|  |  |
| --- | --- |
| 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) | Categorical |

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 | Interval |
| Hair Color | Nominal |
| **Socioeconomic Status** | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Interval |
| **Type of living accommodation** | Ordinal |
| **Level of Agreement** | Ordinal |
| **IQ(Intelligence Scale)** | Interval |
| Sales Figures | Interval |
| Blood Group | Nominal |
| **Time Of Day** | Nominal |
| **Time on a Clock with Hands** | Ordinal |
| Number of Children | Interval |
| Religious Preference | Nominal |
| **Barometer Pressure** | Ratio |
| **SAT Scores** | Ratio |
| **Years of Education** | Ratio |

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

Let H denote heads and T denote tails. Then,

Total Number of possible outcomes: {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT} = 8.

Number of expected outcomes: {HHT, HTH, THH} = 3.

Therefore, probability of two heads and one tail =

=

**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**

Let H denote heads and T denote tails. Then,

Total Number of possible outcomes: {(1,1), (1,2),… (6,6) } = 36.

Probability =

1. Equal to 1

Number of expected outcomes: { }=0.

Therefore, the probability that the sum is equal to one = 0

1. Less than or equal to 4

Number of expected outcomes: {(1,1), (1,2),(1,3), (2,1), (2,2), (3,1) }=6.

Therefore, the probability that the sum is less than or equal to 4 =

1. Sum is divisible by 2 and 3

Number of expected outcomes: {(1,5), (2,4),(3,3), (4,2), (5,1)}=5.

Therefore, the probability that the sum is divisible by 2 and 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?**

Total number of balls = 2 (red) + 3 (green) + 2 (blue) = 7 balls

Number of ways to choose 2 balls out of 7 without considering blue balls = C(7, 2) = 21 ways

Number of ways to choose 2 balls out of 5 (since we are excluding the blue balls) = C(5, 2) = 10 ways

So, the probability that none of the balls drawn is blue = Number of ways without blue balls / Total number of ways = 10 / 21

Therefore, the probability is approximately 0.4762 or 47.62%.

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.)** The formula expected number of candies for a randomly selected child:

Expected Value = Σ (X \* P(X))

where X is the value (number of candies in this case) and P(X) is the probability of that value occurring.

The probabilities for the number of candies each child has:

Child A: P(1) = 0.015

Child B: P(4) = 0.20

Child C: P(3) = 0.65

Child D: P(5) = 0.005

Child E: P(6) = 0.01

Child F: P(2) = 0.120

Therefore,

Expected Value = (1 \* 0.015) + (4 \* 0.20) + (3 \* 0.65) + (5 \* 0.005) + (6 \* 0.01) + (2 \* 0.120)

= 0.015 + 0.80 + 1.95 + 0.025 + 0.06 + 0.24

= 4.105

Hence, the expected number of candies for a randomly selected child is 4.105 candies.

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.)**

For the ‘Points’ column, the mean and median are approximately the same. So, we can conclude that this column approximately follows Normal distribution. The standard deviation (or variance) is also considerably low which is reflected in the range of 2.17 (Max=4.93 and Min=2.76). The mode is 3.07 and 3.92 with frequency 3.

The same is the case with the ‘Score’ column. The mean and median are approximately the same which suggests Normal distribution. However, the standard deviation is about 3 times more which is reflected in its range (3.91). This column is more deviated than the ‘Points’ column. The mode is 3.44 with a frequency of 3.

Along the same lines, the ‘Weigh’ column is also normally distributed. However, it has the highest standard deviation(1.786) and correspondingly the range (21.387) is also very high. The mode is 17.02 and 18.90 with frequency 2.

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?

The Expected Value of the Weight of that patient is: 145.34

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

1. **Cars speed and distance: Use Q9\_a.csv**

The skewness of the speed column is -0.1175, i.e., it is negatively skewed as evident from the histogram. The Kurtosis is -0.50899 which means its platy kurtosis.

When it comes to the dist column it is positively skewed(0.8069) as seen in its histogram and the Kurtosis is 0.4052 which means its Lefto Kurtosis.

The scatter plot and correlation between both variables reveal that they have a strong positive relationship. The correlation is equal to 0.8069.

1. **SP and Weight(WT): Use Q9\_b.csv**

The skewness of the SP column is 1.6114, i.e., it is positively skewed as evident from the histogram. The Kurtosis is 2.9773 which means its Lefto kurtosis. From the histogram, we could see most of the data fits in the interval 110-120.

However, when it comes to the WT column it is negatively skewed(-0.6147) as seen in its histogram and the Kurtosis is 0.9503 which means its platy Kurtosis. We find gaps in the histogram in the intervals 20-25 and 45-50.

The scatter plot and correlation between both variables suggest that they have a weak positive relationship. The correlation is equal to 0.1024.

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



Histogram

The histogram is about chickweight. About half of the data points lie under 100 pounds. The graph is positively skewed. Since it isn’t symmetrical, it does not follow normal distribution.

Boxplot

The boxplot also conveys the same patterns as above. The mean or 50th percentile lies closer to the minimum supporting almost half of the data points are less than 100 pounds. There are also outliers present in the data, around 8 which might fall in the interval 350-400.

**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?

94% CI => ( 143.58 , 256.42 )

96% CI => ( 130.21, 269.79 )

96% CI => ( 138.39, 261.61 )

**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?

Mean, median, variance, standard deviation = 41,40.5,25.53,5.05The average score of students is 41. We could find two outliers from the boxplot, 49 and 56. Those are students who scored exceptionally in the tests. Score 41 is the mode of the data as it was scored by four students followed by 42,40,39, and 38 that was scored by 2 students each.

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

When mean = median, the curve is normally distributed or symmetrical and skewness=0

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

Skewness is positive when mean> median.

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

Skewness is negative when median> mean.

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

The distribution has a taller central peak and fatter tails compared to normal distribution which has a kurtosis=0. (leptokurtic)

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

The distribution has a flatter central peak and lighter tails compared to normal distribution which has a kurtosis=0. (Platykurtic)

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



What can we say about the distribution of the data?

1. About 50% of the data point lie above 15

What is nature of skewness of the data?

1. The data is negatively skewed

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

1. IQR = Q3-Q1 =18-10=8 units.

Q19) Comment on the below Boxplot visualizations?



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

Both the data1(represented by Boxplot 1) and data2 (represented by Boxplot 2) have same mean and free from any outliers. However, the maximum and minimum of each dataset vary significantly, dataset1 has very little range compared to dataset2. We can conclude that the standard deviation or variance of dataset1 is very much lesser than the dataset2.

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)

a)0.41, b)0.75, and c)1.00

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

MPG of Cars is not normally distributed as p\_value>0.05

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

Dataset: wc-at.csv

AT and Waist column are normally distributed as p\_value<0.05

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

1. Z-score for a 90% confidence interval: 1.6448536269514722
2. Z-score for a 94% confidence interval: 1.8807936081512509
3. Z-score for a 60% confidence interval: 0.8416212335729143

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

1. The t-score for a 95% confidence interval: 2.0638985616280205
2. The t-score for a 96% confidence interval: 2.1715446760080677
3. The t-score for a 99% confidence interval: 2.796939504772804

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

If the CEO claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days: 0.32