ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
DEPARTMENT OF INFORMATICS

ΣΤΑΤΙΣΤΙΚΗ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

Στατιστική Συμπερασματολογία

Εργασία 2

Μαρία Σχοινάκη

Σοφία Παπαϊωάννου

Άσκηση 1

a) Καταλληλόλητα δεδομένων

Τα δεδομένα είναι κατάλληλα για να χρησιμοποιήσουμε τις μεθόδους συμπερασματολογίας που γνωρίζουμε, καθώς:

1. Πρόκειται για τυχαία δειγματοληψία από πληθυσμό (*ημερήσιες αιτήσεις από βάση δεδομένων*).
2. Το δείγμα έχει ικανοποιητικό μέγεθος, με **n = 20** παρατηρήσεις (n ≥ 15).
3. Υπάρχει μία ακραία τιμή (*284 ms*), που μπορεί να υποδεικνύει ότι η κατανομή δεν είναι κανονική. Ωστόσο, λόγω του μεγέθους του δείγματος, μπορούμε να χρησιμοποιήσουμε την κατανομή t, καθώς είναι αρκετά ανθεκτική ακόμα και σε αποκλίσεις από την κανονικότητα.

b) Διάστημα εμπιστοσύνης

Ο υπολογισμός του διαστήματος εμπιστοσύνης περιλαμβάνει:

1. Υπολογισμό του μέσου όρου ($\bar{x}$) και της τυπικής απόκλισης (s) των δεδομένων.
2. Χρήση του τύπου διαστήματος εμπιστοσύνης:

$$\Delta.\text{E.} = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

όπου Δ.Ε, Διάστημα Εμπιστοσύνης (C).

- Επίπεδο εμπιστοσύνης: 95%
- Βαθμός ελευθερίας: n−1 = 19

Χρησιμοποιώντας την R, υπολογίσαμε το διάστημα εμπιστοσύνης **[51.41365, 103.38635]**. Αυτό δείχνει ότι η μέση τιμή του χρόνου διεκπεραίωσης βρίσκεται με 95% εμπιστοσύνη εντός αυτού του διαστήματος.

```
> data <- c(82, 55, 58, 94, 86, 45, 42, 36, 41, 130, 284, 96, 39, 107, 52, 54, 45, 81, 83, 38)
>
> n <- length(data)
> n
[1] 20
> xbar <- mean(data)
> xbar
[1] 77.4
> sd <- sd(data)
> sd
[1] 55.52467
> tstar <- -qt(0.025, df=n-1)
> tstar
[1] 2.093024
> xbar + c(-1, 1) * tstar * sd / sqrt(n)
[1]  51.41365 103.38635
>
```

Άσκηση 2

a) **Το λάθος:** Στον τύπο για την τυπική απόκλιση του δειγματικού μέσου ($\sigma_{\bar{x}}$), έχει χρησιμοποιηθεί λάθος τύπος: $\sigma_{\bar{x}} = \sigma / n$ **αντί** για $\sigma / \sqrt{n}$.

Η σωστή προσέγγιση: Η τυπική απόκλιση του δειγματικού μέσου υπολογίζεται ως:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Στην προκειμένη περίπτωση:

$$\sigma_{\bar{x}} = \frac{12}{\sqrt{20}} \quad \text{και όχι} \quad \frac{12}{20}.$$

b) **Το λάθος:** Ο ερευνητής διατυπώνει τη μηδενική υπόθεση με βάση τον δειγματικό μέσο ($\bar{x}$), δηλαδή $H_0 : \bar{x} = 10$.

Η σωστή προσέγγιση: Στους ελέγχους σημαντικότητας, οι υποθέσεις αναφέρονται πάντα στον **πληθυσμιακό μέσο όρο** μ. Η μηδενική υπόθεση θα έπρεπε να είναι:

$$H_0 : \mu = 10$$

c) **Το λάθος:** Η εναλλακτική υπόθεση είναι $H_a : \mu > 54$, αλλά η δειγματική μέση τιμή είναι $\bar{x} = 45$, που είναι μικρότερη από 54. Παρ' όλα αυτά, απορρίπτεται η μηδενική υπόθεση $H_0 : \mu = 54$.

Η σωστή προσέγγιση: Για να απορρίψουμε τη H_0 υπέρ της H_a, η δειγματική μέση τιμή πρέπει να ανήκει στο εύρος που καθορίζει η εναλλακτική υπόθεση, δηλαδή πρέπει $\bar{x} > 54$.

d) **Το λάθος:** Η μηδενική υπόθεση απορρίπτεται, αν και το p-value είναι 0.52, δηλαδή πολύ μεγαλύτερο από τα συνηθισμένα επίπεδα σημαντικότητας (π.χ., 0.05 ή 0.01).

Η σωστή προσέγγιση: Για να απορρίψουμε τη μηδενική υπόθεση, το p-value πρέπει να είναι μικρότερο από το επίπεδο σημαντικότητας α (συνήθως 0.05 ή 0.01).

Άσκηση 3

Κάνουμε έλεγχο σημαντικότητας με μηδενική υπόθεση H_a : μ = μ0 γνωρίζοντας ότι η τιμή του στατιστικού ελέγχου z είναι ίση με 1.34.

```
> z <- 1.34
```

a) Το pvalue για την εναλλακτική υπόθεση **H_a : μ > μ0** είναι:

```
> pvalue <- 1 - pnorm(z)
> pvalue
[1] 0.09012267
```

b) Το pvalue για την εναλλακτική υπόθεση **H_a : μ < μ0** είναι:

```
> pvalue <- pnorm(z)
> pvalue
[1] 0.9098773
```

c) Το pvalue για την εναλλακτική υπόθεση **H_a : μ ≠ μ0** είναι:

```
> pvalue <- 2 * (1 - pnorm(abs(z)))
> pvalue
[1] 0.1802453
```

Άσκηση 4

a)

- Το p-value = 0.04 είναι μικρότερο από το επίπεδο σημαντικότητας α = 0.05.
- Αυτό σημαίνει ότι η μηδενική υπόθεση H0: μ = 30 **απορρίπτεται** με επίπεδο σημαντικότητας 5%.
- Στη στατιστική, αν απορρίπτουμε την H0 με βάση το p-value, η τιμή της H0 (δηλαδή, μ = 30) δεν περιλαμβάνεται στο αντίστοιχο 95% διάστημα εμπιστοσύνης.

Άρα, η τιμή 30 **δεν περιλαμβάνεται** στο 95% διάστημα εμπιστοσύνης, γιατί η απόρριψη της H0 συνεπάγεται ότι το μ = 30 είναι εξαιρετικά απίθανο με βάση τα δεδομένα.

b)

- Το p-value = 0.04 είναι μικρότερο από το επίπεδο σημαντικότητας α = 0.10.
- Επομένως, απορρίπτουμε επίσης τη μηδενική υπόθεση H0: μ = 30 στο 10% επίπεδο σημαντικότητας.
- Εξ ορισμού, ένα 90% διάστημα εμπιστοσύνης είναι μεγαλύτερο από ένα 95% διάστημα εμπιστοσύνης. Ωστόσο, αν το p-value = 0.04 υποδηλώνει απόρριψη της H0, η τιμή μ = 30 **δεν μπορεί να περιλαμβάνεται ούτε στο 90% διάστημα εμπιστοσύνης**.

Η τιμή 30 **δεν περιλαμβάνεται** ούτε στο 90% διάστημα εμπιστοσύνης. Παρόλο που το διάστημα είναι μεγαλύτερο, η απόρριψη της H0 δείχνει ότι το μ = 30 είναι εξαιρετικά απίθανο.

Άσκηση 5

Παρατηρούμε ότι υπάρχει τιμή 6, ενώ δεν γίνεται κάποιος ενήλικας να ζυγίζει 6 κιλά. Οπότε δεν λαμβάνουμε υπόψη αυτή την περίπτωση στους υπολογισμούς μας. Άρα **n = 24**.

a) Τα δεδομένα φαίνεται να είναι σχετικά συμμετρικά, και έχουμε n = 24, το οποίο είναι αρκετό για χρήση της κατανομής t.

Ο υπολογισμός του διαστήματος εμπιστοσύνης περιλαμβάνει:

1. Υπολογισμό του μέσου όρου ($\bar{x}$) και της τυπικής απόκλισης (s) των δεδομένων.
2. Χρήση του τύπου διαστήματος εμπιστοσύνης:

$$\Delta.\text{E.} = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

όπου Δ.Ε, Διάστημα Εμπιστοσύνης (C).

- Επίπεδο εμπιστοσύνης: 95%
- Βαθμός ελευθερίας: n−1 = 23

Περνώντας τα δεδομένα του βάρους στην μεταβλητή weight και εκτελώντας τις παρακάτω εντολές στην R παίρνουμε το διάστημα **[69.57826, 78.00507]** ως το διάστημα εμπιστοσύνης 95% για το μέσο βάρος των κατοίκων της Αθήνας.

Stemplot

5 | 459
6 | 5789
7 | 012233357
8 | 012336
9 | 12

```
> 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)
> n
[1] 24
> mean_weight <- mean(weight)
> mean_weight
[1] 73.79167
> sd_weight <- sd(weight)
> sd_weight
[1] 9.978146
> tstar_weight <- -qt(0.025, df=n-1)
> tstar_weight
[1] 2.068658
> C <- mean_weight + c(-1, 1) * tstar_weight * sd_weight / sqrt(n)
> C
[1] 69.57826 78.00507
```

b) Από τις 24 περιπτώσεις, οι περιπτώσεις που έχουν τιμή Α στην μεταβλητή **"ΦΥΛΟ"** είναι το ανδρικό τυχαίο δείγμα και οι υπόλοιπες το γυναικείο. Οπότε έχουμε n_m = 13 (*male*) και n_f = 11 (*female*).

Τα δεδομένα φαίνεται να είναι σχετικά συμμετρικά, άρα και πάλι μπορούμε να χρησιμοποιήσουμε την κατανομή t.

Stemplot (Male)
6 | 8
7 | 223357
8 | 0136
9 | 12

Stemplot (Female)
5 | 459
6 | 579
7 | 013
8 | 23

Χρησιμοποιώντας την **R** φορτώνουμε κατάλληλα τα δεδομένα στις μεταβλητές **weight_f** και **weight_m** αντίστοιχα και κάνουμε τους απαραίτητους υπολογισμούς.

```
> weight_m <- c(80, 81, 75, 83, 73, 77, 91, 92, 86, 73, 68, 72, 72)
> n_weight_m <- length(weight_m)
> n_weight_m
[1] 13
> mean_weight_m <- mean(weight_m)
> mean_weight_m
[1] 78.69231
> sd_weight_m <- sd(weight_m)
> sd_weight_m
[1] 7.598077

> weight_f <- c(71, 65, 67, 54, 55, 83, 73, 82, 69, 70, 59)
> n_weight_f <- length(weight_f)
> n_weight_f
[1] 11
> mean_weight_f <- mean(weight_f)
> mean_weight_f
[1] 68
> sd_weight_f <- sd(weight_f)
> sd_weight_f
[1] 9.570789
```

Για να βρούμε ένα διάστημα εμπιστοσύνης για την διαφορά του μέσου βάρους των δύο φύλων χρησιμοποιούμε τον ακόλουθο τύπο:

$$(\bar{x}_m - \bar{x}_f) \pm t^* \cdot \sqrt{\frac{s_m^2}{n_m} + \frac{s_f^2}{n_f}}$$

- Επίπεδο εμπιστοσύνης: 80%
- Βαθμός ελευθερίας: min(n_weight_m - 1, n_weight_f - 1) = 10

Χρησιμοποιώντας την R, υπολογίσαμε ως το διάστημα εμπιστοσύνης 80% για την διαφορά της μέσης τιμής του βάρους μεταξύ ανδρών και γυναικών (ενηλίκους κατοίκους Αθηνών) το **[5.789155, 15.595460].**

```
> tstar <- -qt(0.1, df=min(n_weight_m - 1, n_weight_f - 1))
> tstar
[1] 1.372184
> dif_c <- (mean_weight_m - mean_weight_f) + c(-1, 1) * tstar * sqrt(((sd_weight_m ^ 2) / n_weight_m)
+ ((sd_weight_f ^ 2) / n_weight_f))
> dif_c
[1]  5.789155 15.595460
```

c) Θα εξετάσουμε αν υπάρχει στατιστικά σημαντική διαφορά μεταξύ του μέσου βάρους των καπνιστών (μ1) και του μέσου βάρους των μη καπνιστών (μ2) στο δείγμα μας. Συνεπώς, θα πραγματοποιήσουμε δίπλευρο έλεγχο υποθέσεων με:

- Μηδενική υπόθεση: **H0: μ1 = μ2** , δηλαδή ότι τα μέσα βάρη των δύο ομάδων είναι ίσα.
- Εναλλακτική υπόθεση: **Ha: μ1 ≠ μ2** , δηλαδή ότι υπάρχει διαφορά στα μέσα βάρη.

Τα δεδομένα που έχουμε είναι κατάλληλα για να εξαγάγουμε στατιστικά συμπεράσματα, καθώς προέρχονται από απλή τυχαία δειγματοληψία, ο αριθμός των παρατηρήσεων είναι επαρκής ($n \geq 15$), και οι κατανομές των δεδομένων δείχνουν συμμετρία, όπως φαίνεται από τα stemplots. Χωρίζουμε το δείγμα σε δύο ομάδες, **καπνιστές** και **μη καπνιστές**, και τα αποθηκεύουμε στις μεταβλητές **weight_s** (*καπνιστές*) και **weight_ns** (*μη καπνιστές*), ώστε να προχωρήσουμε στους απαραίτητους υπολογισμούς για τον έλεγχο.

Stemplot (*Smoker*)

5 | 9
6 | 5
7 | 137
8 | 0236
9 | 2

Stemplot(*Non Smoker*)

5 | 45
6 | 789
7 | 022335
8 | 13
9 | 1

```
> weight_s <- c(80, 83, 71, 73, 65, 77, 92, 86, 82, 59)
> n_weight_s <- length(weight_s)
> n_weight_s
[1] 10
> mean_weight_s <- mean(weight_s)
> mean_weight_s
[1] 76.8
> sd_weight_s <- sd(weight_s)
> sd_weight_s
[1] 9.975526
```

```
> weight_ns <- c(81, 75, 67, 54, 55, 83, 91, 73, 69, 73, 70, 68, 72, 72)
> n_weight_ns <- length(weight_ns)
> n_weight_ns
[1] 14
> mean_weight_ns <- mean(weight_ns)
> mean_weight_ns
[1] 71.64286
> sd_weight_ns <- sd(weight_ns)
> sd_weight_ns
[1] 9.76341
```

Το **t-στατιστικό** υπολογίζεται με τον τύπο:

$$t = \frac{\bar{x}_s - \bar{x}_{ns}}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_{ns}^2}{n_{ns}}}}$$

```
> t <- (mean_weight_s - mean_weight_ns) / sqrt(((sd_weight_s ^ 2) / n_weight_s) + ((sd_weight_ns ^ 2)
/ n_weight_ns))
> t
[1] 1.259715
```

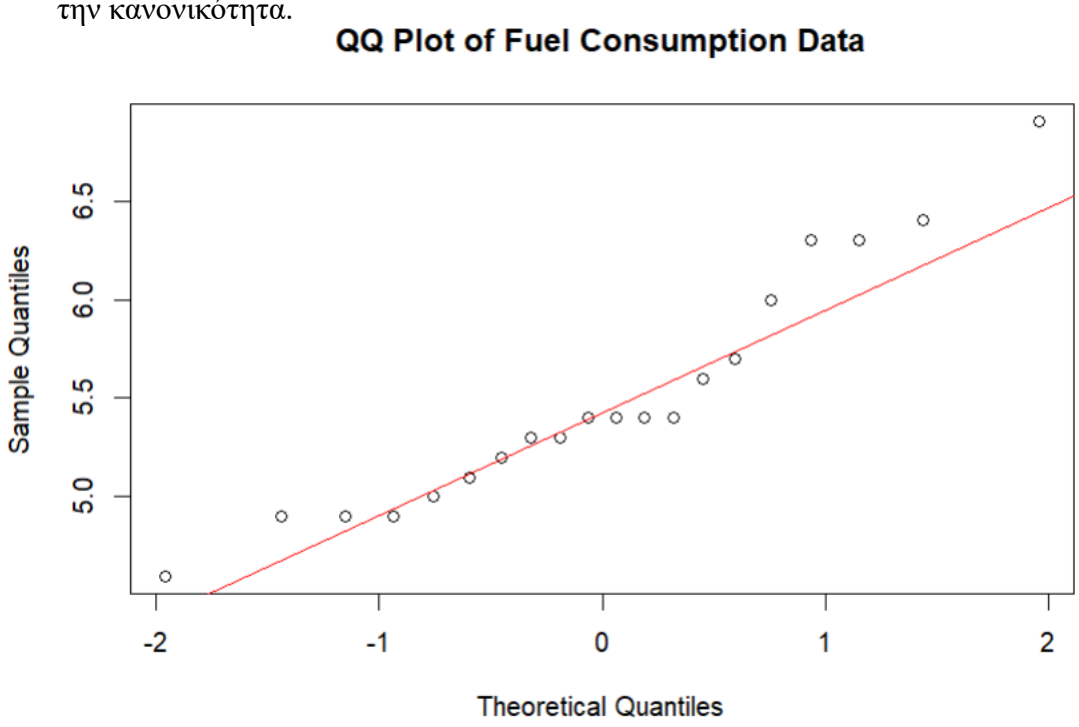
Ως αποτέλεσμα, ο υπολογισμός μας δίνει p-value = **0.2394573**. Αυτή η τιμή υποδεικνύει ότι δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση. Σημειώνεται ότι για τον υπολογισμό του p-value χρησιμοποιήθηκαν 9 βαθμοί ελευθερίας, που αντιστοιχούν στο μικρότερο από τα **n_weight_s − 1** και **n_weight_ns − 1**.

```
> p <- (1 - pt(t, df=min(n_weight_s - 1, n_weight_ns - 1))) * 2
> p
[1] 0.2394573
```

Άσκηση 6

a) Τα δεδομένα που παρέχονται είναι κατάλληλα για συμπερασματολογία επειδή:

1. Πρόκειται για τυχαίο δείγμα (SRS).
2. Ο αριθμός των παρατηρήσεων είναι επαρκής (n = 20, που είναι μεγαλύτερο ή ίσο με 15).
3. Αν και η κατανομή των δεδομένων δεν είναι απόλυτα κανονική, όπως φαίνεται από το normal quantile plot, το μέγεθος του δείγματος επιτρέπει τη χρήση μεθόδων που βασίζονται στην κατανομή t, καθώς αυτές είναι ανθεκτικές σε μέτριες αποκλίσεις από την κανονικότητα.



b) Εκτελούμε τις παρακάτω εντολές στην R.

```
> mean_fuel_cons <- mean(fuel_cons)
> mean_fuel_cons
[1] 5.5
> sd_fuel_cons <- sd(fuel_cons)
> sd_fuel_cons
[1] 0.6008766
```

c)

- Επίπεδο εμπιστοσύνης: 95%
- Βαθμός ελευθερίας: n_fuel_cons – 1 = 19

Εκτελώντας τις παρακάτω εντολές στην **R** παίρνουμε το διάστημα **[5.218781, 5.781219]** ως το διάστημα εμπιστοσύνης 95% για την μέση τιμή της απόδοσης του αυτοκινήτου.

```
> n_fuel_cons <- length(fuel_cons)
> n_fuel_cons
[1] 20
> tstar <- -qt(0.025, df=n_fuel_cons - 1)
> tstar
[1] 2.093024
> c <- mean_fuel_cons + c(-1, 1) * tstar * sd_fuel_cons / sqrt(n_fuel_cons)
> c
[1] 5.218781 5.781219
```

Η ακρίβεια του υπολογισμένου διαστήματος εμπιστοσύνης εξαρτάται από την υπόθεση κανονικότητας της κατανομής των δεδομένων. Αν και χρησιμοποιούμε την κατανομή ttt λόγω του περιορισμένου μεγέθους του δείγματος (n = 20), η υπόθεση της κανονικότητας είναι κρίσιμη για την εγκυρότητα του διαστήματος εμπιστοσύνης.

Σύμφωνα με την ανάλυση στο ερώτημα **α)**, τα δεδομένα παρουσιάζουν αποκλίσεις από την κανονικότητα (*όπως φάνηκε και στο normal quantile plot*). Παρόλα αυτά, το μέγεθος του δείγματος είναι επαρκές (*20 παρατηρήσεις*) για να χρησιμοποιηθεί η προσέγγιση της κατανομής t, καθώς η κατανομή t είναι σχετικά ανθεκτική σε μέτριες αποκλίσεις από την κανονικότητα.

Ωστόσο, πρέπει να σημειωθεί ότι σε περιπτώσεις όπου η κατανομή των δεδομένων αποκλίνει σημαντικά από την κανονικότητα, το αποτέλεσμα του διαστήματος εμπιστοσύνης ενδέχεται να μην είναι απόλυτα ακριβές. Επομένως, η ερμηνεία του 95% διαστήματος εμπιστοσύνης πρέπει να γίνεται με προσοχή, λαμβάνοντας υπόψη την ποιότητα της κατανομής των δεδομένων.

Άσκηση 7

Για να εξετάσουμε τη μηδενική υπόθεση ότι δεν υπάρχει υπερεκτίμηση των ζημιών από το συνεργείο, συγκρίνουμε τη μέση εκτίμηση ζημιών του συνεργείου με εκείνη του εμπειρογνώμονα. Αυτό σημαίνει ότι ελέγχουμε αν, κατά μέσο όρο, οι εκτιμήσεις ζημιών του συνεργείου είναι μεγαλύτερες από αυτές του εμπειρογνώμονα.

Παρατηρούμε ότι οι εκτιμήσεις των ζημιών για κάθε αυτοκίνητο είναι **αλληλένδετες**, καθώς όταν το συνεργείο εκτιμά μια υψηλή ζημιά, είναι πιθανό και ο εμπειρογνώμονας να κάνει μια υψηλή εκτίμηση. Επομένως, οι εκτιμήσεις δεν είναι ανεξάρτητες.

Για να λάβουμε υπόψη αυτή τη συσχέτιση, εργαζόμαστε με τη διαφορά των εκτιμήσεων για κάθε αυτοκίνητο. Τα δεδομένα που θα χρησιμοποιήσουμε είναι οι διαφορές (*Συνεργείο - Εμπειρογνώμονας*) για κάθε αυτοκίνητο. Αυτές οι διαφορές αποτελούν το νέο σύνολο δεδομένων, το οποίο θα χρησιμοποιηθεί για τον έλεγχο.

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

Το δείγμα αποτελείται από **10** αυτοκίνητα, και θεωρείται τυχαίο καθώς αφορά ανεξάρτητες μετρήσεις εκτιμήσεων ζημιών. Εξετάζοντας το παρακάτω stemplot των διαφορών (*Συνεργείο - Εμπειρογνώμονας*), παρατηρούμε ότι η κατανομή τους είναι σχετικά συμμετρική, γεγονός που υποδηλώνει ότι μπορεί να ανήκει σε μια κανονική κατανομή.

Ωστόσο, το μέγεθος του δείγματος (n = 10) είναι μικρό, κάτι που ενδέχεται να επηρεάσει την ακρίβεια των υπολογισμών. Συνεπώς, συνεχίζουμε την ανάλυση με κάποια επιφυλακτικότητα ως προς τη γενίκευση των αποτελεσμάτων.

Stemplot

-0 | 55
 0 | 05
 1 | 05
 2 | 005
 3 | 0

Για να ελέγξουμε τη μηδενική υπόθεση H0: μ = 0 με εναλλακτική υπόθεση Ha: μ > 0 όπου μ είναι η μέση τιμή των διαφορών, χρησιμοποιούμε τη μέθοδο ελέγχου σημαντικότητας. Η εναλλακτική υπόθεση Ha: μ > 0 εστιάζει μόνο στη θετική διαφορά, καθώς στόχος μας είναι να εξετάσουμε αν το συνεργείο υπερεκτιμά τις ζημιές.
Παρακάτω είναι οι εντολές στην **R** για τον υπολογισμό του στατιστικού ελέγχου **t**:

```
> dif <- c(100, 50, -50, 0, -50, 200, 250, 200, 150, 300)
> n <- length(dif)
> n
[1] 10
> mean <- mean(dif)
> mean
[1] 115
> sd <- sd(dif)
> sd
[1] 124.8332
> t <- mean / (sd / sqrt(n))
> t
[1] 2.913182
```

Τέλος, με την παραπάνω ανάλυση, υπολογίσαμε ότι το p-value είναι ίσο με 0.008610911. Αυτή η εξαιρετικά μικρή τιμή του p-value υποδεικνύει ότι απορρίπτουμε τη μηδενική υπόθεση H0 με υψηλή βεβαιότητα. Ακόμα και αν η κατανομή των διαφορών στον πληθυσμό δεν είναι απόλυτα κανονική, η απόκλιση της p-τιμής από τα συνηθισμένα επίπεδα σημαντικότητας (α=0.05 ή α=0.01) είναι τόσο μεγάλη που το αποτέλεσμα παραμένει ισχυρό.

Συνεπώς, τα δεδομένα υποστηρίζουν ότι το συνεργείο **υπερεκτιμά** συστηματικά τις ζημιές σε σχέση με τον εμπειρογνώμονα. Σημειώνεται ότι για τον υπολογισμό του p-value χρησιμοποιήθηκαν **n−1 = 9** βαθμοί ελευθερίας, καθώς το δείγμα περιλαμβάνει **10** παρατηρήσεις.

```
> p <- 1 -pt(t, df=n - 1)
> p
[1] 0.008610911
```

Άσκηση 8

Φορτώνουμε στην **R** τα δεδομένα απο το ερωτηματολόγιο του **2024**.

a) Για να υπολογίσουμε το διάστημα εμπιστοσύνης για τη διαφορά του μέσου ύψους μεταξύ ανδρών και γυναικών φοιτητών πληροφορικής, πρώτα χρειάζεται να διαχωρίσουμε τα δεδομένα για κάθε φύλο. Ειδικότερα, εξάγουμε τις μεταβλητές **male_height** και **female_height**, που αντιπροσωπεύουν το ύψος των ανδρών και των γυναικών αντίστοιχα. Αυτό μπορούμε να το επιτύχουμε χρησιμοποιώντας τις παρακάτω εντολές στην **R**:

```
> male_height <- data$height[data$gender == "M"]
> female_height <- data$height[data$gender == "F"]
```

Στην συνέχεια, κάνουμε τους απαραίτητους υπολογισμούς ώστε να βρούμε το διάστημα εμπιστοσύνης.

```
> n_male_height <- length(male_height)
> n_male_height
[1] 88
> mean_male_height <- mean(male_height)
> mean_male_height
[1] 1.795795
> sd_male_height <- sd(male_height)
> sd_male_height
[1] 0.06802166
```

```
> n_female_height <- length(female_height)
> n_female_height
[1] 49
> mean_female_height <- mean(female_height)
> mean_female_height
[1] 1.668776
> sd_female_height <- sd(female_height)
> sd_female_height
[1] 0.06392159
```

```
> tstar <- -qt(0.025, df=min(n_male_height - 1, n_female_height - 1))
> tstar
[1] 2.010635
```

Ως αποτέλεσμα των υπολογισμών με τον τύπο που χρησιμοποιήθηκε, προκύπτει ότι το 95% διάστημα εμπιστοσύνης για τη διαφορά του μέσου ύψους μεταξύ ανδρών και γυναικών φοιτητών πληροφορικής του ΟΠΑ είναι:

[0.1035750, 0.1504648]

Αυτό σημαίνει ότι με 95% εμπιστοσύνη, η διαφορά του μέσου ύψους ανδρών και γυναικών φοιτητών κυμαίνεται εντός αυτού του διαστήματος, με το μέσο ύψος των ανδρών να είναι μεγαλύτερο.

```
> c <- (mean_male_height - mean_female_height) + c(-1, 1) * tstar * sqrt(((sd_male_height ^ 2) /
n_male_height) + ((sd_female_height ^ 2) / n_female_height))
> c
[1] 0.1035750 0.1504648
```

Για να υποστηρίξουμε με στατιστική βεβαιότητα ότι υπάρχει διαφορά στο μέσο ύψος ανδρών και γυναικών, χρειάζεται να κάνουμε **t-test**:

```
> t.test(male_height, female_height)

        Welch Two Sample t-test

data:  male_height and female_height
t = 10.893, df = 104.66, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1038985 0.1501413
sample estimates:
mean of x mean of y
 1.795795  1.668776
```

b) Για να διερευνήσουμε αν ο μέσος βαθμός των ανδρών φοιτητών πληροφορικής, που έχουν παρακολουθήσει το μάθημα «Στατιστική στην Πληροφορική», είναι μεγαλύτερος από τον αντίστοιχο μέσο βαθμό των γυναικών στο μάθημα «Πιθανότητες», εκτελούμε έναν μονόπλευρο έλεγχο υποθέσεων.

Οι υποθέσεις μας είναι:

- Μηδενική υπόθεση (H0): μ1 = μ2 , δηλαδή ότι οι μέσοι βαθμοί των δύο φύλων είναι ίσοι.
- Εναλλακτική υπόθεση (Ha): μ1 > μ2, δηλαδή ότι ο μέσος βαθμός των ανδρών είναι μεγαλύτερος από τον μέσο βαθμό των γυναικών.

Χρησιμοποιώντας τις παρακάτω εντολές στην **R**, διαχωρίζουμε τα δεδομένα στις αντίστοιχες μεταβλητές για άνδρες και γυναίκες, και στη συνέχεια εκτελούμε τους απαραίτητους υπολογισμούς για να πραγματοποιήσουμε τον έλεγχο σημαντικότητας.

```
> male_prob <- data[data$gender == "M",]$prob
> male_prob <- male_prob[!is.na(male_prob)]
> female_prob <- data[data$gender == "F",]$prob
> female_prob <- female_prob[!is.na(female_prob)]
```

```
> n_male_prob <- length(male_prob)
> n_male_prob
[1] 78
> mean_male_prob <- mean(male_prob)
> mean_male_prob
[1] 6.155128
> sd_male_prob <- sd(male_prob)
> sd_male_prob
[1] 2.566601
```

```
> n_female_prob <- length(female_prob)
> n_female_prob
[1] 42
> mean_female_prob <- mean(female_prob)
> mean_female_prob
[1] 5.869048
> sd_female_prob <- sd(female_prob)
> sd_female_prob
[1] 2.454611
```

Στη συνέχεια, υπολογίζουμε το στατιστικό ελέγχου **t** χρησιμοποιώντας τον γνωστό τύπο που εφαρμόσαμε και προηγουμένως. Το αποτέλεσμα του ελέγχου δίνει p-value = 0.2761521, το οποίο είναι μεγαλύτερο από το επίπεδο σημαντικότητας α=5% (0.05).

Αυτό σημαίνει ότι δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση (H0). Συνεπώς, δεν υπάρχουν στατιστικά σημαντικές ενδείξεις ότι οι άνδρες φοιτητές πληροφορικής επιτυγχάνουν υψηλότερες βαθμολογίες στο μάθημα «Πιθανότητες» σε σύγκριση με τις γυναίκες φοιτήτριες του τμήματος.

```
> t <- (mean_male_prob - mean_female_prob) / sqrt(((sd_male_prob ^ 2) / n_male_prob) +
((sd_female_prob ^ 2) / n_female_prob))
> t
[1] 0.5992488
> p <- 1 - pt(t, df=min(n_male_prob - 1, n_female_prob - 1))
> p
[1] 0.2761521
```

Με την βοήθεια της **t-test**, καταλήγουμε στο ίδιο αποτέλεσμα.

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

        Welch Two Sample t-test

data:  male_prob and female_prob
t = 0.59925, df = 87.362, p-value = 0.2753
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.5075867        Inf
sample estimates:
mean of x mean of y
 6.155128  5.869048
```

c) Πριν προχωρήσουμε στον έλεγχο σημαντικότητας, είναι σημαντικό να παρατηρήσουμε ότι ο βαθμός ενός φοιτητή στο μάθημα «Μαθηματικά 1» σχετίζεται με τον βαθμό του στο μάθημα «Πιθανότητες», καθώς και οι δύο βαθμοί αφορούν το ίδιο άτομο. Επομένως, για να απαντήσουμε στο ερώτημα, θα πρέπει να περιορίσουμε την ανάλυσή μας μόνο στις περιπτώσεις όπου υπάρχουν διαθέσιμοι βαθμοί και στα δύο μαθήματα για τον ίδιο φοιτητή. Αυτό μπορεί να γίνει χρησιμοποιώντας την παρακάτω εντολή στην R:

```
> filtered_data <- na.omit(data[c('math', 'prob')])
> dim(filtered_data)
[1] 114   2
> prob <- filtered_data$prob
> math <- filtered_data$math
```

Μετά τον περιορισμό στα δεδομένα που περιλαμβάνουν βαθμούς και στα δύο μαθήματα, το μέγεθος του δείγματος μειώνεται σε 114 περιπτώσεις αντί για 137, το οποίο παραμένει επαρκές για τη στατιστική ανάλυση. Στη συνέχεια, διατυπώνουμε τις υποθέσεις:

- **Μηδενική υπόθεση (H0):** $\mu 1 - \mu 2 = 0$, δηλαδή δεν υπάρχει διαφορά στους μέσους βαθμούς.
- **Εναλλακτική υπόθεση (Ha):** $\mu 1 - \mu 2 \neq 0$, δηλαδή υπάρχει διαφορά στους μέσους βαθμούς.

Οι μέσοι βαθμοί $\mu 1$ και $\mu 2$ αντιστοιχούν στα μαθήματα «Μαθηματικά 1» και «Πιθανότητες», αντίστοιχα.

```
> grade_dif <- math - prob
> n_gd <- length(grade_dif)
> n_gd
[1] 114
> mean_gd <- mean(grade_dif)
> mean_gd
[1] 0.4859649
> sd_gd <- sd(grade_dif)
> sd_gd
[1] 1.665734
```

Στη συνέχεια, υπολογίζουμε το στατιστικό ελέγχου **t** χρησιμοποιώντας τον τύπο που εφαρμόσαμε, ο οποίος βασίζεται στη διαφορά των μέσων βαθμών των δύο μαθημάτων («Μαθηματικά 1» και «Πιθανότητες»), διαιρεμένη με το σφάλμα αυτής της διαφοράς.

Αυτό σημαίνει ότι απορρίπτουμε τη μηδενική υπόθεση (H0: μmath = μprob)

Υπάρχουν στατιστικά σημαντικές ενδείξεις ότι οι μέσοι βαθμοί στα δύο μαθήματα διαφέρουν.

Δεδομένου ότι η μέση διαφορά είναι θετική (0.4859649), αυτό υποδηλώνει ότι, κατά μέσο όρο, οι φοιτητές επιτυγχάνουν ελαφρώς υψηλότερους βαθμούς στα Μαθηματικά 1 σε σύγκριση με τις Πιθανότητες.

$$t = \frac{\bar{x}-\bar{y}}{s/\sqrt{n}}$$

```
> t <- mean_gd / (sd_gd / sqrt(n_gd))
> t
[1] 3.114955
> p <- (1 - pt(t, df=n_gd - 1)) * 2
> p
[1] 0.002331974
```

Με την χρήση της εντολής **t.test** βγάζουμε το ακόλουθο pvalue, και καταλήγουμε στο ίδιο συμπέρασμα:

```
> t.test(grade_dif)

        One Sample t-test

data:  grade_dif
t = 3.115, df = 113, p-value = 0.002332
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1768805 0.7950494
sample estimates:
mean of x
0.4859649
```