STATISTICS

The processing of data is the most crucial part of any Data Science strategy. When we talk about gaining insights from data, we're basically talking about exploring the possibilities. These possibilities in Data Science are referred to as statistical analysis.

Statistics is the heart of advanced machine learning algorithms in data science, identifying and converting data patterns into usable evidence. Statistics are used by data scientists to collect, assess, analyse, and derive conclusions from data, as well as to apply quantitative mathematical models to applicable variables.

There are two main types of statistics:

Descriptive statistics:

- It is used by researchers to report on population and samples.
- It gives summary descriptions of measurements taken about a group.

Descriptive statistics are a series of short descriptive coefficients that summarise a data set, which might be a representation of the complete population or a sample of the population.

Descriptive statistics include measurements of central tendency and measures of variability (spread). Central tendency metrics include the mean, median, and mode, whereas variability measures include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.A data set's properties are summarised or described using descriptive statistics.

- There are two types of measures in descriptive statistics: measures of central tendency and measures of variability (or spread).
- The centre of data collection is defined using central tendency measures.
- In short, it helps to describe and understand the features of a particular data set by using various samples and measures of the data.

Examples of descriptive statistics are:

- 5 point summary
- Box plot
- Correlation
- Central tendency
- Dispersion

Inferential statistics:

Inferential statistics assist you to arrive at conclusions and make predictions based on your data, whereas descriptive statistics outline the properties of a data collection.

There are two main use cases where you use inferential statistics.

- To make estimates about population.
- To draw conclusions about the populations.

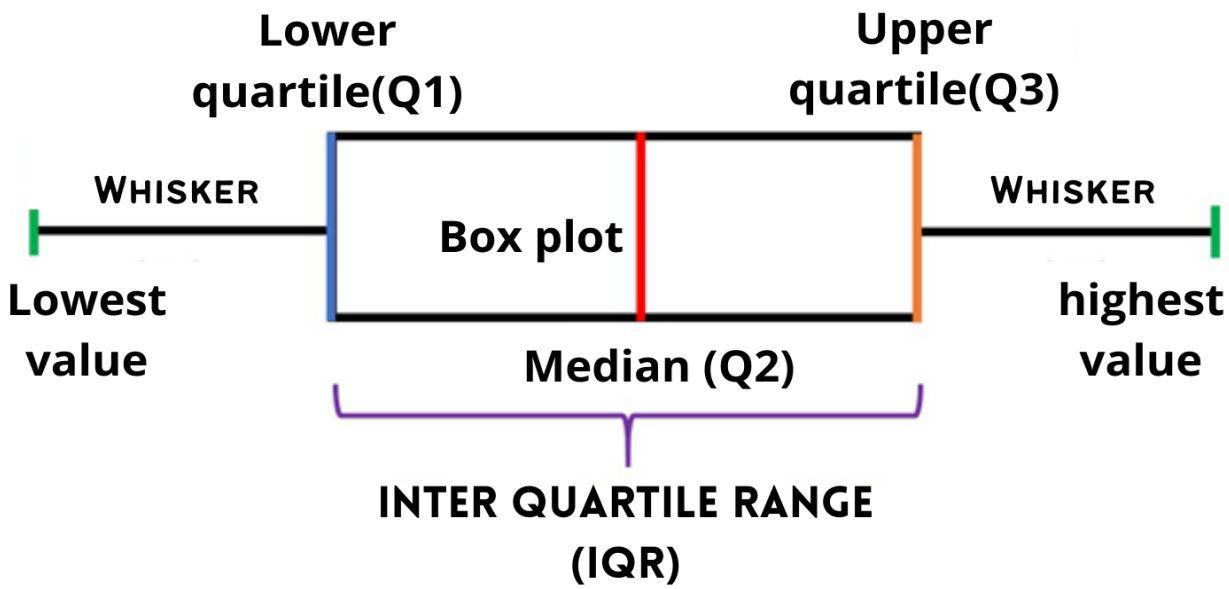
Box and Whiskers Plots and the 5 number summary

The box plot has all of the features such as Mean, Median, Lower quartile, Upper quartile, and extreme values, and it essentially allows you to see the data's significant qualities in a visual format.

5 number summary

It is a numerical representation of the box plot. It consists of:

- The Median (2nd Quartile)
- The 1st Quartile
- 3rd Quartile
- Maximum value
- Minimum value



Let's see how to draw a box plot

Suppose you have a data set

34, 18, 100, 27, 54, 52, 93, 59, 61, 87, 68, 85, 78, 82, 91

Step 1: You'll have to re-arrange the data set in ascending order.

18, 27, 34, 52, 54, 59, 61, 68, 78, 82, 85, 87, 91, 93, 100

Step 2: Find the below values from the data set.

18, 27, 34, 52, 54, 59, 61, 68, 78, 82, 85, 87, 91, 93, 100

Minimum value: 18

Maximum value: 100

Q1: 52

Median: 68

Q3: 87

One simple trick first find out the median then it'll be easy to find the 25^{th} and the 75^{th} percentile values.

Step 3: Now according to the data set we have, we might be guessing what if we have an even number of data sets.

2, 4, 5, 6, 7, 8, 9, 11, 19, 20

So now we have 10 total values and we cannot just split it into two halves

{2, 4, 5, 6,} 7, 8, {9, 11, 19, 20}

7+8 = 15

15 divided by 2 = 7.5

Hence the median is 7.5

2, 4, 5, 6, 7, 7.5, 8, 9, 11, 19, 20

Minimum value: 2

Maximum value: 20

Q1: 5

Median: 7.5

Q3: 11

Points to remember.

- Once you have calculated all these values you can easily plot the box plot.
- All of these values will be calculated in the background in python, we have to understand what happens in the background.
- Always place the Median line according to the median value which we get then it will make the box plot clear visually.
- To create the box plot use code {pd.boxplot(data)}

The Measure of Center Tendency:

The measure of central tendency is a statistical summary that indicates the dataset's center point or typical value. These measurements show where the majority of values in a distribution fall, as well as the distribution's center.

Mean:

The sum of the numbers' average values. A mean is a number that represents the center of the data, denoted by **μ** for the population mean and $\bar{\mathbf{X}}$ for the sample mean.

Example: Consider the following sequence of numbers: 10, 10, 20, 40, and 70.

The mean (sometimes known as the "average") is calculated by multiplying all of the numbers by the number of elements in the set: 10 + 10 + 20 + 40 + 70 / 5 = 30.

Therefore Mean = 30

Median:

The median is the value that divides the data into two equal sections, i.e. the number of terms on the right side equals the number of terms on the left side. It should only be used after the data has been organized in ascending or descending order.

For example, arranging the group from lowest to highest and finding the precise midpoint yields the median. Simply said, the median is the number in the middle: 20.

Therefore Median = 20

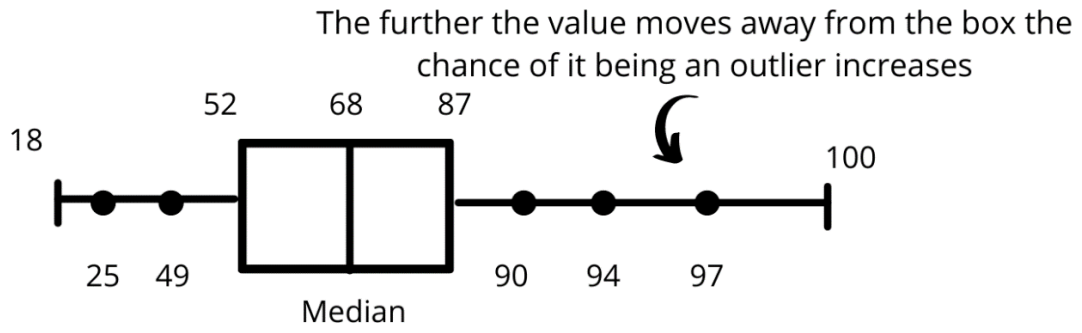
Mode: The mode is the value that appears the most times in the data collection.

Example: The mode of {4, 2, 4, 3, 2, and 2} is 2 because it occurs three times, which is more than any other number.

Outlier

The values that are lying on the whiskers are known as outliers

Whenever the value is going away from the box the probability that a particular or an observation is an outlier keeps increasing.



For example:

If I want to fill the data set

5, 5, 5, 5, 5, __, 5

Which value should I need to put in the blank? It'll be 5

And if the data set consists of outliers like

5, 5, 5, 5, 5, __, 5, 2000

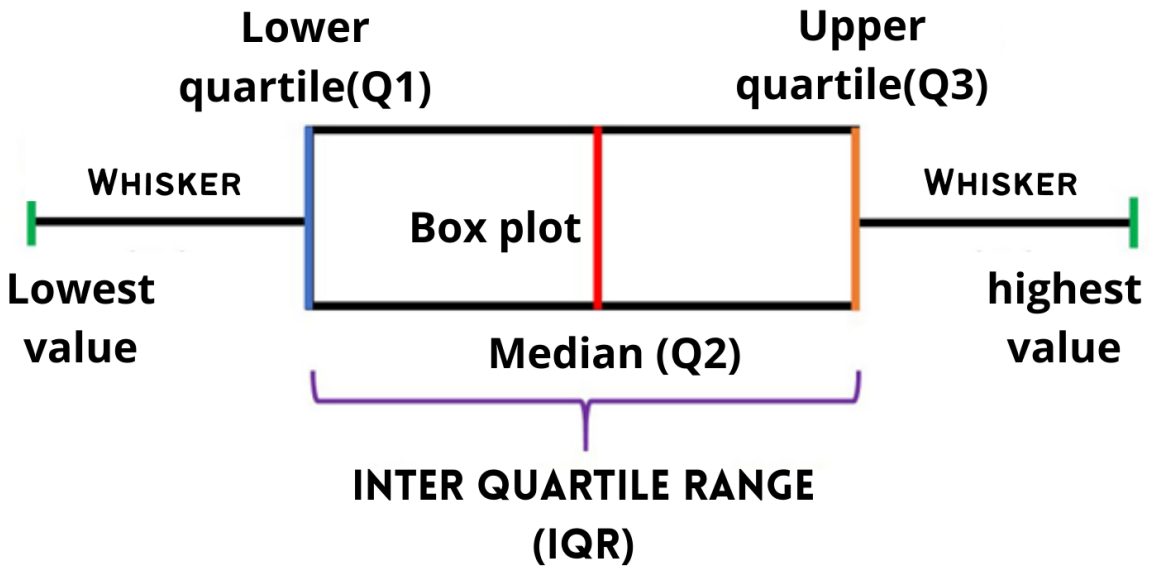
Here the value 2000 is an outlier it'll affect our overall true center and make out data bias. Hence it might be better to drop such values.

Methods for determining an outlier:

- Sorting the data and looking for extreme values
- By data visualizing (Boxplot, Scatterplot, etc.)
- IQR method
- Z-Score method

Inter Quartile Range (IQR): It represents the center of the data set

Formulae for IQR is Q3-Q1



Standard deviation: A standard deviation is a number that describes how far measurements for a group differ from the mean (mean or expected value).

$$\sigma = \sqrt[2]{\frac{\sum(Xi - \mu)^2}{N}}$$

Where σ = Population standard deviation

N = the size of the population

Xi = each value from the population

μ = the population mean

Example: Let us consider three groups {0, 0, 14, 14}, {0, 6, 8, 14} and {6, 6, 8, 8}

Their standard deviations are 7, 5, and 1 accordingly; however, the third group's standard deviation is substantially lower than the other two because its numbers are all near to 7. The standard deviation, in general, informs us how far the remainder of the numbers deviate from the average, and it uses the same units as the numbers themselves.

Variance: A variance is a measure of how much a group of data is spread out from its mean value, and it is a measure of how far data points differ from the mean.

$$\sigma^2 = \frac{1}{n}\sum(Xi - \mu)^2$$

Example: Let's say the heights (in mm) are {610, 450, 160, 420, 310}

Therefore the Mean is $\frac{610+450+160+420+310}{5} = 390$

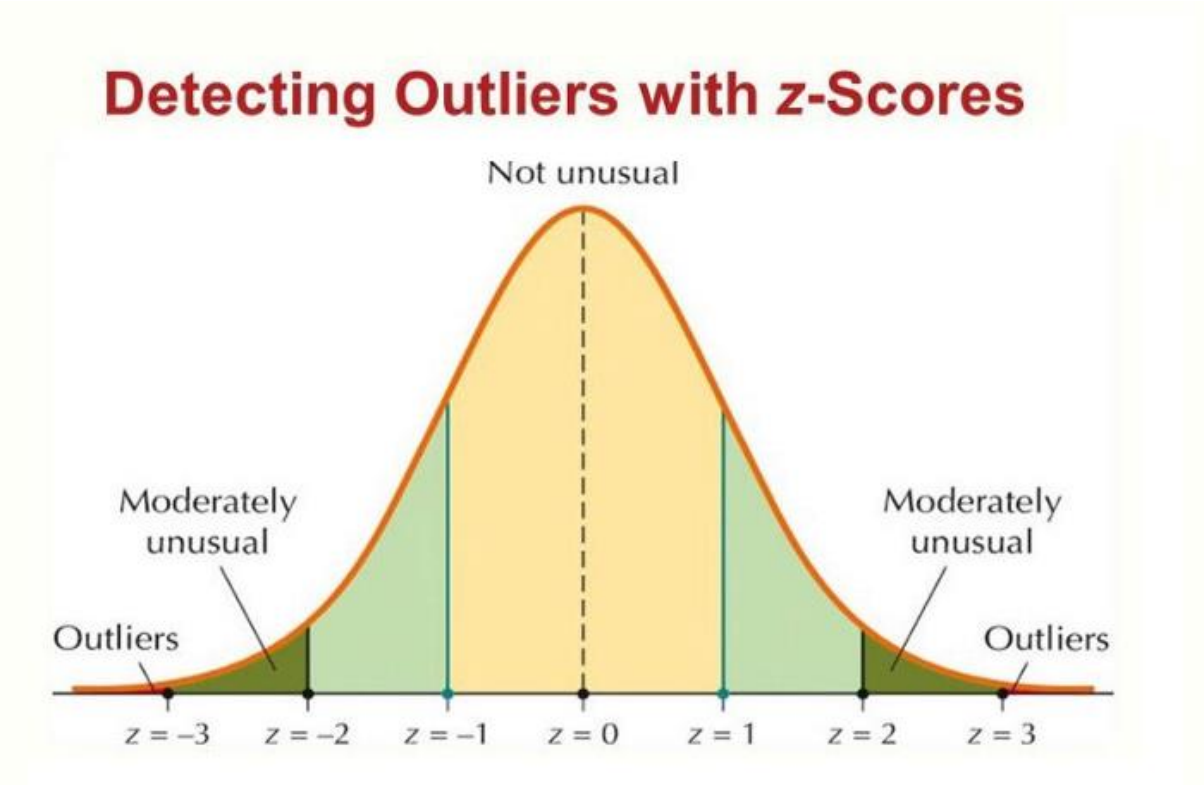
To calculate the variance we have to square the value of the standard deviation.

Range: In a data set the range is the difference between the highest and lowest values.

Formulae to calculate range is given by:

Range = Highest value – The lowest value

Z-Score:



Z-score is also known as standard score it gives us an idea of how far a data point is from the mean. It expresses how far an element deviates from the mean in standard deviations.

For example, the standard deviation of 2 indicates the value is 2 standard deviations away from the mean.

$$\boldsymbol{Z} = \frac{(\boldsymbol{X} - \mu)}{\boldsymbol{\sigma}}$$

Where Z = Z-Score

X = the value of the element

μ = the population mean

σ = the population standard deviation

Types of data

- Quantitative
- Qualitative

Quantitative data: The data that can be measured in numbers, deals with numbers that make sense to perform arithmetic calculations with. {Measured in amounts}

- Quantitative variables such as height, weight, and midterm score, etc.
- Continuous Quantitative: when the data is measured continuously. Example: temperature, sales.
- Discrete Quantitative: When the data is discrete or countable or any data which has definite values. Example: No of units, No of people, etc.

Categorical data: Refers to the values that place "Things" into different groups of categories. {Measured in categories}

- Categorical variables such as skin colour, breed of dog, Low, Medium, and high.
- Nominal categorical: Whenever we arrange data in some categories which cannot be compared. Example: beauty basis (Very beautiful, most beautiful etc.)
- Ordinal categorical: Whenever we arrange data in some categories which can be compared. Example: Grade (A, B, C), etc.

Data Visualization

Since the purpose of a data scientist is to provide insights into the data.

The graphical representation of information and data is known as data visualization, basically making it easy for the human brain to imagine the data. Using visual features like as charts, graphs, and maps, data visualization tools make it simple to explore and comprehend trends, outliers, and patterns in data.

- It is a very important step as it provides a quick and effective way to communicate information.
- It increases the ability to improve the insights and helps us in making decisions faster.
- It also helps people to concentrate more

We have different libraries in python for data visualization such as

Pandas: Panel data

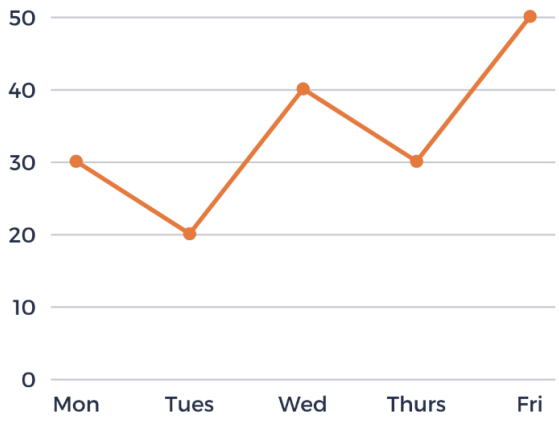
Numpy: Numerical python

Matplotlib: 2D visualization

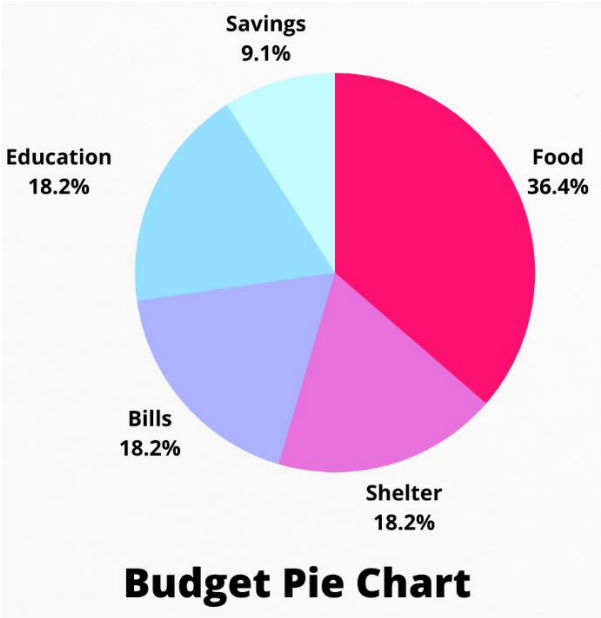
Seaborn: 3D visualization

Different types of data visualization graphs which we commonly use are:

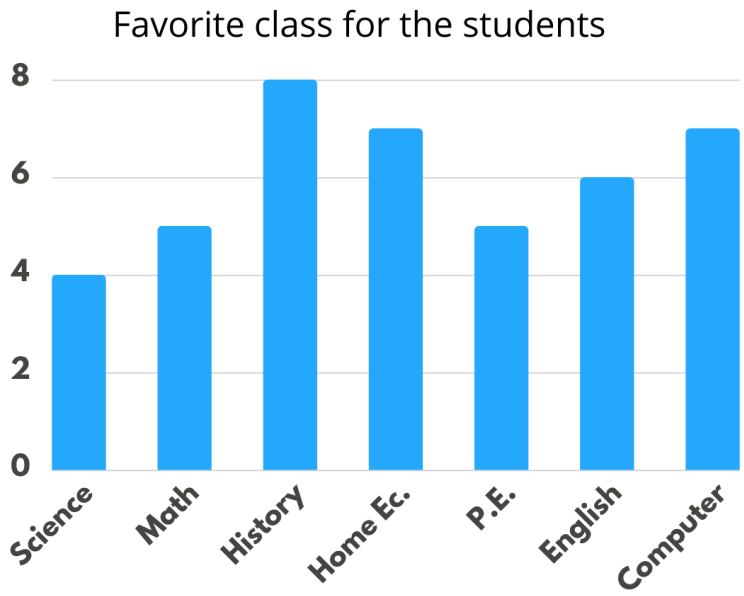
- **Line Chart**: It demonstrates how a single or several lines depicting various variables rise or decrease when the other variable increases or declines. Quantitative variables are the dependent variables in this case.
  It is basically used for time series data.
  Example: Share Markets, Daily Covid cases, etc.



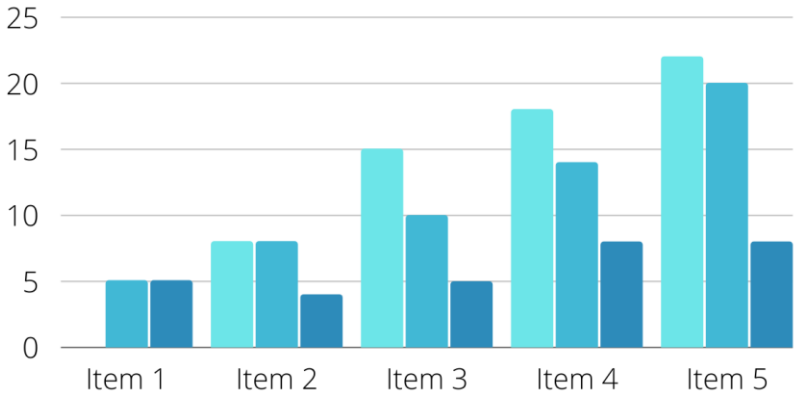
- **Pie chart**: It's a circular graph with slices. The larger the slice, the larger the slice's fraction of the data set. It basically signifies how much of each category in the data is represented by a percentage. A pie chart can be used if there is just one category.
  Used to evaluate the share of each category.
  Example: Percentage of water in the world.



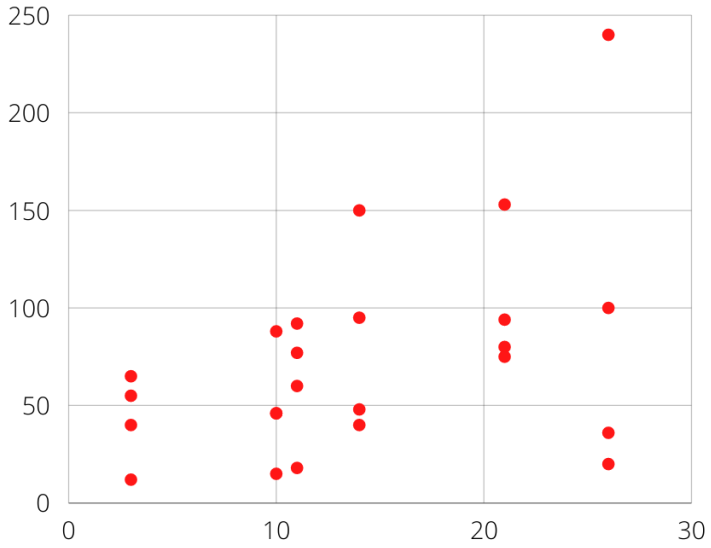
- **Bar graph**: The bar graph is a chart or graph that uses rectangular bars with heights or lengths proportionate to the values they represent to depict categorical data. The bars can be plotted vertically or horizontally.
  Used to compare different categories.



- **Histogram**: A histogram is a graphical representation that divides a set of data points into ranges defined by the user.
  Used for continuous data and reveals the distribution.



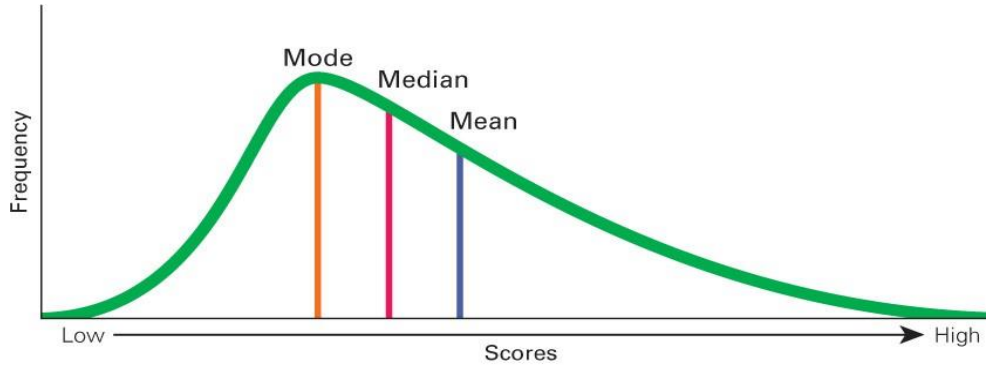
- **Scatterplots**: A scatter plot (also known as a scatter graph, scatter chart, scattergram, or scatter diagram) is a form of plot or mathematical diagram that uses Cartesian coordinates to depict values for two or more variables for a collection of data.
  Used to evaluate the relationship between two variables.



Skewness: Skewness is a measure of how far a random variable's probability distribution deviates from the normal distribution. The probability distribution with no skewness is known as the normal distribution.

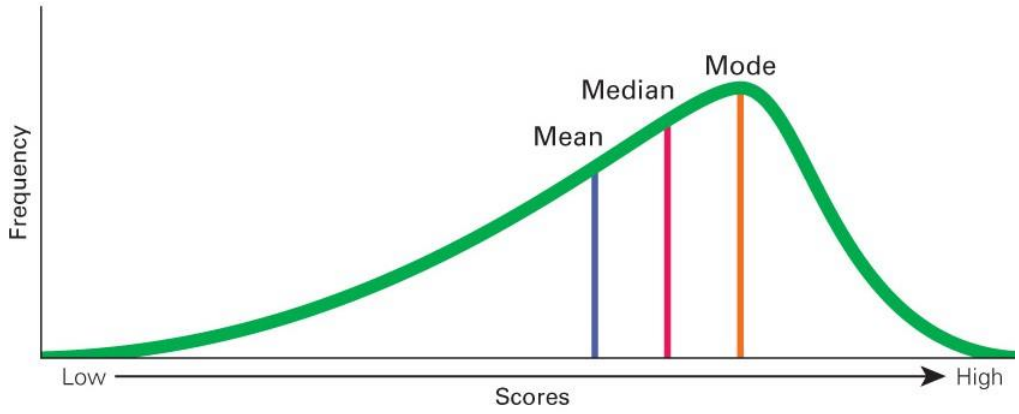
Types of skewness:

Right skewed: This is also called positive skewness, in right-skewed distribution data falls to the right or positive side of the graph's peak. The right side tail is longer than the left side. Data skewed to right is usually a result of a lower boundary of the data set.



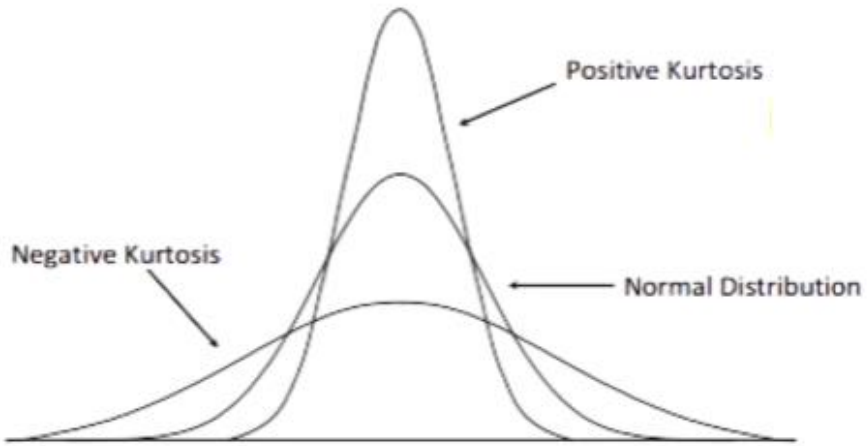
(a) Right-skewed distribution

Left skewed: It is also called negative skewness, in left-skewed distribution, data falls to the left side of the graph's peak, It has a long left tail and is also known as negatively skewed distribution. It is because the long tail is in the negative direction on the number line. The mean is also to the left of the peak



(b) Left-skewed distribution

Kurtosis: It is a measure of whether the data is heavily tailed or light-tailed relative to the normal distribution, Data sets with high kurtosis tend to have heavy tails or outliers.



Covariance: Covariance is a measure of the connection between two random variables. It is expressed in signs, with the units of the two variables ranging from positive to negative.

Positive covariance – It shows two variables moving in the same directions.

Negative covariance – It shows two variables moving in an inverse direction.

Covariance is a metric for determining how much a variable fluctuates at random.

The covariance is a product of the two variables' units. Covariance has a value between -∞ and +∞. The covariance of two variables (x and y) can be represented by cov(x,y). E[x] is the

Expected value or also called as a means of sample 'x'.

Formulae:

$$\boldsymbol{cov(x,y)} = \frac{\sum(\boldsymbol{Xi} - \bar{\boldsymbol{X}})(\boldsymbol{Yi} - \bar{Y})}{\boldsymbol{n}}$$

Where $\bar{\boldsymbol{X}}$ = sample mean of x

$\bar{Y}$= sample mean of y

Xi and Yi are the values of x and y for the I th record in the sample

N = number of records in the sample

The numerator is the product of the amount of variance in x and the amount of variance in y. A unit of x multiplied by a unit of y is the unit of covariance. As a result, if we change the unit of variables, the covariance will change, but the sign will not.

If it is positive, both variables will fluctuate in the same direction; if it is negative, they will vary in the opposite direction.

Correlation: Correlation is a normalized version of covariance and is defined as the correlation of two variables. The coefficients of correlation are usually between -1 and 1. Pearson's correlation coefficient is another name for the correlation coefficient. You will only learn about the direction of Covariance when you read about it, which is insufficient to fully appreciate the relationship. As a result, the covariance is divided by the x and y standard deviations.

$$correlation = \frac{cov(x, y)}{\sigma x * \sigma y}$$

Pearson's correlation coefficient is given by

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2] * [n\sum y^2 - (\sum y)^2]}}$$

Where, **n** = Number of values or elements

$\sum \mathbf{x}$ = Sum of 1st values list

$\sum \mathbf{y}$ = Sum of 2nd values list

$\sum \mathbf{xy}$ = Sum of the product of 1st and 2nd values

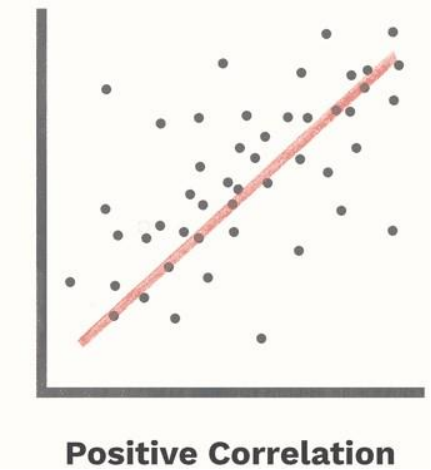
$\sum \mathbf{x^2}$ = Sum of squares of 1st values

$\sum \mathbf{y^2}$ = Sum of squares of 2nd values

Positive correlation: The value of one variable rises linearly as the value of another rise. This suggests that the two variables have a similar relationship. In this situation, the correlation coefficient would be positive, or 1.

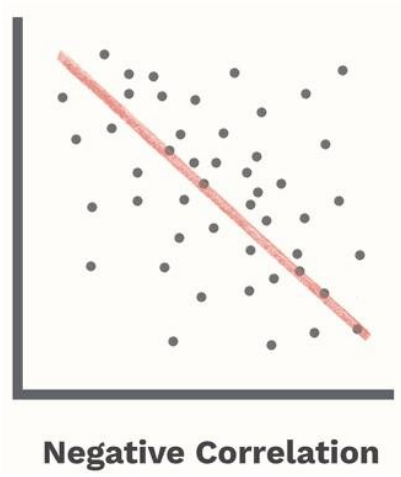
The strength of the correlation is understood as how close the value is to 1

Example: If we have 2 variables with correlations such as r = 0.8 and r = 0.7, Then we can say that the variable with 0.8 has more strong relation with the target variable.

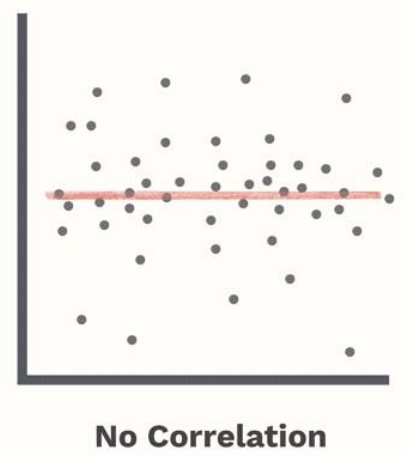


Positive Correlation

Negative correlation: When the values of one variable fall while the values of the other variable rise. The correlation coefficient would be negative in that situation.



No relation or Zero relation: Another scenario is when there is no clear relationship between two variables.



Thumb rule: Any relationship with a magnitude of r greater than 0.75 can be considered to be a strong correlation.

E.g.: -0.84 is a strong Negative correlation and 0.90 is a strong positive correlation.

Difference between Covariance and Correlation

- Correlation is simply a normalized form of covariance. It is obviously important to be precise with language when discussing the two, but conceptually they are almost identical.
- The value of the correlation coefficient ranges from [-1 – 1]. -1 is indicated for a negative relationship. 1 means a positive relationship. 0 means no relationship.

Probability

It denotes the probability of an event occurring in a random experiment. Probability is expressed as a number of 0s and 1s, with 0 denoting impossibility and 1 denoting possibility. When the likelihood of an event is higher, it means that the event is more likely to occur.

Formulae to calculate the probability

$$\boldsymbol{P(A) = \frac{Number\ of\ favorable\ outcomes\ to\ A}{Total\ number\ of\ outcomes}}$$

Example:

Imagine you are tossing a coin and what is the probability of getting head?

Then you'll calculate the probability

There are 2 outcomes that can occur head or tails

Number of favorable outcomes is 1 (head)

Therefore:

$$P(Head) = \frac{Number\ of\ favorable\ outcome(Head)}{Total\ number\ of\ outcomes}$$

$$P(Head) = \frac{1}{2} = 50\%$$

Random experiment: A random experiment in probability is one in which the outcome cannot be predicted with certainty.

For example, the outcome of a coin toss is uncertain, hence it is a random experiment.

Sample space: Sample space refers to the collection of all possible outcomes of an experiment.

For example, while throwing dice the top side can be any one of 1, 2, 3, 4, 5, and 6

Hence here S = {1, 2, 3, 4, 5, 6} and n(S) = 6

Multiplicative Theorem of Probability

- For Independent events

The likelihood of two independent events occurring at the same time is given by the product of their individual probabilities, according to the theorem.

$$\boldsymbol{P(A\ and\ B) = P(A) * P(B)}$$

$$\boldsymbol{P(AB) = P(A) * P(B)}$$

- For Dependent

Given that the first event has occurred, the likelihood of both A and B occurring at the same time is equal to the product of their probabilities. This is known as the Probability Multiplication Theorem.

$$\boldsymbol{P(A|B) = \frac{P(A \cap B)}{P(B)}}$$

$$\boldsymbol{P(B|A) = \frac{P(B \cap A)}{P(A)}}$$

Example

Question: An urn contains 20 red and 10 blue balls. Two balls are drawn from a bag one after the other without replacement. What is the probability that both the balls are drawn are red?

Solution:

Let A and B denote the events that the first and the second balls are drawn are red balls. We have to find P(A∩B) or P(AB).

P(A) = P(red balls in first draw) = 20/30

Now, only 19 red balls and 10 blue balls are left in the bag. The probability of drawing a red ball in the second draw too is an example of conditional probability where the drawing of the second ball depends on the drawing of the first ball.

Hence Conditional probability of B on A will be,

P(B|A) = 19/29

By multiplication rule of probability,

P(A∩B) = P(A) × P(B|A)

$$P(B \cap A) = \frac{20}{30} * \frac{19}{29} = \frac{38}{87}$$

Bayes Theorem

As the name suggests this theorem was introduced by Thomas Bayes, It describes the probability of an event based on the prior knowledge of conditions that might be related to the event.

Formulae:

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

Where,

P(A|B) = the probability of event A occurring, given that event B has already occurred.

P(B|A) = the probability of event B occurring, given event A has occurred

P(A) = the probability of event A

P(B) = the probability of event B

Example: Suppose the weather is cloudy, now we need to know whether it would rain today, given the cloudiness of the day. Therefore you are supposed to calculate the probability of rainfall, given the evidence of previous days of cloudiness.

HYPOTHESIS TESTING

Hypothesis testing is a statistical approach for confirming a hypothesis about a population parameter. The analyst's approach is determined by the type of the data and the purpose of the study. It can be used to decide whether a statement concerning a population parameter's value should be rejected or not.

Hypothesis is usually belief (for example, I believe that the average salary of the data scientist has increased post covid)

There are two types of hypothesis testing.

- Alternate hypothesis (Ha)
- Null hypothesis (Ho)

Example:

A new teaching method is developed that is proved to be better than the current method.

Alternate hypothesis (Ha) – The new teaching is better

Null hypothesis (Ho) – The new teaching method is no better than the old one.

Example:

In order to boost sales, a new sales force bonus plan is being designed.

Alternate hypothesis (Ha) – The new bonus plan increases sales

Null hypothesis (Ho) – The new bonus plan does not increase sales

Example:

A new drug is developed with the goal of lowering blood pressure more than the existing one.

Alternate hypothesis (Ha) – It does

Null hypothesis (Ho) – It does not

How to test the hypothesis

Example:

I believe that the average salary of the data scientist post Covid is $1, 00,000

Case 1:

$\bar{X}$ (Avg salary of the data scientist) = $90,000

This deviation might be due to random variability, Hence we can say our belief is correct as the difference is not that significant.

Case 2:

$\bar{X}$ (Avg salary of the data scientist) = $1, 10,000

This deviation might be due to random variability, Hence we can say our belief is correct as the difference is not that significant.

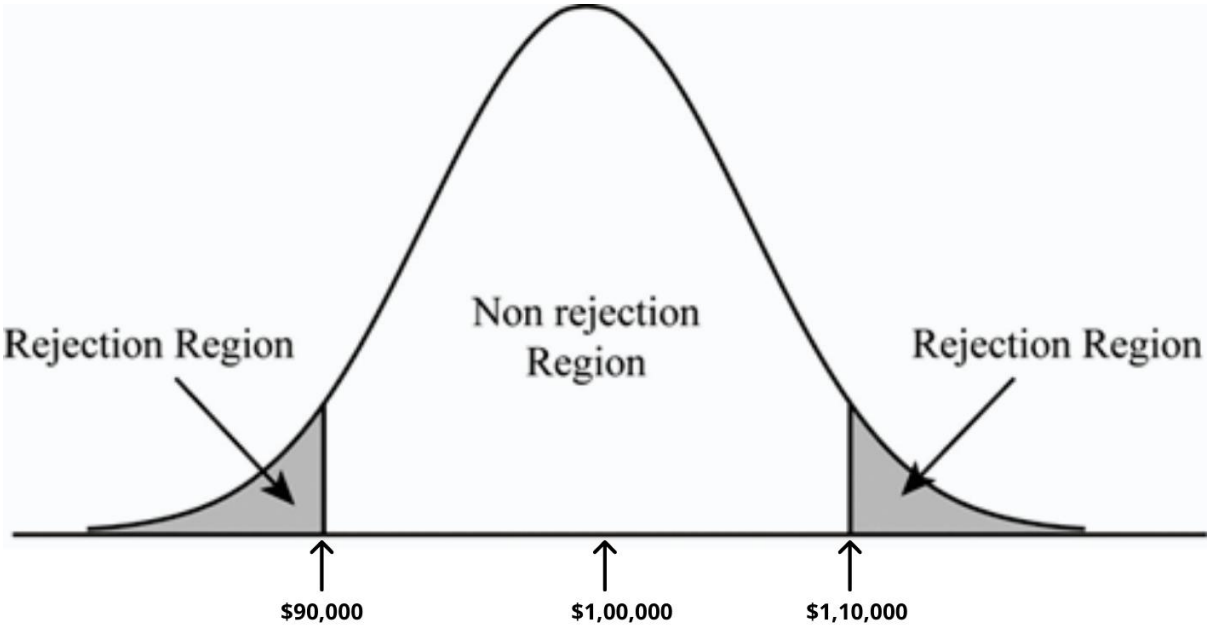
Case 3:

$\bar{X}$ (Avg salary of the data scientist) = $1, 25,000

Here the difference is significant, hence we can say that our hypothesis is a null hypothesis.

Level of significance: If the data falls within this range, the hypothesis is accepted; if the value falls outside of this range, the hypothesis is rejected.

It is denoted by α



Confidence Interval:

The degree of uncertainty or certainty in a sampling process is measured by confidence intervals. They can use any number of confidence levels, with a 95 percent or 99 percent confidence level being the most prevalent. Statistical tools such as the t-test are used to calculate confidence intervals.

Example: Assume you are given time to get to the office by 8:00 a.m. Employees who arrive between 7.55 and 8.05 a.m. are the most loyal, while those who arrive later receive a half-day pay decrease.

A level of significance is the chance of rejecting the null hypothesis when it is true, whereas a confidence interval is a range of values derived from the sample data so that the population can be estimated.

Z-TEST

The z-test is a statistical test that is used to see if two population means are different when the variances are known and the sample size is big. In order to execute an accurate z-test, the test statistic is expected to have a nuisance parameter, and the normal distribution, such as standard deviation, should be known. In the z-test, the standard deviation is expected to be known.

When the sample size is more than 30, the Z-test is employed. Considering the central limit theorem

When the number of samples is great enough, the samples are thought to be roughly regularly distributed.

Formulae

$$z = (x - \mu) / \sigma$$

Where:

- X = Individual data value
- μ = Mean of population
- σ = Standard deviation of the population

Example:

In academic settings, z-scores are used to analyze how well a student's score compares to the mean score on a given exam.

Let's say the score of entrance exams in a college is roughly normally distributed with a mean of 72 and a standard deviation of 5.

If a certain student received an 80 on the exam, we would calculate their z-score to be:

- $Z = (x-\mu)/\sigma$
- $Z = (80-72)/5$
- $Z = 1.6$

This means that this student received a score that was 1.6 standard deviations above the mean.

T-test

A t-test is an inferential statistic that determines if the means of two groups with similar characteristics differ significantly. A t-test is a hypothesis-testing tool that can be used to evaluate a population-based assumption. An analysis of variance must be employed to execute a test with three or more means.

To determine the statistical significance a t-test looks at the t-statistics, t-distribution, and the degrees of freedom.

How t-test is performed

Calculating a t-test requires three key data values.

- The mean values from each data set.
- The standard deviation of each group and
- The number of data values of each group.

Formulae:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{\Delta}}}$$

Where

$$s_{\bar{\Delta}} = \sqrt{\frac{s_1^2}{n1} + \frac{s_2^2}{n2}}$$

Where,

$\bar{x}$ = Mean of the first set of values

$\bar{x}_2$ = Mean of the second set of values

s_2 = Standard deviation of the second set of values

n_1 = Total number of values in the first set

n_2 = Total number of values in the Second set

ANOVA

ANOVA, or analysis of variance, is a statistical method for separating observed variance data into multiple components for use in further testing. Random factors have no statistical impact on the data set, whereas systematic factors do. In a regression analysis, ANOVA is used to determine the impact of independent factors on the dependent variable.

There are two types of ANOVA

- One way ANOVA or unidirectional ANOVA.
- Two-way ANOVA is an extension of one-way ANOVA.

Formulae:

$$F = \frac{MST}{MSE}$$

Where,

F = ANOVA coefficient

MST = Mean sum of squares due to treatment.

MSE = Mean sum of squares due to error.

Why do we use ANOVA

ANOVA examines the means of different samples to see how one or more factors influence the outcome.

Example

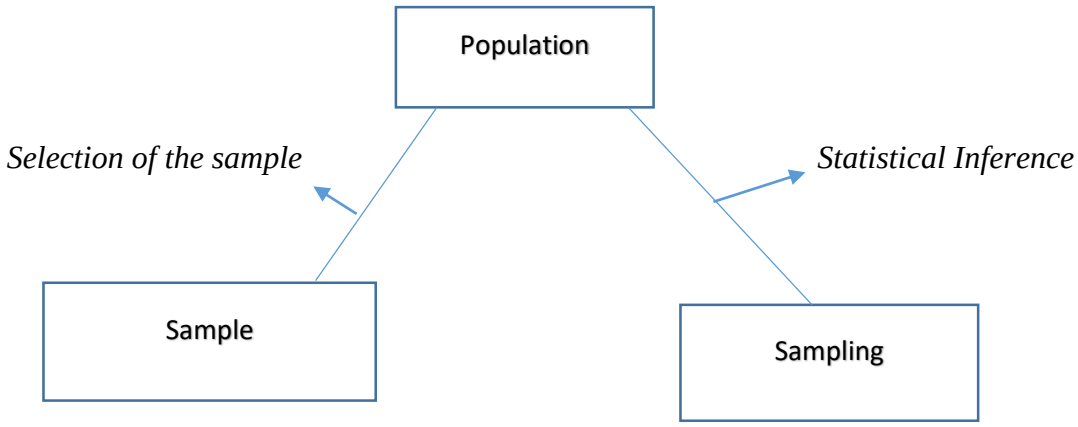
In a corporate setting, an R&D researcher might compare two distinct product development procedures to discover if one is more cost-effective than the other

Example

The combined effects of vaccination (vaccinated or unvaccinated) and health status (healthy or pre-existing condition) on the rate of flu infection in a population are being investigated.

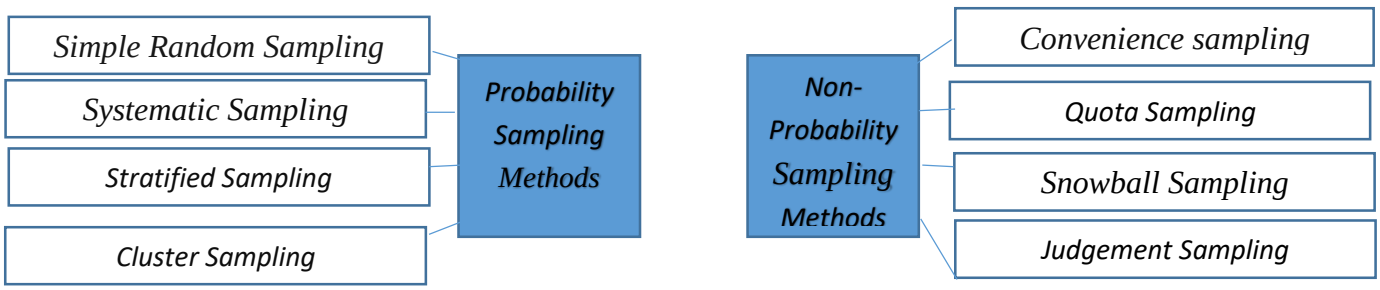
SAMPLING TECHNIQUES

Using sampling techniques, you can deduce information about a population without having to look at every single person. Sampling minimises the number of people in a study, lowering costs and reducing workload. If there is a problem with your sample, it will be reflected in the final result. Sampling makes obtaining high-quality data simple. Having a large enough sample size with enough power to identify actual association is balanced against sampling.



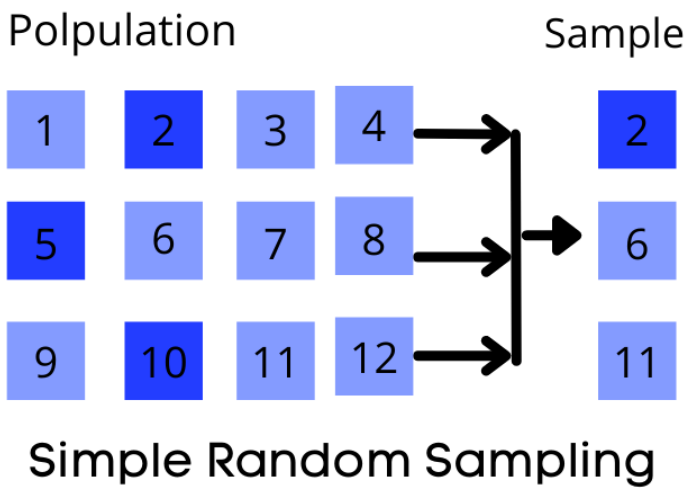
Sampling methods can be categorized into two type:

- Probability sampling methods
- Non- probability sampling methods



Probability Sampling Methods

Simple Random Sampling: In simple random sampling, every observation in the population has the same chance of being chosen, and any sample of a given size has the same chance of being chosen.

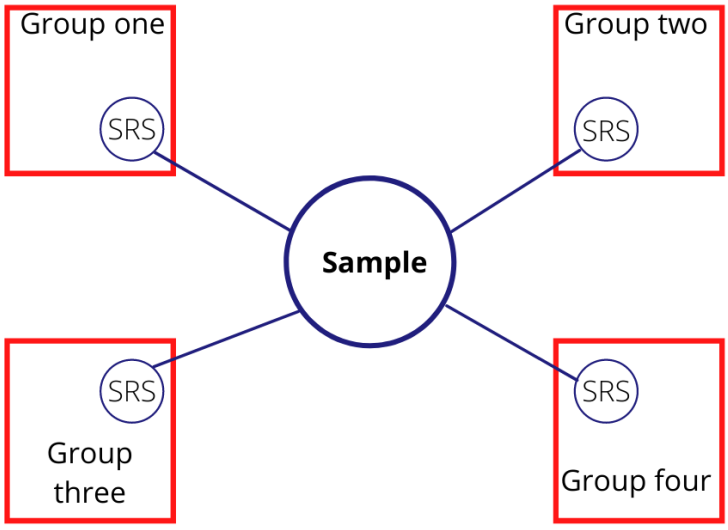


Simple Random Sampling

Example:

A total of 20 students were chosen at random from a class of 50. Every student has an equal chance of being chosen.

Stratified Random Sampling: In this method, the population is first divided into subgroups (or strata) who all share a similar characteristic. It's utilised when there's a chance that the measurement of interest will differ between subgroups, and we want to make sure that all of them are represented.



Stratified Random Sampling

Example: If there are three hospitals in a county, each with a different number of nurses (hospital A has 500 nurses, hospital B has 1000, and hospital C has 2000), it would be appropriate to choose the sample numbers from each hospital proportionally in a research of the health outcomes of nursing staff (For example, ten from hospital A, twenty from hospital B, and forty from hospital C). Nurses from hospitals A and B would be overrepresented in simple random sample, resulting in a more realistic and accurate estimate of nurses' health outcomes across the county.

Systematic Sampling: Sampling in a systematic manner the systematic sampling approach is used by researchers to select sample members of a community at regular periods. Simple random sampling is often more inconvenient than systematic sampling.

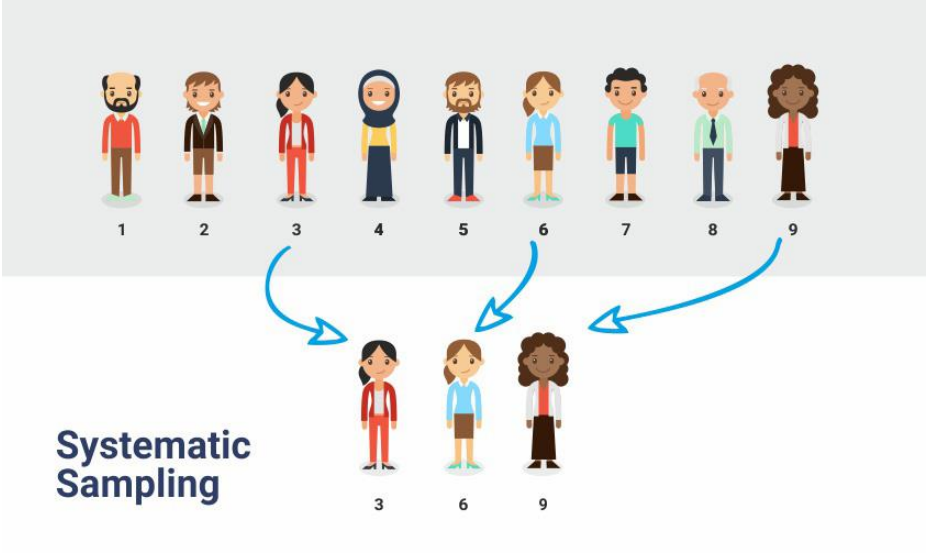
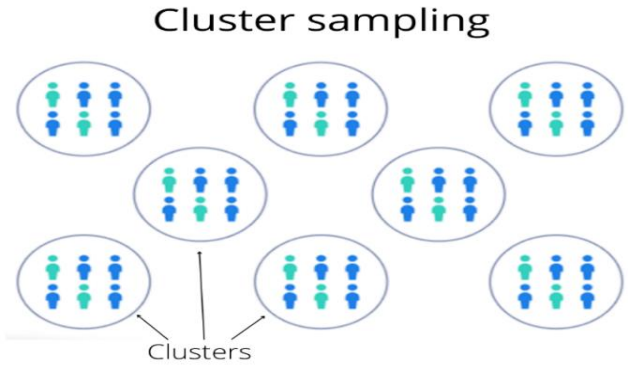


Image source: https://www.questionpro.com/blog/probability-sampling/

Example:

Suppose a statistician selects every 3rd person in a population of 9 persons for the sample. Intervals of sampling can also be systematic, such as selecting a new sample every 3rd person.

Cluster Sampling: In a clustered sample, rather than individuals, subgroups of the population are utilized as the sampling unit. The population is divided into subgroups, known as clusters, which are randomly selected to be included in the study. Clusters are usually already defined, for example, individual GP practices or towns could be identified as clusters.



https://www.scribbr.com/methodology/cluster-sampling/

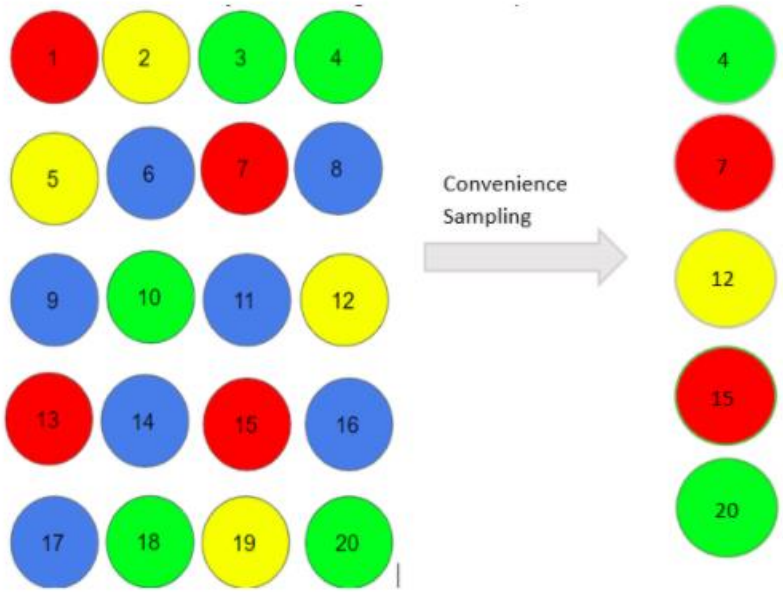
Example:

A notable example of a cluster sample is the General Household Survey, which is conducted annually in England. The survey includes all members of the selected households (clusters).

Non-Probability Sampling Methods

Convenience sampling: Because participants are chosen based on their availability and willingness to participate, convenience sampling is possibly the simplest technique of sampling. Although useful results can be obtained, they are subject to severe bias since individuals who volunteer to participate may differ from those who do not (volunteer bias).

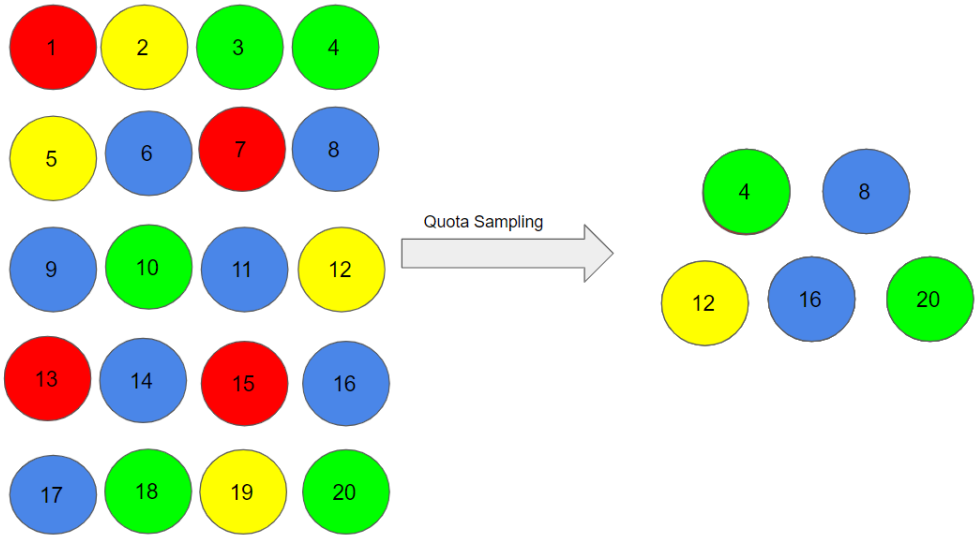
Here, let's say individuals numbered 4, 7, 12, 15 and 20 want to be part of our sample, and hence, we will include them in the sample.



For example, start-ups and NGOs frequently do convenience sampling at a mall to distribute flyers about future events or cause promotion – they do it by standing at the mall entrance and randomly handing out pamphlets.

Convenience sampling is prone to significant bias, because the sample may not be the representation of the specific characteristics such as religion or, say the gender, of the population.

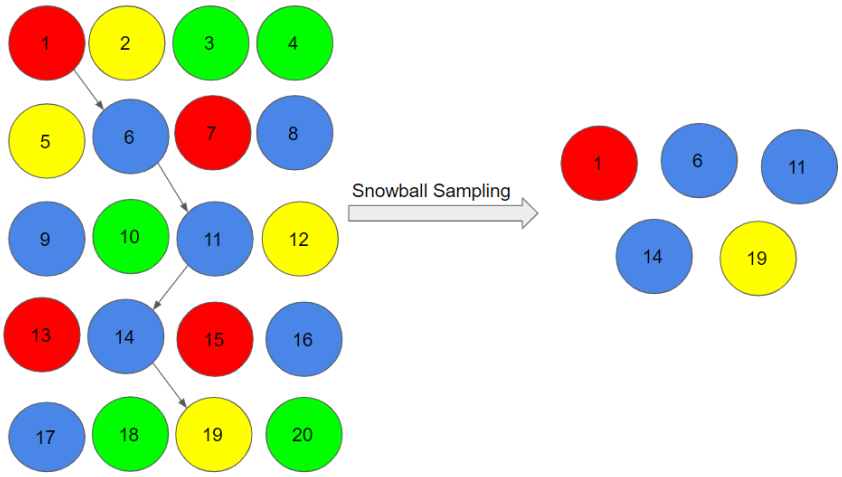
Quota sampling: In this type of sampling, we choose items based on predetermined characteristics of the population. For example, you could divide the population into strata and then select from each strata based on Quota. Consider that we have to select individuals having a number in multiples of four for our sample:



Therefore, the individuals numbered 4, 8, 12, 16, and 20 are already reserved for our sample.

In quota sampling, the chosen sample might not be the best representation of the characteristics of the population that weren't considered.

Snowball sampling: In this technique, Existing people are asked to nominate further people known to them so that the sample increases in size like a rolling snowball. This method of sampling is effective when a sampling frame is difficult to identify.

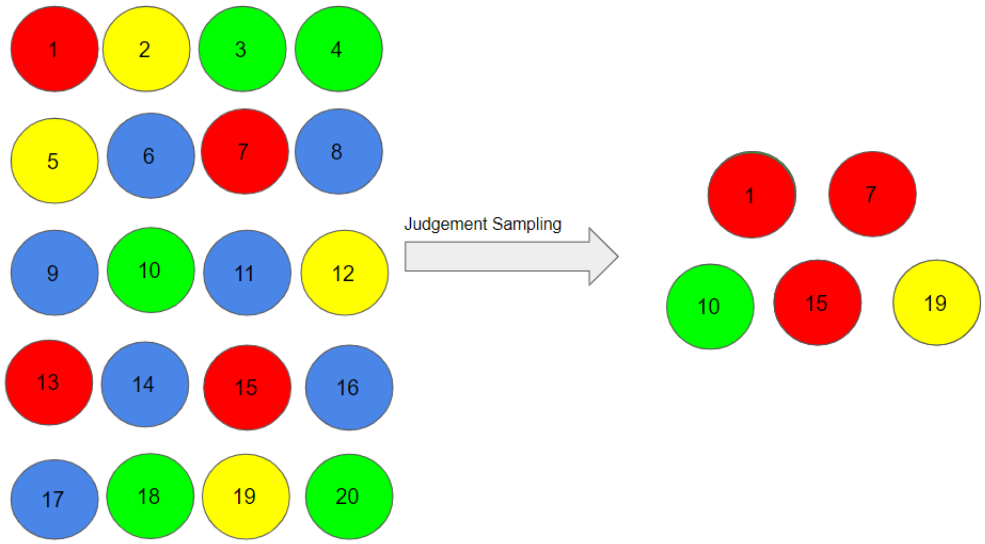


Here, we had randomly chosen person 1 for our sample, and then he/she recommended person 6, and person 6 recommended person 11, and so on.

1 > 6 > 11 > 14 > 19

Because the mentioned persons will share common features with the person who recommends them, there is a considerable risk of selection bias in snowball sampling.

Judgment sampling: Selective sampling is another name for it. When deciding who to invite to join, the experts' judgement is crucial.



Assume that our experts agree that people with the numbers 1, 7, 10, 15, and 19 should be included in our sample since they can help us better infer the population. As you might expect, quota sampling is subject to expert prejudice and may not be entirely representative.