Κεφάλαιο 3 Εισαγωγικές έννοιες στατιστικής

Η στατιστική είναι εφαρμοσμένος κλάδος της πιθανοθεωρίας ο οποίος ασχολείται με πραγματικά προβλήματα, επιδιώκοντας την εξαγωγή συμπερασμάτων βασισμένων σε παρατηρήσεις. Τα συμπεράσματα αφορούν ένα *πληθυσμό*, ενώ εξάγονται με βάση ένα *δείγμα*. Αν και το περιεχόμενο του όρου *πληθυσμός* δεν ορίζεται με ενιαίο και αυστηρό τρόπο στη στατιστική βιβλιογραφία, γενικά θεωρούμε ότι ο όρος περιγράφει οποιαδήποτε συλλογή αντικειμένων των οποίων ενδιαφέρουν οι μετρήσιμες ιδιότητες. Ο πληθυσμός μπορεί να είναι συγκεκριμένος και πεπερασμένος (π.χ. ο πληθυσμός των ετήσιων απορροών του υδρολογικού έτους 1990-91 για όλες τις υδρολογικές λεκάνες της γης με μέγεθος άνω των 100 km^2) ή άπειρος και αφαιρετικά ορισμένος (π.χ. ο πληθυσμός των δυνατών ετήσιων απορροών μιας υδρολογικής λεκάνης). Ο όρος *δείγμα* αναφέρεται σε ένα σύνολο μετρήσεων για το συγκεκριμένο πληθυσμό (βλ. και ορισμό στο εδάφιο 3.1.1).

Το πιο σημαντικό αντικείμενο της στατιστικής είναι η *εκτίμηση*. Διακρίνεται σε *εκτίμηση παραμέτρων* και *πρόγνωση*. Για να διευκρινίσουμε τις έννοιες αυτές, θεωρούμε κάποιο πληθυσμό που αντιπροσωπεύεται από την τυχαία μεταβλητή X με συνάρτηση κατανομής $F(X, \eta)$, όπου η παράμετρος. Στην περίπτωση που η παράμετρος είναι άγνωστη και επιδιώκουμε την εκτίμηση της, έχουμε ένα πρόβλημα εκτίμησης παραμέτρων. Στην περίπτωση που είναι γνωστή η παράμετρος και επιδιώκουμε την εκτίμηση της μεταβλητής X ή μιας συνάρτησής της, έχουμε ένα πρόβλημα πρόγνωσης. Όπως θα δούμε παρακάτω, και τα δύο

προβλήματα αντιμετωπίζονται με παρόμοιες μεθόδους της στατιστικής, γι' αυτό άλλωστε αποδίδονται και τα δύο με τον όρο *εκτίμηση*.

Η εκτίμηση μπορεί να είναι *σημειακή* ή *διαστήματος*. Στην πρώτη περίπτωση το ζητούμενο μέγεθος περιγράφεται με μια αριθμητική τιμή. Στη δεύτερη περιγράφεται από ένα διάστημα μέσα στο οποίο περιέχεται το ζητούμενο μέγεθος με δεδομένο βαθμό ακρίβειας ή αξιοπιστίας. Αντίστροφα, για ένα δεδομένο διάστημα διακύμανσης ενός μεγέθους, η στατιστική μπορεί να υπολογίσει τον αντίστοιχο βαθμό ακρίβειας ή, αντίθετα, το βαθμό αβεβαιότητας που υπάρχει γύρω από το ζητούμενο μέγεθος.

Μια άλλη σπουδαία περιοχή της στατιστικής είναι ο *έλεγχος υποθέσεων*, που αποτελεί τη βάση της *θεωρίας αποφάσεων*. Η διαδικασία του ελέγχου προϋποθέτει τη διατύπωση δύο υποθέσεων: μιας βασικής H_0, που αναφέρεται με τον όρο *μηδενική υπόθεση*, και μιας *εναλλακτικής υπόθεσης* H_1. Ξεκινάμε τη διαδικασία ελέγχου θεωρώντας ότι η μηδενική υπόθεση είναι σωστή και χρησιμοποιούμε τις παρατηρήσεις, προκειμένου να αποφασίσουμε αν η υπόθεση αυτή πρέπει να απορριφθεί. Αυτό γίνεται με τη χρήση στατιστικών θεωρήσεων. Παρόλο που και ο έλεγχος υποθέσεων βασίζεται στο ίδιο θεωρητικό υπόβαθρο όπως η εκτίμηση, η διαφορά εδώ είναι ότι εξετάζουμε δύο εναλλακτικά μοντέλα, ενώ στην εκτίμηση χρησιμοποιούμε ένα μόνο μοντέλο.

Το βασικό υπόβαθρο για όλα τα παραπάνω δίνεται στις ενότητες αυτού του κεφαλαίου, αναλυτικότερα όμως εφαρμόζεται στα κεφάλαια που ακολουθούν, όπου δίνονται και συγκεκριμένα αριθμητικά παραδείγματα. Βεβαίως, υπάρχουν και άλλες περιοχές της στατιστικής, όπως για παράδειγμα η ανάλυση Bayes, που δεν καλύπτονται σε αυτό το κείμενο.

3.1 Εννοιολογία και ορισμοί

3.1.1 Δείγμα

Θεωρούμε μια τυχαία μεταβλητή X με πυκνότητα πιθανότητας $f(X)$, η οποία ορίζεται βάσει ενός δειγματικού χώρου Ω. Θα θεωρούμε ότι η μεταβλητή αυτή ταυτίζεται εννοιολογικά με κάποιο πληθυσμό. *Δείγμα* μεγέθους (ή μήκους) n της μεταβλητής είναι μια ακολουθία n ανεξάρτητων τυχαίων μεταβλητών $X_1, X_2, \ldots, X_n$ με κοινή πυκνότητα πιθανότητας $f(x)$

που ορίζεται στο δειγματικό χώρο $\Omega^n = \Omega \times \cdots \times \Omega$ (Papoulis, 1990, σ. 238). Καθεμιά από τις μεταβλητές X_i αντιστοιχεί στις δυνατές εκβάσεις μιας μέτρησης ή παρατήρησης της μεταβλητής X. Αφού εκτελεστούν οι μετρήσεις, για καθεμιά μεταβλητή θα έχουμε *μία* μέτρηση, άρα συνολικά θα έχουμε την αριθμητική ακολουθία $x_1, x_2, \ldots, x_n$, την οποία λέμε *παρατηρημένο δείγμα* ή απλώς *παρατηρήσεις*.

Η έννοια του δείγματος είναι, λοιπόν, συνυφασμένη με δύο ειδών ακολουθίες: μια αφαιρετική ακολουθία τυχαίων μεταβλητών και την αντίστοιχη συγκεκριμένη ακολουθία των αριθμητικών τιμών τους. Πολλές φορές στην τεχνική υδρολογία ο όρος *δείγμα* χρησιμοποιείται αδιακρίτως και για τη δεύτερη ακολουθία, με παράλειψη του όρου *παρατηρημένο*. Για το λόγο αυτό πρέπει να είμαστε προσεκτικοί ώστε να διακρίνουμε κάθε φορά αν ο όρος αναφέρεται με την αφαιρετική ή τη συγκεκριμένη έννοιά του.

Η λήψη ενός δείγματος μεγέθους n ή *δειγματοληψία* στη στατιστική γίνεται κατά κανόνα με την εκτέλεση n επαναλήψεων ενός πειράματος τύχης, δηλαδή είναι μια πειραματική διαδικασία. Οι επαναλήψεις πρέπει να είναι ανεξάρτητες μεταξύ τους και να εκτελούνται κάτω από ουσιαστικώς ισοδύναμες συνθήκες. Αντίθετα, στην τεχνική υδρολογία δεν υπάρχει η δυνατότητα του πειράματος, και έτσι η δειγματοληψία είναι διαδικασία πολλαπλών μετρήσεων ενός φυσικού φαινομένου σε διάφορες χρονικές στιγμές. Κατά συνέπεια δεν είναι δυνατό να εξασφαλιστούν πάντα οι προϋποθέσεις της ανεξαρτησίας και των ισοδύναμων συνθηκών. Ωστόσο, για ορισμένες υδρολογικές μεταβλητές εξασφαλίζονται κατά προσέγγιση οι προϋποθέσεις αυτές (οι οποίες ισοδυναμούν με αυτές της ανεξαρτησίας, στασιμότητας και εργοδικότητας, βλ. κεφάλαιο 4) και έτσι μπορούμε να χρησιμοποιούμε για τη μελέτη τους τις τρέχουσες μεθόδους της στατιστικής.

3.1.2 Στατιστική συνάρτηση

Με τον όρο *στατιστική συνάρτηση* εννοούμε κάθε συνάρτηση των τυχαίων μεταβλητών του δείγματος, της μορφής $\Theta = g(X_1, \ldots, X_n)$. Από τις παρατηρήσεις του δείγματος μπορούμε να υπολογίσουμε άμεσα την αριθμητική τιμή $\theta = g(x_1, \ldots, x_n)$ της στατιστικής συνάρτησης. Προφανώς η στατιστική συνάρτηση δεν ταυτίζεται με την αριθμητική της τιμή, αφού η πρώτη, ως συνάρτηση τυχαίων μεταβλητών, είναι και η ίδια τυχαία

μεταβλητή, με συγκεκριμένη συνάρτηση κατανομής. Ενώ η αριθμητική τιμή της στατιστικής συνάρτησης υπολογίζεται με απλό αριθμητικό τρόπο από τις παρατηρήσεις του δείγματος, η συνάρτηση κατανομή της συνάγεται βάσει θεωρημάτων της στατιστικής, όπως θα δούμε σε επόμενες ενότητες. Παραδείγματα τυπικών στατιστικών συναρτήσεων δίνονται στα επόμενα εδάφια.

3.1.3 Εκτιμήτριες και εκτιμήσεις

Οι στατιστικές συναρτήσεις χρησιμοποιούνται για την εκτίμηση παραμέτρων του πληθυσμού. Για κάθε παράμετρο η του πληθυσμού μπορούν να βρεθούν μία ή περισσότερες στατιστικές συναρτήσεις της μορφής $\Theta = g(X_1, \ldots, X_n)$ κατάλληλες για την εκτίμηση αυτής της παραμέτρου. Σε αυτή την περίπτωση λέμε ότι η $\Theta = g(X_1, \ldots, X_n)$ είναι *εκτιμήτρια* της παραμέτρου η και ότι η αριθμητική τιμή της $\theta = g(x_1, \ldots, x_n)$ αποτελεί *εκτίμηση* της η.

Δεν υπάρχει αυστηρό κριτήριο για το πότε μια στατιστική συνάρτηση μπορεί να χρησιμοποιηθεί για την εκτίμηση μιας παραμέτρου ενός πληθυσμού. Πολύ συχνά ο σχηματισμός τέτοιων συναρτήσεων είναι εμπειρικός και συνηθέστατα ο τύπος της στατιστικής συνάρτησης Θ είναι ο ίδιος με αυτόν που δίνει την παράμετρο η για μια τυχαία μεταβλητή με διακριτό και πεπερασμένο σύνολο τιμών. Για παράδειγμα, έστω ότι μας ενδιαφέρει να βρούμε μια εκτιμήτρια της μέσης τιμής $\eta = m_X$ μιας μεταβλητής X, με βάση το δείγμα $(X_1, \ldots, X_n)$ με παρατηρήσεις $(x_1, \ldots, x_n)$. Θεωρούμε προς στιγμήν ότι η X, είναι διακριτή μεταβλητή με δυνατές τιμές $(x_1, \ldots, x_n)$, σε καθεμιά από τις οποίες αντιστοιχεί η ίδια πιθανότητα $P(X = x_i) = 1/n$. Από τον τύπο της μέσης τιμής διακριτής μεταβλητής (εξισώσεις (2.16) και (2.18)) βρίσκουμε $\eta = (x_1 + \cdots + x_n)/n$. Στην τελευταία εξίσωση αντικαθιστούμε τις αριθμητικές τιμές με τις αντίστοιχες μεταβλητές και παίρνουμε τη στατιστική συνάρτηση $\Theta = (X_1 + \cdots + X_n)/n$. Όπως θα δούμε και παρακάτω, αυτή η στατιστική συνάρτηση είναι πράγματι εκτιμήτρια της μέσης τιμής οποιασδήποτε μεταβλητής, λέγεται δειγματική μέση τιμή και συμβολίζεται με $\overline{X}$. Πάντως, αυτή η εμπειρική μέθοδος δεν δίνει πάντα την καλύτερη δυνατή εκτιμήτρια.

Παρόλο που, όπως είδαμε παραπάνω, η εκτιμήτρια στη γενική περίπτωση δεν ορίζεται με αυστηρό μαθηματικό τρόπο, διάφορες κατηγορίες εκτιμητριών έχουν αυστηρούς ορισμούς. Έτσι:

1. Μια στατιστική συνάρτηση Θ είναι *αμερόληπτη εκτιμήτρια* της παραμέτρου η αν $E[\Theta] = \eta$. Διαφορετικά είναι *μεροληπτική εκτιμήτρια* και η διαφορά $E[\Theta] - \eta$ λέγεται *μεροληψία*.
2. Μια στατιστική συνάρτηση Θ είναι *συνεπής εκτιμήτρια* της παραμέτρου η αν το σφάλμα εκτίμησης $\Theta - \eta$ τείνει στο μηδέν με πιθανότητα 1 όταν $n \to \infty$. Διαφορετικά είναι *ασυνεπής εκτιμήτρια*.
3. Μια στατιστική συνάρτηση Θ είναι *βέλτιστη εκτιμήτρια* της παραμέτρου η αν το μέσο τετραγωνικό σφάλμα εκτίμησης $(\Theta - \eta)^2$ είναι ελάχιστο.
4. Μια στατιστική συνάρτηση Θ είναι *η πιο αποτελεσματική εκτιμήτρια* της παραμέτρου η αν είναι αμερόληπτη και έχει την ελάχιστη διασπορά.

Ας σημειωθεί ότι η εκτιμήτρια $\overline{X}$ του πιο πάνω παραδείγματος είναι αμερόληπτη και συνεπής, και μάλιστα για ορισμένες συναρτήσεις κατανομής είναι ταυτόχρονα και βέλτιστη και η πιο αποτελεσματική.

Στην πράξη γίνεται προσπάθεια να χρησιμοποιούνται αμερόληπτες και συνεπείς εκτιμήτριες, ενώ ο υπολογισμός της βέλτιστης και της πιο αποτελεσματικής εκτιμήτριας έχει περισσότερο θεωρητικό ενδιαφέρον. Για κάθε παράμετρο μπορεί να υπάρχουν περισσότερες από μία αμερόληπτες ή συνεπείς εκτιμήτριες. Συχνά ο προσδιορισμός αμερόληπτων εκτιμητριών είναι αρκετά δύσκολος, οπότε καταφεύγουμε στη χρήση μεροληπτικών.

3.1.4 Εκτίμηση διαστήματος και όρια εμπιστοσύνης

Εκτίμηση διαστήματος μιας παραμέτρου η είναι ένα διάστημα της μορφής (θ_1, θ_2), όπου $\theta_1 = g_1(x_1, \ldots, x_n)$ και $\theta_2 = g_2(x_1, \ldots, x_n)$ είναι συναρτήσεις των παρατηρήσεων του δείγματος. Το διάστημα (Θ_1, Θ_2) που ορίζουν οι αντίστοιχες στατιστικές συναρτήσεις $\Theta_1 = g_1(X_1, \ldots, X_n)$ και $\Theta_2 = g_2(X_1, \ldots, X_n)$ λέγεται *εκτιμήτρια διαστήματος* της η.

Το διάστημα (Θ_1, Θ_2) ονομάζεται *διάστημα εμπιστοσύνης γ της παραμέτρου η* αν

$$P(\Theta_1 < \eta < \Theta_2) = \gamma \tag{3.1}$$

όπου γ είναι δεδομένη σταθερά ($0 < \gamma < 1$). Η σταθερά αυτή ονομάζεται *συντελεστής εμπιστοσύνης* και κατά κανόνα παίρνει τιμές κοντά στο 1 (π.χ. 0.9, 0.95, 0.99, έτσι ώστε η πιθανότητα στην (3.1) να γίνεται "σχεδόν βεβαιότητα"). Η διαφορά $\alpha = 1 - \gamma$ ονομάζεται *επίπεδο εμπιστοσύνης* και τα όρια Θ_1 και Θ_2 ονομάζονται *όρια εμπιστοσύνης*. Καταχρηστικώς, ο όρος *όρια εμπιστοσύνης* χρησιμοποιείται στην πράξη και για τις αριθμητικές τιμές θ_1 και θ_2 των στατιστικών συναρτήσεων και το ίδιο συμβαίνει και για τον όρο *διάστημα εμπιστοσύνης*.

Προκειμένου να δώσουμε ένα γενικό τρόπο υπολογισμού του διαστήματος εμπιστοσύνης, ας θεωρήσουμε ότι η στατιστική συνάρτηση $\Theta = g(X_1, \ldots, X_n)$ είναι αμερόληπτη (σημειακή) εκτιμήτρια της παραμέτρου η και ότι η συνάρτηση κατανομής της είναι η $F_\Theta(\theta)$. Με βάση αυτή τη συνάρτηση κατανομής μπορούν να υπολογιστούν δύο θετικοί αριθμοί ξ_1 και ξ_2, έτσι ώστε το σφάλμα εκτίμησης $\Theta - \eta$ να βρίσκεται στο διάστημα $(-\xi_1, \xi_2)$ με πιθανότητα γ, ήτοι

$$P(\eta - \xi_1 < \Theta < \eta + \xi_2) = \gamma \tag{3.2}$$

και παράλληλα το διάστημα $(-\xi_1, \xi_2)$ να είναι το ελάχιστο δυνατό.* Η τελευταία εξίσωση μετασχηματίζεται άμεσα στην

$$P(\Theta - \xi_1 < \eta < \Theta + \xi_2) = a \tag{3.3}$$

Κατά συνέπεια τα ζητούμενα όρια εμπιστοσύνης είναι $\Theta_1 = \Theta - \xi_1$ και $\Theta_2 = \Theta + \xi_2$.

Παρόλο που οι εξισώσεις (3.2) και (3.3) είναι ισοδύναμες, η στατιστική ερμηνεία τους είναι διαφορετική. Η πρώτη είναι προγνωστική, με την έννοια ότι μας δίνει όρια εμπιστοσύνης† για την τυχαία μεταβλητή Θ. Η δεύτερη είναι εκτιμητική, αφού μας δίνει όρια εμπιστοσύνης για την άγνωστη παράμετρο η, η οποία προφανώς *δεν* είναι τυχαία μεταβλητή.

---

* Αν η κατανομή της Q είναι συμμετρική τότε η απαίτηση αυτή ικανοποιείται για $\xi_1 = \xi_2$. Διαφορετικά, επειδή είναι δύσκολος ο υπολογισμός του ελάχιστου διαστήματος για ασύμμετρες κατανομές, απλοποιούμε το πρόβλημα διασπώντας την (3.2) στις εξισώσεις $P(\Theta < \eta - \xi_1) = P(\Theta > \eta + \xi_2) = (1 - \gamma) / 2$.

† Οι όροι *όρια εμπιστοσύνης*, *διάστημα εμπιστοσύνης*, *συντελεστής εμπιστοσύνης* κτλ. χρησιμοποιούνται και για αυτή την προγνωστική μορφή της εξίσωσης.

3.2 Τυπικές σημειακές εκτιμήτριες

Στην ενότητα αυτή δίνουμε τις πιο τυπικές εκτιμήτριες που αναφέρονται σε στατιστικές ροπές του πληθυσμού και είναι ανεξάρτητες της συνάρτησης κατανομής του $F(x)$. Συγκεκριμένα, δίνουμε εκτιμήτριες για τη μέση τιμή, τη διασπορά και την τρίτη κεντρική ροπή μιας μεταβλητής. Δεν επεκτεινόμαστε σε ροπές μεγαλύτερης τάξης, γιατί και η εξαγωγή αμερόληπτων εκτιμητριών είναι αρκετά πολύπλοκη, αλλά και η αποτελεσματικότητά τους είναι περιορισμένη για μικρά δείγματα (δηλαδή οι εκτιμήτριες έχουν μεγάλη διασπορά) έτσι ώστε οι αντίστοιχες εκτιμήσεις να είναι ασαφείς. Γι' αυτό, άλλωστε, στην τεχνική υδρολογία δεν χρησιμοποιούνται ροπές τάξης μεγαλύτερης από τρία. Ακόμη και η εκτίμηση της τρίτης ροπής είναι αρκετά ασαφής για μικρά δείγματα, αλλά όμως αυτή χρησιμοποιείται συχνά στην τεχνική υδρολογία για το λόγο ότι περιγράφει την ασυμμετρία της κατανομής, που είναι βασικό χαρακτηριστικό για τις, κατά κανόνα θετικά ασύμμετρες, υδρολογικές μεταβλητές.

Πέρα από τις παραπάνω ροπές μιας μεταβλητής, δίνουμε και εκτιμήτριες της συνδιασποράς και του συντελεστή συσχέτισης δύο μεταβλητών, που χρησιμοποιούνται κατά την ταυτόχρονη στατιστική ανάλυση δύο μεταβλητών.

3.2.1 Γενικά για τις εκτιμήτριες ροπών

Οι εκτιμήτριες των ροπών περί την αρχή για μία ή δύο μεταβλητές, ήτοι των $m_X^{(r)}$ και $m_{XY}^{(rs)}$ (όπου τα r και s είναι τυχόντες ακέραιοι), κατασκευασμένες με την εμπειρική μέθοδο που περιγράφηκε στο εδάφιο 3.1.3, είναι οι εξής:

$$\widetilde{M}_X^{(r)} = \frac{\sum_{i=1}^{n} X_i{}^r}{n} \qquad \widetilde{M}_{XY}^{(rs)} = \frac{\sum_{i=1}^{n} X_i{}^r Y_i^s}{n} \tag{3.4}$$

Αποδεικνύεται (Kendall and Stewart, 1968, σ. 229) ότι

$$E\left[\widetilde{M}_X^{(r)}\right] = m_X^{(r)} \qquad E\left[\widetilde{M}_{XY}^{(rs)}\right] = m_{XY}^{(rs)} \tag{3.5}$$

Κατά συνέπεια οι εκτιμήτριες είναι αμερόληπτες. Οι διασπορές των εκτιμητριών αυτών είναι

$$\mathrm{Var}\left[\tilde{M}_X^{(r)}\right]=\frac{1}{n}\left[m_X^{(2r)}-\left(m_X^{(r)}\right)^2\right] \qquad E\left[\tilde{M}_{XY}^{(rs)}\right]=\frac{1}{n}\left[m_{XY}^{(2r,2s)}-\left(m_{XY}^{(rs)}\right)^2\right] \quad (3.6)$$

Οι τελευταίες εξισώσεις δείχνουν ότι η διασπορές τείνουν στο 0 όταν $n \to \infty$ και κατά συνέπεια ότι οι εκτιμήτριες είναι συνεπείς.

Για την κατασκευή εκτιμητριών των κεντρικών ροπών $\mu_X^{(r)}$ ή $\mu_{XY}^{(rs)}$ μίας ή δύο μεταβλητών, αντίστοιχα, χρησιμοποιούνται ως βάση οι στατιστικές συναρτήσεις $\hat{M}_X^{(r)}$ και $\hat{M}_{XY}^{(rs)}$ που ορίζονται από τις εξισώσεις

$$\hat{M}_X^{(r)}=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^r}{n} \qquad \hat{M}_{XY}^{(rs)}=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^r\left(Y_i-\overline{Y}\right)^s}{n} \quad (3.7)$$

Οι εκτιμήτριες αυτές έχουν κατασκευαστεί με την εμπειρική μέθοδο που περιγράφηκε στο εδάφιο 3.1.3. Ωστόσο, είναι μεροληπτικές εκτιμήτριες (για $r + s > 1$). Αν $\hat{M}_X^{*(r)}$ και $\hat{M}_{XY}^{*(rs)}$ είναι κάποιες αμερόληπτες εκτιμήτριες των ίδιων ροπών του πληθυσμού, τότε οι λόγοι

$$\phi\left(\hat{M}_X^{(r)},n\right)=\frac{\hat{M}_X^{*(r)}}{\hat{M}_X^{(r)}} \qquad \phi\left(\hat{M}_{XY}^{(rs)},n\right)=\frac{\hat{M}_{XY}^{*(rs)}}{\hat{M}_{XY}^{(rs)}} \quad (3.8)$$

ονομάζονται *συντελεστές διόρθωσης μεροληψίας*. Αντιστρέφοντας και γενικεύοντας το συμβολισμό, αν δοθεί μια μεροληπτική εκτιμήτρια Θ και ο αντίστοιχος συντελεστής διόρθωσης μεροληψίας $\phi(\Theta, n)$ τότε προσδιορίζεται άμεσα και η αντίστοιχη αμερόληπτη εκτιμήτρια $\Theta^* = \phi(\Theta, n)\ \Theta$. Ο προσδιορισμός συντελεστών διόρθωσης μεροληψίας δεν είναι πάντα εύκολος και γι' αυτό σε ορισμένες περιπτώσεις δίνονται προσεγγιστικοί συντελεστές. Επίσης, για τυχούσα μη γραμμική συνάρτηση $g()$, η $g(\Theta^*)$ δεν παραμένει γενικά αμερόληπτη, γεγονός που δυσκολεύει τον προσδιορισμό συντελεστών διόρθωσης για μετασχηματισμούς μιας ή περισσότερων στατιστικών συναρτήσεων.

3.2.2 Δειγματική μέση τιμή

Η πιο κοινή στατιστική συνάρτηση είναι η δειγματική μέση τιμή, η οποία, όπως είδαμε και παραπάνω, αποτελεί εκτιμήτρια της μέσης τιμής και ορίζεται από τη σχέση

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.9}$$

που αποτελεί ειδική περίπτωση της (3.4) για $r = 1$. Η αριθμητική τιμή της

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.10}$$

ονομάζεται *παρατηρημένη* ή *αριθμητική μέση τιμή* ή απλώς *μέσος όρος* του δείγματος. Τα μεγέθη $\overline{X}$ και $\overline{x}$ δεν πρέπει να συγχέονται εννοιολογικά μεταξύ τους, ούτε με τη μέση τιμή της μεταβλητής X, δηλαδή $m_X = E[X]$ που ορίζεται με βάση τις εξισώσεις (2.15) ή (2.16) και (2.18). Ωστόσο τα τρία μεγέθη είναι στενά συνδεδεμένα. Εφαρμόζοντας τις (3.5) και (3. 6) βρίσκουμε

$$E\left[\overline{X}\right] = E\left[X\right] \qquad \mathrm{Var}\left[\overline{X}\right] = \frac{\mathrm{Var}\left[X\right]}{n} \tag{3.11}$$

και αυτό ανεξάρτητα από το ποια είναι η κατανομή της X. Οι παραπάνω εξισώσεις δείχνουν ότι η εκτιμήτρια είναι αμερόληπτη και συνεπής.

3.2.3 Διασπορά και τυπική απόκλιση

Η μεροληπτική εκτιμήτρια της διασποράς σ_X^2 του πληθυσμού είναι η ακόλουθη:

$$S_X^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n} \tag{3.12}$$

για την οποία αποδεικνύεται (Kendall and Stewart, 1968, σ. 277) ότι

$$E\left[S_X^2\right]=\frac{n-1}{n}\sigma_X^2$$

$$Var\left[S_X^2\right]=\frac{\mu_X^{(4)}-\sigma_X^4}{n}-\frac{2\left(\mu_X^{(4)}-2\sigma_X^4\right)}{n^2}+\frac{\mu_X^{(4)}-3\sigma_X^4}{n^3} \tag{3.13}$$

όπου $\mu_X^{(4)}$ η τέταρτη κεντρική ροπή του πληθυσμού. Στην τελευταία εξίσωση μπορούν να παραλειφθούν οι δύο τελευταίοι όροι για μεγάλο n. Από την πρώτη από τις πιο πάνω εξισώσεις προκύπτει άμεσα ότι ο συντελεστής διόρθωσης μεροληψίας είναι

$$\phi\left(S_X^2,n\right)=\frac{n}{n-1} \tag{3.14}$$

και κατά συνέπεια η αμερόληπτη (και συνεπής) εκτιμήτρια της σ_X^2 είναι η ακόλουθη, γνωστή ως *δειγματική διασπορά*:

$$S_X^{*2}=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n-1} \tag{3.15}$$

Για μεγάλο μέγεθος δείγματος n οι δύο εκτιμήτριες πρακτικώς ταυτίζονται. Στην περίπτωση κανονικής κατανομής του πληθυσμού αποδεικνύεται ότι

$$\mathrm{Var}\left[S_X^2\right]=\frac{2(n-1)\sigma_X^4}{n^2} \qquad \mathrm{Var}\left[S_X^{*2}\right]=\frac{2\sigma_X^4}{(n-1)} \tag{3.16}$$

Ως εκτιμήτριες της τυπικής απόκλισης του πληθυσμού χρησιμοποιούνται οι τετραγωνικές ρίζες των παραπάνω, δηλαδή οι S_X και S_X*, οι οποίες πάντως δεν είναι αμερόληπτες. Έτσι (Yevjevich, 1972, σ. 193· Kendall and Stewart, 1968, σ. 233)

$$E\left[S_X\right]=\sigma_X+O\left(\frac{1}{n}\right) \qquad Var\left[S_X\right]=\frac{\mu_X^{(4)}-\sigma_X^4}{4\sigma_X^2\,n}+O\left(\frac{1}{n^2}\right) \tag{3.17}$$

όπου οι όροι $O(1/n)$ και $O(1/n^2)$ είναι ποσότητες ανάλογες των $1/n$ και $1/n^2$, αντίστοιχα, και μπορούν να παραλειφθούν για μέγεθος δείγματος σχετικά μεγάλο ($n \geq 20$). Στην περίπτωση κανονικής κατανομής του πληθυσμού, για την S_X χρησιμοποιούνται οι ακόλουθες προσεγγιστικές εξισώσεις

$$E[S_X] \approx \sigma_X \sqrt{\frac{n-1}{n}} \qquad Var[S_X] \approx \frac{\sigma_X^2}{2\,n} \tag{3.18}$$

Τα σφάλματα αυτών των εξισώσεων είναι μικρότερα από 2.5% και 2.7%, αντίστοιχα, για $n \geq 10$ και πρακτικώς μηδενίζονται για $n \geq 100$. Οι αντίστοιχες εξισώσεις για την S_X^* είναι

$$E\left[S_X^*\right] \approx \sigma_X \qquad Var\left[S_X^*\right] \approx \frac{\sigma_X^2}{2\,(n-1)} \tag{3.19}$$

ενώ ακριβέστερες προσεγγίσεις δίνουν οι

$$E\left[S_X^*\right] \approx \sigma_X \sqrt{\frac{n-\frac{5}{4}}{n-\frac{3}{4}}} \qquad Var\left[S_X^*\right] \approx \frac{\sigma_X^2}{2\left(n-\frac{3}{4}\right)} \tag{3.20}$$

των οποίων τα σφάλματα είναι μικρότερα από 0.005% και 0.2%, αντίστοιχα, για $n \geq 10$.*

Τέλος, για το συντελεστή μεταβλητότητας χρησιμοποιείται μια από τις δύο εκτιμήτριες

$$\hat{C}_{v_X} = \frac{S_X}{\overline{X}} \qquad \hat{C}^*_{v_X} = \frac{S_X^*}{\overline{X}} \tag{3.21}$$

Αν η X είναι θετική, τότε αποδεικνύεται ότι οι εκτιμήτριες αυτές είναι άνω φραγμένες ($\hat{C}_{v_X} \leq \sqrt{n-1}$) χωρίς να ισχύει το ίδιο και για τις αντί-

---

* Παραθέτουμε εδώ και τις ακριβείς εξισώσεις, που είναι

$$E\left[S_X^*\right] = \sigma_X \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\frac{n-1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)} \qquad Var\left[S_X^*\right] = \sigma_X^2 \left[1 - \frac{\Gamma^2\left(\frac{n}{2}\right)}{\frac{n-1}{2}\,\Gamma^2\left(\frac{n-1}{2}\right)}\right]$$

στοιχες παραμέτρους του πληθυσμού. Αυτό εισάγει ασφαλώς μεροληψία, η οποία πάντως για τις τρέχουσες εφαρμογές είναι αμελητέα.*

3.2.4 Τρίτη κεντρική ροπή και συντελεστής ασυμμετρίας

Η μεροληπτική εκτιμήτρια της τρίτης κεντρικής ροπής $\mu_X^{(3)}$ του πληθυσμού είναι η ακόλουθη

$$\hat{M}_X^{(3)} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{n} \tag{3.22}$$

για την οποία αποδεικνύεται (Kendall and Stewart, σ. 278-281) ότι†

$$E\left[\hat{M}_X^{(3)}\right] = \frac{(n-1)(n-2)}{n^2}\mu_X^{(3)} \tag{3.23}$$

Από την παραπάνω εξίσωση προκύπτει άμεσα ότι ο συντελεστής διόρθωσης μεροληψίας είναι‡

---

* Η έκφραση για τη διασπορά της εκτιμήτριας είναι αρκετά πολύπλοκη και παραλείπεται (βλ. Kendall and Stewart, 1968, σ. 233). Στην περίπτωση κανονικής κατανομής της X ισχύει

$$\mathrm{Var}\left[\hat{C}_{v_X}\right] \approx C_{v_X}^2 / 2n$$

† Η έκφραση για τη διασπορά της εκτιμήτριας είναι αρκετά πολύπλοκη:

$$\mathrm{Var}\left[\hat{M}_X^{(3)}\right] \approx \frac{1}{n}\left[\mu_X^{(6)} - \left(\mu_X^{(3)}\right)^2 - 6\mu_X^{(4)}\sigma_X^2 + 9\sigma_X^6\right]$$

Για κανονική κατανομή της X ισχύει (Kendall and Stewart, 1968, σ. 243)

$$\mathrm{Var}\left[\hat{M}_X^{(3)}\right] = 6\,\sigma_X^6 / n$$

‡ Όπως προαναφέρθηκε, για μεγαλύτερης τάξης ροπές, ο υπολογισμός αμερόληπτων εκτιμητριών καθίσταται ιδιαίτερα πολύπλοκος. Για παράδειγμα, ο συντελεστής διόρθωσης μεροληψίας της τέταρτης κεντρικής ροπής είναι

$$\phi\left(\hat{M}_X^{(4)}, n\right) = \frac{n^3}{(n-1)(n-2)(n-3)}\left[1 - \frac{2n-3}{n^2}\left(\frac{3}{\hat{C}_{k_X}} + 1\right)\right]$$

όπου

$$\phi\left(\hat{M}_X^{(3)}, n\right) = \frac{n^2}{(n-1)(n-2)} \tag{3.24}$$

και κατά συνέπεια η αμερόληπτη (και συνεπής) εκτιμήτρια της $\mu_X^{(3)}$ είναι η

$$\hat{M}_X^{*(3)} = \frac{n\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{(n-1)(n-2)} \tag{3.25}$$

Για μεγάλο μέγεθος δείγματος *n* οι δύο εκτιμήτριες πρακτικώς ταυτίζονται.

Για την εκτίμηση του συντελεστή ασυμμετρίας C_{s_X} του πληθυσμού χρησιμοποιείται κατά βάση η εκτιμήτρια

$$\hat{C}_{s_X} = \frac{\hat{M}_X^{(3)}}{S_X^3} \tag{3.26}$$

η οποία δεν είναι αμερόληπτη. Η μεροληψία δεν οφείλεται μόνο στο γεγονός ότι οι δύο εκτιμήτριες ροπών από τις οποίες υπολογίζεται (αριθμητής και παρονομαστής) δεν είναι αμερόληπτες, αλλά κυρίως στο γεγονός ότι η $\hat{C}_{s_X}$ είναι άνω και κάτω φραγμένη σε αντίθεση με την C_{s_X} του πληθυσμού που δεν είναι φραγμένη. Αυτό οφείλεται στο πεπερασμένο μέγεθος του δείγματος, το οποίο καθορίζει και το άνω και κάτω όριο. Έτσι αποδεικνύεται (Kirby, 1974· Wallis et al., 1974) ότι $\hat{C}_{s_X} \le (n-2)/\sqrt{n-1}$.

Έχουν προταθεί διάφοροι προσεγγιστικοί συντελεστές διόρθωσης μεροληψίας, χωρίς όμως κανένας απ' αυτούς να οδηγεί σε αυστηρώς αμε-

---

$$\hat{C}_{k_X} = \hat{M}_X^{(4)} / \left(\hat{M}_X^{(2)}\right)^2$$

(Η εξίσωση αυτή προκύπτει από τις εξισώσεις (12.29), (12.16), (12.13) και (3.43) του Kendall and Stewart (1968)). Ας σημειωθεί ότι, σε πολλά βιβλία στατιστικής υδρολογίας, εσφαλμένα παραλείπεται ο όρος που βρίσκεται μέσα στην αγκύλη, πράγμα που οδηγεί σε υπερεκτίμηση της τέταρτης ροπής του πληθυσμού.

ρόληπτη εκτιμήτρια του συντελεστή ασυμμετρίας. Παραθέτουμε τους σημαντικότερους απ' αυτούς:

1. $$\phi\left(\hat{C}_{s_X}, n\right) = \frac{\sqrt{n(n-1)}}{n-2} \tag{3.27}$$

Αυτός προκύπτει αν στην (3.26) αντικατασταθούν οι μεροληπτικές εκτιμήτριες των ροπών από τις αντίστοιχες αμερόληπτες.

2. $$\phi\left(\hat{C}_{s_X}, n\right) = \frac{n^2}{(n-1)(n-2)} \tag{3.28}$$

Αυτός προκύπτει αν στην (3.26) αντικατασταθεί η μεροληπτική εκτιμήτρια της τρίτης κεντρικής ροπής από την αντίστοιχη αμερόληπτη (Yevjevich, 1978, σ. 110).

3. $$\phi\left(\hat{C}_{s_X}, n\right) = \frac{\sqrt{n(n-1)}}{n-2}\left(1 + \frac{8.5}{n}\right) \tag{3.29}$$

Αυτός έχει προταθεί από τον Hazen (1924).

4. $$\phi\left(\hat{C}_{s_X}, n\right) = 1 + \left(\frac{6.51}{n} + \frac{20.20}{n^2}\right) + \left(\frac{1.48}{n} + \frac{6.77}{n^2}\right)\hat{C}_{s_X}^2 \tag{3.30}$$

Αυτός έχει προταθεί από τους Bobıe and Robitaille (1975), οι οποίοι στηρίχτηκαν σε αποτελέσματα των Wallis et al. (1974).

3.2.5 Συνδιασπορά και συσχέτιση

Η μεροληπτική εκτιμήτρια της συνδιασποράς σ_{XY} ενός πληθυσμού δύο μεταβλητών X και Y είναι η ακόλουθη:

$$S_{XY} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n} \tag{3.31}$$

για την οποία αποδεικνύεται (Papoulis, 1990, σ. 295) ότι

$$E\left[S_{XY}\right]=\frac{n-1}{n}\sigma_{XY} \tag{3.32}$$

και επομένως ο συντελεστής διόρθωσης μεροληψίας είναι

$$\phi\left(S_{XY},n\right)=\frac{n}{n-1} \tag{3.33}$$

Έτσι, η αμερόληπτη (και συνεπής) εκτιμήτρια της σ_X^2 είναι η ακόλουθη, γνωστή ως *δειγματική συνδιασπορά*:*

$$S_{XY}^{*}=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{n-1} \tag{3.34}$$

Για μεγάλο μέγεθος δείγματος n οι δύο εκτιμήτριες πρακτικώς ταυτίζονται.

Η εκτιμήτρια του συντελεστή συσχέτισης ρ_{XY} είναι η ακόλουθη, γνωστή ως *δειγματικός συντελεστής συσχέτισης*:

$$R_{XY}=\frac{S_{XY}}{S_X S_Y}=\frac{S_{XY}^{*}}{S_X^{*}S_Y^{*}}=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\ \sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}} \tag{3.35}$$

Η ακριβής συνάρτηση κατανομής της και οι ροπές της είναι δύσκολο να προσδιοριστούν αναλυτικά, αλλά πάντως θεωρείται κατά προσέγγιση αμερόληπτη εκτιμήτρια.

* Σε πολλά βιβλία στον παρονομαστή της (3.34) υπάρχει ο όρος $n-2$, ο οποίος όμως δεν είναι σωστός.

3.3 Τυπικά όρια εμπιστοσύνης

3.3.1 Μέση τιμή - γνωστή διασπορά πληθυσμού

Έστω X τυχαία μεταβλητή που αντιπροσωπεύει κάποιο πληθυσμό, με μέση τιμή μ_X και τυπική απόκλιση σ_X. Σύμφωνα με το κεντρικό οριακό θεώρημα και την εξίσωση (3.11), η δειγματική μέση τιμή $\overline{X}$ (ως μέσος όρος n τυχαίων μεταβλητών) ακολουθεί κανονική κατανομή $N\left(\mu_X, \sigma_X / \sqrt{n}\right)$, υπό την προϋπόθεση ότι το n είναι αρκετά μεγάλο. Επίσης ακολουθεί την ίδια κανονική κατανομή, ανεξαρτήτως του μεγέθους n, αν η X είναι κανονική.

Το πρόβλημα που μας απασχολεί εδώ είναι ο προσδιορισμός των ορίων εμπιστοσύνης της μ_X για συντελεστή εμπιστοσύνης γ. Συμβολίζουμε με $z_{(1+\gamma)/2}$ το $((1+\gamma)/2)$-ποσοστημόριο της τυποποιημένης κανονικής κατανομής $N(0, 1)$ (δηλαδή το σημείο z που αντιστοιχεί σε πιθανότητα μη υπέρβασης $(1+\gamma)/2$). Μάλιστα, λόγω συμμετρίας ισχύει $z_{(1-\gamma)/2} = -z_{(1+\gamma)/2}$ (βλ. Σχ. 3.1). Θα έχουμε

$$P\left(\mu_X - \frac{z_{(1+\gamma)/2}\,\sigma_X}{\sqrt{n}} < \overline{X} < \mu_X + \frac{z_{(1+\gamma)/2}\,\sigma_X}{\sqrt{n}}\right) = \gamma \qquad (3.36)$$

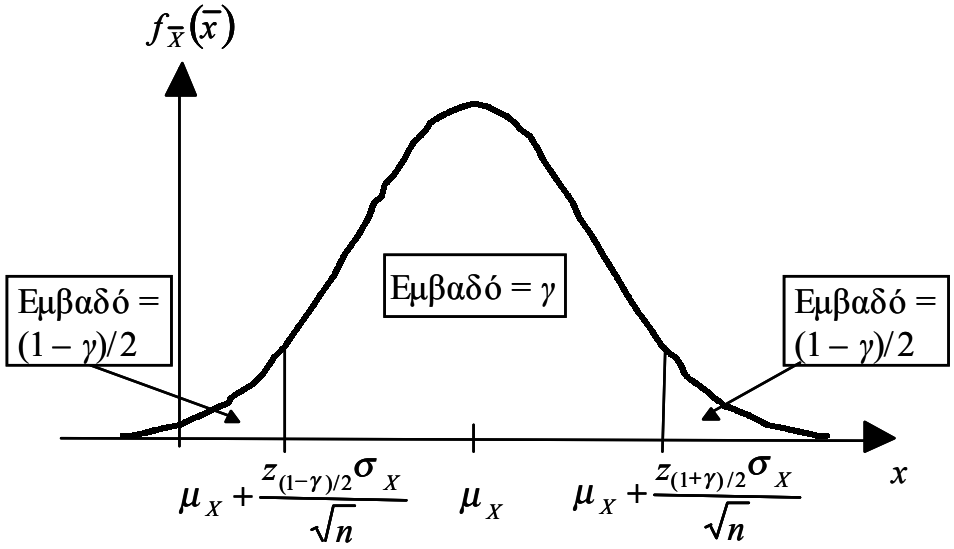
ή ισοδύναμα

$$P\left(\overline{X} - \frac{z_{(1+\gamma)/2}\,\sigma_X}{\sqrt{n}} < \mu_X < \overline{X} + \frac{z_{(1+\gamma)/2}\,\sigma_X}{\sqrt{n}}\right) = \gamma \qquad (3.37)$$

Η (3.37) δίνει τα ζητούμενα όρια εμπιστοσύνης. Για τον αριθμητικό τους υπολογισμό αντικαθιστούμε στην (3.37) την εκτιμήτρια $\overline{X}$ με την αντίστοιχη εκτίμηση $\bar{x}$.

Στον Πίν. 3.1 δίνονται για τους πιο συνηθισμένους συντελεστές εμπιστοσύνης οι αντίστοιχες τιμές των κανονικών ποσοστημορίων $z_{(1+\gamma)/2}$ (οι ίδιες τιμές προκύπτουν και από τον πίνακα της κανονικής κατανομής του Παρατήματος). Παρατηρούμε ότι όσο πλησιέστερος προς το 1 είναι ο συντελεστής εμπιστοσύνης, δηλαδή όσο αυξάνει η αξιοπιστία της εκτίμησης, τόσο μεγαλύτερο είναι διάστημα εμπιστοσύνης, δηλαδή τόσο ασαφέστερη είναι η εκτίμηση. Αντίθετα, αν μικρύνουμε το συντελεστή

εμπιστοσύνης, παίρνουμε πιο "συγκεντρωμένη εκτίμηση", δηλαδή μικρό διάστημα εμπιστοσύνης, αλλά τότε μειώνεται η αξιοπιστία της εκτίμησης, ή, ισοδύναμα, αυξάνεται ο βαθμός αβεβαιότητας.



Σχ. 3.1 Επεξηγηματικό σκαρίφημα για τα όρια εμπιστοσύνης της μέσης τιμής.

Από την εξίσωση (3.37) διαπιστώνουμε ότι υπάρχει τρόπος να αυξηθεί η αξιοπιστία χωρίς να διευρυνθεί το διάστημα εμπιστοσύνης, ή να μειωθεί το διάστημα εμπιστοσύνης χωρίς να μειωθεί η αξιοπιστία της εκτίμησης. Ο τρόπος αυτός είναι να αυξηθεί το μέγεθος του δείγματος n με εκτέλεση και άλλων μετρήσεων (βλ. και την Εφαρμογή που ακολουθεί).

Πίν. 3.1 Τυπικές τιμές κανονικών ποσοστημορίων $z_{(1+\gamma)/2}$ για τον υπολογισμό ορίων εμπιστοσύνης.

γ	0.90	0.95	0.99	0.999
$(1+\gamma)/2$	0.95	0.975	0.995	0.9995
$z_{(1+\gamma)/2}$	1.645	1.960	2.576	3.291

Η προϋπόθεση της γνωστής διασποράς του πληθυσμού, στην οποία στηρίχτηκε η προηγούμενη ανάλυση, στην τεχνική υδρολογία δεν είναι γενικά ρεαλιστική. Αυτό επειδή η πληροφορία που έχουμε για μια υδρολογική μεταβλητή προέρχεται πάντα από ένα δείγμα. Ωστόσο τα αποτελέσματα έχουν πρακτικό ενδιαφέρον, δεδομένου ότι η (3.37) μπορεί να χρησιμοποιηθεί ως προσέγγιση αν το n είναι αρκετά μεγάλο (> 30), με αντικατάσταση της διασποράς του πληθυσμού από την εκτίμησή της από το δείγμα.

3.3.2 Μέση τιμή - άγνωστη διασπορά πληθυσμού

Η ανάλυση που παρουσιάζουμε εδώ μπορεί να χρησιμοποιηθεί για άγνωστη διασπορά πληθυσμού και για οποιοδήποτε μέγεθος δείγματος. Ωστόσο, και η ανάλυση αυτή έχει μια περιοριστική προϋπόθεση: ότι η τυχαία μεταβλητή X που αντιπροσωπεύει τον πληθυσμό είναι κανονική $N(\mu_X, \sigma_X)$. Σε αυτή την περίπτωση προκύπτουν τα ακόλουθα συμπεράσματα:

1. Η δειγματική μέση τιμή ακολουθεί κανονική κατανομή $N\left(\mu_X, \sigma_X / \sqrt{n}\right)$. Το συμπέρασμα αυτό είναι συνέπεια της ιδιότητας της κανονικής κατανομής να διατηρείται με την άθροιση (ιδιότητα συμβατή με το κεντρικό οριακό θεώρημα).
2. Η συνάρτηση της δειγματικής διασποράς $(n-1)S_X^{*2} / \sigma_X^2$ ακολουθεί κατανομή $\chi^2(n-1)$. Το συμπέρασμα αυτό προκύπτει από το θεώρημα του εδαφίου 2.8.3, σύμφωνα με το οποίο το άθροισμα των τετραγώνων ενός αριθμού τυποποιημένων κανονικών μεταβλητών ακολουθεί κατανομή χ^2.
3. Οι τυχαίες μεταβλητές $\overline{X}$ και S_X^{*2} είναι ανεξάρτητες. Αυτό το συμπέρασμα είναι απόρροια θεωρήματος της στατιστικής (βλ. π.χ. Papoulis, 1990, σ. 222).
4. Ο λόγος $(\overline{X} - \mu_X) / (S_X^* / \sqrt{n})$ ακολουθεί κατανομή Student $t(n-1)$. Το συμπέρασμα αυτό βασίζεται στο θεώρημα του εδαφίου 2.8.4.

Συμβολίζουμε με $t_{(1+\gamma)/2}$ το $[(1+\gamma)/2]$-ποσοστημόριο της κατανομής Student $t(n-1)$ (δηλαδή το σημείο t που αντιστοιχεί σε πιθανότητα μη υπέρβασης $(1+\gamma)/2$, για $n-1$ βαθμούς ελευθερίας). Λόγω της συμμετρίας ισχύει $t_{(1-\gamma)/2} = -t_{(1+\gamma)/2}$. Θα έχουμε

$$P\left(-t_{(1+\gamma)/2} < \frac{\overline{X} - \mu_X}{S_X^* / \sqrt{n}} < t_{(1+\gamma)/2}\right) = \gamma \tag{3.38}$$

ή ισοδύναμα

$$P\left(\overline{X} - \frac{t_{(1+\gamma)/2}\ S_X^*}{\sqrt{n}} < \mu_X < \overline{X} + \frac{t_{(1+\gamma)/2}\ S_X^*}{\sqrt{n}} \right) = \gamma \qquad (3.39)$$

Η (3.39) δίνει τα ζητούμενα όρια εμπιστοσύνης. Για τον αριθμητικό τους υπολογισμό αντικαθιστούμε στην (3.39) τις εκτιμήτριες $\overline{X}$ και S_X^* με τις αντίστοιχες εκτιμήσεις $\bar{x}$ και s_X^*.

Παρόλο που οι (3.37) και (3.39) διαφέρουν σημαντικά ως προς τη θεωρητική τους υπόσταση και τις παραδοχές βάσει των οποίων προκύπτουν, υπολογιστικά είναι όμοιες. Μάλιστα, για μεγάλο n (>30) πρακτικώς ταυτίζονται δεδομένου ότι ισχύει $t_{(1+\gamma)/2} \approx z_{(1+\gamma)/2}$ (ακριβέστερα $t_{(1+\gamma)/2} \approx z_{(1+\gamma)/2}\sqrt{(n-1)/(n-3)}$, για $n-1$ βαθμούς ελευθερίας).

Οι παραπάνω αναλύσεις αφήνουν ένα κενό, ήτοι την περίπτωση μικρού δείγματος, άγνωστης διασποράς του πληθυσμού και μη κανονικής κατανομής του πληθυσμού. Αυτή η περίπτωση δεν καλύπτεται με γενικό και αυστηρό τρόπο από τη στατιστική.* Προσεγγιστικά, πάντως, μπορούμε να χρησιμοποιήσουμε την πιο πάνω μεθοδολογία και σε αυτές τις περιπτώσεις, αρκεί η κατανομή του πληθυσμού να είναι κωδωνοειδής, δηλαδή όχι ιδιαίτερα ασύμμετρη. Άλλωστε ελάχιστες είναι οι περιπτώσεις που, όπως στις αναλύσεις που προηγήθηκαν, υπολογίζονται θεωρητικώς ακριβή όρια εμπιστοσύνης. Στα περισσότερα από τα προβλήματα. που θα ακολουθήσουν θα αρκεστούμε σε προσεγγιστικά όρια εμπιστοσύνης.

Εφαρμογή 3.3.2

Από δείγμα ετήσιων παροχών ποταμού (χρονικά μέσων) μήκους 15 ετών βρέθηκε η δειγματική μέση τιμή ίση με 10.05 m^3/s και η δειγματική τυπική απόκλιση 2.80 m^3/s. Ζητούνται (1) τα όρια εμπιστοσύνης 95% της μέσης υπερετήσιας παροχής και (2) το μέγεθος του δείγματος που, με βαθμό εμπιστοσύνης 95%, εξασφαλίζει ακρίβεια 10% στην εκτίμηση της μέσης υπερετήσιας παροχής.

* Υπάρχουν βέβαια ειδικές περιπτώσεις που υπολογίζονται θεωρητικώς ακριβή όρια εμπιστοσύνης. Μια τέτοια περίπτωση αναφέρεται στις συναρτήσεις κατανομής μιας παραμέτρου (π.χ. εκθετική). Το θέμα αυτό καλύπτεται από βιβλία μαθηματικής στατιστικής, π.χ., Papoulis (1990, σ. 280).

(1) Θεωρούμε ότι η μέση ετήσια παροχή ακολουθεί κανονική κατανομή (βλ. και ενότητα 2.8 καθώς και κεφάλαιο 5) και χρησιμοποιούμε την εξίσωση (3.39). Από τον πίνακα της κατανομής Student (Παράρτημα Π3) βρίσκουμε ότι για $n - 1 = 14$ βαθμούς ελευθερίας $t_{(1+\gamma)/2} = t_{0.975} = 2.14$. Κατά συνέπεια τα όρια εμπιστοσύνης 95% είναι*

$$10.05 - 2.14 \times 2.80/\sqrt{15} < \mu_X < 10.05 + 2.14 \times 2.80/\sqrt{15}$$

ήτοι

$$8.50 < \mu_X < 11.60$$

Για λόγους σύγκρισης θα υπολογίσουμε τα όρια εμπιστοσύνης και με τη (3.37), αν και δεν είναι σωστό. Από τον Πίν. 3.1 βρίσκουμε $z_{(1+\gamma)/2} = z_{0.975} = 1.96$. Κατά συνέπεια τα όρια εμπιστοσύνης 95% είναι

$$10.05 - 1.96 \times 2.80/\sqrt{15} < \mu_X < 10.05 + 1.96 \times 2.80/\sqrt{15}$$

ήτοι

$$8.63 < \mu_X < 11.47$$

Το διάστημα εμπιστοσύνης σε αυτή την περίπτωση είναι ελαφρώς στενότερο.

(2) Υποθέτουμε ότι $n \geq 30$ οπότε μπορούμε να χρησιμοποιήσουμε την (3.37). Θα πρέπει

$$1.96 \times 2.8 / \sqrt{n} = 10\% \times 10.05,$$

άρα $n = 30$. Διαπιστώνουμε ότι καλύπτεται οριακά η προϋπόθεση που τέθηκε ($n \geq 30$). (Αν δεν καλυπτόταν θα συνεχίζαμε με δοκιμές, χρησιμοποιώντας πλέον την (3.39)).

---

* Εφαρμόζοντας τη (3.39), δηλαδή αντικαθιστώντας τις εκτιμήτριες με τις εκτιμήσεις, δεν θα ήταν μαθηματικώς ορθό να γράψουμε

$$\mathrm{P}(10.05 - 2.62 \times 2.80/\sqrt{15} < \mu_X < 10.05 - 2.62 \times 2.80/\sqrt{15}) = 0.95$$

γιατί η μ_X είναι (άγνωστη) παράμετρος και όχι τυχαία μεταβλητή, άρα δεν ορίζεται συνάρτηση πιθανότητας για αυτή. Επίσης, δεν είναι σωστό να πούμε π.χ. ότι "με πιθανότητα 95% η μέση τιμή βρίσκεται στο διάστημα (8.16, 11.94)". Το σωστό είναι να χρησιμοποιήσουμε τον όρο "με εμπιστοσύνη 95%".

3.3.3 Διασπορά και τυπική απόκλιση

Όπως και στο προηγούμενο εδάφιο, θα υποθέσουμε και εδώ ότι η μεταβλητή X ακολουθεί κανονική κατανομή $N(\mu_X, \sigma_X)$. Όπως ήδη αναφέρθηκε, σε αυτή την περίπτωση η συνάρτηση της δειγματικής διασποράς $(n-1)S_X^{*2} / \sigma_X^2$ ακολουθεί κατανομή $\chi^2(n-1)$.

Συμβολίζουμε με $\chi^2_{(1+\gamma)/2}$ το $[(1+\gamma)/2]$-ποσοστημόριο της κατανομής $\chi^2(n-1)$ και με $\chi^2_{(1-\gamma)/2}$ το $[(1-\gamma)/2]$-ποσοστημόριο της ίδιας κατανομής (εδώ τα δύο αυτά μεγέθη δεν συνδέονται με σχέση συμμετρίας, γιατί η κατανομή χ^2 δεν είναι συμμετρική). Θα έχουμε

$$P\left(\chi^2_{(1-\gamma)/2} < \frac{(n-1)S_X^{*2}}{\sigma_X^2} < \chi^2_{(1+\gamma)/2} \right) = \gamma \tag{3.40}$$

ή ισοδύναμα

$$P\left(\frac{(n-1)S_X^{*2}}{\chi^2_{(1+\gamma)/2}} < \sigma_X^2 < \frac{(n-1)S_X^{*2}}{\chi^2_{(1-\gamma)/2}} \right) = \gamma \tag{3.41}$$

Η (3.41) δίνει τα ζητούμενα όρια εμπιστοσύνης. Εύκολα προκύπτει ότι τα όρια εμπιστοσύνης της τυπικής απόκλισης δίνονται από την εξίσωση

$$P\left(\frac{\sqrt{n-1}\, S_X^*}{\sqrt{\chi^2_{(1+\gamma)/2}}} < \sigma_X < \frac{\sqrt{n-1}\, S_X^*}{\sqrt{\chi^2_{(1-\gamma)/2}}} \right) = \gamma \tag{3.42}$$

Εφαρμογή 3.3.3

Να υπολογιστούν τα όρια εμπιστοσύνης 95% της τυπικής απόκλισης της μέσης ετήσιας παροχής στο πρόβλημα της Εφαρμογής 3.3.2.

Υπενθυμίζουμε ότι η δειγματική τυπική απόκλιση στο παραπάνω πρόβλημα ήταν 2.8 m^3/s. Θεωρούμε και πάλι ότι η μέση ετήσια παροχή ακολουθεί κανονική κατανομή και χρησιμοποιούμε την εξίσωση (3.42). Από τον πίνακα της κατανομής χ^2 (Παράρτημα Π2) βρίσκουμε ότι για $n-1 = 14$ βαθμούς ελευθερίας $\chi^2_{(1+\gamma)/2} = \chi^2_{0.975} = 26.12$ και $\chi^2_{(1-\gamma)/2} = \chi^2_{0.025} = 5.63$. Κατά συνέπεια τα όρια εμπιστοσύνης 95% είναι

$$\frac{\sqrt{14}*2.80}{\sqrt{26.12}} < \sigma_X < \frac{\sqrt{14}*2.80}{\sqrt{5.63}}$$

ήτοι

$$2.05 < \sigma_X < 4.41$$

3.3.4 Ποσοστημόριο κανονικής κατανομής – Τυπικό σφάλμα

Στην τεχνική υδρολογία το συχνότερο πρόβλημα ορίων εμπιστοσύνης που αντιμετωπίζουμε αφορά σε μεγέθη υδρολογικού σχεδιασμού, π.χ. στην παροχή σχεδιασμού ενός έργου. Έστω η υδρολογική μεταβλητή X, η οποία ακολουθεί δεδομένη κατανομή $F_X(x)$. Εδώ θα θεωρήσουμε ότι η $F_X(x)$ είναι κανονική $N(\mu_X, \sigma_X)$, της οποίας η αντιμετώπιση είναι σχετικά απλή, ενώ στο κεφάλαιο 6 θα δώσουμε αναλυτικούς τύπους για όλες τις κατανομές που χρησιμοποιούνται στην τεχνική υδρολογία. Για δεδομένη πιθανότητα μη υπέρβασης $u = F_X(x)$, η αντίστοιχη τιμή της μεταβλητής X (συμβολικά $x_u = u$-ποσοστημόριο) θα είναι

$$x_u = \mu_X + z_u \, \sigma_X \tag{3.43}$$

όπου z_u το u-ποσοστημόριο της τυποποιημένης κανονικής κατανομής $N(0, 1)$. Ωστόσο, στην παραπάνω εξίσωση οι παράμετροι πληθυσμού μ_X και σ_X στην πράξη είναι άγνωστες. Χρησιμοποιώντας τις σημειακές εκτιμήσεις τους, παίρνουμε την αντίστοιχη εκτίμηση $\hat{x}_u = \overline{x} + z_u \, s_X$, η οποία μπορεί να θεωρηθεί ως τιμή της τυχαίας μεταβλητής

$$\hat{X}_u = \overline{X} + z_u \, S_X \tag{3.44}$$

Η εξίσωση αυτή μπορεί να χρησιμοποιηθεί για τον προσδιορισμό ορίων εμπιστοσύνης της x_u. Ωστόσο, ο ακριβής προσδιορισμός είναι πρακτικά αδύνατος, λόγω της πολυπλοκότητας της συνάρτησης κατανομής της $\hat{X}_u$. Θα αρκεστούμε λοιπόν (όπως και σε επόμενες ακόμη πιο πολύπλοκες περιπτώσεις) σε προσεγγιστικά όρια εμπιστοσύνης, βασισμένα στην παραδοχή ότι η $\hat{X}_u$ ακολουθεί κανονική κατανομή.

Η μέση τιμή της $\hat{X}_u$ δίνεται από την εξίσωση (2.43), η οποία αν συνδυαστεί και με τις (3.11) και (3.17) γράφεται

$$E\left[\hat{X}_u\right] = E\left[\overline{X}\right] + z_u \, E\left[S_X\right] \approx \mu_X + z_u \, \sigma_X = x_u \tag{3.45}$$

όπου θεωρήσαμε ότι το n είναι αρκετά μεγάλο και παραλείψαμε τον όρο $O(1/n)$ στην $E[S_X]$.* Αντίστοιχα, η διασπορά της $\hat{X}_u$ δίνεται από την εξίσωση (2.45), η οποία γράφεται

$$\mathrm{Var}\left[\hat{X}_u\right] = \mathrm{Var}\left[\overline{X}\right] + z_u^2\ \mathrm{Var}\left[S_X\right] + 2z_u\ \mathrm{Cov}\left[\overline{X}, S_X\right] \tag{3.46}$$

Δεδομένου ότι η X ακολουθεί κανονική κατανομή, ο τρίτος όρος της (3.46) είναι μηδέν (όπως αναφέρθηκε πιο πάνω, οι μεταβλητές $\overline{X}$ και S_X είναι ανεξάρτητες). Συνδυάζοντας και τις (3.11) και (3.18), η (3.46) γράφεται

$$\varepsilon_u^2 := \mathrm{Var}\left[\hat{X}_u\right] \approx \frac{\sigma_X^2}{n} + z_u^2\ \frac{\sigma_X^2}{2n} = \frac{\sigma_X^2}{n}\left(1 + \frac{z_u^2}{2}\right) \tag{3.47}$$

Η ποσότητα ε_u είναι γνωστή στη βιβλιογραφία με τον όρο *τυπικό σφάλμα ποσοστημορίου* ή απλώς τυπικό σφάλμα.

Χρησιμοποιώντας την παραδοχή ότι η $\hat{X}_u$ ακολουθεί κανονική κατανομή $N(x_u, \varepsilon_u)$ μπορούμε να γράψουμε

$$P\left(-z_{(1+\gamma)/2} < \frac{\hat{X}_u - x_u}{\varepsilon_u} < z_{(1+\gamma)/2}\right) = \gamma \tag{3.48}$$

όπου γ ο συντελεστής εμπιστοσύνης. Ισοδύναμα

$$P\left(\hat{X}_u - z_{(1+\gamma)/2}\ \varepsilon_u < x_u < \hat{X}_u + z_{(1+\gamma)/2}\ \varepsilon_u\right) = \gamma \tag{3.49}$$

Αν αντικαταστήσουμε στην παραπάνω το ε_u από την (3.47), και στη συνέχεια αντικαταστήσουμε την τυπική απόκλιση σ_X με την αντίστοιχη εκτιμήτρια, παίρνουμε την ακόλουθη τελική σχέση

$$P\left(\hat{X}_u - z_{(1+\gamma)/2}\ \sqrt{1 + \frac{z_u^2}{2}}\ \frac{S_X}{\sqrt{n}} < x_u < \hat{X}_u + z_{(1+\gamma)/2}\ \sqrt{1 + \frac{z_u^2}{2}}\ \frac{S_X}{\sqrt{n}}\right) = \gamma \tag{3.50}$$

---

* Λόγω του προσεγγιστικού χαρακτήρα της ανάλυσης, δεν κάνουμε διάκριση των εκτιμητριών S_X και S_X^*, θεωρώντας ότι το n είναι αρκετά μεγάλο, ώστε πρακτικά να ταυτίζονται.

Η τελευταία εξίσωση είναι προσεγγιστική, με βαθμό ακρίβειας που αυξάνεται με την αύξηση του n. Επίσης, ισχύει μόνο όταν η κατανομή της X είναι κανονική. Ωστόσο, η (3.49) χρησιμοποιείται και για άλλες κατανομές της X, με τη διαφορά ότι το τυπικό σφάλμα έχει διαφορετική έκφραση και διαφορετικό τρόπο υπολογισμού. Ο ενδιαφερόμενος αναγνώστης για μια γενική έκφραση του τυπικού σφάλματος παραπέμπεται στον Kite (1988, σ. 33-38).

Οι σημειακές εκτιμήσεις των ορίων εμπιστοσύνης

$$\hat{x}_{u1} = \hat{x}_u - z_{(1+\gamma)/2}\sqrt{1+\frac{z_u^2}{2}}\,\frac{s_X}{\sqrt{n}} \qquad \hat{x}_{u2} = \hat{x}_u + z_{(1+\gamma)/2}\sqrt{1+\frac{z_u^2}{2}}\,\frac{s_X}{\sqrt{n}} \tag{3.51}$$

είναι προφανώς συναρτήσεις του u ή, ισοδύναμα, της πιθανότητας υπέρβασης, $1 - u$. Τα γραφήματα των συναρτήσεων αυτών βρίσκονται εκατέρωθεν της καμπύλης x_u και λέγονται *καμπύλες εμπιστοσύνης* (ή *καμπύλες ασφάλειας* της x_u).

Εφαρμογή 3.3.4

Να υπολογιστούν τα όρια εμπιστοσύνης 95% της παροχής που έχει πιθανότητα υπέρβασης (α) 1% και (β) 99% στο πρόβλημα της Εφαρμογής 3.3.2.

Κατ' αρχήν, επειδή το μέγεθος του δείγματος είναι πολύ μικρό, δεν περιμένουμε πολύ καλό βαθμό ακρίβειας στους υπολογισμούς μας.

Θα υπολογίσουμε πρώτα τις σημειακές εκτιμήσεις. Για την παροχή πιθανότητας υπέρβασης $F_1 = 0.01$ έχουμε $u = 1 - F_1 = 0.99$ και $z_u = 2.326$ (από τον πίνακα της κανονικής κατανομής του Παραρτήματος Π1). Άρα η σημειακή εκτίμηση είναι $\hat{x}_u = 10.05 + 2.326 \times 2.80 = 16.56$. Αντίστοιχα, για την παροχή πιθανότητας υπέρβασης $F_1 = 0.99$ έχουμε $u = 1 - F_1 = 0.01$ και $z_u = -2.326$, άρα $\hat{x}_u = 10.05 - 2.326 \times 2.80 = 3.54$.

Προχωρούμε τώρα στον υπολογισμό των ορίων εμπιστοσύνης. Για $\gamma = 95\%$ και $z_{(1+\gamma)/2} = 1.96$, τα όρια για την παροχή πιθανότητας υπέρβασης 1% θα είναι:

$$\hat{x}_{u1} = 16.56 - 1.96\sqrt{1+\frac{2.326^2}{2}}\,\frac{2.80}{\sqrt{15}} = 13.83$$

$$\hat{x}_{u2} = 16.56 + 1.96\sqrt{1+\frac{2.326^2}{2}}\,\frac{2.80}{\sqrt{15}} = 19.29$$

Αντίστοιχα, τα όρια για την παροχή πιθανότητας υπέρβασης 99% θα είναι:

$$\hat{x}_{u1} = 3.54 - 1.96\sqrt{1+\frac{2.326^2}{2}}\,\frac{2.80}{\sqrt{15}} = 0.81$$

$$\hat{x}_{u2} = 3.54 + 1.96\sqrt{1+\frac{2.326^2}{2}}\,\frac{2.80}{\sqrt{15}} = 6.27$$

3.3.5 Συντελεστής συσχέτισης

Για τον υπολογισμό των ορίων εμπιστοσύνης του συντελεστή συσχέτισης ρ ενός πληθυσμού δύο μεταβλητών X και Y, χρησιμοποιούμε τη βοηθητική μεταβλητή Z που ορίζεται από το λεγόμενο *μετασχηματισμό Fisher*:

$$Z = \frac{1}{2}\ln\frac{1+R}{1-R} \Leftrightarrow R = \frac{e^{2Z}-1}{e^{2Z}+1} = \tanh Z \qquad (3.52)$$

όπου R ο δειγματικός συντελεστής συσχέτισης. Παρατηρούμε ότι για $-1 < R < 1$ το διάστημα μεταβολής της Z είναι $-\infty < Z < \infty$, ενώ για $R = 0$ έχουμε $Z = 0$. Αποδεικνύεται ότι αν οι X και Y ακολουθούν κανονικές κατανομές, τότε η Z ακολουθεί κατά προσέγγιση κανονική κατανομή $N(\mu_Z, \sigma_Z)$ όπου

$$\mu_Z = E[Z] \approx \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \qquad \sigma_Z^2 = \mathrm{Var}[Z] \approx \frac{1}{n-3} \qquad (3.53)^*$$

Κατά συνέπεια, αν $\zeta_{(1+\gamma)/2}$ είναι το $(1+\gamma)/2$-ποσοστημόριο της τυποποιημένης κανονικής κατανομής,† θα έχουμε

---

* Ακριβέστερες προσεγγίσεις δίνονται από τις ακόλουθες εξισώσεις

$$\mu_Z = E[Z] \approx \frac{1}{2}\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} \qquad \sigma_Z^2 = \mathrm{Var}[Z] \approx \frac{1}{n-1} + \frac{4-\rho^2}{2(n-1)^2}$$

πράγμα που σημαίνει ότι οι προσεγγίσεις της (3.53) ισχύουν για μικρές τιμές του ρ ή μεγάλες τιμές του n (βλ. Kendall and Stuart, 1963, σ. 391).

† Εδώ χρησιμοποιήσαμε το ελληνικό σύμβολο $\zeta_{(1+\gamma)/2}$ αντί του συνήθους λατινικού $z_{(1+\gamma)/2}$, προκειμένου να αποφευχθεί η σύγχυση με τη μεταβλητή του Fisher Z η οποία επίσης ακολουθεί κανονική κατανομή, αλλά όχι τυποποιημένη.

$$P\left(\mu_Z - \frac{\zeta_{(1+\gamma)/2}}{\sqrt{n-3}} < Z < \mu_Z + \frac{\zeta_{(1+\gamma)/2}}{\sqrt{n-3}}\right) = \gamma \tag{3.54}$$

ή ισοδύναμα

$$P\left(Z - \frac{\zeta_{(1+\gamma)/2}}{\sqrt{n-3}} < \mu_Z < Z + \frac{\zeta_{(1+\gamma)/2}}{\sqrt{n-3}}\right) = \gamma \tag{3.55}$$

Αντικαθιστώντας στην (3.55) το μ_Z από την (3.53) και επιλύοντας ως προς ρ, αφού λάβουμε υπόψη και τη μονοτονικότητα του μετασχηματισμού (3.52), παίρνουμε

$$P\left(R_1 < \rho < R_2\right) = \gamma \tag{3.56}$$

όπου

$$\begin{aligned} R_1 &= \frac{e^{2Z_1} - 1}{e^{2Z_1} + 1} \qquad R_2 = \frac{e^{2Z_2} - 1}{e^{2Z_2} + 1} \\ \left.\begin{matrix} Z_2 \\ Z_1 \end{matrix}\right\} &= Z \pm \frac{\zeta_{(1+\gamma)/2}}{\sqrt{n-3}} \qquad Z = \frac{1}{2}\ln\frac{1+R}{1-R} \end{aligned} \tag{3.57}$$

Για τον αριθμητικό υπολογισμό των ορίων εμπιστοσύνης εφαρμόζουμε τις εξισώσεις (3.57) ξεκινώντας από το τέλος προς την αρχή και αντικαθιστώντας τις εκτιμήτριες με τις αντίστοιχες εκτιμήσεις (π.χ. $R = r$, κτλ.).

3.4 Εκτίμηση παραμέτρων συναρτήσεων κατανομής

Έστω τυχαία μεταβλητή X με γνωστό τύπο συνάρτησης κατανομής, ο οποίος περιέχει τις άγνωστες παραμέτρους $\theta_1, \theta_2, \ldots, \theta_r$. Έτσι, η πυκνότητα πιθανότητας της X είναι μια συνάρτηση $f_X(x, \theta_1, \theta_2, \ldots, \theta_r)$. Θα εξετάσουμε το πρόβλημα της εκτίμησης αυτών των παραμέτρων με βάση ένα δείγμα $X_1, X_2, \ldots, X_n$. Συγκεκριμένα θα παραθέσουμε τις δύο πιο κλασικές μεθόδους της στατιστικής για την εκτίμηση των παραμέτρων: τη *μέθοδο των ροπών* και τη *μέθοδο της μέγιστης πιθανοφάνειας*.

Η στατιστική έχει αναπτύξει και άλλες γενικές μεθόδους για την εκτίμηση παραμέτρων κατανομών, όπως π.χ. τη μέθοδο της μέγιστης εντροπίας ή τη μέθοδο των L-ροπών. Τέτοιες μέθοδοι έχουν συχνά χρησιμοποιηθεί και στην τεχνική υδρολογία, αλλά δεν παρατίθενται στο κείμενο αυτό (ο ενδιαφερόμενος αναγνώστης παραπέμπεται στους Singh and Rajagopal (1986) και Stedinger et al. (1993), αντίστοιχα). Επιπλέον, στην τεχνική υδρολογία χρησιμοποιούνται κατά περίπτωση και ειδικότερες μέθοδοι, γραφικές ή υπολογιστικές και κατά κανόνα εμπειρικές ή ημιεμπειρικές. Δόκιμα παραδείγματα τέτοιων μεθόδων θα παραθέσουμε στο κεφάλαιο 6 για συγκεκριμένες κατανομές.

3.4.1 Η μέθοδος των ροπών

Η μέθοδος των ροπών βασίζεται στην εξίσωση των θεωρητικών ροπών της κατανομής της X με τις αντίστοιχες δειγματικές εκτιμήσεις των ροπών. Έτσι αν r είναι ο αριθμός των άγνωστων παραμέτρων της κατανομής, μπορούμε να γράψουμε r εξισώσεις της μορφής

$$m_X^{(k)} = \hat{m}_X^{(k)} \qquad k = 1,2,\ldots,r \tag{3.58}$$

όπου τα $m_X^{(k)}$ είναι οι θεωρητικές ροπές περί την αρχή, οι οποίες είναι συναρτήσεις των άγνωστων παραμέτρων και δίνονται από τη σχέση

$$m_X^{(k)} = \int_{-\infty}^{\infty} x^k f_X\left(x,\theta_1,\ldots,\theta_r\right)dx \tag{3.59}$$

ενώ τα $\hat{m}_X^{(k)}$ είναι αριθμητικές εκτιμήσεις που υπολογίζονται από το δείγμα σύμφωνα με τη σχέση

$$\hat{m}_X^{(k)} = \frac{1}{n}\sum_{i=1}^{n} x_i^k \tag{3.60}$$

Με την κατάστρωση και στη συνέχεια επίλυση των r εξισώσεων υπολογίζονται οι άγνωστες παράμετροι θ_1, θ_2, …, θ_r. Σημειώνουμε ότι κατά κανόνα το σύστημα των εξισώσεων δεν είναι γραμμικό και έτσι μπορεί να μην επιλύεται αναλυτικά αλλά μόνο αριθμητικά.

Ισοδύναμα μπορούμε να χρησιμοποιήσουμε τις κεντρικές ροπές, αντί των ροπών περί την αρχή, για $k > 1$, οπότε το σύστημα των εξισώσεων γίνεται

$$\mu_X = \bar{x} \qquad \mu_X^{(k)} = \hat{\mu}_X^{(k)} \qquad k = 2,\ldots,r \tag{3.61}$$

όπου $\mu_X = m_X^{(1)}$ η θεωρητική μέση τιμή του πληθυσμού, $\bar{x} = \hat{m}_X^{(1)}$ η δειγματική μέση τιμή, $\mu_X^{(k)}$ οι θεωρητικές κεντρικές ροπές που δίνονται από τη σχέση

$$\mu_X^{(k)} = \int_{-\infty}^{\infty} \left(x - \mu_X\right)^k f_X\left(x,\theta_1,\ldots,\theta_r\right)dx \tag{3.62}$$

και $\hat{\mu}_X^{(k)}$ οι αντίστοιχες δειγματικές εκτιμήσεις που υπολογίζονται από τη σχέση

$$\hat{\mu}_X^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^k \tag{3.63}$$

Υπενθυμίζεται ότι οι ροπές περί την αρχή που υπολογίζονται από την (3.60) είναι αμερόληπτες εκτιμήσεις, ενώ οι κεντρικές ροπές της (3.63) είναι μεροληπτικές. Πάντως, συχνά χρησιμοποιούνται οι αμερόληπτες εκτιμήσεις των κεντρικών ροπών αντί των μεροληπτικών. Σημειώνεται ότι η μέθοδος των ροπών γενικά δεν οδηγεί σε αμερόληπτες εκτιμήσεις των παραμέτρων θ_1, θ_2, …, θ_r. (εκτός από ορισμένες ειδικές περιπτώσεις), όποιες και αν είναι οι εκτιμήσεις των ροπών που χρησιμοποιούνται.

Υπολογισμός παραμέτρων της κανονικής κατανομής με τη μέθοδο των ροπών

Ως παράδειγμα εφαρμογής της μεθόδου των ροπών, θα υπολογίσουμε τις παραμέτρους της κανονικής κατανομής. Η συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής είναι:

$$f_X\left(x,\mu,\sigma\right) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

και περιλαμβάνει δύο παραμέτρους, τις μ και σ. Κατά συνέπεια χρειάζονται δύο εξισώσεις. Με βάση την (3.61) οι εξισώσεις αυτές θα είναι

$$\mu_X = \bar{x} \qquad \sigma_X^2 = s_X^2$$

όπου για τη δεύτερη εξίσωση έχουμε χρησιμοποιήσει τα πιο κοινά σύμβολα σ_X^2 και s_X^2 για τη θεωρητική και δειγματική διασπορά (δηλαδή δεύτερη κεντρική ροπή) της X, αντίστοιχα. Γνωρίζουμε όμως για τις θεωρητικές ροπές (βλ. εδάφιο 2.8.2) ότι

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

Αντικαθιστώντας τις τιμές αυτές στην προηγούμενη εξίσωση βρίσκουμε άμεσα τις τελικές εκτιμήσεις

$$\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \sigma = s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Η παραπάνω εκτίμηση της σ είναι μεροληπτική, ενώ της μ είναι αμερόληπτη. Αν χρησιμοποιούσαμε την αμερόληπτη εκτίμηση της διασποράς, στην τελευταία εξίσωση θα είχαμε στον παρονομαστή (n-1) στη θέση του n. Ακόμη και τότε όμως δεν θα παίρναμε αμερόληπτη εκτίμηση της σ για τους λόγους που εξηγήσαμε στην ενότητα 3.2.

Όπως είδαμε σε αυτό το συγκεκριμένο παράδειγμα η εφαρμογή της μεθόδου των ροπών είναι απλούστατη και αυτό επεκτείνεται και στους άλλους τύπους συναρτήσεων κατανομής (χωρίς βέβαια η λύση να είναι πάντα τόσο προφανής όσο στο παραπάνω παράδειγμα).

3.4.2 Η μέθοδος της μέγιστης πιθανοφάνειας

Έστω τυχαία μεταβλητή X με πυκνότητα πιθανότητας $f_X(x, \theta_1, \theta_2, \ldots, \theta_r)$ όπου $\theta_1, \theta_2, \ldots, \theta_r$ παράμετροι, και δείγμα $X_1, X_2, \ldots, X_n$ της μεταβλητής. Λόγω της ανεξαρτησίας των μεταβλητών $X_1, X_2, \ldots, X_n$, η από κοινού συνάρτηση πυκνότητας πιθανότητάς τους είναι

$$f_{X_1,\ldots,X_n}\left(x_1,\ldots,x_n,\theta_1,\ldots\theta_r\right) = \prod_{i=1}^{n} f_X\left(x_i,\theta_1,\ldots\theta_r\right) \qquad (3.64)$$

Για δεδομένες τιμές των τυχαίων μεταβλητών ίσες με τις παρατηρήσεις $x_1, \ldots, x_n$ η $f_{X_1,\ldots,X_n}(\ )$ είναι συνάρτηση των παραμέτρων $\theta_1, \theta_2, \ldots, \theta_r$. και λέγεται *συνάρτηση πιθανοφάνειας* αυτών των παραμέτρων.

Η μέθοδος της μέγιστης πιθανοφάνειας εκτιμά τις παραμέτρους $\theta_1, \theta_2, \ldots, \theta_r$ σε τρόπο ώστε η συνάρτηση πιθανοφάνειας να γίνει μέγιστη. Στην περίπτωση αυτή θα πρέπει

$$\frac{\partial f_{X_1,\ldots,X_n}(x_1,\ldots,x_n,\theta_1,\ldots,\theta_r)}{\partial\theta_k}=0 \qquad k=1,\ldots,r \tag{3.65}$$

Από τις r εξισώσεις της παραπάνω μορφής προκύπτουν οι r άγνωστες παράμετροι. Επειδή όμως ο χειρισμός των εξισώσεων αυτών είναι πολύπλοκος, αντί της μεγιστοποίησης της συνάρτησης πιθανοφάνειας, επιδιώκουμε τη μεγιστοποίηση του λογαρίθμου της

$$\begin{aligned} L(x_1,\ldots,x_n,\theta_1,\ldots\theta_r) &= \ln f_{X_1,\ldots,X_n}(x_1,\ldots,x_n,\theta_1,\ldots\theta_r) \\ &= \sum_{i=1}^{n}\ln f_X(x_i,\theta_1,\ldots\theta_r) \end{aligned} \tag{3.66}$$

Η συνάρτηση $L(\ )$ λέγεται *λογαριθμική συνάρτηση πιθανοφάνειας*. Για να είναι μέγιστη θα πρέπει

$$\frac{\partial L(x_1,\ldots,x_n,\theta_1,\ldots\theta_r)}{\partial\theta_k}=\sum_{i=1}^{n}\frac{1}{f_X(x_i,\theta_1,\ldots\theta_r)}\frac{\partial f_X(x_i,\theta_1,\ldots\theta_r)}{\partial\theta_k}=0 \tag{3.67}$$

για $k = 1, \ldots, r$. Από τη λύση αυτών των r εξισώσεων προκύπτουν οι r άγνωστες παράμετροι.

Υπολογισμός παραμέτρων της κανονικής κατανομής με τη μέθοδο της μέγιστης πιθανοφάνειας

Ως παράδειγμα εφαρμογής της μεθόδου της μέγιστης πιθανοφάνειας, θα υπολογίσουμε τις παραμέτρους της κανονικής κατανομής με συνάρτηση πυκνότητας πιθανότητας

$$f_X(x,\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Με βάση την (3.64) σχηματίζουμε τη συνάρτηση πιθανοφάνειας

$$f(x_1,\ldots,x_n,\mu,\sigma)=\frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Λογαριθμίζοντας παίρνουμε τη λογαριθμική συνάρτηση πιθανοφάνειας:

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

Υπολογίζουμε τις παραγώγους ως προς τις άγνωστες παραμέτρους μ και σ και τις εξισώνουμε με 0, οπότε παίρνουμε

$$\frac{\partial L}{\partial\mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-m)=0 \qquad \frac{\partial L}{\partial\sigma}=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2=0$$

Λύνοντας το σύστημα παίρνουμε τις τελικές εκτιμήσεις των παραμέτρων:

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

Συγκρίνοντας με τα αποτελέσματα του προηγούμενου ένθετου εδαφίου, παρατηρούμε κατ' αρχήν ότι η εφαρμογή της μεθόδου της μέγιστης πιθανοφάνειας είναι πιο πολύπλοκη σε σχέση με τη μέθοδο των ροπών. Επίσης, παρατηρούμε ότι τα αποτελέσματα των δύο μεθόδων είναι τα ίδια. Αυτό όμως δεν είναι κανόνας για όλες τις συναρτήσεις κατανομής. Αντίθετα, στις περισσότερες περιπτώσεις οι δύο μέθοδοι δίνουν διαφορετικά αποτελέσματα.

3.5 Έλεγχος υποθέσεων

Μια στατιστική υπόθεση είναι μια υπόθεση σχετικά με τις τιμές μιας ή περισσοτέρων παραμέτρων ενός στατιστικού μοντέλου, το οποίο περιγράφεται από μια συνάρτηση κατανομής. Ο έλεγχος μιας υπόθεσης είναι μια δισήμαντη στατιστική διαδικασία που οδηγεί είτε στην απόρριψη είτε στη αποδοχή (ακριβέστερα: μη απόρριψη) της υπόθεσης.

Στην ενότητα αυτή δίνουμε πολύ περιληπτικά τη σχετική ορολογία και διαδικασία, ενώ σε επόμενα κεφάλαια θα δούμε ορισμένες εφαρμο-

γές. Πάντως, στην τεχνική υδρολογία υπάρχει πληθώρα περιπτώσεων όπου αξιοποιούμε τη θεωρία του ελέγχου υποθέσεων, οι οποίες δεν καλύπτονται στο εισαγωγικό αυτό κείμενο. Ο αναγνώστης που ενδιαφέρεται για λεπτομερέστερη παρουσίαση της θεωρίας παραπέμπεται στα βιβλία στατιστικής (όπως Papoulis, 1990, σ. 321-387, Freund et al., 1988, σ. 310-542), ενώ για μια περιεκτική επισκόπηση των υδρολογικών εφαρμογών παραπέμπεται στους Hirsch et al. (1993, σ. 17.11-29).

3.5.1 Ορολογία

- *Μηδενική υπόθεση* είναι η υπόθεση που ελέγχεται (συμβολικά H_0). Συνήθως πρόκειται για μια υπόθεση της μορφής $\theta = \theta_0$, όπου θ μια παράμετρος που συνδέεται με την κατανομή δεδομένης μεταβλητής και θ_0 μια συγκεκριμένη αριθμητική τιμή.

- *Εναλλακτική υπόθεση* είναι μια δεύτερη υπόθεση η οποία δεν πρέπει να συναληθεύει με τη μηδενική (συμβολικά H_1). Μπορεί να είναι είτε *απλή*, όπως $\theta = \theta_1$, είτε (συνηθέστερα) *σύνθετη*, όπως $\theta \neq \theta_0$, $\theta > \theta_0$ ή $\theta < \theta_0$.

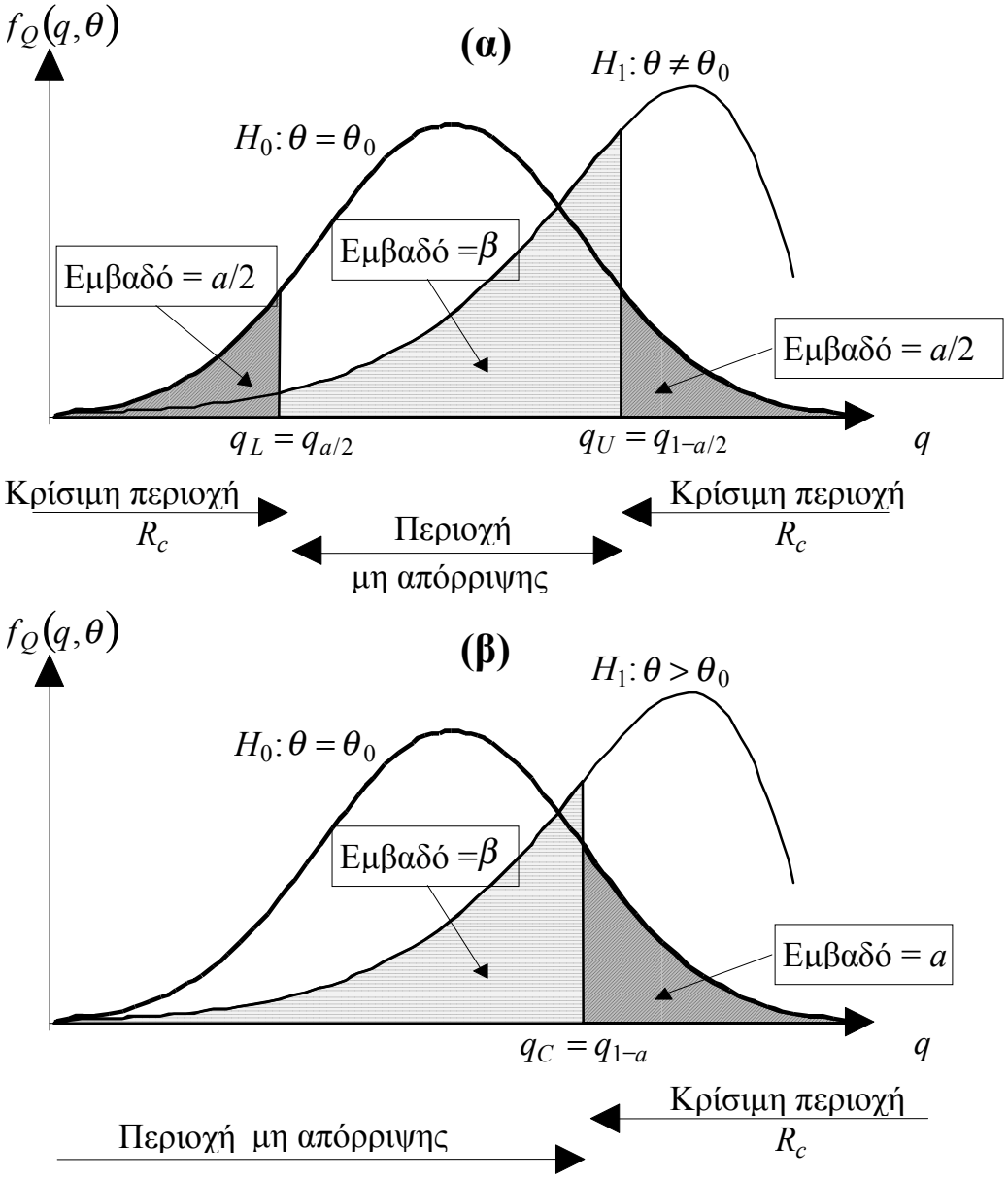
- *Στατιστική συνάρτηση ελέγχου* είναι μια κατάλληλα επιλεγμένη στατιστική συνάρτηση του δείγματος, η οποία χρησιμοποιείται για τον έλεγχο (συμβολικά Q).

- *Κρίσιμη περιοχή* είναι ένα διάστημα πραγματικών αριθμών που, όταν βρίσκεται σε αυτό η τιμή της στατιστικής συνάρτησης ελέγχου, τότε απορρίπτουμε τη μηδενική υπόθεση (συμβολικά R_c, βλ. Σχ. 3.2).

- *Μονόπλευρος* λέγεται ένας έλεγχος όταν η εναλλακτική υπόθεση είναι της μορφής $\theta > \theta_0$ ή $\theta < \theta_0$. Στην περίπτωση αυτή η κρίσιμη περιοχή είναι μια ημιευθεία της μορφής $(q > q_C)$ ή $(q < q_C)$, αντίστοιχα.

- *Αμφίπλευρος* λέγεται ένας έλεγχος όταν η εναλλακτική υπόθεση είναι της μορφής $\theta \neq \theta_0$. Στην περίπτωση αυτή η κρίσιμη περιοχή αποτελείται από τις δύο ημιευθείες $(q < q_L)$ και $(q > q_U)$.

- *Παραμετρικός* λέγεται ένας έλεγχος όταν μπορεί να εφαρμοστεί μόνο για συγκεκριμένο τύπο συνάρτησης κατανομής του πληθυσμού.



Σχ. 3.2 Επεξηγηματικό σκαρίφημα για τις έννοιες που υπεισέρχονται στο στατιστικό έλεγχο: (α) αμφίπλευρος έλεγχος, (β) μονόπλευρος έλεγχος.

- *Μη παραμετρικός* λέγεται ένας έλεγχος όταν εφαρμόζεται για οποιοδήποτε τύπο συνάρτησης κατανομής του πληθυσμού.
- *Κανόνας απόφασης* είναι ο κανόνας βάσει του οποίου απορρίπτουμε ή όχι τη μηδενική υπόθεση. Κατά κανόνα εκφράζεται ως εξής:

απορρίπτουμε την H_0 αν $q \in R_c$

- *Σφάλμα τύπου I* είναι η απόρριψη (με βάση τον κανόνα απόφασης) μιας αληθούς μηδενικής υπόθεσης.
- *Σφάλμα τύπου II* είναι η μη απόρριψη (με βάση τον κανόνα απόφασης) μιας ψευδούς μηδενικής υπόθεσης.
- *Επίπεδο σημαντικότητας* του ελέγχου είναι η πιθανότητα σφάλματος τύπου I, δηλαδή η πιθανότητα απόρριψης αληθούς μηδενικής υπόθεσης. Συμβολικά

$$\alpha = P\left(Q \in R_c | H_0\right) \tag{3.68}$$

- *Ισχύς* του ελέγχου είναι η πιθανότητα απόρριψης ψευδούς μηδενικής υπόθεσης. Συμβολικά

$$P = 1 - \beta = P\left(Q \in R_c | H_1\right) \tag{3.69}$$

όπου β είναι η πιθανότητα σφάλματος τύπου II, δηλαδή*

$$\beta = P\left(Q \notin R_c | H_1\right) \tag{3.70}$$

3.5.2 Διαδικασία ελέγχου

Η διαδικασία ελέγχου περιλαμβάνει τα ακόλουθα βήματα:

1. Σχηματοποίηση της μηδενικής υπόθεσης H_0 και της εναλλακτικής H_1.
2. Επιλογή της στατιστικής συνάρτησης ελέγχου $Q = g(X_1, \ldots, X_n)$ και καθορισμός της συνάρτησης πυκνότητας πιθανότητάς της $f_Q(q, \theta)$.

---

* Διευκρινίζεται ότι τα μεγέθη P και β είναι συναρτήσεις της άγνωστης υπό έλεγχο παραμέτρου θ και όχι αριθμητικές σταθερές. Ο προσδιορισμός τους γίνεται με βάση τις εξισώσεις

$$\beta(\theta) = \int_{R-R_c} f_Q(q,\theta)\,dq \qquad P(\theta) = 1 - \beta(\theta)$$

όπου $f_Q(q, \theta)$ η συνάρτηση πυκνότητας πιθανότητας της Q και $R–R_c$ το διάστημα μη απόρριψης της H_0.

3. Επιλογή του επιπέδου σημαντικότητας α του ελέγχου και καθορισμός της κρίσιμης περιοχής R_c.
4. Υπολογισμός της τιμής $q = g(x_1, \ldots, x_n)$ της Q από το δείγμα.
5. Εφαρμογή του κανόνα απόφασης και απόρριψη ή αποδοχή της H_0.
6. Υπολογισμός της ισχύος P του ελέγχου.

Το τελευταίο βήμα, λόγω της πολυπλοκότητας του, στις πρακτικές εφαρμογές συνήθως παραλείπεται. Τα πιο πάνω βήματα καθώς και η ορολογία διασαφηνίζονται στο εδάφιο που ακολουθεί.

3.5.3 Έλεγχος σημαντικότητας του συντελεστή συσχέτισης

Ως παράδειγμα εφαρμογής της πιο πάνω διαδικασίας θα δώσουμε τον έλεγχο της συμαντικότητας του συντελεστή συσχέτισης δύο τυχαίων μεταβλητών X και Y, βάσει του οποίου μπορούμε να αποφαινόμαστε αν οι μεταβλητές είναι γραμμικά συσχετισμένες.

Αν οι μεταβλητές είναι γραμμικά συσχετισμένες τότε ο συντελεστής συσχέτισής τους $\rho = \rho_{XY}$ θα είναι διαφορετικός από 0, αλλιώς θα είναι 0. Με βάση αυτή την παρατήρηση, προχωρούμε στα ακόλουθα βήματα της διαδικασίας στατιστικού ελέγχου.

1. Η μηδενική υπόθεση θα είναι $\rho = 0$ και η εναλλακτική υπόθεση $\rho \neq 0$. Κατά συνέπεια θα προβούμε σε αμφίπλευρο έλεγχο. (Αν θέλαμε να αποφασίσουμε σχετικά με το αν οι δύο μεταβλητές είναι θετικά γραμμικά συσχετισμένες θα διατυπώναμε την εναλλακτική υπόθεση ως $\rho > 0$, δηλαδή θα προβαίναμε σε μονόπλευρο έλεγχο).
2. Επιλέγουμε ως στατιστική συνάρτηση ελέγχου την

$$Q = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)\sqrt{n-3} = Z\sqrt{n-3} \tag{3.71}$$

όπου R ο δειγματικός συντελεστής συσχέτισης και Z η βοηθητική μεταβλητή του Fisher (βλ.εδάφιο 3.3.5), η οποία, με την υπόθεση H_0, έχει κατά προσέγγιση κανονική κατανομή με μέση τιμή 0 και τυπική

απόκλιση $1/\sqrt{n-3}$. Κατά συνέπεια η Q ακολουθεί την τυποποιημένη κανονική κατανομή $N(0, 1)$.†

3. Επιλέγουμε μια τιμή του επιπέδου σημαντικότητας, έστω $\alpha = 0.05$. Αν $z_{1-\alpha/2}$ είναι το $(1-\alpha/2)$-ποσοστημόριο της κανονικής κατανομής, τότε η αντίστοιχη κρίσιμη περιοχή R_c είναι η $|q| > z_{1-\alpha/2}$ ή $|q| > z_{0.975}$, ή τελικά $|q| > 1.96$, δεδομένου ότι

$$P(|Q| > z_{1-\alpha/2}) = P(Q < -z_{1-\alpha/2}) + P(Q > z_{1-\alpha/2})$$
$$= 2P(Q < z_{\alpha/2}) = 2\ \alpha/2 = \alpha$$

(Υπενθυμίζουμε ότι λόγω συμμετρίας της κανονικής πυκνότητας πιθανότητας ισχύει $z_{1-u} = z_u$. Ακόμη, σημειώνουμε ότι στην περίπτωση του μονόπλευρου ελέγχου η κρίσιμη περιοχή θα ήταν $q > z_{1-\alpha}$).

4. Η αριθμητική τιμή q προκύπτει από το δείγμα ως εξής:

$$q = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)\sqrt{n-3} \qquad r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{3.72}$$

5. Ο κανόνας απόφασης θα είναι ο εξής:

απορρίπτουμε την H_0 αν $|q| > z_{1-\alpha/2}$

και για $\alpha = 0.05$

---

† Συχνά στα βιβλία στατιστικής χρησιμοποιείται για τον ίδιο σκοπό η μεταβλητή

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

η οποία έχει κατανομή $t(n-1)$. Πρακτικώς τα αποτελέσματα των δύο μεθόδων ταυτίζονται (δεν διαφέρουν περισσότερο από 1% για συνήθη επίπεδα σημαντικότητας). Εδώ έχουμε προτιμήσει τον έλεγχο βάσει της μεταβλητής του Fisher λόγω της γενικότερης εφαρμογής της για κάθε ρ, σε αντίθεση με την T που εφαρμόζεται μόνο για $\rho = 0$.

$$\text{απορρίπτουμε την } H_0 \text{ αν } |q| = \frac{1}{2}\left|\ln\left(\frac{1+r}{1-r}\right)\right|\sqrt{n-3} > 1.96$$

Επιλύοντας την παραπάνω εξίσωση ως προς r βρίσκουμε ότι το κρίσιμο όριο r_c του δειγματικού συντελεστή συσχέτισης, το οποίο ορίζει την κρίσιμη περιοχή R_c για τη στατιστική συνάρτηση R, είναι

$$r_c = \frac{e^{3.92/\sqrt{n-3}} - 1}{e^{3.92/\sqrt{n-3}} + 1} \tag{3.73}$$

Ένας απλός και εύκολος στην απομνημόνευση τύπος που προσεγγίζει ικανοποιητικά την παραπάνω εξίσωση είναι ο ακόλουθος:

$$r_c = \frac{2}{\sqrt{n}} \tag{3.74}$$

Κατά συνέπεια, μπορούμε να κάνουμε τον έλεγχο της υπόθεσης με πιο άμεσο τρόπο, συγκρίνοντας την απόλυτη τιμή του r με την κρίσιμη τιμή r_c. Αν $|r| > r_c$ τότε θεωρούμε ότι υπάρχει στατιστικά σημαντική συσχέτιση ανάμεσα στις μεταβλητές

Εφαρμογή 3.5.3

Από ταυτόχρονα δείγματα ετήσιας βροχόπτωσης και απορροής μιας λεκάνης, μήκους 18 ετών, υπολογίστηκε συντελεστής συσχέτισης 0.58. Υπάρχει γραμμική συσχέτιση ανάμεσα στην ετήσια βροχόπτωση και στην απορροή της λεκάνης;

Υπολογίζουμε την κρίσιμη τιμή r_c χρησιμοποιώντας μία από τις εξισώσεις (3.73) ή (3.74). Εδώ για λόγους σύγκρισης χρησιμοποιούμε και τις δύο.

$$r_c = \frac{e^{3.92/\sqrt{15}} - 1}{e^{3.92/\sqrt{15}} + 1} = 0.470 \qquad r_c = \frac{2}{\sqrt{18}} = 0.471$$

Παρατηρούμε ότι και οι δύο εξισώσεις δίνουν πρακτικώς το ίδιο αποτέλεσμα. Αφού $0.58 > 0.47$ συμπεραίνουμε ότι υπάρχει στατιστικά σημαντική γραμμική συσχέτιση ανάμεσα στην ετήσια βροχόπτωση και απορροή της λεκάνης.