Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Ανάλυση Δεδομένων

Μάθημα 7ου εξαμήνου ΤΗΜΜΥ

Δημήτρης Κουγιουμτζής

E-mail: dkugiu@auth.gr

1 Νοεμβρίου 2022

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Έννοιες των πιθανοτήτων και της στατιστικής καθώς και μέθοδοι που στηρίζονται σε αυτές είναι χρήσιμα και απαραίτητα εργαλεία για να καταλάβουμε τον κόσμο γύρω μας, να μελετήσουμε φυσικά μεγέθη, φαινόμενα και διαδικασίες. Αν για ένα μέγεθος ή σύστημα που μελετάμε δεν υπάρχει καθόλου αβεβαιότητα ή τυχαιότητα δε χρειάζεται η πιθανοκρατική και στατιστική προσέγγιση αφού **καθοριστικά μοντέλα** (deterministic models) μπορούν να περιγράψουν το φαινόμενο επακριβώς. Για παράδειγμα τέτοιο σύστημα είναι αυτό των κινήσεων των πλανητών του ηλιακού συστήματος. Μπορούμε με μεγάλη ακρίβεια και χρησιμοποιώντας καθοριστικά μοντέλα να προσδιορίσουμε τη θέση των πλανητών που συμφωνεί με την πραγματική παρατήρηση.

Όμως τα δεδομένα που συλλέγουμε από τα περισσότερα φυσικά φαινόμενα και πραγματικές διαδικασίες δε μπορούν να εξηγηθούν ικανοποιητικά με μαθηματικά καθοριστικά μοντέλα. Αυτό συμβαίνει γιατί υπάρχει ο παράγοντας της αβεβαιότητας ή τυχαιότητας και για αυτό χρειάζεται να περιγράψουμε το σύστημα **πιθανοκρατικά** (probabilistically) και να καταφύγουμε σε **στατιστικές μεθόδους και μοντέλα** (statistical methods and models) για να λύσουμε προβλήματα εκτίμησης και πρόβλεψης σε τέτοια συστήματα.

Η **θεωρία πιθανοτήτων** μας επιτρέπει να μελετήσουμε τη μεταβλητότητα του αποτελέσματος ενός πειράματος (ή γενικά μιας πραγματοποίησης ενός φαινομένου ή μιας διαδικασίας) για το οποίο το ακριβές αποτέλεσμα δεν είναι δυνατόν να προβλεφθεί με ακρίβεια. Από την άλλη μεριά, η **στατιστική** συνίσταται στη συλλογή δεδομένων που λέγεται *δειγματοληψία* (sampling), στην περιγραφή τους, που αναφέρεται ως *περιγραφική στατιστική* (descriptive statistics) και κυρίως στην ανάλυση των δεδομένων που οδηγεί και στην απόκτηση συμπερασμάτων και για αυτό αναφέρεται ως *στατιστική συμπερασματολογία* (statistical inference). Ο συνδυασμός των εννοιών και ιδιοτήτων από τη θεωρία πιθανοτήτων με τις τεχνικές της στατιστικής είναι το αντικείμενο της **ανάλυσης δεδομένων** (data analysis). Σκοπός της ανάλυσης

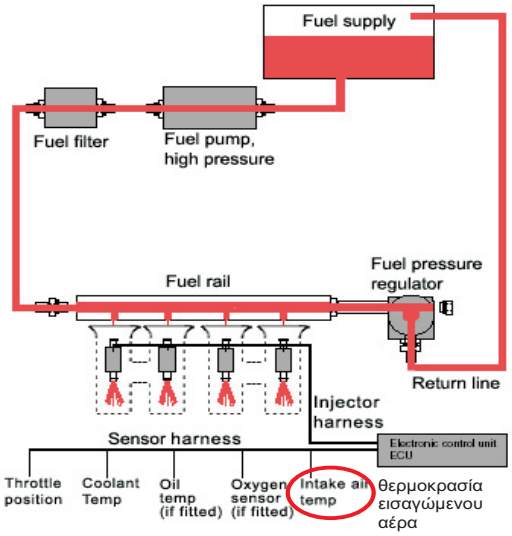
δεδομένων είναι η σύνοψη σε λίγες παραμέτρους της πληροφορίας από ένα σύνολο δεδομένων, το οποίο μπορεί να αποτελείται από πολλές παρατηρήσεις και να αφορά περισσότερα από ένα μεγέθη. Αυτή είναι η προσέγγιση που ακολουθείται στο πλαίσιο του μαθήματος με τίτλο 'Ανάλυση Δεδομένων' στο 7ο εξάμηνο του Νέου Προγράμματος Σπουδών του ΤΗΜΜΥ, ΑΠΘ.

Γενικότερα η ανάλυση δεδομένων, όπως ορίζεται στη Wikipedia, είναι η διαδικασία της επιθεώρησης, καθαρισμού, μετασχηματισμού και μοντελοποίησης των δεδομένων που έχει ως σκοπό να ανακαλύψει χρήσιμη πληροφορία, να δώσει συμπεράσματα και να υποστηρίξει τη λήψη αποφάσεων. Τα τελευταία έτη η ανάλυση δεδομένων έχει γνωρίσει ιδιαίτερο ενδιαφέρον και μέσα από τη χρήση νέων όρων όπως 'αναλύσεις δεδομένων' (data analytics) και 'αναλύσεις δεδομένων μεγάλης κλίμακας' (big data analytics), 'εξόρυξη δεδομένων' (data mining) και 'επιστήμη δεδομένων' (data science). Οι όροι αυτοί σχετίζονται άμεσα ή λιγότερα άμεσα με την ανάλυση δεδομένων. Δεν είναι ξεκάθαρη η διαφορά του όρου 'αναλύσεις δεδομένων' από την ανάλυση δεδομένων και περισσότερο χρησιμοποιείται στο χώρο της επιχειρηματικότητας και αγοράς και έχει περισσότερο επικοινωνιακό χαρακτήρα, δηλαδή να παρουσιάζει, οπτικοποιεί και επικοινωνεί τα αποτελέσματα της ανάλυσης. Ο όρος 'αναλύσεις δεδομένων μεγάλης κλίμακας' αναφέρεται σε δεδομένα μεγάλης κλίμακας, όπου πέρα από την ανάγκη της ανάλυσης δεδομένων (με την έννοια της χρήσης εργαλείων κυρίως της στατιστικής για την ανάλυση των δεδομένων), επεκτείνεται και σε άλλα θέματα που προκύπτουν λόγω του μεγάλου όγκου και πολυ-τροπικότητας των δεδομένων, όπως η απόκτηση και αποθήκευση τους, η αναζήτηση και μεταφορά στοιχείων από το σύνολο των δεδομένων, η ενημέρωση και οπτικοποίηση τους, καθώς και θέματα ιδιωτικότητας και προσβασιμότητας. Ο όρος 'εξόρυξη δεδομένων' έχει επικάλυψη με την ανάλυση δεδομένων, αλλά επικεντρώνεται περισσότερο στην ανακάλυψη προτύπων στα δεδομένα και χρησιμοποιεί μεθόδους της στατιστικής και πληροφορικής (μηχανική μάθηση). Τέλος ο όρος 'επιστήμη δεδομένων' είναι γενικότερος όρος, συμπεριλαμβάνει την ανάλυση δεδομένων, αλλά επεκτείνεται στη διαχείριση δεδομένων που μπορεί να είναι από διαφορετικές πηγές και διαφορετικών τύπων. Στο μάθημα αυτό, θα θεωρήσουμε ότι τα δεδομένα προς ανάλυση είναι καλά ορισμένα και αναφέρονται σε συγκεκριμένες μεταβλητές ενδιαφέροντος, χωρίς να μας απασχολεί αν είναι μικρού ή μεγάλου όγκου.

Ας δούμε κάποια παραδείγματα που αναδεικνύουν τα προβλήματα που αφορούν την ανάλυση δεδομένων, τις έννοιες και τα θέματα που θα μας απασχολήσουν στη συνέχεια.

Παράδειγμα 1.1. Μια μηχανή με αυτόματο ψεκασμό καυσίμου έχει τη μονάδα ελέγχου της μηχανής που συμπεριλαμβάνει αισθητήρες και μετρη-

τές, όπως αισθητήρα για τη θέση της βαλβίδας εισαγωγής ατμοποιημένου καυσίμου, τη θερμοκρασία αέρα, το οξυγόνο και το χρόνο ανάφλεξης (δες Σχήμα 1.1). Μπορούμε λοιπόν να συλλέξουμε μετρήσεις για όλα αυτά τα



Σχήμα 1.1: Διάγραμμα συστήματος ψεκασμού καυσίμου (αντιγραφή από τη διεύθυνση `http://www.twminduction.com`).

μεγέθη και ας σταθούμε για παράδειγμα στη θερμοκρασία του αέρα που εισάγεται στη μηχανή. Η θερμοκρασία του αέρα είναι *τυχαίο μέγεθος* που μπορεί να αλλάζει σε διάφορες χρονικές στιγμές ή σε διαφορετικές καταστάσεις λειτουργίας της μηχανής. Για τις μετρήσεις της θερμοκρασίας αέρα μπορεί να μας ενδιαφέρουν δύο βασικά χαρακτηριστικά. Το πρώτο είναι η *ακρίβεια επανάληψης* (precision), δηλαδή αν μεταβάλλονται πολύ ή λίγο οι μετρήσεις θερμοκρασίας αέρα (για τις ίδιες συνθήκες λειτουργίας). Το δεύτερο είναι η *ακρίβεια (ορθότητα)* (accuracy), δηλαδή κατά πόσο οι μετρήσεις θερμοκρασίας αέρα είναι κοντά στην επιθυμητή τιμή ή υπάρχουν συστηματικές αποκλίσεις. Η περιγραφή της τυχαίας μεταβολής της θερμοκρασίας αέρα είναι ένα θέμα της θεωρίας των πιθανοτήτων, που αναφέρεται ως *κατανομή μιας τυχαίας μεταβλητής*. Όταν έχουμε ένα πλήθος μετρήσεων της θερμοκρασίας αέρα μπορούμε να εκτιμήσουμε συγκεκριμένα χαρακτηριστικά της, όπως η μέση τιμή και η διασπορά της. Επίσης μπορούμε να συγκρίνουμε τα χαρακτηριστικά της θερμοκρασίας αέρα σε διαφορετικές συνθήκες λειτουργίας ή σε διαφορετικές μηχανές (κάτω από τις ίδιες συνθήκες λειτουργίας), ή ακόμα να διορθώσουμε τη μηχανή για να πετύχουμε καλύτερη ορθότητα στην τιμή της θερμοκρασίας αέρα.

Σε συνέχεια του παραδείγματος, θα θέλαμε επίσης να προσδιορίσουμε την ποσότητα καυσίμου που χρησιμοποιείται γνωρίζοντας άλλα μεγέθη σχετικά

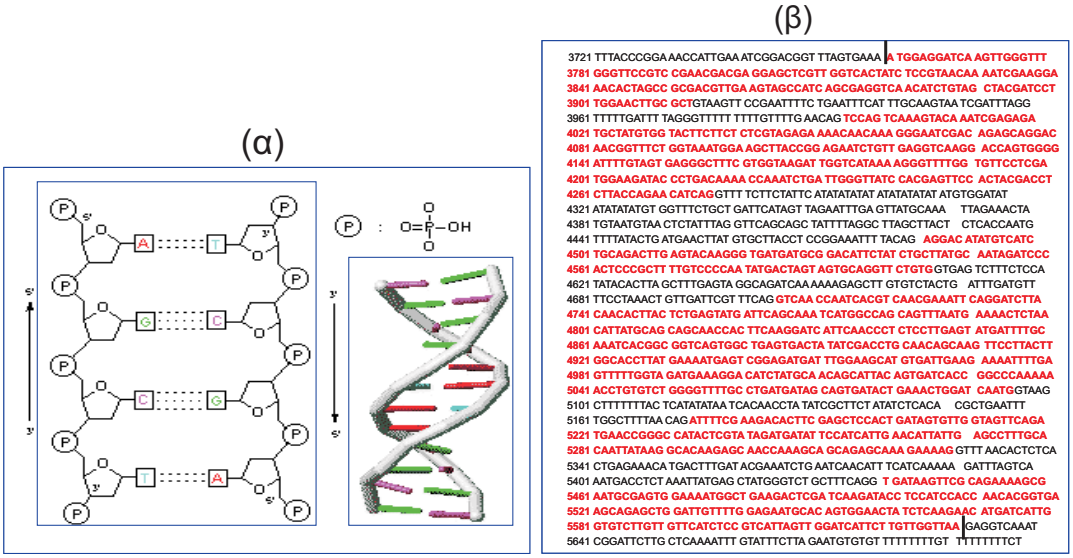
με τη λειτουργία της μηχανής την ίδια χρονική στιγμή ή διάρκεια, όπως είναι το άνοιγμα της βαλβίδας (ή ο ρυθμός ανοίγματος), ο χρόνος ανάφλεξης και η πίεση ψεκασμού. Ένα τέτοιο μοντέλο λέγεται *μουτέλο παλινδρόμησης* (regression model).

Ένα άλλο πρόβλημα, ίσως πιο δύσκολο να το διερευνήσουμε είναι η περιγραφή της χρονικής εξέλιξης της ποσότητας καυσίμου που καταναλώνεται κατά τη λειτουργία της μηχανής. Μια προσέγγιση είναι να θεωρήσουμε την εξέλιξη αυτή ως μια *καθοριστική διαδικασία* (deterministic process), δηλαδή να αποκλείσουμε την παρουσία τυχαίων διακυμάνσεων και παρεμβολών, που συνήθως αναφέρονται ως *θόρυβος* (noise), ή πιο ρεαλιστικά, να αγνοήσουμε την επίδραση τους. Αλλά αν θα θέλαμε να μελετήσουμε τη χρονική μεταβολή της θερμοκρασίας του αέρα στο σύστημα ψεκασμού της μηχανής θα ήταν πιο κατάλληλο να θεωρήσουμε τη διαδικασία ως *στοχαστική* (stochastic process), δηλαδή να συμπεριλάβουμε στην περιγραφή και το θόρυβο. Και οι δύο προσεγγίσεις αφορούν την *ανάλυση χρονοσειράς* (time series analysis) ενός μεγέθους (την ποσότητα καυσίμου ή τη θερμοκρασία αέρα), αποτελούν το αντικείμενο άλλου μαθήματος και για αυτό δε θα επεκταθούμε σε ανάλυση δεδομένων από χρονοσειρές.

Για τη θερμοκρασία του αέρα σε μια πολύπλοκη μηχανή θα μπορούσαν να υπάρχει πληθώρα άλλων παρατηρούμενων μεγεθών που μπορεί να επηρεάζει τη θερμοκρασία αέρα. Σε μια τέτοια περίπτωση θα θέλαμε να επιλέξουμε τα πιο σχετικά μεγέθη, είτε με απευθείας επιλογή από το σύνολο των μεγεθών ή μέσω κάποιου μετασχηματισμού. Γενικά η μείωση διάστασης σε δεδομένα από μεγάλο πλήθος μεγεθών είναι ένα θέμα που θα μας απασχολήσει στο τέλος του μαθήματος.

Παράδειγμα 1.2. Ένα άλλο παράδειγμα είναι η περιγραφή της δομής των στοιχείων στη σειρά του DNA (δες Σχήμα 1.2α).

Είναι γνωστό πως η σειρά του DNA αποτελείται από τέσσερα αμινοξέα, την Αδενίνη (A), κυτοσίνη (C), γουανίνη (G) και θυμίνη (T). Η μεταβλητή ενδιαφέροντος είναι το στοιχείο της σειράς DNA που δεν παίρνει αριθμητικές τιμές αλλά ένα από τα τέσσερα αυτά σύμβολα (δες Σχήμα 1.2β). Τμήματα της σειράς ορίζουν τα γονίδια (genes) που αποτελούνται από τις λεγόμενες κωδικοποιημένες περιοχές (με κόκκινο είναι οι κωδικοποιημένες περιοχές του γονιδίου που δίνεται στο Σχήμα 1.2β μεταξύ των κάθετων γραμμών) ενώ άλλα δεν έχουν κάποια γνωστή κωδικοποίηση (το μεγαλύτερο τμήμα των ανθρωπίνων χρωμοσωμάτων). Η περιγραφή της θέσης, της συχνότητας εμφάνισης και γενικά της δομής των τεσσάρων βάσεων όπως και συνδυασμών αυτών στη σειρά του DNA μπορεί να γίνει με τη βοήθεια πιθανοκρατικών (στατιστικών) μεθόδων για την ανάλυση διακριτών (κατηγορηματικών) δεδομένων ή συμβολοσειρών. Και εδώ υπάρχουν δύο προσεγγίσεις, η πρώτη θεωρώντας τη



Σχήμα 1.2: (α) Η δομή του DNA. (β) Ένα τμήμα της σειράς DNA που αποτελεί ένα γονίδιο.

συμβολοσειρά ως πραγματοποίηση κάποιας στοχαστικής αλυσίδας (διακριτής διαδικασίας) ή κάποιου δυναμικού συστήματος ορισμένο σε σύμβολα.

Παραθέτονται στη συνέχεια κάποιοι βασικοί ορισμοί που χρησιμοποιούνται στη θεωρία πιθανοτήτων και στη στατιστική ανάλυση:

τυχαία μεταβλητή (τ.μ.) (random variable): οποιοδήποτε χαρακτηριστικό του οποίου η τιμή αλλάζει στα διάφορα στοιχεία του πληθυσμού. Η τ.μ. μπορεί να είναι:

- **συνεχής (continuous):** να παίρνει τιμές σ' ένα διάστημα, όπως είναι η θερμοκρασία αέρα σε μια μηχανή ψεκασμού,
- **διακριτή (discrete):** να παίρνει μια τιμή σε ένα αριθμήσιμο σύνολο διακριτών τιμών, όπως ένα στοιχείο της σειράς DNA.

δεδομένα (data): ένα σύνολο τιμών μιας τ.μ. που έχουμε στη διάθεση μας, π.χ. μετρήσεις της θερμοκρασίας αέρα σε διάφορες χρονικές στιγμές ή σε διάφορες μηχανές ψεκασμού για τις ίδιες συνθήκες λειτουργίας, ή ένα κομμάτι της σειράς DNA.

πληθυσμός (population): μια ομάδα ή μια κατηγορία στην οποία αναφέρεται η τ.μ., το χρωμόσωμα No 22 ή ένας τύπος αυτόματης μηχανής με σύστημα ψεκασμού.

δείγμα (sample): ένα υποσύνολο του πληθυσμού που μελετάμε, π.χ. ένα κομμάτι της σειράς DNA του χρωμοσώματος Νο 22 ή 20 αυτόματες μηχανές με σύστημα ψεκασμού ίδιου τύπου.

παράμετρος (parameter): ένα μέγεθος που συνοψίζει με κάποιο τρόπο τις τιμές της τ.μ. στον πληθυσμό, π.χ. το ποσοστό εμφάνισης του στοιχείου Α στο χρωμόσωμα Νο 22, ή η μέση θερμοκρασία αέρα κάτω από κάποιες συνθήκες για έναν τύπο μηχανής με σύστημα ψεκασμού.

στατιστικό (statistic): ένα μέγεθος που συνοψίζει με κάποιο τρόπο τις τιμές της τ.μ. στο δείγμα, π.χ. το ποσοστό εμφάνισης του στοιχείου Α σ' ένα κομμάτι 1000 στοιχείων της σειράς DNA από το χρωμόσωμα Νο 22, ή ο μέσος όρος της θερμοκρασίας αέρα που υπολογίσαμε σε 20 αυτόματες μηχανές με σύστημα ψεκασμού κάτω από τις ίδιες συνθήκες.

Η μελέτη μιας τ.μ. με τη βοήθεια της πιθανοθεωρίας προϋποθέτει ότι γνωρίζουμε (ή υποθέτουμε) την κατανομή της τ.μ., και άρα και τον πληθυσμό στον οποίο παίρνει τιμές, καθώς και τις παραμέτρους της. Στην πράξη βέβαια κάτι τέτοιο δε συμβαίνει, αλλά σε κάποια προβλήματα μπορούμε να υποθέτουμε γνωστές κατανομές. Σε κάθε περίπτωση, ένα από τα κύρια προβλήματα στην ανάλυση δεδομένων είναι να εκτιμήσουμε τις άγνωστες (αλλά σταθερές) παραμέτρους του πληθυσμού από τα γνωστά αλλά μεταβλητά στατιστικά του δείγματος δεδομένων που έχουμε στη διάθεση μας.

Στο γενικό πλαίσιο της στοχαστικής προσέγγισης που χρησιμοποιείται στην ανάλυση δεδομένων, υπάρχουν δύο κύριες κατευθύνσεις. Η πρώτη αναφέρεται ως ανάλυση κατά Bayes ή Μπεϋζιανή ανάλυση ή προσέγγιση (Bayesian approach) και υποθέτει από πριν κάποια κατανομή ή πιθανότητες σχετικά με το πρόβλημα που μελετάμε. Η δεύτερη κατεύθυνση δε χρησιμοποιεί κάποια υπόθεση κατανομής αλλά υπολογίζει τις πιθανότητες από τη συχνότητα εμφάνισης που υπολογίζεται απευθείας από τα δεδομένα και αναφέρεται ως ανάλυση με βάση τις συχνότητες (frequentist approach). Στα θέματα που θα μελετήσουμε δε θα εμβαθύνουμε στη Μπεϋζιανή προσέγγιση.

Κεφάλαιο 2

Πιθανότητες και Τυχαίες Μεταβλητές

Μπορούμε να καταλάβουμε την έννοια της πιθανότητας από τη σχετική συχνότητα εμφάνισης n_i κάποιας τιμής x_i μιας διακριτής τ.μ. X. Αν είχαμε τη δυνατότητα να συλλέξουμε αυθαίρετα πολλές n παρατηρήσεις ($n \to \infty$), τότε το όριο της σχετικής συχνότητας είναι η **πιθανότητα** η τ.μ. X να πάρει την τιμή x_i

$$\mathrm{P}(x_i) \equiv \mathrm{P}(X = x_i) = \lim_{n \longrightarrow \infty} \frac{n_i}{n} \tag{2.1}$$

(το σύμβολο ≡ σημαίνει ισοδυναμία συμβολισμού). Για να είναι έγκυρος αυτός ο ορισμός πρέπει επίσης να υποθέσουμε ότι οι συνθήκες για την τ.μ. X σε κάθε επανάληψη της παρατήρησης παραμένουν οι ίδιες, και αυτή η ιδιότητα ονομάζεται *στατιστική ομαλότητα* (statistical regularity). Για παράδειγμα, η πιθανότητα βροχής σε μια περιοχή (σε μια τυχαία μέρα του χρόνου ή ενός συγκεκριμένου μήνα) μπορεί να έχει αλλάξει τα τελευταία χρόνια λόγω του φαινομένου του θερμοκηπίου.

Για συνεχή τ.μ. X δεν έχει νόημα να μιλάμε για την πιθανότητα η X να πάρει μια συγκεκριμένη τιμή αλλά για την πιθανότητα η X να ανήκει σε ένα διάστημα τιμών $\mathrm{d}x$. Ποια είναι η πιθανότητα να έχει κάποιος συμφοιτητής σας ένα συγκεκριμένο ύψος που ορίζεται με ακρίβεια πολλών (άπειρων) δεκαδικών, π.χ. 1.80123256538634255; Μπορείτε όμως να προσδώσετε μη μηδενική πιθανότητα για το γεγονός ότι ένας συμφοιτητής σας έχει ύψος στα 1.80 μέτρα (όπου με βάση τη στρογγυλοποίηση του εκατοστού έχουμε $\mathrm{d}x = [1.795, 1.805)$).

2.1 Κατανομή πιθανότητας

2.1.1 Κατανομή πιθανότητας μιας τ.μ.

Η πιθανότητα η τ.μ. X να πάρει κάποια τιμή x_i, αν είναι διακριτή, ή να βρίσκεται σε ένα διάστημα τιμών $\mathrm{d}x$, αν είναι συνεχής, μπορεί να μεταβάλλεται στο σύνολο των διακεκριμένων τιμών ή σε διαφορετικά διαστήματα και δίνεται ως συνάρτηση της τ.μ. X. Για διακριτή τ.μ. X που παίρνει τις τιμές $x_1, x_2, \ldots, x_m$, η συνάρτηση αυτή λέγεται **συνάρτηση μάζας πιθανότητας, σμπ** (probability mass function), ορίζεται ως $f_X(x_i) = \mathrm{P}(X = x_i)$ και ικανοποιεί τις συνθήκες

$$f_X(x_i) \geq 0 \qquad \text{και} \qquad \sum_{i=1}^{m} f_X(x_i) = 1. \tag{2.2}$$

Αντίστοιχα, για συνεχή τ.μ. X ($X \in \mathbf{R}$) ορίζεται η **συνάρτηση πυκνότητας πιθανότητας, σππ** (probability density function) $f_X(x)$ που ικανοποιεί τις συνθήκες

$$f_X(x) \geq 0 \qquad \text{και} \qquad \int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1. \tag{2.3}$$

Η *κατανομή πιθανότητας* (probability distribution) της τ.μ. X ορίζεται επίσης από την **αθροιστική συνάρτηση κατανομής, ασκ** (cumulative distribution function) $F_X(x)$, που δηλώνει την πιθανότητα η τ.μ. X να πάρει τιμές μικρότερες ή ίσες από κάποια τιμή x. Για διακριτή τ.μ. X είναι

$$F_X(x_i) = \mathrm{P}(X \leq x_i) = \sum_{x \leq x_i} f_X(x) \tag{2.4}$$

και για συνεχή τ.μ. X

$$F_X(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)\mathrm{d}u. \tag{2.5}$$

Σημειώνεται ότι μια συνεχής μεταβλητή μπορεί να μετατραπεί σε διακριτή με κατάλληλη διαμέριση του πεδίου τιμών της. Αν η συνεχής τ.μ. X ορίζεται στο διάστημα $[a, b]$ μια διαμέριση Σ σε m κελιά δίνεται ως

$$\Sigma = \{a = r_0, r_1, \ldots, r_{m-1}, r_m = b\}, \quad \text{όπου} \quad r_0 < r_1 < \ldots < r_m.$$

Αντιστοιχίζοντας διακεκριμένες τιμές x_i, $i = 1, \ldots, m$, σε κάθε κελί (διάστημα) $[r_{i-1}, r_i)$, η πιθανότητα εμφάνισης μιας τιμής x_i της διακριτικοποιημένης τ.μ. X', $f_{X'}(x_i) = \mathrm{P}(X' = x_i)$, δίνεται από την πιθανότητα η συνεχής τ.μ. X να παίρνει τιμές στο διάστημα $[r_{i-1}, r_i)$, $\mathrm{P}(r_{i-1} \leq X \leq r_i) = F_X(r_i) - F_X(r_{i-1})$.

2.1.2 Από κοινού πιθανότητα δύο τ.μ.

Σε πολλά προβλήματα χρειάζεται να ορίσουμε την συνδυασμένη μεταβλητότητα δύο τ.μ. X και Y, δηλαδή την από κοινού κατανομή πιθανότητας τους. Έστω η διακριτή τ.μ. X με δυνατές διακεκριμένες τιμές $x_1, x_2, \ldots, x_n$ και Y με δυνατές διακεκριμένες τιμές $y_1, y_2, \ldots, y_m$, αντίστοιχα. Η **από κοινού συνάρτηση μάζας πιθανότητας** (joint probability mass function) $f_{XY}(x, y)$ ορίζεται για κάθε ζεύγος δυνατών τιμών (x_i, y_i) ως

$$f_{XY}(x_i, y_i) = \mathrm{P}(X = x_i, Y = y_i) \tag{2.6}$$

και η **από κοινού αθροιστική συνάρτηση κατανομής** (joint cumulative density function) ορίζεται ως

$$F_{XY}(x_i, y_i) = \mathrm{P}(X \leq x_i, Y \leq y_i) = \sum_{x \leq x_i} \sum_{y \leq y_i} f_{XY}(x_i, y_i). \tag{2.7}$$

Η **από κοινού συνάρτηση πυκνότητας πιθανότητας** (joint probability density function) $f_{XY}(x, y)$ για δύο συνεχείς τ.μ. X και Y θα πρέπει να ικανοποιεί τις συνθήκες

$$f_{XY}(x, y) \geq 0 \qquad \text{και} \qquad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) \mathrm{d}y \mathrm{d}x = 1. \tag{2.8}$$

Η **από κοινού (αθροιστική) συνάρτηση κατανομής** για δύο συνεχείς τ.μ. X και Y ορίζεται ως

$$F_{XY}(x, y) = \mathrm{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(u, v) \mathrm{d}v \mathrm{d}u. \tag{2.9}$$

Δύο τ.μ. X και Y (συνεχείς ή διακριτές) είναι **ανεξάρτητες** (independent) αν για κάθε δυνατό ζεύγος τιμών τους (x, y) ισχύει

$$f_{XY}(x, y) = f_X(x) f_Y(y) \tag{2.10}$$

Με ανάλογο τρόπο ορίζονται οι συναρτήσεις από κοινού κατανομής για περισσότερες τ.μ. καθώς και η ανεξαρτησία πολλών τ.μ..

2.2 Παράμετροι κατανομής τυχαίων μεταβλητών

Η κατανομή πιθανότητας περιγράφει πλήρως τη συμπεριφορά της τ.μ., αλλά συνήθως στην πράξη δεν είναι γνωστή ή απαραίτητη. Όταν μελετάμε μια τ.μ. μας ενδιαφέρει κυρίως να προσδιορίσουμε κάποια βασικά χαρακτηριστικά της κατανομής της, όπως η *κεντρική τάση* και η *μεταβλητότητα* της τ.μ.. Αυτά τα χαρακτηριστικά είναι οι παράμετροι της κατανομής της τ.μ..

2.2.1 Μέση τιμή

Αν X είναι μια διακριτή τ.μ. που παίρνει m διακριτές τιμές $x_1, x_2, \ldots, x_m$, με σμπ $f_X(x)$, η μέση τιμή της, που συμβολίζεται $\mu_X \equiv \mathrm{E}[X]$ ή απλά μ, δίνεται ως

$$\mu \equiv \mathrm{E}[X] = \sum_{i=1}^{m} x_i f_X(x_i). \tag{2.11}$$

Αν η X είναι συνεχής τ.μ. με σπп $f_X(x)$, η μέση τιμή της δίνεται ως

$$\mu \equiv \mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x) \mathrm{d}x. \tag{2.12}$$

Κάποιες βασικές ιδιότητες της μέσης τιμής είναι:

1. Αν η τ.μ. X παίρνει μόνο μια σταθερή τιμή c είναι $\mathrm{E}[X] = c$.
2. Αν X είναι μια τ.μ. και c είναι μια σταθερά: $\mathrm{E}[cX] = c\mathrm{E}[X]$.
3. Αν X και Y είναι δύο τ.μ.: $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$.
4. Αν X και Y είναι δύο ανεξάρτητες τ.μ.: $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$.

Οι ιδιότητες (2) και (3) δηλώνουν πως η μέση τιμή έχει τη γραμμική ιδιότητα, δηλαδή ισχύει $\mathrm{E}[aX + bY] = a\mathrm{E}[X] + b\mathrm{E}[Y]$.

Άλλα χαρακτηριστικά της τ.μ. X εκτός της μέσης τιμής είναι τα εκατοστιαία σημεία που μπορούν να προσδιοριστούν από την αθροιστική συνάρτηση κατανομής. Η **διάμεσος** (median) $\tilde{\mu}$ μιας τ.μ. X είναι το 50-εκατοστιαίο σημείο, δηλαδή η $\tilde{\mu}$ ικανοποιεί τη σχέση $F_X(\tilde{\mu}) = 0.5$.

2.2.2 Διασπορά

Η **διασπορά** ή **διακύμανση** (variance) μιας τ.μ. X και κυρίως η **τυπική απόκλιση** (standard deviation) (που είναι η τετραγωνική ρίζα της διασποράς), εκφράζουν τη μεταβλητότητα της τ.μ. X γύρω από τη μέση τιμή. Αν X είναι μια τ.μ. (διακριτή ή συνεχής) με μέση τιμή μ, τότε η διασπορά της που συμβολίζεται $\sigma_X^2 \equiv \mathrm{Var}[X]$ ή απλά σ^2 δίνεται ως

$$\sigma^2 \equiv \mathrm{E}[(X - \mu)^2] = \mathrm{E}[X^2] - \mu^2. \tag{2.13}$$

Κάποιες βασικές ιδιότητες της διασποράς είναι:

1. Αν η τ.μ. X παίρνει μόνο μια σταθερή τιμή c είναι $\mathrm{Var}[c] = 0$.
2. Αν X είναι μια τ.μ. και c είναι μια σταθερά: $\mathrm{Var}[X + c] = \mathrm{Var}[X]$ και $\mathrm{Var}[cX] = c^2\mathrm{Var}[X]$.

3. Αν X και Y είναι δύο ανεξάρτητες τ.μ.: $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$.

Οι ιδιότητες (2) και (3) δηλώνουν πως η διασπορά δεν έχει τη γραμμική ιδιότητα. Όταν όμως οι δύο τ.μ. είναι ανεξάρτητες η διασπορά έχει τη ψευδογραμμική ιδιότητα, δηλαδή ισχύει $\mathrm{Var}[aX + bY] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y]$.

2.2.3 Ροπές μιας τ.μ.

Συχνά για την περιγραφή των ιδιοτήτων μιας τ.μ. X δεν αρκεί μόνο η μέση τιμή και η διασπορά (που βασίζονται στην πρώτη και δεύτερη δύναμη της X), αλλά πρέπει να καταφύγουμε σε μεγαλύτερες δυνάμεις της X. Η μέση τιμή και η διασπορά αναφέρονται και ως ροπή πρώτης τάξης και (κεντρική) ροπή δεύτερης τάξης, αντίστοιχα. Γενικά ορίζονται οι ροπές $\mathrm{E}[X^n]$ και οι κεντρικές ροπές $\mu_n \equiv \mathrm{E}[(x-\mu_X)^n]$ για κάθε τάξη n.

Από τις σημαντικότερες ροπές είναι η κεντρική ροπή τρίτης τάξης που χρησιμοποιείται στον ορισμό του συντελεστή λοξότητας (coefficient of skewness) λ

$$\lambda = \frac{\mu_3}{\sigma^3} = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]. \tag{2.14}$$

Ο συντελεστής λοξότητας λ εκφράζει τη λοξότητα της κατανομής της τ.μ. X (δηλαδή της σμπ για διακριτή X ή της σππ για συνεχή X). Για $\lambda = 0$ η κατανομή είναι συμμετρική.

Από τη κεντρική ροπή τέταρτης τάξης ορίζεται ο συντελεστή κύρτωσης (coefficient of kurtosis) κ

$$\kappa = \frac{\mu_4}{\sigma^4} - 3 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3. \tag{2.15}$$

Ο συντελεστής κύρτωσης κ δηλώνει τη σχέση του πλάτους της κατανομής γύρω από τη κεντρική τιμή με τις ουρές της. Πιο συγκεκριμένα δηλώνει κατά πόσο αυτή η σχέση αποκλίνει από αυτήν της τυπικής κανονικής κατανομής όπου ορίζεται να είναι ίση με 3 (και για αυτό αφαιρείται) (Η τυπική κανονική κατανομή παρουσιάζεται παρακάτω).

2.2.4 Συνδιασπορά και συντελεστής συσχέτισης

Όταν μελετάμε δύο τ.μ. X και Y που δεν είναι ανεξάρτητες, έχει ενδιαφέρον να προσδιορίσουμε πόσο ισχυρά συσχετίζεται η μια με την άλλη. Γι αυτό ορίζουμε τη **συνδιασπορά** ή **συνδιακύμανση** (covariance) των τ.μ. X και Y, που συμβολίζεται $\sigma_{XY} \equiv \mathrm{Cov}[X,Y]$, και ορίζεται ως

$$\sigma_{XY} \equiv \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \mathrm{E}[XY] - \mu_X\mu_Y. \tag{2.16}$$

Αν οι δύο τ.μ. X και Y συσχετίζονται ισχυρά και θετικά, δηλαδή όταν αυξάνει η μία αυξάνει και η άλλη, τότε η συνδιασπορά παίρνει μεγάλη θετική τιμή σε σχέση με τις τιμές των X και Y. Αντίθετα αν οι δύο τ.μ. X και Y συσχετίζονται ισχυρά και αρνητικά, δηλαδή όταν αυξάνει η μία μειώνεται η άλλη, τότε η συνδιασπορά παίρνει μεγάλη αρνητική τιμή. Αν οι X και Y είναι ανεξάρτητες εύκολα μπορεί να δειχθεί από την (2.16) ότι $\sigma_{XY} = 0$.

Το μειονέκτημα της συνδιασποράς είναι ότι η τιμή της εξαρτάται από τις μονάδες μέτρησης των τ.μ. X και Y. Γι αυτό όταν θέλουμε να μετρήσουμε τη συσχέτιση δύο τ.μ. X και Y, χρησιμοποιούμε συνήθως το **συντελεστή συσχέτισης** (correlation coefficient), που συμβολίζεται $\rho_{XY} \equiv \mathrm{Corr}[X, Y]$ ή απλά ρ, και προκύπτει από την κανονικοποίηση της συνδιασποράς με το γινόμενο των τυπικών αποκλίσεων των X και Y

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{2.17}$$

Παραθέτουμε κάποιες ιδιότητες του συντελεστή συσχέτισης:

1. $-1 \leq \rho \leq 1$.

2. Αν οι τ.μ. X και Y είναι ανεξάρτητες είναι $\rho = 0$, αλλά $\rho = 0$ δε δηλώνει ότι οι X και Y είναι ανεξάρτητες αλλά απλά ότι δεν είναι γραμμικά συσχετισμένες (μπορεί δηλαδή να είναι μη-γραμμικά συσχετισμένες).

3. $\rho = -1$ ή $\rho = 1$ αν και μόνο αν $Y = a + \beta X$ για κάποιους αριθμούς a και β.

2.3 Γνωστές κατανομές μιας τ.μ.

Στη μελέτη μιας τ.μ. μας βοηθάει να έχουμε κάποια πρότυπα για την κατανομή της, δηλαδή κάποιες γνωστές συναρτήσεις $f_X(x)$ της τ.μ. X με γνωστές παραμέτρους. Επίσης σε πολλά πραγματικά προβλήματα η κατανομή μιας τ.μ. μπορεί να περιγραφεί ικανοποιητικά από κάποια γνωστή κατανομή. Θα παραθέσουμε εδώ δύο τέτοια παραδείγματα πολύ γνωστών κατανομών, μια για διακριτή και μια για συνεχή τ.μ..

2.3.1 Διωνυμική κατανομή

Πολλά προβλήματα και κυρίως πειράματα εμπεριέχουν **επαναλαμβανόμενες δοκιμές** (repeated trials). Για παράδειγμα μπορεί να θέλουμε να γνωρίζουμε την πιθανότητα η μια στις 5 βελόνες χαρακτικής να σπάσει σε ένα πείραμα αντοχής τάνυσης. Σε κάθε δοκιμή ορίζουμε δύο μόνο δυνατά

αποτελέσματα που συνήθως τα χαρακτηρίζουμε συμβολικά ως 'επιτυχία' και 'αποτυχία', χωρίς το όνομα να έχει απαραίτητα πραγματική σημασία ('επιτυχία' μπορεί να είναι το σπάσιμο της βελόνας). Υποθέτουμε ότι έχουμε κάνει n δοκιμές και η πιθανότητα 'επιτυχίας' p σε κάθε προσπάθεια είναι ίδια. Δοκιμές που τηρούν αυτές τις προϋποθέσεις λέγονται **δοκιμές Bernoulli**.

Ορίζουμε την τ.μ. X ως τον αριθμό των επιτυχίων σε n δοκιμές. Η X παίρνει τιμές στο σύνολο $\{0, 1, \ldots, n\}$. Η πιθανότητα να έχουμε x 'επιτυχίες' δίνεται από τη **διωνυμική** (binomial) σμπ, που συμβολίζεται $\mathrm{B}(n, p)$, και ορίζεται ως

$$f_X(x) = \mathrm{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{2.18}$$

όπου $\binom{n}{x} \equiv \frac{n!}{x!(n-x)!}$ είναι ο *διωνυμικός συντελεστής* (binomial coefficient). Τα n και p είναι οι παράμετροι που ορίζουν τη διωνυμική κατανομή. Δηλώνουμε ότι μια τ.μ. X ακολουθεί διωνυμική κατανομή ως $X \sim \mathrm{B}(n, p)$. Η μέση τιμή και η διασπορά της X είναι

$$\mu = \mathrm{E}[X] = np \qquad \text{και} \qquad \sigma^2 = \mathrm{Var}[X] = np(1-p). \tag{2.19}$$

Παράδειγμα 2.1. Σε ένα πείραμα αντοχής τάνυσης δοκιμάζουμε 4 βελόνες χαρακτικής σε ένα συγκεκριμένο όριο τάνυσης. Η πιθανότητα να σπάσει η βελόνα σε μια δοκιμή είναι $p = 0.2$. Οι δοκιμές είναι τύπου Bernoulli. Μπορούμε να ορίσουμε την πιθανότητα να μη σπάσει καμιά βελόνα στις 4 δοκιμές από τη διωνυμική κατανομή ως

$$f_X(0) = \mathrm{P}(X = 0) = \binom{4}{0} 0.2^0 0.8^4 = 0.4096$$

Όμοια υπολογίζονται οι πιθανότητες όταν στις 4 δοκιμές σπάζει μια βελόνα κι όταν σπάζουν 2, 3 και 4 βελόνες

$$f_X(1) = 0.4096 \quad f_X(2) = 0.1536 \quad f_X(3) = 0.0256 \quad f_X(4) = 0.0016.$$

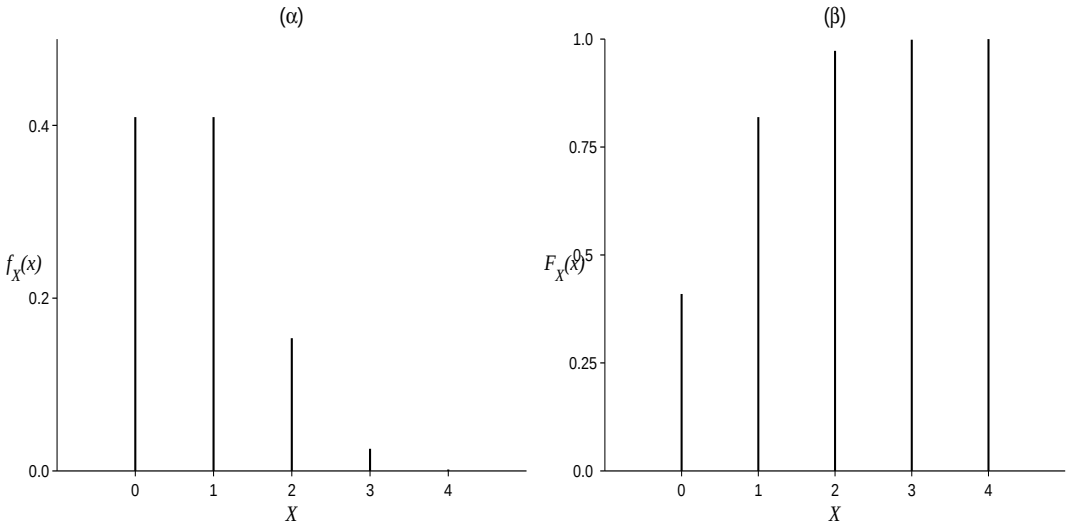
Με αυτόν τον τρόπο έχουμε ορίσει τη συνάρτηση σμπ $f_X(x)$ γι αυτό το παράδειγμα. Από την $f_X(x)$ ορίζεται εύκολα και η αθροιστική συνάρτηση $F_X(x) \equiv \mathrm{P}(X \leq x)$. Η γραφική παράσταση των συναρτήσεων $f_X(x)$ και $F_X(x)$ δίνονται στο Σχήμα 2.1.

Η πιθανότητα να σπάσει η βελόνα τουλάχιστον μια φορά είναι

$$\mathrm{P}(X \geq 1) = 1 - \mathrm{P}(X = 0) = 1 - 0.4096 = 0.5904$$

ενώ η πιθανότητα να σπάσει η βελόνα το πολύ δύο φορές δίνεται από την αθροιστική συνάρτηση κατανομής

$$F_X(2) \equiv \mathrm{P}(X \leq 2) = \sum_{x=0}^{2} \mathrm{P}(X = x) = 0.4096 + 0.4096 + 0.1536 = 0.9728.$$



Σχήμα 2.1: Γραφική παράσταση της συνάρτησης $f_X(x)$ στο (α) και της συνάρτησης $F_X(x)$ στο (β) για τον αριθμό βελόνων που σπάζουν σε n δοκιμές.

Η μέση τιμή για τον αριθμό 'επιτυχιών' (όπου επιτυχία είναι το σπάσιμο της βελόνας στη δοκιμή) είναι $\mathrm{E}[X] = 4 \cdot 0.2 = 0.8$ δηλαδή στις 4 δοκιμές περίπου μια φορά θα σπάζει η βελόνα. Η τυπική απόκλιση από αυτήν την τιμή είναι

$$\sigma = \sqrt{\mathrm{VarX}} = \sqrt{4 \cdot 0.2 \cdot 0.8} = 0.8.$$

2.3.2 Ομοιόμορφη κατανομή

Η πιο απλή συνεχής κατανομή είναι η ομοιόμορφη κατανομή που ορίζεται σε πεπερασμένο διάστημα $[a, b]$ και έχει σππ (δες Σχήμα 2.2)

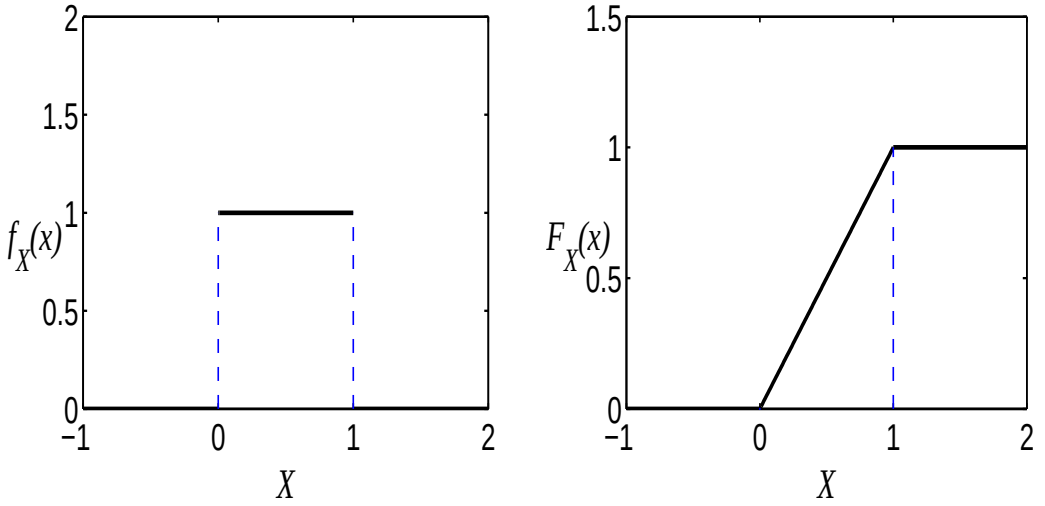
$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{αλλού} \end{cases} \tag{2.20}$$

και ασκ

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x \ge b \end{cases} \tag{2.21}$$

Ο συμβολισμός που χρησιμοποιείται για να δείξουμε ότι μια τ.μ. X ακολουθεί ομοιόμορφη κατανομή στο διάστημα $[a, b]$ είναι $X \sim \mathrm{U}[a, b]$. Η μέση τιμή και διασπορά της X είναι

$$\mu = \mathrm{E}[X] = \frac{a+b}{2} \qquad \text{και} \qquad \sigma^2 = \mathrm{Var}[X] = \frac{(b-a)^2}{12}. \tag{2.22}$$



Σχήμα 2.2: Η συνάρτηση πυκνότητας πιθανότητας στο (α) και η αθροιστική συνάρτηση στο (β) της ομοιόμορφης κατανομής στο διάστημα $[a, b]$.

2.3.3 Δημιουργία τυχαίων αριθμών από δεδομένη κατανομή μέσω της ομοιόμορφης κατανομής

Συχνά σε προσομοιώσεις χρειάζεται να δημιουργήσουμε τυχαίους αριθμούς από κάποια δεδομένη κατανομή συνεχούς τ.μ., δηλαδή κατανομή που ορίζεται με κάποια γνωστή σππ ή εναλλακτικά ασκ και για ορισμένες τιμές των παραμέτρων της. Ένα χρήσιμο αποτέλεσμα που χρησιμοποιείται στη δημιουργία τυχαίων αριθμών από δεδομένη κατανομή είναι το παρακάτω θεώρημα.

Θεώρημα 2.1. Αν $X \sim \mathrm{U}[0, 1]$ τότε η τ.μ. $Y = F_Y^{-1}(X)$ έχει ασκ $F_Y(y)$.

Αν λοιπόν γνωρίζουμε την ασκ $F_Y(y)$ μιας τ.μ. Y για την οποία θέλουμε να παράγουμε τυχαίους αριθμούς, μπορούμε να υπολογίσουμε την αντίστροφη ασκ $F_Y^{-1}(.)$, που είναι πάντα εφικτό αφού η $F_Y(y)$ είναι μονότονη. Θεωρούμε πως έχουμε κάποια γεννήτρια συνάρτηση ψευδο-τυχαίων αριθμών για να παράγουμε τυχαίους αριθμούς x στο διάστημα $[0, 1]$, δηλαδή να παράγουμε τιμές x της $X \sim \mathrm{U}[0, 1]$ (δες άσκηση 1). Εφαρμόζοντας την $F_Y^{-1}(.)$ σε κάθε τιμή x της $X \sim \mathrm{U}[0, 1]$ θα μας δώσει την αντίστοιχη τιμή y της Y που ακολουθεί τη δεδομένη κατανομή με ασκ $F_Y(y)$.

Παράδειγμα 2.2. Η εκθετική κατανομή δίνεται από την σππ $f_Y(y) = \lambda e^{-\lambda y}$ και ασκ $F_Y(y) = 1 - e^{-\lambda y}$, όπου λ η παράμετρος της εκθετικής κατανομής (που είναι και η μέση τιμή). Θέτοντας $X \equiv F_Y(y)$, έχουμε $X \sim \mathrm{U}[0, 1]$ και

μπορούμε να δημιουργήσουμε τυχαίους αριθμούς από ομοιόμορφη κατανομή. Τότε για κάθε τέτοια τιμή x υπολογίζουμε την αντίστοιχη τιμή y από εκθετική κατανομή με παράμετρο λ από την αντίστροφη της $F_Y(y)$

$$y = -\frac{1}{\lambda}\ln(1-x).$$

2.3.4 Κανονική κατανομή

Η κανονική κατανομή είναι η σπουδαιότερη συνεχής κατανομή και αποτελεί τη βάση για πολλά στατιστικά μοντέλα και συμπεράσματα. Η σπουδαιότητα της οφείλεται κυρίως στο ότι περιγράφει ικανοποιητικά την κατανομή πολλών τυχαίων πραγματικών μεγεθών που παίρνουν συνεχείς αριθμητικές τιμές, αλλά προσεγγίζει ικανοποιητικά και πολλές διακριτές κατανομές. Σε πολλά τυχαία μεγέθη που μελετάμε παρατηρούμε ότι οι τιμές τους 'μαζεύονται' συμμετρικά γύρω από μια κεντρική τιμή και 'αραιώνουν' καθώς απομακρύνονται από αυτήν την κεντρική τιμή. Η κατάλληλη συνάρτηση πυκνότητας πιθανότητας για μια κατανομή τέτοιου τύπου 'τομή καμπάνας' είναι αυτή της **κανονικής κατανομής** (normal distribution) που ορίζεται ως

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty, \tag{2.23}$$

όπου οι παράμετροι μ και σ^2 που ορίζουν την κανονική κατανομή είναι η μέση τιμή και η διασπορά αντίστοιχα και η κατανομή συμβολίζεται ως $N(\mu, \sigma^2)$. Η αθροιστική συνάρτηση κατανομής $F_X(x)$ δίνεται από το ολοκλήρωμα της $f_X(x)$ όπως ορίστηκε στην (2.5). Στο Σχήμα 2.3 δίνονται σχηματικά η $f_X(x)$ και η $F_X(x)$.

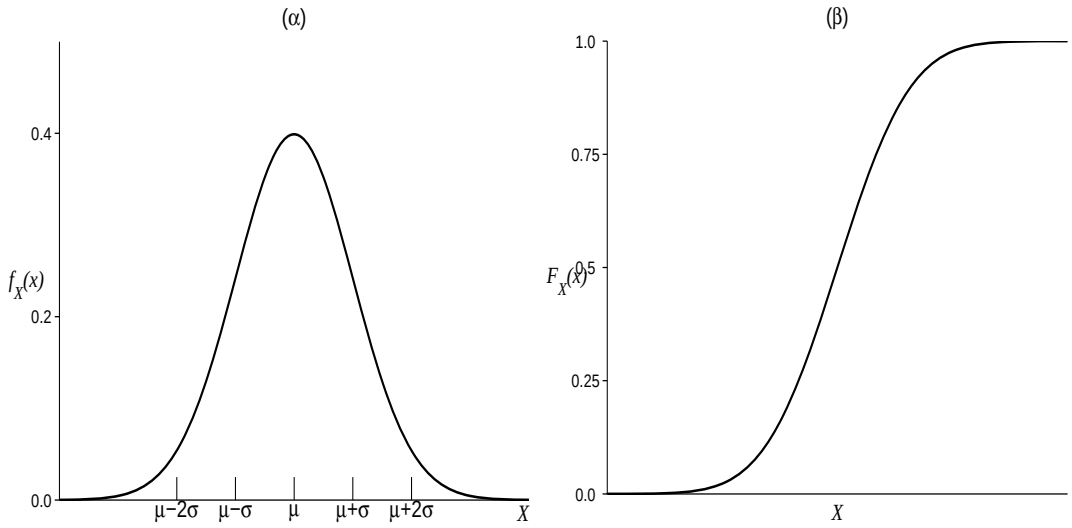
Φαίνεται ότι περίπου το 70% των τιμών της X βρίσκονται στο διάστημα $[\mu-\sigma, \mu+\sigma]$ και περίπου το 95% των τιμών της X βρίσκονται στο διάστημα $[\mu-2\,\sigma, \mu+2\,\sigma]$.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η **τυπική ή τυποποιημένη κανονική κατανομή** (standard normal distribution) που είναι η πιο απλή μορφή της κανονικής κατανομής, δηλαδή για $\mu = 0$ και $\sigma = 1$. Για να ξεχωρίσουμε την τ.μ. που ακολουθεί τυπική κανονική κατανομή τη συμβολίζουμε με Z ή z και είναι $Z \sim N(0,1)$. Η σππ είναι

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \qquad -\infty < z < \infty. \tag{2.24}$$

Η αθροιστική συνάρτηση συμβολίζεται $\Phi(z)$ και είναι

$$\Phi(z) \equiv F_Z(z) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}}\,du \qquad -\infty < z < \infty. \tag{2.25}$$



Σχήμα 2.3: Η συνάρτηση πυκνότητας πιθανότητας στο (α) και η αθροιστική συνάρτηση στο (β) της κανονικής κατανομής.

Οι τιμές της $\Phi(z)$ για διάφορες τιμές του z είναι πολύ χρήσιμες στη στατιστική γι αυτό και δίνονται σε στατιστικό πίνακα σε κάθε βιβλίο στατιστικής. Κάθε τ.μ. X που ακολουθεί κανονική κατανομή μπορεί να μετασχηματιστεί στη Z με τον απλό μετασχηματισμό

$$X \sim \mathrm{N}(\mu, \sigma^2) \quad \Longrightarrow \quad Z \equiv \frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1). \tag{2.26}$$

Μπορούμε λοιπόν να υπολογίσουμε οποιαδήποτε πιθανότητα για τη X από την $\Phi(z)$. Γενικά η πιθανότητα για $X \in [a,b]$ είναι

$$\mathrm{P}(a \le X \le b) = \mathrm{P}\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \tag{2.27}$$

Παράδειγμα 2.3. Το πάχος ενός κυλινδρικού σωλήνα είναι σχεδιασμένο από το εργοστάσιο να είναι μ, αλλά παρατηρείται ότι το πάχος δεν είναι σταθερό σε κάθε παραγόμενο σωλήνα αλλά αποκλίνει από το μ με τυπική απόκλιση $\sigma = 0.1\,\mathrm{mm}$. Υποθέτουμε λοιπόν ότι το πάχος του κυλινδρικού σωλήνα είναι τυχαία μεταβλητή X που ακολουθεί κανονική κατανομή, δηλαδή $X \sim \mathrm{N}(\mu, 0.1^2)$ (σε mm).

Έστω ότι θέλουμε να υπολογίσουμε την πιθανότητα η απόκλιση του πάχους του κυλινδρικού σωλήνα από το προδιαγεγραμμένο πάχος να μην είναι μεγαλύτερη από 0.1 mm. Έχουμε σύμφωνα με την (2.27)

$$\begin{aligned} \mathrm{P}(\mu - 0.1 \le X \le \mu + 0.1) &= \mathrm{P}(-1 \le Z \le 1) = \Phi(1) - \Phi(-1) \\ &= 0.8413 - 0.1587 = 0.6826, \end{aligned}$$

που συμφώνει με το αποτέλεσμα που αναφέρθηκε παραπάνω, δηλαδή ότι περίπου το 70% των τιμών της X βρίσκονται στο διάστημα $[\mu - \sigma, \mu + \sigma]$. Αντίστροφα, μπορούμε να προσδιορίσουμε ένα όριο για το σφάλμα (πάνω και κάτω από την προδιαγεγραμμένη τιμή μ) που αντιστοιχεί σε κάποια πιθανότητα, ας πούμε 0.05. Αν ονομάσουμε το σφάλμα ϵ έχουμε

$$\begin{aligned} \mathrm{P}(X \le \mu - \epsilon \text{ ή } X \ge \mu + \epsilon) &= 0.05 \Rightarrow \\ \mathrm{P}(\mu - \epsilon \le X \le \mu + \epsilon) &= 0.95 \Rightarrow \\ \Phi\left(\frac{\epsilon}{0.1}\right) - \Phi\left(-\frac{\epsilon}{0.1}\right) &= 0.95 \Rightarrow \\ 2\Phi\left(\frac{\epsilon}{0.1}\right) - 1 &= 0.95 \Rightarrow \\ \Phi\left(\frac{\epsilon}{0.1}\right) &= 0.975. \end{aligned}$$

Από τον πίνακα της τυπικής κανονικής κατανομής βρίσκουμε πως η τιμή z που αντιστοιχεί για $\Phi(z) = 0.975$ είναι $z = 1.96$. Άρα με πιθανότητα 0.95 το πάχος του κυλινδρικού σωλήνα δεν αποκλίνει από τη μέση τιμή μ περισσότερο από $0.196\,\mathrm{mm}$.

2.3.5 Κανονικότητα

Στους λόγους που αναφέρθηκαν παραπάνω για τη σπουδαιότητα της κανονικής κατανομής, θα πρέπει να προστεθεί και η ιδιότητα της κανονικής κατανομής να 'έλκει' τα αθροίσματα τυχαίων μεταβλητών, που δεν είναι απαραίτητα κανονικές, δηλαδή η κατανομή των αθροισμάτων τ.μ. να προσεγγίζει την κανονική κατανομή. Οι περιορισμοί για να ισχύει αυτό είναι οι τυχαίες μεταβλητές να είναι ανεξάρτητες μεταξύ τους, να έχουν πεπερασμένη διασπορά και το άθροισμα να είναι αρκετά μεγάλο. Κάτω από αυτές τις συνθήκες ισχύει το *κεντρικό οριακό θεώρημα, ΚΟΘ,* (central limit theorem, CLT):

Θεώρημα 2.2. Έστω οι τ.μ. X_i, $i = 1, \ldots, n$, για n μεγάλο (συνήθως θεωρούμε $n > 30$) που έχουν κατανομές με μέσες τιμές μ_i και διασπορές σ_i^2 για $i = 1, \ldots, n$, αντίστοιχα. Τότε ισχύει

$$Y = \sum_{i=1}^{n} X_i \sim \mathrm{N}(\mu_Y, \sigma_Y^2), \tag{2.28}$$

όπου η μέση τιμή της τ.μ. του αθροίσματος Y είναι $\mu_Y = \sum_{i=1}^{n} \mu_i$ και η διασπορά είναι $\sigma_Y^2 = \sum_{i=1}^{n} \sigma_i^2$.

Προφανώς όταν οι τ.μ. X_i έχουν την ίδια κατανομή με μέση τιμή μ και διασπορά σ^2, τότε ισχύει $\mu_Y = n\mu$ και $\sigma_Y^2 = n\sigma^2$.

Για το μέσο όρο των τ.μ. X_i, $i = 1, \ldots, n$, με ίδια κατανομή, από το ΚΟΘ ισχύει

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathrm{N}(\mu, \sigma^2/n), \tag{2.29}$$

δηλαδή ο μέσος όρος ακολουθεί κανονική κατανομή με την ίδια μέση τιμή όπως οι τ.μ. X_i και διασπορά $\sigma^2_{\bar{X}} = \sigma^2/n$.

Ασκήσεις Κεφαλαίου 2

1. Επιβεβαίωσε τον ορισμό της πιθανότητας ως το όριο της σχετικής συχνότητας για αριθμό επαναλήψεων να τείνει στο άπειρο. Προσομοίωσε τη ρίψη ενός νομίσματος n φορές χρησιμοποιώντας τη γενέτειρα συνάρτηση τυχαίων αριθμών, είτε από ομοιόμορφη διακριτή κατανομή (δίτιμη για 'κορώνα' και 'γράμματα'), ή από ομοιόμορφη συνεχή κατανομή στο διάστημα $[0, 1]$ χρησιμοποιώντας κατώφλι 0.5 (π.χ. αριθμός μικρότερος του 0.5 είναι 'κορώνα' και μεγαλύτερος 'γράμματα'). Επανέλαβε το πείραμα για αυξανόμενα n και υπολόγισε κάθε φορά την αναλογία των 'γραμμάτων' στις n επαναλήψεις. Κάνε την αντίστοιχη γραφική παράσταση της αναλογίας για τα διαφορετικά n.

   Βοήθεια (`matlab`*):* Για τη δημιουργία των τυχαίων αριθμών χρησιμοποίησε τη συνάρτηση `rand` ή `unidrnd`.

2. Δημιούργησε 1000 τυχαίους αριθμούς από εκθετική κατανομή με παράμετρο $\lambda = 1$ χρησιμοποιώντας την τεχνική που δίνεται στην Παρ. 2.3.3. Κάνε το ιστόγραμμα των τιμών και στο ίδιο σχήμα την καμπύλη της εκθετικής σππ $f_X(x) = \lambda e^{-\lambda x}$.

   Βοήθεια (`matlab`*):* Το ιστόγραμμα δίνεται με τη συνάρτηση `hist`.

3. Δείξε με προσομοίωση ότι όταν δύο τ.μ. X και Y δεν είναι ανεξάρτητες δεν ισχύει η ιδιότητα $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$. Για να το δείξεις θεώρησε μεγάλο πλήθος τιμών n από X και Y που ακολουθούν τη διμεταβλητή κανονική κατανομή.

   Βοήθεια (`matlab`*):* Για τον υπολογισμό της διασποράς από n παρατηρήσεις, χρησιμοποίησε τη συνάρτηση `var`. Για να δημιουργήσεις παρατηρήσεις από διμεταβλητή κανονική κατανομή χρησιμοποίησε τη συνάρτηση `mvnrnd`.

4. Ισχύει $\mathrm{E}[1/X] = 1/\mathrm{E}[X]$; Διερεύνησε το υπολογιστικά για X από ομοιόμορφη συνεχή κατανομή στο διάστημα $[1, 2]$ υπολογίζοντας τους αντίστοιχους μέσους όρους για αυξανόμενο μέγεθος επαναλήψεων n. Κάνε κατάλληλη γραφική παράσταση για τις δύο μέσες τιμές και τα διαφορετικά n. Τι συμβαίνει αν το διάστημα της ομοιόμορφης κατανομής είναι $[0, 1]$ ή $[-1, 1]$;

5. Το μήκος X των σιδηροδοκών που παράγονται από μια μηχανή, είναι γνωστό ότι κατανέμεται κανονικά $X \sim \mathrm{N}(4, 0.01)$. Στον ποιοτικό έλεγχο που ακολουθεί αμέσως μετά την παραγωγή απορρίπτονται όσοι σιδηροδοκοί έχουν μήκος λιγότερο από 3.9. Ποια είναι η πιθανότητα μια

σιδηροδοκός να καταστραφεί; Που πρέπει να μπει το όριο για να καταστρέφονται το πολύ το 1% των σιδηροδοκών;

Βοήθεια (`matlab`*):* Η αθροιστική συνάρτηση κανονικής κατανομής δίνεται με τη συνάρτηση `normcdf`. Η αντίστροφη της δίνεται με τη συνάρτηση `norminv`.

6. Δείξε ότι ισχύει το ΚΟΘ με προσομοίωση. Έστω $n = 100$ τ.μ. από ομοιόμορφη κατανομή στο διάστημα $[0, 1]$ και έστω Y η μέση τιμή τους. Υπολόγισε $N = 10000$ τιμές της Y και σχημάτισε το ιστόγραμμα των τιμών μαζί με την καμπύλη της κανονικής κατανομής.

   Βοήθεια (`matlab`*):* Το ιστόγραμμα δίνεται με τη συνάρτηση `hist`.

Κεφάλαιο 3

Στοιχεία Στατιστικής

Η στατιστική ασχολείται με τις εφαρμογές της θεωρίας των πιθανοτήτων τυχαίων μεταβλητών σε πραγματικά προβλήματα και συνίσταται στην εξαγωγή συμπερασμάτων που βασίζονται στις παρατηρήσεις. Εύκολα μπορεί να καταλάβει κάποιος τη σύνδεση πιθανολογικών εννοιών με την πραγματικότητα από την προσέγγιση της πιθανότητας που δώσαμε στην (2.1). Στην πραγματικότητα έχουμε n παρατηρήσεις που αποτελούν το δείγμα και αν η τιμή x_i μιας διακριτής τ.μ. X εμφανίζεται n_i φορές στο δείγμα, μπορούμε να προσεγγίσουμε την πιθανότητα εμφάνισης της x_i, $p = \mathrm{P}(X = x_i)$, ως

$$\hat{p} = n_i/n. \tag{3.1}$$

Η τιμή $\hat{p}$ αποτελεί την εκτίμηση της p με βάση το δείγμα. Σημειώνεται ότι το p μπορούμε να το ονομάσουμε και αναλογία εμφάνισης της τιμής x_i στο σύνολο των δυνατών τιμών της X. Η **εκτίμηση** αυτή είναι **σημειακή** (point estimation) και δίνει την καλύτερη προσέγγιση με μια τιμή που μπορούμε να δώσουμε στην πραγματική αλλά άγνωστη αναλογία p με βάση το δείγμα. Σε πολλές περιπτώσεις θα θέλαμε να εκτιμήσουμε ένα **διάστημα εμπιστοσύνης** (confidence interval) σε κάποιο επίπεδο σημαντικότητας a (ή αντίστοιχα επίπεδο εμπιστοσύνης $1 - a$) που να περιέχει την πραγματική αλλά άγνωστη αναλογία p. Σε άλλες περιπτώσεις μας ενδιαφέρει μόνο να ελέγξουμε αν η αναλογία μπορεί να πάρει ή να υπερβεί κάποια τιμή και για αυτό κάνουμε **έλεγχο υπόθεσης** (hypothesis test).

Σε πολλές μελέτες τα δεδομένα είναι αριθμητικά και το ενδιαφέρον είναι στον προσδιορισμό του κέντρου (που ορίζεται με τη μέση τιμή ή διάμεσο) της κατανομής της παρατηρούμενης τ.μ. ή της διασποράς της. Γενικά για να εκτιμήσουμε με σημειακή εκτίμηση ή διάστημα εμπιστοσύνης κάποια άγνωστη παράμετρο θ (π.χ. μέση τιμή ή διασπορά) ή να ελέγξουμε αν αυτή μπορεί να πάρει κάποια τιμή υπάρχουν τρεις προσεγγίσεις. Η **παραμετρική προσέγγιση** (parametric approach) υποθέτει ότι τα δεδομένα του προβλήματος

προέρχονται από κάποια γνωστή κατανομή. Αυτή η προσέγγιση είναι απλή στην πραγματοποίηση της. Από τη γνωστή κατανομή υπολογίζεται η κατανομή του εκτιμητή και στη συνέχεια το διάστημα εμπιστοσύνης ή σχηματίζεται η πιθανότητα της ορθότητας της μηδενικής υπόθεσης του ελέγχου. Η προσέγγιση αυτή είναι η πιο ακριβής αν η υπόθεση για την κατανομή είναι σωστή. Αντίθετα η **μη-παραμετρική** (nonparametric) προσέγγιση δεν υποθέτει κάποια γνωστή κατανομή για τα δεδομένα. Είναι λιγότερη ακριβής για γνωστές κατανομές αλλά είναι πιο κατάλληλη από την παραμετρική όταν τα δεδομένα δεν προσαρμόζονται καλά σε κάποια γνωστή κατανομή. Η τρίτη προσέγγιση επίσης δε θεωρεί γνωστή κατανομή για τα δεδομένα και χρησιμοποιεί **επαναδειγματοληψία** (resampling) για να δημιουργήσει νέα δείγματα. Από αυτά τα δείγματα υπολογίζεται ένα πλήθος τιμών του εκτιμητή, ένα για κάθε δείγμα, σχηματίζεται η κατανομή του και υπολογίζεται έτσι το διάστημα εμπιστοσύνης ή γίνεται ο έλεγχος υπόθεσης. Αυτή είναι η πιο ακριβής προσέγγιση για πραγματικά προβλήματα, όπου συνήθως τα δεδομένα δεν προσαρμόζονται πιστά σε γνωστές κατανομές, αλλά η ακρίβεια της απαιτεί πολλούς περισσότερους υπολογισμούς. Για αυτό και αυτή η προσέγγιση άρχισε να χρησιμοποιείται ευρέως τα τελευταία χρόνια με την ανάπτυξη ισχυρής υπολογιστικής τεχνολογίας.

Στο κεφάλαιο αυτό θα επικεντρωθούμε κυρίως στην πρώτη προσέγγιση για τον υπολογισμό διαστημάτων εμπιστοσύνης παραμέτρου και την πραγματοποίηση ελέγχου υπόθεσης παραμέτρου. Ειδικότερα θα μελετήσουμε τις παραμέτρους της μέσης τιμής και της διασποράς. Τέλος θα γενικεύσουμε τη χρήση του ελέγχου υπόθεσης και σε άλλα προβλήματα, όπως στην προσαρμογή της κατανομής των δεδομένων σε κάποια γνωστή κατανομή. Δε θα αναφερθούμε στη χρήση μη-παραμετρικών διαστημάτων εμπιστοσύνης αλλά στο τέλος του κεφαλαίου θα περιγράψουμε συνοπτικά την εκτίμησης διαστημάτων εμπιστοσύνης και έλεγχο υπόθεσης με μεθόδους επαναδειγματοληψίας.

3.1 Σημειακή εκτίμηση

Η σημειακή εκτίμηση μιας παραμέτρου θ είναι το στατιστικό (ή η στατιστική) $\hat{\theta}$ που υπολογίζουμε από το δείγμα για να προσδιορίσουμε την άγνωστη θ, δηλαδή είναι μια τιμή, που υπολογίζεται με βάση τα δεδομένα του δείγματος και αντιπροσωπεύει την πραγματική τιμή της αντίστοιχης παράμετρου του πληθυσμού.

Έστω X μια τ.μ. με αθροιστική συνάρτηση κατανομής $F_X(x;\theta)$ που εξαρτάται από την παράμετρο θ την οποία θέλουμε να εκτιμήσουμε (το ';' στον παραπάνω συμβολισμό ξεχωρίζει μεταβλητές από παραμέτρους). Έστω ακόμα

ότι έχουμε παρατηρήσεις $\{x_1, \ldots, x_n\}$ της X από ένα δείγμα μεγέθους n. Τότε η σημειακή εκτίμηση της ϑ δίνεται από μια *εκτιμήτρια συνάρτηση* των τιμών του δείγματος, $\hat{\vartheta} = g(x_1, \ldots, x_n)$ και το $\hat{\vartheta}$ ονομάζεται **εκτιμητής** (estimator) της ϑ.

Το βασικό (και λεπτό σημείο) στην εκτίμηση παραμέτρων είναι να καταλάβουμε ότι η $\hat{\vartheta}$ είναι τυχαία μεταβλητή. Οι παρατηρήσεις $\{x_1, \ldots, x_n\}$ παίρνουν τυχαίες τιμές σε κάθε δείγμα μεγέθους n και πάντα με την ίδια κατανομή $F_X(x; \vartheta)$. Άρα μπορούμε να θεωρήσουμε τις $\{x_1, \ldots, x_n\}$ ως τ.μ. και τότε και η $\hat{\vartheta}$ είναι τ.μ. ως συνάρτηση n τ.μ.. Βέβαια για ένα συγκεκριμένο δείγμα οι $\{x_1, \ldots, x_n\}$ παίρνουν πραγματικές τιμές που δίνουν την τιμή της $\hat{\vartheta}$ για αυτό το δείγμα.

Ως τ.μ. η $\hat{\vartheta}$ ακολουθεί κάποια κατανομή με μέση τιμή $\mu_{\hat{\vartheta}} \equiv \mathrm{E}[\hat{\vartheta}]$ και διασπορά $\sigma^2_{\hat{\vartheta}} \equiv \mathrm{Var}[\hat{\vartheta}]$.

3.1.1 Μέση τιμή και διασπορά

Δύο σημαντικές παράμετροι μιας τ.μ. X που θέλουμε να εκτιμήσουμε είναι η μέση τιμή της μ και η διασπορά της σ^2. Ο εκτιμητής της μ δίνεται από το γνωστό μέσο όρο των $\{x_1, \ldots, x_n\}$ και ονομάζεται *δειγματική μέση τιμή* ή απλά *μέσος όρος* $\bar{x}$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{3.2}$$

Ο εκτιμητής της διασποράς σ^2 είναι η *δειγματική διασπορά* s^2

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

Ένας άλλος εκτιμητής της σ^2 δίνεται ως

$$\tilde{s}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{3.4}$$

Οι εκτιμητές 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 εκτιμητές συγκλίνουν στην ίδια τιμή.

3.1.2 Βαθμοί ελευθερίας

Η χρήση του $n-1$ στον τύπο της διασποράς στην (3.3) είναι σε συμφωνία με τους *βαθμούς ελευθερίας* (degrees of freedom) του προβλήματος εκτίμησης της διασποράς. Οι βαθμοί ελευθερίας δηλώνουν τις ελεύθερες (τυχαίες)

τιμές που υπάρχουν στο πρόβλημα που μελετάμε. Εδώ αρχικά οι βαθμοί ελευθερίας είναι n, όσες και οι παρατηρήσεις στο δείγμα, αλλά επειδή στον ορισμό της δειγματικής διασποράς περιλαμβάνεται η δειγματική μέση τιμή δεσμεύονται οι n ελεύθερες τιμές με την συνθήκη να ικανοποιούν την εξίσωση (3.2), δηλαδή να δίνουν την $\bar{x}$. Έτσι χάνεται ένας βαθμός ελευθερίας και οι βαθμοί έλευθερίας είναι $n-1$.

3.1.3 Κριτήρια καλών εκτιμητών

Παραπάνω ορίσαμε κάπως αυθαίρετα τους εκτιμητές της μέσης τιμής μ και της διασποράς σ^2 χωρίς να γνωρίζουμε αν είναι 'καλοί' εκτιμητές ή όχι. Γενικά για τον ορισμό βέλτιστου εκτιμητή $\hat{\theta}$ κάποιας παραμέτρου θ θέτουμε κάποια βασικά κριτήρια που αποτελούν και ιδιότητες του εκτιμητή $\hat{\theta}$. Επικεντρώνουμε την προσοχή μας σε δύο βασικές ιδιότητες ενός εκτιμητή, όπου η πρώτη έχει να κάνει με τη μέση τιμή του $\mu_{\hat{\theta}}$ και η δεύτερη με τη διασπορά του $\sigma^2_{\hat{\theta}}$.

Ο εκτιμητής $\hat{\theta}$ είναι **αμερόληπτος** (unbiased) αν η μέση τιμή του είναι ίση με την παράμετρο θ, δηλαδή αν ισχύει

$$\mathrm{E}[\hat{\theta}] = \theta.$$

Αλλιώς λέγεται μεροληπτικός με μεροληψία

$$b(\hat{\theta}) = \mathrm{E}[\hat{\theta}] - \theta.$$

Ένα δεύτερο σημαντικό κριτήριο καλού εκτιμητή είναι η **αποτελεσματικότητα** (efficiency) αναφέρεται στη διασπορά του εκτιμητή και δίνεται συγκριτικά. Ένας εκτιμητής $\hat{\theta}_1$ της θ είναι πιο αποτελεσματικός (effective) από έναν άλλο εκτιμητή $\hat{\theta}_2$ αν έχει μικρότερη διασπορά, αν δηλαδή ισχύει $\sigma^2_{\hat{\theta}_1} < \sigma^2_{\hat{\theta}_2}$.

Οι εκτιμητές $\bar{x}$ για την παραμέτρο μ και s^2 για την παράμετρο σ^2, που ορίσαμε αυθαίρετα, είναι και οι δύο αμερόληπτοι και οι πιο αποτελεσματικοί (δες επίσης 3.2.1). Ο εκτιμητής $\tilde{s}^2$ είναι μεροληπτικός εκτιμητής της σ^2 με μεροληψία $b(\tilde{s}^2) = -\sigma^2/n$. Είναι κατανοητό ότι καθώς το μέγεθος του δείγματος αυξάνει η μεροληψία τείνει προς το μηδέν και για αυτό θεωρούμε πως ο εκτιμητής $\tilde{s}^2$ είναι *ασυμπτωτικά* (asymptotic) αμερόληπτος.

Σε κάποια προβλήματα είναι δύσκολο ή αδύνατο να βρούμε εκτιμητή που είναι και αμερόληπτος και ο πιο αποτελεσματικός. Για αυτό συνθέτουμε το κριτήριο του **μέσου τετραγωνικού σφάλματος** (mean square error) της εκτίμησης ως το άθροισμα της διασποράς του εκτιμητή και του τετραγώνου της μεροληψίας

$$\mathrm{MSE}[\hat{\theta}] = b(\hat{\theta})^2 + \sigma^2_{\hat{\theta}} = (\mathrm{E}[\hat{\theta}] - \theta)^2 + \mathrm{E}[\hat{\theta}^2] - (\mathrm{E}[\hat{\theta}])^2 = \mathrm{E}[(\hat{\theta} - \theta)^2]. \quad (3.5)$$

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

3.1.4 Μέθοδος της μέγιστης πιθανοφάνειας

Η σημαντικότερη μέθοδος της στατιστικής για την εκτίμηση παραμέτρων είναι η μέθοδος της **μέγιστης πιθανοφάνειας** (maximum likelihood). 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 δείγμα.

Υποθέτουμε ότι η τ.μ. X έχει κάποια γνωστή κατανομή, δηλαδή γνωρίζουμε τη γενική μορφή της ασκ $F_X(x;\theta)$ και της σππ $f_X(x;\theta)$. Η παράμετρος θ της κατανομής είναι άγνωστη και θέλουμε να την εκτιμήσουμε από ένα δείγμα ανεξάρτητων παρατηρήσεων $\{x_1,\dots,x_n\}$.

Επειδή $\{x_1,\dots,x_n\}$ είναι ανεξάρτητες τ.μ., η πιθανότητα να τις παρατηρήσουμε σε ένα τυχαίο δείγμα μεγέθους n δίνεται από τη **συνάρτηση πιθανόφανειας** (likelihood function) ως προς θ

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

Αν λοιπόν $L(x_1,\dots,x_n;\theta_1) > L(x_1,\dots,x_n;\theta_2)$ για δύο τιμές θ_1 και θ_2 της θ, τότε η τιμή θ_1 είναι πιο αληθοφανής από τη θ_2 γιατί δίνει μεγαλύτερη πιθανότητα να παρατηρήσουμε το συγκεκριμένο δείγμα των $\{x_1,\dots,x_n\}$. Θέλουμε λοιπόν να βρούμε την 'πιο αληθοφανή' τιμή της θ, δηλαδή την τιμή $\hat{\theta}$ που μεγιστοποιεί τη $L(x_1,\dots,x_n;\theta)$ ή καλύτερα (για ευκολότερους υπολογισμούς) τη $\log L(x_1,\dots,x_n;\theta)$. Άρα ο **εκτιμητής μέγιστης πιθανοφάνειας** (maximum likelihood estimator) $\hat{\theta}$ βρίσκεται από τη σχέση

$$\frac{\partial \log L(x_1,\dots,x_n;\theta)}{\partial\theta} = 0. \tag{3.6}$$

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

$$\frac{\partial \log L(x_1,\dots,x_n;\theta_1,\dots,\theta_m)}{\partial\theta_j} = 0 \quad \text{για} \quad j = 1,\dots,m. \tag{3.7}$$

Παράδειγμα 3.1. Έχουμε ένα τυχαίο δείγμα $\{x_1,\dots,x_n\}$ από κανονική κατανομή $\mathrm{N}(\mu,\sigma^2)$ και θέλουμε να εκτιμήσουμε τη μέση τιμή μ θεωρώντας τη σ^2 γνωστή. Η συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής

δίνεται από την (2.23). Η συνάρτηση πιθανόφανειας (για την οποία μόνο η παράμετρος μ είναι άγνωστη) είναι

$$L(x_1, \ldots, x_n; \mu) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right],$$

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

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

Ο εκτιμητής μέγιστης πιθανοφάνειας $\hat{\mu}$ βρίσκεται μηδενίζοντας την παράγωγο της $\log L$

$$\frac{\partial \log L}{\partial \mu} = 0 \quad \Rightarrow \quad \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \tag{3.8}$$

που δίνει τη λύση

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

δηλαδή είναι ίδιος με τον εκτιμητή $\bar{x}$ της μέσης τιμής μ που ορίσαμε για οποιαδήποτε κατανομή της τ.μ. X.

Παράδειγμα 3.2. Ας υποθέσουμε στο προηγούμενο παράδειγμα πως και η διασπορά σ^2 είναι άγνωστη. Τότε στην παραπάνω εξίσωση (3.8) προστίθεται και η εξίσωση

$$\frac{\partial \log L}{\partial \sigma^2} = 0 \quad \Rightarrow \quad -\frac{n}{2\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0. \tag{3.9}$$

Η επίλυση του συσήματος των εξισώσεων (3.8) και (3.9) δίνει την ίδια λύση για τη μ και για τη σ^2 έχουμε

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

Οι εκτιμητές μέγιστης πιθανοφάνειας λοιπόν για τη μέση τιμή μ και τη διασπορά σ^2 μιας τ.μ. που ακολουθεί κανονική κατανομή είναι απλά η δειγματική μέση τιμή και η δειγματική διασπορά αντίστοιχα, αλλά για τη διασπορά έχουμε τον ασυμπτωτικά αμερόληπτο εκτιμητή $\tilde{s}^2$ (σχέση (3.4)).

Η μέθοδος μέγιστης πιθανοφάνειας είναι η καλύτερη μέθοδος εκτίμησης αν γνωρίζουμε την κατανομή της τ.μ. X και μπορεί να εφαρμοσθεί σε οποιοδήποτε πρόβλημα εκτίμησης παραμέτρων από δείγμα ανεξάρτητων παρατηρήσεων.

3.2 Εκτίμηση διαστήματος εμπιστοσύνης

Η σημειακή εκτίμηση $\hat{\theta}$ από κάποιο δείγμα δεν περιέχει καμιά πληροφορία για την ακρίβεια της εκτίμησης της θ. Η εκτίμηση $\hat{\theta}$ που παίρνουμε από ένα δείγμα είναι μια τιμή που δε γνωρίζουμε πόσο κοντά είναι στην πραγματική τιμή της θ και επίσης η τιμή αυτή αλλάζει με το δείγμα. Για παράδειγμα, υπολογίζουμε τη δειγματική μέση τιμή $\bar{x}$ από ένα τυχαίο δείγμα μεγέθους n. Αν 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 μ απ' ότι αυτή από το προηγούμενο δείγμα. Γι αυτό στην εκτίμηση της θ είναι σημαντικό εκτός από τη σημειακή εκτίμηση $\hat{\theta}$ να υπολογίσουμε και διάστημα $[\theta_1, \theta_2]$ που να μπορούμε να πούμε με κάποια πιθανότητα $1-a$ ότι θα περιέχει την πραγματική τιμή της παραμέτρου θ (δηλαδή η πιθανότητα σφάλματος είναι a).

Στη συνέχεια θα παρουσιάσουμε την παραμετρική διαδικασία υπολογισμού διαστημάτων εμπιστοσύνης χρησιμοποιώντας ως παράμετρο θ τη μέση τιμή μ και τη διασπορά σ^2. Θεωρούμε και πάλι πως οι παρατηρήσεις $x_1, \ldots, x_n$ είναι ανεξάρτητες. Για τον υπολογισμό του διαστήματος εμπιστοσύνης θα πρέπει να γνωρίζουμε την κατανομή του εκτιμητή (π.χ. $\bar{x}$ ή s^2) και με βάση αυτήν την κατανομή ορίζεται το διάστημα εμπιστοσύνης για την παράμετρο.

3.2.1 Διάστημα εμπιστοσύνης της μέσης τιμής μ

Το διάστημα εμπιστοσύνης της μ υπολογίζεται με βάση την κατανομή της τ.μ. $\bar{x}$, που είναι ο καλύτερος εκτιμητής της μ. Πράγματι στην παράγραφο 3.1.3 αναφέρθηκε ότι η $\bar{x}$ είναι αμερόληπτος εκτιμητής της μ, δηλαδή ισχύει

$$\mu_{\bar{x}} \equiv \mathrm{E}[\bar{x}] = \mu \tag{3.10}$$

και άρα η μέση τιμή της $\bar{x}$ είναι το ίδιο το μ. Η διασπορά της $\bar{x}$ είναι

$$\sigma^2_{\bar{x}} \equiv \mathrm{Var}[\bar{x}] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[x_i] = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}. \tag{3.11}$$

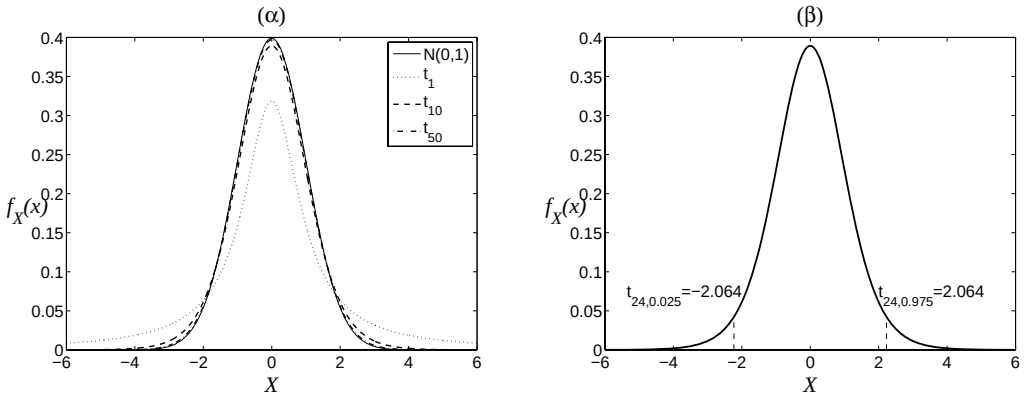
και άρα η τυπική απόκλιση της είναι $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ και ονομάζεται **τυπικό σφάλμα** (standard error) του εκτιμητή $\bar{x}$. Η μορφή της κατανομής της $\bar{x}$ εξαρτάται από το μέγεθος του δείγματος n, από το αν η κατανομή της X είναι κανονική και επίσης από το αν γνωρίζουμε τη διασπορά της.

Αν η κατανομή της τ.μ. X είναι κανονική, $X \sim N(\mu, \sigma^2)$, τότε οι παρατηρήσεις $x_1, \ldots, x_n$ ακολουθούν την ίδια κατανομή και άρα ο μέσος όρος τους $\bar{x}$ ακολουθεί επίσης κανονική κατανομή και ισχύει $\bar{x} \sim N(\mu, \sigma^2/n)$. Από το ΚΟΘ στην παράγραφο 2.3.5 το ίδιο ισχύει ακόμα και όταν δε γνωρίζουμε την

κατανομή της τ.μ. X αλλά το δείγμα είναι μεγάλο ($n > 30$). Στην πράξη όμως δε γνωρίζουμε τη διασπορά σ^2 της τ.μ. X. Αν θεωρήσουμε αντί για σ^2 τη δειγματική διασπορά s^2, η υπόθεση $\bar{x} \sim N(\mu, s^2/n)$ δεν είναι πλέον ακριβής αλλά η $\bar{x}$ ακολουθεί μια άλλη κατανομή, επίσης συμμετρική και με κωνοειδές σχήμα, αλλά με πιο παχιές ουρές από την κανονική. Ειδικότερα η κανονικοποίηση της $\bar{x}$, $(\bar{x} - \mu)/(s/\sqrt{n})$, ακολουθεί την **κατανομή Student** ή t-κατανομή με $n - 1$ βαθμούς ελευθερίας

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}. \tag{3.12}$$

Οι βαθμοί ελευθερίας είναι $n - 1$ γιατί υπάρχουν n ελεύθερες παρατηρήσεις στο δείγμα που δεσμεύονται με μια συνθήκη, να δίνουν την s^2. Στο Σχήμα 3.1α δίνεται το γράφημα της σππ της κατανομής Student για διαφορετικούς βαθμούς ελευθερίας. Για ένα βαθμό ελευθερίας (δείγμα δύο μόνο παρατηρήσεων)



Σχήμα 3.1: (α) Η σππ της τυπικής κανονικής κατανομής και της Student για βαθμούς ελευθερίας όπως δίνονται στο ένθετο. (β) Η σππ της Student για 24 βαθμούς ελευθερίας και οι κρίσιμες τιμές για $a = 0.05$.

η κατανομή Student διαφέρει φανερά από την τυπική κανονική κατανομή και έχει πολύ παχιές ουρές. Αυτή η κατανομή είναι γνωστή με πολλά ονόματα (**Cauchy, Lorentzian, Breit-Wigner**) και χρησιμοποιείται συχνά στη φυσική, όπως στη φυσική υψηλής ενέργειας ως μοντέλο σππ για την ενέργεια που εμφανίζεται συντονισμός. Καθώς οι βαθμοί ελευθερίας πληθαίνουν η κατανομή Student συγκλίνει στην τυπική κανονική κατανομή. Πρακτικά για μεγάλα δείγματα δεν υπάρχει διαφορά μεταξύ της κατανομής Student και της τυπικής κανονικής κατανομής.

Στην Παράγραφο 2.3.4 είχε δειχθεί ότι για την τυπική κανονική κατανομή σε κάθε πιθανότητα, έστω $1 - a$, αντιστοιχεί ένα διάστημα τιμών της z

συμμετρικό ως προς το 0, $[-z_{1-a/2}, z_{1-a/2}]$, έτσι ώστε

$$P(-z_{1-a/2} < z \leq z_{1-a/2}) = \Phi(z_{1-a/2}) - \Phi(-z_{1-a/2}) = 1 - a,$$

όπου $\Phi(z)$ είναι η ασκ της τυπικής κανονικής κατανομής. Αντίστοιχα για την κατανομή Student με $n-1$ βαθμούς ελευθερίας σε κάθε πιθανότητα $1-a$ αντιστοιχεί ένα διάστημα τιμών της t, $[-t_{n-1,1-a/2}, t_{n-1,1-a/2}]$, έτσι ώστε ίσχύει

$$P(-t_{n-1,1-a/2} < t \leq t_{n-1,1-a/2}) = 1 - a, \tag{3.13}$$

Για παράδειγμα για $a = 0.05$ και 24 βαθμούς ελευθερίας το διάστημα $[-2.064, 2.064]$ περιέχει την τ.μ. t με πιθανότητα 0.95, όπου $t_{24,0.975} = 2.064$, όπως φαίνεται στο Σχήμα 3.1β. Η τιμή που ορίζει την ουρά της κατανομής λέγεται και *κρίσιμη τιμή* (critical value) και παραδοσιακά δίνεται από κατάλληλο στατιστικό πίνακα, αλλά μπορεί να υπολογισθεί εύκολα (π.χ. στο `matlab` η κρίσιμη τιμή για την κατανομή Student δίνεται από τη συνάρτηση `tinv`).

Για να βρούμε το διάστημα που με πιθανότητα $1-a$ περιέχει την παραμέτρο μ αντικαθιστούμε το $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ στη σχέση (3.13) και λύνουμε τις ανισότητες ως προς μ

$$P(\bar{x} - t_{n-1,1-a/2}\frac{s}{\sqrt{n}} < \mu \leq \bar{x} - t_{n-1,1-a/2}\frac{s}{\sqrt{n}}) = 1 - a. \tag{3.14}$$

Η παραπάνω σχέση ορίζει πως το διάστημα

$$\bar{x} \pm t_{n-1,1-a/2}\frac{s}{\sqrt{n}} \quad \text{ή} \quad \left[\bar{x} - t_{n-1,1-a/2}\frac{s}{\sqrt{n}}, \quad \bar{x} + t_{n-1,1-a/2}\frac{s}{\sqrt{n}}\right] \tag{3.15}$$

περιέχει την πραγματική μέση τιμή μ για κάποια δοθείσα πιθανότητα $1-a$ που είναι το προκαθορισμένο **επίπεδο (ή στάθμη) εμπιστοσύνης** (confidence level) και λέγεται **διάστημα εμπιστοσύνης** (confidence interval) της μ σε επίπεδο εμπιστοσύνης $1-a$.

Όταν το δείγμα είναι μεγάλο η κατανομή Student συγκλίνει στην τυπική κανονική κατανομή και άρα στο διάστημα εμπιστοσύνης η κρίσιμη τιμή από την κατανομή Student $t_{n-1,1-a/2}$ μπορεί να αντικατασταθεί από αυτήν της τυπικής κανονικής $z_{1-a/2}$. Αν το δείγμα είναι μικρό και δε μπορούμε να υποθέσουμε πως η κατανομή της τ.μ. X είναι κανονική, π.χ. γιατί φαίνεται από ένα ιστόγραμμα να είναι ισχυρά ασύμμετρη (λοξή), τότε δε μπορούμε να χρησιμοποιήσουμε το παραμετρικό διάστημα εμπιστοσύνης στην (3.15) για τη μέση τιμή μ. Σε αυτήν την περίπτωση θα πρέπει να καταφύγουμε στη μη-παραμετρική προσέγγιση. Η πιο συχνή μη-παραμετρική μέθοδος χρησιμοποιεί τις *τάξεις* (ranks) των δεδομένων, δηλαδή τη σειρά τους όταν αυτά κατατάσσονται σε αύξουσα σειρά.

Παράδειγμα 3.3. Θέλουμε να μελετήσουμε κατά πόσο ασφάλειες ενός τύπου καίγονται σε ένταση ρεύματος 40 αμπέρ όπως είναι η ένδειξη τους. Στον Πίνακα 3.1 δίνονται οι μετρήσεις έντασης του ηλεκτρικού ρεύματος στις οποίες κάηκαν 25 ασφάλειες που δοκιμάσαμε. Η τ.μ. X που μας ενδιαφέρει

40.9	40.3	39.8	40.1	39.0	41.4	39.8	41.5	40.0	40.6
38.3	39.0	40.9	39.1	40.3	39.3	39.6	38.4	38.4	40.7
39.7	38.9	38.9	40.6	39.6					

Πίνακας 3.1: Δεδομένα ορίου έντασης ηλεκτρικού ρεύματος που κάηκαν 25 ασφάλειες των 40 αμπέρ.

είναι το όριο έντασης ηλεκτρικού ρεύματος που καίγονται ασφάλειας των 40 αμπέρ. Η δειγματική μέση τιμή της είναι

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

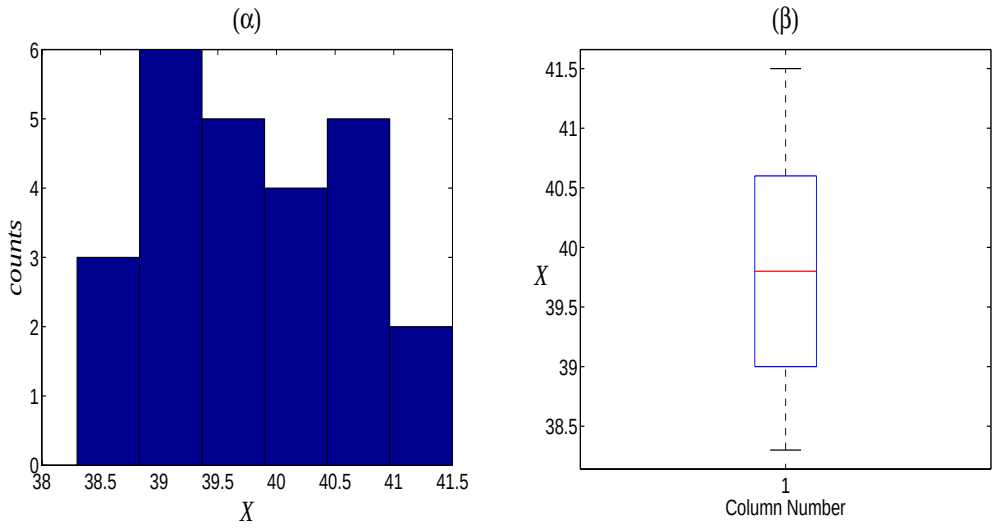
και η δειγματική διασπορά της είναι

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

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

Έστω τώρα ότι θέλουμε να εκτιμήσουμε διάστημα εμπιστοσύνης σε επίπεδο 95% για το μέσο όριο έντασης ηλεκτρικού ρεύματος για ασφάλειες των 40 αμπέρ. Το δείγμα είναι μικρό ($n = 25 < 30$). Εξετάζουμε τη δειγματική κατανομή του ορίου έντασης ηλεκτρικού ρεύματος από τα δεδομένα μας. Για αυτό σχεδιάζουμε το ιστόγραμμα και το θηκόγραμμα των δεδομένων του Πίνακα 3.1, τα οποία παρουσιάζονται στο Σχήμα 3.2.

Το **θηκόγραμμα** (boxplot) είναι η απεικόνιση των τεταρτομορίων των δεδομένων, και δίνεται από τους παρακάτω 5 αριθμούς: τη μικρότερη παρατήρηση, την παρατήρηση με τάξη στο 25% των παρατηρήσεων (πρώτο τεταρτομόριο), τη διάμεσο, την παρατήρηση με τάξη στο 75% των παρατηρήσεων (τρίτο τεταρτομόριο) και τη μεγαλύτερη τιμή. Στο θηκόγραμμα, η θήκη (κουτί) διαγράφεται μεταξύ του πρώτου και τρίτου τεταρτομορίου και οι μύστακες ενώνουν τα τεταρτομόρια με τα άκρα (ή διακόπτεται η γραμμή και οι ακραίες τιμές δηλώνονται με κάποιο σύμβολο όταν είναι απόμακρες). Τέλος σχηματίζεται μια γράμμη στη θήκη στη θέση της διαμέσου. Από το θηκόγραμμα μπορούμε να κρίνουμε τη συμμετρία της κατανομής που αναφέρονται τα δεδομένα (αν η γραμμή της διαμέσου βρίσκεται προς το κέντρο της θήκης και



Σχήμα 3.2: Ιστόγραμμα στο (α) και θηκόγραμμα στο (β) των δεδομένων του ορίου έντασης ηλεκτρικού ρεύματος για ασφάλειες των 40 αμπέρ του Πίνακα 3.1.

οι μύστακες έχουν περίπου στο ίδιο μήκος) και κατ' επέκταση αν η κατανομή είναι κανονική.

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). Άρα μπορούμε να υποθέσουμε ότι το όριο έντασης ηλεκτρικού ρεύματος X ακολουθεί κανονική κατανομή και μπορούμε να χρησιμοποιήσουμε το διάστημα εμπιστοσύνης από την κατανομή Student που δίνεται στην (3.15). Βρίσκουμε την κρίσιμη τιμή $t_{n-1,1-a/2}$ για $1 - a/2 = 0.975$ και $n - 1 = 24$, $t_{24,0.975} = 2.064$ (από το στατιστικό πίνακα για την κατανομή Student ή υπολογίζοντας απευθείας την αντίστροφη ασκ σε κάποιο πρόγραμμα όπως το `matlab`, δες επίσης Σχήμα 3.1β). Το διάστημα εμπιστοσύνης για τη μ είναι

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

Με βάση το παραπάνω 95% διάστημα εμπιστοσύνης μπορούμε να πούμε ότι η σημειακή εκτίμηση $\bar{x} = 39.80$ είναι αρκετά ακριβής αφού το αντίστοιχο 95% διάστημα εμπιστοσύνης είναι αρκετά μικρό. Επίσης παρατηρούμε ότι το 95% διάστημα εμπιστοσύνης περιέχει το 40, δηλαδή με 95% εμπιστοσύνη μπορούμε να συμπεράνουμε ότι κατά 'μέσο όρο' οι ασφάλειες διασφαλίζουν την ένδειξη τους και καίγονται πράγματι σε ένταση ηλεκτρικού ρεύματος 40 αμπέρ.

3.2.2 Διάστημα εμπιστοσύνης της διασποράς σ^2

Για να ορίσουμε διάστημα εμπιστοσύνης για τη μέση τιμή μ κανονικοποιήσαμε τον εκτιμητή $\bar{x}$ ως $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ ώστε να ακολουθεί κατανομή Student. Το ίδιο θα κάνουμε και εδώ για τον εκτιμητή s^2 της διασποράς σ^2. Η κανονικοποίηση είναι $(n-1)s^2/\sigma^2$ που ακολουθεί την κατανομή $\mathcal{X}^2$ με $n-1$ βαθμούς ελευθερίας. Για να δείξουμε αυτό το αποτέλεσμα ας δούμε πρώτα πως προκύπτει η κατανομή αυτή.

Η κατανομή $\mathcal{X}^2$ χρησιμοποιείται σε πολλά προβλήματα στατιστικής συμπερασματολογίας. Έστω ότι η τ.μ. χ^2 είναι το άθροισμα τετραγώνων των διαφορών μεταξύ παρατηρούμενων και προσδοκώμενων ή μέσων τιμών διαιρούμενο με τη διασπορά τους. Είναι φανερό από τον ορισμό της ότι η χ^2 εξαρτάται από το πλήθος των παρατηρήσεων n. Ειδικότερα όταν οι παρατηρήσεις προέρχονται από την ίδια κατανομή η χ^2 ορίζεται ως

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2} \tag{3.16}$$

και ακολουθεί κατανομή $\mathcal{X}^2$ με $n-1$ βαθμούς ελευθερίας (ένας βαθμός ελευθερίας χάνεται λόγω της συνθήκης της δειγματικής μέσης τιμής $\bar{x}$). Συνδυάζοντας της σχέση (3.16) με τη σχέση (3.3) που ορίζει τη δειγματική διασπορά s^2 έχουμε το ζητούμενο αποτέλεσμα

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \mathcal{X}^2_{n-1}. \tag{3.17}$$

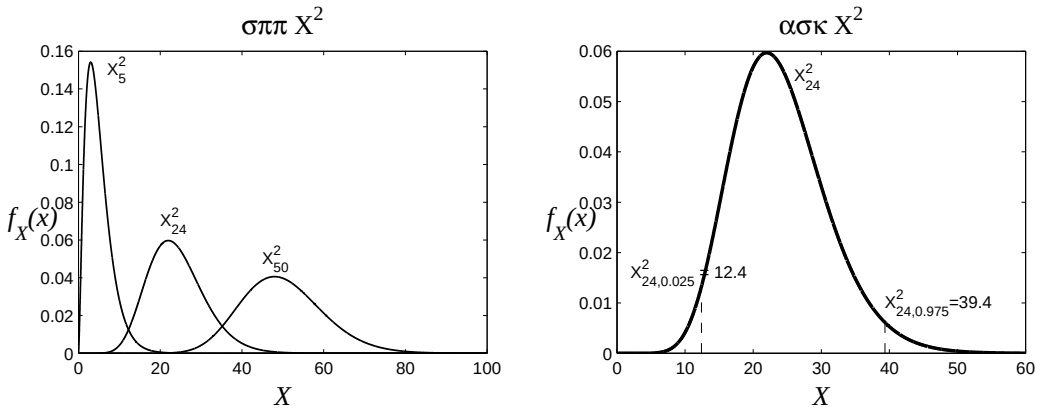
Στο Σχήμα 3.3α δίνεται το γράφημα της σππ της κατανομής $\mathcal{X}^2$ για διαφορετικούς βαθμούς ελευθερίας. Για λίγους βαθμούς ελευθερίας η κατανομή $\mathcal{X}^2$ είναι φανερά ασύμμετρη αλλά καθώς οι βαθμοί ελευθερίας πληθαίνουν γίνεται πιο συμμετρική και κωνοειδής. Για πολλούς βαθμούς ελευθερίας προσεγγίζει την κανονική κατανομή. Γενικά επειδή η κατανομή $\mathcal{X}^2$ δεν είναι συμμετρική γύρω από το 0, οι ουρές της ορίζονται με δύο κρίσιμες τιμές, την αριστερή κρίσιμη τιμή $\chi^2_{n-1,a/2}$ και τη δεξιά κρίσιμη τιμή $\chi^2_{n-1,1-a/2}$ για κάποιο επίπεδο σημαντικότητας a. Στο Σχήμα 3.3β φαίνονται οι ουρές για $n = 25$ και $a = 0.05$.

Η πιθανότητα $1-a$ η τ.μ. χ^2 να βρίσκεται μεταξύ των ορίων που δίνονται από την αριστερή και δεξιά κρίσιμη τιμή είναι

$$\mathrm{P}(\chi^2_{n-1,a/2} < \chi^2 < \chi^2_{n-1,1-a/2}) = 1 - a. \tag{3.18}$$

Αντικαθιστώντας $\chi^2 = (n-1)s^2/\sigma^2$ και λύνοντας τις δύο ανισότητες ως προς σ^2 στη σχέση (3.18) βρίσκουμε το $(1-a)\%$ διάστημα εμπιστοσύνης για τη σ^2

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1,1-a/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,a/2}}\right]. \tag{3.19}$$



Σχήμα 3.3: (α) Η σππ της κατανομής $\mathcal{X}^2$ για βαθμούς ελευθερίας όπως δίνονται στο σχήμα. (β) Η σππ της κατανομής $\mathcal{X}^2$ για 24 βαθμούς ελευθερίας και οι κρίσιμες τιμές για $a = 0.05$.

Το 95% δ.ε. για την τυπική απόκλιση σ έχει ως άκρα τις τετραγωνικές ρίζες των αντίστοιχων άκρων του 95% δ.ε. για τη διασπορά σ^2.

Παράδειγμα 3.4. Από τα δεδομένα για το όριο έντασης ηλεκτρικού ρεύματος ασφάλειας που χρησιμοποιήθηκαν στο Παράδειγμα 3.3 θέλουμε να εκτιμήσουμε τη διασπορά σ^2 του ορίου έντασης. Η σημειακή εκτίμηση βρέθηκε να είναι $s^2 = 0.854\,(\text{αμπέρ})^2$. Για $n-1 = 24$ και $a = 0.05$ από τον στατιστικό πίνακα για τη $\mathcal{X}^2$ (ή με απευθείας υπολογισμό στο `matlab` με τη συνάρτηση `chi2inv`) βρίσκουμε $\chi^2_{24,\,0.025} = 12.4$ και $\chi^2_{24,\,0.975} = 39.4$ (δες επίσης Σχήμα 3.3β). Το 95% δ.ε. για τη διασπορά σ^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 σ του ορίου ηλεκτρικού ρεύματος είναι

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

που δείχνει ότι οι ασφάλειες δεν καίγονται με μεγάλη ακρίβεια γύρω από το όριο των 40 αμπέρ αλλά με μια τυπική απόκλιση που περιμένουμε να κυμαίνεται μεταξύ 0.7 και 1.3 αμπέρ. Αυτό το συμπέρασμα φαίνεται να είναι σε αντίθεση με το συμπέρασμα για την καλή ακρίβεια εκτίμησης του μέσου ορίου ηλεκτρικού ρεύματος μ που βρήκαμε στο Παράδειγμα 3.3. Εδώ θα πρέπει να τονισθεί ότι η κατανομή του εκτιμητή $\bar{x}$ της μέσης τιμής είναι πολύ πιο 'στενή' από την κατανομή της ίδιας της τ.μ. X και μάλιστα η διασπορά της όπως εκτιμάται από το δείγμα είναι s^2/n δηλαδή η τυπική της απόκλιση (το τυπικό σφάλμα) είναι $\sqrt{n}$ φορές μικρότερη από την τυπική απόκλιση

της X. Μια τυπική μέτρηση του ορίου έντασης ηλεκτρικού ρεύματος που καίγεται η ασφάλεια των 40 αμπέρ με βάση το δείγμα των 25 παρατηρήσεων, θα περιμέναμε να κυμαίνεται στο διάστημα $\bar{x} \pm s = 39.80 \pm \sqrt{0.854} = 39.80 \pm 0.925$. Επίσης το 95 % των μετρήσεων του ορίου ηλεκτρικού ρεύματος θα περιμέναμε να βρίσκεται στο διάστημα

$$\bar{x} \pm t_{n-1,1-a/2} s. \tag{3.20}$$

δηλαδή $39.80 \pm 2.064 \cdot 0.925 = 39.80 \pm 1.763$, που δηλώνει ότι θα εμφανίζονται ασφάλειες με σοβαρές αποκλίσεις από την ονομαστική ένδειξη της ασφάλειας.

3.3 Έλεγχος υπόθεσης

Σε πολλά προβλήματα δεν ενδιαφερόμαστε να εκτιμήσουμε με κάποια ακρίβεια την τιμή της παραμέτρου αλλά να διαπιστώσουμε αν η παραμέτρος είναι μικρότερη ή μεγαλύτερη από μια δεδομένη τιμή που έχει φυσική σημασία για το πρόβλημα μας. Στο προηγούμενο παράδειγμα μπορεί να μας ενδιαφέρει να ελέγξουμε αν το μέσο όριο έντασης ηλεκτρικού ρεύματος που καίγονται οι ασφάλειες των 40 αμπέρ διαφέρει σημαντικά από 40 αμπέρ. Σε όργανα μέτρησης μας ενδιαφέρει να ελέγξουμε αν η μέτρηση είναι σωστή και δεν υπάρχει συστηματικό λάθος στο όργανο μέτρησης. Βέβαια την απάντηση σε τέτοια ερωτήματα μπορεί να τη δώσει η εκτίμηση κατάλληλου διαστήματος εμπιστοσύνης ελέγχοντας αν η δεδομένη τιμή ανήκει σ' αυτό το διάστημα ή όχι, αλλά εδώ θα δούμε μια διαφορετική προσέγγιση, θέτοντας κατάλληλη στατιστική υπόθεση και ελέγχοντας αν είναι αποδεκτή ή όχι.

Ο **έλεγχος υπόθεσης** (hypothesis testing) επεξεργάζεται στατιστικά εγαλεία (τον εκτιμητή και την κατανομή του) σε μια διαδικασία λήψης απόφασης. Για τη διαδικασία ελέγχου μιας στατιστικής υπόθεσης πρώτα ορίζουμε τη στατιστική υπόθεση, μετά υπολογίζουμε το στατιστικό ελέγχου και την περιοχή απόρριψης και τέλος αποφασίζουμε για την υπόθεση με βάση την ένδειξη που έχουμε από το δείγμα.

Η *στατιστική υπόθεση* (statistical hypothesis) μπορεί να είναι μια οποιαδήποτε 'στατιστική' δήλωση ή πρόταση που θέτουμε υπό έλεγχο με βάση τις παρατηρήσεις. Στην αρχή θα μελετήσουμε υποθέσεις για την τιμή μιας παραμέτρου και στη συνέχεια για την κατανομή μιας τ.μ.. Η **μηδενική υπόθεση** (null hypothesis) την οποία θέτουμε υπό έλεγχο συμβολίζεται H_0 ενώ η **εναλλακτική υπόθεση** (alternative hypothesis) την οποία δεχόμαστε αν απορρίψουμε τη H_0 συμβολίζεται H_1. Οι δυνατές αποφάσεις του ελέγχου είναι:

1. *Σωστή απόφαση*: Αποδεχόμαστε την H_0 όταν η H_0 είναι σωστή. Η πιθανότητα αυτής της απόφασης είναι

$$P(\text{αποδοχή της } H_0 \mid H_0 \text{ σωστή}) = 1 - a.$$

2. **Σφάλμα τύπου II** (type II error): Αποδεχόμαστε την H_0 όταν η H_0 είναι λανθασμένη. Η πιθανότητα αυτού του σφάλματος είναι

$$P(\text{αποδοχή της } H_0 \mid H_0 \text{ λανθασμένη}) = \beta.$$

3. **Σφάλμα τύπου I** (type I error): Απορρίπτουμε την H_0 όταν η H_0 είναι σωστή. Η πιθανότητα αυτού του σφάλματος είναι το επίπεδο σημαντικότητας

$$P(\text{απόρριψη της } H_0 \mid H_0 \text{ σωστή}) = a.$$

4. *Σωστή απόφαση*: Απορρίπτουμε την H_0 και η H_0 είναι λανθασμένη. Η πιθανότητα αυτής της απόφασης είναι

$$P(\text{απόρριψη της } H_0 \mid H_0 \text{ λανθασμένη}) = 1 - \beta$$

και δηλώνει την **ισχύ του ελέγχου** (power of the test).

Οι 4 δυνατές περιπτώσεις στην αποφάση του ελέγχου δίνονται στον Πίνακα 3.2. Για να είναι ένας έλεγχος ακριβής θα πρέπει το πραγματικό σφάλμα

	Αποδοχή της H_0	Απόρριψη της H_0
H_0 σωστή	ορθή απόφαση $(1-a)$	σφάλμα τύπου I (a)
H_0 λανθασμένη	σφάλμα τύπου II (β)	ορθή απόφαση $(1-\beta)$

Πίνακας 3.2: Οι 4 περιπτώσεις στην απόφαση ελέγχου με την αντίστοιχη πιθανότητα σε παρένθεση.

τύπου I να είναι στο επίπεδο σημαντικότητας a στο οποίο γίνεται ο έλεγχος. Στην πράξη αυτό δεν είναι βέβαια εφικτό αλλά μπορούμε να το διαπιστώσουμε αν έχουμε τη δυνατότητα να κάνουμε προσομοιώσεις. Για να υπολογίσουμε το πραγματικό σφάλμα τύπου I θα πρέπει να γνωρίζουμε ότι η H_0 είναι σωστή και να επαναλάβουμε τον έλεγχο σε M διαφορετικά δείγματα ίδιου τύπου και στο ίδιο επίπεδο σημαντικότητας a. Αν η H_0 απορρίπτεται m φορές σε επίπεδο σημαντικότητας a θα πρέπει για να είναι ο έλεγχος ακριβής (να έχει σωστή σημαντικότητα) η αναλογία m/M να είναι κοντά στο a.

Επίσης μας ενδιαφέρει ο έλεγχος να έχει μεγάλη ισχύ. Την ισχύ του ελέγχου μπορούμε υπολογιστικά να τη μετρήσουμε και πάλι με προσομοιώσεις

όπου τώρα θα πρέπει να γνωρίζουμε ότι η H_0 δεν είναι σωστή και επιπλέον ότι ο έλεγχος έχει σωστή σημαντικότητα (σύμφωνα με τα παραπάνω).

Οι παραπάνω προσομοιώσεις συνήθως ακολουθούνται για να αξιολογήσουμε το στατιστικό που χρησιμοποιείται στον έλεγχο. Στη συνέχεια θα παρουσιάσουμε τη διαδικασία του παραμετρικού ελέγχου για τη μέση τιμή.

3.3.1 Έλεγχος μέσης τιμής

Θέλουμε να ελέγξουμε με βάση ένα δείγμα παρατηρήσεων $\{x_1, \ldots, x_n\}$ μιας τ.μ. X αν η μέση τιμή μ της X μπορεί να πάρει κάποια τιμή μ_0, δηλαδή η μηδενική υπόθεση είναι $H_0 : \mu = \mu_0$. Φυσικά όταν υποθέτουμε $\mu = \mu_0$ δεν εννοούμε αυστηρά την ισότητα και θα θέλαμε ο έλεγχος να αποφασίζει ότι η H_0 είναι ορθή όταν η μ βρίσκεται 'κοντά' στην τιμή μ_0 και λανθασμένη αλλιώς. Έτσι η τυχόν απόρριψη της H_0 δεν ερμηνεύεται ως αποδοχή της πρότασης $\mu = \mu_0$ αλλά μη-απόρριψη της, πάντα με βάση το δείγμα.

Ο κατάλληλος εκτιμητής της μ είναι η δειγματική μέση τιμή $\bar{x}$ που υπολογίζεται από το δείγμα. Σύμφωνα και με τα παραπάνω, τιμές της $\bar{x}$ 'κοντά' στη μ_0 υποστηρίζουν την ορθότητα της H_0 και σχηματίζουν την περιοχής αποδοχής της H_0, ενώ τιμές της $\bar{x}$ 'μακριά' από τη μ_0 δεν την υποστηρίζουν και σχηματίζουν την *περιοχή απόρριψης* (rejection region) που συμβολίζουμε R.

Η απόφαση για την αποδοχή ή απόρριψη της H_0 γίνεται με βάση τις πιθανότητες και όπως για τα διαστήματα εμπιστοσύνης έτσι και εδώ ορίζουμε επίπεδο σημαντικότητας a (η επίπεδο εμπιστοσύνης $1 - a$) για την απόφαση ελέγχου. Το a καθορίζει το 'κοντά' και 'μακριά' που αναφέραμε παραπάνω. Στην παραμετρική προσέγγιση που ακολουθούμε εδώ το a ορίζει το εύρος των ουρών της κατανομής του εκτιμητή $\bar{x}$.

Είχαμε δείξει πως η κανονικοποίηση του εκτιμητή $\bar{x}$, $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$, ακολουθεί κατανομή Student θεωρώντας ότι η τ.μ. X ακολουθεί κανονική κατανομή ή ότι το δείγμα είναι μεγάλο. Θεωρώντας ότι ισχύει η H_0, έχουμε $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$, όπου t είναι το **στατιστικό ελέγχου** (test statistic). Οι κρίσιμες τιμές του t για δεδομένο a δίνουν τα όρια του R. Τιμές του t που είναι στις ουρές της κατανομής t_{n-1} ανήκουν στο R, δηλαδή δεν είναι πιθανές όταν ισχύει η H_0 και άρα η εμφάνιση τους συνιστά απόρριψη της H_0.

Υπολογίζουμε την τιμή του στατιστικού t από το δείγμα, έστω $\tilde{t}$. Αν το $\tilde{t}$ ανήκει στην περιοχή απόρριψης R που ορίσαμε για κάποιο επίπεδο σημαντικότητας a απορρίπτουμε την H_0. Σημειώνεται ότι αυτή η απόρριψη ισχύει για το a που επιλέχτηκε και μπορεί η απόφαση του ελέγχου να αλλάξει για μικρότερο a. Αντίστροφα αν δεν απορρίπτουμε την H_0 για κάποιο a, μπορεί να την απορρίψουμε για μεγαλύτερο a. Η μικρότερη τιμή του a που δίνει απόρριψη της H_0 λέγεται ***p*-τιμή** (*p*-value) και είναι η πιθανότητα να παρατηρήσουμε για το t μια τιμή τόσο ακραία όσο το $\tilde{t}$ όταν ισχύει η H_0. Άρα

p-τιμή είναι η πιθανότητα το t να είναι στο R που ορίζεται με κρίσιμη τιμή το $\tilde{t}$, δηλαδή

$$p = 2\,\mathrm{P}(t > |\tilde{t}|) = 2\,(1 - \mathrm{P}(t < |\tilde{t}|)). \tag{3.21}$$

Όσο πιο κοντά στο 0 είναι η p-τιμή τόσο πιο σίγουρη είναι η απόρριψη της H_0. Τιμές $p > 0.05$ δηλώνουν πως η H_0 δε μπορεί να απορριφθεί. Η p-τιμή υπολογίζεται εύκολα από την ασκ της κατανομής t_{n-1} (στο `matlab` δίνεται από τη συνάρτηση `tcdf`), αλλά παλιότερα που τέτοιοι υπολογισμοί δεν ήταν εύκολα εφικτοί δε χρησιμοποιούταν η p-τιμή και ο έλεγχος γινόταν σε κάποιο επίπεδο σημαντικότητας a (με απάντηση τύπου ναι / όχι).

Σε κάποιες περιπτώσεις ο έλεγχος μπορεί να είναι **μονόπλευρος** (one-sided), δηλαδή η απορριπτική περιοχή που ενισχύει την H_1 σχηματίζεται μόνο από τη μια ουρά της κατανομής γιατί θεωρούμε πως είναι αδύνατον για το πρόβλημα μας η μ να παίρνει τιμές στην άλλη ουρά της κατανομής του $\bar{x}$. Η επιλογή μονόπλευρου ή δίπλευρου ελέγχου εξαρτάται από την έρευνα που θέλουμε να κάνουμε και από το κατά πόσο μπορούμε να προβλέψουμε το αποτέλεσμα της έρευνας. Για παράδειγμα, έστω ότι θέλουμε να ελέγξουμε αν η μέση απόδοση μ ενός μηχανήματος που σχεδιάσαμε μπορεί να φθάσει την απόδοση αναφοράς μ_0 του 'άριστου' μηχανήματος (μηχάνημα αναφοράς). Σε αυτήν την περίπτωση ο έλεγχος πρέπει να είναι μονόπλευρος ($\mathrm{H}_0 : \mu = \mu_0$ ή ισοδύναμα $\mathrm{H}_0 : \mu \geq \mu_0$ και $\mathrm{H}_1 : \mu < \mu_0$) γιατί *γνωρίζουμε* πως δε μπορεί $\mu > \mu_0$ αφού η απόδοση του νέου μηχανήματος δε μπορεί να ξεπεράσει αυτή του μηχανήματος αναφοράς.

Παράδειγμα 3.5. Με αναφορά στο Παράδειγμα 3.3 έστω ότι θέλουμε να ελέγξουμε αν το μέσο όριο του ηλεκτρικού ρεύματος που καίγονται οι ασφάλειες των 40 αμπέρ είναι πράγματι 40. Οι υποθέσεις του ελέγχου είναι $\mathrm{H}_0 : \mu = 40$ και $\mathrm{H}_1 : \mu \neq 40$.

Είχαμε δεχθεί πως με βάση το δείγμα των 25 ασφαλειών η κατανομή του οριού του ηλεκτρικού ρεύματος που καίγονται οι ασφάλειες των 40 αμπέρ μπορεί να είναι κανονική. Χρησιμοποιούμε λοιπόν ως στατιστικό ελέγχου το $t = \frac{\bar{x}-40}{s/\sqrt{n}}$ που ακολουθεί την t_{n-1} σύμφωνα με την H_0. Για επίπεδο σημαντικότητας $a = 0.05$ η απορριπτική περιοχή ορίζεται από την κρίσιμη τιμή $t_{n-1,1-a/2}$ για $1-a/2 = 0.975$ και $n-1 = 24$, $t_{24,0.975} = 2.064$. Η απορριπτική περιοχή είναι

$$R = \{t \mid t < -2.064 \vee t > 2.064\} = \{t \mid |t| > 2.064\}.$$

Η τιμή του στατιστικού από το δείγμα είναι ($\bar{x} = 39.8$, $s = 0.925$)

$$\tilde{t} = \frac{39.8 - 40}{0.925/5} = -1.081$$

που προφανώς δεν ανήκει στην απορριπτική περιοχή και άρα δε μπορούμε να απορρίψουμε την H_0, ότι οι ασφάλειες καίγονται στο όριο των 40 αμπέρ. Από την ασκ της t_{24}-κατανομής για $\tilde{t} = -1.081$ βρίσκουμε την p-τιμή

$$p = 2(1 - P(t \leq |\tilde{t}|)) = 2(1 - P(t \leq 1.081)) = 2(1 - 0.855) = 0.29$$

που δηλώνει ποσοστό εμπιστοσύνης της απόρριψης της H_0 σε επίπεδο περίπου 70%. Πρακτικά αυτό σημαίνει πως δε μπορούμε να απορρίψουμε την H_0.

Αν για κάποιο λόγο αποκλείουμε ότι οι ασφάλειες μπορούν να καίγονται σε υψηλότερο όριο από 40 αμπέρ, τότε ο έλεγχος γίνεται μονόπλευρος, $H_0 : \mu \geq 40$ και $H_1 : \mu < 40$. Η απορριπτική περιοχή για $a = 0.05$ τότε είναι

$$R = \{t \mid t < t_{n-1,a/2}\} = \{t \mid t < t_{24,0.05}\} = \{t \mid t < -1.71\}.$$

Και πάλι όμως η H_0 δεν απορρίπτεται αφού $\tilde{t} \notin R$. Η p-τιμή για το μονόπλευρο έλεγχο είναι

$$p = P(t \leq \tilde{t}) = 0.145$$

που και πάλι είναι αρκετά υψηλό και δηλώνει πως η H_0 δε μπορεί να απορριφθεί.

3.3.2 Έλεγχος διασποράς

Η στατιστική υπόθεση για τη διασπορά σ^2 είναι όπως και για τη μέση τιμή, δηλαδή $H_0 : \sigma^2 = \sigma_0^2$ με κατάλληλη εναλλακτική υπόθεση H_1 ανάλογα αν ο έλεγχος είναι δίπλευρος ή μονόπλευρος. Το στατιστικό ελέγχου είναι το χ^2 που χρησιμοποιήθηκε στο διάστημα εμπιστοσύνης της σ^2 (δες Παράγραφο 3.2.2)

$$\chi^2 \equiv \frac{(n-1)s^2}{\sigma_0^2} \sim \mathcal{X}^2_{n-1}, \tag{3.22}$$

όπου s^2 είναι ο εκτιμητής της διασποράς. Από την κατανομή $\mathcal{X}^2$ με $n-1$ βαθμούς ελευθερίας και για επίπεδο σημαντικότητας a βρίσκουμε τις κρίσιμες τιμές και η περιοχή απόρριψης R δίνεται για τον κάθε τύπο ελέγχου ως

1. $H_1 : \sigma^2 \neq \sigma_0^2, \quad R = \{\chi^2 |\ \chi^2 < \chi^2_{n-1,a/2} \ \vee \ \chi^2 > \chi^2_{n-1,1-a/2}\}.$
2. $H_1 : \sigma^2 < \sigma_0^2, \quad R = \{\chi^2 |\ \chi^2 < \chi^2_{n-1,a}\}.$
3. $H_1 : \sigma^2 > \sigma_0^2, \quad R = \{\chi^2 |\ \chi^2 > \chi^2_{n-1,1-a}\}.$

Το στατιστικό ελέγχου από το δείγμα $\tilde{\chi}^2$ υπολογίζεται θέτοντας στη σχέση (3.22) την εκτίμηση s^2 από το δείγμα. Αν $\tilde{\chi}^2 \in R$ η H_0 απορρίπτεται. Η p-τιμή για τα τρία είδη ελέγχου είναι

1. $H_1 : \sigma^2 \neq \sigma_0^2, \quad p = P(\chi^2 < \tilde{\chi}^2 \ \vee \ \chi^2 > \tilde{\chi}^2\}.$
2. $H_1 : \sigma^2 < \sigma_0^2, \quad p = P(\chi^2 < \tilde{\chi}^2\}.$
3. $H_1 : \sigma^2 > \sigma_0^2, \quad p = P(\chi^2 > \tilde{\chi}^2\}.$

Παράδειγμα 3.6. Θέλουμε να ελέγξουμε σε επίπεδο εμπιστοσύνης 99% αν η τυπική απόκλιση σ του όριου έντασης ηλεκτρικού ρεύματος των ασφαλειών 40 αμπέρ μπορεί να είναι 0.7 αμπέρ. Χρησιμοποιούμε τα 25 δεδομένα για το όριο έντασης ηλεκτρικού ρεύματος που χρησιμοποιήθηκαν στο Παράδειγμα 3.3.

Εφαρμόζουμε δίπλευρο έλεγχο για τη διασπορά: $H_0 : \sigma^2 = 0.49$, $H_1 : \sigma^2 \neq 0.49$. Το μέγεθος του δείγματος είναι $n = 25$ και η δειγματική διασπορά είναι $s^2 = 0.854\,(\text{αμπέρ})^2$. Η κρίσιμη τιμή του στατιστικού ελέγχου χ^2 για $a = 0.01$ είναι $\chi^2_{24,0.005} = 9.886$ και $\chi^2_{24,0.995} = 45.558$. Η περιοχή απόρριψης είναι

$$R = \{\chi^2 \mid \chi^2 < 9.886 \vee \chi^2 > 45.558\}.$$

Η στατιστική ελέγχου που παίρνουμε από το δείγμα είναι

$$\tilde{\chi}^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{24 \cdot 0.854}{0.49} = 41.829.$$

Η $\tilde{\chi}^2$ δεν ανήκει στην περιοχή απόρριψης και άρα δε μπορούμε να απορρίψουμε την H_0 στο επίπεδο σημαντικότητας $a = 0.01$. Φαίνεται όμως να μπορούμε να απορρίψουμε την H_0 για λίγο υψηλότερο a. Η τιμή αυτή είναι

$$\begin{aligned} p &= P(\chi^2 < \tilde{\chi}^2 \ \vee \ \chi^2 > \tilde{\chi}^2\} = P(\chi^2 > 41.829\} \\ &= 1 - P(\chi^2 < 41.829\} = 1 - 0.986 = 0.014. \end{aligned}$$

και άρα μπορούμε να απορρίψουμε την H_0 σε επίπεδο σημαντικότητας ως και 0.014 (αλλά όχι 0.01). Στην παραπάνω σχέση της πιθανότητας επιλέγουμε μια από τις δύο ανισότητες (για την αριστερή ή δεξιά κρίσιμη τιμή που δίνεται από το $\tilde{\chi}^2$) για να υπολογίσουμε την p-τιμή.

Γενικά όταν θέλουμε να κάνουμε στατιστικό έλεγχο για την τυπική απόκλιση σ υψώνουμε στο τετράγωνο την τιμή της τυπικής απόκλισης που θέλουμε να ελέγξουμε και κάνουμε έλεγχο διασποράς γι αυτήν την τιμή.

3.3.3 Έλεγχος καταλληλότητας $\mathcal{X}^2$

Μια άλλη περίπτωση που χρησιμοποιείται η κατανομή $\mathcal{X}^2$ είναι ο **έλεγχος καλής προσαρμογής** (goodness-of-fit test) γνωστής κατανομής στα δεδομένα, δηλαδή για να ελέγξουμε αν το δείγμα προέρχεται από πληθυσμό με κάποια γνωστή κατανομή.

Για τον έλεγχο $\mathcal{X}^2$ η τ.μ. X πρέπει να είναι διακριτή, ή αν είναι συνεχής να μετατραπεί σε διακριτή (δες Παράγραφο 2.1.1). Πρακτικά αυτό σημαίνει πως αν τα δεδομένα είναι αριθμητικά θα πρέπει να ορισθεί πρώτα διαμέριση των τιμών και να ομαδοποιηθούν. Έστω οι K παρατηρούμενες διακριτές τιμές O_j, $j = 1, \ldots, K$, και οι αντίστοιχες αναμενόμενες τιμές E_j από τη γνωστή κατανομή. Στην περίπτωση της διακριτικοποίησης, O_j είναι η συχνότητα εμφάνισης παρατηρήσεων σε κάθε ομάδα (διάστημα) j από τις K ομάδες της διαμέρισης και E_j η αντίστοιχη αναμενόμενη συχνότητα. Το στατιστικό ελέγχου είναι

$$\chi^2 = \sum_{j=1}^{K} \frac{(O_j - E_j)^2}{E_j}. \tag{3.23}$$

Όταν η παρατηρούμενη τ.μ. X είναι διακριτή τότε οι αναμένομενες τιμές E_j δίνονται από το γινόμενο του πλήθους των δεδομένων n με την αντίστοιχη πιθανότητα της διακριτής κατανομής $f_X(x_j) = \mathrm{P}(X = x_j)$. Όταν η X είναι συνεχής οι αναμένομενες τιμές E_j υπολογίζονται ως

$$E_j = n(F_X(x_j^u) - F_X(x_j^l)), \tag{3.24}$$

όπου $F_X(x)$ είναι η ασκ της X για την τιμή x και x_j^l και x_j^u είναι το κάτω και πάνω άκρο του διαστήματος j, αντίστοιχα.

Οι βαθμοί ελευθερίας της κατανομής $\mathcal{X}^2$ είναι $K - c$, όπου για διακριτή κατανομή $c = 1$ επειδή οι αναμενόμενες συχνότητες E_j δεν είναι όλες ανεξάρτητες αφού το άθροισμα τους πρέπει να είναι n. Για συνεχή κατανομή, οι παράμετροι της εκτιμούνται από τα δεδομένα και άρα χάνονται τόσοι βαθμοί ελευθερίας επιπλέον όσες και οι παράμετροι της κατανομής. Για παράδειγμα για να ελέγξουμε αν τα δεδομένα προσαρμόζονται σε κανονική κατανομή, υπολογίζονται πρώτα οι εκτιμήσεις της μέσης τιμής $\bar{x}$ και διασποράς s^2 από τα δεδομένα για να ορίσουμε την ασκ της κανονικής κατανομής και να υπολογίσουμε στη συνέχεια τις αναμενόμενες τιμές E_j από την σχέση (3.24). Άρα σε αυτήν την περίπτωση οι βαθμοί ελευθερίας της $\mathcal{X}^2$ κατανομής είναι $K - 3$.

Έχοντας ότι κάτω από την H_0 είναι $\chi^2 \sim \mathcal{X}^2_{K-c}$, εξετάζουμε αν το στατιστικό $\tilde{\chi}^2$ από το δείγμα που δίνεται από την σχέση (3.23) ανήκει στην απορριπτική περιοχή $R = \{\chi^2 |\ \chi^2 > \chi^2_{K-c,1-a/2}\}$. Εναλλακτικά μπορούμε να χρησιμοποιήσουμε την p-τιμή που ορίζεται ως $p = \mathrm{P}(\chi^2 > \tilde{\chi}^2\}$.

Ο έλεγχος $\mathcal{X}^2$ μπορεί να εφαρμοσθεί για οποιαδήποτε μονομεταβλητή κατανομή, διακριτή ή συνεχή, για την οποία η ασκ μπορεί να υπολογισθεί. Έχει όμως και κάποια μειονεκτήματα: εξαρτάται από τη διαμέριση για τη διακριτικοποίηση των δεδομένων (όταν είναι αριθμητικά) και απαιτεί αρκετά μεγάλο δείγμα. Υπάρχουν και άλλοι έλεγχοι καλής προσαρμογής κατανομής, όπως ο έλεγχος Kolmogorov-Smirnov που εφαρμόζεται όμως μόνο για συνεχή κατανομή.

Παράδειγμα 3.7. Έστω ένα παιχνίδι ζαριών, όπου ο παίχτης πετάει το ζάρι τρεις φορές και κερδίζει ανάλογα με το πλήθος των εξαριών που φέρνει. Ας υποθέσουμε ότι ένας παίχτης παίζει 100 φορές και οι παρατηρούμενες εμφανίσεις εξαριών δίνονται στον Πίνακα 3.3. Αν το παιχνίδι είναι δίκαιο θα

πλήθος εξαριών	παρατηρούμενο πλήθος	αναμενόμενο πλήθος
0	47	57.9
1	36	34.7
2	14	6.9
3	4	0.5

Πίνακας 3.3: Οι εμφανίσεις εξαριών σε 3 ρίψεις ζαριών για 100 επαναλήψεις. Στη δεύτερη στήλη είναι οι παρατηρούμενες συχνότητες και στην τρίτη οι αναμενόμενες από τη διωνυμική κατανομή $B(3, 1/6)$.

πρέπει να εμφανίζεται εξάρι σε κάθε ζαριά με πιθανότητα 1/6 (πιθανότητα 'επιτυχίας'). Στις 3 ζαριές η πιθανότητα να έρθουν 0,1,2, ή 3 εξάρια δίνεται από τη διωνυμική κατανομή $B(m, p)$ για πλήθος επαναλήψεων $m = 3$, πιθανότητα επιτυχίας $p = 1/6$ και δυνατές τιμές της τ.μ. X εμφάνισης εξαριών 0,1,2,και 3. Θέλουμε να ελέγξουμε αν πράγματι μπορούμε να δεχτούμε ότι $X \sim B(3, 1/6)$ με βάση αυτό το δείγμα.

Οι πιθανότητες εμφάνισης 0,1,2, και 3 εξαριών από την $B(3, 1/6)$ είναι (δες Παράγραφο 2.3.1): $P(X = 0) = 0.579$, $P(X = 1) = 0.347$, $P(X = 2) = 0.069$, $P(X = 3) = 0.005$. Οι αντίστοιχες αναμενόμενες συχνότητες είναι $(n \cdot P(X = x))$ και δίνονται στην τρίτη στήλη του Πίνακα 3.3. Συγκρίνοντας τις παρατηρούμενες και τις αναμενόμενες τιμές φαίνεται να υπάρχουν σοβαρές διαφορές για τις πετυχημένες ζαριές με 2 και 3 εξάρια.

Εφαρμόζοντας τη σχέση (3.23) βρίσκουμε $\chi^2 = 36.28$. Η τιμή αυτή είναι πολύ μεγαλύτερη από την κρίσιμη τιμή της $\mathcal{X}^2_3$ (βαθμοί ελευθερίας $K - c = 4 - 1$) για $a = 0.05$ που είναι $\chi^2_{3,0.95} = 7.815$. Η p-τιμή του ελέγχου είναι $p = 6.5 \cdot 10^{-8}$, δηλαδή είναι πάρα πολύ απίθανο το δείγμα αυτό να προέρχεται από τη διωνυμική κατανομή $B(3, 1/6)$. Μάλλον λοιπόν ο παίχτης δεν ρίχνει σωστά τα ζάρια αλλά κάτι κάνει και φέρνει πιο συχνά εξάρες!

3.4 Μέθοδοι επαναδειγματοληψίας για διαστήματα εμπιστοσύνης και ελέγχους υπόθεσης

Πριν προχωρήσουμε στην περιγραφή των μεθόδων επαναδειγματοληψίας ας ανακεφαλαιώσουμε τα βήματα της εκτίμησης διαστήματος εμπιστοσύνης (δ.ε.) και έλεγχου υπόθεσης, χρησιμοποιώντας ως παράδειγμα την εκτίμηση

της μέσης τιμής και έλεγχο υπόθεσης μέσης τιμής με χρήση του στατιστικού που ακολουθεί την κατανομή student. Όπως αναφέρθηκε στην αρχή του κεφαλαίου, όταν για τον υπολογισμό του δ.ε. ή τον έλεγχο υπόθεσης γίνεται χρήση γνωστής κατανομής αυτά ονομάζονται **παραμετρικό διάστημα εμπιστοσύνης** και **παραμετρικός έλεγχο υπόθεσης**.

Παραμετρική εκτίμηση παραμέτρων

1. Επιλέγουμε ένα κατάλληλο στατιστικό για την παράμετρο που θα εκτιμήσουμε, π.χ. $\bar{x}$ για την παράμετρο μ μιας τ.χ. X.

2. Επιλέγουμε κατάλληλη κανονικοποίηση του στατιστικού ώστε να ακολουθεί κάποια γνωστή κατανομή, π.χ. στην περίπτωση άγνωστης διασποράς σ^2 γνωρίζουμε πως αν η τ.μ. X ακολουθεί κανονική κατανομή ή το δείγμα n είναι μεγάλο τότε $t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$.

3. Με βάση τη γνωστή κατανομή υπολογίζουμε τις κρίσιμες τιμές που καθορίζουν τις ουρές της κατανομής σε επίπεδο εμπιστοσύνης $(1-a)\%$, π.χ. $-t_{n-1,1-a/2}$ για την αριστερή ουρά και $t_{n-1,1-a/2}$ για τη δεξιά ουρά της κατανομής.

4. Υπολογίζουμε το $(1-a)\%$ διάστημα εμπιστοσύνης της παραμέτρου με βάση τη γνωστή βασική κατανομή, π.χ. το $(1-a)\%$ δ.ε. για μ είναι $\bar{x} \pm t_{n-1,1-a/2}\frac{s}{\sqrt{n}}$.

Παραμετρικός έλεγχος υπόθεσης

1. Ορίζουμε την τιμή της παραμέτρου στη μηδενική υπόθεση καθώς και στην εναλλακτική υπόθεση, π.χ. H_0: $\mu = \mu_0$ και H_0: $\mu \neq \mu_0$ (για δίπλευρο έλεγχο).

2. Επιλέγουμε ένα κατάλληλο στατιστικό για τον έλεγχο, π.χ. $\bar{x}$ για την παράμετρο μ μιας τ.χ. X.

3. Επιλέγουμε κατάλληλη κανονικοποίηση του στατιστικού ώστε να ακολουθεί κάποια γνωστή κατανομή κάτω από τη μηδενική υπόθεση, π.χ. στην περίπτωση άγνωστης διασποράς σ^2 γνωρίζουμε πως αν η τ.μ. X ακολουθεί κανονική κατανομή ή το δείγμα n είναι μεγάλο τότε $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$.

4. Υπολογίζουμε το κανονικοποιημένο στατιστικό στο δείγμα, $\tilde{t} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$.

5. Με βάση τη γνωστή κατανομή υπολογίζουμε τις κρίσιμες τιμές που καθορίζουν τις ουρές της κατανομής σε επίπεδο εμπιστοσύνης $(1-a)\%$, π.χ. $-t_{n-1,1-a/2}$ για την αριστερή ουρά και $t_{n-1,1-a/2}$ για τη δεξιά ουρά της κατανομής.

6. Ορίζουμε την απορριπτική περιοχή και αποφασίζουμε την απόρριψη ή όχι της H_0, π.χ. απόρριψη της H_0 αν $|\tilde{t}| > t_{n-1,1-a/2}$.

Στο παραμετρικό διάστημα εμπιστοσύνης και στον παραμετρικό έλεγχο υπόθεσης θεωρούμε την κανονικοποίηση του στατιστικού, όπως $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ για το στατιστικό $\bar{x}$ της μέσης τιμής μ. Στο παράδειγμα αυτό, η κανονικοποίηση γίνεται αφαιρώντας τη μέση τιμή (ή την υποτιθέμενη μέση τιμή μ_0 για τον έλεγχο υπόθεσης μέσης τιμής) και διαιρώντας με το τυπικό σφάλμα (τυπική απόκλιση) του εκτιμητή $\bar{x}$, $s/\sqrt{n}$. Ενώ για τον εκτιμητή της μ, $\bar{x}$, είναι γνωστό το τυπικό σφάλμα, $s/\sqrt{n}$, για άλλους εκτιμητές δεν είναι γνωστό. Επίσης θεωρούμε πως η κατανομή του κανονικοποιημένου στατιστικού είναι γνωστή π.χ. student στο παράδειγμα μας, $t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$. Δε γνωρίζουμε όμως πάντα την κατανομή του (κανονικοποιημένου) στατιστικού. Υπάρχουν λοιπόν περιορισμοί που δεν επιτρέπουν την εκτίμηση παραμέτρου με παραμετρικό διάστημα εμπιστοσύνης και την πραγματοποίηση παραμετρικού ελέγχου υπόθεσης. Οι μέθοδοι επαναδειγματοληψίας μας επιτρέπουν να ορίσουμε διάστημα εμπιστοσύνης και να κάνουμε έλεγχο υπόθεσης για κάθε περίπτωση χωρίς περιορισμούς.

Οι **μέθοδοι επαναδειγματοληψίας** είναι υπολογιστικές μέθοδοι που μας επιτρέπουν να προσεγγίσουμε την κατανομή οποιουδήποτε στατιστικού με βάση μόνο τις παρατηρήσεις του δείγματος.

3.4.1 Η μέθοδος επαναδειγματοληψίας bootstrap

Θεωρούμε το δείγμα $\mathbf{x} = \{x_1, \ldots, x_n\}$ της τ.μ. X που ακολουθεί κάποια άγνωστη κατανομή με ασκ $F_X(x)$. Θέλουμε να εκτιμήσουμε την παράμετρο ϑ της κατανομής F και έστω ότι έχουμε ένα στατιστικό $\hat{\vartheta} = g(\mathbf{x})$, π.χ. $\hat{\vartheta} \equiv \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ για την παράμετρο $\vartheta \equiv \mu$.

Η εμπειρική κατανομή Δεδομένου του $\mathbf{x} = \{x_1, \ldots, x_n\}$ μπορούμε να ορίσουμε την εμπειρική κατανομή της X με ασκ $\hat{F}_X(x)$. Θεωρούμε πως αντί να έχουμε τις τιμές της τ.μ. X στον πληθυσμό, όπου ορίζεται η κατανομή της X με ασκ $F_X(x)$, έχουμε τις τιμές της X στο δείγμα μεγέθους n, δηλαδή πληθυσμός είναι το δείγμα των n τιμών. Ορίζουμε πρώτα πως όλες οι τιμές στο δείγμα έχουν την ίδια πιθανότητα εμφάνισης $1/n$, δηλαδή σε κάθε παρατήρηση x_i, $x_i \in \mathbf{x}$, $\hat{p}(x_i) = 1/n$ (χρησιμοποιείται ο συμβολισμός $\hat{p}$ για να

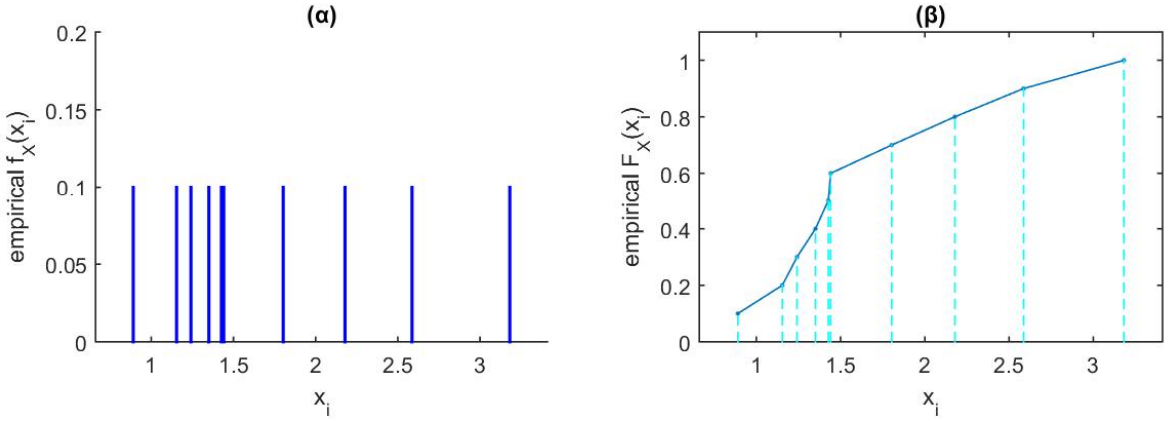
δηλώσει την αναφορά στο δείγμα και όχι στον πληθυσμό της X). Η **εμπειρική αθροιστική συνάρτηση κατανομής (εασκ)** (empirical cumulative density function, ecdf) για κάποια τιμή x_i ορίζεται από την αναλογία των τιμών στο δείγμα που είναι μικρότερες ή ίσες με τη x_i,

$$\hat{F}_n(x) = \hat{p}(X < x) = \frac{\text{πλήθος στοιχείων στο δείγμα} \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)$$

όπου $I(.)$ είναι η δείκτρια συνάρτηση που δίνεται ως

$$I(x_i \leq x) = \begin{cases} 1 & \text{όταν } x_i \leq x \\ 0 & \text{όταν } x_i > x \end{cases}$$

Στο Σχήμα 3.4, δίνεται ένα δείγμα 10 παρατηρήσεων και η αντίστοιχη εασκ.



Σχήμα 3.4: (α) Δείγμα $n = 10$ παρατηρήσεων. (β) Η εασκ για το δείγμα στο (α).

Όπως φαίνεται στο Σχήμα 3.4β, η εασκ $\hat{F}_n(x)$ είναι βηματική και αυξάνει κατά $1/n$ με κάθε παρατήρηση σε αύξουσα σειρά. Οι 10 παρατηρήσεις είναι από κανονική κατανομή και η εασκ φαίνεται να διαφέρει αρκετά από τη σιγμοειδή καμπύλη της ασκ της κανονικής κατανομής (δες Σχήμα 2.3). Αυτό συμβαίνει γιατί το δείγμα είναι πολύ μικρό και για $n = 10$ η εκτίμηση της ασκ, από την οποία εξάγαμε τις παρατηρήσεις, με την εασκ έχει μεγάλη διασπορά. Είναι γνωστό πως η εασκ είναι εκτιμητής της ασκ και συγκλίνει με πιθανότητα ένα στην ασκ (θεώρημα Glivenko - Cantelli) για $n \to \infty$.

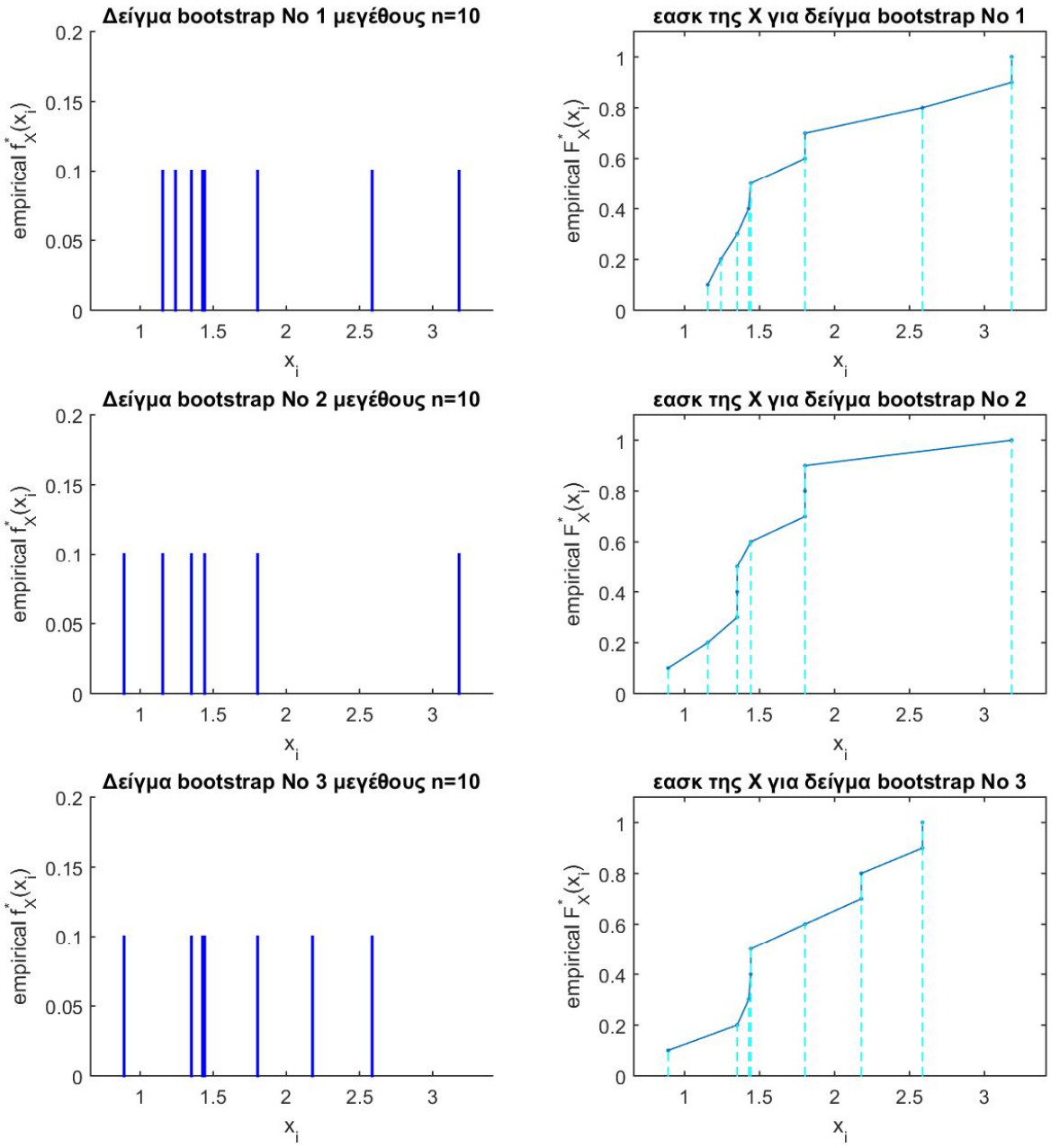
Δείγμα bootstrap Η μέθοδος bootstrap αντικαθιστά την (άγνωστη) κατανομή του στατιστικού ενδιαφέροντος με την εμπειρική κατανομή που σχηματίζεται από πολλά δείγματα που δημιουργούνται από το αρχικό δείγμα με

επαναδειγματοληψία. Η μέθοδος bootstrap βασίζεται στην ιδέα του δείγματος bootstrap. Για το δείγμα $\mathbf{x} = \{x_1, \ldots, x_n\}$ της τ.μ. X ορίσαμε παραπάνω την εασκ $\hat{F}_n(x)$. Αυτή είναι η ασκ που αντιστοιχεί (όχι στον αρχικό πληθυσμό στον οποίο αναφέρεται η X αλλά) στον πληθυσμό n στοιχείων που είναι τα στοιχεία του δείγματος $\mathbf{x} = \{x_1, \ldots, x_n\}$. Σε αυτόν τον 'πληθυσμό' μπορούμε να κάνουμε δειγματοληψία και να παράγουμε νέα δείγματα. Αυτά είναι τα δείγματα bootstrap. Το **δείγμα bootstrap** $\mathbf{x}^*$ έχει n παρατηρήσεις που εξάγονται από την $\hat{F}_n(x)$. Ο αστερίσκος στο $\mathbf{x}^*$ δηλώνει πως το δείγμα bootstrap δεν είναι το ίδιο με το αρχικό δείγμα $\mathbf{x}$. Ισοδύναμα και πιο απλά, το δείγμα bootstrap $\mathbf{x}^* = \{x_1^*, \ldots, x_n^*\}$ είναι ένα δείγμα με n παρατηρήσεις, όπου η κάθε παρατήρηση επιλέγεται τυχαία και **με επανάθεση** από τις παρατηρήσεις του $\mathbf{x} = \{x_1, \ldots, x_n\}$. Λόγω της επανάθεσης κάποιες παρατηρήσεις του $\mathbf{x}$ μπορεί να μην εμφανιστούν στο $\mathbf{x}^*$ ενώ κάποιες άλλες μπορεί να εμφανιστούν περισσότερες φορές. Υπολογιστικά η δειγματοληψία με επανάθεση γίνεται με την επιλογή n τυχαίων τιμών από τη λίστα των ακεραίων τιμών $\{1, 2, \ldots, n\}$ που είναι οι δείκτες των αντίστοιχων παρατηρήσεων $\{x_1, \ldots, x_n\}$.

Στο Σχήμα 3.5, δίνονται τρία δείγματα bootstrap που σχηματίστηκαν από το αρχικό δείγμα $n = 10$ παρατηρήσεων (δες Σχήμα 3.4) και η αντίστοιχη εασκ για κάθε δείγμα bootstrap. Από τη σύγκριση των γραφημάτων εασκ στο Σχήμα 3.5 μεταξύ τους καθώς και με αυτό στο Σχήμα 3.4β φαίνεται οι εασκ $\hat{F}_n(x^*)$ των δειγμάτων bootstrap να διαφέρουν αρκετά μεταξύ τους καθώς και με την εασκ $\hat{F}_n(x)$ του αρχικού δείγματος. Αυτό συμβαίνει γιατί το δείγμα είναι μικρό. Για μεγάλα δείγματα οι εασκ των δειγμάτων bootstrap συγκλίνουν στην εασκ του αρχικού δείγματος, καθώς όλες οι εασκ συγκλίνουν στην ασκ της τ.μ. X.

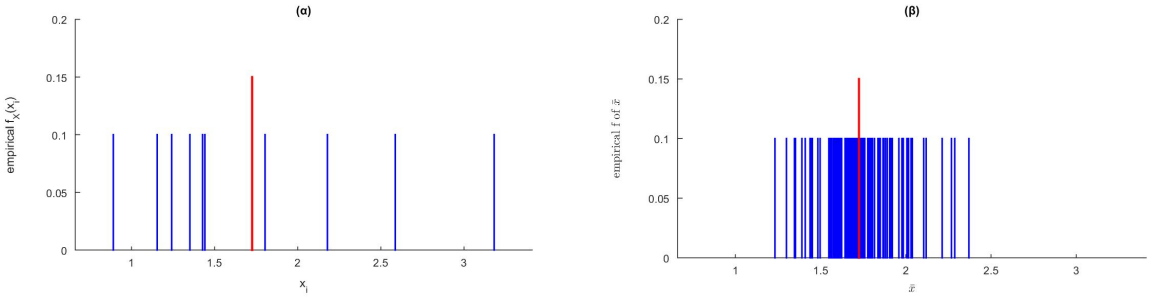
Η επανάληψη bootstrap του στατιστικού Θεωρούμε τώρα ένα στατιστικό $\hat{\theta}$ μιας παραμέτρου θ που δίνεται ως κάποια συνάρτηση g του δείγματος $\mathbf{x} = \{x_1, \ldots, x_n\}$, $\hat{\theta} = g(\mathbf{x})$, π.χ. $\hat{\theta} \equiv \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Για κάθε δείγμα bootstrap $\mathbf{x}^* = \{x_1^*, \ldots, x_n^*\}$ μπορούμε κατά τον ίδιο τρόπο να υπολογίσουμε την τιμή του στατιστικού στο δείγμα bootstrap, $\hat{\theta}^* = g(\mathbf{x}^*)$, π.χ. $\bar{x}^* = \frac{1}{n}\sum_{i=1}^{n} x_i^*$. Η τιμή $\hat{\theta}^*$ είναι μια **επανάληψη bootstrap του** $\hat{\theta}$. Για B δείγματα bootstrap έχουμε B επαναλήψεις bootstrap του στατιστικού $\hat{\theta}$, $\{\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}\}$.

Για το παράδειγμα της παραμέτρου της μέσης τιμής, $\theta \equiv \mu$, με στατιστικό το μέσο όρο, $\hat{\theta} \equiv \bar{x}$, δίνεται στο Σχήμα 3.6 ένα δείγμα 10 παρατηρήσεων (και ο μέσος όρος του), καθώς και οι μέσοι όροι από $B = 100$ δείγματα bootstrap. Παρατηρούμε πως οι $B = 100$ επαναλήψεις bootstrap του μέσου όρου, $\{\bar{x}^{*1}, \bar{x}^{*2}, \ldots, \bar{x}^{*B}\}$, είναι γύρω από τον μέσο όρο $\bar{x}$ του αρχικού δείγματος. Αυτό ακριβώς είναι το χαρακτηριστικό της μεθόδου bootstrap που μας επιτρέπει να εκτιμούμε την παράμετρο από τις επαναλήψεις bootstrap του



Σχήμα 3.5: Τρία δείγματα bootstrap (αριστερά) και η αντίστοιχη εασκ (δεξιά) για δείγμα $n = 10$ παρατηρήσεων.

στατιστικού χωρίς να υποθέτουμε κάποια γνωστή κατανομή για το στατιστικό. Η μεταβλητότητα των τιμών $\bar{x}^{*i}$ γύρω από το $\bar{x}$ καθορίζεται από τη διασπορά σ^2 της τ.μ. X και το μέγεθος n του δείγματος.



Σχήμα 3.6: (α) Δείγμα $n = 10$ παρατηρήσεων και ο μέσος όρος (με κόκκινο, η γραμμή που προεξέχει). (β) Ο μέσος όρος του αρχικού δείγματος στο (α) και $B = 100$ επαναλήψεις bootstrap του μέσου όρου.

3.4.2 Bootstrap εκτίμηση του τυπικού σφάλματος εκτιμητή

Η μεταβλητότητα ενός στατιστικού $\hat{\theta}$ δίνεται από το **τυπικό σφάλμα (standard error)** $\text{se}(\hat{\theta}) = \sigma_{\hat{\theta}}$ και είναι απλά η τυπική απόκλιση του εκτιμητή $\hat{\theta}$. Γενικά για ένα οποιοδήποτε στατιστικό $\hat{\theta}$ δεν είναι γνωστό το τυπικό σφάλμα $\sigma_{\hat{\theta}}$. Οι επαναλήψεις bootstrap μας επιτρέπουν να εκτιμήσουμε το τυπικό σφάλμα $\sigma_{\hat{\theta}}$ για οποιοδήποτε στατιστικό $\hat{\theta}$.

Η εκτίμηση του τυπικού σφάλματος $\text{se}(\hat{\theta}) = \sigma_{\hat{\theta}}$ με τη μέθοδο bootstrap δίνεται στα παρακάτω βήματα:

1. Επιλέγουμε B δείγματα bootstrap $\mathbf{x}^{*1}, \ldots, \mathbf{x}^{*B}$.

2. Υπολογίζουμε B επαναλήψεις bootstrap του στατιστικού $\hat{\theta}$ στα B δείγματα bootstrap $\mathbf{x}^{*1}, \ldots, \mathbf{x}^{*B}$

$$\{\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}\}$$

3. Εκτιμούμε το τυπικό σφάλμα $\text{se}(\hat{\theta})$ από την τυπική απόκλιση του $\hat{\theta}$ στις B επαναλήψεις

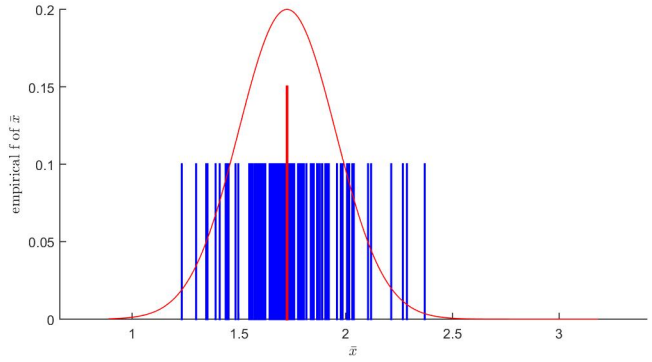
$$\hat{\text{se}}_B(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^{*b} - \bar{\theta}^{*}\right)^2}$$

όπου $\bar{\theta}^{*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}$ ο μέσος όρος των επαναλήψεων bootstrap του στατιστικού $\hat{\theta}$.

Για την παράμετρο της μέσης τιμής, $\theta \equiv \mu$, γνωρίζουμε την αναλυτική έκφραση του τυπικού σφάλματος του εκτιμητή $\hat{\theta} \equiv \bar{x}$ και είναι $\text{se}(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Είναι λοιπόν δυνατόν σε αυτήν την περίπτωση να εκτιμήσουμε το τυπικό

σφάλμα του $\bar{x}$ από την εκτίμηση της τυπικής απόκλισης σ της τ.μ. X, $\hat{\text{se}}(\bar{x}) = s_{\bar{x}} = \frac{s}{\sqrt{n}}$. Μια δεύτερη εκτίμηση μπορεί να γίνει από την τυπική απόκλιση των επαναλήψεων bootstrap για το μέσο όρο $\bar{x}$, $\{\bar{x}^{*1}, \bar{x}^{*2}, \ldots, \bar{x}^{*B}\}$.

Στο Σχήμα 3.7, φαίνονται και πάλι οι μέσοι όροι από $B = 100$ δείγματα bootstrap (μπλε μπάρες) καθώς και ο μέσος όρος του αρχικού δείγματος των $n = 10$ παρατηρήσεων (δες Σχήμα 3.6). Στο ίδιο σχήμα έχει σχηματιστεί



Σχήμα 3.7: Ο μέσος όρος δείγματος μεγέθους $n = 10$, οι $B = 100$ επαναλήψεις bootstrap του μέσου όρου και η κανονική κατανομή με κέντρο το μέσο όρο του δείγματος και τυπική απόκλιση την εκτίμηση του τυπικού του σφάλματος $s/\sqrt{n}$.

και η κανονική κατανομή με κέντρο το μέσο όρο του δείγματος $\bar{x}$ και τυπική απόκλιση την εκτίμηση του τυπικού του σφάλματος $\hat{\text{se}}(\bar{x}) = s/\sqrt{n}$. Για το συγκεκριμένο δείγμα η παραμετρική εκτίμηση του τυπικού σφάλματος του μέσου όρου είναι $\hat{\text{se}}(\bar{x}) = 0.227$, ενώ η bootstrap εκτίμηση είναι $\hat{\text{se}}_B(\bar{x}) = 0.213$. Οι δύο εκτιμήσεις είναι σχετικά κοντά μεταξύ τους αλλά όχι κοντά στο πραγματικό τυπικό σφάλμα, αφού το αρχικό δείγμα δημιουργήθηκε από τυχαία δειγματοληψία από τυπική κανονική κατανομή και άρα το τυπικό σφάλμα του $\bar{x}$ είναι $\sigma/\sqrt{n} = 1/\sqrt{10} = 0.316$.

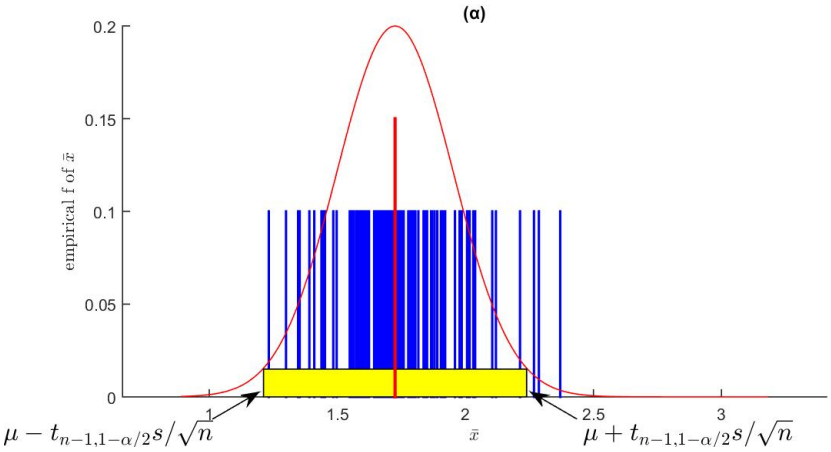
Παρατηρούμε στο Σχήμα 3.7 πως οι $B = 100$ επαναλήψεις bootstrap $\{\bar{x}^{*1}, \bar{x}^{*2}, \ldots, \bar{x}^{*B}\}$ δεν απλώνονται συμμετρικά γύρω από το μέσο όρο του δείγματος $\bar{x}$, ενώ η κανονική κατανομή είναι κεντραρισμένη στο $\bar{x}$. Αυτό οδηγεί σε διαφορετικά διαστήματα εμπιστοσύνης για τη μέση τιμή μ όπως θα δούμε στη συνέχεια.

3.4.3 Bootstrap εκτίμηση του διαστήματος εμπιστοσύνης

Έχουν προταθεί διαφορετικοί τρόποι να χρησιμοποιηθούν οι B επαναλήψεις bootstrap του στατιστικού $\hat{\theta}$, $\{\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}\}$, για να σχηματιστούν

διαστήματα εμπιστοσύνης (δ.ε.) για την παράμετρο ϑ. Εδώ θα εστιάσουμε στην **εκτίμηση του διαστήματος εμπιστοσύνης με ποσοστιαία bootstrap**.

Ας θυμηθούμε πρώτα πως γίνεται η παραμετρική εκτίμηση διαστήματος εμπιστοσύνης. Αυτή η εκτίμηση υποθέτει γνωστή κατανομή για το στατιστικό $\hat{\vartheta}$. Για παράδειγμα, για την εκτίμηση δ.ε. για τη μέση τιμή $\vartheta \equiv \mu$, θεωρούμε το στατιστικό του μέσου όρου, $\hat{\vartheta} \equiv \bar{x}$. Έχει βρεθεί πως αν η τ.μ. X ακολουθεί κανονική κατανομή τότε θεωρώντας το μετασχηματισμό $t = (\bar{x} - \mu)/(s/\sqrt{n})$ του $\bar{x}$, η τ.μ. t ακολουθεί την κατανομή student με $n-1$ βαθμούς ελευθερίας, $t \sim t_{n-1}$. Έτσι το $(1-a)\%$ δ.ε. για μ ορίζεται ως $\bar{x} \pm t_{n-1,1-a/2}\frac{s}{\sqrt{n}}$. Στο ίδιο παράδειγμα που χρησιμοποιήθηκε στις προηγούμενες ενότητες, δίνεται αυτό το δ.ε. στο Σχήμα 3.8. Παρατηρούμε πως στο συγκεκριμένο παράδειγμα,



Σχήμα 3.8: Ο μέσος όρος δείγματος μεγέθους $n = 10$, οι $B = 100$ επαναλήψεις bootstrap του μέσου όρου, η κανονική κατανομή με κέντρο το μέσο όρο του δείγματος και τυπική απόκλιση την εκτίμηση του τυπικού του σφάλματος $s/\sqrt{n}$, καθώς και το 95% παραμετρικό διάστημα εμπιστοσύνης για τη μέση τιμή με βάση το συγκεκριμένο δείγμα σε κίτρινο πλαίσιο.

το παραμετρικό 95% διάστημα εμπιστοσύνης για τη μέση τιμή (δίνεται με το κίτρινο πλαίσιο στο Σχήμα 3.8) δεν είναι σε πλήρη συμφωνία με τις $B = 100$ επαναλήψεις bootstrap $\{\bar{x}^{*1}, \bar{x}^{*2}, \ldots, \bar{x}^{*B}\}$, π.χ. ενώ υπάρχουν τιμές των επαναλήψεων bootstrap στα δεξιά του παραμετρικού δ.ε. δεν υπάρχουν στα αριστερά. Οι $B = 100$ επαναλήψεις bootstrap $\{\bar{x}^{*1}, \bar{x}^{*2}, \ldots, \bar{x}^{*B}\}$ σχηματίζουν μια άλλη κατανομή που φαίνεται να είναι διαφορετική από την κατανομή που δίνεται με βάση την κατανομή student. Αυτήν την κατανομή θα χρησιμοποιήσουμε στη συνέχεια για το σχηματισμό του δ.ε. με ποσοστιαία bootstrap.

Η εκτίμηση του διαστήματος εμπιστοσύνης της παραμέτρου ϑ με ποσο-

στιαία bootstrap από το δείγμα n παρατηρήσεων της τ.μ. X, δίνεται στα παρακάτω βήματα:

1. Επιλέγουμε B δείγματα bootstrap $\mathbf{x}^{*1}, \ldots, \mathbf{x}^{*B}$.

2. Υπολογίζουμε B επαναλήψεις bootstrap του στατιστικού $\hat{\vartheta}$ στα B δείγματα bootstrap $\mathbf{x}^{*1}, \ldots, \mathbf{x}^{*B}$

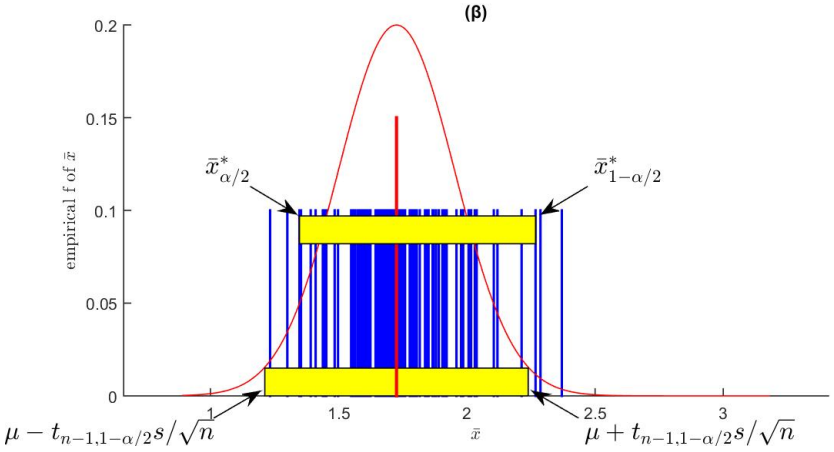
$$\{\hat{\vartheta}^{*1}, \hat{\vartheta}^{*2}, \ldots, \hat{\vartheta}^{*B}\}$$

   Οι τιμές αυτές σχηματίζουν την εμπειρική κατανομή του $\hat{\vartheta}$.

3. Οι ουρές αυτής της εμπειρικές κατανομής δίνονται από τα $a/2$ και $1 - a/2$ ποσοστιαία σημεία $\hat{\vartheta}^*_{a/2}$ και $\hat{\vartheta}^*_{1-a/2}$ του δείγματος των $\{\hat{\vartheta}^{*1}, \hat{\vartheta}^{*2}, \ldots, \hat{\vartheta}^{*B}\}$

4. Το διάστημα $[\hat{\vartheta}^*_{a/2}, \hat{\vartheta}^*_{1-a/2}]$ είναι το $(1-a)\%$ δ.ε. για ϑ με τη μέθοδο των ποσοστιαίων bootstrap.

Για το ίδιο παράδειγμα δίνεται στο Σχήμα 3.9 το δ.ε. ποσοστιαίων bootstrap για τη μέση τιμή μαζί με το παραμετρικό δ.ε.. Παρατηρούμε πως το



Σχήμα 3.9: Το ίδιο σχήμα με αυτό στο Σχήμα 3.8, όπου έχει προστεθεί και το δ.ε. ποσοστιαίων bootstrap για τη μέση τιμή.

δ.ε. ποσοστιαίων bootstrap για τη μέση τιμή δεν είναι κεντραρισμένο γύρω από το μέσο όρο $\bar{x}$ όπως το παραμετρικό δ.ε.. Αυτό συμβαίνει γιατί τυχαίνει τα ποσοστιαία σημεία $a/2\%$ και $(1 - a/2)\%$ να μη βρίσκονται συμμετρικά γύρω από το $\bar{x}$.

Συγκεκριμένα τα $a/2\%$ και $(1-a/2)\%$ ποσοστιαία σημεία από τα $\{\hat{\vartheta}^{*1}, \hat{\vartheta}^{*2}, \ldots, \hat{\vartheta}^{*B}\}$ ορίζονται ως εξής. Το $a/2\%$ ποσοστιαίο σημείο είναι το σημείο στη θέση

$k = [(B+1)a/2]$, όπου $[x]$ είναι το ακέραιο μέρος του x και θεωρώντας πως τα $\hat{\vartheta}^{*i}$ έχουν διαταχθεί πρώτα σε αύξουσα σειρά. Το $(1-a/2)\%$ ποσοστιαίο σημείο είναι το $B+1-k$ σημείο.

Για $\vartheta \equiv \mu$ το παραμετρικό δ.ε. $\bar{x} \pm t_{n-1,1-a/2}\frac{s}{\sqrt{n}}$ και το δ.ε. ποσοστιαίων bootstrap $[\bar{x}^*_{a/2}, \bar{x}^*_{1-a/2}]$ γενικά συμφωνούν και συγκλίνουν για $n \to \infty$. Για άλλα στατιστικά μπορεί αυτό να μη συμβαίνει. Στο ερώτημα 'ποιο από τα δύο θα επιλέξουμε;', η απάντηση δεν είναι απλή. Τα παραμετρικά δ.ε. είναι ακριβή όταν ικανοποιούνται οι υποθέσεις για τη χρήση τους, όπως π.χ. κανονική κατανομή της τ.μ. X για το δ.ε. της μ με χρήση της κατανομής student. Σε αυτήν την περίπτωση δεν υπάρχει λόγος να προτιμηθεί το bootstrap δ.ε.. Σε περιπτώσεις όμως που οι υποθέσεις για τη χρήση του παραμετρικού δ.ε. δεν ικανοποιούνται, το bootstrap δ.ε. αποτελεί τη μέθοδο που θα πρέπει να προτιμηθεί.

Διάστημα εμπιστοσύνης για τη διαφορά μέσων τιμών. Θα επαναλάβουμε την εκτίμηση δ.ε. για τη διαφορά μέσω τιμών, πρώτα με παραμετρικό δ.ε. και μετά με δ.ε. ποσοστιαίων bootstrap. Έστω ότι έχουμε δείγμα $\{x_1, \ldots, x_n\}$ της τ.μ. X και $\{y_1, \ldots, y_m\}$ της τ.μ. Y και θέλουμε να ελέγξουμε αν οι μέσες τιμές των X και Y, μ_X και μ_Y, είναι ίσες. Η παράμετρος είναι $\vartheta \equiv \mu_X - \mu_Y$ και το αντίστοιχο στατιστικό $\hat{\vartheta} \equiv \bar{x} - \bar{y}$, όπου $\bar{x}$ και $\bar{y}$ οι μέσοι όροι των X και Y στα δύο δείγματα. Το παραμετρικό $(1-a)\%$ δ.ε. για $\mu_X - \mu_Y$ θεωρώντας κανονική κατανομή για τις X και Y και κοινή διασπορά σ^2 δίνεται με βάση την κατανομή student, δηλαδή θεωρούμε το μετασχηματισμένο στατιστικό $t \equiv \frac{(\bar{x}-\bar{y})-(\mu_X-\mu_Y)}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$ ακολουθεί κατανομή student με $n+m-2$ βαθμούς ελευθερίας, όπου s_p^2 είναι η εκτίμηση της κοινής (pooled) διασποράς, $s_p^2 = \frac{(n-1)s_X^2+(m-1)s_Y^2}{n+m-2}$. Το παραμετρικό $(1-a)\%$ δ.ε. για $\mu_X - \mu_Y$ είναι $(\bar{x}-\bar{y}) \pm t_{n+m-2,1-a/2}\, s_p\sqrt{\frac{1}{n}+\frac{1}{m}}$.

Το αντίστοιχο $(1-a)\%$ δ.ε. για $\vartheta \equiv \mu_X - \mu_Y$ με ποσοστιαία bootstrap δίνεται ως εξής.

1. Επιλέγουμε B δείγματα bootstrap για τη X, $\{\mathbf{x}^{*1}, \ldots, \mathbf{x}^{*B}\}$, και για τη Y, $\{\mathbf{y}^{*1}, \ldots, \mathbf{y}^{*B}\}$.

2. Υπολογίζουμε B επαναλήψεις bootstrap του στατιστικού $\hat{\vartheta} \equiv \bar{x} - \bar{y}$, $\{\bar{x}^{*1} - \bar{y}^{*1}, \ldots, \bar{x}^{*B} - \bar{y}^{*B}\}$, όπου $\bar{x}^{*i}$ είναι ο μέσος όρος του δείγματος $\mathbf{x}^{*i}$ και $\bar{y}^{*i}$ είναι ο μέσος όρος του δείγματος $\mathbf{y}^{*i}$.

3. Τα ποσοστιαία σημεία $(\bar{x}^* - \bar{y}^*)_{a/2}$ και $(\bar{x}^* - \bar{y}^*)_{1-a/2}$ ορίζουν το $(1-a)\%$ δ.ε. για $\mu_X - \mu_Y$ με τη μέθοδο των ποσοστιαίων bootstrap.

Άλλα bootstrap διαστήματα εμπιστοσύνης Το δ.ε. ποσοστιαίων bootstrap $[\hat{\vartheta}^*_{a/2}, \hat{\vartheta}^*_{1-a/2}]$ ορίζεται εύκολα όταν δίνονται οι B επαναλήψεις bootstrap του στατιστικού $\hat{\vartheta}$, $\{\hat{\vartheta}^{*1}, \hat{\vartheta}^{*2}, \ldots, \hat{\vartheta}^{*B}\}$. Δεν είναι όμως το πιο ακριβές. Υπάρχουν άλλα δ.ε. με τη μέθοδο bootstrap, που βασίζονται δηλαδή στις B επαναλήψεις bootstrap του στατιστικού $\hat{\vartheta}$, $\{\hat{\vartheta}^{*1}, \hat{\vartheta}^{*2}, \ldots, \hat{\vartheta}^{*B}\}$, και είναι πιο ακριβή αλλά και πιο σύνθετα από το δ.ε. ποσοστιαίων bootstrap. Τα πιο γνωστά είναι τα παρακάτω:

- Το **Studentized bootstrap** ή **bootstrap-t** διορθώνει τις κρίσιμες τιμές για $a/2$ και $1 - a/2$ στο παραμετρικό δ.ε. με αυτές από επαναλήψεις bootstrap.

- Το **bias corrected and accelerated (BCa) bootstrap** διορθώνει τη μεροληψία και λοξότητα στην κατανομή bootstrap του στατιστικού. Αυτό χρησιμοποιείται ως προεπιλογή στη συνάρτηση `bootci` του Matlab για τον υπολογισμό bootstrap διαστήματος εμπιστοσύνης.

3.4.4 Έλεγχος υπόθεσης με μεθόδους επαναδειγματοληψίας

Για την εκτίμηση διαστήματος εμπιστοσύνης με τη μέθοδο bootstrap δημιουργούμε B δείγματα bootstrap $\mathbf{x}^{*1}, \ldots, \mathbf{x}^{*B}$ από το αρχικό δείγμα και υπολογίζουμε B επαναλήψεις bootstrap του στατιστικού $\hat{\vartheta}$, $\{\hat{\vartheta}^{*1}, \hat{\vartheta}^{*2}, \ldots, \hat{\vartheta}^{*B}\}$. Οι τιμές αυτές σχηματίζουν την εμπειρική κατανομή του στατιστικού $\hat{\vartheta}$. Στον έλεγχο υπόθεσης θέλουμε αντίστοιχα να σχηματίσουμε την κατανομή του στατιστικού, αλλά κάτω από τη μηδενική υπόθεση, δηλαδή όταν ισχύει $\mathrm{H}_0 : \vartheta = \vartheta_0$. Όμως δεν είναι πάντα εφικτό να σχηματιστεί αυτή η κατανομή εμπειρικά με επαναλήψεις bootstrap.

Όπως είδαμε στον παραμετρικό έλεγχο προσεγγίζουμε την κατανομή του στατιστικού $\hat{\vartheta}$ κάτω από την H_0 με κάποια παραμετρική κατανομή και εξετάζουμε αν η τιμή του στατιστικού από το δείγμα είναι στα άκρα της κατανομής αυτής (για απόρριψη της H_0) ή όχι (για μη-απόρριψη της H_0). Για παράδειγμα για τον παραμετρικό έλεγχο της υπόθεσης ότι η μέση τιμή μιας τ.μ. X είναι κάποια τιμή μ_0 ($\vartheta \equiv \mu$ και $\mathrm{H}_0 : \mu = \mu_0$), χρησιμοποιούμε την κατανομή student. Μέσα από κατάλληλο μετασχηματισμό του αρχικού στατιστικού, $\bar{x}$, έχουμε $t \equiv (\bar{x} - \mu_0)/(s/\sqrt{n}) \sim t_{n-1}$. Αυτή είναι η κατανομή που θα περιμέναμε για το t αν $\mathrm{H}_0 : \mu = \mu_0$. Στη συνέχεια εξετάζουμε πόσο μακριά είναι το στατιστικό υπολογισμένο στο δείγμα, $\tilde{t}$, από το 0 χρησιμοποιώντας ως όριο την κρίσιμη τιμή της t_{n-1} για κάποιο επίπεδο σημαντικότητας a. Στο ίδιο αποτέλεσμα (με αυτό του δίπλευρου ελέγχου) καταλήγουμε και με το διάστημα εμπιστοσύνης, εξετάζοντας πόσο μακριά είναι το 0 από το στατιστικό

$t \sim t_{n-1}$ χρησιμοποιώντας ως όριο την ίδια κρίσιμη τιμή. Άρα ένας τρόπος να ελέγξουμε την τιμή μιας παραμέτρου με τη μέθοδο bootstrap είναι να δημιουργήσουμε το bootstrap διάστημα εμπιστοσύνης και να ελέγξουμε αν περιέχει αυτήν την τιμή. Αν θέλουμε όμως να κάνουμε τον έλεγχο υπόθεσης η μέθοδος bootstrap δε μπορεί πάντα να σχηματίσει την κατανομή κάτω από την H_0. Στη συνέχεια θα εστιάσουμε στον έλεγχο ισότητας μέσων τιμών, όπου μπορούμε να εκτελέσουμε τον έλεγχο και με τη μέθοδο bootstrap, δηλαδή να προσεγγίσουμε την κατανομή κάτω από την H_0 με επαναλήψεις bootstrap.

Έλεγχος υπόθεσης ισότητας δύο μέσων τιμών Έστω ότι έχουμε δείγμα $\{x_1, \ldots, x_n\}$ της τ.μ. X και $\{y_1, \ldots, y_m\}$ της τ.μ. Y και θέλουμε να ελέγξουμε αν οι μέσες τιμές των X και Y είναι ίσες. Η μηδενική υπόθεση είναι H_0 : $\mu_X - \mu_Y = 0$.

Για τον παραμετρικό έλεγχο είχαμε δει πως το κατάλληλο στατιστικό προκύπτει από την κανονικοποίηση του εκτιμητή της παραμέτρου $\vartheta = \mu_X - \mu_Y$:

$$t \equiv \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathrm{t}_{n+m-2}$$

Για την H_0 έχουμε $\mu_X - \mu_Y = 0$ και το στατιστικό ακολουθεί κατανομή student, δηλαδή ισχύει

$$t \equiv \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathrm{t}_{n+m-2}$$

Η τιμή του στατιστικού στο δείγμα είναι $\tilde{t} \equiv (\bar{x} - \bar{y})/(s_p \sqrt{1/n + 1/m})$ (με αντικατάσταση των τιμών $\bar{x}$, $\bar{y}$, s_p, n και m από το δείγμα των X και Y) και η απορριπτική περιοχή για την απόφαση του ελέγχου είναι $R = \{|\tilde{t}| > t_{n+m-2,1-a/2}\}$.

Για την H_0 : $\mu_X - \mu_Y = 0$ μπορούμε να κάνουμε έλεγχο που κάνει χρήση της επαναδειγματοληψίας. Πρώτα θα ορίσουμε τον **έλεγχο αντιμετάθεσης** (permutation test). Αυτός ο έλεγχος μπορεί να χρησιμοποιηθεί για μηδενικές υποθέσεις που αναφέρονται σε δύο δείγματα και κάτω από τη μηδενική υπόθεση προέρχονται από τον ίδιο πληθυσμό. Ουσιαστικά ελέγχουμε αν οι δύο κατανομές είναι ίδιες, συνήθως όμως ο έλεγχος αφορά κάποια παράμετρο κατανομής, όπως έλεγχος για ίσες μέσες τιμές, ίσες διαμέσους και ίσες διασπορές (ή τυπικές αποκλίσεις). Θεωρώντας πως οι δύο κατανομές των X και Y είναι ίδιες, οποιαδήποτε από τις n παρατηρήσεις του X θα μπορούσε να ήταν στο δείγμα των m παρατηρήσεων της Y και το αντίθετο. Άρα μπορούμε να συνδυάσουμε τις $n + m$ παρατηρήσεις σε ένα δείγμα και να επιλέξουμε τυχαία και **χωρίς επανάθεση** n παρατηρήσεις από αυτό να είναι οι νέες

παρατηρήσεις της X και οι υπόλοιπες m παρατηρήσεις να είναι οι νέες παρατηρήσεις της Y. Το πλήθος των νέων δειγμάτων που μπορούν να παραχθούν με αυτόν τον τρόπο είναι όλοι οι δυνατοί τρόποι να χωριστεί ένα σύνολο $n+m$ στοιχείων σε δύο υποσύνολα μεγέθους n και m, που δίνεται από το διωνυμικό συντελεστή $\binom{n+m}{n} = \frac{(n+m)!}{n!m!}$, όπου $n!$ είναι το παραγοντικό ακεραίου n. Για το στατιστικό ελέγχου μπορούν να υπολογιστούν $\binom{n+m}{n}$ τιμές του στατιστικού που σχηματίζουν εμπειρικά την κατανομή του στατιστικού κάτω από την H_0 και λέγεται **κατανομή αντιμετάθεσης** (permutation distribution). Η απορριπτική περιοχή μπορεί να οριστεί από τα άκρα της κατανομής αντιμετάθεσης για να παρθεί στη συνέχεια η απόφαση ελέγχου.

Ο έλεγχος αντιμετάθεσης χρησιμοποιείται μόνο για πολύ μικρά μεγέθη δειγμάτων n και m, καθώς θα πρέπει να σχηματιστεί μεγάλο πλήθος δειγμάτων $\binom{n+m}{n}$ στα οποία θα υπολογιστεί το στατιστικό ελέγχου. Για παράδειγμα για $n = m = 10$ το πλήθος δειγμάτων από αντιμετάθεση είναι περίπου 184756 ενώ για $n = m = 15$ πάνω από 150 εκατομμύρια. Πρακτικά ένα ικανά μεγάλο υποσύνολο μεγέθους B του συνόλου όλων των δυνατών δειγμάτων αντιμετάθεσης χρησιμοποιούνται για να σχηματίσουν την κατανομή του στατιστικού κάτω από την H_0. Αυτός ο έλεγχος λέγεται **έλεγχος τυχαίας αντιμετάθεσης ή τυχαιοποίησης** (random permutation or randomization test).

Για την $H_0 : \mu_X - \mu_Y = 0$, θεωρούμε πως οι X και Y έχουν την ίδια κατανομή και ο έλεγχος τυχαιοποίησης δίνεται στα παρακάτω βήματα:

1. Επιλέγουμε **χωρίς επανάθεση** B δείγματα μεγέθους $n+m$ από τυχαία αντιμετάθεση των τιμών στο κοινό δείγμα $\{x_1, \ldots, x_n, y_1, \ldots, y_m\}$. Οι πρώτες n τιμές είναι για το δείγμα της X και οι υπόλοιπες m για το δείγμα της Y. Τα σύνολο των B δειγμάτων είναι

$$\{[\mathbf{x}\ \mathbf{y}]^{*1}, \ldots, [\mathbf{x}\ \mathbf{y}]^{*B}\}.$$

2. Υπολογίζουμε B επαναλήψεις τυχαίας αντιμετάθεσης του στατιστικού $\hat{\delta} \equiv \bar{x} - \bar{y}$. Το σύνολο των B επαναλήψεων είναι

$$\{(\bar{x} - \bar{y})^{*1}, \ldots, (\bar{x} - \bar{y})^{*B}\}.$$

3. Βρίσκουμε τη θέση (rank) r του στατιστικού $\bar{x}-\bar{y}$ του αρχικού δείγματος στη λίστα των $B+1$ τιμών του στατιστικού

$$\{\bar{x} - \bar{y}, (\bar{x} - \bar{y})^{*1}, \ldots, (\bar{x} - \bar{y})^{*B}\}$$

αφού πρώτα έχουν διαταχθεί τα στοιχεία της λίστας σε αύξουσα σειρά.

4. Η θέση r ορίζει την απόφαση του ελέγχου σε επίπεδο σημαντικότητας a:

   - Αν $r < (B+1)a/2$ ή $r > (B+1)(1-a/2)$, απορρίπτεται η H_0 και συμπεραίνουμε πως οι μέσες τιμές μ_X και μ_Y διαφέρουν και άρα οι X και Y δεν ακολουθούν την ίδια κατανομή.
   - Αν $(B+1)a/2 \leq r \leq (B+1)(1-a/2)$, δεν απορρίπτεται η H_0 και συμπεραίνουμε πως οι μέσες τιμές μ_X και μ_Y δε διαφέρουν. Αυτό βέβαια δεν είναι ικανή ένδειξη να θεωρήσουμε πως οι X και Y ακολουθούν την ίδια κατανομή, καθώς μπορεί να υπάρχει διαφορά σε κάποιο άλλο χαρακτηριστικό (παράμετρο) της κατανομής.

Στον έλεγχο τυχαιοποίησης (τυχαίας αντιμετάθεσης) επιλέγουμε τα στοιχεία σε κάθε νέο δείγμα μεγέθους $n+m$ χωρίς επανάθεση. Μπορούμε να επιλέξουμε τα στοιχεία **με επανάθεση** δηλαδή με τη μέθοδο bootstrap. Όπως είδαμε οι επαναλήψεις bootstrap του στατιστικού σχηματίζουν την εμπειρική κατανομή του στατιστικού. Στην περίπτωση που εξετάζουμε εδώ, αυτή είναι η κατανομή της διαφοράς $X-Y$. Ο **έλεγχος bootstrap** γίνεται με ακριβώς τον ίδιο τρόπο όπως και ο έλεγχος τυχαιοποίησης με τη διαφορά πως η επιλογή των στοιχείων του δείγματος γίνεται με επανάθεση.

Στον παραμετρικό έλεγχο ως στατιστικό ελέγχου ορίστηκε ο κατάλληλος μετασχηματισμός του στατιστικού $\bar{x}-\bar{y}$ για να ακολουθεί την κατανομή student. Στον έλεγχο τυχαιοποίησης και στον έλεγχο bootstrap δε χρειάζεται να εφαρμόσουμε τον μετασχηματισμό και χρησιμοποιούμε ως στατιστικό ελέγχου απευθείας το $\bar{x}-\bar{y}$. Μπορούμε βέβαια να θεωρήσουμε τον μετασχηματισμό και να χρησιμοποιήσουμε το στατιστικό $t \equiv \frac{(\bar{x}-\bar{y})-(\mu_X-\mu_Y)}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$.

Ασκήσεις Κεφαλαίου 3

1. Η κατανομή **Poisson** χρησιμοποιείται αντί της διωνυμικής όταν ο αριθμός m των επαναλήψεων των δοκιμών είναι μεγάλος και η πιθανότητα 'επιτυχίας' p σε κάθε δοκιμή είναι μικρή και τότε το γινόμενο $\lambda = mp$ ορίζει το μέσο αριθμό επιτυχιών. Η κατανομή **Poisson** χρησιμοποιείται επίσης για να περιγράψει την εμφάνιση πλήθους γεγονότων (επιτυχιών) σε ένα χρονικό διάστημα ή γενικότερα στο πεδίο αναφοράς, π.χ. αριθμός διακοπών σύνδεσης δικτύου σε μια μέρα, αριθμός καμμένων εικονοστοιχείων (pixel) σε μια οθόνη. Συμβολίζοντας X την τ.μ. του αριθμού εμφανίσεων των γεγονότων ενδιαφέροντος (επιτυχιών), η σππ της κατανομής Poisson είναι

$$f_X(x;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!} \tag{3.25}$$

όπου x είναι το πλήθος εμφανίσεων 'επιτυχιών', $x! = 1 \cdot 2 \cdots x$ είναι το παραγοντικό του x και λ η παράμετρος της κατανομής.

(α΄) Έστω τυχαίο δείγμα ανεξάρτητων παρατηρήσεων $\{x_1, \ldots, x_n\}$ από κατανομή Poisson με άγνωστη παράμετρο λ. Δείξετε ότι ο εκτιμητής μέγιστης πιθανοφάνειας του λ είναι η δειγματική μέση τιμή.

(β΄) Φτιάξτε μια συνάρτηση στο `matlab` που θα δημιουργεί M δείγματα μεγέθους n από κατανομή Poisson με δεδομένη παράμετρο λ και θα υπολογίζει τη δειγματική μέση τιμή για κάθε ένα από τα M δείγματα. Στη συνέχεια θα κάνει κατάλληλο ιστόγραμμα των M δειγματικών μέσων τιμών και θα δίνει ως έξοδο το μέσο όρο από τις M δειγματικές μέσες τιμές. Καλέστε τη συνάρτηση για διαφορετικούς συνδυασμούς των M, n και λ. Είναι πάντα (και σύμφωνα με το αποτέλεσμα στο υποερώτημα 1α΄) το κέντρο της κατανομής της δειγματικής μέσης τιμής (που περιγράφεται από το ιστόγραμμα) στην τιμή του λ;

Βοήθεια (`matlab`*):* Για τη δημιουργία των τυχαίων αριθμών χρησιμοποίησε τη συνάρτηση `poissrnd`.

2. Στην ανάλυση αξιοπιστίας (reliability analysis) στη μηχανική εξετάζεται συχνά ο ρυθμός αποτυχίας (failure rate) που συνήθως συμβολίζεται με λ. Έστω ο χρόνος ζωής του συστήματος X και έστω ότι η διαδικασία που ορίζει το χρόνο ζωής του συστήματος δεν έχει μνήμη. Αυτό σημαίνει πως η πιθανότητα να εμφανιστεί αποτυχία σε λιγότερο από κάποιο χρόνο s δεν εξαρτάται από το χρόνο που λειτουργούσε το σύστημα ως τώρα t,

$P(X < t + s|t) = P(X < s)$. Τότε ο χρόνος ζωής X ακολουθεί εκθετική κατανομή με παράμετρο λ και σππ

$$f_X(x;\lambda) = \lambda e^{-\lambda x}, \tag{3.26}$$

Επαναλάβετε τα ερωτήματα 1α΄ και 1β΄ για την εκθετική κατανομή.

Βοήθεια (`matlab`*):* Για τη δημιουργία των τυχαίων αριθμών χρησιμοποίησε τη συνάρτηση `exprnd`.

3. Σε συνέχεια της προηγούμενης άσκησης, προσομοιώστε το χρόνο ζωής n μηχανικών συστημάτων δημιουργώντας n τιμές από εκθετική κατανομή με μέσο χρόνο ζωής $1/\lambda = 15$ μήνες. Με βάση αυτό το δείγμα υπολογίστε το 95% παραμετρικό διάστημα εμπιστοσύνης για το μέσο χρόνο ζωής και εξετάστε αν περιέχεται σε αυτό η τιμή $1/\lambda = 15$.

   (α΄) Υπολογίστε $M = 1000$ δείγματα μεγέθους $n = 5$. Σε τι ποσοστό βρίσκεται ο πραγματικός μέσος χρόνος ζωής μέσα στο 95% διάστημα εμπιστοσύνης;

   (β΄) Κάνετε το ίδιο για $M = 1000$ αλλά $n = 100$. Διαφέρει το ποσοστό αυτό από το παραπάνω;

   Βοήθεια (`matlab`*):* Για τον υπολογισμό διαστήματος εμπιστοσύνης και ελέγχου για τη μέση τιμή με χρήση της κατανομής Student κάλεσε τη συνάρτηση `ttest`.

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

   (β΄) Από παλιότερες μετρήσεις είχαμε βρει πως η τυπική απόκλιση της τάσης διακοπής παρόμοιου κυκλώματος ήταν περίπου 5 kV. Με βάση το δείγμα κάνετε έλεγχο για την υπόθεση πως αυτή είναι η τυπική απόκλιση της τάσης διακοπής.

   (γ΄) Βρείτε 95% διάστημα εμπιστοσύνης για τη μέση τάση διακοπής του κυκλώματος.

(δ΄) Μπορούμε να αποκλείσουμε ότι η μέση τάση διακοπής είναι 52 kV;

(ε΄) Κάνετε έλεγχο χ^2 καλής προσαρμογής σε κανονική κατανομή και βρείτε την p-τιμή του ελέγχου.

Βοήθεια (`matlab`*):* Για τον υπολογισμό διαστήματος εμπιστοσύνης και ελέγχου για τη διασπορά με χρήση της κατανομής χ^2 κάλεσε τη συνάρτηση `vartest`. Για τον έλεγχο χ^2 καλής προσαρμογής κάλεσε τη συνάρτηση `chi2gof`.

5. Ο θερμοπίδακας Old Faithful στην Αμερική είναι από τους πιο γνωστούς θερμοπίδακες για το μέγεθος αλλά και την κανονικότητα των εξάρσεων του (eruptions)
(δες `http://en.wikipedia.org/wiki/Old_Faithful`).
Στο αρχείο δεδομένων `eruption.dat` στην ιστοσελίδα του μαθήματος δίνονται στην πρώτη και δεύτερη στήλη 298 μετρήσεις (σε λεπτά) του διαστήματος αναμονής (waiting time) και της διάρκειας του ξεσπάσματος (duration) για το 1989 και στην τρίτη στήλη 298 μετρήσεις του διαστήματος αναμονής έξαρσης για το 2006. Για κάθε ένα από τα τρία μετρούμενα μεγέθη κάνετε τα παρακάτω.

   (α΄) Βρείτε 95% διάστημα εμπιστοσύνης για την τυπική απόκλιση του μεγέθους και ελέγξτε αν είναι 10′ για την αναμονή και 1′ για τη διάρκεια.

   (β΄) Βρείτε 95% διάστημα εμπιστοσύνης για τη μέση τιμή του μεγέθους και ελέγξτε αν είναι 75′ για την αναμονή και 2.5′ για τη διάρκεια.

   (γ΄) Κάνετε έλεγχο χ^2 καλής προσαρμογής σε κανονική κατανομή και βρείτε την p-τιμή του ελέγχου.

   Με βάση τις 298 μετρήσεις για το χρόνο αναμονής και διάρκειας έξαρσης το 1989, εξετάστε αν μπορείτε να δεχθείτε τον παρακάτω ισχυρισμό (αντιγραφή από τη διεύθυνση της Wikipedia): "With an error of 10 minutes, Old Faithful will erupt 65 minutes after an eruption lasting less than 2.5 minutes or 91 minutes after an eruption lasting more than 2.5 minutes."

6. Θεωρείστε δείγμα μεγέθους $n = 10$ από τ.μ. $X \sim N(0, 1)$.

   (α΄) Δημιουργείστε $B = 1000$ δείγματα bootstrap από το αρχικό δείγμα και υπολογίστε το μέσο όρο τους. Σχηματίστε το ιστόγραμμα του μέσου όρου $\bar{x}$ από τα δείγματα bootstrap (σχεδιάστε στο ίδιο σχήμα και το μέσο όρο του αρχικού δείγματος).

(β΄) Υπολογίστε την εκτίμηση bootstrap $\hat{\text{se}}_B(\bar{x})$ του τυπικού σφάλματος του $\bar{x}$ από τα ίδια $B = 1000$ δείγματα bootstrap. Συγκρίνετε την εκτίμηση αυτή με αυτήν του τυπικού σφάλματος $\hat{\text{se}}(\bar{x})$ του $\bar{x}$ με βάση το αρχικό δείγμα.

(γ΄) Επαναλάβετε τα δύο παραπάνω σημεία για το δείγμα των τιμών που προκύπτουν από το αρχικό δείγμα με το μετασχηματισμό $y = e^x$, δηλαδή για δείγμα από την μεταβλητή Y, όπου $Y = e^X$.

7. Δημιουργείστε $M = 100$ δείγματα μεγέθους $n = 10$ από τ.μ. $X \sim N(0, 1)$.

(α΄) Για κάθε ένα από τα M δείγματα κάνετε τα παρακάτω:

i. Υπολογίστε το παραμετρικό 95% δ.ε. για τη μέση τιμή της X.

ii. Υπολογίστε το 95% δ.ε. με τη μέθοδο ποσοστιαίων bootstrap για τη μέση τιμή της X.

Εξετάστε αν συμφωνούν οι δύο τρόποι υπολογισμού δ.ε., π.χ. παρουσιάζοντας ιστογράμματα των άνω και κάτω άκρων των δ.ε. υπολογισμένα με τους δύο τρόπους στα $M = 100$ δείγματα.

(β΄) Θεωρείστε τον μετασχηματισμό $Y = X^2$ και εφαρμόστε τον στις παρατηρήσεις των $M = 100$ δειγμάτων της τ.μ. X. Επαναλάβετε την παραπάνω διαδικασία.

8. Κάνετε τα βήματα της παραπάνω άσκησης για τον υπολογισμό δ.ε. για την τυπική απόκλιση.

9. Δημιουργείστε $M = 100$ δείγματα μεγέθους $n = 10$ από τ.μ. $X \sim N(0, 1)$ και $M = 100$ δείγματα μεγέθους $m = 12$ από τ.μ. $Y \sim N(0, 1)$.

(α΄) Για κάθε ζευγάρι δειγμάτων των X και Y κάνετε τα παρακάτω:

i. Υπολογίστε το παραμετρικό 95% δ.ε. για τη διαφορά μέσων τιμών των X και Y.

ii. Υπολογίστε το 95% δ.ε. με τη μέθοδο ποσοστιαίων bootstrap για τη διαφορά μέσων τιμών των X και Y.

Μετρήστε το ποσοστό που οι μέσες τιμές των X και Y διαφέρουν με τους δύο τρόπους υπολογισμού του δ.ε. διαφοράς μέσων τιμών.

(β΄) Θεωρείστε τον μετασχηματισμό $Y = X^2$ και εφαρμόστε τον στις παρατηρήσεις των $M = 100$ δειγμάτων της τ.μ. X και Y. Επαναλάβετε την παραπάνω διαδικασία.

(γ΄) Επαναλάβετε τα παραπάνω βήματα θεωρώντας πως $Y \sim N(0.2, 1)$.

10. Η μέθοδος bootstrap μπορεί να χρησιμοποιηθεί και στον έλεγχο μέσης τιμής ως εξής. Τα bootstrap δείγματα θα πρέπει να είναι σύμφωνα με τη μηδενική υπόθεση $H_0 : \mu = \mu_0$. Το αρχικό δείγμα $\mathbf{x} = \{x_1, \ldots, x_n\}$ με δειγματική μέση τιμή $\bar{x}$ κεντράρεται και προστίθεται το μ_0

$$\tilde{x}_i = x_i - \bar{x} + \mu_0, \qquad i = 1, \ldots, n$$

ώστε το δείγμα $\tilde{\mathbf{x}} = \{\tilde{x}_1, \ldots, \tilde{x}_n\}$ που δημιουργείται να είναι σύμφωνο με τη μηδενική υπόθεση, να προέρχεται δηλαδή από κατανομή με μέση τιμή μ_0. Η δειγματοληψία με επανάθεση γίνεται στο δείγμα $\tilde{\mathbf{x}}$ και δημιουργούνται M δείγματα bootstrap. Το στατιστικό ελέγχου υπολογίζεται σε κάθε ένα από αυτά τα M δείγματα καθώς και στο αρχικό δείγμα $\mathbf{x}$. Από τη θέση (rank) της τιμής του στατιστικού στο αρχικό δείγμα στη λίστα όλων των $M + 1$ τιμών προκύπτει η απόφαση του ελέγχου, όπως και για τον έλεγχο bootstrap για την ισότητα δύο μέσων τιμών (δες ενότητα 3.4.4).

Να συγκριθεί αυτός ο έλεγχος bootstrap με τον παραμετρικό έλεγχο μέσης τιμής, όπως έγινε για διάστημα εμπιστοσύνης στην Άσκηση 7. Δημιουργείστε $M = 100$ δείγματα μεγέθους $n = 10$ από τ.μ. $X \sim N(0, 1)$.

(α′) Για κάθε ένα από τα M δείγματα κάνετε τα παρακάτω:

i. Υπολογίστε την p-τιμή του παραμετρικού ελέγχου για τη μηδενική υπόθεση πως η μέση τιμή της X είναι 0 και επίσης για τη μηδενική υπόθεση πως η μέση τιμή της X είναι 0.5.

ii. Υπολογίστε την p-τιμή του ελέγχου bootstrap για τις ίδιες μηδενικές υποθέσεις.

Εξετάστε αν συμφωνούν οι αποφάσεις των δύο τύπων ελέγχων υπολογίζοντας το ποσοστό των απορρίψεων στις M επαναλήψεις για συγκεκριμένες τιμές του επιπέδου σημαντικότητας a.

(β′) Θεωρείστε τον μετασχηματισμό $Y = X^2$ και εφαρμόστε τον στις παρατηρήσεις των $M = 100$ δειγμάτων της τ.μ. X. Επαναλάβετε την παραπάνω διαδικασία για τις μηδενικές υποθέσεις πως η μέση τιμής της X είναι 1 και 2.

11. Για τη μηδενική υπόθεση H_0 ισότητας δύο μέσων τιμών παρουσιάστηκαν έλεγχοι bootstrap και τυχαίας αντιμετάθεσης στην ενότητα 3.4.4, που μπορούν να χρησιμοποιηθούν αντί του παραμετρικού ελέγχου (t-test). Να συγκριθούν αυτοί οι δύο έλεγχοι (bootstrap και τυχαίας αντιμετάθεσης) με τον παραμετρικό έλεγχο, όπως έγινε για το διάστημα

εμπιστοσύνης διαφοράς μέσων τιμών στην Άσκηση 9 (για bootstrap και παραμετρικό διάστημα εμπιστοσύνης). Ειδικότερα να εξετάσετε αν συμφωνούν οι αποφάσεις των τριών τύπων ελέγχων (παραμετρικός, bootstrap και τυχαίας αντιμετάθεσης) υπολογίζοντας το ποσοστό των απορρίψεων στις M επαναλήψεις για συγκεκριμένες τιμές του επιπέδου σημαντικότητας a.

12. Σε συνέχεια της παραπάνω Άσκησης 11, θα χρησιμοποιήσουμε ένα δεύτερο έλεγχο bootstrap για τη μηδενική υπόθεση H_0 ισότητας δύο μέσω τιμών. Οι παραπάνω έλεγχοι θεωρούν τη μηδενική υπόθεση πως τα δύο δείγματα προέρχονται από την ίδια κατανομή και άρα οι δύο μέσες τιμές είναι ίσες. Τώρα θα θεωρήσουμε πως οι κατανομές μπορούν να διαφέρουν, π.χ. να έχουν διαφορετικές διασπορές, και η μηδενική υπόθεση H_0 παραμένει η ίδια, δηλαδή οι δύο μέσες τιμές είναι ίσες. Ο αντίστοιχος παραμετρικός έλεγχος είναι και πάλι με βάση την κατανομή Student αλλά για άνισες διασπορές. Συγκεκριμένα ο παραμετρικός έλεγχος θεωρεί το στατιστικό

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2}}$$

όπου $s_{\bar{x}}^2 = s_X^2/n$ είναι η διασπορά του μέσου όρου $\bar{x}$. Το στατιστικό ακολουθεί την κατανομή Student με βαθμούς ελευθερίας (degrees of freedom, dfe)

$$\text{dfe} = \frac{(s_{\bar{x}}^2 + s_{\bar{y}}^2)^2}{s_{\bar{x}}^2/(n-1) + s_{\bar{y}}^2/(m-1)}$$

Για τον έλεγχο bootstrap τα δύο δείγματα μπορούν να είναι από διαφορετικές κατανομές αλλά θα πρέπει να έχουν την ίδια μέση τιμή για να είναι σύμφωνα με την H_0. Υπολογίζουμε πρώτα το μέσο όρο $\bar{z}$ από όλες τις $n + m$ παρατηρήσεις των δειγμάτων $\mathbf{x}$ και $\mathbf{y}$. Το αρχικό δείγμα $\mathbf{x} = \{x_1, \ldots, x_n\}$ κεντράρεται και προστίθεται το $\bar{z}$ και το ίδιο για το $\mathbf{y} = \{y_1, \ldots, y_n\}$

$$\begin{aligned} \tilde{x}_i &= x_i - \bar{x} + \bar{z}, \qquad i = 1, \ldots, n \\ \tilde{y}_i &= y_i - \bar{y} + \bar{z}, \qquad i = 1, \ldots, m \end{aligned}$$

έτσι ώστε τα δείγματα $\tilde{\mathbf{x}} = \{\tilde{x}_1, \ldots, \tilde{x}_n\}$ και $\tilde{\mathbf{y}} = \{\tilde{y}_1, \ldots, \tilde{y}_n\}$ να έχουν και τα δύο μέσο όρο $\bar{z}$ και να είναι έτσι σύμφωνα με την H_0. Στη συνέχεια ακολουθούνται τα ίδια βήματα του ελέγχου bootstrap για την απόφαση του ελέγχου.

Επαναλάβετε την παραπάνω Άσκηση 11 με αυτούς τους δύο ελέγχους, δηλαδή τον παραμετρικό έλεγχο και τον έλεγχο bootstrap για την H_0 ίσων μέσων τιμών χωρίς να θεωρείται ίδια κατανομή για τα δύο δείγματα.

Κεφάλαιο 4

Αβεβαιότητα και σφάλμα μέτρησης

Από την κλασσική αρχαία εποχή που ο Αριστοτέλης θεώρησε τη βεβαιότητα (και αβεβαιότητα) ενός αποτελέσματος ως σήμερα ο χαρακτηρισμός της αβεβαιότητας της μέτρησης είναι ένα θέμα που έχει απασχολήσει τους επιστήμονες. Σήμερα υπάρχουν ερευνητικά κέντρα και οργανισμοί, όπως ο Παγκόσμιος Οργανισμός Μέτρων (International Organization for Standards, ISO), που αναπτύσσουν μεθοδολογίες για τον καθορισμό μέτρων και σταθμών καθώς και τον καθορισμό της αβεβαιότητας των μετρήσεων.

Κάθε φορά που προσπαθούμε να ποσοτικοποιήσουμε μια φυσική διαδικασία, εμφανίζεται η **αβεβαιότητα** που μπορεί να σχετίζεται με το μοντέλο και την προσομοίωση στον υπολογιστή που υποθέτουμε για τη διαδικασία, όπως επίσης και με τη μέτρηση της διαδικασίας. Αντίστοιχα έχουμε λοιπόν την **αβεβαιότητα του μοντέλου ή / και της προσομοίωσης** (model or simulation uncertainty) και την **αβεβαιότητα της μέτρησης** (measurement uncertainty). Ο πρώτος τύπος αβεβαιότητας είναι δύσκολο να προσδιοριστεί αφού η πραγματική φυσική διαδικασία μας είναι άγνωστη και δεν υπάρχουν διεθνώς αναγνωρισμένα μέτρα (standards) για αυτόν τον τύπο αβεβαιότητας. Για παράδειγμα η αβεβαιότητα του μοντέλου εμφανίζεται όταν υποθέτουμε ένα απλουστευμένο μαθηματικό μοντέλο για μια φυσική διαδικασία που περιέχει παράγοντες που δεν έχουμε συμπεριλάβει στο μοντέλο. Σε αυτόν τον τύπο αβεβαιότητας θα πρέπει να συμπεριλάβουμε και την *υπολογιστική (αριθμητική) αβεβαιότητα* (numerical uncertainty) που αναφέρεται στην αριθμητική επίλυση μαθηματικών εξισώσεων. Ο δεύτερος τύπος αβεβαιότητας που αναφέρεται στο πείραμα και τη μέτρηση μπορεί πιο εύκολα να προσδιοριστεί και να προτυποποιηθεί.

Σχετικά με την ονοματολογία για την αβεβαιότητα, στη βιβλιογραφία συνήθως γίνεται διαχωρισμός της αβεβαιότητας και του **σφάλματος** (error) και

αυτό θα ακολουθήσουμε εδώ. Θα δεχτούμε ότι η αβεβαιότητα του μοντέλου εκφράζει την εν δυνάμει ανεπάρκεια του μοντέλου λόγω έλλειψης γνώσης για τη διαδικασία που μελετάμε, ενώ το **σφάλμα του μοντέλου** (modeling error) είναι η αναγνωρισμένη ανεπάρκεια που δίνει το μοντέλο όταν το εφαρμόζουμε στην πράξη. Αντίστοιχα, το **σφάλμα μέτρησης** (measurement error) είναι η διαφορά πραγματικής και παρατηρούμενης τιμής ενώ η αβεβαιότητα μέτρησης είναι η εκτίμηση για το σφάλμα μέτρησης, που ορίζει ένα σύνολο δυνατών τιμών για το σφάλμα για μια συγκεκριμένη μέτρηση. Σε αυτό το κεφάλαιο θα μελετήσουμε την αβεβαιότητα στη μέτρηση και στα επόμενα κεφάλαια θα συζητήσουμε την αβεβαιότητα στο μοντέλο όταν θα μελετήσουμε μοντέλα για φυσικές διαδικασίες διαφόρων μορφών.

4.1 Συστηματικά και τυχαία σφάλματα

Το αποτέλεσμα της μέτρησης ενός φυσικού μεγέθους είναι κατά κανόνα ακαθόριστο σε μικρό ή μεγάλο βαθμό, δηλαδή αν επαναλάβουμε τη μέτρηση δε θα πάρουμε ακριβώς το ίδιο αποτέλεσμα. Ακόμα και αν επαναλάβουμε ένα πείραμα διατηρώντας τις ίδιες ακριβώς συνθήκες, κάποιος παράγοντας που δεν ελέγχουμε μπορεί να επηρεάζει λίγο ή πολύ το αποτέλεσμα της μέτρησης. Θα θέλαμε λοιπόν να εκτιμήσουμε τις πειραματικές αυτές αποκλίσεις στη μέτρηση και το πρώτο βήμα είναι να εντοπίσουμε τους διαφορετικούς τύπους σφάλματος μέτρησης.

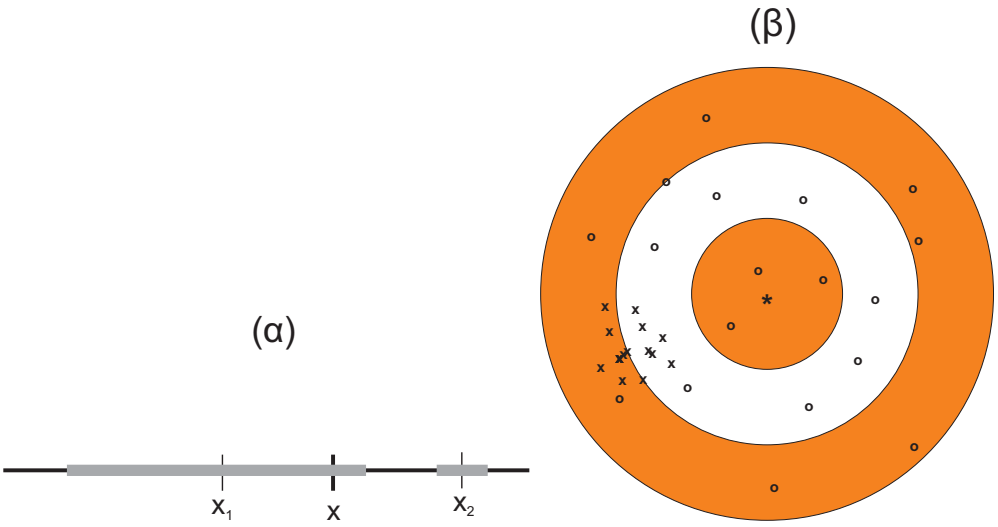
Οι δύο τύποι σφαλμάτων μέτρησης είναι τα **συστηματικά σφάλματα** (systematic errors) και τα **τυχαία σφάλματα** (random errors) που προσδίδουν και αντίστοιχα χαρακτηριστικά στις μετρήσεις.

- Τα συστηματικά σφάλματα επαναλαμβάνονται και υπάρχει κάποιο αίτιο που τα δημιουργεί. Πολλές φορές είναι δύσκολο να εντοπισθούν αλλά μπορούν να εξουδετερωθούν με κατάλληλη *βαθμονόμηση* (calibration), συγκρίνοντας με κάποιον τρόπο μετρήσεις και πραγματικές τιμές. Τα συστηματικά σφάλματα ορίζουν την **ακρίβεια (ορθότητα)** (accuracy) της μέτρησης, δηλαδή κατά πόσο οι μετρήσεις είναι κοντά στις πραγματικές τιμές ή υπάρχουν συστηματικές αποκλίσεις. Με αναφορά στην εκτίμηση παραμέτρων τα συστηματικά σφάλματα συνδέονται με τη μεροληψία (bias), όπου η εκτίμηση του μεγέθους (ή παραμέτρου) δεν είναι ίδια με την πραγματική τιμή του μεγέθους.

- Τα τυχαία σφάλματα δεν επαναλαμβάνονται με το πείραμα αλλά αντιπροσωπεύουν την τυχαιότητα που χαρακτηρίζει το μέγεθος που μετράμε. Για αυτό και αυτού του τύπου τα σφάλματα δε μπορούν να απαλειφθούν. Τα τυχαία σφάλματα ορίζουν την **ακρίβεια επανάληψης**

(precision) της μέτρησης, δηλαδή το μέγεθος της μεταβολής των τιμών μέτρησης σε κάθε επανάληψη της μέτρησης (για τις ίδιες συνθήκες του πειράματος).

Η ορθότητα και η ακρίβεια επανάληψης της μέτρησης δίνονται σχηματικά στο Σχήμα 4.1α και Σχήμα 4.1β σε μια και σε δύο διαστάσεις αντίστοιχα. Στο

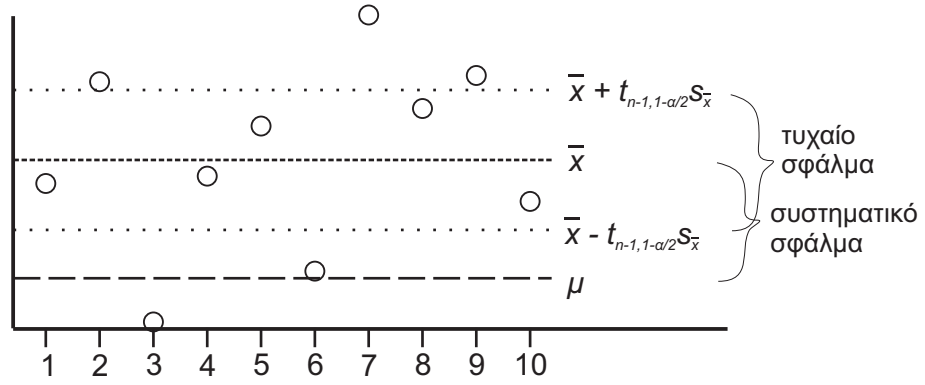


Σχήμα 4.1: Σχηματική παράσταση της ορθότητας και ακρίβειας επανάληψης της μέτρησης σε μια διάσταση στο (α) και σε δύο διαστάσεις στο (β). Στο (α) για το μέγεθος x η μέτρηση x_1 έχει μεγάλο τυχαίο σφάλμα ενώ η μέτρηση x_2 έχει μεγάλο συστηματικό σφάλμα. Στο (β) αντίστοιχα το μέγεθος x είναι στο κέντρο των κύκλων ενώ η μέτρηση με μεγάλο τυχαίο σφάλμα δίνεται με ανοιχτούς κύκλους και με συστηματικό σφάλμα με σύμβολα 'x'.

Σχήμα 4.1α διαγράφονται τα σφάλματα δύο μετρήσεων x_1 και x_2. Η πρώτη μέτρηση x_1 έχει μεγάλο τυχαίο σφάλμα (μικρή ακρίβεια επανάληψης) που καλύπτει το συστηματικό σφάλμα. Η δεύτερη μέτρηση x_2 έχει μικρό τυχαίο σφάλμα που αναδεικνύει μεγάλο συστηματικό σφάλμα (μικρή ορθότητα ή μεγάλη μεροληψία). Τα ίδια χαρακτηριστικά δίνονται στο Σχήμα 4.1β σε δύο διαστάσεις, όπου για μεγάλο τυχαίο σφάλμα οι μετρήσεις είναι διάσπαρτες (ανοιχτοί κύκλοι) γύρω από την πραγματική τιμή στο κέντρο των κύκλων. Αν υποθέσουμε πως το Σχήμα 4.1β απεικονίζει έναν πίνακα σκοποβολής, τότε οι ανοιχτοί κύκλοι δηλώνουν έναν μάλλον άστοχο σκοπευτή. Στη δεύτερη περίπτωση, το συστηματικό σφάλμα είναι μεγάλο και το τυχαίο σφάλμα μικρό και οι μετρήσεις μαζεύονται όχι όμως γύρω από το κέντρο των κύκλων (σύμβολα 'x'). Ο σκοπευτής εδώ είναι ακριβής αλλά σκοπεύει σε λάθος σημείο. Στην πρώτη περίπτωση δε μπορούμε να διορθώσουμε την επιτυχία της σκόπευσης

(ενδεχομένως μπορούμε να εκπαιδεύσουμε τον σκοπευτή να σκοπεύει με μεγαλύτερη ακρίβεια αλλά αυτό είναι άλλο θέμα). Στη δεύτερη περίπτωση όμως μπορούμε να εντοπίσουμε το αίτιο της σκόπευσης σε λάθος σημείο, π.χ. λανθασμένος εστίαση του οργάνου σκόπευσης, και να το διορθώσουμε. Αυτή η διαδικασία αναφέρεται ως βαθμοσκόπηση.

Ας δούμε τις έννοιες της ορθότητας μέτρησης και ακρίβειας επανάληψης μέτρησης στην περίπτωση της εκτίμησης της μέσης τιμής ενός μεγέθους, δηλαδή μιας τ.μ. X. Έστω ότι έχουμε 10 παρατηρήσεις (μετρήσεις) της τ.μ. X. Αν υπάρχει συστηματικό σφάλμα στη διαδικασία μέτρησης τότε υπάρχει μεροληψία στην εκτίμηση της μέσης τιμής και η δειγματική μέση τιμή (μέσος όρος των 10 παρατηρήσεων) $\bar{x}$ αποκλίνει από την πραγματική μέση τιμή. Σε αυτήν την περίπτωση το διάστημα εμπιστοσύνης (για κάποιο επίπεδο σημαντικότητας a) μπορεί να μην περιέχει την πραγματική μέση τιμή μ με πιθανότητα μεγαλύτερη του a, όπως στο Σχήμα 4.2. Το εύρος του διαστήματος εμπιστοσύνης καθορίζει σε αυτήν την περίπτωση την ακρίβεια επανάληψης για τη μέση τιμή μ και η απόσταση του $\bar{x}$ από την μ την ορθότητα.



Σχήμα 4.2: Σχηματική παράσταση της ορθότητας (μεροληψίας) και ακρίβειας επανάληψης στην εκτίμηση της μέσης τιμής από 10 παρατηρήσεις.

Αν έχουμε συλλέξει n μετρήσεις μιας τ.μ. X η αβεβαιότητα της μέτρησης είναι η εκτίμηση του σφάλματος μέτρησης που δίνεται από την τυπική απόκλιση s (δες την έκφραση της διασποράς s^2 στην (3.3)). Κάνοντας χρήση της κρίσιμης τιμής από την κατανομή Student και κάτω από προϋποθέσεις (X από κανονική κατανομή ή μεγάλο n), το *όριο της ακρίβειας επανάληψης* (random uncertainty) για κάθε (επόμενη) μέτρηση σε επίπεδο σημαντικότητας a είναι

$$\bar{x} \pm t_{n-1,1-a/2}s.$$

Αντίστοιχα, η αβεβαιότητα για την εκτίμηση της μέσης τιμής μ είναι η εκτίμη-

ση του σταθερού σφάλματος του μέσου όρου $s_{\bar{x}} = s/\sqrt{n}$ και το όριο της ακρίβειας για τη μέση τιμή είναι σε επίπεδο σημαντικότητας a

$$\bar{x} \pm t_{n-1,1-a/2} s/\sqrt{n}.$$

Παράδειγμα 4.1. Η διαφορά τάσης (voltage difference) V σε έναν αντιστάτη (resistor) μετρήθηκε 10 φορές και έδωσε τις παρακάτω τιμές (σε mV)

i	1	2	3	4	5	6	7	8	9	10
V_i	123.5	125.3	124.1	123.9	123.7	124.2	123.2	123.7	124.0	123.2

Η μέση διαφορά τάσης υπολογίζεται από τις 10 μετρήσεις ως

$$\bar{V} = \frac{1}{10}\sum_{i=1}^{10} V_i = 123.880 \text{ mV}$$

Υποθέτοντας ότι η διαφορά τάσης V είναι τ.μ. που ακολουθεί κανονική κατανομή, η αβεβαιότητα για κάθε μέτρηση V_i είναι

$$s_V = \sqrt{\frac{1}{10-1}\sum_{i=1}^{n}(V_i - \bar{V})^2} = 0.607 \text{ mV}.$$

και θα έχουμε

$V_1 = (123.5 \pm 0.6)\text{mV},\ V_2 = (125.3 \pm 0.6)\text{mV},\ \ldots,\ V_{10} = (123.2 \pm 0.6)\text{mV}.$

Η αβεβαιότητα για το $\bar{V}$ είναι

$$s_{\bar{V}} = \frac{s_V}{\sqrt{10}} = 0.192 \text{ mV}$$

και θα έχουμε

$$\bar{V} = (123.880 \pm 0.192) \text{ mV}.$$

Το όριο αβεβαιότητας σε επίπεδο σημαντικότητας $a = 0.05$ για κάθε νέα μέτρηση θα είναι

$$\begin{aligned} \bar{V} \pm t_{n-1,1-a/2} s_V &= 123.88 \pm t_{9,0.975} \cdot 0.607 = 123.88 \pm 2.2622 \cdot 0.607 \\ &= 123.88 \pm 1.373 \text{ mV}, \end{aligned}$$

ενώ το όριο αβεβαιότητας για τη μέση τιμή μ είναι το 95% διάστημα εμπιστοσύνης

$$\bar{V} \pm t_{n-1,1-a/2} s_V/\sqrt{n} = 123.88 \pm 2.2622 \cdot 0.607/\sqrt{10} = 123.88 \pm 0.434 \text{ mV}.$$

4.2 Διάδοση σφάλματος μέτρησης

Ας υποθέσουμε τώρα πως έχουμε μετρήσεις ενός φυσικού μεγέθους X που είναι τυχαία μεταβλητή, δηλαδή γνωρίζουμε το X με κάποια αβεβαιότητα σ_X που είναι η τυπική του απόκλιση. Έστω ότι μας ενδιαφέρει ένα άλλο φυσικό μέγεθος Y που δίνεται ως συνάρτηση του X, $Y = f(X)$. Η μεταβολή του Y, $\mathrm{d}Y$, για κάθε μικρή μεταβολή $\mathrm{d}X$ γύρω από κάποια τιμή x δίνεται ως

$$\mathrm{d}Y \simeq \left(\frac{\mathrm{d}f}{\mathrm{d}X}\right)_{X=x} \mathrm{d}X,$$

όπου η προσέγγιση του αναπτύγματος Ταψλορ έγινε μόνο ως την πρώτη τάξη (πρώτη παράγωγο). Θεωρώντας την μεταβολή $\mathrm{d}X = x - \bar{x}$ και $\mathrm{d}Y = y - \bar{y}$, όπου $\bar{x}$ και $\bar{y}$ οι μέσοι όροι των παρατηρήσεων της X και Y αντίστοιχα, παίρνουμε την παρακάτω έκφραση για την αβεβαιότητα στην Y

$$\sigma_Y^2 \simeq \left(\frac{\mathrm{d}f}{\mathrm{d}X}\right)_{X=x}^2 \sigma_X^2 \;\Leftrightarrow\; \sigma_Y \simeq \left|\frac{\mathrm{d}f}{\mathrm{d}X}\right|_{X=x} \sigma_X. \tag{4.1}$$

Η απόλυτη τιμή χρησιμοποιείται για να δίνει πάντα θετική τιμή στην αβεβαιότητα σ_Y, ακόμα και αν η παράγωγος της f είναι αρνητική.

Η παραπάνω σχέση μπορεί να επεκταθεί και όταν η τ.μ. Y που μας ενδιαφέρει είναι συνάρτηση m παρατηρούμενων τ.μ. $X_1, X_2, \ldots, X_m$ και δίνεται ως

$$\sigma_Y^2 \simeq \sum_{i=1}^{m} \left(\frac{\mathrm{d}f}{\mathrm{d}X_i}\right)_{X_i=x_i}^2 \sigma_{X_i}^2 + 2 \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \left(\frac{\mathrm{d}f}{\mathrm{d}X_i}\right)_{X_i=x_i} \left(\frac{\mathrm{d}f}{\mathrm{d}X_j}\right)_{X_j=x_j} \sigma_{X_i,X_j}, \tag{4.2}$$

όπου $x_1, x_2, \ldots, x_m$ είναι οι μετρήσεις των $X_1, X_2, \ldots, X_m$ αντίστοιχα και σ_{X_i,X_j} είναι η συνδιασπορά των X_i και X_j. Ο παραπάνω τύπος ονομάζεται **νόμος διάδοσης των σφαλμάτων** (law of propagation of errors) και δίνει μια εκτίμηση της μέγιστης αβεβαιότητας σ_Y της Y για δεδομένες αβεβαιότητες $\sigma_{X_1}, \sigma_{X_2}, \ldots, \sigma_{X_m}$. Αυτός ο νόμος είναι αυστηρά ακριβής όταν η συνάρτηση f είναι γραμμική (η ανάπτυξη της f κατά Taylor συμπίπτει με την ίδια την f). Η f μπορεί να έχει μη-γραμμική μορφή που δεν επιτρέπει καλή προσέγγιση της αβεβαιότητας σ_Y αν δεν επεκταθεί το ανάπτυγμα Taylor σε μεγαλύτερη τάξη.

Όταν οι τ.μ. $X_1, X_2, \ldots, X_m$ είναι ανεξάρτητες, το δεύτερο διπλό άθροισμα στη σχέση (4.2) απαλείφεται αφού οι συνδιασπορές μηδενίζονται και η αβεβαιότητα δίνεται ως

$$\sigma_Y \simeq \sqrt{\sum_{i=1}^{m} \left(\frac{\mathrm{d}f}{\mathrm{d}X_i}\right)_{X_i=x_i}^2 \sigma_{X_i}^2}. \tag{4.3}$$

Ειδικότερα ας θεωρήσουμε ότι η f είναι γραμμική, δηλαδή η Y είναι γραμμικός συνδυασμός των $X_1, X_2, \ldots, X_m$

$$Y = \sum_{i=1}^{m} a_i X_i = a^{\mathrm{T}} X$$

όπου $a_1, a_2, \ldots, a_m$ είναι οι συντελεστές του γραμμικού συνδυασμού και τα διανύσματα είναι $a = [a_1, a_2, \ldots, a_m]^{\mathrm{T}}$, $X = [X_1, X_2, \ldots, X_m]^{\mathrm{T}}$. Γενικά για συσχετισμένες τ.μ. $X_1, X_2, \ldots, X_m$, ο πίνακας συνδιασποράς είναι

$$\Sigma = \begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1,X_2} & \ldots & \sigma_{X_1,X_m} \\ \sigma_{X_2,X_1} & \sigma_{X_2}^2 & \ldots & \sigma_{X_2,X_m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_m,X_1} & \sigma_{X_m,X_2} & \ldots & \sigma_{X_m}^2 \end{bmatrix}.$$

Η διασπορά της Y είναι

$$\sigma_Y^2 = \sum_{i=1}^{m} \sum_{i=1}^{m} a_i \sigma_{X_i,X_j} a_j = a^{\mathrm{T}} \Sigma a.$$

Για κάθε δύο διαφορετικές τ.μ. X_i και X_j ο συντελεστής συσχέτισης είναι $\rho_{X_i,X_j} = \sigma_{X_i,X_j}/(\sigma_{X_i}\sigma_{X_j})$ και η παραπάνω έκφραση μπορεί να γραφεί ως

$$\sigma_Y^2 = \sum_{i=1}^{m} a_i \sigma_{X_i}^2 + 2 \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} a_i a_j \rho_{X_i,X_j} \sigma_{X_i} \sigma_{X_j}.$$

Όταν οι τ.μ. $X_1, X_2, \ldots, X_m$ είναι ανεξάρτητες η παραπάνω σχέση απλοποιείται ως

$$\sigma_Y^2 = \sum_{i=1}^{m} a_i \sigma_{X_i}^2.$$

Συχνά χρησιμοποιείται ο συμβολισμός ΔX αντί σ_X για την αβεβαιότητα της X. Επίσης χρησιμοποιείται η **σχετική αβεβαιότητα** (relative uncertainty σ_X/X και συχνά σε ποσοστιαίες μονάδες.

Παράδειγμα 4.2. Ένα απλό και πρακτικό παράδειγμα διάδοσης σφάλματος είναι ο νόμος του Ωμ $R = V/I$, όπου μετρώντας την ένταση του ρεύματος I και την τάση V σε έναν αντιστάτη θέλουμε να προσδιορίσουμε την αντίσταση R.

Αν υποθέσουμε ότι οι παρατηρούμενη ένταση και τάση του ρεύματος δίνονται με αβεβαιότητα σ_I και σ_V αντίστοιχα, η αβεβαιότητα σ_R είναι

$$\sigma_R = \sqrt{\left(\frac{\sigma_V}{I}\right)^2 + \left(\frac{V}{I^2}\sigma_I\right)^2} = R\sqrt{\left(\frac{\sigma_V}{V}\right)^2 + \left(\frac{\sigma_I}{I}\right)^2}.$$

Σε αυτήν την απλή περίπτωση, η σχετική αβεβαιότητα σ_R/R είναι η τετραγωνική ρίζα του αθροίσματος των τετραγώνων των σχετικών αβεβαιοτήτων της έντασης και τάσης.

Ασκήσεις Κεφαλαίου 4

1. Ο συντελεστής αποκατάστασης e (coefficient of restitution) μιας μπάλας, όπου εδώ έχει το νόημα της αναπήδησης, ορίζεται από το λόγο της ταχύτητας μετά την επαφή με την επιφάνεια προς την ταχύτητα πριν την επαφή. Ισοδύναμα ο συντελεστής αποκατάστασης ορίζεται ως $e = \sqrt{h_2/h_1}$, όπου h_1 είναι το ύψος από το οποίο αφήνουμε τη μπάλα και h_2 είναι το ύψος που φτάνει η μπάλα μετά την επαφή με την επιφάνεια (έδαφος).

   (α΄) Αφήνουμε μια μπάλα μπάσκετ από ύψος 100 cm και μετράμε το ύψος που φτάνει μετά την αναπήδηση. Επαναλαμβάνουμε αυτό το πείραμα 5 φορές και βρίσκουμε (σε cm) 60, 54, 58, 60, 56. Για μια σωστά φουσκωμένη μπάλα θα πρέπει ο συντελεστής αναπήδησης να είναι 0.76. Εκτιμήστε την αβεβαιότητα ορθότητας και ακρίβειας επανάληψης για το συντελεστή αναπήδησης για αυτήν τη μπάλα.

   (β΄) Προσομοιώστε $M = 1000$ φορές το παραπάνω πείραμα. Σε κάθε πείραμα δημιουργείστε τις 5 μετρήσεις του ύψους h_2 μετά την αναπήδηση της μπάλας από κανονική κατανομή με μέση τιμή $\mu_2 = 58\text{cm}$ και τυπική απόκλιση $\sigma_2 = 2\text{cm}$. Υπολογίστε σε κάθε προσομοίωση το μέσο όρο και την τυπική απόκλιση του ύψους h_2 καθώς και του συντελεστή αποκατάστασης e. Εξετάστε αν οι τιμές αυτές είναι συνεπείς με τις αναμενόμενες από τη σχέση $e = \sqrt{h_2/h_1}$ για σταθερό $h_1 = 100\text{cm}$.

   (γ΄) Αφήνουμε τη μπάλα μπάσκετ 5 φορές από διαφορετικά αρχικά ύψη h_1 και βρίσκουμε τις παρακάτω τιμές (σε cm)

h_1	80	100	90	120	95
h_2	48	60	50	75	56

   Εκτιμήστε την αβεβαιότητα στα δύο ύψη και υπολογίστε την αβεβαιότητα στο συντελεστή αναπήδησης. Αν ο συντελεστής αναπήδησης πρέπει να είναι 0.76, μπορούμε να δεχθούμε ότι η μπάλα είναι κατάλληλα φουσκωμένη;

2. Στη μέτρηση αγροτεμαχίων σε παραλληλόγραμμο σχήμα μετράμε το μήκος και πλάτος με αβεβαιότητα 5m.

   (α΄) Ποια είναι η αβεβαιότητα στη μέτρηση της επιφάνειας αγροτεμαχίου με μήκος 500m και πλάτος 300m; Για ποια μήκη και πλάτη αυτή η αβεβαιότητα είναι ίδια;

(β΄) Σχηματίστε την αβεβαιότητα ως συνάρτηση του μήκους και πλάτους της έκτασης αγροτεμαχίου.

Βοήθεια (`matlab`*):* Για το σχηματισμό επιφάνειας χρησιμοποίησε την εντολή `surf`.

3. Η ισχύς που εκλύεται (is dissipated) από ένα κύκλωμα εναλλασσόμενου ρεύματος δίνεται ως $P = VI\cos f$, όπου V και I είναι η ημιτονοειδής τάση και ένταση ρεύματος στο κύκλωμα και f είναι η διαφορά φάσης μεταξύ των V και I.

(α΄) Θεωρώντας ότι τα V, I και f δε συσχετίζονται υπολογίστε την αβεβαιότητα της ισχύος από τις τιμές των V, I και f και την αβεβαιότητα τους.

(β΄) Θεωρείστε ότι η ακρίβεια των V, I και f δίνεται ως (μέση τιμή και τυπική απόκλιση)

$$\begin{aligned} V &= (77.78 \pm 0.71)\text{V} \\ I &= (1.21 \pm 0.071)\text{A} \\ f &= (0.283 \pm 0.017)\text{rad} \end{aligned}$$

Θεωρώντας κανονική κατανομή για κάθε μια από τις V, I και f, ελέγξετε με $M = 1000$ προσομοιώσεις αν η ισχύς έχει την αναμενόμενη ακρίβεια.

(γ΄) Κάνετε το ίδιο με παραπάνω αλλά θεωρώντας επιπλέον πως η διαφορά φάσης συσχετίζεται με την τάση με συντελεστή συσχέτισης $\rho_{V,f} = 0.5$.

Βοήθεια (`matlab`*):* Για τη δημιουργία τυχαίων αριθμών από πολυμεταβλητή κανονική κατανομή χρησιμοποίησε την εντολή `mvnrnd`.

Κεφάλαιο 5

Συσχέτιση και Παλινδρόμηση

Στο προηγούμενο κεφάλαιο μελετήσαμε τη διάδοση του σφάλματος από μια τυχαία μεταβλητή X σε μια τ.μ. Y που δίνεται ως συνάρτηση της X. Σε αυτό το κεφάλαιο θα διερευνήσουμε τη συσχέτιση των δύο μεταβλητών X και Y και θα εκτιμήσουμε τη συνάρτηση που δίνει τη Y γνωρίζοντας τη X. Επίσης θα επεκτείνουμε τη μελέτη για την περίπτωση που θέλουμε να εκτιμήσουμε τη Y γνωρίζοντας περισσότερες από μια μεταβλητές.

Πρώτα θα μελετήσουμε τη σχέση δύο τ.μ. X και Y. Συχνά στη μελέτη ενός τεχνικού συστήματος ή φυσικού φαινομένου 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 και 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 μας ενδιαφέρει η μεταβολή της μέσης τιμής (και ορισμένες φορές και η διασπορά), για αυτό και 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 πρώτα την απλή γραμμική παλινδρόμηση, δηλαδή όταν η συνάρτηση παλινδρόμησης είναι γραμμική ως προς μια τ.μ. X. Στη συνέχεια θα μελετήσουμε κάποιες μορφές μη-γραμμικής παλινδρόμησης, καθώς και πολλαπλής παλινδρόμησης, όπου η Y εξαρτάται από περισσότερες από μια μεταβλητές.

5.1 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 και οι δύο από άλλες μεταβλητές, όπως η ταχύτητα του κινητήρα. Η συσχέτιση λοιπόν δεν υποδηλώνει απαραίτητα κάποια αιτιατή σχέση των δύο μεταβλητών. Η παρατήρηση ότι σε κάποια περιοχή ο πληθυσμός των πελαργών συσχετίζεται με το πλήθος των γεννήσεων, δε σημαίνει πως οι πελαργοί προκαλούν γεννήσεις.

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. Στην παράγραφο 2.2.4 ορίσαμε τη συνδιασπορά σ_{XY} και το συντελεστή συσχέτισης ρ δύο τ.μ. X και Y με διασπορά σ_X^2 και σ_Y^2 αντίστοιχα (δες τη σχέση (2.16) για το σ_{XY} και (2.17) για το ρ). Ο συντελεστής συσχέτισης $\rho = \mathrm{Corr}(X, Y)$ αποτελεί κανονικοποίηση της συνδιασποράς σ_{XY} και 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. Ο συντελεστής συσχέτισης λέγεται και συντελεστής συσχέτισης Pearson για να το διαχωρίσουμε από άλλους συντελεστές συσχέτισης, όπως του Spearman και του Kendall.

Όταν $\rho = \pm 1$ η σχέση είναι αιτιοκρατική και 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 ισχυρή τη συσχέτιση όταν $|\rho| > 0.9$) ενώ όταν είναι κοντά στο μηδέν οι τ.μ. είναι πρακτικά ασυσχέτιστες.

Όπως φαίνεται από τον ορισμό στη σχέση (2.17), ο συντελεστής συσχέτισης ρ δεν εξαρτάται από τη μονάδα μέτρησης των X και Y και είναι συμμετρικός ως προς τις X και Y.

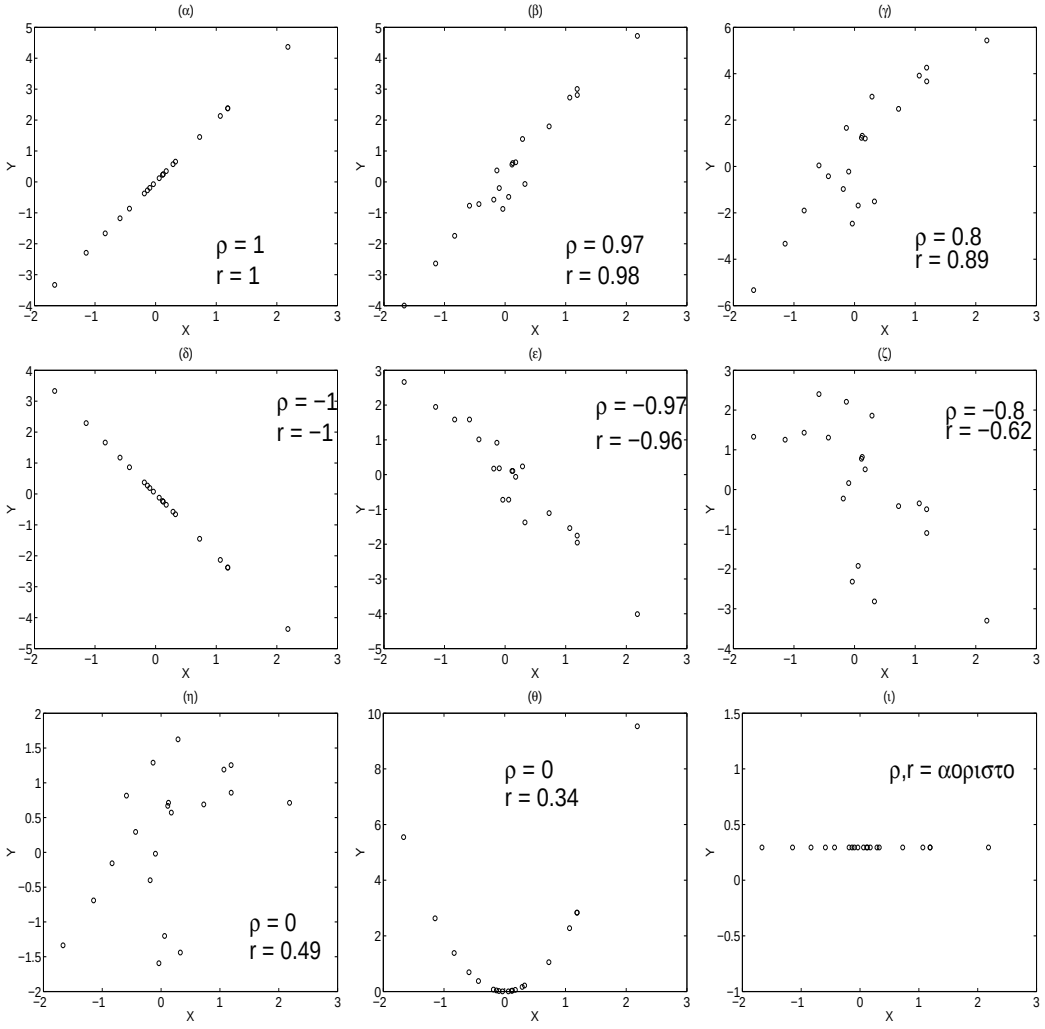
5.1.1 Δειγματικός συντελεστής συσχέτισης

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 πάρουμε μια πρώτη εντύπωση για τη συσχέτιση τους 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 5.1 παρουσιάζονται τυπικά διαγράμματα διασποράς για ισχυρές και ασθενείς συσχετίσεις δύο τ.μ. X και Y. Στα Σχήματα 5.1α και 5.1δ



Σχήμα 5.1: 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 ρ και η δειγματική r. Στο (ι) ο συντελεστής συσχέτισης δεν ορίζεται.

η σχέση είναι τέλεια ($\rho = 1$ και $\rho = -1$ αντίστοιχα), στα Σχήματα 5.1β και 5.1ε είναι ισχυρή (θετική με $\rho = 0.97$ και αρνητική με $\rho = -0.97$ αντίστοιχα)

και στα Σχήματα 5.1γ και 5.1ζ είναι λιγότερο ισχυρή (θετική με $\rho = 0.8$ και αρνητική με $\rho = -0.8$ αντίστοιχα). Στο Σχήμα 5.1η είναι $\rho = 0$ γιατί οι τ.μ. X και Y είναι ανεξάρτητες ενώ στο Σχήμα 5.1θ είναι πάλι $\rho = 0$ αλλά οι X και Y δεν είναι ανεξάρτητες αλλά συσχετίζονται μόνο μη-γραμμικά. Τέλος για το Σχήμα 5.1ι ο συντελεστής συσχέτισης δεν ορίζεται γιατί η Y είναι σταθερή ($\sigma_Y = 0$ στον ορισμό του ρ στην (2.17)).

Η **σημειακή εκτίμηση** (point estimation) του συντελεστή συσχέτισης ρ του πληθυσμού από το δείγμα των n ζευγαρωτών παρατηρήσεων των X και Y γίνεται με την αντικατάσταση στη σχέση (2.17) της συνδιασποράς σ_{XY} και των τυπικών αποκλίσεων σ_X και σ_Y από τις αντίστοιχες εκτιμήσεις από το δείγμα

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

Ο αμερόληπτος εκτιμητής s_{XY} του σ_{XY} δίνεται όπως και ο αμερόληπτος εκτιμητής της συνδιασποράς σ_X^2 (δες σχέση 3.3) ως

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right), \tag{5.1}$$

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

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}. \tag{5.2}$$

Είναι προφανές πως η παραπάνω σχέση για το r δεν αλλάζει αν θεωρήσουμε τους μεροληπτικούς εκτιμητές των σ_{XY}, σ_X και σ_Y. Το r λέγεται **δειγματικός συντελεστής συσχέτισης (του Pearson** (sample Pearson correlation coefficient).

Καλύτερη φυσική ερμηνεία της συσχέτισης δύο τ.μ. επιτυγχάνεται με το r^2 που λέγεται **συντελεστής προσδιορισμού** (coefficient of determination) (εκφράζεται συνήθως σε ποσοστό, δηλαδή $100r^2$). Ο συντελεστής προσδιορισμού δίνει το ποσοστό μεταβλητότητας των τιμών της Y που υπολογίζεται από τη X (και αντίστροφα) και είναι ένας χρήσιμος τρόπος να συνοψίσουμε τη σχέση δύο τ.μ..

Στο Σχήμα 5.1 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 ρ. Αυτό συμβαίνει γιατί ο εκτιμητής r όπως δίνεται στη σχέση (5.2) είναι μια τ.μ. που εξαρτάται από τις τιμές και το πλήθος των ζευγών των παρατηρήσεων.

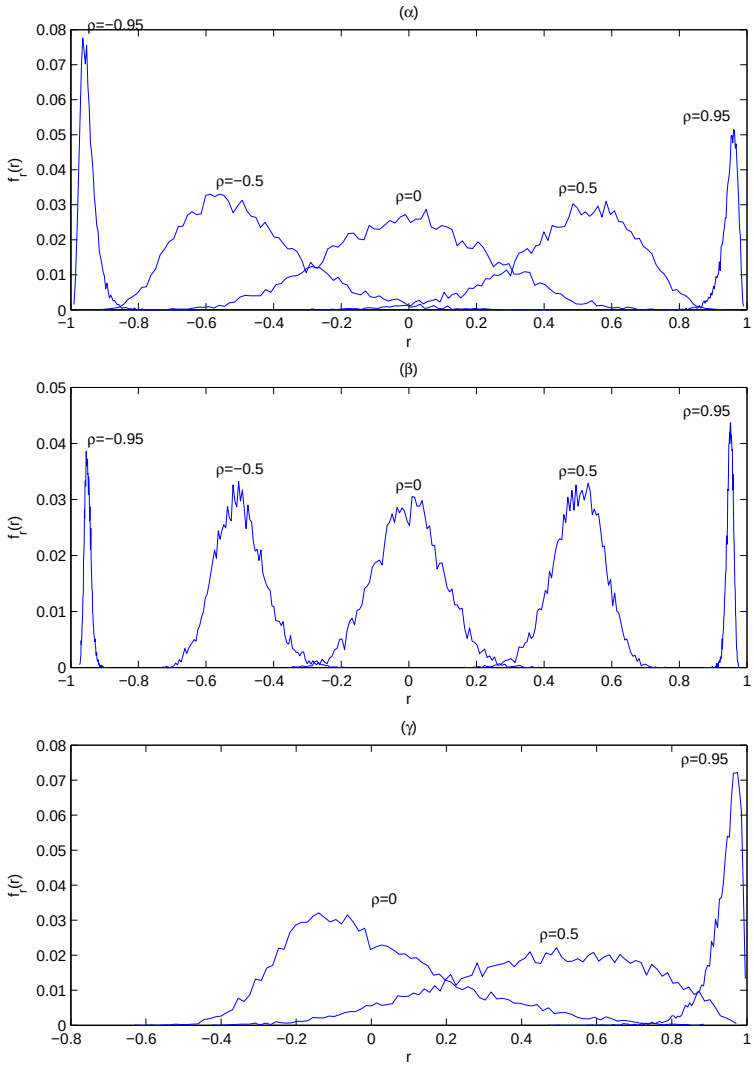
5.1.2 Κατανομή του εκτιμητή r

Η κατανομή του εκτιμητή r δε μπορεί εύκολα να περιγραφτεί καθώς το r δεν έχει την ίδια κατανομή για κάθε τιμή του πραγματικού συντελεστή ρ. Η κατανομή του r δεν είναι κάποια γνωστή, όπως κανονική ή άλλη κατανομή που φθίνει εκθετικά ή με νόμο δύναμης, αφού φράζεται από το -1 και το 1. Για $\rho = 0$, η πυκνότητα των τιμών του r κατανέμεται συμμετρικά γύρω από το 0, και η μέση τιμή του είναι 0. Καθώς όμως το ρ αποκλίνει προς το -1 ή το 1, η κατανομή του r γίνεται θετικά ή αρνητικά ασύμμετρη, αντίστοιχα. Όταν το ρ πλησιάζει το -1 ή το 1, η διασπορά του r μικραίνει για να γίνει 0 όταν το ρ ισούται με -1 ή 1. Η μορφή της κατανομής του r για κάθε τιμή του ρ εξαρτάται από το μέγεθος του δείγματος n αλλά και από την κατανομή των τ.μ. X και Y.

Η κατανομή του r δίνεται αναλυτικά όταν το ζεύγος τ.μ. (X, Y) ακολουθεί διμεταβλητή κανονική κατανομή και υπάρχουν επίσης προσεγγιστικές αναλυτικές εκφράσεις όταν η διμεταβλητή κατανομή δεν είναι κανονική, αλλά όλες αυτές οι αναλυτικές εκφράσεις είναι αρκετά σύνθετες και δεν παρουσιάζονται εδώ. Στο Σχήμα 5.2 δίνεται η εμπειρική κατανομή του r για διαφορετικές τιμές του ρ, που σχηματίστηκε ως εξής. Για σταθερές μέσες τιμές και τυπικές αποκλίσεις των X και Y και για κάθε τιμή του συντελεστή συσχέτισης ρ, πραγματοποιούνται M δείγματα μεγέθους n και σε καθένα από αυτά υπολογίζεται ο δειγματικός συντελεστής συσχέτισης r. Από τις M τιμές του r σχηματίζεται το κανονικοποιημένο ιστόγραμμα (για σχετικές συχνότητες) που προσομοιώνει τη συνάρτηση πυκνότητας πιθανότητας του r, $f_r(r)$. Στο Σχήμα 5.2α οι παράμετροι είναι $\mu_X = 0$, $\mu_Y = 0$, $\sigma_X = 1$, $\sigma_Y = 1$, $M = 10000$ και $n = 20$. Παρατηρούμε πως η μορφή της εμπειρικής $f_r(r)$ διαφέρει για τις διαφορετικές τιμές του ρ και εμφανίζει συμμετρία μόνο για $\rho = 0$. Στο Σχήμα 5.2β, όπου το μέγεθος των δειγμάτων πενταπλασιάστηκε ($n = 100$), η $f_r(r)$ γίνεται πιο συμμετρική και με μικρότερη διασπορά (όπως αναμένεται για μεγαλύτερο δείγμα). Σημειώνεται ότι η σύγκλιση της $f_r(r)$ στην κανονική κατανομή που περιμένουμε σύμφωνα με το Κεντρικό Οριακό Θεώρημα είναι πιο αργή για μεγαλύτερες τιμές του $|\rho|$. Για παράδειγμα για τιμές του ρ κοντά στα όρια -1 και 1, η λοξότητα 'διορθώνεται' όταν η κατανομή γίνεται πολύ στενή, δηλαδή για μεγάλα δείγματα.

Η μορφή της $f_r(r)$ αλλάζει για το ίδιο ρ όταν το δείγμα δεν προέρχεται από διμεταβλητή κανονική κατανομή. Υποθέτοντας τα τετράγωνα των κανονικών τ.μ. X και Y, $X' = X^2$ και $Y' = Y^2$, μπορεί να δειχθεί ότι $\rho' = \mathrm{Corr}(X', Y') = \rho^2$. Σε αυτήν την περίπτωση η μορφή της $f_r(r)$ για την εκτίμηση του ρ' διαφέρει, όπως προκύπτει από τη σύγκριση των Σχημάτων 5.2α και 5.2γ.

Σε πολλά προβλήματα της στατιστικής και ανάλυσης δεδομένων προσπα-



Σχήμα 5.2: (α) Συνάρτηση πυκνότητας πιθανότητας του r για διαφορετικές τιμές του ρ, εκτιμώμενη από το ιστόγραμμα των τιμών του r υπολογισμένο σε 10000 ζευγαρωτά δείγματα μεγέθους $n = 20$ από κανονική κατανομή με μέσες τιμές 0 και διασπορά 1 και για τις δύο τ.μ. X και Y. (β) Όπως στο (α) αλλά για $n = 100$. (γ) Όπως το (α) αλλά μόνο για θετικό συντελεστή ρ των τετραγώνων των X και Y. Σε αυτήν την περίπτωση αυτός ο συντελεστής συσχέτισης είναι το τετράγωνο του συντελεστή συσχέτισης των X και Y.

θούμε να μετασχηματίσουμε την παρατηρούμενη μεταβλητή ώστε να ακολουθεί κάποια γνωστή κατανομή, συνήθως κανονική. Εδώ η μεταβλητή είναι ο

εκτιμητής r. Προτάθηκε από τον Fisher ο μετασχηματισμός

$$z = \tanh^{-1}(r) = 0.5 \ln \frac{1+r}{1-r}, \tag{5.3}$$

ώστε όταν το δείγμα είναι μεγάλο και από διμεταβλητή κανονική κατανομή το z να τείνει προς την κανονική κατανομή με μέση τιμή $\mu_z \equiv \mathrm{E}(z) = \tanh^{-1}(\rho)$ και διασπορά $\sigma_z^2 \equiv \mathrm{Var}(z) = 1/(n-3)$, δηλαδή η διασπορά είναι ανεξάρτητη του ρ. Μπορούμε λοιπόν να υπολογίσουμε διάστημα εμπιστοσύνης και να κάνουμε έλεγχο υπόθεσης χρησιμοποιώντας την προσεγγιστικά κανονική κατανομή του z.

5.1.3 Διάστημα εμπιστοσύνης για το συντελεστή συσχέτισης

Για τον υπολογισμό παραμετρικού διαστήματος εμπιστοσύνης για το ρ, το πρώτο βήμα είναι ο μετασχηματισμός του εκτιμητή r στο z από τη σχέση (5.3). Σύμφωνα και με το διάστημα εμπιστοσύνης για τη μέση τιμή μιας τ.μ., υποθέτοντας ότι το z ακολουθεί προσεγγιστικά κανονική κατανομή, το $(1-a)\%$ διάστημα εμπιστοσύνης για το z δίνεται ως

$$z \pm z_{1-a/2} \sqrt{1/(n-3)}.$$

Το δεύτερο βήμα λοιπόν είναι να υπολογίσουμε από την παραπάνω σχέση αυτό το διάστημα και ας το συμβολίσουμε $[\zeta_l, \zeta_u]$. Τέλος, στο τρίτο βήμα παίρνουμε τον αντίστροφο μετασχηματισμό για τα άκρα του διαστήματος ζ_l και ζ_u, που δίνονται ως

$$r_l = \tanh(\zeta_l) = \frac{\exp(2\zeta_l) - 1}{\exp(2\zeta_l) + 1}, \quad r_u = \frac{\exp(2\zeta_u) - 1}{\exp(2\zeta_u) + 1} \tag{5.4}$$

και ορίζουν το $(1-a)\%$ διαστήματος εμπιστοσύνης για το ρ.

5.1.4 Έλεγχος μηδενικής συσχέτισης

Σε πολλά προβλήματα μας ενδιαφέρει να ελέγξουμε αν δύο τ.μ. συσχετίζονται. Ένας τρόπος να το επιτύχουμε είναι από τον υπολογισμό του διαστήματος εμπιστοσύνης του ρ. Αν το $(1-a)\%$ διαστήματος εμπιστοσύνης του ρ, $[r_l, r_u]$ (δες σχέση (5.4)), δεν περιέχει το 0, τότε μπορούμε να δεχθούμε στο ίδιο επίπεδο εμπιστοσύνης ότι οι δύο τ.μ. συσχετίζονται.

Αν θέλουμε να κάνουμε έλεγχο υπόθεσης για το $\rho = 0$ τότε μπορούμε να χρησιμοποιήσουμε την κατανομή του r κάτω από τη μηδενική υπόθεση H_0:

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 5.1: Δεδομένα αντίστασης x_i και χρόνου αποτυχίας y_i για 20 αντιστάτες.

$\rho = 0$. Για $\rho = 0$, ισχύει

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}, \tag{5.5}$$

δηλαδή η τ.μ. t που προκύπτει από τον παραπάνω μετασχηματισμό του r ακολουθεί κατανομή Student με $n-2$ βαθμούς ελευθερίας. Άρα η απόφαση του ελέγχου γίνεται με βάση το στατιστικό t της σχέσης (5.5) και η p-τιμή του ελέγχου μπορεί να υπολογισθεί από την αθροιστική κατανομή Student για την τιμή του στατιστικού από το δείγμα.

Παράδειγμα 5.1. Θέλουμε να διερευνήσουμε τη συσχέτιση της αντίστασης και του χρόνου αποτυχίας κάποιου υπερφορτωμένου αντιστάτη. Για αυτό πήραμε μετρήσεις αντίστασης (σε Ωμ, ohm) και χρόνου αποτυχίας (σε λεπτά, min) από δείγμα 20 αντιστατών, οι οποίες παρουσιάζονται στον Πίνακα 5.1.

Για να βρούμε τον συντελεστή συσχέτισης r υπολογίζουμε πρώτα τα πα-

ρακάτω

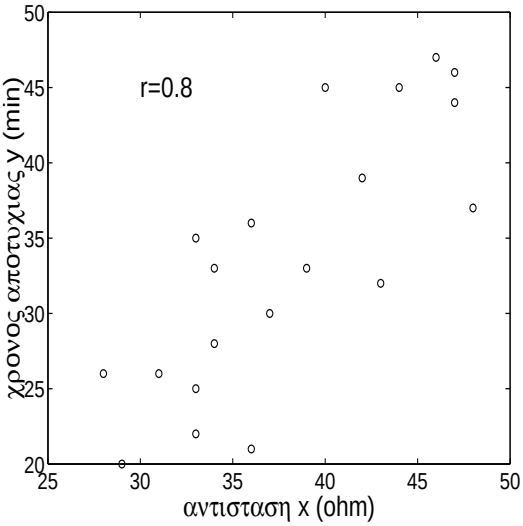
$$\bar{x} = 38 \qquad \bar{y} = 33.5$$

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

Από τη σχέση (5.2) 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 και ο χρόνος αποτυχίας αντιστάτη έχουν γραμμική θετική συσχέτιση αλλά όχι πολύ ισχυρή. Αυτό φαίνεται και από το διάγραμμα διασποράς στο Σχήμα 5.3. Η μεταβλητότητα της μιας τ.μ. (αντίσταση ή χρόνος αποτυχίας) μπορεί να



Σχήμα 5.3: Διάγραμμα διασποράς για το δείγμα παρατηρήσεων αντίστασης και χρόνου αποτυχίας 20 αντιστατών του Πίνακα 5.1.

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.

Για τον υπολογισμό του 95% διαστήματος εμπιστοσύνης του συντελεστή συσχέτισης ρ υπολογίζουμε πρώτα το μετασχηματισμό Fisher του r από τη

σχέση (5.3), $z = 1.110$. Τα άκρα του 95% διαστήματος εμπιστοσύνης του z είναι $[0.634, 1.585]$ και ο αντίστροφος μετασχηματισμός στα άκρα αυτού του διαστήματος δίνει το 95% διάστημα εμπιστοσύνης του ρ ως $[0.561, 0.919]$. Τα όρια αυτά δηλώνουν ότι για τόσο μικρό δείγμα δε μπορούμε να εκτιμήσουμε με ακρίβεια το συντελεστή συσχέτισης.

Το κάτω άκρο του 95% διαστήματος εμπιστοσύνης του ρ είναι 0.561 και είναι πολύ μεγαλύτερο του 0, άρα μπορούμε να ισχυριστούμε με μεγάλη σιγουριά πως η αντίσταση και ο χρόνος αποτυχίας συσχετίζονται. Ο έλεγχος της υπόθεσης $\rho = 0$ δίνει το Student στατιστικό $t = 5.736$ (δες σχέση (5.5)) και $p = 0.0000194$, δηλαδή είναι εντελώς απίθανο η αντίσταση και ο χρόνος αποτυχίας να είναι ανεξάρτητες τ.μ. με βάση το δείγμα των 20 ζευγαρωτών παρατηρήσεων.

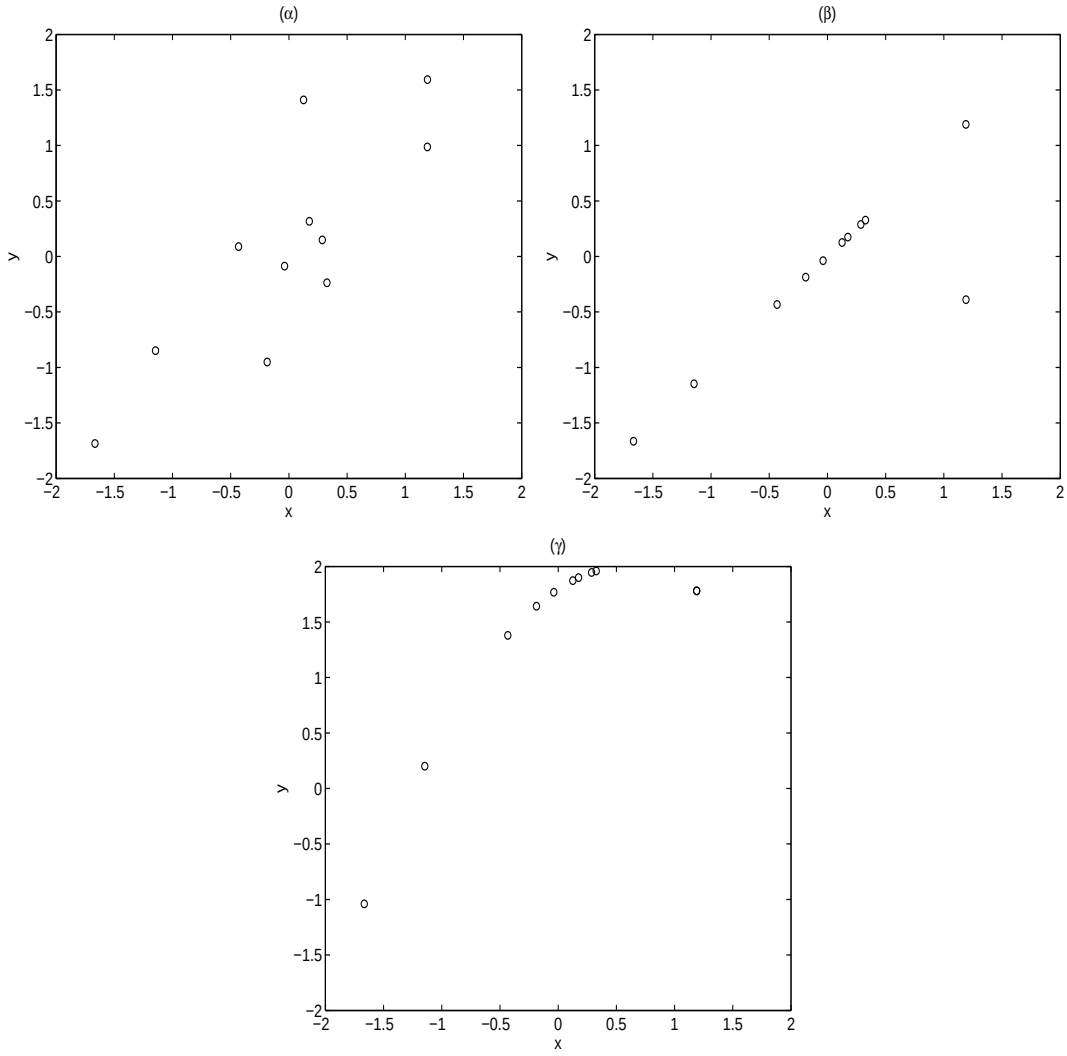
Πρέπει τέλος να σημειωθεί ότι η εκτίμηση r του συντελεστή συσχέτισης μπορεί να αλλάξει σημαντικά με την πρόσθεση ή αφαίρεση λίγων παρατηρήσεων γιατί το μέγεθος του δείγματος είναι μικρό.

5.1.5 Συσχέτιση και γραμμικότητα

Ο συντελεστής συσχέτισης Pearson ρ καθώς και ο εκτιμητής του r αναφέρονται στο μέγεθος της γραμμικής συσχέτισης μεταξύ δύο τ.μ. X και Y. Για ένα δείγμα του ζεύγους (X, Y), η τιμή του r δεν αποδίδει πάντα σωστά τη συσχέτιση, καθώς η συσχέτιση τους μπορεί να μήν είναι απλά γραμμική.

Στο Σχήμα 5.4 δίνονται τα διαγράμματα διασποράς για τρία δείγματα του ζεύγους (X, Y). Ο δειγματικός συντελεστής συσχέτισης για τα τρία δείγματα είναι ο ίδιος, $r = 0.84$. Για το πρώτο δείγμα στο Σχήμα 5.4α, το ζεύγος (X, Y) είναι από διμεταβλητή κανονική κατανομή και άρα το r αποδίδει σωστά το μέγεθος της συσχέτισης τους. Το δεύτερο και τρίτο δείγμα όμως δεν αντιστοιχεί σε γραμμικά συσχετισμένα X και Y. Είναι φανερό πως για το δεύτερο δείγμα στο Σχήμα 5.4β η σχέση των X και Y είναι απόλυτα γραμμική ($\rho = 1$) για όλα τα ζευγάρια εκτός από ένα απόμακρο σημείο (outlier), το οποίο μειώνει την εκτίμηση στο $r = 0.84$. Στο τρίτο δείγμα στο Σχήμα 5.4γ η συσχέτιση των X και Y είναι απόλυτη αλλά μη-γραμμική με αποτέλεσμα η συσχέτιση τους πάλι να υποεκτιμάται.

Από τα παραπάνω παραδείγματα είναι φανερό πως ο συντελεστής συσχέτισης Pearson δεν είναι πάντα κατάλληλο μέτρο συσχέτισης. Για την αντιμετώπιση απόμακρων σημείων άλλα μέτρα είναι πιο κατάλληλα, όπως ο εκτιμητής του συντελεστή συσχέτισης Spearman και Kendall. Όταν η συσχέτιση μπορεί να είναι μη-γραμμική άλλα μέτρα πρέπει να αναζητηθούν. Ένα τέτοιο μέτρο είναι η **αμοιβαία πληροφορία** (mutual information) που μετράει την πληροφορία που μπορούμε να έχουμε για τη μια τ.μ. γνωρίζοντας την άλλη.



Σχήμα 5.4: Διαγράμματα διασποράς δειγμάτων που δίνουν όλα τον ίδιο δειγματικό συντελεστή συσχέτισης $r = 0.84$. (α) Τα X και Y είναι από διμεταβλητή κανονική κατανομή. (β) $Y = X$ για όλα εκτός από ένα ζευγάρι παρατηρήσεων του δείγματος. (γ) $Y = 2 - 0.6(X - 0.585)^2$.

5.2 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 στη συνέχεια σχεδιάζουμε την εξάρτηση μιας τ.μ. Y, που την ονομάζουμε **εξαρτημένη μεταβλητή** (dependent variable), από άλλες τυχαίες μεταβλητές που τις ονομάζουμε **ανεξάρτητες μεταβλητές** (independent

variables). Γενικά θεωρούμε πως τις τιμές των ανεξάρτητων μεταβλητών τις ορίζουμε εμείς και δεν εμπεριέχουν σφάλματα, δεν είναι δηλαδή τυχαίες μεταβλητές. Σε πολλά πρακτικά προβλήματα όμως αυτό είναι περισσότερο μια παραδοχή για να εφαρμόσουμε τη μέθοδο των ελαχίστων τετραγώνων για να εκτιμήσουμε το μοντέλο παλινδρόμησης που δίνει τη μαθηματική έκφραση της εξάρτησης της εξαρτημένης από τις ανεξάρτητες μεταβλητές.

Στην αρχή θα θεωρήσουμε την εξάρτηση της Y από μία μόνο ανεξάρτητη μεταβλητή X και η περίπτωση αυτή αναφέρεται ως *απλή παλινδρόμηση*. Ενώ η συσχέτιση είναι συμμετρική ως προς τα X και Y, στην απλή παλινδρόμηση η εξαρτημένη μεταβλητή Y 'καθοδηγείται' από την ανεξάρτητη μεταβλητή X. Για αυτό και στην ανάλυση που κάνουμε παίζει ρόλο ποιόν από τους δύο παράγοντες που μετράμε ορίζουμε ως ανεξάρτητη μεταβλητή και ποιόν ως 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.

Στο πρώτο μέρος της μελέτης θα θεωρήσουμε πως η εξάρτηση είναι γραμμική, δηλαδή έχουμε την περίπτωση της *απλής γραμμικής παλινδρόμησης*.

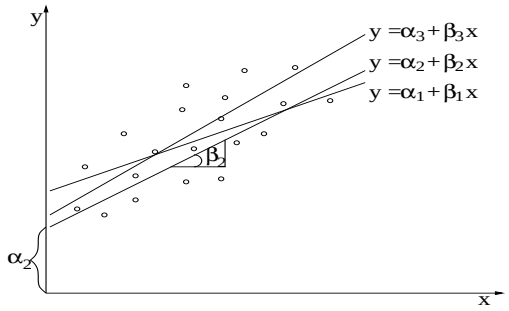
5.2.1 Το πρόβλημα της απλής γραμμικής παλινδρόμησης

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

$$\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x \tag{5.6}$$

και η σχέση αυτή λέγεται **απλή γραμμική παλινδρόμηση της** Y **στη** X (linear regression). Το πρόβλημα της παλινδρόμησης είναι η εύρεση των παραμέτρων β_0 και β_1 που εκφράζουν καλύτερα τη γραμμική εξάρτηση της Y από τη X. Κάθε ζεύγος τιμών (β_0, β_1) καθορίζει μια διαφορετική γραμμική σχέση που εκφράζεται γεωμετρικά από ευθεία γραμμή. Η διαφοράς ύψους β_0 είναι η τιμή του y για $x = 0$ και λέγεται **διαφορά ύψους** (intercept) κι ο συντελεστής του x, β_1, 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 5.5.



Σχήμα 5.5: Ευθείες απλής γραμμικής παλινδρόμησης

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 ως 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

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{5.7}$$

όπου ϵ_i είναι και αυτή τ.μ., λέγεται **σφάλμα παλινδρόμησης** (regression error) και 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:

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

- Η σχέση (5.6) 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

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

δηλαδή η διασπορά της εξαρτημένης μεταβλητής Y είναι η ίδια για κάθε τιμή της X και μάλιστα είναι $\sigma_{Y|X}^2 = \sigma_\epsilon^2$. Η τελευταία σχέση προκύπτει από τη σχέση (5.7), αφού οι παράμετροι β_0 και β_1 είναι σταθερές και το x_i γνωστό. Η ιδιότητα αυτή λέγεται *ομοσκεδαστικότητα* και αντίθετα έχουμε *ετεροσκεδαστικότητα* όταν η διασπορά της Y (ή του σφάλματος ϵ) 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 παραμετρικά διαστήματα εμπιστοσύνης για τις παραμέτρους ή να κάνουμε παραμετρικούς στατιστικούς ελέγχους 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

$$Y|X = x \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma_\epsilon^2).$$

Αν η κατανομή της Y δεν είναι κανονική, μπορούμε να χρησιμοποιήσουμε κάποιο μετασχηματισμό που την κάνει κανονική και για αυτό συχνά χρησιμοποιείται ο λογάριθμος.

5.2.2 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 β_0,

2. της κλίσης της ευθείας παλινδρόμησης β_1,

3. της διασποράς σφάλματος της παλινδρόμησης σ_ϵ^2.

Τα β_0 και β_1 προσδιορίζουν την ευθεία παλινδρόμησης και άρα καθορίζουν τη γραμμική σχέση εξάρτησης της τ.μ. Y από τη μεταβλητή X. Η παράμετρος σ_ϵ^2 προσδιορίζει το βαθμό μεταβλητότητας γύρω από την ευθεία παλινδρόμησης και εκφράζει την αβεβαιότητα της γραμμικής σχέσης.

Εκτίμηση των παραμέτρων της ευθείας παλινδρόμησης

Η εκτίμηση των παραμέτρων β_0 και β_1 γίνεται με τη μέθοδο των **ελαχίστων τετραγώνων** (method of least squares). Η μέθοδος λέγεται έτσι γιατί βρίσκει την ευθεία παλινδρόμησης με παραμέτρους b_0 και b_1 έτσι ώστε το άθροισμα των τετραγώνων των κατακόρυφων αποστάσεων των σημείων από την ευθεία να είναι το ελάχιστο. Οι εκτιμήσεις των β_0 και β_1 δίνονται από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \epsilon_i^2 \qquad \text{ή} \qquad \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{5.8}$$

Για να λύσουμε αυτό το πρόβλημα θέτουμε τις μερικές παραγώγους ως προς τα β_0 και β_1 ίσες με το μηδέν και καταλήγουμε στο σύστημα δύο εξισώσεων με δύο αγνώστους

$$\left.\begin{array}{l} \frac{\partial \sum (y_i-\beta_0-\beta_1 x_i)^2}{\partial \beta_0} = 0 \\ \frac{\partial \sum (y_i-\beta_0-\beta_1 x_i)^2}{\partial \beta_1} = 0 \end{array}\right\} \begin{array}{l} \sum_{i=1}^n y_i = n\beta_0 + \beta_1 \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i = \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 \end{array}$$

από το οποίο με αντικατάσταση του β_0 παίρνουμε την εκτίμηση για την κλίση

$$\hat{\beta}_1 \equiv b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}. \tag{5.9}$$

Το άθροισμα που συμβολίζεται στην παραπάνω σχέση ως S_{xy} διαιρώντας με $n-1$ δίνει τη δειγματική συνδιασπορά s_{XY} (δες σχέση (5.1)) και αντίστοιχα το S_{xx} διαιρώντας με $n-1$ δίνει τη δειγματική διασπορά της X s_{XY} (δες σχέση (3.3)). Με αντικατάσταση του b_1 στην πρώτη εξίσωση του παραπάνω συστήματος παίρνουμε την εκτίμηση του σταθερού όρου ως

$$\hat{\beta}_0 \equiv b_0 = \frac{\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i}{n} \tag{5.10}$$

και άρα οι εκτιμήσεις των β_0 και β_1 δίνονται ως

$$b_1 = \frac{s_{XY}}{s_X^2}, \qquad b_0 = \bar{y} - b_1 \bar{x}. \tag{5.11}$$

Τα b_0 και b_1 ορίζουν την ευθεία

$$\hat{y} = b_0 + b_1 x,$$

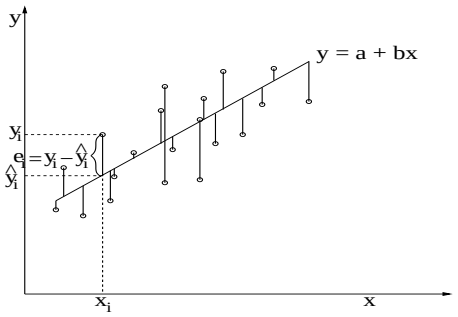
που λέγεται και **ευθεία ελαχίστων τετραγώνων**.

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

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$$

είναι η κατακόρυφη απόσταση της πραγματικής τιμής από την ευθεία ελαχίστων τετραγώνων και λέγεται σφάλμα ελαχίστων τετραγώνων ή απλά **υπόλοιπο** (residual). Στο Σχήμα 5.6 απεικονίζονται τα υπόλοιπα της παλινδρόμησης.



Σχήμα 5.6: Ευθεία ελαχίστων τετραγώνων και υπόλοιπα

Το υπόλοιπο e_i είναι η εκτίμηση του σφάλματος παλινδρόμησης ϵ_i αντικαθιστώντας απλά τις παραμέτρους παλινδρόμησης με τις εκτιμήσεις ελαχίστων τετραγώνων στον ορισμό του σφάλματος $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$. Άρα η εκτίμηση της διασποράς σ_ϵ^2 του σφάλματος (που είναι και η δεσμευμένη διασπορά της Y ως προς X) δίνεται από τη δειγματική διασπορά s_ϵ^2 των υπολοίπων e_i

$$s_\epsilon^2 \equiv \hat{\sigma_\epsilon^2} = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{5.12}$$

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_ϵ^2 μπορεί να εκφραστεί ως προς τις δειγματικές διασπορές των X και Y και της συνδιασποράς τους, αν αντικαταστήσουμε τις εκφράσεις των b_0 και b_1 από την (5.11) στην παραπάνω σχέση (όπου θέτουμε $\hat{y}_i = b_0 + b_1 x_i$)

$$s_\epsilon^2 = \frac{n-1}{n-2}\left(s_Y^2 - \frac{s_{XY}^2}{s_X^2}\right) = \frac{n-1}{n-2}(s_Y^2 - b_1^2 s_X^2) \tag{5.13}$$

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

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

   Άρα η ευθεία ελαχίστων τετραγώνων μπορεί επίσης να οριστεί ως

   $$y_i - \bar{y} = b_1(x_i - \bar{x}).$$

2. Η σημειακή εκτίμηση των β_0 και β_1 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 b_0 και b_1 είναι οι εκτιμήτριες μεγίστης πιθανοφάνειας (και 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.

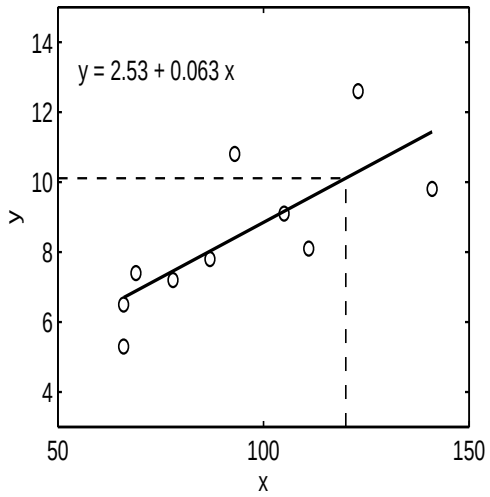
4. Για κάθε τιμή x της X μπορούμε να *προβλέψουμε* την αντίστοιχη τιμή y της Y από την ευθεία ελαχίστων τετραγώνων, $\hat{y} = b_0 + b_1 x$. Εδώ πρέπει να προσέξουμε ότι η τιμή x πρέπει να ανήκει στο εύρος τιμών της X που έχουμε από το δείγμα. Για τιμές έξω από αυτό το διάστημα η πρόβλεψη δεν είναι αξιόπιστη.

Παράδειγμα 5.2. Θέλουμε να μελετήσουμε σε ένα ολοκληρωμένο κύκλωμα την εξάρτηση της απολαβής ρεύματος κρυσταλλολυχνίας (τρανζίστορ) από την αντίσταση του στρώματος της κρυσταλλολυχνίας. Στον Πίνακα 5.2 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 5.2: 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 στρώματος της και το διάγραμμα διασποράς από το δείγμα στο Σχήμα 5.7 επιβεβαιώνει αυτήν την υπόθεση. Για να



Σχήμα 5.7: Διάγραμμα διασποράς για τα δεδομένα του Πίνακα 5.2 και ευθεία ελαχίστων τετραγώνων.

εκτιμήσουμε τις παραμέτρους b_0 και b_1 της ευθείας ελαχίστων τετραγώνων

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 (5.1) 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_1 και b_0 είναι

$$\begin{aligned} b_1 &= \frac{41.81}{662.1} = 0.063 \\ b_0 &= 8.46 - 0.063 \cdot 93.9 = 2.53. \end{aligned}$$

Από τη σχέση (5.13) υπολογίζουμε την εκτίμηση διασποράς των σφαλμάτων παλινδρόμησης

$$s_\epsilon^2 = \frac{9}{8}(4.66 - 0.063^2 \cdot 41.81) = 2.271.$$

Τα αποτελέσματα αυτά ερμηνεύονται ως εξής:

1. b_1: Για αύξηση της αντίστασης στρώματος κατά μία μονάδα μέτρησης (1 ohm/cm) η απολαβή του ρεύματος της κρυσταλλολυχνίας αυξάνεται κατά 0.063.

2. b_0: 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_ϵ^2: Η εκτίμηση της διασποράς γύρω από την ευθεία παλινδρόμησης για κάθε τιμή της X (στο διάστημα τιμών του πειράματος) είναι 2.271, ή αλλιώς το τυπικό σφάλμα της εκτίμησης της παλινδρόμησης είναι 1.507 μονάδες, που είναι σχετικά 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 5.7 απεικονίζεται η

πρόβλεψη της απολαβής ρεύματος για αντίσταση στρώματος $x = 120\,\text{ohm/cm}$ και είναι

$$\hat{y} = 2.53 + 0.063 \cdot 120 = 10.11.$$

Στο παραπάνω παράδειγμα η πρόβλεψη της απολαβής ρεύματος μπορεί να διαφέρει σημαντικά αν αλλάξουν οι σημειακές εκτιμήσεις των παραμέτρων παλινδρόμησης. Στη συνέχεια θα μελετήσουμε την αβεβαιότητα στην εκτίμηση των παραμέτρων παλινδρόμησης και της πρόβλεψης της εξαρτημένης μεταβλητής.

5.2.3 Σχέση του συντελεστή συσχέτισης και παλινδρόμησης

Η παλινδρόμηση ορίζεται θεωρώντας την ανεξάρτητη μεταβλητή X ελεγχόμενη 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 ρ όπως και πριν. Η σχέση μεταξύ του r (της εκτιμήτριας του ρ από το δείγμα) και του b_1 (της εκτίμησης του συντελεστή της παλινδρόμησης β_1 με τη μέθοδο των ελαχίστων τετραγώνων) δίνεται ως εξής (συνδυάζοντας τις σχέσεις $r = \frac{s_{XY}}{s_X s_Y}$ και $b_1 = \frac{s_{XY}}{s_X^2}$)

$$r = b_1 \frac{s_X}{s_Y} \quad \text{ή} \quad b_1 = r \frac{s_Y}{s_X}. \tag{5.14}$$

Και τα δύο μεγέθη, r και b_1, εκφράζουν ποσοτικά τη γραμμική συσχέτιση των μεταβλητών X και Y, αλλά το b_1 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_1 περιγράφεται ως εξής:

- Αν η συσχέτιση είναι θετική ($r > 0$) τότε η κλίση της ευθείας παλινδρόμησης b_1 είναι επίσης θετική.

- Αν η συσχέτιση είναι αρνητική ($r < 0$) τότε η κλίση της ευθείας παλινδρόμησης b_1 είναι επίσης αρνητική.

- Αν οι μεταβλητές X και Y δε συσχετίζονται ($r = 0$) τότε η ευθεία παλινδρόμησης είναι οριζόντια ($b_1 = 0$).

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

$$s_\epsilon^2 = \frac{n-1}{n-2} s_Y^2 (1 - r^2) \quad \text{ή} \quad r^2 = 1 - \frac{n-2}{n-1} \frac{s_\epsilon^2}{s_Y^2}. \tag{5.15}$$

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 5.3. Στο παραπάνω παράδειγμα 5.2, ο συντελεστής συσχέτισης της απολαβής 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 και από τις σχέσεις (5.14) ή (5.15). Ο συντελεστής συσχέτισης δηλώνει την ασθενή θετική συσχέτιση της απολαβής ρεύματος και της αντίστασης στρώματος, που δε μπορούμε όμως να συμπεράνουμε από την τιμή του συντελεστή παλινδρόμησης $b_1 = 0.063$ ή την τιμή της διασποράς των σφαλμάτων $s_\epsilon^2 = 2.249$ γιατί εξαρτιούνται από τις μονάδες μέτρησης των δύο μεταβλητών.

5.2.4 Διάστημα εμπιστοσύνης των παραμέτρων της απλής γραμμικής παλινδρόμησης

Όπως η δειγματική μέση τιμή $\bar{x}$ και η τυπική απόκλιση s μπορούν να διαφέρουν από δείγμα σε δείγμα παρατηρήσεων μιας τ.μ. X, έτσι και η εκτιμώμενη κλίση b_1 και διαφορά ύψους b_0 μπορούν να διαφέρουν επίσης από δείγμα σε δείγμα ζευγαρωτών παρατηρήσεων των (X, Y). Σε αντιστοιχία με την προσέγγιση για τη μέση τιμή και τυπική απόκλιση, για να υπολογίσουμε διαστήματα εμπιστοσύνης (ή να κάνουμε στατιστικό έλεγχο όπως θα δούμε αμέσως μετά) για τις παραμέτρους β_1 και β_0 θα μελετήσουμε την κατανομή των b_1 και b_0 αντίστοιχα. Σημειώνεται ότι μόνο η Y είναι τυχαία μεταβλητή, καθώς θεωρήσαμε πως η X είναι ελεγχόμενη (ανεξάρτητη) μεταβλητή.

Μπορεί να δειχθεί με βάση της σχέση (5.9) πως ο εκτιμητής b_1 της κλίσης δίνεται ως γραμμικός συνδυασμός των $y_1, \ldots, y_n$, που κάθε ένα από αυτά είναι τ.μ. όπως το Y. Επιπλέον το b_1 έχει μέση τιμή

$$\mu_{b_1} \equiv \mathrm{E}(b_1) = \beta_1$$

και διασπορά

$$\sigma_{b_1}^2 \equiv \mathrm{Var}(b_1) = \frac{\sigma_\epsilon^2}{S_{xx}}$$

ή αντίστοιχα τυπική απόκλιση $\sigma_{b_1} = \sigma_\epsilon / \sqrt{S_{xx}}$. Καθώς η διασπορά των σφαλμάτων εκτιμάται από την ευθεία ελαχίστων τετραγώνων ως s_ϵ^2, η εκτίμηση της τυπικής απόκλισης του b_1 είναι

$$s_{b_1} = \frac{s_\epsilon}{\sqrt{S_{xx}}}. \tag{5.16}$$

Θεωρώντας πως η Y ακολουθεί κανονική κατανομή, ο εκτιμητής κλίσης b_1 ακολουθεί επίσης κανονική κατανομή και άρα το $(1-a)\%$ διάστημα εμπιστοσύνης της κλίσης β_1 δίνεται όπως και για τη μέση τιμή ως

$$b_1 \pm t_{n-2,1-a/2} s_{b_1} \quad \text{ή} \quad b_1 \pm t_{n-2,1-a/2} \frac{s_\epsilon}{\sqrt{S_{xx}}}. \tag{5.17}$$

Η κατανομή Student που δίνει την κρίσιμη τιμή έχει $n-2$ βαθμούς ελευθερίας όσους θεωρούμε και στην εκτίμηση του s_ϵ (δες σχέση (5.12)).

Με αντίστοιχη προσέγγιση μπορεί να δειχθεί πως η εκτιμώμενη τυπική απόκλιση της εκτίμησης της διαφοράς ύψους b_0 είναι

$$s_{b_0} = s_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \tag{5.18}$$

και το $(1-a)\%$ διάστημα εμπιστοσύνης του σταθερού όρου β_0 είναι

$$b_0 \pm t_{n-2,1-a/2} s_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}. \tag{5.19}$$

5.2.5 Έλεγχος υπόθεσης για τις παραμέτρους της απλής γραμμικής παλινδρόμησης

Κάτω από την υπόθεση της κανονικότητας μπορούν να γίνουν αντίστοιχοι παραμετρικοί έλεγχοι υπόθεσης για την κλίση και τη διαφορά ύψους.

Για τη μηδενική υπόθεση H_0: $\beta_1 = \beta_1^0$, δηλαδή ότι η κλίση μπορεί να είναι β_1^0, ο παραμετρικός έλεγχος γίνεται με το στατιστικό

$$t = \frac{b_1 - \beta_1^0}{s_b} = \frac{(b_1 - \beta_1^0)\sqrt{S_{xx}}}{s_\epsilon}, \tag{5.20}$$

όπου $t \sim t_{n-2}$, δηλαδή το στατιστικό κάτω από την H_0 ακολουθεί κατανομή Student με $n-2$ βαθμούς ελευθερίας. Ιδιαίτερο ενδιαφέρον έχει η μηδενική υπόθεση H_0: $\beta_1 = 0$, δηλαδή ότι η γραμμή παλινδρόμησης είναι οριζόντια και άρα η τ.μ. Y δεν εξαρτάται από την X.

Αντίστοιχα ο έλεγχος υπόθεσης για τη διαφορά ύψους, H_0: $\beta_0 = \beta_0^0$, δίνεται από το στατιστικό

$$t = \frac{b_0 - \beta_0^0}{s_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \tag{5.21}$$

όπου και πάλι $t \sim t_{n-2}$.

5.2.6 Διαστήματα πρόβλεψης

Έχοντας μελετήσει την κατανομή των παραμέτρων της ευθείας ελαχίστων τετραγώνων b_0 και b_1 μπορούμε να μελετήσουμε την κατανομή της εκτίμησης της μέσης τιμής του Y για κάποιο x, $E(Y|X = x) = \beta_0 + \beta_1 x$, από την ευθεία ελαχίστων τετραγώνων $\hat{y}$ που δίνεται ως $\hat{y} = b_0 + b_1 x$.

Μπορεί να δειχθεί και πάλι πως ο εκτιμητής $\hat{y}$ δίνεται ως γραμμικός συνδυασμός των τ.μ. $y_1, \ldots, y_n$ και των σταθερών $x_1, \ldots, x_n$ και x. Το $\hat{y}$ είναι αμερόληπτος εκτιμητής του $E(Y|X = x)$, δηλαδή

$$\mu_{\hat{y}} \equiv E(\hat{y}) = E(Y|X = x) = \beta_0 + \beta_1 x.$$

Η διασπορά του $\hat{y}$ μπορεί να δειχθεί ότι είναι

$$\sigma^2_{\hat{y}} \equiv \mathrm{Var}(\hat{y}) = \sigma^2_\epsilon \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right)$$

Αντικαθιστώντας τη διασπορά των σφαλμάτων από την ευθεία ελαχίστων τετραγώνων ως s^2_ϵ, η εκτίμηση της τυπικής απόκλισης του $\hat{y}$ είναι

$$s_{\hat{y}} = s_\epsilon \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}. \tag{5.22}$$

Όπως και για τις παραμέτρους b_0 και b_1, ο εκτιμητής της ευθείας ελαχίστων τετραγώνων για κάποιο x ακολουθεί επίσης κανονική κατανομή. Άρα το $(1 - a)$**% διάστημα εμπιστοσύνης της μέσης τιμής του** Y **για κάποιο** x είναι

$$\hat{y} \pm t_{n-2,1-a/2} s_{\hat{y}} \quad \text{ή} \quad (b_0 + b_1 x) \pm t_{n-2,1-a/2} s_\epsilon \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}. \tag{5.23}$$

Το διάστημα εμπιστοσύνης για μέσης τιμής του Y για κάποιο x εξαρτάται από την απόσταση του x από το μέσο όρο των τιμών της μεταβλητής Q του δείγματος. Υπάρχει λοιπόν μεγαλύτερη βεβαιότητα για το Y όταν το η τιμή x είναι κοντά στο κέντρο των δεδομένων της Q που ήδη γνωρίζουμε (βάσει των οποίων έγινε η εκτίμηση της ευθείας ελαχίστων τετραγώνων).

Το παραπάνω διάστημα εμπιστοσύνης της μέσης τιμής του Y για κάποιο x, $\mathrm{E}(Y|X=x)$, είναι το **διάστημα της μέσης πρόβλεψης**, δηλαδή δίνει σε επίπεδο εμπιστοσύνης $(1-a)\%$ τα όρια της πρόβλεψης για τη μέση (αναμενόμενη) τιμή της Y όταν δίνεται μια συγκεκριμένη τιμή x για τη μεταβλητή X. Συχνά μας ενδιαφέρει να γνωρίζουμε και τα όρια της πρόβλεψης για μια (μελλοντική) τιμή y της Y που θα πάρουμε για κάποια τιμή x της X. Το ζητούμενο διάστημα εμπιστοσύνης δεν αναφέρεται στην παράμετρο της δεσμευμένης μέσης τιμής $\mathrm{E}(Y|X=x)$ αλλά κάποιας τιμής y της Y για το ίδιο x και άρα περιμένουμε τα όρια να είναι μεγαλύτερα. Η διαφορά είναι ίδια με την ακρίβεια της μέσης τιμής κάποιας τ.μ. X και μιας παρατήρησης της X. Πράγματι το $(1-a)$**% διάστημα πρόβλεψης για μια παρατήρηση** y **της** Y **για κάποιο** x είναι

$$\hat{y} \pm t_{n-2,1-a/2}\sqrt{s_\varepsilon^2 + s_{\hat{y}}^2} \quad \text{ή} \quad (b_0+b_1x) \pm t_{n-2,1-a/2}s_\varepsilon\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}. \quad (5.24)$$

Παράδειγμα 5.4. Στο παραπάνω παράδειγμα 5.2, βρήκαμε τις εκτιμήσεις των παραμέτρων της ευθείας παλινδρόμησης $b_1 = 0.063$, $b_0 = 2.53$ και την εκτίμηση της τυπικής απόκλισης των σφαλμάτων παλινδρόμησης $s_\varepsilon = 1.507$. Θα εξετάσουμε στη συνέχεια την ακρίβεια και σημαντικότητα των παραμέτρων της ευθείας παλινδρόμησης και θα δώσουμε διαστήματα πρόβλεψης για την απολαβή ρεύματος όταν η αντίσταση στρώματος είναι $x = 120\,\mathrm{ohm/cm}$.

Η εκτίμηση της τυπικής απόκλιση της κλίσης της ευθείας ελαχίστων τετραγώνων b_1 είναι από τη σχέση (5.16) $s_{b_1} = 0.0195$. Σύμφωνα με την σχέση (5.17) το 95% διάστημα εμπιστοσύνης για την κλίση είναι

$$0.063 \pm 2.306 \cdot 0.0195 \quad \Rightarrow \quad [0.018, 0.108].$$

Το διάστημα περιέχει μόνο θετικές τιμές που δείχνει πως η κλίση β_1 της ευθείας παλινδρόμησης είναι σημαντικά διάφορη του μηδενός, τουλάχιστον σε επίπεδο σημαντικότητας $a = 0.05$.

Ο έλεγχος για τη μηδενική υπόθεση H_0: $\beta_1 = 0$ επιβεβαιώνει τη σημαντικότητα της κλίσης. Το στατιστικό από το δείγμα από τη σχέση (5.20) έχει την τιμή

$$t = \frac{0.063}{0.0195} = 3.235$$

που είναι αρκετά μεγαλύτερη από την κρίσιμη τιμή $t_{0.975,8} = 2.306$ που δίνει το όριο της t για να δεχθούμε την H_0. Η τιμή $t = 3.235$ από το δείγμα αντιστοιχεί στην πιθανότητα (p-τιμή) $p = 0.012$. Η p-τιμή δηλώνει πως δε θα μπορούσαμε να απορρίψουμε την H_0 ότι η κλίση είναι μηδενική σε επίπεδο σημαντικότητας $a = 0.01$, δηλαδή το μοντέλο παλινδρόμησης δεν έχει μεγάλη σημαντικότητα. Αυτό οφείλεται σε μεγάλο βαθμό στο μικρό αριθμό των

παρατηρήσεων σε συνδυασμό με την όχι τόσο ισχυρή εξάρτηση της απολαβής ρεύματος από την αντίσταση στρώματος ($r = 0.753$).

Αντίστοιχα η εκτίμηση της τυπικής απόκλισης της διαφοράς ύψους β_0 από τη σχέση (5.18) είναι $s_{b_0} = 1.894$ και το 95% διάστημα εμπιστοσύνης είναι σύμφωνα με τη σχέση (5.19)

$$2.53 \pm 2.306 \cdot 1.894 \quad \Rightarrow \quad [-1.837, 6.898]$$

Το διάστημα περιέχει το μηδέν και άρα η διαφορά ύψους δεν είναι σημαντική σε επίπεδο σημαντικότητας $a = 0.05$. Αυτό επιβεβαιώνεται από τον έλεγχο της μηδενικής υπόθεσης H_0: $\beta_0 = 0$. Το στατιστικό από το δείγμα από τη σχέση (5.21) είναι

$$t = \frac{2.53}{1.894} = 1.336$$

που είναι μικρότερο της κρίσιμης τιμής $t_{0.975,8} = 2.306$ και δίνει p-τιμή $p = 0.218$. Το αποτέλεσμα αυτό δηλώνει πως η εξάρτηση της απολαβής ρεύματος από την αντίσταση στρώματος μπορεί να περιγραφτεί με μοντέλο γραμμικής παλινδρόμησης χωρίς σημαντική διαφορά ύψους.

Στη συνέχεια ας εξετάσουμε τα όρια πρόβλεψης για αντίσταση στρώματος $x = 120\,\text{ohm/cm}$. Από τη σχέση (5.22) υπολογίζουμε την τυπική απόκλιση της μέσης απολαβής ρεύματος για αντίσταση στρώματος $x = 120\,\text{ohm/cm}$ ως

$$s_{\hat{y}} = 1.507 \sqrt{\frac{1}{10} + \frac{(120 - 93.9)^2}{9 \cdot 662.1}} = 0.698.$$

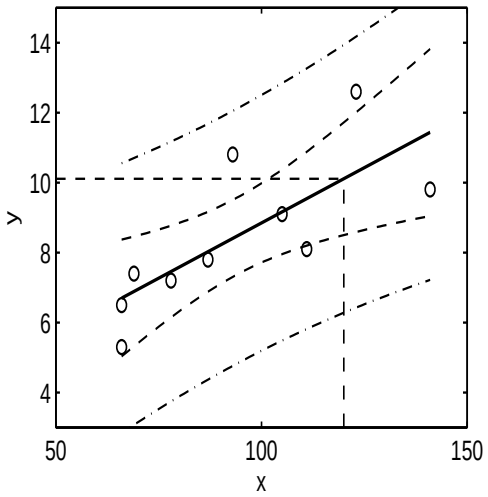
Το 95% διάστημα πρόβλεψης της μέσης απολαβής ρεύματος για αντίσταση στρώματος $x = 120\,\text{ohm/cm}$ δίνεται από τη σχέση (5.23) ως

$$10.108 \pm 2.306 \cdot 0.698 \quad \Rightarrow \quad [8.499, 11.717]$$

Το αντίστοιχο 95% διάστημα πρόβλεψης για μια (μελλοντική) παρατήρηση y της απολαβής ρεύματος για αντίσταση στρώματος $x = 120\,\text{ohm/cm}$ είναι από τη σχέση (5.24)

$$10.108 \pm 2.306 \cdot 1.507 \sqrt{1 + \frac{1}{10} + \frac{(120 - 93.9)^2}{9 \cdot 662.1}} \quad \Rightarrow \quad [6.279, 13.937]$$

Τα διαστήματα πρόβλεψης της μέσης απολαβής ρεύματος και μελλοντικής απολαβής για τιμές αντίστασης στρώματος στο διάστημα που ορίζεται από το δείγμα δίνονται στο Σχήμα 5.8. Παρατηρούμε πως και τα δύο διαστήματα πρόβλεψης είναι πιο μικρά για τιμές αντίστασης στρώματος κοντά στη δειγματική μέση τιμή της αντίστασης στρώματος. Σε αυτό το παράδειγμα των 10 ζευγαρωτών παρατηρήσεων το 95% διάστημα πρόβλεψης των τιμών απολαβής ρεύματος περιέχει και τις 10 τιμές.



Σχήμα 5.8: Διάγραμμα διασποράς για τα δεδομένα του Πίνακα 5.2, ευθεία ελαχίστων τετραγώνων και διαστήματα πρόβλεψης μέσης και μελλοντικής απολαβής ρεύματος.

5.2.7 Επάρκεια μοντέλου απλής γραμμικής παλινδρόμησης

Γενικά όταν έχουμε ένα πραγματικό πρόβλημα, όπως εδώ το πρόβλημα της απλής γραμμικής παλινδρόμησης, δε γνωρίζουμε από πριν ποιο είναι το κατάλληλο μοντέλο. Θα πρέπει λοιπόν να διερευνήσουμε αν το μοντέλο είναι επαρκές, δηλαδή αν εξηγεί ικανοποιητικά τις σχέσεις των μεταβλητών στο πρόβλημα μας. Για την περίπτωση της απλής γραμμικής παλινδρόμησης που μελετάμε εδώ θα πρέπει να εξετάσουμε αν η προσαρμογή ευθείας ερμηνεύει πλήρως τη σχέση της εξαρτημένης μεταβλητής Y από την ανεξάρτητη μεταβλητή X, ή μπορεί μια άλλη σχέση (πιο πολύπλοκη από την απλή ευθεία) να εξηγεί καλύτερα. Κάποιος μπορεί να προσπαθήσει να απαντήσει σε αυτό το ερώτημα μελετώντας το διάγραμμα διασποράς της Y προς τη X, δηλαδή αν φαίνεται κάποια άλλη καμπύλη να προσαρμόζεται καλύτερα στα σημεία (X, Y), αλλά την απάντηση καλύτερα μας τη δίνουν τα υπόλοιπα $e_i = y_i - \hat{y}_i$ της προσαρμογής της ευθείας των ελαχίστων τετραγώνων.

Η διερεύνηση της καταλληλότητας του μοντέλου γίνεται στα τυποποιημένα σφάλματα προσαρμογής, $e_i^* = e_i/s_e$, όπου s_e είναι η εκτιμώμενη τυπική απόκλιση του σφάλματος e_i που δίνεται από τη ρίζα της διασποράς στη σχέση (5.12) ή (5.13) και η μέση τιμή του είναι 0. Θεωρώντας κανονική κατανομή για το σφάλμα e_i υπάρχει πιο ακριβής έκφραση για τη διασπορά του αλλά θα περιοριστούμε εδώ στο s_e^2 όπως το ορίσαμε γενικά.

Κατάλληλα γραφήματα των σφαλμάτων μπορούν να διαγνώσουν την καταλληλότητα και επάρκεια του μοντέλου. Το πιο σημαντικό **διαγνωστικό γράφημα** (diagnostic plot) είναι το διάγραμμα διασποράς των τυποποιημένων σφαλμάτων ως προς την εκτιμώμενη εξαρτημένη μεταβλητή $\hat{y}_i$. Σε αυτό συνήθως σχηματίζονται και δύο οριζόντιες γραμμές στο επίπεδο 2 και -2 (για την ακρίβεια ± 1.96) που αντιστοιχούν στα 95% όρια των τιμών του e_i^* αν αυτό ακολουθεί τυπική κανονική κατανομή. Κάποια από τα προβλήματα στην καλή προσαρμογή του μοντέλου απλής γραμμικής παλινδρόμησης δίνονται στο Σχήμα 5.9 και εξηγούνται παρακάτω.

- Το ποσοστό των σημείων στο γράφημα e_i^* προς $\hat{y}_i$ έξω από τα όρια υπερβαίνει το 5% (περίπου). Τότε η κανονικότητα των σφαλμάτων αμφισβητείται και άρα τα παραμετρικά διαστήματα και ο έλεγχος για τις παραμέτρους, καθώς και τα διαστήματα πρόβλεψης, δεν είναι ακριβή (δες Σχήμα 5.9α).

- Η διασπορά των σφαλμάτων δεν είναι σταθερή αλλά αλλάζει με το $\hat{y}_i$. Η υπόθεση της σταθερής διασποράς των σφαλμάτων δεν ισχύει και άρα και πάλι τα παραμετρικά διαστήματα δεν ισχύουν (δες Σχήμα 5.9β).

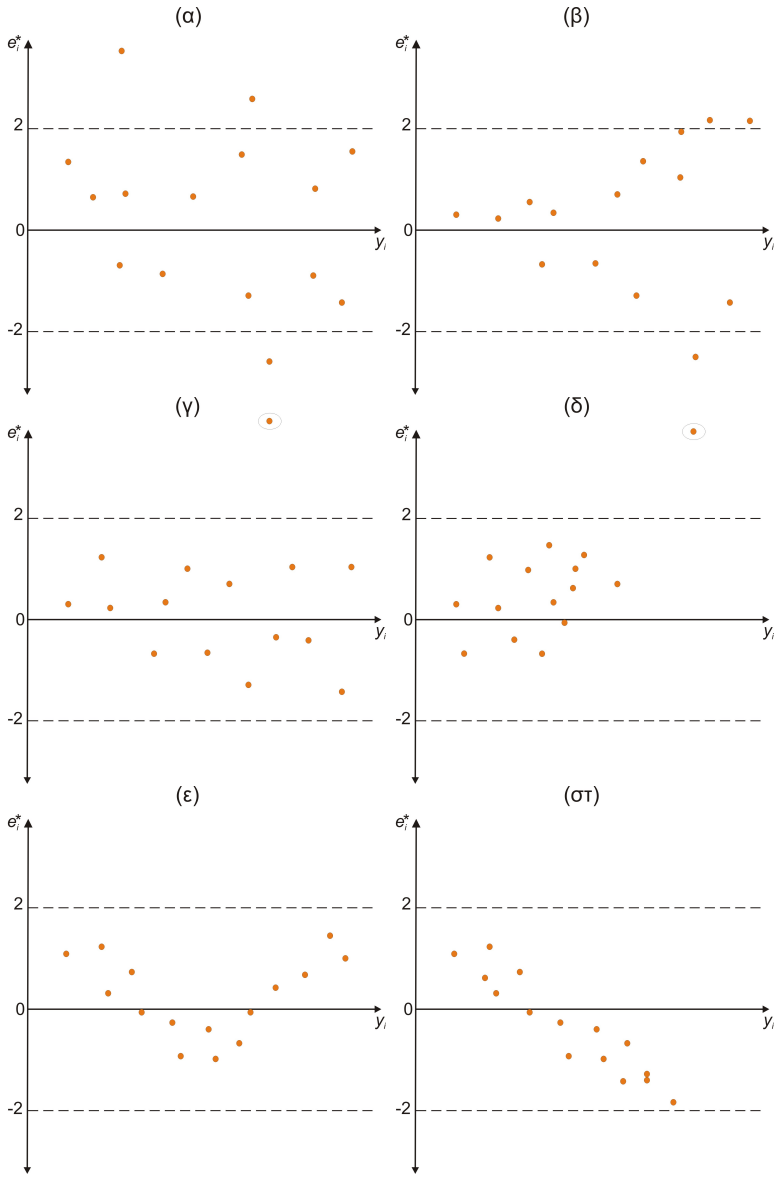
- Κάποιο σφάλμα είναι πολύ μεγαλύτερο από όλα τα άλλα. Αυτό έχει επηρεάσει σημαντικά την εκτίμηση των παραμέτρων του μοντέλου (δηλαδή τους συντελεστές της ευθείας ελαχίστων τετραγώνων). Άρα αν παραλείψουμε το ζεύγος τιμών των (X, Y) που αντιστοιχεί στο μεγάλο σφάλμα, το εκτιμώμενο μοντέλο στο νέο δείγμα μπορεί να είναι σημαντικά διαφορετικό (δες Σχήμα 5.9γ).

- Υπάρχει απόμακρο σημείο (outlier). Αυτό και πάλι επηρεάζει σημαντικά την εκτίμηση των παραμέτρων του μοντέλου και αν το απαλείψουμε μπορεί το μοντέλο της απλής γραμμικής παλινδρόμησης να μη φαίνεται πλέον κατάλληλο (δες Σχήμα 5.9δ).

- Τα σφάλματα διαμορφώνουν κάποιο σχήμα καμπύλης. Η υπόθεση της γραμμικής εξάρτησης δεν ισχύει και η εξάρτηση μπορεί να είναι μηγραμμική (δες Σχήμα 5.9ε).

- Τα σφάλματα φαίνεται να συγκεντρώνονται γύρω από μια γραμμή. Η εξαρτημένη μεταβλητή μπορεί να εξαρτιέται γραμμικά και από κάποια δεύτερη μεταβλητή που πρέπει να συμπεριληφθεί στο μοντέλο γραμμικής παλινδρόμησης (δες Σχήμα 5.9στ).

Ειδικότερα οι δύο τελευταίες δυσκολίες που παρουσιάστηκαν στην προσαρμογή του μοντέλου απλής γραμμικής παλινδρόμησης στα Σχήματα 5.9ε



Σχήμα 5.9: Διαγράμματα διασποράς των τυποποιημένων σφαλμάτων προς τις εκτιμήσεις της εξαρτημένης μεταβλητής, που δείχνουν ανεπάρκεια του μοντέλου. (α) Μη-κανονική κατανομή (πολλά σημεία εκτός των ορίων του 95% της κανονικής κατανομής. (β) Η διασπορά των σφαλμάτων δεν είναι σταθερή. (γ) Ακραία τιμή σφάλματος. (δ) Απομακρυσμένη τιμή σφάλματος και ανεξάρτητης μεταβλητής. (ε) Ύπαρξη μη-γραμμικής εξάρτησης. (στ) Ύπαρξη επιπλέον ανεξάρτητης μεταβλητής.

και 5.9στ, συνιστούν τη χρήση άλλου τύπου μοντέλου, απλής μη-γραμμικής παλινδρόμησης και πολλαπλής γραμμικής παλινδρόμησης, αντίστοιχα, που θα δούμε παρακάτω.

5.3 Μη-Γραμμική Παλινδρόμηση

Σε κάποια προβλήματα μπορεί να έχουμε θεωρητικές ενδείξεις ότι η εξάρτηση μιας εξαρτημένης τ.μ. Y από μια ανεξάρτητη μεταβλητή X είναι κάποιας συγκεκριμένης μη-γραμμικής μορφής. Σε κάποιες περιπτώσεις, μπορεί τη μη-γραμμική μορφή να μας την υποδείξει το διαγνωστικό γράφημα για ένα μοντέλο απλής γραμμικής παλινδρόμησης, όπως στο Σχήμα 5.9ε. Σε κάθε περίπτωση έχουμε να εκτιμήσουμε τις παραμέτρους μιας μη γραμμικής συνάρτησης.

5.3.1 Εγγενής γραμμική συνάρτηση παλινδρόμησης

Κάποιες μη-γραμμικές συναρτήσεις μπορεί με κατάλληλο μετασχηματισμό να γίνουν γραμμικές, και τότε μια τέτοια συνάρτηση λέγεται **εγγενής γραμμική** (intrinsically linear). Στον Πίνακα 5.3 δίνονται τέσσερις από τις πιο γνωστές εγγενείς γραμμικές συναρτήσεις και για κάθε μια δίνεται ο κατάλληλος μετασχηματισμός και η γραμμική μορφή που δίνει ο μετασχηματισμός.

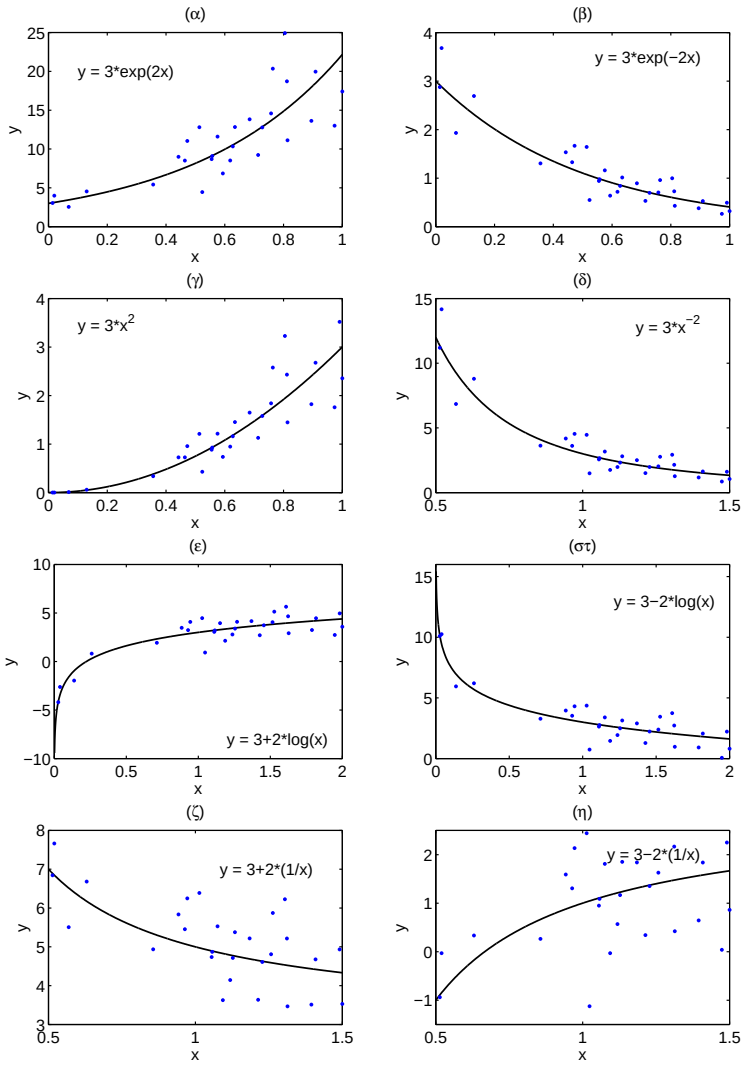
Εγγενής συνάρτηση	*Μετασχηματισμός*	*Γραμμική συνάρτηση*
1. Εκθετική: $y = ae^{\beta x}$	$y' = \ln(y)$	$y' = \ln(a) + \beta x$
2. Δύναμης: $y = ax^{\beta}$	$y' = \log(y)$, $x' = \log(x)$	$y' = \log(a) + \beta x'$
3. $y = a + \beta \log(x)$	$x' = \log(x)$	$y = a + \beta x'$
4. Αντίστροφη: $y = a + \beta \frac{1}{x}$	$x' = \frac{1}{x}$	$y = a + \beta x'$

Πίνακας 5.3: Εγγενείς γραμμικές συναρτήσεις, οι κατάλληλοι μετασχηματισμοί και οι γραμμικές συναρτήσεις που προκύπτουν από τους μετασχηματισμούς. Όπου δίνεται ο δεκαδικός λογάριθμος μπορεί ισοδύναμα να χρησιμοποιηθεί ο νεπέρειος λογάριθμος.

Το βασικό πλεονέκτημα που έχουμε όταν γνωρίζουμε πως η μορφή της συνάρτησης παλινδρόμησης δεν είναι μια οποιαδήποτε μη-γραμμική συνάρτηση αλλά είναι εγγενής γραμμική είναι πως μπορούμε να εκτιμήσουμε τις παραμέτρους της συνάρτησης με τη μέθοδο των ελαχίστων τετραγώνων το ίδιο

εύκολα όπως στη γραμμική παλινδρόμηση. Αυτό συμβαίνει γιατί η συνάρτηση του αθροίσματος των τετραγώνων των σφαλμάτων παραμένει γραμμική ως προς τις παραμέτρους.

Για κάθε εγγενή γραμμική συνάρτηση, η αντίστοιχη στοχαστική συνάρτηση που σχηματίζεται προσθέτοντας θόρυβο ϵ δεν είναι πάντα εγγενής γραμμική. Για παράδειγμα το εκθετικό στοχαστικό μοντέλο $y = ae^{\beta x} + \epsilon$ ή το στοχαστικό μοντέλο δύναμης $y = ax^{\beta} + \epsilon$ (και τα δύο θεωρώντας προσθετικό θόρυβο ϵ) δεν είναι εγγενείς στοχαστικές συναρτήσεις γιατί ο μετασχηματισμός της λογαρίθμησης εφαρμόζεται σε άθροισμα με έναν όρο την ανεξάρτητη μεταβλητή x και δεύτερο όρο το θόρυβο ϵ και άρα δε μπορούν αυτά να διαχωριστούν. Αν όμως θεωρήσουμε πολλαπλασιαστικό θόρυβο ϵ, δηλαδή $y = ae^{\beta x} \cdot \epsilon$ ή $y = ax^{\beta} \cdot \epsilon$, τότε ο διαχωρισμός είναι δυνατός. Μάλιστα αν ο θόρυβος ϵ έχει λογαριθμική κανονική κατανομή (lognormal distribution) τότε ο μετασχηματισμός δίνει θόρυβο $\epsilon' = \ln \epsilon$ με κανονική κατανομή. Για τις δύο άλλες εγγενείς γραμμικές συναρτήσεις (τις δύο τελευταίες στον Πίνακα 5.3) ο θόρυβος πρέπει να είναι προσθετικός και τότε οι στοχαστικές συναρτήσεις $y = a + \beta \log(x) + \epsilon$ και $y = a + \beta \frac{1}{x} + \epsilon$ είναι εγγενείς γραμμικές, δηλαδή ο μετασχηματισμός δίνει ισοδύναμο γραμμικό μοντέλο παλινδρόμησης. Στο Σχήμα 5.10 παρουσιάζονται οι εγγενείς συναρτήσεις και παρατηρήσεις που προκύπτουν από αυτές προσθέτοντας κατάλληλο θόρυβο (από λογαριθμική κατανομή στις δύο πρώτες εγγενείς συναρτήσεις και από κανονική κατανομή για τις άλλες δύο).



Σχήμα 5.10: Οι 4 εγγενείς γραμμικές συναρτήσεις του Πίνακα 5.3 για θετικό και αρνητικό συντελεστή β, καθώς και παρατηρήσεις από αυτές με θόρυβο: εγγενής συνάρτηση 1 με $\beta > 0$ και $\beta < 0$ στο (α) και (β), αντίστοιχα συνάρτηση 2 στο (γ) και (δ), συνάρτηση 3 στο (ε) και (στ), και συνάρτηση 4 στο (ζ) και (η).

Παράδειγμα 5.5. Για ένα ιδανικό αέριο ισχύει $pV^\gamma = C$ για κάποια σταθερά C, όπου p είναι η απόλυτη πίεση του αερίου, V ο όγκος του και γ είναι ένας εκθέτης χαρακτηριστικός για το ιδανικό αέριο (λέγεται και λόγος των ειδικών θερμοτήτων (ratio of the specific heats)). Θέλουμε να εκτιμήσουμε τον εκθέτη γ καθώς και τη σταθερά C από τις παρακάτω μετρήσεις απόλυτης πίεσης και όγκου του ιδανικού αερίου στον Πίνακα 5.4. Επίσης θέλουμε να προβλέψουμε την απόλυτη πίεση για όγκο ιδανικού αερίου 25 in.^3.

A/A	p [psi]	V [in.^3]
1	16.6	50
2	39.7	30
3	78.5	20
4	115.5	15
5	195.3	10
6	546.1	5

Πίνακας 5.4: Μετρήσεις απόλυτη πίεσης και όγκου για ένα ιδανικό αέριο.

Πράγματι στο Σχήμα 5.11α οι μετρήσεις υποδεικνύουν μια μη-γραμμική σχέση της απόλυτης πίεσης και του όγκου του ιδανικού αερίου. Η σχέση του προβλήματος δηλώνει εγγενή γραμμική συνάρτηση της μορφής δύναμης (συνάρτηση 2 του Πίνακα 5.3)

$$pV^\gamma = C \quad \Leftrightarrow \quad y = ax^\beta,$$

όπου η εξαρτημένη μεταβλητή είναι $y = p$, η ανεξάρτητη μεταβλητή είναι $x = V$, και οι παράμετροι είναι $a = C$ και $\beta = -\gamma$. Εφαρμόζοντας λογάριθμο και στα δύο μέρη της ισότητας και θέτοντας $y' = \ln(y) = \ln(p)$ και $x' = \ln(x) = \ln(V)$ παίρνουμε το γραμμικό μοντέλο

$$y' = \ln(a) + \beta x' \quad \Leftrightarrow \quad \ln(p) = \ln(C) - \gamma \ln(V).$$

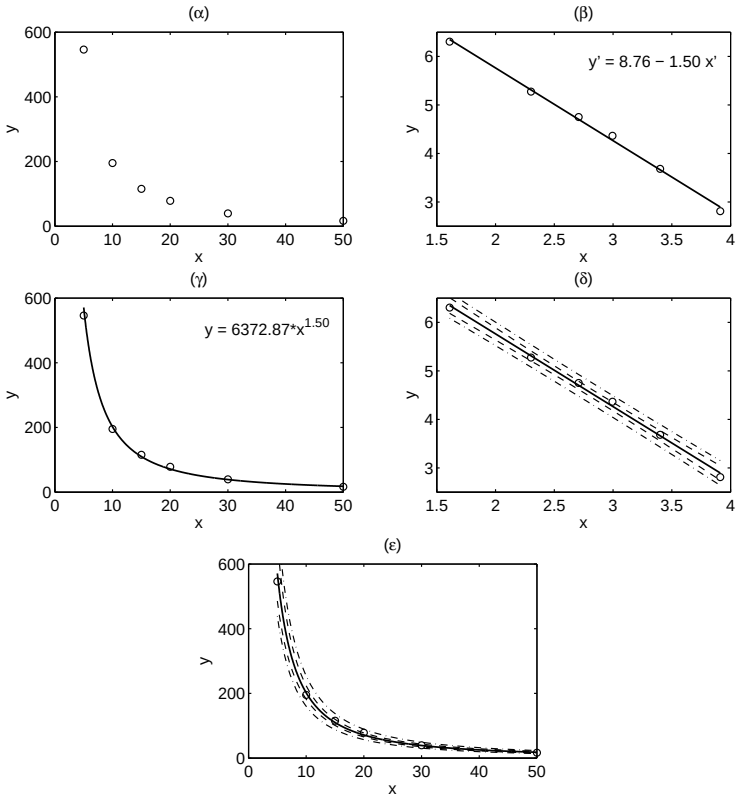
Θεωρώντας ότι ο θόρυβος στις παρατηρήσεις του Πίνακα 5.4 εισέρχεται πολλαπλασιαστικά στο αρχικό μοντέλο, δηλαδή

$$pV^\gamma = C \cdot \epsilon \quad \Leftrightarrow \quad y = ax^\beta \cdot \epsilon,$$

και θέτοντας $\epsilon' = \ln(\epsilon)$ έχουμε το πρόβλημα της απλής γραμμικής παλινδρόμησης του y' προς τη x' στη μορφή

$$y' = \ln(a) + \beta x' + \epsilon' \quad \Leftrightarrow \quad \ln(p) = \ln(C) - \gamma \ln(V) + \ln(\epsilon).$$

Τα μετασχηματισμένα δεδομένα για x' και y' δίνονται στον Πίνακα 5.5. Το διάγραμμα διασποράς τους στο Σχήμα 5.11β υποδηλώνει τη γραμμική



Σχήμα 5.11: (α) Διάγραμμα διασποράς των 6 παρατηρήσεων απόλυτης πίεσης p και όγκου V για ένα ιδανικό αέριο. (β) Διάγραμμα διασποράς των μετασχηματισμένων p και V και προσαρμογή ευθείας ελαχίστων τετραγώνων. (γ) Διάγραμμα διασποράς όπως στο (α) και προσαρμογή της συνάρτησης που προκύπτει από τον αντίστροφο μετασχηματισμό της ευθείας ελαχίστων τετραγώνων. (δ) Όπως στο (β) αλλά με τα 95% διαστήματα πρόβλεψης για τη μέση τιμή και για μια τιμή της y' για κάθε x'. (ε) Όπως στο (γ) αλλά με τα 95% διαστήματα πρόβλεψης για τη μέση τιμή και για μια τιμή της απόλυτης πίεσης για κάθε τιμή του όγκου του ιδανικού αερίου.

σχέση του λογαρίθμου της απόλυτης πίεσης y' και του λογαρίθμου του όγκου x' του ιδανικού αερίου.

Εκτιμούμε τις παραμέτρους b_0 και b_1 της ευθείας ελαχίστων τετραγώνων. Έχουμε $\bar{x} = 2.82$, $\bar{y} = 4.53$, $s_X^2 = 0.6614$ και $s_{XY} = -0.9915$, οπότε εκτιμήσεις των συντελεστών της ευθείας είναι

$$\begin{aligned} b_1 &= \frac{-0.9915}{0.6614} = -1.4991 \\ b_0 &= 4.53 + 1.4991 \cdot 2.82 = 8.7598. \end{aligned}$$

A/A	x'	y'
1	1.609	6.303
2	2.303	5.275
3	2.708	4.749
4	2.996	4.363
5	3.401	3.681
6	3.912	2.809

Πίνακας 5.5: Τιμές λογαρίθμων των μετρήσεων απόλυτη πίεσης και όγκου για ένα ιδανικό αέριο.

Από τη σχέση (5.13) και έχοντας $s_Y^2 = 1.4908$, υπολογίζουμε την εκτίμηση διασποράς των σφαλμάτων παλινδρόμησης

$$s_\epsilon^2 = \frac{5}{4}(1.4908 - 1.4991^2 \cdot 0.6614) = 0.00554,$$

και αντίστοιχα οι τυπικές αποκλίσεις των σφαλμάτων είναι $s_\epsilon = 0.07442$.

Για να προβλέψουμε την απόλυτη πίεση για όγκο ιδανικού αερίου 25 in.^{3}, χρησιμοποιούμε το λογάριθμο της τιμής του όγκου $x' = \ln(25) = 3.219$ και η πρόβλεψη του λογαρίθμου της απόλυτης πίεσης είναι

$$\ln(p) = y' = 8.7598 - 1.4991 \cdot 3.219 = 3.9344.$$

Άρα η πρόβλεψη της απόλυτης πίεσης για όγκο ιδανικού αερίου 25 in.^{3} είναι $p = \exp(3.9344) = 51.13$ psi. Στο Σχήμα 5.11γ δίνεται η καμπύλη που δίνει την απόλυτη πίεση p ως συνάρτηση του όγκου V.

Μπορούμε επίσης να υπολογίσουμε διαστήματα πρόβλεψης για τη μέση απόλυτη πίεση και για μια πρόβλεψη της απόλυτης πίεσης για κάποιο δεδομένο όγκο του ιδανικού αερίου. Υπολογίζουμε πρώτα τα διαστήματα πρόβλεψης για το μετασχηματισμένο γραμμικό μοντέλο και παίρνουμε τον αντίστροφο μετασχηματισμό στα όρια του διαστήματος πρόβλεψης.

Για $x' = \ln(25) = 3.219$ από τη σχέση (5.23) υπολογίζουμε το 95% διάστημα πρόβλεψης της μέσης y' ως $[3.839, 4.030]$ και εφαρμόζοντας την εκθετική συνάρτηση στα άκρα προκύπτει το 95% διάστημα πρόβλεψης της μέσης απόλυτης πίεσης για όγκο ιδανικού αερίου 25 in.^{3} ως $[46.465, 56.264]$. Αντίστοιχα τα όρια του 95% διαστήματος πρόβλεψης για μια τιμή του y' όταν $x' = \ln(25) = 3.219$ είναι $[3.707, 4.162]$ και τα αντίστοιχα όρια για μια τιμή απόλυτης πίεσης για όγκο ιδανικού αερίου 25 in.^{3} είναι $[40.718, 64.205]$. Τα 95% διαστήματα πρόβλεψης (μέσης και μιας τιμής) για το γραμμικοποιημένο και το αρχικό μοντέλο παλινδρόμησης δίνονται στο Σχήμα 5.11δ και 5.11ε, αντίστοιχα. Παρατηρούμε πως η πρόβλεψη της απόλυτης πίεσης είναι λιγότερη ακριβής για μικρούς όγκους ιδανικού αερίου.

Θα πρέπει να σημειωθεί πως στην περίπτωση εγγενούς γραμμικής συνάρτησης παλινδρόμησης, οι εκτιμήσεις των παραμέτρων στο μετασχηματισμένο γραμμικό μοντέλο παλινδρόμησης είναι οι καλύτερες. Από αυτές μπορούμε να εκτιμήσουμε τις παραμέτρους του αρχικού μη-γραμμικού μοντέλου παλινδρόμησης, αλλά αυτές οι εκτιμήσεις δεν είναι οι καλύτερες, με την έννοια της ελαχιστοποίησης των σφαλμάτων. Για να το πετύχουμε αυτό θα πρέπει να εφαρμόσουμε τη μέθοδο ελαχίστων τετραγώνων απευθείας στο αρχικό μη-γραμμικό σύστημα. Η μέθοδος αυτή απαιτεί τη λύση ενός μη-γραμμικού συστήματος εξισώσεων ως προς τις παραμέτρους, που ανάλογα με τη μορφή της μη-γραμμικής συνάρτησης παλινδρόμησης μπορεί να είναι αρκετά σύνθετη.

Στη συνέχεια θα δούμε έναν τύπο μη-γραμμικής παλινδρόμησης, την πολυωνυμική παλινδρόμηση, που δεν αντιμετωπίζεται με μετασχηματισμό σε γραμμική συνάρτηση αλλά εκτιμάται απευθείας με χρήση της μεθόδου ελαχίστων τετραγώνων.

5.3.2 Πολυωνυμική παλινδρόμηση

Τα μη-γραμμικά μοντέλα παλινδρόμησης που δίνονται από εγγενείς γραμμικές συναρτήσεις της εξαρτημένης μεταβλητής Y προς την ανεξάρτητη μεταβλητή X έχουν ως κοινό χαρακτηριστικό ότι οι συναρτήσεις είναι μονότονες, αύξουσες ή φθίνουσες (δες Σχήμα 5.10). Σε πολλά προβλήματα η θεωρητική προσέγγιση ή το διάγραμμα διασποράς συνιστά ότι η συνάρτηση έχει ένα ή περισσότερα σημεία καμπής. Σε τέτοιες περιπτώσεις η πολυωνυμική συνάρτηση κάποιου βαθμού k μπορεί να αποτελεί ικανοποιητική προσέγγιση της πραγματικής συνάρτησης παλινδρόμησης.

Η **πολυωνυμική παλινδρόμηση βαθμού** k (k-th degree polynomial regression) δίνεται από την εξίσωση

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon. \tag{5.25}$$

για y τιμή της εξαρτημένης μεταβλητής Y και x τιμή της ανεξάρτητης μεταβλητής X. Όπως και στην γραμμική παλινδρόμηση υποθέτουμε πως το σφάλμα παλινδρόμησης με τιμή ϵ ακολουθεί κανονική κατανομή με μέση τιμή 0 και διασπορά σ_ϵ^2 και είναι ανεξάρτητο της X.

Το πρόβλημα της πολυωνυμικής παλινδρόμησης, όπως και για την γραμμική παλινδρόμηση, ορίζεται για τη δεσμευμένη μέση τιμή $\mathrm{E}(Y|X = x)$ υποθέτοντας πως η εξάρτηση δίνεται από την πολυωνυμική συνάρτηση ως

$$\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$$

Η υπόθεση κανονικής κατανομής του σφάλματος παλινδρόμησης μας επιτρέπει να εκτιμήσουμε διαστήματα εμπιστοσύνης και να κάνουμε ελέγχους

για τις παραμέτρους του μοντέλου, όπως και για να εκτιμήσουμε διαστήματα πρόβλεψης. Η μέθοδος ελαχίστων τετραγώνων όμως που χρησιμοποιείται για την σημειακή εκτίμηση των παραμέτρων $\beta_0, \beta_1, \ldots, \beta_k$ δεν προϋποθέτει κανονικότητα των σφαλμάτων για να δώσει τις καλύτερες σημειακές εκτιμήσεις των παραμέτρων. Η εκτίμηση των παραμέτρων $\beta_0, \beta_1, \ldots, \beta_k$, με τη μέθοδο ελαχίστων τετραγώνων γίνεται όπως και για το γραμμικό μοντέλο παλινδρόμησης γιατί ενώ η πολυωνυμική συνάρτηση παλινδρόμησης είναι μη-γραμμική ως προς την ανεξάρτητη μεταβλητή X, είναι γραμμική ως προς τους συντελεστές $\beta_0, \beta_1, \ldots, \beta_k$. Το άθροισμα των τετραγώνων των σφαλμάτων για κάποιο διμεταβλητό δείγμα μεγέθους n των (X, Y), είναι

$$f(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k) \right)^2 .$$

Οι μερικοί παράγωγοι της συνάρτησης αυτής ως προς κάθε παράμετρο β_0, $\beta_1, \ldots, \beta_k$, δίνει ένα σύστημα k εξισώσεων, γνωστό και ως σύστημα **κανονικών εξισώσεων** (normal equations) από το οποίο βρίσκουμε τις εκτιμήσεις $b_0, b_1, \ldots, b_k$

$$\begin{aligned} b_0 n + b_1 \sum x_i + b_2 \sum x_i^2 + \cdots b_k \sum x_i^k &= \sum y_i \\ b_0 \sum x_i + b_1 \sum x_i^2 + b_2 \sum x_i^3 + \cdots b_k \sum x_i^{k+1} &= \sum x_i y_i \\ \vdots \qquad \vdots \qquad \vdots \qquad \vdots & \qquad \vdots \\ b_0 \sum x_i^k + b_1 \sum x_i^{k+1} + b_2 \sum x_i^{k+2} + \cdots b_k \sum x_i^{2k} &= \sum x_i^k y_i \end{aligned} \tag{5.26}$$

Τα σφάλματα του μοντέλου πολυωνυμικής παλινδρόμησης που εκτιμήθηκε με τη μέθοδο ελαχίστων τετραγώνων είναι $e_i = y_i - \hat{y}_i$, όπου

$$\hat{y}_i = b_0 + b_1 x_i + b_2 x_i^2 + \cdots + b_k x_i^k .$$

Η εκτίμηση της διασποράς των σφαλμάτων e_i ορίζεται ως

$$s_e^2 = \frac{1}{n - (k+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{5.27}$$

Όπως και για την απλή γραμμική παλινδρόμηση και συσχέτιση, ορίζουμε τον **συντελεστή του πολλαπλού προσδιορισμού** (coefficient of multiple determination) R^2 ως

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{5.28}$$

που δηλώνει την αναλογία της μεταβλητότητας που εξηγείται από το μοντέλο.

Είναι φυσικό πως η πρόσθεση περισσότερων μη-γραμμικών όρων (δυνάμεων) της ανεξάρτητης μεταβλητής x στο πολυωνυμικό μοντέλο παλινδρόμησης θα βελτιώσει την προσαρμογή του στις n ζευγαρωτές παρατηρήσεις, χωρίς αυτό να σημαίνει ότι ένας μεγάλος βαθμός k είναι πάντα ο πιο κατάλληλος. Γι αυτό χρησιμοποιούμε τον **προσαρμοσμένο συντελεστή του πολλαπλού προσδιορισμού** (adjusted coefficient of multiple determination)

$$\text{adj}R^2 = 1 - \frac{n-1}{n-(k+1)} \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.29}$$

που δίνει μικρότερες τιμές από το R^2 της σχέσης (5.28) που το ποσό μείωσης προσαρμόζεται στο πλήθος των μη-γραμμικών όρων k.

Για τις παραμέτρους $\beta_0, \beta_1, \ldots, \beta_k$ μπορούμε να εκτιμήσουμε διαστήματα εμπιστοσύνης και να κάνουμε στατιστικούς ελέγχους όπως και για τις παραμέτρους του γραμμικού μοντέλου παλινδρόμησης, αφού υποθέσουμε πως τα σφάλματα ακολουθούν κανονική κατανομή με σταθερή διασπορά για κάθε τιμή της μεταβλητής X. Για την πολυωνυμική παλινδρόμηση η διασπορά της εκτίμησης για κάθε μια από τις $\beta_0, \beta_1, \ldots, \beta_k$ δίνεται από σχετικά πολύπλοκους τύπους που εξαρτώνται και από το βαθμό του πολυώνυμου. Το ίδιο ισχύει και για τα διαστήματα πρόβλεψης.

Παράδειγμα 5.6. Στον Πίνακα 5.6 δίνονται τα δεδομένα για την ημέρα της συγκομιδής (αριθμός ημερών αφού ανθίσει) και του μεγέθους σοδειάς (σε kg/ha) ενός είδους Ινδικού ρυζιού που λέγεται paddy.

Το διάγραμμα διασποράς δίνεται στο Σχήμα 5.12α. Η καμπύλη που φαίνεται να προσαρμόζεται καλύτερα είναι παραβολή, δηλαδή πολυώνυμο δευτέρου βαθμού. Προσαρμόζουμε πολυώνυμα από πρώτο ως τέταρτο βαθμό. Οι προσαρμογές δίνονται στο Σχήμα 5.12. Οι εκτιμήσεις των παραμέτρων δίνονται ως λύση των κανονικών εξισώσεων στη σχέση (5.31). Υπολογίζουμε τα σφάλματα προσαρμογής του κάθε μοντέλου και αντίστοιχα τον συντελεστή προσδιορισμού και τον προσαρμοσμένο συντελεστή προσδιορισμού, τα οποία και εμφανίζοντα με κάθε προσαρμογή μοντέλου στο Σχήμα 5.12.

Είναι φανερό πως το γραμμικό μοντέλο παλινδρόμησης δεν είναι κατάλληλο. Πράγματι τα σφάλματα παλινδρόμησης δεν είναι τυχαία αλλά θα σχηματίζουν καμπύλη, όπως στο Σχήμα 5.9ε. Επίσης ο συντελεστής προσδιορισμού είναι πολύ κοντά στο 0, που σημαίνει ότι το μοντέλο αδυνατεί πλήρως να ερμηνεύσει τις παρατηρήσεις. Η πρόσθεση του όρου του τετραγώνου των ημερών για τη συγκομιδή (ανεξάρτητη μεταβλητή) δίνει την παραβολή που προσαρμόζεται πολύ καλά στα δεδομένα. Ο συντελεστής προσδιορισμού R^2, καθώς και ο προσαρμοσμένος συντελεστής προσδιορισμού $\text{adj}R^2$, εκτοξεύτηκαν στο επίπεδο του 0.8, που σημαίνει ότι με το πολυωνυμικό μοντέλο δευτέρου βαθμού μπορούμε να εκτιμήσουμε τη σοδειά του paddy όταν δίνεται

Α/Α	*Ημέρες*	*Σοδειά*
1	16	2508
2	18	2518
3	20	3304
4	22	3423
5	24	3057
6	26	3190
7	28	3590
8	30	3883
9	32	3823
10	34	3646
11	36	3708
12	38	3333
13	40	3517
14	42	3241
15	44	3103
16	46	2776

Πίνακας 5.6: Τιμές ημερών για τη συγκομιδή και μεγέθους σοδειάς για το Ινδικό ρύζι paddy.

ο αριθμός ημερών για τη συγκομιδή. Η παραπέρα αύξηση του βαθμού του πολυωνύμου δε φαίνεται να βελτιώνει την παλινδρόμηση του μεγέθους σοδειάς του paddy προς τον αριθμό ημερών για τη συγκομιδή. Πράγματι, το R^2 παραμένει το ίδιο ενώ το adjR^2 μειώνεται.

Το πολυωνυμικό μοντέλο δευτέρου βαθμού εκτιμήθηκε να είναι

$$y = -1.1242 + 0.2979x - 0.0046x^2$$

και μπορούμε να το χρησιμοποιήσουμε για να κάνουμε προβλέψεις του μεγέθους της σοδειάς για κάθε δεδομένη χρονική περίοδο ως τη συγκομιδή.

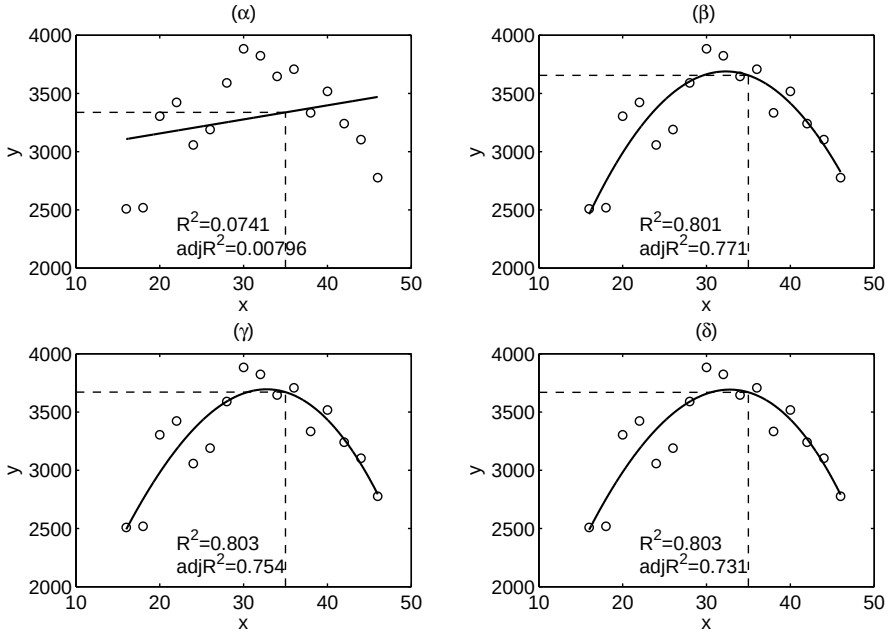
5.4 Πολλαπλή Παλινδρόμηση

Στην εξίσωση την πολυωνυμικής παλινδρόμησης που δίνεται στη σχέση (5.25), αν θέσουμε

$$x_1 = x, \quad x_2 = x^2, \quad \ldots \quad x_k = x^k,$$

η έκφραση της παλινδρόμησης γίνεται

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \tag{5.30}$$



Σχήμα 5.12: (α) Διάγραμμα διασποράς των 16 παρατηρήσεων του μεγέθους σοδειάς για το Ινδικό ρύζι paddy για δεδομένες ημέρες για τη συγκομιδή. Σχηματίζεται η ευθεία ελαχίστων τετραγώνων που προσαρμόζεται στα σημεία. Ο συντελεστής προσδιορισμού και ο προσαρμοσμένος συντελεστής προσδιορισμού δίνονται μέσα στο σχήμα. (β) Όπως το (α) αλλά για μοντέλο παλινδρόμησης από πολυώνυμο δευτέρου βαθμού. (γ) Όπως το (α) αλλά για πολυώνυμο τρίτου βαθμού. (δ) Όπως το (α) αλλά για πολυώνυμο τετάρτου βαθμού.

όπου και πάλι υποθέτουμε πως το σφάλμα παλινδρόμησης με τιμή ϵ έχει μέση τιμή 0 και διασπορά σ_e^2 (για να εκτιμήσουμε παραμετρικά διαστήματα εμπιστοσύνης ή να κάνουμε παραμετρικό έλεγχο θα χρειαστεί να υποθέσουμε πως το σφάλμα παλινδρόμησης ακολουθεί κανονική κατανομή).

Θεωρώντας πως οι τιμές $x_1, x_2, \ldots, x_k$ δεν είναι δυνάμεις της ίδιας μεταβλητής X αλλά είναι τιμές των ανεξάρτητων μεταβλητών $X_1, X_2, \ldots, X_k$, αντίστοιχα, ορίζεται η **γραμμική πολλαπλή παλινδρόμηση** (liner multiple regression) της εξαρτημένης μεταβλητής Y στις ανεξάρτητες μεταβλητές $X_1, X_2, \ldots, X_k$. Το μοντέλο της γραμμικής πολλαπλής παλινδρόμησης ορίζεται όπως και πριν για τη δεσμευμένη μέση τιμής της Y ως

$$\mathrm{E}(Y|X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Γενικά $X_1, X_2, \ldots, X_k$ αντιπροσωπεύουν διαφορετικά μεγέθη που πιθανόν να επηρεάζουν την Y και δε γνωρίζουμε αν είναι μεταξύ τους ανεξάρτητα.

Ένα μοντέλο πολλαπλής παλινδρόμησης μπορεί να περιέχει και μη-γραμμικούς όρους, όπως δυνάμεις των (υποθετικά) ανεξάρτητων μεταβλητών $X_1, X_2, \ldots, X_k$ και όρους αλληλεπίδρασης τους. Στη γενική του μορφή ένα τέτοιο μοντέλο λέγεται **μοντέλο προσθετικής πολλαπλής παλινδρόμησης** (additive multiple regression model), όπου ό όρος 'προσθετικής' τονίζει ότι όλοι οι όροι του μοντέλου συμπεριλαμβάνονται αθροιστικά στο μοντέλο. Ας δούμε ένα απλό παράδειγμα με δύο ανεξάρτητες μεταβλητές X_1 και X_2. Τα δυνατά μοντέλα προσθετικής πολλαπλής παλινδρόμησης είναι:

1. Το μοντέλο πρώτου πολυωνυμικού βαθμού (γραμμικής πολλαπλής παλινδρόμησης)
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

2. Το μοντέλο δεύτερου πολυωνυμικού βαθμού χωρίς αλληλεπίδραση
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \epsilon.$$

3. Το μοντέλο πρώτου πολυωνυμικού βαθμού με αλληλεπίδραση
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

4. Το πλήρες μοντέλο δευτέρου πολυωνυμικού βαθμού (με αλληλεπίδραση)
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon.$$

Υπάρχουν διαφορές μεταξύ των τεσσάρων αυτών μοντέλων ως προς τις μεταβλητές X_1 και X_2 αλλά όλα είναι γραμμικά ως προς τις παραμέτρους και η εκτίμηση των παραμέτρων μπορεί να γίνει με την κλασσική μέθοδο των ελαχίστων τετραγώνων. Το ίδιο βέβαια ισχύει και όταν οι μεταβλητές είναι περισσότερες από δύο.

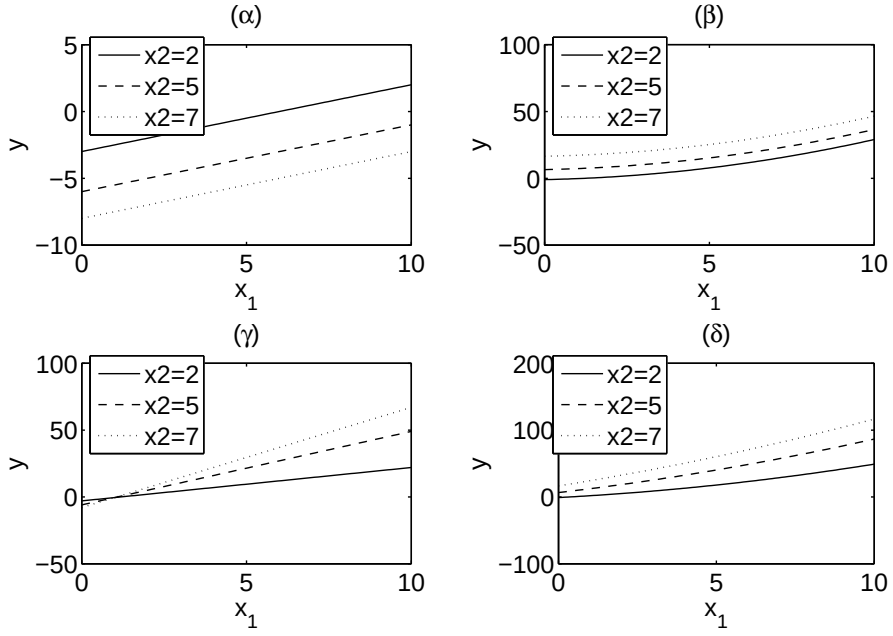
Σε προβλήματα πολλαπλής παλινδρόμησης όπου εμπλέκονται περισσότερες από μια ανεξάρτητες μεταβλητές δε μπορούμε να αποφασίσουμε εύκολα για τη μορφή του μοντέλου, όπως μπορούμε να κάνουμε για την απλή παλινδρόμηση της Y ως προς τη X, σχηματίζοντας το διάγραμμα διασποράς του ζεύγους (X, Y). Όταν οι ανεξάρτητες μεταβλητές είναι δύο, γραφική εκτίμηση του κατάλληλου μοντέλου μπορεί να γίνει από το γράφημα της (X_1, Y) για διαφορετικές τιμές της X_2 (ή ισοδύναμα αντιστρέφοντας τις θέσεις των X_1 και X_2). Για τα παραπάνω 4 μοντέλα δύο ανεξάρτητων μεταβλητών θα περιμέναμε τα εξής για τα γραφήματα του αιτιοκρατικού μέρους του κάθε μοντέλου (αγνοώντας την ύπαρξη του θορύβου ϵ):

1. Το γράφημα των σημείων (x_1, y) είναι σε ευθείες παράλληλες για κάθε τιμή του X_2 γιατί η μεταβολή της Y για κάποια μεταβολή της X_1 (π.χ. κατά μια μονάδα) είναι ανεξάρτητη της X_2 (δες Σχήμα 5.13α).

2. Το γράφημα των σημείων (x_1, y) είναι σε καμπύλες παραβολής παράλληλες για κάθε τιμή του X_2 λόγω της παρουσίας των τετραγωνικών όρων (δες Σχήμα 5.13β).

3. Το γράφημα των σημείων (x_1, y) είναι σε ευθείες για κάθε τιμή του X_2 που τέμνονται γιατί η μεταβολή της Y ως προς τη X_1 δεν είναι τώρα ανεξάρτητη της X_2 λόγω της παρουσίας του όρου αλληλεπίδρασης (δες Σχήμα 5.13γ).

4. Το γράφημα των σημείων (x_1, y) είναι σε καμπύλες παραβολής για κάθε τιμή του X_2 που δεν είναι παράλληλες λόγω της παρουσίας του όρου αλληλεπίδρασης (δες Σχήμα 5.13γ).



Σχήμα 5.13: Προβολή του γραφήματος της συνάρτησης παλινδρόμησης δύο ανεξάρτητων μεταβλητών (X_1, X_2) στο επίπεδο των (y, x_1) για τρεις τιμές του X_2. Η έκφραση της συνάρτησης είναι: (α) $y = -1 + 0.5x_1 - x_2$, (β) $y = -1 + 0.5x_1 + 25x_1^2 - x_2 + 0.5x_2^2$, (γ) $y = -1 + 0.5x_1 - x_2 + x_1x_2$, (δ) $y = -1 + 0.5x_1 + 25x_1^2 - x_2 + 0.5x_2^2 + x_1x_2$.

Όταν η παλινδρόμηση αφορά περισσότερες από δύο μεταβλητές τότε δεν έχουμε γραφικά εργαλεία για να καθορίσουμε τη μορφή του μοντέλου προσθετικής πολλαπλής παλινδρόμησης και πρέπει να δοκιμάσουμε διαφορετικά μοντέλα και να επιλέξουμε από αυτά που βρέθηκαν να είναι κατάλληλα (με-

τά από διαγνωστικό έλεγχο) και προσαρμόζονται καλά στα δεδομένα το πιο απλό.

5.4.1 Εκτίμηση μοντέλου πολλαπλής γραμμικής παλινδρόμησης

Είδαμε παραπάνω ότι σε κάθε περίπτωση ένα μοντέλο προσθετικής πολλαπλής παλινδρόμησης μπορεί να γραφτεί στη μορφή της σχέσης (5.30), όπου οι μεταβλητές $X_1, X_2, \ldots, X_k$, μπορεί να αντιπροσωπεύουν διαφορετικά μεγέθη ή δυνάμεις και αλληλεπιδράσεις τους.

Ας υποθέσουμε πως το πολυμεταβλητό δείγμα μεγέθους n είναι $\{x_{1i}, x_{2i}, \ldots, x_{ki}, y_i\}_{i=1}^{n}$. Η εκτίμηση των παραμέτρων του μοντέλου στη σχέση (5.30) γίνεται με τη μέθοδο ελαχίστων τετραγώνων. Το άθροισμα των τετραγώνων των σφαλμάτων είναι

$$f(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki})\right)^2 .$$

Το σύστημα κανονικών εξισώσεων που προκύπτει από τις μερικές παραγώγους της συνάρτησης αυτής ως προς κάθε παράμετρο $\beta_0, \beta_1, \ldots, \beta_k$, είναι

$$\begin{aligned}
b_0 n + b_1 \sum x_{1i} + b_2 \sum x_{2i} + \cdots b_k \sum x_{ki} &= \sum y_i \\
b_0 \sum x_{1i} + b_1 \sum x_{1i}^2 + b_2 \sum x_{1i}x_{2i} + \cdots b_k \sum x_{1i}x_{ki} &= \sum x_{1i}y_i \\
\vdots \qquad \vdots \qquad \vdots \qquad \vdots & \qquad \vdots \\
b_0 \sum x_{ki} + b_1 \sum x_{1i}x_{ki} + b_2 \sum x_{2i}x_{ki} + \cdots b_k \sum x_{ki}^2 &= \sum x_{ki}y_i
\end{aligned} \tag{5.31}$$

από το οποίο βρίσκουμε τις εκτιμήσεις $b_0, b_1, \ldots, b_k$.

Η εκτίμηση της εξαρτημένης μεταβλητής με το μοντέλο πολλαπλής παλινδρόμησης που εκτιμήθηκε με τη μέθοδο ελαχίστων τετραγώνων είναι

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}$$

και τα σφάλματα του μοντέλου είναι $e_i = y_i - \hat{y}_i$. Η εκτίμηση της διασποράς των σφαλμάτων s_e^2 ορίζεται όπως και για την πολυωνυμική παλινδρόμηση από τη σχέση (5.27) και αντίστοιχα ορίζεται ο συντελεστής του πολλαπλού προσδιορισμού R^2 από τη σχέση (5.28) και ο προσαρμοσμένος συντελεστής πολλαπλού προσδιορισμού $\text{adj}R^2$ από τη σχέση (5.29).

Είναι δυνατόν να υπολογίσουμε παραμετρικά διαστήματα εμπιστοσύνης για τις παραμέτρους $\beta_0, \beta_1, \ldots, \beta_k$ για τις οποίες όμως η εκτίμηση της διασποράς είναι σύνθετη. Γενικά αν η εκτίμηση της διασποράς είναι $s_{b_j}^2$ για

$j = 0, 1, \ldots, k$ τότε το $(1 - a)\%$ διάστημα εμπιστοσύνης για το συντελεστή β_j είναι

$$b_j \pm t_{n-(k+1),1-a/2} s_{b_j}.$$

Ο έλεγχος για μια τιμή β_j^0 της β_j, H_0: $\beta_j = \beta_j^0$ γίνεται με το στατιστικό

$$t = \frac{\beta_j - \beta_j^0}{s_{b_j}} \sim t_{n-(k+1)}.$$

Το $(1 - a)\%$ διάστημα εμπιστοσύνης για τη μέση τιμή της y όταν δίνονται τα $x_1, \ldots, x_k$ είναι

$$\hat{y} \pm t_{n-(k+1),1-a/2} s_{\hat{y}}$$

όπου η διασπορά της εκτίμησης $\hat{y}$, $s^2_{\hat{y}}$, δίνεται από επίσης σύνθετη έκφραση. Αντίστοιχα το $(1 - a)\%$ διάστημα πρόβλεψης μιας (μελλοντικής) τιμής της y είναι

$$\hat{y} \pm t_{n-(k+1),1-a/2} \sqrt{s_e^2 + s^2_{\hat{y}}}.$$

Παράδειγμα 5.7. Σε μελέτη της επίδρασης γεωργικών χημικών στην προσρόφηση ιζημάτων και εδάφους, δίνονται στον Πίνακα 5.7 13 δεδομένα για το δείκτη προσρόφησης φωσφορικού άλατος (Y), για το εξαγώγιμο σίδηρο (X_1) και το εξαγώγιμο αργίλιο (X_2).

Α/Α	*Εξαγώγιμο σίδηρο*	*Εξαγώγιμο αργίλιο*	*Δείκτης προσρόφησης*
1	61	13	4
2	175	21	18
3	111	24	14
4	124	23	18
5	130	64	26
6	173	38	26
7	169	33	21
8	169	61	30
9	160	39	28
10	244	71	36
11	257	112	65
12	333	88	62
13	199	54	40

Πίνακας 5.7: Τιμές του δείκτη προσρόφησης, του εξαγώγιμου σιδήρου και εξαγώγιμου αργιλίου για τη μελέτη της επίδρασης γεωργικών χημικών στο έδαφος.

Το μοντέλο που θα εκτιμήσουμε είναι

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

Οι εκτιμήσεις των συντελεστών του μοντέλου με τη μέθοδο των ελαχίστων τετραγώνων καθώς και η εκτίμηση της τυπικής τους απόκλισης δίνονται στον Πίνακα 5.8.

Παράμετρος	*Εκτιμητής* b_i	*Εκτίμηση SD* s_{b_i}
β_0	-7.351	3.485
β_1	0.11273	0.02969
β_2	0.34900	0.07131

Πίνακας 5.8: Εκτίμηση παραμέτρων και τυπική απόκλισης τους.

Το 95% διάστημα εμπιστοσύνης για το συντελεστή του εξαγώγιμου σιδήρου β_1 είναι ($t_{10,0.975} = 2.228$)

$$0.11273 \pm 2.228 \cdot 0.02969 = [0.0466, 0.1789]$$

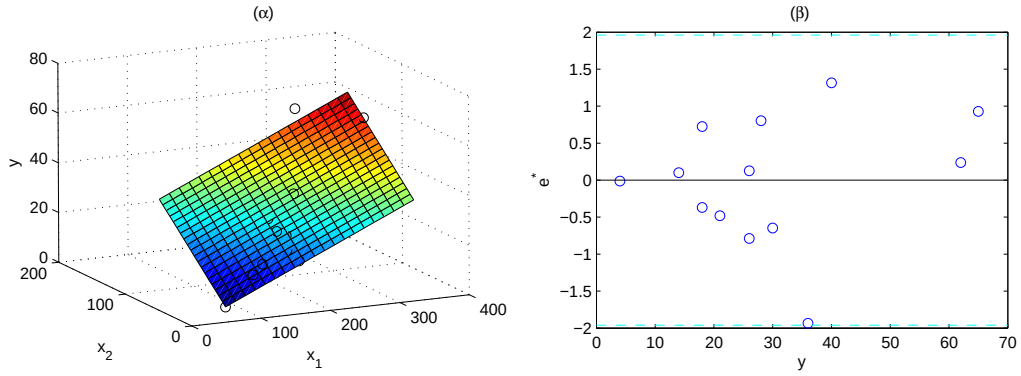
και αντίστοιχα για το συντελεστή του εξαγώγιμου αργιλίου β_2 είναι

$$0.34900 \pm 2.228 \cdot 0.07131 = [0.1901, 0.5079].$$

Παρατηρούμε πως και τα δύο διαστήματα εμπιστοσύνης δεν περιέχουν το 0, αλλά μπορούμε με βεβαιότητα σε επίπεδο 95% να συμπεράνουμε πως το εξαγώγιμο σίδηρο και αργίλιο επηρεάζουν σημαντικά το δείκτη προσρόφησης φωσφορικού άλατος και καλώς συμπεριλαμβάνονται στο μοντέλο. Πράγματι το γράφημα του μοντέλου στο χώρο που ορίζεται από τα σημεία (x_1, x_2, y) είναι ένα επίπεδο που δεν είναι παράλληλο ούτε προς τον άξονα x_1 (εξαγώγιμο σίδηρο) ούτε και προς τον άξονα x_2 (εξαγώγιμο αργίλιο), όπως φαίνεται στο Σχήμα 5.14α. Το επίπεδο φαίνεται να προσαρμόζεται ικανοποιητικά στα 13 σημεία του δείγματος. Πράγματι η τυπική απόκλιση των σφαλμάτων είναι $s_e = 4.616$, ο συντελεστής του πολλαπλού προσδιορισμού είναι $R^2 = 0.948$ και ο προσαρμοσμένος συντελεστής είναι $adjR^2 = 0.931$. Επίσης, όπως φαίνεται στο διαγνωστικό διάγραμμα διασποράς των τυποποιημένων σφαλμάτων του μοντέλου προς τις τιμές του y στο Σχήμα 5.14β, τα σφάλματα είναι μέσα στα 95% όρια της κανονικής κατανομής και κατανέμονται τυχαία ως προς y.

Έστω ότι θέλουμε να προβλέψουμε το δείκτη προσρόφησης y όταν ο εξαγώγιμος σίδηρος είναι $x_1 = 160$ και ο εξαγώγιμος αργίλιου είναι $x_2 = 39$. Η πρόβλεψη είναι

$$\hat{y} = -7.351 + 0.11273 \cdot 160 + 0.34900 \cdot 39 = 24.30.$$



Σχήμα 5.14: (α) Διάγραμμα διασποράς των 13 παρατηρήσεων του δείκτη προσρόφησης (Y) για δοθείσες τιμές του εξαγώγιμου σιδήρου (X_1) και του εξαγώγιμου αργιλίου (X_2). Δίνεται το γράφημα του μοντέλου γραμμικής πολλαπλής παλινδρόμησης που εκτιμήθηκε με τη μέθοδο ελαχίστων τετραγώνων. (β) Διαγνωστικό διάγραμμα διασποράς των τυποποιημένων σφαλμάτων του μοντέλου στο (α) προς τις αντίστοιχες τιμές του Y.

Η εκτίμηση της τυπικής απόκλισης για αυτήν την πρόβλεψη $\hat{y}$ βρέθηκε να είναι $s_{\hat{y}} = 1.30$. Το 95% διάστημα εμπιστοσύνης για το μέσο δείκτη προσρόφησης y όταν ο εξαγώγιμος σίδηρος είναι $x_1 = 160$ και ο εξαγώγιμος αργίλιου είναι $x_2 = 39$ βρίσκεται ως

$$24.30 \pm 2.228 \cdot 1.30 = [21.40, 27.20]$$

και το αντίστοιχο 95% διάστημα πρόβλεψης για μια μελλοντική τιμή του y (για $x_1 = 160$ και $x_2 = 39$) είναι

$$24.30 \pm 2.228 \cdot \sqrt{(4.616)^2 + (1.30)^2} = [13.62, 34.98]$$

5.4.2 Επιλογή μεταβλητών

Σε κάποια προβλήματα παλινδρόμησης έχουμε στη διάθεση μας δεδομένα από πολλούς παράγοντες που μπορεί να επηρεάζουν την εξαρτημένη μεταβλητή που μας ενδιαφέρει να καθορίσουμε ή να προβλέψουμε. Θα θέλαμε λοιπόν να επιλέξουμε το μικρότερο δυνατόν υποσύνολο ανεξάρτητων (επεξηγηματικών) μεταβλητών που εξηγεί σχεδόν το ίδιο καλά την εξαρτημένη μεταβλητή όπως και μεγαλύτερα υποσύνολα ανεξάρτητων μεταβλητών ή ακόμα και ολόκληρο το σύνολο ανεξάρτητων μεταβλητών.

Σε μια πρώτη προσέγγιση η λύση είναι να προσαρμόσει κάποιος όλα τα δυνατά μοντέλα, δηλαδή για όλους του συνδυασμούς υποσυνόλων των ανεξάρτητων μεταβλητών, και να βρει αυτό που προσαρμόζεται καλύτερα. Το

κριτήριο προσαρμογής θα πρέπει να δίνει κάποια μορφή ποινής σε πιο πολύπλοκα μοντέλα (με περισσότερες ανεξάρτητες μεταβλητές) όπως ο προσαρμοσμένος συντελεστής πολλαπλού προσδιορισμού adjR^2. Αυτή η απλοϊκή προσέγγιση μπορεί να είναι αρκετά επίπονη όταν το πλήθος των ανεξάρτητων μεταβλητών είναι μεγάλο και συνήθως δεν ακολουθείται.

Έχουν προταθεί διάφορες μέθοδοι επιλογής των ανεξάρτητων μεταβλητών που υπολογίζουν το βέλτιστο μοντέλο πολλαπλής παλινδρόμησης βηματικά (stepwise regression). Όλες αυτές οι μέθοδοι εφαρμόζουν διαδοχικούς ελέγχους για το αν κάποια ανεξάρτητη μεταβλητή X_j είναι σημαντική, δηλαδή H_0: $\beta_j = 0$. Για παράδειγμα η **μέθοδος απαλοιφής προς τα πίσω** (backward elimination) αρχίζει με το μοντέλο να περιέχει όλες τις ανεξάρτητες μεταβλητές $X_1, X_2, \ldots, X_k$ και με διαδοχικούς ελέγχους απαλοίφει κάθε φορά μια ανεξάρτητη μεταβλητή όταν φαίνεται να 'περισσεύει', δηλαδή να μην εξηγεί την y όταν παρουσιάζεται με άλλες ανεξάρτητες μεταβλητές στο μοντέλο. Οι στατιστικοί έλεγχοι μπορούν επίσης να γίνουν με αντίστροφη πορεία, ξεκινώντας από το απλό μοντέλο σταθερού όρου και προσθέτοντας κάθε φορά μια ανεξάρτητη μεταβλητή στο μοντέλο που φαίνεται να είναι η πιο σημαντική για να εξηγήσει τη Y όταν ήδη υπάρχουν κάποιες άλλες στο μοντέλο. Αυτή είναι η μέθοδος της **επιλογής προς τα μπρος** (forward selection).

Ένα πρόβλημα που είναι συνδεδεμένο με την επιλογή μεταβλητών είναι η **πολλαπλή συγραμμικότητα** (multicollinearity), δηλαδή δύο ή περισσότερες από τις k ανεξάρτητες μεταβλητές του μοντέλου, να είναι κατά όνομα ανεξάρτητες αλλά να είναι ισχυρά αλληλεξαρτημένες. Αυτό μπορεί κάποιος να το διαπιστώσει προσαρμόζοντας ένα μοντέλο παλινδρόμησης σε κάθε μια από τις ύποπτες για αλληλεξάρτηση μεταβλητές X_j ως προς όλες τις υπόλοιπες. Αν η X_j μπορεί να προβλεφθεί καλά από τις υπόλοιπες $k-1$ ανεξάρτητες μεταβλητές (μεγάλο R^2 ή adjR^2), τότε τα δεδομένα έχουν το πρόβλημα της πολλαπλής συγραμμικότητας. Πολλές φορές το πρόβλημα της συγραμμικότητας παραβλέπεται σε προβλήματα παλινδρόμησης. Γενικά δεν υπάρχει συγκεκριμένη μέθοδος αντιμετώπισης της πολλαπλής συγραμμικότητας.

5.4.3 Άλλα μη-γραμμικά μοντέλα

Τα μοντέλα προσθετικής (πολλαπλής) παλινδρόμησης που εξετάσαμε έχουν ως κοινό σημείο ότι είναι γραμμικά ως προς τις παραμέτρους τους. Σε κάποιο πρόβλημα όπου οι παράμετροι φαίνονται να εμπλέκονται μη-γραμμικά, είναι ενδεχομένως δυνατόν με κατάλληλους μετασχηματισμούς να φέρουμε το μοντέλο σε προσθετική μορφή. Τέτοια μοντέλα αποτελούν τη γενίκευση της εγγενής γραμμική συνάρτησης παλινδρόμησης σε περισσότερες ανεξάρτητες μεταβλητές. Σε άλλες περιπτώσεις η μορφή του μοντέλου δεν απλοποιείται και τότε πρέπει να υπολογιστούν οι παράμετροι με κάποια μέθοδο

μη-γραμμικής βελτιστοποίησης.

Υπάρχουν και άλλες κλάσεις μοντέλων που δεν έχουν κάποια γνωστή αναλυτική μορφή αλλά δίνονται ως άθροισμα διαφορετικών βασικών συναρτήσεων, όπως τα **νευρωνικά δίκτυα** (neural networks). Τέλος υπάρχουν και μη παραμετρικά μοντέλα που κάνουν εκτίμηση ή πρόβλεψη για τις δεδομένες τιμές των ανεξάρτητων μεταβλητών χρησιμοποιώντας από τα υπάρχοντα δεδομένα αυτά που είναι 'γειτονικά'. Τέτοια μοντέλα είναι τα μοντέλα **πυρήνων** (kernels).

Ασκήσεις Κεφαλαίου 5

1. Δημιουργείστε $M = 1000$ δείγματα μεγέθους $n = 20$ ζευγαρωτών παρατηρήσεων των (X, Y) από διμεταβλητή κανονική κατανομή με μέσες τιμές $\mu_X = 0$, $\mu_Y = 0$, τυπικές αποκλίσεις $\sigma_X = 1$, $\sigma_Y = 1$ και για δύο τιμές του συντελεστή συσχέτισης, $\rho = 0$ και $\rho = 0.5$.

   (α΄) Υπολογίστε το παραμετρικό 95% διάστημα εμπιστοσύνης κάνοντας χρήση του μετασχηματισμού Fisher για κάθε ένα από τα M δείγματα. Κάνετε το ιστόγραμμα για κάθε ένα από τα δύο άκρα του διαστήματος εμπιστοσύνης. Σε τι ποσοστό το διάστημα εμπιστοσύνης περιέχει τη πραγματική τιμή του ρ; Κάνετε τους υπολογισμούς ξεχωριστά για $\rho = 0$ και $\rho = 0.5$.

   (β΄) Κάνετε έλεγχο της υπόθεσης για μηδενική συσχέτιση των X και Y χρησιμοποιώντας το στατιστικό της κατανομής Student t της σχέσης (5.5) για κάθε ένα από τα M δείγματα. Σε τι ποσοστό απορρίπτεται η μηδενική υπόθεση; Κάνετε τους υπολογισμούς ξεχωριστά για $\rho = 0$ και $\rho = 0.5$.

   (γ΄) Επαναλάβετε τους παραπάνω υπολογισμούς για δείγματα μεγέθους $n = 200$. Υπάρχει διαφορά στα αποτελέσματα του διαστήματος εμπιστοσύνης και στατιστικής υπόθεσης;

   (δ΄) Επαναλάβετε τους παραπάνω υπολογισμούς για δείγματα μεγέθους $n = 20$ και $n = 200$ αλλά παίρνοντας τα τετράγωνα των παρατηρήσεων, δηλαδή θεωρείστε τις τ.μ. X^2 και Y^2. Υπάρχει διαφορά στα αποτελέσματα του διαστήματος εμπιστοσύνης και στατιστικής υπόθεσης από τα αντίστοιχα για τις τ.μ. X και Y;

2. Μελετήσαμε τον παραμετρικό έλεγχο για ανεξαρτησία (ή καλύτερα μηδενική συσχέτιση) δύο τ.μ. X και Y κάνοντας χρήση του στατιστικού t της σχέσης (5.5) και θεωρώντας ότι ακολουθεί κατανομή Student. Μπορούμε να κάνουμε τον έλεγχο χωρίς να θεωρήσουμε γνωστή κατανομή του στατιστικού κάτω από τη μηδενική υπόθεση και αυτός λέγεται μη-παραμετρικός έλεγχος. Θα χρησιμοποιήσουμε έναν τέτοιο έλεγχο που λέγεται έλεγχος τυχαιοποίησης και θα δημιουργήσουμε L τυχαιοποιημένα δείγματα από το αρχικό μας διμεταβλητό δείγμα των X και Y σύμφωνα με την μηδενική υπόθεση. Για αυτό θα αλλάξουμε τυχαία τη σειρά όλων των παρατηρήσεων της μιας από τις δύο τ.μ. στο δείγμα, και θα το κάνουμε αυτό L φορές. Στη συνέχεια θα υπολογίσουμε το στατιστικό t της σχέσης (5.5) στο αρχικό δείγμα, έστω t_0, αλλά και στα L τυχαιοποιημένα δείγματα, έστω $t_1, \ldots, t_L$. Η μηδενική υπόθεση απορρίπτεται αν η τιμή t_0 δεν περιέχεται στην κατανομή των $t_1, \ldots, t_L$,

δηλαδή στην εμπειρική (μη-παραμετρική) κατανομή του t κάτω από τη μηδενική υπόθεση της ανεξαρτησίας των X και Y. Συγκεκριμένα για επίπεδο σημαντικότητας a θα εξετάσουμε αν το t_0 είναι μεταξύ των $a/2\%$ και $(1 - a/2)\%$ ποσοστιαίων σημείων, δηλαδή μεταξύ των σημείων με σειρά $La/2$ και $L(1 - a/2)$ (προσεγγιστικά στον πλησιέστερο ακέραιο), όταν βάζουμε τα $t_1, \ldots, t_L$ σε αύξουσα σειρά.

Θεωρείστε $L = 1000$ τυχαιοποιημένα δείγματα και επαναλάβετε τον έλεγχο, αλλά τώρα τυχαιοποιημένο αντί για παραμετρικό, για την άσκηση 1, για την περίπτωση $n = 20$, και για τα ζευγάρια (X, Y) και (X^2, Y^2). Υπάρχουν διαφορές στα αποτελέσματα;

Βοήθεια (`matlab`*):* Για τη δημιουργία ενός τυχαιοποιημένου δείγματος αρκεί να αλλάξετε τυχαία τη σειρά των παρατηρήσεων της μιας τ.μ.. Μπορείτε να δημιουργείστε μια τυχαιοποιημένη σειρά δεικτών $1, \ldots, n$ με την συνάρτηση `randperm` και όρισμα n. Έτσι αν x είναι το διάνυσμα των n παρατηρήσεων της X και y το αντίστοιχο της Y, τότε αν r = randperm(n) είναι το διάνυσμα τυχαίων δεικτών, θέτοντας xr = x(r) δημιουργούμε το τυχαιοποιημένο δείγμα των xr και y.

3. Δίνονται οι μέσες μηνιαίες τιμές θερμοκρασίας και βροχόπτωσης στη Θεσσαλονίκη για την περίοδο 1959 - 1997. Ελέγξετε αν υπάρχει συσχέτιση μεταξύ της θερμοκρασίας και της βροχόπτωσης για κάθε μήνα ξεχωριστά. Χρησιμοποιείστε το στατιστικό t της σχέσης (5.5) και κάνετε παραμετρικό και έλεγχο τυχαιοποίησης (σύμφωνα με την άσκηση 2). Τα δεδομένα δίνονται στην ιστοσελίδα του μαθήματος σε δύο αρχεία πινάκων, ένας για τη θερμοκρασία και ένας για τη βροχόπτωση, που έχουν 39 γραμμές για τα 39 έτη και 12 στήλες για τους 12 μήνες κάθε έτους, από Ιανουάριο ως Δεκέμβριο.

4. Στο αρχείο `lightair.dat` στην ιστοσελίδα του μαθήματος δίνονται 100 μετρήσεις της πυκνότητας του αέρα (σε $\mathrm{kg/m^3}$) σε διαφορετικές συνθήκες θερμοκρασίας και πίεσης (στην πρώτη στήλη) και οι αντίστοιχες μετρήσεις της ταχύτητας φωτός (-299000 km/sec) στη δεύτερη στήλη.

   (α′) Σχεδιάστε το κατάλληλο διάγραμμα διασποράς και υπολογίστε τον αντίστοιχο συντελεστή συσχέτισης.

   (β′) Εκτιμήστε το μοντέλο γραμμικής παλινδρόμησης με τη μέθοδο ελαχίστων τετραγώνων για τη γραμμική εξάρτηση της ταχύτητας φωτός από την πυκνότητα του αέρα. Υπολογίστε παραμετρικό διάστημα εμπιστοσύνης σε επίπεδο 95% για τους δύο συντελεστές

της ευθείας ελαχίστων τετραγώνων (διαφορά ύψους β_0 και κλίση β_1).

(γ΄) Σχηματίστε στο διάγραμμα διασποράς την ευθεία ελαχίστων τετραγώνων, τα όρια πρόβλεψης σε επίπεδο 95% για τη μέση ταχύτητα φωτός καθώς και για μια τιμή της ταχύτητας φωτός. Επίσης κάνετε πρόβλεψη για πυκνότητα αέρα 1.29 $\mathrm{kg/m^3}$ δίνοντας και τα όρια μέσης τιμής και μιας παρατήρησης της ταχύτητας φωτός.

(δ΄) Η πραγματική σχέση της ταχύτητας φωτός στον αέρα με την πυκνότητα του αέρα είναι:

$$c_{air} = c\left(1 - 0.00029\frac{d}{d_0}\right),$$

όπου οι δύο σταθερές είναι:

- $c = 299792.458$ km/sec, η ταχύτητα φωτός στο κενό, και
- $d_0 = 1.29$ $\mathrm{kg/m^3}$, η πυκνότητα του αέρα σε θερμοκρασία και πίεση δωματίου.

Από την παραπάνω σχέση υπολογίστε την εξίσωση της πραγματικής ευθείας παλινδρόμησης της ταχύτητας φωτός στον αέρα ως προς την πυκνότητα του αέρα. Στη συνέχεια κάνετε έλεγχο ξεχωριστά για κάθε συντελεστή της πραγματικής ευθείας παλινδρόμησης αν τον δεχόμαστε με βάση το δείγμα των 100 ζευγαρωτών παρατηρήσεων (σύμφωνα με τις εκτιμήσεις τους στο 4β΄). Είναι οι πραγματικές μέσες τιμές της ταχύτητας φωτός μέσα στα όρια μέσης πρόβλεψης για κάθε τιμή πυκνότητας αέρα στο δείγμα; (σύμφωνα με το διάστημα μέσης πρόβλεψης που υπολογίσατε και σχηματίσατε στο 4γ΄).

5. Θα υπολογίσουμε μη παραμετρικό διάστημα εμπιστοσύνης για τους δύο συντελεστές της ευθείας ελαχίστων τετραγώνων (διαφορά ύψους β_0 και κλίση β_1) και θα το εφαρμόσουμε στα δεδομένα της προηγούμενης άσκησης (εξάρτηση της ταχύτητας φωτός από την πυκνότητα του αέρα). Η μέθοδος που θα χρησιμοποιήσουμε λέγεται bootstrap και για τον υπολογισμό των διαστημάτων εμπιστοσύνης των β_0 και β_1 ορίζεται ως εξής:

   - Πάρε ένα νέο δείγμα 100 τυχαίων ζευγαρωτών παρατηρήσεων από το δείγμα των 100 ζευγαρωτών παρατηρήσεων (πυκνότητας αέρα και ταχύτητας φωτός). Το κάθε ζευγάρι επιλέγεται τυχαία με επανάθεση, δηλαδή το ίδιο ζευγάρι μπορεί να εμφανιστεί πολλές φορές στο νέο δείγμα (και άλλο ζευγάρι να μην εμφανιστεί καθόλου).

- Υπολόγισε τις εκτιμήσεις b_0 και b_1 των συντελεστών της ευθείας ελαχίστων τετραγώνων για αυτό το νέο δείγμα.
- Επανέλαβε τα παραπάνω δύο βήματα $M = 1000$ φορές.
- Από τις M τιμές για το b_0 υπολόγισε τα όρια του $(1 - a)\%$ (εδώ $a = 0.05$) διαστήματος εμπιστοσύνης για το β_0 από τα ποσοστιαία σημεία $a/2\%$ και $(1-a/2)\%$. Για αυτό βάλε τις $M = 1000$ τιμές σε αύξουσα σειρά και βρες τις τιμές για τάξη $Ma/2\%$ και $M(1-a/2)\%$. Κάνε το ίδιο για το b_1.

Σύγκρινε τα bootstrap διαστήματα εμπιστοσύνης των β_0 και β_1 με τα παραμετρικά που βρήκες στην προηγούμενη άσκηση.

Βοήθεια (`matlab`*):* Για τη δημιουργία n τυχαίων αριθμών από 1 ως N με επανάθεση, χρησιμοποίησε τη συνάρτηση `unidrnd` με κατάλληλα ορίσματα ως `unidrnd(N,n,1)`. Θα δώσει το διάνυσμα δεικτών για το νέο δείγμα από τα n στοιχεία του αρχικού δείγματος (εδώ τα στοιχεία είναι οι ζευγαρωτές παρατηρήσεις).

6. Στον παρακάτω πίνακα δίνονται τα δεδομένα για το ποσοστό υψηλής επίδοσης που ακόμα έχουν ελαστικά (με ακτινωτή ενίσχυση) ενώ έχουν ήδη χρησιμοποιηθεί για τα αντίστοιχα χιλιόμετρα.

Α/Α	*Απόσταση* σε χιλιάδες km	*ποσοστό δυνατότητας* χρήσης
1	2	98.2
2	3	91.7
3	8	81.3
4	16	64.0
5	32	36.4
6	48	32.6
7	64	17.1
8	80	11.3

(α΄) Κάνε το διάγραμμα διασποράς και προσάρμοσε το κατάλληλο μοντέλο για τον προσδιορισμό του ποσοστού δυνατότητας χρήσης (υψηλής επίδοσης) προς τα αντίστοιχα χιλιόμετρα χρήσης του ελαστικού. Για να ελέγξεις την καταλληλότητα του μοντέλου, κάνε διαγνωστικό διάγραμμα διασποράς των τυποποιημένων σφαλμάτων προσαρμογής του επιλεγμένου μοντέλου προς την εξαρτημένη μεταβλητή (ποσοστό δυνατότητας χρήσης).

(β΄) Πρόβλεψε το ποσοστό δυνατότητας χρήσης για ελαστικό που χρησιμοποιήθηκε για 25000km.

7. Ο θερμοστάτης είναι αντιστάτης με αντίσταση που εξαρτιέται από τη θερμοκρασία. Είναι φτιαγμένος (συνήθως) από ημιαγωγό υλικό με ενεργειακό διάκενο E_g. Η αντίσταση R του θερμοστάτη αλλάζει σύμφωνα με τη σχέση

$$R \propto R_0 e^{E_g/2kT},$$

όπου T είναι η θερμοκρασία (σε ^{o}K) και R_0, k είναι σταθερές. Για κατάλληλες παραμέτρους β_0 και β_1 ($\beta_1 = 2k/E_g$) η παραπάνω εξίσωση μπορεί να απλοποιηθεί στη γραμμική μορφή

$$\frac{1}{T} = \beta_0 + \beta_1 \ln(R).$$

Το ενεργειακό διάκενο E_g έχει κάποια μικρή εξάρτηση από τη θερμοκρασία έτσι ώστε η παραπάνω έκφραση να μην είναι ακριβής. Διορθώσεις μπορούν να γίνουν προσθέτοντας πολυωνυμικούς όρους του $\ln(R)$.

Στον παρακάτω πίνακα δίνονται 32 μετρήσεις της αντίστασης R (σε Ωμ) και της θερμοκρασίας σε ^{o}C (θα πρέπει να μετατραπούν σε ^{o}K, δηλαδή να προστεθεί σε κάθε τιμή 273.15).

(α΄) Βρείτε το κατάλληλο πολυωνυμικό μοντέλο της παλινδρόμησης του $1/T$ ως προς $\ln(R)$, κάνοντας διαγνωστικό έλεγχο με το διάγραμμα διασποράς των τυποποιημένων υπολοίπων προς $1/T$ για κάθε μοντέλο που δοκιμάζετε (πρώτου βαθμού, δευτέρου βαθμού κτλ).

(β΄) Συγκρίνετε την προσαρμογή και καταλληλότητα του μοντέλου που καταλήξατε με το μοντέλο του Steinhart-Hart που δίνεται από την εξίσωση

$$\frac{1}{T} = \beta_0 + \beta_1 \ln(R) + \beta_3 (\ln(R))^3.$$

Α/Α	*Αντίσταση*	*θερμοκρασία (σε ^oC)*
1	0.76	110
2	0.86	105
3	0.97	100
4	1.11	95
5	1.45	85
6	1.67	80
7	1.92	75
8	2.23	70
9	2.59	65
10	3.02	60
11	3.54	55
12	4.16	50
13	4.91	45
14	5.83	40
15	6.94	35
16	8.31	30
17	10.00	25
18	12.09	20
19	14.68	15
20	17.96	10
21	22.05	5
22	27.28	0
23	33.89	-5
24	42.45	-10
25	53.39	-15
26	67.74	-20
27	86.39	-25
28	111.30	-30
29	144.00	-35
30	188.40	-40
31	247.50	-45
32	329.20	-50

8. Μετρήθηκε το βάρος και 10 δείκτες σώματος σε 22 άνδρες νεαρής ηλικίας. Τα δεδομένα δίνονται στην ιστοσελίδα του μαθήματος στο αρχείο `physical.txt`. Η πρώτη γραμμή του πίνακα του αρχείου έχει τα ονόματα των δεικτών σε κάθε στήλη δεδομένων και δίνονται στον παρακάτω πίνακα.

Α/Α	*Όνομα*	*Περιγραφή*
1	Mass	Βάρος σε κιλά
2	Fore	μέγιστη περιφέρεια του πήχη χεριού
3	Bicep	μέγιστη περιφέρεια του δικέφαλου μυ
4	Chest	περιμετρική απόσταση στήθους (στο ύψος κάτω από τις μασχάλες)
5	Neck	περιμετρική απόσταση λαιμού (στο μέσο ύψος λαιμού)
6	Shoulder	περιμετρική απόσταση ώμου
7	Waist	περιμετρική απόσταση μέσης (οσφίου)
8	Height	ύψος από την κορυφή στα δάχτυλα ποδιού
9	Calf	μέγιστη περιφέρεια κνήμης
10	Thigh	περιμετρική απόσταση γοφού
11	Head	περιμετρική απόσταση κεφαλιού

Διερευνήστε το κατάλληλο μοντέλο γραμμικής παλινδρόμησης για το βάρος. Δοκιμάστε το μοντέλο με τις 10 ανεξάρτητες μεταβλητές και συγκρίνετε το με το μοντέλο που δίνει κάποια μέθοδος βηματικής παλινδρόμησης. Υπολογίστε για το κάθε μοντέλο τις εκτιμήσεις των παραμέτρων, τη διασπορά των σφαλμάτων και το συντελεστή προσδιορισμού (καθώς και τον προσαρμοσμένο συντελεστή προσδιορισμού).

Βοήθεια (`matlab`*):* Για να εφαρμόσετε βηματική παλινδρόμηση, το `matlab` παρέχει γραφικό περιβάλλον με την εντολή `stepwise` και κάνει τους ίδιους υπολογισμούς στη συνάρτηση `stepwisefit`. Για να φορτώσετε τα δεδομένα του αρχείου με την εντολή `load` θα πρέπει πρώτα να διαγράψετε την πρώτη σειρά με τα ονόματα των μεταβλητών.

9. Μετρήθηκαν σε 12 νοσοκομεία των ΗΠΑ οι μηνιαίες ανθρωποώρες που σχετίζονται με την υπηρεσία αναισθησιολογίας, καθώς και άλλοι δείκτες που ενδεχομένως επηρεάζουν την απασχόληση προσωπικού στην υπηρεσία αναισθησιολογίας. Τα δεδομένα δίνονται στην ιστοσελίδα του μαθήματος στο αρχείο `hospital.txt`. Η πρώτη γραμμή του πίνακα του αρχείου έχει τα ονόματα των δεικτών σε κάθε στήλη δεδομένων και δίνονται στον παρακάτω πίνακα.

Α/Α	*Όνομα*	*Περιγραφή*
1	ManHours	οι ανθρωποώρες στην υπηρεσία αναισθησιολογίας
2	Cases	τα περιστατικά χειρουργείου μηνιαία
3	Eligible	ο πληθυσμός που εξυπηρετείται ανά χιλιάδες
4	OpRooms	οι αίθουσες χειρουργείου

Διερευνήστε το κατάλληλο μοντέλο γραμμικής παλινδρόμησης για τις ανθρωποώρες. Δοκιμάστε το μοντέλο με τις 3 ανεξάρτητες μεταβλητές και συγκρίνετε το με το μοντέλο που δίνει κάποια μέθοδος βηματικής παλινδρόμησης. Υπολογίστε για το κάθε μοντέλο τις εκτιμήσεις των παραμέτρων, τη διασπορά των σφαλμάτων και το συντελεστή προσδιορισμού (καθώς και τον προσαρμοσμένο συντελεστή προσδιορισμού). Επίσης διερευνήστε το φαινόμενο πολλαπλής συγραμμικότητας για τους 3 δείκτες.

Κεφάλαιο 6

Μείωση διάστασης

Ασκήσεις Κεφαλαίου 6

1. Δημιουργείστε δείγμα $n = 1000$ παρατηρήσεων από διμεταβλητή κανονική κατανομή με διασπορά της πρώτης τ.μ. 1 και της δεύτερης τ.μ. 4. Στη συνέχεια μετασχηματίστε τα δεδομένα από το $\mathbf{R}^2$ στο $\mathbf{R}^3$ με τον πίνακα

$$W = \begin{bmatrix} 0.2 & 0.8 \\ 0.4 & 0.5 \\ 0.7 & 0.3 \end{bmatrix}$$

   (α΄) Βρείτε τις ιδιοτιμές και ιδιοδιανύσματα του πίνακα διασπορών - συνδιασπορών. Σχηματίστε τα σκορ κυρίων συνιστωσών στο $\mathbf{R}^3$.

   (β΄) Κάνετε το scree plot. Δείξτε γραφικά πως τα σκορ κυρίων συνιστωσών περιορίζονται στο $\mathbf{R}^2$ (κάνετε κατάλληλη περιστροφή στο τρισδιάστατο σχήμα).

   (γ΄) Σχηματίστε τα σκορ κυρίων συνιστωσών στο $\mathbf{R}^2$ και συγκρίνετε αυτό το σχήμα με το αρχικό των σημείων που δημιουργήσατε (στο $\mathbf{R}^2$).

2. Εφαρμόστε την ανάλυση PCA σε ένα σύνολο επιπέδων εκφράσεων $p = 384$ (μεταβλητές) γονιδίων σε $n = 17$ χρονικές στιγμές (παρατηρήσεις), που δίνονται στο αρχείο `yeast.dat`.

   (α΄) Εκτιμήστε τη διάσταση $d \leq p$ για τη μείωση διάστασης με ΠΆ.

   (β΄) Σχηματίστε τα σκορ κυρίων συνιστωσών στο $\mathbf{R}^2$ και $\mathbf{R}^3$.

3. Εφαρμόστε την ανάλυση PCA στα δεδομένα της Άσκησης 5.8 (στο αρχείο `physical.txt`).

   (α΄) Εκτιμήστε τη διάσταση $d \leq p$ για τη μείωση διάστασης με ΠΆ.

   (β΄) Σχηματίστε τα σκορ κυρίων συνιστωσών στο $\mathbf{R}^2$ και $\mathbf{R}^3$.

4. Φορτώστε τα σήματα ήχου από 1) τερέτισμα και 2) κρουστό από τα αρχεία chirp.mat και gong.mat που υπάρχουν ως παραδείγματα στο Matlab. Κρατήστε τις πρώτες 10000 παρατηρήσεις από κάθε σήμα.

   (α΄) Επιλέξτε ένα τυχαίο πίνακα μίξης A μεγέθους 2×2 και μετασχηματίστε τα αρχικά σήματα σε δύο αναμεμιγμένα σήματα. Εφαρμόστε την μέθοδο ICA με και χωρίς προλεύκανση και ελέγξετε αν ανακτώνται τα δύο αρχικά σήματα (τερέτισμα και κρουστό).

(β΄) Επαναλάβετε το ίδιο με παραπάνω αλλά για πίνακα μίξης A μεγέθους 2×3.

5. Δημιουργήστε τον πίνακα δεδομένων X μεγέθους $n \times p$ από $p = 5$ μεταβλητές που ακολουθούν εκθετική κατανομή με διαφορετικές μέσες τιμές η κάθε μια (ελεύθερη επιλογή). Δημιουργήστε το διάνυσμα απόκρισης $\mathbf{y}$ από τη σχέση $\mathbf{y} = X\beta + \epsilon$, όπου το διάνυσμα συντελεστών β έχει μόνο δύο μη-μηδενικά στοιχεία, π.χ. $\beta = [0, 2, 0, -3, 0]^{\mathrm{T}}$, και η τυχαία μεταβλητή θορύβου, που δίνει τις τιμές στο ϵ, ακολουθεί κανονική κατανομή με δεδομένη τυπική απόκλιση, π.χ. $\sigma_\epsilon = 5$.

   (α΄) Εκτιμήστε το μοντέλο πολλαπλής γραμμικής παλινδρόμησης με κάθε μια από τις μεθόδους OLS, PCR, PLS, RR και LASSO.

   (β΄) Για κάθε μέθοδο σχηματίστε το διάγραμμα διασποράς των παρατηρούμενων και εκτιμώμενων τιμών της εξαρτημένης μεταβλητής καθώς και το διάγραμμα των τυποποιημένων σφαλμάτων παλινδρόμησης.

   (γ΄) Συγκρίνετε τις εκτιμήσεις του β με κάθε μια μέθοδο.

6. Στην Άσκηση 5.8 εκτιμήσαμε το μοντέλο γραμμικής παλινδρόμησης με την OLS και τη βηματική παλινδρόμηση, όπου χρησιμοποιήθηκαν τα δεδομένα στο αρχείο `physical.txt`. Στα ίδια δεδομένα εκτιμήστε το μοντέλο γραμμικής παλινδρόμησης με τις μεθόδους PCR, PLS, RR και LASSO. Συγκρίνετε τα αποτελέσματα με αυτά των OLS και βηματικής παλινδρόμησης.