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

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) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ordinal |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

1. The probability that two heads and one tail are obtained is 3/8 when three coins are tossed.

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
4. a) The probability that sum is equal to 1 is 0, when two dice are rolled.

b) The probability that sum is less than or equal to 4 is .

c) The probability that sum is divisible by 2 and 3 is i.e., .

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?

1. The total no. of balls the bag contains = 7

The total no. of required balls to be drawn,

total no. of balls – total no. of blue balls

7 – 2

5

Now, the probability that none of the blue balls are drawn

Therefore, the probability is 10/21.

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

1. The probabilities of count of candies for children are given.

The formula for expectation is .

Now, the expected number of candies for a randomly selected child,

.

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**

1. (Refer code)

For ‘**Points**’ column:

The Mean value of the data is 3.5965.

The Median value of the data is 3.695.

The Mode of the data are 3.07 and 3.92. Hence, it is bi-modal series.

The Variance of the data is 0.2858.

The Standard Deviation of the data is 0.5346.

The Range of the data is from 2.76 to 4.93.

For ‘**Score**’ column:

The Mean value of the data is 3.21725.

The Median value of the data is 3.325.

The Mode of the data is 3.44.

The Variance of the data is 0.9573.

The Standard Deviation of the data is 0.978457.

The Range of the data is from 1.513 to 5.424.

For ‘**Weigh**’ column:

The Mean value of the data is 17.8487.

The Median value of the data is 17.7100.

The Mode of the data are 17.02 and 18.90.

Hence, it is bi-modal series.

The Variance of the data is 3.1931.

The Standard Deviation of the data is 1.7869.

The Range of the data is from 14.5 to 22.9.

Inference:

As we observe, the data doesn’t consist any null values.

The data type of 1st column is object values and the data type of the other columns are float values.

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?

1. The weights of patients at a clinic are 108, 110, 123, 134, 135, 145, 167, 187, 199

The probability that one patient is chosen = 1/9

The formula for expectation is .

The Expected value of the weight of that patient is;

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

**Cars’ speed and distance**

**Use Q9\_a.csv**

1. **(Refer code)**

**Speed**

The Skewness of car’s speed is -0.1175

The Kurtosis of car’s speed is -0.5089

**Distance**

The Skewness of car’s distance is 0.8068

The Kurtosis of car’s distance is 0.40505

* Hence, the Speed and distance of car are positively skewed (mean>median) and are platykurtic (<3).

**SP and Weight (WT)**

**Use Q9\_b.csv**

1. **(Refer code)**

**SP**

The Skewness of car’s SP is 1.6114

The Kurtosis of car’s SP is 2.9773

* Hence, the SP of car is positively skewed (mean>median) and platykurtic (<3).

**Weight (WT)**

The Skewness of car’s Weight is -0.6147

The Kurtosis of car’s Weight is 0.9502

* Hence, the Weight of car is negatively skewed (mean<median) and platykurtic (<3).

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



* The histogram has more data on the left side and hence mean is greater than median. So, it is positively skewed.
* The boxplot has more distance from the median to the maximum (upper limit) than from the median to the minimum (lower limit). So, it is positively skewed.

**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?

1. (Refer Code)

Given,

The sample size of men, n = 2000

The population size, N = 3,000,000

Mean of the sample = 200 pounds

Standard Deviation of the sample = 30 pounds

The confidence interval estimates are as follows;

At 94% - (143.5761, 256.4238)

At 98% - (130.2095, 269.7904)

At 96% - (138.3875, 261.6124)

**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. (Refer code)

The Mean of the student scores – 41.0

The Median of the student scores – 40.5

The Variance of the student scores – 24.1111

The Standard Deviation of the student scores– 4.9103

1. What can we say about the student marks?
2. According to the student marks’ mean, we can say that he’s an average student. Since the mean is greater than median, we can say that the

data is positively skewed.

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

1. Skewness is zero as the data is perfectly symmetrical.

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

1. The skewness is positive, which means there is more data on the left side.

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

1. The skewness is negative, which means there is more data on the right side.

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

1. The positive kurtosis value indicates that it is Leptokurtic.

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

1. The negative kurtosis value indicates that it is Platykurtic.

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



What can we say about the distribution of the data?

1. Since the median is not equidistant from the minimum and the maximum, the boxplot is asymmetrical.

What is nature of skewness of the data?

1. The skewness of data is negative.

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

1. The IQR of the data is:

Q3 – Q1 = 18 – 10 = 8 (approximately).

**Q19)** Comment on the below Boxplot visualizations?



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

1. Inference:

* Since the medians of the two boxplots are same, there is no much difference between them.
* The Inter Quartile Range of Boxplot 1 is lesser than the Boxplot 2.
* The Boxplot 1 consists of smaller range of whiskers which indicates less scattered data than the Boxplot 2.
* The Boxplot 1 is positively skewed whereas the Boxplot 2 is symmetrical.

**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)
  3. P (20<MPG<50)

1. (Refer code)

The probability of MPG of Cars for the below cases are as follows:

1. P(MPG>38) = 0.34759
2. P(MPG<40) = 0.72935
3. P(20<MPG<50) = 0.89886

**Q 21)** Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

1. (Refer code)

The mean, median and mode of the data column MPG are not equal.

The Skewness of the data is not equal to zero i.e., MPG is asymmetric and the Kurtosis is less than 3 (platykurtic).

Hence, the MPG of Cars doesn’t follow Normal Distribution.

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

Dataset: wc-at.csv

1. (Refer code)

The mean, median and mode of the dataset are not equal.

The skewness is not equal to zero and kurtosis is less than 3.

Hence, the dataset doesn’t follow Normal Distribution.

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

* 1. (Refer code)

The Z score values for the following confidence intervals;

At 90% confidence interval – 1.6448

At 94% confidence interval – 1.8807

At 60% confidence interval – 0.8416

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

* 1. (Refer code)

The t-score values for the given confidence intervals are as follows;

At 95% confidence interval – 2.0638

At 96% confidence interval – 2.1715

At 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

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

* + 1. (Refer code)

The null hypothesis, H0: Average life of bulb >= 260

The alternative hypothesis, H1: Average life of bulb < 260

Mean, = 270

Sample size, n = 18

Sample mean, x = 260

Sample standard deviation, s = 90

Now, the test statistic is;

= = -0.4714

The statistic value, t is -0.4714

The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 0.32167

Since, p\_value is greater than significant level i.e.,

0.32167 > 0.05

Hence, we reject alternative hypothesis H1.

Therefore, we accept null hypothesis i.e.,

The average life of bulb is >= 260.