ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

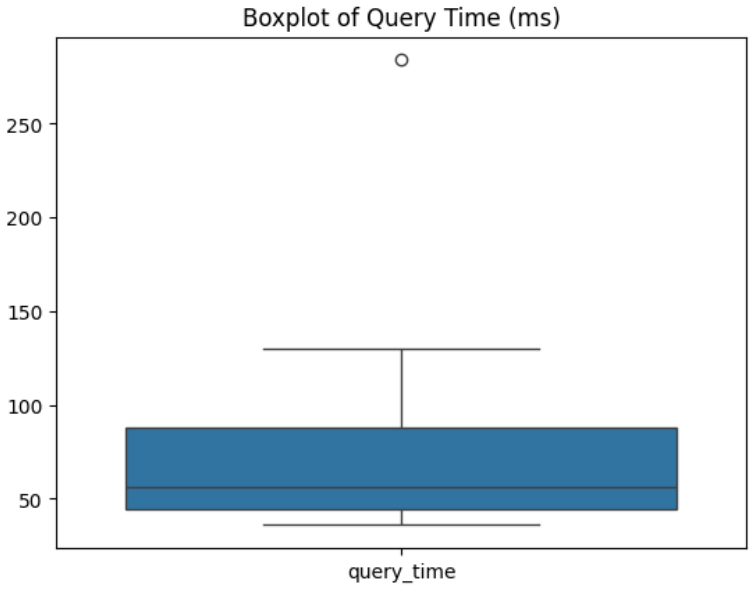
2η Σειρά Ασκήσεων – Στατιστική στην Πληροφορική, 2023-2024

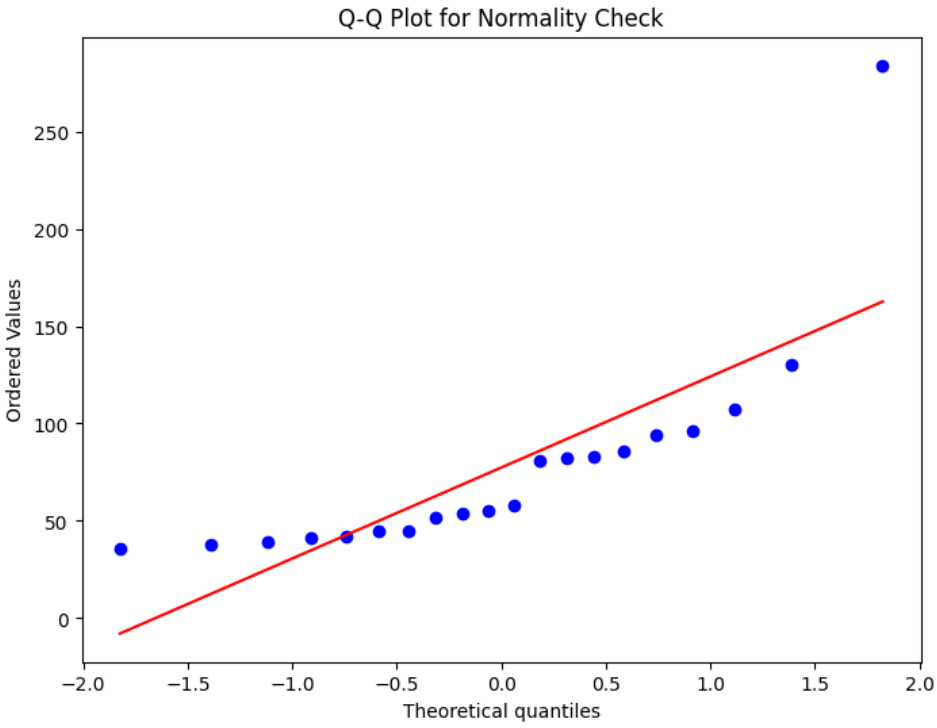
- *Όλες οι ασκήσεις είναι λυμένες επεξηγηματικά σε μορφή jupyter- notebooks στα αρχεία askisi_i.ipynb, i->[1, 8]*

Άσκηση 1).

a) Τα δεδομένα είναι κατάλληλα για τις μεθόδους συμπερασματολογίας που γνωρίζουμε;

- Εφόσον οι χρόνοι επιλέγονται τυχαία και ανεξάρτητα, τα δεδομένα φαίνονται κατάλληλα για τις βασικές μεθόδους συμπερασματολογίας.





- Παρατηρούμε ότι έχουμε outlier με τιμή query_time = 284 milliseconds.
- Τα δεδομένα μας δεν ακολουθούν ιδιαίτερα την κανονική κατανομή.
- Το δείγμα μας όμως είναι αρκετά μεγάλο (n=20 > 15), ώστε μέθοδοι συμπερασματολογίας που βασίζονται σε κατανομή t να έχουν αρκετά καλή ακρίβεια.

b) Δώστε ένα 95% διάστημα εμπιστοσύνης για τη μέση τιμή του χρόνου διεκπεραίωσης

(51.413652441807386, 103.38634755819263)

Άσκηση 2).

a) Λαμβάνεται ένα τυχαίο δείγμα μεγέθους 20 από πληθυσμό με τυπική απόκλιση 12. Η τυπική απόκλιση του δειγματικού μέσου είναι 12/20.

- Έχει γίνει λάθος ο υπολογισμός της τυπικής απόκλισης του δειγματικού μέσου. Θα έπρεπε να είναι 12 / √20 και όχι 12 / 20.

b) Ένας ερευνητής χρησιμοποιεί σε έναν έλεγχο σημαντικότητας τη μηδενική υπόθεση H_0: μ= 10.

- Είναι ασυνήθιστο για μία μηδενική υπόθεση να αναφέρεται σε δειγματική παράμετρο. Συνήθως, μία μηδενική υπόθεση αφορά κάποια παράμετρο σχετική με τον πληθυσμό. Στην συγκεκριμένη περίπτωση θα ήταν πιο αποδεκτό η μηδενική υπόθεση να ήταν η ακόλουθη: μ=10, δεδομένου ότι μ η μέση τιμή του πληθυσμού.

c) Σε μια στατιστική έρευνα με μ= 45 απορρίπτεται η μηδενική υπόθεση H_0: μ= 54 όταν η εναλλακτική είναι H_α: μ>54.

- Η μηδενική υπόθεση μ=54 απορρίπτεται και επιλέγεται η εναλλακτική υπόθεση μ>54, ενώ ισχύει ότι ο δειγματικός μέσος μ είναι μικρότερος από την υποτιθέμενη μέση τιμή (του πληθυσμού) [μ=45 < 54]. Αυτό δεν είναι σωστό, διότι για να αποδεχτούμε την εναλλακτική υπόθεση (H_α: μ>54) θα πρέπει να ισχύει μόνο αν ο δειγματικός μέσος είναι αρκετά μεγαλύτερος του 54 και σίγουρα όχι αν είναι μικρότερος.

d) Σε μια στατιστική έρευνα όπου p value = 0.52 απορρίπτεται η μηδενική υπόθεση.

- Η απόρριψη της μηδενικής υπόθεσης με p-value = 0.52 είναι λανθασμένη. Γενικά μία p-value μεγαλύτερη από το επίπεδο σημαντικότητας (συνήθως 0.05) δεν δύναται να μας οδηγήσει στην απόρριψη της μηδενικής υπόθεσης. Απορρίπτουμε την μηδενική υπόθεση αν η p-value είναι μικρότερη από το

επίπεδο σημαντικότητας (πχ. 0.05) και όχι όταν είναι υψηλότερη (όπως στην συγκεκριμένη περίπτωση που είναι 0.52)

Άσκηση 3).

Σε έλεγχο σημαντικότητας με μηδενική υπόθεση H_0: μ = μ_0 η τιμή του στατιστικού ελέγχου z (z-statistic) είναι 1.34

- p value για την εναλλακτική υπόθεση H_α: μ > μ_0 ? [δεξιόστροφος έλεγχος]
  - p = 0.09012267246445238
- p value για την εναλλακτική υπόθεση Ha: $\mu < \mu 0$? [αριστερόστροφος έλεγχος]
  - p = 0.9098773275355476
- p value για την εναλλακτική υπόθεση H_α: $\mu \neq \mu_0$? [έλεγχος δύο κατευθύνσεων]
  - p = 0.18024534492890476

Άσκηση 4).

Έστω ότι το p value για ένα δίπλευρο έλεγχο με μηδενική υπόθεση H_0 : μ = 30 είναι 0.04.

a). Η τιμή 30 περιέχεται στο 95% διάστημα εμπιστοσύνης για τη μέση τιμή μ; Γιατί;

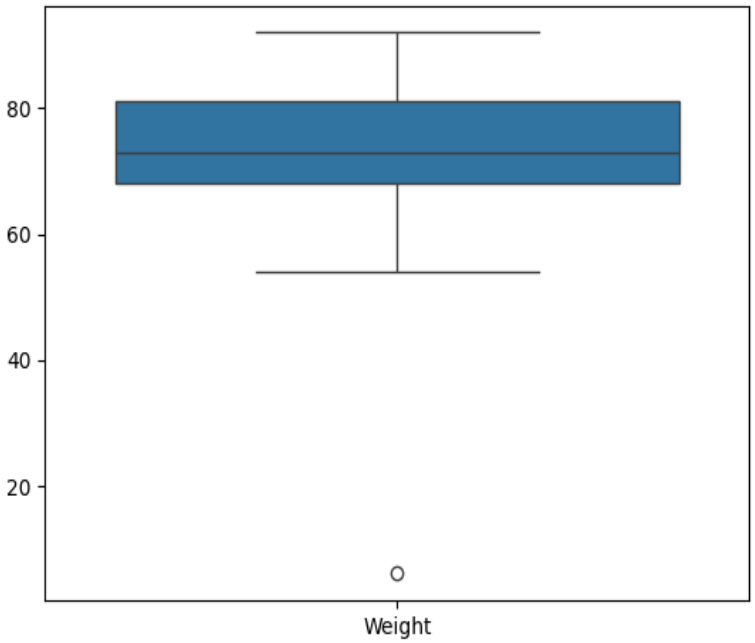
- Εάν η μηδενική υπόθεση (H0: μ = 30) είναι αληθής, τότε η τιμή 30 θα περιλαμβάνεται στο διάστημα εμπιστοσύνης του 95%. Εδώ, το p-value 0.04 είναι μικρότερο από 0.05, έτσι απορρίπτουμε τη μηδενική υπόθεση, που σημαίνει ότι η τιμή 30 δεν περιλαμβάνεται στο 95% διάστημα εμπιστοσύνης.

b). Η τιμή 30 περιέχεται στο 90% διάστημα; Γιατί;

- Εφόσον το p-value του 0.04 είναι μικρότερο και από αυτό το επίπεδο, πάλι απορρίπτουμε τη μηδενική υπόθεση, που σημαίνει ότι η τιμή 30 δεν περιλαμβάνεται ούτε στο 90% διάστημα εμπιστοσύνης.

ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

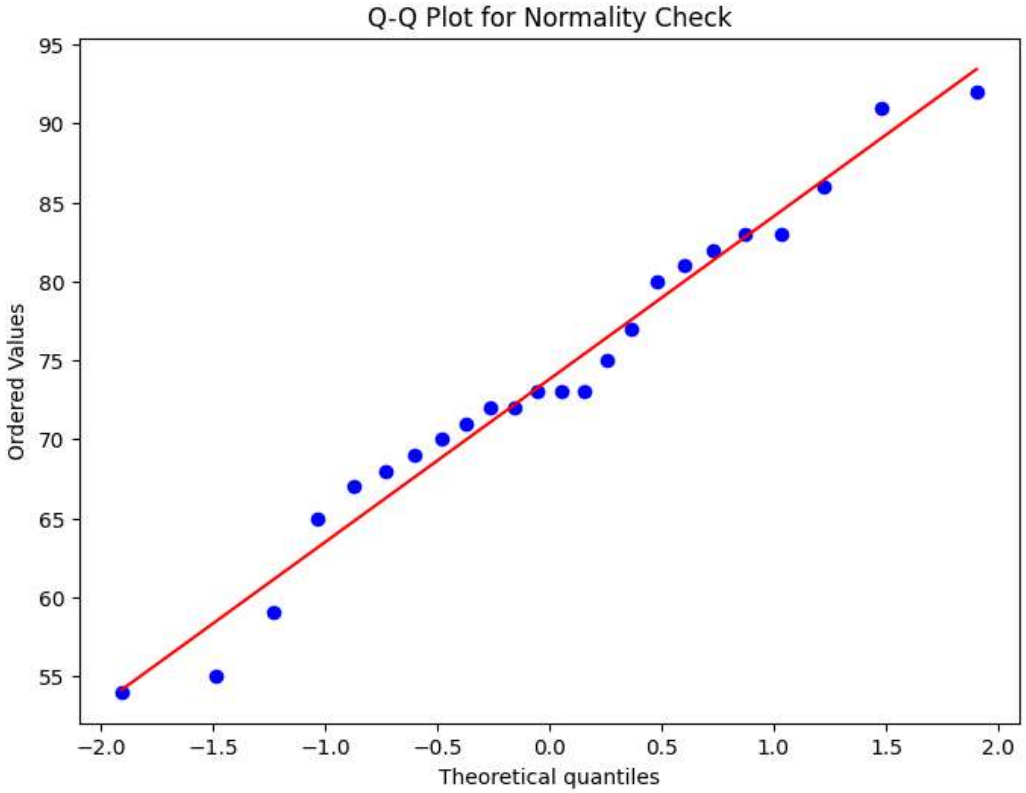
Άσκηση 5).



Παρατηρούμε την ύπαρξη ενός ατυπικού σημείου:

Δεν μπορεί άνθρωπος να έχει βάρος 6 ! (Το αφαιρούμε από το δείγμα μας)

```
# Remove outliers - just one (weight cannot be 6 !)
df = df[df['Weight'] > 10]
```



Από τα παραπάνω γραφήματα συμπαιρένουμε τα εξής:

- Η κατανομή δεν φαίνεται να είναι ιδιαίτερα ασυμμετρική.
- Το δείγμα προέκυψε από απλή τυχαία δειγματοληψία (SRS)
- Τελικό δείγμα, αρκετά μεγάλο. n=24 > 15

Επομένως μπορούμε να εφαρμόσουμε t στατιστικούς ελέγχους και να περιμένουμε αρκετά καλή ακρίβεια.

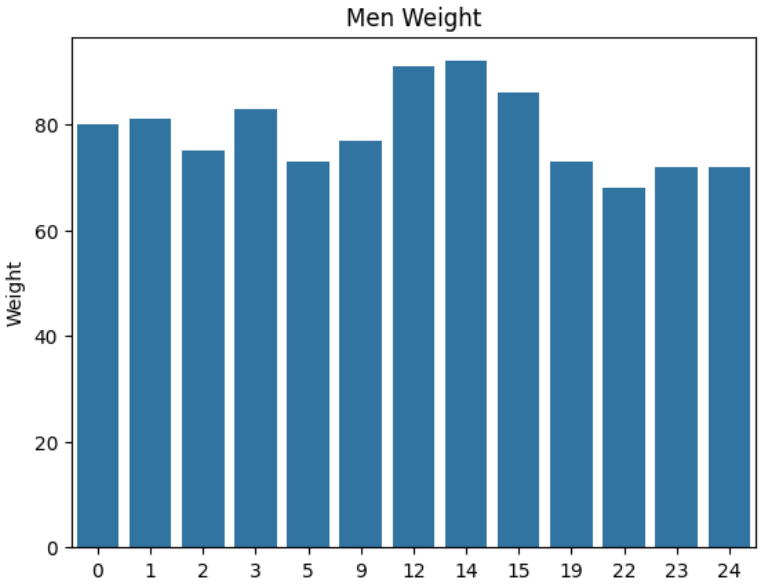
ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

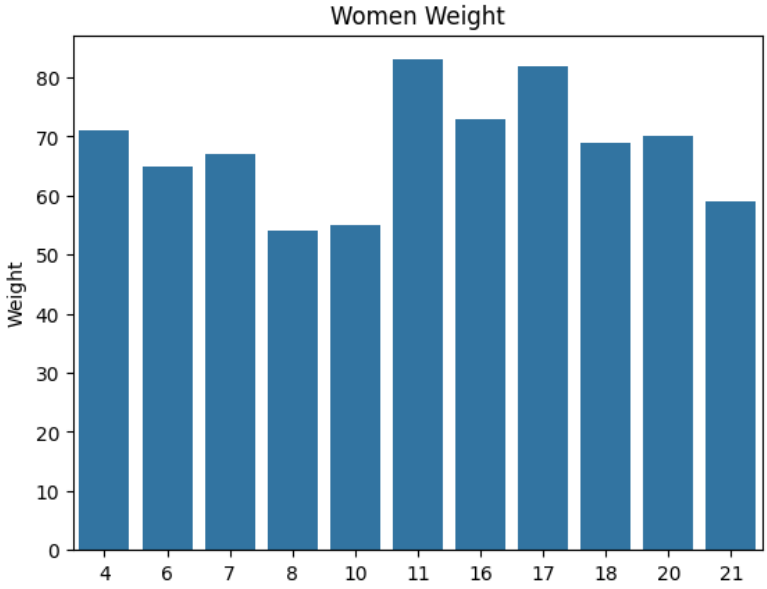
a). Δώστε ένα 95% διάστημα εμπιστοσύνης για το μέσο βάρος των ενηλίκων κατοίκων Αθήνας.

(69.57826496299103, 78.00506837034231)

b). Δώστε ένα 80% διάστημα εμπιστοσύνης για τη διαφορά του μέσου βάρους μεταξύ ανδρών και γυναικών (ενηλίκους κατοίκους Αθηνών).

Παρατηρούμε από τα παρακάτω barplots ότι οι κατανομές των βαρών τόσο των ανδρών όσο και των γυναικών είναι αρκετά συμμετρικές.

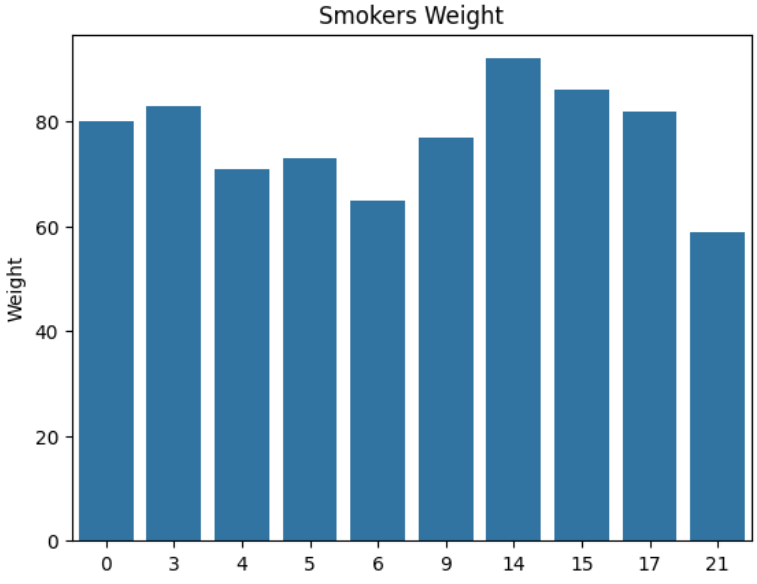


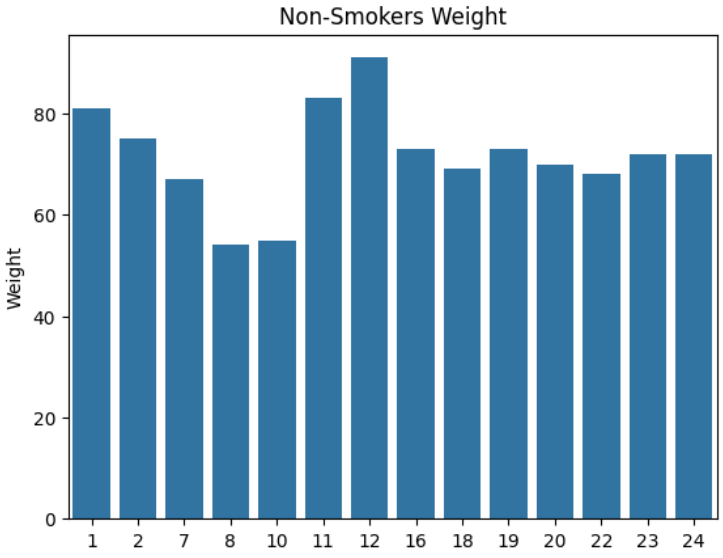


80% Διάστημα εμπιστοσύνης για τη διαφορά των μέσων βαρών των 2 ομάδων :

(5.789155267932715, 15.595460116682672)

c). Το κάπνισμα έχει σχέση με το βάρος; Διατυπώστε έναν κατάλληλο έλεγχο σημαντικότητας και σχολιάστε τα ευρήματά σας.





ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

Παρατηρούμε από τα παραπάνω barplots ότι οι κατανομές των βαρών τόσο των ανδρών όσο και των γυναικών είναι αρκετά συμμετρικές.

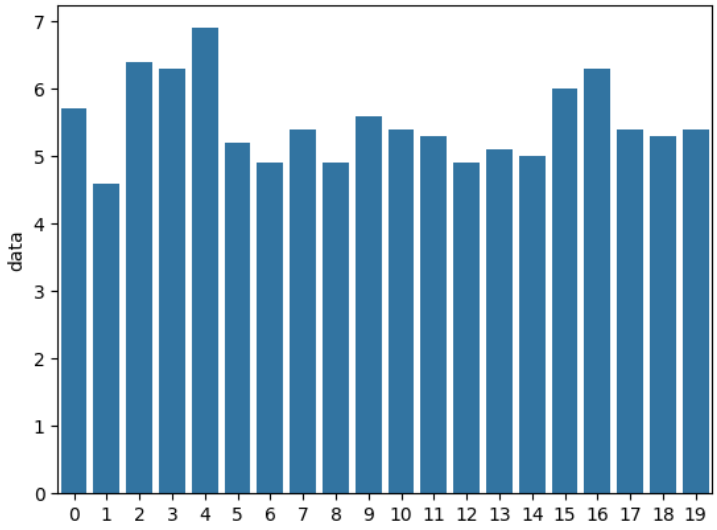
t-statistic (σε απόλυτη τιμή): 1.259715297625312

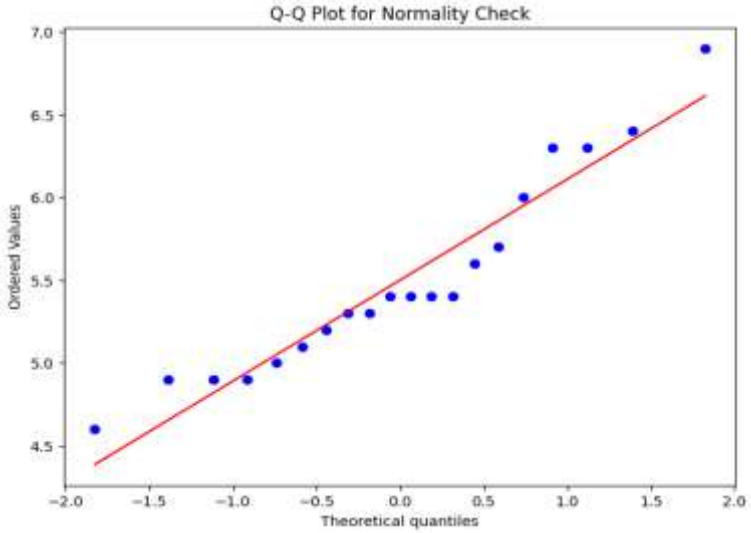
p-value: 0.22280799190887193

- Αφού η τιμή του p-value είναι μεγαλύτερη από το τυπικά αποδεκτό 5% κατώφλι (0.22281 > 0.05), δεν υπάρχει αρκετή στατιστική απόδειξη για να απορρίψουμε την μηδενική υπόθεση, η οποία λέει ότι **ΔΕΝ** υπάρχει διαφορά στο μέσο βάρος καπνιστών και μη καπνιστών.

Άσκηση 6).

a) Τα δεδομένα είναι κατάλληλα για τις μεθόδους συμπερασματολογίας που γνωρίζουμε;





Από τα παραπάνω γραφήματα παρατηρούμε ότι τα δοσμένα δεδομένα είναι αρκετά συμμετρικά, δεν αποκλίνουν πάρα πολύ από την κανονική κατανομή και έχουν ικανοποιητικό μέγεθος (n=20 >15)

b) Βρείτε τη μέση τιμή και τυπική απόκλιση για τα δεδομένα αυτά.

$\bar{x}$ = 5.500000000000001

s = 0.6008765526952665

c) Εκτιμήστε τη μέση τιμή μ της απόδοσης του αυτοκινήτου, με ένα 95% διάστημα εμπιστοσύνης χρησιμοποιώντας τα παραπάνω δεδομένα.

(5.218781116858684, 5.781218883141317)

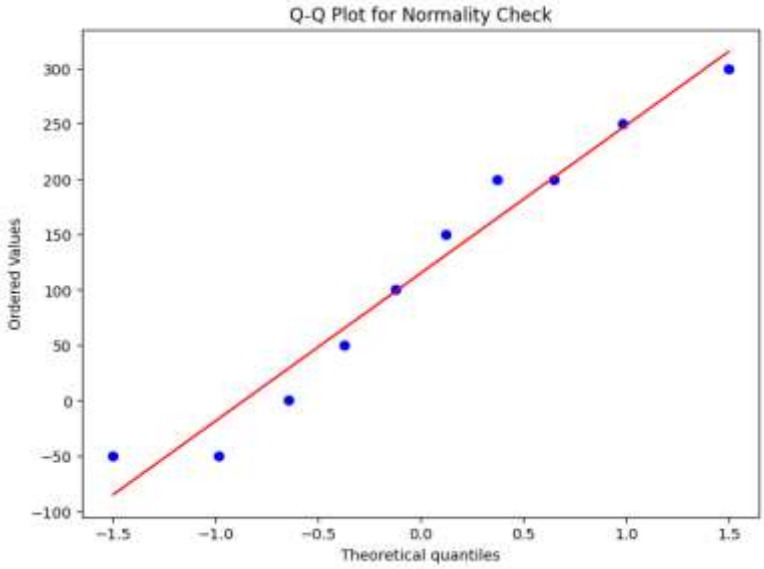
```
confidence_interval_95 = stats.t.interval(confidence=0.95, df=len(df)-1, loc=mean, scale=sem)
confidence_interval_95
```

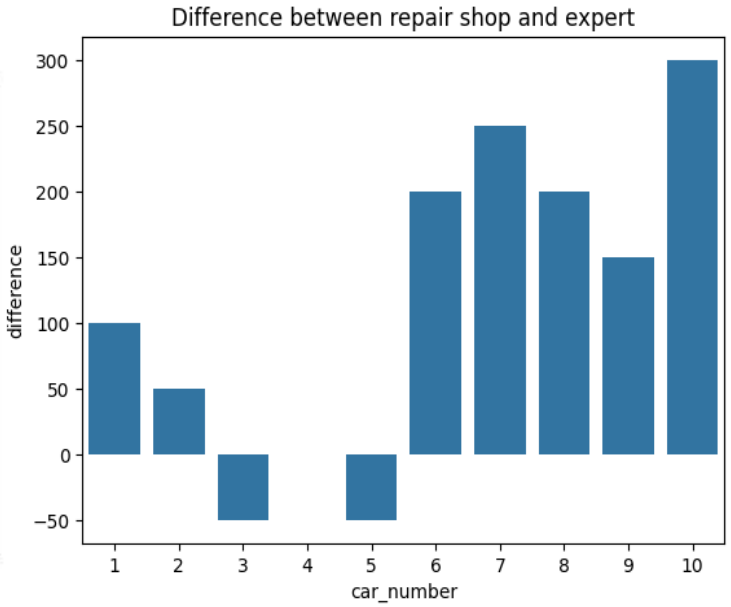
ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

Άσκηση 7).

- (null hypothesis) H0: μ=0 | Δεν υπάρχει υπερεκτίμηση ζημιών από το συνεργείο.
    - Δηλ. Mean(εκτίμηση συνεργείου - εκτίμηση εμπειρογνώμονα) = 0 ή ισοδύναμα Mean(εκτίμηση συνεργείου) = Mean(εκτίμηση εμπειρογνώμονα)
- (alternative hypothesis) H1: μ > 0 (αφού η υπολογισμένη διαφορά είναι συνεργείο - εμπειρογνώμονας) | Το συνεργείο υπερεκτιμά τις ζημιές.
    - Δηλ. Mean(εκτίμηση συνεργείου - εκτίμηση εμπειρογνώμονα) > 0 ή ισοδύναμα Mean(εκτίμηση συνεργείου) > Mean(εκτίμηση εμπειρογνώμονα)

Από τα παρακάτω γραφήματα παρατηρούμε ότι δείγμα μας είναι μικρό (n=10 < 15), όμως η κατανομή είναι αρκετά κοντά στην κανονική. Επομένως, η ακρίβεια των υπολογισμών μας δεν θα είναι πολύ μεγάλη.






```
t_stat, p_value = stats.ttest_1samp(df['difference'], 0, alternative = 'greater')
```

t-statistic (σε απόλυτη τιμή): 2.9131822833531684

p-value: 0.008610910984738962

- Αφού το p-value είναι πολύ μικρότερη από 0.05 (0.008610910984738962 < 0.05), παρατηρούμε ότι υπάρχει στατιστικά σημαντική διαφορά μεταξύ των εκτιμήσεων του συνεργείου και των εκτιμήσεων του εμπειρογνώμονα σε επίπεδο σημαντικότητας 5%. Άρα απορρίπτουμε τη μηδενική υπόθεση.

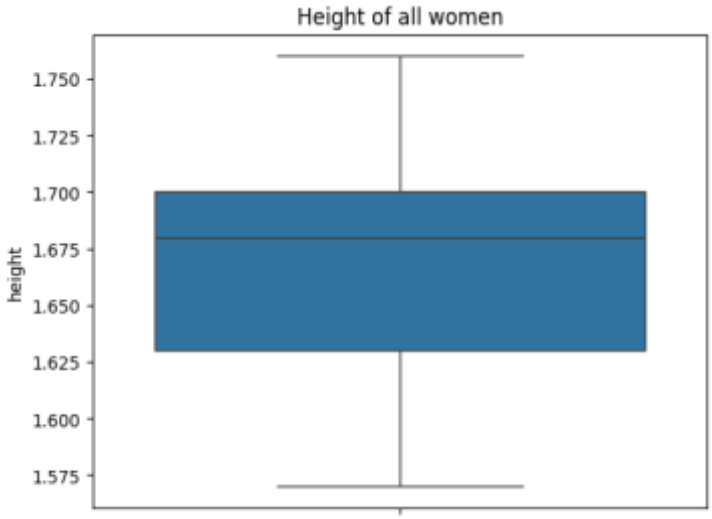
Επομένως συμπεραίνουμε ότι υπάρχει υπερεκτίμηση των ζημιών από το συνεργείο.

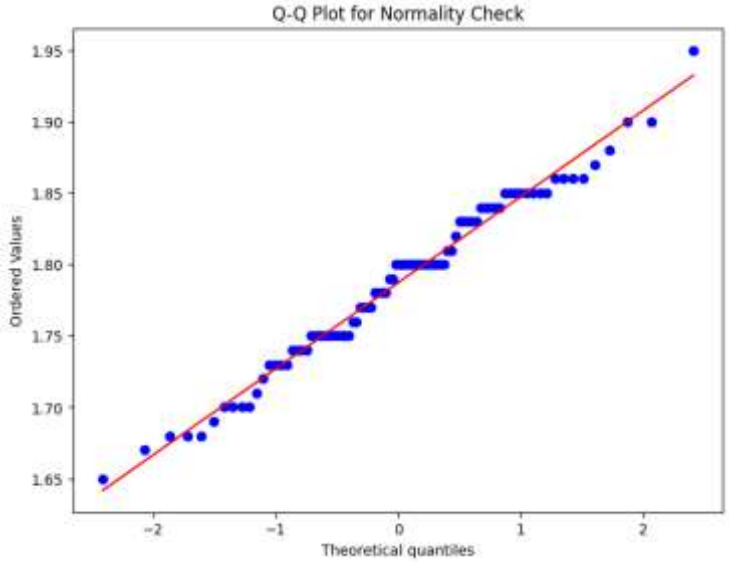
ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

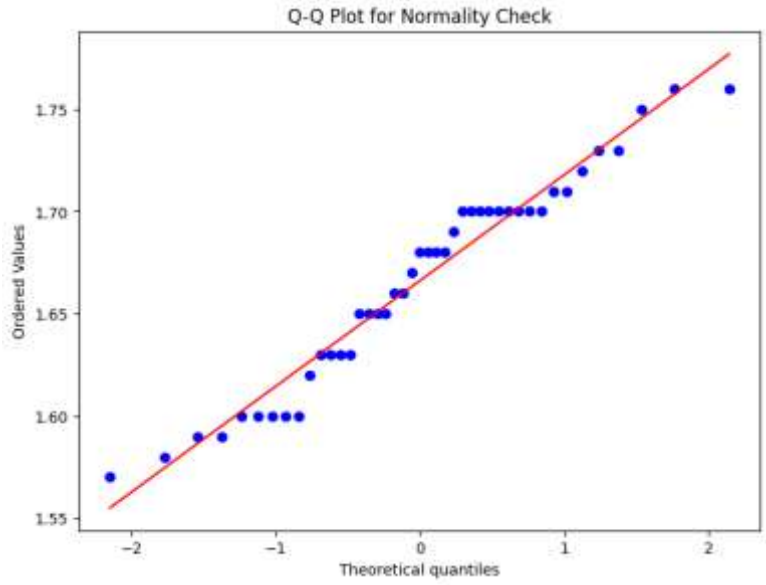
Άσκηση 8).

a) Βρείτε ένα 95% διάστημα εμπιστοσύνης για τη διαφορά του μέσου ύψους μεταξύ ανδρών και γυναικών φοιτητών πληροφορικής του ΟΠΑ.

- Χωρίζουμε τα δεδομένα μας σε δύο ομάδες, σε άνδρες και σε γυναίκες και κρατάμε μόνο την στήλη που περιέχει το ύψος του καθενός.







Από τα παραπάνω γραφήματα για τις δύο ομάδες (άνδρες και γυναίκες) παρατηρούμε ότι τα δείγματά μας είναι αρκετά μεγάλα, δεν έχουν ατυπικά σημεία και δεν απέχουν πολύ από την κανονική κατανομή. Επομένως, μπορούμε να εφαρμόσουμε στατιστικός έλεγχο t με ικανοποιητική ακρίβεια.

95% διάστημα εμπιστοσύνης: (0.10062808763257364, 0.1414649356232402)

b) Οι άνδρες φοιτητές πληροφορικής -που έχουν πάρει ή θα έπαιρναν το μάθημα «Στατιστική στην Πληροφορική»-, επιτυγχάνουν μεγαλύτερο μέσο βαθμό στο μάθημα των Πιθανοτήτων από τον αντίστοιχο πληθυσμό γυναικών; Απαντήστε σε επίπεδο σημαντικότητας 5%.

- Χωρίζουμε τα δεδομένα μας σε δύο ομάδες, σε άνδρες και σε γυναίκες και κρατάμε μόνο την στήλη που περιέχει τον βαθμό του καθενός στις Πιθανότητες.
- Επιπλέον, όσες γραμμές είναι άδειες (Nan τιμές), τις αφαιρούμε.

Εκτέλεση μονόπλευρου t-ελέγχου, θεωρώντας τα 2 δείγματα ανεξάρτητα:

```
t_stat, p_value = stats.ttest_ind(men_prob, women_prob, equal_var=False)
```

t-statistic (σε απόλυτη τιμή) : 0.2514972094471861

p-value : 0.8023000854883574

- Εφόσον το p value είναι πολύ μεγαλύτερο από το επίπεδο σημαντικότητας 0.05, δεν απορρίπτουμε την μηδενική υπόθεση.

Επομένως, δεν μπορούμε να συμπεράνουμε από τα δεδομένα μας, ότι ο μέσος βαθμός των γυναικών είναι μεγαλύτερος από αυτόν των γυναικών.

c) Ο μέσος βαθμός στα Μαθηματικά 1 διαφέρει από το μέσο βαθμό στις Πιθανότητες - μεταξύ των φοιτητών που έχουν πάρει ή θα έπαιρναν το μάθημα «Στατιστική στην Πληροφορική»-;

- Αρχικά, φιλτράρουμε τα δεδομένα μας, αφαιρώντας όλες τις εγγραφές που έχουν άδεια κελιά είτε στην στήλη των Μαθηματικών, είτε στην στήλη των Πιθανοτήτων.

```
# keep only rows that have values for both math and prob
df = df.dropna(subset=['math', 'prob'])
df.head()
```

Εκτέλεση t-ελέγχου για ένα δείγμα, αφού έχουμε ζευγαρώσει τα δεδομένα μας με n=106:

```
t_stat, p_value = stats.ttest_1samp(data['difference'], 0)
```

T-statistic (σε απόλυτη τιμή) : 0.9543203248437889

P-Value : 0.34209277861965337

- Αφού το p-value είναι αρκετά μεγαλύτερο από το επίπεδο σημαντικότητας 0.05, παρατηρούμε ότι δεν υπάρχει στατιστικά σημαντική διαφορά στους μέσους βαθμούς μεταξύ των δύο μαθημάτων.

Επομένως, μπορούμε να συμπεράνουμε ότι οι μέσοι βαθμοί στα Μαθηματικά 1 και στις Πιθανότητες δεν διαφέρουν σημαντικά μεταξύ των φοιτητών που έχουν πάρει ή θα έπαιρναν το μάθημα "Στατιστική στην Πληροφορική".