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

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

**Answer:**

When three coins are tossed, there are 2^3 = 8 possible outcomes, since each coin can land heads or tails.

Let H denote heads and T denote tails. The possible outcomes are:

Out of these 8 outcomes, there are 3 outcomes where two heads and one tail are obtained, namely:

* H H T
* H T H
* T H H

Therefore, the probability of getting two heads and one tail is 3/8 = 37.5%.

Top of Form

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

**Answer:**

* 1. The probability that the sum is equal to 1 is zero 0 as the minimum sum is 2.
  2. Total number of outcomes from rolling two dice is 6^2 = 36. Total number of outcomes where sum is less than or equal to four are (1,1), (1,2), (2,1), (2,2), (1,3), (3,1)

So the probability is 6/36 = 1/6 = 16.66%

* 1. Total number of ways the sum is divisible by 2 and 3 both is the sum is divisible by 6. Number of possible ways for that is (1,5), (5,1),(6,6),(2,4), (4,2),(3,3)

So the probability is 6/36 = 1/6 = 16.66%

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?

**Answer:**

Total number of ways two balls can be drawn at random from 7 balls in the bag is

7C2 = 7!/2!(7-2)! = 21.

Total number of blue balls is 2

Total number of ways to draw the balls which are not blue i.e. green and red balls is 5C2 = 10

So the probability of drawing 2 balls and none of them is blue is 10/21 = 47.6%

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

**Answer:**

To calculate the expected number of candies for a randomly selected child, multiply the candies count for each child by their respective probability and then sum up the results:

Expected number of candies = (1 \* 0.015) + (4 \* 0.20) + (3 \* 0.65) + (5 \* 0.005) + (6 \* 0.01) + (2 \* 0.120) = 0.015 + 0.80 + 1.95 + 0.025 + 0.06 + 0.24 = 3.085

Therefore, the expected number of candies for a randomly selected child is 3.085

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**

**Answer:**

For Points

1. Mean = 3.5965625000000006
2. Median = 3.6950000000000003
3. Mode = 3.92
4. Variance = 0.2858813508064516
5. Standard Deviation = 0.5346787360709715
6. Range = 2.17
7. Inference: Typical value in the array is 3.59, the central tendency of the points values is 3.69, the most frequently occurring value 3.92, the values from array vary from average value by 0.28, the low variance indicates values are tightly clumped together, low standard deviation indicates the data is tightly clustered around mean, the range of 2.17 tells that the data is not spread out widely.

For Score

1. Mean = 3.2172499999999995
2. Median = 3.325
3. Mode = 3.44
4. Variance = 0.9573789677419355
5. Standard Deviation = 0.9784574429896967
6. Range = 3.9110000000000005
7. Inference: : Typical value in the array is 3.2, the central tendency of the points values is 3.325, the most frequently occurring value 3.44, the values from array vary from average value by 0.95, the low variance indicates values are tightly clumped together, low standard deviation indicates the data is tightly clustered around mean, the range of 3.911 tells that the data is not spread out widely.

For Weight

1. Mean = 17.848750000000003
2. Median = 17.71
3. Mode = 17.02
4. Variance = 3.1931661290322575
5. Standard Deviation = 1.786943236096843
6. Range = 8.399999999999999
7. Inference: : Typical value in the array is 17.84, the central tendency of the points values is 17.71, the most frequently occurring value 17.02, the values from array vary from average value by 3.19, the low variance indicates values are tightly clumped together, low standard deviation indicates the data is tightly clustered around mean, the range of 8.33 tells that the data is not spread out widely.

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?

**Answer:**

Assuming that each patient is equally likely to be chosen at random, the probability of choosing any one patient is 1/9 (since there are 9 patients in total). Therefore, the expected value of the weight can be calculated as follows:

E(X) = (108 \* 1/9) + (110 \* 1/9) + (123 \* 1/9) + (134 \* 1/9) + (135 \* 1/9) + (145 \* 1/9) + (167 \* 1/9) + (187 \* 1/9) + (199 \* 1/9)

E(X) = 134.22 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**

**Answer:**

For Q9\_a.csv

Speed

1. Skew = -0.11395477012828319
2. Kurtosis = -0.5771474239437371

Inference: If the skewness of an array is -0.11395477012828319, it indicates that the distribution of the data is slightly negatively skewed. This means that the data is skewed to the left, with the left tail being longer than the right tail. In other words, the mean of the data is slightly lower than the median.

If the kurtosis of an array is -0.5771474239437371, it indicates that the distribution of the data is slightly platykurtic. This means that the data has a flatter peak and thinner tails than a normal distribution. In other words, the distribution has fewer outliers or extreme values compared to a normal distribution.

Distance

1. Skew = 0.7824835173114966
2. Kurtosis = 0.24801865717051808
3. Inference: If the skewness of an array is 0.7824835173114966, it indicates that the distribution of the data is positively skewed. This means that the data is skewed to the right, with the right tail being longer than the left tail. In other words, the mean of the data is higher than the median.

If the kurtosis of an array is 0.24801865717051808, it indicates that the distribution of the data is slightly leptokurtic. This means that the data has a more peaked shape and heavier tails than a normal distribution. In other words, the distribution has more outliers or extreme values compared to a normal distribution.

For Q9\_b.csv

SP

1. Skew = 1.5814536794423764
2. Kurtosis = 2.7235214865269244

Inference: : If the skewness of an array is 1.5814536794423764 , it indicates that the distribution of the data is positively skewed. This means that the data is skewed to the right, with the right tail being longer than the left tail. In other words, the mean of the data is higher than the median.

If the kurtosis of an array is 2.7235214865269244, it indicates that the distribution of the data is slightly leptokurtic. This means that the data has a more peaked shape and heavier tails than a normal distribution. In other words, the distribution has more outliers or extreme values compared to a normal distribution.

Weight

1. Skew = -0.6033099322115126
2. Kurtosis = 0.8194658792266849
3. Inference: If the skewness of an array is -0.6033099322115126, it indicates that the distribution of the data is slightly negatively skewed. This means that the data is skewed to the left, with the left tail being longer than the right tail. In other words, the mean of the data is slightly lower than the median.  
   If the kurtosis of an array is 0.8194658792266849, it indicates that the distribution of the data is slightly leptokurtic. This means that the data has a more peaked shape and heavier tails than a normal distribution. In other words, the distribution has more outliers or extreme values compared to a normal distribution.

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



**Answer**:

**For Histogram**

1. Shape of the distribution:

The histogram is not in bell shape so it is not a normal distribution. The data is positively skewed as it has long tail to the right. The right tail is larger than the left tail which means the mean of the data is larger than the median.

1. Center of Distribution

200 chicks have the weight between 50 and 100. The mode of the data is 200. The center of the histogram is situated at 50.

1. Spread of the distribution

The data has a large spread.

1. Outliers

There are outliers in the data for weights between 300 to 400.

**For Box plot:**

1. The plot is positively skewed.
2. There are outliers in the plot greater than the mean and median.
3. The IQR is small so the data has a small spread.

**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?

**Answer**:

For a 94% confidence interval, the corresponding z-score is 1.88.

For a 98% confidence interval, the corresponding z-score is 2.33.

For a 96% confidence interval, the corresponding z-score is 1.96.

Formula for calculating the Confidence interval is CI = X̄ ± z\*(σ/√n)

Plugging in the values given in the question, we get:

For a 94% confidence interval: CI = 200 ± 1.88\*(30/√2000) = (198.45, 201.55)

For a 98% confidence interval: CI = 200 ± 2.33\*(30/√2000) = (197.08, 202.92)

For a 96% confidence interval: CI = 200 ± 1.96\*(30/√2000) = (198.00, 202.00)

**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?

**Answer:**

* Mean = 41.0
* Median = 40.5
* Mode = 41
* Variance = 25.529411764705884
* Standard Deviation = 5.05266382858645
* Range = 22

Inference:

1. Mean and median suggests that student's marks are clustered around 40-41.
2. The Standard deviation, variance and range suggest that suggest that the marks have a relatively small range, but there is still some variability in the dataset.
3. There are no significant outliers

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

**Answer**:

If the mean and median of a dataset are equal, it indicates that the data is symmetrically distributed around the central value. In such a case, the skewness of the data would be zero, indicating that the distribution is perfectly symmetrical.

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

**Answer**:

If the mean of a dataset is greater than the median, it indicates that the distribution is positively skewed. This means that the data is skewed to the right, with the right tail being longer than the left tail. A positive skewness can also impact measures of central tendency, such as the mean, making them larger than measures such as the median.

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

If the median of a dataset is greater than the mean, it indicates that the distribution is negatively skewed. This means that the data is skewed to the left, with the left tail being longer than the right tail.

A negative skewness can also impact measures of central tendency, such as the mean, making them smaller than measures such as the median.

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

**Answer**:

A positive kurtosis value indicates that a dataset has a distribution that is more peaked and has heavier tails than a normal distribution. This is known as leptokurtic distribution.

A leptokurtic distribution indicates that the data has a higher concentration of values around the mean, with more extreme values than a normal distribution. This can be caused by several factors, such as outliers or the presence of clusters in the data.

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

**Answer**:

A negative kurtosis value indicates that a dataset has a distribution that is less peaked and has lighter tails than a normal distribution. This is known as platykurtic distribution.

A platykurtic distribution indicates that the data has a lower concentration of values around the mean and more values that are farther from the mean than a normal distribution. This can be caused by several factors, such as a large sample size, data that has a high degree of variability, or the absence of outliers.

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



What can we say about the distribution of the data?

**Answer**:

There are no outliers in the dataset. The minimum of the data set is at 1 and the maximum if the data set is 19. The first quartile is at 10, the median is at 15 and the third quartile is at 18. The IQR of the data set is 8. Since Q1 is 10, 25% of data is below 10 and since Q3 is 18, 75% of data is below 18.

What is nature of skewness of the data?

**Answer**: The data is negatively skewed as the left tail is greater than the right tail.

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

The IQR of the data set is 8.

Q19) Comment on the below Boxplot visualizations?



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

**Answer:**

The box plot 1 has a small spread than box plot 2. Both the plots have skew of 0 i.e. they are normally distributed. The median of both the plots is the equal. The IQR of box plot 1 is smaller than IQR of box plot 2. The first quartile of box plot 1 is greater than the first quartile of plot 2 and the third quartile of plot 1 is smaller than plot 2.

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)

**Answer:**

1. P(MPG>38) = The probability that the value of elements greater than 38 is **0.41**
2. P(MPG<40) = The probability that the value of elements less than 40 is **0.75**
3. P (20<MPG<50) = The probability that the value of element is > 20 and <50 is 0.8518518518518519

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Answer:** Using the Shapiro-Wilks test in python the MPG of cars in the given data set follows the 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

**Answer:**

For WC : The data does not follow a normal distribution.

For AT: The data follows a normal distribution.

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

**Answer :**

**Using the following Python code :**

*from scipy.stats import norm*

*# 90% confidence interval*

*z\_90 = norm.ppf(0.95)*

*# 94% confidence interval*

*z\_94 = norm.ppf(0.97)*

*# 60% confidence interval*

*z\_60 = norm.ppf(0.8)*

*print(f"Z score for a 90% confidence interval: {z\_90:.2f}")*

*print(f"Z score for a 94% confidence interval: {z\_94:.2f}")*

*print(f"Z score for a 60% confidence interval: {z\_60:.2f}")*

Z score for a 90% confidence interval: 1.64

Z score for a 94% confidence interval: 1.88

Z score for a 60% confidence interval: 0.84

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

**Using the following Python code**

*from scipy.stats import t*

*# degrees of freedom for a sample of size 25*

*df = 24*

*# 95% confidence interval*

*t\_95 = t.ppf(0.975, df)*

*# 96% confidence interval*

*t\_96 = t.ppf(0.98, df)*

*# 99% confidence interval*

*t\_99 = t.ppf(0.995, df)*

*print(f"t-score for a 95% confidence interval: {t\_95:.2f}")*

*print(f"t-score for a 96% confidence interval: {t\_96:.2f}")*

*print(f"t-score for a 99% confidence interval: {t\_99:.2f}")*

t-score for a 95% confidence interval: 2.06

t-score for a 96% confidence interval: 2.17

t-score for a 99% confidence interval: 2.80

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

very difficult problem….cannot solve…