Assignment – 1:

Basic Statistics – 1

Bhagyashree S. D.

deshpandebhagya1997@gmail.com

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	Categorical
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	Ratio
Time on a Clock with Hands	Ratio
Number of Children	Nominal
Religious Preference	Nominal
Barometer Pressure	Interval
SAT Scores	Ordinal
Years of Education	Ratio

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

Ans: When 3 coins are tossed, then the total possible sample spaces are $2^3 = 8$

To find the probability that two heads and one tail are obtained.

Possible sample space:

(HHH), (HHT), (HTH), (THH), (TTH), (THT), (HTT), (TTT)

From this Sample Space, we need to find the probability that 2 heads and 1 tail is obtained.

We have 3 possible outcomes.

$\therefore$ P (Getting 2 heads and 1 tail) = $\frac{3}{8} = 0.375$

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

a) Equal to 1
b) Less than or equal to 4
c) Sum is divisible by 2 and 3

Ans: 2 dice are rolled, the possible outcomes are $6^2 = 36$.

They are,

{(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)}

We have to find the probability that the sum is

a) Equal to 1
   If two dice are rolled then we will not get sum equal to 1.
   i.e., The event that the sum is equal to 1 = 0
   ∴ The probability that the sum is equal to 1 = 0/36 = 0
b) Less than or equal to 4
   From the sample space we have to choose the event of getting sum less than or equal to 4.
   We have 6 possible outcomes. They are:
   (1,1), (1,2), (1,3), (2,1), (2,2), (3,1)
   ∴ The probability that getting sum less than or equal to 4 = 6/36
   = 1/6 = 0.166
c) Sum is divisible by 2 and 3
   From the sample space we have to choose the event of getting sum is divisible by 2 and 3.
   We have 6 possible outcomes. They are:
   (1,5), (2,4), (3,3), (4,2), (5,1), (6,6)
   ∴ The probability that getting sum divisible by 2 and 3 = 6/36
   = 1/6 = 0.166

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: Total number of balls in the bag = 7
Number of ways in which 2 balls can be drawn = ${}^{7}C_{2} = 21$
Now, we have to pick 2 balls out of 5 balls (Because we are not considering blue balls)
∴ The number of ways in which 2 balls are drawn from 5 balls = ${}^{5}C_{2} = 10$
∴ The probability that none of the balls picked are blue = 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

Ans: We know that, Expectation of X $E(X) = \Sigma(X*P(X))$

$\therefore$ We have

CHILD	Candies count (X)	Probability (P(X))	X*P(X)
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
SUM			3.09

$\therefore$ The Expected number of Candies = $3.09 \approx 3$

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: Using Python:

```
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
data = pd.read_csv('Q7.csv')
data.mean() #Mean of the data
Points 3.596563
Score 3.217250
Weigh 17.848750
data.median() #Median of the data
Points 3.695
Score 3.325
Weigh 17.710
data['Points'].mode
3.90
data['Score'].mode
2.620
data['Weigh'].mode
16.46
data.var() #Variance of the data
Points 0.285881
Score 0.957379
Weigh 3.193166
data.std() #Standard deviation of the data
Points 0.534679
Score 0.978457
Weigh 1.786943
```

We have tabulated the answers below:

	Points	Score	Weigh
Mean	3.596563	3.217250	17.848750
Median	3.695	3.325	17.710
Mode	3.90	2.620	16.46
Variance	0.285881	0.957379	3.193166
Standard Deviation	0.534679	0.978457	1.786943

Inference: From the data, we observe that, mean, median and mode are not equal.

∴ We conclude that the given data is skewed and also there may be a chance of presence of the outliers.

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?

Ans: Given, the weights of the patients at the clinic are:

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

The probability of choosing one person = 1/9

∴ The Expected value is $E(X) = \Sigma(X*P(X))$

∴ We have,

X	P(X)	X*P(X)
108	1/9	12
110	1/9	12.2222
123	1/9	13.6666
134	1/9	14.8888
135	1/9	15
145	1/9	16.1111
167	1/9	18.5555
187	1/9	20.7777
199	1/9	22.1111
SUM		145.3330

The Expected weight of the patient is 145.333 pounds.

Q9) Calculate Skewness, Kurtosis & draw inferences on the following data Cars speed and distance

Use Q9_a.csv

Ans: Using Python:

```
import pandas as pd
data1 = pd.read_csv('Q9_a.csv')
data1.skew()
Index 0.000000
speed -0.117510
dist 0.806895
data1.kurtosis()
Index -1.200000
speed -0.508994
dist 0.405053
dtype: float64
```

	Speed	Distance
Skewness	-0.117510	0.806895
Kurtosis	-0.508994	0.405053

From the skewness of speed, we observe that the data of speed is fairly symmetrical and from the distance, we observe that the data is moderately positively skewed.

SP and Weight (WT)

Use Q9_b.csv

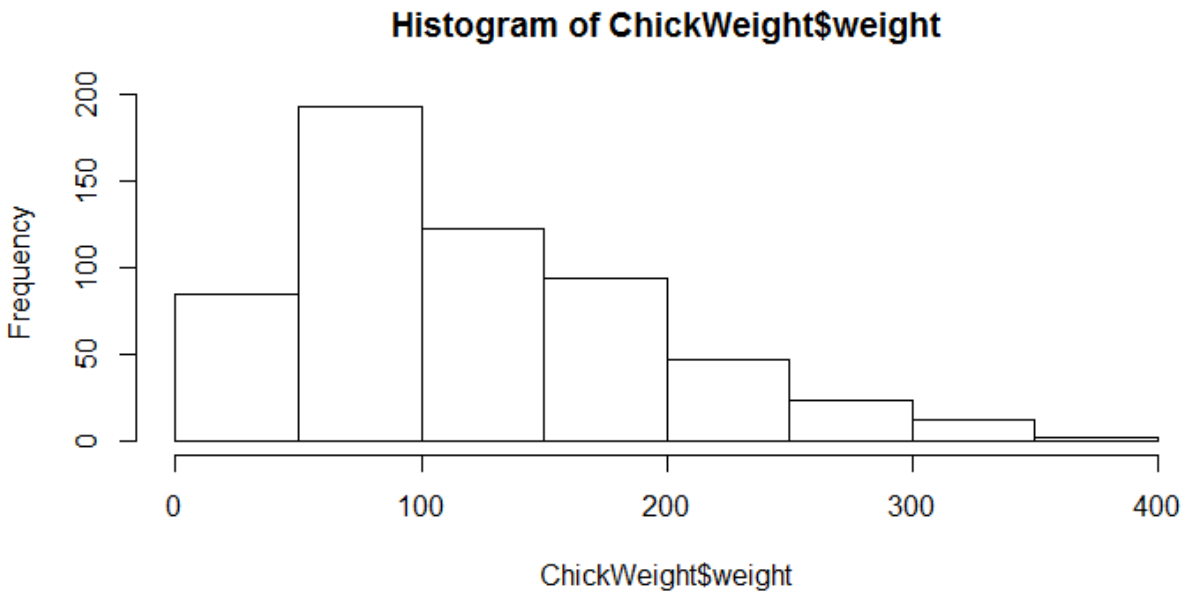
Ans: Using python:

```
import pandas as pd
data2 = pd.read_csv('Q9_b.csv')
data2.skew()
Unnamed: 0 0.000000
SP 1.611450
WT -0.614753
dtype: float64
data2.kurtosis()
Unnamed: 0 -1.200000
SP 2.977329
WT 0.950291
dtype: float64
```

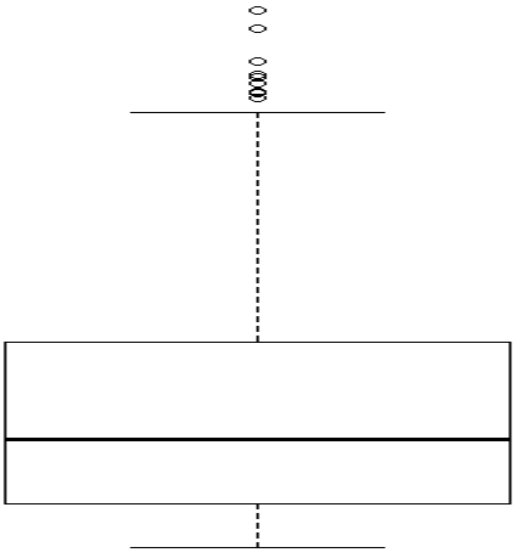
	Speed	Weight
Skewness	1.611450	-0.614753
Kurtosis	2.977329	0.950291

From the skewness of speed we observe that the data of speed is positively skewed and the skewness of weight we observe that the data is moderately negatively skewed.

Q10) Draw inferences about the following boxplot & histogram



Ans: From the above plot, we can say that the data is distributed symmetrically (Positively symmetric).



From the above plot, we can say that the data is symmetrically distributed and we have noticed that there are some outliers.

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?

Ans: Sample size, n = 2000

Sample mean, $\bar{X} = 200$

Sample variance, $s^2 = 30$

$\therefore$ Class interval for mean is given by,

$[\ \bar{X} - Z_{\alpha/2}\ S/\sqrt{n},\ \bar{X} + Z_{\alpha/2}\ S/\sqrt{n}\ ]$

- 94% class interval where $Z_{\alpha/2} = 1.89$ is [198.7383, 201.2616]
- 96% class interval where $Z_{\alpha/2} = 2.33$ is [198.6223, 201.3776]
- 98% class interval where $Z_{\alpha/2} = 2.96$ is [198.4394, 201.5605]

Using Python:

```
import numpy as np
import scipy.stats as st
st.norm.interval(alpha = 0.94, loc = 200, scale = 30/np.sqrt(2000)) #94% CI
(198.738325292158, 201.261674707842)
st.norm.interval(alpha = 0.96, loc = 200, scale = 30/np.sqrt(2000)) #96% CI
(198.62230334813333, 201.37769665186667)
st.norm.interval(alpha = 0.98, loc = 200, scale = 30/np.sqrt(2000)) #98% CI
(198.43943840429978, 201.56056159570022)
```

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?

Ans: Using Python:

```
import pandas as pd
x = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]
data = pd.DataFrame(x)
print(data.mean())
print(data.median())
print(data.var())
print(data.std())
```

Result:

Mean = 41

Median = 40.5

Variance = 25.529412

Standard Deviation = 5.052664

On an average a student scores 41 marks.

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

Ans: The nature of skewness is perfectly symmetric, i.e., zero skewed.

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

Ans: The nature of skewness is positively skewed.

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

Ans: The nature of skewness is negatively skewed.

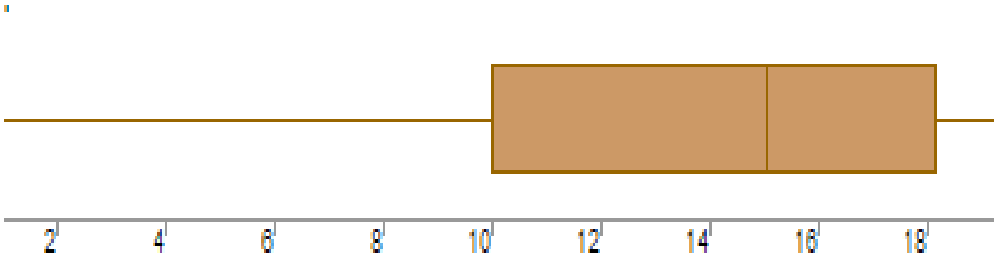
Q16) What does positive kurtosis value indicates for a data?

Ans: Positive tail indicates that we have heavy tails that is lot data lies in tails.

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

Ans: Negative tail indicates that we have light tails that is little data lies in the tails.

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



What can we say about the distribution of the data?

Ans: Here, the distribution is skewed distribution.

What is nature of skewness of the data?

Ans: The nature of skewness is negatively skewed.

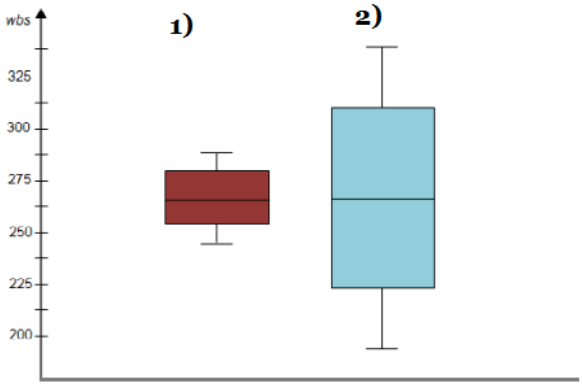
What will be the IQR of the data (approximately)?

Ans: $IQR = Q_3 - Q_1$

$= 18 - 10$

$IQR = 8$

Q19) Comment on the below Boxplot visualizations?



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

Ans: Here both the plots indicate that they follow normal distribution. The difference is Boxplot 1 has lesser range when compared to Boxplot 2.

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

a. P(MPG>38)
b. P(MPG<40)
c. P (20<MPG<50)

Ans: Using Python:

```
import pandas as pd

import scipy.stats as st

data = pd.read_csv('Cars.csv')

Mean = data['MPG'].mean()

SD = data['MPG'].std()

# P(MPG > 38)

1 - (st.norm.cdf(38, loc = Mean, scale = SD))

0.3475939251582705
# P(MPG < 40)

1 - (st.norm.cdf(40, loc = Mean, scale = SD))

0.27065012378483844
# P(20 < MPG < 50)

st.norm.cdf(50, loc = Mean, scale = SD) - st.norm.cdf(20, loc = Mean,
scale = SD)

0.8988689169682046
```

We have,

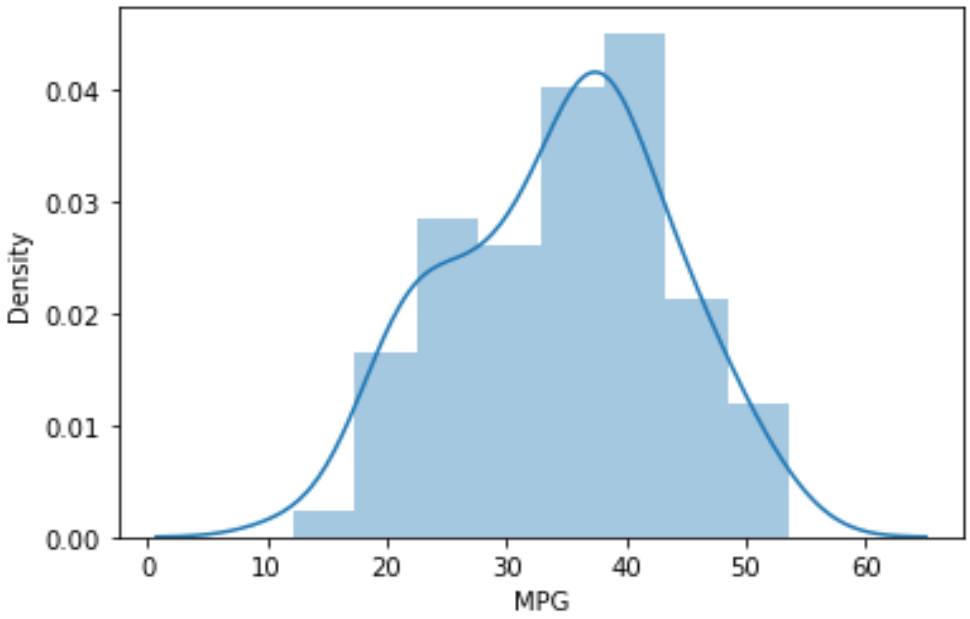
P (MPG > 38) = 0.3475939251582705
P (MPG < 40) = 0.27065012378483844
P (20 < MPG < 50) = 0.8988689169682046

Q 21) Check whether the data follows normal distribution

a) Check whether the MPG of Cars follows Normal Distribution
Dataset: Cars.csv

Ans: Using Python:

```
import pandas as pd
import seaborn as sns
import scipy.stats as st
data = pd.read_csv('Cars.csv')
sns.distplot(data['MPG'])
```

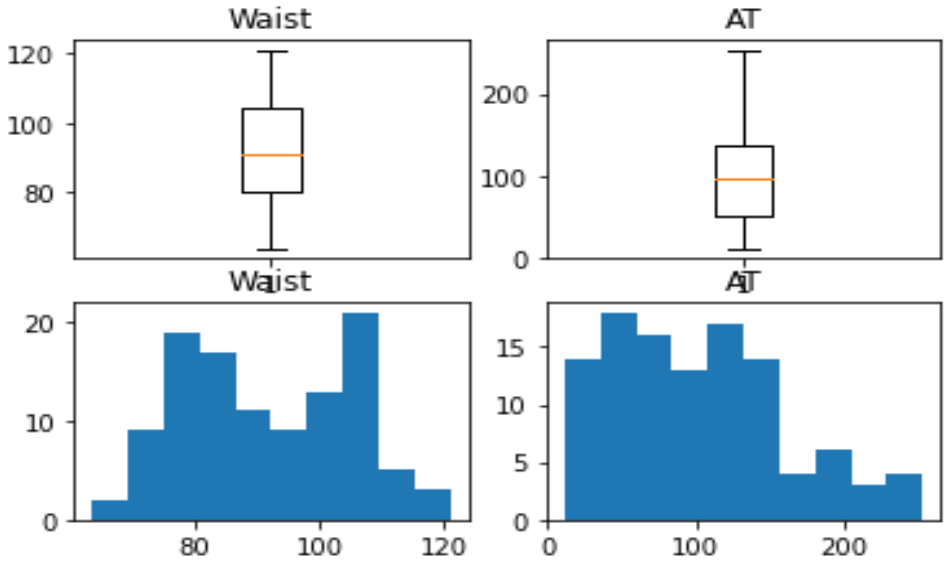


From the plot, we can say that the given data approximately follows normal distribution.

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

Ans: Using Python:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data = pd.read_csv('wc-at.csv')
figure, axis = plt.subplots(2,2)
axis[0,0].boxplot(data['Waist'])
axis[0,0].set_title('Waist')
axis[0,1].boxplot(data['AT'])
axis[0,1].set_title('AT')
axis[1,0].hist(data['Waist'])
axis[1,0].set_title('Waist')
axis[1,1].hist(data['AT'])
axis[1,1].set_title('AT')
```



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

Ans: Using Python:

```
import scipy.stats as st
print(st.norm.ppf(0.90))
print(st.norm.ppf(0.94))
print(st.norm.ppf(0.60))
1.2815515655446004
1.5547735945968535
0.2533471031357997
```

We get,

For 90% Confidence Interval, Z score is 1.2815515655446004
For 94% Confidence Interval, Z score is 1.5547735945968535
For 60% Confidence Interval, Z score is 0.2533471031357997

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

Ans: Using Python:

```
import scipy.stats as st
print(st.t.ppf(0.95,24))
print(st.t.ppf(0.96,24))
print(st.t.ppf(0.99,24))
1.7108820799094275
1.8280511719596342
2.4921594731575762
```

We get,

For 95% Confidence Interval, t score is 1.7108820799094275

For 96% Confidence Interval, t score is 1.8280511719596342

For 99% Confidence Interval, t score is 2.4921594731575762

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: The hypotheses formulation:

H_0 : Number of days an average light bulb lasts = 270

H_1 : Number of days an average light bulb lasts < 270

Level of significance – 5%

The test statistic is given by,

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

$$= \frac{260-270}{90/\sqrt{18}}$$

t = -0.4714

t table (0.05,17) = 1.771

Since t score value is less than t table value, we do not reject the null hypothesis.