ΣΤΑΤΙΣΤΙΚΗ ΣΤΗ ΠΛΗΡΟΦΟΡΙΚΗ

ΕΡΓΑΣΙΑ 2

ΜΕΛΗ ΟΜΑΔΑΣ:

ΑΘΑΝΑΣΙΟΣ ΚΛΕΤΤΑΣ: 3180079

ΙΑΣΩΝ ΖΙΩΓΑΣ: 3180057

ΑΣΚΗΣΗ 1:

a)

Τα δεδομένα είναι κατάλληλα για τις μεθόδους συμπερασματολογίας που γνωρίζουμε διότι η δειγματοληψία είναι τυχαία σύμφωνα μα την εκφώνηση, ο αριθμός των δεδομένων είναι n = 20 > 15 και ο πληθυσμός δεν έχει πολύ ασύμμετρη μορφή σύμφωνα με το παρακάτω stemplot.

```
> stem(x)

  The decimal point is 2 digit(s) to the right of the |

  0 | 44444
  0 | 55556688899
  1 | 013
  1 |
  2 |
  2 | 8
```

b)

Για να υπολογίσουμε το 95% διάστημα εμπιστοσύνης για τη μέση τιμή του χρόνου διεκπεραίωσης χρησιμοποιήσαμε τον εξής τύπο:

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

όπου:

- $\bar{x}$ = 77.4
- t_* = 2.093024
- s = 55.52467
- n = 20

Το διάστημα εμπιστοσύνης που προκύπτει τελικά είναι το:

(51.41365, 103.38635)

ΑΣΚΗΣΗ 2:

a)

Η τυπική απόκλιση του δειγματικού μέσου είναι: $\frac{12}{\sqrt{20}}$

b)

Οι υποθέσεις έχουν ως στόχο να εξάγουν συμπέρασμα γενικό και όχι από την ίδια την δειγματοληψία.

c)

Στην στατιστική ερευνά με δειγματική μέση τιμή ίση με 45 και μέση τιμή του πληθυσμού ίση με 54 αποκλείεται να απορριφθεί η αρχική υπόθεση για την **συγκεκριμένη εναλλακτική** καθώς το p-value θα ήταν αρκετά πιο υψηλό από τα συνηθισμένα επίπεδα σημαντικότητας.

d)

Το λάθος είναι η απόρριψη της μηδενικής υπόθεσης καθώς για να απορρίπτουμε μια μηδενική υπόθεση το p-value θα έπρεπε να είναι μικρότερο από τα συνηθισμένα επίπεδα σημαντικότητας (10%,5%,1%).

ΑΣΚΗΣΗ 3:

a)

Έχουμε ότι :

- $H_0 : \mu = \mu_0$
- $H_a : \mu > \mu_0$
- $z = 1.34$

Υπολογίζουμε το Φ(z): Φ(1.34) = 0.9099

Άρα p-value = 1 - Φ(z) = 0.0901

b)

Έχουμε ότι :

- $H_0 : \mu = \mu_0$
- $H_a : \mu < \mu_0$
- z = 1.34

Υπολογίζουμε το Φ(z): Φ(1.34) = 0.9099

Άρα p-value = Φ(z) = 0.9099

c)

Έχουμε ότι :

- $H_0 : \mu = \mu_0$
- $H_a : \mu \neq \mu_0$
- z = 1.34

Υπολογίζουμε το Φ(-|z|): Φ(-1.34) = 0.0901

Άρα p-value = 2Φ(-|z|) = 2(0.0901) = 0.1802

ΑΣΚΗΣΗ 4:

a)

Εφόσον το διάστημα εμπιστοσύνης είναι 95% τότε α = 0.05

Παρατηρούμε ότι: p-value = 0.04 < 0.05 = α

Δηλαδή p-value < α

Άρα μπορούμε να απορρίψουμε την H_0. Επομένως δεν ξέρουμε αν το 30 περιέχεται στο 95% διάστημα εμπιστοσύνης για τη μέση τιμή μ.

b)

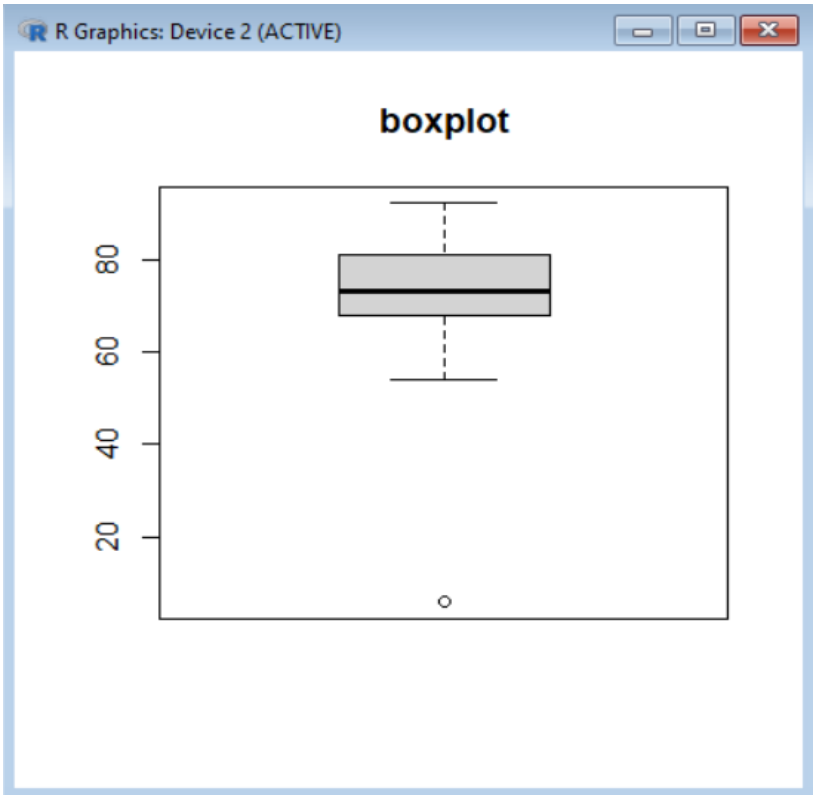
Παρατηρούμε ότι: p-value = 0.04 < 0.1 = α

Άρα η απάντηση είναι ίδια με το ερώτημα (a).

ΑΣΚΗΣΗ 5:

a)

Αρχικά παρατηρήσαμε ότι υπάρχει ένα ισχυρό outlier (A/A: 14) όπου το βάρος του είναι 6 κιλά το οποίο είναι αδύνατο για κάποιον ενήλικα. Επομένως αφαιρέσαμε την περίπτωση 14 . Παρακάτω φαίνεται και το boxplot με το οποίο το εντοπίσαμε.



Τα δεδομένα είναι κατάλληλα για τις μεθόδους συμπερασματολογίας που γνωρίζουμε διότι η δειγματοληψία είναι τυχαία σύμφωνα μα την εκφώνηση, ο αριθμός των δεδομένων είναι n = 24 > 15 και ο πληθυσμός δεν έχει πολύ ασύμμετρη μορφή σύμφωνα με το παρακάτω stemplot.

```
> stem(x3)

  The decimal point is 1 digit(s) to the right of the |

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

Για να υπολογίσουμε το 95% διάστημα εμπιστοσύνης για το μέσο βάρος των ενηλίκων κατοίκων Αθήνας χρησιμοποιήσαμε τον εξής τύπο:

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

όπου:

- $\bar{x}$ = 73.79167
- t_* = 2.068658
- s = 9.978146
- n = 24

Το διάστημα εμπιστοσύνης που προκύπτει τελικά είναι το:

(69.57826, 78.00507)

b)

Αρχικά παρατηρούμε ότι η δειγματοληψία είναι τυχαία σύμφωνα μα την εκφώνηση και ότι τα δεδομένα των βαρών των αγοριών και των κοριτσιών δεν έχουν ασύμμετρη μορφή όπως φαίνεται και από τα παρακάτω stemplots.

```
> stem(female)

  The decimal point is 1 digit(s) to the right of the |

  5 | 459
  6 | 579
  7 | 013
  8 | 23

> stem(male)

  The decimal point is 1 digit(s) to the right of the |

  6 | 8
  7 | 2233
  7 | 57
  8 | 013
  8 | 6
  9 | 12
```

Στην συνέχεια με την βοήθεια της R και της συνάρτησης t.test() υπολογίζουμε ότι το 80% διάστημα εμπιστοσύνης για τη διαφορά του μέσου βάρους μεταξύ ανδρών και γυναικών είναι:

(5.948055, 15.436561)

c)

Τα δεδομένα είναι κατάλληλα για τις μεθόδους συμπερασματολογίας που γνωρίζουμε διότι η δειγματοληψία είναι τυχαία σύμφωνα μα την εκφώνηση και ο πληθυσμός δεν έχει πολύ ασύμμετρη μορφή σύμφωνα με τα παρακάτω stemplots.

```
> stem(smokers)

  The decimal point is 1 digit(s) to the right of the |

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

> stem(not_smokers)

  The decimal point is 1 digit(s) to the right of the |

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

Στην συνέχεια παίρνουμε σαν μηδενική υπόθεση την:

- $H_0 : \mu_s = \mu_{n\_s}$

και σαν ενναλακτική υπόθεση την:

- $H_a : \mu_s \neq \mu_{n\_s}$

Όπου μ_s τα άτομα που καπνίζουν και μ_{n_s} τα που δεν καπνίζουν.

Με την βοήθεια της R και της συνάρτησης t.test() βρίσκουμε ότι:

$$p\text{-value} = 0.2228$$

Το p_value είναι πολύ μεγάλο ώστε να απορρίψουμε την μηδενική υπόθεση (H_0). Άρα φαίνεται ότι το κάπνισμα δεν έχει σχέση με το βάρος.

ΑΣΚΗΣΗ 6:

a)

Τα δεδομένα είναι κατάλληλα για τις μεθόδους συμπερασματολογίας που γνωρίζουμε διότι η δειγματοληψία είναι τυχαία σύμφωνα μα την εκφώνηση, ο αριθμός των δεδομένων είναι n = 20 > 15 και ο πληθυσμός δεν έχει πολύ ασύμμετρη μορφή σύμφωνα με το παρακάτω stemplot.

```
> stem(x2)

  The decimal point is at the |

  4 | 6999
  5 | 012334444
  5 | 67
  6 | 0334
  6 | 9
```

b)

Μέση τιμή: $\bar{x} = 5.5$
τυπική απόκλιση: $s = 0.6008766$

c)

Για να υπολογίσουμε το 95% διάστημα εμπιστοσύνης για τη μέση τιμή χρησιμοποιήσαμε τον εξής τύπο:

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

όπου:

- $\bar{x}$ = 5.5
- t_* = 2.093024
- s = 0.6008766
- n = 20

Το διάστημα εμπιστοσύνης που προκύπτει τελικά είναι το:

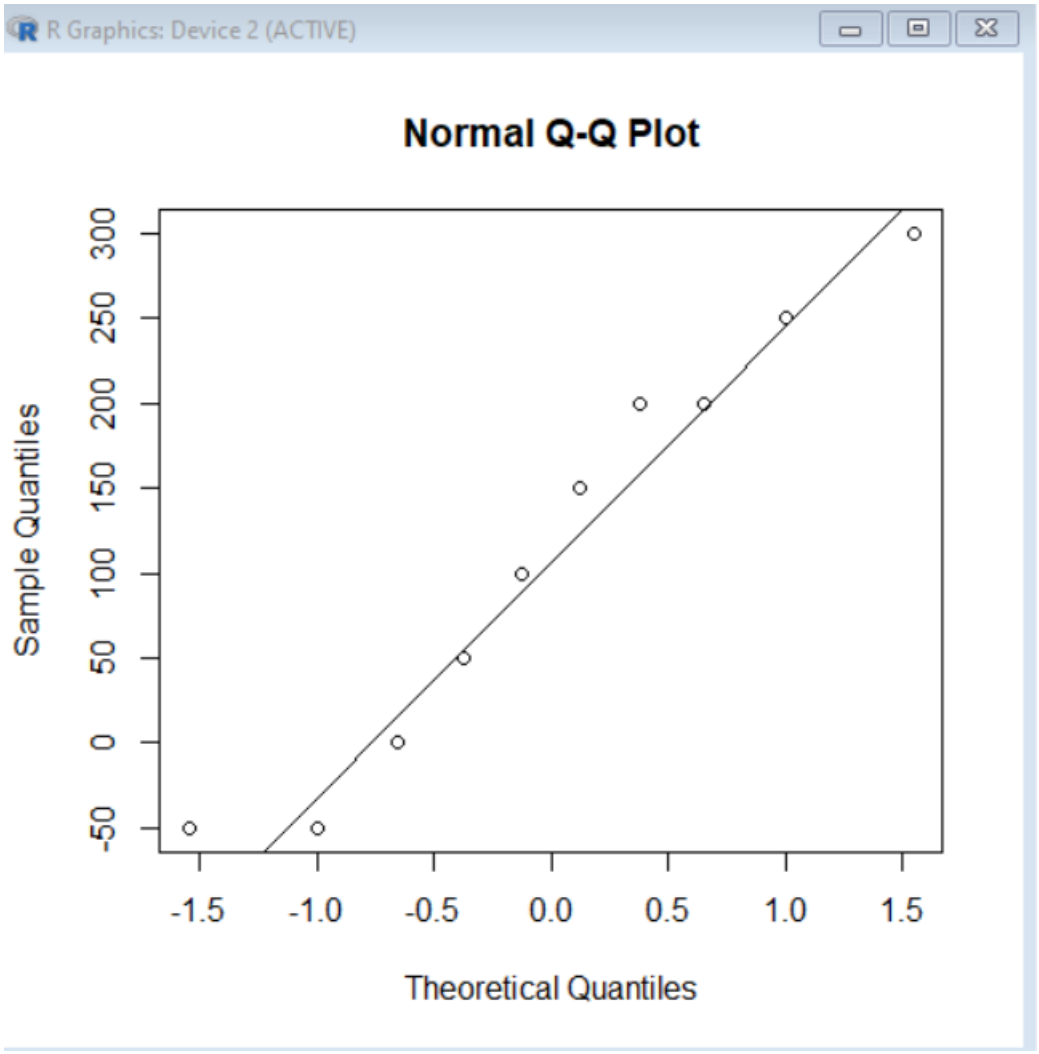
(5.218781, 5.781219)

ΑΣΚΗΣΗ 7:

Αρχικά δημιουργούμε ένα πίνακα με την διαφορά των τιμών του συνεργείου και του εμπειρογνώμονα. Ο πίνακας που προκύπτει είναι ο εξής:

100	50	-50	0	-50	200	250	200	150	300

Τα δεδομένα είναι κατάλληλα για τις μεθόδους συμπερασματολογίας που γνωρίζουμε διότι η δειγματοληψία είναι τυχαία σύμφωνα μα την εκφώνηση, ο αριθμός των δεδομένων είναι n = 10 <15 αλλά ο πληθυσμός προσεγγίζει την κανονική κατανομή όπως φαίνεται στο παρακάτω normal quantile plot.



Επίσης ο πληθυσμός δεν έχει πολύ ασύμμετρη μορφή σύμφωνα με το παρακάτω stemplot.

```
> stem(s_e)

  The decimal point is 2 digit(s) to the right of the |

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

Με μ η μέση τιμή της διαφοράς παίρνουμε

- H_0 : $\mu = 0$
- H_a: $\mu > 0$

Με την βοήθεια της R και της συνάρτησης t.test() υπολογίζουμε ότι:

p-value = 0.008611

Παρατηρούμε ότι το p-value είναι πολύ μικρό οπότε μπορούμε να απορρίψουμε την μηδενική υπόθεση (H_0).

ΑΣΚΗΣΗ 8:

a)

Αρχικά παρατηρούμε ότι τα δεδομένα μας είναι κατάλληλα για τις μεθόδους συμπερασματολογίας που γνωρίζουμε διότι η δειγματοληψία είναι τυχαία, ο αριθμός των δεδομένων είναι n > 15 και ο πληθυσμός δεν έχει πολύ ασύμμετρη μορφή σύμφωνα με τα παρακάτω stemplots.

```
> stem(men_height)

  The decimal point is 1 digit(s) to the left of the |

  16 | 3
  16 | 579
  17 | 000112333444444
  17 | 5556677888888999
  18 | 000000000011122233333344
  18 | 55555577789
  19 | 00134

> stem(women_height)

  The decimal point is 1 digit(s) to the left of the |

  15 | 4
  15 | 688
  16 | 00000122334
  16 | 555556777889
  17 | 0000
  17 | 57788
  18 |
  18 | 55
```

Με την βοήθεια της R και της συνάρτησης t.test() υπολογίζουμε ότι το 95% διάστημα εμπιστοσύνης για τη διαφορά του μέσου ύψους μεταξύ ανδρών και γυναικών φοιτητών πληροφορικής του ΟΠΑ είναι :

(0.09846512, 0.15469278)

b)

Κάνουμε τις εξής υποθέσεις:

- $H_0: \mu_M = \mu_F$
- $H_a: \mu_M > \mu_F$

Όπου μ_M η μέση τιμή των βαθμών των αγοριών στις Πιθανότητες και μ_F η μέση τιμή των βαθμών των κοριτσιών στις Πιθανότητες.

Με την βοήθεια της R και της συνάρτησης t.test() υπολογίζουμε ότι:

$$\text{p-value} = 0.8798$$

Παρατηρούμε ότι p-value = 0.8798 $< 0.5 = \alpha$

Επομένως δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση (H_0).

c)

Κάνουμε τις εξής υποθέσεις:

- $H_0: \mu_П = \mu_M$
- $H_a: \mu_П \neq \mu_M$

Όπου $\mu_П$ η μέση τιμή των βαθμών των φοιτητών που έχουν πάρει ή θα έπαιρναν το μάθημα Στατιστική στην Πληροφορική στις Πιθανότητες και μ_F η μέση τιμή των βαθμών των φοιτητών που έχουν πάρει ή θα έπαιρναν το μάθημα Στατιστική στην Πληροφορική στα Μαθηματικά 1.

Με την βοήθεια της R και της συνάρτησης t.test() υπολογίζουμε ότι:

$$\text{p-value} = 0.5445$$

Παρατηρούμε ότι το p-value είναι πού μεγάλο. Επομένως δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση (H_0).