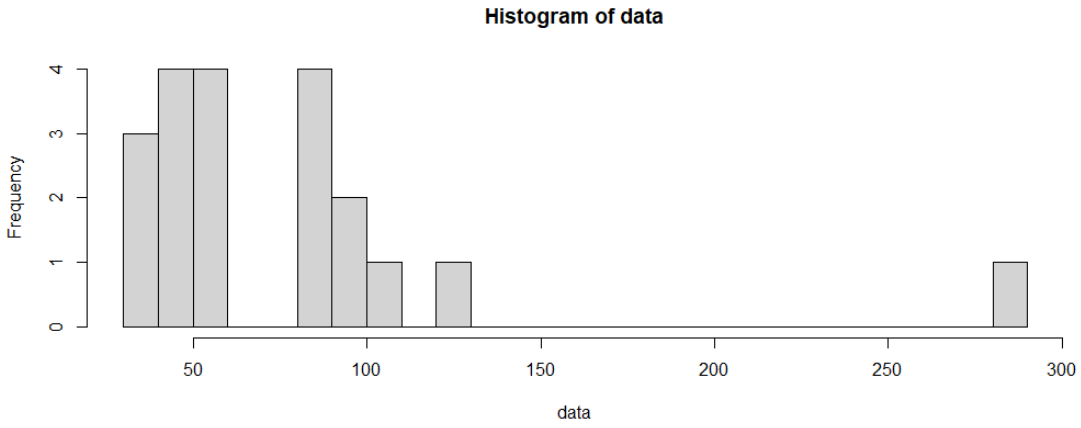
2^{H} Σειρά Ασκήσεων

Ομάδα:

Άγγελος Τσελές (Α.Μ: 3170160)

Ανδρέας Πολυχρονάκης (Α.Μ: 3170140)

Άσκηση 1

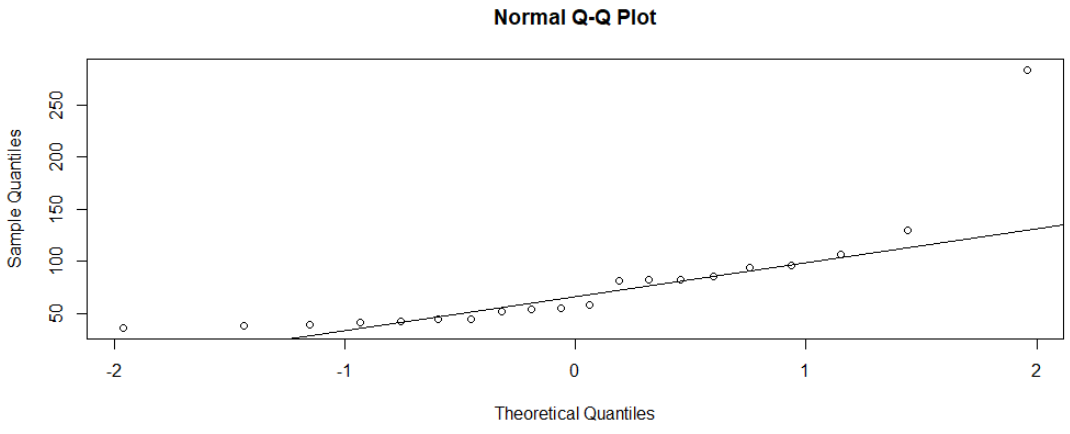


a) Προκειμένου να διαπιστώσουμε αν τα δεδομένα είναι κατάλληλα για τις μεθόδους συμπερασματολογίας εξετάζουμε τρεις παραμέτρους:

-Πρέπει τα δεδομένα να προέρχονται από ένα απλό τυχαίο δείγμα(SRS),κάτι που ισχύει στην άσκηση αυτή.

-Πρέπει το μέγεθος δείγματος να είναι μεγαλύτερο του 15 ,ώστε να πετύχουμε όσο το δυνατόν καλύτερη ακρίβεια. Στην άσκηση αυτή έχουμε μέγεθος δείγματος ίσο με 20.

- Εξετάζουμε τα ατυπικά σημεία και την συμμετρικότητα των δεδομένων. Παρατηρούμε στο παραπάνω ιστόγραμμα ότι υπάρχει ένα ατυπικό σημείο, ενώ επίσης ο πληθυσμός έχει ασύμμετρη κατανομή.



Επομένως, τα δεδομένα δεν είναι κατάλληλα.

b) Δεν γνωρίζουμε την τυπική απόκλιση οπότε θα χρησιμοποιήσουμε την Κατανομή t.

Δειγματικός μέσος x= 77,4

n=20

df=n-1=19

Για C=0.95 : t*=2.093024

Ελάχιστο Όριο = 51.41365

Μέγιστο Όριο = 103.3863

```
> mean(values,na.rm=TRUE) -> x
> x
[1] 77.4
> abs(qt(0.025,df=19)) -> t
> t
[1] 2.093024
> mt <- t*sd(values,na.rm=TRUE)/sqrt(20)
> mt
[1] 25.98635
> x+mt
[1] 103.3863
> x-mt
[1] 51.41365
```

Επομένως το διάστημα εμπιστοσύνης είναι το [51.41365,103.3863].

Άσκηση 2

a) Είναι λάθος διότι η τυπική απόκλιση του δειγματικού μέσου είναι ίση με

Τυπική Απόκλιση / Ρίζα(Δείγμα) δηλαδή ίση με 12/sqrt(20) = 2.683282

b) Είναι λάθος γιατί η μηδενική υπόθεση δεν βασίζεται στον εκτιμητή αλλά στην παράμετρο του πληθυσμού.

c) Είναι λάθος γιατί πρέπει να γνωρίζουμε εκ των προτέρων το επίπεδο σημαντικότητας του ελέγχου ώστε να το συγκρίνουμε με το p value και τελικά να απορρίψουμε ή όχι την μηδενική υπόθεση.

d) Είναι λάθος γιατί το p value πρέπει να είναι πολύ μικρότερο του 0.52 για να απορριφθεί η μηδενική υπόθεση.

2^{H} Σειρά Ασκήσεων

Άσκηση 3

3) Ho : μ=μ0

z=1.34

a) Ηα : μ > μ0

p value = 1-Φ(z) = 1-Φ(1.34) = 1-0.91 = 0.09 = 0.090123

b) Ηα : μ < μ0

p value = Φ(z) = Φ(1.34) = 0.909877

c) Ηα : μ ≠ μ0

p value = 2Φ(−|z|) = 2Φ(-|1.34|)=2*Φ(-1.34) = 0.180254

Άσκηση 4

H0: μ=30

p value=0.04

Α) Στο διάστημα εμπιστοσύνης 95% αντιστοιχεί επίπεδο σημαντικότητας ελέγχου α=5%

Έχουμε ότι: p value <= a <=> 4% <= 5% ΙΣΧΥΕΙ

Άρα η μηδενική υπόθεση απορρίπτεται οπότε η τιμή 30 δεν περιέχεται στο 95% διάστημα εμπιστοσύνης για τη μέση τιμή μ.

Β) Στο διάστημα εμπιστοσύνης 90% αντιστοιχεί επίπεδο σημαντικότητας ελέγχου α=10%

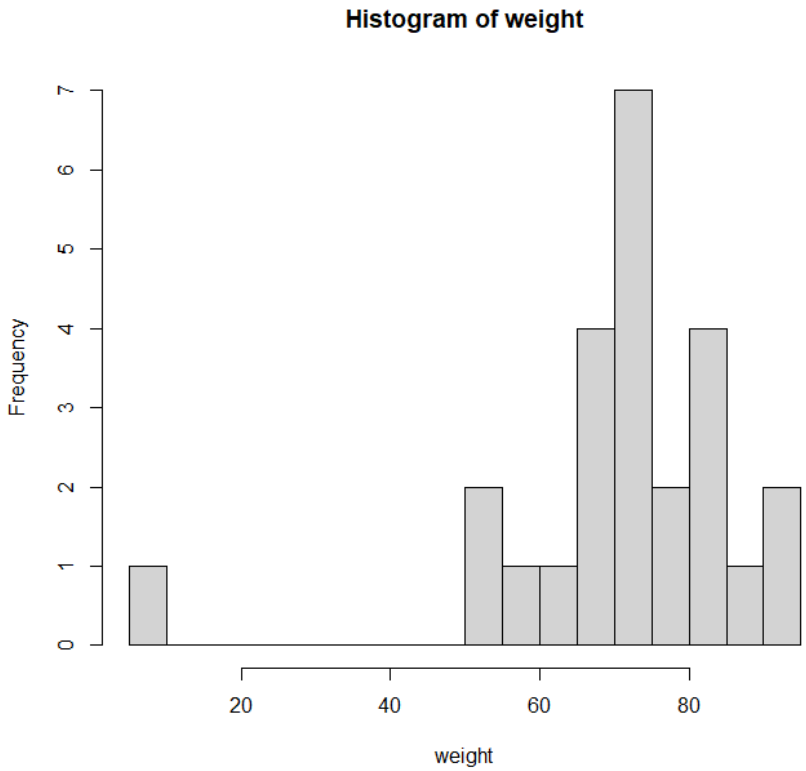
Έχουμε ότι : p value <= a <=> 4% <= 10% ΙΣΧΥΕΙ

Άρα η μηδενική υπόθεση απορρίπτεται οπότε η τιμή 30 δεν περιέχεται στο 90% διάστημα εμπιστοσύνης για τη μέση τιμή μ.

Άσκηση 5

α) Παρατηρούμε ότι υπάρχει ατυπικό σημείο (outliers) που επηρεάζει σημαντικά το διάστημα, όπως φαίνεται και στην παρακάτω εικόνα. Η τιμή 6 δεν υφίσταται για

βάρος, οπότε τα δεδομένα δεν είναι κατάλληλα. Θα αφαιρέσουμε την τιμή αυτή και θα ελέγξουμε τα νέα δεδομένα.

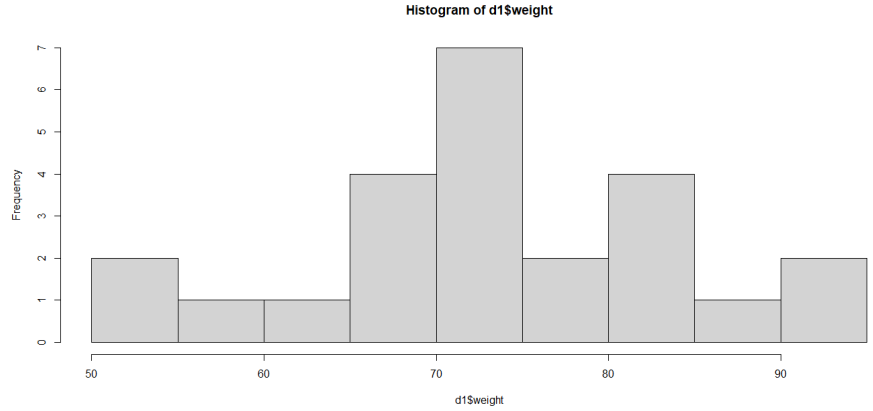


Για τα νέα δεδομένα, ο τύπος για τα διαστήματα εμπιστοσύνης είναι ακριβής διότι:

1. Τα δεδομένα προήλθαν από απλή τυχαία δειγματοληψία.

2. Το δείγμα n είναι μεγαλύτερο του 15

3. Δεν υπάρχουν, πλέον, ατυπικά σημεία (outliers) που επηρεάζουν σημαντικά το διάστημα, όπως φαίνεται και στην παρακάτω εικόνα, ενώ η κατανομή είναι αρκετά συμμετρική.



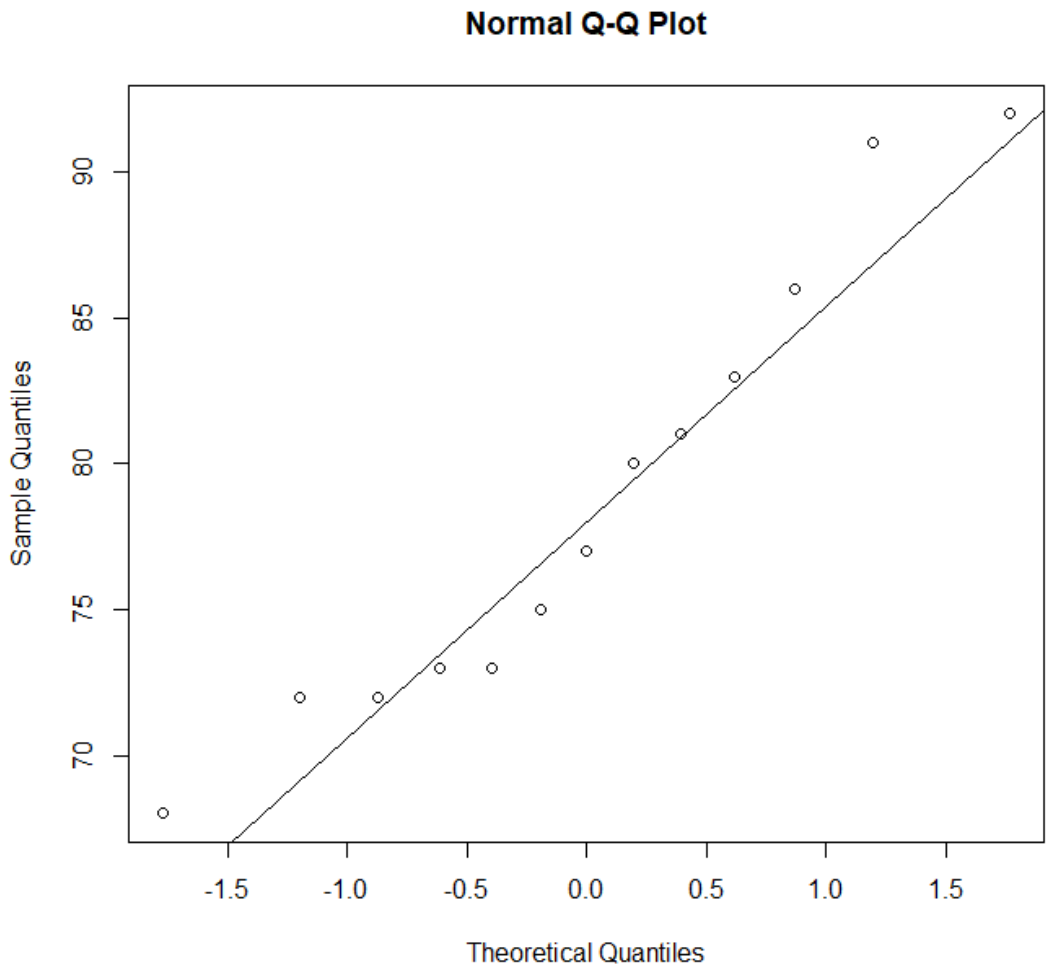
2Η Σειρά Ασκήσεων

Κάνοντας χρήση της t.test βρίσκουμε τα εξής:

```
> t.test(dl$weight)

        One Sample t-test

data:  dl$weight
t = 36.23, df = 23, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 69.57826 78.00507
sample estimates:
mean of x
 73.79167
```

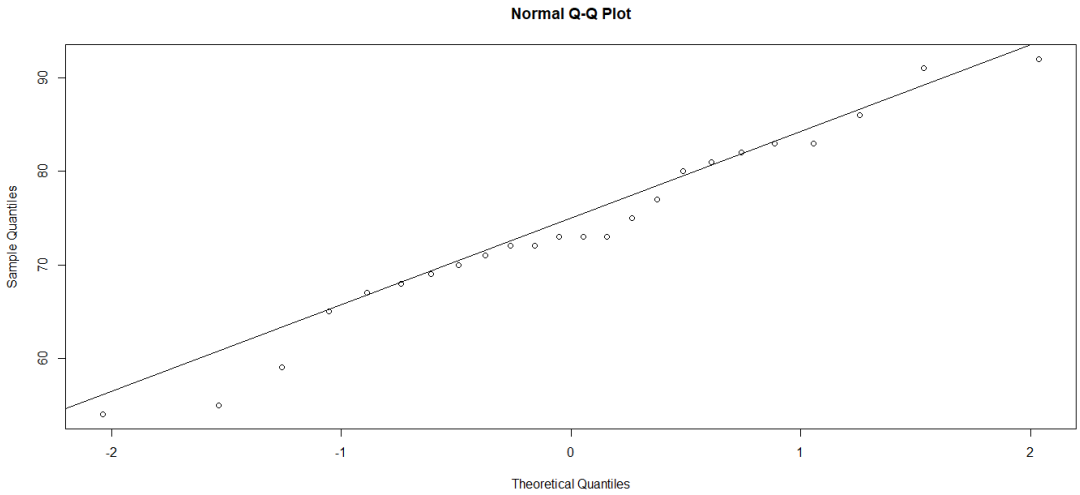
Άρα διάστημα εμπιστοσύνης είναι [69.57826,78.00507] με t* = 36.23

b)



Στην παραπάνω εικόνα, απεικονίζεται η κανονική κατανομή για το βάρος των αντρών. Παρατηρούμε ότι δεν υπάρχει κάποιο ατυπικό σημείο και η κατανομή δεν είναι και πολύ ασύμμετρη.

2Η Σειρά Ασκήσεων



Στην παραπάνω εικόνα, απεικονίζεται η κανονική κατανομή για το βάρος των γυναικών. Παρατηρούμε ότι δεν υπάρχει κάποιο ατυπικό σημείο και η κατανομή δεν είναι και πολύ ασύμμετρη.

Επομένως, τα δεδομένα είναι κατάλληλα.

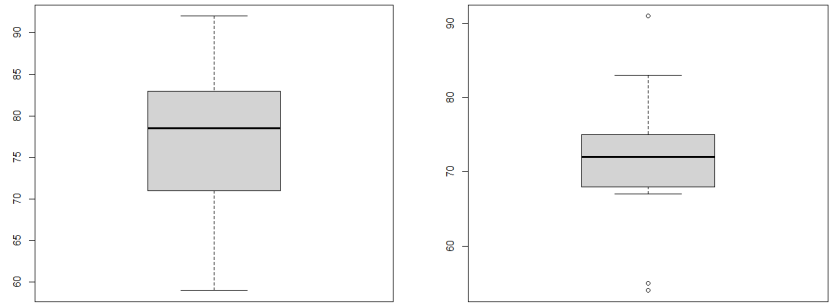
Χρησιμοποιούμε την t test για να βρούμε το διάστημα εμπιστοσύνης 80%

```
> t.test(weight[sex=="M"],weight[sex=="F"],conf.level=0.80)

        Welch Two Sample t-test

data:  weight[sex == "M"] and weight[sex == "F"]
t = 2.5701, df = 13.875, p-value = 0.02235
alternative hypothesis: true difference in means is not equal to 0
80 percent confidence interval:
  7.555799 24.162149
sample estimates:
mean of x mean of y
 78.69231  62.83333
```

Άρα διάστημα εμπιστοσύνης είναι [7.555799,24.162149] με t* = 36.23

c) Φτιάχνουμε τα boxplots του βάρους για καπνίζοντες και μη καπνίζοντες (αριστερό boxplot για καπνίζοντες και δεξί boxplot για μη καπνίζοντες)

2Η Σειρά Ασκήσεων



Καταρχάς ,υπάρχουν κάποια λίγα ατυπικά σημεία στους μη καπνίζοντες τα οποία δεν επηρεάζουν πουθενά. Έστω μ1 το μέσο βάρος των καπνιζόντων και μ2 το μέσο βάρος των μη καπνιζόντων.

Μηδενική Υπόθεση Η0: μ1=μ2

Εναλλακτική Υπόθεση Ηα: μ1 > μ2

Κάνουμε το αντίστοιχο t.test και προκύπτει ότι:

```
> t.test(weight[smoker=="YES"],weight[smoker=="NO"],alternative='greater')

        Welch Two Sample t-test

data:  weight[smoker == "YES"] and weight[smoker == "NO"]
t = 1.6114, df = 21.945, p-value = 0.0607
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.6270107        Inf
sample estimates:
mean of x mean of y
 76.80000  67.26667
```

Για p value= 0.0607 >= a=0.05, δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση. Επομένως, δεν μπορούμε να απαντήσουμε με βεβαιότητα ότι ο μέσος βάρους των καπνιζόντων είναι μεγαλύτερος του μέσου βάρους των μη καπνιζόντων. Άρα το κάπνισμα δεν έχει σχέση με το βάρος.

Άσκηση 6

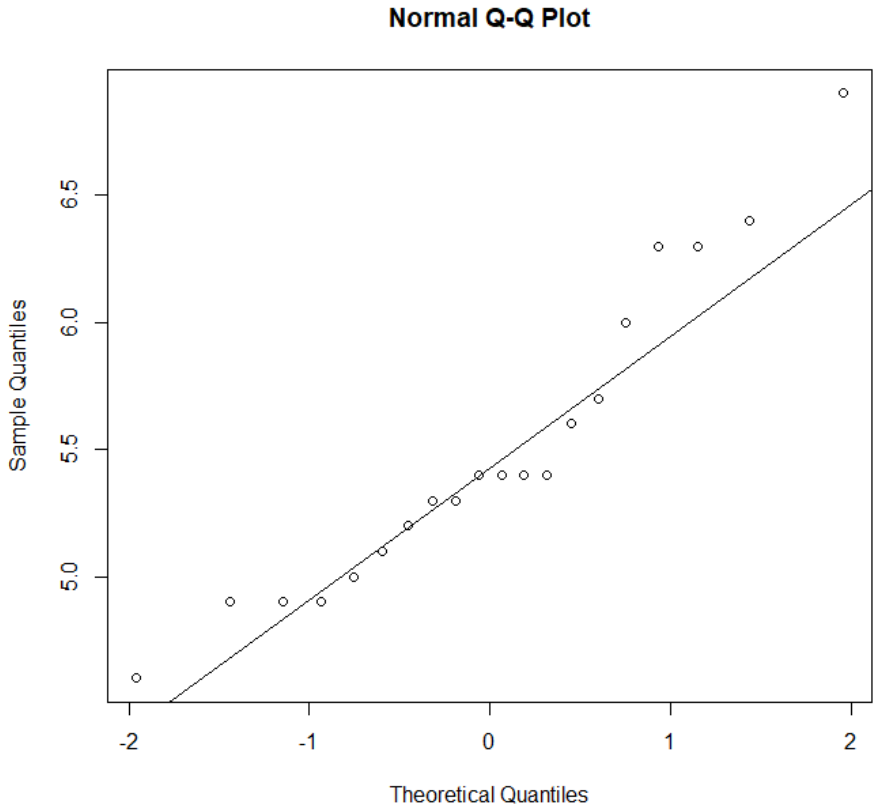
a) Προκειμένου να διαπιστώσουμε αν τα δεδομένα είναι κατάλληλα για τις μεθόδους συμπερασματολογίας εξετάζουμε τρεις παραμέτρους:

-Πρέπει τα δεδομένα να προέρχονται από ένα απλό τυχαίο δείγμα(SRS),κάτι που ισχύει στην άσκηση αυτή.

-Πρέπει το μέγεθος δείγματος να είναι μεγαλύτερο του 15 ,ώστε να πετύχουμε όσο το δυνατόν καλύτερη ακρίβεια. Στην άσκηση αυτή έχουμε μέγεθος δείγματος ίσο με 20.

-Δημιουργούμε την κανονική κατανομή στην οποία εξετάζουμε τα ατυπικά σημεία και την συμμετρικότητα των δεδομένων. Δεν υπάρχουν ατυπικά σημεία που να

επηρεάζουν σημαντικά. Υπάρχει ασυμμετρία αλλά το γεγονός ότι το μέγεθος δείγματος είναι 20 το υπερκαλύπτει.



Άρα τα δεδομένα είναι κατάλληλα για συμπερασματολογία.

b) Μέση τιμή = 5.5

Τυπική Απόκλιση = 0.6

c)

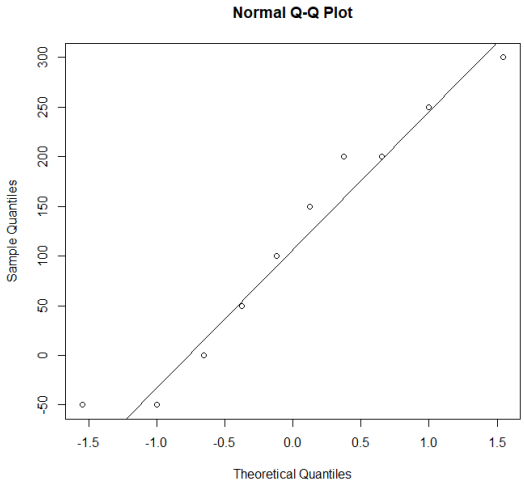
```
> t.test(values)

        One Sample t-test

data:  values
t = 40.935, df = 19, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 5.218781 5.781219
sample estimates:
mean of x
      5.5
```

Διάστημα εμπιστοσύνης 95% είναι [5.2118781,5.781219] με t*= 40.935

2Η Σειρά Ασκήσεων

Άσκηση 7

Θα ελέγξουμε πρώτα αν τα δεδομένα είναι κατάλληλα. Εξετάζουμε την κανονική κατανομή της αφαίρεσης συνεργείου και εμπειρογνώμονα(έστω μ).



Παραπάνω φαίνεται η κανονική κατανομή δεν είναι ασύμμετρη και δεν υπάρχουν ατυπικά σημεία που να επηρεάζουν σημαντικά. Επομένως, τα δεδομένα είναι κατάλληλα.

Ζητείται να εξετάσουμε την υπερεκτίμηση του συνεργείου οπότε έχουμε:

Μηδενική Υπόθεση H0 : μ=0

Εναλλακτική Υπόθεση Ηα : μ>0

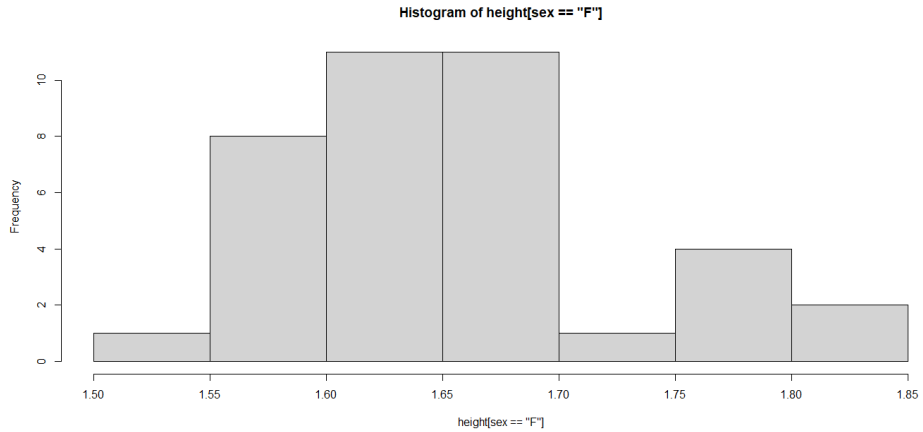
```
> #H0 : μ=0 ,Hα: μ>0
> t.test(service-expert,alternative='greater')

        One Sample t-test

data:  service - expert
t = 2.9132, df = 9, p-value = 0.008611
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 42.63653      Inf
sample estimates:
mean of x
      115
```

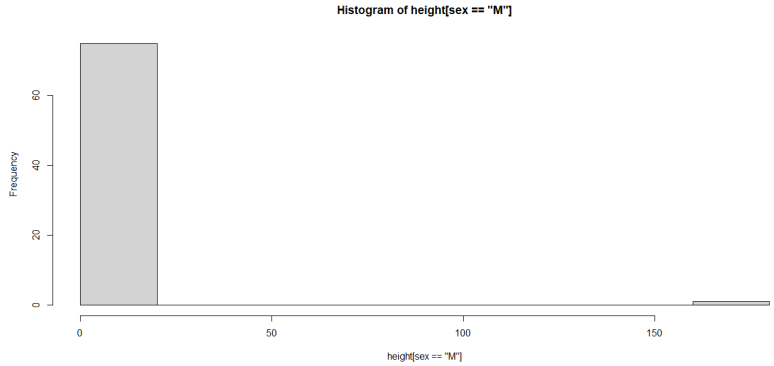
Παρατηρούμε ότι η τιμή του p value είναι πολύ μικρή οπότε θα απορρίψουμε την μηδενική υπόθεση .Επομένως, υπάρχει υπερεκτίμηση ζημιών από το συνεργείο.

2^{H} Σειρά Ασκήσεων

Άσκηση 8

a) Καταρχάς, θα ελέγξουμε τα δεδομένα. Δημιουργούμε ιστόγραμμα για το ύψος των αντρών και για το ύψος των γυναικών.



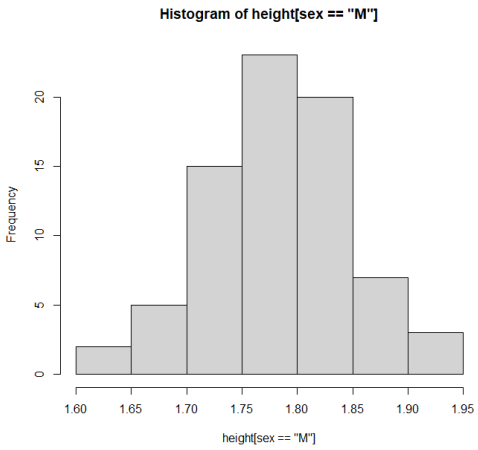


Παρατηρούμε ότι ,ως προς τις γυναίκες, δεν υπάρχουν ατυπικά σημεία που να έχουν κάποια πολύ μεγάλη σημασία και δεν υπάρχει πολύ μεγάλη ασυμμετρία .Επομένως, τα δεδομένα των γυναικών είναι κατάλληλα.

Παρατηρούμε ότι ,ως προς τους άντρες, υπάρχει ατυπικό σημείο με τιμή 176,το οποίο εννοείται δεν ανταποκρίνεται στην πραγματικότητα και επηρεάζει σε μεγάλο βαθμό .Επομένως, τα δεδομένα των ανδρών δεν είναι κατάλληλα.

Μία καλή λύση σε αυτό το πρόβλημα θα ήταν να αφαιρέσουμε την τιμή 176.

2Η Σειρά Ασκήσεων



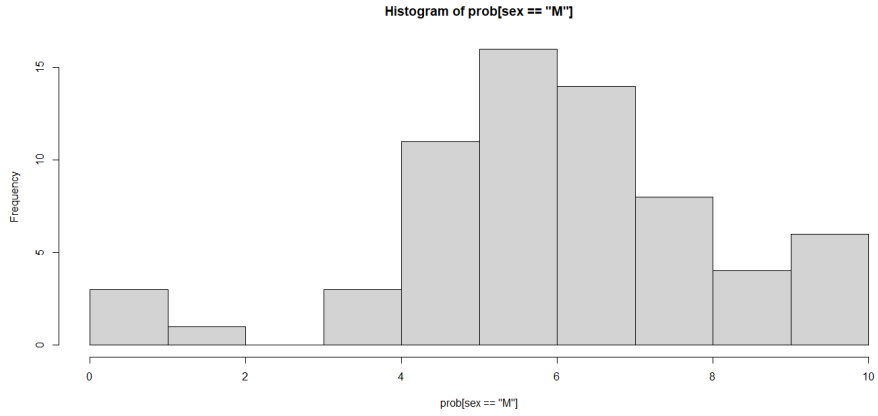
Παρατηρούμε ότι δεν υπάρχουν σημαντικά ατυπικά σημεία και υπάρχει καλή συμμετρία. Επομένως, πλέον, τα δεδομένα είναι κατάλληλα.

Εκτελούμε το κατάλληλο t.test και παίρνουμε ότι:

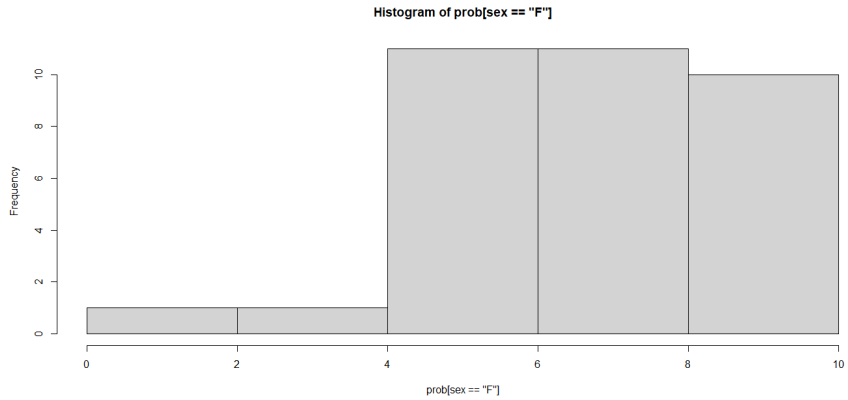
```
> t.test(height[sex=="M"],height[sex=="F"])

        Welch Two Sample t-test

data:  height[sex == "M"] and height[sex == "F"]
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 διάστημα εμπιστοσύνης 95% είναι [0.09882401,0.15521810]

b) Ελέγχουμε πρώτα τα δεδομένα για τον βαθμό των ανδρών και των γυναικών στο μάθημα των Πιθανοτήτων.



2Η Σειρά Ασκήσεων



Έστω ότι μ1 = ο μέσος βαθμός των ανδρών φοιτητών πληροφορικής που έχουν πάρει ή θα έπαιρναν το μάθημα «Στατιστική στην Πληροφορική» και μ2 = ο αντίστοιχος πληθυσμός των γυναικών.

Μηδενική Υπόθεση H0 : μ1 = μ2
Εναλλακτική Υπόθεση Ηα: μ1 > μ2

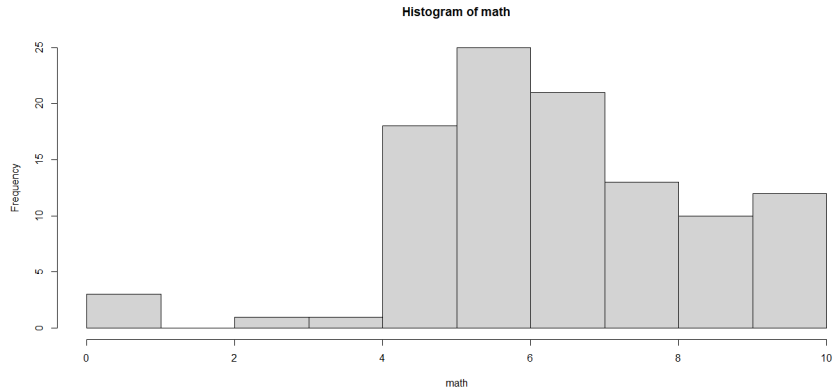
Κάνουμε το παρακάτω t test.

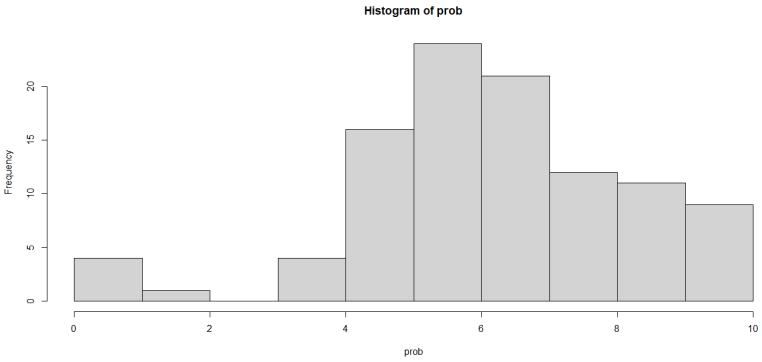
```
> t.test(prob[sex=="M"],prob[sex=="F"],alternative='great')

        Welch Two Sample t-test

data:  prob[sex == "M"] and prob[sex == "F"]
t = -1.284, df = 71.142, p-value = 0.8983
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -1.282117       Inf
sample estimates:
mean of x mean of y
 6.280303  6.838235
```

Παρατηρούμε ότι το p value = 0.8983 και είναι κατά πολύ μεγαλύτερο του a=0.05 ,οπότε δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση.

c) Ελέγχουμε τα ιστογράμματα για τους φοιτητές που έδωσαν Μαθηματικά 1 και Πιθανότητες.

2Η Σειρά Ασκήσεων





Παρατηρούμε ότι και στα δύο ιστογράμματα, υπάρχουν κάποια ατυπικά σημεία τα οποία δεν έχουν κάποια μεγάλη σημασία και επίσης υπάρχει μια ασυμμετρία , αλλά όχι σε τέτοιο βαθμό ώστε τα δεδομένα μας να θεωρηθούν ακατάλληλα.

Έστω μ1 = ο μέσος βαθμός στα Μαθηματικά 1 των φοιτητών που έχουν πάρει ή θα έπαιρναν το μάθημα «Στατιστική στην Πληροφορική» και μ2 = ο μέσος βαθμός του αντίστοιχου πληθυσμού στις Πιθανότητες.

Μηδενική Υπόθεση H0: μ1 -μ2 = 0
Εναλλακτική Υπόθεση Ηα : μ1-μ2 != 0
Κάνουμε t test και παίρνουμε ότι:

```
> #H0: μ1 = μ2 , Hα: μ1 != μ2
> t.test(math-prob)

        One Sample t-test

data:  math - prob
t = 0.97077, df = 96, p-value = 0.3341
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.2100256  0.6120874
sample estimates:
mean of x
0.2010309
```

Βλέπουμε ότι το p value=0.3341, το οποίο σίγουρα δεν είναι ασήμαντο, οπότε η μηδενική υπόθεση δεν απορρίπτεται.

Σημείωση: Το t test έγινε σε ένα δείγμα ,διότι ο βαθμός ενός φοιτητή στα Μαθηματικά 1 μπορεί να έχει επιρροή στον βαθμό των Πιθανοτήτων του.