**Data science assignment -1**

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

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

Solution: You can calculate it, but for such a small number of possible combinations of independent events (8), let’s look at them all.

H = heads

T = tails

Possible events with equal probability (order matters):

HHH

HHT

HTH

HTT

THH

THT

TTH

TTT

Number with 2 heads: 3

Total number: 8

From the definition of probability, the number you are looking for is 3/8 = 0.375 = 37.5%

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 2and 3

Solution: With two dice, there are (6 ) \* ( 6 ) = ( 36 ) possible combinations of numbers.

The minimum sum possible for the two dice thrown is (1, 1) = a sum of (2 )

The maximum sum possible for the two dice thrown is (6, 6) = a sum of (12).

**Sum = (1).**

The minimum possible sum is (1, 1) = (2 ).

Therefore P( 1 ) = ( 0 )/( 36 ) = 0

**Less than or equal to 4**

The possible outcomes are:{(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)}

Therefore P(<=4) = (6)/(36) = 1/6 = 0.16666

**Sum is divisible by 2 and 3**

The possible outcomes are:{(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)}

Therefore P = (24)/(36) = 2/3 = 0.6666

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: None of the balls drawn is blue ,This can only happen when the two balls drawn at random are either red and green or both

Total number of balls = 2+3+2 = 7

=>Number of ways of drawing 2 balls out of 7 = 7C2 = (7\*6)/(2\*1) = 42/2 = 21

Number of balls other than blue = 5

=>Number of ways of drawing 2 balls out of 5 = (5\*4)/(2\*1) = 20/2 = 10

Therefore Required Probability = 10/21= 0.47619.

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

Solu:Expected number of candies for a randomly selected child

= 1 \* 0.015 + 4\*0.20 + 3 \*0.65 + 5\*0.005 + 6 \*0.01 + 2 \* 0.12

= 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**

**Solution:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weight |
| Mean | 3.596563 | 3.21725 | 17.84875 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.92 | 3.44 | 17.02 |
| Variance | 0.285881 | 0.957379 | 3.193166 |
| Standard Deviation | 0.534679 | 0.9784574 | 1.786943 |
| Maximum | 4.93 | 5.424 | 22.9 |
| Minimum | 2.76 | 1.513 | 14.5 |
| Range | 2.17 | 3.911 | 8.4 |

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:**

Expected Value  =  ∑ ( probability  \* Value )

 ∑ P(x).E(x)

there are 9 patients

Probability of selecting each patient = 1/9

Ex  108, 110, 123, 134, 135, 145.,167,187, 199

P(x)  1/9  1/9   1/9  1/9   1/9   1/9   1/9   1/9  1/9

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

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

= (1/9)  (  1308)

= 145.33

Expected Value of the Weight of that patient = 145.33

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

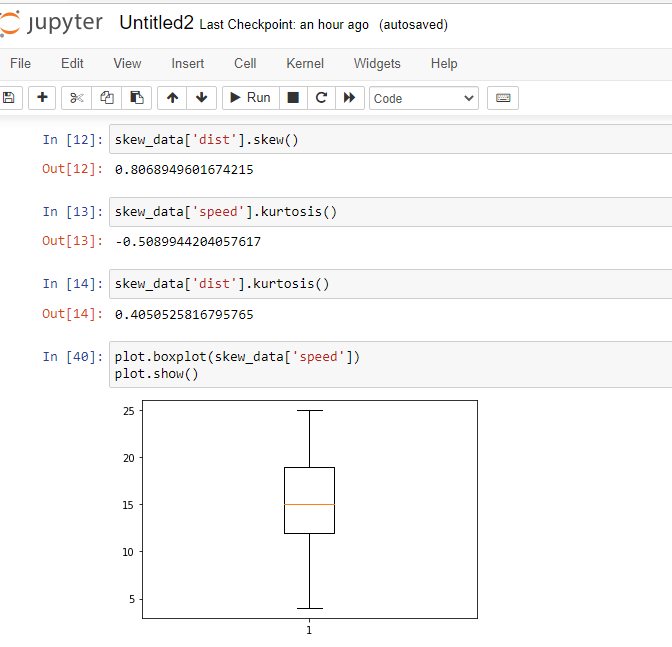
**Cars speed and distance**

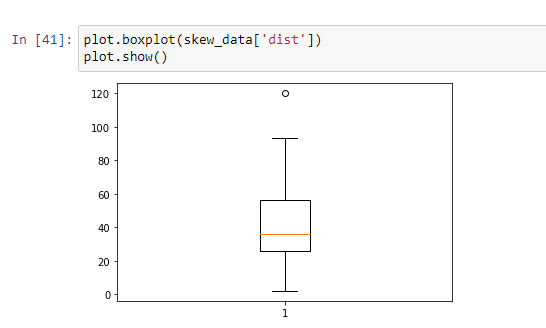
**Use Q9\_a.csv**

**SP and Weight(WT)**

**Use Q9\_b.csv**

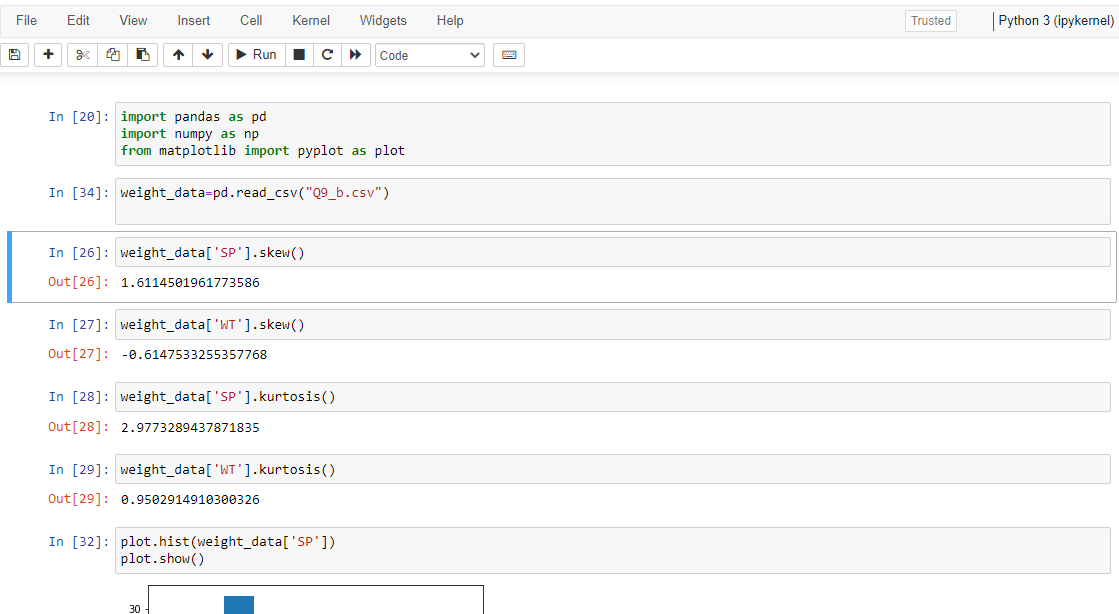
Solution:

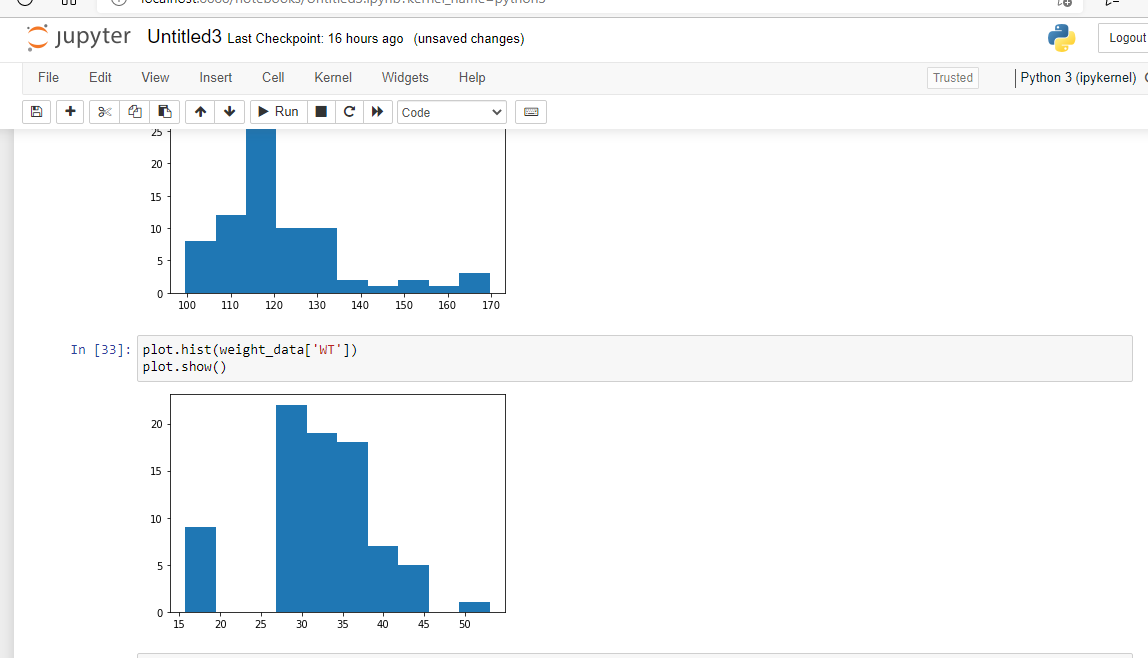




**SP and Weight(WT)**

**Use Q9\_b.csv**





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





**Solution:**

The given data is Right skew data or positive skewed where mean>median.

**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:**

Using the **t-distribution**, it is found that:

* The **94%** confidence interval is **(198.73, 201.27).**
* The **96%** confidence interval is **(198.61, 201.39).**
* The **98%** confidence interval is **(198.43, 201.57).**

The **information given** is:

* Sample **mean**of https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20%3D%20200.
* Sample **standard deviation** of https://tex.z-dn.net/?f=s%20%3D%2030.
* Sample **size**of https://tex.z-dn.net/?f=n%20%3D%202000.

The **interval** is:

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20%5Cpm%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D

* In which **t** is the critical value for the two-tailed confidence interval.

Considering a **94%** confidence level, using a calculator, with 200 - 1 = **199 df**, the critical value is **t = 1.8916**, hence:

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20-%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20-%201.8916%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20198.73

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20%2B%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20%2B%201.8916%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20201.27

The **94%** confidence interval is **(198.73, 201.27).**

Considering a **96%** confidence level, using a calculator, with 200 - 1 = **199 df**, the critical value is **t = 2.0673**, hence:

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20-%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20-%202.0673%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20198.61

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20%2B%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20%2B%202.0673%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20201.39

The **96%** confidence interval is **(198.61, 201.39).**

Considering a **98%** confidence level, using a calculator, with 200 - 1 = **199 df**, the critical value is **t = 2.3452**, hence:

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20-%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20-%202.3452%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20198.43

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20%2B%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20%2B%202.3452%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20201.57

The **98%** confidence interval is **(198.43, 201.57).**

|  |  |  |
| --- | --- | --- |
| **Confidence interval** | **Z-value** | **Range** |
| Confidence interval 94% | 1.880794 | 198.74,201.26 |
| Confidence interval 96% | 2.053749 | 198.62,201.38 |
| Confidence interval 98% | 2.326348 | 198.43,201.56 |

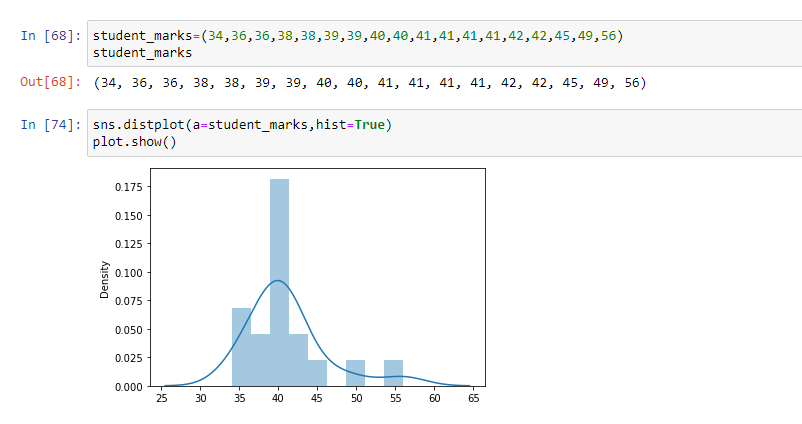
**Q12)** Below are the scores obtained by a student in tests.

**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:

|  |  |
| --- | --- |
| Mean | 41 |
| Median | 40.5 |
| Mode | 41 |
| Variance | 25.52941176 |
| Standard Deviation | 5.052663829 |
| Skew | 1.686841192 |
| Kurtosis | 3.953278897 |



2) What can we say about the student marks?

Solution: Mass of the Student marks between 38-42 .

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

Solution: Data is normalized and there is no skewness.

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

Solution: Negative Skewness implies mass of the Distribution concentrated on right side.

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

Solution: Positive skewness implies mass of the Distribution concentrated on the left side.

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

Solution: Positive Kurtosis value indicate that thinner peak and wider tails.

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

Solution: Negative kurtosis value indicate that wider peak and thinner tails.

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



What can we say about the distribution of the data?

Solution: Not Normally Distributed.

What is nature of skewness of the data?

Solution: Negative skewness.

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

Solution: 10-18.

Q19) Comment on the below Boxplot visualizations?



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

Solution: from the given boxplot 1,boxplot 2 the data is normally distributed where

Mean is equal to median .when the data is normally distributed then skewness and kurtosis are nearly equal to zero.

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:

P(MPG>38)

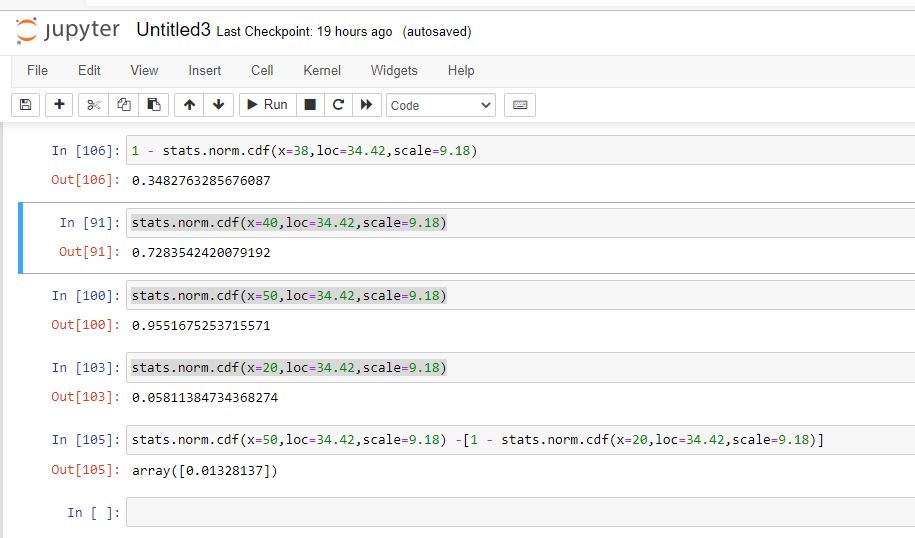
=0.348276.

P(MPG<40)

=0.728354.

P (20<MPG<50)

=0.01328.

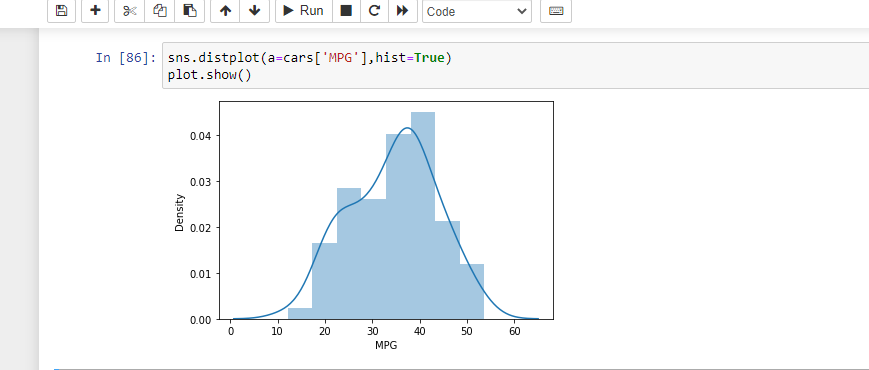


Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

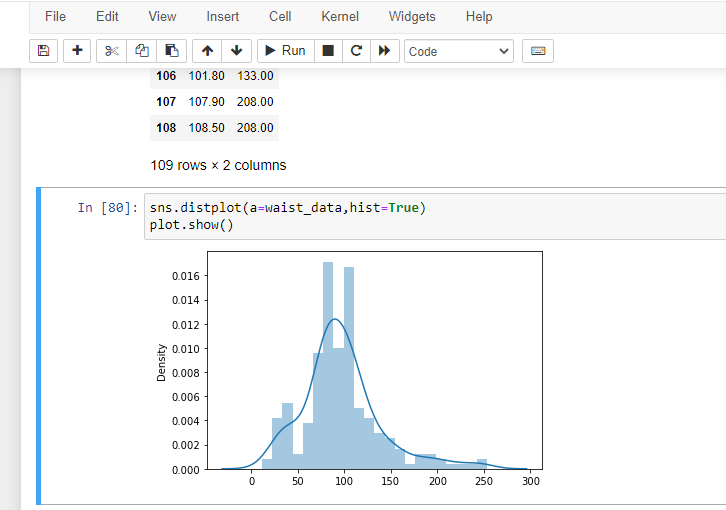
Solution: The data follows Normal Distribution.



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

Dataset: wc-at.csv

**Solution**: The Data set follows Normal Distribution



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

Solution:

|  |  |
| --- | --- |
| Confidence interval | Z Score |
| 60% | 0.8416212 |
| 90% | 1.644854 |
| 94% | 1.880794 |

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

Solution:

|  |  |
| --- | --- |
| Confidence interval | T Score |
| 95% | 2.063899 |
| 96% | 2.171545 |
| 99% | 2.79694 |

Q 24**)**A Government companyclaims 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: **Given:**

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.

**To find:**

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

**Solution:**

t - statistics for the data is given as follows:

https://tex.z-dn.net/?f=t%3D%5Cdfrac%7Bx-%5Cmu%7D%7B%5Cfrac%7Bs%7D%7B%5Csqrt%20n%7D%7D

x = mean of the sample of bulbs = 260

μ = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18

https://tex.z-dn.net/?f=t%3D%5Cdfrac%7B260-270%7D%7B%5Cfrac%7B90%7D%7B%5Csqrt%2018%7D%7D

https://tex.z-dn.net/?f=t%20%3D%20%5Cdfrac%7B-10%7D%7B%5Cfrac%7B90%7D%7B3%5Csqrt%202%7D%7D

https://tex.z-dn.net/?f=t%20%3D%20%5Cdfrac%7B-10%7D%7B%5Cfrac%7B30%7D%7B%5Csqrt%202%7D%7D

https://tex.z-dn.net/?f=t%20%3D%20%5Cdfrac%7B-1%20%5Ctimes%20%5Csqrt%202%7D%7B3%7D

t = - 0.471

For probability calculations, the number of degrees of freedom is n - 1, so here you need the t-distribution with 17 degrees of freedom.

The probability that **t < - 0.471 with 17 degrees of freedom** assuming the population mean is true, the t-value is less than the t-value obtained With 17 degrees of freedom and a t score of - 0.471, the probability of the bulbs lasting less than 260 days on average of **0.3218** assuming the mean life of the bulbs is 300 days.

