Q1) Identify the Data type for the Following:

|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | int |
| Results of rolling a dice | int |
| Weight of a person | float |
| Weight of Gold | float |
| Distance between two places | Float |
| Length of a leaf | float |
| Dog's weight | Int |
| Blue Color | Boolean |
| Number of kids | Int |
| Number of tickets in Indian railways | Int |
| Number of times married | Int |
| Gender (Male or Female) | Boolean |

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

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

**Ans:** As three coins are tossed so outcome will be 9

the probability of two heads and one tails is

* 1st coin is head, 2nd coin is head, 3rd coin is tails
* 1st coin is tails, 2nd coin is head, 3rd coin is head
* 1st coin is head, 2nd coin is tail, 3rd coin is head

total outcome is 9

desired outcome is 3

so, probability is 3/9= 1/3

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

**ANS:** When rolling two dice, the sample space consists of all possible combinations of outcomes, which is 36 in total (6 outcomes for the first die, and 6 outcomes for the second die).

a) The sum of two dice cannot be 1, as the minimum value that can be obtained is 2 (by getting a 1 on one die and a 1 on the other). Therefore, the probability of getting a sum of 1 is 0.

b) To get a sum less than or equal to 4, the possible outcomes are (1,1), (1,2), (2,1), and (2,2), which have a total of 4 possibilities. Therefore, the probability of getting a sum less than or equal to 4 is 4/36 or 1/9.

c) To get a sum that is divisible by both 2 and 3, we need to find the common multiples of 2 and 3, which are the multiples of 6. The possible outcomes are (1,5), (2,4), (3,3), (4,2), (5,1), and (6,6), which have a total of 6 possibilities. Therefore, the probability of getting a sum that is divisible by both 2 and 3 is 6/36 or 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?

**ANS:** Total number of balls =7 balls

N (Event (2 balls are drawn randomly from bag) = 7! / 2! \* 5! = (7654321) / (21) \* (54321)

N (Event (2 balls are drawn randomly from bag) = (76)/ (21) = 21

If none of them drawn 2 balls are blue = 7 – 2 = 5 N (Event (None of the balls drawn is blue) = 5! / 2! \* 3! = (54) / (2\*1) = 10 P (None of the balls drawn is blue) = N (Event (None of the balls drawn is blue) / N (Event (2 balls are drawn randomly from bag) = 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

**Ans**: Expected number of candies for a randomly selected child

= 1 \* 0.015 + 4\*0.20 + 3 \*0.65 + 5\*0.005 + 6 \*0.01 + 2 \* 0.12

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 3.090

= 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:**

* + **Mean for** Points = 3.59, Score = 3.21, Weigh = 17.84,
  + **Median for** Points = 3.69, Score = 3.32, Weigh = 17.71
  + **Mode for** Points = 3.07, Score = 3.44, Weigh = 17.02
  + **Variance for** Points = 0.28, Score = 0.95, Weigh = 3.19
  + **Standard Deviation for** Points = 0.53, Score = 0.97, Weigh = 1.78
  + **Range [Min-Max] for** Points [3.59 – 4.93], Score [3.21 – 5.42] and Weigh [17.84 – 22.9]

**Based on the given information, we can make the following inferences:**

1. The mean score for Points is 3.59, for Score it is 3.21, and for Weigh it is 17.84.

2. The median score for Points is 3.69, for Score it is 3.32, and for Weigh it is 17.71.

3. The mode score for Points is 3.07, for Score it is 3.44, and for Weigh it is 17.02.

4. The variance for Points is relatively low at 0.28, indicating that the data points are clustered close to the mean. In contrast, the variance for Score is much higher at 0.95, suggesting greater variability in the data points. Similarly, the variance for Weigh is 3.19.

5. The standard deviation for Points is 0.53, for Score it is 0.97, and for Weigh it is 1.78. This indicates that the data points for Weigh have a larger spread than Points and Score.

6. The range for Points is from 3.59 to 4.93, for Score it is from 3.21 to 5.42, and for Weigh it is from 17.84 to 22.9.

7. The median score for Points is higher than the mean score, which indicates that the distribution of data points for Points is negatively skewed. In contrast, the median score for Score is lower than the mean score, indicating a positively skewed distribution.

**Overall, the data suggests that Points has a relatively narrow range of values with low variance and negative skewness, whereas Score and Weigh have wider ranges of values with higher variance and different skewness.**

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 = Sum (X \* Probability of X) = (1/9) (108) + (1/9) (110) + (1/9) (123) + (1/9) (134) + (1/9) (145) + (1/9) (167) + (1/9) (187) + (1/9) (199) = 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans:**

For Cars Speed Skewness value= -0.12 and Kurtosis value= 0.81

For Cars Distance Skewness value = 0.81 and Kurtosis value = 0.41 for Cars Distance

**Based on the given information, we can make the following inferences:**

1. The skewness value for Cars Speed is -0.12, which indicates a slightly negatively skewed distribution. This means that the majority of the data points are clustered towards higher speeds, with a relatively small number of lower-speed outliers.
2. The skewness value for Cars Distance is 0.81, which indicates a moderately positively skewed distribution. This suggests that the majority of the data points are clustered towards lower distances, with a relatively small number of higher-distance outliers.
3. The kurtosis value for Cars Speed is 0.81, which suggests a distribution that is more peaked than a normal distribution. This indicates that the data for Cars Speed is concentrated around the mean and has relatively fewer outliers.
4. The kurtosis value for Cars Distance is 0.41, which suggests a distribution that is flatter than a normal distribution. This means that the data for Cars Distance is more spread out than Cars Speed and has more outliers.
5. Overall, the data suggests that Cars Speed has a relatively narrow range of values, with a moderately negatively skewed distribution that is more peaked than a normal distribution. In contrast, Cars Distance has a wider range of values, with a moderately positively skewed distribution that is flatter than a normal distribution.

**SP and Weight (WT) Use Q9\_b.csv**

**Ans:**

For SP Skewness = 1.61 kurtosis = 0.95

For WT Skewness = 1.61 Kurtosis = 0.95

1. The given information suggests that both SP and WT have the same values for skewness and kurtosis. Skewness refers to the measure of symmetry or asymmetry of the distribution, while kurtosis measures the peak or flatness of the distribution.
2. A skewness value of 1.61 indicates that the distribution of data for both SP and WT is highly positively skewed. This means that the majority of the data points are clustered towards lower values, with a long tail of higher values.
3. A kurtosis value of 0.95 suggests that the distribution of data for both SP and WT is closer to a normal distribution than some other types of distributions. This indicates that the data has a moderate amount of peak, and is not overly flat or peaked.
4. Overall, the data for both SP and WT suggests that the majority of the data points are clustered towards lower values, with a long tail of higher values. However, the data is still relatively close to a normal distribution, indicating a moderate amount of peak.

**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 intervals, we need to use the formula:

CI = X̄ ± z\*(σ/√n)

where X̄ is the sample mean, σ is the population standard deviation (which we don't know), n is the sample size, and z is the critical value from the standard normal distribution corresponding to the desired level of confidence.

For 94% confidence interval, the z-value is 1.88 (obtained from the standard normal distribution table).

For 98% confidence interval, the z-value is 2.33.

For 96% confidence interval, the z-value is 1.96.

Now, we can calculate the confidence intervals as follows:

94% confidence interval:

CI = 200 ± 1.88\*(30/√2000)

CI = 200 ± 2.19

CI = [197.81, 202.19]

98% confidence interval:

CI = 200 ± 2.33\*(30/√2000)

CI = 200 ± 2.71

CI = [197.29, 202.71]

96% confidence interval:

CI = 200 ± 1.96\*(30/√2000)

CI = 200 ± 2.28

CI = [197.72, 202.28]

Therefore, we can say with 94% confidence that the average weight of an adult male in Mexico is between 197.81 and 202.19 pounds, with 98% confidence that it is between 197.29 and 202.71 pounds, and with 96% confidence that it is between 197.72 and 202.28 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.

**Ans:** Mean =41, Median =40.5, Variance =25.52 and Standard Deviation =5.05

1. What can we say about the student marks?

**Ans:** We don’t have outliers and the data is slightly skewed towards right because mean is greater than median.

Based on the mean, median, variance, and standard deviation, we can say that the student's scores are clustered around the mean score of 41, with a small variation of about 9.62. The median score of 40.5 also supports this observation. The highest score is 56, which is not too far from the mean, and the lowest score is 34, which is also not too far from the mean.

Overall, we can say that the student's marks are relatively consistent and do not vary greatly from the mean, indicating that the student is performing consistently well. However, we cannot make any further conclusions about the student's performance without more information about the grading criteria and how the scores compare to those of other students.

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

**Ans:** No skewness is present we have a perfect symmetrical distribution

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

**Ans:** Skewness and tail is towards Right

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

**Ans:** Skewness and tail is towards left

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

**Ans:** Positive kurtosis means the curve is more peaked and it is Leptokurtic

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

**Ans:** Negative Kurtosis means the curve will be flatter and broader

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 a skewed towards left. The whisker range of minimum value is greater than 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)
  3. P (20<MPG<50)

ANS:

* 1. P(MPG>38) = 0.348
  2. P(MPG<40) = 0.729
  3. p(MPG>20) = 0.943
  4. P(MPG<50) = 0.956
  5. P(20<MPG<50) = 0.013000000000000012

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans:** a.) MPG of cars follows 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) and Waist does not follow Normal Distribution

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

**ANS:**

To calculate the z-scores for different confidence intervals, we need to use the standard normal distribution with a mean of 0 and a standard deviation of 1.

Here are the z-scores for the given confidence intervals:

1. For a 90% confidence interval, we need to use a significance level of α = 0.10/2 = 0.05 for a two-tailed test. Using a standard normal distribution table or calculator, the z-score for a 90% confidence interval is 1.645.
2. For a 94% confidence interval, we need to use a significance level of α = 0.06/2 = 0.03 for a two-tailed test. Using a standard normal distribution table or calculator, the z-score for a 94% confidence interval is 1.880.
3. For a 60% confidence interval, we need to use a significance level of α = 0.40/2 = 0.20 for a two-tailed test. Using a standard normal distribution table or calculator, the z-score for a 60% confidence interval is 0.253.

Note that as the confidence level decreases, the z-score also decreases, indicating that we need a narrower interval to capture the true population mean with lower confidence.

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

ANS:

To calculate the t-scores for different confidence intervals for a sample size of 25, we need to use the t-distribution with degrees of freedom equal to n - 1 = 25 - 1 = 24. The t-distribution is used when the population standard deviation is unknown and we estimate it using the sample standard deviation.

Here are the t-scores for the given confidence intervals:

1. For a 95% confidence interval, we need to use a significance level of α = 0.05/2 = 0.025 for a two-tailed test. The degrees of freedom are 24. Using a t-distribution table or calculator, the t-score for a 95% confidence interval is 2.064.
2. For a 96% confidence interval, we need to use a significance level of α = 0.04/2 = 0.02 for a two-tailed test. The degrees of freedom are 24. Using a t-distribution table or calculator, the t-score for a 96% confidence interval is 2.171.
3. For a 99% confidence interval, we need to use a significance level of α = 0.01/2 = 0.005 for a two-tailed test. The degrees of freedom are 24. Using a t-distribution table or calculator, the t-score for a 99% confidence interval is 2.797.

Note that as the confidence level increases, the t-score also increases, indicating that we need a wider interval to capture the true population mean with higher confidence.

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:

To calculate the probability that 18 randomly selected bulbs would have an average life of no more than 260 days, we can use a one-sample t-test. The null hypothesis is that the population mean lifespan of the bulbs is equal to 270 days, and the alternative hypothesis is that it is less than 270 days.

We can calculate the t-score using the formula:

t = (x̄ - μ) / (s / √n)

where x̄ is the sample mean (260), μ is the population mean (270), s is the sample standard deviation (90), and n is the sample size (18).

t = (260 - 270) / (90 / √18)

t = -1.697

The degrees of freedom (df) for this test is n - 1 = 18 - 1 = 17. Using the R code pt(tscore, df), we can find the probability of getting a t-score of -1.697 or lower with 17 degrees of freedom:

pt (-1.697, 17)

This gives us a probability of 0.057, or 5.7%. Therefore, if the CEO's claim were true, there is a 5.7% chance that 18 randomly selected bulbs would have an average life of no more than 260 days. This is not a very low probability, so it is possible that the CEO's claim is not accurate based on this sample. However, we would need more information, such as the level of significance or the desired margin of error, to make a more definitive conclusion.