|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Continuous |
| 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 | Interval |
| Celsius Temperature | Ratio |
| Weight | Ratio |
| Hair Color | Ordinal |
| Socioeconomic Status | Interval |
| Fahrenheit Temperature | Ratio |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of Agreement | Interval |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Ordinal |
| Time Of Day | Ratio |
| Time on a Clock with Hands | Ratio |
| Number of Children | Interval |
| Religious Preference | Ordinal |
| 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?

Solution: In tossing three coins, the sample space is given by

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

Total number of outcomes = 2^3 = 8

P(2H+1T) = Number of possible events/ Total number of events = 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

Solution : when two dice are rolled the below are the total number of possible outcomes

(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)

1. 0
2. <or=4 = 6/36 = 1/6
3. 6/36 = 1/6

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?

Total number of balls = 7

Number of ways 2 balls can be drawn = 7C2 = 21 ways

Number of ways 2 balls can be red = 2c2 = 2 ways

Number of ways 2 balls can be green = 3c2 = 3 ways

P(x) = Number of interested events / total number of events

P(none of the balls drawn is blue) = 2/21 + 3/21 – 2/21 = 3/21

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

Below is the calculation of Expected Value

|  |  |  |  |
| --- | --- | --- | --- |
| CHILD | Candies count | Probability | x p(x) |
| A | 1 | 0.015 | 0.015 |
| B | 4 | 0.2 | 0.8 |
| C | 3 | 0.65 | 1.95 |
| D | 5 | 0.005 | 0.025 |
| E | 6 | 0.01 | 0.06 |
| F | 2 | 0.12 | 0.24 |
|  |  | **Expected Value** | **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**

|  |  |  |  |
| --- | --- | --- | --- |
| **Measure** | **Points** | **Score** | **Weigh** |
| Mean | 3.60 | 3.22 | 17.85 |
| Median | 3.70 | 3.33 | 17.71 |
| Mode | 3.92 | 3.44 | 17.02 |
| Variance | 0.29 | 0.96 | 3.19 |
| Standard Deviation | 0.53 | 0.98 | 1.79 |
| Range | 2.17 | 3.91 | 8.40 |

Inferences –

1. The population has no outliers which is reflected by the minimal difference in the Measures of Central Tendency/ First Moment Business Decisions.
2. The Measures of Dispersion or the Second Moment Business Decisions – the Points have a standard deviation of 0.53 which is indicating that there is a lesser spread in the data, with 1.79 for the weight.
3. The range for Points and Score indicates a lesser dispersion in the dataset, however, the range for weigh is 8.40.

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 = 145.333.

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

**Cars speed and distance**

**Use Q9\_a.csv**

|  |  |  |
| --- | --- | --- |
|  | speed | dist |
| Skewness | -0.1175099 | 0.80689496 |
| Kurtosis | -0.5089944 | 0.40505258 |

Skewness represents that speed is negatively skewed implying that the mass of the distribution is concentrated on the right, where as the Dist is positively skewed.

Kurtosis results represent that Kurt value for speed is flatter than the normal distribution where as the peakedness for the dist is towards the normal distribution.

**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
|  | SP | WT |
| Skewness | 1.6114502 | -0.6147533 |
| Kurtosis | 2.97732894 | 0.95029149 |

Negative skewness for WT. Kurtosis for SP represents that the distribution is sharply peaked.

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



Inferences –

1. Histogram indicates a positively skewed distribution as the mass of the distribution is concentrated on the left.
2. The joint bars represent that the data is continuous



Inferences for Box Plot

1. This data set is not normally distributed as the Whisker is not representing 25% of the data, Lower quartile is less than 25%, and most of the data is above the median.
2. The outliers exist above the upper extreme.
3. Most of the data median to upper extreme.
4. The range of the data set will not give us a true picture of the values existing within the data set due to outliers.
5. IQR is high

**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?

Confidence interval is +/- of the below Confidence values

|  |  |  |  |
| --- | --- | --- | --- |
| Significance | 0.06 | 0.02 | 0.04 |
| Standard Deviation of Sample | 30 | 30 | 30 |
| sample | 2000 | 2000 | 2000 |
| sample mean | 200 | 200 | 200 |
| Confidence Interval for a Sample Mean | 1.26167471 | 1.5605616 | 1.37769665 |

**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 |
| Median | 40.5 |
| Standard Deviation | 5.05266383 |
| Variance | 25.5294118 |

1. What can we say about the student marks?

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

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

Positively Skewed Distribution

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

Negatively Skewed Distribution

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

Heavier tails than normal distribution

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

Lighter tails and a flatter peak

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



What can we say about the distribution of the data?

Percentage distribution of data in lower quartile is greater than the spread in upper quartile.

What is nature of skewness of the data?

Negatively skewed as the mass of the distribution is on the right.

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

Q3=35%

Q1=60%

Q3-Q1=IQR = 25%

Q19) Comment on the below Boxplot visualizations?



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

Answer: Boxplot 2 is more normally distributed than 1 as the Quartiles seem to be equal in 2. The upper quartile is greater than the lower Quartile in Boxplot 1.

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) = 33/81
  2. P(MPG<40)= 61/81

c. P (20<MPG<50) = 69/81

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Yes the data is normally distributed as the mean and median values are similar and the skewness and kutosis values are in agreement with the normal distribution levels.

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

Dataset: wc-at.csv

Yes normally distributed as the skewness and kutosis values are in agreement with the normal distribution levels.

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

|  |  |
| --- | --- |
| Confidence Interval | Z Scores |
| 90% | -1.6448536 |
| 94% | -1.8807936 |
| 60% | -0.8416212 |

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