

Q1) Identify the Data type for the Following:

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

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	Ordinal
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?

Ans:

By Probability Method

Total Number of events

= Number of Possibilities per Experiments ^{Number of Experiments} = $2^3 = 8$

Total Number of Interested Events

$$= {}^nC_r = {}^3C_2 = \frac{3!}{(3-2)! * 2!} = 3$$

Probability of Interested Events

$$P(X=2H) = \frac{\text{Total Number of events}}{\text{Total Number of Interested Events}} = \frac{3}{8} = 0.375 = 3.75 \%$$

By Probability Mass Function formula

$$p.m.f. = {}^nC_x * P^x * (1-P)^{n-x}$$

where, n = Number of trails

x = Number of success required

p = probability of getting success in one trail

$$P(X=2H) = {}^3C_2 * (0.5)^2 * (1-0.5)^{3-2}$$

$$P(X=2H) = \frac{3!}{(3-2)! * 2!} * 0.25 * 0.5$$

$$P(X=2H) = 3 * 1.25$$

$$P(X=2H) = 3.75\%$$

➤ By Python

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

```
In [1]: 1 from scipy import stats as st
```

probability of getting 2 head

st.binom.pmf(k, n, p, loc)

- k = Number of success required
- n = Number of trials
- p = probability of getting success in one trial

```
In [2]: 1 st.binom.pmf(2,3,0.5)
        2
```

```
Out[2]: 0.375
```

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

a) Equal to 1

When we roll two dies,

Total Number of events

= Number of Possibilities per Experiments $\text{Number of Experiments} = 6^2 = 36$

We will get minimum sum of 2, Therefore

Total Number of Interested Events = 0

Probability of Interested Events

$$P(\text{sum}=1) = \frac{\text{Total Number of events}}{\text{Total Number of Interested Events}} = \frac{0}{36} = 0 = 0 \%$$

b) Less than or equal to 4

When we roll two dies,

Total Number of events

$$= \text{Number of Possibilities per Experiments} \text{ Number of Experiments} = 6^2 = 36$$

Total Number of Interested Events

= Number below that gives sum of 4 or less

$$= [(3,1), (2,2), (1,3)] = 3$$

Probability of Interested Events

$$P(\text{sum} \leq 4) = \frac{\text{Total Number of events}}{\text{Total Number of Interested Events}} = \frac{3}{36} = 0.0833 = 8.33 \%$$

c) Sum is divisible by 2 and 3

When we roll two dies,

Total Number of events

$$= \text{Number of Possibilities per Experiments} \text{ Number of Experiments} = 6^2 = 36$$

Number below that divisible by 2&3 both = [6,12]

Number of Combinations that gives sum 6

$$= [(5,1), (4,2), (3,3), (2,4), (1,5)] = 5$$

Number of Combinations that gives sum 12

$$= (6,6) = 1$$

$$\text{Total Number of Interested Events} = 5 + 1 = 6$$

Probability of Interested Events

$$= \frac{\text{Total Number of events}}{\text{Total Number of Interested Events}} = \frac{6}{36} = 0.166 = 1.66 \%$$

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?

Total Number of events

$$= {}^7C_2 = {}^7C_2 = \frac{7*6}{2} = 21$$

Total Number of Interested Events

$$= {}^nC_r = {}^5C_2 = \frac{5*4}{2} = 10$$

Probability of Interested Events

$$P(\text{none of the ball is blue}) = \frac{\text{Total Number of events}}{\text{Total Number of Interested Events}} = \frac{10}{21} = 0.476 = 47.6\%$$

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:

$$\text{Expected Random value} = \sum p(X_i) * X_i$$

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

$$= 3.09$$

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 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

	FORMULA(STEPS)	POINTS	SCORE	WEIGHT	INFERENCE
MEAN	$\frac{1}{N} \sum Xi$	3.6	3.22	17.85	* More number of observations near mean *Probability of Getting Mean is High *Mean will be influenced by Outliers.
MEDIAN	*ARRANGR IN ASSENDIG OR DESSENDING ORDER *Finding Middle Value	3.7	3.33	17.71	*Media will Not be Influenced by the outliers, there for the we are seeing variations in Mean & Median *IN points +ve Outliers are there as Mean < Median *IN Score +ve Outliers are there as Mean < Median *IN Weight -ve Outliers are there as Mean > Median
MODE	*MOST FREQUENTLY OCCURRED NUMBER	3.92	3.44	17.02	*Median will be useful in the case of Categorical data
VARIENCE	For Population $\sigma^2 = \frac{1}{N} \sum_{i=1}^n (Xi - \mu)^2$ For Sample $S^2 = \frac{1}{n-1} \sum_{i=1}^n (Xi - \bar{X})^2$	0.29	0.96	3.19	*A large number of Variance indicates that the data in data set are far from mean and far from each other and smaller Variance is opposite to larger number's indication. *Here Weight is having more variance and it indicates that data points in weight are far from mean & far from each other as compared to Points & Score
STANDARD DIVIATION	$\sigma = \sqrt{\sigma^2}$	0.53	0.98	1.79	*Standard Deviation tells us how the spread of data around the Mean of data *Here Point's data are closely clustered around the Mean as compared to Others.
RANGE	Range = Max(xi) - Min (Xi)	2.17	3.91	8.4	*Range tells us that how much spread is there from the lowest value to the Highest values.

By Using Python

```
In [1]: from scipy import stats
import pandas
```

```
In [2]: data= pandas.read_csv("Q7.csv")
data.head(5)
```

```
Out[2]:
```

	Unnamed: 0	Points	Score	Weigh
0	Mazda RX4	3.90	2.620	16.46
1	Mazda RX4 Wag	3.90	2.875	17.02
2	Datsun 710	3.85	2.320	18.61
3	Hornet 4 Drive	3.08	3.215	19.44
4	Hornet Sportabout	3.15	3.440	17.02

```
In [3]: data.describe()
```

```
Out[3]:
```

	Points	Score	Weigh
count	32.000000	32.000000	32.000000
mean	3.596563	3.217250	17.848750
std	0.534679	0.978457	1.786943
min	2.760000	1.513000	14.500000
25%	3.080000	2.581250	16.892500
50%	3.695000	3.325000	17.710000
75%	3.920000	3.610000	18.900000
max	4.930000	5.424000	22.900000

Points	Score	Weigh
In [4]: <code>data["Points"].mean()</code>	In [10]: <code>data["Score"].mean()</code>	In [17]: <code>data["Weigh"].mean()</code>
Out[4]: 3.5965625000000006	Out[10]: 3.2172499999999995	Out[17]: 17.848750000000003
In [5]: <code>data["Points"].median()</code>	In [11]: <code>data["Score"].median()</code>	In [18]: <code>data["Weigh"].median()</code>
Out[5]: 3.6950000000000003	Out[11]: 3.325	Out[18]: 17.71
In [6]: <code>data["Points"].mode()</code>	In [12]: <code>data["Score"].mode()</code>	In [19]: <code>data["Weigh"].mode()</code>
Out[6]: 0 3.07 1 3.92 dtype: float64	Out[12]: 0 3.44 dtype: float64	Out[19]: 0 17.02 1 18.90 dtype: float64
In [7]: <code>data["Points"].var()</code>	In [13]: <code>data["Score"].var()</code>	In [20]: <code>data["Weigh"].var()</code>
Out[7]: 0.28588135080645166	Out[13]: 0.9573789677419356	Out[20]: 3.193166129032258
In [8]: <code>data["Points"].std()</code>	In [14]: <code>data["Score"].std()</code>	In [21]: <code>data["Weigh"].std()</code>
Out[8]: 0.5346787360709716	Out[14]: 0.9784574429896967	Out[21]: 1.7869432360968431
In [9]: <code>np.ptp(data["Points"]) #RAGNE</code>	In [15]: <code>np.ptp(data["Score"]) #RANGE</code>	In [16]: <code>np.ptp(data["Weigh"]) # RANGE</code>
Out[9]: 2.17	Out[15]: 3.9110000000000005	Out[16]: 8.399999999999999

Q8) Calculate Expected Value for the problem below

a) 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?

Xi	P(Xi)	P(Xi)*Xi
108	0.111111	12
110	0.111111	12.22222
123	0.111111	13.66667
134	0.111111	14.88889
135	0.111111	15
145	0.111111	16.11111
167	0.111111	18.55556
187	0.111111	20.77778
199	0.111111	22.11111
	$\sum (P(Xi) * (Xi))$	145.3333

Expected Value of the Weight of Randomly chosen patient = 145.33 Pounds

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
<pre>In [1]: import pandas data = pandas.read_csv("Q9_a.csv")</pre> <p>SPEED</p> <pre>In [2]: pandas.Series.skew(data["speed"]) Out[2]: -0.11750986144663393</pre> <pre>In [3]: pandas.Series.kurt(data["speed"]) Out[3]: -0.5089944204057617</pre> <p>Distance</p> <pre>In [4]: pandas.Series.skew(data["dist"]) Out[4]: 0.8068949601674215</pre> <pre>In [5]: pandas.Series.kurt(data["dist"]) Out[5]: 0.4050525816795765</pre>	<p>ST</p> <pre>In [10]: pandas.Series.skew(data["SP"]) Out[10]: 1.6114501961773586</pre> <pre>In [11]: pandas.Series.kurt(data["SP"]) Out[11]: 2.9773289437871835</pre> <p>WT</p> <pre>In [13]: pandas.Series.skew(data["WT"]) Out[13]: -0.6147533255357768</pre> <pre>In [14]: pandas.Series.kurt(data["WT"]) Out[14]: 0.9502914910300326</pre>

Inference:

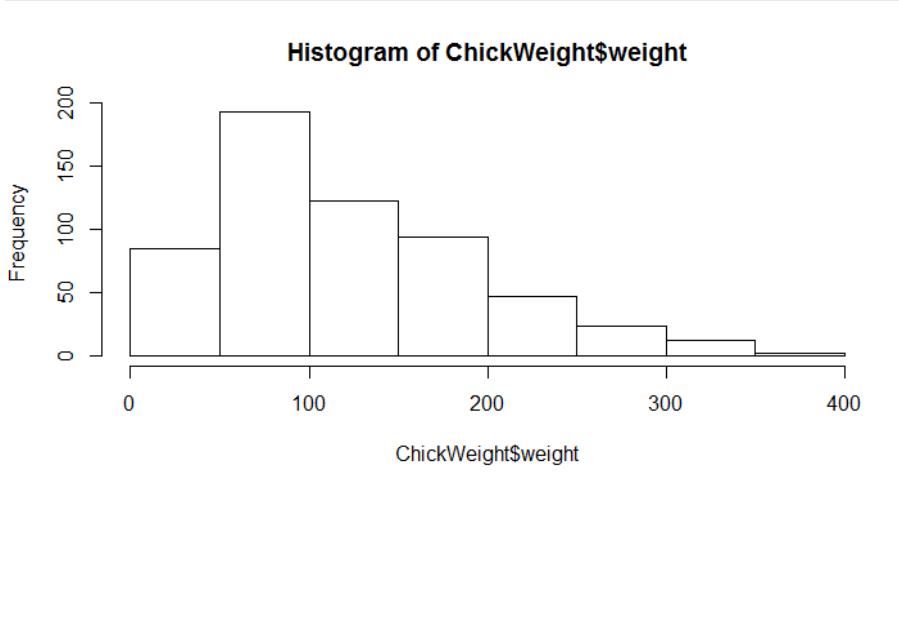
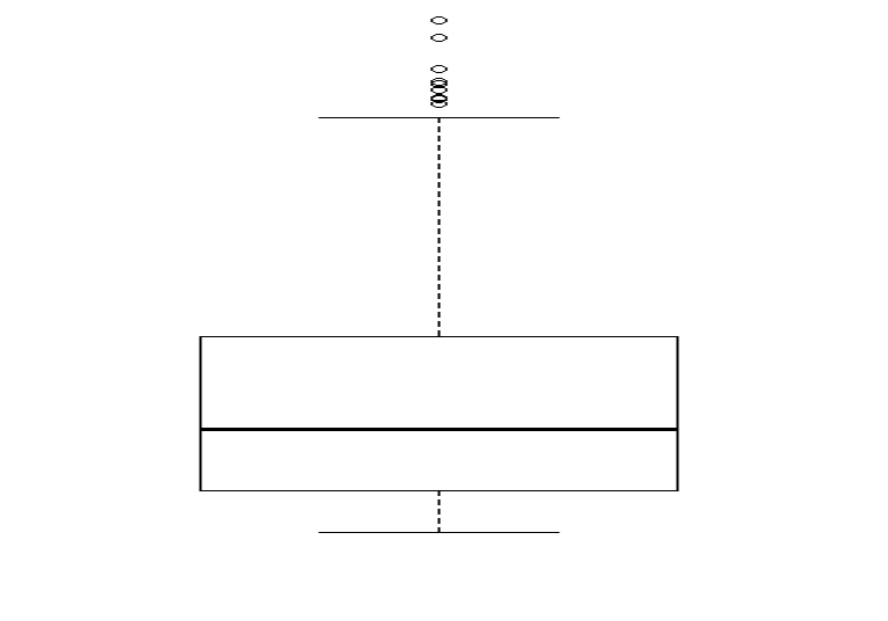
Speed:

- Skewness = -0.117
 - data is slightly Negatively Skewed or Left Skewed data (Mass of data is on right side of median),
 - means data spared is More on left side of the Median
- Kurtosis = -0.508
 - Data has platykurtic distribution& has thin tails compared to Normal dist.,
 - The distribution is flat as compared to Normal distribution.

Distance:

- Skewness = 0.806
 - data is skewed Positively or Right skewed data (Mass of data is on left side of median),
 - Means data spared is more on right side of the Median
- Kurtosis = 0.405
 - Data has Leptokurtic distribution & has thick tails as compared to normal dist.,
 - The distribution is peak as compared to Normal Distribution.

Q10) Draw inferences about the following boxplot & histogram

	
<p>Inference :</p> <ul style="list-style-type: none"> • Positively Skewed data(Right Skewed data) • Spared of the data on right side of the distribution is More & Mass of data is on left side of Median • Frequency of the data between 50 – 100 is more 	<p>Interance:</p> <ul style="list-style-type: none"> • Positively Skewed data or Right skewed data (Whisker is More on right side on median) • Spared of the data on right side of the distribution is More & Mass of data is on left side of Median • Positive Outliers are there on Right side of the distribution

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?

n = 2000 Pounds, N = 3000000, \bar{X} = 200, S = 30 Pounds

$$t = \frac{(X_i - \bar{X})}{\frac{S}{\sqrt{n}}}$$

$$CI = \bar{X} \pm t^* \frac{S}{\sqrt{n}}$$

94%	96%	98%
<p>In [5]: <code>stats.t.ppf(q=0.03,df=1999,loc=0,scale=1)</code> <code>#ppf(q, df, loc=0, scale=1)</code></p> <p>Out[5]: -1.8818614764780115</p> <p>In [6]: <code>200-1.8818614764780115*(30/np.sqrt(2000))</code></p> <p>Out[6]: 198.7376089443071</p> <p>In [7]: <code>200+1.8818614764780115*(30/np.sqrt(2000))</code></p> <p>Out[7]: 201.2623910556929</p>	<p><code>stats.t.ppf(q=0.02,df=1999,loc=0,scale=1)</code> <code>#ppf(q, df, loc=0, scale=1)</code></p> <p>-2.055089962825778</p> <p><code>200-2.055089962825778*(30/np.sqrt(2000))</code></p> <p>198.6214037429732</p> <p><code>200+2.055089962825778*(30/np.sqrt(2000))</code></p> <p>201.3785962570268</p>	<p><code>stats.t.ppf(q=0.01,df=1999,loc=0,scale=1)</code> <code>#ppf(q, df, loc=0, scale=1)</code></p> <p>-2.3282147761069725</p> <p><code>200-2.3282147761069725*(30/np.sqrt(2000))</code></p> <p>198.4381860483216</p> <p><code>200+2.3282147761069725*(30/np.sqrt(2000))</code></p> <p>201.5618139516784</p>

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?

```
In [1]: from scipy import stats

In [2]: stats.t.interval(alpha=0.06,df=1999,loc=200,scale=30)
Out[2]: (197.74162011566807, 202.25837988433193)

In [3]: stats.t.interval(alpha=0.04,df=1999,loc=200,scale=30)
Out[3]: (198.49520384079835, 201.50479615920165)

In [4]: stats.t.interval(alpha=0.02,df=1999,loc=200,scale=30)
Out[4]: (199.24783863179837, 200.75216136820163)
```

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 = $\frac{1}{N} \sum Xi = \frac{738}{18} = 41$
- Median = $\frac{40+41}{2} = 40.5$
- Variance = $S^2 = \frac{1}{n-1} \sum_{i=1}^n (Xi - \bar{X})^2 = \frac{1}{18-1} * 434 = 25.52941$
- Standard Deviation = $\sigma = \sqrt{\sigma^2} = \sqrt{25.52941} = 5.052664$

```
In [1]: import pandas
import numpy

In [2]: Marks = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]

In [3]: Marks = pandas.Series(Marks)

In [4]: Marks.mean() # MEAN
Out[4]: 41.0

In [5]: Marks.median() # MEDIAN
Out[5]: 40.5

In [6]: Marks.mode() # MODE
Out[6]: 0    41
dtype: int64

In [7]: Marks.var() # VARIANCE
Out[7]: 25.529411764705884

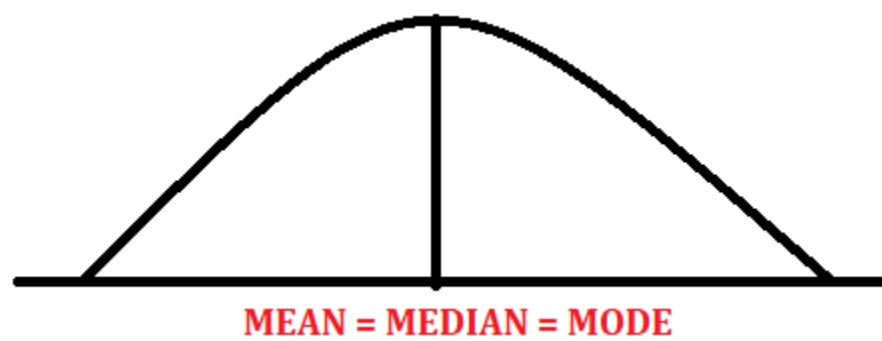
In [8]: Marks.std() # STANDARD DEVIATION
Out[8]: 5.05266382858645

In [9]: numpy.ptp(Marks) # RANGE
Out[9]: 22
```

2) What can we say about the student marks?

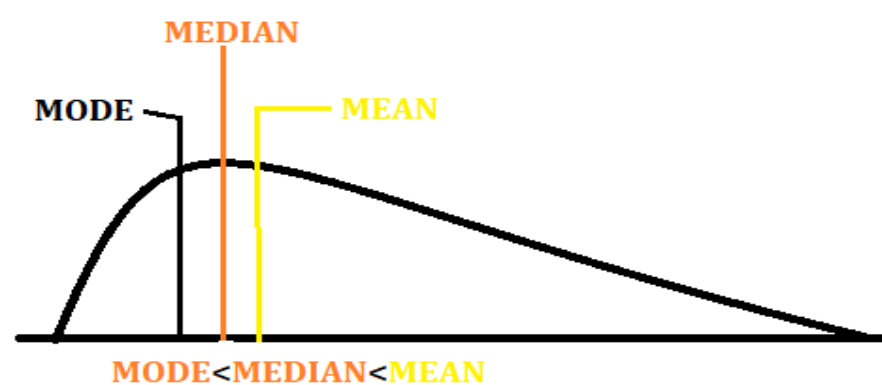
Mean = 41 <ul style="list-style-type: none">• Most of students' marks are nearer to 41
Median = 40.5 \cong Mean <ul style="list-style-type: none">• There is no too high (like 98,76) & too low marks (like 0,2) (Outliers) present
Standard deviation = 5.05 <p>As mean is approximately equal to median follows Normal distribution,</p> <ul style="list-style-type: none">• $1\sigma = (41-5=36, 41+5=47)$• 68% of students are scored between 36 to 47• $2\sigma = (41-10=31, 41+10=51)$• 95% of students are scored between 31 to 51• $3SD = (41-15=26, 41+16=57)$• All most all (99.7%) students are scored between 26 to 57

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



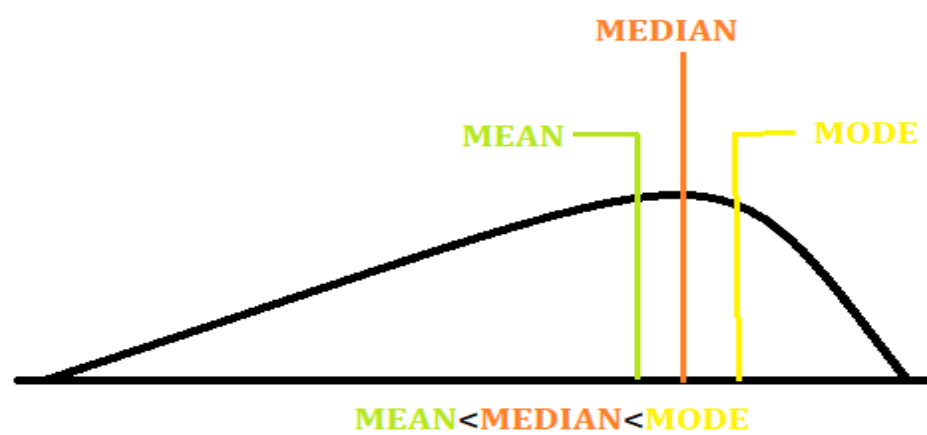
When Mean = Median, we can say data is Normally Distributed.

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



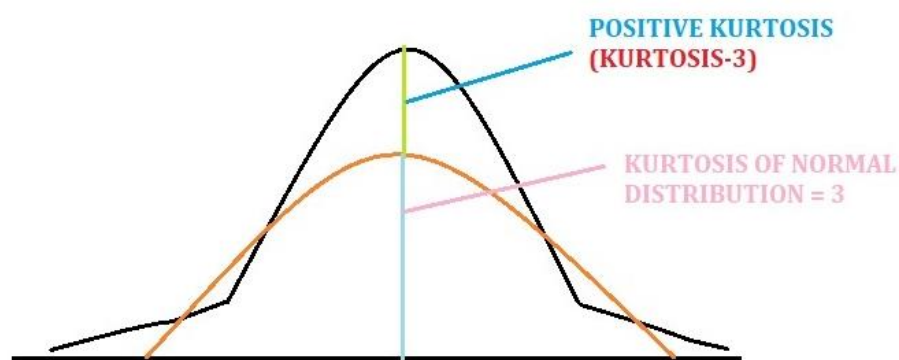
When Mean > Median, we can say Positively Skewed data (Right Skewed data).

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



When Mean < Median, we can say Negatively Skewed data (left Skewed data).

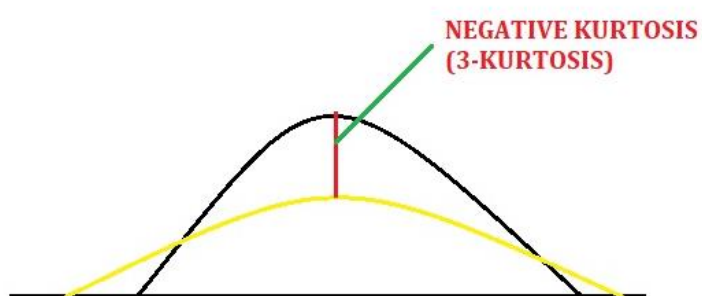
Q16) What does positive kurtosis value indicates for a data?



Positive Kurtosis (Excess Kurtosis) indicates that,

- Distribution is Leptokurtic (peak of bell curve is more as compared to Normal distribution)
- Spread There are more values around mean.

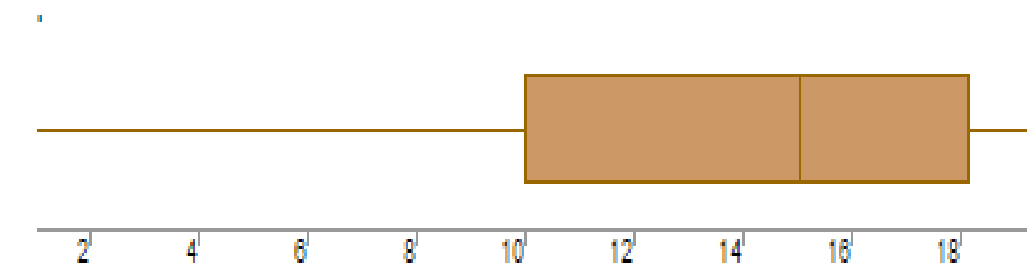
Q17) What does negative kurtosis value indicates for a data?



Negative Kurtosis indicates that,

- Distribution is Platykurtic (peak of bell curve is less as compared to Normal distribution)
- Spread of the data is More (There are more far values from mean).

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



What can we say about the distribution of the data?

- Most of the data lies between 10 to 18.
- Q1 = Quartile 1 = 10
- Q2 = Quartile 2 = 15 = MEDIAN = 50th Percentile
- Q3 = Quartile 3 = 18

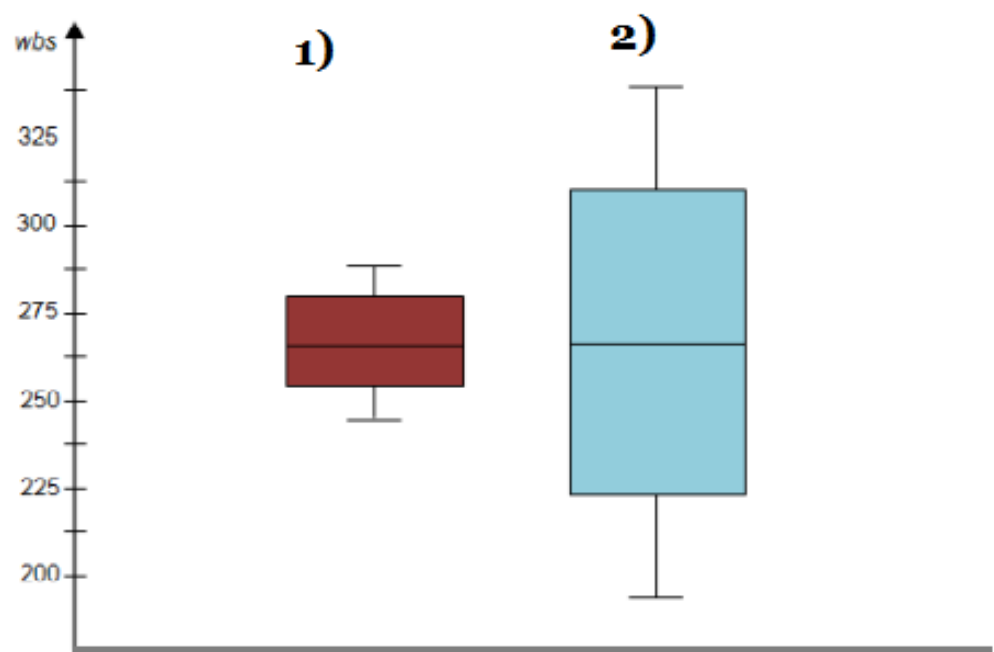
What is nature of skewness of the data?

Negatively skewed data: There are negative outliers present in the data

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.

Boxplot 1	Boxplot 2
Data ranges between 240 to 280	Data ranges between 190 to 340
Mean = Median = Mode = Qurtile2(Q2) =260	Mean = Median = Mode= Qurtile2(Q2) =260
Normally Distributed	Normally Distributed
Quartile1 = 255	Quartile1 = 220
Quartile3 = 280	Quartile3 = 310
IQR (INTER QURTAIL RANGE) is less = 280-255 =25	IQR (INTER QURTAIL RANGE) is more = 310-220 =90

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

- P(MPG>38)
- P(MPG<40)
- P (20<MPG<50)

Mean = 34.42

STANDARD DEVIATION = S = 9.131445

$P(X \leq X_i) = P(Z_{X_i})$

<p>P(MPG>38)</p> <p>= 1 - P(MPG≤38)</p> <p>= 1 -P$\left(Z\left(\frac{(Xi - \bar{X})}{s}\right)\right)$</p> <p>= 1 – P$\left(Z\left(\frac{(38 - 34.42)}{9.1314}\right)\right)$</p> <p>= 1-P(Z_{0.392})</p> <p>=1-0.6517</p> <p>=0.3483</p> <p>=34.83%</p>	<p>P(MPG≤ 40)</p> <p>= P(MPG≤ 40)</p> <p>= P$\left(Z\left(\frac{(Xi - \bar{X})}{s}\right)\right)$</p> <p>= P$\left(Z\left(\frac{(40 - 34.42)}{9.1314}\right)\right)$</p> <p>= P(Z_{0.611})</p> <p>= 0.7291</p> <p>=72.91%</p>	<p>P(20<MPG<50)</p> <p>=P(MPG≤50) – P(MPG≤ 20)</p> <p>= P$\left(Z\left(\frac{(50 - 34.42)}{9.1314}\right)\right)$ -</p> <p>P$\left(Z\left(\frac{(20 - 34.42)}{9.1314}\right)\right)$</p> <p>=P(Z_{1.707}) – P(Z_{-1.579})</p> <p>=0.9564-0.0571</p> <p>=0.8993</p> <p>=89.93%</p>
--	--	--

In [1]:

from scipy import stats
import pandas

In [2]:

data = pandas.read_csv("Cars.csv")
data.head(5)

Out[2]:

	HP	MPG	VOL	SP	WT
0	49	53.700681	89	104.185353	28.762059
1	55	50.013401	92	105.461264	30.466833
2	55	50.013401	92	105.461264	30.193597
3	70	45.696322	92	113.461264	30.632114
4	53	50.504232	92	104.461264	29.889149

In [3]:

MPG=data["MPG"]
MPG.head()

Out[3]:

0	53.700681
1	50.013401
2	50.013401
3	45.696322
4	50.504232

Name: MPG, dtype: float64

a. P(MPG>38)

In [4]:

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

Out[4]:

0.3475939251582705

b. P(MPG<40)

In [5]:

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

Out[5]:

0.7293498762151616

c. P (20<MPG<50)

In [6]:

X1_20 = stats.norm.cdf(x=20,loc=MPG.mean(),scale=MPG.std())
X1_20

Out[6]:

0.05712377632115936

In [7]:

X2_58 = stats.norm.cdf(x=50,loc=MPG.mean(),scale=MPG.std())
X2_58

Out[7]:

0.955992693289364

In [8]:

P = X2_58 - X1_20
P

Out[8]:

0.8988689169682046

Q 21) Check whether the data follows normal distribution
a) Check whether the MPG of Cars follows Normal Distribution
Dataset: Cars.csv

MPG OF CARS



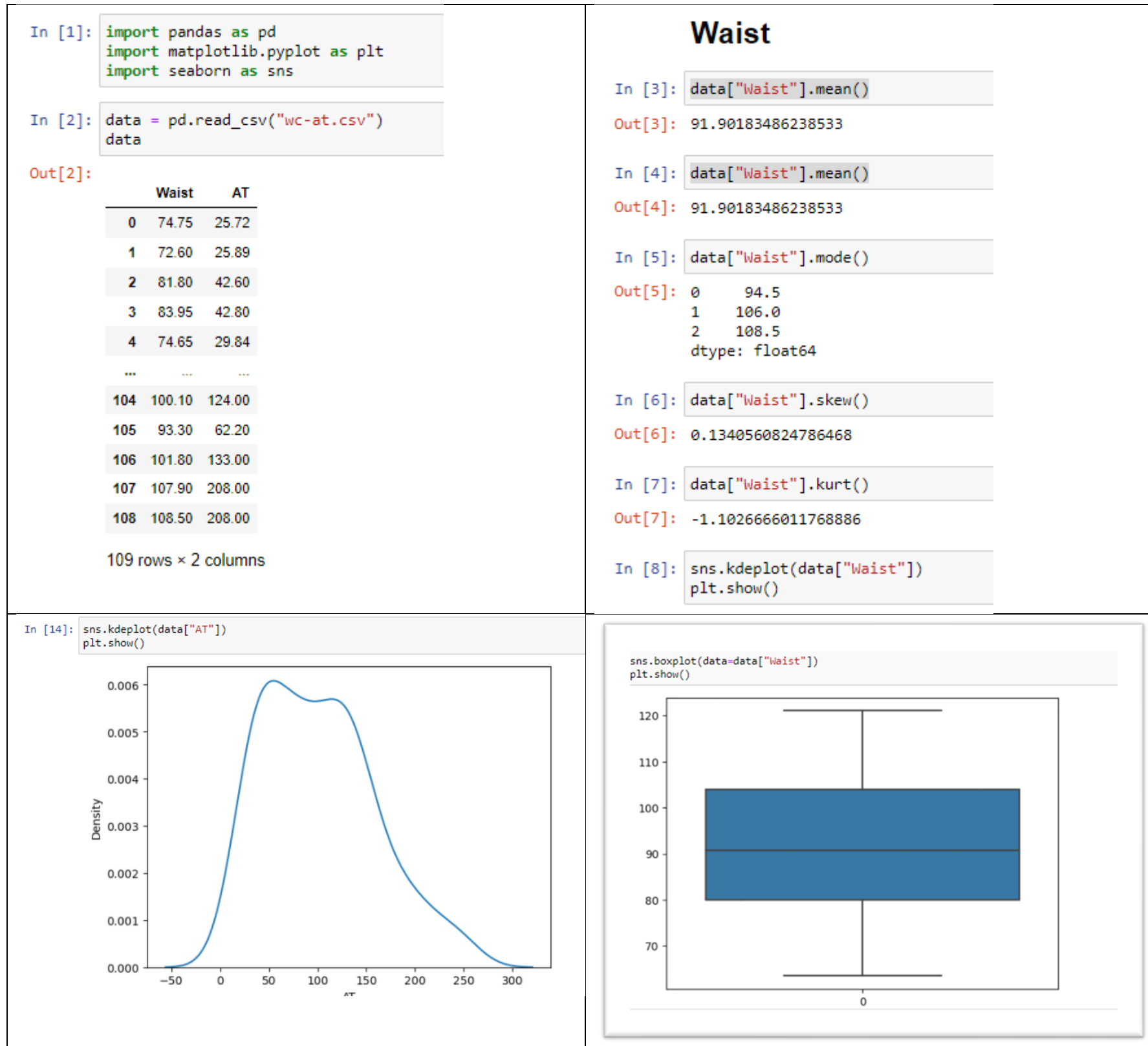
Since,

- 1)MEAN \neq MEDIAN,
- 2)Skewness = 0.177
- 3)Kurtosis = 0.6116
- 4)IN Box plot Q2 is not at center, whisker is more negative side , Midian(Q2) is nearer to Q3 and in bell curve skewed towards negative numbers

We can Say That the “MPG” data is Sightly Right skewed or Negatively Skewed data.

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

Waist:



Since,

- 1) MEAN = MEDIAN = 91.9018,
- 2) Skewness = 0.134 ≈ 0
- 3) Kurtosis = -1.01
- 4) IN Box plot Q2 is approximately at center

We can Say That the “Waist” data is Normally Distributed

AT

```
In [1]: import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
In [2]: data = pd.read_csv("wc-at.csv")
data
```

Out[2]:

	Waist	AT
0	74.75	25.72
1	72.60	25.89
2	81.80	42.60
3	83.95	42.80
4	74.65	29.84
...
104	100.10	124.00
105	93.30	62.20
106	101.80	133.00
107	107.90	208.00
108	108.50	208.00

109 rows × 2 columns

AT

```
In [9]: data["AT"].mean()
```

Out[9]: 101.89403669724771

```
In [10]: data["AT"].median()
```

Out[10]: 96.54

```
In [11]: data["AT"].mode()
```

Out[11]: 0 121.0
1 123.0
dtype: float64

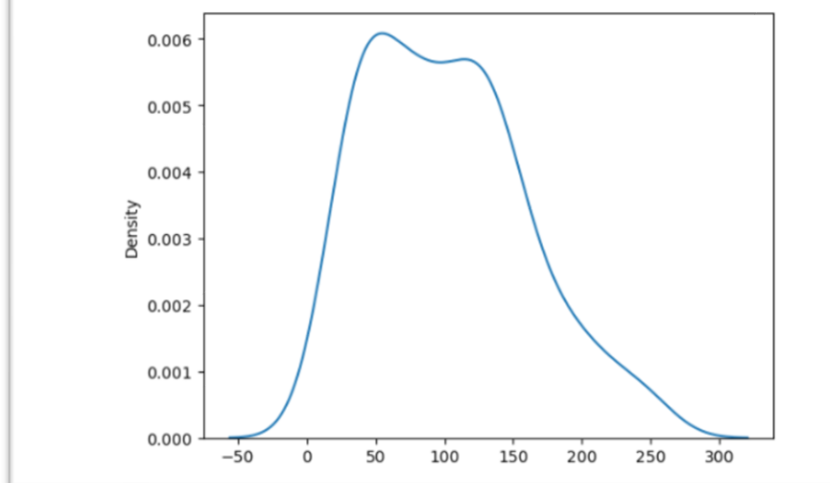
```
In [12]: data["AT"].skew()
```

Out[12]: 0.584869324127853

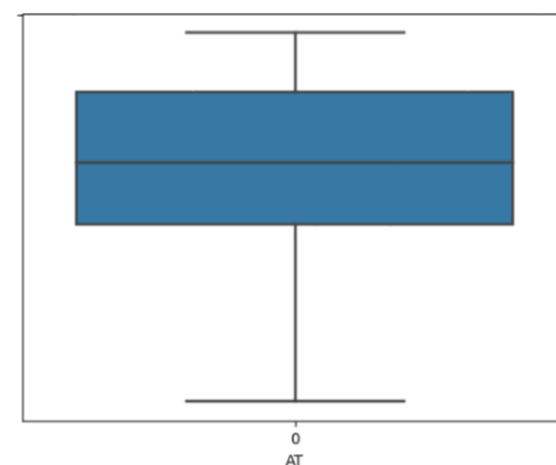
```
In [13]: data["AT"].kurt()
```

Out[13]: -0.28557567504584425

```
In [14]: sns.kdeplot(data["AT"])
plt.show()
```



```
In [28]: sns.boxplot(data=data["AT"])
plt.show()
```



Since,

- 1) MEAN \neq MEDIAN
- 2) Skewness, is not nearer zero
- 3) Kurtosis = -2.855 is not nearer to zero
- 4) IN Box plot Q2 is not at center and whisker is More in Positive side

We can Say That the “AT” data is Moderately Positively Skewed Data.

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

Confidence Interval	Alpha(α) =(1-CL)/2	Z score (Z table)
90%	0.10/2=0.05	± 1.64
94%	0.06/2=0.03	± 1.88
60%	0.40/2=0.20	± 0.84

Using Python:

<pre>from scipy import stats</pre> <p>90%</p> <pre>stats.norm.ppf(q=0.05, loc=0, scale=1) #ppf(q, loc=0, scale=1)</pre> <p>-1.6448536269514729</p> <pre>stats.norm.ppf(q=0.95, loc=0, scale=1)</pre> <p>1.6448536269514722</p> <p>or</p> <pre>stats.norm.interval(alpha=0.90, loc=0, scale=1) #interval(alpha, loc=0, scale=1)</pre> <p>(-1.6448536269514729, 1.6448536269514722)</p>	<p>94%</p> <pre>stats.norm.ppf(q=0.03, loc=0, scale=1) #ppf(q, loc=0, scale=1)</pre> <p>-1.880793608151251</p> <pre>stats.norm.ppf(q=0.97, loc=0, scale=1) #ppf(q, loc=0, scale=1)</pre> <p>1.8807936081512509</p> <p>or</p> <pre>stats.norm.interval(alpha=0.94, loc=0, scale=1) #interval(alpha, loc=0, scale=1)</pre> <p>(-1.8807936081512509, 1.8807936081512509)</p>	<p>60%</p> <pre>stats.norm.ppf(q=0.2, loc=0, scale=1) #ppf(q, loc=0, scale=1)</pre> <p>-0.8416212335729142</p> <pre>stats.norm.ppf(q=0.80, loc=0, scale=1) #ppf(q, loc=0, scale=1)</pre> <p>0.8416212335729143</p> <p>or</p> <pre>stats.norm.interval(alpha=0.60, loc=0, scale=1) #interval(alpha, loc=0, scale=1)</pre> <p>(-0.8416212335729142, 0.8416212335729143)</p>
--	--	--

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

Confidence Interval	Df	T score (t table)
95%	25	2.060
96%		2.060
99%		2.787

<pre>from scipy import stats</pre> <p>95%</p> <pre>stats.t.ppf(q=0.025, df=25, loc=0, scale=1) #ppf(q, df, loc=0, scale=1)</pre> <p>-2.0595385527532946</p> <pre>stats.t.ppf(q=0.975, df=25, loc=0, scale=1) #ppf(q, df, loc=0, scale=1)</pre> <p>2.059538552753294</p> <p>or</p> <pre>stats.t.interval(alpha=0.95, df=25, loc=0, scale=1) #interval(alpha, df, loc=0, scale=1)</pre> <p>(-2.059538552753294, 2.059538552753294)</p>	<p>96%</p> <pre>stats.t.ppf(q=0.02, df=25, loc=0, scale=1) #ppf(q, df, loc=0, scale=1)</pre> <p>-2.1665866344527567</p> <pre>stats.t.ppf(q=0.98, df=25, loc=0, scale=1) #ppf(q, df, loc=0, scale=1)</pre> <p>2.1665866344527562</p> <p>or</p> <pre>stats.t.interval(alpha=0.96, df=25, loc=0, scale=1) #interval(alpha, df, loc=0, scale=1)</pre> <p>(-2.1665866344527562, 2.1665866344527562)</p>	<p>99%</p> <pre>stats.t.ppf(q=0.005, df=25, loc=0, scale=1) #ppf(q, df, loc=0, scale=1)</pre> <p>-2.7874358136758515</p> <pre>stats.t.ppf(q=0.995, df=25, loc=0, scale=1) #ppf(q, df, loc=0, scale=1)</pre> <p>2.787435813675851</p> <p>or</p> <pre>stats.t.interval(alpha=0.99, df=25, loc=0, scale=1) #interval(alpha, df, loc=0, scale=1)</pre> <p>(-2.787435813675851, 2.787435813675851)</p>
---	---	--

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:

Claim: an average light bulb lasts 270 days

$= \mu = 270$

Number of Sample bulbs = $n = 18$

Average days of sample = $\bar{X} = 260$ days

Standard Deviation of Sample = $S = 90$ days

To find probability that 18 randomly selected bulbs would have an average life of no more than 260 days,

We need to calculate t statistics for given data,

$$t = \frac{(\bar{X} - \mu)}{\frac{S}{\sqrt{n}}}$$

$$t = \frac{(260 - 270)}{\frac{90}{\sqrt{18}}}$$

$$t = -0.4714$$

$$P_{t=-0.471, df=17}$$

```
In [1]: import scipy.stats as st  
  
In [2]: st.t.sf(abs(-.4714), 17)  
= Out[2]: 0.32167411684460556
```

18 randomly selected bulbs would have an average life of no more than 260 days = 0.3216 = 32.16 %