Υπόλοιπα, παράδειγμα ΑΒΕΖ

6.4.3 Η Ανάλυση της Διακύμανσης

Η μεθοδολογία της Ανάλυσης της Διακύμανσης για το μοντέλο της Παλινδρόμησης βασίζεται όπως είδαμε στην § 5.10 στη διάσπαση της ολικής μεταβλητότητας, που μετράται με το ολικό άθροισμα τετραγώνων, στα επί μέρους αθροίσματα τετραγώνων της Παλινδρόμησης και του σφάλματος. Μία ανάλογη διάσπαση της

ολικής μεταβλητότητας οδηγεί στην αντίστοιχη μεθοδολογία για το μοντέλο της Ανάλυσης της Διακύμανσης κατά ένα παράγοντα.

Η ολική μεταβλητότητα των παρατηρήσεων Y_{ij} , που ισούται με το ολικό άθροισμα τετραγώνων (SS_{tot}) των αποκλίσεων $Y_{ij} - \bar{Y}_{..}$ των παρατηρήσεων από την δειγματική γενική μέση τιμή, δίνεται από τη σχέση

$$SS_{tot} = \sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2.$$

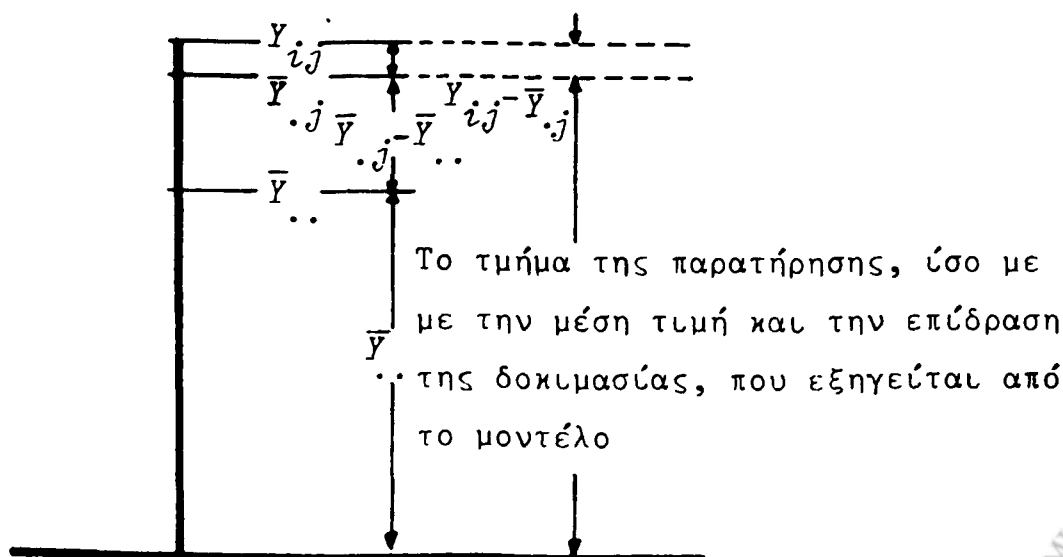
Η ολική απόκλιση $Y_{ij} - \bar{Y}_{..}$ κάθε μίας από τις παρατηρήσεις μπορεί να διασπαστεί σε δύο επί μέρους τμήματα. Στην απόκλιση των δειγματικών μέσων των δοκιμασιών από την δειγματική γενική μέση τιμή και στην απόκλιση των παρατηρήσεων από τους δειγματικούς μέσους των αντίστοιχων δοκιμασιών.

Ισχύει δηλαδή,

$$Y_{ij} - \bar{Y}_{..} = \bar{Y}_{.j} - \bar{Y}_{..} + Y_{ij} - \bar{Y}_{.j}. \quad (6.6)$$

Η σχέση (6.6) δίνεται στο Σχήμα 6.3.

Το τμήμα της παρατήρησης που δεν εξηγείται από το μοντέλο



Σχήμα 6.3 Η διάσπαση μίας παρατήρησης στο μοντέλο της Διακύμανσης κατά ένα παράγοντα



Αν υψώσουμε την (6.6) στο τετράγωνο, αθροίσουμε και λάβουμε υπόψη ότι

$$\sum_j \sum_i 2(\bar{y}_{.j} - \bar{y}_{..})(y_{ij} - \bar{y}_{.j}) = 2 \sum_j (\bar{y}_{.j} - \bar{y}_{..}) \sum_i (y_{ij} - \bar{y}_{.j}) = 0,$$

έχουμε

$$\sum_j \sum_i (y_{ij} - \bar{y}_{..})^2 = \sum_j \sum_i (\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_j \sum_i (y_{ij} - \bar{y}_{.j})^2 \quad (6.7)$$

ή συμβολικά

$$SS_{tot} = SS_{tre} + SS_{res}.$$

Η αξιοσημείωτη σχέση (6.7) μας λέει ότι το άθροισμα τετραγώνων των αποκλίσεων των y_{ij} από την δειγματική γενική μέση τιμή (ολικό άθροισμα τετραγώνων), ισούται με το άθροισμα τετραγώνων των αποκλίσεων των δειγματικών μέσων των δοκιμασιών από την δειγματική γενική μέση τιμή (άθροισμα τετραγώνων των δοκιμασιών, SS_{tre}) και το άθροισμα τετραγώνων των υπολοίπων e_{ij} (SS_{res}).

Τα αθροίσματα τετραγώνων που εμφανίζονται στην (6.7) υπολογίζονται ευκολώτερα από τους παρακάτω ισοδύναμους τύπους:

$$SS_{tot} = \sum_j \sum_i y_{ij}^2 - \frac{y_{..}^2}{n}$$

$$SS_{tre} = \sum_j \frac{y_{.j}^2}{n_j} - \frac{y_{..}^2}{n}$$

$$SS_{res} = \sum_j \sum_i y_{ij}^2 - \sum_j \frac{y_{.j}^2}{n_j}$$

$$= SS_{tot} - SS_{tre}$$

Στα αθροίσματα τετραγώνων της (6.7) αντιστοιχούν οι παρακάτω



βαθμοί ελευθερίας: Στο SS_{tot} , $n-1$ β.ε. γιατί υπάρχουν n όροι στο άθροισμα και η συνθήκη $\sum_j \sum_i (Y_{ij} - \bar{Y}_{..}) = 0$, στο SS_{tre} , $r-1$ β.ε. γιατί υπάρχουν r όροι και η συνθήκη $\sum_j n_j (\bar{Y}_{.j} - \bar{Y}_{..}) = 0$ και στο SS_{res} , $n-r$ β.ε. γιατί υπάρχουν n όροι και r συνθήκες $\sum_i (Y_{ij} - \bar{Y}_{.j}) = 0$, $j=1, \dots, r$. Σημειώστε την ανάλογη προσθετικότητα με τα αθροίσματα τετραγώνων των αντίστοιχων β.ε.

Τα μέσα τετράγωνα δημιουργούνται όπως και στην Παλινδρόμηση αν διαιρέσουμε τα αθροίσματα τετραγώνων με τους αντίστοιχους β.ε. Έτσι έχουμε

$$MS_{tre} = \frac{SS_{tre}}{r-1}$$

και

$$MS_{res} = \frac{SS_{res}}{n-r}$$

όπου MS_{tre} είναι το μέσο τετράγωνο των δοκιμασιών και MS_{res} το μέσο τετράγωνο των σφαλμάτων.

Τα αθροίσματα τετραγώνων, οι β.ε. και τα μέσα τετράγωνα αποτελούν τον λεγόμενο πίνακα της Ανάλυσης της Διακύμανσης (ΑΝΑΔΙΑ).

ΑΝΑΔΙΑ

Προέλευση Μεταβλητότητας	Άθροισμα Τετραγώνων	Βαθμοί Ελευθερίας	Μέσο Τετράγωνο
Δοκιμασίες ή Μεταξύ των Δοκιμ.	$SS_{tre} = \sum_j n_j (\bar{Y}_{.j} - \bar{Y}_{..})^2$	$r-1$	$MS_{tre} = \frac{SS_{tre}}{r-1}$
Υπόλοιπο ή Εντός των Δοκιμ.	$SS_{res} = \sum_i \sum_j (Y_{ij} - \bar{Y}_{.j})^2$	$n-r$	$MS_{res} = \frac{SS_{res}}{n-r}$
Ολική Μεταβλητότητα	$SS_{tot} = \sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2$	$n-1$	

Για τα δεδομένα του παραδείγματος της ΑΒΕΖ μετά τους σχετικούς υπολογισμούς προκύπτει ο παρακάτω πίνακας:



ΑΝΑΔΙΑ

Προέλευση Μεταβλητότητας	Άθροισμα Τετραγώνων	Βαθμοί Ελευθερίας	Μέσο Τετράγωνο
Δοκιμασίες	258	3	86
Υπόλοιπο	46	6	7.67
Ολική Μεταβλητότητα	304	9	

Το MS_{res} μπορεί διαφορετικά να γραφτεί

$$MS_{res} = \frac{1}{n-r} \sum_j \left((n_j - 1) \frac{\sum (y_{ij} - \bar{y}_{.j})^2}{n_j - 1} \right)$$

ή

$$MS_{res} = \frac{1}{n-r} \sum_j (n_j - 1) S_j^2$$

όπου S_j^2 είναι η δειγματική διακύμανση των παρατηρήσεων που προέρχονται από την j δοκιμασία. Έτσι έχουμε

$$E(MS_{res}) = \frac{1}{n-r} \sum_j (n_j - 1) E(S_j^2) = \sigma^2.$$

Είναι δηλαδή το MS_{res} πάντα ένας αμερόληπτος εκτιμητής της κοινής διακύμανσης σ^2 των r πληθυσμών.

Επίσης, μπορεί να αποδειχθεί ότι

$$E(MS_{tre}) = \sigma^2 + \frac{\sum n_j (\mu_j - \mu)^2}{r-1}$$

Η σχέση αυτή οδηγεί στην $E(MS_{tre}) = \sigma^2$ όταν $\mu_1 = \dots = \mu_r = \mu$ ή ισοδύναμα όταν $\tau_1 = \dots = \tau_r = 0$. Στην περίπτωση αυτή τα MS_{res} και MS_{tre} εκτιμούν την ίδια παράμετρο σ^2 . Όταν όμως οι μέσες τιμές διαφέρουν, η $E(MS_{tre})$ είναι μεγαλύτερη από την $E(MS_{res})$ δηλαδή το MS_{tre} φυσιολογικά θα τείνει να γίνεται

μεγαλύτερο του MS_{res} . Αν τώρα τα μ_j είναι ίσα, τα $\bar{Y}_{.j}$ ακολουθούν την ίδια δειγματική κατανομή με $E(\bar{Y}_{.j}) = \mu$ και $Var(\bar{Y}_{.j}) = \sigma^2/n_j$. αν διαφέρουν, τα $\bar{Y}_{.j}$ ακολουθούν διαφορετικές κατανομές η καθεμιά με την ίδια διακύμανση αλλά με διαφορετικές μέσες τιμές. Τα $\bar{Y}_{.j}$ δηλαδή θα τείνουν να διαφέρουν μεταξύ τους περισσότερο όταν τα μ_j διαφέρουν παρά όταν είναι ίσα μεταξύ τους.

Η υπόθεση που αρχικά μας ενδιαφέρει να ελέγξουμε είναι κατά πόσον οι μέσες επιδράσεις των δοκιμασιών είναι ίδιες. Για παράδειγμα, αν οι τέσσερις τύποι ζυμαρικών της ΑΒΕΖ οδηγούν σε ίσες ποσότητες πωλήσεων, δεν θα υπήρχε κανένα όφελος από μία λεπτομερέστερη ανάλυση των δεδομένων με σκοπό τον προσδιορισμό του τύπου με τις περισσότερες προτιμήσεις.

Έτσι βασικά μας ενδιαφέρει να ελέγξουμε την $H_0: \mu_1 = \dots = \mu_r = \mu$ έναντι της H_a : Τα μ_j δεν είναι όλα ίσα μεταξύ τους ή ισοδύναμα την $H_0: \tau_1 = \dots = \tau_r = 0$ έναντι της H_a : Τα τ_j δεν είναι όλα ίσα με μηδέν. Για τον έλεγχο της H_0 χρησιμοποιούμε το στατιστικό MS_{tre}/MS_{res} για το οποίο σύμφωνα με όσα προαναφέραμε περιμένουμε να παίρνει τιμές κοντά στην μονάδα όταν τα μ_j είναι ίσα και μεγαλύτερες από την μονάδα όταν τα μ_j διαφέρουν.

Όταν ισχύει η H_0 , μπορεί να αποδειχθεί (βλέπε π.χ. Wilks, 1962) ότι τα στατιστικά SS_{tre} και SS_{res} είναι ανεξάρτητα και επί πλέον

$$\frac{SS_{tre}}{\sigma^2} \sim \chi_{r-1}^2$$

και

$$\frac{SS_{res}}{\sigma^2} \sim \chi_{n-r}^2$$

Έτσι έχουμε

$$F = \frac{MS_{tre}}{MS_{res}} \sim F_{r-1, n-r}.$$

(6.8)



Η σχέση (6.8), από τις σημαντικότερες στην Στατιστική, αποτελεί επιγραμματικά το γνωστό σαν θεώρημα του Cochran. Αφού λοιπόν γνωρίζουμε ότι όταν ισχύει η H_0 το στατιστικό F έχει την $F_{r-1, n-r}$ κατανομή και ότι μεγάλες τιμές του F προκύπτουν όταν ισχύει η H_a , ο κανόνας είναι να απορρίπτουμε την H_0 με επίπεδο σημαντικότητας α όταν

$$F > F_{\alpha, r-1, n-r}.$$

Έστω τώρα ότι μας ενδιαφέρει να ελέγξουμε με $\alpha=0.05$ κατά πόσον οι μέσες πωλήσεις για τους τέσσερες τύπους ζυμαρικών της ΑΒΕΖ είναι ίσες. Από τον αντίστοιχο πίνακα ΑΝΑΔΙΑ, $F = MS_{\text{tre}}/MS_{\text{res}} = 11.2$. Επειδή $F = 11.2 > F_{0.05, 3, 6} = 4.76$ απορρίπτουμε την $H_0: \mu_1 = \dots = \mu_4$, συμπεραίνουμε δηλαδή με βάση τα δεδομένα που έχουμε στη διάθεσή μας ότι οι 4 τύποι ζυμαρικών δεν οδηγούν σε ίσες μέσες πωλήσεις.

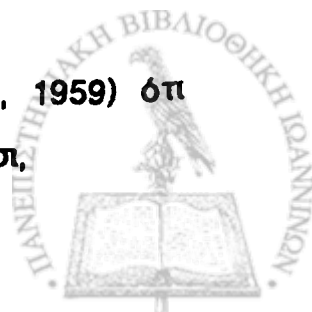
Παρακάτω θα δούμε το δεύτερο στάδιο της Ανάλυσης, πως δηλαδή μπορούμε να προχωρήσουμε σε επιμέρους ελέγχους με σκοπό τον προσδιορισμό των δοκιμασιών που προκαλούν τις διαφορετικές επιδράσεις.

6.4.4 Ανάλυση των μέσων τιμών

Είδαμε στο εδάφιο 6.4.2 ότι ένας αμερόληπτος εκτιμητής της παραμέτρου μ_j είναι ο $\hat{\mu}_j = \bar{Y}_{\cdot j}$ με διακύμανση, λόγω της ανεξαρτησίας των Y_{ij} , σ^2/n_j . Επομένως,

$$\bar{Y}_{\cdot j} \sim N(\mu_j, \sigma^2/n_j).$$

Επί πλέον, μπορεί να αποδειχθεί (βλέπε π.χ. Scheffé, 1959) ότι $SS_{\text{res}}/\sigma^2 \sim \chi^2_{n-r}$ με τα $\bar{Y}_{\cdot j}$ και SS_{res} ανεξάρτητα. Έτσι,



$$\frac{\bar{Y}_{.j} - \mu_j}{\sqrt{MS_{res}/n_j}} = t_{n-r}. \quad (6.9)$$

Αν χρησιμοποιήσουμε την σχέση (6.9) μπορούμε να ελέγξουμε διάφορες υποθέσεις για το μ_j .

Ένα $(1 - \alpha)100\%$ διάστημα εμπιστοσύνης για το μ_j είναι το $\bar{Y}_{.j} \pm t_{\alpha/2, n-r} \sqrt{MS_{res}/n_j}$.

Ένα 95% διάστημα εμπιστοσύνης για τις μέσες πωλήσεις των ζυμαρικών τύπου 1 για τα δεδομένα της ABEZ είναι το $15.0 \pm 2.447 \sqrt{7.67/2}$ ή το $[10.2, 19.8]$. Έτσι, με συντελεστή εμπιστοσύνης 95%, εκτιμάμε ότι οι μέσες πωλήσεις ζυμαρικών του τύπου 1 είναι ανά κατάσταση κάπου μεταξύ των 10.2 και 19.8 πακέτων.

Η σχέση (6.9) χρησιμοποιεί τον αμερόληπτο εκτιμητή MS_{res} της κοινής διακύμανσης σ^2 των r πληθυσμών αντί της δειγματικής διακύμανσης S_j^2 του j δείγματος που θα οδηγούσε στο γνωστό t -τεστ. Το MS_{res} στηριζόμενο σε περισσότερες παρατηρήσεις είναι καλύτερος εκτιμητής από το S_j^2 .

Γενικά, περισσότερο ενδιαφέρον παρουσιάζει η σύγκριση δύο δοκιμασιών, ο έλεγχος δηλαδή υποθέσεων για την διαφορά δύο μέσων. Συγκρίσεις τέτοιας μορφής είναι γνωστές σαν πολλαπλές συγκρίσεις. Οι αντίστοιχες μεθοδολογίες περιλαμβάνουν την μέθοδο της ελάχιστης σημαντικής διαφοράς του Fisher, την μέθοδο του εύρους του Tukey, την μέθοδο των γραμμικών αντιθέσεων ή συγκρίσεων του Scheffé και την μέθοδο του Bonferroni. Εμείς εδώ θα αναφέρουμε τα βασικά στοιχεία για δύο από τις παραπάνω μεθόδους πολλαπλών συγκρίσεων για τις οποίες λεπτομέρειες δίνονται, μεταξύ άλλων, στα βιβλία των Neter and Wasserman (1975) και Scheffé (1959).

(1) Μέθοδος της ελάχιστης σημαντικής διαφοράς

Έστω ότι μας ενδιαφέρει να συγκρίνουμε τις δοκιμασίες μ και ν . Ένας (αμερόληπτος) εκτιμητής της διαφοράς $\mu_u - \mu_\nu$ δίνεται



από τη μεταβλητή $\bar{Y}_{.u} - \bar{Y}_{.v}$ που έχει κανονική κατανομή με $E(\bar{Y}_{.u} - \bar{Y}_{.v}) = \mu_u - \mu_v$ και $\text{Var}(\bar{Y}_{.u} - \bar{Y}_{.v}) = \sigma^2(1/n_u + 1/n_v)$. Επίσης, $SS_{\text{res}}/\sigma^2 \sim \chi^2_{n-r}$ με τα $\bar{Y}_{.u} - \bar{Y}_{.v}$ και SS_{res} ανεξάρτητα. Έτσι

$$\frac{\bar{Y}_{.u} - \bar{Y}_{.v} - (\mu_u - \mu_v)}{\sqrt{MS_{\text{res}}(1/n_u + 1/n_v)}} \sim t_{n-r}. \quad (6.10)$$

Αν χρησιμοποιήσουμε την σχέση (6.10) μπορούμε να ελέγξουμε διάφορες υποθέσεις για την διαφορά $\mu_u - \mu_v$ για οποιοδήποτε ζευγάρι δοκιμασιών (u, v) . Έτσι, με επίπεδο σημαντικότητας α , η μηδενική υπόθεση $H_0: \mu_u - \mu_v = 0$ με εναλλακτική $H_a: \mu_u - \mu_v \neq 0$ απορρίπτεται αν

$$|\bar{Y}_{.u} - \bar{Y}_{.v}| > t_{\alpha/2, n-r} \sqrt{MS_{\text{res}}(1/n_u + 1/n_v)}$$

ή πάλι η $H_0: \mu_u - \mu_v \leq 0$ έναντι της $H_a: \mu_u - \mu_v > 0$ απορρίπτεται αν

$$\bar{Y}_{.u} - \bar{Y}_{.v} > t_{\alpha, n-r} \sqrt{MS_{\text{res}}(1/n_u + 1/n_v)}.$$

Έστω ότι, για το παράδειγμα της ΑΒΕΖ, θα μας ενδιέφερε να διαπιστώσουμε κατά πόσον οι μέσες πωλήσεις των τεσσάρων τύπων ζυμαρικών, συγκρινόμενες ανά δύο, παρουσιάζουν σημαντικές διαφορές. Για την σύγκριση μεταξύ του τύπου ζυμαρικών 3 και του τύπου 4, με $\alpha=0.05$, έχουμε: $u=3$, $v=4$, $n_3=3$, $n_4=2$, $\bar{Y}_{.3} = 19$, $\bar{Y}_{.4} = 27$, $MS_{\text{res}} = 7,67$ και $t_{0.025, 6} = 2.447$. Επειδή

$$|\bar{Y}_{.3} - \bar{Y}_{.4}| = 8 > t_{0.025, 6} \sqrt{7.67(1/3 + 1/2)} = 6.187$$

συμπεραίνουμε ότι $\mu_3 - \mu_4 \neq 0$ δηλαδή ότι υπάρχει σημαντική διαφορά στις μέσες επιδράσεις μεταξύ των τύπων (δοκιμασιών) 3 και 4 στις αντίστοιχες πωλήσεις. Ανάλογα γίνονται και οι υπόλοιπες συγκρίσεις

από τις οποίες προκύπτει ότι δεν υπάρχει σημαντική διαφορά μόνο στις μέσες επιδράσεις μεταξύ των τύπων 1 και 2.

Ένα $(1-\alpha) \cdot 100\%$ διάστημα εμπιστοσύνης για την διαφορά $\mu_U - \mu_V$ είναι το

$$\bar{Y}_{\cdot U} - \bar{Y}_{\cdot V} \pm t_{\alpha/2, n-r} \sqrt{MS_{res}(1/n_U + 1/n_V)}.$$

Ένα 95% διάστημα εμπιστοσύνης για την διαφορά των μέσων πωλήσεων των τύπων 3 και 4, $\mu_3 - \mu_4$, για τα δεδομένα της ABEZ είναι το $-8 \pm 2.447 \sqrt{7.67(1/3 + 1/2)}$ ή το $[-14.2, -1.8]$. Εκτιμάμε δηλαδή με συντελεστή εμπιστοσύνης 95% ότι οι μέσες πωλήσεις για τον τύπο 3 είναι λιγότερες από αυτές του τύπου 4 κατά έναν αριθμό που βρίσκεται μεταξύ του 1.8 και του 14.2.

Σημειώνουμε εδώ ότι η αντικατάσταση του αμερόληπτου εκτιμητή MS_{res} με τον αθροιστικό εκτιμητή

$$S^2 = [(n_U - 1)S_U^2 + (n_V - 1)S_V^2] / (n_U + n_V - 2)$$

στη σχέση (6.10) θα οδηγούσε στο γνωστό t -τεστ που χρησιμοποιείται για συγκρίσεις των μέσων δύο ανεξάρτητων $N(\mu_U, \sigma^2)$ και $N(\mu_V, \sigma^2)$ πληθυσμών. Η συστηματική όμως χρήση ενός τέτοιου τεστ για όλες τις δυνατές συγκρίσεις των r δοκιμασιών είναι επικίνδυνη γιατί το επίπεδο σημαντικότητας της όλης διαδικασίας είναι κατά πολύ μεγαλύτερο του αρχικά επιλεγμένου. Για παράδειγμα, έστω ότι σε μία μελέτη συγκρίνονται τέσσερις δοκιμασίες. Τότε θα υπάρχουν συνολικά έξι ζεύγη μέσων τιμών για σύγκριση. Ας υποθέσουμε ότι τα έξι t -τεστ για τον έλεγχο των $H_{01}: \mu_1 = \mu_2, \dots, H_{06}: \mu_3 = \mu_4$ είναι ανεξάρτητα και ότι $\alpha = 0.05$. Τότε το επίπεδο σημαντικότητας α' της όλης διαδικασίας ελέγχου της $H_{01}: \mu_1 = \mu_2, \dots, H_{06}: \mu_3 = \mu_4$ είναι

$$\begin{aligned} \alpha' &= P(\text{η } H_0 \text{ απορρίπτεται} / H_0 \text{ αληθεύει}) \\ &= P(\text{τουλάχιστον μία από τις έξι } H_{0i} \text{ απορρίπτεται} / H_0 \text{ αληθεύει}) \\ &= P(\text{τουλάχιστον μία επιτυχία σε έξι δοκιμές με } p=\alpha) \\ &= 1 - P(\text{καμία επιτυχία σε έξι δοκιμές με } p=\alpha) \\ &= 1 - \binom{6}{0} \alpha^0 (1-\alpha)^6 = 0.265, \end{aligned}$$



που είναι πολύ μεγαλύτερο από το $\alpha=0.05$.

(11) Μέθοδος του εύρους

Έστω $\bar{Y}_{\max}$ και $\bar{Y}_{\min}$ η μεγαλύτερη και η μικρότερη από τις δειγματικές μέσες τιμές $\bar{Y}_{.j}$, $j=1, \dots, r$ αντίστοιχα. Αν $n_1=n_2=\dots=n_r=m$, η στατιστική συνάρτηση

$$Q = \frac{\bar{Y}_{\max} - \bar{Y}_{\min}}{\sqrt{MS_{res}/m}} \quad (6.11)$$

έχει την λεγόμενη Studentized Range κατανομή της οποίας το αντίστροφο (από πάνω) εκατοστιαίο σημείο Q_α εξαρτάται από τον αριθμό των μέσων r και τους βαθμούς ελευθερίας του MS_{res} και δίνεται από τον πίνακα V.

Αν χρησιμοποιήσουμε το στατιστικό (6.11) η μέθοδος του εύρους συνίσταται στην απόρριψη της μηδενικής $H_0: \mu_U - \mu_V = 0$ με εναλλακτική την $H_a: \mu_U - \mu_V \neq 0$ αν, για κάποιο επίπεδο σημαντικότητας α , είναι

$$|\bar{Y}_{.u} - \bar{Y}_{.v}| > Q_\alpha \sqrt{\frac{MS_{res}}{m}}.$$

Ένα $(1-\alpha)100\%$ διάστημα εμπιστοσύνης για την διαφορά $\mu_U - \mu_V$ είναι το

$$\bar{Y}_{.u} - \bar{Y}_{.v} \pm Q_\alpha \sqrt{MS_{res}/m}.$$

Η μέθοδος του εύρους που ισχύει όταν $n_1 = \dots = n_r$, είναι συντηρητική με την έννοια ότι δύσκολα δίνει σημαντικές διαφορές γι' αυτό και βασικά χρησιμοποιείται όταν η διαφορά μεταξύ των δειγματικών μέσων που αντιστοιχούν στις δοκιμασίες που θα θέλαμε να συγκρίνουμε είναι μεγάλη. Αντίθετα, η μέθοδος της ελάχιστης σημαντικής διαφοράς γενικά τείνει να δίνει σημαντικότερες

διαφορές από αυτές που πραγματικά υπάρχουν, γι'αυτό και χρησιμοποιείται σε περιπτώσεις όπου το ενδιαφέρον για κάποια σύγκριση έχει εκδηλωθεί πριν από την επεξεργασία των δεδομένων.

6.4.5 Έλεγχος για την καταλληλότητα του μοντέλου

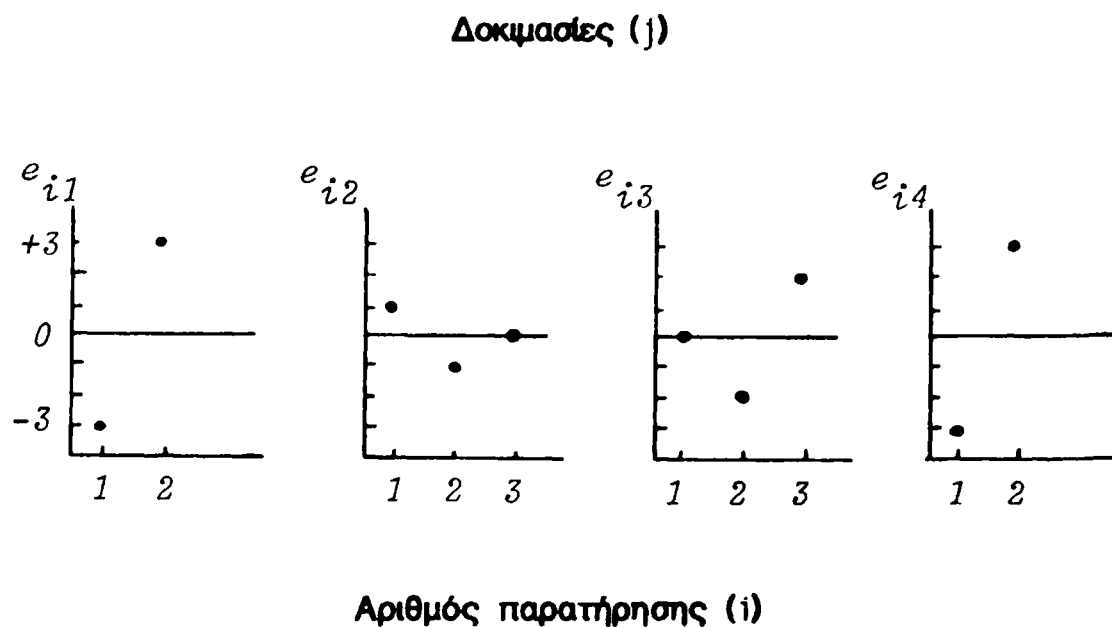
Στην Παλινδρόμηση, ιδιαίτερη έμφαση δόθηκε στην σημαντικότητα της εξέτασης της καταλληλότητας του μοντέλου όπως και στην αποτελεσματικότητα των γραφικών παραστάσεων των υπολοίπων στον εντοπισμό μεγάλων αποκλίσεων από τις υποθέσεις του μοντέλου (§ 5.11).

Ο έλεγχος της καταλληλότητας του μοντέλου της Ανάλυσης της Διακύμανσης κατά ένα παράγοντα είναι το ίδιο σημαντικός με αυτόν της Παλινδρόμησης οι δε γραφικές παραστάσεις των υπολοίπων είναι σε θέση, όπως στην Παλινδρόμηση, να εντοπίσουν τυχόν αποκλίσεις από τις υποθέσεις του μοντέλου. Σημειώνουμε εδώ ότι πριν από την χρησιμοποίηση ενός οποιουδήποτε στατιστικού μοντέλου θα πρέπει προηγουμένα να διαπιστωθεί κατά πόσον οι υποθέσεις του μοντέλου ικανοποιούνται από τα δεδομένα του εξετάζονται. Στην περίπτωση που υπάρξουν αποκλίσεις από τις υποθέσεις θα πρέπει ή να δοκιμαστούν διάφορα μέτρα επαναφοράς στην καταλληλότητα (μετασχηματισμοί δεδομένων κλπ.) ή να γίνει χρήση κάποιου άλλου στατιστικού μοντέλου (βλέπε π.χ. Neter and Wasserman, 1975) και μόνο όταν δεν διαπιστωθούν μεγάλες αποκλίσεις συνιστάται η παραπέρα ανάλυση των δεδομένων. Τα υπόλοιπα e_{ij} για το μοντέλο (6.1) ορίζονται, όπως είδαμε στην § 6.4.2 ως $e_{ij} = Y_{ij} - \bar{Y}_{.j}$ και επειδή ισχύει η (6.5) δεν είναι ανεξάρτητα. Τα κανονικοποιημένα υπόλοιπα $e_{ij}/\sqrt{MS_{res}}$ χρησιμοποιούνται συχνά αντί των e_{ij} στην ανάλυση των υπολοίπων.

Παράδειγμα γραφικών παραστάσεων των υπολοίπων ως



προς τον αριθμό των παρατηρήσεων για το μοντέλο της Ανάλυσης της Διακύμανσης κατά ένα παράγοντα αποτελεί το Σχήμα 6.4 στο



Σχήμα 6.4 Γραφικές παραστάσεις υπολοίπων, παράδειγμα ABEZ

οποίο περιέχονται τα υπόλοιπα του Πίνακα 6.2 για τα δεδομένα της ABEZ. Είναι χρήσιμο οι γραφικές παραστάσεις να γίνονται ξεχωριστά για κάθε δοκιμασία και να τοποθετούνται η μία δίπλα στην άλλη. Πριν αναφερθούμε περιληπτικά στις διάφορες μορφές που εμφανίζουν οι γραφικές παραστάσεις των υπολοίπων όταν υπάρχουν μεγάλες αποκλίσεις από τις υποθέσεις του μοντέλου, να σημειώσουμε ότι αποκλίσεις δεν φαίνεται να υπάρχουν για τα δεδομένα της ABEZ.

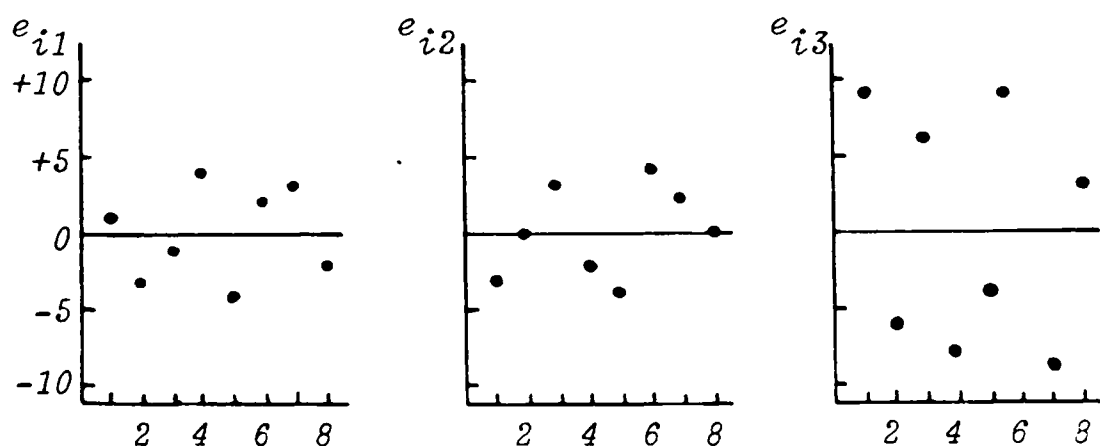
Όταν δεν υπάρχουν αποκλίσεις από την υπόθεση της σταθερότητας της διακύμανσης των τυχαίων σφαλμάτων για κάθε δοκιμασία, η γραφική παράσταση κάθε δοκιμασίας θα πρέπει να εμφανίζει μια διασπορά γύρω από το μηδέν της ίδιας μορφής. Παράδειγμα απόκλισης από την υπόθεση της σταθερότητας της διακύμανσης αποτελεί το Σχήμα 6.5 το οποίο παρέχει ενδείξεις ότι τα τυχαία σφάλματα της τρίτης δοκιμασίας έχουν μεγαλύτερη διακύμανση από αυτά των άλλων δύο.

Η γραφική παράσταση των υπολοίπων ως προς την χρονική σειρά των παρατηρήσεων, αν τα δεδομένα έχουν παρθεί

κατά χρονολογική σειρά, παρέχει ενδείξεις για τυχόν απόκλιση από την υπόθεση της ανεξαρτησίας των τυχαίων σφαλμάτων. Οι γραφικές παραστάσεις του Σχήματος 6.6 παρέχουν ενδείξεις ότι τα τυχαία σφάλματα δεν είναι ανεξάρτητα με τον χρόνο. Μια τέτοια περίπτωση συνήθως αντιμετωπίζεται συμπεριλαμβάνοντας στο μοντέλο μία επιπλέον μεταβλητή που να αντιπροσωπεύει την επίδραση του χρόνου. Οι επιδράσεις του χρόνου έχουν συχνά σαν αποτέλεσμα και την αύξηση ή την ελάττωση της διακύμανσης των τυχαίων σφαλμάτων με τον χρόνο.

Οι γραφικές παραστάσεις των υπολοίπων μπορούν επίσης εύκολα να δείξουν αν υπάρχουν ακραίες παρατηρήσεις, δηλαδή τιμές στα δεδομένα που διαφέρουν, συγκριτικά με τις άλλες, κατά πολύ από τον αντίστοιχο δειγματικό τους μέσο $\bar{Y}_{.j}$. Έτσι έχει επικρατήσει, όπως και στην Παλινδρόμηση, να απορρίπτεται μία ακραία παρατήρηση όταν αποδειγμένα αντιπροσωπεύει λάθος ανιγγραφής, υπολογιστικό λάθος ή παρόμοιες καταστάσεις.

Δοκμασίες (j)



Αριθμός παρατήρησης (i)

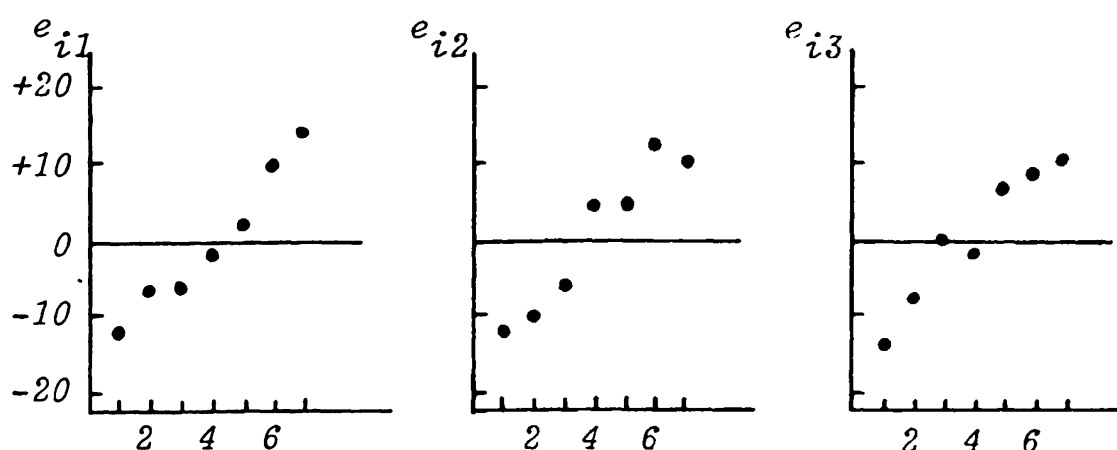
Σχήμα 6.5 Παράδειγμα γραφικής παράστασης υπολοίπων όταν η διακύμανση δεν είναι σταθερή για όλες τις δοκμασίες

Ενδείξεις για απόκλιση από την υπόθεση της



κανονικότητας των τυχαίων σφαλμάτων δίνονται, για σχετικά μεγάλα n_j , από τα κανονικοποιημένα υπόλοιπα $e_{ij}/\sqrt{MS_{res}}$. Όταν, για τα δεδομένα που εξετάζονται, το 95% περίπου από τα κανονικοποιημένα υπόλοιπα βρίσκονται στο διάστημα $(-2, 2)$ η υπόθεση της κανονικότητας δεν παραβιάζεται στην συγκεκριμένη μελέτη.

Δοκιμασίες (j)



Αριθμός παρατήρησης (i) κατά χρονολογική σειρά

Σχήμα 6.6 Παράδειγμα γραφικής παράστασης υπολοίπων όταν τα τυχαία σφάλματα δεν είναι ανεξάρτητα του χρόνου

6.4.6 Στατιστικά πακέττα

Έτοιμα προγράμματα ηλεκτρονικού υπολογιστή, ανάλογα με αυτά της Παλινδρόμησης (§ 5.16), κατάλληλα να επεξεργαστούν αριθμητικά τα δεδομένα σύμφωνα με την μεθοδολογία της Ανάλυσης της Διακύμανσης κατά ένα παράγοντα διατίθενται στο εμπόριο. Η χρήση των προγραμμάτων αυτών που περιέχονται, μεταξύ άλλων, στα στατιστικά πακέττα BMDP και SPSS είναι απλή και δεν απαιτεί ιδιαίτερες γνώσεις προγραμματισμού.

Στις προηγούμενες παραγράφους οι διάφοροι υπολογισμοί

που υπεισέρχονται στην μεθοδολογία της Ανάλυσης της Διακύμανσης κατά ένα παράγοντα χρησιμοποιήθηκαν για την στατιστική ανάλυση των δεδομένων της ΑΒΕΖ. Εδώ θα δώσουμε τα αποτελέσματα για τα δεδομένα της ΑΒΕΖ όπως πάρθηκαν από τον ηλεκτρονικό υπολογιστή του Πανεπιστημίου Ιωαννίνων χρησιμοποιώντας το πρόγραμμα BMDP1V που αντιστοιχεί στην μεθοδολογία της Ανάλυσης της Διακύμανσης κατά ένα παράγοντα.

Τα αριθμητικά αποτελέσματα που ακολουθούν έχουν παραχθεί με χρήση ηλεκτρονικού υπολογιστή και περιέχουν:

- (i) Τους εκτιμητές ελαχίστων τετραγώνων των μέσων μ_j που δίνονται από τις δειγματικές μέσες τιμές $\bar{Y}_{.j}$ (§ 6.4.2).
- (ii) Τον πίνακα ΑΝΑΔΙΑ και το F-τεστ για τον έλεγχο της σημαντικότητας της υπόθεσης ότι $\mu_1 = \dots = \mu_r$, ότι δηλαδή οι επιδράσεις των r δοκιμασιών είναι ίσες (§ 6.4.3).
- (iii) Τα t-τεστ για τον έλεγχο της σημαντικότητας της υπόθεσης ότι $\mu_i = \mu_j$, $i \neq j$ ότι δηλαδή οι μέσες επιδράσεις των δοκιμασιών i και j είναι ίσες (§ 6.4.4).

ΠΑΡΑΔΕΙΓΜΑ ΑΒΕΖ

(i)

ESTIMATES OF MEANS

	P1	P2	P3	P4	TOTAL
	1	2	3	4	5
PACKAGES	1 15.0000	13.0000	19.0000	27.0000	18.0000

(ii)

ONE WAY ANALYSIS OF VARIANCE FOR VARIABLE PACKAGES

ANALYSIS OF VARIANCE TABLE

SOURCE OF VARIANCE	DF.	SUM OF SQ.	MEAN SQ.	F-VALUE	PROB(TAL)
EQUALITY OF CELL MEANS	3	258.0000	86.0000	11.2174	.0071
ERROR	6	46.0000	7.6667		



(iii)

T-TEST MATRIX FOR GROUP MEANS ON 6 DEGREES OF FREEDOM

		P1	P2	P3	P4
		1	2	3	4
P1	1	0.0000			
P2	2	-.7913	0.0000		
P3	3	1.5825	2.6540	0.0000	
P4	4	4.3339	5.5388	3.1650	0.0000

PROBABILITIES FOR THE T-VALUES ABOVE

		P1	P2	P3	P4
		1	2	3	4
P1	1	1.0000			
P2	2	.4589	1.0000		
P3	3	.1646	.0378	1.0000	
P4	4	.0049	.0015	.0194	1.0000

6.5 Αρχές σχεδιασμού πειραμάτων κατά ένα και δύο παράγοντες

Στο παράδειγμα της ABEZ που αναφερθήκαμε στα προηγούμενα, μπορεί να φαίνεται λογικό το συμπέρασμα που καταλήξαμε ότι δηλαδή οι τέσσερις τύποι ζυμαρικών επιδρούν διαφορετικά στις μέσες πωλήσεις. Πόσο "λογικό" είναι όμως αυτό το συμπέρασμα; Μήπως αποκλείεται οι διαφορές στις μέσες πωλήσεις να μην οφείλονται στις διαφορετικές προτιμήσεις του πληθυσμού στους τέσσερις τύπους ζυμαρικών αλλά σε άλλες αιτίες, όπως: σε κάποια ίσως διαφήμιση των τύπων 3 και 4 που έτυχε να γίνεται την περίοδο της συλλογής των δεδομένων ή στο γεγονός ότι τα καταστήματα που εκλέχτηκαν να πωλούν τον τύπο 4 (κοφτό μακαρονάκι) τροφοδοτούν, για παράδειγμα, τους δημόσιους βρεφονηπιακούς σταθμούς της πόλης; Είναι βέβαια πολύ πιθανό οι διαφορές μεταξύ

των τεσσάρων μέσων $\bar{Y}_{.j}$ να οφείλονται βασικά στις διαφορετικές επιδράσεις των τεσσάρων τύπων όμως εμείς απλώς αναφέραμε δύο άλλους παράγοντες που θα μπορούσαν να θεωρηθούν υπεύθυνοι για την διαφορά στις μέσες πωλήσεις που προέκυψε. Είναι σημαντικό να μην ξεχνάμε ότι ένα στατιστικό τεστ ελέγχου μπορεί να δείξει ότι οι διαφορές που υπάρχουν σε κάποιους δειγματικούς μέσους είναι πολύ μεγάλες για να αποδοθούν στην τύχη, δεν μπορεί όμως να δώσει καμιά ένδειξη για το πως προέκυψαν αυτές οι διαφορές.

Γενικά, αν θέλουμε να δείξουμε ότι ένας παράγοντας (μεταξύ διαφόρων που πιθανά επιδρούν) μπορεί να θεωρηθεί η αιτία κάποιου φαινομένου που μελετάμε, θα πρέπει με κάποιο τρόπο να σιγουρευτούμε ότι κανένας από τους άλλους παράγοντες που λέγονται και εξωγενείς δεν μπορεί να θεωρηθεί υπεύθυνος. Για παράδειγμα θα μπορούσαμε να σχεδιάσουμε ένα αυστηρά ελεγχόμενο πείραμα, στο οποίο όλες οι μεταβλητές εκτός από αυτήν που μας ενδιαφέρει να μελετήσουμε μένουν σταθερές (ελεγχόμενες). Θα μπορούσε π.χ. η ΑΒΕΖ την βδομάδα της συλλογής των δεδομένων να σταματήσει τις διαφημίσεις για τους τύπους ζυμαρικών 3 και 4 ώστε να αποτρέψει τυχόν επιδράσεις στις πωλήσεις του παράγοντα διαφήμιση. Κάτω από τέτοιες προσεκτικά ελεγχόμενες συνθήκες, διαφορές στους δειγματικούς μέσους θα μπορούσαν πράγματι να αποδοθούν σε διαφορετικές επιδράσεις των επιπέδων του παράγοντα που μελετάται. Από την άλλη μεριά όμως, ένα τέτοιο συμπέρασμα θα μπορούσε να γενικευθεί μόνο για καταστάσεις εντελώς παρόμοιες με αυτή του πειράματος αφού δεν γνωρίζουμε αν, κάτω από διαφορετικές συνθήκες, θα υπήρχαν οι ίδιες διαφορές.

Σε πολλές περιπτώσεις, αυστηρά ελεγχόμενα πειράματα σαν αυτό που περιγράψαμε δεν μας παρέχουν πραγματικά το είδος της πληροφορίας που θέλουμε. Έτσι εξετάζουμε διαφορετικές λύσεις, μία από τις οποίες βρίσκεται σε πλήρη αντίθεση με αυτή του αυστηρά ελεγχόμενου πειράματος. Μπορούμε δηλαδή να κατασκευάσουμε πειράματα στα οποία κανένας από τους εξωγενείς παράγοντες δηλαδή παράγοντες που δεν μας ενδιαφέρουν να ελέγχεται, αλλά στα οποία "προστατευόμαστε" από την επίδραση τους



με την τυχαιοποίηση. Σχεδιάζουμε δηλαδή τα πειράματα έτσι ώστε να μπορούμε να αποδώσουμε στην τύχη το μέρος της μεταβλητότητας που δεν οφείλεται στον παράγοντα που μελετάμε. Έτσι, επιλέγουμε τυχαία τις πειραματικές μονάδες (τα δέκα καταστήματα για το παράδειγμα της ΑΒΕΖ) και επίσης τυχαία επιλέγουμε σε ποιές από αυτές θα δοθεί η κάθε δοκιμασία (οι τέσσερις τύποι ζυμαρικών για το παράδειγμα της ΑΒΕΖ).

Όταν όλες οι μη ελεγχόμενες πηγές μεταβλητότητας, εκτός αυτής που μελετάμε, μπορούν να συμπεριληφθούν στον όρο "τύχη" το σχετικό σχέδιο πειράματος είναι γνωστό σαν π λ ή ρ ε ς τ υ χ α ι ο π ο ι η μέ ν ο π ε ί ρ α μ α και η ανάλυση δεδομένων που προέρχονται από αυτό δίνεται από την Ανάλυση της Διακύμανσης κατά ένα παράγοντα.

Δεν είναι δύσκολο βέβαια να δούμε ότι η τυχαιοποίηση προστατεύει από τις επιδράσεις των εξωγενών παραγόντων μόνο από πιθανοθεωρητικής σκοπιάς. Στην περίπτωση της ΑΒΕΖ για παράδειγμα, είναι δυνατόν παρ'όλο ότι τα τρία καταστήματα που θα πωλούσαν τον τύπο 1 επιλέχθηκαν τυχαία, αυτό δεν αποκλείει την περίπτωση τα τρία αυτά καταστήματα να βρίσκονται σε περιοχή με άτομα που δεν αρέσκονται γενικά σε οποιονδήποτε τύπο ζυμαρικών. Αυτή είναι και μία από τις αιτίες που συνήθως προσπαθούμε να ελέγχουμε μερικούς από τους εξωγενείς παράγοντες και να επιβάλλουμε την τυχαιοποίηση στους άλλους, να χρησιμοποιήσουμε δηλαδή έναν σχεδιασμό που θα αποτελεί συνδιασμό των δύο ακραίων μορφών που περιγράψαμε.

Για να εισάγουμε τώρα ένα άλλο σημαντικό στοιχείο που υπεισέρχεται στον σχεδιασμό πειραμάτων, ας θεωρήσουμε τα παρακάτω δεδομένα που δίνουν τον χρόνο σε λεπτά που κάποιο άτομο χρειάστηκε για να πάει με το αυτοκίνητό του στην δουλειά από Δευτέρα μέχρι Παρασκευή, ακολουθώντας τέσσερις διαφορετικές πορείες (δεδομένα τεσσάρων εβδομάδων).



Ημέρα Πορεία	Δ	Τ	Τ	Π	Π
I	22	26	25	25	31
II	25	27	28	26	29
III	26	29	33	30	33
IV	26	28	27	30	30

Οι μέσες τιμές των τεσσάρων δειγμάτων είναι 25.8, 27.0, 30.2, 28.2 και καθώς οι διαφορές μεταξύ αυτών είναι αρκετά μεγάλες θα φαινόταν αρκετά λογικό αν συμπεραίναμε ότι υπάρχουν κάποιες πραγματικές διαφορές στον αληθή μέσο χρόνο που ένα άτομο χρειάζεται μέχρι να πάει οδηγώντας στην δουλειά του δια μέσου των τεσσάρων διαφορετικών δρόμων του παραδείγματος. Αυτό όμως δεν προκύπτει από την Ανάλυση της Διακύμανσης κατά έναν παράγοντα. Οι σχετικοί υπολογισμοί δίνουν $F=2.8 < F_{0.05,3,16}=3.24$ και η H_0 ότι δηλαδή δεν υπάρχουν διαφορές στους μέσους χρόνους των τεσσάρων δρόμων δεν μπορεί να απορριφθεί.

Βέβαια, η H_0 μπορεί πράγματι να αληθεύει, παρατηρούμε όμως ότι υπάρχουν όχι μόνο σημαντικές διαφορές μεταξύ των τεσσάρων μέσων, αλλά επίσης και μεταξύ των τιμών μέσα στα δείγματα. Στο πρώτο δείγμα οι τιμές ποικίλουν από 22 μέχρι 31, στο δεύτερο από 25 μέχρι 29, στο τρίτο από 26 μέχρι 33 και στο τέταρτο από 26 μέχρι 30. Παρατηρούμε επίσης ότι, σε κάθε δείγμα, η τιμή της Δευτέρας είναι η μικρότερη και αυτή της Παρασκευής η μεγαλύτερη. Αυτό ίσως δείχνει ότι η μεταβλητότητα μέσα στα δείγματα μπορεί να οφείλεται σε διαφορές των συνθηκών οδήγησης που επικρατούν τις διάφορες μέρες της βδομάδας. Αν πράγματι συμβαίνει αυτό τότε η μεταβλητότητα που οφείλεται στις συνθήκες οδήγησης συμπεριελήφθηκε στο άθροισμα τετραγώνων των υπολοίπων (SS_{res}) της Ανάλυσης της Διακύμανσης κατά ένα παράγοντα, επηρέασε έτσι το στατιστικό F και μπορεί να συνέβαλε στο ότι το αποτέλεσμα δεν ήταν σημαντικό.

Για να αποφύγουμε παρόμοιες καταστάσεις θα μπορούσαμε να κρατήσουμε σταθερό τον (εξωγενή) δεύτερο παράγοντα όμως μια τέτοια τακτική σπάνια δίνει την πληροφορία που



θέλουμε. Στο παράδειγμα μας λόγου χάριν, θα μπορούσαμε να περιοριστούμε σε δεδομένα Δευτέρας μόνο, αλλά τότε δεν θα μπορούσαμε με βεβαιότητα να γενικεύσουμε τα συμπεράσματα της μελέτης μας για όλες τις μέρες της βδομάδας. Μία άλλη δυνατότητα μας παρέχεται αν μεταβάλλουμε σκόπιμα τον εξωγενή παράγοντα όσο το δυνατόν περισσότερο έτσι ώστε η μεταβλητότητα που οφείλεται σ' αυτόν να μπορεί να μετρηθεί ξεχωριστά και έτσι να εξαληφτεί από το άθροισμα τετραγώνων των υπολοίπων. Αυτό σημαίνει ότι πρέπει να σχεδιάσουμε το πείραμα έτσι ώστε η ανάλυση των δεδομένων να βασιστεί στη μεθοδολογία της Ανάλυσης της Διακύμανσης κατά δύο παράγοντες στην οποία η ολική μεταβλητότητα των δεδομένων χωρίζεται σε τρεις συνιστώσες που οφείλονται, αντίστοιχα, στις δοκιμασίες (στο παράδειγμά μας, οι τέσσερις πορείες), τον εξωγενή παράγοντα (στο παράδειγμά μας, συνθήκες οδήγησης στις διαφορετικές μέρες της βδομάδας) και στο πειραματικό ή τυχαίο σφάλμα.

Αυτό που εδώ προτείνουμε λέγεται τετραγωνοποίηση με τις μέρες της βδομάδας να αναφέρονται σαν τα τετράγωνα. Γενικά, τα τετράγωνα αποτελούνται από τα επίπεδα στα οποία κρατάμε σταθερό τον εξωγενή παράγοντα έτσι ώστε να μπορέσουμε να υπολογίσουμε την συνεισφορά του στην συνολική μεταβλητότητα των δεδομένων. Όταν κάθε δοκιμασία εμφανίζεται τον ίδιο αριθμό φορές σε κάθε τετράγωνο (στο παράδειγμά μας, κάθε πορεία ακολουθείται μία φορά για κάθε μία μέρα της βδομάδας) το σχέδιο του πειράματος λέγεται *πλήρες σχέδιο τετραγώνων*. Αν, επιπλέον, οι δοκιμασίες κατανέμονται τυχαία μέσα στο κάθε τετράγωνο (στο παράδειγμά μας, θα κατανέμαμε τυχαία τις τέσσερις πορείες ανάμεσα στις τέσσερις Δευτέρες, τέσσερις Τρίτες κλπ.) τότε λέμε ότι το σχέδιο του πειράματος είναι ένα (πλήρες) *τυχαίοποιημένο σχέδιο τετραγώνων* και η ανάλυση δεδομένων που προέρχονται από αυτό δίνεται από την Ανάλυση της Διακύμανσης κατά δύο παράγοντες.



6.6 Ανάλυση της Διακύμανσης κατά δύο παράγοντες

Στα προηγούμενα μελετήσαμε την επίδραση ενός παράγοντα πάνω σε μία εξαρτημένη μεταβλητή. Το ενδιαφέρον μας στρέφεται τώρα σε μελέτες στις οποίες διερευνάται η ταυτόχρονη επίδραση δύο παραγόντων πάνω στην εξαρτημένη μεταβλητή με τον περιορισμό ότι οι δύο παράγοντες δεν αλληλοεπιδρούν. Με την έννοια "αλληλοεπίδραση" εννοούμε μία ειδική αλληλοεξάρτηση των δύο παραγόντων το αποτέλεσμα της οποίας είναι οι παρατηρήσεις (αποκρίσεις) να φανερώνουν μία συνεισφορά που να οφείλεται στην αλληλοεπίδραση αυτή, επι πλέον της συνεισφοράς του κάθε παράγοντα. Σαν ενδεικτικό παράδειγμα, έστω ότι δύο λιπασματα δοκιμάζονται στην παραγωγή καλαμποκιού σε δύο είδη εδάφους. Οι παράγοντες "λίπασμα" και "έδαφος" δεν αλληλοεπιδρούν όταν η διαφορά παραγωγής εξαιτίας των δύο λιπασμάτων στο ένα είδος εδάφους είναι ίδια με την αντίστοιχη στο άλλο είδος εδάφους.

Εδώ θα περιοριστούμε στην μεθοδολογία που εφαρμόζεται όταν οι δύο παράγοντες δεν αλληλοεπιδρούν, όταν δηλαδή η συνεισφορά τους είναι προσθετική. Λεπτομέρειες για την περίπτωση που οι δύο παράγοντες αλληλοεπιδρούν, όπως και για μελέτες με περισσότερους από δύο παράγοντες ή πιο πολύπλοκα μοντέλα Ανάλυσης της Διακύμανσης που ξεφεύγουν από τον σκοπό ενός εισαγωγικού βιβλίου, δίνονται μεταξύ άλλων στα βιβλία των Cochran and Cox (1957), Neter and Wasserman (1975) και Scheffé (1959).

Τα βασικά στοιχεία στην Ανάλυση της Διακύμανσης κατά δύο παράγοντες είναι απλά και ανάλογα με αυτά της Ανάλυσης της Διακύμανσης κατά ένα παράγοντα. Η εξαρτημένη μεταβλητή παίρνει τιμές που εξαρτώνται από την παρουσία ή απουσία των επιπέδων των παραγόντων. Με άλλα λόγια σε κάθε ζεύγος επιπέδων των δύο παραγόντων αντιστοιχούν μία ή περισσότερες τιμές της εξαρτημένης μεταβλητής. Για κάθε συνδυασμό επιπέδων (δοκιμασία) οι τιμές της εξαρτημένης μεταβλητής θεωρούνται τυχαίες παρατηρήσεις από μία κατανομή της οποίας η μέση τιμή εξαρτάται από τον συνδυασμό



επιπέδων των παραγόντων. Στην πιο απλή περίπτωση οι τυχαίες παρατηρήσεις θεωρούνται ανεξάρτητες και οι κατανομές της εξαρτημένης μεταβλητής για κάθε ζεύγος επιπέδων έχουν την ίδια διακύμανση. Οι τιμές της εξαρτημένης μεταβλητής προκύπτουν είτε από παρατηρητικές μελέτες όπου εξετάζονται μεταβλητές ή παράγοντες όπως εμφανίζονται στη φύση (π.χ. δειγματοληψίες, δημοσκοπήσεις) είτε από πειραματικές μονάδες κάτω από ελεγχόμενες συνθήκες.

Στις μελέτες κατά ένα παράγοντα δεν κάναμε καμία υπόθεση για την μορφή των r δοκιμασιών. Έτσι, τυπικά το σύνολο των ζευγών των επιπέδων των παραγόντων σε μελέτες κατά δύο παράγοντες θα μπορούσε να αποτελέσει τις r δοκιμασίες, η ανάλυση των οποίων θα βασιζονταν στη μεθοδολογία που αναπτύχθηκε στις προηγούμενες παραγράφους. Η ανάγκη για καινούργιες μεθόδους έγκειται στο γεγονός ότι επιδίωξη μας είναι η μελέτη της επίδρασης κάθε παράγοντα ξεχωριστά.

6.6.1 Το στατιστικό μοντέλο

Έστω ότι οι παράγοντες που εξετάζονται είναι ο A σε l επίπεδα και ο B σε m επίπεδα. Ο δείκτης i θα συμβολίζει τα επίπεδα του παράγοντα A , $i=1, \dots, l$ και ο δείκτης j θα συμβολίζει τα επίπεδα του παράγοντα B , $j=1, \dots, m$. Η παρατήρηση που αντιστοιχεί στην (i,j) δοκιμασία, που έχει δηλαδή δεχθεί ταυτόχρονα τις επιδράσεις του επιπέδου i του A παράγοντα και του επιπέδου j του B παράγοντα θα συμβολίζεται με Y_{ij} . Στο μοντέλο που εξετάζουμε δεχόμαστε ότι για κάθε δοκιμασία (i,j) υπάρχει μία μόνο παρατήρηση, δηλαδή συνολικά τα δεδομένα αποτελούνται από lm παρατηρήσεις.

Το μοντέλο έχει την μορφή

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

(6.12)



όπου Y_{ij} είναι η τιμή (απόκριση) της εξαρτημένης μεταβλητής για την (i,j) δοκιμασία. Τα ε_{ij} λέγονται τυχαία σφάλματα και υποτίθενται ανεξάρτητες $N(0, \sigma^2)$ τυχαίες μεταβλητές. Η παράμετρος μ (σταθερά, κοινή για όλες τις παρατηρήσεις) λέγεται μεγάλη μέση τιμή, η παράμετρος α_i (σταθερά για κάθε επίπεδο του παράγοντα A) συμβολίζει την επίδραση του i επιπέδου του A παράγοντα και η παράμετρος β_j (σταθερά για κάθε επίπεδο του παράγοντα B) συμβολίζει την επίδραση του j επιπέδου του B παράγοντα. Για τις παραμέτρους α_i και β_j υποθέτουμε ότι

$$\sum \alpha_i = 0 \quad \text{και} \quad \sum \beta_j = 0.$$

Το αρχικό ενδιαφέρον στις μελέτες κατά δύο παράγοντες έγκειται στο κατά πόσον οι μέσες επιδράσεις των επιπέδων του παράγοντα A είναι ίσες και κατά πόσον οι μέσες επιδράσεις των επιπέδων του παράγοντα B είναι ίσες. Οι υποθέσεις δηλαδή που ενδιαφέρουν είναι οι

$$H_0^A: \alpha_1 = \alpha_2 = \dots = \alpha_l = 0 \text{ έναντι της } H_a^A: \text{όχι όλα τα } \alpha_i \text{ ίσα με μηδέν,}$$

και

$$H_0^B: \beta_1 = \beta_2 = \dots = \beta_m = 0 \text{ έναντι της } H_a^B: \text{όχι όλα τα } \beta_j \text{ ίσα με μηδέν.}$$

6.6.2 Εκτίμηση των παραμέτρων με τη μέθοδο των ελαχίστων τετραγώνων

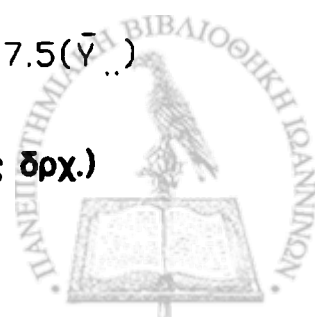
Έστω ότι έχουμε στην διάθεση μας τις παρατηρήσεις Y_{ij} , $i=1, \dots, l$, $j=1, \dots, m$ που πάρθηκαν μετά από κάποιο πείραμα. Οι τιμές των παραμέτρων του μοντέλου (6.12) είναι άγνωστες και κατά συνέπεια θα πρέπει να τις εκτιμήσουμε από τα δεδομένα. Πριν περιγράψουμε την μέθοδο των ελαχίστων τετραγώνων την οποία θα χρησιμοποιήσουμε για να βρούμε τους εκτιμητές, θα δώσουμε ορισμένους συμβολισμούς σε συνδιασμό με ένα ενδεικτικό παράδειγμα: Η βιομηχανία ΦΑΓΕ ενδιαφέρεται να



διαπιστώσει κατά πόσο υπάρχουν επιδράσεις της τιμής πώλησης και του τρόπου διαφήμισης στις πωλήσεις ενός προϊόντος της. Τρεις τιμές πώλησης (45, 50, 55 δρχ.) και δύο τύποι διαφήμισης (τοπικός τύπος, ραδιόφωνο) αποφασίστηκε να μελετηθούν. Έστω Α ο παράγων "τιμή πώλησης" και Β ο παράγων "διαφήμιση". Έξι καταστήματα με περίπου ίσες πωλήσεις επιλέχθηκαν, ένα σε κάθε μία από έξι διαφορετικές πόλεις. Σε κάθε ένα από τα καταστήματα επιδόθηκε τυχαία μία από τις δοκιμασίες, δηλαδή ένας από τους έξι δυνατούς συνδιασμούς των τριών επιπέδων του παράγοντα Α και των δύο επιπέδων του παράγοντα Β. Οι εισπράξεις (σε χιλιάδες δρχ.) από τις πωλήσεις του προϊόντος σε διάστημα μίας βδομάδας, δίνονται στον Πίνακα 6.3 (υποθετικά δεδομένα). Υποθέτουμε ότι οι παράγοντες Α και Β δεν αλληλοεπιδρούν κάτι που μπορεί να αποδειχθεί (βλέπε π.χ. Neter and Wasserman, 1975) ότι πράγματι ισχύει. Στον ίδιο πίνακα δίνονται οι συμβολισμοί (σε παρένθεση) που αντιστοιχούν στα δεδομένα της ΦΑΓΕ. Όπως προαναφέραμε, το Y_{ij} παριστά την παρατήρηση (εισπράξεις) της πειραματικής μονάδας (κατάστημα) στην οποία επιδρά ο συνδιασμός επιπέδων (i, j) του i επιπέδου του Α παράγοντα (τιμή πώλησης) και του j επιπέδου του Β παράγοντα (τύπος διαφήμισης). Το άθροισμα των παρατηρήσεων που αντιστοιχεί στο i επίπεδο του παράγοντα Α συμβολίζεται με $Y_{i.} = \sum_j Y_{ij}$.

Παράγοντας Α (τιμή πώλησης)	Παράγοντας Β (τύπος διαφήμισης)		Μέση Τιμή
	Τύπος (j-1)	Ραδιόφωνο (j-2)	
55 δρχ (i=1)	14(Y_{11})	10(Y_{12})	24($Y_{1.}$) 12.0($\bar{Y}_{1.}$)
50 (i=2)	21(Y_{21})	18(Y_{22})	39($Y_{2.}$) 19.5($\bar{Y}_{2.}$)
45 (i=3)	22(Y_{31})	20(Y_{32})	42($Y_{3.}$) 21.0($\bar{Y}_{3.}$)
Άθροισμα	57($Y_{.1}$)	48($Y_{.2}$)	105($Y_{..}$)
Μέση τιμή	19($\bar{Y}_{.1}$)	16($\bar{Y}_{.2}$)	17.5($\bar{Y}_{..}$)

Πίνακας 6.3 Δεδομένα εισπράξεων (σε χιλιάδες δρχ.)
ανά κατάστημα, παράδειγμα ΦΑΓΕ



Ο αντίστοιχος δειγματικός μέσος συμβολίζεται με $\bar{Y}_{i.}; \bar{Y}_{i.} = Y_{i.}/m$. Ανάλογοι είναι και οι συμβολισμοί για τα επίπεδα του παράγοντα Β. Τέλος το συνολικό άθροισμα των παρατηρήσεων συμβολίζεται με $Y_{..}; Y_{..} = \sum \sum Y_{ij}$ και η δειγματική γενική μέση τιμή με $\bar{Y}_{..}; \bar{Y}_{..} = Y_{..}/1m$.

Οι εκτιμητές των ελαχίστων τετραγώνων των παραμέτρων προκύπτουν αν ελαχιστοποιήσουμε ως προς μ, α_i, β_j την συνάρτηση

$$Q(\mu, \alpha_1, \dots, \alpha_l, \beta_1, \dots, \beta_m) = \sum \sum (Y_{ij} - \mu - \alpha_i - \beta_j)^2$$

υπό την προϋπόθεση ότι $\sum \alpha_i = 0$ και $\sum \beta_j = 0$.

Μπορεί εύκολα ναδειχθεί ότι οι εκτιμητές ελαχίστων τετραγώνων των παραμέτρων μ, α_i, β_j δίνονται από τις σχέσεις:

$$\hat{\mu} = \bar{Y}_{..}$$

$$\hat{\alpha}_i = \bar{Y}_{i.} - \bar{Y}_{..}, \quad i=1, \dots, l \quad (6.13)$$

$$\hat{\beta}_j = \bar{Y}_{.j} - \bar{Y}_{..}, \quad j=1, \dots, m.$$

Για τα δεδομένα της ΦΑΓΕ έχουμε

$$\hat{\mu} = 17.5, \quad \hat{\alpha}_1 = -5.5, \quad \hat{\alpha}_2 = 2.0, \quad \hat{\alpha}_3 = 3.5, \quad \hat{\beta}_1 = 1.5, \quad \hat{\beta}_2 = -1.5.$$

Μπορεί να αποδειχθεί ότι οι εκτιμητές ελαχίστων τετραγώνων που δίνονται από τις σχέσεις (6.13) είναι αμερόληπτοι, ταυτίζονται δε με τους αντίστοιχους εκτιμητές μεγίστης πιθανοφάνειας.

Τα υπόλοιπα e_{ij} είναι, όπως και στην Ανάλυση της Διακύμανσης κατά ένα παράγοντα, πολύ χρήσιμα στην μελέτη της καταλληλότητας του μοντέλου. Ορίζονται από την σχέση

$$e_{ij} = Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j = Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$$



Παράγοντας Α (τιμή πώλησης)	Παράγοντας Β (τύπος διαφήμισης)		Σύνολο
	j-1	j-2	
i=1	0.5	-0.5	0
i=2	0.0	0.0	0
i=3	-0.5	0.5	0
Σύνολο	0	0	0

Πίνακας 6.4 Υπόλοιπα, παράδειγμα ΦΑΓΕ

Για τα δεδομένα (Πίνακας 6.3) του παραδείγματος της ΦΑΓΕ τα υπόλοιπα e_{ij} δίνονται στον Πίνακα 6.4. Όπως διαπιστώνουμε και από τον Πίνακα τα υπόλοιπα ικανοποιούν τις συνθήκες $\sum_i e_{ij}=0$ και $\sum_j e_{ij}=0$.

6.6.3 Η Ανάλυση της Διακύμανσης

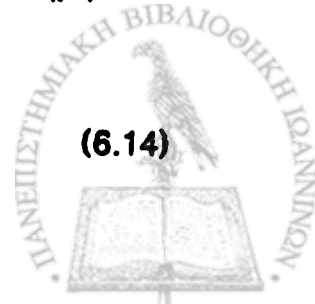
Η ολική μεταβλητότητα των παρατηρήσεων Y_{ij} , που ισούται με το ολικό άθροισμα τετραγώνων (SS_{tot}) των αποκλίσεων $Y_{ij}-\bar{Y}_{..}$, δίνεται από την σχέση

$$SS_{tot} = \sum \sum (Y_{ij} - \bar{Y}_{..})^2.$$

Η ολική απόκλιση $Y_{ij}-\bar{Y}_{..}$ μπορεί να διασπασθεί σε τρία επί μέρους τμήματα. Στην απόκλιση των δειγματικών μέσων των επιπέδων του παράγοντα Α από την δειγματική γενική μέση τιμή, στην απόκλιση των δειγματικών μέσων των επιπέδων του παράγοντα Β από την δειγματική γενική μέση τιμή και στην απόκλιση των παρατηρήσεων από την εκτιμώμενη τιμή τους (υπόλοιπα). Ισχύει δηλαδή,

$$Y_{ij} - \bar{Y}_{..} = \bar{Y}_{i.} - \bar{Y}_{..} + \bar{Y}_{.j} - \bar{Y}_{..} + Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$$

(6.14)



Αν υψώσουμε την (6.14) στο τετράγωνο και αθροίσουμε τελικά έχουμε

$$\begin{aligned} \sum \sum (y_{ij} - \bar{y}_{..})^2 &= \sum \sum (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum \sum (\bar{y}_{.j} - \bar{y}_{..})^2 \\ &\quad + \sum \sum (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 \end{aligned} \quad (6.15)$$

ή συμβολικά

$$SS_{tot} = SS_A + SS_B + SS_{res}.$$

Η σχέση (6.15) λέει ότι ένα μέρος από την ολική μεταβλητότητα των παρατηρήσεων y_{ij} , που εκφράζεται με το ολικό άθροισμα τετραγώνων, οφείλεται στην επίδραση του παράγοντα A (άθροισμα τετραγώνων του A παράγοντα, SS_A), ένα μέρος στην επίδραση του παράγοντα B (άθροισμα τετραγώνων του B παράγοντα, SS_B) και ένα μέρος είναι τυχαίας προέλευσης (άθροισμα τετραγώνων των υπολοίπων e_{ij} , SS_{res}).

Τα αθροίσματα τετραγώνων που εμφανίζονται στην (6.15) υπολογίζονται ευκολότερα από τους παρακάτω ισοδύναμους τύπους:

$$SS_{tot} = \sum_i \sum_j y_{ij}^2 - \frac{y_{..}^2}{1m}$$

$$SS_A = \sum_i \frac{y_{i.}^2}{m} - \frac{y_{..}^2}{1m}$$

$$SS_B = \sum_j \frac{y_{.j}^2}{1} - \frac{y_{..}^2}{1m}$$

$$\begin{aligned} SS_{res} &= \sum_i \sum_j y_{ij}^2 - \sum_i \frac{y_{i.}^2}{m} - \sum_j \frac{y_{.j}^2}{1} + \frac{y_{..}^2}{1m} \\ &= SS_{tot} - SS_A - SS_B. \end{aligned}$$



Στο SS_{tot} αντιστοιχούν $lm-1$ β.ε., στο SS_A , $l-1$ β.ε., στο SS_B , $m-1$ β.ε., και στο SS_{res} , $(l-1)(m-1)$ β.ε. Σημειώστε την ανάλογη προσθετικότητα με τα αθροίσματα τετραγώνων των αντίστοιχων βαθμών ελευθερίας.

Τα μέσα τετράγωνα δημιουργούνται αν διαιρέσουμε τα αθροίσματα τετραγώνων με τους αντίστοιχους β.ε. Έτσι έχουμε

$$MS_A = \frac{SS_A}{l-1}$$

$$MS_B = \frac{SS_B}{m-1}$$

και

$$MS_{res} = \frac{SS_{res}}{(l-1)(m-1)}.$$

Τα αθροίσματα τετραγώνων, οι β.ε. και τα μέσα τετράγωνα αποτελούν τον λεγόμενο πίνακα της Ανάλυσης της Διακύμανσης (ΑΝΑΔΙΑ).

ΑΝΑΔΙΑ

Προέλευση Μεταβλητότητας	Άθροισμα Τετραγώνων	Βαθμοί Ελευθερίας	Μέσο Τετράγωνο
Παράγοντας Α	$SS_A = m \sum (\bar{Y}_{i.} - \bar{Y}_{..})^2$	$l-1$	$MS_A = \frac{SS_A}{l-1}$
Παράγοντας Β	$SS_B = l \sum (\bar{Y}_{.j} - \bar{Y}_{..})^2$	$m-1$	$MS_B = \frac{SS_B}{m-1}$
Υπόλοιπο	$SS_{res} = \sum \sum (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$	$(l-1)(m-1)$	$MS_{res} = \frac{SS_{res}}{(l-1)(m-1)}$
Ολική Μεταβλητότητα	$SS_{tot} = \sum \sum (Y_{ij} - \bar{Y}_{..})^2$	$lm-1$	

Για τα δεδομένα του παραδείγματος της ΦΑΓΕ, μετά τους σχετικούς υπολογισμούς, προκύπτει ο παρακάτω πίνακας:



ΑΝΑΔΙΑ

Προέλευση Μεταβλητότητας	Άθροισμα Τετραγώνων	Βαθμοί Ελευθερίας	Μέσο Τετράγωνο
Παράγοντας Α (Τιμή πώλησης)	93.0	2	46.5
Παράγοντας Β (Τύπος διαφήμισης)	13.5	1	13.5
Υπόλοιπο	1.0	2	0.5
Ολική Μεταβλητότητα	107.5	5	

Μπορεί να αποδειχθεί ότι το MS_{res} είναι πάντα αμερόληπτος εκτιμητής του σ^2 . Είναι επίσης

$$E(MS_A) = \sigma^2 + \frac{m}{1-1} \sum \alpha_i^2$$

και

$$E(MS_B) = \sigma^2 + \frac{1}{m-1} \sum \beta_j^2.$$

Οι παραπάνω αναμενόμενες τιμές δείχνουν ότι όταν δεν υπάρχουν επιδράσεις του παράγοντα Α, όταν δηλαδή $\alpha_i = 0$, τα MS_A και MS_{res} είναι περίπου ίσα· διαφορετικά το MS_A τείνει να γίνει μεγαλύτερο. Όμοια, όταν δεν υπάρχουν επιδράσεις του παράγοντα Β, όταν δηλαδή $\beta_j = 0$, τα MS_B και MS_{res} είναι περίπου ίσα· διαφορετικά το MS_B τείνει να γίνει μεγαλύτερο. Έτσι, οι ποσότητες MS_A/MS_{res} και MS_B/MS_{res} μπορούν να δώσουν πληροφορίες για το κατά πόσον υπάρχουν επιδράσεις των παραγόντων Α και Β.

Οι υποθέσεις που αρχικά μας ενδιαφέρει να ελέγξουμε είναι κατά πόσον οι επιδράσεις των επιπέδων των παραγόντων Α ή/και Β είναι ίδιες. Έτσι, για τον παράγοντα Α βασικά μας ενδιαφέρει να ελέγξουμε την

$$H_0^A: \alpha_1 = \alpha_2 = \dots = \alpha_l = 0 \text{ έναντι της } H_a^A: \text{όχι όλα τα } \alpha_i \text{ ίσα με μηδέν.}$$

Για τον παράγοντα Β μας ενδιαφέρει να ελέγξουμε την



$H_0^B: \beta_1 = \beta_2 = \dots = \beta_m = 0$ έναντι της H_a^B : όχι όλα τα β_j ίσα με μηδέν.

Μπορεί να αποδειχθεί (βλέπε π.χ. Scheffé, 1959) ότι τα στατιστικά SS_A , SS_B , SS_{res} είναι ανεξάρτητα και επί πλέον

$$\frac{SS_{res}}{\sigma^2} \sim \chi^2_{(1-1)(m-1)},$$

$$\frac{SS_A}{\sigma^2} \sim \chi^2_{1-1} \text{ όταν } H_0^A \text{ αληθής}$$

και

$$\frac{SS_B}{\sigma^2} \sim \chi^2_{m-1} \text{ όταν } H_0^B \text{ αληθής.}$$

Έτσι, όταν H_0^A αληθής έχουμε

$$F^A = \frac{MS_A}{MS_{res}} \sim F_{1-1, (1-1)(m-1)}.$$

Ο κανόνας δηλαδή είναι να απορρίπτουμε, σε επίπεδο σημαντικότητας α , τη H_0^A με εναλλακτική H_a^A όταν

$$F^A > F_{\alpha, 1-1, (1-1)(m-1)}.$$

Επίσης, όταν H_0^B αληθής, έχουμε

$$F^B = \frac{MS_B}{MS_{res}} \sim F_{m-1, (1-1)(m-1)}.$$

Έτσι η H_0^B με εναλλακτική H_a^B απορρίπτεται, σε επίπεδο σημαντικότητας α , όταν



$$F^B > F_{\alpha, m-1, (1-1)(m-1)} .$$

Για τα δεδομένα της ΦΑΓΕ είναι

$$F_A = \frac{46.5}{0.5} = 93 \quad \text{και} \quad F_{0.05, 2, 2} = 19 .$$

Επειδή $93 > 19$, συμπεραίνουμε, σε επίπεδο σημαντικότητας $\alpha=0.05$, ότι υπάρχουν επιδράσεις του παράγοντα "τιμή πώλησης". Επίσης, $F^B = 13.5/0.5 = 27$ και $F_{0.05, 1, 2} = 18.5$. Επειδή $27 > 18.5$ συμπεραίνουμε ότι υπάρχουν επιδράσεις του παράγοντα "διαφήμιση" στις πωλήσεις του προϊόντος.

Σημειώστε ότι για το παράδειγμα του εδαφίου 6.5 αν συμβολίσουμε με Α τον παράγοντα "πορείες" και με Β τον παράγοντα "ημέρες" οι σχετικοί υπολογισμοί δίνουν $F_A = 7.75 > F_{0.05, 3, 12} = 3.49$ και $F_B = 8.06 > F_{0.05, 4, 12} = 3.26$. Έτσι, η H_0 ότι δεν υπάρχουν διαφορές στους μέσους χρόνους των τεσσάρων δρόμων απορρίπτεται (σε αντίθεση με το συμπέρασμα που μας είχε οδηγήσει η Ανάλυση της Διακύμανσης κατά ένα παράγοντα).

6.6.4. Ανάλυση των μέσων τιμών

Όταν διαπιστώνεται ότι υπάρχουν σημαντικές διαφορές μεταξύ των επιπέδων του παράγοντα Α ή/και μεταξύ των επιπέδων του παράγοντα Β, μπορούμε να προχωρήσουμε σε επιμέρους ελέγχους με σκοπό τον προσδιορισμό των επιπέδων που προκαλούν τις διαφορετικές επιδράσεις.

Οι συγκρίσεις των επιπέδων ανά ζεύγη γίνονται χρησιμοποιώντας τις μεθόδους που αναφέρθηκαν με λεπτομέρεια στην § 6.4.4.

Ενδεικτικά αναφέρουμε ότι, αν χρησιμοποιήσουμε την μέθοδο της ελάχιστης σημαντικής διαφοράς, ένα $(1-\alpha) \cdot 100\%$ διάστημα εμπιστοσύνης για την διαφορά $\mu_u - \mu_v$ των επιπέδων u και v του παράγοντα Α είναι το



$$\bar{Y}_{u.} - \bar{Y}_{v.} \pm t_{\alpha/2, (1-1)(m-1)} \sqrt{\frac{2MS_{res}}{m}}.$$

Επίσης ένα $(1-\alpha)$ 100% διάστημα εμπιστοσύνης για την διαφορά $\beta_U - \beta_V$ των επιπέδων u και v του παράγοντα B είναι το

$$\bar{Y}_{.u} - \bar{Y}_{.v} \pm t_{\alpha/2, (1-1)(m-1)} \sqrt{\frac{2MS_{res}}{1}}.$$

Ένα 95% διάστημα εμπιστοσύνης για τη διαφορά των μέσων εισπράξεων που αντιστοιχούν στις τιμές 55 και 45 για τα δεδομένα της ΦΑΓΕ είναι το $12.0-21.0 \pm 4.303 \sqrt{\frac{2 \times (0.5)}{2}}$ ή το $[-12.04, -5.96]$.

Αν χρησιμοποιήσουμε την μέθοδο του εύρους τα $(1-\alpha)$ 100% διαστήματα εμπιστοσύνης για τις διαφορές $\alpha_U - \alpha_V$ και $\beta_U - \beta_V$ είναι αντίστοιχα

$$\bar{Y}_{u.} - \bar{Y}_{v.} \pm Q_{\alpha} \sqrt{\frac{MS_{res}}{m}} \text{ και } \bar{Y}_{.u} - \bar{Y}_{.v} \pm Q_{\alpha} \sqrt{\frac{MS_{res}}{1}}.$$

6.6.5 Έλεγχος για την καταλληλότητα του μοντέλου

Δεν παρουσιάζονται καινούργια προβλήματα εξετάζοντας την καταλληλότητα του μοντέλου της Ανάλυσης της Διακύμανσης κατά δύο παράγοντες. Τα υπόλοιπα

$$e_{ij} = Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$$

μπορούν να εξεταστούν για τον εντοπισμό αποκλίσεων από τις υποθέσεις της κανονικότητας, σταθερότητας της διακύμανσης και ανεξαρτησίας των τυχαίων σφαλμάτων κατά τρόπο ίδιο με αυτόν που περιγράφεται στην § 6.4.5 για την περίπτωση του μοντέλου της Ανάλυσης της Διακύμανσης κατά ένα παράγοντα.



6.6.6 Στατιστικά πακέττα

Έτοιμα προγράμματα ηλεκτρονικού υπολογιστή κατάλληλα να επεξεργαστούν αριθμητικά τα δεδομένα σύμφωνα με την μεθοδολογία της Ανάλυσης της Διακύμανσης κατά δύο παράγοντες διατίθενται στο εμπόριο και περιέχονται, μεταξύ άλλων, στα στατιστικά πακέττα BMDP και SPSS.

Στις προηγούμενες παραγράφους οι διάφοροι υπολογισμοί που υπεισέρχονται στην μεθοδολογία της Ανάλυσης της Διακύμανσης κατά δύο παράγοντες χρησιμοποιήθηκαν για την στατιστική ανάλυση των δεδομένων της ΦΑΓΕ. Εδώ θα δώσουμε τα αποτελέσματα για τα δεδομένα της ΦΑΓΕ όπως πάρθηκαν από τον ηλεκτρονικό υπολογιστή του Πανεπιστημίου Ιωαννίνων χρησιμοποιώντας το πρόγραμμα BMDP2V που αντιστοιχεί στην μεθοδολογία της Ανάλυσης της Διακύμανσης κατά δύο παράγοντες.

Τα αριθμητικά αποτελέσματα που ακολουθούν έχουν παραχθεί με χρήση ηλεκτρονικού υπολογιστή και περιέχουν:

- (i) Τα δεδομένα εισπράξεων ανά κατάσταση, Y_{ij} (§ 6.6.2).
- (ii) Τον Πίνακα ΑΝΑΔΙΑ και τα δύο F-τεστ για τον έλεγχο της σημαντικότητας των υποθέσεων ότι οι επιδράσεις των επιπέδων των παραγόντων Α ή/και Β είναι ίδιες (§ 6.6.3).

ΠΑΡΑΔΕΙΓΜΑ ΦΑΓΕ

(i)

CELL MEANS FOR 1-ST DEPENDENT VARIABLE

PRICE	- P1	P1	P2	P2	P3	P3
ADVERTIC	- A1	A2	A1	A2	A1	A2
ENCASH	14.00000	10.00000	21.00000	18.00000	22.00000	20.00000
COUNT	1	1	1	1	1	1

(ii)

ANALYSIS OF VARIANCE FOR 1-ST DEPENDENT VARIABLE - ENCASH

SOURCE	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F
MEAN	1837.50000	1	1837.50000	3675.00
PRICE	93.00000	2	46.50000	93.00
ADVERTIC	13.50000	1	13.50000	27.00
ERROR	1.00000	2	.50000	



ΑΣΚΗΣΕΙΣ

- 6.1 Για το μοντέλο της Ανάλυσης της Διακύμανσης κατά ένα παράγοντα ναδειχτεί ότι

$$E(MS_{tre}) = \sigma^2 + \frac{\sum n_j(\mu_j - \mu)^2}{r-1}.$$

- 6.2 Για το μοντέλο της Ανάλυσης της Διακύμανσης κατά δύο παράγοντες ναδειχτεί ότι

$$E(MS_A) = \sigma^2 + \frac{m}{l-1} \sum \alpha_i^2$$

και

$$E(MS_B) = \sigma^2 + \frac{1}{m-1} \sum \beta_j^2.$$

- 6.3 Μια ομάδα από 28 υπέρβαρα κορίτσια αποφάσισαν λίγο πριν τις καλοκαιρινές διακοπές να χάσουν βάρος ακολουθώντας τέσσερις διαφορετικές μεθόδους διαίτας $\Delta_1, \Delta_2, \Delta_3, \Delta_4$. Χωρίστηκαν σε 4 ίσες ομάδες και κάθε ομάδα ακολούθησε μία διαίτα. Λόγω αποχωρήσεων (αρρώστιες, όψιμη προτίμηση βουνού αντί θάλασσας κλπ.) ο αριθμός των κοριτσιών που τελικά ολοκλήρωσαν τη διαίτα διέφερε από ομάδα σε ομάδα. Τα δεδομένα που ακολουθούν αποτελούν τα κιλά που έχασαν τα κορίτσια μετά από δύο μήνες διαίτα.

Δ_1	7	9	7	8	8	9
Δ_2	7	7	8	8	7	7
Δ_3	6	8	5	6	5	6
Δ_4	8	9	10	9		

Υπάρχει διαφορά ως προς την αποτελεσματικότητα των



τεσσάρων μεθόδων δίατας ($\alpha=0.05$) και αν ναι ποια μέθοδος είναι η καλύτερη;

- 6.4 Ένας ψυχολόγος επιθυμώντας να συγκρίνει τρεις μεθόδους (M_1, M_2, M_3) ελάττωσης της επιθετικότητας χρησιμοποίησε 11 φοιτητές και μετά με κάποιο τεστ μέτρησε το βαθμό επιθετικότητας. Η M_1 εφαρμόστηκε σε 4 φοιτητές που επιλέχτηκαν τυχαία, η M_2 σε 4 φοιτητές που επίσης επιλέχτηκαν τυχαία και η M_3 στους υπόλοιπους 3 φοιτητές. Μετά δύο μήνες από την εφαρμογή των μεθόδων τα αποτελέσματα ήταν:

M_1	80	92	87	83
M_2	70	81	78	74
M_3	63	76	70	

Υπάρχει σημαντική διαφορά στις επιδράσεις των τριών μεθόδων;

- 6.5 Τα παρακάτω δεδομένα αποτελούν τους αριθμούς λαθών που έκαναν σε πέντε διαδοχικές βδομάδες τέσσερις τεχνικοί (T_1, T_2, T_3, T_4) που εργάζονται σε κάποιο εργαστήριο Βιολογίας.

T_1	13	16	12	14	15
T_2	14	16	11	19	15
T_3	13	18	16	14	18
T_4	18	10	14	15	12

Υπάρχει σημαντική διαφορά στην απόδοση των τεσσάρων τεχνικών;



- 6.6 Για να μελετηθεί η επίδραση της απώλειας ύπνου στην ικανότητα οδήγησης, μετρήθηκαν οι παραβάσεις στην οδήγηση διάρκειας μίας ώρας σε πόλη, οδηγών αυτοκινήτων, που δεν είχαν κοιμηθεί για ορισμένες ώρες. Να γίνει στατιστική ανάλυση των δεδομένων που δίνονται στον πίνακα που ακολουθεί.

Ώρες χωρίς ύπνο			
4	12	20	28
11	13	14	26
7	15	26	22
6	16	21	17
8	9	15	19

- 6.7 Ένα πείραμα γίνεται για να διαπιστωθεί η επίδραση στο γυναικείο δέρμα τεσσάρων απορρυπαντικών (A_1, A_2, A_3, A_4). Τρία τετραγωνικά εκατοστά δέρματος εκτίθενται στη χρήση απορρυπαντικού και βαθμολογείται η επίδραση σε κλίμακα 1-3 ανάλογα με τον βαθμό ερεθισμού που προκαλείται. Τρεις γυναίκες εθελόντριες ($\Gamma_1, \Gamma_2, \Gamma_3$) χρησιμοποιούνται και εφαρμόζονται, σε τυχαία σειρά, και τα τέσσερα απορρυπαντικά σε κάθε μία από αυτές. Τα αποτελέσματα δίνονται στον πίνακα που ακολουθεί.

	A_1	A_2	A_3	A_4
Γ_1	2	3	3	2
Γ_2	3	2	1	3
Γ_3	1	2	2	1

Αν γνωρίζουμε ότι οι τρεις γυναίκες έχουν αντίστοιχα ξηρό, κανονικό και λιπαρό τύπο δέρματος, υπάρχουν ενδείξεις για διαφορετική μέση επίδραση των τεσσάρων απορρυπαντικών;



- 6.8 Σε γεωργικό πειραματικό σταθμό καλλιεργήθηκαν χρησιμοποιώντας το πλήρες τυχαιοποιημένο σχέδιο τετραγώνων τέσσερις ποικιλίες σιταριού ($\Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4$) και χρησιμοποιήθηκαν τρία είδη λιπάσματος ($\Lambda_1, \Lambda_2, \Lambda_3$). Τα αποτελέσματα της παραγωγής (σε Kgr) δίνονται στον πίνακα που ακολουθεί.

	Λ_1	Λ_2	Λ_3
Σ_1	35	41	27
Σ_2	39	47	36
Σ_3	42	43	31
Σ_4	51	49	30

Να ελεγχθεί κατά πόσον υπάρχει διαφορά στην παραγωγή που να οφείλεται (i) στα τρία λιπάσματα, (ii) στις τέσσερις ποικιλίες.

- 6.9 Μία αντιπροσωπία διαθέτει για πώληση τέσσερα αυτοκίνητα (A_1, A_2, A_3, A_4). Ένας πελάτης θέλοντας να τα συγκρίνει ως προς την κατανάλωση βενζίνης χρησιμοποιεί τρεις ποιότητες βενζίνης (B_1, B_2, B_3) και κάνει το ακόλουθο πείραμα: Στο άδειο τεπόζιτο του κάθε αυτοκινήτου προστίθεται διαδοχικά και σε τυχαία σειρά ένα λίτρο από τους τρεις τύπους βενζίνης και οδηγείται το αυτοκίνητο μέχρι να σταματήσει. Ο πίνακας που ακολουθεί δείχνει τις αποστάσεις (σε Km) που διανύθηκαν σε κάθε μία από τις εννιά δοκιμαστικές οδηγήσεις.

	A_1	A_2	A_3	A_4
B_1	22	23	30	21
B_2	14	15	22	15
B_3	20	18	27	22

Υπάρχει διαφορά (i) μεταξύ των αυτοκινήτων και (ii)

μεταξύ των ποιοτήτων βενζίνης ως προς τα χιλιόμετρα που διανύθηκαν;

- 6.10 Τρία είδη αντισκωριακού (A_1, A_2, A_3) χρησιμοποιήθηκαν σε τέσσερα είδη σιδήρου ($\Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4$) σύμφωνα με το πλήρες τυχαιοποιημένο σχέδιο τετραγώνων. Τα αποτελέσματα που ακολουθούν αποτελούν τον χρόνο αντίδρασης σε δευτερόλεπτα.

Σ_1	Σ_2	Σ_3	Σ_4
A_1	A_3	A_1	A_2
1.7	2.1	2.1	2.2
A_3	A_1	A_2	A_1
2.3	1.5	2.3	0.6
A_2	A_2	A_3	A_3
3.4	2.6	0.8	1.6

Υπάρχουν ενδείξεις για διαφορά (i) στην ποιότητα των τριών αντισκωριακών και (ii) στα τέσσερα είδη σιδήρου;



ΚΕΦΑΛΑΙΟ 7

μη παραμετρική στατιστική

7.1 Γενικότητες

Τα στατιστικά τεστ που μελετήσαμε στα προηγούμενα κεφάλαια στηρίζονται σε διάφορες υποθέσεις σχετικά με την κατανομή του πληθυσμού από τους οποίους προέρχονται τα δεδομένα. Στις περισσότερες από τις περιπτώσεις υποθέσαμε ότι ο πληθυσμός είναι κανονικός. Άλλες φορές υποθέσαμε ότι οι τυπκές αποκλίσεις είχαν κάποια γνωστή τιμή ή ότι ήσαν ίσες μεταξύ τους. Για δύο ή περισσότερα δείγματα πολλές φορές ήταν απαραίτητη η υπόθεση της ανεξαρτησίας. Επειδή υπάρχουν περιπτώσεις που τέτοιες υποθέσεις δεν ικανοποιούνται, έχουν αναπτυχθεί διαφορετικές τεχνικές που δεν στηρίζονται σε υποθέσεις συγκεκριμένης μορφής γύρω από την μορφή του πληθυσμού ή των παραμέτρων αυτού. Οι τεχνικές αυτές είναι γνωστές σαν μη παραμετρικές (non parametric) μέθοδοι ή μέθοδοι ελεύθερες κατανομών (distribution free). Επί πλέον, πολλές από τις μη παραμετρικές μεθόδους είναι σε θέση να αναλύσουν δεδομένα όχι μόνο ποσοτικών μεταβλητών αλλά και μεταβλητών των οποίων οι τιμές έχουν απλά την δυνατότητα της διάταξης. Η ανάγκη για μη παραμετρικές μεθόδους ξεκίνησε από την ψυχολογία όπου οι περισσότεροι από τους πληθυσμούς είναι μη κανονικοί, ενώ ταυτόχρονα δεν υπάρχουν ή είναι ασύμφορη η συλλογή μεγάλου αριθμού δεδομένων που θα καθιστούσε δυνατή τη χρήση του κεντρικού οριακού θεωρήματος. Εκτός από το γεγονός ότι οι μη παραμετρικές μέθοδοι μπορούν να χρησιμοποιηθούν κάτω από πολύ γενικότερες συνθήκες σε σχέση με τις αντίστοιχες κλασσικές (παραμετρικές) μεθόδους που μελετήσαμε στα προηγούμενα εδάφια



είναι αρκετά "δημοφιλείς" γιατί είναι απλές και εύχρηστες. Το κύριο μειονέκτημα τους έναντι των κλασσικών μεθόδων είναι ότι συνήθως υπάρχει απώλεια πληροφορίας στην επεξεργασία των δεδομένων και έτσι γίνονται λιγώτερο αποτελεσματικές (αξιόπιστες). Η σύγκριση σε αποτελεσματικότητα μειονεκτεί βέβαια μόνο όταν οι συνθήκες που απαιτούνται για τις κλασσικές μεθόδους ικανοποιούνται. Έτσι δεν είναι σωστό να υποτιμηθεί η αξία των μη παραμετρικών μεθόδων. Από διαφορετική σκοπιά, όσο λιγώτερες υποθέσεις γίνονται τόσο λιγώτερα είναι τα συμπεράσματα που μπορούν να βγουν από τα δεδομένα που αναλύονται, αλλά τόσο γενικώτερα είναι τα συμπεράσματα αυτά.

Η αναγκαιότητα των μη παραμετρικών μεθόδων οδήγησε στη δημιουργία ενός ξεχωριστού κλάδου της Στατιστικής γνωστού ως Μη Παραμετρική Στατιστική. Εδώ θα παρουσιάσουμε με απλό μεθοδολογικό τρόπο μερικές από τις κυριώτερες μη παραμετρικές μεθόδους. Λεπτομέρειες γι'αυτές τις μεθόδους όπως και για άλλες λιγώτερο γνωστές που ξεφεύγουν από τον σκοπό ενός εισαγωγικού βιβλίου δίνονται, μεταξύ άλλων, στα βιβλία των Conover (1971), Gibbons (1971), Kendall (1970) και Siegel (1956).

7.2 Κλίμακες μετρήσεων

Είναι γνωστή η διάκριση των μεταβλητών σε διακριτές και συνεχείς, ποιοτικές και ποσοτικές. Υπάρχουν όμως και επιπρόσθετες διακρίσεις που διευκολύνουν την μελέτη των ερευνητικών θεμάτων και βοηθούν στην επιλογή της στατιστικής μεθοδολογίας.

Όταν οι τιμές της μεταβλητής αναφέρονται σε ονόματα, κατηγορίες, είδη κ.λ.π. και χρησιμοποιούνται σύμβολα ή αριθμοί για να την περιγράψουν, τότε λέμε ότι η μεταβλητή είναι ονοματική και χρησιμοποιούμε μία ονοματική (nominal) κλίμακα μετρήσεων. Για παράδειγμα, όταν ο ψυχίατρος χαρακτηρίζει κάποιον ως σχιζοφρενή, παρανοϊκό, καταθλιπτικό ή ψυχονευροτικό χρησιμοποιεί κάποιο για να αναπαραστήσει την ομάδα που ανήκει, χρησιμοποιεί

δηλαδή μία ονοματική κλίμακα. Άλλο παράδειγμα δεδομένων ονοματικής κλίμακας αποτελούν τα σύμβολα (γράμματα) στις πινακίδες των αυτοκινήτων που προσδιορίζουν την περιοχή που εκδόθηκε η άδεια κυκλοφορίας τους.

Όταν κάποιου είδους σχέση (καλύτερο, περισσότερο, προτιμητέο) εφαρμοσμένη σε δεδομένα της ονοματικής κλίμακας έχει νόημα, όταν δηλαδή οι ομάδες συνδέονται με κάποια σχέση συγκριτικής μορφής, τότε είναι δυνατές μετρήσεις στην διατάξιμη (ordinal, ranking) κλίμακα. Για παράδειγμα, μετρήσεις που σχετίζονται με τους βαθμούς των αξιωματικών του στρατού όπου λοχαγός < ταγματάρχης < ... < στρατηγού, είναι δεδομένα διατάξιμης κλίμακας.

Μετρήσεις διατάξιμης κλίμακας που επιπλέον, έχουν την ιδιότητα η απόσταση οποιουδήποτε ζεύγους να είναι γνωστού μεγέθους, αποτελούν δεδομένα της λεγόμενης συνεχούς (interval) κλίμακας. Η συνεχής κλίμακα χαρακτηρίζεται από μία σταθερή μονάδα μέτρησης που προσδίδει ένα πραγματικό αριθμό στη διαφορά οποιουδήποτε ζεύγους από τα διατεταγμένα αντικείμενα. Το σημείο μηδέν όπως και η μονάδα μέτρησης είναι αυθαίρετα έτσι, ο λόγος δύο διαστημάτων της συνεχούς κλίμακας είναι ανεξάρτητος της μονάδας μέτρησης και του σημείου μηδέν. Για παράδειγμα την θερμοκρασία την μετράμε σε συνεχή κλίμακα. Στην πραγματικότητα υπάρχουν δύο κλίμακες - Κελσίου και Φερενάϊτ - που συνήθως χρησιμοποιούνται, με διαφορετική μονάδα μέτρησης και σημείο μηδέν η κάθε μία. Όμως και οι δύο κλίμακες παρέχουν την ίδια πληροφορία αφού μία μέτρηση της κλίμακας Κελσίου (Κ) μπορεί να μετατραπεί ισοδύναμα σε μέτρηση της κλίμακας Φερενάϊτ (Φ) αν χρησιμοποιηθεί η σχέση $\Phi = 1.8K + 32$. Είναι εύκολο ναδειχθεί ότι ο λόγος των διαφορών μεταξύ μετρήσεων θερμοκρασίας είναι ανεξάρτητος από την μονάδα μέτρησης και το σημείο μηδέν. Ο λόγος των διαφορών μεταξύ των μετρήσεων 30 και 10, και 10 και 0 της κλίμακας Κελσίου είναι $(30-10)/(10-0)=2$. Ο λόγος των διαφορών μεταξύ των ισοδύναμων μετρήσεων της κλίμακας Φερενάϊτ είναι $(86-50)/(50-32)=2$. Με άλλα λόγια σε μία συνεχή κλίμακα, ο λόγος οποιονδήποτε δύο διαστημάτων είναι ανεξάρτητος της



μονάδας μέτρησης που χρησιμοποιήθηκε και του σημείου μηδέν τα οποία είναι αυθαίρετα.

Όταν μία κλίμακα μετρήσεων έχει όλα τα χαρακτηριστικά της συνεχούς και επιπλέον το πραγματικό σημείο μηδέν στην αρχή ονομάζεται κλίμακα λόγου (ratio). Ο λόγος δύο οιαδήποτε σημείων της κλίμακας είναι ανεξάρτητος της μονάδας μέτρησης η οποία είναι αυθαίρετη. Για παράδειγμα κλίμακες λόγου χρησιμοποιούμε για να μετρήσουμε τον όγκο ή το βάρος.

Μετρήσεις στην ονομαστική ή στην διατάξιμη κλίμακα εμφανίζονται συχνά στις επιστήμες συμπεριφοράς και στις κοινωνικές επιστήμες. Δεδομένα που αποτελούνται από μετρήσεις αυτού του είδους αναλύονται με την βοήθεια μη παραμετρικών μεθόδων. Δεδομένα που αποτελούνται από μετρήσεις στην συνεχή κλίμακα ή στην κλίμακα λόγου μπορούν να αναλυθούν και με την βοήθεια των παραμετρικών (κλασσικών) μεθόδων όταν δεν διαπιστώνονται αποκλίσεις από τις υποθέσεις που γίνονται στο παραμετρικό στατιστικό μοντέλο.

7.3 Ένας πληθυσμός - Ένα δείγμα

Συχνά επιθυμούμε να ελέγξουμε κατά πόσο ένα τυχαίο δείγμα προέρχεται από πληθυσμό με κάποια γνωστή κατανομή. Τα σημαντικότερα στατιστικά τεστ για ελέγχους τέτοιων υποθέσεων είναι το χ^2 τεστ καλής προσαρμογής και το τεστ των Kolmogorov-Smirnov που βασίζουν τον έλεγχο σε κάποια σύγκριση (προσαρμογή) μεταξύ της κατανομής συχνοτήτων των δεδομένων και της υποθετικής κατανομής. Αν η κατανομή του πληθυσμού είναι κανονική τότε το t-τεστ (βλ. Κεφ. 4) μπορεί να χρησιμοποιηθεί για να ελέγξει υποθέσεις γύρω από τη μέση τιμή του πληθυσμού. Αντίστοιχες υποθέσεις, όταν υπάρχουν ενδείξεις για μεγάλες αποκλίσεις από την κανονικότητα της κατανομής, ελέγχονται με το προσημικό τεστ που δεν απαιτεί γνώση της κατανομής του πληθυσμού.



7.3.1 Το χ^2 τεστ καλής προσαρμογής

Έστω ότι οι παρατηρήσεις $x_1, x_2, \dots, x_n$ αποτελούν τυχαίο δείγμα μεγέθους n από κάποιο πληθυσμό με α.σ.κ. $F(x)$ και $F_0(x)$ κάποια γνωστή κατανομή. Μας ενδιαφέρει να ελέγξουμε την μηδενική υπόθεση $H_0: F(x) = F_0(x)$ με εναλλακτική $H_a: F(x) \neq F_0(x)$. Διαιρούμε τον δειγματικό χώρο σε k κατηγορίες (τμήματα) και υπολογίζουμε την πιθανότητα κάθε μίας από τις κατηγορίες με τη βοήθεια της κατανομής που δηλώνει η μηδενική υπόθεση. Έτσι έχουμε $p_i = P(x \in i\text{-οστή κατηγορία} / F_0(x))$, $0 < p_i < 1$, $\sum p_i = 1$. Αν η H_0 είναι σωστή τότε ο αριθμός n_i των παρατηρήσεων που ανήκουν στην i -οστή κατηγορία, $i=1, \dots, k$, δεν θα πρέπει να διαφέρει πολύ από τον αντίστοιχο αναμενόμενο (θεωρητικό) αριθμό $e_i = np_i$ που βασίζεται στην $F_0(x)$.

Ένα κριτήριο που μπορεί να μετρήσει την συνολική διαφορά μεταξύ παρατηρούμενων και αναμενόμενων τιμών οφείλεται στον Karl Pearson και δίνεται από τη σχέση

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i} \quad (7.1)$$

για το οποίο μπορεί να αποδειχθεί (Pearson, 1900) ότι κάτω από ορισμένες προϋποθέσεις έχει ασυμπτωματικά χ^2 κατανομή με $k-1$ βαθμούς ελευθερίας, αφού $\sum n_i = \sum e_i = n$, όταν ισχύει η H_0 .

Επειδή μεγάλες τιμές του χ^2 σημαίνουν μεγάλες διαφορές μεταξύ παρατηρούμενων και αναμενόμενων τιμών απορρίπτουμε την υπόθεση ότι τα δεδομένα προέρχονται από την $F_0(x)$ σε επίπεδο σημαντικότητας α , όταν $\chi^2 \geq \chi^2_{\alpha, k-1}$. Αν s παράμετροι της $F_0(x)$ είναι άγνωστες, τότε θα πρέπει να εκτιμηθούν από τα δεδομένα με αποτέλεσμα την μείωση κατά s των βαθμών ελευθερίας της χ^2 κατανομής. Η προσέγγιση της χ^2 κατανομής είναι καλύτερη όταν τα p_i είναι σχεδόν ίσα, η επιλογή των κατηγοριών γίνεται ανεξάρτητα από τη μορφή των δεδομένων και τα $e_i \geq 5$ ή τουλάχιστον το 80% των $e_i \geq 5$ και κανένα από



τα $e_i < 1$. Το κριτήριο χ^2 που είναι γνωστότερο σαν χι-τετράγωνο τεστ καλής προσαρμογής είναι από τα πλέον δημοφιλή στατιστικά τεστ και μπορεί να εφαρμοστεί σε ανάλυση ονοματικών δεδομένων. Η ευκολία στη χρησιμοποίηση του φαίνεται στο παράδειγμα που ακολουθεί.

Για να διαπιστωθεί κατά πόσο ένα ζάρι είναι αμερόληπτο (ισοβαρές), το συγκεκριμένο ζάρι ρίχτηκε 90 φορές με τα παρακάτω αποτελέσματα:

πλευρά	1	2	3	4	5	6
συχνότητα εμφάνισης	13	17	16	18	12	14

Μπορούμε να ισχυριστούμε, σε επίπεδο σημαντικότητας $\alpha=0.05$, ότι το ζάρι είναι ισοβαρές; Η μηδενική υπόθεση ότι το ζάρι είναι ισοβαρές υποδηλώνει ότι οι πιθανότητες εμφανίσεων των πλευρών είναι ίσες, έτσι για κάθε μία από τις $k=6$ φυσικές κατηγορίες έχουμε $p_i=1/6$, δηλαδή $e_i=90/6=15$, $i=1, \dots, 6$. Το χ^2 κριτήριο της (7.1) δίνει $\chi^2 = (13-15)^2/15 + \dots + (14-15)^2/15 = 1.87$. Επειδή $\chi^2 \leq \chi^2_{0.05,5} = 11.1$ η μηδενική υπόθεση ότι το ζάρι είναι ισοβαρές δεν μπορεί να απορριφθεί.

7.3.2 Το τεστ των Kolmogorov - Smirnov

Έστω $x_1, x_2, \dots, x_n$ τυχαίο δείγμα μεγέθους n από κάποιο πληθυσμό με α.σ.κ. $F(x)$. Μας ενδιαφέρει να ελέγξουμε την μηδενική υπόθεση ότι το δείγμα προέρχεται από κάποια γνωστή κατανομή $F_0(x)$.

Το στατιστικό

$$D_n = \max_x |F_n(x) - F_0(x)| \quad (7.2)$$

όπου η

$$F_n(x) = (\text{αριθμός παρατηρήσεων του δείγματος με τιμές } \leq x)/n$$



αποτελεί την εμπειρική ή παρατηρηθείσα συνάρτηση κατανομής, μπορεί να χρησιμοποιηθεί για να αποδώσει τον βαθμό συμφωνίας μεταξύ της εμπειρικής κατανομής των δεδομένων και της θεωρητικής κατανομής $F_0(x)$ που προσδιορίζει η H_0 . Το κριτήριο D_n οδηγεί σε ένα τεστ καλής προσαρμογής και οφείλεται στους Kolmogorov (1941) και Smirnov (1948).

Είναι φανερό ότι μικρές τιμές του D_n προκύπτουν όταν δεν υπάρχουν σημαντικές διαφορές μεταξύ των κατανομών $F_n(x)$ και $F_0(x)$ έτσι, σε επίπεδο σημαντικότητας α , η H_0 απορρίπτεται όταν $D_n \geq D_{n,\alpha}$ όπου $P(D_n \geq D_{n,\alpha}) = \alpha$. Ο πίνακας VI δίνει τις τιμές $D_{n,\alpha}$ για διάφορα α . Για τον υπολογισμό του D_n αρκεί κανείς να περιοριστεί στα x_i του δείγματος. Το τεστ των Kolmogorov - Smirnov μπορεί να εφαρμοστεί σε ανάλυση δεδομένων της διατάξης κλίμακας και είναι ακριβές και αποτελεσματικότερο του χ^2 όταν η $F_0(x)$ είναι συνεχής.

Στο τεστ των Kolmogorov - Smirnov δεν υπάρχει απώλεια πληροφορίας αφού κάθε παρατήρηση χρησιμοποιείται ξεχωριστά σε αντίθεση με το χ^2 τεστ που βασίζεται στις συχνότητες των ομαδοποιημένων παρατηρήσεων. Ο τρόπος που εφαρμόζεται το τεστ των Kolmogorov - Smirnov δίνεται στο παρακάτω παράδειγμα.

Έστω ότι μας ενδιαφέρει να ελέγξουμε κατά πόσο οι μετρήσεις 0.79, 0.15, 0.43, 0.01, 0.93, 0.15, 0.27, 0.19, 0.32, 0.96 προέρχονται από την ομοιόμορφη κατανομή $U(0,1)$.

Η υπόθεση που θέλουμε να ελέγξουμε είναι η

$$H_0: F(x) = F_0(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x < 1 \\ 0, & x \geq 1 \end{cases} \quad \text{έναντι της } H_a: F(x) \neq F_0(x).$$

Από τα δεδομένα παίρνουμε

$F_0(x)$	0.01	0.15 (δύο μετρ)	0.19	0.27	0.32	0.43	0.79	0.93	0.96
$F_n(x)$	1/10	3/10	4/10	5/10	6/10	7/10	8/10	9/10	10/10
$F_n(x) - F_0(x)$	0.09	0.15	0.21	0.23	0.28	0.27	0.01	0.03	0.04



Έτσι, $D_{10} = \max_x |F_n(x) - F_0(x)| = 0.28$ και $D_{n,0.05} = 0.322$. Επειδή το 0.28 δεν ανήκει στην κρίσιμη περιοχή μεγέθους $\alpha=0.05$, δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση ότι τα δεδομένα προέρχονται από την $U(0,1)$ κατανομή.

7.3.3 Το προσημικό τεστ

Έστω $x_1, x_2, \dots, x_n$ τυχαίο δείγμα μεγέθους n από κάποιο πληθυσμό και έστω ότι μας ενδιαφέρει να ελέγξουμε διάφορες υποθέσεις γύρω από τη μέση τιμή του πληθυσμού.

Αν το δείγμα προέρχεται από κανονικό πληθυσμό τότε για ελέγχους αυτής της μορφής μπορούμε να χρησιμοποιήσουμε το κλασσικό t -τεστ. Διαφορετικά, αν ο πληθυσμός είναι συνεχής και συμμετρικός έτσι ώστε η πιθανότητα να έχουμε στο δείγμα τιμές μικρότερες της μέσης τιμής μ και η πιθανότητα να έχουμε τιμές μεγαλύτερες του μ να είναι η κάθε μία ίση με 0.5, μπορούμε να χρησιμοποιήσουμε το λεγόμενο προσημικό (sign) κριτήριο.

Για να ελέγξουμε την μηδενική υπόθεση

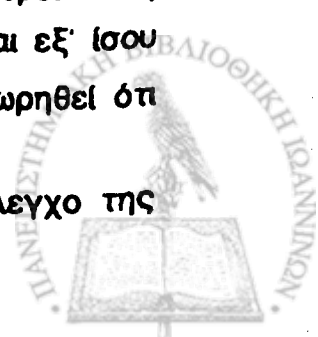
$$H_0 : \mu = \mu_0$$

ως προς τις εναλλακτικές

$$H_a : \mu \neq \mu_0, \quad H_a : \mu < \mu_0, \quad H_a : \mu > \mu_0$$

αντικαθιστούμε κάθε μία από τις τιμές του δείγματος με ένα $+$ ή ένα $-$ ανάλογα με το αν η τιμή είναι μεγαλύτερη ή μικρότερη του μ_0 . Τις μετρήσεις που είναι ίσες με μ_0 δεν τις υπολογίζουμε. Αν η H_0 είναι σωστή, τα πρόσημα $+$ και $-$ θα πρέπει να είναι εξ' ίσου κατανεμημένα και το σύνολο των σημείων $+$ μπορεί να θεωρηθεί ότι προέρχεται από τον διωνυμικό πληθυσμό $B(n, p=1/2)$.

Έτσι, ο αρχικός έλεγχος μεταφέρεται στον έλεγχο της μηδενικής υπόθεσης



$$H_0 : p = \frac{1}{2}$$

ως προς τις εναλλακτικές

$$(i) H_a : p \neq \frac{1}{2}, (ii) H_a : p < \frac{1}{2}, (iii) H_a : p > \frac{1}{2}$$

και γίνεται με το στατιστικό

$$X = \text{άθροισμα των σημείων} + \text{στο δείγμα}$$

που έχει την $B(n, 1/2)$ κατανομή όταν αληθεύει η H_0 .

Οι αντίστοιχες κρίσιμες περιοχές επιπέδου σημαντικότητας α είναι

$$(i) X \geq k_{\alpha/2} \text{ ή } X \leq k'_{\alpha/2}$$

$$(ii) X \leq k'_{\alpha}$$

$$(iii) X \geq k_{\alpha},$$

όπου k_{α} είναι ο μικρότερος ακέραιος για τον οποίο

$$\sum_{y=k_a}^n \binom{n}{y} \left(\frac{1}{2}\right)^n \leq \alpha \text{ και } k'_{\alpha} \text{ είναι ο μεγαλύτερος ακέραιος για τον}$$

$$\text{οποίο } \sum_{y=k'_{\alpha}}^n \binom{n}{y} \left(\frac{1}{2}\right)^n \leq \alpha.$$

Για σχετικά μεγάλα n ($n > 10$) η κανονική προσέγγιση της διωνυμικής κατανομής και η διόρθωση συνεχείας οδηγούν στο στατιστικό



$$Z = \begin{cases} \frac{X - n/2 - 1/2}{\sqrt{n/2}}, & X > n/2 \\ \frac{X - n/2 + 1/2}{\sqrt{n/2}}, & X < n/2 \end{cases}$$

που έχει την $N(0,1)$ κατανομή. Σ' αυτή την περίπτωση οι αντίστοιχες κρίσιμες περιοχές επιπέδου σημαντικότητας α είναι

$$(i) \quad |z| \geq z_{\alpha/2}$$

$$(ii) \quad z \leq -z_{\alpha}$$

$$(iii) \quad z \geq z_{\alpha}$$

Το προσημικό κριτήριο μπορεί να χρησιμοποιηθεί για να αναλύσει δεδομένα διατάξιμης κλίμακας. Είναι απλό στην εφαρμογή του αλλά δεν λαμβάνει υπόψη του το μέγεθος των παρατηρήσεων με αποτέλεσμα να υστερεί σε αποτελεσματικότητα του t -τεστ όταν ο πληθυσμός είναι κανονικός.

Σημειώνουμε ότι αν ο πληθυσμός δεν είναι συμμετρικός μπορούμε να χρησιμοποιήσουμε την ίδια μέθοδο για να ελέγξουμε διάφορες υποθέσεις γύρω από την διάμεσο του πληθυσμού. Μία εφαρμογή του προσημικού κριτηρίου δίνεται στο παράδειγμα που ακολουθεί.

Σε κάθε μία από 19 γάτες αφού δόθηκε δόση 10 mgr προκαΐνης μετρήθηκαν οι παλμοί της καρδιάς της. Τα αποτελέσματα ήταν: 163, 165, 160, 189, 161, 171, 158, 169, 162, 163, 139, 172, 165, 148, 166, 172, 163, 187, 170. Έστω ότι μας ενδιαφέρει να ελέγξουμε την $H_0: \mu = 160$ ως προς την $H_a: \mu > 160$, με $\alpha = 0.05$. Το προσημικό κριτήριο μετά τον προσδιορισμό της H_0 χρησιμοποιεί αντί του αρχικού δείγματος, την διαδοχή προσήμων

+ + + + + - + + + - + + - + + + + +



από την οποία παίρνουμε $X=15$. Επειδή η κρίσιμη περιοχή είναι $X \geq k_{0.05} = 14$, η H_0 απορρίπτεται. Στο ίδιο συμπέρασμα καταλήγουμε χρησιμοποιώντας το προσεγγιστικό τεστ αφού, $z = (15-0.5-18/2)/(1/2)\sqrt{18} = 2.59$ και η κρίσιμη περιοχή είναι $z > z_{0.05} = 1.65$.

7.4 Δύο πληθυσμοί - Δύο δείγματα (συσχετισμένα)

Αρκετά συχνά τα προς ανάλυση δεδομένα αποτελούν ένα τυχαίο δείγμα μεγέθους n από συσχετισμένα ζεύγη μετρήσεων (x_i, y_i) . Δεδομένα αυτής της μορφής εμφανίζονται συνήθως όταν το ίδιο αντικείμενο χρησιμοποιείται ως πειραματική μονάδα και ως μονάδα ελέγχου (control), π.χ. x : το βάρος πριν από κάποια δίαιτα, y : το βάρος έξι βδομάδες μετά την εφαρμογή της. Σκοπός είναι ο έλεγχος της αποτελεσματικότητας κάποιας δοκιμασίας, ο έλεγχος δηλαδή της υπόθεσης ότι τα x και y προέρχονται από τον ίδιο πληθυσμό.

Επειδή τα x_i συσχετίζονται με τα y_i , χρησιμοποιούμε τις διαφορές $d_i = x_i - y_i$. Αν υποθέσουμε ότι τα d_i έχουν κανονική κατανομή τότε για να ελέγξουμε την υπόθεση ότι οι μέσες τιμές των πληθυσμών X και Y είναι ίσες μπορούμε να χρησιμοποιήσουμε το t -τεστ για συγκρίσεις κατά ζεύγη (βλ. Κεφ. 4).

Ο αντίστοιχος έλεγχος, χωρίς την υπόθεση της κανονικότητας, μπορεί να γίνει είτε με το προσημικό τεστ (sign test) για συγκρίσεις κατά ζεύγη, είτε με το προσημικό τεστ των τάξεων του Wilcoxon.

7.4.1 Προσημικό τεστ για συγκρίσεις κατά ζεύγη

Έστω ότι το τυχαίο δείγμα αποτελείται από τα



συσχετισμένα ζεύγη μετρήσεων $(x_1, y_1), \dots, (x_n, y_n)$. Αν για κάθε ζεύγος είναι δυνατή η σύγκριση μεταξύ των x και y , αν δηλαδή τα δεδομένα είναι της διατάξιμης κλίμακας εντός των ζευγών, το προσημικό τεστ (§ 7.3.3) μπορεί να χρησιμοποιηθεί για να ελεγχθεί η υπόθεση ότι οι μέσες τιμές των πληθυσμών X και Y είναι ίσες. Αν υποτεθεί ότι οι διαφορές $d_i = x_i - y_i$ προέρχονται από ένα συνεχή και συμμετρικό πληθυσμό η μηδενική υπόθεση είναι ισοδύναμη με την

$$H_0 : P(d_i > 0) = P(d_i < 0) = 1/2$$

και ελέγχεται με το στατιστικό

X - άθροισμα των θετικών d_i στο δείγμα,

το οποίο έχει $B(n, 1/2)$ κατανομή όταν αληθεύει η H_0 . Οι μετρήσεις για τις οποίες $d_i = 0$ δεν υπολογίζονται, με ανάλογη μείωση του μεγέθους του δείγματος, αφού θεωρητικά τα d_i δεν μπορεί να είναι μηδέν. Οι κρίσιμες περιοχές για μονόπλευρο ή δίπλευρο έλεγχο δίνονται στο εδάφιο 7.3.3.

Ο τρόπος εφαρμογής του προσημικού τεστ για συγκρίσεις κατά ζεύγη φαίνεται στο παράδειγμα που ακολουθεί.

Για να προσδιορισθεί αν μία επεξεργασία θερμότητας είναι αποτελεσματική στο να ελαττώνει ή όχι τον αριθμό των βακτηριδίων του γάλακτος παίρνονται 10 δείγματα και για κάθε δείγμα γίνεται μικροσκοπική καταμέτρηση των βακτηριδίων πριν και μετά την επεξεργασία. Οι διαφορές των λογάριθμων των αριθμών των βακτηριδίων (βλ. Παράδειγμα 4.6) είναι 0.03, 0.21, 1.17, 0.15, -0.02, -0.04, 0.52, 0.22, -0.01, 0.13. Έτσι $X=7$, $n=10$ και η H_0 : η επεξεργασία δεν είναι αποτελεσματική ως προς την H_a : η επεξεργασία οδηγεί σε ελάττωση των βακτηριδίων, δεν απορρίπτεται με $\alpha=0.05$ αφού, $z = \frac{7-0.5-10/2}{\sqrt{10/2}} = 0.95 < z_{0.05} = 1.65$.



7.4.2 Προσημικό τεστ των τάξεων του Wilcoxon

Έστω ότι το τυχαίο δείγμα αποτελείται από τα συσχετισμένα ζεύγη $(x_1, y_1), \dots, (x_n, y_n)$ και έστω οι διαφορές $d_i = x_i - y_i$. Αν υποτεθεί ότι τα δεδομένα προέρχονται από συνεχείς πληθυσμούς και επιπλέον είναι της διατάξεως κλίμακας τόσο εντός των ζευγών όσο και μεταξύ των ζευγών, το προσημικό τεστ των τάξεων (signed ranks test) του Wilcoxon μπορεί να χρησιμοποιηθεί για να ελέγξει την υπόθεση

H_0 : τα d_i προέρχονται από μια συμμετρική περί το μηδέν κατανομή.

Το τεστ του Wilcoxon, σε αντίθεση με το προσημικό τεστ που περιγράψαμε προηγούμενα, χρησιμοποιεί το μέγεθος των d_i σύμφωνα με την παρακάτω διαδικασία.

Αφού δημιουργήσουμε τις διαφορές d_i , τις τοποθετούμε σε αύξουσα σειρά μεγέθους με πρώτη τη μικρότερη χωρίς να λαμβάνουμε υπόψη το πρόσημο και μετά αντικαθιστούμε την j κατά σειρά διαφορά με $+j$ αν $d_i > 0$ ή $-j$ αν $d_i < 0$. Όταν δύο ή περισσότερα d_i είναι ίσα κατ'απόλυτη τιμή, τα d_i αντικαθίστανται με την μέση τιμή των αντίστοιχων τάξεων. Ζεύγη τιμών που οδηγούν σε $d_i = 0$ απορρίπτονται με ανάλογη μείωση του μεγέθους του δείγματος.

Για τον έλεγχο της H_0 χρησιμοποιούμε το στατιστικό

T - άθροισμα των θετικών τάξεων

και απορρίπτουμε την H_0 σε επίπεδο σημαντικότητας α αν

$$T' < T_{\alpha/2, n} \quad , \quad T' = \min\left(T, \frac{n(n+1)}{2} - T\right).$$

Τα αντίστροφα εκατοστιαία σημεία, $T_{\alpha, n}$, της κατανομής του T δίνονται, για $n=6, \dots, 25$, στον πίνακα VII. Για μονόπλευρα τεστ



μεγέθους α η κρίσιμη περιοχή είναι $T' \leq T_{\alpha,n}$.

Όταν $n > 25$ μπορεί ναδειχθεί ότι η κανονική του T όταν αληθεύει η H_0 προσεγγίζει την κανονική με μέση τιμή $\mu_T = n(n+1)/4$ και διακύμανση $\sigma_T^2 = n(n+1)(2n+1)/24$. Έτσι το στατιστικό $Z = (T - \mu_T)/\sigma_T$ έχει προσεγγιστικά τυπική κανονική κατανομή όταν αληθεύει η H_0 και μπορεί να χρησιμοποιηθεί ανάλογα. Σημειώνουμε ότι η προσέγγιση είναι αρκετά καλή ακόμα και για $n \geq 8$.

Το τεστ του Wilcoxon χρησιμοποιεί τα μεγέθη των διαφορών έτσι, είναι αποτελεσματικότερο του προσημικού τεστ. Επίσης, θεωρώντας τις διαφορές $d_i = x_i - m$, το τεστ του Wilcoxon μπορεί να χρησιμοποιηθεί για να ελέγξει κατά πόσον ένα τυχαίο δείγμα $x_1, x_2, \dots, x_n$ προέρχεται από ένα συμμετρικό πληθυσμό με διάμεσο m .

Αν χρησιμοποιήσουμε το τεστ του Wilcoxon για να ελέγξουμε κατά πόσον η επεξεργασία θερμοκρασίας οδηγεί σε ελάττωση των βακτηριδίων του γάλακτος του παραδείγματος της προηγούμενης παραγράφου έχουμε

$$d_i = -0.01, -0.02, 0.03, -0.04, 0.13, 0.15, 0.21, 0.22, 0.52, 1.17$$

$$\text{τάξη} = -1, -2, 3, -4, 5, 6, 7, 8, 9, 10$$

$$T = 3 + 5 + 6 + 7 + 8 + 9 + 10 = 48$$

$$\text{και } T' = \min\left(T, \frac{10(10+1)}{2} - 48\right) = 7.$$

Επειδή $T' = 7 < T_{0.025,10} = 8$ η H_0 : η επεξεργασία δεν είναι αποτελεσματική, απορρίπτεται με $\alpha = 0.025$.



7.5 Δύο πληθυσμοί - Δύο δείγματα (ανεξάρτητα)

Συχνά τα δεδομένα αποτελούνται από δύο ανεξάρτητα τυχαία δείγματα και ενδιαφερόμαστε να ελέγξουμε κατά πόσο προέρχονται από την ίδια κατανομή. Έλεγχοι αυτής της μορφής μπορούν να γίνουν, μεταξύ άλλων και με τα στατιστικά τεστ χ^2 (ομογένειας - ανεξαρτησίας) και Kolmogorov - Smirnov. Αν τα τυχαία δείγματα προέρχονται από κανονικούς πληθυσμούς με κοινή διακύμανση, το γνωστό t-τεστ θα μπορούσε να χρησιμοποιηθεί για να ελέγξει αν οι μέσες τιμές των δύο πληθυσμών είναι ίσες. Όταν υπάρχουν ενδείξεις ότι οι πληθυσμοί δεν είναι κανονικοί, ανάλογοι έλεγχοι γίνονται με το τεστ των Wilcoxon-Mann-Whitney.

7.5.1 χ^2 τεστ ομογένειας ποσοστών - ανεξαρτησίας

Όταν τα δεδομένα αποτελούνται από δύο ανεξάρτητα τυχαία δείγματα και μας ενδιαφέρει να ελέγξουμε αν αυτά προέρχονται από τον ίδιο πληθυσμό το χ^2 -τεστ μπορεί να χρησιμοποιηθεί για να προσδιορίσει τυχόν διαφορές μεταξύ των πληθυσμών των δύο δειγμάτων. Η χρησιμοποίηση του χ^2 -τεστ προϋποθέτει ότι τα δεδομένα, που μπορεί να είναι ακόμα και της ονοματικής κλίμακας αποτελούνται από συχνότητες σε διακριτές κατηγορίες. Έτσι, όταν οι πληθυσμοί διαφέρουν ως προς κάποιο χαρακτηριστικό, αυτή η διαφορά θα εμφανίζεται και στις σχετικές συχνότητες με τις οποίες τα μέλη των πληθυσμών εμπίπτουν σε κάθε κατηγορία. Επειδή η γενίκευση σε περισσότερους από δύο πληθυσμούς είναι άμεση, θα δώσουμε την σχετική μεθοδολογία για την γενική περίπτωση.

Έστω ότι τα δεδομένα συνιστούν r ανεξάρτητα δείγματα από r πληθυσμούς από συχνότητες n_{ij} που αντιστοιχούν σε k κατηγορίες όπως φαίνεται στον παρακάτω $r \times k$ πίνακα



| | | | | |
|----------|----------|---------|----------|----------|
| n_{11} | n_{12} | $\dots$ | n_{1k} | $n_{1.}$ |
| e_{11} | e_{12} | $\dots$ | e_{1k} | |
| n_{21} | n_{22} | $\dots$ | n_{2k} | $n_{2.}$ |
| e_{21} | e_{22} | $\dots$ | e_{2k} | |
| $\vdots$ | | | | |
| $\vdots$ | | | | |
| n_{r1} | n_{r2} | $\dots$ | n_{rk} | $n_{r.}$ |
| e_{r1} | e_{r2} | $\dots$ | e_{rk} | |
| <hr/> | | | | |
| $n_{.1}$ | $n_{.2}$ | $\dots$ | $n_{.k}$ | n |

όπου, $n_{i.} = \sum_j n_{ij}$, $n_{.j} = \sum_i n_{ij}$, $n = \sum_i \sum_j n_{ij}$, $i=1, \dots, r$, $j=1, \dots, k$.

Έστω p_{ij} η πιθανότητα (παράμετρος) τα δεδομένα της κατηγορίας j ($j=1, \dots, k$) να προέρχονται από τον i -οστό πληθυσμό ($i=1, \dots, r$). Κατά την αρχική υπόθεση ότι τα r δείγματα προέρχονται από τον ίδιο πληθυσμό που επιθυμούμε να ελέγξουμε,

$$p_{1j} = p_{2j} = \dots = p_{rj}, \quad j=1, \dots, k.$$

Έτσι, όταν ισχύει η H_0 ,

$$p_{ij} = n_{.j}/n$$

και οι αντίστοιχες αναμενόμενες συχνότητες e_{ij} (δίνονται στον πίνακα) υπολογίζονται από τη σχέση

$$e_{ij} = n_{i.} \frac{n_{.j}}{n}.$$

Ο έλεγχος της μηδενικής υπόθεσης μπορεί να γίνει με το στατιστικό

$$\chi^2 = \sum_i \sum_j (n_{ij} - e_{ij})^2 / e_{ij}$$

που μπορεί να δειχτεί ότι προσεγγιστικά έχει την χ^2 κατανομή με $(r-1)(k-1)$ βαθμούς ελευθερίας. Η προσέγγιση είναι καλύτερη



όταν οι συχνότητες e_{ij} δεν είναι πολύ μικρές (βλέπε σχετικά Εδάφιο 7.3.1). Τώρα, αν οι παρατηρούμενες συχνότητες προέρχονται από τον ίδιο πληθυσμό, δεν θα πρέπει να διαφέρουν κατά πολύ από τις αντίστοιχες αναμενόμενες συχνότητες. Σ'αυτή την περίπτωση οι διαφορές $n_{ij} - e_{ij}$ θα είναι μικρές και κατά συνέπεια η τιμή του στατιστικού θα είναι μικρή.

Έτσι, σε επίπεδο σημαντικότητας α , απορρίπτουμε την H_0 αν

$$\chi^2 \geq \chi^2_{(r-1)(k-1), \alpha}.$$

Σημειώνουμε, ότι τα r ανεξάρτητα δείγματα μπορεί να αποτελούν παρατηρήσεις ενός πληθυσμού με δύο χαρακτηριστικά (X : χρώμα μαλλιών, Y : χρώμα ματιών) σε r και k κατηγορίες αντίστοιχα (X_1 : μαύρο, X_2 : καστανό, Y_1 : μαύρο, Y_2 : καστανό, Y_3 : πράσινο). Ο πίνακας με δεδομένα αυτής της μορφής είναι γνωστός σαν $r \times k$ πίνακας συνάφειας. Ο έλεγχος της υπόθεσης ότι τα δύο χαρακτηριστικά είναι ανεξάρτητα που συνήθως ενδιαφέρει σ'αυτές τις περιπτώσεις γίνεται με το χ^2 -τεστ σύμφωνα με τα παραπάνω.

Όταν $r=k=2$, το στατιστικό χ^2 υπολογίζεται ευκολότερα από την ισοδύναμη σχέση

$$\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$$

που, επιπλέον, εμπεριέχει την λεγόμενη διόρθωση συνέχειας.

Μία εφαρμογή του χ^2 τεστ δίνεται στο παράδειγμα που ακολουθεί.

Μία εταιρεία πρέπει να επιλέξει μεταξύ τριών σχεδίων συνταξιοδότησης. Ο αρμόδιος διευθυντής θέλει να γνωρίζει αν η προτίμηση για τα σχέδια είναι ανεξάρτητη από τη σχέση εργασίας. Γι'αυτό το λόγο πάρθηκε δείγμα από 500 εργαζόμενους οι οποίοι σε ένα ερωτηματολόγιο συμπλήρωσαν τη σχέση εργασίας τους και την προτίμησή τους στο σχέδιο συνταξιοδότησης. Τα δεδομένα συνοψίζονται στον παρακάτω 2×3 πίνακα.



| Σχέδιο
συνταξιοδότησης | 1 | 2 | 3 | Σύνολο |
|---------------------------|----------|----------|--------|--------|
| Σχέση εργασίας | | | | |
| Μισθωτοί | 160(136) | 140(136) | 40(68) | 340 |
| Ωρομίσθιοι | 40 (64) | 60 (64) | 60(32) | 160 |
| Σύνολο | 200 | 200 | 100 | 500 |

Ο έλεγχος της H_0 : η προτίμηση για τα σχέδια είναι ανεξάρτητη από τη σχέση εργασίας, ως προς την H_a : η προτίμηση για τα σχέδια δεν είναι ανεξάρτητη, μπορεί να γίνει με το χ^2 -τεστ. Οι αναμενόμενες συχνότητες e_{ij} όταν αληθεύει η H_0 υπολογίζονται από τη σχέση $e_{ij} = n_{i.} \cdot n_{.j} / n$, ($e_{11} = 340 \times 200 / 500 = 136$, κ.λ.π.) και δίνονται στον πίνακα μέσα στις παρενθέσεις. Από τα δεδομένα του πίνακα έχουμε

$$\begin{aligned}\chi^2 &= \sum_{i=1}^2 \sum_{j=1}^3 (n_{ij} - e_{ij})^2 / e_{ij} \\ &= (160 - 136)^2 / 136 + \dots + (60 - 32)^2 / 32 \\ &= 49.63\end{aligned}$$

Επειδή $\chi^2 = 49.63 \geq \chi^2_{(r-1)(k-1), \alpha} = \chi^2_{2, 0.05} = 5.99$, η αρχική υπόθεση απορρίπτεται, άρα η προτίμηση για τα σχέδια συνταξιοδότησης εξαρτάται από τη σχέση εργασίας.

7.5.2 Το τεστ των Kolmogorov - Smirnov

Έστω τα ανεξάρτητα τυχαία δείγματα $x_1, x_2, \dots, x_m$ και $y_1, y_2, \dots, y_n$ από τους πληθυσμούς $F_X(x)$ και $F_Y(y)$ αντίστοιχα. Για τον έλεγχο της $H_0: F_X(x) = F_Y(y)$ ως προς την $H_a: F_X(x) \neq F_Y(y)$ το τεστ των Kolmogorov - Smirnov χρησιμοποιεί το στατιστικό



$$D = \max_t |S_m(t) - S_n(t)|$$

όπου,

$$S_m(t) = (\text{αριθμός των } x_i \leq t) / m$$

και

$$S_n(t) = (\text{αριθμός των } y_i \leq t) / n,$$

που έχει κατανομή ανεξάρτητη από την κοινή κατανομή της αρχικής υπόθεσης. Ο υπολογισμός των εμπειρικών συναρτήσεων κατανομής $S_m(t)$ και $S_n(t)$ γίνεται εύκολα είτε με ομαδοποίηση των δεδομένων X και Y σε κοινές κατηγορίες, είτε με ανάμειξη των x_i και y_i σε ένα δείγμα και διάταξη του δείγματος κατά πραγματικό μέγεθος (βλ. παράδειγμα).

Όταν η αρχική υπόθεση αληθεύει το D θα πρέπει να είναι μικρό έτσι, απορρίπτουμε την H_0 σε επίπεδο σημαντικότητας α , όταν $D \geq D_\alpha$ όπου τα D_α είναι τα αντίστροφα εκατοσπιαία σημεία της κατανομής του D και δίνονται, προσεγγιστικά, για σχετικά μεγάλα m, n από τις σχέσεις του Πίνακα VIII.

Το τεστ των Kolmogorov - Smirnov είναι ισχυρότερο του χ^2 -τεστ αφού χρησιμοποιεί περισσότερη πληροφορία, αλλά οι πίνακες με τα ακριβή D_α για μικρά δείγματα δεν είναι ευρέως διαδεδομένοι και επιπλέον μπορεί να χρησιμοποιηθεί σε ανάλυση δεδομένων τουλάχιστον της διατάξιμης κλίμακας.

Έστω για παράδειγμα ότι θα μας ενδιέφερε να ελέγξουμε κατά πόσον οι παρατηρήσεις x_i : -2, 1, 3, 4, 8, 9, 10 και y_i : -10, -8, -7, -4, -3, -1, 7, -3, 28, -3 προέρχονται από την ίδια κατανομή. Αν χρησιμοποιήσουμε το τεστ των Kolmogorov - Smirnov θα έχουμε:

| | y | y | y | y | y | y | y | x | y |
|-------------|------|------|------|------|----|----|------|------|------|
| | -10 | -8 | -7 | -4 | -3 | -3 | -3 | -2 | -1 |
| $S_7(t)$ | 0 | 0 | 0 | 0 | | | 0 | 1/7 | 1/7 |
| $S_{10}(t)$ | 1/10 | 2/10 | 3/10 | 4/10 | | | 7/10 | 7/10 | 8/10 |



| | x | x | x | y | x | x | x | y |
|-------------|------|------|------|------|------|------|------|----|
| | 1 | 3 | 4 | 7 | 8 | 9 | 10 | 28 |
| $S_7(t)$ | 2/7 | 3/7 | 4/7 | 4/7 | 5/7 | 6/7 | 1 | 1 |
| $S_{10}(t)$ | 8/10 | 8/10 | 8/10 | 9/10 | 9/10 | 9/10 | 9/10 | 1 |

Ετσι

$$D = \max_t |S_7(t) - S_{10}(t)| = 0.7$$

και

$$D_{0.05} = 1.36 \sqrt{\frac{m+n}{mn}} = 1.36 \sqrt{\frac{7+10}{7 \times 10}} = 0.67$$

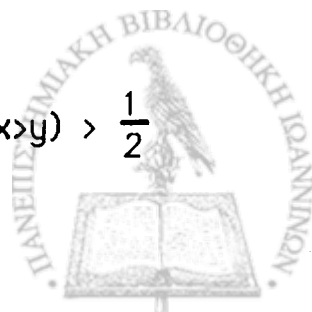
Επειδή $D=0.7 > D_{0.05}=0.67$ η $H_0: F_X(x)=F_Y(y)$ απορρίπτεται σε επίπεδο σημαντικότητας $\alpha=0.05$.

7.5.3 Το τεστ των Wilcoxon - Mann - Whitney

Έστω τα ανεξάρτητα τυχαία δείγματα $x_1, x_2, \dots, x_m$ και $y_1, y_2, \dots, y_n$ από τους πληθυσμούς X και Y αντίστοιχα. Όταν τα δεδομένα είναι της διατάξιμης κλίμακας το τεστ των Wilcoxon - Mann - Whitney μπορεί να χρησιμοποιηθεί για να ελέγξει την αρχική υπόθεση ότι τα δύο δείγματα προέρχονται από τον ίδιο πληθυσμό. Το τεστ αυτό είναι ένα από τα ισχυρότερα και δημοφιλέστερα τεστ της μη παραμετρικής Στατιστικής και, ίσως, η πιο χρήσιμη εναλλακτική λύση του γνωστού t -τεστ (βλ. Κεφ. 4) για τις περιπτώσεις που οι πληθυσμοί δεν είναι κανονικοί ή τα δεδομένα δεν είναι της συνεχούς κλίμακας.

Οι εναλλακτικές υποθέσεις ως προς τις οποίες ελέγχεται η H_0 είναι

$$H_a : P(x > y) = \frac{1}{2}, \quad H_a : P(x > y) < \frac{1}{2}, \quad H_a : P(x > y) > \frac{1}{2}$$



όπου, για παράδειγμα, η τρίτη από τις εναλλακτικές σημαίνει ότι ο πληθυσμός X είναι "στοχαστικά" μεγαλύτερος του Y , δηλαδή ότι ο κύριος όγκος του πληθυσμού X είναι υψηλότερα από ότι του Y . Έτσι, όταν μία τέτοια εναλλακτική υπόθεση αληθεύει, η πιθανότητα μια τυχαία παρατήρηση από τον πληθυσμό X να είναι μεγαλύτερη από μια τυχαία παρατήρηση του πληθυσμού Y , είναι μεγαλύτερη του $1/2$. Ανάλογη είναι και η ερμηνεία για τις άλλες εναλλακτικές.

Αν τα δύο δείγματα αναμειχθούν, διαταχθούν κατ'αύξουσα τάξη μεγέθους και αντικατασταθούν οι αρχικές παρατηρήσεις με μια ακολουθία απλά των x και y , το τεστ των Wilcoxon - Mann - Whitney για να ελέγξει την H_0 , χρησιμοποιεί το στατιστικό

U - αριθμός φορές που μια x παρατήρηση ακολουθεί μια y παρατήρηση.

Όταν αληθεύει η H_0 τα x και y θα πρέπει να είναι έτσι αναμειγμένα ώστε το U να είναι ίσο με $mn/2$ αφού υπάρχουν m παρατηρήσεις x , n παρατηρήσεις y και η πιθανότητα ένα x να ακολουθεί ένα y είναι $1/2$.

Το στατιστικό U μπορεί ευκολότερα να υπολογιστεί αν αθροίσουμε τις τάξεις των x_i στο αναμειγμένο διατεταγμένο δείγμα και αφαιρέσουμε το $m(m+1)/2$. Αυτό προκύπτει από το γεγονός ότι αν η i -οστή κατά σειρά μεγέθους x παρατήρηση έχει τάξη R_i στο αναμειγμένο διατεταγμένο δείγμα, τότε εύκολα μπορεί να δείχτει ότι

$$U = \sum (R_i - i) = \sum R_i - m(m+1)/2.$$

Για παρατηρήσεις x ή/και y που είναι ίσες μεταξύ τους, οι κοινές τους τάξεις στο διατεταγμένο δείγμα προκύπτουν από τον μέσο όρο των τάξεων που θα έπαιρναν οι παρατηρήσεις αυτές αν διέφεραν μεταξύ τους χωρίς να επηρεάζουν την διατάξιμη σχέση τους με τις υπόλοιπες x και y παρατηρήσεις.

Ο πίνακας IX δίνει για $m=1, \dots, n$, $n=3, \dots, 7$ τις πιθανότητες p , για μονόπλευρο τεστ, που αντιστοιχούν, όταν αληθεύει η H_0 , σε τιμές του



$$U' = \min(U, mn - U)$$

μικρότερες ή ίσες με την παρατηρούμενη. Έτσι η H_0 απορρίπτεται, σε επίπεδο σημαντικότητας α , όταν $p \leq \alpha$. Όταν ο έλεγχος γίνεται ως προς την δίπλευρη εναλλακτική, η H_0 απορρίπτεται όταν $2p \leq \alpha$.

Για μεγάλα m και n η κατανομή του U , όταν αληθεύει η H_0 είναι ασυμπτωτικά κανονική με μέση τιμή και διακύμανση

$$\mu_U = mn/2, \quad \sigma_U^2 = mn(m+n+1)/12.$$

Έτσι, προσεγγιστικά, για σχετικά μεγάλα m και n , ο έλεγχος της αρχικής υπόθεσης ως προς κάποια από τις τρεις εναλλακτικές υποθέσεις που δώσαμε, γίνεται κατά τα γνωστά με το στατιστικό

$$Z = \frac{U - \mu_U}{\sigma_U} \text{ που έχει } N(0,1) \text{ όταν αληθεύει η } H_0.$$

Η προσέγγιση είναι αρκετά καλή ακόμα και για $m, n \geq 8$.

Σαν μια εφαρμογή, θα χρησιμοποιήσουμε το τεστ των Wilcoxon - Mann - Whitney για να ελέγξουμε κατά πόσο τα δεδομένα x : 9, 11, 15 και y : 6, 8, 10, 13 προέρχονται από την ίδια κατανομή.

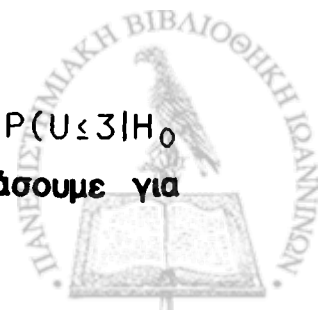
Για τον υπολογισμό του στατιστικού U' έχουμε:

| | y | y | x | y | x | y | x |
|-------------------------------------|-----|-----|-----|-----|-----|-----|-----|
| Αναμετρημένο
διατεταγμένο δείγμα | 6 | 8 | 9 | 10 | 11 | 13 | 15 |
| Τάξεις R_i | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

$$U = \sum R_i - m(n+1)/2 = 3+5+7-3(3+1)/2 = 9$$

$$U' = \min(U, mn - U) = \min(9, 3) = 3.$$

Επειδή, από τον πίνακα IX, έχουμε $p = P(U \leq 3 | H_0 \text{ αληθής}) = 0.20$, πιθανότητα που πρέπει να διπλασιάσουμε για



δίπλευρο τεστ, και $0.40 > \alpha = 0.05$ δεν μπορούμε να απορρίψουμε την αρχική υπόθεση ότι τα δύο δείγματα προέρχονται από την ίδια κατανομή.

7.6 Μέτρα συσχέτισης

Στο εδάφιο 5.14 δώσαμε την έννοια της συσχέτισης για δύο τυχαίες μεταβλητές X και Y και ορίσαμε σαν ένα μέτρο συσχέτισης των τον "θεωρητικό" συντελεστή συσχέτισης $\rho(X,Y)$ όπως και τον δειγματικό συντελεστή συσχέτισης του Pearson $r(X,Y)$ που είναι ένας εκτιμητής του $\rho(X,Y)$ και υπολογίζεται από τα δεδομένα. Όταν έχουμε ένα τυχαίο δείγμα από ζεύγη παρατηρήσεων $(x_1, y_1), \dots, (x_n, y_n)$ των τ.μ. X, Y και επιπλέον οι τ.μ. X, Y είναι κανονικές είδαμε πως μπορούμε να ελέγξουμε την υπόθεση ότι οι τ.μ. X και Y είναι ασυσχέτιστες με διάφορα παραμετρικά τεστ που βασίζονται στον δειγματικό συντελεστή συσχέτισης $r(X,Y)$.

Όταν το δείγμα είναι μικρού μεγέθους και οι τ.μ. X, Y δεν είναι κανονικές ή τα δεδομένα είναι το πολύ της διατάξιμης κλίμακας τα τεστ που στηρίζονται στον δειγματικό συντελεστή $r(X,Y)$ δεν μπορούν να χρησιμοποιηθούν. Γι'αυτές τις περιπτώσεις έχουν διαμορφωθεί κατάλληλα μη παραμετρικά τεστ. Από αυτά, το τεστ που βασίζεται στον συντελεστή συσχέτισης των τάξεων του Spearman, r_s και το τεστ που βασίζεται τον συντελεστή συσχέτισης του Kendall, τ είναι τα σημαντικότερα (Kendall, 1970).

Εδώ θα περιοριστούμε στο τεστ που χρησιμοποιεί τον συντελεστή συσχέτισης του Spearman που αποτελεί το πλέον εύχρηστο και διαδεδομένο από όλα τα σχετικά μη παραμετρικά τεστ.

7.6.1 Συντελεστής συσχέτισης του Spearman

Έστω ένα τυχαίο δείγμα που αποτελείται από τα ζεύγη



των παρατηρήσεων $(x_1, y_1), \dots, (x_n, y_n)$ δύο τυχαίων μεταβλητών X και Y και έστω $R(X_i)$ και $R(Y_i)$ οι τάξεις των μετρήσεων x_i και y_i αντίστοιχα που προκύπτουν αν διατάξουμε τις x και y παρατηρήσεις κατ'αύξουσα σειρά μεγέθους.

Ο συντελεστής συσχέτισης των τάξεων του Spearman, που συμβολίζεται με $r_S(X, Y)$ ή απλά r_S , ορίζεται από τη σχέση

$$r_S = \frac{\sum (R(X_i) - \bar{R}(X))(R(Y_i) - \bar{R}(Y))}{\sqrt{\sum (R(X_i) - \bar{R}(X))^2 \sum (R(Y_i) - \bar{R}(Y))^2}},$$

όπου $\bar{R}(X) = \sum R(X_i)/n$ και $\bar{R}(Y) = \sum R(Y_i)/n$.

Ο συντελεστής συσχέτισης του Spearman r_S παίρνει τιμές μεταξύ του -1 και 1 και έχει ανάλογες ιδιότητες με τον συντελεστή συσχέτισης r του Pearson (§ 5.14). Εύκολα μπορεί ναδειχτεί ότι ο συντελεστής r_S δίνεται από την ισοδύναμη και πιο εύχρηστη σχέση

$$r_S = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

όπου $d_i = R(X_i) - R(Y_i)$, $i=1, \dots, n$.

Όταν δύο ή περισσότερες x ή δύο ή περισσότερες y παρατηρήσεις είναι ίσες μεταξύ τους οι κοινές τους τάξεις στο διατεταγμένο δείγμα των x ή y παρατηρήσεων προκύπτουν από τον μέσο όρο των τάξεων που θα έπαιρναν αν διέφεραν μεταξύ τους χωρίς να επηρεάζουν τη συγκριτική τους σχέση με τις υπόλοιπες x ή y παρατηρήσεις.

Επειδή για τα τυχαία δείγματα από συσχετισμένες τ.μ. ο συντελεστής r_S παίρνει τιμές κοντά στο ± 1 μπορούμε να χρησιμοποιήσουμε το στατιστικό r_S για να διαπιστώσουμε κατά πόσον δύο τ.μ. είναι ασυσχέτιστες.

Στον πίνακα X δίνονται για διάφορα n τα αντίστροφα εκατοστιαία σημεία r_α της κατανομής του r_S που προσδιορίζουν την

κρίσιμη περιοχή για μονόπλευρες εναλλακτικές. Έτσι, για επίπεδο σημαντικότητας α , απορρίπτουμε την $H_0: \rho=0$ όταν $|r_s| > r_{\alpha}$. Για δίπλευρο τεστ η κρίσιμη περιοχή είναι $|r_s| > r_{\alpha/2}$.

Μπορεί να αποδειχθεί ότι, όταν $\rho=0$, το στατιστικό

$$t_s = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$

έχει ασυμπτωτικά την t_{n-2} κατανομή και ότι η προσέγγιση είναι καλή για $n > 10$.

Έτσι, για σχετικά μεγάλα n , μπορούμε να χρησιμοποιήσουμε το στατιστικό t_s για να ελέγξουμε κατά τα γνωστά την αρχική υπόθεση. Για παράδειγμα, αν ο έλεγχος της $H_0: \rho=0$ γίνεται ως προς την εναλλακτική $H_a: \rho \neq 0$, σε επίπεδο σημαντικότητας α , η περιοχή απορρίψης της H_0 είναι

$$|t_s| > t_{\alpha/2, n-2}.$$

Μια εφαρμογή του συντελεστή συσχέτισης του Spearman δίνεται στο παράδειγμα που ακολουθεί.

Με δύο διαφορετικά πιεσόμετρα εξετάσαμε την πίεση αίματος έξι υποτασικών ασθενών. Οι τιμές των δύο οργάνων ήταν:

(8.2, 8.7), (9.6, 9.6), (7.0, 6.9), (9.4, 8.5), (10.9, 11.3), (7.1, 7.6).

Μας ενδιαφέρει να ελέγξουμε κατά πόσον οι ενδείξεις των δύο οργάνων είναι ασυσχέτιστες. Για να υπολογίσουμε τον συντελεστή συσχέτισης του Spearman, έχουμε

| | | | | | | |
|----------|-----|-----|-----|-----|------|-----|
| x_i | 8.2 | 9.6 | 7.0 | 9.4 | 10.9 | 7.1 |
| y_i | 8.7 | 9.6 | 6.9 | 8.5 | 11.3 | 7.6 |
| $R(X_i)$ | 3 | 5 | 1 | 4 | 6 | 2 |
| $R(Y_i)$ | 4 | 5 | 1 | 3 | 6 | 2 |
| d_i | -1 | 0 | 0 | 1 | 0 | 0 |



ΑΣΚΗΣΕΙΣ

Κατά την επαργύρωση χαλκίνου ελάσματος πλάτους 30 cm παρατηρήθηκαν δέκα κενά μικρής επιφάνειας, των οποίων οι αποστάσεις σε cm από την μία άκρη του ελάσματος προς την άλλη και κατά την διεύθυνση του πλάτους ήταν: 4.8, 14.8, 28.2, 23.1, 4.4, 28.7, 19.5, 2.4, 25.0, 6.2. Ελέγξατε την υπόθεση ότι οι παρατηρήσεις προέρχονται από την ομοιόμορφη κατανομή $U(0, 30)$ με τη χρήση (i) του χ^2 -τεστ, (ii) του τεστ των Kolmogorov - Smirnov.

Η συστολική πίεση του αίματος δεκαπέντε ατόμων που εκλέχτηκαν τυχαία από ένα πληθυσμό ήταν: 10.5, 13.5, 12.5, 11.0, 14.0, 15.5, 16.0, 15.0, 12.5, 11.5, 13.0, 14.5, 15.0, 16.5, 18.0. Με τη χρήση του προσημικού τεστ, ελέγξατε κατά πόσον η μέση τιμή του πληθυσμού είναι ίση με 12.

Για να διαπιστωθεί η αποτελεσματικότητα ενός νέου συστήματος ελέγχου της κυκλοφορίας σε μία περιοχή, μετρήθηκε ο αριθμός των σοβαρών ατυχημάτων που έγιναν σε 12 επικίνδυνες διασταυρώσεις στις 4 τελευταίες βδομάδες πριν την εγκατάσταση και στις 4 πρώτες βδομάδες μετά την εγκατάσταση του νέου συστήματος. Ο αριθμός ατυχημάτων πριν και μετά ήταν αντίστοιχα: (3,1), (5,2), (2,0), (3,2), (3,2), (3,0), (0,2), (4,3), (1,3), (6,4), (4,1), (1,0). Χρησιμοποιώντας (i) το προσημικό τεστ των ζευγών, (ii) το προσημικό τεστ του Wilcoxon, ελέγξατε αν το καινούργιο σύστημα είναι πιο αποτελεσματικό.



- 7.4 Οι μαθητές της πρώτης τάξης του Δημοτικού σχολείου χωρίστηκαν σε δύο τμήματα. Στο ένα τμήμα 40 μαθητών διδάχτηκε ανάγνωση με τη μέθοδο M_1 και στο άλλο τμήμα 50 μαθητών, με τη μέθοδο M_2 . Τα ομαδοποιημένα αποτελέσματα των τελικών εξετάσεων στην ανάγνωση δίνονται στον παρακάτω πίνακα.

| | Βαθμός | | | | | | | | | |
|---------|--------|---|---|----|----|----|---|---|---|----|
| Μέθοδοι | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| M_1 | 0 | 0 | 4 | 8 | 13 | 10 | 4 | 0 | 1 | 0 |
| M_2 | 1 | 2 | 3 | 10 | 20 | 10 | 3 | 0 | 0 | 1 |

Με τη χρήση (i) του χ^2 -τεστ, (ii) του τεστ των Kolmogorov - Smirnov, να ελεγχθεί η υπόθεση ότι οι δύο μέθοδοι είναι το ίδιο αποτελεσματικές.

- 7.5 Η διάρκεια ζωής (σε ώρες) για 12 λάμπες φωτισμού που επιλέχτηκαν τυχαία από το είδος Λ_1 και για 12 λάμπες φωτισμού που επιλέχτηκαν τυχαία από το είδος Λ_2 , μετά από συνεχή χρήση ήταν:

| | | | | | | | | | | | | |
|-------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Λ_1 | 407 | 426 | 453 | 378 | 434 | 399 | 441 | 373 | 393 | 386 | 415 | 418 |
| Λ_2 | 403 | 424 | 383 | 445 | 439 | 417 | 412 | 462 | 439 | 432 | 413 | 433 |

Ελέγξατε, με τη χρήση του τεστ των Wilcoxon - Mann - Whitney, αν ευσταθεί η άποψη ότι δεν υπάρχει διαφορά στις μέσες ώρες ζωής για τις λάμπες των δύο ειδών.

- 7.6 Παρακάτω δίνονται οι ώρες προετοιμασίας για μια εξέταση 10 μαθητών και η αντίστοιχη βαθμολογία που προέκυψε από την εξέταση.

| | | | | | | | | | | |
|--------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Ώρες | 8 | 5 | 11 | 13 | 10 | 5 | 18 | 15 | 2 | 8 |
| Βαθμός | 5.6 | 4.4 | 7.9 | 7.2 | 7.0 | 5.4 | 9.4 | 8.5 | 3.3 | 6.5 |

Να ελεγχθεί κατά πόσον οι ώρες μελέτης των μαθητών σχετίζονται με τις επιδόσεις τους στις εξετάσεις.



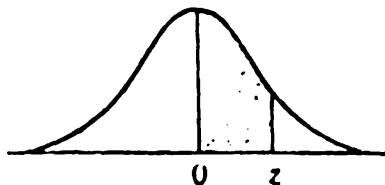
ΠΑΡΑΡΤΗΜΑ

Στατιστικοί Πίνακες

- I Τυπική Κανονική Κατανομή
- II Κατανομή χ^2
- III Κατανομή t
- IV Κατανομή F
- V Τιμές $Q_{\alpha,n}$: Studentized Range
- VI Τιμές $D_{n,\alpha}$: Τεστ των Kolmogorov-Smirnov, ένας πληθυσμός
- VII Τιμές $T_{\alpha,n}$: Τεστ του Wilcoxon
- VIII Τιμές D_{α} : Τεστ των Kolmogorov-Smirnov, δύο πληθυσμοί
- IX Τεστ των Wilcoxon-Mann-Whitney
- X Τιμές r_{α} : Τεστ Συσχέτισης του Spearman.



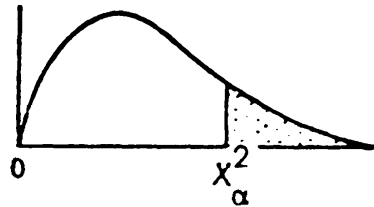
ΠΙΝΑΚΑΣ Ι: Τυπική Κανονική Κατανομή



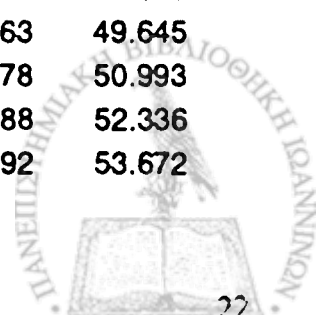
| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.0 | .0000 | .0040 | .0080 | .0120 | .0160 | .0199 | .0239 | .0279 | .0319 | .0359 |
| 0.1 | .0398 | .0438 | .0478 | .0517 | .0557 | .0596 | .0636 | .0675 | .0714 | .0753 |
| 0.2 | .0793 | .0832 | .0871 | .0910 | .0948 | .0987 | .1026 | .1064 | .1103 | .1141 |
| 0.3 | .1179 | .1217 | .1255 | .1293 | .1331 | .1368 | .1406 | .1443 | .1480 | .1517 |
| 0.4 | .1554 | .1591 | .1628 | .1664 | .1700 | .1736 | .1772 | .1808 | .1844 | .1879 |
| 0.5 | .1915 | .1950 | .1985 | .2019 | .2054 | .2088 | .2123 | .2157 | .2190 | .2224 |
| 0.6 | .2257 | .2291 | .2324 | .2357 | .2389 | .2422 | .2454 | .2486 | .2517 | .2549 |
| 0.7 | .2580 | .2611 | .2642 | .2673 | .2704 | .2734 | .2764 | .2794 | .2823 | .2852 |
| 0.8 | .2881 | .2910 | .2939 | .2967 | .2995 | .3023 | .3051 | .3078 | .3106 | .3133 |
| 0.9 | .3159 | .3186 | .3212 | .3238 | .3264 | .3289 | .3315 | .3340 | .3365 | .3389 |
| 1.0 | .3413 | .3438 | .3461 | .3485 | .3508 | .3534 | .3554 | .3577 | .3599 | .3621 |
| 1.1 | .3643 | .3665 | .3686 | .3708 | .3729 | .3749 | .3770 | .3790 | .3810 | .3830 |
| 1.2 | .3849 | .3869 | .3888 | .3907 | .3925 | .3944 | .3962 | .3980 | .3997 | .4015 |
| 1.3 | .4032 | .4049 | .4066 | .4082 | .4099 | .4115 | .4131 | .4147 | .4162 | .4177 |
| 1.4 | .4192 | .4207 | .4222 | .4236 | .4251 | .4265 | .4279 | .4292 | .4306 | .4319 |
| 1.5 | .4332 | .4345 | .4357 | .4370 | .4382 | .4394 | .4406 | .4418 | .4429 | .4441 |
| 1.6 | .4452 | .4463 | .4474 | .4484 | .4495 | .4505 | .4515 | .4525 | .4535 | .4545 |
| 1.7 | .4554 | .4564 | .4573 | .4582 | .4591 | .4599 | .4608 | .4616 | .4625 | .4633 |
| 1.8 | .4641 | .4649 | .4656 | .4664 | .4671 | .4678 | .4686 | .4693 | .4699 | .4706 |
| 1.9 | .4713 | .4719 | .4726 | .4732 | .4738 | .4744 | .4750 | .4756 | .4761 | .4767 |
| 2.0 | .4772 | .4778 | .4783 | .4788 | .4793 | .4798 | .4803 | .4808 | .4812 | .4817 |
| 2.1 | .4821 | .4826 | .4830 | .4834 | .4838 | .4842 | .4846 | .4850 | .4854 | .4857 |
| 2.2 | .4861 | .4864 | .4868 | .4871 | .4875 | .4878 | .4881 | .4884 | .4887 | .4890 |
| 2.3 | .4893 | .4896 | .4898 | .4901 | .4904 | .4906 | .4909 | .4911 | .4913 | .4916 |
| 2.4 | .4918 | .4920 | .4922 | .4925 | .4927 | .4929 | .4931 | .4932 | .4934 | .4936 |
| 2.5 | .4938 | .4940 | .4941 | .4943 | .4945 | .4946 | .4948 | .4949 | .4951 | .4952 |
| 2.6 | .4953 | .4955 | .4956 | .4957 | .4959 | .4960 | .4961 | .4962 | .4963 | .4964 |
| 2.7 | .4965 | .4966 | .4967 | .4968 | .4969 | .4970 | .4971 | .4972 | .4973 | .4974 |
| 2.8 | .4874 | .4975 | .4976 | .4977 | .4977 | .4978 | .4979 | .4979 | .4980 | .4981 |
| 2.9 | .4981 | .4982 | .4982 | .4983 | .4984 | .4984 | .4985 | .4985 | .4986 | .4986 |
| 3.0 | .4987 | .4987 | .4987 | .4988 | .4988 | .4989 | .4989 | .4989 | .4990 | .4990 |

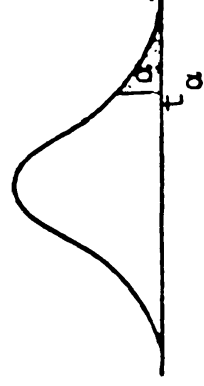
ΠΙΝΑΚΑΣ II: Κατανομή χ^2

Τιμές $\chi^2_{\alpha, \nu}$:



| ν | $\alpha=.995$ | $\alpha=.99$ | $\alpha=.975$ | $\alpha=.95$ | $\alpha=.05$ | $\alpha=.025$ | $\alpha=.01$ | $\alpha=.005$ |
|-------|---------------|--------------|---------------|--------------|--------------|---------------|--------------|---------------|
| 1 | .0000393 | .000157 | .000982 | .00393 | 3.841 | 5.024 | 6.635 | 7.879 |
| 2 | .0100 | .0201 | .0506 | .103 | 5.991 | 7.378 | 9.210 | 10.597 |
| 3 | .0717 | .115 | .216 | .352 | 7.815 | 9.348 | 11.345 | 12.838 |
| 4 | .207 | .297 | .484 | .711 | 9.488 | 11.143 | 13.277 | 14.860 |
| 5 | .412 | .554 | .831 | 1.145 | 11.070 | 12.832 | 15.086 | 16.750 |
| 6 | .676 | .872 | 1.237 | 1.635 | 12.592 | 14.449 | 16.812 | 18.548 |
| 7 | .989 | 1.239 | 1.690 | 2.167 | 14.067 | 16.013 | 18.475 | 20.278 |
| 8 | 1.344 | 1.646 | 2.180 | 2.733 | 15.507 | 17.535 | 20.090 | 21.955 |
| 9 | 1.735 | 2.088 | 2.700 | 3.325 | 16.919 | 19.023 | 21.666 | 23.589 |
| 10 | 2.156 | 2.558 | 3.247 | 3.940 | 18.307 | 20.483 | 23.209 | 25.188 |
| 11 | 2.603 | 3.053 | 3.816 | 4.575 | 19.675 | 21.920 | 24.725 | 26.757 |
| 12 | 3.074 | 3.571 | 4.404 | 5.226 | 21.026 | 23.337 | 26.217 | 28.300 |
| 13 | 3.565 | 4.107 | 5.009 | 5.892 | 22.362 | 24.736 | 27.688 | 29.819 |
| 14 | 4.075 | 4.660 | 5.629 | 6.571 | 23.685 | 26.119 | 29.141 | 31.319 |
| 15 | 4.601 | 5.229 | 6.262 | 7.261 | 24.996 | 27.488 | 30.578 | 32.801 |
| 16 | 5.142 | 5.812 | 6.908 | 7.962 | 26.296 | 28.845 | 32.000 | 34.267 |
| 17 | 5.697 | 6.408 | 7.564 | 8.672 | 27.587 | 30.191 | 33.409 | 35.718 |
| 18 | 6.265 | 7.015 | 8.231 | 9.390 | 28.869 | 31.526 | 34.805 | 37.156 |
| 19 | 6.844 | 7.633 | 8.907 | 10.117 | 30.144 | 32.852 | 36.191 | 38.582 |
| 20 | 7.434 | 8.260 | 9.591 | 10.851 | 31.410 | 34.170 | 37.566 | 39.997 |
| 21 | 8.034 | 8.897 | 10.283 | 11.591 | 32.671 | 35.479 | 38.932 | 41.401 |
| 22 | 8.643 | 9.542 | 10.982 | 12.338 | 33.924 | 36.781 | 40.289 | 42.796 |
| 23 | 9.260 | 10.196 | 11.689 | 13.091 | 35.172 | 38.076 | 41.638 | 44.181 |
| 24 | 9.886 | 10.856 | 12.401 | 13.848 | 36.415 | 39.364 | 42.980 | 45.558 |
| 25 | 10.520 | 11.524 | 13.120 | 14.611 | 37.652 | 40.646 | 44.314 | 46.928 |
| 26 | 11.160 | 12.198 | 13.844 | 15.379 | 38.885 | 41.923 | 45.642 | 48.290 |
| 27 | 11.808 | 12.879 | 14.573 | 16.151 | 40.113 | 43.194 | 46.963 | 49.645 |
| 28 | 12.461 | 13.565 | 15.308 | 16.928 | 41.337 | 44.461 | 48.278 | 50.993 |
| 29 | 13.121 | 14.256 | 16.047 | 17.708 | 42.557 | 45.722 | 49.588 | 52.336 |
| 30 | 13.787 | 14.953 | 16.791 | 18.493 | 43.773 | 46.979 | 50.892 | 53.672 |



ΠΙΝΑΚΑΣ III: Κατανομή t Τιμές $t_{\alpha, \nu}$:

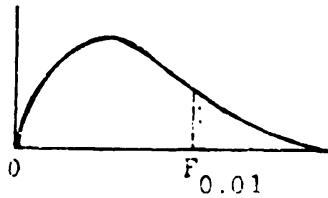
| ν | $\alpha=10$ | $\alpha=05$ | $\alpha=025$ | $\alpha=01$ | $\alpha=005$ | ν |
|-------|-------------|-------------|--------------|-------------|--------------|-------|
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 1 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 2 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 3 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 4 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 6 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 7 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 8 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 9 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 10 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 11 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 12 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 13 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 14 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 15 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 16 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 17 |
| 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 18 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 19 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 20 |
| 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 21 |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 22 |
| 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 23 |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 24 |
| 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 25 |
| 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 26 |
| 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 27 |
| 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 28 |
| 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 29 |
| inf. | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | inf. |

ΠΙΝΑΚΑΣ IV: Κατανομή F

ν_1 -Βαθμοί Ελευθερίας Αριθμητή

Τιμές

$F_{0.01, \nu_1, \nu_2}$



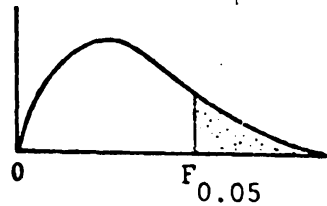
ν_2 -Βαθμοί Ελευθερίας Παρονομαστή

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ |
|----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|----------|
| 1 | 4052 | 5000 | 5403 | 5625 | 5764 | 5859 | 5928 | 5982 | 6023 | 6056 | 6106 | 6157 | 6209 | 6235 | 6261 | 6287 | 6313 | 6339 | 6366 |
| 2 | 98.5 | 99.1 | 99.2 | 99.2 | 99.3 | 99.3 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.5 | 99.5 | 99.5 | 99.5 | 99.5 | 99.5 |
| 3 | 34.1 | 30.8 | 29.5 | 28.7 | 28.2 | 27.9 | 27.7 | 27.5 | 27.3 | 27.2 | 27.1 | 26.9 | 26.7 | 26.6 | 26.5 | 26.4 | 26.3 | 26.2 | 26.1 |
| 4 | 21.2 | 18.1 | 16.7 | 16.1 | 15.5 | 15.2 | 15.1 | 14.8 | 14.7 | 14.5 | 14.4 | 14.2 | 14.1 | 13.9 | 13.8 | 13.7 | 13.7 | 13.6 | 13.5 |
| 5 | 16.3 | 13.3 | 12.1 | 11.4 | 11.1 | 10.7 | 10.5 | 10.3 | 10.2 | 10.1 | 9.89 | 9.72 | 9.55 | 9.47 | 9.38 | 9.29 | 9.20 | 9.11 | 9.02 |
| 6 | 13.7 | 10.9 | 9.78 | 9.15 | 8.75 | 8.47 | 8.26 | 8.10 | 7.98 | 7.87 | 7.72 | 7.56 | 7.40 | 7.31 | 7.23 | 7.14 | 7.06 | 6.97 | 6.88 |
| 7 | 12.2 | 9.55 | 8.45 | 7.85 | 7.46 | 7.19 | 6.99 | 6.84 | 6.72 | 6.62 | 6.47 | 6.31 | 6.16 | 6.07 | 5.99 | 5.91 | 5.82 | 5.74 | 5.65 |
| 8 | 11.3 | 8.65 | 7.59 | 7.01 | 6.63 | 6.37 | 6.18 | 6.03 | 5.91 | 5.81 | 5.67 | 5.52 | 5.36 | 5.28 | 5.20 | 5.12 | 5.03 | 4.95 | 4.86 |
| 9 | 10.6 | 8.02 | 6.99 | 6.42 | 6.06 | 5.80 | 5.61 | 5.47 | 5.35 | 5.26 | 5.11 | 4.96 | 4.81 | 4.73 | 4.65 | 4.57 | 4.48 | 4.40 | 4.31 |
| 10 | 10.1 | 7.56 | 6.55 | 5.99 | 5.64 | 5.39 | 5.20 | 5.06 | 4.94 | 4.85 | 4.71 | 4.56 | 4.41 | 4.33 | 4.25 | 4.17 | 4.08 | 4.00 | 3.91 |
| 11 | 9.65 | 7.21 | 6.22 | 5.67 | 5.32 | 5.07 | 4.89 | 4.74 | 4.63 | 4.54 | 4.40 | 4.25 | 4.10 | 4.02 | 3.94 | 3.86 | 3.78 | 3.69 | 3.60 |
| 12 | 9.33 | 6.93 | 5.95 | 5.41 | 5.06 | 4.82 | 4.64 | 4.50 | 4.39 | 4.30 | 4.16 | 4.01 | 3.86 | 3.78 | 3.70 | 3.62 | 3.54 | 3.45 | 3.36 |
| 13 | 9.07 | 6.70 | 5.74 | 5.21 | 4.86 | 4.62 | 4.44 | 4.30 | 4.19 | 4.10 | 3.96 | 3.82 | 3.66 | 3.59 | 3.51 | 3.43 | 3.34 | 3.25 | 3.17 |
| 14 | 8.86 | 6.51 | 5.56 | 5.04 | 4.70 | 4.46 | 4.28 | 4.14 | 4.03 | 3.94 | 3.80 | 3.66 | 3.51 | 3.43 | 3.35 | 3.27 | 3.18 | 3.09 | 3.00 |
| 15 | 8.68 | 6.36 | 5.42 | 4.89 | 4.56 | 4.32 | 4.14 | 4.00 | 3.89 | 3.80 | 3.67 | 3.52 | 3.37 | 3.29 | 3.21 | 3.13 | 3.05 | 2.96 | 2.87 |
| 16 | 8.53 | 6.23 | 5.29 | 4.77 | 4.44 | 4.20 | 4.03 | 3.89 | 3.78 | 3.69 | 3.55 | 3.41 | 3.26 | 3.18 | 3.10 | 3.02 | 2.93 | 2.84 | 2.75 |
| 17 | 8.40 | 6.11 | 5.19 | 4.67 | 4.34 | 4.10 | 3.93 | 3.79 | 3.68 | 3.59 | 3.46 | 3.31 | 3.16 | 3.08 | 3.00 | 2.92 | 2.83 | 2.75 | 2.65 |
| 18 | 8.29 | 6.01 | 5.09 | 4.58 | 4.25 | 4.01 | 3.84 | 3.71 | 3.60 | 3.51 | 3.37 | 3.23 | 3.08 | 3.00 | 2.92 | 2.84 | 2.75 | 2.66 | 2.57 |
| 19 | 8.19 | 5.93 | 5.01 | 4.50 | 4.17 | 3.94 | 3.77 | 3.63 | 3.52 | 3.43 | 3.30 | 3.15 | 3.00 | 2.92 | 2.84 | 2.76 | 2.67 | 2.58 | 2.49 |
| 20 | 8.10 | 5.85 | 4.94 | 4.43 | 4.10 | 3.87 | 3.70 | 3.56 | 3.46 | 3.37 | 3.23 | 3.09 | 2.94 | 2.86 | 2.78 | 2.69 | 2.61 | 2.52 | 2.42 |
| 21 | 8.02 | 5.78 | 4.87 | 4.37 | 4.04 | 3.81 | 3.64 | 3.51 | 3.40 | 3.31 | 3.17 | 3.03 | 2.88 | 2.8 | 2.72 | 2.64 | 2.55 | 2.46 | 2.36 |
| 22 | 7.95 | 5.72 | 4.82 | 4.31 | 3.99 | 3.76 | 3.59 | 3.45 | 3.35 | 3.26 | 3.12 | 2.98 | 2.83 | 2.75 | 2.67 | 2.58 | 2.50 | 2.40 | 2.31 |
| 23 | 7.88 | 5.66 | 4.76 | 4.26 | 3.94 | 3.71 | 3.54 | 3.41 | 3.30 | 3.21 | 3.07 | 2.93 | 2.78 | 2.70 | 2.62 | 2.54 | 2.45 | 2.35 | 2.26 |
| 24 | 7.82 | 5.61 | 4.72 | 4.22 | 3.90 | 3.67 | 3.50 | 3.36 | 3.26 | 3.17 | 3.03 | 2.89 | 2.74 | 2.66 | 2.58 | 2.49 | 2.4 | 2.31 | 2.21 |
| 25 | 7.77 | 5.57 | 4.68 | 4.18 | 3.86 | 3.63 | 3.46 | 3.32 | 3.22 | 3.13 | 2.99 | 2.85 | 2.70 | 2.62 | 2.53 | 2.45 | 2.36 | 2.27 | 2.17 |
| 30 | 7.56 | 5.39 | 4.51 | 4.02 | 3.70 | 3.47 | 3.30 | 3.17 | 3.07 | 2.98 | 2.84 | 2.70 | 2.55 | 2.47 | 2.39 | 2.30 | 2.21 | 2.11 | 2.01 |
| 40 | 7.31 | 5.18 | 4.31 | 3.83 | 3.51 | 3.29 | 3.12 | 2.99 | 2.89 | 2.80 | 2.66 | 2.52 | 2.37 | 2.29 | 2.20 | 2.11 | 2.02 | 1.92 | 1.80 |
| 60 | 7.08 | 4.98 | 4.13 | 3.65 | 3.34 | 3.12 | 2.95 | 2.82 | 2.72 | 2.63 | 2.50 | 2.35 | 2.20 | 2.12 | 2.03 | 1.94 | 1.84 | 1.73 | 1.60 |
| 120 | 6.85 | 4.79 | 3.95 | 3.48 | 3.17 | 2.96 | 2.79 | 2.66 | 2.56 | 2.47 | 2.34 | 2.19 | 2.03 | 1.95 | 1.86 | 1.76 | 1.66 | 1.53 | 1.38 |
| ∞ | 6.63 | 4.61 | 3.78 | 3.32 | 3.02 | 2.80 | 2.64 | 2.51 | 2.41 | 2.32 | 2.18 | 2.04 | 1.88 | 1.79 | 1.70 | 1.59 | 1.47 | 1.32 | 1.00 |

ΠΙΝΑΚΑΣ IV: Κατανομή F

ν_1 -Βαθμοί Ελευθερίας Αριθμητή

Τιμές
 $F_{0.05, \nu_1, \nu_2}$



ν_2 -Βαθμοί Ελευθερίας Παρονομαστή

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ |
|----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|----------|
| 1 | 161 | 200 | 216 | 225 | 230 | 234 | 237 | 239 | 241 | 242 | 244 | 246 | 248 | 249 | 250 | 251 | 252 | 253 | 254 |
| 2 | 18.5 | 19.1 | 19.2 | 19.2 | 19.3 | 19.3 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 | 19.5 | 19.5 | 19.5 | 19.5 | 19.5 | 19.5 |
| 3 | 10.1 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 | 8.79 | 8.74 | 8.70 | 8.66 | 8.64 | 8.62 | 8.59 | 8.57 | 8.55 | 8.53 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.40 | 6.00 | 5.96 | 5.91 | 5.86 | 5.80 | 5.77 | 5.75 | 5.72 | 5.69 | 5.66 | 5.63 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 | 4.68 | 4.62 | 4.56 | 4.53 | 4.50 | 4.46 | 4.43 | 4.40 | 4.37 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 | 4.00 | 3.94 | 3.87 | 3.84 | 3.81 | 3.77 | 3.74 | 3.7 | 3.67 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 | 3.57 | 3.51 | 3.44 | 3.41 | 3.38 | 3.34 | 3.3 | 3.27 | 3.23 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 | 3.28 | 3.22 | 3.15 | 3.12 | 3.08 | 3.04 | 3.01 | 2.97 | 2.93 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 | 3.07 | 3.01 | 2.94 | 2.90 | 2.86 | 2.83 | 2.79 | 2.75 | 2.71 |
| 10 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 | 2.91 | 2.85 | 2.77 | 2.74 | 2.70 | 2.66 | 2.62 | 2.58 | 2.54 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.85 | 2.79 | 2.72 | 2.65 | 2.61 | 2.57 | 2.53 | 2.49 | 2.45 | 2.40 |
| 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 | 2.75 | 2.69 | 2.62 | 2.54 | 2.51 | 2.47 | 2.43 | 2.38 | 2.34 | 2.30 |
| 13 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | 2.67 | 2.60 | 2.53 | 2.46 | 2.42 | 2.38 | 2.34 | 2.30 | 2.25 | 2.21 |
| 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 | 2.60 | 2.53 | 2.46 | 2.39 | 2.35 | 2.31 | 2.27 | 2.22 | 2.18 | 2.13 |
| 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 | 2.54 | 2.48 | 2.40 | 2.33 | 2.29 | 2.25 | 2.20 | 2.16 | 2.11 | 2.07 |
| 16 | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 | 2.49 | 2.42 | 2.35 | 2.28 | 2.24 | 2.19 | 2.15 | 2.11 | 2.06 | 2.01 |
| 17 | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 | 2.45 | 2.38 | 2.31 | 2.23 | 2.19 | 2.15 | 2.10 | 2.06 | 2.01 | 1.96 |
| 18 | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 | 2.41 | 2.34 | 2.27 | 2.19 | 2.15 | 2.11 | 2.06 | 2.02 | 1.97 | 1.92 |
| 19 | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 | 2.38 | 2.31 | 2.23 | 2.16 | 2.11 | 2.07 | 2.03 | 1.98 | 1.93 | 1.88 |
| 20 | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.6 | 2.51 | 2.45 | 2.39 | 2.35 | 2.28 | 2.20 | 2.12 | 2.08 | 2.04 | 1.99 | 1.95 | 1.90 | 1.84 |
| 21 | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.37 | 2.32 | 2.25 | 2.18 | 2.10 | 2.05 | 2.01 | 1.96 | 1.92 | 1.87 | 1.81 |
| 22 | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.34 | 2.30 | 2.23 | 2.15 | 2.07 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 1.78 |
| 23 | 4.28 | 3.42 | 3.03 | 2.8 | 2.64 | 2.53 | 2.44 | 2.37 | 2.32 | 2.27 | 2.20 | 2.13 | 2.05 | 2.01 | 1.96 | 1.91 | 1.86 | 1.81 | 1.76 |
| 24 | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.30 | 2.25 | 2.18 | 2.11 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 1.79 | 1.73 |
| 25 | 4.24 | 3.39 | 2.99 | 2.76 | 2.6 | 2.49 | 2.40 | 2.34 | 2.28 | 2.24 | 2.16 | 2.09 | 2.01 | 1.86 | 1.92 | 1.87 | 1.82 | 1.77 | 1.71 |
| 30 | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.21 | 2.16 | 2.09 | 2.01 | 1.93 | 1.89 | 1.84 | 1.79 | 1.74 | 1.68 | 1.62 |
| 40 | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 | 2.08 | 2.00 | 1.92 | 1.84 | 1.79 | 1.74 | 1.69 | 1.64 | 1.58 | 1.51 |
| 60 | 4.00 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 | 1.99 | 1.92 | 1.84 | 1.75 | 1.70 | 1.65 | 1.59 | 1.53 | 1.47 | 1.39 |
| 120 | 3.92 | 3.07 | 2.68 | 2.45 | 2.29 | 2.18 | 2.09 | 2.02 | 1.96 | 1.91 | 1.83 | 1.75 | 1.66 | 1.61 | 1.55 | 1.50 | 1.43 | 1.35 | 1.25 |
| ∞ | 3.84 | 3.00 | 2.60 | 2.37 | 2.21 | 2.10 | 2.01 | 1.94 | 1.88 | 1.83 | 1.75 | 1.67 | 1.57 | 1.52 | 1.46 | 1.39 | 1.32 | 1.22 | 1.00 |

ΠΙΝΑΚΑΣ V: Κατανομή Studentized Range

Τιμές $Q_{0,05}$

Αριθμός Δοκιμασιών r

Βαθμοί Ελευθερίας n

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|----------|
| 1 | 18.1 | 27.1 | 32.8 | 37.2 | 40.5 | 43.1 | 45.4 | 47.3 | 49.1 | 50.6 | 51.9 | 53.2 | 54.3 | 55.4 | 56.3 | 57.2 | 58.1 | 58.8 | 59.6 |
| 2 | 6.09 | 8.33 | 9.80 | 10.9 | 11.7 | 12.4 | 13.1 | 13.5 | 14.1 | 14.4 | 14.8 | 15.1 | 15.4 | 15.7 | 15.9 | 16.1 | 16.4 | 16.6 | 16.8 |
| 3 | 4.50 | 5.91 | 6.83 | 7.51 | 8.04 | 8.47 | 8.85 | 9.18 | 9.46 | 9.72 | 9.75 | 10.2 | 10.4 | 10.5 | 10.7 | 10.8 | 10.9 | 11.1 | 11.2 |
| 4 | 3.93 | 5.04 | 5.76 | 6.29 | 6.71 | 7.06 | 7.35 | 7.60 | 7.83 | 8.03 | 8.21 | 8.37 | 8.52 | 8.67 | 8.80 | 8.92 | 9.03 | 9.14 | 9.24 |
| 5 | 3.64 | 4.60 | 5.22 | 5.67 | 6.03 | 6.33 | 6.58 | 6.80 | 6.99 | 7.17 | 7.32 | 7.47 | 7.60 | 7.72 | 7.83 | 7.93 | 8.03 | 8.12 | 8.21 |
| 6 | 3.46 | 4.34 | 4.90 | 5.31 | 5.63 | 5.89 | 6.12 | 6.32 | 6.49 | 6.65 | 6.79 | 6.92 | 7.04 | 7.14 | 7.24 | 7.34 | 7.43 | 7.51 | 7.59 |
| 7 | 3.34 | 4.16 | 4.68 | 5.06 | 5.35 | 5.59 | 5.80 | 5.99 | 6.15 | 6.29 | 6.42 | 6.54 | 6.65 | 6.75 | 6.84 | 6.93 | 7.01 | 7.08 | 7.16 |
| 8 | 3.26 | 4.04 | 4.53 | 4.89 | 5.17 | 5.40 | 5.60 | 5.77 | 5.92 | 6.05 | 6.18 | 6.29 | 6.39 | 6.48 | 6.57 | 6.65 | 6.73 | 6.80 | 6.87 |
| 9 | 3.20 | 3.95 | 4.42 | 4.76 | 5.02 | 5.24 | 5.43 | 5.60 | 5.74 | 5.87 | 5.98 | 6.09 | 6.19 | 6.28 | 6.36 | 6.44 | 6.51 | 6.58 | 6.65 |
| 10 | 3.15 | 3.88 | 4.33 | 4.66 | 4.91 | 5.12 | 5.30 | 5.46 | 5.60 | 5.72 | 5.83 | 5.93 | 6.03 | 6.12 | 6.20 | 6.27 | 6.34 | 6.41 | 6.47 |
| 11 | 3.11 | 3.82 | 4.26 | 4.58 | 4.82 | 5.03 | 5.20 | 5.35 | 5.49 | 5.61 | 5.71 | 5.81 | 5.90 | 5.98 | 6.06 | 6.14 | 6.20 | 6.27 | 6.33 |
| 12 | 3.08 | 3.77 | 4.20 | 4.51 | 4.75 | 4.95 | 5.12 | 5.27 | 5.40 | 5.51 | 5.61 | 5.71 | 5.80 | 5.88 | 5.95 | 6.02 | 6.09 | 6.15 | 6.21 |
| 13 | 3.06 | 3.73 | 4.15 | 4.46 | 4.69 | 4.88 | 5.05 | 5.19 | 5.32 | 5.43 | 5.53 | 5.63 | 5.71 | 5.79 | 5.86 | 5.93 | 6.00 | 6.06 | 6.11 |
| 14 | 3.03 | 3.70 | 4.11 | 4.41 | 4.64 | 4.83 | 4.99 | 5.13 | 5.25 | 5.36 | 5.46 | 5.56 | 5.64 | 5.72 | 5.79 | 5.86 | 5.92 | 5.98 | 6.03 |
| 15 | 3.01 | 3.67 | 4.08 | 4.37 | 4.59 | 4.78 | 4.94 | 5.08 | 5.20 | 5.31 | 5.40 | 5.49 | 5.57 | 5.65 | 5.72 | 5.79 | 5.85 | 5.91 | 5.96 |
| 16 | 3.00 | 3.65 | 4.05 | 4.34 | 4.56 | 4.74 | 4.90 | 5.03 | 5.15 | 5.26 | 5.35 | 5.44 | 5.52 | 5.59 | 5.66 | 5.73 | 5.79 | 5.84 | 5.90 |
| 17 | 2.98 | 3.62 | 4.02 | 4.31 | 4.52 | 4.70 | 4.86 | 4.99 | 5.11 | 5.21 | 5.31 | 5.39 | 5.47 | 5.55 | 5.61 | 5.68 | 5.74 | 5.79 | 5.84 |
| 18 | 2.97 | 3.61 | 4.00 | 4.28 | 4.49 | 4.67 | 4.83 | 4.96 | 5.07 | 5.17 | 5.27 | 5.35 | 5.43 | 5.50 | 5.57 | 5.63 | 5.69 | 5.74 | 5.79 |
| 19 | 2.96 | 3.59 | 3.98 | 4.26 | 4.47 | 4.64 | 4.79 | 4.92 | 5.04 | 5.14 | 5.23 | 5.32 | 5.39 | 5.46 | 5.53 | 5.59 | 5.65 | 5.70 | 5.75 |
| 20 | 2.95 | 3.58 | 3.96 | 4.24 | 4.45 | 4.62 | 4.77 | 4.90 | 5.01 | 5.11 | 5.20 | 5.28 | 5.36 | 5.43 | 5.50 | 5.56 | 5.61 | 5.66 | 5.71 |
| 24 | 2.92 | 3.53 | 3.90 | 4.17 | 4.37 | 4.54 | 4.68 | 4.81 | 4.92 | 5.01 | 5.10 | 5.18 | 5.25 | 5.32 | 5.38 | 5.44 | 5.50 | 5.55 | 5.59 |
| 30 | 2.89 | 3.48 | 3.84 | 4.11 | 4.30 | 4.46 | 4.60 | 4.72 | 4.83 | 4.92 | 5.00 | 5.08 | 5.15 | 5.21 | 5.27 | 5.33 | 5.38 | 5.43 | 5.48 |
| 40 | 2.86 | 3.44 | 3.79 | 4.04 | 4.23 | 4.39 | 4.52 | 4.63 | 4.74 | 4.82 | 4.90 | 4.98 | 5.05 | 5.11 | 5.17 | 5.22 | 5.27 | 5.32 | 5.36 |
| 60 | 2.83 | 3.40 | 3.74 | 3.98 | 4.16 | 4.31 | 4.44 | 4.55 | 4.65 | 4.73 | 4.81 | 4.88 | 4.94 | 5.00 | 5.06 | 5.11 | 5.15 | 5.20 | 5.24 |
| 120 | 2.80 | 3.36 | 3.69 | 3.92 | 4.10 | 4.24 | 4.36 | 4.47 | 4.56 | 4.64 | 4.71 | 4.78 | 4.84 | 4.90 | 4.95 | 5.00 | 5.04 | 5.09 | 5.13 |
| ∞ | 2.77 | 3.32 | 3.63 | 3.86 | 4.03 | 4.17 | 4.29 | 4.39 | 4.47 | 4.55 | 4.62 | 4.68 | 4.74 | 4.80 | 4.84 | 4.89 | 4.93 | 4.97 | 5.01 |

Πίνακας VI: Κρίσιμη περιοχή για το τεστ των
Kolmogorov - Smirnov για ένα πληθυσμό

| n | Τιμές $D_{\alpha,n}$ | | | | |
|-----|---------------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| | Επίπεδο σημαντικότητας α | | | | |
| | .20 | .15 | .10 | .05 | .01 |
| 1 | .900 | .925 | .950 | .975 | .995 |
| 2 | .684 | .726 | .776 | .842 | .929 |
| 3 | .565 | .597 | .642 | .708 | .828 |
| 4 | .494 | .525 | .564 | .624 | .733 |
| 5 | .446 | .474 | .510 | .565 | .669 |
| 6 | .410 | .436 | .470 | .521 | .618 |
| 7 | .381 | .405 | .438 | .486 | .577 |
| 8 | .358 | .381 | .411 | .457 | .543 |
| 9 | .339 | .360 | .388 | .432 | .514 |
| 10 | .322 | .342 | .368 | .410 | .490 |
| 11 | .307 | .326 | .352 | .391 | .468 |
| 12 | .295 | .313 | .338 | .375 | .450 |
| 13 | .284 | .302 | .325 | .361 | .433 |
| 14 | .274 | .292 | .314 | .349 | .418 |
| 15 | .266 | .283 | .304 | .338 | .404 |
| 16 | .258 | .274 | .295 | .328 | .392 |
| 17 | .250 | .266 | .286 | .318 | .381 |
| 18 | .244 | .259 | .278 | .309 | .371 |
| 19 | .237 | .252 | .272 | .301 | .363 |
| 20 | .231 | .246 | .264 | .294 | .356 |
| 25 | .21 | .22 | .24 | .27 | .32 |
| 30 | .19 | .20 | .22 | .24 | .29 |
| 35 | .18 | .19 | .21 | .23 | .27 |
| >35 | $\frac{1.07}{\sqrt{n}}$ | $\frac{1.14}{\sqrt{n}}$ | $\frac{1.22}{\sqrt{n}}$ | $\frac{1.36}{\sqrt{n}}$ | $\frac{1.63}{\sqrt{n}}$ |



Πίνακας VII: Κρίσιμη περιοχή για το προσημικό τεστ του Wilcoxon

| n | Τιμές T _{α,n} | | |
|----|--------------------------------|--------------|---------------|
| | Επίπεδο σημαντικότητας α | | |
| | για μονόπλευρο (δίπλευρο) τεστ | | |
| | .025
(.05) | .01
(.02) | .005
(.01) |
| 6 | 0 | - | - |
| 7 | 2 | 0 | - |
| 8 | 4 | 2 | 0 |
| 9 | 6 | 3 | 2 |
| 10 | 8 | 5 | 3 |
| 11 | 11 | 7 | 5 |
| 12 | 14 | 10 | 7 |
| 13 | 17 | 13 | 10 |
| 14 | 21 | 16 | 13 |
| 15 | 25 | 20 | 16 |
| 16 | 30 | 24 | 20 |
| 17 | 35 | 28 | 23 |
| 18 | 40 | 33 | 28 |
| 19 | 46 | 38 | 32 |
| 20 | 52 | 43 | 38 |
| 21 | 59 | 49 | 43 |
| 22 | 66 | 56 | 49 |
| 23 | 73 | 62 | 55 |
| 24 | 81 | 69 | 61 |
| 25 | 89 | 77 | 68 |



Πίνακας VIII: Κρίσιμη περιοχή για το τεστ των
Κολμογορον - Σμιρνον για δύο πληθυσμούς
(μεγάλα δείγματα - δίπλευρο τεστ)

| Επίπεδο σημαντικότητας α | Τιμές D_α |
|---------------------------------|-----------------------------|
| .10 | $1.22\sqrt{\frac{m+n}{mn}}$ |
| .05 | $1.36\sqrt{\frac{m+n}{mn}}$ |
| .025 | $1.48\sqrt{\frac{m+n}{mn}}$ |
| .01 | $1.63\sqrt{\frac{m+n}{mn}}$ |
| .005 | $1.73\sqrt{\frac{m+n}{mn}}$ |
| .001 | $1.95\sqrt{\frac{m+n}{mn}}$ |



Πίνακας ΙΧ: Πιθανότητες που αντιστοιχούν σε τιμές του U μικρότερες ή ίσες με τις παρατηρούμενες για το τεστ των Wilcoxon - Mann - Whitney

| n-3 | | | | n-4 | | | | | | |
|-----|---|------|------|------|---|---|------|------|------|------|
| U | m | 1 | 2 | 3 | U | m | 1 | 2 | 3 | 4 |
| 0 | | .250 | .100 | .050 | 0 | | .200 | .067 | .028 | .014 |
| 1 | | .500 | .200 | .100 | 1 | | .400 | .133 | .057 | .029 |
| 2 | | .750 | .400 | .200 | 2 | | .600 | .267 | .114 | .057 |
| 3 | | | .600 | .350 | 3 | | | .400 | .200 | .100 |
| 4 | | | | .500 | 4 | | | .600 | .314 | .171 |
| 5 | | | | .650 | 5 | | | | .429 | .243 |
| | | | | | 6 | | | | .571 | .343 |
| | | | | | 7 | | | | | .443 |
| | | | | | 8 | | | | | .557 |

| n-5 | | | | | | n-6 | | | | | | | | |
|-----|---|------|------|------|------|------|----|---|------|------|------|------|------|------|
| U | m | 1 | 2 | 3 | 4 | 5 | U | m | 1 | 2 | 3 | 4 | 5 | 6 |
| 0 | | .167 | .047 | .018 | .008 | .004 | 0 | | .143 | .036 | .012 | .005 | .002 | .001 |
| 1 | | .333 | .095 | .036 | .016 | .008 | 1 | | .286 | .071 | .024 | .010 | .004 | .002 |
| 2 | | .500 | .190 | .071 | .032 | .016 | 2 | | .428 | .143 | .048 | .019 | .009 | .004 |
| 3 | | .667 | .286 | .125 | .056 | .028 | 3 | | .571 | .214 | .083 | .033 | .015 | .008 |
| 4 | | | .429 | .196 | .095 | .048 | 4 | | | .321 | .131 | .057 | .026 | .013 |
| 5 | | | .571 | .286 | .143 | .075 | 5 | | | .429 | .190 | .086 | .041 | .021 |
| 6 | | | | .393 | .206 | .111 | 6 | | | .571 | .274 | .129 | .063 | .032 |
| 7 | | | | .500 | .278 | .155 | 7 | | | | .357 | .176 | .089 | .047 |
| 8 | | | | .607 | .365 | .210 | 8 | | | | .452 | .238 | .123 | .066 |
| 9 | | | | | .452 | .274 | 9 | | | | .548 | .305 | .165 | .090 |
| 10 | | | | | .548 | .345 | 10 | | | | | .381 | .214 | .120 |
| 11 | | | | | | .421 | 11 | | | | | .457 | .268 | .155 |
| 12 | | | | | | .500 | 12 | | | | | .545 | .331 | .197 |
| 13 | | | | | | .579 | 13 | | | | | | .396 | .242 |
| | | | | | | | 14 | | | | | | .465 | .294 |
| | | | | | | | 15 | | | | | | .535 | .350 |
| | | | | | | | 16 | | | | | | | .409 |
| | | | | | | | 17 | | | | | | | .469 |
| | | | | | | | 18 | | | | | | | .531 |



Πίνακας ΙΧ: (συνέχεια)

| | | n=7 | | | | | | |
|----|---|------|------|------|------|------|------|------|
| U | m | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 0 | | .125 | .028 | .008 | .003 | .001 | .001 | .000 |
| 1 | | .250 | .056 | .017 | .006 | .003 | .001 | .001 |
| 2 | | .375 | .111 | .033 | .012 | .005 | .002 | .001 |
| 3 | | .500 | .167 | .058 | .021 | .009 | .004 | .002 |
| 4 | | .625 | .250 | .092 | .036 | .015 | .007 | .003 |
| 5 | | | .333 | .133 | .055 | .024 | .011 | .006 |
| 6 | | | .444 | .192 | .082 | .037 | .017 | .009 |
| 7 | | | .556 | .258 | .115 | .053 | .026 | .013 |
| 8 | | | | .333 | .158 | .074 | .037 | .019 |
| 9 | | | | .417 | .206 | .101 | .051 | .027 |
| 10 | | | | .500 | .264 | .134 | .069 | .036 |
| 11 | | | | .583 | .324 | .172 | .090 | .049 |
| 12 | | | | | .394 | .216 | .117 | .064 |
| 13 | | | | | .464 | .265 | .147 | .082 |
| 14 | | | | | .538 | .319 | .183 | .104 |
| 15 | | | | | | .378 | .223 | .130 |
| 16 | | | | | | .438 | .267 | .159 |
| 17 | | | | | | .500 | .314 | .191 |
| 18 | | | | | | .562 | .365 | .228 |
| 19 | | | | | | | .418 | .267 |
| 20 | | | | | | | .473 | .310 |
| 21 | | | | | | | .527 | .355 |
| 22 | | | | | | | | .402 |
| 23 | | | | | | | | .451 |
| 24 | | | | | | | | .500 |
| 25 | | | | | | | | .549 |



Πίνακας Χ: Κρίσιμη περιοχή για τον συντελεστή συσχέτισης r_s του Spearman.

| n | Τιμές r_α | |
|----|---------------------------------|-----------|
| | Επίπεδο σημαντικότητας α | |
| | για μονόπλευρο (δίπλευρο) τεστ | |
| | .05 (.01) | .01 (.02) |
| 4 | 1.000 | |
| 5 | .900 | 1.000 |
| 6 | .829 | .943 |
| 7 | .714 | .893 |
| 8 | .643 | .833 |
| 9 | .600 | .783 |
| 10 | .564 | .746 |
| 12 | .506 | .712 |
| 14 | .456 | .645 |
| 16 | .425 | .601 |
| 18 | .399 | .564 |
| 20 | .377 | .534 |
| 22 | .359 | .508 |
| 24 | .343 | .485 |
| 26 | .329 | .465 |
| 28 | .317 | .448 |
| 30 | .306 | .432 |



TABLE 2. THE EFFECT OF THE TYPE OF SOIL ON THE GROWTH OF PLANTS

| No. | Type of soil | | Height of plants (cm) | Weight of plants (g) |
|-----|--------------|-----|-----------------------|----------------------|
| | 1 | 2 | | |
| 1 | 100 | 100 | 100 | 100 |
| 2 | 100 | 100 | 100 | 100 |
| 3 | 100 | 100 | 100 | 100 |
| 4 | 100 | 100 | 100 | 100 |
| 5 | 100 | 100 | 100 | 100 |
| 6 | 100 | 100 | 100 | 100 |
| 7 | 100 | 100 | 100 | 100 |
| 8 | 100 | 100 | 100 | 100 |
| 9 | 100 | 100 | 100 | 100 |
| 10 | 100 | 100 | 100 | 100 |
| 11 | 100 | 100 | 100 | 100 |
| 12 | 100 | 100 | 100 | 100 |
| 13 | 100 | 100 | 100 | 100 |
| 14 | 100 | 100 | 100 | 100 |
| 15 | 100 | 100 | 100 | 100 |
| 16 | 100 | 100 | 100 | 100 |
| 17 | 100 | 100 | 100 | 100 |
| 18 | 100 | 100 | 100 | 100 |
| 19 | 100 | 100 | 100 | 100 |
| 20 | 100 | 100 | 100 | 100 |
| 21 | 100 | 100 | 100 | 100 |
| 22 | 100 | 100 | 100 | 100 |
| 23 | 100 | 100 | 100 | 100 |
| 24 | 100 | 100 | 100 | 100 |
| 25 | 100 | 100 | 100 | 100 |

ΒΙΒΛΙΟΓΡΑΦΙΑ

Α. ΕΛΛΗΝΙΚΗ

1. Αθανασόπουλος, Δ. (1975). *Εισαγωγή εις την Στατιστικήν και την Θεωρίαν των Πιθανοτήτων Ι*, Αθήνα.
2. Αναστασιάδης, Ι. (1978). *Στοιχεία Στατιστικής*, Θεσσαλονίκη.
3. Δρακάτος, Κ. (1979). *Στατιστική*, Τόμοι Ι, ΙΙ, ΙΙΙ και ΙV, Αθήνα.
4. Κακούλλος, Θ. (1972). *Στατιστική - Θεωρία και Εφαρμογαι*, Αθήνα.
5. Κεβόρκ, Κ. (1972). *Στατιστική*, Τόμοι Ι, ΙΙ, και ΙΙΙ, Αθήνα.
6. Κουνιάς, Σ., Κολύβα-Μαχαίρα, Φ., Μπαγιάτης, Κ. και Μπόρα-Σέντα, Ε. (1985). *Εισαγωγή στη Στατιστική*, Θεσσαλονίκη.
7. Λαμπράκης, Δ. (1980). *Στατιστική*, Δεύτερη έκδοση, Αθήνα.
8. Περισίδης, Σ.Κ. (1977). *Πιθανότητες και Στατιστική*, ΕΣΠΙ, Αθήνα, (Μετάφραση του Β. 27).
9. Ρούσσας, Γ. (1976). *Στατιστική Συμπερασματολογία*, Τόμοι 1 και 2, Πάτρα.



B. ΞΕΝΗ

1. Armitage, P. (1971). *Statistical Methods in Medical Research*. Blackwell Scientific Publications, Inc. Oxford.
2. Bhattacharyya, C. and Johnson, R. (1977). *Statistical Concepts and Methods*. John Wiley & sons, Inc., New York.
3. Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. Worth Wads Press, Boston.
4. Cochran, W.G. and Cox, G. (1957). *Experimental Designs, Second Edition*. John Wiley & sons, Inc., New York.
5. Conover, W.J. (1971). *Practical Nonparametric Statistics*. John Wiley & sons, Inc., New York.
6. Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis*. John Wiley & sons, Inc., New York.
7. Fisz, M. (1963). *Probability Theory and Mathematical Statistics, Third Edition*. John Wiley & sons, Inc., New York.
8. Freund, J. (1971). *Mathematical Statistics, Second Edition*. Prentice-Hall, Inc., New Jersey.
9. Gibbons, J.D. (1971). *Nonparametric Statistical Inference, International Student Edition*. McGraw-Hill Kogakusha Ltd., Tokyo-London.
10. Hennequin, P.L. et Tortat, A. (1965). *Theorie de Probabilités et Quelques Applications*. Masson et Cie, Paris.
11. Hogg, R. and Craig, A. (1970). *Introduction to Mathematical Statistics, Third Edition*. The Macmillan Co., New York.



12. Huntsberger, D.V. and Billingsley, P. (1973). *Elements of Statistical Inference, Third Edition*. Allyn and Bacon, Inc., Boston.
13. Kempthorne, O. and Folks, L. (1971). *Probability, Statistics and Data Analysis*. Iowa State University Press, Ames, Iowa.
14. Kendall, M.G. (1970). *Rank Correlation Methods, Fourth Edition*. Griffin, London.
15. Kendall, M.G. and Stuart, A. (1977). *The Advanced Theory of Statistics, Vol. 1*. Griffin, London.
16. Kolmogorov, A. (1941). "Confidence Limits for an Unknown distribution function". *Annals of Mathematical Statistics*, 12, 461-463.
17. Lindgren, B.W. (1976). *Statistical Theory, Third Edition*. Macmillan Publishing Co., Inc., New York.
18. Mendenhall, W. (1975). *Introduction to Probability and Statistics, Fourth Edition*. Duxbury Press, Massachusetts.
19. Mendenhall, W. and Scheaffer, R. (1973). *Mathematical Statistics with Applications*. Duxbury Press, Massachusetts.
20. Mood, A., Graybill, F. and Boes, D. (1974). *Introduction to the Theory of Statistics, Third Edition*. McGraw-Hill, Book Co., Inc., New York.
21. Neter, J. and Wasserman, W. (1975). *Applied Linear Statistical Models*. Richard D. Irwing, Inc., Illinois.
22. Pearson, K. (1900). "On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". *Philosophical Magazine*, 50, 157-175.



23. Roussas, G. (1973). *A First Course in Mathematical Statistics*. Addison-Wesley Publishing Co., Reading Massachussetts.
24. Sheffé, H. (1959). *The Analysis of Variance*. John Wiley & sons, Inc., New York.
25. Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Book Co., Inc., New York.
26. Smirnov, N.V. (1948). "Table of estimating the goodness of fit of empirical distributions". *Annals of Mathematical Statistics*, 19, 279-281.
27. Spiegel, M.R. (1972). *Statistics, Schaum's Outline Series*. McGraw-Hill Book Co., Inc., New York.
28. Wilks, S.S. (1962). *Mathematical Statistics*. John Wiley & sons, Inc., New York.



ΑΠΑΝΤΗΣΕΙΣ ΣΤΙΣ ΑΣΚΗΣΕΙΣ

ΚΕΦΑΛΑΙΟ 2

- 2.1 (i) $\bar{X}=4$, $M=4,3$ και 6 , $R=5$, $S^2=4$, $S=2$.
(ii) $\bar{X}=0$, $M=0$, $-4,0$ και 2 , $R=12$, $S^2=16$, $S=4$.
(iii) $\bar{X}=14$, $M=15$, καμμία κορυφή, $R=22$, $S^2=64$,
 $S=8$.
- 2.2 (i) Μήκος ομάδας 36 .
(ii) $\bar{X}=396.5$, $M=385.78$, $S=34.78$
(iii) $x_{0.25}=324$, $x_{0.50}=373$ $x_{0.75}=448$,
 $x_{0.10}=273$.

ΚΕΦΑΛΑΙΟ 3

- 3.1 (i) $\mu=4.5$, $\sigma^2=5/4$.
(ii) $(3,3),(3,4),(4,3)$ κ.λ.π. 16 δείγματα.
(iii) $E\bar{X}=4.5$, $Var(\bar{X})=5/8$.
(iv) $E(S^2)=5/4$.
- 3.2 (i) $N(0, (1-1/n)\sigma^2)$.
(ii) χ_1^2 .
(iii) $(n-1)\sigma^2/n$.
(iv) χ_n^2 .
- 3.3 $32/9$.
- 3.4 χ_n^2 .



314 ΑΠΑΝΤΗΣΕΙΣ

3.5 $\bar{X} - \bar{Y} \sim N(-1, 1)$ και $P(\bar{X} - \bar{Y} > 0) = 0.1587.$

3.6 $X_1 + \dots + X_{40} \sim N(36 \times 10^3, 10^5), 0.0571.$

3.7 (i) $\mu = 5, \sigma = \sqrt{10}.$
(ii) $P(|t_A| \leq c) = 0.90, c = 2.132.$

3.8 (i) $P(X_3^2 \leq 1) = 0.20.$
(ii) $F_{1,2}.$

3.9 (i) $N(0, 1).$
(ii) $F_{1,1}.$
(iii) $t_1.$
(iv) $F_{1,1}.$

3.10 (i) $N(0, n/[4k(n-k)]).$
(ii) $\chi_2^2.$
(iii) $F_{1,1}.$

3.11 $\mu = 50, \sigma^2 = 100, K0\theta: = 0.6826.$

3.12 $\mu = 0, \sigma^2 = 1/24, K0\theta: = 0.9954.$

3.13 $\mu = 1, \sigma^2 = 1/3, K0\theta: = 0.4514.$

3.14 $K0\theta: = 0.1587.$

3.15 $\mu = 0, \sigma^2 = 1/12, K0\theta: = 0.7698.$

3.16 $K0\theta: = 0.0162.$



- 3.17 (i) $P(80 \leq X < 120) \geq 0.75$.
 (ii) $K0\theta: \approx 0.9596$.
- 3.18 (i) $E(S_n) = \text{Var}(S_n) = n\theta$.
 (ii) $K0\theta: \approx 0.2145$.
- 3.19 (i) $X \sim B(10, 0.01): 0.09135$.
 (ii) $X \sim P(3): 0.6472$.
 (iii) $X \sim N(6, 5.94): 0.8435$.
- 3.20 $Y = \text{αριθ. τιμών} < 3$.
 $Y \sim B(72, 2/3), K0\theta: \approx 0.26595$.
- 3.21 (i) $X \sim B(10, 0.05) = 0.0105$.
 (ii) $X \sim NB(2, 0.05) = 0.01797$.
 (iii) $K0\theta: \approx 0.697$.
- 3.22 $X = \text{αριθ. επιβατών με εισητήρια που δεν εμφανίζονται}, X \sim B(160, 0.05), K0\theta: 0.2942$.

ΚΕΦΑΛΑΙΟ 4

- 4.2 $\bar{x}, s^2$.
- 4.3 $n_A = 500/20 = 25, n_B = 10, \text{Var}(\bar{X}_A) = 0.16$
 $\text{Var}(\bar{X}_B) = 0.3$, προτιμ. η Α.
- 4.4 $\mu = 5, \sigma^2 = 2/3, \sigma_{\bar{X}}^2 = 1/3$.
 3 από τα 9 Δ.Ε. περιέχουν το μ , έτσι $1 - \alpha = 1/3$.



314 ΑΠΑΝΤΗΣΕΙΣ

3.5 $\bar{X} - \bar{Y} \sim N(-1, 1)$ και $P(\bar{X} - \bar{Y} > 0) = 0.1587.$

3.6 $X_1 + \dots + X_{40} \sim N(36 \times 10^3, 10^5), 0.0571.$

3.7 (i) $\mu = 5, \sigma = \sqrt{10}.$
(ii) $P(|t_A| \leq c) = 0.90, c = 2.132.$

3.8 (i) $P(X_3^2 \leq 1) = 0.20.$
(ii) $F_{1,2}.$

3.9 (i) $N(0, 1).$
(ii) $F_{1,1}.$
(iii) $t_1.$
(iv) $F_{1,1}.$

3.10 (i) $N(0, n/[4k(n-k)]).$
(ii) $\chi_2^2.$
(iii) $F_{1,1}.$

3.11 $\mu = 50, \sigma^2 = 100, K0\theta: = 0.6826.$

3.12 $\mu = 0, \sigma^2 = 1/24, K0\theta: = 0.9954.$

3.13 $\mu = 1, \sigma^2 = 1/3, K0\theta: = 0.4514.$

3.14 $K0\theta: = 0.1587.$

3.15 $\mu = 0, \sigma^2 = 1/12, K0\theta: = 0.7698.$

3.16 $K0\theta: = 0.0162.$



- 3.17 (i) $P(80 \leq X < 120) \geq 0.75$.
 (ii) $K0\theta: \approx 0.9596$.
- 3.18 (i) $E(S_n) = \text{Var}(S_n) = n\theta$.
 (ii) $K0\theta: \approx 0.2145$.
- 3.19 (i) $X \sim B(10, 0.01): 0.09135$.
 (ii) $X \sim P(3): 0.6472$.
 (iii) $X \sim N(6, 5.94): 0.8435$.
- 3.20 $Y = \text{αριθ. τιμών} < 3$.
 $Y \sim B(72, 2/3), K0\theta: \approx 0.26595$.
- 3.21 (i) $X \sim B(10, 0.05) = 0.0105$.
 (ii) $X \sim NB(2, 0.05) = 0.01797$.
 (iii) $K0\theta: \approx 0.697$.
- 3.22 $X = \text{αριθ. επιβατών με εισιτήρια που δεν εμφανίζονται}, X \sim B(160, 0.05), K0\theta: 0.2942$.

ΚΕΦΑΛΑΙΟ 4

- 4.2 $\bar{X}, S^2$.
- 4.3 $n_A = 500/20 = 25, n_B = 10, \text{Var}(\bar{X}_A) = 0.16$
 $\text{Var}(\bar{X}_B) = 0.3$, προτιμ. η Α.
- 4.4 $\mu = 5, \sigma^2 = 2/3, \sigma_{\bar{X}}^2 = 1/3$.
 3 από τα 9 Δ.Ε. περιέχουν το μ , έτσι $1 - \alpha = 1/3$.



316 ΑΠΑΝΤΗΣΕΙΣ

- 4.5 (1) 10.156, 0.367 .
 (11) (9.348, 10.964) .
 (111) (0.9015, 4.6684) .
- 4.6 (17.640, 62.3597).
- 4.7 $\alpha=0.0013$, $\beta=0.0668$, $\gamma=0.9332$.
- 4.8 $\alpha=0.013$, $\beta=0.1587$.
- 4.9 97.
- 4.10 (1) (99.457, 112.143).
 (11) $t=2.068 < t_{0.025, 9}=2.262$, δεν απορρίπτεται η H_0 .
- 4.11 $H_0: \mu=9.75$, $H_a: \mu \neq 9.75$
 $t=1.239 < t_{0.025, 5}=2.571$, δεν απορρίπτεται η H_0 .
- 4.12 2017 περίπου.
- 4.13 (0.8979, 0.9421).
- 4.14 (-1.521, 0.0213).
- 4.15 $H_0: \mu_1=\mu_2$, $H_a: \mu_1 \neq \mu_2$,
 $z=7.018 > 1.96$, απορρίπτεται η H_0 .
- 4.16 (1) $H_0: \mu_1=\mu_2$, $H_a: \mu_1 \neq \mu_2$,
 $z=-7.042 > -1.96$, απορρίπτεται η H_0 ,
 (11) $1-\Phi(1.96-6/\sqrt{0.000244})+$
 $\Phi(-1.96-6/\sqrt{0.000244})$.
- 4.17 $H_0: \mu_1=\mu_2$, $H_a: \mu_1 \neq \mu_2$,
 $t=1.651 < t_{0.025, 16}=2.12$, δεν απορρίπτεται η H_0



- 4.18 $H_0: \mu_1 = \mu_2, H_a: \mu_1 \neq \mu_2$,
 $t = 2.339 < t_{0.025, 6} = 2.447$ δεν το μεταβάλλει.
- 4.19 $H_0: \mu_{\text{πρόσ}} \leq \mu_{\text{μετά}}, H_a: \mu_{\text{πρόσ}} > \mu_{\text{μετά}},$
 $t = 4.129 > t_{0.01, 8} = 2.896$, ναι.
- 4.20 $H_0: \mu_B \leq \mu_A, H_a: \mu_B > \mu_A$
 $t = 3.592 > t_{0.01, 8} = 2.896$, το B δίνει μεγαλύτερες
 ενδείξεις.
- 4.21 $H_0: \sigma_A^2 = \sigma_B^2, H_a: \sigma_A^2 \neq \sigma_B^2$,
 $F = 4.464 > F_{0.05, 9, 7} = 3.68$ υπάρχει.
- 4.22 $H_0: \sigma_A^2 = \sigma_B^2, H_a: \sigma_A^2 \neq \sigma_B^2$,
 $F_{0.95, 15, 15} = 0.4166 < F = 0.5203 < F_{0.05, 15, 15} = 2.40$,
 δεν υπάρχει διαφορά.
- 4.23 $H_0: p_A = p_B, H_a: p_A \neq p_B$,
 $z = 0.649 < z_{0.025} = 1.96$, όχι.
- 4.24 $H_0: p_1 = p_2, H_a: p_1 \neq p_2$,
 $z = 1.155 < z_{0.025} = 1.96$, όχι.
- 4.25 $H_0: p_1 = p_2, H_a: p_1 < p_2$,
 $z = -1.3652 > -z_{0.05} = -1.645$, ίδια ανθεκτικότητα.
- 4.26 $H_0: p_A = p, H_a: p_A \neq p$
 $z = 2.062 > z = 1.96$, υπάρχει διαφορά.



ΚΕΦΑΛΑΙΟ 5

$$5.1 \quad - \frac{\bar{x}\sigma^2}{\sum(x_1 - \bar{x})^2}, \quad 0.$$

$$5.2 \quad \hat{\beta}_0 = \bar{y}, \quad \hat{\beta}_1 = \frac{\sum(x_1 - \bar{x})(y_1 - \bar{y})}{\sum(x_1 - \bar{x})^2}.$$

- 5.3 (i) $\hat{y} = 9.27 + 1.44x$.
 (ii) $F = 96.17 > F_{0.05, 1, 19} = 5.12$, απορρίπτεται η H_0 .
 (iii) [1.11, 1.77].
 (iv) [12.15, 15.03].
 (v) όχι.

- 5.4 (i) $\hat{y} = 4.15 + 0.095x$.
 (ii) $F = 41.65 > F_{0.05, 1, 5} = 6.01$, απορρίπτεται η H_0 .
 (iii) [0.058, 0.134].

- 5.5 (i) $\hat{y} = 343.71 + 3.22x$.
 (ii) $F = 6.75 > F_{0.05, 1, 10} = 4.96$, απορρίπτεται η H_0 .

- 5.6 (i) $\hat{y} = -5.96 + 1.55x$.
 (ii) 5.

- 5.7 (i) $\hat{y} = -21.94 + 0.44x$.
 (ii) $F = 34.05 > F_{0.05, 1, 8} = 5.32$, απορρίπτεται η H_0 .
 (iii) [21.98, 34.36].

- 5.8 (i) $\hat{y} = 36 + 13x$.

- 5.9 $t = 5.895 > t_{0.005, 31} = 2.576$, απορρίπτεται η H_0 .

- 5.10 $t = 4.22 > t_{0.025, 10} = 2.23$, απορρίπτεται η H_0 .



ΚΕΦΑΛΑΙΟ 6

- 6.3 $F=8.87 > F_{0.05,3,19}=3.13$, απορρίπτεται η H_0 .
Καλύτερες Δ_1, Δ_4 .
- 6.4 $F=7.72 > F_{0.05,2,8}=4.46$, απορρίπτεται η H_0 .
- 6.5 $F=0.68 < F_{0.05,3,16}=3.24$, δεν απορρίπτεται η H_0 .
- 6.6 $F=9.08 > F_{0.05,3,12}=3.49$, απορρίπτεται η H_0 .
- 6.7 $F_A=0.01 < F_{0.05,3,6}=4.76$, δεν απορρίπτεται η H_0 .
- 6.8 $F_\Sigma=2.84 < F_{0.05,3,6}=4.76$, δεν απορρίπτεται η H_0 .
 $F_A=14.13 > F_{0.05,2,6}=5.14$, απορρίπτεται η H_0 .
- 6.9 $F_A=34.85 > F_{0.05,2,6}=5.14$, απορρίπτεται η H_0 .
 $F_B=24.62 > F_{0.05,3,6}=4.76$, απορρίπτεται η H_0 .
- 6.10 $F_A=36.53 > F_{0.05,2,6}=5.14$, απορρίπτεται η H_0 .
 $F_\Sigma=15.47 > F_{0.05,3,6}=4.76$, απορρίπτεται η H_0 .

ΚΕΦΑΛΑΙΟ 7

- 7.1 (i) $\chi^2=5.625 < \chi^2_{3,0.05}=7.815$, δεν απορρίπτεται η H_0 .
(ii) $D_{10}=0.194 < D_{0.05,10}=0.409$, δεν απορρίπτεται η H_0 .
- 7.2 (i) $z=2.066 > z_{0.025}=1.96$, απορρίπτεται η H_0 .



320 ΑΠΑΝΤΗΣΕΙΣ

- 7.3 (I) $z=1.49 < z_{0.05}=1.65$, δεν απορρίπτεται η H_0 .
 (II) $T'=20 > T_{0.025,12}=14$, δεν απορρίπτεται η H_0 .

- 7.4 (I) $\chi^2=5.692 < \chi_{9,0.05}^2=18.307$, δεν απορρίπτεται η H_0 .
 (II) $D=0.2 < D_{0.05}=0.6$, δεν απορρίπτεται η H_0 .

- 7.5 $|z|=1.33 < z_{0.025}=1.96$, δεν απορρίπτεται η H_0 .

- 7.6 $r_s=0.98 > r_{0.02}=0.746$, απορρίπτεται η H_0 .



ΕΥΡΕΤΗΡΙΟ

- Αξιώματα του Kolmogorov 4
Ακραία τιμή 196, 241
Αλληλοεπίδραση 249
Αμερόληπτος 96
Αμερόληπτος ομοιόμορφα ελάχιστης διακύμανσης 97
Αμεροληψία 96
Ανάλυση της διακύμανσης 215, 216
Ανάλυση των υπολοίπων 190, 239
Αναμενόμενη τιμή 19
Ανισότητα του Chebushen 20
Αντίστροφα εκατοσπαία σημεία 68
Αντίστροφος μετασχηματισμός 198
Απλή γραμμική παλινδρόμηση 169
Αριθμητικά μεγέθη 40
Αριθμητικός μέσος όρος 41
Αριθμητικά χαρακτηριστικά του δείγματος 27
Αρνητική διωνυμική κατανομή 15, 21
Αθροιστική συνάρτηση κατανομής 8
Αθροιστική συχνότητα 27, 28
Αυστηρά ελεγχόμενο πείραμα 245
Βήτα κατανομή 17
Βαθμός ελευθερίας 49
Γάμμα κατανομή 17
Γεωμετρική κατανομή 15, 21
Γραφήματα 27
Γραφικές μέθοδοι σύμπτυξης 35
Γραμμική συνάρτηση παλινδρόμησης 169
Δείγμα 11
Δείκτρια μεταβλητή 205
Δειγματικές κατανομές 59
Δειγματική: διάμεσος 42· διασπορά 47· κορυφή 45· μέση τιμή 41· τυπική απόκλιση 50
Δειγματικό εύρος 47
Δειγματικός: συντελεστής κυρτότητας 53· συντελεστής λοξότητας 53· συντελεστής συσχέτισης 202· χώρος 4
Δεσμευμένη πιθανότητα 5
Διάγραμμα διασποράς 165
Διάστημα εμπιστοσύνης 99
Διόρθωση συνεχείας 83
Διακύμανση 19
Διατάξιμη κλίμακα 269
Διατεταγμένος πίνακας συχνοτήτων 30
Διωνυμική κατανομή 14, 20
Διωνυμικό πείραμα 14
Δοκιμασία 220
Εκατοσπαίο σημείο 44
Εκτιμήτρια συνάρτηση 95
Εκτιμώμενη συνάρτηση παλινδρόμησης 174, 175
Εκτιμητές ελαχίστων τετραγώνων 173
Εκτιμητής 95, 97
Εκτιμητική με διαστήματα εμπιστοσύνης 95



- Εκθετική κατανομή 17, 21
 Έλεγχος στατιστικών υποθέσεων 104
 Ενδεχόμενα 4
 Επίπεδο εμπιστοσύνης 99
 Επίπεδο σημαντικότητας στατιστικού
 τεστ 107
 Επίπεδο παράγοντα 219
 F τεστ 143, 188
 Θεώρημα: Cochran 234· Gauss-
 Markov 179
 Ιδιότητες των εκτιμητών 178
 Ιστογράμματα 29, 35
 Ισχύς του τεστ 107
 Καμπύλη παλινδρόμησης 168
 Κανόνας του Bayes 6
 Κανονικές εξισώσεις 172
 Κανονική: κατανομή 18, 21·
 προσέγγιση της διωνυμικής 81
 Κανονικοποιημένα υπόλοιπα 190
 Κατανομή: Βήτα 21· Γάμμα 21· F_{v_1, v_2}
 6 6· κανονική 18, 21·
 πιθανότητας 7· Poisson 16, 21·
 συχνοτήτων 32· t_n 64· χ^2 62
 Κατηγορικές μεταβλητές 10
 Κεντρικό οριακό θεώρημα 74, 75
 Κλίμακα λόγου 270
 Κλίμακες μετρήσεων 268
 Κλαδεμένη μέση τιμή 45
 Κορυφή 22
 Κρίσιμη περιοχή 104
 Λογαριθμικός μετασχηματισμός 197
 Μέσα τετράγωνα 256
 Μέση απόλυτη απόκλιση 49
 Μέση τετραγωνική απόκλιση 48
 Μέση τιμή 19
 Μέση τιμή του Winsor 45
 Μέσο εύρος 46
 Μέσο τετράγωνο 187, 231
 Μέσο τετραγωνικό σφάλμα 179
 Μέτρα: θέσης 40· μεταβλητότητας 47
 Μέθοδος: ελάχιστης σημαντικής
 διαφοράς 235, 259· ελαχίστων
 τετραγώνων 171, 226, 251·
 εύρους 238, 260
 Μη παραμετρικές μέθοδοι 267
 Μη παραμετρική στατιστική 267
 Μοντέλα: ανάλυσης διακύμανσης 206,
 219· ανάλυσης συνδιακύμανσης
 206· παλινδρόμησης 166, 169
 Ομαδοποιημένες μετρήσεις 31
 Ομοιόμορφη κατανομή 16, 21
 Ονοματική κλίμακα 268
 Πίνακας ανάλυσης διακύμανσης 231
 Πίνακες: στατιστικών δεδομένων 27·
 συχνοτήτων 29, 30
 Παλινδρόμηση 161, 216
 Παράγοντας 219
 Παράμετροι 58
 Περιγραφική στατιστική 2, 27
 Πιθανότητας 4
 Πλήρες σχέδιο τετραγώνων 248
 Πλήρες τυχαίοποιημένο πείραμα 246
 Πλήρως τυχαίοποιημένο σχέδιο 126,
 148
 Πληθυσμός 11
 Ποιοτικές μεταβλητές 10



- Πολύγωνο συχνοτήτων 38
 Πολλαπλές συγκρίσεις 235
 Πολλαπλή γραμμική παλινδρόμηση 200
 Ποσοστιαία σημεία 22
 Ποσοτικές μεταβλητές 10
 Προσημικό τεστ 274, 277
 Προσημικό τεστ των τάξεων 279
 Ροπές του δείγματος 52
 Ροπογεννήτρια συνάρτηση 23
 Σύμπτυξη των αριθμητικών δεδομένων 29
 Σφάλμα τύπου I 105
 Σφάλμα τύπου II 106
 Σημειοεκτιμητική 95
 Στατιστικά πακέττα 206, 207, 242, 261
 Στατιστική: συμπερασματολογία 2· συνάρτηση 58· σχέση 163
 Στατιστικό: 58· τεστ 104
 Στοχαστική διαδικασία του Poisson 16
 Σχέδια τυχαιοποιημένων τετραγώνων 135, 137
 Σχετική συχνότητα 27
 Συνάρτηση: παλινδρόμησης 168, 170· πιθανότητας 7· πυκνότητας 8
 Συναρτησιακή σχέση 162
 Συνδιακύμανση 201
 Συνεχής κλίμακα 269
 Συντελεστές παλινδρόμησης 170
 Συντελεστής: κύρτωσης 19· λοξότητας 19· μεταβλητότητας 50· προσδιορισμού 186· συσχέτισης 201, 289
 Συσχέτιση 201
 Συχνότητα: 27· ομάδας 32
 t τεστ 116, 189
 Τεστ: Kolmogorov - Smirnov 272, 284· Mc Nemar 146· σημαντικότητας 112· Wilcoxon - Mann - Whitney 286· χ^2 141
 Τιμή P 114
 Τυπική απόκλιση 20
 Τυπική κανονική κατανομή 18
 Τυπικό σφάλμα του εκτιμητή 97
 Τυχαία μεταβλητή 7
 Τυχαία πειράματα 3
 Τυχαίο δείγμα 12
 Τυχαίο σφάλμα 170, 222, 251
 Τυχαιοποιημένο σχέδιο τετραγώνων 248
 Υπόλοιπα 175, 227, 253
 Υπεργεωμετρική κατανομή 15
 Υπεργεωμετρική κατανομή 20
 χ^2 τεστ: ανεξαρτησίας 281· καλής προσαρμογής 271
 χ^2 κατανομή 62





Τυπώθηκε στο Πανεπιστημιακό Τυπογραφείο με δαπάνη
του Πανεπιστημίου Ιωαννίνων

Κ.Α. Πανεπιστημιακού Τυπογραφείου.

Copyright : Πανεπιστήμιο Ιωαννίνων

Απαγορεύεται η μερική ή ολική ανατύπωση, καθώς και η
λήψη φωτοαντιγράφων από το βιβλίο χωρίς τη γραπτή
άδεια του Τμήματος Δημοσιευμάτων του Πανεπιστημίου
Ιωαννίνων και του συγγραφέα.

Διατίθεται και στο Βιβλιοπωλείο του Πανεπιστημίου
Ιωαννίνων, Δομπόλη, 451 10 Ιωάννινα τηλ. 21801.

ΔΙΑΝΕΜΕΤΑΙ ΔΩΡΕΑΝ στους φοιτητές.



140
 ΤΑΤΑΙΩΝΑ, Τ.
 Εργαστήριον
 102/95
 ΔΑΝΕΙΣΜΕΝΟΣ ΗΜΕΡΟΜΗΝΙΑ

| | | |
|--------|---------|---|
| 80/3 | 27/3 | 8 |
| 1/3/01 | 15/3 | 2 |
| | 11/3/02 | |

