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

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

HHH, HTT, THH, HTH, THT, HHT, TTH, TTT

P (two heads and one tail) = 3

P = 3/8

So the probability = 3/8 or 0.375

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

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

Ans a) Possible outcomes = (1,1)

Total outcomes = 6 \* 6 = 36

Hence, sum is equal to 1 = 1/36

Ans b) possible outcomes = (1,1), (1,2), (1,3), (2,1), (2,2), (3,1)

total outcomes = 6\*6 = 36

hence, sum is less than or equal to 4 = 6/36 or 1/6

Ans c) possible outcomes = (1, 5), (2, 4), (3, 3), (4, 2), (5, 1) (2, 6), (3, 5), (4, 4), (5, 3), (6, 2)

Total outcomes = 6\*6 = 36 Hence, Sum is divisible by 2 and 3 = 10/36 = 5/18

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 - total balls = 2(red) + 3(green) + 2(blue) = 7

Number of ways of drawing 2 balls out of 7  
7C2 ​ = (7\*6)/(2\*1)​ = 21

Event of drawing 2 balls, none of which is blue.  
Number of ways of drawing 2 balls out of (2 + 3) balls.  
5C2 ​ = (5\*4)/(2\*1)​ = 10  
probability of none of balls drawn is blue = 21/10​

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 – to calculate the expected number of candies for a randomly selected child

We have to multiply candies count for each child by their respective probability and then sum up the results.

Expected number of candies = (candies count for child A \* probability of child A) + (candies count for child B \* probability of child B) + (candies count for child C \* probability of child C) + (candies count for child D \* probability of child D) + (candies count for child E \* probability of child E) + (candies count for child F \* probability of child F)

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

Expected number of candies = 0.015 + 0.8 + 1.95 + 0.025 + 0.01 + 0.24

Expected number of candies = 3.09

Therefore, 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**

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?

Ans – expected value(mean) = 108+110+123+134+135+145+167+187+199/9

Expected value = 1313/9

Expected value = 145.89 pounds

**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 - The histograms peak has right skew and tail is on right. Mean > Median. We have outliers on the higher side.

Ans - The boxplot has outliers on the maximum side.



**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 – to calculate the confidence interval we have to use formula

Confidence interval = sample mean ± (critical value \* standard error)

Sample size (n) = 2000

Sample mean = 200 pounds

Standard deviation(s) = 30 pounds

First, lets calculate the standard error

Standard error = s/ √ n

Standard error = 30/√ 2000

Standard error = 0.67

Now we can calculate the confidence interval for different confidence levels:

1. 94% confidence intervals:

Since the confidence level is 94% we have to calculate the critical value corresponding to that level, we will used two tailed test so the alpha value is (1-0.94) / 2 = 0.03

The critical value for a sample size of 2000 and alpha = 0.03 is approximately 1.96.

Confidence interval = 200 ± (1.96 \* 0.671)

Confidence interval = 200 ± (1.316)

Confidence interval = (198.684, 201.316)

So the 94% confidence interval for the average weight of adult males in Mexico is approximately (198.684 pounds, 201.316 pounds)

1. 98% confidence intervals:

Since the confidence level is 98% we have to calculate the critical value corresponding to that level, we will used two tailed test so the alpha value is (1-0.98)/ 2 = 0.01

The critical value for a sample size of 2000 and alpha = 0.01 is

1 – (0.01/2) is approximately 2.576.

Confidence interval = 200 ± (2.576 \* 0.671)

Confidence interval = 200 ± (1.728)

Confidence interval = (198.272, 201.728)

So the 98% confidence interval for the average weight of adult males in Mexico is approximately (198.272 pounds, 201.728 pounds)

1. 96% confidence intervals:

Since the confidence level is 96% we have to calculate the critical value corresponding to that level, we will used two tailed test so the alpha value is (1-0.96)/ 2 = 0.02

The critical value for a sample size of 2000 and alpha = 0.02 is approximately 2.326.

Confidence interval = 200 ± (2.326 \* 0.671)

Confidence interval = 200 ± (1.561)

Confidence interval = (198.439, 201.561)

So the 96% confidence interval for the average weight of adult males in Mexico is approximately (198.439 pounds, 201.561 pounds)

**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?

Ans – 1)

Mean – to find the mean sum up all the scores and divide it by the total number of scores

Mean = 34 + 36 + 36 + 38 + 38 + 39 + 39 + 40 + 40 + 41 + 41 + 41 + 41 + 42 + 42 + 45 + 49 + 56 / 18

Mean = sum of scores/ total number of scores = 693/18 = 38.5

Median – to find the median first arrange all the scores in ascending order and determine the middle value. If there are two middle values calculate their average.

Arranging the scores in ascending order: 34, 36, 36, 38, 38, 39, 39, 40, 40, 41, 41, 41, 41, 42, 42, 45, 49, 56

The middle value is the 9th score, which is 40.

Median = 40

Variance – to find the variance calculate the average of the squared differences between each score and the mean

Subtract the mean from each score, square the result, and find the average of these squared differences

Step 1: Calculate the squared differences from the mean for each score:

(34 - 38.5)^2 = 26.25

(36 - 38.5)^2 = 6.25

(36 - 38.5)^2 = 6.25

(38 - 38.5)^2 = 0.25

(38 - 38.5)^2 = 0.25

(39 - 38.5)^2 = 0.25

(39 - 38.5)^2 = 0.25

(40 - 38.5)^2 = 2.25

(40 - 38.5)^2 = 2.25

(41 - 38.5)^2 = 6.25

(41 - 38.5)^2 = 6.25

(41 - 38.5)^2 = 6.25

(41 - 38.5)^2 = 6.25

(42 - 38.5)^2 = 12.25

(42 - 38.5)^2 = 12.25

(45 - 38.5)^2 = 42.25

(49 - 38.5)^2 = 110.25

(56 - 38.5)^2 = 306.25

Step 2: Calculate the average of the squared differences:

(26.25 + 6.25 + 6.25 + 0.25 + 0.25 + 0.25 + 0.25 + 2.25 + 2.25 + 6.25 + 6.25 + 6.25 + 6.25 + 12.25 + 12.25 + 42.25 + 110.25 + 306.25) / 18 = 17.8

Variance = 17.8

Standard deviation – is the square root of the variance, it measures the dispersion of the data points from the mean.

Standard deviation = √ variance = √17.8 = 4.216

Ans – 2) - Interpretation:

Based on the calculations, we can make the following observations about the student marks:

The mean score is 38.5, which represents the average performance of the student across all the tests.

The median score is 40, which indicates that half of the scores are below 40 and half are above 40. It suggests that the distribution of scores is slightly skewed to the left.

The variance of 17.8 implies that the scores are spread out or vary considerably from the mean. Higher variance indicates more significant dispersion.

The standard deviation of approximately 4.216 quantifies the average amount of deviation or dispersion of scores from the mean. It provides a measure of the overall consistency or variability in the student's performance.

In summary, the student's marks indicate an average performance with some variability or inconsistency across the tests. The majority of scores are clustered around the mean, but there are also some higher and lower scores, resulting in a relatively large standard deviation and variance.

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

ANS - When the mean and median are equal, it suggests that the distribution is symmetrical. In such cases, the skewness of the data is zero or close to zero.

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

ANS – when the mean is greater than the median, it indicates that the data distribution is positively skewed or right skewed.

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

ANS – when the median is greater than mean, it indicates that the data distribution is negatively skewed or left skewed.

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

ANS – A positive kurtosis value indicates that the data distribution has heavier tails and a more peaked or concentrated central region compared to a normal distribution.

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

ANS – a negative kurtosis value indicates that the data distribution has lighter tails and a flatter or more spread-out central region compared to a normal distribution. It suggest that there is fewer extreme value in the distribution than would be expected in a normal distribution.

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



What can we say about the distribution of the data?

ANS - The above Boxplot is not normally distributed the median is towards the higher value

What is nature of skewness of the data?

ANS - The data is skewed towards the left. The whisker range of minimum value is greater than the maximum.

What will be the IQR of the data (approximately)?   
Ans: The Inter Quantile Range = Q3 Upper quartile – Q1 Lower Quartile = 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.

Ans - First there are no outliers. Second both the box plot shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range.

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)

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

1. 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

Ans – first we calculate the t-score using the formula,

T = (sample mean – population mean) /( (sample standard deviation)/sqrt(sample size))

Given,

Sample mean = 260 days

Population mean = 270 days

Sample standard deviation = 90 days

Sample size (n) = 18 bulbs

T = (260-270) / ((90)/sqrt(18))

T = -10/(90/sqrt(18))

T = -1.609

Next, we need to find the degrees of freedom(df) for the t distribution, which is (n-1), in this case,

Df = n-1 = 18-1 = 17

Using the R code pt(tscore, df), we can calculate the probability Here’s the calculation

Probability = pt(-1.609, 17)

We can execute the following code in R code

Pt(tscore, df)

Then we will get probability = 0.063

Probability = 0.063