Statistics for Data Science

Need of statistics?

Statistical knowledge helps you use the proper methods to collect the data, employ the correct analyses, and effectively present the results. Statistics is a crucial process behind how we make discoveries in science, make decisions based on data, and make predictions

What are statistics and Parameter:

In our day-in and day-out, we keep speaking about the **Population** and **sample** even though we use statistics most of the time but we are unaware of that. So, it is very important to know the terminology to represent the population and the sample.

A parameter is a number that describes the data from the population. And, the statistic is a number that describes the data from a sample.

Statistics and its types:

The Wikipedia definition of Statistics states that "it is a discipline that concerns the collection, organization, analysis, interpretation, and presentation of data."

It means, as part of statistical analysis, we collect, organize, and draw meaningful insights from the data either through visualizations or mathematical explanations.

There are majorly 2 types of statistics:
1. Descriptive Statistics
2. Inferential Statistics

Basically, as part of descriptive Statistics, we measure the following:
1. **Frequency**: No. of times a data point occurs
2. **Central tendency**: the centrality of the data – mean, median, and mode
3. **Dispersion**: the spread of the data – range, variance, and standard deviation
4. **The measure of position**: percentiles and quantile ranks

Inferential Statistics:

In Inferential statistics, we estimate the population parameters. Or we run Hypothesis testing to assess the assumptions made about the population parameters.

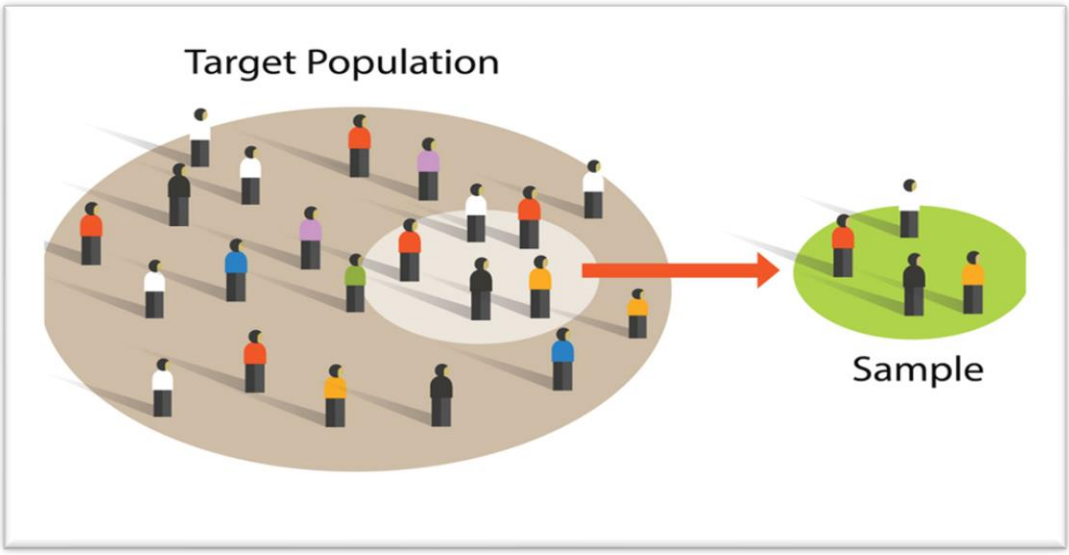
In simple terms, we interpret the meaning of the descriptive statistics by inferring them from the population. In simple words after running different tests and experiments like Z, T-test, P-value, etc. over sample data to understand the behavior and results, from that data we infer some conclusion that data is mostly accurate for population data most of the time.

For example, we are conducting a survey on the number of two-wheelers in a city. Assume the city has a total population(N) of 5L people. So, we take a sample(n) of 1000 people
as it is impossible to run an analysis on the entire population data.

From the survey conducted, it is found that 800 people out of 1000 (800 out of 1000 is 80%) are two-wheelers. So, we can infer these results from the population and conclude that 4L people out of the 5L population are two-wheelers.

Statistics for Data Science

Population (N) and sample (n):-



Data Types

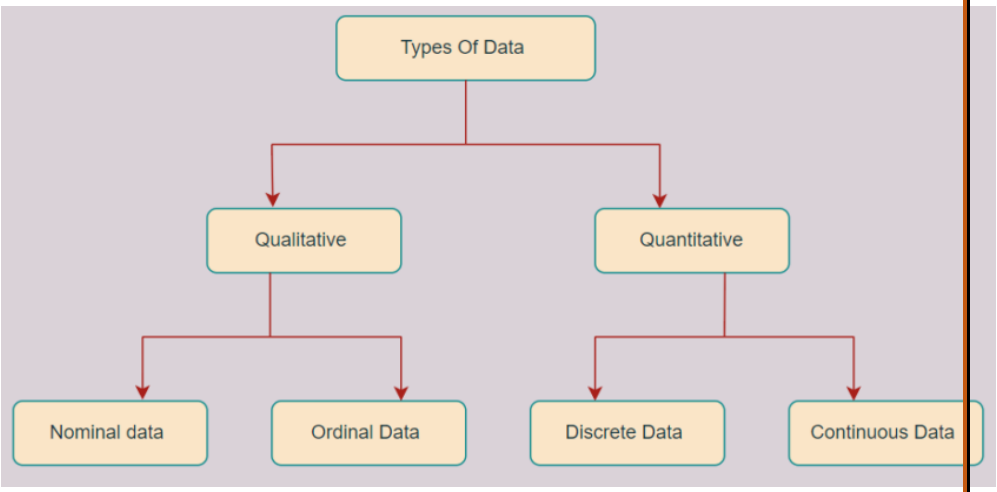
In general, two types of data available that is *Qualitative and Quantitative*

Level of Measurement

In statistics, the level of measurement is a *classification* that describes the relationship between the values of a variable.

We have **four** fundamental levels of measurement. They are:

- Nominal Scale
- Ordinal Scale
- Interval Scale
- Ratio Scale



1. Nominal Scale: This scale contains the least information since the data have names/labels only.
It can be used for classification. We cannot perform mathematical operations on nominal data because there is no numerical value to the options (numbers associated with the names can only be used as tags).
Example: Which country do you belong to? India, Japan, and Korea.

2. **Ordinal Scale**: In comparison to the nominal scale, the ordinal scale has more information because along with the labels, it has order/direction.
Example: Income level – High income, medium income, low income.

3. **Interval Scale**: It is a numerical scale. The Interval scale has more information than the

nominal, and ordinal scales. Along with the order, we know the difference between the two variables (interval indicates the distance between two entities).
Mean, median, and mode can be used to describe the data.

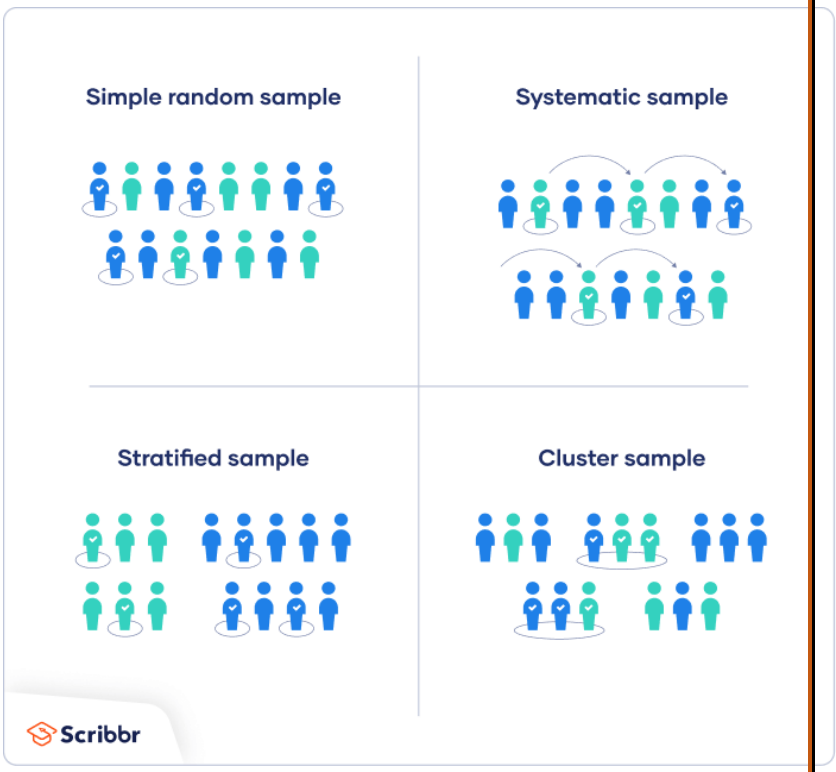
Example: Temperature, income, etc

4. **Ratio Scale**: The ratio scale has the most information about the data. Unlike the other three scales, the ratio scale can accommodate a true *zero* point. The ratio scale is simply said to be the combination of Nominal, Ordinal, and Interval scales.
Example: Current weight, height, etc.

Sampling Types:

simple random sample

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.
To conduct this type of sampling, you can use tools like *random number generators* or other techniques that are based entirely on chance.



Credit:

Systematic sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct.

Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample

Stratified sampling

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you to draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g. gender, age range, income bracket, job role)

Cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.

Moments of Business Decision

We have four moments of business decision that help us understand the data.

The Measure of central tendency: (It is also known as **First Moment Business Decision**)
Mean: It is the sum of all the data points divided by the total number of values in the data set.
Mean cannot always be relied upon because it is influenced by ***outliers***.

Example: dataset= {1,2,3,4,5,6,7,8,9,10,11,12} mean= 1+2+3+4……+12/12= 6.5

Median: It is the middlemost value of a sorted/ordered dataset. If the size of the dataset is even, then the median is calculated by taking the average of the two middle values.
Example: Median= 6+7/2= 6.5
Mode: It is the most repeated value in the dataset. Data with a single mode is called unimodal, data with two modes is called bimodal, and data with more than two modes is called multimodal.
Example: Dataset= {1,2,3,4,5,6,3,2,1,2,3,8,3,9,3} Mode= 3

Measures of Dispersion: (***Second Moment*** *Business Decision)*
Talks about the spread of data from its center.

Dispersion can be measured using:
Variance: It is the average squared distance of all the data points from their mean. The problem with Variance is, that the units will also get squared.

Let's an example: Data_1= {2,2,4,4} -> Mean= 3 Widely spread
Data_2= {1,1,5,5} -> Mean= 3 Widely spread (variance more that means each data points are far away from its mean as compared to *Dataset_1*)

Observation: In both, the above dataset Mean is the same but on other hand, the variance differs not
Similar

Standard Deviation: It is the *square root of Variance*. Helps in retrieving the original units.
Range: It is the difference between the **maximum and the minimum** values of a dataset

	Population	Sample
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i-\mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i-\bar{x})^2}{n-1}$
Standard Deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i-\mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i-\bar{x})^2}{n-1}}$

Statistics for Data Science

Skewness:(***Third Moment*** *Business Decision*)
It measures the asymmetry in the data. The two types of Skewness are:

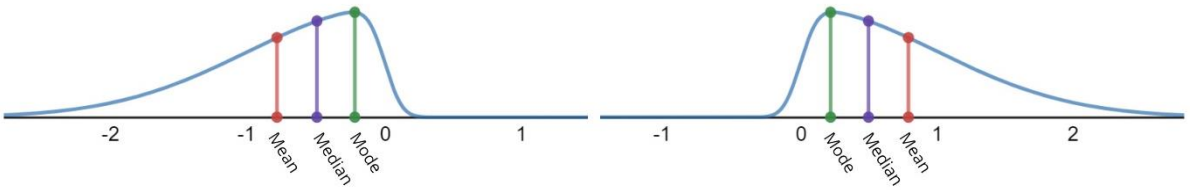
Positive/right-skewed: Data is said to be positively skewed if most of the data is concentrated on the left side and has a tail towards the right.

$$\text{skewness} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{(N-1)s^3}$$

where:

- s is the standard deviation
- $\bar{x}$ is the mean of the distribution
- N is the number of observations of the sample

Negative/left-skewed: Data is said to be negatively skewed if most of the data is concentrated on the right side and has a tail towards the left



Left-Skewed

Right-Skewed

In positive Skewness: - Mean>Median>Mode
In positive Skewness: - Mean<Median<Mode

Kurtosis
(It is also known as **Fourth Moment** Business Decision)

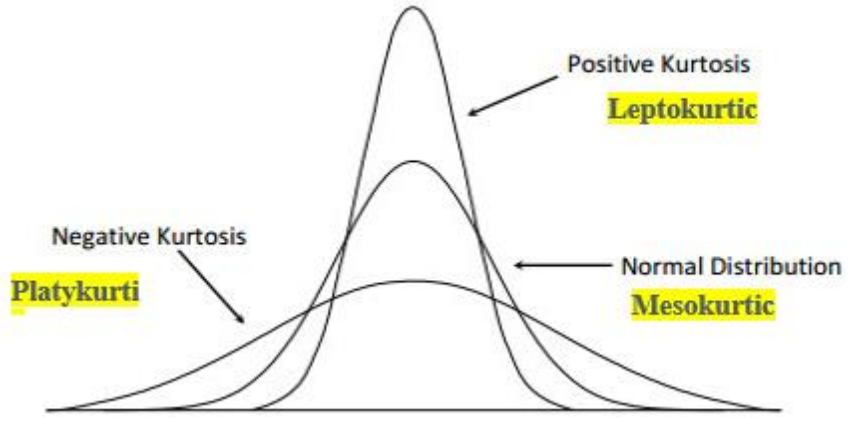
Talks about the central peakiness or fatness of tails. The three types of Kurtosis are:

- **Positive/leptokurtic:** Has sharp peaks and lighter tails
- **Negative/Platokurtic:** Has wide peaks and thicker tails
- **MesoKurtic:** Normal distribution

$$\text{Kurtosis} = \frac{\sum(x_i - \bar{x})^4}{nS^4}$$

$\bar{x}$ = mean of the given data
S = standard deviation of the data
n = total number of observations

Graph of Kurtosis



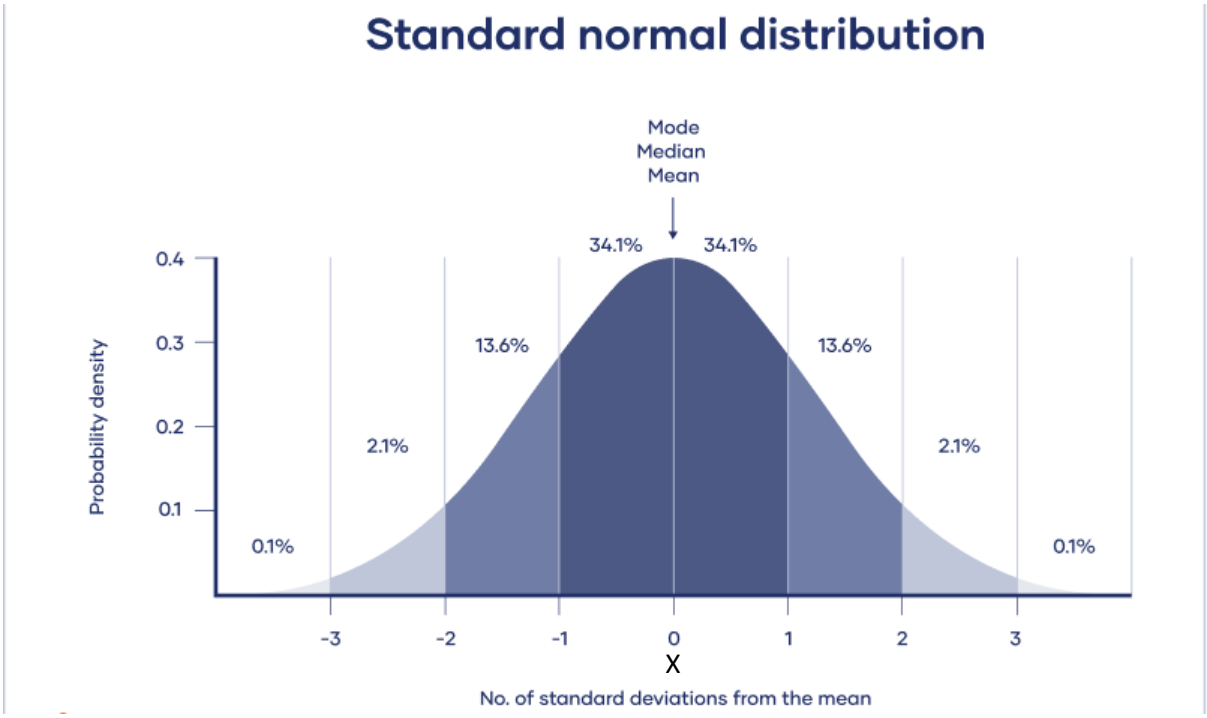
Statistics for Data Science

Gaussian/Normal Distribution curve – Data spread on Symmetrical distribution

The data was distributed in **68%, 95%, and 99.7%** manner (Empirical Rule)

The **standard normal distribution**, also called the **z-distribution**, is a special normal distribution where the **mean** is 0 and the **standard deviation** is 1.

Any normal distribution can be standardized by converting its values into *z*-scores. *Z*-scores tell you how many standard deviations from the mean each value lies.



Credit: Z-Score

When you standardize a distribution, the **mean** becomes 0 and the **standard deviation** becomes 1 [SND= (μ=0 and standard deviation=0)]

- A positive *z*-score means that your *x*-value is *greater* than the mean.
- A negative *z*-score means that your *x*-value is *less* than the mean.
- A *z*-score of zero means that your *x*-value is *equal* to the mean.

Z-score formula	Explanation
$z = \frac{x - \mu}{\sigma}$	• x = individual value • μ = mean • σ = standard deviation

Suppose you are enrolled in three classes, **statistics, biology, and Math**, and you just took the first exam in each. You receive a grade of 82 on your statistics exam, where the mean grade was 74 and the standard deviation was 12. You receive a grade of 72 on your biology exam, where the mean grade was 65 and the standard deviation was 10. Finally, you receive a grade of 91 on your Math exam, where the mean grade was 88 and the standard deviation was 6. Although your highest test score was 91 (Math), in which class did you score the best, *relative to the rest of the class*? We can answer this using a **z-score!**

Statistics for Data Science

Statistics: A grade of 82, and a z-score of: $z = \frac{82-74}{12} = 0.67$

Biology: A grade of 72, and a *z*-score of: $z = \frac{72-65}{10} = 0.70$

Math : A grade of 81, and a *z*-score of: $z = \frac{91-88}{6} = 0.50$

Results:

Your statistics, Biology & Math exam score was 0.67,0.70,0.50 standard deviations better than the class average respectively. Therefore, even though your actual score on the biology exam was the lowest of the three exam scores, ***relative to the distribution of all class exam scores***, your biology exam score was the highest relative grade.

If Dataset is not *normally distributed* or different measured with different metrics then it is a matter to bring that data points to a standard normal distribution. Let's get into it.
We have two ways to do it.

- Standardization
- Normalization

Standardization

In statistics, standardization is **the process of putting different variables on the same scale**. This process allows you to compare scores between different types of variables. Typically, to standardize variables, you calculate the mean and standard deviation for a variable.

Standardization comes into the picture when features of the input data set have large differences between their ranges, or simply when they are measured in different units.

How to standardize the data by using **Z-score.**

Example question: A hot dog stand has mean daily sales of $420 with a standard deviation of $50. The income has a normal distribution. What is the standardized value for daily sales of $520?

Here the values are…

- X = 520
- μ = 420
- σ = 50

Standardized value = X – μ / σ = 520 – 420 / 50.

So, the z-score is **2**

when should you standardize your data, and why?

1. **BEFORE PRINCIPAL COMPONENT ANALYSIS (PCA)**

In principal component analysis, features with high **variances** or wide ranges get more weight than those with low variances, and consequently, they end up illegitimately

dominating the first principal components (components with maximum variance). I used the word "illegitimately" here because the reason these features have high variances compared to the other ones is just that they were measured in different scales.

Standardization can prevent this, by giving the same weightage to all features.

2. **BEFORE CLUSTERING**

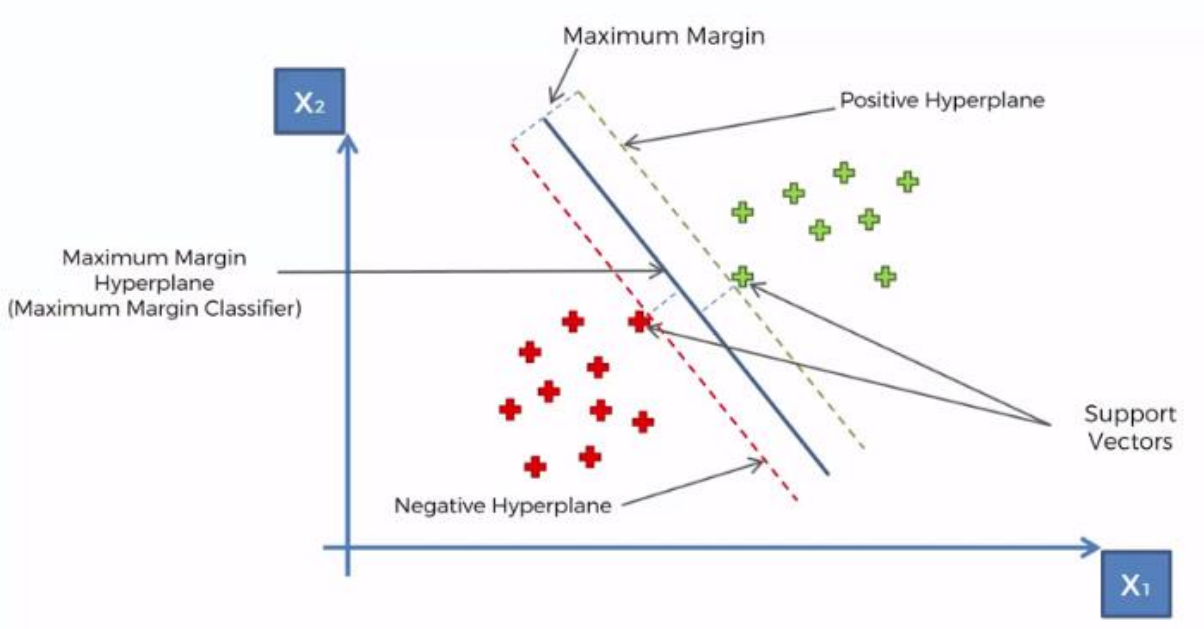
Clustering models are **distance-based** algorithms. In order to measure similarities between observations and form clusters they use a distance metric. So, features with high ranges will have a bigger influence on the clustering. Therefore, standardization is required before building a **clustering model**.

3. **BEFORE K-NEAREST NEIGHBORS (KNN)**

K-nearest neighbor is a distance-based classifier that classifies new observations based on similar measures (e.g., distance metrics) with labeled observations of the training set. Standardization makes all variables contribute equally to the similarity measures.

4. **BEFORE SUPPORT VECTOR MACHINE (SVM)**

Support vector machine tries to maximize the distance between the *separating plane and the support vectors*. If one feature has very large values, it will dominate over other features when calculating the distance. Standardization gives all features the same influence on the distance metric.



5. **BEFORE MEASURING VARIABLE IMPORTANCE IN REGRESSION MODELS**

You can measure variable importance in regression analysis by fitting a regression model using the standardized independent variables and comparing the absolute value of their standardized coefficients. But, if the independent variables are not standardized, comparing their coefficients becomes meaningless.

Statistics for Data Science

6. BEFORE LASSO AND RIDGE REGRESSIONS

Lasso and ridge regressions place a penalty on the magnitude of the coefficients associated with each variable, and the scale of variables will affect how much of a penalty will be applied on their coefficients. The coefficients of variables with a large variance are small and thus less penalized. Therefore, standardization is required before fitting both regressions.

Cases When Standardization Is Not Needed

LOGISTIC REGRESSIONS AND TREE-BASED MODELS

Logistic regressions and tree-based algorithms such as decision trees, random forests, and gradient boosting are not sensitive to the magnitude of variables. So, standardization is not needed before fitting these kinds of models.

Normalization

The goal of normalization is to transform features to be on a similar scale. This improves the performance and training stability of the model.

Four common normalization techniques may be useful:

- scaling to a range (min-max scaler) Scaling from(0 to 1)
- clipping
- log scaling
- z-score

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Normalization is used when the data doesn't have Gaussian distribution whereas Standardization is used on data having Gaussian distribution.

- Normalization scales in a range of [0,1] or [-1,1].
- Standardization is not bounded by range.
- Normalization is highly affected by **outliers**

There we have to handle Outliers.

What is Outlier: - An outlier is **an observation that lies an abnormal distance from other values in a random sample from a population**

Two graphical techniques for identifying outliers, **scatter and box plots** (IQR method)

Ways to handle outliers: There are many ways to handle outliers I have mentioned here some famous techniques

- Quantile based flooring and capping (IQR)
- Through Z-score
- Mean/median imputation
- Trimming/removing the outlier

Statistics for Data Science

Example: **How to detect outliers in a dataset: -**

Data points that lie 1.5 times of IQR above Q3 (75 percentile) and below Q1 (25 percentile) are outliers.

- Sort the dataset in ascending order
- calculate the 1st and 3rd quartiles (Q1, Q3) [Box plot]
- compute IQR=Q3-Q1 (Q1, Q3- lower, higher fence)
- compute lower fence = (Q1–1.5*IQR), upper fence = (Q3+1.5*IQR)
- loop through the values of the dataset and check for those who fall below the lower bound and above the upper bound and mark them as outliers

Dataset= [30, 171, 184, 201, 212, 250, 265, 270, 272, 289, 305, 306, 322, 322, 336, 346, 351, 370, 390, 404, 409, 411, 436, 437, 439, 441, 444, 448, 451, 453, 470, 480, 482, 487, 494, 495, 499, 503, 514, 521, 522, 527, 548, 550, 559, 560, 570, 572, 574, 578, 585, 592, 592, 607, 616, 618, 621, 629, 637, 638, 640, 656, 668, 707, 709, 719, 737, 739, 752, 758, 766, 792, 792, 794, 802, 818, 830, 832, 843, 858, 860, 869, 918, 925, 953, 991, 1000, 1005, 1068, 1441]

Lower quartile = 0.25(N+1)th ordered point = 22.75th ordered point = 411 + 0.75(436-411) = 429.75
Upper quartile = 0.75(N+1)th ordered point = 68.25th ordered point = 739 +0.25(752-739) = 742.25

Interquartile range (IQR) = 742.25 - 429.75 = 312.5
Lower fence = 429.75 - 1.5 (312.5) = -39.0
Upper fence = 742.25 + 1.5 (312.5) = 1211.0

Clearly observed the upper fence is **1211** but, in our Dataset, only 1441 is available after point 1211, so 1441 is an **outlier**.

What is Probability?

Probability denotes the possibility of the **outcome** of any random event. The meaning of this term is to check the extent to which any event is likely to happen. For example, when we flip a coin in the air, what is the possibility of getting a head? The answer to this question is based on the number of possible outcomes.

Permutation and combination

There are basically two types of permutation:

1. **Repetition is Allowed**: such as the lock above. It could be "333".
2. **No Repetition**: for example the first three people in a running race. You can't be first *and* second.

1. Permutations with Repetition

These are the easiest to calculate.
When a thing has ***n*** different types ... we have ***n*** choices each time!
For example: choosing ***3*** of those things, the permutations are:

Statistics for Data Science

n × n × n
(n multiplied 3 times)

More generally: choosing ***r*** of something that has ***n*** different types, the permutations are:

n × n × ... *(r times)*

(In other words, there are **n** possibilities for the first choice, THEN there are **n** possibilites for the second choice, and so on, multiplying each time.)

Which is easier to write down using an exponent of **r**: **n × n × ... (r times) = n^r**

2.Permutations without Repetition

There are also two types of combinations (remember the order does **not** matter now):

1. **Repetition is Allowed**: such as coins in your pocket (5,5,5,10,10)
2. **No Repetition**: such as lottery numbers (2,14,15,27,30,33)

$$\frac{n!}{(n-r)!}$$

where ***n*** is the number of things to choose from,
and we choose ***r*** of them,
no repetitions,
order matters.

	Repeats allowed	No Repeats
Permutations (order matters):	n^r	$\frac{n!}{(n-r)!}$
Combinations (order doesn't matter):	$\frac{(r+n-1)!}{r!(n-1)!}$	$\frac{n!}{r!(n-r)!}$

Variance:

Variance is the expected value of the squared variation of a random variable from its mean value, in probability and statistics. Informally, variance estimates how far a set of numbers (random) are spread out from their mean value

Variance is symbolically represented by σ^2, s^2, or **Var(X)**.it is square of standard deviation

Covariance:

Covariance is a measure of the relationship between two **random variables**, in statistics. The covariance indicates the relation between the two variables and helps to know if the two variables vary together

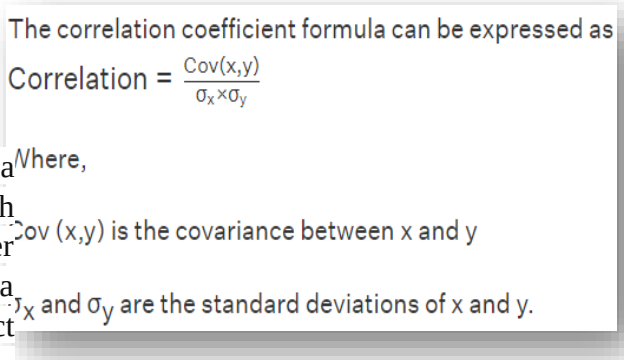
$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$cov_{x,y}$ = covariance between variable x and y
x_i = data value of x
y_i = data value of y
$\bar{x}$ = mean of x
$\bar{y}$ = mean of y
N = number of data values

Statistics for Data Science

Pearsons Correlation Coefficient (-1 to 1)

The correlation coefficient is a statistical measure of the strength of a *linear relationship* between two variables. Its values can range from **-1 to 1**



A correlation coefficient of -1 describes a perfect *negative, or inverse*, correlation, with values in one series rising as those in the other decline, and vice versa. A coefficient of 1 shows a ***perfect positive correlation***, or a direct relationship. A correlation coefficient of 0 means there is no linear relationship.

Correlation Coefficient is zero when there is no relation between the variable

Spearman's rank correlation coefficient

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.

It is used for Non-linear Datasets. For linear dataset, person correlation coefficient works better.

Formula:

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

ρ = Spearman's rank correlation coefficient

di = Difference between the two ranks of each observation

n = Number of observations

Introduction to Univariate, Bivariate and Multivariate Analysis

In the field of data, there is nothing more important than understanding the data that you are trying to analyze. In order to understand the data is it important to understand the nature and dependency of data.

You may have seen sometimes a single data type can extract any meaningful insight, other times 2 data combined to produce any useful information, and same way 2 or more than that can have some info. On that dataset.

There Univariate, Bivariate and Multivariate Analysis come into the picture.

Statistics for Data Science

Univariate analysis

Uni means one and variate means variable, so in univariate analysis, there is only one dependable variable. The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately. The dataset may be *categorical or Numerical*.

Example: Here we have the IQ of students

The list of IQ scores is 118, 139, 124, 125, 127, 128, 129, 130, 130, 133, 136, 138, 141, 142,

149, 130, 154.

IQ Range Number
118-125 (3)
126-133 (7)
134-141 (4)
142-149 (2)
150-157 (1)
In this case, the dependable variable IQ can establish the meaning in the student database.

Bar charts, Pie charts, Box plots, frequency table, and Histograms are used for Univariate analysis of data visualization.

Bivariate analysis

Bi means two and variate means variable, so here there are two variables. Now again the variables can be numeric and categorical. Bivariate analysis helps in studying the relationship between two variables.

Types of bivariate analysis

- Numerical feature vs Numerical feature
- Categorical feature vs Categorical feature
- Numerical feature vs Categorical feature

Visualization plots/test type used in bivariate analysis

Scatter plot (Numeric vs Numeric)
Chi-Squared Test (Categorical vs Categorical)
ANOVA (Continuous vs categorical)
Z/T-test (Numerical vs Categorical)

Multivariate analysis

Multivariate analysis is required when more than two variables have to be analyzed simultaneously. It is a tremendously hard task for the human brain to visualize a relationship among 4 variables in a graph and thus multivariate analysis is used to study more complex sets of data.
Types of Multivariate Analysis includes

- Cluster Analysis

Statistics for Data Science

- Factor Analysis
- Multiple Regression Analysis
- PCA & MCA

What is Distribution

The distribution of a statistical dataset is the spread of the data which shows all possible values or intervals of the data and how they occur. The distribution provides a parameterized mathematical function which will calculate the probability of any individual observation from the sample space.

The most common measures of any distribution, how sample differs from each other is the standard deviation, and the standard error of the mean.

Categories of Distribution

Frequency Distribution

The number of times each numerical/categorical value occurs.

Probability Distribution

List of Probabilities associated with each of its possible numerical values.

Types of Distribution:

- Bernoulli Distribution
- Uniform Distribution
- Binomial Distribution
- Normal Distribution
- Poisson Distribution
- Exponential Distribution

Bernoulli Distribution:

Bernoulli Distribution is a type of discrete probability distribution where every experiment conducted asks a question that can be answered only in *yes or no*. In other words, the random variable can be **1** with a probability **p** or it can be **0** with a probability **(1 - p).** Such an experiment is called a Bernoulli trial.

Bernoulli Distribution Example

Suppose there is an experiment where you **flip a coin** that is fair. If the outcome of the flip heads, then you will win. This means that the probability of getting heads is p = 1/2. If X is the random variable following a Bernoulli Distribution, we get P (X = 1) = p = 1/2.

Uniform Distribution:

It is a continuous or rectangular distribution. It describes an experiment where an outcome lies between certain boundaries.

Example:
Time to fly from Newark to Atlanta ranges from 120 to 150 minutes if we monitor the fly time for many commercial flights it will follow more or less the uniform distribution.

Statistics for Data Science

Time for Pizza delivery from Nanganallur to Alandur may range from 20 to 30 mins uniformly from the time delivery man leaves the Pizza Hut.

Binomial Distribution:

The most widely known **discrete probability distribution**. It has been used hundreds of years.

Assumptions:

1. The experiment involves n identical trials.
2. Each trial has only two possible outcomes – success or failure.
3. Each trial is independent of the previous trials.
4. The terms p and q remain constant throughout the experiment, where p is the probability of getting a success on any one trial and q = (1 – p) is the probability of getting a failure on any one trial.

Poisson Distribution

It is the discrete probability distribution of the number of times an event is likely to occur within a specified period of time. It is used for independent events which occur at a constant rate within a given interval of time.

The occurrences in each interval can range from zero to infinity **(0 to α).**

Examples:

1. How many black colours are there in a random sample of 50 cars
2. No of cars arriving at a car wash during a 20 minute time interval

Exponential Distribution

It is concerned with the amount of time until some specific event occurs.
Example:

- The amount of time until an earthquake occurs has an exponential distribution
- The amount of time in business telephone calls
- The car battery lasts.
- The amount of money customers spend on one trip to the supermarket follows an exponential distribution. There are more people who spend small amounts of money and fewer people who spend large amounts of money.

The exponential distribution is widely used in the field of reliability.

PDF & CDF

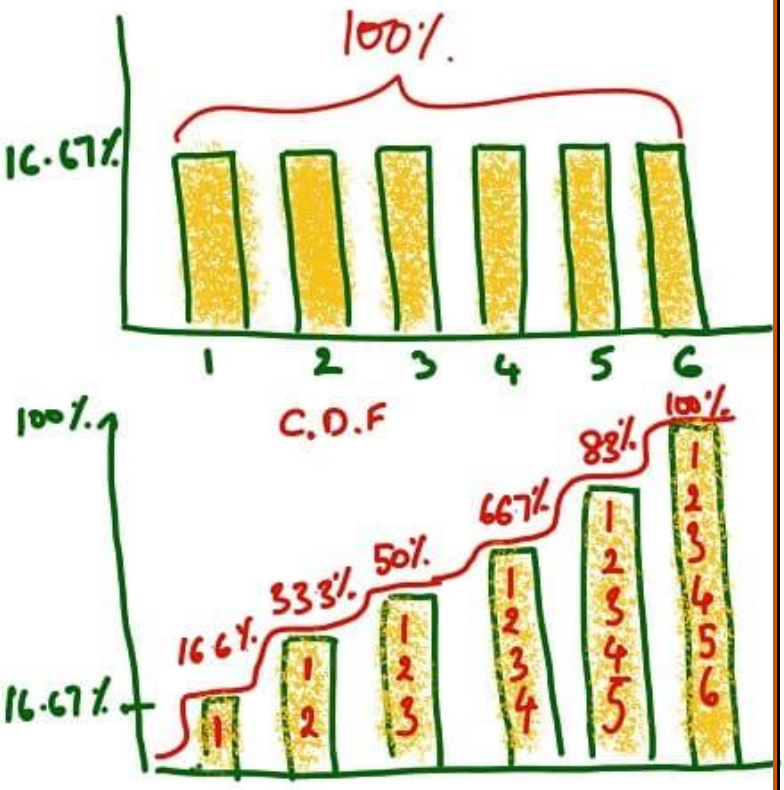
Two of the most important statistical functions in a nutshell. The PDF and CDF statistical functions are widely used techniques in the Exploratory Data Analysis **to find the probabilistic relationships between the variables**.

Statistics for Data Science

Probability Density Function (PDF)

The probability density function (PDF) is the probability that a random variable, say X, will take a value **exactly equal** to X. – The PDF focuses on ***one specific value***. Whereas, for the cumulative distribution function, we are interested in the probability of taking on a value equal to or less than the specified value. The probability density function is also referred to as **the probability mass function** (PMF).

For example, if you roll a die, the probability of obtaining 1, 2, 3, 4, 5, or 6 is 16.667% (=1/6). The probability density function (PDF) or the probability that you will get exactly 2 will be 16.667%. *Whereas, the cumulative distribution function (CDF) of 2 is 33.33% as described above.*



Credit: Graduatetutor

Cumulative Distribution Function (CDF)

The cumulative distribution function (CDF) is the probability that a random variable, say X, will take a value equal to or less than x.

For example, if you roll a die, the probability of obtaining a 1 or 2 or 3 or 4 or 5 or 6 is 16.667% (=1/6) individually. The cumulative distribution function (CDF) of 1 is the probability that the next roll will take a value less than or equal to 1 and is 16.667%. There is only one possible way to get a 1.

Why should we use PDF, CDF?

PDFs, CDFs are widely used in data analysis. Once the distribution of data, mean, and variance of the data is known, it is easy to plot PDF and CDF. These plots can help find solutions to many problems in data analysis.

Consider an example of placing a purchase order for shirts for an organization. Let the available sizes of shirts be small, medium, or large. A very common question will be how many of each size to order. Such questions can be answered easily using PDF and CDF.

CDF will give the percentage of the population having a size greater than a certain value or lesser than a certain value. From this interpretation, we can easily find the percentage of the population in each category.

Box-Cox and Log-Normal Distribution

Box-Cox transformation is a statistical technique that transforms your target variable **so that your data closely resembles a normal distribution**. In many statistical techniques, we assume that the errors are normally distributed. This assumption allows us to construct confidence intervals and conduct hypothesis tests

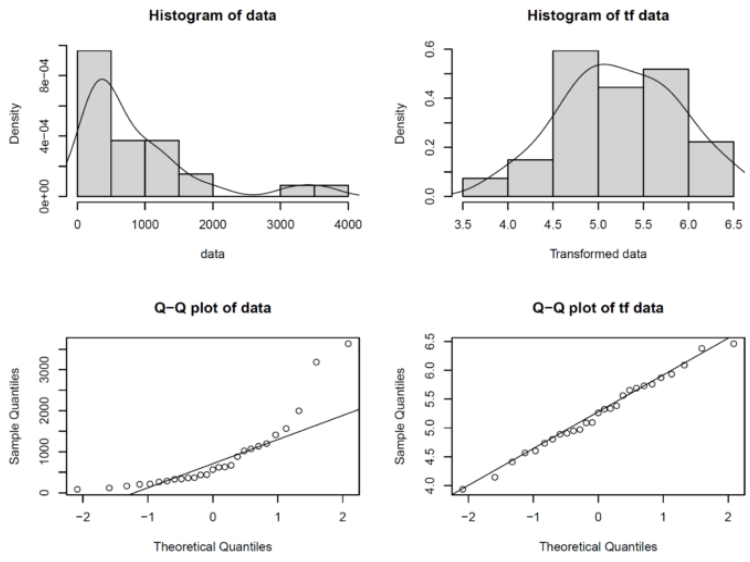
The Box-Cox transformation (Box and Cox, 1964) is a way to transform data that **ordinarily do not follow a normal distribution** so that it then conforms to it. The transformation is a piecewise function of the power parameter λ:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & (\lambda \neq 0) \\ \ln y & (\lambda = 0) \end{cases}$$

The Lambda value indicates the power to which all data should be raised. In order to do this, the Box-Cox power transformation searches from Lambda = -5 to Lamba = +5 until the best value is found

Is a Box-Cox transformation a log transformation?

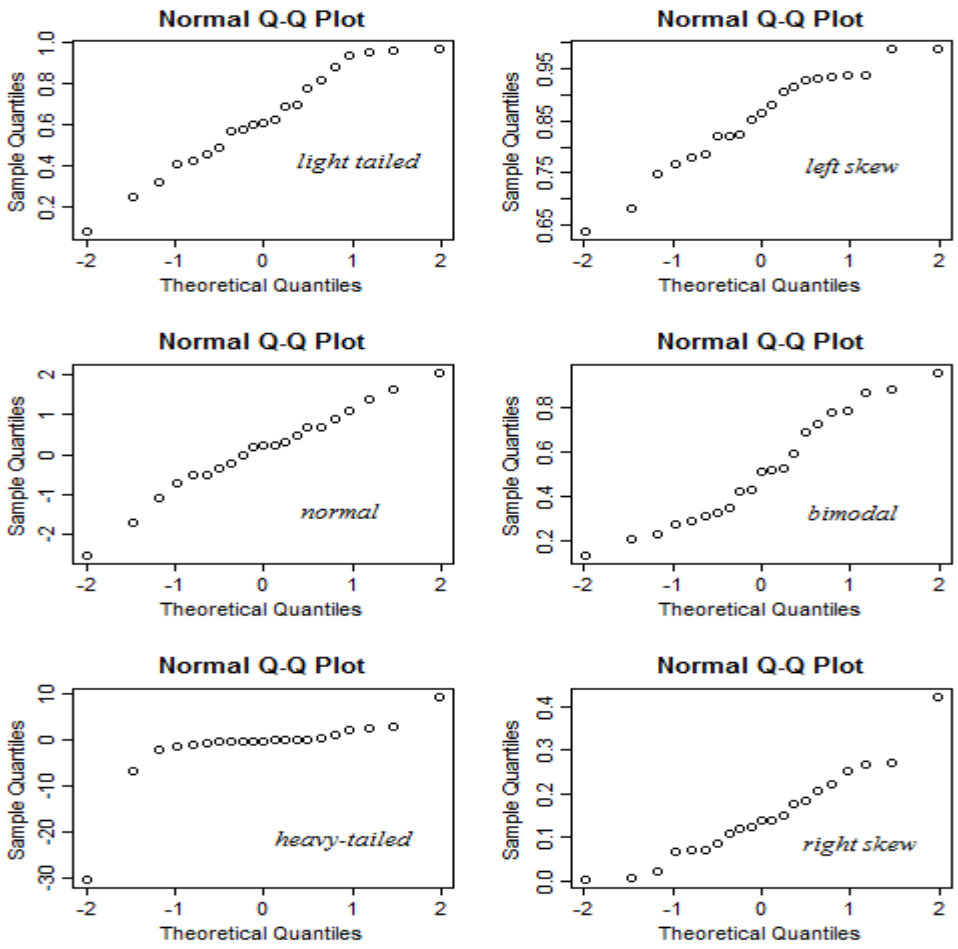
The log transformation is actually a special case of the Box-Cox transformation when λ = 0; the transformation is as follows: Y(s) = ln(Z(s)), for Z(s) > 0, and ln is the natural logarithm.



Statistics for Data Science

Q-Q plot (Quantile-Quantile plot)

Being a data scientist or analyst, it's very important to know whether the *distribution is normal* or not so as to apply various statistical measures on the data and interpret it in much more human-understandable visualization and there Q-Q plot comes into the picture. The most fundamental question answered by the Q-Q plot is: Curve normally distributed or not…



Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc.

You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot. In general, we are talking about **Normal distributions** only because we have a very beautiful concept of **68–95–99.7 rule**.

Usage:

The Quantile-Quantile plot is used for the following purpose:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behavior

How to Draw a Q-Q plot

- Collect the data for plotting the quantile-quantile plot.
- Sort the data in ascending or descending order.
- Draw a normal distribution curve.
- Find the z-value (cut-off point) for each segment.
- Plot the dataset values against the normalizing cut-off points.

Statistics for Data Science

Inferential statistics

Why do we need Inferential Statistics?

Suppose, you want to know the average salary of Data Science professionals in India. Which of the following methods can be used to calculate it?

- Meet every Data Science professional in India. Note down their salaries and then calculate the total average
- Or hand pick a number of professionals in a city like Bangalore. Note down their salaries and use them to calculate the Indian average.

In simple language, **Inferential Statistics** is used to draw inferences beyond the immediate data available.

With the help of **inferential statistics**, we can answer the following questions:

- Making inferences about the population from the sample.
- Concluding whether a sample is significantly different from the population. For example, let's say you collected the *salary* details of Data Science professionals in Bangalore. And you observed that the average salary of Bangalore's data scientists is more than the average salary across India. Now, we can conclude if the difference is statistically significant.
- If adding or removing a feature from a model will really help to improve the model.
- If one model is significantly better than the other?
- Hypothesis testing in general.

When sample means are equals to the population mean we are getting a normally distributed shape of the dataset.

Points to note:

- *Central Limit Theorem* holds true irrespective of the type of distribution of the population.
- Now, we have a way to estimate the population mean by just making repeated observations of samples of a fixed size.
- Greater the sample size, lower the standard error and greater the accuracy in determining the population mean from the sample mean.

Points to cover:

- Confidence Interval
- Hypothesis Testing
- P-value
- Significance value
- Regression & ANNOVA
- Chi square test

Statistics for Data Science

There are two important types of estimates you can make about the population: point estimates and interval estimates.

- A **point estimate** is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean.
- An **interval estimate** gives you a range of values where the parameter is expected to lie. A **confidence interval** is the most common type of interval estimate.

Confidence Interval

The confidence interval is a type of interval estimate from the sampling distribution which gives a range of values in which the population statistic may lie.

Where:

$$CI = \bar{X} \pm Z^* \frac{\sigma}{\sqrt{n}}$$

- CI = confidence interval
- $\bar{X}$ = population mean (point of estimate)
- Z* = critical value of the *z*-distribution
- σ = population standard deviation
- √n = square root of the population size

P-values in scientific studies are used to determine whether a null hypothesis formulated before the performance of the study is to be *accepted or rejected*. In exploratory studies, p-values enable the recognition of any statistically noteworthy findings. Confidence intervals provide information about a range in which the true value lies with a certain degree of probability, as well as about the direction and strength of the demonstrated effect.

The **lower the p-value**, the greater the statistical significance of the observed difference. A p-value of **0.05** or lower is generally considered statistically significant.

For example, if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.

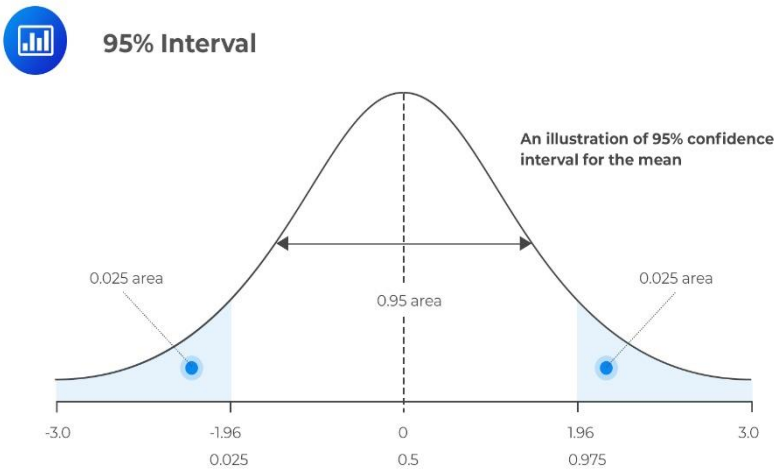
Your desired confidence level is usually one minus the alpha (*a*) value you used in your statistical test:

Confidence level = 1 – *a*

Significance value = alpha

So, if you use an alpha value of $p < 0.05$ for statistical significance, then your confidence level would be 1 − 0.05 = 0.95, or 95%

Here in this figure, we have a 95% confidence interval and



the significance value is 0.025 on each tail.

We know that 95% of the values lie within 2 (1.96 to be more accurate) standard deviation of a normal distribution curve. So, for the above curve, the blue-shaded portion represents the confidence interval for a sample mean of 0

For a *z*-statistic, some of the most common values are shown in this table:

Confidence level	90%	95%	99%
alpha for one-tailed CI	0.1	0.05	0.01
alpha for two-tailed CI	0.05	0.025	0.005
z-statistic	1.64	1.96	2.57

Hypothesis Testing

Hypothesis testing is the process used **to evaluate the strength of evidence from the sample and provides a framework for making determinations related to the population**, ie, it provides a method for understanding how reliably one can extrapolate observed findings in a sample under study to the larger population.

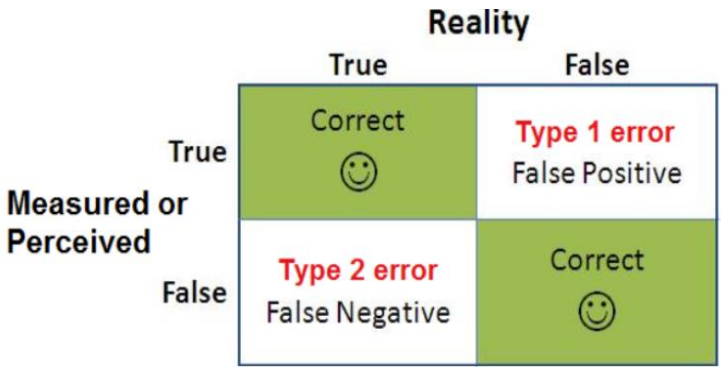
Example: Class 8th has a **mean** score of 40 marks out of 100. The principal of the school decided that extra classes are necessary in order to improve the performance of the class. The class scored an average of 45 marks out of 100 after taking extra classes. Can we be sure whether the increase in marks is a result of extra classes or is it just random?

Hypothesis testing is defined in two terms – **Null Hypothesis** and **Alternate Hypothesis**.

Error 1: The probability and confidence level where we will reject a **null hypothesis** when the significance level (alpha) is True.

Error 2: The probability and confidence level where we will fail to reject a **null hypothesis** when the significance level (alpha) is False

To be clear when we get the Hypothesis test value within the confidence interval/level that means we conclude that we **fail to reject** the **null** hypothesis on the other hand if the hypothesis-test value is outside of the confidence interval there we reject the null hypothesis.



Statistics for Data Science

How to select the best Hypothesis Test

Type of predictor variable	Distribution type	Desired Test	Attributes
Quantitative	Normal Distribution	Z – Test	• Large sample size • Population standard deviation known
Quantitative	T Distribution	T-Test	• Sample size less than 30 • Population standard deviation unknown
Quantitative	Positively skewed distribution	F – Test	• When you want to compare 3 or more variables
Quantitative	Negatively skewed distribution	NA	• Requires feature transformation to perform a hypothesis test
Categorical	NA	Chi-Square test	• Test of independence • Goodness of fit

t-Test formula

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

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

Z-Test formula

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

$\bar{x}$ = sample mean

μ = population mean

σ = population standard deviation

n = sample size

We will understand how to identify which **t-test** to be used and then proceed on to solve it. The other t-tests will follow the same argument.

Example 1: A population has a mean weight of 68 kg. A random sample of size 25 has a mean weight of 70 with a standard deviation of 4. Identify whether this sample is representative of the population.

Step 0: Identifying the type of t-test

Number of samples in question = 1

Number of times the sample is in study = 1

Any intervention on sample = No

Recommended t-test = 1- sample t-test.

Had there been 2 samples, we would have opted for a 2-sample t-test and if there would have been 2 observations on the same sample, we would have opted for the paired t-test.`

Step 1: State the Null and Alternate Hypothesis

Null Hypothesis: The sample mean and population mean are the same.

Alternate Hypothesis: The sample mean and population mean are different.

Step 2: Calculate the appropriate test statistic

df = 25-1 =24 ***[df (degrees of freedom)= sample size-1]***

t= (70-68)/(4/√25) = 2.5

Now, for a 95% confidence level, the t-critical (two-tail) for rejecting Null Hypothesis for 24 d.f is 2.06 . Hence, we can **reject** the Null Hypothesis and conclude that the two means are different.

Example 2:

An automatic cutter machine must cut steel strips of 1200 mm lenght. From preliminary data, we checked that the lenghts of the pieces produced by the machine can be considered as normal random variables with a 3mm standard deviation. We want to make sure that the machine is set currectly. Therefore 16 pieces of the products are randomly selected and weight. The figures were in mm:
1193,1196,1198,1195,1198,1199,1204,1193,1203,1201,1196,1200,1191,1196,1198,1191

Solution:

From the question we understand (step 1)

The standard deviation of the sample is given

the population is said to be normally distributed

So according to test rule we can use Z-test as the SD is known.

Step 2: State the null and the alternate hypothesis

H0: $\mu 0 = \mu 1$
Ha: $\mu 0 \neq \mu 1$

Step 3: Calculate the mean of the sample
Mean = 1197

Step 4: Calculate the test statistic

$$Z = \frac{(1197-1200)}{3/\sqrt{16}}$$

Z= -3/(3/4) = -4

Step 5: If we Look up the value in z table, and found the p-value is more than the significance level 0.05 that is 0.975. Hence, we **reject** the null hypothesis.

Statistics for Data Science

STATS HELP – HYPOTHESIS TESTING and CONFIDENCE INTERVALS

by Preston

z-SCORE	$z = \frac{X - \mu}{\sigma}$	X-SCORE	$X = \mu + (z \times \sigma)$

DESCRIPTIVE STATISTICS

CALCULATE THE SIZE OF ONE STANDARD DEVIATION

DISTRIBUTION		POPULATION	SAMPLE	DIFFERENCE SCORES	INDEPENDENT SAMPLES
SCORE		X	X	D	X
SIZE		N	n	n	n
DEGREES OF FREEDOM		$df = N$	$df = n - 1$	$df = n - 1$	$df = n_1 + n_2 - 2$
MEAN		$\mu = \frac{\Sigma X}{N}$	$\bar{X} = \frac{\Sigma X}{n}$	$\bar{D} = \frac{\Sigma D}{n}$	$\bar{X}_1 = \frac{\Sigma X_1}{n_1}$
SUM OF SQUARES		$SS = \Sigma(X - \mu)^2$	$SS = \Sigma(X - \bar{X})^2$	$SS_D = \Sigma(D - \bar{D})^2$	$SS_1 = \Sigma(X_1 - \bar{X}_1)^2$
VARIANCE		$\sigma^2 = \frac{SS}{N}$	$s^2 = \frac{SS}{n-1}$	$s_D^2 = \frac{SS_D}{n-1}$	$s_p^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2}$
STANDARD DEVIATION	SD	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$	$s_D = \sqrt{s_D^2}$	THE STANDARD DEVIATIONS HERE ARE EMBEDDED BELOW WITHIN THE SEM

INFERENTIAL STATISTICS

USE THE SAMPLING DISTRIBUTION TO CACULATE THE HYPOTHESIS TEST STATISTIC AND/OR THE CONFIDENCE INTERVAL

SAMPLING DISTRIBUTION		SAMPLING DIST. OF z	SAMPLING DIST. OF t	SAMPLING DIST. OF $\bar{D}$	SAMPLING DIST. OF $\bar{X}_1 - \bar{X}_2$
STANDARD ERROR	SEM	STANDARD ERROR OF THE MEAN	ESTIMATED STANDARD ERROR OF THE MEAN	ESTIMATED STANDARD ERROR OF THE MEAN DIFFERENCE SCORES	ESTIMATED STANDARD ERROR OF THE DIFFERENCE BETWEEN TWO SAMPLE MEANS
		$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$	$s_{\bar{X}} = \frac{s}{\sqrt{n}}$	$s_{\bar{D}} = \frac{s_D}{\sqrt{n}}$	$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$

HYPOTHESIS TESTING

HYPOTHESIS TEST		ONE SAMPLE z-TEST	ONE SAMPLE t-TEST	RELATED SAMPLES t-TEST	INDEPENDENT SAMPLES t-TEST
HYPOTHESISED MEAN		$\mu_{hyp} = \mu_X$	$\mu_{hyp} = \mu_X$	$\mu_{hyp} = \mu_D$	$\mu_{hyp} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2}$
DATA RESULT	RATIO	$z_{obs} = \frac{\bar{X} - \mu_{hyp}}{\sigma_{\bar{X}}}$	$t_{obs} = \frac{\bar{X} - \mu_{hyp}}{s_{\bar{X}}}$	$t_{obs} = \frac{\bar{D} - \mu_{hyp}}{s_{\bar{D}}}$	$t_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_{hyp}}{s_{\bar{X}_1 - \bar{X}_2}}$
	AREA	p	p	p	p
CRITICAL CUTOFF	RATIO	z_{crit}	t_{crit}	t_{crit}	t_{crit}
	AREA	α	α	α	α

ESTIMATION

CONFIDENCE INTERVAL	$CI = \bar{X} \pm (z_{conf})(\sigma_{\bar{X}})$	$CI = \bar{X} \pm (t_{conf})(s_{\bar{X}})$	$CI = \bar{D} \pm (t_{conf})(s_{\bar{D}})$	$CI = \bar{X}_1 - \bar{X}_2 \pm (t_{conf})(s_{\bar{X}_1 - \bar{X}_2})$

Statistics for Data Science

ANOVA Test

ANOVA (Analysis of Variance) is used to check if at least one of two or more groups have statistically different means. Now, the question arises – Why do we need another test for checking the difference of means between ***independent groups***? Why can we not use multiple t-tests to check for the difference in means?

The answer is simple. Multiple t-tests will have a compound effect on the error rate of the result. Performing t-test thrice will give an error rate of ~15% which is too high, whereas ANOVA keeps it at 5% for a 95% confidence interval.

Analysis of Variance(ANOVA)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares (MS)	F
Within	$SS_w = \sum_{j=1}^{k}\sum_{j=1}^{l}(X-\bar{X}_j)^2$	$df_w = \mathrm{k}-1$	$MS_w = \frac{SS_w}{df_w}$	$F = \frac{MS_b}{MS_w}$
Between	$SS_b = \sum_{j=1}^{k}(\bar{X}_j - \bar{X})^2$	$df_b = \mathrm{n}-\mathrm{k}$	$MS_b = \frac{SS_b}{df_b}$	
Total	$SS_t = \sum_{j=1}^{n}(\bar{X}_j - \bar{X})^2$	$df_t = \mathrm{n}-1$		

Chi-square test

Sometimes, the variable under study is not a continuous variable but a categorical variable. Chi-square test is used when we have one single categorical variable from the population.

Let us understand this with help of an example. Suppose a company that manufactures chocolates, states that they manufacture 30% dairy milk, 60% temptation and 10% kit-kat. Now suppose a random sample of 100 chocolates has 50 dairy milk, 45 temptation and 5 kitkats. Does this support the claim made by the company?

Let us state our Hypothesis first.

Null Hypothesis: The claims are True

Alternate Hypothesis: The claims are False.

Chi-Square Test is given by:

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

Where:

O= observed Value

E= expected value

Let us now calculate the Expected values of all the levels.

E (dairy milk)= 100 * 30% = 30

E (temptation) = 100 * 60% =60

E (kitkat) = 100 * 10% = 10

Calculating chi-square = [(50-30)^2/30+(45-60)^2/60+(5-10)^2/10] =19.58

Now, checking for p (chi-square >19.58) using chi-square calculator, we get p=0.0001. This is significantly lower than the alpha(0.05).

So, we reject the Null Hypothesis.

_ END _

Date: 05-11-2022

Made by- Biranchi Narayan Maharana

[Digital Marketer & Data scientist]