Q1) Identify the Data type for the Following:

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

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 |
| Socio-economic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| 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?

P(x) = P(HHT) + P(HTH) + P(THH)

= () + () + ()

=

Q4) Two Dice are rolled, find the probability that sum is

| S/No. | 1 | 2 | 3 | 4 | 5 | 6 |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 6 | 7 | 8 | 9 | 10 | 11 | 12 |

1. Equal to 1

No events with sum equal to 1 are ***Zero***

1. Less than or equal to 4

Possible events with sum <= 4 are [(1,1), (2,1), (1,2), (3,1), (1,3), (2,2)]

P[x<=4] =

1. Sum is divisible by 2 and 3

6 & 12 are divisible by both 2 and 3

Frequency of 6 = 5

Frequency of 12 = 1

P(x) =

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?

P[x] =

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

P[x] =

= 3.09

Expected number of candies for a randomly selected child = 3.09

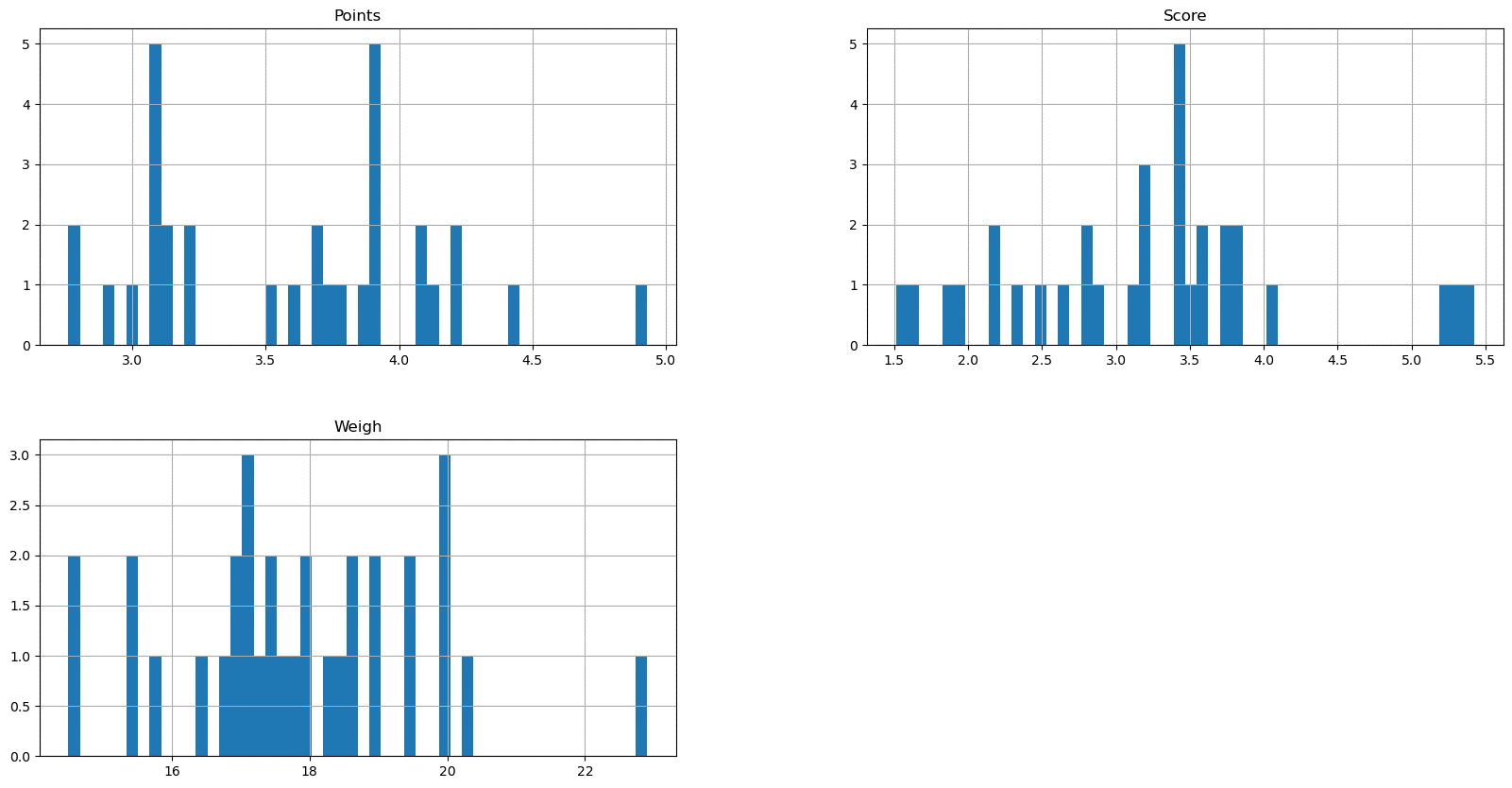
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given data-set

* For Points, Score, Weigh>

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

|  |  |  |  |
| --- | --- | --- | --- |
| Characteristics | Points | Score | Weigh |
| Mean | 3.5965625 | 3.21725 | 17.84875 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | [3.07, 3.92] | 3.44 | [17.02, 18.9] |
| Standard Deviation | 0.534678736 | 0.978457443 | 1.786943236 |
| Variance | 0.285881351 | 0.957378968 | 3.193166129 |
| Range | 2.17 | 3.911 | 8.4 |

1. No column has mean = median = mode to form a symmetrical curve.
2. ‘Points’ and ‘Weigh’ are bimodular data.
3. In ‘Score’ column mode>median>mean, we say it is negatively skewed distribution.
4. ‘Weigh’ has highest standard deviation. So, this data is more spread across mean and also possible to have more outliers than other column’s data.
5. In the histograms below, we can see the weighted distribution tilted towards right, so we expect positive skewness in all histograms.



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?

We can take mean as central measure, so excepted value of the weight of that patient is mean

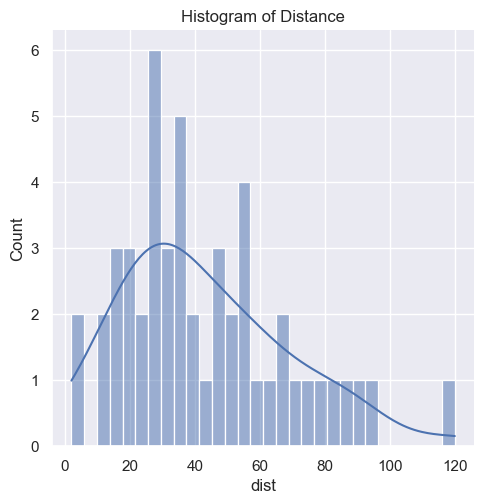
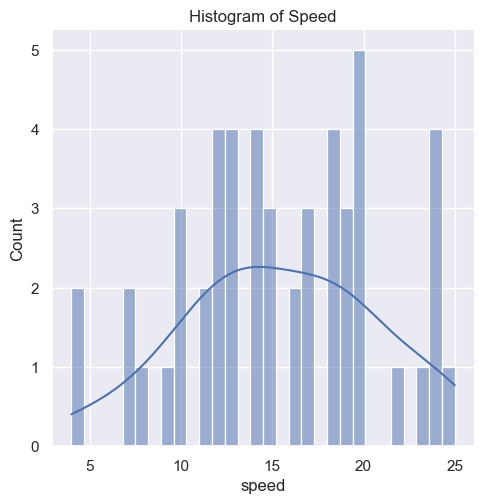
Mean =

=

**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data Car’s speed and distance**

**Use Q9\_a.csv**

|  |  |  |
| --- | --- | --- |
| **Columns** | **Skewness** | **Kurtosis** |
| **Speed** | -0.117509861 | -0.50899442 |
| **Distance** | 0.80689496 | 0.405052582 |

****

Speed:

1. Skewness is negative and in range of (-0.5, 0.5) so we can assume that the distribution to be approximately symmetric.
2. Kurtosis is negative (or less than 3), it refers to Platykurtic. The distribution set follows the subtle or pale curve, and that curve indicates the small number of outliers in a distribution.

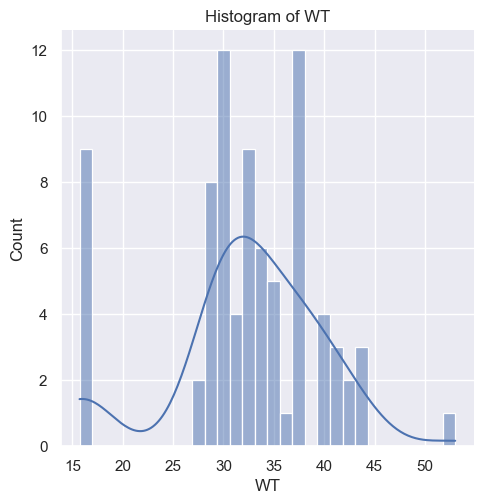
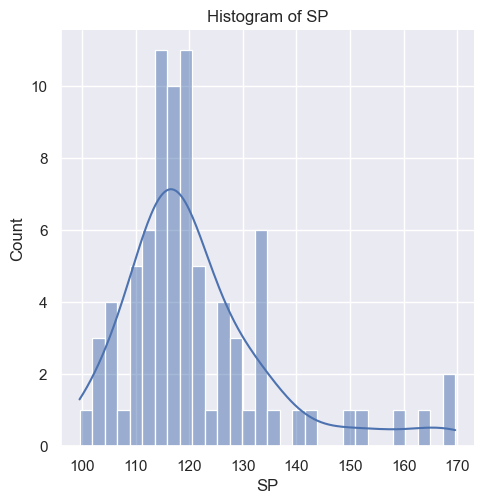
Distance:

1. Skewness is greater than 0.5, we consider the distribution to be **positively skewed or right-skewed** where data points cluster on the left side and the tails are longer on the right side of the distribution.
2. Kurtosis is 0.405 (or less than 3), it refers to Platykurtic. The distribution set follows the subtle or pale curve, and that curve indicates the small number of outliers in a distribution.

**SP and Weight (WT)**

**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
| **Columns** | **Skewness** | **Kurtosis** |
| **SP** | 1.611450196 | 2.9773289 |
| **WT** | -0.61475333 | 0.9502915 |

****

SP:

1. Skewness is greater than 0.5, we consider the distribution to be **positively skewed or right-skewed** where data points cluster on the left side and the tails are longer on the right side of the distribution.
2. Kurtosis is nearly equal to 3, it refers to **Mesokurtic** It means that the data set follows a **normal distribution**.

WT:

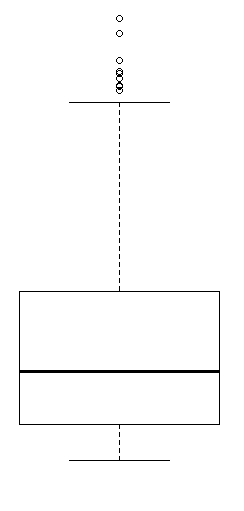
1. Skewness is negative and in range of (-0.5, 0.5) so we can assume that the distribution to be approximately symmetric
2. Kurtosis is 0.405 (or less than 3), it refers to Platykurtic. The distribution set follows the subtle or pale curve, and that curve indicates the small number of outliers in a distribution.

**Q10) Draw inferences about the following box-plot & histogram**



Inferences:

1. Mode is in between 50 to 100.
2. It is positively skewed and this means majority of the data distribution will be on the left side of the mean, while the lower ranging values will be on the right side of the curve.
3. Direction of most of the outliers, which is on the right side of the curve in the tail.



Inferences:

1. It is positively skewed data as median tends toward lower quartile.
2. Outliers are above upper extreme.

**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?

Confidence interval (CI) =

Where : Sample mean

z: z-scores from z-score table

S: Standard Deviation

n: Sample Size

N: Population Size

Given

: 200; S: 30; n:2,000; N: 3,000,000

With a 94% confidence level, (100 -minus 94) ÷ 2 = 3 is the area in each tail of the standard normal curve. 100 -minus 3 = 97 indicates the cumulative area up to the second tail. Find 97 ÷ 100 = 0.97 in the z-table to get a z-score of 1.8 + 0.08 = 1.88.

(CI) =

= 200 ± 1.261

Out of 100 samples, on average in 94 of them the population mean would be between 198.739 and 201.261.

|  |  |  |
| --- | --- | --- |
| **Confidence** | **Z\_Scores** | **Confidence Interval** |
| 0.94 | 1.881861476 | [198.73802960115347, 201.26197039884653] |
| 0.98 | 2.328214776 | [198.43870647955293, 201.56129352044707] |
| 0.96 | 2.055089963 | [198.6218631219836, 201.3781368780164] |

**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.

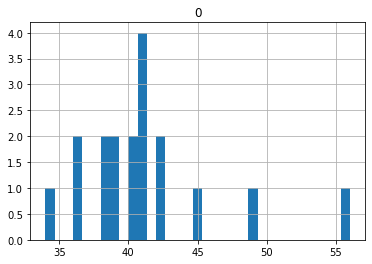
Mean =

= 41

Median = 40.5(

Variance = 24.11111111111111

Standard Deviation = 4.910306620885412

1. What can we say about the student marks?
   1. From the histogram we can say the distribution is towards left and Mean = Mode > Median, so it is having positive skewness.
   2. ‘56’ is outlier.
   3. Median is central point of skewed distribution. i.e., 40.5

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

When Mean = Mode = Median then it is normal distribution. In normal distribution skewness of distribution is Zero and is symmetric shape or bell curve.

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

Positively skewed and the distribution is more towards the left.

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

Negatively skewed and the distribution is more towards the right.

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

A distribution with a 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?

Negative values of kurtosis indicate that a distribution is flat and has thin tails

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



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

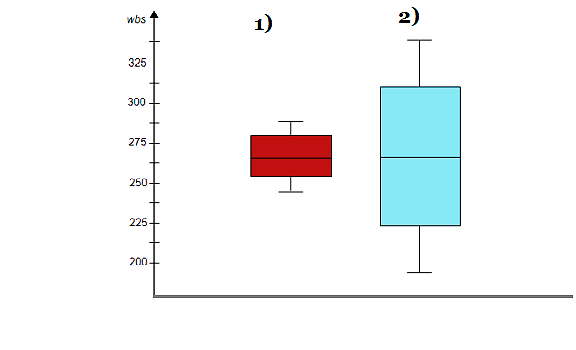
The weight of distribution tends towards higher values.

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

From the plot we can see median is 15.5 approx. Median lies more towards third quartile, so the distribution is negatively skewed.

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

The interquartile range is indicated by the length of the box.  
IQR (approx.) = 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.

1. The spread of data in Boxplot-1 is less compare to Boxplot-2.
2. Median for both plots is nearly same.
3. Outlier aren’t present in Boxplot-1 and Boxplot-2
4. Inter-quartile range is less in Boxplot-1 compare to Boxplot-2
5. Both plots are nearly symmetric and normal distribution.

Q 20) Calculate probability from the given dataset for the below case. Data \_set: Cars.csv

Calculate the probability of MPG of Cars for the below cases. MPG <- Cars $ MPG

Python code:

import pandas as pd

from scipy.stats import norm

df = pd.read\_csv('path’)

df.Head()

m=df['MPG'].mean();s=df['MPG'].std();

print(m,'\n' ,s)

# P(MPG>38)

print('P[MPG>38]:',1-norm.cdf(38,m,s))

# P(MPG<40)

print('P[MPG>38]:',norm.cdf(38,m,s))

# P(20>MPG>50)

print('P[20>MPG>50]:',norm.cdf(20,m,s)-(1-norm.cdf(50,m,s)))

* 1. P(MPG>38)

P[MPG>38] = 0.3475939251582705

* 1. P(MPG<40)

P[MPG>38] = 0.6524060748417295

* 1. P (20<MPG<50)

P[MPG>38] = 0.013116469610523339

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

from scipy. stats import shapiro

# Conduct the Shapiro-Wilk Test

statistic, pvalue =shapiro(df['MPG'])

#Null Hypothesis

print(pvalue)

alpha= 0.05

#H0 Data is normal distribution

#H1 Data is not normal distribution

if pvalue < alpha:

print ("Ho is rejected and H1 is accepted")

else:

print ("H1 is rejected and Ho is accepted")

Output:

0.17639249563217163

H1 is rejected and Ho is accepted

We can use Shapiro will test for normality test.

“Data is nearly following normal distribution.”

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

Dataset: wc-at.csv

import matplotlib.pyplot as plt

import seaborn as sns

from scipy.stats import shapiro

df1 = pd.read\_csv('C:\\Users\\nanir\\OneDrive\\excelR\\Assignments\\Assig-1\\wc-at.csv')

df1.head()

# Conduct the Shapiro-Wilk Test

statistic,pvalue =shapiro(df1['Waist'])

#Null Hypothesis

print(pvalue)

alpha= 0.05

#H0 Data is normal distribution

#H1 Data is not normal distribution

if pvalue < alpha:

print ("Ho is rejected and H1 is accepted")

print ('Data of Waist is not normal distribution')

else:

print ("H1 is rejected and Ho is accepted")

print ('Data of Waist is normal distribution')

plt.figure(figsize=(16,9))

sns.set()

sns.displot(df1['Waist'], kde=True, bins = 30)

plt.title("Histogram of Waist") #for histogram title

statistic, pvalue =shapiro(df1['AT'])

#Null Hypothesis

print(pvalue)

alpha= 0.05

#H0 Data is normal distribution

#H1 Data is not normal distribution

if pvalue < alpha:

print ("Ho is rejected and H1 is accepted")

print ('Data of AT is not normal distribution')

else:

print ("H1 is rejected and Ho is accepted")

print ('Data of AT is normal distribution')

sns.displot(df1['AT'], kde = True, bins = 30)

plt.title("Histogram of AT") # for histogram title

output:

0.0011704873759299517

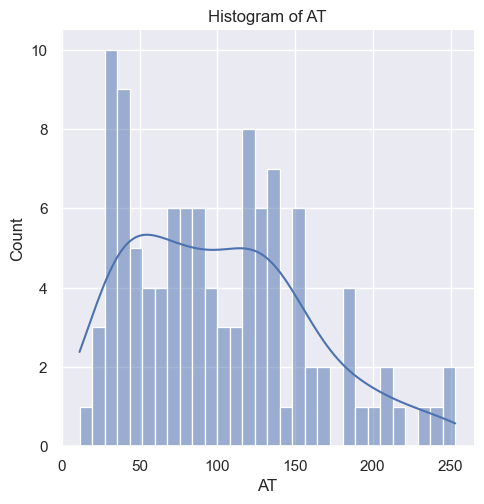
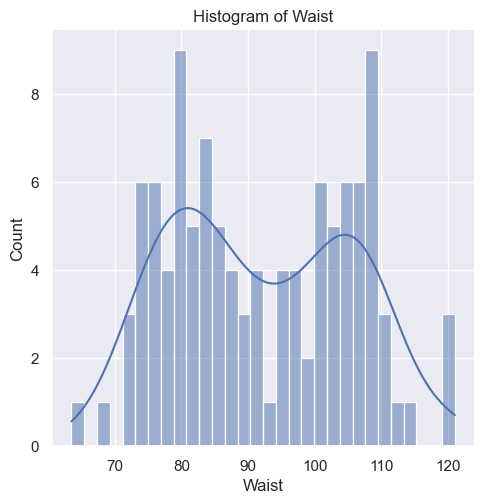
Ho is rejected and H1 is accepted

Data of Waist is not normal distribution

0.0006539996829815209

Ho is rejected and H1 is accepted

Data of AT is not normal distribution



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

We get z-score from z-score tables. From python we can use stats.norm.ppf function.

from scipy import stats

stats.norm.ppf (0.94)

Output: 1.5547735945968535

|  |  |
| --- | --- |
| **Confidence Interval** | **z-scores** |
| 90% | 1.281551566 |
| 94% | 1.554773595 |
| 60% | 0.253347103 |

Q23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25?

By using python scipy.stats module

from scipy import stats

stats.t.ppf(0.95,24)

Output: 1.71088

|  |  |
| --- | --- |
| **Confidence Interval** | **t-values** |
| 95% | 1.71088 |
| 96% | 1.82805 |
| 99% | 2.49216 |

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

x̄ = 260 days

μ = 270 days

s = 90 day

n = 18

df = 18-1 = 17

Therefore,s

t =

t =

= -0.471

> pt(-0.47,17)

[1] 0.3221639

Req Prob= 0.32= 32%