**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 |
| 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 | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Interval |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Interval |
| Years of Education | Interval |

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

Ans = The Probability of getting two heads & one tail are 3 times & if converted into percentage that will be approx 37.5%.

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

a) sum equal to 1 = 1/36

b) sum less than or equal to 4 = 1/9

c) sum divisible by 2 and 3 = 1/18

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- the probability of drawing two balls that are not blue is 4/7 or approximately 0.57.

**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 need to find the weighted average of the candies count.

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

= 3.14

The expected number of candies for a randomly selected child is 3.14 candies.

**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 - **For Points-**

Mean = 3.59, Median = 3.69, Mode = 3.07, Variance = 0.28 ,

Standard Deviation = 0.53, Range = 2.17

The mean, median, and mode are relatively close in value, indicating that the data is not heavily skewed.

The standard deviation is relatively low compared to the mean, indicating that there is not a wide range of values in the data set.

The range is also quite small, supporting the idea that there is not a wide range of values present.

The mode is a particularly useful measure in this data set, as there are several values that occur multiple times (3.9, 3.92, and 3.07).

**For Score-**

Mean = 3.21, Median = 3.32, Mode = 3.44, Variance = 0.95 ,

Standard Deviation = 0.97, Range = 3.91

The mean, median, and mode are relatively close in value, indicating that the data is not heavily skewed.

The standard deviation is relatively high compared to the mean, indicating that there is a wide range of values in the data set.

The range is also quite large, supporting the idea that there is a wide range of values present.

There are some outliers in the data set, as indicated by the relatively high values of 5.424 and 5.345 compared to the other values in the set.

**For Weigh-**

Mean = 17.84, Median = 17.71, Mode = 17.02, Variance = 3.19 ,

Standard Deviation = 1.7, Range = 8.4

The relatively high standard deviation suggests that there is a wide range of values in the data set.

The range is quite large, supporting the idea that there is a wide range of values present.

There are some outliers in the data set, as indicated by the relatively high values of 22.9 and 20.22 compared to the other values in the set. These outliers may be driving up the standard deviation and range values.

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 - To calculate the expected value of the weight of a patient chosen at random from the given weights, we need to first find the average weight of the patients.

We can find the mean .

Mean weight = (108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199) / 9

= 135.33 pounds

The expected value of the weight of a patient chosen at random is 135.33 pounds.

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans –** Car’s Speed-

Skewness: -0.11751, Kurtosis: -0.5089.

The negative value of the skewness indicates that the data is slightly skewed to the left, with a longer tail on the left side of the distribution. However, since the absolute value of the skewness is less than 1, we can conclude that the skewness is not very significant and the distribution is roughly symmetrical.

The negative value of the kurtosis indicates that it has fewer outliers and is less peaked than a normal distribution. However, the absolute value of the kurtosis is less than 1, which indicates that the kurtosis is not very significant .

Distance –

Skewness: 0.806895, Kurtosis: 0.405053.

A positive skewness value of 0.806895 indicates that the tail of the distribution is longer on the right-hand side, meaning that there are more data points with higher values. In other words, the distribution is skewed towards the right. A commonly used rule of thumb is that a skewness value between -0.5 and 0.5 indicates a roughly symmetric distribution, while a value greater than 0.5 indicates a moderately skewed distribution.

A kurtosis value of 0.405053 indicates that the distribution is platykurtic, meaning that it has a flatter peak compared to a normal distribution. This indicates that the data has fewer extreme values, or outliers, compared to a normal distribution. A value of 0 for kurtosis indicates a normal distribution, while a positive value indicates a more peaked shape.

**SP and Weight(WT)**

**Use Q9\_b.csv**

SP-

Skewness: 1.61145, Kurtosis: 2.977329.

A positive skewness value of 1.61145 indicates that the tail of the distribution is longer on the right-hand side and that there are more data points with higher values. The distribution is heavily skewed towards the right. A commonly used rule of thumb is that a skewness value greater than 1 or less than -1 indicates a highly skewed distribution.

A kurtosis value of 2.977329 indicates that the distribution is leptokurtic, meaning that it has a very peaked shape compared to a normal distribution. This indicates that the data has more extreme values, or outliers, compared to a normal distribution.

Weight (WT)-

Skewness: -0.614753, Kurtosis: 0.950291.

A negative skewness value of -0.614753 indicates that the tail of the distribution is longer on the left-hand side and that there are more data points with lower values. In other words, the distribution is skewed towards the left.

A kurtosis value of 0.950291 indicates that the distribution is mesokurtic, meaning that it has a similar shape compared to a normal distribution. This indicates that the data has a similar number of extreme values, or outliers, compared to a normal distribution.

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



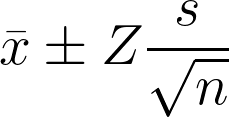
The data trails off to the right has a longer right tail, the shape is right-skewed . When data are right-skewed, or positive-skewed, many of the values are near the lower end of the range, and higher values are infrequent.

A right-skewed graph can be thought of as starting with a symmetric data set and pulling on its right tail. We can appear to pull the graph to the right by adding more values near the maximum of the distribution.



We can conclude by observing the boxplot, that the upper whisker is more wide than lower whisker as it is Positively Skewed. They are outliers on the upper side of the whiskers. The Median line is slightly lower in 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?**

Ans – The formula for Confidence Interval is 

For 94%, confidence interval is 198.7 to 201.3 pounds

Exp - Sample mean (x)= 200, Standard deviation (s) is 30

Sample size (n)= 2000, Z score = 1.880794 of 94% CL

For 98%, confidence interval is 198.4 to 201.6 pounds

Exp - Sample mean (x)= 200, Standard deviation (s) is 30

Sample size (n)= 2000, Z score = 2.326348 of 98% CL

For 96%, confidence interval is 198.6 to 201.4 pounds

Exp - Sample mean (x)= 200, Standard deviation (s) is 30

Sample size (n)= 2000, Z score = 2.053749 of 98% CL

**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) First, find the mean:

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

The mean is 41.

There are two values in Median

Median = (40+41)/2 = 40.5

To calculate variance need to find the deviations from the mean for each value.

34-41 = -7 , 36-41 = -5, 36-41 = -5 , 38-41 = -3 , 38-41 = -3, 39-41 = -2,

39-41 = -2, 40-41 = -1, 40-41 = -1, 41-41 = 0, 41-41 = 0 , 41-41 = 0 ,

41-41 = 0, 42-41 = 1, 42-41 = 1, 45-41 = 4, 49-41 = 8, 56-41 = 15

Now, we square each deviation: (-7)^2 = 49 , (-5)^2 = 25, (-5)^2 = 25,

(-3)^2 = 9 , (-3)^2 = 9, (-2)^2 = 4, (-2)^2 = 4 , (-1)^2 = 1 , (-1)^2 = 1, 0^2 = 0 0^2 = 0, 0^2 = 0 , 0^2 = 0 , 1^2 = 1, 1^2 = 1, 4^2 = 16, 8^2 = 64,

15^2 = 225

Then, we add up the squared deviations: 49 + 25 + 25 + 9 + 9 + 4 + 4 + 1 + 1 + 0 + 0 + 0 + 0 + 1 + 1 + 16 + 64 + 225 = 433

divide by the number of values minus one (n-1): Variance = 433 / 17 Variance = 25.47

The variance is 25.47.

Standard deviation = square root of 25.47 Standard deviation = 5.05

The standard deviation = 5.05.

1. What can we say about the student marks?

The student has taken multiple tests, as there are 18 scores listed.

Student's scores range from 34 to 56, which is a range of 22 points.

This will be a normal Distribution as mean=median=mode.

Mean is 41, median is 40.5, mode is 41, standard deviation is 5.05

Based on info, student has performed consistently well.

Top of Form

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

**Ans –** when mean, median data are equal then it is Symmetric in nature, he distribution is symmetric and has zero skewness.

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

**Ans** – when mean > median the nature of skewness is positively skewed or right-skewed.

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

**Ans**- when median>mean the nature of skewness is negatively skewed or left-skewed

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

**Ans -** Positive kurtosis indicates that a data set has a peaked or heavy-tailed distribution, meaning that the data has a higher proportion of extreme values in the tails of the distribution.

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

**Ans-** Negative kurtosis indicates the data set has a has a "flat" or "light-tailed" distribution, meaning that the data has a lower proportion of values in the tails of the distribution.

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



1. **What can we say about the distribution of the data?**

Ans – The distribution of data is asymmetric in nature.

**B) What is nature of skewness of the data?**

Ans- The median is closer to the top of the box, and the whisker is shorter on the upper end of the box, the distribution is Negatively skewed (skewed left).

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

Ans - The range of IQR consist from 10 to 18

Q1 is 12, Q2 is 15, Q3 is 18

IQR = Q3-Q1= 18-12= 6

The IQR of the data is 6

**Q19) Comment on the below Boxplot visualizations?**



Ans – No1. Boxplot min score is between 235 -240 highest score is 285-290

Lower Quartile is at 250 & Upper Quartile is at 275

No2 Boxplot min score is below 200 & highest score is above 325

Lower Quartile is at 225 & Upper Quartile is at 300

No1 Boxplot is smaller than No2 Boxplot but the median of both boxplot are at the same score.

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

Inference we draw from plot is No1 Boxplot IQR range is much Lower than No2 Boxplot IQR range.

They both have Symmetric distribution in nature.

**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

Ans-

* 1. P(MPG>38) - TRUE
  2. P(MPG<40) - FALSE

c. P (20<MPG<50)- FALSE

**Q 21) Check whether the data follows normal distribution**

1. **Check whether the MPG of Cars follows Normal Distribution**

Dataset: Cars.csv

Ans- This is the graph of MPG

It doesn’t follow a 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

Ans – Adipose Tissue (AT) graph

It doesn’t follow a normal Distribution.

Waist Circumference (WT) graph

It follows a Normal Distribution.

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

**Ans-**  To calculate 90% confidence interval , formula is

AL = 1+CL/2 = 1+0.90/2 = 0.95

According to Z score table the AL of 0.95 is 1.644854

So, Z score 90% CI = 1.644854

To calculate 94% confidence interval , formula is

AL = 1+CL/2 = 1+0.94/2 = 0.97

According to Z score table the AL of 0.97 is 1.880794

So, Z score 94% CI = 1.880794

To calculate 60% confidence interval , formula is

AL = 1+CL/2 = 1+0.60/2 = 0.53

According to Z score table the AL of 0.53 is 0.841621

So, Z score 60% CI = 0.841621

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

**Ans** - Calculate the t score for a 95% confidence interval with a sample size of 25

t\_score , qt(0.975, 24)

2.063899

Calculate the t score for a 96% confidence interval with a sample size of 25

t\_score , qt(0.98, 24)

2.171012

Calculate the t score for a 99% confidence interval with a sample size of 25

t\_score <- qt(0.995, 24)

2.79694

**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- Before using rcode, we have to find t value, we can perform a t-test using the null hypothesis that the population mean is equal to 270 days.

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

t =(260-270)/(90/(18)) = -1.8974

Now using rcode , where tscore= -1.8974, degrees of freedom (sample\_size – 1)= 18-1=17

pt(-1.8974,17) = 0.03744685

Since this probability is less than the significance level of 0.05, we would reject the null hypothesis and conclude that the average life of light bulbs is significantly less than 270 days.