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`.

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 $\mathcal{X}^2$ 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 $\mathcal{X}^2$ κάλεσε τη συνάρτηση `vartest`. Για τον έλεγχο $\mathcal{X}^2$ 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 $\mathcal{X}^2$ 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 $\hat{\mathrm{se}}_B(\bar{x})$ του τυπικού σφάλματος του $\bar{x}$ από τα ίδια $B = 1000$ δείγματα bootstrap. Συγκρίνετε την εκτίμηση αυτή με αυτήν του τυπικού σφάλματος $\hat{\mathrm{se}}(\bar{x})$ 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 $\mathrm{H}_0 : \mu = \mu_0$. 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 α.

(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)

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

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 $\bar{\mathbf{z}}$ 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 $\bar{\mathbf{z}}$ και το ίδιο για το $\mathbf{y} = \{y_1, \ldots, y_n\}$

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

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 $\bar{\mathbf{z}}$ 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.

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 ($h_1 = 100$cm). Σε κάθε πείραμα δημιουργείστε τις 5 μετρήσεις του ύψους h_2 μετά την αναπήδηση της μπάλας από κανονική κατανομή με μέση τιμή $\mu_2 = 58$cm και τυπική απόκλιση $\sigma_2 = 2$cm. Υπολογίστε σε κάθε προσομοίωση το μέσο όρο και την τυπική απόκλιση του ύψους h_2 ($\bar{h}_2$, s_{h_2}), καθώς και του συντελεστή αποκατάστασης e ($\bar{e}$, s_e). Εξετάστε αν οι τιμές αυτές είναι συνεπείς με τις αναμενόμενες (πόσο κοντά αριθμητικά είναι) από τη σχέση $e = \sqrt{h_2/h_1}$ για σταθερό $h_1 = 100$cm (μ_{h_2}, σ_{h_2} για h_2 και μ_e, σ_e για e). Για να το κάνετε αυτό μπορείτε να σχηματίσετε ιστογράμματα για κάθε στατιστικό μαρκάροντας την τιμή της αντίστοιχης παραμέτρου.

   (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 έχουμε αυτήν την τιμή της αβεβαιότητας;

(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`.

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 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 $(1-\alpha)\%$ (εδώ $\alpha = 0.05$) διαστήματος εμπιστοσύνης για το β_0 από τα ποσοστιαία σημεία $\alpha/2\%$ και $(1-\alpha/2)\%$. Για αυτό βάλε τις $M = 1000$ τιμές σε αύξουσα σειρά και βρες τις τιμές για τάξη $M\alpha/2\%$ και $M(1-\alpha/2)\%$. 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.