Statistics

Ans 1) Six Sigma is a statistical methodology used to improve processes by reducing defects or errors to a level that corresponds to no more than 3 - 4 defects per million opportunities (DPMO).

The term "Sigma" refers to the standard deviation of a process. In a Six Sigma process, the goal is to reduce process variation so that it operates within six standard deviations of the process mean, allowing for very little variation and thus very high quality.

Here's an example to illustrate the concept of Six Sigma:

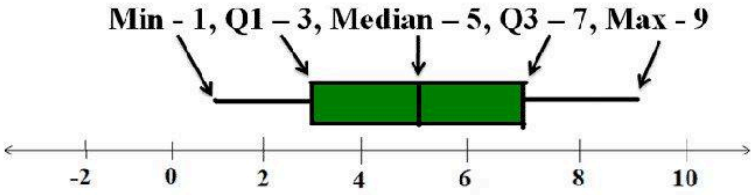
Let's say you work in a manufacturing company that produces widgets. Your company has implemented Six Sigma to improve the quality of widgets produced. The defect in this case could be a widget that doesn't meet certain specifications, such as incorrect size, shape, or functionality.

Before implementing Six Sigma, your process might produce widgets with an average of 10 defects per million widgets produced, which is equivalent to a 3 Sigma level.

After implementing Six Sigma methodologies, your process has been improved to produce widgets with an average of only 3.4 defects per million widgets, which corresponds to the 6 Sigma level.

So, in this example, Six Sigma has helped the company significantly reduce defects in its manufacturing process, leading to higher quality products and increased customer satisfaction.

Ans 3: Five Number summary describes given data in 5 numbers and they are: 1. Minimum
2. Q1
3. Q2 (or Median)
4. Q3
5. Maximum



Minimum : All the data lies above it. I.e., 100% data lies above this point.
Q1 : 75% of the data lies above this value.
Q2 : 50% of the data lies above this point.
Q3. 25% of the data lies above this point.
Maximum : All the data is below this point.

Example: In a class of 100 students

Minimum: This represents the lowest mark attained by any student in the class.

Q1 (First Quartile): This value separates the lowest 25% of the marks from the rest. In other words, 25% of the students scored below Q1, and 75% scored above it.

Median (Q2): This is the middle value of the dataset when it is ordered from smallest to largest. It represents the mark that divides the class into two equal parts, with 50% of the students scoring below it and 50% scoring above it.

Q3 (Third Quartile): This value separates the lowest 75% of the marks from the highest 25%. In other words, 75% of the students scored below Q3, and only 25% scored above it.

Maximum: This represents the highest mark attained by any student in the class.

For example, let's say you have a class of 100 students, and their marks in a test are as follows (ordered from lowest to highest): In this case:

- Minimum = 35
- Q1 = 52.5 (the average of the 25th and 26th marks, which are 50 and 55)
- Median = 72.5 (the average of the 50th and 51st marks, which are 70 and 75)
- Q3 = 87.5 (the average of the 75th and 76th marks, which are 85 and 90)
- Maximum = 100

This five-number summary provides a concise overview of the distribution of marks in the class, indicating the spread of marks and the central tendency.

Ans 2) Non-Gaussian Distribution with example are :
Pareto Distribution

The Pareto distribution is commonly used to describe phenomena where a small number of observations have very high values while the majority have much lower values. It often arises in economics, finance, and other fields. An example could be the distribution of wealth, where a small percentage of the population holds the majority of the wealth.

Example:

Suppose you're analyzing the distribution of income among individuals in a country. You find that a small percentage of the population controls the majority of the wealth, while the majority of the population has much lower income.

Let's say you collect data and find that:

- 20% of the population controls 80% of the wealth.
- 80% of the population shares the remaining 20% of the wealth.

This distribution of wealth follows a Pareto distribution, where a small percentage of observations (in this case, the wealthy individuals) have much higher values compared to the majority of observations.

Exponential Distribution

The exponential distribution describes the time between events in a Poisson process, where events occur continuously and independently at a constant rate. It has a single parameter, the rate parameter, which determines the rate at which events occur. An example could be the distribution of the time between arrivals of customers at a service desk, where the rate parameter represents the average rate of customer arrivals.

Example:

Imagine you're studying the waiting times between consecutive arrivals of customers at a service desk in a bank. You find that customers arrive randomly, and the time between arrivals follows a particular pattern.

After collecting data, you observe that on average, a customer arrives every 5 minutes. The exponential distribution describes this scenario, where the time between arrivals (interarrival times) follows an exponential pattern.

For instance, if you plot the probability density function (PDF) of interarrival times, it will have a characteristic shape where the probability of short interarrival times is high initially and decreases exponentially as the time between arrivals increases. This reflects the randomness and independence of customer arrivals in the process.

—----------

Ans 4) Correlation is a statistical measure that describes the strength and direction of a relationship between two variables. It indicates how much one variable changes when the other variable changes.

There are several types of correlation coefficients, but one of the most commonly used is the Pearson correlation coefficient, denoted by ‘r’. The Pearson correlation coefficient measures the linear relationship between two continuous variables. It ranges from -1 to 1:

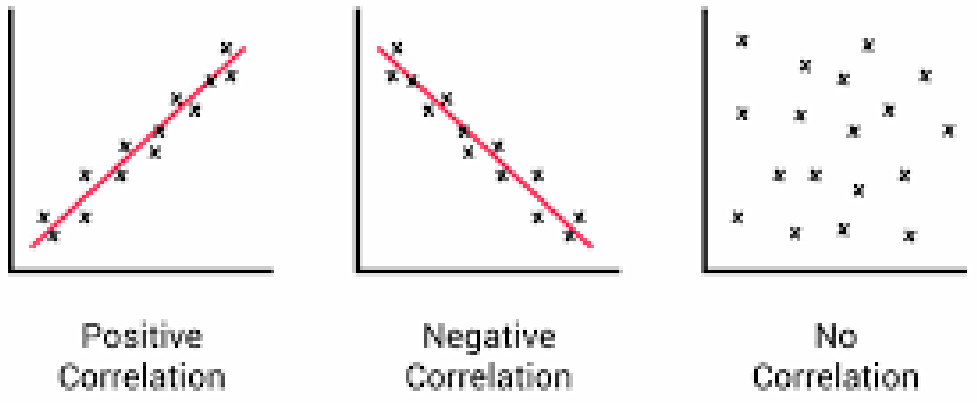
r = 1: Perfect positive correlation. It means that as one variable increases, the other variable also increases proportionally.

r = -1: Perfect negative correlation. It means that as one variable increases, the other variable decreases proportionally.

r = 0 : No correlation. It means that there is no linear relationship between the two variables.

It's important to note that correlation does not imply causation. Even if two variables are strongly correlated, it doesn't necessarily mean that changes in one variable cause changes in the other. Correlation only measures the strength and direction of the relationship between variables.

For example, consider a dataset that measures the amount of rainfall and the yield of crops in a particular region over several years. If there's a strong positive correlation between rainfall and crop yield, it means that when rainfall increases, crop yield tends to increase as well. However, correlation alone cannot tell us whether increased rainfall directly causes higher crop yields or if there are other factors at play.



Furthermore this is Explained with proper dataset in the notebook attached