Question: 1

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

Answer:

Six Sigma refers to a measure of quality that strives for near perfection. It's a disciplined, data-driven approach and methodology for eliminating defects in any process, from manufacturing to transactional and from product to service.

The statistical meaning of Sigma (standard deviation) measures how much a set of values varies. In a normal distribution, Six Sigma means that 99.99966% of all opportunities to produce some feature of a part are statistically expected to be free of defects. Practically, this translates to 3.4 defects per million opportunities (DPMO).

Here's a real-life example:

Let's say a pharmaceutical company manufactures tablets, and they want each tablet to contain 500mg of a particular active ingredient. In a Six Sigma process:

The average dose would be very close to 500mg.

The variation from tablet to tablet would be extremely small, such that virtually all tablets (99.99966%) would have a dosage between 499.999mg and 500.001mg.

This would mean that out of a million tablets produced, only 3.4 would have a dosage outside the 499.999mg to 500.001mg range.

This level of precision ensures that the medication is safe and effective for consumers, with almost no chance of a dosage error. Six Sigma processes in pharmaceutical manufacturing help maintain stringent quality control, ensuring the safety and efficacy of medications.

Question: 2

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

Answer:

Poisson Distribution:

This distribution is often used to model count data. For example, the number of emails received in an hour by a company's support desk or the number of cars passing through a toll booth in a day.

A key characteristic of Poisson distribution is that the mean and variance are equal, which is not the case in a Gaussian distribution. This kind of data is typically discrete and represents the number of times an event occurs in a fixed interval of time or space.

Bernoulli Distribution

This is a simple discrete distribution having only two outcomes: success or failure (1 or 0, true or false). An example would be flipping a coin, where you can only get heads or tails.

Unlike a Gaussian distribution, which is continuous and symmetric, the Bernoulli distribution is discrete and can be asymmetric if the probability of success is not equal to the probability of failure.

Uniform Distribution:

In a uniform distribution, all outcomes are equally likely. An example is the roll of a fair die where each of the six outcomes (1, 2, 3, 4, 5, or 6) has an equal probability of occurring.

Gaussian distributions have a bell curve shape, where values are more likely to occur near the mean, and the likelihood decreases symmetrically as you move away from the mean. In contrast, a uniform distribution does not have this property, as no single outcome is more likely than another.

Exponential Distribution:

The exponential distribution models the time between events in a process where events occur continuously and independently at a constant average rate. An example is the amount of time until a radioactive particle decays, or the time between arrivals at a service station.

It is a skewed distribution with a peak at the beginning and a long tail, which is different from the symmetric bell curve of a Gaussian distribution and doesn't follow the multiplicative process of a log-normal distribution.

Question: 3

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

Answer

Minimum: The smallest number in the dataset.

First Quartile (Q1): Also known as the lower quartile, is the median of the lower half of the dataset. It cuts off the lowest 25% of data.

Median (Q2): The middle value when the data are arranged in ascending order. If there is an even number of observations, it is the average of the two middle numbers.

Third Quartile (Q3): Also known as the upper quartile, is the median of the upper half of the dataset. It cuts off the highest 25% of data.

Maximum: The largest number in the dataset.

Eg: [78, 85, 96, 80, 70, 65, 92, 95, 88, 76]

Minimum: 65
Q1: 76 (the median of the lower half [65, 70, 76, 78, 80])
Median (Q2): 81.5 (the average of the two middle values 80 and 83 from the ordered list)
Q3: 92.5 (the median of the upper half [85, 88, 92, 95, 96])
Maximum: 96

Question: 4

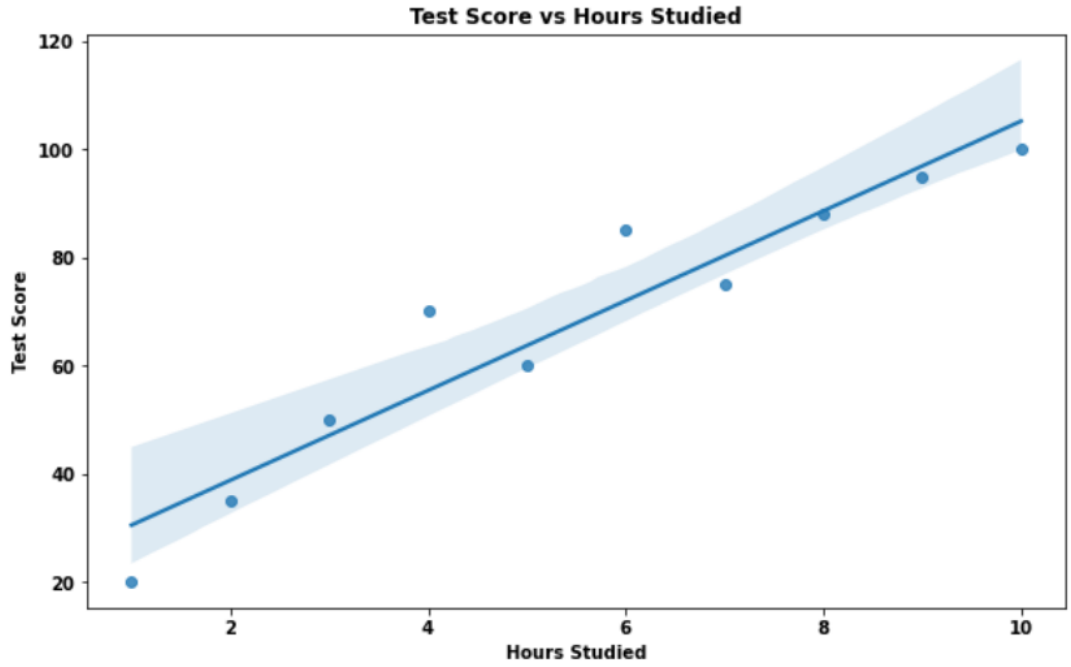
What is correlation? Give an example with a dataset & graphical representation on jupyter Notebook

Answer:
Correlation is a statistical measure that expresses the extent to which two variables are linearly related. It's a common tool for describing simple relationships without making a statement about cause and effect.

The correlation coefficient, often denoted as r, ranges from -1 to +1. A value of +1 implies a perfect positive correlation, meaning that as one variable increases, the other variable also increases. A value of -1 implies a perfect negative correlation, meaning that as one variable increases, the other variable decreases. A correlation of 0 means no linear relationship between the variables.

In [8]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

# Example dataset of study hours vs test scores
data = {
    'Hours_Studied': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
    'Test_Score': [20, 35, 50, 70, 60, 85, 75, 88, 95, 100]
}

df = pd.DataFrame(data)

# Calculate the correlation coefficient
correlation = df.corr()

# Create a scatter plot with a regression line
plt.figure(figsize=(10, 6))
sns.regplot(x='Hours_Studied', y='Test_Score', data=df)
plt.title('Test Score vs Hours Studied')
plt.xlabel('Hours Studied')
plt.ylabel('Test Score')
plt.show()

print("Correlation Coefficient:")
print(correlation)
```



Correlation Coefficient:

	Hours_Studied	Test_Score
Hours_Studied	1.000000	0.951998
Test_Score	0.951998	1.000000