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

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

**Ans:** let S be the sample space,

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

Then the probability of getting two heads and one tail is :

P = 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:**

Let S be the sample space

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), (5, 3), (5, 4), (5, 5), (5, 6),  
 (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)}.

The probability of getting

1. Sum is equal to 1 = 0/8 = 0
2. Sum is less than or equal to 4 = 6/36 = 1/6
3. Sum is divisible by 2 or 3 = 24/36 = 2/3

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:**

The probability that none of the balls drawn is blue is :

P = total number of red and green balls

Total number of balls

P = 5/7

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. 0.015 = n

Child B – probability of having 4 candies = 0.20

**Ans:**

The Formula for calculation of expected cadies is :

E[X]= Summation of (P(Xi​)) \* Xi

So calculation for the given data,

E[X] = 0.015 \* 1 + 0.20 \*4 + 0.65\*3 + 0.005\*5 + 0.01\*6 + 0.120\*2

E[X] = 3.125

So the Expected number of candies for a randomly selected child for the given data is 3.125

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**

**Ans:**

1. Mean:

Points mean =

3.9+3.9+3.9+3.8+3.08+3.15+2.76+3.21+3.69+3.92+3.92+3.92+3.07+3.07+3.07+2.93+3+3.23+4.08+4.93+4.22+3.7+2.76+2.76+3.15+3.73+3.08+4.08+4.43+3.77+4.22+3.62+3.54

32

Points mean = 3.5965

In python code it can be done like:

Data[‘Points’].mean() = 3.5965

1. Score mean:

Score mean =

2.62 + 2.875 + 2.32 + 3.215+ 3.44 + 3.46 + 3.57 + 3.19+3.15 + 3.44 + 3.44 +4.07 +3.73 + 3.73 + 3.78+ 5.25 +5.424 + 5.345 +2.2 + 1.615 +1.835 + 2.465 + 3.52 +3.435 +3.84 + 3.845 + 1.935+ 2.14 +1.513 +3.17 +2.77+ 3.57 + 2.78

32

Score mean = 3.217

In python code it can be done like:

Data[‘Score’].mean() = 3.5965

1. Weigh Mean:

Weigh Mean =

16.46 + 17.02 + 18.61 + 19.44+ 17.02 + 20.22 +15.84 + 20 + 22.9 + 18.3 + 18.9+ 17.4+ 17.6 +18 + 17.98 + 17.82 + 17.42+ 19.47 + 18.52 + 19.9+ 20.01+ 16.7 + 16.9 + 14.5 + 15.5 + 14.6 + 18.6

32

Weigh mean = 17.848

In python code it can be done like:

Data[‘Weigh].mean() = 3.5965

Q8) Calculate Expected Value for the problem below

1. 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 :

The expected value for the weights of patients chosen randomly is given as = (108 + 110 + 123 + 134 + 135 +145 + 167 + 187 + 199) / 9

E[X] = 145.33

**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

**Use Q9\_a.csv**

**Use Q9\_b.csv**

**Ans:**

First import stats module from scipy

Skewness of speed in python can be given as

df['speed'].skew()

skewness of speed = -0.117

Kurtosis of speed in python can be given as

kurtosis(df, axis=0, bias=True)

Kurtosis of speed = -0.577

Skewness of distance in python can be given as

df['distance'].skew()

skewness of distance = 0.806

Kurtosis of distance in python can be given as

kurtosis(df, axis=0, bias=True)

Kurtosis of distance = 0.248

SP and Weight(WT)

Skewness of SP in python can be given as

df['SP'].skew()

skewness of SP= 1.611

Kurtosis of SP in python can be given as

kurtosis(df2, axis=0, bias=True)

Kurtosis of SP= 2.723

Skewness of Weight in python can be given as

df['Weight'].skew()

skewness of Weight = -0.614

Kurtosis of Weight in python can be given as

kurtosis(df2, axis=0, bias=True)

Kurtosis of Weight= 0.819

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



Ans :

The data is highly positive skewed

Most of the data point are present between 0 to 200



Ans:

The data is skewed on the upper side and outliers are present on upper side

There are 7 outliers present on the upper 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?

Ans:

Sample size = 2000

Sample mean = 200 pounds

SD of the sample = 30

For 94 % confidence level the critical value is 1.88

So confidence interval in python is

stats.norm.interval(0.94,200,30/np.sqrt(2000))

confidence interval 198.73 - 201.76

For 98 % confidence level the critical value is 2.33

So confidence interval in python is

stats.norm.interval(0.98,200,30/np.sqrt(2000))

confidence interval 198.43 - 201.56

For 96 % confidence level the critical value is 2.05

So confidence interval in python is

stats.norm.interval(0.96,200,30/np.sqrt(2000))

confidence interval 198.62 - 201.37

**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:

Mean = 34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+42+45+49+56

18

Mean = 41

Median = (40+41)/2 = 40.5

Variance = 25.529

Standard Deviation = summation of [(Xi – Xmean)]/(n-1)

The standard deviation for given data = 5.025

Following conclusions can be done on the data:

* By calculating mean we can say that the data is centered about the value 41
* There are no outliers present in the data

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

Ans:

When mean and median of the data are equal then the skewness in the data can be considered as zero.

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

Ans: when mean is greater than the median then the distribution of the data is positively skewed and the outliers are present at the upper side of the box plot if plotted

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

Ans :

When the mean < median then there is high negative skewness present in the data and outliers are presents at the lower side of the box plot when plotted.

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

Ans:

The kurtosis value indicates the peak ness of the data means how high the data is spread on the y axis of the normal distribution in short that how much the data distribution is fat tailed at the ends

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

Ans:  
A negative value of kurtosis indicates that the distribution has flatter peak than the normal distribution and the data is plainly closed to the x variable.

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



What can we say about the distribution of the data?

What is nature of skewness of the data?

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

1. The distribution of the data is highly skewed to left means but there are no outliers present in the data as box plot doesn’t show it.
2. There is negative skewness in the data
3. The IQR of the data can be approximately given as 9

Q19) Comment on the below Boxplot visualizations?



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

Ans:

In the first boxplot the distribution seems to be near normal distribution bus has a very high value of kurtosis but in second boxplot the distribution is normal with medium value of kurtosis.in short the peak ness for first data set is more as compared to the second dataset

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)
  2. P(MPG<40)

c. P (20<MPG<50)

Ans:

1. P(MPG>38) = Prob\_1 = 1 - stats.norm.cdf(38,mean,SD)

P(MPG>38) = 34.75%

1. P(MPG<40) = prob\_2 = stats.norm.cdf(40,mean,SD)

P(MPG<40) = 72.93 %

c. P (20<MPG<50) = prob\_3 = (1 - stats.norm.cdf(20,mean,SD)) - (1-stats.norm.cdf(50,mean,SD))

P (20<MPG<50) = 89.88 %

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans:

The mean of the MPG column is 34.42 and median is 35.15 and hence the difference is very less so we can say that its close to normal distribution. Also the histogram plotted for the mpg column then is shows that the distribution for the column is nearly normal distribution and the density plot shows the same.

A graph with blue bars

Description automatically generated with medium confidence

A graph of a line

Description automatically generated with medium confidence

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

Dataset: wc-at.csv

Ans:

For Adipose tissue column:

Mean and median is 101.89 and 96.54 respectively so as we can see the difference between mean and median is somewhat less so we can say that the distribution for the adipose tissue column is somewhat normal distribution. Also the histogram and density plot shows the same.

A blue graph with numbers

Description automatically generated

A graph with a line

Description automatically generated

For the Waist Circumference column:

Mean and median is 90.901 and 90.8 respectively so as we can see the difference between mean and median is very less so we can say that the distribution for the adipose tissue column is similar to normal distribution. Also the histogram and density plot shows the same.

A graph with blue bars

Description automatically generated with medium confidenceA graph with a line

Description automatically generated

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

Ans:

The Z-scores are

For 90% confidence interval:

stats.norm.ppf(0.9)

1.28155

For 94% confidence interval:

stats.norm.ppf(0.9)

1.55477

For 60% confidence interval:

stats.norm.ppf(0.9)

0.25334

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

Ans:

The t scores can be given as

For 95% confidence interval:

stats.t.ppf(0.95,df = 25)

t-score = 1.7081

For 96% confidence interval:

stats.t.ppf(0.96,df = 25)

t-score = 1.8242

For 99% confidence interval:

stats.t.ppf(0.99,df = 25)

t-score = 2.4851

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

here we have

population mean = 270

sample size = 18

sample mean = 260

sample standard deviation = 90

as we don’t have population standard deviation we have to use t distribution method

so from given values we can calculate t-score and It can be given as

tscore = (sample mean -population mean) / (sample standard deviation / square root (sample size))

tscore = (260-270)/(90/square root(18))

tscore = 0.4714

by using r code pt(0.4714,17) we can have probability

as 32.16 %