**Q1) Identify the Data type for the Following:**

|  |  |
| --- | --- |
| **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 |

**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 | Interval |
| IQ(Intelligence Scale) | Ordinal |
| Sales Figures | Ratio |
| Blood Group | Ordinal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| Religious Preference | Ordinal |
| 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?**

*SOLUTION:*

**Head= H Tail= T**

Three Coins are tossed, then sample space is

S={(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)}

n(S)= 8

A=Two heads and One tail are obtained, then sample space is,

A={(HHT), (HTH), (THH)}

n(A)=3

probability of two heads and one tail are obtained is,

p= n(A)/n(S)

= 3/8

**= 0.375**

**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**

*SOLUTION:*

Two dice are rolled, then sample space is,

S= {(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. **Sum is Equal to 1**

A= {}

n(A)= 0

p = n(A)/n(S)

**= 0**

1. **Less than or equal to 4**

B= {(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)}

n(B)= 6

p= n(B)/n(S)

= 6/36

=1/6

**=0.167**

1. **Sum is divisible by 2 and 3**

C= {(1,5), (2,4), (3,3), (4,2), (5,1),(6,6)}

n(C)= 6

p=n(C)/n(S)

=6/36

=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?**

*SOLUTION:*

A bag contains 2 red, 3 green and 2 blue balls, then total number of balls is,

2+3+2= 7

Then sample space of two balls are drawn from a bag contains 7 balls is,

7C2 =

= = = **= 21** *n(S)=21*

Suppose, X be the event that 2 balls are drawn from bag which is not blue then, number of ways of drawing 2 balls out of 2+3 balls is,

5C2 =

= = =  **= 10** n(X)=10

P(X) = n(X)/n(S)

= 10/21

**= 0.477**

**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**

*SOLUTION:*

Expected value of candies for randomly selected child =

1\*0.015 + 4\*0.20 + 3\*0.65 + 5\*0.005 + 6\* 0.01 + 2\* 0.120

= 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**

*SOLUTION:*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weigh** |
| **Mean** | 3.5966 | 3.2173 | 17.8488 |
| **Median** | 3.6950 | 3.3250 | 17.7100 |
| **Mode** | 0: 3.92, 1: 3.07 | 3.44 | 0:17.02, 1: 18.90 |
| **Variance** | 0.2858 | 0.9573 | 3.1931 |
| **Standard Deviation** | 0.5847 | 0.9764 | 1.7869 |
| **Range** | 2.17 | 3.911 | 8.399 |

**Interpretation:**

1. Standard deviation of Weigh is more that is 1.78, from this we conclude that, maximum and minimum distance between each and every data points is **17.84 + 1.78 and 17.84 – 1.78**

I.e some data points are **1.78 units** away from the mean.

1. For **‘Score’**, some data points are **0.97 unit** away from the mean.
2. The **‘Points’** Column is bimodal. Points column has **2 modes, 3.92 and 3.07** respectively
3. Similarly, **‘Weigh’** column is also bimodal. It has **2 modes, 17.02 and 18.90** respectively.
4. Range for the ‘Weigh’ column is more, from this we say that, spread of
5. data points is more in weigh column.
6. For ‘Points’ Column median is slightly greater than mean, from this we can say that, the data is slightly ‘skewed to the left’.
7. Also, for the ‘Weigh’ column, median is slightly greater, i.e. the data is slightly ‘skewed to the left’.
8. The mean for Points, **Score and Weigh variable is 3.5966, 3.2173 and 17.8488** respectively.

**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?**

*SOLUTION:*

One person is chosen at random out of 9 patients, then probability is, n(S)=1/9

**=0.111**

Then expected value of the weight of that patient is,

E(X) = 0.111\*108 + 0.111\*110 + 0.111\*123 + 0.111\*134 + 0.111\* 135 + 0.111\*145 + 0.111\*167 + 0.111\*187 + 0.111\* 199

=**145.333**

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

**Cars speed and distance**

**Use Q9\_a.csv**

**SP and Weight(WT)**

**Use Q9\_b.csv**

*SOLUTION:*

1. **For Q9\_a.csv (Cars speed and distance)**
2. Skewness for Cars speed is -0.11. This indicates that, Cars speed data is left skewed. Majority of the data is on the right side.
3. Skewness for distance is 0.8068. This indicates that, distance data is slightly right skewed. Majority of the data is on the left side.
4. Kurtosis for Cars speed is -0.5089. This indicates that, the distribution has lighter tail than normal distribution. The distribution is platykurtic.
5. Kurtosis for distance is 0.4050. This indicates that, the distribution has more in the tail than the normal distribution. The distribution is leptokurtic.
6. **For Q9\_b.csv (SP and Weight(WT))**
7. Skewness for SP is 1.6114. This indicates that, SP data is right skewed. Majority of the data is on the left side of the distribution.
8. Skewness for WT is -0.61. This indicates that, WT data is left skewed. Majority of the data is on the right side of the distribution.
9. Kurtosis for SP is 2.9773. This indicates that, the distribution has more in the tail than the normal distribution. The distribution is leptokurtic.
10. Kurtosis for WT is 0.9502. This indicates that, the distribution has more in the tail than the normal distribution. The distribution is leptokurtic.

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



**Interpretation:**

1. From above histogram, we see that, majority of the data is in left side. From this we can conclude that, ChickWeight$Weight data are right skewed.
2. Majority of the data points are in the interval (50 – 200). About to 50% of the data are in this interval.
3. From histogram, we see that, the longer tail is on the right side of the distribution, from this we can conclude that, outliers are present and they are on the upper extreme side.



**Interpretation:**

1. From above boxplot, we see that the value of upper extreme is very high and the value of lower extreme is very low as compare to each other.
2. From boxplot, we see that the outliers are present in the data. Outliers are present in upper extreme side.
3. We see that the median Is slightly closer to bottom side of the box, so we can conclude that, the distribution is right skewed or 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?**

*SOLUTION:*

Sample = 2000

Sample mean = 200 pounds

Sample standard deviation (s) = 30

**For this we use t-distribution**

94% confidence interval is = (198.738325292158, 201.261674707842)

96% confidence interval is = (198.62230334813333, 201.3776966518666)

98% confidence interval is = (198.43943840429978, 201.5605615957002)

**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?**

*SOLUTION:*

1. **Find mean, median, variance, standard deviation.**

Mean = 41

Median = 40.5

Variance = 25.5294

Standard deviation = 5.0526

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

Average of student’s marks is 41 and standard deviation of the students’ marks is 5.05. From this we can conclude that, the student’s marks deviate 5.05 units from the mean value i.e. 41 – 5.05 and 41 + 5.05

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

**Ans:**  When mean and median are equal then we can say that, the data is symmetric in nature or the data is normally distributed. Their left- and right-hand side tails are equally balanced around the mean.

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

**Ans:**  When mean > median then we can say that, the distribution is positively skewed or the distribution has long tail to the right side of the distribution. Majority of the datapoints are is in left side of the distribution.

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

**Ans:** When median > mean then we can say that, the distribution is negatively skewed or the distribution has long tail to the left side of the distribution. Majority of the data points are is in right side of the distribution.

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

**Ans:** Positive kurtosis value indicates the distribution is Leptokurtic. The distribution show longer tails on either side, indicating large outliers. The distribution has longer tail than normal distribution.

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

**Ans:**  Negative value of kurtosis indicates that the distribution is Platykurtic. The distribution has lighter tail than normal distribution.

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



**What can we say about the distribution of the data?**

From above boxplot, we see that median is slightly closer to the lower end of the box, then we can say that the distribution is positively skewed or right skewed.

**What is nature of skewness of the data?**

The distribution is positively skewed or right skewed. Majority of the data points are in the left side of the distribution.

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

The IQR will be approximately equal to 18 - 10 = 8 units.

**Q19) Comment on the below Boxplot visualizations?**



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

**Ans:**

1. From above two boxplots we see that, both boxplots have same median value
2. One boxplot has short box, this indicates that their datapoints are closer to center value.
3. Another boxplot has taller box, this indicates that their datapoints are far from the center value. They have more dispersion in data.

**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)**

*SOLUTION:*

Number of samples = N = 81

**P(MPG>38) =** Number of samples (MPG) > 38/ Total number of samples

= 67/ 81

= 0.8271

**P(MPG<40) =** Number of samples (MPG) <40 / Total number of samples

= 61 / 81

= 0.7530

**P (20<MPG<50) =** Number of samples (MPG) is in between 20-50 / Total

number of samples

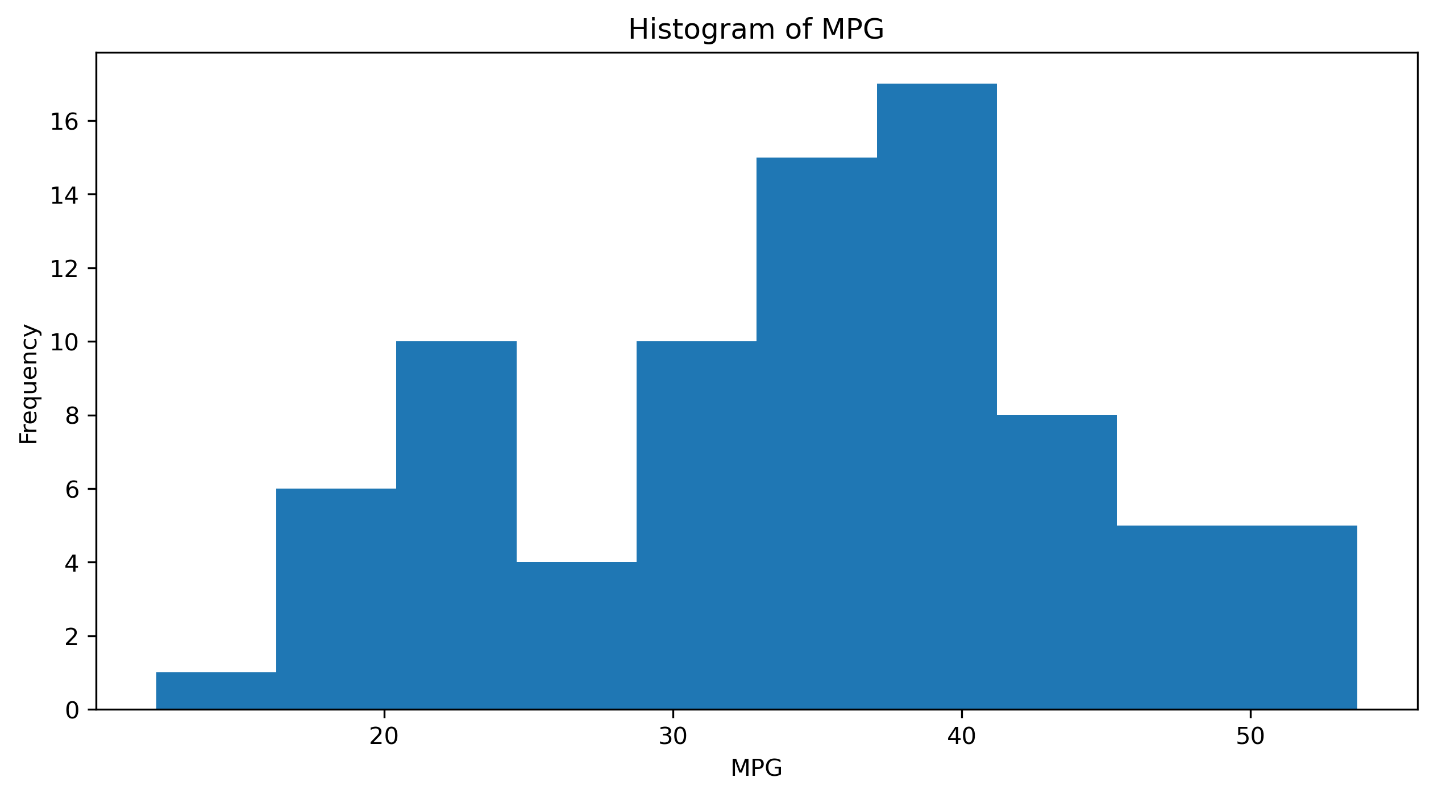
= 69 / 81

= 0.8518

**Q 21) Check whether the data follows normal distribution**

1. **Check whether the MPG of Cars follows Normal Distribution**

**Dataset: Cars.csv**

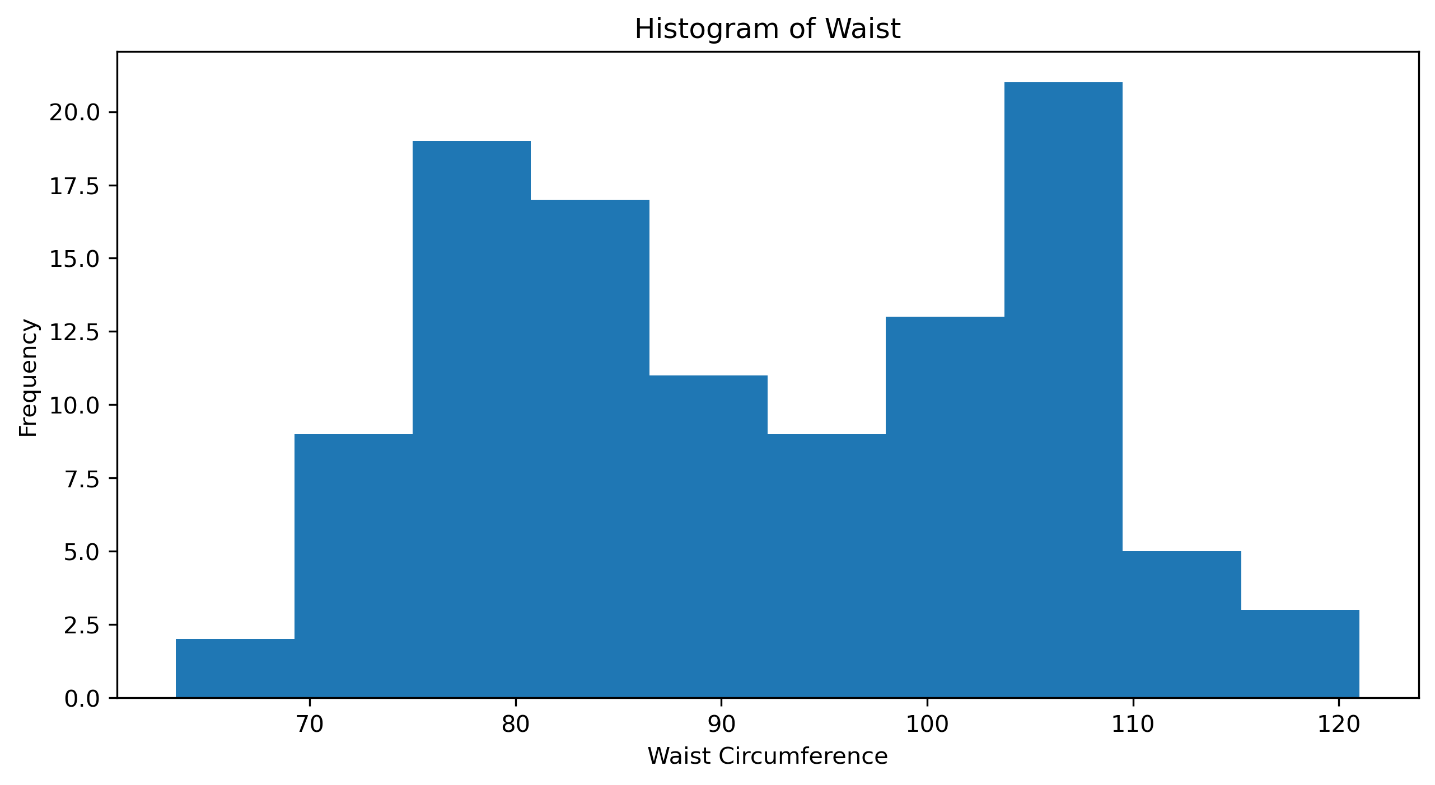
*SOLUTION:*****

From above histogram, we see that, out MPG data is nearly distributed symmetric. There is no log tail to right and left side.

From this we can conclude that, our data (MPG) is normally distributed.

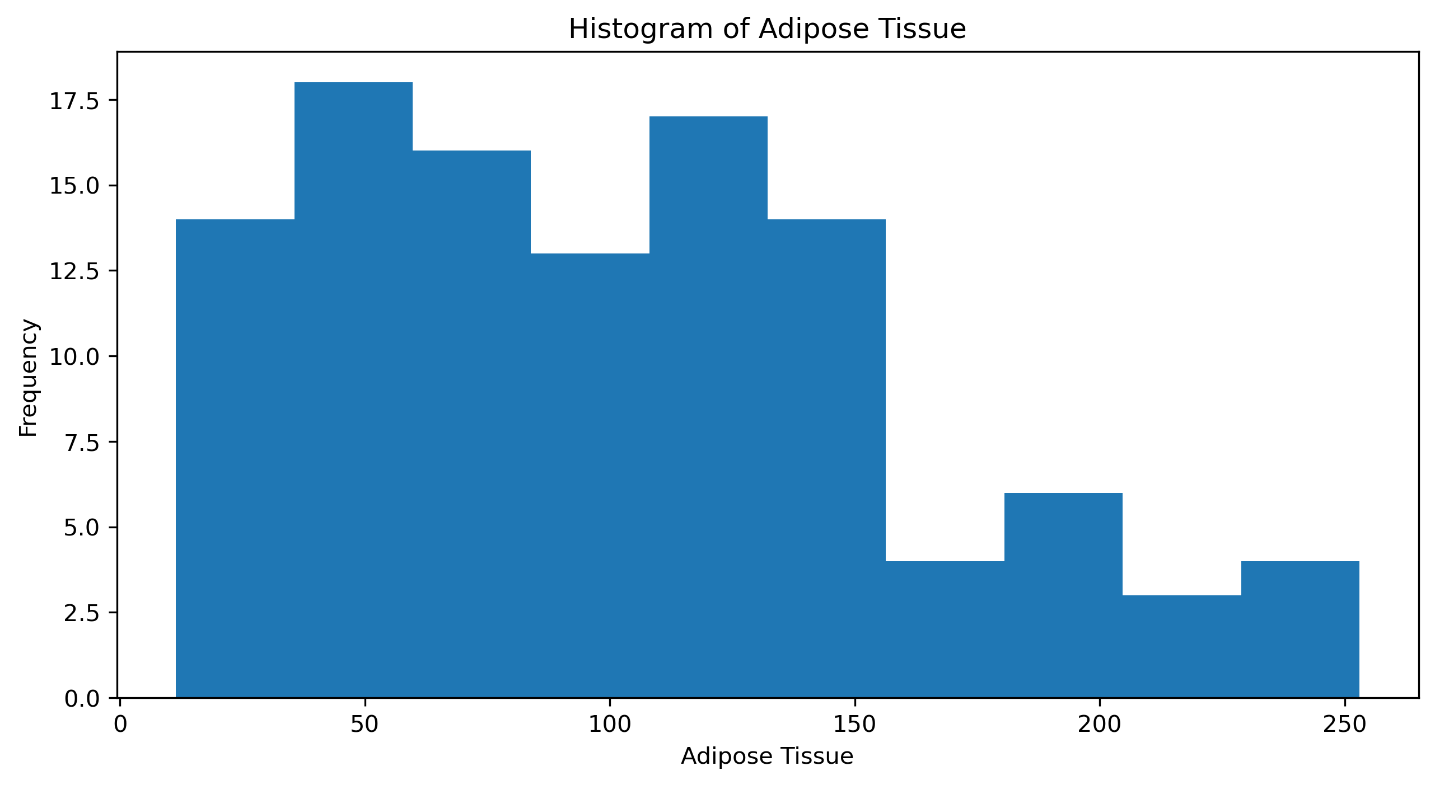
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv



From above histogram, we see that our data is spread everywhere. There is no pattern that show that our data is follows normal distributed or not.

From this, we can concluded that, our data (Waist Circumference) does not follow normal distribution.



From above histogram, we see that the majority of the data is on the left side of the distribution. i.e. Our data is right skewed

We can conclude that, out data (AT) does not follow normal distribution.

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

**Ans:**

Z score value for 90% confidence interval is +-1.96

Z score value for 94% confidence interval is +-1.88

Z score value for 60% confidence interval is +- 0.253

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

**Ans: For one-tailed**

t score value for 95% confidence interval is 1.71

t score value for 96% confidence interval is 1.82

t score value for 99% confidence interval is 2.49

**For Two-tailed**

t score value for 95% confidence interval is +-2.06

t score value for 96% confidence interval is +-2.17

t score value for 99% confidence interval is +-2.79

**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**

*SOLUTION:*

Population mean = u = 270

Sample mean = X = 260

Sample standard deviation = s = 90

Number of samples = n = 18

Then t- statistic is,

**t = x-u/(s/)**

**= 260-270/(90/)**

**= -0.471**

Degrees of freedom is n-1 = 18-1=17

Then the probability that, 18 randomly selected bulbs would have an average life of no more than 260 days is that,

The probability that the t-value we obtained with 17 degrees of freedom

**t< -0.472 with 17 d.f**

**= 0.3218**