|  |  |
| --- | --- |
| 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 | Discrete |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Discrete |

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 | Interval |
| 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 | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Interval |

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:

* 1. Ans: Equal to 1 - **0 Probability**
  2. Less than or equal to 4 - **1/6 (0.167)**
  3. Sum is divisible by 2 and 3 - **1/6 (0.167)**

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: 0.47**

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 = 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.

**Ans:**

**For Points:  
 Mean = 3.596  
 Median = 3.695  
 Mode = 3.92 & 3.07  
 Variance = 0.286  
 Std Deviation = 0.535  
 Range = 2.76 to 4.93 (2.17)  
Comments: Multi Modes. Variance is veery small and the range is also small. Median & Mode is more than mean, hence the distribution is Left Skewed.**

**For Score:  
 Mean = 3.217  
 Median = 3.325  
 Mode = 3.44  
 Variance = 0.957  
 Std Deviation = 0.978  
 Range = 1.153 to 5.424 (3.911)  
Comments: Median & Mode is higher than Mean, hence the distribution is Left Skewed.**

**For Height:  
 Mean = 17.849  
 Median = 17.71  
 Mode = 17.02 & 18.9  
 Variance = 3.193  
 Std Deviation = 1.787  
 Range = 14.5 to 22.9 (8.4)  
Comments: Multi Modes. Mean is higher than Median & Mode, hence the distribution is Right Skewed.**

Inference Drawn:

* The mean is useful for spotting trends in the data because we can compare means over a time period to spot trends. The mean is the most common measure of central tendency.
* The **median** divides a sample of data in half; it is the middle score. The median is a useful statistic if we think our data have some extreme cases. The median is not impacted by extreme cases, but the mean is.

**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 = =145.33**

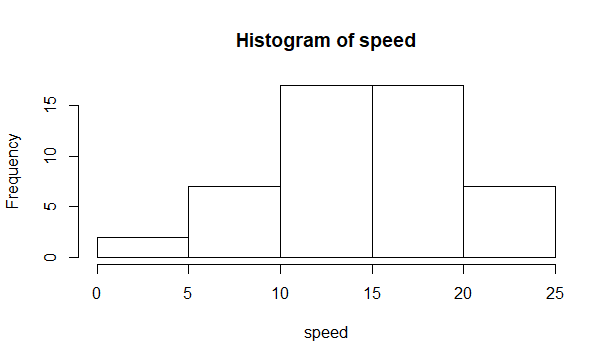
Q9) Calculate Skewness, Kurtosis & draw inferences on the following data

Cars speed and distance

Use Q9\_a.csv

**Ans: For Speed:** For Speed Column

|  |
| --- |
| > attach(a)  > skewness(speed)  [1] -0.1139548 ## left skewness  > kurtosis(speed)  [1] 2.422853 ## negative Kurtosis |
|  |
| |  | | --- | |  | |



Skewness is negative, that tells us that the distribution is skewed towards left. Mean of distribution is less than the Median. Kurtosis Value is less than 3, that tells us that the distribution has broad peak and thin tails as evident from the histogram.

For dist column

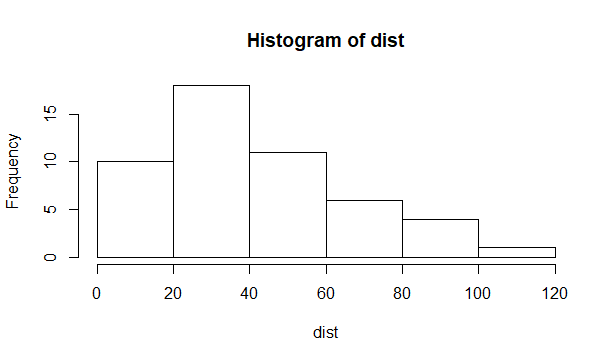
> attach(a)

> skewness(dist)

[1] 0.7824835 ## positive/Right skewness

> kurtosis(dist)

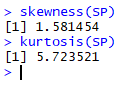
[1] 3.248019 ##positive Kurtosis

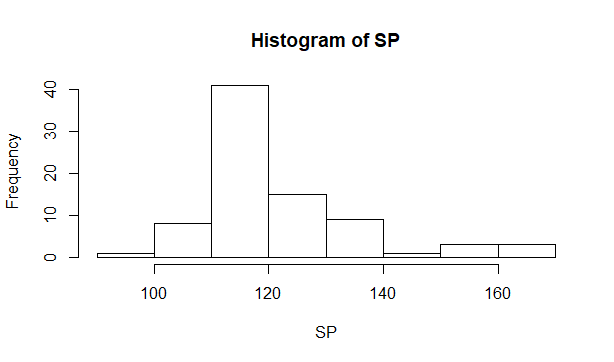


Skewness is positive, that tells us that the distribution is skewed towards right. Mean of distribution is more than the Median. Kurtosis Value is more than 3, that tells us that the distribution has sharp peak and wide tails as evident from the histogram.

Use Q9\_b.csv

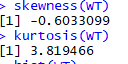
**Ans:** For SP:

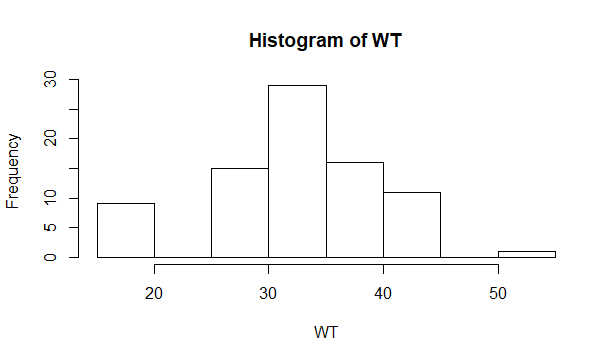
****



Skewness is positive, that tells us that the distribution is skewed towards right. Mean of distribution is more than the Median. Kurtosis Value is more than 3, that tells us that the distribution has sharp peak and wide tails as evident from the histogram.

**For WT:**





Skewness is negative, that tells us that the distribution is skewed towards left. Mean of distribution is less than the Median.Kurtosis Value is more than 3, that tells us that the distribution has sharp peak and wide tails as evident from the histogram.

Q10) Draw inferences about the following boxplot & histogram



**Ans: The distribution is right skewed. Most of the data is concentrated between 50 to 100 with a frequency of 200. The sharp peak indicates that the Kurtosis is positive.**



**Ans: There are outliers towards upper extreme. In box plot the box contains the middle 50% of the data points and each of the two whisker contain upper and lower 25th percentile of data of the data. Data is positively skewed.**

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: Population (N) = 3,000,000  
 Sample Size (n) = 2000  
 Sample Mean (x̄) = 200  
 Sample Std Dev (s) = 30  
 Standard Error = 0.671  
   
For 94% C.I.: α =1-0.94=0.06  
+ t 0.94,1999\*s/=200+1.881861\*0.671=200+1.26272873=201.262729**

**- t 0.94,1999\*s/=200- 1.881861\*0.671=200-1.26272873=198.737271**

**[198.737271, 201.104466]**

**For 98% C.I.: α =1-0.98=0.02  
+ t 0.98,1999\*s/=200+2.328215\*0.671=200+1.56223227=201.562232**

**- t 0.98,1999\*s/=200-2.328215\*0.671=200-1.56223227=198.437768**

**[198.437768, 201.562232]**

**For 96% C.I.: α =1-0.96=0.04  
+ t 0.96,1999\*s/=200+2.05509\*0.671=200+1.37896539=201.378965**

**- t 0.96,1999\*s/=200-2.05509\*0.671=200-1.37896539=198.621035**

**[198.621035, 201.378965]**

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:**

**Mean = 41  
 Median = 40.5  
 Variance = 25.53  
 Std. Dev = 5.053  
The average score by students is 41, while the scores ranges between 34 to 56. The score distribution is left skewed.**

Q13) What is the nature of skewness when mean, median of data are equal?  
**Ans: No Skewness**

Q14) What is the nature of skewness when mean > median ?  
**Ans: Right / Positive Skewness**

Q15) What is the nature of skewness when median > mean?  
**Ans: Left / Negative Skewness**

Q16) What does positive kurtosis value indicates for a data ?  
**Ans: Positive kurtosis value indicates that the distribution has heavier tails and a sharper peak than the normal distribution.**

Q17) What does negative kurtosis value indicates for a data?  
**Ans: Negative kurtosis value indicates that the distribution has lighter tails and a flatter peak 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: 25th percentile is at 10, 50th Percentile value is 15, 75th Percentile value is 18. It is not a normal distribution.**

What is nature of skewness of the data?  
**Ans: The distribution is Left / Negative skewed. Median is greater than Mean.**

What will be the IQR of the data (approximately)?   
**Ans: IQR = 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: Boxplot 1 & Boxplot 2 have the same Median value.   
Boxplot 1 has Positive Kurtosis value, Boxplot 2 has Negative Kurtosis value.  
There are no outliers in both plots.  
Both plots have no skewness, Normal Distribution. Median = Mean.**

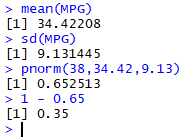
Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

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

MPG <- Cars$MPG

* 1. (MPG>38)



P(MPG>38) = **0.35**

* 1. (MPG<40)



P(MPG<40) = **0.73**

* 1. (20<MPG<50)

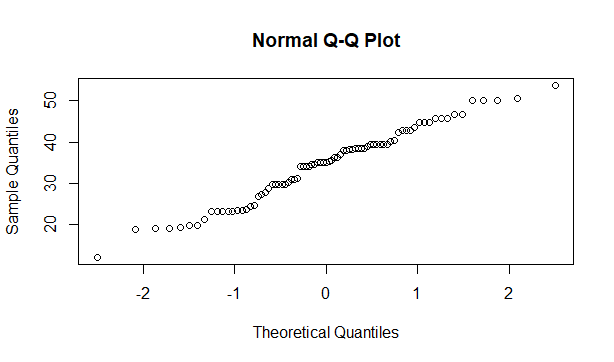


P(20<MPG<50) = 0.898

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

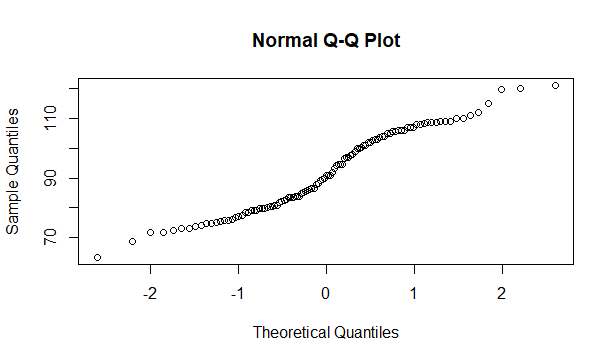
Dataset: Cars.csv

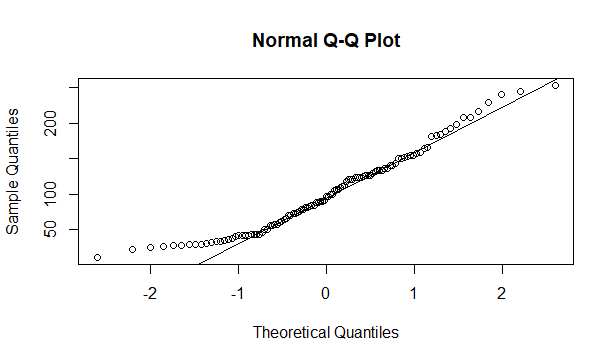


**Follows Normal Distribution as indicated by the qqplot.**

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

Dataset: wc-at.csv

**Wc-at $waist follows Normal Distribution**

**Wc-at$AT follows Normal Distribution**

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

**Ans: Z-Score for 90% C.I. = 1.645**

**Z-Score for 94% C.I. = 1.881**

**Z-Score for 60% C.I. = 0.842**

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

**Ans: t-score for 95% C.I. = 2.064  
t-score for 96% C.I. = 2.172  
t-score for 99% C.I. = 2.797**

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: µ=270, n=18, =260, s=90, df=n-1=17**

**tscore= = = -10/21.23= -0.4714**

**> pt(-0.4714,17)**

**0.3217**