Στατιστική στην Πληροφορική

Φέκα Αγγελική A.M:3140290
Σταυρουλάκη Μαρία A.M:3160168

1η σειρά ασκήσεων

Άσκηση 1η :
a)

a)

Δεδομένα I : m=32,65

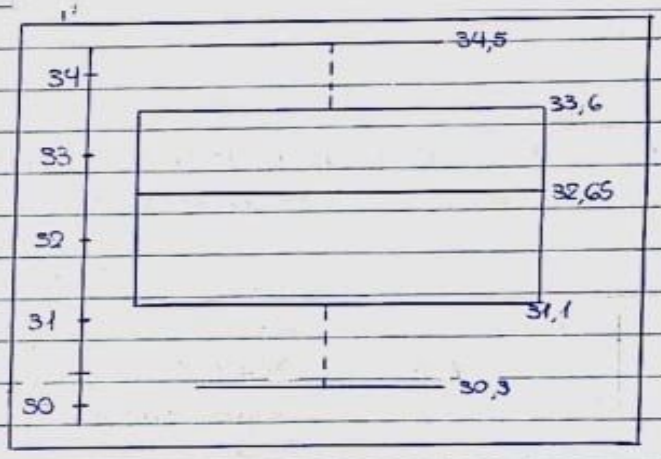
30,3 | 31,0 | 31,1 | 32,1 | 32,6 | 32,7 | 33,4 | 33,6 | 34,2 | 34,5

Stemplot

30	3
31	0, 1
32	6, 7
33	4, 6
34	2, 5

Boxplot

Διάμεση τιμή m = 32,65
Q1 = 31,1
Q3 = 33,6
min = 30,3
max = 34,5



Δεδομένα II $m=1,5$

0,0 | 0,0 | 0,2 | 0,8 | 1,2 | 1,4 | 3,2 | 4,2 | 6,4 | 9,0

Stemplot:

0	0, 2, 8
1	2, 4
2	
3	2
4	2
5	
6	4
7	
8	
9	0

Boxplot:

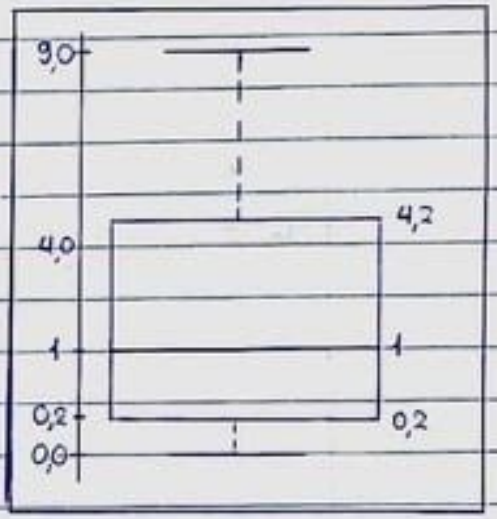
min = 0,0

$Q_1 = 0,2$

Διάμεση τιμή $m = 1,5$

$Q_3 = 4,2$

max = 9,0



Δεδομένα III $m=39,5$

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, 55, 59, 59, 60, 66, 81, 86, 87, 88, 89, 94, 96

Stemplot:

0	0, 1, 6, 8
1	0, 3, 5, 6, 7, 8
2	0, 1, 5, 6
3	0, 5, 9
4	0, 1, 3, 4, 6, 8
5	2, 4, 8, 9, 9
6	0, 6
7	
8	1, 6, 7, 8, 9
9	4, 6

Boxplot:

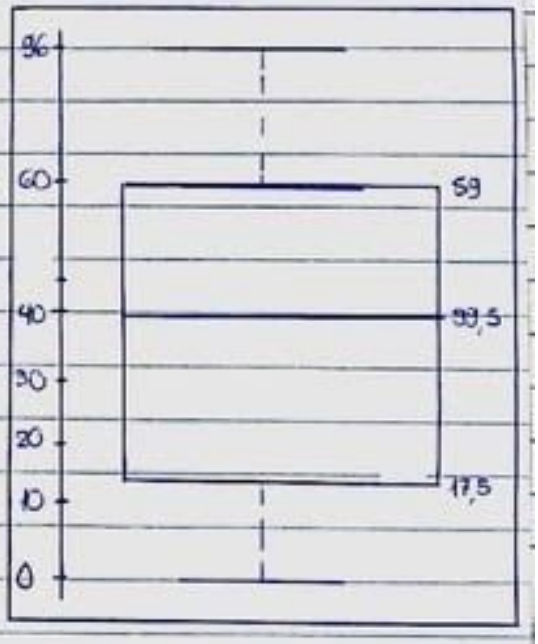
min = 0

$Q_1 = 17,5$

Διάμεση τιμή $m = 39,5$

$Q_3 = 59$

max = 96

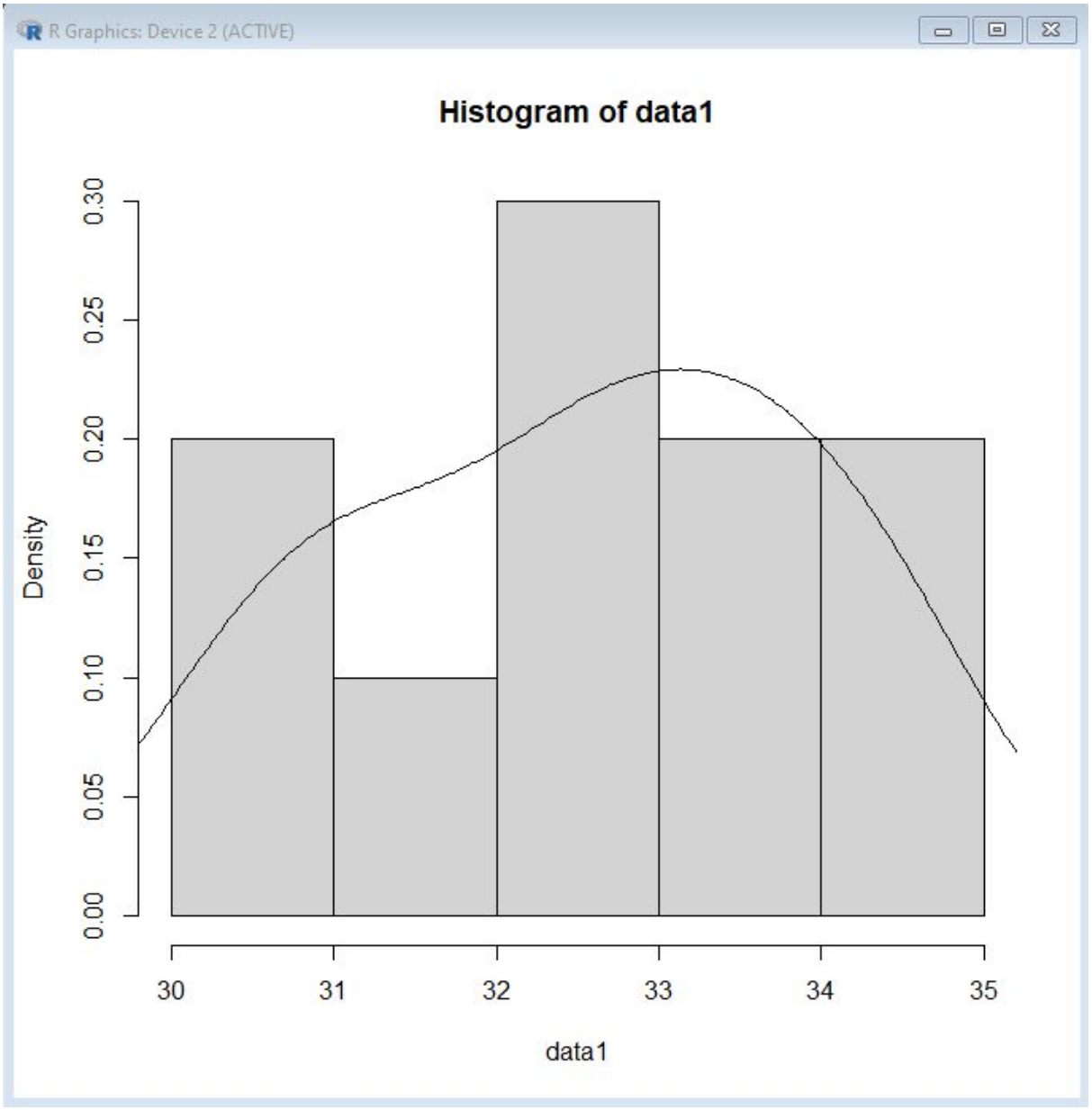


b)
Και για τις τρεις ομάδες δεδομένων ισχύει ότι η ομάδα τιμών που συνοψίζει καλύτερα την κατανομή είναι η σύνοψη των 5 αριθμών. Αυτό συμβαίνει διότι λαμβάνοντας υπόψην μόνο την μέση τιμή και την τυπική απόκλιση δεν γίνεται να οριστεί με σιγουριά το εύρος των τιμών των δεδομένων που δόθηκαν, καθώς η κατανομή τους μπορεί να είναι εντελώς ανομοιόμορφη και διασκορπισμένη “άσχημα” μακριά από την μέση τιμή και την τυπική απόκλιση. Αντιθέτως, με την χρήση της σύνοψης των 5 αριθμών, έχουμε περισσότερες πληροφορίες και έτσι υπάρχει μια γενικότερη εικόνα για την κάθε κατανομή που μας απασχολεί.

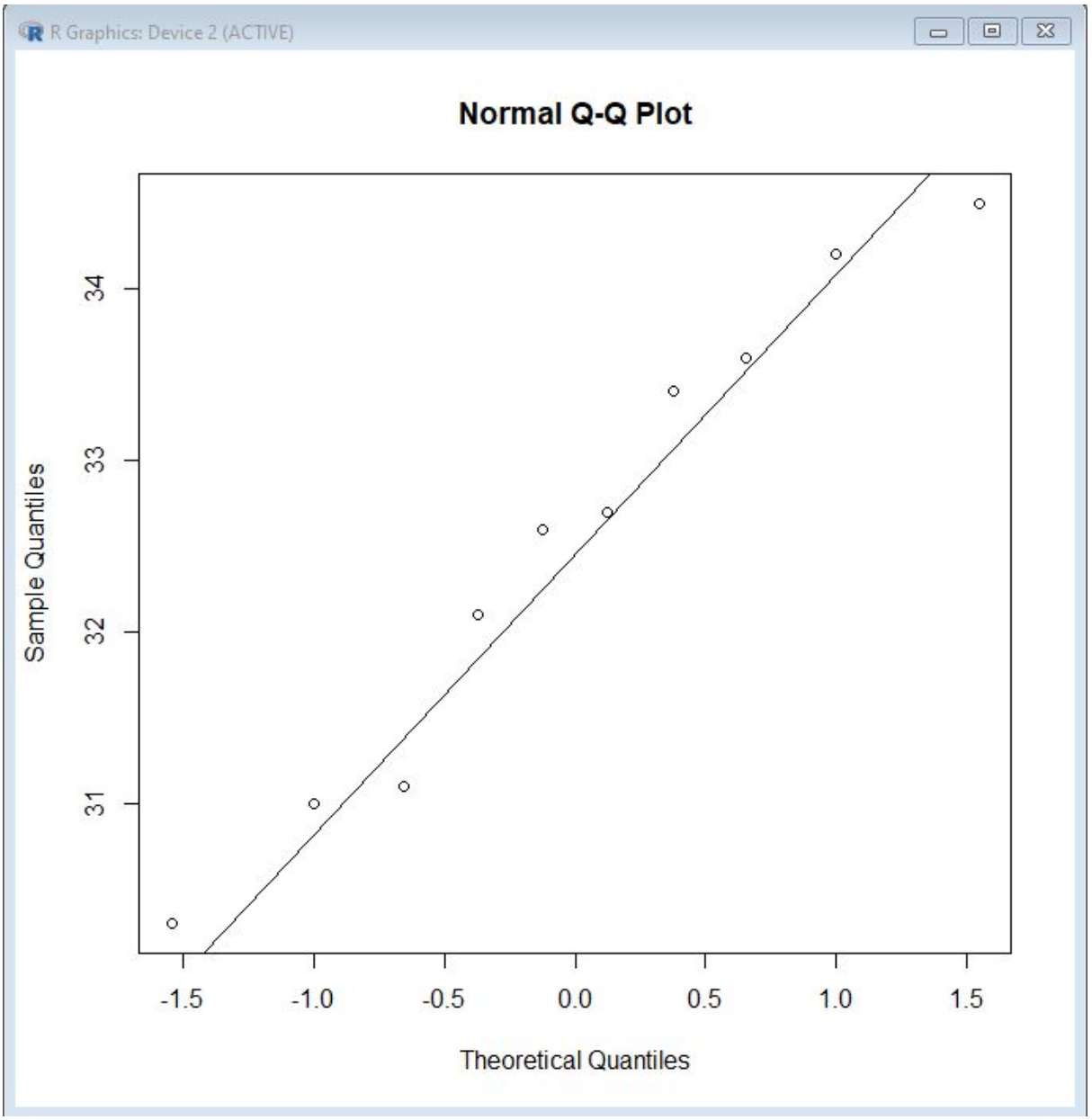
c)

Δεδομένα Ι:

Το ιστόγραμμα της κατανομής των δεδομένων από μια καμπύλη πυκνότητας της Κανονικής κατανομής είναι το παρακάτω:



Μέσω της γραφικής παράστασης της σύγκρισης κατανομής με κανονική καμπύλη πυκνότητας θα είναι πιο φανερά τα παραδείγματα εγγύτητας ή απόκλισης των ποσοστημορίων τους. Παρακάτω δίνεται το διάγραμμα:

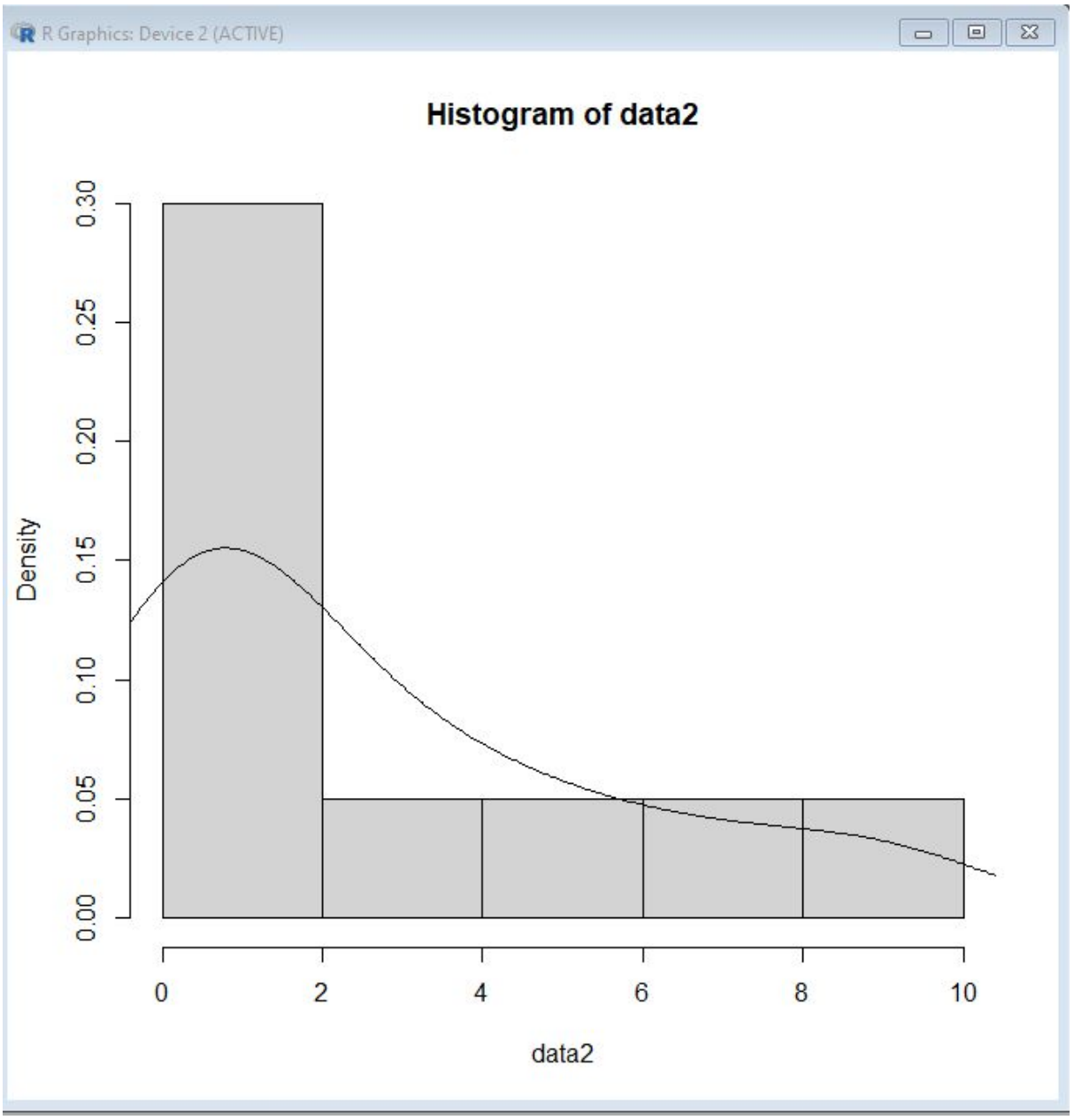


Από τον κανόνα 68-95-99.7, γνωρίζουμε ότι στην κανονική κατανομή ισχύουν τα εξής:

- 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)

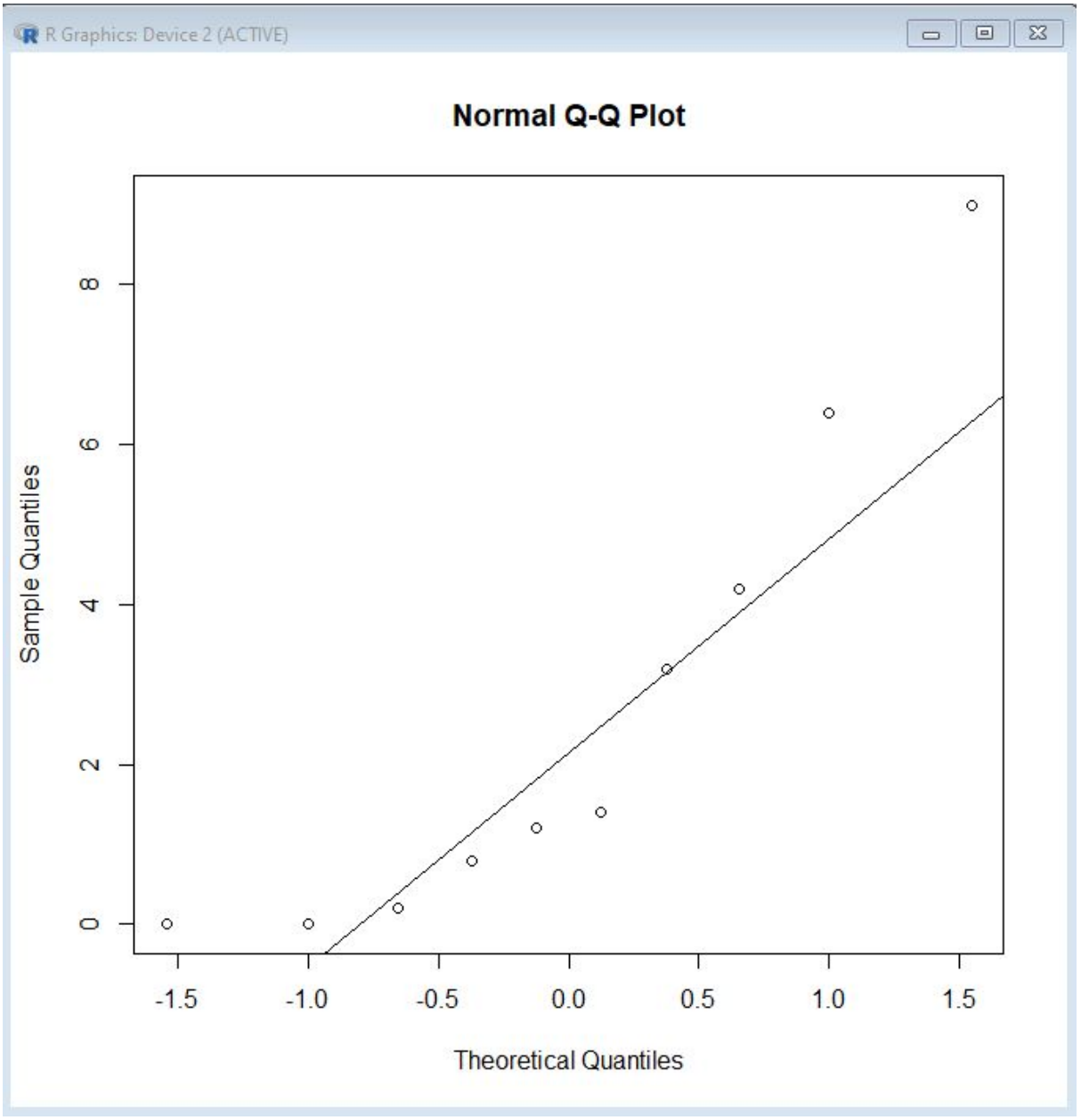
όπου μ = μέση τιμή και σ = τυπική απόκλιση.
Με την βοήθεια των εντολών mean(x) και sd(x) μπορούμε να υπολογίσουμε την μέση τιμή = 32.55 και την τυπική απόκλιση = 1.419898. Παρατηρούμε ότι το ποσοστό των τιμών που βρίσκονται στο διάστημα (0.3, 0.8), δηλαδή το αντίστοιχο (μ – σ, μ + σ) για τα δικά μας δεδομένα, είναι 50% και όχι 68% που είναι το αναμενόμενο. Έτσι, μπορούμε να συμπεράνουμε ότι η προσέγγιση της κατανομής των δεδομένων μας από μια καμπύλη πυκνότητας της Κανονικής Κατανομής θα παρουσιάζει σημαντικές αποκλίσεις.
Δεδομένα ΙΙ:

Το ιστόγραμμα της κατανομής των δεδομένων από μια καμπύλη πυκνότητας της Κανονικής κατανομής είναι το παρακάτω:



Μέσω της γραφικής παράστασης της σύγκρισης κατανομής με κανονική καμπύλη πυκνότητας θα είναι πιο φανερά τα παραδείγματα εγγύτητας ή απόκλισης των

ποσοστημορίων τους. Παρακάτω δίνεται το διάγραμμα:



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

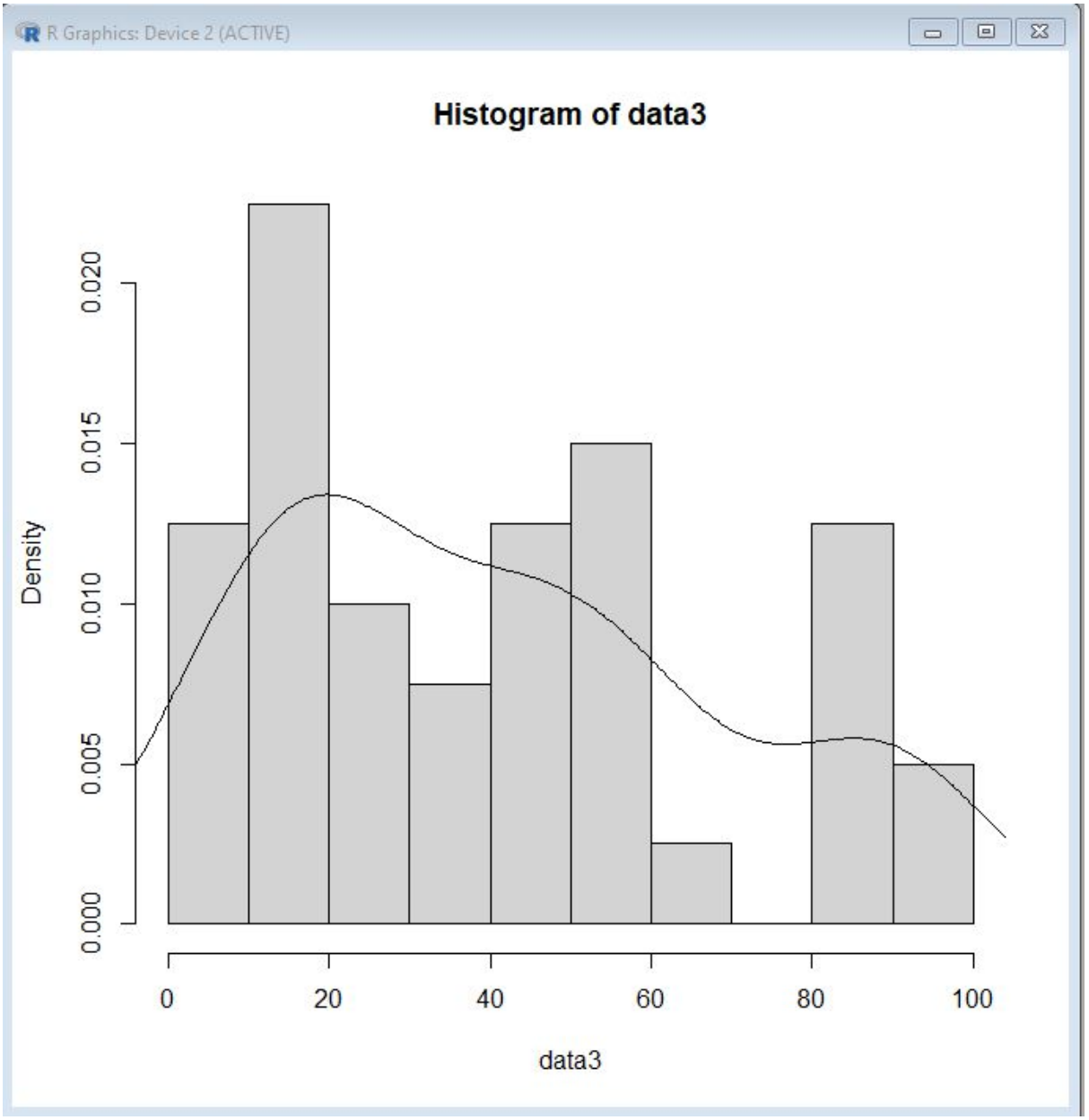
- το 68% των παρατηρήσεων βρίσκεται στο διάστημα (μ – σ, μ + σ)
- το 95% των παρατηρήσεων βρίσκεται στο διάστημα (μ – 2σ, μ + 2σ)
- το 99,5% των παρατηρήσεων βρίσκεται στο διάστημα (μ – 3σ, μ + 3σ)

όπου μ = μέση τιμή και σ = τυπική απόκλιση.

Με την βοήθεια των εντολών mean(x) και sd(x) μπορούμε να υπολογίσουμε την μέση τιμή = 2.64 και την τυπική απόκλιση = 3.059121. Παρατηρούμε ότι το ποσοστό των τιμών που βρίσκονται στο διάστημα (0, 0.8), δηλαδή το αντίστοιχο (μ – σ, μ + σ) για τα δικά μας δεδομένα, είναι 80% και όχι 68% που είναι το αναμενόμενο. Έτσι, μπορούμε να συμπεράνουμε ότι η προσέγγιση της κατανομής των δεδομένων μας από μια καμπύλη πυκνότητας της Κανονικής Κατανομής θα παρουσιάζει σημαντικές αποκλίσεις.

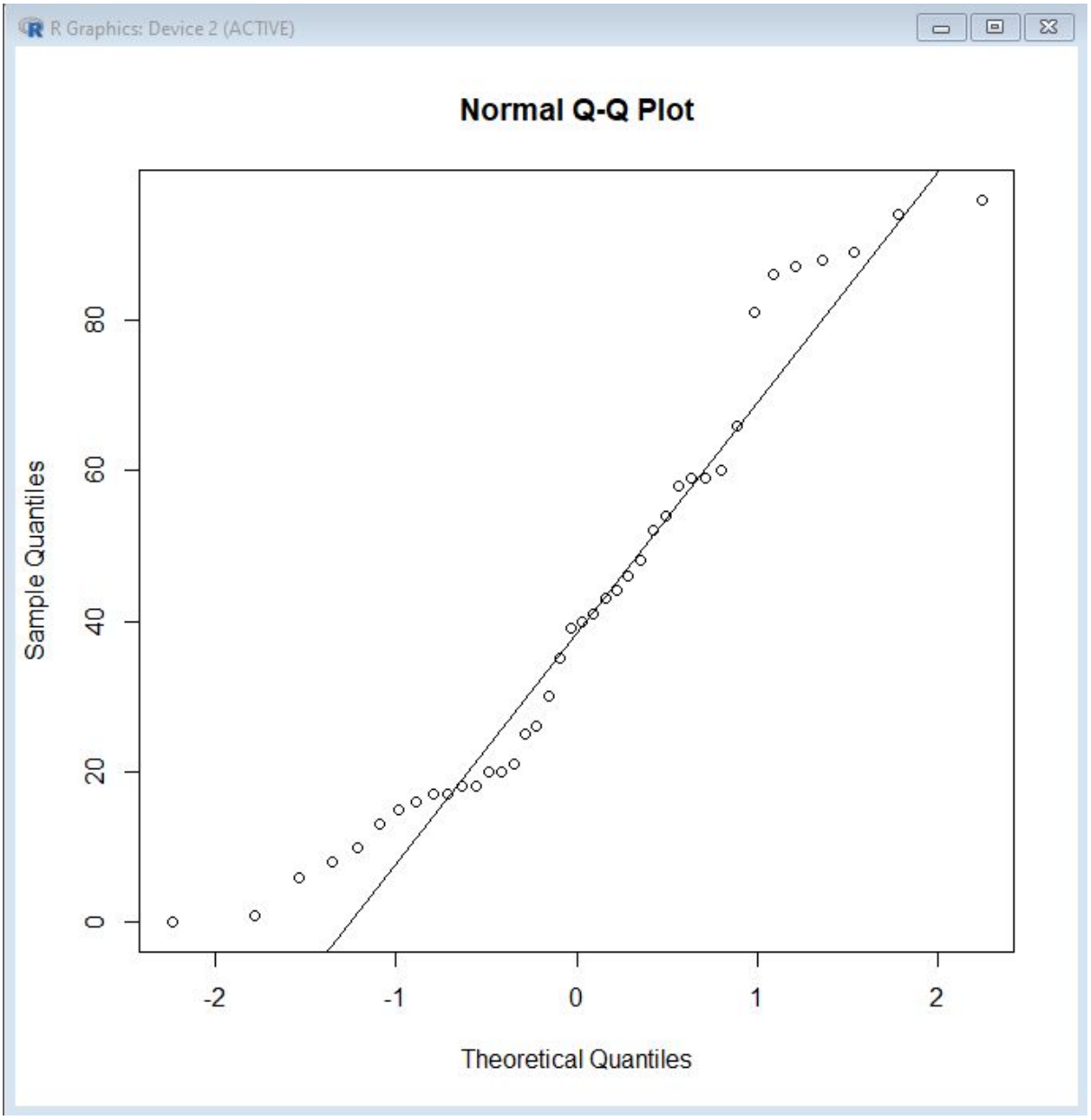
Δεδομένα III:

Το ιστόγραμμα της κατανομής των δεδομένων από μια καμπύλη πυκνότητας της Κανονικής κατανομής είναι το παρακάτω:



Μέσω της γραφικής παράστασης της σύγκρισης κατανομής με κανονική καμπύλη πυκνότητας θα είναι πιο φανερά τα παραδείγματα εγγύτητας ή απόκλισης των

ποσοστημορίων τους. Παρακάτω δίνεται το διάγραμμα:



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.

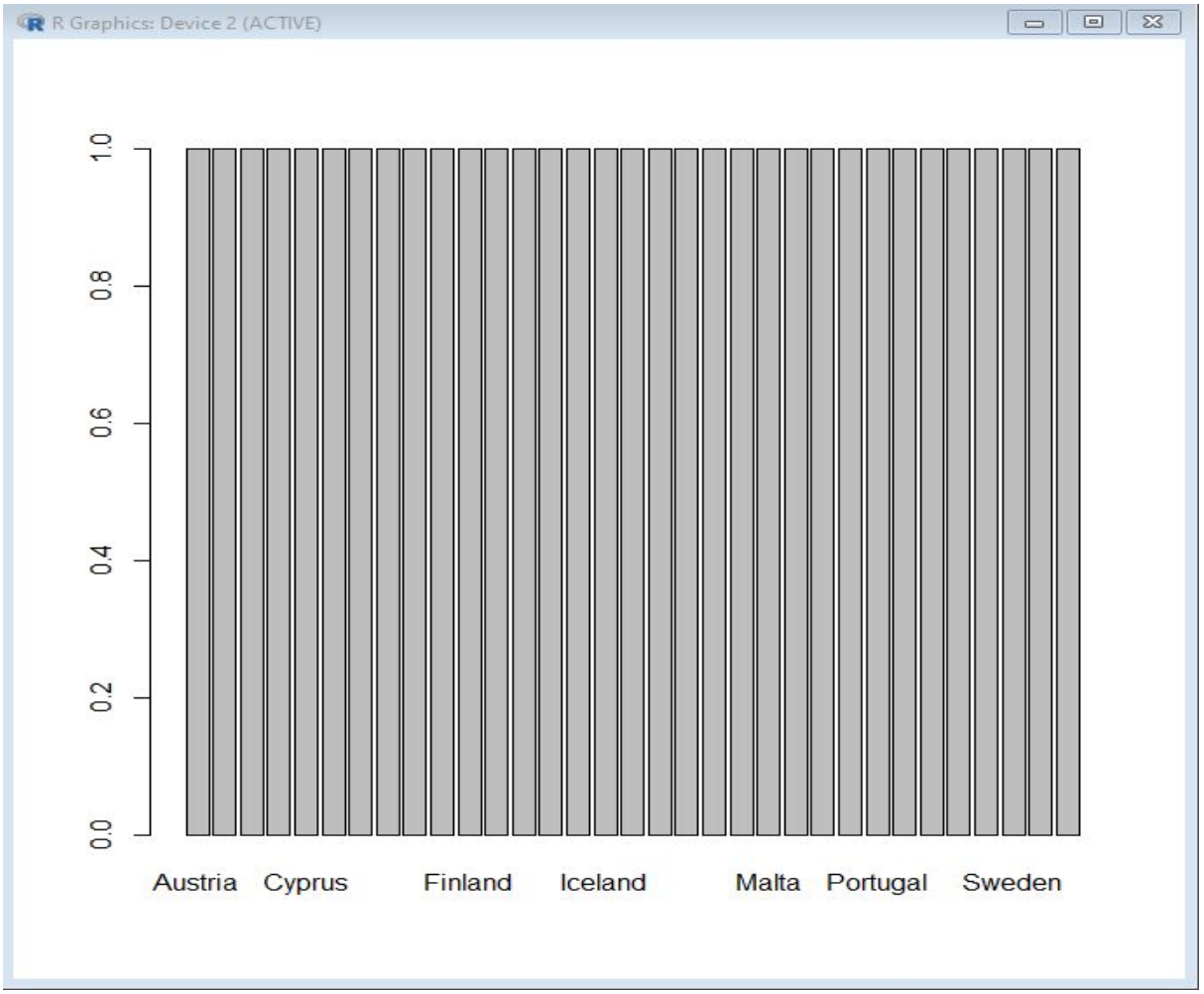
Με την βοήθεια των εντολών mean(x) και sd(x) μπορούμε να υπολογίσουμε την μέση τιμή = 41.15 και την τυπική απόκλιση = 28.26754. Παρατηρούμε ότι το ποσοστό των τιμών που βρίσκονται στο διάστημα (0.125, 0.825), δηλαδή το αντίστοιχο (μ – σ, μ + σ) για τα δικά μας δεδομένα, είναι 70% και όχι 68% που είναι το αναμενόμενο. Έτσι, μπορούμε να συμπεράνουμε ότι η προσέγγιση της κατανομής των δεδομένων μας από μια καμπύλη πυκνότητας της Κανονικής Κατανομής θα παρουσιάζει σημαντικές αποκλίσεις.

Άσκηση 2η :

a) Τα δεδομένα προέρχονται από το site Statistics | Eurostat , το οποίο είναι το site της Ευρωπαϊκής Στατιστικής Υπηρεσίας (EUROSTAT). Τα δεδομένα που χρησιμοποιήθηκαν αφορούν τα πόσα χρόνια ζει ένας άνθρωπος, κατά μέσο όρο ανά χώρα, σύμφωνα με τις στατιστικές μελέτες που έχουν γίνει από το 2007 μέχρι το 2018.

b) Ο πίνακας που δημιουργείται από τα παραπάνω δεδομένα περιέχει μία κατηγορική μεταβλητή την *Country* που περιέχει τα ονόματα των χωρών της Ευρωπαϊκής Ένωσης. Όλες οι υπόλοιπες μεταβλητές είναι ποσοτικές και έχουν ως όνομα τις χρονολογίες *2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018* που περιέχουν τον μέσο όρο προσδόκιμου ζωής των ανθρώπων για την κάθε χώρα.

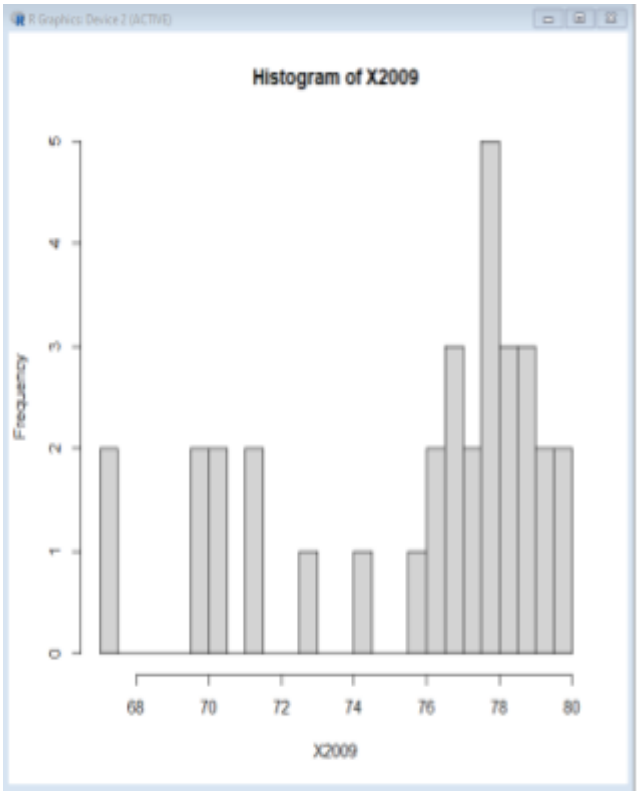
c) Η γραφική παράσταση κατανομής της μεταβλητής *Country* είναι:

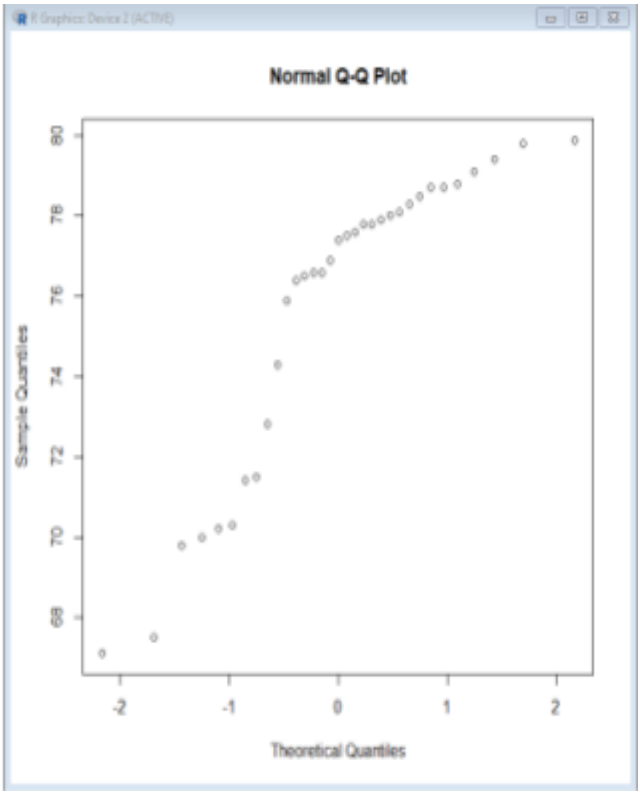


Επειδή η μεταβλητή *Country* είναι κατηγορική, η γραφική της παράσταση κατανομής της οπτικοποιείται με ραβδόγραμμα.

Για τις ποσοτικές μεταβλητές θα παρουσιαστούν ενδεικτικά κάποιες από τις κατανομές τους σε γραφική μορφή. Συγκεκριμένα θα παρουσιαστουν οι γραφικές παραστάσεις των μεταβλητών 2009, 2013, 2018.

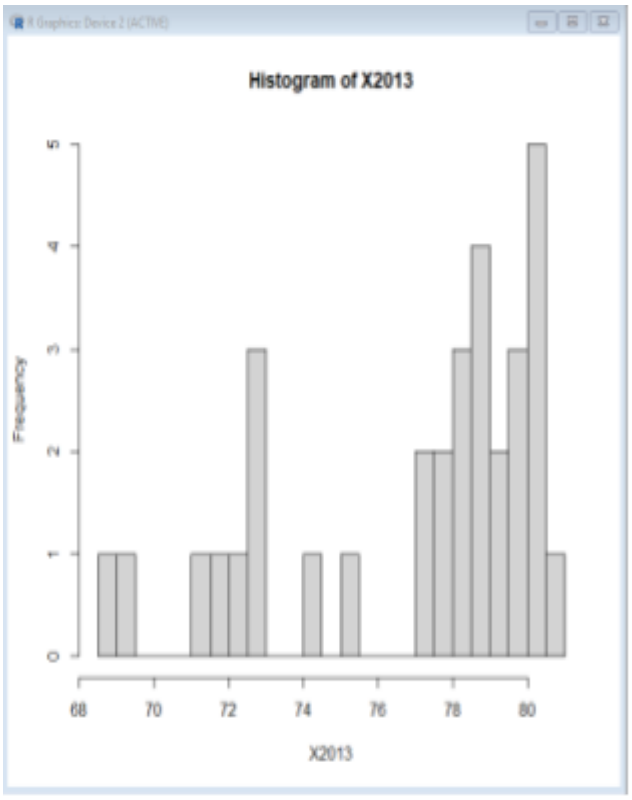
Η γραφική παράσταση κατανομής της μεταβλητής *2009* είναι:

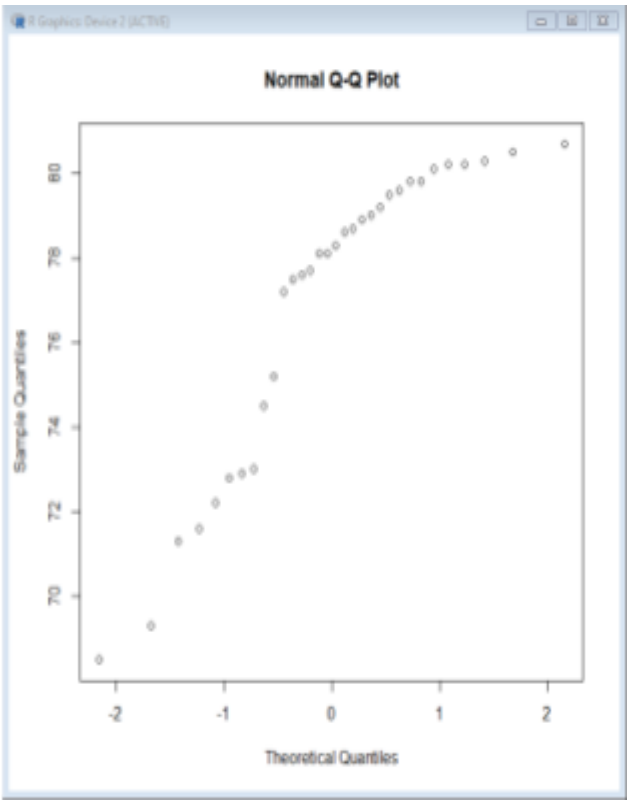




Επειδή η μεταβλητή 2009 είναι ποσοτική η γραφική της παράσταση είναι ιστόγραμμα.

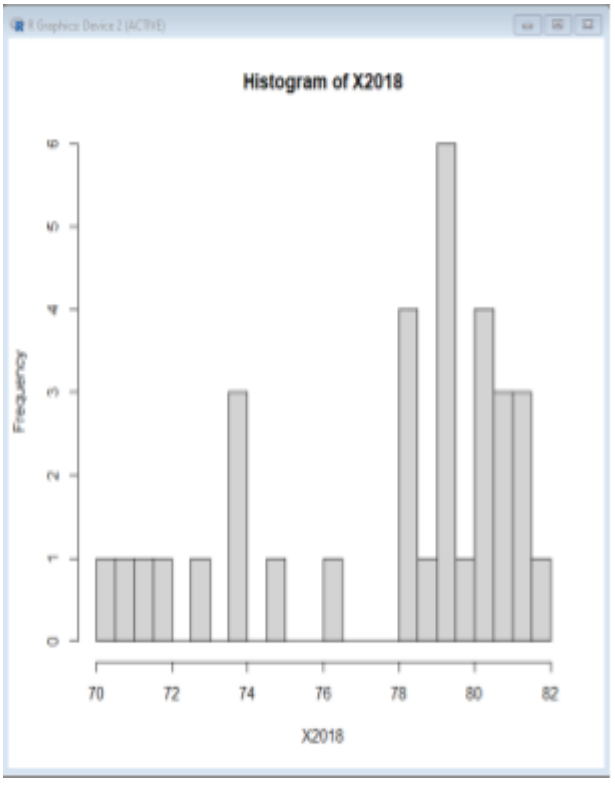
Η γραφική παράσταση κατανομής της μεταβλητής *2013* είναι:

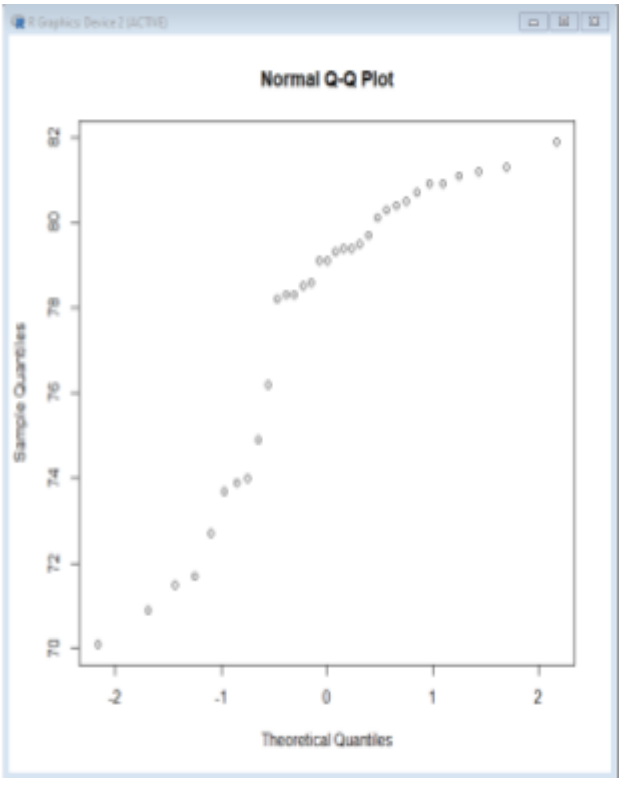




Επειδή η μεταβλητή 2013 είναι ποσοτική η γραφική της παράσταση είναι ιστόγραμμα.

Η γραφική παράσταση κατανομής της μεταβλητής *2018* είναι:





Επειδή η μεταβλητή 2018 είναι ποσοτική η γραφική της παράσταση είναι ιστόγραμμα.

Στα παραπάνω σχεδιαγράμματα δεν παρατηρούμε την ύπαρξη ατυπικών σημείων. Παρατηρείται επίσης ότι με την πάροδο του χρόνου το προσδόκιμο ζωής αυξάνεται, κάτι που μπορεί να οφείλεται στην ανάπτυξη της ιατρικής και της τεχνολογίας.

d) Η μεταβλητή 2007 έχει μέση τιμή = 74.91212 και τυπική απόκλιση = 4.171687 και η σύνοψη των 5 αριθμών της είναι min = 64.5, Q1= 72.2, διάμεση τιμή m = 76.7, Q3= 77.6 και max = 79.6.

Η μεταβλητή 2008 έχει μέση τιμή = 75.33333 και τυπική απόκλιση = 3.964819 και η σύνοψη των 5 αριθμών της είναι min = 65.9, Q1 = 72.3, διάμεση τιμή m = 76.9, Q3 = 78.1 και max = 80.0

Η μεταβλητή 2009 έχει μέση τιμή = 75.6697 και τυπική απόκλιση = 3.77065 και η σύνοψη των 5 αριθμών της είναι min = 67.1, Q1 = 72.8, διάμεση τιμή m = 77.4, Q3 = 78.3 και max = 79.9

Η μεταβλητή 2010 έχει μέση τιμή = 75.9375 και τυπική απόκλιση = 3.740342 και η σύνοψη των 5 αριθμών της είναι min = 67.60, Q1 = 72.80, διάμεση τιμή m = 77.35 , Q3 = 78.75 και max = 80.30

Η μεταβλητή 2011 έχει μέση τιμή = 76.40606 και τυπική απόκλιση = 3.630336 και η σύνοψη των 5 αριθμών της είναι min = 68.1, Q1 = 73.8, διάμεση τιμή m = 77.9 , Q3 = 79.0 και max = 80.7

Η μεταβλητή 2012 έχει μέση τιμή = 76.45625 και τυπική απόκλιση = 3.615285 και η σύνοψη των 5 αριθμών της είναι min = 68.40, Q1 = 73.25, διαμεση τιμη m = 77.90 , Q3 = 79.00 και max = 81.60

Η μεταβλητή 2013 έχει μέση τιμή = 76.90313 και τυπική απόκλιση = 3.567911 και η σύνοψη των 5 αριθμών της είναι min = 68.50, Q1 = 73.75, διάμεση τιμή m = 78.20, Q3 = 79.70 και max = 80.70

Η μεταβλητή 2014 έχει μέση τιμή = 77.25758 και τυπική απόκλιση = 3.621035 και η σύνοψη των 5 αριθμών της είναι min = 69.1, Q1 = 74.7, διάμεση τιμή m = 78.7, Q3 = 79.9 και max = 81.3

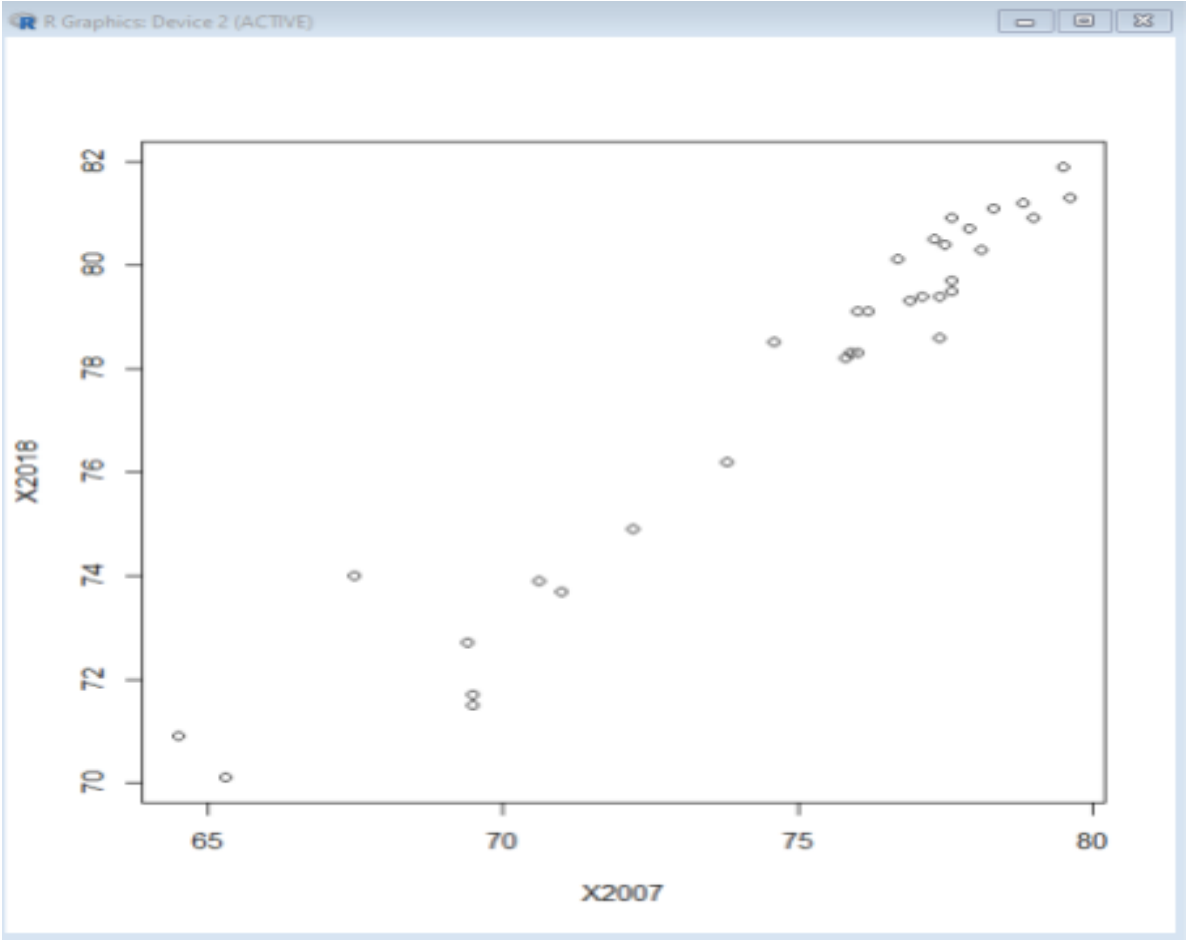
Η μεταβλητή 2015 έχει μέση τιμή = 77.20909 και τυπική απόκλιση = 3.517666 και η σύνοψη των 5 αριθμών της είναι min = 69.2, Q1 = 74.4, διάμεση τιμή m = 78.7, Q3 = 79.9 και max = 81.2

Η μεταβλητή 2016 έχει μέση τιμή = 77.50909 και τυπική απόκλιση = 3.525475 και η σύνοψη των 5 αριθμών της είναι min = 69.5, Q1 = 75.0, διάμεση τιμή m = 78.9, Q3 = 80.1 και max = 81.7

Η μεταβλητή 2017 έχει μέση τιμή = 77.62727 και τυπική απόκλιση = 3.466561 και η σύνοψη των 5 αριθμών της είναι min = 69.8, Q1 = 74.9, διάμεση τιμή m = 78.9, Q3 = 80.2 και max = 81.6

Η μεταβλητή 2018 έχει μέση τιμή = 77.76667 και τυπική απόκλιση = 3.495414 και η σύνοψη των 5 αριθμών της είναι min = 70.1, Q1 = 74.9, διάμεση τιμή m = 79.1 , Q3 = 80.4 και max = 81.9

e) Επιλέγω να συγκρίνω τις μεταβλητές 2007, 2018. Θα θεωρήσω ότι το έτος 2007 καθορίζει το έτος 2018, επομένω το 2007 θα είναι η επεξηγηματική μεταβλητή και το 2018 είναι η μεταβλητή απόκρισης. Παρακάτω παρουσιάζεται το scatterplot για τη σχέση 2007 και 2018:

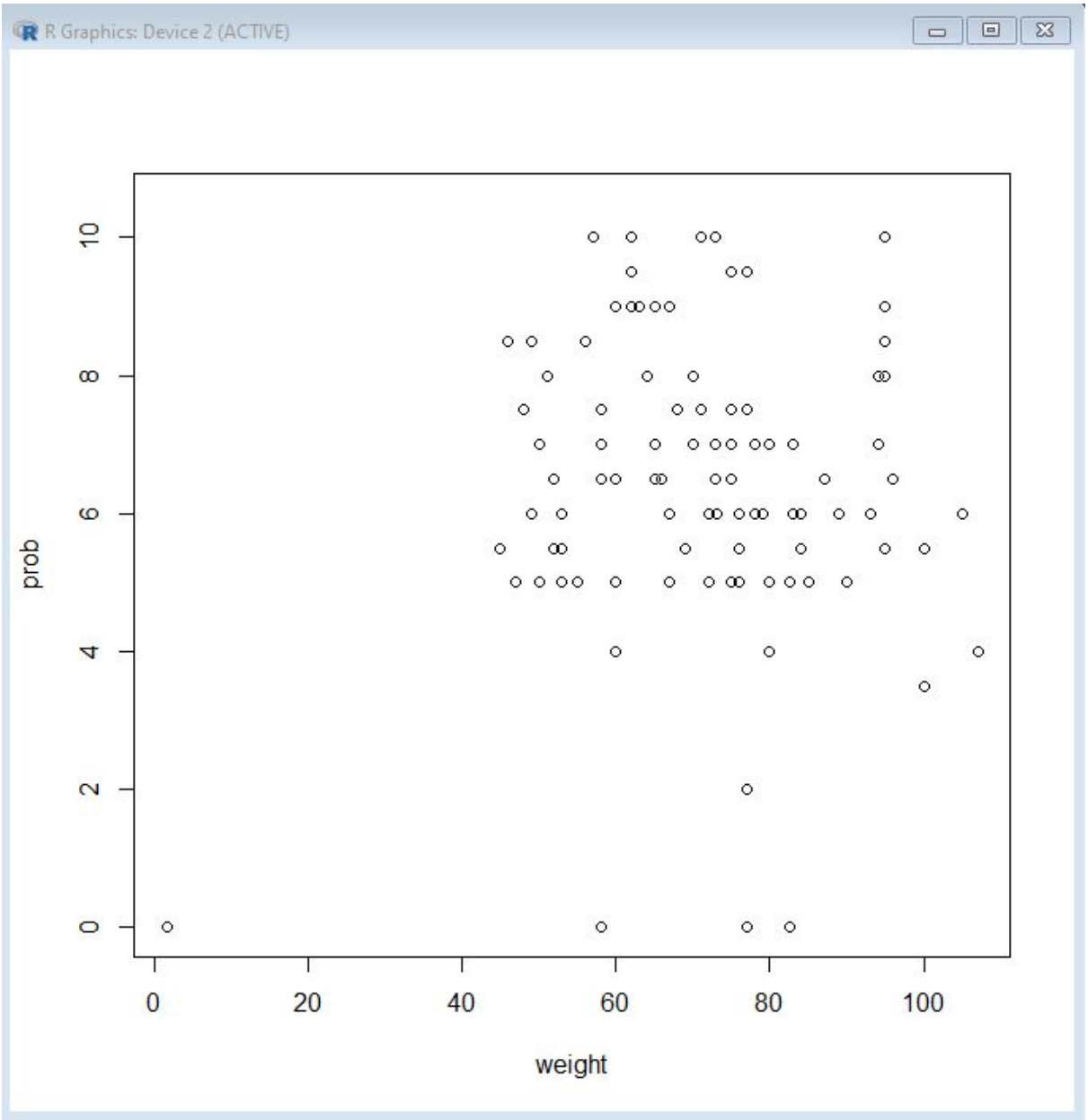


Όπως φαίνεται στο διάγραμμα η σχέση είναι θετική και ισχυρή γραμμική, καθώς ο συντελεστής συσχέτισης είναι r = 0.9701396.

Άσκηση 3η :

Για την άσκηση αυτή θα επιλέξω να συγκρίνω τις ποσοτικές μεταβλητές *weight* και *prob* των δεδομένων που μας δίνονται απο τις απαντήσεις του ερωτηματολογίου 2020. Θα θεωρήσω ότι το βάρος καθορίζει το βαθμό στο μάθημα των Πιθανοτήτων στον κάθε άνθρωπο, επομένως το βάρος(*weight*) θα είναι η επεξηγηματική μεταβλητή και ο βαθμός στο μάθημα των Πιθανοτήτων(*prob*) η μεταβλητή απόκρισης.

a) Το scatterplot για της σχέσης weight και prob είναι το παρακάτω:



Σχόλια:

- Για την μορφή: Όπως παρατηρούμε από το scatterplot η μορφή της σχέσης των δύο μεταβλητών είναι συμβατή με γραμμική μορφή, αν και είναι ασθενής. Επιπλέον παρατηρώ ότι υπάρχουν 4 ατυπικά σημεία με μηδενικό βαθμό στο μάθημα των Πιθανοτήτων. Ο λόγος που συμβαίνει αυτό είναι το γεγονός ότι υπάρχει μια ασάφεια στο ερωτηματολόγιο για το αν θα πρέπει να συμπληρωθεί ο βαθμός του φοιτητή στις Πιθανότητες ακόμα και αν δεν το έχει περάσει/δώσει.
- Για την κατεύθυνση: Η κατεύθυνση της σχέσης φαίνεται να είναι αύξουσα.
- Δύναμη της σχέσης: Η δύναμη της σχέσης είναι αρκετά ασθενής.

b) Ο συντελεστής συσχέτισης της σχέσης height και shoe είναι r = 0.03003128 που μας επιβεβαιώνει την πολύ ασθενή και αύξουσα γραμμική συσχέτιση που παρατηρείται.

Η γραμμική παλινδρόμηση ελαχίστων τετραγώνων είναι η παρακάτω: