Assignment 1   : Basic Statistics 1

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?

<u>Answer:</u>

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:
S = {HHH, HHT, HTH, THH, TTH, THT, HTT, TTT}
Therefore, n(S) = 8
now we need to choose the event of 2 heads and a tail. We have 3 possible outcomes.
i.e.

A = {HHT,HTH,THH}

n(A) =3

Therefore,

P (getting 2 heads and a tail) = n(A) / n(S)
= 3/8
= 0.375

P (getting 2 heads and a tail) = 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**

Answer:

When two dice are rolled the possible sample space are: $6^2 = 36$.

{(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 sum is

a. Sum is equal to 1 :
If two dice are rolled then 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. Sum 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 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
=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 ,
{(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
=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))

CHIL D	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
Total			3.09

Therefore,

The expected number of candies = Σ(x*P(x)) = 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

<u>Answer:</u>

Python code :

```
import pandas as pd
data=pd.read_csv('Q7.csv')
data.mean() ,
data.median()
data['Points'].mode()
data['Score'].mode()
data['Weigh'].mode()
data.var()
data.std()
```

Output is :

	Points	Score	Weigh
Mean	3.5965	3.2172	17.8487
Median	3.695	3.325	17.710
Mode	3.92	3.44	18.90
Variance	0.2858	0.9573	3.1931
Standard deviation	0.5346	0.9784	1.7869

Inference:

From the data, we observe that mean, median and mode are not equal.Hence, we can say that our data is skewed .

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:

We have to find mean for the given data,

Python Code :

```
import numpy as np
weights_of_patients = [108, 110, 123, 134, 135, 145, 167, 187, 199]
np.mean(weight_of_persons)
```

Output : 145.33333333333334

Expected Value of the Weight of that patient = 145.3333

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

Answer :

Python Code :

```
data_1 = pd.read_csv('Q9_a.csv')
data_1.skew()
data_1.kurtosis()
```

Output :

	Speed	Distance
Skewness	-0.1175	0.8069
Kurtosis	-0.5090	0.4050

From the skewness of speed we can say that the data of speed is symmetric and from the distance we can say that the data is positively skewed.

SP and Weight(WT)
Use Q9_b.csv

Answer :

Python Code :

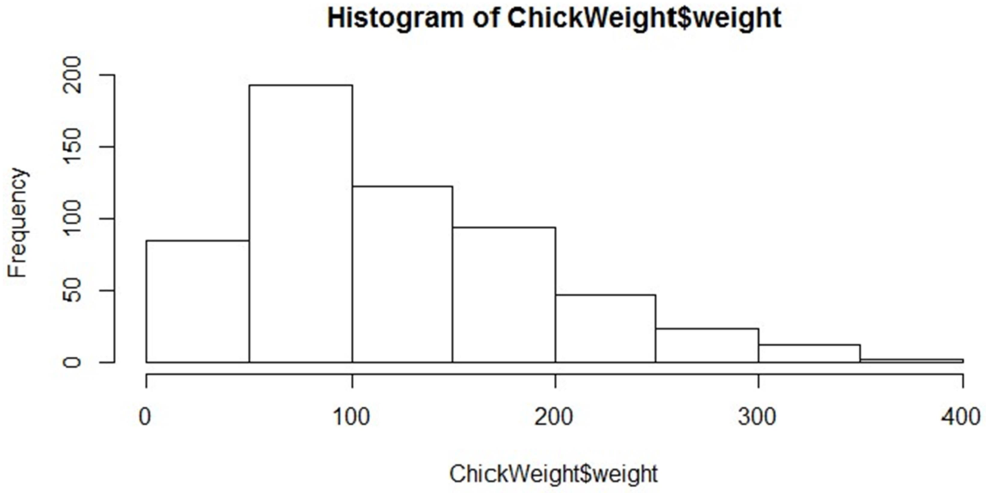
```
data_2 = pd.read_csv('Q9_b.csv')
data_2.skew()
data_2.kurtosis()
```

Output :

	SP	Weight
Skewness	1.6114	-0.6147
Kurtosis	2.9773	0.9502

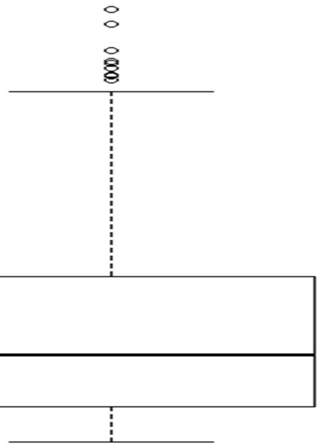
From the skewness of SP we can say that the data of SP is positively skewed and from the skewness of weight we can say that the data is negatively skewed.

Q10) Draw inferences about the following box plot & histogram



Answer :

From the above histogram we can say that the data is Positively symmetric.



Answer :

From the above plot, we can say 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?

<u>Answer :</u>

```
import numpy as np
from scipy import stats

stats.norm.interval(alpha=0.94, loc=200, scale= 30/np.sqrt(2000))    # 94% CI
stats.norm.interval(alpha=0.96, loc=200, scale= 30/np.sqrt(2000))    # 96% CI
stats.norm.interval(alpha=0.98, loc=200, scale= 30/np.sqrt(2000))    # 98% CI
```

Output :

C.I.	Lower limit	Upper Limit
94%	198.7383	201.2617
96%	198.6223	201.3777
98%	198.4394	201.5606

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:

Python code :

```
import numpy as np
x=[34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]
np.mean(x)
np.median(x)
np.var(x)
```

```
            np.std(x)
Output :
        Mean = 41
        Median = 40.5
        Variance =
        24.1111
        Standard Deviation = 4.9103
```

Student scores 41 marks as an average.

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

Answer :

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

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

Answer :

The nature of skewness is positively skewed.

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

Answer :

The nature of skewness is negatively skewed.

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

Answer :

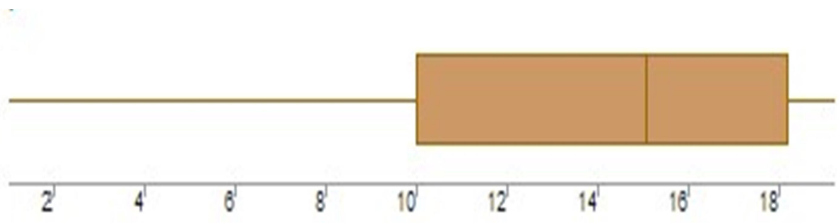
Positive tail indicates that we have heavy tails that is lot of data lies in tails.

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

Answer :

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?

Answer :

Here the distribution is skewed distribution.

What is nature of skewness of the data?

Answer :

The nature of skewness is negatively skewed.

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

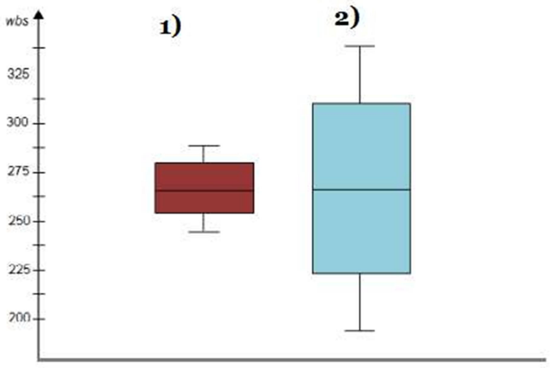
Answer :

IQR = Q3-Q1

IQR = 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.

Answer :

Here both the plots follows normal distribution. The difference is first boxplot have lesser range compared to second boxplot.

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 :

Python Code :

```
import pandas as pd
from scipy import stats
data=pd.read_csv("Cars (1).csv")
average=data['MPG'].mean()
std=data['MPG'].std()
1-(stats.norm.cdf(38, loc=average, scale=std))        # P(MPG>38)
stats.norm.cdf(40, loc=average, scale=std)            # P(MPG<40)
stats.norm.cdf(50, loc=average, scale=std)- stats.norm.cdf(20, loc=average,
scale=std)
                                                      # P(20<MPG<50)
```

Output :

a) P (MPG>38) = 0.3475
b) P (MPG<40) = 0.7295
c) P (20<MPG<50) = 0.8988

Q 21) Check whether the data follows normal distribution

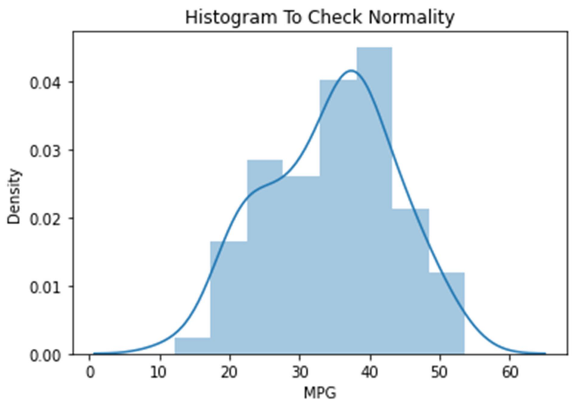
a). Check whether the MPG of Cars follows Normal Distribution Dataset: Cars.csv

Answer :

Python Code :

```
import pandas as pd
data=pd.read_csv("Cars (1).csv")

import matplotlib.pyplot as plt
import seaborn as sns
sns.distplot(data["MPG"])
plt.title('Histogram To Check Normality')
plt.show()
```



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

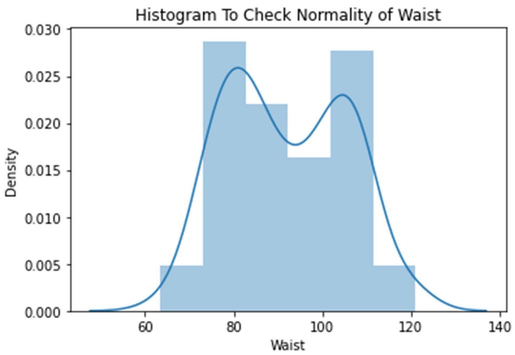
b). Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution Dataset: wc-at.csv

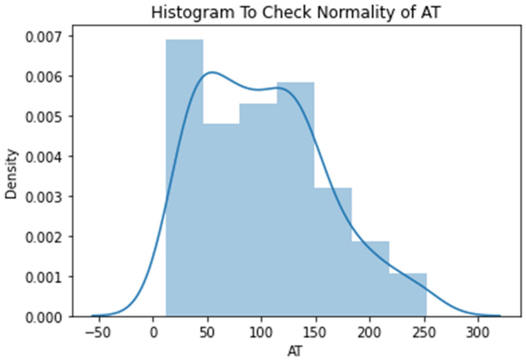
Answer:

Python Code :

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data_set = pd.read_csv('wc-at (1).csv')
sns.distplot(data_set['Waist'])
plt.title("Histogram To Check Normality of Waist")
plt.show( )
```

```
sns.distplot(data_set['AT'])
plt.title("Histogram To Check Normality of Waist")
plt.show( )
```





From the above plots we can say that, the data is normally distributed for 'Waist' and data is positively skewed for 'AT'.

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

Answer:

Python Code :

```
from scipy import stats
stats.norm.ppf(0.95)
stats.norm.ppf(0.97)
stats.norm.ppf(0.60)
```

Outpput :

Z scores (two tailed) for ,
90% confidence interval = 1.64485
94% confidence interval = 1.88079
60% confidence interval = 0.2533

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

Answer:

Python Code :

```
        from scipy import stats
        stats.t.ppf(0.975,df=24)
        stats.t.ppf(0.98,df=24)
        stats.t.ppf(0.995,df=24)
```

Output :

```
t scores for ,
              95% confidence interval =  2.0636
              96% confidence interval =  2.1715
              99% confidence interval =  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 are,

Ho: The average light bulb lasts 270

vs

H1: The average light bulb lasts < 270

level of significance : 5%

Now

Test statistic is

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

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

$$t = 0.47$$

by using t-table

los = 5% =0.05 and df = n-1 = 18-1 = 17

t (0.05,17) = 1.771

since , Calculated value of t is less than tabulat value ,hence we do not reject the null hypothesis.