ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ»

ΥΠΟΛΟΓΙΣΤΙΚΕΣ ΣΤΑΤΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ

ΕΡΓΑΣΙΑ ΠΡΩΤΗ

ΜΠΟΥΡΑΝΤΑΝΗΣ ΑΝΔΡΕΑΣ

mpouras.math@gmail.com

ΑΜ:ΜΕΣ22003

Άσκηση 1 *(40 Βαθμοί)*

Μία ερευνητική ομάδα θέλει να μελετήσει τις σχέσεις 5 συστατικών του αίματος και για το λόγο αυτό μελέτησε 100 άτομα και κατέγραψε τις ποσότητες από κάθε συστατικό σε 20 *ml* αίματος (*DoctorData.txt*).

(α΄) (10 Βαθμοί) Σύμφωνα με τον παγκόσμιο οργανισμό υγείας, η τιμή του *RBCs* πρέπει να είναι 5. Χρησιμοποιώντας $B = 1000$ δείγματα *Bootstrap* να κάνετε τον έλεγχο της υπόθεσης

$$H_0 : \ \mu = 5 \qquad H_1 : \ \mu < 5$$

Να υπολογιστεί η τιμή p του ελέγχου.

Με τη χρήση του στατιστικού πακέτου βρήκαμε ότι η τιμή p-value είναι 28.9% επομένως δεν απορρίπτουμε τη μηδενική υπόθεση για τα σύνηθες επίπεδα σημαντικότητας.

(β΄) (10 Βαθμοί) Ο αναλυτής θεωρεί ότι τα δεδομένα ακολουθούν την κατανομή[1] *LogNormal*(μ, σ^2) όπου οι παράμετροι (μ, σ^2) εκτιμώνται με τη μέθοδο μεγίστης πιθανοφάνειας[2]. Χρησιμοποιώντας παραμετρικό *Bootstrap* με $B = 1000$ δείγματα και θέτοντας *set.seed*(1), να βρεθεί η αναμενόμενη μέση τιμή και η τυπική απόκλιση για την μεταβλητή *RBCs*.

Με τη μέθοδο της πιθανοφάνειας εκτιμούμε τις παραμέτρους μ και σίγμα. Αφού τις βρούμε υπολογίζουμε τη μέση τιμή της RBC και καταλήγουμε πως ο μέση τιμή και η τυπική απόκλιση για τη μεταβλητή RBC είναι 4.9458 και 1.04585 αντίστοιχα.

(γ΄) (20 Βαθμοί) Ένας ολικός δείκτης ελέγχου συσχέτισης των 5 συστατικών του αίματος είναι η μέση τιμή των συντελεστών συσχέτισης όλων των πιθανών συνδυασμών των 5 συστατικών (εδώ φυσικά πρόκειται για 10 διαφορετικές συσχετίσεις). Να κατασκευαστούν και τα 4 διαφορετικά 95% διαστήματα εμπιστοσύνης ώστε να εκτιμηθούν τα όρια του ολικού δείκτη ελέγχου χρησιμοποιώντας $B = 1000$ δείγματα *Bootstrap* και θέτοντας *set.seed*(1).

```
> r_total_CIs(data)
                       Lower     Upper
Boostrap           0.2941911 0.4689015
Bootstrap-t        0.2897384 0.4733680
Quantile bootstrap 0.2827561 0.4574665
Bca bootstrap      0.2897469 0.4496969
```

Άσκηση 2 (*10 Βαθμοί*)

Δίνονται τα ύψη 50 φοιτητών του πανεπιστημίου Πειραιώς (*HeightsData.txt*). Χρησιμοποιώντας τον *Jackknife* εκτιμητή, να εκτιμήσετε το εύρος του ύψους των φοιτητών καθώς και την πιθανότητα το ύψος να είναι μεγαλύτερο από 165 εκατοστά. Να εκτιμηθεί επίσης η τυπική απόκλιση και να κατασκευάσετε 98% διαστήματα εμπιστοσύνης.

Ο jack knife εκτιμητής του εύρους του ύψους των φοιτητών είναι:

```
> jknife(Height,euros)
$unbiased.est
[1] 14.3602

$biased.est
[1] 12.3702
```

Ο jack knife εκτιμητής της πιθανότητας το ύψος να είναι μεγαλύτερος από 165 εκατοστά είναι:

```
> jknife(data,prob_165,fxn_dim=1)
$unbiased.est
[1] 0.84

$biased.est
[1] 0.84
```

Τα 98% διαστήματα εμπιστοσύνης για την τυπική απόκλιση του Heights είναι:

```
> bootstrap_CIs(Height,fxn=sd,fxn_dim = 1,alfa = 0.02)
$est
[1] 2.742843

$se
[1] 0.2475712

$CIs
                       Lower    Upper
Bootstrap           2.270866 3.350419
Bootstrap-t         2.256688 3.529801
Quantile bootstrap  2.205300 3.284852
Bca bootstrap       2.330968 3.263379
```

Τα 98% διαστήματα εμπιστοσύνης για τον jack-knife εκτιμητή της απόκλισης είναι:

```
> bootstrap_CIs(data=Height,fxn=jknife_2,fxn_dim=1,alfa=0.02)
$est
[1] 2.754391

$se
[1] 0.2475793

$CIs
                       Lower    Upper
Bootstrap           2.282619 3.360485
Bootstrap-t         2.270565 3.550313
Quantile bootstrap  2.219158 3.297024
Bca bootstrap       2.349760 3.272786
```

Άσκηση 3 (*10 Βαθμοί*)

Δίνονται τα ακόλουθα δεδομένα που αφορούν τον αριθμό των νικών που έχουν κάνει 50 ομάδες σε ένα πρωτάθλημα:

12	*19*	*8*	*15*	*21*	*16*	*10*	*8*	*19*	*16*
24	*31*	*20*	*22*	*25*	*29*	*31*	*29*	*29*	*37*
16	*15*	*12*	*20*	*18*	*14*	*11*	*10*	*13*	*10*
28	*38*	*21*	*24*	*27*	*24*	*23*	*23*	*21*	*29*
17	*19*	*14*	*14*	*18*	*14*	*24*	*23*	*23*	*23*

Χρησιμοποιώντας *Bootstrap* με $B = 1000$ δείγματα και θέτοντας *set.seed*(1), να κατασκευαστούν και τα 4 διαφορετικά 95% διαστήματα εμπιστοσύνης για την διαφορά του μέσου αριθμού νικών μεταξύ των 10 πρώτων ομάδων και των 10 τελευταίων ομάδων (σε πλήθος νικών).

```
> bootstrap_CIs(data,fxn=diff_win,alfa=0.05,fxn_dim = 1)
$est
[1] 19.413

$se
[1] 1.798946

$CIs
                       Lower    Upper
Bootstrap           17.10000 24.20000
Bootstrap-t         17.18818 25.26342
Quantile bootstrap  15.80000 22.90000
Bca bootstrap       17.20000 23.70000
```

Άσκηση 4 (*40 Βαθμοί*)

Θεωρήστε το σύνολο δεδομένων *petrol* που βρίσκεται στη βιβλιοθήκη *MASS* της *R*. Για την εν λόγω άσκηση θα χρησιμοποιηθούν τα δεδομένα των στηλών 2 έως 6. Έστω Σ είναι ο πίνακας διακύμανσης-συνδιακύμανσης των δεδομένων και οι ιδιοτιμές του $\lambda_1 > \ldots > \lambda_5 > 0$. Τότε,

$$\theta = \frac{\lambda_1}{\sum_{i=1}^{5} \lambda_i}$$

είναι η προς εκτίμηση ποσότητα. Έστω επίσης $\hat{\lambda}_1 > \ldots > \hat{\lambda}_5$ οι ιδιοτιμές από το δειγματικό πίνακα διακύμανσης-συνδιακύμανσης $\hat{\Sigma}$.

(α΄) (10 Βαθμοί) Να χρησιμοποιήσετε τον *jackknife* εκτιμητή ώστε να εκτιμήσετε την μεροληψία και το τυπικό σφάλμα του εκτιμητή του θ που είναι ο

$$\hat{\theta} = \frac{\hat{\lambda}_1}{\sum_{i=1}^{5} \hat{\lambda}_i}.$$

```
> jknife(data,lambda1,fxn_dim=2)
$unbiased.est
[1] 0.81042

$biased.est
[1] 0.8129821

$se
[1] 0.04959806

$ci
[1] 0.7092641 0.9115759

$pseudos
 [1] 1.3799922 0.7691053 0.3885014 0.1445471 1.2694760 0.9167903 0.6531896 1.2899693 0.9294329 0.7109595 1.1208033
[12] 0.8340982 0.7149975 0.7084384 0.8330999 0.7885426 0.8025999 0.9065804 0.8049816 0.8594911 0.6585532 0.6564455
[23] 0.7944668 1.0512078 0.5754918 0.7894160 1.2553506 0.6536093 1.0838234 0.3043166 0.4276258 0.8575372

$partials
 [1] 0.7946088 0.8143148 0.8265924 0.8344619 0.7981738 0.8095508 0.8180540 0.7975128 0.8091430 0.8161905 0.8029697
[12] 0.8122183 0.8160602 0.8162718 0.8122505 0.8136878 0.8132343 0.8098801 0.8131575 0.8113991 0.8178810 0.8179490
[23] 0.8134967 0.8052147 0.8205604 0.8136596 0.7986295 0.8180405 0.8041626 0.8293080 0.8253303 0.8114622

$bias
[1] 0.002482014
```

Από το στατιστικό πακέτο βλέπουμε πως ο jack-knife εκτιμητής για το τυπικό σφάλμα και τη μεροληψία είναι 0.1382 και 0.233 αντίστοιχα.

(β΄) (20 Βαθμοί) Χρησιμοποιήστε $B = 1000$ δείγματα *Bootstrap* (με *set.seed*(1)) ώστε να δημιουργήσετε τα 4 βασικά 95% διαστήματα εμπιστοσύνης για την ποσότητα θ.

```
$CIs
                       Lower     Upper
Bootstrap           0.7290392 0.9091493
Bootstrap-t         0.6807985 0.8956084
Quantile bootstrap  0.7166547 0.8967649
Bca bootstrap       0.7153203 0.8841876
```

(γ') (10 Βαθμοί) Χρησιμοποιώντας ως εξαρτημένη μεταβλητή την Y, και ανεξάρτητες τις SG, VP, $V10$, EP, να βρείτε το μοντέλο που έχει τη μεγαλύτερη ικανότητα πρόβλεψης.

```
+ }
1 : 111.3774 118.0935 118.0935 118.6789
1 2 : 104.6194 120.1337 120.1337 121.6575
1 3 : 94.94811 108.2786 112.4015 109.3685
1 4 : 100.3243 112.502 118.2527 113.1741
1 5 : 54.99042 61.47409 61.69866 61.80275
1 2 3 : 94.93658 116.1747 116.1747 118.2877
1 2 4 : 100.1793 121.4799 121.4799 123.3208
1 2 5 : 26.93584 31.94398 32.03198 32.04375
1 3 4 : 94.26835 112.6539 121.0251 113.7291
1 3 5 : 11.55835 13.96677 13.98381 14.04549
1 4 5 : 5.331611 6.432492 6.43583 6.510818
1 2 3 4 : 94.02363 122.2871 122.2871 124.1794
1 2 3 5 : 8.296235 10.56698 10.58164 10.48264
1 2 4 5 : 4.562535 5.677598 5.680997 5.6723
1 3 4 5 : 5.01936 6.410202 6.415518 6.402391
1 2 3 4 5 : 4.212624 5.531695 5.536439 5.462749
>
```

Βλέπουμε πως το μοντέλο με τη μεγαλύτερη προβλεπτική αξία είναι το πλήρες μοντέλο καθώς επιτυγχάνει ελάχιστες τιμές για όλα τα κριτήρια. Ωστόσο το μοντέλο με επεξηγηματικές τις V10 και EP δείχνει να έχει κοντινή(αν και μικρότερη) προβλεπτική αξία από το πλήρες επομένως αν θέλουμε ένα πιο οικονομικό μοντέλο μπορούμε να επιλέξουμε αυτό με μόνο 2 επεξηγηματικές.

(δ') (10 Bonus) Στο ερώτημα αυτό θα χρειαστούμε και τα δεδομένα της πρώτης στήλης (*No*). Χρησιμοποιήστε *Bootstrap* με $B = 1000$ δείγματα ώστε να ελέγξετε την υπόθεση η μέση τιμή της μεταβλητής Y για την κατηγορία D είναι ίση με τα $2/3$ της μέσης τιμής της μεταβλητής Y για την κατηγορία A.

Αρχικά θα δούμε αν τα δείγματα προέρχονται από ομοσκεδαστικούς πληθυσμούς με το Welch Test.

```
> welch_test(data)
[1] 0.8729599
>
```

Επομένως δε μπορούμε να απορρίψουμε τη μηδενική υπόθεση πως οι πληθυσμοί είναι ομοσκεδαστικοί.

Θα ελέγξουμε αν $\mu_\chi-\mu_y=0$ όπου y=3/2 της αρχικής μεταβλητής.

```
> mu_test(data)
$est
[1] -0.05658815

$p_value
[1] 0.8948949

$CIs
[1] -3.470712  3.438415
```

Βλέπουμε πως από το p-value καθώς και από το διάστημα εμπιστοσύνης πως δε μπορούμε να απορρίψουμε τη μηδενική υπόθεση πως η μέση τιμή της Y για την κατηγορία D είναι ίση με τα 2/3 της μέσης τιμής της μεταβλητής Y την κατηγορία A.

Προσέγγιση με λόγο μέσων

$H_0: \frac{\mu_A}{\mu_D} = 3/2$

```
> bootstrap_CIs(data,fxn=mean_ratio,fxn_dim = 2)
$est
[1] 1.423176

$se
[1] 0.06312473

$CIs
                      Lower    Upper
Bootstrap          1.172061 1.463736
Bootstrap-t        1.234792 2.030229
Quantile bootstrap 1.348993 1.640669
Bca bootstrap      1.348993 1.640669
```

Λόγω μικρού αριθμού δείγματος(n=4) στο εσωτερικό bootstrap-t για εκτίμηση της τυπικής απόκλισης θα υπάρχουν δείγματα με 4 φορές την ίδια παρατήρηση επομένως θα έχουμε μηδενική διασπορά για αυτό και το 2^o δ.ε. διαφέρει σημαντικά από τα άλλα.

Στο πρώτο Δ.Ε. το 3/2 οριακά δεν ανήκει ωστόσο στο 3^o και 4^o βλέπουμε πως δεν απορρίπτουμε. Επομένως και μέσω αυτού του ελέγχου δεν απορρίπτουμε τον ισχυρισμό της υπόθεσης.