Στατιστική στη Πληροφορική, 1η Ομαδική Εργασία

Νικηφόρος Βλάχος, p3170018

Κωνσταντίνος Νικολούτσος, p-3170122

Οικονομικό Πανεπιστήμιο Αθηνών 2019-2020

Σημείωση Συγγραφέα:

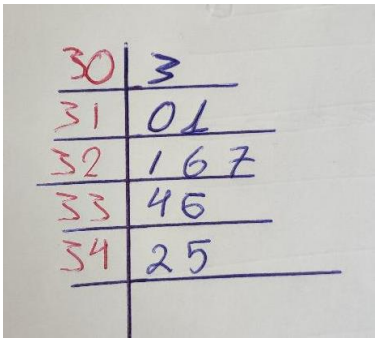
Έχουν ακολουθηθεί οι οδηγίες τις εργασίας περί των τρόπων λύσεις των ασκήσεων και συγκεκριμένα στην 1.α. τα stem/box-plots έχουν λυθεί και σχεδιαστεί σε χαρτί με γνωστές από το μάθημα μεθόδους και αυτό φαίνεται καθώς τα σχήματα είναι φωτογραφίες που έχουν επισημανθεί σε αυτό το έγγραφο για πρακτικούς λόγους. Επίσης στο Ερώτημα 3, το μέρος του β) που ζητάει να εκτελεστεί γραμμική παλινδρόμηση ελαχίστων τετραγώνων έχει ολοκληρωθεί μαζί με την "απάντηση" της α). Τελος να πουμε οτι τα δεδομένα έγιναν extract απο τα csv files με την χρήση regex!

Ερώτημα 1ο:

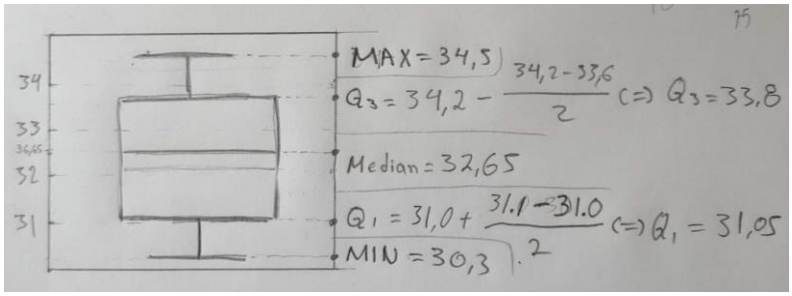
Δίνονται τα ακόλουθα δεδομένα:

Δεδομένα I									
30.3	31.0	31.1	32.1	32.6	32.7	33.4	33.6	34.2	34.5
Δεδομένα I									
30.3	31.0	31.1	32.1	32.6	32.7	33.4	33.6	34.2	34.5
Δεδομένα II									
0.0	0.0	0.2	0.8	1.2	1.4	3.2	4.2	6.4	9.0
Δεδομένα III									
0	1	6	8	10	13	15	16	17	17
18	18	20	20	21	25	26	30	35	39
40	41	43	44	46	48	52	54	58	59
59	60	66	81	86	87	88	89	94	96

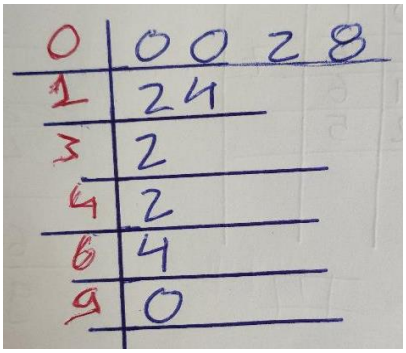
a. Τα ζητούμενα stemplot και τα boxplot:



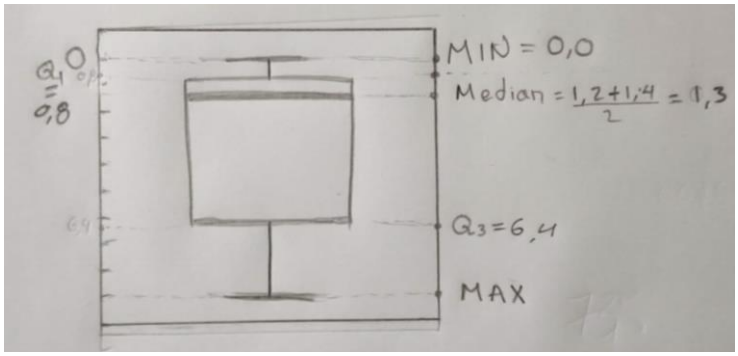
Εικόνα 1α) stemplot Δεδομένων I



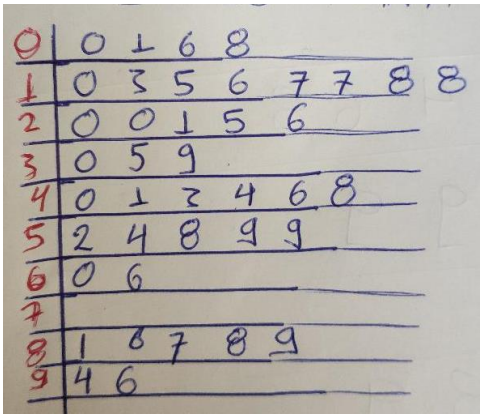
Εικόνα 1β) boxplot Δεδομένων II



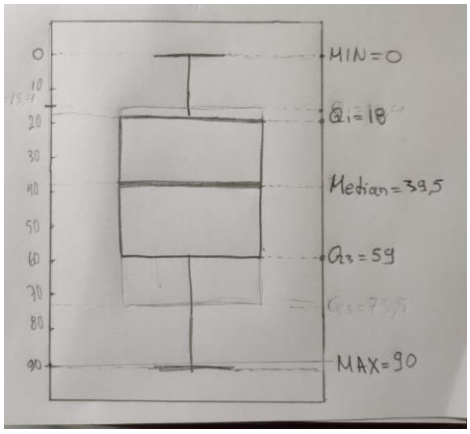
Εικόνα 2α) stemplot Δεδομένων II



Εικόνα 2β) boxplot Δεδομένων II



Εικόνα 3α) stemplot Δεδομένων III



Εικόνα 2β) boxplot Δεδομένων III

Σημείωση: Δεν βρέθηκαν ατυπικές τιμές χρησιμοποιώντας τους τύπους:

- UPPER fence: $Q_3 + 1.5(Q_3 – Q_1)$
- LOWER fence: $Q_3 - 1.5(Q_3 – Q_1)$ δηλαδή Q_3 (+ ή –) 1.5(IQS)

b.

1. Το σύνολο των Δεδομένων Ι μπορεί να αναπαρασταθεί επαρκώς και με τους δύο τρόπους, αλλά εμείς θεωρήσαμε βέλτιστο αυτόν του συνδυασμού της Μέσης τιμής με την Τυπική απόκλιση αφού τα Δεδομένα Ι είναι ομοιόμορφα κατανεμημένα με Μέση τιμή $\mu = 32.55$ και Τυπική απόκλιση $\sigma = 1.41$.
2. Το σύνολο των Δεδομένων ΙΙ αναπαριστάτε βέλτιστα με την σύνοψη των 5 αριθμών (δηλαδή boxplot) καθώς με αυτό είναι ευδιάκριτο το γεγονός πως τα στοιχεία έχουν μια κλίση προς τις χαμηλότερες τιμές δημιουργώντας έτσι μία αραίωση στις υψηλές και μια συσσώρευση στις χαμηλές, κάτι που φαίνεται καθαρά και από την τιμή της Διαμέσου και μόνο. Επίσης η Τυπική απόκλιση των Δεδομένων ΙΙ είναι αρκετά μεγάλη σχετικά με την Μέση τιμή τους ($\mu = 2.64 < \sigma = 3.05$).
3. Το σύνολο των Δεδομένων ΙΙΙ ερμηνεύονται και με τους δύο τρόπους καλά αλλά πιστεύουμε βέλτιστα με την σύνοψη των 5 αριθμών (boxplot) μιας και είναι πιο περιγραφική και ειδικά με την αναπαράσταση boxplot φαίνεται ξεκάθαρα η πραγματική κατανομή των τιμών (Αν και με την Μέση τιμή $\mu = 41.15$ και την Τυπική απόκλιση $\sigma = 28.26$, με ένα σφάλμα περίπου 8 μονάδων στην Τυπική απόκλιση θεωρητικά δεν είναι και πάρα πολύ λάθος εκπροσώπηση).

c.

Πόσο ακριβής θα ήταν η προσέγγιση της κατανομής των δεομένων από μια καμπύλη πυκνότητας της Κανονικής κατανομής; (Δώστε συγκεκριμένα παραδείγματα εγγύτητας ή απόκλισης των ποσοστημορίων τους.)

Ακολουθώντας τον αλγόριθμο του είδαμε στις διαλέξεις, συγκεκριμένα αυτόν της κατασκευής ενός Normal-Quantile plot έχουμε:

1) Διάταση των Δεδομένα I σε αύξουσα σειρά:

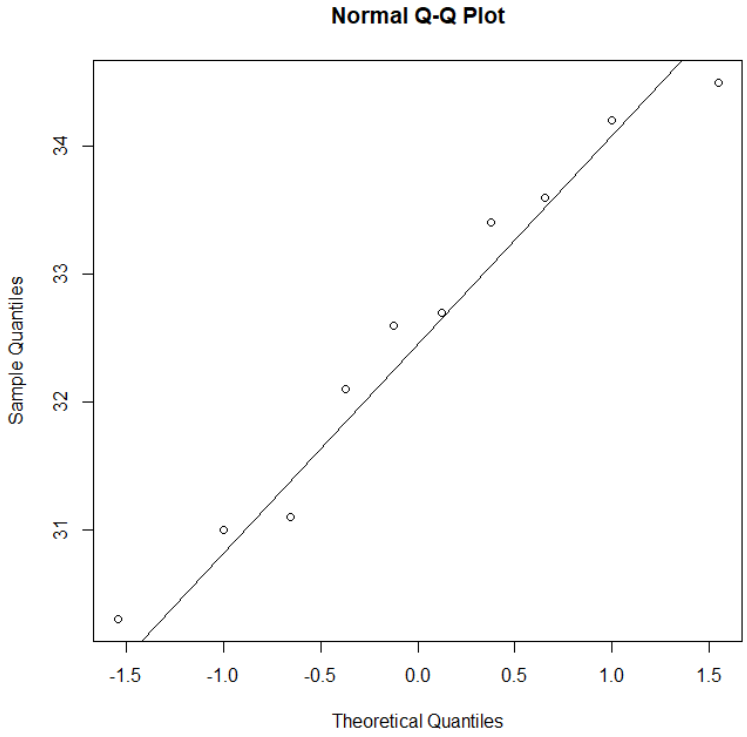
$X_1 = 30$
$X_2 = 31$
$X_3 = 31$
$X_4 = 32$
$X_5 = 33$
$X_6 = 33$
$X_7 = 33$
$X_8 = 34$
$X_9 = 34$
$X_{10} = 35$

2) Εύρεση των επί της εκατό τιμών αριστερά κάθε τιμής (δηλαδή τα ποσοστιμόρια):

$X_1 = 0\%$
$X_2 = 10\%$
$X_3 = 20\%$
$X_4 = 30\%$
$X_5 = 40\%$
$X_6 = 50\%$
$X_7 = 60\%$
$X_8 = 70\%$
$X_9 = 80\%$
$X_{10} = 90\%$

3) Υπολογισμός των $x_1 < x_2 < \ldots < x_{10}$ όπου $x_1 = p_1$, $x_2 = p_3$, $\ldots$, $x_{10} = p_{10}$ ποσοστιμόρια κανονικής κατανομής

4) Και τέλος τοποθετούμε τις τιμές των δεδομένων στον κάθετο άξονα του y και τις τιμές των ποσοστημορίων της κανονικής κατανομής στον οριζόντιο άξονα του x:

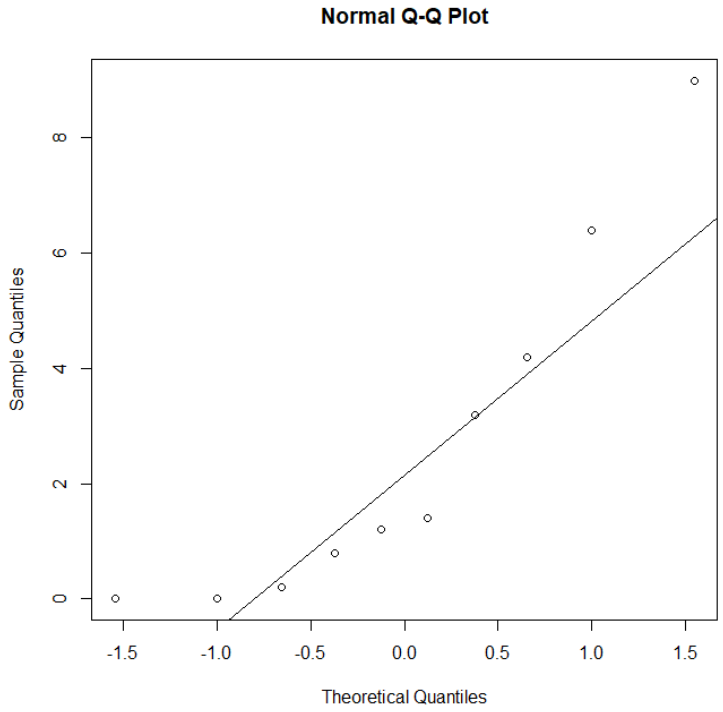


Εικόνα 1γ) Το Normal-Quantile plot των Δεδομένων Ι.

Από την μορφή (σχήμα) των σημείων στο επίπεδο μπορούμε να συμπεράνουμε πως έχοντας ίσως παραπάνω παρατηρήσεις (πλήθος Δεδομένων Ι) θα φαινόταν ακόμα πιο ισχυρή η σχέση με τον άξονα

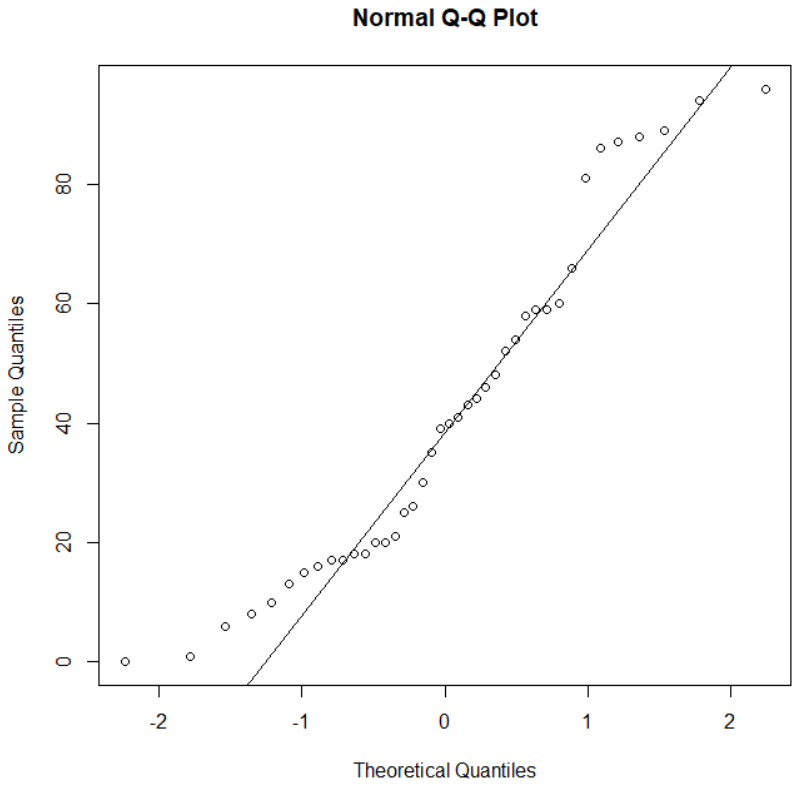
x = y. Με άλλα λόγια είναι ξεκάθαρο πως τα Δεδομένα Ι θα προσέγγιζαν επαρκώς την καμπύλη πυκνότητας της κανονικής κατανομής.

Για τα Δεδομένα II και III ακολουθώντας τον ίδιο αλγόριθμο μπορούμε να παράγουμε τα δύο ακόλουθα Normal-Quantile plots:



Εικόνα 2γ) Το Normal-Quantile plot των Δεδομένων I.

Αν και σε πλήθος τα Δεδομένα II είναι ίσα με τα Δεδομένα I φαίνεται ακόμα πιο αμφιλεγόμενο θέμα το αν η μορφή (σχήμα) των δεδομένων αυτών ακολουθεί μια ευθεία γραμμή, καθώς με αρκετά από τα ήδη λίγα στοιχεία φαίνεται να αποκλίνουν από την ευθεία. Κατά την γνώμη μας τα δεδομένα αυτά δεν προσεγγίζουν την καμπύλη πυκνότητας της κανονικής κατανομής επαρκώς με το συγκεκριμένο πλήθος παρατηρήσεων. Μπορεί όμως με τις διπλάσιες ακόμα μόνο παρατηρήσεις να μπορούσαμε να δούμε μια κάπως πιο ισχυρή σχέση.



Εικόνα 3γ) Το Normal-Quantile plot των Δεδομένων ΙΙΙ.

Με μία πρώτη ματιά στο Normal-Quantile plot των Δεδομένων ΙΙΙ θα μπορούσε κάποιος να πει πως η μορφή (σχήμα) των στοιχείων προσεγγίζουν την μορφή μια ευθείας γραμμής, όμως παρατηρώντας τα λίγο πιο προσεκτικά μπορούμε να δούμε πως στις τιμές στο πρώτο τέταρτο του άξονα x υπάρχουν όχι απλά ατυπικές τιμές αλλά pattern από στοιχεία τα οποία αποκλίνουν από την ευθεία. Ακόμα φαίνεται το σχήμα των στοιχείων να έχει μια ελικοειδή μορφή που καθιστά ακόμα δυσκολότερη την ανάλυση, μίας και οι ελικοειδείς μορφές γενικά ακολουθούν την φορά ευθείας όμως με ένα μικρό εμβαδόν (ενός υποθετικού ελικοειδές σχήματος) να επικαλύπτεται με την ευθεία αυτή, μίας και το μεγαλύτερο αποκλίνει δημιουργώντας τα ημικύκλιά. Συνεπώς κατά την γνώμη

μας τα Δεδομένα ΙΙΙ δεν προσεγγίζουν την καμπύλη πυκνότητας της κανονικής κατανομής επαρκώς.

Ερώτημα 2

α. Δώστε μια σύντομη περιγραφή από που προέρχονται τα δεδομένα και πόσες περιπτώσεις περιέχονται.

Απάντηση:

Τα στατιστικά δεδομένα προέρχονται από το επίσημο website του OECD (https://stats.oecd.org/).

Επιλέξαμε δεδομένα που δείχνουν πόσα σοβαρά ατυχήματα έχουμε ανά χρόνο στην Ελλάδα και στο Ισραήλ (2 περιπτώσεις). Πιο συγκεκριμένα περιέχονται πληροφορίες από το 1970 έως 2018.

β. Ποιες είναι κατηγορικές και ποιες ποσοτικές μεταβλητές; Δώστε μια σύντομη περιγραφή κάθε μίας από αυτές (ή ορισμένων εάν είναι πάρα πολλές).

Απάντηση:

Τα δεδομένα μας αποτελούνται από 3 columns/μεταβλητές μια Κατηγορική και δύο ποσοτικές:

- Χώρα -> Κατηγορική μεταβλητή
- Χρονιά -> Ποσοτική μεταβλητή (Αν και μπορει θεωρηθεί κατηγορική)
- Αριθμός ατυχημάτων -> Ποσοτική μεταβλητή

Αξίζει να σημειώσουμε ότι: στην μεταβλητή *Αριθμός ατυχημάτων* προσμετρούνται ατυχήματα τα οποία χρίζουν εξαιρετικής σοβαρότητας.

Όπως αναφέρθηκε και παραπάνω οι χρονιές κυμαίνονται από το 1970 έως και το 2018, και οι χώρες παίρνουν τις τιμές “Ελλάδα” ή “Ισραήλ”.

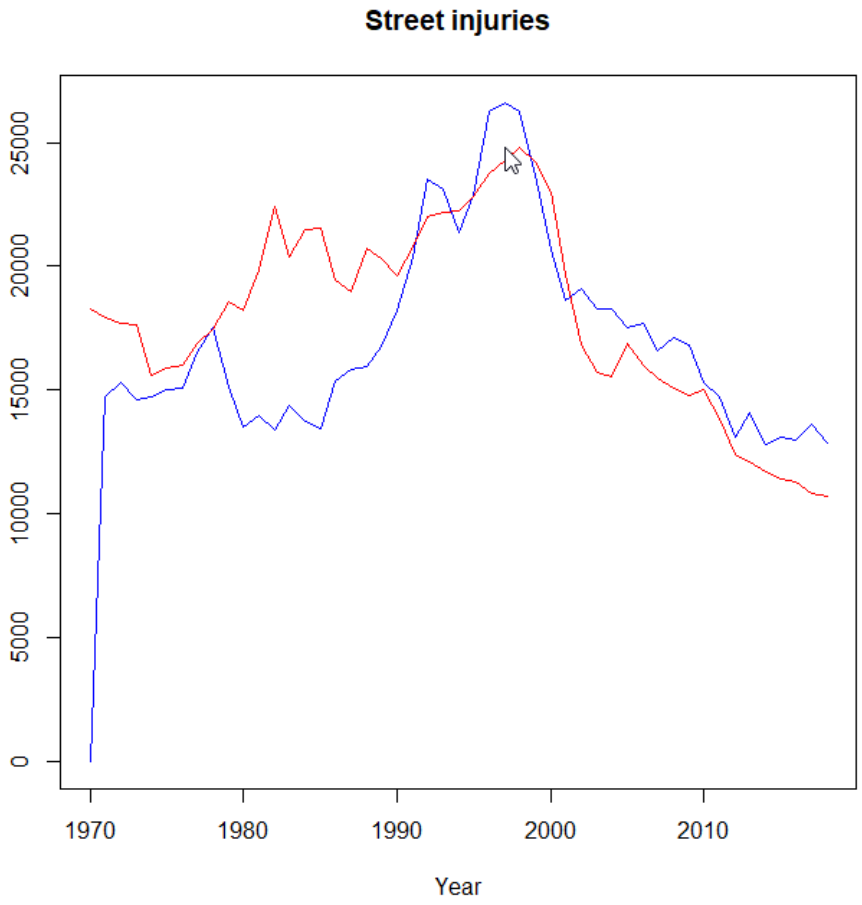
γ. Δώστε τις κατανομές των μεταβλητών σε γραφική μορφή. Σχολιάστε τη μορφή των κατανομών, πιθανούς λόγους που έχουν αυτή τη μορφή, την ύπαρξη ατυπικών σημείων (outliers) κτλ.

Απάντηση:

- Ύπαρξη ατυπικών τιμών (outliers):

Παρατηρούμε ότι στα δεδομένα μας υπάρχει ένα προφανές outliers το οποίο είναι **(ISRAEL, 1970, 0)** οπού δείχνει ότι δεν έγινε κανένα ατύχημα τότε. Αυτό ίσως οφείλεται σε κάποιο λάθος της καταγραφής τον δεδομένων.

- Τα ατυχήματα παρατηρούμε ότι αρχίζουν να μειώνονται στο πέρασμα του χρόνου. Αυτό ίσως οφείλετε στις καινούργιες τεχνολογίες των αυτοκινήτων (Αερόσακους, ταχύτερη επιβράδυνση κλπ.) καθώς και στην επένδυση από το κράτος σε καινούργιους αυτοκινητόδρομους καθώς και την συντήρηση παλαιότερων. Επίσης υπήρχε εάν spike σοβαρών ατυχημάτων μεταξύ 1990-2000 και στις δύο χώρες, το οποίο ίσως να οφειλόταν στην απότομη αύξηση των αγορών αυτοκινήτων από τα νοικοκυριά λόγο του καλύτερου παγκοσμίου οικονομικού κλίματος (μιας και που η Ελλάδα και το Ισραήλ δεν έχουν πολλές πολιτιστικές ούτε κοινωνικοπολιτικές ομοιότητες.



Εικόνα 1) Η κατανομή των δεδομένων όπου μπλε: Ισραήλ και κόκκινο: Ελλάδα.

(ανάποδα έπρεπε το ξέρουμε άλλη φορά τώρα :p)

```
> summary(y_g)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0   14114   15858   16752   18288   26628
> summary(y_i)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10743   15547   17950   17971   20764   24819
> sd(y_g)
[1] 4507.971
> sd(y_i)
[1] 3878.802
```

Και τέλος οι μεταβλητές η σύνοψη των 5 αριθμών καθώς και η τυπική απόκλιση δοσμένες από την R όπου y_g: Ελλάδα και y_i: Ισραήλ.

δ. Για κάθε ποσοτική μεταβλητή, υπολόγισε α) τη μέση τιμή και τυπική απόκλιση β) τη σύνοψη των πέντε αριθμών. Σχολιάστε την καταλληλότητα των α), β) για κάθε μεταβλητή.

```
> summary(data)
 sex        height          weight           month        colour       number           prob
 F:28   Min.   :1.530   Min.   :  1.85   Aug    :15   black  :19   Min.   :  1.00   Min.   : 0.000
 M:49   1st Qu.:1.700   1st Qu.: 60.00   May    : 8   blue   :19   1st Qu.: 13.00   1st Qu.: 5.500
        Median :1.760   Median : 70.00   Nov    : 8   purple :11   Median : 34.00   Median : 6.000
        Mean   :1.748   Mean   : 70.79   Dec    : 7   green  : 6   Mean   : 37.69   Mean   : 6.279
        3rd Qu.:1.800   3rd Qu.: 80.00   Mar    : 7   red    : 6   3rd Qu.: 59.00   3rd Qu.: 8.000
        Max.   :1.900   Max.   :120.00   Jan    : 6   white  : 5   Max.   :100.00   Max.   :10.000
                                         (Other):26   (Other):11                    NA's   :7
      math            size            pet
 Min.   : 0.00   Min.   :35.00   dog    :53
 1st Qu.: 5.50   1st Qu.:39.50   cat    :17
 Median : 7.00   Median :42.00   parrot : 2
 Mean   : 6.71   Mean   :42.05          : 1
 3rd Qu.: 8.50   3rd Qu.:44.00   hamster: 1
 Max.   :10.00   Max.   :49.50   horse  : 1
 NA's   :8                       (Other): 2
```

Προφανώς εξαιρούνται οι μεταβλητές sex, month, colour και pet καθώς είναι κατηγορικές και όχι ποσοτικές, απλά για λόγους ευχρηστίας όπως φαίνεται χρησιμοποιήσαμε κατευθείαν την εντολή summary(data) που διδάχτηκε στο εργαστήριο.

Επίσης οι τυπικές αποκλίσεις είναι οι εξής:

Height: 0.084847

Weight: 16.6188

Number: 26.72928

Prob: 2.488576

Math: 2.600065

Size: 3.008958

ε. Επιλέξτε δύο μεταβλητές και διερευνήστε τη σχέση τους. Εάν θεωρήσετε ότι υπάρχει σχέση, αυτή είναι αιτιατή ή όχι: Σχολιάστε αναλόγως.

Απάντηση:

Επιλέγουμε τις μεταβλητές Χρόνος και ατυχήματα. Υπάρχει σχέση μεταξύ τους αλλά φυσικά δεν είναι αιτιατή αφού όπως είπαμε και στο μάθημα αυτά μέσα τους κρύβουν μεταβλητές όπως τεχνολογία, οικονομική κατάσταση κλπ..

Ερώτημα 3

Εδώ θα διερευνήσετε τη σχέση μεταξύ ύψους (μεταβλητή height) και μεγέθους παπουτσιού (μεταβλητή size) στα δεδομένα των απαντήσεων ερωτηματολογίου 2019 που βρίσκονται στο eclass.

a. Δώστε το scatterplot και σχολιάστε τη μορφή, κατεύθυνση και δύναμη της σχέσης των δύο μεταβλητών.

- Load the heights in a variable

```
x = c(1.73, 1.7, 1.79, 1.61, 1.7, 1.64, 1.68, 1.87, 1.79, 1.58, 1.61, 1.67, 1.82, 1.78, 1.7,
1.74, 1.75, 1.76, 1.75, 1.84, 1.77, 1.71, 1.6, 1.78, 1.77, 1.79, 1.72, 1.68, 1.71, 1.68,
1.7, 1.8, 1.9, 1.73, 1.76, 1.76, 1.8, 1.83, 1.73, 1.7, 1.87, 1.65, 1.83, 1.74, 1.77, 1.77,
1.75, 1.67, 1.58, 1.76, 1.9, 1.88, 1.6, 1.84, 1.8, 1.75, 1.7, 1.8, 1.81, 1.63, 1.82, 1.65,
1.85, 1.88, 1.85, 1.9, 1.64, 1.8, 1.7, 1.83, 1.78, 1.53, 1.82, 1.84, 1.8, 1.71, 1.65)
```

- Load feet size in a variable

```
y = c(42, 45, 45, 38, 41, 35, 39, 45, 43, 38, 38, 40, 43, 42, 39, 43, 44, 42, 44, 44, 39,
40, 36, 43, 42, 44, 42, 39, 42, 39, 39.5, 43, 49, 43, 44, 42, 43, 44, 43, 39, 45, 40, 44,
42, 43, 45.5, 43, 38, 37, 41, 43, 46, 38, 45, 44, 41.5, 41, 42, 43, 40, 47, 38, 46, 44, 44,
49.5, 35, 46.5, 39, 45, 42, 39, 45, 46, 43, 42, 39)
```

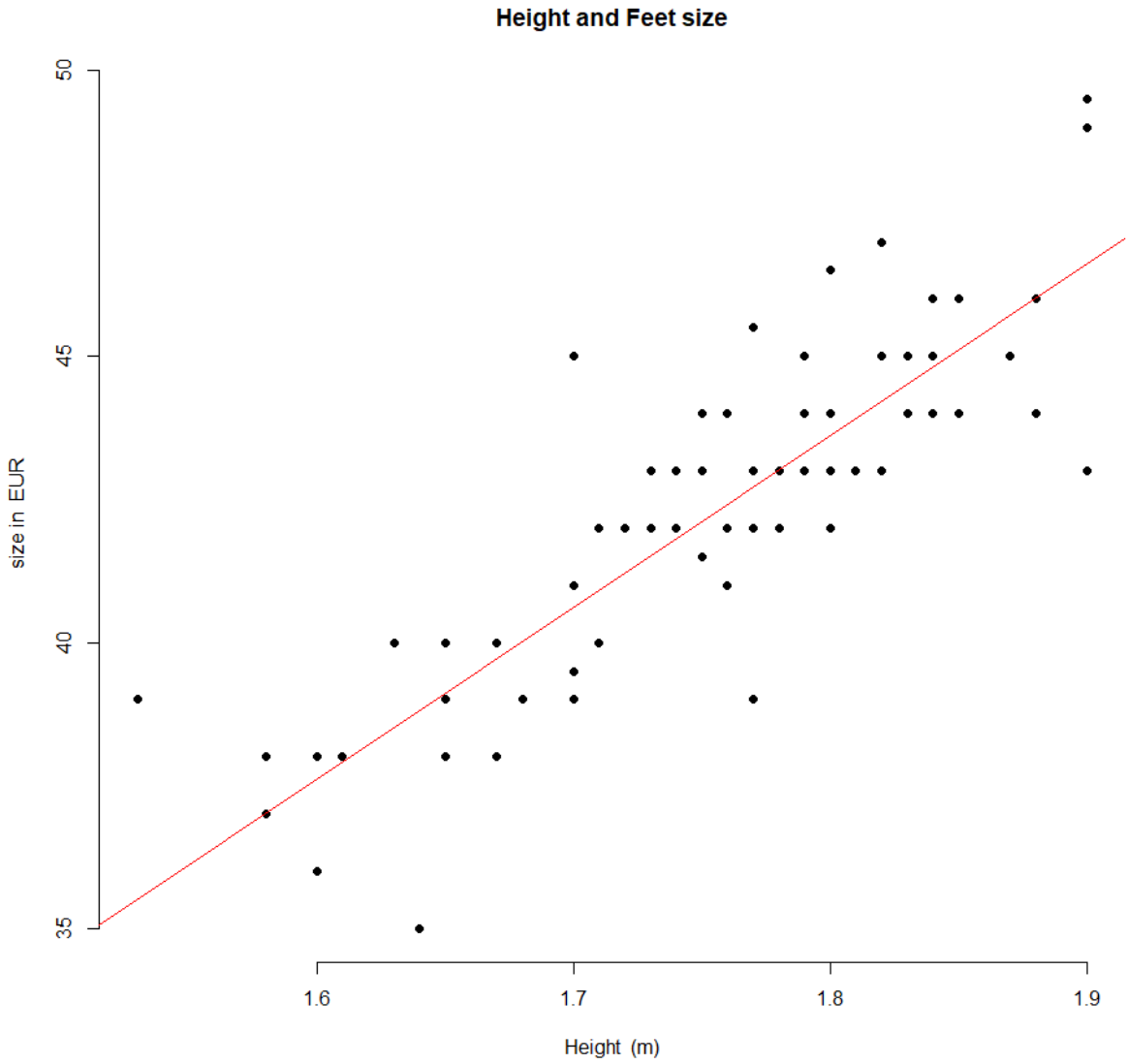
- Make the scatter plot

```
plot(x, y, main = "Height and Feet size", xlab = "Height (m)", ylab = "size in EUR", pch =
16, frame = FALSE)
```

- Apply linear regression (γραμμική παλινδρόμηση ελαχίστων τετραγώνων.)

```
abline(lm(y ~ x, data = mtcars), col = "blue")
```

Εικόνα 3α) Τα βήματα που ακολουθήσαμε για να παράγουμε το Scatterplot στην R.



Εικόνα 3β) Το Scatterplot των Δεδομένων του ερωτηματολογίου 2019

Ως προς την μορφή των δεδομένων, από την παραπάνω κλίμακα φαίνεται να έχουν μια νεφελώδες μορφή με σχήμα καθαρής ευθείας, με εξαίρεση μόνο ελάχιστα outliners.

Ως προς την κατεύθυνση, είναι ξεκάθαρο πως η θεωρητική ευθεία που παριστάνουν έχει μια κλίση που τείνει τις 45^{o}.

Τέλος η δύναμη της σχέσης των δύο μεταβλητών (x = ύψος και y = μέγεθος παπουτσιού) είναι θετική, καθώς η συνδιακύμανση τους cov(x, y) = 0.2162, δηλαδή κάτι θετικό αλλά και πάλι κάτι πιο κοντά στο 0 παρά

στο 1. Όμως από την τάση τους να σχηματίζουν ευθεία με θετική κλίση που τείνει σε αυτή της ευθείας x = y μπορούμε εντέλει να συμπεράνουμε πως η δύναμη της σχέσης τους είναι θετική ισχυρή.

b. Ο συντελεστής συσχέτισης αφού φορτώσαμε τα δεδομένα στην R και εκτελέσαμε την Συνάρτηση Cor(x, y) με x = height και y = size βγήκε r = 0.846. (Όπως αναφέρθηκε και παραπάνω η γραμμική παλινδρόμηση έχει εκτελεστή παραπάνω μαζί με την απάντηση του υποερωτήματος α. .