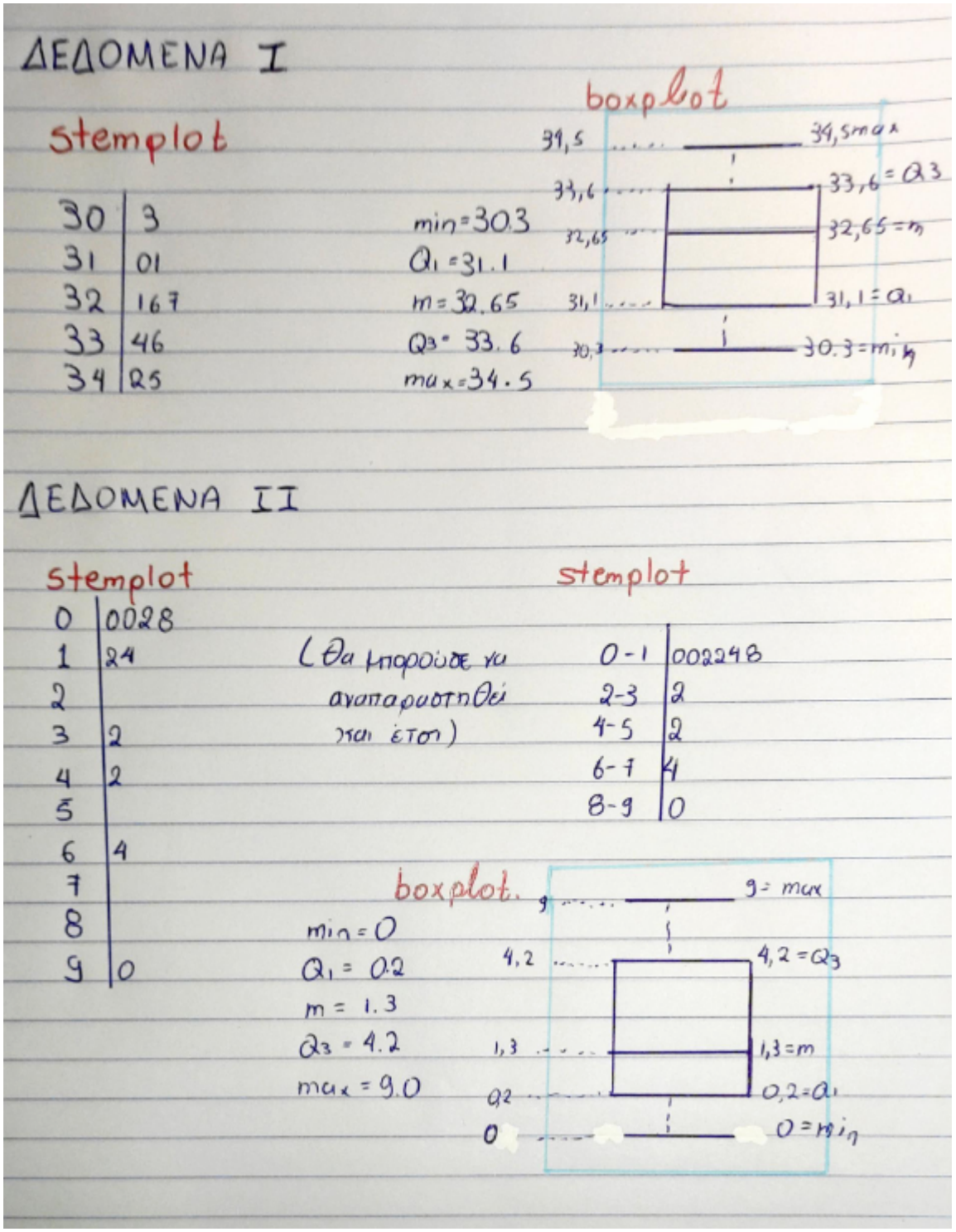
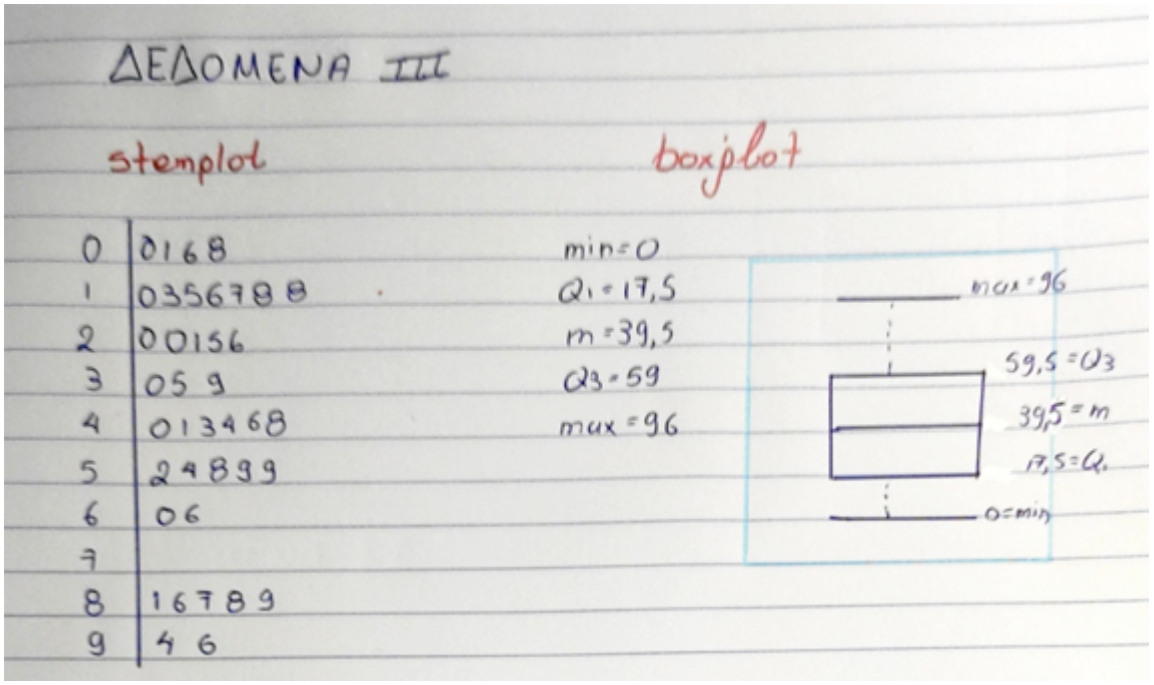
Εργασία 1

Άσκηση 1.

α)





β)

Ομάδα Δεδομένων I
Μέση Τιμή: **32.55**
Τυπική Απόκλιση: **1.41**

Τα δεδομένα είναι συμμετρικά κατανεμημένα γύρω από την μέση τιμή, επομένως η **μέση τιμή** και η **τυπική απόκλιση** συνοψίζουν καλύτερα την κατανομή.

Ομάδα Δεδομένων II
Μέση Τιμή: **2.64**
Τυπική Απόκλιση: **3.05**

Η ομάδα τιμών που περιγράφει καλύτερα τα δεδομένα είναι η σύνοψη των 5 αριθμών. Λαμβάνοντας υπόψη αποκλειστικά το ζεύγος μέση τιμή - τυπική απόκλιση δε μπορούμε να οδηγηθούμε σε συμπέρασμα για το εύρος των τιμών του πειράματός μας καθώς η κατανομή τους μπορεί να είναι εντελώς ανομοιόμορφη και σχεδόν όλες οι τιμές εμφανίζονται στο 0 και στο 1 (δηλαδή κάτω από την διάμεσο) και η μέση τιμή είναι μικρότερη από την τυπική απόκλιση. Αντίθετα, χρησιμοποιώντας την σύνοψη των 5 αριθμών, έχουμε πολύ περισσότερες πληροφορίες (min, Q1, median, Q3, max) και μπορούμε να έχουμε μια γενικότερη εικόνα της κατανομής όπως την μεγαλύτερη και τη μικρότερη τιμή, τη διάμεσο, καθώς και τις διαμέσους των επιμέρους τμημάτων (min,median), (median,max), κοινώς τα Q1, Q3.Επομένως η **σύνοψη των 5 αριθμών** συνοψίζει καλύτερα την κατανομή.

Ομάδα Δεδομένων III
Μέση Τιμή: **41.15**
Τυπική Απόκλιση: **28.26**

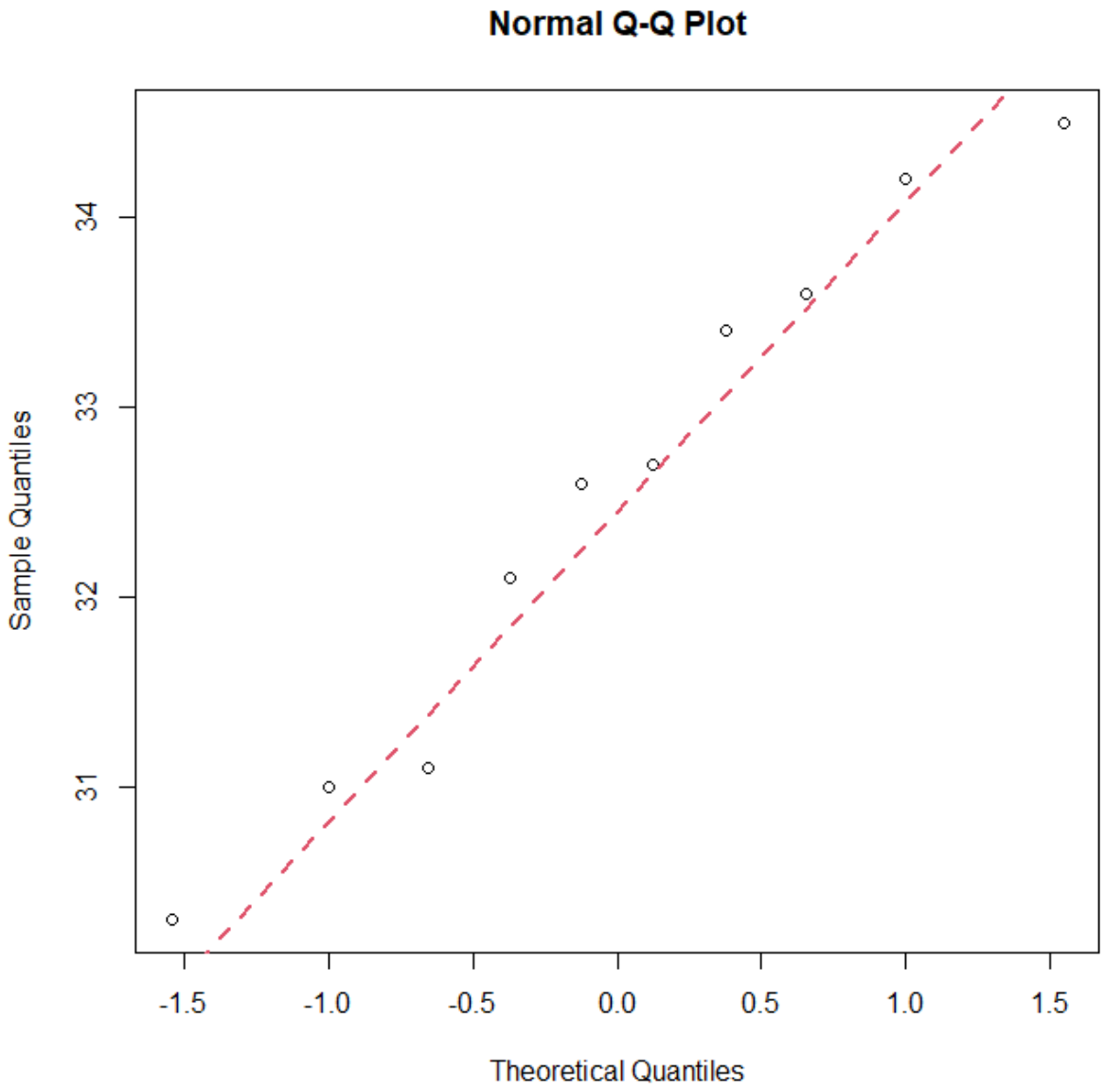
Τα δεδομένα είναι ομοιόμορφα κατανεμημένα γύρω από την μέση τιμή, επομένως η **μέση τιμή** και η **τυπική απόκλιση** συνοψίζουν την κατανομή, παρ' όλ' αυτά, το πλήθος των δεδομένων είναι μεγάλο και γιαυτό το λόγο μπορούν να αναπαρασταθούν ακριβέστερα και με την **σύνοψη των 5 αριθμών** λαμβάνοντας επιπλέον πληροφορίες οπως το (min, Q1, median, Q3, max).

γ)

Ομάδα Δεδομένων Ι

yi	30.3	31	31.1	32.1	32.6	32.7	33.4	33.6	34.2	34.5
pi	9%	18%	27%	36%	45%	54%	63%	72%	81%	90%

```
> datal <- c(30.3,31.0,31.1,32.1,32.6,32.7,33.4,33.6,34.2,34.5)
> qqnorm(datal)
> qqline(datal, col = 2, lwd=2, lty =2)
```



Η ομάδα δεδομένων Ι προσεγγίζει πολύ την καμπύλη πυκνότητας της κανονικής κατανομής.

*Από τον κανόνα **68-95-99.7**, γνωρίζουμε ότι:στην κανονική κατανομή ισχύουν τα εξής:*

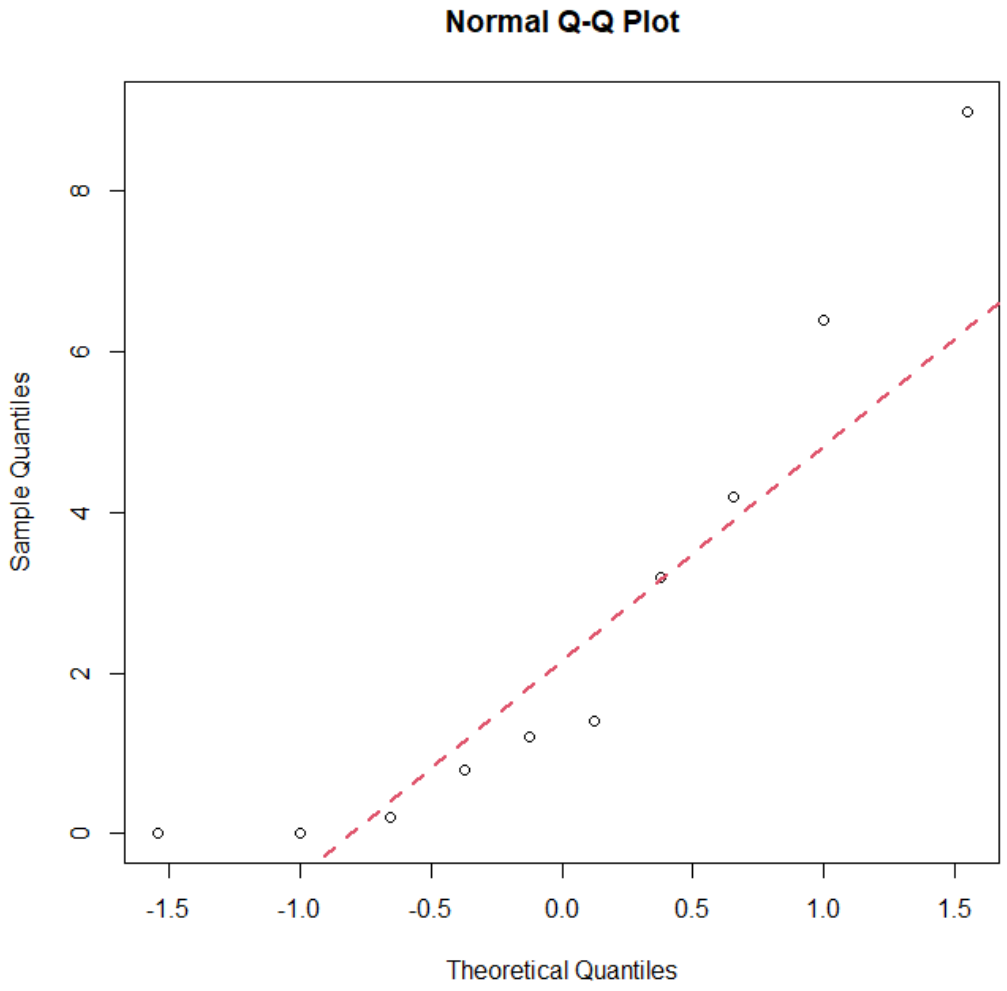
- *το 68% των παρατηρήσεων βρίσκεται στο διάστημα (μ − σ, μ + σ)*
- *το 95% των παρατηρήσεων βρίσκεται στο διάστημα (μ − 2σ, μ + 2σ)*
- *το 99,5% των παρατηρήσεων βρίσκεται στο διάστημα (μ − 3σ, μ + 3σ),*

όπου μ = μέση τιμή και σ = τυπική απόκλιση.

Όπως έχουμε ήδη βρει μ =2.64, σ =3.05
Παρατηρούμε ότι στο διάστημα (-0.41, 5,69) δηλαδή το αντίστοιχο (μ − σ, μ + σ) για τα δικά μας δεδομένα, βρίσκονται οι 8 από τις 10 παρατηρήσεις, ποσοστό της τάξης του 80% και όχι του αναμενόμενου 68% ώστε να έχουμε κανονική κατανομή ή κάποιας κοντινής σε αυτή.Έτσι, καταλήγουμε στο συμπέρασμα ότι η προσέγγιση της κατανομής των δεδομένων μας από την καμπύλη πυκνότητας της Κανονικής Κατανομής θα παρουσιάζει σημαντικές αποκλίσεις . Κάνοντας δοκιμές για τα αντίστοιχα διαστήματα (μ − 2σ, μ + 2σ)δηλαδή στο διάστημα(-3.46,8,74) βρίσκεται το 90% των παρατηρήσεών μας έναντι του 95% της Κανονικής Κατανομής, και στο (μ − 3σ, μ +3σ) δηλαδή στο διάστημα (-6.51,11.79) βρίσκεται και εκεί το 100% αυτών έναντι του 99,5%.

Ομάδα Δεδομένων II

yi	0.0	0.0	0.2	0.8	1.2	1.4	3.2	4.2	6.4	9.0
pi	9%	18%	27%	36%	45%	54%	63%	72%	81%	90%



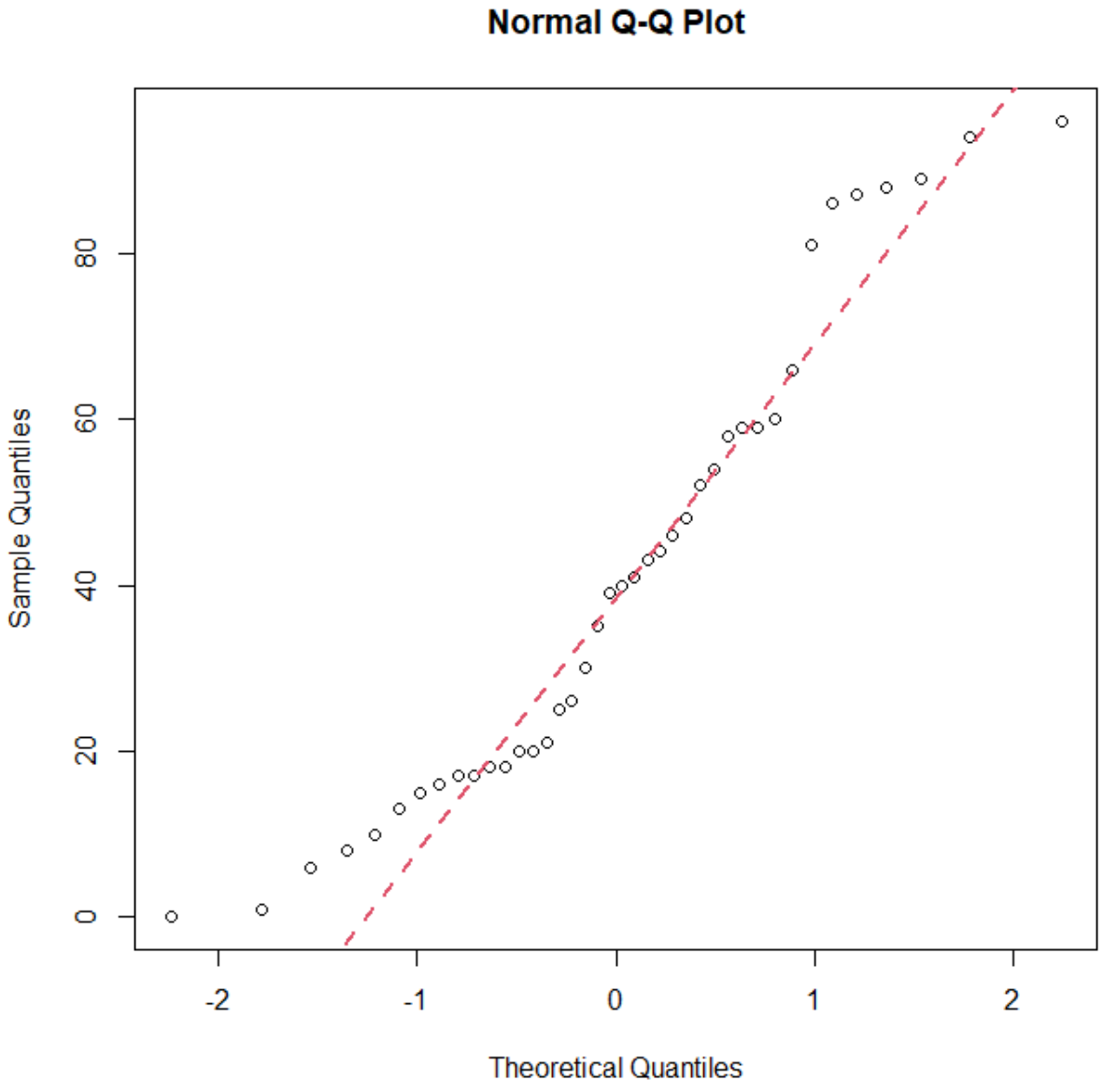
Τα δεδομένα της ομάδας II αποκλίνουν από την καμπύλη πυκνότητας της κανονικής κατανομής.

*Από τον κανόνα **68-95-99.7**, γνωρίζουμε ότι:στην κανονική κατανομή ισχύουν τα εξής:*
- *το 68% των παρατηρήσεων βρίσκεται στο διάστημα (μ − σ, μ + σ)*
- *το 95% των παρατηρήσεων βρίσκεται στο διάστημα (μ − 2σ, μ + 2σ)*
- *το 99,5% των παρατηρήσεων βρίσκεται στο διάστημα (μ − 3σ, μ + 3σ),*

όπου μ = μέση τιμή και σ = τυπική απόκλιση.

Όπως έχουμε ήδη βρει μ =32.55, σ = 1.41
Παρατηρούμε ότι στο διάστημα (31.14, 33.96) δηλαδή το αντίστοιχο (μ − σ, μ + σ) για τα δικά μας δεδομένα, βρίσκονται οι 5 απο τις 10 παρατηρήσεις, ποσοστό της τάξης του 50% και όχι του αναμενόμενου 68% ώστε να έχουμε κανονική κατανομή ή κάποιας κοντινής σε αυτή.Έτσι, καταλήγουμε στο συμπέρασμα ότι η προσέγγιση της κατανομής των δεδομένων μας από την καμπύλη πυκνότητας της Κανονικής Κατανομής θα παρουσιάζει αποκλίσεις . Κάνοντας δοκιμές για τα αντίστοιχα διαστήματα (μ − 2σ, μ + 2σ)δηλαδή στο διάστημα(29.73,35.37) βρίσκεται το 100% των παρατηρήσεών μας έναντι του 95% της Κανονικής Κατανομής, και στο (μ − 3σ, μ +3σ) δηλαδή στο διάστημα (28.32,36.78) βρίσκεται και εκεί το 100% αυτών έναντι του 99,5%.

Ομάδα Δεδομένων III
Ομοίως και με παραπάνω.



Κάποια δεδομένα της ομάδας III συμπίπτουν με την καμπύλη αλλά πολλά άλλα αποκλίνουν. Ακολουθούν σχεδόν ελικοειδές σχήμα με αποτέλεσμα να μην προσεγγίζουν επαρκώς την καμπύλη πυκνότητας της κανονικής κατανομής.

*Από τον κανόνα **68-95-99.7**, γνωρίζουμε ότι:στην κανονική κατανομή ισχύουν τα εξής:*
- *το 68% των παρατηρήσεων βρίσκεται στο διάστημα (μ − σ, μ + σ)*
- *το 95% των παρατηρήσεων βρίσκεται στο διάστημα (μ − 2σ, μ + 2σ)*
- *το 99,5% των παρατηρήσεων βρίσκεται στο διάστημα (μ − 3σ, μ + 3σ),*

όπου μ = μέση τιμή και σ = τυπική απόκλιση.

Όπως έχουμε ήδη βρει μ =41,15, σ = 28,26
Παρατηρούμε ότι στο διάστημα (12.89 ,69.41) δηλαδή το αντίστοιχο (μ − σ, μ + σ) για τα δικά μας δεδομένα, βρίσκονται οι 16 απο τις 37 παρατηρήσεις, ποσοστό της τάξης του 43,24% και όχι του αναμενόμενου 68% ώστε να έχουμε κανονική κατανομή ή κάποιας κοντινής σε αυτή.Έτσι, καταλήγουμε στο συμπέρασμα ότι η προσέγγιση της κατανομής των δεδομένων μας από την καμπύλη πυκνότητας της Κανονικής Κατανομής θα παρουσιάζει σημαντικές Αποκλίσεις . Κάνοντας δοκιμές για τα αντίστοιχα διαστήματα (μ − 2σ, μ + 2σ)δηλαδή στο διάστημα (-15,37,97.67) βρίσκεται το 100% των παρατηρήσεών μας έναντι του 95% της Κανονικής Κατανομής, και στο (μ − 3σ, μ +3σ) δηλαδή στο διάστημα (-43.63,125.93) βρίσκεται και εκεί το 100% αυτών έναντι του 99,5%.

Άσκηση 2

α)

Τα δεδομένα που χρησιμοποιούμε προέρχονται από την Ελληνική Στατιστική Υπηρεσία (ΕΛΣΤΑΤ) και αφορούν τα ατυχήματα που έλαβαν χώρα κατά τη χρονική περίοδο 1991-2001 στην Ελλάδα και αναλύουν πόσα από τα ατυχήματα αυτά ήταν θανατηφόρα ή μη καθώς και τους τραυματίες και τους θανόντες που προκλήθηκαν από αυτά.

β)

Κατηγορικές Μεταβλητές

Χρονολογία: Τα στοιχεία μας αφορούν το εύρος χρονολογιών 1991-2019

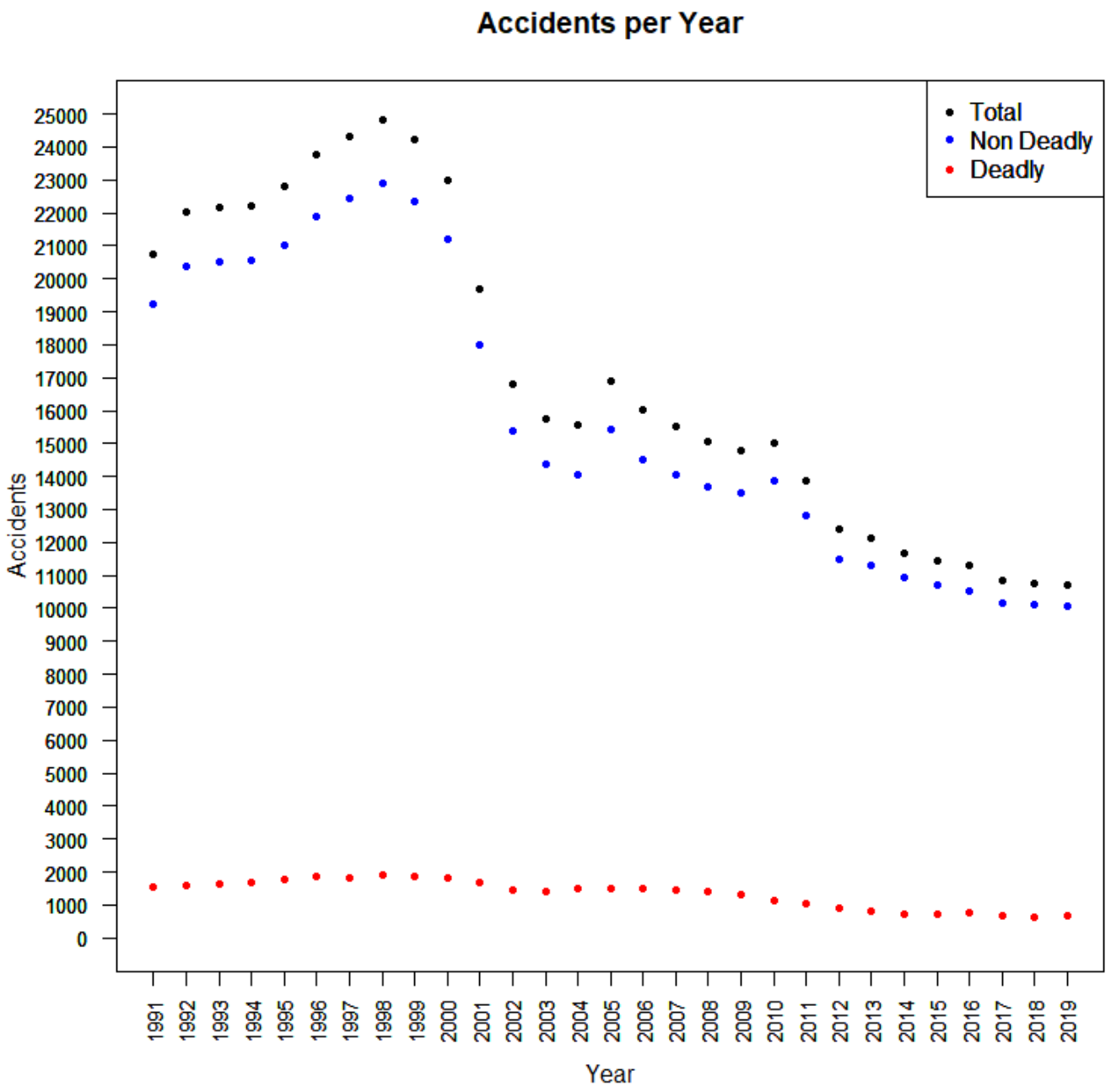
Ποσοτικές Μεταβλητές

Σύνολο τροχαίων ατυχημάτων: Ο αριθμός τροχαίων ατυχημάτων που έλαβαν μέρος. (ανά έτος)

Σύνολο θανατηφόρων ατυχημάτων: Ο αριθμός θανατηφόρων τροχαίων ατυχημάτων που έλαβαν μέρος. (ανά έτος)

Σύνολο μη θανατηφόρων ατυχημάτων: Ο αριθμός μη θανατηφόρων τροχαίων ατυχημάτων που έλαβαν μέρος. (ανά έτος)

γ)



Στα στοιχεία μας δεν παρουσιάζονται ατυπικές τιμές.

Όπως είναι αναμενόμενο, παρατηρείται μείωση των ατυχημάτων με την πάροδο του χρόνου, ειδικότερα μετά από το 1998. Όσο η τεχνολογία εξελίσσεται και οι νόμοι ενισχύονται για την αποφυγή τους. Παρόλο που περισσότερα οχήματα υπάρχουν στους δρόμους η οδήγηση σήμερα είναι ασφαλότερη από ότι παλαιότερα.

Η μείωση των ατυχημάτων μετά από το 1998 θα μπορούσε να οφείλεται στην αναπροσαρμογή που επήλθε στον ΚΟΚ το 1998-1999, και τις καινούργιες αυστηρότερες διατάξεις του. Σύμφωνα με αυτές, μεταξύ άλλων αλλαγών, πολλά πρόστιμα υπερδιπλασιάστηκαν.

```
> summary(TA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10712   12398   15751   17113   22165   24819
> summary(NDA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10056   11490   14351   15774   20531   22898
> summary(DA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    645     908    1442    1340    1669    1921
```

δ)

Σύνολο τροχαίων ατυχημάτων
Τυπική Απόκλιση: **4871.59**
Min: **10712**
Q1: **12398**
Μέση Τιμή: **17113**
Q3: **22165**
Max: **24819**

Σύνολο μη θανατηφόρων τροχαίων ατυχημάτων
Τυπική Απόκλιση: **4470.38**
Min: **10056**
Q1: **11490**
Μέση Τιμή: **15774**
Q3: **20531**
Max: **22898**

Σύνολο θανατηφόρων τροχαίων ατυχημάτων
Τυπική Απόκλιση: **427.40**
Min: **645**
Q1: **908**
Μέση Τιμή: **1340**
Q3: **1669**
Max: **1921**

```
> NDA <- (data$Non.Deadly.Accidents)
> DA <- (data$Deadly.Accidents)
> TA <- (data$Total.Accidents)
> sd(NDA)
[1] 4470.389
> sd(DA)
[1] 427.4083
> sd(TA)
[1] 4871.593
> mean(NDA)
[1] 15773.93
> mean(DA)
[1] 1339.552
> mean(TA)
[1] 17113.48
> quantile(NDA)
   0%   25%   50%   75%  100%
10056 11490 14351 20531 22898
> quantile(DA)
  0%  25%  50%  75% 100%
 645  908 1442 1669 1921
> quantile(TA)
   0%   25%   50%   75%  100%
10712 12398 15751 22165 24819
> min(NDA)
[1] 10056
> min(DA)
[1] 645
> min(TA)
[1] 10712
> max(NDA)
[1] 22898
> max(DA)
[1] 1921
> max(TA)
[1] 24819
```

```
> stem(NDA)

  The decimal point is 3

  10 | 11257035
  12 | 8579
  14 | 114544
  16 |
  18 | 02
  20 | 456029
  22 | 459
```

```
> stem(TA)

  The decimal point is 3

  10 | 778347
  12 | 148
  14 | 801558
  16 | 089
  18 | 7
  20 | 8
  22 | 022808
  24 | 238
```

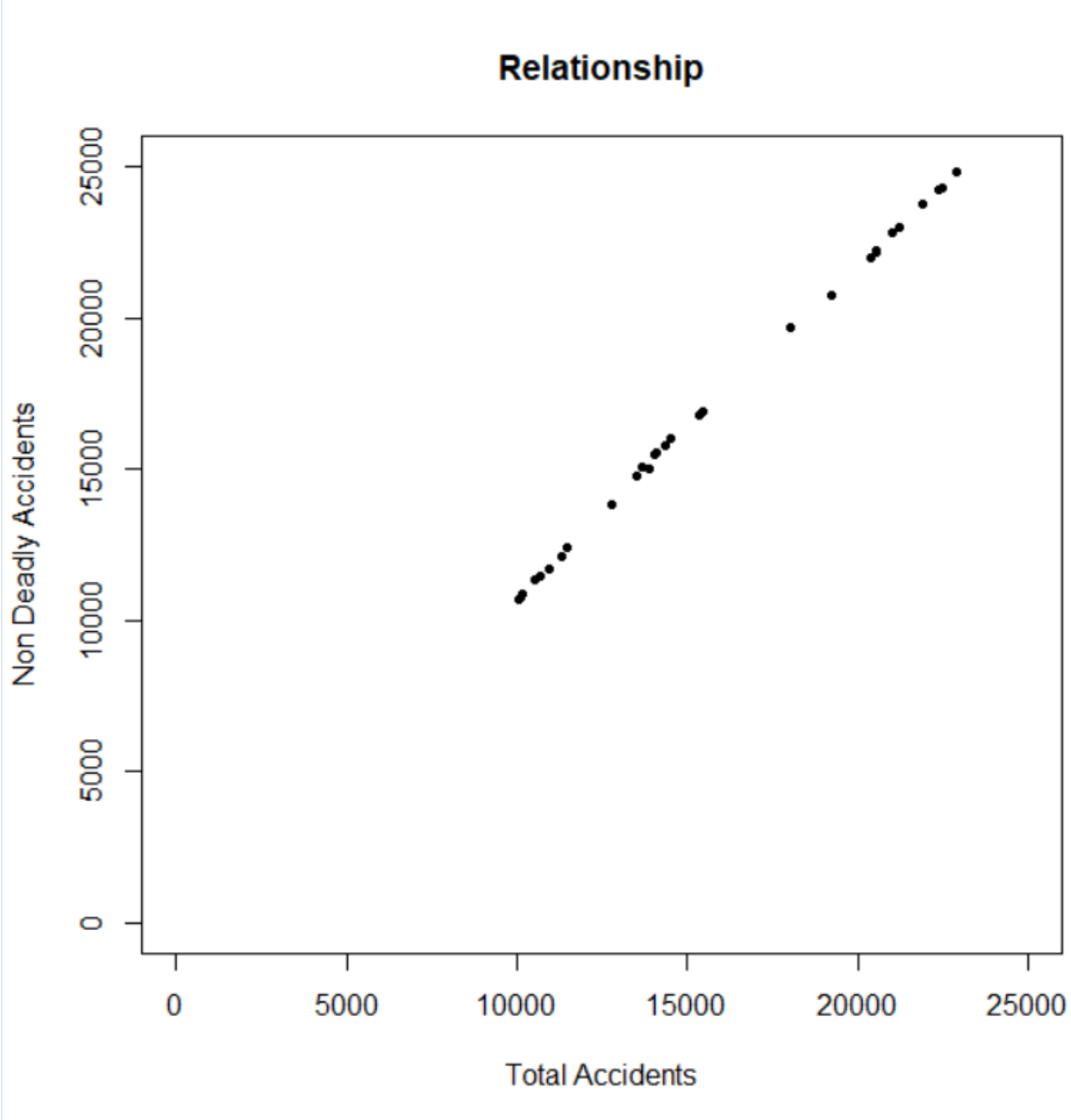
```
> stem(DA)

  The decimal point is 2

   6 | 568447
   8 | 11
  10 | 54
  12 | 0
  14 | 01448806
  16 | 1377
  18 | 004782
```

Η ομάδα τιμών που περιγράφει καλύτερα τα δεδομένα και των τριών περιπτώσεων είναι η σύνοψη των 5 αριθμών. Λαμβάνοντας υπόψη αποκλειστικά το ζεύγος μέση τιμή - τυπική απόκλιση δε μπορούμε να οδηγηθούμε σε συμπέρασμα για το εύρος των τιμών των πειραμάτων μας καθώς η κατανομή τους μπορεί να είναι εντελώς ανομοιόμορφη. Επιπλέων το πλήθος των δεδομένων μας είναι αρκετά μεγάλο. Αντίθετα, χρησιμοποιώντας την σύνοψη των 5 αριθμών, έχουμε πολύ περισσότερες πληροφορίες (min, Q1, median, Q3, max) και μπορούμε να έχουμε μια γενικότερη εικόνα της κατανομής όπως την μεγαλύτερη και τη μικρότερη τιμή, τη διάμεσο, καθώς και τις διαμέσους των επιμέρους τμημάτων (min,median), (median,max), κοινώς τα Q1, Q3. Επομένως η **σύνοψη των 5 αριθμών** συνοψίζει καλύτερα την κατανομή.

ε) Θα μελετήσουμε τις μεταβλητές "Συνολικά Ατυχήματα" και "Μη Θανατηφόρα Ατυχήματα". Όπως παρατηρούμε και στο scatterplot που ακολουθεί βλεπουμε πως, προφανώς, όσο αυξάνονται τα ατυχήματα αυξάνονται και τα μη θανατηφόρα ατυχήματα.

Επεξηγηματική μεταβλητή → Σύνολο Ατυχημάτων
Μεταβλητή απόκρισης → Σύνολο Μη Θανατηφόρων Ατυχημάτων

Κάθε περίπτωση i δίνει σημείο (x,y) όπου x= Total accidents και y = Non deadly accidents

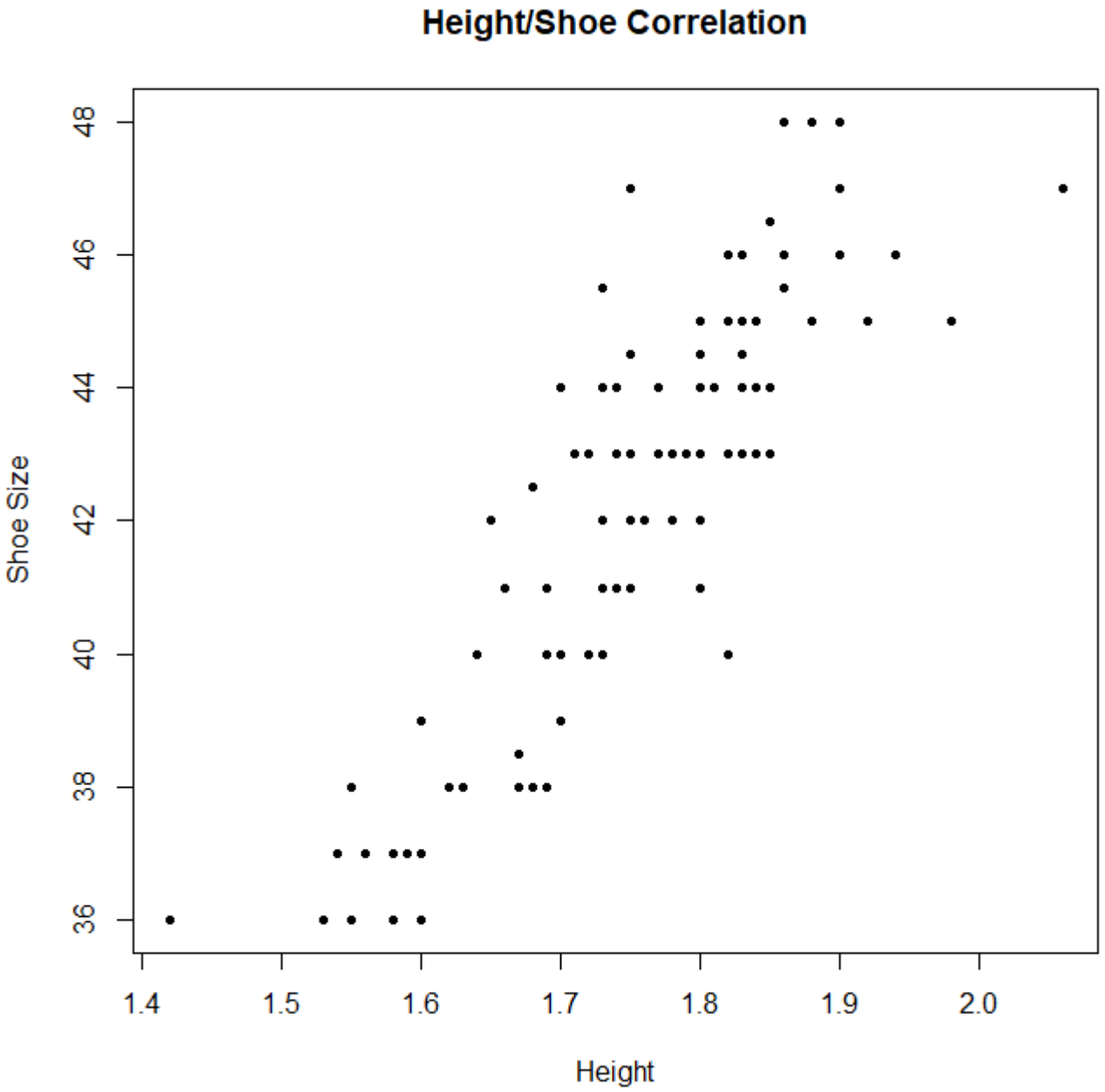


```
> cor(data$Total.Accidents,data$Non.Deadly.Accidents)
[1] 0.9995015
```

Ο συντελεστής συσχέτισης r είναι μεγαλύτερός του μηδενος και περίπου 1 οπότε πρόκειται για **αύξουσα, γραμμική συσχέτιση με πολύ μεγάλη ισχύ** και κανένα ατυπικό σημείο, όπως ήταν και αναμενόμενο εφόσον τα ατυχήματα χωρίς θανόντες αποτελούν υποσύνολο όλων των ατυχημάτων.

Άσκηση 3

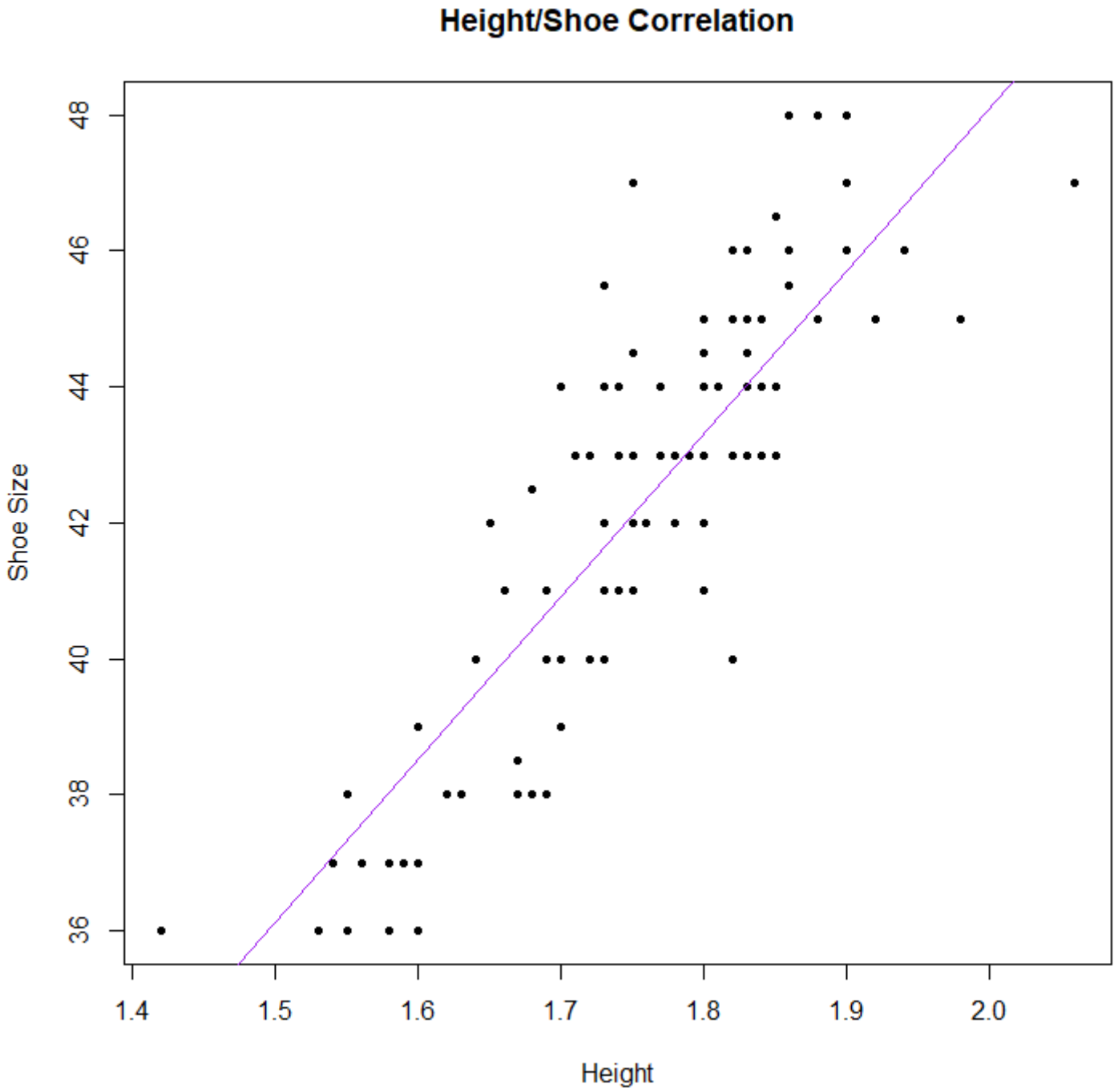
α) Τα αποτελέσματα που ακολουθούν αφορούν μόνο όσους απάντησαν και για τις δύο τιμές, ενας εκ των 123 υποψηφίων δεν έδωσε απάντηση σε ένα εκ των δύο αυτών πεδίων.




```
> data = read.csv("C:\\Users\\Debs\\Downloads\\data.csv", header=T)
> shoe <- (data$shoe)
> height <- (data$height)
> cor(height, shoe)
[1] 0.8461584
> plot(height, shoe,pch=20, ylab= "Shoe Size", xlab = "Height", main = "Height/Shoe Correlation")
```

Ο συντελεστής συσχέτισης r είναι μεγαλύτερός του μηδενος με τιμή r = 0.84, οπότε πρόκειται για **αύξουσα, γραμμική συσχέτιση με μέτρια ισχύ** και 2 ατυπικά σημεία τα οποία όμως δεν είναι λαθεμένα. Η συσχέτιση είναι αναμενόμενη εφόσον συνήθως όσο πιο ψηλός είναι κάποιος τόσο πιο μεγάλο νούμερο παπούτσι φοράει.

β) Οι περιπτώσεις στις οποίες τα στοιχεία δεν είχαν απαντηθεί αφαιρέθηκαν και συντελεστής συσχέτισης r είναι μεγαλύτερός του μηδενος με τιμή r = 0.84. Η γραμμική παλινδρόμηση ελαχίστων τετραγώνων είναι αυτή:




```
> cor(height,shoe,use="complete.obs")
[1] 0.8461584
> with(data,lm(shoe ~ height))-> m
> abline(m, col="purple")
```