Assignment – 1: **M. S. Jayanth**

Basic Statistics 1 **msjayanth185@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?

Answer:

When 3 coins are tossed, then the possible sample spaces are: $2^3=8$
Here we need to find probability of 2 heads and 1 tail.
Let us see the possible sample space:
(HHH), (HHT), (HTH), (THH), (TTH), (THT), (HTT), (TTT)
From this we need to choose the event of 2 heads and a tail.
We have 3 possible outcomes.
Therefore P (getting 2 heads and a tail) = 3/8 = 0.375.

Alter:

Since we have tossing of coin it follows binomial distribution
Therefore, $P(X=2) = {}^3C_2*(0.5)^2*(1-0.5)^1$

$=3*0.25*0.5$

$= 0.375$

Therefore P (getting 2 heads and a tail) = 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 2and 3

Answer:

When two dice are rolled the possible sample space are: $6^2 = 36$. Those 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 need to find the probability that sum is

a. Equal to 1
   If two dice are rolled then we will not get sum equal to 1.
   That is the event that the sum is equal to 1 is 0.
   Therefore, the probability that sum is equal to 1 is 0/36=0

b. Less than or equal to 4
   From the sample space we need to choose the event of getting sum less than or equal to 4.
   We have 6 possible outcomes that are
   (1,1), (1,2), (1,3), (2,1), (2,2), (3,1).
   Therefore, the probability that getting sum less than or equal to 4 = 6/36
   =1/6=0.166

c. Is divisible by 2 and 3
   From the sample space we need to choose the event of getting sum is divisible by 2 and 3.
   We have 6 possible outcomes that are
   (1,5), (2,4), (3,3), (4,2), (5,1), (6,6).
   Therefore, the probability that getting sum is 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?

Answer:

Total number of balls contain in the bag = 7
In which 2 balls are drawn out of 7
Then the number of ways in which 2 balls are drawn from 7= $^{7}C_{2}$ = 21
Now, we need to pick 2 balls out of 5 balls as we are not considering blue balls.
Therefore, the number of ways in which 2 balls are drawn from 5= $^{5}C_{2}$ = 10.
Therefore, the probability that none of the balls are blue are = 10/21=0.4761.

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

Answer:

We know that, Expected value E(X) = Σ(X*P(X)).
Therefore,

CHILD	Candidates 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

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

Answer:

Using python:

```
import pandas as pd
data= pd.read_csv("C:/Users/jayanth/Downloads/Q7.csv")

data.mean() # Calculating Mean

data.median() # calculating Median

data[“Points”].mode
data[“Scores”].mode
data[“Weigh”]. ,mode

data.var() # Calculating Variance

data.std() # Calculating standard deviation
```

We have tabulated the answers below:

	Points	Scores	Weighs
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
Standard deviation	0.5346	0.9784	1.7869

Inference:

From our data, we observe that mean, median and mode are not equal Therefore we can say that our data is skewed and also we can say that there may be chance of outliers present in our data.

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?

Answer:

Given that the weights of the patient at the clinic are: 108, 110, 123, 134, 135, 145, 167, 187, and 199.
The probability of choosing a person is 1/9
Then the Expected value is given by $E(X) = \Sigma(X*P(X))$.
Therefore,

X	P(X)	X*P(X)
108	1/9	12
110	1/9	12.2222
123	1/9	13.6667
134	1/9	14.8889
135	1/9	15
145	1/9	16.1111
167	1/9	18.5556
187	1/9	20.7778
199	1/9	22.1111
Sum=		145.3334

The expected weight of the patient is 145.3334 pounds.

Alter:

We can simply find mean for the given data

Using python:

```
X=[108, 110, 123, 134, 135, 145, 167, 187, 199]
data=pd.DataFrame(X)
data.mean()
```

We get mean of 145.3333.

Q9) Calculate Skewness, Kurtosis & draw inferences on the following data
Cars speed and distance
Use Q9_a.csv

Answer:

Using python:

```
import pandas as pd
data1=( C:\\Users\\jayanth\\Downloads\\Q9_a.csv")
data1.skew()
data1.kurtosis()
```

	Speed	Distance
Skewness	0.1175	0.8069
kurtosis	-0.5090	0.4050

From the skewness of speed we came to know that the data of speed is fairly symmetrical and from the distance we came to know that the data is moderately positively skewed.

SP and Weight(WT)
Use Q9_b.csv

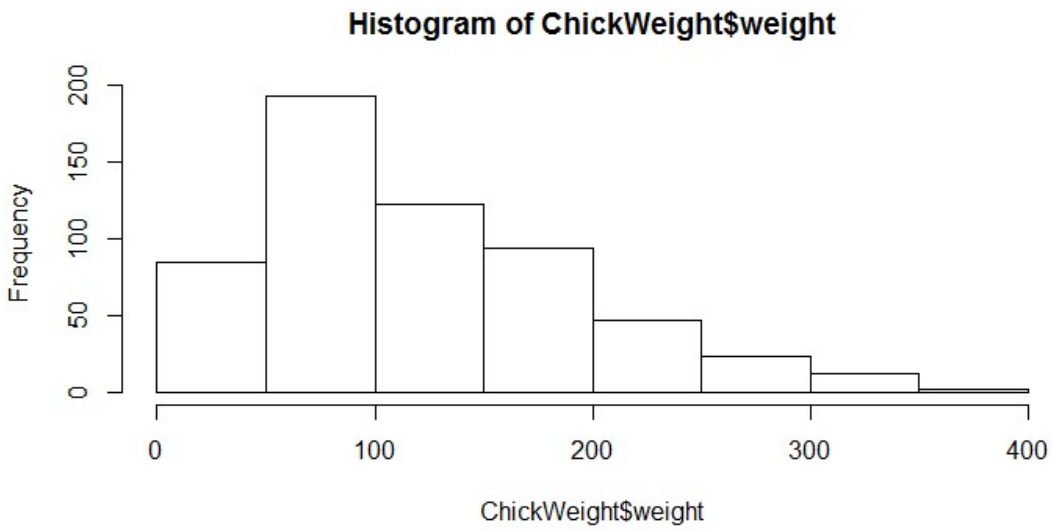
Using python:

```
import pandas as pd
data2=( C:\\Users\\jayanth\\Downloads\\Q9_b.csv")
data2.skew()
data2.kurtosis()
```

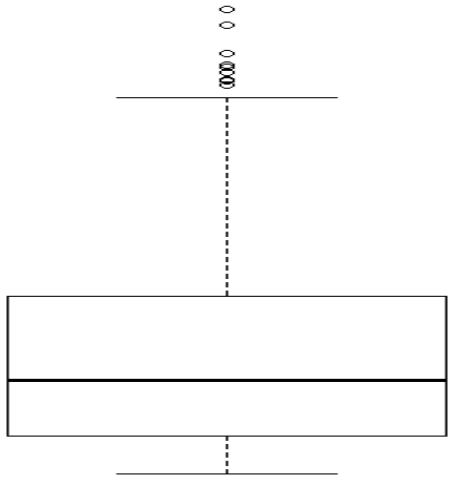
	SP	Weight
Skewness	1.6114	-0.6147
kurtosis	2.9773	0.9502

From the skewness of SP we came to know that the data of SP is positively skewed and from the skewness of weight we came to know that the data is moderately negatively skewed.

Q10) Draw inferences about the following box plot & histogram



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



From the 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?

Answer:

Sample size 'n' = 2000
Sample mean '$\bar{x}$ '= 200
Sample variance 's^2 '= 30

Therefore, class interval for mean is given by
$[\ \bar{x} - z_{\alpha/2}\ s/\sqrt{n}\ ,\ \bar{x} + z_{\alpha/2}\ s/\sqrt{n}\ ]$

The 94% class interval where $z_{\alpha/2}$ =1.89 is [198.73, 201.27]
The 96% class interval where $z_{\alpha/2}$ =2.33 is [198.43, 201.56]
The 98% class interval where $z_{\alpha/2}$ =2.06 is [198.62, 201.38]

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
st.norm.interval(alpha=0.96, loc=200, scale= 30/np.sqrt(2000))  #  94% CI
st.norm.interval(alpha=0.98, loc=200, scale= 30/np.sqrt(2000))  #  94% CI
```

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?

Answer:

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())
pritn(data.std())
```

The results are:
Mean = 42
Median = 40.5
Variance = 25.5294
Standard Deviation = 5.0527

On an average a student scores 42 marks.

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

A: The nature of skewness is perfectly symmetric that is it is zero skewed.

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

A: The nature of skewness is positively skewed.

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

A: The nature of skewness is negatively skewed.

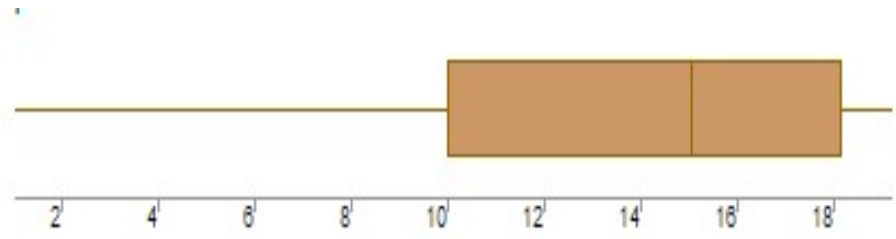
Q16) What does positive kurtosis value indicates for a data?

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

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

A: Negetive 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?

A: Here the distribution is skewed distribution.

What is nature of skewness of the data?

A: The nature of skewness is negatively skewed.

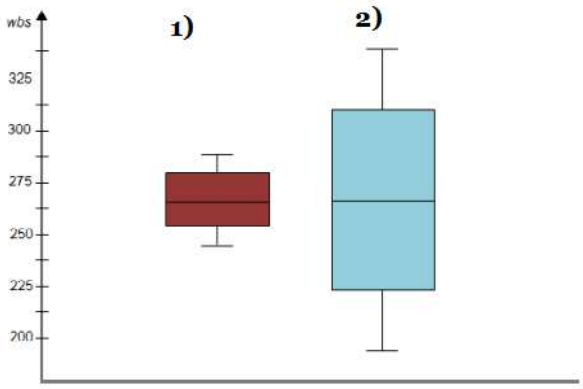
What will be the IQR of the data (approximately)?

A: IQR = Q3-Q1

= 18-10

IQR = 8

Q19) Comment on the below Box plot visualizations?



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

A: Here both the plots indicate that they follow normal distribution. The difference is Boxplot1 have lesser range when compared to Boxplot2.

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)

Answer:

Using Python:

```
import pandas as pd
import scipy.stats as st
data=pd.read_csv("F:/ExcelR/Python and Csv files/cars.csv")
```

```
Mean=data['MPG'].mean()
SD=data['MPG'].std()
# P(MPG>38)
1-(st.norm.cdf(38, loc=Mean, scale=SD))
# P(MPG<40)
st.norm.cdf(40, loc=Mean, scale=SD)
# P(20<MPG<50)
st.norm.cdf(50, loc=Mean, scale=SD)-st.norm.cdf(20, loc=Mean, scale=SD)
```

We have got

P (MPG>38) = 0.3475

P (MPG<40) = 0.7295

P (20<MPG<50) = 0.8988.

Q 21) Check whether the data follows normal distribution

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

Dataset: Cars.csv

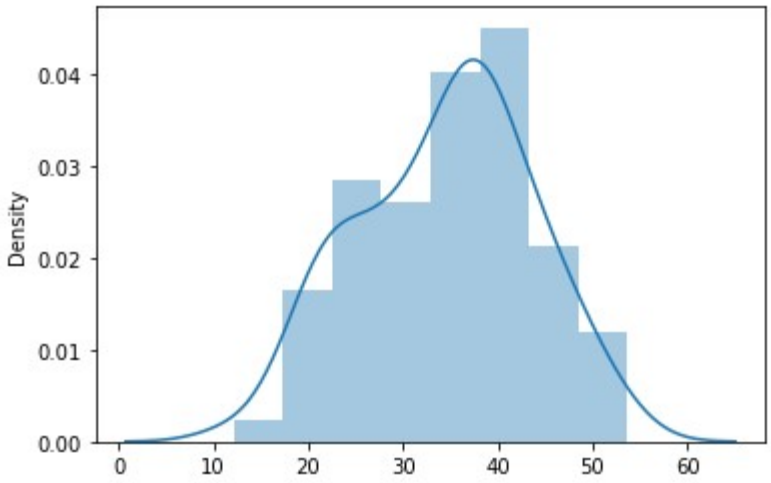
Answer:

Using python:

```
import pandas as pd
import seaborn as sns
import scipy.stats as st

data=pd.read_csv("F:/ExcelR/Python and Csv files/cars.csv")
sns.distplot(data['MPG'])
```

We have got



From the plot we can say that the data follows approximately 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

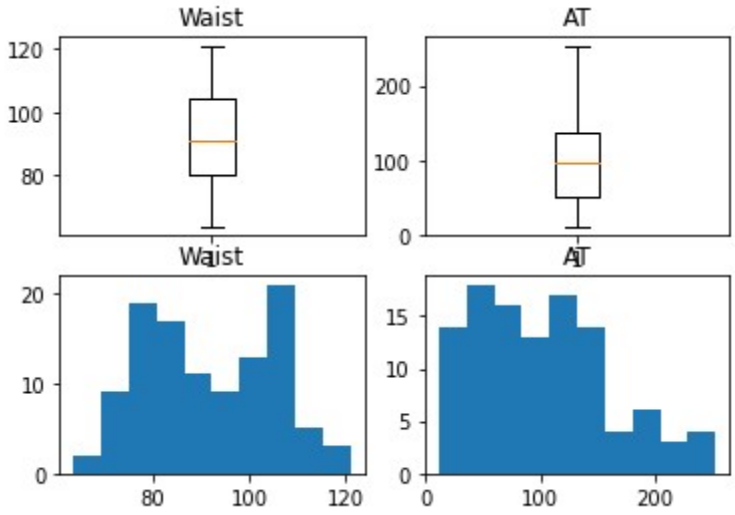
Answer:

Using python:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

data=pd.read_csv("C:/Users/santh/Downloads/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")
```



From the plot of waist we can say that the data is normally distributed and from the plot of AT we observe that it has long right tail. Therefore AT is positively skewed.

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

Answer:

Using Python:

```
import scipy.stats as st

print(st.norm.ppf(0.95))
print(st.norm.ppf(0.97))
print(st.norm.ppf(0.60))
```

We got
For 90% CI, Z score is 1.64485
For 97% CI, Z score is 1.88079
For 60% CI, Z score is 0.25334.

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

Answer:

Using Python:

```
import scipy.stats as st

print(st.t.ppf(0.975,24))
print(st.t.ppf(0.98,24))
print(st.t.ppf(0.995,24))
```

We got
For 95% CI, Z score is 2.0636
For 96% CI, Z score is 2.1715
For 99% CI, Z score is 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

Answer:

The hypothesis of interest is

Ho: The average light bulb lasts=270
H1: The average light bulb lasts <270
Let the level of significance be 5%
The test statistic is

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

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

t = -0.47.

t table(0.05,17) = 1.771

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