ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Στατιστική στην Πληροφορική (2023-2024)

Διονύσιος Ρηγάτος (P3200262)

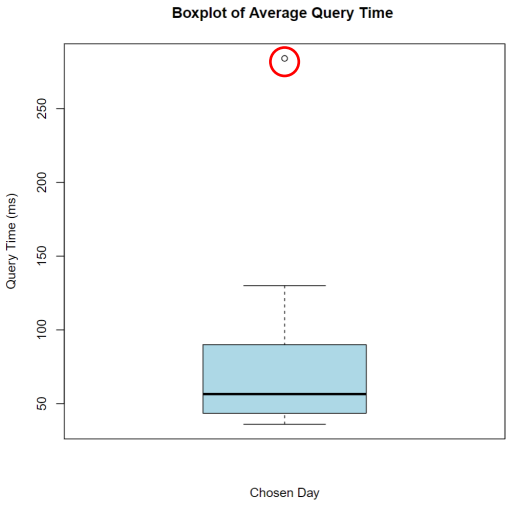
Εργασία 2

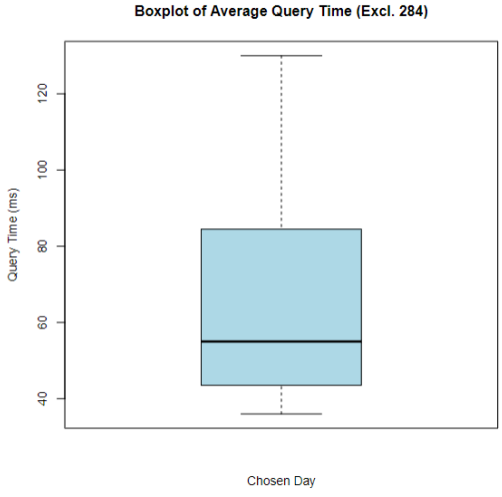
Άσκηση 1

Α) Το δεδομένα μας αποτελούν δείγμα τυχαίας δειγματοληψίας (SRS) αφού επιλέχθηκαν τυχαία από έναν πληθυσμό (queries της μέρας). Το δείγμα μας είναι επαρκές, με n = 20 στοιχεία (20 >= 15).

Το δείγμα μας δεν είναι κανονικά κατανεμημένο καθώς υπάρχει ένα ατυπικό σημείο (284) και δεν είναι αρκετά μεγάλο ώστε να δικαιολογείται αυτό (20 < 40).

Ιδανικά αφαιρούμε το ατυπικό σημείο έτσι ώστε να εφαρμόσουμε μεθόδους συμπερασματολογίας στα δεδομένα μας χωρίς πρόβλημα, αλλιώς το αποτέλεσμα δεν θα είναι ιδανικό. Τα στοιχεία μας είναι τώρα 19 δεν είναι συμμετρικά όμως είναι σίγουρα πιο ορθά για να εξάγουμε συμπεράσματα.





Β) Έχουμε:

```
Standard Deviation (s): 27.5345609513503
Mean (x̄): 66.5263157894737
t*: 2.10092204024104
```

Για 95% διάστημα εμπιστοσύνης, θα έχουμε από τον τύπο:

$$\bar{x} \pm t_* \frac{s}{\sqrt{n}}$$

```
95% Confidence Interval: [53.2550822233055, 79.7975493556419]
```

Το οποίο διάστημα επηρεαζόταν σημαντικά στο upper limit από το ατυπικό σημείο, που ήταν ~103.

Άσκηση 2

Α) Η τυπική απόκλιση του δειγματικού μέσου, βάσει ορισμού, είναι:

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

Και συνεπώς λείπει η ρίζα στον παρονομαστή, άρα έχουμε $\frac{12}{\sqrt{20}}$.

Β) Ο ερευνητής χρησιμοποίησε την δειγματική μέση τιμή για την μηδενική υπόθεση που είναι εκτιμητής, ενώ κανονικά πρέπει να χρησιμοποιηθεί η μέση τιμή του πληθυσμού (μ).

Γ) Για να απορριφθεί η μηδενική υπόθεση θα πρέπει ο δειγματικός μέσος να είναι μεγαλύτερος του 54, έτσι ώστε να δεχτούμε την εναλλακτική υπόθεση.

Όμως εδώ είναι αδύνατο να δεχτούμε την εναλλακτική υπόθεση καθώς ο δειγματικός μέσος είναι 45 < 54.

Δ) Το p-value εδώ είναι 52%, που σημαίνει ότι υπάρχει 52% πιθανότητα να εμφανιστούν παρόμοιες τιμές με το στατιστικό ελέγχου |z|, εφόσον ίσχυε η μηδενική υπόθεση. Άρα είναι παράλογο να απορρίψουμε την μηδενική υπόθεση αφού οι τιμές που έχουμε έχουν υψηλή πιθανότητα να είναι παρόμοιες.

Άσκηση 3

Α) Ha : μ > μ0

```
z: 1.34
Φ(z): 0.909877327535548
p-value (1 - Φ(z)): 0.0901226724644524
```

Β) Ha : μ < μ0

```
z: 1.34
Φ(z): 0.909877327535548
p-value (Φ(z)): 0.909877327535548
```

Γ) Ha : μ ≠ μ0

```
z: 1.34
Φ(z): 0.909877327535548
p-value (2*Φ(-|z|)): 0.180245344928905
```

Άσκηση 4

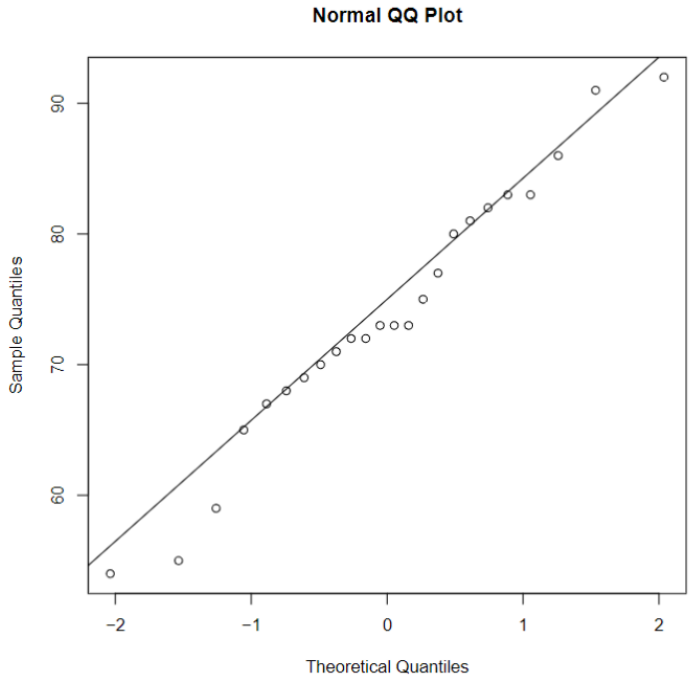
Α) Για διάστημα εμπιστοσύνης 95% έχουμε επίπεδο σημαντικότητας α = 5%, και η μηδενική υπόθεση απορρίπτεται αφού p-value = 0.04 < α = 0.05 και συνεπώς το 30 δεν περιέχεται.

Β) Για διάστημα εμπιστοσύνης 90% έχουμε επίπεδο σημαντικότητας α = 10%, και η μηδενική υπόθεση απορρίπτεται αφού p-value = 0.04 < α = 0.10 και συνεπώς το 30 δεν περιέχεται.

Άσκηση 5

Τα δεδομένα προήλθαν από απλή τυχαία δειγματοληψία (SRS). Περιέχουν, όμως, ένα λάθος entry καθώς ο A/A είναι 6 κιλά και ενήλικας, κάτι προφανώς αδύνατο. Συνεπώς θα ήταν ιδανικό να αφαιρέσουμε την λανθασμένη τιμή από τα δεδομένα πριν προχωρήσουμε στην περαιτέρω ανάλυση τους.

Έπειτα από την αφαίρεση της λανθασμένης τιμής, έχουμε n = 24 entries, κάτι που καθιστά το δείγμα επαρκώς μεγάλο (n >= 15). Ο πληθυσμός επίσης φαίνεται να είναι κανονικά κατανεμημένος από το Normal QQ-plot.



Α) Για 95% διάστημα εμπιστοσύνης, θα έχουμε από τον τύπο:

$$\bar{x} \pm t_* \frac{s}{\sqrt{n}}$$

```
Sample Size: 24
Mean: 73.7916666666667
Standard Deviation: 9.97814641088722
t*: 2.06865761041905
95% Confidence Interval: [69.578264962991, 78.0050683703423]
```

Β) Όμοια με τον ίδιο τύπο.

```
Men Mean: 78.6923076923077
Women Mean: 68
Difference of Means: 10.6923076923077
Men Standard Deviation: 7.59807667971107
Women Standard Deviation: 9.57078889120432
t*: 1.37218364111034
error: 4.90315242437515
80% Confidence Interval: [5.78915526793254, 15.5954601166828]
```

Γ) Αρχικά είναι σημαντικό να τονίσουμε ότι ιδανικά θα εφαρμόσουμε μεθόδους συμπερασματολογίας ξεχωριστά στους άντρες και στις γυναίκες καθώς υπάρχει μια βιολογική διαφορά βάρους που μπορεί να επηρεάσει το αποτέλεσμα. Συνεπώς θα διερευνήσουμε και τις 3 περιπτώσεις για πληρότητα, παρ' όλα αυτά η σύγκριση θα γίνει μεταξύ των περιπτώσεων 2 & 3.

Είναι σημαντικό να τονιστεί πως στις περιπτώσεις που πήραμε τους άντρες και τις γυναίκες ξεχωριστά έχουμε μικρά δείγματα (13 και 11 respectively) και συνεπώς δεν είναι ιδανικά για μεθόδους συμπερασματολογίας.

<u>Περίπτωση 1 – Άντρες & Γυναίκες:</u>

Έστω μΚ το μέσο βάρος των καπνιστών και μΜ το μέσο βάρος των μη καπνιστών.

Ho: μΚ = μΜ (Δεν υπάρχει διαφορά βάρους μεταξύ καπνιστών και μη)

Ha: μΚ ≠ μΜ (Υπάρχει διαφορά βάρους μεταξύ καπνιστών και μη)

```
t = 1.2597, df = 19.281, p-value = 0.2228
```

Με p-value = 0.2228, που είναι και αρκετά υψηλό, αποδεχόμαστε την μηδενική υπόθεση και συνεπώς πως δεν υπάρχει διαφορά βάρους μεταξύ καπνιστών και μη.

<u>Περίπτωση 2 – Άντρες:</u>

Έστω μΚΑ το μέσο βάρος των αντρών καπνιστών και μΜΑ το μέσο βάρος των αντρών μη καπνιστών.

Ho: μΚΑ = μΜΑ (Δεν υπάρχει διαφορά βάρους μεταξύ αντρών καπνιστών και μη)

Ha: μΚΑ ≠ μΜΑ (Υπάρχει διαφορά βάρους μεταξύ αντρών καπνιστών και μη)

```
p-value = 0.1732
```

Με p-value = 0.1732, που είναι και αρκετά υψηλό, αποδεχόμαστε την μηδενική υπόθεση και συνεπώς πως δεν υπάρχει διαφορά βάρους μεταξύ αντρών καπνιστών και μη.

<u>Περίπτωση 3 – Γυναίκες:</u>

Έστω μΚΓ το μέσο βάρος των γυναικών καπνιστών και μΜΓ το μέσο βάρος των γυναικών μη καπνιστών.

Ho: μΚΓ = μΜΓ (Δεν υπάρχει διαφορά βάρους μεταξύ γυναικών καπνιστών και μη)

Ha: μΚΓ ≠ μΜΓ (Υπάρχει διαφορά βάρους μεταξύ γυναικών καπνιστών και μη)

```
t = 0.3155, df = 6.5609, p-value = 0.7622
```

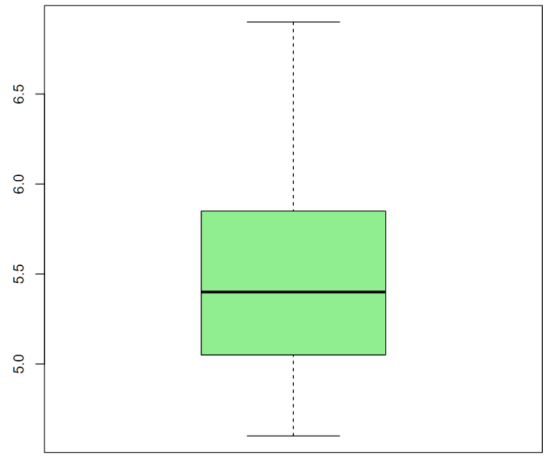
Με p-value = 0.7622, που είναι και αρκετά υψηλό, αποδεχόμαστε την μηδενική υπόθεση και συνεπώς πως δεν υπάρχει διαφορά βάρους μεταξύ γυναικών καπνιστών και μη.

Τελικά, μπορούμε να συμπεραίνουμε ότι οι καπνιστές δεν έχουν διαφορά στο βάρος από τους μη καπνιστές καθώς και τα 2 p-value στις περιπτώσεις 2 & 3 είναι αρκετά υψηλά και συνεπώς η μηδενική υπόθεση δεν απορρίπτεται.

Άσκηση 6

Α) Τα δεδομένα μας επιλέχθηκαν τυχαία (SRS), δεν έχουν ατυπικά σημεία και το πλήθος τους (20) είναι επαρκώς μεγάλο (n >= 15). Επίσης είναι συμμετρικά όπως και πρέπει (20 < 40). Συνεπώς τα δεδομένα μας είναι κατάλληλα για μεθόδους συμπερασματολογίας.

Β)



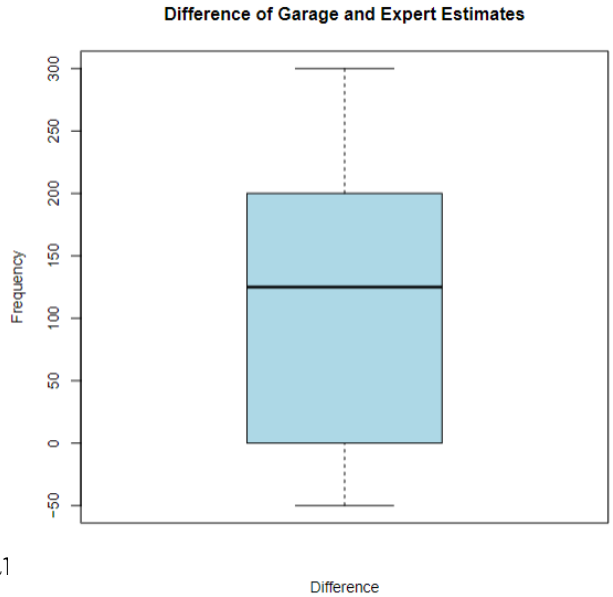

```
Sample S
Mean: 5.5
Standard Deviation: 0.600876552695267
```

Γ)

```
t*: 2.09302405440831
95% Confidence Interval: [5.21878111685868, 5.78121888314132]
```

Άσκηση 7

Το sample μας είναι μικρό (10 < 15). Μας ενδιαφέρει να δοκιμάσουμε τεχνικές συμπερασματολογίας στην διαφορά των δύο samples και παρατηρούμε ότι είναι συμμετρικά, όχι όμως ιδανικά.



Έστω δ η μέση τιμ κόστους του συνεργείου με του δειγματικού μέσου κόστους του εμπειρογνώμονα.

```
Garage Mean: 1220
Expert Mean: 1105
Mean difference (δ): 115
Mean difference standard deviation: 124.833222073827
```

Ho: δ = 0 (Το δειγματικό μέσο κόστος της διαφοράς είναι το ίδιο)

Ha: δ > 0 (Το συνεργείο χρεώνει περισσότερα)

Έστω διάστημα εμπιστοσύνης 95% και επίπεδο εμπιστοσύνης α 5%.

```
p-value = 0.008611
```

Άρα έχουμε ότι p-value < α ⇔ 0.008611 < 0.05 και συνεπώς η εναλλακτική υπόθεση ισχύει. Επομένως το συνεργείο υπερεκτιμάει το κόστος των ζημιών.

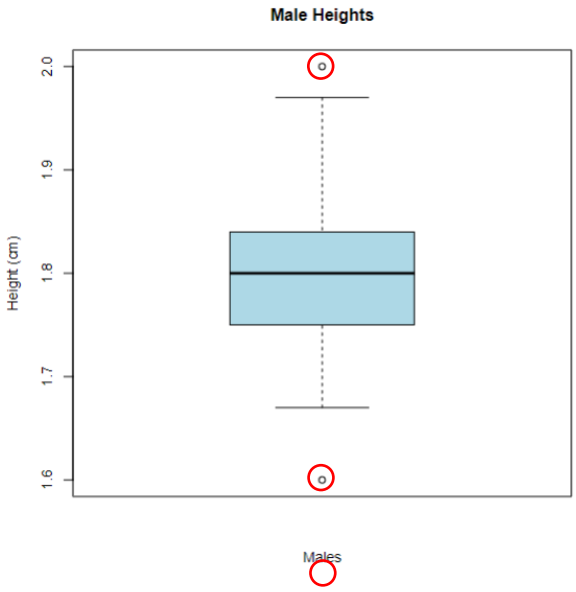
Άσκηση 8

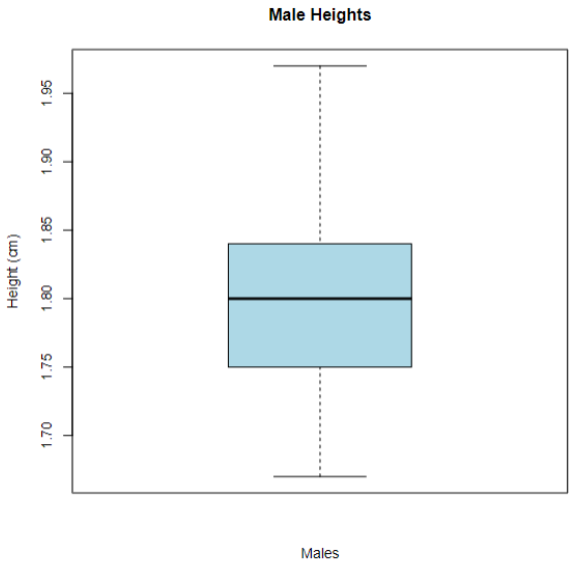
Α) Το sample μας για ύψη φοιτητών είναι μεγάλο, καθώς περιέχει συνολικά 117 (η 115 χωρίς ατυπικά) τιμές. Αφαιρέθηκαν οι φοιτητές που επέλεξαν σαν φύλο OTHER (1) και όσοι δεν απάντησαν στην ερώτηση (0).

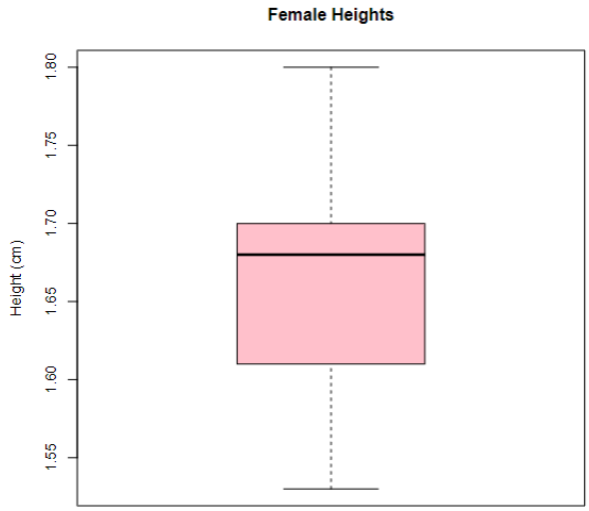
80 φοιτητές απάντησαν MALE (78 χωρίς ατυπικά).

37 φοιτητές απάντησαν FEMALE.

Ακολουθούν τα boxplots που μας δείχνουν τα ατυπικά σημεία, όμως δεν είναι αναγκαίο να είναι συμμετρικά κατανεμημένα καθώς το sample size μας είναι μεγάλο (117 η 115 > 40). Υπάρχουν 2 ατυπικά σημεία, τα οποία όμως δεν επηρεάζουν κάτι όπως φαίνεται και στο διάστημα εμπιστοσύνης. Συνεπώς, με τις παραπάνω πληροφορίες, συμπεραίνουμε ότι μπορούμε να χρησιμοποιήσουμε μεθόδους συμπερασματολογίας σε αυτά.







Το 95% διάστημα εμπιστοσύνης για τη διαφορά του μέσου ύψους μεταξύ αρσενικών και θηλυκών φοιτητών του τμήματος πληροφορικής του ΟΠΑ είναι:

Με ατυπικά σημεία:

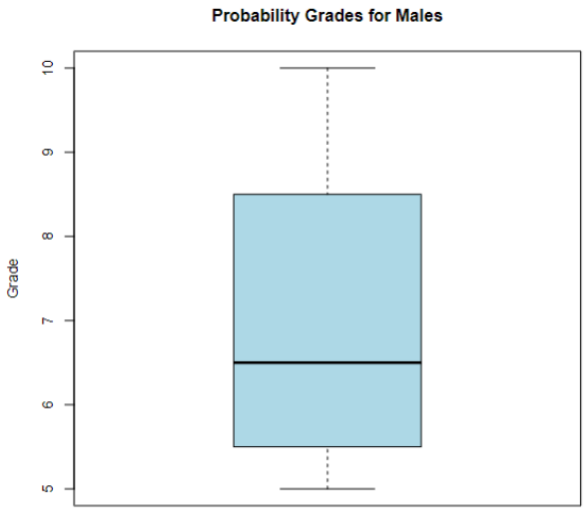
```
95 percent confidence interval:
 0.1004419 0.1548554
```

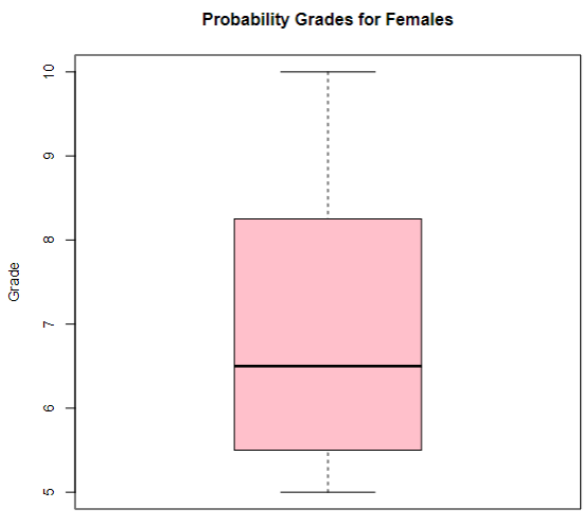
Χωρίς ατυπικά σημεία:

```
95 percent confidence interval:
 0.1011276 0.1541184
```

Β) Φιλτράρουμε τα δεδομένα και αφαιρούμε τους φοιτητές που δεν έχουν δηλώσει βαθμούς στις Πιθανότητες η δεν πέρασαν το μάθημα (βαθμός < 5 η κενό, καθώς δεν υπήρχε τρόπος να ξεχωρίσουμε το εάν κάποιος δεν πέρασε το μάθημα και δεν δήλωσε βαθμό στο ερωτηματολόγιο, επειδή μερικοί δήλωσαν βαθμούς που δεν είναι προβιβάσιμοι) η δήλωσαν OTHER gender. Μένουμε με 95 φοιτητές, εκ των οποίων 31 δήλωσαν FEMALE και 64 MALE.

Το sample μας είναι αρκετά μεγάλο (95 > 40) και συνεπώς είναι καλό για





Θέλουμε να συμπεράνουμε σε επίπεδο σημαντικότητας α = 5% εάν τα males επιτυγχάνουν μεγαλύτερο μέσο βαθμό (μm) στις Πιθανότητες από τα females (μf). Άρα:

Ho: μm <= μf (Δεν επιτυγχάνουν μεγαλύτερο βαθμό τα MALES)

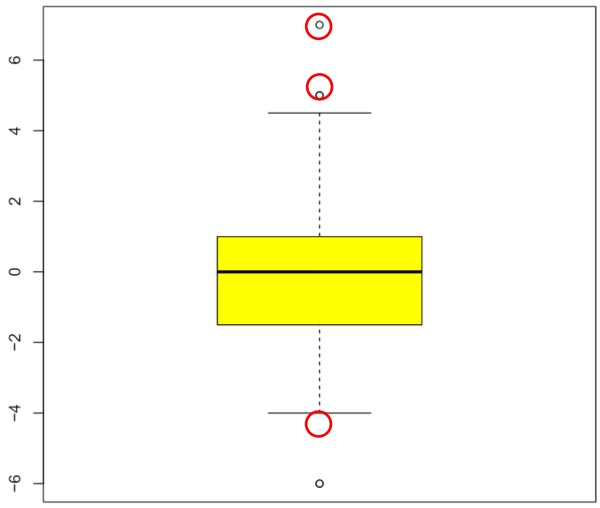
Ha: μm > μf (Επιτυγχάνουν μεγαλύτερο βαθμό από τα FEMALES)

```
t = 0.019055, df = 60.737, p-value = 0.4924
alternative hypothesis: true difference in means is greater than 0
```

Παρατηρούμε πως το p-value είναι ίσο με 0.4924, κάτι που σημαίνει ότι p-value > a ⇔ 0.4924 > 0.05 και συνεπώς η μηδενική υπόθεση δεν απορρίπτεται. Άρα τα males δεν επιτυγχάνουν μεγαλύτερο μέσο βαθμό στις πιθανότητες από τα females σε επίπεδο σημαντικότητας 5%.

Γ) Φιλτράρουμε τα δεδομένα και αφαιρούμε τους φοιτητές που δεν έχουν δηλώσει βαθμούς στις Πιθανότητες ή στα Μαθηματικά 1. Μένουμε με 106 φοιτητές, εκ των οποίων συμπεριλάβαμε και το 1 άτομο που δήλωσε OTHER σαν φύλο μιας και δεν μας αφορά εδώ. Παρ' όλα αυτά, παρατηρούμε ότι υπάρχουν 5 ατυπικά σημεία (μετρήθηκαν στην R, επικαλύπτονται) όπως φαίνεται στο παρακάτω boxplot, τα οποία όμως δεν θα αφαιρέσουμε καθώς πρόκειται για βαθμούς και σίγουρα δεν πρόκειται για κάποιου είδους λάθος η ακραία τιμή.

Τα δεδομένα μας είναι πολλά και συνεπώς μπορούμε να δοκιμάσουμε μεθόδους συμπερασματολογίας (106 > 40).



Έστω μΜ η μέση τιμή των βαθμών στα Μαθηματικά 1 και μΠ η μέση τιμή των βαθμών στις Πιθανότητες. Θέλουμε να δούμε εάν το μΠ - μΜ είναι μεγαλύτερο του 0, δηλαδή εάν υπάρχει διαφορά μεταξύ των βαθμών. Θα το ελέγξουμε σε επίπεδο σημαντικότητας 5%.

Ηο: μΜ = μΠ (μΜ δεν διαφέρει από μΠ)

Ηα: μΜ ≠ μΠ (μΜ διαφέρει από μΠ)

Υπολογίζουμε το p-value (t.test) και παρατηρούμε ότι p-value > a ⇔ 0.1506 > 0.05 και συνεπώς η μηδενική υπόθεση δεν απορρίπτεται.

```
t = -1.4481, df = 105, p-value = 0.1506
alternative hypothesis: true mean is not equal to 0
```

Άρα δεν υπάρχει διαφορά μεταξύ του μέσου βαθμού στα Μαθηματικά 1 και στις Πιθανότητες μεταξύ των φοιτητών που συμπλήρωσαν το ερωτηματολόγιο.

Κώδικας R

Google Drive Link

Ο κώδικας δεν γράφτηκε για να βαθμολογηθεί και συνεπώς είναι προχειρογραμμένος, χωρίς comments και μπορεί να περιέχει περιττές (η να λείπουν) και ακόμα και λανθασμένες εντολές και συμπεριλαμβάνεται μόνο για πληρότητα και ακεραιότητα ώστε οι πράξεις να μην φανούν ουρανοκατέβατες.