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

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** |
| 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 | **Ratio** |
| Blood Group | **Nominal** |
| Time Of Day | **Interval** |
| 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?

**Solution** :Total number of elements in sample space = 23

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

n(S)=8

A={HHT, HTH, HHT}

n(A)=3

**Ans : Probability that two heads and one tails = 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

Solution :

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

Total number in sample space= 62 = 36

a)Event A be the sum equal to 1

**Ans a) Probability of the sum equal to 1 = 0**

b) Let event B be the set of sum less than or equal to 4

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

**Ans b) Probability of sum is Less than or equal to 4 = 6/36 =1/6**

c) Let event C be the set of Sum is divisible by 2 and 3

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

**Ans c) probability of Sum is divisible by 2 and 3 = 5/36**

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: n(sample space)= 7C2(where n is the number of items and r is the total number items to be selected.) =7\*6\*5! /2! \*(5!) = 42/2 =21

Possibilities with 2red and 3 green balls = 5= 5C2 =20/2=10

**Ans: Probability that none of the drawn balls are blue = 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 | xipi |
| A | 1 | 0.015 | 0.015 |
| B | 4 | 0.20 | 0.8 |
| C | 3 | 0.65 | 1.95 |
| D | 5 | 0.005 | 0.025 |
| E | 6 | 0.01 | 0.06 |
| F | 2 | 0.120 | 0.24 |

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

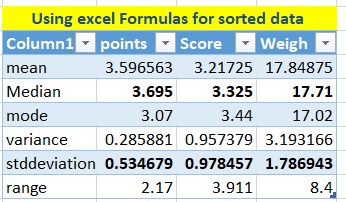
Expected number of candies = summation of xipi = 3.09

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset –For Points,Score,Weight>

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

**Use Q7.csv file**

**Ans** [**solved file**](file:///C:\Users\HP\Desktop\DATASCIENCE\ASSIGNMENTS\basic%20statistic%20level1\Q7solved.csv.xlsx)

****

**Ans :Comments: The mean ,median and mode have approximately equal values.**

**Inferences:1) since the mean,median and mode for point,score and weight all three variables are approximately equal hence they show symmetrical frequency curve symmetrical distribution of data.**

**2)There are no outliers in this data.**

**3)Since the standard deviation value is low hence the spread is small.All the values are near the mean value.**

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?

|  |  |
| --- | --- |
|  | 108 |
|  | 110 |
|  | 123 |
|  | 134 |
|  | 135 |
|  | 145 |
|  | 167 |
|  | 187 |
|  | 199 |
| Sum | 1308 |
| Count | 9 |
| expected value | 145.3333 |

**Expected Value = summation of all the values / number of values =1308/9 =145.33**

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

**Cars speed and distance**

**Use Q9\_a.csv**

[**Q9-solved**](file:///C:\Users\HP\Desktop\DATASCIENCE\ASSIGNMENTS\basic%20statistic%20level1\Q9_aSolved.csv)

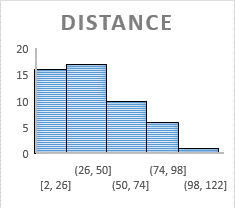
**Solution**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Dist* |  |  | *speed* |  |
|  |  |  |  |  |
| Mean | 42.98 |  | Mean | 15.4 |
| Standard Error | 3.644340314 |  | Standard Error | 0.747785847 |
| Median | 36 |  | Median | 15 |
| Mode | 26 |  | Mode | 20 |
| Standard Deviation | 25.76937749 |  | Standard Deviation | 5.287644435 |
| Sample Variance | 664.0608163 |  | Sample Variance | 27.95918367 |
| Kurtosis | 0.405052582 |  | Kurtosis | -0.50899442 |
| Skewness | 0.80689496 |  | Skewness | -0.117509861 |

**Inference for distance 3rd and 4 th moments ;-**

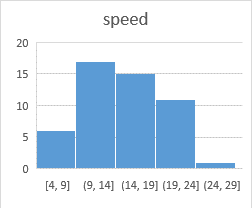
**Since the skewness is positive number hence this is right skewed with a value of skewness = 0.80689**

**Kurtosis is the measure of tailedness it also is a positive value = 0.4050 show the graph will be show right tailedness with a value = 0.4050**



**Inference for speed 3rd and 4 th moments ;-**

**Since the skewness and kurtosis values are –ve number hence this is slightly left skewed with a slight left tailedness .but since the values are -0.5 and -0.11 we can consider this to be a normally distributed data following the bell curve.**



**Histograms for speed**

**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |  |  |
| --- | --- | --- | --- |
| *SP* | | *WT* | |
|  |  |  |  |
| Kurtosis | 2.977329 | Kurtosis | 0.950291 |
| Skewness | 1.61145 | Skewness | -0.61475 |

|  |  |
| --- | --- |
|  |  |

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



**Ans : In the above data range is highest –lowest = 400 -0 =400.**

**Definition skewness** is a measure of lack of symmentry of the [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution) of a [real](https://en.wikipedia.org/wiki/Real_number)-valued [random variable](https://en.wikipedia.org/wiki/Random_variable) about its mean. The right tail is longer; the mass of the distribution is concentrated on the left of the figure. The distribution is said to be *right-skewed*, *right-tailed*, or *skewed to the right*, despite the fact that the curve itself appears to be skewed or leaning to the left; *right* instead refers to the right tail being drawn out and, often, the mean being skewed to the right of a typical center of the data.

**This is assymetrical distribution of data with a positive skewness hence in this case mean >median>mode**



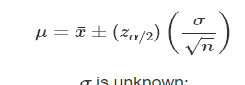
**This box plot identifies the outliers present in the data set .**

**Since the median is on the left side to the center of the box hence this is positively skewed. The right whisker - line above is larger than the lower whisker line hence the data is having more variance is very high and data shows the spread of data on the right of the Q3 because more data lies in 25% on the right than that in the left 25% .**

**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?

Za/2

**Solution: Population count N=3000000 ,sample count S= 2000,average weight in sample = 200 pounds ,std deviation = 30 pounds. for 94 % interval, significance level = 6 % =.6 Z(0.06/2) = Z0.03 = 1.88,**

**Formula **

**To calculate interval for 94% interval**

Explaination : *A 94 % confidence interval*

*has two tails of 6/2 = 3 %*

*so it goes from 3% to 97 % which leaves 94 % in the middle*

*so look up the Z for*

*P(z<Z) = 0.97*

*two closest values in the z-table*

*P(z<1.88 ) = 0.96995*

*P(z< 1.89) = 0.97062*

*interpolating*

*1.88 + ( 0.97- 0.96995)\*(0.01)/ (0.97062- 0.96995)*

*Z\_critical = approx. 1.880746*

**lower level for class interval = 200 – (1.88) \*(30/sqroot(2000))=** 198.738325292158

**higher level for class interval = 200 + (1.88) \*(30/sqroot(2000))=** 201.261674707842

94% confidence interval with sqrt(n)= ( 198.738325292158 , 201.261674707842 )

**Z(0.01)=2.22**

**to calculate for 98 % interval Z (Z value of significance level) = 2.326**

**lower level for class interval = 200 – 2.22\*(30/sqroot(2000))=** 198.43943840429978

**HIGHER level for class interval = 200 + 2.22\*(30/sqroot(2000))=** 201.56056159570022

98% confidence interval with sqrt(n)= ( 198.43943840429978 , 201.56056159570022 )

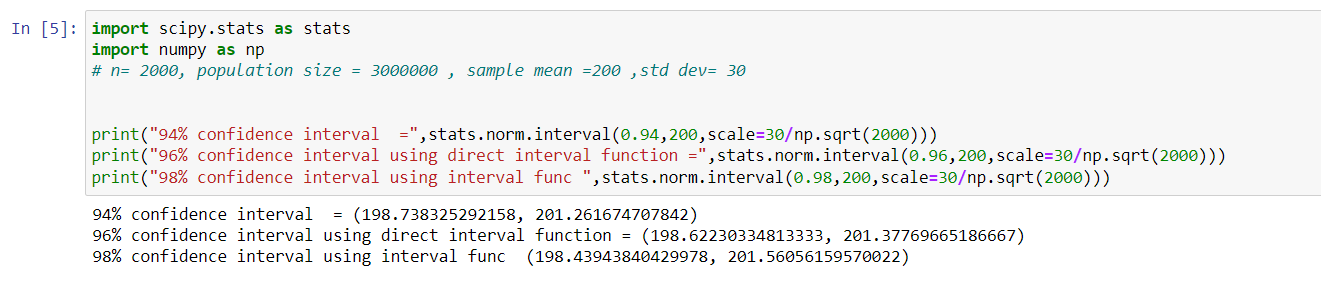
**Z(0.02)=2.05**

**To calculate for 96% intevel.**

**lower level for class interval = 200 – 2.05\*(30/sqroot(2000))=** 198.62230334813333

**higher level for class interval = 200 + 2.05 \*(30/sqroot(2000))=** 201.37769665186667

96% confidence interval with sqt(n) = ( 198.62230334813333 , 201.37769665186667 )



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

Median = (40+41)/2 =40.5

Variance=sum((xi-mean)2)/(17)=(49+25+25+9+9+4+4+1+1+0+0+0+0+1+1+16+64+225)/17 =25.52

Stddeviation =square root of variance =5.0526

**Ans1) Mean=41,unimodal Mode =41,Median= 40.5,std deviation=5.0526 ,variance = 25.52**

**Ans2)Mean ,Median and Mode have same values hence the data is distributed symmentrically and there are no outliers.Hence this represents a Gaussian curve.**

**Hence 68 % people fall in the one std deviation which is (41-5.0526 ) = 35.9474 to (41 +5.0526)=46.0526**

**95% people fall in the second std deviation which is (41-2\*5.0526) =30.8948 to (41+2\*5.0526) = 51.10526**

**99.5% people fall is the third std deviation which is ( 41-3\*5.0526)= 25.8422 to (41+3\*5.0526)=56.1578**

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

**Ans : when the measure or mean and median are equal then the distribution is symmentric and the distribution has zero skewness as skewness is the measure of lack of symmentry.This data will have a** **normal frequency distribution graph which will have a symmetrical frequency curve.**

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

**Ans : Skewness is a measure of the asymmetry of a distribution.when Mean is greater than the Median ,the graph has positive skewness.**

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

**Ans : Skewness is a measure of the asymmetry of a distribution.when Mean is less than the Median ,the graph has negative skewness.**

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

**Kurtosis is the measure of tailedness.the positive value of kurtosis means it has a right tailed graph.**

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

**Kurtosis is the measure of tailedness.The negative value of kurtosis means it has a left tailed graph.**

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



What can we say about the distribution of the data?

Ans : **50% of the data lies between 10 to 18 values .since Q3 = 18,Q1=10.**

IQR Interquartile range has 50 **% of the data distributed.It is not normally distributed data.**

What is nature of skewness of the data?

**Ans:since the median is on the right side of the box hence , Negative skewness of data is shown in the above graph. hence here the mode > median >**

**Mean .median value is approximately =15.5**

What will be the IQR of the data (approximately)?   
Ans :Q1 = 10 and Q3= 18 (approximately) ,inter quartile range = 18-10 =8

Q2 median = 15.5 (approximately)

Q19) Comment on the below Boxplot visualizations?



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

**Ans Inference Both the box plot have the same median value. Around 256 (approx.)**

**Both the box plot show a normal distribution with median passing from the center of the IQR**

**Details for 1);- The concentration of data in IQR for the 1)st boxplot is between 245 to 280(approx.),Q1=253 ,Q4 =280 (approx.)**

**Details for 2)The concentration of data in IQR for the 2)st boxplot is between 225 to 320(approx.) ,Q1=224,Q4=320 (approx.)**

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)

Solution

Manual calculation

|  |  |
| --- | --- |
| *MPG* | |
|  |  |
| Mean | 34.42207573 |
| Standard Error | 1.01460497 | Standard Deviation=9.14 |
| Median | 35.15272697 | MEAN=34.42 |

**P(MPG < = 38)**

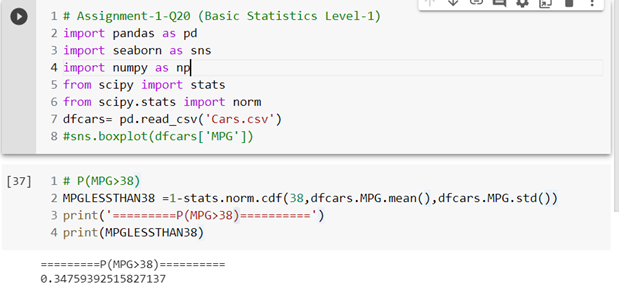
**Z value for x is 38= (38 - 34.42)/9.14 =0.391684902**

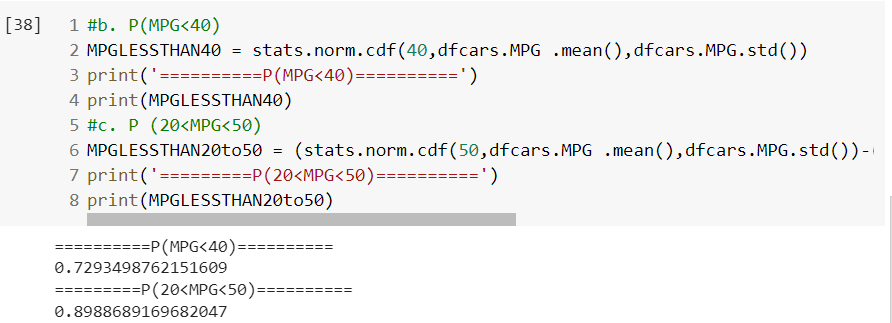
**Using z score table area under (p > 0.39) = 0.6517 =65%**

**Hence P(MPG>38) = 1- area P(MPG < = 38)**

**Hence the area under this region P(MPG>38) = 1 – 0.6517 =0.3483 =34% (=35% approx.)**

**Using python code ;All are solved**

****

****

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

|  |  |
| --- | --- |
| USING EXCEL | USING PYTHON CODE |
|  |  |
| |  |  | | --- | --- | | Mean | 34.42207573 | | Median | 35.15272697 | |  |

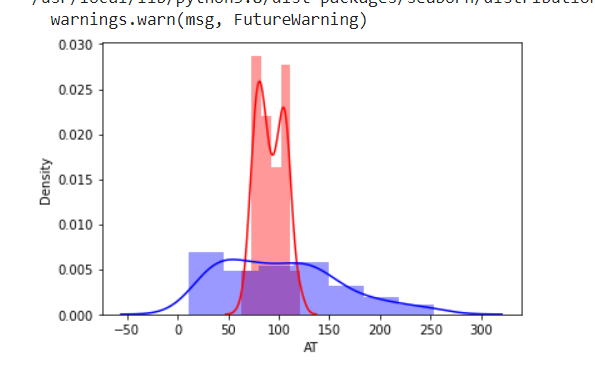
Inference ;- SINCE MEAN AND MEDIAN are approximately equal hence distribution can be said as normal distribution.

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

Dataset: wc-at.csv

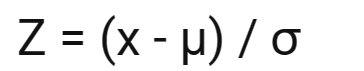
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | *Waist* |  | |  |  | | Mean | 91.90183 | | Standard Error | 1.298728 | | Median | 90.8 | | Mode | 94.5 | | Standard Deviation | 13.55912 | | Sample Variance | 183.8496 | | Kurtosis | -1.10267 | | Skewness | 0.134056 | | Range | 57.5 | | Minimum | 63.5 | | Maximum | 121 | | Sum | 10017.3 | | Count | 109 | | Confidence Level(95.0%) | 2.574304 | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | *AT* |  | |  |  | | Mean | 80.49753086 | | Median | 75.08 | | Mode | 123 | | Standard Deviation | 43.4739481 | | Sample Variance | 1889.984164 | | Kurtosis | 0.096425478 | | Skewness | 0.672653804 | | Range | 205.56 | | Minimum | 11.44 | | Maximum | 217 | | Sum | 6520.3 | | Count | 81 | | Confidence Level(95.0%) | 9.612879322 | |



INFERENCE;- No they do not follow normal distribution

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

Manual calculation ;

Explaination : *A 90 % confidence interval*

*has two tails of 10/2 =5 %*

*so it goes from 5% to 95 % which leaves 90 % in the middle*

*so look up the Z for*

*P(z<Z) = 0.95*

*two closest values in the z-table*

*P(z<1.65 ) =* 0.9505

*P(z< 1.64) =* 0.9495

*P(z<Z) = 0.05*

*two closest values in the z-table*

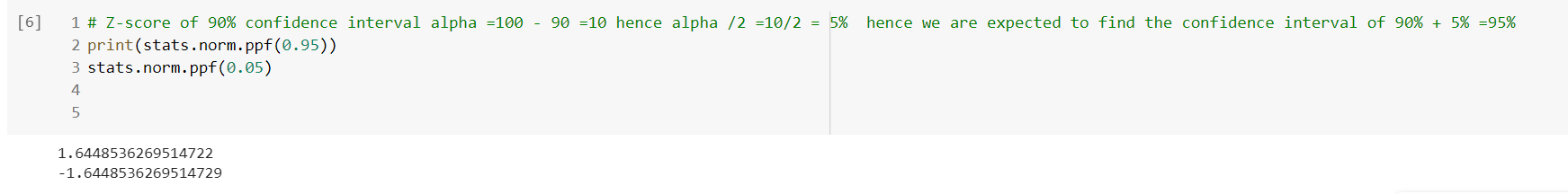
*P(z<1.64) =* 0.0505

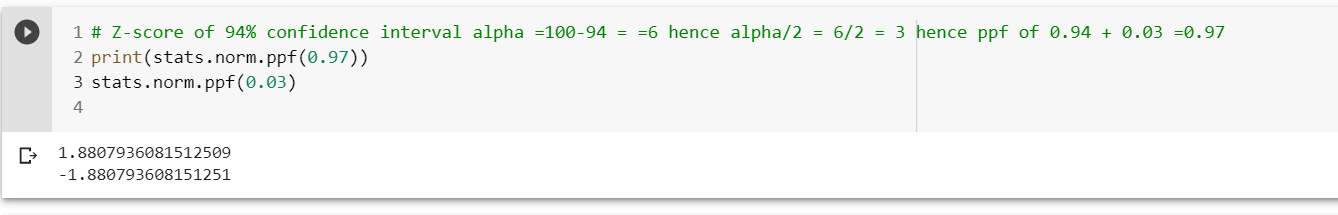
*P(z< 1.65) =* 0.0516

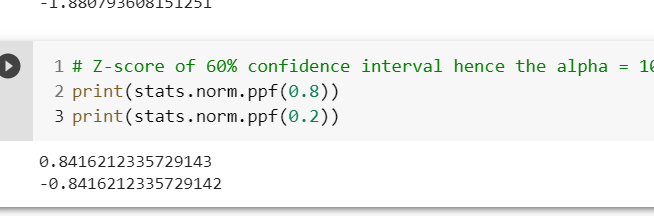
*Z\_scores for 90 % = approx. 1.64*

*Hence the z scores are (-1.64 and +1.64*

the Z scores for 90% confidence interval,94% confidence interval, 60% confidence interval using python code.

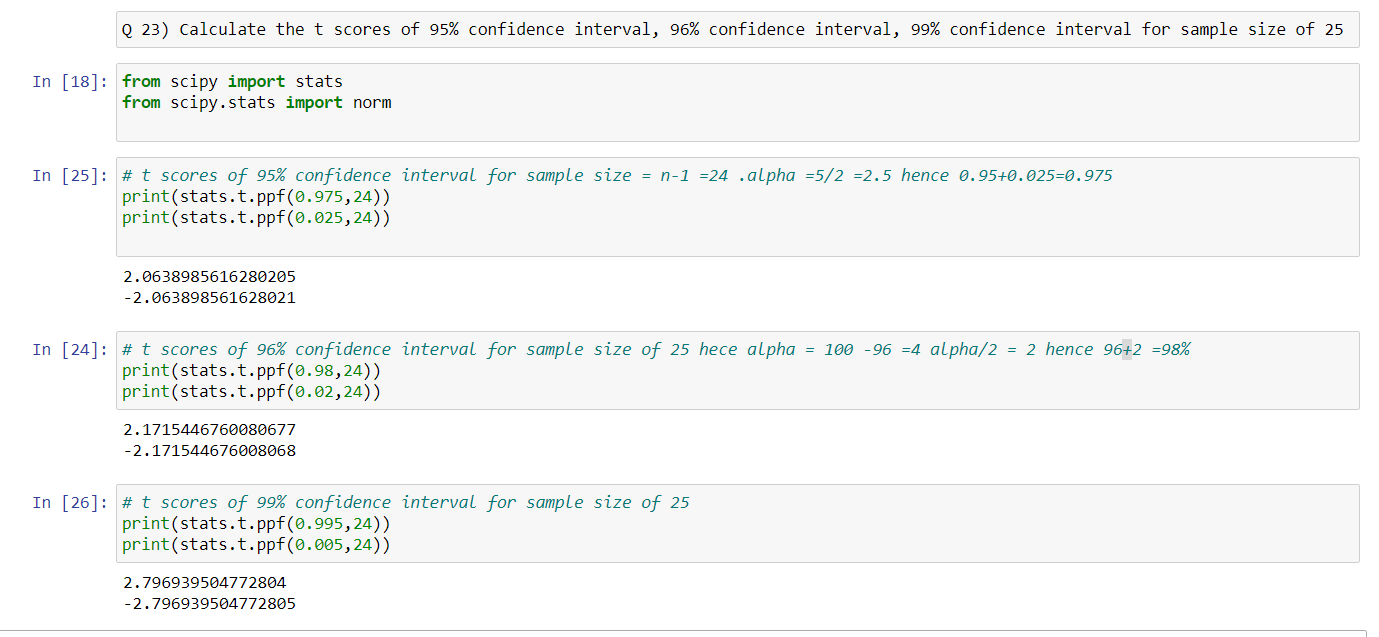






Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

Using code



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 =n-1

solution ; -

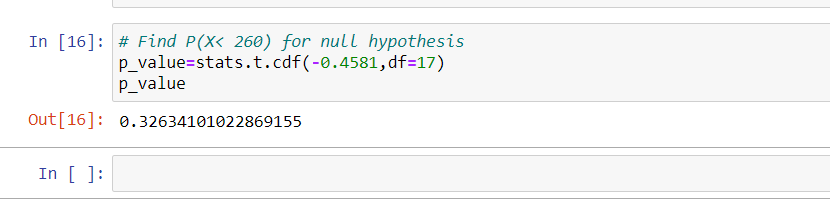
population mean =270, sample size n=18 ,sample average = 260 ,SD=90.

Since the sample size is less than 30 hence we will use the t test.

Manual calculation(t statistics ) t= (VAL-EV)/Sd

260-270/21.828 = -0.4581

Step 3 ) t .cdf(-0.4581, 17) = 0.6783



Hence the pvalue = 32.63%

Probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 32.63%

Python code:

