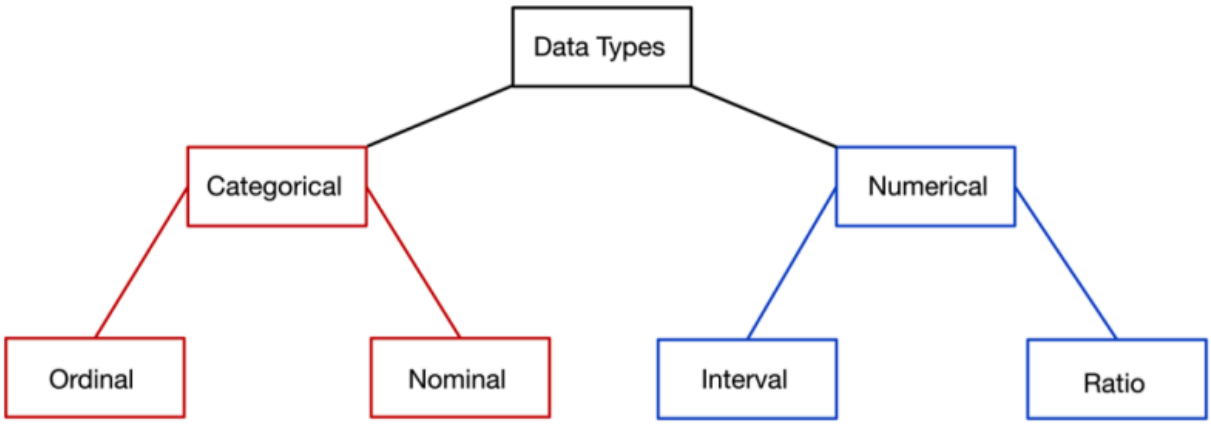
Descriptive Stats

1. Types of data



Categorical → Qualitative

Numerical → Quantitative

Nominal Data (Qualitative)

- **Definition**: Categories with no inherent order.
- **Examples**:
  - Types of fruits: `Apple`, `Banana`, `Cherry`
  - Colours: `Red`, `Blue`, `Green`
  - Gender: `Male`, `Female`

- State: `Haryana`, `West Bengal`

Ordinal Data (Qualitative)

- **Definition**: Categories with a meaningful order, but no consistent difference between them.
- **Examples**:
  - Movie ratings: `Poor`, `Fair`, `Good`, `Excellent`
  - Education level: `High School`, `Bachelor's`, `Master's`, `PhD`
  - Feedback: Bad, Average, Good, Awesome

Interval Data

- **Definition**: Ordered data with equal intervals, but no true zero.
- **Examples**:
  - Temperature in Celsius: `10°C`, `20°C`, `30°C`
  - Dates: `2000`, `2010`, `2020`

Ratio Data

- **Definition**: Ordered data with equal intervals and a true zero point.
- **Examples**:
  - Height in cm: `150 cm`, `160 cm`, `170 cm`
  - Weight in kg: `50 kg`, `60 kg`, `70 kg`

Discrete Data

- **Definition**: Data that can take only specific, separate values. These are countable and often whole numbers.
- **Characteristics**:
  - Cannot take fractional values.
  - Often results from counting.
- Examples:

- Number of students in a class: `25`, `30`, `32`
- Number of cars in a parking lot: `10`, `15`, `20`
- Number of goals scored in a match: `0`, `1`, `2`, `3`

Continuous Data

- **Definition**: Data that can take any value within a given range. These are measurable and can include fractions and decimals.
- **Characteristics**:
  - Infinite possibilities within a range.
  - Often results from measuring.
- Examples:
  - Height of a person: `170.5 cm`, `165.2 cm`
  - Temperature: `36.6°C`, `22.3°C`
  - Time taken to run a race: `12.45 seconds`, `10.78 seconds`
  - Weight: 35.3
  - Height: 172.3

---

2. Measure of Central Tendency

It represents typical or central value of the dataset. It provides the summary of the data by identifying a single value that is most representative of the dataset as a whole. Basically find the centre of the data.

1. Mean
2. Median
3. Mode
4. Trimmed mean
5. Weighted mean

Mean:

Formula:

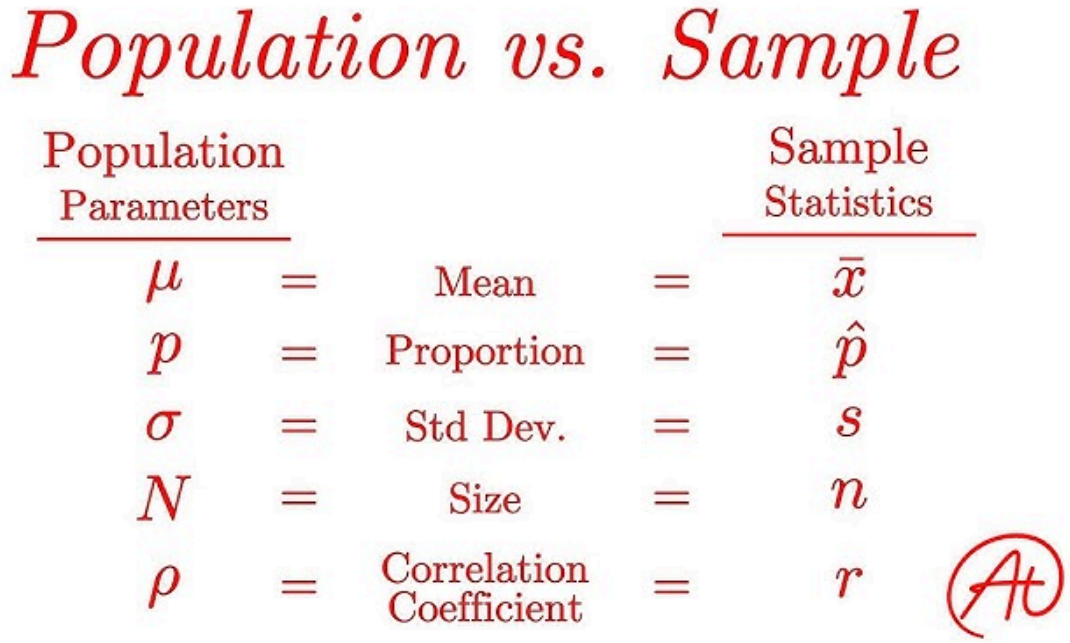
$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

where:

- xi = each value in the sample
- N = total number of values in the sample

Formula for population and sample are same.

Notation:



Example:

Sample: 12, 14, 16

$$(12 + 14 + 16)/3 = 14$$

The problem in mean is it's prone to outliers.

Median:

It's basically a middle value in my dataset, when the data is arranged in order (both ascending or descending order). Use Median whenever you have outliers instead of mean. Always focus on median because the outlier goes to last.

Notations:

Mode:

A value appears most frequently in the dataset. Use whenever you have a categorical data, it tells that which categories are so frequently, you can understand this using dot plot. You use mode for discrete data.

Notations:

Weighted Mean:

Let's say you have $[1, 2, 3]$, instead of calculating directly you assign weights to each element in the input $[0.5, 0.2, 0.3]$. The weighted mean is the sum of products of each value and it's weight, divided by the sum of weights. So,

$$1 * 0.5 + 2 * 0.2 + 3 * 0.3/0.5 + 0.2 + 0.3 = 3.2$$

Trimmed Mean:

It’s an another technique where you can remove outliers. 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.

Let’s say you specified 20% as your trimmed percentage. It removes 10% data from beginning of the data and 10% from end of the data.

---

3. 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.

a = [x, x, x]

a = [x, x, x]

In the bellow examples, assume that it gives same mean value, but the spread in the last is bigger than the first, to measure the spread we use measure of dispersion. some of the techniques are

1. Range (max - min is the range)
2. Variance
3. Standard Deviation
4. Coefficient of Variation

Range:

It’s **affected** by **outliers**, to avoid this we use Variance. We don’t use Range when we are looking for measure of dispersion.

Variance:

Variance is a 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. It's highly **prone** to **outliers**, if you square the big numbers, it will be super big.

Population:

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

Sample:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

It's not exactly spread, i**t's proportional measure of spread.** Basically you're finding what's the distance between mean to each value.

There is an another concept called **Mean Absolute Deviation** basically a same formula but it's a absolute value. You don't use this often because you can't infer the population properly. It's less prone to outliers compare to variance.

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

Standard Deviation

It's basically a square root of variance. It's a widely used measure of dispersion that is useful in describing the shape of a distribution.

Why do we need standard deviation? SD units are same as data units.

For example:

- If your data is in **centimeters**, SD is in **centimeters**.
- If your data is in **seconds**, SD is in **seconds**.

$$(5-7)^2 = 4, \quad (7-7)^2 = 0, \quad (9-7)^2 = 4 s^2 = \frac{4+0+4}{2} = 4 s = \sqrt{4} = 2$$

Population:

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

Sample:

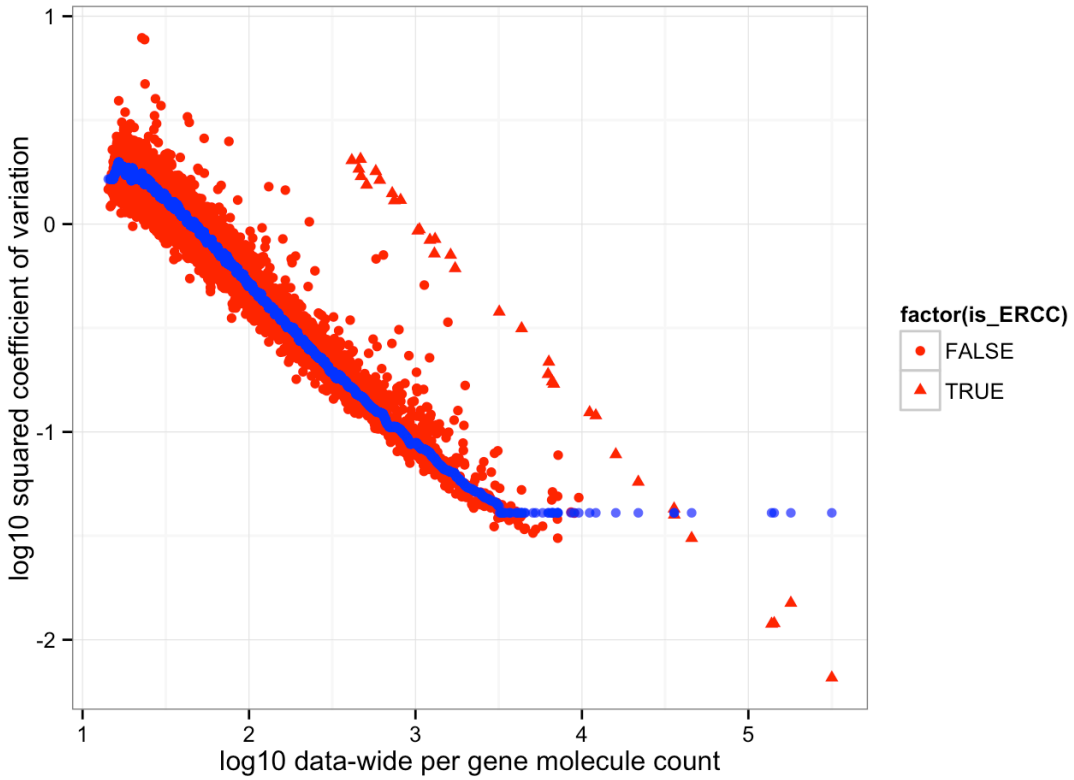
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Coefficient of Variation (CV)

Let's say you have two columns, you can't compare standard deviation of the two columns with each other (salary, experience) it's irrelevant. So you get CV which is uniformed score of how the data is spread to the mean. After getting the CV, you can compare two columns.

The core idea of CV: To **measure how much variability (spread)** a dataset has **relative to its mean**, so that you can **compare the consistency or stability** of different datasets, even if they are on **different scales or units**.

$$CV = (Standard.deviation * mean) * 100$$



4. Univariate Analysis for Categorical data

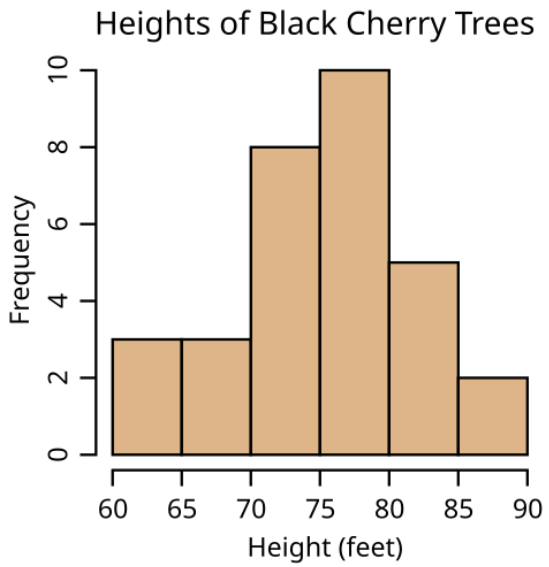
Frequency distribution table: It summarises the number of times that each value occurs in a dataset. It's for **categorical data**, once you have frequency table you can easily create a **bar graph** easily.

Relative Frequency table: After calculating the frequency of each categorical variable, you calculate proportion or percentage of the frequency. It is calculated by dividing the frequency of a category by the total number of observations in the dataset or sample. You can use this for **pie chart**.

Cumulative Frequency: Basically it's a running total of frequencies of a variable or category in a dataset or sample. It's calculated by adding up the frequencies of the current category and all previous categories in the dataset or sample. You can use OGive for this.

5. Univariate Analysis for Numerical data

In previously you calculated the frequencies for the categories or category values, instead in numerical we calculate frequencies for **bins/bucket**. Then you can create **Histogram** but there is no space between the bins/bucket but in box plot we have spaces because those are categories in nature.



Here you have to select the bucket size on your own, it's hyper-parameter in this plot. You can similarly calculate this values for relative frequency and cumulative frequency.

---

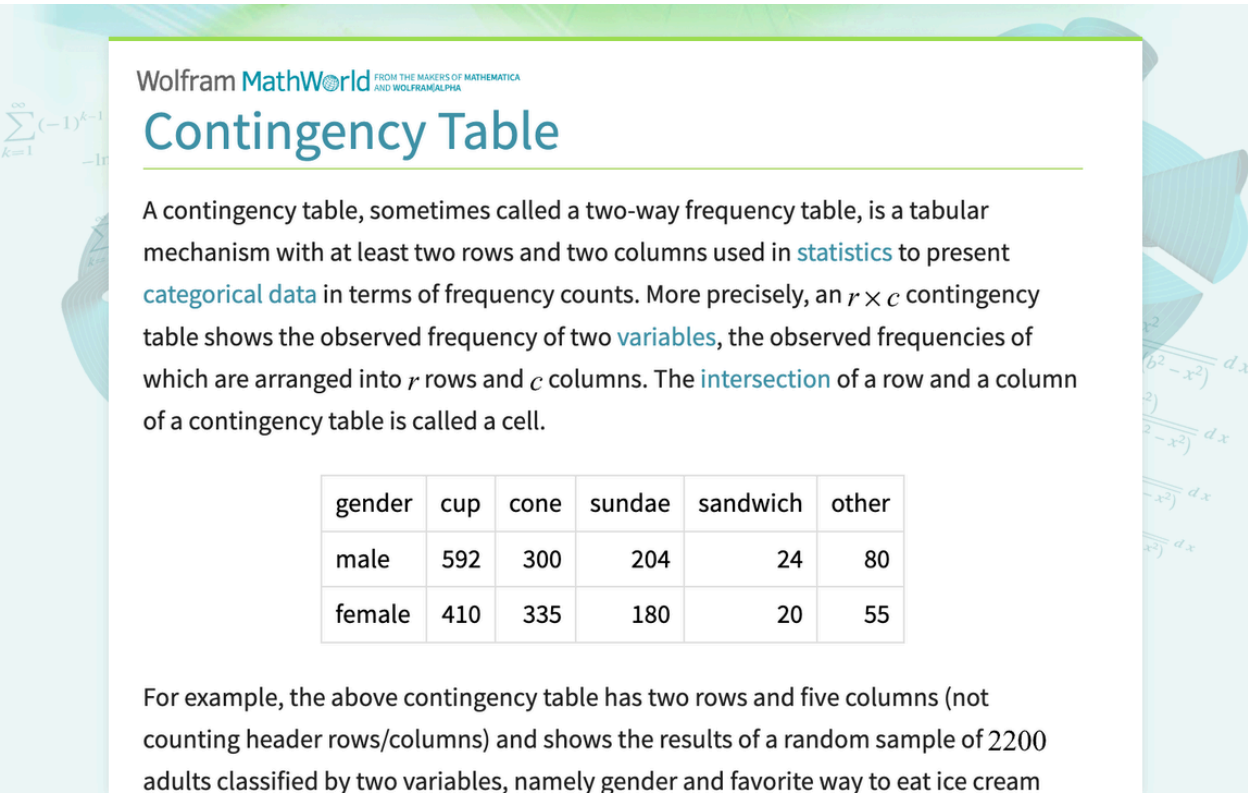
6. Bivariate Analysis Intro

There are three different possibilities you analyse Bivariate data

1. Categorical with Categorical
2. Numerical with Numerical
3. Categorical with Numerical

Categorical with Categorical plots

Contingency Table/ Cross Tab: It's used to summarise the relationship between two categorical variables. A contingency table displays the frequencies or relative frequencies of the observed values of the two variables, organised into rows and columns. It's also called two way frequency table.



Wolfram MathWorld FROM THE MAKERS OF MATHEMATICA AND WOLFRAM|ALPHA

Contingency Table

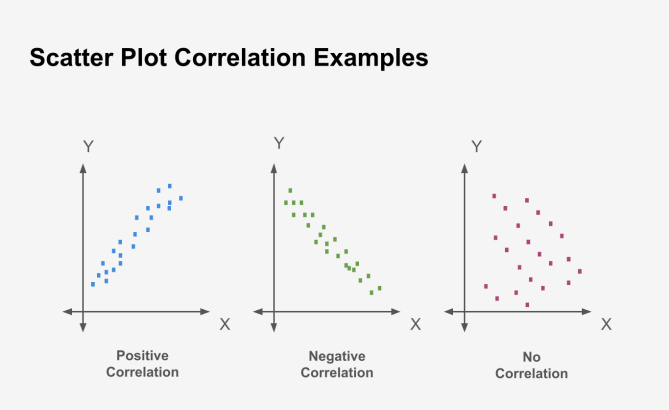
A contingency table, sometimes called a two-way frequency table, is a tabular mechanism with at least two rows and two columns used in statistics to present categorical data in terms of frequency counts. More precisely, an $r \times c$ contingency table shows the observed frequency of two variables, the observed frequencies of which are arranged into r rows and c columns. The intersection of a row and a column of a contingency table is called a cell.

gender	cup	cone	sundae	sandwich	other
male	592	300	204	24	80
female	410	335	180	20	55

For example, the above contingency table has two rows and five columns (not counting header rows/columns) and shows the results of a random sample of 2200 adults classified by two variables, namely gender and favorite way to eat ice cream

Numerical with Numerical plots

You can use **scatter plot.** From scatter plot you can understand covariance and correlation. relatino



Numerical with Categorical plots

1. **Bar chart**
2. **Contingency table** (you can convert one column to buckets)

---

7. Quantiles and Percentiles

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

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

There are different types of quantiles used in statistical analysis,

1. Quartiles: divide the data into **4** equal parts, Q1, Q2, Q3, Q4
2. Deciles: Divide the data into **10** equal parts Q1,...Q10
3. Percentiles: Divide the data into **100** equal parts.
4. Quintiles: Divide the data into **5** equal parts

- You can calculate quantiles after sorting the data and also you're basically finding the **location** of the **data point**.
- And one thing to remember all other tiles can be easily derived from Percentiles.
- They are not actual values in the data.

Percentiles

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

$$P_k = \frac{k}{100} \times (N + 1)$$

Where:

- Pk = position (rank) of the k-th percentile
- k = desired percentile (e.g., 25 for 25th percentile)
- N = total number of observations

Step 1: Sort the data (ascending order)

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

$$r = \frac{75}{100} \times (10 + 1) = 8.25$$

It means, it's between 8th or 9th index which is 96, 98. so to calculate the accurate we have to find the distance

Step 2: find the distance

$96 + 0.25(98 - 96) = 96.5$

0.25 is coming from 8.25. So 96.5 is the 75th percentile.

Percentiles to Values

$$P = x + \frac{0.5y}{n}$$

- x: the lower data value (below the percentile position)
- y: number of values equal to the given value
- n: the number of total observations (or sometimes a scaling factor)
- 0.5y: adjusts for the fractional position between ranks

---

8) 5 Number Summary

1. Minimum value
2. Q1
3. Median Q2
4. Q3
5. Maximum value

The five number summary is often represented visually using a box plot, which displays the range of the dataset, the median and the quartiles. It's used to quickly summarise the **central tendency**, **variability** and **distribution** of a dataset.

IQR = [Q3 - Q2] (basically it's centre 50% of data)

Box plot is also known as box-and-whisker plot. It shows the 5 Number summary.

Outlier detection:

$$\text{Min} = Q_1 - 1.5 \times IQR$$
$$\text{Max} = Q_3 + 1.5 \times IQR$$

---

9. 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?

Basically if you two variables it finds whether it has linear relationship or not whether it's **positive** or **negative** or **zero**.

If it's positive, it means that both variables tend to move together in the same direction.

If it's negative, it means that both variable tend to move in opposite direction.

If it's 0, that means variables are not linearly related.

Population Formula

$$\sigma_{xy} = \frac{\sum(X - \mu_x)(Y - \mu_y)}{N} \tag{1}$$

mu → population mean

N → total number of population observation

Sample Formula

$$S_{xy} = \frac{\sum(X - \bar{x})(Y - \bar{y})}{n - 1} \tag{2}$$

x bar and y bar → sample mean

n → total number of sample observations

X, Y → respective value from variables or features

Why We Need It?

1. **To detect relationships between variables**
    - Covariance tells us **direction** (positive or negative relationship). **Mean** and **variance** can't.
    - Example: revenue vs. marketing spend, height vs. weight.
2. **Used in machine learning and finance**
    - **Covariance matrix** helps measure how features co-vary → critical for PCA (Principal Component Analysis).

Dis advantage of Covariance:

1. Basically it gives the linear relationship of two variables like positive, negative or zero but 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**.

2. In simple terms, let's say a = 30, b = 30 both are having same correlation but b has more spread in the data compare to the a, we can't capture this in covariance.
3. Mostly no one uses covariance in data science, people uses something called correlation.

Covariance of a variable itself?

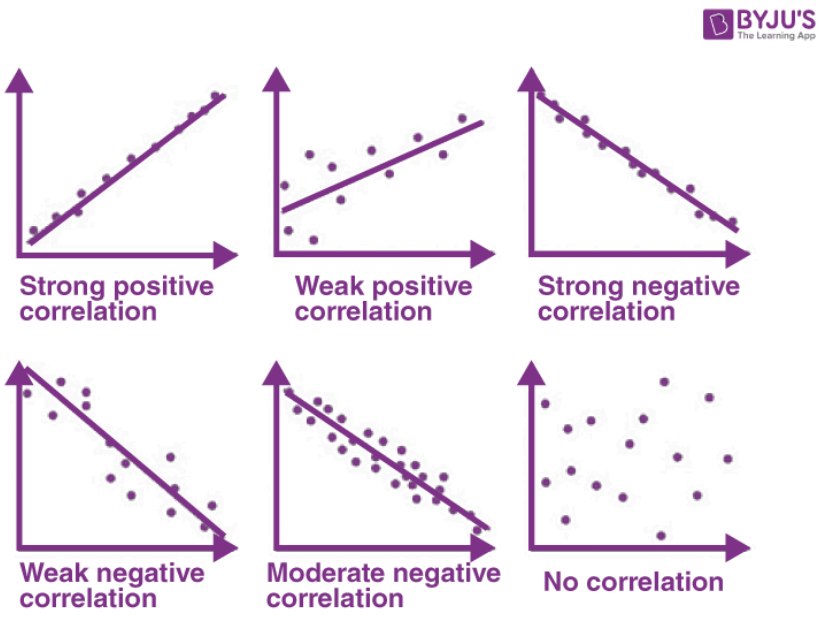
It's variance, it's so bad.

---

10. Correlation

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

It gives both strength + direction = awesome.



Correlation is often measured using a statistical tool called `correlation coefficient`. which ranges from `-1` to `+1` .

`-1` indicates the negative correlation

`+1` indicates the positive correlation

`0` indicates the zero correlation or no correlation

Formula:

$$Correlation = \frac{cov(X,Y)}{\sigma x * \sigma y}$$

sigma x → standard deviation of x

sigma y → standard deviation of y

correlation and causation

The phrase `correlation does not imply causation` . It means the correlation between two variables does not necessarily imply that one variable is the reason for another variable's behaviours.

Correlation

Correlation means **two variables move together** — either in the same or opposite direction.

It shows **a statistical relationship**, but **not necessarily a cause-effect link**.

Causation

Causation means **one variable directly influences or causes changes in another**.

⚠️ The Phrase:

> "Correlation does not imply causation."

means:

Just because two variables are correlated (i.e., they occur together), it **doesn't mean one causes the other**.

There could be:

- A **third hidden variable** (confounder)
- A **coincidence**
- Or **reverse causation**

Example:

Observation:

Children who sleep with the lights on are more likely to develop nearsightedness.

Explanation:

It's not the **night light** causing myopia —

> Parents who are nearsighted are more likely to keep a night light on **and** pass down the genetic tendency for myopia.
>
> So the **real cause** is genetics, not the light.

Thus while correlation can provide valuable insights into how different variables are related, they cannot be used to establish causality. To establish causality often required additional evidences such as experiments, or well designed observational studies.

Relation ≠ cause

Inferential Statistics

Inferential Statistics is a branch of statistics that focuses on making predictions, estimations or generalizations about a larger population based on a sample of data taken from the population. It involves the use of probability theory to make inferences and draw conclusions about the characteristics of a population by analysing a smaller subset or sample.

The key idea behind inferential statistics is that it is often impractical or impossible to collect data from every member of a population, so instead we use a representative sample to make inferences about the entire group. Inferential statistical technique include hypothesis testing, confidence intervals, and regression analysis, among other.

These methods help researchers answer questions like,

1. Is there a significant difference between two groups?
2. Can we predict the outcome of a variable based on the values of other variable ?
3. What is the relationship between two or more variables?

1. Central Limit Theorem

Sampling Distribution

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

Let's say you have **salary data for the entire population of India** — millions of data points. Since analysing every single record is impractical, you decide to take **samples**.

- Suppose your **sample size** is 50.

  You randomly select 50 salary values from the population:

  $X_1 = \{x_{11}, x_{12}, \ldots, x_{1,50}\}$
- You then repeat this process **100 times**, each time drawing a new random sample of 50 salaries:

  $X2, X3, \ldots, X100$

Now, you have **100 different samples**:

$X1, X2, \ldots, X100$

Together, these samples form what we call the **sampling distribution** — a collection of samples drawn from the same population.

If you calculate the **sample mean** for each of these samples —

$\bar{X}1, \bar{X}2, \ldots, \bar{X}100$

and plot all these mean values, the resulting curve is known as the **sampling distribution of the sample mean**.

If you're calculating for variance then this gonna be sampling distribution of the sample variance.

Why Sampling distribution is Important?

Sampling distribution is important in statistics and machine learning 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.

What "variability of a sample statistic" means?

When we take a **sample** from a population, the statistic we calculate (like mean, median, or variance) will almost never be *exactly* the same each time — even if we use the same sample size and sampling method.

This **variation** in the values of a sample statistic (like the sample mean) **across different samples** is what we call the **variability of a sample statistic**.

Intuition of CLT

The central Limit Theorem states that the distribution of the sample means of a large number of independent and identically distributed random variables will approach a normal distribution, regardless of the underlying distribution of the variables. e

Setup — population and notation

- Let the **population** salary (a single random draw) be the random variable with meanμ and variance σ.

$$X \text{ has } E[X] = \mu,\ \mathrm{Var}(X) = \sigma^2$$

- We will draw n=100 salaries for each sample (this is your `sample_size`).
- We will repeat the sampling process T=1000 times (this is your `n_trails`, i.e. T trials).

Step 1 — Draw the sample 1000 times and calculate mean for each sample

$$X_1 = \bar{x_1}, ...X_{1000} = \bar{x}_{1000}$$

Step 3 - Plot all sample mean in graph

- If you plot all the sample mean in graph, you will get a gaussian distribution.
- This holds **regardless of the shape** of the original salary distribution (it could be skewed, heavy-tailed, etc.), provided the observations are i.i.d. and n is large enough.

Conditions for CLT

1. The sample size is large enough, typically greater than or equal to 30.
2. The sample is drawn from a finite population or an infinite population with a finite variant.
3. The random variable in the sample are independent and identically distributed.

The **mean** and **standard deviation** of the formed gaussian distribution will be **exactly same** as **population**.

The CLT is important in statistics and machine learning because it allows us to make probabilistic inferences about a population based on a sample of data. For example, we can use the CLT to construct confidence intervals, perform hypothesis testing, and make predictions about the population mean based on the sample data. The CLT also provides a theoretical justification for many commonly used statistical techniques, such as t-tests, ANOVA, and linear regression.

Example Case Study "What is the average income of Indians"

Step-by-step process:

1. Collect multiple random samples of salaries from a representative group of Indians. Each sample should be large enough (usually, n > 30) to ensure the CLT holds. Make sure the samples are representative and unbiased to avoid skewed results.
2. Calculate the sample mean (average salary) and sample standard deviation for each sample.
3. Calculate the average of the sample means. This value will be your best estimate of the population mean (average salary of all Indians).
4. Calculate the standard error of the sample means, which is the standard deviation of the sample means divided by the square root of the number of samples.
5. Calculate the confidence interval around the average of the sample means to get a range within which the true population mean likely falls. For a 95% confidence interval:

lower_limit = average_sample_means - 1.96 * standard_error
upper_limit = average_sample_means + 1.96 * standard_error

`1.96` is two standard deviation for standard normal distribution.

6. Report the estimated average salary and the confidence interval.

---

2. Confidence Intervals

Some key terms before moving into the CI.

Population

A population is the entire group or set of individuals, objects, or events that a researcher wants to study or draw conclusion about. It can be people, animals, plants, or even inanimate objects, depending on the context of the study. The population usually represents the complete set of possible data points or observations.

Sample

A sample is a subset of population that is selected for study. It is a smaller group that is intended to be representative of the large population. Researches collect data from the sample and use it to make inferences about population as a whole. Since it often impractical or impossible to collect data from every member of a population, samples are used as an efficient and cost-effective way to gather information.

> *A sample should be **representative** and it should be **random**.*

Parameter

A Parameter is a numerical value that describes a characteristic of a population. Parameters are usually denoted using Greek letters, such as μ for the population mean or σ for population standard deviation. since it's often difficult or impossible to obtain data from an entire population, parameters are usually unknown and must estimated based on available sample data .

Statistic or Estimate

A statistic is a numerical value that describes a characteristic of a sample, which is a subset of population. By using statistics calculated from a representative sample, researchers can make inferences about the unknown respective parameter of the population. Common statistics include the sample mean $\bar{x}$, s for sample standard deviation.

> If you infer mean, median, ... from population it's called **parameter**. Similarly, if you infer mean, median, ... from sample it's called **statistic** or **estimate**.

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 mean, std. Point estimates are often used in statistics when we want to make inferences about a population based on a sample.

Population Parameter (unknown)	Sample Statistic (Point Estimate)	Example
Population Mean (μ)	Sample Mean $(\bar{X})$	You take the **average salary** of 100 employees to estimate the **average salary of all employees** in India.
Population Proportion (p)	Sample Proportion ($\hat{p}$)	You survey 200 people; 60 say they prefer "Product A". So $(\hat{p} = \frac{60}{200} = 0.3)$. That's your point estimate for the **population proportion** who prefer Product A.
Population Variance (σ^2)	Sample Variance (s^2)	You calculate variance of test scores for a sample of 50 students to estimate the variance for all students.
Population Standard Deviation (σ)	Sample Standard Deviation (s)	You compute the **sample standard deviation** of monthly spending from 100 customers to estimate the population's spending variability.
Population Correlation (ρ)	Sample Correlation (r)	You compute correlation between **age** and **income** for a subset of 500 people to estimate correlation in the full population.

The problem with **point estimates** is that they are not always reliable. Since they are calculated from a **sample**, there's no guarantee that this single value accurately represents the **entire population**. To address this uncertainty, researchers recommend providing an **interval estimate** instead of a single value. This range, known as a **confidence interval**, gives a span of values within which the **true population parameter** is likely to lie.

Intuition of 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 a estimate obtained from a sample of data. Confidence interval is created for **parameters**; **not for statistics**. Statistics help us get the **confidence interval** for a **parameter**.

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

$$\text{Confidence Interval} = \text{Population Parameter} \pm \text{Margin of Error}$$

There are multiple ways to calculate

1. Z procedure (*you have Population std*)
2. t procedure (*you don't have population std*)

Z Procedure (Sigma Known)

It’s used when population standard deviation is available however to use this procedure you have to satisfy **three assumptions**

1. **Random sampling;**
2. **Known population standard deviation;** The population standard deviation must be known or accurately estimated. In practice, the population standard deviation is often unknown, and the sample standard deviation is used as an estimate. However if the sample size is large enough, the sample standard deviation can provide a reasonably accurate approximation.
3. **Normal distribution or large sample size;** The Z-procedure assumes that the underlying population is normally distributed. However, If the population distribution is not normal, the central limit theorem can be applied when the sample size is large n≥ 30.

Interpreting Confidence Intervals:

A confidence interval is a range of values within which a population parameter, such as the population mean, is estimated to lie with a certain level of confidence. The confidence interval provides an indication of the precision and uncertainty associated with the estimate. To interpret the confidence interval values, consider the following points:

1. **Confidence level:** The confidence level (commonly set at 90%, 95%, or 99%) represents the probability that the confidence interval will contain the true population parameter if the sampling and estimation process were repeated multiple times. For example, a 95% confidence interval means that if you were to draw 100 different samples from the population and calculate the confidence interval for each, approximately 95 of those intervals would contain the true population parameter.
2. **Interval range:** The width of the confidence interval gives an indication of the precision of the estimate. A narrower confidence interval suggests a more precise estimate of the population parameter, while a wider interval indicates greater uncertainty. The width of the interval depends on the sample size, variability in the data, and the desired level of confidence.
3. **Interpretation:** To interpret the confidence interval values, you can say that you are "X% confident that the true population parameter lies within the range (lower limit, upper limit)." Keep in mind that this statement is about the interval, not the specific point estimate, and it refers to the confidence level you chose when constructing the interval.

Margin of Error

The **margin of error** represents the **amount of uncertainty** or **possible error** in an estimate calculated from a sample.

It tells you **how far** your sample statistic (like sample mean $\bar{X}$) is likely to be from the **true population parameter** (like population mean μ).

$$CI = \text{Point Estimate} \pm \text{Margin of Error}$$

$$\text{Margin of Error} = Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$CI = \bar{X} \pm Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$\bar{X}$ = sample mean

$Z_{\alpha/2}$= critical value from the standard normal distribution (for example, 1.96 for 95% confidence)

σ = population standard deviation

n = sample size

CI depends on `critical value`, `sample size`, `population standard deviation`

Calculating critical value

$Z_{\alpha/2} = Z_{(1-\frac{1-\text{CL}}{2})}$

Calculating Z for 95% confidence

$$\alpha = 1 - 0.95 = 0.05$$
$$Z_{\alpha/2} = Z_{(1-0.025)} = Z_{0.975} \approx 1.96$$

Calculating Z for 99% confidence

$$\alpha = 1 - 0.99 = 0.01$$
$$Z_{\alpha/2} = Z_{(1-0.005)} = Z_{0.995} \approx 2.576$$

T Procedure (Sigma UnKnown)

Assumptions

1. **Random sampling:** The data must be collected using a random sampling method to ensure that the sample is representative of the population. This helps to minimize biases and ensures that the results can be generalized to the entire population.
2. **Sample standard deviation:** The population standard deviation (σ) is unknown, and the sample standard deviation (s) is used as an estimate. The t-distribution is specifically designed to account for the additional uncertainty introduced by using the sample standard deviation instead of the population standard deviation.
3. **Approximately normal distribution:** The t-procedure assumes that the underlying population is approximately normally distributed, or the sample size is large enough for the Central Limit Theorem to apply. If the population distribution is heavily skewed or has extreme outliers, the t-procedure may not be accurate, and non-parametric methods should be considered.

4. **Independent observations**: The observations in the sample should be independent of each other. In other words, the value of one observation should not influence the value of another observation. This is particularly important when working with time series data or data with inherent dependencies.

Here we don't have population standard deviation, so the next possible value we can take is estimate standard deviation (sample standard deviation). Normally when you're calculating CI, you convert your normal distribution to standard normal distribution for the symmetrical purpose, the formula is $Z = \frac{\bar{x}}{\frac{\sigma}{\sqrt{n}}}$, here we don't have the σ, so we have to sample s standard deviation. So, $Z = \frac{\bar{x}}{\frac{s}{\sqrt{n}}}$.

The another problem is whenever you're converting from normal distribution to standard normal distribution using new formula $Z = \frac{\bar{x}}{\frac{s}{\sqrt{n}}}$, it does not form the normal distribution, it looks like normal distribution but it's not. The new distribution which is formed called `student T distribution`, it's a theoretical distribution named by the after some researcher named as student. Student T distribution is not exists in nature like normal distribution, it's created so we call it as theoretical distribution.

It's a theoretical distribution, which looks similar to normal distribution but it's not. In Student T distribution we have one parameter called `degree of fredom`, basically it's `n-1` . If the sample is 50, the `df = 49` . It is **theoretical**, not naturally occurring like the normal distribution.

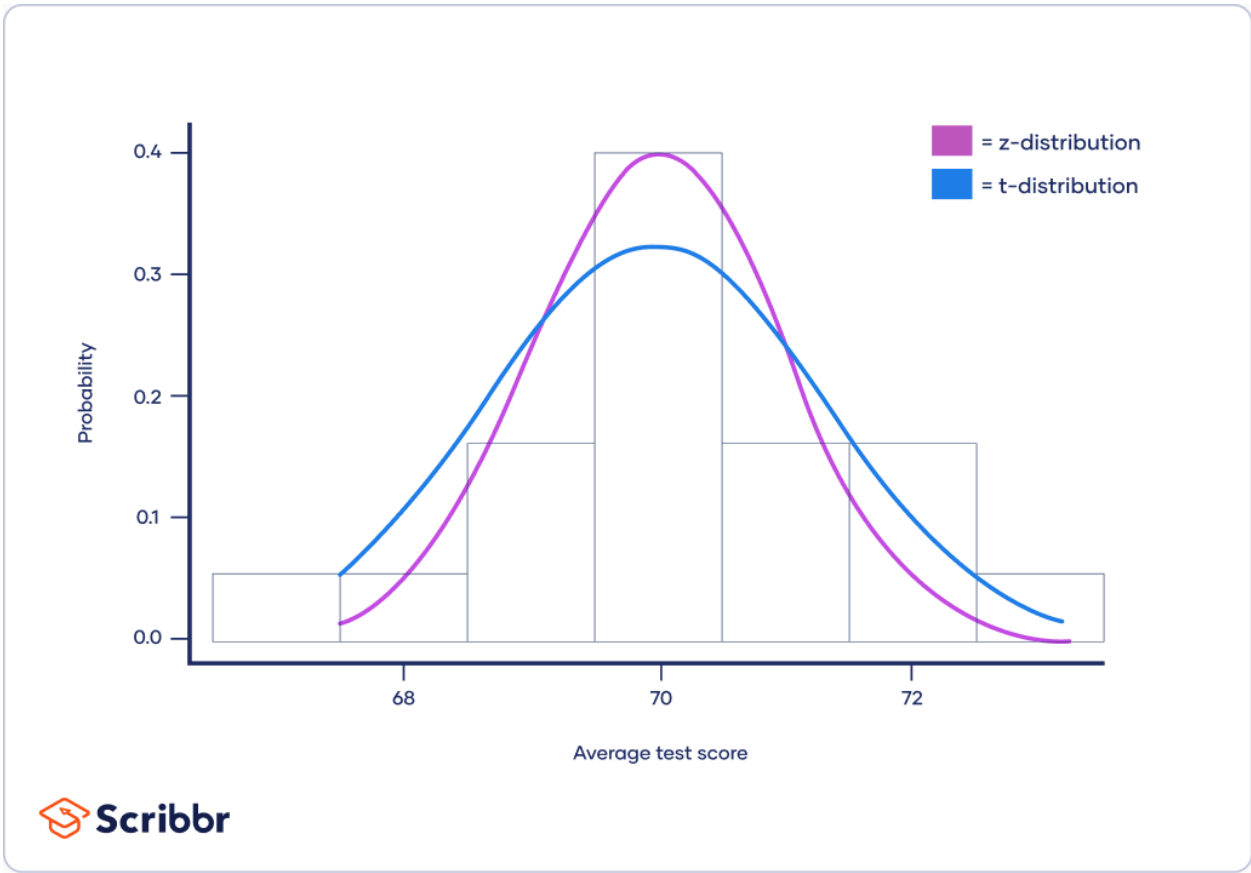
$$df = n - 1$$

As n increases, the **T-distribution approaches** the **normal distribution. Degrees of freedom** adjust the shape of the T-distribution.

So the final formula looks like

$$CI = \bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

The tail in the student t distribution is fatter, means near to outlier we will have a lot of values.



If the $df = \infty$ near to infinity, you will the both distribution match with each other. Play here: t distribution tool

For lower sample size and lower degree of freedom, the t statistics will be greater than z statistics because you're uncertain about this sample standard deviation, so it will increase the confidence interval range more.

Normally the confidence interval for `t procedure` is high because you're unsure about the `t distribution` . So you have a buffer range.

3. 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.

Hypothesis testing is a way to **use data** to check whether an **idea** or **claim** about a **group** of things is likely to be **true**. It helps us make informed decisions based on evidence from a sample instead of looking at the entire group.

Why it's used:

- To see if a new idea, product, or method works better than the old one.
- To check claims or statements about a population.
- To make decisions based on limited data rather than guessing.

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 to **either reject** or **fail** to **reject** the **null hypothesis** in favour of the alternative hypothesis, which claims there is significant effect or relationship.

This is called **Status Quo**

Alternate Hypothesis (H1 or Ha)

The null hypothesis is a **statement** that assumes there is **no 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.

This is called **research hypothesis**

Important points

- How to decide what will be Null hypothesis and what will be Alternate Hypothesis (Typically the Null hypothesis says nothing new is happening)
- We try to **gather evidence** to **reject null hypothesis**
- It's important to note that failing to reject the null hypothesis doesn't necessarily mean that the null hypothesis is true; it just means that there isn't enough evidence to support the alternative hypothesis.

Hypothesis tests are similar to jury trials, in a sense. In a jury trial, $H0$ is similar to the not-guilty verdict, and Ha is the guilty verdict. You assume in a jury trial that the defendant isn't guilty unless the prosecution can show beyond a reasonable doubt that he or she is guilty. If the jury says the evidence is beyond a reasonable doubt, they reject H0, not guilty, in favour of Ha , guilty.

There are two ways you can do Hypothesis testing

1. Rejection Region Approach (basic level testing)
2. P value approach

Steps involved in **Rejection Region Approach**

1. Formulate a Null and Alternate Hypothesis
2. Select a significance level (this is the probability of rejecting the null hypothesis when it's actually true, usually set at 0.05 (5%) or 0.01(1%))
3. Check assumptions (example distribution)
4. Decide which test is appropriate (Z-test, T-test, chi-square test, ANOVA)
5. State the relevant test statistic
6. Conduct the test
7. Reject or not reject the null hypothesis
8. Interpret the result

Significance level

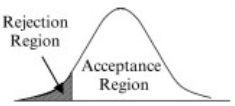
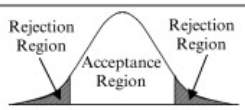
It's denoted by α ,is a predetermined threshold used in hypothesis testing to determine whether the null hypothesis should be rejected or not. It represents the probability of rejecting null hypothesis when it is actually true, also known as **Type 1 error.**

Observations: if you increase the significance level, the rejection area will increase, if you decrease the significance level the rejection area will decrease.

α (alpha) is the probability of making a mistake by rejecting H 0 when the null hypothesis really is true in the population. **Example:** "If the null hypothesis is in fact correct (true in reality), I am willing to accept a 5% chance of incorrectly rejecting it based on my sample data."

Rejection Region

The rejection region is the region of values that corresponds to the rejection of the null hypothesis at some chosen probability level.

One-Tailed Test (Left Tail)	Two-Tailed Test	One-Tailed Test (Right Tail)
$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X < \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X \neq \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X > \mu_0$
		

Problems with Rejection Region Approach

- The rejection region approach only tells you **"reject"** or **"don't reject"**. It doesn't communicate *how strongly* your data supports or contradicts H0. Let's say you're critical region value is 1.96 and your z value is 15, definitely you will reject but here; you have very strong evidence, let's say your z value is 2, here also you will reject but the evidence is not strong like before. You **can't capture** the **strong** or **effectiveness** of **your evidence** in rejection region approach.

- We still reject the HO if your values are very near to each other. Your critical region is 1.96 and your Z value is 1.95 still you reject the H0, that's another problem.

In hypothesis testing there are two types of errors that can occur when making a decision about the null hypothesis.

1. Type 1 Error (False Positive)
2. Type 2 Error (False Negative)

Type 1 Error

Error occurs when the sample results, lead to the rejection of the null hypothesis when it is in fact true. In other words, it's the mistake of finding a significant effect or relationship when there is none.

The probability of committing a Type I error is denoted by α (alpha), which is also known as the significance level. By choosing a significance level, researchers can control the risk of making a Type I error.

$$Type\ I\ error = Reject\ H0\ when\ H0\ is\ true.$$

Type 2 Error

Error occurs when based on the sample results, the null hypothesis is not rejected when it is in fact false. This means that the researcher fails to detect a significant effect or relationship when one actually exists.

The probability of committing a Type II error is denoted by β (beta). Beta is the probability of committing type 2 error.

$$Type\ II\ error = Fail\ to\ reject\ H0\ when\ H0\ is\ false.$$

Trade off between Type 1 and Type 2

- To **reduce the risk of Type I error**, you **decrease the significance level α**.
- To **accept more risk of Type I error**, you **increase**
- Reducing α makes it harder to reject H0.
- This **increases the probability of Type II error (β)**.
- There's always a trade-off: lowering Type I error often increases Type II error, unless you increase sample size.

One sided test

A one sided test (aka one tailed test) is used when the researcher is interested in testing the effect in a specific direction (either greater than or less than the values specified in the null hypothesis). the alternative hypothesis in a one sided test contains an inequality (either "<" or ">")

Example: A Researcher wants to test whether a new medication increases the average recovery rate compared to the existing medication.

It has two types of one sided test

1. Right tail test $H_a : \mu > value$
2. Left tail test $H_a : \mu < value$

In this context, μ (mu) represents the population mean.

In a right tail test, the alternative hypothesis H_a: μ > value means you're testing whether the population mean is greater than a specific value stated in the null hypothesis .

For example, if you're testing whether a new medication increases the average recovery rate compared to an existing one, μ would represent the true average recovery rate in the population, and you'd be testing if it's significantly higher than the current value.

Two sided test

A two-sided test is used when the researcher is interested in testing the effect in both directions (i.e., whether the value specified in the null hypothesis is different, either greater or lesser). The alternative hypothesis in a two-sided test contains a "not equal to" sign (≠).

Example: A researcher wants to test whether a new medication has a different average recovery rate compared to the existing medication.

The main difference between them lies in the directionality of the alternative hypothesis and how the significance level is distributed in the critical regions.

Two-tailed test (two-sided):

Advantages:

1. **Detects effects in both directions:** Two-tailed tests can detect effects in both directions, which makes them suitable for situations where the direction of the effect is uncertain or when researchers want to test for any difference between the groups or variables.
2. **More conservative:** Two-tailed tests are more conservative because the significance level (α) is split between both tails of the distribution. This reduces the risk of Type I errors in cases where the direction of the effect is uncertain.

Disadvantages:

1. Less powerful: Two-tailed tests are generally less powerful than one-tailed tests because the significance level (α) is divided between both tails of the distribution. This means the test requires a larger effect size to reject the null hypothesis, which could lead to a higher risk of Type II errors (failing to reject the null hypothesis when it is false).
2. **Not appropriate for directional hypotheses:** Two-tailed tests are not ideal for cases where the research question or hypothesis is directional, as they test for differences in both directions, which may not be of interest or relevance.

One-tailed test (one-sided):

Advantages:

1. **More powerful:** One-tailed tests are generally more powerful than two-tailed tests, as the entire significance level (α) is allocated to one tail of the distribution. This means that the test is more likely to detect an effect in the specified direction, assuming the effect exists.

2. **Directional hypothesis:** One-tailed tests are appropriate when there is a strong theoretical or practical reason to test for an effect in a specific direction.

Disadvantages:

1. **Missed effects:** One-tailed tests can miss effects in the opposite direction of the specified alternative hypothesis. If an effect exists in the opposite direction, the test will not be able to detect it, which could lead to incorrect conclusions.

2. **Increased risk of Type I error:** One-tailed tests can be more prone to Type I errors if the effect is actually in the opposite direction than the one specified in the alternative hypothesis.

Where hypothesis testing is used?

1. Testing the effectiveness of interventions or treatments: Hypothesis testing can be used to determine whether a new drug, therapy, or educational intervention has a significant effect compared to a control group or an existing treatment.
2. Comparing means or proportions: Hypothesis testing can be used to compare means or proportions between two or more groups to determine if there's a significant difference. This can be applied to compare average customer satisfaction scores, conversion rates, or employee performance across different groups.
3. Analysing relationships between variables: Hypothesis testing can be used to evaluate the association between variables, such as the correlation between age and income or the relationship between advertising spend and sales.
4. Evaluating the goodness of fit: Hypothesis testing can help assess if a particular theoretical distribution (e.g., normal, binomial, or Poisson) is a good fit for the observed data.
5. Testing the independence of categorical variables: Hypothesis testing can be used to determine if two categorical variables are independent or if there's a significant association between them. For example, it can be used to test if there's a relationship between the type of product and the likelihood of it being returned by a customer.
6. A/B testing: In marketing, product development, and website design, hypothesis testing is often used to compare the performance of two different versions (A and B) to determine which one is more effective in terms of conversion rates, user engagement, or other metrics.

Hypothesis Testing for ML

1. Model comparison: Hypothesis testing can be used to compare the performance of different machine learning models or algorithms on a given dataset. For example, you can use a paired t-test to compare the accuracy or error rate of two models on multiple cross-validation folds to determine if one model performs significantly better than the other.

2. Feature selection: Hypothesis testing can help identify which features are significantly related to the target variable or contribute meaningfully to the model's performance. For example, you can use a t-test, chi-square test, or ANOVA to test the relationship between

individual features and the target variable. Features with significant relationships can be selected for building the model, while non-significant features may be excluded.

3. Hyperparameter tuning: Hypothesis testing can be used to evaluate the performance of a model trained with different hyperparameter settings. By comparing the performance of models with different hyperparameters, you can determine if one set of hyperparameters leads to significantly better performance.
4. Assessing model assumptions: In some cases, machine learning models rely on certain statistical assumptions, such as linearity or normality of residuals in linear regression. Hypothesis testing can help assess whether these assumptions

---

P Values

P - value is the **probability** of getting a **sample as** or **more extreme** (having more evidence against H0) than our **own sample** given the Null Hypothesis (H0) is true.

In simple words p-value is a measure of strength of the evidence against the Null Hypothesis that is provided by our sample data.

The p-value tells us **how strong the sample evidence is against the null hypothesis.**

If this evidence is strong enough (i.e., p-value < α), we reject H_0.

If it's weak, we don't reject H_0.

Feature	Significance Level (α)	p-value
Definition	The predetermined probability of rejecting the null hypothesis when it is actually true (a Type I error).	The probability of observing a result as extreme or more extreme than your data, assuming the null hypothesis is true.

You toss a coin 100 times and observe 53 heads. You compute a p-value of 0.3086.

If the coin is truly fair, there is a **30.86% chance** of observing a result **as extreme as or more extreme than 53 heads in 100 flips** *purely due to random chance*.

- You flipped a coin 100 times and got 53 heads (instead of the expected 50 if the coin were fair)
- You calculated a p-value of 0.3086 (or 30.86%)

What the p-value means:

The p-value answers this question: "If the coin is actually fair, what's the probability of getting a result at least as extreme as 53 heads?"

In this case, 0.3086 means there's about a 31% chance of observing 53 or more heads (or 47 or fewer heads) purely by random chance when flipping a fair coin 100 times.

The conclusion:

Since 31% is a fairly high probability, this result isn't unusual enough to conclude the coin is biased. You'd typically reject the null hypothesis (that the coin is fair) only if the p-value is very small—usually less than 0.05 (5%).

Think of it this way: if something happens 31% of the time by random chance, it's not particularly surprising or strong evidence of unfairness.

Some Rule of thumb If Significance value is available

If $P - value \leq \alpha$ reject Ho

P value gives the magnitude of strength for rejecting null

Some Rule of thumb If Significance value is not available

1. Very small p-values (e.g., p < 0.01) indicate **strong evidence against** the null hypothesis, suggesting that the observed effect or difference is unlikely to have occurred by chance alone.
2. Small p-values (e.g., 0.01 ≤ p < 0.05) **indicate moderate evidence** against the null hypothesis, suggesting that the observed effect or difference is less likely to have occurred by chance alone.
3. Large p-values (e.g, 0.05 ≤ p < 0.1) indicate **weak evidence** against the null hypothesis, suggesting that the observed effect or different might have occurred by chance alone, but there is still some level of uncertainty.
4. Very large p-values (e.g. p≥ 0.1) indicate **weak or no evidence** agains the null hypothesis, suggesting that the observed effect or difference is likely to have occurred by chance alone.

Z Test

If you know population standard deviation, use this. Most of the times, you won't be having population standard deviation, so you mostly use Z test.

Z-Value (Z-Score)

A **Z-value** (or **Z-score**) is a statistical measure that describes how many standard deviations a data point is from the mean of a distribution. It is used to standardize values from different normal distributions, allowing for comparison across different scales.

Formula:

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

Where:

- X is the individual data point
- μ is the population mean
- σ is the population standard deviation

Interpretation:

- A Z-score of **0** means the data point is exactly at the mean
- A **positive Z-score** indicates the data point is above the mean
- A **negative Z-score** indicates the data point is below the mean

- A Z-score of **±1** means the value is one standard deviation away from the mean
- A Z-score of **±2** means the value is two standard deviations away from the mean

Uses:

1. **Standardization:** Converting different distributions to a standard normal distribution (mean = 0, standard deviation = 1)
2. **Identifying outliers:** Values with Z-scores beyond ±3 are typically considered outliers
3. **Probability calculation:** Using Z-tables to find probabilities and percentiles
4. **Hypothesis testing:** In Z-tests to determine if a sample mean is significantly different from a population mean

Example:

If a test score distribution has a mean (μ) of 70 and standard deviation (σ) of 10, and a student scores 85:

$Z = \frac{85-70}{10} = \frac{15}{10} = 1.5$

This means the student scored 1.5 standard deviations above the mean.

Connection to Normal Distribution:

In a standard normal distribution:

- Approximately **68%** of data falls within Z = ±1
- Approximately **95%** of data falls within Z = ±2
- Approximately **99.7%** of data falls within Z = ±3

T Test

If you don't know population standard deviation, use this and also If you have **smaller sample size** it works so good.

T test is a statistical test used in hypothesis testing to compare the **means** of **two sample** or to **compare** a **sample mean** to a **known population mean**. The t-test is based on the t-distribution, which is used when the population standard deviation is unknown and the sample size is small.

There are three types

1. **One sample t-test:** The one sample t-test is used to compare the mean of a single sample to a known population mean. The null hypothesis states that there is no significant difference between the sample mean and the population mean, while the alternative hypothesis states that there is a significant difference.
2. **Independent two sample t-test:** The independent two sample t-test is used to compare the means of two independent groups. The null hypothesis states that there is no significant difference between the means of the two samples, while the alternative hypothesis states that there is a significant difference.
3. **Paired t-test (dependent two-sample t-test):** The paired t-test is used to compare the means of two samples that are dependent or paired, such as pre-test and post-test scores for the same group of subjects or measurements taken on the same subjects under two different conditions. The null hypothesis states that there is no significant difference between the means of the paired differences, while the alternative hypothesis states that there is a significant difference.

One sample t-test:

A one sample t-test checks whether the sample mean differs from population mean

Assumptions

1. **Normality;** Population from which the sample is drawn is normally distributed.
2. **Independence;** The observations in the sample must be independent, which means the value of one observation should not influence the value of another observation.
3. **Random sampling;** The sample must be random and representative subset of the population.
4. **Unknown population std**; The population std is not known

Example problem

> Suppose a manufacturer claims that the average weight of their new chocolate bars is 50 grams, we highly doubt that and want to check this so we drew out a sample of 25 chocolate bars and measured their weight, the sample mean came out to be 49.7 grams and the sample std deviation was 1.2 grams. Consider the significance level to be 0.05

Variable Definitions:

$$\bar{x} = 49.7$$
$$mu = 50$$
$$s = 1.2$$
$$n = 25$$
$$alpha = 0.05$$

Hypotheses:

$$H_0 : \mu = 50 \text{ The mean weight is 50 grams}$$
$$H_1 : \mu \neq 50 \text{ The mean weight is not 50 grams}$$

T-statistic Formula for one sample t-test:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

T-statistic Calculation:

$$t = \frac{49.7 - 50}{1.2/\sqrt{25}} = \frac{-0.3}{0.24} \approx -1.25$$

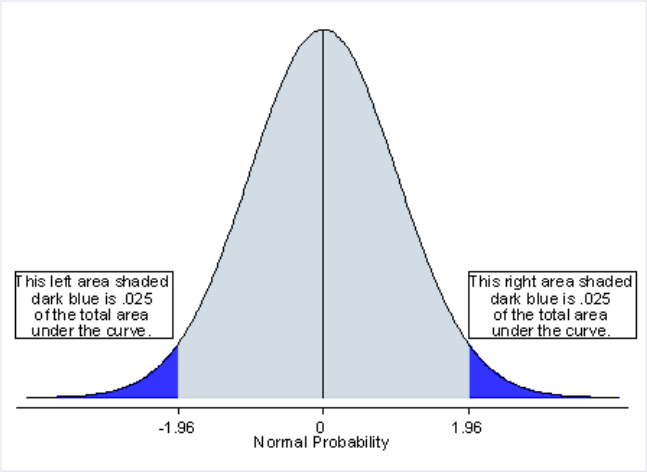
Degrees of Freedom:

$$df = n - 1 = 24$$

CDF at -1.25: $(Z \leq -1.25) \approx 0.106$ (you can calculate using cdf function)

Two-tailed p-value: $0.2112 \times 0.106 \approx 0.211$ (because we want to infer in both the side)

So, $p - value \approx 0.211$, which is greater than α = 0.05 confirming that we **fail** to **reject** the **null hypothesis**.




```
# python code to calculate the area (cdf) for t statitic for two sampled test
from scipy.stats import t

t_value = -1.25
df = 24

# Calculate the CDF value
cdf_value = t.cdf(t_value, df)
print(cdf_value*2) # *2 because we are calculating for both the sides

# ------
# one sample t test using stats module
import scipy.stats as stats

t_statistic, p_value = stats.ttest_1samp(sample_age, pop_mean)
print("t-statistic:", t_statistic)
print("p-value:", p_value)
```

Independent two Sample t-test

An independent two-sample t-test also known as unpaired t-test, is a statistical method used to **compare** the means of **two independent groups** to determine if there is a significant difference between them.

Assumptions

1. **Independence of observations;** The two samples must be independent, meaning there is no relationship between the observations in one group and the observations in other group. the subjects in the two groups should be selected randomly and independently.
2. **Normality;** The data in each of the two groups should be approximately normally distributed. The t-test is considered robust to mild violations of normality, especially when the sample sizes are large(typically $n \geq 30$), and the sample sizes of the two groups are similar. If the data is highly skewed or has substantial outliers, consider using a non parametric test, such as the Mann-Whitney U test.
3. **Equal Variances (Homoscedasticity):** The variances of the two populations should be approximately equal. This assumptions can be checked using F-test for equality of variances. If the assumptions is not met, you can use Welch's t-test, which does not require equal variances.
4. **Random Sampling;** The data should be collected using a random sampling method from the respective populations, This ensures the sample is representative of the populations and reduces the risk of selection bias.

If your sample size is less than 30, please use some testing method to check whether the sample is normally distributed or not. (tests like shapiro wilk)

Example Problem

Suppose a website owner claims that there is no difference in the average time spent on their website between desktop and mobile users. To test this claim, we collect data from 30 desktop users and 30 mobile users regarding the time spent on the website in minutes. Consider the significance level to be 0.05. The sample statistics are as follows:

desktop users =
$[12, 15, 18, 16, 20, 17, 14, 22, 19, 21, 23, 18, 25, 17, 16, 24, 20, 19, 22, 18, 15, 14, 23, 16, 12, 21, 19, 17, 20, 14]$
mobile_users =
$[10, 12, 14, 13, 16, 15, 11, 17, 14, 16, 18, 14, 20, 15, 14, 19, 16, 15, 17, 14, 12, 11, 18, 15, 10, 16, 15, 13, 16, 11]$

Desktop users:

- Sample size (n1): 30
- Sample mean (mean1): 18.5 minutes
- Sample standard deviation (std_dev1): 3.5 minutes

Mobile users:

- Sample size (n2): 30
- Sample mean (mean2): 14.3 minutes
- Sample standard deviation (std_dev2): 2.7 minutes

Variables definition

$$
\begin{gathered}
n_1 = 30 \\
n_2 = 30 \\
\bar{x_1} = 18.5 \\
\bar{x_2} = 14.3 \\
s_1 = 3.5 \\
s_2 = 2.7 \\
\alpha = 0.05
\end{gathered}
$$

State Hypothesis

$$
\begin{gathered}
H_0 = (\mu_1 == \mu_2) \\
H_1 = (\mu_1 \neq \mu_2)
\end{gathered}
$$

T formula for independent two sample t test

$$
t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}
$$

T calculation

$$
t = \frac{18.5 - 14.3}{\sqrt{\frac{3.5^2}{30} + \frac{2.7^2}{30}}} \approx 5.204
$$

Degree of freedom

$$
\begin{gathered}
df = n_1 + n_2 - 2 \\
df = 58
\end{gathered}
$$

CDF at 5.204 $p - value = 6.256e^{-04}$ It's close to zero

The p-value is lesser than alpha. so we can **reject** the **null hypothesis**.

```
# python code to calculate the area (cdf) for t statitic for two sampled test
from scipy.stats import t

t_value = -5.204  # because cdf gives left to right so I am calculating for the firs
t
# portion then I can multiply the output with 2 because it's symmetrical in nature.
df = 58

# Calculate the CDF value
cdf_value = t.cdf(t_value, df)
print(cdf_value*2) # *2 because we are calculating for both the sides
```

```
# The entire ttest_independent using stats module
import scipy.stats as stats
```

```
t_statistic, p_value = stats.ttest_ind(sample_male, sample_female)

print("t-statistic:", t_statistic)
print("p-value:", p_value) # divide by two if you're doing for single sample t test
```

Paired t-test (dependent two-sample t-test)

A Paired two sample t-test, also known as a dependent or paired samples t-test, is a statistical test used to compare the means of two related or dependent groups.

Assumptions

1. **Paired observations**: The two sets of observations must be related or paired in some way, such as before-and-after measurements on the same subjects or observations from matched or correlated groups.
2. **Normality**: The differences between the paired observations should be approximately normally distributed. This assumption can be checked using graphical methods (e.g., histograms, Q-Q plots) or statistical tests for normality (e.g., Shapiro-Wilk test). Note that the t-test is generally robust to moderate violations of this assumption when the sample size is large.
3. **Independence of pairs**: Each pair of observations should be independent of other pairs. In other words, the outcome of one pair should not affect the outcome of another pair. This assumption is generally satisfied by appropriate study design and random sampling.

Common scenarios where a paired two-sample t-test is used include:

1. Before-and-after studies: Comparing the performance of a group before and after an intervention or treatment.
2. Matched or correlated groups: Comparing the performance of two groups that are matched or correlated in some way, such as siblings or pairs of individuals with similar characteristics.

Example Problem

> Let's assume that a fitness center is evaluating the effectiveness of a new 8-week weight loss program. They enroll 15 participants in the program and measure their weights before and after the program. The goal is to test whether the new weight loss program leads to a significant reduction in the participants' weight.
>
> Before the program:
> $[80, 92, 75, 68, 85, 78, 73, 90, 70, 88, 76, 84, 82, 77, 91]$
>
> After the program:
> $[78, 93, 81, 67, 88, 76, 74, 91, 69, 88, 77, 81, 80, 79, 88]$
>
> Significance level (α) = 0.05

Variable Definitions

$$n = 15$$
$$\alpha = 0.05$$

Hypothesis

$$H_O = (\mu_{before} == \mu_{after})$$
$$H_1 = (\mu_{before} > \mu_{after})$$

Calculate diff and check whether the difference follows normality or not

$$diff = before - after$$

Calculate T

$$t = \frac{\bar{x}_{diff}}{\frac{s_{diff}}{\sqrt{n}}}$$

$$\bar{x}_{diff} = \text{np.mean(diff)}$$
$$s_{diff} = \text{np.std(diff)}$$

Apply it in formula

```
n = len(differences)
t_statistic = mean_diff / (std_diff / np.sqrt(n))
df = n - 1

alpha = 0.05
p_value = stats.t.cdf(t_statistic, df)
# p_value = 0.54
```

$p - value > \alpha$ so we are **not rejecting** the **null hypothesis**.

Chi Square Test

It's used to determine if there is a significant association between categorical variables or if an observed distribution of categorical data differs from an expected theoretical distribution. It's based on chi square distribution, and it's commonly used in two main scenarios:

1. **Chi square Goodness-of-fit test:** This test is used to determine if the observed distribution of a **single categorical variable** matches an **expected theoretical distribution.** It's often applied to check if the data follows a specific probability distribution, such as the uniform or binomial distribution or poisson or among others.
2. **Chi square Test for independence** (chi-square test for association): This test is used to determine whether there is a significant association between two categorical variables in a sample.

The Chi-Square Goodness-of-Fit test is a **non-parametric test.** Non- parametric tests do not assume that the data comes from a specific probability distribution or make any assumptions about population parameters like the mean or standard deviation.

In the Chi-Square Goodness-of-Fit test, we compare the observed frequencies of the categorical data to the expected frequencies based on a hypothesized distribution. The **test doesn't rely on any assumptions about the underlying distribution's parameters**. Instead, it focuses on comparing observed counts to expected counts, making it a non- parametric test.

It's the **non-parametric test**, it will not take any assumptions about the data. Chi square based on **chi square distribution**.

Chi Square Distribution

It's also called χ^2, is a **continuous probability** distribution that is widely used in statistical hypothesis testing, particularly in the context of goodness of fit tests and test for independence in contingency tables. It arises when the sum of the squares of independent standard normal random variables follows this distribution.

$\chi^2 = Z^2$ The distribution of chi square forms once you square all the values of standard normal distribution. Here the $\text{degree of freedom} = 1$

If $\chi^2 = Z^2 + Z^2$. Here the degree of freedom = 2

If $\chi^2 = Z^2 + Z^2 + Z^3$. Here the degree of freedom = 3

Denoted as $\chi^2 = \sum_{i=1}^{k} Z_k^2, df = k$

If you increase the degree of freedom, the chi square distribution will approximate the normal distribution, it's very similar to normal distribution.

The chi square distribution has a **single parameter**, the `degree of freedom (df)` which influences the shape and spared of the distribution. The degrees of freedom are typically associated with the number of independent variables or constraints in a statistical problems.

Key Points

1. It's a continuous distribution, defined for non-negative values
2. It's a positively skewed, with the degree of skewness decreasing as the degrees of freedom increases.

3. The **mean** of **chi-square distribution** is equal to it's **degree of freedom** K and variance is $2k$
4. As the degree of freedom increases, the chi square distribution approaches the normal distribution in shape.

The chi square distribution is used in various statistical tests, such as the chi square goodness of fit test which evaluates whether an observed frequency distribution fits an expected theoretical distribution, and the chi square test for independence which checks the association between categorical variables in a contingency table.

Goodness of Fit Test

The Chi-Square **Goodness-of-Fit test** is a statistical hypothesis test used to determine if the observed distribution of a single categorical variable matches an expected theoretical distribution.

It helps to evaluate whether the data follows a specific probability distribution, such as uniform, binomial, or Poisson distribution, among others. This test is particularly useful when you want to assess if the **sample data** is **consistent** with an **assumed distribution** or if there are **significant deviations** from the **expected pattern**.

Assumptions

1. **Independence**: The observations in the sample must be independent of each other. This means that the outcome of one observation should not influence the outcome of another observation.
2. **Categorical data:** The variable being analysed must be categorical, not continuous or ordinal. The data should be divided into mutually exclusive and exhaustive categories.
3. **Expected frequency:** Each category should have an expected frequency of at **least 5.** This guideline helps ensure that the Chi- Square distribution is a reasonable approximation for the distribution of the test statistic. Having small expected frequencies can lead to an inaccurate estimation of the Chi-Square distribution, potentially increasing the likelihood of a Type I error (incorrectly rejecting the null hypothesis) or a Type II error (incorrectly failing to reject the null hypothesis).

4. **Fixed distribution:** The theoretical distribution being compared to the observed data should be specified before the test is conducted. It is essential to avoid choosing a distribution based on the observed data, as doing so can lead to biased results.

Expected value or expected frequencies and Theoretical value or observed value :

Let's say a candy company claims that their candy colors are **evenly distributed**:

- Red, Blue, Green, and Yellow — each 25%.

You open a bag of **100 candies**, count them, and get:

Color	Observed (O)
Red	30
Blue	20
Green	25
Yellow	25

Now, if the company's claim is true, the **expected count** (E) for each color =

$$E = \text{Total Candies} \times \text{Expected Proportion} = 100 \times 0.25 = 25$$

So,

Color	Observed (O)	Expected (E)
Red	30	25
Blue	20	25
Green	25	25
Yellow	25	25

So, in short

Term	Meaning
Observed (O)	What you actually saw or counted in data
Expected (E)	What you would expect if the null hypothesis were true

Steps

1. Define the null hypothesis (H0) and the alternative hypothesis (H1):
   H0: The observed data follows the expected theoretical distribution.
   H1: The observed data does not follow the expected theoretical distribution.
2. Calculate the expected frequencies for each category based on the theoretical distribution and the sample size.
3. Compute the Chi-Square test statistic (χ2) by comparing the observed and expected frequencies. The test statistic is calculated as:

   $\chi 2 = \frac{\sum (O_i - E_i)^2}{E_i}$

   Oi is the observed frequency in category i

   Ei is the expected frequency in category i

   $\sum$ is taken over all categories.
4. Determine the degrees of freedom (df), which is typically the number of categories minus one (df = k - 1), where k is the number of categories.
5. Calculate the p-value for the test statistic using the Chi-Square distribution with the calculated degrees of freedom.
6. Compare the test statistic to the critical value or the p-value

Example Problems

Suppose we have a six-sided fair die, and we want to test if the die is indeed fair. We roll the die 60 times and record the number of times each side comes up. We'll use the Chi-Square

Goodness-of-Fit test to determine if the observed frequencies are consistent with a fair die (i.e., a uniform distribution of the sides).

Observed frequencies:
- Side 1: 12 times
- Side 2: 8 times
- Side 3: 11 times
- Side 4: 9 times
- Side 5: 10 times
- Side 6: 10 times

Hypothesis

H_0 : The die is fair (uniform distribution)
H_1 : The die is not fair

Variable Definition & Expected frequency calculation

Side	Obserev Frequency	Expected Frequency
1	12	10
2	8	10
3	11	10
4	9	10
5	10	10
6	10	10

$alpha = 0.05$

Chi Square Statistic for goodness of fit

$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(12-10)^2+(8-10)^2+(11-10)^2+(9-10)^2+(10-10)^2+(10-10)^2}{10} = \frac{4+4+1+1+0+0}{10} = \frac{10}{10} = 1$

Degrees of Freedom

$df = n - 1 = 6 - 1 = 5$

P value calculation (look chi table or use python function)

$\text{p-value for test static 1 is} = 0.37059$

Conclusion

$\text{Since } p > 0.05, \text{ we fail to reject } H_0. \text{ The die is fair.}$

Test for independence

The Chi-Square test for independence, also known as the Chi-Square test for association, is a statistical test used to determine whether there is a significant association between two categorical variables in a sample. It helps to identify if the occurrence of one variable is dependent on the occurrence of the other variable, or if they are independent of each other.

The test is based on comparing the observed frequencies in a contingency table (a table that displays the frequency distribution of the variables) with the frequencies that would be expected under the assumption of independence between the two variables.

Assumptions

1. **Independence of observations:** The observations in the sample should be independent of each other. This means that the occurrence of one observation should not affect the occurrence of another observation. In practice, this usually implies that the data should be collected using a simple random sampling method.
2. **Categorical variables**: Both variables being tested must be categorical, either ordinal or nominal. The Chi-Square test for independence is not appropriate for continuous variables.
3. **Adequate sample size:** The sample size should be large enough to ensure that the expected frequency for each cell in the contingency table is sufficient. A common rule of thumb is that the expected frequency for each cell should be **at least 5**. If some cells have expected frequencies less than 5, the test may not be valid, and other methods like Fisher's exact test may be more appropriate.
4. **Fixed marginal totals:** The marginal totals (the row and column sums of the contingency table) should be fixed before the data is collected. This is because the Chi-Square test for independence assesses the association between the two variables under the assumption that the marginal totals are fixed and not influenced by the relationship between the variables.

Steps

1. State the null hypothesis (H0) and alternative hypothesis (H1):
   H0: There is no association between the two categorical variables (they are independent).
   H1: There is an association between the two categorical variables (they are dependent).

2. Create a contingency table with the observed frequencies for each combination of the categories of the two variables.

3. Calculate the expected frequencies for each cell in the contingency table assuming that the null hypothesis is true (i.e., the variables are independent).
4. Compute the Chi-Square test statistic:
   $\chi 2 = \Sigma[(O_i j - E_i j)2/E_i j]$

   where $O_i j$ is the observed frequency in each cell and $E_i j$ is the expected frequency.
5. Determine the degrees of freedom: df = (number of rows - 1) * (number of columns - 1)
   Obtain the critical value or p-value using the Chi-Square distribution table or a statistical software/calculator with the given degrees of freedom and significance level (commonly α = 0.05).

6. Compare the test statistic to the critical value or the p-value to the significance level to decide whether to reject or fail to reject the null hypothesis. If the test statistic is greater than the critical value, or if the p-value is less than the significance level, we reject the null hypothesis and conclude that there is a significant association between the two variables.

One way ANNOVA

One-way ANOVA (Analysis of Variance) is a statistical method used to compare the **means** of **three** or **more independent groups** to determine if there are any significant differences between them. It is an extension of the t-test, which is used for comparing the means of two independent groups. **In simple terms, If you have more than 2 independent groups and you want to compare the means between them then use ANNOVA.**

The term "one-way" refers to the fact that there is only one independent variable (factor) with multiple levels (groups) in this analysis. Ex: Gender (you have Male, Female, Non binary, some other categories)

The primary purpose of one-way ANOVA is to test the null hypothesis that all the group means are equal. The alternative hypothesis is that at least one group mean is significantly different from the others.

It works based on the F distribution

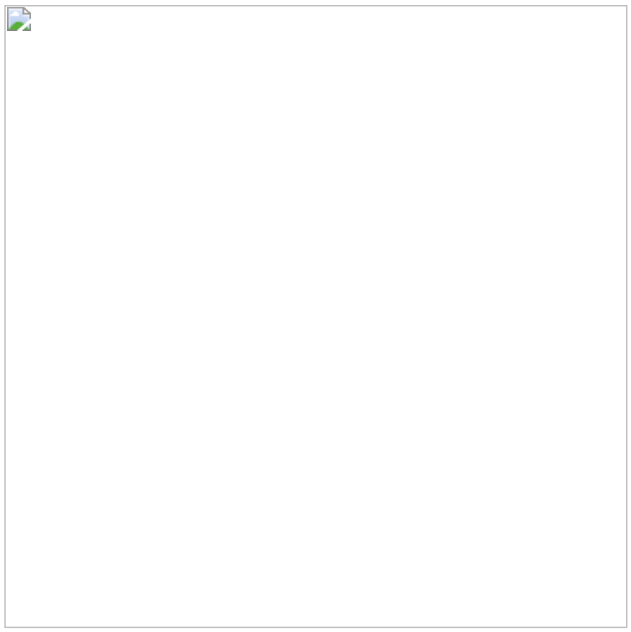
F Distribution

It's actually related to chi-square distribution (chi-square is actually related to normal distribution).

How to form F distribution?

Take two chi square χ_1^2, χ_2^2 and respective degrees of freedom df_1, df_2 and put this in this formula, then you get the f distribution.

$$F = \frac{\frac{\chi_1^2}{df_1}}{\frac{\chi_2^2}{df_2}}$$



1. **Continuous probability distribution:** The F-distribution is a continuous probability distribution used in statistical hypothesis testing and analysis of variance (ANOVA).
2. **Fisher-Snedecor distribution**: It is also known as the Fisher-Snedecor distribution, named after Ronald Fisher and George Snedecor, two prominent statisticians.
3. **Degrees of freedom:** The F-distribution is defined by **two parameters** - the degrees of freedom for the numerator (df1) and the degrees of freedom for the denominator (df2).
4. **Positively skewed and bounded**: The shape of the F-distribution is positively skewed, with its left **bound** at **zero**. The distribution's shape depends on the values of the degrees of freedom. (bounded means zero)
5. **Testing equality of variances:** The F-distribution is commonly used to test hypotheses about the equality of two variances in different samples or populations.
6. **Comparing statistical models:** The F-distribution is also used to compare the fit of different statistical models, particularly in the context of ANOVA.
7. **F-statistic:** The F-statistic is calculated by dividing the ratio of two sample variances or mean squares from an ANOVA table. This value is then compared to critical values from the F-distribution to determine statistical significance.
8. **Applications**: The F-distribution is widely used in various fields of research, including psychology, education, economics, and the natural and social sciences, for hypothesis testing and model comparison.

ONE WAY ANNOVA STEPS

Data

A	B	C
3	1	8
6	8	6
3	9	10

$n = 9; k = 3$

k is total categories

STEP 1 (Create Hypothesis)

$H_0 : \mu_A = \mu_B = \mu_C \quad \text{(All group means are equal)}$
$H_1 : \text{At least one group mean is significantly different}$

Ho basically what we are trying to prove is all three groups are coming from same population.

STEP 2

Calculate the overall mean (grand mean) of all the groups combined and mean of all the groups individually.

$$\text{Grand Mean calculation}$$
$$\bar{X}_A = \frac{3+6+3}{3} = 4$$
$$\bar{X}_B = \frac{1+8+9}{3} = 6$$
$$\bar{X}_C = \frac{8+6+10}{3} = 8$$
$$\bar{X}_{\text{grand}} = \frac{3+6+3+1+8+9+8+6+10}{9} = \frac{54}{9} = 6$$
$$\text{Individual Group Means}$$
$$\bar{X}_A = \frac{3+6+3}{3} = 4$$
$$\bar{X}_B = \frac{1+8+9}{3} = 6$$
$$\bar{X}_C = \frac{8+6+10}{2} = 8$$

STEP 3

3.1 Calculate SST

SST → sum of squared total $\sum(x_i - \bar{X}_{grand})^2$

$$\text{Sum of Squares Total (SST)}$$
$$\text{Grand Mean} = 6$$
$$\text{SST} = (6-3)^2 + (6-6)^2 + (6-3)^2 + (6-1)^2 + (6-8)^2 + (6-9)^2 + (6-5)^2 + (6-6)^2 + (6-10)^2$$
$$= 9 + 0 + 9 + 25 + 4 + 9 + 4 + 0 + 16 = 76$$

This is called **overall variance.**

$df = n - 1$
$df = 9 - 1 = 8$

3.2 Calculate SSW

SSW → Sum of Squares within

"To calculate the **within-group variance,** subtract each value in a group from that group's mean. For example, for Category A, subtract each value from the mean of Category A; similarly, for Category B, subtract each value from the mean of Category B. Repeat this process for all groups, then square the differences and sum them to compute the Sum of Squares Within (SSW)."

$$\text{Sum of Squares Within (SSW)}$$
$$\text{SSW} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

$$\text{Sum of Squares Within (SSW)}$$
$$\text{Group A: } (3-4)^2 + (6-4)^2 + (3-4)^2 = 1 + 4 + 1 = 6$$
$$\text{Group B: } (1-6)^2 + (8-6)^2 + (9-6)^2 = 25 + 4 + 9 = 38$$
$$\text{Group C: } (8-8)^2 + (10-8)^2 + (6-8)^2 = 0 + 4 + 4 = 8$$
$$\text{SSW} = 6 + 38 + 8 = 52$$

3.3 Calculate Degrees of freedom for SSW

$$\text{df}_{\text{within}} = n - k$$
$$\text{df}_{\text{within}} = 9 - 3 = 6$$

Another way to calculate

- $\text{Group A: } 3 \text{ values} \rightarrow \text{df}_A = 3 - 1 = 2$
- $\text{Group B: } 3 \text{ values} \rightarrow \text{df}_B = 3 - 1 = 2$
- $\text{Group C: } 3 \text{ values} \rightarrow \text{df}_C = 3 - 1 = 2$

3.4 Calculate Sum of Squares Between

$$\text{Sum of Squares Between (SSB)}$$
$$\text{SSB} = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2$$
$$\text{expansion}$$
$$\text{SSB} = n_A(\bar{X}_A - \bar{X}_{grand})^2 + n_B(\bar{X}_B - \bar{X}_{grand})^2 + n_C(\bar{X}_C - \bar{X}_{grand})^2 + ..n$$

Where:

- n_A, n_B, n_C are the number of observations in groups A, B, and C respectively
- $\bar{X}_A, \bar{X}_B, \bar{X}_C$ are the means of each group
- $\bar{X}_{grand}$ is the grand mean

$$\text{Sum of Squares Between (SSB)}$$
$$\bar{X} = 6 \quad \text{(Grand Mean)}$$
$$\text{Group A: } n_A = 3, \bar{X}_A = 4$$
$$\text{Group B: } n_B = 3, \bar{X}_B = 6$$
$$\text{Group C: } n_C = 3, \bar{X}_C = 8$$
$$\text{SSB} = 3(6-4)^2 + 3(6-6)^2 + 3(6-8)^2$$

$= 3(4) + 3(0) + 3(4) = 12 + 0 + 12 = 24$

$\text{df}_{\text{between}} = k - 1$
$\text{df}_{\text{between}} = 3 - 1 = 2$

All the values

Quantity	Value	df
SSB	24	2
SSW	52	6
SST	76	8

$SST = SSW + SSB$

This is the universal fact.

SST (total variance) = SSW (internal variance groups) + SSB (main variance with local variance)

3.5 Calculate F Statistic

You calculate F static by MSB and MSW

1. **MSB (Mean Square Between):**

   MSB represents the variance between different groups. It is calculated by dividing the Sum of Squares Between (SSB) by its degrees of freedom.

   $\text{MSB} = \frac{\text{SSB}}{\text{df}_{\text{between}}}$

   In our example: $\text{MSB} = \frac{24}{2} = 12$

2. **MSW (Mean Square Within):**

   MSW represents the variance within groups. It is calculated by dividing the Sum of Squares Within (SSW) by its degrees of freedom.

   $\text{MSW} = \frac{\text{SSW}}{\text{df}_{\text{within}}}$

   In our example: $\text{MSW} = \frac{52}{6} = 8.67$

The F-statistic is the ratio of these two mean squares:

$F = \frac{\text{MSB}}{\text{MSW}}$
$F = \frac{\text{MSB}}{\text{MSW}} = \frac{\text{SSB}/\text{df}_{\text{between}}}{\text{SSW}/\text{df}_{\text{within}}}$

This F static follows the f distribution.

$F = \frac{\text{MSB}}{\text{MSW}} = \frac{24/2}{52/6} = \frac{12.00}{8.67} = 1.38$

3.6 Calculate P Value

You basically find the right side area using f table, are you can use this function bellow.

```
import scipy.stats as stats

f_statistic = 1.4  # The F-statistic value you've calculated
df1 = 2            # Degrees of freedom for the numerator (between groups)
df2 = 6           # Degrees of freedom for the denominator (within groups)

p_value = stats.f.sf(f_statistic, df1, df2)
print("P-value:", p_value)
# p-value = 0.31
```

Null hypothesis can't be rejected, $\mu_a = \mu_b = \mu_c$

// need to work on geometric intuition

Geometric intuition

In ANOVA, we start with a large population and aim to determine whether three subgroups (or categories) are derived from the same population (**null hypothesis, H_0**) or from different populations (**alternative hypothesis, H_1**).

Each category is assumed to follow a normal distribution. We visualize these distributions on a graph and compute two key metrics:

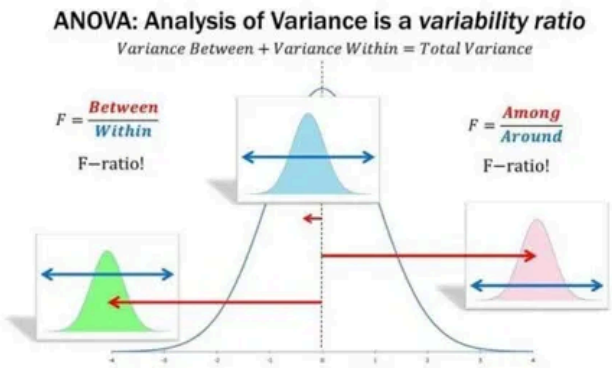
1. **SSB (Sum of Squares Between)**: This measures the variation between the means of each category and the overall population mean. It captures how far each group mean is from the grand mean.
2. **SSW (Sum of Squares Within)**: This measures the variation within each category — essentially the variance inside each group.

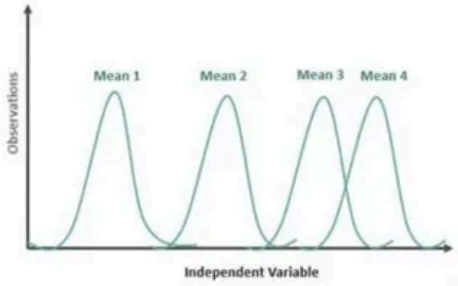
The total variation is the sum of SSB and SSW.

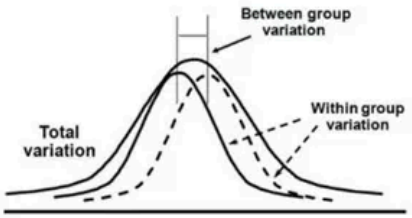
A high **SSB** indicates that the group means are far from the overall mean, suggesting that the groups are distinct. In such cases, **SSW** tends to be lower because the individual data points within each group are closer to their respective group means. This inverse relationship makes **SSB** a critical factor in determining the **F-statistic**.

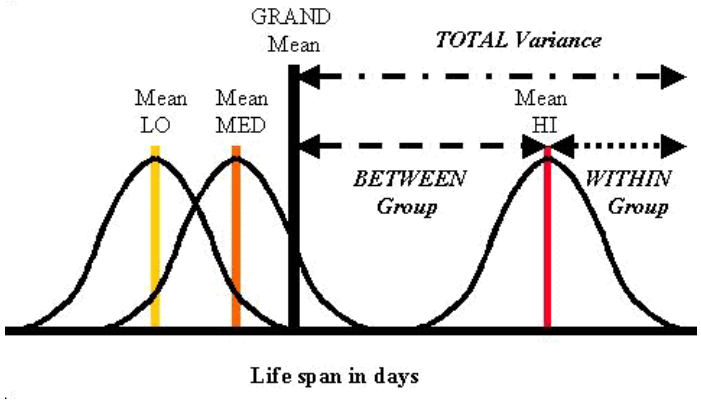
If **SSB** is large relative to **SSW**, the **F-statistic** will be high, which increases the likelihood of rejecting the null hypothesis — indicating that the groups are likely from different populations.

Analysis of Variance (ANOVA)









Assumptions:

1. **Independence:** The observations within and between groups should be independent of each other. This means that the outcome of one observation should not influence the outcome of another. Independence is typically achieved through random sampling or random assignment of subjects to groups.
2. **Normality:** The data within each group should be approximately normally distributed. While one-way ANOVA is considered to be robust to moderate violations of normality, severe deviations may affect the accuracy of the test results. If normality is in doubt, non-parametric alternatives like the Shapiro-wilk test can be considered.

3. **Homogeneity of variances:** The variances of the populations from which the samples are drawn should be equal, or at least approximately so. This assumption is known as homoscedasticity. If the variances are substantially different, the accuracy of the test results may be compromised. Levene's test or Bartlett's test can be used to assess the homogeneity of variances. If this assumption is violated, alternative tests such as Welch's ANOVA can be used.

Post Hoc Test

Post hoc tests, also known as post hoc **pairwise comparisons** or **multiple comparison tests**, are used in the context of ANOVA when the overall test indicates a significant difference among the group means. These tests are performed after the initial one-way ANOVA to determine which **specific groups** or **pairs of groups have significantly different means**.

In simple terms, you conducted an one way ANNOVA test and it's failed, now you want to know that which category (A,B,C) is failed or all the categories are failed ?, to find that you use Post Hoc test.

The main purpose of post hoc tests is to control the **family-wise error rate (FWER)** and adjust the significance level for multiple comparisons to avoid inflated **Type I errors**. There are several post hoc tests available, each with different characteristics and assumptions. Some common post hoc tests include:

1. **Bonferroni correction:** This method adjusts the significance level (α) by dividing it by the number of comparisons being made. It is a conservative method that can be applied when making multiple comparisons, but it may have lower statistical power when a large number of comparisons are involved.

   Example:

   You have 3 categories [A, B, C], you do three independent 2 sample t test

   t test ⇒ (a,b) ⇒ p value

   t test ⇒ (b, c) ⇒ p value

   t test ⇒ (c, a) ⇒ p value

   You compare p-value and you find the culprit, but the problem is you are doing the t-test for 3 time, for each time you have a significance value of 0.05 (5%) means you can make 5% mistakes False positives, you are combinely doing three independent events so 0.05 * 3 = 0.15, it means now you can make 15% mistakes False positives. that's the draw back in this error. So to solve this **we divide by total categories** for **each test** to get the **finite p-value**.
2. **Tukey's HSD (Honestly Significant Difference) test**: This test controls the FWER and is used when the sample sizes are equal and the variances are assumed to be equal across the groups. It is one of the most commonly used post hoc tests.

When performing post hoc tests, it is essential to choose a test that aligns with the assumptions of your data (e.g., equal variances, equal sample sizes) and provides an appropriate balance between controlling Type I errors and maintaining statistical power.

Significance value: 0.05% basically you can make a mistake 5% of time. it's basically a TYPE 1 error.

Why t-test is not used for more than 3 categories?

1. **Increased Type I error**: When you perform multiple comparisons using individual t-tests, the probability of making a Type I error (false positive) increases. The more tests you perform, the higher the chance that you will incorrectly reject the null hypothesis in at least one of the tests, even if the null hypothesis is true for all groups.
2. **Difficulty in interpreting results:** When comparing multiple groups using multiple t-tests, the interpretation of the results can become complicated. For example, if you have 4 groups and you perform 6 pairwise t-tests, it can be challenging to interpret and summarize the overall pattern of differences among the groups.
3. **Inefficiency:** Using multiple t-tests is less efficient than using a single test that accounts for all groups, such as one-way ANOVA. One-way ANOVA uses the information from all the groups simultaneously to estimate the variability within and between the groups, which can lead to more accurate conclusions.

Applications in machine learning

1. **Hyperparameter tuning:** When selecting the best hyperparameters for a machine learning model, one-way ANOVA can be used to compare the performance of models with different hyperparameter settings. By treating each hyperparameter setting as a group, you can perform one-way ANOVA to determine if there are any significant differences in performance across the various settings.
2. **Feature selection:** One-way ANOVA can be used as a univariate feature selection method to identify features that are significantly associated with the target variable, especially when the target variable is categorical with more than two levels. In this context, the one-way ANOVA is performed for each feature, and features with low p-values are considered to be more relevant for prediction.
3. **Algorithm comparison**: When comparing the performance of different machine learning algorithms, one-way ANOVA can be used to determine if there are any significant differences in their performance metrics (e.g., accuracy, F1 score, etc.) across multiple runs or cross-validation folds. This can help you decide which algorithm is the most suitable for a specific problem.
4. **Model stability assessment:** One-way ANOVA can be used to assess the stability of a machine learning model by comparing its performance across different random seeds or initializations. If the model's performance varies significantly between different initializations, it may indicate that the model is unstable or highly sensitive to the choice of initial conditions.

Test of Variances

To check whether the variances of the two populations are the same (homoscedasticity) or not (Heteroscedasticity), we can use **Levene's test** or **Bartlett's test**.

In simple words:

- **Homoscedasticity** = equal variances across groups.
- **Heteroscedasticity** = unequal variances.

So we test:

- **Null hypothesis (H_0):** Variances are equal.
- **Alternative hypothesis (H_1):** Variances are not equal.

Rule of thumb

If $p > 0.05$: then the variances are equal (Ho)

If $p < 0.05$: then the variances are not-equal (H1 or Ha)

```
from scipy.stats import levene

# Example data for two groups
group1 = [10, 12, 13, 9, 11, 10]
group2 = [20, 21, 19, 22, 20, 18]

# Perform Levene's test
stat, p = levene(group1, group2)

print('Statistic = %.3f, p-value = %.3f' % (stat, p))

# Interpretation
if p > 0.05:
    print("Variances are equal (homoscedasticity)")
else:
    print("Variances are not equal (heteroscedasticity)")
```

Probability Distributions

1. Random Variables

A Random Variables is a set of possible outcomes from random experiment. Random experiment is an any kind of experiment whose outcome is random. You denote random variable as **CAPITAL** letters.

Example: tossing a coin X = {1, 0} here X is a random variable and the set is called **sample space**.

There are **two** types,

1. Discrete; X = {1, 2, 3}
2. Continuous (range of values), Y = {0, 10}, it's between 0 to 10

---

2. Probability Distributions

A probability distribution is a list of possible outcome of a random variable along with their corresponding probability values. To get the corresponding probability

values for the random variable, we use a function called `probability distribution function`. Which takes outcome as input and outputs the corresponding probability values.

Generating Probability distribution for rolling two dice in a same time

Step 1: Cross tab of all possible outcomes:

	1	2	3	4	5	6
1	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
2	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
3	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
4	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
5	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
6	(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

Step 2: Sum all outcomes

$+$	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

This table is called Probability distribution

Step 3 Calculate Results

If $P(X = 2) = \frac{1}{36}$
because two occured only one

If $P(X = 3) = \frac{2}{36}$
because3 occurred two times

This is how you generate a probability for particular value, but this process is tedious and also for continuous values, this table will expand like hell.

Instead of creating a PDF, we can create a cross tab by having all the possible outcomes but it will become huge whenever you're trying to find the probability for the continuous random values. So, it's better to use the **probability distribution function** which will output the probability values for the specific outcome.

PDF: it's a mathematical relationship between all the possible outcomes of a random variable with the probabilities of the outcome. Probability distribution and probability distribution functions words are used interchangeably.

Types of data for PDF :

1. Discrete Random variable
2. Continuous Random Variable

Why PDF is so important?

1. It gives you the idea about the shape / distribution of the data.
2. And if our data follows a any distribution, then you can automatically know a lot about your data.

Parameter in PDF

In every probability distribution you have some parameters, it's basically a tuning knobs. If you change this you will see the graph structure will change.

Parameter in probability distributions are numerical values that determine the shape, location and scale of the distribution. Different probability distributions

have different sets of parameters that determine their shape and characteristics, and understanding these parameters is essential in statistical analysis.

Types of PDF

1. Probability Mass Function (PMF); it's for **Discrete random** variable
2. Probability Density Function (PDF); it's for **continuous random** variable
3. Cumulative Distribution Function (CDF); it's used for **both** variables.

Probability distribution function is a umbrella term, inside it you have two types.

---

3. Probability Mass Functions (PMF)

It's a mathematical function that describes the probability distribution of a **discrete** random variable. It assigns a probability to each possible value of the random variable. The probabilities assigned by the PMF must satisfy two conditions:

1. Must be non-negative; `>0`
2. The sum of all probabilities must equal to `1`.

For a discrete random variable X:

$$P(X = x_i) = \frac{\text{Number of times outcome } x_i \text{ occurs}}{\text{Total number of experiments}}$$

```
# Import necessary library
from collections import Counter

# Simulate outcomes of rolling a die 60 times
rolls = [1, 2, 3, 4, 5, 6, 1, 3, 2, 6, 5, 4, 3, 2, 1, 6, 5,
4, 1, 2,
```

```
          3, 4, 5, 6, 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, 1,
2, 3, 4,
          5, 6, 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, 1, 2, 3,
4, 5, 6]

# Count frequency of each outcome
counts = Counter(rolls)

# Total number of rolls
n = len(rolls)

# Calculate PMF
pmf = {outcome: freq/n for outcome, freq in counts.items()}

print("PMF of rolling a die:")
for outcome, prob in pmf.items():
    print(f"P(X={outcome}) = {prob:.2f}")
```

CDF for PMF

The Cumulative Distribution Function describes the probability that a random variable X with a given probability distribution (in our case PMF) will be found at a value less than or equal to x.

In PMF you calculate probability for F(3) = 3/6 but in CMF you calculate probability for `F(x≤3)` means what's the probability of it **being 3** or **less than 3.**

so F(3) =

$$F(x) = P(X \leq x)$$
$$f(x <= 3) = f(3) + f(2) + f(1) + f(0)$$

Summary:

- **PMF** gives the probability of a **single value**.
- **CDF** gives the **cumulative probability** up to and including that value.

---

4. Probability Density Function (PDF)

It's a mathematical function that describes the probability distribution of a **continuous random variable**

In PMF **Y axis** has **probabilities** but in PDF the **Y axis** has **Probability density** instead of probability, this is the main difference between PDF and PMF.

1. Why Probability density and why not probability?

   ▼ answer

   Here we are dealing with continuous values, let's take `7.912` , if you want to find probability for the value it would be zero, because here we are dealing with infinite values, it's impossible to find probabilities for all the values in the range. So that's the reason we use probability density function.

2. What does the are of this graph represents?

   ▼ answer

3. What is probability density?

   Probability density describes how likely a continuous random variable (outcome) is to take on a value within a certain range. So you basically calculate **area** between the range using Integration.

   To find the probability that X lies between two values A and B, you calculate:

$$PDF = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Density Estimation

Density Estimation is a statistical technique used to estimate the probability density of a random variable based on a set of observations of data. It can be used in hypothesis testing, data analysis, and in machine learning used in estimate probability distribution of input data or to model the likelihood of certain events or outcomes.

In simpler terms, it involves estimating the underlying distribution of a set of data points. Again this is not a actual probability it's an estimation.

There are various methods;

1. Parametric (it assumes the data follows a specific probability distribution)
2. non-parametric (does not make any assumption about the distribution and instead estimate it directly from data)

Commonly used for density estimation include **kernel density estimation(KDE)**, **histogram estimation**, and **Gaussian mixture models**.

The choice of method depends on the specific characteristics of the data and the intended use of the density estimate.

Parametric Density Estimation

Parametric density estimation is a method of estimating the probability density function (PDF) of a random variable by assuming that underlying distribution belongs to a specific parametric family of probability distribution, such as the normal, exponential, or poisson distribution.

Let's say you have a sample data which is very similar to normal distribution. First step you to is to estimate the population mean and population standard deviation. Then you can take the PDF equation to calculate the probability Density estimation for all outcome. You calculate using this formula;

$$PDF = \frac{1}{\sigma\sqrt{2}\pi} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Non Parametric Density Estimation

But sometimes the distribution is not clear or it's not one of the famous distribution. You use Non Parametric Density Estimation.

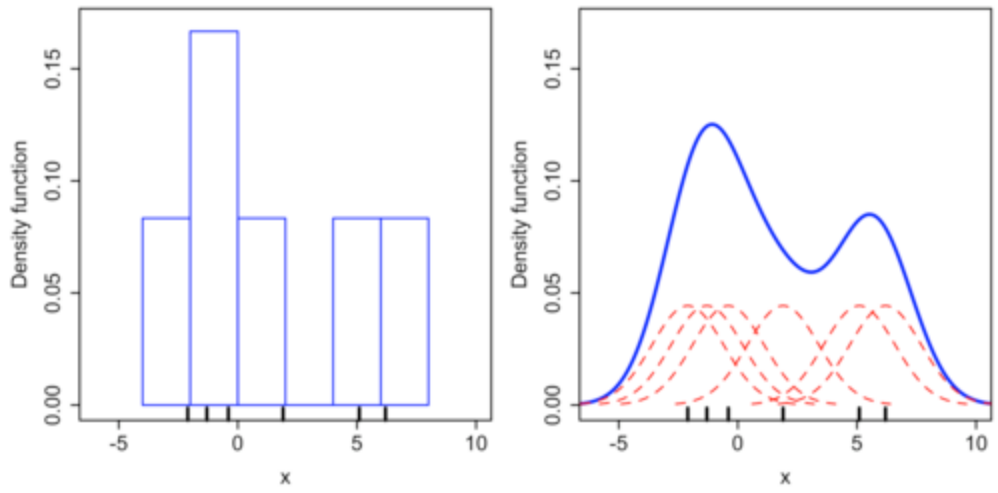
It's a Statistical technique used to estimate the probability density function of a random variable without making any assumptions about the underlying distribution. It's non-parametric density estimation because it does not require the use of a pre defined probability distribution function, as opposed to parametric methods such as the **Gaussian distribution**.

The non parametric density estimation technique involves constructing an estimate of the probability density function using the available data. this is typically by creating **Kernel density estimate**.

This is computationally intensive and may require more data to achieve accurate estimates compared to parametric methods.

Kernel Density Estimation

The KDE technique involves using a kernel function to **smooth out** the data and create a **continuous estimate** of the underlying density function.



Step 1: For each data point, create a small normal (Gaussian) distribution

Each observation contributes a **kernel** — typically a normal distribution centered at that data point. These small curves represent local density around each point.

Step 2: Choose a bandwidth (smoothing parameter)

The **bandwidth (h)** controls the **width** of each kernel:

- Small h → very narrow, noisy (overfitting)
- Large h → too wide, oversmoothed (underfitting)

Step 3: Sum (or average) all the small distributions

You sum up all the small normal curves to obtain the overall smooth density estimate. In simpler terms, if you're currently evaluating the density at a specific value — say `x = 5` you look at how many of those small distributions overlap at that point and add up their contributions to get the final `y-value`.

You're calculating the **density at x=5**.

- You have many small normal curves centered at your data points (say 3.2, 4.8, 5.1, 6.0...).
- Some of these curves have nonzero height at x=5.
- Each curve gives you a small **value** (like 0.03, 0.07, 0.05...).
- You **add up all these small values** — these are their **contributions** — to get the final height of the KDE curve at x=5.

A “contribution” is the **amount each kernel adds to the estimated density** at a given point.

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where

- $K(.)$ = kernel function (often Gaussian)
- h = bandwidth
- n = number of data points

Step 4: Normalize it

Ensure the total area under the KDE curve equals 1 — just like a probability density function.

Step 5: Plot or use it to estimate probabilities

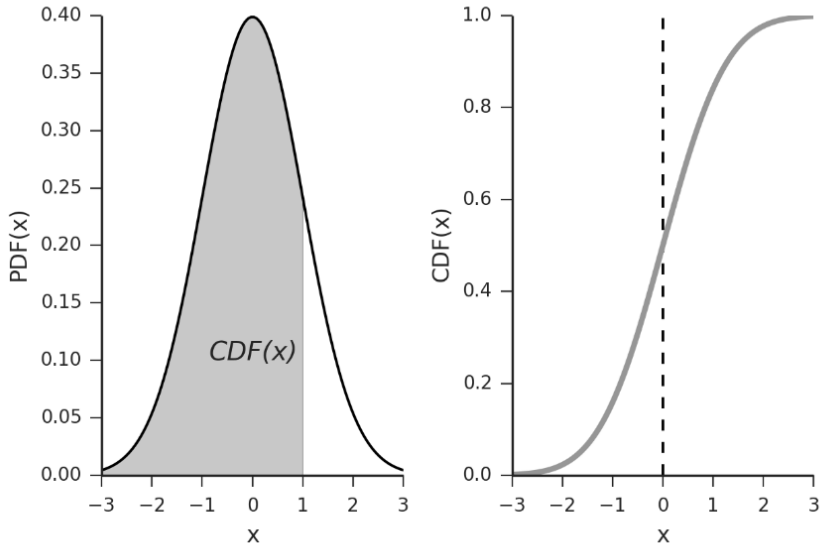
Now you can:

- Plot the smooth curve to visualize the underlying distribution.
- Estimate probabilities over intervals using the KDE function.

Cumulative Distribution Function for PDF

$$F(x) = P(X \leq x)$$
$$f(x <= 3) = f(3) + f(2) + f(1) + f(0)$$

Here in the Y axis you will get probability instead of probability density because you add the current x and previous values.



Let's say you're at 1 in PDF function, to calculate CDF for PDF, you calculate the area until 1 in PDF, so you get the probability.

How to Use PDF in Data Science

Used in analysing the pattern in the data.

How to Use CDF in Data Science

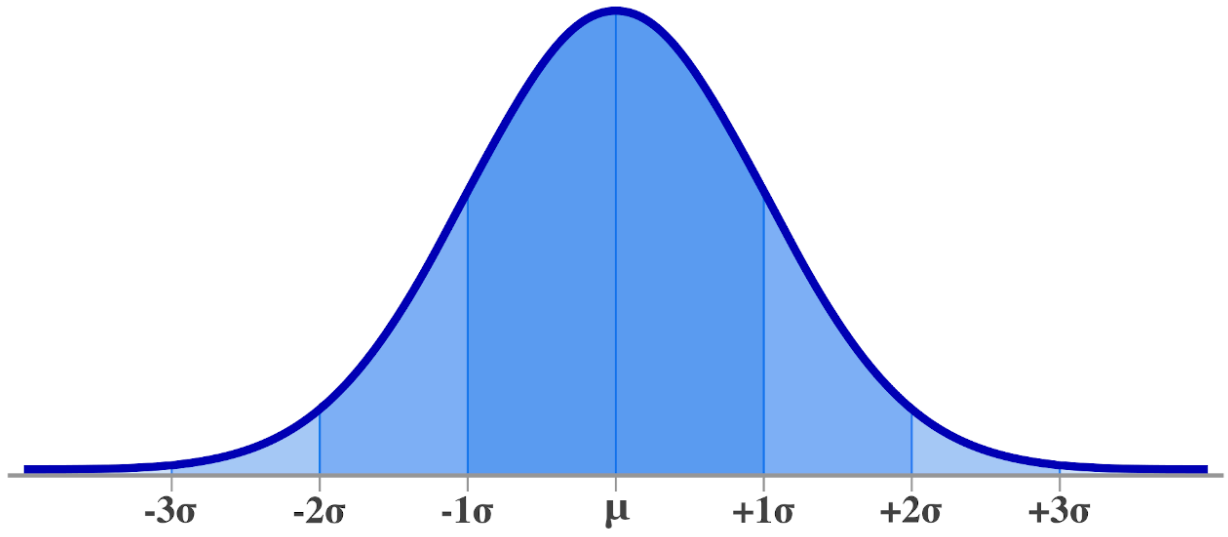
You make a decision boundaries based on PDF, but to validate whether the decision is correct or not we use CDF.

95% versicolor flowers will be less than 1.7 (current data point) whereas only 10% vergeninca flowers will be less than 1.7

---

5. Normal Distribution

Normal Distribution is also known as **Gaussian Distribution**, is a probability distribution that is commonly used in statistical analysis. It's a continuous probability distribution (PDF) that is symmetrical around the mean, with a bell-shaped curve.



Observations:

1. center is Mean
2. Y axis is probability density values

3. Tail (3, -3) it's not touching x axis, it means asymptomatic, It goes until infinity and touch infinity.
4. There are many points near to centre and some points far away.

It has two parameters

1. mean μ (represents center of the distribution)
2. Standard deviation σ (represents spread of the distribution)

Denoted as $X \sim \mathcal{N}(\mu, \sigma)$ in lectures

Why it's so important?

Commonality in Nature. Many Natural Phenomena follow a normal distribution, such as heights, weights, IQ and many more. Thus the Normal distribution provides a convenient way to model and analyse such data.

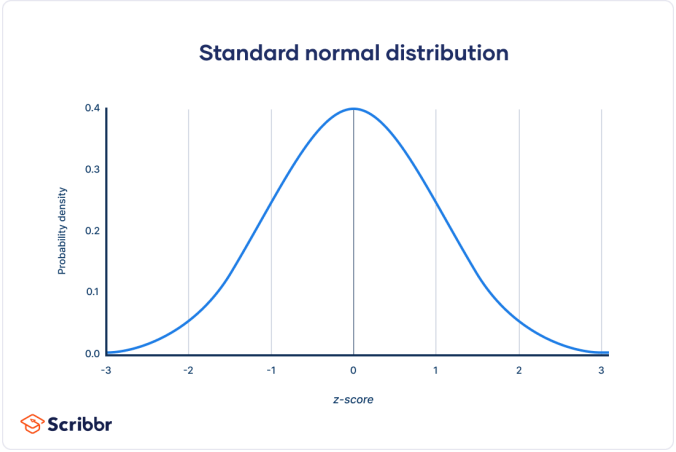
PDF equation of Normal Distribution

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Standard Normal Variate(Z) or Standard Normal Distribution

If $\mu = 0, \sigma = 1$ it's a special type of Normal distribution called Standard Normal Variate.

$$Z \sim \mathcal{N}(0, 1)$$



Benefits

- Standardising a **normal distribution** allows us to compare **different distributions** with each other, and to calculate probabilities using **standardised tables** or software.

$$PDF = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$
$$\mu = 0; \sigma = 1$$
$$SDF = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{1}{2}x^2\right)}$$

To convert any Normal distribution to Standard Normal Distribution

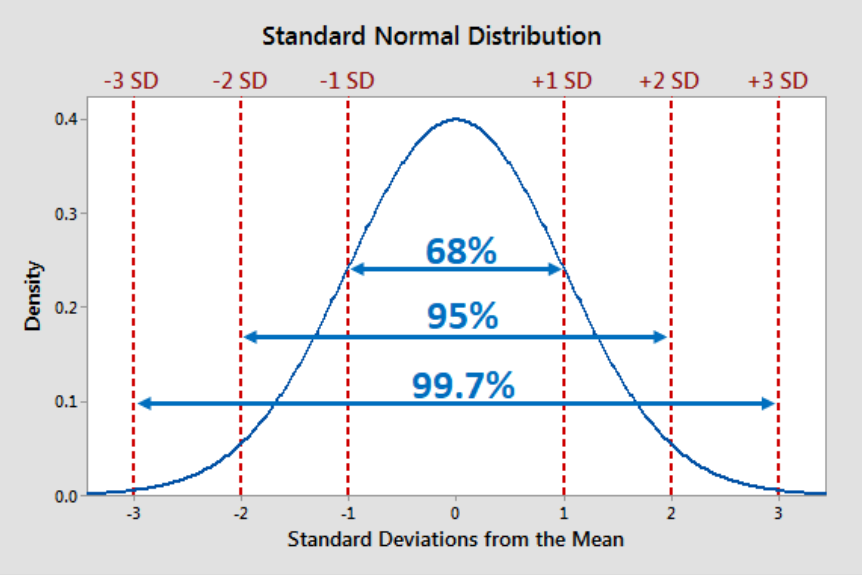
$$Z = \frac{x - \mu}{\sigma}$$

After doing this operation in all the data points, you will get standard normal distribution.

Why are we converting to Standard Normal Distribution? Let's say If you want to solve any question or you need to infer something from the normal distribution it's very difficult so we convert that to Standard Normal Distribution and we use `Z` table to find the probability for some specific point.

Empirical Rule

The normal distribution has well-known empirical rule, also known as 68-95-99.7 rule, which states that approximately 68% of the data falls within one standard deviation of the mean, about 95% of the data falls within two standard deviation of the mean, and about 99.7% of the data falls within three standard deviations of the mean.



Properties of Normal distribution

1. Symmetrical

   The normal distribution s symmetric about it's mean which means that the probability of observing a value above the means is the same as the

probability of observing a value below the mean. The bell shape curve of the normal distribution reflects this symmetry.

2. Measure of Central Tendencies are equal `mean==median==mode`.
3. Empirical Rule
4. The area under the curve is 1.

Skewness

A normal distribution is a bell-shaped, symmetrical distribution with a specific mathematical formula that describes how the data is spread out. Skewness indicates that the data is **not symmetrical**, which means it is **not normally distributed**.

Skewness measures the **asymmetry** of a probability distribution. It tells us whether the data is concentrated more on one side of the mean than the other. In other terms, it's a statistical measure that describes the degree to which a dataset deviates from the normal distribution.

In a Symmetrical distribution, the mean, median, and mode are all equal. In contrast, in a skewed distribution, the mean, median, and mode are not equal, and the distribution tends to have a longer tail on one side than the other side.

The greater the skew the greater the distance between mode, median and mean.

Types of Skewness

1. **Positive Skew (Right Skew)** The tail on the `right side` is longer. Most data points are concentrated on the left. Mean > Median > Mode.Example: Income distribution, where most people earn moderate amounts but a few earn extremely high salaries.
2. **Negative Skew (Left Skew)**The tail on the `left side` is longer. Most data points are concentrated on the right. Mean < Median < Mode.Example: Age at retirement, where most people retire around 65 but some retire much earlier.
3. **Zero Skew** The distribution is perfectly symmetrical. Mean = Median = Mode.This is characteristic of a normal distribution.

Formula for Skewness

$$Skewness = \frac{3(Mean - Median)}{Standard\,Deviation}$$

Or using the third moment:

$$Skewness = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^3}{\sigma^3}$$

Interpreting Skewness Values

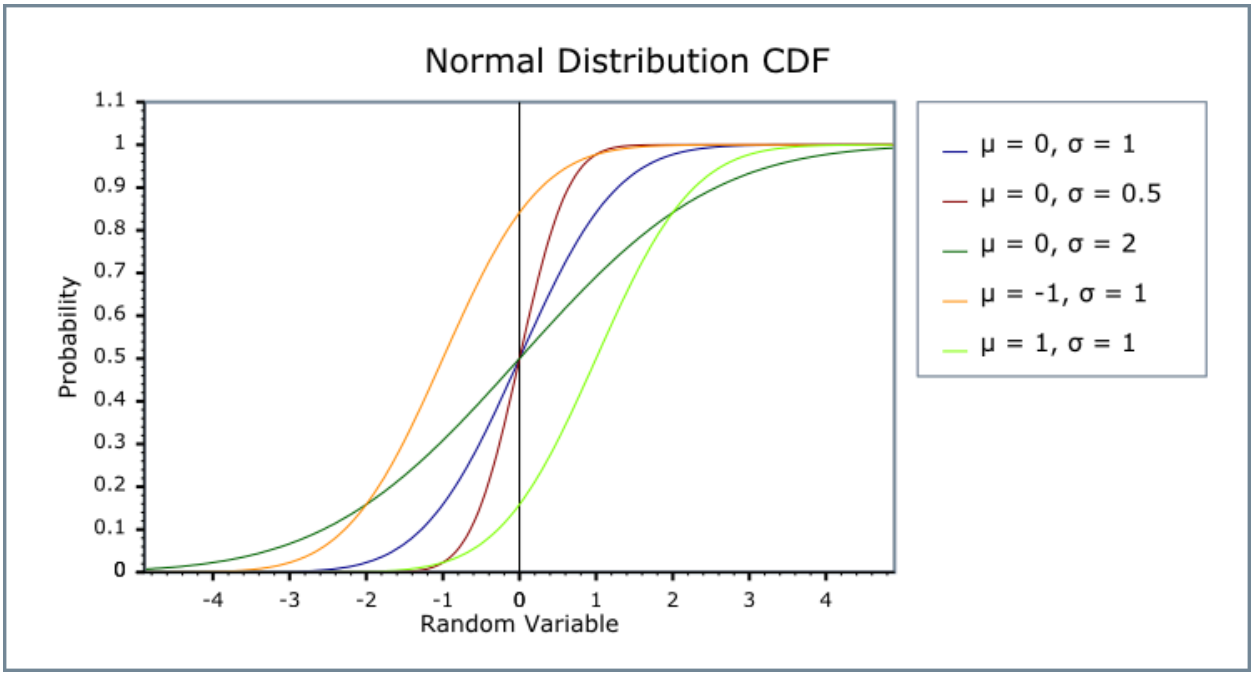
- Skewness > 0: Positive skew (right-tailed)
- Skewness = 0: Symmetric distribution
- Skewness < 0: Negative skew (left-tailed)

Why Skewness Matters in Data Science

- **Model Selection:** Many machine learning algorithms assume normally distributed data. High skewness may require transformation.
- **Feature Engineering:** Skewed features can be transformed using log, square root, or Box-Cox transformations to improve model performance.
- **Outlier Detection:** Skewness often indicates the presence of outliers that may need special handling.
- **Statistical Testing:** Many statistical tests assume normality; skewed data may violate these assumptions.

CDF of Normal Distribution Function

Probability up till X point



Use of Normal Distribution in Data Science

1. outlier detection
2. Assumptions on data for ML algorithms → Linear Regression and GMM
3. Hypothesis Testing
4. Central Limit Theorem

---

6. Kurtosis:

Kurtosis is a 4th Statistical Moment. In probability theory and statistics, kurtosis (meaning “curved”, “arching”) is a measure of the “**tailedness**” of the probability distribution of a real-valued random variable. Basically Kurtosis will help us understand the **fatness** of the tail.

Practical Use case of Kurtosis:

In finance, kurtosis risk refers to the risk associated with the possibility of extreme outcome or “fat tails” in the distribution of returns of a particular asset or portfolio.

If a distribution has high kurtosis, it means that there is a higher likelihood of extreme events occurring, either positive or negative, compared to a normal distribution.

In finance, kurtosis risk is important to consider because it indicates that there is a greater probability of large losses or gains occurring, which can have significant implications for investors. As a result, investors may want to adjust their investment strategies to account for kurtosis risk.

Excess Kurtosis

It's a measure of how much more peaked or flat a distribution is compared to a normal distribution, which is considered to have a kurtosis is 0. It's calculated by subtracting 3 from the sample kurtosis coefficient.

Types of Kurtosis

1. Leptokurtic (slender)

   A distribution with **positive excess kurtosis** is called leptokurtic. You subtract 3 in the **kurtosis formula** if the value is greater than 0, it's called Leptokurtic. In terms of shape, a leptokurtic distribution has fatter tails. That indicates that there are more extreme values or outliers in the distribution.

   Here the tail is **super fatter** than any other distribution.

2. Platykurtic (broad)

   A distribution with **negative excess** kurtosis is called platykurtic. You subtract 3 in the kurtosis formula if the value is lesser than 0. it's called PlatyKurtic. "Platy" means - 'broad'. In terms of shape, a platykurtic distribution has thinner tails. This indicates that there are fewer extreme values or outliers in the distribution.

   Assets with negative kurtosis are less risky and less volatile than those with a normal distribution, and they may experience more gradual price movements that are less likely to result in large gains or losses.

3. Mesokurtic

   Distribution with zero excess kurtosis are called mesokurtic. The most prominent example of mesokurtic distribution is the normal distribution family, regardless of the values of parameters.

   Mesokurtic is a term used to describe a distribution with a excess kurtosis of 0, indicating that it has the same degree of “peakedness” or “flatness” as a normal distribution.

   Example: In finance, a mesokurtic distribution is considered to be the ideal distribution for assets or portfolios, as it represents a balance between risk and return.

---

7. How to find the given distribution is normal or not?

You can find whether the distribution is gaussian or nor using three ways;

1. Visual inspection
2. QQ Plot (Quantile Quantile plot)
3. Statistical tests. (Shapiro-Wilk test, Anderson Darling test, Kolmogorov Smirnov test)

QQ Plot:

A QQ Plot is Quantile Quantile Plot is a graphical tool used to assess the similarity of the distribution of two sets of data. It is particularly useful for determining whether a set of data follows a normal distribution or not or given two distribution it finds both are comparable or not, you can use this to find uniform distribution, pareto distribution and more. .

You have `X, Y` distribution. `X` is your current data and `Y` is theoretical Normal distribution. You basically check $[X, Y]$ is comparable or not. If it’s comparable

then this is normal distribution.

Step 1

Sort the `X` and `Y` (theoretical normal distribution) data

Step 2

Calculate quantiles(percentiles) for both `X` and `Y`

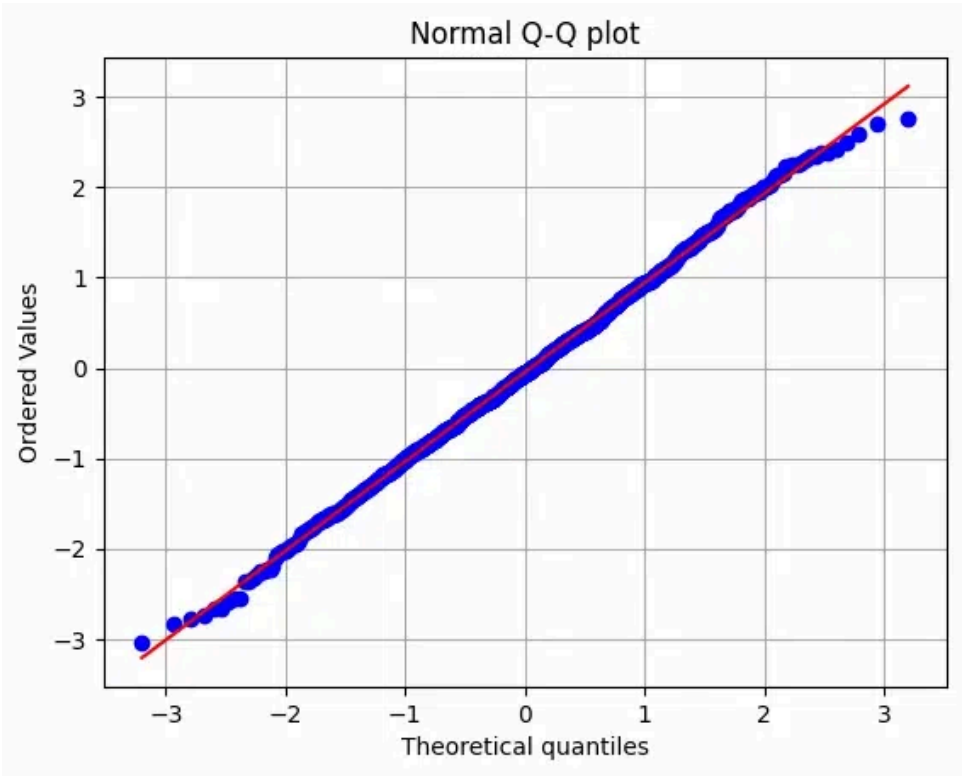
Step 3

Plot X_i, Y_i in the scatter plot

Step 4

Draw a 45° reference line $y = x$

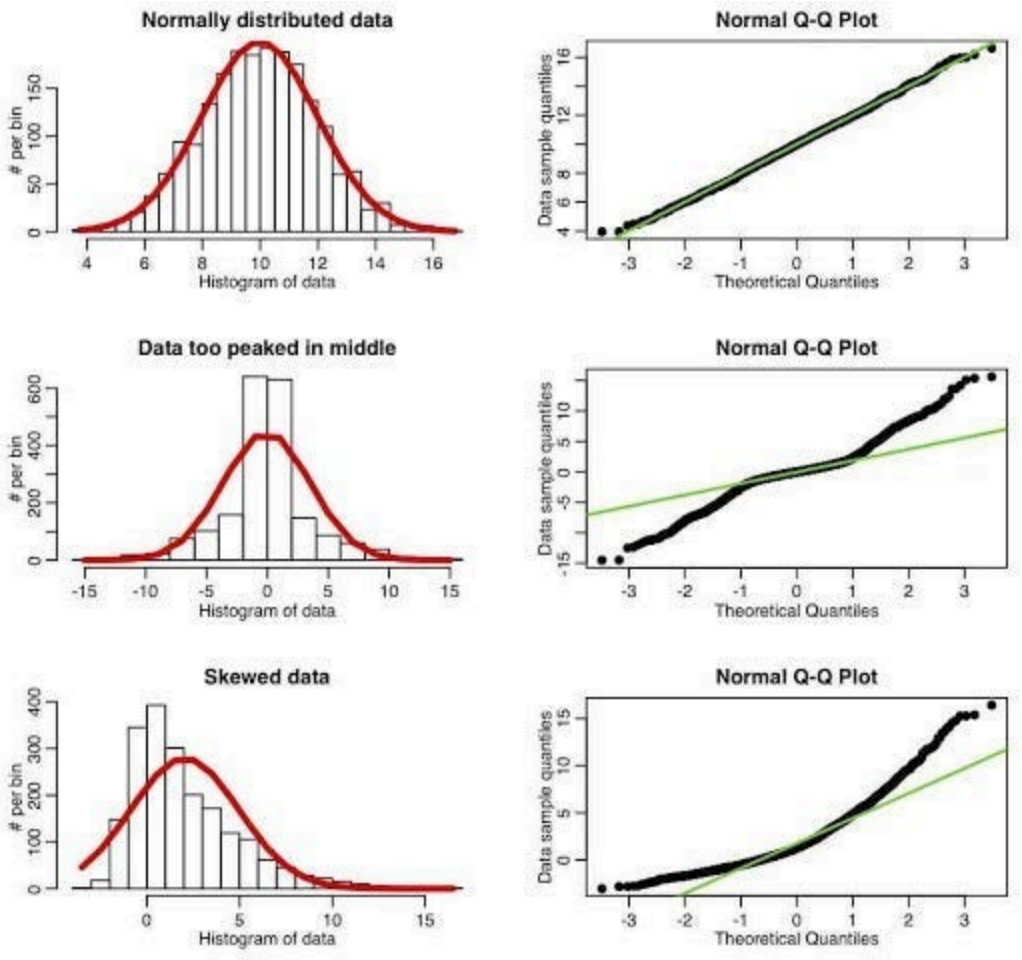
If your data is normally distributed, **the points will approximately lie on this line**. Deviations from the line indicate non-normality (e.g., skewness or heavy tails).



```
import numpy as np
import matplotlib.pyplot as plt

# Step 1: Generate sample data (try changing to np.random.exp
onential for non-normal)
data = np.random.normal(50, 10, 100)   # mean=50, std=10, n=1
00

# Step 2: Sort your sample data
data_sorted = np.sort(data)

# Step 3: Generate theoretical quantiles from a standard norm
al distribution
# Compute percentiles (from 1/(n+1) to n/(n+1))
n = len(data_sorted)
percentiles = (np.arange(1, n+1) - 0.5) / n

# Get corresponding theoretical quantiles from standard norma
l
theoretical_quantiles = np.quantile(np.random.normal(0, 1, 10
000), percentiles)

# Step 4: Plot the theoretical vs actual quantiles
plt.scatter(theoretical_quantiles, data_sorted, color='steelb
lue', label='Data points')
plt.plot(theoretical_quantiles, np.mean(data) + np.std(data)
* theoretical_quantiles, color='red', label='Reference line
(Normal)')
plt.xlabel('Theoretical Quantiles')
plt.ylabel('Sample Quantiles')
plt.title('Q–Q Plot using NumPy')
plt.legend()
```

```
plt.grid(True)
plt.show()
```

This is how you can interpret the QQ Plot.



Shapiro-Wilk test

This code will output tuple `(statistic, p-value)` . You only refer the p-value, if the $p-value < 0.5$ then this is **not randomly distributed**

If the $p-value > 0.5$ then is **normally distributed.**

```
from scipy.stats import shapiro

# Example data
data = [12.1, 12.5, 12.7, 12.4, 12.3, 12.6, 12.2]

# Perform Shapiro-Wilk test
stat, p = shapiro(data)

print('Statistic = %.3f, p-value = %.3f' % (stat, p))

# Interpret result
if p > 0.05:
    print("Sample looks Gaussian (normal distribution)")
else:
    print("Sample does not look Gaussian")
```

---

8. Non Gaussian Distribution

Basically it’s a non normal distribution and it has two types of data,

1. Continuous Non-gaussian distribution
2. Discrete Non gaussian distribution

Uniform Distribution

It’s a probability distribution where all outcomes are equally likely within a given range. This means that if you were to select a random value from this range, any value would be likely as any other value.

Types

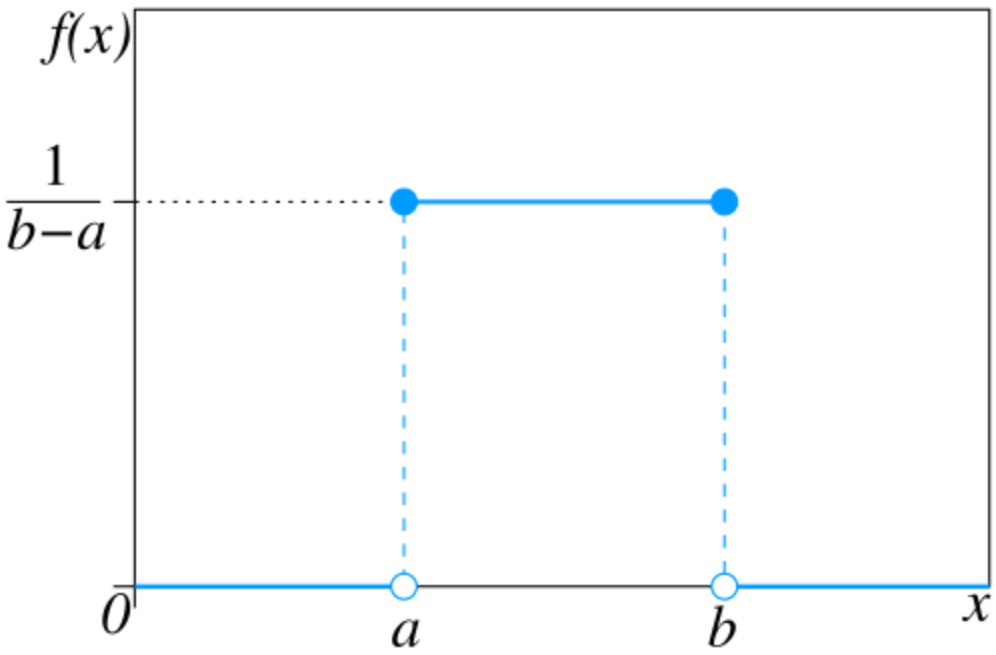
1. Discrete uniform distribution

2. Continuous uniform distribution (consider selecting from range of values)

Denoted as; $X \sim u(a, b)$ `a` and `b` are parameters.

Formula for Continuous

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise } x < a \text{ or } x > b \end{cases}$$



Skewness → 0

Examples of Uniform Distribution

1. Rolling a Fair Die (Discrete Uniform Distribution)

When you roll a fair six-sided die, each outcome (1, 2, 3, 4, 5, 6) has an equal probability of 1/6. This is a discrete uniform distribution where $X \sim \mathcal{U}(1, 6)$.

2. Selecting a Random Number (Continuous Uniform Distribution)

If you randomly select a number between 0 and 10, any value in that range is equally likely. For example, the probability of selecting a number between 2 and 3 is the same as selecting a number between 7 and 8. This is represented as $X \sim \mathcal{U}(0, 10)$.

3. Random Arrival Time

Suppose a bus arrives at a stop every 15 minutes, and you arrive at a random time. The waiting time is uniformly distributed between 0 and 15 minutes, where $X \sim \mathcal{U}(0, 15)$.

4. Random Number Generation in Programming

When using functions like `random.uniform(a, b)` in Python, the generated numbers follow a continuous uniform distribution between `a` and `b`.

5. Lottery Numbers (Discrete Uniform Distribution)

In a lottery where numbers from 1 to 100 are drawn, each number has an equal probability of 1/100 of being selected. This follos a discrete uniform distribution $X \sim \mathcal{U}(1, 100)$.

Application in machine learning

1. **Random initialisation**; In many machine learning algorithms, such as neural networks and k-means clustering, the initial values of the parameters can have significant impact on the final result. Uniform distribution is often used to randomly initialise the parameters, as it ensures that all values in the range have an equal probability of being selected.
2. **Sampling;** If you have a dataset with an equal number of samples from each class, you can use uniform distribution to randomly select a subset of data that is representative of all the classes.
3. **Data augmentation;** Uniform distribution can be used to generate new data points that are within a specified range of the original data. Specifically used in deep learning

4. **Hyper parameter tuning;** You need to search for the best combination of hyper parameters for a machine learning model. By defining a uniform prior distribution for each hyper parameter, you can sample from the distribution to explore the hyper parameter space.
5. **Generating random numbers;**

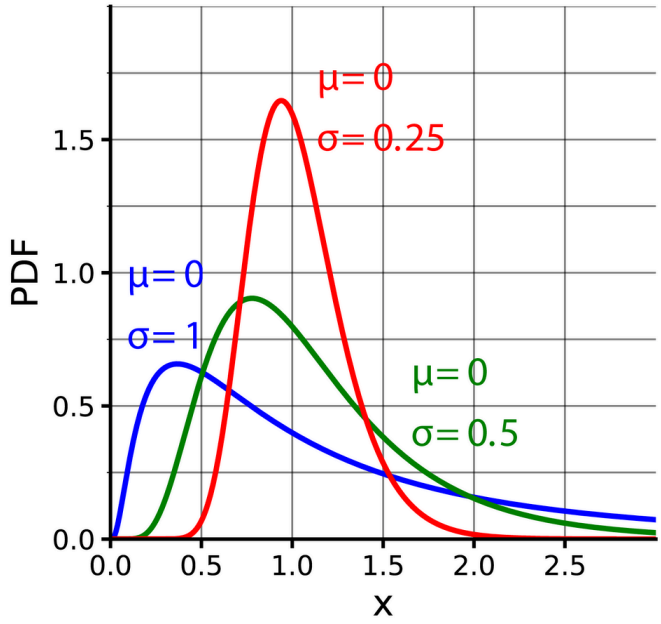
Log Normal Distribution

A log normal distribution is a heavily tailed continuous probability distribution of a random variable whose logarithm is normally distributed. You will see this distribution very commonly when you're building **internet applications.** (it's continuous distribution)

It's right skewed. If X is log normally distributed, then $log(x)$ is normally distributed.

$$X \sim \log \text{normal}$$
$$log(x) \sim N(\mu, \sigma)$$

Log normal distribution is denoted as $log(x) \sim N(\mu, \sigma)$



Steps to Calculate Log Normal Distribution

Step 1: Take the Natural Logarithm of Your Data

If you have data X that follows a log normal distribution, transform it by taking the natural logarithm: $Y = \ln(X)$. This transformed data Y should follow a normal distribution.

Step 2: Calculate Parameters of the Normal Distribution

Once you have $Y = \ln(X)$, calculate the mean (μ) and standard deviation (σ) of Y:

$$\mu = \frac{1}{n}\sum_{i=1}^{n}\ln(x_i)$$

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

Step 3: Define the Log Normal PDF

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

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \quad x > 0$$

Step 4: Calculate Probabilities or Generate Values

Use the PDF to calculate probabilities for specific values, or generate random samples from the log normal distribution using the parameters μ and σ.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import lognorm

# Step 1: Define parameters (mu and sigma of the underlying n
ormal distribution)
mu = 0      # mean of log(X)
sigma = 1   # standard deviation of log(X)

# Step 2: Generate log normal random samples
# In scipy, lognorm uses s=sigma and scale=exp(mu)
samples = lognorm.rvs(s=sigma, scale=np.exp(mu), size=1000)

# Step 3: Plot the distribution
plt.hist(samples, bins=50, density=True, alpha=0.6, color='or
ange', label='Log Normal Samples')

# Step 4: Plot the theoretical PDF
x = np.linspace(0.01, 10, 1000)
pdf = lognorm.pdf(x, s=sigma, scale=np.exp(mu))
plt.plot(x, pdf, 'r-', linewidth=2, label='Theoretical PDF')

plt.xlabel('X')
plt.ylabel('Probability Density')
plt.title('Log Normal Distribution')
plt.legend()
plt.grid(True)
plt.show()
```

Applications

- The length of the comments posted in Internet discussion forums follows a log normal distribution.
- The length of chess games tends to follow a log normal distribution.
- Users' dwell time on online articles (jokes, news etc.) follows a log normal distribution.
- In Income of 97%-99% of the population is distributed log normally.

How to check if a random variable is log normally distributed or not?

Method 1: Visual Inspection Using Histogram

Plot a histogram of your data. If it shows a right-skewed distribution with a long tail, it might be log normal.

```
import numpy as np
import matplotlib.pyplot as plt

# Sample data
data = np.random.lognormal(mean=0, sigma=1, size=1000)

# Plot histogram
plt.hist(data, bins=50, density=True, alpha=0.6, color='blu
e')
plt.xlabel('X')
plt.ylabel('Frequency')
plt.title('Histogram of Data')
plt.grid(True)
plt.show()
```

Method 2: Take Logarithm and Check for Normality

Transform your data by taking the natural logarithm. If the transformed data follows a normal distribution, then the original data is log normal.

```
# Take natural logarithm
log_data = np.log(data)

# Plot histogram of log-transformed data
plt.hist(log_data, bins=50, density=True, alpha=0.6, color='g
reen')
plt.xlabel('log(X)')
plt.ylabel('Frequency')
plt.title('Histogram of Log-Transformed Data')
plt.grid(True)
plt.show()
```

Method 3: Q-Q Plot

Create a Q-Q plot of the log-transformed data against a normal distribution. If the points fall approximately on a straight line, the original data is log normal.

```
from scipy import stats

# Q-Q plot for log-transformed data
stats.probplot(log_data, dist="norm", plot=plt)
plt.title('Q-Q Plot of Log-Transformed Data')
plt.grid(True)
plt.show()
```

Method 4: Shapiro-Wilk Test

Perform a Shapiro-Wilk test on the log-transformed data to test for normality. If p-value > 0.05, the log-transformed data is normally distributed, meaning the original data is log normal.

```
# Shapiro-Wilk test on log-transformed data
stat, p_value = stats.shapiro(log_data)

print(f"Shapiro-Wilk Test Statistics: {stat}")
print(f"P-value: {p_value}")
```

```
if p_value &gt; 0.05:
    print("The log-transformed data is normally distributed
(original data is log normal)")
else:
    print("The log-transformed data is NOT normally distribut
ed")
```

Method 5: Kolmogorov-Smirnov Test

Use the Kolmogorov-Smirnov test to compare your data against a log normal distribution.

```
from scipy.stats import lognorm

# Fit log normal distribution to data
shape, loc, scale = lognorm.fit(data)

# Perform K-S test
ks_stat, ks_p_value = stats.kstest(data, lambda x: lognorm.cd
f(x, shape, loc, scale))

print(f"K-S Test Statistics: {ks_stat}")
print(f"P-value: {ks_p_value}")

if ks_p_value &gt; 0.05:
    print("The data follows a log normal distribution")
else:
    print("The data does NOT follow a log normal distributio
n"
```

Pareto Distribution

It’s used to model the distribution of wealth, income, and other quantities that exhibit a similar power-law behaviour. (it’s continuous distribution)

Power law

In mathematics, a power law is a functional relationship between two variables, where one variable is proportional to a power of the other. Specifically, if y an x are two variables related by a power law, then the relationship can be written as;

Denoted as $y = K * x^a$; K is constant

Whenever something is following power law only 20% of the data has 80% of results or occupancy; `80-20` RULE.

Example 1: Occupancy Analytics in a Hotel

Imagine you're a **data scientist at a hotel chain** analyzing room occupancy rates.

You notice this pattern:

- 20% of your **loyal customers** (frequent travelers, business clients)

  → generate 80% of your **total room bookings** (occupancy).

Example 2: Office Space Occupancy

You're analyzing **sensor data** from office rooms to understand how efficiently space is used.

- You find that **20% of rooms** account for **80% of total occupancy time**.

This means:

- Most employees use only a few key spaces frequently.
- You can **optimize resource allocation** — maybe convert underused rooms into shared zones or collaboration areas.

Example3:

In computer science the Pareto principle can be applied to optimization efforts. [19] For example, Microsoft noted that by fixing the top 20% of the most-reported

bugs, 80% of the related errors and crashes in a given system would be eliminated.

Pareto distribution is based on an mathematical law called power law.

Observations

- The more the alpha value (parameter) the higher it’s in the y axis also the more the alpha value the thinner in the y axis.

- The less the alpha value (parameters) the lower it's in the y axis also the less the alpha value the broader in the y axis.
- If you make α value ∞ you get vertical line (line in black), it has no tail in it.

How to detect if a distribution is Pareto Distribution or not

1. Use log-log plot
2. QQ Plot

Bernoulli Distribution

This is a non gaussian distribution. It models **binary outcome** where the outcome can be either success (1), or failure (0). It has a single parameters `p` (probability of success).

Examples

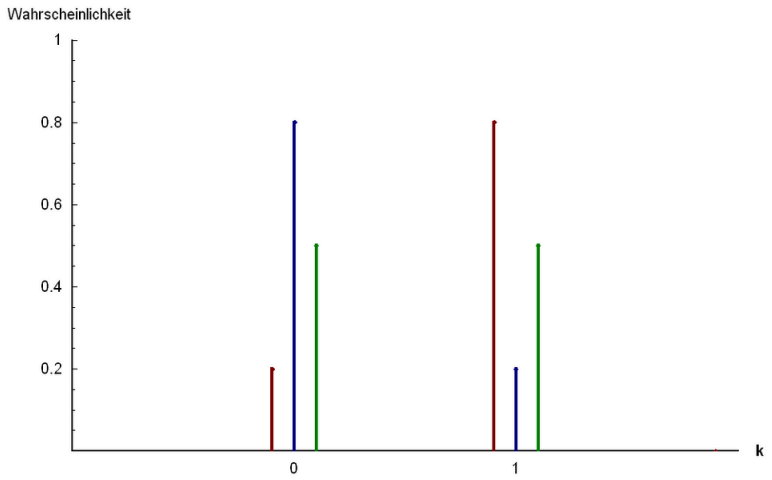
1. Coin toss (H or T)
2. Spam classifier (Spam or Ham)
3. Rolling a dice and getting a 5 (Getting 5 or No)

PMF

Denoted as $P(X = x) = P^x(1 - p)^{1-x}$

P ⇒ Probability of success

1-p ⇒ Probability of failure



It’s used in machine learning for modelling binary outcomes, such a whether customer will make purchase or not, whether an email is spam or not, or whether a patient will have a certain disease or not. One more thing, it will be used in Binomial Distribution.

Binomial Distribution

Binomial distribution is a probability distribution that describes the number of successes in a fixed number of independent Bernoulli trails with two possible outcomes, where the probability of success is constant for each trail.

If n=1 ⇒ Bernoulli

If n>1 ⇒ Binomial

Let’s say you’re reading a single email and determining whether it’s **spam or not spam** — that’s a **Bernoulli trial**, since there are only two possible outcomes.

Now, if you’re reading **5 emails at once** and counting **how many of them are spam**, that becomes a **Binomial distribution**, because it represents the **number of successes (spam emails)** across **multiple independent Bernoulli trials**.

Quick Intuition

- **Bernoulli** → one trial (spam / not spam)

- **Binomial** → multiple trials (how many spams out of 5 emails)

Examples:

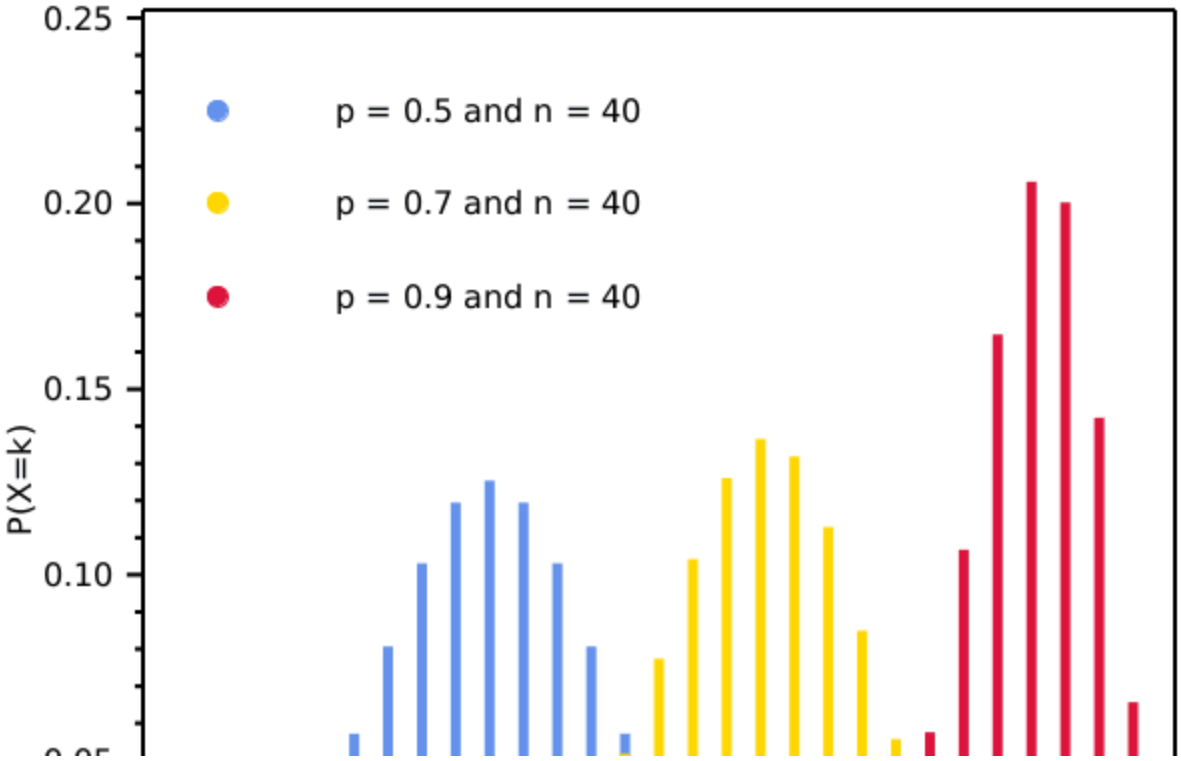
1. No-one out of 3 people will like it?
2. 1 out of 3 people will like it ? (what's the probability)
3. 2 out of 3 people will like it ?
4. 3 out of 3 will like it?

PDF:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n$$

Here:

- n → number of trials
- k → number of successes
- p → probability of success
- $\binom{n}{k}$ → number of ways k successes can occur among n trials



Criteria

1. The process should consist of n trails (n>1)
2. only 2 exclusive outcomes are possible, a success and a failure.
3. P(success) = p and P(failure) = 1 - p and it’s fixed from trail to trail
4. The trail are independent

Applications

1. **Binary Classification problems**; we may model the probability of an email being spam or not spam using a binomial distribution
2. **Hypothesis testing;** We use binomial distribution to calculate the probability of observing a certain number of success in a given number of trails, assuming a null hypothesis is true.

3. **Logistic Regression**; It models the probability of an event happening as a logistic function of the input variables. Since the logistic function can be viewed as a transformation of a linear combination of inputs, the output of logistic regression can be thought of as a binomial distribution.
4. A/B testing

---

9. Transformation

Transformation is nothing but mathematical transformation which takes non-gaussian distribution and converts it to gaussian distribution.

Log Transform

You Basically take a log of your entire data, the output will approximate the gaussian distribution.

When to Use?

- If you don't want any negative values, because of log of Negative is undefined.
- If you have a **right skewed data**, you can use this.
- If you have a bigger values in your data it will scale it up, so you can use this, the output will be linearly separated values, because you're taking a log of big values.

Reciprocal transform ($\frac{1}{x}$)

Small values becomes higher and higher values becomes lower. It's a general transform, you can use it for **all type of data**.

Square transform (x^2)

You use this when you have **left skewed** data.

Box Cox Transform

Given any distribution it converts into the normal distribution. It's a General Transform method. You have one parameter called λ.

$$x_i^\lambda = \begin{cases} \frac{x_i^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \ln(x_i), & \text{if } \lambda = 0, \end{cases}$$

λ is the power parameter basically here you try to calculate power for the proper normal distribution. It varies `-5` to `5` in the process of searching, we examine all values of λ. Finally we choose the optimal value (resulting in the best approximation to a normal distribution) for your variable.

This is applicable only the `number > 0`, if you have value is less than 0, you have to use `Yeo-Johnson Transform`.

Yeo Johnson Transform

It's a adjustment to the Box-cox transformation, by which we can apply it to negative numbers.

---

10. Resources

1. More Continuous distribution pdf: click here
2. More about Binomial distributions pdf: click here
3.

Maximum Likelihood Estimation

Why are we studying MLE?

- It's a general topic used in many areas, some times you used in deep learning, linear regression.
- This topic is really important it will help us to understand many ML algorithms.
- And also this is the important favourite interview topic.

Difference between Probability and Likelihood:

Probability Definition:

> We know that, the probability of getting a *head* or *tail* from a fair coin (random experiment) is 0.5. You can also prove this by mathematically, this probability

distribution is basically a Bernoulli distribution because the outputs are either 1 or 0.

Here you have a random experiment (tossing a fair coin) and also you know the distribution (Bernoulli), If you know the experiment, distribution and parameter then you can find the chance for particular event (getting a tail). To **finding** a ***chance*** for a **particular event** is called "**Probability**"

Formal definition: This is a measure of the chance that a certain event will occur out
of all possible events. It's usually presented as a ratio or fraction, and it ranges from 0 (meaning the event will not happen) to 1 (meaning the event is certain to happen).

A probability quantifies how often you observe a certain outcome of a test, given a certain understanding of the underlying data.

Likelihood Definition:

Now you have a fair coin, you have tossed 5 times, the outcomes are 5 heads `{ H, H, H, H, H }`. Here you get a simple question in your mind that; How come all the outcomes are **5** heads, does it really a fair coin? If it's really a fair coin, what is the plausibility (chances) that **P (**parameter for the Bernoulli distribution) parameter is 0.5 (or; coin is fair) ?

Again simply how come the coin is fair if we are getting all the outcomes as 5 heads. *The doubt is called likelihood*.

Let's say you calculated likelihood for the outcomes (5 heads) = 0.03125. It means that given that 5H outcomes for the 5 toss (random experiment), the likelihood that coin is **fair** is **very low**.

Here you have a event (you tossed 5 times and you got 5 Heads as output) and the parameter is P = 0.5 (Bernoulli distribution).

- **Probability** → when the parameter is known and outcomes are random.
- **Likelihood** → when the outcome is observed and the parameter is questioned.

Likelihood does **not** represent the probability that the coin is fair. Instead, it measures how plausible a particular parameter value is in light of the observed outcome. In this case, observing 5 consecutive heads is possible under fairness, but not very plausible.

Formal definition: In statistical context, likelihood is a function that measures the plausibility of a particular parameter value given some observed data. It quantifies how well a specific outcome supports specific parameter values.

A likelihood quantifies how good one's model is, given a set of data that's been observed.

Maximum Likelihood Estimation:

You need to find the parameter value for the random experiment which has the more likelihood, that is called maximum likelihood. To estimate the maximum likelihood we use maximum likelihood estimation.

> **MLE is the method of estimating the parameters of the statistical model given some observed data.**

Most Plausible means Among all parameter values, this one makes the observed data the least surprising (or most consistent). "Most plausible" just means **this parameter fits the observed data best compared to others**.

For many models, these two perspectives are equivalent - minimizing the loss function is the same as maximizing the likelihood function. In fact, many common loss functions can be derived from the principle of MLE under certain assumptions about the data.

MLE for Discrete Random Variable

// need to add the description for this

MLE for Normal Distribution

Let’s assume the data follows a normal distribution. Suppose you have already calculated the probability that a person’s height lies between 160 cm and 170 cm using the normal distribution formula.

Now, imagine you observe a new data point: a random variable with value 110 cm. The question becomes: **for this observed value (110 cm), which parameters of the normal distribution—mean (μ) and standard deviation (σ)—make this observation most likely?**

Finding the values of μ and σ that maximize the likelihood of observing 110 cm is exactly what **Maximum Likelihood Estimation (MLE)** does.

// need to add the image of this gaussian distribution for this problem statement

One of the rules for log is

log(a.b) = Log(a) + log(b)

So this will be "*log likelihood".*

MLE in ML

Usual steps in MLE:

Normally given some observed data first you assume some distribution for the observed data then you use likelihood function (mostly PMF or PDF formula) and then you find the values for the parameters that maximises the likelihood function using differentiation.

1. Get the observed data
2. Assume the distribution for the observed data
3. Calculate Likelihood function (mostly PDF or PMF formula based on the observed distribution)
4. Find the parameters which maximises the likelihood function. The parameters for the particular distribution.

How this will use in the Machine Learning?

→ First step you find out the distribution of the Y (dependent variable) | given X (independent variable) in the given dataset.

→ Based on distribution you decide to apply a ML model which is **parametric** (parametric models) in **nature**.

→ You randomly decide some values for the parameters.

→ Select a Likelihood function.

→ Find out the values of parameters which maximises the likelihood function.

Parametric model means, the model assumes the data is coming from some specific distribution like Linear Regression, Neural Network and Logistic Regression, Decision Trees are Non Parametric model.

MLE in Logistic Regression

Training a model using MLE in logistic regression.

Probability

1. Basic Terms

Random Experiment

An Experiment is called random experiment if it's satisfies two conditions

1. It has more than one possible outcome
2. It's not possible to predict the outcome in advance.

e.g, **tossing a coin** it's a Random experiment because it has more than one possible outcome (head or tail) also you can't predict the outcome in advance.

Trail

Trail refers to a single execution of a random experiment. For every trail you will get outcome.

Outcome

Outcome refers to single possible result of a trail

Sample Space

Sample space of a random experiment is the set of all possible outcomes that can occur from an experiment. Generally one random experiment will have one set of sample space.

e.g., **tossing a coin** is an experiment, the set of possible outcomes are $S = \{Head, Tail\}$

Event

Event is a **specific set of outcomes** from a random experiment or process. Essentially, it's a
subset of the **sample space**. An event can include a single outcome, or it can include multiple outcomes. One random experiments can have multiple events.

e.g., **tossing a coin** is an experiment, what's the probability of getting a head. This "*what's the probability of getting a coin*" is one event. This is a subset of sample space.

e.g., **rolling a die** is an experiment, now you want to measure the probability of getting a number less than 3. Here even is "*getting a number less than 3*". This is a subset of sample space.

Example of Random Experiment

Context: Consider the Titanic dataset with **891 passengers** categorized by **class**.

Random Experiment (RE): Randomly drawing out a passenger and finding their class.

Trial: One instance of drawing a passenger.

Outcome: The class of the selected passenger (e.g., 1, 2, or 3).

Sample Space (S): {1, 2, 3}

Events: Event A: The passenger is from class 1 A = {1}

Event B: The passenger is not from class 2 B = {1, 3}

Types of Events

1. **Simple Event:** Also known as an elementary event, a simple event is an event that consists
   of exactly one outcome. For example, when rolling a fair six-sided die, getting a 3 is a simple event.
2. **Compound Event:** A compound event consists of two or more simple events. For example, when rolling a die, the event "rolling an odd number" is a compound event
   because it consists of three simple events: rolling a 1, rolling a 3, or rolling a 5.
3. **Independent Events:** Two events are independent if the occurrence of one event does not
   affect the probability of the occurrence of the other event.
   For example, if you flip a coin and roll a die, the outcome of the coin flip does not affect the
   outcome of the die roll.
4. **Dependent Events:** Events are dependent if the occurrence of one event **does affect** the
   probability of the occurrence of the other event.
   For example, if you draw two cards from a deck without replacement, the outcome of the
   first draw affects the outcome of the second draw because there are fewer cards left in the
   deck.
5. **Mutually Exclusive Events**: Two events are mutually exclusive (or disjoint) if they cannot
   both occur at the same time.
   For example, when rolling a die, the events "roll a 2" and "roll a 4" are mutually exclusive
   because a single roll of the die cannot result in both a 2 and a 4.
6. **Exhaustive Events:** A set of events is exhaustive if at least one of the events must occur
   when the experiment is performed.
   For example, when rolling a die, the events "roll an even number" and "roll

7. **Impossible Events:** You know that the specific outcome won't occur. Example: getting a seven, when rolling an one die. It's impossible right ?
8. **Certain event or sure event:** A certain event in probability is an event that is guaranteed to happen, with a probability of 1 or 100%. It is also called a sure event and represents the entire sample space of possible outcomes.

   Example; Rolling a die**:** The event of rolling a number between 1 and 6 is a certain event because every possible outcome falls within this range.

Probability

In simplest terms, probability is a **measure** of the **likelihood** that a **particular event** will **occur**. It
is a fundamental concept in statistics and is used to make predictions and informed decisions
in a wide range of disciplines, including science, engineering, medicine, economics, and social
sciences.

Probability is usually expressed as a number between 0 and 1, inclusive:

- A probability of 0 means that an event will not happen.
- A probability of 1 means that an event will certainly happen.
- A probability of 0.5 means that an event will happen half the time (or that it is as likely to
  happen as not to happen).

There are two types of probability

1. Empirical Probability
2. Theoretical Probability

Empirical Probability

Empirical means experimental. Empirical probability, also known as experimental probability, is a probability measure that is based on observed data, rather than theoretical assumptions. It's calculated as the ratio of the number of times a particular event occurs to the total number of trials.

e.g., Suppose that, in our 100 tosses, we get heads 55 times and tails 45 times. What is the empirical probability of getting a head ? **Answer**: $P(\frac{55}{100})$ that's the answer.

e.g., Let's say you have a bag with 50 marbles. Out of these 50 marbles, 20 are red, 15 are blue, and 15 are green. You start to draw marbles one at a time, replacing the marble back into the bag after each draw. After 200 draws, you find that you've drawn a red marble 80 times, a blue marble 70 times, and a green marble 50 times. What is the empirical probability of getting a red marble? **Answer**: $P(\frac{80}{200})$ that's the answer.

Theoretical Probability

Theoretical (or classical) probability is used when each outcome in a sample space is equally
likely to occur. If we denote an event of interest as Event A, we calculate the theoretical
probability of that event as:

Theoretical Probability of Event A = Number of Favourable Outcomes (that is, outcomes in
Event A) / Total Number of Outcomes in the Sample Space

Example of Theoretical Probability

Context: A fair six-sided die is rolled once.

Sample Space (SS): {1, 2, 3, 4, 5, 6}

Event A: Rolling an even number A = {2, 4, 6}

Number of Favourable Outcomes: 3 (since 2, 4, and 6 are even)

Total Number of Outcomes: 6

Theoretical Probability of Event A: Number of Favourable Outcomes / Total Number of Outcomes = 3 / 6 = 0.5

Fact: If you do Empirical probability nearly infinity times, the probability of empirical probability will approximate theoretical probability (means both the probability will be very similar)

Random Variable

Random variable is **not** a **variable**, it's a **function**. Function is basically takes some input and based on some logic and it outputs something.

In the context of probability theory, a random variable is a **function** that maps the outcomes of a random process (known as the sample space) to a set of real numbers.

Input: The input to the function is an outcome from the **sample space** of a r**andom process**.

Output: The output of the function is a real number that we assign to each possible outcome.

Example 1: Rolling a Die

Random Process: Roll a fair six-sided die once

Sample Space: {1, 2, 3, 4, 5, 6}

Random Variable X: X maps each outcome to its square

Logic: Squaring (it's an example, it might be anything)

Mapping:

1 → 1

2 → 4

3 → 9

4 → 16

5 → 25

6 → 36

Example 2: Tossing Two Coins

Random Process: Toss two fair coins

Sample Space: {HH, HT, TH, TT}

Random Variable Y: Y maps each outcome to the number of heads

Logic: Head count (this logic is example, it might be anything)

Mapping:

HH → 2

HT → 1

TH → 1

TT → 0

Random variable is a function, the input is **sample space;** based on some logic you will get real number outputs. The logic is defined by **events**. The creation of logic depends up on the events and it varies a lot.

Logic Creation example:

Random Process: Rolling a two dice once

Event: Get total 7 as outcome

Sample Space: { (1,1), (1,2), (1,3), ... (6,6) } ⟶ total 36

Logic: sum each element in sample space { (2, 3, 4, 5, ...12) }

Mapping:

(1,1) → 2

(5,6) → 11

Basically a idea is Random variable is a way to map an event by converting an sample space to some numerical output.

Two types of Random Variable

1. Continuous Random Variable
2. Discrete Random Variable

---

2. Probability Distributions

A probability distribution is a list of all of the possible outcomes of a random variable along with their corresponding probability values.

Refer statistics notes for probability distribution.

---

3. Random Variable Statistics

Mean of a Random Variable

The mean of a random variable, often called the **expected value,** is essentially the average
outcome of a random process that is repeated many times. More technically, it's a weighted
average of the possible outcomes of the random variable, where each outcome is weighted by
its probability of occurrence.

Example: Expected Value of a Random Variable

Random Process: Roll a fair six-sided die once

Sample Space: $\{1, 2, 3, 4, 5, 6\}$

Random Variable X: X = outcome of the die roll

Probabilities: $P(X = k) = \frac{1}{6}$ for each $k \in \{1, 2, 3, 4, 5, 6\}$

Expected Value: $E[X] = \sum_{k=1}^{6} k \cdot P(X = k)$

$$E[X] = (1)(\frac{1}{6}) + (2)(\frac{1}{6}) + (3)(\frac{1}{6}) + (4)(\frac{1}{6}) + (5)(\frac{1}{6}) + (6)(\frac{1}{6})$$

$$E[X] = \frac{1 + 2 + 3 + 4 + 5 + 6}{6}$$

$$E[X] = \frac{21}{6}$$

$$E[X] = 3.5$$

Variance of a Random Variable

The variance of a random variable is a statistical measurement that describes how much
individual observations in a group differ from the mean (expected value).

Example: Variance of a Random Variable

Random Process: Roll a fair six-sided die once

Sample Space: $\{1, 2, 3, 4, 5, 6\}$

Random Variable X: X = outcome of the die roll

Probabilities: $P(X = k) = \frac{1}{6}$ for each $k \in \{1, 2, 3, 4, 5, 6\}$

Expected Value: $E[X] = 3.5$

Variance Formula: $Var(X) = E[(X - E[X])^2] = \sum_{k=1}^{6} (k - 3.5)^2 \cdot P(X = k)$

$Var(X) = \frac{(1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2}{6}$

$Var(X) = \frac{(6.25) + (2.25) + (0.25) + (0.25) + (2.25) + (6.25)}{6}$

$Var(X) = \frac{17.5}{6}$

$Var(X) \approx 2.92$

Example is for population (because we are not dividing by n-1).

---

4. Prerequisites for types of probability

Venn Diagram

A **Venn diagram** is a visual representation of **events** and their **relationships** within a **sample space**.

It helps illustrate concepts like **union**, **intersection**, and **complement** of events.

- The **rectangle** represents the **sample space (S)** — all possible outcomes.
- **Circles** inside the rectangle represent **events** (like A, B, etc.).

(a) Union of Events

Represents the event that **A or B or both** occur.

Symbolically:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

In the Venn diagram, the **shaded area** includes both circles A and B.

(b) Intersection of Events

Represents the event that **both A and B** occur together.

Symbolically:

$$P(A \cap B)$$

In the Venn diagram, this is the **overlapping region** of A and B.

(c) Complement of an Event

Represents the event that **A does not occur**.

Symbolically:

A' or A^c

Formula:

$$P(A') = 1 - P(A)$$

In the Venn diagram, this is the **area outside** circle A but still inside the rectangle (sample space).

(d) Mutually Exclusive Events

Two events that **cannot occur together**.

Symbolically:

$$P(A \cap B) = 0$$

In the Venn diagram, circles A and B **do not overlap**.

(e) Exhaustive Events

A set of events that **cover the entire sample space**.

Symbolically:

$$P(A \cup B \cup C ...) = 1$$

In the Venn diagram, all regions within the rectangle are covered by the circles.

Formulas

(i) For Two Events

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

(ii) For Three Events

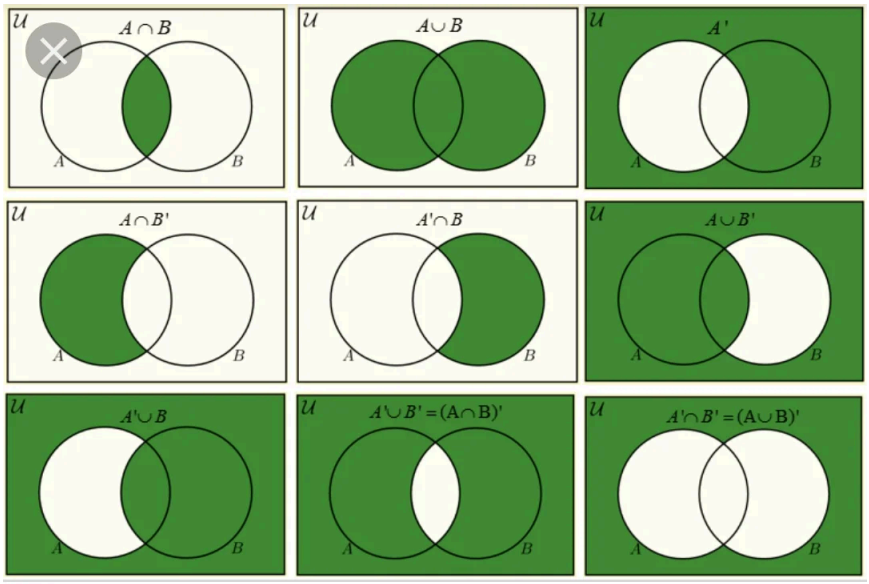
$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

$$-P(A \cap B) - P(A \cap C) - P(B \cap C)$$

- P(A ∩ B ∩ C)

Visual Intuition

- **A only:** Area inside A but outside B
- **B only:** Area inside B but outside A
- **A ∩ B:** Overlapping region of A and B
- **A' ∩ B:** B occurs but A does not
- **A ∩ B':** A occurs but B does not
- **A' ∩ B':** Neither A nor B occur (outside both circles)
- A U B: Both A and B occurs (inside both circles)



Contingency Table

Whatever information you show in the Venn diagram, you can show that in Contingency table.

A **contingency table** (also called a **cross-tabulation** or **two-way table**) is a tabular method used to display the **joint frequencies** (or probabilities) of two categorical variables. It helps us analyze **relationships between two events** — for example, between *gender* and *smoking habit*, or *test result* and *disease presence*.

	Event B occurs	Event B' (does not occur)	Total
Event A occurs	n(A ∩ B)	n(A ∩ B')	n(A)
Event A' (does not occur)	n(A' ∩ B)	n(A' ∩ B')	n(A')
Total	n(B)	n(B')	n(S)

Each cell value can be converted to probability by dividing by total outcomes:

$$P(A \cap B) = n(A \cap B)/n(S)$$
$$P(A \cap B') = n(A \cap B')/n(S)$$
$$P(A') = n(A')/n(S)$$

Similarly you can convert contingency table to Venn diagram.

---

5. Types of Probability

Joint Probability

We have 2 Random Variables X and Y. The joint probability of X and Y, denoted as $P(X = x, Y = y)$, is the Probability that X takes the value of x and Y takes the value of y (small y) at the same time.

This explanation for machine learning, in books you see joint probability is the two events occurs at a same time.

	B occurs	B' (does not occur)	Total
A occurs	n(A ∩ B)	n(A ∩ B')	n(A)
A' does not occur	n(A' ∩ B)	n(A' ∩ B')	n(A')
Total	n(B)	n(B')	n(S)

From this table:

P(A ∩ B) = n(A ∩ B) / n(S)

P(A ∩ B') = n(A ∩ B') / n(S)

P(A' ∩ B) = n(A' ∩ B) / n(S)

P(A' ∩ B') = n(A' ∩ B') / n(S)

Left side axis (all rows) is Y and right side axis (all columns) X

Example

Pclass	Survived = 0	Survived = 1	Total
1	80	136	216
2	97	87	184
3	372	119	491
Total	**549**	**342**	**891**

$$P(X = 1, Y = 0) = n(X = 1 \cap Y = 0)/n(S)$$

n(X=1 ∩ Y=0) = 80

n(S) = 891

P(X=1, Y=0) = 80 / 891 ≈ 0.0898 = 8.98 % chance

```
# Joint probability calculation
import seaborn as sns

titanic = sns.load_dataset('titanic')
joint_prob = len(titanic[(titanic['pclass'] == 1) & (titanic
['survived'] == 0)]) / len(titanic)

# another approach
pd.crosstab( df['survived'], df['pclass'], normalize = 'all')
```

Joint Probability Distribution: Get all joint probability outcomes and plot it in graph. Getting all joint probability for all outcomes is called joint probability distribution.

Marginal Probability

This is also called **Simple Probability / Unconditional Probability**

Marginal Probability refers to probability of an event occurring irrespective of outcome of some other event. In a contingency table, these probabilities are found in the **margins** (totals) of the table — hence the name *marginal*.

It helps answer questions like:

- What is the probability that a passenger survived (irrespective of class)?
- What is the probability that a passenger was in 1st class (irrespective of survival)?

Pclass	Survived = 0	Survived = 1	Total
1	80	136	216
2	97	87	184
3	372	119	491
Total	**549**	**342**	**891**

Here,

- Total passengers (n(S)) = 891
- Pclass = 1, 2, or 3
- Survived = 0 (did not survive), 1 (survived)

(a) Marginal Probability of Pclass = 1

P(P class = 1) = n(P class = 1) / n(S)

n(P class = 1) = 216

n(S) = 891

P(P class = 1) = 216 / 891 ≈ 0.2425

Interpretation:

About **24.25%** of passengers were in **1st class**.

You can also divide margins with total numbers for proper visualisation

Pclass	Survived = 0	Survived = 1	Total
1	80	136	216 / 891
2	97	87	184 / 891
3	372	119	491 / 891
Total	**549 / 891**	**342 / 891**	**891**

- Marginal probability looks at **a single variable**, not combinations.

- It is obtained by **summing joint probabilities** or using **totals** from a contingency table.
- The sum of all marginal probabilities across a variable = 1.

The sum of X axis always gonna be 1 ((216/891) + (184/891) + (491+891)) == 1

Similarly, The sum of Y axis always gonna be 1 ((549 / 891) + (342/ 891)) == 1

Marginal probability distribution of Y == ((549 / 891) + (342/ 891))

Marginal Probability distribution of X = ((216/891) + (184/891) + (491+891))

```
pd.crosstab( df['survived'], df['pclass'], normalize = 'all',
margins=True)

# normalize = True ( you get joint probability )
# margins = True ( you get marginal probability )
# you get both
```

Conditional Probability

Conditional Probability is a measure of the probability of an event occurring, given that another event has already occurred. If the event of interest is A and the event B has already occurred, the conditional probability of A given B is usually written is P(A|B).

P(A|B) = $\frac{\text{Joint Probability of A and B}}{\text{Marginal Probability of B}} = \frac{P(A \cap B)}{P(B)}$

Pclass	Survived = 0	Survived = 1	Total
1	80	136	216

Pclass	Survived = 0	Survived = 1	Total
2	97	87	184
3	372	119	491
Total	**549**	**342**	**891**

n(S) = 891

n(Survived = 1) = 342

n(Pclass = 1) = 216

Goal: probability a passenger survived given they were in 1st class.

Step A — numerator: P(Pclass = 1 ∩ Survived = 1) = n(Pclass = 1 ∩ Survived = 1) / n(S) = 136 / 891

Step B — denominator: P(Pclass = 1) = n(Pclass = 1) / n(S) = 216 / 891

Step C — apply formula:

$$P(Survived = 1 | Pclass = 1) = \frac{P(Pclass=1 \cap Survived=1)}{P(Pclass=1)} = \frac{136/891}{216/891} = \frac{136}{216}$$

Answer: P(Survived = 1 | Pclass = 1) = 136/216 = 17/27 ≈ 0.6296

Interpretation: Given a passenger is 1st class, there is about a 62.96% chance they survived.

```
# conditional probability
pd.crosstab( df['survived'], df['pclass'], normalize = 'colum
ns')

# conditional probabilty reverse (conditional probability bas
ed on the rows)
```

```
pd.crosstab( df['survived'], df['pclass'], normalize = 'inde
x')
```

Intuition behind the conditional probability formula

The Intuition behind the conditional probability formula is based on the **concept** of **reducing** our **sample space**.

The denominator, P(B), is the probability of event B occurring. When we say we want to know the probability of A given B, we're effectively saying that B has occurred and therefore B is our new "universe" or sample space. So we're not considering cases when B didn't occur anymore, and we're normalizing by the probability of B.

The numerator, P(A ∩ B), is the joint probability of A and B, meaning both A and B occur. So in the context of our new universe where B has occurred, P(A ∩ B) represents the cases where A also occurs.

By dividing the joint probability P(A ∩ B) by the probability of B (P(B)), we effectively find the proportion of times that A occurs given that B has occurred.

In summary, the conditional probability of A given B is just the joint probability of A and B happening (A and B together in the "universe" where everything happens), normalized by the probability of B (the new "universe" where only B happens).

---

6. Bayes Theorem

Bayes' theorem is a fundamental concept in the field of probability and statistics that describes how to update the probabilities of hypotheses when given evidence. It's used in a wide variety of fields, including machine learning, statistics, and game theory.

Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new evidence.

$$P(A \mid B) = \frac{P(A) \times P(B|A)}{P(B)}$$

P(B|A) → likelihood

P(B) → Marginal

P(A) → Prior

P(A|B) → Posterior