|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete Count |
| Results of rolling a dice | Continuous |
| Weight of a person | Continuous |
| Weight of Gold | Continuous |
| Distance between two places | Continuous |
| Length of a leaf | Continuous |
| Dog's weight | Continuous |
| Blue Color | Discrete |
| Number of kids | Discrete Count |
| Number of tickets in Indian railways | Discrete Categorical |
| Number of times married | Discrete Count |
| Gender (Male or Female) | Discrete binary |

Q1) Identify the Data type for the Following:

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

|  |  |
| --- | --- |
| Data | Data Type |
| Gender | Nominal |
| High School Class Ranking | Ordinal |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ordinal |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Ratio |
| Years of Education | Ordinal |

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?

Ans. 3/8

Q4) Two Dice are rolled, find the probability that sum is

1. Equal to 1
2. Less than or equal to 4
3. Sum is divisible by 2 and 3

Ans. (a) the probability of equal to 1=0;

(b) Less than or equal to 4 = (1,1),(1,2),(1, 3),(2,1),(2,2)(3,1) =6/36=1/6

(C) Sum is divisible by 2 and 3 = (24/36)= (2/3)

Q5) A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?

Ans. Total number of balls = 7 Balls

Number of ways drawing two balls= 7C2= 21

If we draw out 2 blue balls= 7-2 =5 balls

Number of ways drawing two balls after drawing out blue balls= 5C2=10

Total Probability = 10/21

Q6) Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)

|  |  |  |
| --- | --- | --- |
| CHILD | Candies count | Probability |
| A | 1 | 0.015 |
| B | 4 | 0.20 |
| C | 3 | 0.65 |
| D | 5 | 0.005 |
| E | 6 | 0.01 |
| F | 2 | 0.120 |

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

Ans. 1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.120=3.09

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points,Score,Weigh>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

> mean(Points) Weight variable has highest mean value a compare to other variable

[1] 3.596563

> mean(Score)

[1] 3.21725

> mean(Weigh)

[1] 17.84875

> mode(Points)

[1] "numeric"

> mode(Score)

[1] "numeric"

> mode(Weigh)

[1] "numeric"

> var(Points)

[1] 0.2858814

> var(Score)

[1] 0.957379

> var(Weigh)

[1] 3.193166

> sd(Points)

[1] 0.5346787

> sd(Score)

[1] 0.9784574

> sd(Weigh)

[1] 1.786943

> range(Points)

[1] 2.76 4.93

> range(Score)

[1] 1.513 5.424

> range(Weigh)

[1] 14.5 22.9

**1. Minimum value for points variable is 2.76 and maximum value is 4.93 similar for others variables**

Q8) Calculate Expected Value for the problem below

1. The weights (X) of patients at a clinic (in pounds), are

108, 110, 123, 134, 135, 145, 167, 187, 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

Ans. The Expected value of random = 145.34

**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

**Use Q9\_a.csv**

> skewness(speed)

[1] -0.1139548

Explanation - speed skewness is –ve that means that long tails lies towards left side and speed of the cars is more than the average speed.

> skewness(dist)

[1] 0.7824835

Explanation - speed skewness is +ve that means that long tails lies towards right side and distance travelled by the cars is less than the average distance.

> kurtosis(speed)

[1] 2.422853

Explanation - speed kurtosis is +ve that means that there is high peakness in the data,

which indicates that there are lots outliers present in the data.

> kurtosis(dist)

[1] 3.248019

Explanation - distance kurtosis is +ve that means that there is high peakness in the data,

Which indicates that there are lots outliers present in the data.

**SP and Weight(WT)**

**Use Q9\_b.csv**

> skewness(SP)

[1] 1.581454

Postive skewness means long tail lie towards right side and SP value highest value is less than average value of data set

> skewness(WT)

[1] -0.6033099

Negative skewness means long tail lie towards left side and SP value highest value is more than average value of data set

> kurtosis(SP)

[1] 5.723521

Positive means that there will high peakedness in the data and data is not spreaded that much. So there will lots outlier present in the data set

> kurtosis(WT)

[1] 3.819466

Positive means that there will high peakedness in the data and data is not spreaded that much. So there will lots outlier present in the data set

**Q10) Draw inferences about the following boxplot & histogram**



Explanation – According to histogram graph, it is showing a positive skewness. This means that most the weight data is less than the average weight of chcikweight. Moreover, it might have positive kurtosis because data is not spreaded that much.



Explanation – According to histogram graph, it is showing a positive skewness. This means that most the weight data is less than the average weight of chcikweight. Moreover, it might have positive kurtosis because data is not spreaded that much. Additionally, data is left skewed, which mean that most the data lies towards left side but average data of weight more than the chickweight data.

**Q11)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

Ans. Xbar=200, n= 2000, sigma=30 …….Z value at 94%, left at 0.97=1.88

CI= xbar+-Z(1-alpha)\*sigma/squrt(n)

xbar<-200

> n<-2000

> sigma<-30

> error1<-qnorm(0.97)\*sigma/sqrt(n)

> left<-xbar-error1

> right<-xbar+error1

> left

[1] 198.7383

> right

[1] 201.2617

**Q12)** Below are the scores obtained by a student in tests

**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

1. Find mean, median, variance, standard deviation.
2. > X2<-c(34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56)
3. > mean(X2)
4. [1] 41
5. > median(X2)
6. [1] 40.5
7. > mode(X2)
8. [1] "numeric"
9. > var(X2)
10. [1] 25.52941
11. > sd(X2)
12. [1] 5.052664
13. What can we say about the student marks?

Explanations—The average marks of student is 41

Q13) What is the nature of skewness when mean, median of data are equal?

Explanation—When mean= median= mode then it is consider as normal distribution and it form a bell curve shaped graph.

Q14) What is the nature of skewness when mean > median ?

It is positively skewed, because most of the data will lie right side.

Q15) What is the nature of skewness when median > mean?

It is negatively skewed, because most of the data will lie left side.

Q16) What does positive kurtosis value indicates for a data ?

Positive value indicates that the data is not widely spreaded and there is a high number of outlier present in data.

Q17) What does negative kurtosis value indicates for a data?

Negative value indicates that the data is widely spreaded and there is a high probability of error between the data. It means that there is chance of multicollinearity in data

Q18) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

It is a negatively skewed data, most of the lies towards left side

What is nature of skewness of the data?

It is negatively skewed

What will be the IQR of the data (approximately)?

Q3-Q1, 18-10 = 8

Q19) Comment on the below Boxplot visualizations?



Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

Explanation—Boxplot1 indicates that has positive kurtosis, because the distribution of data is not widely spreaded as compare to Boxplot2, which is widely spreaded and consider as negative kurtosis

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

* 1. P(MPG>38)

> d<-mean(MPG)

> q<-sd(MPG)

> s<-38

> Cardatadist<-pnorm(s,d,q)

> Cardatadist

[1] 0.6524061

> Probgreaterthan38<-1-convertzscore

> Probgreaterthan38

[1] 0.169538

[1] 0.2570696, so greater 38 value have 16.9% of probability

lying in that region

* 1. P(MPG<40)

> d<-mean(MPG)

> q<-sd(MPG)

> s<-40

> Cardatadist<-pnorm(s,d,q)

> Cardatadist

[1] 0.7293499

> Problessthan40<-Cardatadist

> Problessthan40

[1] 0.7293499

So probability of values which fall in the range of less than 40 range is 72.9%

c. P (20<MPG<50)

X> d<-mean(MPG)

> q<-sd(MPG)

> s<-20

> Cardatadist<-pnorm(s,d,q)

> Cardatadist

X[1] 0.05712378, so 5.7% data lye in the less than 20 range

> d<-mean(MPG)

> q<-sd(MPG)

> s<-50

> Cardatadist<-pnorm(s,d,q)

> Cardatadist

X[2] 0.9559927

So for the range 20<MPG<50, we will minus X2 from X1

> Probgreaterthan20<-Cardatadist

> X1<-Probgreaterthan20

> Probbetween20and50<-(X2-X1)\*100

> Probbetween20and50

[1] 89.88689

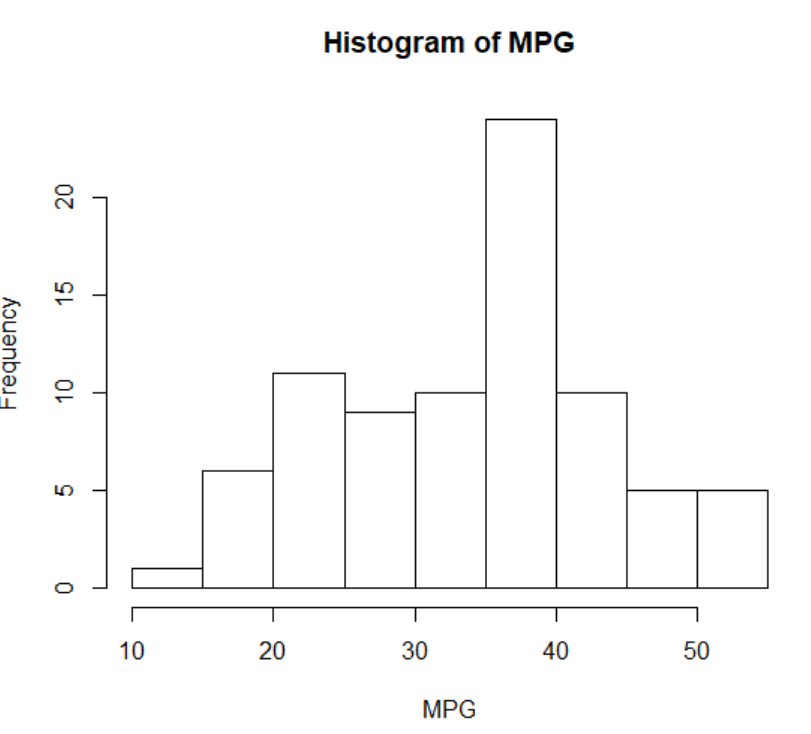
So 89.9% data will lye in the range of 20 to 50

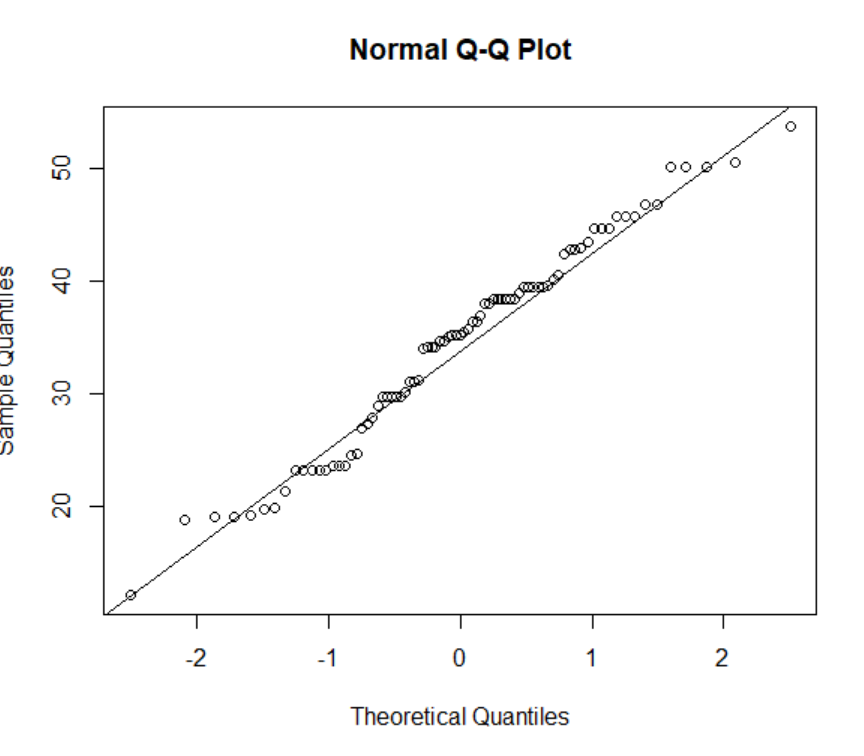
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

No it doesn’t follow the normal distribution curve, data is left skewed and have a positive kurtosis.

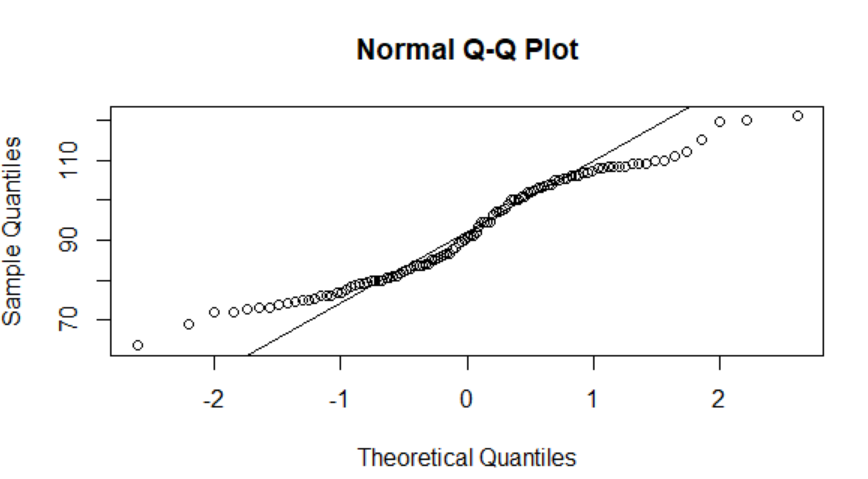




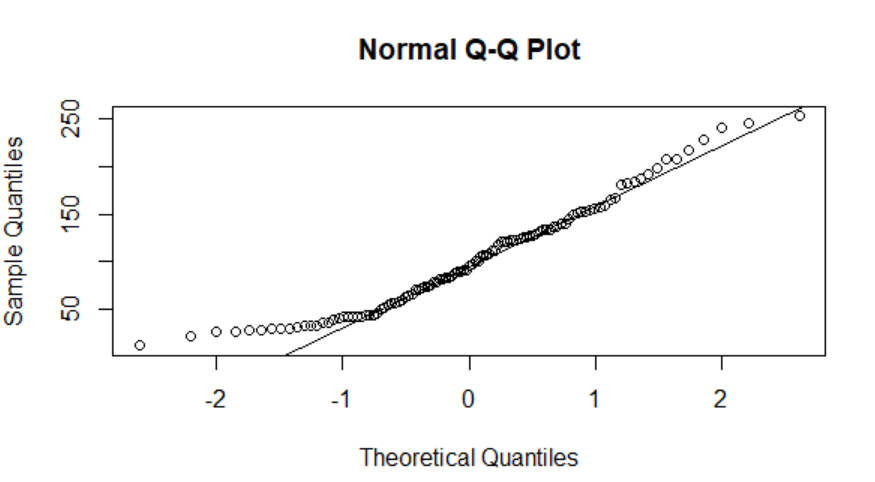
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

Wc\_at$Waist Graph



For Wc\_at$PT graph



Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

For 90%

> d<-mean(MPG)

> q<-sd(MPG)

> n1<-81

> error2<-qnorm(0.95)\*q/sqrt(n1)

> left22<-d-error2

> left22

[1] 32.7532

> right22<-d+error2

> right22

[1] 36.09095

For 94%

> d<-mean(MPG)

> q<-sd(MPG)

> n1<-81

> error2<-qnorm(0.97)\*q/sqrt(n1)

> left22<-d-error2

> left22

[1] 32.51381

> right22<-d+error2

> right22

[1] 36.33034

For 60%

> d<-mean(MPG)

> q<-sd(MPG)

> n1<-81

> error2<-qnorm(0.75)\*q/sqrt(n1)

> left22<-d-error2

> left22

[1] 33.73774

> right22<-d+error2

> right22

[1] 35.10642

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

For 95%

> d<-mean(MPG) ### T distribution

> q<-sd(MPG)

> n2<-25

> error2<-qt(0.975,df=n2-1)\*q/sqrt(n2)

> left22<-d-error2

> left22

[1] 30.6528

> right22<-d+error2

> right22

[1] 38.19135

For 96%

> d<-mean(MPG) ### T distribution

> q<-sd(MPG)

> n2<-25

> error2<-qt(0.98,df=n2-1)\*q/sqrt(n2)

> left22<-d-error2

> left22

[1] 30.45621

> right22<-d+error2

> right22

[1] 38.38794

For 99%

> d<-mean(MPG) ### T distribution

> q<-sd(MPG)

> n2<-25

> error2<-qt(0.995,df=n2-1)\*q/sqrt(n2)

> left22<-d-error2

> left22

[1] 29.31406

> right22<-d+error2

> right22

[1] 39.5301

Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

> muu=270

> xbarr=260

> n2=17 ## sample number

> s=90

> terror<-s/sqrt(n2)

> tscore<-(xbarr-muu)/terror

> tscore

[1] -0.4581228

> pt(tscore,df=17)

[1] 0.326333

So, the average life of bulb not more than 260 days will have 32.6% probability