Q1) Identify the Data type for the Following:

Answers Below:

|  |  |
| --- | --- |
| 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 | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Interval |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

The total number of possible outcomes when tossing three coins is 23 = 8 ( As coin has two sides: Heads and Tails)

Let is determine the numbers of outcomes that satisfying the condition of getting two heads and one tails: HHT, HTH, THH

The probability if getting two heads and one tail is the number of favorable outcomes divided by the total number of outcomes:

Probability = Number of favorable outcomes/total number of outcomes = 3/8

Hence Probability of two heads and one tail are 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

Since 6 sides

1. Equal to 1: When two dice are rolled, the minimum sum you can get is 2 (1 on each die). It's not possible to get a sum of 1. So, the probability of getting a sum equal to 1 is 0.
2. Less than or equal to 4 : To find the probabilities for the sum less than or equal to 4,

let's list out the possible combinations:

Possible combinations for sums less than or equal to 4:

Sum = 2: (1, 1)

Sum = 3: (1, 2), (2, 1)

Sum = 4: (1, 3), (2, 2), (3, 1)

There are a total of 3 favorable outcomes out of the total 36 possible outcomes (since there are 6 sides on each dice).

So, the probability of getting a sum less than or equal to 4 is

3/36 = 1/12.

1. Sum is divisible by 2 and 3: The sums divisible by both 2 and 3 are multiples of 6. Let's find the combinations for sums that are multiples of 6:

Possible combinations for sums divisible by 6: Sum = 6: (1, 5), (2, 4), (3, 3), (4, 2), (5, 1)

There are 5 favorable outcomes out of the total 36 possible outcomes.

So, the probability of getting a sum divisible by both 2 and 3 is 6/36 = 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?

To find the probability that none of the balls drawn is blue, you'll want to determine the total number of ways to draw two balls from the bag without getting any blue balls, divided by the total number of ways to draw two balls from the bag.

Total number of balls = 2 (red) + 3 (green) + 2 (blue) = 7 balls

Total number of ways to draw 2 balls out of 7 = 7C2 = 21 combinations

Number of ways to draw 2 balls without getting any blue balls: There are 5 balls that are not blue (2 red + 3 green) = 5C2 = 10 combinations

Therefore, the probability that none of the balls drawn is blue is 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

Expected number of candies = sum of (Candies count) × (Probability) for each child

For the given probabilities:

Child A: Candies count = 1, Probability = 0.015

Child B: Candies count = 4, Probability = 0.20

Child C: Candies count = 3, Probability = 0.65

Child D: Candies count = 5, Probability = 0.005

Child E: Candies count = 6, Probability = 0.01

Child F: Candies count = 2, Probability = 0.120

Expected number of candies = (1 \* 0.015) + (4 \* 0.20) + (3 \* 0.65) + (5 \* 0.005) + (6 \* 0.01) + (2 \* 0.120)

= 0.015 + 0.80 + 1.95 + 0.025 + 0.06 + 0.24 = 3.095

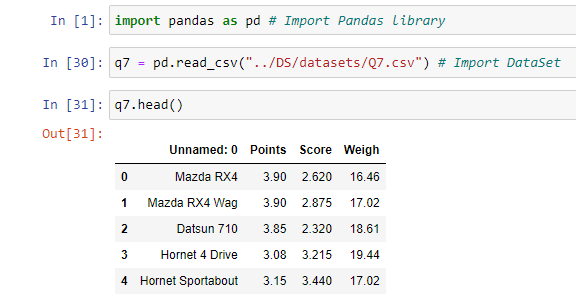
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points, Score, Weigh>

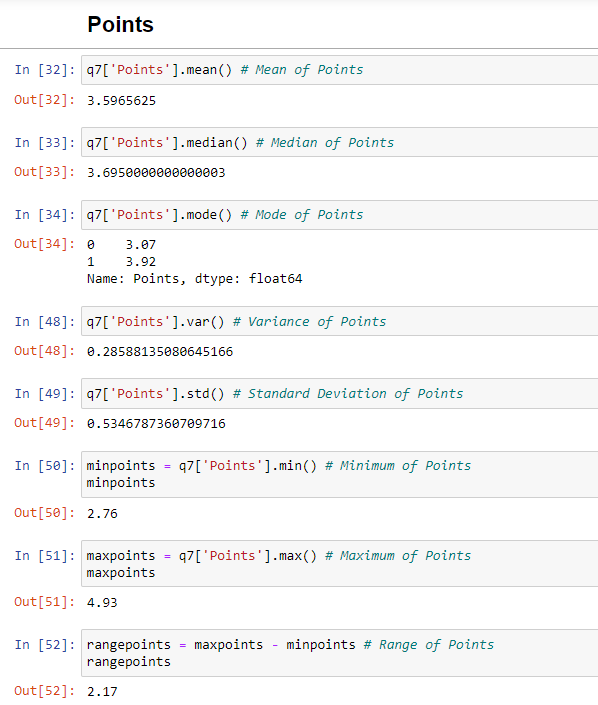
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

**Answer:**

****

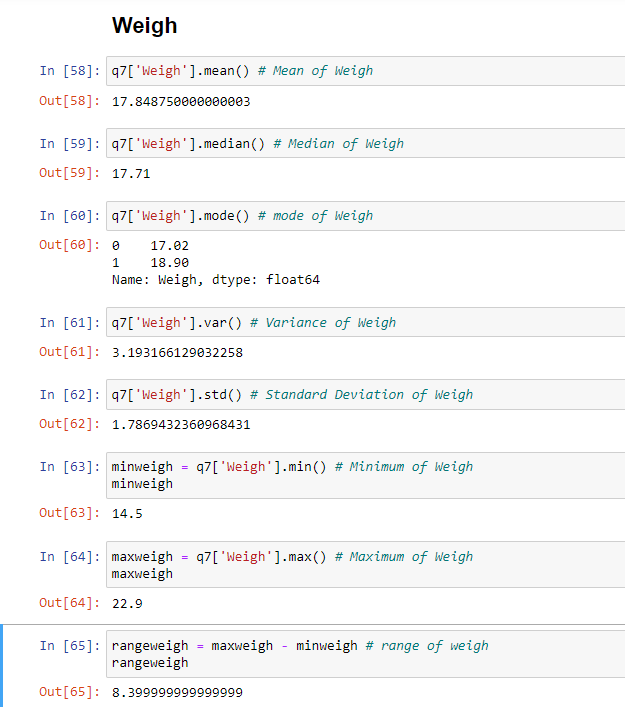
**Points -** Mean, Median, Mode, Variance, Standard Deviation, and Range

****

**Score -** Mean, Median, Mode, Variance, Standard Deviation, and Range



**Weigh -** Mean, Median, Mode, Variance, Standard Deviation, and Range

****

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?

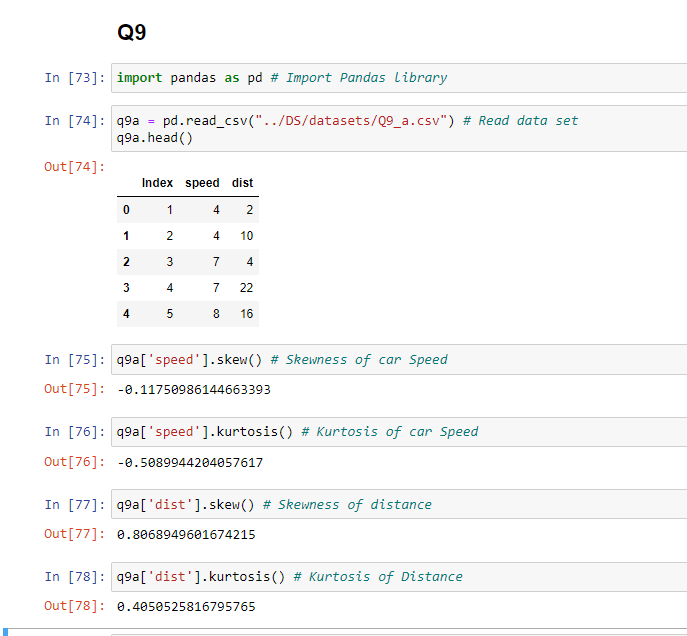
Expected Value = (1/9)\*108 + (1/9)\*110 + (1/9)\*123 + (1/9)\*134 + (1/9)\*135 + (1/9)\*145 + (1/9)\*167 + (1/9)\*187 + (1/9)\*199

= 145.33

**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**

****

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



Answer:

Box plot and histogram used to represent data Visually to find outliers and outlier values.

Box plot represents that data is symmetrical, grouping of data and skewness of data.

Histogram represents frequency distribution of continuous data.

From the above graph histogram:

Histogram is right skewed, Highest values on the right side and mean is greater than median.

From the above graph box plot:

Boxplot is positive skewed and outliers are above the third 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?

Answer:

N = 3000000

n = 2000

mean = 200

Std = 30

Degree of freedom df = 2000 – 1 = 1999

from scipy.stats import t

# Degrees of freedom

df = n-1 = 200 -1 = 199

# Confidence level

confidence\_level = 0.94

# Calculate the critical t-value

critical\_t = t.ppf((1 + confidence\_level) / 2, df)

print("Critical t-value:", critical\_t)

Confidence Interval = mean +|- t \* std/sqrt(n)

1. 94% confidence interval

t c,df  = 1.89

=1.89 \* 30/sqrt(2000) = 56.7 / 44.72 = 1.267

Range estimate [200+1.267, 200-1.267]

Range estimate for 94% confidence interval is [201.27. 198.73]

1. 96% confidence interval

t c,df  = 2.067

=2.067 \* 30/sqrt(2000) = 62.01 / 44.72 = 1.386

Range estimate [200+1.386, 200-1.386]

Range estimate for 96% confidence interval is [201.39. 198.61]

1. 98% confidence interval

t c,df  = 2.345

=2.345 \* 30/sqrt(2000) = 70.35/ 44.72 = 1.573

Range estimate [200+1.573, 200-1.573]

Range estimate for 98% confidence interval is [201.57. 198.43]

**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?

Answer:

1. Mean = 738/18 = 41

Median = (40+41)/2 = 40.5

Variance = 1/N(x-mean)2 = 433/17 = 25.47

Standard Deviation = sqrt(variance) = 5.05

1. Most o the students marks lied between 35 to 45

Average marks obtained by students is 40.5 and Standard deviation of +5.05 to -5.05 marks in a class.

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

Answer: Symmetric and Skewness is Zero.

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

Answer: Skewness is Positive when mean>median.

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

Answer: Skewness is Negative when median> mean

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

Positive values of kurtosis indicate that distribution is peaked and has thick tails.

An extreme positive kurtosis indicates a distribution where most of the numbers are located in the tails of the distribution instead of mean.

Peak of distribution is higher than the peak of the Normal distribution.

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

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

Peak of distribution is lower than the peak of the Normal distribution.

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



What can we say about the distribution of the data?

Considering Q1 = 10, Q2 = 15, Q3 = 18

Median is closure to Q3. Whisker is shorter at the upper end. Mean<Median

What is nature of skewness of the data?

Negative Skewness

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

* **Compare the Medians —**The Median is at the center of the Box. Here the Median line of Plot 2 and Plot 1 are the same.
* **Compare the Dispersion or Spread of data —** The Inter Quartile range (length of the box) gives us an idea about how dispersed the data is. Here Plot 2 has a longer length than Plot 1 which means that the dispersion of data is more in plot 2 as compared to plot 1. The length of whiskers also gives an idea of the overall spread of data. The extreme values (minimum & maximum) give the range of data distribution. Larger the range, the more scattered the data. Plot 2 has a larger range than Plot 1.

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)

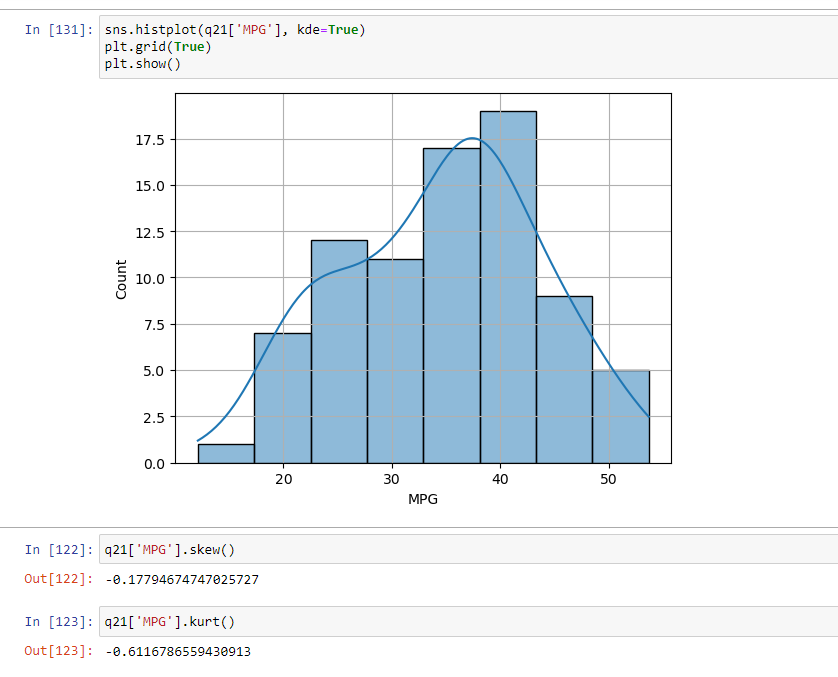
c. P (20<MPG<50)



Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

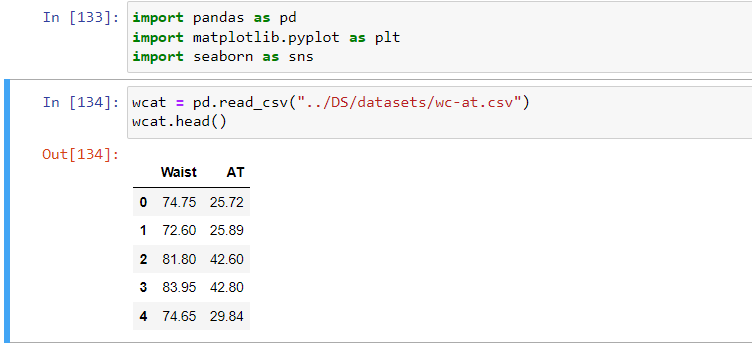
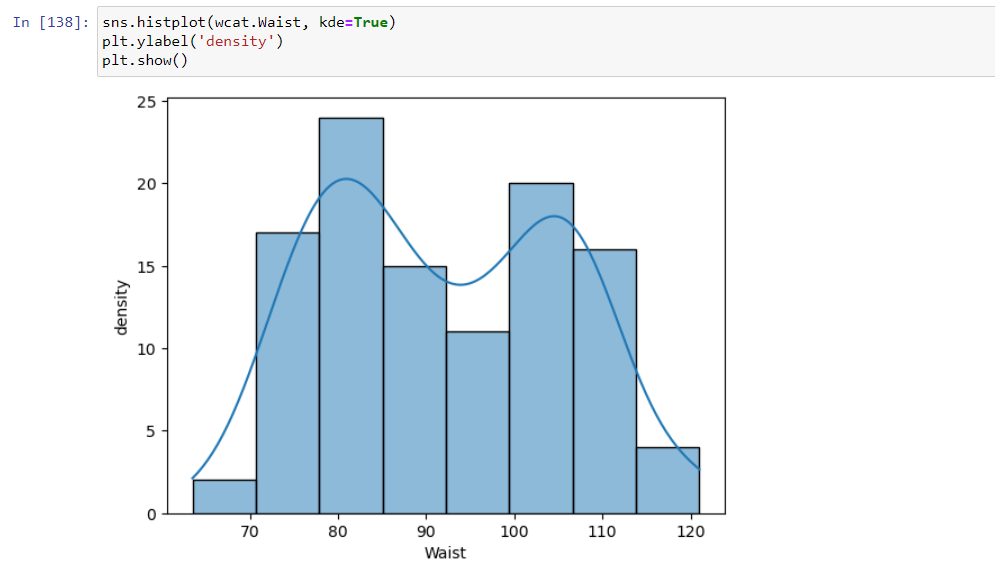
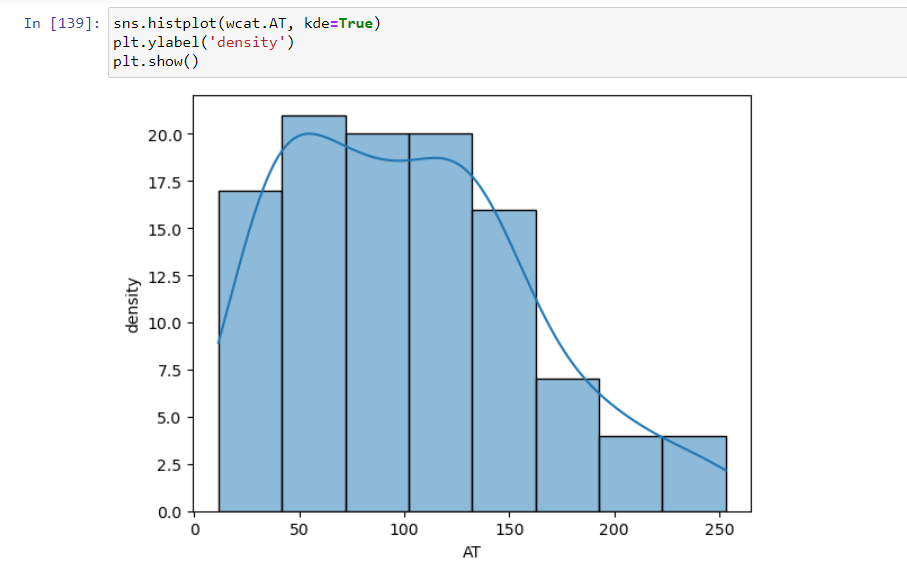
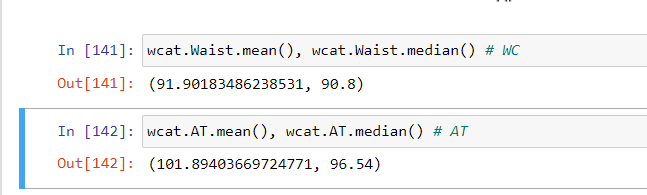
 

From above plot and values, we can say that data is fairly symmetrical, that fairly normally distributed.

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

Dataset: wc-at.csv

Answer:

From above plot and values, we can say that data is asymmetrical, that not normally distributed.

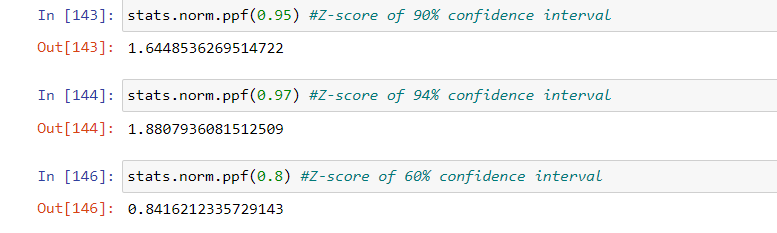
Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

For 90% confidence interval: We have the significance level at 5 % (as it is a two tailed test) that is: α = 5 % = 0.05 z at α = 0.05 from the z table will be: z = 1.645.

For 94 % confidence interval, we get: We have the significance level at 3 % (as it is a two tailed test) that is: α = 3 % = 0.03 z at α = 0.03 from the z table will be: z = 1.555.

For 60 % confidence interval, we get: We have the significance level at 20 % (as it is a two tailed test) that is: α =20 % = 0.2 z at α = 0.2 from the z table will be: z = 0.253

Therefore, we get that the z score at 90% confidence interval is 1.645, at 94% confidence interval is 1.555 and at 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

a) The sample size is n=25, So the degrees of freedom is n−1=25−1=24 Thus, we are interested in the quantity t(α/2)=t(0.05/2)=t(0.025) for a t-distribution with 24 degrees of freedom. Upon using a t-table, we see that the critical t-value for this 95% confidence interval is t(α/2)=2.064.

b)Upon using a t-table, we see that the critical t-value for this 96% confidence interval is t(α/2)=2.164. c)Upon using a t-table, we see that the critical t-value for this 99% confidence interval is t(α/2)=2.797

t-scores of 95% confidence interval for sample size of 25

(df = n-1 = 24)

2.0638985616280205

t-scores of 96% confidence interval for sample size of 25

2.1715446760080677

t-scores of 99% confidence interval for sample size of 25

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

Answer:

Ho = Avg life of Bulb >= 260 days

Ha = Avg life of Bulb < 260 days

find t-scores at x=260;

t= -0.4714045207910317

Find P(X>=260) for Ho

p\_value= 0.32167411684460556 \* 100

= 32.17%