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

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

Outcome = {(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)}

Probability = Number of favorable outcomes**/**total number of possible outcomes. Probability = 3/8.

Hence when 3 coins are tossed, the probability of two heads and one tail is 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

Outcome = {(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. Probability of sum equal to 1 = 0/36 = 0
2. Probability of sum less than or equal to 4 = 6/36 = 1/6 = 16.67
3. Probability of sum divisible by 2 & 3 = 6/36 = 1/6 = 16.67

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?

Red balls = R1, R2; Green Balls = G1, G2 & G3; Blue Balls = B1, B2

Outcome = {R1R2, R1G1, R1G2, R1G3, R2G1, R2G2, R2G3, G1G2, G1G3, G2G3, R1B1, R1B2, R2B1, R2B2, G1B1, G1B2, G2B1, G2B2, G3B1, G3B2, B1B2}

Probability that none drawn are blue = 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

Expected number = Total (Candies count \* probability) 🡪 row wise

Expected number of candies for a randomly selected child = 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**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data/Measures** | **Mean** | **Median** | **Mode** | **Variance** | **Standard Deviation** | **Range (Max – Min)** |
| **Points** | 3.60 | 3.70 | 3.07, 3.92 | 0.29 | 0.53 | 2.17 |
| **Score** | 3.22 | 3.32 | 3.44 | 0.96 | 0.98 | 3.91 |
| **Weigh** | 17.85 | 17.71 | 17.02, 18.90 | 3.19 | 1.79 | 8.4 |

Mean is also known as average of all the numbers in the data set.

Median is mid value in the ordered (ascending sorted) data set.

Mode is the number which occur most often in the data set. ‘Points’ & ‘weigh’ have bimodal values. (two observations are repeated most frequently)

Variance is the numerical values that describe the variability of the observations from its arithmetic mean and denoted by sigma-squared (σ2). Variance measure how far individuals in the group are spread out, in the set of data from the mean.

Standard deviation measure how much observations of the data set differs from its mean.

‘Points’ has less standard deviation compared to ‘score’ & ‘weigh’. It is recommended to have a lower standard deviation/variance for better insights & prediction quality of data.

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?

Probability of selecting a patient at random = 1/9

Expected Value = 1/9 (108+110+123+134+135+145+167+187+199) = 145.34

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

**Cars speed and distance**

**Use Q9\_a.csv**

|  |  |  |
| --- | --- | --- |
| **Car Data/Calculations** | **Skewness** | **Kurtosis** |
| **Speed** | -0.12 | -0.51 |
| **Distance** | 0.81 | 0.41 |

‘Car speed’ is negatively/left skewed. Skewness shows the direction of the outliers. Outliers are present on the left side of distribution. This means majority of the data distribution will be on the right side of the mean (if plotted using histogram). The skewness lies in the range of -0.5 to 0.5 which is said to be moderately symmetrical.

‘Car speed’ has negative kurtosis value and also less than 3 which means it tends to produce fewer and less extreme outliers than the normal distribution.

‘Car distance’ is positively/right skewed. Skewness shows the direction of the outliers. Outliers are present on the right side of distribution. This means majority of the data distribution will be on the left side of the mean (if plotted using histogram). The skewness lies in the range of -0.5 to 0.5 which is said to be moderately symmetrical.

‘Car distance’ has positive kurtosis value and also less than 3 which means it tends to produce fewer and less extreme outliers than the normal distribution.

**SP and Weight (WT)**

**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
| **Car Data/Calculations** | **Skewness** | **Kurtosis** |
| **Speed** | 1.61 | 2.98 |
| **Weight** | -0.62 | 0.95 |

‘Car speed’ is highly positive/right skewed. Skewness shows the direction of the outliers. Outliers are present on the right side of distribution. This means majority of the data distribution will be on the left side of the mean (if plotted using histogram). The skewness in the range of -0.5 to 0.5 which is said to be moderately symmetrical.

‘Car speed’ has positive kurtosis value and also close to 3 which means it tends to follow a normal distribution.

‘Car weight’ is negatively/left skewed. Skewness shows the direction of the outliers. Outliers are present on the left side of distribution. This means majority of the data distribution will be on the right side of the mean (if plotted using histogram).

‘Car weight’ has positive kurtosis value and also less than 3 which means it tends to produce fewer and less extreme outliers than the normal distribution.

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



The Histogram as per above graph is right/positively skewed. Skewness shows the direction of outliers, which means from right side of mean distribution there are outliers.

The Boxplot as per above graph, generally divided in 5 parts and also depicts that there are outliers (extreme values) on the top side (in circles). The box in boxplot describes about the Q1, Q2 & Q3 values. Q1 is the lower quartile, which has 25% of data. Q2 is the median or has 50% of data is below this. And Q3 is the upper quartile. The distribution is positively skewed as median is closer to the bottom of the box.

**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? (t-test should be used for calculation as more information about the data available is w.r.t sample only)

94% = [143.54, 256.46]

96% = [138.35, 261.65] ; 98% = [130.15, 269.85]

**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 = 41

Median = 40.5

Variance = 25.53

Standard Deviation = 5.05

The scores of most of the students are in range from 34 to 42. Few students have scored more than 42 marks in tests which are basically the outliers of data.

Q13) What is the nature of skewness when mean, median of data are equal? – Perfectly symmetrical, no skewness

Q14) What is the nature of skewness when mean > median? – Positively skewed or Right skewed

Q15) What is the nature of skewness when median > mean? – Negatively skewed or Left skewed

Q16) What does positive kurtosis value indicates for a data? – It indicates there are more outliers present in data

Q17) What does negative kurtosis value indicates for a data? – It indicates there are more outliers present in data

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



What can we say about the distribution of the data?

The distribution is not a normal distribution. More number/observations of data present below 50% of distribution.

What is nature of skewness of the data?

It is negatively skewed or left skewed.

What will be the IQR of the data (approximately)?   
IQR of data = Q3 – Q1 = 18 – 10 = 8 (approx.)

Q19) Comment on the below Boxplot visualizations?



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

The boxplot 1 & boxplot 2 have same mean of 262.5 (approx.).

The box plot1 & boxplot 2 are normally distributed (no skewness) as the median (Q2) is equidistant from Q1(lower quartile) & Q3(upper quartile) and there are no outliers in the data (which means kurtosis value would be close to 3).

The range of boxplot 1 is lower compared to range of boxplot 2 which means more values/data are considered in boxplot 2 compared to 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) – 34.75%
  2. P(MPG<40) – 72.95%

c. P (20<MPG<50) – 89.89%

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 follows a fairly symmetrical distribution.

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

Dataset: wc-at.csv

Adipose Tissue (AT) does not follows a normal distribution, its rightly skewed.

Waist Circumference (Waist) follows a fairly symmetrical distribution.

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

* For 90%: Z-Score: 1.64; For 94%: Z-Score: 1.88; For 60%: Z-Score: 0.84; (1 tailed)
* For 90%: Z-Score: 1.28; For 94%: Z-Score: 1.55; For 60%: Z-Score: 0.25 (2 tailed)

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

* For 95%: t-Score: 2.06; For 96%: t-Score: 2.17; For 96%: t-Score: 2.52; (1 tailed)
* For 95%: t-Score: 1.71; For 96%: t-Score: 1.83; For 96%: t-Score: 2.49; (2 tailed)

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

The t-value = -0.4714, df = 17 (n-1)

After substituting the values in python code, we get p-value = 0.3217

There is 32.17% probability that a randomly selected bulb would have an average life of less than 260 days.