**Statistics**

**Q1)ans**

Six Sigma is a methodology used in business and statistics to improve the quality of process outputs by identifying and removing the causes of defects and minimizing variability in manufacturing and business processes. It aims to achieve near-perfect quality by reducing defects to a level where they occur at a rate of fewer than 3.4 defects per million opportunities (DPMO). The term "Six Sigma" comes from the statistical concept of standard deviation (σ), where "Six Sigma" refers to a process that operates within six standard deviations from the mean, indicating that it has a very low defect rate

Here's a breakdown of the meaning of Six Sigma:

* 1.Define: Define the problem, the process, and the customer requirements clearly.
* 2.Measure: Measure key aspects of the current process and collect relevant data.
* 3.Analyze: Analyze the data to determine the root causes of defects and identify areas for improvement.
* 4.Improve: Improve the process by implementing solutions aimed at reducing defects and variability.
* 5.Control: Control the process to sustain the improvements and prevent defects from recurring.

An example of Six Sigma in action could be a manufacturing company aiming to reduce defects in its production process for a particular product. Suppose the company identifies that defects occur due to variations in temperature during the manufacturing process. They implement Six Sigma principles to address this issue:

* 1.Define: The company defines the problem as high defect rates in the production process and sets a goal to reduce defects to less than 3.4 DPMO.
* 2.Measure: They collect data on temperature variations during the production process and identify key metrics related to defect rates.
* 3.Analyze: Analyzing the data, they discover that temperature fluctuations beyond a certain range lead to defects in the final product.
* 4.Improve: They implement solutions such as installing temperature control systems and refining production processes to minimize temperature variations.
* 5.Control: They establish control measures to monitor temperature continuously during production and implement procedures to adjust settings if variations exceed acceptable limits.

By applying the Six Sigma methodology, the company aims to achieve a significant reduction in defects, resulting in higher-quality products and improved customer satisfaction.

**Q2)ans:**

Data that does not follow a log-normal distribution or a Gaussian (normal) distribution is often referred to as non-normal data. This type of data may exhibit asymmetry, heavy tails, or other characteristics that deviate from the bell-shaped curve associated with normal distributions.

Examples of data that do not have a log-normal or Gaussian distribution include:

**1..Skewed Data**: Skewed data is asymmetrically distributed, meaning that the tail on one side of the distribution is longer or heavier than the tail on the other side. For example, income data often exhibits right-skewness, where a few individuals have very high incomes, resulting in a long right tail.

**2.**.**Bimodal or Multimodal Data:** Bimodal or multimodal data have multiple peaks or modes in the distribution, indicating distinct subgroups within the data. For instance, if we measure the heights of individuals from two different populations (e.g., children and adults), the resulting distribution may have two peaks, one for each subgroup.

**3.**.**Heavy-Tailed Data:** Heavy-tailed distributions have tails that decay more slowly than those of normal distributions. These distributions exhibit more extreme values or outliers than would be expected under a Gaussian distribution. An example of heavy-tailed data is the distribution of stock market returns, which often display fat tails due to the occurrence of extreme market events.

**4.Discrete Data**: Discrete data take on distinct, separate values rather than forming a continuous distribution. Examples include counts of events (e.g., number of accidents in a day) or categorical variables (e.g., type of vehicle involved in an accident).

**5.Categorical Data**: Categorical data represent categories or groups rather than numerical values. Examples include gender, race, or types of products.

**5.Mixed Data:** Mixed data contain a combination of continuous, discrete, and categorical variables. For example, in a medical study, variables such as age (continuous), blood type (categorical), and number of hospital visits (discrete) may be collected.

**Q3)ans:**

The five-number summary is a set of descriptive statistics that provides a concise summary of the distribution of a dataset. It consists of five key values: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. These values divide the dataset into four equal parts, providing insights into the spread and central tendency of the data.

**Minimum:** The smallest value in the dataset. It represents the lower extreme of the data distribution.

**First Quartile (Q1):** The value below which 25% of the data fall. It marks the first quarter of the dataset when it is ordered from smallest to largest. Q1 is also known as the lower quartile.

**Median (Q2):** The middle value of the dataset when it is ordered from smallest to largest. It represents the 50th percentile or the value that separates the lower 50% of the data from the upper 50%.

**Third Quartile (Q3):** The value below which 75% of the data fall. It marks the third quarter of the dataset when it is ordered from smallest to largest. Q3 is also known as the upper quartile.

**Maximum:** The largest value in the dataset. It represents the upper extreme of the data distribution.

The five-number summary is particularly useful for understanding the central tendency, spread, and skewness of a dataset, as well as identifying potential outliers.

Example:

Consider the following dataset representing the ages of 10 individuals:

22,25,28,30,35,38,40,45,50,55

To find the five-number summary:

1.Minimum: 22

2.Q1 (First Quartile): 28

3.Median (Q2): 35

4.Q3 (Third Quartile): 45

5.Maximum: 55

Using the five-number summary, we can quickly understand that the dataset ranges from 22 to 55, with a median age of 35. Additionally, we can see that 25% of the individuals are below the age of 28 (Q1) and 75% are below the age of 45 (Q3). This summary provides a clear picture of the distribution of ages in the dataset.

**Q4)ans:**

Correlation is a statistical measure that describes the strength and direction of the relationship between two variables. It indicates whether and how much two variables change together. A correlation coefficient quantifies the degree to which changes in one variable are associated with changes in another variable.

The correlation coefficient, denoted by "r," ranges from -1 to 1:

* A correlation coefficient of 1 indicates a perfect positive correlation, meaning that as one variable increases, the other variable also increases linearly.
* A correlation coefficient of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other variable decreases linearly.
* A correlation coefficient of 0 indicates no linear relationship between the variables.

Notebook link: https://github.com/Kaveriborse/DS-assessment.git