Statistics

What is Statistics?

Statistics is a set of mathematical methods and tools that enable us to answer important questions about data or Statistics is the science of collecting, organising and analysing the data.

It is divided into two categories:

Descriptive Statistics		Inferential Statistics	
Measures of Central Tendency	Measures of Dispersion	Hypothesis Testing	Regression Analysis
Mean Median Mode	Range Standard Deviation Variance Absolute Deviation	Z test F test T test	Linear Regression

1. **Descriptive Statistics** - this Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

   Univariate Analysis

   Univariate analysis involves the examination across cases of one variable at a time. There are three major characteristics of a single variable that we tend to look at:

   - the distribution
   - the central tendency
   - the dispersion

2. **Inferential Statistics** – It consists of collecting sample data and making conclusion about population data using some experiments.

   i. **Simple Random Sampling**
   A sample of size n from a population of size N is obtained through simple random sampling if every possible sample of size n has an equally likely chance of occurring.

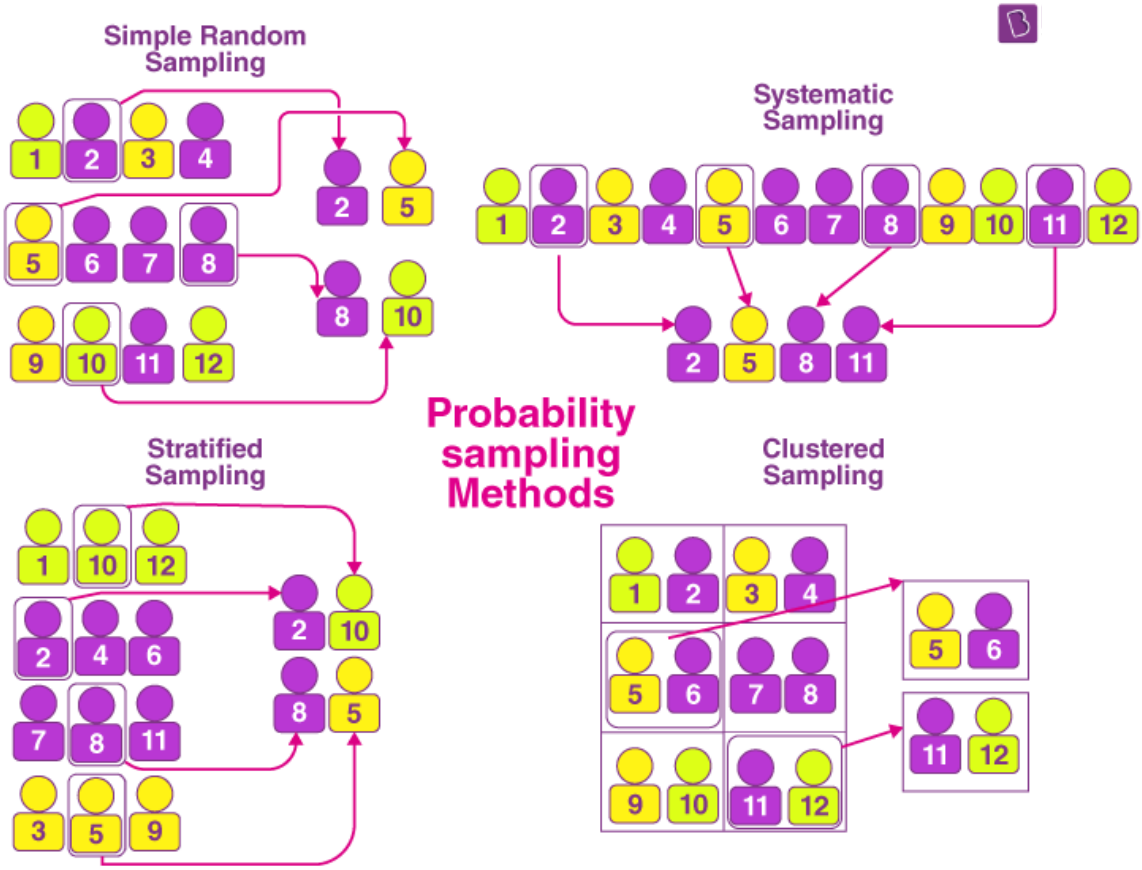
   ii. **Stratified Sampling**
   The items are selected from the target population by selecting the random selection point and selecting the other methods after a fixed sample interval. It is calculated by dividing the total population size by the desired population size.

iii. **Systematic Sampling**

The total population is divided into smaller groups to complete the sampling process. The small group is formed based on a few characteristics in the population. After separating the population into a smaller group, the statisticians randomly select the sample.

iv. **Cluster Sampling**

The cluster or group of people are formed from the population set. The group has similar significatory characteristics. Also, they have an equal chance of being a part of the sample. This method uses simple random sampling for the cluster of population.



Probability sampling vs Non-probability Sampling Methods

Probability Sampling Methods	**Non-probability Sampling Methods**
Probability Sampling is a sampling technique in which samples taken from a larger population are chosen based on probability theory.	Non-probability sampling method is a technique in which the researcher chooses samples based on subjective judgment, preferably random selection.

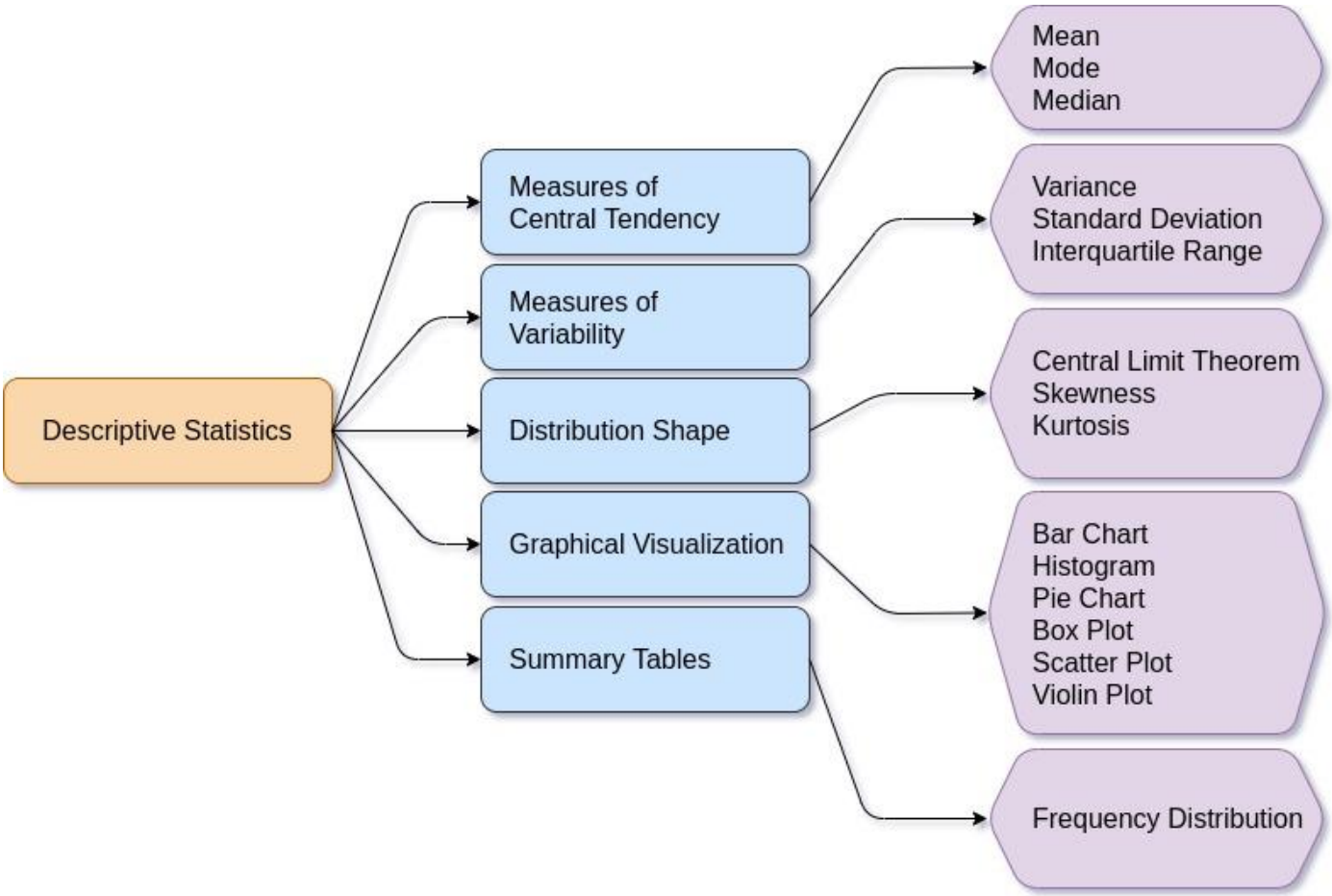
These are also known as Random sampling methods.	These are also called non-random sampling methods.
These are used for research which is conclusive.	These are used for research which is exploratory.
These involve a long time to get the data.	These are easy ways to collect the data quickly.
There is an underlying hypothesis in probability sampling before the study starts. Also, the objective of this method is to validate the defined hypothesis.	The hypothesis is derived later by conducting the research study in the case of non-probability sampling.

From Data to Knowledge

In isolation, raw observations are just data. We use **descriptive statistics** to transform these observations into insights that make sense.
Then we can use **inferential statistics** to study small samples of data and extrapolate our findings to the entire population.

Types of variables			
Quantitative (a.k.a. Numerical)		**Qualitative** (a.k.a. Categorical)	
Continuous	**Discrete**	**Ordinal**	**Nominal**
Consists of numerical values that can be measured but not counted.	Consists of numerical values that can be counted.	Consists of text or labels that have a logical order.	Consists of text or labels that have no logical order.
e.g. Weight, Height, Age, Speed {56.06 Kg, 87 Kg}	e.g. whole no. pin code, No. of children {0, 1, 2, 3}	e.g. Beverage size {small, medium, large}	e.g. Profession {chemist, carpenter}

Descriptive Statistics



Measures of Central Tendency

Measure of Central Tendency is a single value that attempt to describe a set of data identifying the central position.

Mean – The sum of total number devided by how many numbers there are.

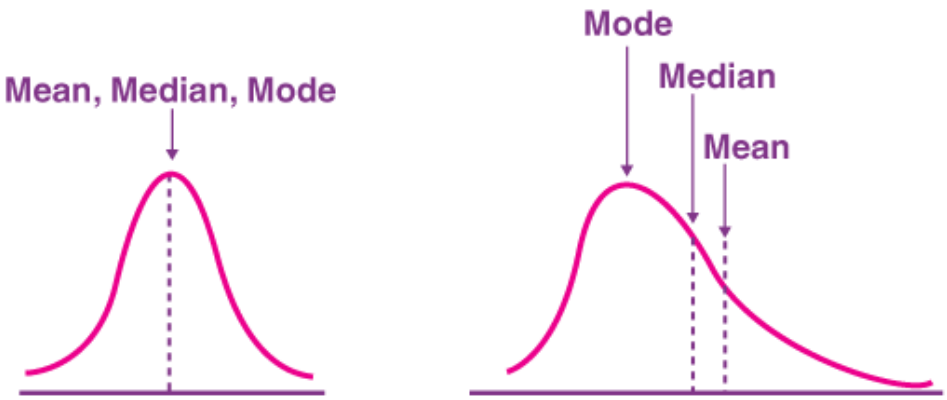
Median – Firstly short the numbers, Find the central number.

- If the no. of elements are even we find the average of central element.
- If the no. of elements are odd we find the central elements.

Mode – The frequent occuring element.

Uses in EDA

- *No Outliers -> Mean*
- *With Outliers -> Median*
- *With Categorical -> Mode*



Mean Median Mode Formula

$$\text{Mean} = \frac{\text{Sum of Observations}}{\text{Total Number of Observations}}$$

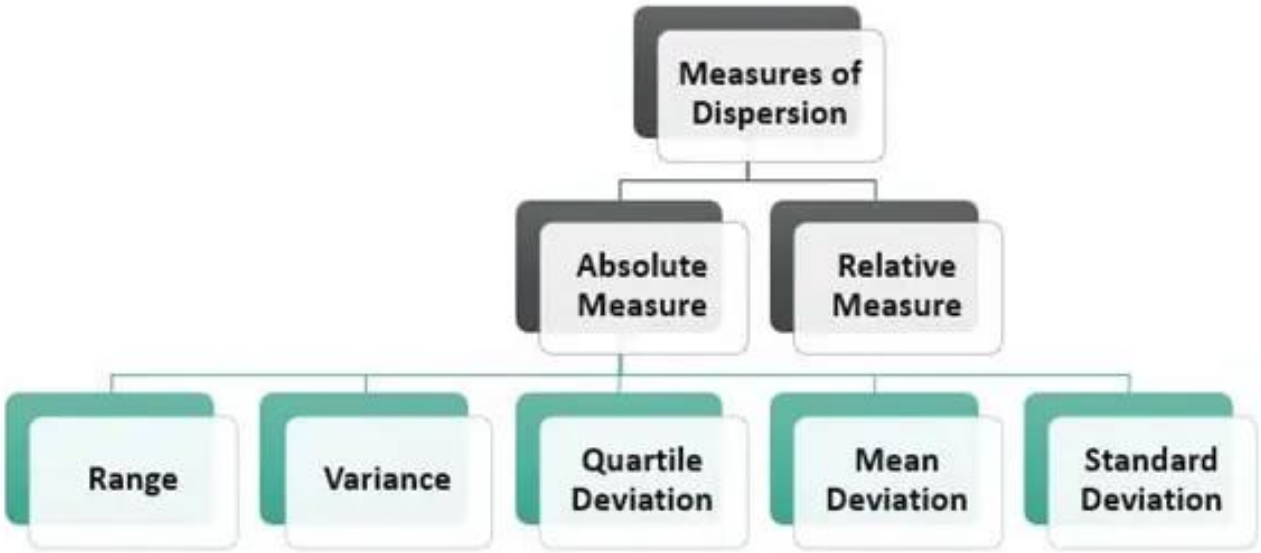
If 'n' is odd: $\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term}$

If 'n' is even: $\text{Median} = \frac{\left(\frac{n}{2}\right)^{th}\text{term} + \left(\frac{n}{2}+1\right)^{th}\text{term}}{2}$

$$\text{Mode} = L + h\,\frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)}$$

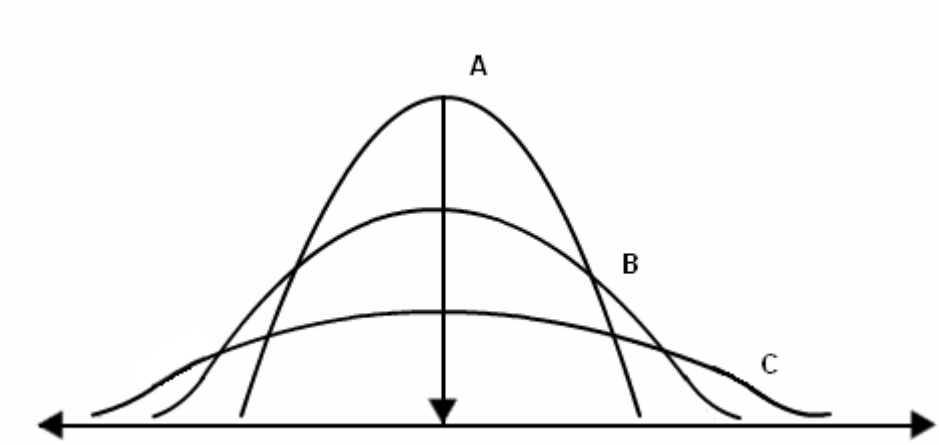
Population Mean	Sample Mean
$\mu = \frac{\sum_{i=0}^{N} x_i}{N}$	$\overline{x} = \frac{\sum_{i=0}^{n} x_i}{n}$
μ = population mean. x_i = individual values in the population data. N = number of values in the population data.	$\overline{x}$ = sample mean. x_i = individual values in the sample data. n = number of values in the sample data.

Measures of Dispersion



Dispersion is the extent to which values in a distribution differ from the average of the distribution. It gives us an idea about the extent to which individual items vary from one another, and from the central value.

Dispersion means the distance of the scattered data from the mean or average value of the data.



Absolute Measure

- Absolute measures of dispersion are expressed in the unit of variable itself, like kilograms, rupees, centimetres, marks etc.

Relative measures

- Relative measures of dispersion are obtained as ratios or percentages of the average.
- These are also known as coefficients of dispersion.

1) Range

- It is the simplest method of measurement of dispersion.
- It is defined as the difference between the largest and the smallest item in a given distribution.

Range = Largest item (L) – Smallest item (S)

(2) Interquartile Range (IQR)

- It is defined as the difference between the Upper Quartile and Lower Quartile of a given distribution.

IQR = Upper Quartile (Q_3) – Lower Quartile (Q_1)

(3) Quartile Deviation

- It is known as Semi-Inter-Quartile Range, i.e. half of the difference between the upper quartile and lower quartile.

Quartile Deviation *= (Upper Quartile (Q_3) – Lower Quartile (Q_1)) / 2*

(4) Mean Deviation

- Mean deviation is the arithmetic mean (average) of deviations |D| of observations from a central value {Mean or Median}.

$$\text{MD about Mean} = \frac{\sum|X-\bar{X}|}{n} = \frac{\sum|D|}{n}$$

where $|D| = |X-\bar{X}|$ and n is the number of observations.

$$\text{MD about Median} = \frac{\Sigma|x-median|}{n} = \frac{\Sigma|D|}{n}$$

where $|D| = |X-median|$ and n is the number of observations.

(5) Variance / Mean-square deviation

According to layman's words, the variance is a measure of how far a set of data are dispersed out from their mean or average value. It is denoted as 'σ^2'.

- It is always non-negative since each term in the variance sum is squared and therefore the result is either positive or zero.
- Variance always has squared units.

The sample variance formula is given as:

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

The population variance formula is given by:

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

Here,

σ^2 = Population variance

s^2 = Sample variance

N = Number of observations in population/sample

X_i = i th observation in the population/sample

μ = Population mean

(6) Standard Deviation / Root mean-square deviation

The spread of statistical data is measured by the standard deviation. Distribution measures the deviation of data from its mean or average position. The degree of dispersion is computed by the method of estimating the deviation of data points. It is denoted by the symbol, 'σ'.

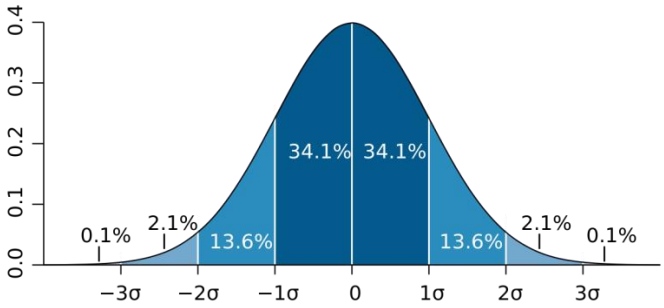
- It describes the square root of the mean of the squares of all values in a data set and is also called the root-mean-square deviation.
- The smallest value of the standard deviation is 0 since it cannot be negative.
- When the data values of a group are similar, then the standard deviation will be very low or close to zero. But when the data values vary with each other, then the standard variation is high or far from zero
- While standard deviation is the square root of the variance, variance is the average of all data points within a group.

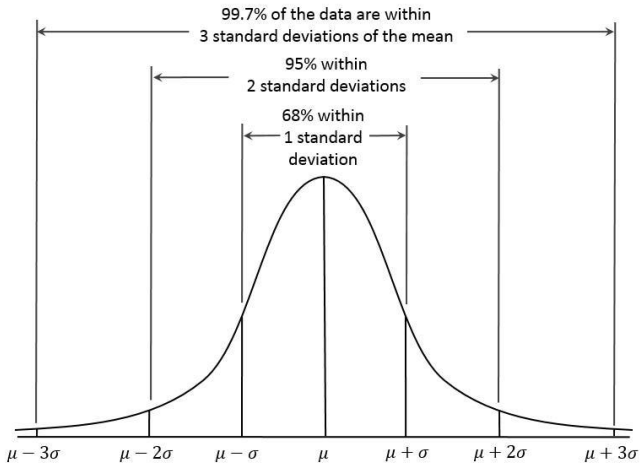
$$\sigma = \sqrt{\sigma^2} \quad \text{-----> Population}$$

$$s = \sqrt{s^2} \quad \text{-----> Sample}$$

Example: if our 5 dogs are just a sample of a bigger population of dogs, we divide by 4 instead of 5 like this:

Sample Variance = 108,520 / 4 = 27,130
Sample Standard Deviation = √27,130 = 165 (to the nearest mm)

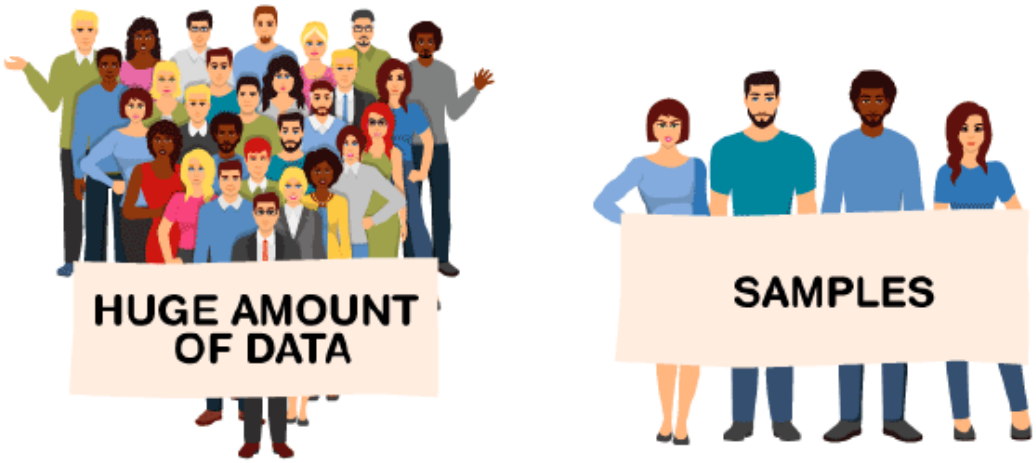




Distribution Shape

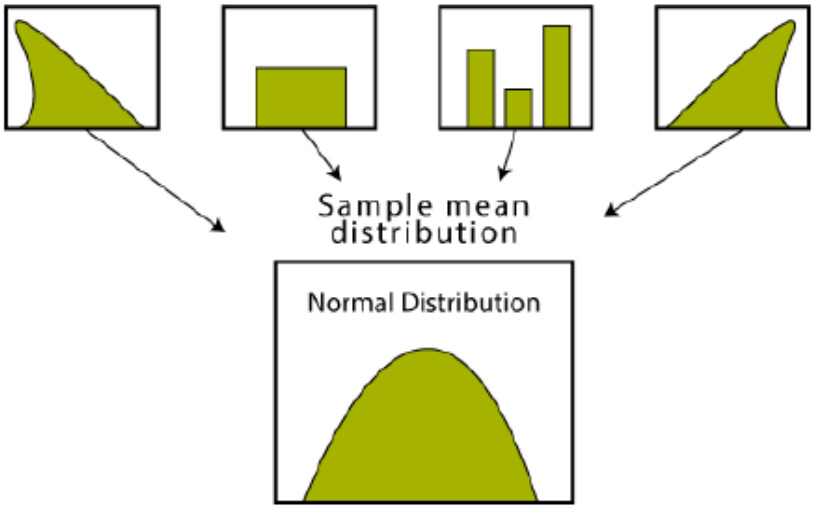
Central Limit Theorem Definition

The Central Limit Theorem (CLT) states that the distribution of a sample mean that approximates the normal distribution, as the sample size becomes larger, assuming that all the samples are similar, and no matter what the shape of the population distribution.



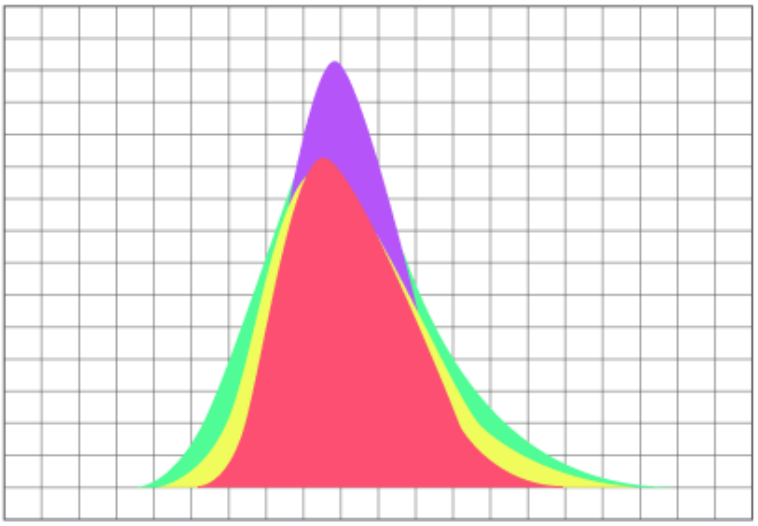
Example:

Suppose you have 10 teams in your school (Sports). Each team will have 100 students in it. Now, we want to measure the average height of the students in the sports team. The simplest way to do would be to find the average of their heights. The first step in this would be to measure the weight of all the students individually and then add them. Then, Divide the sum of their weights with the total number of students. This way we will get the average height. But this method will not make sense for long calculations as it would be tiresome and very long.



So, we will use CTL(Central Limit Theorem) to make the calculation easy. In this method, we will randomly pick students from different teams and make a sample. Each sample will include 20 students. Then, we will follow the following steps to solve it.

1. Take all these samples and find the mean for each individual sample.
2. Now, Find the mean of the sample means.
3. This way we will get the approximate mean height of the students in the sports team.
4. We will get a bell curve shape if we will find the histogram of these sample mean heights.



Note: The sample taken should be sufficient by size. When the sample size gets larger, the sample means distribution will become normality as we calculate it using repeated sampling.

The central limit theorem is applicable for a sufficiently large sample size (n≥30). The formula for central limit theorem can be stated as follows:

Central Limit Theorem Formula

$$\text{Sample mean} = \text{Population mean} = \mu$$

$$\text{Sample standard deviation} = \frac{(\text{Standard deviation})}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

Where,

μ = Population mean

σ = Population standard deviation

μ_x = Sample mean

σ_x= Sample standard deviation

n = Sample size

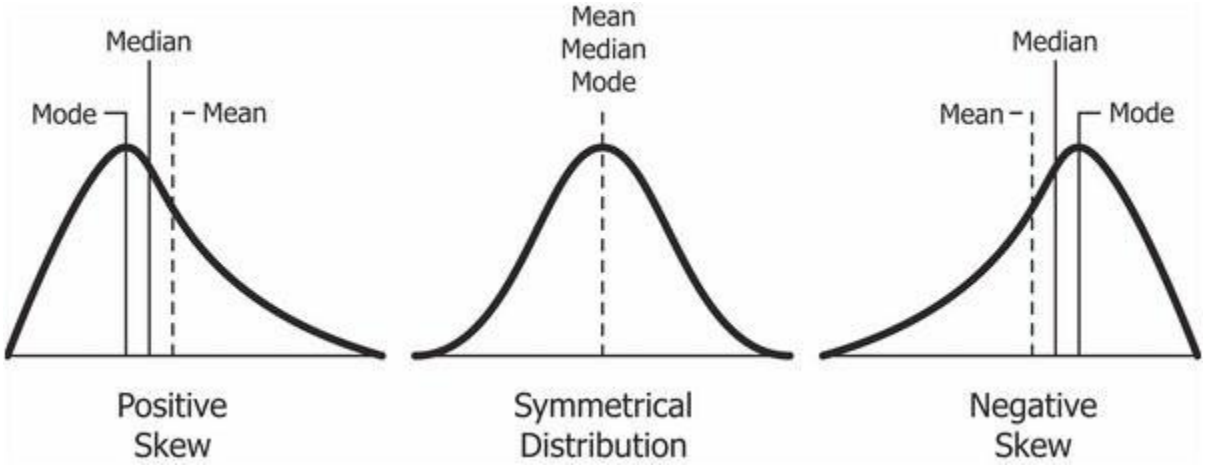
Applications of Central Limit Theorem

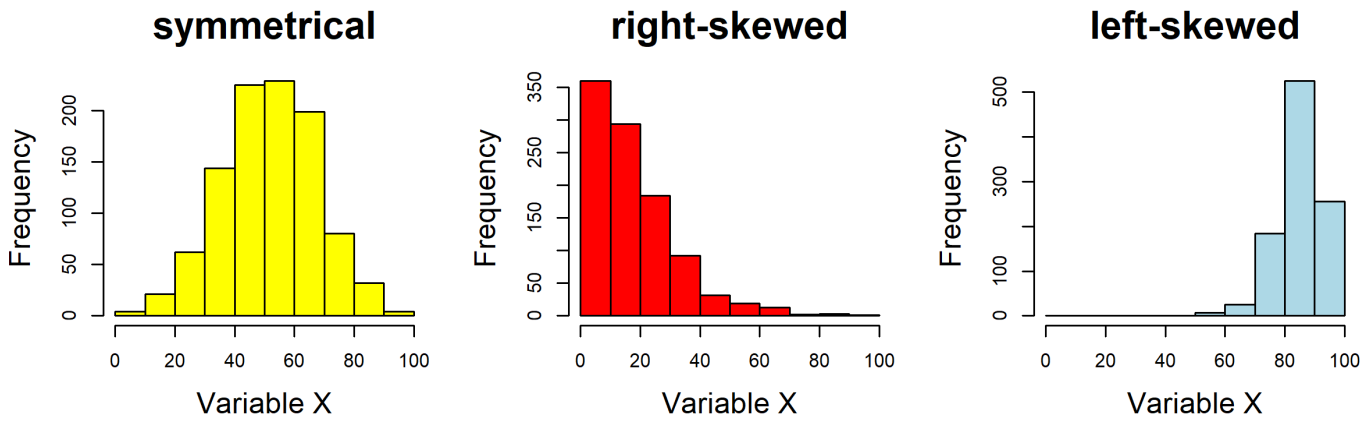
Statistical Application of CLT	Practical Significance of CLT
If the distribution is not known or not normal, we consider the sample distribution to be normal according to CTL. As this method assume that the population given is normally distributed. This helps in analyzing data in methods like constructing confidence intervals.	One of the most common applications of CLT is in election polls. To calculate the percentage of persons supporting a candidate which are seen on news as confidence intervals.
To estimate the population mean more accurately, we can increase the samples taken from the population which will ultimately decrease the sample means deviation.	It is also used to measure the mean or average family income of a family in a particular region.
To create a range of values which is likely to include the population mean, we can use the sample mean.	

Skewness

Skewness is the measure of how much the probability distribution of a random variable deviates from the normal distribution.

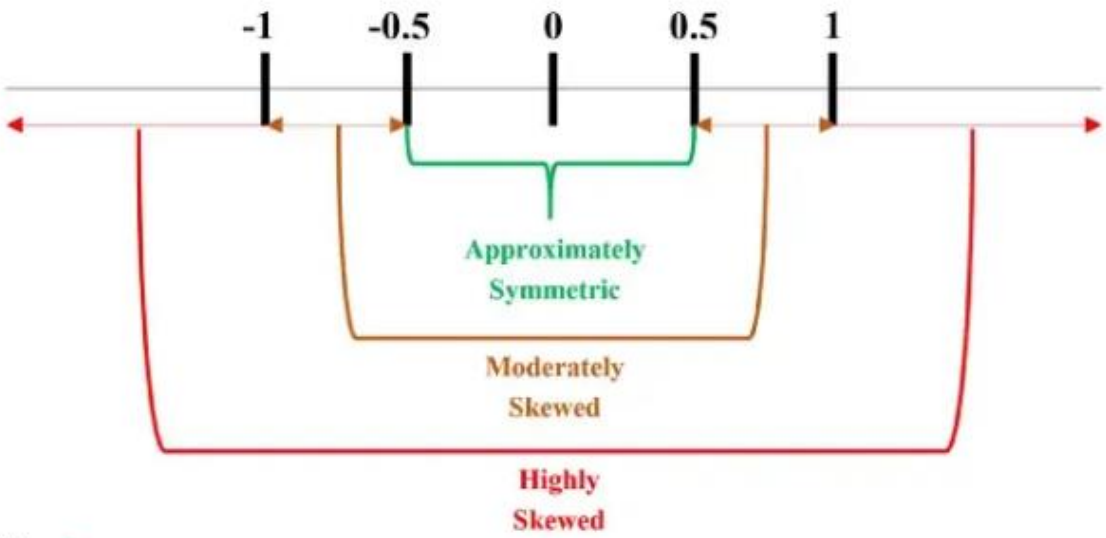
A normal distribution is a bell curve with a perfect symmetric shape. If the curve becomes asymmetric or extends toward the right or left, it is called a skewed bell curve. What is skewness? Skewness in statistics represents an imbalance of a normal distribution. This means that the data set has outliers or extreme values in its distribution.





Skewness = 3(Mean−Median) / Standard Deviation

Scale of Skewness:



(-0.5, 0.5) = Low

(-1, -0.5) U (0.5, 1) = Moderate

(-1 & beyond) U (1 & beyond) = High

$$\frac{(Mean - Mode)}{Standard\ Deviation}$$

Pearson Measure of Skewness

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

Pearson Measure of Skewness (Alternative Form)

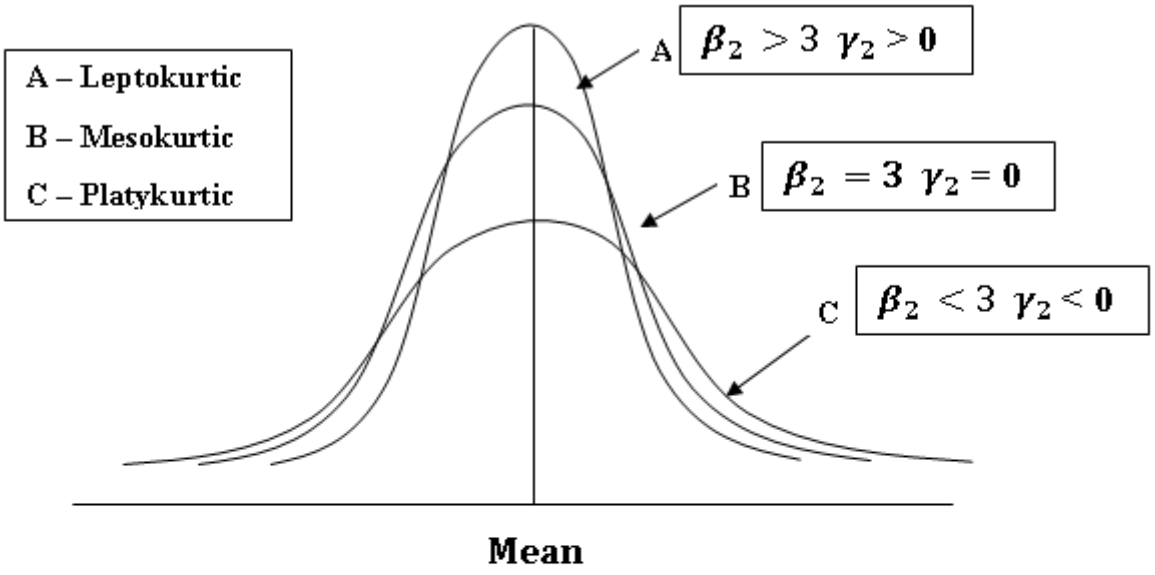
$$\frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_1 - 2Q_2 + Q_3}{Q_3 - Q_1}$$

* where Q_1=First Quartile, Q_2=Second Quartile, Q_3= Third Quartile

Bowley's Coefficient of Skewness

Kurtosis

Kurtosis is a statistical measure that's used to describe the distribution, or skewness, of observed data around the mean, sometimes referred to as the volatility of volatility. Kurtosis is used generally in the statistical field to describe trends in charts. Kurtosis can be present in a chart with fat tails and a low, even distribution, as well as be present in a chart with skinny tails and a distribution concentrated toward the mean.



$$k = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$

Probability

Probability means possibility. It is a branch of mathematics that deals with the occurrence of a random event. The value is expressed from zero to one.

Probability is a measure of the likelihood of an event to occur. Many events cannot be predicted with total certainty. We can predict only the chance of an event to occur i.e., how likely they are going to happen, using it. Probability can range from 0 to 1, where 0 means the event to be an impossible one and 1 indicates a certain event.

Formula for Probability

The probability formula is defined as the possibility of an event to happen is equal to the ratio of the number of favourable outcomes and the total number of outcomes.

Probability of event to happen P(E) = Number of favourable outcomes/Total Number of outcomes

1) There are 6 pillows in a bed, 3 are red, 2 are yellow and 1 is blue. What is the probability of picking a yellow pillow?

Ans: The probability is equal to the number of yellow pillows in the bed divided by the total number of pillows, i.e. 2/6 = 1/3.

2) There is a container full of coloured bottles, red, blue, green and orange. Some of the bottles are picked out and displaced. Sumit did this 1000 times and got the following results:

- No. of blue bottles picked out: 300
- No. of red bottles: 200
- No. of green bottles: 450
- No. of orange bottles: 50

a) What is the probability that Sumit will pick a green bottle?

Ans: For every 1000 bottles picked out, 450 are green.

Therefore, P(green) = 450/1000 = 0.45

b) If there are 100 bottles in the container, how many of them are likely to be green?

Ans: The experiment implies that 450 out of 1000 bottles are green.

Therefore, out of 100 bottles, 45 are green.

Types of Probability

There are three major types of probabilities:

- Theoretical Probability
- Experimental Probability
- Axiomatic Probability

Theoretical Probability

It is based on the possible chances of something to happen. The theoretical probability is mainly based on the reasoning behind probability. For example, if a coin is tossed, the theoretical probability of getting a head will be ½.

Experimental Probability

It is based on the basis of the observations of an experiment. The experimental probability can be calculated based on the number of possible outcomes by the total number of trials. For example, if a coin is tossed 10 times and head is recorded 6 times then, the experimental probability for heads is 6/10 or, 3/5.

Axiomatic Probability

In axiomatic probability, a set of rules or axioms are set which applies to all types. These axioms are set by Kolmogorov and are known as Kolmogorov's three axioms. With the axiomatic approach to probability, the chances of occurrence or non-occurrence of the events can be quantified. The axiomatic probability lesson covers this concept in detail with Kolmogorov's three rules (axioms) along with various examples.

Conditional Probability is the likelihood of an event or outcome occurring based on the occurrence of a previous event or outcome.

Probability of an Event

Assume an event E can occur in r ways out of a sum of n probable or possible **equally likely ways**. Then the probability of happening of the event or its success is expressed as;

$P(E) = r/n$

The probability that the event will not occur or known as its failure is expressed as:

$P(E') = (n-r)/n = 1-(r/n)$

E' represents that the event will not occur.

Therefore, now we can say;

$P(E) + P(E') = 1$

This means that the total of all the probabilities in any random test or experiment is equal to 1.

What are Equally Likely Events?

When the events have the same theoretical probability of happening, then they are called equally likely events. The results of a sample space are called equally likely if all of them have the same probability of occurring. For example, if you throw a die, then the probability of getting 1 is 1/6. Similarly, the probability of getting all the numbers from 2,3,4,5 and 6, one at a time is 1/6. Hence, the following are some examples of equally likely events when throwing a die:

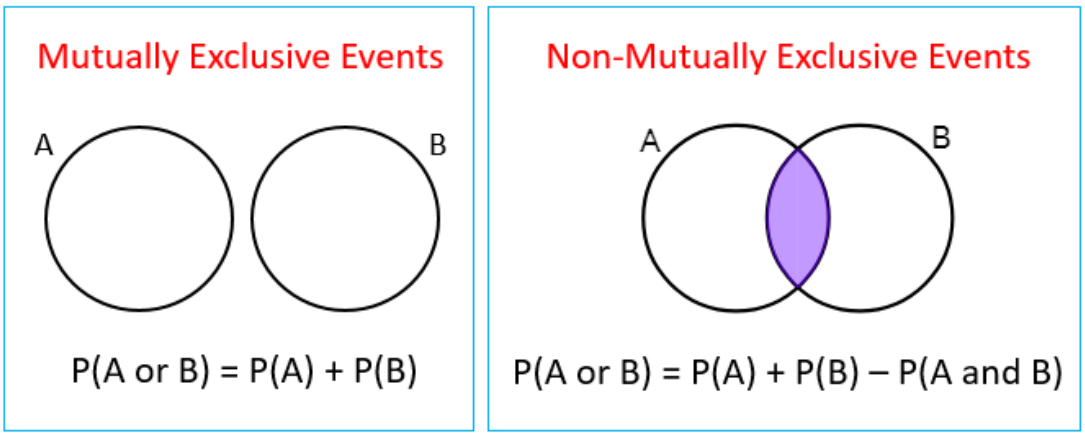
- Getting 3 and 5 on throwing a die
- Getting an even number and an odd number on a die
- Getting 1, 2 or 3 on rolling a die

are equally likely events, since the probabilities of each event are equal.

Complementary Events

The possibility that there will be only two outcomes which states that an event will occur or not. Like a person will come or not come to your house, getting a job or not getting a job, etc. are examples of complementary events. Basically, the complement of an event occurring in the exact opposite that the probability of it is not occurring. Some more examples are:

- It will rain or not rain today
- The student will pass the exam or not pass.
- You win the lottery or you don't.



Mutually exclusive events are those events that do not occur at the same time. For example, when a coin is tossed then the result will be either head or tail, but we cannot get both the results. Such events are also called disjoint events since they do not happen simultaneously.

Question 1: What is the probability of a die showing a number 3 or number 5?

Solution: Let,

P(3) is the probability of getting a number 3

P(5) is the probability of getting a number 5

P(3) = 1/6 and P(5) = 1/6

So,

P(3 or 5) = P(3) + P(5)

P(3 or 5) = (1/6) + (1/6) = 2/6

P(3 or 5) = 1/3

Therefore, the probability of a die showing 3 or 5 is 1/3.

How to Find Mutually Exclusive Events?

In probability, the specific addition rule is valid when two events are mutually exclusive. It states that the probability of either event occurring is the sum of probabilities of each event occurring. If A and B are said to be mutually exclusive events then the probability of an event A occurring or the probability of event B occurring that is P (a ∪ b) formula is given by P(A) + P(B), i.e.,

- **P (A Or B) = P(A) + P(B)**
- **P (A ∪ B) = P(A) + P(B)**

Note: If the events A and B are not mutually exclusive, the probability of getting A or B that is

P (A ∪ B) formula is given as follows:

P (A ∪ B) = P(A) + P(B) – P (A and B)

Dependent and Independent Events

Two events are said to be dependent if the occurrence of one event changes the probability of another event. Two events are said to be independent events if the probability of one event does not affect the probability of another event. If two events are mutually exclusive, they are not independent. Also, independent events cannot be mutually exclusive.

From the definition of mutually exclusive events, certain rules for probability are concluded.

- Addition Rule: P (A + B) = 1
- Subtraction Rule: $P (A \cup B)' = 0$
- Multiplication Rule: $P (A \cap B) = 0$

Events Associated with "OR"

If two events E_1 and E_2 are associated with **OR** then it means that either E_1 or E_2 or both. The union symbol **(U)** is used to represent OR in probability.

Events Associated with "AND"

If two events E_1 and E_2 are associated with **AND** then it means the intersection of elements which is common to both the events. The intersection symbol **(∩)** is used to represent AND in probability.

Event E_1 but not E_2

It represents the difference between both the events. Event E_1 but not E_2 represents all the outcomes which are present in E_1 but not in E_2. Thus, the event E_1 but not E_2 is represented as

$E_1, E_2 = E_1 – E_2$

Probability Density Function

The Probability Density Function (PDF) is the probability function which is represented for the density of a continuous random variable lying between a certain range of values. Probability Density Function explains the normal distribution and how mean and deviation exists.
The standard normal distribution is used to create a database or statistics, which are often used in science to represent the real-valued variables, whose distribution is not known.

Probability Terms and Definition

Some of the important probability terms are discussed here:

Term	Definition	Example
Sample Space	The set of all the possible outcomes to occur in any trial	1. Tossing a coin, Sample Space (S) = {H,T} 2. Rolling a die, Sample Space (S) = {1,2,3,4,5,6}

Term	Definition	Example
Sample Point	It is one of the possible results	In a deck of Cards: • 4 of hearts is a sample point. • The queen of clubs is a sample point.
Experiment or Trial	A series of actions where the outcomes are always uncertain.	The tossing of a coin, Selecting a card from a deck of cards, throwing a dice.
Event	It is a single outcome of an experiment.	Getting a Heads while tossing a coin is an event.
Outcome	Possible result of a trial/experiment	T (tail) is a possible outcome when a coin is tossed.
Complimentary event	The non-happening events. The complement of an event A is the event, not A (or A')	In a standard 52-card deck, A = Draw a heart, then A' = Don't draw a heart
Impossible Event	The event cannot happen	In tossing a coin, impossible to get both head and tail at the same time

Applications of Probability

Probability has a wide variety of applications in real life. Some of the common applications which we see in our everyday life while checking the results of the following events:

- Choosing a card from the deck of cards
- Flipping a coin
- Throwing a dice in the air
- Pulling a red ball out of a bucket of red and white balls
- Winning a lucky draw

Other Major Applications of Probability

- It is used for risk assessment and modelling in various industries
- Weather forecasting or prediction of weather changes
- Probability of a team winning in a sport based on players and strength of team
- In the share market, chances of getting the hike of share prices.

General Probability Rules

Rule 1: The probability of an impossible event is zero; the probability of a certain event is one. Therefore, for any event A, the range of possible probabilities is: $0 \leq P(A) \leq 1$

Rule 2: For S the sample space of all possibilities, P(S) = 1. That is the sum of all the probabilities for all possible events is equal to one. Recall the party affiliation above: if you have to belong to one of the three designated political parties, then the sum of P(R), P(D) and P(I) is equal to one.

Rule 3: For any event A, $P(A^c) = 1 - P(A)$. It follows then that $P(A) = 1 - P(A^c)$

Rule 4 (Addition Rule): This is the probability that **either one or both** events occur

a. If two events, say A and B, are **mutually exclusive** - that is A and B have no outcomes in common - then P(A or B) = P(A) + P(B)

b. If two events are NOT mutually exclusive, then P(A or B) = P(A) + P(B) - P(A and B)

Rule 5 (Multiplication Rule): This is the probability that **both** events occur

a. P(A and B) = P(A) • P(B|A) or P(B)*P(A|B)

Note: this straight line symbol, |, does not mean divide! This symbols means "conditional" or "given". For instance P(A|B) means the probability that event A occurs *given* event B has occurred.

b. If A and B are independent - neither event influences or affects the probability that the other event occurs - then P(A and B) = P(A)*P(B). This particular rule extends to more than two independent events. For example, P(A and B and C) = P(A)*P(B)*P(C)

Rule 6 (Conditional Probability): P(A|B)=P(A and B)/P(B) or P(B|A)=P(A and B)/P(A)

Permutation

- Permutation is the arrangement of items in which **order matters**
- Number of ways of **selection and arrangement of items** in which Order Matters

$$ {}^{n}P_{r} = \frac{n!}{(n-r)!} $$

Combination

- Combination is the selection of items in which **order does not matters** .
- Number of ways of **selection of items** in which Order does not Matters

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

Permutations = Position or Placement matter
Combinations = Couldn't Care Less about order

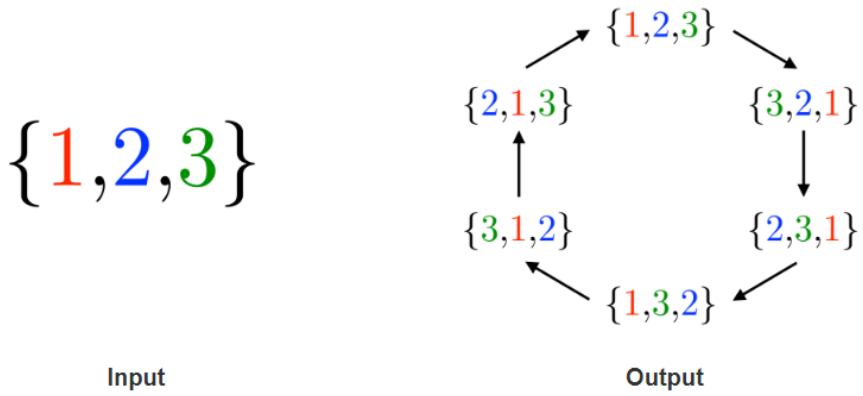
In mathematics, **permutation** relates to the act of arranging all the members of a set into some sequence or order.

The **combination** is a way of selecting items from a collection, such that (unlike permutations) the order of selection does not matter.

Permutation Formula: A permutation is the choice of r things from a set of n things without replacement and where the order matters.
nPr = n!/(n-r)!

Combination Formula: A combination is the choice of r things from a set of n things without replacement and where order does not matter.
nCr = n!/(r! (n-r)!)

Generating Permutations



Input Description: An integer n.
Problem: Generate (1) all, or (2) a random, or (3) the next permutation of length n.

$$\frac{{}_nP_r}{x_1!} = \frac{{}_5P_5}{2!}$$

There are 5 letters, n = 5, and you are choosing all 5 digits, r = 5

$$\frac{{}_5P_5}{2!} = \frac{5\times4\times3\times2\times1}{2\times1} = \frac{120}{2}$$

There are 2 letters repeating (P's), therefore divide by 2!

$$\frac{{}_5P_5}{2!} = 60\ arrangements$$

Combination Formula

A combination is a grouping or subset of items.
For a combination, the order does not matters.

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

Number of items in set

Number of items selected from the set

Example of counting combinations [edit]

As a specific example, one can compute the number of five-card hands possible from a standard fifty-two card deck as:[7]

$$\binom{52}{5} = \frac{52 \times 51 \times 50 \times 49 \times 48}{5 \times 4 \times 3 \times 2 \times 1} = \frac{311{,}875{,}200}{120} = 2{,}598{,}960.$$

Alternatively one may use the formula in terms of factorials and cancel the factors in the numerator against parts of the factors in the denominator, after which only multiplication of the remaining factors is required:

$$\begin{aligned}
\binom{52}{5} &= \frac{52!}{5!47!} \\
&= \frac{52 \times 51 \times 50 \times 49 \times 48 \times \cancel{47!}}{5 \times 4 \times 3 \times 2 \times \cancel{1} \times \cancel{47!}} \\
&= \frac{52 \times 51 \times 50 \times 49 \times 48}{5 \times 4 \times 3 \times 2} \\
&= \frac{(26 \times \cancel{2}) \times (17 \times \cancel{3}) \times (10 \times \cancel{5}) \times 49 \times (12 \times \cancel{4})}{\cancel{5} \times \cancel{4} \times \cancel{3} \times \cancel{2}} \\
&= 26 \times 17 \times 10 \times 49 \times 12 \\
&= 2{,}598{,}960.
\end{aligned}$$

Another alternative computation, equivalent to the first, is based on writing

$$\binom{n}{k} = \frac{(n-0)}{1} \times \frac{(n-1)}{2} \times \frac{(n-2)}{3} \times \cdots \times \frac{(n-(k-1))}{k},$$

Selecting 4 fruits out of 10 fruits

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

$${}^{10}C_{4} = C(n, r) = C(10, 4)$$

$$= \frac{10!}{(4!(10-4)!)}$$

$$= \frac{10!}{4! \times 6!}$$

$$= 210 \text{ ways}$$

- Factorial of any negative quantity is not valid.
- If a particular thing can be done in m ways and another thing can be done in n ways, then
  - Either one of the two can be done in ***m + n*** ways and
  - Both of them can be done in ***m × n*** ways
- 0! = 1
- 1! = 1
- If from the total set of n objects and 'p_1' are of one kind and 'p_2' and 'p_3' and so on …. till p_r are others respectively then *nPr* = {n!}/{p_1 ! × q_2 ! × ……. p_r ! }
- ${}^nP_n = n!$
- ${}^nc_n = 1$
- ${}^nc_0 = 1$
- ${}^nc_r = {}^nc_{(n-r)}$
- ${}^nc_0+{}^nc_1+{}^nc_2+{}^nc_3+\ldots{}^nc_n = 2^n$

Permutation and Combination Formulas- Factorial

n ! = n(n-1)(n-2) …… 1

Eg. – 5! = 5(5-1)(5-2)(5-3)(5-4) = 5(4)(3)(2)(1)

Standard Truths

- 0! = 1
- n! only exists of n >= 0 and doesn't exist for n < 0

n	n!
0	1
1	1
2	2
3	6
4	24
5	120
6	720
7	5 040
8	40,320
9	362 880
10	3 628 800

Permutations Formulas

Number of ways in which Permutations out of n things r things can be SELECTED & ARRANGED (denoted by ${}^nP_r$).

${}^nP_r$ = number of permutations (arrangements) of n things taken r at a time.

nPr = {n!}/{(n-r)!} ***n>=r***

Eg.

- Arrangement of Letters/Alphabets to form words with meaning or without meaning.
- Arrangements of balls on a table.

Formulas for Combinations

The number of ways in which r things at a time can be SELECTED from from n things is Combinations (represented by ${}^{n}C_{r}$)..

${}^{n}C_{r}$ = Number of combinations (selections) of n things taken r at a time.

- ***nPr* = {n!}/{(r)! (n-r)!} ; where n ≥ r (n is greater than or equal to r).**

Eg.

- Selections for people from total numbers who want to go out on a picnic.
- Filling posts with people
- Selection for a sports team out of available players
- Selection of balls from a bag

Important Properties:

Property 1

Number of permutations (or arrangements) of n different things taken all at a time = n!

Property 2

For Objects in which P1 are alike and are of one type, P2 are alike or other different type and P3 are alike or another different type and the rest must be all different,

***Number of permutations* = {n!}/(*p*1)!(*p*2)!(*p*3)!**

Property 3

When repetition is allowed number of permutations of n different things taken r at a time = n × n× n ×… (r times) = n^{r}

Property 4

Here, we are counting the number of ways in which ***k balls*** can be distributed into ***n boxes*** under various conditions.

The conditions which are generally asked are

1. The balls are either distinct or identical.
2. No box can contain more than one ball or any box may contain more than one ball.
3. No box can be empty or any box can be empty.

Covariance

In the study of covariance only sign matters. A positive value shows that both variables vary in the same direction and negative value shows that they vary in the opposite direction.

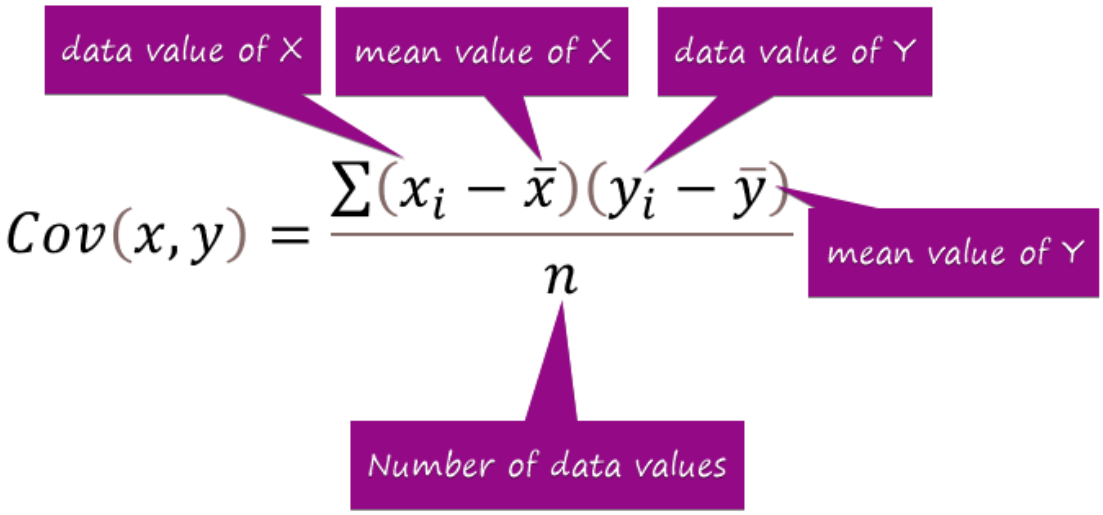
Types of Covariance

Covariance can have both positive and negative values. Based on this, it has two types:

1. Positive Covariance
2. Negative Covariance

Covariance Formula

Covariance formula is a statistical formula, used to evaluate the relationship between two variables. It is one of the statistical measurements to know the relationship between the variance between the two variables. Let us say X and Y are any two variables, whose relationship has to be calculated. Thus the covariance of these two variables is denoted by Cov(X,Y). The formula is given below for both population covariance and sample covariance.



Significance of the formula:

- Numerator: Quantity of variance in x multiplied by the quantity of variance in y.
- Unit of covariance: Unit of x multiplied by a unit of y
- Hence if we change the unit of variables, covariance will have new value however sign will remain the same.
- Therefore the numerical value of covariance does not have any significance however if it is positive then both variables vary in the same direction else if it is negative then they vary in the opposite direction.

Population Covariance Formula

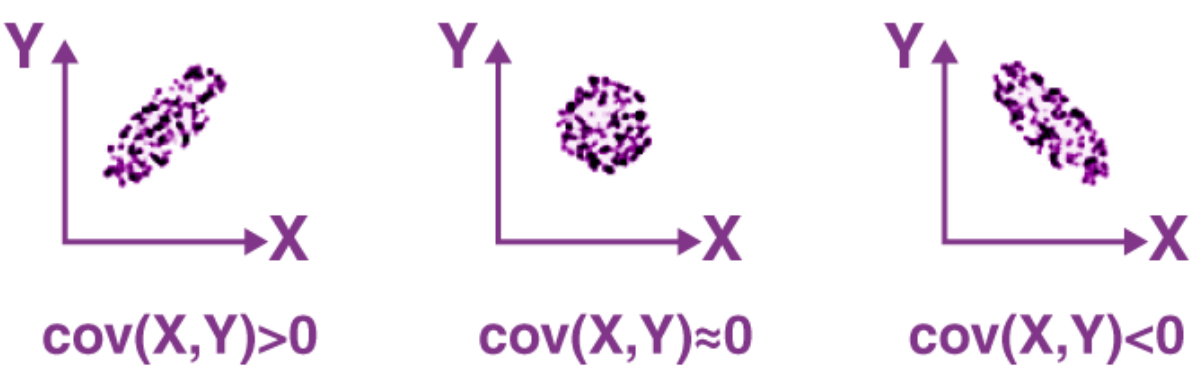
$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample Covariance

$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - y)}{N - 1}$$

Where,

- xi = data value of x
- yi = data value of y
- $\bar{x}$ = mean of x
- $\bar{y}$ = mean of y
- N = number of data values.



Correlation

As covariance only tells about the direction which is not enough to understand the relationship completely, we divide the covariance with a standard deviation of x and y respectively and get correlation coefficient which varies between -1 to +1.

$$\text{Correlation}, \rho(X,Y) = \text{Cov}(X,Y)/\sigma_x \sigma_y$$

- -1 and +1 tell that both variables have a perfect linear relationship.
- Negative means they are inversely proportional to each other with the factor of correlation coefficient value.
- Positive means they are directly proportional to each other mean vary in the same direction with the factor of correlation coefficient value.
- if the correlation coefficient is 0 then it means there is no linear relationship between variables however there could exist other functional relationship.
- if there is no relationship at all between two variables then correlation coefficient will certainly be 0 however if it is 0 then we can only say that there is no linear relationship but there could exist other functional relationship.

Formula for Correlation Coefficient

cuemath
THE MATH EXPERT

Population Correlation Coefficient

$$P_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sqrt{\left(\sum (xi - \bar{x})^2\right)\left(\sum (yi - \bar{y})^2\right)}}$$

Where, $\sigma_x, \sigma_y \rightarrow$ Population Standard Deviation

$\sigma_{xy} \rightarrow$ Population Covariance

$\bar{X}, \bar{Y} \rightarrow$ Population Mean

Sample Correction, coefficient between x and y

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sqrt{\left(\sum (xi - \bar{x})^2\right)\left(\sum (yi - \bar{y})^2\right)}}$$

Where, $S_x, S_y \rightarrow$ Sample Standard Deviation

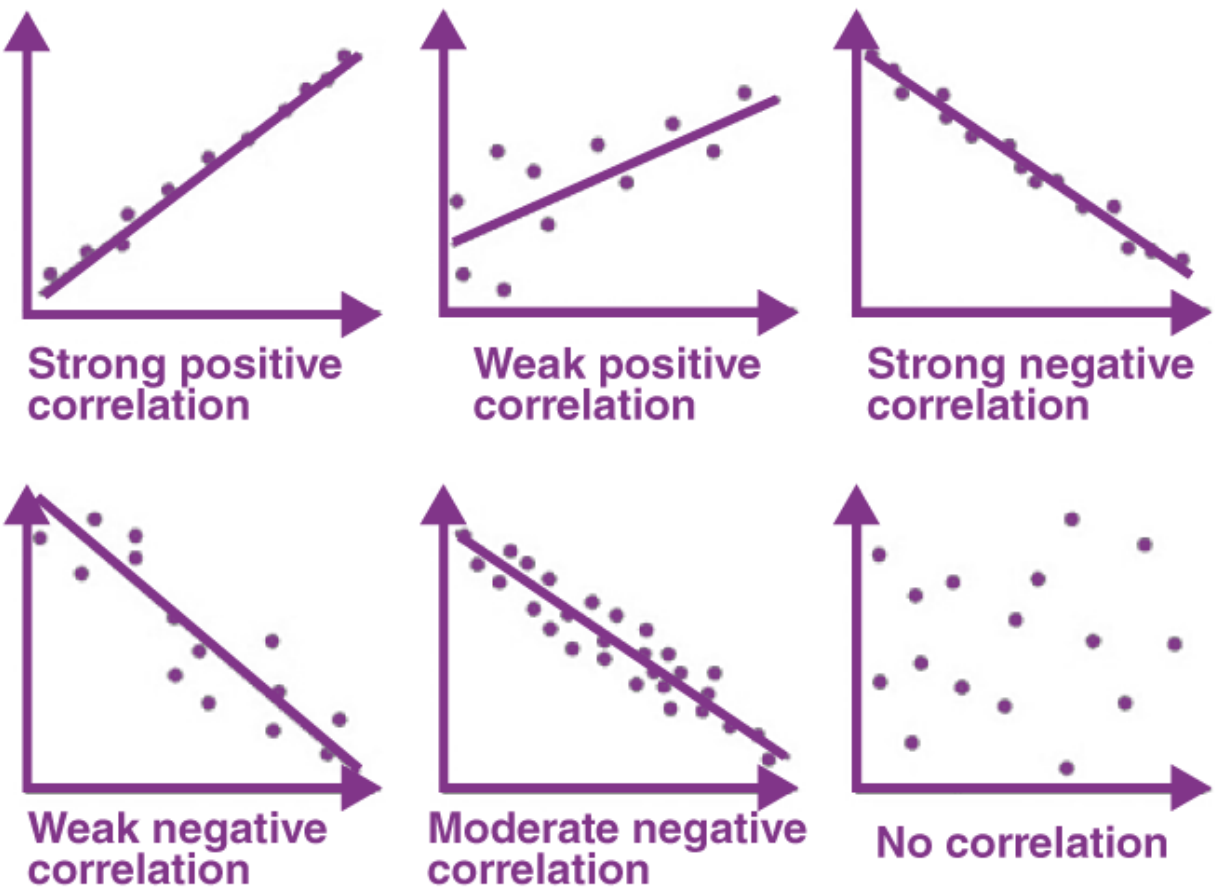
$S_{xy} \rightarrow$ Sample Covariance

$\bar{X}, \bar{Y} \rightarrow$ SampleMean

Where:

- S_xy is the covariance between x and y.
- S_x and S_y are the standard deviations of x and y respectively.
- r_xy is the correlation coefficient.
- The correlation coefficient is a dimensionless quantity. Hence if we change the unit of x and y then also the coefficient value will remain the same.

Let’s understand what is the significance of the correlation coefficient with the help of the below graph:



Covariance and Correlation

Below table shows the comparison among covariance and correlation in brief.

Covariance	**Correlation**
It is a measure to show the extent to which given two random variables change with respect to each other.	It is a measure used to describe how strongly the given two random variables are related to each other.
It is a measure of correlation.	It is defined as the scaled form of covariance.
The value of covariance lies between -∞ and +∞.	The value of correlation lies between -1 and +1.
It indicates the direction of the linear relationship between the given two variables.	It measures the direction and strength of the linear relationship between the given two variables.

Covariance and Variance

Covariance and variance both are the terms used in statistics. Variance is the measure of spread of data around its mean value but covariance measures the relation between two random variables. Learn Variance in statistics at BYJU'S.

Covariance Example

Below example helps in better understanding of the covariance of among two variables.

Question:

Calculate the coefficient of covariance for the following data:

X	2	8	18	20	28	30
Y	5	12	18	23	45	50

Solution:

Number of observations = 6

Mean of X = 17.67

Mean of Y = 25.5

Cov(X, Y) = (⅙) [(2 – 17.67)(5 – 25.5) + (8 – 17.67)(12 – 25.5) + (18 – 17.67)(18 – 25.5) + (20 – 17.67)(23 – 25.5) + (28 – 17.67)(45 – 25.5) + (30 – 17.67)(50 – 25.5)]

= 157.83

Spearman's Rank Correlation

Spearman's rank correlation measures the strength and direction of association between two ranked variables. It basically gives the measure of monotonicity of the relation between two variables i.e. how well the relationship between two variables could be represented using a monotonic function.

The formula for Spearman's rank coefficient is:

$$\rho = 1 - \frac{6\Sigma\, d_i^2}{n(n^2 - 1)}$$

$\boldsymbol{\rho}$ = Spearman's rank correlation coefficient

di = Difference between the two ranks of each observation

n = Number of observations

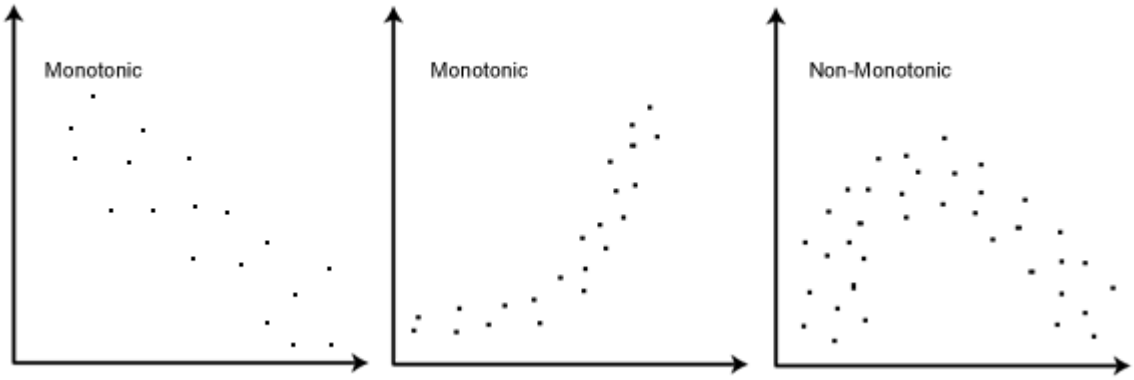
The Spearman Rank Correlation can take a value from +1 to -1 where,

- A value of +1 means a perfect association of rank
- A value of 0 means that there is no association between ranks
- A value of -1 means a perfect negative association of rank

What Is Monotonic Function?

To understand Spearman's rank correlation, it is important to understand monotonic function. A monotonic function is one that either never increases or never decreases as its independent variable changes.

The following graph illustrates the monotonic function:



- Monotonically Increasing: As the variable X increases, the variable Y never decreases.
- Monotonically Decreasing: As the variable X increases, the variable Y never increases.
- Not Monotonic: As the X variable increases, the Y variable sometimes decreases and sometimes increases.

Example of Spearman's Rank Correlation

Consider the score of 5 students in Maths and Science that are mentioned in the table.

Students	Maths	Science
A	35	24
B	20	35
C	49	39
D	44	48
E	30	45

Step 1: Create a table for the given data.

Step 2: Rank both the data in descending order. The highest marks will get a rank of 1 and the lowest marks will get a rank of 5.

Step 3: Calculate the difference between the ranks (d) and the square value of d.

Step 4: Add all your d square values.

Students	Maths	Rank	Science	Rank	d	d square
A	35	3	24	5	2	4
B	20	5	35	4	1	1
C	49	1	39	3	2	4
D	44	2	48	1	1	1
E	30	4	45	2	2	4
						14

Step 5: Insert these values into the formula.

$$\rho = 1 - \frac{6\Sigma\, d_i^2}{n(n^2 - 1)}$$

= 1 - (6 * 14) / 5(25 - 1)

= 0.3

The Spearman's Rank Correlation for the given data is 0.3. The value is near 0, which means that there is a weak correlation between the two ranks.

Percentiles and Quartiles

A **percentile** is a value below below which a certain percentage of observation lie.

Note: *First we need to short the data/population before starting the process.*

Percentile ----> GAT, CAT, JEE, NEET

Eg: 99 percentile ---> It means the person has got better marks than 99% of the entire marks.

Percentile(x) = (Number of values fall under 'x'/total number of values) × 100

$$P = (n/N) \times 100$$

Where :

n = no. of values below , N = total count of population

Eg: What is percentile value for the score 80 for the given population 50,100,70,80,56,60,80,75.

Sol = *The given data is not sorted. So first sort the data in ascending order.*

Sorted data: 50, 56, 60, 70, 75, 80, 80, 100

Number of values fall under 80 (n) = 5

Total count of values (N) = 8

Percentile = (n/N) × 100

= (5/8) × 100

= 62.5

The percentile of value 80 for the given population is 62.5

Eg: What is the percentile value for the value 60 in a given population of weights of persons 50, 55, 40, 60, 100, 95, 90, 60, 80, 75.

Solution: The given data is not sorted. So first sort the data in ascending order.

Sorted data: 40,50,55,60,60,75,80,90,95,100

Number of values fall under 60 (n)= 3

Total count of values (N)= 10

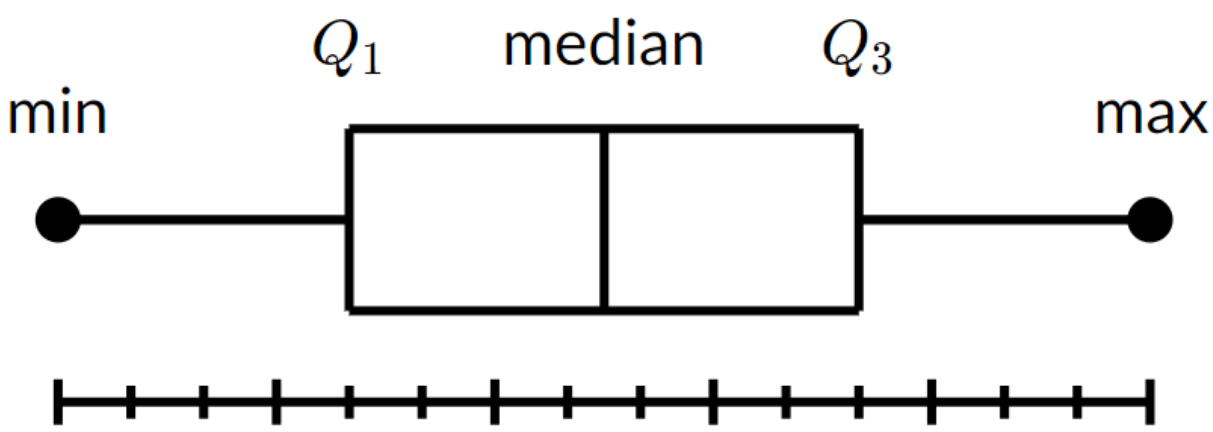
Percentile = (n/N) x 100

= (3/10) x 100

= 30

The percentile of value 60 for the given population is 30.

Five number summary for box plot



The five-number summary is the minimum, first quartile, median, third quartile, and maximum.

The Box and Whisker Plot also called a box plot, uses the following five-number summary:

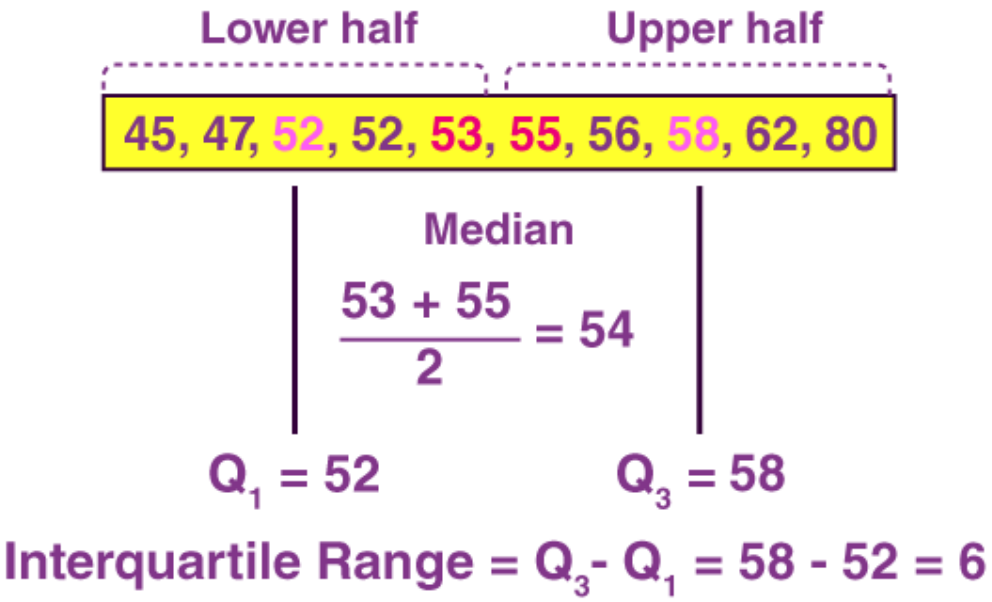
1. Minimum
2. Quartile 1
3. Median
4. Quartile 3
5. Maximum

These formula gives index of the data value if the index is float then we get average.

$$\text{Lower Quartile (Q1)} = (N+1) \times \frac{1}{4}$$

$$\text{Middle Quartile (Q2)} = (N+1) \times \frac{2}{4}$$

$$\text{Upper Quartile (Q3)} = (N+1) \times \frac{3}{4}$$



Example: Find the median, lower quartile, upper quartile and interquartile range of the following data set of scores:

19 21 24 21 24 28 25 24 30

Solution: Arrange the score values in ascending order of magnitude:

19 21 21 24 24 24 25 28 30

There are 9 values in the data set.

$$\text{Now, median} = \left(\frac{n+1}{2}\right)\text{th value}$$
$$= \left(\frac{9+1}{2}\right)\text{th value}$$
$$= \frac{10}{2}\text{th value}$$
$$= 5\text{th value}$$
$$= 24$$

$$\text{Lower quartile} = \frac{1}{4}(n+1)\text{th value}$$
$$= \frac{1}{4}(9+1)\text{th value}$$
$$= \frac{1}{4}(10)\text{th value}$$
$$= 2.5\text{th value}$$
$$= \frac{21+21}{2} \qquad \text{(Average of the 2nd and 3rd values)}$$
$$= \frac{42}{2}$$
$$= 21$$

$$\text{Upper quartile} = \frac{3}{4}(n+1)\text{th value}$$
$$= \frac{3}{4}(9+1)\text{th value}$$
$$= \frac{3}{4}(10)\text{th value}$$
$$= 7.5\text{th value}$$
$$= \frac{25+28}{2} \qquad \text{(Average of the 7th and 8th values)}$$
$$= \frac{53}{2}$$
$$= 26.5$$

$$\text{Interquartile range} = \text{Upper quartile} - \text{Lower quartile}$$
$$= 26.5 - 21$$
$$= 5.5$$

This means the middle 50% of the data values range from 21 to 26.5.

Interquartile Range (IQR) :

the IQR is the difference between the upper quartile (Quartile 3) and the lower quartile (Quartile 1), and by using the example above we find that the interquartile range for this dataset is

$$IQR = Q3 - Q1$$

Outliers :

Outliers are those values that don't seem to fit the rest of the dataset. To locate outliers, we need to find our "fences," or those numbers that enclose the data and indicate the acceptable range for our data set. If a number falls outside of the fence, then it is an outlier. We locate our fences by using our upper and lower quartiles and our interquartile range as follows:

Ex: { 8, 12, 6, 13, 15, 18, 14, 20, 11, 7, 15 },

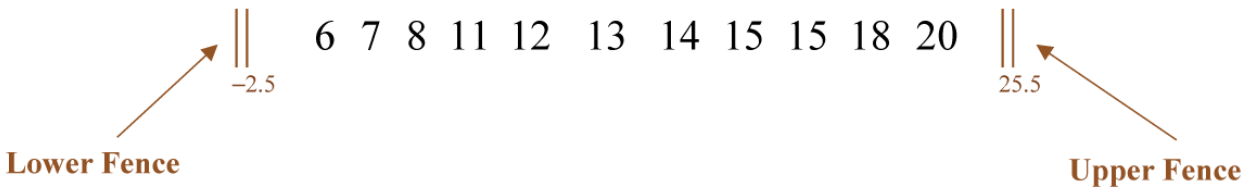
where we determined Q1 is 8, Q3 is 15, and the IQR is 7, then our lower and upper fences are:

$$Lower\ \ Fence: Q1 - 1.5IQR$$
$$Upper\ \ Fence: Q3 + 1.5IQR$$

$$6\ \ 7\ \ \underset{Q1}{8}\ \ 11\ \ 12\ \ \underset{Median}{13}\ \ 14\ \ 15\ \ \underset{Q3}{15}\ \ 18\ \ 20$$

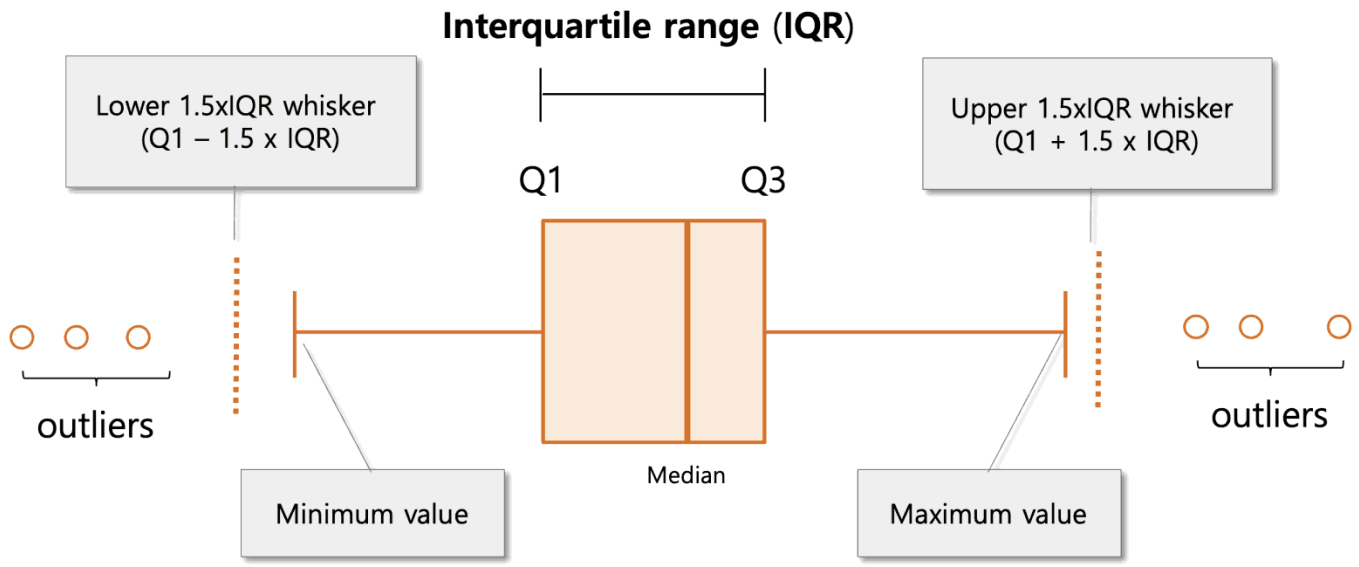
$$IQR = Q3 - Q1 = 15 - 8 = 7$$

$$Lower\ \ Fence: Q1 - 1.5IQR \rightarrow \quad 8 - 1.5(7) = -2.5$$
$$Upper\ \ Fence: Q3 + 1.5IQR \rightarrow \quad 15 + 1.5(7) = 25.5$$



As we can see, all of our data set falls within the fences, so we don't have any outliers!

Note : After the lower fence and higher fence all values are outliers.



Distributions

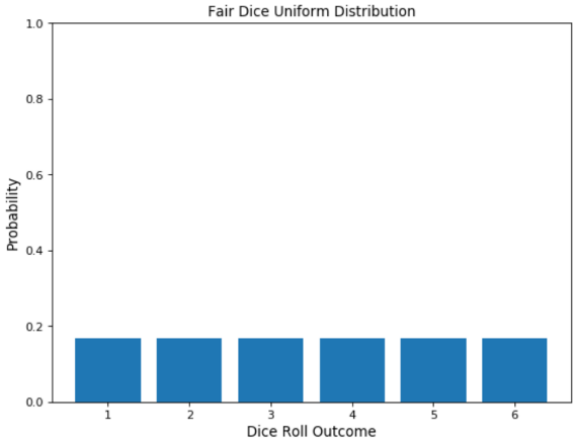
Depending on the type of data we use, we have grouped distributions into two categories, discrete distributions for discrete data (finite outcomes) and continuous distributions for continuous data (infinite outcomes).

Discrete distributions

1. ***Discrete uniform distribution: All outcomes are equally likely***

In statistics, uniform distribution refers to a statistical distribution in which all outcomes are equally likely. Consider rolling a six-sided die. You have an equal probability of obtaining all six numbers on your next roll, i.e., obtaining precisely one of 1, 2, 3, 4, 5, or 6, equaling a probability of 1/6, hence an example of a discrete uniform distribution.

As a result, the uniform distribution graph contains bars of equal height representing each outcome. In our example, the height is a probability of 1/6 (0.166667).

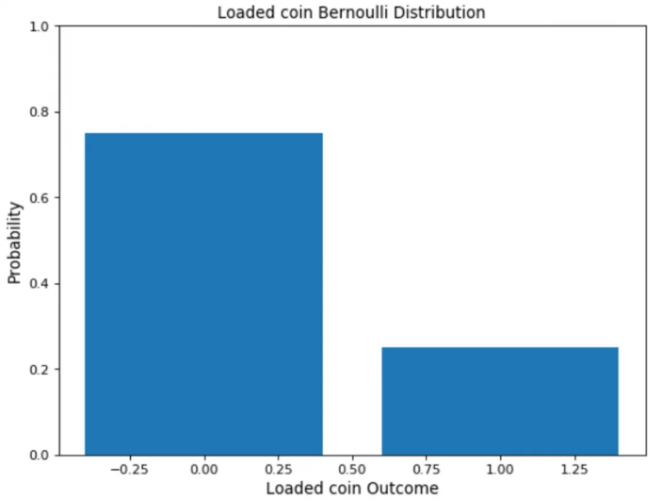


2. *Bernoulli Distribution: Single-Trial with Two Possible Outcomes*

The Bernoulli distribution is one of the easiest distributions to understand. It can be used as a starting point to derive more complex distributions. Any event with a single trial and only two possible outcomes follow a Bernoulli distribution. Flipping a coin or choosing between True and False in a quiz are examples of a Bernoulli distribution.

They have a single trial and only two outcomes. Let's assume you flip a coin once; this is a single trail. The only two possible outcomes are either heads or tails. This is an example of a Bernoulli distribution.

It is represented by bern(p), where p is the probability of success. The expected value of a Bernoulli trial 'x' is represented as, E(x) = p, and similarly Bernoulli variance is, Var(x) = p(1-p).



3. *Binomial Distribution: A sequence of Bernoulli events*

The Binomial Distribution can be thought of as the sum of outcomes of an event following a Bernoulli distribution. Therefore, Binomial Distribution is used in binary outcome events, and the probability of success and failure is the same in all successive trials.

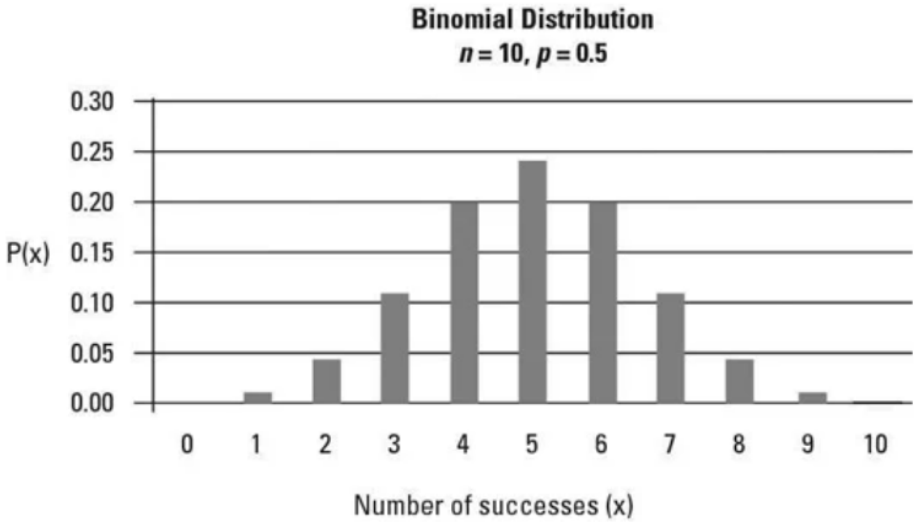
An example of a binomial event would be flipping a coin multiple times to count the number of heads and tails.

$$P\Big(X = x\Big) = \frac{n!}{x!(n-x)!}\, p^x \Big(1-p\Big)^{(n-x)}$$

Binomial vs Bernoulli distribution.

The difference between these distributions can be explained through an example. Consider you're attempting a quiz that contains 10 True/False questions. Trying a single T/F question would be considered a Bernoulli trial, whereas attempting the entire quiz of 10 T/F questions would be categorized as a Binomial trial. The main characteristics of Binomial Distribution are:

- Given multiple trials, each of them is independent of the other. That is, the outcome of one trial doesn't affect another one.
- Each trial can lead to just two possible results (e.g., winning or losing), with probabilities p and (1 – p).



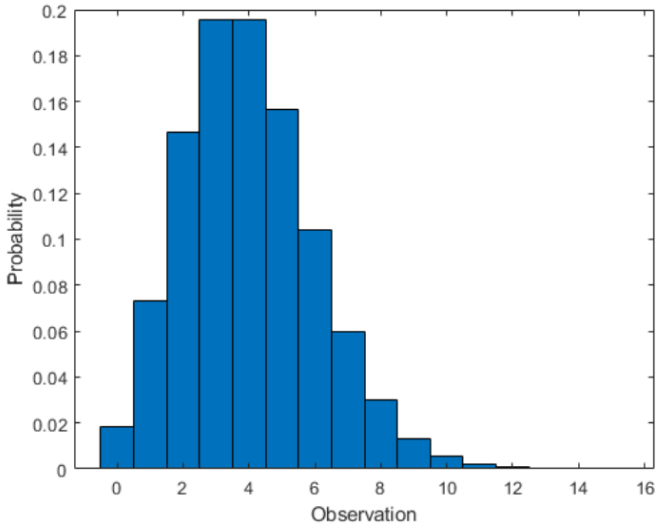
4. Poisson Distribution: The probability that an event May or May not occur

Poisson distribution deals with the frequency with which an event occurs within a specific interval. Instead of the probability of an event, Poisson distribution requires knowing how often it happens in a particular period or distance. For example, a cricket chirps two times in 7 seconds on average. We can use the Poisson distribution to determine the likelihood of it chirping five times in 15 seconds.

A Poisson process is represented with the notation Po(λ), where λ represents the expected number of events that can take place in a period. The expected value and variance of a Poisson process is λ. X represents the discrete random variable. A Poisson Distribution can be modeled using the following formula.

The main characteristics which describe the Poisson Processes are:

- The events are independent of each other.
- An event can occur any number of times (within the defined period).
- Two events can’t take place simultaneously.

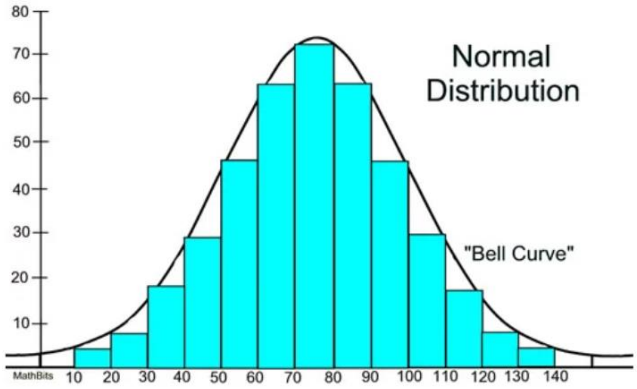


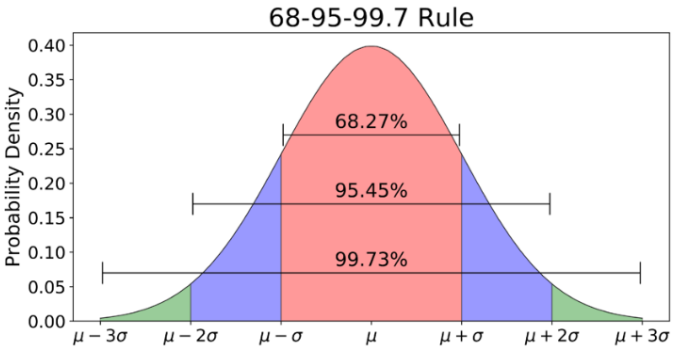
The graph of Poisson distribution plots the number of instances an event occurs in the standard interval of time and the probability of each one.

Continuous Distributions

5. Normal Distribution: Symmetric Distribution of Values Around the Mean

Normal distribution is the most used distribution in data science. In a normal distribution graph, data is symmetrically distributed with no skew. When plotted, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center.

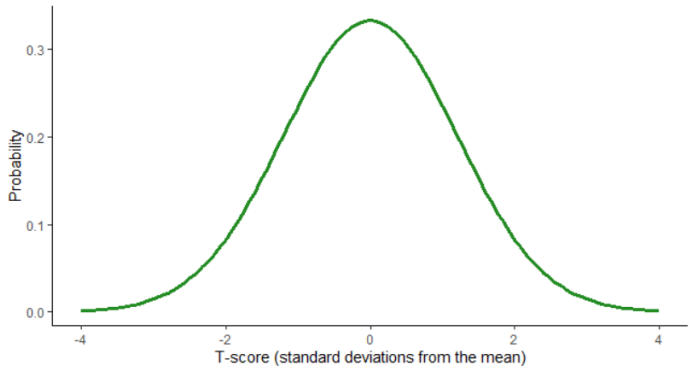




Probability Density and 68-95-99.7 Rule

6. Student t-Test Distribution: Small sample size approximation of a normal distribution

The student's t-distribution, also known as the t distribution, is a type of statistical distribution similar to the normal distribution with its bell shape but has heavier tails. The t distribution is used instead of the normal distribution when you have small sample sizes.



Student t-Test Distribution Curve

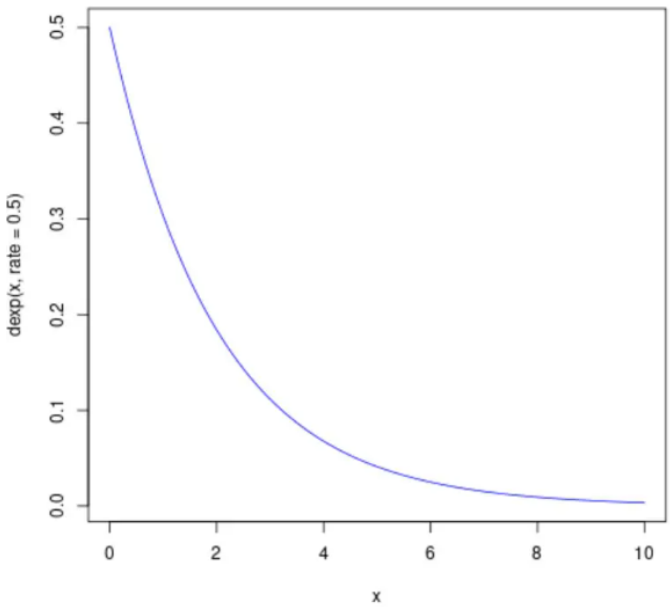
T-Distribution Table

df	a = 0.1	0.05	0.025
∞	t_a = 1.282	1.645	1.960
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182
4	1.533	2.132	2.776
5	1.476	2.015	2.571

Overall, the student t distribution is frequently used when conducting statistical analysis and plays a significant role in performing hypothesis testing with limited data.

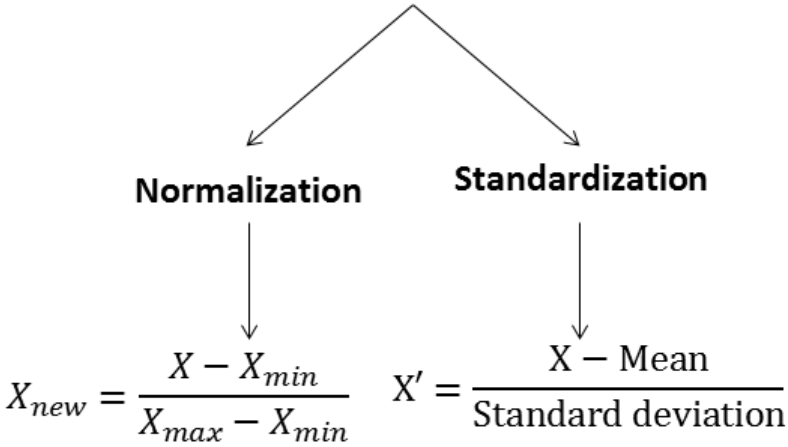
7. Exponential distribution: Model elapsed time between two events

Exponential distribution is one of the widely used continuous distributions. It is used to model the time taken between different events. For example, in physics, it is often used to measure radioactive decay; in engineering, to measure the time associated with receiving a defective part on an assembly line; and in finance, to measure the likelihood of the next default for a portfolio of financial assets. Another common application of Exponential distributions in survival analysis (e.g., expected life of a device/machine).



Exponential Distribution Curve

Feature Scaling



Normalization

Normalization is a feature scaling technique to bring the features in the data to a common range say [0, 1] or [-1, 0] or [-1, 1]. In this section, we'll go through 3 popular normalization methods as discussed below.

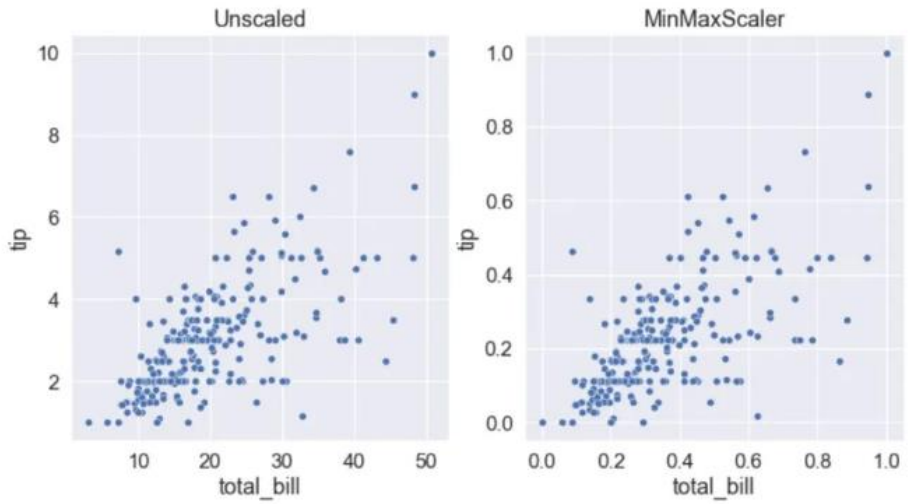
MinMaxScaler

This method scales each feature individually such that it is in the range [0,1]. Each feature value is subtracted with the min value and divided by the difference between max and min.
It uses the minimum and maximum values for scaling and both minimum & maximum are sensitive to outliers. As a result, the MinMaxScaler method is also sensitive to outliers. Note that MinMaxScaler doesn't change the distribution of the data.

$$X_i' = \frac{X_i - X_{\min}}{X_{max} - X_{min}}$$

You can use MinMaxScaler from Sklearn as shown below.

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

minmaxscaler_df = pd.DataFrame(scaler.fit_transform(df), columns=df.columns)
```



When to use MinMaxScaler?

- MinMaxScaler is preferred when the distribution of the features is unknown (i.e. if the features are not normally distributed).
- MinMaxScaler can also be considered if the underlying machine learning algorithms you are using don't make any assumptions about the distribution of the data (eg. kNN, Neural Nets, etc.).
- Consider using MinMaxScaler only if features have very few or no outliers.

By default, MinMaxScaler scales the data in the range [0,1]. However, you can modify this range as per your need by setting feature_range parameter.

Standardization

Standardization is the most commonly used feature scaling technique in machine learning. This is because some of the algorithms assume the normal or near-normal distribution of the data. If the features are normally distributed then the model behaves badly. The StandardScaler and standardization both refer to the same thing.

StandardScaler

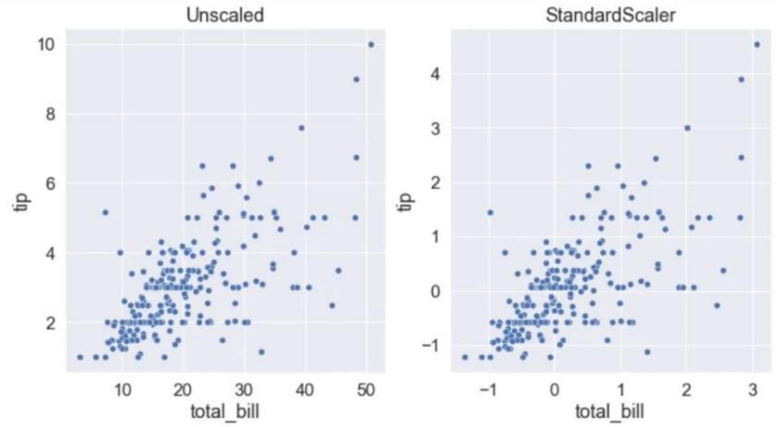
This method removes the mean and scales the data with unit variance (or standard deviation). The calculated mean and standard deviation are stored so that they can be used during the transformation of the test set. The scaling happens independently for each feature in the data.

$$X_i' = \frac{X_i - \mu}{\sigma} = \frac{X_i - X_{\text{mean}}}{X_{std}}$$

Scikit-learn implementation

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

standardscaler_df = pd.DataFrame(scaler.fit_transform(df), columns = df.columns)
```

The StandardScaler uses mean and mean is sensitive to outliers. Hence, outliers have an influence on the StandardScaler.



When to use StandardScaler?

- If the features are normally distributed then StandardScaler will be your first choice.
- Consider using StandardScaler if the underlying machine learning algorithms you are using make assumptions about the normal distribution of the data (eg. linear regression, logistic regression, etc.)
- If there are outliers in the data, then you can remove those outliers and use either MinMaxScaler/MaxAbsScaler/StandardScaler.

Summary

Normalization and Standardization are the two popular feature scaling techniques. The below table gives the summary of both methods.

Sl. No	Normalization	Standardization
1	Feature scaling method to bring the data into common range such as [0, 1], [-1, 1], etc.	Feature scaling method bring the data with mean 0 and unit variance
2	Scikit-learn provides MinMaxScaler, MaxAbsScaler and RobustScaler methods for normalization	Scikit-learn provides StandardScaler for standardization
3	MinMaxScaler and MaxAbsScaler are sensitive to outliers whereas RobustScaler is more robust to outliers	Standardization is less sensitive to outliers compared to MinMaxScaler and MaxAbsScaler
4	Useful when we don't know about the distribution of features and there are no or little outliers - MinMaxScaler: if features don't follow normal distriubtion and if there are no or less outliers - MaxAbsScaler: if the data is sparse - RobustScaler: if the data contains outliers	Useful when we know features are normally distributed (Gaussian distribution)

Normalization Vs Standardization

In [1]:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import MaxAbsScaler
from sklearn.preprocessing import RobustScaler
from sklearn.preprocessing import StandardScaler
sns.set(font_scale=1.5)
```

In [2]:

```
df = sns.load_dataset('tips')
df = df[['total_bill', 'tip']]
```

In [3]:

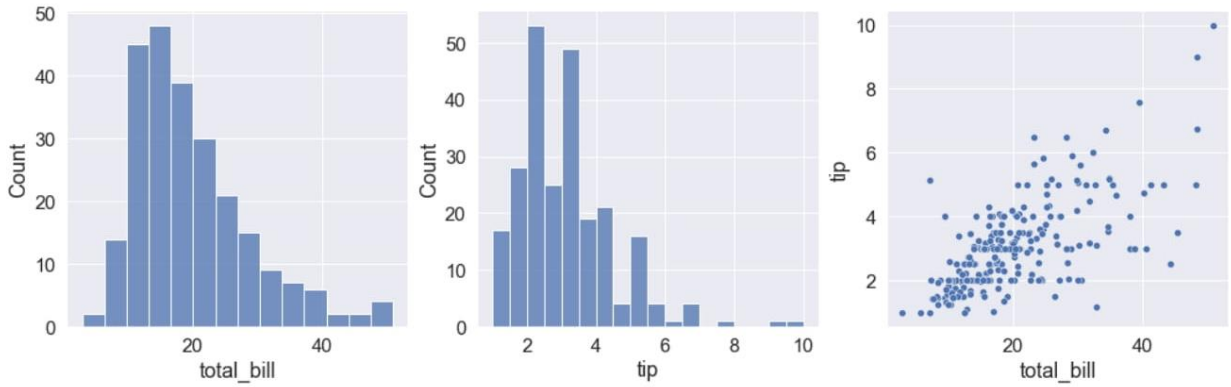
```
df.head()
```

Out[3]:

	total_bill	tip
0	16.99	1.01
1	10.34	1.66
2	21.01	3.50
3	23.68	3.31
4	24.59	3.61

In [4]:

```
fig, axes = plt.subplots(1, 3, figsize=(18, 5))
sns.histplot(data=df, x='total_bill', ax=axes[0])
sns.histplot(data=df, x='tip', ax=axes[1])
sns.scatterplot(data=df, x='total_bill', y='tip', ax=axes[2]);
```



MinMaxScaler

In [5]:

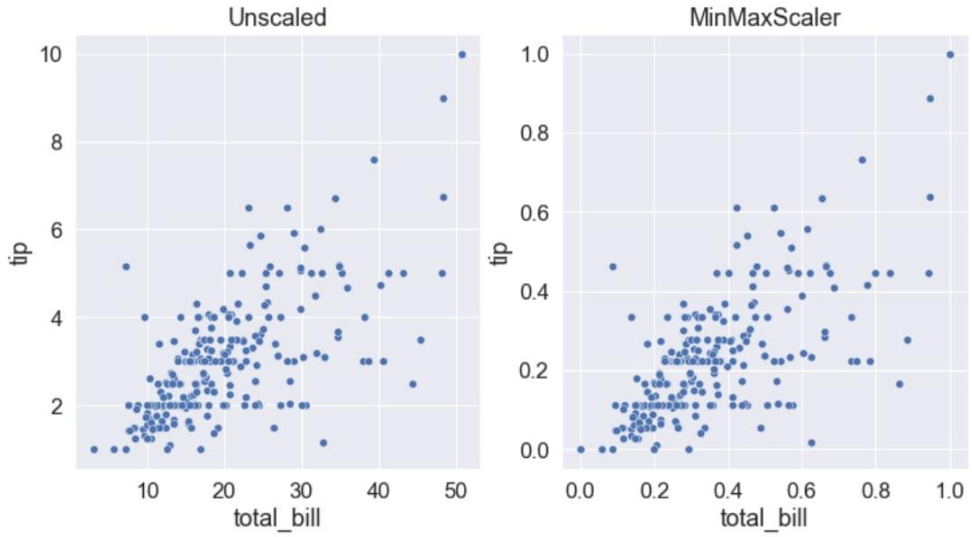
```
scaler = MinMaxScaler()
minmaxscaler_df = pd.DataFrame(scaler.fit_transform(df), columns = df.columns)
minmaxscaler_df.head()
```

Out[5]:

	total_bill	tip
0	0.291579	0.001111
1	0.152283	0.073333
2	0.375786	0.277778
3	0.431713	0.256667
4	0.450775	0.290000

In [6]:

```
fig, axes = plt.subplots(1, 2, figsize=(12, 6))

sns.scatterplot(data=df, x='total_bill', y='tip', ax=axes[0])
axes[0].set(title='Unscaled');

sns.scatterplot(data=minmaxscaler_df, x='total_bill', y='tip', ax=axes[1]);
axes[1].set(title='MinMaxScaler');
```



MaxAbsScaler

In [7]:

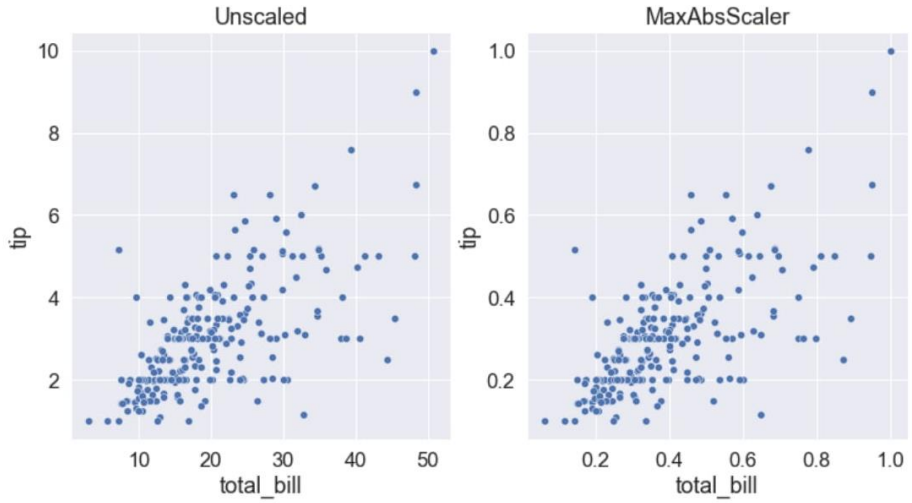
```
scaler = MaxAbsScaler()
maxabsscaler_df = pd.DataFrame(scaler.fit_transform(df), columns = df.columns)
maxabsscaler_df.head()
```

Out[7]:

	total_bill	tip
0	0.334383	0.101
1	0.203503	0.166
2	0.413501	0.350
3	0.466050	0.331
4	0.483960	0.361

In [8]:

```
fig, axes = plt.subplots(1, 2, figsize=(12, 6))

sns.scatterplot(data=df, x='total_bill', y='tip', ax=axes[0])
axes[0].set(title='Unscaled');

sns.scatterplot(data=maxabsscaler_df, x='total_bill', y='tip', ax=axes[1]);
axes[1].set(title='MaxAbsScaler');
```



RobustScaler

In [9]:

```
scaler = RobustScaler()
robustscaler_df = pd.DataFrame(scaler.fit_transform(df), columns = df.columns)
robustscaler_df.head()
```

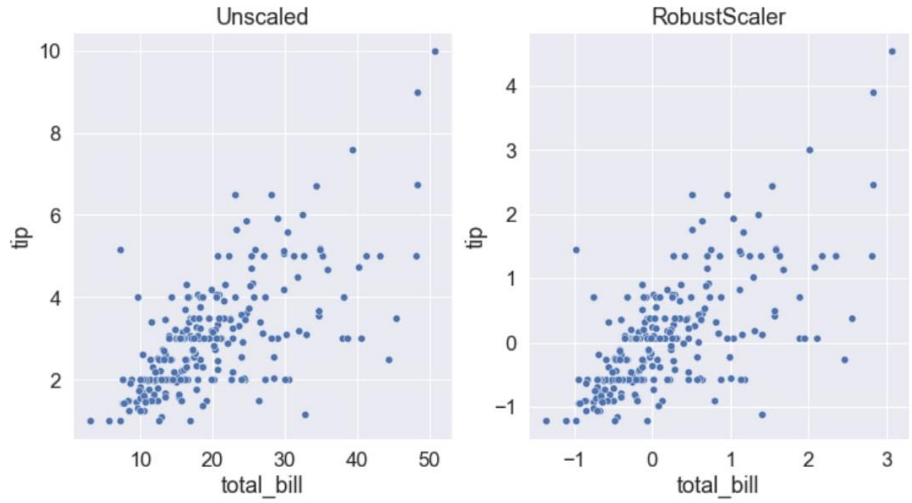
Out[9]:

	total_bill	tip
0	-0.074675	-1.2096
1	-0.691558	-0.7936
2	0.298237	0.3840
3	0.545918	0.2624
4	0.630334	0.4544

However, note that feature scaling is not mandatory for all the algorithms. The tree-based algorithms such as the Decision Tree algorithm, Random Forest algorithm, Gradient Boosted Trees, etc. don’t need feature scaling.

In [10]:

```
fig, axes = plt.subplots(1, 2, figsize=(12, 6))

sns.scatterplot(data=df, x='total_bill', y='tip', ax=axes[0])
axes[0].set(title='Unscaled');

sns.scatterplot(data=robustscaler_df, x='total_bill', y='tip', ax=axes[1]);
axes[1].set(title='RobustScaler');
```



StandardScaler

In [11]:

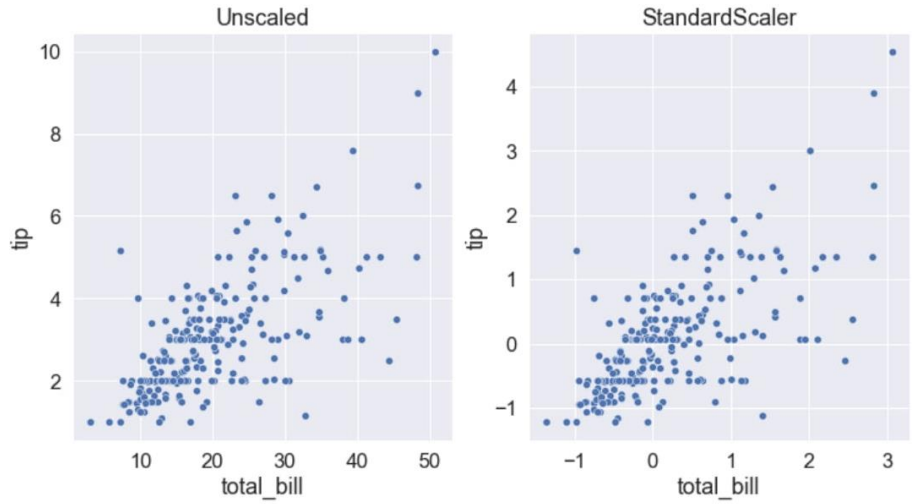
```
scaler = StandardScaler()
standardscaler_df = pd.DataFrame(scaler.fit_transform(df), columns = df.columns)
standardscaler_df.head()
```

Out[11]:

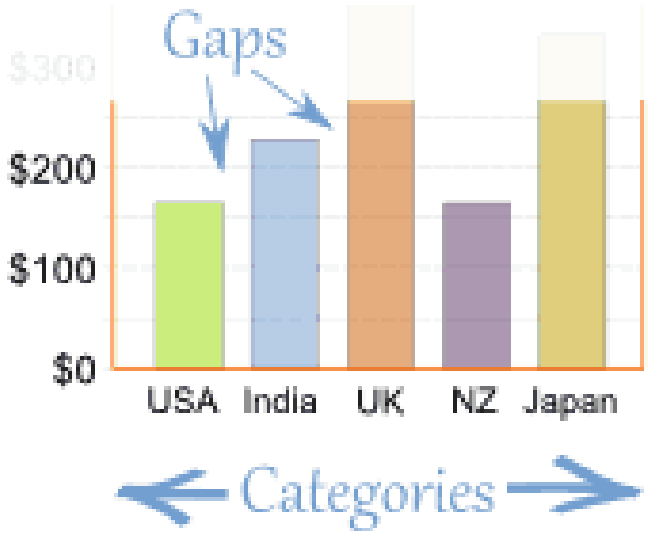
	total_bill	tip
0	-0.314711	-1.439947
1	-1.063235	-0.969205
2	0.137780	0.363356
3	0.438315	0.225754
4	0.540745	0.443020

In [12]:

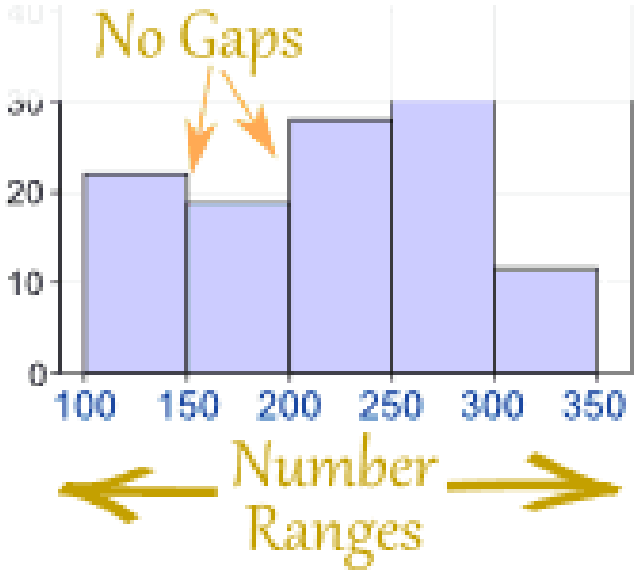
```
fig, axes = plt.subplots(1, 2, figsize=(12, 6))

sns.scatterplot(data=df, x='total_bill', y='tip', ax=axes[0])
axes[0].set(title='Unscaled');

sns.scatterplot(data=robustscaler_df, x='total_bill', y='tip', ax=axes[1]);
axes[1].set(title='StandardScaler');
```



Graphical Visualization



Bar Graph



Histogram

Composition

Pie chart

Stacked column chart

Stacked area chart

Waterfall charts

Comparison

Column chart

Bar chart

Line chart

Radar charts

Relationship

Scatter chart

Bubble chart

Distribution

Histogram chart

Scatter chart

Frequency Distribution

The **frequency** of a value is the number of times it occurs in a dataset. A **frequency distribution** is the pattern of frequencies of a variable. It's the number of times each possible value of a variable occurs in a dataset.

Types of frequency distributions

There are four types of frequency distributions:

- **Ungrouped frequency distributions:** The number of observations of each **value** of a variable.
  - You can use this type of frequency distribution for categorical variables.
- **Grouped frequency distributions:** The number of observations of each **class interval** of a variable. Class intervals are ordered groupings of a variable's values.
  - You can use this type of frequency distribution for quantitative variables.
- **Relative frequency distributions:** The proportion of observations of each value or class interval of a variable.
  - You can use this type of frequency distribution for **any type of variable** when you're more interested in **comparing frequencies** than the actual number of observations.
- **Cumulative frequency distributions:** The sum of the frequencies less than or equal to each value or class interval of a variable.
  - You can use this type of frequency distribution for **ordinal or quantitative variables** when you want to understand **how often observations fall below certain values**.

Example: Cumulative frequency distribution

Cumulative frequency table of the ages of survey participants

Age, *a* (years)	Frequency	Cumulative frequency	Cumulative relative frequency
$19 \leq a < 29$	4	4	4 / 20 = .2
$29 \leq a < 39$	9	9 + 4 = 13	.65
$39 \leq a < 49$	3	9 + 4 + 3 = 16	.8
$49 \leq a < 59$	3	19	.95
$59 \leq a < 69$	1	20	1

Example: Relative frequency distribution

Relative frequency table of the frequency of bird species at a bird feeder

Bird species	Frequency	Relative frequency
Chickadee	3	$= \frac{3}{(3 + 1 + 4 + 2 + 4 + 2)}$ $= \frac{3}{16}$ $= .19$
Dove	1	.06
Finch	4	.25
Grackle	2	.13
Sparrow	4	.25
Starling	2	.13

Inferential Statistics

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.

Inferential Statistics	
Hypothesis Testing	Regression Analysis
Z test	Linear Regression
F test	Nominal Regression
T test	Logistic Regression
ANOVA Test	Ordinal Regression
Wilcoxon Signed Rank Test	
Mann-Whitney U Test	

Hypothesis

In statistics, a hypothesis is a claim or statement about a property of a population. A hypothesis test is a standard procedure for testing a claim about a property of a population.

Terminology used for Hypothesis Testing
Null hypothesis (H0): It is denoted by; H0: $\mu 1 = \mu 2$, which shows that there is no difference between the two population means.

Alternative hypothesis (H1 or Ha): Contrary to the null hypothesis, the alternative hypothesis shows that observations are the result of a real effect.

Four Outcomes of hypothesis:

Outcome 1: We reject the null hypothesis, when in really it is **false**. ---> Yes

Outcome 2: We reject the null hypothesis, when in reality it is **true**. ----> Type 1 Error

Outcome 3: We accept the null hypothesis, when in reality it is **false**. ----> Type 2 Error

Outcome 4: We accept the null hypothesis, when in reality it is **true**. ----> Good

Components of a Formal Hypothesis Test

	2-Tailed Test	Right-Tailed	Left Tailed
Null hypothesis	$H_0: \mu = \mu_0$	$H_0: \mu \leq \mu_0$	$H_0: \mu \geq \mu_0$
Alternative hypothesis	$H_a: \mu \neq \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$

Null Hypothesis: H0 the null hypothesis (denoted by H0) is a statement that the value of a population parameter (such as proportion, mean, or standard deviation) is equal to some claimed value. We test the null hypothesis directly. Either reject H0 or fail to reject H (in other words, accept H0).Emphasize "equal to".

Alternative Hypothesis: H1
The alternative hypothesis (denoted by H1) is the statement that the parameter has a value that somehow differs from the null hypothesis. The symbolic form of the alternative hypothesis must use one of these symbols: ≠, <, > (not equal, less than, greater than) Give examples of different ways to word "not equal to," < and >, such as 'is different from', 'fewer than', 'more than', etc.

Example 1Claim: the XSORT method of gender selection increases the likelihood of having a baby girl. We express this claim in symbolic form: p>0.5 (here p denotes the proportion of baby girls) Null hypothesis must say "equal to", so H0: p=0.5 Alternative hypothesis must express difference:

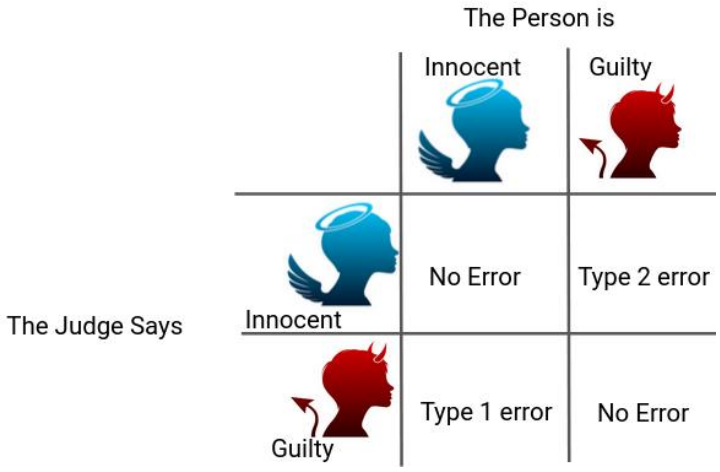
H1 : p>0.5 Original claim is now the alternative hypothesis

Let's take an example to understand the concept of Hypothesis Testing.

A person is on trial for a criminal offense and the judge needs to provide a verdict on his case.

Now, there are four possible combinations in such a case:

- First Case: The person is innocent and the judge identifies the person as innocent
- Second Case: The person is innocent and the judge identifies the person as guilty
- Third Case: The person is guilty and the judge identifies the person as innocent
- Fourth Case: The person is guilty and the judge identifies the person as guilty

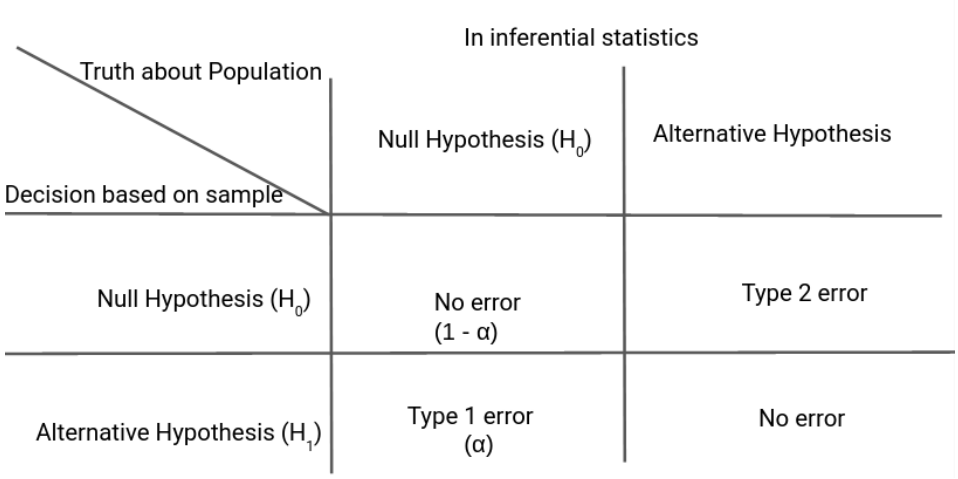


As you can clearly see, there can be two types of error in the judgment – Type 1 error, when the verdict is against the person while he was innocent and Type 2 error, when the verdict is in favor of Person while he was guilty

According to the Presumption of Innocence, the person is considered innocent until proven guilty. That means the judge must find the evidence which convinces him “beyond a reasonable doubt”. This phenomenon of **“Beyond a reasonable doubt”** can be understood as **Probability (Judge Decided Guilty | Person is Innocent) should be small.**

The basic concepts of Hypothesis Testing are actually quite analogous to this situation.

We consider **the Null Hypothesis** to be true until we find strong evidence against it. Then. we accept the **Alternate Hypothesis**. We also determine the **Significance Level (α)** which can be understood as the Probability of (Judge Decided Guilty | Person is Innocent) in the previous example. Thus, if α is smaller, it will require more evidence to reject the Null Hypothesis. Don’t worry, we’ll cover all of this using a case study later.



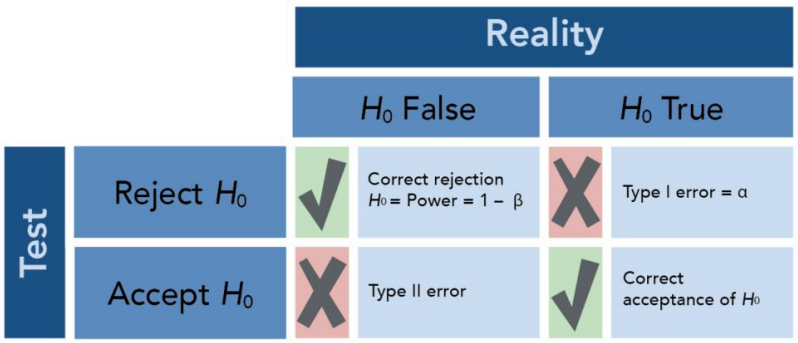
General Rules

If the null hypothesis is rejected, the alternative hypothesis is accepted. If the null hypothesis is accepted, the alternative hypothesis is rejected. Acceptance or rejection of the null hypothesis is an initial conclusion. Always state the final conclusion expressed in terms of the original claim, not in terms of the null hypothesis or the alternative hypothesis.

Type I Error: A Type I error is the mistake of rejecting the null hypothesis when it is actually true. The symbol α (alpha) is used to represent the probability of a type I error.

Type II Error: A Type II error is the mistake of accepting the null hypothesis when it is actually false. The symbol β (beta) is used to represent the probability of a type II error.

Type I and Type II Errors



ExampleClaim: a new medicine has a greater success rate, p>p0, than the old (existing) one. Null hypothesis: H0 : p=p0Alternative hypothesis: H1 : p>p (agrees with the original claim)Page 398 of Elementary Statistics, 10th Edition

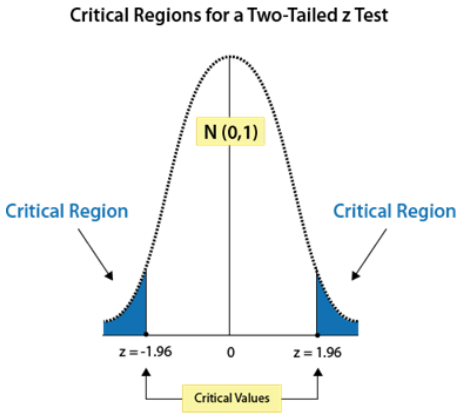
Example (continued)

Type I error: the null hypothesis is true, but we reject it => we accept the claim, hence we adopt the new (inefficient, potentially harmful) medicine. This is a critical error, should be avoided!

Type II error: the alternative hypothesis is true, but we reject it => we reject the claim, hence we decline the new medicine and continue using the old one (no harm…).Page 398 of Elementary Statistics, 10th Edition

Significance Level The probability of the type I error (denoted by α) is also called the significance level of the test. It characterizes the chances that the test fails (i.e., type I error occurs) It must be a small number. Typical values used in practice: α = 0.1, 0.05, or 0.01 (in percents, 10%, 5%, or 1%).

Compute the test statistic:



Draw the diagram (the normal curve)

On the diagram, mark a region of extreme values that agree with the alternative hypothesis:Sample proportion of:orTest Statistic z = 3.21

Critical Region The critical region (or rejection region) is the set of all values of the test statistic that cause us to reject the null hypothesis.

Critical Value A critical value is a value that separates the critical region (where we reject the null hypothesis) from the values of the test statistic that do not lead to rejection of the null hypothesis. See the previous figure where the critical value is z = It corresponds to a significance level of α = 0.05.

Significance Level The significance level (denoted by α) is the probability that the test statistic will fall in the critical region (when the null hypothesis is actually true).

Types of Hypothesis Tests:

Two-tailed, Left-tailed, Right-tailed
The tails in a distribution are the extreme regions where values of the test statistic agree with the alternative hypothesis.

Right-tailed Test H0: p=0.5 H1: p>0.5 α is in the right tail
Points Right

Critical value for a right-tailed test
A right-tailed test requires one (positive) critical value: za

Left-tailed Test H0: p=0.5 H1: p<0.5 α is in the left tail
Points Left

Critical value for a left-tailed test
A left-tailed test requires one (negative) critical value: –za

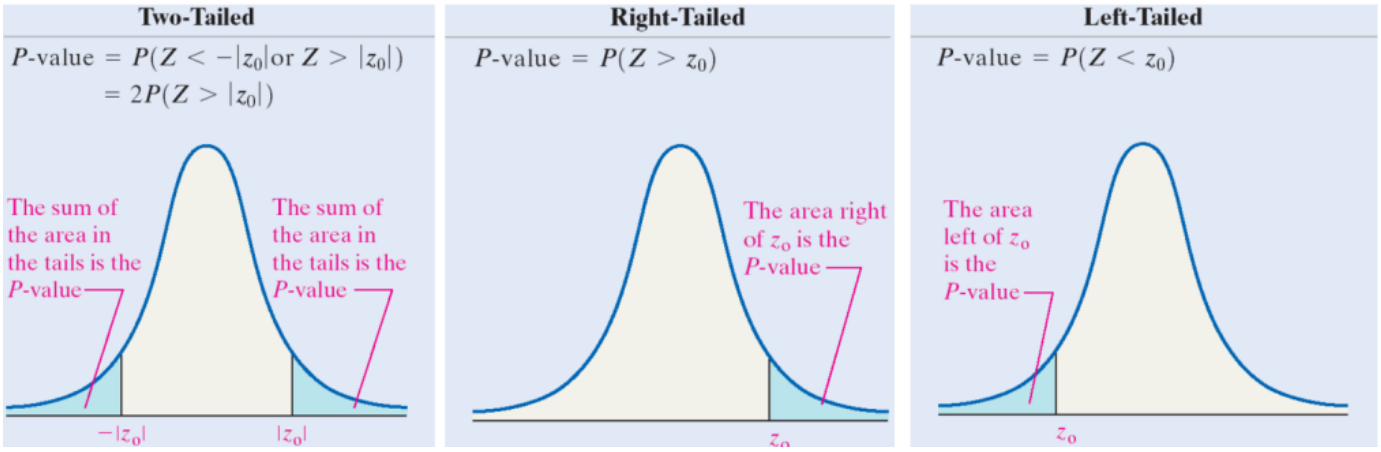
α is divided equally between the two tails of the critical
Two-tailed TestH0: p=0.5H1: p≠0.5α is divided equally between the two tails of the critical region
Means less than or greater than

Critical values for a two-tailed test
A two-tailed test requires two critical values: za/2 and –za/2

P-Value Approach

Assume that the null hypothesis is true.
The P-Value is the probability of observing a sample mean that is as or more extreme than the observed.
How to compute the P-Value for each type of test:
Step 1: Compute the test statistic $z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

Two-tail | Right Tail | Left Tail



P-Value: The P-value (or p-value or probability value) is the probability of getting a value of the test statistic that is at least as extreme as the one representing the sample data, assuming that the null hypothesis is true.

Example 1 (continued) P-value is the area to the right of the test statistic z = 3.21.We refer to Table A-2 (or use calculator) to find that the area to the right of z = 3.21 isP.

P-Value method: If P-value ≤ α , reject H0.
If P-value > α , fail to reject H0.If the P is low, the null must go.If the P is high, the null will fly.

Hence the null hypothesis must be rejected
Example 1 (continued)P-value =It is smaller than α = 0.05.Hence the null hypothesis must be rejected.

Hypothesis Testing: One Population

Hypothesis Tests for the Population Mean

Null Hypothesis (Ho)	Alternative Hypothesis (Ha)	Test Statistic	Rejection Region
Case 1: σ^2 is known $\mu = \mu_o$	$\mu < \mu_o$ $\mu > \mu_o$ $\mu \neq \mu_o$	$Z = \frac{\bar{X} - \mu_o}{\sigma/\sqrt{n}}$	$z < -z_\alpha$ $z > z_\alpha$ $\lvert z \rvert > z_{\frac{\alpha}{2}}$
Case 2: σ^2 is unknown and n ≤ 30 $\mu = \mu_o$	$\mu < \mu_o$ $\mu > \mu_o$ $\mu \neq \mu_o$	$T = \frac{\bar{X} - \mu_o}{s/\sqrt{n}}$	$t < -t_{\alpha, n-1}$ $t > t_{\alpha, n-1}$ $\lvert t \rvert > t_{\frac{\alpha}{2}, n-1}$
Case 3: σ^2 is unknown and n > 30 $\mu = \mu_o$	$\mu < \mu_o$ $\mu > \mu_o$ $\mu \neq \mu_o$	$Z = \frac{\bar{X} - \mu_o}{s/\sqrt{n}}$	$z < -z_\alpha$ $z > z_\alpha$ $\lvert z \rvert > z_{\frac{\alpha}{2}}$

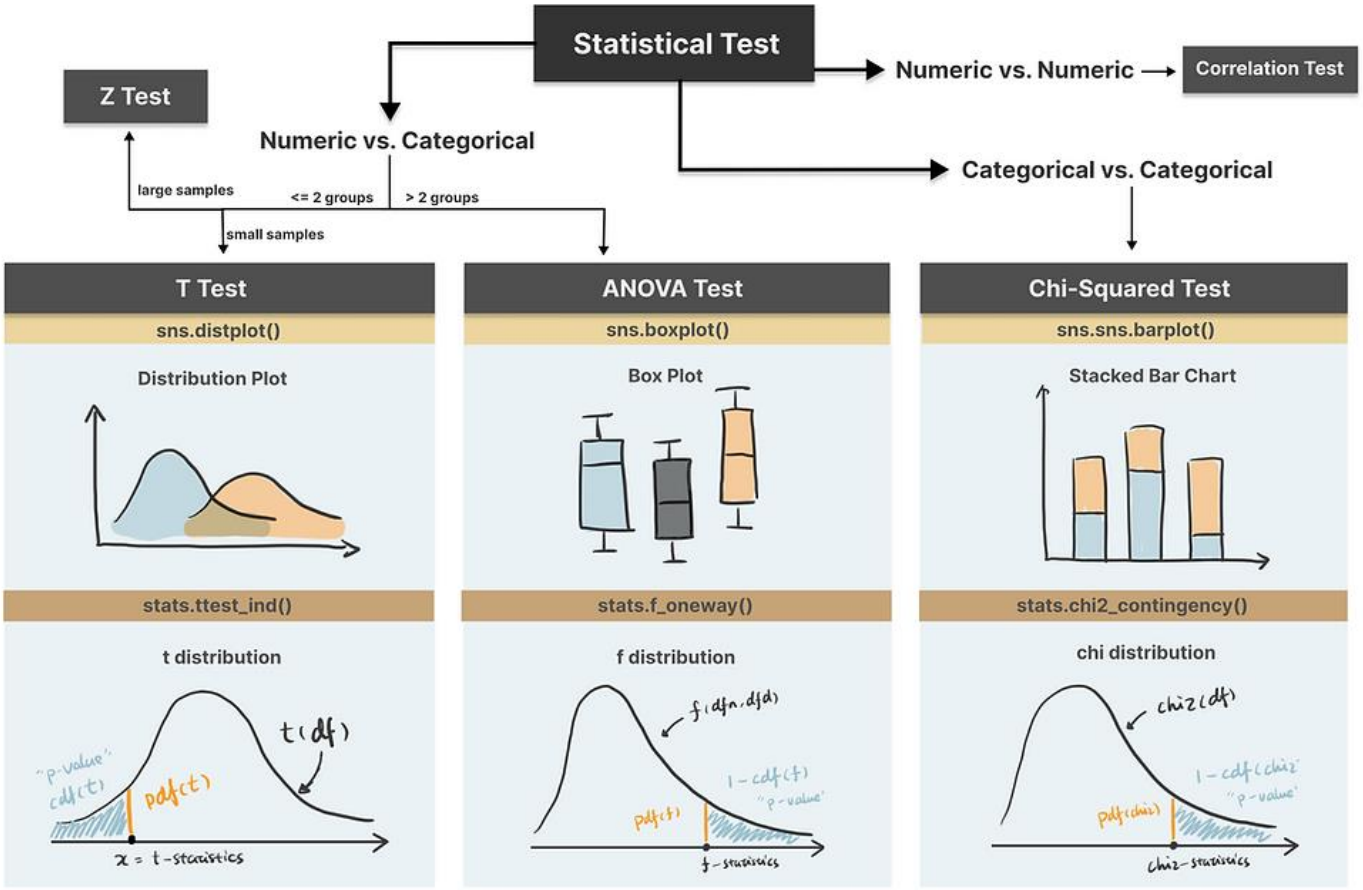
Hypothesis Tests for the Population Proportion

Null Hypothesis (Ho)	Alternative Hypothesis (Ha)	Test Statistic	Rejection Region
$P = P_o$	$P < P_o$ $P > P_o$ $P \neq P_o$	$Z = \frac{Y - np_o}{\sqrt{np_o(1-p_o)}}$ *where Y is the number of successes in a random sample of size n*	$z < -z_\alpha$ $z > z_\alpha$ $\lvert z \rvert > z_{\frac{\alpha}{2}}$

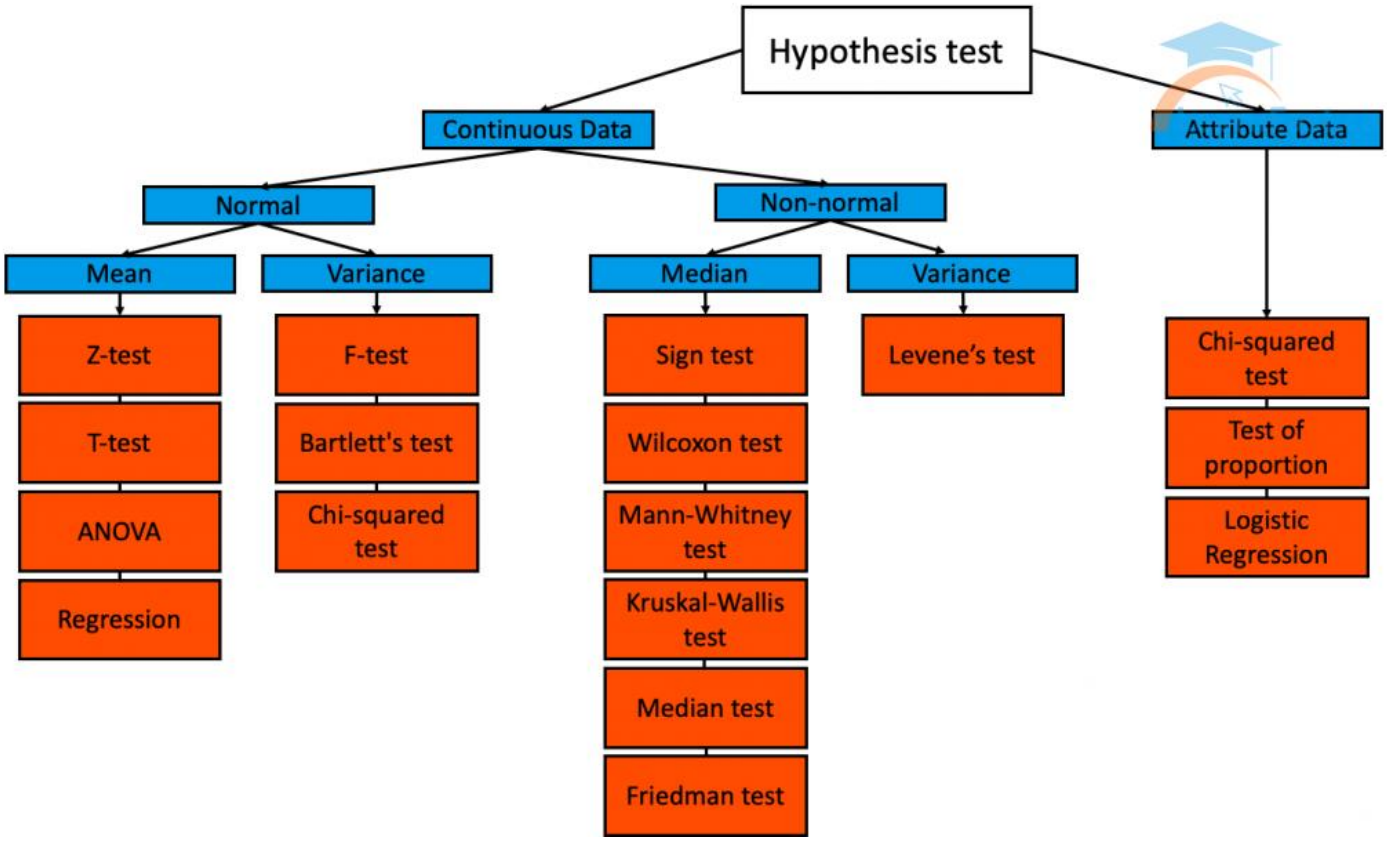
Hypothesis Testing: Two Populations

Hypothesis Tests for the Difference of Means (Independent Samples)

Null Hypothesis (Ho)	Alternative Hypothesis (Ha)	Test Statistic	Rejection Region
Case 1: σ_X^2 and σ_Y^2 are known $\mu_X - \mu_Y = d_o$	$\mu_X - \mu_Y < d_o$ $\mu_X - \mu_Y > d_o$ $\mu_X - \mu_Y \neq d_o$	$Z = \frac{(\bar{X} - \bar{Y}) - d_o}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}$	$z < -z_\alpha$ $z > z_\alpha$ $\lvert z \rvert > z_{\frac{\alpha}{2}}$



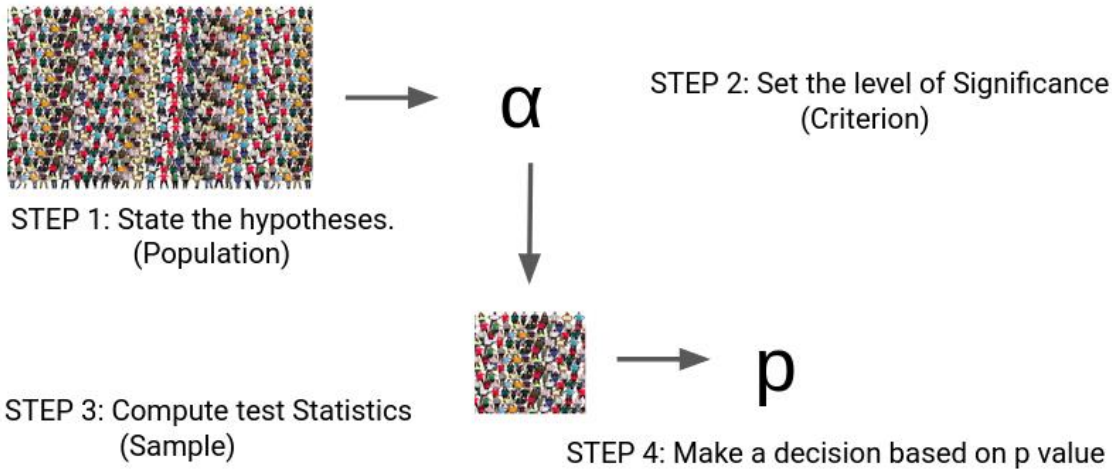
visit www.visual-design.net for step by step guide



Steps to Perform Hypothesis testing

There are four steps to perform Hypothesis Testing:

- Set the Hypothesis
- Set the Significance Level, Criteria for a decision
- Compute the test statistics
- Make a decision



Steps 1 to 3 are quite self-explanatory but on what basis can we make a decision in step 4?

What does this p-value indicate?

We can understand this p-value as the measurement of the Defense Attorney's argument. If the p-value is less than α , we reject the Null Hypothesis or if the p-value is greater than α, we fail to reject the Null Hypothesis.

Critical Value, p-value

Let's understand the logic of Hypothesis Testing with the graphical representation for Normal Distribution.

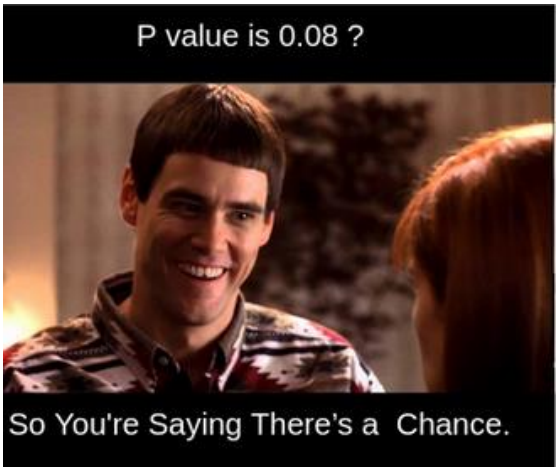
Typically, we set the Significance level at 10%, 5%, or 1%. If our test score lies in the Acceptance Zone we fail to reject the Null Hypothesis. If our test score lies in the critical zone, we reject the Null Hypothesis and accept the Alternate Hypothesis.

Critical Value is the cut off value between Acceptance Zone and Rejection Zone. We compare our test score to the critical value and if the test score is greater than the critical value, that means our test score lies in the Rejection Zone and we reject the Null Hypothesis.

On the opposite side, if the test score is less than the Critical Value, that means the test score lies in the Acceptance Zone and we fail to reject the null Hypothesis.

But why do we need p-value when we can reject/accept hypotheses based on test scores and critical value?

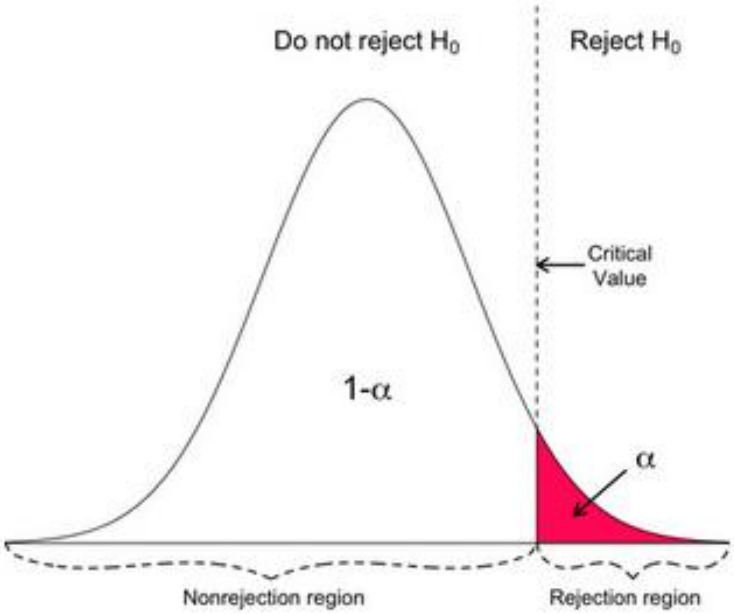
p-value has the benefit that we **only need one value** to make a decision about the hypothesis. We don't need to compute two different values like critical value and test scores. Another benefit of using p-value is that we can test at **any desired level of significance** by comparing this directly with the significance level.



This way we don't need to compute test scores and critical value for each significance level. We can get the p-value and directly compare it with the significance level.

Directional Hypothesis

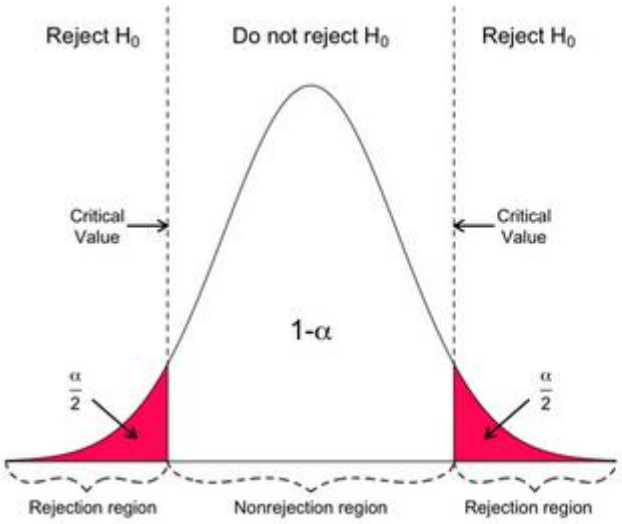
In the Directional Hypothesis, the null hypothesis is rejected if the test score is too large (for right-tailed and too small for left tailed). Thus, the rejection region for such a test consists of one part, which is right from the center.



Ex. Females will score significantly higher than Males on IQ score

Non-Directional Hypothesis

In a Non-Directional Hypothesis test, the Null Hypothesis is rejected if the test score is either too small or too large. Thus, the rejection region for such a test consists of two parts: one on the left and one on the right.



Ex. Females will score significantly different than Males on IQ score

Z-Test

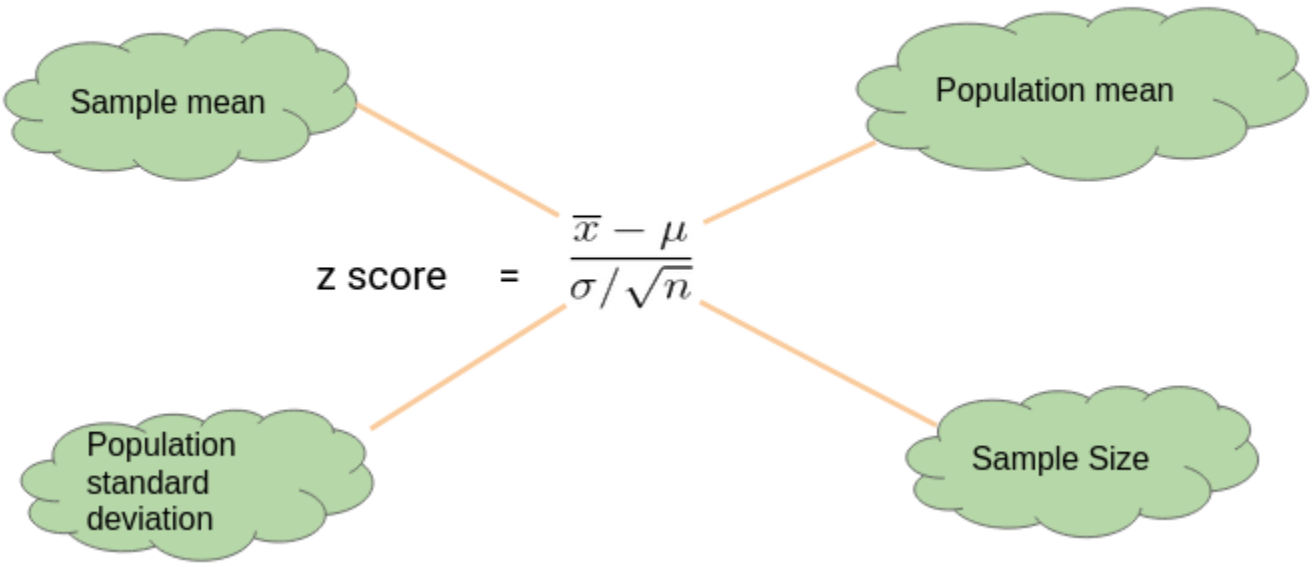
Z- tests are a statistical way of testing a hypothesis when either:

- We know the population variance, or
- We do not know the population variance but our sample size is large n ≥ 30

If we have a sample size of less than 30 and do not know the population variance, then we must use a t-test.

<u>One-Sample Z test</u>

We perform the One-Sample Z test when we want to compare **a sample mean with the population mean**



.

Here's an Example to Understand a One Sample Z Test

Let's say we need to determine if girls on average score higher than 600 in the exam. We have the information that the standard deviation for girls' scores is 100. So, we collect the data of 20 girls by using random samples and record their marks. Finally, we also set our α value (significance level) to be 0.05.

Score	650	730	510	670	480	800	690	530	590	620	710	670	640	780	650	490	800	600	510	700

In this example:

- Mean Score for Girls is 641
- The size of the sample is 20
- The population mean is 600
- Standard Deviation for Population is 100

$$\text{z score} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{641 - 600}{100/\sqrt{20}}$$

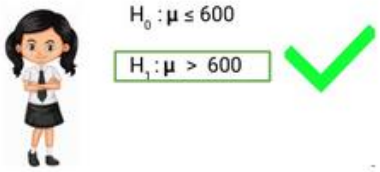
$$= 1.8336$$

p value = .033357.

Critical Value = 1.645

Z score > Critical Value

P value < 0.05



Since the P-value is less than 0.05, we can reject the null hypothesis and conclude based on our result that Girls on average scored higher than 600.

<u>Two Sample Z Test</u>

We perform a Two Sample Z test when we want to compare **the mean of two samples.**

Difference bw Sample mean $\bar{x}_1 - \bar{x}_2$

Difference bw population mean $\mu_1 - \mu_2$

$$\text{z score} = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Population standard deviation σ_1, σ_2

Sample Size n_1, n_2

Here's an Example to Understand a Two Sample Z Test

Here, let's say we want to know if Girls on average score 10 marks more than the boys. We have the information that the standard deviation for girls' Score is 100 and for boys' score is 90. Then we collect the data of 20 girls and 20 boys by using random samples and record their marks. Finally, we also set our α value (significance level) to be 0.05.

Score
650
730
510
670
480
800
690
530
590
620
710
670
640
780
650
490
800
600
510
700

Score
630
720
462
631
440
783
673
519
543
579
677
649
632
768
615
463
781
563
488
650

In this example:

- Mean Score for Girls (Sample Mean) is 641
- Mean Score for Boys (Sample Mean) is 613.3
- Standard Deviation for the Population of Girls' is 100
- Standard deviation for the Population of Boys' is 90
- Sample Size is 20 for both Girls and Boys
- Difference between Mean of Population is 10

$$\text{z score} = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{(641 - 613.3) - (10)}{\sqrt{\frac{100^2}{20} + \frac{90^2}{20}}}$$

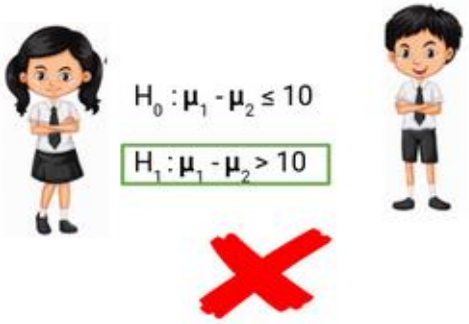
$$= 0.588$$

P value = 0.278

Critical Value = 1.645

Z score < Critical Value

P value > 0.05



Thus, we can **conclude based on the P-value that we fail to reject the Null Hypothesis**. We don't have enough evidence to conclude that girls on average score of 10 marks more than the boys.

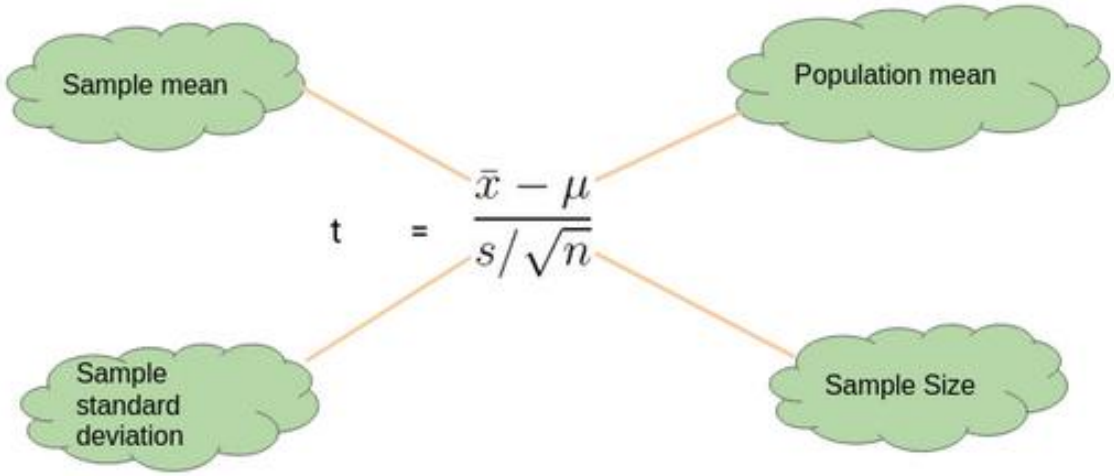
T-Test

t-tests are a statistical way of testing a hypothesis when:

- We do not know the population variance
- Our sample size is small, n < 30

<u>One-Sample t-Test</u>

We perform a One-Sample t-test when we want to **compare a sample mean with the population mean**. The difference from the Z Test is that we do **not have the information on Population Variance** here. We use the **sample standard deviation** instead of population standard deviation in this case.



Here's an Example to Understand a One Sample t-Test

Let's say we want to determine if on average girls score more than 600 in the exam. We do not have the information related to variance (or standard deviation) for girls' scores. To a perform t-test, we randomly collect the data of 10 girls with their marks and choose our α value (significance level) to be 0.05 for Hypothesis Testing.

Girls_Score
587
602
627
610
619
622
605
608
596
592

In this example:

- Mean Score for Girls is 606.8
- The size of the sample is 10
- The population mean is 600
- Standard Deviation for the sample is 13.14

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

$$= \frac{606.8 - 600}{13.14/\sqrt{10}}$$

$$= 1.64$$

$H_0 : \mu \leq 600$

$H_1 : \mu > 600$

Critical Value = 1.833

t score < Critical Value

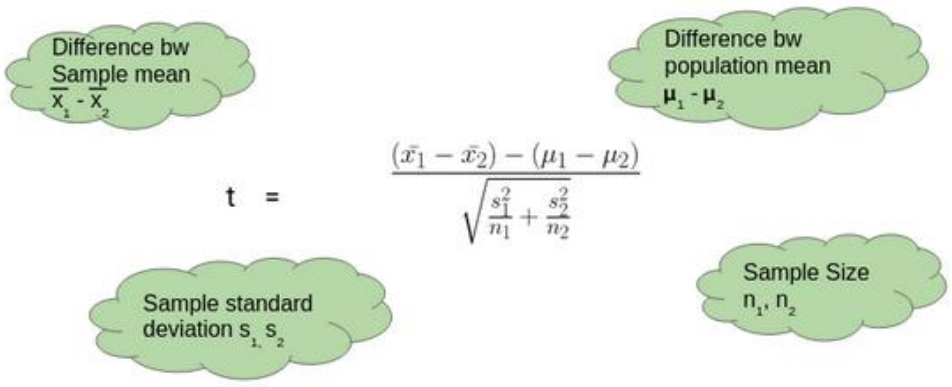
P value = 0.0678

P value > 0.05

Our **P-value is greater than 0.05 thus we fail to reject the null hypothesis** and don't have enough evidence to support the hypothesis that on average, girls score more than 600 in the exam.

Two-Sample t-Test

We perform a Two-Sample t-test when we want to compare the mean of two samples.



Here's an Example to Understand a Two-Sample t-Test

Here, let's say we want to determine if on average, boys score 15 marks more than girls in the exam. We do not have the information related to variance (or standard deviation) for girls' scores or boys' scores. To perform a t-test. we randomly collect the data of 10 girls and boys with their marks. We choose our α value (significance level) to be 0.05 as the criteria for Hypothesis Testing.

Girls_Score
587
602
627
610
619
622
605
608
596
592

Boys_Score
626
643
647
634
630
649
625
623
617
607

In this example:

- Mean Score for Boys is 630.1
- Mean Score for Girls is 606.8
- Difference between Population Mean 15
- Standard Deviation for Boys' score is 13.42
- Standard Deviation for Girls' score is 13.14

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\frac{(630.1 - 606.8) - (15)}{\sqrt{\frac{(13.42)^2}{10} + \frac{(13.14)^2}{10}}}$$

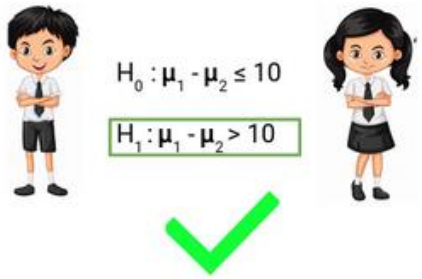
Critical Value = 1.833

t = 2.23

P value = 0.019

Critical Value > t score

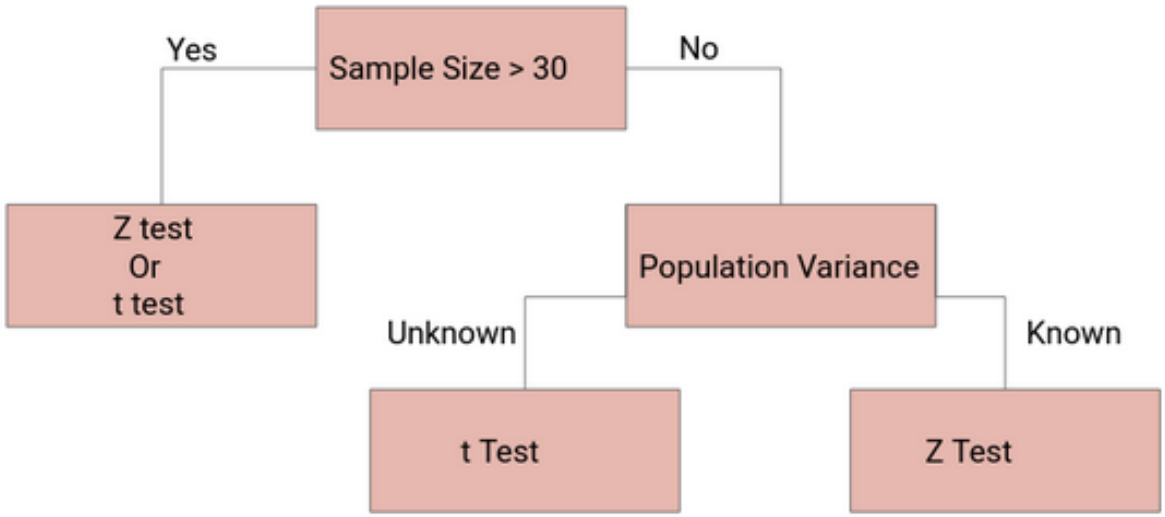
P value < 0.05



Thus, **P-value is less than 0.05 so we can reject the null hypothesis** and conclude that on average boys score 15 marks more than girls in the exam.

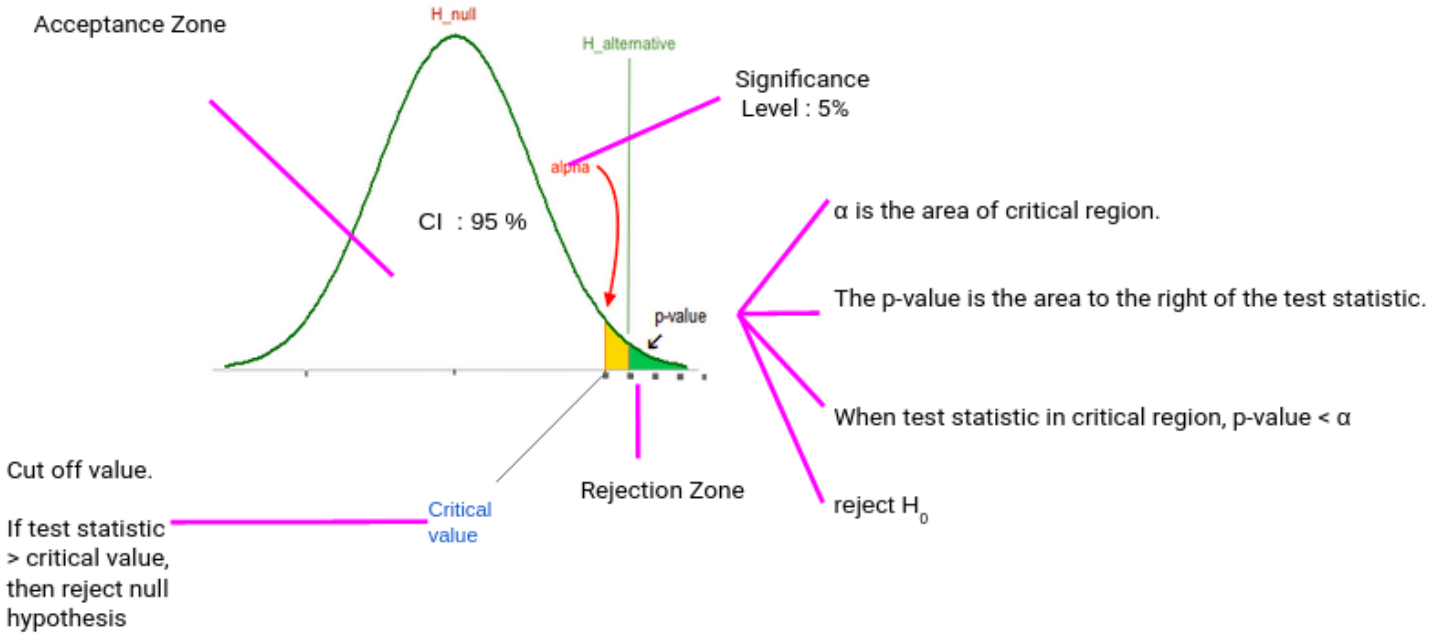
Deciding between Z Test and T-Test

So when we should perform the Z test and when we should perform t-Test? It's a key question we need to answer if we want to master statistics.



If the sample size is large enough, then the Z test and t-Test will conclude with the same results. For a **large sample size**, **Sample Variance will be a better estimate** of Population variance so even if population variance is unknown, we can **use the Z test using sample variance.**

Similarly, for a **Large Sample**, we have a high degree of freedom. And since t-**distribution approaches the normal distribution**, **t**he difference between the z score and t score is negligible.



Test statistic	Associated test	Sample size	Information given	Distribution	Test question
z-score	z-test	Two populations or large samples (n > 30)	• Standard deviation of the population (this will be given as σ) • Population mean or proportion	Normal	Do these two populations differ?
t-statistic	t-test	Two small samples (n < 30)	• Standard deviation of the sample (this will be given as *s*) • Sample mean	Normal	Do these two samples differ?
f-statistic	ANOVA	Three or more samples	• Group sizes • Group means • Group standard deviations	Normal	Do any of these three or more samples differ from each other?
chi-squared	chi-squared test	Two samples	• Number of observations for each categorical variable	Any	Are these two categorical variables independent?

F-Test

F test is to find out whether the two independent estimates of population variance differ significantly. In this case F ratio is

$$F = \frac{\sigma_1^2}{\sigma_2^2}$$

$$where\ \sigma^2 = \frac{\sum(x-\mu)^2}{n}$$

or

To find out whether the two samples drawn from the normal population having the same variance. In this case F ratio is

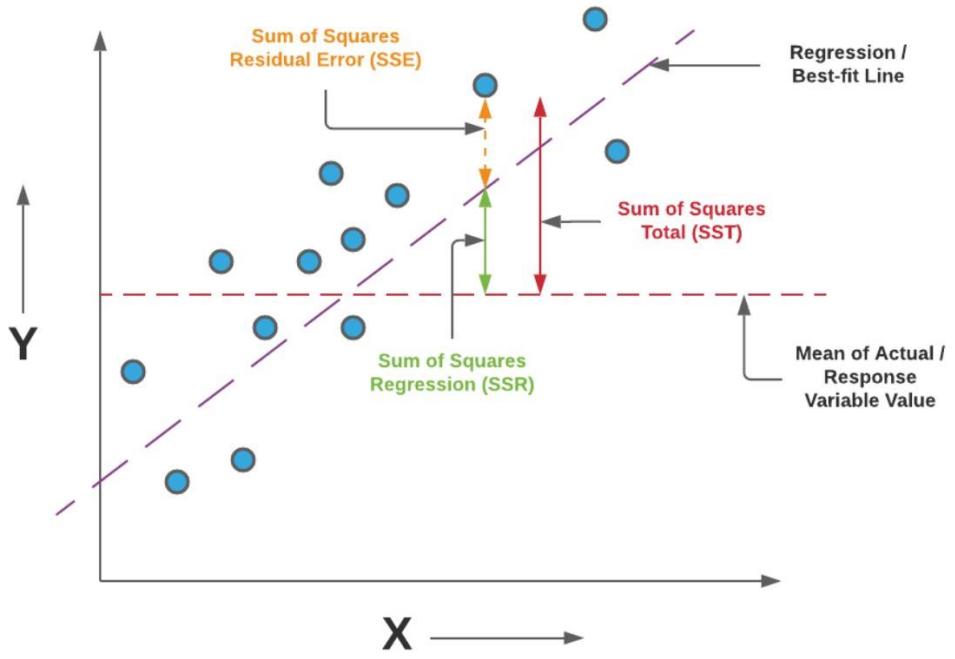
$$F = \frac{s_1^2}{s_2^2}$$

$$where\ s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

In both the cases $\sigma_1^2 > \sigma_2^2$, $S_1^2 > S_2^2$ in other words larger estimate of variance always be in numerator and smaller estimate of variance in denominator

Degrees of freedom (ϑ)

- DF of larger variance (i.e numerator) $=n_1-1$
- DF of smaller variance (i.e denominator) $=n_2-1$



What is an F Statistic?

F statistic also known as F value is used in ANOVA and regression analysis to identify the means between two populations are significantly different or not. In other words F statistic is ratio of two variances (Variance is nothing but measure of dispersion, it tells how far the data is dispersed from the mean). F statistic accounts corresponding degrees of freedom to estimate the population variance.

F statistic is almost similar to t statistic. t-test states a single variable is statistically significant or not whereas F test states a group of variables are statistically significant or not.

F statistics are based on the ratio of mean squares. F statistic is the ratio of the mean square for treatment or between groups with the Mean Square for error or within groups.

F = MS Between / MS Within

Distribution	H_a	Rejection Region
Right-tailed	$Ha: \sigma^2_1 > \sigma^2_2$	$F \geq F_\alpha$
Left-tailed	$Ha: \sigma^2_1 < \sigma^2_2$	$F \leq F_{1-\alpha}$
Two-tailed	$Ha: \sigma^2_1 \neq \sigma^2_2$	$F \leq F_{1-\alpha/2}$ or $F \geq F_{\alpha/2}$

If calculated F value is greater than the appropriate value of the F critical value (found in a table or provided in software), then the null hypothesis can be rejected. (helpful in ANOVA)

The calculated F-statistic for a known source of variation is found by dividing the mean square of the known source of variation by the mean square of the unknown source of variation.

When would you use an F Test?

There are different types of F tests are exists for different purpose.

- In statistics, an F-test of equality of variances is a test for the null hypothesis that two normal populations have the same variance.
- F-test is to test equality of several means. While ANOVA uses to test the equality of means.
- F-test for linear regression model is to tests any of the independent variables in a multiple linear regression are significant or not. It also indicates a linear relationship between dependent variable and at least one of the independent variable.

Steps to conduct F test

- Choose the test: Note down the independent variables and dependent variable and also assume the samples are normally distributed
- Calculate the F statistic, choose the highest variance in the numerator and lowest variance in the denominator with a degrees of freedom (n-1)
- Determine the statistical hypothesis
- State the level of significance
- Compute the critical F value from F table. (use $\alpha/2$ for two tailed test)
- Calculate the test statistic

- Finally, draw the statistical conclusion. reject the null hypothesis; If the test statistic falls in the critical region.

Example of an F Test in DMAIC

In Measure and Analyze phase of DMAIC. F test is to find out whether the two independent estimates of population variance differ significantly (or) to find out whether the two samples drawn from the normal population having the same variance002E

A step-by-step procedure for using the F-test

To accomplish the above goals, we will follow these steps:

STEP 1: Developing the intuition for the test statistic

Recollect that the F-test measures how much better a complex model is as compared to a simpler version of the same model in its ability to explain the variance in the dependent variable.

Consider two regression models 1 and 2 operating over a sample of *n* values:

- Let Model 1 has ***k_1*** parameters. Model 2 has ***k_2*** parameters.
- Let ***k_1*** < ***k_2***
- Thus, Model 1 is the simpler version of model 2. i.e. model 1 is the restricted model and model 2 is the unrestricted model. Model 1 can be nested within model 2.
- Let ***RSS_1*** and ***RSS_2*** be the sum of squares of residual errors after Model 1 and Model 2 are fitted to the same data set. **The residual error is the difference between the observed value and the predicted value.**

$$\begin{aligned} &Residual\ error\ \epsilon_i \\ &\quad = observed\ value - predicted\ value \\ &\quad = y_i - \widehat{\mu_i} \end{aligned}$$

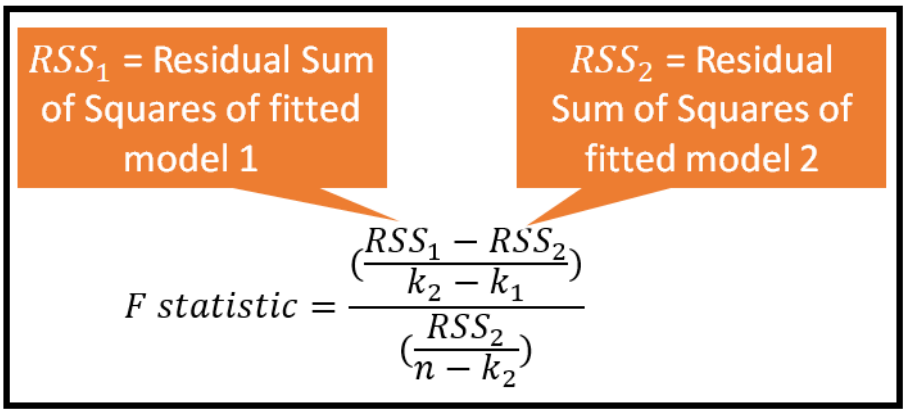
Residual error

The sum of squares of residuals (**RSS**) is expressed as follows:

$$Residual\ Sum\ of\ Squares\ (RSS) = \sum_{i=0}^{n} (y_i - \widehat{\mu_i})^2$$

Residual Sum of Squares (RSS)

With the above definitions in place, the test statistic of the F-test for regression can be expressed as a ratio as follows:



Formula for the F-statistic when applied to regression analysis

The F-statistic formula lets you calculate how much of the variance in the dependent variable, the simpler model is *not* able to explain as compared to the complex model, expressed as a fraction of the unexplained variance from the complex model.
In regression analysis, the mean squared error of the fitted model is an excellent measure of unexplained variance.

Which explains the RSS terms in the numerator and the denominator.

The numerator and the denominator are suitably scaled using the corresponding available degrees of freedom.

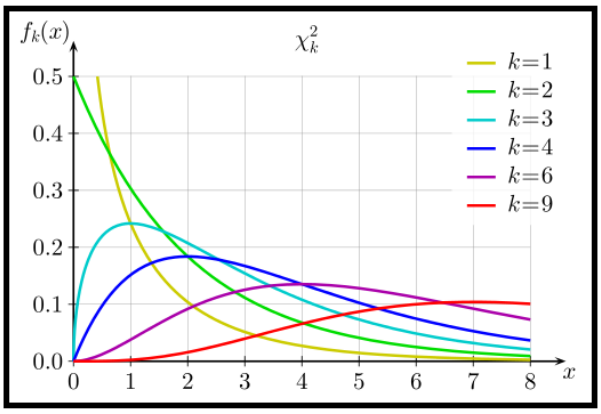
The F-statistic is itself a random variable.
Let's determine which **P**robability **D**ensity **F**unction the F-statistic obeys.

STEP 2: Identifying the Probability Density Function of the F-statistic

Notice that both the numerator and denominator of the test statistic contain sums of squares of residual errors. Also recollect that in regression, a residual error happens to be a random variable with some probability density (or probability mass) function, i.e. a PDF or PMF depending on whether it is continuous or discrete. In this case we are concerned with finding the PDF of the F-statistic.

If we assume that the residual errors from the two models are 1) independent and 2) normally distributed, which incidentally happen to be requirements of **O**rdinary **L**east **S**quares regression, then it can be seen that the numerator and denominator of the F-statistic formula contain sums of squares of independent, normally distributed random variables.

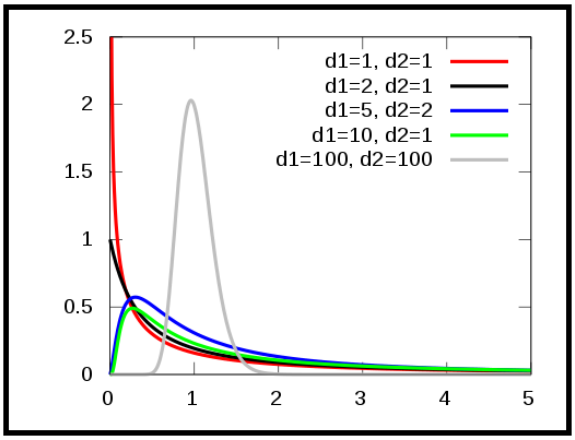
It can be proved that the sum of squares of k independent, standard normal random variables follow the PDF of the Chi-squared(k) distribution.



PDF of the Chi-Squared distribution (Source: Wikimedia Commons under CC BY 3.0)

Thus the numerator and denominator of the F-statistic formula can be shown to each obey scaled versions of two chi-squared distributions.

With a little bit of math, it can also be shown that the ratio of two suitably scaled Chi-squared distributed random variables is itself a random variable that follows the **F-distribution**, whose PDF is shown below.



The F-distribution (Source: Wikimedia Commons under CC BY-SA 4.0)

In other words:

If the random variable ***X*** has the PDF of the F-distribution with parameters *d_1* and *d_2*, i.e. :

$$If\ X \sim F(d_1, d_2)$$

then, ***X*** can be shown to be expressed as the ratio of two suitably scaled random variables ***X_1*** and ***X_2***, each of which has the PDF of a Chi-squared distribution. i.e. :

$$X = \frac{X_1/d_1}{X_2/d_2}$$

$$Where:$$

$$X_1 \sim \chi^2_{d_1} \text{ and } X_2 \sim \chi^2_{d_2}$$

An F-distributed random variable X, expressed as the ratio of two scaled Chi-squared distributed random variables X_1 and *X_2*

Now recollect that k_1 and k_2 are the number of variables in the simple and complex models M1 and M2 introduced earlier, and *n* is the number of data samples.
Substitute d_1 and d_2 as follows:

d_1 = (k_2 — k_1) which is the difference in degrees of freedom of the residuals of the two models M1 and M2 to be compared, and

d_2 = (n — k_2) which is the degrees of freedom of the residuals of the complex model M2,

With these substitutions, we can rewrite the F-distribution's formula as follows:

$$X = \frac{\frac{X_1}{(k_2 - k_1)}}{\frac{X_2}{(n - k_2)}}$$

$$Where:$$

$$X_1 \sim \chi^2_{(k_2 - k_1)} and\ X_2 \sim \chi^2_{(n - k_2)}$$

Alternate formula for the F-distribution's PDF

Let's compare the above formula with the formula for the F-statistic (reproduced below), where we know that the numerator and denominator contain suitably scaled PDFs of Chi-squared distributions:

$$(RSS_1 - RSS_2) \sim \chi^2_a$$

$$F\ statistic = \frac{(\frac{RSS_1 - RSS_2}{k_2 - k_1})}{(\frac{RSS_2}{n - k_2})}$$

$$RSS_2 \sim \chi^2_b$$

Formula for the F-test's test statistic

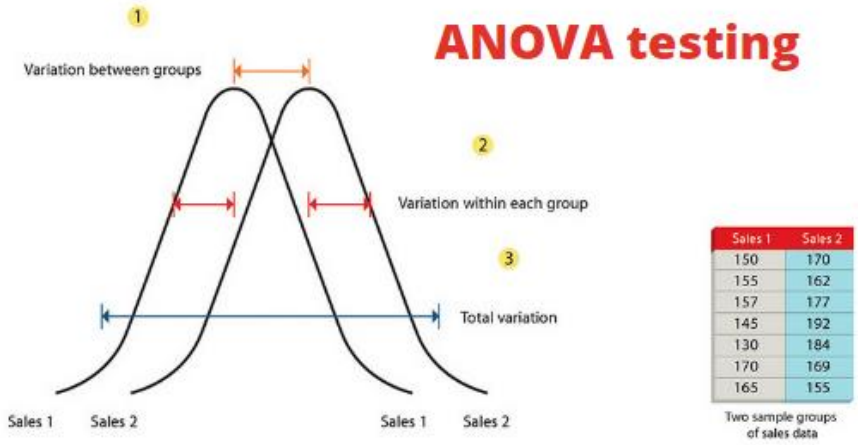
Comparing these two formulae, it is clear that:

1. The degree of freedom '*a*' of the Chi-squared distribution in the numerator is (k1 — k2).
2. The degree of freedom '*b*' of the Chi-squared distribution in the denominator is (n — k2).
3. The test statistic of the F-test has the same PDF as that of the F-distribution.

In other words, the F-statistic follows the F-distribution.

ANOVA

ANOVA is a type of hypothesis testing which is used to find out the experimental results by analyzing the variance of the different survey groups. It is usually used for deciding the result of the dataset.
Analysis of variance(ANOVA) is a statistical method to find out if the means of two or more groups are significantly different from each other. It checks the impact of one or more factors by comparing the means of different samples.
When we have two samples/groups we use a t-test to find out the mean between those samples but it is not that much reliable for more than two samples, therefore, we use ANOVA.



Why do we use ANOVA testing?
In machine learning, the biggest problem is selecting the best features or attributes for training the model. We only require those features that are highly dependent on the response variable so that our model can able to predict the actual outcome after training the model. ANOVA is used to figure out the result when we have a continuous response variable and the target feature is categorical.

For example, we set up an experience of three groups of people, the very first group gets water drinks, second get some sugary juice and the third one like to take coffee or tea. Now, we need to test everyone's reaction time and want to know if there is any difference between the groups or not.

29	28	25
30	29	28
31	27	29
31	30	27
29	29	29

The **null hypothesis** tells that all the three groups have the same reaction time, we have three groups here to experiment and find out the result so we need to apply the ANOVA testing in case of two groups we could use the t-test when we experiment we would notice that the result won't be same.

The total variance of all these scores is made up of two parts:

1. The variance within the groups: As people have different reaction time in each group.
2. The variance between the groups: As the drinks are different which people prefer.

Example one:

10	11	12
12	14	13
18	19	17
24	23	25
36	38	37

As we can see here, there is a lot of variation in each sample/group, some of them are faster and some of them are slower but the groups are quite to one another, there is not much variation between the groups. So we can say that people are making a difference but not the type of drinks, in this case, we need to accept the null hypothesis we can't reject that as the type of drink doesn't put any effect on reaction time.

Example two:

29	17	10
29	18	11
30	19	12
31	19	13
31	20	13

Here we can see that there is not much difference within the groups but there is a lot of f=deifference between the groups. The people's reaction time doesn't make any effect on the groups, so here we will reject the null hypothesis.

In the example, we have seen a term hypothesis, what is the Hypothesis? ANOVA uses many terminologies with it.

Mean:

There are two types of mean that we used in ANOVA

1. Mean of each sample
2. Grand mean that is the mean of all the observation combined.

Types of ANOVA

<u>One way ANOVA:</u>

The one-way ANOVA is used to find out the statistically significant difference between the mean of more than two independent groups.

More specifically it is used to test the null hypothesis.

In one-way ANOVA μ = group means and k is a number of groups, if one-way ANOVA returns the significant result, in this case, we accept the alternative hypothesis, this means that the mean of two groups is not equal.

<u>Two-way ANOVA:</u>

A two-way is used to determine the effect of two nominal predictor features on a continuous outcome feature. It tests the effect of two independent variables on the expected outcome with the outcome itself.

<u>F-value for ANOVA:</u>

The F-value os ANOVA is a tool to help you to determine that, Is the variance between the means of two samples significantly different or not. The ratio of the between the groups and within the groups. It also helps us to find out the p-Value. The P-value is the probability of getting the result at least at the point where the null hypothesis should be true.

The formula for f-value:

$$\textbf{F-value=}\frac{\text{Mean between the groups}}{\text{Mean Within the groups}}$$

ANOVA Test Table

cuemath
THE MATH EXPERT

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma\, n_j(\bar{X}_j - \bar{X})^2$	$df_1 = k - 1$	$MSB = SSB / (k - 1)$	$f = MSB / MSE$
Error	$SSE = \Sigma\Sigma(X - \bar{X}_j)^2$	$df_2 = N - k$	$MSE = SSE / (N - k)$	
Total	$SST = SSB + SSE$	$df_3 = N - 1$		

Chi-Square Test

When we consider, the null speculation is true, the sampling distribution of the test statistic is called as **chi-squared distribution**. The chi-squared test helps to determine whether there is a notable difference between the normal frequencies and the observed frequencies in one or more classes or categories. It gives the probability of independent variables.

Note: Chi-squared test is applicable only for categorical data, such as men and women falling under the categories of Gender, Age, Height, etc.

Finding P-Value

P stands for probability here. To calculate the p-value, the chi-square test is used in statistics. The different values of p indicates the different hypothesis interpretation, are given below:

- P≤ 0.05; Hypothesis rejected
- P>.05; Hypothesis Accepted

Probability is all about chance or risk or uncertainty. It is the possibility of the outcome of the sample or the occurrence of an event. But when we talk about statistics, it is more about how we handle various data using different techniques. It helps to represent complicated data or bulk data in a very easy and understandable way. It describes the collection, analysis, interpretation, presentation, and organization of data. The concept of both probability and statistics is related to the chi-squared test.

Properties

The following are the important properties of the chi-square test:

- Two times the number of degrees of freedom is equal to the variance.
- The number of degree of freedom is equal to the mean distribution
- The chi-square distribution curve approaches the normal distribution when the degree of freedom increases.

Formula

The chi-squared test is done to check if there is any difference between the observed value and expected value. The formula for chi-square can be written as;

$$X^2 = \sum \frac{(\text{Observed value - Expected value})^2}{\text{Expected value}}$$

or

$$\chi^2 = \sum(O_i - E_i)^2/E_i$$

where O_i is the observed value and E_i is the expected value.

Chi-Square Test of Independence

The chi-square test of independence also known as the chi-square test of association which is used to determine the association between the categorical variables. It is considered as a non-parametric test. It is mostly used to test statistical independence.

The chi-square test of independence is not appropriate when the categorical variables represent the pre-test and post-test observations. For this test, the data must meet the following requirements:

- Two categorical variables
- Relatively large sample size
- Categories of variables (two or more)
- Independence of observations

Example of Categorical Data

Let us take an example of a categorical data where there is a society of 1000 residents with four neighbourhoods, P, Q, R and S. A random sample of 650 residents of the society is taken whose occupations are doctors, engineers and teachers. The null hypothesis is that each person's neighbourhood of residency is independent of the person's professional division. The data are categorised as:

Categories	P	Q	R	S	Total

Doctors	90	60	104	95	349
Engineers	30	50	51	20	151
Teachers	30	40	45	35	150
Total	150	150	200	150	650

Assume the sample living in neighbourhood P, 150, to estimate what proportion of the whole 1,000 people live in neighbourhood P. In the same way, we take 349/650 to calculate what ratio of the 1,000 are doctors. By the supposition of independence under the hypothesis, we should "expect" the number of doctors in neighbourhood P is;

150 x 349/650 ≈ 80.54

So by the chi-square test formula for that particular cell in the table, we get;

(Observed – Expected)^2/Expected Value = (90-80.54)^2/80.54 ≈ 1.11

Some of the exciting facts about the Chi-square test are given below:

The Chi-square statistic can only be used on numbers. We cannot use them for data in terms of percentages, proportions, means or similar statistical contents. Suppose, if we have 20% of 400 people, we need to convert it to a number, i.e. 80, before running a test statistic.

A chi-square test will give us a p-value. The p-value will tell us whether our test results are significant or not.

However, to perform a chi-square test and get the p-value, we require two pieces of information:

(1) Degrees of freedom. That's just the number of categories minus 1.

(2) The alpha level(α). You or the researcher chooses this. The usual alpha level is 0.05 (5%), but you could also have other levels like 0.01 or 0.10.

In elementary statistics, we usually get questions along with the degrees of freedom(DF) and the alpha level. Thus, we don't usually have to figure out what they are. To get the degrees of freedom, count the categories and subtract 1.

What is the chi-square test write its formula?

When we consider the null hypothesis is true, the test statistic's sampling distribution is called chi-squared distribution. The formula for chi-square is:
χ^2 = ∑(O_i – E_i)^2/E_i
Here,
O_i = Observed value
E_i = Expected value

How do you calculate chi squared?

The value of the Chi-squared statistic can be calculated using the formula given below:
$\chi^2 = \sum(O_i – E_i)^2/E_i$
This can be done as follows.
For each observed number in the data, subtract the corresponding expected value, i.e. (O — E).
Square the difference, (O — E)^2
Divide these squares by the expected value of each observation, i.e. [(O – E)^2 / E].
Finally, take the sum of these values.
Thus, the obtained value will be the chi-squared statistic.

What is a chi-square test used for?

The chi-squared test is done to check if there is any difference between the observed value and the expected value.

How do you interpret a chi-square test?

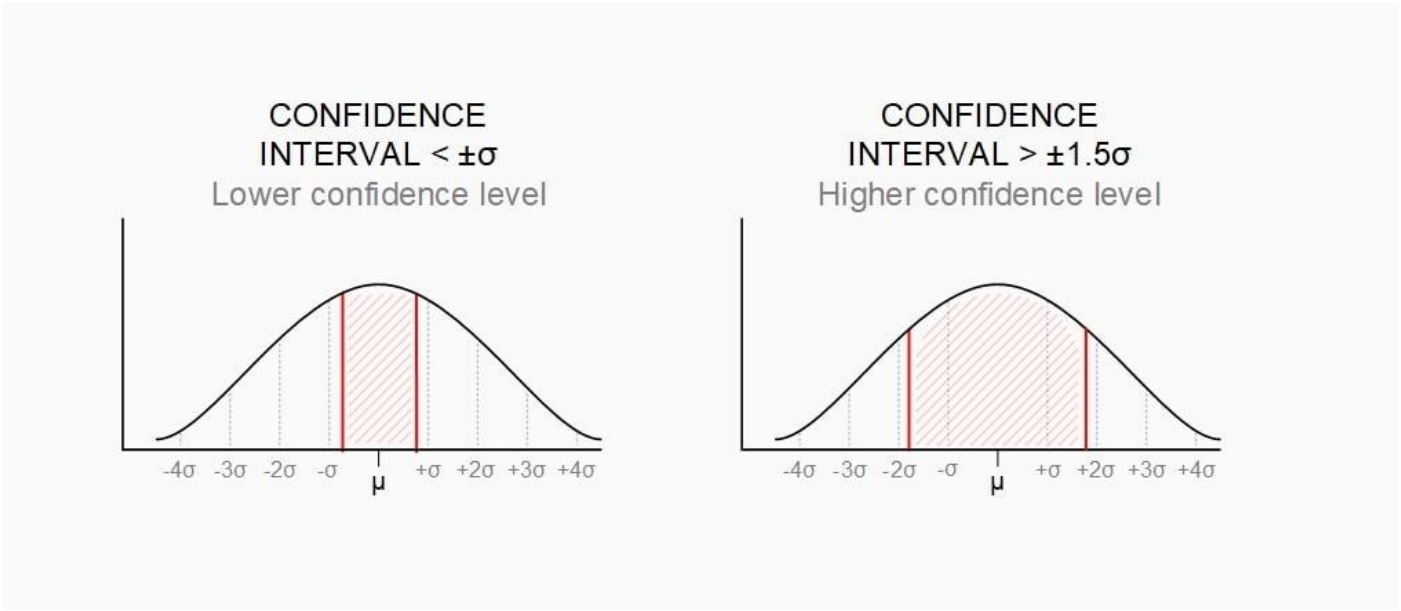
For a Chi-square test, a p-value that is less than or equal to the specified significance level indicates sufficient evidence to conclude that the observed distribution is not the same as the expected distribution. Here, we can conclude that a relationship exists between the given categorical variables.

What is a good chi-square value?

A good chi-square value is assumed to be 5. As we know, for the chi-square approach to be valid, the expected frequency should be at least 5.

Confidence Interval

Confidence interval is a kind of interval calculation, obtained from the observed data that holds the actual value of the unknown parameter. It is associated with the confidence level that quantifies the confidence level in which the interval estimates the deterministic parameter. Also, we can say, it is based on Standard Normal Distribution, where Z value is the z-score.



Definition

The confidence level represents the proportion (frequency) of acceptable confidence intervals that contain the true value of the unknown parameter. In other terms, the confidence intervals are evaluated using the given confidence level from an endless number of independent samples. So that the proportion of the range contains the true value of the parameter that will be equal to the confidence level.

Mostly, the confidence level is selected before examining the data. The commonly used confidence level is 95% confidence level. However, other confidence levels are also used, such as 90% and 99% confidence levels.

Confidence Interval Formula

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

CI = confidence interval

$\bar{x}$ = sample mean

z = confidence level value

s = sample standard deviation

n = sample size

The value after the **±** symbol is known as the margin of error.

Note: This interval is only accurate when the population distribution is normal. But, in the case of large samples from other population distributions, the interval is almost accurate by the Central Limit Theorem.

Confidence Interval Table

The confidence interval table for Z values are given as follows

Confidence Interval	Z Value
80%	1.282
85%	1.440
90%	1.645
95%	1.960

99%	2.576
99.5%	2.807
99.9%	3.291

How to Calculate Confidence Interval?

To calculate the confidence interval, go through the following procedure.

Step 1: Find the number of observations n(sample space), mean $\bar{X}$, and the standard deviation σ.

Step 2: Decide the confidence interval of your choice. It should be either 95% or 99%. Then find the Z value for the corresponding confidence interval given in the table.

Step 3: Finally, substitute all the values in the formula.

Confidence Interval Example

Question: In a tree, there are hundreds of apples. You are randomly choosing 46 apples with a mean of 86 and a standard deviation of 6.2. Determine that the apples are big enough.

Solution:

Given: Mean, $\bar{X} = 86$

Standard deviation, $\sigma = 6.2$

Number of observations, $n = 46$

Take the confidence level as 95%. Therefore, the value of $z = 1.960$ (from the table)

The formula to find the confidence interval is

$$\bar{X} \pm Z\alpha/2 \times [\ \sigma / \sqrt{n}\ ]$$

Now, substitute the values in the formula, we get

$86 \pm 1.960 \times [\ 6.2 / \sqrt{46}\ ]$

$86 \pm 1.960 \times [\ 6.2 / 6.78]$

$86 \pm 1.960 \times 0.914$

86 ± 1.79

Here, the margin of error is 1.79

Therefore, all the hundreds of apples are likely to be between in the range of 84. 21 and 87.79.

Estimates of Variability

Location is just one dimension in summarizing a feature. A second dimension, *variability*, also referred to as *dispersion*, measures whether the data values are tightly clustered or spread out. At the heart of statistics lies variability: measuring it, reducing it, distinguishing random from real variability, identifying the various sources of real variability, and making decisions in the presence of it.

KEY TERMS FOR VARIABILITY METRICS

Deviations

The difference between the observed values and the estimate of location.

Synonyms

errors, residuals

Variance

The sum of squared deviations from the mean divided by $n - 1$ where n is the number of data values.

Synonyms

mean-squared-error

Standard deviation

The square root of the variance.

Synonyms

l2-norm, Euclidean norm

Mean absolute deviation

The mean of the absolute value of the deviations from the mean.

Synonyms

l1-norm, Manhattan norm

Median absolute deviation from the median

The median of the absolute value of the deviations from the median.

Range

The difference between the largest and the smallest value in a data set.

Order statistics

Metrics based on the data values sorted from smallest to biggest.

Synonyms

ranks

Percentile

The value such that *P* percent of the values take on this value or less and (100–P) percent take on this value or more.

Synonyms

quantile

Interquartile range

The difference between the 75th percentile and the 25th percentile.

Synonyms

IQR

Just as there are different ways to measure location (mean, median, etc.) there are also different ways to measure variability.

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

$$\text{Variance} = s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

$$\text{Standard deviation} = s = \sqrt{\text{Variance}}$$

Exploring the Data Distribution

Each of the estimates we've covered sums up the data in a single number to describe the location or variability of the data. It is also useful to explore how the data is distributed overall.

KEY TERMS FOR EXPLORING THE DISTRIBUTION

Boxplot

A plot introduced by Tukey as a quick way to visualize the distribution of data.

Synonyms

Box and whiskers plot

Frequency table

A tally of the count of numeric data values that fall into a set of intervals (bins).

Histogram

A plot of the frequency table with the bins on the x-axis and the count (or proportion) on the y-axis.

Density plot

A smoothed version of the histogram, often based on a *kernal density estimate*.

Exploring Binary and Categorical Data

For categorical data, simple proportions or percentages tell the story of the data.

KEY TERMS FOR EXPLORING CATEGORICAL DATA

Mode

The most commonly occurring category or value in a data set.

Expected value

When the categories can be associated with a numeric value, this gives an average value based on a category's probability of occurrence.

Bar charts

The frequency or proportion for each category plotted as bars.

Pie charts

The frequency or proportion for each category plotted as wedges in a pie.

KEY IDEAS

- Categorical data is typically summed up in proportions, and can be visualized in a bar chart.
- Categories might represent distinct things (apples and oranges, male and female), levels of a factor variable (low, medium, and high), or numeric data that has been binned.
- Expected value is the sum of values times their probability of occurrence, often used to sum up factor variable levels.

Correlation

Exploratory data analysis in many modeling projects (whether in data science or in research) involves examining correlation among predictors, and between predictors and a target variable. Variables X and Y (each with measured data) are said to be positively correlated if high values of X go with high values of Y, and low values of X go with low values of Y. If high values of X go with low values of Y, and vice versa, the variables are negatively correlated.

KEY TERMS FOR CORRELATION

Correlation coefficient

A metric that measures the extent to which numeric variables are associated with one another (ranges from –1 to +1).

Correlation matrix

A table where the variables are shown on both rows and columns, and the cell values are the correlations between the variables.

Scatterplot

A plot in which the x-axis is the value of one variable, and the y-axis the value of another.

Correlation measures the relationship between two variables. We mentioned that a function has a purpose to predict a value, by converting input (x) to output (f(x)). We can say also say that a function uses the relationship between two variables for prediction.
The correlation coefficient is a standardized metric so that it always ranges from –1 (perfect negative correlation) to +1 (perfect positive correlation). A correlation coefficient of 0 indicates no correlation, but be aware that random arrangements of data will produce both positive and negative values for the correlation coefficient just by chance.

Exploring Two or More Variables

Familiar estimators like mean and variance look at variables one at a time (*univariate analysis*). Correlation analysis (see "Correlation") is an important method that compares two variables (*bivariate analysis*). In this section we look at additional estimates and plots, and at more than two variables (*multivariate analysis*).

KEY TERMS FOR EXPLORING TWO OR MORE VARIABLES

Contingency tables

A tally of counts between two or more categorical variables.

Hexagonal binning

A plot of two numeric variables with the records binned into hexagons.

Contour plots

A plot showing the density of two numeric variables like a topographical map.

Violin plots

Similar to a boxplot but showing the density estimate.

KEY IDEAS

- Hexagonal binning and contour plots are useful tools that permit graphical examination of two numeric variables at a time, without being overwhelmed by huge amounts of data.
- Contingency tables are the standard tool for looking at the counts of two categorical variables.
- Boxplots and violin plots allow you to plot a numeric variable against a categorical variable.

Data and Sampling Distributions

A popular misconception holds that the era of big data means the end of a need for sampling. In fact, the proliferation of data of varying quality and relevance reinforces the need for sampling as a tool to work efficiently with a variety of data and to minimize bias. Even in a big data project, predictive models are typically developed and piloted with samples. Samples are also used in tests of various sorts (e.g., pricing, web treatments).

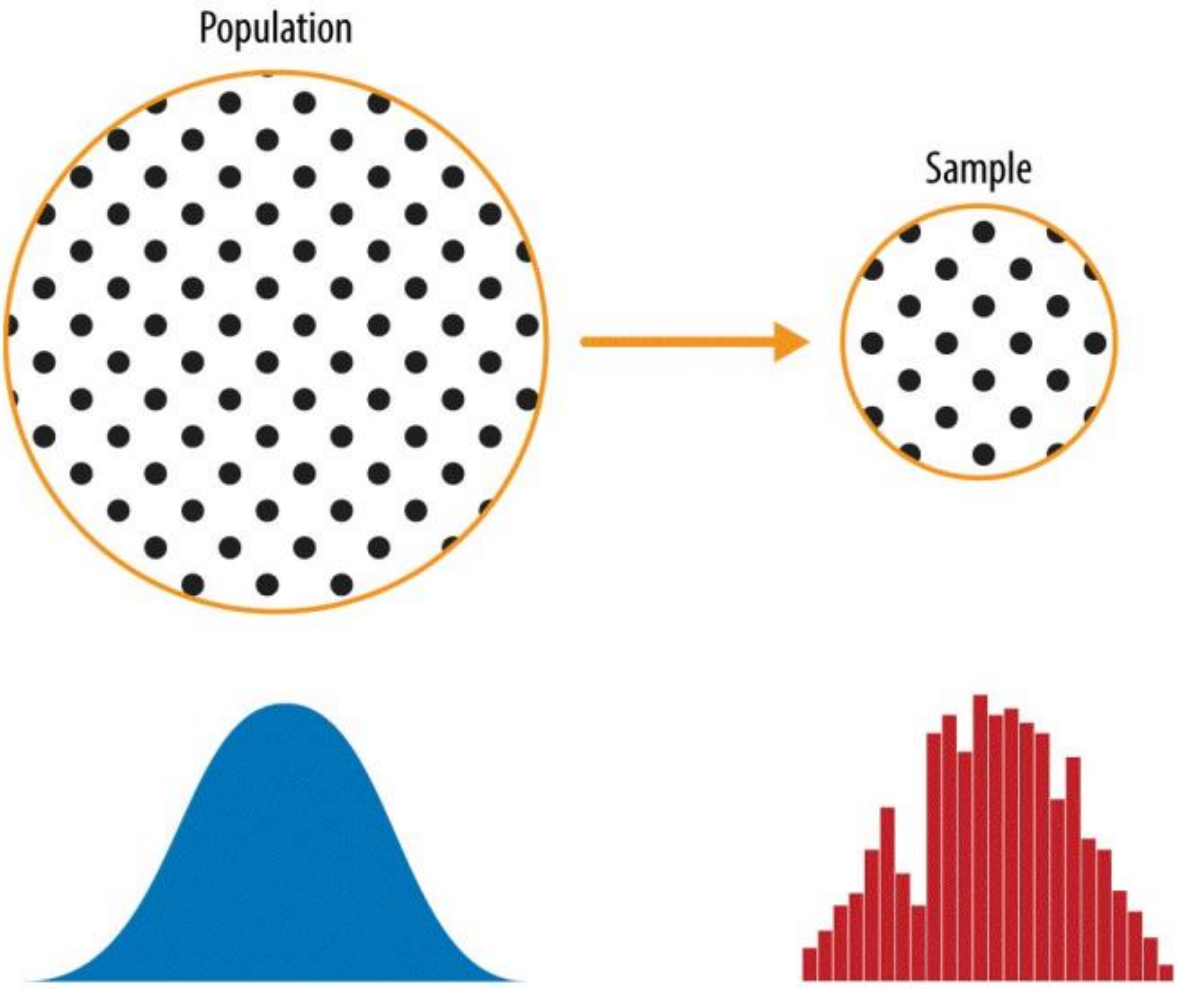


Figure 2-1. Population versus sample

KEY TERMS FOR RANDOM SAMPLING

Sample

A subset from a larger data set.

Population

The larger data set or idea of a data set.

N (n)

The size of the population (sample).

Random sampling

Drawing elements into a sample at random.

Stratified sampling

Dividing the population into strata and randomly sampling from each strata.

Simple random sample

The sample that results from random sampling without stratifying the population.

Sample bias

A sample that misrepresents the population.

KEY IDEAS

- Even in the era of big data, random sampling remains an important arrow in the data scientist's quiver.
- Bias occurs when measurements or observations are systematically in error because they are not representative of the full population.
- Data quality is often more important than data quantity, and random sampling can reduce bias and facilitate quality improvement that would be prohibitively expensive.

Bias

Statistical bias refers to measurement or sampling errors that are systematic and produced by the measurement or sampling process. An important distinction should be made between errors due to random chance, and errors due to bias.

Eg: It will not hit the absolute center of the target every time, or even much at all.

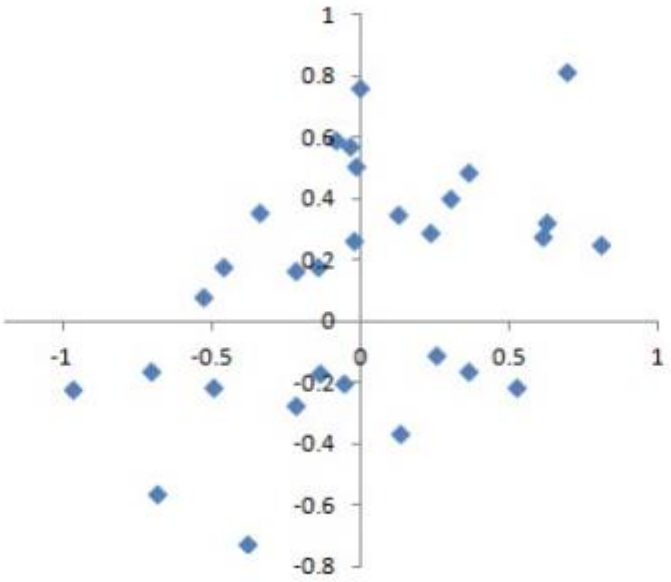


Figure 2-2. Scatterplot of shots from a gun with true aim

Selection Bias

To paraphrase Yogi Berra, “If you don’t know what you’re looking for, look hard enough and you’ll find it.”

Selection bias refers to the practice of selectively choosing data — consciously or unconsciously — in a way that that leads to a conclusion that is misleading or ephemeral.

KEY TERMS

Bias

Systematic error.

Data snooping

Extensive hunting through data in search of something interesting.

Vast search effect

Bias or nonreproducibility resulting from repeated data modeling, or modeling data with large numbers of predictor variables.

Sampling Distribution of a Statistic

The term *sampling distribution* of a statistic refers to the distribution of some sample statistic, over many samples drawn from the same population. Much of classical statistics is concerned with making inferences from (small) samples to (very large) populations.

KEY TERMS

Sample statistic
A metric calculated for a sample of data drawn from a larger population.

Data distribution
The frequency distribution of individual *values* in a data set.

Sampling distribution
The frequency distribution of a *sample statistic* over many samples or resamples.

Central limit theorem
The tendency of the sampling distribution to take on a normal shape as sample size rises.

Standard error
The variability (standard deviation) of a sample *statistic* over many samples (not to be confused with *standard deviation*, which, by itself, refers to variability of individual data *values*).

Resampling versus Bootstrapping

Sometimes the term *resampling* is used synonymously with the term *bootstrapping*, as just outlined. More often, the term *resampling* also includes permutation procedures (see "Permutation Test"), where multiple samples are combined and the sampling may be done without replacement. In any case, the term *bootstrap* always implies sampling with replacement from an observed data set.

KEY IDEAS

- The bootstrap (sampling with replacement from a data set) is a powerful tool for assessing the variability of a sample statistic.
- The bootstrap can be applied in similar fashion in a wide variety of circumstances, without extensive study of mathematical approximations to sampling distributions.
- It also allows us to estimate sampling distributions for statistics where no mathematical approximation has been developed.
- When applied to predictive models, aggregating multiple bootstrap sample predictions (bagging) outperforms the use of a single model.

Confidence Intervals

Frequency tables, histograms, boxplots, and standard errors are all ways to understand the potential error in a sample estimate. Confidence intervals are another.

KEY TERMS

Confidence level

The percentage of confidence intervals, constructed in the same way from the same population, expected to contain the statistic of interest.

Interval endpoints

The top and bottom of the confidence interval.

Given a sample of size n, and a sample statistic of interest, the algorithm for a bootstrap confidence interval is as follows:

1. Draw a random sample of size n with replacement from the data (a resample).
2. Record the statistic of interest for the resample.
3. Repeat steps 1–2 many (R) times.
4. For an x% confidence interval, trim [(1 – [x/100]) / 2]% of the R resample results from either end of the distribution.
5. The trim points are the endpoints of an x% bootstrap confidence interval.

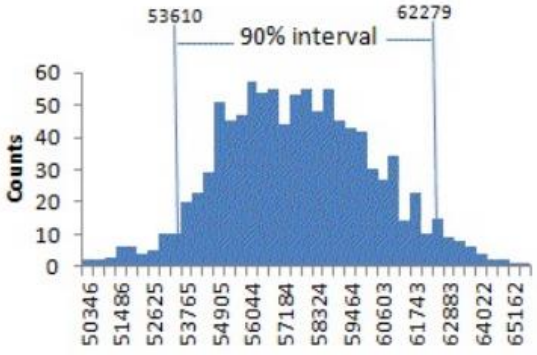


Figure 2-9. Bootstrap confidence interval for the annual income of loan applicants, based on a sample of 20

NOTE

For a data scientist, a confidence interval is a tool to get an idea of how variable a sample result might be. Data scientists would use this information not to publish a scholarly paper or submit a result to a regulatory agency (as a researcher might), but most likely to communicate the potential error in an estimate, and, perhaps, learn whether a larger sample is needed.

KEY IDEAS

- Confidence intervals are the typical way to present estimates as an interval range.
- The more data you have, the less variable a sample estimate will be.
- The lower the level of confidence you can tolerate, the narrower the confidence interval will be.
- The bootstrap is an effective way to construct confidence intervals.

Normal Distribution

The bell-shaped normal distribution is iconic in traditional statistics.[1] The fact that distributions of sample statistics are often normally shaped has made it a powerful tool in the development of mathematical formulas that approximate those distributions.

KEY TERMS

Error
: The difference between a data point and a predicted or average value.

Standardize
: Subtract the mean and divide by the standard deviation.

z-score
: The result of standardizing an individual data point.

Standard normal
: A normal distribution with mean = 0 and standard deviation = 1.

QQ-Plot
: A plot to visualize how close a sample distribution is to a normal distribution.

In a normal distribution (Figure 2-10), 68% of the data lies within one standard deviation of the mean, and 95% lies within two standard deviations.

> **WARNING**
>
> It is a common misconception that the normal distribution is called that because most data follows a normal distribution — that is, it is the normal thing. Most of the variables used in a typical data science project — in fact most raw data as a whole — are *not* normally distributed: see "Long-Tailed Distributions". The utility of the normal distribution derives from the fact that many statistics *are* normally distributed in their sampling distribution. Even so, assumptions of normality are generally a last resort, used when empirical probability distributions, or bootstrap distributions, are not available.

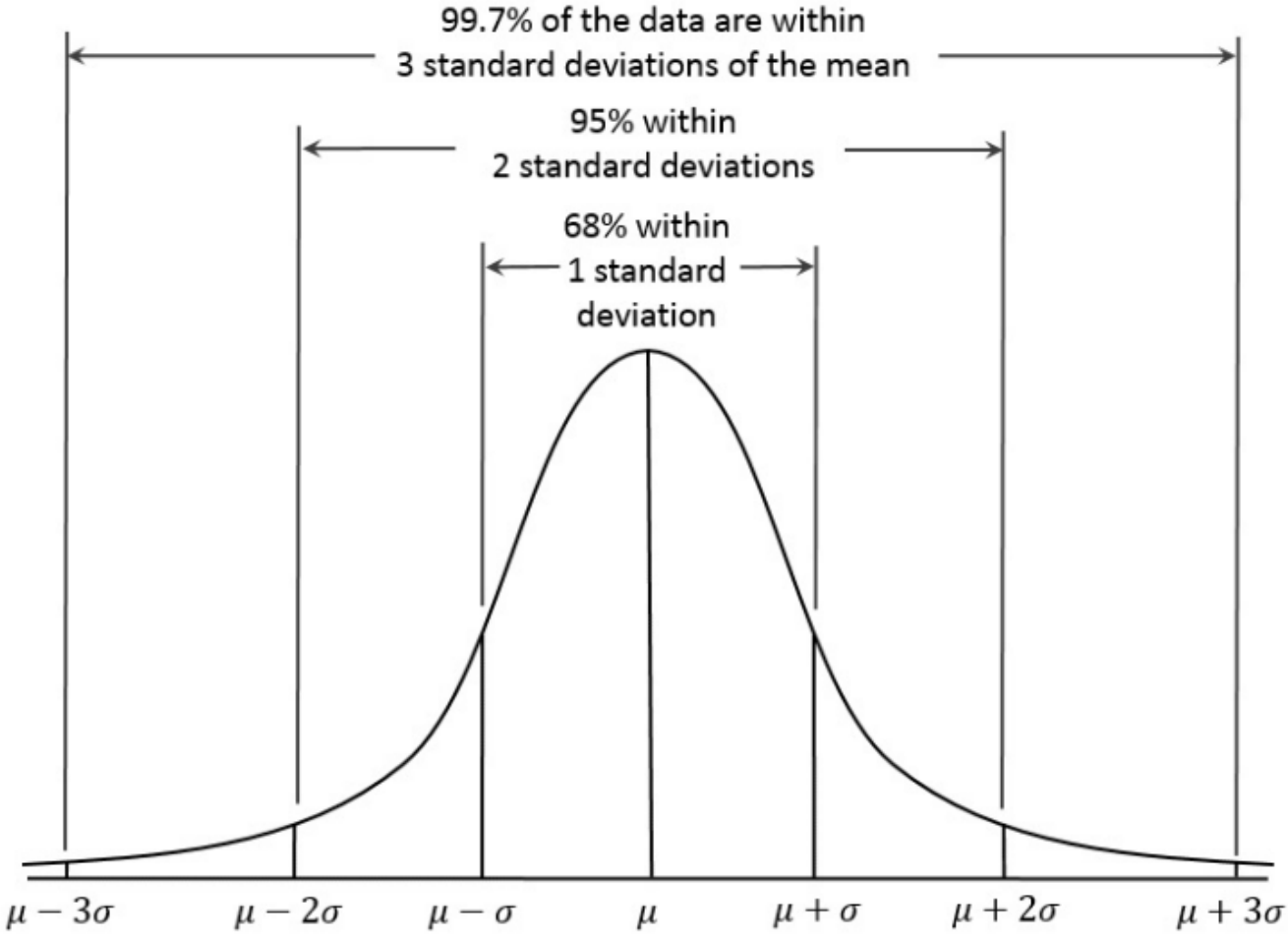


Figure 2-10. Normal curve

NOTE

The normal distribution is also referred to as a *Gaussian* distribution after Carl Friedrich Gauss, a prodigous German mathematician from the late 18th and early 19th century. Another name previously used for the normal distribution was the "error" distribution. Statistically speaking, an *error* is the difference between an actual value and a statistical estimate like the sample mean. For example, the standard deviation (see "Estimates of Variability") is based on the errors from the mean of the data. Gauss's development of the normal distribution came from his study of the errors of astronomical measurements that were found to be normally distributed.

KEY IDEAS

- The normal distribution was essential to the historical development of statistics, as it permitted mathematical approximation of uncertainty and variability.
- While raw data is typically not normally distributed, errors often are, as are averages and totals in large samples.
- To convert data to *z*-scores, you subtract the mean of the data and divide by the standard deviation; you can then compare the data to a normal distribution.

Long-Tailed Distributions

Despite the importance of the normal distribution historically in statistics, and in contrast to what the name would suggest, data is generally not normally distributed.

KEY TERMS FOR LONG-TAIL DISTRIBUTION

Tail

The long narrow portion of a frequency distribution, where relatively extreme values occur at low frequency.

Skew

Where one tail of a distribution is longer than the other.

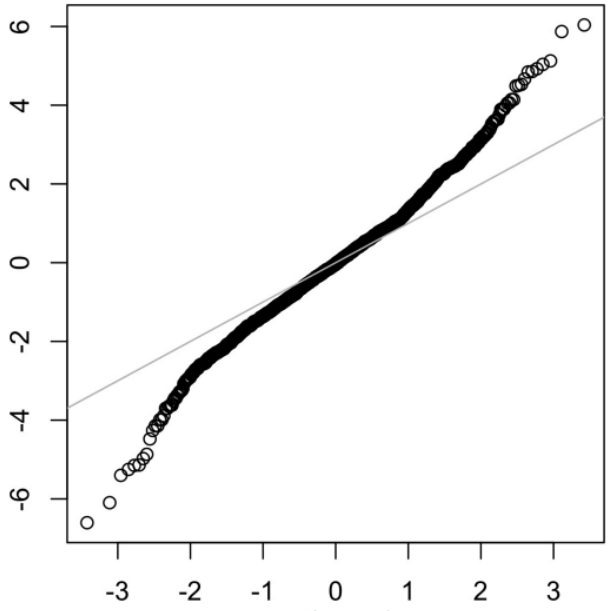


Figure 2-12. QQ-Plot of the returns for NFLX

NOTE

There is much statistical literature about the task of fitting statistical distributions to observed data. Beware an excessively data-centric approach to this job, which is as much art as science. Data is variable, and often consistent, on its face, with more than one shape and type of distribution. It is typically the case that domain and statistical knowledge must be brought to bear to determine what type of distribution is appropriate to model a given situation. For example, we might have data on the level of internet traffic on a server over many consecutive 5-second periods. It is useful to know that the best distribution to model "events per time period" is the Poisson (see "Poisson Distributions").

KEY IDEAS FOR LONG-TAIL DISTRIBUTION

- Most data is not normally distributed.
- Assuming a normal distribution can lead to underestimation of extreme events ("black swans").

Student's t-Distribution

The *t-distribution* is a normally shaped distribution, but a bit thicker and longer on the tails. It is used extensively in depicting distributions of sample statistics. Distributions of sample means are typically shaped like a t-distribution, and there is a family of t-distributions that differ depending on how large the sample is. The larger the sample, the more normally shaped the t-distribution becomes.

KEY TERMS FOR STUDENT'S T-DISTRIBUTION

n

Sample size.

Degrees of freedom

A parameter that allows the t-distribution to adjust to different sample sizes, statistics, and number of groups.

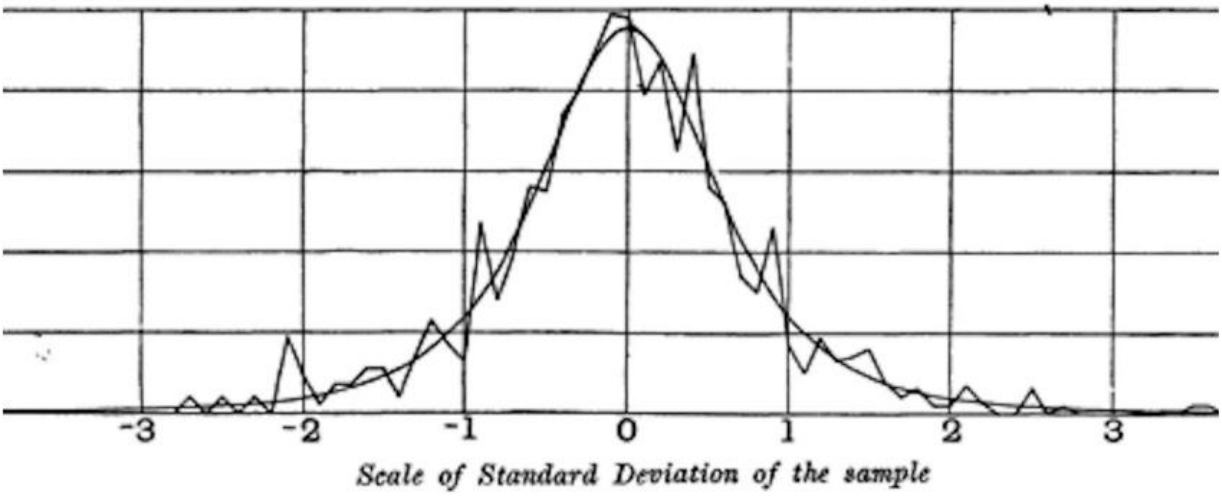


Figure 2-13. Gossett's resampling experiment results and fitted t-curve (from his 1908 Biometrika paper)

NOTE

What do data scientists need to know about the t-distribution and the central limit theorem? Not a whole lot. These distributions are used in classical statistical inference, but are not as central to the purposes of data science. Understanding and quantifying uncertainty and variation are important to data scientists, but empirical bootstrap sampling can answer most questions about sampling error. However, data scientists will routinely encounter t-statistics in output from statistical software and statistical procedures in R, for example in A-B tests and regressions, so familiarity with its purpose is helpful.

KEY IDEAS

- The t-distribution is actually a family of distributions resembling the normal distribution, but with thicker tails.
- It is widely used as a reference basis for the distribution of sample means, differerences between two sample means, regression parameters, and more.

Binomial Distribution

KEY TERMS FOR BINOMIAL DISTRIBUTION

Trial

An event with a discrete outcome (e.g., a coin flip).

Success

The outcome of interest for a trial.

Synonyms

"1" (as opposed to "0")

Binomial

Having two outcomes.

Synonyms

yes/no, 0/1, binary

Binomial trial

A trial with two outcomes.

Synonym

Bernoulli trial

Binomial distribution

Distribution of number of successes in x trials.

Synonym

Bernoulli distribution

Yes/no (binomial) outcomes lie at the heart of analytics since they are often the culmination of a decision or other process; buy/don't buy, click/don't click, survive/die, and so on. Central to understanding the binomial distribution is the idea of a set of trials, each trial having two possible outcomes with definite probabilities.

For example, flipping a coin 10 times is a binomial experiment with 10 trials, each trial having two possible outcomes (heads or tails);

KEY IDEAS

- Binomial outcomes are important to model, since they represent, among other things, fundamental decisions (buy or don't buy, click or don't click, survive or die, etc.).
- A binomial trial is an experiment with two possible outcomes: one with probability p and the other with probability $1 - p$.
- With large n, and provided p is not too close to 0 or 1, the binomial distribution can be approximated by the normal distribution.

Poisson and Related Distributions

Many processes produce events randomly at a given overall rate — visitors arriving at a website, cars arriving at a toll plaza (events spread over time), imperfections in a square meter of fabric, or typos per 100 lines of code (events spread over space).

KEY TERMS FOR POISSON AND RELATED DISTRIBUTIONS

Lambda
The rate (per unit of time or space) at which events occur.

Poisson distribution
The frequency distribution of the number of events in sampled units of time or space.

Exponential distribution
The frequency distribution of the time or distance from one event to the next event.

Weibull distribution
A generalized version of the exponential, in which the event rate is allowed to shift over time.

A/B Testing

An A/B test is an experiment with two groups to establish which of two treatments, products, procedures, or the like is superior. Often one of the two treatments is the standard existing treatment, or no treatment. If a standard (or no) treatment is used, it is called the *control*. A typical hypothesis is that treatment is better than control.

KEY TERMS FOR A/B TESTING

Treatment
Something (drug, price, web headline) to which a subject is exposed.

Treatment group
A group of subjects exposed to a specific treatment.

Control group
A group of subjects exposed to no (or standard) treatment.

Randomization
The process of randomly assigning subjects to treatments.

Subjects
The items (web visitors, patients, etc.) that are exposed to treatments.

Test statistic
The metric used to measure the effect of the treatment.

KEY IDEAS

- Subjects are assigned to two (or more) groups that are treated exactly alike, except that the treatment under study differs from one to another.
- Ideally, subjects are assigned randomly to the groups.

Hypothesis Tests

Hypothesis tests, also called *significance tests*, are ubiquitous in the traditional statistical analysis of published research. Their purpose is to help you learn whether random chance might be responsible for an observed effect.

KEY TERMS

Null hypothesis
The hypothesis that chance is to blame.

Alternative hypothesis
Counterpoint to the null (what you hope to prove).

One-way test
Hypothesis test that counts chance results only in one direction.

Two-way test
Hypothesis test that counts chance results in two directions.

In a properly designed A/B test, you collect data on treatments A and B in such a way that any observed difference between A and B must be due to either: Random chance in assignment of subjects A true difference between A and B A statistical hypothesis test is further analysis of an A/B test, or any randomized experiment, to assess whether random chance is a reasonable explanation for the observed difference between groups A and B.

Resampling

Resampling in statistics means to repeatedly sample values from observed data, with a general goal of assessing random variability in a statistic. It can also be used to assess and improve the accuracy of some machine-learning models (e.g., the predictions from decision tree models built on multiple bootstrapped data sets can be averaged in a process known as *bagging*: see "Bagging and the Random Forest").

There are two main types of resampling procedures: the *bootstrap* and *permutation* tests. The bootstrap is used to assess the reliability of an estimate; it was discussed in the previous chapter (see "The Bootstrap"). Permutation tests are used to test hypotheses, typically involving two or more groups, and we discuss those in this section.

KEY TERMS

Permutation test

The procedure of combining two or more samples together, and randomly (or exhaustively) reallocating the observations to resamples.

Synonyms

Randomization test, random permutation test, exact test.

With or without replacement

In sampling, whether or not an item is returned to the sample before the next draw.

Statistical Significance and P-Values

Statistical significance is how statisticians measure whether an experiment (or even a study of existing data) yields a result more extreme than what chance might produce. If the result is beyond the realm of chance variation, it is said to be statistically significant.

KEY TERMS

P-value

Given a chance model that embodies the null hypothesis, the p-value is the probability of obtaining results as unusual or extreme as the observed results.

Alpha

The probability threshold of "unusualness" that chance results must surpass, for actual outcomes to be deemed statistically significant.

Type 1 error

Mistakenly concluding an effect is real (when it is due to chance).

Type 2 error

Mistakenly concluding an effect is due to chance (when it is real).

The ASA statement stressed six principles for researchers and journal editors:

1. P-values can indicate how incompatible the data are with a specified statistical model.

2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

4. Proper inference requires full reporting and transparency.

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Type 1 and Type 2 Errors

In assessing statistical significance, two types of error are possible:

- Type 1 error, in which you mistakenly conclude an effect is real, when it is really just due to chance
- Type 2 error, in which you mistakenly conclude that an effect is not real (i.e., due to chance), when it really is real

Actually, a Type 2 error is not so much an error as a judgment that the sample size is too small to detect the effect. When a p-value falls short of statistical significance (e.g., it exceeds 5%), what we are really saying is "effect not proven." It could be that a larger sample would yield a smaller p-value.

The basic function of significance tests (also called *hypothesis tests*) is to protect against being fooled by random chance; thus they are typically structured to minimize Type 1 errors.

Data Science and P-Values

The work that data scientists do is typically not destined for publication in scientific journals, so the debate over the value of a p-value is somewhat academic. For a data scientist, a p-value is a useful metric in situations where you want to know whether a model result that appears interesting and useful is within the range of normal chance variability. As a decision tool in an experiment, a p-value should not be considered controlling, but merely another point of information bearing on a decision. For example, p-values are sometimes used as intermediate inputs in some statistical or machine learning models — a feature night be included in or excluded from a model depending on its p-value.

KEY IDEAS

- Significance tests are used to determine whether an observed effect is within the range of chance variation for a null hypothesis model.
- The p-value is the probability that results as extreme as the observed results might occur, given a null hypothesis model.
- The alpha value is the threshold of "unusualness" in a null hypothesis chance model.
- Significance testing has been much more relevant for formal reporting of research than for data science (but has been fading recently, even for the former).

t-Tests

There are numerous types of significance tests, depending on whether the data comprises count data or measured data, how many samples there are, and what's being measured. A very common one is the *t-test*, named after Student's t-distribution, originally developed by W. S. Gossett to approximate the distribution of a single sample mean (see "Student's t-Distribution").

KEY TERMS

Test statistic

A metric for the difference or effect of interest.

t-statistic

A standardized version of the test statistic.

t-distribution

A reference distribution (in this case derived from the null hypothesis), to which the observed t-statistic can be compared.

KEY IDEAS

- Before the advent of computers, resampling tests were not practical and statisticians used standard reference distributions.
- A test statistic could then be standardized and compared to the reference distribution.
- One such widely used standardized statistic is the t-statistic.

In a resampling mode, we structure the solution to reflect the observed data and the hypothesis to be tested, not worrying about whether the data is numeric or binary, sample sizes are balanced or not, sample variances, or a variety of other factors. In the formula world, many variations present themselves, and they can be bewildering. Statisticians need to navigate that world and learn its map, but data scientists do not — they are typically not in the business of sweating the details of hypothesis tests and confidence intervals the way a researcher preparing a paper for presentation might.

Multiple Testing

As we've mentioned previously, there is a saying in statistics: "torture the data long enough, and it will confess." This means that if you look at the data through enough different perspectives, and ask enough questions, you can almost invariably find a statistically significant effect.

KEY TERMS

Type 1 error
Mistakenly concluding that an effect is statistically significant.

False discovery rate
Across multiple tests, the rate of making a Type 1 error.

Adjustment of p-values
Accounting for doing multiple tests on the same data.

Overfitting
Fitting the noise.

Degrees of Freedom

In the documentation and settings to many statistical tests, you will see reference to "degrees of freedom." The concept is applied to statistics calculated from sample data, and refers to the number of values free to vary. For example, if you know the mean for a sample of 10 values, and you also know 9 of the values, you also know the 10th value. Only 9 are free to vary.

KEY TERMS

n or sample size
The number of observations (also called *rows* or *records*) in the data.

d.f.
Degrees of freedom.

KEY IDEAS

- The number of degrees of freedom (d.f.) forms part of the calculation to standardize test statistics so they can be compared to reference distributions (t-distribution, F-distribution, etc.).
- The concept of degrees of freedom lies behind the factoring of categorical variables into $n - 1$ indicator or dummy variables when doing a regression (to avoid multicollinearity).

ANOVA

Suppose that, instead of an A/B test, we had a comparison of multiple groups, say A-B-C-D, each with numeric data. The statistical procedure that tests for a statistically significant difference among the groups is called *analysis of variance*, or *ANOVA*.

KEY TERMS FOR ANOVA

Pairwise comparison

A hypothesis test (e.g., of means) between two groups among multiple groups.

Omnibus test

A single hypothesis test of the overall variance among multiple group means.

Decomposition of variance

Separation of components. contributing to an individual value (e.g., from the overall average, from a treatment mean, and from a residual error).

F-statistic

A standardized statistic that measures the extent to which differences among group means exceeds what might be expected in a chance model.

SS

"Sum of squares," referring to deviations from some average value.

F-Statistic

Just like the t-test can be used instead of a permutation test for comparing the mean of two groups, there is a statistical test for ANOVA based on the *F-statistic*. The F-statistic is based on the ratio of the variance across group means (i.e., the treatment effect) to the variance due to residual error. The higher this ratio, the more statistically significant the result. If the data follows a normal distribution, then statistical theory dictates that the statistic should have a certain distribution. Based on this, it is possible to compute a p-value.

DECOMPOSITION OF VARIANCE

Observed values in a data set can be considered sums of different components. For any observed data value within a data set, we can break it down into the grand average, the treatment effect, and the residual error. We call this a "decomposition of variance."

1. Start with grand average (173.75 for web page stickiness data).
2. Add treatment effect, which might be negative (independent variable = web page).
3. Add residual error, which might be negative.

Thus, the decomposition of the variance for the top-left value in the A-B-C-D test table is as follows:

1. Start with grand average: 173.75
2. Add treatment (group) effect: –1.75 (172 – 173.75).
3. Add residual: –8 (164 – 172).
4. Equals: 164.

Two-Way ANOVA

The A-B-C-D test just described is a "one-way" ANOVA, in which we have one factor (group) that is varying. We could have a second factor involved — say, "weekend versus weekday" — with data collected on each combination (group A weekend, group A weekday, group B weekend, etc.). This would be a "two-way ANOVA," and we would handle it in similar fashion to the one-way ANOVA by identifying the "interaction effect." After identifying the grand average effect, and the treatment effect, we then separate the weekend and the weekday observations for each group, and find the difference between the averages for those subsets and the treatment average.

You can see that ANOVA, then two-way ANOVA, are the first steps on the road toward a full statistical model, such as regression and logistic regression, in which multiple factors and their effects can be modeled (see Chapter 4).

KEY IDEAS

- ANOVA is a statistical proecdure for analyzing the results of an experiment with multiple groups.
- It is the extension of similar procedures for the A/B test, used to assess whether the overall variation among groups is within the range of chance variation.
- A useful outcome of an ANOVA is the identification of variance components associated with group treatments, interaction effects, and errors.

Chi-Square Test

Web testing often goes beyond A/B testing and tests multiple treatments at once. The chi-square test is used with count data to test how well it fits some expected distribution. The most common use of the *chi-square* statistic in statistical practice is with $r \times c$ contingency tables, to assess whether the null hypothesis of independence among variables is reasonable.

The chi-square test was originally developed by Karl Pearson in 1900. The term "chi" comes from the greek letter ξ used by Pearson in the article.

KEY TERMS

Chi-square statistic
A measure of the extent to which some observed data departs from expectation.

Expectation or expected
How we would expect the data to turn out under some assumption, typically the null hypothesis.

d.f.
Degrees of freedom.

NOTE

$r \times c$ means "rows by columns" — a 2×3 table has two rows and three columns.

Chi-Square Test: A Resampling Approach

Suppose you are testing three different headlines — A, B, and C — and you run them each on 1,000 visitors, with the results shown in Table 3-4.

Table 3-4. Web testing results of three different headlines

	Headline A	Headline B	Headline C
Click	14	8	12
No-click	986	992	988

The *Pearson residual* is defined as:

$$R = \frac{\text{Observed} - \text{Expected}}{\sqrt{\text{Expected}}}$$

The chi-squared statistic is defined as the sum of the squared Pearson residuals:

$$\xi = \sum_{i}^{r}\sum_{j}^{c} R^2$$

Chi-Squared Test: Statistical Theory

Asymptotic statistical theory shows that the distribution of the chi-squared statistic can be approximated by a *chi-square distribution*. The appropriate standard chi-square distribution is determined by the *degrees of freedom* (see "Degrees of Freedom"). For a contingency table, the degrees of freedom are related to the number of rows (r) and columns (s) as follows:

$$\text{degrees of freedom} = (r-1) \times (c-1)$$

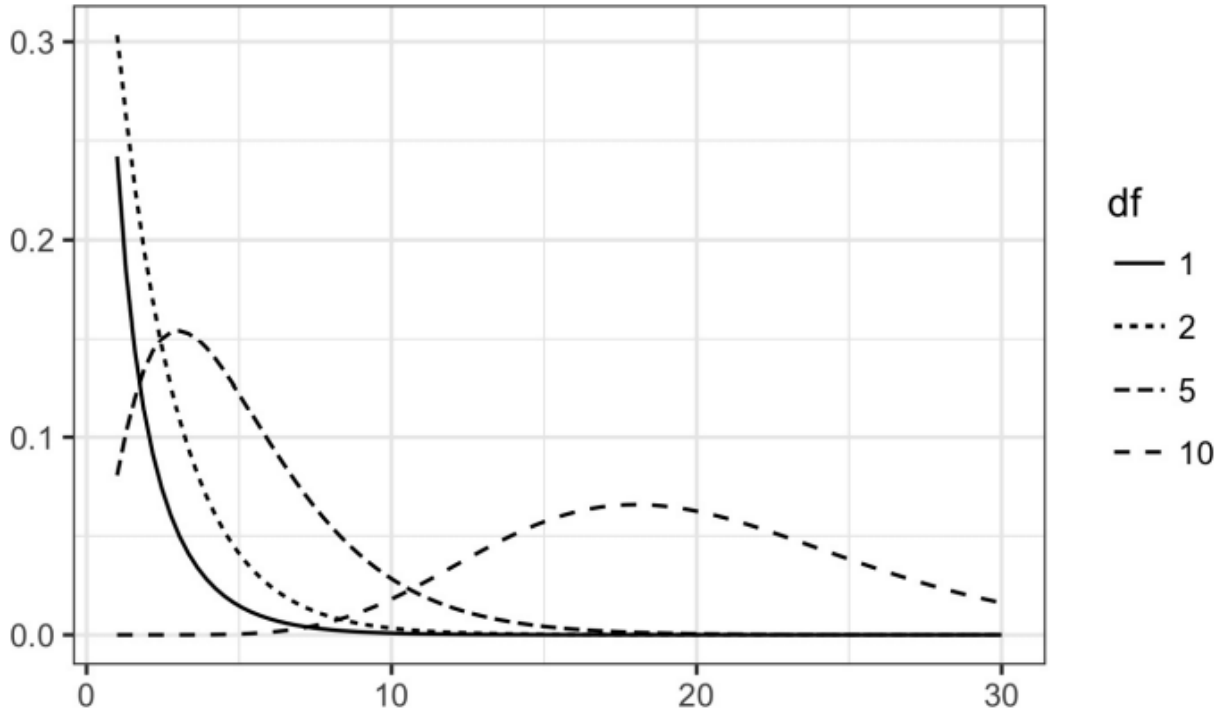


Figure 3-7. Chi-square distribution with various degrees of freedom (probability on y-axis, value of chi-square statistic on x-axis)

Fisher's Exact Test

The chi-square distribution is a good approximation of the shuffled resampling test just described, except when counts are extremely low (single digits, especially five or fewer). In such cases, the resampling procedure will yield more accurate p-values. In fact, most statistical software has a procedure to actually enumerate *all* the possible rearrangements (permutations) that can occur, tabulate their frequencies, and determine exactly how extreme the observed result is. This is called *Fisher's exact test* after the great statistician R. A. Fisher. R code for Fisher's exact test is simple in its basic form:

DETECTING SCIENTIFIC FRAUD

An interesting example is provided by Tufts University researcher Thereza Imanishi-Kari, who was accused in 1991 of fabricating data in her research. Congressman John Dingell became involved, and the case eventually led to the resignation of her colleague, David Baltimore, from the presidency of Rockefeller University.

Imanishi-Kari was ultimately exonerated after a lengthy proceeding. However, one element in the case rested on statistical evidence regarding the expected distribution of digits in her laboratory data, where each observation had many digits. Investigators focused on the *interior* digits, which would be expected to follow a *uniform random* distribution. That is, they would occur randomly, with each digit having equal probability of occurring (the lead digit might be predominantly one value, and the final digits might be affected by rounding). Table 3-7 lists the frequencies of interior digits from the actual data in the case.

Power and Sample Size

If you run a web test, how do you decide how long it should run (i.e., how many impressions per treatment are needed)? Despite what you may read in many guides to web testing on the web, there is no good general guidance — it depends, mainly, on the frequency with which the desired goal is attained.

KEY TERMS

Effect size
: The minimum size of the effect that you hope to be able to detect in a statistical test, such as "a 20% improvement in click rates".

Power
: The probability of detecting a given effect size with a given sample size.

Significance level
: The statistical significance level at which the test will be conducted.

KEY IDEAS

- Finding out how big a sample size you need requires thinking ahead to the statistical test you plan to conduct.
- You must specify the minimum size of the effect that you want to detect.
- You must also specify the required probability of detecting that effect size (power).
- Finally, you must specify the significance level (alpha) at which the test will be conducted.

Simple Linear Regression

Simple linear regression models the relationship between the magnitude of one variable and that of a second — for example, as *X* increases, *Y* also increases. Or as *X* increases, *Y* decreases.[1] Correlation is another way to measure how two variables are related: see the section "Correlation". The difference is that while correlation measures the strength of an association between two variables, regression quantifies the nature of the relationship.

KEY TERMS FOR SIMPLE LINEAR REGRESSION

Response

The variable we are trying to predict.

Synonyms

dependent variable, Y-variable, target, outcome

Independent variable

The variable used to predict the response.

Synonyms

independent variable, X-variable, feature, attribute

Record

The vector of predictor and outcome values for a specific individual or case.

Synonyms

row, case, instance, example

Intercept

The intercept of the regression line — that is, the predicted value when $X = 0$.

Synonyms

b_0, β_0

Regression coefficient

The slope of the regression line.

Synonyms

slope, b_1, β_1, parameter estimates, weights

Fitted values

The estimates $\hat{Y}_i$ obtained from the regression line.

Synonyms

predicted values

Residuals

The difference between the observed values and the fitted values.

Synonyms

errors

Least squares

The method of fitting a regression by minimizing the sum of squared residuals.

Synonyms

ordinary least squares

Multiple Linear Regression

When there are multiple predictors, the equation is simply extended to accommodate them:

$$Y = b_0 + b_1X_1 + b_2X_2 + \ ... + b_pX_p + e$$

Instead of a line, we now have a linear model — the relationship between each coefficient and its variable (feature) is linear.

KEY TERMS FOR MULTIPLE LINEAR REGRESSION

Root mean squared error

The square root of the average squared error of the regression (this is the most widely used metric to compare regression models).

Synonyms

RMSE

Residual standard error

The same as the root mean squared error, but adjusted for degrees of freedom.

Synonyms

RSE

R-squared

The proportion of variance explained by the model, from 0 to 1.

Synonyms

coefficient of determination, R^2

t-statistic

The coefficient for a predictor, divided by the standard error of the coefficient, giving a metric to compare the importance of variables in the model.

Weighted regression

Regression with the records having different weights.

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1X_{1,i} + \hat{b}_2X_{2,i} + \ ... + \hat{b}_pX_{p,i}$$

Cross-Validation

Classic statistical regression metrics (R^2, F-statistics, and p-values) are all "in-sample" metrics — they are applied to the same data that was used to fit the model. Intuitively, you can see that it would make a lot of sense to set aside some of the original data, not use it to fit the model, and then apply the model to the set-aside (holdout) data to see how well it does. Normally, you would use a majority of the data to fit the model, and use a smaller portion to test the model.

This idea of "out-of-sample" validation is not new, but it did not really take hold until larger data sets became more prevalent; with a small data set, analysts typically want to use all the data and fit the best possible model.

Using a holdout sample, though, leaves you subject to some uncertainty that arises simply from variability in the small holdout sample. How different would the assessment be if you selected a different holdout sample?

Cross-validation extends the idea of a holdout sample to multiple sequential holdout samples. The algorithm for basic *k-fold cross-validation* is as follows:

1. Set aside *1/k* of the data as a holdout sample.
2. Train the model on the remaining data.
3. Apply (score) the model to the *1/k* holdout, and record needed model assessment metrics.
4. Restore the first *1/k* of the data, and set aside the next *1/k* (excluding any records that got picked the first time).
5. Repeat steps 2 and 3.
6. Repeat until each record has been used in the holdout portion.
7. Average or otherwise combine the model assessment metrics.

The division of the data into the training sample and the holdout sample is also called a *fold*.

AIC, BIC AND MALLOWS CP

The formula for AIC may seem a bit mysterious, but in fact it is based on asymptotic results in information theory. There are several variants to AIC:

- AICc: a version of AIC corrected for small sample sizes.
- BIC or Bayesian information criteria: similar to AIC with a stronger penalty for including additional variables to the model.
- Mallows Cp: A variant of AIC developed by Colin Mallows.

Data scientists generally do not need to worry about the differences among these in-sample metrics or the underlying theory behind them.

The coefficents in the weighted regression are slightly different from the original regression.

KEY IDEAS

- Multiple linear regression models the relationship between a response variable Y and multiple predictor variables $X_1, \ldots, X_p$.
- The most important metrics to evaluate a model are root mean squared error (RMSE) and R-squared (R^2).
- The standard error of the coefficients can be used to measure the reliability of a variable's contribution to a model.
- Stepwise regression is a way to automatically determine which variables should be included in the model.
- Weighted regression is used to give certain records more or less weight in fitting the equation.

Prediction Using Regression

The primary purpose of regression in data science is prediction. This is useful to keep in mind, since regression, being an old and established statistical method, comes with baggage that is more relevant to its traditional explanatory modeling role than to prediction.

KEY TERMS FOR PREDICTION USING REGRESSION

Prediction interval

An uncertainty interval around an individual predicted value.

Extrapolation

Extension of a model beyond the range of the data used to fit it.

PREDICTION INTERVAL OR CONFIDENCE INTERVAL?

A prediction interval pertains to uncertainty around a single value, while a confidence interval pertains to a mean or other statistic calculated from multiple values. Thus, a prediction interval will typically be much wider than a confidence interval for the same value. We model this individual value error in the bootstrap model by selecting an individual residual to tack on to the predicted value. Which should you use? That depends on the context and the purpose of the analysis, but, in general, data scientists are interested in specific individual predictions, so a prediction interval would be more appropriate. Using a confidence interval when you should be using a prediction interval will greatly underestimate the uncertainty in a given predicted value.

KEY IDEAS

- Extrapolation beyond the range of the data can lead to error.
- Confidence intervals quantify uncertainty around regression coefficients.
- Prediction intervals quantify uncertainty in individual predictions.
- Most software, R included, will produce prediction and confidence intervals in default or specified output, using formulas.
- The bootstrap can also be used; the interpretation and idea are the same.

Factor Variables in Regression

Factor variables, also termed *categorical* variables, take on a limited number of discrete values. For example, a loan purpose can be “debt consolidation,” “wedding,” “car,” and so on. The binary (yes/no) variable, also called an *indicator* variable, is a special case of a factor variable. Regression requires numerical inputs, so factor variables need to be recoded to use in the model. The most common approach is to convert a variable into a set of binary *dummy* variables.

KEY TERMS FOR FACTOR VARIABLES

Dummy variables

Binary 0–1 variables derived by recoding factor data for use in regression and other models.

Reference coding

The most common type of coding used by statisticians, in which one level of a factor is used as a reference and other factors are compared to that level.

Synonyms

treatment coding

One hot encoder

A common type of coding used in the machine learning community in which all factors levels are retained. While useful for certain machine learning algorithms, this approach is not appropriate for multiple linear regression.

Deviation coding

A type of coding that compares each level against the overall mean as opposed to the reference level.

Synonyms

sum contrasts

Interpreting the Regression Equation

In data science, the most important use of regression is to predict some dependent (outcome) variable. In some cases, however, gaining insight from the equation itself to understand the nature of the relationship between the predictors and the outcome can be of value. This section provides guidance on examining the regression equation and interpreting it.

KEY TERMS FOR INTERPRETING THE REGRESSION EQUATION

Correlated variables

When the predictor variables are highly correlated, it is difficult to interpret the individual coefficients.

Multicollinearity

When the predictor variables have perfect, or near-perfect, correlation, the regression can be unstable or impossible to compute.

Synonyms

collinearity

Confounding variables

An important predictor that, when omitted, leads to spurious relationships in a regression equation.

Main effects

The relationship between a predictor and the outcome variable, independent from other variables.

Interactions

An interdependent relationship between two or more predictors and the response.

MODEL SELECTION WITH INTERACTION TERMS

In problems involving many variables, it can be challenging to decide which interaction terms should be included in the model. Several different approaches are commonly taken:

- In some problems, prior knowledge and intuition can guide the choice of which interaction terms to include in the model.
- Stepwise selection (see "Model Selection and Stepwise Regression") can be used to sift through the various models.
- Penalized regression can automatically fit to a large set of possible interaction terms.
- Perhaps the most common approach is the use *tree models*, as well as their descendents, *random forest* and *gradient boosted trees*. This class of models automatically searches for optimal interaction terms; see "Tree Models".

KEY IDEAS

- Because of correlation between predictors, care must be taken in the interpretation of the coefficients in multiple linear regression.
- Multicollinearity can cause numerical instability in fitting the regression equation.
- A confounding variable is an important predictor that is omitted from a model and can lead to a regression equation with spurious relationships.
- An interaction term between two variables is needed if the relationship between the variables and the response is interdependent.

Polynomial and Spline Regression

The relationship between the response and a predictor variable is not necessarily linear. The response to the dose of a drug is often nonlinear: doubling the dosage generally doesn't lead to a doubled response. The demand for a product is not a linear function of marketing dollars spent since, at some point, demand is likely to be saturated. There are several ways that regression can be extended to capture these nonlinear effects.

KEY TERMS FOR NONLINEAR REGRESSION

Polynomial regression
Adds polynomial terms (squares, cubes, etc.) to a regression.

Spline regression
Fitting a smooth curve with a series of polynomial segments.

Knots
Values that separate spline segments.

Generalized additive models
Spline models with automated selection of knots.

Synonyms
GAM

NONLINEAR REGRESSION

When statisticians talk about *nonlinear regression*, they are referring to models that can't be fit using least squares. What kind of models are nonlinear? Essentially all models where the response cannot be expressed as a linear combination of the predictors or some transform of the predictors. Nonlinear regression models are harder and computationally more intensive to fit, since they require numerical optimization. For this reason, it is generally preferred to use a linear model if possible.