What is Statistics

Statistics is a branch of mathematics that involves collecting, analysing, interpreting, and presenting data. It provides tools and methods to understand and make sense of large amounts of data and to draw conclusions and make decisions based on the data.

In practice, statistics is used in a wide range of fields, such as business, economics, social sciences, medicine, and engineering. It is used to conduct research studies, analyse market trends, evaluate the effectiveness of treatments and interventions, and make forecasts and predictions.

Examples:

1. Business - Data Analysis(Identifying customer behavior) and Demand Forecasting
2. Medical - Identify efficacy of new medicines(Clinical trials), Identifying risk factor for diseases(Epidemiology)
3. Government & Politics - Conducting surveys, Polling
4. Environmental Science - Climate research

Types of Statistics

1. Descriptive:

Descriptive statistics deals with the collection, organization, analysis, interpretation, and presentation of data. It focuses on summarizing and describing the main features of a set of data, without making inferences or predictions about the larger population.

2. Inferential:

Inferential statistics deals with making conclusions and predictions about a population based on a sample. It involves the use of probability theory to estimate the likelihood of certain events occurring, hypothesis testing to determine if a certain claim about a population is supported by the data, and regression analysis to examine the relationships between variables.

Population Vs Sample:

Population refers to the entire group of individuals or objects that we are interested in studying. It is the complete set of observations that we want to make inferences about. For example, the population might be all the students in a particular school or all the cars in a particular city.

A sample, on the other hand, is a subset of the population. It is a smaller group of individuals or objects that we select from the population to study. Samples are used to estimate characteristics

of the population, such as the mean or the proportion with a certain attribute. For example, we might randomly select 100 students.

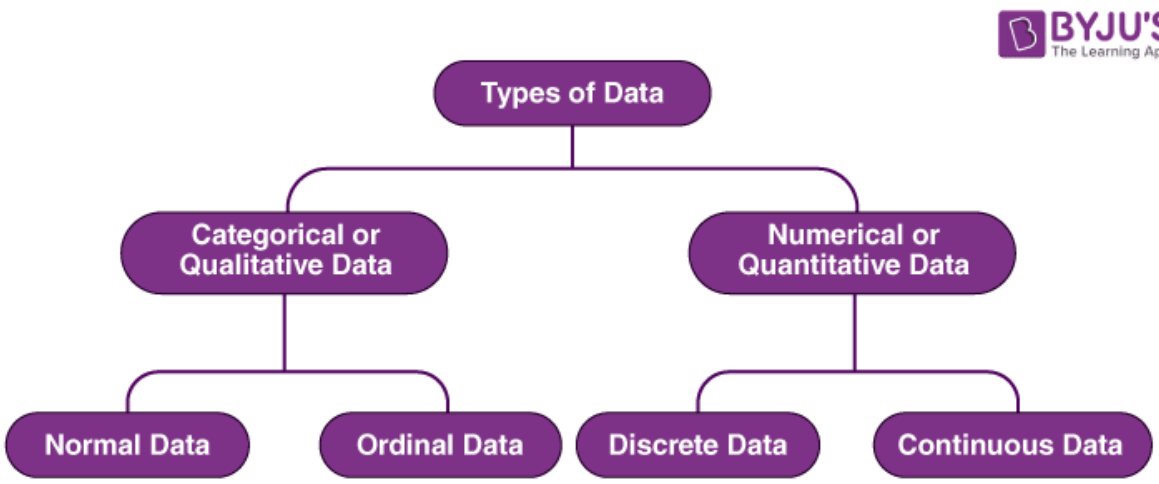
Examples:

1. All cricket fans vs fans who were present in the stadium
2. All students vs who visit college for lectures

Things to be careful about which creating samples

1. Sample Size
2. Random
3. Representative

Types of Data



Measure of Central Tendency:

A measure of central tendency is a statistical measure that represents a typical or central value for a dataset. It provides a summary of the data by identifying a single value that is most representative of the dataset as a whole.

1. Mean:

The mean is the sum of all values in the dataset divided by the number of values.

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

2. Median:

The median is the middle value in the dataset when the data is arranged in order.

3. Mode:

The mode is the value that appears most frequently in the dataset.

Note: Mode is usefull with category data or with discrete data to check frequent values.

Trimmed Mean:

A trimmed mean is calculated by removing a certain percentage of the smallest and largest values from the dataset and then taking the mean of the remaining values. The percentage of values removed is called the trimming percentage.

Example :

Values:

20,22,23,25,28,30,32,35,50,80

Actual mean : 34.5

Trimmed Values:

25,28,30,32,35

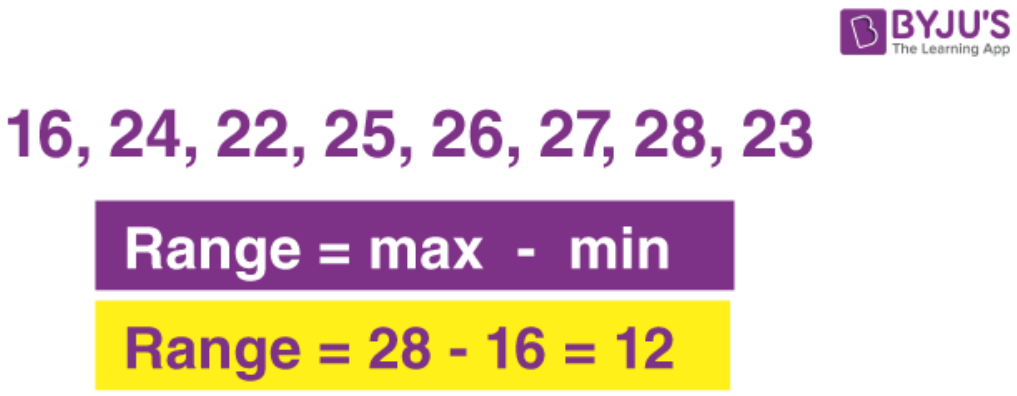
Actual mean : 30

Measure of Dispersion:

A measure of dispersion is a statistical measure that describes the spread or variability of a dataset. It provides information about how the data is distributed around the central tendency (mean, median or mode) of the dataset.

1. Range:

The range is the difference between the maximum and minimum values in the dataset. It is a simple measure of dispersion that is easy to calculate but can be affected by outliers.



2. Variance:

The variance is the average of the squared differences between each data point and the mean. It measures the average distance of each data point from the mean and is useful in comparing the dispersion of datasets with different means.

x	(x-mean)	(x-mean)^2
3	3 - 3 = 0	0
2	2 - 3 = -1	1
1	1 - 3 = -2	4
5	5 - 3 = 2	4
4	4 - 3 = 1	1

Population	Sample
$\sigma^2 = \frac{\Sigma(x_i-\mu)^2}{n}$	$S^2 = \frac{\Sigma(x_i-\bar{x})^2}{n-1}$
μ - Population Average xi - Individual Population Value n - Total Number of Population σ^2 - Variance of Population	$\bar{x}$ - Sample Average x_i - Individual Population Value n - Total Number of Sample S^2 - Variance of Sample

3. Standard Deviation:

The standard deviation is the square root of the variance. It is a widely used measure of dispersion that is useful in describing the shape of a distribution.

x	(x-mean)^2
15	(15 -14)^2 = 1
17	(17 - 14)^2 = 9
13	(13 - 14)^2 = 1
11	(11 - 14)^2 = 9

Standard Deviation Formula

cuemath
THE MATH EXPERT

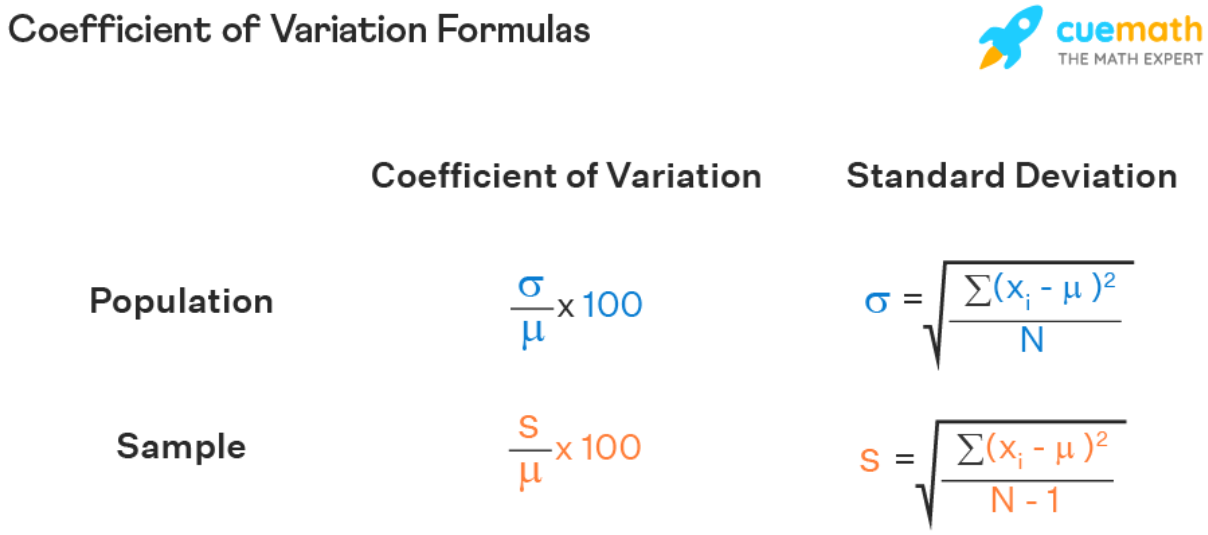
Population	Sample
$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$	$s = \sqrt{\frac{\Sigma(X-\bar{x})^2}{n-1}}$
X - The Value in the data distribution μ - The population Mean N - Total Number of Observations	X - The Value in the data distribution $\bar{x}$ - The Sample Mean n - Total Number of Observations

Coefficient of Variation:

Coefficient of Variation (CV): The CV is the ratio of the standard deviation to the mean expressed as a percentage. It is used to compare the variability of datasets with different means and is commonly used in fields such as biology, chemistry, and engineering.

The coefficient of variation (CV) is a statistical measure that expresses the amount of variability in a dataset relative to the mean. It is a dimensionless quantity that is expressed as a percentage.

The formula for calculating the coefficient of variation is:



Quantiles and Percentiles :

Quantiles are statistical measures used to divide a set of numerical data into equal-sized groups, with each group containing an equal number of observations.

Quantiles are important measures of variability and can be used to: understand distribution of data, summarize and compare different datasets. They can also be used to identify outliers.

There are several types of quantiles used in statistical analysis, including:

1. Quartiles: Divide the data into four equal parts, Q1 (25th percentile) Q2 (50th percentile or median), and Q3 (75th percentile).
2. Deciles: Divide the data into ten equal parts, D1 (10th percentile), D2 (20th percentile), ..., D9 (90th percentile).
3. Percentiles: Divide the data into 100 equal parts, P1 (1st percentile), P2 (2nd percentile), ..., P99 (99th percentile).
4. Quintiles: Divides the data into 5 equal parts

Things to remember while calculating these measures:

1. Data should be sorted from low to high
2. You are basically finding the location of an observation
3. They are not actual values in the data
4. All other tiles can be easily derived from Percentiles

Percentile :

A percentile is a statistical measure that represents the percentage of observations in a dataset that fall below a particular value. For example, the 75th percentile is the value below which 75% of the observations in the dataset fall.

Formula to calculate the percentile value:

$$PL = \frac{P}{100}(N + 1)$$

PL = the desired percentile value location
N = the total number of observations in the dataset
p = the percentile rank (expressed as a percentage)

Example: Find the 75th percentile score from the below data :

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

5 number summary:

The five-number summary is a descriptive statistic that provides a summary of a dataset. It consists of five values that divide the dataset into four equal parts, also known as quartiles. The five-number summary includes the following values:

1. Minimum value:
The smallest value in the dataset.

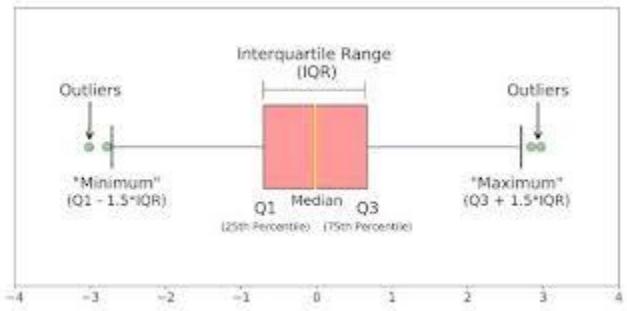
2. First quartile (Q1): The value that separates the lowest 25% of the data from the rest of the dataset.

3. Median (Q2): The value that separates the lowest 50% from the highest 50% of the data.

4. Third quartile (Q3): The value that separates the lowest 75% of the data from the highest 25% of the data.

5. Maximum value: The largest value in the dataset.

The five-number summary is often represented visually using a box plot, which displays the range of the dataset, the median, and the quartiles. The five-number summary is a useful way to quickly summarize the central tendency, variability, and distribution of a dataset.



Interquartile Range

The interquartile range (IQR) is a measure of variability that is based on the five-number summary of a dataset. Specifically, the IQR is defined as the difference between the third quartile (Q3) and the first quartile (Q1) of a dataset.

Random Variables:

A random variable is a *numerical value* that is determined by the outcome of a random experiment. It assigns a number to each possible outcome.

Example:

Toss a coin : If Heads = 1, Tails = 0
Here, the outcome of the coin toss is random, so the number we assign (0 or 1) is a random variable.

Example of rolling dice:

The experiment have random possibilities: Y=[1,2,3,4,5,6] = **Random Variables.**

Types Of Random Variables:

1. **Discrete Random Variables:**

   The Random variables are based on discrete data.

   Example: Head or Tail, Rolling Dice.

2. **Continuous Random Variables:**

   The Random variables are based on continuous data.

   Example: Height, Distance, Temperature.

Probability Distributions:

A probability distribution is a list of all of the possible outcomes of a random variable

along with their corresponding probability values.

Example: Toss of a Coin with probability

Coin Toss	**Head**	**Tail**
Probability	**1/2**	**1/2**

Problem with Distribution?

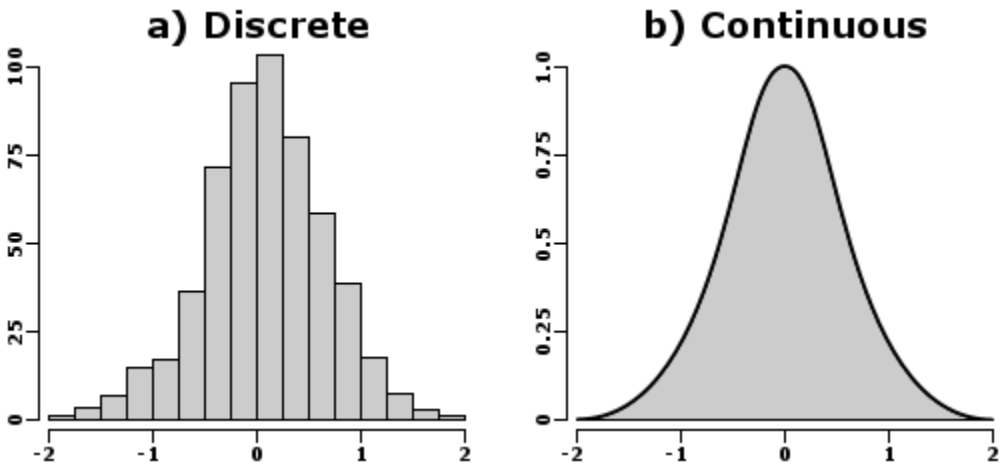
In many scenarios, the number of outcomes can be much larger and hence a table would

be tedious to write down. Worse still, the number of possible outcomes could be infinite,

Example : Height of people, Rolling 10 dice together.

So, we create a mathematical function that establishes the relationship between outcomes and probabilities which is know as probability distribution function.

Types of Probability Distributions:



Why Probability Distribution Is Important:

- Gives an idea about the shape/distribution of the data.

Probability Distribution Function:

A probability distribution function is a mathematical function that describes the probability of obtaining different values of a random variable in a particular probability distribution.

Types Of Probability Distribution Function:

1. Probability Mass Function (PMF)
2. Probability Density Function (PDF)
3. Commulitative Distribution Function of PMF
4. Commulitative Distribution Function of PDF

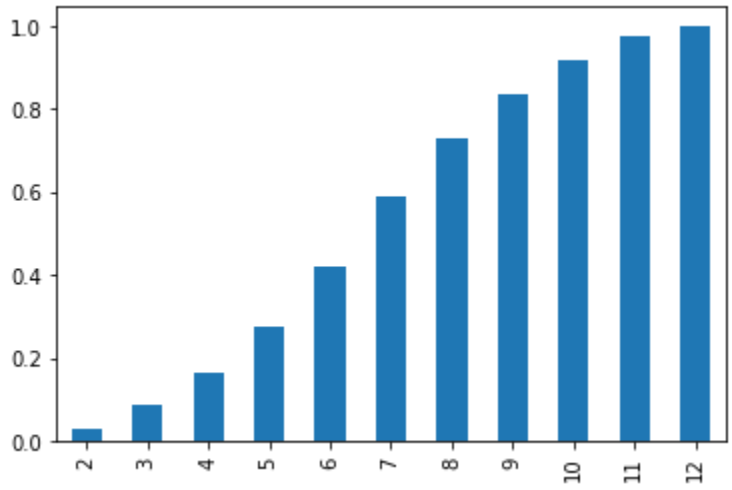
Probability Mass Function (PMF)

PMF stands for Probability Mass Function. It is a mathematical function that describes the probability distribution of a discrete random variable. The PMF of a discrete random variable assigns a probability to each possible value of the random variable. The probabilities assigned by the PMF must satisfy two conditions:

1. The probability assigned to each value must be non-negative (i.e., greater than or equal to zero).
2. The sum of the probabilities assigned to all possible values must equal 1.

Commulitative Distribution Function(CDF) of PMF:

A CDF is a function that shows the probability that a random variable can take a value up to x. We keep adding the PMF probabilities step by step, and that gives us the CDF.



Probability Density Function (PDF)

PDF stands for Probability Density Function. It is a mathematical function that describes the probability distribution of a continuous random variable.

Why Probability density instead of Probability mass function:

For continuous random variables, the probability of any exact value is 0 or closer to 0. To overcome this, we use the probability density function (PDF), which allows us to calculate the probability over an interval.

Density In PDF:

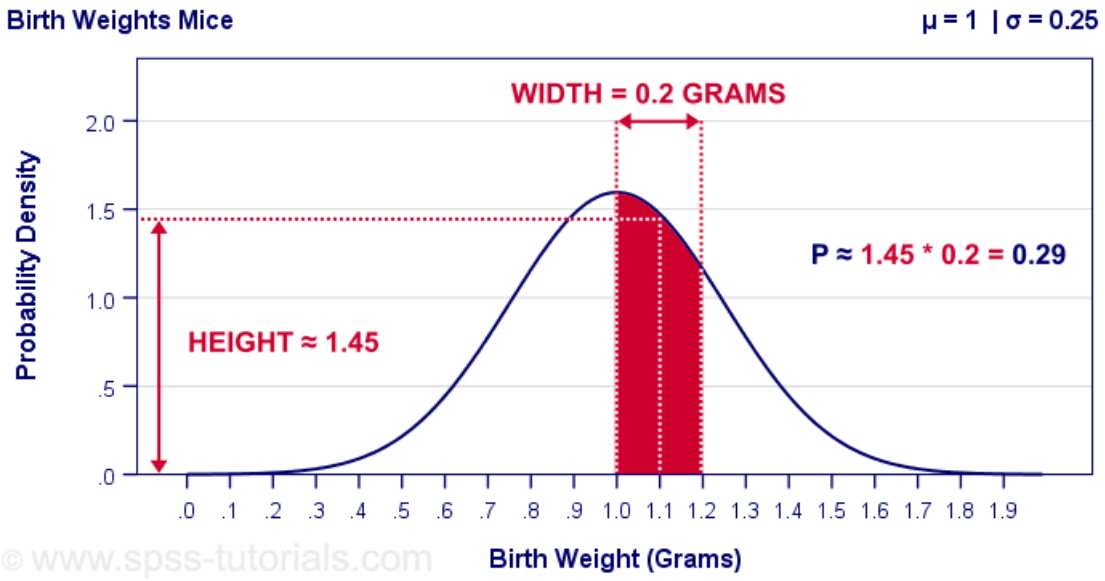
Density is a function that shows how probability is spread out over the values of a continuous random variable.

Example:

The probability that a student's exact height is 165.234 cm is 0. But the chance of having a height around 165 cm can be higher or lower. This intensity of chance is what we call density.

Formula : $\text{Density} = \frac{Count}{N \times Bin\ width}$ N = Total no of values

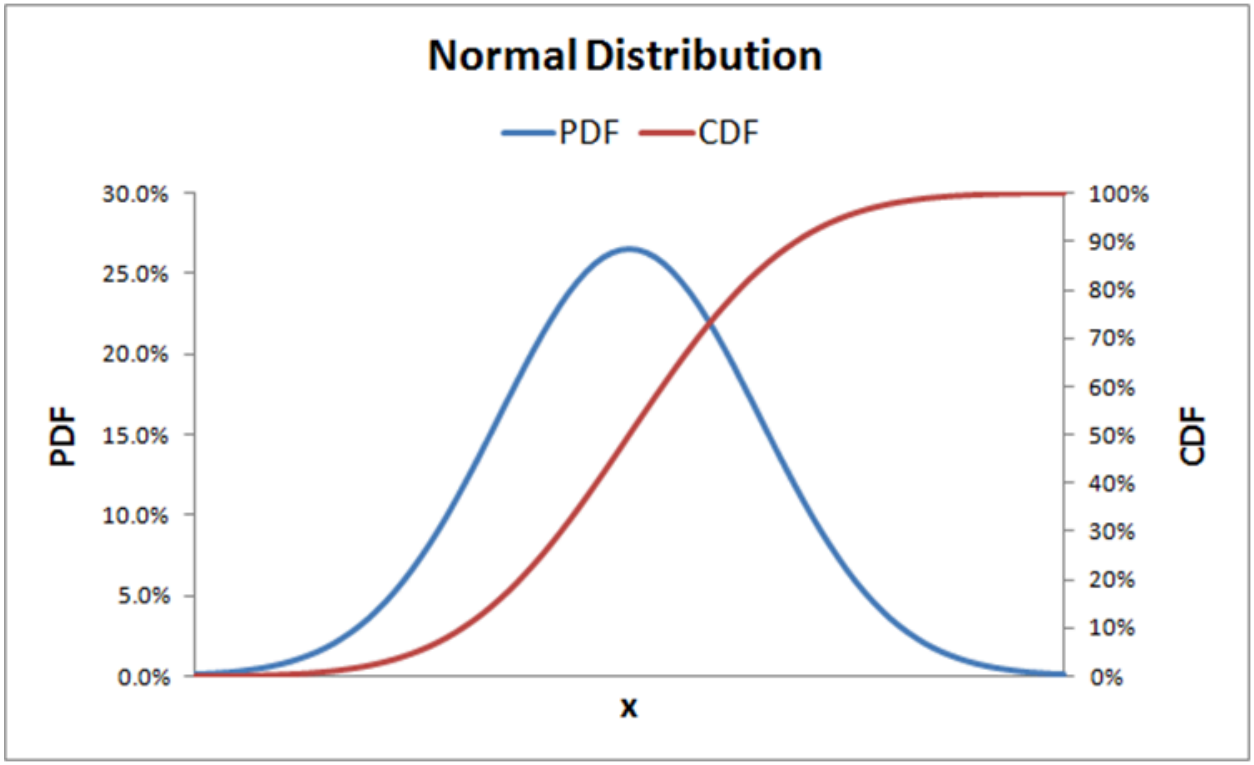
Present The Actual Probability Finding:



Probability Formula : **Prob=Density×bin width**

Commulitative Distribution Function of PDF:

A CDF is a function that shows the probability that a continuous random variable can take a value up to x. We keep adding (integrating) the PDF values step by step over the interval, and that gives us the CDF.

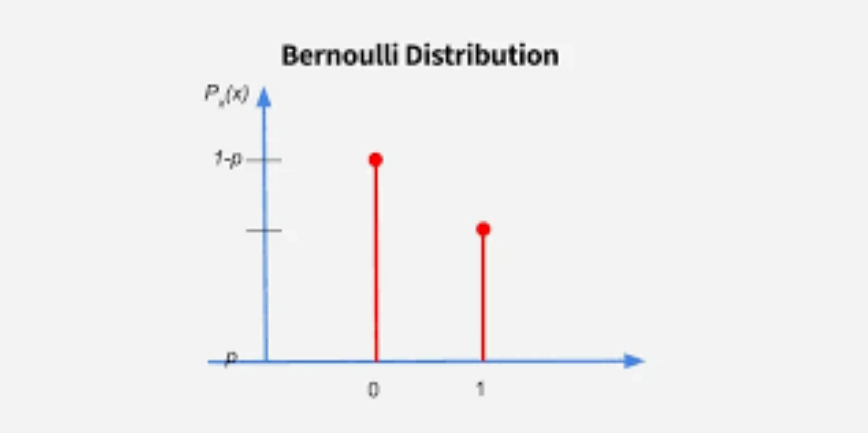


PMF Common Distributions:

1. Bernoulli Distribution:

   A Bernoulli distribution models a single trial with only two outcomes — success (1) or failure (0).

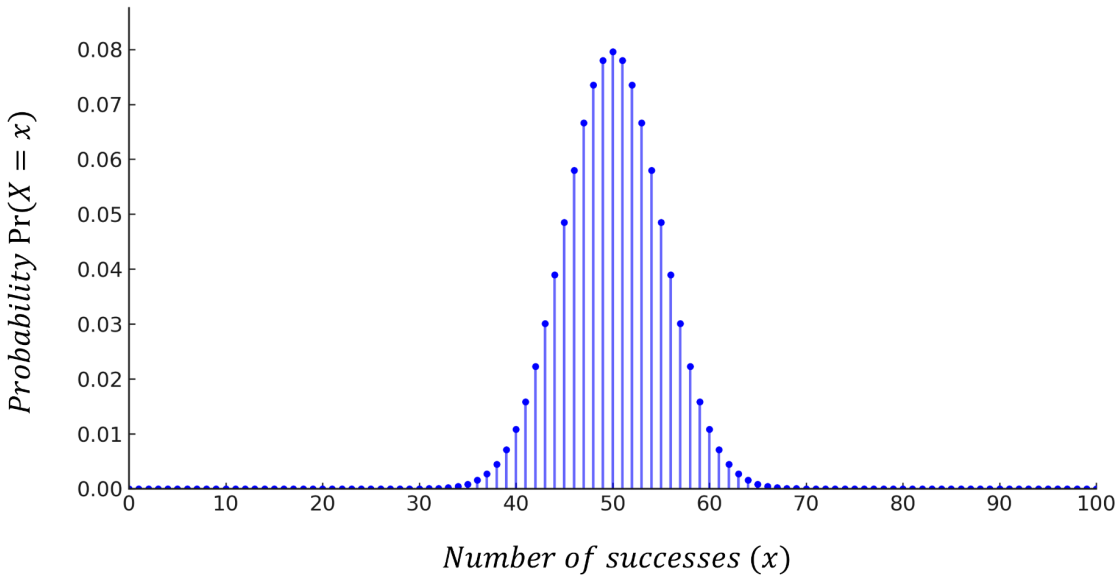
   Example: `Clickadd one time` : 1 = clicked , 0 = not clicked



2. Binomial Distribution:

Binomial distribution models the **number of successes in n independent Bernoulli trials**.

Example : 50 students Exam: (Pass , Fail)



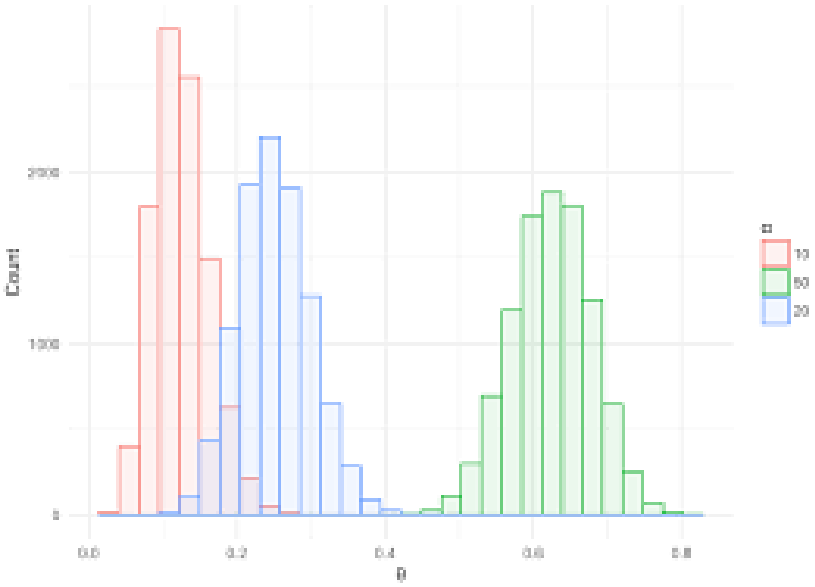
3. Multinomial Distribution:

A multinomial distribution models multiple independent trials, where each trial can have more than two possible outcomes, and it describes the counts of each outcome over all the trials.

Example :

`DiceRollCounts` : counts of 1, 2, 3, 4, 5, 6 when dice is rolled n times

`SurveyResponses` : counts of responses: Yes, No, Maybe



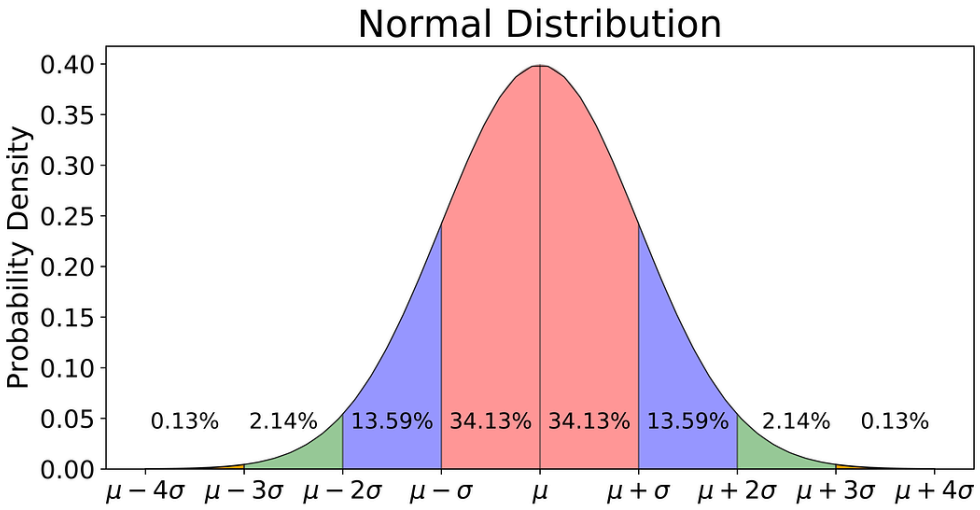
PDF Common Distributions:

1. Normal Distribution:

A normal distribution is a continuous probability distribution that is symmetric around its mean, forming a bell-shaped curve, where most of the data is concentrated around the mean, and probabilities for values further away from the mean decrease smoothly.

Characteristics:

- Symmetric around mean (μ)
- Mean = Median = Mode



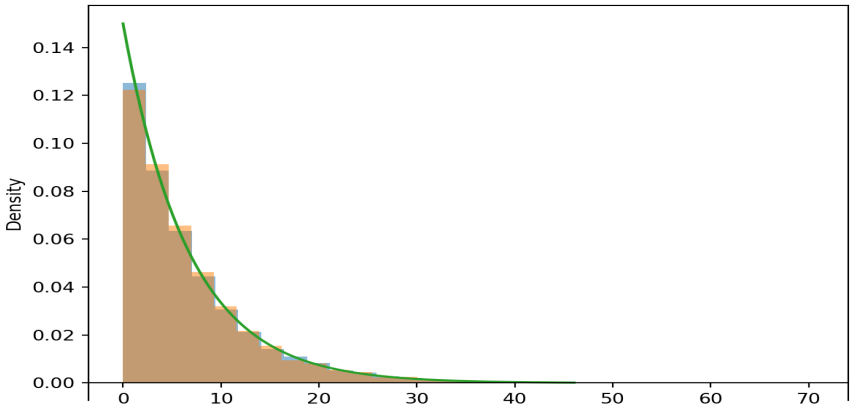
2. Exponential Distribution:

Exponential distribution is a continuous probability distribution that models the time between events, where events occur continuously and independently at a constant average rate.

It is right-skewed, with most values near zero and a long tail extending to the right.

Example:

- Time between arrivals of customers at a bank.
- Waiting time in a queue.
- Mean > Median > Mode.



Sampling Distribution

Sampling distribution is a probability distribution that describes the statistical properties of a sample statistic (such as the sample mean or sample proportion) computed from multiple independent samples of the same size from a population.

- Population : The entire data.
- Sample : Some part of the data.

Why Sampling Distribution is important?

Sampling distribution is important in statistics because it allows us to estimate the variability of a sample statistic, which is useful for making inferences about the population. By analysing the properties of the sampling distribution, we can compute confidence intervals, perform hypothesis tests, and make predictions about the population based on the sample data.

Central Limit Theorem

The Central Limit Theorem (CLT) states that the distribution of the sample means of a large number of independent and identically distributed random variables will approach a normal distribution.

The conditions required for the CLT:

1. The sample size is large enough, typically greater than or equal to 30.
2. The random variables in the sample are independent and identically distributed.

Point Estimate

A point estimate is a single value, calculated from a sample, that serves as the best guess or approximation for an unknown population parameter, such as the mean or standard deviation. Point estimates are often used in statistics when we want to make inferences about a population based on a sample.

Confidence Interval

Confidence interval, in simple words, is a range of values within which we expect a particular population parameter, like a mean, to fall. It's a way to express the uncertainty around an estimate obtained from a sample of data.

Confidence level, usually expressed as a percentage like 95%, indicates how sure we are that the true value lies within the interval.

Confidence Interval = Point Estimate ± Margin of Error

Two Types of CI :

- Confidence Interval (Z Procedure)
- Confidence Interval (T Procedure)

1. Confidence Interval With Z Procedure:

Assumptions :

1. Random Sampling
2. Known population Standard Daviation
3. Normal Distribution , if not normal than apply central limit theorem to make distribution normal.

Formula : CI = $\overline{X} \pm Z\alpha/2 \ \frac{\delta}{\sqrt{n}}$

Confidence Interval With T Procedure:

Assumptions :

1. Random Sampling
2. Unknown population Standard Daviation
3. Normal Distribution , if not normal than apply central limit theorem to make distribution normal.

Formula : CI = $\overline{X} \pm t\alpha/2 \ \frac{S}{\sqrt{n}}$

Scientific Reason Why CI Work

Even with one sample, the CI works because:

1. **Sample is random is unbiased:**
   Each random sample is equally likely to be above or below the population mean.

2. **CLT guarantees:**
   The distribution of sample means is predictable (approximately normal).

3. **Probability interpretation:**
   If you repeated the experiment many times, 95% of the confidence intervals calculated from different samples would contain the true population mean.

Hypothesis Testing:

A statistical hypothesis test is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis. Hypothesis testing allows us to make probabilistic statements about population parameters.

Example : You are a social media manager. You changed the thumbnail of your video to see if more people will click on it. You want to know, Does the new thumbnail increase the click rate compared to the old one?

- ### Null Hypothesis(Ho):

  The null hypothesis is a statement that assumes there is no significant effect or relationship between the variables being studied. It serves as the starting point for hypothesis testing and represents the status quo or the assumption of no effect until proven otherwise. The purpose of hypothesis testing is to gather evidence (data) to either reject or fail to reject the null hypothesis in favour of the alternative hypothesis, which claims there is a significant effect or relationship.

  Example : The new thumbnail does not increase the click rate.

- ## Alternate Hypothesis(H1):

  The alternative hypothesis, is a statement that contradicts the null hypothesis and claims there is a significant effect or relationship between the variables being studied. It represents the research hypothesis or the claim that the researcher wants to support through statistical analysis.

  Example : The new thumbnail increases the click rate.

Steps involved in Hypothesis Testing

1. Formulate a Null and Alternate hypothesis

2. Select a significance level(This is the probability of rejecting the null hypothesis when it is actually true, usually set at 0.05 or 0.01)

3. Check assumptions (make distribution, check population std, data type working with, single or multiple columns

4. Decide which test is appropriate(Z-test, T-test)

5. Conduct the test

6. Reject or not reject the Null Hypothesis.

7. Interpret the result

Example : Performing Test

Suppose a snack food company claims that their Lays wafer packets contain an average weight of 50 grams per packet. To verify this claim, a consumer watchdog organization decides to test a random sample of Lays wafer packets. The organization wants to determine whether the actual average weight differs significantly from the claimed 50 grams. The organization collects a random sample of 40 Lays wafer packets and measures their weights. They find that the sample has an average weight of 49 grams, with a known population standard deviation of 4 grams.

Covariance:

Covariance is a statistical measure that describes the degree to which two variables are linearly related. It measures how much two variables change together, such that when one variable increases, does the other variable also increase, or does it decrease?

If the covariance between two variables is positive, it means that the variables tend to move together in the same direction. If the covariance is negative, it means that the variables tend to move in opposite directions. A covariance of zero indicates that the variables are not linearly related.

Covariance Formula	
Population	**Sample**
$\sigma_{xy} = \frac{\sum(X-\mu_x)(Y-\mu_y)}{N}$	$s_{xy} = \frac{\sum(X-\bar{x})(Y-\bar{y})}{n-1}$
X, Y – The Value of X and Y in the Population *μ_x, μ_y – The population Mean of X and Y* *N – Total Number of Observations*	*X, Y – The Value of X and Y in the Sample Data* *$\bar{x}, \bar{y}$ – The Sample Mean of X and Y* *n - Total Number of Observations*

Example:

Experience (X)	**Salary (Y)**	**X -** $\overline{X}$	**Y -** $\overline{Y}$	**(X -** $\overline{X}$**) × (Y -** $\overline{Y}$**)**
2	1			
5	2			
8	5			
12	12			
13	10			

Disadvantage of using Covariance

One limitation of covariance is that it does not tell us about the strength of the relationship between two variables, since the magnitude of covariance is affected by the scale of the variables.

Correlation:

Correlation refers to a statistical relationship between two or more variables. Specifically, it measures the degree to which two variables are related and how they tend to change together.

Correlation is often measured using a statistical tool called the correlation coefficient, which ranges from -1 to 1. A correlation coefficient of -1 indicates a perfect negative correlation, a correlation coefficient of 0 indicates no correlation, and a correlation coefficient of 1 indicates a perfect positive Correlation.

$$r = \frac{cov(x, y)}{\sigma_x \cdot \sigma_y}$$

Derived :

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

r = Pearson correlation coefficient

x_i = x variable sample

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ = mean of values in y variable