|  |  |
| --- | --- |
| 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 | interval |
| Celsius Temperature | interval |
| Weight | ratio |
| Hair Color | nominal |
| Socioeconomic Status | nominal |
| Fahrenheit Temperature | interval |
| Height | ratio |
| Type of living accommodation | nominal |
| Level of Agreement | nominal |
| IQ(Intelligence Scale) | interval |
| Sales Figures | interval |
| Blood Group | nominal |
| Time Of Day | nominal |
| Time on a Clock with Hands | ordinal |
| Number of Children | nominal |
| Religious Preference | ordinal |
| Barometer Pressure | ratio |
| SAT Scores | interval |
| Years of Education | nominal |

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

The required probability is = 0.375

Q4) Two Dice are rolled, find the probability that sum is

1. Equal to 1

The required probability is 0

1. Less than or equal to 4

The required probability is = 0.166

1. Sum is divisible by 2 and 3

The required probability is 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?

Total number of balls are 7.

If two balls are drawn at random, the number of possible combinations such that none of the ball is blue will be 3.

Hence the probability is = 0.428

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

Expected number of candies = 1\*0.015+ 4\*0.20 +3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.120 = 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**

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?

Expected value = (108+110+123+134+135+145+167+187+199)/9

= 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**SP and Weight(WT)**

**Use Q9\_b.csv**

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



The above histogram represents the data about the weight of the chick and the frequency(number of chicks having that particular weight).The weight of the chick ranges from 0 to 400.The mode of the data seems to lie somewhere between 50 to 100.Also here the mean is greater than the mode, hence the data is positively skewed.



In the above box plot, the median divides the box into two unequal parts,the longer part of the box is above the median,hence the data represented by the above box plot is positively skewed or right skewed.Also some data points appear outside the 100th percentile,hence the data contains the 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?

Sample size = 2000

Population size = 3000000

Sample mean = 200

Std deviation = 30

94% CI = (198.74, 201.26)

98% CI = (198.44, 201.56)

96% CI = (198.622, 201.378)

**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.

Mean = 41.0

Median = 40.5

Variance = 24.111

Standard deviation = 4.9103

1. What can we say about the student marks?

From the data we can say that the mean of the marks is 41 which is slightly greater than the median.

Most of the students got marks between 41 and 42 and there are two outliers in the data which are 49 and 56.

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

The data is symmetric.

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

The data is positively skewed.

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

The data is negatively skewed.

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

The positive kurtosis value for a data indicates that the data is peaked or leptokurtic that is the graph of the data will have a peak. If the value is more positive, the data will contain more values that lie near the tail of the graph. This type of data may contain the outliers.

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

The negative kurtosis value for a data indicates that the data is platykurtic that is the graph of the data will be some what flat.

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



What can we say about the distribution of the data?

The data does not seem to be normally distributed. It seems to be skewed.

What is nature of skewness of the data?

The data is negatively skewed as the median is closer to the upper end and the whisker at the upper end is shorter.

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

The IQR will be approximately 18 – 10 = 8

Q19) Comment on the below Boxplot visualizations?



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

Data represented by both the box plots seem to have normal distribution.The median of both the box plots is same however the IQR varies.

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) = 41%
  2. P(MPG<40) = 75%

c. P (20<MPG<50) =85%

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

The MPG of cars does not tend to follow the 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

The Adipose tissue and Waist Circumference does not seem to follow the normal distribution.

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

Z score for 90% CI =1.644

Z score for 94% CI =1.880

Z score for 60% CI =0.8416

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

T score of 95% CI=2.063

T score of 96% CI=2.171

T score of 99% CI=2.796

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, mean of sample bulbs = 260

Population mean = 270

Std deviation = 90

Sample size = 18

Therefore, df = sample size - 1

= 18 – 1

= 17

T score = - 0.471

The required probability is 0.32167 that is around 32.16%