Statistics

Question: 1

What is the meaning of six sigma in statistics? Give proper example

Ans- Six Sigma is a method used by companies to make their processes as close to perfect as possible. It focuses on reducing mistakes and making sure products or services meet high standards.

Benefits of Six Sigma:------

1. Improved product quality and customer satisfaction
2. Reduced costs due to fewer defects and rework
3. Increased efficiency and productivity
4. Enhanced process control and risk management

Example-

Imagine a pizza delivery company wants to make sure they deliver pizzas on time every time. Using Six Sigma, they analyze their process and find ways to make it more efficient. As a result, they aim to have very few late deliveries—maybe only three or four late deliveries out of every million orders. This level of accuracy and consistency is what Six Sigma aims for.

Question: 2

What type of data does not have a log-normal distribution or a Gaussian distribution? Give proper example

Ans- There are many types of data that don't follow a log-normal or Gaussian (normal) distribution. Here are some examples:

1. Discrete Data with Limited Values:

Shoe Sizes- Shoe sizes are typically whole or half numbers (e.g., 7, 7.5, 8). They represent discrete data with a limited range of possible values. While you could technically calculate a mean and standard deviation, these wouldn't accurately capture the distribution of shoe sizes in a population.

Customer Satisfaction Ratings- Customer satisfaction ratings might be on a scale of 1 to 5 (or similar), resulting in discrete data with a limited range. A normal distribution wouldn't accurately depict the distribution of these ratings, as there might be a natural clustering towards the center (average) or extremes (very satisfied/dissatisfied).

2. Data with a Natural Lower Bound (Zero or Positive):

Waiting Times- Waiting times (e.g., at a bank or call center) can't be negative. This creates a skewed distribution where most values are concentrated near zero (short waiting times) with a longer tail towards higher wait times. While a log-normal distribution might capture the positive skew, it might not always be the best fit for all waiting time data.

Income Levels-   Income levels are typically non-negative. A normal distribution wouldn't be suitable because it allows for negative values, which wouldn't be realistic for income. Instead, distributions like Pareto or log-normal might be more appropriate depending on the data's skewness.

3. Data with an Upper Bound:

Number of website visitors per day- The number of visitors to a website on a given day can't be infinitely high. There's a practical limit based on the website's popularity and available resources. A normal distribution wouldn't be suitable here, and alternative distributions like Poisson or truncated distributions might be more appropriate.

Income data- Income data is often right-skewed, with a long tail of high earners. While it cannot be negative, it does not follow a log-normal or Gaussian distribution due to the unequal distribution of income across the population.

Question: 3
What is the meaning of the five-number summary in Statistics? Give proper example

Ans- The five-number summary is a concise way to describe the spread and center of a dataset using five key values. It provides an overview of how our data is distributed without getting into complex calculations.

We explain of the five numbers summary—-

Minimum: The smallest value in the dataset.
First Quartile (Q1): The value that separates the lowest 25% of the data from the rest.
Median: The middle value when the data is ordered from least to greatest. If we have an even number of data points, the median is the average of the two middle values.
Third Quartile (Q3): The value that separates the highest 25% of the data from the rest.
Maximum: The largest value in the dataset.
Example:

Consider the following dataset representing the scores of students in a class:

60,65,70,75,80,85,90,95,100

The five-number summary for this dataset would be:

- Minimum: 60
- Q1: 70
- Median: 80
- Q3: 90
- Maximum: 100

This summary provides a concise layout of the distribution of scores, indicating that the majority of scores fall between 70 and 90, with equal numbers of scores below and above the median.

Question: 4

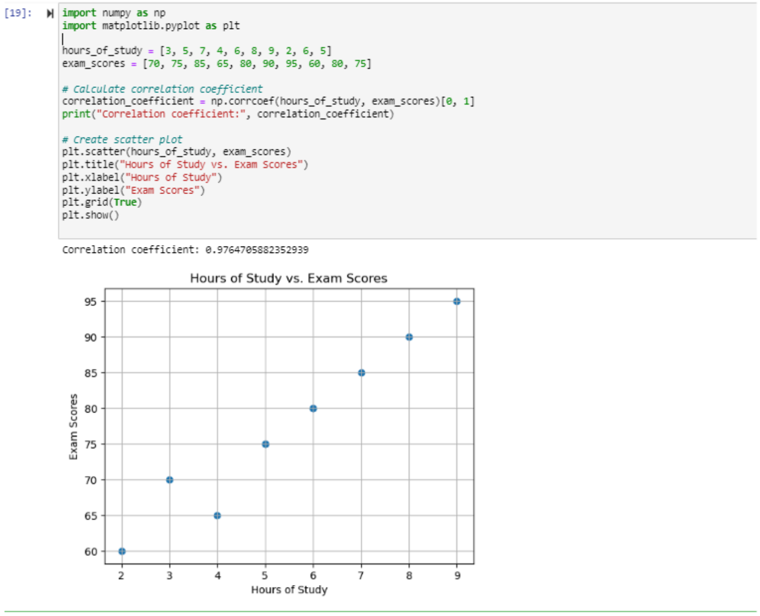
What is correlation? Give an example with a dataset & graphical representation on jupyter Notebook

Ans- Correlation is a statistical measure that indicates the extent to which two variables change together. It doesn't necessarily imply causation, but rather the existence of a relationship between the variables. There are different types of correlation—------

1. Positive Correlation- When one variable increases, the other tends to increase as well (e.g., study hours and exam scores).
2. Negative Correlation- When one variable increases, the other tends to decrease (e.g., age and eyesight).
3. No Correlation- No clear relationship exists between the two variables.

The strength of the correlation is measured by a correlation coefficient, the range from -1 (perfect negative correlation) to +1 (perfect positive correlation). A value of 0 indicates no correlation.

Python Code—



```
[19]:  import numpy as np
       import matplotlib.pyplot as plt

       hours_of_study = [3, 5, 7, 4, 6, 8, 9, 2, 6, 5]
       exam_scores = [70, 75, 85, 65, 80, 90, 95, 60, 80, 75]

       # Calculate correlation coefficient
       correlation_coefficient = np.corrcoef(hours_of_study, exam_scores)[0, 1]
       print("Correlation coefficient:", correlation_coefficient)

       # Create scatter plot
       plt.scatter(hours_of_study, exam_scores)
       plt.title("Hours of Study vs. Exam Scores")
       plt.xlabel("Hours of Study")
       plt.ylabel("Exam Scores")
       plt.grid(True)
       plt.show()
```

```
Correlation coefficient: 0.9764705882352939
```