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

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

**Ans**: 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

**Ans Q4a)**: is **zero**. This is because the minimum possible sum of two dice is 2, when both dice show 1. There is **no way to get a sum of 1 with two dice**.

**Ans Q4b)**: is the same as the probability that the sum is 2, 3, or 4. There are 36 possible outcomes when rolling two dice, and only 6 of them have a sum of 2, 3, or 4. These are (1,1), (1,2), (2,1), (1,3), (3,1), and (2,2). Therefore, the probability is 6/36, which simplifies to **1/6**

**Ans Q5c)**: is the same as the probability that the sum is divisible by 6, since 6 is the least common multiple of 2 and 3. There are only four possible outcomes that have a sum divisible by 6, namely (1,5), (5,1), (3,3), and (6,6). There are 36 possible outcomes in total when rolling two dice, so the probability is 4/36, which simplifies to **1/9**

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 that none of the balls drawn is blue is 10/21. This is because there are 10 ways to choose two balls from the 5 non-blue balls, and 21 ways to choose two balls from the 7 total balls

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**: 1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*.0120 =**3.090**

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: Q7.ipynb**

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: Q8.ipynb**

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans: Q9\_a.ipynb**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans: Q9\_b.ipynb**

**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?

**Explanation:**

conf\_94 =stats.t.interval(alpha = 0.94, df=1999, loc=200, scale=30/np.sqrt(2000)) print(np.round(conf\_94,0)) print(conf\_94) For 94% confidence interval Range is [ 198.73 – 201.26] For 98% confidence interval range is [198.43 – 201.56] For 96% confidence interval range is [198.62 – 201.37]

**Ans**:  the following confidence intervals:

| **Confidence Level** | **Confidence Interval** |
| --- | --- |
| 94% | (198.29, 201.71) |
| 98% | (197.61, 202.39) |
| 96% | (198.02, 201.98) |

**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?

**Explanation:**

1. Find mean, median, variance, standard deviation. Ans: Mean =41, Median =40.5, Variance =25.52 and Standard Deviation =5.05
2. 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

**Q12-1 - Ans:**

Mean = 41

Median = 40.5

Variance= 25.52

Standard Deviation ​= 5.05

**Q12 -2 – Ans**:

**Slightly skewed to the right**: The mean is higher than the median, which indicates that there are some high scores that pull the mean up. The highest score is 56, which is far from the mean and the median.

**Moderately dispersed**: The variance and the standard deviation are not very large, which means that most of the scores are close to the mean. However, they are not very small either, which means that there is some variation in the scores. The range of the scores is 22, which is about 53% of the mean.

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

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

**Ans**: The nature of skewness when the mean and the median of data are equal is **zero skew**. This means that the distribution of data is symmetrical, and its left and right sides are mirror images of each other. A normal distribution is an example of a symmetrical distribution with zero skew

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

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

**Ans**: The nature of skewness when the mean is greater than the median is **positive skew or right skew**. This means that the distribution of the data is asymmetrical and has a long tail on the right side, where the larger values are. In a positively skewed distribution, most of the data points are concentrated on the left side of the distribution, and there are some extreme values on the right side

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

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

**Ans:** The nature of skewness when the median is greater than the mean is **negative skew or left skew**. This means that the distribution of the data is asymmetrical and has a long tail on the left side, where the smaller values are. In a negatively skewed distribution, most of the data points are concentrated on the right side of the distribution, and there are some extreme values on the left side

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

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

**Ans:** A positive kurtosis value indicates that **the data has a leptokurtic distribution,** which means that it has a sharper peak and heavier tails than a normal distribution. This implies that there are more outliers or extreme values in the data than in a normal distribution. A positive kurtosis value is also known as **positive excess kurtosis**, which is the difference between the kurtosis of the data and the kurtosis of a normal distribution

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

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

**Ans:** A negative kurtosis value indicates that the data has a **platykurtic distribution,** which means that it has a flatter peak and thinner tails than a normal distribution. This implies that more data values are located near the mean and less data values are located on the tails

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



What can we say about the distribution of the data?

**Explanation:**

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

**Ans:** we can say that the distribution of the data is:

**Symmetrical**: The median is in the middle of the box, which means that the data is evenly distributed on both sides of the center.

**Unimodal**: The box plot has only one peak, which is the median. This means that there is only one most frequent value in the data.

**Not skewed**: The mean and the median are equal, which means that there is no skewness in the data. The data does not have a long tail on either side.

**Moderately dispersed**: The interquartile range (IQR) is relatively small, which means that most of the data values are close to the median. However, the range of the data is relatively large, which means that there is some variation in the data. The data does not have any outliers or extreme values.

What is nature of skewness of the data?

**Explanation**:

* The data is a skewed towards left. The whisker range of minimum value is greater than maximum

**Ans:** we can say that the nature of skewness of the data is **positive skew or right skew.** This means that the distribution of the data is asymmetrical and has a long tail on the right side, where the larger values are. In a positively skewed distribution, most of the data points are concentrated on the left side of the distribution, and there are some extreme values on the right side

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

**Explanation**:

* The Inter Quantile Range = Q3 Upper quartile – Q1 Lower Quartile = 18 – 10 =8

**Ans**:   
we can say that the IQR (Interquartile Range) of the data is approximately 4. The IQR is the difference between the third quartile (Q3) and the first quartile (Q1) of the data. In this box plot, Q3 appears to be around 14 and Q1 appears to be around 10. So, the IQR is:

**IQR=Q3−Q1=14−10=4**

The IQR measures the spread of the middle 50% of the data. A smaller IQR means that the data is more consistent around the median

Q19) Comment on the below Boxplot visualizations?



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

**Explanation**: 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.

**Ans:**

* **Boxplot 1** shows a more compact distribution of data with a smaller interquartile range (IQR) compared to Boxplot 2. The median value is also lower in Boxplot 1. This suggests that the dataset represented by Boxplot 1 has a lower central tendency and less variability than that of **Boxplot 2**.
* **Boxplot 2** has a wider interquartile range (IQR), indicating greater variability in the data, and its median value is higher. This suggests that the dataset represented by **Boxplot 2** has a higher central tendency and more variability than that of **Boxplot 1**.
* Both boxplots have no visible outliers, which means that there are no extreme values in either dataset that are far from the rest of the data.
* The y-axis label “wbs” could indicate the name of the variable being measured or compared by the boxplots

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)

**Ans:**

**a. P(MPG>38) Ans**: Prob\_MPG\_greater\_than\_38 = np.round(1 - stats.norm.cdf(38, loc= q20.MPG.mean(), scale= q20.MPG.std()),3) print('P(MPG>38)=',Prob\_MPG\_greater\_than\_38)

P(MPG>38)= 0.348

**b. P(MPG<40) Ans**: prob\_MPG\_less\_than\_40 = np.round(stats.norm.cdf(40, loc = q20.MPG.mean(), scale = q20.MPG.std()),3) print('P(MPG<40)=',prob\_MPG\_less\_than\_40)

P(MPG<40)= 0.729

**c. P (20<MPG<50) Ans:** prob\_MPG\_greater\_than\_20 = np.round(1-stats.norm.cdf(20, loc = q20.MPG.mean(), scale = q20.MPG.std()),3) print('p(MPG>20)=',(prob\_MPG\_greater\_than\_20)) p(MPG>20)= 0.943

prob\_MPG\_less\_than\_50 = np.round(stats.norm.cdf(50, loc = q20.MPG.mean(), scale = q20.MPG.std()),3) print('P(MPG<50)=',(prob\_MPG\_less\_than\_50)) P(MPG<50)= 0.956

prob\_MPG\_greaterthan20\_and\_lessthan50= (prob\_MPG\_less\_than\_50) - (prob\_MPG\_greater\_than\_20) print('P(20<MPG<50)=',(prob\_MPG\_greaterthan20\_and\_lessthan50)) 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: 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

Explanation:

z value for 90% confidence interval

print('Z score for 60% Conifidence Intervla =',np.round(stats.norm.ppf(.05),4)) Z score for 60% Conifidence Intervla = -1.6449

z value for 94% confidence interval

print('Z score for 60% Conifidence Intervla =',np.round(stats.norm.ppf(.03),4)) Z score for 60% Conifidence Intervla = -1.8808

z value for 60% confidence interval

print('Z score for 60% Conifidence Intervla =',np.round(stats.norm.ppf(.2),4)) Z score for 60% Conifidence Intervla = -0.8416

**Ans:** The Z scores of a confidence interval are the values of Z that correspond to the desired level of confidence. They can be found by using a Z-table or a calculator. For example, for a 90% confidence interval, we need to find the value of Z that leaves 5% of the area in each tail of the standard normal curve. This value is 1.645. Similarly, we can find the Z scores for other confidence levels:

| **Confidence Level** | **Z Score** |
| --- | --- |
| 90% | 1.645 |
| 94% | 1.881 |
| 60% | 0.842 |

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

**Explanation**:

print('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.025,df=24),4)) T score for 95% Confidence Interval = -2.0639

print('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.005,df=24),4)) T score for 95% Confidence Interval = -2.7969

**Ans:** The t scores of a confidence interval are the values of t that correspond to the desired level of confidence. They can be found by using a t-table or a calculator. For **example**, for a 95% confidence interval with a sample size of 25, we need to find the value of t that leaves 2.5% of the area in each tail of the t-distribution with 24 degrees of freedom. This value is 2.064. Similarly, we can find the t scores for other confidence levels:

| **Confidence Level** | **t Score** |
| --- | --- |
| 95% | 2.064 |
| 96% | 2.172 |
| 99% | 2.797 |

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

**Explanation**:

import numpy as np

Import scipy as stats

t\_score = (x - pop mean) / (sample standard daviation / square root of sample size) (260-270)/90/np.sqrt(18))

t\_score = -0.471

stats.t.cdf(t\_score, df = 17) 0.32 = 32%

**Ans:**

This is a question about **hypothesis testing** using the t-distribution. To answer it, we need to calculate the t-statistic and the corresponding p-value for the given data. The t-statistic is given by the formula:

where xˉ is the sample mean, μ is the population mean, s is the sample standard deviation, and n is the sample size. Plugging in the given values, we get:

The p-value is the probability of getting a t-statistic less than or equal to the observed value, assuming the null hypothesis (that the population mean is 270) is true. To find the p-value, we need to use a t-table or a calculator. The degrees of freedom for the t-distribution are n−1=17. Using a t-table 1, we can see that the p-value is approximately 0.3218. This means that there is a 32.18% chance of getting a sample mean of 260 or less, if the true population mean is 270.

Alternatively, we can use a calculator to find the p-value. For example, using the R programming language, we can use the command pt(-0.471, 17) to get the same result