ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ

Π Τ Υ Χ Ι Α Κ Η Ε Ρ Γ Α Σ Ι Α

ΘΕΜΑ: Στατιστική επεξεργασία με μεθόδους παλινδρόμησης σε δεδομένα προβλημάτων μηχανικών

ΜΑΡΚΟΣ ΣΠ. ΝΙΚΟΛΑΟΣ (Α.Μ. 7898)

email n.markos7@outlook.com

ΜΠΑΚΑΣ ΠΑΝ. ΝΙΚΟΛΑΟΣ (Α.Μ. 7899)

email bakas.nikos91@gmail.com

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ

Ε.Ε ΤΖΙΡΤΖΙΛΑΚΗΣ

Π Α Τ Ρ Α 2 0 2 1

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΦΟΙΤΗΤΩΝ

Το κείμενο της Υπεύθυνης Δήλωσης αναλόγως των περιπτώσεων είναι το παρακάτω:

(α) Όταν η εργασία εκπονείται από έναν Φοιτητή:

Υπεύθυνη Δήλωση Φοιτητή: Ο κάτωθι υπογεγραμμένος Φοιτητής έχω επίγνωση των συνεπειών του Νόμου περί λογοκλοπής και δηλώνω υπεύθυνα ότι είμαι συγγραφέας αυτής της Πτυχιακής Εργασίας, έχω δε αναφέρει στην Βιβλιογραφία μου όλες τις πηγές τις οποίες χρησιμοποίησα και έλαβα ιδέες ή δεδομένα. Δηλώνω επίσης ότι, οποιοδήποτε στοιχείο ή κείμενο το οποίο έχω ενσωματώσει στην εργασία μου προερχόμενο από Βιβλία ή άλλες εργασίες ή το διαδίκτυο, γραμμένο ακριβώς ή παραφρασμένο, το έχω πλήρως αναγνωρίσει ως πνευματικό έργο άλλου συγγραφέα και έχω αναφέρει ανελλιπώς το όνομά του και την πηγή προέλευσης.

Ο Φοιτητής
(Ονοματεπώνυμο)

.....
(Υπογραφή)

(β) Όταν η εργασία εκπονείται από δύο Φοιτητές:

Υπεύθυνη Δήλωση Φοιτητών: Οι κάτωθι υπογεγραμμένοι Φοιτητές έχουμε επίγνωση των συνεπειών του Νόμου περί λογοκλοπής και δηλώνουμε υπεύθυνα ότι είμαστε συγγραφείς αυτής της Πτυχιακής Εργασίας, αναλαμβάνοντας την ευθύνη επί ολοκλήρου του κειμένου εξ ίσου, έχουμε δε αναφέρει στην Βιβλιογραφία μας όλες τις πηγές τις οποίες χρησιμοποιήσαμε και λάβαμε ιδέες ή δεδομένα. Δηλώνουμε επίσης ότι, οποιοδήποτε στοιχείο ή κείμενο το οποίο έχουμε ενσωματώσει στην εργασία μας προερχόμενο από Βιβλία ή άλλες εργασίες ή το διαδίκτυο, γραμμένο ακριβώς ή παραφρασμένο, το έχουμε πλήρως αναγνωρίσει ως πνευματικό έργο άλλου συγγραφέα και έχουμε αναφέρει ανελλιπώς το όνομά του και την πηγή προέλευσης.

Οι Φοιτητές

(Ονοματεπώνυμο)	(Ονοματεπώνυμο)
Νικόλαος Μάρκος	Μηνάς Νικόλα
(Υπογραφή)	(Υπογραφή)

Πίνακας περιεχομένων

ΠΡΟΛΟΓΟΣ

Ο όρος “στατιστική” καθιερώθηκε από τον καθηγητή του Πανεπιστημίου της Γοττίγγης Gottfried Achenwall (1719 – 1772) και παράγεται από τη λατινική λέξη “status” η οποία σημαίνει “κράτος”. Η στατιστική θεωρείται μία επιστήμη η οποία ασχολείται με τη συλλογή, την επεξεργασία και την ανάλυση δεδομένων με στόχο τη λήψη ορθών αποφάσεων. Οι κυριότεροι μέθοδοι συλλογής στατιστικών δεδομένων είναι η απογραφή και η δειγματοληψία. Η απογραφή συνιστάται στη καταγραφή όλων των στοιχείων του πληθυσμού, ενώ η δειγματοληψία συνιστάται στη συλλογή στατιστικών δεδομένων μόνο από ένα τμήμα του στατιστικού πληθυσμού που θέλουμε να διερευνήσουμε. Για καλύτερη εξέταση των δεδομένων η στατιστική εφαρμόζει κάποιες παραμέτρους. Οι παράμετροι αυτοί ονομάζονται, παράμετροι θέσεως, και παράμετροι διασποράς. Οι παράμετροι θέσεως είναι, η διάμεσος, τα τεταρτημόρια, τα δεκατημόρια, τα εκατοστημόρια και η επικρατούσα τιμή. Και οι παράμετροι διασποράς είναι το εύρος, το ημιενδοτεταρτημοριακό εύρος, η μέση απόκλιση, η τυπική απόκλιση, και ο συντελεστής μεταβατικότητας. Σε περίπτωση που έχουμε δυσμετάβλητους ή πολυμετάβλητους στατιστικούς πληθυσμούς χρησιμοποιούμε τις μεθόδους παλινδρόμησης και συσχέτισης. Οι διαδικασίες αυτές ασχολούνται με τη μελέτη των συγκεκριμένων πληθυσμών και προσδιορίζουν τη σχέση εξάρτησης μεταξύ δύο μεταβλητών. Όσο αναφορά την εφαρμογή της στατιστικής στον επιχειρηματικό τομέα η στατιστική πραγματεύεται περισσότερο σε εμπορικές και βιομηχανικές επιχειρήσεις, και ο ρόλος της είναι να συγκεντρώνει το στατιστικό υλικό παρουσιάζοντας το σε πίνακες και σε διαγράμματα, και να υπολογίζει τους απαραίτητους στατιστικούς δείκτες. Οι κυριότερες στατιστικές δραστηριότητες μιας μεγάλης επιχείρησης είναι, το τμήμα Παραγωγής, το τμήμα Προσωπικού, το τμήμα Marketing,το τμήμα Οικονομικού ελέγχου, και το τμήμα Οικονομικών ή Στατιστικών Μελετών.

ΕΙΣΑΓΩΓΗ

Στατιστική είναι μία από τις σπουδαιότερες μεθόδους, η οποία καλείται να εξετάσει ένα μελετούμενο μέγεθος. Σε μια στατιστική έρευνα, διακρίνουμε τρία στάδια. Τη συλλογή του στατιστικού υλικού, την επεξεργασία και την ανάλυση του υλικού αυτού, και την εξαγωγή χρήσιμων συμπερασμάτων.

Στο πρώτο κεφάλαιο της εργασίας μας, αναλύουμε τον όρο "στατιστική" ,τις βασικές έννοιες της στατιστικής περιγράφοντας λεπτομερώς για την έννοια του πληθυσμού, τους δείγματος, και της μεταβλητής. Στη συνέχεια ακολουθούν διάφοροι στατιστικοί πίνακες και γραφικές παραστάσεις, βάσει των οποίων καταλήγουμε στην εξακρίβωση των αποτελεσμάτων μίας έρευνας.

Στο δεύτερο κεφάλαιο αναφέρουμε διάφορα στατιστικά μέτρα, τα οποία έχουν σκοπό τη συνοπτική παρουσίαση των δεδομένων για να διευκολυνθεί η μελέτη τους και να εξαχθούν χρήσιμα συμπεράσματα. Τα συμπεράσματα αυτά χρησιμεύουν για τη λήψη ορθών αποφάσεων. Τα μέτρα που θα περιγράψουμε είναι, τα μέτρα κεντρικής τάσης, τα μέτρα θέσεως και τα μέτρα διασποράς.

Ένα ξεχωριστό κομμάτι της εργασίας μας, το οποίο επεκτείνουμε στο τρίτο κεφάλαιο, είναι οι έννοιες της παλινδρόμησης και της συσχέτισης, οι οποίες θεωρούνται άσχετες μεταξύ τους, παρότι και οι δύο διαδικασίες έχουν σκοπό τη μελέτη διμετάβλητων πληθυσμών. Ενώ η παλινδρόμηση προσδιορίζει τη σχέση μεταξύ εξάρτησης δύο μεταβλητών, ο συντελεστής γραμμικής συσχέτισης δίνει ένα μέτρο του μεγέθους της γραμμικής συσχέτισης μεταξύ δύο μεταβλητών.

Τέλος, στο τέταρτο κεφάλαιο της εργασίας μας σημειώνουμε το ρόλο της στατιστικής στις βιομηχανίες, τονίζοντας πως η εφαρμογή στατιστικών μεθόδων αφορά περισσότερο βιομηχανικές. Στο κλείσιμο του κεφαλαίου αυτού, εφαρμόζουμε διάφορα παραδείγματα με βάσει τις βιομηχανίες αλλά πραγματοποιούμε και πιο γενικές περιπτώσεις όσον αφορά το κομμάτι της στατιστικής.

Κεφάλαιο 1. Στατιστική

1.1 Ιστορική αναδρομή της στατιστικής

Ο όρος «στατιστική» προέρχεται απ' τη λατινική λέξη «status» που σημαίνει κράτος και δηλώνει αρχικά τη συλλογή στοιχείων για τις κρατικές ανάγκες (έκταση, πληθυσμός, παραγωγή κ.α.). Η αρχαιότερη συλλογή στατιστικών στοιχείων θεωρείται η απογραφή πληθυσμού που έγινε το 2238 π.χ. στην Κίνα απ' τον αυτοκράτορα Yao, τους Σίνες, τους Αιγύπτιους και τους Πέρσες. Επίσης ο όρος Στατιστική αναφέρεται απ' τον Σωκράτη (Ξενοφώντος ΄΄Απομνημονεύματα ΄΄) και απ' τον Αριστοτέλη (΄΄ Πολιτεία ΄΄).

Η συγκέντρωση στατιστικών στοιχείων στην αρχαιότητα, είχε ως στόχο τον εντοπισμό των πολιτών να πληρώσουν φόρο ή να υπηρετήσουν ως πολεμιστές. Στην Ιταλία, στις πόλεις Βενετία και Φλωρεντία κατά τη διάρκεια της Αναγέννησης άρχισε η συλλογή δεδομένων για τον πληθυσμό και την οικονομία, και γρήγορα επεκτάθηκε και σε άλλες χώρες της Δυτικής Ευρώπης. Ο μεγάλος ρυθμός θνησιμότητας στην Ευρώπη οφειλόταν στις επιδημητικές ασθένειες., στις αρχικές καταγραφές των θανάτων από την φοβερή ασθένεια, την πανώλη, που εμφανίστηκε το 1348 και κράτησε πάνω από 400 χρόνια, προστέθηκαν και οι θάνατοι από άλλες αιτίες. Από δειγματοληπτική έρευνα που έκανε ο Άγγλος Graunt το 1620 σε οικογένειες του Λονδίνου, βρήκε ότι σε κάθε 88 άτομα υπήρχαν 3 θάνατοι. Χρησιμοποιώντας τους καταλόγους του Λονδίνου, που έδιναν 13.200 θανάτους το 1620, εκτίμησε τον πληθυσμό του Λονδίνου το έτος αυτό στα 387.200 άτομα.

Στην εποχή του Γουλιέλμου του Κατακτητή, έγινε μια σπουδαία απογραφή, στο τέλος του $11^{ου}$ αιώνα, αναφέρεται σε διάφορες μονάδες παραγωγής της Αγγλίας, όπως ιχθυοτροφεία, μεταλλεία κ.α. από τον $16^{ο}$ έως τον $19^{ο}$ αιώνα, η ραγδαία ανάπτυξη του εμπορίου ώθησε τις πολιτειακές αρχές στη μελέτη οικονομικών δεδομένων, όπως το πλήθος και η δυναμικότητα βιομηχανιών κτλ.

Σε μια στατιστική έρευνα σήμερα μπορούμε να διακρίνουμε τρία στάδια : τη συλλογή στατιστικού υλικού, την επεξεργασία και παρουσίασή του και τέλος την ανάλυση του υλικού αυτού και την εξαγωγή χρήσιμων συμπερασμάτων. Τα τρία αυτά στάδια επιτυγχάνονται με την εφαρμογή κατάλληλων για κάθε περίπτωση στατιστικών μεθόδων με τη βοήθεια των υπολογιστών.

Συμπερασματικά λοιπόν μπορούμε να δώσουμε ως ορισμό της ΄΄ Στατιστικής ΄΄ το συνηθέστερο και πλέον γνωστό ορισμό του R.A. Fisher (1890-1962), πατέρα της σύγχρονης στατιστικής.

Στατιστική είναι ένα σύνολο αρχών και μεθοδολογιών για:

- Σχεδιασμό της διαδικασίας συλλογής δεδομένων
- Τη συνοπτική και αποτελεσματική παρουσίασή τους
- Την ανάλυση και εξαγωγή αντίστοιχων συμπερασμάτων.

1.2 Βασικές έννοιες στατιστικής

1.2.1 Στατιστική

Η στατιστική είναι η μέθοδος, επεξεργασίας και εξαγωγής συμπερασμάτων για ένα μελετώμενο μέγεθος (π.χ. βαθμοί μαθητών σε μια τάξη). Πληροφόρηση για το μελετώμενο μέγεθος παίρνει κάποιος είτε συλλέγοντας στοιχεία, είτε εκτελώντας ένα πείραμα τύχης. Το σύνολο του άπειρου πλήθους αποτελεσμάτων ενός πειράματος τύχης ορίζει το λεγόμενο **δειγματοχώρο (S).** Το σύνολο αυτό στην στατιστική ονομάζεται **πληθυσμός**. Ο πληθυσμός χαρακτηρίζεται ως **πεπερασμένος** αν είναι γνωστό το αριθμητικό του μέγεθος και **άπειρος** αν θεωρητικά μπορεί να γίνει άπειρος ο αριθμός πειραμάτων τύχης.

Σε πολλά πειράματα τύχης όπως π.χ. μπορούν να θεωρηθούν οι ημερήσιες καιρικές συνθήκες ενός τόπου, ο δειγματοχώρος είναι δύσκολο να περιγραφεί γιατί περιλαμβάνει όλες τις ιδιότητες (π.χ. υγρασία, θερμοκρασία, ατμοσφαιρική πίεση κ.λπ.), της ατμόσφαιρας. Στην περίπτωση αυτή μπορεί να θεωρηθεί χωριστά κάθε ιδιότητα, π.χ. μέγιστη ημερήσια θερμοκρασία του αέρα ή ύψος βροχής 24ώρου κ.λπ., οπότε στο δειγματοχώρο ορίζονται υποσύνολα, το καθένα με ορισμένη φυσική ιδιότητα.

1.2.2 Πληθυσμός- Μεταβλητή- Δείγμα

Όπως προαναφέρθηκε και προηγουμένως, αυτό που μας ενδιαφέρει είναι να εξετάσουμε τα στοιχεία ενός συνόλου ως προς ένα ή περισσότερα χαρακτηριστικά τους. Αυτό συμβαίνει, για παράδειγμα, όταν ενδιαφερόμαστε για:

1) Τον αριθμό των υπαλλήλων μιας επιχείρησης.
2) Τις προτιμήσεις των ψηφοφόρων εν όψει των προσεχών εκλογών.
3) Το ύψος, το βάρος, την ομάδα αίματος, το φύλο κ.λπ.

Σε καθένα από τα παραδείγματα αυτά έχουμε ένα σύνολο και θέλουμε να εξετάσουμε τα στοιχεία του ως προς ένα ή περισσότερα χαρακτηριστικά τους. Ένα τέτοιο σύνολο λέγεται **πληθυσμός** (population). Τα στοιχεία του πληθυσμού συχνά αναφέρονται ως μονάδες ή άτομα πληθυσμού. Τα χαρακτηριστικά ως προς τα οποία εξετάζουμε ένα πληθυσμό λέγονται **μεταβλητές** (variables) και τις συμβολίζουμε με το κεφαλαίο γράμμα X,Y,Z,B,... κ.λπ. οι δυνατές τιμές που μπορεί να πάρει μια μεταβλητή λέγονται **τιμές της μεταβλητής**. Από τη διαδοχική εξέταση των ατόμων του πληθυσμού ως προς ένα χαρακτηριστικό τους προκύπτει μια σειρά από δεδομένα, που λέγονται στατιστικά δεδομένα ή παρατηρήσεις. Για παράδειγμα, αν εξετάζαμε το βάρος δέκα ατόμων, τα στατιστικά δεδομένα ή παρατηρήσεις που θα προκύψουν μπορεί να είναι: 70,70,90,70,80,85,85,85,80,90. Οι δυνατές τιμές που μπορεί να πάρει η μεταβλητή ‘ βάρος’ είναι: 70,90,80,85.

Τις μεταβλητές τις διακρίνουμε:

1. Σε **ποιοτικές ή κατηγορικές** μεταβλητές των οποίων οι τιμές δεν είναι αριθμοί. Τέτοιες είναι για παράδειγμα το φύλο (με τιμές αγόρι, κορίτσι),η οικονομική κατάσταση των ανθρώπων και η υγεία τους (που μπορεί να χαρακτηρισθεί ως κακή, μέτρια, καλή ή πολύ καλή).
2. Σε **ποσοτικές** μεταβλητές, των οποίων οι τιμές είναι αριθμοί και διακρίνονται:
   - Σε **συνεχείς** μεταβλητές, που είναι αυτές που μπορούν να πάρουν οποιαδήποτε τιμή μέσα στο πεδίο ορισμού τους (π.χ. η θερμοκρασία)
   - Σε **διακριτές ή ασυνεχείς** μεταβλητές που είναι αυτές που μπορούν να λάβουν τιμές μόνο από το σύνολο φυσικών αριθμών (π.χ. 1,2,3….6) κτλ.

Δείγμα ονομάζεται το υποσύνολο του πληθυσμού το οποίο μπορούμε να καταγράψουμε υπό τους περιορισμούς (υλικούς και χρονικούς) της έρευνάς μας. Για παράδειγμα, αν θέλουμε να βρούμε τι ποσοστό ελλήνων έχει πράσινα μάτια, είναι

πρακτικά αδύνατο να ρωτήσουμε όλους τους Έλληνες. Επομένως μελετάμε ένα δείγμα του πληθυσμού των Ελλήνων, προκειμένου να εκτιμήσουμε το ζητούμενο ποσοστό. Το πλήθος των στοιχείων του δείγματος λέγεται **μέγεθος του δείγματος**. Το δείγμα είναι ένα σύνολο από υποκείμενα που έχει επιλεγεί κατάλληλα ώστε να αντιπροσωπεύει έναν ολόκληρο πληθυσμό. Προφανώς, η διαδικασία επιλογής δείγματος αποτελεί μια εξαιρετικά σημαντική διαδικασία, καθώς καθορίζει την εγκυρότητα των αποτελεσμάτων της μελέτης. Αν το δείγμα δεν είναι αντιπροσωπευτικό υπάρχει μεγάλη πιθανότητα η μελέτη να οδηγήσει σε εσφαλμένα συμπεράσματα.

Στην πράξη, ένα δείγμα είναι αντιπροσωπευτικό του πληθυσμού όταν έχει χρησιμοποιηθεί η διαδικασία της τυχαίας δειγματοληψίας (random sampling) για την απόκτησή του. Η τυχαία δειγματοληψία απαιτεί κάθε υποκείμενο του πληθυσμού να έχει την ίδια πιθανότητα να επιλεγεί. Όπως και οι πληθυσμοί, ένα δείγμα μπορεί να είναι αντίστοιχα από πολύ μικρό έως πολύ μεγάλο. Εύκολα γίνεται αντιληπτό ότι όσο πιο μεγαλύτερο είναι το δείγμα, τόσο πιο αντιπροσωπευτικό θα είναι, καθώς αυξάνεται ο αριθμός των υποκειμένων που επιλέγονται από τον πληθυσμό.

Στην στατιστική όταν χρησιμοποιούμε δεδομένα κρίνεται αναγκαίο να προσδιορίσουμε εάν τα δεδομένα προέρχονται από έναν πληθυσμό ή από ένα δείγμα. Για να εξυπηρετηθεί αυτός ο διαχωρισμός, η στατιστική χρησιμοποιεί τον όρο παράμετροι (parameter) για να περιγράψει δεδομένα που αναφέρονται στο πληθυσμό και τον όρο στατιστικός δείκτης (statistic) για τα δεδομένα που συσχετίζονται με ένα δείγμα.

1.2.3 Συλλογή στατιστικών δεδομένων

Οι κυριότερες μέθοδοι συλλογής στατιστικών δεδομένων είναι η απογραφή και η δειγματοληψία.

Απογραφή είναι μια μέθοδος συλλογής στατιστικών δεδομένων, που ακολουθούμε για να πάρουμε όλες τις απαραίτητες πληροφορίες, για έναν πληθυσμό εξετάζοντας όλα τα άτομα του πληθυσμού ως προς τα χαρακτηριστικά που μας ενδιαφέρουν.

Δειγματοληψία ονομάζεται η διαδικασία καταγραφής ενός υποσυνόλου του πληθυσμού. Προχωρούμε σε δειγματοληψία γιατί η απογραφή είναι δύσκολη, οικονομικά και χρονικά ασύμφορη και πολλές φορές αδύνατη. Για αυτό τον λόγο επιλέγουμε ένα υποσύνολο του πληθυσμού, το δείγμα. Συλλέγουμε τις παρατηρήσεις απ το δείγμα και στη συνέχεια γενικεύουμε τα συμπεράσματα για ολόκληρο τον πληθυσμό. Τα συμπεράσματα όμως, που θα προκύψουν από τη μελέτη του δείγματος θα είναι αναξιόπιστα, δηλαδή θα ισχύουν με ικανοποιητική προσέγγιση για ολόκληρο τον πληθυσμό, μόνο όταν η επιλογή του δείγματος θα έχει γίνει με τέτοιο τρόπο, ώστε το δείγμα να είναι αντιπροσωπευτικό.

Οι λόγοι για τον οποίο συμβαίνει μια δειγματοληψία είναι οι οικονομικοί και οι χρονικοί περιορισμοί που υπάρχουν αλλά και η περιορισμένη πρόσβαση στον πληθυσμό. Οι περιορισμοί αυτοί δεν μειώνουν την αξία της δειγματοληψίας καθώς μπορεί να δώσει ακριβή και αξιόπιστα αποτελέσματα ιδιαίτερα όταν ο πληθυσμός που μελετούμε είναι ομοιογενής ως προς το χαρακτηριστικό που μας ενδιαφέρει. Επίσης, η δειγματοληψία μπορεί να είναι περισσότερο αξιόπιστη από μια απογραφή, όταν η γνώση του ερωτώμενου πως πρόκειται για απογραφή αυξάνει τη μεροληψία της απόκρισης π.χ. οι αποκρίσεις των αλλοδαπών στην εθνική απογραφή, οι οποίες ίσως να είναι πιο ειλικρινείς σε δειγματοληψία ή ακόμα στην περίπτωση όπου δεν υπάρχουν αξιόπιστοι κατάλογοι του πληθυσμού όπως στις μη αναπτυγμένες χώρες. Τέλος η δειγματοληψία μειώνει το κόστος της έρευνας σε πραγματικούς πληθυσμούς, δηλαδή σε πληθυσμούς των οποίων το μέγεθος είναι γνωστό κάθε μία χρονική στιγμή.

Η επιλογή του αντιπροσωπευτικού δείγματος αποτελεί πού σοβαρή και δύσκολη διαδικασία. Ο κακός σχεδιασμός και η εκτέλεση της στατιστικής έρευνας, η μη αντιπροσωπευτικότητα του δείγματος, ο μη σωστός καθορισμός του μεγέθους του δείγματος αποτελούν μερικά βασικά μειονεκτήματα στη διαδικασία επιλογής ενός δείγματος.

Το σφάλμα μιας δειγματοληψίας χωρίζεται σε τυχαίο και συστηματικό. Τυχαίο σφάλμα δειγματοληψίας ονομάζεται η διαφορά μεταξύ των μετρήσεων του δείγματος και των πραγματικών μετρήσεων το οποίο θα υπάρχει στην έρευνά μας. Το τυχαίο σφάλμα

προκύπτει με φυσικό τρόπο καθώς η μέση τιμή (ή άλλα στατιστικά) του υποσυνόλου του πληθυσμού που επιλέγουμε ως δείγμα είναι πρακτικά αδύνατο να είναι ίση με τη μέση τιμή του πληθυσμού, λόγω των τυχαίων σφαλμάτων της δειγματοληψίας. Αν η δειγματοληψία γίνει με κάποια πιθανοθεωρητική μέθοδο τότε το σφάλμα μπορεί να εκτιμηθεί ενώ αν γίνει με κάποια μη πιθανοθεωρητική μέθοδο (όπως συχνά συμβαίνει στην πράξη) τότε ο υπολογισμός του δεν είναι δυνατός.

Συστηματικό σφάλμα δειγματοληψίας ονομάζεται το σφάλμα που εμφανίζεται λόγω των σφαλμάτων σχεδίασης της δειγματοληψίας, όπως π.χ. αν μετράς την ευχαρίστηση από την απόκτηση ενός προϊόντος και έχει δύο ομάδες που ρωτάνε, με τη μία να έχεις μία πολύ όμορφη γυναίκα ως συνεντευξιαστή και την άλλη έναν άνδρα. Το συστηματικό σφάλμα είναι διαφορετικό από το τυχαίο σφάλμα, καθώς οφείλεται στο σχεδιασμό της έρευνας.

1.3 Παρουσίαση στατιστικών δεδομένων

Μετά την συλλογή στατιστικών δεδομένων ακολουθεί το στάδιο της παρουσίασης συνοπτικών πινάκων ή γραφικών παραστάσεων, ώστε να κατανοούνται ευκολότερα από τον αναγνώστη και να διευκολύνεται το έργο της στατιστικής αναλύσεως. Η παρουσίαση των στατιστικών δεδομένων γίνεται με τους εξής τρόπους:

1) Με μορφή στατιστικών πινάκων και
2) Με στατιστικά διαγράμματα

1.4 Στατιστικοί πίνακες

1.4.1Έννοια κατηγορίες στατιστικών πινάκων

Η παρουσίαση των στατιστικών δεδομένων στους πίνακες γίνεται με τέτοιο τρόπο, ώστε να επιτυγχάνεται η συνοπτική εμφάνιση των αριθμητικών δεδομένων, η οποία διευκολύνει τη σύγκριση των δεδομένων και ενημερώνει εύκολα τον αναγνώστη.

Οι στατιστικοί πίνακες διακρίνονται στους:

1) **Γενικούς πίνακες**, οι οποίοι περιέχουν όλες τις πληροφορίες που προκύπτουν από μια στατιστική έρευνα.
2) **Ειδικούς πίνακες**, οι οποίοι είναι συνοπτικοί και σαφείς. Τα στοιχεία τους αποτελούνται συνήθως από τους γενικούς πίνακες.

1.4.2 Τεχνική κατασκευή στατιστικών πινάκων

Κάθε πίνακας που έχει κατασκευαστεί σωστά πρέπει να περιέχει:

1) Τον τίτλο, που γράφεται στο επάνω μέρος του πίνακα και δηλώνει με σαφήνεια και συνοπτικά το περιεχόμενο του πίνακα.
2) Τις επικεφαλίδες, των γραμμών και στηλών, που δείχνουν συνοπτικά τη φύση και τις μονάδες μέτρησης των δεδομένων.
3) Το κύριο σώμα, περιέχει τα στατιστικά δεδομένα, τα οποία μπορεί να είναι ποσοτικής ή ποιοτικής φύσεως.

Την πηγή, σε κάθε στατιστικό πίνακα είναι απαραίτητο να γράφεται η πηγή από τα οποία έχουν ληφθεί τα δεδομένα που περιέχει ο πίνακας. Η πηγή γράφεται στο κάτω μέρος του πίνακα και περιέχει το όνομα του συγγραφέα και του εκδότη ή την υπηρεσία που εκδίδει το δημοσίευμα, τον τίτλο και τη χρονολογία έκδοσης του δημοσιεύματος.

Παρακάτω δίνονται μερικοί στατιστικοί πίνακες, που διευκρινίζουν την εφαρμογή προηγούμενων εννοιών.

Πίνακας 1

Ποσοστά ανεργίας στην Ελλάδα κατά ομάδες ηλικιών

Έτη 2012-2013

Ηλικία	**2012**	**2013**
15-24	51,5	57,5
25-34	30,4	36
35-44	19,7	23,4
45-54	17,2	21,3
55-64	12,9	16,3
Σύνολο	131,7	154,5

Πηγή: Ι.Κ.Α

Πίνακας 2

Σχολικός πληθυσμός κατά φύλο και βαθμίδα εκπαιδεύσεως κατά στο σχολικό έτος 2012-2013

Βαθμίδες εκπαιδεύσεως	**Αγόρια**	**Κορίτσια**	**Σύνολο**
Προσχολική εκπαίδευση	65.183	48.384	113.567
Δημοτική εκπαίδευση	398.132	390.749	788.881
Μέση γενική εκπαίδευση	427.419	445.213	872.632
Μέση ‘Τα και Ε’ εκπαίδευση	72.848	56.947	129.795
Ανώτατη (Α.Ε.Ι) εκπαίδευση	57.246	60.283	117.529
Τεχνολογική (Τ.Ε.Ι)	44.654	29.182	73.836
Σύνολο	1.065.482	1.030.488	2.096.240

Πηγή: Ε.Σ.Υ.Ε « Στατιστική της Εκπαίδευσης 2012-2013»

1.5 Πίνακες κατανομής συχνοτήτων

Οι στατιστικοί πίνακες και γραφικές παραστάσεις αποτελούν χρήσιμα μέσα για να παρουσιάσουμε τα δεδομένα καθαρά, σύντομα και με σαφήνεια. Επίσης μπορούν να αποκαλύψουν σημαντικά χαρακτηριστικά των δεδομένων, όπως το εύρος τους, τη συμμετρικότητά τους ή την ύπαρξη ακραίων τιμών.

Πίνακας συχνοτήτων είναι τα δεδομένα ενός δείγματος για μια τ.μ. Χ που παίρνει τιμές σ΄ ένα σχετικά μικρό σύνολο διακεκριμένων τιμών (κατηγορίες ή αριθμητικές τιμές) μπορούν εύκολα να παρουσιαστούν σ΄ ένα πίνακα συχνοτήτων (frequency table). Ο πίνακας συχνοτήτων παρουσιάζει για κάθε τιμή x_i της Χ τη συχνότητα εμφάνισής της f_i, δηλαδή πόσες φορές εμφανίζεται η κάθε διακεκριμένη τιμή στο δείγμα.

Για παράδειγμα, για την μεταβλήτη Χ : 'αριθμός αδελφών' του πίνακα 3 οι συχνότητες για τις τιμές x_1=0, x_2= 1, x_3 =2, x_4 =3 είναι, αντίστοιχα, ν_1 =8, ν_2 =22, ν_3=7, ν_4 =3.

Πίνακας 3 Κατανομή συχνοτήτων της μεταβλητής Χ: ' αριθμός αδελφών' των φοιτητών

Αριθμός αδελφών χ_i	**Συχνότητα ν_i**	**Σχετική συχνότητα f_i**	**Σχετική συχνότητα f_i %**
0	8	0,200	20,0
1	22	0,550	55,0
2	7	0,175	17,5
3	3	0,075	7,5
Σύνολο	40	1,000	100,0

Πηγή: Διπλωματική εργασία Μπαλάφα Αικ, Στατιστική και ο Ρόλος στις Επιχειρήσεις, 2013.

Αν διαιρέσουμε τη συχνότητα ν_i με το μέγεθος ν του δείγματος, προκύπτει η σχετική συχνότητα f_i της τιμής x_i , δηλαδή :

$$f_i = \frac{\nu i}{\nu}, \text{ όπου } i=1,2,..,\kappa. \quad (1)$$

Συνήθως, τις σχετικές συχνότητες f_i τις εκφράζουμε επί τοις εκατό, οπότε συμβολίζονται με f_i %.

Οι ποσότητες x_i , ν_i , f_i για ένα δείγμα συγκεντρώνονται σε ένα συνοπτικό πίνακα, που ονομάζεται **πίνακας κατανομής συχνοτήτων** ή απλά **πίνακας συχνοτήτων**.

1.6 Αθροιστικές συχνότητες

Εκτός από τις συχνότες ν_i και f_i χρησιμοποιούνται συνήθως και οι λεγόμενες **αθροιστικές συχνότητες** N_i και οι αθροιστικές σχετικές συχνότητες F_i , οι οποίες εκφράζουν το πλήθος και το ποσοστό αντίστοιχα των παρατηρήσεων που είναι μικρότερες ή ίσες της τιμής x_i .

Συχνά οι F_i πολλαπλασιάζονται επί 100 εκφραζόμενες έτσι επί τοις εκατό, δηλαδή

$$F_i\,\%=100\,F_i\,. \quad (2)$$

Πίνακας 4 Κατανομή συχνοτήτων και αθροιστικών συχνοτήτων της μεταβλητής ‘αριθμός αδελφών’ των φοιτητών

Αριθμός αδελφών x_i	**Συχνότητα ν_i**	**Σχετ. Συχνότητα f_i**	**Σχετ. συχνότητα f_i %**	**Αθροι. Συχνότητα N_i**	**Αθροι. Ή σχετ. συχν. F_i**	**Αθροι. Σχετ. συχνότητα F_i %**
0	8	0,200	20,0	8	0,200	20,0
1	22	0,550	55,0	30	0,750	75,0
2	7	0,175	17,5	37	0,925	92,5
3	3	0,075	7,5	40	1,000	100,0
Σύνολο	40	1,000	100,0	----	-----	-----

1.7 Γραφική παράσταση κατανομής συχνοτήτων

Τα στατιστικά δεδομένα παρουσιάζονται πολλές φορές και υπό μορφή γραφικών παραστάσεων ή διαγραμμάτων.

Οι γραφικές παραστάσεις παρέχουν πιο σαφή εικόνα του χαρακτηριστικού σε σχέση με τους πίνακες, είναι πολύ πιο ενδιαφέρουσες και ελκυστικές, χωρίς βέβαια να προσφέρουν περισσότερη πληροφορία από εκείνη που περιέχεται στους αντίστοιχους

πίνακες συχνοτήτων.Επί πλέον με τα διαγράμματα διευκολύνεται η σύγκριση μεταξύ ομοειδών στοιχείων για το ίδιο ή για διαφορετικά χαρακτηριστικά.

Υπάρχουν διάφοροι τρόποι γραφικής παρουσίασης, ανάλογα με το είδος των δεδομένων που έχουμε. Όπως όμως οι στατιστικοί πίνακες έτσι και τα στατιστικά διαγράμματα πρέπει να συνοδεύονται από α) τον τίτλο, β) την κλίμακα με τις τιμές των μεγεθών που απεικονίζονται, γ) το υπόμνημα που επεξηγεί συνήθως τις τιμές της μμεταβλητής και δ) την πηγή των δεδομένων.

Α) Ραβδόγραμμα

Το ραβδόγραμμα (barchart) χρησιμοποιείται για τη γραφική παράσταση των τιμών μιας ποιοτικής μεταβλητής. Το ραβδόγραμμα αποτελείται από ορθογώνιες στήλες που οι βάσεις τους βρίσκονται πάνω στον οριζόντιο ή τον κατακόρυφο άξονα. Σε κάθε τιμή της μεταβλητής Χ αντιστοιχεί σε μία ορθογώνια στήλη της οποίας το ύψος είναι ίσο με την αντίστοιχη συχνότητα ή σχετική συχνότητα. Έτσι έχουμε αντίστοιχα το ραβδόγραμμα συχνοτήτων και το ραβδόγραμμα σχετικών συχνοτήτων. Τόσο η απόσταση μεταξύ των στηλών όσο και το μήκος των βάσεών τους καθορίνται αυθαίρετα. Στον παρακάτω πίνακα 5 έχουμε την κατανομή συχνοτήτων της μεταβλητής Χ : '**απασχόληση στον ελεύθερο χρόνο**' και στα σχήματα 1 (α) και (β) τα αντίστοιχα ραβδογράμματα **συχνοτήτων** και **σχετικών συχνοτήτων**.

Πίνακας 5

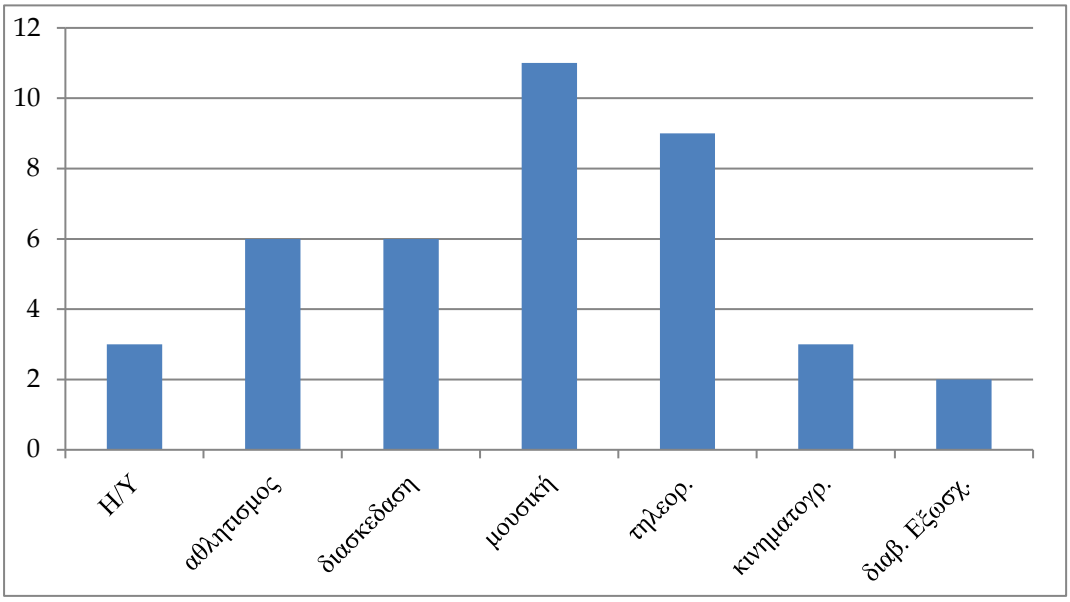
Κατανομή συχνοτήτων για την απασχόληση στον ελεύθερο χρόνο τους των φοιτητών

i	**Απασχόληση x_i**	**Συχνότητα v_i**	**Σχετική συχνότητα f_i**	**Σχετική συχνότητα f_i %**
1	Υπολογιστές	3	0,075	7,5
2	Αθλητισμός	6	0,150	15,0

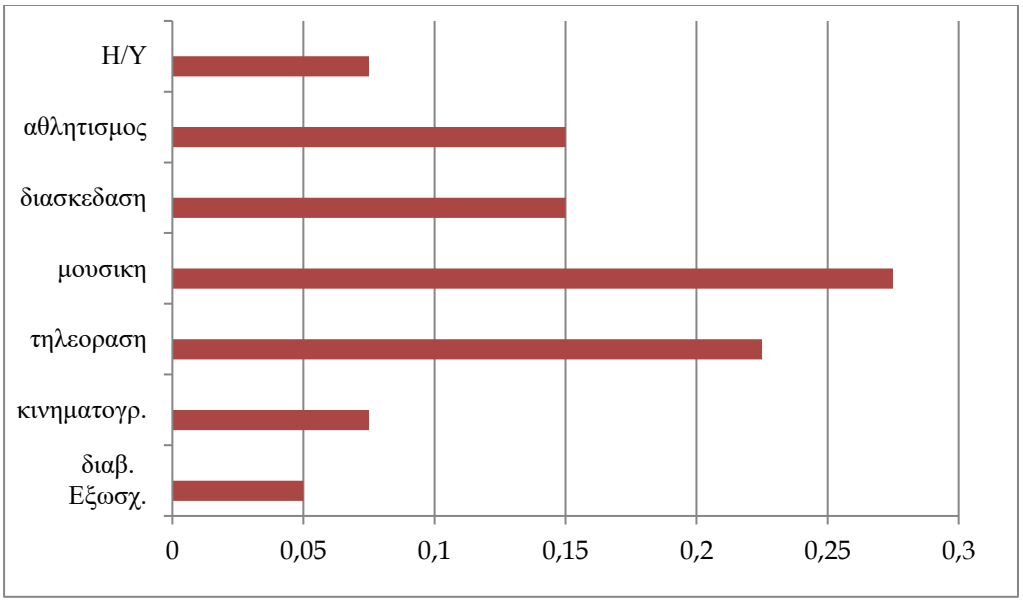
3	Διασκέδαση	6	0,150	15,0
4	Μουσική	11	0,275	27,5
5	Τηλεόραση	9	0,225	22,5
6	Κινηματογράφος	3	0,075	7,5
7	Διάβασμα εξωσχ. Βιβλίων και άλλα	2	0,050	5,0
Σύνολο		40	1,000	100,0

Πηγή: Διπλωματική εργασία Μπαλάφα Αικ, Στατιστική και ο Ρόλος στις Επιχειρήσεις, 2013.

Μερικές φορές σε ένα ραβδόγραμμα συχνοτήτων ο ρόλος των δύο αξόνων είναι δυνατών να αντιστραφεί, όπως φαίνεται στο σχήμα 1 (β), που παριστάνεται το ραβδόγραμμα σχετικών συχνοτήτων της ίδιας μεταβλητής.



Σχήμα 1 (α) Σύμφωνα με τα δεδομένα του πίνακα 5



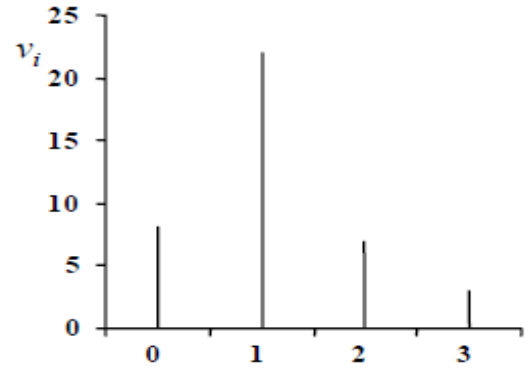
Σχήμα 1 (β)

Ραβδόγραμμα συχνοτήτων (α) και σχετικών συχνοτήτων (β) για την απασχόληση των φοιτητών με τα δεδομένα του πίνακα 5.

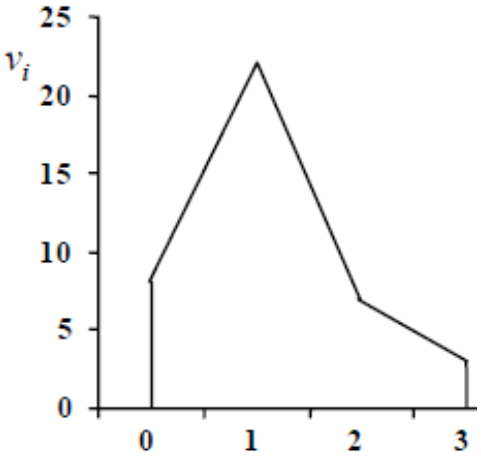
Β) Διάγραμμα συχνοτήτων

Όταν έχουμε μια ποσοτική μεταβλητή αντί του ραβδογράμματος χρησιμοποιείται το **διάγραμμα συχνοτήτων**. Αυτό μοιάζει με το ραβδόγραμμα με μόνη διαφορά ότι αντί να χρησιμοποιούμε συμπαγή ορθογώνια υψώνουμε σε κάθε x_i (υποθέτοντας ότι $X_1<X_2<\ldots<X_k$) μία κάθετη γραμμή με μήκος ίσο προς την αντίστοιχη συχνότητα όπως φαίνεται στο σχήμα 2.

Ενώνοντας τα σημεία (x_i , ν_i) ή (x_i , f_i) έχουμε το λεγόμενο **πολύγωνο συχνοτήτων** ή **πολύγωνο σχετικών συχνοτήτων**, αντίστοιχα, που μας δίνουν μια γενική ιδέα για την μεταβολή της συχνότητας ή της σχετικής συχνότητας όσο μεγαλώνει η τιμή της μεταβλητής που εξετάζουμε, όπως φαίνεται στο σχήμα 2 (β).



Σχήμα 2 (α) σύμφωνα με τα δεδομένα του Πίνακα



Σχήμα 2 (β) σύμφωνα με τα δεδομένα του πίνακα

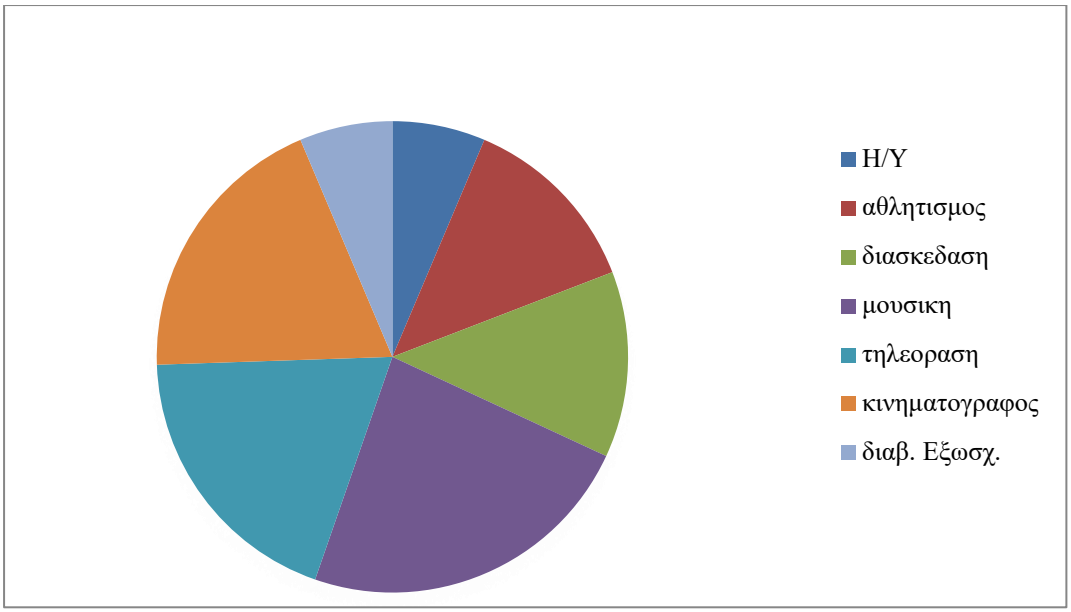
Γ) Κυκλικό διάγραμμα

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. Για να κατασκευάσουμε ένα κυκλικό διάγραμμα χωρίζουμε τον κυκλικό δίσκο σε κυκλικούς τομείς των οποίων οι επίκεντρες γωνίες βγαίνουν σε τόξα α_i ανάλογα με τις συχνότητες v_i των τιμών της μεταβλητής.

Επειδή τα ποσά α_i και v_i είναι ανάλογα, ισχύει:

$$\alpha_i = \frac{vi\ 360}{\nu} = 360^o\ f_i \text{ για } i=1,2,3,\ldots,\kappa. \quad (3)$$

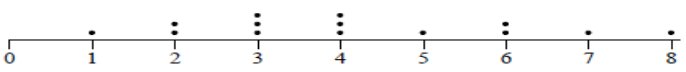
Στο σχήμα 3 παριστάνεται το αντίστοιχο κυκλικό διάγραμμα σχετικών συχνοτήτων της ‘απασχόλησης των φοιτητών’.



Σχήμα 3 Κυκλικό διάγραμμα σχετικών συχνοτήτων της απασχόλησης φοιτητών

Δ) Σημειόγραμμα

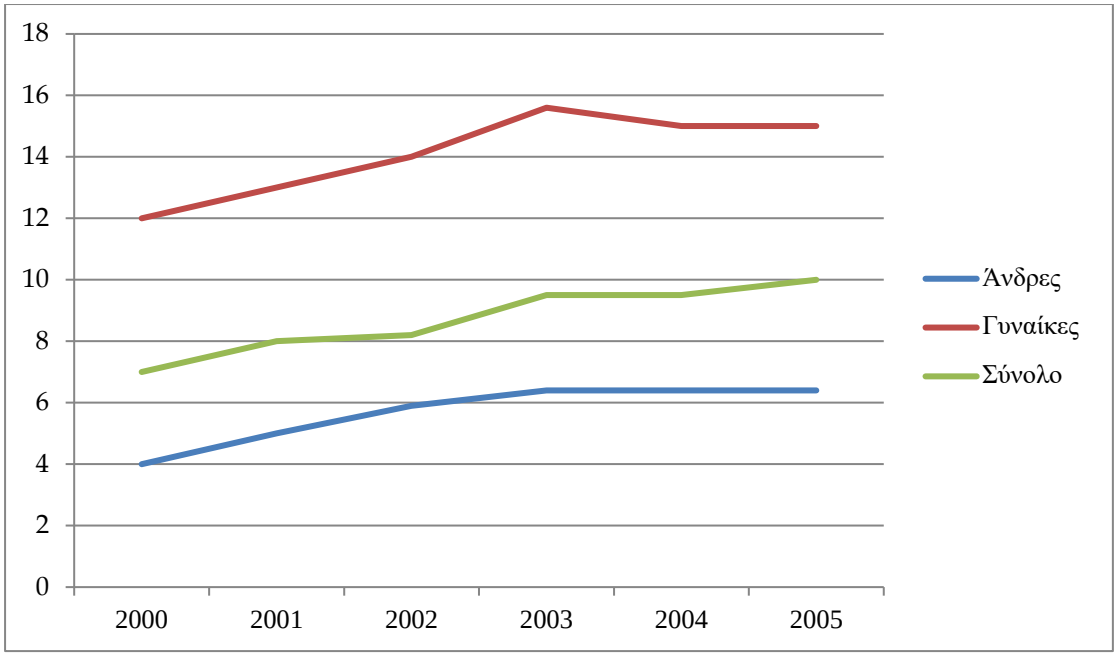
Όταν έχουμε λίγες παρατηρήσεις, η κατανομή τους μπορεί να περιγραφεί με το **σημειόγραμμα**, στο οποίο οι τιμές παριστάνονται γραφικά σαν σημεία υπεράνω ενός οριζόντιου άξονα. Στο παρακάτω σχήμα 4 έχουμε το σημειόγραμμα των χρόνων (σε λεπτά) 4,2,3,1,5,6,4,2,3,4,7,4,8,6,3 που χρειάστηκαν 15 φοιτητές για να λύσουν ένα πρόβλημα.



Σχήμα 4

Ε) Χρονόγραμμα

Το **χρονόγραμμα** ή **χρονολογικό διάγραμμα** χρησιμοποιείται για την γραφική απεικόνιση ενός οικονομικού, δημογραφικού ή άλλου μεγέθους. Ο οριζόντιος άξονας χρησιμοποιείται συνήθως ως άξονας μέτρησης του χρόνου και ο κάθετος ως άξονας μέτρησης της εξεταζόμενης μεταβλητής. Στο παρακάτω σχήμα 5 έχουμε το χρονόγραμμα του ποσοστού ανεργίας στη χώρα μας από το 2000 έως το 2005.



Σχήμα 5 Ποσοστό ανεργίας στην Ελλάδα, πηγή Ε.Σ.Υ.Ε

Στο σχήμα παρατηρούμε ότι στο γυναικείο πληθυσμό υπάρχει συστηματικά ποσοστό ανεργίας, γύρω στις 8 εκατοστιαίες μονάδες. Στο διάστημα 2003-2005 το ποσοστό ανεργίας έχει σταθεροποιηθεί γύρω στο 6,5% για τους άνδρες και γύρω στο 15% για τις γυναίκες.

1.8 Ιστόγραμμα συχνοτήτων

Στην περίπτωση διακριτής μεταβλητής, όταν το πλήθος των τιμών της είναι μεγάλο, αλλά πολύ περισσότερο σε συνεχή μεταβλητή X, ταξινομούμε τα δεδομένα σε ίσα διαδοχικά διαστήματα της μορφής [,) (κλάσεις ή τάξεις) και καταγράφουμε τις συχνότητες των παρατηρήσεων που ανήκουν σε κάθε κλάση. Για την i κλάση [α, b) η κεντρική τιμή xi είναι:

$$Xi=\frac{\alpha+b}{2} \quad (4)$$

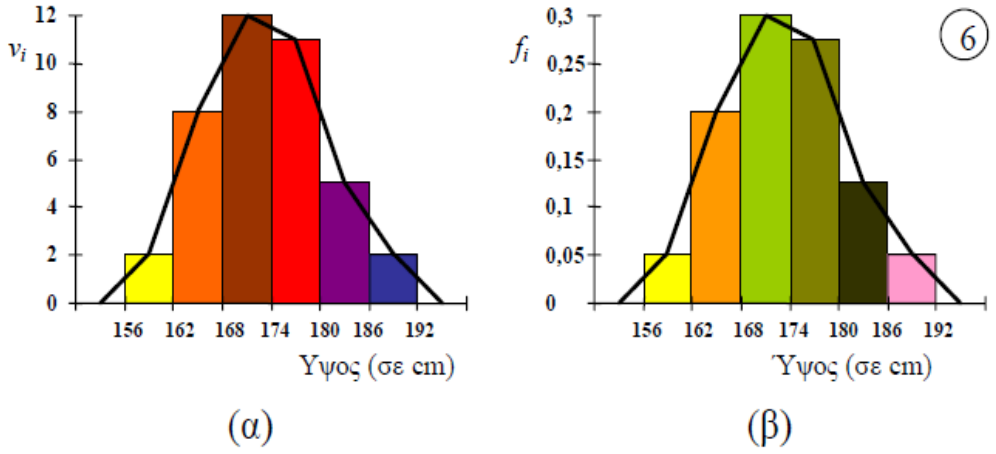
Ιστόγραμμα συχνοτήτων ονομάζεται η γραφική παράσταση ενός πίνακα συχνοτήτων με ομαδοποιημένα δεδομένα. Στον οριζόντιο άξονα ενός συστήματος ορθογωνίων αξόνων σημειώνουμε, με κατάλληλη κλίμακα, τα όρια των κλάσεων. Στη συνέχεια κατασκευάζουμε διαδοχικά ορθογώνια (ιστούς), από καθένα από τα οποία έχει βάση ίση με το πλάτος της κλάσης και ύψος τέτοιο, ώστε **το εμβαδόν του ορθογωνίου να ισούται με την συχνότητα** της κλάσης αυτής.

Α) Κλάσεις ίσου πλάτους

Θεωρώντας το πλάτος c ως μονάδα μέτρησης στον οριζόντιο άξονα, το ύψος κάθε ορθογωνίου είναι ίσο προς τη συχνότητα της αντίστοιχης κλάσης, έτσι ώστε να ισχύει ότι το εμβαδόν των ορθογωνίων είναι ίσο με τις αντίστοιχες συχνότητες. Με ανάλογο τρόπο γίνεται και το **ιστόγραμμα σχετικών συχνοτήτων**.

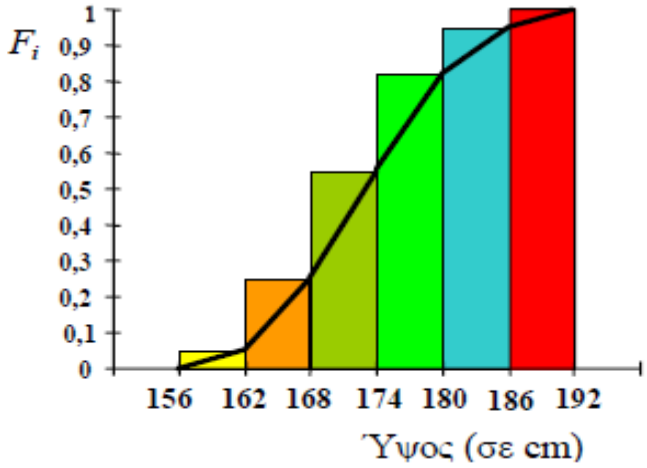
Αν στα ιστογράμματα συχνοτήτων προσθέσουμε ακόμα δύο κλάσεις, στην αρχή και στο τέλος, με συχνότητα μηδέν και στην συνέχεια ενώσουμε τα μέσα των άνω βάσεων των ορθογωνίων, σχηματίζεται το λεγόμενο **πολύγωνο συχνοτήτων**. Το εμβαδόν του χωρίου που ορίζεται από το πολύγωνο συχνοτήτων και τον οριζόντιο άξονα είναι ίσο με το άθροισμα των συχνοτήτων, δηλαδή με το μέγεθος του δείγματος ν.

Στο παρακάτω **σχήμα 6** κατασκευάζεται από το ιστόγραμμα σχετικών συχνοτήτων και το πολύγωνο σχετικών συχνοτήτων με εμβαδόν ίσο με 1.



Ιστόγραμμα και πολύγωνο (α) συχνοτήτων και (β) σχετικών συχνοτήτων.

Το **ιστόγραμμα αθροιστικών συχνοτήτων** και **σχετικών συχνοτήτων** κατασκευάζονται με τον ίδιο τρόπο. Αν ενώσουμε σε ένα ιστόγραμμα αθροιστικών συχνοτήτων τα δεξιά άκρα των άνω βάσεων των ορθογωνίων με ευθύγραμμα τμήματα βρίσκουμε το **πολύγωνο αθροιστικών συχνοτήτων της κατανομής**. Στο σχήμα 7 παριστάνεται το ιστόγραμμα και το πολύγωνο αθροιστικών σχετικών συχνοτήτων για το ύψος των φοιτητών.



Σχήμα 7

Β) Κλάσεις άνισου πλάτους

Συνήθως επιλέγουμε κλάσεις ίσου πλάτους, υπάρχουν όμως και περιπτώσεις που είναι απαραίτητο να έχουμε και κλάσεις διαφορετικού πλάτους όπως, για παράδειγμα στην

περίπτωση όπου οι συχνότητες σε κάποιες κλάσεις να είναι πολύ μικρές τότε γίνεται συγχώνευση κελιών ή ακόμα και στην κατανάλωση νερού.

Για παράδειγμα, η διάρκεια (σε sec) v=80 τηλεφωνημάτων που έγιναν τυχαία από ένα κινητό τηλέφωνο όπως δίνεται στον παρακάτω πίνακα.

Πίνακας 6

Διάρκεια τηλεφωνημάτων σε sec	**Συχνότητα v_i**
0-20	20
20-25	20
25-30	24
30-40	16
Σύνολο	V=80

Πηγή: Διπλωματική εργασία Μπαλάφα Αικ, Στατιστική και ο Ρόλος στις Επιχειρήσεις, 2013.

Το αντίστοιχο ιστόγραμμα συχνοτήτων κατασκευάζεται πάλι, έτσι ώστε το εμβαδόν κάθε ορθογωνίου να ισούται με την συχνότητα της αντίστοιχης κλάσης. Άρα αν c_i είναι το πλάτος της κλάσεις i με συχνότητα v_i, το ύψος του ορθογωνίου θα είναι:

$$u_i = \frac{vi}{ci} \text{ , i=1,2,…, κ. (5)}$$

Για την κατασκευή του ιστογράμματος συχνοτήτων χρειαζόμαστε τα πλάτη των κλάσεων και τα ύψη των ορθογωνίων. Αυτά δίνονται στον επόμενο πίνακα.

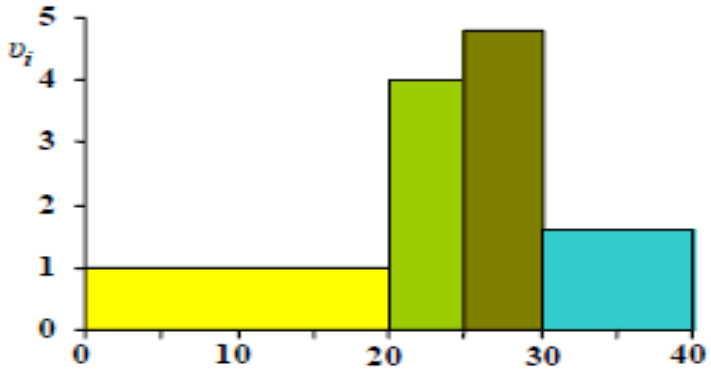
Πίνακας 7

Διάρκεια τηλεφωνημάτων σε sec	**Πλάτος κλάσης c_i**	**Συχνότητα v_i**	**Ύψος $u_i=\frac{vi}{ci}$**	**Ύψος $u_i=\frac{fi\%}{ci}$**

0-20	20	20	1,0	1,25
20-25	5	20	4,0	5,00
25-30	5	24	4,8	6,00
30-40	10	16	1,6	2,00

Πηγή: Διπλωματική εργασία Μπαλάφα Αικ, Στατιστική και ο Ρόλος στις Επιχειρήσεις, 2013.

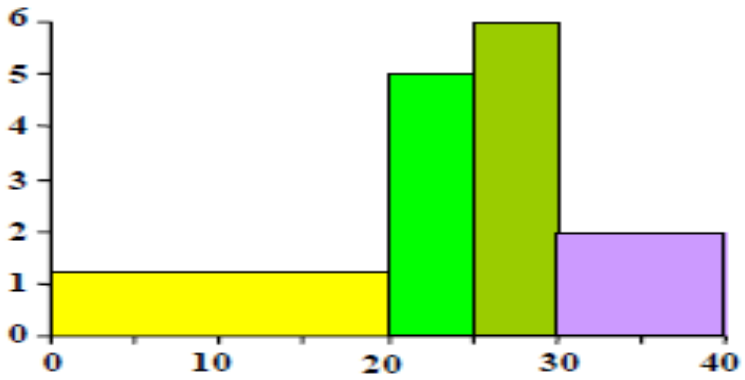
Το ιστόγραμμα συχνοτήτων δίνεται στο **σχήμα 8** (α). Παρατηρούμε ότι το άθροισμα των εμβαδών όλων των ορθογωνίων είναι ίσο με το συνολικό μέγεθος δείγματος ν, όπως δηλαδή συμβαίνει και στο ιστόγραμμα με κλάσεις ίσου πλάτους



Σχήμα 8 (α)

Ιστόγραμμα συχνοτήτων σύμφωνα με τα δεδομένα του πίνακα 7

Με ανάλογο τρόπο κατασκευάζεται και το ιστόγραμμα σχετικών συχνοτήτων σχήμα 8 (β) αρκεί να χρησιμοποιήσουμε ως ύψος των ορθογωνίων το λόγο των σχετικών συχνοτήτων προς το πλάτος των κλάσεων

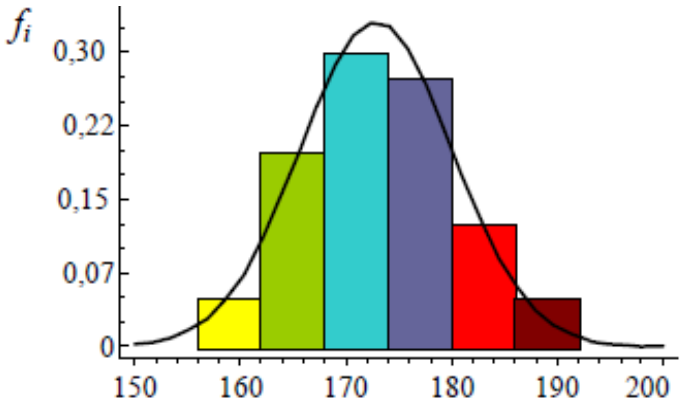


Σχήμα 8 (β)

Ιστόγραμμα σχετικών συχνοτήτων σύμφωνα με τα δεδομένα του πίνακα 7

1.9 Καμπύλες συχνοτήτων

Εάν ο αριθμός των κλάσεων είναι αρκετά μεγάλος δηλαδή τείνει στο άπειρο και το πλάτος των κλάσεων είναι πολύ μικρό δηλαδή τείνει στο μηδέν, τότε η πολυγωνική γραμμή συχνοτήτων τείνει να πάρει τη μορφή μιας ομαλής καμπύλης, η οποία ονομάζεται **καμπύλη συχνοτήτων**, όπως φαίνεται στο παρακάτω σχήμα 9 όπου καμπύλη συχνοτήτων για το ύψος των φοιτητών.



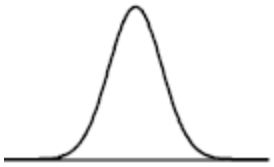
Σχήμα 9 με τα δεδομένα του πίνακα 7

Οι καμπύλες συχνοτήτων έχουν μεγάλη εφαρμογή στη στατιστική, οι ιδιότητες τους μπορούν να χρησιμοποιηθούν για την εξαγωγή χρήσιμων συμπερασμάτων.

Μερικές κατανομές συχνοτήτων:

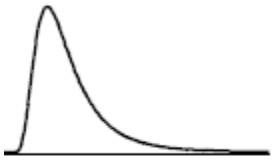
(α)

Στο σχήμα (α) οι παρατηρήσεις 'κατανέμονται' ομοιόμορφα σε ένα διάστημα [α,β], η κατανομή αυτή λέγεται **ομοιόμορφη**.



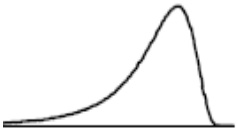
(β)

Στο σχήμα (β), η κατανομή με ' κωδωνοειδή' μορφή λέγεται **κανονική κατανομή**.



(γ)

Όταν οι παρατηρήσεις δεν είναι συμμετρικά κατανεμημένες , η κατανομή λέγεται ασυμμετρία θετική ασυμμετρία ως προς την κατανομή σχήμα (γ).



(δ)

Στο σχήμα (δ) βλέπουμε την κατανομή που λέγεται αρνητική ασυμμετρία.

Κεφάλαιο 2: Στατιστικά μέτρα

2.1 Μέτρα κεντρικής τάσης

Ένα από τα σημαντικότερα μέτρα κεντρικής τάσης αλλά και γενικότερα της στατιστικής θεωρείται ο μέσος αριθμητικός (Arithmetic Mean or Average) ή αλλιώς μέση τιμή.

2.1.1 Μέσος αριθμητικός ή μέση τιμή

Ο **μέσος αριθμητικός** ή **μέση τιμή** είναι ένα από τα σπουδαιότερα μέτρα της στατιστικής .Συμβολίζεται με $\bar{x}$ όταν τα αριθμητικά δεδομένα αφορούν ένα δείγμα, και με "μ" όταν τα δεδομένα αφορούν ολόκληρο πληθυσμό. Όταν έχουμε ένα σύνολο ν παρατηρήσεων, ο αριθμητικός μέσος είναι το άθροισμα των παρατηρήσεων δια το πλήθος των παρατηρήσεων. Για τον υπολογισμό του αριθμητικού μέσου διακρίνουμε δύο περιπτώσεις:

1)Υπολογισμός αριθμητικού μέσου βάσει αταξινόμητων δεδομένων.

2) Υπολογισμός αριθμητικού μέσου βάσει ταξινομημέων δεδομένων.

Α. Υπολογισμός αριθμητικού μέσου βάσει αταξινόμητων δεδομένων

Σε περίπτωση αταξινόμητων δεδομένων ο μέσος αριθμητικός διακρίνεται δύο τρόπους υπολογισμού. Στον απλό ή αστάθμητο μέσο αριθμητικό και στον σταθμικό μέσο αριθμητικό.

1) ΑΠΛΟΣ Ή ΑΣΤΑΘΜΗΤΟΣ ΑΡΙΘΜΗΤΙΚΟΣ ΜΕΣΟΣ

Όταν έχουμε μια σειρά 10 τιμών : $x_1,x_2,x_3,...,x_{10}$ ο αστάθμητος αριθμητικός μέσος είναι το πηλίκο τους αθροίσματος των τιμών δια το πλήθος των τιμών και υπολογίζεται με το τύπο:

$$\bar{x}=\frac{x1+x2+x3+\ldots+xn}{n} = \frac{1}{n} \times \sum_{i=1}^{n} xi \quad (1)$$

όπου i=1,2,3,...,n

Σημείωση:

Σ = το άθροισμα των τιμών της μεταβλητής.

ΠΑΡΑΔΕΙΓΜΑ : Έστω η μισθοδοσία 6 υπαλλήλων για το μήνα Δεκέμβριο : 1000€, 900€, 750€, 1100€, 800€, 700€.

Τότε ο αστάθμητος αριθμητικός μέσος είναι :

$$\bar{x}=\frac{x1+x2+x3+\ldots+x5}{n}=\frac{1000+900+750+1100+800+700}{6}=\frac{5300}{6}=883{,}33€$$

Αυτό σημαίνει πως αν οι υπάλληλοι είχαν την ίδια μισθοδοσία, θα έπαιρναν και οι 6 υπάλληλοι 883,33 ο καθένας.

1) ΣΤΑΘΜΙΚΟΣ ΑΡΙΘΜΗΤΙΚΟΣ ΜΕΣΟΣ

Όταν έχουμε μια σειρά n τιμών: $x_1,x_2,x_3,\ldots,x_n$ στις οποίες δίνεται διαφορετική σημασία, που εκφράζονται από κάποιους αριθμούς που λέγονται συντελεστές βαρύτητας (ή σταθμίσεως) : $w_1,w_2,w_3,\ldots,w_n$ τότε χρησιμοποιούμε το **σταθμικό αριθμητικό μέσο** και υπολογίζεται με το τύπο:

$$\bar{x}=\frac{x1w1+x2w2+x3w3+\cdots+xnwn}{w1+w2+w3+\cdots+wn} \quad (2)$$

όπου i=1,2,3,...,n

Σxiwi είναι το άθροισμα των γινομένων κάθε τιμής x επί τον αντίστοιχο συντελεστή βαρύτητας.

Σwi είναι το άθροισμα των συντελεστών βαρύτητας.

ΠΑΡΑΔΕΙΓΜΑ : Έστω οι βαθμολογίες ενός μαθητή στις πανελλαδικές εξετάσεις. Μαθηματικά 16,Βιολογία 12,Φυσική 11,Έκθεση 12,Χημεία 15.

Και οι συντελεστές βαρύτητας των μαθημάτων, αντίστοιχα 2, 1.5, 1, 0.5, 1.

Τότε έχουμε

$\Sigma x_i w_i = x_1w_1 + x_2w_2 + x_3w_3 + x_4w_4 + x_5w_5$ =16 x2 + 12 x1,5 + 11 x1 + 12 x0,5 + 15 x1 =32 + 18 + 11 + 6 + 15 = 82

και

$\Sigma w_i = w1+w2+w3+w4+w5$ = 2 + 1.5 + 1 + 0.5 + 1 = 6

Άρα ο σταθμικός αριθμητικός μέσος είναι:

$$\bar{x} = \frac{\Sigma xiwi}{\Sigma wi} = \frac{82}{6} = 13,6.$$

Β) ΥΠΟΛΟΓΙΣΜΟΣ ΑΤΙΘΜΗΤΙΚΟΥ ΜΕΣΟΥ ΒΑΣΕΙ ΤΑΞΙΝΟΜΗΜΕΝΩΝ ΔΕΔΟΜΕΝΩΝ ΣΕ ΚΑΤΑΝΟΜΗ ΣΥΧΝΟΤΗΤΩΝ ΜΕ ΙΣΑ ΔΙΑΣΤΗΜΑΤΑ ΤΑΞΕΩΝ.

Όταν τα δεδομένα που έχουμε είναι ταξινομημένα σε μια κατανομή συχνοτήτων, τότε λέμε ότι ο μέσος αριθμητικός είναι σταθμικός. Στη φάση αυτή ως τιμές της μεταβλητής (x_i) θεωρούνται οι κεντρικοί όροι των τάξεων και οι συντελεστές σταθμίσεως οι αντίστοιχες συχνότητες κάθε κατανομής (f_i). Για να υπολογίσουμε το μέσο αριθμητικό σε αυτή τη περίπτωση έχουμε 2 τρόπους υπολογισμού. Τον άμεσο και τον έμμεσο.

ΑΜΕΣΟΣ ΤΡΟΠΟΣ : Έστω $x_1,x_2,x_3,...,x_n$ οι κεντρικοί όροι των τάξεων μιας κατανομής συχνοτήτων και $f_1,f_2,f_3,...,f_n$ οι αντίστοιχες συχνότητες της κατανομής. Για τον υπολογισμό, λοιπόν του μέσου αριθμητικού χρησιμοποιούμε τον τύπο:

$$\bar{x} = \frac{X1f1+X2f2+\cdots+Xnfn}{f1+f2+f3+\cdots+fn} = \frac{1}{N} * \sum_{i=1}^{n} xifi \quad (3)$$

Όπου

$$N = f_1 + f_2 + \ldots + f_n = \sum_{i=1}^{n} xifi$$

ΠΑΡΑΔΕΙΓΜΑ : Έστω το μέσο ημερομίσθιο μιάς κατανομής 100 εργατών.

Πίνακας 1

Τάξεις ημερομισθίων σε ευρώ (κλάσεις)	**Κεντρικοί όροι τάξεων** x_i	**Αριθμός εργατών** f_i	**Γινόμενα** $x_i f_i$
1500-2500	2000	5	10000
2500-3500	3000	13	39000
3500-4500	4000	20	80000
4500-5500	5000	35	175000

5500-6500	6000	18	108000
6500-7500	7000	7	49000
7500-8500	8000	2	16000
Σύνολο	----------	100= N	477000

Πηγή: Διπλωματική εργασία Μπαλάφα Αικ, Στατιστική και ο Ρόλος στις Επιχειρήσεις, 2013.

Το άθροισμα των γινομένων X_iF_i δηλώνει το συνολικό ποσό της ημερήσιας μισθοδοσίας των 100 εργατών.

Με βάση τον τύπο έχουμε:

$$\bar{x} = \frac{\Sigma xifi}{\Sigma fi} = \frac{477000}{100} = 4700$$

Ερμηνεία:

Αν όλοι οι εργάτες αμοίβονταν εξίσου, έπαιρναν δηλαδή το ίδιο ημερομίσθιο, τότε το ημερομίσθιο κάθε εργάτη θα ήταν 4700 ευρώ.

ΕΜΜΕΣΟΣ ΤΡΟΠΟΣ : Για τον υπολογισμό μέσου αριθμητικού με έμμεσο τρόπο ακολουθούμε την εξής διαδικασία :

α)Αντικαθιστούμε τη μεταβλητή x_i (δηλαδή τους κεντρικούς όρους των τάξεων) με μια νέα μεταβλητή ξ_i. Η διαφορά ανάμεσα στη μεταβλητή x_i με τη μεταβλητή ξ_i είναι πως η μεταβλητή x_i αποτελείται από μεγάλους αριθμούς, ενώ η μεταβλητή ξ_i αποτελείται από νομοψήφιους αριθμούς.

β)Επιλέγουμε μια τιμής της μεταβλητής x_i, και αυτή είναι συνήθως η τιμή που αντιστοιχεί στη μεγαλύτερη συχνότητα της κατανομής, την οποία συμβολίζουμε με το X_o.

γ)Στη συνέχεια, από κάθε τιμή της μεταβλητής x_i αφαιρούμε την τιμή X_o και τις διαφορά αυτών τη διαιρούμε με το διάστημα των τάξεων, το οποίο συμβολίζεται με δ.

Έτσι, έχουμε τον τύπο:

$$\xi_i = \frac{Xi - Xo}{\delta} \quad (4)$$

δ) Έπειτα πολλαπλασιάζουμε κάθε τιμή της μεταβλητής ξ_i με τις αντίστοιχες συχνότητες ($=f_i$), και τα γινόμενα τα γράφουμε σε μια στήλη που συμβολίζεται με $\xi_i f_i$.

ε) Μετά αθροίζουμε τα γινόμενα της στήλης $\xi_i f_i$ και το άθροισμα τους το συμβολίζουμε με $\Sigma \xi_i f_i$.

στ) Τέλος για να υπολογίσουμε το μέσο αριθμητικό εφαρμόζουμε τον τύπο :

$$\overline{X} = \frac{Xo + \delta * \Sigma \xi i fi}{\Sigma fi} \quad (5)$$

ΠΑΡΑΔΕΙΓΜΑ : Έστω το μέσο ημερομίσθιο μιάς κατανομής 100 εργατών.

Πίνακας 2

Τάξεις ημερομισθίων	Κεντρικοί όροι x_i	Συχνότητες f_i	ξ_i	$\xi_i f_i$

σε ευρώ (κλάσεις)				
1500-2500	2000	5	-3	-15
2500-3500	3000	13	-2	-26
3500-4500	4000	20	-1	-20
4500-5500	5000	35	0	0
5500-6500	6000	18	1	18
6500-7500	7000	7	2	14
7500-8500	8000	2	3	6
Σύνολο	---------	100	-------	-23

Με βάση τη παραπάνω διαδικασία, στο πίνακα έχουμε:

$$X_o = 5000,\ \delta = 1000,\ \Sigma f_i = 100,\ \ \Sigma \xi_i f_i = -23$$

Σύμφωνα με το τύπο του αριθμητικού μέσου αντικαθιστούμε τα δεδομένα και βρίσκουμε:

$$\overline{X} = \frac{5000+1000*(-23)}{100} = \text{ 4770 ευρώ.}$$

Έτσι παρατηρούμε ότι και με τον άμεσο και με τον έμμεσο τρόπο έχουμε το ίδιο αποτέλεσμα.

Γ) ΥΠΟΛΟΓΙΣΜΟΣ ΑΡΙΘΜΗΤΙΚΟΥ ΜΕΣΟΥ ΒΑΣΕΙ ΤΑΞΙΝΟΜΗΜΕΝΩΝ ΔΕΔΟΜΕΝΩΝ ΣΕ ΚΑΤΑΝΟΜΗ ΣΥΧΝΟΤΗΤΩΝ ΜΕ ΑΝΙΣΑ ΔΙΑΣΤΗΜΑΤΑ ΤΑΞΕΩΝ

Σε περίπτωση που το διάστημα των τάξεων (δ) δεν είναι ίδιο για όλες τις τάξεις, τότε στον άμεσο τρόπο ακολουθούμε την ίδια διαδικασία, όπως και στη κατανομή συχνοτήτων με ίσα διαστήματα τάξεως, όταν θέλουμε να υπολογίσουμε το μέσο αριθμητικό. Στον έμμεσο τρόπο, για τον υπολογισμό του μέσου αριθμητικού ισχύουν τα παραπάνω με τη διαφορά ότι για το σχηματισμό της στήλης ξ_i, επιλέγουμε ένα διάστημα τάξεως δ, έτσι ώστε οι διαιρούμενες με το δ διαφορές (X_i-X_o) να δίνουν ολιγόψηφα πηλίκα. Αλλιώς για τον υπολογισμό αριθμητικού μέσου χρησιμοποιούμε τον άμεσο τρόπο.

ΠΑΡΑΔΕΙΓΜΑ : Ο παρακάτω πίνακας μας δείχνει τη διαδικασία υπολογισμού του μέσου αριθμητικού σε περίπτωση που τα διαστήματα των τάξεων είναι άνισα. Έστω λοιπόν, το μέσο οικογενειακό εισόδημα 1000000 οικογενειών.

Πίνακας 3

Τάξεις οικογενειακού εισοδήματος σε ευρώ (κλάσεις)	Κεντρικοί όροι x_i	Αριθμός οικογενειών f_i	ξ_i	$\xi_i f_i$
0-50000	2500	7200	-1,25	-90000

50000-100000	7500	165200	-0,75	-123000
100000-200000	150000	360800	0	0
200000-400000	300000	360000	1,5	540000
400000-1000000	700000	42000	5,5	231000
Σύνολο	-------	1000000	------	557100

Υπολογίζοντας, ξέρουμε ότι

X_o=150000, δ=100000, Σf_i=1000000, $\Sigma \xi_i f_i$=557100.

Έτσι αντικαθιστώντας τα παραπάνω δεδομένα έχουμε δύο επιλογές:

1) Με τον άμεσο τρόπο το μέσο οικογενειακό εισόδημα είναι :

$$\bar{\chi} = \frac{\Sigma xifi}{\Sigma fi} = \frac{205710000000}{1000000} = 205710 \text{ ευρώ}$$

2) Με τον έμμεσο τρόπο το μέσο οικογενειακό εισόδημα είναι :

$$\bar{\chi} = \frac{Xo+\delta*\Sigma\xi ifi}{\Sigma fi} = \frac{150000+100000*557100}{1000000} = 205710 \text{ ευρώ.}$$

2.2 Μέτρα παράμετροι θέσεως

Με βάση τα προηγούμενα ,είπαμε πως ο μέσος αριθμητικός είναι ένα από τα σπουδαιότερα στατιστικά μέτρα. Αυτό βέβαια συμβαίνει όταν έχουμε μια κατανομή συχνοτήτων της οποίας η καμπύλη συχνοτήτων είναι συμμετρική. Σε περίπτωση που η

κατανομή συχνοτήτων είναι ασυμμετρική, τότε ο μέσος αριθμητικός δεν είναι αντιπροσωπευτικό στατιστικό μέτρο για τη λήψη ορθών αποφάσεων. Για την αντιμετώπιση της συγκεκριμένης περίπτωσης υπάρχουν άλλα στατιστικά μέτρα, τα οποία οδηγούν σε αξιόπιστες πληροφορίες για τον εξεταζόμενο στατιστικό πληθυσμό. Τα στατιστικά αυτά μέτρα ονομάζονται παράμετροι θέσεως και είναι :

α) η διάμεσος,

β) τα τεταρτημόρια,

γ) τα δεκατημόρια,

δ) τα εκατοστημόρια, και

ε) η επικρατούσα τιμή ή τύπος.

2.2.1. Διάμεσος

Όταν έχουμε μια σειρά n τιμών τις οποίες τις ταξινομούμε με αύξουσα σειρά, κατά τη φυσική τους κατάσταση, δηλαδή από τη μικρότερη προς τη μεγαλύτερη, τότε η διάμεσος ορίζεται ως η μεσαία τιμή όταν το n είναι περιττός αριθμός, ή ο μέσος όρος των δύο μεσαίων τιμών όταν ο n είναι άρτιος αριθμός. Η διάμεσος συμβολίζεται με το σύμβολο "Μ" και θεωρείται το σπουδαιότερο στατιστικό μέτρο θέσεως. Για τον υπολογισμό της διαμέσου διακρίνουμε τις παρακάτω περιπτώσεις:

Α) ΥΠΟΛΟΓΙΣΜΟΣ ΔΙΑΜΕΣΟΥ ΑΤΑΞΙΝΟΜΗΤΩΝ ΔΕΔΟΜΕΝΩΝ

Σε περίπτωση που τα δεδομένα είναι αταξινόμητα και έχουν τοποθετηθεί κατά αύξουσα σειρά για τον υπολογισμό της διακρίνουμε εξίσου 2 περιπτώσεις:

α) Όταν το πλήθος των τιμών είναι μονός αριθμός, τότε η μεσαία τιμή που κατέχει τη κεντρική θέση είναι η διάμεσος.

ΠΑΡΑΔΕΙΓΜΑ : Έστω τιμές :

121, 135, 127, 139, 114, 129, 148, 137, 161.

Αρχικά κατατάσσουμε τις τιμές κατά αύξουσα σειρά:

114, 121, 127, 129, 135, 137, 139, 148, 161.

Η κεντρική θέση των τιμών αυτών είναι:

$$\frac{n+1}{2} = \frac{9+1}{2} = \frac{10}{2} = 5 \quad (6)$$

Αυτό σημαίνει ότι η διάμεσος βρίσκεται στη 5η θέση, και είναι η τιμή 135.

β) Όταν το πλήθος των τιμών είναι ζυγός αριθμός, τότε σε αυτή τη περίπτωση υπάρχουν δύο τιμές που κατέχουν τη κεντρική θέση και όχι μία. Ο μέσος όρος των τιμών αυτών που κατέχουν τη κεντρική θέση είναι η διάμεσος.

ΠΑΡΑΔΕΙΓΜΑ : Έστω οι παρακάτω τιμές, κατά αύξουσα σειρά:

114,121,127,129,135,137,139,148.

Η κεντρική θέση των τιμών αυτών είναι **:**

$$\frac{n+1}{2} = \frac{8+1}{2} = 4{,}5$$

Αυτό σημαίνει ότι η διάμεσος βρίσκεται ανάμεσα στη 4η και στη 5η θέση.

Επομένως, βάσει του τύπου, η διάμεσος είναι:

$$M = \frac{129+135}{2} = 132. \quad (7)$$

Β) ΥΠΟΛΟΓΙΣΜΟΣ ΔΙΑΜΕΣΟΥ ΤΑΞΙΝΟΜΗΜΕΩΝ ΔΕΔΟΜΕΝΩΝ ΣΕ ΚΑΤΑΝΟΜΗ ΣΥΧΝΟΤΗΤΩΝ

Σε περίπτωση που τα δεδομένα είναι ταξινομημένα σε κατανομή συχνοτήτων η διάμεσος υπολογίζεται βάσει του τύπου:

$$\text{M} = \frac{xi+\delta}{fi\,[\frac{n}{2}+\Phi i]} \quad (8)$$

Σημείωση:

X_i είναι το κατώτερο όριο της τάξεως που εντοπίζεται η διάμεσος.

δ είναι το διάστημα της τάξεως που εντοπίζεται η διάμεσος.

f_i είναι η συχνότητα της τάξεως που εντοπίζεται η διάμεσος.

Ν είναι το σύνολο των συχνοτήτων της κατανομής.

Φ_i είναι η μικρότερη του $\frac{N}{2}$ δεξιόστροφη αθροιστική συχνότητα, δηλαδή $\Phi_i <= \frac{N}{2}$.

Μ είναι η διάμεσος.

ΠΑΡΑΔΕΙΓΜΑ: Έστω, μια κατανομή συχνοτήτων 100 εργατών.

Πίνακας 4

Τάξεις ημερομίσθιων σε ευρώ (κλάσεις)	Συχνότητες f_i	Δεξιόστροφη αθροιστική σειρά Φ_i	Φ_i
2000-3000	6	6	6
3000-4000	14	20	20
4000-5000	21	41	41
5000-6000	36	74	74
6000-7000	18	92	92
7000-8000	5	97	97
8000-9000	3	100	100
Σύνολο	100	-------	--------

Για τον υπολογισμό της διαμέσου ακολουθούμε την εξής διαδικασία :

α) Μετασχηματίζουμε τις απόλυτες συχνότητες f_i της κατανομής στη δεξιόστροφη αθροιστική σειρά Φ_i.

β) Προσδιορίζουμε την τιμή $\frac{N}{2}$,όπου N= Σf_i.

γ) Η τιμή $\frac{N}{2} = \frac{100}{2}$ =50,περιλαμβάνεται μεταξύ των ορών 41= Φ_i και 74 = Φ_i+1.

Άρα η διάμεσος εντοπίζεται μεταξύ του 5000 και 6000.

δ) Ο όρος 74=Φi+1 ανήκει στην τάξη 5000-6000 το κατώτερο αυτής της τάξεως είναι η τιμή του X_i.

ε) Η τάξη 5000-6000 περιέχει 36 συχνότητες, αυτές είναι η τιμή του f_i.

στ) Το δ είναι το πλάτος της τάξεως που εντοπίζεται η Διάμεσος, σε κατανομή με ίσα ή άνισα διαστήματα τάξεων.

Έτσι έχουμε,X_i=5000,

δ=1000,

f_i=36,

N=100($\frac{N}{2}$ = 50)

και Φ_i=41.

Αντικαθιστώντας τα δεδομένα στο τύπο που ξέρουμε, έχουμε:

$$M = \frac{Xi+\delta}{fi\,[\frac{N}{2}-\Phi\iota]} = \frac{5000+1000}{36[\,50-41\,]} = 5000 + (27{,}7^* 9) = 5000+250 = 5250 \text{ ευρώ.}$$

Ερμηνεία :

Αυτό σημαίνει ότι το 50% των εργατών έχουν ημερομίσθιο μέχρι 5250 ευρώ, και το υπόλοιπο 50% έχει ημερομίσθιο άνω των 5250 ευρώ.

2.2.2 Τεταρτημόρια – Δεκατημόρια – Εκατοστημόρια

Παράμετροι θέσεως εκτός από τη διάμεσο, είναι επίσης και τα **τεταρτημόρια** (Quartiles), **τα δεκατημόρια** (Deciles) και **τα εκατοστημόρια** (Centiles). Ο ρόλος των τεταρτημορίων είναι να υποδιαιρούν το σύνολο των τιμών σε τέσσερις ισοπληθείς ομάδες, των εκατοστημορίων σε δέκα ισοπληθείς ομάδες, και των εκατοστημορίων σε

εκατό ισοπληθείς ομάδες. Τα μέτρα αυτά χρησιμοποιούνται αρκετά συχνά για τη μελέτη ενός συνόλου δεδομένων.

Α) ΤΕΤΑΡΤΗΜΟΡΙΑ

Το πρώτο τεταρτημόρια συμβολίζεται με το Q_1,το δεύτερο τεταρτημόριο με Q_2,και το τρίτο τεταρτημόριο συμβολίζεται με το Q_3.

Για το Q_1,έχουμε αριστερά το πολύ 25% των παρατηρήσεων και δεξιά το πολύ 75% των παρατηρήσεων. Ομοίως και για το Q_3,έχουμε αριστερά το πολύ 75% των παρατηρήσεων και δεξιά το πολύ 25% των παρατηρήσεων. Για το Q_2 το οποίο αντιστοιχεί στο 50% συμπεραίνουμε πως είναι και η διάμεσος.

Για τον υπολογισμό των τεταρτημορίων έχουμε τις εξής περιπτώσεις:

α) Όταν τα δεδομένα είναι αταξινόμητα το Q_1 το εντοπίζουμε σε ποιά θέση είναι,με το τύπο

$$\frac{n+1}{4}, \quad (9)$$

ενώ το Q_3 το εντοπίζουμε σε ποιά θέση είναι, με το τύπο

$$\frac{3\,(\,n+1)}{4}\,. \quad (10)$$

β) Όταν τα δεδομένα είναι αταξινόμητα τότε για το πρώτο τεταρτημόριο χρησιμοποιούμε τον τύπο

$$Q_1 = \frac{Xi+\delta}{fi\,[\frac{N}{4}-\Phi\iota]} \quad (11)$$

και για το τρίτο τεταρτημόριο χρησιμοποιούμε τον τύπο

$$Q_3 = \frac{Xi+\delta}{fi\,[\frac{3N}{4}-\Phi\iota]}\,. \quad (12)$$

Σημείωση : Για τον υπολογισμό των τεταρτημορίων ακολουθούμε την ίδια διαδικασία που ακολουθήσαμε και για τη διάμεσο.

ΠΑΡΑΔΕΙΓΜΑ : Με βάση τα δεδομένα του προηγούμενου πίνακα έχουμε :

Για το πρώτο τεταρτημόριο :

- $N = \frac{100}{4} = 25$,
- $\Phi_i = 20$,
- $X_i = 4000$,
- $f_i = 21$
- $\delta = 1000$.

Βάσει του τύπου λοιπόν έχουμε :

$Q_1 = \frac{4000+1000}{21\ [\ 25-20]}$ = 4000 + (47,61 * 5) = 4000 + 238 = 4238 ευρώ

Ερμηνεία :

Αυτό σημαίνει ότι τα πρώτα 25% των εργατών έχουν ημερομίσθιο μέχρι 4238 ευρώ.

Για το τρίτο τεταρτημόριο :

- $\frac{3\ N}{4} = \frac{3*100}{4} = 75$,
- $\Phi_i = 74$,
- $X_i = 6000$,
- $f_i = 18$,

- $\delta = 1000$.

Βάσει του τύπου λοιπόν έχουμε :

$Q3 = \frac{6000+1000}{18\ [\ 75-74\ ]} = 6000 + (56 * 1) = 6000 + 56 = 6056$ ευρώ

Ερμηνεία :

Αυτό σημαίνει ότι τα πρώτα 25% των εργατών παίρνουν ημερομίσθιο μέχρι 6056 ευρώ, ενώ τα υπόλοιπα 25% των εργατών παίρνουν πάνω από 6056 ευρώ.

Β) ΔΕΚΑΤΗΜΟΡΙΑ ΚΑΙ ΕΚΑΤΟΣΤΗΜΟΡΙΑ

Έτσι αντίστοιχα, και τα δεκατημόρια υπολογίζονται βάσει του τύπου :

$$Dk = \frac{Xi+\delta}{fi\ [\ k*\frac{N}{10}-\Phi i\ ]}\ , (13)$$

όπου k = 1,2,3,...,9

και τα εκατοστημόρια βάσει του τύπου :

$$Ck = \frac{Xi+\delta}{fi\ [\ k*\frac{N}{100}-\Phi i\ ]}\ , (14)$$

όπου k = 1,2,3,...,9

ΠΑΡΑΔΕΙΓΜΑ : Με βάσει πάλι τα δεδομένα του προηγούμενου για το 10ο εκατοστημόριο έχουμε :

- $\frac{N}{10} = \frac{100}{10} = 10$,
- $\Phi_i = 6$,
- $X_i = 3000$,
- $f_i = 14$
- $\delta = 1000$.

βάσει του τύπου λοιπόν έχουμε :

$C_{10} = 3000 + \frac{1000}{14\ (10-6)} = 3000 + (\ 71{,}42 * 4\ ) = 3000 + 285 = 3285$ ευρώ

Ερμηνεία :

Αυτό σημαίνει ότι τα πρώτα 10% των εργατών έχουν ημερομίσθιο μέχρι 3285 ευρώ.

2.2.3 Επικρατούσα τιμή ή τύπος

Επικρατούσα τιμή ή Τύπος είναι η τιμή εκείνη της μεταβλητής που αντιστοιχεί στη μεγαλύτερη συχνότητα της κατανομής. Συμβολίζεται με M_o και λέγεται και αλλιώς **Σημείο Μεγαλύτερης Συχνότητας**. Για τον υπολογισμό της επικρατούσας τιμής διακρίνουμε δύο περιπτώσεις. Όταν τα δεδομένα είναι ταξινομημένα, και όταν είναι αταξινόμητα.

Στη περίπτωση που τα δεδομένα είναι ταξινομημένα, τότε η επικρατούσα τιμή δίνεται από το τύπο:

$$M_o = \alpha_{i-1} + \delta * \frac{\Delta 1}{\Delta 1 + \Delta 2} \quad (15)$$

Όπου:

α_{i-1} : το κατώτερο όριο της τάξεως στην οποία αντιστοιχεί η μεγαλύτερη συχνότητα (f_i).

δ: το πλάτος της τάξεως με τη μεγαλύτερη συχνότητα, σε κατανομή με ίσα ή άνισα διαστήματα τάξεως.

$\Delta_1 = f_i - f_{i-1}$: η διαφορά της μεγαλύτερης συχνότητας μείον τη προηγούμενη.

$\Delta_2 = f_i - f_{i-1}$: η διαφορά της μεγαλύτερης συχνότητας μείον την επόμενη.

ΠΑΡΑΔΕΙΓΜΑ : Με βάση το προηγούμενο πίνακα έχουμε :

- $\alpha_{i-1} = 5000$
- $\delta = 1000$
- $\Delta_1 = f_i - f_{i-1} = 36 - 21 = 15$
- $\Delta_2 = f_i - f_{i-1} = 36 - 18 = 18$

Με βάσει το τύπο λοιπόν, έχουμε:

$M_0 = \alpha_{i-1} + \delta * \frac{\Delta 1}{\Delta 1 + \Delta 2} = 5000 + 1000 * \frac{15}{15+18} = 5450$ ευρώ.

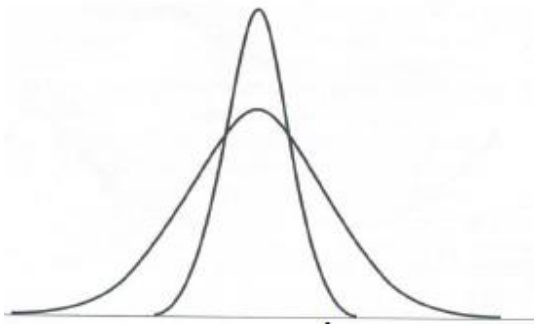
Ερμηνεία :

Αυτό σημαίνει ότι το συχνότερο ημερομίσθιο είναι περίπου 4531 ευρώ.

Στη περίπτωση που τα δεδομένα είναι αταξινόμητα τότε ο τύπος εντοπίζεται στη μεγαλύτερη συχνότητα των δεδομένων.

2.3 Μέτρα διασποράς

Τα μέτρα κεντρικής τάσης από μόνα τους δεν επαρκούν για την επαρκή περιγραφή ενός συνόλου αριθμητικών μετρήσεων. Η αντιπροσωπευτικότητά τους εξαρτάται σε μεγάλο βαθμό από την ετερογένεια που παρουσιάζουν οι μετρήσεις. Στο σχήμα 1εμφανίζονται δύο κατανομές εντελώς διαφορετικές η μία με την άλλη, οι οποίες όμως έχουν την ίδια μέση τιμή, διάμεσο και επικρατούσα τιμή. Είναι προφανές ότι η περιγραφή των δύο αυτών κατανομών, διά μέσου μόνο των μέτρων κεντρικής τάσης, δεν επαρκεί για την εξαγωγή ασφαλών συμπερασμάτων.



Σχήμα 1

Αυτό που διαφοροποιεί τις δύο κατανομές και δεν προσδιορίζεται από τα μέτρα κεντρικής τάσης, είναι η διασπορά των τιμών τους γύρω από το κέντρο τους. Ο προσδιορισμός αυτός γίνεται με τη βοήθεια των μέτρων διασποράς .

Για να γίνει κατανοητό ο λόγος χρήσης των μέτρων διασποράς, παραθέτουμε το ακόλουθο παράδειγμα.

ΠΑΡΑΔΕΙΓΜΑ:

Δυο όμιλοι επιχειρήσεων, που ο καθένας αποτελείται από 10 επιχειρήσεις, είχαν ετήσια έσοδα (σε χιλιάδες ευρώ) για το οικονομικό έτος 2013 τα ποσά που αναγράφονται στον παρακάτω πίνακα:

Πίνακας 5

Όμιλος Α	502	500	496	503	499	500	504	498	501	497
Όμιλος Β	500	498	510	495	500	492	501	497	503	504

Οι μέσες τιμές των εσόδων είναι:

$$\bar{x}_{\text{A}} = \frac{1}{10}\sum_{i=1}^{10} x_{\text{i}} = \frac{502+500+496+503+499+500+504+498+501+497}{10} = 500$$

Και

$$\bar{x}_{\text{B}} = \frac{1}{10}\sum_{i=1}^{10} x_{\text{I}} = \frac{500+498+510+495+500+492+501+497+503+504}{10} = 500$$

Όπως προκύπτει από τις τιμές των δειγμάτων, και οι δύο όμιλοι έχουν μέση τιμή εσόδων 500000 ευρώ. Όμως τα έσοδα των επιχειρήσεων του ομίλου Α είναι 496000 έως 504000, ενώ του ομίλου Β από 492000 μέχρι 510000. Από τα διαστήματα αυτά προκύπτει ότι τα έσοδα των επιχειρήσεων του ομίλου Α βρίσκονται κοντά στη μέση τιμή, ενώ το αντίστοιχο διάστημα για τον όμιλο Β απλώνεται σε τιμές που βρίσκονται πιο μακριά από τη μέση τιμή.

Επομένως είναι φανερό ότι δεν είναι αρκετή η γνώση της μέσης τιμής.

Τα μέτρα διασποράς στοχεύουν στον προσδιορισμό της μεταβλητότητας που παρουσιάζει ένα σύνολο μετρήσεων. Τα μέτρα αυτά χρησιμοποιούνται σε συνδυασμό με τα μέτρα θέσης και από κοινού περιγράφουν τις κατανομές δεδομένων με τρόπο συμπληρωματικό. Τα μέτρα διασποράς είναι :

1) Το εύρος
2) Τα εκατοστημόρια
3) Το ενδοτετρτημοριακό εύρος
4) Μέση απόκλιση
5) Η τυπική απόκλιση και
6) Ο συντελεστής μεταβλητότητας

2.3.1 Εύρος

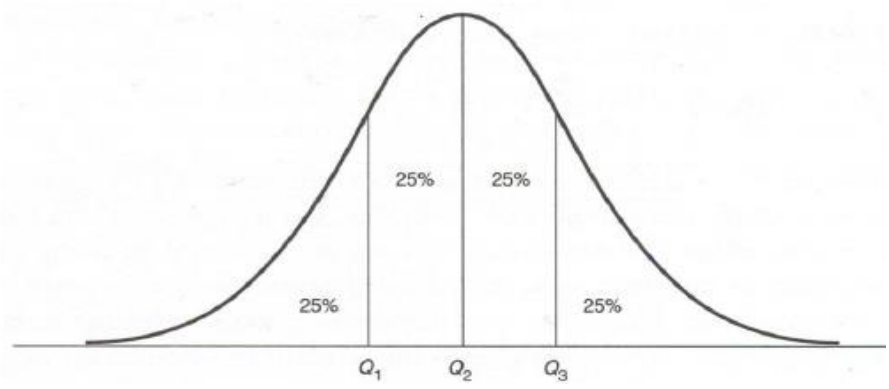
Το εύρος είναι το απλούστερο από όλα τα μέτρα διασποράς και ορίζεται ως η διαφορά μεταξύ μέγιστης και ελάχιστης τιμής ενός συνόλου μετρήσεων.

$$\text{Εύρος } R = \max - \min \quad (16)$$

Αν και το εύρος είναι εύκολο στον προσδιορισμό του, η χρηστικότητά του είναι εξαιρετικά περιορισμένη. Και αυτό διότι στον υπολογισμό του υπεισέρχονται δύο μόνο τιμές, οι πλέον ακραίες. Εξαιτίας αυτού του γεγονότος είναι εξαιρετικά ευαίσθητο στην ύπαρξη τιμών που διαφοροποιούνται πολύ των υπολοίπων, επομένως, ο προσδιορισμός της διασποράς των μετρήσεων διά μέσου του εύρους μπορεί να οδηγήσει σε παραπλανητικά συμπεράσματα.

2.3.2 Ενδοτετρτημοριακό εύρος

Η διαφορά τρίτου και πρώτου τεταρτημορίου $H = Q_3 - Q_1$ ονομάζεται ενδοτεταρτημοριακό εύρος. Το ενδοτεταρτημοριακό εύρος, όπως γίνεται άμεσα αντιληπτό από τον ορισμό του, δεν επηρεάζεται από πιθανές ακραίες τιμές που μπορεί να υπάρχουν στο σύνολο των μετρήσεων (Σχήμα 2).



Σχήμα 2

2.3.3 Μέση απόκλιση

Τα μέτρα διασποράς που μέχρι στιγμής αναφέρθηκαν δεν δίνουν καμία πληροφορία για την διασπορά των δεδομένων γύρω από το μέσο της κατανομής τους. Επιπλέον, κανένα από αυτά δεν λαμβάνει υπόψη κατά τον υπολογισμό του το σύνολο των τιμών της κατανομής

Έστω χ_1 , χ_2 ,…, χ_k οι τιμές μιας μεταβλητής X, που αφορά τα στοιχεία ενός δείγματος μεγέθους ν,k ≤ ν και $\bar{x}$ η μέση τιμή. Η **μέση απόκλιση** ή **μέση απόλυτη απόκλιση** δίνεται από τον τύπο:

$$\text{MAD} = \frac{V1*|X1-\bar{x}|+V2*|X2-\bar{x}|+\cdots+Vk*|Xk-\bar{x}|}{v-1} = \frac{1}{v-1}\sum_{i=1}^{k} vi * |xi-\bar{x}| \quad (17)$$

2.3.4 Διακύμανση και τυπική απόκλιση

Ένας εναλλακτικός τρόπος για να αποφύγουμε το πρόβλημα των πρόσημων στον υπολογισμό των αποκλίσεων $x_i - \bar{x}$, και επιπλέον για να μην εμπλακούμε με τη χρήση των απόλυτων τιμών, είναι να χρησιμοποιήσουμε ως αποστάσεις των τιμών από τη μέση τιμή τους τα τετράγωνα των αποκλίσεων. Η μέση τιμή των τετραγώνων των αποκλίσεων ονομάζεται **διακύμανση.**

Η διακύμανση όταν υπολογίζεται στο σύνολο των στοιχείων ενός πληθυσμού, δίδεται από τον τύπο :

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \quad (18)$$

Όπου x_1, x_2,….., χ_N οι πληθυσμιακές τιμές , μ η πληθυσμιακή μέση τιμή και N το πλήθος των στοιχείων του πληθυσμού.

Σε περίπτωση δειγματικών δεδομένων, ο υπολογισμός της διακύμανσης γίνεται με την βοήθεια του τύπου :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (19)$$

Όπου x_1, x_2,…., x_n οι δειγματικές τιμές , $\bar{x}$ η δειγματική μέση τιμή και n το μέγεθος του δείγματος.

Ο λόγος για τον οποίο, στον τύπο της δειγματικής διακύμανσης αντί της διαίρεσης του αθροίσματος των τετραγώνων των αποκλίσεων διά του n χρησιμοποιείται το n-1, είναι η υποεκτίμηση που προκύπτει για τη διακύμανση του πληθυσμού, όταν αυτή εκτιμηθεί από τα δειγματικά δεδομένα. Η διακύμανση εκφράζει τον ίδιο τύπο πληροφορίας με τη μέση απόκλιση, έχει όμως ορισμένες πολύ σημαντικές ιδιότητες, όπως π.χ. η δυνατότητά της να αναλύεται σε επιμέρους συνιστώσεις, που την κάνουν να υπάρχει σαφώς της μέσης απόκλισης. Ο υπολογισμός της απλουστεύεται όταν χρησιμοποιείται ο τύπος :

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} xi^{\,2} - \frac{\left(\sum_{i=1}^{n} xi\right)}{n}\right] \quad \textbf{(20)}$$

Ο τύπος αυτός προκύπτει εύκολα από τον αρχικό τύπο της διακύμανσης με κατάλληλους μετασχηματισμούς. Επειδή δε κατά τον υπολογισμό του απαιτούνται πολύ λιγότερες πράξεις από αυτές που απαιτεί ο αρχικός τύπος, συνιστάται ως πλέον αποτελεσματικός και λιγότερος χρονοβόρος .

Η τετραγωνική ρίζα της διακύμανσης ονομάζεται **τυπική απόκλιση** και στην πράξη είναι το ευρύτερα χρησιμοποιούμενο μέτρο διασποράς. Η τυπική απόκλιση χρησιμοποιούμενη από κοινού με τη μέση τιμή, μπορούν να περιγράψουν αποτελεσματικά την κατανομή ενός συνόλου μετρήσεων. Η μέση τιμή ορίζει το σημείο γύρω από το οποίο τείνουν να συσσωρεύονται οι τιμές μιας κατανομής, ενώ η τυπική απόκλιση προσδιορίζει το βαθμό της διασποράς των τιμών γύρο από το σημείο αυτό. Ο τύπος της τυπικής απόκλισης είναι :

$$S = \sqrt{s^2} \quad (21)$$

2.3.5 Συντελεστής μεταβλητότητας

Επειδή, η τυπική απόκλιση έχει μονάδες μέτρησης οι οποίες είναι ίδιες με τις μονάδες του μεγέθους στο οποίο αναφέρεται, στερείται νοήματος η σύγκριση τυπικών αποκλίσεων μεταβλητών που μετρώνται σε διαφορετικές μονάδες. Δεν έχει νόημα για παράδειγμα να συγκρίνουμε την τυπική απόκλιση μετρήσεων βάρους και μετρήσεων θερμοκρασίας. Για την σύγκριση της μεταβλητότητας, χρησιμοποιείται ο **συντελεστής μεταβλητότητας**. Ο συντελεστής μεταβλητότητας είναι ένα σχετικό μέτρο διασποράς και εκφράζει την τυπική απόκλιση ενός συνόλου μετρήσεων ως ποσοστό (%) επί της μέσης τιμής και δίνεται από τον τύπο :

$$\text{CV} = \frac{S}{\bar{x}} \quad \bar{x} \neq 0. \quad (22)$$

Συμπερασματικά μπορούμε να πούμε ότι όσο πιο μικρός είναι ο συντελεστής μεταβλητότητας, τόσο μεγαλύτερη ομοιογένεια υπάρχει στις τιμές της μεταβλητής. Γενικά, δεχόμαστε ότι ένα δείγμα θα είναι ομοιογενές, εάν ο συντελεστής μεταβλητότητας δεν είναι μεγαλύτερος του 10%.

Κεφάλαιο 3: Παλινδρόμηση και συσχέτιση

3.1 Διαγράμματα διασποράς

Διάγραμμα διασποράς ονομάζουμε την παράσταση σε ορθογώνιο σύστημα αξόνων όλων των σημείων με συντεταγμένες (x_i , y_i). Τα σημεία αυτά σχηματίζουν ένα ‘νέφος’ ή ‘σμήνος’ σημείων, από την προσεκτική παρατήρηση του οποίου μπορούμε να πάρουμε πληροφορίες για τη σχέση εξάρτησης που ενδεχομένως υπάρχει μεταξύ των μεταβλητών Χ και Υ.

Ο παρακάτω πίνακας δίνει τα ύψη Χ (σε cm) και τα βάρη Υ (σε kg) των 18 αγοριών της Γ' Λυκείου.

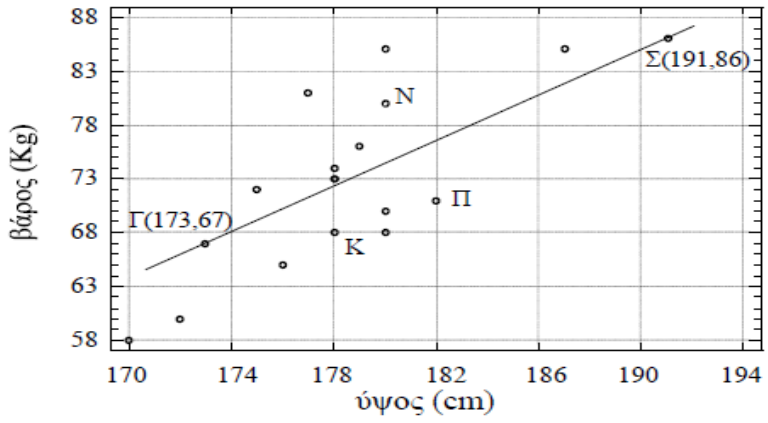
Πίνακας1

Λίστα υψών (σε cm) και βάρος (σε kg) των 18 αγοριών

Μαθητής	**Ύψος X**	**Βάρος Y**	**Μαθητής**	**Ύψος X**	**Βάρος Y**
Α	170	58	Κ	178	68
Β	172	60	Λ	179	76
Γ	173	67	Μ	180	68
Δ	175	72	Ν	180	80
Ε	176	65	Ξ	180	70
Ζ	177	81	Ο	180	85
Η	178	73	Π	182	71
Θ	178	74	Ρ	187	85
Ι	178	73	Σ	191	86

Πηγή :Αναστάσιος Χ. Μπάρλας, Μαθηματικά Γενικής Παιδείας

Στο παράδειγμα αυτό έχουμε την περίπτωση όπου σε κάθε άτομο (μαθητή) γίνονται δύο μετρήσεις. Δηλαδή το δείγμα αποτελείται από τα ζεύγη τιμών των συνεχών μεταβλητών Χ (ύψος) και Υ (βάρος).



Σχήμα 1

Διάγραμμα διασποράς και ευθεία για τα δεδομένα του πίνακα 1

Στο παράδειγμα αυτό παρατηρούμε ότι οι ψηλοί μαθητές είναι συνήθως και πιο βαρείς. Η προσεκτική παρατήρηση ενός διαγράμματος διασποράς μπορεί να μας δώσει σημαντικές πληροφορίες για τη σχέση εξάρτησης που ενδεχομένως υπάρχει μεταξύ των μεταβλητών τις οποίες εξετάζουμε.

3.2 Απλή γραμμική παλινδρόμηση

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

3.2.1 Μέθοδος ελάχιστων τετραγώνων

Από το παραπάνω διάγραμμα διασποράς φαίνεται καθαρά ότι υπάρχει μία σχέση ανάμεσα στο βάρος και στο ύψος των 18 μαθητών. Τα σημεία (x_i , y_i) είναι συγκεντρωμένα περίπου γύρω από μια ευθεία, δηλαδή η σχέση μεταξύ του Χκαι Υ είναι κατά προσέγγιση γραμμική. Η ευθεία που θα προσαρμόζεται καλύτερα στα σημεία αυτά καλείται **ευθεία παλινδρόμησης** της Υ πάνω στη Χ. Όπως γνωρίζουμε, η εξίσωση μιας ευθείας δίνεται από τη σχέση:

$$y = \alpha + bx \qquad (1)$$

όπου α και b είναι παράμετροι τις οποίες θέλουμε να υπολογίσουμε ή όπως λέμε, να "εκτιμήσουμε" , έτσι ώστε η ευθεία που θα προκύψει να μας δίνει όσο το δυνατόν την καλύτερη περιγραφή της σχέσης (εξάρτησης) που υπάρχει μεταξύ των μεταβλητών X και Y.

Η πιο απλή διαδικασία προσαρμογής μιας ευθείας γραμμής σε ένα διάγραμμα διασποράς είμαι με το μάτι. Αυτή όμως έχει πολλά μειονεκτήματα παρά την απλότητά της. Το κυριότερο είναι η έλλειψη αντικειμενικότητας, αφού διάφορα άτομα μπορούν να χαράξουν διαφορετικές μεταξύ τους ευθείες. Χρειαζόμαστε λοιπόν μια ακριβέστερη μέθοδο για την προσαρμογή μιας ευθείας γραμμής σε τέτοιου είδους δεδομένα. Μια μέθοδος που χρησιμοποιείται για την εκτίμηση των παραμέτρων α και b, άρα και για την εύρεση της εξίσωσης της καλύτερης ευθείας που προσαρμόζεται στα δεδομένα, είναι η "μέθοδος ελάχιστων τετραγώνων ".

Η μέθοδος των ελάχιστων τετραγώνων συνίσταται στον προσδιορισμό των παραμέτρων α και b, έτσι ώστε να ελαχιστοποιείται το άθροισμα των τετραγώνων των κατακόρυφων αποστάσεων των σημείων (x_i , y_i),i = 1, 2,…, n από την ευθεία y = α + bx, δηλαδή το

$$\sum_{i=1}^{n}(y_i - a - bx_i)^2 \qquad (2)$$

να γίνεται ελάχιστο.

Οι τιμές των παραμέτρων α και b, που ελαχιστοποιούν την (1), καλούνται **εκτιμήτριες ελάχιστων τετραγώνων**, συμβολίζονται με $\hat{a}$ και $\hat{b}$, αντιστοίχως και αποδεικνύεται ότι δίνονται από τις σχέσεις:

$$\hat{b} = \frac{n\sum_{i=1}^{n} x_i\, y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \qquad (3)$$

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x} \quad (4)$$

Η ευθεία :

$$\hat{y} = \hat{a} + \hat{b}x \quad (5)$$

Καλείται **ευθεία ελάχιστων τετραγώνων** ή **ευθεία παλινδρόμησης** της Υ πάνω στη Χ.

3.2.2 Ερμηνεία των εκτιμητριών ελάχιστων τετραγώνων

Στην εξίσωση ελάχιστων τετραγώνων $\hat{y} = \hat{a} + \hat{b}x$ η τιμή της εκτιμήτριας $\hat{a}$ της παραμέτρου α παριστάνει την τεταγμένη του σημείου στο οποίο η ευθεία τέμνει τον άξονα y΄y, δηλαδή την τιμή της εξαρτημένης μεταβλητής Υ όταν x = 0.

Έστω τώρα δύο τιμές χ_1 και $\chi_2 = \chi_1 + 1$ της ανεξάρτητης μεταβλητής. Τότε, λαμβάνοντας τη διαφορά των αντίστοιχων προβλεπόμενων τιμών της εξαρτημένης μεταβλητής βρίσκουμε:

$$\hat{y}_2 - \hat{y}_1 = \hat{a} + \hat{b}x_2 - (\hat{a} + \hat{b}x_1) = \hat{a} + \hat{b}(x_1 + 1) - (\hat{a} + \hat{b}x_1) = \hat{b}. \quad (6)$$

Δηλαδή :

$$\hat{y}_2 = \hat{y}_1 + \hat{b} \quad (7)$$

Συνεπώς, ο συντελεστής διεύθυνσης $\hat{b}$ της ευθείας $\hat{y} = \hat{a} + \hat{b}x$ παριστάνει τη μεταβολή της εξαρτημένης μεταβλητής Υ όταν το Χ μεταβληθεί κατά μία μονάδα. Έτσι, όταν το Χ αυξηθεί κατά μία μονάδα, τότε το $\hat{y}$ αυξάνεται κατά $\hat{b}$ μονάδες όταν $\hat{b} > 0$ ή ελαττώνεται κατά $\hat{b}$ μονάδες όταν $\hat{b} < 0$.

3.2.3 Συντελεστής γραμμικής συσχέτισης του Pearson

Ο **συντελεστής γραμμικής συσχέτισης** δύο μεταβλητών X και Y ορίζεται με βάση ένα δείγμα n ζευγών παρατηρήσεων (x_i,y_i), i= 1,2,…,n, συμβολίζεται με r(X,Y) ή απλά με r και δίνεται από τον τύπο:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^{2}}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^{2}}} \quad (8)$$

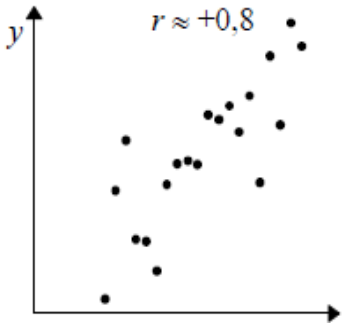
Από τον ορισμό του r παρατηρούμε ότι για μεγάλες τιμές x_i της X και y_i της Y (μεγαλύτερες από τη μέση τιμής τους) οι διαφορές $x_i - \bar{x}$ και $y_i - \bar{y}$ είναι θετικές, οπότε το γινόμενό τους είναι θετικό. Όμοια και για τις μικρές τιμές x_i και y_i οι διαφορές $x_i - \bar{x}$ και $y_i - \bar{y}$ είναι αρνητικές, οπότε το γινόμενό τους είναι πάλι θετικό. Επομένως, όταν σε μεγάλες τιμές της μεταβλητής X αντιστοιχούν και μεγάλες τιμές της Y, ή σε μικρές τιμές της X αντιστοιχούν μικρές τιμές της Y, τότε ο συντελεστής συσχέτισης είναι θετικός και λέμε ότι X και Y είναι θετικός συσχετισμένες.

Ο συντελεστής συσχέτισης είναι καθαρός αριθμός, δηλαδή δεν εκφράζεται σε συγκεκριμένες μονάδες μέτρησης, επομένως είναι ανεξάρτητος των χρησιμοποιούμενων μονάδων μέτρηση των μεταβλητών X και Y. Επιπλέον ισχύει πάντοτε ότι:

$$-1 \leq r \leq 1. \quad (9)$$

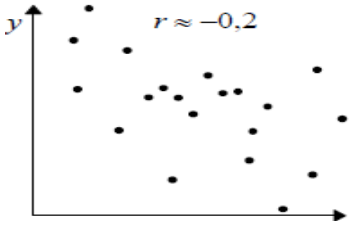
Πιο συγκεκριμένα όταν:

- $0< r < +1$, τότε οι X, Y είναι **θετικά** γραμμικά συσχετισμένες (σχήμα 2. (α)).



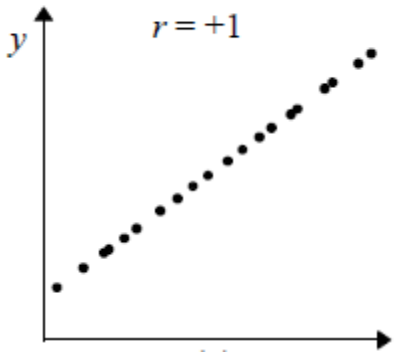
Σχήμα 2 (α) Διάγραμμα διασποράς και συντελεστής συσχέτισης για διάφορα ζεύγη παρατηρήσεων (x_i, y_i)

- -1< r <0, τότε οι X,Y είναι **αρνητικά** γραμμικά συσχετισμένες (σχήμα 2 (β)).



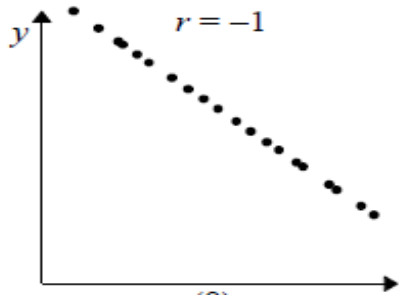
Σχήμα 2 (β) Διάγραμμα διασποράς και συντελεστής συσχέτισης για διάφορα ζεύγη παρατηρήσεων (x_i, y_i)

- r= +1, τότε έχουμε **τέλεια θετική γραμμική συσχέτιση** και όλα τα σημεία βρίσκονται πάνω σε μία θετική κλίση (σχήμα 2 (γ)). Δηλαδή $y = \alpha + \beta_\chi$, $\beta>0$.



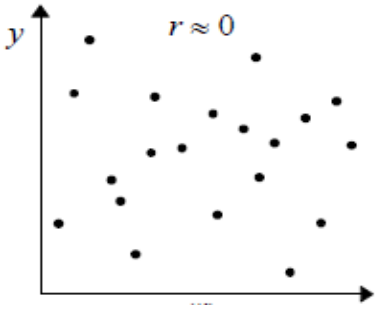
Σχήμα 2 (γ) Διάγραμμα διασποράς και συντελεστής συσχέτισης για διάφορα ζεύγη παρατηρήσεων (x_i, y_i)

- r= -1, τότε έχουμε **τέλεια αρνητική γραμμική συσχέτιση** και όλα τα σημεία βρίσκονται πάνω σε μία αρνητική κλίση (σχήμα 2 (δ)). Δηλαδή y = α +β_χ, β<0.



Σχήμα 2 (δ) Διάγραμμα διασποράς και συντελεστής συσχέτισης για διάφορα ζεύγη παρατηρήσεων (x_i, y_i)

- r =0, τότε δεν υπάρχει γραμμική συσχέτιση μεταξύ των μεταβλητών. Οι μεταβλητές X,Y είναι γραμμικά ασυσχέτιστες (σχήμα 2 (ε)).



Σχήμα 2 (ε) Διάγραμμα διασποράς και συντελεστής συσχέτισης για διάφορα ζεύγη παρατηρήσεων (x_i, y_i)

Αποδεικνύεται ότι ο συντελεστής γραμμικής συσχέτισης r δίνεται ισοδύναμα και από τον παρακάτω τύπο, η χρήση του οποίου διευκολύνει συχνά τους υπολογισμούς τους υπολογισμούς κυρίως στην περίπτωση που οι $\overline{x}, \overline{y}$ δεν είναι ακέραιοι:

$$r= \frac{n\ \sum_{i=1}^{n} x_i\ y_i - \ (\sum_{i=1}^{n} x_i\ )(\sum_{i=1}^{n} y_i\ )}{\sqrt{n\sum_{i=1}^{n} x_i^2\ -(\sum_{i=1}^{n}\ x_i\ )^2}\ \sqrt{n\sum_{i=1}^{n} y_i^2\ -(\sum_{i=1}^{n}\ y_i\ )^2}} \qquad (10)$$

Όταν ο συντελεστής συσχέτισης είναι κοντά στο -1 ή 1, η γραμμική συσχέτιση των μεταβλητών X και Y είναι ισχυρή (συνήθως χαρακτηρίζουμε ισχυρές τις συσχετίσεις όταν $|r| > 0.9$ ενώ όταν είναι κοντά στο 0 οι μεταβλητές X και Yείναι πρακτικά ασυσχέτιστες.

3.2.4 Συντελεστής προσδιορισμού

Στην προηγούμενη παράγραφο μιλήσαμε για το συντελεστή συσχέτισης (= r), ο οποίος μετράει την ένταση (το βαθμό) της εξαρτήσεως μεταξύ των μεταβλητών X και Y, όταν η σχέση εξαρτήσεως είναι γραμμικής μορφής. Θέλουμε τώρα να καθορίσουμε ένα

στατιστικό μέτρο, το οποίο θα μας δείξει όχι μόνο το συνοδευτικό κρίκο μεταξύ παλινδρομήσεως και συσχετίσεως, αλλά θα δώσει μια κατάλληλη ερμηνεία της τιμής των συντελεστών συσχετίσεως. Το στατιστικό αυτό μέτρο ονομάζεται **συντελεστής προσδιορισμού**.

Ένας τρόπος για να αξιολογήσουμε την προσαρμογή της ευθείας των ελάχιστων τετραγώνων είναι να υπολογίσουμε το συντελεστή προσδιορισμού. Ο συντελεστής προσδιορισμού της δειγματικής ευθείας της παλινδρόμησης, συμβολίζεται με το R^2, ορίζεται ως το τετράγωνο του δειγματικού συντελεστή συσχέτισης, δηλαδή:

$$R^2 = r^2 \quad (11)$$

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 των X και Y.

Αποδεικνύεται ότι ισχύει ο τύπος:

$$\mathrm{R}^2 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\hat{\mathrm{y}}_\mathrm{i} - \bar{\mathrm{y}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y}_\mathrm{i} - \bar{\mathrm{y}})^2} \quad (12)$$

Πράγματι, έχουμε:

$$\mathrm{R}^2 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\hat{\mathrm{y}}_\mathrm{i} - \bar{\mathrm{y}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y}_\mathrm{i} - \bar{\mathrm{y}})^2} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\hat{\alpha} + \hat{\mathrm{b}}\mathrm{x}_\mathrm{i} - \hat{\alpha} - \hat{\mathrm{b}}\bar{\mathrm{x}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y}_\mathrm{i} - \bar{\mathrm{y}})^2} =$$

$$\mathrm{b}^2 \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_\mathrm{i} - \bar{\mathrm{x}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y}_\mathrm{i} - \bar{\mathrm{y}})^2} = \left[\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_\mathrm{i} - \bar{\mathrm{x}})(\mathrm{y}_\mathrm{i} - \bar{\mathrm{y}})}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_\mathrm{i} - \bar{\mathrm{x}})^2}\right]^2 \left[\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_\mathrm{i} - \bar{\mathrm{x}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y}_\mathrm{i} - \bar{\mathrm{y}})^2}\right] =$$

$$\frac{[\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_\mathrm{i} - \bar{\mathrm{x}})(\mathrm{y}_\mathrm{i} - \bar{\mathrm{y}})]^2}{[\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_\mathrm{i} - \bar{\mathrm{x}})^2][\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y}_\mathrm{i} - \bar{\mathrm{y}})^2]} = \mathrm{r}^2$$

Ο συντελεστής προσδιορισμού, επομένως, μπορεί να ερμηνευθεί ως ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής το οποίο εξηγείται από την ανεξάρτητη μεταβλητή.

Κεφάλαιο 4: Η στατιστική και ο ρόλος της στη βιομηχανία

Πρώτη φορά που εφαρμόστηκαν στατιστικές μέθοδοι για την αντιμετώπιση προβλημάτων βιομηχανικής παραγωγής ήταν τον προηγούμενο αιώνα.

Συνέπεια αυτού ήταν η διασπορά βιομηχανικών μονάδων σε όλη την έκταση των αναπτυγμένων χωρών . Το τελευταίο διάστημα παρατηρήθηκε το φαινόμενο μεγάλης ανάπτυξης βιομηχανικών μονάδων ,που με βοήθεια της στατιστικής και των μεθόδων της οργανώθηκε καλυτέρα η βιομηχανική παραγωγή.

Η γνώση στατιστικών μεθόδων δεν είναι απολύτως απαραίτητη για ένα μικρό βιομήχανο δεδομένου του μικρού εύρους δραστηριοτήτων που οφείλει να διεκπεραιώσει.

Αν όμως έχουμε να κάνουμε με τον επικεφαλή μιας μεγάλης βιομηχανίας , η χρήση στατιστικής είναι απολύτως αναγκαία ώστε να δρομολογηθούν σωστότερες και αποτελεσματικότερες λύσεις στα προβλήματα παραγωγής που θα προκύψουν.

Με τη χρήση στατιστικών πινάκων, διαγραμμάτων(ιστογράμματα, πολύγωνα, κυκλικά διαγράμματα κ.τ.λ.) καθώς και διαφόρων δεικτών , όπως μεσών τιμών , τυπικών αποκλίσεων συντελεστών μεταβολής , καθίσταται ευκολότερη η λήψη των βέλτιστων αποφάσεων , ως προς την παραγωγή βιομηχανικών προϊόντων , από τον επικεφαλή της εκάστοτε μεγάλης βιομηχανίας.

Οι κυριότερες στατιστικές δραστηριότητες μιας βιομηχανίας επί της παραγωγής είναι οι κάτωθι:

- Το τμήμα παραγωγής
- Το τμήμα προσωπικού
- Το τμήμα στατιστικών μελετών

Στο τμήμα παραγωγής ελέγχεται η υλοποίηση των προδιαγραφών παραγωγής του εκάστου προϊόντος , ώστε να ελεγχθούν και να υπολογιστούν με ακρίβεια οι πιθανές αποκλίσεις επί των αρχικών προδιαγραφών, που δημιουργούνται κατά την παραγωγή ρυθμίζοντας το περιβάλλον παραγωγής, ώστε να περιορισθούν τα προϊόντα με ελαττώματα, μειώνοντας το κόστος ,με απόρροια να αυξάνονται τα κέρδη.

Στο τμήμα προσωπικού καθορίζονται οι αμοιβές, τα όποια επιδόματα, ο χρόνος εργασίας τα bonus, αντιστοιχίζοντας τα με τις ικανότητες του εκάστοτε εργαζομένου.

Στο τμήμα στατιστικών μελετών αναλύονται προβλέψεις ως προς τους διάφορους τομείς της παραγωγής, όπως ο αριθμός των παραγόμενων προϊόντων, οι διαστάσεις τους, το κόστος και τέλος ο συντονισμός των άλλων τμημάτων της βιομηχανίας.

4.1 Εφαρμογές στο χώρο της βιομηχανίας

ΑΣΚΗΣΗ 1

Από την ημερήσια παραγωγή ενός εργοστασίου λαμπτήρων επιλέγονται κατά καιρούς μερικοί και ελέγχονται ως προς την ποιότητά τους. Σε 40 τέτοιους ελέγχους ο αριθμός των λαμπτήρων ήταν:

5	3	1	0	2	2	4	1	0	3	3	2	4	5	1	1	0	2	2	3
2	4	4	1	0	0	1	0	2	2	5	3	3	2	5	1	1	2	3	4

α) Να κατασκευάσετε πίνακα συχνοτήτων αθροιστικών συχνοτήτων και σχετικών συχνοτήτων.

β) Σε πόσους ελέγχους βρέθηκαν 3 ελαττωματικοί λαμπτήρες;

γ) Σε ποιο ποσοστό ελέγχων δεν βρέθηκε κανένας λαμπτήρας ελαττωματικός;

ΛΥΣΗ

α)

χ_i	ν_i	f_i %	N_i
0	6	15	6
1	8	20	14
2	10	25	24
3	7	17,5	31
4	5	12,5	36
5	4	10	40
Άθροισμα	**40**	**100**	----------

Αν διαιρέσουμε τη συχνότητα ν_i με το μέγεθος ν του δείγματος, προκύπτει η σχετική συχνότητα f_iτης τιμής x_i, δηλαδή :

$f_i = \frac{\nu i}{\nu}$, i=1,2,..,κ. Τις σχετικές συχνότητες f_iτις εκφράζουμε επί τοις εκατό, οπότε συμβολίζονται με f_i %.

Οι ποσότητες x_i,ν_i,f_i για ένα δείγμα συγκεντρώνονται σε ένα συνοπτικό πίνακα, που ονομάζεται **πίνακας κατανομής συχνοτήτων** ή απλά **πίνακας συχνοτήτων.**

β) Εφόσον ζητείται ο αριθμός των ελέγχων με τουλάχιστον 3 ελαττωματικούς λαμπτήρες τότε από το σύνολο (40 λαμπτήρες) αφαιρούμε τους ελέγχους με μέχρι και 2 λαμπτήρες ελαττωματικούς (24 λαμπτήρες) δηλαδή :

40 -24 = 16 , λαμπτήρες .

γ) 0 ελαττωματικούς λαμπτήρες είχαμε σε 6 ελέγχους δηλαδή $\frac{6}{40} * 100\% = 40\%$,

των λαμπτήρων .

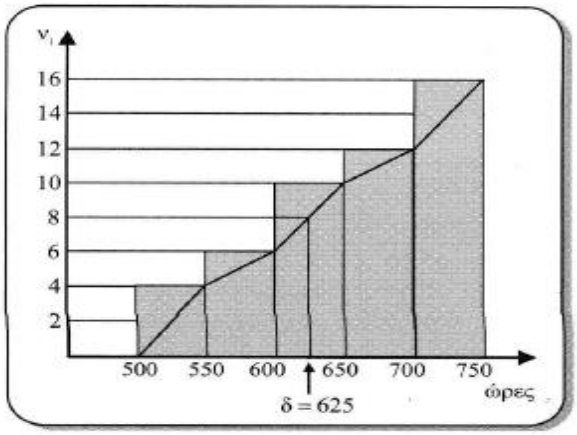
ΑΣΚΗΣΗ 2

Στο παρακάτω ιστόγραμμα αθροιστικών συχνοτήτων αναφερεται στη διάρκεια ζωής δείγματος λαμπτήρων

i) Ποιο το μέγεθος του δείγματος;
ii) Να σχεδιάσετε το πολύγωνο αθροιστικών συχνοτήτων και να υπολογίσετε την διάμεσο.
iii) Να κατασκευάσετε πίνακα συχνοτήτων και σχετικών συχνοτήτων.
iv) Τι ποσοστο λαμπτήρων ‘έζησε ‘ τουλάχιστον 650 ώρες;
v) Να υπολογίσετε τη μέση διάρκεια ζωής των λαμπτήρων.

ΛΥΣΗ

i) Η τελευταία αθροιστική συχνότητα είναι 16 , όσο το ύψος της τελευταίας ράβδου , αρά οι λαμπτήρες είναι 16.
ii) Αφού σχεδιάσουμε το πολύγωνο εντοπίζουμε το 8 στον κατακόρυφο άξονα όσο και ο μισός πληθυσμός. Από το σχήμα φαίνεται ότι οι μισοί λαμπτήρες <<έζησαν>> μέχρι 625 ώρες
δηλαδή δ= 625



ΣΧΟΛΙΟ

Η διάμεσος συνεχούς μεταβλητής βρίσκεται από το πολύγωνο αθροιστικών συχνοτήτων στο 50% του κατακόρυφου άξονα .

iii) Από το ιστόγραμμα έχουμε $N_5 = 16$ $N_4 = 12$ $N_3 = 10$ $N_2 = 6$ $N1 = 4$

$v_1 = N_1 = 4$, $v_2 = N_2 - N_1 = 6 - 4 = 2$, $v_3 = N_3 - N_2 = 10 - 6 = 4$,

$v_4 = N_4 - N_3 = 12\text{-}10$

- $f_1 = \frac{v1}{\nu} = \frac{4}{16} = 0{,}250$
- $f_2 = \frac{v2}{\nu} = \frac{2}{16} = 0{,}125$
- $f_3 = \frac{v3}{\nu} = \frac{4}{16} = 0{,}250$
- $f_4 = \frac{v4}{\nu}$ $f_i = \frac{2}{16} = 0{,}125$
- $f_5 = \frac{v5}{\nu}$ $f_i = \frac{4}{16} = 0{,}250$

Κλάσεις [-)	ν_ι	f_i
500-550	4	0.250
550-600	2	0.125
600-650	4	0.250
650-700	2	0.125
700-750	4	0.250
Σύνολο	**16**	**1**

iv) $F4\% + f_5\% = 100f_4 + 100f_5 = 12,5\% + 25\% = 37,5\%$

vi) Ξαναγράφουμε τον πινάκα με τις κεντρικές τιμές και ην στήλη $\chi_{i*}\ \nu_i$.Η μέση διάρκεια ζωής των λαμπτήρων είναι $\overline{x} = \frac{\Sigma xi*\nu i}{n} = \frac{10000}{16} = 625$ ώρες.

Κλάσεις [-)	χ_ι	ν_i	$\chi_\iota \nu_i$
500-550	525	4	2.100
550-600	575	2	1.15
600-650	625	4	2.500
650-700	675	2	1.350
700-750	725	4	2.900
Σύνολο	-	16	10.000

ΑΣΚΗΣΗ 3

Σε μια βιομηχανία γάλακτος εξετάζονται καθημερινά η ποιότητα του προϊόντος .Τα αποτελέσματα των τελευταίων ημερών 25 φαίνονται στο παρακάτω πινάκα.

α) Να βρεθεί ο μέσος Όρος ελαττωματικών μονάδων προϊόντος.

β) Να βρεθεί η διάμεσος.

x_i	v_i	$f_i\%$
0	7	28
1	7	28
2	6	24
3	4	16
4	1	4
Άθροισμα	**25**	**100**

ΛΥΣΗ

α)

x_i	v_i	f_i%	$x_i v_i$
0	7	28	0
1	7	28	7
2	6	24	12
3	4	16	12
4	1	4	4
Άθροισμα	**25**	**100**	**35**

$\bar{x}=\frac{\Sigma xi*vi}{n}=>\quad \bar{x}=\frac{0*7+1*7+2*6+3*4+4*1}{25}=>\ \bar{x}=\frac{35}{25}=>\bar{x}=1,$

β) Δεδομένου ότι ν =25 , τότε $\frac{\nu}{2}=\frac{25}{2}$=12,5. => δ = 1

ΑΣΚΗΣΗ 4

Μια επιχείρηση έχει προς ενοικίαση αυτοκίνητα με μέσο χρόνο λειτουργίας πριν την πρώτη βλάβη να είναι 12 μήνες , με τυπική απόκλιση 3 μήνες.

α) Να αποδείξετε ότι το δείγμα δεν είναι ομοιογενές.

β) Αν η επιχείρηση φροντίσει να μεγαλώσει το χρόνο λειτουργίας κάθε αυτοκινήτου πριν την πρώτη βλάβη κατά c μήνες ώστε το δείγμα να είναι ομοιογενές να βρείτε το c.

ΛΥΣΗ

α) Έχουμε $\bar{x}$ = 12 μήνες και s = 3μήνες . Άρα ο CV = $\frac{s}{\bar{x}}=\frac{3}{12}$= 0,25= 25% >10% δηλαδή το δείγμα δεν είναι ομοιογενές.

β) Αυξάνοντας το χρόνο λειτουργίας κατά c μήνες πριν την $1^η$ βλάβη, έχουμε:

$\bar{x}$ =12 + c , s = 3 ⇔ CV = $\frac{s}{\bar{x}+c}$= CV = $\frac{3}{12+c}$, με το δείγμα να είναι ομοιογενές αν

CV ≤ 0,1 ⇔ $\frac{3}{12+c} \leq \frac{1}{10}$ ⇔ 30 ≤ 12+c ⇔ 18 ≤ c.

ΑΣΚΗΣΗ 5

Μια βιομηχανία παρασκευάζει 4 είδη σοκολάτας, σε ίσες ποσότητες καθαρού βάρους 60,70 , 100,150gr.

α) Να βρείτε το μέσο βάρος και τον συντελεστή μεταβολής

β) Αν το βάρος κάθε σοκολάτας αυξηθεί κατά 10% , να βρεθούν το νέο μέσο βάρος, ο νέος συντελεστής μεταβολής.

ΛΥΣΗ

α) Για το μέσο βάρος ισχύει:

$\bar{x} = \frac{\sum_{i=1}^{\nu} xi}{\nu} = \frac{60+70+100+150}{4} = \frac{380}{4} = 95,$

$s^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}(x_i - \bar{x})^2 = \frac{(60-95)2+(70-95)2+(100-95)2+(150-95)2}{4} = \frac{4900}{4} = 1225$

Άρα s = 35, CV = $\frac{35}{95}$= 0,3684>0,1.

β) Αν το βάρος της κάθε σοκολάτας αυξηθεί κατά 10% , τότε

$\psi_i = \chi_i + 0{,}1\,\chi_{i,,}$

$\bar{\psi} = \frac{\sum_{i=1}^{\nu} \psi i}{\nu} = \frac{\sum_{i=1}^{\nu} xi + 0{,}1\chi i}{\nu} = \bar{x} + 0{,}1\bar{x} = 104{,}5$ gr.

$s^2_{\psi} = \frac{1}{\nu}\sum_{i=1}^{\nu}(\psi_i - \bar{\psi})^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}(1.1(x_i - \bar{x}))^2 = 1{,}21 * s^2 = 1482{,}25$ ⇔ s = 38,5gr.

CV = $\frac{38,5}{104,5}$= 0.3684.

ΑΣΚΗΣΗ 6

Από μια έρευνα που έγινε σχετικά με τους μισθούς των εργατών

Μιας επιχείρησης προέκυψε ότι το 2,5% αυτών έχουν μηνιαίο μισθό μικρότερο των 500 ευρώ, το 84% μισθό μικρότερο των 800 ευρώ. Αν η κατανομή μισθών ακολουθεί περίπου την κανονική κατανομή τότε:

α) Να βρείτε την μέση τιμή και την τυπική απόκλιση των μισθών.

β) Να εξετασθεί αν το δείγμα είναι ομοιογενές

γ)Αν η επιχείρηση απασχολεί 4000 εργάτες να βρείτε:

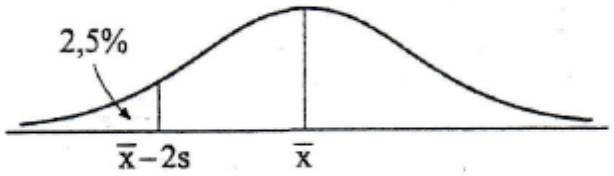
i) Πόσοι εργάτες παίρνουν μισθό από 500 ως 800 ευρώ

ii) Πόσοι εργάτες έχουν μισθό μεγαλύτερο από 900.

ΛΥΣΗ

α)

- **Σ**ε μια κανονική κατανομή το 95% περίπου των παρατηρήσεων βρίσκεται στο διάστημα ($\bar{x}$-2s, $\bar{x}$+2s) . Επομένως εκτός του διαστήματος αυτού βρίσκεται το 5% (2,5% μικρότερες από το κάτω όριο και 2,5% μεγαλύτερες από το μεγαλύτερο . επειδή το 2,5% έχει μισθό μικρότερο των 500 ευρώ έχουμε : $\bar{x}$ -2s = 500 (1).



- Σε μια κανονική κατανομή το 84% περίπου των παρατηρήσεων είναι μικρότερες από $\bar{x}$+s. Επειδή το 84% των εργατών έχει μισθό το πολύ 800 ευρώ έχουμε:

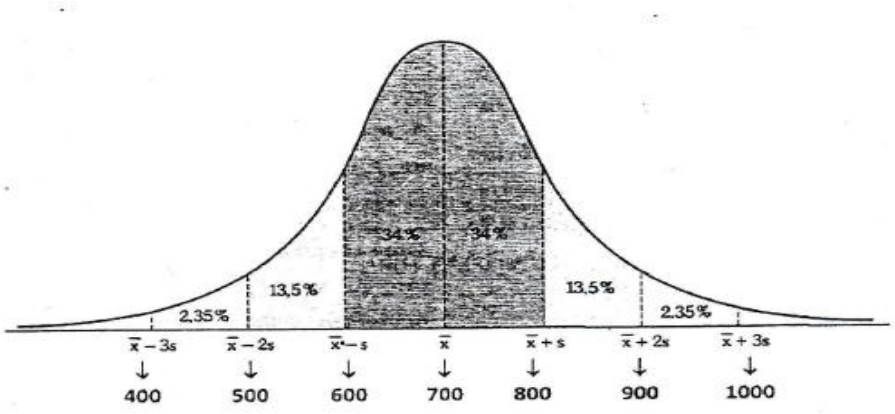
  $\bar{x}$+s=800 (2) => $\bar{x}$=800-s (1)=> 800-s -2s= 500 => 3s = 300 => s = 100 => $\bar{x}$ = 700

β) Για να βρούμε αν το δείγμα είναι ομοιογενές υπολογίζουμε το

CV= $\frac{S}{\bar{x}}$, CV = $\frac{100}{700}$, CV = 0,143 ή CV = 14,3% > 10% . Άρα το δείγμα είναι ανομοιογενές .

γ) i) Αν η επιχείρηση απασχολεί 400 εργάτες τότε το ποσοστό των εργατών με μισθό από 500 έως 800 ευρώ είναι 13,5% +34% +34% = 81,5%,με πλήθος 81,5%*400= 326, όπως φαίνεται από το παρακάτω διάγραμμα.

ii) Το ποσοστό εργατών με μισθό μεγαλύτερο από 900 ευρώ είναι $\frac{2,5\%}{2}$, με πλήθος $\frac{2,5\%}{2}$*400 = 5.



ΑΣΚΗΣΗ 7

Οι βαθμοί 100 μαθητών μιας τάξης ενός λυκείου που ομαδοποιήθηκαν σε 4 κλάσεις ίσου πλάτους δίνονται στον παρακάτω πίνακα:

Κλάσεις	f_i%
[0, 4)	35
[4, 8)	
[8, 12)	30
[12, 16)	

Αν ο μέσος όρος βαθμολογίας των μαθητών είναι 7

α) Να υπολογίσετε τις συχνότητες f_2 % και f_4% που λείπουν

β) Να υπολογίσετε την τυπική απόκλιση

γ) Να εξετάσετε το δείγμα ως προς την ομοιογένεια

δ) Αν ανέβει η βαθμολογία όλων των μαθητών κατά 1 μονάδα πόσο θα είναι η μέση βαθμολογία τους

ΛΥΣΗ

α) Επειδή ν =100 και f_1 % = 35% , f_1 = 0,35 δηλαδή ν_1= f_1* ν

ν_1= 35. Ομοίως ν_3= f_3*ν => ν_3 = 0,30*100 => ν_3 = 30 .

Όμως ν = ν_1 + ν_2 + ν_3 + ν_4 , έχουμε 35 + ν_2 + 30 +ν_4 = 100

ν_2 + ν_4 = 35 (1) , επίσης

$\bar{x} = 7 \Rightarrow \bar{x} = \frac{2*35+6*\nu 2+10*30+14*\nu 4}{100} = 7 \Leftrightarrow 70 + 6\nu_2 + 300 + 14\nu_4 = 700 \Leftrightarrow 6\nu_2 + 14\nu_4 = 330$ (2) .

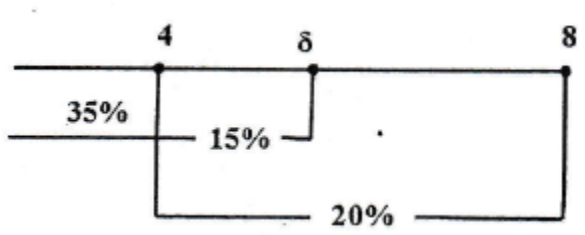
Από τη σχέση (1) έχουμε $\nu_2 = 35 - \nu_4$, (2) $\Leftrightarrow 6(35 - \nu_4) + 14\nu_4 = 330$

$\Leftrightarrow 210 - 6\nu_4 + 14\nu_4 = 330 \Leftrightarrow 8\nu_4 = 120 \Leftrightarrow \nu_4 = 15$, $\nu_2 = 35-15$

$\Leftrightarrow \nu_2 = 20$. Άρα $f_2\% = 15\%$, $f_4\% = 20\%$.

Κλάσεις	x_i	ν_i	f_i	$f_i\%$	$F_i\%$	$x_i * \nu_i$	x_i^2	$x_i^2 * \nu_i$
[0,4)	2	35	0,35	35	35	70	4	140
[4,8)	6	20	0,20	20	55	120	36	720
[8,12)	10	30	0,30	30	85	300	100	3000
[12,16)	14	15	0,15	15	100	210	196	2940
Σύνολο		**100**	**1,00**	**100**		**700**	**336**	**6800**

β) Η διάμεσος , αντιστοιχίζεται στην τιμή $\chi = \delta$ της μεταβλητής Χ , ώστε το 50% των παρατηρήσεων να είναι μικρότερες ή ίσες αυτής . Από τον πίνακα παρατηρούμε ότι βρίσκεται στην 2η κλάση , ώστε το διάστημα από το 4 ως την δ να ανήκει το 15% των παρατηρήσεων σχετικά παρατίθεται το εξής



$\frac{\delta-4}{8-4} = \frac{15\%}{20\%} \Leftrightarrow \frac{\delta-4}{8-4} = \frac{15}{20} \Leftrightarrow \delta - 4 = 3 \Leftrightarrow \delta = 7$.

Για τον υπολογισμό της τυπικής απόκλισης έχουμε:

$s^2 = \frac{1}{n}[\sum_{i=1}^{n} xi^{2*} \nu i \ - \frac{(\sum_{i=1}^{n} xi*\nu i)^2}{n}] = \frac{1}{100}[6800 - \frac{700^2}{100}] = 19$, $s = \sqrt{s^2}$

$s = \sqrt{19}$ <=> s = 4,36 .

γ) Για την ομοιογένεια του δείγματος θα βρούμε το συντελεστή μεταβλητότητας

CV= $\frac{S}{\bar{x}} = \frac{4,36}{7}$ = 0,623 ή 62,3% > 10% , επομένως το δείγμα είναι

ανομοιογενές.

δ) Αν ο καθηγητής ανεβάσει τη βαθμολογία όλων των μαθητών 1 μονάδα

τότε θα ανέβει κατά 1 μονάδα.

ΑΣΚΗΣΗ 8

Οι ν τιμές t_1 , t_2,…, t_v , μεταβλητής X ακολουθούν την κανονική , εύρους R ≈ 6 , το 2,5% αυτών είναι μικρότερες από

Το 10 . Επιπλέον $\sum_{i=1}^{n}(ti)^2$= 58000 , τότε :

α) Να βρείτε το συντελεστή μεταβολής των τιμών του δείγματος και να εξετάσετε αν το δείγμα είναι ομοιογενές.

β) Να βρείτε το πλήθος των παρατηρήσεων.

γ) Να βρείτε το πλήθος των παρατηρήσεων στο διάστημα (13,14)

δ) Αν οι τιμές αυξηθούν 12 μονάδες , να δείξετε ότι ο συντελεστής μεταβολής μένει ο μισός του αρχικού.

ΛΥΣΗ

α) Γνωρίζουμε ότι το εύρος R στην κανονική κατανομή είναι περίπου

6*s ⇔ 6*s= 6 ⇔ s = 1 . Επίσης το 95% των παρατηρήσεων βρίσκονται στο διάστημα

$(\bar{x} - 2s, \bar{x} + 2s)$, οπότε λόγω συμμετρίας της κατανομής το 2,5% των παρατηρήσεων είναι μικρότερες του $\bar{x} - 2s$.

$\bar{x} - 2s$= 10 ⇔ $\bar{x} - 2$= 10 ⇔$\bar{x} = 12$. CV = $\frac{S}{\bar{x}} = \frac{1}{12} < \frac{1}{10}$, το δείγμα ομοιογενές.

β) Μετασχηματίζοντας τον τύπο $s^2 = \frac{1}{\nu}[\sum_{i=1}^{\nu} ti^2 - \frac{(\sum_{i=1}^{\nu} ti)}{\nu}]$,παίρνουμε :

$s^2 = \frac{1}{\nu}\sum_{i=1}^{n} ti^2 - [\frac{(\sum_{i=1}^{\nu} ti)}{\nu}]^2$ ⇔ $s^2 = \frac{1}{\nu}\sum_{i=1}^{\nu} ti^2 - (\bar{x})^2$ ⇔ $\nu = \frac{\sum_{i=1}^{n} ti2}{s\,2+(\bar{x})2}$

$\nu = \frac{58000}{1+144}$ = 400.

γ) Το διάστημα (13,14) είναι το $(\bar{x} + s, \bar{x} + 2s)$. Επίσης το 95% των παρατηρήσεων βρίσκονται στο διάστημα $(\bar{x} - 2s, \bar{x} + 2s)$ και το 68% στο διάστημα $(\bar{x} - s, \bar{x} + s)$ οπότε λόγω συμμετρίας θα έχουμε το $\frac{95\%-68\%}{2}$ =13,5% των παρατηρήσεων να βρίσκεται στο $(\bar{x} + s, \bar{x} + 2s)$.

Τελικά στο διάστημα (13,14) είναι το $\frac{13,5}{100}$*400 = 54 παρατηρήσεις .

Αν οι τιμές αυξηθούν κατά 12 μονάδες τότε $\bar{x}'$= $\bar{x}$+12 =24 και η νέα τυπική απόκλιση

s΄ = s= 1 . Τελικά ο νέος συντελεστής μεταβολής θα είναι CV΄= $\frac{S'}{\bar{x}'} = \frac{1}{24} = \frac{CV}{2}$.

ΑΣΚΗΣΗ 9

Επιχείρηση παραγωγής μηχανημάτων εκτιμά ότι η πιθανότητα ένα μηχάνημα να επιστραφεί για αντικατάσταση λόγω ελαττώματος είναι 5%. Η επιχείρηση παρήγαγε 20 μηχανήματα τον προηγούμενο μήνα.

i. Ποια η πιθανότητα κανένα να μην χρειαστεί αντικατάσταση;

ii. Ποια η πιθανότητα να χρειαστούν αντικατάσταση το πολύ 4 μηχανήματα;

iii. Ποιος είναι ο αναμενόμενος αριθμός των μηχανημάτων που θα χρειαστούν αντικατάσταση;

ΛΥΣΗ

Το παραπάνω πείραμα είναι πείραμα Bernoulli με πιθανότητα επιτυχίας p=0,05.
Αν συμβολίσουμε με X το συνολικό αριθμό επιτυχιών στις 20 (ανεξάρτητες) επαναλήψεις του πειράματος, δηλαδή στη παραγωγή 20 μηχανημάτων, τότε είναι προφανές ότι η τυχαία μεταβλητή X ακολουθεί τη διωνυμική κατανομή με παραμέτρους n = 20 και p = 0,05 δηλαδή, X~B(20 , 0,05). Έχουμε λοιπόν:

i) $$P(X=0)=\frac{20!}{0!(20-0)!}(0,05)^0(1-0,05)^{20-0}=\frac{20!}{1*20!}(0,05)^0(0,95)^{20}\cong 0,3585$$

ή χρησιμοποιούμε τους πίνακες διωνυμικής κατανομής

ii) $$P(X\le 4)=P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4)$$

επειδή

$$P(X=1)=\frac{20!}{1!(20-1)!}(0,05)^1(1-0,05)^{20-1}=\frac{20!}{1*19!}(0,05)^1(0,95)^{19}\cong 0,3774$$

$$P(X=2)=\frac{20!}{2!(20-2)!}(0,05)^2(1-0,05)^{20-2}=\frac{20!}{2*18!}(0,05)^2(0,95)^{18}\cong 0,1887$$

$$P(X=3)=\frac{20!}{3!(20-3)!}(0,05)^3(1-0,05)^{20-3}=\frac{20!}{6*17!}(0,05)^3(0,95)^{17}\cong 0,0596$$

$$P(X=4)=\frac{20!}{4!(20-4)!}(0,05)^4(1-0,05)^{20-4}=\frac{20!}{24*16!}(0,05)^4(0,95)^{16}\cong 0,0133$$

είναι P(X≤4)=0,3585+0,3774+0,1887+0,0596+0,0133=0,9974

ή χρησιμοποιούμε τους πίνακες διωνυμικής κατανομής που δίνουν απευθείας την πιθανότητα P(X≤4)=0,9974

iii) $E(X)=np=20*0,05=1$ μηχανήματα

ΑΣΚΗΣΗ 10

Μια επιχείρηση που παράγει ηλεκτρικά καλώδια γνωρίζει πως ο αριθμός των ελαττωμάτων σε κάποιο συγκεκριμένο τύπο καλωδίων ακολουθεί την κατανομή Poisson με μέσο πλήθος ελαττωμάτων 4 ανά 100 μέτρα καλωδίου.

i. Ποια είναι η πιθανότητα να βρει κανείς ακριβώς 5 ελαττώματα σε ένα καλώδιο 50 μέτρων;
ii. Να υπολογισθεί ο συντελεστής μεταβλητότητας του αριθμού των ελαττωμάτων που εμφανίζονται σε καλώδιο 50 μέτρων, καθώς και σε καλώδιο 100 μέτρων. Τι συμπεραίνετε;
iii. Ποια είναι η πιθανότητα να βρει κάποιος 2 ελαττώματα, σε ένα καλώδιο 100 μέτρων, ξέροντας πως υπάρχει τουλάχιστον ένα ελάττωμα;
iv. Ποια η πιθανότητα σε δυο διαφορετικά καλώδια 50 μέτρων το καθένα να μην εμφανιστεί ελάττωμα.

<u>ΛΥΣΗ</u>

Έστω η τ.μ. X: αριθμός ελαττωμάτων σε καλώδιο 100 μέτρων.

Γνωρίζουμε ότι X ~ Poisson (4).

i) Θεωρούμε την τ.μ. Y: αριθμός ελαττωμάτων σε καλώδιο 50 μέτρων.

Θα ισχύει ότι Y~ Poisson ($\lambda' = \frac{1}{2}\,\lambda = \frac{1}{2}4 = 2$).

Η ζητούμενη πιθανότητα υπολογίζεται ως εξής:

$$\text{P(y=5)} = \frac{e^{-\lambda}\,\lambda^5}{5!} = \frac{e^{-2}\,2^5}{5!} \Rightarrow P(y=5) = 0{,}0360894.$$

ii) Γνωρίζω ότι στην Poisson ισχύει : $\mu = \sigma^2 = \lambda$. Στην περίπτωση των καλωδίων 50 μέτρων έχω λ=2, και επομένως CV = $\frac{\sigma}{\mu} = \frac{\sqrt{2}}{2} \approx 0{,}7$, ενώ στην περίπτωση των καλωδίων 100 μέτρων λ=4.

Επομένως CV = $\frac{\sigma}{\mu} = \frac{2}{4} = 0{,}5$.

Στην πρώτη περίπτωση έχω μεγαλύτερη μεταβλητότητα.

iii) Είμαστε στην περίπτωση όπου μας ενδιαφέρει ο αριθμός των ελαττωμάτων σε καλώδιο 100 μέτρων, δηλ. $X \sim \text{Poisson}(4)$. Η ζητούμενη πιθανότητα είναι η $P(x=2|x \geq 1)$ η οποία μπορεί να υπολογισθεί χρησιμοποιώντας τον ορισμό της δεσμευμένης πιθανότητας:

$$P(x=2|x \geq 1)=\frac{P(x=2, x \geq 1)}{P(x \geq 1)} = \frac{P(x=2)}{P(x \geq 1)} = \frac{P(x=2)}{1-P(x=0)} = \frac{e^{-\lambda}\frac{\lambda^2}{2!}}{1-e^{-\lambda}\frac{\lambda^0}{0!}} = \frac{e^{-4}\frac{4^2}{2!}}{1-e^{-4}} = \frac{8}{e^4-1} =$$

$$0{,}149259$$

iv) Τα ενδεχόμενα να υπάρχει ελάττωμα σε δυο διαφορετικά καλώδια 50 μέτρων είναι μεταξύ τους ανεξάρτητα. Άρα η ζητούμενη πιθανότητα ισούται με:

$$P(x=0)P(x=0) = e^{-2} * e^{-2} = 0{,}018$$

ΑΣΚΗΣΗ 11

Μια εταιρεία επιθυμεί να εκτιμήσει το αναμενόμενο κόστος παραγωγής όταν γνωρίζει το επίπεδο παραγωγής. Για το λόγο αυτό κατέγραψε από τα αρχεία της το επίπεδο παραγωγής και το αντίστοιχο κόστος παραγωγής ανά εβδομάδα, για 16 εβδομάδες που επιλέχθηκαν τυχαία από τους προηγούμενους έξι μήνες. Το κόστος (σε ευρώ) και το επίπεδο παραγωγής (σε τεμάχια) δίνονται στον πίνακα που ακολουθεί.

Επίπεδα Παραγωγής	**Κόστος (σε χιλ. Ευρώ)**

(X)	(Y)
114	2637
101	2177
84	1920
94	1910
98	1984
97	1787
77	1689
92	1866
96	1896
81	1684
79	1633
80	1657
77	1569
64	1390
64	1387
56	1289

i. Να κατασκευασθεί στο Excel το διάγραμμα διασποράς των παραπάνω ζευγών τιμών (X,Y) και με βάση αυτό να εξετασθεί η γραμμική σχέση του επιπέδου και του κόστους παραγωγής.

ii. Δεδομένου ότι η σχέση μεταξύ των μεταβλητών X και Y είναι γραμμική να εκτιμηθούν οι συντελεστές a_0 και a_1 του γραμμικού υποδείγματος $Y = a_0 + a_1 X + \varepsilon$ με την μέθοδο των ελαχίστων τετραγώνων και να ερμηνευθούν. Οι υπολογισμοί θα πρέπει να γραφούν αναλυτικά και να γίνουν με δύο τρόπους, δηλαδή τόσο με τη χρήση των τύπων που περιλαμβάνουν τις μεταβλητές σε αποκλίσεις από τους μέσους τους, όσο και με τη χρήση των τύπων που δεν απαιτούν να εκφρασθούν οι μεταβλητές σε αποκλίσεις από τους μέσους τους. Για το σκοπό αυτό να δημιουργηθεί στο Excel κατάλληλος πίνακας στον οποίο να γίνουν οι προαπαιτούμενοι υπολογισμοί.

iii. Να υπολογισθεί ο συντελεστής συσχέτισης και να ερμηνευθεί.

iv. Στο παραπάνω γραμμικό υπόδειγμα ($Y = a_0 + a_1 X + \varepsilon$) να υπολογισθεί ο συντελεστής προσδιορισμού R^2 και να ερμηνευθεί. Ο υπολογισμός να γίνει και με το Excel, με τη βοήθεια των στοιχείων του πίνακα που κατασκευάσατε στο ερώτημα ii.

v. Να κατασκευάσετε ξανά το διάγραμμα διασποράς του υπό-ερωτήματος i προσθέτοντας την γραμμή παλινδρόμησης που εκτιμήσατε στο ερώτημα ii.

vi. Να υπολογίσετε στο excel όλες τις προβλεπόμενες τιμές και να εντοπίσετε τις περιπτώσεις με τη μεγαλύτερη θετική και αρνητική απόκλιση μεταξύ προβλεπόμενων και παρατηρούμενων τιμών. Τι σημαίνει αυτό για την εταιρεία;

vii. Με βάση το παραπάνω μοντέλο, να υπολογίσετε το αναμενόμενο κόστος της επιχείρησης σε επίπεδο παραγωγής 50 προϊόντων.

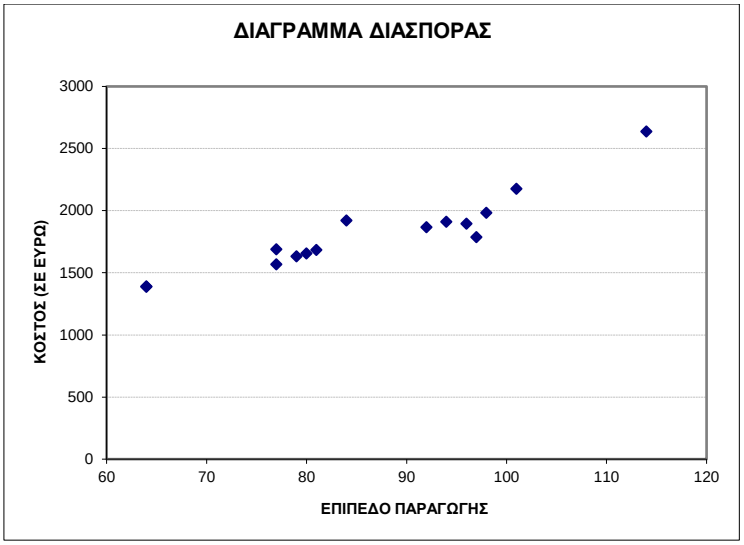
viii. Ας υποθέσουμε ότι ενδιαφερόμαστε και για την εκτίμηση των παραμέτρων b_0 και b_1 του γραμμικού υποδείγματος $X = b_0 + b_1 Y + u$. Να εκτιμηθούν οι συντελεστές b_0, b_1 με την μέθοδο των ελαχίστων τετραγώνων. Οι υπολογισμοί θα πρέπει να γραφούν αναλυτικά και να γίνουν με τη χρήση των τύπων που απαιτούν να εκφρασθούν οι μεταβλητές σε αποκλίσεις από τους μέσους τους.

Για το σκοπό αυτό να συμπληρωθεί ο Πίνακας του Excel που κατασκευάσατε στο ερώτημα ii ώστε να γίνουν οι προαπαιτούμενοι υπολογισμοί.

ix. Να επαληθεύσετε ότι ικανοποιείται η απλή σχέση $a_1^2 S_x^2 = R^2 S_y^2$ για οποιαδήποτε δεδομένα X και Y .

Λ Υ Σ Η

i) Το Διάγραμμα Διασποράς εμφανίζεται στο Αρχείο Excel. Από αυτό προκύπτει καταρχήν μία θετική σχέση μεταξύ των δύο μεταβλητών. Επίσης, από το διάγραμμα φαίνεται ότι η σχέση των δυο μεταβλητών είναι γραμμική και ισχυρή.



ii) Όπως ήδη αναφέρθηκε στο ερώτημα a η σχέση των δυο μεταβλητών είναι γραμμική και ισχυρή. Κατά συνέπεια μπορούμε να προχωρήσουμε στο υπολογισμό των συντελεστών της γραμμικής εξίσωσης $Y = a_0 + a_1 X + \varepsilon$ με την μέθοδο των ελαχίστων τετραγώνων.

Κατά συνέπεια :

<u>Α' Τρόπος (με αποκλίσεις από τους μέσους)</u>

Σύμφωνα με τον πρώτο τρόπο πρέπει να χρησιμοποιήσουμε τους τύπους

$$\alpha_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a_0 = \bar{y} - a_1\bar{x}$$

Από τα στοιχεία μας μπορούμε να υπολογίσουμε τους αριθμητικούς μέσους των μεταβλητών *X* και *Y* οι οποίοι είναι

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1354}{16} = 84{,}625 \text{ και } \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{28475}{16} = 1779{,}6875$$

Για τη διευκόλυνση των πράξεων σχηματίζουμε τον παρακάτω Πίνακα.

Πίνακας 1

$\mathbf{i}$	$\mathbf{x_i}$	$\mathbf{y_i}$	$\mathbf{x_i - \bar{x}}$	$\mathbf{(x_i - \bar{x})^2}$	$\mathbf{(x_i - \bar{x})(y_i - \bar{y})}$
1	114	2637	29,38	862,89	25183,55
2	101	2177	16,38	268,14	6505,99
3	84	1920	-0,63	0,39	-87,70
4	94	1910	9,38	87,89	1221,68
5	98	1984	13,38	178,89	2732,68
6	97	1787	12,38	153,14	90,49
7	77	1689	-7,63	58,14	691,49
8	92	1866	7,38	54,39	636,55
9	96	1896	11,38	129,39	1323,05
10	81	1684	-3,63	13,14	346,87
11	79	1633	-5,63	31,64	825,12
12	80	1657	-4,63	21,39	567,43
13	77	1569	-7,63	58,14	1606,49

14	64	1390	-20,63	425,39	8037,30
15	64	1387	-20,63	425,39	8099,18
16	56	1289	-28,63	819,39	14045,93
ΑΘΡΟΙΣΜΑ	**1354**	**28475**	**0,00**	**3587,75**	**71826,125**

Κατά συνέπεια έχουμε ότι:

N=16,

$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ = 71826,125

$\sum_{i=1}^{n}(x_i - \bar{x})^2$ = 3587,75

Επομένως,

$$a_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{71826,175}{3587,75} = 20,02$$

Για δε την εκτίμηση της σταθεράς έχουμε:

$$a_0 = \bar{y} - a_1\bar{x} = 1779,6875 - 20,02 * 84,625 = 85,5$$

Άρα η εξίσωση παλινδρόμησης είναι: $\hat{y} = 85,5 + 20X$

<u>Β' Τρόπος (χωρίς τις αποκλίσεις από τους μέσους)</u>

Ο εναλλακτικός τύπος υπολογισμού του συντελεστή a_1 που δεν απαιτεί να εκφρασθεί το Υ σε αποκλίσεις από το μέσο του είναι

$$a_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Συμπληρώνουμε τον Πίνακα 6.1 και δημιουργούμε τις απαραίτητες στήλες για τον υπολογισμό του a_1 με τον εναλλακτικό τρόπο. Έτσι δημιουργείται ο παρακάτω Πίνακας:

Πίνακας 2

$\mathbf{i}$	$\mathbf{x_i}$	$\mathbf{y_i}$	x_i^2	$\mathbf{x_i y_i}$
1	114	2637	12996	300618
2	101	2177	10201	219877
3	84	1920	7056	161280
4	94	1910	8836	179540
5	98	1984	9604	194432
6	97	1787	9409	173339
7	77	1689	5929	130053
8	92	1866	8464	171672
9	96	1896	9216	182016
10	81	1684	6561	136404
11	79	1633	6241	129007
12	80	1657	6400	132560
13	77	1569	5929	120813

14	64	1390	4096	88960
15	64	1387	4096	88768
16	56	1289	3136	72184
ΑΘΡΟΙΣΜΑ	**1354**	**28475**	**118170**	**2481523**

Κατά συνέπεια έχουμε ότι:

$$a_1 = \frac{2481523 - 16 * (84{,}625) * (1779{,}6875)}{118170 - 16 * (84{,}625)^2} = 20{,}02$$

και επομένως,

$a_0 = \bar{y} - a_1\bar{x} = 1779{,}6875 - 20{,}02*84{,}625 = 85{,}5$

Άρα η εξίσωση παλινδρόμησης είναι: $\hat{y} = 85{,}5 + 20X$

Ερμηνεία των συντελεστών: Η σταθερά (a_0) εκφράζει την αναμενόμενη τιμή της *Y* όταν το *X* είναι μηδέν, δηλαδή μπορούμε να πούμε ότι όταν η επιχείρηση δεν παράγει προϊόντα τότε το σταθερό κόστος θα είναι περίπου ίσο με 85,5 ευρώ.

Ο συντελεστής κλίσης a_1 εκφράζει την επίδραση στην αναμενόμενη τιμή της *Y* που προκαλεί η μεταβολή της *X* κατά μια μονάδα. Επομένως αν αυξηθεί η παραγωγή κατά 1 προϊόν τότε το κόστος παραγωγής θα αυξηθεί περίπου κατά 20 Ευρώ.

iii) Ο συντελεστής συσχέτισης δίνεται από τη σχέση:

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

Υπολογίζουμε ότι

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 1637459{,}4375$$

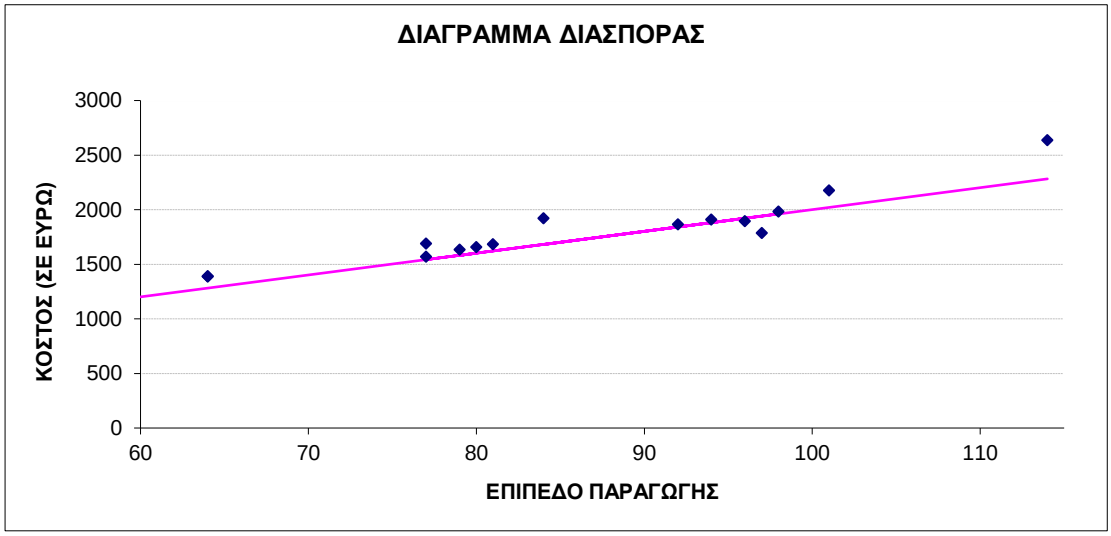
Από τον Πίνακα 1 έχουμε ότι $r = \frac{71826{,}125}{\sqrt{(3587{,}75)(1637459{,}4375)}} = 0{,}937$

Η τιμή αυτή δηλώνει την ισχυρή γραμμική εξάρτηση (όπως αναμέναμε) μεταξύ του επιπέδου και του κόστους παραγωγής.

iv)Γνωρίζουμε ότι ο συντελεστής προσδιορισμού R^2 ισούται με το τετράγωνο του συντελεστή συσχέτισης. Κατά συνέπεια, στην περίπτωσή μας θα έχουμε ότι: $R^2 = (0{,}937)^2 = 0{,}878$.

Η τιμή αυτή δηλώνει ότι το 87,8% της μεταβλητότητας του κόστους (Y) ερμηνεύεται από το αριθμό των παραγόμενων προϊόντων δηλαδή το επίπεδο παραγωγής (X).

v)



vii) Με τη βοήθεια του excel υπολογίζουμε τον παρακάτω πίνακα:

i	x_i	y_i	$\hat{y}_i$	$\hat{y}_i - y_i$
1	114	2637	2367,8	-269,2
2	101	2177	2107,5	-69,5
3	84	1920	1767,2	-152,8

4	94	1910	1967,4	57,4
5	98	1984	2047,5	63,5
6	97	1787	2027,4	240,4
7	77	1689	1627,0	-62,0
8	92	1866	1927,3	61,3
9	96	1896	2007,4	111,4
10	81	1684	1707,1	23,1
11	79	1633	1667,1	34,1
12	80	1657	1687,1	30,1
13	77	1569	1627,0	58,0
14	64	1390	1366,8	-23,2
15	64	1387	1366,8	-20,2
16	56	1289	1206,6	-82,4
			Ελάχιστη τιμή	-269,23
			Μέγιστη τιμή	240,43

Συνεπώς για την πρώτη παρτίδα, το πραγματικό κόστος ήταν 269,23 ευρώ μεγαλύτερο από το εκτιμώμενο ενώ η $6^{η}$ παρτίδα το πραγματικό κόστος ήταν 240,43 ευρώ μικρότερο από αυτό που αναμενόταν σύμφωνα με το μοντέλο της παλινδρόμησης.

vii) $\hat{y}_{500}$=85,5 + 20*50 = 1085,5

Δηλαδή για την παραγωγή 50 προϊόντων το αναμενόμενο κόστος παραγωγής είναι ίσο με 1085,5 ευρώ .

viii) Οι εκτιμήσεις της μεθόδου των ελαχίστων τετραγώνων είναι

$$b_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$b_0 = \bar{x} - b_1\bar{y}$$

Στον Πίνακα 1 προσθέτουμε δύο στήλες με τις αποκλίσεις $(Y_i\text{-}\bar{y})$ και $(Y_i\text{-}\bar{y})^2$.

Πίνακας 3

i	$\mathbf{x_i}$	$\mathbf{y_i}$	$\mathbf{x_i - \bar{x}}$	$\mathbf{(x_i - \bar{x})^2}$	$\mathbf{y_i - \bar{y}}$	$\mathbf{(Y_i\text{-}\bar{y})^2}$	$\mathbf{(x_i - \bar{x})}$ $\mathbf{(y_i - \bar{y})}$
1	114	2637	29,38	862,89	857,31	734984,72	25183,55
2	101	2177	16,38	268,14	397,31	157857,22	6505,99
3	84	1920	-0,63	0,39	140,31	19687,60	-87,70
4	94	1910	9,38	87,89	130,31	16981,35	1221,68
5	98	1984	13,38	178,89	204,31	41743,60	2732,68
6	97	1787	12,38	153,14	7,31	53,47	90,49
7	77	1689	-7,63	58,14	-90,69	8224,22	691,49
8	92	1866	7,38	54,39	86,31	7449,85	636,55
9	96	1896	11,38	129,39	116,31	13528,60	1323,05
10	81	1684	-3,63	13,14	-95,69	9156,10	346,87
11	79	1633	-5,63	31,64	-146,69	21517,22	825,12

12	80	1657	-4,63	21,39	-122,69	15052,22	567,43
13	77	1569	-7,63	58,14	-210,69	44389,22	1606,49
14	64	1390	-20,63	425,39	-389,69	151856,35	8037,30
15	64	1387	-20,63	425,39	-392,69	154203,47	8099,18
16	56	1289	-28,63	819,39	-490,69	240774,22	14045,93
ΑΘΡΟΙΣΜΑ	**1354**	**28475**	**0,00**	**3587,75**	**0,00**	**1637459,4375**	**71826,125**

και υπολογισμούς που έγιναν στο ερώτημα ii είναι εύκολο να υπολογίσουμε ότι

$$b_1 = \frac{71826,125}{1637569,4375} = 0,04386$$

Και

$$b_0 = \bar{x} - b_1\bar{y} = 84,625 - 0,04386*1773,6875 = 6,57$$

ix) Από το τυπολόγιο έχουμε

$$a_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ και } s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Συνεπώς

$$a_1 s_x^2 = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^2 \frac{1}{n-1}\sum_{i=1}^{n}(x_1 - \bar{x})^2 = \frac{\left[\sum_{i}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]^2}{(n-1)\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Όμοια από το τυπολόγιο έχουμε

$$r^2 = \frac{\left\{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right\}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{και } s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Συνεπώς

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

Συνεπώς ισχύει η ισότητα του ερωτήματος (ix).

ΑΣΚΗΣΗ 12

Ο υπεύθυνος ποιότητας μιας βιομηχανίας παραγωγής γιαουρτιού υποστηρίζει ότι το βάρος των συσκευασιών γιαουρτιού που παρασκευάζει ακολουθεί κατανομή με μέση τιμή 100 γραμμάρια και τυπική απόκλιση 3 γραμμάρια.

i. Αν η βιομηχανία παράγει 1000 συσκευασίες γιαουρτιών την ημέρα να εκτιμήσετε το ποσοστό αυτών που έχουν βάρος από 91 έως 109 γραμμάρια.

ii. Αν υποθέσουμε ότι η κατανομή του βάρους είναι η κανονική να βρεθεί το ποσοστό των συσκευασιών που έχουν βάρος από 91 έως 109 γραμμάρια.

iii. Οι υπεύθυνοι μιας οργάνωσης για την προστασία του καταναλωτή επιλέγουν τυχαία 10 συσκευασίες γιαουρτιού από τη συγκεκριμένη βιομηχανία. Ο παρακάτω πίνακας δίνει το βάρος (σε γραμμάρια) των 10 συσκευασιών.

Βάρος συσκευασιών (σε γραμμάρια)	102	94	98	99	102	101	98	96	102	98

Με βάση τα στοιχεία αυτά να κατασκευασθεί το θηκόγραμμα του βάρους των συσκευασιών και να εξαχθούν συμπεράσματα για τη μορφή της κατανομής του με βάση το δείγμα που επιλέχθηκε.

ΛΥΣΗ

i) Σύμφωνα με το Θεώρημα του Chebyshev σε απόσταση τριών τυπικών αποκλίσεων από τον μέσο (δηλαδή στο διάστημα από 91 έως 109 γραμμάρια) βρίσκεται τουλάχιστον το $\left(1-\frac{1}{\kappa^2}\right) = \left(1-\frac{1}{3^2}\right) = 0{,}89$ των συσκευασιών. Κατά συνέπεια, το ποσοστό των συσκευασιών που έχουν βάρος από 91 έως 109 γραμμάρια είναι 89%, δηλαδή τουλάχιστον 890 συσκευασίες γιαουρτιών .

ii) Έστω X η τυχαία μεταβλητή που εκφράζει το βάρος των συσκευασιών. Σύμφωνα με το πρόβλημα η τυχαία αυτή μεταβλητή ακολουθεί κανονική κατανομή με μέσο 100 γραμμάρια και τυπική απόκλιση 3 γραμμάρια, δηλαδή

$X \sim N(100,\ 3^2)$

Κατά συνέπεια, η πιθανότητα μία συσκευασία να έχει βάρος από 91 έως 109 γραμμάρια θα είναι

$$P(91 < X < 109) = P\left(\frac{91-100}{3} < Z < \frac{109-100}{3}\right) = \Phi(-3)-\Phi(3) = 0{,}9987-(1-0{,}9987)=0{,}9974$$

Άρα, το ποσοστό των συσκευασιών με βάρος από 91 έως 109 γραμμάρια θα είναι 99,74%.

iii) Για την κατασκευή του θηκογράμματος χρησιμοποιούνται οι παρακάτω πέντε τιμές: ελάχιστη και μέγιστη τιμή των δεδομένων, πρώτο και τρίτο τεταρτημόριο και διάμεσος.

Διατάσσω τα δεδομένα κατά αύξουσα τάξη μεγέθους:

ΤΑΞΗ	1	2	3	4	5	6	7	8	9	10
Βάρος συσκευασιών (σε γραμμάρια)	94	96	98	98	98	99	101	102	102	102

Διάμεσος: n = 10, άρτιος άρα

$$M=\frac{1}{2}\left(x_{\frac{n}{2}}+x_{\frac{n}{2}+1}\right)=\frac{98+99}{2}=98,5$$

1ο Τεταρτημόριο:

$$\frac{n+1}{4}=\frac{11}{4}=2.75$$

$$Q_1 = x_2 + 0,75\,(x_3 - x_2) = 96 + ,75(98 - 96) = 97,5$$

3ο Τεταρτημόριο:

$$\frac{3(n+1)}{4} = 3*11/4 = 8,25$$

$$Q_3 = x_3 + 0,25\,(x_9 - x_8) = 102 + 0,25(12\text{-}102) = 102$$

Άρα καταλήξαμε ότι:

x_{min}	Q_1	M	Q_3	x_{max}
94	97,5	98,5	102	102

Με βάση τα στοιχεία αυτά δημιουργούμε το Θηκόγραμμα των οφειλών.

Q_1 =97,5 M = 98,5 Q_3 = 102

X_{max} =102 X_{min} =94

Από το Θηκόγραμμα παρατηρούμε ότι:

- $Q_3 - M = 3,5 > 1$ = M - Q_1 (δηλαδή ότι το διαχωριστικό ενδιάμεσο τμήμα στο εσωτερικό του ορθογώνιου πλαισίου βρίσκεται πλησιέστερα προς το αριστερό άκρο του γεγονός το οποίο παραπέμπει σε δεξιά ασυμμετρία).

  Όμως

- $x_{max} - Q_3 = 0 < 3{,}5 = Q_1 - x_{min}$ (δηλαδή το μήκος της δεξιάς απόληξης είναι μικρότερο από το αντίστοιχο της αριστερής γεγονός το οποίο παραπέμπει σε αριστερή ασυμμετρία)

Κατά συνέπεια με βάση το θηκόγραμμα και μόνο δεν μπορούμε να συνάγουμε ασφαλές συμπέρασμα για τη μορφή της κατανομής των δεδομένων.

ΑΣΚΗΣΗ 13

Σε μια εταιρία μεταφορών, από την εμπειρία, είναι γνωστό ότι στα ελαστικά των φορτηγών που διαθέτει παρουσιάζεται πρόβλημα μια φορά, κατά μέσο όρο, την εβδομάδα . Έστω ότι ο αριθμός των φορτηγών που παρουσιάζουν βλάβη στα ελαστικά κατά την διάρκεια μιας εβδομάδας ακολουθεί την κατανομή Poisson.

Ποια η πιθανότητα σε μια εβδομάδα :

i) ο αριθμός των φορτηγών που παρουσιάζουν βλάβη στα ελαστικά να είναι μηδενικός;

ii) ο αριθμός των φορτηγών που παρουσιάζουν βλάβη στα ελαστικά να είναι το πολύ δύο;

iii. Να υπολογιστούν η μέση τιμή, μ, η τυπική απόκλιση, σ, του αριθμού X των φορτηγών που παρουσιάζουν πρόβλημα στα ελαστικά τους. Να υπολογισθεί η πιθανότητα όπως η X να παίρνει τιμές σε απόσταση δυο τυπικές μονάδες από τη μέση τιμή.

v. Αν ο αριθμός των φορτηγών που παρουσιάζουν μηχανική βλάβη κατά τη διάρκεια μιας εβδομάδας ακολουθεί την κατανομή Poisson με παράμετρο λ=3, τότε ποια είναι η πιθανότητα, μέσα σε μια εβδομάδα, ένα φορτηγό να μην παρουσιάσει κάποια βλάβη.

Λ Υ Σ Η

Δίνεται ότι ο αριθμός Χ των φορτηγών που παρουσιάζουν βλάβη στα ελαστικά κατά την διάρκεια μιας εβδομάδας ακολουθεί την κατανομή Poisson με λ=1.

i) Έστω *X* ο αριθμός των φορτηγών που παρουσιάζει βλάβη στα ελαστικά κατά τη διάρκεια μιας εβδομάδας. Με βάση την υπόθεση Χ: P(λ=1)και άρα η συνάρτηση πιθανότητας είναι:

$$P(X=x)\ e^{-\lambda}\frac{\lambda^{\chi}}{\chi!} = e^{-1}\frac{1^{\chi}}{\chi!}\ ,\ \chi = 0,1,2,K,$$

Ζητείται η πιθανότητα ώστε σε μια εβδομάδα να μην παρουσιασθεί βλάβη στα ελαστικά , άρα $P(x=0) = e^{-1} \cong 0,367$

iv) Ζητείται η πιθανότητα

$$P(x\leq 2) = P(x=0) + P(x=1) + P(x=2) = 0,917$$

$$P(x=0) = e^{-1}\frac{1^0}{0!} = 0,367$$

$$P(x=1) = e^{-1}\frac{1^1}{1!} = 0,367$$

$$P(x=2) = e^{-1}\frac{1^2}{2!} = 0,183$$

v) Γνωρίζω ότι για την κατανομή Poisson ισχύει: $\mu = \sigma^2 = \lambda$,

Στην περίπτωση μας λ = 1. Άρα,

$$P(\mu - 2\sigma \leq \chi \leq \mu + 2\sigma) = P(1 - 2 \leq x \leq 1 + 2)$$

$$= P(-1 \leq x \leq 3)$$

$$= P(x=0) + P(x=1) + P(x=2) + P(x=3) = 0,977$$

vi) Έστω Y ο αριθμός των φορτηγών που παρουσιάζει μηχανική βλάβη κατά τη διάρκεια μιας εβδομάδας. Με βάση την υπόθεση Y : P(λ=3). Είναι επίσης γνωστό ότι ο αριθμός X των φορτηγών που παρουσιάζει βλάβη στα ελαστικά κατά την διάρκεια μιας εβδομάδας ακολουθεί την κατανομή Poisson με λ=1, δηλαδή X : P(λ = 1).

Ζητούμε την πιθανότητα P (x =0, y =0) η οποία λόγω της ανεξαρτησίας των ενδεχομένων ισούται με P(x =0) P(y = 0).

Κατά συνέπεια προκύπτει ότι:

P(x = 0, y = 0) = P(x =0)P(y = 0)= e^{-1} e^{-3}= 0,018

ΑΣΚΗΣΗ 14

Σε μια βιομηχανία χάρτου, η παραγόμενη μονάδα προϊόντος είναι χαρτί διαστάσεων 100×1 μέτρων (δηλαδή 100 τετραγωνικών μέτρων) το οποίο συσκευάζεται σε μορφή κυλίνδρου. Η μονάδα ποιοτικού ελέγχου του εργοστασίου διαπίστωσε ότι ο αριθμός των ελαττωμάτων (στίγματα) στο χαρτί περιγράφονται από την κατανομή Poisson με μέσο όρο εμφάνισης 0,035 στίγματα ανά τετραγωνικό μέτρο.

A. Σε έναν τυχαία επιλεγμένο κύλινδρο:

i. Ποια η πιθανότητα να υπάρχουν τουλάχιστον 3 στίγματα;

ii. Ποια είναι η πιθανότητα, ο αριθμός των στιγμάτων να βρίσκεται εντός του διαστήματος [μ–σ, μ+σ] (όπου μ είναι αναμενόμενος αριθμός και σ η τυπική απόκλιση των στιγμάτων);

B. Κάθε κύλινδρος χαρτιού (ανεξάρτητα από τους άλλους) ο οποίος έχει 6 ή περισσότερα στίγματα χαρακτηρίζεται ως μη αποδεκτής ποιότητας. Οι κύλινδροι χαρτιού συσκευάζονται σε παλέτες των 10 (κυλίνδρων) και κατόπιν προωθούνται στην αγορά προς πώληση. Μια παλέτα (ανεξάρτητη από κάθε άλλη παλέτα) χαρακτηρίζεται ως μη αποδεκτή από έναν πελάτη, αν έχει τουλάχιστον τρεις (από τους δέκα) κυλίνδρους να είναι μη αποδεκτής ποιότητας.

i. Ποιος είναι ο αναμενόμενος αριθμός των μη αποδεκτών κυλίνδρων σε μια παλέτα;

ii. Ποια η πιθανότητα, μια τυχαία επιλεγμένη παλέτα να είναι μη αποδεκτή από τον πελάτη;

Γ. Αν το καθαρό βάρος των παλετών ακολουθεί κανονική κατανομή με μέση τιμή 60,2 Kgr με τυπική απόκλιση 0,6 Kgr, να υπολογιστεί η πιθανότητα μια τυχαία επιλεγμένη παλέτα να ζυγίζει:

i. περισσότερο από 59 Kgr και λιγότερο από 61,4 Kgr.

ii. Το πολύ 61,4 Kgr

iii. Τουλάχιστον 59 Kgr.

Λ Υ Σ Η

Ai) Δίνεται ότι ο αριθμός των ελαττωμάτων ακολουθεί την κατανομή Poisson με παράμετρο λ=0,035 ελαττώματα ανά τετραγωνικό μέτρο (τμ). Επομένως για έναν κύλινδρο χαρτιού (100 τμ) ο μέσος όρος εμφάνισης ελαττωμάτων θα είναι: λ=3,5 ελαττώματα ανά κύλινδρο (100 τμ), δηλαδή αν αποκαλέσουμε X την τυχαία μεταβλητή που καταγράφει τον αριθμό των ελαττωμάτων σε ένα κύλινδρο έχουμε $X \sim P(\lambda = 3{,}5)$ και επομένως:

$$P(x=x \mid \lambda = 3{,}5) = e^{-\lambda}\ \frac{\lambda^{\chi}}{\chi!} = e^{-3{,}5}\ \frac{3{,}5^{\chi}}{\chi!}, \quad \chi = 1,2,K$$

Η πιθανότητα να έχουμε τουλάχιστον 3 στίγματα σε έναν τυχαία επιλεγμένο κύλινδρο είναι:

$P(x \geq 3 \mid \lambda = 3,5) = 1 - P(x < 3 \mid \lambda = 3,5) =$

$= 1 - [P(x=0 \mid \lambda=3,5) - P(x=1 \mid \lambda = 3,5) - P(x=2 \mid \lambda = 3,)]=$

$= 1 - [\ e^{-3,5}\frac{3,5^0}{0!} + e^{-3,5}\frac{3,5^1}{1!} + e^{-3,5}\frac{3,5^2}{2!}\ ] = 1 - e^{-3,5}\ [\ 1 + 3,5 + \frac{3,5^2}{2!}\ ] =$

$= 1 - e^{-3,5}\ [\ 10,625] = \cong 1 - 0,321 = 0,679$

ii) Εφόσον για έναν τυχαία επιλεγμένο κύλινδρο έχουμε ότι ο αριθμός των ελαττωμάτων περιγράφεται από την κατανομή P(λ= 3,5) ισχύει ότι:

$$\mu=\sigma^2 = \lambda = 3,5.$$

Άρα για το διάστημα :

$[\mu-\sigma, \mu+\sigma]= [3,5-\sqrt{3,5},\ 3,5+\sqrt{3,5}] \cong [1,629,\ 5,371]$

θα έχουμε:

$P(\mu-\sigma \leq \chi \leq \mu + \sigma|\lambda = 3,5) = P(1,629 \leq x \leq 5,371|\lambda = 3,5) =$

$P(x=2|\lambda = 3,5) + P(x = 3|\lambda=3,5) + P(x=4|\ \lambda=3,5) +\ P(x=5 \mid \lambda=3,5)=$

$$e^{-3,5}\frac{3,5^2}{2!} + e^{-3,5}\frac{3,5^3}{3!} + e^{-3,5}\frac{3,5^4}{4!} + e^{-3,5}\frac{3,5^5}{5!} =$$

$$e^{-3,5}\frac{3,5^2}{2}\left[1 + \frac{3,5}{3} + \frac{3,5^2}{12} + \frac{3,5^3}{60}\right] \cong 0.722$$

<u>B</u>i) Η πιθανότητα ένας κύλινδρος να είναι μη αποδεκτής ποιότητας είναι:

$P(x \geq 6|\lambda = 3,5) = 1 - P(x < 6|\lambda = 3,5) =$

$=1-P(x=0 \mid \lambda=3,5) - P(x=1|\lambda=3,5) - \sum_{i=2}^{5} P(x = i \mid \lambda = 3,5)$

Όμως από το ερώτημα (A ii) έχουμε:

$$\sum_{i=2}^{5} P(x = i|\lambda = 3,5) = 0,722$$

Άρα

$P(x \geq 6 | \lambda = 3,5 =$

$=1-[P(x=0 |\lambda=3,5) + P(x=1|\lambda=3,5) + \sum_{i=2}^{5} P(x = i|\lambda = 3,5)] =$

$=1-[e^{-3,5}\frac{3,5^0}{0!} + e^{-3,5}\frac{3,5^1}{1!} + 0.722] = 1-e^{-3,5}[1 + 3,5] - 0,722 \cong 0,142$

Επομένως κάθε κύλινδρος είναι είτε αποδεκτός είτε μη αποδεκτός, ανεξάρτητα από τους υπόλοιπους. Η πιθανότητα να είναι μη αποδεκτός είναι 0.142. Άρα η τυχαία μεταβλητή Y που εκφράζει το πλήθος των κυλίνδρων που είναι μη αποδεκτοί σε μια παλέτα ακολουθεί την διωνυμική κατανομή, δηλαδή:$Y \sim B(n,p) = B(10, 0,142)$.

Επομένως, ο αναμενόμενος αριθμός των μη αποδεκτών κυλίνδρων σε μια παλέτα είναι:E(y)= μ-np=10*0,142=1,42.

ii) Μια τυχαία επιλεγμένη παλέτα θα είναι μη αποδεκτή αν τουλάχιστον 3 κύλινδροι είναι μη αποδεκτής ποιότητας. Εφόσον το πλήθος των μη αποδεκτών κυλίνδρων σε μια παλέτα είναι $Y \sim B(n,p) = B(10, 0,142)$.έχουμε την πιθανότητα μια παλέτα να είναι μη αποδεκτή:

$P(y \geq 3) = 1 - [P(y = 0) + P(y = 1) + P(y = 2)] =$

$=1-\left[\binom{10}{0}(0,142)^0(1-0,142)^{10-0} + \binom{10}{1}(0,142)^1(1-0,142)^{10-1} + \binom{10}{2}(0,142)^2(1-0,142)^{10-2}\right] =$

$=1-\left[\frac{10!}{0!(10-0)!}(0,858)^{10} + \frac{10!}{1!(10-1)!}(0,142)^1(0,858)^9 + \frac{10!}{2!(10-2)!}(0,142)^2(0,858)^2\right] =$

$=1-(0,858)^8[(0,858)^2 + 10(0,142)(0,858) + 45(0,142)^2] \cong 0,16$

Γi) Έστω X η τυχαία μεταβλητή που εκφράζει το καθαρό βάρος μιας παλέτας. Δίνεται ότι $X \sim N(60,2, 0,6^2$. Ζητείται η πιθανότητα $P(59 < X < 61,4)$.

$P(59< X < 61{,}4)= P\left(\frac{59-60{,}2}{0{,}6} < \frac{\chi-\mu}{\sigma} < \frac{61{,}4-60{,}2}{0{,}6}\right)=$

$=P(-2 < Z < 2) =$

$=\Phi(2)- \Phi(-2)$

$=2\Phi(2)-1$

$=2*0{,}9772 – 1 = 0{,}9544$

iii) Ζητείται η πιθανότητα $P(x\leq 61{,}4)$

$P(x\leq 61{,}4)= P\left(\frac{\chi-\mu}{\sigma} \leq \frac{61{,}4-60{,}2}{0{,}69}\right)$

$= P(Z\leq 2) = \Phi(2) = 0{,}9772$

iii)Ζητείται η πιθανότητα $P\ (x\geq 59)$

$P\ (x\geq 59)= 1-P(x< 59) = 1 - P\left(z < \frac{59-60{,}2}{0{,}6}\right)=$

$=1-P(Z< -2)$

$=1-\Phi(-2)$

$=1-1+\Phi(2)= 0{,}9772$

ΣΥΜΠΕΡΑΣΜΑΤΑ

Η στατιστική είναι ένα ισχυρό εργαλείο στην υπηρεσία οποιασδήποτε επιστήμης και παρέχει πολλές δυνατότητες, όσον αφορά στον προσδιορισμό της μεταβλητότητας, στην αντιμετώπιση, στην πρόβλεψη, στο σχεδιασμό και λήψη αποφάσεων ενώ ταυτόχρονα μας εξασφαλίζει κέρδος, χρόνο και χρήματος. Αυτό είναι σημαντικό σε μια βιομηχανία.

Καταλήγουμε στο συμπέρασμα πως η εφαρμογή των στατιστικών μεθόδων στις βιομηχανίες έδωσε λύσεις σε προβλήματα που αφορούσαν το τμήμα της παραγωγής και το τεχνικό τμήμα της βιομηχανίας. Η εφαρμογή της στατιστικής όμως, δεν είναι απαραίτητη σε μια μικρή βιομηχανία ενώ για μια μεγάλη βιομηχανία η στατιστική είναι χρήσιμη ως πληροφοριακό εργαλείο που καθιστά πιο εύκολη την λήψη αποφάσεων σε όλους τους τομείς της βιομηχανίας (π.χ. το τμήμα παραγωγής ,το τμήμα προσωπικού, το τμήμα στατιστικών μελετών).

Σε μια μεγάλη βιομηχανία ο στατιστικός συγκεντρώνει τα στατιστικά στοιχεία και έπειτα από τον απαραίτητο υπολογισμό των στατιστικών δεικτών τα παρουσιάζει σε πίνακες ή διαγράμματα με σκοπό την επεξεργασία τους και την χρησιμοποίησή τους

στην λήψη ορθών αποφάσεων. Κατά συνέπεια, η εφαρμογή της στατιστικής οδηγεί στην βελτίωση της βιομηχανίας σε οποιοδήποτε τομέα του οποίου έγινε συγκέντρωση των στατιστικών στοιχείων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Λ. Αδαμάπουλος, Χ. Δαμιανού και Α. Σβέρκος (1999) « *Μαθηματικά και στοιχεία στατιστικής» Αθήνα* , ινστιτούτο τεχνολογίας υπολογιστών και εκδόσεων ΔΙΟΦΑΝΤΟΣ.

[2] Αποστολόπουλος Θ. , (2003), *Περιγραφική στατιστική επιχειρήσεων, Αθήνα, Σύγχρονη εκδοτική*

[3] Γναρδέλης Χ., (2003) *Εφαρμοσμένη στατιστική*, Αθήνα, Εκδόσεις Παπαζήση.

[4] Λιώκη – Λειβαδά Η. και Ασημακόπουλος Δ.Ν., (2007) εισαγωγή στην εφαρμοσμένη στατιστική, Τεύχος 1 Μεθοδολογίες, Αθήνα ,Συμμετρία.

[5] Λιώκη – Λειβαδά Η. και Ασημακόπουλος Δ.Ν., (2007) εισαγωγή στην εφαρμοσμένη στατιστική, Τεύχος 2 Ασκήσεις, Αθήνα, Συμμετρία.

[6] Λεωνίδας Θαρραλίδης , Γιώργος Μαυρίδης (2016)« Μαθηματικά Γ ΕΠΑΛ » Θεσσαλονίκη ,Μαθηματική βιβλιοθήκη

[7] ΕΑΠ ΠΑΤΡΩΝ « Θέματα εργασιών»

[8] Αναστάσιος Χ.Μπάρλας (2008)« Μαθηματικά Γενικής Παιδείας» Αθήνα, Ελληνοεκδοτική

[9] Μυλωνάς Ν., Παπαδόπουλος Β., (2016) Πιθανότητες και στατιστική για μηχανικούς, Θεσσαλονίκη, Τζιόλα.

[10] Ζαφειρόπουλος Κ., (2017) Εισαγωγή στην στατιστική και τις πιθανότητες 2η έκδοση, Κριτική.