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 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 |

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

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

Ans:- Sample space = {HHH, HHT, HTH, HTT, TTT, TTH, THT, THH}

Count of sample space = 8

Probability = No of favorable outcome / total no of outcome

Favorable outcome = {HHT, HTH, THH}, count = 3

Probability = 3/8 = 0.375 or 37.5%

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

* Equal to 1
* Less than or equal to 4
* Sum is divisible by 2 and 3

Ans:- a) Sum cannot be equal to 1 because the minimum number on both the dices would be 1 and 1+1 = 2. So, probability P( sum equal to 1) = 0

b) There are 6 favorable outcomes & 36 total outcomes

P(X) = 6/36

P(X) = 1/6 i.e. 0.16

c) Favorable outcome = {(1,5), (2,4), (3,3), (4,2), (5,1), (6,6)}, count = 6

Probability = 6/36 = 1/6 = 0.166 = 16.6%

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:- balls except blue / total number of balls

P(X) = 5/7

P(X) = 0.71

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:-

Expected values when the probability is provided for a data point is given as

E(X) = ∑ X \* P(X)

|  |  |  |  |
| --- | --- | --- | --- |
| CHILD | Candies count | Probability | Expected Value |
| A | 1 | 0.015 | 1 \* 0.015 = 0.015 |
| B | 4 | 0.20 | 4 \* 0.20 = 0.80 |
| C | 3 | 0.65 | 3 \* 0.65 = 1.95 |
| D | 5 | 0.005 | 5 \* 0.005 = 0.025 |
| E | 6 | 0.01 | 6 \* 0.01 = 0.06 |
| F | 2 | 0.120 | 2 \* 0.120 = 0.240 |

To get the expected number of candies for a randomly selected child we sum up all the calculated expected values

0.015 + 0.80 + 1.95 + 0.025 + 0.06 + 0.240 = 3.09

the Expected number of candies for a randomly selected child is **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**

**Ans:-**

Q8) Calculate Expected Value for the problem below

* 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:-

Expected values when the probability is provided for a data point is given as

E(X) =∑ (X \* P(X))

Count of weight of patient at a clinic = 9

Probability of patient selected randomly = 1/9

E(X) = (108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199) \* 1/9

= 1308 \* 1/9

= 1308/9

= 145.33

the Expected Value of the Weight of that patient chosen randomly = **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**



**Ans:-**

* This tells us that: maximum number (i.e approximately 200) of Chicks has weight ranging from around 50 to 100.
* The mean of the weights of the chicks is in between 100 to 150.
* This is a positively skewed distribution, as the tail on the right side (higher values) is longer. This means that the majority of data points are concentrated on the left side of the distribution, while a few higher values are more spread out to the right.
* As this is positively skewed distribution, we can say that the mean is greater than the median.
* And there might be the presence of outliers on the right tail.



**Ans:-**

* Outliers are present at upper most extrem, and by counting them we can say that there are 7 outliers present.
* Also we can say that this is a positively skewed, as most of the data points are present left hand side.
* By using 5 number summary: here we have lower extrem, Q1, mean, Q2 and, upper extrem.
* **Upper Whisker:** The upper whisker shows the maximum value within the range of 1.5 times the IQR above the upper quartile (Q3). Any data points beyond this upper whisker are considered potential outliers.
* **Lower Whisker:** The lower whisker shows the minimum value within the range of 1.5 times the IQR below the lower quartile (Q1). Any data points beyond this lower whisker are considered potential outliers.
* The IQR indicates the spread of the middle 50% of the data and provides a measure of our data's variability.

**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:-**

**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**

* Find mean, median, variance, standard deviation.
* What can we say about the student marks?

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

**Ans:-**  When the mean and median of a dataset are equal, it generally indicates that the data is symmetrically distributed/ has Normal distribution and has a skewness of approximately zero.

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

**Ans:-** positive skewness

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

**Ans:-** Negative skewness

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

**Ans:-** This indicates that the dataset has fatter tails (more extreme values) than a normal distribution. A positive kurtosis suggests that the data has a peak that is taller and sharper than the normal distribution's peak, resulting in more values clustered around the mean and more extreme values in the tails.

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

**Ans:-** Negative kurtosis suggests that the dataset has lighter tails and is flatter compared to a normal distribution. In a negatively kurtotic distribution, values are more spread out, and there are fewer extreme values in the tails.

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



a) What can we say about the distribution of the data?

b) What is nature of skewness of the data?

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

Ans:- a) Negatively skewed distribution

b) left skewness/ negative skewness

c) Q1 : 10, Q2 : 15.2 (approx), Q3 : 18

Q19) Comment on the below Boxplot visualizations? 

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

Ans:- 1) Q2 for both the boxplots is same

2) variance or we can say spread of the data in boxplot 2 is more than boxplot 1.

3) both are normally distributed.

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

* P(MPG>38)
* P(MPG<40)

c. P (20<MPG<50)

Q 21) Check whether the data follows normal distribution

* Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

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

Dataset: wc-at.csv

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

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

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