# Στατιστική

για **Ηλεκτρολόγους Μηχανικούς**

ΜΕΡΟΣ Β

**Δημήτρης Κουγιουμτζής**

http://www.users.auth.gr/dkugiu/Teach/ElectricEngineer/

E-mail: dkugiu@gen.auth.gr

Απρίλιος 2010

Στο Μέρος Α ασχοληθήκαμε με την τυχαία μεταβλητή (τ.μ.) $X$, μελετήσαμε τη συνάρτηση πυκνότητας $f_X(x)$ και τη συνάρτηση κατανομής $F_X(x)$ καθώς και κύριες παραμέτρους της κατανομής, όπως τη μέση (προσδοκώμενη) τιμή $\mathrm{E}(X) \equiv \mu$ και τη διασπορά (διακύμανση) $\mathrm{Var}(X) \equiv \sigma^2$. Η μελέτη μιας τ.μ. με τη βοήθεια της πιθανοθεωρίας προϋποθέτει ότι γνωρίζουμε (ή υποθέτουμε) την κατανομή της τ.μ. και τις παραμέτρους της.

Όταν εξετάζουμε ένα τυχαίο φαινόμενο, π.χ. το όριο ηλεκτρικού ρεύματος που καίγεται μια ασφάλεια συγκερκιμένου τύπου, δεν γνωρίζουμε από πριν ποιές τιμές ηλεκτρικού ρεύματος είναι πιο πιθανές, δηλαδή την κατανομή πιθανότητας του ορίου για την ασφάλεια. Για να μελετήσουμε τέτοια προβλήματα πρώτα συλλέγουμε δεδομένα. **Στατιστική** (statistics) είναι η επιστήμη ή 'τέχνη' του να μαθαίνουμε από τα δεδομένα. Η στατιστική συνίσταται στη συλλογή δεδομένων, στην περιγραφή τους και κυρίως στην ανάλυση τους που οδηγεί και στην απόκτηση συμπερασμάτων. Η συλλογή των δεδομένων αναφέρεται κι ως **δειγματοληψία** (sampling) κι αποτελεί ξεχωριστό πεδίο της στατιστικής που δεν θα μας απασχολήσει εδώ. Η περιγραφή των δεδομένων και η παρουσίαση συνοπτικών στοιχείων και γραφημάτων αναφέρεται ως **περιγραφική στατιστική** (descriptive statistics). Η ανάλυση στατιστικών δεδομένων και η ερμηνεία των αποτελεσμάτων αναφέρεται ως **στατιστική συμπερασματολογία** (statistical inference) και αποτελεί την βάση της στατιστικής και γι αυτό συχνά αναφέρεται απλά και ως **στατιστική**.

Στο Μέρος Β θα ασχοληθούμε με κάποια κύρια θέματα της στατιστικής. Στο Κεφάλαιο 1 θα παρουσιάσουμε συνοπτικά κύριες τεχνικές της περιγραφικής στατιστικής. Στο Κεφάλαιο 2 θα μελετήσουμε την εκτίμηση κυρίων παραμέτρων κατανομής μιας τ.μ. ενός πληθυσμού και στο Κεφάλαιο 3 θα μελετήσουμε τη συσχέτιση δύο τ.μ. καθώς και την εξάρτηση (παλινδρόμηση) μιας τ.μ. από μια άλλη.

Παραθέτονται κάποιοι βασικοί ορισμοί για την στατιστική ανάλυση:

**δεδομένα (data):** ένα σύνολο τιμών μιας τ.μ. που έχουμε στη διάθεση μας, π.χ. μετρήσεις της τάσης διάσπασης κάποιου ηλεκτρικού κυκλώματος (η τ.μ. εδώ είναι η τάση διάσπασης του κυκλώματος).

**πληθυσμός (population):** μια ομάδα ή μια κατηγορία στην οποία αναφέρεται η τ.μ., π.χ. γενικά τα ηλεκτρικά κυκλώματα ή κάποιος τύπος ηλεκτρικών κυκλωμάτων.

**δείγμα (sample):** ένα υποσύνολο του πληθυσμού που μελετάμε, π.χ. 25 κυκλώματα που χρησιμοποιήσαμε για να κάνουμε μετρήσεις της τάσης διάσπασης.

**παράμετρος (parameter):** ένα μέγεθος που συνοψίζει με κάποιο τρόπο τις τιμές της τ.μ. στον πληθυσμό, π.χ. η μέση τιμή τάσης διάσπασης κυκλώματος.

**στατιστική (statistic):** ένα μέγεθος που συνοψίζει με κάποιο τρόπο τις τιμές της τ.μ. στο δείγμα, π.χ. ο μέσος όρος τάσης διάσπασης από τα 25 ηλεκτρικά κυκλώματα που μετρήσαμε.

Οι παράμετροι πληθυσμού είναι *σταθερές* αλλά *άγνωστες* ενώ οι στατιστικές δείγματος είναι *γνωστές* αλλά *μεταβλητές*. Θα ασχοληθούμε μόνο με τις πιο γνωστές από την πιθανοθεωρία παραμέτρους κατανομής μιας τ.μ., δηλαδή τη μέση τιμή $\mu$, τη διασπορά $\sigma^2$ και την αναλογία $p$.

# Κεφάλαιο 1

# ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Σ' αυτό το κεφάλαιο θα δούμε πρώτα τρόπους να παρουσιάσουμε τα δεδομένα με στατιστικούς πίνακες και διαγράμματα και μετά να συνοψίσουμε τα δεδομένα υπολογίζοντας συνοπτικά μέτρα.

## 1.1 Περιγραφή Στατιστικών Δεδομένων

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 συμμετρικότητα τους ή την ύπαρξη ακραίων τιμών.

**Πίνακας συχνοτήτων** Τα δεδομένα ενός δείγματος για μια τ.μ. $X$ 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 $X$ τη συχνότητα εμφάνισής της $f_i$, δηλαδή πόσες φορές εμφανίζεται η κάθε διακεκριμένη τιμή στο δείγμα. Εύκολα μπορούμε επίσης να υπολογίσουμε και τη **σχετική συχνότητα** (relative frequency) εμφάνισης ή αλλιώς το **ποσοστό** (percent) $p_i$ που ορίζεται από το λόγο της συχνότητας εμφάνισης $f_i$ μιας τιμής $x_i$ προς το σύνολο των παρατηρήσεων $n$ του δείγματος

$$p_i = \frac{f_i}{n}. \tag{1.1}$$

Ορίζεται επίσης η **αθροιστική συχνότητα** (cumulative frequency) $F_i$ μιας τιμής $x_i$ ως το άθροισμα των συχνοτήτων όλων των τιμών που είναι μικρότερες ή ίσες της $x_i$,

$$F_i = \sum_{j=1}^{i} f_j \quad \text{όπου} \quad x_j \leq x_i \quad \text{για} \quad j \leq i.$$

Με τον ίδιο τρόπο ορίζεται και η αθροιστική σχετική συχνότητα $P_i$

$$P_i = \sum_{j=1}^{i} p_j \quad \text{όπου} \quad x_j \leq x_i \quad \text{για} \quad j \leq i.$$

Ο πίνακας της σχετικής συχνότητας αντιστοιχεί στην εμπειρική ή δειγματική συνάρτηση μάζας πιθανότητας της διακριτής τ.μ. $X$, δηλαδή στην εκτίμηση της $f_X(x)$ βασισμένη στο δείγμα. Όμοια ο πίνακας της αθροιστικής σχετικής συχνότητας μας δίνει μια εκτίμηση της αθροιστικής συνάρτησης κατανομής $F_X(x)$ από το δείγμα.

**Ραβδόγραμμα** Τα δεδομένα του πίνακα συχνοτήτων εύκολα μπορούν να παρασταθούν γραφικά σ' ένα **ραβδόγραμμα** (bar chart), όπου η κάθε ράβδος παρουσιάζει τη συχνότητα (ή τη σχετική συχνότητα ή την αθροιστική συχνότητα, ή ακόμα την αθροιστική σχετική συχνότητα) για κάθε τιμή $x_i$. Το ραβδόγραμμα σχετικής συχνότητας είναι το γράφημα της εμπειρικής ή δειγματικής συνάρτησης μάζας πιθανότητας (δηλαδή της εκτίμησης της $f_X(x)$ από το δείγμα). Αντίστοιχα το ραβδόγραμμα της αθροιστικής σχετικής συχνότητας απεικονίζει τη δειγματική αθροιστική συνάρτηση κατανομής. Η ίδια πληροφορία μπορεί να δοθεί και με άλλου είδους γραφήματα, όπως μ' ένα **κυκλικό διάγραμμα** ή **διάγραμμα πίτας** (pie chart) όπου το κάθε κομμάτι της επιφάνειας του κύκλου ('πίτα') παρουσιάζει τη συχνότητα της αντίστοιχης τιμής.

**Παράδειγμα 1.1** *Μια εταιρεία τροφοδοσίας ηλεκτρικών προϊόντων πουλάει ένα προϊόν στους πελάτες της σε παρτίδες των 5 λίτρων. Η εταιρεία θέλει να μελετήσει την ποσότητα του προϊόντος σε κάθε παραγγελία. Από τα αρχεία της εταιρείας βρέθηκαν 120 πρόσφατες παραγγελίες αυτού του χημικού προϊόντος και ο αριθμός των παρτίδων σε κάθε παραγγελία δίνεται στον Πίνακα 1.1.*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 2 | 2 | 2 | 3 | 4 | 3 | 1 | 1 | 3 | 3 |
| 1 | 2 | 1 | 2 | 1 | 1 | 2 | 3 | 3 | 5 | 1 | 2 |
| 2 | 3 | 2 | 1 | 1 | 4 | 3 | 4 | 1 | 1 | 6 | 2 |
| 1 | 3 | 2 | 1 | 2 | 2 | 3 | 2 | 4 | 3 | 3 | 5 |
| 1 | 3 | 5 | 3 | 1 | 2 | 2 | 3 | 1 | 2 | 6 | 4 |
| 1 | 2 | 5 | 4 | 3 | 1 | 2 | 4 | 2 | 1 | 3 | 4 |
| 2 | 2 | 2 | 3 | 2 | 1 | 3 | 3 | 4 | 2 | 1 | 5 |
| 2 | 2 | 3 | 3 | 2 | 4 | 6 | 3 | 2 | 3 | 1 | 3 |
| 2 | 1 | 5 | 1 | 1 | 4 | 4 | 2 | 5 | 4 | 2 | 2 |
| 4 | 2 | 1 | 2 | 2 | 2 | 3 | 2 | 3 | 2 | 1 | 4 |

Πίνακας 1.1: Αριθμός παρτίδων σε δείγμα 120 παραγγελιών ενός προϊόντος.

*Η τ.μ. $X$ είναι ο αριθμός των παρτίδων του προϊόντος ανά παραγγελία. Τα δεδομένα του Πίνακα 1.1 είναι διακριτά αριθμητικά (ως αριθμοί από το 1 ως το ανώτατο αριθμό παρτίδων). Με κατάλληλη επεξεργασία θα μπορούσαμε να οργανώσουμε τα δεδομένα σε διατακτικά κατηγορικά, δηλαδή ως κατηγορίες μεγέθους παραγγελίας με βάση τον αριθμό των παρτίδων, μικρού μεγέθους για 1 ή 2 παρτίδες, μεσαίου μεγέθους για 3 ή 4 παρτίδες και μεγάλου μεγέθους για περισσότερες από 4 παρτίδες.*

*Με βάση τον Πίνακα 1.1 δεν είναι εύκολο να μελετήσουμε την συχνότητα εμφάνισης των διαφόρων αριθμών παρτίδων. Ποιος αριθμός παρτίδων εμφανίζεται συχνότερα; Είναι περισσότερες παραγγελίες των 2 παρτίδων ή των 3 παρτίδων; Για να απαντήσουμε σε τέτοια ερωτήματα μπορούμε να μετρήσουμε πόσες φορές εμφανίζεται ο κάθε αριθμός παρτίδων και να φτιάξουμε έτσι τον πίνακα συχνοτήτων. Αυτούς τους απλούς υπολογισμούς μπορούμε εύκολα να τους κάνουμε*

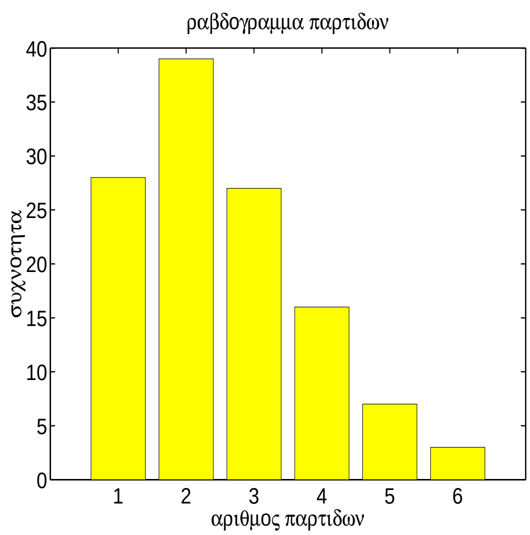
| $x_i$ | $f_i$ | $p_i$ | $F_i$ | $P_i$ |
|---|---|---|---|---|
| 1 | 28 | 0.23 | 28 | 0.23 |
| 2 | 39 | 0.33 | 67 | 0.56 |
| 3 | 27 | 0.23 | 94 | 0.78 |
| 4 | 16 | 0.13 | 110 | 0.92 |
| 5 | 7 | 0.06 | 117 | 0.97 |
| 6 | 3 | 0.03 | 120 | 1.00 |
| Άθροισμα | 120 | 1.00 | | |

Πίνακας 1.2: Πίνακας συχνοτήτων για τον αριθμό παρτίδων σε 120 παραγγελίες του Πίνακα 1.1 που περιλαμβάνει τη συχνότητα $f_i$, τη σχετική συχνότητα $p_i$, τη αθροιστική συχνότητα $F_i$ και τη σχετική αθροιστική συχνότητα $P_i$).

*μόνοι μας αλλά όταν το δείγμα είναι μεγάλο θα χρειαστεί να χρησιμοποιήσουμε κάποιο υπολογιστικό πρόγραμμα (όπως το στατιστικό πακέτο SPSS).*

*Ο Πίνακας 1.2 παρουσιάζει τις τιμές της $X$ (αριθμός παρτίδων $x_i$ για $i = 1, \dots, 6$) στην πρώτη στήλη, τη συχνότητα $f_i$ της κάθε τιμής $x_i$ στη δεύτερη στήλη, τη σχετική συχνότητα (ποσοστό) $p_i$ στην τρίτη στήλη, την αθροιστική συχνότητα $F_i$ στην τέταρτη στήλη και τη σχετική αθροιστική συχνότητα $P_i$ στην πέμπτη στήλη.*

*Στο Σχήμα 1.1 παρουσιάζεται το ραβδόγραμμα που προκύπτει από τις συχνότητες εμφάνισης του κάθε αριθμού παρτίδων. Από τον πίνακα συχνοτήτων και το ραβδόγραμμα είναι φανερό πως*




Σχήμα 1.1: Ραβδόγραμμα του αριθμού παρτίδων σε 120 παραγγελίες του Πίνακα 1.1.

*οι περισσότερες παραγγελίες στο δείγμα μας είναι 2 παρτίδων, λιγότερες είναι 1 παρτίδας ή 3 παρτίδων και η συχνότητα φθείνει καθώς αυξάνει ο αριθμός παρτίδων, δηλαδή για παραγγελίες 4, 5 και 6 παρτίδων.*

**Ομαδοποίηση** Όταν τα δεδομένα είναι αριθμητικά και είτε ο αριθμός των διακεκριμένων τιμών που παίρνει η τ.μ. $X$ είναι μεγάλος, ή η $X$ είναι συνεχής (δηλαδή παίρνει τιμές σ' ένα διάστημα τιμών), είναι προφανές ότι οι πίνακες και τα γραφήματα συχνοτήτων των τιμών δεν προσφέρονται για την απεικόνιση των δεδομένων. Σε τέτοιες περιπτώσεις πρώτα χωρίζουμε τα δεδομένα σε $k$ **ομάδες** (groups), ή κλάσεις διαστημάτων, και μετά παρουσιάζουμε σε πίνακα ή σε γράφημα τη συχνότητα της κάθε ομάδας, δηλαδή τον αριθμό των παρατηρήσεων που ανήκουν σε κάθε ομάδα.

Το εύρος τιμών $r_i$ της κάθε $i$ ομάδας είναι συνήθως το ίδιο ($r_i = r$). Δεν υπάρχει συγκεκριμένος τρόπος να το καθορίσουμε και το διαλέγουμε ανάλογα με την κλίμακα τιμών για την οποία μας ενδιαφέρει να δούμε διαφορές. Γενικά φροντίζουμε να είναι τέτοιο ώστε να μην προκύπτουν πολλές ομάδες με αποτέλεσμα να έχουμε μικρές συχνότητες γιατί τότε δε μπορούμε να διακρίνουμε κάποιο σχηματισμό στα δεδομένα. Από την άλλη δε θα πρέπει οι ομάδες να είναι πολύ λίγες γιατί τότε δε θα μπορούμε να διακρίνουμε διαφορές παρά μόνο για μεγάλες κλίμακες τιμών.

Για το χωρισμό των δεδομένων σε ομάδες βρίσκουμε πρώτα τη μικρότερη (ελάχιστη) τιμή $x_{\min}$ και μεγαλύτερη (μέγιστη) τιμή $x_{\max}$ και υπολογίζουμε το εύρος των δεδομένων

$$R = x_{\max} - x_{\min}.$$

Διαιρώντας το $R$ με τον αριθμό των ομάδων $k$ που επιλέγουμε έχουμε το εύρος τιμών $r$ της κάθε ομάδας το οποίο συνήθως στρογγυλοποιούμε για να έχουμε εύχρηστα νούμερα. Η πρώτη ομάδα έχει σαν κάτω άκρο του διαστήματος κάποιον κατάλληλα στρογγυλοποιημένο αριθμό μικρότερου ή ίσου του $x_{\min}$, τα διαστήματα των ομάδων είναι ισομήκη και το διάστημα της τελευταίας ομάδας περιλαμβάνει το $x_{\max}$. Για τα διαστήματα διαλέγουμε συμβατικά να είναι κλειστά από αριστερά (να περιέχουν την ακραία μικρότερη τιμή) κι ανοιχτά από δεξιά (να μην περιέχουν την ακραία μεγαλύτερη τιμή).

**Ιστόγραμμα** Έχοντας ομαδοποιήσει τα αριθμητικά δεδομένα μπορούμε να κάνουμε τον πίνακα και τα γραφήματα συχνοτήτων όπως και πριν. Ειδικότερα για το ραβδόγραμμα για την κάθε ράβδο παίρνουμε το κέντρο του διαστήματος που αντιστοιχεί σε κάθε ομάδα. Επίσης δεν υπάρχει κενό διάστημα μεταξύ των ράβδων και το γράφημα αυτό λέγεται **ιστόγραμμα** (histogram). Στον κάθετο άξονα του ιστογράμματος μπορεί να είναι η συχνότητα $f_i$, η σχετική συχνότητα (ποσοστό) $p_i$, η αθροιστική συχνότητα $F_i$, ή ακόμα η σχετική αθροιστική συχνότητα $P_i$ για την κάθε $i$ ομάδα. Σε πλήρη αντιστοιχία με το ραβδόγραμμα σχετικής συχνότητας το ιστόγραμμα σχετικής συχνότητας (ή απλά το ιστόγραμμα συχνότητας) δίνει μια εκτίμηση του γραφήματος της συνάρτησης πυκνότητας πιθανότητας $f_X(x)$ της συνεχούς τ.μ. $X$, βασισμένη στο δείγμα και στην επιλεγμένη ομαδοποίηση των παρατηρήσεων.

Το ιστόγραμμα είναι χρήσιμο για να κρίνουμε αν μπορούμε να δεχθούμε ότι η τ.μ. $X$, όπως την παρατηρήσαμε από το δείγμα, ακολουθεί κάποια γνωστή κατανομή. Εδώ κυρίως ενδιαφερόμαστε για την **κανονική κατανομή** γιατί τότε μπορούμε να χρησιμοποιήσουμε μια ολόκληρη μεθοδολογία εκτιμητικής που βασίζεται στην κανονική κατανομή της $X$. Για μικρά δείγματα (με λιγότερες από 30 παρατηρήσεις) δεν περιμένουμε το ιστόγραμμα να δίνει το σχήμα καμπάνας της κανονικής κατανομής (ακόμα και για δεδομένα που προέρχονται από κανονική κατανομή είναι δυνατόν να παρατηρήσουμε σημαντικές αποκλίσεις). Συνήθως δεχόμαστε ότι η τ.μ. $X$ ακολουθεί κανονική κατανομή όταν το ιστόγραμμα φαίνεται να διατηρεί κάποια σχετική συμμετρία με τις υψηλότερες συχνότητες να παρουσιάζονται στα κεντρικά διαστήματα τιμών.

Υπάρχουν κι άλλα γραφήματα που απεικονίζουν την εμπειρική κατανομή της τ.μ. $X$ με διαφορετικό τρόπο. Αναφέρουμε ενδεικτικά χωρίς να τα περιγράψουμε το *φυλλογράφημα* (stem and leaf plot) και το *σημειογράφημα* (dotplot).

**Παράδειγμα 1.2** *Στον Πίνακα 1.3 δίνονται οι χρόνοι 'ζωής' 25 μπαταριών αυτοκινήτου μιας εταιρίας Α (σε έτη με το τους μήνες τροποποιημένους σε δεκαδικά της μονάδας του έτους) που διαλέχτηκαν τυχαία (και αντίστοιχα για* $20$ *μπαταρίες αυτοκινήτων από μια άλλη εταιρία Β που θα μελετήσουμε αργότερα).*

| Α/Α | εταιρία Α | εταιρία Β |
|---|---|---|
| 1 | 5.3 | 5.0 |
| 2 | 4.5 | 4.2 |
| 3 | 5.7 | 5.4 |
| 4 | 5.8 | 5.5 |
| 5 | 4.8 | 4.6 |
| 6 | 6.4 | 6.1 |
| 7 | 6.4 | 6.1 |
| 8 | 5.6 | 5.3 |
| 9 | 5.8 | 5.5 |
| 10 | 5.7 | 5.4 |
| 11 | 5.5 | 5.2 |
| 12 | 6.1 | 5.8 |
| 13 | 5.2 | 4.9 |
| 14 | 7.0 | 6.7 |
| 15 | 5.5 | 5.2 |
| 16 | 5.7 | 5.4 |
| 17 | 6.3 | 6.0 |
| 18 | 5.6 | 5.3 |
| 19 | 5.5 | 5.2 |
| 20 | 5.0 | 4.8 |
| 21 | 5.8 | |
| 22 | 4.7 | |
| 23 | 6.1 | |
| 24 | 6.7 | |
| 25 | 5.1 | |
| Σύνολο | 141.8 | 107.6 |

Πίνακας 1.3: Δεδομένα χρόνου 'ζωής' μπαταριών αυτοκινήτου δύο εταιριών Α και Β.

*Ονομάζουμε* $X$ *την τυχαία μεταβλητή του χρόνου ζωής της μπαταρίας από την εταιρία Α. Ο μικρότερος χρόνος ζωής είναι* $x_{min} = 4.5$*, ο μεγαλύτερος χρόνος ζωής είναι* $x_{max} = 7.0$ *και το εύρος των δεδομένων του δείγματος είναι*

$$R = x_{max} - x_{min} = 7.0 - 4.5 = 2.5.$$

*Διαλέγουμε να χωρίσουμε τα δεδομένα σε 10 ομάδες (*$k = 10$*) κι άρα η κάθε ομάδα καλύπτει εύρος τιμών*

$$r = \frac{R}{k} = \frac{2.5}{10} = 0.25,$$

*αρχίζοντας από την τιμή* $4.5$. *Στη συνέχεια μπορούμε να υπολογίσουμε τον πίνακα συχνοτήτων μετρώντας τον αριθμό των δεδομένων σε κάθε ομάδα.*

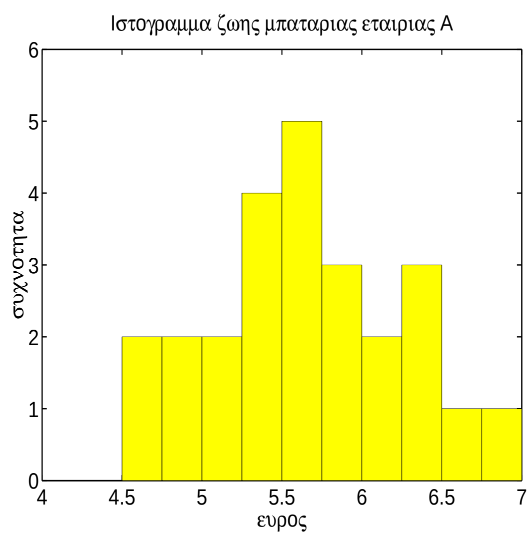
*Στον Πίνακα 1.4 δίνεται ο πίνακας συχνοτήτων, που περιλαμβάνει τη συχνότητα, τη σχετική συχνότητα, την αθροιστική συχνότητα και την σχετική αθροιστική συχνότητα. Παρατηρούμε ότι οι*

| *Διάστημα τιμών* | $f_i$ | $p_i$ | $F_i$ | $P_i$ |
|---|---|---|---|---|
| 4.50 – 4.75 | 2 | 0.08 | 2 | 0.08 |
| 4.75 – 5.00 | 2 | 0.08 | 4 | 0.16 |
| 5.00 – 5.25 | 2 | 0.08 | 6 | 0.24 |
| 5.25 – 5.50 | 4 | 0.16 | 10 | 0.40 |
| 5.50 – 5.75 | 5 | 0.20 | 15 | 0.60 |
| 5.75 – 6.00 | 3 | 0.12 | 18 | 0.72 |
| 6.00 – 6.25 | 2 | 0.08 | 20 | 0.80 |
| 6.25 – 6.50 | 3 | 0.12 | 23 | 0.92 |
| 6.50 – 6.75 | 1 | 0.04 | 24 | 0.96 |
| 6.75 – 7.00 | 1 | 0.04 | 25 | 1.00 |
| Άθροισμα | 25 | 1.00 | | |

Πίνακας 1.4: Πίνακας συχνοτήτων για τα δεδομένα του χρόνου ζωής μπαταρίας της εταιρίας Α που περιλαμβάνει τη συχνότητα $f_i$, τη σχετική συχνότητα $p_i$, την αθροιστική συχνότητα $F_i$ και τη σχετική αθροιστική συχνότητα $P_i$).

*χρόνοι ζωής μπαταρίας στο δείγμα των 25 μπαταρίων συγκεντρώνονται σε ένα κεντρικό διάστημα τιμών (περίπου οι μισές παρατηρήσεις εμφανίζονται στις τρεις ομάδες που καλύπτουν το διάστημα* $[5.25, 6.0]$*).*

*Μπορούμε να σχηματίσουμε καλύτερη εντύπωση για την κατανομή της τυχαίας μεταβλητής* $X$ *της ζωής της μπαταρίας από το ιστόγραμμα συχνοτήτων, που παρουσιάζεται στο Σχήμα 1.2. Πράγματι από το ιστόγραμμα φαίνεται ότι οι παρατηρήσεις συγκεντρώνονται σ' ένα κεντρικό*




Σχήμα 1.2: Ιστόγραμμα των δεδομένων χρόνου ζωής μπαταριών της εταιρίας Α του Πίνακα 1.3.

*διάστημα τιμών κι απλώνονται με κάποια συμμετρία γύρω από αυτό. Η εικόνα που δίνει το* *ιστόγραμμα είναι συνεπής με την γραφική παράσταση κανονικής κατανομής (σχήμα καμπάνας ) και μας επιτρέπει να δεχτούμε ότι η κατανομή του χρόνου ζωής μπαταρίας από την εταιρία Α είναι κανονική.*

## 1.2 Περιγραφικά Μέτρα Στατιστικών Δεδομένων

Ο πίνακας συχνοτήτων και το ραβδόγραμμα ή ιστόγραμμα δίνουν μια συνοπτική παρουσίαση των δεδομένων και μας επιτρέπουν να μελετήσουμε ποιοτικά την κατανομή της τυχαίας μεταβλητής που παρατηρήσαμε. Στη συνέχεια θα ορίσουμε ποσοτικά μεγέθη που περιγράφουν περιληπτικά τα βασικά χαρακτηριστικά της κατανομής της τ.μ. $X$ και λέγονται **συνοπτικά ή περιγραφικά μέτρα** (summarizing or descriptive statistics). Κάθε τέτοιο μέτρο υπολογίζεται από τις παρατηρήσεις του δείγματος κι όπως θα δούμε στα επόμενα κεφάλαια αποτελεί εκτίμηση κάποιας παραμέτρου της κατανομής της τ.μ. που μελετάμε.

Θα ασχοληθούμε με δύο τύπους περιγραφικών μέτρων:

- τα **μέτρα θέσης** (measures of location) που προσδιορίζουν χαρακτηριστικές θέσεις μέσα στο εύρος των δεδομένων και

- τα **μέτρα μεταβλητότητας** (variability measures) που δίνουν περιληπτικά τη διασκόρπιση και μεταβλητότητα των δεδομένων.

### 1.2.1 Μέτρα θέσης

Ως μέτρα θέσης εννοούμε κυρίως τα μέτρα κεντρικής τάσης που προσδιορίζουν ένα κεντρικό σημείο γύρω από το οποίο τείνουν να συγκεντρώνονται τα δεδομένα. Τα κυριότερα μέτρα κεντρικής τάσης είναι:

- η **δειγματική μέση τιμή** (sample mean value) ή **αριθμητικός μέσος** (arithmetic mean), ή **μέσος όρος** (average),

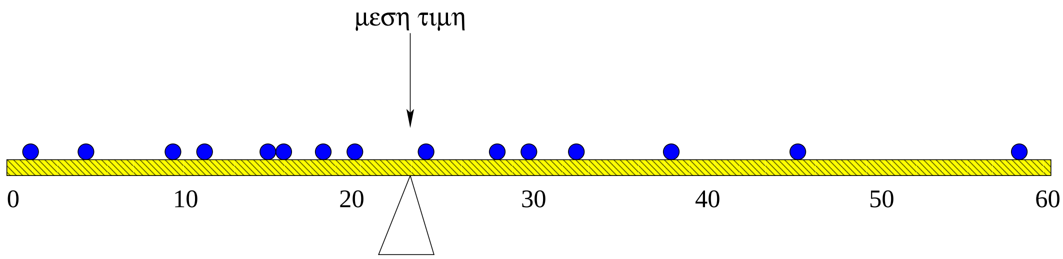
- η **δειγματική διάμεσος** (sample median),

- η **δειγματική επικρατούσα τιμή** (sample mode).

**Μέση τιμή** Η δειγματική μέση τιμή είναι το πιο γνωστό και χρήσιμο μέτρο του κέντρου των δεδομένων. Έστω $x_1, x_2, \dots, x_n$, οι τιμές των παρατηρήσεων του δείγματος για μια τ.μ. $X$ που μελετάμε. Η δειγματική μέση τιμή συμβολίζεται $\bar{x}$ κι ορίζεται ως

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.2}$$

Η μέση τιμή είναι το 'κέντρο ισορροπίας' των δεδομένων. Για να καταλάβουμε τη φυσική της σημασία ας φανταστούμε μία αβαρή σανίδα πάνω στην οποία σκορπίζουμε ένα αριθμό $n$ ίδιων βαριδιών. Το σημείο στήριξης της σανίδας (ώστε να ισορροπεί σε οριζόντια θέση) είναι η μέση τιμή της θέσης των βαριδίων πάνω στη σανίδα, όπως φαίνεται και στο Σχήμα 1.3.




Σχήμα 1.3: Σχηματική παρουσίαση της μέσης τιμής.

**Διάμεσος** Η δειγματική διάμεσος είναι ένα άλλο μέτρο του κέντρου των δεδομένων και ορίζεται ως η κεντρική τιμή όταν διατάξουμε τα δεδομένα σε αύξουσα σειρά. Θα τη συμβολίζουμε ως $\tilde{x}$. Αν ο αριθμός $n$ των δεδομένων είναι περιττός τότε η διάμεσος είναι η τιμή στη θέση $(n+1)/2$, ενώ αν το $n$ είναι άρτιος τότε είναι το ημιάθροισμα των τιμών στις θέσεις $n/2$ και $n/2+1$, δηλαδή

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & n = 2k+1 \\ \dfrac{x_{n/2} + x_{n/2+1}}{2} & n = 2k. \end{cases} \tag{1.3}$$

Για παράδειγμα σε δείγμα τριών τιμών η διάμεσος είναι η δεύτερη μικρότερη τιμή και σε δείγμα τεσσάρων τιμών η διάμεσος είναι ο μέσος όρος της δεύτερης και τρίτης μικρότερης τιμής.

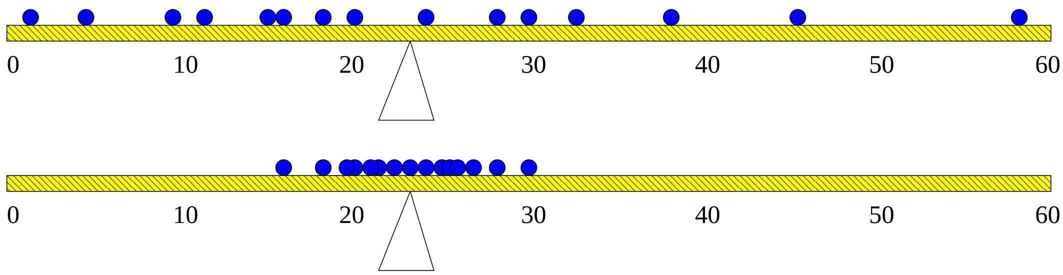
**Επικρατούσα τιμή** Η δειγματική επικρατούσα τιμή χρησιμοποιείται επίσης για να δηλώσει την κεντρική τάση των δεδομένων κι ορίζεται ως η τιμή που εμφανίζεται με τη μεγαλύτερη συχνότητα. Αν υπάρχουν πάνω από μία τέτοιες τιμές, τότε όλες αυτές θεωρούνται επικρατούσες τιμές. Είναι φανερό πως η επικρατούσα τιμή δεν έχει νόημα όταν το δείγμα δεν αποτελείται από διακεκριμένες επαναλαμβονόμενες τιμές.

Η δειγματική μέση τιμή είναι το πιο σημαντικό από τα τρία μέτρα κεντρικής τάσης και θα μας απασχολήσει ιδιαίτερα καθώς θα τη χρησιμοποιήσουμε στη στατιστική συμπερασματολογία (στα επόμενα κεφάλαια) για να βγάλουμε συμπεράσματα για τη μέση τιμή $\mu$ του πληθυσμού. Για τον υπολογισμό της μέσης τιμής χρησιμοποιούνται όλες οι τιμές του δείγματος, ενώ για τη διάμεσο μόνο η τάξη τους. Γι αυτό και η μέση τιμή επηρεάζεται από μακρινές τιμές αλλά η διάμεσος όχι. Όταν η κατανομή των αριθμητικών δεδομένων είναι μονοκόρυφη και συμμετρική, τότε και τα τρία μέτρα κεντρικής τάσης συμπίπτουν.

Η ύπαρξη μακρινών παρατηρήσεων στο δείγμα δυσκολεύει τη στατιστική περιγραφή κι ανάλυση. Γι αυτό πριν προχωρήσουμε θα πρέπει να αποφασίσουμε αν θα συμπεριλάβουμε τη μακρινή παρατήρηση (αν πιστεύουμε ότι είναι σωστή) ή αν θα την αγνοήσουμε (αν έχουμε λόγους να πιστεύουμε ότι δεν είναι ακριβής).

### 1.2.2 Μέτρα μεταβλητότητας

Εκτός από την κεντρική τάση μας ενδιαφέρει επίσης και η μεταβλητότητα ή διασπορά των παρατηρήσεων. Όταν τα δεδομένα είναι συγκεντρωμένα γύρω από μια κεντρική τιμή, δηλαδή η διασπορά των δεδομένων είναι μικρή, τότε η κεντρική τιμή αντιπροσωπεύει ικανοποιητικά τα δεδομένα. Από την άλλη, όταν τα δεδομένα είναι πολύ σκορπισμένα τα μέτρα κεντρικής τιμής δε δίνουν καλή περιληπτική περιγραφή των δεδομένων. Επίσης, διαφορετικά δείγματα από τον ίδιο πληθυσμό μπορεί να έχουν το ίδιο μέτρο κεντρικής τάσης αλλά να διαφέρουν κατά κάποιο σημαντικό τρόπο ως προς τη διασπορά των παρατηρήσεων. Χρησιμοποιώντας το παράδειγμα με τα βαρίδια και τη σανίδα βλέπουμε στο Σχήμα 1.4 πώς δύο δείγματα που έχουν την ίδια μέση τιμή μπορεί να διαφέρουν σημαντικά και κατά χαρακτηριστικό τρόπο ως προς τη διασπορά τους.




Σχήμα 1.4: Σχηματική παρουσίαση δύο δειγμάτων ίσου πλήθους με ίδια μέση τιμή και διαφορετική μεταβλητότητα.

Τα κυριότερα μέτρα διασποράς είναι:

- το **δειγματικό εύρος** (sample range) $R$,
- η **δειγματική διακύμανση** ή **δειγματική διασπορά** (sample variance) $s^2$ και η **δειγματική τυπική απόκλιση** (standard deviation) $s$.
- τα **εκατοστιαία σημεία** (percentiles) και το **ενδοτεταρτομοριακό εύρος** (interquartile range).

**Εύρος** Όπως αναφέρθηκε παραπάνω το εύρος των δεδομένων $R = x_{\max} - x_{\min}$ είναι η διαφορά της ελάχιστης από τη μέγιστη τιμή του δείγματος. Το εύρος υπολογίζεται εύκολα αλλά δεν είναι ανθεκτικό μέτρο μεταβλητότητας. Εξαρτάται μόνο από τις δύο ακραίες παρατηρήσεις $x_{\min}$ και $x_{\max}$ και αγνοεί τις υπόλοιπες παρατηρήσεις. Γι αυτό μπορεί να αλλάζει σημαντικά από δείγμα σε δείγμα (ίδιου πλήθους κι από τον ίδιο πληθυσμό). Γενικά το εύρος αυξάνει όταν μεγαλώνει το δείγμα καθώς αναμένεται να συμπεριληφθούν πιο ακραίες τιμές.

**Διασπορά** Η διασπορά ή διακύμανση μετράει τη μεταβλητότητα των παρατηρήσεων γύρω από τη μέση τιμή. Αν ορίσουμε την απόκλιση μιας παρατήρησης $x_i$ από τη μέση τιμή ως $x_i - \bar{x}$, είναι φανερό πως το άθροισμα όλων αυτών των αποκλίσεων είναι 0 γιατί χρησιμοποιώντας τον ορισμό της δειγματικής μέσης τιμής (1.2) έχουμε

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

Η δειγματική μέση τιμή $\bar{x}$ έχει οριστεί έτσι ώστε οι θετικές αποκλίσεις για τιμές μεγαλύτερες του $\bar{x}$ να είναι αθροιστικά ίδιες με τις αρνητικές αποκλίσεις για τιμές μικρότερες του $\bar{x}$. Για να μετρήσουμε λοιπόν τη μεταβλητότητα των παρατηρήσεων γύρω από τη μέση τιμή διαλέγουμε να αθροίσουμε όχι τις ίδιες τις αποκλίσεις αλλά τα τετράγωνα των αποκλίσεων. Επίσης για να πάρουμε ένα μέτρο της μέσης απόκλισης που δεν εξαρτάται από το πλήθος των παρατηρήσεων θα πρέπει να διαιρέσουμε με το πλήθος $n$ των παρατηρήσεων. Όμως για τεχνικούς λόγους που θα εξηγήσουμε παρακάτω διαιρούμε με $n-1$ αντί για $n$ και η δειγματική διασπορά $s^2$ ορίζεται ως

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2. \tag{1.4}$$

Αναπτύσσοντας τα τετράγωνα της (1.4) έχουμε τον ισοδύναμο τύπο

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right), \tag{1.5}$$

που είναι πιό εύχρηστος και χρησιμοποιείται στους υπολογισμούς.

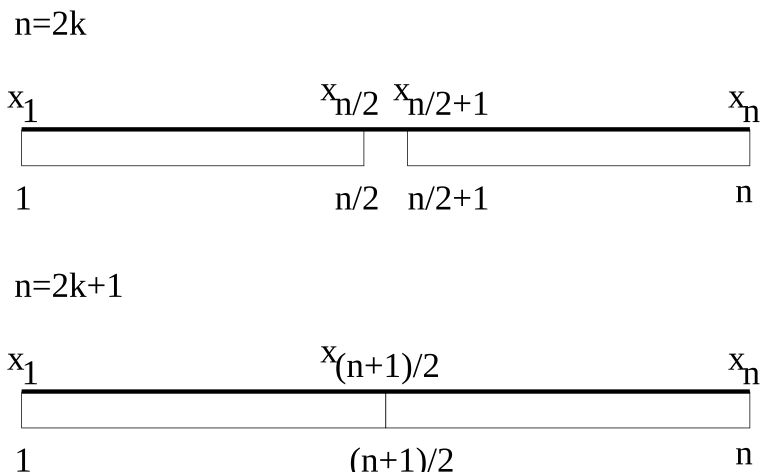
Η διασπορά $s^2$ προκύπτει από τα τετράγωνα των παρατηρήσεων και συχνά είναι δύσκολο να την ερμηνεύσουμε ως πραγματικό φυσικό μέγεθος. Γι αυτό ορίζουμε τη δειγματική τυπική απόκλιση $s$, που είναι απλά η θετική ρίζα της δειγματικής διασποράς $s^2$. Η τυπική απόκλιση $s$ μετριέται με τη μονάδα μέτρησης της τ.μ. $X$ κι εκφράζει (όπως δηλώνει η ονομασία της) την τυπική απόκλιση των δεδομένων από τη δειγματική μέση τιμή, δηλαδή μέχρι πόσο περίπου περιμένουμε μια τυπική τιμή της $X$ να απέχει από τη μέση τιμή.

*Σημείωση:* Χρήση του $n-1$ αντί του $n$

Όπως η δειγματική μέση τιμή εκτιμά τη μέση τιμή του πληθυσμού $\mu$, έτσι και η δειγματική διασπορά $s^2$ εκτιμά τη διασπορά του πληθυσμού $\sigma^2$. Αν γνωρίζαμε τη $\mu$ τότε θα τη χρησιμοποιούσαμε στον τύπο για τον υπολογισμό του $s^2$, αλλά συνήθως η $\mu$ είναι άγνωστη. Οι παρατηρήσεις $x_i$, τείνουν να είναι πιο κοντά στη $\bar{x}$ παρά στη $\mu$ κι άρα οι υπολογισμοί με βάση τις αποκλίσεις $x_i-\bar{x}$ δίνουν μικρότερες τιμές απ' ότι αν χρησιμοποιούσαμε τις αποκλίσεις $x_i-\mu$. Για να αντισταθμίσουμε αυτήν την τάση για υποεκτίμηση της διασποράς του πληθυσμού $\sigma^2$ διαιρούμε με $n-1$ αντί με $n$.

Μία άλλη πιο τεχνική εξήγηση βασίζεται στους *βαθμούς ελευθερίας* (degrees of freedom). Οι $n$ 'ελεύθερες' παρατηρήσεις αποτελούν τους $n$ βαθμούς ελευθερίας. Για τον υπολογισμό της $\bar{x}$ σχηματίζουμε το μέσο όρο διαιρώντας το άθροισμα των παρατηρήσεων με τους βαθμούς ελευθερίας $n$ αφού δεν έχουμε καμιά συνθήκη για τις $n$ παρατηρήσεις που χρησιμοποιούμε. Για τον υπολογισμό της $s^2$ όμως έχουμε της συνθήκη $\sum_{i=1}^{n}(x_i-\bar{x})=0$, δηλαδή αν ξέρουμε $n-1$ από τις αποκλίσεις μπορούμε να βρούμε αυτήν που απομένει. Άρα για τον υπολογισμό της $s^2$ οι βαθμοί ελευθερίας είναι $n-1$ και γι αυτό διαιρούμε με $n-1$.

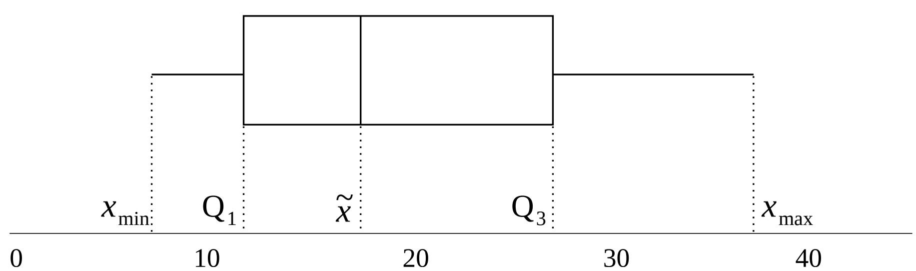
**Εκατοστιαία σημεία – ενδοτεταρτομοριακό εύρος – θηκόγραμμα** Η διάμεσος χωρίζει τα δεδομένα στα δύο. Μπορούμε να ορίσουμε άλλα σημεία χωρισμού του διατεταγμένου συνόλου τιμών που παίρνουμε από το δείγμα. Τέτοια σημεία είναι τα εκατοστιαία σημεία. Μια παρατήρηση καλείται το **$p$-εκατοστιαίο σημείο** ($p$–percentile) όταν ποσοστό παρατηρήσεων το πολύ $p\%$ είναι μικρότερες απ' αυτήν την παρατήρηση ($0 \leq p < 1$). Η διάμεσος είναι το $50$-εκατοστιαίο σημείο. Αλλά χαρακτηριστικά εκατοστιαία σημεία είναι αυτά που ορίζουν τέταρτα ή *τεταρτομόρια* (quartiles). Το $25$-εκατοστιαίο σημείο είναι το **πρώτο** ή **κατώτερο τεταρτομόριο** (first or lower quartile) και το συμβολίζουμε $Q_1$, ενώ το $75$-εκατοστιαίο σημείο είναι το **τρίτο** ή **ανώτερο τεταρτομόριο** (third or upper quartile) και το συμβολίζουμε $Q_3$. Το πρώτο και τρίτο τεταρτομόριο ορίζονται όπως η διάμεσος αλλά περιορίζοντας το σύνολο των δεδομένων στα αντίστοιχα υποσύνολα (κατώτερο ή ανώτερο μισό). Ειδικότερα, έχοντας πρώτα διατάξει τις παρατηρήσεις σε αύξουσα σειρά, αν το σύνολο των παρατηρήσεων $n$ είναι άρτιος αριθμός τότε το κατώτερο υποσύνολο περιέχει τις παρατηρήσεις από 1 ως $n/2$ και το ανώτερο από $n/2+1$ ως $n$, ενώ αν το $n$ είναι περιττός το κατώτερο υποσύνολο περιέχει τις παρατηρήσεις από 1 ως $(n+1)/2$ και το ανώτερο από $(n+1)/2$ ως $n$ (δες Σχήμα 1.5).




Σχήμα 1.5: Σχηματική παρουσίαση των δύο υποσυνόλων του δείγματος που σχηματίζονται για άρτιο και περιττό πλήθος παρατηρήσεων. Το πρώτο και τρίτο τεταρτομόριο υπολογίζονται ως οι διάμεσοι του πρώτου και δεύτερου υποσυνόλου αντίστοιχα.

Η διαφορά $I = Q_3 - Q_1$ λέγεται **ενδοτεταρτομοριακό εύρος** (interquartile range) και δίνει το εύρος που καλύπτουν τα μισά από τα δεδομένα που είναι πιο κοντά στην κεντρική τιμή (διάμεσο). Το $I$ είναι ένα άλλο μέτρο διασποράς των δεδομένων (πιο ανθεκτικό από το εύρος $R$) που δεν ορίζεται ως προς τη δειγματική μέση τιμή κι άρα δεν επηρεάζεται απ' αυτήν.

Η διάμεσος $\tilde{x}$, το πρώτο και τρίτο τεταρτομόριο $Q_1$ και $Q_3$ αντίστοιχα, καθώς και η ελάχιστη και μέγιστη τιμή των δεδομένων $x_{\min}$ και $x_{\max}$, αποτελούν τη **σύνοψη των 5 αριθμών** (five number summary). Γραφικά η παρουσίαση της σύνοψης των 5 αριθμών γίνεται με το **θηκόγραμμα** (box plot) σε οριζόντια ή κάθετη θέση όπως δείχνει το Σχήμα 1.6.




Σχήμα 1.6: Σχηματική παρουσίαση οριζόντιου θηκογράμματος.

*Σημείωση:* Θηκόγραμμα στον υπολογιστή

Το θηκόγραμμα που παράγει κάποιο στατιστικό πρόγραμμα, όπως το SPSS, μπορεί να διακόπτει τις γραμμές που ενώνουν τα άκρα του κουτιού με την ελάχιστη και μέγιστη τιμή (που λέγονται *μύστακες* (whiskers)) σε κάποια άλλα σημεία και να παρουσιάζει με ειδικά σύμβολα τις υπόλοιπες μακρινές τιμές. Για το σκοπό αυτό χρησιμοποιείται κάποιο κριτήριο για να χαρακτηριστεί μια μακρινή τιμή $x_i$ σαν *ύποπτη ακραία τιμή* (outlier), ή *ακραία τιμή* (extreme). Το κριτήριο είναι η απόσταση της τιμής από τα άκρα του κουτιού (δηλαδή από το $Q_1$ ή $Q_3$). Τυπικά αν η απόσταση αυτή είναι μεγαλύτερη από $1.5I$ (όπου τελειώνει ο μύστακας) τότε η τιμή χαρακτηρίζεται *ύποπτη ακραία*, ενώ αν είναι μεγαλύτερη από $3I$ χαρακτηρίζεται *ακραία*.

**Θηκόγραμμα και κανονική κατανομή** Το θηκόγραμμα, όπως και το ιστόγραμμα, μας επιτρέπει να κρίνουμε αν μπορούμε να δεχτούμε ότι η κατανομή της συνεχούς τυχαίας μεταβλητής που παρατηρήσαμε είναι κανονική. Για να κάνουμε αυτήν την παραδοχή θα πρέπει:

- η διάμεσος να μην αποκλίνει σημαντικά προς το πρώτο ή το τρίτο τεταρτομόριο, δηλαδή η γραμμή που αντιστοιχεί στη διάμεσο να μην πλησιάζει σε κάποιο από τα δύο άκρα του κουτιού (γιατί αλλιώς αυτό θα σήμαινε πως η κατανομή δεν είναι συμμετρική και δείχνει λοξότητα),

- το εύρος των τιμών στα δύο ακραία τεταρτομόρια να μη διαφέρει σημαντικά, δηλαδή τα μήκη των δύο μυστάκων να είναι συγκρίσιμα (για τη διατήρηση της συμμετρίας).

- να μην υπάρχουν ακραίες τιμές, δηλαδή να μην υπάρχουν σημεία μακριά από τους δύο μύστακες (η ύπαρξη ακραίων σημείων δηλώνει πως οι ουρές της κατανομής είναι ‘παχιές ’ που δε συμφωνεί με την κανονική κατανομή).

Είναι σημαντικό να αναλογιστούμε ότι με λίγα δεδομένα δεν είναι δυνατόν να τηρούνται αυστηρά οι παραπάνω προϋποθέσεις για το θηκόγραμμα ακόμα κι αν τα δεδομένα προέρχονται πράγματι από κανονική κατανομή. Αν όμως το θηκόγραμμα (όπως και το ιστόγραμμα, ή οποιοδήποτε άλλο γράφημα της κατανομής των δεδομένων) δίνει ενδείξεις σημαντικής απόκλισης από συμμετρική κατανομή τότε δεν θα πρέπει να θεωρήσουμε πως η κατανομή είναι κανονική στην στατιστική ανάλυση που θα κάνουμε στη συνέχεια.

**Παράδειγμα 1.3** *Θέλουμε να εκτιμήσουμε την περιεκτικότητα σε ραδιενέργεια του χάλυβα που παράγεται από ένα εργοστάσιο Α. Γι αυτό έγιναν μετρήσεις της ραδιενέργειας (σε Bq/g) σε 10 δοκίμια από το εργοστάσιο Α. Τα αποτελέσματα δίνονται στον Πίνακα 1.5.*

| Α/Α | εργοστάσιο Α | εργοστάσιο Β |
|---|---|---|
| 1 | 0.40 | 0.11 |
| 2 | 0.51 | 0.13 |
| 3 | 0.51 | 0.26 |
| 4 | 0.54 | 0.27 |
| 5 | 0.55 | 0.33 |
| 6 | 0.59 | 0.37 |
| 7 | 0.63 | 0.52 |
| 8 | 0.67 | 0.65 |
| 9 | 0.75 | |
| 10 | 2.10 | |
| Σύνολο | 7.25 | 2.64 |

Πίνακας 1.5: Δεδομένα περιεκτικότητας ραδιενέργειας (σε μονάδα μέτρησης Bq/g) σε δοκίμια από χάλυβα κατασκευασμένα από δύο εργοστάσια Α και Β.

*Θα ασχοληθούμε με τις παρατηρήσεις της περιεκτικότητας ραδιενέργειας στο χάλυβα από το εργοστάσιο Α κι ας ονομάσουμε αυτήν την τυχαία μεταβλητή $X$. Η δειγματική μέση τιμή υπολογίζεται από το άθροισμα των 10 παρατηρήσεων που δίνονται στη δεύτερη στήλη του Πίνακα 1.5 ως*

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{7.25}{10} = 0.725.$$

*Η δειγματική διάμεσος εύκολα μπορεί να βρεθεί μια κι οι παρατηρήσεις δίνονται σε αύξουσα σειρά. Αφού το $n$ είναι άρτιο η διάμεσος δίνεται ως*

$$\tilde{x} = \frac{x_{n/2} + x_{n/2+1}}{2} = \frac{x_5 + x_6}{2} = \frac{0.55 + 0.59}{2} = 0.57.$$

*Παρατηρούμε ότι η δειγματική μέση τιμή δίνει μεγαλύτερη τιμή στην εκτίμηση της κεντρικής τάσης των δεδομένων από τη δειγματική διάμεσο. Αυτό συμβαίνει γιατί η μέση τιμή επηρεάζεται από την ακραία τιμή της δεκάτης παρατήρησης που είναι πολύ μεγαλύτερη από όλες τις άλλες. Θα πρέπει να γνωρίζουμε αν αυτή η ακραία τιμή είναι πραγματική ή οφείλεται σε κάποιο σφάλμα της παρατήρησης (σφάλμα του μηχανήματος μέτρησης, σφάλμα στην καταγραφή κτλ).*

*Η μικρότερη περιεκτικότητα ραδιενέργειας στο δείγμα είναι $x_{min} = 0.40$ Bq/g κι η μεγαλύτερη είναι $x_{max} = 2.10$ Bq/g. Άρα το εύρος των δεδομένων είναι $R = 1.70$ Bq/g, που είναι μεγάλο εξαιτίας της ακραίας τιμής που συμπεριλάβαμε στο δείγμα.*

*Για να βρούμε τη διασπορά $s^2$ της περιεκτικότητας της ραδιενέργειας στο χάλυβα στο δείγμα μας υπολογίζουμε πρώτα το άθροισμα των τετραγώνων των παρατηρήσεων*

$$\sum_{i=1}^{10} x_i^2 = 0.40^2 + 0.51^2 + \cdots + 0.75^2 + 2.10^2 = 7.44$$

*κι αντικαθιστώντας το στον τύπο της δειγματικής διασποράς (1.5) βρίσκουμε*

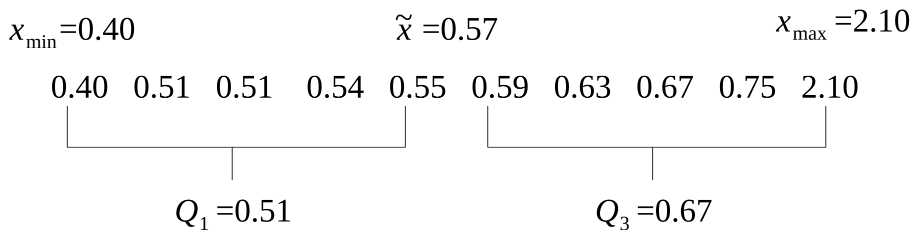
$$s^2 = \frac{1}{9}\left(\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2\right) = \frac{1}{9}\left(7.44 - 10\cdot 0.725^2\right) = 0.243.$$

*Η δειγματική τυπική απόκλιση είναι*

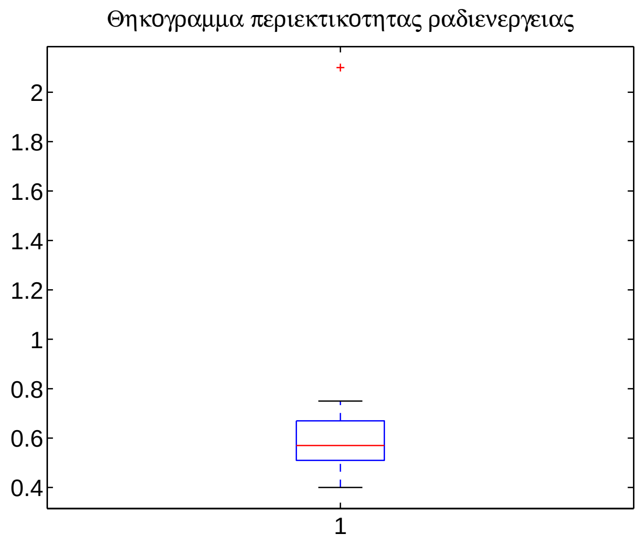
$$s = \sqrt{s^2} = \sqrt{0.243} = 0.493,$$

*δηλαδή η αντιπροσωπευτική τυπική απόκλιση από τη μέση περιεκτικότητα ραδιενέργειας είναι περίπου $0.5$ Bq/g, που είναι πολύ μεγάλη (όση περίπου και η διάμεσος).*

*Η σύνοψη των 5 αριθμών δίνεται στο Σχήμα 1.7 όπου παρουσιάζεται σχηματικά και η εύρεση του πρώτου και τρίτου τεταρτομορίου. Από το πρώτο και τρίτο τεταρτομόριο υπολογίζουμε το ενδοτεταρτομοριακό εύρος, $I = Q_3 - Q_1 = 0.67 - 0.51 = 0.16$ Bq/g. Έχοντας βρεί τη σύνοψη των 5 αριθμών μπορούμε εύκολα να παραστήσουμε το θηκόγραμμα. Στο Σχήμα 1.8 δίνεται το θηκόγραμμα σε κατακόρυφη θέση. Παρατηρούμε ότι ο άνω μύστακας δεν προεκτείνεται ως τη μέγιστη τιμή των δεδομένων που είναι 2.10. Αυτή η τιμή δηλώνεται ως ακραία τιμή με ιδιαίτερο σύμβολο, αφού η απόσταση του αντίστοιχου σημείου από το πάνω μέρος του κουτιού, $Q_3 = 0.67$, είναι μεγαλύτερη από $3I = 3\cdot 0.16 = 0.48$.*




Σχήμα 1.7: Σχηματική παρουσίαση της σύνοψης των 5 αριθμών μαζί με τις 10 παρατηρήσεις της περιεκτικότητας ραδιενέργειας.




Σχήμα 1.8: Θηκόγραμμα των 10 παρατηρήσεων περιεκτικότητας ραδιενέργειας σε χάλυβα του εργοστασίου Α.

**Παράδειγμα 1.4** *Θεωρούμε πως η ακραία τιμή στο προηγούμενο παράδειγμα οφείλεται πράγματι σε κάποιο σφάλμα της μέτρησης και την απαλείφουμε. Για το νέο δείγμα των 9 παρατηρήσεων υπολογίζουμε τα μέτρα θέσης και μεταβλητότητας.*

*Η μέση τιμή είναι*

$$\bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = \frac{5.15}{9} = 0.572$$

*και η διάμεσος (το $n$ είναι περιττό)*

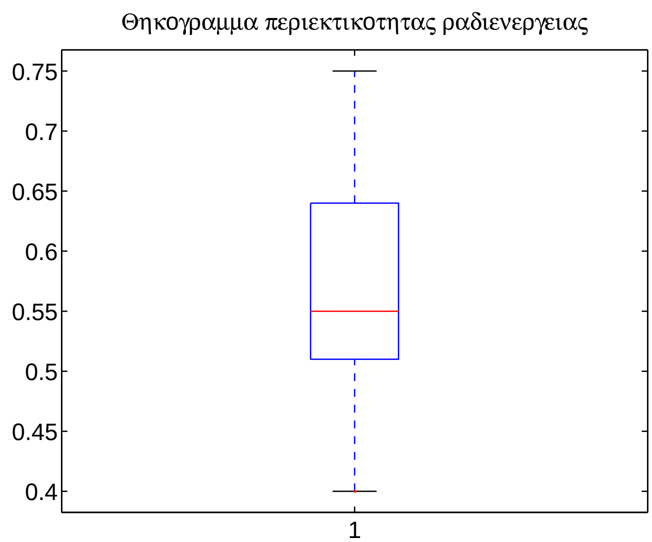
$$\tilde{x} = x_{(n+1)/2} = x_5 = 0.55.$$

*Βλέπουμε λοιπόν ότι μετά την απαλοιφή της ακραίας τιμής η δειγματική μέση τιμή και η διάμεσος δίνουν περίπου την ίδια εκτίμηση της κεντρικής τάσης των δεδομένων, που είναι ένδειξη συμμετρίας της κατανομής.*

*Για τα άλλα μέτρα έχουμε:*

| | |
|---|---|
| *Διασπορά* | $s^2 = \frac{1}{8}(5.15 - 9 \cdot 0.572^2) = 0.010.$ |
| *Τυπική απόκλιση* | $s = \sqrt{0.010} = 0.10$ |
| *Ελάχιστη τιμή* | $x_{min} = 0.40$ |
| *Μέγιστη τιμή* | $x_{max} = 0.75$ |
| *Εύρος* | $R = 0.75 - 0.40 = 0.35$ |
| *Πρώτο τεταρτομόριο* | *(διάμεσος των* $\{x_1, \ldots, x_5\}$*)* $Q_1 = x_3 = 0.51$ |
| *Τρίτο τεταρτομόριο* | *(διάμεσος των* $\{x_5, \ldots, x_9\}$*)* $Q_3 = x_7 = 0.63$ |
| *Ενδοτεταρτομοριακό εύρος* | $I = 0.63 - 0.51 = 0.12$ |

*Το αντίστοιχο θηκόγραμμα δίνεται στο Σχήμα 1.9. Από τα μέτρα θέσης και μεταβλητότητας*




Σχήμα 1.9: Θηκόγραμμα των 9 παρατηρήσεων περιεκτικότητας ραδιενέργειας σε χάλυβα του εργοστασίου Α.

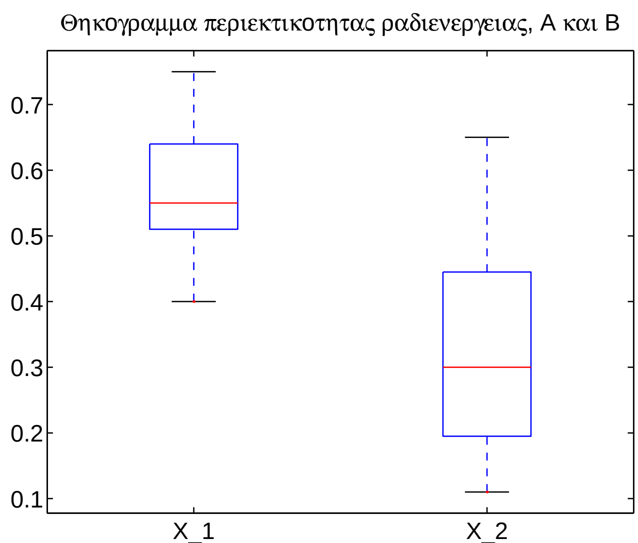
*καθώς κι από το θηκόγραμμα παρατηρούμε ότι με την απαλοιφή της ακραίας τιμής η κατανομή της περιεκτικότητας ραδιενέργειας στο χάλυβα του εργοστασίου Α φαίνεται να είναι συμμετρική και μπορούμε να κάνουμε τώρα την παραδοχή ότι η κατανομή είναι κανονική με βάση το δείγμα.*

**Παράδειγμα 1.5** *Στον Πίνακα 1.5 παρουσιάζονται επίσης οι μετρήσεις της περιεκτικότητας ραδιενέργειας σε 8 δοκίμια χάλυβα που κατασκευάστηκαν από ένα άλλο εργοστάσιο Β. Θέλουμε να συγκρίνουμε την περιεκτικότητα ραδιενέργειας στο χάλυβα από τις δύο μονάδες παραγωγής με βάση τα δείγματα των 9 και 8 παρατηρήσεων που έχουμε από το εργοστάσιο Α και από το εργοστάσιο Β αντίστοιχα. Έστω* $X_1$ *η τ.μ. της περιεκτικότητας ραδιενέργειας στο χάλυβα για το εργοστάσιο Α και* $X_2$ *η αντίστοιχη τ.μ. για το εργοστάσιο Β. Στο προηγούμενο παράδειγμα υπολογίσαμε τα μέτρα θέσης και μεταβλητότητας για τη* $X_1$*. Κάνουμε το ίδιο για τη* $X_2$*. Τα συγκεντρωτικά αποτελέσματα δίνονται στον Πίνακα 1.6. Επίσης στο Σχήμα 1.10 δίνεται το συνδυασμένο θηκόγραμμα για το δείγμα περιεκτικότητας ραδιενέργειας σε χάλυβα κι από τα δύο εργοστάσια.*

*Από τα μέτρα θέσης (μέση τιμή και διάμεσο) φαίνεται ότι η κεντρική τάση της περιεκτικότητας ραδιενέργειας είναι μεγαλύτερη για το χάλυβα του εργοστασίου Α. Από τα μέτρα μεταβλητότητας (τυπική απόκλιση, εύρος δεδομένων και ενδοτεταρτομοριακό εύρος) φαίνεται πως η περιεκτικότητα ραδιενέργειας μεταβάλλεται λιγότερο στο χάλυβα του εργοστασίου Α.*

| Μέτρο | $X_1$ | $X_2$ |
|---|---|---|
| Μέση τιμή | $\bar{x}_1 = 0.572$ | $\bar{x}_2 = 0.33$ |
| Διάμεσος | $\tilde{x}_1 = 0.55$ | $\tilde{x}_2 = 0.30$ |
| Διασπορά | $s_1^2 = 0.010$ | $s_2^2 = 0.034$ |
| Τυπική απόκλιση | $s_1 = 0.10$ | $s_2 = 0.18$ |
| Ελάχιστη τιμή | $x_{1,\min} = 0.40$ | $x_{2,\min} = 0.11$ |
| Μέγιστη τιμή | $x_{1,\max} = 0.75$ | $x_{1,\max} = 0.65$ |
| Εύρος | $R_1 = 0.35$ | $R_2 = 0.54$ |
| Πρώτο τεταρτομόριο | $Q_{1,1} = 0.51$ | $Q_{2,1} = 0.195$ |
| Τρίτο τεταρτομόριο | $Q_{1,3} = 0.63$ | $Q_{2,3} = 0.445$ |
| Ενδοτεταρτομοριακό εύρος | $I_1 = 0.12$ | $I_2 = 0.250$ |

Πίνακας 1.6: Μέτρα θέσης και μεταβλητότητας για τα δεδομένα περιεκτικότητας ραδιενέργειας σε 9 και 8 δοκίμια χάλυβα κατασκευασμένα από τα εργοστάσια Α και Β στη στήλη 2 και 3 αντίστοιχα.




Σχήμα 1.10: Θηκόγραμμα των 9 παρατηρήσεων περιεκτικότητας ραδιενέργειας σε χάλυβα του εργοστασίου Α και των 8 παρατηρήσεων περιεκτικότητας ραδιενέργειας σε χάλυβα του εργοστασίου Β.

# Κεφάλαιο 2

# ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ

Οι στατιστικές δείγματος, όπως η δειγματική μέση τιμή $\bar{x}$ και η δειγματική διασπορά $s^2$, που θα δούμε παρακάτω, υπολογίζονται από τα στατιστικά δεδομένα που έχουμε συλλέξει και χρησιμοποιούνται για την *εκτίμηση* των σχετικών παραμέτρων πληθυσμού ($\mu$ και $\sigma^2$ αντίστοιχα).

Τα στατιστικά δεδομένα πρέπει να προέρχονται από *τυχαίο* κι *αντιπροσωπευτικό* δείγμα του πληθυσμού. Για παράδειγμα, αν θέλουμε να μελετήσουμε το όριο της έντασης ηλεκτρικού ρεύματος που καίγονται ασφάλειες 40 αμπέρ θα πρέπει να διαλέξουμε στην τύχη ασφάλειες του ίδιου τύπου και να μετρήσουμε σε ποια ένταση του ρεύματος καίγεται η κάθε μια από αυτές.

Από τις μετρήσεις του δείγματος μπορούμε να υπολογίσουμε τη **σημειακή εκτίμηση** (point estimation) και την **εκτίμηση διαστήματος** (interval estimation) της παραμέτρου μιας τ.μ..

## 2.1 Σημειακή Εκτίμηση

Η *σημειακή εκτίμηση* είναι μια τιμή, που υπολογίζεται με βάση τα δεδομένα και αντιπροσωπεύει την πραγματική τιμή της σχετικής παραμέτρου του πληθυσμού. Για παράδειγμα, ο μέσος όρος των 25 τιμών τάσης διάσπασης που μετρήθηκαν σε δείγμα 25 ηλεκτρικών κυκλωμάτων αποτελεί μια σημειακή εκτίμηση της μέσης τάσης διάσπασης κυκλώματος.

Έστω $X$ μια τ.μ. με συνάρτηση κατανομής $F_X(x;\theta)$ που εξαρτάται από την παράμετρο $\theta$ την οποία θέλουμε να εκτιμήσουμε. Έστω ακόμα ότι έχουμε ανεξάρτητες μεταξύ τους παρατηρήσεις $\{x_1,\dots,x_n\}$ της $X$ από ένα δείγμα μεγέθους $n$. Τότε η σημειακή εκτίμηση της $\theta$ δίνεται από τη συνάρτηση $g(x_1,\dots,x_n)$ των τιμών του δείγματος που λέγεται **εκτιμήτρια συνάρτηση**. Η **εκτιμήτρια** (estimator) της $\theta$ από το δείγμα είναι $\hat{\theta}=g(x_1,\dots,x_n)$.

Επειδή οι παρατηρήσεις $\{x_1,\dots,x_n\}$ αλλάζουν κάθε φορά που μελετάμε διαφορετικό δείγμα μεγέθους $n$, μπορούμε να υποθέσουμε ότι οι παρατηρήσεις $\{x_1,\dots,x_n\}$ είναι τιμές των τ.μ. $\{X_1,\dots,X_n\}$, που είναι ανεξάρτητες μεταξύ τους κι ακολουθούν την ίδια κατανομή $F(x;\theta)$. Η παράμετρος $\hat{\theta}$ είναι συνάρτηση αυτών των τ.μ.. Για ευκολία θα χρησιμοποιούμε το συμβολισμό $\{x_1,\dots,x_n\}$ και θεωρητικά (εννοώντας τις τ.μ. $\{X_1,\dots,X_n\}$) και πρακτικά (εννοώντας τις παρατηρούμενες αριθμητικές τιμές αυτών των τ.μ.). Θα αναφερόμαστε στις $\{x_1,\dots,x_n\}$ ως **τυχαίο δείγμα**.

Είναι φανερό ότι για διαφορετικά τυχαία δείγματα η εκτιμήτρια συνάρτηση της παραμέτρου $\hat{\theta}$ παίρνει διαφορετικές τιμές, δηλαδή η $\hat{\theta}$ είναι η ίδια τ.μ. με κάποια κατανομή κι έχει μέση τιμή $\mu_{\hat{\theta}}\equiv \mathrm{E}(\hat{\theta})$ και διασπορά $\sigma^2_{\hat{\theta}}\equiv \mathrm{Var}(\hat{\theta})$.

Δύο σημαντικές παράμετροι μιας τ.μ. $X$ που θέλουμε να εκτιμήσουμε είναι η μέση τιμή $\mu$ κι η διασπορά $\sigma^2$.

**Εκτίμηση μέσης τιμής** Είναι φυσικό σαν εκτιμήτρια της $\mu$ να ορίσουμε τον αριθμητικό μέσο όρο των $n$ παρατηρήσεων, που λέγεται και **δειγματική μέση τιμή**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.1}$$

**Εκτίμηση διασποράς** Όπως για τη μέση τιμή έτσι και για τη διασπορά $\sigma^2$ η εκτιμήτρια είναι η δειγματική διασπορά που ορίζεται με δύο τρόπους, έχουμε δηλαδή τις δύο εκτιμήτριες της $\sigma^2$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.2}$$

$$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{2.3}$$

Οι εκτιμήτριες $s^2$ και $\tilde{s}^2$ διαφέρουν μόνο ως προς το συντελεστή του αθροίσματος ($\frac{1}{n-1}$ και $\frac{1}{n}$ αντίστοιχα). Για μεγάλο $n$ οι δύο εκτιμήτριες συγκλίνουν στην ίδια τιμή. Αναπτύσσοντας τα τετράγωνα της (2.2) έχουμε

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

και χρησιμοποιώντας τον ορισμό της $\bar{x}$ από την (2.1) παίρνουμε τον ισοδύναμο τύπο (όμοια για τη (2.3))

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right). \tag{2.4}$$

Ο τύπος (2.4) είναι πιο εύχρηστος για τον υπολογισμό της $s^2$.

### 2.1.1 Κριτήρια καλών εκτιμητριών

Παραπάνω ορίσαμε κάπως αυθαίρετα εκτιμήτριες της μέσης τιμής $\mu$ και της διασποράς $\sigma^2$ χωρίς να γνωρίζουμε αν είναι 'καλές' εκτιμήτριες ή όχι. Γενικά όταν ορίζουμε μια εκτιμήτρια $\hat{\theta}$ κάποιας παραμέτρου $\theta$ θέλουμε να ελέγξουμε αν είναι κατάλληλη και γι αυτό θέτουμε κάποια κριτήρια ή ιδιότητες που πρέπει να πληρεί μια 'καλή' εκτιμήτρια. Παρακάτω περιγράφονται ορισμένες επιθυμητές ιδιότητες μιας εκτιμήτριας $\hat{\theta}$.

**Αμεροληψία** Η $\hat{\theta}$ είναι αμερόληπτη (unbiased) αν η μέση τιμή της είναι ίση με την παράμετρο $\theta$, δηλαδή $\mathrm{E}(\hat{\theta}) = \theta$. Αλλιώς λέγεται μεροληπτική με μεροληψία $b(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta$.

Αμεροληψία μέσης τιμής

Η δειγματική μέση τιμή $\bar{x}$ που ορίστηκε στη (2.1) είναι αμερόληπτη εκτιμήτρια της μέσης τιμής $\mu$ μιας τ.μ. $X$ ενός πληθυσμού.

Έχουμε $\theta \to \mu$ και $\hat{\theta} \to \bar{x}$ και θέλουμε να δείξουμε ότι $\mathrm{E}(\bar{x}) = \mu$.

$$\mathrm{E}(\bar{x}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(x_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

Αμεροληψία διασποράς

Για την εκτίμηση της διασποράς $\sigma^2$ μιας τ.μ. $X$ του πληθυσμού ισχύουν οι παρακάτω δύο προτάσεις:

1. Η δειγματική διασπορά $s^2$ που ορίστηκε στη (2.2) είναι αμερόληπτη εκτιμήτρια της $\sigma^2$, $\mathrm{E}(s^2) = \sigma^2$.

2. Η δειγματική διασπορά $\tilde{s}^2$ που ορίστηκε στη (2.3) είναι μεροληπτική εκτιμήτρια της $\sigma^2$ με μεροληψία $b(\tilde{s}^2) = -\frac{\sigma^2}{n}$.

Αποδεικνύουμε πρώτα την πρόταση (1). Υπενθυμίζουμε ότι για μια τ.μ. $X$ με μέση τιμή $\mu$ και διασπορά $\sigma^2$ ισχύουν

$$\sigma^2 = \mathrm{E}(X^2) - \mu^2 \qquad \text{και} \qquad \mathrm{E}(X^2) = \sigma^2 + \mu^2. \tag{2.5}$$

Επίσης επειδή οι $\{x_1, \ldots, x_n\}$ είναι ανεξάρτητες κι έχουν την ίδια διασπορά $\sigma^2$ με τη $X$ ισχύει

$$\mathrm{Var}\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n}\mathrm{Var}(x_i) = n\sigma^2$$

και άρα η διασπορά της δειγματικής μέσης τιμής $\bar{x}$ είναι

$$\sigma_{\bar{x}}^2 \equiv \mathrm{Var}(\bar{x}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}. \tag{2.6}$$

Χρησιμοποιώντας την παραπάνω ταυτότητα για τη διασπορά καθώς κι ότι $\mu_{\bar{x}} \equiv \mathrm{E}(\bar{x}) = \mu$ έχουμε από την (2.5)

$$\mathrm{E}(\bar{x}^2) = \sigma_{\bar{x}}^2 + \mu_{\bar{x}}^2 = \frac{\sigma^2}{n} + \mu^2.$$

Στη συνέχεια υπολογίζουμε τη μέση τιμή της $s^2$ που δίνεται από την (2.4):

$$\begin{aligned}
\mathrm{E}(s^2) &= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathrm{E}(x_i^2) - n\mathrm{E}(\bar{x}^2)\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n}(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right) \\
&= \frac{1}{n-1}(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2) = \sigma^2
\end{aligned}$$

κι αποδεικνύεται η αμεροληψία της εκτιμήτριας $s^2$.

Για την πρόταση (2), από παραπάνω προκύπτει (διαιρώντας με $n$ αντί του $n-1$) ότι $\mathrm{E}(\tilde{s}^2) = \frac{n-1}{n}\sigma^2$ κι άρα η αμεροληψία είναι

$$b(\tilde{s}^2) = \mathrm{E}(\tilde{s}^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

**Συνέπεια** Η ιδιότητα αυτή ορίζει πως όσο αυξάνει το μέγεθος $n$ του δείγματος τόσο μεγαλώνει η πιθανότητα η εκτίμηση να είναι 'κοντά' στην πραγματική τιμή της παραμέτρου, όπου το 'κοντά' δίνεται από κάποια απόσταση $\epsilon$. Δηλαδή η $\hat{\theta}$ είναι συνεπής (consistent) αν

$$\mathrm{P}(|\theta - \hat{\theta}| \leq \epsilon) \to 1 \quad \text{όταν} \quad n \to \infty,$$

όπου $\epsilon$ είναι αυθαίρετα μικρός θετικός αριθμός. Όταν μια εκτιμήτρια είναι συνεπής τότε με την αύξηση του δείγματος οι τιμές της θα συγκλίνουν στην πραγματική τιμή της παραμέτρου.

Η εκτιμήτρια $\bar{x}$ της μέσης τιμής $\mu$ μιας τ.μ. $X$ είναι συνεπής. Αν όμως αντί της δειγματικής μέσης τιμής διαλέξουμε σαν εκτιμήτρια της $\mu$ τον αριθμητικό μέσο της μικρότερης και μεγαλύτερης τιμής του δείγματος

$$x_d = \frac{x_{\min} + x_{\max}}{2},$$

τότε μπορεί να δειχθεί ότι η εκτιμήτρια $x_d$ δεν είναι συνεπής εκτιμήτρια της $\mu$.

**Αποτελεσματικότητα** Η αποτελεσματικότητα αναφέρεται στη διασπορά της εκτιμήτριας και δίνεται συγκριτικά. Μια εκτιμήτρια $\hat{\theta}_1$ της $\theta$ είναι πιο αποτελεσματική (effective) από μια άλλη εκτιμήτρια $\hat{\theta}_2$ αν έχει μικρότερη διασπορά, $\sigma^2_{\hat{\theta}_1} < \sigma^2_{\hat{\theta}_2}$.

Για παράδειγμα, η δειγματική μέση τιμή $\bar{x}$ και η $x_d$ είναι δύο εκτιμήτριες της μέσης τιμής $\mu$ κι έχουν διασπορές $\sigma^2_{\bar{x}}$ και $\sigma^2_{x_d}$ αντίστοιχα. Μπορεί να δειχθεί ότι $\sigma^2_{\bar{x}} < \sigma^2_{x_d}$ κι άρα η εκτιμήτρια $\bar{x}$ είναι πιό αποτελεσματική από τη $x_d$.

**Επάρκεια** Μια εκτιμήτρια της παραμέτρου $\theta$ είναι επαρκής (adequate) όταν χρησιμοποιεί όλη την πληροφορία από το δείγμα που σχετίζεται με τη $\theta$.

Η δειγματική μέση τιμή $\bar{x}$, εκτιμήτρια της μέσης τιμής $\mu$ μιας τ.μ. $X$, είναι επαρκής γιατί χρησιμοποιεί όλες τις παρατηρήσεις που μετρήθηκαν στο δείγμα, ενώ η $x_d$ δεν είναι επαρκής γιατί χρησιμοποιεί μόνο δύο τιμές των παρατηρήσεων του δείγματος ($x_{\min}$ και $x_{\max}$).

**Παρατηρήσεις**

Οι εκτιμήτριες $\bar{x}$ για την παράμετρο $\mu$ και $s^2$ για την παράμετρο $\sigma^2$, που ορίσαμε αυθαίρετα, πληρούν όλες τις τέσσερις ιδιότητες και είναι καλές εκτιμήτριες.

Όταν θέλουμε να βρούμε εκτιμήτρια για μια παράμετρο, μας ενδιαφέρει κυρίως να είναι αμερόληπτη και να έχει μικρή διασπορά για να εξασφαλίσουμε ότι οι τιμές της εκτιμήτριας συγκεντρώνονται στην πραγματική τιμή της παραμέτρου. Γι αυτό ορίζουμε την **αμερόληπτη εκτιμήτρια ελάχιστης διασποράς** (minimum variance ubiased estimator), δηλαδή αυτή που απ' όλες τις αμερόληπτες εκτιμήτριες έχει την μικρότερη διασπορά.

Στον ορισμό των εκτιμητριών $\bar{x}$ και $s^2$ δεν κάναμε κάποια υπόθεση για την κατανομή της τ.μ. $X$ κι άρα μπορούμε να τις χρησιμοποιήσουμε για οποιαδήποτε τ.μ. $X$ που παρατηρούμε.

Στη συνέχεια θα υποθέσουμε πως η κατανομή της $X$ είναι γνωστή (ως προς τη γενική μορφή της) αλλά δεν είναι γνωστή κάποια παράμετρο $\theta$ της κατανομής και θα δούμε πως μπορούμε γενικά να υπολογίσουμε την εκτιμήτρια της $\theta$.

## 2.1.2 Μέθοδος υπολογισμού της σημειακής εκτίμησης

Υποθέτουμε ότι η τ.μ. $X$ έχει κάποια γνωστή κατανομή, δηλαδή γνωρίζουμε τη γενική μορφή της αθροιστικής συνάρτησης κατανομής $F(x;\theta)$ και της $f(x;\theta)$, που είναι η συνάρτηση πυκνότητας πιθανότητας αν η $X$ είναι συνεχής και η συνάρτηση μάζας πιθανότητας αν η $X$ είναι διακριτή. Η παράμετρος $\theta$ της κατανομής είναι άγνωστη και θέλουμε να την εκτιμήσουμε από το τυχαίο δείγμα $\{x_1, \ldots, x_n\}$. Θα θεωρήσουμε επίσης τη γενική περίπτωση να έχουμε περισσότερες από μία άγνωστες παραμέτρους.

**Μέθοδος των Ροπών** Για συνήθεις κατανομές, μια παράμετρος $\theta$ της κατανομής $F(x;\theta)$ σχετίζεται με τις δύο κύριες παραμέτρους $\mu$ και $\sigma^2$. Για παράδειγμα, για την κανονική κατανομή η μέση τιμή $\mu$ και η διασπορά $\sigma^2$ αποτελούν και τις (μοναδικές) παραμέτρους της. Για την ομοιόμορφη κατανομή σε διάστημα $[a,b]$, η σχέση των παραμέτρων της κατανομής $a$ και $b$ με τη μέση τιμή $\mu$ και τη διασπορά $\sigma^2$ δίνεται ως $\mu = \frac{a+b}{2}$ και $\sigma^2 = \frac{(b-a)^2}{12}$.

Γενικά όταν υπάρχει κάποια σχέση που μας επιτρέπει να υπολογίσουμε την παράμετρο $\theta$ (ή τις παραμέτρους $\theta_1, \theta_2$) από τις $\mu$ και $\sigma^2$, τότε υπολογίζουμε πρώτα τις εκτιμήσεις $\bar{x}$ και $s^2$ και μετά υπολογίζουμε τη $\theta$ (ή τις $\theta_1, \theta_2$) από την σχέση, όπου αντικαθιστούμε τις $\mu$ και $\sigma^2$ με τις εκτιμήσεις $\bar{x}$ και $s^2$ αντίστοιχα. Αυτή είναι η *μέθοδος των ροπών* (method of moments). Η ονομασία προκύπτει από τη χρήση των ροπών στην εκτίμηση των παραμέτρων: τη μέση τιμή $\mu$ που είναι η πρώτη ροπή και τη διασπορά $\sigma^2$ που είναι η δεύτερη κεντρική ροπή. Αν οι δύο αυτές ροπές δεν επαρκούν, δηλαδή έχουμε να εκτιμήσουμε περισσότερες από δύο παραμέτρους ή οι σχέσεις δε δίνουν μοναδικότητα λύσης για τις παραμέτρους, χρησιμοποιούμε και ροπές μεγαλύτερου βαθμού, αλλά δε θα ασχοληθούμε με τέτοια προβλήματα.

**Παράδειγμα 2.1** *Μια εταιρεία Α παράγει ασφάλειες των 40 αμπέρ και θέλουμε να μελετήσουμε κατά πόσο πράγματι καίγονται οι ασφάλειες σε ένταση ρεύματος 40 αμπέρ όπως είναι η ένδειξη τους. Στον Πίνακα 2.1 δίνονται οι μετρήσεις έντασης του ηλεκτρικού ρεύματος στις οποίες κάηκαν 25 ασφάλειες που δοκιμάσαμε (ο πίνακας περιέχει και ίδιου τύπου δεδομένα γι ασφάλειες από άλλη εταιρεία που θα μελετήσουμε αργότερα). Στον Πίνακα 2.1 δίνεται επίσης το άθροισμα των τιμών, τα τετράγωνα των τιμών καθώς και το άθροισμα των τετραγώνων των τιμών. Για το όριο έντασης ηλεκτρικού ρεύματος ασφάλειας 40 αμπέρ της εταιρείας Α η δειγματική μέση τιμή είναι σύμφωνα με τη σχέση (2.1)*

$$\bar{x} = \frac{1}{25}\sum_{i=1}^{25} x_i = \frac{1}{25}995.1 = 39.80$$

*Η δειγματική διασπορά είναι σύμφωνα με τη σχέση (2.4)*

$$s^2 = \frac{1}{24}\left(\sum_{i=1}^{25} x_i^2 - 25\bar{x}^2\right) = \frac{1}{24}(39629 - 25 \cdot 39.80^2) = 0.854.$$

*Με βάση αυτό το δείγμα η εκτίμηση της μέση τιμής $\mu$ είναι $\bar{x} = 39.80$ αμπέρ και της διασποράς $\sigma^2$ είναι $s^2 = 0.854\,$(αμπέρ)$^2$.*

*Αν η τ.μ. $X$ (το όριο έντασης ηλεκτρικού ρεύματος ασφάλειας 40 αμπέρ) ακολουθεί κανονική κατανομή $\mathrm{N}(\mu, \sigma^2)$, τότε είναι φανερό πως αυτές οι εκτιμήσεις περιγράφουν πλήρως την κανονική κατανομή της $X$ με βάση αυτό το δείγμα.*

*Αν η τ.μ. $X$ ακολουθεί ομοιόμορφη κατανομή σε κάποιο διάστημα $[a,b]$, τότε μπορούμε να εκτιμήσουμε τις παραμέτρους $a$ και $b$ από τις σχέσεις $\mu = \frac{a+b}{2}$ και $\sigma^2 = \frac{(b-a)^2}{12}$. Αντικαθιστούμε τις $\mu$ και $\sigma^2$ με τις εκτιμήσεις $\bar{x}$ και $s^2$ και λύνουμε το παρακάτω σύστημα εξισώσεων*

$$\begin{array}{rcl} \bar{x} & = & \frac{a+b}{2} \\ s^2 & = & \frac{(b-a)^2}{12} \end{array} \quad \Rightarrow \quad \begin{array}{rcl} \hat{a} & = & \bar{x} - \sqrt{3}s \\ \hat{b} & = & \bar{x} + \sqrt{3}s \end{array} \quad \Rightarrow \quad \begin{array}{l} \hat{a} = 38.20 \\ \hat{b} = 41.40. \end{array}$$

| Α/Α ($i$) | $x_{1i}$ (αμπέρ) | $x_{1i}^2$ | $x_{2i}$ (αμπέρ) | $x_{2i}^2$ |
|---|---|---|---|---|
| 1 | 40.9 | 1672.8 | 41.6 | 1730.6 |
| 2 | 40.3 | 1624.1 | 41.0 | 1681.0 |
| 3 | 39.8 | 1584.0 | 40.5 | 1640.2 |
| 4 | 40.1 | 1608.0 | 40.8 | 1664.6 |
| 5 | 39.0 | 1521.0 | 39.7 | 1576.1 |
| 6 | 41.4 | 1714.0 | 42.1 | 1772.4 |
| 7 | 39.8 | 1584.0 | 40.5 | 1640.2 |
| 8 | 41.5 | 1722.2 | 42.2 | 1780.8 |
| 9 | 40.0 | 1600.0 | 40.7 | 1656.5 |
| 10 | 40.6 | 1648.4 | 41.3 | 1705.7 |
| 11 | 38.3 | 1466.9 | 39.0 | 1521.0 |
| 12 | 39.0 | 1521.0 | 39.7 | 1576.1 |
| 13 | 40.9 | 1672.8 | 41.6 | 1730.6 |
| 14 | 39.1 | 1528.8 | 39.8 | 1584.0 |
| 15 | 40.3 | 1624.1 | 41.0 | 1681.0 |
| 16 | 39.3 | 1544.5 | 40.0 | 1600.0 |
| 17 | 39.6 | 1568.2 | 40.3 | 1624.1 |
| 18 | 38.4 | 1474.6 | 39.1 | 1528.8 |
| 19 | 38.4 | 1474.6 | 39.1 | 1528.8 |
| 20 | 40.7 | 1656.5 | 41.4 | 1714.0 |
| 21 | 39.7 | 1576.1 | | |
| 22 | 38.9 | 1513.2 | | |
| 23 | 38.9 | 1513.2 | | |
| 24 | 40.6 | 1648.4 | | |
| 25 | 39.6 | 1568.2 | | |
| Σύνολο | 995.1 | 39629 | 811.4 | 32937 |

Πίνακας 2.1: Δεδομένα ορίου έντασης ηλεκτρικού ρεύματος για ασφάλειες 40 αμπέρ από δύο εταιρείες: $x_{1i}$ είναι για την εταιρεία Α και $x_{2i}$ για την εταιρεία Β. Στις στήλες 3 και 5 δίνονται και τα τετράγωνα των $x_{1i}$ και $x_{2i}$ και στην τελευταία σειρά δίνεται το σύνολο για κάθε στήλη.

**Μέθοδος της Μεγίστης Πιθανοφάνειας** Η μέθοδος αυτή δίνει την εκτίμηση που έχει τη μέγιστη πιθανοφάνεια, δηλαδή δίνει την τιμή της παραμέτρου η οποία, μεταξύ όλων των δυνατών τιμών της παραμέτρου, είναι η πιο πιθανή με βάση το τυχαίο δείγμα.

Η συνάρτηση πυκνότητας πιθανότητας για κάποια τιμή $X = x_i$ που εξαρτάται από κάποια παράμετρο $\theta$ παίρνει την τιμή $f(x_i; \theta)$ (αν η $X$ είναι διακριτή τότε αυτή η τιμή εκφράζει την πιθανότητα $\mathrm{P}(X = x_i)$). Επειδή $\{x_1, \dots, x_n\}$ είναι ανεξάρτητες, η πιθανότητα να τις παρατηρήσουμε ταυτόχρονα σ' ένα τυχαίο δείγμα μεγέθους $n$ δίνεται από τη **συνάρτηση πιθανοφάνειας** (likelihood function) ως προς $\theta$

$$L(x_1, \dots, x_n; \theta) = f(x_1; \theta) \cdots f(x_n; \theta).$$

Στα προβλήματα εκτίμησης, θεωρούμε τα $\{x_1, \dots, x_n\}$ δεδομένα και ενδιαφερόμαστε για τη

$\theta$. Άν λοιπόν $L(x_1,\ldots,x_n;\theta_1) > L(x_1,\ldots,x_n;\theta_2)$ τότε η $\theta_1$ είναι πιο αληθοφανής από τη $\theta_2$ γιατί δίνει μεγαλύτερη πιθανότητα στα παρατηρούμενα $\{x_1,\ldots,x_n\}$. Θέλουμε λοιπόν να βρούμε την 'πιο αληθοφανή' τιμή της $\theta$, δηλαδή την τιμή $\hat{\theta}$ που μεγιστοποιεί τη $L(x_1,\ldots,x_n;\theta)$ ή καλύτερα, για ευκολότερους υπολογισμούς, τη $\log L(x_1,\ldots,x_n;\theta)$. Άρα η **εκτιμήτρια μεγίστης πιθανοφάνειας** (maximum likelihood estimator) $\hat{\theta}$ βρίσκεται μηδενίζοντας την παράγωγο της $\log L$ ως προς $\theta$

$$\frac{\partial \log L(x_1,\ldots,x_n;\theta)}{\partial\theta} = 0. \tag{2.7}$$

Άν θέλουμε να εκτιμήσουμε δύο ή περισσότερες παραμέτρους $\theta_1,\ldots,\theta_m$, η συνάρτηση πιθανοφάνειας είναι $L(x_1,\ldots,x_n;\theta_1,\ldots,\theta_m)$ και οι εκτιμήτριες $\hat{\theta}_1,\ldots,\hat{\theta}_m$ βρίσκονται λύνοντας το σύστημα των $m$ εξισώσεων

$$\frac{\partial \log L(x_1,\ldots,x_n;\theta_1,\ldots,\theta_m)}{\partial\theta_j} = 0 \quad \text{για} \quad j=1,\ldots,m. \tag{2.8}$$

**Παράδειγμα 2.2** *Έχουμε ένα τυχαίο δείγμα $\{x_1,\ldots,x_n\}$ από κανονική κατανομή $N(\mu,\sigma^2)$ και θέλουμε να εκτιμήσουμε τη μέση τιμή $\mu$ θεωρώντας τη $\sigma^2$ γνωστή. Η συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής είναι*

$$f_X(x;\mu) \equiv f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

*Η συνάρτηση πιθανοφάνειας (για την οποία μόνο η παράμετρος $\mu$ είναι άγνωστη) είναι*

$$\begin{aligned} L(x_1,\ldots,x_n;\mu) &= \left(\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_1-\mu)^2}{2\sigma^2}}\right)\cdot\left(\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_2-\mu)^2}{2\sigma^2}}\right)\cdots\left(\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_n-\mu)^2}{2\sigma^2}}\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right], \end{aligned}$$

*όπου $\exp(x) \equiv e^x$. Ο λογάριθμος της συνάρτησης πιθανοφάνειας είναι*

$$\log L(x_1,\ldots,x_n;\mu) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

*Η εκτιμήτρια μεγίστης πιθανοφάνειας $\hat{\mu}$ βρίσκεται μηδενίζοντας την παράγωγο της $\log L$ (σχέση (2.7))*

$$\frac{\partial \log L}{\partial\mu} = 0 \quad \Rightarrow \quad \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \tag{2.9}$$

*που δίνει τη λύση*

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},$$

*δηλαδή είναι ίδια με την εκτιμήτρια $\bar{x}$ της μέσης τιμής $\mu$ που ορίσαμε για οποιαδήποτε κατανομή της τ.μ. $X$.*

**Παράδειγμα 2.3** *Άν υποθέσουμε στο προηγούμενο παράδειγμα πως κι η διασπορά $\sigma^2$ είναι άγνωστη, τότε στην παραπάνω εξίσωση $\frac{\partial \log L}{\partial \mu} = 0$ προστίθεται κι η εξίσωση*

$$\frac{\partial \log L}{\partial \sigma^2} = 0 \quad \Rightarrow \quad -\frac{n}{2\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0. \tag{2.10}$$

*Η επίλυση του συστήματος των δύο εξισώσεων (2.9) και (2.10) ως προς $\mu$ και $\sigma^2$ δίνει την ίδια λύση για τη $\mu$ και για τη $\sigma^2$ είναι*

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

*Οι εκτιμήτριες μεγίστης πιθανοφάνειας λοιπόν για τη μέση τιμή $\mu$ και τη διασπορά $\sigma^2$ μιας τ.μ. που ακολουθεί κανονική κατανομή είναι απλά η δειγματική μέση τιμή και διασπορά αντίστοιχα, αλλά για τη διασπορά έχουμε τη 'μεροληπτική' δειγματική διασπορά $\tilde{s}^2$ (σχέση (2.3)). Ασυμπτωτικά όμως (για μεγάλο $n$) η εκτιμήτρια $\hat{\sigma^2} = \tilde{s}^2$ είναι αμερόληπτη.*

Η μέθοδος μεγίστης πιθανοφάνειας είναι η καλύτερη μέθοδος εκτίμησης αν γνωρίζουμε την κατανομή της τ.μ. $X$ και μπορεί να εφαρμοσθεί σε οποιοδήποτε πρόβλημα εκτίμησης παραμέτρων.

**Μέθοδος των Ελαχίστων Τετραγώνων** Αυτή η μέθοδος εφαρμόζεται στην περίπτωση που οι άγνωστες παράμετροι εμφανίζονται σε σχέσεις τυχαίων μεταβλητών και οι σχέσεις αυτές είναι *γραμμικές* ως προς τις παραμέτρους που θέλουμε να εκτιμήσουμε. Μια απλή περίπτωση είναι να έχουμε μια τ.μ. $Y$ και η κάθε τιμή της $y$ να δίνεται από τη σχέση

$$y = \theta_1 x_1 + \cdots + \theta_m x_m + \epsilon,$$

όπου οι τιμές $x_1, \ldots, x_m$ είναι γνωστές, $\theta_1, \ldots, \theta_m$ είναι οι άγνωστες παράμετροι και $\epsilon$ είναι μια άλλη τ.μ. με $\mathrm{E}(\epsilon) = 0$. Θα ασχοληθούμε με τη μέθοδο αυτή στο Κεφάλαιο 2.

**Παρατηρήσεις**
Η μέθοδος μεγίστης πιθανοφάνειας μπορεί να εφαρμοσθεί για οποιοδήποτε $\theta$ αν γνωρίζουμε την κατανομή $F_X(x;\theta)$ ενώ η μέθοδος των ροπών δεν εφαρμόζεται αν η $\theta$ δε μπορεί να υπολογισθεί από τις ροπές.

Η εκτίμηση μεγίστης πιθανοφάνειας έχει όλες τις ιδιότητες καλής εκτιμήτριας, δηλαδή είναι αμερόληπτη (ασυμπτωτικά, δηλαδή η μεροληψία $b(\theta)$ τείνει στο μηδέν για μεγάλα $n$), συνεπής, αποτελεσματική κι επαρκής. Γι αυτό κι αυτή η μέθοδος είναι η καλύτερη μέθοδος εκτίμησης παραμέτρων αν γνωρίζουμε την κατανομή $F_X(x;\theta)$.

## 2.2 Εκτίμηση Διαστήματος Εμπιστοσύνης

Η σημειακή εκτίμηση $\hat{\theta}$ από κάποιο δείγμα δεν περιέχει καμιά πληροφορία για την ακρίβεια της εκτίμησης της $\theta$. Είναι ένας αριθμός που δε γνωρίζουμε πόσο κοντά είναι στην πραγματική τιμή της $\theta$ κι αλλάζει με το δείγμα. Για παράδειγμα, υπολογίζουμε τη δειγματική μέση τιμή $\bar{x}$

από ένα τυχαίο δείγμα μεγέθους $n$. Άν πάρουμε ένα άλλο τυχαίο δείγμα ίδιου μεγέθους, η τιμή της $\bar{x}$ θα είναι διαφορετική. Μπορεί να είναι πιο κοντά ή πιο μακριά στην πραγματική τιμή της $\mu$ απ' ότι αυτή από το προηγούμενο δείγμα. Γενικά λοιπόν η εκτιμήτρια $\hat{\theta}$ εξαρτάται από το δείγμα και είναι η ίδια τ.μ. με κάποια κατανομή. Γι αυτό στην εκτίμηση της $\theta$ είναι σημαντικό εκτός από τη σημειακή εκτίμηση $\hat{\theta}$ να υπολογίσουμε και διάστημα $[\theta_1, \theta_2]$ που να μπορούμε να πούμε με μεγάλη εμπιστοσύνη ότι θα περιέχει την πραγματική τιμή της παραμέτρου $\theta$. Στη συνέχεια θα δούμε τέτοια διαστήματα εμπιστοσύνης για διάφορες παραμέτρους, αρχίζοντας από τη μέση τιμή.

### 2.2.1 Διάστημα εμπιστοσύνης της μέσης τιμής $\mu$

Η σημειακή εκτίμηση της μέσης τιμής $\mu$ μιας τ.μ. $X$ είναι η δειγματική μέση τιμή $\bar{x}$ (σχέση (2.1)) που είναι κι αμερόληπτη εκτιμήτρια της $\mu$, δηλαδή η μέση τιμή της $\bar{x}$ είναι η πραγματική τιμή $\mu$ που είναι άγνωστη όμως σε μας, $\mu_{\bar{x}} \equiv \mathrm{E}(\bar{x}) = \mu$. Παρ' όλο που η εκτιμήτρια $\bar{x}$ είναι διαφορετική από δείγμα σε δείγμα, επειδή η $\bar{x}$ είναι συνεπής εκτιμήτρια όταν αυξάνεται το μέγεθος $n$ του δείγματος πλησιάζει τη μέση τιμή $\mu$. Η διασπορά λοιπόν της $\bar{x}$ θα πρέπει να εξαρτάται από το $n$. Πράγματι για τη διασπορά $\sigma_{\bar{x}}^2$ έχουμε βρει στη σχέση (2.6)

$$\sigma_{\bar{x}}^2 \equiv \mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n},$$

δηλαδή η διασπορά της εκτιμήτριας $\bar{x}$ είναι ανάλογη της διασποράς $\sigma^2$ της $X$ κι αντιστρόφως ανάλογη του αριθμού των παρατηρήσεων $n$. Στην παραπάνω σχέση υποθέτουμε πως οι παρατηρήσεις $\{x_1, \ldots, x_n\}$ είναι ανεξάρτητες (εννοώντας και πάλι τις τ.μ. $X_1, \ldots, X_n$). Την τυπική απόκλιση (τετραγωνική ρίζα της διασποράς) $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ της $\bar{x}$ θα την ονομάζουμε **σταθερό σφάλμα** (standard error), γιατί ορίζει το τυπικό σφάλμα εκτίμησης της $\mu$ με τη $\bar{x}$.

Η τ.μ. $\bar{x}$ λοιπόν έχει κάποια κατανομή με μέση τιμή $\mu_{\bar{x}} = \mu$ και διασπορά $\sigma_{\bar{x}}^2 = \sigma^2/n$. Με βάση την κατανομή της $\bar{x}$ θέλουμε να υπολογίσουμε το διάστημα εμπιστοσύνης της $\mu$. Για να ορίσουμε την κατανομή της $\bar{x}$ χρειάζεται να ελέγξουμε το μέγεθος του δείγματος, αν η κατανομή της $X$ είναι κανονική κι αν γνωρίζουμε τη διασπορά της. Στη συνέχεια θα ορίσουμε την κατανομή της $\bar{x}$ και το διάστημα εμπιστοσύνης για τη $\mu$ που προκύπτει κάτω από τις διάφορες συνθήκες.

**Γνωστή διασπορά** Θεωρούμε εδώ ότι γνωρίζουμε τη διασπορά $\sigma^2$ της τ.μ. $X$ στον πληθυσμό. Για την κατανομή της $\bar{x}$ διακρίνουμε δύο περιπτώσεις:

1. Αν η τ.μ. $X$ ακολουθεί κανονική κατανομή $\mathrm{N}(\mu, \sigma^2)$ **ή** το δείγμα είναι μεγάλο ($n > 30$) τότε και η τ.μ. $\bar{x}$ ακολουθεί κανονική κατανομή $\mathrm{N}(\mu, \sigma^2/n)$.

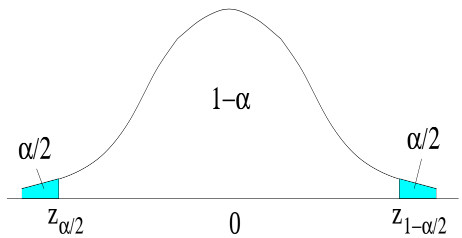
2. Αν δε συμβαίνει το παραπάνω, δηλαδή αν η τ.μ. $X$ δεν ακολουθεί κανονική κατανομή **και** το δείγμα είναι μικρό τότε γενικά δε γνωρίζουμε την κατανομή της $\bar{x}$.

Αν το δείγμα είναι μεγάλο η κανονική κατανομή της $\bar{x}$ δίνεται από το *Κεντρικό Οριακό Θεώρημα*. Αν η τ.μ. $X$ ακολουθεί κανονική κατανομή με γνωστή διασπορά $\sigma^2$ τότε και το άθροισμα τέτοιων τ.μ. $x_1, \ldots, x_n$ ακολουθεί κανονική κατανομή κι έτσι προκύπτει πως και η $\bar{x}$ ακολουθεί κανονική κατανομή.

Υποθέτοντας ότι η $\bar{x}$ ακολουθεί κανονική κατανομή $\mathrm{N}(\mu, \sigma^2/n)$, η τ.μ. $z$ που προκύπτει από τον απλό μετασχηματισμό $z \equiv \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ ακολουθεί την τυπική κανονική κατανομή

$$\bar{x} \sim \mathrm{N}(\mu, \sigma^2/n) \quad \Rightarrow \quad z \equiv \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1). \tag{2.11}$$

Για την τυπική κανονική κατανομή μπορούμε να ορίσουμε ένα διάστημα $[z_{\alpha/2}, z_{1-\alpha/2}]$, στο οποίο θα ανήκει η $z$ με κάποια δοθείσα πιθανότητα $1-\alpha$, όπως φαίνεται στο Σχήμα 2.1. Τα




Σχήμα 2.1: Συνάρτηση πυκνότητας πιθανότητας της τυπικής κανονικής κατανομής και οι ουρές της για κάποιο $\alpha$.

άκρα του διαστήματος, $z_{\alpha/2}$ και $z_{1-\alpha/2}$, λέγονται *κρίσιμες τιμές*. Οι δείκτες $\alpha/2$ και $1-\alpha/2$ δηλώνουν τις τιμές της αθροιστικής συνάρτησης για $z_{\alpha/2}$ και $z_{\alpha/2}$ αντίστοιχα, δηλαδή ισχύει

$$\begin{aligned} \Phi(z_{\alpha/2}) &= \mathrm{P}(z < z_{\alpha/2}) = \alpha/2 \\ \Phi(z_{1-\alpha/2}) &= \mathrm{P}(z < z_{1-\alpha/2}) = 1-\alpha/2 \end{aligned}$$

όπου $\Phi(z)$ είναι η αθροιστική συνάρτηση της τυπικής κανονικής κατανομής. Άρα η πιθανότητα να είναι $z < z_{\alpha/2}$ και $z > z_{1-\alpha/2}$ είναι $\alpha$. Οι δύο σκιασμένες περιοχές στο Σχήμα 2.1 κατέχουν μαζί ποσοστό $\alpha\%$ του συνολικού εμβαδού του ολοκληρώματος της συνάρτησης πυκνότητας πιθανότητας. Αντίστοιχα η πιθανότητα να συμβαίνει $z \in [z_{\alpha/2}, z_{1-\alpha/2}]$ είναι $1-\alpha$. Γενικά λοιπόν ισχύει

$$\mathrm{P}(z_{\alpha/2} < z \leq z_{1-\alpha/2}) = \Phi(z_{1-\alpha/2}) - \Phi(z_{\alpha/2}) = 1-\alpha.$$

Επειδή η συνάρτηση πυκνότητας πιθανότητας της τυπικής κανονικής κατανομής είναι συμμετρική ως προς το 0 ισχύει

$$z_{\alpha/2} = -z_{1-\alpha/2}.$$

Άρα στην ουσία για να ορίσουμε το διάστημα $[z_{\alpha/2}, z_{1-\alpha/2}]$ χρειαζόμαστε μία μόνο κρίσιμη τιμή.

Ανακεφαλαιώνοντας, βρήκαμε ότι σε κάθε δοθείσα πιθανότητα $1-\alpha$ αντιστοιχεί ένα διάστημα τιμών της τ.μ. $z$ που ορίζεται από την κρίσιμη τιμή $z_{1-\alpha/2}$ που την υπολογίζουμε ως $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ από τον στατιστικό πίνακα της τυπικής κανονικής κατανομής. Για παράδειγμα για $\alpha = 0.05$ το διάστημα $[-1.96, 1.96]$ περιέχει την τ.μ. $z$ με πιθανότητα $0.95$, όπου $z_{0.975} = \Phi^{-1}(0.975) = 1.96$. [Ο συμβολισμός που χρησιμοποιούμε εδώ, όπως και για άλλες κατανομές που θα δούμε παρακάτω, βασίζεται στην αρχή της αντιστοίχισης του δείκτη της κρίσιμης τιμής στην τιμή της αντίστοιχης αθροιστικής συνάρτησης. Σε κάποια βιβλία η θετική κρίσιμη τιμή $z_{1-\alpha/2}$ συμβολίζεται ως $z_{\alpha/2}$ και η αρνητική κρίσιμη τιμή $z_{\alpha/2}$ συμβολίζεται ως $-z_{\alpha/2}$].

Θέλουμε να μετασχηματίσουμε το διάστημα $[z_{\alpha/2}, z_{1-\alpha/2}]$ για πιθανότητα $1-\alpha$ στο αντίστοιχο διάστημα που περιέχει την παράμετρο $\mu$. Γι αυτό λύνουμε τις σχέσεις

$$z_{\alpha/2} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \qquad z_{1-\alpha/2} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

ως προς $\mu$ και βρίσκουμε τα άκρα του διαστήματος για τη μέση τιμή $\mu$

$$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \rightarrow \quad \left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]. \tag{2.12}$$

Το διάστημα αυτό υπολογίστηκε για κάποια δοθείσα πιθανότητα $1-\alpha$ που είναι το προκαθορισμένο **επίπεδο εμπιστοσύνης** (confidence level, λέγεται και στάθμη εμπιστοσύνης) και λέγεται **διάστημα εμπιστοσύνης (δ.ε.)** (confidence interval) της $\mu$ σε επίπεδο $1-\alpha$. Η ερμηνεία αυτού του διαστήματος θέλει προσοχή. Σε μια πρώτη προσέγγιση θα λέγαμε ότι σημαίνει 'με πιθανότητα (εμπιστοσύνη) $1-\alpha$ η μέση τιμή $\mu$ βρίσκεται μέσα σ' αυτό το διάστημα', το οποίο δεν είναι ορθό αφού η $\mu$ είναι σταθερό μέγεθος κι όχι τ.μ. για να μιλάμε για την τιμή της με πιθανότητες. Το μέγεθος που αλλάζει (ανάλογα με το δείγμα) είναι το διάστημα και σ' αυτό πρέπει να αναφέρεται η πιθανότητα ή εμπιστοσύνη. Σωστότερη λοιπόν ερμηνεία είναι ότι 'αν χρησιμοποιούσαμε πολλά τέτοια διαστήματα από διαφορετικά δείγματα, ποσοστό $(1-\alpha)\%$ απ' αυτά θα περιείχαν τη $\mu$' ή 'με $1-\alpha$ πιθανότητα (εμπιστοσύνη) το διάστημα αυτό θα περιέχει την πραγματική $\mu$'.

Συνοπτικά η διαδικασία για τον προσδιορισμό του διαστήματος εμπιστοσύνης της $\mu$ είναι

Επιλογή του επιπέδου εμπιστοσύνης $1-\alpha$

$\Downarrow$

Υπολογισμός του $z_{1-\alpha/2}$ από τον αντίστοιχο στατιστικό
πίνακα για την τυπική κανονική κατανομή

$\Downarrow$

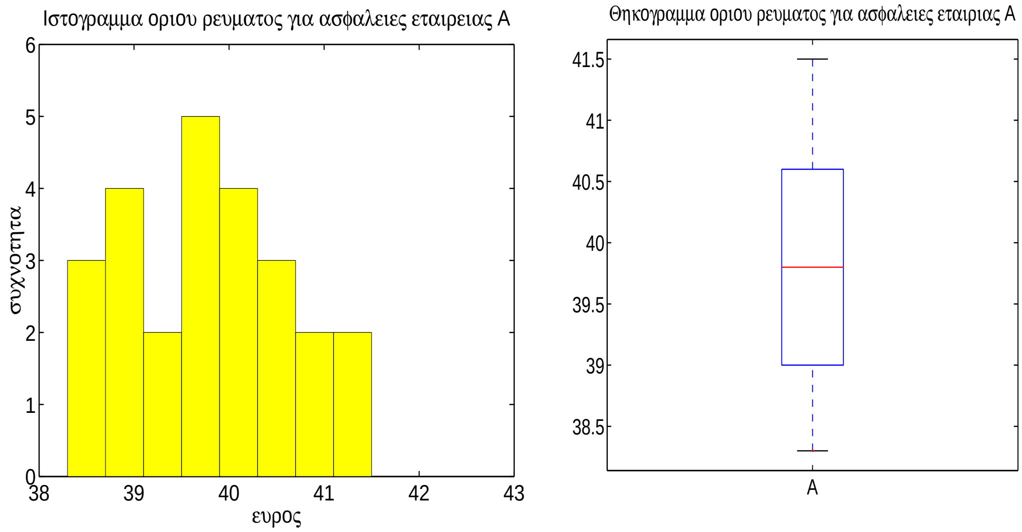
Υπολογισμός του διαστήματος $[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$, όπου

- το $\sigma$ είναι γνωστό,
- το $\bar{x}$ υπολογίζεται από το δείγμα των $n$ παρατηρήσεων.

**Παράδειγμα 2.4** *Θέλουμε να εκτιμήσουμε διάστημα εμπιστοσύνης σε επίπεδο $95\%$ για το μέσο όριο έντασης ηλεκτρικού ρεύματος για ασφάλειες των 40 αμπέρ που παράγει η εταιρία Α από τα δεδομένα του Πίνακα 2.1. Υποθέτουμε ότι από παλιότερες μετρήσεις γνωρίζουμε ότι η διασπορά για αυτό το όριο είναι $\sigma^2 = 1$ (αμπέρ)$^2$.*

*Το δείγμα είναι μικρό ($n = 25 < 30$). Εξετάζουμε τη δειγματική κατανομή του ορίου έντασης ηλεκτρικού ρεύματος από τα δεδομένα μας. Γι αυτό σχεδιάζουμε το ιστόγραμμα και το θηκόγραμμα των δεδομένων του Πίνακα 2.1, τα οποία παρουσιάζονται στο Σχήμα 2.2. Από το ιστόγραμμα και το θηκόγραμμα βλέπουμε ότι η κατανομή του ορίου έντασης ηλεκτρικού ρεύματος φαίνεται να είναι κανονική (είναι συμμετρική και δεν έχει μακριές ουρές). Ειδικά για το θηκόγραμμα φαίνεται να τηρούνται τα τρία χαρακτηριστικά που συνιστούν κανονική κατανομή όπως τα θέσαμε στην Παράγραφο 1.2:*

- *Η διάμεσος δεν τείνει προς το $Q_1$ ή το $Q_3$.*
- *Οι μύστακες έχουν περίπου το ίδιο μήκος.*
- *Δεν υπάρχουν ακραίες τιμές ή απομακρισμένες τιμές.*




Σχήμα 2.2: Ιστόγραμμα και θηκόγραμμα των δεδομένων του ορίου έντασης ηλεκτρικού ρεύματος για ασφάλειες της εταιρίας Α του Πίνακα 2.1.

*Άρα μπορούμε να υποθέσουμε ότι το όριο έντασης ηλεκτρικού ρεύματος $X$ ακολουθεί κανονική κατανομή N$(\mu, 1)$ και τότε η δειγματική μέση τιμή $\bar{x}$ του ορίου έντασης ηλεκτρικού ρεύματος ακολουθεί επίσης κανονική κατανομή N$(\mu, 1/25)$, όπου $n = 25$ είναι το μέγεθος του δείγματος.*

*Για την εκτίμηση του δ.ε. της $\mu$ ακολουθούμε τα παρακάτω βήματα:*

1. *Το επίπεδο εμπιστοσύνης είναι $1 - \alpha = 0.95$ ($\alpha = 0.05$).*

2. *Από τον στατιστικό πίνακα έχουμε $z_{0.975} = \Phi^{-1}(0.975) = 1.96$.*

3. *Η διασπορά είναι γνωστή, $\sigma^2 = 1\,(\text{αμπέρ})^2$. Η δειγματική μέση τιμή έχει υπολογιστεί στο Παράδειγμα 2.1 και είναι $\bar{x} = 39.80$ αμπέρ. Το $95\%$ δ.ε. για τη μέση τιμή $\mu$ είναι*

$$39.80 \pm 1.96\frac{1}{5} \quad \rightarrow \quad [39.41, 40.20].$$

*Άρα το 95% δ.ε. του μέσου ορίου έντασης ηλεκτρικού ρεύματος για ασφάλειες των 40 αμπέρ της εταιρίας Α με βάση το δείγμα είναι $[39.41, 40.20]$. Μπορούμε να πούμε ότι η σημειακή εκτίμηση $\bar{x} = 39.80$ είναι αρκετά ακριβής αφού το αντίστοιχο 95% δ.ε. είναι αρκετά μικρό. Επίσης παρατηρούμε ότι το 95% δ.ε. περιέχει το 40, δηλαδή με 95% εμπιστοσύνη μπορούμε να συμπεράνουμε ότι κατά μέσο όρο οι ασφάλειες που παράγει η εταιρία Α διασφαλίζουν την ένδειξή τους και καίγονται πράγματι σε ένταση ηλεκτρικού ρεύματος 40 αμπέρ.*

**Άγνωστη διασπορά** Γενικά η διασπορά $\sigma^2$ της τ.μ. $X$ είναι άγνωστη και την εκτιμούμε από το δείγμα με τη δειγματική διασπορά $s^2$ (π.χ. δες την (2.4) που χρησιμοποιούμε για τους υπολογισμούς).

Μεγάλο $n$

Αν το δείγμα είναι μεγάλο ($n > 30$) η εκτιμήτρια $s^2$ είναι αρκετά ακριβής κι απλά μπορούμε να αντικαταστήσουμε τη διασπορά $\sigma^2$ με τη δειγματική διασπορά $s^2$ στην παραπάνω διαδικασία για να πάρουμε το διάστημα εμπιστοσύνης για τη $\mu$. Άρα για μεγάλο $n$ κι άγνωστη διασπορά η διαδικασία υπολογισμού του δ.ε. της $\mu$ δίνεται ως εξής.

**δ.ε. για $\mu$: άγνωστη διασπορά και $n$ μεγάλο**

Επιλογή του επιπέδου εμπιστοσύνης $1-\alpha$

$\Downarrow$

Υπολογισμός του $z_{1-\alpha/2}$ από τον αντίστοιχο στατιστικό πίνακα για την τυπική κανονική κατανομή

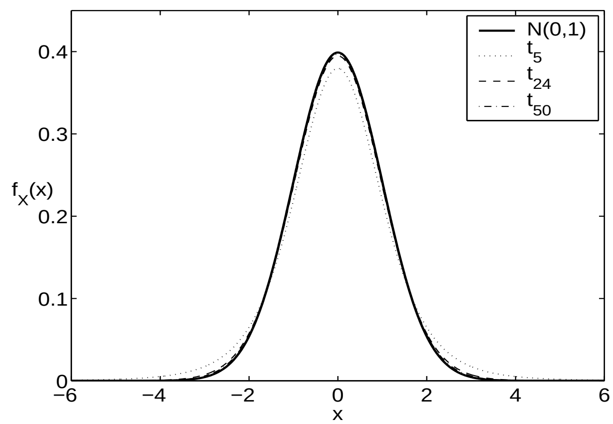
$\Downarrow$

Υπολογισμός του διαστήματος $[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}]$, όπου

- το $s$ υπολογίζεται από το δείγμα των $n$ παρατηρήσεων,
- το $\bar{x}$ υπολογίζεται από το δείγμα των $n$ παρατηρήσεων.

Μικρό $n$

Αν το δείγμα είναι μικρό τότε η προσέγγιση δεν είναι καλή και το διάστημα μπορεί να είναι αρκετά ανακριβές ακόμα κι αν γνωρίζουμε ότι η τ.μ. $X$ ακολουθεί κανονική κατανομή.



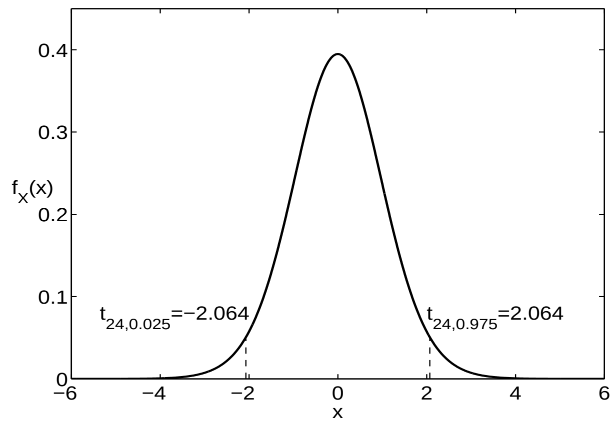

Σχήμα 2.3: Συνάρτηση πυκνότητας πιθανότητας για την τυπική κανονική κατανομή και την κατανομή student με 5, 24 και 50 βαθμούς ελευθερίας, όπως δείχνει το επεξήγημα των γραμμών.

Για μικρό $n$ και υποθέτοντας πως η τ.μ. $X$ ακολουθεί κανονική κατανομή, η τ.μ. $t$ που ορίζεται ως $t \equiv \frac{\bar{x}-\mu}{s/\sqrt{n}}$ ακολουθεί την **κατανομή student** ή **t-κατανομή** με $v = n-1$ βαθμούς ελευθερίας

$$t \equiv \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim \mathrm{t}_{n-1}.$$

κατανομή αυτή μοιάζει με την τυπική κανονική κατανομή και την προσεγγίζει καθώς αυξάνει ο αριθμός των βαθμών ελευθερίας όπως φαίνεται στο Σχήμα 2.3. Για μεγάλο $n$ η προσέγγιση είναι πολύ καλή κι οι τιμές των $t$ και $z$ είναι πρακτικά ίδιες.

Η διαδικασία για την εκτίμηση του διαστήματος εμπιστοσύνης είναι όπως παραπάνω αλλά η κρίσιμη τιμή είναι $t_{n-1,1-\alpha/2}$ αντί για $z_{1-\alpha/2}$ και τη βρίσκουμε από το στατιστικό πίνακα της κατανομής student. [Η κρίσιμη τιμή της $t$ ορίζεται από το $1-\alpha/2$ αλλά και από τους βαθμούς ελευθερίας $v = n-1$. Στο Σχήμα 2.4 παρουσιάζεται η κατανομή student για $v = 24$ και οι κρίσιμες τιμές της για $1-\alpha = 0.95$.] Το διάστημα εμπιστοσύνης σε επίπεδο $1-\alpha$ είναι




Σχήμα 2.4: Συνάρτηση πυκνότητας πιθανότητας για την κατανομή student με 24 βαθμούς ελευθερίας κι η δεξιά κι αριστερή κρίσιμη τιμή της για $1-\alpha = 0.95$.

$$\bar{x} \pm t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \qquad \text{ή} \qquad [\bar{x} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}, \, \bar{x} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}]. \tag{2.13}$$

Αναλυτικά η διαδικασία υπολογισμού του δ.ε. της $\mu$ δίνεται ως εξής.

**δ.ε. για $\mu$: άγνωστη διασπορά, μικρό $n$ και $X \sim \mathbf{N}(\mu, \sigma^2)$**

Επιλογή του επιπέδου εμπιστοσύνης $1-\alpha$

$\Downarrow$

Υπολογισμός του $t_{n-1,1-\alpha/2}$ από τον αντίστοιχο στατιστικό πίνακα για την κατανομή student

$\Downarrow$

Υπολογισμός του διαστήματος $\bar{x} \pm t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}$, όπου

- το $s$ υπολογίζεται από το δείγμα των $n$ παρατηρήσεων.
- το $\bar{x}$ υπολογίζεται από το δείγμα των $n$ παρατηρήσεων.

**Παράδειγμα 2.5** *Για το προηγούμενο παράδειγμα 2.4 υποθέτουμε πως η διασπορά είναι άγνωστη (που είναι και η πιο πιθανή περίπτωση για το πραγματικό πρόβλημα). Το δείγμα είναι αρκετά μεγάλο ($n = 25$) κι ίσως θα μπορούσαμε να υποθέσουμε πως η $\bar{x}$ σαν τ.μ. ακολουθεί κανονική κατανομή και να χρησιμοποιήσουμε στους υπολογισμούς του διαστήματος εμπιστοσύνης την κρίσιμη τιμή $z_{1-\alpha/2}$. Για να εκτιμήσουμε όμως το διάστημα εμπιστοσύνης για το μέσο όριο έντασης ηλεκτρικού ρεύματος της ασφάλειας με την καλύτερη δυνατή ακρίβεια χρησιμοποιούμε την κρίσιμη τιμή $t_{n-1,1-\alpha/2}$ από την κατανομή student.*

*Στο Παράδειγμα 2.1 είχαμε υπολογίσει από το δείγμα των 25 ασφαλειών τη δειγματική μέση $\bar{x} = 39.80$ αμπέρ και τη δειγματική διασπορά $s^2 = 0.854\,(\text{αμπέρ})^2$. Από το στατιστικό πίνακα για την κατανομή student, για $1-\alpha/2 = 0.975$ και $n-1 = 24$, βρίσκουμε $t_{24,0.975} = 2.064$ και το δ.ε. για τη $\mu$ είναι*

$$39.80 \pm 2.064\frac{\sqrt{0.854}}{5} \quad \rightarrow \quad [39.42,\, 40.18].$$

*Η χρήση της κρίσιμης τιμής $t_{24,\,0.975} = 2.064$ υπαγορεύεται από τη χρήση της εκτίμησης $s^2$ αντί της πραγματικής διασποράς $\sigma^2$ που δεν τη γνωρίζουμε. Αν αποφασίζαμε εσφαλμένα να χρησιμοποιήσουμε στους παραπάνω υπολογισμούς τη $z_{0.975} = 1.96$ της κανονικής κατανομής θα βρίσκαμε το διάστημα $[39.44, 40.16]$ πού είναι ελάχιστα μικρότερο (αφού $z_{0.975} < t_{24,\,0.975}$). Το διάστημα όμως αυτό δεν είναι ακριβές γιατί κάναμε την υπόθεση για κανονική κατανομή της $\bar{x}$ που δεν εφαρμόζεται στην περίπτωση που η διασπορά είναι άγνωστη και το δείγμα είναι μικρό.*

Γενικά όταν τον $n$ δεν είναι μεγάλο το διάστημα εμπιστοσύνης της $\bar{x}$ που παίρνουμε υποθέτοντας ότι η $\bar{x}$ ακολουθεί κατανομή student είναι πιό μεγάλο από αυτό που παίρνουμε υποθέτοντας ότι η $\bar{x}$ ακολουθεί κανονική κατανομή και η διασπορά παραμένει η ίδια.

**Μη κανονική κατανομή και μικρά δείγματα** Σε κάποιες περιπτώσεις το δείγμα μπορεί να είναι μικρό **και** η κατανομή της τ.μ. $X$ που παρατηρούμε να μην είναι κανονική. [Όταν δε ξέρουμε τίποτε για την κατανομή της $X$ αυτό το εκτιμούμε από τα δεδομένα του δείγματος, π.χ. από τη μορφή του ιστογράμματος των δεδομένων.] Σε μια τέτοια περίπτωση (κι ανεξάρτητα από το αν η διασπορά είναι γνωστή ή όχι) δε μπορούμε να υποθέσουμε πως η $\bar{x}$ ακολουθεί κάποια συγκεκριμένη κατανομή και στη συνέχεια να εκτιμήσουμε διάστημα εμπιστοσύνης για τη μέση τιμή $\mu$.

Όταν η κατανομή δεν είναι κανονική δεν είναι και συμμετρική. Σε τέτοιες περιπτώσεις ενδιαφερόμαστε για τη διάμεσο $\delta = \delta_X$ της τ.μ. $X$ αντί της $\mu = \mu_X$. Η σημειακή εκτίμηση $\tilde{x}$ της $\delta$ είναι απλά η κεντρική τιμή των παρατηρήσεων διαταγμένες σε αύξουσα σειρά (δες Παράγραφο 1.2.1). Το διάστημα εμπιστοσύνης για τη $\delta$ βρίσκεται με τη μέθοδο Wilcoxon που βασίζεται στην τάξη των παρατηρήσεων παρά στις τιμές τους. Αυτή η εκτίμηση διαστήματος εμπιστοσύνης λέγεται **μη παραμετρική** (επειδή δεν υποθέτουμε κάποια κατανομή και τις παραμέτρους της για την εκτιμήτρια).

**Παρατηρήσεις**

Γενικά θα θέλαμε το διάστημα εμπιστοσύνης να είναι όσο το δυνατό μικρότερο για να έχουμε όσο το δυνατόν πιο ακριβή σημειακή εκτίμηση. Το διάστημα εμπιστοσύνης εξαρτάται από:

- την κατανομή και τη διασπορά $\sigma^2$ της τ.μ. $X$ [αυτά δε μπορούμε να τα αλλάξουμε, είναι χαρακτηριστικά της τ.μ. που μελετάμε].
- το μέγεθος $n$ του δείγματος [αύξηση του $n$ έχει σαν αποτέλεσμα τη μείωση του εύρους του δ.ε., που είναι φυσικά επιθυμητό αλλά όχι πάντοτε εφικτό, αφού η απόκτηση πολλών παρατηρήσεων μπορεί να είναι εργασία επίπονη και πολυέξοδη].
- το επίπεδο εμπιστοσύνης $1-\alpha$ [αυτό το καθορίζουμε εμείς, αλλά δε θα θέλαμε να μικρύνουμε το διάστημα μειώνοντας την εμπιστοσύνη μας σ' αυτό γιατί τότε τα αποτελέσματά μας δε θα είχαν την επιθυμητή στατιστική σημαντικότητα]. Το επίπεδο εμπιστοσύνης που συνήθως χρησιμοποιείται στην πράξη είναι 95%.

Στον Πίνακα 2.2 συνοψίζεται η εκτίμηση του διαστήματος εμπιστοσύνης της $\mu$ στις διάφορες περιπτώσεις.

| *διασπορά* | *κατανομή της $X$* | $n$ | *κατανομή της $\bar{x}$* | *διάστημα εμπιστοσύνης* |
|---|---|---|---|---|
| γνωστή | κανονική | | $z \equiv \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$ | $\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ |
| γνωστή | μη κανονική | μεγάλο | $z \equiv \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$ | $\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ |
| γνωστή | μη κανονική | μικρό | — | — |
| άγνωστη | | μεγάλο | $z \equiv \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim \mathrm{N}(0,1)$ | $\bar{x} \pm z_{1-\alpha/2}\frac{s}{\sqrt{n}}$ |
| άγνωστη | κανονική | μικρό | $t \equiv \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim \mathrm{t}_{n-1}$ | $\bar{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$ |
| άγνωστη | μη κανονική | μικρό | — | — |

Πίνακας 2.2: Εκτίμηση του διαστήματος εμπιστοσύνης της $\mu$ ανάλογα με τη γνώση της διασποράς και κατανομής της τ.μ. $X$ και το μέγεθος $n$ του δείγματος.

**Εύρος διαστήματος εμπιστοσύνης** Πολλές φορές πριν να κάνουμε το πείραμα και συλλέξουμε τις μετρήσεις προκαθορίζουμε ένα συγκεκριμένο εύρος για το δ.ε. ή ζητάμε το εύρος του δ.ε. να μην ξεπερνάει κάποιο ανώτατο όριο για να έχουν νόημα τα αποτελέσματα. Για να το πετύχουμε αυτό χωρίς να αλλάξουμε τη σημαντικότητα των στατιστικών αποτελεσμάτων, βρίσκουμε το μέγεθος $n$ του δείγματος που μας δίνει αυτό το εύρος του δ.ε. Αυτό υπολογίζεται θέτοντας το εύρος του δ.ε. ίσο με την τιμή που ζητάμε και λύνοντας την εξίσωση ως προς το $n$. Για παράδειγμα, ας υποθέσουμε πως το δείγμα είναι μικρό και η τ.μ. $X$ ακολουθεί κανονική κατανομή με άγνωστη διασπορά, χρησιμοποιούμε δηλαδή τη σχέση (2.13) για να υπολογίσουμε το δ.ε. της μέσης τιμής $\mu$. Το εύρος του δ.ε. είναι

$$w = 2\,t_{n-1,1-\alpha/2} \cdot \frac{s}{\sqrt{n}} \tag{2.14}$$

και λύνοντας ως προς $n$ βρίσκουμε ότι για να είναι το εύρος του δ.ε. ίσο με $w$ πρέπει το δείγμα να έχει μέγεθος

$$n = \left(2\,t_{n-1,1-\alpha/2} \cdot \frac{s}{w}\right)^2 . \tag{2.15}$$

Στην παραπάνω σχέση η τιμή $t_{n-1,1-\alpha/2}$ δεν είναι γνωστή αφού το $n$ είναι άγνωστο και ζητούμενο. Θα πρέπει λοιπόν να χρησιμοποιήσουμε την τιμή $t_{n-1,1-\alpha/2}$ που συμφωνεί καλύτερα με το αποτέλεσμα για το $n$ από την (2.15). Αν το $n$ παίρνει μεγάλες τιμές το παραπάνω πρόβλημα δεν υφίσταται αφού για μεγάλα $n$ η κρίσιμη τιμή $t_{n-1,1-\alpha/2}$ είναι πρακτικά ίδια με την αντίστοιχη κρίσιμη τιμή της τυπικής κανονικής κατανομής $z_{1-\alpha/2}$. Άρα για μεγάλο $n$ η σχέση (2.15) δίνεται ως

$$n = \left(2\,z_{1-\alpha/2} \cdot \frac{s}{w}\right)^2 . \tag{2.16}$$

**Παράδειγμα 2.6** *Στο προηγούμενο παράδειγμα για το όριο έντασης ηλεκτρικού ρεύματος που καίγονται οι ασφάλειες 40 αμπέρ υπολογίσαμε το 95% δ.ε. του μέσου ορίου έντασης κάνοντας χρήση της κατανομή student και βρήκαμε ότι είναι* $[39.42,\ 40.18]$ *αμπέρ. Το εύρος του δ.ε. είναι*

$w = 40.18 - 39.42 = 0.76$ *αμπέρ. Αν θέλουμε το εύρος του δ.ε. να μην ξεπερνάει* $0.5$ *αμπέρ τότε αντί για 25 ασφάλειες πρέπει να δοκιμάσουμε*

$$n = \left(2 \cdot 1.96 \cdot \frac{\sqrt{0.854}}{0.5}\right)^2 = 52.5 \to 53,$$

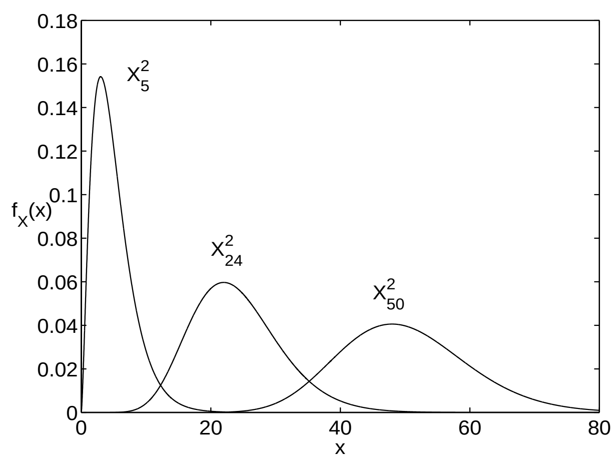
*δηλαδή πρέπει να αυξήσουμε το δείγμα σε 53 ασφάλειες (στρογγυλοποιούμε πάντα στον αμέσως μεγαλύτερο ακέραιο).*

### 2.2.2 Διάστημα εμπιστοσύνης της διασποράς $\sigma^2$

Όπως για να βρούμε διάστημα εμπιστοσύνης για τη μέση τιμή $\mu$ ορίσαμε πρώτα την κατανομή της εκτιμήτριας $\bar{x}$ έτσι και για να βρούμε διάστημα εμπιστοσύνης για τη διασπορά $\sigma^2$ ορίζουμε πρώτα την κατανομή της αμερόληπτης εκτιμήτριας $s^2$ της $\sigma^2$ (σχέση (2.2)). Γνωρίζουμε ότι η $\chi^2 \equiv \frac{(n-1)s^2}{\sigma^2}$ ακολουθεί την κατανομή $\mathcal{X}^2$ με $v = n-1$ βαθμούς ελευθερίας

$$\chi^2 \equiv \frac{(n-1)s^2}{\sigma^2} \sim \mathcal{X}^2_{n-1}.$$

Στο Σχήμα 2.5 παρουσιάζεται η μορφή της κατανομής $\mathcal{X}^2$ για χαρακτηριστικούς βαθμούς ελευθερίας. Για λίγους βαθμούς ελευθερίας η κατανομή $\mathcal{X}^2$ είναι αρκετά λοξή και γίνεται πιο




Σχήμα 2.5: Συνάρτηση πυκνότητας πιθανότητας για την κατανομή $\mathcal{X}^2$ με $5$, $24$ και $50$ βαθμούς ελευθερίας.

συμμετρική καθώς αυξάνουν οι βαθμοί ελευθερίας. Για πολύ μεγάλο $n$ η $\mathcal{X}^2$ προσεγγίζει την κανονική κατανομή.

Για δοθείσα πιθανότητα $1-\alpha$ μπορούμε να βρούμε από τον στατιστικό πίνακα για τη κατανομή $\mathcal{X}^2$ τις δύο κρίσιμες τιμές $\chi^2_{n-1,\alpha/2}$ και $\chi^2_{n-1,1-\alpha/2}$ για τις οποίες ισχύει

$$\mathrm{P}(\chi^2_{n-1,\alpha/2} < \chi^2 < \chi^2_{n-1,1-\alpha/2}) = 1-\alpha. \tag{2.17}$$

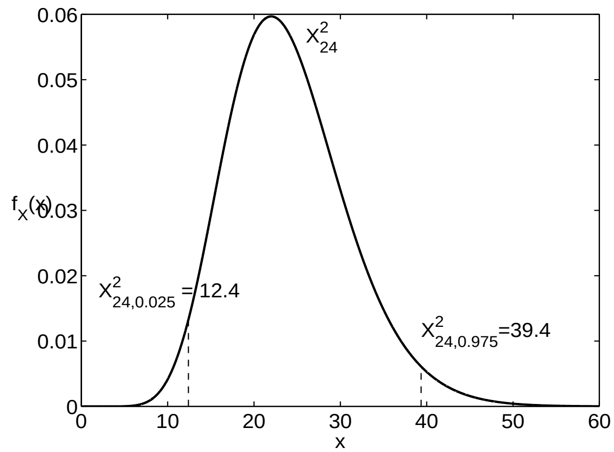
Επειδή η $\mathcal{X}^2$ δεν είναι συμμετρική έχουμε δύο κρίσιμες τιμές: την αριστερή κρίσιμη τιμή $\chi^2_{n-1,\alpha/2}$ που είναι τέτοια ώστε

$$\mathrm{P}(\chi^2 < \chi^2_{n-1,\alpha/2}) = \alpha/2$$

και τη δεξιά κρίσιμη τιμή $\chi^2_{n-1,1-\alpha/2}$ που είναι τέτοια ώστε

$$\mathrm{P}(\chi^2 < \chi^2_{n-1,1-\alpha/2}) = 1 - \alpha/2.$$

[Οι δείκτες $\alpha/2$ και $1-\alpha/2$ των κρίσιμων τιμών είναι κι οι τιμές της αντίστοιχης αθροιστικής συνάρτησης κατανομής.] Στο Σχήμα 2.6 παρουσιάζεται η κατανομή $\mathcal{X}^2$ για $n-1=24$ βαθμούς ελευθερίας καθώς κι οι κρίσιμες τιμές της για $1-\alpha=0.95$.




Σχήμα 2.6: Συνάρτηση πυκνότητας πιθανότητας για την κατανομή $\mathcal{X}^2$ με $24$ βαθμούς ελευθερίας κι η δεξιά κι αριστερή κρίσιμη τιμή της για $1-\alpha=0.95$.

Στη σχέση (2.17), λύνοντας τις δύο ανισότητες

$$\chi^2_{n-1,\alpha/2} < \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{n-1,1-\alpha/2}$$

ως προς $\sigma^2$ βρίσκουμε το $(1-\alpha)\%$ διάστημα εμπιστοσύνη για τη $\sigma^2$

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right], \tag{2.18}$$

όπου η δειγματική διασπορά $s^2$ υπολογίζεται από το δείγμα των $n$ παρατηρήσεων.

Αναλυτικά η διαδικασία υπολογισμού του δ.ε. της $\sigma^2$ δίνεται ως εξής.

**δ.ε. για $\sigma^2$: $X \sim \mathbf{N}(\mu, \sigma^2)$**

Επιλογή του επιπέδου εμπιστοσύνης $1-\alpha$

$\Downarrow$

Υπολογισμός των $\chi^2_{n-1,\alpha/2}$ και $\chi^2_{n-1,1-\alpha/2}$ από τον αντίστοιχο στατιστικό πίνακα για την κατανομή $\mathcal{X}^2$

$\Downarrow$

Υπολογισμός του διαστήματος $\left[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right]$, όπου

- το $s^2$ υπολογίζεται από το δείγμα των $n$ παρατηρήσεων.

Το 95% δ.ε. για την τυπική απόκλιση $\sigma$ έχει σαν άκρα τις τετραγωνικές ρίζες των αντίστοιχων άκρων του 95% δ.ε. για τη διασπορά $\sigma^2$.

**Παράδειγμα 2.7** *Από τα δεδομένα για το όριο έντασης ηλεκτρικού ρεύματος ασφάλειας της εταιρίας Α στον Πίνακα 2.1 θέλουμε να εκτιμήσουμε τη διασπορά $\sigma^2$ του ορίου έντασης. Η σημειακή εκτίμηση βρέθηκε να είναι $s^2 = 0.854\,(αμπέρ)^2$. Για $n-1=24$ και $\alpha = 0.05$ από τον στατιστικό πίνακα για τη $\mathcal{X}^2$ βρίσκουμε $\chi^2_{24,\,0.025} = 12.4$ και $\chi^2_{24,\,0.975} = 39.4$ (δες επίσης Σχήμα 2.6). Το $95\%$ δ.ε. για τη $\sigma^2$ είναι*

$$\left[\frac{24\cdot 0.854}{39.4},\ \frac{24\cdot 0.854}{12.4}\right] = [0.52,\ 1.65].$$

*Το 95% δ.ε. για την τυπική απόκλιση $\sigma$ του ορίου ηλεκτρικού ρεύματος είναι*

$$[\sqrt{0.52},\ \sqrt{1.65}] = [0.72,\ 1.28].$$

Στο παραπάνω παράδειγμα παρατηρούμε ότι η εκτίμηση $s^2 = 0.854\,(\text{αμπέρ})^2$ είναι πιο κοντά στο αριστερό άκρο του διαστήματος εμπιστοσύνης. Γενικά το διάστημα εμπιστοσύνης για τη διασπορά $\sigma^2$ δεν είναι συμμετρικό ως προς τη σημειακή εκτίμηση $s^2$ κι αυτό γιατί η κατανομή $\mathcal{X}^2$ δεν είναι συμμετρική όπως είναι η κανονική κατανομή κι η κατανομή student. Όσο αυξάνουν όμως οι βαθμοί ελευθερίας (δηλαδή το μέγεθος δείγματος) η $\mathcal{X}^2$ κατανομή προσεγγίζει την κανονική κατανομή (δες Σχήμα 2.5). Γι αυτό για πολύ μεγάλα δείγματα το διάστημα εμπιστοσύνης μπορεί να υπολογισθεί κι από άλλο τύπο που περιέχει κρίσιμες τιμές της τυπικής κανονικής κατανομής.

### 2.2.3 Διάστημα εμπιστοσύνης της διαφοράς δύο μέσων τιμών $\mu_1 - \mu_2$

Θέλουμε να εκτιμήσουμε τη διαφορά των μέσων τιμών $\mu_1$ και $\mu_2$ δύο ανεξάρτητων τ.μ. $X_1$ και $X_2$ έχοντας δύο δείγματα μεγέθους $n_1$ και $n_2$ από τον πληθυσμό της $X_1$ και τον πληθυσμό της $X_2$ αντίστοιχα.

Η σημειακή εκτίμηση της διαφοράς $\mu_1 - \mu_2$ είναι απλά η διαφορά των δειγματικών μέσων τιμών $\bar{x}_1 - \bar{x}_2$. Για το διάστημα εμπιστοσύνης της διαφοράς $\mu_1 - \mu_2$ πρέπει να ελέγξουμε την κατανομή της εκτιμήτριας $\bar{x}_1 - \bar{x}_2$, όπως κάναμε στην εκτίμηση διαστήματος εμπιστοσύνης για τη μέση τιμή $\mu$.

**Γνωστές διασπορές** Θεωρούμε ότι γνωρίζουμε τις διασπορές $\sigma_1^2$ και $\sigma_2^2$ των $X_1$ και $X_2$. Υποθέτουμε επίσης ότι τα δείγματα είναι αρκετά μεγάλα **ή** η κατανομή των $X_1$ και $X_2$ είναι κανονική. Τότε η εκτιμήτρια $\bar{x}_1 - \bar{x}_2$ ακολουθεί κανονική κατανομή με μέση τιμή

$$\mu_{\bar{x}_1-\bar{x}_2} \equiv \mathrm{E}(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$$

και διασπορά

$$\sigma^2_{\bar{x}_1-\bar{x}_2} \equiv \mathrm{Var}(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Αν οι $X_1$ και $X_2$ είναι *ομοσκεδαστικές*, δηλαδή $\sigma_1^2 = \sigma_2^2 = \sigma^2$, τότε η διασπορά είναι $\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})$.

Η διαδικασία είναι ίδια όπως για την εύρεση του διαστήματος εμπιστοσύνης για τη μέση τιμή $\mu$ αν κάνουμε τις παρακάτω αντικαταστάσεις:

$$\begin{array}{rccc} \text{εκτιμήτρια} & \bar{x} & \longrightarrow & \bar{x}_1 - \bar{x}_2 \\ \text{μέση τιμή εκτιμήτριας} & \mu & \longrightarrow & \mu_1 - \mu_2 \\ \text{διασπορά εκτιμήτριας} & \dfrac{\sigma^2}{n} & \longrightarrow & \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}. \end{array}$$

Επομένως το διάστημα εμπιστοσύνης της διαφοράς $\mu_1 - \mu_2$ είναι

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \tag{2.19}$$

**δ.ε. για $\mu_1 - \mu_2$: γνωστές διασπορές και είτε $n_1$,$n_2$ μεγάλα ή $X_1 \sim \mathbf{N}(\mu_1, \sigma_1^2)$ και $X_2 \sim \mathbf{N}(\mu_2, \sigma_2^2)$**

Επιλογή του επιπέδου εμπιστοσύνης $1 - \alpha$

$\Downarrow$

Υπολογισμός του $z_{1-\alpha/2}$ από τον αντίστοιχο στατιστικό πίνακα για την τυπική κανονική κατανομή

$\Downarrow$

Υπολογισμός του διαστήματος $(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, όπου

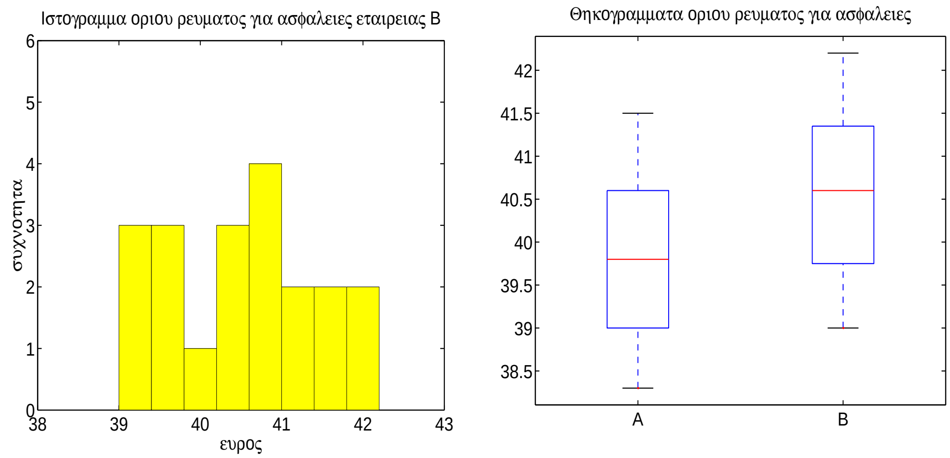
- το $\sigma_1^2$ και $\sigma_2^2$ είναι γνωστά και
- το $\bar{x}_1$ και $\bar{x}_2$ υπολογίζονται από τα δείγματα των $n_1$ και $n_2$ παρατηρήσεων.

Συχνά στην πράξη με την εκτίμηση του διαστήματος εμπιστοσύνης της διαφοράς $\mu_1 - \mu_2$ θέλουμε να διαπιστώσουμε αν κατά μέσο όρο η μια τ.μ. είναι διαφορετική (μεγαλύτερη ή μικρότερη) από την άλλη. Στη συνέχεια, αν διαπιστώσουμε πως η διαφορά είναι στατιστικά σημαντική, το διάστημα εμπιστοσύνης δίνει επίσης εκτίμηση του μεγέθους αυτής της διαφοράς. Το διάστημα εμπιστοσύνης λοιπόν το ερμηνεύουμε ως εξής:

- Αν περιέχει το μηδέν τότε δε μπορούμε να πούμε ότι οι μέσες τιμές των τ.μ. $X_1$ και $X_2$ διαφέρουν με στατιστική σημαντικότητα για το επίπεδο σημαντικότητας $\alpha$ που χρησιμοποιήσαμε και με βάση τα συγκεκριμένα δεδομένα.
- Αν είναι θετικό τότε μπορούμε να πούμε πως για το επίπεδο εμπιστοσύνης $1 - \alpha$ που χρησιμοποιήσαμε η τ.μ. $X_1$ είναι κατά μέσο όρο μεγαλύτερη από τη $X_2$ κατά ένα ποσό που καθορίζεται από τα όρια του διαστήματος που εκτιμήσαμε. Αντίστοιχα ερμηνεύουμε το διάστημα εμπιστοσύνης όταν είναι αρνητικό.

**Παράδειγμα 2.8** *Ενδιαφερόμαστε να διαπιστώσουμε αν το όριο έντασης ηλεκτρικού ρεύματος που καίγονται ασφάλειες 40 αμπέρ της εταιρίας Α είναι κατά μέσο όρο διαφορετικό από το όριο για ίδιου τύπου ασφάλειες κατασκευασμένες από μια άλλη εταιρία Β. Γι αυτό θέλουμε να* *εκτιμήσουμε το $95\%$ δ.ε. για τη διαφορά των μέσων τιμών $\mu_1$ και $\mu_2$ του ορίου έντασης ηλεκτρικού ρεύματος για τις ασφάλειες της εταιρίας Α και της εταιρίας Β αντίστοιχα. Δίνεται ότι η διασπορά είναι κοινή και γνωστή και είναι $\sigma^2 = 1\,(\alpha\mu\pi\acute{\varepsilon}\rho)^2$.*

*Στον Πίνακα 2.1 παρουσιάζονται 25 μετρήσεις του ορίου έντασης ηλεκτρικού ρεύματος για τις ασφάλειες 40 αμπέρ της εταιρίας Α και 20 μετρήσεις σε ασφάλειες της εταιρίας Β. Στο Παράδειγμα 2.4 είδαμε με τη βοήθεια ιστογράμματος και θηκογράμματος (Σχήμα 2.2) πως το όριο έντασης ηλεκτρικού ρεύματος για τις ασφάλειες της εταιρίας Α θα μπορούσε να ακολουθεί κανονική κατανομή. Από το ιστόγραμμα και το θηκόγραμμα του Σχήματος 2.7 μπορούμε να πούμε το ίδιο και για το όριο ηλεκτρικού ρεύματος για τις ασφάλειες της εταιρίας Β. Άρα μπορούμε να υποθέσουμε πως οι τ.μ. του ορίου έντασης ηλεκτρικού ρεύματος $X_1$ και $X_2$ και για τους δύο πληθυσμούς, δηλαδή τις ασφάλειες 40 αμπέρ και των δύο εταιριών, ακολουθούν κανονική κατανομή. Συγκρίνοντας τα θηκογράμματα για τις ασφάλειες των εταιριών Α και Β (δες Σχήμα 2.7) φαίνεται ότι η κεντρική τάση (εδώ διάμεσος) του ορίου έντασης ηλεκτρικού ρεύματος για τις ασφάλειες της εταιρίας Β είναι υψηλότερη από αυτή για τις ασφάλειες της εταιρίας Α, αλλά ίσως όχι σημαντικά αφού οι δύο θήκες (τα διαστήματα των 50% κεντρικών τιμών του κάθε δείγματος) επικαλύπτονται κατά μεγάλο ποσοστό. Στη συνέχεια θα εξετάσουμε αν αυτή η διαφορά είναι στατιστικά σημαντική κάνοντας χρήση του διαστήματος εμπιστοσύνης της $\mu_1 - \mu_2$.*




Σχήμα 2.7: Ιστόγραμμα των δεδομένων του ορίου έντασης ηλεκτρικού ρεύματος για ασφάλειες των 40 αμπέρ της εταιρίας Β του Πίνακα 2.1 και θηκογράμματα των δεδομένων του ορίου έντασης ηλεκτρικού ρεύματος για ασφάλειες των 40 αμπέρ των δύο εταιριών από τον ίδιο πίνακα.

*Οι δειγματικές μέσες τιμές υπολογίζονται σε $\bar{x}_1 = 39.80$ αμπέρ και $\bar{x}_2 = 40.57$ αμπέρ. Η διαφορά τους είναι $\bar{x}_1 - \bar{x}_2 = -0.77$ αμπέρ. Από τη σχέση (2.19), όπου $z_{1-\alpha/2} = z_{0.975} = 1.96$ και $\sigma_1^2 = \sigma_2^2 = \sigma^2 = 1$ βρίσκουμε ότι το $95\%$ δ.ε. είναι*

$$-0.77 \pm 1.96\sqrt{1 \cdot \left(\frac{1}{25} + \frac{1}{20}\right)} \quad \rightarrow \quad [-1.36,\ -0.18].$$

*Συμπεραίνουμε λοιπόν πως σε επίπεδο εμπιστοσύνης $95\%$ μπορούμε να πούμε πως το μέσο όριο έντασης ηλεκτρικού ρεύματος διαφέρει σημαντικά στις ασφάλειες των δύο εταιριών. Συγκεκριμένα οι ασφάλειες των 40 αμπέρ της εταιρίας Α καίγονται κατά μέσο όρο σε μικρότερη ένταση* *ηλεκτρικού ρεύματος από ότι οι ασφάλειες της εταιρίας Β, κατά ποσό μεταξύ 0.18 και 1.36 αμπέρ.*

**Άγνωστες διασπορές** Συνήθως όταν δε γνωρίζουμε τις $\mu_1$, $\mu_2$ αγνοούμε και τις $\sigma_1^2$, $\sigma_2^2$.

Μεγάλο $n$

Όταν τα δείγματα είναι μεγάλα ($n_1 > 30$ και $n_2 > 30$) μπορούμε να αντικαταστήσουμε στη σχέση (2.19) τις $\sigma_1^2$, $\sigma_2^2$ με τις δειγματικές μέσες τιμές $s_1^2$, $s_2^2$ και να εκτιμήσουμε έτσι το διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_2$.

**δ.ε. για $\mu_1 - \mu_2$: άγνωστες διασπορές και $n_1$,$n_2$ μεγάλα**

Επιλογή του επιπέδου εμπιστοσύνης $1-\alpha$

$\Downarrow$

Υπολογισμός του $z_{1-\alpha/2}$ από τον αντίστοιχο στατιστικό πίνακα για την τυπική κανονική κατανομή

$\Downarrow$

Υπολογισμός του διαστήματος $(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, όπου

- το $\bar{x}_1$ και $\bar{x}_2$ υπολογίζονται από τα δείγματα των $n_1$ και $n_2$ παρατηρήσεων.
- το $s_1^2$ και $s_2^2$ υπολογίζονται επίσης από τα δείγματα.

Μικρό $n$

Όταν το μέγεθος του ενός ή και των δύο δειγμάτων είναι μικρό η εκτίμηση του διαστήματος εμπιστοσύνης είναι πιο περίπλοκη. Στην περίπτωση που οι κατανομές των $X_1$ και $X_2$ δε δίνονται να είναι κανονικές (ή δεν προκύπτει από τη περιγραφική μελέτη των δεδομένων), δε μπορούμε γενικά να προσδιορίσουμε την κατανομή της διαφοράς $\bar{x}_1 - \bar{x}_2$ για να εκτιμήσουμε έτσι το διάστημα εμπιστοσύνης και πρέπει να χρησιμοποιήσουμε τη μή-παραμετρική προσέγγιση.

Υποθέτουμε λοιπόν πως οι κατανομές των $X_1$ και $X_2$ είναι κανονικές κι επιπλέον ομοσκεδαστικές ($\sigma_1^2 = \sigma_2^2 = \sigma^2$). Σ' αυτήν την περίπτωση ορίζουμε πρώτα τη *δειγματική κοινή διασπορά* $s_p^2$ ως συνάρτηση των $s_1^2$ και $s_2^2$

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}. \tag{2.20}$$

Μπορεί να δειχθεί ότι η $s_p^2$ είναι αμερόληπτη εκτιμήτρια της κοινής διασποράς $\sigma^2$. Με τη βοήθεια της $s_p^2$ η εκτίμηση της διασποράς της διαφοράς $\bar{x}_1 - \bar{x}_2$ είναι $s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$. Για την εκτιμήτρια $\bar{x}_1 - \bar{x}_2$ μπορούμε να ορίσουμε τον παρακάτω μετασχηματισμό που ακολουθεί κατανομή student με $n_1 + n_2 - 2$ βαθμούς ελευθερίας

$$t \equiv \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathrm{t}_{n_1+n_2-2}$$

από το οποίο προκύπτει πως το $(1-\alpha)\%$ δ.ε. της διαφοράς $\mu_1 - \mu_2$ είναι

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2,1-\alpha/2}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \tag{2.21}$$

**δ.ε. για $\mu_1 - \mu_2$: άγνωστες και ίσες διασπορές, $n_1$,$n_2$ μικρά και $X_1 \sim \mathbf{N}(\mu_1, \sigma_1^2)$ και $X_2 \sim \mathbf{N}(\mu_2, \sigma_2^2)$**

Επιλογή του επιπέδου εμπιστοσύνης $1-\alpha$

$\Downarrow$

Υπολογισμός του $t_{n_1+n_2-2,1-\alpha/2}$ από τον αντίστοιχο στατιστικό πίνακα για την κατανομή student

$\Downarrow$

Υπολογισμός του διαστήματος $(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2,1-\alpha/2}\, s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$, όπου

- το $\bar{x}_1$ και $\bar{x}_2$ υπολογίζονται από τα δείγματα των $n_1$ και $n_2$ παρατηρήσεων.
- το $s_1^2$ και $s_2^2$ υπολογίζονται επίσης από τα δείγματα.

**Παράδειγμα 2.9** *Για το προηγούμενο παράδειγμα ας υποθέσουμε πως η διασπορά του ορίου έντασης ηλεκτρικού ρεύματος για τις ασφάλιες 40 αμπέρ των δύο εταιριών A και B είναι άγνωστη. Από τα δύο ιστογράμματα και θηκογράμματα στα Σχήματα 2.2 και 2.7 μπορούμε να δεχτούμε ότι οι τ.μ. ορίου έντασης ηλεκτρικού ρεύματος για τις ασφάλειες και των δύο εταιριών ακολουθούν κανονική κατανομή και μάλιστα έχουν την ίδια διασπορά (το εύρος των τιμών των δύο δειγμάτων είναι περίπου το ίδιο). Επειδή όμως τα δείγματα είναι σχετικά μικρά θα υποθέσουμε ότι η διαφορά $\bar{x}_1 - \bar{x}_2$ ακολουθεί κατανομή student κι όχι κανονική.*

*Η δειγματική διασπορά του ορίου έντασης ηλεκτρικού ρεύματος για τις ασφάλειες της εταιρίας A είναι $s_1^2 = 0.854\,(\text{αμπέρ})^2$ και για την εταιρεία B είναι $s_2^2 = 0.952\,(\text{αμπέρ})^2$ (σχετικά κοντά). Εφαρμόζοντας τη σχέση (2.20) βρίσκουμε ότι η δειγματική κοινή διασπορά είναι ($n_1 = 25$ και $n_2 = 20$)*

$$s_p^2 = \frac{(25-1)\cdot 0.854 + (20-1)\cdot 0.952}{25+20-2} = 0.947 \text{ αμπέρ}.$$

*Οι βαθμοί ελευθερίας είναι $n_1+n_2-2 = 43$ και για επίπεδο εμπιστοσύνης $95\%$ βρίσκουμε από τον στατιστικό πίνακα για την κατανομή student την κρίσιμη τιμή $t_{43,\,0.975} = 2.02$ (πολύ κοντά στην αντίστοιχη κρίσιμη τιμή $z_{0.975} = 1.96$ της τυπικής κανονικής κατανομής γιατί οι βαθμοί ελευθερίας είναι πολλοί). Το $95\%$ διάστημα εμπιστοσύνης είναι ($\bar{x}_1 - \bar{x}_2 = -0.77$)*

$$-0.77 \pm 2.02 \cdot 0.947\sqrt{\frac{1}{25}+\frac{1}{20}} \quad \rightarrow \quad [-1.35,\ -0.19].$$

*Όπως και πριν που η διασπορά ήταν γνωστή συμπεραίνουμε πως σε επίπεδο εμπιστοσύνης 95% οι ασφάλειες της εταιρίας A κατά μέσο όρο καίγονται σε μικρότερη ένταση ηλεκτρικού ρεύματος απ' ότι οι ασφάλειες της εταιρίας B κατά ένα ποσό που κυμαίνεται μεταξύ 0.19 και 1.35 αμπέρ.*

### Παρατηρήσεις

Οι παράγοντες που επηρεάζουν τον υπολογισμό του διαστήματος εμπιστοσύνης της $\mu_1 - \mu_2$ είναι οι ίδιες όπως για τη $\mu$. Σ' αυτές προστίθεται και ο παράγοντας της ισότητας των διασπορών των $X_1$ και $X_2$. Στον Πίνακα 2.3 συνοψίζεται η εκτίμηση του διαστήματος εμπιστοσύνης της $\mu_1 - \mu_2$ στις διάφορες περιπτώσεις.

| *διασπορές* *των* $X_1$,$X_2$ | *κατανομή* *των* $X_1$,$X_2$ | $n_1$, $n_2$ | *κατανομή* της $\bar{x}_1 - \bar{x}_2$ | *διάστημα εμπιστοσύνης* |
|---|---|---|---|---|
| γνωστές | κανονική | | $z \equiv \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1)$ | $(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ |
| γνωστές | μη κανονική | μεγάλα | $z \equiv \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1)$ | $(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ |
| γνωστές | μη κανονική | μικρά | — | — |
| άγνωστες άνισες/ίσες | | μεγάλα | $z \equiv \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim \mathrm{N}(0,1)$ | $(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ |
| άγνωστες ίσες | κανονική | μικρά | $t \equiv \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim \mathrm{t}_{n_1+n_2-2}$ | $(\bar{x}_1-\bar{x}_2) \pm t_{n_1+n_2-2,1-\alpha/2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ |
| άγνωστες ίσες | μη κανονική | μικρά | — | — |
| άγνωστες άνισες | | μικρά | — | — |

Πίνακας 2.3: Εκτίμηση του διαστήματος εμπιστοσύνης της διαφοράς $\mu_1 - \mu_2$ ανάλογα με τη γνώση των διασπορών και κατανομών των τ.μ. $X_1$ και $X_2$ καθώς και των μεγεθών $n_1$ και $n_2$ των αντιστοίχων δειγμάτων.

Υπάρχει φανερή αντιστοιχία των περιπτώσεων για το διάστημα εμπιστοσύνης της μέσης τιμής (Πίνακας 2.2) και για το διάστημα εμπιστοσύνης της διαφοράς δύο μέσων τιμών (Πίνακας 2.3, οι έξι πρώτες σειρές). Για τη διαφορά μέσων τιμών υπάρχει ακόμα η περίπτωση των άνισων κι αγνώστων διασπορών σε συνδυασμό με μικρά δείγματα (τελευταία σειρά του Πίνακα 2.3) για την οποία δεν μπορούμε να καθορίσουμε την κατανομή της εκτιμήτριας $\bar{x}_1 - \bar{x}_2$ και απ' αυτήν να βρούμε το διάστημα εμπιστοσύνης. Σ' αυτήν την περίπτωση ούτε η μη παραμετρική εκτίμηση μπορεί να δώσει διάστημα εμπιστοσύνης (για τη διάμεσο).

Για το δ.ε. της $\mu$ είχαμε ορίσει τη σχέση του εύρους $w$ του δ.ε. και του μεγέθους $n$ του δείγματος. Επειδή για το δ.ε. της $\mu_1 - \mu_2$ έχουμε μεγέθη δύο δειγμάτων $n_1$ και $n_2$ είναι δύσκολο να καθορίσουμε τα $n_1$ και $n_2$ για κάποιο εύρος $w$ του διαστήματος εμπιστοσύνης. Αυτό το πρόβλημα δε θα μας απασχολήσει εδώ.

Για το διάστημα εμπιστοσύνης της διαφοράς μέσων τιμών υποθέσαμε οτι οι τ.μ. $X_1$ και $X_2$ είναι ανεξάρτητες. Δε θα ασχοληθούμε με την περίπτωση που οι $X_1$ και $X_2$ είναι εξαρτημένες (όταν δηλαδή έχουμε *ζευγαρωτές παρατηρήσεις*) γιατί τέτοια προβλήματα δεν παρουσιάζονται συχνά στη μηχανική.

Τα διαστήματα εμπιστοσύνης, πέρα από το ότι δίνουν ένα εύρος τιμών για την παράμετρο που μελετάμε, μας δίνουν τη δυνατότητα να ελέγξουμε αν μια συγκεκριμένη τιμή της παραμέτρου είναι πιθανή βλέποντας αν ανήκει στο διάστημα εμπιστοσύνης. Έτσι, στο τελευταίο παράδειγμα, ελέγχοντας αν η τιμή μηδέν ανήκει στο διάστημα εμπιστοσύνης μας επιτρέπει στην ουσία να ελέγξουμε αν οι δύο μέσες τιμές διαφέρουν σημαντικά ή όχι. Για τον στατιστικό έλεγχο υποθέσεων υπάρχει συγκεκριμένη μεθοδολογία που αποτελεί σημαντικό κομμάτι της στατιστικής με το οποίο όμως δε θα ασχοληθούμε εδώ. Θα πρέπει να τονισθεί ότι τα διαστήματα εμπιστοσύνης που ορίσαμε μας επιτρέπουν να ελέγξουμε αν μια παράμετρος μπορεί να πάρει κάποια τιμή και τα αποτελέσματα είναι ισοδύναμα με αυτά του στατιστικού ελέγχου της ίδιας υπόθεσης.

# Κεφάλαιο 3

# ΣΥΣΧΕΤΙΣΗ ΚΑΙ ΠΑΛΙΝΔΡΟΜΗΣΗ

Στα προηγούμενα κεφάλαια ορίσαμε και μελετήσαμε την τ.μ. με τη βοήθεια της πιθανοθεωρίας (κατανομή, ροπές) και της στατιστικής (εκτίμηση, στατιστική υπόθεση). Σ' αυτό το κεφάλαιο θα μελετήσουμε πως μια τ.μ. μεταβάλλεται όταν αλλάζει μια άλλη μεταβλητή (τυχαία ή μη).

Πρώτα θα μελετήσουμε τη σχέση δύο τ.μ. $X$ και $Y$. Συχνά στη μελέτη ενός τεχνικού συστήματος ενδιαφερόμαστε να προσδιορίσουμε τη σχέση μεταξύ δύο μεταβλητών του συστήματος. Για παράδειγμα στη λειτουργία μια μηχανής μας ενδιαφέρει η σχέση του χρόνου ως την αποτυχία κάποιου στοιχείου της μηχανής και της ταχύτητας του κινητήρα (σε περιστροφές ανά λεπτό). Θα προσδιορίσουμε και θα εκτιμήσουμε το συντελεστή συσχέτισης που μετράει τη γραμμική συσχέτιση δύο τ.μ..

Στη συνέχεια θα μελετήσουμε τη συναρτησιακή σχέση εξάρτησης μιας τ.μ. $Y$ ως προς μια άλλη μεταβλητή $X$. Η σχέση αυτή είναι πιθανοκρατική κι ορίζεται με την κατανομή της $Y$ για κάθε τιμή της $X$. Για παράδειγμα η απόδοση κάποιου εργαστηριακού πειράματος μπορεί να θεωρηθεί τ.μ. με κατανομή που μεταβάλλεται με τη θερμοκρασία στην οποία πραγματοποιείται. Συνήθως η μεταβολή αφορά μόνο τη μέση τιμή (κι ορισμένες φορές και τη διασπορά), γι αυτό κι η περιγραφή της κατανομής της $Y$ ως προς τη $X$ περιορίζεται στη δεσμευμένη μέση τιμή $\mathrm{E}(Y|X)$ και γίνεται με τη λεγόμενη ανάλυση παλινδρόμησης. Θα μελετήσουμε την απλή γραμμική παλινδρόμηση, δηλαδή θα περιοριστούμε να εκτιμήσουμε τη γραμμική σχέση για τη μέση τιμή $\mathrm{E}(Y|X)$ ως προς μια τ.μ. $X$.

## 3.1 Συσχέτιση δύο τ.μ.

Δύο τ.μ. $X$ και $Y$ μπορεί να συσχετίζονται με κάποιο τρόπο. Αυτό συμβαίνει όταν επηρεάζει η μία την άλλη, ή αν δεν αλληλοεπηρεάζονται όταν επηρεάζονται και οι δύο από μια άλλη μεταβλητή. Για παράδειγμα ο χρόνος ως την αποτυχία ενός στοιχείου κάποιας μηχανής και η ταχύτητα του κινητήρα της μηχανής μπορούν να θεωρηθούν σαν δύο τ.μ. που συσχετίζονται, όπου ο χρόνος αποτυχίας εξαρτάται από την ταχύτητα του κινητήρα (το αντίθετο δεν έχει πρακτική σημασία). Μπορούμε επίσης να θεωρήσουμε τη συσχέτιση του χρόνου αποτυχίας και της θερμοκρασίας του στοιχείου της μηχανής, αλλά τώρα δεν εξαρτάται η μια από την άλλη παρά εξαρτιούνται κι οι δύο από άλλες μεταβλητές, όπως η ταχύτητα του κινητήρα.

Στη συνέχεια θα θεωρήσουμε ότι οι δύο τ.μ. $X$ και $Y$ είναι συνεχείς. Για διακριτές τ.μ. μπορούμε πάλι να ορίσουμε μέτρο συσχέτισης τους αλλά δε θα μας απασχολήσει εδώ.

### 3.1.1 Ο συντελεστής συσχέτισης $\rho$

Ας θεωρήσουμε δύο τ.μ. $X$ και $Y$ με διασπορά $\sigma_X^2$ και $\sigma_Y^2$ αντίστοιχα και συνδιασπορά

$$\sigma_{XY} \equiv \mathrm{Cov}(X,Y) = \mathrm{E}(X,Y) - \mathrm{E}(X)\mathrm{E}(Y).$$

Η συνδιασπορά εκφράζει τη γραμμική συσχέτιση δύο τ.μ., δηλαδή την αναλογική μεταβολή (αύξηση ή μείωση) της μιας τ.μ. που αντιστοιχεί σε μεταβολή της άλλης μεταβλητής. Η συνδιασπορά είναι ένα ποσοτικό μέγεθος κι η μονάδα μέτρησης της εξαρτάται από τις μονάδες μέτρησης των δύο τ.μ. $X$ και $Y$. Γι αυτό για να μετρήσουμε καλύτερα τον βαθμό της *γραμμικής* συσχέτισης δύο τ.μ. χρησιμοποιούμε τον **συντελεστή συσχέτισης** (correlation coefficient) $\rho$ που ορίζεται ως

$$\rho = \frac{\sigma_{XY}}{\sigma_X\, \sigma_Y}. \tag{3.1}$$

Ο συντελεστής συσχέτισης $\rho$ παίρνει τιμές στο διάστημα $[-1, 1]$:

- $\rho = 1$: υπάρχει *τέλεια θετική* σχέση μεταξύ των $X$ και $Y$,
- $\rho = 0$: δεν υπάρχει καμιά (γραμμική) σχέση μεταξύ των $X$ και $Y$,
- $\rho = -1$: υπάρχει *τέλεια αρνητική* σχέση μεταξύ των $X$ και $Y$.

Όταν $\rho = \pm 1$ η σχέση είναι αιτιοκρατική κι όχι πιθανοκρατική γιατί γνωρίζοντας την τιμή της μιας τ.μ. γνωρίζουμε και την τιμή της άλλης τ.μ. ακριβώς. Όταν ο συντελεστής συσχέτισης είναι κοντά στο $-1$ ή $1$ η γραμμική συσχέτιση των δύο τ.μ. είναι ισχυρή (συνήθως χαρακτηρίζουμε ισχυρές τις σχέσεις όταν $|\rho| > 0.9$) ενώ όταν είναι κοντά στο $0$ οι τ.μ. είναι πρακτικά ασυσχέτιστες.

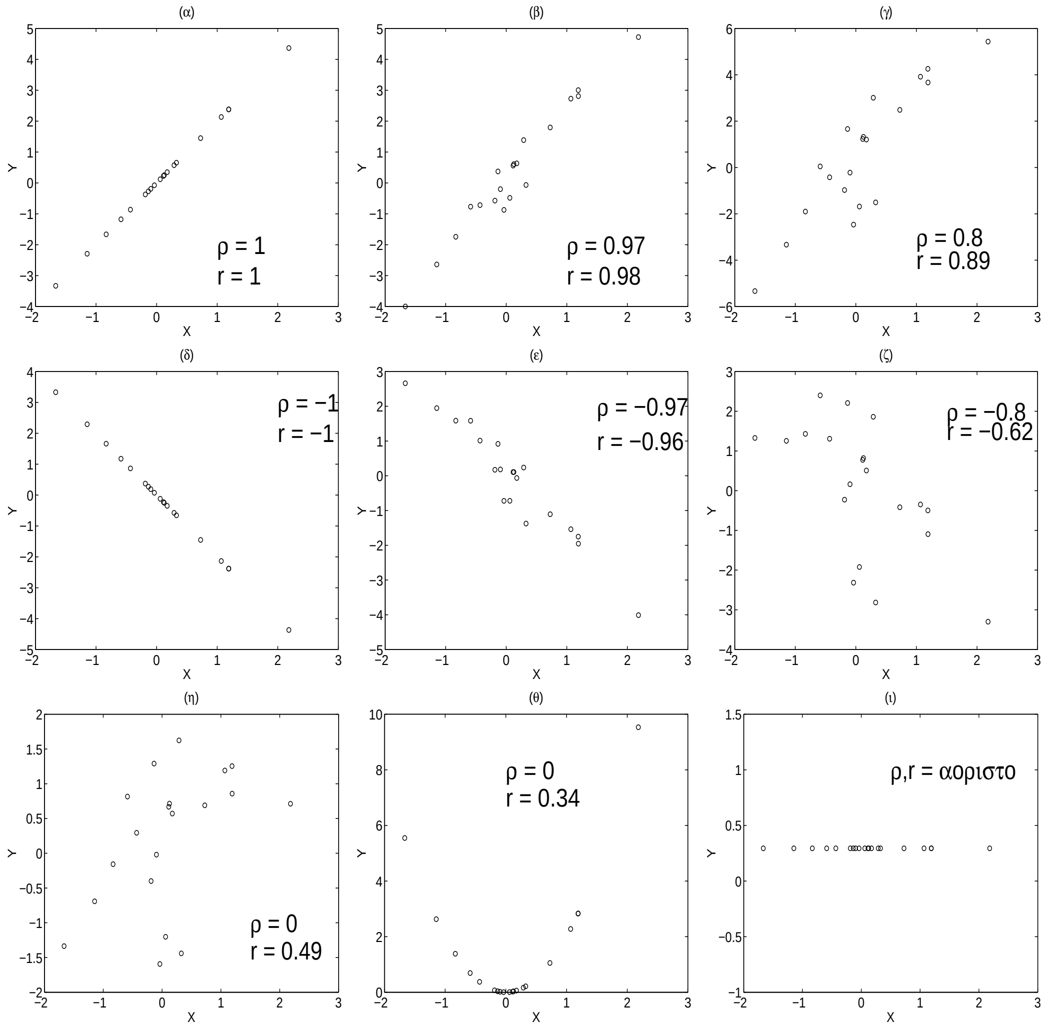
Όπως φαίνεται από τον ορισμό στη σχέση (3.1), ο συντελεστής συσχέτισης $\rho$ δεν εξαρτάται από τη μονάδα μέτρησης των $X$ και $Y$ και είναι συμμετρικός ως προς τις $X$ και $Y$.

### 3.1.2 Σημειακή εκτίμηση του συντελεστή συσχέτισης

Όταν έχουμε παρατηρήσεις των δύο τ.μ. $X$ και $Y$ κατά ζεύγη

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\},$$

μπορούμε να εκτιμήσουμε τη συσχέτιση τους ποιοτικά από το **διάγραμμα διασποράς** (scatter diagram), που είναι η απεικόνιση των σημείων $(x_i, y_i)$, $i = 1, \ldots, n$, σε καρτεσιανό σύστημα συντεταγμένων. Στο Σχήμα 3.1 παρουσιάζονται τυπικά διαγράμματα διασποράς για ισχυρές κι ασθενείς συσχετίσεις δύο τ.μ. $X$ και $Y$. Στα Σχήματα 3.1α και 3.1δ η σχέση είναι τέλεια ($\rho = 1$ και $\rho = -1$ αντίστοιχα), στα Σχήματα 3.1β και 3.1ε είναι ισχυρή (θετική με $\rho = 0.97$ κι αρνητική με $\rho = -0.97$ αντίστοιχα) και στα Σχήματα 3.1γ και 3.1ζ είναι λιγότερο ισχυρή (θετική με $\rho = 0.8$ κι αρνητική με $\rho = -0.8$ αντίστοιχα). Στο Σχήμα 3.1η είναι $\rho = 0$ γιατί οι τ.μ. $X$ και $Y$ είναι ανεξάρτητες ενώ στο Σχήμα 3.1θ είναι πάλι $\rho = 0$ αλλά οι $X$ και $Y$ δεν είναι ανεξάρτητες αλλά συσχετίζονται μόνο μη-γραμμικά. Τέλος για το Σχήμα 3.1ι ο συντελεστής συσχέτισης δεν ορίζεται γιατί η $Y$ είναι σταθερή ($\sigma_Y = 0$ στον ορισμό του $\rho$ στην (3.1)).




Σχήμα 3.1: Διάγραμμα διασποράς δύο τ.μ. $X$ και $Y$ από $n = 20$ παρατηρήσεις που παρουσιάζουν θετική σχέση στα σχήματα (α), (β) και (γ), αρνητική σχέση στα σχήματα (δ), (ε) και (ζ) και καμιά συσχέτιση στα σχήματα (η), (θ) και (ι). Σε κάθε σχήμα δίνεται η πραγματική τιμή του συντελεστή συσχέτισης $\rho$ κι η δειγματική $r$. Στο (ι) ο συντελεστής συσχέτισης δεν ορίζεται.

Η σημειακή εκτίμηση του συντελεστή συσχέτισης $\rho$ του πληθυσμού από το δείγμα των $n$ ζευγαρωτών παρατηρήσεων των $X$ και $Y$ γίνεται με την αντικατάσταση στη σχέση (3.1) της συνδιασποράς $\sigma_{XY}$ και των διασπορών $\sigma_X^2$ και $\sigma_Y^2$ από τις αντίστοιχες εκτιμήσεις από το δείγμα

$$\hat{\rho} \equiv r = \frac{s_{XY}}{s_X\, s_Y}.$$

Οι αμερόληπτες εκτιμήτριες $s_{XY}$, $s_X^2$ και $s_Y^2$ δίνονται ως

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) \tag{3.2}$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \tag{3.3}$$

όπου $\bar{x}$ και $\bar{y}$ είναι οι δειγματικές μέσες τιμές των $X$ και $Y$. Από τα παραπάνω προκύπτει η έκφραση της εκτιμήτρια $r$

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}. \tag{3.4}$$

Είναι προφανές πως η παραπάνω σχέση για το $r$ δεν αλλάζει αν θεωρήσουμε τις μεροληπτικές εκτιμήτριες των $\sigma_{XY}$, $\sigma_X$ και $\sigma_Y$.

Στο Σχήμα 3.1 δίνεται ο δειγματικός συντελεστής συσχέτισης $r$ για κάθε περίπτωση. Επειδή το δείγμα είναι μικρό ($n = 20$) η τιμή του $r$ δεν είναι πάντα κοντά στην πραγματική τιμή $\rho$. Αυτό συμβαίνει γιατί η εκτιμήτρια $r$ όπως δίνεται στη σχέση (3.4) είναι μια τ.μ. που εξαρτάται από τις τιμές και το πλήθος των ζευγών των παρατηρήσεων.

Για να εκφράσουμε τη συσχέτιση δύο τ.μ. χρησιμοποιούμε επίσης την ποσότητα $r^2$ που λέγεται και **συντελεστής προσδιορισμού** (coefficient of determination) (εκφράζεται συνήθως σε ποσοστό $\%$, $100r^2$). Ο συντελεστής προσδιορισμού δίνει το ποσοστό μεταβλητότητας των τιμών της $Y$ που υπολογίζεται από τη $X$ (κι αντίστροφα) κι είναι ένας χρήσιμος τρόπος να συνοψίσουμε τη σχέση δύο τ.μ..

**Παράδειγμα 3.1** *Θέλουμε να διερευνήσουμε τη συσχέτιση της αντίστασης και του χρόνου αποτυχίας κάποιου υπερφορτωμένου αντιστάτη. Γι αυτό πήραμε μετρήσεις αντίστασης (σε Ωμ, ohm) και χρόνου αποτυχίας (σε λεπτά, min) από δείγμα 20 αντιστατών, οι οποίες παρουσιάζονται στον Πίνακα 3.1. Για να βρούμε τον συντελεστή συσχέτισης $r$ υπολογίζουμε πρώτα τα παρακάτω*

$$\bar{x} = 38 \qquad \bar{y} = 33.5$$

$$\sum_{i=1}^{20} x_i^2 = 29634 \qquad \sum_{i=1}^{20} y_i^2 = 23910 \qquad \sum_{i=1}^{20} x_i y_i = 26305.$$

*Από τη σχέση (3.4) βρίσκουμε*

$$r = \frac{26305 - 20 \cdot 38 \cdot 33.5}{\sqrt{(29634 - 20 \cdot 38^2) \cdot (23910 - 20 \cdot 33.5^2)}} = 0.804.$$

*Η τιμή του συντελεστή συσχέτισης $r \simeq 0.8$ υποδηλώνει ότι η αντίσταση κι ο χρόνος αποτυχίας αντιστάτη έχουν γραμμική θετική συσχέτιση αλλά όχι πολύ ισχυρή. Αυτό φαίνεται κι από το διάγραμμα διασποράς στο Σχήμα 3.2. Η μεταβλητότητα της μιας τ.μ. (αντίσταση ή χρόνος αποτυχίας) μπορεί να εξηγηθεί από τη συσχέτιση της με την άλλη κατά ποσοστό που δίνεται από το συντελεστή προσδιορισμού, που είναι*

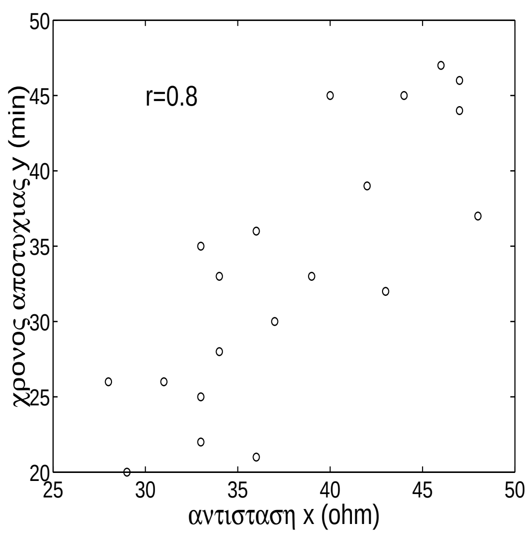
$$r^2 \cdot 100 = 0.804^2 \cdot 100 = 64.64 \quad \rightarrow \quad \simeq 65\%.$$

*Συμπεραίνουμε λοιπόν πως η γνώση της μιας τ.μ. δε μας επιτρέπει να προσδιορίσουμε την άλλη με μεγάλη ακρίβεια. Πρέπει επίσης να σημειωθεί ότι η εκτίμηση $r$ του συντελεστή συσχέτισης μπορεί να αλλάξει σημαντικά με την πρόσθεση ή αφαίρεση λίγων παρατηρήσεων γιατί το μέγεθος του δείγματος είναι μικρό.*

Για τον συντελεστή συσχέτισης $\rho$ μπορούμε να υπολογίσουμε διαστήματα εμπιστοσύνης κάτω από την υπόθεση κανονικής κατανομής των $X$ και $Y$, αλλά εδώ δε θα ασχοληθούμε με αυτά τα θέματα.

| Α/Α ($i$) | Αντίσταση $x_i$ (ohm) | Χρόνος αποτυχίας $y_i$ (min) |
|---|---|---|
| 1 | 28 | 26 |
| 2 | 29 | 20 |
| 3 | 31 | 26 |
| 4 | 33 | 22 |
| 5 | 33 | 25 |
| 6 | 33 | 35 |
| 7 | 34 | 28 |
| 8 | 34 | 33 |
| 9 | 36 | 21 |
| 10 | 36 | 36 |
| 11 | 37 | 30 |
| 12 | 39 | 33 |
| 13 | 40 | 45 |
| 14 | 42 | 39 |
| 15 | 43 | 32 |
| 16 | 44 | 45 |
| 17 | 46 | 47 |
| 18 | 47 | 44 |
| 19 | 47 | 46 |
| 20 | 48 | 37 |

Πίνακας 3.1: Δεδομένα αντίστασης $x_i$ και χρόνου αποτυχίας $y_i$ για 20 αντιστάτες.




Σχήμα 3.2: Διάγραμμα διασποράς για το δείγμα παρατηρήσεων αντίστασης και χρόνου αποτυχίας 20 αντιστατών του Πίνακα 3.1.

## 3.2 Απλή Γραμμική Παλινδρόμηση

Στη *συσχέτιση* που μελετήσαμε παραπάνω μετρήσαμε με το συντελεστή συσχέτισης τη γραμμική σχέση δύο τ.μ. $X$ και $Y$. Στην *παλινδρόμηση* που θα μελετήσουμε τώρα σχεδιάζουμε την εξάρτηση μιας τ.μ. $Y$, που την ονομάζουμε **εξαρτημένη μεταβλητή** (dependent variable), από κάποια άλλη μεταβλητή $X$ που την ονομάζουμε **ανεξάρτητη μεταβλητή** (independent variable). Ενώ λοιπόν η συσχέτιση είναι συμμετρική ως προς τα $X$ και $Y$, στην παλινδρόμηση η εξαρτημένη μεταβλητή $Y$ 'καθοδηγείται' από την ανεξάρτητη μεταβλητή $X$. Γι αυτό και στην ανάλυση που κάνουμε παίζει ρόλο ποιόν από τους δύο παράγοντες που μετράμε ορίζουμε σαν ανεξάρτητη μεταβλητή και ποιόν σαν εξαρτημένη. Για παράδειγμα, σε μια μονάδα παραγωγής ηλεκτρικής ενέργειας από λιγνίτη, για να προσδιορίσουμε το κόστος της παραγωγής ενέργειας, μελετάμε την εξάρτηση του από το κόστος του λιγνίτη. Είναι φυσικό λοιπόν ως εξαρτημένη μεταβλητή $Y$ να θεωρήσουμε το κόστος παραγωγής ηλεκτρικής ενέργειας κι ως ανεξάρτητη μεταβλητή $X$ το κόστος του λιγνίτη.

### 3.2.1 Το πρόβλημα της απλής γραμμικής παλινδρόμησης

Η εξαρτημένη τ.μ. $Y$ ακολουθεί κάποια κατανομή. Επειδή μας ενδιαφέρει η συμπεριφορά της $Y$ για κάθε δυνατή τιμή της ανεξάρτητης μεταβλητής $X$ θέλουμε να μελετήσουμε τη δεσμευμένη κατανομή της $Y$ για κάθε τιμή $x$ της $X$. Με αναφορά στη δεσμευμένη αθροιστική συνάρτηση κατανομής θέλουμε να προσδιορίσουμε την $F_Y(y|X=x)$ για κάθε τιμή $x$ της $X$. Αυτό είναι αρκετά περίπλοκο πρόβλημα που στην πράξη συχνά δε χρειάζεται να λύσουμε. Περιορίζουμε λοιπόν τη μελέτη του προβλήματος της παλινδρόμησης στη δεσμευμένη μέση τιμή $\mathrm{E}(Y|X=x)$. Υποθέτοντας ότι η εξάρτηση εκφράζεται από γραμμική σχέση έχουμε

$$\mathrm{E}(Y|X=x) = \alpha + \beta x \tag{3.5}$$

και η σχέση αυτή λέγεται **απλή γραμμική παλινδρόμηση της $Y$ στη $X$** (linear regression). [Λέγεται *απλή* γιατί η εξάρτηση είναι προς μόνο μια μεταβλητή.] Το πρόβλημα της παλινδρόμησης είναι η εύρεση των παραμέτρων $\alpha$ και $\beta$ που εκφράζουν καλύτερα τη γραμμική εξάρτηση της $Y$ από τη $X$. Κάθε ζεύγος τιμών $(\alpha, \beta)$ καθορίζει μια διαφορετική γραμμική σχέση που εκφράζεται γεωμετρικά από ευθεία γραμμή. Ο σταθερός όρος $\alpha$ είναι η τιμή του $y$ για $x=0$ και λέγεται **διαφορά ύψους** (intercept) κι ο συντελεστής του $x$, $\beta$, είναι η **κλίση** (slope) της ευθείας ή αλλιώς ο **συντελεστής παλινδρόμησης** (regression coefficient). Αν θεωρήσουμε τις παρατηρήσεις $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ και το διάγραμμα διασποράς που τις απεικονίζει σαν σημεία, μπορούμε να σχηματίσουμε πολλές τέτοιες ευθείες που προσεγγίζουν την υποτιθέμενη γραμμική εξάρτηση της $\mathrm{E}(Y|X=x)$ ως προς $X$, όπως φαίνεται στο Σχήμα 3.3.

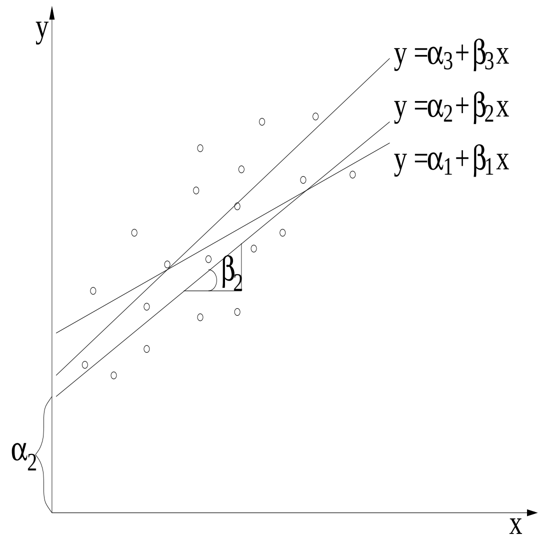
Για κάποια τιμή $x_i$ της $X$ αντιστοιχούν διαφορετικές τιμές $y_i$ της $Y$, σύμφωνα με κάποια κατανομή πιθανότητας $F_Y(y_i|X=x_i)$, δηλαδή μπορούμε να θεωρήσουμε την $y_i$ σαν τ.μ. [θα ήταν σωστότερο να χρησιμοποιούσαμε το συμβολισμό $Y_i$ αντί $y_i$, όπου ο δείκτης $i$ ορίζει την εξάρτηση από $X=x_i$, αλλά θα χρησιμοποιήσουμε εδώ τον ίδιο συμβολισμό $y_i$ για την τ.μ. και την παρατήρηση]. Η τ.μ. $y_i$ για κάποια τιμή $x_i$ της $X$ θα δίνεται κάτω από την υπόθεση της γραμμικής παλινδρόμησης ως

$$y_i = \alpha + \beta x_i + \epsilon_i, \tag{3.6}$$

όπου $\epsilon_i$ είναι κι αυτή τ.μ., λέγεται **σφάλμα παλινδρόμησης** (regression error) κι ορίζεται ως η διαφορά της $y_i$ από τη δεσμευμένη μέση τιμή της

$$\epsilon_i = y_i - \mathrm{E}(Y|X=x_i).$$

Για την ανάλυση της γραμμικής παλινδρόμησης κάνουμε τις παρακάτω υποθέσεις:




Σχήμα 3.3: Ευθείες απλής γραμμικής παλινδρόμησης

- Η μεταβλητή $X$ είναι *ελεγχόμενη* για το πρόβλημα που μελετάμε, δηλαδή γνωρίζουμε τις τιμές της χωρίς καμιά αμφιβολία.

- Η σχέση (3.5) ισχύει, δηλαδή η εξάρτηση της $Y$ από τη $X$ είναι γραμμική.

- $\mathrm{E}(\epsilon_i) = 0$ και $\mathrm{Var}(\epsilon_i) = \sigma_\epsilon^2$ για κάθε τιμή $x_i$ της $X$, δηλαδή το σφάλμα παλινδρόμησης έχει μέση τιμή μηδέν για κάθε τιμή της $X$ και η διασπορά του είναι σταθερή και δεν εξαρτάται από τη $X$.

Η τελευταία συνθήκη είναι ισοδύναμη με τη συνθήκη

$$\mathrm{Var}(Y|X = x) \equiv \sigma_{Y|X}^2 = \sigma^2,$$

δηλαδή η διασπορά της εξαρτημένης μεταβλητής $Y$ είναι η ίδια για κάθε τιμή της $X$ και μάλιστα είναι $\sigma_{Y|X}^2 = \sigma_\epsilon^2 \equiv \sigma^2$. Η τελευταία σχέση προκύπτει από τη σχέση (3.6), αφού οι παράμετροι $\alpha$ και $\beta$ είναι σταθερές και το $x_i$ γνωστό. Η ιδιότητα αυτή λέγεται *ομοσκεδαστικότητα* (δες επίσης 2.2.3) και αντίθετα έχουμε *ετεροσκεδαστικότητα* όταν η διασπορά της $Y$ (ή του σφάλματος $\epsilon$) μεταβάλλεται με τη $X$.

Γενικά για να εκτιμήσουμε τις παραμέτρους της γραμμικής παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων, όπως θα δούμε παρακάτω, δεν είναι απαραίτητο να υποθέσουμε κάποια συγκεκριμένη δεσμευμένη κατανομή $F_Y(y_i|X = x_i)$ της $Y$ ως προς τη $X$. Αν θέλουμε όμως να υπολογίσουμε διαστήματα εμπιστοσύνης για τις παραμέτρους θα χρειαστούμε να υποθέσουμε κανονική δεσμευμένη κατανομή για τη $Y$. Επίσης οι παραπάνω υποθέσεις για γραμμική σχέση και σταθερή διασπορά αποτελούν χαρακτηριστικά πληθυσμών με κανονική κατανομή. Συνήθως λοιπόν σε προβλήματα γραμμικής παλινδρόμησης υποθέτουμε ότι η δεσμευμένη κατανομή της $Y$ είναι κανονική

$$Y|X = x \sim \mathrm{N}(\alpha + \beta x, \sigma^2).$$

### 3.2.2 Σημειακή εκτίμηση παραμέτρων της απλής γραμμικής παλινδρόμησης

Το πρόβλημα της απλής γραμμικής παλινδρόμησης με τις υποθέσεις που ορίστηκαν παραπάνω συνίσταται στην εκτίμηση των τριών παραμέτρων της παλινδρόμησης:

1. της διαφοράς ύψους της ευθείας παλινδρόμησης $\alpha$,

2. της κλίσης της ευθείας παλινδρόμησης $\beta$,

3. της διασποράς σφάλματος της παλινδρόμησης $\sigma^2$.

Τα $\alpha$ και $\beta$ προσδιορίζουν την ευθεία παλινδρόμησης κι άρα καθορίζουν τη γραμμική σχέση εξάρτησης της τ.μ. $Y$ από τη μεταβλητή $X$. Η παράμετρος $\sigma^2$ προσδιορίζει το βαθμό μεταβλητότητας γύρω από την ευθεία παλινδρόμησης κι εκφράζει την αβεβαιότητα της γραμμικής σχέσης.

**Εκτίμηση των παραμέτρων της ευθείας παλινδρόμησης** Η εκτίμηση των παραμέτρων $\alpha$ και $\beta$ γίνεται με τη μέθοδο των **ελαχίστων τετραγώνων** (method of least squares). Η μέθοδος λέγεται έτσι γιατί βρίσκει την ευθεία παλινδρόμησης με παραμέτρους $a$ και $b$ έτσι ώστε το άθροισμα των τετραγώνων των κατακόρυφων αποστάσεων των σημείων από την ευθεία να είναι το ελάχιστο. Οι εκτιμήσεις των $\alpha$ και $\beta$ δίνονται από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \epsilon_i^2 \qquad \text{ή} \qquad \min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2. \tag{3.7}$$

Για να λύσουμε αυτό το πρόβλημα θέτουμε τις μερικές παραγώγους ως προς τα $\alpha$ και $\beta$ ίσες με το μηδέν και καταλήγουμε στο σύστημα δύο εξισώσεων με δύο αγνώστους

$$\left.\begin{array}{l} \frac{\partial \sum (y_i-\alpha-\beta x_i)^2}{\partial \alpha} = 0 \\ \frac{\partial \sum (y_i-\alpha-\beta x_i)^2}{\partial \beta} = 0 \end{array}\right\} \begin{array}{l} \sum_{i=1}^{n} y_i = n\alpha + \beta \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i = \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 \end{array}$$

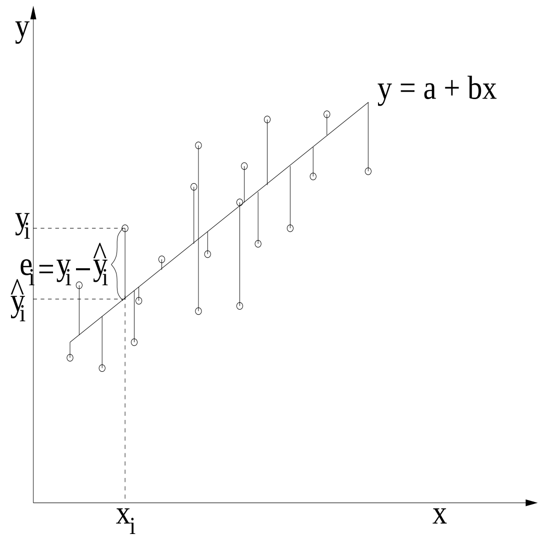
από το οποίο παίρνουμε τις εκτιμήσεις των $\alpha$ και $\beta$

$$b = \frac{s_{XY}}{s_X^2}, \qquad a = \bar{y} - b\bar{x}, \tag{3.8}$$

όπου $s_{XY}$ και $s_X^2$ είναι η δειγματική συνδιασπορά των $X$ και $Y$ και η δειγματική διασπορά της $X$ που ορίστηκαν στις σχέσεις (3.2) και (3.3) αντίστοιχα. Τα $a$ και $b$ ορίζουν την ευθεία

$$\hat{y} = a + bx,$$

που λέγεται κι **ευθεία ελαχίστων τετραγώνων**.




Σχήμα 3.4: Ευθεία ελαχίστων τετραγώνων και υπόλοιπα

**Εκτίμηση της διασποράς των σφαλμάτων παλινδρόμησης** Για κάθε δοθείσα τιμή $x_i$ με τη βοήθεια της ευθείας ελαχίστων τετραγώνων εκτιμούμε την τιμή $\hat{y}_i$ που γενικά είναι διαφορετική από την πραγματική τιμή $y_i$. Η διαφορά

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

είναι η κατακόρυφη απόσταση της πραγματικής τιμής από την ευθεία ελαχίστων τετραγώνων και λέγεται σφάλμα ελαχίστων τετραγώνων ή απλά **υπόλοιπο** (residual). Στο Σχήμα 3.4 απεικονίζονται τα υπόλοιπα της παλινδρόμησης.

Το υπόλοιπο $e_i$ είναι η εκτίμηση του σφάλματος παλινδρόμησης $\epsilon_i$ αντικαθιστώντας απλά τις παραμέτρους παλινδρόμησης με τις εκτιμήσεις ελαχίστων τετραγώνων στον ορισμό του σφάλματος $\epsilon_i = y_i - \alpha - \beta x_i$. Άρα η εκτίμηση της διασποράς $\sigma^2$ του σφάλματος (που είναι κι η δεσμευμένη διασπορά της $Y$ ως προς $X$) δίνεται από τη δειγματική διασπορά $s^2$ των υπολοίπων $e_i$

$$s^2 \equiv \sigma_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{3.9}$$

όπου διαιρούμε με $n-2$ γιατί από τους βαθμούς ελευθερίας $n$ του μεγέθους του δείγματος αφαιρούμε δύο για τις δύο παραμέτρους που έχουν ήδη εκτιμηθεί. Η δειγματική διασπορά $s^2$ μπορεί να εκφραστεί ως προς τις δειγματικές διασπορές των $X$ και $Y$ και της συνδιασποράς τους, αν αντικαταστήσουμε τις εκφράσεις των $a$ και $b$ από την (3.8) στην παραπάνω σχέση (όπου θέτουμε $\hat{y}_i = a + bx_i$)

$$s^2 = \frac{n-1}{n-2}\left(s_Y^2 - \frac{s_{XY}^2}{s_X^2}\right) = \frac{n-1}{n-2}(s_Y^2 - b^2 s_X^2) \tag{3.10}$$

όπου και πάλι υποθέτουμε τις αμερόληπτες εκτιμήτριες για τις διασπορές.

**Παρατηρήσεις**

1. Η ευθεία ελαχίστων τετραγώνων περνάει από το σημείο $(\bar{x}, \bar{y})$ γιατί

$$a + b\bar{x} = \bar{y} - b\bar{x} + b\bar{x} = \bar{y}.$$

   Άρα η ευθεία ελαχίστων τετραγώνων μπορεί επίσης να οριστεί ως

$$y_i - \bar{y} = b(x_i - \bar{x}).$$

2. Η εκτίμηση των $\alpha$ και $\beta$ με τη μέθοδο των ελαχίστων τετραγώνων δεν προϋποθέτει σταθερή διασπορά και κανονική κατανομή της εξαρτημένης μεταβλητής $Y$ για κάθε τιμή της ανεξάρτητης μεταβλητής $X$. Όταν όμως ισχύουν οι δύο αυτές συνθήκες οι εκτιμήτριες ελαχίστων τετραγώνων $a$ και $b$ είναι οι εκτιμήτριες μεγίστης πιθανοφάνειας (κι άρα έχουν και τις επιθυμητές ιδιότητες εκτιμητριών).

3. Αν η διασπορά της $Y$ αλλάζει με το $X$, τότε η διαδικασία των ελαχίστων τετραγώνων πρέπει να διορθωθεί έτσι ώστε να δίνει περισσότερο βάρος στις παρατηρήσεις που αντιστοιχούν σε μικρότερη διασπορά.

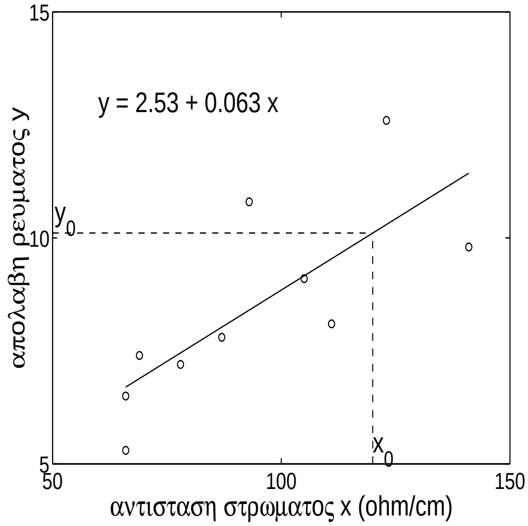
4. Για κάθε τιμή $x_0$ της $X$ μπορούμε να *προβλέψουμε* την αντίστοιχη τιμή $y_0$ της $Y$ από την ευθεία ελαχίστων τετραγώνων, $\hat{y}_0 = a + bx_0$. Εδώ πρέπει να προσέξουμε ότι η τιμή $x_0$ πρέπει να ανήκει στο εύρος τιμών της $X$ που έχουμε από το δείγμα. Για τιμές έξω από αυτό το διάστημα η πρόβλεψη δεν είναι αξιόπιστη.

**Παράδειγμα 3.2** *Θέλουμε να μελετήσουμε σ' ένα ολοκληρωμένο κύκλωμα την εξάρτηση της απολαβής ρεύματος κρυσταλλολυχνίας (τρανζίστορ) από την αντίσταση του στρώματος της κρυσταλλολυχνίας. Στον Πίνακα 3.2 παρουσιάζονται 10 μετρήσεις της απολαβής ρεύματος για αντίστοιχες τιμές της αντίστασης στρώματος της κρυσταλλολυχνίας.*

| Α/Α ($i$) | Αντίσταση στρώματος $x_i$ (ohm/cm) | Απολαβή ρεύματος $y_i$ |
|---|---|---|
| 1 | 66 | 5.3 |
| 2 | 66 | 6.5 |
| 3 | 69 | 7.4 |
| 4 | 78 | 7.2 |
| 5 | 87 | 7.8 |
| 6 | 93 | 10.8 |
| 7 | 105 | 9.1 |
| 8 | 111 | 8.1 |
| 9 | 123 | 12.6 |
| 10 | 141 | 9.8 |

Πίνακας 3.2: Δεδομένα απολαβής ρεύματος τρανζίστορ ($y_i$) για διαφορετικές τιμές της αντίστασης στρώματος κρυσταλλολυχνίας ($x_i$).

*Υποθέτουμε πως η απολαβή ρεύματος της κρυσταλλολυχνίας εξαρτάται γραμμικά από την αντίσταση του στρώματός της και το διάγραμμα διασποράς από το δείγμα στο Σχήμα 3.5 επιβεβαιώνει αυτήν την υπόθεση. Για να εκτιμήσουμε τις παραμέτρους $a$ και $b$ της ευθείας ελαχίστων τετραγών-*




Σχήμα 3.5: Διάγραμμα διασποράς για τα δεδομένα του Πίνακα 3.2 κι ευθεία ελαχίστων τετραγώνων.

*ων υπολογίζουμε πρώτα τα παρακάτω*

$$\bar{x} = 93.9 \qquad \bar{y} = 8.46$$

$$\sum_{i=1}^{10} x_i^2 = 94131 \qquad \sum_{i=1}^{10} y_i^2 = 757.64 \qquad \sum_{i=1}^{10} x_i y_i = 8320.2$$

*και χρησιμοποιώντας τις σχέσεις (3.2) και (3.3) για τη δειγματική συνδιασπορά και διασπορά αντίστοιχα, βρίσκουμε*

$$s_{XY} = 41.81 \qquad s_X^2 = 662.1 \qquad s_Y^2 = 4.66.$$

*Οι εκτιμήσεις $b$ και $a$ είναι*

$$\begin{aligned} b &= \frac{41.81}{662.1} = 0.063 \\ a &= 8.46 - 0.063 \cdot 93.9 = 2.53. \end{aligned}$$

*Από τη σχέση (3.10) υπολογίζουμε την εκτίμηση διασποράς των σφαλμάτων παλινδρόμησης*

$$s^2 = \frac{9}{8}(4.66 - 0.063^2 \cdot 41.81) = 5.056.$$

*Τα αποτελέσματα αυτά ερμηνεύονται ως εξής:*

1. $b$: *Για αύξηση της αντίστασης στρώματος κατά μία μονάδα μέτρησης (1 ohm/cm) η απολαβή του ρεύματος της κρυσταλλολυχνίας αυξάνεται κατά* $0.063$.

2. $a$: *Όταν δεν υπάρχει καθόλου αντίσταση στρώματος ($x = 0$), η απολαβή του ρεύματος είναι* $2.53$ *μονάδες αλλά βέβαια είναι αδύνατο να θεωρήσουμε στρώμα χωρίς αντίσταση. Δε θα πρέπει λοιπόν να επιχειρήσουμε προβλέψεις για τιμές της αντίστασης στρώματος μικρότερης του* $66$ *ohm/cm και μεγαλύτερης του* $141$ *ohm/cm, που είναι οι ακραίες τιμές της αντίστασης του δείγματος.*

3. *$s^2$: Η εκτίμηση της διασποράς γύρω από την ευθεία παλινδρόμησης για κάθε τιμή της $X$ (στο διάστημα τιμών του πειράματος) είναι $5.056$, ή αλλιώς το τυπικό σφάλμα της εκτίμησης της παλινδρόμησης είναι $2.249$ μονάδες, που είναι μεγάλο σε σχέση με το επίπεδο τιμών της $Y$.*

*Με βάση το μοντέλο παλινδρόμησης που εκτιμήσαμε μπορούμε να προβλέψουμε την απολαβή ρεύματος για κάθε αντίσταση στρώματος κρυσταλλολυχνίας στο διάστημα $[66, 141]$ ohm/cm. Στο Σχήμα 3.5 απεικονίζεται η πρόβλεψη της απολαβής ρεύματος για αντίσταση στρώματος $x_0 = 120$ ohm/cm και είναι*

$$y_0 = 2.53 + 0.063 \cdot 120 = 10.11.$$

Για τις παραμέτρους $\alpha$ και $\beta$, καθώς και για τη διασπορά $\sigma^2$, μπορούμε να εκτιμήσουμε διαστήματα εμπιστοσύνης και να κάνουμε στατιστικούς ελέγχους αλλά δε θα ασχοληθούμε με αυτά τα θέματα.

### 3.2.3 Σχέση του συντελεστή συσχέτισης και παλινδρόμησης

Η παλινδρόμηση ορίζεται θεωρώντας την ανεξάρτητη μεταβλητή $X$ σταθερή και την εξαρτημένη μεταβλητή $Y$ τυχαία, ενώ για τη συσχέτιση θεωρούμε και τις δύο μεταβλητές $X$ και $Y$ τυχαίες. Για τις μεταβλητές $X$ και $Y$ της παλινδρόμησης, μπορούμε να αγνοήσουμε ότι η $X$ δεν είναι τ.μ. και να ορίσουμε το συντελεστή συσχέτισης $\rho$ όπως και πριν. Η σχέση μεταξύ του $r$ (της εκτιμήτριας του $\rho$ από το δείγμα) και του συντελεστή της παλινδρόμησης $b$ δίνεται ως εξής (συνδυάζοντας τις σχέσεις $r = \frac{s_{XY}}{s_X s_Y}$ και $b = \frac{s_{XY}}{s_X^2}$)

$$r = b\frac{s_X}{s_Y} \quad \text{ή} \quad b = r\frac{s_Y}{s_X}. \tag{3.11}$$

Και τα δύο μεγέθη, $r$ και $b$, εκφράζουν ποσοτικά τη γραμμική συσχέτιση των μεταβλητών $X$ και $Y$, αλλά το $b$ εξαρτάται από τη μονάδα μέτρησης των $X$ και $Y$ ενώ το $r$, επειδή προκύπτει από το λόγο της συνδιασποράς προς τις τυπικές αποκλίσεις των $X$ και $Y$, δεν εξαρτάται από τη μονάδα μέτρησης των $X$ και $Y$ και δίνει τιμές στο διάστημα $[-1, 1]$. Η σχέση των $r$ και $b$ περιγράφεται ως εξής:

- Αν η συσχέτιση είναι θετική ($r > 0$) τότε η κλίση της ευθείας παλινδρόμησης $b$ είναι επίσης θετική.
- Αν η συσχέτιση είναι αρνητική ($r < 0$) τότε η κλίση της ευθείας παλινδρόμησης $b$ είναι επίσης αρνητική.
- Αν οι μεταβλητές $X$ και $Y$ δε συσχετίζονται ($r = 0$) τότε η ευθεία παλινδρόμησης είναι οριζόντια ($b = 0$).

Επίσης μπορούμε να εκφράσουμε το $r^2$ ως προς τη δειγματική διασπορά του σφάλματος $s^2$ και αντίστροφα

$$s^2 = \frac{n-1}{n-2}s_Y^2(1-r^2) \quad \text{ή} \quad r^2 = 1 - \frac{n-2}{n-1}\frac{s^2}{s_Y^2}. \tag{3.12}$$

Η παραπάνω σχέση δηλώνει πως όσο μεγαλύτερο είναι το $r^2$ (ή το $|r|$) τόσο μειώνεται η διασπορά του σφάλματος της παλινδρόμησης, δηλαδή τόσο ακριβέστερη είναι η πρόβλεψη που βασίζεται στην ευθεία παλινδρόμησης.

**Παράδειγμα 3.3** *Στο παραπάνω παράδειγμα, ο συντελεστής συσχέτισης του κέρδους ρεύματος της κρυσταλλολυχνίας και της αντίστασης στρώματος είναι*

$$r = \frac{s_{XY}}{s_X\, s_Y} = \frac{41.81}{\sqrt{662.1 \cdot 4.66}} = 0.753$$

*που θα μπορούσαμε να υπολογίσουμε κι από τις σχέσεις (3.11) ή (3.12). Ο συντελεστής συσχέτισης δηλώνει την ασθενή θετική συσχέτιση του κέρδους ρεύματος και της αντίστασης στρώματος, που δε μπορούμε όμως να συμπεράνουμε από την τιμή του συντελεστή παλινδρόμησης* $b = 0.063$ *ή την τιμή της διασποράς των σφαλμάτων* $s^2 = 5.056$ *γιατί εξαρτιούνται από τις μονάδες μέτρησης των δύο μεταβλητών.*

Σ' αυτό το κεφάλαιο ασχοληθήκαμε μόνο με την απλή γραμμική παλινδρόμηση. Η μελέτη της παλινδρόμησης επεκτείνεται στη μη-γραμμική παλινδρόμηση, που αποτελεί γενικότερη και πιο ρεαλιστική (αλλά και πιο πολύπλοκη) προσέγγιση για τον προσδιορισμό της εξάρτησης της $Y$ από τη $X$. Επίσης η τ.μ. $Y$ μπορεί να εξαρτάται από περισσότερες από μια μεταβλητές που είναι το πρόβλημα της πολλαπλής παλινδρόμησης (γραμμική ή μη-γραμμική).

# ΑΣΚΗΣΕΙΣ

*Περιγραφική Στατιστική, Εκτίμηση Παραμέτρων*

1. Δείξτε ότι η εκτιμήτρια $s^2$ της διασποράς $\sigma^2$ είναι αμερόληπτη.

2. Δύο τυχαίες μεταβλητές $X_1$ και $X_2$ έχουν κοινή διασπορά $\sigma^2$ και $s_1^2$, $s_2^2$ είναι οι αμερόληπτες δειγματικές διασπορές των $X_1$ και $X_2$, αντίστοιχα, από δείγματα μεγέθους $n_1$ και $n_2$. Δείξτε ότι η εκτιμήτρια $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ της $\sigma^2$ είναι επίσης αμερόληπτη.

3. Για τον έλεγχο της περιεκτικότητας του χάλυβα σε ραδιενέργεια σε δύο εργοστάσια παραγωγής χάλυβα Α και Β έγιναν οι παρακάτω μετρήσεις ραδιενέργειας σε τυχαία δοκίμια χάλυβα (οι μετρήσεις είναι σε Bq/g):

| Δοκίμια | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A (Bq/g) | 0.37 | 0.00 | 0.54 | 0.59 | 0.16 | 0.86 | 0.86 | 0.49 | 0.60 | 0.55 | | | | | |
| B (Bq/g) | 0.24 | 0.52 | 0.12 | 0.95 | 0.26 | 0.33 | 0.62 | 0.32 | 0.27 | 0.05 | 0.39 | 0.10 | 0.51 | 0.79 | 0.09 |

Θεωρείται ότι η διασπορά της ραδιενέργειας στο χάλυβα είναι ίδια για τα δύο εργοστάσια ($\sigma_1^2 = \sigma_2^2 = \sigma^2$).

(α΄) Σχηματίστε το θηκόγραμμα για τα δεδομένα ραδιενέργειας στο χάλυβα των δύο δειγμάτων και σχολιάστε αν η ραδιενέργεια στους χάλυβες των δύο εργοστασίων φαίνεται να ακολουθούν κανονική κατανομή. Σχολιάστε επίσης αν φαίνεται να διαφέρουν αυτές οι δύο κατανομές.

(β΄) Εκτιμείστε τη μέση ραδιενέργεια στο χάλυβα για το εργοστάσιο Α (σημειακή εκτίμηση και $95\%$ διάστημα εμπιστοσύνης).

(γ΄) Κάνετε την ίδια εκτίμηση για το εργοστάσιο Β.

(δ΄) Έστω ότι το ανώτατο επιτρεπτό όριο για τη μέση ραδιενέργεια στο χάλυβα είναι 0.5 Bq/g. Με βάση τα παραπάνω διαστήματα εμπιστοσύνης σε επίπεδο εμπιστοσύνης $95\%$ απαντήστε αν θα γινόταν αποδεκτός στην αγορά ο χάλυβας από το εργοστάσιο Α και από το εργοστάσιο Β.

(ε΄) Ελέγξτε χρησιμοποιώντας διάστημα εμπιστοσύνης σε επίπεδο εμπιστοσύνης $95\%$ αν η μέση ραδιενέργεια στους χάλυβες των δύο εργοστασίων είναι ίδια.

4. **(*)** Έγιναν 15 μετρήσεις της συγκέντρωσης διαλυμένου οξυγόνου (Δ.Ο.) σ΄ ένα ποτάμι (σε $mg/l$)

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Οξυγόνο Α | 1.8 | 2.0 | 2.1 | 1.7 | 1.2 | 2.3 | 2.5 | 2.9 | 1.6 | 2.2 | 2.3 | 1.8 | 2.4 | 1.6 | 1.9 |

Από παλιότερες μετρήσεις γνωρίζουμε ότι η διασπορά του Δ.Ο. είναι $0.1\,(mg/l)^2$.

(α΄) Εκτιμείστε τη διασπορά της συγκέντρωσης Δ.Ο. από το δείγμα καθώς και τα διαστήματα εμπιστοσύνης σε επίπεδο $99\%$ και $90\%$. Εξετάστε και για τα δύο επίπεδα εμπιστοσύνης αν μπορούμε να δεχτούμε την εμπειρική τιμή της διασποράς γι αυτό το δείγμα.

(β΄) Υπολογίστε τα μέτρα θέσης και μεταβλητότητας για τα δεδομένα του δείγματος και σχηματίστε το κατάλληλο θηκόγραμμα. Σχολιάστε αν φαίνεται η συγκέντρωση Δ.Ο. στο νερό του ποταμού να ακολουθεί κανονική κατανομή.

(γ΄) Εκτιμείστε τη μέση συγκέντρωση Δ.Ο. από το δείγμα και δώστε γι αυτήν $95\%$ διάστημα εμπιστοσύνης υποθέτοντας πρώτα ότι η διασπορά είναι γνωστή και μετά χρησιμοποιώντας αυτήν του δείγματος.

(δ΄) Αν υποθέσουμε ότι για ένα εργοστάσιο δίπλα στο ποτάμι είναι σημαντικό η μέση συγκέντρωση Δ.Ο. να μην πέφτει κάτω από $1.8\,mg/l$, θα προκαλούσαν ανησυχία αυτές οι παρατηρήσεις (χρησιμοποιείστε τη διασπορά από το δείγμα);

(ε΄) Αν δε μας ικανοποιεί το πλάτος του τελευταίου παραπάνω διαστήματος και θέλουμε να το μειώσουμε σε $0.2\,mg/l$ πόσες επιπρόσθετες ημερήσιες μετρήσεις πρέπει να γίνουν;

5. **(*)** Σε συνέχεια της Άσκησης 1, σ΄ έναν άλλο ποταμό έγιναν οι παρακάτω παρόμοιες μετρήσεις

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Οξυγόνο Β | 2.3 | 2.1 | 1.9 | 2.6 | 2.9 | 1.5 | 3.1 | 2.1 | 2.7 | 2.3 | 2.6 | 2.5 |

(α΄) Σχηματίζοντας κατάλληλο θηκόγραμμα σχολιάστε αν φαίνεται η συγκέντρωση Δ.Ο. στον ποταμό Β να ακολουθεί κανονική κατανομή. Από τα θηκογράμματα φαίνεται να διαφέρει η συγκέντρωση Δ.Ο. στο νερό των δύο ποταμών;

(β΄) Βρείτε $95\%$ διάστημα εμπιστοσύνης για τη διαφορά των μέσων συγκεντρώσεων Δ.Ο. στα δύο ποτάμια υποθέτοντας πρώτα ότι η διασπορά είναι γνωστή ($0.1\,(mg/l)^2$) και ίδια για τα δύο δείγματα και μετά χρησιμοποιώντας τις εκτιμήσεις των διασπορών από τα δείγματα. Μπορούμε να πούμε πως η μέση συγκέντρωση Δ.Ο. είναι ίδια στα δύο ποτάμια (στην κάθε περίπτωση);

6. Η τάση διακοπής εναλλασσόμενου ρεύματος ενός μονωτικού υγρού δηλώνει τη διηλεκτρική ανθεκτικότητα του. Πήραμε τις παρακάτω παρατηρήσεις της τάσης διακοπής (kV) σε κάποιο κύκλωμα κάτω από ορισμένες συνθήκες.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 41 | 46 | 47 | 47 | 48 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| 48 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 52 | 52 | 53 | 55 |
| 50 | 50 | 50 | 50 | 52 | 52 | 53 | 53 | 53 | 53 | 53 | 57 |
| 52 | 52 | 53 | 53 | 53 | 53 | 53 | 53 | 54 | 54 | 55 | 68 |

(α΄) Σχηματίστε το ιστόγραμμα των δεδομένων της τάσης διακοπής του κυκλώματος και σχολιάστε αν η τάση διακοπής φαίνεται να ακολουθεί κανονική κατανομή.

(β΄) Βρείτε 95% διάστημα εμπιστοσύνης για τη διασπορά της τάσης διακοπής του κυκλώματος.

(γ΄) Από παλιότερες μετρήσεις είχαμε βρει πως η τυπική απόκλιση της τάσης διακοπής παρόμοιου κυκλώματος ήταν περίπου 5 kV. Μπορούμε να δεχθούμε αυτήν την τυπική απόκλιση και γι αυτό το κύκλωμα;

(δ΄) Βρείτε 95% διάστημα εμπιστοσύνης για τη μέση τάση διακοπής του κυκλώματος, υποθέτοντας

(i) τυπική απόκλιση της τάσης διακοπής 5 kV,

(ii) ότι η τυπική απόκλιση της τάσης διακοπής είναι άγνωστη.

(ε΄) Μπορούμε να αποκλείσουμε ότι η μέση τάση διακοπής είναι 52 kV για τις περιπτώσεις (i) και (ii) του προηγούμενου ερωτήματος;

7. Πριν να συμφωνήσει την παραγγελία μιας μεγάλης παρτίδας καλύμματος πολυαιθυλενίου για την προστασία κάποιου ειδικού τύπου καλωδίου, μια εταιρία θέλει να σιγουρευτεί ότι η πραγματική τυπική απόκλιση του πάχους των καλυμμάτων είναι περίπου 0.05 mm. Για το σκοπό αυτό μετρήθηκε το πάχος σε 20 καλύμματα κι οι τιμές (σε mm) είναι

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6.711 | 6.730 | 6.746 | 6.860 | 6.861 | 6.872 | 6.928 | 6.973 | 7.012 | 7.015 |
| 6.746 | 6.860 | 6.861 | 6.872 | 6.928 | 6.973 | 7.012 | 7.015 | 7.053 | 7.359 |

(α΄) Υπολογίστε τα μέτρα θέσης και μεταβλητότητας για τα δεδομένα του δείγματος και σχηματίστε το κατάλληλο θηκόγραμμα. Σχολιάστε αν φαίνεται το πάχος των καλυμμάτων να ακολουθεί κανονική κατανομή.

(β΄) Ελέγξτε με βάση τα κατάλληλα διαστήματα εμπιστοσύνης αν η εταιρία θα δεχτεί ότι το μέσο πάχος των καλυμμάτων είναι 0.05 mm σε επίπεδο εμπιστοσύνης 95% και 99%.

8. Σε δύο τυχαία δείγματα ελασμάτων που χρησιμοποιούνται στην κατασκευή πινάκων κυκλωμάτων μετρήθηκε το ποσό της στρέβλωσης (σε mm) κάθε ελάσματος κάτω από συγκεκριμένες συνθήκες (διαφορετικές για το κάθε ένα από τα δύο δείγματα). Οι μετρήσεις έδωσαν τα παρακάτω αποτελέσματα

| Δείγμα | Πλήθος ελασμάτων | Μέση τιμή (mm) | Τυπική απόκλιση (mm) |
|---|---|---|---|
| 1 | 25 | 0.161 | 0.0165 |
| 2 | 22 | 0.150 | 0.0130 |

Υποθέστε ότι το ποσό στρέβλωσης των ελασμάτων για τις δύο διαφορετικές συνθήκες (με ή χωρίς συγκόλληση) ακολουθεί κανονική κατανομή.

(α΄) Βρείτε 95% διάστημα εμπιστοσύνης για το μέσο ποσό στρέβλωσης των ελασμάτων για τις δύο διαφορετικές συνθήκες ξεχωριστά.

(β′) Ελέγξτε με βάση το 95% διάστημα εμπιστοσύνης αν το μέσο ποσό στρέβλωσης των ελασμάτων είναι το ίδιο κάτω από τις δύο διαφορετικές συνθήκες.

9. Ο υπολογισμός της φόρτωσης του δικτύου υπολογιστών γίνεται εύκολα από τη μέτρηση του χρόνου που κάνει ένα πακέτο πληροφορίας να πάει από μια υπολογιστική μονάδα (host) (μονάδα αποστολής) σε μια άλλη μονάδα (μονάδα παραλαβής) και να επιστρέψει πίσω (ping time). Από την κεντρική υπολογιστική μονάδα του εργαστηρίου και μέσα σε μικρό χρονικό διάστημα μετρήθηκαν οι χρόνοι επιστροφής πακέτων πληροφορίας σε 10 μακρινές μονάδες παραλαβής και δίνονται παρακάτω σε χιλιοστά του δευτερολέπτου (ms).

| Μονάδα παραλαβής | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Χρόνος (ms) | 25 | 38 | 58 | 52 | 69 | 101 | 132 | 148 | 202 | 232 |

(α′) Εκτιμείστε σε επίπεδο εμπιστοσύνης $95\%$ το μέσο χρόνο επιστροφής πακέτων πληροφορίας με βάση το δείγμα, υποθέτοντας πως η κατανομή του χρόνου επιστροφής είναι κανονική.

(β′) Αν το δίκτυο χαρακτηρίζεται υπερφορτωμένο όταν ο μέσος χρόνος επιστροφής πακέτων πληροφορίας ξεπερνάει τα 200 ms θα θεωρούσατε το δίκτυο υπερφορτωμένο τη χρονική περίοδο που έγιναν οι μετρήσεις;

(γ′) Πόσες μετρήσεις ακόμα πρέπει να κάνουμε για να έχουμε ακρίβεια εκτίμησης του μέσου χρόνου επιστροφής πακέτων πληροφορίας $\pm 20$ ms (δηλαδή το εύρος του $95\%$ διαστήματος εμπιστοσύνης να είναι 40 ms);

10. Μετρήθηκε η χωρητικότητα (σε αμπερώρες) 10 μπαταριών και τα αποτελέσματα δίνονται ως εξής:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 140 | 136 | 150 | 144 | 148 | 152 | 138 | 141 | 143 | 151 |

Από παλιότερες μετρήσεις γνωρίζουμε ότι η χωρητικότητα της μπαταρίας ακολουθεί κανονική κατανομή με τυπική απόκλιση 6 αμπερώρες.

(α′) Εκτιμείστε σε επίπεδο εμπιστοσύνης $95\%$ και σε επίπεδο εμπιστοσύνης $99\%$ τη μέση χωρητικότητα της μπαταρίας.

(β′) Η μπαταρία θεωρείται μη αποδεκτή αν η μέση χωρητικότητα της είναι κάτω από 140 αμπερώρες. Με βάση τα διαστήματα εμπιστοσύνης στο (10α′) θα γινόταν η μπαταρία αποδεκτή; [απαντήστε για το κάθε ένα διάστημα εμπιστοσύνης ξεχωριστά].

11. Σ ένα πείραμα μετρήθηκε η αντίσταση 14 καλωδίων τύπου Α και η μέση τιμή αντίστασης καλωδίων του δείγματος βρέθηκε να είναι $\bar{x}_1 = 0.137$ Ohms. Η κατανομή του μεγέθους αντίστασης καλωδίου θεωρείται κανονική και η τυπική απόκλιση της αντίστασης καλωδίου θεωρείται γνωστή και είναι 0.01 Ohms.

(α′) Εκτιμώντας το κατάλληλο διάστημα εμπιστοσύνης, εξετάστε αν μπορούμε να δεχτούμε σε επίπεδο εμπιστοσύνης $95\%$ ότι το καλώδιο τύπου Α έχει μέση αντίσταση 1.40 Ohms.

(β') Σ ένα δεύτερο πείραμα μετρήθηκε επίσης η αντίσταση 10 καλωδίων από έναν άλλο τύπο Β. Η κατανομή και η τυπική απόκλιση της αντίστασης καλωδίου τύπου Β είναι όπως και για το καλώδιο τύπου Α. Η μέση τιμή αντίστασης των 10 καλωδίων τύπου Β του δείγματος βρέθηκε να είναι $\bar{x}_2 = 0.133$ Ohms. Εκτιμώντας το κατάλληλο διάστημα εμπιστοσύνης, εξετάστε αν μπορούμε να δεχτούμε σε επίπεδο εμπιστοσύνης $95\%$ ότι το καλώδιο τύπου Α έχει μεγαλύτερη αντίσταση από το καλώδιο τύπου Β.

12. Σ ένα ολοκληρωμένο κύκλωμα μετρήθηκε η απολαβή κρυσταλλολυχνίας (τρανζίστορ) μεταξύ εκπομπού και δέκτη κατά τη διαδικασία εναπόθεσης. Έγιναν 14 επαναλήψεις της διαδικασίας εναπόθεσης σε διαφορετικές συνθήκες εκπομπής και οι μετρήσεις της απολαβής κρυσταλλολυχνίας (σε hFE) δίνονται παρακάτω:

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1004 | 1636 | 1260 | 1270 | 1272 | 903 | 1555 | 1276 | 1321 | 852 | 1506 | 1269 | 1146 | 1225 |

(α') Για τα παραπάνω δεδομένα, σχηματίστε το θηκόγραμμα, υπολογίζοντας πρώτα τα κατάλληλα συνοπτικά μέτρα, κι εξετάστε αν η κατανομή της απολαβής κρυσταλλολυχνίας φαίνεται να είναι κανονική.

(β') Υποθέτοντας πως η κατανομή της απολαβής κρυσταλλολυχνίας είναι κανονική, εκτιμείστε σε επίπεδο εμπιστοσύνης $95\%$ διάστημα εμπιστοσύνης για τη μέση απολαβή της κρυσταλλολυχνίας.

(γ') Σχολιάστε πως θα εκτιμούσατε το παραπάνω διάστημα εμπιστοσύνης στο (12β'), αν το δείγμα είχε μέγεθος 100 (αντί για 14).

13. Σε 144 δοκιμές σε κάποιο εργαστήριο, 48 κατέληξαν σε ανάφλεξη κάποιου συγκεκριμένου τύπου υποστρώματος από αναμμένο τσιγάρο. Έστω $p$ η αναλογία όλων των δοκιμών που καταλήγουν σε ανάφλεξη του υποστρώματος.

(α') Υπολογίστε το $99\%$ διάστημα εμπιστοσύνης για την αναλογία $p$.

(β') Αν θέλουμε το εύρος του $99\%$ διαστήματος εμπιστοσύνης για την αναλογία να είναι $0.1$ πόσες ακόμα δοκιμές πρέπει να κάνουμε;

14. Σε μια μελέτη για ένα καινούριο σύστημα εκτόξευσης ρουκετών μικρού βεληνεκούς έγιναν δοκιμές με το παλιό και το καινούριο σύστημα. Σε δείγμα 60 πειραματικών εκτοξεύσεων με το παλιό σύστημα 44 ήταν πετυχημένες και σε 80 πειραματικές εκτοξεύσεις με το καινούριο σύστημα 72 ήταν πετυχημένες.

(α') Υπολογίστε τα $95\%$ διαστήματα εμπιστοσύνης για τις αναλογίες πετυχημένων εκτοξεύσεων με τα δύο συστήματα.

(β') Σε επίπεδο εμπιστοσύνης $90\%$ κάνετε στατιστικό έλεγχο για την υπόθεση ότι οι δύο αναλογίες δε διαφέρουν.

(γ') Αν στον παραπάνω έλεγχο βρήκατε ότι υπάρχει διαφορά εκτιμείστε πόση είναι αυτή η διαφορά.

15. Μετρήθηκε ο χρόνος απόκρισης μιας υπολογιστικής μονάδας μέσου δικτύου σε επικοινωνία με την κεντρική υπολογιστική μονάδα. Παρακάτω δίνονται οι μετρήσεις σε 21 διαφορετικές χρονικές στιγμές (σε ms).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 16.1 | 9.6 | 24.9 | 20.4 | 12.7 | 21.2 | 30.2 | 25.8 | 18.5 | 10.3 | 25.3 |
| 14.0 | 27.1 | 45.0 | 23.3 | 24.2 | 14.6 | 8.9 | 32.4 | 11.8 | 28.5 | |

Θεωρούμε ότι ο χρόνος απόκρισης ακολουθεί κανονική κατανομή αλλά με άγνωστη για μας διασπορά.

(α′) Υπολογίστε το $95\%$ διάστημα εμπιστοσύνης καθώς και το $99\%$ διάστημα εμπιστοσύνης για το μέσο χρόνο απόκρισης με βάση το δείγμα.

(β′) Με βάση τα παραπάνω διαστήματα εμπιστοσύνης στο 15α′ (ή αντίστοιχα και ισοδύναμα με κατάλληλο έλεγχο υπόθεσης) εξετάστε αν μπορεί ο μέσος χρόνος απόκρισης να είναι μεγαλύτερος των 30 ms. [Δώστε απαντήσεις σε επίπεδο εμπιστοσύνης $95\%$ και $99\%$.]

16. Μετρήθηκε ο χρόνος ζωής 10 λαμπτήρων τύπου Α σε ώρες:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 430 | 530 | 590 | 580 | 740 | 800 | 830 | 1000 | 1020 | 1100 |

(α′) Για τα παραπάνω δεδομένα, σχηματίστε το θηκόγραμμα, υπολογίζοντας πρώτα τα κατάλληλα συνοπτικά μέτρα, κι εξετάστε αν η κατανομή του χρόνου ζωής των λαμπτήρων φαίνεται να είναι κανονική.

(β′) Υπολογίστε το $95\%$ διάστημα εμπιστοσύνης για την τυπική απόκλιση του χρόνου ζωής των λαμπτήρων τύπου Α.

(γ′) Μετρήθηκε επίσης ο χρόνος ζωής 19 λαμπτήρων τύπου Β σε ώρες και βρέθηκαν τα εξής:
δειγματική μέση τιμή $\bar{x}_2 = 500$
δειγματική τυπική απόκλιση $s_2 = 150$
Υποθέτοντας πως η κατανομή του χρόνου ζωής λαμπτήρων είναι κανονική (και για τους δύο τύπους και ανεξάρτητα από την απάντηση σας στο 16α′), εκτιμείστε σε επίπεδο εμπιστοσύνης $99\%$ αν υπάρχει διαφορά και πόση στο χρόνο ζωής των δύο τύπων λαμπτήρων.

17. Μετρήθηκε η χωρητικότητα (σε αμπερώρες) 100 μπαταριών και υπολογίστηκε για το δείγμα αυτό η μέση χωρητικότητα και τυπική απόκλιση και βρέθηκαν να είναι $\bar{x} = 140$ αμπερώρες και $s = 30$ αμπερώρες, αντίστοιχα.

(α′) Εκτιμείστε σε επίπεδο εμπιστοσύνης $95\%$ και σε επίπεδο εμπιστοσύνης $90\%$ τη μέση χωρητικότητα της μπαταρίας.

(β′) Η μπαταρία θεωρείται μη αποδεκτή αν η μέση χωρητικότητα της είναι κάτω από 135 αμπερώρες. Με βάση τα διαστήματα εμπιστοσύνης που υπολογίσατε στο 17α′ θα γινόταν η μπαταρία αποδεκτή; [απαντήστε για το κάθε ένα επίπεδο εμπιστοσύνης ξεχωριστά]

(γ′) Για να απαντήσουμε με μεγαλύτερη σιγουριά στο ερώτημα 17β′, θέλουμε να μικρύνουμε το εύρος του $95\%$ διαστήματος εμπιστοσύνης σε 6 αμπερώρες (δηλαδή να δηλώσουμε τη μέση χωρητικότητα με ακρίβεια $\pm 3$ αμπερώρες). Για να το πετύχουμε αυτό σε πόσες ακόμα μπαταρίες πρέπει να μετρήσουμε τη χωρητικότητα;

18. Από παλιά μελέτη (πριν αρκετά χρόνια) για το χρόνο λειτουργίας ηλεκτρικής μηχανής μέχρι την πρώτη βλάβη που έγινε σε δείγμα 169 μηχανών, γνωρίζουμε πως το $95\%$ διάστημα εμπιστοσύνης για το μέσο χρόνο μέχρι την πρώτη βλάβη έχει εύρος 20 μέρες.

    (α΄) Πόση ήταν η τυπική απόκλιση του χρόνου μέχρι την πρώτη βλάβη (σε μέρες) στο δείγμα των 169 μηχανών; Στα πλαίσια νέας μελέτης αλλά σε μικρότερη κλίμακα μετρήσαμε το χρόνο μέχρι την πρώτη βλάβη σε 61 ηλεκτρικές μηχανές ίδιου τύπου αλλά νεότερης τεχνολογίας και βρήκαμε το μέσο όρο του χρόνου μέχρι την πρώτη βλάβη $\bar{x} = 720$ μέρες και την τυπική απόκλιση $s = 100$ μέρες.

    (β΄) Υπολογίστε το $95\%$ διάστημα εμπιστοσύνης για το μέσο χρόνο μέχρι την πρώτη βλάβη με βάση το νέο δείγμα.

    (γ΄) Θεωρούμε πως η εκτίμηση της τυπικής απόκλισης του χρόνου μέχρι την πρώτη βλάβη από το νέο δείγμα ($s = 100$) είναι ακριβής (δηλαδή είναι πολύ κοντά στην πραγματική τυπική απόκλιση). Κάνετε τους απαραίτητους υπολογισμούς για να διερευνήσετε αν μπορεί να γίνει η εκτίμηση του μέσου χρόνου μέχρι την πρώτη βλάβη με μεγαλύτερη ακρίβεια (μικρότερο εύρος του $95\%$ διαστήματος εμπιστοσύνης) στις νέες μηχανές απ΄ ότι στις παλιές μηχανές (όπου είχαμε ακρίβεια 10 ημερών με 169 παλιές μηχανές).

19. Εκτενής μελέτη ενός συστήματος υπολογιστών καταμερισμού υπολογιστικού χρόνου (computer time-sharing system) έδειξε ότι ο χρόνος απόκρισης σε κάποια εντολή του λειτουργικού συστήματος ακολουθεί κανονική κατανομή με τυπική απόκλιση 15 millisec.

    (α΄) Θέλουμε να προσδιορίσουμε τον πραγματικό μέσο χρόνο απόκρισης $\mu$ με ακρίβεια 3 millisec σε επίπεδο εμπιστοσύνης $95\%$ (δηλαδή το διάστημα εμπιστοσύνης να έχει εύρος 6 millisec). Πόσες μετρήσεις του χρόνου απόκρισης πρέπει να κάνουμε;

    Δοκιμάστηκε ένα νέο λειτουργικό σύστημα και μετρήθηκε ο χρόνος απόκρισης του λειτουργικού συστήματος σε 61 εντολές του λειτουργικού συστήματος. Από τις 61 μετρήσεις υπολογίστηκαν τα παρακάτω μέτρα γι αυτό το δείγμα των παρατηρήσεων του χρόνου απόκρισης:
    μέση τιμή $\bar{x} = 25$ millisec
    τυπική απόκλιση $s = 18$ millisec.

    (β΄) Υπολογίζοντας κατάλληλο διάστημα εμπιστοσύνης εξετάστε αν σε επίπεδο εμπιστοσύνης $99\%$ μπορούμε να δεχθούμε ότι η τυπική απόκλιση του χρόνου απόκρισης για το νέο λειτουργικό σύστημα παραμένει στο επίπεδο των 15 millisec.

    (γ΄) Ο προμηθευτής του λειτουργικού συστήματος ισχυρίζεται ότι ο μέσος χρόνος απόκρισης είναι 20 millisec. Υπολογίζοντας κατάλληλο διάστημα εμπιστοσύνης για το ίδιο δείγμα εξετάστε αν σε επίπεδο εμπιστοσύνης $90\%$ μπορούμε να δεχθούμε τον ισχυρισμό του προμηθευτή.

*Συσχέτιση - Παλινδρόμηση*

1. **(*)** Η πρώτη ύλη που χρησιμοποιείται στην παραγωγή κάποιας συνθετικής ίνας αποθηκεύεται σε κάποιο χώρο χωρίς έλεγχο υγρασίας. Έγιναν μετρήσεις για 15 μέρες της σχετικής υγρασίας στον αποθηκευτικό χώρο και της περιεκτικότητας σε υγρότητα σε δείγμα της πρώτης ύλης. Τα αποτελέσματα (σε ποσοστά) είναι

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Σχετική υγρασία | 46 | 53 | 29 | 61 | 36 | 39 | 47 | 49 | 52 | 38 | 55 | 32 | 57 | 54 | 44 |
| Υγρότητα | 12 | 15 | 7 | 17 | 10 | 11 | 11 | 12 | 14 | 9 | 16 | 8 | 18 | 14 | 12 |

   (α′) Σχεδιάστε το διάγραμμα διασποράς και υπολογίστε τον συντελεστή συσχέτισης της σχετικής υγρασίας στον αποθηκευτικό χώρο και της περιεκτικότητας σε υγρότητα της πρώτης ύλης. Μπορούμε να υποθέσουμε ότι η συσχέτιση είναι γραμμική;

   (β′) Εκτιμείστε την ευθεία ελαχίστων τετραγώνων και τη διασπορά των σφαλμάτων.

   (γ′) Προβλέψτε αν γίνεται την περιεκτικότητα σε υγρότητα της πρώτης ύλης όταν η σχετική υγρασία στον αποθηκευτικό χώρο είναι 10%, 40% και 70%.

2. Ο χρόνος ως την αποτυχία κάποιου μηχανικού στοιχείου σχετίζεται με την ηλεκτρική τάση λειτουργίας της μηχανής. Εκτελέστηκε ένα σχεδιασμένο πείραμα σ′ ερευνητικό εργαστήριο που έδωσε τα παρακάτω δεδομένα

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Χρόνος αποτυχίας (min) | 2145 | 2155 | 2220 | 2225 | 2260 | 2266 | 2334 | 2340 | 2212 | 2180 |
| Τάση λειτουργίας (αμπέρ) | 110 | 110 | 110 | 110 | 120 | 120 | 120 | 130 | 115 | 115 |

   (α′) Σχεδιάστε το διάγραμμα διασποράς και σχολιάστε αν μπορεί να θεωρηθεί ότι ο χρόνος ως την αποτυχία του μηχανικού στοιχείου εξαρτάται γραμμικά από την τάση λειτουργίας της μηχανής.

   (β′) Υποθέτοντας γραμμική παλινδρόμηση του χρόνου αποτυχίας στην τάση λειτουργίας, εκτιμείστε την ευθεία ελαχίστων τετραγώνων και τη διασπορά των σφαλμάτων.

   (γ′) Σχολιάστε αν η γραμμική εξάρτηση είναι ισχυρή υπολογίζοντας πρώτα το συντελεστή συσχέτισης.

   (δ′) Μπορούμε να υποθέσουμε ότι ο μέσος χρόνος αποτυχίας για τάση λειτουργίας 115 αμπέρ είναι μικρότερος από 2000 min; Μπορεί να είναι μικρότερος από 2200 min; Ποια είναι η καλύτερη εκτίμηση του μέσου χρόνου αποτυχίας γι αυτήν την τάση με βάση τη γραμμική παλινδρόμηση κι αυτό το δείγμα;

3. Το κόστος παραγωγής ισχύος (ανά κιλοβατώρα) πιστεύεται ότι εξαρτάται (εκτός άλλων παραγόντων) από το κόστος του λιγνίτη (σε cent ανά εκατομμύριο Btu). Οι παρακάτω παρατηρήσεις έγιναν από 12 σπαστήρες λιγνίτη

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Κόστος λιγνίτη | 14 | 16 | 22 | 24 | 20 | 29 | 26 | 15 | 29 | 24 | 25 | 13 |
| Κόστος ισχύος | 4.1 | 4.4 | 5.6 | 5.1 | 5.0 | 5.3 | 5.4 | 4.8 | 6.1 | 5.5 | 4.7 | 3.9 |

(α′) Σχεδιάστε το διάγραμμα διασποράς και υπολογίστε τον συντελεστή συσχέτισης του κόστους της παραγωγής ισχύος και του κόστους του λιγνίτη. Μπορούμε να υποθέσουμε ότι υπάρχει συσχέτιση;

(β′) Εκτιμείστε την ευθεία ελαχίστων τετραγώνων και τη διασπορά των σφαλμάτων.

(γ′) Με βάση το μοντέλο παλινδρόμησης, αν το κόστος λιγνίτη είναι 18 cent/millBtu θα μπορούσε το τυπικό κόστος ισχύος ανά κιλοβατώρα να είναι 6 μονάδες; Θα μπορούσε για το ίδιο κόστος λιγνίτη να είναι 5 μονάδες;

4. Θέλουμε να μελετήσουμε κατά πόσο εξαρτάται ο χρόνος επιστροφής πακέτων πληροφορίας (ping time) από τον αριθμό των ενδιαμέσων υπολογιστικών μονάδων της διαδρομής της πληροφορίας στο δίκτυο από τη μονάδα αποστολής στη μονάδα παραλαβής. Από την κεντρική υπολογιστική μονάδα του εργαστηρίου και μέσα σε μικρό χρονικό διάστημα μετρήθηκαν οι χρόνοι επιστροφής (σε χιλιοστά του δευτερολέπτου, ms) πακέτων πληροφορίας στην ίδια μακρινή μονάδα παραλαβής αλλά με διαφορετικό αριθμό ενδιαμέσων σταθμών. Τα αποτελέσματα δίνονται παρακάτω

| Ενδιάμεσοι σταθμοί | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|---|---|
| Χρόνος (ms) | 25 | 38 | 58 | 52 | 69 | 101 | 132 | 148 | 202 | 232 |

(α′) Σχηματίστε το κατάλληλο διάγραμμα διασποράς και σχολιάστε αν η υπόθεση, δηλαδή ότι ο χρόνος επιστροφής πακέτων πληροφορίας εξαρτάται γραμμικά από τον αριθμό των ενδιαμέσων υπολογιστικών μονάδων της διαδρομής, φαίνεται σωστή με βάση το δείγμα των παρατηρήσεων από τους δύο πίνακες.

(β′) Για το παραπάνω πρόβλημα στο ερώτημα (4α′), υπολογίστε τις σημειακές εκτιμήσεις $a$ και $b$ των παραμέτρων $\alpha$ και $\beta$ της ευθείας παλινδρόμησης (με τη μέθοδο των ελαχίστων τετραγώνων).

(γ′) Προβλέψτε το χρόνο επιστροφής πακέτων πληροφορίας (αν γίνεται) όταν δεν υπάρχει ενδιάμεσος σταθμός κι όταν υπάρχουν 5 ενδιάμεσοι σταθμοί.

5. Η απευθείας μέτρηση της ποσότητας πρωτεΐνης σε δείγμα (κομμάτι) συκωτιού είναι δύσκολη κι απαιτεί πολύ χρόνο. Πιστεύεται ότι η ποσότητα της πρωτεΐνης σχετίζεται με την ποσότητα του φωτός που θα απορροφιόταν από το δείγμα συκωτιού. Γι αυτό έγινε στο εργαστήριο το ακόλουθο πείραμα. Στάλθηκε από ένα φασματόμετρο φως σε διάλυμα που περιείχε δείγμα συκωτιού και μετρήθηκε το φως που απορροφήθηκε. Αυτή η διαδικασία εφαρμόσθηκε σε 5 δείγματα συκωτιού για τα οποία η ποσότητα της πρωτεΐνης ήταν γνωστή. Τα αποτελέσματα δίνονται στον παρακάτω πίνακα.

| Φως που απορροφήθηκε | 0.44 | 0.82 | 1.20 | 1.61 | 1.83 |
|---|---|---|---|---|---|
| Ποσότητα πρωτεΐνης (mg) | 2 | 16 | 30 | 46 | 55 |

(α′) Σχηματίστε το κατάλληλο διάγραμμα διασποράς και σχολιάστε αν η υπόθεση, ότι η ποσότητα πρωτεΐνης σε δείγμα συκωτιού εξαρτάται γραμμικά από το αντίστοιχο φως που απορροφιέται από το δείγμα, φαίνεται σωστή με βάση το δείγμα των παρατηρήσεων από τον πίνακα.

(β′) Για το παραπάνω πρόβλημα στο ερώτημα (5α′), υπολογίστε τις σημειακές εκτιμήσεις $a$ και $b$ των παραμέτρων $\alpha$ και $\beta$ της ευθείας παλινδρόμησης (με τη μέθοδο των ελαχίστων τετραγώνων) και προβλέψτε την ποσότητα πρωτεΐνης σε δείγμα συκωτιού για το οποίο η απορρόφηση του φωτός είναι 1.5.

6. Στα πλαίσια μιας μελέτης σε πετρελαιοκίνητες μηχανές (diesel) ελαφρών φορτηγών για την έκλυση πρωτοξείδιου του αζώτου, διερευνήθηκε η εξάρτηση της από την υγρασία. Έγινε γι αυτό ένα πείραμα και μετρήθηκε σε 10 διαφορετικές χρονικές στιγμές η υγρασία (σε ποσοστό επί των κορεσμένων ατμών) και η έκλυση πρωτοξείδιου του αζώτου (σε ppm). Οι μετρήσεις δίνονται στον παρακάτω πίνακα:

| Υγρασία | 8.3 | 10.7 | 12.9 | 20.1 | 24.0 | 34.3 | 35.1 | 41.6 | 72.2 | 72.4 |
|---|---|---|---|---|---|---|---|---|---|---|
| Άζωτο | 1.15 | 1.00 | 1.10 | 1.03 | 1.07 | 0.96 | 0.89 | 0.91 | 0.77 | 0.90 |

(α′) Σχηματίστε το κατάλληλο διάγραμμα διασποράς και σχολιάστε αν η υπόθεση, ότι η έκλυση πρωτοξείδιου του αζώτου $Y$ εξαρτάται γραμμικά από την υγρασία $X$ φαίνεται σωστή με βάση το δείγμα των παρατηρήσεων.

(β′) Για το παραπάνω πρόβλημα στο ερώτημα (6α′), υπολογίστε τις σημειακές εκτιμήσεις $a$ και $b$ των παραμέτρων $\alpha$ και $\beta$ της ευθείας παλινδρόμησης (με τη μέθοδο των ελαχίστων τετραγώνων). Σχηματίστε την ευθεία ελαχίστων τετραγώνων στο διάγραμμα διασποράς και προβλέψτε την έκλυση μονοξειδίου του αζώτου για υγρασία $40\%$.

7. Πιστεύεται ότι σ′ ένα ολοκληρωμένο κύκλωμα η απολαβή κρυσταλλολυχνίας (τρανζίστορ) μεταξύ εκπομπού και δέκτη κατά τη διαδικασία εναπόθεσης εξαρτάται από τη δόση εκπομπής, η οποία ελέγχεται κατά τη διαδικασία εναπόθεσης. Έγιναν 14 επαναλήψεις της διαδικασίας εναπόθεσης σε διαφορετικές συνθήκες εκπομπής και οι μετρήσεις της απολαβής κρυσταλλολυχνίας (σε hFE) και της δόσης εκπομπής (σε ιόντα) δίνονται στον παρακάτω πίνακα:

| Δόση εκπομπής | Απολαβή |
|---|---|
| 4.00 | 1004 |
| 4.00 | 1636 |
| 4.00 | 1260 |
| 4.10 | 1270 |
| 4.20 | 1272 |
| 4.30 | 903 |
| 4.30 | 1555 |
| 4.30 | 1276 |
| 4.30 | 1321 |
| 4.60 | 852 |
| 4.60 | 1506 |
| 4.60 | 1269 |
| 4.70 | 1146 |
| 4.72 | 1225 |

(α′) Σχηματίστε το κατάλληλο διάγραμμα διασποράς και σχολιάστε αν η υπόθεση, ότι η απολαβή κρυσταλλολυχνίας εξαρτάται γραμμικά από τη δόση εκπομπής φαίνεται σωστή με βάση το δείγμα των παρατηρήσεων.

(β′) Για το παραπάνω πρόβλημα στο ερώτημα (7α′), υπολογίστε τις σημειακές εκτιμήσεις $a$ και $b$ των παραμέτρων $\alpha$ και $\beta$ της ευθείας παλινδρόμησης (με τη μέθοδο των ελαχίστων τετραγώνων), καθώς και την εκτίμηση $s^2$ της διασποράς των σφαλμάτων $\sigma^2$. Σχολιάστε με βάση τις εκτιμήσεις αυτές, αν μπορούμε να κάνουμε ακριβείς προβλέψεις της απολαβής κρυσταλλολυχνίας όταν γνωρίζουμε τη δόση εκπομπής για τιμές δόσης εκπομπής μεταξύ 4.00 και 4.70.

8. Στον παρακάτω πίνακα δίνεται για 10 σταθμούς ο αριθμός των ημερών σ′ ένα χρόνο που η θερμοκρασία έπεσε κάτω από $0^o$C και το υψόμετρο τους.

| Υψόμετρο (m) | 1000 | 1050 | 1110 | 1220 | 1320 | 1380 | 1420 | 1560 | 1670 | 1950 |
|---|---|---|---|---|---|---|---|---|---|---|
| Αριθμός ημερών | 32 | 29 | 36 | 38 | 43 | 53 | 52 | 63 | 73 | 100 |

(α′) Υποθέτουμε ότι ο αριθμός των ημερών $Y$ εξαρτάται γραμμικά από το υψόμετρο $X$ [$\mathrm{E}(Y|X = x) = \alpha + \beta x$]. Σχηματίστε το κατάλληλο διάγραμμα διασποράς και σχολιάστε αν αυτή η υπόθεση φαίνεται σωστή με βάση το δείγμα των παρατηρήσεων του πίνακα.

(β′) Υπολογίστε τις σημειακές εκτιμήσεις $a$ και $b$ των παραμέτρων $\alpha$ και $\beta$ της ευθείας παλινδρόμησης (με τη μέθοδο των ελαχίστων τετραγώνων).

(γ′) Με βάση το δείγμα, μπορείτε να εκτιμήσετε το μέσο αριθμό ημερών το χρόνο που η θερμοκρασία πέφτει κάτω από $0^o$C σε υψόμετρο 1500m; Σε υψόμετρο 2200m;

9. Θέλουμε να διερευνήσουμε αν η μηνιαία κατανάλωση ηλεκτρικού ρεύματος (σε χιλιάδες ΩΧΒ) σε μια χημική βιομηχανία εξαρτάται γραμμικά από την παραγωγή του προϊόντος (σε τόνους). Οι μετρήσεις σε 9 μήνες δίνονται παρακάτω:

| Προϊόν | 98 | 104 | 105 | 106 | 107 | 109 | 110 | 115 | 120 |
|---|---|---|---|---|---|---|---|---|---|
| Κατανάλωση | 27 | 30 | 29 | 43 | 30 | 31 | 35 | 35 | 39 |

(α′) Σχηματίστε το κατάλληλο διάγραμμα διασποράς και σχολιάστε αν η υπόθεση, ότι η κατανάλωση ηλεκτρικού ρεύματος εξαρτάται γραμμικά από την παραγωγή του προϊόντος, φαίνεται σωστή με βάση το δείγμα.

(β′) Για το παραπάνω πρόβλημα στο ερώτημα 9α′, υπολογίστε τις σημειακές εκτιμήσεις $a$ και $b$ των παραμέτρων $\alpha$ και $\beta$ της ευθείας παλινδρόμησης (με τη μέθοδο των ελαχίστων τετραγώνων).

(γ′) Προβλέψτε την κατανάλωση ρεύματος για παραγωγή 110 τόνων και παραγωγή 150 τόνων του προϊόντος.

10. Στα πλαίσια μιας μελέτης για την ποιότητα ελασμάτων που χρησιμοποιούνται στην κατασκευή πινάκων κυκλωμάτων μετρήθηκε σε 9 ελάσματα το ποσό στρέβλωσης (σε mm) κάτω από συγκεκριμένη θερμοκρασία περιβάλλοντος για το κάθε έλασμα. Οι μετρήσεις δίνονται στον παρακάτω πίνακα:

| Θερμοκρασία (σε $^o$C) | 15 | 20 | 23 | 28 | 30 | 30 | 32 | 36 | 40 |
|---|---|---|---|---|---|---|---|---|---|
| Στρέβλωση (σε mm) | 0.11 | 0.08 | 0.17 | 0.12 | 0.16 | 0.14 | 0.18 | 0.12 | 0.17 |

(α′) Σχηματίστε το κατάλληλο διάγραμμα διασποράς και σχολιάστε αν η υπόθεση, ότι η στρέβλωση ελάσματος εξαρτάται γραμμικά από την θερμοκρασία φαίνεται σωστή με βάση το δείγμα των παρατηρήσεων.

(β′) Για το παραπάνω πρόβλημα στο ερώτημα 10α′, υπολογίστε τις σημειακές εκτιμήσεις $a$ και $b$ των παραμέτρων $\alpha$ και $\beta$ της ευθείας παλινδρόμησης (με τη μέθοδο των ελαχίστων τετραγώνων).

(γ′) Σχηματίστε την ευθεία ελαχίστων τετραγώνων στο διάγραμμα διασποράς και προβλέψτε τη στρέβλωση ελάσματος για θερμοκρασία $30^o$C.

11. Σε ένα φορτηγό μηχανής diesel έγιναν μετρήσεις της εκπομπής νιτρικού οξέος και της υγρασίας σε 10 διαφορετικές συνθήκες υγρασίας, και τα αποτελέσματα δίνονται παρακάτω.

| Υγρασία (σε %) | 72 | 41 | 34 | 35 | 10 | 12 | 8 | 20 | 72 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|
| Νιτρικό οξύ (σε ppm) | 0.90 | 0.91 | 0.96 | 0.89 | 1.00 | 1.10 | 1.15 | 1.03 | 0.77 | 1.07 |

(α′) Σχηματίστε το κατάλληλο διάγραμμα διασποράς και σχολιάστε αν φαίνεται να εξαρτάται, και πως, η εκπομπή νιτρικού οξέος από την υγρασία.

(β′) Υπολογίστε τις σημειακές εκτιμήσεις $a$ και $b$ των παραμέτρων $\alpha$ και $\beta$ της ευθείας παλινδρόμησης (με τη μέθοδο των ελαχίστων τετραγώνων) για το πρόβλημα της γραμμικής εξάρτησης της εκπομπής νιτρικού οξέος από την υγρασία. Σχηματίστε την ευθεία ελαχίστων τετραγώνων στο διάγραμμα διασποράς.

12. Ένας οδηγός ισχυρίζεται ότι η κατανάλωση καυσίμου του αυτοκινήτου του δεν εξαρτάται από την ταχύτητα του αυτοκινήτου. Για να ελέγξει αυτήν την υπόθεση, οδήγησε το αυτοκίνητο στον αυτοκινητόδρομο με σταθερή ταχύτητα και για ταχύτητες από 80km/h μέχρι 110km/h. Μέτρησε επίσης την κατανάλωση καυσίμου (σε km/l) σε κάθε περίπτωση και τα αποτελέσματα δίνονται στον παρακάτω πίνακα.

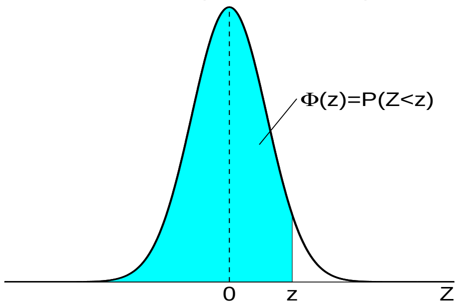
| Ταχύτητα (km/h) | 80 | 85 | 90 | 95 | 100 | 105 | 110 |
|---|---|---|---|---|---|---|---|
| Κατανάλωση καυσίμου (km/l) | 9.1 | 8.7 | 9.2 | 8.5 | 8.0 | 8.7 | 8.3 |

(α′) Σχηματίστε το κατάλληλο διάγραμμα διασποράς και σχολιάστε αν η υπόθεση του οδηγού φαίνεται σωστή και αν όχι, τότε σχολιάστε πως φαίνεται να εξαρτάται η κατανάλωση καυσίμου από την ταχύτητα που τρέχει το αυτοκίνητο.

(β′) Για το παραπάνω πρόβλημα υπολογίστε τις σημειακές εκτιμήσεις $a$ και $b$ των παραμέτρων $\alpha$ και $\beta$ της ευθείας παλινδρόμησης (με τη μέθοδο των ελαχίστων τετραγώνων). Σχηματίστε την ευθεία ελαχίστων τετραγώνων στο διάγραμμα διασποράς.

(γ′) Υποδείξτε αν μπορούμε να κάνουμε πρόβλεψη της κατανάλωσης καυσίμου για ταχύτητα 130 km/h, και αν ναι υπολογίστε την πρόβλεψη της μέσης κατανάλωσης καυσίμου σε αυτή την περίπτωση.

## Στατιστικός Πίνακας Τυπικής Κανονική Κατανομής

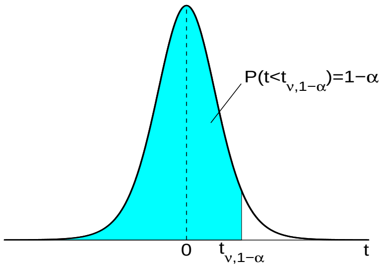



Παράδειγμα:

$$z = 1.28 \Longleftrightarrow \Phi(z) = 0.90$$
$$z = 1.65 \Longleftrightarrow \Phi(z) = 0.95$$
$$z = 2.33 \Longleftrightarrow \Phi(z) = 0.99$$
$$z = 3.08 \Longleftrightarrow \Phi(z) = 0.999$$

| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7852 |
| 0.8 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8810 | 0.8830 |
| 1.2 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.3 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.4 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.5 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.6 | 0.9452 | 0.9463 | 0.9474 | 0.9484 | 0.9495 | 0.9505 | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.7 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.8 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.9 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 | 0.9767 |
| 2.0 | 0.9772 | 0.9778 | 0.9783 | 0.9788 | 0.9793 | 0.9798 | 0.9803 | 0.9808 | 0.9812 | 0.9817 |
| 2.1 | 0.9821 | 0.9826 | 0.9830 | 0.9834 | 0.9838 | 0.9842 | 0.9846 | 0.9850 | 0.9854 | 0.9857 |
| 2.2 | 0.9861 | 0.9864 | 0.9868 | 0.9871 | 0.9875 | 0.9878 | 0.9881 | 0.9884 | 0.9887 | 0.9890 |
| 2.3 | 0.9893 | 0.9896 | 0.9898 | 0.9901 | 0.9904 | 0.9906 | 0.9909 | 0.9911 | 0.9913 | 0.9916 |
| 2.4 | 0.9918 | 0.9920 | 0.9922 | 0.9925 | 0.9927 | 0.9929 | 0.9931 | 0.9932 | 0.9934 | 0.9936 |
| 2.5 | 0.9938 | 0.9940 | 0.9941 | 0.9943 | 0.9945 | 0.9946 | 0.9948 | 0.9949 | 0.9951 | 0.9952 |
| 2.6 | 0.9953 | 0.9955 | 0.9956 | 0.9957 | 0.9959 | 0.9960 | 0.9961 | 0.9962 | 0.9963 | 0.9964 |
| 2.7 | 0.9965 | 0.9966 | 0.9967 | 0.9968 | 0.9969 | 0.9970 | 0.9971 | 0.9972 | 0.9973 | 0.9974 |
| 2.8 | 0.9974 | 0.9975 | 0.9976 | 0.9977 | 0.9977 | 0.9978 | 0.9979 | 0.9979 | 0.9980 | 0.9981 |
| 2.9 | 0.9981 | 0.9982 | 0.9982 | 0.9983 | 0.9984 | 0.9984 | 0.9985 | 0.9985 | 0.9986 | 0.9986 |
| 3.0 | 0.9987 | 0.9987 | 0.9987 | 0.9988 | 0.9988 | 0.9989 | 0.9989 | 0.9989 | 0.9990 | 0.9990 |
| 3.1 | 0.9990 | 0.9991 | 0.9991 | 0.9991 | 0.9992 | 0.9992 | 0.9992 | 0.9992 | 0.9993 | 0.9993 |
| 3.2 | 0.9993 | 0.9993 | 0.9994 | 0.9994 | 0.9994 | 0.9994 | 0.9994 | 0.9995 | 0.9995 | 0.9995 |
| 3.3 | 0.9995 | 0.9995 | 0.9995 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9997 |
| 3.4 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9998 |

## Στατιστικός Πίνακας Κατανομής Student

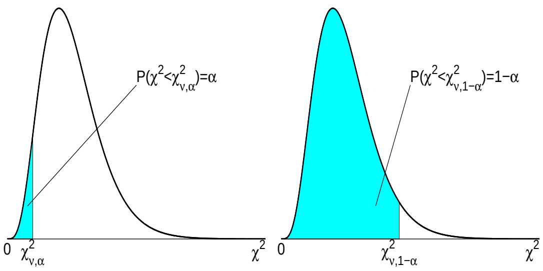



$\nu$: βαθμοί ελευθερίας
$1-\alpha$: τιμή της αθροιστικής συνάρτησης

Παράδειγμα:
$\nu = 10,\ \alpha = 0.05 \implies \mathrm{t}_{10,0.95} = 1.812$
$\nu = 10,\ \alpha = 0.001 \implies \mathrm{t}_{10,0.999} = 4.144$
$\nu = 20,\ \alpha = 0.025 \implies \mathrm{t}_{20,0.975} = 2.086$

| | $1-\alpha$ | | | | | | |
|---|---|---|---|---|---|---|---|
| $\nu$ | 0.900 | 0.950 | 0.975 | 0.990 | 0.995 | 0.999 | 0.9995 |
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 318.309 | 636.619 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 | 31.599 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 | 12.924 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 8.610 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 | 6.869 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 5.959 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 | 5.408 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 | 5.041 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 | 4.781 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 4.587 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 | 4.437 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 | 4.318 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 | 4.221 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 | 4.140 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 | 4.073 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 | 4.015 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 | 3.965 |
| 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 | 3.922 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 | 3.883 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 3.850 |
| 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 | 3.819 |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 | 3.792 |
| 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 | 3.768 |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 | 3.745 |
| 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 | 3.725 |
| 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 | 3.707 |
| 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 | 3.690 |
| 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 | 3.674 |
| 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 | 3.659 |
| 30 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 | 3.646 |
| 32 | 1.309 | 1.694 | 2.037 | 2.449 | 2.738 | 3.365 | 3.622 |
| 34 | 1.307 | 1.691 | 2.032 | 2.441 | 2.728 | 3.348 | 3.601 |
| 36 | 1.306 | 1.688 | 2.028 | 2.434 | 2.719 | 3.333 | 3.582 |
| 38 | 1.304 | 1.686 | 2.024 | 2.429 | 2.712 | 3.319 | 3.566 |
| 40 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.307 | 3.551 |
| 50 | 1.299 | 1.676 | 2.009 | 2.403 | 2.678 | 3.261 | 3.496 |
| 60 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.232 | 3.460 |
| 100 | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 | 3.174 | 3.390 |
| 120 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 | 3.160 | 3.373 |
| 150 | 1.287 | 1.655 | 1.976 | 2.351 | 2.609 | 3.145 | 3.357 |
| 200 | 1.286 | 1.653 | 1.972 | 2.345 | 2.601 | 3.131 | 3.340 |
| 300 | 1.284 | 1.650 | 1.968 | 2.339 | 2.592 | 3.118 | 3.323 |
| 400 | 1.284 | 1.649 | 1.966 | 2.336 | 2.588 | 3.111 | 3.315 |
| $\infty$ | 1.282 | 1.646 | 1.962 | 2.330 | 2.581 | 3.098 | 3.300 |

# Στατιστικός Πίνακας Κατανομής $\mathcal{X}^2$




$\nu$: βαθμοί ελευθερίας

$\alpha$ ή $1-\alpha$: τιμή αθροιστικής συνάρτησης

Παράδειγμα:

$$\nu = 10, \alpha = 0.05 \implies \begin{array}{l} \chi^2_{10,0.05} = 3.94 \\ \chi^2_{10,0.95} = 18.31 \end{array}$$

$$\nu = 18, \alpha = 0.025 \implies \begin{array}{l} \chi^2_{18,0.025} = 8.23 \\ \chi^2_{18,0.975} = 31.53 \end{array}$$

| | $\alpha$ | | | | | $1-\alpha$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\nu$ | 0.001 | 0.005 | 0.010 | 0.025 | 0.050 | 0.950 | 0.975 | 0.990 | 0.995 | 0.999 |
| 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 3.84 | 5.02 | 6.63 | 7.88 | 10.83 |
| 2 | 0.00 | 0.01 | 0.02 | 0.05 | 0.10 | 5.99 | 7.38 | 9.21 | 10.60 | 13.82 |
| 3 | 0.02 | 0.07 | 0.11 | 0.22 | 0.35 | 7.81 | 9.35 | 11.34 | 12.84 | 16.27 |
| 4 | 0.09 | 0.21 | 0.30 | 0.48 | 0.71 | 9.49 | 11.14 | 13.28 | 14.86 | 18.47 |
| 5 | 0.21 | 0.41 | 0.55 | 0.83 | 1.15 | 11.07 | 12.83 | 15.09 | 16.75 | 20.52 |
| 6 | 0.38 | 0.68 | 0.87 | 1.24 | 1.64 | 12.59 | 14.45 | 16.81 | 18.55 | 22.46 |
| 7 | 0.60 | 0.99 | 1.24 | 1.69 | 2.17 | 14.07 | 16.01 | 18.48 | 20.28 | 24.32 |
| 8 | 0.86 | 1.34 | 1.65 | 2.18 | 2.73 | 15.51 | 17.53 | 20.09 | 21.95 | 26.12 |
| 9 | 1.15 | 1.73 | 2.09 | 2.70 | 3.33 | 16.92 | 19.02 | 21.67 | 23.59 | 27.88 |
| 10 | 1.48 | 2.16 | 2.56 | 3.25 | 3.94 | 18.31 | 20.48 | 23.21 | 25.19 | 29.59 |
| 11 | 1.83 | 2.60 | 3.05 | 3.82 | 4.57 | 19.68 | 21.92 | 24.72 | 26.76 | 31.26 |
| 12 | 2.21 | 3.07 | 3.57 | 4.40 | 5.23 | 21.03 | 23.34 | 26.22 | 28.30 | 32.91 |
| 13 | 2.62 | 3.57 | 4.11 | 5.01 | 5.89 | 22.36 | 24.74 | 27.69 | 29.82 | 34.53 |
| 14 | 3.04 | 4.07 | 4.66 | 5.63 | 6.57 | 23.68 | 26.12 | 29.14 | 31.32 | 36.12 |
| 15 | 3.48 | 4.60 | 5.23 | 6.26 | 7.26 | 25.00 | 27.49 | 30.58 | 32.80 | 37.70 |
| 16 | 3.94 | 5.14 | 5.81 | 6.91 | 7.96 | 26.30 | 28.85 | 32.00 | 34.27 | 39.25 |
| 17 | 4.42 | 5.70 | 6.41 | 7.56 | 8.67 | 27.59 | 30.19 | 33.41 | 35.72 | 40.79 |
| 18 | 4.90 | 6.26 | 7.01 | 8.23 | 9.39 | 28.87 | 31.53 | 34.81 | 37.16 | 42.31 |
| 19 | 5.41 | 6.84 | 7.63 | 8.91 | 10.12 | 30.14 | 32.85 | 36.19 | 38.58 | 43.82 |
| 20 | 5.92 | 7.43 | 8.26 | 9.59 | 10.85 | 31.41 | 34.17 | 37.57 | 40.00 | 45.31 |
| 21 | 6.45 | 8.03 | 8.90 | 10.28 | 11.59 | 32.67 | 35.48 | 38.93 | 41.40 | 46.80 |
| 22 | 6.98 | 8.64 | 9.54 | 10.98 | 12.34 | 33.92 | 36.78 | 40.29 | 42.80 | 48.27 |
| 23 | 7.53 | 9.26 | 10.20 | 11.69 | 13.09 | 35.17 | 38.08 | 41.64 | 44.18 | 49.73 |
| 24 | 8.08 | 9.89 | 10.86 | 12.40 | 13.85 | 36.42 | 39.36 | 42.98 | 45.56 | 51.18 |
| 25 | 8.65 | 10.52 | 11.52 | 13.12 | 14.61 | 37.65 | 40.65 | 44.31 | 46.93 | 52.62 |
| 26 | 9.22 | 11.16 | 12.20 | 13.84 | 15.38 | 38.89 | 41.92 | 45.64 | 48.29 | 54.05 |
| 27 | 9.80 | 11.81 | 12.88 | 14.57 | 16.15 | 40.11 | 43.19 | 46.96 | 49.64 | 55.48 |
| 28 | 10.39 | 12.46 | 13.56 | 15.31 | 16.93 | 41.34 | 44.46 | 48.28 | 50.99 | 56.89 |
| 29 | 10.99 | 13.12 | 14.26 | 16.05 | 17.71 | 42.56 | 45.72 | 49.59 | 52.34 | 58.30 |
| 30 | 11.59 | 13.79 | 14.95 | 16.79 | 18.49 | 43.77 | 46.98 | 50.89 | 53.67 | 59.70 |
| 32 | 12.81 | 15.13 | 16.36 | 18.29 | 20.07 | 46.19 | 49.48 | 53.49 | 56.33 | 62.49 |
| 34 | 14.06 | 16.50 | 17.79 | 19.81 | 21.66 | 48.60 | 51.97 | 56.06 | 58.96 | 65.25 |
| 36 | 15.32 | 17.89 | 19.23 | 21.34 | 23.27 | 51.00 | 54.44 | 58.62 | 61.58 | 67.99 |
| 38 | 16.61 | 19.29 | 20.69 | 22.88 | 24.88 | 53.38 | 56.90 | 61.16 | 64.18 | 70.70 |
| 40 | 17.92 | 20.71 | 22.16 | 24.43 | 26.51 | 55.76 | 59.34 | 63.69 | 66.77 | 73.40 |
| 50 | 24.67 | 27.99 | 29.71 | 32.36 | 34.76 | 67.50 | 71.42 | 76.15 | 79.49 | 86.66 |
| 60 | 31.74 | 35.53 | 37.48 | 40.48 | 43.19 | 79.08 | 83.30 | 88.38 | 91.95 | 99.61 |
| 100 | 61.92 | 67.33 | 70.06 | 74.22 | 77.93 | 124.34 | 129.56 | 135.81 | 140.17 | 149.45 |
| 120 | 77.76 | 83.85 | 86.92 | 91.57 | 95.70 | 146.57 | 152.21 | 158.95 | 163.65 | 173.62 |
| 150 | 102.11 | 109.14 | 112.67 | 117.98 | 122.69 | 179.58 | 185.80 | 193.21 | 198.36 | 209.26 |
| 200 | 143.84 | 152.24 | 156.43 | 162.73 | 168.28 | 233.99 | 241.06 | 249.45 | 255.26 | 267.54 |
| 300 | 229.96 | 240.66 | 245.97 | 253.91 | 260.88 | 341.40 | 349.87 | 359.91 | 366.84 | 381.43 |
| 400 | 318.26 | 330.90 | 337.16 | 346.48 | 354.64 | 447.63 | 457.31 | 468.72 | 476.61 | 493.13 |
| $\infty$ | 867.48 | 888.56 | 898.91 | 914.26 | 927.59 | 1074.68 | 1089.53 | 1106.97 | 1118.95 | 1143.92 |