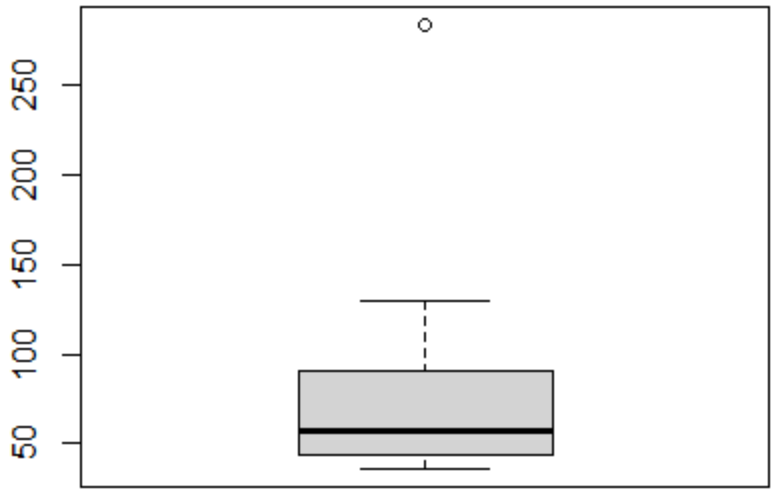
Στατιστική Στην Πληροφορική - 2η Σειρά Ασκήσεων
ΟΠΑ , Ακαδημαϊκό Έτος: 2020-2021

Ομάδα #0

- ΤΣΙΟΜΠΙΚΑΣ ΔΗΜΗΤΡΙΟΣ , 3180223
- ΠΑΝΑΓΙΩΤΟΥ ΠΑΝΑΓΙΩΤΗΣ , 3180139

1η Άσκηση

a. Τα δεδομένα που έχουμε είναι κατάλληλα για τις μεθόδους συμπερασματολογίας που ξέρουμε και συγκεκριμένα για το SRS (“*Simple Random Sample*”), διότι παίρνουμε ουσιαστικά ένα τυχαίο δείγμα ενός πληθυσμού (20 χρόνοι από 20 queries τα οποία ανήκουν στο γενικό πληθυσμό queries που έγινε εκείνη την ημέρα). Επίσης και το Boxplot δείχνει μια ασυμμετρία στα δεδομένα το οποίο είναι φυσιολογικό αφού n ≥ 20.



b. Οι υπολογισμοί θα γίνουν με την *R*:

- Αρχικά βρίσκουμε την **Δειγματική μέση τιμή**:

```
> data <- c(82,55,58,94,86,45,42,36,41,130,284,96,39,107,52,54,45,81,83,38)
> m <- mean(data)
> m
[1] 77.4
```

Η οποία είναι **m=77.4 milliseconds**.

- Μετά την **Δειγματική τυπική απόκλιση**:

```
> s <- sd(data)
> s
[1] 55.52467
```

Η οποία είναι **s=55.52467 milliseconds**.

- Και τέλος βρίσκουμε το **t*** :

```
> tstar <- qt(.975,19)
> tstar
[1] 2.093024
```

Το οποίο είναι **t*=2.093024 milliseconds**. (df = 19)

Χρησιμοποιούμε τον παρακάτω τύπο :

➢ **Διάστημα εμπιστοσύνης C% για τη μέση τιμή:** $\bar{x} \pm t_* \frac{s}{\sqrt{n}}$

Διότι η τυπική απόκλιση του γενικού πληθυσμού queries είναι άγνωστη.

```
> upperBound <- m + tstar*s/sqrt(20)
> upperBound
[1] 103.3863
```

```
> lowerBound <- m - tstar*s/sqrt(20)
> lowerBound
[1] 51.41365
```

Άρα **το 95% διάστημα εμπιστοσύνης** είναι **[51.41365,103.3863].**

2η Άσκηση

a. Είναι λάθος, διότι η δειγματική τυπική απόκλιση ως γνωστόν είναι:

$$s = \frac{\sigma}{\sqrt{n}}$$

Άρα εδώ:

$$s = \frac{12}{\sqrt{20}}$$

b. Στον έλεγχο σημαντικότητας η μηδενική υπόθεση και γενικώς οι υποθέσεις δεν μπορούν να χρησιμοποιήσουν τις μεταβλητές δειγματοληψίας. Πρέπει να χρησιμοποιούν αυτές του γενικού πληθυσμού. Δηλαδή εδώ θα έπρεπε να χρησιμοποιήσει την μ και όχι την xbar.

c. Την αποδεχόμαστε την μηδενική υπόθεση διότι όπως έχουμε δει και στις διαλέξεις αν η τιμή της μεταβλητής z του ελέγχου z βγει αρνητική θα έχουμε p-value = 99.9999% άρα εφόσον έχουμε ΠΟΛΥ μεγάλο p-value (μεγαλύτερο σίγουρα και από τον βαθμό σημαντικότητας α), αποδεχόμαστε την μηδενική υπόθεση.

d. Το 52% για pvalue είναι αρκετά μεγάλο ποσοστό, ωστόσο το αν θα απορρίψουμε την μηδενική υπόθεση εξαρτάται και από τον βαθμό σημαντικότητας (α) των δεδομένων.
Αν 52% < α τότε θα την απορρίπταμε την μηδενική υπόθεση και θα κοιτάγαμε την εναλλακτική, τώρα δεν μπορούμε να την απορρίψουμε.

3η Άσκηση

Χρησιμοποιήθηκε ο **πίνακας z** (*"z table"*) για αυτή την άσκηση.

a) **P-value** = P(z ≥ 1.34) = 1 – Φ(1.34) = 1 – 0.9099 = 0.0901
b) **P-value** = P(z ≤ 1.34) = Φ(1.34) = 0.9099
c) **P-value** = P(z <> 1.34) = 2*Φ(-1.34) = 2 * 0.0901 = 0.1802

4η Άσκηση

a. Παρατηρούμε ότι ο βαθμός σημαντικότητας α = 100-95 = 5% = 0.05. Άρα εφόσον pvalue < α απορρίπτεται η μηδενική υπόθεση, άρα δεν μπορούμε να είμαστε σίγουροι για την ύπαρξη της τιμής 30 στο διάστημα εμπιστοσύνης 95%.

b. Όμοια εφόσον α = 100-90 = 10 = 0.10 και pvalue < α δεν είμαστε σίγουροι αν βρίσκεται το 30 στο διάστημα εμπιστοσύνης 90%.

5η Άσκηση

a. Αρχικά, να σημειώσουμε ότι παρατηρήθηκε ένα δεδομένο το οποίο λογικά είναι λάθος, καθώς δεν υπάρχει ενήλικας 6kg οπότε θα το αγνοήσουμε.

```
> weight <- c(80,81,75,83,71,73,65,67,54,77,55,83,91,92,86,73,82,
69,73,70,59,68,72,72)
> n <- length(weight)
> t <- -qt(0.025,df = n - 1)
> mean(weight) + c(-1,1) * t * sd(weight)/sqrt(n)
[1] 69.57826 78.00507
```

Χρησιμοποιώντας λοιπόν τις εντολές που μάθαμε στο εργαστήριο παρατηρούμε ότι **το 95% διάστημα εμπιστοσύνης** είναι το **[69.57826,78.00507].**

Μεταβλητές που βρέθηκαν :

- **n** = 24
- **t*** = 2.068658
- **xbar** (*Δειγματική μέση τιμή*) = 73.79167
- **s** (*Δειγματική τυπική απόκλιση*) = 9.978146
- **df** = 23

b. Χωρίζουμε τα δεδομένα αντρών γυναικών σε πίνακες **mdata, fdata** αντίστοιχα.

```
> fdata <- data[sex == "F",]
```

```
> mdata <- data[sex == "M",]
```

Βρίσκουμε τις εξής τιμές (**m = male , f = female**):

- **nm** = 13
- **nf** = 11
- **xbarm** = 78.69231 kg
- **xbarf** = 68 kg
- **sm** = 7.598077 kg
- **sf** = 9.570789 kg

Διαχωρίσαμε τα βάρη γυναικών και ανδρών σε **vectors** :

```
> fweight <- c(71,65,67,54,55,83,73,82,69,70,59)
```

```
> mweight <- c(80,81,75,83,73,77,91,92,86,73,68,72,72)
```

Και εκτελέσαμε την εντολή *t.test* για 80% διάστημα εμπιστοσύνης :

```
> t.test(mweight,fweight,conf.level = 0.80)

        Welch Two Sample t-test

data:  mweight and fweight
t = 2.9923, df = 19.005, p-value = 0.007486
alternative hypothesis: true difference in means is not equal to
 0
80 percent confidence interval:
  5.948055 15.436561
sample estimates:
mean of x mean of y
 78.69231  68.00000
```

Άρα **το 80% διάστημα εμπιστοσύνης** είναι: **[5.948055,15.436561].**

c. Θα εκτελέσουμε δίπλευρο έλεγχο σημαντικότητας με τις εξής υποθέσεις :

- **μΝαι** = καπνιστές
- **μΟχι** = μη-καπνιστές

και θέλουμε να δούμε αν **μΝαι = μΟχι** άρα μηδενική υπόθεση

- **Η0**: μΝαι = μΟχι
- **Ηα**: μΝαι <> μΟχι

Και θα χρησιμοποιήσουμε τα κιλά των κατηγοριών για να βρούμε αν εν τέλει σχετίζεται το κάπνισμα με το βάρος.

Αρχικά διαχωρίζουμε τα δεδομένα :

```
> smokers <- data[smoker == "Y",]
```

```
> nonSmokers <- data[smoker == "N",]
```

Παίρνουμε τα βάρη της κάθε κατηγορίας σε **vectors** για ευκολία στις εντολές :

```
> smokeWeight <- c(80,83,71,73,65,77,92,86,82,59)
```

```
> nonSmokeWeight <- c(81,75,67,54,55,83,91,73,69,73,70,69,72,72)
```

Και θα χρησιμοποιήσουμε τον εξής τύπο για να βρούμε το ζητούμενο :

$$\text{Στατιστικό ελέγχου } t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- **XbarSmoke** = 76.8 kg
- **xbarNonSmoke** = 71.71429
- **sSmoke** = 9.975526
- **sNonSmoke** = 9.738335
- **nSmoke** = 10
- **nNonSmoke** = 14
- **df** = min{nSmoke - 1,nNonSmoke - 1} = min{9,13} = 9

```
> xbarSmoke <- mean(smokeWeight)
> xbarNonSmoke <- mean(nonSmokeWeight)
> sSmoke <- sd(smokeWeight)
> sNonSmoke <- sd(nonSmokeWeight)
> nSmoke <- length(smokeWeight)
> nNonSmoke <- length(nonSmokeWeight)
> t <- (xbarSmoke - xbarNonSmoke) / sqrt((sSmoke ^ 2/nSmoke) + (sNo
nSmoke ^ 2/nNonSmoke))
> t
[1] 1.243564
```

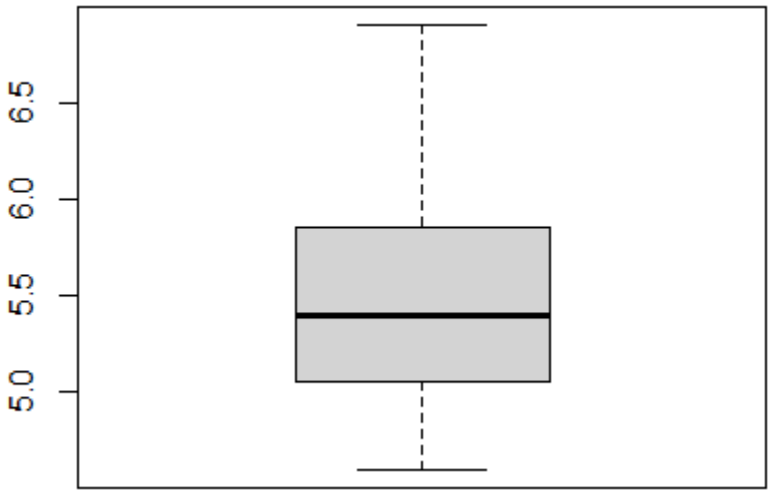
- **t** = 1.243564
- **Pvalue** = 2Φ(-1.243564) = 2 * 0.1075 = 0.215

Βλέπουμε πως ότι τιμή και να βάλουμε στο α (1%,5% ή 10%) η pvalue θα είναι πάντα μεγαλύτερη από το α οπότε δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση. Αυτό μας οδηγεί στο συμπέρασμα ότι το κάπνισμα **μάλλον** δεν σχετίζεται με τα κιλά.

6η Άσκηση

a. Παρατηρούμε πολύ μικρή ασυμμετρία στα δεδομένα μας μέσω του boxplot καθώς n ≥ 20 και καθόλου ατύπικες τιμές. Επίσης, εφόσον έχουμε πάρει ένα τυχαίο δείγμα από έναν γενικό πληθυσμό και το boxplot έχει καλή ένδειξη μπορούμε να εφαρμόσουμε μεθόδους συμπερασματολογίας (SRS).

Το **boxplot** είναι:



b.

```
> fuel <- c(5.7,4.6,6.4,6.3,6.9,5.2,4.9,5.4,4.9,5.6,5.4,5.3,4.9,5.
1,5.0,6.0,6.3,5.4,5.3,5.4)
```

```
> mean(fuel)
[1] 5.5
> sd(fuel)
[1] 0.6008766
```

c.

```
> t <- -qt(0.025,df = 19)
> mean(fuel) + c(-1,1) * t * sd(fuel) / sqrt(20)
[1] 5.218781 5.781219
```

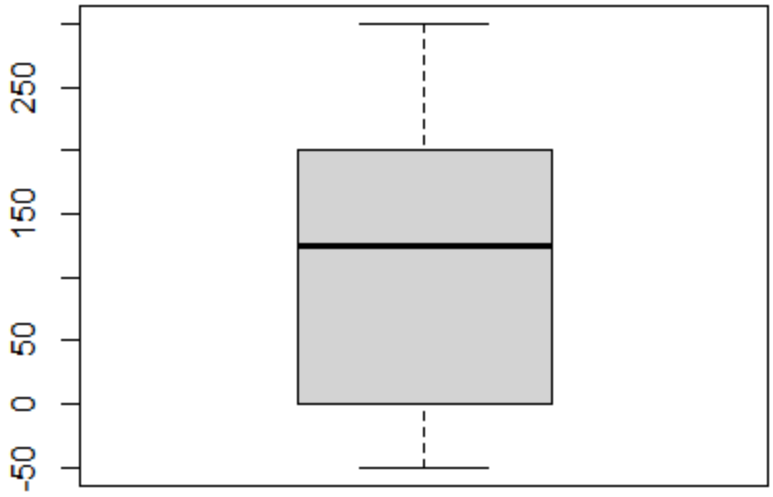
Άρα το διάστημα εμπιστοσύνης 95% είναι **[5.218781 , 5.781219].**

7η Άσκηση

Αρχικά , εφόσον τα δεδομένα μας ΔΕΝ είναι ανεξάρτητα πρέπει να φτιάξουμε ένα **vector** με τις διαφορές τους ώστε να χρησιμοποιήσουμε έλεγχο σημαντικότητας.

```
> car <- c(100,50,-50,0,-50,200,250,200,150,300)
```

Ελέγχουμε την καταλληλόλητα των δεδομένων με boxplot ώστε να εφαρμόσουμε μεθόδους συμπερασματολογίας :



Παρατηρούμε ότι δεν υπάρχει ασυμμετρία και τα δεδομένα ακολουθούν κανονική κατανομή (βάσει θεωρίας κιόλας καθώς έχουμε n < 15) άρα μπορούμε να εφαρμόσουμε μεθόδους συμπερασματολογίας.
Θα χρησιμοποιήσουμε τις εξής υποθέσεις :

- **H0**: μ = 0 (το συνεργείο ΔΕΝ υπερεκτιμά τις ζημιές)
- **Hα**: μ > 0 (το συνεργείο ΥΠΕΡΕΚΤΙΜΑ τις ζημιές)

Έχουμε :

- **t** = 2.913 (Χρησιμοποιώντας τον τύπο xbar – μ / (s/sqrt(n)))
- **Pvalue** = 1 – Φ(2.913) = 1 – 0.9982 = 0.0018.

Παρατηρούμε ότι το Pvalue είναι ΠΟΛΥ μικρό άρα απορρίπτουμε την μηδενική υπόθεση , το οποίο σημαίνει ότι εν τέλει το συνεργείο υπερεκτιμά τις ζημιές. Τέλος , να σημειώσουμε ότι θα ήταν καλύτερο να είχαμε μεγαλύτερο δείγμα από μετρήσεις ώστε να βγάλουμε ένα πιο αντιπροσωπευτικό συμπέρασμα.

8η Άσκηση

a. Αρχικά , διαχωρίζουμε τα δεδομένα σε αντρών και γυναικών με τις κλασικές μεθόδους και βγάζουμε *NA* όπου υπάρχουν

Χρησιμοποιώντας το *t.test* function παίρνουμε :

```
data:  mdata$height and fdata$height
t = 8.9954, df = 65.471, p-value = 4.745e-13
alternative hypothesis: true difference in means is not equal to
 0
95 percent confidence interval:
 0.09882401 0.15521810
sample estimates:
mean of x mean of y
 1.793600  1.666579
```

Άρα **το 95% διάστημα εμπιστοσύνης** είναι : **[0.09882401, 0.15521810].**

b. Παίρνουμε τις εξής υποθέσεις :

- **μ1** = μέσος όρος βαθμών ανδρών στο μάθημα Πιθανότητες
- **μ2** = μέσος όρος βαθμών γυναικών στο μάθημα Πιθανότητες
- **H0**: μ1 = μ2 (οι άνδρες πήραν περίπου ίδιους βαθμούς κατά μέσο όρο με γυναίκες)
- **Ha**: μ1 > μ2 (οι άνδρες πήραν μεγαλύτερους βαθμούς κατά μέσο όρο από τις γυναίκες).

Διαχωρίζουμε τα δεδομένα και ξεκινάμε να βρίσκουμε τις τιμές :

- **xbarMaleGrades** = 6.280303
- **xbarFemGrades** = 6.838235
- **sMaleGrades** = 2.155724
- **sFemGrades** = 2.006507
- **nMaleGrades** = 66
- **nFemGrades** = 34
- **df** = min{66-1,34-1} = min{65,33} = 33

```
> t <- (xbarMaleGrades - xbarFemGrades)/sqrt((sMaleGrades ^ 2/nMa
leGrades) + (sFemGrades ^ 2/nFemGrades))
> t
[1] 1.234485
```

- **t** = -1.28396
- **Pvalue** = 1-Φ(t) = 1 – Φ(-1.28396) = 1 – 0.1003 = 0.8997.

Άρα Pvalue > α (το οποίο ισούται με 0.05) , το οποίο σημαίνει ότι δεν γίνεται να απορρίψουμε τη μηδενική υπόθεση. Συμπεραίνουμε λοιπόν ότι **δεν** μπορούμε να ξέρουμε ΣΙΓΟΥΡΑ ότι οι άντρες έχουν κατά μέσο όρο μεγαλύτερη βαθμολογία από τις γυναίκες στο μάθημα των Πιθανοτήτων.

c.

```
> probG <- probG[!is.na(probG)]
> mathG <- mathG[!is.na(mathG)]
```

Διαχωρίζουμε τα δεδομένα κατάλληλα (χωρίς NA κλπ).
Και θέλουμε να βρούμε το Pvalue για τις εξής υποθέσεις:

- **H0** : μ1 = μ2
- **Hα** : μ1 <> μ2

Μέσω των κλασικών μεθοδολογιών σε R βρίσκουμε ότι:

- **t** = -0.662008
- **Pvalue** = 2Φ(-0.662008) = 2 * 0.2546 = 0.5092

Παρατηρούμε πως το Pvalue είναι πολύ μεγάλο άρα δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση. Άρα , καταλήγουμε στο συμπέρασμα ότι οι μέσοι βαθμοί των 2 αυτών μαθημάτων δεν διαφέρουν και τόσο.