**Statistics:**

**Question: 1**

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

Answer:

In statistics, sigma represents the standard deviation, which measures how spread out data points are from the average value (mean) in a particular set of data.

Six Sigma uses sigma as a metric for process capability. It refers to a process that has very minimal variability, resulting in almost no defects. A six sigma process achieves a defect rate of only 3.4 DPMO (defects per million opportunities), which translates to a 99.99966% success rate.

Eg:

A normal shop might have some pizzas with slightly burnt crusts (a few standard deviations away from ideal). Six Sigma aims to get nearly every pizza perfect, with almost no burnt crusts (very few standard deviations from ideal). This translates to super happy customers.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Question: 2**

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

Answer:

a. Uniform Distribution:

This distribution represents data where all values within a specific range have an equal probability of occurring.

Example: Imagine rolling a fair six-sided die. Each number (1 to 6) has an equal chance of appearing.

This scenario wouldn't be suited for a normal distribution because the normal curve concentrates probability around the center, with lower probabilities at the extremes. In contrast, the uniform distribution spreads probability evenly across the entire range.

b. Exponential Distribution:

This distribution describes the time between events in situations where events occur continuously and independently at a constant average rate.

Example: The time between the arrivals of buses at a stop follows an exponential distribution. We know buses will arrive eventually, but the exact timing between them is uncertain.

An exponential distribution wouldn't be ideal for a normal distribution because the normal curve is symmetrical, with values on either side of the center balancing each other out. The exponential distribution, however, depicts a one-sided trend, with a higher probability of shorter wait times and a decreasing probability of longer waits.

c. Poisson Distribution:

This distribution represents the number of events occurring in a fixed interval of time or space. It's particularly useful for modeling rare events that occur randomly and independently of each other.

Example: The number of customer complaints received by a call center in an hour can be modeled with a Poisson distribution. Complaints are unlikely to be clustered or dependent on one another.

A normal distribution wouldn't be optimal for Poisson data because the normal curve allows for negative values, which don't make sense for counts. Additionally, the normal distribution predicts a smooth, continuous curve across all possible values, whereas complaint counts tend to cluster around smaller whole numbers with a gradual decrease as the numbers increase.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Question: 3**

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

Answer:

The Essential Five:

1. Minimum Value: This is the smallest value in the dataset, representing the absolute lower limit.
2. First Quartile (Q1): This value marks the 25th percentile. It indicates that 25% of the data points fall at or below this value.
3. Median (Q2): The median is the middle value when the data is arranged in ascending or descending order. It signifies the 50th percentile, with half the data falling below and half above it.
4. Third Quartile (Q3): This value represents the 75th percentile. It indicates that 75% of the data points fall at or below this value.
5. Maximum Value: This is the largest value in the dataset, representing the absolute upper limit.

In a dataset containing the test scores of 50 students, for example:

Minimum Value = 55

Q1 (First Quartile) = 72

Median (Q2) = 84

Q3 (Third Quartile) = 92

Maximum Value = 98

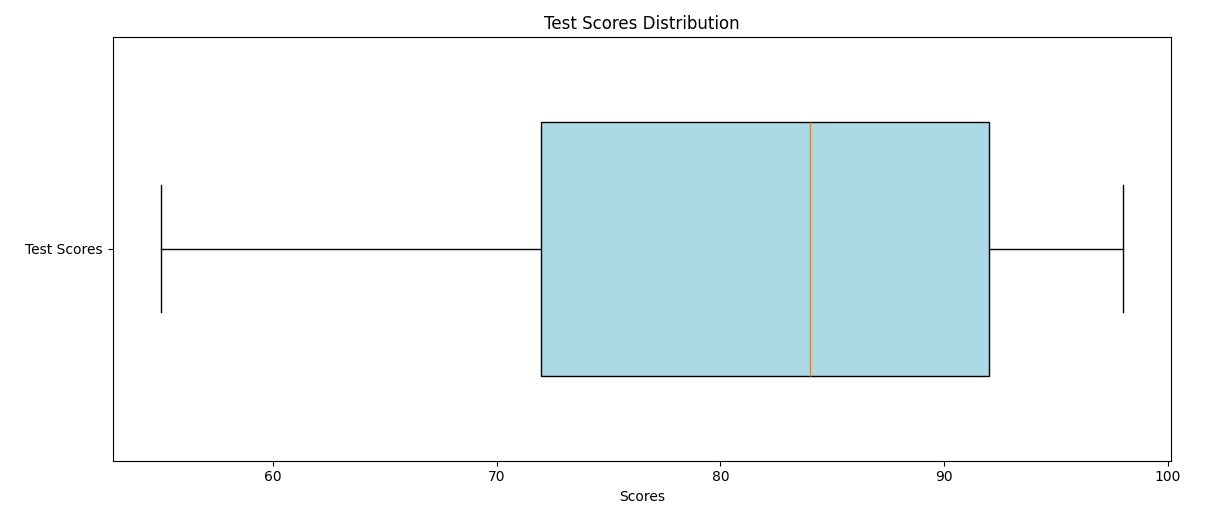
Here,

The lowest score was 55, and the highest score was 98.

The spread of scores ranges from 55 to 98.

The median score of 84 indicates that half the class scored below 84 and the other half scored above.

Q1 (72) and Q3 (92) reveal that the middle 50% of the scores fall between 72 and 92.



```

import matplotlib.pyplot as plt

import numpy as np

data = [55, 72, 84, 92, 98] # Minimum, Q1, Median, Q3, Maximum

plt.boxplot(data, vert=False, patch\_artist=True, widths=0.6)

plt.title('Test Scores Distribution')

plt.xlabel('Scores')

plt.yticks([1], ['Test Scores'])

box = plt.gca().get\_children()[0]

box.set(facecolor='lightblue')

plt.show()

```

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Question: 4**

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

Please refer to the below link for the answer.

<https://github.com/robinbanerje/iNeuron-DS-Assessment/blob/main/Statistics_question_3_%26_4.ipynb> ”

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_