|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Discrete |
| 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 | Qualitative |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Qualitative |

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 | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Interval |
| Blood Group | Nominal |
| Time Of Day | Ratio |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ordinal |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Interval |

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

Ans : (HHH, HHT, HTT, TTT, THH, TTH, THT, HTH)

N(s) = 8

P(x) for getting two head and one tail = 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 : The data set for the two dice

(1,1) (1,2) (1,3) (1,4) (1,5) (1,6)

(2,1) (2,2) (2,3) (2,4) (2,5) (2,6)

(3,1) (3,2) (3,3) (3,4) (3,5) (3,6)

(4,1) (4,2) (4,3) (4,4) (4,5) (4,6)

(5,1) (5,2) (5,3) (5,4) (5,5) (5,6)

(6,1) (6,2) (6,3) (6,4) (6,5) (6,6)

N(s) = 36

1. The sum is equal to 1 = p(E1) = p(E1)/ N(s) = 0
2. Sum less than or equal to 4 = p(E2) = 6/36= 1/6
3. Sum is divisible by 2 and 3 = p(E3) = 25/36

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?

Total number of balls = 7

n(S)= Number of ways of drawing 2 balls out of 7 = 7C2 =21

Let E = Event of drawing 2 balls, none of which is blue.

n(E)= Number of ways of drawing 2 balls out of (2 + 3) balls = 5C2=10

P(E)= N(E)/N(S)

=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

Expected number of candies for a randomly selected child

= 1 \*0.015 + 4\*0.20 + 3 \*0.65 + 5\*0.005 + 6 \*0.01 + 2 \* 0.12

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 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**

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 : Expected Value = ∑ ( probability \* Value )

= ∑ P(x).E(x)

P(x) = 1/9

E(x) = 108, 110, 123, 134, 135, 145, 167, 187, 199

Expected Value = 1/9(108+110+ 123 +134+ 135+ 145+ 167+ 187+ 199)

= 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

Ans: > #Cars speed and distance

>

> df = read.csv("C:\\Users\\ganes\\Desktop\\DS\\Q9\_a.csv")

> names(df)

[1] "Index" "speed" "dist"

> print(df)

> ##\*\*\*\*\*\*\*\*\*\*\*\*\*\*For Speed\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> SP=df['speed']

> library(moments)

> skewness(speed)

SP

1.581454

> kurtosis(speed)

SP

5.723521

> ##\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*For Distance\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> distance=df['dist']

> skewness(distance)

dist

0.7824835

> kurtosis(distance)

dist

3.248019

> ##Draw point plot

> plot(df$speed, df$dist, xlab = 'Speed',

+ ylab = 'Distance', main = 'Inference between car speed and distance',

+ col = 'blue')

> ##Histogram

> hist(df$speed)

> hist(df$dist)

For speed

> skewness(speed)

SP

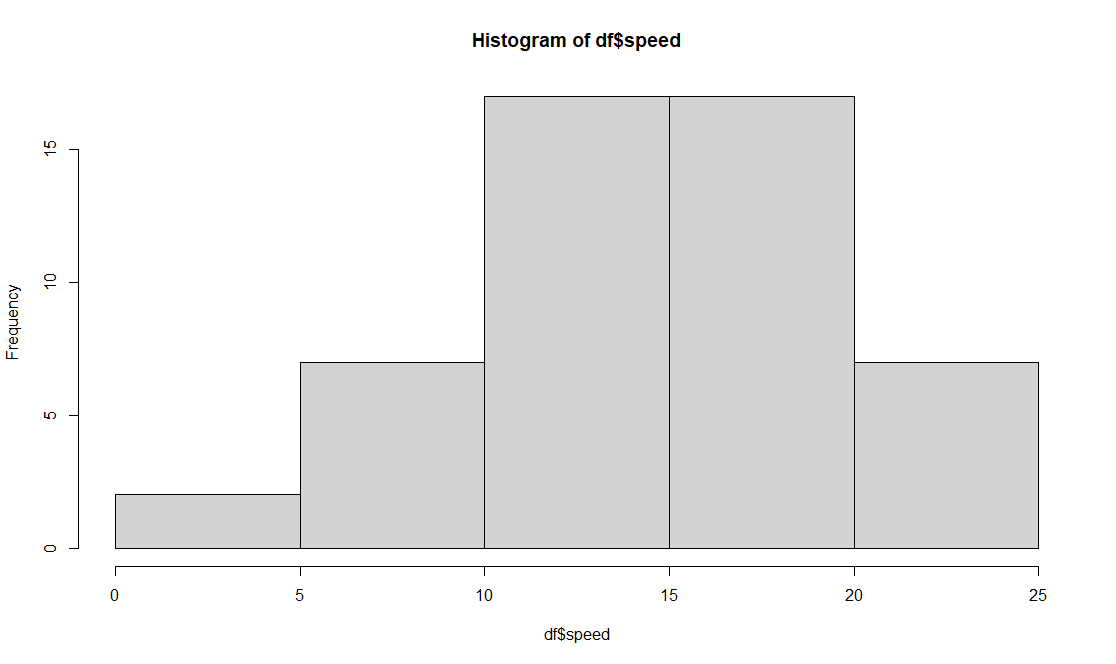
-0.9931548

> kurtosis(speed)

SP

2.4228

Negative Skewness as Distribution is skewed towards left. The kurtosis value is less than 3 the distribution having broader peak and thin tail as marked from the histogram the Kurtosis is negative. Mean of the distribution is less than the median.



For distance

> distance=df['dist']

> skewness(distance)

dist

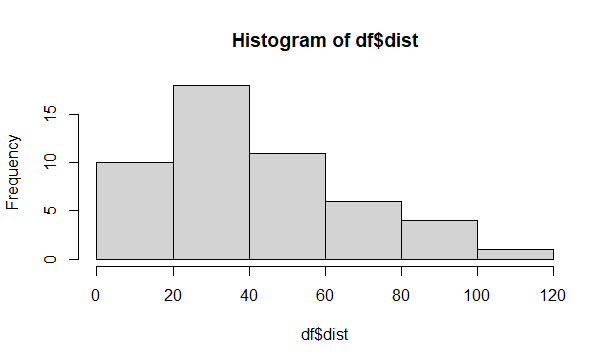
0.7824835

> kurtosis(distance)

dist

3.248019

Skewness is positive as the distribution is skewed towards right also the kurtosis is also positive as the kurtosis value is greater than 3 the distribution having sharp peak and wide tails. Mean of the distribution is greater then the median.



**SP and Weight(WT)**

**Use Q9\_b.csv**

Ans.

> df = read.csv("C:\\Users\\ganes\\Desktop\\DS\\Q9\_b.csv")

> names(df)

[1] "X" "SP" "WT"

> print(df)

> ##\*\*\*\*\*\*\*\*\*\*\*\*\*\*For Speed\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> Speed=df['SP']

> library(moments)

> skewness(Speed)

SP

1.581454

> kurtosis(Speed)

SP

5.723521

> ##\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*For Weight\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> Weight=df['WT']

> skewness(Weight)

WT

-0.6033099

> kurtosis(Weight)

WT

3.819466

> ##Draw point plot

> plot(df$SP, df$WT, xlab = 'Speed',

+ ylab = 'Weight', main = 'Inference between car speed and Weight',

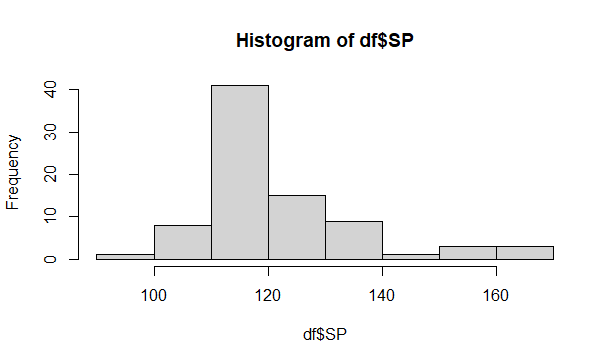
+ col = 'blue')

> ##histogram

> hist(df$SP)

> hist(df$WT)

Skewness is positive distribution is skewed towards the right also the kurtosis value is greater than 3 so the distribution having sharp peak and wider tail evident from the histogram also the kurtosis is positive. Mean of the distribution is greater than the median.



**For** Weight

> ##\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*For Weight\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> Weight=df['WT']

> skewness(Weight)

WT

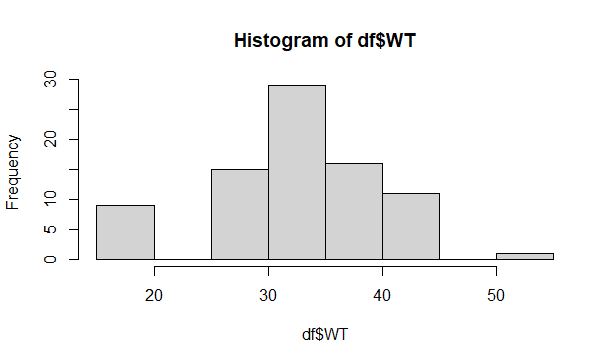
-0.6033099

> kurtosis(Weight)

WT

3.819466

The skewness negative, it is shows the distribution is left skewed also the kurtosis value is greater than 3 so the distribution is having sharp peak and wider tail as evident from histogram. Positive kurtosis. Mean of the distribution is less than the median.



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



Ans:

The distribution is right skewed. The mean of the distribution is greater than the median. The distribution having sharp peak and wider tail.



**Ans** :

Above box plot shows that its having multiple outliers 7 number of outlier .

**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.

Given data

For 94% interval

N=2000

X-bar= 200

Sd= 30

Confidence interval= X-bar+-z\*sd/sqrt(n)

=200-1.88\*30/sqrt(2000)

=198.74-201.26

Now for 98%

Confidence interval= 200-2.33\*30/sqrt(2000)

=198.43- 201.56

Now for 96%

Confidence interval=200-2.05\*30/sqrt(2000)

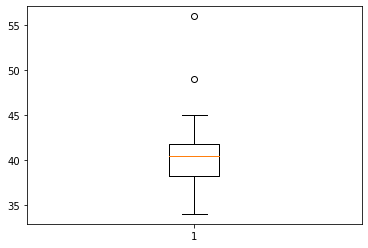
=198.62-201.37

**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. What can we say about the student marks?

Ans: 1)**Mean=41 Median=40.5 Variance=25.52 Standard deviation=5.05**

****

**Students marks** having two outliers 49,56

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

**Ans: It is called zero skewed Skewness=0,Symmetric**

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

**Ans: Right skewed distribution**

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

**Ans: Left skewed distribution**

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

Ans : It indicates that the distribution having sharp peak and having thick tail than the normal distribution.

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

Ans: Negative kurtosis shows that the distribution having broader peak and thin tail than normal distribution.

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



What can we say about the distribution of the data?

**Ans: asymmetrical distribution ,Not normal distribution**

What is nature of skewness of the data?

**Ans: Left skewed**

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

Ans= Inter Quartile Range= 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.

Ans. Both are normally distributed.

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)
  2. P(MPG<40)
  3. P (20<MPG<50)

Ans : a) # p(MPG>38)

stats.norm.cdf(0.38,df.MPG.mean(),df.MPG.std())

9.650060262669067e-05

b) # p(MPG<40)

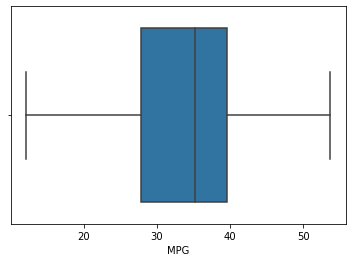
stats.norm.cdf(40,df.MPG.mean(),df.MPG.std())

0.7293498762151616

c) # p(20<MPG>50)

stats.norm.cdf(50,df.MPG.mean(),df.MPG.std())-stats.norm.cdf(20,df.MPG.mean(),df.MPG.std())

0.8988689169682046



Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans : df= read.csv("C:\\Users\\ganes\\Desktop\\DS\\Cars.csv")

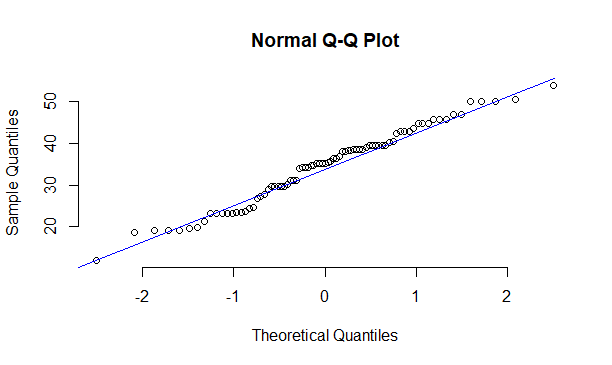
df

names(df)

plot.new()

qqnorm(df$MPG, pch = 1, frame = FALSE)

qqline(df$MPG, col = "Blue")



*As all point falls approximately along the reference line which is 45% incline the distribution follows normality.*

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

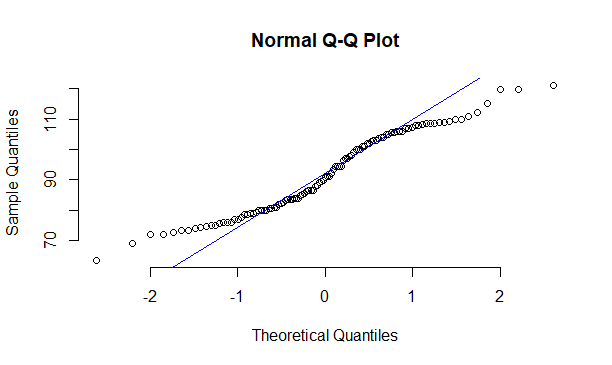
Dataset: wc-at.csv

Ans: df= read.csv("C:\\Users\\ganes\\Desktop\\DS\\WC\_AT.csv")

df

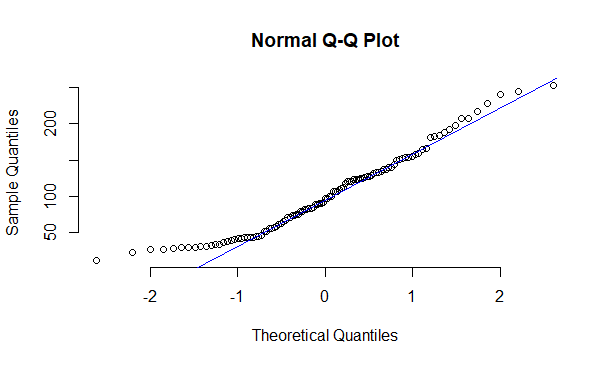
qqnorm(df$Waist, pch =1, frame = FALSE)

qqline(df$Waist, col= 'blue')

**

*For Waist size it follows the normal distribution as all the points falls approximately along the reference line.*

*For AT all the points falls approximately along the reference line so it follows the normal distribution.*



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

Ans : from scipy import stats

from scipy.stats import norm

# Z -score of the 90% confidence interval

stats.norm.ppf(0.90)

1.2815515655446004

# Z- score of the 94% confidence interval

stats.norm.ppf(0.94)

1.5547735945968535

# Z score of the 60% confidence interval

stats.norm.ppf(0.60)

0.2533471031357997

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

Ans : from scipy import stats

from scipy.stats import norm

# t score value for the 95 % confidence interval

# df = n-1 = 25-1

stats.t.ppf(0.95, 24)

1.7108820799094275

# t score value for the 96% confidence interval

stats.t.ppf(0.96, 24)

1.8280511719596342

# t score value for the 99% confidence interval

stats.t.ppf(0.99, 24)

2.4921594731575762

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

Àns: from scipy import stats

from scipy.stats import norm

# Assume Null Hypothesis is: Ho = Avg life of the Bulb >=260 days

# Alternate Hypothesis is :Ha = Avg life of the Bulb <260 days

# to find t-score given data is

# x-bar= 260; n=18; SD= 90 days; pop mean= 270; df= n-1=18-1=17

# t=(s\_mean-P\_mean/s\_SD/sqrt(n))

t=(260-270)/(90/18\*\*0.5)

t

-0.4714045207910317

# find the P(x>=260) for null hypothesis

# p\_value=1-stats.t.cdf(t\_score,df)...using cdf function

p\_value=1-stats.t.cdf(abs(-0.4714), 17)

p\_value =0.32167411684460556

## The probability of that 18 selected bulb of having avergae life not more than 260 days is 0.3216 which is

>than alpha value 0.05 so if p>0.05(standard value) we accept the Ho and reject the null Ha and (vice versa)

so here the p>0.05 so we accept the Ho also the CEO claim is false and the average life of the bulb is> 260 days.