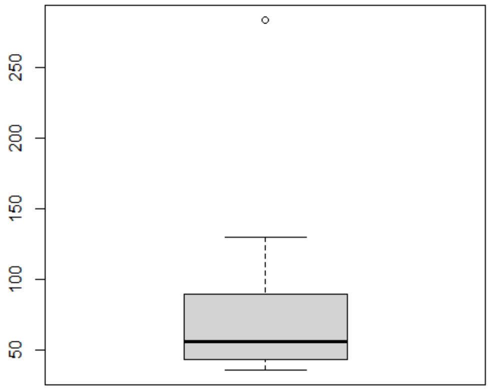
Χρήστος Πατρινόπουλος
3200150
Αλβιόνα Μάντσο
3200098

Στατιστική στην Πληροφορική

1. a) Αρχικά παρατηρούμε ότι δεν είναι γνωστή η τυπική απόκλιση του πληθυσμού. Οπότε οι γνωστές μέθοδοι συμπερασματολογίας απαιτούν για να είναι ακριβείς τα εξής:
   - Απλή τυχαία δειγματοληψία (SRS)
   - Μέγεθος δείγματος (n): 20 >= 15
   - Ισχύει ότι n=20 < 40 άρα ο πληθυσμός δεν πρέπει να έχει πολύ ασύμμετρη κατανομή. Δεν γνωρίζουμε την κατανομή του πληθυσμού, αλλά του δείγματος δεν φαίνεται να είναι συμμετρική:

   Οπότε θα πρέπει να είμαστε κάπως επιφυλακτικοί με τα αποτελέσματα.
   - **Υπάρχει ένα ατυπικό σημείο (τιμή 284 ms).**

   b)

$$\boxed{\overline{x} \pm t * \frac{s}{\sqrt{n}}}$$

Αφαιρώντας το ατυπικό σημείο έχουμε τα εξής:

$\overline{x}$ = { `mean(sample)` } = **66.52632**

s = { `sd(sample)` } = **27.53456**

t*= { `-qt(0.025, df=19-1)` } = **2.100922**

{ `mean(x) +c(-1, 1)* t*sd(x)/sqrt(19)` }

Διάστημα εμπιστοσύνης: **[53. 25508, 79.79755]** για επίπεδο εμπιστοσύνης 95%.

(αν αντίθετα είχαμε συμπεριλάβει το ατυπικό σημείο, θα παίρναμε ένα αρκετά μεγαλύτερο διάστημα, δλδ μια συντηρητικότερη προσέγγιση: $[51.41365,\ 103.38635]$).

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

   Με βάση τα γνωστά για την τυπική απόκλιση του δειγματικού μέσου δίνεται από τον τύπο $\sigma/\sqrt{n}$ οπότε θα έπρεπε να είναι $12/\sqrt{20}$. (Αν ο πληθυσμός είναι κανονικά κατανεμημένος ή το δείγμα είναι αρκούντως μεγάλο ώστε ο δειγματικός μέσος ακολουθεί την $N(\mu, \frac{\sigma}{\sqrt{n}})$).

---

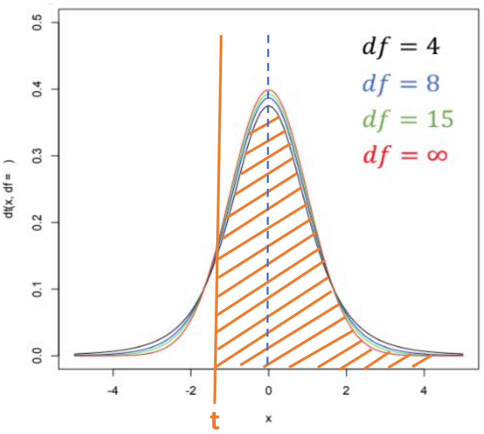
b) «*Ένας ερευνητής χρησιμοποιεί σε έναν έλεγχο σημαντικότητας τη μηδενική υπόθεση H_0 : $\bar{x}$= 10.»*

Οι υποθέσεις που διατυπώνουμε πρέπει να αφορούν παραμέτρους του πληθυσμού και όχι εκτιμητές. Θα ήταν σωστή εάν αναφερόταν στη μέση τιμή μ.

---

c) *«Σε μια στατιστική έρευνα με $\bar{x}$= 45 απορρίπτεται η μηδενική υπόθεση H_0 : μ = 54 όταν η εναλλακτική είναι H_a: μ > 54.»*

Έστω ότι η τυπική απόκλιση δεν ήταν γνωστή και χρησιμοποιήθηκε t-έλεγχος (εντελώς ανάλογη θα ήταν η περίπτωση που χρησιμοποιήθηκε z-έλεγχος).



$$t = \frac{\bar{x}-54}{s/\sqrt{n}} = \frac{45-54}{s/\sqrt{n}} = \frac{-9}{s/\sqrt{n}} < 0$$

$$\text{p value } = P(t \geq \frac{-9}{s/\sqrt{n}}) = 1 - F\left(\frac{-9}{\frac{s}{\sqrt{n}}}\right) > 0.5$$

Οπότε με ένα τόσο μεγάλο p-value δεν θα έπρεπε να απορριφθεί η μηδενική υπόθεση.

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

Το p-value είναι πολύ μεγάλο οπότε θα ήταν παράλογο να απορριφθεί η μηδενική υπόθεση.

3. a) - H_a: $\mu > \mu_0$

$$p \text{ value} = 1 - \Phi(z) = 1 - \Phi(1.34) =$$
$$= 1 - 0.9098773 = 0.09012267$$

b) - H_a: $\mu < \mu_0$

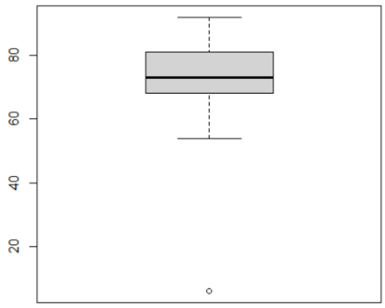
$$p \text{ value} = \Phi(z) = \Phi(1.34) = 0.9098773$$

c) - H_a: $\mu \neq \mu_0$

$$p \text{ value} = 2 * \Phi(-|z|) = 2 * \Phi(-1.34) =$$
$$= 2 * 0.09012267 = 0.1802453$$

4. a) Με επίπεδο σημαντικότητας α =1 - C=1 - 0.95 = 0.05 η μηδενική υπόθεση θα απορριπτόταν (αφού p-value=0.04 < α=0.05). Αφού έχουμε δίπλευρο έλεγχο μπορούμε να συμπεράνουμε ότι το μ_0=30 <u>δεν</u> περιέχεται στο διάστημα εμπιστοσύνης C=0.95=95%

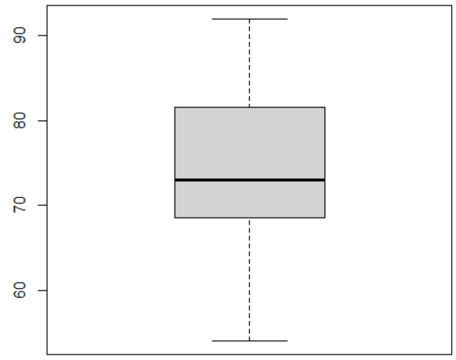
b) Με επίπεδο σημαντικότητας α =1 - C=1– 0.90 = 0.10 η μηδενική υπόθεση θα απορριπτόταν (αφού p-value=0.04 < α=0.10). Αφού έχουμε δίπλευρο έλεγχο μπορούμε να συμπεράνουμε ότι το μ_0=30 <u>δεν</u> περιέχεται στο διάστημα εμπιστοσύνης C=0.90=90%

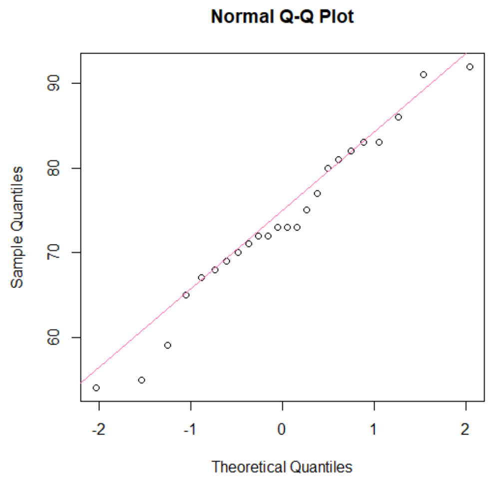


5. Αρχικά παρατηρούμε ότι υπάρχει ένα ατυπικό σημείο (outlier) στο δείγμα,

καθώς το άτομο με α/α 14 έχει βάρος 6. Προφανώς πρόκειται για σφάλμα, οπότε επιλέγουμε να αφαιρέσουμε το στοιχείο αυτό για να προχωρήσουμε τους υπολογισμούς.
Πλέον μέγεθος δείγματος: n = 25-1 = 24.

a)





Για τα δεδομένα μας παρατηρούμε τα εξής:

- Προέρχονται από απλή τυχαία δειγματοληψία (SRS)
- Μέγεθος δείγματος (n): 24 >= 15
- Ισχύει ότι n=24 < 40 άρα ο πληθυσμός πρέπει να έχει κανονική κατανομή ή τουλάχιστον να μην είναι ασύμμετρα κατανεμημένος. Από τα διαγράμματα παραπάνω επιβεβαιώνεται το ζητούμενο αυτό.

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

$\bar{x}$ = {`mean(sample)`} = **73.79167**

s= {`sd(sample)`} = **9.978146**

t*= {`-qt(0.025, df=24-1`} = **2.068658**

```
{ mean(x) +c(-1, 1)* t*sd(x)/sqrt(24) }
```

Διάστημα εμπιστοσύνης: **[69.57826, 78.00507]** για επίπεδο εμπιστοσύνης 95%.

b)

Έχουμε δύο ανεξάρτητα απλά τυχαία δείγματα: ένα από κάθε πληθυσμό (ανδρών και γυναικών). Για τα δεδομένα μας παρατηρούμε τα εξής:

- Προέρχονται από απλή τυχαία δειγματοληψία (SRS)
- Συνολικό πλήθος (n): $n_{male}+n_{female}$ = 13 + 11 = 24>15
- Ισχύει ότι n=24 < 40 άρα ο πληθυσμός πρέπει να έχει κανονική κατανομή ή τουλάχιστον να μην είναι ασύμμετρα κατανεμημένος. Αυτό ισχύει γενικά για «πληθυσμούς» βαρών (όπως και υψών κλπ...)

$$\boxed{(\overline{\boldsymbol{x1}}-\overline{\boldsymbol{x2}}) \pm \boldsymbol{t} * \sqrt{\frac{\boldsymbol{s1}^2}{\boldsymbol{n1}}+\frac{\boldsymbol{s2}^2}{\boldsymbol{n2}}}}$$

$n_1 = 13$

$n_2 = 11$

$\overline{x1}$ = {mean(sample_M)} = 78.69231
$\overline{x2}$ = {mean(sample_F)} = 68

s_1 = {sd(sample_M)} = 7.598077

s_2 = {sd(sample_F)} = 9.570789

t^* = {-qt(0.1, df=min(n1-1, n2-1))} = 1.372184

{ x1_bar-x2_bar + c(-1, 1)*t*sqrt(s1^2/n1 + s2^2/n2) }

Διάστημα εμπιστοσύνης: **[5.789155, 15.595460]** για επίπεδο εμπιστοσύνης 80%.

c)

Θεωρούμε δύο ανεξάρτητα απλά τυχαία δείγματα: ένα από κάθε πληθυσμό (καπνιστών και μη καπνιστών).

$n_{smokers} = 10$

$n_{non\_smokers} = 14$

$$H_0: \mu_{smokers} = \mu_{non\_smokers} \Rightarrow \mu_{smokers} - \mu_{non\_smokers} = 0$$
$$H_a: \mu_{smokers} \neq \mu_{non\_smokers} \Rightarrow \mu_{smokers} - \mu_{non\_smokers} \neq 0$$

```
> t.test(smokers, non_smokers)

        Welch Two Sample t-test

data:  smokers and non_smokers
t = 1.2597, df = 19.281, p-value = 0.2228
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.403049 13.717334
sample estimates:
mean of x mean of y
 76.80000  71.64286
```

<u>Παρατήρηση:</u>

- Για την κατανομή t χρησιμοποιούμε (έμμεσα, μέσω της t.test) το df που παράγει το λογισμικό αντί να υπολογίσουμε το df=min{$n_{smokers}$, $n_{non\_smokers}$} που θα ήταν λιγότερο ακριβές.

Επειδή ωστόσο το φύλο επηρεάζει το βάρος, είναι ορθότερο να εφαρμόσουμε τον έλεγχο στους άντρες και τις γυναίκες ξεχωριστά (με υποπληθυσμούς: καπνιστές και μη καπνιστές).

$n_{male\_smokers} = 6$

$n_{male\_non\_smokers} = 7$

(αρκετά μικρά)

$$H_0: \mu_{male\_smokers} - \mu_{male\_non\_smokers} = 0$$
$$H_a: \mu_{male\_smokers} - \mu_{male\_non\_smokers} \neq 0$$

```
> t.test(m_smokers, m_non_smokers)

        Welch Two Sample t-test

data:  m_smokers and m_non_smokers
t = 1.4566, df = 10.987, p-value = 0.1732
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.982106 14.648773
sample estimates:
mean of x mean of y
 81.83333  76.00000
```

$n_{fem\_smokers} = 4$

$n_{fem\_non\_smokers} = 7$

(αρκετά μικρά)

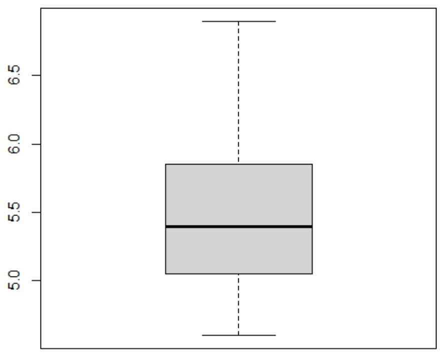
$$H_0: \mu_{female\_smokers} - \mu_{female\_non\_smokers} = 0$$
$$H_a: \mu_{female\_smokers} - \mu_{female\_non\_smokers} \neq 0$$

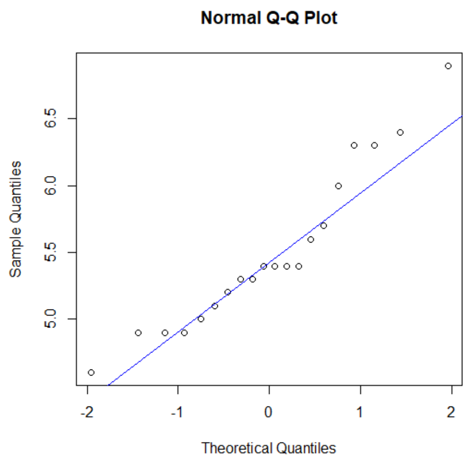
```
> t.test(f_smokers, f_non_smokers)

        Welch Two Sample t-test

data:  f_smokers and f_non_smokers
t = 0.3155, df = 6.5609, p-value = 0.7622
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.95998  16.88855
sample estimates:
mean of x mean of y
 69.25000  67.28571
```

<u>Συμπέρασμα:</u> Τόσο στην περίπτωση των ανδρών όσο και των γυναικών, (αλλά και στον πρώτο κοινό τους έλεγχο) το p-value είναι αρκετά μεγάλο ώστε να μην απορρίψουμε τη μηδενική υπόθεση. Άρα το κάπνισμα δεν πρέπει να επηρεάζει το βάρος.

6. a) Αρχικά παρατηρούμε ότι δεν είναι γνωστή η τυπική απόκλιση του πληθυσμού (όπου ο πληθυσμός είναι οι μετρήσεις). Οπότε οι γνωστές μέθοδοι συμπερασματολογίας απαιτούν για να είναι ακριβείς τα εξής:

- Απλή τυχαία δειγματοληψία (SRS)
- Μέγεθος δείγματος (n): 20 >= 15
- Ισχύει ότι n=20 < 40 άρα ο πληθυσμός δεν πρέπει να έχει πολύ ασύμμετρη κατανομή. Η κατανομή του δείγματος φαίνεται να είναι σχετικά συμμετρική:





b) $\bar{x}$ = {`mean(sample)`} = 5.5

s = {`sd(sample)`} = 0.6008766

c) t^*= {`-qt(0.025, df=20-1)`} = 2.093024

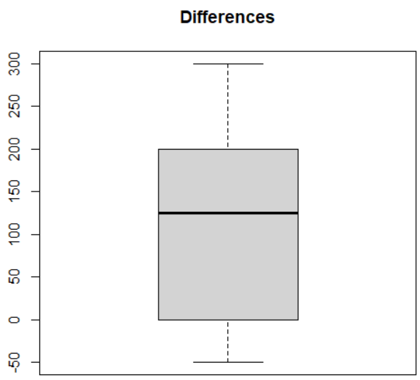
```
{ mean(x) +c(-1, 1)* t*sd(x)/sqrt(20) }
```

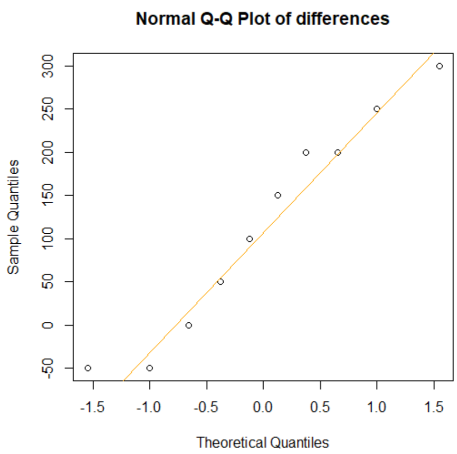
Διάστημα εμπιστοσύνης: [5.218781, 5.781219] για επίπεδο εμπιστοσύνης 95%.

7. Βρισκόμαστε στην περίπτωση ενός συγκριτικού τυχαιοποιημένου πειράματος με ταίριασμα σε ζεύγη όπου οι παράμετροι που αφορούν το αυτοκίνητο και τις ζημιές είναι κοινές σε κάθε ζεύγος (γιατί πρόκειται για το ίδιο αυτοκίνητο) και το ζεύγος διαφέρει μόνο στην εκτίμηση των ζημιών (συνεργείο ή εμπειρογνώμονας). Θεωρούμε ότι οι διαφορές ανάμεσα στα «μέλη» κάθε ζεύγους είναι απλό τυχαίο δείγμα (μεγέθους n=10).

Για να είναι ακριβής ο έλεγχος σημαντικότητας θα πρέπει να ισχύουν τα εξής:

- Απλή τυχαία δειγματοληψία (SRS)
- Μέγεθος δείγματος (n): 10 < 15,
- Ισχύει ότι n=10 < 40 άρα ο πληθυσμός δεν πρέπει να έχει πολύ ασύμμετρη κατανομή. Η κατανομή του δείγματος φαίνεται να είναι σχετικά συμμετρική:





Θα πρέπει να είμαστε κάπως επιφυλακτικοί με τα αποτελέσματα.

<u>*Σημείωση:*</u>

- s: Αφορά το συνεργείο
- e: Αφορά τον εμπειρογνώμονα

H_0: $\mu_s \leq \mu_e \Rightarrow \mu_s - \mu_e \leq 0$ «το συνεργείο <u>δεν</u> υπερεκτιμά τη ζημιά»

H_a: $\mu_s > \mu_e \Rightarrow \mu_s - \mu_e > 0$ «το συνεργείο υπερεκτιμά τη ζημιά»

$\overline{xs}$ = {mean(sample_S)} = 1220

$\overline{xe}$ = {mean(sample_E)} = 1105

s = {sd(diffs)} = 124.8332

t = {(xs_bar-xe_bar)/(s/sqrt(10))} = 2.9132

p-value = {1-pt(t, df=10-1)} = 0.008610911

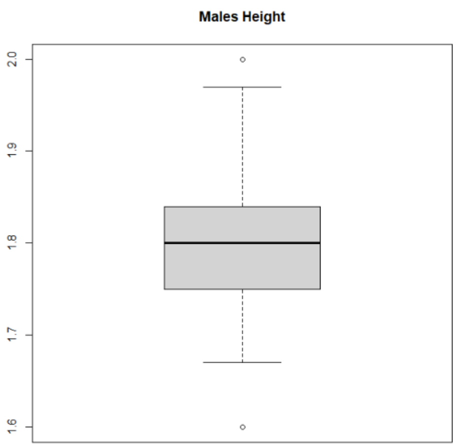
```
> t.test(diffs, alternative='greater')

        One Sample t-test

data:  diffs
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
```

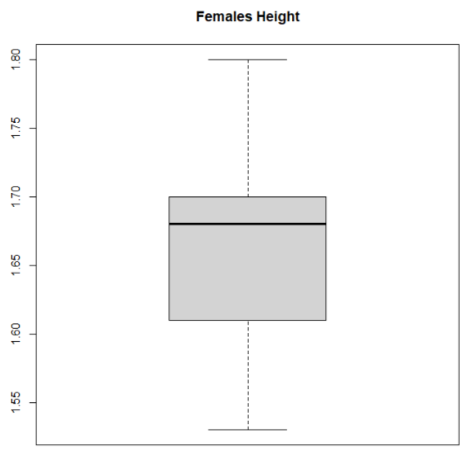
Το p-value είναι αρκετά μικρό ώστε μπορούμε να απορρίψουμε τη μηδενική υπόθεση (μ_s - $\mu_e \leq 0$). Συνεπώς τα δεδομένα είναι στατιστικά σημαντικά για την εναλλακτική υπόθεση (μ_s - $\mu_e > 0$). Οπότε επιβεβαιώνεται η υποψία ότι το συνεργείο τείνει να υπερεκτιμά τις ζημιές.

8. a) Έχουμε δύο ανεξάρτητα απλά τυχαία δείγματα: ένα από κάθε πληθυσμό (αγοριών φοιτητών και κοριτσιών φοιτητριών).

Για ακριβή αποτελέσματα χρειαζόμαστε τα εξής:

- Απλή τυχαία δειγματοληψία (SRS)
- Συνολικό πλήθος (n): $n_{m\_height}+n_{f\_height}$ = 80 + 37 = 117 αρκετά μεγάλο
- **Υπάρχουν ατυπικά σημεία στα ύψη των ανδρών**





Ωστόσο τα ατυπικά αυτά σημεία δεν επηρεάζουν τους υπολογισμούς.

Συμπεριλαμβανομένων των ατυπικών σημείων έχουμε

```
m <- height[gender == 'M']
f <- height[gender == 'F']
```

```
> t.test(m, f)

	Welch Two Sample t-test

data:  m and f
t = 9.3536, df = 71.715, p-value = 4.774e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1004419 0.1548554
sample estimates:
mean of x mean of y 
 1.799000  1.671351 
```

(ενώ <u>μη</u> συμπεριλαμβανομένων των ατυπικών σημείων παίρνουμε παρόμοιο διάστημα εμπιστοσύνης: $[0.1011276, 0.1541184]$)

<u>Παρατηρήσεις:</u>

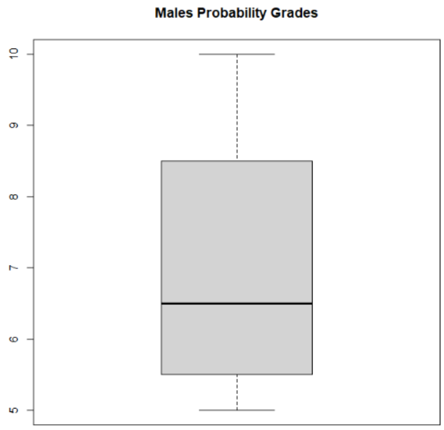
- Η συνάρτηση t.test της R εξαλείφει τα NA.
- Για την κατανομή t χρησιμοποιούμε (έμμεσα, μέσω της t.test) το df που παράγει το λογισμικό αντί να υπολογίσουμε το ζητούμενο διάστημα με df=min{n_{m_height}, n_{f_height}} καθώς το δεύτερο θα μας οδηγούσε σε ένα πιο συντηρητικό διάστημα. Εκμεταλλευόμαστε το γεγονός ότι η συνάρτηση t.test παρουσιάζει το διάστημα εμπιστοσύνης για επίπεδο 95% όπως ήταν εδώ το ζητούμενο.

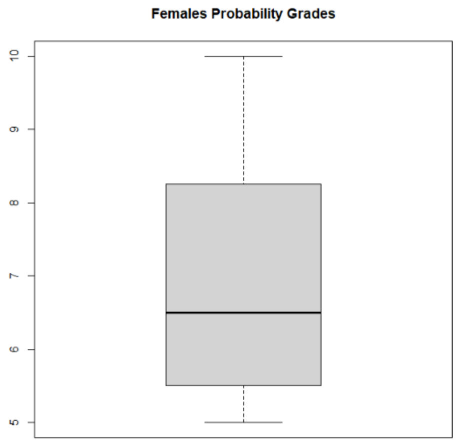
Διάστημα εμπιστοσύνης: $\mathbf{[0.1004419, 0.1548554]}$ για επίπεδο εμπιστοσύνης 95%.

b) Έχουμε δύο ανεξάρτητα απλά τυχαία δείγματα: ένα από κάθε πληθυσμό (αγοριών φοιτητών και κοριτσιών φοιτητριών).

Για ακριβή αποτελέσματα χρειαζόμαστε τα εξής:

- Απλή τυχαία δειγματοληψία (SRS)
- Συνολικό πλήθος (n): $n_{m\_prob}+n_{f\_prob}$ = 64 + 31 = 95 αρκετά μεγάλο





$$\begin{cases} H_0: \mu_m \leq \mu_f \Rightarrow \mu_m - \mu_f \leq 0 \\ H_a: \mu_m > \mu_f \Rightarrow \mu_m - \mu_f > 0 \end{cases}$$

H_0: «οι άνδρες φοιτητές πληροφορικής <u>δεν επιτυγχάνουν</u> μεγαλύτερο βαθμό στο μάθημα των πιθανοτήτων από τον αντίστοιχο πληθυσμό των γυναικών»

H_a: «οι άνδρες φοιτητές πληροφορικής <u>επιτυγχάνουν</u> μεγαλύτερο βαθμό στο μάθημα των πιθανοτήτων από τον αντίστοιχο πληθυσμό των γυναικών»

Επειδή στο ερωτηματολόγιο δεν διευκρινίζεται το αν κάποιος που δεν έχει περάσει το μάθημα πρέπει να δηλώσει τον βαθμό του (μικρότερος του 5) ή να μην απαντήσει, θεωρούμε ορθότερο να λάβουμε υπόψιν μας μόνο τους βαθμούς που είναι υψηλότεροι της βάσης (5).

```
m <- prob[gender == 'M' & prob>=5]
m <- m[!is.na(m)]
f <- prob[gender == 'F' & prob>=5]
f <- f[!is.na(f)]
```

```
> t.test(m,f,alternative="greater")

        Welch Two Sample t-test

data:  m and f
t = 0.019055, df = 60.737, p-value = 0.4924
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.6333336        Inf
sample estimates:
mean of x mean of y 
 7.023438  7.016129 
```

<u>Συμπέρασμα:</u> Παρατηρούμε ότι p-value=0.4924 ~ 49% και συνεπώς αρκετά μεγαλύτερο από το επίπεδο σημαντικότητας 5%. Επομένως δεν απορρίπτουμε τη μηδενική υπόθεση ότι οι άνδρες φοιτητές πληροφορικής επιτυγχάνουν μικρότερο ή ίσο μέσο βαθμό στο μάθημα των πιθανοτήτων από τον αντίστοιχο πληθυσμό των γυναικών. Τα δεδομένα δεν είναι στατιστικά σημαντικά για την εναλλακτική υπόθεση.

c) Τα δείγματα είναι εξαρτημένα επειδή ανά ζεύγος βαθμών ο βαθμός του κάθε μαθήματος εξαρτάται από τις δυνατότητες του φοιτητή και πιθανώς ο βαθμός των πιθανοτήτων να εξαρτάται και από τις ικανότητές του στα μαθηματικά 1. Βρισκόμαστε στην περίπτωση ενός συγκριτικού τυχαιοποιημένου πειράματος με ταίριασμα σε ζεύγη όπου οι παράμετροι που αφορούν τον φοιτητή (ικανότητες…κλπ) είναι κοινές σε κάθε ζεύγος (γιατί πρόκειται για τον ίδιο φοιτητή). Θεωρούμε ότι οι διαφορές ανάμεσα στα «μέλη» κάθε ζεύγους είναι απλό τυχαίο δείγμα (μεγέθους n=106).

Για ακριβή αποτελέσματα χρειαζόμαστε τα εξής:

- Απλή τυχαία δειγματοληψία (SRS) ✓
- Συνολικό πλήθος (n): 106, αρκετά μεγάλο ✓

$$\begin{cases} H_0: \mu_{\text{prob}} = \mu_{\text{math}} \Rightarrow \mu_{\text{prob}} - \mu_{\text{math}} = 0 \\ H_a: \mu_{\text{prob}} \neq \mu_{\text{math}} \Rightarrow \mu_{\text{prob}} - \mu_{\text{math}} \neq 0 \end{cases}$$

H_0: «Ο μέσος βαθμός στα Μαθηματικά 1 <u>δεν διαφέρει</u> από το μέσο βαθμό στις Πιθανότητες»

H_a: «Ο μέσος βαθμός στα Μαθηματικά 1 <u>διαφέρει</u> από το μέσο βαθμό στις Πιθανότητες»

n = `length(diffs)` = **106**

s = `sd(diffs)` = **2.079226**

t = `(mean(diffs))/(s/sqrt(n))`

= **-1.448128**

p-value = `2*pt(t, df=n-1)` = **0.1505607**

```
> t.test(prob-math)

        One Sample t-test

data:  prob - math
t = -1.4481, df = 105, p-value = 0.1506
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.6928869  0.1079812
sample estimates:
 mean of x
-0.2924528
```

<u>Συμπέρασμα:</u> Παρατηρούμε ότι p-value=0.1506 ~ 15% και συνεπώς όχι πολύ μικρό. Επομένως δεν θα απορρίψουμε τη μηδενική υπόθεση ότι ο μέσος βαθμός στα Μαθηματικά 1 ισούται με το μέσο βαθμό στις Πιθανότητες. Τα δεδομένα θεωρούμε ότι δεν είναι στατιστικά σημαντικά για την εναλλακτική υπόθεση.