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

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

A) Number of favorable outcomes = HHT + HTH + THH = 3

Total number of possible outcomes = 23 = 8

P (Two heads and one tail) = Number of favorable outcomes / Total number of possible outcomes = 3/8 = 0.375 = 37.5%

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
4. Number of possible outcomes for the above event is N (Event (Two dice rolled)) = 62 = 36

Probability = No of favorable Outcomes / Total number of possible outcomes

a) P (sum = 1) = 0/36 = 0

b) P (Sum <= 4) = 6 / 36 = 1/6 = 0.166 = 16.66%

c) P (Sum is divisible by 2 and 3) = 6 / 36 = 1/6 = 0.166 = 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?

1. There are a total of 7 balls in a bag, and 2 of them are blue.

We need to choose 2 balls out of 7, and none of them should be blue.

This means we need to choose 2 balls out of 5 non-blue balls.

The probability of choosing 2 balls out of 5 non-blue balls is:

P(choosing 2 non-blue balls) = 5C2 / 7C2

where 5C2 is the number of combinations of 2 balls out of 5, and 7C2 is the number of combinations of 2 balls out of 7.

5C2 = 5! / (2! \* 3!) = 10

7C2 = 7! / (2! \* 5!) = 21

Therefore, the probability of choosing 2 balls out of 5 non-blue balls is:

P(choosing 2 non-blue balls) = 10 / 21 = 0.476

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

A)  Expected number of candies for a randomly selected child = Sum (Candies count \* Probability of candies count) = 1\*0.015+4\*0.20+3\*0.65 +5\*0.005+6\*0.01+2\*0.120 = 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.

1. **Mean**

\* Points = 3.59

\* Score = 3.21

\* Weigh = 17.84

**Median**

\* Points = 3.69

\* Score = 3.32

\* Weigh = 17.71

**Mode**

\* Points = 3.07

\* Score = 3.44

\* Weigh = 17.02

**Variance**

\* Points = 0.28

\* Score = 0.95

\* Weigh = 3.19

**Standard Deviation**

\* Points = 0.53

\* Score = 0.97

\* Weigh = 1.78

**Range**

\* Points = 2.17

\* Score = 3.911

\* Weigh = 8.4

**Inferences:**

1. The mean score for all three columns is relatively close together, with Points having the highest mean (3.46), followed by Score (3.36), and then Weigh (17.65).
2. The median score for all three columns is also relatively close together, with Points having the highest median (3.44), followed by Score (3.32), and then Weigh (17.71).
3. The mode score for all three columns is not as close together, with Points having the highest mode (3.07), followed by Weigh (17.02), and then Score (3.44).
4. The variance for Points is the lowest (0.28), followed by Score (0.95), and then Weigh (3.19).
5. The standard deviation for Points is the lowest (0.53), followed by Score (0.97), and then Weigh (1.78).
6. The range for Points is the lowest (2.17), followed by Score (3.91), and then Weigh (8.4).

**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?

1. Given weights of the patients at a clinic (in pounds) are 108, 110, 123, 134, 135, 145, 167, 187, 199

The Probability of selecting a person at random is 1/9

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

Therefore, the expected value of the weight of a person chosen at random is **145.33**

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

**Cars speed and distance**

**Use Q9\_a.csv**

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Skewness | Kurtosis | Inference |
| Speed | -0.12 | -0.51 | The distribution is slightly **negatively skewed**, meaning there is a longer tail on the left side indicating more low speeds than high speeds.  The distribution is **platykurtic**, meaning it has thinner tails than a normal distribution. |
| Distance | 0.81 | 0.41 | The distribution is **positively skewed**, meaning there is a longer tail on the right side indicating more high distances than low distances.  The distribution is **platykurtic**, meaning it has thinner tails than a normal distribution. |

**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Skewness | Kurtosis | Inference |
| SP | 1.61 | 2.98 | The distribution is **positively skewed**, meaning there are more values on the right tail of the distribution than on the left.  The distribution is **Leptokurtic**, meaning it has heavier tails and a higher peak than a normal distribution. This means there are more extreme values (outliers) than in a normal distribution. |
| WT | -0.61 | 0.95 | The distribution is **negatively skewed**, meaning there are more values on the left tail of the distribution than on the right.  The distribution is **platykurtic**, meaning it has lighter tails and a lower peak than a normal distribution. This means there are fewer extreme values (outliers) than in a normal distribution. |

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



1. **Histogram:** The histogram of a Chick Weights is Positively Skewed i.e., Mean > median > Mode. This means that there are more chicks that are lighter than average than there are chicks that are heavier than average. The skewness is not too extreme, but it is noticeable.

**Boxplot:** Data is Positively Skewed and 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?

**A)** import numpy as np

from scipy.stats import norm

CI\_94 = norm.interval(0.94,

loc=200,

scale=30/np.sqrt(2000))

print("Confidence interval at 94%: ",CI\_94)

CI\_96 = norm.interval(0.96,

loc=200,

scale=30/np.sqrt(2000))

print("Confidence interval at 96%: ",CI\_96)

CI\_98 = norm.interval(0.98,

loc=200,

scale=30/np.sqrt(2000))

print("Confidence interval at 98%: ",CI\_98)

**Results**

Confidence interval at 94% : (198.738325292158, 201.261674707842)

Confidence interval at 96% : (198.62230334813333, 201.37769665186667)

Confidence interval at 98%: (198.43943840429978, 201.56056159570022)

**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?

**A)** (1) Mean =41, Median =40.5, Variance =25.53 and Standard Deviation =5.05

(2) The scores appear to be normally distributed and there are no outliers. The scores show an upward trend and 83.33% of the scores are within 1 standard deviation of the mean. The student is performing well and consistently above average.

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

A) No skewness, the data follows a normal distribution

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

A) When the mean > median, the distribution is said to be positively skewed (right Skewed). This means that the "tail" of the distribution extends more towards the larger values, with a greater mass of data on the right side of the distribution as compared to the left.

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

A) When the median > mean, the data is said to have a Negatively Skewed (Left Skewed) .This means that the "tail" of the distribution extends more towards the smaller values, with a greater mass of data on the left side of the distribution as compared to the right.

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

A) A positive kurtosis value indicates that a data set has a **peaked** distribution with **heavy tails**. Compared to a normal distribution, there are more data points concentrated around the center (the peak) and more extreme values (outliers) in the tails on both ends.

In Statistical terms, a positive kurtosis value indicates that the dataset has leptokurtic characteristics

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

A) A negative kurtosis value indicates that a data set has a **flattened** distribution with **light tails**. Compared to a normal distribution, there are fewer data points concentrated around the center and fewer extreme values (outliers) in the tails on both ends.

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



What can we say about the distribution of the data?

A) The data is not symmetric. It is more concentrated towards right side

What is nature of skewness of the data?

A) Negatively Skewed

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

A) Inter Quartile Range (IQR) = Q3 – Q1

= 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.

A) **Boxplot 1 :** Data is Normally Distributed. There are no Outliers. Median is around 263, it has less IQR as compared to 2nd Boxplot

**Boxplot 2 :** Data is Normally Distributed. There are no Outliers. Median is around 265, it has more IQR as compared to 1st Boxplot

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

a. P(MPG>38)

import numpy as np

import pandas as pd

from scipy import stats

df = pd.read\_csv("C:\\Users\\marpaka pradeep sai\\Downloads\\Cars.csv")

df

MPG\_38 = 1 - stats.norm.cdf(38, loc=df['MPG'].mean(), scale= df['MPG'].std())

print('P(MPG>38) :',np.round(MPG\_38,3))

P(MPG>38) = 0.348

b. P(MPG<40)

MPG\_40 = stats.norm.cdf(40, loc=df['MPG'].mean(), scale= df['MPG'].std())

print('P(MPG<38) :',np.round(MPG\_40,3))

P(MPG<40) = 0.729

c. P (20<MPG<50)

MPG\_20 = stats.norm.cdf(20, loc=df['MPG'].mean(), scale= df['MPG'].std())

print('P(MPG>20) :',np.round(MPG\_20,3))

MPG\_50 = stats.norm.cdf(50, loc=df['MPG'].mean(), scale= df['MPG'].std())

print('P(MPG<50) :',np.round(MPG\_50,3))

print('P(20<MPG<50) :', np.round((MPG\_50 - MPG\_20),4))

P (20<MPG<50) = 0.899

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

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

1. Adipose Tissue (AT) and Waist Circumference (Waist) doesn’t follow normal distribution

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

A) import numpy as np

from scipy import stats

from scipy.stats import norm

print("Z score at 90% confidence interval : ", np.round(stats.norm.ppf(.95),3))

**Output** : Z score at 90% confidence interval : 1.645

print("Z score at 94% confidence interval : ", np.round(stats.norm.ppf(.97),3))

**Output :** Z score at 94% confidence interval : 1.881

print("Z score at 60% confidence interval : ", np.round(stats.norm.ppf(.80),3))

**Output :** Z score at 60% confidence interval : 0.842

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

A) import numpy as np

import scipy.stats as stats

print('T-score for 95% Confidence Interval :',np.round(stats.t.ppf(0.025,df=24),3))

**Output :** T-score for 95% Confidence Interval : -2.064

print('T score for 94% Confidence Interval =',np.round(stats.t.ppf(0.03,df=24),3))

**Output :** T-score for 94% Confidence Interval : -1.974

print('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.005,df=24),3))

**Output :** T-score for 95% Confidence Interval = -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

A) import numpy as np

import scipy as stats

x\_bar = 260

mu = 270

SD = 90

n = 18

t\_score = (x\_bar - mu) / (SD / np.sqrt(n))

print("The Probability that 18 randomly selected bulbs would have an average life of no more than 260 days is ", np.round((stats.t.cdf(t\_score, df = n-1)\*100),2),'%')

**Output :** The Probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 32.17 %