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

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

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

Answer: If three coins are tossed then total probability are

\*Here H-Head, T-Tail

(H,H,H),(H,H,T),(H,T,H),(T,H,H),(H,T,T),(T,H,T),(T,T,H),(T,T,T)

So the total probability is 8 , out of it two heads and one tail are (H,H,T),(H,T,H),(T,H,H) is 3… So clealy

P(Having two heads 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

When two dices are rolled , the output will be like (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)

So, the total probabilities are 36.

1. For equal to one, its 0 Probability (Equal to 1) =0/36
2. Less than or equal to 4 is, its (1, 1), (1, 2), (2, 1), (1, 3), (2, 2), (3, 1) so

Probability (Less than or equal to 4) =6/36=1/6.

1. Sum divisible by 2 and 3 is (1, 1), (1, 2), (1, 3), (1, 5), (2, 1), (2, 2), (2, 4), (2, 6), (3, 1), (3, 3),(3, 5), (3, 6), (4, 2), (4, 4), (4, 5), (4, 6), (5, 1), (5, 3), (5, 4), (5, 5), (6, 2), (6, 3), (6, 4), (6, 6)

So the no of terms divisible by 2 and 3 is 24=24/36 or 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. Here formula is nCr=n! /(n-r)!\*r!

Probability (Two balls selected) =7C2=7! /5!\*2!=21

Probability (Excluding two blue balls, other two-color balls selected) = 2red+3green=5

Hence 5C2=10

Formula,P(X)=Favorable Outcome/Total Outcome

Hence, Probability (Two balls selected except blue balls) =

Probability (Excluding two blue balls, other two-color balls selected)/Probability (Two balls selected of any color) =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.

=> No of candies child A expect is 1 \* 0.015=0.015---------------(1)

=> No of candies child B expect is 4\*0.20=0.8-----------(2)

=> No of candies child C expect is 3 \*0.65=1.95-----------(3)

=> No of candies child D expect is 5\*0.005=0.025-----------(4)

=> No of candies child E expect is 6 \*0.01=0.06-----------(5)

=> No of candies child F expect is 2 \* 0.12=0.24-----------(6)

From adding all the equations from 1-6, The numbers of candy for a randomly expected child are

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

=> 3.090

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**

Refer: Q7-U. AGHIL MENON (aghilmenon@gmail.com). IPYNB

**Inferences:**

**Points:** From the above distribution we got, Mean=3.5965, Median=3.6950, Mode=(0):3.07,(1):3.92,Standard Deviation=0.53467 and Variance=0.2858…. this implies

(i)Since Mean is closer to Median and Mode , it is relatively normally distributed.

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?

The expected value of the weight of a random patient will be the average of the all 9 patients= (108+110+ 123+134+135+145+167+187+199)/9

=>1308/9 = 145.33

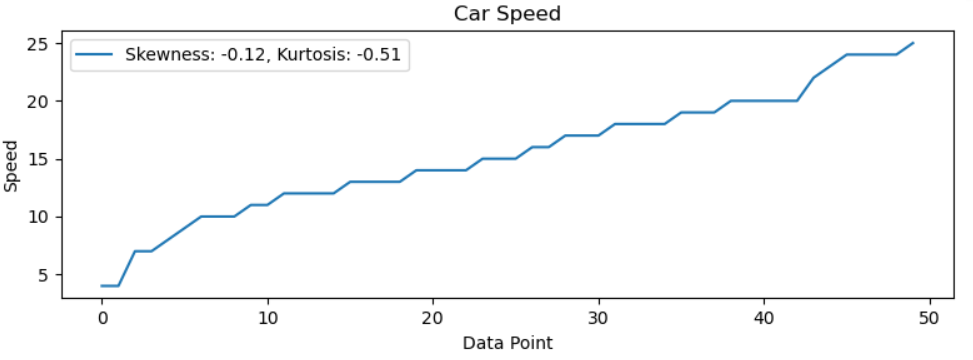
Q9) Calculate Skewness, Kurtosis & draw inferences on the following data

Car’s speed and distance

**Use Q9\_a.csv**

Refer: Q9\_a U. AGHIL MENON ([aghilmenon@gmail.com](mailto:aghilmenon@gmail.com)). IPYNB

**Inferences:**

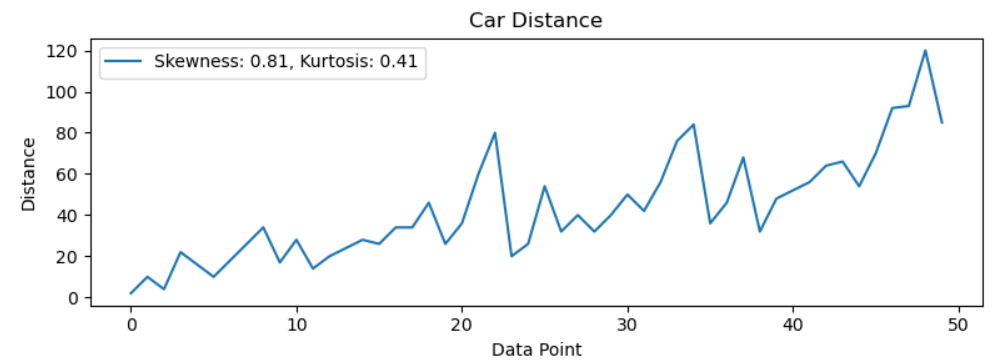
****

1. **Speed:** From the above plot and distribution, for speed we got Skewness=-0.12 and Kurtosis=-0.51…. from that since skewness value is slightly negative, we can infer that distribution is slightly left skewed with longer tails on left end and regarding Kurtosis it is platykurtic …there are fewer extreme values or outliers in the dataset with lower tail ends compared to normal distribution.
2. **Distance:** From the above plot and distribution, for distance we got Skewness=0.81 and Kurtosis=0.41…. from that we can infer since skewness value is slightly positive, we can say distribution is slightly right skewed and longer tail on right end and regarding Kurtosis it is Leptokurtic…. there are many extreme values or outliers in the dataset with heavy tails compared to normal distribution.

**SP and Weight (WT)**

**Use Q9\_b.csv**

Refer: Q9\_b U. AGHIL MENON (aghilmenon@gmail.com). IPYNB



**Inferences:**

1. **SP:** From the above plot and distribution, for speed we got Skewness=1.61 and Kurtosis=2.98…. from that since skewness value is positive, we can infer that distribution is right skewed with longer tails on right end and regarding Kurtosis it is leptokurtic …there are more extreme values or outliers in the dataset with higher tail ends compared to normal distribution.

**WT:** From the above plot and distribution, for distance we got Skewness=

-0.61 and Kurtosis=0.95…. from that we can infer since skewness value is slightly negative, we can say distribution is slightly left skewed and longer tail on left end and regarding Kurtosis it is Leptokurtic…. there are many extreme values or outliers in the dataset with heavy tails compared to normal distribution.

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



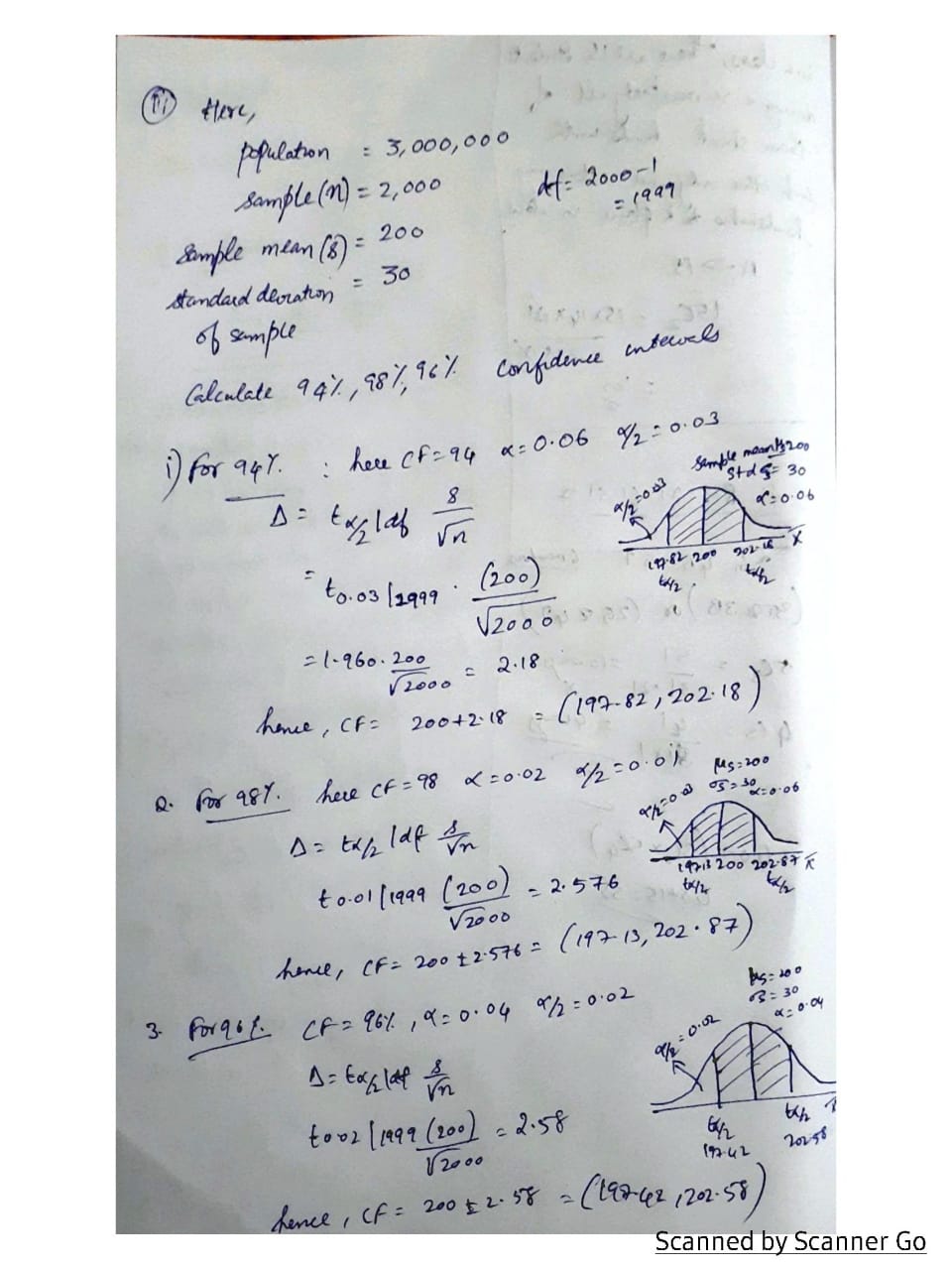
**Answer:**

**a)Histogram plot on Chick weight**:From the above plot, it appears that distribution is right skewed distribution with more chicks weighing in range of 50 to 150.Here in this distribution ,Mean > Median >Mode.

**b) Boxplot:** From the above boxplot it appears that it is a right skewed distribution, and here median is closer to first quartile.

**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?

Refer: Q11-ASSIGNMENT1, U. AGHIL MENON (aghilmenon@gmail.com). IPYNB

Answer:

**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: From the above distribution there are 18 students,

1.1) Mean = Sum of marks of all 18 students /18

=> (34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+45+49+56)/18

=>697/18=38.72, Formula,

1.2) Median=

=> In these marks 34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56 middle value is 8th mark and 9th mark that is 40 and 41, so median is 40+41=81 which is divided by 2 = 40.5 marks

1.3) Variance= Formula,,here which is mean 38.72

((34-38.72) + (34-38.72) + (34-38.72) + (34-38.72) + (34-38.72) + (34-38.72) + (34-38.72) + (34-38.72) + (34-38.72) + (34-38.72) + (34-38.72) + (34-38.72) + (34-38.72) + (34-38.72) + (34-38.72) + (34-38.72) + (34-38.72) + (34-38 / 18 = 30.3333

1.4) Standard Deviation, STD= , which is 5.5066

2. **Inference:** Since the mean 38.72 and meridian 40.5 is almost similar, we can say the marks of the student’s distribution is symmetrically distributed. Similarly, from variance and standard distribution, marks in distribution are moderately variable.

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

Answer: When the mean and median of a data set are equal, it indicates that the data is approximately symmetrically distributed. It implies that skewness is Zero skewed.

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

Answer: When the mean is greater than the median, it indicates that the data set is positively skewed or right-skewed. Also, we can witness the following character of right skewed distribution

* Distribution’s Tail will extend more towards right.
* There will be many values towards the right side of distribution
* Median will be attracted towards the lower end of distribution’s data series.
* Mean will be attracted towards the right end of distribution’s data series.

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

Answer: When the median is greater than the mean, it indicates that the data set is negatively skewed or left-skewed. Also, we can witness the following character of left skewed distribution

* Distribution’s Tail will extend more towards left.
* There will be many values towards the left side of distribution
* Median will be attracted towards the higher end of distribution’s data series.
* Mean will be attracted towards the left end of distribution’s data series.

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

Answer: Positive kurtosis also known as Leptokurtic in which dataset has longer tails and have extreme values along outliers than the normal distribution. This type of distributions will have highest peaks and also mean will be mostly concentrated on it.

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

Answer: Negative kurtosis also known as Platykurtic in which dataset has shorter tails and have fewer extreme values along outliers. This type of distributions will have flat peaks and data will be dispersed and also less concentrated along mean.

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



What can we say about the distribution of the data?

Answer: In this boxplot data is heavily distributed along the left side of the plot.

What is nature of skewness of the data?

Answer: The boxplot is left skewed since median is located much near to third quartile.

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

Answer: In the boxplot it is clear that Q1 here is approx. 10 and Q3 is approx. 18, so the IQR will be difference of Quartiles 3 and 1, Q3-Q1 i.e.) 18-10 which will be around 8. Hence IQR=8  
  
  
Q19) Comment on the below Boxplot visualizations?



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

Answer: From the Boxplot 1, we get the following data, that is Minimum=237.5, Quartile Q1=250, Q3=275 and Maximum will be 287. 5.. from that we will get Median around (250+275)/2=262.5.

From the Boxplot 1, we get the following data, that is Minimum=200, Quartile Q1=225, Q3=300 and Maximum will be 337. 5.. from that we will get Median around (225+300)/2=262.5.

Hence from the above two boxplot distributions, the data distribution is less in Boxplot 1 compared to Boxplot 2. Even though both the boxplots have same value as median , but the IQR range of Boxplot1 is 25 and Boxplot2 is 75 which says IQR of Boxplot2 is 3 times bigger than Boxplot1…Nevertheless the range both the boxplots 1 and 2 are no skewed.

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)

Refer: Q20.Assignment1(U. AGHIL MENON) ([aghilmenon@gmail.com).IPYNB](mailto:aghilmenon@gmail.com).IPYNB)

Answer: From the IPYNB file above we got Mean of MPG=34.422 and Standard Deviation as 9. 1314.Here µ is mean and is Standard Deviation

**20.a) P(MPG>38):**

Standardizing the above data, we get ===0.39183=0.65173(From Z-Table)

Since P(MPG>38) it is 1-0.65=0.35=35%

**20.b) P(MPG<40):**

Standardizing the above data, we get ===0.6108=0.72907(From Z-Table)

Since P(MPG>38) it is 0.72907=72.9%

**20.c) P(20<MPG<50):**

Here we have to calculate z values of both the 20 and 50 and we have to take difference of Z\_50 and Z\_20

For the value of Z\_50

Standardizing the above data, we get ===1.705=0.95994(From Z-Table)

For the value of Z\_20

Standardizing the above data, we get ===-1.579=0.05821(From Z-Table)

Since P(20>MPG>50) it is 0.95994-0.05821=0.8988=89%

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

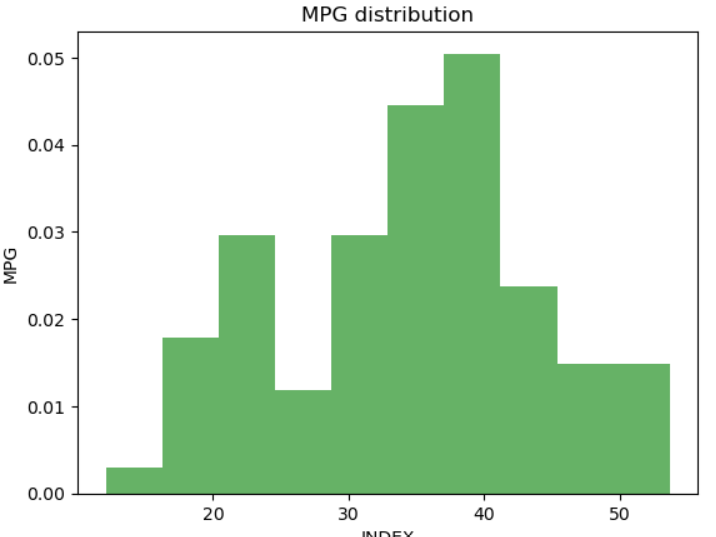
Dataset: Cars.csv

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

Dataset: wc-at.csv

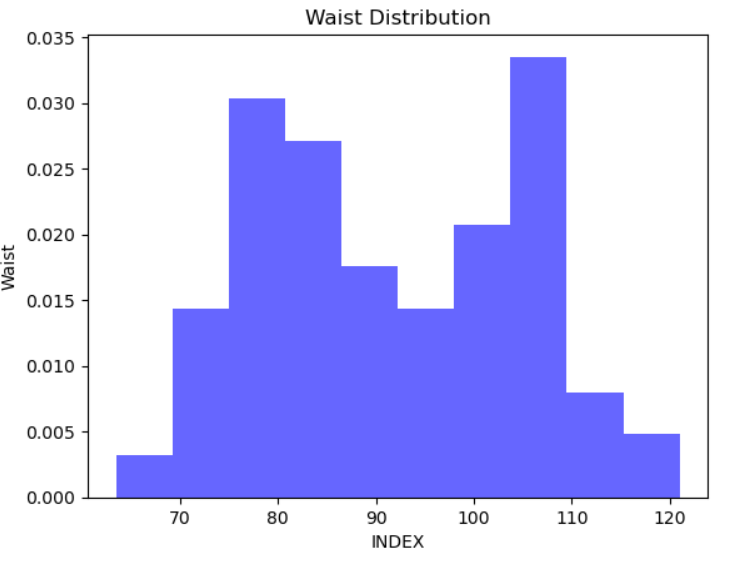
21.a) Refer: Q20.Assignment1(U. AGHIL MENON) ([aghilmenon@gmail.com).IPYNB](mailto:aghilmenon@gmail.com).IPYNB)

From the above distribution , the histogram plot will appear like this

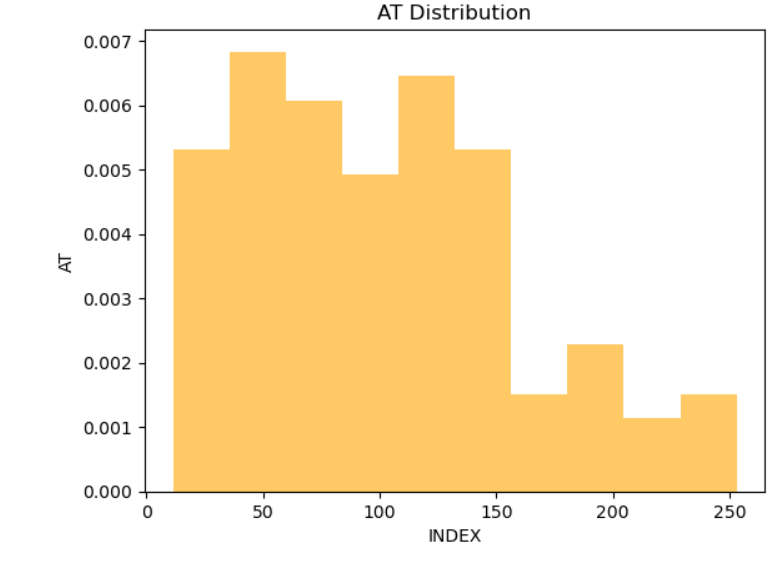


The mean and median values are 34.12 and 35.1527 it is approximately normal distribution as the distribution follows the bell curve…But is slightly left skewed but nevertheless it has bell curve and Gausian distribution hence it is normal curve.

21.b) For Waist distribution, the following Histogram plot appears like this



Which clearly shows it has two peaks and follows a bimodal distribution, hence it is clearly not a normal distribution

For AT (Adipose Tissue) distribution, the following Histogram plot appears like this

Which clearly shows it right skewed distribution as many data lies on the right side of the plot , hence it is clearly not a normal distribution

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

REFER: Q22&Q23 ASSIGNMENT1(U.AGHIL MENON),aghilmenon@gmail.com.IPYNB

Answer: In order to calculate the Z-scores of a confidence interval, we have to standardize the given plot. Here Z score will be a standard deviation from mean.

1. For 90% confidence interval, In the scale of 1.0 in standardized distribution 90%, we will have the value as 0.90 and the rest is given as significance level (α)=0.10…since in a distribution we have to find the interval, it is given as α/2=0.05

In the Z-table the value of 0.05 is 1.64

1. For 94% confidence interval, In the scale of 1.0 in standardized distribution 94%, we will have the value as 0.94 and the rest is given as significance level (α)=0.06…since in a distribution we have to find the interval, it is given as α/2=0.03

In the Z-table the value of 0.03 is 1.88

1. For 60% confidence interval, In the scale of 1.0 in standardized distribution 60%, we will have the value as 0.60 and the rest is given as significance level (α)=0.40…since in a distribution we have to find the interval, it is given as α/2=0.20

In the Z-table the value of 0.20 is 0.84.

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

REFER: Q22&Q23 ASSIGNMENT1(U.AGHIL MENON),aghilmenon@gmail.com.IPYNB

Answer: In order to calculate the t-scores of a confidence interval, we have to find the value from the respective t table ... Here t-score will be a standard deviation from mean with also the consideration of Degrees of freedom (Df). Since the sample value is 25, the respective (Df) will be kept as 24.

1. For 95% confidence interval, In the scale of 1.0 in t-distribution 95%, we will have the value as 0.95 and the rest is given as significance level (α)=0.05…since in a distribution we have to find the interval, it is given as α/2=0.025.

In the t-table the value of 0.025 and Degrees of freedom (Df) of 24 is 2.064

1. For 96% confidence interval, In the scale of 1.0 in t-distribution 96%, we will have the value as 0.96 and the rest is given as significance level (α)=0.04…since in a distribution we have to find the interval, it is given as α/2=0.02.

In the t-table the value of 0.02 and Degrees of freedom (Df) of 24 is 2.492

1. For 99% confidence interval, In the scale of 1.0 in t-distribution 99%, we will have the value as 0.99 and the rest is given as significance level (α)=0.01…since in a distribution we have to find the interval, it is given as α/2=0.005.

In the t-table the value of 0.005 and Degrees of freedom (Df) of 24 is 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

REFER: Q24,Assignment1,(U.AGHIL MENON)(aghilmenon@gmail.com)

Answer:

Here ; n=18

; =260

Here ,

Since Confidence interval is not given, we assume it as 95% and α=0.05

Also since Standard Deviation of sample is not given, we will be taking t-test. = ……Here = =-0.4714=P=0.3218, We have considered this by considering Degree of Freedom as 17.

P>α => Null Hypothesis is accepted…the average bulb lasts 270 days.