Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Πληροφορικής
Μάθημα: Στατιστική στην Πληροφορική
Ακαδημαϊκό έτος: 2022–23
Υπεύθυνη φοιτήτρια: Μαρία Κονταράτου (3200078)

$2^{η}$ Σειρά Ασκήσεων

<u>Άσκηση 1:</u>

(a) Τα δεδομένα φαίνεται να είναι κατάλληλα, καθώς μιλάμε για απλή τυχαία δειγματοληψία μιας συγκεκριμένης ημέρας (τυχαία επιλογή πληθυσμού) και το μέγεθος του δείγματος είναι επαρκές (20>15). Ακόμα και αν λόγω της ατυπικής τιμής 284 μπορεί να οδηγηθούμε σε μη κανονικά κατανεμημένο πληθυσμό, μας αρκεί το μέγεθος του δείγματος άρα τα δεδομένα χαρακτηρίζονται κατάλληλα.

(b) Για c = 0.95, s = 55.524674 , $\bar{x}$ = 77.4, t = 2.093024
Άρα, διάστημα εμπιστοσύνης: [77.4 ± 2.093 * 55.52/√20] = [51.41365, 103.3863]

<u>Άσκηση 2:</u>

(a) Λάθος, τυπική απόκλιση είναι σ/√n άρα 12/√20
(b) Λάθος, η H_0 δεν εξαρτάται από το δείγμα, σε αντίθεση με το $\bar{x}$
(c) Λάθος, αν απορρίψουμε την H_0 πρέπει το p-value να είναι μικρότερο από τον βαθμό σημαντικότητας. Όμως, ο δειγματικός μέσος είναι 45 άρα δεν μπορεί να απορριφθεί αφού 45<54.
(d) Λάθος, για να απορριφθεί η μηδενική υπόθεση πρέπει το p value να είναι ελάχιστο

<u>Άσκηση 3:</u>

z = 1.34 άρα
(a) H_a: $\mu>\mu_0$ άρα p value = 1- Φ(Z) = 1-0.9099= 0.0901
(b) H_a: $\mu<\mu_0$ άρα p value = Φ(Z) = 0.9099
(c) H_a: $\mu\neq\mu_0$ άρα p value = 2 Φ(-|Z|) =2*(-|1.34|)= 2*0.0901= 0.1802

<u>Άσκηση 4:</u>

(a) Όχι δεν προέρχεται. Αν ανήκε θα έπρεπε να μην απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 5%, πράγμα που στην παρούσα περίπτωση συμβαίνει αφού το p value είναι 0.04
(b) Όχι. Πάλι απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 10% αφού p value είναι 0.04

Άσκηση 5:

(a) Παρατηρούμε στα δεδομένα μας μια ατυπική τιμή (Α/Α 14) όπου τα κιλά της γυναίκας είναι αδύνατο να είναι 6. Όμως, επειδή το μέγεθος του δείγματος είναι 24 (>15) μπορούμε να εφαρμόσουμε κανονικά τη μέθοδο που βασίζεται στην κατανομή t

Για c = 0.95, s = 9.9781146 , $\bar{x}$ = 73.79167, t = 2.068658
Άρα, διάστημα εμπιστοσύνης: [73.79167± 2.068658* 9.978146 /√24] = [69.57826, 78.00507]

```
> mw <- mean(ΒΑΡΟΣ)
> mw
[1] 73.79167
> sdw <- sd (ΒΑΡΟΣ)
> t1 <- abs(qt(0.025,df=23))
> error1 <- t1*sdw/sqrt(24)
> uplimit1<-mw+error1
> lowlimit1 <- mw-error1
> sdw
[1] 9.978146
> error1
[1] 4.213402
> t1
[1] 2.068658
> uplimit1
[1] 78.00507
> lowlimit1
[1] 69.57826
```

(b) Σε κάθε περίπτωση έχουμε κοινό c = 0.8 όμως για τις γυναίκες και για τους άνδρες έχουμε διαφορετικά:

N_A=13, s_A = 7.598077 , $\overline{x}_A$ = 78.69231

N_Γ=11, s_Γ = 9.570789 , $\overline{x}_\Gamma$ = 68

$\overline{x}_A$ - $\overline{x}_\Gamma$ = 10.69321 , t = 1.372184

Άρα, διάστημα εμπιστοσύνης: [5.789155, 15.59546]

```
> weightW <- ΒΑΡΟΣ[ΦΥΛΟ=='Γ']
> weightM <- ΒΑΡΟΣ[ΦΥΛΟ=='Α']
> meanWeightW <- mean(weightW)
> meanWeightW
[1] 68
> meanWeightM <- mean(weightM)
> meanWeightM
[1] 78.69231
> dif <- meanWeightM - meanWeightW
> dif
[1] 10.69231
> sdW <- sd(weightW)
> sdW
[1] 9.570789
> sdM <- sd(weightM)
> sdM
[1] 7.598077
> t2 <- abs(qt(0.1,df=10))
> t2
[1] 1.372184
> error2 <- t2*(sqrt(((sdM^2)/13)+ ((sdW^2)/11)))
> error2
[1] 4.903152
> uplimit2 <- dif + error2
> lowlimit2 <- dif - error2
> uplimit2
[1] 15.59546
> lowlimit2
[1] 5.789155
>
```

(c) Θα ερευνήσουμε το μέσο βάρος στις κατηγορίες των καπνιστών και μη. Όπως παρατηρούμε , το βάρος του εκάστοτε καπνιστή είναι περίπου μεγαλύτερο κατά 5 κιλά συγκριτικά με τον μη-καπνιστή.
Έστω η μηδενική υπόθεση H_0: $M_{καπνιστών}= M_{μη\text{-}καπνιστών}$
Θα βρούμε σε αυτή τη φάση την τιμή του στατιστικού ελέγχου z για να βρούμε το p-value = 0.2394573

```
> smoking <- ΒΑΡΟΣ[ΚΑΠΝΙΣΤΗΣ =="ΝΑΙ"]
> nonsmoking <- ΒΑΡΟΣ[ΚΑΠΝΙΣΤΗΣ =="ΌΧΙ"]
> meanS <- mean(smoking)
> meanS
[1] 76.8
> meanNS <- mean(nonsmoking)
> meanNS
[1] 71.64286
> sdS <- sd(smoking)
> sdNS <- sd(nonsmoking)
> sdS
[1] 9.975526
> sdNS
[1] 9.76341
> dif2 <- meanS - meanNS
> dif
[1] 10.69231
> dif2
[1] 5.157143
> a <- (sdS^2)/10
> b <- (sdns^2)/14
Error: object 'sdns' not found
> b <- (sdNS^2)/14
> z <- dif2/sqrt(a+b)
> z
[1] 1.259715
> 2*pt(df=9,-abs(z))
[1] 0.2394573
>
```

Άσκηση 6:

(a) Τα δεδομένα φαίνεται να είναι κατάλληλα, καθώς μιλάμε για απλή τυχαία δειγματοληψία (τυχαία επιλογή πληθυσμού) και το μέγεθος του δείγματος είναι επαρκές (20>15). Δεν παρουσιάζονται ατυπικές τιμές και τα δεδομένα είναι συμμετρικά.

(b) M_{V1} = 5.5
S_{V1} = 0.6008766

```
> data<-read.table("data.txt")
> attach(data)
> data
    V1
1  5.7
2  4.6
3  6.4
4  6.3
5  6.9
6  5.2
7  4.9
8  5.4
9  4.9
10 5.6
11 5.4
12 5.3
13 4.9
14 5.1
15 5.0
16 6.0
17 6.3
18 5.4
19 5.3
20 5.4
> m <- mean(V1)
> m
[1] 5.5
> sd1 <- sd(V1)
> sd1
[1] 0.6008766
> |
```

(c) Για c = 0.95, s = 0.6008766 , $\bar{x}$ = 5.5, t = 2.093024
Άρα, διάστημα εμπιστοσύνης: [5.2187, 5.7812]

```
> t <- abs(qt(0.025,df=19))
> t
[1] 2.093024
> error <- t*(sd1/sqrt(20))
> error
[1] 0.2812189
> uplimit <- m + error
> uplimit
[1] 5.781219
> lowlimit <- m - error
> lowlimit
[1] 5.218781
>
```

Άσκηση 7:
Δεν μπορούμε στην παρούσα περίπτωση να εφαρμόσουμε μεθοδολογία για ανεξάρτητο δείγμα αφού υπάρχει εξάρτηση δειγμάτων καθώς η μεγάλη εκτίμηση ζημιάς επηρεάζει το συνεργείο. Συνεπώς δημιουργούμε πίνακα διαφοράς συνεργείου με εμπειρογνώμονα:

ΑΥΤΟΚΙΝΗΤΟ	ΔΙΑΦΟΡΑ ΕΚΤΙΜΗΣΗΣ
1	100
2	50
3	-50
4	0
5	-50
6	200
7	250
8	200
9	150
10	300

Έχουμε ότι s = 124,8332 , $\overline{x}$ = 115, t = 2.093024
Έστω μ: μέση τιμή διαφοράς τότε
H_0: μ=0
$H_α$: μ>0

Άρα p value = 1- Φ(t) = 1-0.9981 = 0.0018
Το p value είναι ελάχιστο για τα συνηθισμένα επίπεδα σημαντικότητα άρα απορρίπτουμε την αρχική υπόθεση άρα το συνεργείο υπερεκτιμά τις ζημιές.

```
> m <- mean(V1)
> m
[1] 115
> sd <- sd(V1)
> sd
[1] 124.8332
> t <- m/(sd/sqrt(10))
> t
[1] 2.913182
>
```

<u>Άσκηση 8:</u>
Τα δεδομένα του ερωτηματολογίου φαίνεται να είναι κατάλληλα, καθώς μιλάμε για απλή τυχαία δειγματοληψία (τυχαία επιλογή πληθυσμού) και το μέγεθος του δείγματος είναι επαρκές (118>15). Γενικά δεν φαίνεται να υπάρχουν ατυπικές τιμές.

(a) Σε κάθε περίπτωση έχουμε κοινό c = 0.95 όμως για τις γυναίκες και για τους άνδρες έχουμε διαφορετικά:
$N_Α$=80, $s_Α$ = 0.06976634 , $\overline{x}_Α$ = 1.799
$N_Γ$=37, $s_Γ$ = 0.068116 , $\overline{x}_Γ$ = 1.671351
$\overline{x}_Α$ - $\overline{x}_Γ$ = 0.1276486 , t = 1.305514

Άρα, διάστημα εμπιστοσύνης: [5.789155, 15.59546]

```
> heightF <- height[gender =='F']
> heightM <- height[gender =='M']
> meanHeightF <- mean(heightF)
> meanHeightM <- mean(heightM)
> meanHeightM
[1] 1.799
> meanHeightF
[1] 1.671351
> dif <- meanHeightM - meanHeightF
> dif
[1] 0.1276486
> sdM <- sd(heightM)
> sdM
[1] 0.06976634
> sdF <- sd(heightF)
> sdF
[1] 0.068116
> t <- abs(qt(0.1,df=36))
> error <- t* (sqrt(((sdM^2)/80) + ((sdF^2)/37)))
> error
[1] 0.01781639
> uplimit <- dif + error
> lowlimit <- dif - error
> uplimit
[1] 0.145465
```

(b) Έστω μ_1 και μ_2 ο μέσος όρος βαθμού πιθανοτήτων στα αγόρια και τα κορίτσια

H_0: $\mu_1 = \mu_2$

H_α: $\mu_1 > \mu_2$

Θα βρούμε σε αυτή τη φάση την τιμή του στατιστικού ελέγχου z για να βρούμε το p-value = 0.3650116

Το p-value > 0.05 άρα δεν απορρίπτεται η μηδενική υπόθεση άρα άντρες και και γυναίκες επιτυγχάνουν ίδιο βαθμό κατά μέσο όρο στις πιθανότητες

```
> probF <- prob[gender=="F"]
> probF <- probF[!is.na(probF)]
> probF
 [1]  6.0  7.0  5.0  4.0  7.5  9.5 10.0  6.0  5.5  7.0  9.5  5.5 10.0  6.0  5.5  5.0  9.5  5.0  7.5  5.5  2.0  9.5  5.0  7.5  9.0
[26]  7.0  5.5  7.0  6.5  5.5  6.0 10.0  6.5  0.0
> meanF <- mean(probF)
> meanF
[1] 6.573529
> probM <- prob[gender=="M"]
> probM <- probM[!is.na(probM)]
> probM
 [1]  9.0  9.5  6.5  8.5  5.0  0.0 10.0  6.5  3.0  5.5  7.5  5.5  3.0  6.0 10.0 10.0  0.0  9.0 10.0  5.0  5.0  5.5  6.5  7.0  7.5
[26]  5.5  6.5  6.5 10.0  9.5 10.0 10.0  6.0 10.0  7.5  7.0  6.0  3.0  5.0  5.0  2.0  2.0  5.0 10.0  5.0  7.5  8.0  5.5  5.0  5.0
[51]  5.0  0.0  5.5  0.0  3.0  5.0  9.0  6.5  8.0  7.0  6.0  8.5  3.0  5.0  8.0  5.5  0.0  5.5  7.0  8.0  5.5  8.0  6.0  6.0  9.0
[76]  5.0  3.0
> sdF <- sd(probF)
> sdF
[1] 2.253389
> sdM <- sd(probM)
> sdM
[1] 2.644702
> meanM <- mean(probM)
> meanM
[1] 6.123377
> dif <- meanF - meanM
> dif
[1] 0.4501528
> a <- (sdF^2)/34
> b <- (sdM^2)/77
> z <- dif/sqrt(a+b)
> z
[1] 0.9185205
> 2*pt(df=33,-abs(z))
[1] 0.3650116
> 
```

(c) Τα δείγματα βαθμών δεν είναι ανεξάρτητα άρα πρέπει να εξετάσουμε τη διαφορά των μέσων $\mu = \mu_{\pi}-\mu_{\mu}$ όπου είναι μέσοι βαθμοί για πιθανότητες και μαθηματικά αντίστοιχα.

Έχουμε ότι s = 2.060823 , $\overline{x}$ = -0.2924528, t = -1.46106
Έστω μ_1 και μ_2 ο μέσος όρος βαθμού πιθανοτήτων στα αγόρια και τα κορίτσια
H_0: $\mu_1 = \mu_2$
H_α: $\mu_1 \neq \mu_2$

Υπολογίζουμε p-value : 0.146985 άρα δεν απορρίπτουμε την αρχική υπόθεση άρα ο μέσος όρος βαθμού πιθανοτήτων είναι περίπου ίδιος.

```
102  1.0
103  0.0
104  2.0
105 -1.5
106 -3.0
> m <- mean(V1)
> m
[1] -0.2924528
> sd <- sd(V1)
> sd
[1] 2.060823
> t <- m/(sd/sqrt(106))
> t
[1] -1.46106
> 2*pt(df=105,-abs(t))
[1] 0.146985
>
```