|  |  |
| --- | --- |
| 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 | Nominal |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Discrete |

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 | Nominal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ratio |
| Time on a Clock with Hands | Ratio |
| Number of Children | Nominal |
| 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?

The tossing of three coins at a time. These are the number of possible outcomes

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

S(A)={ THH , HHT, HTH }

P(A)= S(A)/S

= 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

**Ans** When two dice are rolled. These are the number of possible outcomes

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),(6,3),(4,4),(5,5),(5,6)

(6,1),(6,2),(6,3),(6,4),(6,5),(6,6) } =36

1. S(a) = Sum is equal to 1 = 0

P(a) = S(a)/S = 0/36

P(a) = 0

1. S(b) = Less than or equal to 4 = {(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)}

P(b) =S(b)/S = 6/36

P(b) =1/6

1. S(c) = Sum is divisible by 2 and 3 = {(1,5),(2,4),(3,3),(4,2),(5,1),(6,6)}

P(c) = S(c)/S = 6/36

P(c) = 1/36

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?

**Ans.** nCr =n!/r!\*(n-r)!

n = Total number of balls =7 balls

r= Two balls drawn at random

Sample Space = 7c2 = 7! / 2! \* 5! = 21

Event E be none of the balls is blue = all balls are either red or green or both.

n(E) = 5c2 = 5!/2!\*3! = 10

P(E) =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.** P(x)=Candies count\* Probability)

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

= 3.09

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

* For Points,Score,Weigh>

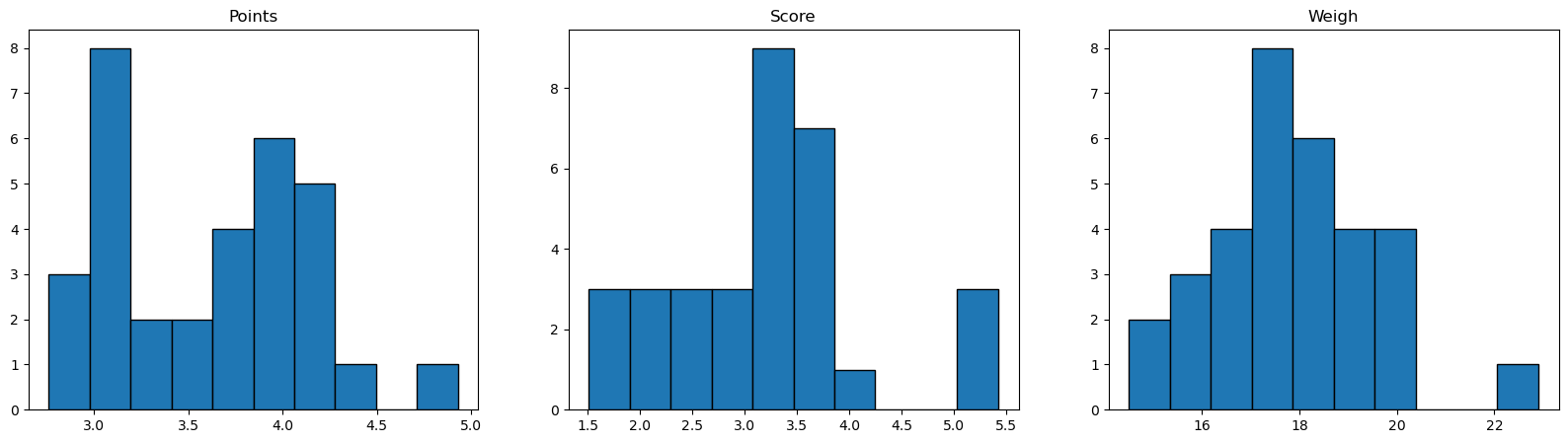
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

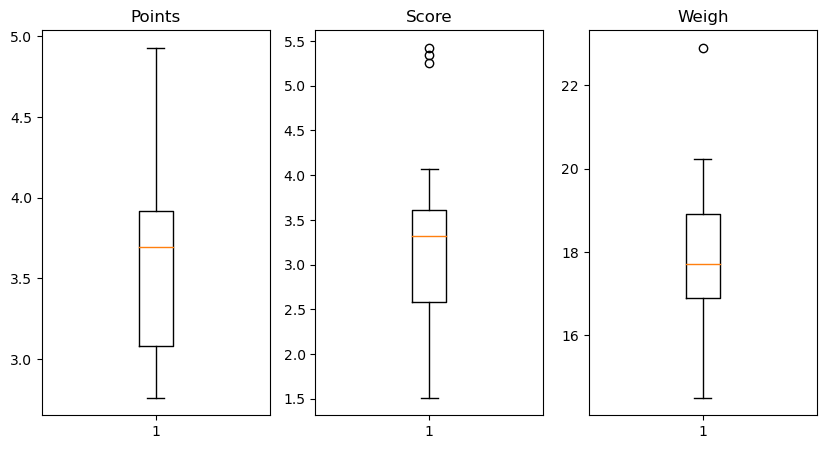
**Use Q7.csv file**

Variance =∑(x-µ)^2/n

Std dev.=√ ∑(x-µ)^2/n

Range= Max.-Min.





|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weigh** |
| **Mean** | 3.5965 | 3.2172 | 17.8487 |
| **Median** | 3.695 | 3.325 | 17.710 |
| **Mode** | 3.92 | 3.44 | 17.02 |
| **Variance** | 0.2858 | 0.9573 | 3.1931 |
| **Std dev** | 0.5346 | 0.9784 | 1.7869 |
| **Range** | 2.17 | 3.91 | 8.399 |

Q8) Calculate Expected Value for the problem below

1. The weights (X) of patients at a clinic (in pounds), are

|  |  |
| --- | --- |
| **E(x)** | **P(x)** |
| 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

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

**Ans.**

Expected Value=∑( probability  \* Value )

  = ∑ P(x).E(x)

Probability of selecting each patient = 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)

= 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**

**Ans.**

Skewness -

Speed= - 0.117510

Distance= 0.806895

Kurtosis -

Speed= - 0.508994

Distance= 0.405053

**SP and Weight(WT)**

**Use Q9\_b.csv**

Skewness -

SP = 1.611450

WT = - 0.614753

Kurtosis -

SP = 2.977329

WT = 0.950291

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



Ans. The histograms peak has right skew and tail is on right side.

Mean > Median or mode . We have outliers on the higher side.



**Ans**. Data is not normally distributed.The boxplot has outliers on the maximum side.

**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=X̅± t(1-α), (n-1) \*S/√n

**Average Weight of Adult in Mexico with 94% CI**

To calculate t score

Cl= 94%

α = 1-0.94 =0.06

[ t=stats.t.ppf(0.97,df=1999)

print(np.round(t,2)) ]

t = 1.88

n=2000

df=(n-1) =1999

mean=200

std dev.(S)=30

upper\_conf=mean+t\*(S/√n)

=200+1.88(30/√2000)

=201.26

lower\_conf=mean-t\*(S/√n)

=200-1.88(30/√2000)

=198.74

Confidence Interval =[198.74,201.26]

**Average Weight of Adult in Mexico with 98% CI**

To calculate t score

Cl= 98%

α = 1-0.98 =0.02

[ t=stats.t.ppf(0.99,df=1999)

print(np.round(t,2)) ]

t = 2.33

n=2000

df=(n-1) =1999

mean=200

std dev.(S)=30

upper\_conf=mean+t\*(S/√n)

=200+2.33(30/√2000)

= 201.56

lower\_conf=mean-t\*(S/√n)

=200-2.33(30/√2000)

=198.44

Confidence Interval =[198.44,201.56]

**Average Weight of Adult in Mexico with 96% CI**

To calculate t score

Cl= 96%

α = 1-0.96 =0.04

[ t=stats.t.ppf(0.98,df=1999)

print(np.round(t,2)) ]

t = 2.06

n=2000

df=(n-1) =1999

mean=200

std dev.(S)=30

upper\_conf=mean+t\*(S/√n)

=200+2.06(30/√2000)

= 201.38

lower\_conf=mean-t\*(S/√n)

=200-2.06(30/√2000)

= 198.62

Confidence Interval =[198.62 ,201.38]

**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.

**Ans.**

Mean(µ)= (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 =∑(x-µ)^2/n

= 24.11

Std dev.=√ ∑(x-µ)^2/n

=4.91

Mean=41 Median=40.5 Variance=24.11 Standard deviation=4.91

1. What can we say about the student marks?

**Ans.**

Data is Slightly Skew towards Right because Mean>Median

And Distribution Boxplot show two outliers i.e two student score high in Test

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

**Ans.** No skewness is present we have a perfect symmetrical distribution

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

**Ans.** Skewness and tail is towards Right i.e Positive Skewness

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

**Ans**. Skewness and tail is towards left i.e.Negative Skewness

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

**Ans.** Positive kurtosis means the curve is more peaked and it is Leptokurtic

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

**Ans**. Negative Kurtosis means the curve is more flatter and broader and it is Platykurtic

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



What can we say about the distribution of the data?

**Ans.**

Above Boxplot is not Normally Distributed

Median>Mean i.e data is skewed toward left.

What is nature of skewness of the data?

**Ans.**

The data is a skewed towards left. The whisker range on left side is greater

than right side

We can say it is Negative skewness or Left skewness

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

**Ans.**

IQR(Inter Quartile Range) = Q3(Upper Quartile)-Q1(Lower Quartile)

IQR =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.

**Ans.**

There are no outliers in both the Boxplot 1 and 2

In Boxplot 1 & 2 There wisker length is also same on both side

We can say Boxplot 1 and 2 both are Normally Distributed

Both the box plot having same median i.e. ~262.5

Boxplot 1 distributed in range from ~[237.5-287.5]

Boxplot 2 distributed in range from ~[200-325]

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)

from scipy import stats

1-stats.norm.cdf(38,loc=cars.MPG.mean(),scale=cars.MPG.std())

P(MPG>38) =0.3475

* 1. P(MPG<40)

stats.norm.cdf(40,loc=cars.MPG.mean(),scale=cars.MPG.std())

P(MPG<40) =0.7293

* 1. P (20<MPG<50)

1-stats.norm.cdf(20,cars.MPG.mean(),cars.MPG.std())

P(MPG>20)=0.9428

stats.norm.cdf(50,cars.MPG.mean(),cars.MPG.std())

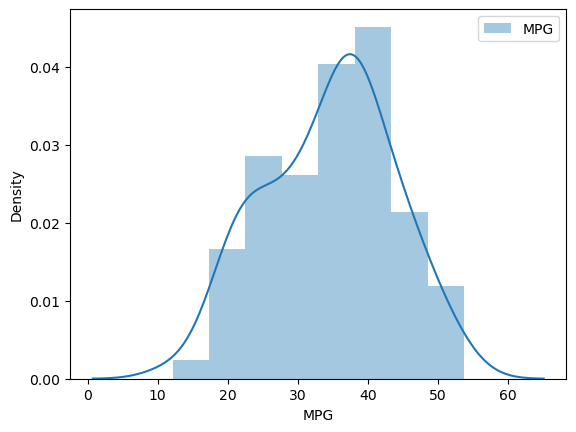
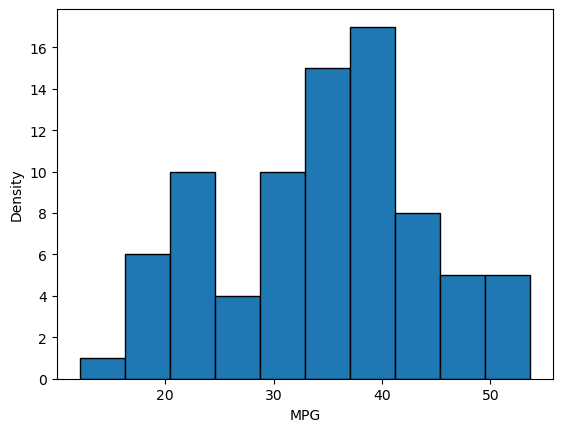
P(MPG<50)= 0.9559

P (20<MPG<50) = 0.9559-0.9428 =0.0131

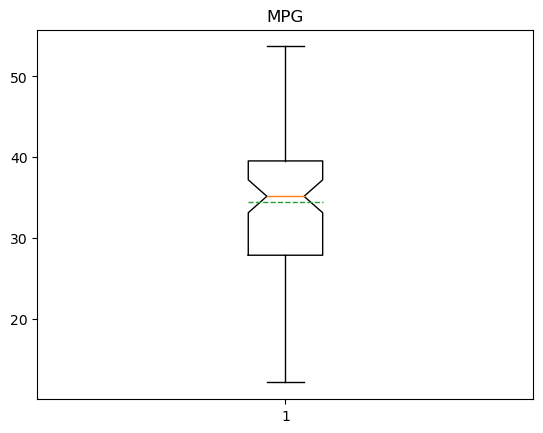
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv



|  |  |
| --- | --- |
|  | MPG |
| **count** | 81.000000 |
| **mean** | **34.422076** |
| **std** | 9.131445 |
| **min** | 12.101263 |
| **25%** | 27.856252 |
| **50%** | **35.152727** |
| **75%** | 39.531633 |
| **max** | 53.700681 |

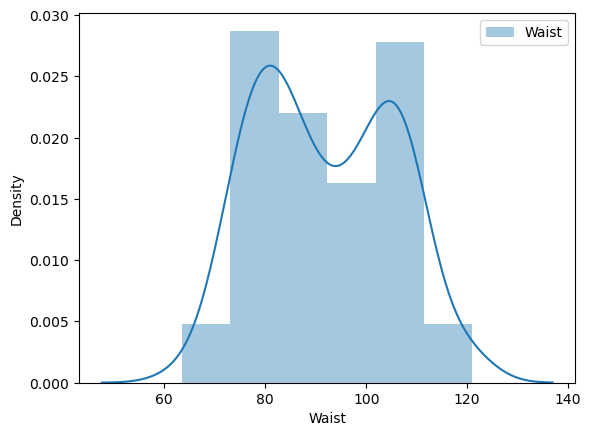
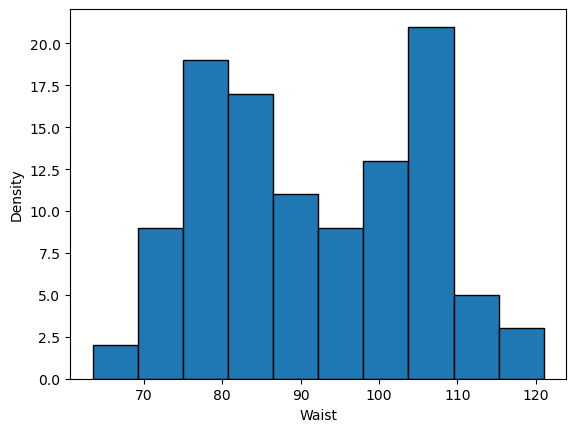


Mean<Median

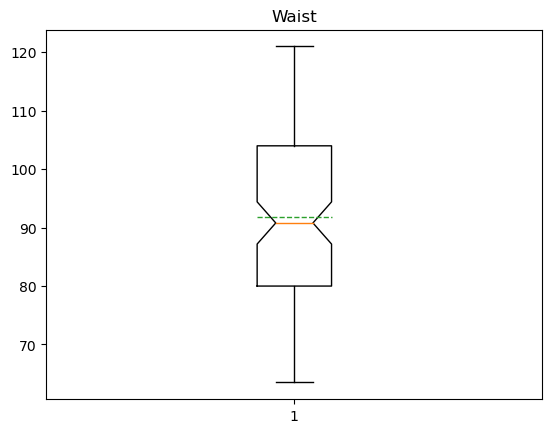
Slightly skewness towards the left i.e.Left Skewness

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

Dataset: wc-at.csv



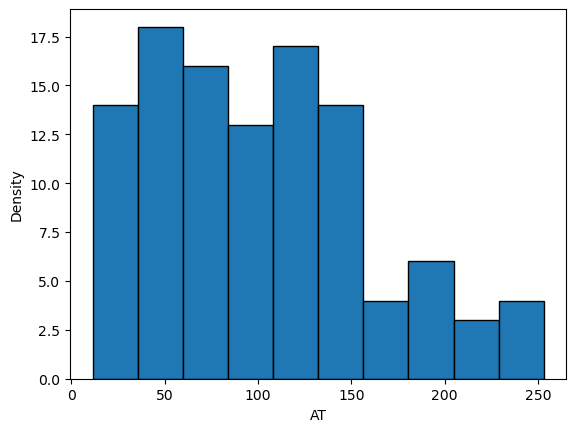
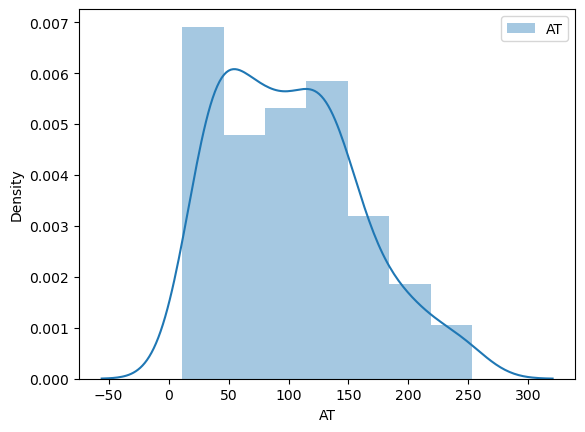
|  |  |
| --- | --- |
|  | Waist |
| **count** | 109.000000 |
| **mean** | **91.901835** |
| **std** | 13.559116 |
| **min** | 63.500000 |
| **25%** | 80.000000 |
| **50%** | **90.800000** |
| **75%** | 104.000000 |
| **max** | 121.000000 |



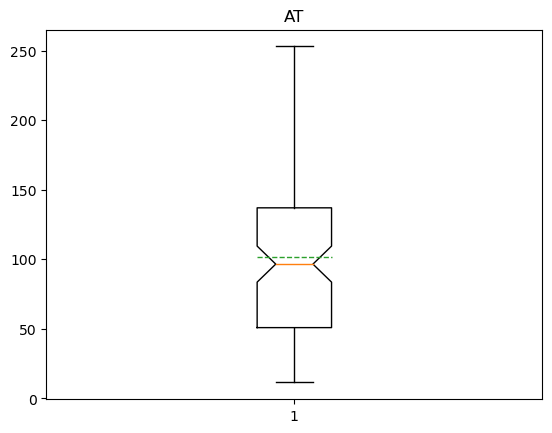
In Waist

Mean>Median

Slightly skewness towards the Right i.e.Right Skewness

|  |  |
| --- | --- |
|  | AT |
| **count** | 109.000000 |
| **mean** | **101.894037** |
| **std** | 57.294763 |
| **min** | 11.440000 |
| **25%** | 50.880000 |
| **50%** | **96.540000** |
| **75%** | 137.000000 |
| **max** | 253.000000 |



In AT

Mean>Median

In boxplot Right Wisker Length is higher than left wisker

Skewness towards the right i.e. Right skewness

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

**Ans.**

Cl=90% Cl=94% Cl=60%

α=(1-Cl) α=(1-Cl) α=(1-Cl)

= 1-0.90 = 0.10 =1-0.94 = 0.06 =1-0.60 =0.40

(1+Cl)/2 =0.95 (1+Cl)/2 =0.97 (1+Cl)/2=0.80

Z(0.95) =1.64 Z(0.97) = 1.88 Z(0.80)=0.84

OR

Cl=90%

z=stats.norm.ppf(0.95)

Cl=94%

z= stats.norm.ppf(0.97)

Cl=60%

z= stats.norm.ppf(0.80)

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

**Ans.**

df=25-1

=24 Cl=95% Cl=96% Cl=99%

α=(1-Cl) α=(1-Cl) α=(1-Cl)

= 1-0.95 = 0.05 =1-0.96 = 0.04 =1-0.99 =0.01

(1+Cl)/2 =0.975 (1+Cl)/2 =0.98 (1+Cl)/2=0.995

t(0.975)(24)=2.06 t(0.98)(24)=2.17 t(0.995)(24)=2.79

OR

Cl=95%

z=stats.t.ppf(0.975,df=24) = 2.06

Cl=96%

z= stats.t.ppf(0.98,df=24) = 2.171

Cl=99%

z= stats.t.ppf(0.995,df=24) = 2.7969

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

**Ans**.

Find P(X>=260)

Sample mean=260 days

Population mean=270 days

Sample Std.dev.=90 days

Sample Size(n)=18

df(degree of freedom)=18-1=17

t score = (sample mean-Pop.Mean) / (Std.dev / √Sample Size)

= (260-270)/(90/√18)

= -0.471

t score = -0.471

For P-value

from scipy import stats

p\_value**=**stats**.**t**.**cdf(t score,df**=**17)

=0.321

=32%