1η Σειρά Ασκήσεων(2021-2022)

Φοιτητές:
Ειρήνη Χρυσικοπούλου(3180208)
Παναγιώτης Παναγιώτου(3180139)

Άσκηση 1η:

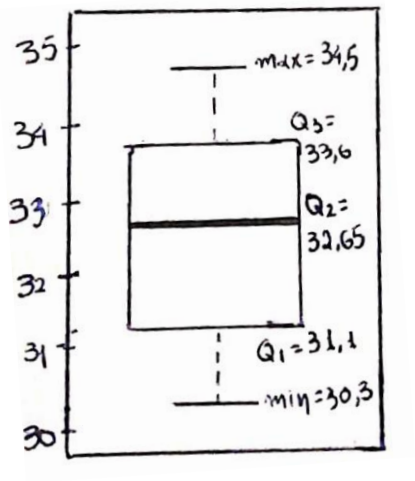
a.Τα stemplot και boxplot για τα δεδομένα που μας δίνονται είναι:

Δεδομένα I:

Stemplot I

30 | 3
31 | 01
32 | 167
33 | 46
34 | 25

Boxplot I



5 number summary για εύρεση boxplot:

Min=30.3

Q1=31.1

Q2=(32.7+32.6)/2=32.65

Q3=33.6

Max=34.5

Έλεγχος για ατυπικά σημεία:

IQR=Q3-Q1=2.5

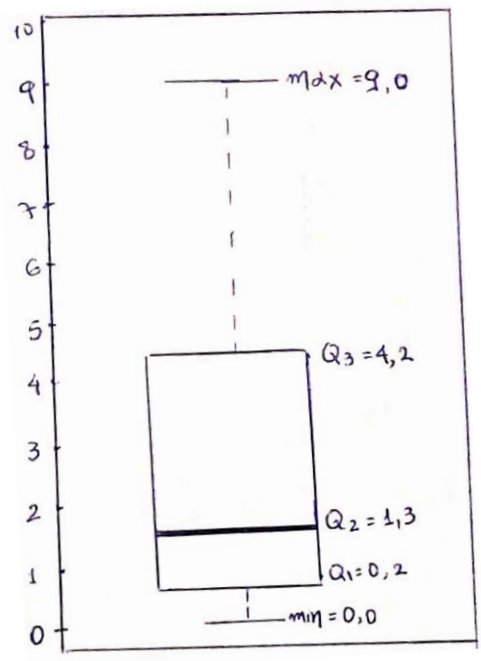
Δεν υπάρχουν τιμές μικρότερες του Q1-1.5*IQR=27.35 ούτε μεγαλύτερες από Q3+1.5*IQR=37.35,συνεπώς δεν υπάρχουν ατυπικά σημεία.

Δεδομένα ΙΙ:

Stemplot II

0-1 | 002824
2-3 | 2
4-5 | 2
6-7 | 4
8-9 | 0

Boxplot II



5 number summary για εύρεση boxplot:

Min=0.0

Q1=0.2

Q2=(1.2+1.4)/2=1.3

Q3=4.2

Max=9.0

Έλεγχος για ατυπικά σημεία:

IQR=Q3-Q1=4

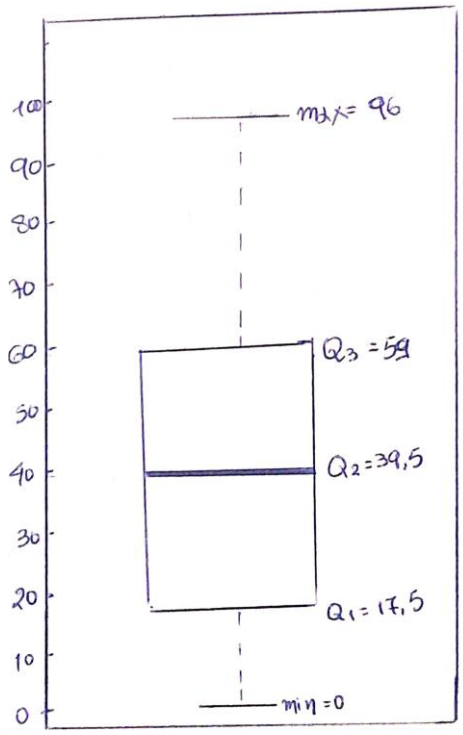
Δεν υπάρχουν τιμές μικρότερες του Q1-1.5*IQR=-5.8 ούτε μεγαλύτερες από Q3+1.5*IQR=10.2,συνεπώς δεν υπάρχουν ατυπικά σημεία.

Δεδομένα III:

Stemplot III

0|0168
1|03567788
2|00156
3|059
4|013468
5|24899
6|06
7
8|16789
9|46

Boxplot III



5 number summary για εύρεση boxplot:

Min=0

Q1=(17+18)/2=17.5

Q2=(39+40)/2=39.5

Q3=(59+59)/2=59

Max=96

Έλεγχος για ατυπικά σημεία:

IQR=Q3-Q1=41.5

Δεν υπάρχουν τιμές μικρότερες του Q1-1.5*IQR=-44.75 ούτε μεγαλύτερες από Q3+1.5*IQR=121.25,συνεπώς δεν υπάρχουν ατυπικά σημεία.

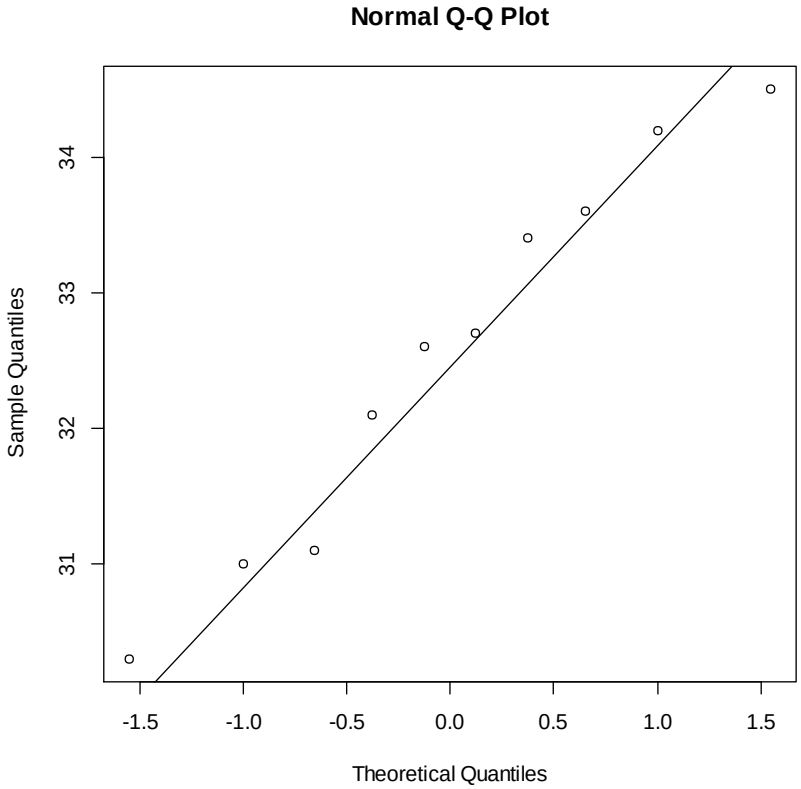
b.Για συμμετρικές κατανομές και κατανομές χωρίς ισχυρά ατυπικά σημεία προτιμάμε την μέση τιμή και την τυπική απόκλιση για την σύνοψη της κατανομής, ενώ σε αντίθετη περίπτωση την σύνοψη των 5 αριθμών.
Για την πρώτη ομάδα δεδομένων προτιμάμε την μέση τιμή και την τυπική απόκλιση καθώς δεν υπάρχουν ατυπικές τιμές και οι τιμές είναι συμμετρικά κατανεμημένες γύρω απο την μέση τιμή.Η μέση τιμή είναι 32.55 και η τυπική απόκλιση 1.419898.

Για την δεύτερη ομάδα δεδομένων προτιμάμε την σύνοψη των 5 αριθμών.Παρότι δεν υπάρχουν ατυπικές τιμές, οι τιμές δεν είναι συμμετρικά κατανεμημένες συνεπώς η σύνοψη των 5 αριθμών περιγράφει καλύτερα την κατανομή τους. Όπως φαίνεται στο boxplot παρατηρούμε μια

συγκέντρωση των τιμών προς τα χαμηλότερα ποσοστημόρια. Η μέση τιμή είναι 2.64 και η τυπική απόκλιση είναι 3.059121.
Για την τριτη ομάδα δεδομένων: Για την περιγραφή της κατανομής προτιμάμε την σύνοψη των 5 αριθμών.Παρόλο που δεν υπάρχουν ούτε εδώ ατυπικά σημεία,παρατηρούμε ότι οι τιμές δεν είναι συμμετρικά κατανεμημένες(συγκέντρωση τιμών στα χαμηλότερα ποσοστομόρια).Η μέση τιμή είναι 41.15 και η τυπική απόκλιση 28.26754.
c.

Για κάθε ομάδα δεδομένων θα σχεδιάσουμε το Normal Quantile Plot και θα ελέγξουμε κατά πόσο τα σημεία (x_i,y_i),όπου x_i=ποσοστημόριο τυπικής κανονικής κατανομής και y_i i-οστή μικρότερη τιμή της ομάδας δεδομένων,είναι συγραμμικά.

I



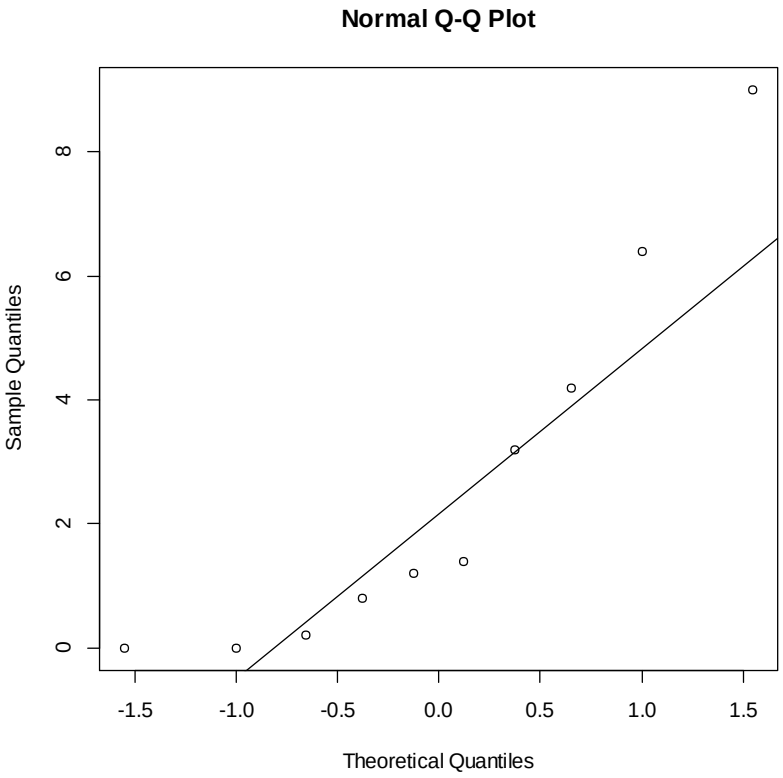
Παρατηρούμε ότι τα δεδομένα βρίσκονται αρκετά κοντά στην ευθεία,γεγονός που σημαίνει ότι η καμπύλη πυκνότητας της κανονικής κατανομής προσεγγίζει επαρκώς την κατανομή των δεδομένων.Η τυπική απόκλιση είναι s=1.419898 και ο μέσος m=32.55.Μπορούμε να ελέγξουμε αν ισχύει ο κανόνας 68-95-99.7 που ισχύει για την κανονική κατανομή.

Στο διάστημα [m-s,m+s] βρίσκεται το 50% των τιμών

Στο διάστημα [m-2*s,m+2*s] βρίσκεται το 100% των τιμών.

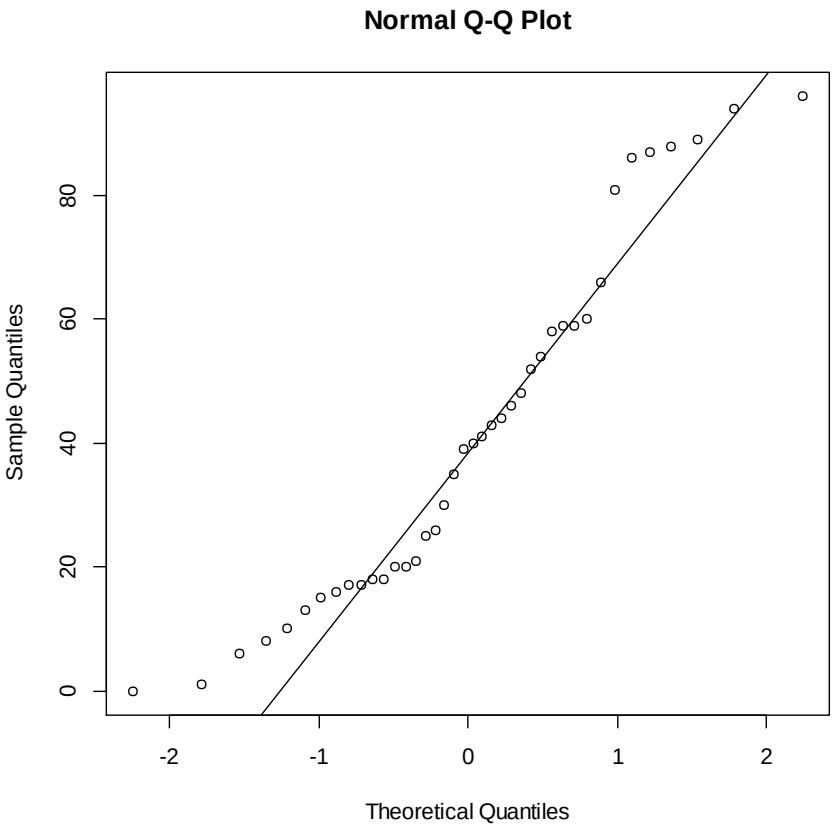
Παρατηρούμε ότι υπάρχει μια απόκλιση σε αυτόν τον κανόνα,επομένως θα υπάρχει μικρή απόκλιση στην προσέγγιση από καμπύλη κανονικής κατανομής.

II.



Παρατηρούμε ότι στα ενδιάμεσα ποσοστημόρια η κατανομή των δεδομένων προσεγγίζεται επαρκώς την καμπύλη πυκνότητας της κανονικής κατανομή.Όμως στα ακραία ποσοστημόρια τα δεδομένα αποκλείνουν πολύ απο την ευθεία ,επομένως δεν θα ήταν ακριβής η προσέγγιση κατανομής των δεδομένων από μια καμπύλη πυκνότητας κανονικής κατανομής.

III



Και εδώ παρατηρούμε ότι στα ενδιάμεσα ποσοστημόρια τα δεδομένα προσεγγίζονται από την καμπύλη πυκνότητας της κανονικής κατανομής επαρκώς.Όμως αυτό δεν ισχύει για τα ακραία ποσοστημόρια όπου διαπιστώνονται μεγάλες αποκλήσεις των σημείων από την ευθεία.Συνεπώς ούτε για αυτά τα δεδομένα θα ήταν ακριβής η προσέγγιση κατανομής των δεδομένων από μια καμπύλη πυκνότητας κανονικής κατανομής.

Άσκηση 2η:

a.

Τα δεδομένα προέρχονται από δημοσιευμένη έρευνα("HarvardX and MITx: Four Years of Open Online Courses") των Harvard και MIT σχετικά με την δημοτικότητα εξ αποστάσεως σύντομων εκπαιδευτικών προγραμμάτων(short courses).Περιέχονται συνολικά 290 περιπτώσεις.

b.

Κατηγορικές Μεταβλητές:

- Institution:Ίδρυμα παροχής του μαθήματος
- Course Title:Τίτλος μαθήματος
- Course Subject:Αντικείμενο μαθήματος

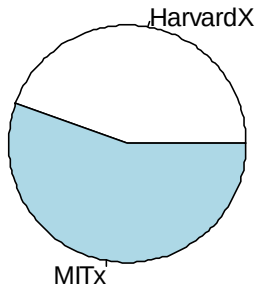
Ποσοτικές Μεταβλητές:

- Year: Χρόνος ολοκλήρωσης μαθήματος
- Participants Course Content Accessed: Αριθμός συμμετεχόντων με πρόσβαση στο υλικό του μαθήματος
- %Certified: Ποσοστό συμμετεχόντων που πήραν πιστοποίηση παρακολούθησης
- Median Age: Μέση ηλικία συμμετεχόντων
- % Male: Ποσοστό ανδρών συμμετεχόντων
- % Female: Ποσοστό γυναικών συμμετεχόντων
- % Bachelor's Degree or Higher:Ποσοστό συμμετεχόντων με πτυχίο Bachelor ή ανώτερο.

c.

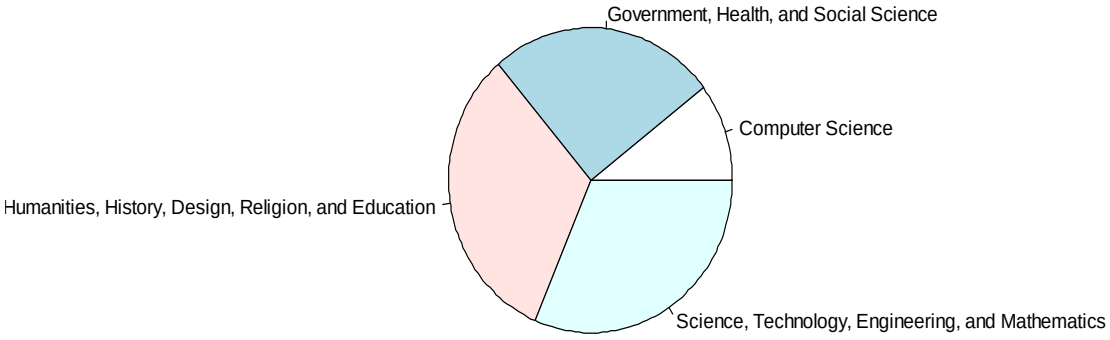
Για τις κατηγορικές μεταβλητές θα παρουσιάσουμε μόνο την κατανομή των Institution και Course Subject.

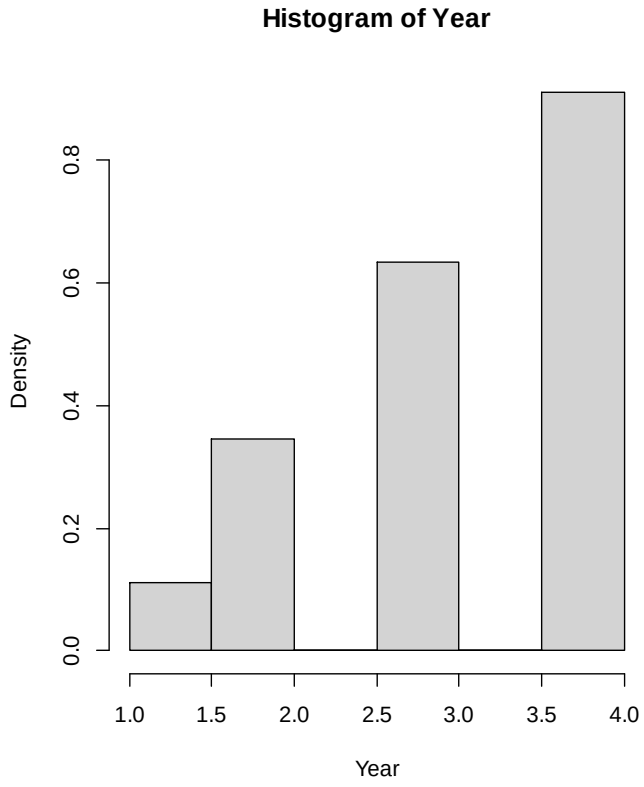
Κατανομή της κατηγορικής μεταβλητής Institution:



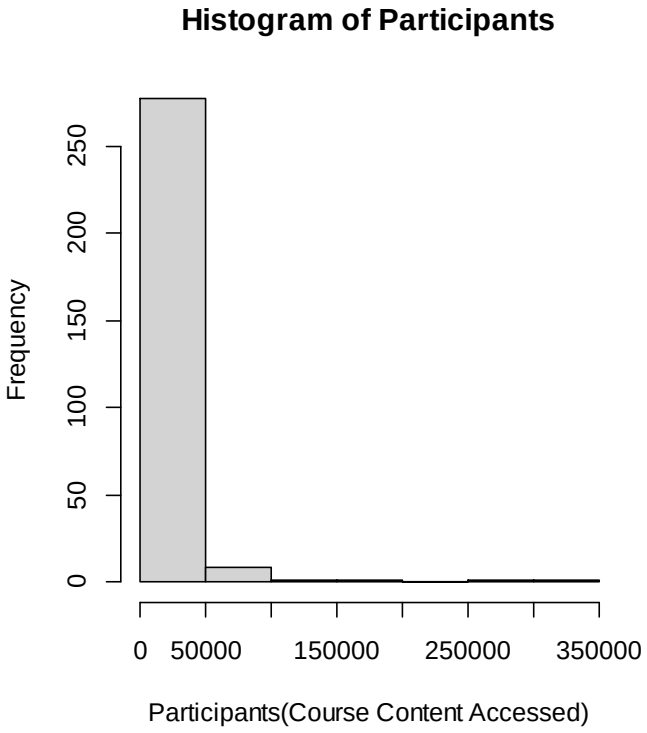
Εδώ παρατηρούμε ότι το MIT(161 από τα 290) προσφέρει περισσότερα online μαθήματα από το Harvard.

Κατανομή της κατηγορικής μεταβλητής Course Subject.Εδώ βλέπουμε σε ποιους τομείς αφορούν τα μαθήματα. Οι δημοφιλέστρεροι κλάδοι ήταν οι ανθρωπιστικές επιστήμες και οι επιστήμες τεχνολογίας.

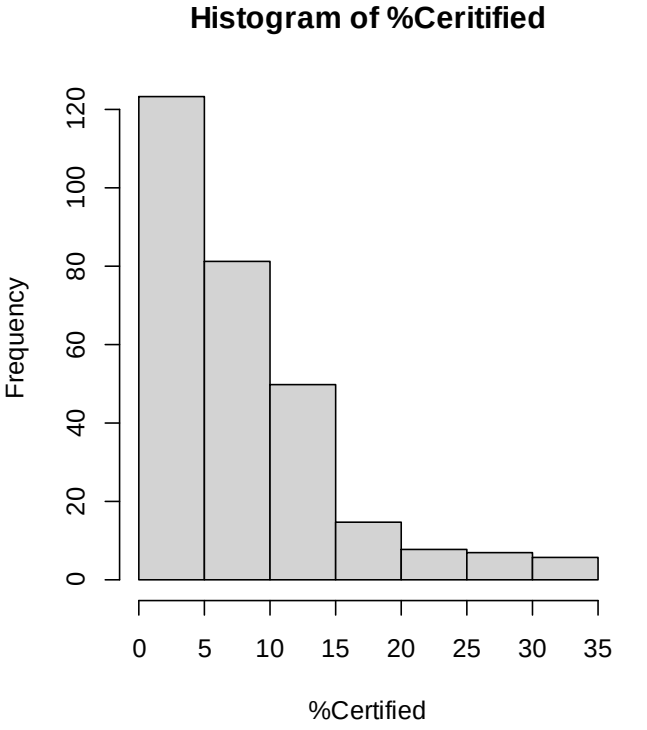




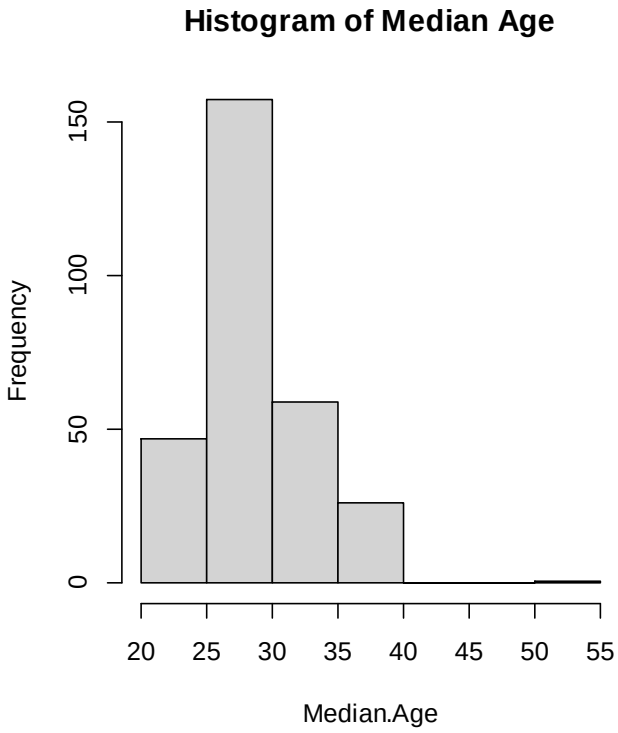
Κατανομή της ποσοτικής μεταβλητής Year. Παρατηρούμε ότι τα περισσότερα μαθήματα θέλουν 2,5-4 χρόνια για να ολοκληρωθούν.



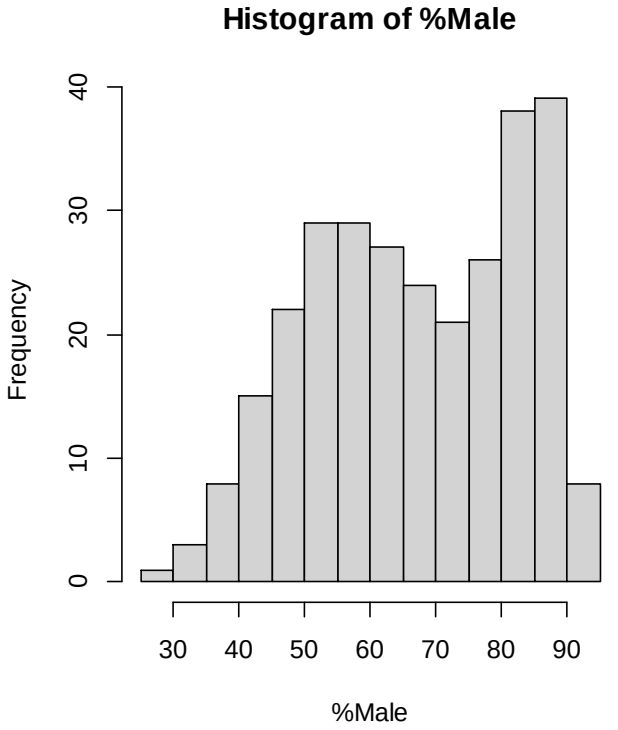
Διάγραμμα κατανομής της μεταβλητής Participants Course Content Accessed. Παρατηρούμε οτι γενικά τα περισσότερα μαθήματα έχουν λιγότερους από 5000 εγγεγραμμένους με πρόσβαση στο υλικό τους.Υπάρχουν όμως τέσσερις ισχυρές ατυπικές τιμές.Υπάρχουν δηλαδή τέσσερα μαθήματα με πάρα πολλούς εγγεγραμμένους,γεγονός που ίσως να οφείλεται στην φύση του μαθήματος(πχ αφορά κάποιο πολύ επίκαιρο θέμα).



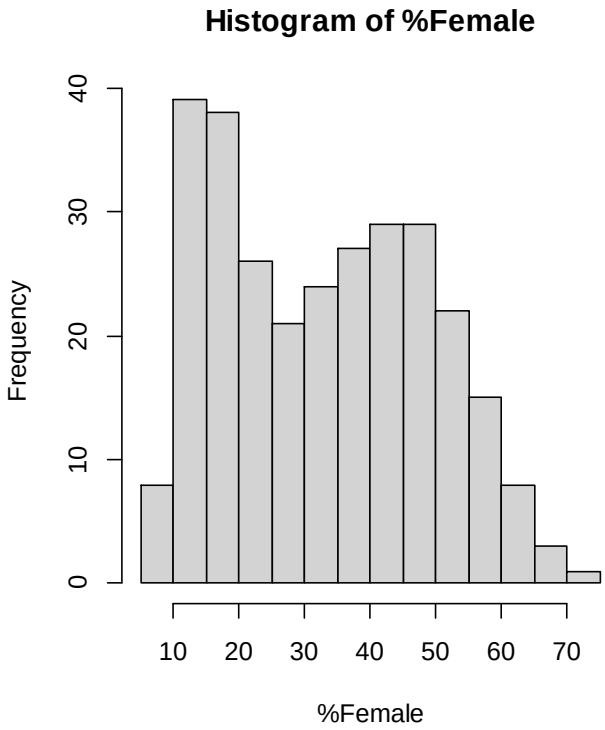
Κατανομή της μεταβλητής %Certified.Παρατηρούμε ότι σε κάθε περίπτωση το ποσοστό των εγγεγραμμέων που παίρνει βεβαίωση συμμετοχής είναι μικρότερο του 35% των αρχικών εγγραμμένων.Η συγκέντρωση των τιμών είναι υψηλότερη στο χαμηλότερο ποσοστημόριο,πράγμα λογικό καθότι είναι σύνθες να εγγράφεται κάποιος σε ένα μάθημα και τελικά να μην φτάνει μέχρι το σημείο να πάρει πιστοποίηση για την παρακολούθηση του. Δεν υπάρχουν ατυπικές τιμές.



Κατανομή μέσης ηλικίας εγγεγραμμένων ανά μάθημα.Παρατηρούμε οτι η κατανομή είναι σχετικά συμμετρική και οι περισσότερες τιμές κατανέμονται γύρω απο τον μέσο.Εξαίρεση αποτελεί ένα ατυπικό σημείο.

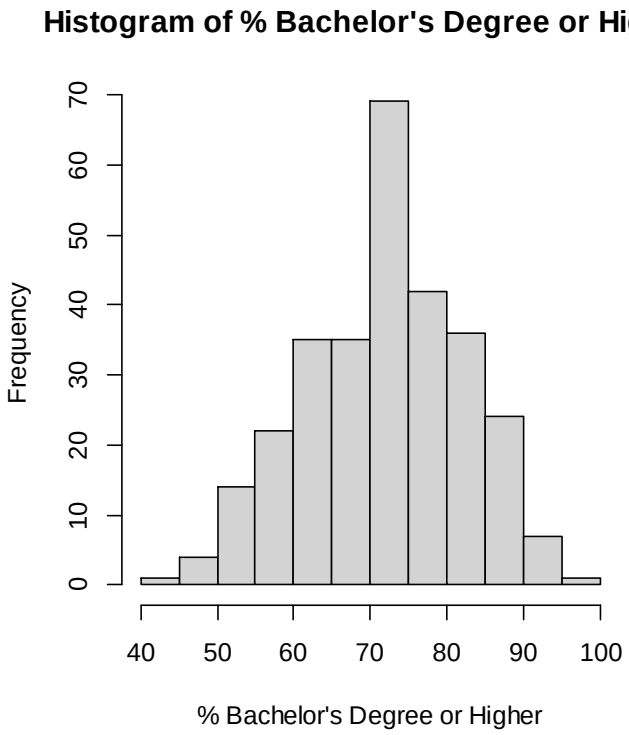


Κατανομή ποσοστού ανδρών μάθημα: Παρατηρούμε ότι τα περισσότερα μαθήματα έχουν ποσοστό εγγεγραμμένων ανδρών>30% (εξαίρεση αποτελεί ένα μάθημα),ενώ σε πολλά μαθήματα αποτελούν την συντριπτική πλειοψηφία(σε περίπου 80 μαθήματα αποτελούν το 80%-90% των συμμετεχόντων.)Η κατανομή των τιμών δεν είναι συμμετρική.



Κατανομή ποσοστού γυναικών μάθημα:Παρατηρούμε ότι στα περισσότερα μαθήματα οι γυναίκες συμμετέχουν σε χαμηλό ποσοστό(σε περίπου 80 μαθήματα αποτελούν ποσοστό 10%-20%).

Και εδώ όπως και στο πάνω ιστόγραμμα οι τιμές δεν είναι συμμετρικά κατανεμημένες



Κατανομή ποσοστού εγγραμμένων με πτυχίο Bachelor ή ανώτερο:Παρατηρούμε οτι τα δεδομένα κατανέμονται σχετικά συμμετρικά γύρω από τον μέσο.Βλέπουμε γενικά οτι σε κάθε μάθημα το ποσοστό εγγραμμένων με πτυχίο Bachelor ή ανώτερο είναι μεγάλο(>40%).

d.

Για την εύρεση του 5 number summary,του μέσου και της τυπικής απόκλισης χρησιμοποιούμε τις εντολες fivenum,mean,sd της r.Για να αξιολογήσουμε την καταλληλότητα της κάθε περιγραφής πρέπει να λάβουμε υπόψιν μας κατά πόσο τα δεδομένα είναι συμμετρικά κατανεμημένα και την ύπαρξη ή οχι ατυπικών σημείων.

Για την μεταβλητή Participants:

> fivenum(Participants..Course.Content.Accessed.)

[1] 322.0 3806.0 7901.5 18183.0 301082.0

Επομένως το 5 number summary είναι:

Min= 322.0	Q1=3806.0	Q2=7901.5	Q3=18183.0	Max=301082.0

Εύρεση μέσου και τυπικής απόκλισης:

> mean(Participants..Course.Content.Accessed.) // Εντολή εύρεσης μέσου

[1] 15344.33 //μέσος

> sd(Participants..Course.Content.Accessed.) // Εντολή τυπικής απόκλισης

[1] 28207.58 // τυπική απόκλιση

Θα χρησιμοποιήσουμε το 5 number summary για την περιγραφή των δεδομένων καθώς όπως παρατηρούμε τα δεδομένα δεν είναι συμμετρικά κατανεμημένα(οι περισσότερες τιμές είναι συγκεντρωμένες στο διάστημα [0,5000]) και υπάρχουν ισχυρές ατυπικές τιμές.

Για την μεταβλητή %Certified:

> fivenum(X..Certified)

[1] 0.00 2.40 5.95 10.71 33.98

Επομένως το 5 number summary είναι:

Min= 0.00	Q1=2.40	Q2=5.95	Q3=10.71	Max=33.98

Εύρεση μέσου και τυπικής απόκλισης:

> mean(X..Certified) // Εντολή εύρεσης μέσου

[1] 7.782586//μέσος

> sd(X..Certified) // Εντολή τυπικής απόκλισης

[1] 6.972437 // τυπική απόκλιση

Θα χρησιμοποιήσουμε το 5 number summary για την περιγραφή των δεδομένων καθώς όπως παρατηρούμε τα δεδομένα δεν είναι συμμετρικά κατανεμημένα.Οι περισσότερες τιμές είναι παρατηρούνται στο διάστημα [0,5] και στα υπόλοιπα διαστήματα ολοένα και μειώνονται.

Για την μεταβλητή Median Age:

> fivenum(Median.Age)

[1] 22 26 29 31 53

Επομένως το 5 number summary είναι:

Min= 22	Q1=26	Q2=29	Q3=31	Max=53

Εύρεση μέσου και τυπικής απόκλισης:

> mean(Median.Age) // Εντολή εύρεσης μέσου

[1] 29.3//μέσος

> sd(Median.Age) // Εντολή τυπικής απόκλισης

[1] 4.047897// τυπική απόκλιση

Για την περιγραφή των δεδομένων λόγω ατυπικής τιμής θα χρησιμοποιήσουμε το 5 number summary.

Για την μεταβλητή % Male:

> fivenum(X..Male)

[1] 25.240 54.140 66.515 81.690 93.440

Επομένως το 5 number summary είναι:

Min= 25.240	Q1=54.140	Q2=66.515	Q3=81.690	Max=93.440

Εύρεση μέσου και τυπικής απόκλισης:

> mean(X..Male) // Εντολή εύρεσης μέσου

[1] 67.01069//μέσος

> sd(X..Male) // Εντολή τυπικής απόκλισης

[1] 15.84364 // τυπική απόκλιση

Θα χρησιμοποιήσουμε το 5 number summary για την περιγραφή των δεδομένων καθώς αν και δεν έχουμε ατυπικά σημεία η κατανομή των τιμών δεν είναι συμμετρική(περισσότερες τιμές στα υψηλότερα ποσοστημόρια).

Για την μεταβλητή % Female:

> fivenum(X..Female)

[1] 6.560 18.310 33.485 45.860 74.760

Επομένως το 5 number summary είναι:

Min= 6.560	Q1=18.310	Q2=33.485	Q3=45.860	Max=74.760

Εύρεση μέσου και τυπικής απόκλισης:

> mean(X..Female)

[1] 32.98931 //μέσος

> sd(X..Female)

[1] 15.84364//τυπική απόκλιση

Θα χρησιμοποιήσουμε το 5 number summary για την περιγραφή των δεδομένων καθώς αν και δεν έχουμε ατυπικά σημεία η κατανομή των τιμών δεν είναι συμμετρική(περισσότερες τιμές στα χαμηλότερα ποσοστημόρια).

Για την μεταβλητή % Bachelor's Degree or Higher:

> fivenum(X..Bachelor.s.Degree.or.Higher)

[1] 44.950 64.490 73.055 79.350 98.110

Επομένως το 5 number summary είναι:

Min= 44.950	Q1=64.490	Q2=73.055	Q3=79.350	Max=98.110

Εύρεση μέσου και τυπικής απόκλισης:

> mean(X..Bachelor.s.Degree.or.Higher)

[1] 72.07872

> sd(X..Bachelor.s.Degree.or.Higher)

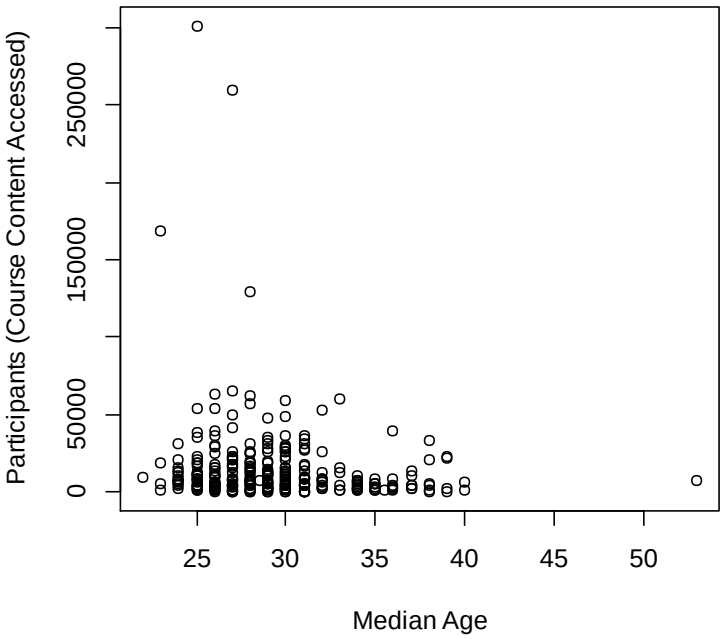
[1] 10.25643

Εδώ μπορούμε να χρησιμοποιήσουμε τον μέσο και την τυπική απόκλιση για την περιγραφή των δεδομένων καθώς δεν υπάρχουν ατυπικά σημεία και η κατανομή των δεδομένων είναι, σχετικά, συμμετρική.

e.

Θα διερευνήσουμε την σχέση μεταξύ της μέσης ηλικίας(επεξηγηματική μεταβλητή) και του αριθμού συμμετεχόντων (μεταβλητή απόκρισης).Θα δούμε δηλαδή κατά πόσο σχετίζεται ο αρθμός των συμμετεχόντων σε ένα μάθημα με την μέση ηλικία των συμμετεχόντων.

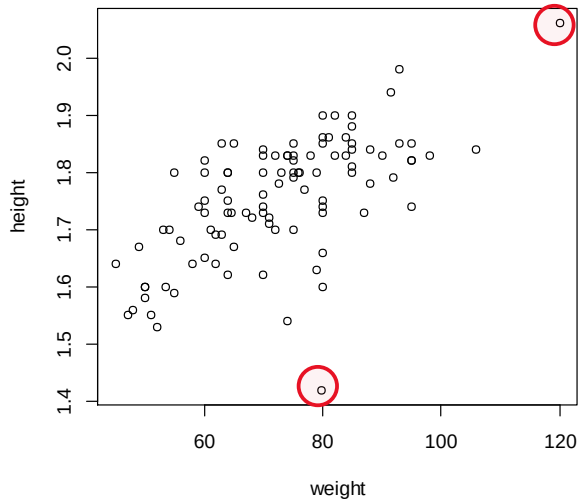
Το scatterplot είναι το εξής:



Ο συντελεστής συσχέτισης r ισούται με -0.1470493.Όπως φαίνεται και από το παραπάνω scatterplot και από το γεγονός οτι r<0 η σχέση είναι φθίνουσα.Παρατηρούμε ότι r≈0,συνεπώς η σχέση είναι ασθενής.Η σχέση ανάμεσα στις δύο μεταβλητές είναι ανύπαρκτη καθώς όπως βλέπουμε μαθήματα με μέση ηλικία συμμετέχοντα 25 έχουν την ίδια συμμετοχή με μαθήματα που ο μέσος όρος ηλικίας είναι πάνω από 35.

Άσκηση 3η

a.Θα μελετήσουμε την σχέση μεταξύ του βάρους(επεξηγηματική μεταβλητή)και του ύψους(μεταβλητή απόκρισης).Το scatterplot είναι το εξής:



Παρατηρούμε οτι η σχέση είναι αύξουσα,γραμμική και ισχυρή.Υπάρχουν δύο ατυπικά σημεία.

b. Για την εύρεση του συντελεστή συσχέτισης χρησιμοποιήθηκε η συνάρτηση cor της r.

> cor(height,weight)//εντολή εύρεσης συντελεστή συσχέτισης

[1] 0.640862 // συντελεστής συσχέτισης

Ο συντελεστής συσχέτισης είναι θετικός αριθμός επομένως η σχέση είναι αύξουσα.

Η τιμή του συντελεστή συσχέτισης δίχνει ότι η σχέση πρόκειται για μια μέτρια ισχυρή γραμμική σχέση(το 0.640862 δεν είναι τόσο κοντά στο 1 ώστε να πούμε οτι η σχέση είναι ισχυρή αλλά απέχει πολύ από το 0 για να θεωρήσουμε την σχέση ασθενή)

Για την εκτέλεση γραμμικής παλινδρόμηση ελαχίστων τετραγώνων χρησιμοποιήσαμε την συνάρτηση της r abline ως εξής (παρακάτω το διάγραμμα που προκύπτει):

> abline(lm(height~weight),col='blue')