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

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 | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

Ans :HHT,HTH,THH

Total possibility = 2\*2\*2 = 8

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 2and 3
4. Zero
5. Total possibilities = 6\*6 = 36

Possibilities for less than or equal to 4 = 6

Probability = 1/6

1. Total possibilities = 6\*6 = 36

Possibilities for sum is divisible by 2 and 3 = 6

Probability = 1/6

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 possibility = 7c2

Possibility for none of the balls drawn is blue = 5c2

Probability = 5c2/7c2 = 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: 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\*0.120

= 3.09

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

* For Points,Score,Weight

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

**Use Q7.csv file**

**Ans:**

**Using R programming**

**Points**

Mean = 3.59

Median =3.69

Mode = Cannot find, because values are numeric type

Varience = 0.2858

Standard Deviation = 0.5346

Range = 2.17

**Score**

Mean = 3.21

Median = 3.325

Mode = Cannot find, because values are numeric type

Varience = 0.95

Standard Deviation = 0.97

Range = 3.91

**Weight**

Mean = 17.84

Median = 17.71

Mode = Cannot find, because values are numeric type

Varience = 3.19

Standard Deviation = 1.78

Range = 8.39

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: (108+110+123+134+135+145+167+187+199)/9 = 145.33

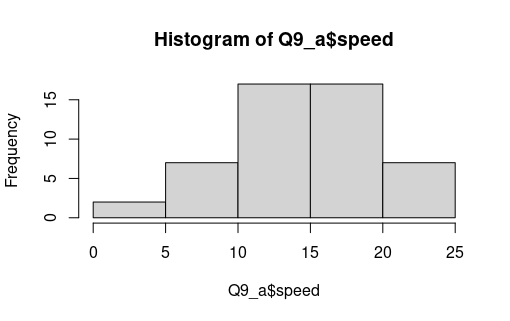
**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

**Use Q9\_a.csv**

**Car Speed**

Ans: Using moments package in Rstudio



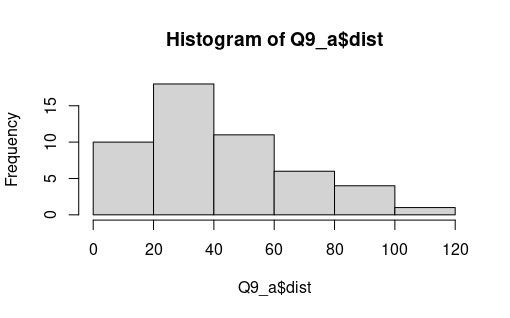
skewness(Q9\_a$speed) = -0.1139548

* Negative skeweness or left skeweness that is outliers are present in left side

kurtosis(Q9\_a$speed) = 2.422853

* Data distribution is wide not peak
* Kurtosis value is less than 3 that is platykartic curve

**Car distance**

****

skewness(Q9\_a$dist) = 0.7824835

* Positive skeweness or right skeweness that is outliers are present in right side

kurtosis(Q9\_a$dist) = 3.248019

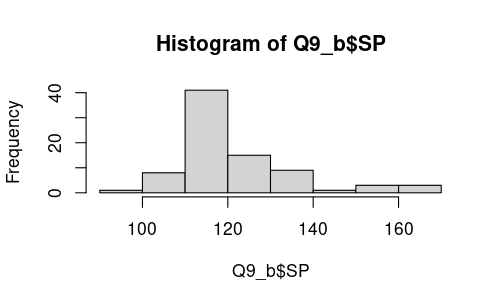
* Data distribution is somewhat peaked (More than normal distribution)
* Kurtosis value is approximately 3 that is mesokartic curve

**SP and Weight(WT)**

**Use Q9\_b.csv**

Ans:

**SP**

****

skewness(Q9\_b$SP) = 1.581454

* Positive skeweness or right skeweness that is outliers are present in right side

kurtosis(Q9\_b$SP) = 5.723521

* Data distribution is peak
* Kurtosis value is greater than 3 that is leptokartic curve

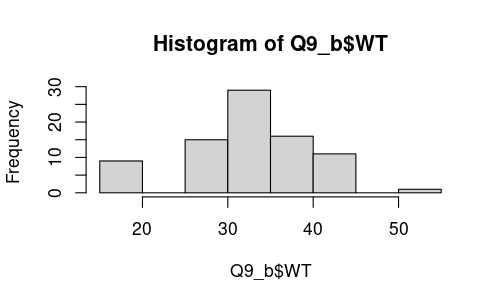
**Weight(WT)**

skewness(Q9\_b$SP) = -0.6033099

* Negative skeweness or left skeweness that is outliers are present in left side

kurtosis(Q9\_b$SP) = 3.819466

* Data distribution is peak
* Kurtosis value is greater than 3 that is leptokartic curve



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



Ans:

* 50-100 weight having high frequency that is more data points are concentrated in this area
* 300-400 weight having low frequency
* Outliers are present in right side
* Right skewed or positive skewed
* Data is not normally distributed



* Median is less than the mean that is right skewed
* 7 Outliers are present in upper side of box plot
* Less data points between Q1 and Q3

**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:

Here population standard deviation is not given so using tscore values for finding confidence interval

Sample size = 2000

Degree of freedom = 1999(2000-1)

Sample mean = 200

Sample standard deviation = 30

Using python coding

**94% Confidence level**

stats.t.interval(0.94,1999,200,30)

Confidence interval = (143.54415570565965, 256.45584429434035)

**98% Confidence level**

stats.t.interval(0.98,1999,200,30)

Confidence interval = (130.15355671679083, 269.84644328320917)

**96% Confidence level**

stats.t.interval(0.96,1999,200,30)

Confidence interval = (138.34730111522666, 261.6526988847733)

**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.Using python coding Make the values as data frame

df = pd.DataFrame({"salary" :[34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]})

Mean = 41 (df['salary'].mean())

Median =40.5(df['salary'].median())

Varience = 25.529(df['salary'].var())

Standard deviation = 5.052(df['salary'].std())

2.

* Score range from 34-56
* Most occurring value is 41(Mode)
* Most student score between 36-42
* Median<Mean, so right skewed

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

Ans:

When the values of mean and median are equal, there is no skewness also can say that the data is in normal distribution

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

Ans:

If the mean is greater than the median, the distribution is positively skewed or right skewed and also outliers are present.

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

Ans:

If the median is greater than the mean, the distribution is negatively skewed or left skewed and also outliers are present

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

Ans:

A distribution with positive kurtosis value that the distribution has heavier tails than the normal distribution

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

Ans:

A distribution with negative kurtosis value that the distribution has lighter tails than the normal distribution

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



What can we say about the distribution of the data?

Ans:

* No outliers
* Median between 15 and 16
* Most data present in the 10-18
* 25 percentage of data lies in below 10 (Q1)
* 25 percentage of data lies in between 10-18 (Q2)
* 25 percentage of data lies in above 18 (Q3)
* Not following normal distribution
* Left skeweness of data

What is nature of skewness of the data?

Ans: Left skeweness/ negatively skewed

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

Ans: IQR = 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 plot infer that their data is normally distributed
* As comparing boxplot 1 and boxplot 2, we can say that boxplot 1 for sample distribution and boxplot 2 is for population or a sample with large size.
* No outliers
* Q1 is 25% ,Q3 is75% and Q2 is 50% for both the box plots. So we can say that both the distribution follow normal distribution

ie, mean = median = mode

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)

c. P (20<MPG<50)

Ans : Using python coding First find out the values of mean and standard deviation

mean = df1["MPG"].mean()

std = df1["MPG"].std()

1. P(MPG>38)

1-stats.norm.cdf(38,mean,std) = 0.34759392515827137

P(MPG>38) = 34.75%

1. P(MPG<40)

stats.norm.cdf(40,mean,std) = 0.7293498762151609

P(MPG<40) = 72.93%

1. P(20<MPG<50)

stats.norm.cdf(50,mean,std) - stats.norm.cdf(20,mean,std) = 0.8988689169682047

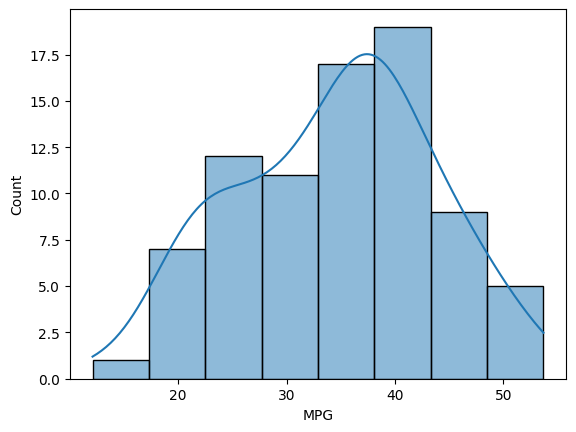
P(20<MPG<50) = 89.88%

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans:



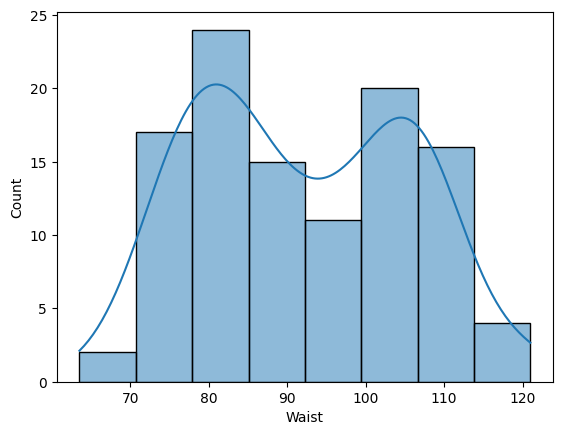
Its skeweness = -0.177, that means it’s slightly negatively skewed and it close to zero. So MPG is not normally distributed

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

Dataset: wc-at.csv

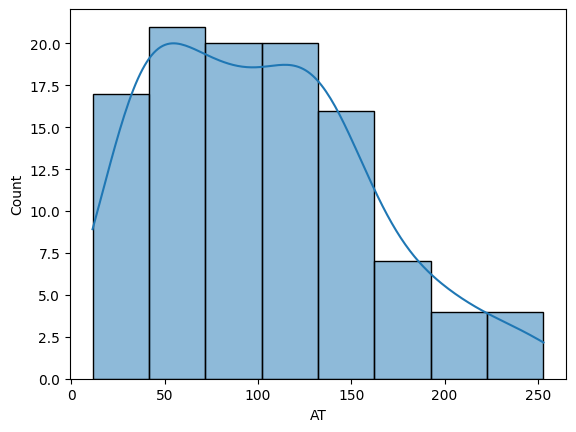
Ans:

Waist Circumference(Waist)



It is normally distributed

Adipose Tissue (AT)



Adipose Tissue dataset is not normally distributed

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

Ans:

Z scoresof 90%

stats.norm.ppf(0.95)(0.90+.05)

1.6448536269514722

Z scoresof 94%

stats.norm.ppf(0.97)(0.94+.03)

1.8807936081512509

Z scoresof 60%

stats.norm.ppf(0.80)(0.60+.20)

0.8416212335729143

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

Degree of freedom = 25-1 = 24

t scores of 95%

stats.t.ppf(0.975,24)

2.0638985616280205

t scores of 96%

stats.t.ppf(0.98,24)

2.1715446760080677

t scores of 99%

stats.t.ppf(0.995,24)

2.796939504772804

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

Ans:

Sample size = 18 = n

Sample mean = 260 days = x

Population mean = 270

Sample standard deviation = s = 90days

Tscore = 260-270/90/SQRT(18)

from numpy.ma.core import sqrt

tscore = (260-270)/(90/sqrt(18))

tscore

-0.4714

1-stats.t.cdf(abs(tscore),df=17)

0.32167253567098353

Probability average life of no more than 260 days = 32.16%