ΤΕΙ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ / ΜΕΣΟΛΟΓΓΙ

Πτυχιακή εργασία

Ο ρόλος της Στατιστικής στις Επιχειρήσεις

Μαντάς Μιλτιάδης

Μεσολόγγι 2015

ΤΕΙ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ / ΜΕΣΟΛΟΓΓΙ

Πτυχιακή εργασία

Ο ρόλος της Στατιστικής στις Επιχειρήσεις

Μαντάς Μιλτιάδης

Επιβλέπων καθηγητής

Αθανάσιος Μεγαρίτης

Μεσολόγγι 2015

Μεσολόγγι 2015

ΠΡΟΛΟΓΟΣ

Κατά τη διάρκεια της φοίτησής μου στο Τ.Ε.Ι. Μεσολογγίου στο τμήμα Πρώην Ε.Π.Δ.Ο. διδάχτηκα πολλά μαθήματα τα οποία είχαν σχέση με τα οικονομικά, τα μαθηματικά και τη στατιστική.

Αυτό μου γέννησε την επιθυμία να εμβαθύνω λίγο στη μαθηματική έννοια της στατιστικής και στη χρήση της στις επιχειρήσεις, βάζοντας ένα μικρό λιθαράκι στις γνώσεις μου γύρω από αυτό το πολύ μεγάλο αντικείμενο. Οι επιχειρήσεις πλέον (και στη χώρα μας) έχουν εκσυγχρονιστεί, χρησιμοποιούν το διαδίκτυο, συλλέγουν πληροφορίες και κάνουν στατιστικές αναλύσεις σχετικά με την πορεία τους. Ο Έλληνας επιχειρηματίας έχει αρχίσει να απομακρύνεται πια από τις λύσεις που βασίζονται στη διαίσθηση και προσανατολίζεται προς τα νέα συστήματα και τις επιστημονικές μεθόδους που παρέχει η επιστήμη της στατιστικής. Έτσι, θέλησα να μπω λίγο στη λογική αυτών των διαδικασιών.

Κατά τη διάρκεια της συγγραφής αυτής της πτυχιακής αντιμετώπισα διάφορες δυσκολίες οι οποίες είχαν να κάνουν περισσότερο με την εύρεση και τη χρήση του προγράμματος SPSS. Τελικά άντλησα πληροφορίες από το διαδίκτυο και τον επιβλέποντα καθηγητή.

Τα παραδείγματα τα οποία χρησιμοποίησα προσπάθησα να είναι κατανοητά σε οποιονδήποτε θα διάβαζε αυτήν την εργασία.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω για την πολύτιμη βοήθειά του τον καθηγητή μου κύριο Μεγαρίτη Αθανάσιο ο οποίος με τις γνώσεις του και την εμπειρία του με καθοδήγησε για τη συγγραφή αυτής της εργασίας.

ΠΕΡΙΛΗΨΗ

Η πτυχιακή εργασία που ακολουθεί αφορά την εφαρμογή της στατιστικής επιστήμης στις επιχειρήσεις. Ξεκινά από μια ιστορική αναδρομή, από τη στιγμή που εμφανίστηκε πρώτη φορά η στατιστική σαν επιστήμη, συνεχίζει με τους θεμελιωτές της που την καταξίωσαν και τις βασικές έννοιες που χρησιμοποιούμε και τα στατιστικά δεδομένα.

Τα στατιστικά μέτρα που χρησιμοποιούνται συγκρίνονται μεταξύ τους ώστε να βγουν κάποια συμπεράσματα. Δύο πολύ βασικές έννοιες της στατιστικής είναι της συσχέτισης και της παλινδρόμησης, οι οποίες μέσα από πίνακες και παραδείγματα δείχνουν τη μεγάλη αξία τους όσον αφορά την ύπαρξη της στατιστικής.

Η χρήση, η θέση και η συμβολή της στατιστικής είναι πολύ σημαντική για την εξέλιξη της επιχείρησης, την απόδοσή της, την αύξηση του κεφαλαίου της, την ανοδική της πορεία και τη σταθεροποίησή της στην αγορά εργασίας.

Όλα αυτά γίνονται κατανοητά μέσα από παραδείγματα, διαγράμματα, πίνακες και σαφείς ορισμούς έτσι ώστε να δοθεί στη στατιστική ο ρόλος που της αξίζει.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Πρόλογος I

Περίληψη II

Εισαγωγή IX

Κεφάλαιο 1 Στατιστική

1.1 Ιστορική αναδρομή της Στατιστικής 1

1.2 Βασικές έννοιες της Στατιστικής 4

1.2.1 Πληθυσμός – μεταβλητές 4

1.2.2 Συλλογή Στατιστικών δεδομένων 6

1.3 Παρουσίαση στατιστικών δεδομένων 7

1.3.1 Στατιστικοί πίνακες 7

1.3.2 Πίνακες κατανομής συχνοτήτων 9

1.3.3 Γραφική παράσταση κατανομής συχνοτήτων 13

1.4 Ομαδοποίηση στατιστικών δεδομένων 21

1.4.1 Ομαδοποίηση των παρατηρήσεων 21

Κεφάλαιο 2 Στατιστικά μέτρα

2.1 Μέτρα θέσης ή κεντρικής τάσης 31

2.1.1 Μέτρα θέσης και διασποράς 31

2.1.2 Μέση τιμή 32

2.1.3 Διάμεσος 38

2.1.4 Εκατοστημόρια – Τεταρτημόρια 42

2.1.5 Επικρατούσα τιμή 44

2.1.6 Σύγκριση μέτρων θέσης 47

2.2 Μέτρα διασποράς 48

2.2.1 Εύρος 48

2.2.2 Διακύμανση 49

2.2.3 Τυπική απόκλιση 52

2.2.4 Ενδοτεταρτημοριακό εύρος 56

2.2.5 Συντελεστής μεταβλητότητας 57

2.2.6 Σύγκριση μέτρων διασποράς 63

Κεφάλαιο 3 Συσχέτιση – Παλινδρόμηση

3.1 Συσχέτιση 65

3.2 Διάγραμμα διασποράς 66

3.3 Συντελεστής συσχέτισης του Pearson 68

3.4 Απλή γραμμική παλινδρόμηση 72

3.4.1 Το υπόδειγμα της απλής γραμμικής παλινδρόμησης 72

3.4.2 Η δειγματική εξίσωση της παλινδρόμησης 73

3.5 Συντελεστής προσδιορισμού 84

Κεφάλαιο 4 Στατιστική και επιχειρήσεις

4.1 Ο ρόλος της Στατιστικής στην επιχείρηση 89

4.1.1 Η σύγχρονη διοικητική των επιχειρήσεων (management) 89

4.1.2 Έρευνες αγορών – Διαφήμιση 90

4.1.3 Λήψη επιχειρηματικών αποφάσεων 91

4.1.4 Ποιότητα διαχείρισης – Αποτελεσματική διοίκηση 92

4.1.5 Η θέση της Στατιστικής στον ενοποιημένο οικονομικό χώρο 93

4.1.6 Η χρησιμοποίηση από τις επιχειρήσεις στατιστικών πινάκων και διαγραμμάτων 93

4.1.7 Η σπουδαιότητα των μέτρων θέσης στην επιχειρηματική δραστηριότητα 94

4.1.8 Η σπουδαιότητα των μέτρων διασποράς για τις επιχειρήσεις 94

4.1.9 Η σημασία της παλινδρόμησης και της συσχέτισης στη σύγχρονη επιχείρηση 95

4.2 Εφαρμογές της Στατιστικής στις επιχειρήσεις 96

Συμπεράσματα 100

Βιβλιογραφία 101

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 1.1 Παραδείγματα για την καλύτερη κατανόηση των εννοιών 5

Πίνακας 1.2 Στοιχεία ανά κλάδο, βιομηχανικών επιχειρήσεων στην Ελλάδα με απασχόληση πάνω από 10 άτομα, το έτος 2010 8

Πίνακας 1.3 Κατανομή του πληθυσμού της Ελλάδας σε γεωγραφικά διαμερίσματα, κατά φύλο και περιφέρεια, σύμφωνα με την απογραφή του 2011 8

Πίνακας 1.4 Πίνακας συχνοτήτων της μεταβλητής: «αριθμός αδελφών» 10

Πίνακας 1.5 Σχολικός πληθυσμός της Ελλάδας ανά βαθμίδα εκπαίδευσης 14

Πίνακας 1.6 Ποσοστιαία κατανομή κατά περιοχές του πληθυσμού της Γης το 1970 15

Πίνακας 1.7 Κατανάλωση ενέργειας το έτος 1991 Πίνακας 1.8 Κέρδη μιας επιχείρησης σε χιλιάδες ευρώ τα έτη 2008-2013 16

Πίνακας 1.8 Κέρδη μιας επιχείρησης σε χιλιάδες ευρώ τα έτη 2008-2013 20

Πίνακας 1.9 Κατανομές συχνοτήτων για τη μεταβλητή Χ: «τιμές πώλησης σε € ενός προϊόντος» 23

Πίνακας 3.1 Ποσοστό ατόμων που ζουν σε αστικά κέντρα και ποσοστό ατόμων που γνωρίζουν να διαβάζουν σε ένα τυχαίο δείγμα χωρών του ΟΗΕ 66

Πίνακας 3.2 Τιμές ημερήσιας ενεργειακής πρόσληψης 40 ενηλίκων και ηλικίες αυτών 76

Πίνακας 4.1 Επενδύσεις και μακροχρόνιο επιτόκιο της ελληνικής οικονομίας 96

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ

Εικόνα 1.1 Ραβδόγραμμα συχνοτήτων των δεδομένων του Πίνακα 1.5 14

Εικόνα 1.2 Ραβδόγραμμα σχετικών συχνοτήτων των δεδομένων του Πίνακα 1.6 15

Εικόνα 1.3 Κυκλικό διάγραμμα σχετικών συχνοτήτων του Πίνακα 1.7 17

Εικόνα 1.4 Διάγραμμα συχνοτήτων και πολύγωνο συχνοτήτων των δεδομένων του παραδείγματος 1.5 18

Εικόνα 1.5 Σημειόγραμμα των δεδομένων του παραδείγματος 1.6 19

Εικόνα 1.6 Χρονόγραμμα των κερδών της επιχείρησης του Πίνακα 1.8 20

Εικόνα 1.7 Ιστόγραμμα συχνοτήτων 24

Εικόνα 1.8 Ιστόγραμμα σχ. Συχνοτήτων 24

Εικόνα 1.9 Ιστόγραμμα συχνοτήτων για τη μεταβλητή Χ: «τιμές πώλησης σε € ενός προϊόντος» 26

Εικόνα 1.10 Πολύγωνο συχνοτήτων για τη μεταβλητή Χ: «τιμές πώλησης σε € ενός προϊόντος» 26

Εικόνα 1.11 Ιστόγραμμα σχετικών συχνοτήτων για τη μεταβλητή Χ: «τιμές πώλησης σε € ενός προϊόντος» 27

Εικόνα 1.12 Πολύγωνο σχετικών συχνοτήτων για τη μεταβλητή Χ: «τιμές πώλησης σε € ενός προϊόντος» 27

Εικόνα 1.13 Ιστόγραμμα αθροιστικών συχνοτήτων για τη μεταβλητή Χ: «τιμές πώλησης σε € ενός προϊόντος» 28

Εικόνα 1.14 Πολύγωνο αθροιστικών συχνοτήτων για τη μεταβλητή Χ: 28

«τιμές πώλησης σε € ενός προϊόντος»

Εικόνα 1.15 Ιστόγραμμα αθροιστικών σχετικών συχνοτήτων για τη μεταβλητή Χ: «τιμές πώλησης σε € ενός προϊόντος» 29

Εικόνα 1.16 Πολύγωνο αθροιστικών σχετικών συχνοτήτων για τη μεταβλητή Χ: «τιμές πώλησης σε € ενός προϊόντος» 29

Εικόνα 3.1 Διάγραμμα διασποράς των δεδομένων του Πίνακα 3.1 67

Εικόνα 3.2-3.2-3.3-3.4 Παραδείγματα πλήρους γραμμικής συσχέτισης 69-70

Εικόνα 3.6 Διάγραμμα διασποράς 74

Εικόνα 3.7 Διάγραμμα διασποράς της ηλικίας και της ημερήσιας ενεργειακής πρόσληψης 40 ενηλίκων 77

Εικόνα 3.8Ευθεία ελαχίστων τετραγώνων για τη σχέση ενεργειακής πρόσληψης με την ηλικία 79

Εικόνα 4.1 Διάγραμμα διασποράς των δεδομένων του Πίνακα 4.1 97

ΕΙΣΑΓΩΓΗ

Η παρούσα πτυχιακή εργασία έχει ως θέμα την Εφαρμογή της Στατιστικής στις Επιχειρήσεις.

Το περιεχόμενο της εργασίας χωρίζεται σε 4 κεφάλαια. Το 1ο κεφάλαιο έχει να κάνει με την ιστορική αναδρομή της στατιστικής, από την προέλευση της λέξης «στατιστική» μέχρι τη χρήση της στην πορεία του χρόνου από διάφορους λαούς, την αναφορά στους θεμελιωτές της, τις βασικές έννοιες και την παρουσίαση στατιστικών δεδομένων (στατιστικοί πίνακες, γραφικές αναπαραστάσεις, μεταβλητές).

Στο 2ο κεφάλαιο γίνεται η ανάλυση των στατιστικών μέτρων (π.χ. μέση τιμή, εύρος, διάμεσος) και η σύγκρισή μεταξύ τους με τη χρήση διαφόρων διαγραμμάτων και ορισμών.

Το 3ο κεφάλαιο ασχολείται με τις έννοιες της συσχέτισης και της παλινδρόμησης δοσμένες μέσα από παραδείγματα, ορισμούς και σχεδιαγράμματα.

Τέλος, στο 4ο κεφάλαιο παρουσιάζεται η συμβολή της στατιστικής στη σύγχρονη επιχείρηση, πως η χρήση της μπορεί να βοηθήσει τον επιχειρηματία να κάνει έρευνα αγοράς, να πάρει επιχειρηματικές αποφάσεις, να έχει σωστή ποιότητα στη διαχείριση και αποτελεσματική διοίκηση.

Η χρησιμοποίηση της στατιστικής και η θέση της δίνονται μέσα από διάφορα παραδείγματα, διαγράμματα, πίνακες και ορισμούς, έτσι ώστε να γίνει κατανοητή η σημασία της χρήσης της στατιστικής από τις επιχειρήσεις.

Σκοπός της πτυχιακής εργασίας ήταν να εισάγει τον αναγνώστη στην έννοια της στατιστικής και στη χρήση της στις επιχειρήσεις δίνοντάς του να καταλάβει τη μεγάλη σημασία της στην ανάπτυξη και εξέλιξη μιας επιχείρησης.

Κεφάλαιο 1 Στατιστική

1.1 Ιστορική αναδρομή της Στατιστικής

Ο όρος «Στατιστική» κατ' άλλους προέρχεται από τη λατινική λέξη «status» που σημαίνει κράτος-κατάσταση και δήλωνε, από τότε που πρωτοχρησιμοποιήθηκε συλλογή στοιχείων για τις κρατικές ανάγκες ενώ κατ' άλλους προέρχεται από την ελληνική λέξη «στατίζω» που σημαίνει τοποθετώ, διαπιστώνω, προσδιορίζω, δημιουργώ.

Η στοιχειώδης συλλογή στατιστικών στοιχείων είναι αρκετά παλιά. Η αρχαιότερη ίσως γνωστή απογραφή πληθυσμού είναι αυτή που έγινε το 2238 π.Χ. στην Κίνα από τον αυτοκράτορα Yao. Στην αρχαιότητα, η συγκέντρωση στατιστικών στοιχείων είχε στόχο τον εντοπισμό των πολιτών που ήταν υποχρεωμένοι να υπηρετήσουν σαν πολεμιστές ή να υποβληθούν σε φορολογία. Μετά το Μεσαίωνα, οι κυβερνήσεις της Δυτικής Ευρώπης ενδιαφερόντουσαν για στατιστικά στοιχεία λόγω του φόβου των επιδημιών και της αντίληψης ότι το μέγεθος του πληθυσμού επηρέαζε σημαντικά την πολιτική και στρατιωτική τους δύναμη. Τέλος από τον 16ο μέχρι τον 18ο αιώνα, η ραγδαία ανάπτυξη του εμπορίου ώθησε τις πολιτειακές αρχές στη μελέτη οικονομικών δεδομένων όπως εξαγωγικό εμπόριο, πλήθος και δυναμικότητα βιομηχανιών, κ.λπ. Δεν είναι μάλιστα τυχαίο ότι πολλά από τα στοιχεία που συγκεντρώθηκαν αυτή την περίοδο θεωρήθηκαν κρατικά μυστικά.

Αξίζει να αναφερθεί ότι ο πρώτος Κυβερνήτης της Ελλάδας, Ι. Καποδίστριας, είχε ενδιαφερθεί σοβαρότατα για τη δημιουργία στατιστικής υπηρεσίας στην Ελλάδα διαβλέποντας τον κρίσιμο ρόλο που θα έπαιζε στη δημιουργία του νέου κράτους.

Έτσι αρκετά νωρίς για την ελληνική και διεθνή πραγματικότητα, το 1833, ιδρύθηκε η Υπηρεσία Γενικής Στατιστικής του Κράτους, η οποία τέθηκε στη δικαιοδοσία του Υπουργείου των Εσωτερικών. Με το Κανονιστικό Διάταγμα (Κ.Δ.) της 29/4 του 1834, πάλι στο Υπουργείο Εσωτερικών, ιδρύθηκε το «Γραφείο Δημοσίας Οικονομίας», το οποίο συμπεριέλαβε και την Υπηρεσία Γενικής Στατιστικής του Κράτους.

Περί τα τέλη του 19ου αιώνα η Στατιστική έχει το κατάλληλο επιστημονικό υπόβαθρο για την ανάπτυξη στατιστικών μεθόδων.

Θα πρέπει εδώ να σημειώσουμε τη συμβολή των Άγγλων Karl Pearson (1857-1936) και R.A. Fisher (1890-1962) στη θεμελίωση της σύγχρονης Στατιστικής, όπως επίσης και του Βέλγου στατιστικολόγου A. Quetelet, ο οποίος ήταν και ο κύριος διοργανωτής του πρώτου διεθνούς συνέδριου Στατιστικής το 1853.

Στις αρχές του 20ου αιώνα, χρησιμοποιήθηκε η στατιστική μέθοδος ελέγχου και ποιότητας των βιομηχανικών προϊόντων και θεωρείται ως η πρώτη σε ευρεία εφαρμογή χρήση στατιστικών μεθόδων στην παραγωγή.

Σήμερα, η στατιστική χρησιμοποιείται σε όλες σχεδόν τις επιστήμες και στους περισσότερους τομείς της ανθρώπινης δραστηριότητας. Ενδεικτικά αναφέρουμε ότι η Στατιστική συνεργάζεται με τις Οικονομικές Επιστήμες (παρουσίαση εθνικού προϊόντος, παρακολούθηση του δείκτη τιμών καταναλωτή και των μεταβολών του, καταγραφή των εμπορικών σχέσεων μιας χώρας με τις άλλες χώρες κτλ.), με την Επιστήμη της Οργάνωσης και Διοίκησης επιχειρήσεων (ανάλυση των συνθηκών αγοράς, παραγωγή και τιμές αγαθών κτλ.), τις Κοινωνικές Επιστήμες (ποσοτική διερεύνηση ανεργίας, μετανάστευσης, εγκληματικότητας κτλ.), την Ιατρική (μελέτη επιδημιών και άλλων νόσων), τις Πολιτικές Επιστήμες (σφυγμομετρήσεις κοινής γνώμης), τη Βιολογία, τη Μετεωρολογία, τη Γεωλογία, τη Σεισμολογία κλπ. Στην επέκταση της χρήσης των στατιστικών μεθόδων συνέβαλε η ταχύτατη τεχνολογική πρόοδος που σημειώθηκε τα τελευταία χρόνια στη δυνατότητα εκτέλεσης πολύπλοκων αριθμητικών υπολογισμών και στη μετάδοση των πληροφοριών.

Συνοψίζοντας λοιπόν μπορούμε να πούμε ότι:
Στατιστική είναι ένα σύνολο αρχών και μεθοδολογιών για
α. το σχεδιασμό της διαδικασίας συλλογής δεδομένων,
β. τη συνοπτική και αποτελεσματική παρουσίασή τους,
γ. την εξαγωγή αντίστοιχων συμπερασμάτων.

Ο κλάδος της Στατιστικής που ασχολείται με τον πρώτο στόχο λέγεται **σχεδιασμός πειραμάτων**, ενώ για το δεύτερο στόχο φροντίζει η **περιγραφική στατιστική**. Τέλος η **στατιστική συμπερασματολογία** μας δίνει τα μέσα για την κατάλληλη ανάλυση των στατιστικών στοιχείων και την εξαγωγή συμπερασμάτων για τον πληθυσμό από τον οποίο προέρχονται.

Δεν συμπεριλαμβάνεται εδώ η λήψη αποφάσεων, γιατί αυτό δεν είναι αρμοδιότητα του επιστήμονα στατιστικού που κάνει τη μελέτη, αλλά αυτού που ανέθεσε τη μελέτη στον στατιστικό.

Σύμφωνα με τον ορισμό της Στατιστικής μπορούμε να απαριθμήσουμε τα βασικά στάδια που ακολουθούμε για τη μελέτη ενός φαινομένου. Αυτά είναι:

1. Ο σχεδιασμός της μελέτης του φαινομένου.
2. Η συγκέντρωση των απαραιτήτων πληροφοριών (στατιστικών στοιχείων).
3. Η επεξεργασία και παρουσίαση των στοιχείων αυτών.
4. Η ανάλυση των στοιχείων με επιστημονικές μεθόδους.
5. Η εξαγωγή χρήσιμων συμπερασμάτων.

Οι έρευνες των ανθρώπινων πληθυσμών (**δημοσκοπήσεις**) αποτελούν σπουδαίες πηγές γνώσης των κοινωνικών επιστημών. Οικονομολόγοι, ψυχολόγοι, κοινωνιολόγοι και πολιτικοί επιστήμονες μελετούν ποικίλα θέματα, όπως πρότυπα εσόδων-εξόδων

των οικογενειών και των επιχειρήσεων, την επίδραση της επαγγελματικής απασχόλησης των γυναικών στην οικογενειακή τους ζωή, την επίδραση των μέσων μαζικής ενημέρωσης στις προτιμήσεις των καταναλωτών, τις προτιμήσεις των ψηφοφόρων κ.α.

1.2 Βασικές έννοιες της Στατιστικής

1.2.1 Πληθυσμός - μεταβλητές

Όπως αναφέρθηκε στον ορισμό της Στατιστικής, αυτό που μας ενδιαφέρει είναι να εξετάσουμε τα στοιχεία ενός συνόλου ως προς ένα ή περισσότερα χαρακτηριστικά τους και να συγκεντρώσουμε τις απαραίτητες πληροφορίες (στατιστικά στοιχεία).
Τέτοιες περιπτώσεις έχουμε, όταν π.χ. εξετάζουμε:

- Τα βιβλία μιας βιβλιοθήκης ως προς το περιεχόμενό τους.
- Τους μαθητές μιας τάξης ως προς το ύψος τους ή ως προς το βάρος τους ή ως προς τη βαθμολογία τους σ' ένα μάθημα.
- Τις οικογένειες μιας πόλης ως προς τον αριθμό των παιδιών και των αριθμό των δωματίων της κατοικίας τους.
- Τις ενδείξεις κατά τις διαδοχικές ρίψεις ενός ζαριού.
- Το χρώμα των αυτοκινήτων.

Πριν προχωρήσουμε κρίνεται απαραίτητο να ορίσουμε με σαφήνεια μερικούς όρους τους οποίους θα χρησιμοποιήσουμε στη συνέχεια.

• **Στατιστικός πληθυσμός** ή **πληθυσμός** είναι κάθε σύνολο, τα στοιχεία του οποίου εξετάζουμε ως προς μία ή περισσότερες χαρακτηριστικές ιδιότητες.

• **Άτομα** ή **στατιστικές μονάδες** είναι τα στοιχεία του συνόλου αυτού.

• **Μεταβλητή** είναι η χαρακτηριστική ιδιότητα ως προς την οποία εξετάζουμε τα άτομα του πληθυσμού. Οι μεταβλητές συμβολίζονται με κεφαλαία γράμματα: X, Y, Z, ...

• **Τιμές της μεταβλητής** είναι οι αριθμοί ή οι άλλες συμβολικές εκφράσεις (συνήθως λέξεις) που μετρούν ή εκφράζουν τις διάφορες καταστάσεις μιας μεταβλητής. Με απλά λόγια μια μεταβλητή θέτει ένα ερώτημα στα στοιχεία του πληθυσμού. Οι απαντήσεις στο ερώτημα αυτό είναι οι τιμές τις μεταβλητής.

• Κάθε τιμή που παίρνει η μεταβλητή X, όταν εξετάζουμε ως προς αυτή τον πληθυσμό, λέγεται **παρατήρηση**.

Οι μεταβλητές διακρίνονται:

1. Σε **ποιοτικές** ή **κατηγορικές μεταβλητές,** που δεν επιδέχονται μέτρηση και οι τιμές τους δεν εκφράζονται με αριθμούς, αλλά με λέξεις. Τέτοιες είναι για παράδειγμα, το επάγγελμα, το φύλλο, η ομάδα αίματος κ.τ.λ.

2. Σε **ποσοτικές μεταβλητές** που επιδέχονται μέτρηση και οι τιμές τους εκφράζονται με αριθμούς σε συγκεκριμένες μονάδες και διακρίνονται:

α) σε **διακριτές μεταβλητές** που παίρνουν μόνο μεμονωμένες τιμές. Τέτοιες είναι για παράδειγμα, ο αριθμός των υπαλλήλων μιας επιχείρησης (με τιμές 1, 2, …), το αποτέλεσμα της ρίψης ενός ζαριού (με τιμές 1, 2, …,6) κ.τ.λ.

β) σε **συνεχείς μεταβλητές** που μπορούν να πάρουν οποιαδήποτε τιμή ενός διαστήματος πραγματικών αριθμών (α, β). Τέτοιες είναι το ύψος, το βάρος των μαθητών κ.τ.λ.

Για την καλύτερη κατανόηση των εννοιών «πληθυσμός», «άτομα πληθυσμού», «μεταβλητή» και «τιμές μεταβλητές» παραθέτουμε τον παρακάτω πίνακα.

Πίνακας 1.1
Παραδείγματα για την καλύτερη κατανόηση των εννοιών

ΠΛΗΘΥΣΜΟΣ	ΑΤΟΜΟ	ΜΕΤΑΒΛΗΤΗ	ΤΙΜΕΣ
1. Βιβλία μιας βιβλιοθήκης	Ένα βιβλίο	Περιεχόμενο Βιβλίων	[Λ, Ι, Μ, Ο, Ε] (1)
2. Μαθητές μιας τάξης	Ένας μαθητής	Ύψος μαθητών	[150, 153, ..., 175]
3. Μαθητές μιας τάξης	Ένας μαθητής	Βάρος μαθητών	[45, 50, ..., 80]
4. Μαθητές μιας τάξης	Ένας μαθητής	Βαθμολογία	[5, 6, ..., 20]
5. Οικογένειες μιας πόλης	Μια οικογένεια	Αριθμός παιδιών	[1, 2, ..., 6]
6. Οικογένειες μιας πόλης	Μια οικογένεια	Αριθμός δωματίων	[2, 3, ...7]
7. Αυτοκίνητα	Ένα αυτοκίνητο	Χρώμα αυτοκινήτου	[Α, Κο, Κι, Μ, Π] (2)

(1) Λογοτεχνικό, Ιστορικό, Μαθηματικό, Οικονομικό, Εγκυκλοπαιδικό

(2) Άσπρο, Κόκκινο, Κίτρινο, Μαύρο, Πράσινο

1.2.2 Συλλογή Στατιστικών δεδομένων

Η συλλογή των στατιστικών στοιχείων μπορεί να γίνει απευθείας από τον πληθυσμό, οπότε τα στατιστικά στοιχεία προέρχονται από όλες τις μονάδες (άτομα) του πληθυσμού που θέλουμε να μελετήσουμε, ή ακόμα από ένα γνήσιο υποσύνολο του πληθυσμού.

- Όταν οι παρατηρήσεις προκύπτουν από όλα τα στοιχεία του πληθυσμού, λέμε ότι κάνουμε **απογραφή**.
- Όταν οι παρατηρήσεις προκύπτουν από ένα υποσύνολο του πληθυσμού, η έρευνα ονομάζεται **δειγματοληψία** και το εξεταζόμενο υποσύνολο **δείγμα**.

Στις περισσότερες περιπτώσεις η εξέταση όλων των ατόμων ενός πληθυσμού, δηλαδή η απογραφή, είναι δύσκολη ή αδύνατη ή ακόμα ασύμφορη. Τότε ο ερευνητής συλλέγει τις πληροφορίες του με δειγματοληψία. Κάνει τις παρατηρήσεις του στο δείγμα και μετά γενικεύει τα συμπεράσματά του για ολόκληρο τον πληθυσμό. Τα συμπεράσματα που θα προκύψουν από τη μελέτη του δείγματος θα είναι αξιόπιστα, θα ισχύουν δηλαδή με ικανοποιητική ακρίβεια για ολόκληρο τον πληθυσμό, αν η επιλογή του δείγματος γίνει με σωστό τρόπο, ώστε το δείγμα να είναι, όπως λέμε, **αντιπροσωπευτικό** του πληθυσμού. Το πλήθος των μονάδων του δείγματος καλείται **μέγεθος (ν) του δείγματος**.

Οι αρχές και οι μέθοδοι για τη συλλογή και ανάλυση δεδομένων από πεπερασμένους πληθυσμούς είναι το αντικείμενο της **Δειγματοληψίας**.

Όταν εξετάζουμε τις μονάδες ενός πληθυσμού ή ενός δείγματος ως προς ένα κοινό χαρακτηριστικό, τότε προκύπτουν πληροφορίες που είναι αποτέλεσμα μετρήσεων ή παρατηρήσεων. Οι πληροφορίες αυτές καλούνται **στατιστικά δεδομένα**.

1.3 Παρουσίαση στατιστικών δεδομένων

1.3.1 Στατιστικοί πίνακες

Μετά τη συλλογή την επεξεργασία και την ταξινόμηση των στατιστικών δεδομένων, ακολουθεί το στάδιο της συνοπτικής παρουσίασης των συγκεντρωθέντων στοιχείων, έτσι ώστε να είναι εύκολη η κατανόησή τους και η εξαγωγή σωστών συμπερασμάτων. Η παρουσίαση των πληροφοριών γίνεται συνήθως με συνοπτικούς **πίνακες** ή **γραφικές παραστάσεις**. Η παρουσίαση των στατιστικών δεδομένων σε πίνακες γίνεται με την κατάλληλη τοποθέτηση των πληροφοριών σε γραμμές και στήλες, με τρόπο που να διευκολύνεται η σύγκριση των στοιχείων και η καλύτερη ενημέρωση του αναγνώστη σχετικά με τη δομή του πληθυσμού που ερευνάται.

Οι πίνακες διακρίνονται στους:

1) **Γενικούς ή λεπτομερείς πίνακες**, οι οποίοι περιέχουν όλες τις πληροφορίες που προκύπτουν από μια στατιστική έρευνα (συνήθως με αρκετά λεπτομερειακά στοιχεία) και αποτελούν πηγές στατιστικών πληροφοριών στη διάθεση των επιστημόνων – ερευνητών για παραπέρα ανάλυση και εξαγωγή συμπερασμάτων.

2) **Ειδικούς ή συνοπτικούς πίνακες**, οι οποίοι παρουσιάζουν συνοπτικά μερικές μόνο πληροφορίες για έναν πληθυσμό. Τα στοιχεία τους συνήθως έχουν ληφθεί από τους γενικούς πίνακες.

Κάθε πίνακας που έχει κατασκευαστεί σωστά πρέπει να περιέχει:

α) τον **τίτλο**, που γράφεται στο επάνω μέρος του πίνακα και δηλώνει με σαφήνεια και συνοπτικά το περιεχόμενο του πίνακα,

β) τις **επικεφαλίδες** των γραμμών και των στηλών, που δείχνουν συνοπτικά τη φύση και τις μονάδες μέτρησης των δεδομένων,

γ) το **κύριο σώμα** (κορμό), που περιέχει διαχωρισμένα μέσα στις γραμμές και στις στήλες τα στατιστικά δεδομένα,

δ) την **πηγή** που γράφεται στο κάτω μέρος του πίνακα και δείχνει την προέλευση των στατιστικών στοιχείων, έτσι ώστε ο αναγνώστης να ανατρέχει σ' αυτήν, όταν επιθυμεί, για επαλήθευση στοιχείων ή για λήψη περισσότερων πληροφοριών.

Τους βασικούς τύπους των ειδικών πινάκων οι οποίοι μας ενδιαφέρουν περισσότερο, τους διακρίνουμε κυρίως σε δύο κατηγορίες:

Α. **Πίνακες απλής εισόδου** (Πίνακας 1.2), όταν αναφέρονται σε πληροφορίες για ένα πληθυσμό ως προς μόνο ένα ποιοτικό ή ποσοτικό χαρακτηριστικό (μεταβλητή).

Β. **Πίνακες διπλής εισόδου** (Πίνακας 1.3), όταν αναφέρονται σε πληροφορίες για ένα πληθυσμό, ως προς δύο ποιοτικά ή ποσοτικά χαρακτηριστικά (μεταβλητές).

Πίνακας 1.2

Στοιχεία ανά κλάδο, βιομηχανικών επιχειρήσεων στην Ελλάδα με απασχόληση πάνω από 10 άτομα, το έτος 2010

Αξίες σε ευρώ

ΚΩΔΙΚΟΣ ΚΛΑΔΟΥ	ΑΡΙΘ. ΕΠΙΧΕΙΡ.	ΣΥΝΟΛΟ ΑΠΑΣΧ/ΝΩΝ	ΑΜΟΙΒΕΣ ΑΠΑΣΧΟΛΟΥΜΕΝΩΝ	ΑΚΑΘΑΡΙΣΤΗ ΑΞΙΑ ΠΑΡΑΓΩΓΗΣ	ΣΥΝΟΛΟ ΑΝΑΛΩΣΕΩΝ
10	833	55.556	1.128.115.726	7.724.532.698	4.761.491.587
11	94	9.114	308.895.222	1.819.240.059	747.374.596
12	5	1.944	82.801.860	363.602.259	179.997.456
13	137	6.062	120.073.793	540.145.649	341.362.113
14	282	8.855	153.025.082	743.835.378	482.866.544
15	49	1.480	25.722.275	98.029.696	55.637.913
16	109	3.881	75.467.320	333.472.561	220.300.193
17	117	6.223	142.074.910	839.586.204	518.578.966
18	180	6.597	154.269.268	596.274.982	280.883.376
19	7	4.333	268.058.516	12.759.630.111	11.732.084.615
20	155	9.478	230.242.444	1.503.207.785	893.998.401
21	47	7.650	188.795.810	1.035.966.480	490.361.304
22	217	9.975	206.739.031	1.217.283.269	739.856.682
23	333	14.670	393.436.568	2.030.599.777	1.170.000.989
24	75	10.337	320.148.897	3.830.718.139	3.064.993.076
25	399	16.539	370.047.136	2.196.958.437	1.305.086.796
26	32	3.249	92.088.078	310.354.317	105.909.735
27	123	6.213	145.521.258	982.579.577	674.115.210
28	208	6.017	114.502.491	490.517.003	252.949.256
29	29	1.764	38.812.079	166.661.622	97.353.403
30	39	6.124	209.065.670	408.590.111	144.116.930
31	236	6.619	106.265.128	393.988.181	202.006.650
32	91	2.131	33.485.885	112.359.832	51.298.408
33	94	4.151	71.276.424	215.180.136	61.363.052
ΣΥΝΟΛΟ	**3.891**	**208.963**	**4.978.930.872**	**40.713.314.264**	**28.573.987.250**

Πηγή: ΕΛΣΤΑΤ

Πίνακας 1.3

Κατανομή του πληθυσμού της Ελλάδας σε γεωγραφικά διαμερίσματα, κατά φύλο και περιφέρεια, σύμφωνα με την απογραφή του 2011

ΠΕΡΙΓΡΑΦΗ	Σύνολο	Άρρενες	Θήλεις
ΣΥΝΟΛΟ ΧΩΡΑΣ	**10.816.286**	**5.303.223**	**5.513.063**
ΠΕΡΙΦΕΡΕΙΑ ΑΝΑΤΟΛΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ ΚΑΙ ΘΡΑΚΗΣ	608.182	299.643	308.539
ΠΕΡΙΦΕΡΕΙΑ ΚΕΝΤΡΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ	1.882.108	912.693	969.415
ΠΕΡΙΦΕΡΕΙΑ ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ	283.689	141.779	141.910
ΠΕΡΙΦΕΡΕΙΑ ΗΠΕΙΡΟΥ	336.856	165.775	171.081
ΠΕΡΙΦΕΡΕΙΑ ΘΕΣΣΑΛΙΑΣ	732.762	362.194	370.568
ΠΕΡΙΦΕΡΕΙΑ ΣΤΕΡΕΑΣ ΕΛΛΑΔΑΣ	547.390	277.475	269.915
ΠΕΡΙΦΕΡΕΙΑ ΙΟΝΙΩΝ ΝΗΣΩΝ	207.855	102.400	105.455
ΠΕΡΙΦΕΡΕΙΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ	679.796	339.310	340.486
ΠΕΡΙΦΕΡΕΙΑ ΠΕΛΟΠΟΝΝΗΣΟΥ	577.903	291.777	286.126
ΠΕΡΙΦΕΡΕΙΑ ΑΤΤΙΚΗΣ	3.828.434	1.845.663	1.982.771
ΠΕΡΙΦΕΡΕΙΑ ΒΟΡΕΙΟΥ ΑΙΓΑΙΟΥ	199.231	99.984	99.247
ΠΕΡΙΦΕΡΕΙΑ ΝΟΤΙΟΥ ΑΙΓΑΙΟΥ	309.015	155.865	153.150
ΠΕΡΙΦΕΡΕΙΑ ΚΡΗΤΗΣ	623.065	308.665	314.400

Πηγή: ΕΛΣΤΑΤ

1.3.2 Πίνακες κατανομής συχνοτήτων

Συχνότητα (ν_i)

Συχνότητα (ή **απόλυτη συχνότητα**) μιας τιμής x_i της μεταβλητής X, λέγεται ο φυσικός αριθμός ν_i που δηλώνει πόσες φορές παρουσιάζεται στο δείγμα (ή σε ολόκληρο τον πληθυσμό, αν αναφερόμαστε σε αυτόν) η τιμή αυτή.

Για τη συχνότητα ν_i ισχύουν οι εξής ιδιότητες:

α) $0 \le \nu_i \le \nu$

β) $\nu_1 + \nu_2 + ... + \nu_\kappa = \nu$

Παράδειγμα 1.1

Σε κάποιο σχολείο θέλουμε να εξετάσουμε τους μαθητές μιας τάξης ως προς το χαρακτηριστικό πόσα αδέλφια έχουν. Από τον πληθυσμό του σχολείου σχηματίζουμε ένα δείγμα 50 ατόμων ($\nu = 50$). Οι απαντήσεις που πήραμε είναι οι εξής:

0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
2	3	3	3	3	3	3	3	4	4

Στο δείγμα αυτό, η μεταβλητή X : «αριθμός αδελφών» παίρνει τις εξής πέντε τιμές:

$$x_1 = 0,\ x_2 = 1,\ x_3 = 2,\ x_4 = 3 \text{ και } x_5 = 4$$

- Η συχνότητα της τιμής $x_1 = 0$ είναι $\nu_1 = 12$, διότι η τιμή 0 εμφανίζεται 12 φορές στο δείγμα.
- Η συχνότητα της τιμής $x_2 = 1$ είναι $\nu_2 = 18$, διότι η τιμή 1 εμφανίζεται 18 φορές στο δείγμα.
- Ανάλογα βρίσκουμε ότι για τις τιμές $x_3 = 2,\ x_4 = 3$ και $x_5 = 4$, οι συχνότητες είναι $\nu_3 = 11,\ \nu_4 = 7$ και $\nu_5 = 2$, αντίστοιχα.

Τα προηγούμενα παρουσιάζονται συνοπτικά στον παρακάτω πίνακα, όπου στην πρώτη στήλη γράφουμε τις τιμές x_i της μεταβλητής X και στη δεύτερη στήλη τις (απόλυτες) συχνότητες ν_i των τιμών x_i της μεταβλητής X.

Πίνακας 1.4

Πίνακας συχνοτήτων της μεταβλητής: «αριθμός αδελφών»

Τιμές της μεταβλητής x_i	**Αριθμός παρατηρήσεων (συχνότητα ν_i)**
0	12
1	18
2	11
3	7
4	2
Σύνολο	**50**

Ο Πίνακας 1.4, λέγεται **πίνακας συχνοτήτων** ή **κατανομής συχνοτήτων**. Είναι πίνακας απλής εισόδου, γιατί τα στατιστικά στοιχεία (παρατηρήσεις) αναφέρονται σε μια μόνο ιδιότητα των στοιχείων του πληθυσμού.

Το σύνολο των ζευγών (x_i, ν_i) (τιμή, συχνότητα), λέμε ότι αποτελεί την **κατανομή των συχνοτήτων**.

Σχετική συχνότητα (f_i)

Θεωρούμε ένα δείγμα μεγέθους ν και έστω x_1, x_2, …, x_κ, με $\kappa \le \nu$, οι τιμές της μεταβλητής X ως προς την οποία μελετάμε το δείγμα. Αν ν_i είναι η συχνότητα της τιμής x_i, τότε **σχετική συχνότητα** της τιμής x_i, λέγεται το πηλίκο της συχνότητας ν_i προς το πλήθος ν των παρατηρήσεων, και συμβολίζεται με f_i. Δηλαδή:

$$f_i = \frac{\nu_i}{\nu}\text{, με i = 1, 2, …, κ.}$$

Για τη σχετική συχνότητα ισχύουν οι ιδιότητες:

α) $0 \le f_i \le 1$ για i = 1, 2, …, κ,

β) $f_1 + f_2 + \ldots + f_\kappa = 1$

Επί τοις εκατό σχετική συχνότητα ($f_i\%$)

Επί τοις εκατό σχετική συχνότητα $f_i\%$ της τιμής x_i ονομάζουμε το ποσοστό των τιμών του δείγματος που έχουν τιμή ίση με x_i. Είναι δηλαδή

$$f_i\% = \frac{\nu_i}{\nu} \cdot 100 = f_i \cdot 100\text{, με i = 1, 2, …, κ}$$

Ισχύουν οι εξής προτάσεις:

α) $0 \le f_i\% \le 100$ για i = 1, 2, …, κ,

β) $f_1\% + f_2\% + \ldots + f_\kappa\% = 100$.

Για μια μεταβλητή, το σύνολο των ζευγών (x_i, f_i), ή $(x_i, f_i\%)$, (τιμή, σχετική συχνότητα) λέμε ότι αποτελεί την **κατανομή σχετικών συχνοτήτων**.

Οι ποσότητες (x_i, ν_i, f_i), (τιμή, συχνότητα, σχετική συχνότητα) για ένα δείγμα μεγέθους ν συγκεντρώνονται σε ένα συνοπτικό πίνακα που ονομάζεται **πίνακας κατανομής συχνοτήτων** ή απλά **πίνακας συχνοτήτων**.

Αθροιστική συχνότητα (N_i)

Θεωρούμε ένα δείγμα μεγέθους ν και έστω $x_1, x_2, \ldots, x_\kappa$, με $\kappa \le \nu$, οι τιμές μιας ποσοτικής μεταβλητής X ως προς την οποία μελετάμε το δείγμα. Θεωρούμε ακόμα ότι οι τιμές $x_1, x_2, \ldots, x_\kappa$ είναι ταξινομημένες σε αύξουσα σειρά.

Αθροιστική συχνότητα N_i της τιμής x_i, ορίζουμε το άθροισμα των συχνοτήτων των τιμών της μεταβλητής που είναι μικρότερες ή ίσες από την τιμή x_i. Είναι δηλαδή:

$$N_i = \nu_1 + \nu_2 + \ldots + \nu_i \text{, για i = 1, 2, ..., κ.}$$

Με βάση τον παραπάνω ορισμό είναι $N_1 = \nu_1$, $N_2 = N_1 + \nu_2$, $N_3 = N_2 + \nu_3$, … και γενικά

$$N_\kappa = N_{\kappa-1} + \nu_k \text{, με } 1 < \kappa \le \nu.$$

Σχόλιο

Αν $x_1, x_2, \ldots, x_\kappa$, με $\kappa \le \nu$ είναι οι τιμές μιας ποσοτικής μεταβλητής X, τότε είναι:

$$N_\kappa = \nu_1 + \nu_2 + \ldots + \nu_\kappa = \nu.$$

Δηλαδή, η αθροιστική συχνότητα της τελευταίας μεταβλητής x_κ είναι όσο το μέγεθος ν του δείγματος.

Αθροιστική σχετική συχνότητα (F_i)

Θεωρούμε ένα δείγμα μεγέθους ν και έστω $x_1, x_2, \ldots, x_\kappa$, με $\kappa \le \nu$, οι τιμές μιας ποσοτικής μεταβλητής X ως προς την οποία μελετάμε το δείγμα. Θεωρούμε ακόμα ότι οι τιμές $x_1, x_2, \ldots, x_\kappa$ είναι ταξινομημένες σε αύξουσα σειρά.

Αθροιστική σχετική συχνότητα F_i της τιμής x_i, με F_i ή $F_i\%$ λέγεται το άθροισμα των σχετικών συχνοτήτων f_i ή $f_i\%$ των τιμών της μεταβλητής που είναι μικρότερες ή ίσες από την τιμή x_i. Δηλαδή είναι:

$$F_i = f_1 + f_2 + \ldots + f_i \text{, για i = 1, 2, ..., κ.}$$

Με βάση τον παραπάνω ορισμό είναι $F_1 = f_1$, $F_2 = F_1 + f_2$, $F_3 = F_2 + f_3$, … και γενικά

$$F_\kappa = F_{\kappa-1} + f_k \text{, με } 1 < \kappa \le \nu.$$

Σχόλιο

Αν x_1, x_2, …, x_κ, είναι οι τιμές μιας ποσοτικής μεταβλητής X, τότε είναι:

$$F_\kappa = f_1 + f_2 + ... + f_\kappa = 1 .$$

Δηλαδή, η αθροιστική συχνότητα της τελευταίας μεταβλητής x_κ είναι 1.

Από τους ορισμούς των αθροιστικών συχνοτήτων και αθροιστικών σχετικών συχνοτήτων προκύπτουν:

- Οι αθροιστικές συχνότητες και οι αθροιστικές σχετικές συχνότητες δεν ορίζονται για ποιοτικές μεταβλητές, γιατί στις ποιοτικές μεταβλητές δεν έχει νόημα η τάξη μεγέθους.
- Ακόμη ισχύουν:

$\nu_1 = N_1$ $\quad f_1 = F_1$ και $\quad f_1\% = F_1\%$

$\nu_2 = N_2 - N_1$ $\quad f_2 = F_2 - F_1$ και $\quad f_2\% = F_2\% - F_1\%$

$\nu_3 = N_3 - N_2$ $\quad f_3 = F_3 - F_2$ και $\quad f_3\% = F_3\% - F_2\%$

………………………………………………………………..

$\nu_i = N_i - N_{i-1}$ $\quad f_i = F_i - F_{i-1}$ και $\quad f_i\% = F_i\% - F_{i-1}\%$

1.3.3 Γραφική παράσταση κατανομής συχνοτήτων

Τα στατιστικά δεδομένα τα οποία συγκεντρώνουμε σε ένα πίνακα κατανομής συχνοτήτων μπορούμε να τα παρουσιάσουμε υπό μορφή γραφικών παραστάσεων ή διαγραμμάτων.

Οι γραφικές παραστάσεις είναι το καλύτερο μέσο στατιστικής παρουσίασης γιατί, δίνουν στους αφηρημένους αριθμούς μια συγκεκριμένη μορφή που μας διευκολύνει να έχουμε, με τη βοήθεια ενός γεωμετρικού σχήματος, μια άμεση αντίληψη της μορφής της μεταβλητής (χαρακτηριστικό) που θέλουμε να μελετήσουμε. Από μια καλά σχεδιασμένη γραφική παράσταση μπορούμε να πάρουμε γρήγορα και συνοπτικά διάφορες χρήσιμες πληροφορίες και να διατηρηθεί στη μνήμη καλύτερα από έναν αριθμητικό πίνακα.

Με τα διαγράμματα διευκολύνεται η σύγκριση μεταξύ ομοειδών στοιχείων για το ίδιο ή για διαφορετικά χαρακτηριστικά.

Η στατιστική χρησιμοποιεί πολλά είδη γραφικών παραστάσεων. Όπως κάθε στατιστικός πίνακας, έτσι και κάθε γραφική παράσταση πρέπει να περιλαμβάνει εκτός από το σχέδιο και:

α) τον τίτλο

β) την κλίμακα και τις τιμές των μεγεθών που απεικονίζονται

γ) το υπόμνημα που εξηγεί τις τιμές της μεταβλητής και

δ) την πηγή των δεδομένων.

Παρουσίαση ποιοτικών μεταβλητών

1. Ραβδόγραμμα

Το **ραβδόγραμμα** αποτελείται από ορθογώνιες στήλες ώστε να ισχύουν τα εξής:

- οι βάσεις τους έχουν ίσα μήκη, βρίσκονται πάνω στον οριζόντιο ή τον κατακόρυφο άξονα και ισαπέχουν μεταξύ τους.
- σε κάθε τιμή x_i της μεταβλητής X αντιστοιχεί μια ορθογώνια στήλη της οποίας το ύψος είναι ίσο με την αντίστοιχη συχνότητα ν_i ή τη σχετική συχνότητα f_i ή $f_i\%$, ανάλογα με το αν έχουμε ραβδόγραμμα συχνοτήτων ή σχετικών συχνοτήτων.

Το ραβδόγραμμα χρησιμοποιείται για τη γραφική παράσταση των τιμών μιας ποιοτικής μεταβλητής.

Παράδειγμα 1.2

Να κατασκευαστεί το ραβδόγραμμα συχνοτήτων για τα δεδομένα του παρακάτω πίνακα κατανομής συχνοτήτων της μεταβλητής: «σχολικός πληθυσμός της Ελλάδας ανά βαθμίδα εκπαίδευσης του έτους 1985».

Πίνακας 1.5

Σχολικός πληθυσμός της Ελλάδας ανά βαθμίδα εκπαίδευσης

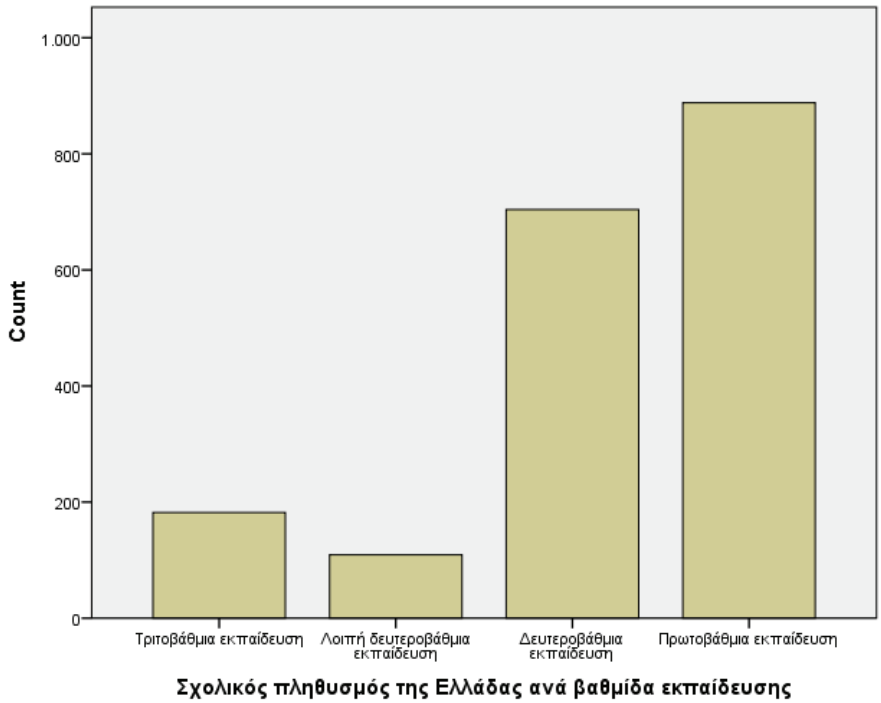
Βαθμίδες Εκπαίδευσης	Αριθμός φοιτούντων (σε χιλιάδες)
Τριτοβάθμια εκπαίδευση	182
Λοιπή δευτεροβάθμια εκπαίδευση	109
Δευτεροβάθμια γενική εκπαίδευση	704
Πρωτοβάθμια εκπαίδευση	888

Πηγή Ε.Σ.Υ.Ε. 1985

Στον οριζόντιο άξονα τοποθετούμε τις τιμές της μεταβλητής «βαθμίδες εκπαίδευσης» και στον κατακόρυφο άξονα τοποθετούμε τις αντίστοιχες συχνότητες. Στη συνέχεια, για κάθε τιμή της μεταβλητής κατασκευάζουμε ορθογώνια που ισαπέχουν μεταξύ τους και έχουν ίσες βάσεις. Τα ύψη των ορθογωνίων αυτών είναι ίσα με τις αντίστοιχες συχνότητες.

Εικόνα 1.1

Ραβδόγραμμα συχνοτήτων των δεδομένων του Πίνακα 1.5



Αν θέλουμε μπορούμε να τοποθετήσουμε τις τιμές της μεταβλητής στον κατακόρυφο άξονα και τις αντίστοιχες συχνότητες στον οριζόντιο.

Παράδειγμα 1.3

Να κατασκευαστεί το ραβδόγραμμα σχετικών συχνοτήτων για τα δεδομένα του παρακάτω πίνακα κατανομής σχετικών συχνοτήτων της μεταβλητής «πληθυσμός της Γης κατά περιοχές το έτος 1970».

Πίνακας 1.6

Ποσοστιαία κατανομή κατά περιοχές του πληθυσμού της Γης το 1970

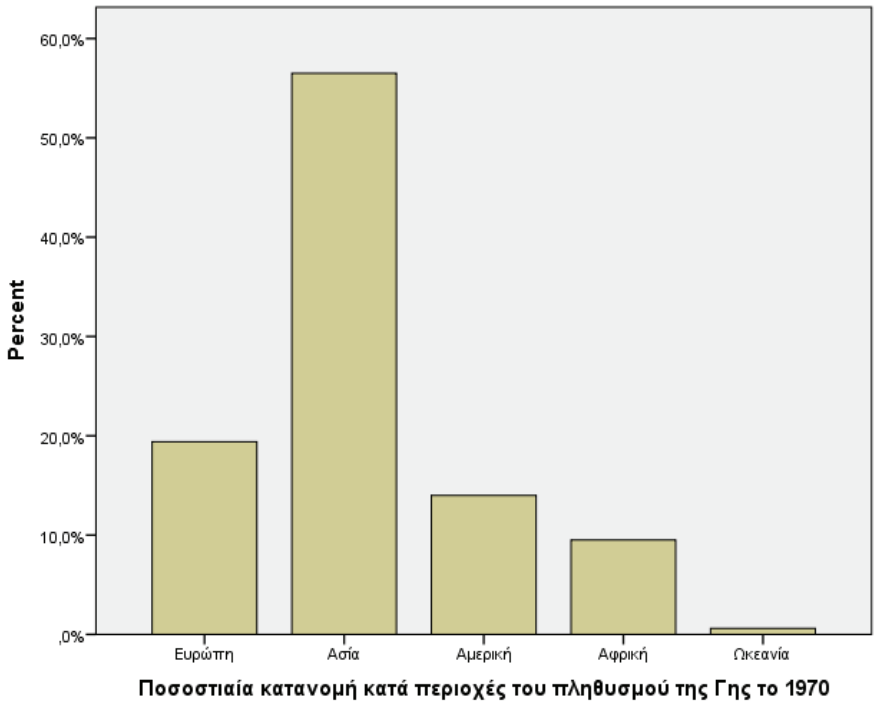
Περιοχές	Ποσοστό
Ευρώπη	19,4
Ασία	56,5
Αμερική	14,0
Αφρική	9,5
Ωκεανία	0,6
Σύνολο	**100,0**

Πηγή: Ε.Σ.Υ.Ε.

Στον οριζόντιο άξονα τοποθετούμε τις τιμές της μεταβλητής «πληθυσμός της Γης κατά περιοχές το έτος 1970» και στον κατακόρυφο άξονα τοποθετούμε τις αντίστοιχες σχετικές συχνότητες. Στη συνέχεια, για κάθε τιμή της μεταβλητής κατασκευάζουμε ορθογώνια που ισαπέχουν μεταξύ τους και έχουν ίσες βάσεις. Τα ύψη των ορθογωνίων αυτών είναι ίσα με τις αντίστοιχες σχετικές συχνότητες.

Εικόνα 1.2

Ραβδόγραμμα σχετικών συχνοτήτων των δεδομένων του Πίνακα 1.6



2. Κυκλικό διάγραμμα

Το **κυκλικό διάγραμμα** χρησιμοποιείται για τη γραφική παράσταση τόσο των ποιοτικών όσο και των ποσοτικών δεδομένων, όταν οι διαφορετικές τιμές της μεταβλητής είναι σχετικά λίγες. Το κυκλικό διάγραμμα είναι ένας κυκλικός δίσκος με επιλογή ακτίνας αυθαίρετη, χωρισμένος σε κυκλικούς τομείς. Κάθε κυκλικός τομέας αντιστοιχεί σε μια τιμή x_i της μεταβλητής και έχει γωνία (ή τόξο) ανάλογη με την αντίστοιχη συχνότητα ν_i ή τη σχετική συχνότητα f_i της τιμής.

Αν συμβολίσουμε με α_i την αντίστοιχη γωνία (ή τόξο) ενός κυκλικού τομέα στο κυκλικό διάγραμμα συχνοτήτων, τότε:

$$\alpha_i = \frac{\nu_i}{\nu} \cdot 360° = f_i \cdot 360°, \text{ i = 1, 2, …, κ.}$$

Παράδειγμα 1.4

Να κατασκευαστεί το κυκλικό διάγραμμα σχετικών συχνοτήτων για τα δεδομένα του παρακάτω πίνακα κατανομής σχετικών συχνοτήτων της μεταβλητής «κατανάλωση ενέργειας έτος 1991».

Πίνακας 1.7

Κατανάλωση ενέργειας το έτος 1991

ΧΡΗΣΕΙΣ	ΠΟΣΟΣΤΟ
Οικιακή	34
Εμπορική	16
Βιομηχανική	40
Γεωργική	5
Λοιπές χρήσεις	5

Πηγή: Ε.Σ.Υ.Ε.

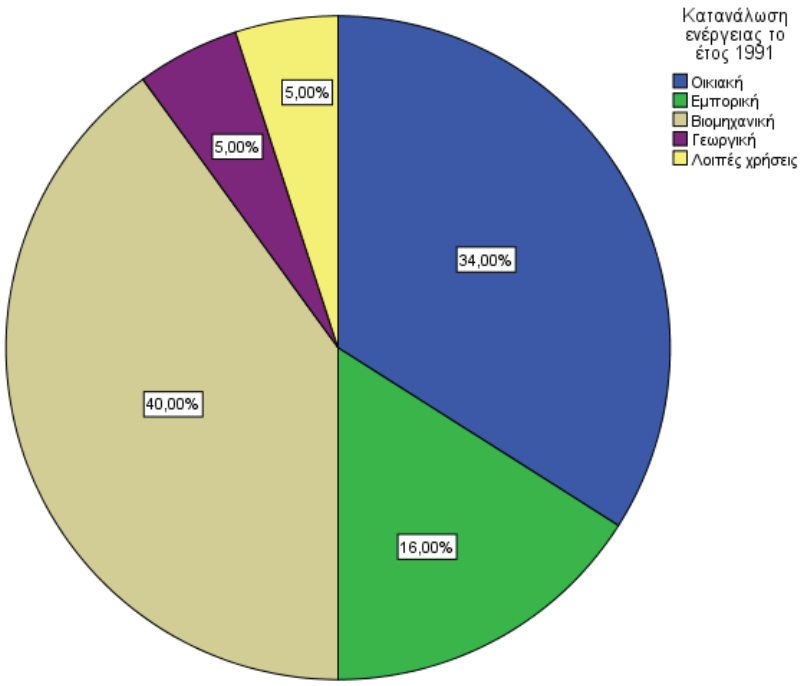
Η τιμή x_1: Οικιακή χρήση, έχει σχετική συχνότητα $f_1 = 0,34$ και σε αυτήν αντιστοιχεί κυκλικός τομέας που έχει γωνία $\alpha_1 = 0,34 \cdot 360° = 122,4°$.

Όμοια βρίσκουμε $\alpha_2 = 57,6°$, $\alpha_3 = 144°$, $\alpha_4 = 18°$ και $\alpha_5 = 18°$

Το ζητούμενο κυκλικό διάγραμμα παρουσιάζεται παρακάτω.

Εικόνα 1.3

Κυκλικό διάγραμμα σχετικών συχνοτήτων του Πίνακα 1.7



Παρουσίαση ποσοτικών μεταβλητών

1. Διάγραμμα – πολύγωνο συχνοτήτων

Το **διάγραμμα συχνοτήτων ή διάγραμμα σχετικών συχνοτήτων** (ν_i, f_i, $f_i\%$, N_i, F_i, $F_i\%$) χρησιμοποιείται για τη γραφική παράσταση των συχνοτήτων ή των σχετικών συχνοτήτων ποσοτικών διακριτών μεταβλητών.

Για να κατασκευάσουμε το διάγραμμα συχνοτήτων κάνουμε τα εξής.

- Τοποθετούμε στον οριζόντιο άξονα τις τιμές x_i της μεταβλητής κατ' αύξουσα σειρά.
- Υψώνουμε κάθετα ευθύγραμμα τμήματα στον οριζόντιο άξονα, ένα σε κάθε τιμή x_i.
- Το μήκος καθενός από τα κάθετα ευθύγραμμα τμήματα είναι ίσο με την αντίστοιχη συχνότητα ή σχετική συχνότητα (ν_i, f_i, $f_i\%$, N_i, F_i, $F_i\%$).

Το **πολύγωνο συχνοτήτων** ή **πολύγωνο σχετικών συχνοτήτων** αντίστοιχα, κατασκευάζεται αν ενώσουμε τα άνω άκρα των γραμμών του αντίστοιχου διαγράμματος συχνοτήτων ή σχετικών συχνοτήτων.

Τα πολύγωνα συχνοτήτων μας δίνουν μια γενική ιδέα για τη μεταβολή της συχνότητας ή της σχετικής συχνότητας όσο μεγαλώνει η τιμή της μεταβλητής που εξετάζουμε.

Παράδειγμα 1.5

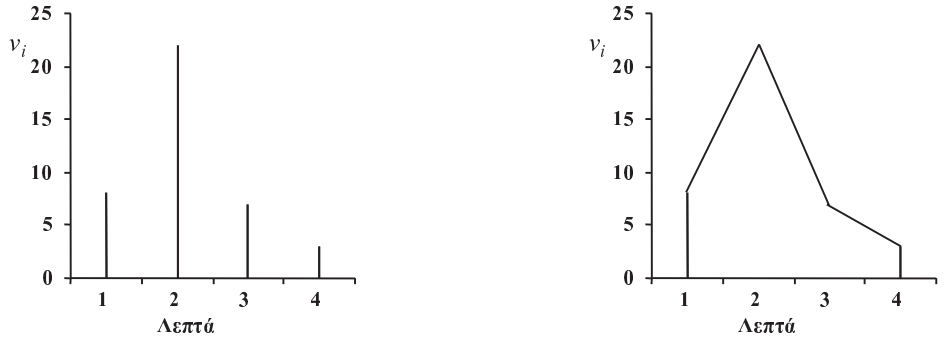
Στο διπλανό πίνακα εμφανίζονται οι χρόνοι (σε λεπτά) τους οποίους χρειάστηκαν 40 μαθητές για να λύσουν μια εξίσωση. Να κατασκευαστεί το διάγραμμα συχνοτήτων και το πολύγωνο συχνοτήτων.

x_i	ν_i
1	8
2	22
3	7
4	3
Σύνολο	**40**

Η διακριτή ποσοτική μεταβλητή X: «χρόνος επίλυσης μιας εξίσωσης» παίρνει τις τιμές: $x_1 = 8,\ x_2 = 22,\ x_3 = 7,\ x_4 = 3$. Στον οριζόντιο άξονα τοποθετούμε τις τιμές x_i της μεταβλητής κατά αύξουσα σειρά και υψώνουμε σε κάθε τιμή x_i κάθετη γραμμή με μήκος ίσο με την αντίστοιχη συχνότητα ν_i. Έτσι προκύπτει το διάγραμμα συχνοτήτων. Αν ενώσουμε τα άνω άκρα των γραμμών του προκύπτει το πολύγωνο συχνοτήτων.

Εικόνα 1.4

Διάγραμμα συχνοτήτων και πολύγωνο συχνοτήτων των δεδομένων του παραδείγματος 1.5



2. Σημειόγραμμα

Όταν το δείγμα είναι μικρό μπορούμε να χρησιμοποιήσουμε το **σημειόγραμμα** προκειμένου να παρουσιάσουμε διάφορα στατιστικά δεδομένα, είτε ποιοτικής είτε ποσοτικής μεταβλητής.

Το σημειόγραμμα αποτελείται από έναν οριζόντιο άξονα στον οποίο τοποθετούμε τις τιμές x_i της μεταβλητής και πάνω από κάθε τιμή βάζουμε κατακόρυφα τόσες τελείες όση είναι και η αντίστοιχη συχνότητα ν_i.

Παράδειγμα 1.6

Στο διπλανό πίνακα εμφανίζονται οι συχνότητες των ενδείξεων ενός ζαριού όταν ρίξαμε το ζάρι 15 φορές. Να κατασκευαστεί το σημειόγραμμα.

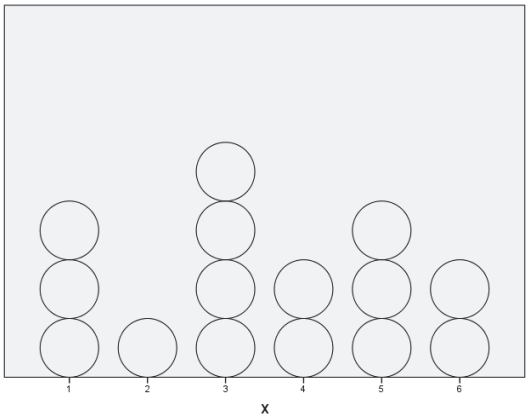
x_i	ν_i
1	3
2	1
3	4
4	2
5	3
6	2
Σύνολο	**15**

Λύση

Στον οριζόντιο άξονα τοποθετούμε τις 6 ενδείξεις x_i του ζαριού. Πάνω από κάθε ένδειξη x_i τοποθετούμε κατακόρυφα τόσες τελείες, όσες και το πλήθος των αντίστοιχων ενδείξεων.

Εικόνα 1.5

Σημειόγραμμα των δεδομένων του παραδείγματος 1.6



3. Χρονόγραμμα

Το **χρονόγραμμα** χρησιμοποιείται για τη γραφική απεικόνιση της διαχρονικής εξέλιξης ενός οικονομικού, δημογραφικού, ή άλλου μεγέθους.

Για να κατασκευάσουμε το χρονόγραμμα κάνουμε τα εξής.

- Στον οριζόντιο άξονα τοποθετούμε τις χρονικές στιγμές (ώρες, ημέρες, έτη, …).
- Στον κατακόρυφο άξονα τοποθετούμε τις τιμές της μεταβλητής που μελετάμε.
- Βρίσκουμε τα σημεία $M_i(t_i, x_i)$, όπου x_i είναι η τιμή της X τη χρονική στιγμή t_i.
- Ενώνουμε τα σημεία διαδοχικά με ευθύγραμμα τμήματα.

Παράδειγμα 1.7

Στον παρακάτω πίνακα παρουσιάζονται τα κέρδη μιας επιχείρησης, σε χιλιάδες ευρώ κατά τα έτη 2008 – 2013. Να κατασκευαστεί συγκριτικό χρονόγραμμα.

Λύση

Στον οριζόντιο άξονα τοποθετούμε τα έτη 2008 – 2013και στον κατακόρυφο άξονα τις τιμές της μεταβλητής. Βρίσκουμε τα σημεία $M_i(t_i, x_i)$ και τα ενώνουμε με διαδοχικά ευθύγραμμα τμήματα.

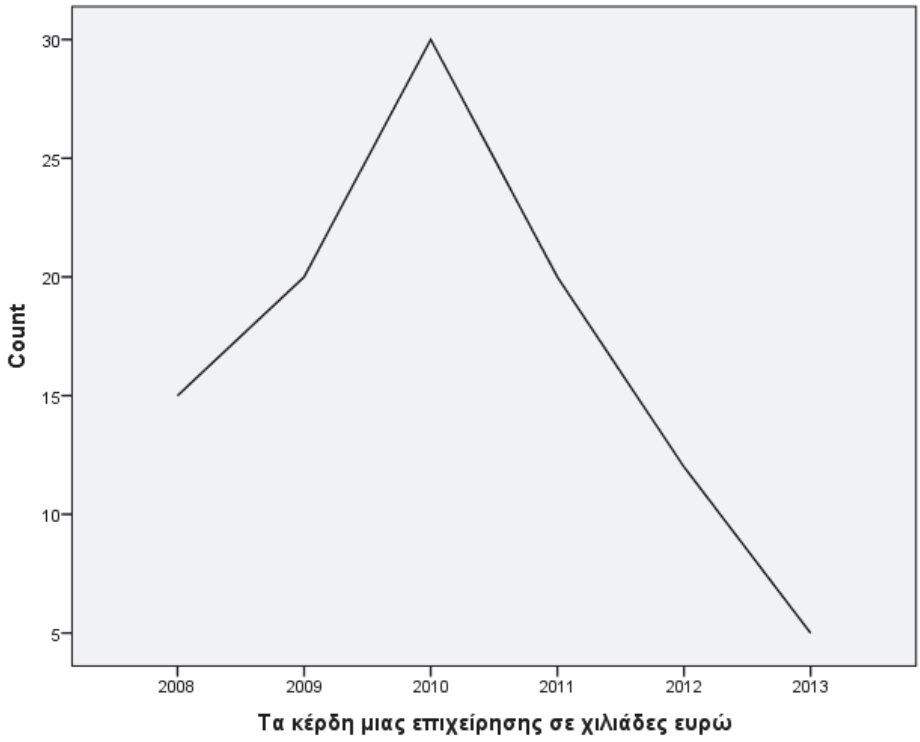
Πίνακας 1.8

Κέρδη μιας επιχείρησης σε χιλιάδες ευρώ τα έτη 2008-2013

Έτη	Κέρδη σε χιλιάδες ευρώ
2008	15
2009	20
2010	30
2011	20
2012	12
2013	5

Εικόνα 1.6

Χρονόγραμμα των κερδών της επιχείρησης του Πίνακα 1.8



1.4 Ομαδοποίηση στατιστικών δεδομένων

1.4.1 Ομαδοποίηση των παρατηρήσεων

Στην περίπτωση που η μεταβλητή X είναι συνεχής ή διακριτή με το μέγεθος του δείγματος να είναι μεγάλο, καταφεύγουμε στην ομαδοποίηση των παρατηρήσεων.

Για το σκοπό αυτό μοιράζουμε τα δεδομένα σε ομάδες που λέγονται **κλάσεις** και τοποθετούμε σε κάθε κλάση το πλήθος των παρατηρήσεων που περιέχονται σε αυτή.

Οι κλάσεις είναι διαδοχικά διαστήματα έτσι, ώστε κάθε τιμή της μεταβλητής να ανήκει σε μια μόνο κλάση.

• Συνήθως, οι κλάσεις είναι διαστήματα της μορφής $[\alpha, \beta)$.

• Κάθε κλάση διακρίνεται από τα **όριά** της, δηλαδή το **ελάχιστο** ή **κατώτερο** (α) και το **μέγιστο** ή **ανώτερο** (β).

• **Πλάτος** (c) κάθε κλάσης λέγεται η διαφορά του κατώτερου ορίου από το ανώτερο. Έτσι για την κλάση $\Delta = [\alpha, \beta)$, το πλάτος c είναι ίσο με $c = \beta - \alpha$.

• Αν $\Delta = [\alpha, \beta)$, τότε ο αριθμός $x_i = \dfrac{\alpha + \beta}{2}$ λέγεται **κεντρική τιμή** ή **κέντρο** της κλάσης και λειτουργεί ως αντιπρόσωπος της συγκεκριμένης κλάσης, δηλαδή έχει την ίδια έννοια και χρήση που έχουν οι τιμές x_i μιας διακριτής μεταβλητής.

• Οι κεντρικές τιμές x_i των κλάσεων αποτελούν διαδοχικούς όρους αριθμητικής προόδου με πρώτο όρο x_1 και διαφορά $\omega = c$, έτσι ώστε: $x_\nu = x_1 + (\nu - 1)c$.

Αν όλες οι κλάσεις έχουν το ίδιο πλάτος, η κατανομή λέγεται **ίσου πλάτους**, ενώ όταν το πλάτος μεταβάλλεται, η κατανομή λέγεται **άνισου πλάτους**. Εμείς, στην εργασία αυτή θα ασχοληθούμε με κλάσεις έχουν το ίδιο πλάτος c.

Εύρος (R)

Εύρος R ενός δείγματος λέγεται η διαφορά της μικρότερης τιμής του δείγματος από τη μεγαλύτερη τιμή του δείγματος.

Όπως προαναφέραμε, οι ομάδες (κλάσεις) στις οποίες θα ταξινομηθούν οι τιμές του δείγματος είναι διαδοχικά διαστήματα πραγματικών αριθμών, τα οποία προκύπτουν σύμφωνα με τα εξής βήματα:

Βήμα 1ο: Προσδιορίζουμε το πλήθος κ των κλάσεων.

Αναλόγως προς το μέγεθος του δείγματος, το χωρίζουμε σε κ ομάδες (ή κλάσεις) σύμφωνα με το παρακάτω πίνακα:

Μέγεθος δείγματος ν	< 20	20-50	50-100	100-200
Αριθμός κλάσεων κ	5	6	7	8

Υπάρχει περίπτωση το πλήθος κ των κλάσεων να καθορίζεται εκ των προτέρων, οπότε δεν χρησιμοποιείται ο παραπάνω πίνακας.

Βήμα 2ο: Υπολογίζουμε το πλάτος c των κλάσεων.

- Βρίσκουμε τη μέγιστη $max(x_i) = \alpha_\kappa$ και την ελάχιστη $min(x_i) = \alpha_0$ των τιμών της μεταβλητής X.
- Βρίσκουμε το εύρος $R = max(x_i) - min(x_i) = \alpha_\kappa - \alpha_0$.
- Υπολογίζουμε το πλάτος $c = \frac{R}{\kappa}$ όπου R το εύρος του δείγματος και κ είναι ο αριθμός των κλάσεων που έχει επιλεγεί στο Βήμα 1. Στην περίπτωση που το c δεν είναι φυσικός αριθμός, στρογγυλοποιείται συνήθως προς τα πάνω.

Βήμα 3ο: Κατασκευάζουμε τις κλάσεις.

- Στην ελάχιστη τιμή α_0 της μεταβλητής προσθέτουμε c και έτσι το διάστημα $[\alpha_0, \alpha_0 + c)$, είναι η πρώτη κλάση.
- Οι επόμενες κλάσεις είναι τα διαστήματα:
  $[\alpha_0 + c, \alpha_0 + 2c)$, $[\alpha_0 + 2c, \alpha_0 + 3c)$, …, κ.λπ.

Το πλήθος ν_i των παρατηρήσεων που προκύπτει από τη διαλογή για την i κλάση καλείται **συχνότητα** της κλάσης αυτής (ή συχνότητα της κεντρικής τιμής x_i).

Τονίζουμε ότι στην ομαδοποίηση δεν έχουμε συχνότητες μιας τιμής αλλά συχνότητες κλάσης οι οποίες βρίσκονται όπως ακριβώς έχει αναλυθεί και στις προηγούμενες περιπτώσεις και παρουσιάζονται και εδώ σε πίνακες κατανομής συχνοτήτων.

Ειδικά, ως **αθροιστική συχνότητα** μιας κλάσης ορίζουμε το άθροισμα των συχνοτήτων της κλάσης αυτής και όλων των προηγουμένων της.

Παρατηρούμε ότι μετά την κατάταξη των παρατηρήσεων σε κλάσεις έχουμε χάσει τις ατομικές πληροφορίες. Η απώλεια των πληροφοριών είναι τόσο μεγαλύτερη όσο περισσότερες είναι οι κλάσεις. Αυτό βέβαια είναι ένα μειονέκτημα της ομαδοποίησης, αλλά εξυπηρετεί την μελέτη των πληροφοριών και τη γρήγορη εξαγωγή συμπερασμάτων.

Παράδειγμα 1.8

Οι παρακάτω παρατηρήσεις δίνουν τις τιμές σε € με τις οποίες πωλείται ένα προϊόν σε 40 εμπορικά καταστήματα της χώρας.

157	187	192	175	156	176	161	174	171	167
177	167	162	179	154	173	180	157	175	172
181	**152**	173	180	187	172	**196**	182	182	178
163	171	172	166	169	171	177	161	164	170

Να ομαδοποιηθούν τα παραπάνω δεδομένα σε κλάσεις ίσου πλάτους και να γίνει ο πίνακας κατανομής συχνοτήτων και σχετικών συχνοτήτων.

Λύση

Από τον παραπάνω πίνακα παρατηρούμε ότι οι τιμές της μεταβλητής είναι πολλές και με μικρή συχνότητα η κάθε μια. Γι' αυτό προχωράμε σε ομαδοποίηση των παρατηρήσεων.

- Επειδή, το δείγμα έχει μέγεθος $\nu = 40$, για την ομαδοποίησή του θα χρειαστούμε $\kappa = 6$ κλάσεις.
- Είναι $max(x_i) = 196$ και $min(x_i) = 152$.
- Υπολογίζουμε το εύρος $R = 196–152 = 44$.
- Υπολογίζουμε το πλάτος κάθε κλάσης $c = \frac{R}{\kappa} = \frac{44}{6} = 7{,}333.. \cong 8$.
- Επειδή η μικρότερη τιμή του δείγματος είναι 152, οι κλάσεις στις οποίες θα ταξινομήσουμε τις τιμές του δείγματος είναι τα διαστήματα:
  [152, 160), [160, 168), [168, 176), [176, 184), [184, 192), [192, 200),

Έτσι, μετά τη διαλογή θα έχουμε τον παρακάτω πίνακα:

Πίνακας 1.9

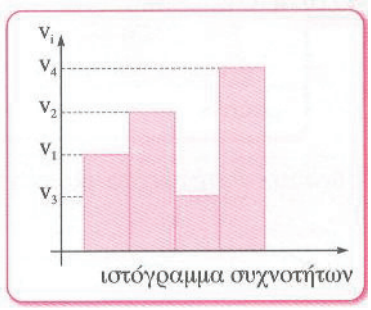
Κατανομές συχνοτήτων για τη μεταβλητή Χ: «τιμές πώλησης σε € ενός προϊόντος»

Κλάσεις [-)	**Κεντρικές τιμές** x_i	**Συχν.** ν_i	**Σχετική συχνότ.** f_i	**Σχετική συχνότητα** $f_i\%$	**Αθροισ. συχνότ.** N_i	**Αθροιστική σχετική συχν.** F_i	**Αθροιστική σχετική συχν.** $F_i\%$
[152-160)	156	5	0,125	12,5	5	0,125	12,5
[160-168)	164	8	0,200	20,0	13	0,325	32,5
[168-176)	172	13	0,325	32,5	26	0,650	65,0
[176-184)	180	10	0,250	25,0	36	0,900	90,0
[184-192)	188	2	0,050	5,0	38	0,950	95,0
[192-200)	196	2	0,050	5,0	40	1,000	100,0
Σύνολο		**40**	**1,000**	**100,0**			

Ιστόγραμμα συχνοτήτων

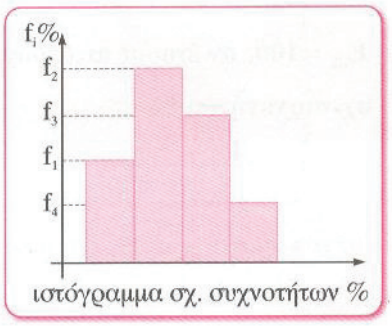
Το ιστόγραμμα χρησιμοποιείται κατά κανόνα για τη γραφική παράσταση ομαδοποιημένων παρατηρήσεων και κατασκευάζεται ως εξής.

- Στον οριζόντιο άξονα ενός συστήματος ορθογωνίων αξόνων σημειώνουμε τα όρια των κλάσεων.
- Κατασκευάζουμε διαδοχικά ορθογώνια παραλληλόγραμμα το καθένα από τα οποία έχει βάση ίση με το πλάτος της κλάσης και ύψος όσο η συχνότητα της κλάσης αυτής.



Εικόνα 1.7
Ιστόγραμμα συχνοτήτων

Με όμοιο τρόπο κατασκευάζεται και το ιστόγραμμα σχετικών συχνοτήτων, με τη διαφορά ότι το ύψος κάθε ορθογωνίου είναι ίσο με την αντίστοιχη σχετική συχνότητα f_i, ή $f_i\%$.



Εικόνα 1.8
Ιστόγραμμα σχ. συχνοτήτων

Παρατηρήσεις

Αν θεωρήσουμε ως μονάδα μέτρησης στον οριζόντιο άξονα το πλάτος c της κάθε κλάσης, τότε κάθε ορθογώνιο του ιστογράμματος συχνοτήτων έχει εμβαδόν ίσο με την αντίστοιχη συχνότητα, ενώ

- το άθροισμα των εμβαδών όλων των ορθογωνίων του ιστογράμματος συχνοτήτων, είναι ίσο με το μέγεθος ν του δείγματος δηλαδή,

$$E = E_1 + E_2 + \ldots + E_\kappa = \nu_1 + \nu_2 + \ldots + \nu_\kappa = \nu$$

- το άθροισμα των εμβαδών όλων των ορθογωνίων του ιστογράμματος σχετικών συχνοτήτων $f_i\%$, είναι

$$E = E_1 + E_2 + \ldots + E_\kappa = f_1\% + f_2\% + \ldots + f_\kappa\% = 100$$

Πολύγωνο συχνοτήτων

Πολύγωνο συχνοτήτων λέγεται η πολυγωνική γραμμή που προκύπτει ως εξής:

- Στο ιστόγραμμα συχνοτήτων τοποθετούμε δύο ακόμα υποθετικές κλάσεις, μια στην αρχή και μια στο τέλος του ιστογράμματος, οι οποίες έχουν συχνότητα μηδέν και πλάτος όσο και οι άλλες κλάσεις.

• Ενώνουμε τα μέσα των άνω βάσεων των ορθογωνίων του ιστογράμματος με τα μέσα των νέων κλάσεων που τοποθετήσαμε.
Με όμοιο τρόπο κατασκευάζεται και το πολύγωνο σχετικών συχνοτήτων.

Παρατηρήσεις

• Το εμβαδόν που περικλείεται μεταξύ του πολυγώνου συχνοτήτων και του οριζόντιου άξονα είναι ίσο με το μέγεθος του δείγματος ν.

•Το εμβαδόν που περικλείεται μεταξύ του πολυγώνου σχετικών συχνοτήτων f_i και του οριζόντιου άξονα είναι ίσο με 1.

• Το εμβαδόν που περικλείεται μεταξύ του πολυγώνου σχετικών συχνοτήτων $f_i\%$ και του οριζόντιου άξονα είναι ίσο με 100.

• Το άθροισμα των εμβαδών όλων των ορθογωνίων του ιστογράμματος συχνοτήτων, ή σχετικών συχνοτήτων, είναι ίσο με το εμβαδόν που περικλείεται από το πολύγωνο συχνοτήτων ή σχετικών συχνοτήτων και τον οριζόντιο άξονα.

Ιστόγραμμα αθροιστικών συχνοτήτων

Για να κατασκευάσουμε το ιστόγραμμα αθροιστικών συχνοτήτων, εργαζόμαστε όπως στην κατασκευή του ιστογράμματος συχνοτήτων. Δηλαδή, στον οριζόντιο άξονα τοποθετούμε τις κλάσεις και σε καθεμία από αυτές σχηματίζουμε ένα ορθογώνιο που έχει βάση ίση με την αντίστοιχη κλάση και ύψος όσο η αθροιστική συχνότητα N_i της κλάσης αυτής.

Με όμοιο τρόπο κατασκευάζεται και το ιστόγραμμα αθροιστικών σχετικών συχνοτήτων, με τη διαφορά ότι το ύψος κάθε ορθογωνίου είναι ίσο με την αντίστοιχη αθροιστική σχετική συχνότητα F_i, ή $F_i\%$.

Πολύγωνο αθροιστικών συχνοτήτων

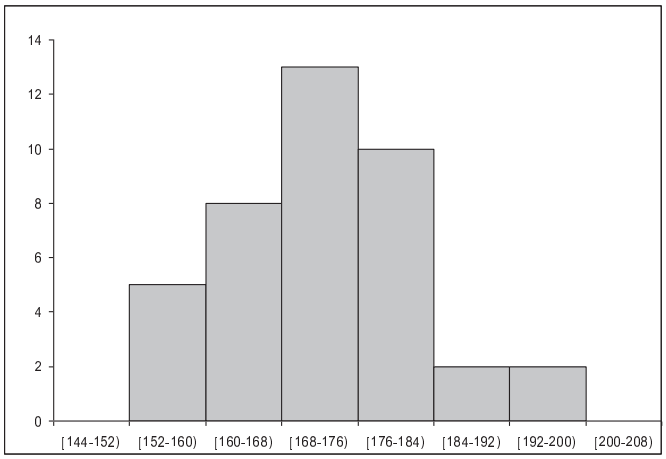
Το πολύγωνο αθροιστικών συχνοτήτων προκύπτει αν ενώσουμε τα δεξιά άκρα των άνω βάσεων των ορθογωνίων με ευθύγραμμα τμήματα.
Με όμοιο τρόπο κατασκευάζεται και το πολύγωνο αθροιστικών σχετικών συχνοτήτων.

Παράδειγμα 1.9

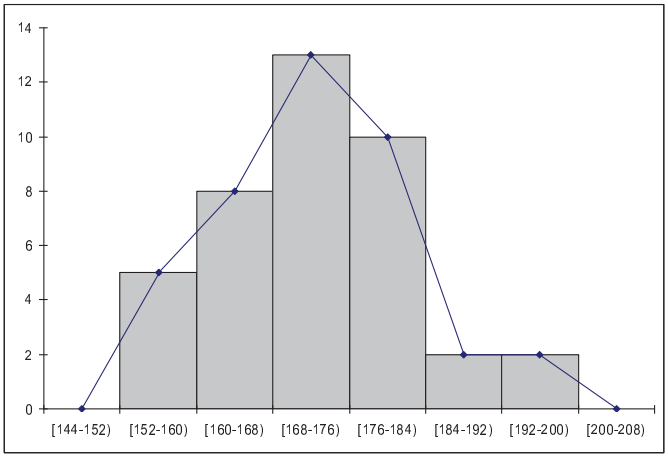
Να κατασκευαστούν: το ιστόγραμμα συχνοτήτων, το πολύγωνο συχνοτήτων, το ιστόγραμμα σχετικών συχνοτήτων, το πολύγωνο σχετικών συχνοτήτων, το ιστόγραμμα αθροιστικών συχνοτήτων, το πολύγωνο αθροιστικών συχνοτήτων, το ιστόγραμμα αθροιστικών σχετικών συχνοτήτων, το πολύγωνο αθροιστικών σχετικών συχνοτήτων για τη μεταβλητή Χ: «τιμές πώλησης σε € ενός προϊόντος» του Παραδείγματος 1.8

Τιμή προϊόντος	[152,160)	[160,168)	[168,176)	[176,184)	[184,192)	[192,200)
Αριθμός καταστημάτων	5	8	13	10	2	2

Εικόνα 1.9
Ιστόγραμμα συχνοτήτων για τη μεταβλητή Χ:
«τιμές πώλησης σε € ενός προϊόντος»



Εικόνα 1.10
Πολύγωνο συχνοτήτων για τη μεταβλητή Χ: «τιμές πώλησης σε € ενός προϊόντος»

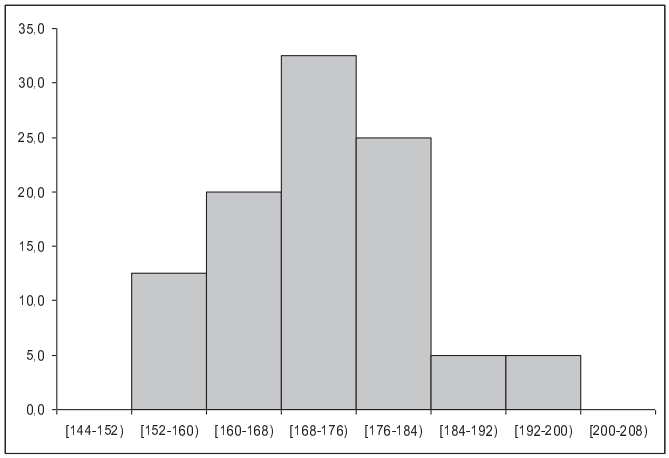


Τιμή προϊόντος	[152,160)	[160,168)	[168,176)	[176,184)	[184,192)	[192,200)
Ποσοστό καταστημ.	12,5	20	32,5	25	5	5

Εικόνα 1.11

Ιστόγραμμα σχετικών συχνοτήτων για τη μεταβλητή Χ:

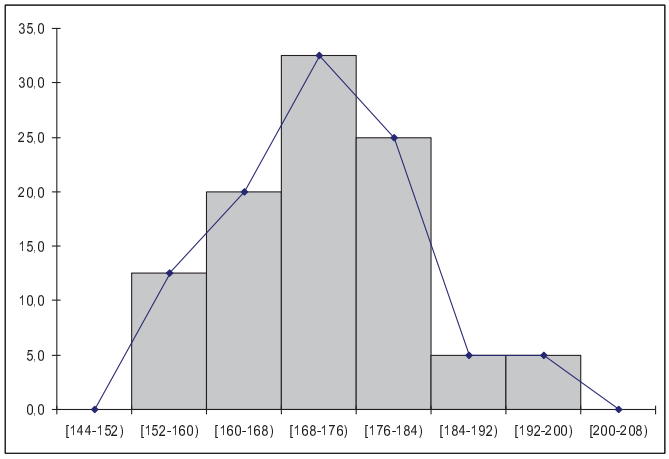
«τιμές πώλησης σε € ενός προϊόντος»



Εικόνα 1.12

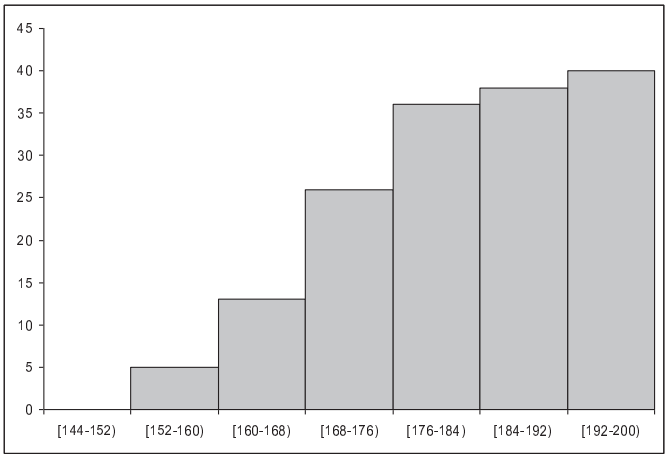
Πολύγωνο σχετικών συχνοτήτων για τη μεταβλητή Χ:

«τιμές πώλησης σε € ενός προϊόντος»

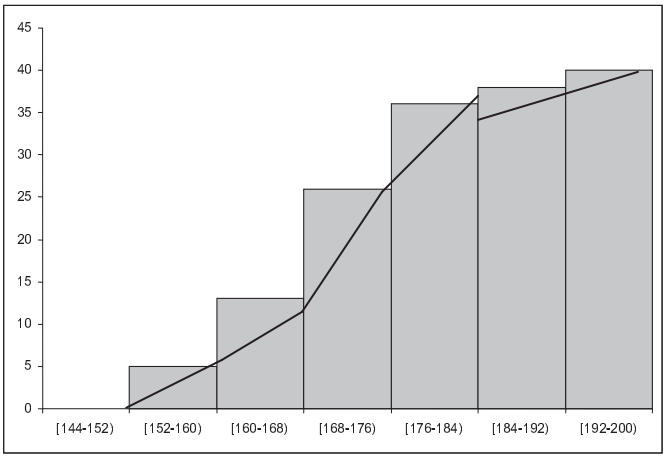


Τιμή προϊόντος	[152,160)	[160,168)	[168,176)	[176,184)	[184,192)	[192,200)
N_i	5	13	26	36	38	40

Εικόνα 1.13
Ιστόγραμμα αθροιστικών συχνοτήτων για τη μεταβλητή Χ:
«τιμές πώλησης σε € ενός προϊόντος»



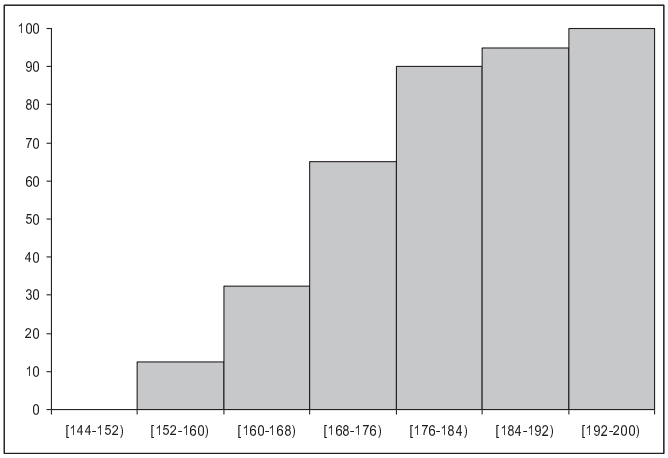
Εικόνα 1.14
Πολύγωνο αθροιστικών συχνοτήτων για τη μεταβλητή Χ:
«τιμές πώλησης σε € ενός προϊόντος»



Τιμή προϊόντος	[152,160)	[160,168)	[168,176)	[176,184)	[184,192)	[192,200)
$F_i\%$	12,5	32,5	65	90	95	100

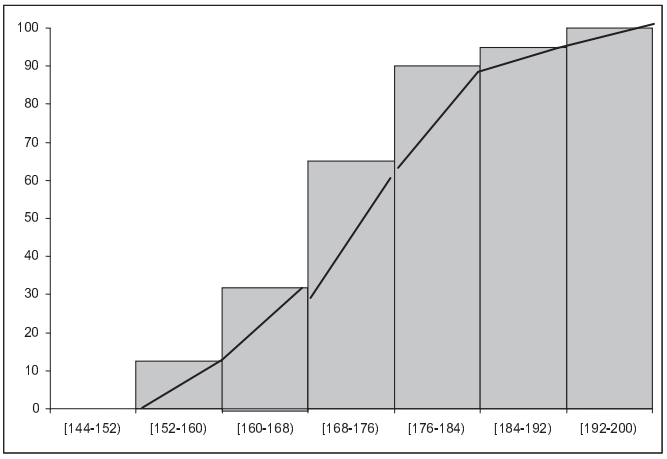
Εικόνα 1.15

Ιστόγραμμα αθροιστικών σχετικών συχνοτήτων για τη μεταβλητή X: «τιμές πώλησης σε € ενός προϊόντος»



Εικόνα 1.16

Πολύγωνο αθροιστικών σχετικών συχνοτήτων για τη μεταβλητή X: «τιμές πώλησης σε € ενός προϊόντος»



Κεφάλαιο 2 Στατιστικά μέτρα

2.1 Μέτρα θέσης ή κεντρικής τάσης

2.1.1 Μέτρα θέσης και διασποράς

Για τη μελέτη ενός πληθυσμού ή ενός δείγματος ως προς μια μεταβλητή X, το πρώτο και βασικό στάδιο είναι η ταξινόμηση και παρουσίαση των στοιχείων με πίνακες συχνοτήτων και στη συνέχεια με γραφική παράσταση έτσι, ώστε να διευκολύνεται η μελέτη των στοιχείων και η εξαγωγή συμπερασμάτων από αυτά.

Όμως, και με τη μορφή αυτή τα στατιστικά στοιχεία εξακολουθούν να παρουσιάζουν σημαντικές δυσκολίες στην απομνημόνευση, αλλά κυρίως δεν διευκολύνουν τις συγκρίσεις στοιχείων από ομοειδείς έρευνες σε διαφορετικούς πληθυσμούς. Για τους λόγους αυτούς παρά την αρχική συνοπτική παρουσίαση των αρχικών δεδομένων, πολλές φορές είναι αναγκαία μια ακόμη περισσότερο συνοπτική παρουσίασή τους. Έτσι λοιπόν, στην περίπτωση που η μεταβλητή είναι ποσοτική, προσπαθούμε να αντικαταστήσουμε τους πίνακες και τις γραφικές παραστάσεις με ορισμένους αντιπροσωπευτικούς αριθμούς οι οποίοι ονομάζονται στατιστικές παράμετροι ή μέτρα της κατανομής.

Ονομάζουμε **στατιστικό μέτρο** της κατανομής τον αριθμό που συνοψίζει βασικά χαρακτηριστικά των παρατηρήσεων του συνόλου των δεδομένων που εξετάζουμε.

Τα μέτρα της κατανομής τα διακρίνουμε σε:

1. **Μέτρα θέσης ή κεντρικής τάσης**
2. **Μέτρα διασποράς**

Τα **μέτρα θέσης** είναι στατιστικές παράμετροι οι οποίες μας πληροφορούν για τη θέση γύρω από την οποία είναι συγκεντρωμένες οι περισσότερες παρατηρήσεις.

Τα **μέτρα διασποράς** είναι αριθμοί οι οποίοι μας πληροφορούν σχετικά με το πόσο συγκεντρωμένες ή διασκορπισμένες είναι οι παρατηρήσεις γύρω από τα μέτρα θέσης.

Τα κυριότερα μέτρα θέσης είναι:

1. **Η μέση τιμή**
2. **Η διάμεσος τιμή**
3. **Τα τεταρτημόρια και**
3. **Η επικρατούσα τιμή**

2.1.2 Μέση τιμή

- **Μέση τιμή των παρατηρήσεων t_i**

Αν οι παρατηρήσεις μιας μεταβλητής X είναι t_1, t_2, …, t_ν ορίζουμε ως **μέση τιμή** ή **αριθμητικό μέσο** των t_i, i = 1, 2, …, ν και συμβολίζουμε με $\bar{x}$, το πηλίκο:

$$\bar{x} = \frac{t_1 + t_2 + \ldots + t_\nu}{\nu} = \frac{\sum_{i=1}^{\nu} t_i}{\nu} = \frac{1}{\nu}\sum_{i=1}^{\nu} t_i \qquad (1)$$

Παρατήρηση

Ο παραπάνω τύπος χρησιμοποιείται σε περίπτωση που το πλήθος των παρατηρήσεων είναι μικρό.

Παράδειγμα 2.1

Η αξία των ετήσιων πωλήσεων δέκα επιχειρήσεων (σε εκατομμύρια €) είναι: 17, 11, 10, 13, 15, 13, 12, 11, 4, 14. Να υπολογιστεί η μέση ετήσια αξία πωλήσεων.

Λύση

Είναι

$$\bar{x} = \frac{t_1 + t_2 + \ldots + t_{10}}{10} = \frac{17+11+10+\ldots+14}{10} = \frac{120}{10} = 12 \text{ εκατομμύρια €.}$$

- **Μέση τιμή από πίνακα συχνοτήτων**

Όταν οι τιμές x_1, x_2, …, x_κ της μεταβλητής X ενός δείγματος μεγέθους ν έχουν συχνότητες ν_1, ν_2, …, ν_κ αντίστοιχα, η **μέση τιμή** $\bar{x}$ υπολογίζεται από τον τύπο:

$$\bar{x} = \frac{x_1\nu_1 + x_2\nu_2 + \ldots + x_\kappa\nu_\kappa}{\nu} = \frac{\sum_{i=1}^{\kappa} x_i\nu_i}{\nu} = \frac{1}{\nu}\sum_{i=1}^{\kappa} x_i\nu_i \qquad (2)$$

Παράδειγμα 2.2

Οι μηνιαίοι μισθοί 40 εργαζομένων μιας επιχείρησης παρουσιάζονται στο διπλανό πίνακα. Να βρεθεί ο μέσος μηνιαίος μισθός των εργαζομένων.

Μισθός (€)	Υπάλληλοι
400	5
500	8
600	16
700	6
800	5
Σύνολο	**40**

Λύση

Είναι $\sum_{i=1}^{5} x_i\nu_i = 400\cdot5 + 500\cdot8 + 600\cdot16 + 700\cdot6 + 800\cdot5 = 23800$.

Άρα, ο μέσος μηνιαίος μισθός των εργαζομένων είναι:

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i \nu_i}{\nu} = \frac{23800}{40} = 595\ €.$$

Παρατήρηση

Στο παραπάνω παράδειγμα, θα μπορούσαμε να συντομεύσουμε τη διαδικασία ως εξής. Στον πίνακα κατανομής συχνοτήτων κατασκευάζουμε μια επιπλέον στήλη αυτή με τα γινόμενα των $x_i \nu_i$, οπότε βρίσκουμε εύκολα το άθροισμα $\sum_{i=1}^{\kappa} x_i \nu_i$ και μετά τη μέση τιμή $\bar{x}$. Δηλαδή.

Μισθός x_i	Συχνότητα ν_i	$x_i \nu_i$
4	5	20
5	8	40
6	16	96
7	6	42
8	5	40
Σύνολο	40	$\sum_{i=1}^{5} x_i \nu_i = 23800$

Συνεπώς, $\bar{x} = \frac{\sum_{i=1}^{5} x_i \nu_i}{\nu} = \frac{23800}{40} = 595\ €.$

- **Μέση τιμή από πίνακα συχνοτήτων ομαδοποιημένης κατανομής**

Όταν οι παρατηρήσεις είναι ομαδοποιημένες, ως τιμές x_i λαμβάνουμε τα κέντρα των κλάσεων $[\alpha, \beta)$, δηλαδή, $x_i = \frac{\alpha + \beta}{2}$.

Παράδειγμα 2.3

Στο διπλανό πίνακα φαίνονται τα έσοδα σε εκατ. € 40 εμπορικών επιχειρήσεων ενός ομίλου κατά τη διάρκεια μια ορισμένης χρονικής περιόδου. Να υπολογιστεί η μέση τιμή των εσόδων των επιχειρήσεων του ομίλου.

Έσοδα εκατ. €	Επιχειρήσεις
[0, 2)	8
[2, 4)	12
[4, 6)	10
[6, 8)	6
[8, 10)	4
Σύνολο	**40**

Λύση

Βρίσκουμε πρώτα τα κέντρα των κλάσεων. Είναι

$$x_1 = \frac{0+2}{2} = \frac{2}{2} = 1.$$

Ομοίως, $x_2 = 3,\ x_3 = 5,\ x_4 = 7$ και $x_5 = 9$.

Στον παρακάτω πίνακα παρουσιάζονται τα κέντρα των παραπάνω κλάσεων. Έχει κατασκευαστεί επίσης μια επιπλέον στήλη με τα γινόμενα $x_i\nu_i$.

Έσοδα εκατ. € [,)	Κέντρα κλάσεων x_i	Συχνότητα ν_i	$x_i\nu_i$
[0, 2)	1	8	8
[2, 4)	3	12	36
[4, 6)	5	10	50
[6, 8)	7	6	42
[8, 10)	9	4	36
Σύνολο		40	$\sum_{i=1}^{5} x_i\nu_i = 172$

Συνεπώς, $\bar{x} = \dfrac{\sum_{i=1}^{5} x_i\nu_i}{\nu} = \dfrac{172}{40} = 4{,}3$ εκατ. €.

- **Μέση τιμή όταν γνωρίζουμε τη σχετική συχνότητα**

Η σχέση $\bar{x} = \dfrac{x_1\nu_1 + x_2\nu_2 + \cdots + x_k\nu_\kappa}{\nu}$, μπορεί ισοδύναμα να γραφεί ως

$$\bar{x} = \frac{x_1\nu_1 + x_2\nu_2 + \cdots + x_k\nu_\kappa}{\nu} = \frac{x_1\nu_1}{\nu} + \frac{x_2\nu_2}{\nu} + \cdots + \frac{x_\kappa\nu_\kappa}{\nu} =$$

$$= x_1\frac{\nu_1}{\nu} + x_2\frac{\nu_2}{\nu} + \cdots + x_\kappa\frac{\nu_\kappa}{\nu} = x_1f_1 + x_2f_2 + \cdots + x_kf_k = \sum_{i=1}^{\kappa} x_if_i$$

Έχουμε δηλαδή:

$$\bar{x} = x_1f_1 + x_2f_2 + \ldots + x_\kappa f_\kappa = \sum_{i=1}^{\kappa} x_if_i \qquad (3)$$

Παράδειγμα 2.4

Στο διπλανό πίνακα παρουσιάζεται η κατανομή σχετικών συχνοτήτων f_i μιας μεταβλητής X. Να υπολογιστεί η μέση τιμή των παρατηρήσεων.

x_i	f_i
2	0,3
4	0,2
6	0,4
8	0,1
Σύνολο	**1,0**

Λύση

Η μέση τιμή των παρατηρήσεων είναι:

$$\bar{x} = \sum_{i=1}^{4} x_if_i = 2\cdot0{,}3 + 4\cdot0{,}2 + 6\cdot0{,}4 + 8\cdot0{,}1 = 4{,}6.$$

- **Μέση τιμή της μεταβλητής $Y = X + \beta$**

Έστω ότι $x_1,\ x_2,\ \ldots,\ x_\nu$ είναι ν παρατηρήσεις μιας μεταβλητής X, που έχουν μέση τιμή $\bar{x}$. Αν $y_1,\ y_2,\ \ldots,\ y_\nu$ είναι οι παρατηρήσεις που προκύπτουν αν αυξήσουμε (ή ελαττώσουμε) κάθε μια από τις $x_1,\ x_2,\ \ldots,\ x_\nu$ κατά μια σταθερά β, δηλαδή,

$$y_1 = x_1 + \beta,\ y_2 = x_2 + \beta,\ \ldots,\ y_\nu = x_\nu + \beta,$$

τότε και η μέση τιμή της νέας μεταβλητής $Y = X + \beta$ με τιμές $y_i = x_i + \beta$ αυξάνεται (ή ελαττώνεται) κατά β, δηλαδή, $\bar{y} = \bar{x} + \beta$. Πράγματι είναι:

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_\nu}{\nu} = \frac{x_1 + \beta + x_2 + \beta + \cdots + x_\nu + \beta}{\nu} =$$

$$= \frac{x_1 + x_2 + \cdots + x_\nu}{\nu} + \frac{\nu\beta}{\nu} = \bar{x} + \beta\ .$$

Δηλαδή, $\bar{y} = \bar{x} + \beta$.

Παράδειγμα 2.5

Οι τιμές 10 βιβλίων σε ευρώ είναι: 15, 9, 6, 18, 21, 6, 18, 27, 9 ,12.

α) Να βρεθεί η μέση τιμή των παραπάνω τιμών.

β) Αν η τιμή κάθε βιβλίου αυξηθεί κατά 0,3 ευρώ, τότε ποια είναι η νέα μέση τιμή;

Λύση

α) Έστω ότι $x_1,\ x_2,\ \ldots,\ x_{10}$ οι τιμές των 10 βιβλίων. Τότε

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_{10}}{10} = \frac{15 + 9 + 6 + \ldots + 12}{10} = \frac{141}{10} = 14{,}1 \text{ ευρώ}$$

β) Επειδή η τιμή του κάθε βιβλίου αυξάνεται κατά 0,3 ευρώ, οι νέες τιμές των βιβλίων είναι $y_1 = x_1 + 0{,}3,\ y_2 = x_2 + 0{,}3,\ \ldots,\ y_{10} = x_{10} + 0{,}3$.

Επομένως, είναι

$\bar{y} = \bar{x} + \beta$ ή $\bar{y} = 14{,}1 + 0{,}3$ ή $\bar{y} = 14{,}4$ ευρώ

- **Μέση τιμή της μεταβλητής $Y = \alpha \cdot X$**

Έστω ότι $x_1,\ x_2,\ \ldots,\ x_\nu$ είναι ν παρατηρήσεις μιας μεταβλητής X που έχουν μέση τιμή $\bar{x}$. Αν $y_1,\ y_2,\ \ldots,\ y_\nu$ είναι οι παρατηρήσεις που προκύπτουν αν πολλαπλασιάσουμε (ή διαιρέσουμε) κάθε μια από τις $x_1,\ x_2,\ \ldots,\ x_\nu$ με τον ίδιο αριθμό α, δηλαδή,

$$y_1 = \alpha \cdot x_1,\ y_2 = \alpha \cdot x_2,\ \ldots,\ y_\nu = \alpha \cdot x_\nu$$

τότε και η μέση τιμή της νέας μεταβλητής $Y = \alpha \cdot X$ με τιμές $y_i = \alpha \cdot x_i$ πολλαπλασιάζεται (ή διαιρείται) με τον ίδιο αριθμό, δηλαδή, $\bar{y} = \alpha \cdot \bar{x}$.

Πράγματι είναι

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_\nu}{\nu} = \frac{\alpha \cdot x_1 + \alpha \cdot x_2 + \cdots + \alpha \cdot x_\nu}{\nu} =$$

$$= \frac{\alpha(x_1 + x_2 + \cdots + x_\nu)}{\nu} = \alpha \frac{x_1 + x_2 + \cdots + x_\nu}{\nu} = \alpha \cdot \bar{x}\,.$$

Δηλαδή, $\bar{y} = \alpha \cdot \bar{x}$.

Παράδειγμα 2.6

Οι τιμές 10 βιβλίων σε ευρώ (χωρίς ΦΠΑ) είναι:

15, 9, 6, 18, 21, 6, 18, 27, 9 ,12.

α) Να βρεθεί η μέση τιμή των παραπάνω τιμών.

β) Αν σε κάθε αρχική τιμή των βιβλίων προσθέσουμε και ΦΠΑ που είναι 20%, τότε ποια είναι η νέα μέση τιμή.

Λύση

α) Έστω ότι x_1, x_2, …, x_{10} είναι οι τιμές των βιβλίων. Τότε

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_{10}}{10} = \frac{15 + 9 + 6 + \ldots + 12}{10} = \frac{141}{10} = 14{,}1 \text{ ευρώ}$$

β) Επειδή η αξία κάθε βιβλίου επιβαρύνεται με το κόστος του ΦΠΑ, η νέα τιμή y_i ενός βιβλίου που είχε αρχική τιμή x_i, θα είναι:

$$y_i = x_i + \frac{20}{100} \cdot x_i \text{ ή } y_i = x_i + 0{,}2 \cdot x_i \text{ ή } y_i = 1{,}2 \cdot x_i$$

Οι νέες τιμές των βιβλίων είναι $y_1 = 1{,}2 \cdot x_1$, $y_2 = 1{,}2 \cdot x_2$, …, $y_{10} = 1{,}2 \cdot x_{10}$

Επομένως, είναι

$$\bar{y} = 1{,}2 \cdot \bar{x} \text{ ή } \bar{y} = 1{,}2 \cdot 14{,}1 \text{ ή } \bar{y} = 16{,}92 \text{ ευρώ}$$

Οι δύο τελευταίες ιδιότητες συμπυκνώνονται στο θεώρημα αλλαγής μεταβλητής.

Θεώρημα αλλαγής μεταβλητής

Έστω $\bar{x}$ η μέση τιμή της μεταβλητής X ως προς την οποία εξετάζουμε ένα δείγμα. Τότε η μέση τιμή $\bar{y}$ της μεταβλητής $Y = \alpha X + \beta$ (α, $\beta \in \mathrm{R}$) είναι $\bar{y} = \alpha\bar{x} + \beta$.

Παρατήρηση

Η μέση τιμή των παρατηρήσεων μιας μεταβλητής X είναι αριθμός μεταξύ της μικρότερης και μεγαλύτερης τιμής της X. Δηλαδή

αν $x_1 \le x_2 \le \ldots \le x_\nu$, τότε $x_1 \le \bar{x} \le x_\nu$

Σταθμικός μέσος

Στις περιπτώσεις που δίνεται διαφορετική βαρύτητα (έμφαση) στις τιμές x_1, x_2, …, x_ν μιας μεταβλητής X, που εκφράζεται με τους λεγόμενους συντελεστές στάθμισης (βαρύτητας) w_1, w_2, …, w_ν, τότε ορίζουμε ως **σταθμικό μέσο** ή **σταθμισμένο αριθμητικό μέσο** το πηλίκο:

$$\bar{x} = \frac{x_1 w_1 + x_2 w_2 + \cdots + x_\nu w_\nu}{w_1 + w_2 + \cdots + w_\nu} = \frac{\sum_{i=1}^{\nu} x_i w_i}{\sum_{i=1}^{\nu} w_i}$$

Παράδειγμα 2.7

Στο διπλανό πίνακα παρουσιάζονται οι βαθμοί x_i ενός μαθητή καθώς και οι αντίστοιχοι συντελεστές w_i στάθμισης των μαθημάτων. Να υπολογιστεί ο σταθμικός μέσος των μαθημάτων.

Βαθμοί μαθημάτων	**Συντελεστές στάθμισης**
10	3
12	2
14	3
16	1
18	1

Λύση

Είναι:

$$\bar{x} = \frac{x_1 w_1 + x_2 w_2 + \cdots + x_\nu w_\nu}{w_1 + w_2 + \cdots + w_\nu} = \frac{10 \cdot 3 + 12 \cdot 2 + 14 \cdot 3 + 16 \cdot 1 + 18 \cdot 1}{3 + 2 + 3 + 1 + 1}$$

$$= \frac{30 + 24 + 42 + 16 + 18}{10} = \frac{130}{10} = 13$$

Άρα, ο σταθμικός μέσος των μαθημάτων είναι $\bar{x} = 13$.

2.1.3 Διάμεσος

◊ **Διάμεσος των παρατηρήσεων t_i**

Αν οι ν παρατηρήσεις t_1, t_2, …, t_ν ενός δείγματος έχουν διαταχθεί σε αύξουσα σειρά, τότε ως **διάμεσο** δ αυτών ορίζουμε:

- τη μεσαία παρατήρηση όταν το ν είναι περιττός αριθμός,
- το ημιάθροισμα των δύο μεσαίων παρατηρήσεων όταν το ν είναι άρτιος αριθμός.

Η διάμεσος ενός δείγματος παρατηρήσεων είναι η τιμή για την οποία το πολύ 50% των παρατηρήσεων είναι μικρότερες από αυτήν και το πολύ 50% των παρατηρήσεων είναι μεγαλύτερες από την τιμή αυτήν.

Επομένως, αν $t_1 \le t_2 \le \ldots \le t_\nu$ είναι οι παρατηρήσεις της μεταβλητής X, τότε:

- $\delta = t_{\frac{\nu+1}{2}}$ όταν ο ν είναι περιττός.
- $\delta = \dfrac{t_{\frac{\nu}{2}} + t_{\frac{\nu}{2}+1}}{2}$ όταν ο ν είναι άρτιος.

Είναι φανερό ότι, στην περίπτωση που ο ν είναι άρτιος, η διάμεσος δ μπορεί να μην είναι ίση με κάποια από τις τιμές t_i.

Παράδειγμα 2.8

α) Αν οι ώρες που εργάσθηκαν 11 υπάλληλοι μιας επιχείρησης σε μια εβδομάδα είναι:

35, 37, 36, 36, 39, 40, 45, 40, 37, 36, 48

να βρεθεί η διάμεσος.

β) Αν οι ώρες που εργάσθηκαν 12 υπάλληλοι μιας επιχείρησης σε μια εβδομάδα είναι:

35, 37, 36, 36, 39, 40, 45, 40, 37, 36, 48, 42

να βρεθεί η διάμεσος.

Λύση

α) Διατάσσουμε κατά αύξουσα σειρά τις παραπάνω τιμές t_i:

35, 36, 36, 36, 37, 37, 39, 40, 40, 45, 48.

Επειδή το πλήθος των τιμών του δείγματος είναι ν = 11, το οποίο είναι περιττός αριθμός η διάμεσος είναι η μεσαία παρατήρηση. Επομένως είναι:

$\delta = t_6 = 37.$

β) Διατάσσουμε κατά αύξουσα σειρά τις παραπάνω τιμές t_i:

35, 36, 36, 36, 37, 37, 39, 40, 40, 42, 45, 48.

Επειδή το πλήθος των τιμών του δείγματος είναι ν = 12, το οποίο είναι άρτιος αριθμός η διάμεσος είναι το ημιάθροισμα των δύο μεσαίων παρατηρήσεων. Επομένως είναι:

$$\delta = \frac{t_6 + t_7}{2} = \frac{37 + 39}{2} = 38.$$

◊ **Διάμεσος από πίνακα συχνοτήτων**

Στην περίπτωση που θέλουμε να βρούμε τη διάμεσο, όταν μας δίνεται ο πίνακας συχνοτήτων εργαζόμαστε όπως στο ακόλουθο παράδειγμα.

Παράδειγμα 2.9

Να βρείτε τη διάμεσο των παρατηρήσεων της μεταβλητής στις επόμενες περιπτώσεις.

α)

x_i	ν_i
2	5
3	14
5	6
6	4
Σύνολο	**29**

β)

x_i	ν_i
1	5
3	10
4	6
6	9
Σύνολο	**30**

Λύση

α) Το μέγεθος του δείγματος είναι $\nu = 29$, δηλαδή περιττός αριθμός, οπότε η διάμεσος δ είναι η μεσαία παρατήρηση

$\delta = t_{\frac{\nu+1}{2}} = t_{15}$.

Θα βρούμε τώρα ποια τιμή έχει η t_{15}.

Παρατηρούμε ότι οι πρώτες 5 παρατηρήσεις έχουν τιμή 2 και ακολουθούν 14 παρατηρήσεις με τιμή 3. Δηλαδή

$t_1 = t_2 = \ldots = t_5 = 2,$

$t_6 = t_7 = \ldots = t_{19} = 3,$

Οπότε η 15η παρατήρηση θα έχει τιμή 3, άρα $\delta = 3$.

β) Το μέγεθος του δείγματος είναι $\nu = 30$, δηλαδή άρτιος αριθμός, οπότε η διάμεσος είναι το ημιάθροισμα των δύο μεσαίων παρατηρήσεων δηλαδή

$$\delta = \frac{t_{\frac{\nu}{2}} + t_{\frac{\nu}{2}+1}}{2} = \frac{t_{15} + t_{16}}{2}.$$

Θα βρούμε τώρα ποια τιμή έχουν οι t_{15} και t_{16}.

Παρατηρούμε ότι οι πρώτες 5 παρατηρήσεις έχουν τιμή 1, ακολουθούν 10 παρατηρήσεις με τιμή 3 και 6 παρατηρήσεις με τιμή 4. Δηλαδή

$t_1 = t_2 = \ldots = t_5 = 1,$

$t_6 = t_7 = \ldots = t_{15} = 3,$

$t_{16} = t_{17} = \ldots = t_{21} = 4,$

Οπότε η 15η παρατήρηση είναι το τελευταίο 3. Η 16η παρατήρηση είναι το πρώτο 4.

Άρα $\delta = \frac{3+4}{2} = 3{,}5.$

◊ Διάμεσος από πίνακα σχετικών συχνοτήτων $f_i\%$

Στην περίπτωση που θέλουμε να βρούμε τη διάμεσο, όταν μας δίνεται ο πίνακας σχετικών συχνοτήτων, χρησιμοποιούμε την ιδιότητα που έχει η διάμεσος να είναι ο αριθμός για τον οποίο το πολύ 50% των παρατηρήσεων είναι μικρότερες από αυτόν και το πολύ 50% των παρατηρήσεων είναι μεγαλύτερες από αυτόν.

Ειδικά, αν για κάποια μεταβλητή υπάρχει τιμή x_i που να έχει αθροιστική συχνότητα $F_i\% = 50$ ή $F_i = 0{,}5$, τότε το πλήθος των παρατηρήσεων είναι άρτιο και η τιμή x_i είναι η πρώτη μεσαία παρατήρηση, οπότε η διάμεσος είναι: $\delta = \frac{x_i + x_{i+1}}{2}$.

Παράδειγμα 2.10

Να βρείτε τη διάμεσο στις επόμενες περιπτώσεις.

α)

x_i	$f_i\%$
2	15
4	35
5	20
6	30

β)

x_i	$f_i\%$
0	15
2	25
3	35
4	25

Λύση

Σε κάθε περίπτωση συμπληρώνουμε τη στήλη $F_i\%$.

α) Παρατηρούμε ότι η τιμή $x_2 = 4$ έχει αθροιστική συχνότητα $F_i\% = 50$. Άρα το πλήθος των παρατηρήσεων είναι άρτιο με πρώτη μεσαία παρατήρηση τη $x_2 = 4$ (το τελευταίο 4) και δεύτερη μεσαία παρατήρηση τη $x_3 = 5$ (το πρώτο 5), οπότε:

x_i	$f_i\%$	$F_i\%$
2	15	15
4	35	50
5	20	70
6	30	100

$$\delta = \frac{4+5}{2} = 4{,}5.$$

β) Το 40% των παρατηρήσεων είναι μικρότερες ή ίσες του 2 και το 75% είναι μικρότερες ή ίσες του 3. Οπότε η διάμεσος είναι $\delta = 3$, αφού αυτή είναι η τιμή της μεταβλητής κάτω από την οποία βρίσκεται το 50% των παρατηρήσεων.

x_i	$f_i\%$	$F_i\%$
0	15	15
2	25	40
3	35	75
4	25	100

◊ **Διάμεσος σε ομαδοποιημένες παρατηρήσεις**

Αν οι παρατηρήσεις είναι ομαδοποιημένες, τότε:

• Προσδιορίζεται αρχικά το διάστημα στο οποίο περιλαμβάνεται η διάμεσος.

• Αφού υπολογιστούν οι αθροιστικές σχετικές συχνότητες κάθε διαστήματος, προσδιορίζεται το πρώτο διάστημα τιμών του οποίου η αθροιστική σχετική συχνότητα είναι μεγαλύτερη ή ίση από το 50%. Στο εσωτερικό αυτού του διαστήματος βρίσκεται η διάμεσος.

• Για τον ακριβή υπολογισμό της διαμέσου χρησιμοποιείται ο τύπος:

$$\delta = L_1 + \frac{\frac{\nu}{2} - C}{\nu_i} \cdot (L_2 - L_1)$$

όπου

L_1: το πραγματικό κατώτερο όριο της τάξης που περιλαμβάνει τη διάμεσο

L_2: το πραγματικό ανώτερο όριο της τάξης που περιλαμβάνει τη διάμεσο

ν: ο συνολικός αριθμός των παρατηρήσεων

C: η αθροιστική συχνότητα της τάξης που βρίσκεται ακριβώς πριν από αυτή στην οποία βρίσκεται η διάμεσος

ν_i: η συχνότητα της τάξης που περιλαμβάνει τη διάμεσο

Παράδειγμα 2.11

Η κατανομή ηλικιών 895 ατόμων είναι η εξής:

Ηλικία σε χρόνια	Συχνότητα ν_i	Αθροιστική συχνότητα N_i	Σχετική συχνότητα $f_i\%$	Σχετική αθροιστική συχνότητα $F_i\%$
[30, 39)	206	206	23,0	23,0
[40, 49)	274	480	30,6	53,6
[50, 59)	194	674	21,7	75,3
[60, 69)	178	852	19,9	95,2
[70, 79)	43	895	4,8	100,0
Σύνολο	895		100,0	

Να υπολογιστεί η διάμεσος.

Λύση

Προσδιορίζουμε το πρώτο διάστημα ηλικιών του οποίου η σχετική αθροιστική συχνότητα είναι μεγαλύτερη ή ίση με 50%. Το διάστημα αυτό είναι το [40, 49) με πραγματικά όρια $L_1 = 39{,}95$ και $L_2 = 49{,}95$

Άρα $\delta = L_1 + \frac{\frac{\nu}{2} - C}{\nu_i} \cdot (L_2 - L_1) = 40 + \frac{\frac{895}{2} - 206}{274}(49{,}95 - 39{,}95) = 48{,}7$

Ιδιότητες της διαμέσου

Έστω ότι οι παρατηρήσεις t_1, t_2, …, t_ν έχουν διάμεσο δ. Τότε οι παρατηρήσεις:

- $t_1+\beta$, $t_2+\beta$, …, $t_\nu+\beta$, έχουν διάμεσο $\delta+\beta$,
- $\alpha\cdot t_1$, $\alpha\cdot t_2$, …, $\alpha\cdot t_\nu$, έχουν διάμεσο $\alpha\cdot\delta$,
- $\alpha\cdot t_1+\beta$, $\alpha\cdot t_2+\beta$, …, $\alpha\cdot t_\nu+\beta$, έχουν διάμεσο $\alpha\cdot\delta+\beta$.

2.1.4 Εκατοστημόρια (P_κ) – Τεταρτημόρια (Q_κ)

Ορίζουμε ως P_κ, κ =1, 2, ..., 99 **εκατοστημόριο** ενός συνόλου παρατηρήσεων, οι οποίες έχουν διαταχθεί σε αύξουσα φυσική σειρά, την τιμή εκείνη για την οποία το πολύ $\kappa\%$ των παρατηρήσεων είναι μικρότερες του P_κ και το πολύ $(100-\kappa)\%$ των παρατηρήσεων είναι μεγαλύτερες από την τιμή αυτή.

Ειδική περίπτωση εκατοστημορίων είναι τα P_{25}, P_{50}, P_{75} τα οποία ονομάζονται τεταρτημόρια και συμβολίζονται με Q_1, Q_2 και Q_3, αντίστοιχα.

Επομένως, $Q_1 = P_{25}$ είναι η τιμή εκείνη για την οποία το 25% των παρατηρήσεων είναι μικρότερες και το 75% είναι μεγαλύτερες από αυτή.

Προφανώς είναι $\delta = Q_2 = P_{50}$.

Συχνά για ευκολία ο υπολογισμός των τεταρτημορίων Q_1 και Q_3 ενός συνόλου δεδομένων γίνεται κατά προσέγγιση υπολογίζοντας τις διαμέσους του πρώτου και του δεύτερου μισού των διατεταγμένων παρατηρήσεων, αντίστοιχα. Για παράδειγμα, προκειμένου να υπολογίσουμε τα τεταρτημόρια των δεδομένων

3, 4, 0, 6, 5, 8, 1, 1, 6, 1, 2, 8, 9,

εργαζόμαστε ως εξής:

- Διατάσσουμε τις παρατηρήσεις σε αύξουσα σειρά μεγέθους:

Έχουμε ν = 13 παρατηρήσεις, οι οποίες σε αύξουσα σειρά είναι:

0 1 1 1 2 3 4 5 6 6 8 8 9.

- Υπολογίζουμε τη διάμεσο, όπως προαναφέραμε:

Η διάμεσος είναι η έβδομη στη σειρά παρατήρηση, δηλαδή δ = 4.

- Υπολογίζουμε τη διάμεσο του πρώτου μισού των διατεταγμένων παρατηρήσεων, δηλαδή των παρατηρήσεων που είναι αριστερά του δ. Η τιμή αυτή είναι το Q_1:

Η διάμεσος των παρατηρήσεων που είναι αριστερά του δ, δηλαδή των

0 1 1 1 2 3, είναι το $Q_1 = \dfrac{1+1}{2} = 1$.

- Υπολογίζουμε τη διάμεσο του δεύτερου μισού των διατεταγμένων παρατηρήσεων,

δηλαδή των παρατηρήσεων που είναι δεξιά του δ. Η τιμή αυτή είναι το Q_3.

Η διάμεσος των παρατηρήσεων που είναι δεξιά του δ, δηλαδή των 5 6 6 8 8 9, είναι το

$$Q_3 = \frac{6+8}{2} = 7.$$

Άλλος τρόπος υπολογισμού των εκατοστημορίων είναι ο εξής:

Όταν η ποσότητα $\frac{\nu\kappa}{100}$ είναι ακέραιος αριθμός, τότε το κ-εκατοστημόριο είναι ο μέσος όρος των τιμών που καταλαμβάνουν τις θέσεις $\frac{\nu\kappa}{100}$ και $\left[\frac{\nu\kappa}{100}+1\right]$ στη διάταξη των μετρήσεων.

Αν η ποσότητα $\frac{\nu\kappa}{100}$ δεν είναι ακέραιος αριθμός, τότε το κ-εκατοστημόριο είναι η τιμή που καταλαμβάνει τη θέση $(m+1)$ στη διάταξη των μετρήσεων, όπου m ο ακέραιος αριθμός, ο αμέσως μικρότερος του $\frac{\nu\kappa}{100}$.

Εκατοστημόρια σε ομαδοποιημένα δεδομένα

Ο προσδιορισμός των εκατοστημορίων σε ομαδοποιημένα δεδομένα γίνεται με αντίστοιχο τρόπο αυτού της διαμέσου. Αφού πρώτα οριστεί, με τη βοήθεια των αθροιστικών συχνοτήτων, το διάστημα που περιλαμβάνει το ζητούμενο κάθε φορά εκατοστημόριο, στη συνέχεια η τιμή του εκατοστημορίου υπολογίζεται ακριβώς με τη βοήθεια του τύπου:

$$P_\kappa = L_1 + \frac{\frac{\kappa\cdot\nu}{100} - C}{\nu_i}\cdot(L_2 - L_1)$$

L_1: το πραγματικό κατώτερο όριο του διαστήματος που περιλαμβάνει το P_κ

L_2: το πραγματικό ανώτερο όριο του διαστήματος που περιλαμβάνει το P_κ

ν: ο συνολικός αριθμός των παρατηρήσεων

C: η αθροιστική συχνότητα του διαστήματος που βρίσκεται ακριβώς πριν από αυτή στην οποία βρίσκεται το P_κ

ν_i: η συχνότητα του διαστήματος που περιλαμβάνει το P_κ

Παράδειγμα 2.12

Για τα δεδομένα του Παραδείγματος 2.11 θα βρούμε το 75^o τεταρτημόριο

Λύση

Το διάστημα που περιλαμβάνει το 75^o τεταρτημόριο είναι το [50, 59)

$$P_{75} = L_1 + \frac{\frac{75 \cdot \nu}{100} - C}{\nu_i} \cdot (L_2 - L_1) = 50 + \frac{\frac{75 \cdot 895}{100} - 480}{194}(59{,}95 - 49{,}95) = 59{,}9$$

2.1.5 Επικρατούσα Τιμή (M_0)

Επικρατούσα τιμή ή **κορυφή** M_0 ορίζεται ως η τιμή που έχει τη μεγαλύτερη συχνότητα.

Εξ ορισμού, όταν οι τιμές της μεταβλητής εμφανίζονται μόνο μια φορά, δεν υπάρχει επικρατούσα τιμή.
Είναι φανερό ότι η επικρατούσα τιμή μπορεί να μην είναι μοναδική. Αν μια κατανομή παρατηρήσεων έχει μια μόνο επικρατούσα τιμή, ονομάζεται μονοκόρυφη, ενώ αν έχει δύο επικρατούσες τιμές, ονομάζεται δικόρυφη κ.ο.κ.

Επικρατούσα τιμή σε ομαδοποιημένα δεδομένα
Αρχικά προσδιορίζεται το διάστημα με τη μεγαλύτερη συχνότητα τιμών (επικρατούσα τάξη). Ο ακριβής προσδιορισμός της επικρατούσας τιμής στο εσωτερικό αυτού του διαστήματος, μπορεί να γίνει κάνοντας την παραδοχή ότι η κατανομή των τιμών στο εσωτερικό του είναι ομοιόμορφη. Σε μια τέτοια περίπτωση, ως επικρατούσα τιμή (M_0) της κατανομής μπορεί να θεωρηθεί η κεντρική τιμή της επικρατούσας τάξης

$$M_0 = \frac{L_1 + L_2}{2}$$

όπου L_1 και L_2 το ανώτερο και το κατώτερο όριο της επικρατούσας τάξης.

Εναλλακτικά, χωρίς την παραδοχή της ομοιόμορφης κατανομής των τιμών, ο προσδιορισμός της επικρατούσας τιμής μπορεί να γίνει ορίζοντας την ακριβή θέση της με παρεμβολή, η οποία βασίζεται στις συχνότητες των δύο τάξεων που βρίσκονται ακριβώς πριν και μετά την επικρατούσα. Στην περίπτωση αυτή, η επικρατούσα τιμή ορίζεται με τη βοήθεια της σχέσης:

$$M_0 = L_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot (L_2 - L_1)$$

L_1: το πραγματικό κατώτερο όριο της επικρατούσας τάξης

L_2: το πραγματικό ανώτερο όριο της επικρατούσας τάξης

Δ_1: η διαφορά της συχνότητας της επικρατούσας τάξης από την αμέσως προηγούμενη αυτής

Δ_2: η διαφορά της συχνότητας της επικρατούσας τάξης από την αμέσως επόμενη αυτής

Παράδειγμα 2.13

Από τους ισολογισμούς 40 επιχειρήσεων πήραμε τα παρακάτω στοιχεία, που είναι οι οφειλές (υποχρεώσεις) αυτών των επιχειρήσεων στους προμηθευτές τους, σε χιλιάδες ευρώ:

2, 4, 3, 7, 18, 16, 13, 62, 66, 68, 68, 74, 76, 76, 77, 51, 52, 54, 54, 56, 58, 22, 26, 26, 28, 32, 32, 33, 34, 37, 38, 38, 42, 44, 46, 48, 48, 48, 48, 49.

α) Να βρεθεί η επικρατούσα τιμή, β) να ομαδοποιηθούν τα παραπάνω δεδομένα σε κλάσεις ίσου πλάτους, γ) να υπολογισθεί η μέση τιμή και δ) να υπολογιστούν η διάμεσος, το πρώτο και το τρίτο τεταρτημόριο.

Λύση

α) Η επικρατούσα τιμή είναι 48, η οποία έχει εμφανιστεί τις περισσότερες φορές από κάθε άλλη τιμή στις τιμές της μεταβλητής που εξετάζουμε. Δηλαδή, 4 επιχειρήσεις οφείλουν από 48.000 ευρώ η κάθε μια.

β) Ομαδοποιούμε τα δεδομένα σε οκτώ κλάσεις ίσου πλάτους c = 10 και σχηματίζουμε τις δύο πρώτες στήλες του παρακάτω πίνακα.

γ) Για τον υπολογισμό της μέσης τιμής κατασκευάζουμε τη στήλη των κέντρων των κλάσεων (x_i) και τη στήλη με τα γινόμενα των $x_i \nu_i$.

Κλάσεις οφειλών [,)	Κέντρα κλάσεων x_i	Συχνότητα ν_i	$x_i \nu_i$
[0, 10)	5	4	20
[10, 20)	15	3	45
[20, 30)	25	4	100
[30, 40)	35	7	245
[40, 50)	45	8	360
[50, 60)	55	6	330
[60, 70)	65	4	260
[70, 80)	75	4	300
Σύνολο		$\nu = 40$	$\sum_{i=1}^{8} x_i \nu_i = 1660$

Συνεπώς, $\bar{x} = \frac{\sum_{i=1}^{8} x_i \nu_i}{\nu} = \frac{1660}{40} = 41{,}5$

Δηλαδή, η μέση τιμή των οφειλών είναι $\bar{x} = 41.500$ ευρώ

δ) Για τον υπολογισμό της διαμέσου και των τεταρτημορίων κατασκευάζουμε τον πίνακα αθροιστικών συχνοτήτων και σχετικών αθροιστικών συχνοτήτων.

Κλάσεις οφειλών	Συχνότητα ν_i	Αθροιστική συχνότητα N_i	Σχετική συχνότητα $f_i\%$	Σχετική αθροιστική συχνότητα $F_i\%$
[0, 10)	4	4	10,0	10,0
[10, 20)	3	7	7,5	17,5
[20, 30)	4	11	10,0	27,5
[30, 40)	7	18	17,5	45,0
[40, 50)	8	26	20,0	65,0
[50, 60)	6	32	15,0	80,0
[60, 70)	4	36	10,0	90,0
[70, 80)	4	40	10,0	100,0
Σύνολο	$\nu = 40$		100,0	

Προσδιορίζουμε το πρώτο διάστημα ηλικιών του οποίου η σχετική αθροιστική συχνότητα είναι μεγαλύτερη ή ίση με 50%. Το διάστημα αυτό είναι το [40, 50) με όρια $L_1 = 40$ και $L_2 = 50$

Η αθροιστική συχνότητα της τάξης που βρίσκεται ακριβώς πριν από αυτή στην οποία βρίσκεται η διάμεσος είναι $C = 18$.

Η συχνότητα της τάξης που περιλαμβάνει τη διάμεσο είναι $\nu_i = 8$

Ο ακριβής υπολογισμός της διαμέσου μπορεί να γίνει με χρήση του τύπου:

$$\delta = L_1 + \frac{\frac{\nu}{2} - C}{\nu_i} \cdot (L_2 - L_1) = 40 + \frac{\frac{40}{2} - 18}{8} \cdot (50 - 40) = 42,5$$

Δηλαδή, η διάμεσος των οφειλών είναι $\delta = 42.500$ ευρώ

Με παρόμοιο τρόπο θα υπολογίσουμε τα τεταρτημόρια.

Για το Q_1 είναι $L_1 = 20$ $L_2 = 30$, $C = 7$ και $\nu_i = 4$

$$Q_1 = L_1 + \frac{\frac{25 \cdot \nu}{100} - C}{\nu_i} \cdot (L_2 - L_1) = 20 + \frac{\frac{25 \cdot 40}{100} - 7}{4} \cdot (30 - 20) = 27,5$$

Για το Q_3 είναι $L_1 = 50$ $L_2 = 60$, $C = 26$ και $\nu_i = 6$

$$Q_3 = L_1 + \frac{\frac{75 \cdot \nu}{100} - C}{\nu_i} \cdot (L_2 - L_1) = 50 + \frac{\frac{75 \cdot 40}{100} - 26}{6} \cdot (60 - 50) = 56,7$$

Από τον υπολογισμό των παραπάνω μέτρων θέσης παρατηρούμε ότι η μέση τιμή των οφειλών των 40 επιχειρήσεων στους προμηθευτές τους είναι 41.500 ευρώ. Το 50% των επιχειρήσεων οφείλει μέχρι και 42.500 ευρώ στους προμηθευτές. Το 25% των επιχειρήσεων οφείλει μέχρι και 27.500, ενώ υπάρχει και ένα 25% των επιχειρήσεων που οφείλει περισσότερο από 56.700 ευρώ.

2.1.6 Σύγκριση μέτρων θέσης

Πλεονεκτήματα και μειονεκτήματα της μέσης τιμής

Πλεονεκτήματα	Μειονεκτήματα
Για τον υπολογισμό της χρησιμοποιούνται όλες οι τιμές του δείγματος.	Επηρεάζεται πολύ από ακραίες τιμές.
Είναι μοναδική σε κάθε δείγμα.	Ενδέχεται να μην αντιστοιχεί σε κάποια τιμή της μεταβλητής.
Ο υπολογισμός της είναι σχετικά εύκολος.	Ενδέχεται να μην είναι ακέραιος σε περίπτωση διακριτής μεταβλητής.
Είναι πολύ χρήσιμη για περαιτέρω στατιστική ανάλυση.	Δεν υπολογίζεται για ποιοτικά δεδομένα.

Πλεονεκτήματα και μειονεκτήματα της διαμέσου

Πλεονεκτήματα	Μειονεκτήματα
Είναι εύκολα κατανοητή.	Δεν συμμετέχουν όλες οι τιμές στον προσδιορισμό της.
Δεν επηρεάζεται πολύ από ακραίες τιμές.	Είναι δύσκολη η εφαρμογή της σε περαιτέρω στατιστική ανάλυση.
Ο υπολογισμός της είναι απλός.	Ενδέχεται να μην αντιστοιχεί σε κάποια τιμή του δείγματος.
Είναι μοναδική σε κάθε δείγμα.	Δεν υπολογίζεται για ποιοτικά δεδομένα.

2.2 Μέτρα διασποράς

Ας υποθέσουμε ότι έχουμε τις τιμές δύο κατηγοριών προϊόντων:

1η κατηγορία: 7 9 9 9 14 14 15 20 20

2η κατηγορία: 10 11 12 12 14 14 14 14 14 15

Οι τιμές των δύο κατηγοριών συμβαίνει να έχουν την ίδια μέση τιμή $\bar{x}$ = 13 και επίσης την ίδια διάμεσο δ = 14. Επομένως, οι πληροφορίες που μας δίνουν οι δύο αυτές παράμετροι είναι ανεπαρκείς για την εξαγωγή κάποιων συμπερασμάτων ώστε να γίνει σύγκριση των τιμών των δύο κατηγοριών.

Αν παρατηρήσουμε, όμως, πιο προσεκτικά τις τιμές των δύο κατηγοριών, θα δούμε ότι οι τιμές της 1ης κατηγορίας, που κυμαίνονται από 7 έως 20, παρουσιάζουν μεγάλη ανομοιογένεια σε αντίθεση με τις τιμές της 2ης κατηγορίας, που κυμαίνονται από 11 έως 15, και παρουσιάζει μεγάλη ομοιογένεια.

Πιο συγκεκριμένα οι τιμές των δύο κατηγοριών προϊόντων έχουν διαφορετικό βαθμό διασποράς των δεδομένων γύρω από κάποιο μέτρο θέσης, συνήθως τη μέση τιμή, δηλαδή τα δεδομένα είναι λιγότερο ή περισσότερο «απλωμένα» γύρω από τη μέση τιμή.

Είναι λοιπόν φανερό ότι για να μπορούμε να εξάγουμε αξιόπιστα συμπεράσματα από τη μελέτη μιας σειράς δεδομένων είναι αναγκαία η χρησιμοποίηση ενός ή περισσοτέρων δεικτών που μας δίνουν το βαθμό συγκέντρωσης ή διασποράς των τιμών της μεταβλητής από τη μέση τιμή. Οι δείκτες αυτοί λέγονται μέτρα διασποράς. Δηλαδή.

Μέτρα διασποράς είναι αριθμοί οι οποίοι μας πληροφορούν σχετικά με το πόσο συγκεντρωμένες ή διασκορπισμένες είναι οι παρατηρήσεις γύρω από τα μέτρα θέσης.

Τα σημαντικότερα μέτρα διασποράς είναι:

1) Το εύρος (R)

2) Η διακύμανση (s^2)

3) Η τυπική απόκλιση (s) και

4) Ο συντελεστής μεταβολής (CV).

2.2.1 Εύρος (R)

• **Εύρος μεταβολής** ή **κύμανση** (R) σε ένα δείγμα λέγεται η διαφορά της μικρότερης παρατήρησης από τη μεγάλη παρατήρηση.

R = μεγαλύτερη παρατήρηση – μικρότερη παρατήρηση

• Όταν έχουμε ομαδοποιημένα δεδομένα, το εύρος δίνεται από τη διαφορά του κατώτερου ορίου της πρώτης κλάσης από το ανώτερο όριο της τελευταίας κλάσης.

2.2.2 Διακύμανση (s^2) των παρατηρήσεων t_i

Ορίζουμε ως **διακύμανση ή διασπορά (s^2)** της μεταβλητής X, τη μέση τιμή των τετραγώνων των αποκλίσεων των παρατηρήσεων t_i από τη μέση τιμή τους $\bar{x}$, δηλαδή

$$s^2 = \frac{(t_1 - \bar{x})^2 + (t_2 - \bar{x})^2 + \ldots + (t_\nu - \bar{x})^2}{\nu} = \frac{1}{\nu}\sum_{i=1}^{\nu}(t_i - \bar{x})^2 \qquad (1)$$

Αποδεικνύεται ότι ο τύπος (1) μπορεί να πάρει την ισοδύναμη μορφή:

$$s^2 = \frac{1}{\nu}\left[\sum_{i=1}^{\nu} t_i^2 - \frac{1}{\nu}\left(\sum_{i=1}^{\nu} t_i\right)^2\right] \qquad (2)$$

που διευκολύνει σημαντικά τους υπολογισμούς παρακάμπτοντας τη μέση τιμή $\bar{x}$, ειδικά όταν η μέση τιμή $\bar{x}$ δεν είναι ακέραιος.

Παράδειγμα 2.14

Δίνονται οι παρατηρήσεις 3, 9 12, 15, 21 μιας μεταβλητής X. Να υπολογίσετε τη διακύμανση αυτών.

Λύση

Η μέση τιμή $\bar{x}$ των παρατηρήσεων αυτών είναι:

$$\bar{x} = \frac{3+9+12+15+21}{5} = \frac{60}{5} = 12$$

Επομένως, η διακύμανση των παρατηρήσεων αυτών είναι:

$$s^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}(t_i - \bar{x})^2 = \frac{1}{5}[(3-12)^2 + (9-12)^2 + (12-12)^2 + (15-12)^2 + (21-12)^2] =$$

$$= \frac{1}{5}(81 + 9 + 0 + 9 + 81) = 36.$$

Διακύμανση (s^2) από πίνακα συχνοτήτων ν_i

Όταν έχουμε πίνακα συχνοτήτων ή ομαδοποιημένες παρατηρήσεις, η διακύμανση s^2 δίνεται από τον τύπο:

$$s^2 = \frac{1}{\nu}\sum_{i=1}^{\kappa}(x_i - \bar{x})^2 \cdot \nu_i \qquad (3)$$

Αποδεικνύεται ότι ο τύπος (3) μπορεί να πάρει την ισοδύναμη μορφή:

$$s^2 = \frac{1}{\nu}\left[\sum_{i=1}^{\kappa} x_i^2 \nu_i - \frac{1}{\nu}\left(\sum_{i=1}^{\kappa} x_i \nu_i\right)^2\right] \qquad (4)$$

όπου x_i, i = 1, 2, …, κ οι τιμές της μεταβλητής ή οι κεντρικές τιμές των κλάσεων και ν_i οι αντίστοιχες συχνότητες.

Παράδειγμα 2.15

Στο διπλανό πίνακα παρουσιάζονται οι ημέρες που απουσίαζαν οι υπάλληλοι μιας εταιρείας κατά τη διάρκεια ενός μήνα. Να υπολογιστεί η διακύμανση της κατανομής αυτής.

Ημέρες x_i	Υπάλληλοι ν_i
1	2
2	3
3	4
4	1
Σύνολο	**10**

Λύση

Θα χρησιμοποιήσουμε τον τύπο (4). Οπότε, στον πίνακα κατανομής συχνοτήτων που δίνεται, προσθέτουμε τρεις επιπλέον στήλες. Μια στήλη με τα τετράγωνα των x_i, μια με τα γινόμενα των $x_i\nu_i$ και μια με τα γινόμενα των $x_i^2\nu_i$. Έτσι βρίσκουμε εύκολα τα αθροίσματα $\sum_{i=1}^{\kappa} x_i\nu_i$ και $\sum_{i=1}^{\kappa} x_i^2\nu_i$.

Ημέρες x_i	Υπάλληλοι ν_i	x_i^2	$x_i\nu_i$	$x_i^2\nu_i$
1	2	1	2	2
2	3	4	6	12
3	4	9	12	36
4	1	16	4	16
Σύνολο	**10**		$\sum_{i=1}^{4} x_i\nu_i = 24$	$\sum_{i=1}^{4} x_i^2\nu_i = 66$

Επομένως,

$$s^2 = \frac{1}{10}\left[\sum_{i=1}^{4} x_i^2\nu_i - \frac{1}{10}\left(\sum_{i=1}^{4} x_i\nu_i\right)^2\right] = \frac{1}{10}\cdot[66 - \frac{1}{10}\cdot 24^2] =$$

$$= \frac{1}{10}\cdot[66 - \frac{1}{10}\cdot 576] = \frac{1}{10}\cdot[66 - 57{,}6] = \frac{1}{10}\cdot 8{,}4 = 0{,}84$$

Παράδειγμα 2.16

Να υπολογιστεί η μέση τιμή $\bar{x}$ και η διακύμανση s^2 της κατανομής:

$[-)$	[0, 5)	[5, 10)	[10, 15)	[15, 20)	[20, 25)	[25, 30)
ν_i	10	25	30	10	15	10

Λύση

Βρίσκουμε πρώτα τα κέντρα των κλάσεων. Είναι:

$$x_1 = \frac{0+5}{2} = \frac{5}{2} = 2{,}5.$$

Ομοίως, $x_2 = 7{,}5$, $x_3 = 12{,}5$, $x_4 = 17{,}5$, $x_5 = 22{,}5$ και $x_6 = 27{,}5$.

Στον παρακάτω πίνακα συχνοτήτων παρουσιάζονται τα κέντρα x_i των παραπάνω κλάσεων, καθώς και τα τετράγωνά τους x_i^2. Έχουν κατασκευαστεί επίσης δύο επιπλέον στήλες, αυτή με τα γινόμενα των $x_i\nu_i$ και αυτή με τα γινόμενα των $x_i^2\nu_i$.

Κλάσεις [–)	Κέντρα κλάσης x_i	x_i^2	Συχνότητα ν_i	$x_i\nu_i$	$x_i^2\nu_i$
[0, 5)	2,5	6,25	10	25,0	62,50
[5, 10)	7,5	56,25	25	187,5	1406,25
[10, 15)	12,5	156,25	30	375,0	4687,50
[15, 20)	17,5	306,25	10	175,0	3062,50
[20, 25)	22,5	506,25	15	337,5	7593,75
[25, 30)	27,5	756,25	10	275,0	7562,50
Σύνολο			100	$\sum_{i=1}^{6} x_i\nu_i = 1.375$	$\sum_{i=1}^{6} x_i^2\nu_i = 24.375$

Επομένως,

$$\bar{x} = \frac{1}{100}\sum_{i=1}^{6} x_i\nu_i = \frac{1}{100}\cdot 1375 = 13,75$$

$$s^2 = \frac{1}{100}\left[\sum_{i=1}^{6} x_i^2\nu_i - \frac{1}{100}\left(\sum_{i=1}^{6} x_i\nu_i\right)^2\right] = \frac{1}{100}\cdot[24.375 - \frac{1}{100}\cdot 1.375^2] =$$

$$= \frac{1}{100}\cdot[24.375 - \frac{1}{100}\cdot 1.890.625] = \frac{1}{100}\cdot[24.375 - 18.906,25] =$$

$$= \frac{1}{100}\cdot 5.468,75 = 54,6875$$

Διακύμανση (s^2) από πίνακα σχετικών συχνοτήτων f_i

Ο τύπος $s^2 = \frac{1}{\nu}\sum_{i=1}^{\kappa}(x_i - \bar{x})^2\cdot\nu_i$ μπορεί να μετασχηματισθεί στον τύπο:

$$s^2 = \sum_{i=1}^{\kappa}(x_i - \bar{x})^2\cdot f_i \qquad (5)$$

Επίσης, ο τύπος $s^2 = \frac{1}{\nu}\left[\sum_{i=1}^{\kappa} x_i^2\nu_i - \frac{1}{\nu}\left(\sum_{i=1}^{\kappa} x_i\nu_i\right)^2\right]$ μπορεί να μετασχηματιστεί στον τύπο:

$$s^2 = \sum_{i=1}^{\kappa} x_i^2\cdot f_i - (\bar{x})^2 \qquad (6)$$

Παράδειγμα 2.17

Για τα δεδομένα του παρακάτω πίνακα να βρείτε τη διακύμανση των παρατηρήσεων.

x_i	2	4	5	6
f_i	0,3	0,2	0,4	0,1

Λύση

Βρίσκουμε πρώτα τη μέση τιμή $\bar{x}$ των παρατηρήσεων. Είναι

$$\bar{x} = \sum_{i=1}^{4} x_i f_i = x_1 f_1 + x_2 f_2 + x_3 f_3 + x_4 f_4$$

$$= 2 \cdot 0,3 + 4 \cdot 0,2 + 5 \cdot 0,4 + 6 \cdot 0,1 = 4$$

Επομένως η διακύμανση είναι

$$s^2 = \sum_{i=1}^{4} (t_i - \bar{x})^2 \cdot f_i =$$

$$= (2-4)^2 \cdot 0,3 + (4-4)^2 \cdot 0,2 + (5-4)^2 \cdot 0,4 + (6-4)^2 \cdot 0,1 =$$

$$= 4 \cdot 0,3 + 0 + 1 \cdot 0,4 + 4 \cdot 0,1 = 1,2 + 0,4 + 0,1 = 1,7$$

2.2.3 Τυπική απόκλιση (s)

Η διακύμανση είναι μια αξιόπιστη παράμετρος διασποράς με ένα μειονέκτημα. Δεν εκφράζεται με τις μονάδες που εκφράζονται οι παρατηρήσεις. Εκφράζεται σε μονάδες οι οποίες είναι τα τετράγωνα των μονάδων των παρατηρήσεων. Για παράδειγμα, αν μετράμε το ύψος ενός πληθυσμού σε cm τότε η διακύμανση εκφράζεται σε cm^2. Για να έχουμε ένα δείκτη ο οποίος να μετρά τη διακύμανση και να εκφράζεται στις ίδιες μονάδες μέτρησης που εκφράζεται η μεταβλητή μας, παίρνουμε την τετραγωνική ρίζα της διακύμανσης. Το μέτρο αυτό ονομάζεται τυπική απόκλιση.

Ορισμός

Ορίζουμε ως **τυπική απόκλιση** την τετραγωνική ρίζα της διακύμανσης.

$$s = \sqrt{s^2}$$

Ιδιότητες διακύμανσης – τυπικής απόκλισης

1. Από τον ορισμό ισχύουν: $s^2 \geq 0$ και $s \geq 0$.

2. Αν οι παρατηρήσεις της μεταβλητής είναι ίσες, η διακύμανση και η τυπική απόκλιση είναι μηδέν.

Πράγματι, αν $t_1 = t_2 = \ldots = t_\nu = \alpha$, τότε $\bar{x} = \dfrac{\alpha + \alpha + \ldots + \alpha}{\nu} = \dfrac{\nu \cdot \alpha}{\nu} = \alpha$.

Επομένως, $t_i - \bar{x} = \alpha - \alpha = 0$, άρα $s^2 = \dfrac{1}{\nu} \sum_{i=1}^{\nu} (t_i - \bar{x})^2 = 0$ και $s = 0$.

3. Ο τύπος της διακύμανσης $s^2 = \frac{1}{\nu}\left[\sum_{i=1}^{\nu} t_i^2 - \frac{1}{\nu}\left(\sum_{i=1}^{\nu} t_i\right)^2\right]$ μπορεί να μετασχηματιστεί ως

$$s^2 = \frac{1}{\nu}\left[\sum_{i=1}^{\nu} t_i^2 - \frac{1}{\nu}\left(\sum_{i=1}^{\nu} t_i\right)^2\right] = \frac{1}{\nu}\sum_{i=1}^{\nu} t_i^2 - \frac{1}{\nu^2}\left(\sum_{i=1}^{\nu} t_i\right)^2$$

$$= \frac{1}{\nu}\sum_{i=1}^{\nu} t_i^2 - \left(\frac{1}{\nu}\sum_{i=1}^{\nu} t_i\right)^2 = \overline{x^2} - (\bar{x})^2$$

Επομένως, είναι

$$s^2 = \overline{x^2} - (\bar{x})^2$$

Διακύμανση και τυπική απόκλιση της μεταβλητής $Y = X + \beta$

Έστω ότι $x_1, x_2, \ldots, x_\nu$ είναι ν παρατηρήσεις μιας μεταβλητής X που έχουν μέση τιμή $\bar{x}$ και διακύμανση s_x^2. Αν $y_1, y_2, \ldots, y_\nu$ είναι οι παρατηρήσεις που προκύπτουν αν αυξήσουμε (ή ελαττώσουμε) κάθε μια από τις $x_1, x_2, \ldots, x_\nu$ κατά μια σταθερά β, δηλαδή,

$$y_1 = x_1 + \beta,\ y_2 = x_2 + \beta, \ldots,\ y_\nu = x_\nu + \beta,$$

τότε η διακύμανση s_y^2 και η τυπική απόκλιση s_y της νέας μεταβλητής $Y = X + \beta$ με τιμές $y_i = x_i + \beta$ είναι:

$$s_y^2 = s_x^2 \text{ και } s_y = s_x$$

Πράγματι είναι

$$s_y^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}(y_i - \bar{y})^2 \text{ με } \bar{y} = \bar{x} + \beta \text{ (ιδιότητα της μέσης τιμής).}$$

Επομένως,

$$s_y^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}(y_i - \bar{y})^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}(x_i + \beta - \bar{x} - \beta)^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}(x_i - \bar{x})^2 = s_x^2.$$

Συνεπώς,

$$s_y^2 = s_x^2$$

Άρα και

$$s_y = s_x.$$

Διακύμανση και τυπική απόκλιση της μεταβλητής $Y = \alpha \cdot X$

Έστω ότι x_1, x_2, …, x_ν είναι ν παρατηρήσεις μιας μεταβλητής X που έχουν μέση τιμή $\bar{x}$ και διακύμανση s_x^2. Αν y_1, y_2, …, y_ν είναι οι παρατηρήσεις που προκύπτουν αν πολλαπλασιάσουμε (ή διαιρέσουμε) κάθε μια από τις x_1, x_2, …, x_ν με τον ίδιο αριθμό α, δηλαδή,

$$y_1 = \alpha \cdot x_1,\ y_2 = \alpha \cdot x_2,\ \ldots,\ y_\nu = \alpha \cdot x_\nu$$

τότε η διακύμανση s_y^2 και η τυπική απόκλιση s_y της νέας μεταβλητής $Y = \alpha \cdot X$ με τιμές $y_i = \alpha \cdot x_i$ είναι

$$s_y^2 = \alpha^2 s_x^2 \text{ και } s_y = |\alpha| \cdot s_x$$

Πράγματι είναι

$$s_y^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}(y_i - \bar{y})^2 \text{ με } \bar{y} = \alpha \cdot \bar{x} \text{ (ιδιότητα της μέσης τιμής).}$$

Επομένως,

$$s_y^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}(y_i - \bar{y})^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}(\alpha x_i - \alpha\bar{x})^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}\alpha^2(x_i - \bar{x})^2 =$$

$$= \alpha^2 \frac{1}{\nu}\sum_{i=1}^{\nu}(x_i - \bar{x})^2 = \alpha^2 s_x^2.$$

Συνεπώς,

$$s_y^2 = \alpha^2 s_x^2$$

Άρα και

$$s_y = \sqrt{s_y^2} = \sqrt{\alpha^2 s_x^2} = |\alpha| \cdot s_x$$

Δηλαδή

$$s_y = |\alpha| \cdot s_x$$

Διακύμανση και τυπική απόκλιση της μεταβλητής $Y = \alpha X + \beta$

Έστω s_x^2 και s_x είναι η διακύμανση και η τυπική απόκλιση αντίστοιχα μιας μεταβλητής Χ ως προς την οποία εξετάζουμε ένα δείγμα. Τότε η διακύμανση s_y^2 και η τυπική απόκλιση s_y της μεταβλητής $Y = \alpha X + \beta$ είναι:

$$s_y^2 = \alpha^2 s_x^2 \text{ και } s_y = |\alpha| \cdot s_x$$

Παράδειγμα 2.18

Ο μέσος μισθός των εργαζόμενων σε μια επιχείρηση είναι $\bar{x}$ = 800 € με τυπική απόκλιση s_x = 40 €. Να βρείτε τη νέα μέση τιμή και τη νέα τυπική απόκλιση στις παρακάτω περιπτώσεις:

α) αν σε κάθε εργαζόμενο δοθεί αύξηση 50 €,

β) αν ο μισθός κάθε εργαζόμενου αυξηθεί κατά 5%,

γ) αν σε κάθε εργαζόμενο δοθεί αύξηση 4% και επίδομα 20 €.

Λύση

Έστω ότι x_1, x_2, …, x_ν είναι οι αρχικοί μισθοί των εργαζομένων.

α) Οι νέοι μισθοί y_1, y_2, …, y_ν των εργαζομένων μετά την αύξηση 50 € θα είναι:

$$y_1 = x_1 + 50,\ y_2 = x_2 + 50,\ \ldots,\ y_\nu = x_\nu + 50$$

Επομένως, η νέα μέση τιμή είναι

$$\bar{y} = \bar{x} + 50 = 800 + 50 = 850\ €$$

και η νέα τυπική απόκλιση είναι

$$s_y = s_x = 40\ €.$$

β) Αν x_i είναι ο αρχικός μισθός ενός εργαζομένου, τότε μετά την αύξηση 5% ο νέος μισθός y_i θα είναι:

$$y_i = x_i + \frac{5}{100}\cdot x_i \text{ ή } y_i = x_i + 0{,}05\cdot x_i \text{ ή } y_i = 1{,}05\cdot x_i$$

Οπότε οι νέοι μισθοί είναι:

$$y_1 = 1{,}05\cdot x_1,\ y_2 = 1{,}05\cdot x_2,\ \ldots,\ y_\nu = 1{,}05\cdot x_\nu$$

Επομένως, η νέα μέση τιμή είναι

$$\bar{y} = 1{,}05\cdot\bar{x} = 1{,}05\cdot 800 = 840\ €$$

και η νέα τυπική απόκλιση είναι

$$s_y = 1{,}05\cdot s_x \text{ ή } s_y = 1{,}05\cdot 40 = 42\ €.$$

γ) Αν x_i είναι ο αρχικός μισθός ενός εργαζομένου, τότε μετά τις αυξήσεις ο νέος μισθός y_i θα είναι:

$$y_i = 1{,}04\cdot x_i + 20$$

οπότε οι νέοι μισθοί είναι:

$$y_1 = 1{,}04\cdot x_1 + 20,\ y_2 = 1{,}04\cdot x_2 + 20\ ,\ \ldots,\ y_2 = 1{,}04\cdot x_2 + 20$$

Άρα, η νέα μέση τιμή είναι

$$\bar{y} = 1{,}04\cdot\bar{x} + 20 = 1{,}04\cdot 800 + 20 = 832 + 20 = 852\ €$$

και η νέα τυπική απόκλιση είναι

$$s_y = |\alpha|\cdot s_x \text{ ή } s_y = 1{,}04\cdot 40 = 41{,}6\ €.$$

2.2.4 Ενδοτεταρτημοριακό εύρος

Το ενδοτεταρτημοριακό εύρος (IQR) είναι η διαφορά μεταξύ του 3ου (Q_3) και του 1ου (Q_1) τεταρτημορίου, $IQR = Q_3 - Q_1$. Έχουμε αναφέρει ότι, τα τεταρτημόρια χωρίζουν τα δεδομένα σε 4 ίσα μέρη (τέταρτα) όπως η διάμεσος διχοτομεί τα δεδομένα σε δύο ίσα μέρη. Το ενδοτεταρτημοριακό εύρος περιλαμβάνει το ενδιάμεσο 50% των παρατηρήσεων. Το υπόλοιπο 50% των παρατηρήσεων βρίσκεται έξω από αυτό το εύρος και μάλιστα το 25% είναι μικρότερες από το Q_1 και το 25% είναι μεγαλύτερες από το Q_3. Όπως είχε αναφερθεί, η διάμεσος θεωρείται το 2ο τεταρτημόριο και συμβολίζεται με Q_2. Ο υπολογισμός των τεταρτημορίων δεν είναι τόσο απλός όσο της διαμέσου και υπάρχουν διάφοροι τρόποι που καταλήγουν σε λίγο διαφορετικά αποτελέσματα.

Παράδειγμα 2.19

Έστω ότι οι ν παρατηρήσεις t_1, t_2, …, t_ν ενός δείγματος έχουν διαταχθεί σε αύξουσα σειρά.

Το 1ο τεταρτημόριο είναι η τιμή στη θέση $Q_1 = t_{\frac{\nu+1}{4}}$ και το 3ο τεταρτημόριο είναι η τιμή στη θέση $Q_3 = t_{\frac{3(\nu+1)}{4}}$

Για παράδειγμα, οι $\nu = 6$ παρατηρήσεις: 7, 1, 3, 6, 3, 7 διατάσσονται ως:
1, 3, 3, 6, 7, 7

Οπότε σε αυτά τα δεδομένα, $t_1 = 1$, $t_2 = 3$, $t_3 = 3$, $t_4 = 6$, $t_5 = 7$, $t_6 = 7$

Το 1ο τεταρτημόριο είναι $Q_1 = t_{\frac{\nu+1}{4}} = t_{\frac{7}{4}} = t_{1\frac{3}{4}}$

Δηλαδή είναι η τιμή που βρίσκεται μεταξύ της $t_1 = 1$ και της $t_2 = 3$ και συγκεκριμένα στα τρία-τέταρτα αυτής της απόστασης. Η διαφορά (η απόσταση) μεταξύ της $t_1 = 1$ και της $t_2 = 3$ είναι 3 – 1 = 2, οπότε

$$Q_1 = t_1 + \frac{3}{4}(t_2 - t_1) = 1 + \frac{3}{4}(3 - 1) = 2{,}5$$

Το 3ο τεταρτημόριο είναι $Q_3 = t_{\frac{3(\nu+1)}{4}} = t_{\frac{21}{4}} = t_{5\frac{1}{4}}$

Δηλαδή είναι η τιμή που βρίσκεται μεταξύ της $t_5 = 7$ και $t_6 = 7$ και συγκεκριμένα στο ένα-τέταρτο αυτής της απόστασης. Τελικά, $Q_3 = 7$, αφού η διαφορά μεταξύ της $t_5 = 7$ και $t_6 = 7$ είναι 7 – 7 = 0.

1 3 3 6 7 7

$Q_1 = 2{,}5$ $Q_2 = 9{,}5$ $Q_3 = 7$

Εφόσον έχουν υπολογιστεί τα τεταρτημόρια, είναι εύκολο να υπολογιστεί το ενδοτεταρτημοριακό εύρος, το οποίο είναι η διαφορά μεταξύ των δύο τεταρτημορίων, $IQR = Q_3 - Q_1 = 7 - 2{,}5 = 4{,}5$.

2.2.5 Συντελεστής μεταβλητότητας

Όταν θέλουμε να συγκρίνουμε δύο κατανομές οι οποίες εκφράζονται σε διαφορετικές μονάδες (π.χ. απόσταση σε μέτρα και μίλια, βάρος σε κιλά και pounds κλπ.) ή όταν οι μέσες τιμές των δύο κατανομών διαφέρουν παρά πολύ μεταξύ τους, τότε δεν μας εξυπηρετούν τα μέτρα διασποράς που έχουμε δει μέχρι τώρα.

Στις περιπτώσεις αυτές χρησιμοποιούμε το συντελεστή μεταβολής ή μεταβλητότητας, ο οποίος είναι ανεξάρτητος από τις μονάδες μέτρησης και, επομένως, επιτρέπει τη σύγκριση τόσο των ομοειδών όσο και των ετεροειδών κατανομών.

Ορισμός

Ορίζουμε ως **συντελεστή μεταβολής** ή **συντελεστή μεταβλητότητας** ενός συνόλου παρατηρήσεων το πηλίκο της τυπικής απόκλισης s προς τη μέση τιμή $\bar{x}$ των παρατηρήσεων. Δηλαδή,

$$CV = \frac{s}{|\bar{x}|} \text{ ή } CV\% = \frac{s}{|\bar{x}|} \cdot 100$$

Παρατηρήσεις

- Εξ ορισμού, ο συντελεστής μεταβολής είναι καθαρός αριθμός, δηλαδή ανεξάρτητος από τις μονάδες μέτρησης.
- Εξ ορισμού είναι $\frac{CV}{100} = \frac{s}{\bar{x}}$, δηλαδή ο συντελεστής μεταβολής είναι το ποσοστό επί τοις εκατό της τυπικής απόκλισης ως προς τη μέση τιμή, συνεπώς $CV = 20\%$ σημαίνει ότι η τυπική απόκλιση είναι το 20% της μέσης τιμής.
- Γενικά δεχόμαστε ότι ένα δείγμα τιμών μιας μεταβλητής είναι ομοιογενές, δηλαδή παρουσιάζει μικρή διασπορά γύρω από τη μέση τιμή, αν ο συντελεστής μεταβολής δεν ξεπερνά το 10%, δηλαδή αν $CV \le 0{,}1$ ή αν $CV\% \le 10\%$.
- Μεταξύ δύο δειγμάτων Α και Β αυτό που έχει μικρότερο συντελεστή μεταβολής θα έχει τη μεγαλύτερη ομοιογένεια.

Συντελεστής μεταβολής της μεταβλητής $Y = X + \beta$

Έστω $\bar{x}$, s_x είναι η μέση τιμή και η τυπική απόκλιση αντίστοιχα μιας μεταβλητής X ως προς την οποία εξετάζουμε ένα δείγμα. Για τη μεταβλητή $Y = X + \beta$ γνωρίζουμε ότι:

◊ η μέση τιμή είναι: $\bar{y} = \bar{x} + \beta$ και

◊ η τυπική απόκλιση είναι: $s_y = s_x$

Συνεπώς, ο συντελεστής μεταβολής CV_y είναι

$$CV_y = \frac{s_y}{|\bar{y}|} = \frac{s_x}{|\bar{x} + \beta|}$$

Συντελεστής μεταβολής της μεταβλητής $Y = \alpha \cdot X$

Έστω $\bar{x}$, s_x είναι η μέση τιμή και η τυπική απόκλιση αντίστοιχα μιας μεταβλητής X ως προς την οποία εξετάζουμε ένα δείγμα. Για τη μεταβλητή $Y = \alpha \cdot X$ γνωρίζουμε ότι:

◊ η μέση τιμή είναι: $\bar{y} = \alpha \cdot \bar{x}$ και

◊ η τυπική απόκλιση είναι: $s_y = |\alpha| \cdot s_x$

Συνεπώς, ο συντελεστής μεταβολής CV_y είναι

$$CV_y = \frac{s_y}{|\bar{y}|} = \frac{|\alpha| \cdot s_x}{|\alpha \cdot \bar{x}|} = \frac{|\alpha| \cdot s_x}{|\alpha| \cdot |\bar{x}|} = \frac{s_x}{|\bar{x}|} = CV_x$$

Συντελεστής μεταβολής της μεταβλητής $Y = \alpha X + \beta$

Έστω $\bar{x}$, s_x είναι η μέση τιμή και η τυπική απόκλιση αντίστοιχα μιας μεταβλητής X ως προς την οποία εξετάζουμε ένα δείγμα. Για τη μεταβλητή $Y = \alpha X + \beta$ γνωρίζουμε ότι:

◊ η μέση τιμή είναι: $\bar{y} = \alpha\bar{x} + \beta$ και

◊ η τυπική απόκλιση είναι: $s_y = |\alpha| \cdot s_x$

Συνεπώς, ο συντελεστής μεταβολής CV_y είναι

$$CV_y = \frac{s_y}{|\bar{y}|} = \frac{|\alpha| \cdot s_x}{|\alpha\bar{x} + \beta|}$$

Υπενθυμίζουμε ότι

Αν $\bar{x}$, s_x, CV_x, είναι η μέση τιμή η τυπική απόκλιση και ο συντελεστής μεταβολής αντίστοιχα μιας μεταβλητής X, τότε για τη μεταβλητή Y ισχύουν:

	$Y = X + \beta$	$Y = \alpha \cdot X$	$Y = \alpha X + \beta$
Μέση τιμή $\bar{y}$	$\bar{y} = \bar{x} + \beta$	$\bar{y} = \alpha \cdot \bar{x}$	$\bar{y} = \alpha\bar{x} + \beta$
Τυπική απόκλιση s_y	$s_y = s_x$	$s_y = \lvert\alpha\rvert \cdot s_x$	$s_y = \lvert\alpha\rvert \cdot s_x$
Συντελεστής μεταβολής CV_y	$CV_y = \frac{s_y}{\lvert\bar{y}\rvert} = \frac{s_x}{\lvert\bar{x} + \beta\rvert}$	$CV_y = CV_x$	$CV_y = \frac{s_y}{\lvert\bar{y}\rvert} = \frac{\lvert\alpha\rvert \cdot s_x}{\lvert\alpha\bar{x} + \beta\rvert}$

Παράδειγμα 2.20

Μια βιομηχανία κατασκευάζει 4 προϊόντα σε ποσοστά: 10%, 20%, 30% και 40% αντίστοιχα και κόστος κατασκευής 5, 4, 3, 2 χιλιάδες € αντίστοιχα.

α) Να υπολογιστεί το μέσο κόστος, η τυπική απόκλιση και ο συντελεστής μεταβολής του κόστους κατασκευής των προϊόντων.

β) Να βρεθεί πόσο τουλάχιστον πρέπει να αυξηθεί το κόστος κατασκευής κάθε προϊόντος, ώστε το κόστος να είναι ομοιογενές.

γ) Αν ελαττωθεί το κόστος κατασκευής κάθε προϊόντος κατά 10% και στη συνέχεια γίνει αύξηση κατά 0,3 χιλιάδες € ανά μονάδα προϊόντος, να βρεθεί ο νέος συντελεστής μεταβολής.

Λύση

α) Κατασκευάζουμε τον πίνακα σχετικών συχνοτήτων με επιπλέον στήλες, τη στήλη με τα x_i^2, τη στήλη με τα γινόμενα των $x_i f_i$ και αυτή με τα γινόμενα των $x_i^2 f_i$.

x_i	f_i	x_i^2	$x_i f_i$	$x_i^2 f_i$
5	0,1	25	0,5	2,5
4	0,2	16	0,8	3,2
3	0,3	9	0,9	2,7
2	0,4	4	0,8	1,6
Σύνολο	1		$\sum_{i=1}^{4} x_i f_i = 3$	$\sum_{i=1}^{4} x_i^2 f_i = 10$

Το μέσο κόστος $\bar{x}$ κατασκευής των προϊόντων είναι:

$$\bar{x} = \sum_{i=1}^{4} x_i f_i = 3 \text{ χιλιάδες €}$$

Η διακύμανση s^2 του κόστους των προϊόντων είναι:

$$s_x^2 = \sum_{i=1}^{4} x_i^2 \cdot f_i - (\bar{x})^2 = 10 - 3^2 = 10 - 9 = 1$$

Οπότε, ο συντελεστής μεταβολής του κόστους των τεσσάρων προϊόντων είναι:

$$CV_x = \frac{s_x}{\bar{x}} = \frac{1}{3} = 0{,}333 \text{ ή } CV_x = 33{,}3\%.$$

β) Έστω ότι x_i είναι το αρχικό κόστος και β η αύξηση που πρέπει να γίνει ανά μονάδα ώστε το κόστος να είναι ομοιογενές. Η νέα τιμή κόστους y_i, θα είναι $y_i = x_i + \beta$.

Συνεπώς, η νέα μέση τιμή $\bar{y} = \bar{x} + \beta$ ή $\bar{y} = 3 + \beta$

και η νέα τυπική απόκλιση $s_y = s_x = 1$

Επομένως, $CV_y = \frac{s_y}{\bar{y}} = \frac{1}{3 + \beta}$

Πρέπει $CV_y \le \frac{10}{100} \Leftrightarrow \frac{1}{3+\beta} \le \frac{1}{10} \Leftrightarrow 3+\beta \ge 10 \Leftrightarrow \beta \ge 7$

Άρα, η αύξηση πρέπει να είναι τουλάχιστον 7 χιλιάδες € ανά μονάδα προϊόντος, ώστε το κόστος να είναι ομοιογενές.

γ) Η νέα τιμή κόστους είναι $y_i = 0{,}9x_i + 0{,}3$.

Συνεπώς, η νέα μέση τιμή $\bar{y} = 0{,}9 \cdot \bar{x} + 0{,}3$ ή $\bar{y} = 0{,}9 \cdot 3 + 0{,}3$ ή $\bar{y} = 3$

και η νέα τυπική απόκλιση $s_y = 0{,}9 \cdot s_x = 0{,}9 \cdot 1 = 0{,}9$.

Οπότε, ο νέος συντελεστής μεταβολής του κόστους των τεσσάρων προϊόντων είναι:

$$CV_y = \frac{s_y}{\bar{y}} = \frac{0{,}9}{3} = 0{,}3 \text{ ή } CV_y = 30\%.$$

Παράδειγμα 2.21

Μια εταιρεία έχει εγκαταστημένα 20 υποκαταστήματα (σημεία πώλησης των προϊόντων της) σε επιλεγμένες περιοχές, κατά το δυνατόν ισοδύναμες μεταξύ τους από πλευράς ευκαιριών και δυνατοτήτων ως προς το ετήσιο ύψος του κύκλου εργασιών τους (ετήσιου τζίρου).

Η εταιρεία προσέλαβε τη φετινή περίοδο ένα νέο γενικό διευθυντή, ο οποίος αναδιοργάνωσε την επιχείρηση και εφάρμοσε νέες επιστημονικές μεθόδους, στην οργάνωση και διοίκηση (management) καθώς και στον τομέα των πωλήσεων της εταιρείας. Μετά τη συμπλήρωση της νέας οικονομικής χρήσης ο γενικός διευθυντής αποφασίζει να μελετήσει τα έσοδα πωλήσεων των 20 υποκαταστημάτων. Έτσι θα αντλήσει συμπεράσματα, που θα τα συγκρίνει με αυτά της περσινής περιόδου, για να μπορέσει να εξάγει κάποια τελικά συμπεράσματα σχετικά με το τι προέκυψε μετά την αναδιοργάνωση της επιχείρησης.

Οι πωλήσεις (σε εκατομμύρια ευρώ) των 20 υποκαταστημάτων της φετινής χρήσης ήταν:

105, 112, 115, 118, 123, 123, 124, 125, 127, 128,
132, 133, 134, 136, 138, 138, 142, 145, 149, 156

Ως προς την περσινή χρήση γνωρίζει ότι για τα 20 καταστήματα, η μέση τιμή των πωλήσεων ήταν $\bar{x}_1 = 90$ εκατ. ευρώ, και η διασπορά των πωλήσεων $s_1 = 10$ εκατ. ευρώ. Ακόμη οι χαμηλότερες πωλήσεις υποκαταστήματος ήταν 70 εκατ. ευρώ και οι υψηλότερες 140 εκατ. ευρώ.

Ο γενικός διευθυντής συντάσσει αρχικά τον παρακάτω πίνακα και υπολογίζει τη νέα μέση τιμή και τη νέα τυπική απόκλιση.

Κλάσεις πωλήσεων	Κέντρα κλάσης x_i	x_i^2	Αριθμός υποκ/των ν_i	$x_i\nu_i$	$x_i^2\nu_i$
[100, 110)	105	11.025	1	105	11.025
[110, 120)	115	13.225	3	345	39.675
[120, 130)	125	15.625	6	750	93.750
[130, 140)	135	18.225	6	810	109.350
[140, 150)	145	21.025	3	435	63.075
[150, 160)	155	24.025	1	155	24.025
Σύνολο			20	$\sum_{i=1}^{6} x_i\nu_i = 2.600$	$\sum_{i=1}^{6} x_i^2\nu_i = 340.900$

Προκύπτει λοιπόν ότι η νέα μέση τιμή $\bar{x}_2$ των πωλήσεων της εταιρείας θα είναι:

$$\bar{x}_2 = \frac{1}{20}\sum_{i=1}^{6} x_i\nu_i = \frac{1}{20}\cdot 2600 = 130 \text{ εκατ. ευρώ}$$

Η νέα διασπορά s_2^2 θα είναι:

$$s_2^2 = \frac{1}{20}\left[\sum_{i=1}^{6} x_i^2\nu_i - \frac{1}{100}\left(\sum_{i=1}^{6} x_i\nu_i\right)^2\right] = \frac{1}{20}\cdot[340900 - \frac{1}{20}\cdot 2600^2] = 145$$

Η νέα τυπική απόκλιση s_2 θα είναι:

$$s_2 = \sqrt{s_2^2} = \sqrt{145} = 12{,}04 \cong 12 \text{ εκατ. ευρώ}$$

Εδώ ο γενικός διευθυντής θα πρέπει να προσέξει. Δεν πρέπει να λάβει υπόψη του την εικόνα των τυπικών αποκλίσεων και να σκεφτεί ότι, επειδή η φετινή τυπική απόκλιση $s_2 = 12$ είναι μεγαλύτερη από την περσινή $s_1 = 10$, η διασπορά των εσόδων φέτος παρουσιάζει χειρότερη εικόνα. Για να είναι σίγουρος πρέπει να υπολογίσει και τους αντίστοιχους συντελεστές μεταβολής CV_1 και CV_2.

$$CV_1 = \frac{s_1}{\bar{x}_1} = \frac{10}{90} = 0{,}111 \text{ ή } CV_1 = 11{,}1\%$$

$$CV_2 = \frac{s_2}{\bar{x}_2} = \frac{12}{130} = 0{,}092 \text{ ή } CV_1 = 9{,}2\%$$

Επομένως, παρόλο που η φετινή τυπική απόκλιση είναι μεγαλύτερη από την περσινή, ο συντελεστής μεταβολής δίνει μεγαλύτερη σχετική διασπορά στην περσινή χρονιά. Αυτό σημαίνει ότι παρατηρείται μεγαλύτερη ομοιογένεια στις πωλήσεις κατά τη φετινή χρονιά.

Στον παρακάτω πίνακα παρουσιάζουμε τις τιμές των στατιστικών μέτρων και τη μεταβολή των δύο διαχειριστικών χρήσεων

Στατιστικά μέτρα	**Προηγούμενη διαχειριστική χρήση (1)**	**Φετινή διαχειριστική χρήση (2)**	**Μεταβολή των χρήσεων (1) και (2)**
1. Μέση τιμή $\bar{x}$	$\bar{x}_1 = 90$	$\bar{x}_2 = 130$	αύξηση (+)
2. Διακύμανση s^2	$s_1^2 = 100$	$s_2^2 = 145$	αύξηση (+)
3. Τυπική απόκλιση s	$s_1 = 10$	$s_2 = 12$	αύξηση (+)
4. Συντελεστής μεταβλητότητας CV	$CV_1 = 11{,}1\%$	$CV_2 = 9{,}2\%$	μείωση (–)
5. Εύρος μεταβολής R	$R_1 = 70$	$R_2 = 51$	μείωση (–)

Συμπέρασμα:

Από τη μέχρι τώρα μελέτη μπορούμε να πούμε ότι η επιχείρηση βελτίωσε την οικονομική της θέση, διότι.

α) Έχει σημαντική αύξηση της μέσης τιμής των πωλήσεων από 90 εκατ. ευρώ πέρυσι σε 130 εκατ. ευρώ φέτος (130 – 90 = 40), ποσοστό αύξησης περίπου 44,44% .

β) Η διασπορά των πωλήσεων φέτος παρουσιάζει καλύτερη εικόνα, δηλαδή έχει μεγαλύτερη ομοιογένεια ως προς τα έσοδα, αφού $CV_2 = 9{,}2\% < CV_{1\,1} = 11{,}1\%$.

γ) Επιπλέον, έχει φέτος ένα εύρος μεταβολής ίσο με: 156 – 105 = 51 εκατ. ευρώ, το οποίο είναι αρκετά μικρότερο από το περσινό, που ήταν 140 – 70 = 70 εκατ. ευρώ.

Φυσικά η μελέτη δεν σταματά εδώ, θα υπάρξει συνέχεια με την επεξεργασία και άλλων στοιχείων και με τη χρήση και άλλων στατιστικών μέτρων, όμως και τα μέχρι τώρα αποτελέσματα που βρέθηκαν είναι σοβαρά και σημαντικά.

2.2.6 Σύγκριση μέτρων διασποράς

Εύρος

Πλεονεκτήματα	Μειονεκτήματα
Είναι πολύ απλό στον υπολογισμό.	Δεν θεωρείται αξιόπιστο αφού επηρεάζεται μόνο από τις δύο ακραίες τιμές.
Χρησιμοποιείται στον έλεγχο ποιότητας.	Δεν χρησιμοποιείται για περαιτέρω στατιστική ανάλυση

Διακύμανση και τυπική απόκλιση

Πλεονεκτήματα	Μειονεκτήματα
Για τον υπολογισμό τους χρησιμοποιούνται όλες οι παρατηρήσεις.	Χρειάζονται συνήθως πολλές πράξεις για τον υπολογισμό τους.
Χρησιμοποιούνται πολύ στη στατιστική ανάλυση.	Η διακύμανση δεν εκφράζεται με τις ίδιες μονάδες που εκφράζονται οι μεταβλητές

Κεφάλαιο 3 Συσχέτιση - Παλινδρόμηση

3.1 Συσχέτιση

Δύο τυχαίες μεταβλητές X και Y μπορεί να συσχετίζονται με κάποιο τρόπο. Αυτό συμβαίνει όταν επηρεάζει η μία την άλλη, ή αν δεν αλληλοεπηρεάζονται, όταν επηρεάζονται και οι δύο από μια άλλη μεταβλητή. Για παράδειγμα ο χρόνος ως την αποτυχία ενός στοιχείου κάποιας μηχανής και η ταχύτητα του κινητήρα της μηχανής μπορούν να θεωρηθούν σαν δύο τυχαίες μεταβλητές που συσχετίζονται, όπου ο χρόνος αποτυχίας εξαρτάται από την ταχύτητα του κινητήρα (το αντίθετο δεν έχει πρακτική σημασία).

Μπορούμε επίσης να θεωρήσουμε τη συσχέτιση του χρόνου αποτυχίας και της θερμοκρασίας του στοιχείου της μηχανής, αλλά τώρα δεν εξαρτάται η μια από την άλλη παρά εξαρτιούνται και οι δύο από άλλες μεταβλητές, όπως η ταχύτητα του κινητήρα. Η συσχέτιση λοιπόν δεν υποδηλώνει απαραίτητα κάποια αιτιακή σχέση των δύο μεταβλητών.

Μια τεχνική, η οποία χρησιμοποιείται ευρύτατα για τον ποσοτικό προσδιορισμό της σχέσης δύο συνεχών τυχαίων μεταβλητών είναι η **συσχέτιση**. Με τον όρο συσχέτιση ορίζεται ο βαθμός στον οποίο συμμεταβάλλονται δύο συνεχείς τυχαίες μεταβλητές, υπό την προϋπόθεση ότι η σχέση τους είναι γραμμική.

Στην πραγματικότητα υπάρχουν διάφοροι τρόποι με τους οποίους μπορούν να συσχετίζονται οι τιμές δύο συνεχών τυχαίων μεταβλητών και είναι απαραίτητο, προτού γίνει οποιοσδήποτε ποσοτικός προσδιορισμός της σχέσης τους, να οριστεί πρώτα η συναρτησιακή μορφή τους. Η συνήθης παραδοχή που γίνεται για τη σχέση δύο συνεχών τυχαίων μεταβλητών X και Y είναι ότι η σχέση τους είναι γραμμική, δηλαδή η συνδυασμένη απεικόνιση των τιμών των δύο μεταβλητών σε ένα ορθογώνιο σύστημα αξόνων, ορίζει ένα σύνολο σημείων τα οποία τείνουν να συσσωρεύονται κατά μήκος μιας ευθείας γραμμής.

3.2 Διάγραμμα Διασποράς

Στον παρακάτω πίνακα δίνεται, για ένα δείγμα 18 χωρών του ΟΗΕ, το ποσοστό (%) των ατόμων κάθε χώρας που ζουν σε αστικά κέντρα (βαθμός αστικότητας) και το ποσοστό (%) των ατόμων που γνωρίζουν να διαβάζουν.

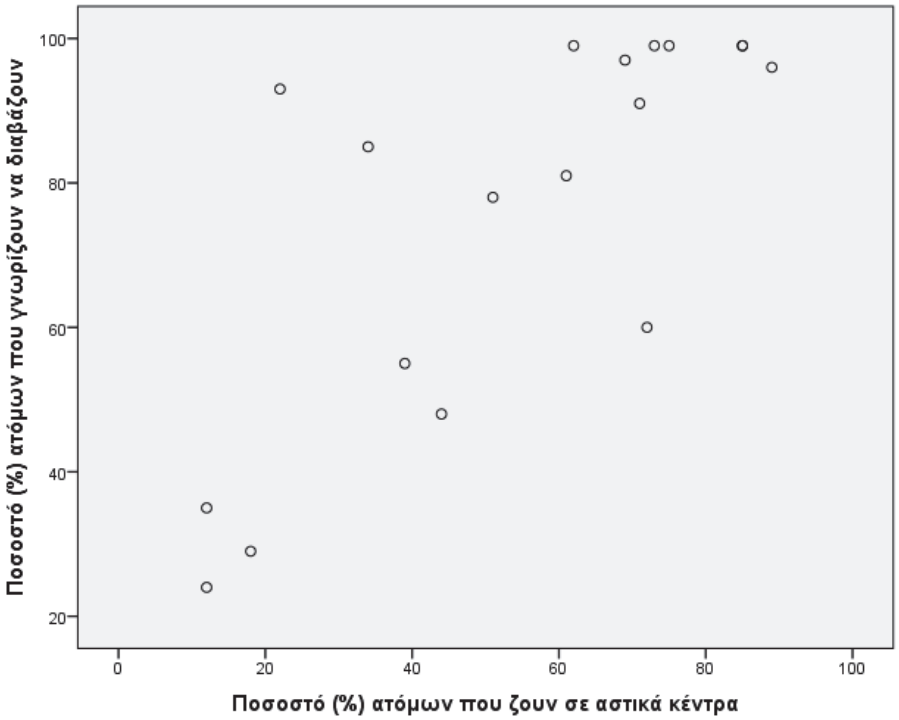
Πίνακας 3.1
Ποσοστό ατόμων που ζουν σε αστικά κέντρα και ποσοστό ατόμων που γνωρίζουν να διαβάζουν σε ένα τυχαίο δείγμα χωρών του ΟΗΕ

	Ποσοστό (%) ατόμων που ζουν σε αστικά κέντρα (X)	Ποσοστό (%) ατόμων που γνωρίζουν να διαβάζουν (Y)
Αίγυπτος	44	48
Αιθιοπία	12	24
Αφγανιστάν	18	29
Βολιβία	51	78
Γαλλία	73	99
Γερμανία	85	99
Γουατεμάλα	39	55
Δανία	85	99
Ελβετία	62	99
Ιράκ	72	60
Ιταλία	69	97
Καμπότζη	12	35
Νορβηγία	75	99
Ουρουγουάη	89	96
Πορτογαλία	34	85
Ταϊβάν	71	91
Ταϊλάνδη	22	93
Τουρκία	61	81

Αν το ποσοστό των ατόμων που ζουν σε αστικά κέντρα συμβολιστεί με X και το ποσοστό των ατόμων που γνωρίζουν να διαβάζουν με Y, ορίζεται για κάθε χώρα ένα ζεύγος τιμών (x_i, y_i). Προκειμένου να διερευνηθεί αν υπάρχει συσχέτιση μεταξύ του βαθμού αστικότητας και του ποσοστού των ατόμων που γνωρίζουν να διαβάζουν, είναι απαραίτητο πριν από οποιαδήποτε προσπάθεια ποσοτικής ανάλυσης των δεδομένων, να απεικονιστεί διαγραμματικά η σχέση των δύο μεταβλητών. Τοποθετώντας στον οριζόντιο άξονα ενός ορθογωνίου συστήματος αξόνων τις τιμές της μεταβλητής X και στον

κατακόρυφο άξονα τις τιμές της μεταβλητής Y, για κάθε ζεύγος τιμών (x_i, y_i) ορίζεται ένα σημείο που αναπαριστά τη συγκεκριμένη χώρα. Το σύνολο των σημείων που προκύπτουν αποτελεί το αντίστοιχο **διάγραμμα διασποράς**.

Εικόνα 3.1
Διάγραμμα διασποράς των δεδομένων του Πίνακα 3.1



Εξετάζοντας τη μορφή του διαγράμματος μπορούμε να συμπεράνουμε για το είδος της σχέσης που πιθανώς να υπάρχει μεταξύ των δύο μεταβλητών, δηλαδή του βαθμού αστικότητας και του ποσοστού των ανθρώπων που γνωρίζουν να διαβάζουν.

Από τη μορφή του διαγράμματος και λόγω της γραμμικής διάταξης των σημείων (χωρών) στο εσωτερικό του προκύπτει ότι το ποσοστό των ατόμων που γνωρίζουν να διαβάζουν στις 18 χώρες τείνει να αυξάνει όσο αυξάνει ο βαθμός αστικότητας των χωρών.

3.3 Συντελεστής συσχέτισης του Pearson

Έστω X και Y ένα ζεύγος συνεχών τυχαίων μεταβλητών με μέσες τιμές μ_X και μ_Y και διακυμάνσεις σ_X^2 και σ_Y^2 οι οποίες σχετίζονται μεταξύ τους με τρόπο γραμμικό. Ως μέτρο της σχέσης των δύο μεταβλητών ορίζεται η ποσότητα:

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

Η ποσότητα αυτή ονομάζεται **συνδιακύμανση** των X και Y.

Αν υπάρχει θετική συσχέτιση μεταξύ των δύο τυχαίων μεταβλητών, δηλαδή οι υψηλές τιμές της X τείνουν να εμφανίζονται με υψηλές τιμές της Y και οι χαμηλές τιμές της X εμφανίζονται με χαμηλές τιμές της Y, όπως στην περίπτωση του Σχήματος 3.2.1 η συνδιακύμανση είναι θετική. Αν υπάρχει αρνητική σχέση μεταξύ των δύο τυχαίων μεταβλητών, δηλαδή οι υψηλές τιμές της X τείνουν να εμφανίζονται με χαμηλές τιμές της Y και οι χαμηλές τιμές της X με υψηλές τιμές της Y, η συνδιακύμανση είναι αρνητική. Αν δεν υπάρχει γραμμική σχέση μεταξύ των X και Y, τότε η συνδιακύμανσή τους είναι 0.

Η συνδιακύμανση εξαρτάται από τις μονάδες των δύο συγκρινόμενων μεταβλητών και επομένως, αποτελεί απόλυτο μέτρο του βαθμού της συσχέτισης που υπάρχει μεταξύ τους. Προκειμένου να οριστεί ένα μέτρο ανεξάρτητο μονάδων, η συνδιακύμανση διαιρείται με το γινόμενο των δύο τυπικών αποκλίσεων σ_X και σ_Y. Η ποσότητα που προκύπτει με αυτόν τον τρόπο ονομάζεται **συντελεστής συσχέτισης** και συμβολίζεται με το ελληνικό γράμμα ρ.

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Ο παραπάνω ορισμός του συντελεστή συσχέτισης αφορά τη σχέση δύο μεταβλητών που ορίζονται σε πληθυσμιακό επίπεδο. Ένα τυχαίο δείγμα n ζευγών παρατηρήσεων (x_1,y_1), (x_2,y_2), …, (x_n,y_n) προερχόμενο από τον αντίστοιχο πληθυσμό μπορεί να χρησιμοποιηθεί για την εκτίμηση του πληθυσμιακού συντελεστή συσχέτισης.

Εκτίμηση του πληθυσμιακού συντελεστή συσχέτισης είναι η ποσότητα:

$$r = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

ή αλλιώς

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}$$

η οποία ονομάζεται **συντελεστής συσχέτισης του Pearson** ή **δειγματικός συντελεστής συσχέτισης**.

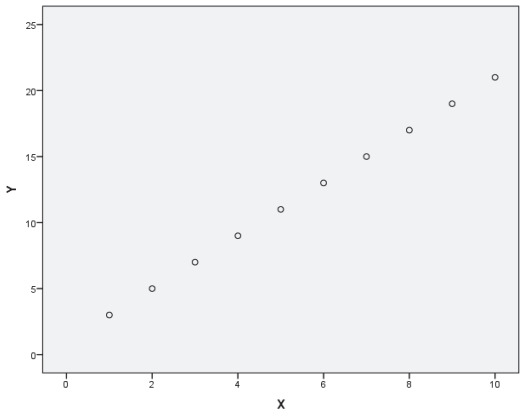
Μια απλούστερη υπολογιστικά έκφραση του συντελεστή συσχέτισης του Pearson δίνεται από τον τύπο:

$$r = \frac{n\sum_{i=1}^{n}x_i y_i - \left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sqrt{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2}}$$

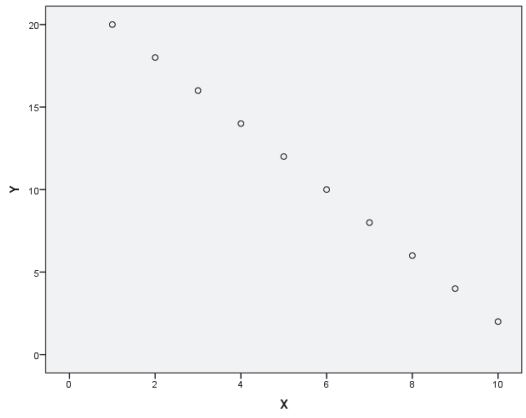
Ο συντελεστής συσχέτισης του Pearson είναι ανεξάρτητος των μονάδων και οι δυνατές τιμές που μπορεί να πάρει ανήκουν στο διάστημα [−1, 1]. Δηλαδή, για κάθε σύνολο ζευγών παρατηρήσεων (x_1, y_1), (x_2, y_2), …, (x_n, y_n) η τιμή του συντελεστή συσχέτισης είναι $-1 \leq r \leq 1$. Οι τιμές $r = -1$ και $r = 1$ προκύπτουν όταν υπάρχει πλήρης γραμμική σχέση μεταξύ των x_i και y_i. Όταν δηλαδή τα σημεία που ορίζονται από τα ζεύγη τιμών (x_i, y_i) βρίσκονται κατά μήκος μιας ευθείας γραμμής.

Παραδείγματα πλήρους γραμμικής συσχέτισης εμφανίζονται στις Εικόνες 3.2 και 3.3.

Εικόνα 3.2



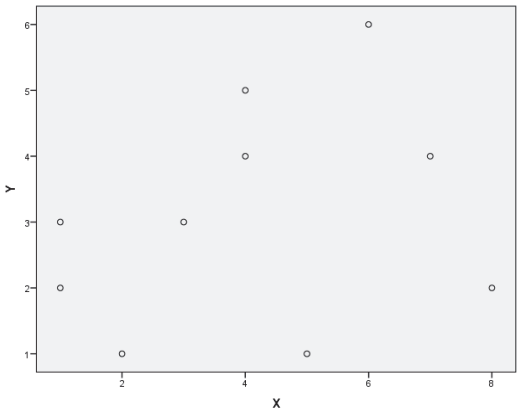
Εικόνα 3.3



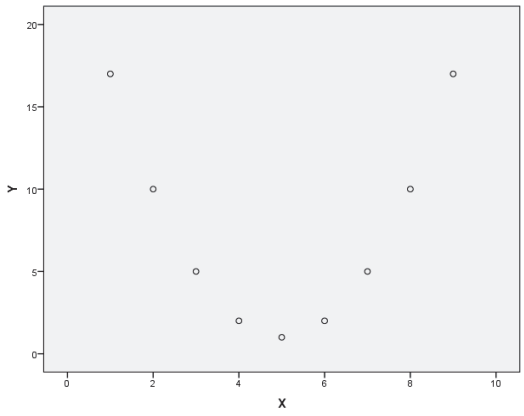
Όσο η σχέση μεταξύ των x_i και y_i αποκλίνει από την πλήρη γραμμικότητα, η τιμή του r τείνει να απομακρύνεται από τις τιμές −1 και 1 και να πλησιάζει το 0. Όταν οι τιμές y_i τείνουν να αυξάνουν όσο αυξάνουν και οι αντίστοιχες τιμές x_i, η τιμή του r είναι

θετική και οι μεταβλητές χαρακτηρίζονται **θετικά συσχετιζόμενες**. Στην αντίστροφη περίπτωση, όπου οι τιμές y_i ελαττώνονται όσο οι τιμές x_i αυξάνουν, ο συντελεστής συσχέτισης r παίρνει αρνητικές τιμές και οι δύο μεταβλητές χαρακτηρίζονται **αρνητικά συσχετιζόμενες**. Αν η τιμή του συντελεστή συσχέτισης είναι $r = 0$, τότε μεταξύ των δύο μεταβλητών δεν υπάρχει γραμμική σχέση (Εικόνα 3.4). Σε μια τέτοια περίπτωση όμως, μπορεί να υπάρχει μη γραμμική σχέση μεταξύ των δύο μεταβλητών όπως φαίνεται στο Εικόνα 3.5.

Εικόνα 3.4



Εικόνα 3.5



Παράδειγμα 3.1

Εφαρμόζοντας τον τύπο του συντελεστή συσχέτισης του Pearson στα δεδομένα του Πίνακα 3.1, έχουμε:

για τη μεταβλητή X, που αναπαριστά το βαθμό αστικότητας των χωρών

$$\bar{x} = \frac{1}{18}\sum_{i=1}^{18} x_i = \frac{44+12+...+61}{18} = 54{,}1$$

για τη μεταβλητή Y, που αναπαριστά το ποσοστό των ατόμων που γνωρίζουν να διαβάζουν

$$\bar{y} = \frac{1}{18}\sum_{i=1}^{18} y_i = \frac{48+24+...+81}{18} = 75{,}9$$

για τις δύο μεταβλητές X και Y

$$\sum_{i=1}^{18}(x_i - \bar{x})(y_i - \bar{y}) =$$

$$= (44-54{,}1)(48-75{,}9)+(12-54{,}1)(24-75{,}9)+...+(61-54{,}1)(81-75{,}9) = 9010{,}1$$

και

$$\sum_{i=1}^{18}(x_i-\bar{x})^2 = (44-54{,}1)^2+(12-54{,}1)^2+\ldots+(61-54{,}1)^2 = 11481{,}8$$

$$\sum_{i=1}^{18}(y_i-\bar{y})^2 = (48-75{,}9)^2+(24-75{,}9)^2+\ldots+(81-75{,}9)^2 = 12184{,}9$$

Άρα ο συντελεστής συσχέτισης Pearson ισούται με:

$$r = \frac{\sum_{i=1}^{18}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{18}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{18}(y_i-\bar{y})^2}} = \frac{9010{,}1}{\sqrt{11481{,}8}\cdot\sqrt{12184{,}9}} = 0{,}76.$$

Από την τιμή του συντελεστή συσχέτισης του Pearson προκύπτει ότι υπάρχει ισχυρή θετική συσχέτιση μεταξύ του βαθμού αστικότητας και του ποσοστού των ατόμων που γνωρίζουν να διαβάζουν. Αυτό σε καμία περίπτωση δεν μπορεί να ερμηνευτεί ως μια σχέση αιτίου-αποτελέσματος, δηλαδή ότι η συγκέντρωση των ανθρώπων στα αστικά κέντρα οδηγεί στη μείωση του αναλφαβητισμού. Ο συντελεστής συσχέτισης μετρά απλά το βαθμό της γραμμικής σχέσης που υπάρχει μεταξύ δύο συνεχών μεταβλητών και, στην προκειμένη περίπτωση, η τιμή του αντικατοπτρίζει αυτήν ακριβώς τη γραμμική σχέση.

3.4 Απλή γραμμική παλινδρόμηση

Ο συντελεστής συσχέτισης που αναφέρθηκε στην προηγούμενη παράγραφο ορίστηκε ως ένα μέτρο του βαθμού της γραμμικής σχέσης που υπάρχει μεταξύ δύο συνεχών τυχαίων μεταβλητών. Κατά τον υπολογισμό και την ερμηνεία του, η συμμετοχή των δύο μεταβλητών είναι απολύτως συμμετρική και δεν έχει κανένα ενδιαφέρον αν ο προσδιορισμός του αφορά τη συσχέτιση της X μα τη Y ή τη συσχέτιση της Y με τη X. Αν ο συμμετρικός ρόλος των δύο μεταβλητών δεν υφίσταται και η γραμμική σχέση των δύο συνεχών μεταβλητών οριστεί με όρους εξάρτησης της μιας από την άλλη, δηλαδή αν η μεταβολή των τιμών της μιας μεταβλητής θεωρηθεί ότι προκύπτει από τη μεταβολή των τιμών της άλλης, τότε η ανάλυση της σχέσης των δύο μεταβλητών πραγματοποιείται με τη βοήθεια της **απλής γραμμικής παλινδρόμησης**.

3.4.1 Το υπόδειγμα της απλής γραμμικής παλινδρόμησης

Η απλή γραμμική παλινδρόμηση ποσοτικοποιεί τη σχέση δύο συνεχών τυχαίων μεταβλητών X και Y υπό τη μορφή ενός γραμμικού υποδείγματος στο οποίο οι τιμές της μιας μεταβλητής εκτιμώνται (ή προβλέπονται) από τις τιμές της άλλης. Αν οι τιμές της μεταβλητής Y εκτιμώνται από τις τιμές της X, τότε η Y ονομάζεται εξαρτημένη μεταβλητή και η X ονομάζεται ανεξάρτητη μεταβλητή.

Η εκτίμηση των τιμών της μεταβλητής Y από τις τιμές της X, διά μέσου του υποδείγματος της απλής γραμμικής παλινδρόμησης μπορεί να γίνει όταν διασφαλίζονται οι εξής προϋποθέσεις:

1. Ο προσδιορισμός των τιμών της μεταβλητής X γίνεται χωρίς σφάλμα. Επειδή στην πραγματικότητα καμία συνεχής μέτρηση δεν είναι απαλλαγμένη σφαλμάτων, η παραδοχή αυτή υπονοεί ότι το μέγεθος του σφάλματος κατά τη μέτρηση της X είναι αμελητέο.
2. Σε κάθε τιμή της X αντιστοιχεί ένας υπο-πληθυσμός τιμών της Y, ο οποίος ακολουθεί την κανονική κατανομή.
3. Οι διακυμάσεις των υπο-πληθυσμών της Y που ορίζονται για διάφορες τιμές της X είναι ίσες. Η κοινή διακύμανση των υπο-πληθυσμών της Y συμβολίζεται με $\sigma^2_{y|x}$. Η παραδοχή της ισότητας των διακυμάνσεων των τιμών της Y, ονομάζεται **ομοσκεδαστικότητα** και είναι ανάλογη με την παραδοχή της ισότητας των διακυμάνσεων που απαιτείται σε ένα t-test για ανεξάρτητα δείγματα ή στην ανάλυση διακύμανσης με έναν παράγοντα.
4. Οι μέσες τιμές των υπο-πληθυσμών της Y συνδέονται με τις αντίστοιχες τιμές της X διά μέσου μιας γραμμικής σχέσης της μορφής $\mu_{y|x} = \alpha + \beta x$, όπου $\mu_{y|x}$ είναι η μέση τιμή

του υπο-πληθυσμού της Y που αντιστοιχεί σε μια συγκεκριμένη τιμή x της μεταβλητής X. Οι ποσότητες α και β ονομάζονται **πληθυσμιακοί συντελεστές** της παλινδρόμησης. Το παραπάνω υπόδειγμα ορίζει μια ευθεία γραμμή, επί της οποίας είναι τοποθετημένες οι μέσες τιμές $\mu_{y|x}$ των διαφόρων υπο-πληθυσμών της Y. Η ευθεία αυτή γραμμή ονομάζεται **πληθυσμιακή ευθεία της παλινδρόμησης**. Γεωμετρικά, ο συντελεστής α αναπαριστά την τεταγμένη στο σημείο 0 και ο συντελεστής β την κλίση της ευθείας της παλινδρόμησης.
5. Οι τιμές της Y είναι ανεξάρτητες η μια της άλλης

Όλες οι προηγούμενες προϋποθέσεις συνοψίζονται στην παρακάτω εξίσωση η οποία ονομάζεται **υπόδειγμα της απλής γραμμικής παλινδρόμησης**:

$$y = \alpha + \beta x + \varepsilon ,$$

όπου y είναι μια οποιαδήποτε τιμή του υπο-πληθυσμού τιμών της Y που αντιστοιχεί στην τιμή x και α, β είναι οι **πληθυσμιακοί συντελεστές** της παλινδρόμησης.

Αν επιλύσουμε την προηγούμενη εξίσωση ως προ ε έχουμε

$$\varepsilon = y - (\alpha + \beta x) = y - \mu_{y|x}$$

Η ποσότητα ε, η οποία ονομάζεται **σφάλμα**, εκφράζει τη διαφοροποίηση της y από τη μέση τιμή του υπο-πληθυσμού της Y στον οποίο αυτή ανήκει, ή αλλιώς, εκφράζει την απόκλιση της y από την ευθεία παλινδρόμησης. Ως συνέπεια της παραδοχής ότι οι διάφοροι υπο-πληθυσμοί της Y ακολουθούν κανονική κατανομή με κοινή διακύμανση, οι ποσότητες ε για κάθε τιμή της X ακολουθούν επίσης κανονική κατανομή με διακύμανση ίση με την κοινή διακύμανση $\sigma^2_{y|x}$ των αντίστοιχων υπο-πληθυσμών της Y. Επιπλέον, από τον ορισμό των σφαλμάτων προκύπτει ότι η μέση τιμή τους είναι ίση με 0.

3.4.2 Η δειγματική εξίσωση της παλινδρόμησης

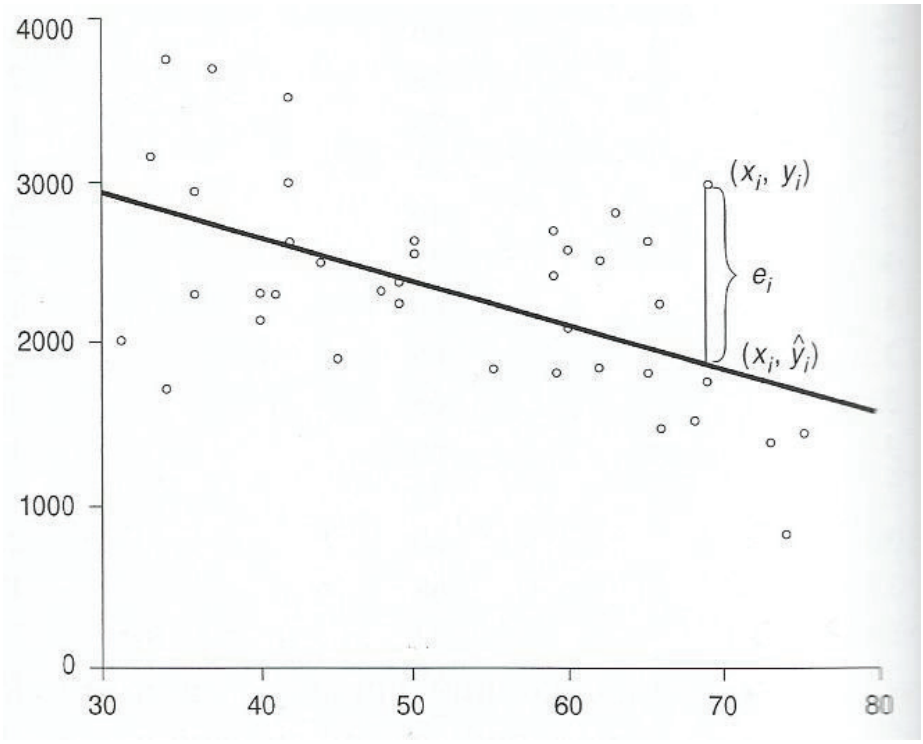
Σε ένα τυπικό πρόβλημα γραμμικής παλινδρόμησης, το ενδιαφέρον εστιάζεται στον προσδιορισμό της πληθυσμιακής ευθείας της παλινδρόμησης, δηλαδή της ευθείας που περιγράφει την πραγματική σχέση που υπάρχει μεταξύ των μεταβλητών X και Y. Ο προσδιορισμός αυτής της ευθείας ισοδυναμεί με την εκτίμηση των συντελεστών παλινδρόμησης α και β. Οι πληθυσμιακοί συντελεστές μπορεί να εκτιμηθούν με τη βοήθεια ενός τυχαίου δείγματος, το οποίο λαμβάνεται από τον πληθυσμό και για το οποίο υπολογίζεται η αντίστοιχη δειγματική ευθεία της παλινδρόμησης. Ο προσδιορισμός των συντελεστών του δειγματικού υποδείγματος αποτελεί τη βάση για την εκτίμηση των αντίστοιχων πληθυσμιακών συντελεστών. Πριν όμως προσδιοριστεί η δειγματική ευθεία της παλινδρόμησης είναι απαραίτητο να επιβεβαιωθεί η γραμμική σχέση που υπάρχει

μεταξύ των δύο μεταβλητών στα δειγματικά δεδομένα. Η διαδικασία αυτή μπορεί να γίνει με τη βοήθεια ενός διαγράμματος διασποράς.

Ο προσδιορισμός της ευθείας είναι απαραίτητο να γίνει με τρόπο αντικειμενικό ώστε να διασφαλίζεται η βέλτιστη προσέγγιση των σημείων από αυτήν. Η μέθοδος η οποία συνήθως χρησιμοποιείται για το σκοπό αυτό είναι γνωστή ως **μέθοδος των ελαχίστων τετραγώνων**, ενώ η ευθεία που ορίζεται από αυτή ονομάζεται **ευθεία των ελαχίστων τετραγώνων**. Ο λόγος για τον οποίο χρησιμοποιείται η συγκεκριμένη ονομασία για τη μέθοδο, προκύπτει από την παρακάτω γεωμετρική διαδικασία προσδιορισμού της ευθείας.

Έστω ένα οποιοδήποτε σημείο του διαγράμματος διασποράς με συντεταγμένες (x_i, y_i) και έστω e_i η κατακόρυφη απόσταση του σημείου από μια οποιαδήποτε ευθεία που προσεγγίζει τα σημεία του διαγράμματος (Εικόνα 3.6)

Εικόνα 3.6



Η απόσταση e_i είναι η διαφορά της τιμής y_i από το σημείο $\hat{y}_i$ που ορίζεται από την κατακόρυφη προβολή του σημείου (x_i, y_i) επί της ευθείας. Δηλαδή

$$e_i = y_i - \hat{y}_i$$

Η απόσταση e_i ονομάζεται **υπόλοιπο (residual)** ή **σφάλμα (error)**. Αν όλα τα υπόλοιπα είναι ίσα με 0, θα έχουμε πλήρη προσαρμογή της ευθείας επί των σημείων του

διαγράμματος. Η πλήρης προσαρμογή της ευθείας είναι απίθανο να προκύψει (εκτός και αν οι δύο μεταβλητές είναι απολύτως εξαρτημένες η μία από την άλλη), μπορούν όμως να ελαχιστοποιηθούν οι κατακόρυφες αποστάσεις (τα υπόλοιπα) των σημείων από την ευθεία. Η ελαχιστοποίηση αυτή ισοδυναμεί με την ελαχιστοποίηση της ποσότητας:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

η οποία ονομάζεται **άθροισμα τετραγώνων των υπολοίπων (residual sum of squares)** ή **άθροισμα τετραγώνων των σφαλμάτων (error sum of squares)**.
Η ευθεία δηλαδή των ελαχίστων τετραγώνων κατασκευάζεται με την ελαχιστοποίηση του αθροίσματος τετραγώνων των σφαλμάτων.

Η διαδικασία προσδιορισμού της ευθείας των ελαχίστων τετραγώνων, η οποία συμβολικά ορίζεται από την εξίσωση

$$\hat{y} = \hat{\alpha} + \hat{\beta} x ,$$

απαιτεί τον προσδιορισμό των ποσοτήτων $\hat{\alpha}$ και $\hat{\beta}$ οι οποίες είναι εκτιμήσεις των πληθυσμιακών συντελεστών της παλινδρόμησης α και β. Από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

προκύπτει (με τη βοήθεια του διαφορικού λογισμού) ότι

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

και

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \text{ (διότι ισχύει } \bar{y} = \hat{\alpha} + \hat{\beta} \bar{x})$$

Οι παραπάνω εξισώσεις δίνουν την κλίση και την τεταγμένη στο σημείο 0 της ευθείας ελαχίστων τετραγώνων.

Εφόσον προσδιοριστούν οι συντελεστές $\hat{\alpha}$ και $\hat{\beta}$ του υποδείγματος μπορούν να αντικατασταθούν οι δειγματικές τιμές x_i στην εξίσωση της ευθείας ελαχίστων τετραγώνων και να υπολογιστούν από την εξίσωση οι αντίστοιχες εκτιμώμενες τιμές $\hat{y}_i$. Τοποθετώντας επί του διαγράμματος διασποράς τα σημεία (x_i, $\hat{y}_i$) προκύπτει η δειγματική ευθεία παλινδρόμησης.

Παράδειγμα 3.2

Στον Πίνακα 3.2 παρουσιάζονται οι τιμές της ημερήσιας ενεργειακής πρόσληψης (σε Kcal) 40 ενηλίκων ατόμων μαζί με την ηλικία τους

Πίνακας 3.2

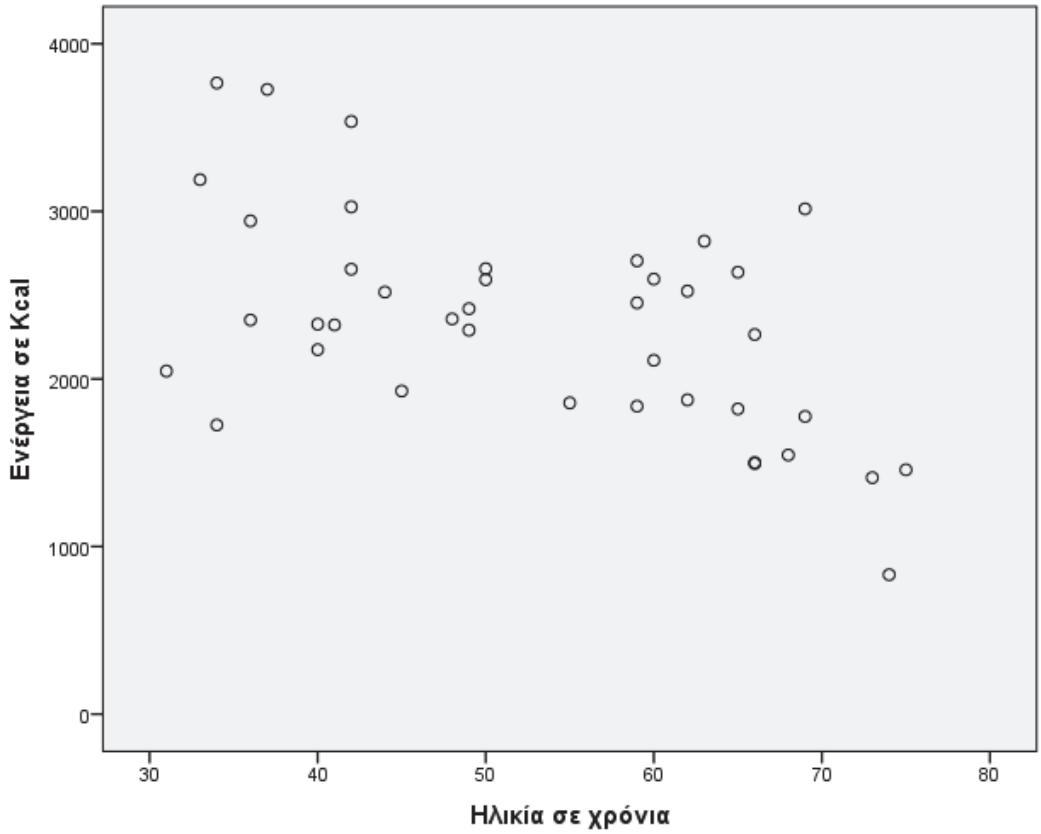
Τιμές ημερήσιας ενεργειακής πρόσληψης 40 ενηλίκων και ηλικίες αυτών

Αριθμός ατόμων	Ηλικία σε χρόνια	Ενέργεια σε Kcal
1	63	2.822
2	49	2.419
3	44	2.518
4	69	3.015
5	55	1.857
6	62	1.875
7	41	2.322
8	42	3.536
9	36	2.943
10	50	2.658
11	60	2.596
12	50	2.593
13	42	2.655
14	37	3.728
15	34	3.766
16	40	2.327
17	65	2.637
18	62	2.524
19	45	1.928
20	42	3.027
21	36	2.352
22	59	2.454
23	40	2.175
24	48	2.358
25	66	1.495
26	59	2.705
27	73	1.411
28	34	1.725
29	49	2.291
30	31	2.047
31	69	1.776
32	66	2.265
33	65	1.821
34	66	1.501
35	75	1.458
36	60	2.111
37	68	1.546
38	33	3.189
39	59	1.837
40	74	832

Μεταξύ της ηλικίας και της ημερήσιας ενεργειακής πρόσληψης υπάρχει γραμμική σχέση, σύμφωνα με την οποία όσο αυξάνεται η ηλικία, η ενεργειακή πρόσληψη ελαττώνεται. Η ύπαρξη της γραμμικής σχέσης μεταξύ των δύο μεταβλητών επιβεβαιώνεται από τη μορφή του διαγράμματος διασποράς που εμφανίζεται στο Εικόνα 3.7.

Εικόνα 3.7

Διάγραμμα διασποράς της ηλικίας και της ημερήσιας ενεργειακής πρόσληψης 40 ενηλίκων



Στο συγκεκριμένο διάγραμμα η ηλικία θεωρείται ανεξάρτητη μεταβλητή και τοποθετείται στον οριζόντιο άξονα, ενώ στον κατακόρυφο άξονα τοποθετείται η ενεργειακή πρόσληψη. Η σχέση των δύο μεταβλητών μπορεί να περιγραφεί με τη βοήθεια μιας ευθείας γραμμής, η οποία θα προσεγγίζει κατά το μέγιστο δυνατόν το σύνολο των σημείων που απεικονίζονται στο εσωτερικό του διαγράμματος.

Με τη βοήθεια του παρακάτω πίνακα θα βρούμε την εξίσωση της ευθείας ελαχίστων τετραγώνων που προσαρμόζεται στα 40 σημεία των τιμών της ηλικίας και της ενεργειακής πρόσληψης

Ηλικία σε χρόνια (x_i)	Ενέργεια σε Kcal (y_i)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
63	2.822	10	495	4.970,98	101,00
49	2.419	-4	92	-361,92	15,60
44	2.518	-9	191	-1.706,09	80,10
69	3.015	16	688	11.036,38	257,60
55	1.857	2	-470	-964,27	4,20
62	1.875	9	-452	-4.093,99	81,90
41	2.322	-12	-5	64,23	142,80
42	3.536	-11	1.209	-13.234,44	119,90
36	2.943	-17	616	-10.434,84	287,30
50	2.658	-3	331	-975,34	8,70
60	2.596	7	269	1.893,81	49,70
50	2.593	-3	266	-783,59	8,70
42	2.655	-11	328	-3.587,49	119,90
37	3.728	-16	1.401	-22.339,97	254,40
34	3.766	-19	1.439	-27.261,94	359,10
40	2.327	-13	0	4,86	167,70
65	2.637	12	310	3.730,98	145,20
62	2.524	9	197	1.779,46	81,90
45	1.928	-8	-399	3.175,03	63,20
42	3.027	-11	700	-7.660,89	119,90
36	2.352	-17	25	-417,39	287,30
59	2.454	6	127	766,08	36,60
40	2.175	-13	-152	1.973,26	167,70
48	2.358	-5	31	-151,59	24,50
66	1.495	13	-832	-10.862,49	170,30
59	2.705	6	378	2.284,63	36,60
73	1.411	20	-916	-18.373,32	402,00
34	1.725	-19	-602	11.415,01	359,10
49	2.291	-4	-36	143,68	15,60
31	2.047	-22	-280	6.154,23	481,80
69	1.776	16	-551	-8.849,57	257,60
66	2.265	13	-62	-813,99	170,30
65	1.821	12	-506	-6.101,82	145,20
66	1.501	13	-826	-10.784,19	170,30
75	1.458	22	-869	-19.169,72	486,20
60	2.111	7	-216	-1.525,44	49,70
68	1.546	15	-781	-11.759,69	226,50
33	3.189	-20	862	-17.189,42	398,00
59	1.837	6	-490	-2.966,77	36,60
74	832	21	-1.495	-31.477,64	443,10
$\sum_{i=1}^{40} x_i =$ 2.118	$\sum_{i=1}^{40} y_i =$ 93.095	0	0	$\sum_{i=1}^{40} (x_i - \bar{x})(y_i - \bar{y}) =$ -184.455	$\sum_{i=1}^{40} (x_i - \bar{x})^2$ = 6.834

Έχουμε

$$\bar{x} = \frac{1}{40}\sum_{i=1}^{40} x_i = \frac{2118}{40} = 52{,}95 \qquad \bar{y} = \frac{1}{40}\sum_{i=1}^{40} y_i = \frac{93095}{40} = 2327{,}38$$

$$\sum_{i=1}^{40}(x_i - \bar{x})(y_i - \bar{y}) = -184.455 \qquad \sum_{i=1}^{40}(x_i - \bar{x})^2 = 6.834$$

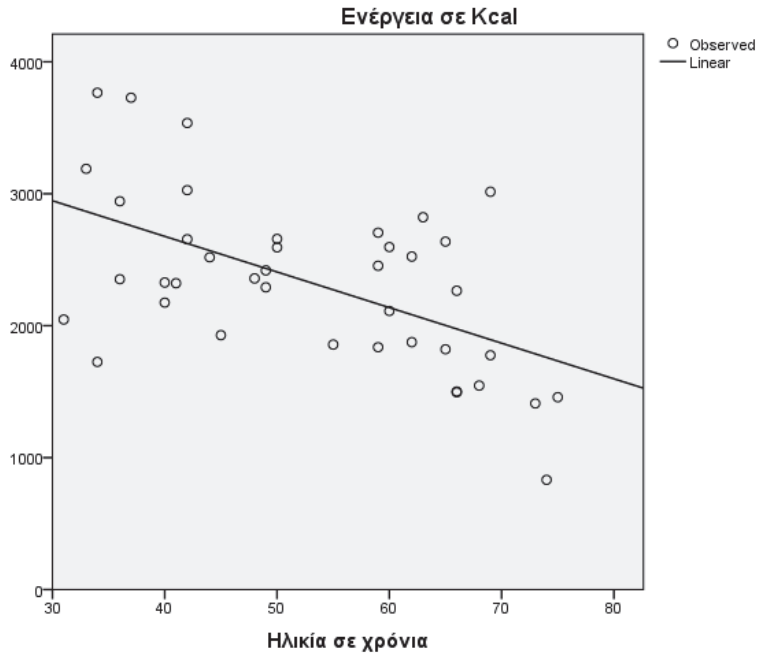
$$\hat{\beta} = \frac{\sum_{i=1}^{40}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{40}(x_i - \bar{x})^2} = \frac{-184455}{6834} = -27$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} = 2327{,}38 - (-27)\cdot 52{,}95 = 3756{,}6$$

Επομένως η ευθεία ελαχίστων τετραγώνων που προσαρμόζεται στα 40 σημεία των τιμών της ηλικίας και της ενεργειακής πρόσληψης έχει εξίσωση $\hat{y} = 3756{,}6 - 27x$.

Εικόνα 3.8

Ευθεία ελαχίστων τετραγώνων για τη σχέση ενεργειακής πρόσληψης με την ηλικία



Αυτή η ευθεία έχει το μικρότερο άθροισμα τετραγώνων των σφαλμάτων από οποιαδήποτε άλλη ευθεία που μπορεί να προσαρμοστεί στο σύνολο των σημείων του διαγράμματος διασποράς.

Η τιμή του συντελεστή $\hat{\alpha}$ είναι ίση με 3756,6 και θεωρητικά αποτελεί τη μέση τιμή της ημερήσιας ενεργειακής πρόσληψης που αντιστοιχεί σε ένα άτομο ηλικίας 0 ετών. Στο

συγκεκριμένο παράδειγμα, η τιμή 0 για την ηλικία δε έχει νόημα.

Η κλίση $\hat{\beta}$ της ευθείας είναι ίση με –27 και ερμηνεύεται ως η μέση μείωση της ημερήσιας ενεργειακής πρόσληψης για κάθε έτος αύξησης ηλικίας.

Παράδειγμα 3.3

Είναι γνωστό ότι ένα τμήμα των εμπορευμάτων που πωλούνται από τις επιχειρήσεις επιστρέφεται, σε μερικές περιπτώσεις, από τους αγοραστές στον πωλητή (επιστροφές εμπορευμάτων) για διάφορους λόγους. Για παράδειγμα, επειδή δεν τηρήθηκαν οι προδιαγραφές που είχαν συμφωνηθεί για το εμπόρευμα, επειδή καθυστέρησε πολύ η παράδοση του εμπορεύματος, ή επειδή υπάρχουν ελαττωματικά εμπορεύματα κ.τ.λ.

Μια εμπορική επιχείρηση είχε την περασμένη χρονιά πωλήσεις εμπορευμάτων αξίας 50 εκατ. ευρώ και επιστροφές εμπορευμάτων αξίας 5 εκατ. ευρώ. Ο νέος γενικός διευθυντής θέλοντας να εξακριβώσει αν οι επιστροφές των εμπορευμάτων γίνονται σε όρια επιτρεπτά, αναθέτει στο διευθυντή πωλήσεων τη διερεύνηση του θέματος.

Η διεύθυνση πωλήσεων, ανάμεσα στα άλλα στοιχεία που συνέλεξε για μελέτη και διερεύνηση του θέματος, συγκέντρωσε και τις περσινές ετήσιες πωλήσεις εμπορευμάτων με τις αντίστοιχες επιστροφές για τις 10 πρώτες σε πωλήσεις επιχειρήσεις του ίδιου κλάδου εμπορίας. Τα στοιχεία παρουσιάζονται στον παρακάτω πίνακα.

Αξία πωλήσεων (X)	20	30	40	40	50	50	60	70	80	90
Αξία επιστροφών (Y)	1	3	3	4	5	6	6	8	9	10

Ο διευθυντής πωλήσεων θέλησε να διαπιστώσει:

α) Το βαθμό συσχέτισης που υπάρχει μεταξύ της αξίας των εμπορευμάτων που πωλούνται ετησίως και της αξίας αυτών που επιστρέφονται στην ομάδα των 10 επιχειρήσεων.

β) Την εξίσωση παλινδρόμησης που θα μπορούσε να εκφράσει την σχέση που υπάρχει μεταξύ της αξίας των πωλήσεων και της αξίας των επιστροφών κατά έτος.

γ) Ποια θα ήταν η αναμενόμενη αξία των εμπορευμάτων που επιστρέφονται, αν οι πωλήσεις αυξάνονταν κατά μια μονάδα, δηλαδή κατά 1 εκατ. ευρώ.

δ) Ποια θα ήταν η αξία των επιστροφών που θα είχε μια επιχείρηση, αν οι ετήσιες πωλήσεις της είχαν ύψος 35 εκατ. ευρώ.

ε) Ποια η διάμεση αξία των πωλήσεων κατά έτος αυτής της ομάδας επιχειρήσεων.

στ) Ποια η διάμεση αξία των επιστροφών κατά έτος.

Λύση

α) Εφαρμόζοντας τον τύπο του συντελεστή συσχέτισης του Pearson στα παραπάνω δεδομένα, έχουμε:

για τη μεταβλητή X, που αναπαριστά την αξία πωλήσεων

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{20+30+...+90}{10} = 53$$

για τη μεταβλητή Y, που αναπαριστά την αξία επιστροφών

$$\bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = \frac{1+3+...+10}{10} = 5,5$$

για τις δύο μεταβλητές X και Y

$$\sum_{i=1}^{10}(x_i-\bar{x})(y_i-\bar{y}) =$$

$$= (20-53)(1-5,5)+(30-53)(3-5,5)+...+(90-53)(10-5,5) = 565$$

και

$$\sum_{i=1}^{10}(x_i-\bar{x})^2 = (20-53)^2+(30-53)^2+...+(90-53)^2 = 4410$$

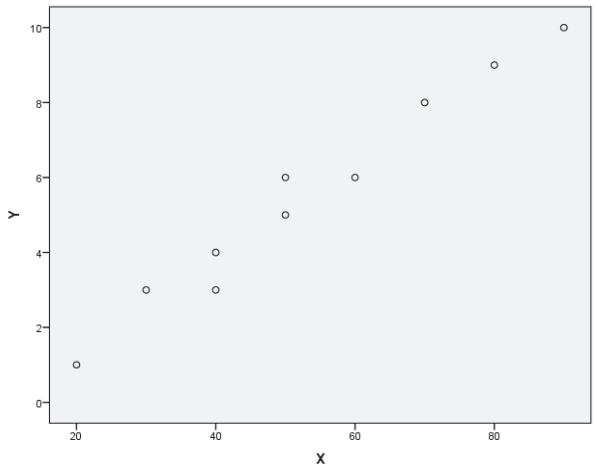
$$\sum_{i=1}^{10}(y_i-\bar{y})^2 = (1-5,5)^2+(3-5,5)^2+...+(10-5,5)^2 = 74,5$$

Άρα ο συντελεστής συσχέτισης Pearson ισούται με:

$$r = \frac{\sum_{i=1}^{10}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{10}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{10}(y_i-\bar{y})^2}} = \frac{565}{\sqrt{4410}\cdot\sqrt{74,5}} = 0,986$$

Από την τιμή του συντελεστή συσχέτισης του Pearson προκύπτει ότι υπάρχει ισχυρή θετική συσχέτιση μεταξύ της αξίας πωλήσεων και της αξίας επιστροφών.

Η ισχυρή αυτή θετική συσχέτιση, επιβεβαιώνεται και από το παρακάτω διάγραμμα διασποράς.



β) Θα υπολογίσουμε τις εκτιμήτριες ελαχίστων τετραγώνων $\hat{a}$, $\hat{\beta}$ και θα προσδιορίσουμε την εξίσωση της ευθείας παλινδρόμησης.

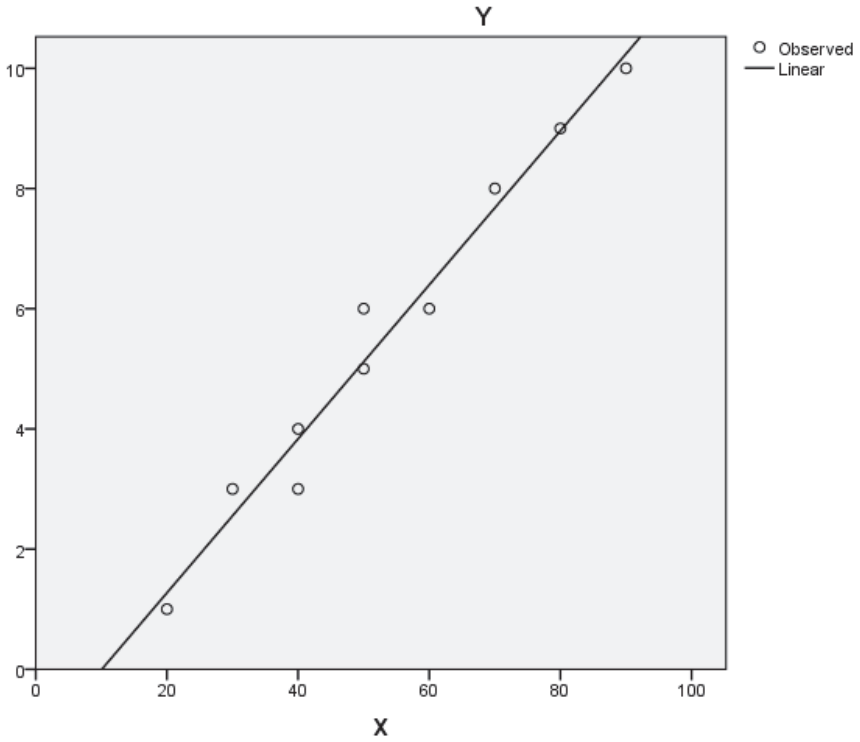
$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{565}{4410} = 0{,}128$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} = 5{,}5 - 0{,}128 \cdot 53 = -1{,}29$$

Άρα η ευθεία που προσαρμόζεται καλύτερα στα δεδομένα είναι:

$$\hat{y} = -1{,}29 + 0{,}128x$$

Η γραφική της παράσταση παρουσιάζεται στο παρακάτω σχήμα.



Παρατηρούμε ότι, αν $\hat{y} = 0$, τότε $x \cong 10$ (γιατί αν $\hat{y} = 0$ τότε $-1{,}29 + 0{,}128x = 0$ ή $0{,}128x = 1{,}29$ ή $x \cong 10$) που σημαίνει ότι για πωλήσεις εμπορευμάτων αξίας περίπου μέχρι και 10 εκατομμυρίων ευρώ δεν υπάρχουν επιστροφές εμπορευμάτων

γ) Αν οι πωλήσεις αυξάνονταν κατά μια μονάδα τότε,

$$\hat{y}^* = -1{,}29 + 0{,}128(x + 1) = \underbrace{-1{,}29 + 0{,}128x}_{\hat{y}} + 0{,}128 = \hat{y} + 0{,}128$$

Συμπεραίνουμε ότι, αν οι πωλήσεις αυξάνονταν κατά μια μονάδα, δηλαδή κατά 1 εκατομμύριο ευρώ, τότε το ύψος των επιστρεφόμενων εμπορευμάτων θα αυξανόταν κατά

0,128 εκατομμύρια ευρώ, όσο ακριβώς και η τιμή του συντελεστή $\hat{\beta}$, ($\hat{\beta} = 0,128$)

δ) Αν οι ετήσιες πωλήσεις μιας επιχείρησης ήταν της τάξης των 35 εκατομμυρίων ευρώ, τότε το ύψος των επιστρεφόμενων εμπορευμάτων θα υπολογιζόταν από τον τύπο:

$\hat{y} = -1,29 + 0,128x$ ή

$\hat{y} = -1,29 + 0,128 \cdot 35 = 3,19$

Δηλαδή, η αξία των εμπορευμάτων που επιστράφηκαν θα ήταν 3,19 εκατ. ευρώ.

ε) Η διάμεση αξία των πωλήσεων κατ' έτος υπολογίζεται ως εξής:

Αξία πωλήσεων: 20, 30, 40, 40, **50**, **50**, 60, 70, 80, 90

Ο αριθμός των επιχειρήσεων είναι 10 (άρτιος) άρα η διάμεσος θα ισούται με $\frac{50+50}{2} = 50$, δηλαδή η διάμεσος αξία πωλήσεων θα είναι 50 εκατ. ευρώ.

στ) Η διάμεση αξία των προϊόντων που επιστράφηκαν κατ' έτος υπολογίζεται ως εξής:

Αξία επιστροφών: 1, 3, 3, 4, **5**, **6**, 6, 8, 9, 10.

Ο αριθμός των επιχειρήσεων είναι 10 (άρτιος) άρα η διάμεσος ισούται με $\frac{5+6}{2} = 5,5$, δηλαδή η διάμεσος αξία των επιστροφών θα είναι 5,5 εκατ. ευρώ.

3.5 Συντελεστής προσδιορισμού

Μετά τον προσδιορισμό της ευθείας ελαχίστων τετραγώνων, δια μέσου της εξίσωσης $\hat{y} = \hat{\alpha} + \hat{\beta} x$ απομένει η αξιολόγηση της προσαρμογής της ευθείας αυτής επί των δειγματικών τιμών. Ένας τρόπος για να αξιολογήσουμε την προσαρμογή της ευθείας ελαχίστων τετραγώνων είναι να υπολογίσουμε το **συντελεστή προσδιορισμού**. Ο συντελεστής προσδιορισμού της δειγματικής ευθείας της παλινδρόμησης συμβολίζεται με R^2 και ορίζεται ως το τετράγωνο του δειγματικού συντελεστή συσχέτισης, δηλαδή:

$$R^2 = r^2$$

Επειδή ο δειγματικός συντελεστής συσχέτισης παίρνει τιμές στο διάστημα [−1, 1], ο συντελεστής προσδιορισμού παίρνει τιμές στο διάστημα [0, 1]. Όταν $R^2 = 1$, όλα τα σημεία που αναπαριστούν τις δειγματικές τιμές των X και Y βρίσκονται τοποθετημένα επί της ευθείας των ελαχίστων τετραγώνων. Όταν $R^2 = 0$, δεν υπάρχει γραμμική σχέση μεταξύ των δειγματικών τιμών X και Y.

Ο συντελεστής προσδιορισμού ως μέτρο της προσαρμογής της ευθείας των ελαχίστων τετραγώνων επί των δειγματικών τιμών, ορίζεται πρωτογενώς από την ανάλυση της συνολικής διασποράς της εξαρτημένης μεταβλητής Y σε επιμέρους συνιστώσες. Χρησιμοποιώντας την ταυτότητα

$(y_i - \hat{y}_i) = (y_i - \bar{y}) - (\hat{y}_i - \bar{y})$, i = 1, 2, …, n,

η οποία ισχύει για τις δειγματικές τιμές της μεταβλητής Y, υψώνοντας και τα δύο μέλη της στο τετράγωνο και αθροίζοντας για i = 1, 2, …, n, παίρνουμε

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[(y_i - \bar{y}) - (\hat{y}_i - \bar{y})]^2 \quad \text{ή}$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[(y_i - \bar{y})^2 + (\hat{y}_i - \bar{y})^2 - 2(y_i - \bar{y})(\hat{y}_i - \bar{y})] \quad \text{ή}$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 - 2\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y}) \qquad (1)$$

Επειδή

$\hat{y} = \hat{\alpha} + \hat{\beta} x$, ισχύει $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

Επίσης ισχύει

$$\bar{y} = \hat{\alpha} + \hat{\beta} \bar{x}$$

Επομένως

$$-2\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y}) = -2\sum_{i=1}^{n}(y_i - \bar{y})(\hat{\alpha} + \hat{\beta} x_i - \hat{\alpha} - \hat{\beta} \bar{x}) =$$

$$= -2\hat{\beta}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x}) \tag{2}$$

Από την ευθεία παλινδρόμησης ισχύει

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \text{ άρα}$$

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \hat{\beta}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{3}$$

Η σχέση (2) λόγω της (3) γράφεται

$$-2\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{y}) = -2\hat{\beta}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x}) = -2\hat{\beta}^2\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{4}$$

Επίσης αν στην ευθεία παλινδρόμησης $\hat{y}=\hat{\alpha}+\hat{\beta}x$ αντικαταστήσουμε $\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}$ προκύπτει $\hat{y}=\bar{y}-\hat{\beta}\bar{x}+\hat{\beta}x$ ή $\hat{\beta}x-\hat{\beta}\bar{x}=\hat{y}-\bar{y}$ ή $\hat{\beta}(x-\bar{x})=\hat{y}-\bar{y}$ ή $\hat{\beta}(x_i-\bar{x})=\hat{y}_i-\bar{y}$

Επομένως,

$$x_i-\bar{x}=\frac{\hat{y}_i-\bar{y}}{\hat{\beta}} \text{ ή } (x_i-\bar{x})^2=\frac{(\hat{y}_i-\bar{y})^2}{\hat{\beta}^2} \tag{5}$$

Η σχέση (4) λόγω της (5) γράφεται

$$-2\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{y}) = -2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

Η σχέση (1), λόγω της τελευταίας γράφεται

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2-2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \text{ ή}$$

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \text{ ή}$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{6}$$

Στην παραπάνω εξίσωση,

- η ποσότητα $\sum_{i=1}^{n}(y_i-\bar{y})^2$ = SST, ονομάζεται **συνολικό άθροισμα τετραγώνων (total sum of squares)** και αποτελεί μέτρο της συνολικής διασποράς των δειγματικών τιμών της Υ γύρω από τη μέση τιμή τους $\bar{y}$.

- η ποσότητα $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ = SSR, ονομάζεται **άθροισμα τετραγώνων επεξηγούμενο**

από τη γραμμική παλινδρόμηση (regression sum of squares) και εκφράζει τη διασπορά των εκτιμώμενων τιμών της Y γύρω από τη δειγματική μέση τιμή $\bar{y}$.

Η ποσότητα αυτή αποτελεί μέτρο της διασποράς των δειγματικών τιμών της Y, που ερμηνεύεται από το υπόδειγμα της γραμμικής παλινδρόμησης (της διασποράς δηλαδή που οφείλεται στη γραμμική επίδραση της X επί της Y). Τέλος,

- η ποσότητα $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 =$ SSE, είναι το γνωστό **άθροισμα τετραγώνων των σφαλμάτων (error sum of squares)** και εκφράζει τη διασπορά των δειγματικών τιμών της Y γύρω από την εκτιμώμενη ευθεία της παλινδρόμησης.

Όσο μικρότερο είναι το άθροισμα τετραγώνων των σφαλμάτων, τόσο πλησιέστερα βρίσκονται οι δειγματικές τιμές της εξαρτημένης μεταβλητής Y στην ευθεία ελαχίστων τετραγώνων.

Ισχύει επομένως ότι:

Συνολικό άθροισμα τετραγώνων = ***Άθροισμα τετραγώνων επεξηγούμενο από τη γραμμική παλινδρόμηση*** + ***Άθροισμα τετραγώνων των σφαλμάτων***

Η (6) ισοδύναμα γράφεται:

SST = SSR + SSE

Για να είναι η προσαρμογή της ευθείας ελαχίστων τετραγώνων επί των δειγματικών δεδομένων όσο το δυνατόν καλύτερη, θα πρέπει το άθροισμα των τετραγώνων των σφαλμάτων να είναι όσο το δυνατόν μικρότερο και επομένως, σύμφωνα με την προηγούμενη εξίσωση, το άθροισμα τετραγώνων το επεξηγούμενο από τη γραμμική παλινδρόμηση να είναι όσο το δυνατόν μεγαλύτερο. Το ποσοστό επομένως, του συνολικού αθροίσματος τετραγώνων που επεξηγείται από τη γραμμική παλινδρόμηση υπολογιζόμενο από το λόγο

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$ ή

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

αποτελεί μέτρο της προσαρμογής της ευθείας των ελαχίστων τετραγώνων επί των δειγματικών τιμών και ορίζει το συντελεστή προσδιορισμού. Ο συντελεστής προσδιορισμού επομένως, μπορεί να ερμηνευτεί ως το ποσοστό της μεταβλητότητας των τιμών της Y που επεξηγείται από το υπόδειγμα της γραμμικής παλινδρόμησης.

Με απλούς μετασχηματισμούς αποδεικνύεται ότι:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = \frac{\sum_{i=1}^{n}(\hat{\alpha} + \hat{\beta}x_i - \hat{\alpha} - \hat{\beta}\bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = \hat{\beta}^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} =$$

$$= \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^2 \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}\right] = \frac{\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y}_i)^2\right]} = r^2$$

δηλαδή ότι $R^2 = r^2$

Παράδειγμα 3.4

Για τις τιμές της σχέσης της ημερήσιας ενεργειακής πρόσληψης με την ηλικία 40 ενηλίκων ατόμων του Παραδείγματος 3.2, η ευθεία ελαχίστων τετραγώνων που προσαρμόζεται στα 40 σημεία έχει εξίσωση:

$$\hat{y} = 3756{,}6 - 27x$$

Στο Παράδειγμα 3.2, υπολογίσαμε τα εξής

$$\sum_{i=1}^{40}(x_i - \bar{x})(y_i - \bar{y}) = -184.455 \qquad \sum_{i=1}^{40}(x_i - \bar{x})^2 = 6.834$$

Θα υπολογίσουμε επίσης το $\sum_{i=1}^{40}(y_i - \bar{y})^2 = 16.240.355$

Άρα ο συντελεστής συσχέτισης Pearson ισούται με:

$$r = \frac{\sum_{i=1}^{40}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{40}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{40}(y_i - \bar{y})^2}} = \frac{-184455}{\sqrt{6834}\cdot\sqrt{16240355}} = -0{,}554$$

$r^2 = (-0{,}554)^2 = 0{,}307$

Επομένως, ο συντελεστής προσδιορισμού είναι

$R^2 = r^2 = 0{,}307$

Η τιμή αυτή δηλώνει ότι το 30,7% της μεταβλητότητας της ημερήσιας ενεργειακής πρόσληψης Y ερμηνεύεται από την ηλικία των ενήλικων ατόμων X.

Κεφάλαιο 4 Στατιστική και επιχειρήσεις

4.1 Ο ρόλος της στατιστικής στην επιχείρηση

Η μέχρι τώρα εμπειρία από τις εφαρμογές της Στατιστικής στον επιχειρηματικό τομέα έχει καταστήσει τις στατιστικές μεθόδους απαραίτητο εργαλείο στην άσκηση της σύγχρονης επιχειρηματικής δραστηριότητας.

Τα στατιστικά στοιχεία που αφορούν την περιουσιακή κατάσταση και εξέλιξη των επιχειρήσεων, όπως αυτά που αναφέρονται στα αποθέματα, τον κεφαλαιουχικό εξοπλισμό, τις αγορές και τις πωλήσεις, τα αποτελέσματα των διαφημιστικών δαπανών κ.τ.λ., μπορούν, αφού μελετηθούν κατάλληλα, να βοηθήσουν σημαντικά στη λήψη ορθών αποφάσεων.

Στις γνωστές έρευνες αγοράς, που διεξάγουν οι σύγχρονες επιχειρήσεις, ο ρόλος της Στατιστικής είναι καθοριστικός και τα συμπεράσματα είναι πολύτιμα για το μέλλον της επιχείρησης.

Με τη βοήθεια της Στατιστικής οι επιχειρήσεις μπορούν να κάνουν προβλέψεις για την εξέλιξη των πωλήσεών τους, να εξασφαλίζουν τον έλεγχο της ποιότητας των παραγόμενων προϊόντων, να ελέγχουν το κόστος παραγωγής κ.τ.λ.

Η ανάπτυξη ολόκληρων επιστημονικών κλάδων, όπως είναι η Οικονομετρία, η Επιχειρησιακή Έρευνα, η Διαφήμιση κ.α., οφείλεται σε μεγάλο βαθμό στη Στατιστική, της οποίας η χρησιμότητα γίνεται κάθε μέρα και περισσότερο εμφανής.

Συνεπώς, η Στατιστική είναι απαραίτητη στη σύγχρονη Διοίκηση Επιχειρήσεων, γιατί η λήψη επιχειρηματικών αποφάσεων πρέπει να βασίζεται σε επιστημονικές μεθόδους και όχι μόνο στη διαίσθηση του επιχειρηματία ή στην τύχη.

4.1.1 Η σύγχρονη διοικητική των επιχειρήσεων (management)

Η οργάνωση και η διοίκηση των επιχειρήσεων (διοικητική των επιχειρήσεων) γίνεται σήμερα με επιστημονικές μεθόδους. Ο ταχύτατος ρυθμός ανάπτυξης της τεχνολογίας, ιδιαίτερα στον τομέα των υπολογιστών και των τηλεπικοινωνιών, η αύξηση του κεφαλαίου των επιχειρήσεων, παράλληλα με τον πολυεθνικό χαρακτήρα που έχουν πολλές από αυτές, και η παγκοσμιοποίηση της οικονομίας που συντελείται, έχουν επηρεάσει σε μεγάλο βαθμό τη σύγχρονη διοίκηση των επιχειρήσεων.

Όλα τα παραπάνω οδηγούν την επιστήμη της διοίκησης επιχειρήσεων σε αναζήτηση νέων προτύπων και μεθόδων της οργανωτικής δομής και λειτουργίας των επιχειρήσεων και σε επαναπροσδιορισμό των θέσεών της. Σήμερα, θα μπορούσαμε να θεωρήσουμε τη

σύγχρονη διοικητική των επιχειρήσεων ως ένα σύστημα δράσης, που μέσα από τον προγραμματισμό, το συντονισμό και τον έλεγχο στην όλη λειτουργία της επιχείρησης, στοχεύει στη σωστή διαχείριση των πόρων της για την επίτευξη του σκοπού της, που είναι το κέρδος.

4.1.2 Έρευνες αγορών - Διαφήμιση

Ένα από τα πλέον σημαντικά και ζωτικής σημασίας πεδία δραστηριότητας των σύγχρονων επιχειρήσεων είναι αυτά της έρευνας αγορών και της διαφήμισης. Εδώ η συνδρομή και η εφαρμογή της Στατιστικής είναι σοβαρότατη. Οι έρευνες αγορών πολύ απλά στοχεύουν στην ακριβή χαρτογράφηση των αναγκών της αγοράς σε αγαθά και, φυσικά, στον προσδιορισμό του βαθμού της ζήτησης τους.

Για μια επιτυχή έρευνα αγοράς απαιτείται κατηγοριοποίηση των υποψηφίων αγοραστών-καταναλωτών, ώστε να προσδιοριστούν σε ποια ή ποιες κατηγορίες αγοραστών θα απευθύνονται τα διαφημιστικά μηνύματά και στη συνέχεια τα προϊόντα της επιχείρησης.

Κριτήρια κατηγοριοποίησης είναι η ηλικία των υποψήφιων αγοραστών, το φύλο, το επάγγελμα, το μορφωτικό επίπεδο, τα ενδιαφέροντά τους κ.α.

Συνήθως στις έρευνες γίνεται χρήση ερωτηματολογίων για να διερευνηθούν οι προθέσεις του αγοραστικού κοινού. Τα βασικά χαρακτηριστικά των ερωτηματολογίων είναι:

1. Η σαφήνεια και απλότητα των ερωτήσεων
2. Η διακριτικότητα στον τρόπο διατύπωσης των ερωτήσεων, ώστε να μη θίγεται ή να μη φοβάται να απαντήσει ο ερωτώμενος
3. Η συντομία στις απαντήσεις των ερωτήσεων
4. Η αποδοχή της σπουδαιότητας της έρευνας
5. Η αποφυγή διπλών ερωτήσεων
6. Η αποφυγή υπόδειξης ορισμένου είδους απάντησης

Στις έρευνες αγοράς ο ρόλος της Στατιστικής είναι καθοριστικός. Πρέπει να απαντήσει για το μέγεθος του δείγματος που θα χρησιμοποιηθεί, να προσδιορίσει τον τρόπο επιλογής του ώστε να είναι αντιπροσωπευτικό, την κοινωνική διαστρωμάτωση που θα έχει, κ.τ.λ. Κατόπιν πρέπει να γίνει η κατάλληλη στατιστική επεξεργασία, και η παρουσίαση των αποτελεσμάτων που προέκυψαν κ.α., ώστε η διοίκηση της επιχείρησης να πάρει ασφαλείς αποφάσεις σε καίρια ζητήματα, όπως για παράδειγμα, την πραγματοποίηση ή όχι επενδύσεων, την επέκταση της εμπορικής δραστηριότητας σε άλλες αγορές κ.τ.λ.

Ένα άλλο σημαντικό βήμα είναι η διαπίστωση των τάσεων της αγοράς, ως προς την τιμή, τη διακίνηση και την προώθηση του προϊόντος που ενδιαφέρει, καθώς και ο τρόπος που θα φτάσει το μήνυμα παραγωγής του νέου προϊόντος σε όλους τους καταναλωτές ώστε να το δεχθεί και ο πλέον δύσπιστος αγοραστής. Στο σημείο αυτό είναι αναγκαία η χρησιμοποίηση της διαφήμισης, για να φτάσει το μήνυμα σε όλους τους αγοραστές, να σπάσει η λεγόμενη αδράνεια της αγοράς και να δοκιμάσει ο καταναλωτής ένα νέο προϊόν.

Φυσικά, υπάρχουν οι έρευνες αγορών και οι διαφημίσεις, που γίνονται, για να διατηρηθεί ένα παλαιό προϊόν στην αγορά ή να καλυτερεύσει τη θέση του στις προτιμήσεις των αγοραστών. Σε κάθε περίπτωση οι στατιστικές έρευνες είναι αναγκαίες ώστε να διαπιστώνονται τα αποτελέσματα μιας διαφήμισης, να επισημαίνονται τυχόν τρωτά σημεία στην όλη διακίνηση του προϊόντος που είναι άγνωστα, με αποτέλεσμα η κάθε επιχείρηση να παρεμβαίνει έγκαιρα, για να αποφεύγονται δυσάρεστες εξελίξεις.

4.1.3 Λήψη επιχειρηματικών αποφάσεων

Η συμβολή της Στατιστικής στη διαδικασία λήψης επιχειρηματικών αποφάσεων είναι ουσιώδης, ιδιαίτερα μάλιστα όταν η αναφορά γίνεται σε συλλογικό επίπεδο υψηλόβαθμων στελεχών και όχι τόσο στις ατομικές καθημερινές και συχνές αποφάσεις που λαμβάνονται από τους υπαλλήλους-στελέχη της επιχείρησης.

Το τμήμα ή η διεύθυνση Στατιστικής είναι το κατεξοχήν αρμόδιο για την επεξεργασία, παρουσίαση και ανάλυση των εσωτερικών δεδομένων της επιχείρησης, που λαμβάνονται κυρίως από το Λογιστήριο, τη διεύθυνση πωλήσεων, τη διεύθυνση προσωπικού κ.τ.λ. και τα οποία αφορούν κυρίως τον έλεγχο της ομαλής εσωτερικής λειτουργίας της επιχείρησης, της πορείας των πωλήσεων κ.τ.λ. Η διεύθυνση Στατιστικής είναι επίσης υπεύθυνη για τη συλλογή πληροφοριών που αφορούν το ύψος των συνολικών πωλήσεων του κλάδου που δραστηριοποιείται η επιχείρηση, το ύψος των διαφημιστικών δαπανών και των προγραμματισμένων επενδύσεων των άλλων ομοειδών επιχειρήσεων κ.τ.λ. Όλα αυτά τα δεδομένα τα αναλύει με σκοπό την εξαγωγή χρήσιμων συμπερασμάτων που θα βοηθήσουν ουσιαστικά στη λήψη επιχειρηματικών αποφάσεων από τη διοίκηση της επιχείρησης.

Επίσης, η Στατιστική αναλύοντας σειρά δεδομένων προηγούμενων χρονικών περιόδων, της επιχείρησης και του κλάδου στον οποίο ανήκει, όπως για παράδειγμα, τις ετήσιες δαπάνες των τελευταίων 10 ετών ή τα ετήσια έσοδα της τελευταίας 10ετίας, έχει τη δυνατότητα με ικανοποιητική προσέγγιση να προβλέψει και να πληροφορήσει τη διοίκηση της επιχείρησης για τις πωλήσεις του επόμενου έτους ή της επόμενης πενταετίας. Έτσι, για παράδειγμα, η επιχείρηση οδηγείται σε μια ορθή λήψη αποφάσεων ως προς τον πενταετή προγραμματισμό των κεφαλαιουχικών επενδύσεών της.

4.1.4 Ποιότητα διαχείρισης – Αποτελεσματική διοίκηση

Προαπαιτούμενο προσόν για την παροχή υπηρεσιών ή αγαθών υψηλής ποιότητας από μέρους της επιχείρησης, ώστε να επιβιώσει και να αναπτυχθεί, είναι να διακρίνεται και η ίδια για την ποιότητα της διαχείρισής της και την αποτελεσματική διοίκηση.

Κατ' αρχήν ας καθορίσουμε πότε ένα προϊόν είναι ποιοτικά καλό. Ένα προϊόν θεωρείται ποιοτικά καλό, όταν ανταποκρίνεται στις προδιαγραφές και στις επιθυμίες του καταναλωτή, τις οποίες οι επιστημονικά και ποιοτικά οργανωμένες επιχειρήσεις έχουν καταγράψει από προηγούμενες έρευνες αγορών. Είναι γνωστό ότι καμία διαφημιστική εταιρεία δεν θα αναλάβει να παρουσιάσει ένα κακό προϊόν ως καλό, αλλά θα προσπαθήσει ένα καλό προϊόν να το προβάλλει σαν καλύτερο.

Για να επιτευχθούν τα παραπάνω, η επιχείρηση θα πρέπει να έχει μία καλή σχέση κόστους-απόδοσης. Αυτό σημαίνει ότι η επιχείρηση έχει μια διαχείριση ποιότητας, η οποία τη βοηθά να διακριθεί από τις υπόλοιπες του κλάδου της. Η επιτυχία των στόχων της (αποτελεσματικότητα) είναι συνάρτηση των ποιοτικών χαρακτηριστικών της. Ενδεικτικά αναφέρουμε μερικά ποιοτικά χαρακτηριστικά μιας επιχείρησης:

- Ορθή διαχείριση των ανθρώπινων πόρων. Αυτό σημαίνει ενεργοποίηση και ουσιαστική συμμετοχή των εργαζομένων σε κάθε εκδήλωση της επιχειρηματικής δραστηριότητας (ενημέρωση για τους στόχους της επιχείρησης, συνεχής επιμόρφωση, ενθάρρυνση στην ανάπτυξη πρωτοβουλιών, κ.τ.λ.).
- Υψηλού επιπέδου χρηματοοικονομική διοίκηση, διακρίνοντας τους ευνοϊκούς για την επιχείρηση τρόπους χρηματοδότησης, κάνοντας σωστή αξιολόγηση επενδύσεων και διαχείριση οικονομικών κινδύνων, κ.τ.λ. Τομέας ιδιαίτερα δύσκολος που απαιτεί λεπτούς χειρισμούς και πολύ εξειδικευμένα στελέχη.
- Δυνατότητες συνεχούς ελέγχου όλων των φάσεων της παραγωγής του προϊόντος και υψηλές προδιαγραφές ποιοτικού ελέγχου των προϊόντων.
- Τακτικές έρευνες των προθέσεων του καταναλωτικού κοινού, καλή επικοινωνιακή σχέση με την πελατεία και τους συνεργάτες.
- Ταχεία ενσωμάτωση νέων τεχνολογιών, που θα βελτιώσουν ποιοτικά το προϊόν, και νέων μεθόδων που θα βελτιώσουν τη λειτουργία της επιχείρησης.
- Προσπάθειες ώστε η επιχείρηση να είναι καινοτόμος στην όλη λειτουργία και πολιτική της, για να γίνει από τις επώνυμες του κλάδου της.
- Εμμονή στην τήρηση των αρχών λειτουργίας που έχει θέσει η επιχείρηση και στην τήρηση προδιαγραφών ποιότητας, ώστε να μπορέσει να δημιουργήσει η επιχείρηση αυτό που λέμε επώνυμο προϊόν.

Η συμβολή της Στατιστικής, για να επιτύχει μια επιχείρηση την ποιοτική διάσταση τόσο στην παραγωγή του προϊόντος της όσο και στη λειτουργία της, είναι σημαντική.

4.1.5 Η θέση της Στατιστικής στον ενοποιημένο οικονομικό χώρο

Για όλους τους διεθνείς οργανισμούς, οικονομικούς ή όχι, για παράδειγμα, ΔΝΤ, ΟΟΣΑ, Ο.Η.Ε, Διεθνής Τράπεζα κ.τ.λ., η Στατιστική αποτελεί ένα από τα σοβαρότερα στηρίγματά τους και τα πιο σημαντικά εργαλεία επίτευξης των σκοπών τους. Η Στατιστική θα τους δώσει στοιχεία επεξεργασμένα για να αποκτήσουν μια σωστή εικόνα για ότι θέλουν να μελετήσουν ή να πιστοποιήσουν την επιτυχία ή όχι των μέτρων και των προγραμμάτων που εφάρμοσαν.

Στη χώρα μας υπάρχει η Ελληνική Στατιστική Αρχή (ΕΛ.ΣΤΑΤ.) πρώην Εθνική Στατιστική Υπηρεσία της Ελλάδος (Ε.Σ.Υ.Ε.), που συγκεντρώνει στατιστικά στοιχεία, τα οποία και δημοσιεύει στην ετήσια στατιστική επετηρίδα ή στο μηνιαίο στατιστικό δελτίο ή στο διαδίκτυο και έτσι μπορεί κάθε ενδιαφερόμενος να έχει τα απαραίτητα στοιχεία για το αντικείμενο έρευνάς του. Εκτός από την ΕΛΣΤΑΤ υπάρχουν και άλλοι φορείς που συγκεντρώνουν στατιστικά στοιχεία. Τέτοιοι φορείς είναι τα ινστιτούτα ερευνών, δημόσιοι και ιδιωτικοί οργανισμοί, τράπεζες κ.τ.λ.

Για την Ελλάδα είναι σημαντικό το γεγονός ότι, λόγω της συμμετοχής της στην Ευρωπαϊκή Ένωση (Ε.Ε.), έχει τη δυνατότητα της εύκολης πρόσβασης στα στοιχεία και τις πληροφορίες που αφορούν τις άλλες χώρες της Ε.Ε., μέσω της Στατιστικής Υπηρεσίας της Ευρώπης της Eurostat. Έτσι, είναι πιθανό, στοιχεία, που δεν υπάρχουν στην Ελληνική Στατιστική Αρχή (ΕΛΣΤΑΤ) και τις άλλες εθνικές στατιστικές υπηρεσίες των άλλων κρατών, να υπάρχουν στην Eurostat. Με τη βοήθεια της Eurostat επιχειρείται η ύπαρξη ενιαίας ονοματολογίας και στατιστικής ορολογίας, ώστε να αποφεύγονται πιθανές παρανοήσεις ακόμη και σε λεπτομερειακά θέματα. Η Eurostat λειτουργεί και ως ο γενικός συντονιστής των στατιστικών υπηρεσιών των χωρών της Ε.Ε. Αυτό βοηθά στην ακόμη μεγαλύτερη προαγωγή της Στατιστικής και στην επένδυση της μεγάλης σημασίας που έχει για την επιβίωση των επιχειρήσεων σήμερα, γεγονός ιδιαίτερα σημαντικό και για τις ελληνικές επιχειρήσεις, που χρειάζονται πληροφορίες για να στηρίξουν την ανάπτυξή τους.

4.1.6 Η χρησιμοποίηση από τις επιχειρήσεις στατιστικών πινάκων και διαγραμμάτων

Σήμερα γίνεται ευρύτατη χρήση των πινάκων και των διαγραμμάτων από τις ιδιωτικές και δημόσιες επιχειρήσεις από τους οργανισμούς, με κυριότερο στόχο τη δημιουργία μιας θετικής εικόνας για την επιχείρηση. Οι επιχειρήσεις που λειτουργούν με επιτυχία προβάλλουν μέσω των πινάκων και των διαγραμμάτων κυρίως: α) Τη θετική οικονομική τους κατάσταση, β) τον ικανοποιητικό βαθμό επίτευξης των στόχων που είχαν θέσει στο παρελθόν και γ) τις ευνοϊκές προϋποθέσεις για μια επιτυχή και ασφαλή πορεία της επιχείρησης.

Η χρήση πινάκων και διαγραμμάτων από τις επιχειρήσεις γίνεται και για εσωτερικούς, ερευνητικούς και επιστημονικούς λόγους που αφορούν την ίδια την επιχείρηση. Με τους πίνακες και τα διαγράμματα γίνεται πιο εύκολη και γρήγορη η εσωτερική ενημέρωση των στελεχών, των εργαζομένων της επιχείρησης και των πολύ στενών συνεργατών της, ιδιαίτερα στις τακτικές και έκτακτες ενημερωτικές συναντήσεις που γίνονται.

Ειδικότερα για τα διαγράμματα, θα πρέπει να γνωρίζουμε ότι λόγω των τεράστιων δυνατοτήτων των ηλεκτρονικών υπολογιστών μπορούν να κατασκευαστούν εξαιρετικής ακρίβειας διαγράμματα με συνθέσεις χρωμάτων και πρωτοποριακά σχήματα που ελκύουν την προσοχή του κοινού.

4.1.7 Η σπουδαιότητα των μέτρων θέσης στην επιχειρηματική δραστηριότητα

Τα μέτρα θέσης που μελετήσαμε, και ιδιαίτερα η μέση τιμή και η διάμεσος, είναι από τα πιο συχνά χρησιμοποιούμενα στατιστικά μέτρα και αποτελούν πολύτιμο πληροφοριακό υλικό για την άσκηση οποιασδήποτε επιχειρηματικής πολιτικής. Ακόμη και από την κρατική πλευρά υπάρχει μεγάλο ενδιαφέρον για αυτά, αφού η εφαρμογή τους βοηθάει πολύ στην άσκηση ορθής κοινωνικής, οικονομικής κ.τ.λ. πολιτικής.

Παρόλο όμως που τα μέτρα θέσης είναι πολύ σημαντικά και απαραίτητα, δεν μας δίνουν πλήρη εικόνα του φαινομένου που εξετάζουμε.

Έτσι τις περισσότερες φορές δεν προχωρούμε σε εξαγωγή συμπερασμάτων και σε λήψη επιχειρηματικών αποφάσεων, αν δεν λάβουμε υπόψη μας και τις πρόσθετες πληροφορίες που παρέχουν τα μέτρα διασποράς.

4.1.8 Η σπουδαιότητα των μέτρων διασποράς για τις επιχειρήσεις

Είναι πλέον προφανές, μετά τη μελέτη που κάναμε πάνω στα μέτρα θέσης και στα μέτρα διασποράς, ότι, αν θέλουμε να έχουμε μία πληρέστερη και ασφαλέστερη εικόνα σχετικά με ένα πρόβλημα ή με ένα θέμα που εξετάζουμε, επιβάλλεται η χρησιμοποίηση των στατιστικών μέτρων θέσης μαζί με αυτά της διασποράς.

Για παράδειγμα, αν ένας οικονομικός ερευνητής μετά τη μελέτη των οικονομικών δεδομένων που έκανε για λογαριασμό ενός οργανισμού ή μιας επιχείρησης, διαπιστώσει την ύπαρξη υψηλού βαθμού διακύμανσης, αυτό θα πρέπει να τον προβληματίσει και να τον οδηγήσει σε περαιτέρω έρευνα και μελέτη των δεδομένων του. Τα μέτρα διασποράς, ως στατιστικά μέτρα που μας πληροφορούν πόσο διάσπαρτες (απλωμένες) γύρω από τα μέτρα θέσης είναι οι παρατηρήσεις, αποτελούν ένα επιπλέον εργαλείο στα χέρια των αναλυτών για να αντιληφθούν το βαθμό αντιπροσωπευτικότητας και αξιοπιστίας των μέτρων θέσης.

Στις επιχειρήσεις και στον οικονομικό χώρο γίνεται ευρύτατη χρησιμοποίηση των μέτρων διασποράς. Για παράδειγμα, ακούμε στην αγορά να γίνεται λόγος για διακύμανση

των ποσοτήτων που πωλούνται από τις επιχειρήσεις. Επίσης, πολύ συχνά ακούμε για το ύψος των διακυμάνσεων που παρουσιάζεται στις τιμές των μετοχών των εταιρειών που είναι στο Χρηματιστήριο και για τη διακύμανση στις συναλλαγματικές ισοτιμίες του Ευρώ σε σχέση με άλλα νομίσματα.

4.1.9 Η σημασία της παλινδρόμησης και της συσχέτισης στη σύγχρονη επιχείρηση

Η παλινδρόμηση και η συσχέτιση είναι δύο πολύ χρήσιμες μέθοδοι για τη σύγχρονη επιχείρηση, που συμβάλλουν σημαντικά στην επιχειρηματική πρόβλεψη, στον προγραμματισμό της επιχειρηματικής δράσης και στη λήψη ορθών επιχειρηματικών αποφάσεων. Χρησιμοποιούνται για τη ποσοτική διερεύνηση των σχέσεων, οι οποίες υπάρχουν μεταξύ των διαφόρων οικονομικών μεγεθών και βοηθούν σημαντικά στη χάραξη της οικονομικής πολιτικής που ασκείται είτε από το κράτος είτε από ιδιωτικές επιχειρήσεις.

Η γραμμική παλινδρόμηση, όπως είδαμε, μελετά και προσδιορίζει μια γραμμική σχέση – εφόσον υπάρχει – μεταξύ των μεταβλητών των διμεταβλητών πληθυσμών. Με βάση τις τιμές της ανεξάρτητης μεταβλητής και τη σχέση $\hat{y} = \hat{a} + \hat{\beta} x$ εκτιμούνται - προβλέπονται με προσέγγιση οι αντίστοιχες τιμές της εξαρτημένης μεταβλητής. Η συσχέτιση είναι ένα μέτρο που φανερώνει, αν η σχέση των μεταβλητών που εξετάζεται είναι έντονη, μέτρια, ασθενής ή και μηδενική, καθώς και αν πρόκειται για μια θετική ή αρνητική συσχέτιση.

Στην πράξη, συνήθως προσδιορίζεται η εξίσωση της ευθείας παλινδρόμησης και υπολογίζεται η συσχέτιση των δύο μεταβλητών, γιατί έτσι η εικόνα που προκύπτει από την από κοινού εξέταση των μεταβλητών αυτών είναι πιο ολοκληρωμένη.

Η παλινδρόμηση και η συσχέτιση είναι, ίσως, οι μέθοδοι εκτίμησης που εφαρμόζονται περισσότερο στις οικονομικές σχέσεις. Αυτό φαίνεται σε διάφορους κλάδους της οικονομίας, όπως στην οικονομετρία, στις τεχνικές έρευνας αγοράς κλπ. Καταλαβαίνουμε λοιπόν ότι αποτελούν ένα πολύ σημαντικό όργανο εργασίας και έρευνας για το σημερινό οικονομολόγο, όπως και για άλλους ερευνητές και επιστήμονες.

4.2 Εφαρμογές της στατιστικής στις επιχειρήσεις

Εφαρμογή 4.1

Ένας ερευνητής θέλει να διερευνήσει τη σχέση μεταξύ των επενδύσεων και των επιτοκίων στην ελληνική οικονομία. Τα δεδομένα που έχει στη διάθεσή του είναι ετήσια, αφορούν το διάστημα 1997 – 2008 και δίνονται στον παρακάτω Πίνακα 4.1.

Πίνακας 4.1
Επενδύσεις και μακροχρόνιο επιτόκιο της ελληνικής οικονομίας

Έτος	Επιτόκιο (%) (X)	Επενδύσεις (δις ευρώ) (Y)
1997	9,763	22,216
1998	8,482	24,562
1999	6,308	27,272
2000	6,108	29,450
2001	5,304	30,874
2002	5,123	33,799
2003	4,268	38,257
2004	4,256	38,987
2005	3,585	38,797
2006	4,070	42,349
2007	4,500	44,439
2008	4,802	39,312

Η μεταβλητή Y εκφράζει τις συνολικές επενδύσεις σε δις ευρώ και η μεταβλητή X το μακροχρόνιο επιτόκιο της ελληνικής οικονομίας.

α) Να κατασκευαστεί το διάγραμμα διασποράς των 12 διαθέσιμων ζευγών τιμών (x_i, y_i) και με βάση αυτό να εξεταστεί αν το επιτόκιο σχετίζεται με τις επενδύσεις.

β) Θεωρώντας ότι η σχέση μεταξύ των μεταβλητών X και Y είναι γραμμική, να εκτιμηθούν οι συντελεστές α και β του υποδείγματος $y = \alpha + \beta x + \varepsilon$ με τη μέθοδο ελαχίστων τετραγώνων και να ερμηνευτούν.

γ) Εάν η Ευρωπαϊκή Κεντρική Τράπεζα μείωνε τα επιτόκια κατά 5% το 2009 (σε σχέση με το 2008), για την τόνωση της οικονομίας, να εκτιμηθεί το ύψος των επενδύσεων το 2009.

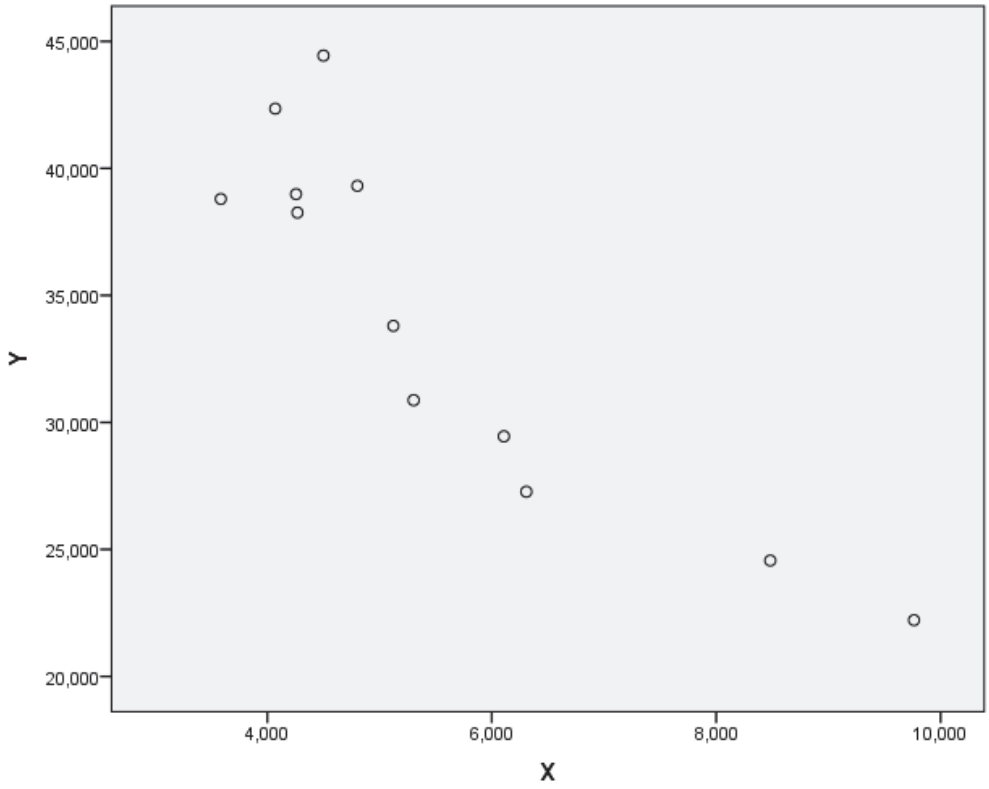
δ) Για τις μεταβλητές X και Y να υπολογιστεί ο συντελεστής συσχέτισης και να ερμηνευτεί.

ε) Για το γραμμικό υπόδειγμα $y = \alpha + \beta x + \varepsilon$, να υπολογιστεί και να ερμηνευτεί ο συντελεστής προσδιορισμού R^2.

Απάντηση

α) Το διάγραμμα διασποράς των ζευγών τιμών (x_i, y_i) παρουσιάζεται στην παρακάτω Εικόνα 4.1.

Εικόνα 4.1
Διάγραμμα διασποράς των δεδομένων του Πίνακα 4.1



Από το διάγραμμα διασποράς προκύπτει ότι υπάρχει μια αρνητική σχέση μεταξύ των δύο μεταβλητών X και Y. Επίσης φαίνεται ότι η σχέση των δύο μεταβλητών είναι γραμμική.

β) Θεωρώντας ότι η σχέση των δύο μεταβλητών είναι γραμμική, όπως φαίνεται στο διάγραμμα, μπορούμε να προχωρήσουμε στον υπολογισμό των συντελεστών α και β του υποδείγματος $y = \alpha + \beta x + \varepsilon$ με τη μέθοδο ελαχίστων τετραγώνων.

Με τη βοήθεια του παρακάτω πίνακα θα βρούμε την εξίσωση της ευθείας ελαχίστων τετραγώνων που προσαρμόζεται στα 12 σημεία των τιμών του επιτοκίου και των επενδύσεων.

Επιτόκιο (%) (x_i)	Επενδύσεις (δις ευρώ) (y_i)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
9,763	22,216	4,216	-11,977	-50,489	17,771
8,482	24,562	2,935	-9,631	-28,262	8,612
6,308	27,272	0,761	-6,921	-5,264	0,578
6,108	29,450	0,561	-4,743	-2,659	0,314
5,304	30,874	-0,243	-3,319	0,808	0,059
5,123	33,799	-0,424	-0,394	0,167	0,180
4,268	38,257	-1,279	4,064	-5,200	1,637
4,256	38,987	-1,291	4,794	-6,191	1,668
3,585	38,797	-1,962	4,604	-9,035	3,851
4,070	42,349	-1,477	8,156	-12,050	2,183
4,500	44,439	-1,047	10,246	-10,732	1,097
4,802	39,312	-0,745	5,119	-3,816	0,556
$\sum_{i=1}^{12} x_i =$ 66,569	$\sum_{i=1}^{12} y_i =$ 410,314	0	0	$\sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y}) =$ -132,724	$\sum_{i=1}^{12}(x_i - \bar{x})^2 =$ 38,506

Έχουμε

$$\bar{x} = \frac{1}{12}\sum_{i=1}^{12} x_i = \frac{66,569}{12} = 5,547 \qquad \bar{y} = \frac{1}{12}\sum_{i=1}^{12} y_i = \frac{410,314}{12} = 34,193$$

$$\sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y}) = -132,724 \qquad \sum_{i=1}^{12}(x_i - \bar{x})^2 = 38,506$$

$$\hat{\beta} = \frac{\sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{12}(x_i - \bar{x})^2} = \frac{-132,724}{38,506} = -3,447$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} = 34,193 - (-3,447)\cdot 5,547 = 53,314$$

Επομένως η ευθεία ελαχίστων τετραγώνων που προσαρμόζεται στα 12 σημεία των τιμών του επιτοκίου και των επενδύσεων έχει εξίσωση

$$\hat{y} = 53,314 - 3,447x \qquad (1)$$

Η τιμή του συντελεστή $\hat{\alpha}$ είναι ίση με 53,314 και εκφράζει την αναμενόμενη τιμή της Y όταν η X είναι μηδέν, δηλαδή μπορούμε να πούμε ότι όταν το επιτόκιο είναι μηδέν, τότε οι αναμενόμενες επενδύσεις θα είναι 53,314 δις ευρώ.

Η κλίση $\hat{\beta}$ της ευθείας είναι ίση με –3,477 και εκφράζει την επίδραση στην αναμενόμενη τιμή της Y που προκαλεί η μεταβολή της X κατά μία μονάδα. Επομένως, αν μειωθεί το

επιτόκιο κατά μια μονάδα (μία ποσοστιαία μονάδα), τότε οι επενδύσεις θα αυξηθούν κατά 3,477 δις ευρώ.

γ) Το 2008, το επιτόκιο ήταν 4,802 και δίνεται ότι 2009 το επιτόκιο θα μειωθεί κατά 5%. Συνεπώς το 2009, το επιτόκιο θα έχει τιμή:

x_{2009} = 0,04802 – 0,05·0,04802 = 0,04802·(1–0,05) = 0,04802·0,95 = 0,04562 = 4,562%

Αντικαθιστώντας στην εξίσωση (1) προκύπτει ότι οι αναμενόμενες επενδύσεις για το έτος 2009 θα είναι y_{2009} = 53,314 – 3,447·4,562 = 37,589 δις ευρώ.

δ) Ο συντελεστής συσχέτισης δίνεται από τη σχέση

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}$$

Από τα δεδομένα του πίνακα που κατασκευάσαμε έχουμε:

$$\sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y}) = -132{,}724 \qquad \sum_{i=1}^{12}(x_i - \bar{x})^2 = 38{,}506$$

Βρίσκουμε επίσης ότι $\sum_{i=1}^{12}(y_i - \bar{y})^2 = 576{,}172$.

Με απλή αντικατάσταση στην παραπάνω σχέση προκύπτει

$$r = \frac{-132{,}724}{\sqrt{38{,}506}\cdot\sqrt{576{,}172}} = -0{,}892$$

Η τιμή αυτή δηλώνει την ισχυρή αρνητική συσχέτιση μεταξύ των επενδύσεων και του επιτοκίου.

ε) Γνωρίζουμε ότι ο συντελεστής προσδιορισμού R^2 ισούται με το τετράγωνο του συντελεστή συσχέτισης, δηλαδή $R^2 = r^2$. Κατά συνέπεια έχουμε:

$R^2 = (-0{,}892)^2 = 0{,}796$

Η τιμή αυτή δηλώνει ότι το 79,6% της μεταβλητότητας των επενδύσεων Y ερμηνεύεται από το επιτόκιο X.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Μέσα από αυτή την πτυχιακή εργασία, γνωρίσαμε μερικώς την πολύ μεγάλη σε μέγεθος και σημασία , επιστήμη της στατιστικής, την εξέλιξή της στο πέρασαμα του χρόνου και τη χρήση της από διάφορους λαούς.Όσον αφορά τη χρήση της από τις επιχειρήσεις διαπιστώνουμε ότι μια επιχείρηση για να έχει θετικά αποτελέσματα, για να ελέγχει την πορεία της, την εξέλιξή της, τη θέση της στην αγόρα εργασίας, την καλύτερη απόδοσή της θα πρέπει να χρησιμοποιεί στατιστικές αναλύσεις και δεδομένα.Τα διάφορα παραδείγματα μας βοηθάνε να καταλάβουμε ότι εξέλιξη και πρόοδος στον επιχειρηματικό τομέα δεν μπορεί να υπάρξει χωρίς τη γνώση και χρήση της στατιστικής .

Βιβλιογραφία

1. Αδαμόπουλος Λ., Δαμιανού Χ. και Σβέρκος Α. (1999). *Μαθηματικά και στοιχεία στατιστικής*. Οργανισμός έκδοσης διδακτικών βιβλίων, Αθήνα

2. Γναρδέλλης Χ. (2003). *Εφαρμοσμένη Στατιστική*. Εκδόσεις Παπαζήση, Αθήνα

3. Δαμιανού Χ. και Κούτρας Μ. (1998). *Εισαγωγή στη Στατιστική, Μέρος Ι*. Εκδόσεις Συμμετρία, Αθήνα

4. Καραγεώργος Δ., Κόκλα Α και Παπακωνσταντίνου Ε. (1999). *Στατιστική Επιχειρήσεων*. Οργανισμός έκδοσης διδακτικών βιβλίων, Αθήνα

5. Κιόχος Π., Κιόχος Α. (2010). *Στατιστική για τις επιχειρήσεις και την οικονομία,* Εκδ. Ελένη Κιόχου

Πνευματικά δικαιώματα