STATISTICS:

Mean:

The mean, also known as the average, is a measure of central tendency that represents the typical value of a set of numbers. It is calculated by summing up all the numbers in the dataset and dividing the sum by the total number of values in the dataset.

Mathematically, the mean of n numbers is calculated as:

$$\mu = \frac{x_1 + x_2 + x_3 \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

In simpler terms, the mean is the sum of all the numbers divided by how many numbers there are. It gives us a single value that represents the "average" of the dataset, indicating its central tendency.

Properties:

1)Sensitive to Outliers:

The mean is sensitive to outliers, which are extreme values that deviate significantly from the rest of the dataset. A single outlier can disproportionately influence the value of the mean, pulling it towards the extreme value.

2)Balancing Property:

The sum of the deviations of each value from the mean is always zero. In other words, for any dataset, the sum of the differences between each value and the mean is equal to zero.

3)Minimising Deviations:

The mean is the value that minimises the sum of the squared deviations from itself. This property makes the mean a suitable measure of central tendency for many statistical analyses.

4)Unique:

The mean is unique for a given dataset. Unlike the median, which may have multiple values in the case of an even number of observations, the mean is always a single value.

Difference between μ & $\bar{x}$

$\mu \rightarrow$ population mean

$\bar{x} \rightarrow$ sample mean

Median:

The median is a measure of central tendency that represents the middle value of a dataset when it is arranged in ascending or descending order. In other words, it is the value that separates the higher half from the lower half of the dataset.

To find the median:

1)Arrange the dataset in ascending or descending order.

2)If the number of values in the dataset is odd, the median is the middle value.

3)If the number of values is even, the median is the average of the two middle values.

For example, in the dataset [2, 3, 6, 7, 10], the median is 6 because it is the middle value.

In the dataset [1, 3, 5, 6], the median is (3 + 5) / 2 = 4 because there are an even number of values, and the median is the average of the two middle values.

Properties:

1)Resistant to Outliers:

Unlike the mean, the median is resistant to outliers or extreme values in the dataset. It is not influenced by extreme values and provides a more robust measure of central tendency in the presence of outliers.

2)Balancing Property:

When the dataset is divided into two equal parts at the median, the sum of the ranks of values below the median is equal to the sum of the ranks of values above the median.

3)Unique or Unique Range:

The median may be a single value if the number of observations is odd, or it may be an interval (range) if the number of observations is even. In the latter case, the range of values between the two middle values represents the median.

4)Simple to Calculate:

The median is easy to calculate and interpret, especially for datasets with a large number of observations. It involves sorting the data and identifying the middle value(s).

5)Applicable to Skewed Distributions:

The median is a suitable measure of central tendency for skewed distributions, where the mean may not accurately represent the central tendency due to the influence of extreme values.

6)Median of Medians:

The median of medians is a concept used in algorithms for finding the median of a dataset efficiently. It involves dividing the dataset into smaller subsets and finding the median of each subset.

Mode:

The mode is a measure of central tendency that represents the most frequently occurring value in a dataset. In other words, it is the value that appears most often.

For example, in the dataset [2, 3, 3, 5, 5, 5, 7, 7, 9], the mode is 5 because it appears more frequently than any other value.

It’s important to note that a dataset can have:

1)No mode: If all values occur with the same frequency.

2)One mode: If there is a single value that occurs more frequently than any other.

3)Multiple modes: If two or more values occur with the same highest frequency.

The mode is particularly useful for categorical data, but it can also be applied to numerical data to identify the most common value or category within the dataset.

Standard Deviation:

Standard deviation is a measure of the dispersion or spread of a set of values in a dataset. It tells us how much the values in the dataset deviate or vary from the mean (average) value.

In essence, the standard deviation tells us how much the values in the dataset are spread out around the mean. A low standard deviation means the values are close to the mean, while a high standard deviation means the values are more spread out.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$

where,

σ = standard deviation

x_i = values of dataset

μ = population mean

n = total no. of obseavation

Variance:

Variance is a statistical measure that quantifies the spread or dispersion of a set of values in a dataset. It represents the average of the squared differences between each value and the mean (average) of the dataset.

$$\mathrm{Var}(x) = \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

Quartiles:

Quartiles are values that divide a dataset into four equal parts, each representing 25% of the data. They are used to understand the distribution of data and identify the central tendency, spread, and skewness of the dataset.

There are three quartiles:

First Quartile (Q1):

Q1 represents the value below which 25% of the data falls. It is the median of the lower half of the dataset.

Second Quartile (Q2):

Q2 represents the median of the dataset, dividing it into two halves. It is the value below which 50% of the data falls.

Third Quartile (Q3):

Q3 represents the value below which 75% of the data falls. It is the median of the upper half of the dataset.

Quartiles are often used in box plots and descriptive statistics to visualise the spread and variability of data. They provide insights into the range of values and the distribution of data points within the dataset.

IQR(Inter Quartile Range)

The Interquartile Range (IQR) is a measure of statistical dispersion that quantifies the spread of data within a dataset. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). Here's how to calculate the IQR:

Calculate Q1 (First Quartile):

Q1 is the median of the lower half of the dataset. It represents the value below which 25% of the data falls.

Calculate Q3 (Third Quartile):

Q3 is the median of the upper half of the dataset. It represents the value below which 75% of the data falls.

Calculate IQR:

IQR = Q3 - Q1

In other words, the IQR represents the range of values that cover the middle 50% of the dataset. It provides a measure of the variability of the data that is less sensitive to outliers compared to the range.

The IQR is commonly used in descriptive statistics and data analysis for several purposes:

Identifying Outliers:

Outliers are data points that lie significantly above or below the bulk of the data. The IQR can be used to identify outliers by defining a range outside of which data points are considered outliers. Typically, data points that fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are considered outliers.

Box Plots:

The IQR is used to create box plots, which visually represent the central tendency and spread of a dataset. The box in a box plot represents the IQR, with the median marked inside the box and the “whiskers” extending to the minimum and maximum values within a certain range (often 1.5 times the IQR).

Comparing Variability:

By comparing the IQRs of different datasets or groups within a dataset, you can assess differences in variability. A larger IQR indicates greater variability in the data, while a smaller IQR indicates less variability.

Distribution:

Gaussian/Normal distribution:

The Gaussian distribution, also known as the normal distribution, is a continuous probability distribution that is widely used in statistics, mathematics, and various fields of science. It is named after the mathematician Carl Friedrich Gauss, who first described it in the early 19th century.

The Gaussian distribution is characterised by its bell-shaped curve when plotted, with the highest point at the mean (average) of the distribution. The curve is symmetric around the mean, and its shape is determined by two parameters: the mean μ (mu) and the standard deviation σ (sigma). The mean represents the centre of the distribution, while the standard deviation measures the spread or dispersion of the values around the mean.

The probability density function (PDF) of the Gaussian distribution is gi

$$P(X=x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{(x_i-\bar{x})^2}{2\sigma^2}\right]}$$

where,

$\bar{x}$ = mean of the distribution

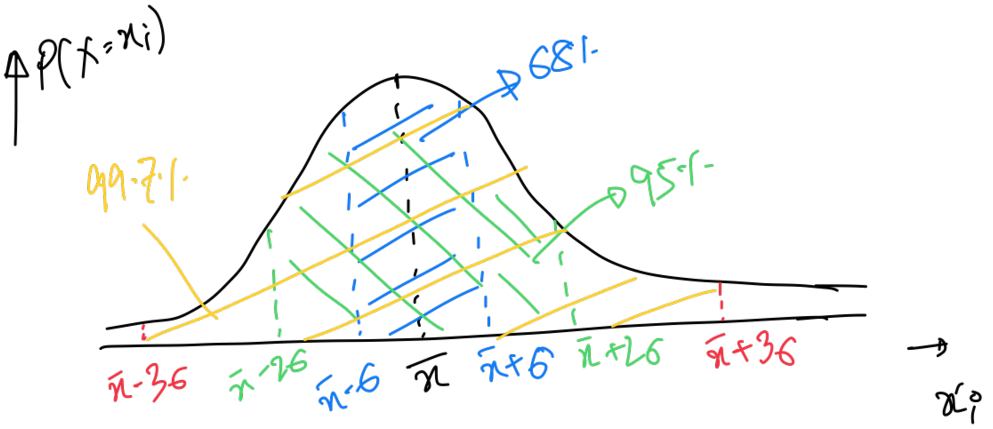
σ = standard deviation of the distribution

X = random variable

e = 2·718

Key properties of the Gaussian distribution include:

Symmetry: The distribution is symmetric around the mean, with approximately 68% of the data falling within one standard deviation of the mean, 95% falling within two standard deviations, and 99.7% falling within three standard deviations.



Central Limit Theorem: The Gaussian distribution arises naturally as the limiting distribution of the sum of independent and identically distributed random variables, according to the Central Limit Theorem.

Ubiquity: The Gaussian distribution is ubiquitous in nature and arises in various contexts, including measurements of physical quantities, errors in experimental data, financial markets, and many more.

Maximum Entropy: Under certain conditions, the Gaussian distribution maximises entropy among all probability distributions with specified mean and variance, making it a natural choice in statistical inference.

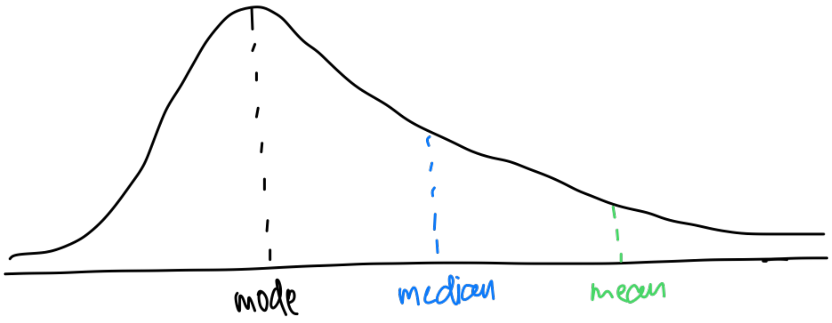
Skewness:

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. In simpler terms, it quantifies the degree to which the distribution of data deviates from symmetry around its mean.

There are two main types of skewness:

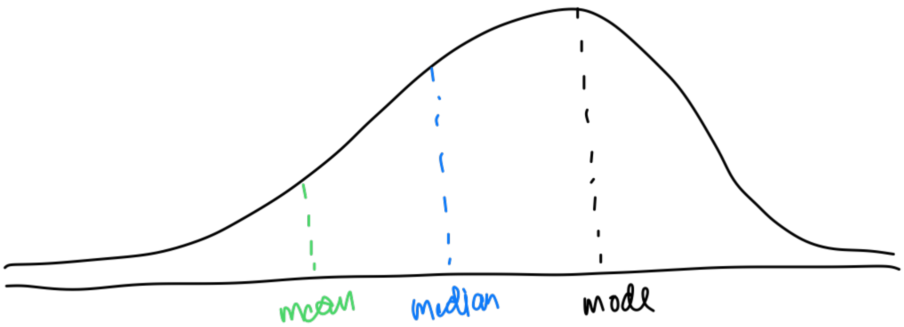
Positive Skewness (Right Skewness):

In a positively skewed distribution, the tail on the right-hand side of the distribution is longer or fatter than the left-hand side. This means that there are more extreme values on the right side of the distribution, pulling the mean (average) towards the higher end of the scale. The median is typically less than the mean, and the mode may be much less than the median.



Negative Skewness (Left Skewness):

In a negatively skewed distribution, the tail on the left-hand side of the distribution is longer or fatter than the right-hand side. This means that there are more extreme values on the left side of the distribution, pulling the mean towards the lower end of the scale. The median is typically greater than the mean, and the mode may be much greater than the median.



Kurtosis:

Kurtosis is a measure that describes the “tailedness” or shape of the probability distribution of a real-valued random variable. In simpler terms, it quantifies how peaked or flat a distribution is compared to a normal distribution.

Leptokurtic Distribution (Positive Kurtosis):

A leptokurtic distribution has “fat tails” and a peak that is higher and sharper than that of a normal distribution. This means that it has more extreme values (both large and small) compared to a normal distribution. Positive kurtosis indicates a distribution with heavy tails and a sharp peak.

Mesokurtic Distribution (Zero Kurtosis):

A mesokurtic distribution has a shape similar to that of a normal distribution. It has moderate tails and a moderate peak. The kurtosis of a normal distribution is typically defined as zero.

Platykurtic Distribution (Negative Kurtosis):

A platykurtic distribution has “lighter tails” and a flatter peak compared to a normal distribution. This means that it has fewer extreme values and a more spread-out shape. Negative kurtosis indicates a distribution with light tails and a flat peak.

Standard Normal Variate:

A standard normal variate refers to a random variable that follows a standard normal distribution. The standard normal distribution, also known as the Z-distribution, is a specific type of normal distribution with a mean (μ) of 0 and a standard deviation (σ) of 1.

The probability density function (PDF) of a standard normal distribution is given by:

Where:

(z) represents the standard normal variate.

(e) is the base of the natural logarithm (approximately equal to 2.71828).

The standard normal distribution is symmetric around its mean of 0 and has a bell-shaped curve. It is widely used in statistical analysis and hypothesis testing, particularly in the context of Z-tests and Z-scores.

When working with a standard normal variate, statisticians often use Z-scores, which represent the number of standard deviations a data point is from the mean of a standard normal distribution. A Z-score can be calculated using the formula:

$$Z = \frac{x - \mu}{\sigma}$$

Where:

is the value of the data point.

is the mean of the distribution (which is 0 for the standard normal distribution).

is the standard deviation of the distribution (which is 1 for the standard normal distribution).

Z-scores allow researchers to compare and analyse data points from different normal distributions on a standardised scale. They are used in various statistical tests and calculations, including hypothesis testing, confidence intervals, and outlier detection.

Kernel Density Estimator:

A kernel density estimator (KDE) is a non-parametric method used to estimate the probability density function (PDF) of a random variable based on a sample of data points. Unlike parametric methods, which assume a specific functional form for the underlying distribution (such as the normal distribution), KDE does not make any assumptions about the shape of the distribution.

How do a kernel density estimator works:

Data Points:

Given a set of data points from a random variable, the goal is to estimate the underlying probability density function.

Kernel Function:

A kernel function is chosen. This is a smooth, symmetric, and non-negative function, typically with a peak at 0.

Estimation at a Point:

To estimate the PDF at a point , the kernel function is centered at and scaled by a bandwidth parameter . The kernel function is then evaluated at each data point, and the results are summed up.

Kernel Weighting:

Each data point contributes to the estimation of the PDF based on the value of the kernel function evaluated at that point. Points closer to have a higher weight in the estimation.

Bandwidth Selection:

The bandwidth parameter controls the smoothness of the estimated PDF. A smaller bandwidth results in a more variable and detailed estimate, while a larger bandwidth results in a smoother estimate. The bandwidth needs to be carefully chosen to balance bias and variance in the estimation.

Final Estimation:

The estimated PDF at point is obtained by normalising the sum of kernel function evaluations by the total number of data points and the bandwidth.

Central Limit Theorem:

The Central Limit Theorem (CLT) is a fundamental concept in statistics that states that the distribution of the sum (or average) of a large number of independent and identically distributed random variables approaches a normal distribution, regardless of the shape of the original distribution.

Eg: Let say distribution of weight of india, which is not following Gaussian distribution.

$S_i \to$ sample taken from the population

$$\begin{tikzcd}
S_1 \arrow[d] & S_2 \arrow[d] & S_3 \arrow[d] & S_4 \arrow[d] & \cdots & S_m \arrow[d] \\
\mu_1 & \mu_2 & \mu_3 & \mu_4 & & \mu_n
\end{tikzcd}$$

$\mu_1 \to$ mean of respective samples.

then, μ_i will follow gaussian distribution

with $\vec{\mu_i} \to N\left(\mu, \frac{\sigma}{n}\right)$

$\mu \to$ mean of population

$\sigma \to$ standard deviation of population

$n \to$ total data points of population

Chebyshev’s Inequality:

Chebyshev’s inequality is a generalised theory that provides an upper bound on the probability that a random variable deviates from its mean by more than a certain number of standard deviations. It applies to any probability distribution, regardless of its shape or characteristics.

$$P(|X-\mu| \geq K\sigma) \leq \frac{1}{K^2}$$

where, $X \to$ random variable

$\mu \to$ population mean

$\sigma \to$ standard deviation

$K \to$ +ve real no.

Log Normal distribution:

The lognormal distribution is a probability distribution of a random variable whose logarithm is normally distributed. In other words, if is a log-normally distributed random variable, then follows a normal distribution.

The lognormal distribution is skewed to the right (positively skewed) because the logarithm of a positive number is negative or zero.

As a result, the lognormal distribution has a long right tail.

Applications:

The lognormal distribution is commonly used to model data that are inherently positive and skewed, such as:

Prices of financial assets (e.g., stocks, commodities).

Sizes of populations (e.g., human populations, animal populations).

Time to failure of mechanical systems.

Sizes of particles in aerosols.

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{[(\ln(x)-\mu)]^2}{2\sigma^2}}$$

Converting log-normal to Gaussian dist.

$x_i \rightarrow$ random variable

$x_1, x_2, x_3 \; . \; — — — \; x_n$

takig log i.e

$\ln(x_1), \ln(x_2), \ln(x_3) \; . \; — . \; \ln(x_n)$

will follow Gaussian distribution.

It is also used in financial mathematics, risk management, and reliability engineering.

Confidence Interval (C.I)

A confidence interval is a range of values derived from sample data that is likely to contain the true population parameter with a certain level of confidence. It is a statistical measure used to estimate the range within which the true value of a population parameter, such as the population mean or proportion, is likely to fall.

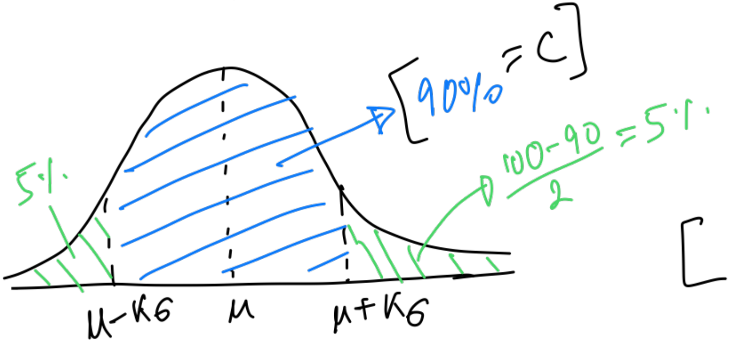
Eg: weight of a sample is

$\{180, 162, 158, 172, 168, 150, 171, 183, 165, 176\}$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 168.5 \qquad \sigma = 6.4$$

$$\underbrace{\mu}_{\text{popl}^n\text{ mean}} \in \underbrace{[162.1, 174.9]}_{\text{Interval}} \text{ with } \underbrace{95\%}_{\text{Confidence}} \text{ probability}$$

Let say for a Gaussian/Normal distribution



[,] with 90%.

The above curve shows that a specific interval contains this much %age of confidence.

Bernoulli Distribution:

The Bernoulli distribution is a discrete probability distribution that models a single experiment or trial with two possible outcomes i.e success (usually denoted by 1) and failure (usually denoted by 0).

Eg. Tossing a coin

$$P(X=x) = \begin{cases} p & , x=1 \\ 1-p & , x=0 \end{cases}$$

where, p = probability of getting 1

Binomial distribution:

The binomial distribution is a discrete probability distribution that describes the number of successes in a fixed number of independent Bernoulli trials, where each trial has only two possible outcomes: success or failure.

$$P(X=k) = {}^{n}C_{k}\,(p)^{k}\,(1-p)^{n-k}$$

where,

p = probability of success

$1-p$ = probability of failure

k = no. of successes

$n-k$ = no. of failure

$${}^{n}C_{k} = \frac{n!}{(n-k)!\,k!}$$

Hypothesis Testing:

Hypothesis testing is a statistical method used to make decisions or draw conclusions about population parameters based on sample data. It involves comparing the observed data to what would be expected under a specific assumption (null hypothesis) and determining whether there is enough evidence to reject that assumption in favour of an alternative hypothesis.

Null Hypothesis (H0): Represents the default assumption. It states that there is no effect or no difference.

Alternative Hypothesis (H1): Represents what the researcher is trying to find evidence for. It states the opposite of Null hypothesis.

Significance level which represents the probability of incorrectly rejecting the null hypothesis when it is actually true. Commonly used significance levels include 0.05, 0.01, or 0.10.

Example:

A pharmaceutical company develops a new drug intended to lower blood pressure in patients with hypertension. Before the drug can be approved for widespread use, it must undergo rigorous testing to determine its effectiveness compared to existing treatments.

Null Hypothesis (): The new drug treatment has no effect on lowering blood pressure in patients.

Alternative Hypothesis (): The new drug treatment is effective in lowering blood pressure in patients.

Steps of Hypothesis Testing:

Formulating Hypotheses:

: The mean reduction in blood pressure after treatment with the new drug is equal to 0.

: The mean reduction in blood pressure after treatment with the new drug is greater than 0.

Choosing a Significance Level ():

Let's choose a significance level of . This means we're willing to accept a 5% chance of incorrectly rejecting the null hypothesis.

Determining the Critical Region or Calculating the p-value:

Based on the chosen significance level and the degrees of freedom, determine the critical t-value or calculate the p-value.

Making a Decision:

If the p-value is less than 0.05, reject the null hypothesis and conclude that the new drug treatment is effective in lowering blood pressure.

If the p-value is greater than or equal to 0.05, fail to reject the null hypothesis.

Interpreting the Results:

If we reject the null hypothesis, we conclude that there is sufficient evidence to support the effectiveness of the new drug treatment in lowering blood pressure.

If we fail to reject the null hypothesis, we do not have enough evidence to conclude that the new drug treatment is effective in lowering blood pressure.