The Statistics Handbook

MISC

Datasheet: <https://docs.google.com/spreadsheets/d/1QZQfAXP-4Zxy3mHXI9Q5HsAUQ4NCZU5AbX_az7w8IKo/edit#gid=1779679046>

TO BE DELETED  
  
<https://docs.google.com/document/d/1GFu31H4RBIG4WY-D7Q0aCdSzuxfOL3Xn21ZGmy3RWZ8/edit#>

LATEX:

<https://www.overleaf.com/learn/latex/Learn_LaTeX_in_30_minutes>

<https://www.learnlatex.org/en/>

<https://nasa.github.io/nasa-latex-docs/html/examples/section.html>

<https://www.overleaf.com/latex/examples/statistics-formula-sheet/kvttpvjhrznh>

<https://www.malinc.se/math/latex/basiccodeen.php>

<https://www.overleaf.com/learn/latex/List_of_Greek_letters_and_math_symbols>

<https://oeis.org/wiki/List_of_LaTeX_mathematical_symbols>

<https://www.papeeria.com/p/d32d16231aed3352be950486d4d68d80?withLastOpenedFile=true#/main.tex>

[**Scope of this handbook**](#_xxcepo8mcgw8) **4**

[**Core Concepts**](#_q55lifxinwb4) **5**

[Let’s start from a question](#_5zhqgxhpypt2) 5

[Property and type of data](#_fzu2iap0dprf) 6

[Random Variable](#_a3vfqg9d8trm) 6

[Continuous vs Discrete](#_pmzu30r55qvo) 6

[Nominal vs Ordinal](#_ob6zdn20svci) 6

[Structured vs Unstructured](#_uuicf4oyzzhv) 7

[Statistical Variables and their properties](#_9zs2rbpaf7uo) 7

[Qualitative statistical variables](#_qn2qcz6rwqc2) 7

[Quantitative statistical variables](#_b6wmc7x93ilm) 7

[Parametric vs Nonparametric](#_u5uyz0wilnt3) 8

[Homoscedasticity vs Heteroscedasticity](#_gylkx3y7cfmz) 8

[Population vs Sample](#_81ftg58gapqr) 9

[Parameters vs Statistics vs Hyperparameters](#_28xqlul0cg0n) 9

[Descriptive and inferential statistics](#_pxwardpejd2j) 9

[Binomial Coefficient](#_eba39cvbjyn) 10

[Binomial Distribution](#_a3qndk8ghjen) 10

[Measurement of Central Tendency](#_tmc92yw26amt) 10

[When to use mean, median and mode?](#_aa757b1r701) 12

[Hat symbols over variables (‘^’)](#_5sl3h3i1nahf) 12

[Measurement of Dispersion](#_b4lcvdfnbxen) 12

[Variance](#_rxkb7z3g5zv7) 12

[Standard Deviation](#_7rq3h250p0kc) 13

[Bessel’s Correction](#_xap1wg7b91md) 13

[Quartiles and IQR](#_hbsrqikwr21f) 13

[**Visualizing Data**](#_1ujzkkizksas) **14**

[Scatter Plot](#_6e7c9lop5jq3) 14

[Linear Regression](#_26bcy1docjqh) 15

[Line Plot](#_drdve0p06jt) 15

[Ogive](#_vezb5bfxedd) 16

[Histograms](#_cln5okoni5xb) 16

[Bar Plots](#_mp07uci6bpk8) 16

[Box and whisker vs candlestick chart](#_vewr5kbi13ln) 17

[Violin Plot](#_8wc1bpopvo27) 17

[KDE Plot](#_budlqlt6ezab) 18

[**Combinatorics**](#_n6y3r9a1jo2w) **18**

[Factorials](#_qjbqxman8k6q) 18

[Properties](#_c66ph5mbehvd) 18

[Factorials and 0](#_h1lesogrkz8m) 19

[Permutations](#_ey8tdoj8rxgj) 19

[Combinations](#_cxsu978bakhb) 19

[Dispositions](#_7majjmg3coha) 20

[Permutations vs Combinations with or without repeats](#_1psa9qeuad5k) 20

[When to use permutations, combinations or dispositions?](#_8zulniv6cl6p) 21

[**Probability**](#_l6puvoppr5tc) **22**

[Notation](#_j3fousbivnrc) 22

[Simple Probability](#_8jhxk3yc6hbd) 23

[Experimental and Expected probability](#_f7hq0sw00ld6) 23

[Law of Large Numbers](#_3xs6b8cjodo2) 23

[Addition Rule](#_2m9ke8yrn8nd) 23

[Fundamental rule for addition or product in probability calculation:](#_2t4tp2f0u7e) 24

[Conditional probability for Independent and Dependent events](#_dvxfzw4qavua) 25

[Bayes Theorem](#_wjy1j1qnlgyj) 25

[Tree Diagrams and Bayes’ Theorem](#_4vn06wuh7qg1) 26

[Discrete Probability](#_jh7v0nitff86) 27

[Discrete variable vs continuous variable](#_p4ni4efuyxv0) 28

[Discrete vs continuous probability distribution examples](#_o8hlki36fxh3) 28

[Transforming Random Variables](#_7cjctacd7611) 29

[Linear Combinations of Random Variables](#_ytcyfeg4cve0) 31

[Fair Game](#_yvfxv82x37aa) 32

[**Joint Distribution**](#_3njhrwpqcb15) **33**

[Covariance](#_a42mpi4h4phv) 33

[Correlation](#_jzrnz39nkl1) 34

[Pearson correlation](#_wrvb1aespcgb) 34

[For a sample](#_4nttlcnmd5ce) 35

[Kendall rank correlation](#_d2tvsu9j25oi) 36

[Spearman rank correlation](#_4t2u3re6pz00) 36

[Point-biserial correlation coefficient](#_uao39kklwrog) 37

[**Data Distribution**](#_nihaam2rv0yi) **37**

[Probability Mass Function (PMF)](#_9j1fg2dxl75e) 37

[Discrete uniform distribution](#_7hkiw07s28ng) 38

[Probability Density Function (PDF)](#_1at4g6y8ufdh) 38

[Continuous uniform distribution](#_u6cdxdbkq8fw) 39

[Cumulative Distribution Functions (CDF)](#_5r5nq7z99gde) 40

[Discrete CDF](#_qvpgjrlzjvp6) 40

[Continuous CDF](#_s56mw7q8mmka) 41

[Binomial Distribution](#_kqydi22fsoqe) 41

[Bernoulli Distribution](#_tscxzci50xrd) 43

[Poisson Distribution](#_kaklu3m1suak) 43

[**Normal Distribution**](#_mpth3qp7lzqa) **45**

[Z-Score](#_hemii53qi13h) 46

[Z-Tables](#_lww7powbobkj) 47

[Normality test](#_tmp3578d6k7y) 47

[Mean, Variance and Standard Deviation for normal distribution](#_a8ojftp5v2ex) 48

[Skewed Distribution (Skewness)](#_jmzg9n2r2wi0) 48

[Kurtosis](#_gxuqdwodevid) 49

[Standard Normal Distribution](#_jxdyet94z31i) 49

[**Sampling**](#_ktf4wd2npivi) **49**

[Sampling Methodologies](#_2ssaa4vxz11t) 49

[Simple Random Sample (SRS)](#_4734q1qeqrwh) 50

[Systematic Random Sample](#_wytjdg5ruzr) 50

[Stratified Random Sample](#_qunclcyvnzj8) 50

[Clustered Random Sample](#_he2apcx5yg5n) 50

[Central Limit Theorem](#_h90v6b2l4eig) 50

[Sampling Distribution of the Sample Mean (SDSM)](#_xikpimxcgvyl) 50

[More on mean of the sampling being equal to the mean of the population](#_euhsrdcoo3w6) 51

[Finite Population Correction Factor (FPC)](#_q77ouk3bo6cz) 51

[The Student's T-Distribution](#_cwb2y0l1g8ck) 51

[Degrees of freedom (DF)](#_ieqedfxopfkr) 52

[T-Score](#_20ubmuxlyvju) 52

[Why “Student Distribution”?](#_i47xiq8if881) 53

[When to use Z or T distribution?](#_fxbr9txd8sx3) 53

[Confidence interval for the mean](#_9tsnzwsh3t5e) 53

[Point Estimation](#_7c6k5e6b6h55) 53

[Interval estimation](#_t364lf7e8zdu) 53

[Confidence interval (CI)](#_d3izi2j07srj) 54

[Confidence Level (CL)](#_9j501tlus6cf) 54

[Confidence interval and Z-scores](#_gjwt43j75xwn) 54

[**Hypothesis Testing**](#_n90tc5nkkdxt) **54**

[Null & Alternative Hypothesis](#_l119k42y7y5h) 55

[Type I and Type II Errors](#_372it23atdru) 55

[Trade-off between Type I and Type II errors](#_q75l7x8j08vt) 55

[Test Statistics](#_tyzadinv93l4) 56

[Two-Tailed and One-Tailed Tests](#_otana98or5u0) 57

[P-Value and Critical Value](#_bvcctxov9x30) 57

[A/B Testing](#_bj2xsq6gmchi) 58

[**Regression Analysis**](#_630ynlhcfvm) **58**

[Ordinary Least Squares (OLS)](#_7u6duwt76jcb) 59

[Correlation Coefficient](#_sm8tdi1vdxjb) 59

[Line Fitting and Residuals](#_y1c1a983ypmw) 60

[Linear Regression Trendlines](#_g0vvcateu1dx) 60

[Regression model evaluation metrics](#_t32bzeb7m2s6) 61

[Chi-Square Test](#_6hkj04e4ir5t) 62

[Analysis of Variance (ANOVA)](#_x0eawc1lee6q) 62

[**License**](#_8jk15lwhpea1) **63**

[**Source Code**](#_u8mp9u4z5bja) **64**

[**Bibliography & Sources**](#_m168hv2w9ylv) **64**

[**Acknowledgement**](#_vd656xwmtl33) **65**

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# Scope of this handbook

We are used to talking generally about mathematical skills, thinking perhaps of derivatives, integrals, theorems, and graphs of functions.

Often we do that in an abstract way, as if they were certainly logical elements, but with just specific applications. Instead, we forget that not only are mathematical elements present in every single action, but that quantitative sciences are components of everyday life.

Specifically, I believe that statistics is among all the mathematical sciences the most fascinating because of the vastness and incredible opportunities for its application.

Every decision we make can be traced back to statistical phenomena, either innate (such as fear of the dark, because in the dark increases the likelihood of dangerous animals) or conscious (today I think it's likely to rain, so I'll take my umbrella).

On the other hand, approaching even basic statistical calculations (e.g., the infamous probability of winning the lottery) requires nontrivial skills in order to apply concepts and formulas that are not always complex but certainly have dissimilar results if used thoughtlessly. I claim for certain that worse than the lack of mathematical thinking is the misuse of mathematical thinking. This paper of mine is also in fact intended to combat my limitations through study and applications.

In this handbook, I wanted to create a path from the basics, including terminology (often one of the main obstacles for the laymen approaching the subject), to formulations of hypotheses, validations, and verification of formulas.

The path was constructed by consulting a large number of sources, cited in the appendix, and during long months of study and in-depth proof of the results and evidence, precisely because first and foremost I wanted to verify my own expertise, even before, of course, I could write about it.

Before releasing this publication, which is distributed under a Creative Common and Free Culture license, I asked for a check from eminent acquaintances with important academic and working backgrounds. I would like to endlessly thank all of them (their names can be found in the appropriate section). Nevertheless, I am staying receptive to additions, insights and corrections, taking full responsibility for any shortcomings and errors, certainly reported in good faith.

Happy reading!

Carlo, 4th of January 2023.

# Core Concepts

## Let’s start from a question

“What is data?”

Data are collected observations and information about a given phenomenon.

“What is statistics?”  
  
Statistics Is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.

“What is a statistical variable?”

It’s the specific characteristic being analyzed among the statistical units on which the statistical analysis is focused on, such as “age” from all the data that may be related to the object “person”. Classification of variables and their measure of scale is paramount to set up the analytical process of statistical analysis.

## Property and type of data

Data doesn’t have to be numeric.

Data can be continuous vs discrete; structured vs unstructured; nominal vs ordinal; population vs sample. And more.

## Random Variable

A random variable (also called random quantity, aleatory variable, or stochastic variable) is a mathematical formalization of a quantity or object which depends on random events.

It is a mapping or a function from possible outcomes in a sample space to a measurable space, often the real numbers.

Informally, randomness typically represents some fundamental element of chance, such as in the roll of a dice.

## Continuous vs Discrete

Discrete means it can only take certain values. There are no “in-between” values.

Discrete data is the number of people: there could be 1 person or 2 people, but not 1,5 people or 0,99 people.

Discrete data are the possible value of rolling a die: 1,2,3,4,5,6 and not 6.5 or 1.5

Continuous means there is an infinite amount of value in between each data point.

Continuous data is the height or weight of a person.

Continuous data are temperature records.

## Nominal vs Ordinal

Nominal data is classified without a natural order or rank. Nominal data can’t be clearly sorted.

Nominal data can’t be “ordered” (from which the term “ordinal”).

Nominal data are animal species: lizard, dog, cat, or the list of ingredients on a recipe.

Ordinal data is data that has a natural order or rank.

Ordinal data can be sorted and ordered.

Ordinal data doesn’t have to be numeric. For example, hot, mild, cold - or even top, low, bottom, can be data attributes that can be ordered and then being considered ordinal.

Ordinal data are the seat numbers on a train.

## Structured vs Unstructured

Structured data is highly specific and stored in a predefined format. It has its own structure.  
Examples are JSON or Excel files, SQL databases.

Unstructured data is data that does not have a specific or well defined format.

Unstructured data are audio data, text data, video data.

Do not confuse “file format” with “formatted data”.

Just because text is in a PDF format doesn't make it structured data.

## Statistical Variables and their properties

### Qualitative Statistical Variables

Qualitative statistical variables are variables whose values are not numbers but modes, or categories.

Examples are: “male” or “female”, “education”, “marital status”, “ethnicity” and such.

Those categories have to be exhaustive and mutually exclusive - a datapoint can’t be both “male” and “female” or both “asian” and “european”. This is a specific problem that may occur in the data preparation and data gathering phase.

Qualitative statistical variables can be classified further in:

**Dichotomic:** variables that have only two kinds of mutually exclusive categories, such as “male” or “female” or “alive” or “dead”.

**Nominal:** variables that have no logical order, are not comparable and not exclusive to each other. Examples of nominal variables are “transportation used for work” or “sport played”.

**Ordinal:** variables that have a logical predefined order, but yet can’t be classified as quantitative.

Example is “education”; High School is surely lower than University, but of how much?

And how far is a MsC from a PhD? They are clearly different, but this difference can’t be clearly measured.

* **Linear ordinal**: they have a clear start and end, such as size “S M L XL”.
* **Cyclical ordinal:** they have no clear start and end and their order is based on convention (such as week days: weeks starts both on Monday, or on Sunday. Seasons).

### Quantitative statistical variables

Quantitative statistical variables are expressed by a numerical quantity.

Quantitative data is naturally ordinable and comparable.

Quantitative data can be further classified in:

**Interval data:** datapoint are expression of a specific point of the dataset (such as result of a test, QI, temperature)

**Ratio scale data:** data that is expressed by a rate, such as age and weight.

### Parametric vs Nonparametric

**Parametric tests**

* Parametric tests assume the presence of distributions of approximately normal type.
* They involve continuous or interval-type variables and a fairly large sample size.
* They assume homogeneity of variances (homoscedasticity).
* They assume estimation of parametric data such as mean, variance and standard deviation.

These tests have higher statistical power because they provide a higher probability of correct rejection of an incorrect statistical hypothesis.

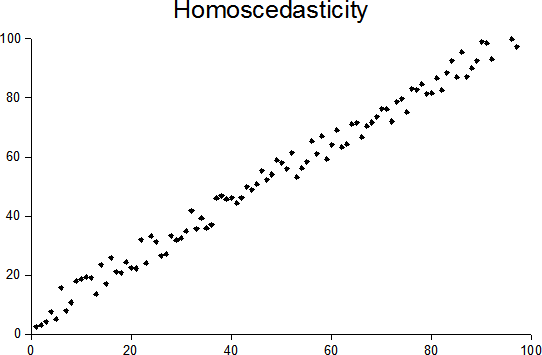
**Nonparametric tests**

Nonparametric tests don’t imply any kind of distribution and don’t imply any kind of parametric estimation such as mean, variance and standard deviation (because, for example, such measures are not estimable).

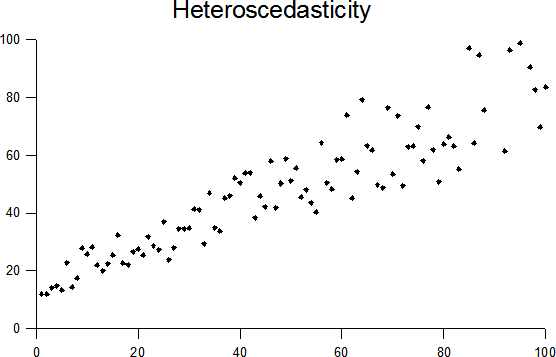
Nonparametric tests should be preferred whenever the dataset is not distributed in a normal (gaussian distribution) way, or, in any case, this specificity is not being demonstrated. A typical example is whenever the dataset is too small to prove a parametric distribution.

### Homoscedasticity vs Heteroscedasticity

**Homoscedasticity** means that all random variables in the dataset have the same finite variance.



**Heteroscedasticity** means that not all random variables in the database have the same finite variance.



## Population vs Sample

Population consists of the representation of every member of a given group or of the entire available data set.

Examples are all the students of a class or all the animals of a specific national park.

Sample refers to a subset of the entire data set. For example, the first 10 students of a class or the top 3 predators from a specific national park.

Population and Sample are data definitions that are heavily dependent from the context.

When analyzing data related to a population, we need to be sure to include a statistically relevant sample. A representative sample.

Sample sizes are a well studied science.

For example 300 is a representative sample out of a population of 1000 students.

Use “population” methods when:

* It’s known the dataset is related to the entire population.
* A generalization to a wider, larger population is not interesting.

Use “sample” methods when:

* It’s known the dataset is related to a subset of the whole dataset
* A generalization to a wider, larger sample or population is interesting

Rule of thumb: statisticians primarily work with samples. Real-world data can be overwhelmingly large.

## Parameters vs Statistics vs Hyperparameters

**Parameters** describe the properties of the entire population.

**Statistics** describe the properties of a sample.  
  
**Hyperparameters[[1]](#footnote-0)** (used in modeling and machine learning processes) are instead tuning values. Hyperparameters are set before the model is trained and are not coming from the dataset.

## Descriptive and inferential statistics

**Descriptive statistics** is a part of statistics that aim to describe data. It is used to summarize the attribute of a dataset, using measures such as Measures of Central Tendency or Measures of Dispersion.

**Inferential statistics** is a part of statistics that is used to test and validate assumptions over a dataset by analyzing a sample, using methods such as Hypothesis Testing or Regression Analysis.

## Binomial Coefficient

Binomial Coefficient is a natural number as defined starting from a pair of natural numbers, usually named *n* and *k.*

Binomial Coefficient represents the number of sub-groups of *k* elements that could be made out of a dataset of *n* objects.

## Binomial Distribution

In probability theory and statistics, the binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes–no question, and each with its own Boolean-valued outcome: success (with probability p) or failure (with probability q=1-p). A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment, and a sequence of outcomes is called a Bernoulli process; for a single trial, i.e., n = 1, the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance.

## Measurement of Central Tendency

**Central tendency** is defined as “the statistical measure that identifies a single value as representative of an entire distribution.” It aims to provide an accurate description of the entire data. It is the single value that is most typical/representative of the collected data.

**Mean:**

Mean is generically expressed as:

sum of all data points / number of data points

Mean is the same concept of “average”, but average is generally used in arithmetic, while “mean” is expressingly considering the central point among a dataset in statistics. Arithmetic Mean is equal to average, while Harmonic or Geometric Mean have different meanings.

Mean can be expressed also with symbols:

called MU or even with x bar x̄

in the specific context of statistical studies,

x̄ is used for mean over a sample

is used for mean over the entire population



There are many different types of Mean, such as:

**Arithmetic Mean**:

It’s expressed as the sum of all data points / number of data points.

It’s the easiest kind of mean.

**Weighted Mean:**

Here the weight of each item is taken into account for computing the mean.  
  
For example, the average weight of an apple, given that you have many apples with different weight clusters.

| Apple (n) | Weight (g) |
| --- | --- |
| 8 | 200 |
| 3 | 250 |
| 8 | 100 |

The weighted mean would be then: ((8\*200)+(3\*250)+(8\*100))/(8+3+8)

**Truncated Mean:**

A truncated mean or trimmed mean is a statistical measure of central tendency, much like the mean and median. It involves the calculation of the mean after discarding given parts of a probability distribution or sample at the high and low end, and typically discarding an equal amount of both.

This number of points to be discarded is usually given as a percentage of the total number of points, but may also be given as a fixed number of points.

I.e. in the given dataset:

16 18 21 27 32 32 33 91 would make sense to trim 16 and 91 since they looks as outliers to our dataset.

**Mode:**

The mode is the value occurring most often in a dataset

dataset = 8, 5, 4, 27, 35, 8, 29

mode = 8

dataset = 8, 5, 4, 27, 35, 8, 29, 35

it’s a bi-modal dataset, mode being 35 and 8

dataset = 5, 4, 27, 35, 8, 29

mode =

**Median:**

The median is the central value of an ordered dataset.

odd number of items dataset:

16, 18, 21, 27, 32, 33, 91

median = 27

even number of items dataset:

16, 18, 21, 27, 32, 32, 33, 91

median = (27 + 32) / 2 = 29.5

### When to use mean, median and mode?

| **DATASET** | **MEAN** | **MEDIAN** | **MODE** |
| --- | --- | --- | --- |
| **Continuous** | YES | YES | YES |
| **Discrete** | YES | YES | YES |
| **Nominal** | MAYBE | NO | YES |
| **Ordinal** | MAYBE | YES | YES |
| **Numeric** | YES | YES | YES |
| **Non-numeric** | NO | YES | YES |

## Hat symbols over variables (‘^’)

The estimated or predicted values in a regression or other predictive model in statistics are termed the “hat values”.

example for ŷ: "Y" because y is the outcome or dependent variable in the model equation, and a "hat" symbol (circumflex) placed over the variable name is the statistical designation of an estimated value.

## Measurement of Dispersion

Measures of dispersion can be defined as positive real numbers that measure how homogeneous or heterogeneous the given data is.

The most common measurements of dispersion are Variance and Standard Deviation.

### Variance

Variance represents the positive distance from a single datapoint for the mean of the dataset. The positive distance is made thanks to the exponential factor applied to the distance of each datapoint.

The exponential factor also magnifies values that are more far from the mean in respect to smaller values, allowing to better understand their impact on the dataset.

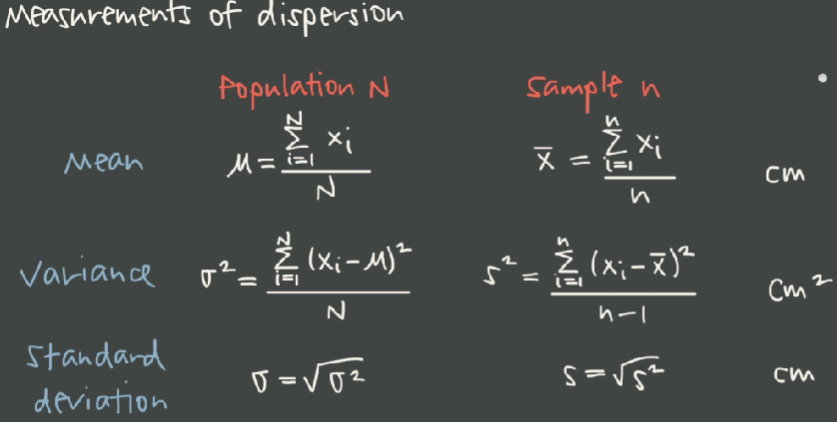
Variance is represented with the greek letter sigma squared (when used for population) or s squared when used for sample of population.

### Standard Deviation

Standard Deviation is the square root of Variance and it’s represented with just the greek sigma () or the letter s.

Being square rooted, the SD returns a value that has again the same scale of our dataset, hence allowing for better comparisons and understanding.

Hence, Mean, Variance, and Standard Deviation, are closely linked together.



### Bessel’s Correction

Why does Sample Variance have n-1 as denominator?

That’s a good question, that leads to a non-trivial answer.

From a mathematical point of view, the -1 correction factor is called Bessel’s correction and it’s used to correct the tendency (that can be demonstrate mathematically or even empirically with a relatively small number of experiment over a dataset) that the biased estimator has a tendency to undershoot (and never to overshoot) the parameter being estimated.

It is possible to think of the Bessel correction as the degrees of freedom of the vector of residuals. When the sample standard deviation is calculated from a sample of n values, sample mean is used which has already been calculated from that same sample of n values. The calculated sample mean has already taken into account one of the degrees of freedom of variability (which is the mean itself) that is available in the sample.

Let's approach the topic with an example: we have a table with 10 dice rolls; we know the result of each die, the overall average of the dataset.

How many elements can we make unknown in our dataset, without altering the goodness of the information we have?

Only one. By eliminating the result of one die roll, we are still able to reconstruct it through the mean of the experiment and the remaining values.

But by eliminating more than one value, we are forced to add approximation, thus invalidating the info we possess.

This is why we can link Bessel correction to degrees of freedom.

<https://towardsdatascience.com/why-sample-variance-is-divided-by-n-1-89821b83ef6d>

<https://stats.stackexchange.com/questions/16008/what-does-unbiasedness-mean>

<https://stats.stackexchange.com/questions/198452/why-sample-variance-has-has-n-1-in-the-denominator>

<https://medium.com/analytics-vidhya/explaining-n-1-the-intuition-behind-bessels-correction-for-sample-variance-daee48bd6fde>

## Quartiles and IQR

In statistics, a quantile is a type of quantile which divides the number of data points into four parts, or quarters, of more-or-less equal size. The data must be ordered from smallest to largest to compute quartiles; as such, quartiles are a form of order statistic. The three main quartiles are as follows:

The first quartile (Q1) is defined as the middle number between the smallest number (minimum) and the median of the data set. It is also known as the lower or 25th empirical quartile, as 25% of the data is below this point.

The second quartile (Q2) is the median of a data set; thus 50% of the data lies below this point.

The third quartile (Q3) is the middle value between the median and the highest value (maximum) of the data set. It is known as the upper or 75th empirical quartile, as 75% of the data lies below this point.

Il secondo quartile coincide con la mediana, e divide la popolazione in due parti di uguale numerosità, delle quali il primo ed il terzo quartile sono le mediane.

Il quartile zero coincide con il valore minimo della distribuzione. Il quarto quartile coincide con il valore massimo della distribuzione.

I quartili equivalgono ai quantili q0 (quartile zero), q1/4 (primo quartile), q2/4=q1/2 (secondo quartile), q3/4 (terzo quartile) e q1 (quarto quartile).

dataset = 1, 2, 3, 5, 8, 8, 9, 10, 15

Q0: 1

Q1: (2 + 3) / 2 = 2.5 (median of first half; 25th percentile)

Q2: 8 (median; 50th percentile)

Q3: (9+10) /2 = 9.5 (median of second half;75th percentile)

Q4: 15

Range = Q4 - Q0 = 15 - 1 = 14

IQR = InterQuartile Range = Q3 - Q1 = 9.5 - 2.5 = 7

# Visualizing Data

## Scatter Plot

A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.

## Linear Regression

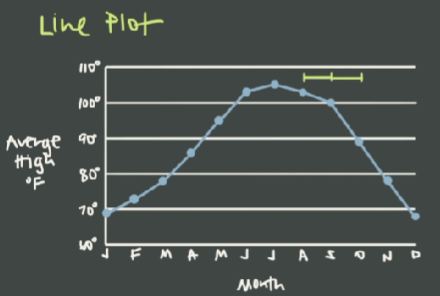
Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Linear Regression is used to find Trendline in a dataset, usually in Scatter Plot charts.

## 

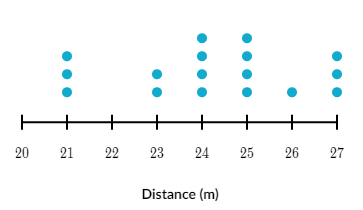
## Line Plot

the line plot is similar to a traditional line chart, with data point connected via a line.



However, online I found another definition, that is the following:

A line plot is a way to display data along a number line. Line plots are also called dot plots.

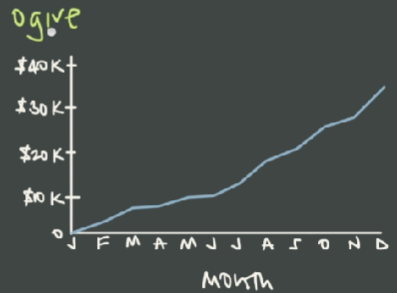


## Ogive

An ogive (oh-jive), sometimes called a cumulative frequency chart, is a type of frequency chart that shows cumulative frequencies. In other words, the cumulative percentages are added on the graph from left to right.

An ogive graph plots cumulative frequency on the y-axis and class boundaries along the x-axis. It’s very similar to a histogram, only instead of rectangles, an ogive has a single point marking where the top right of the rectangle would be.

It is usually easier to create this kind of graph from a frequency table.



https://www.geeksforgeeks.org/ogive-cumulative-frequency-curve-and-its-types/

## Histograms

A histogram is a bar chart that groups continuous data into ranges. Ranges are discretional to the creator of the chart. For example, overall user ages (continuous dataset) can be grouped in clusters such as 0-10, 10-20 and such.

Histogram bars are adicent (no spaces between bars).  
  
Histograms don’t have to be confused with bar charts.

Histograms visualize quantitative data or numerical data, whereas bar charts display categorical (discrete) variables.

In most instances, the numerical data in a histogram will be continuous (having infinite values). Attempting to display all possible values of a continuous variable along an axis would be foolish.

## Bar Plots

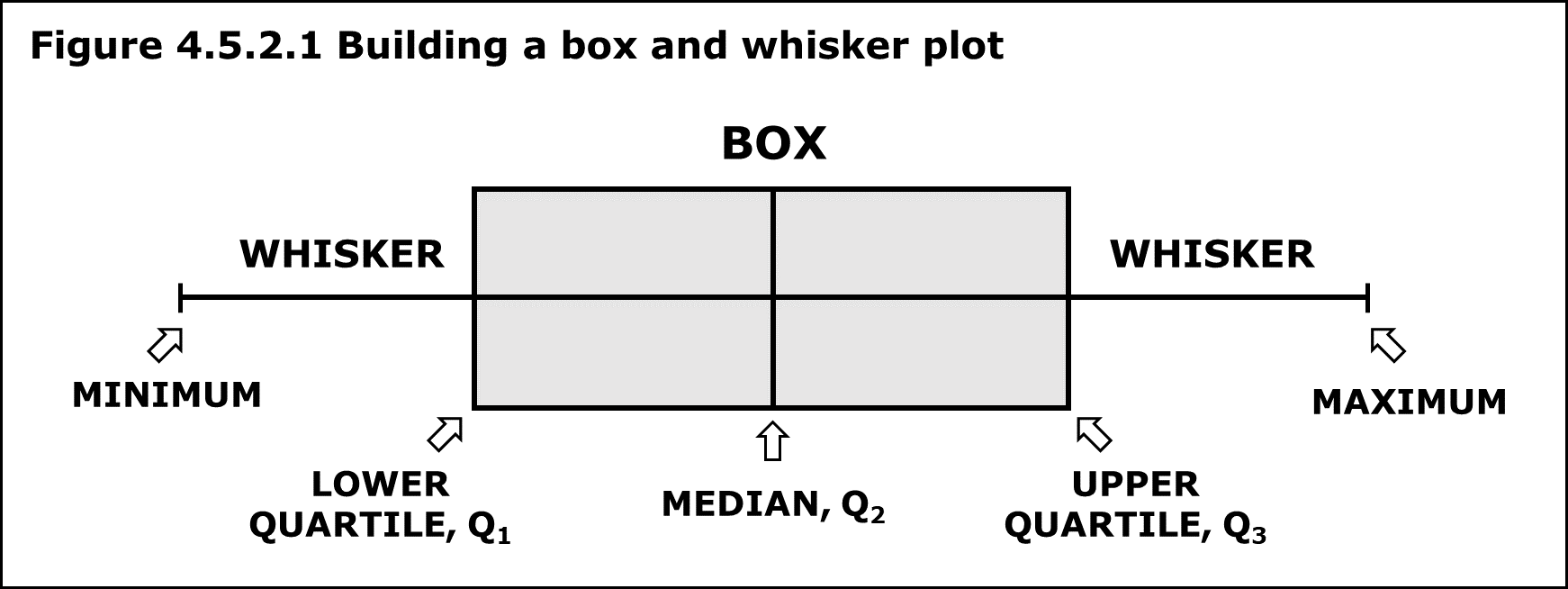
Bar plots are usually used to display categorical data along the horizontal axis - that is, discrete data, such as products, countries, car types and such.

Bar within a bar chart are not adicent. Data on the bar plots are often ordered, in order to enhance chart comprehension.

## Box and whisker Plots

Also called: box plot, box and whisker diagram, box and whisker plot with outliers (in italiano: diagramma a scatola e baffi).

A box and whisker plot is defined as a graphical method of displaying variation in a set of data. It is usually used to display data according to quartile intervals.



### Box and whisker vs candlestick chart

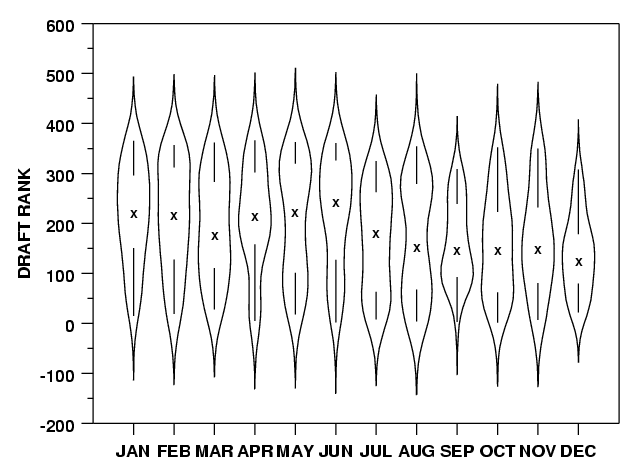
Mathematically speaking there is no difference. Both show an upper and lower boundary and points outside these boundaries. However:

A candlestick chart is mainly used in the finance industry. Its most popular application is to show share price. It is mainly used in the vertical position.

A box and whisker chart tends to be used in non-finance industries. For example, the level of sales of various stores or inventory levels etc. The box and whisker can be shown horizontally as well as vertically. They are often found with labels showing various statistical informations.

## Violin Plot

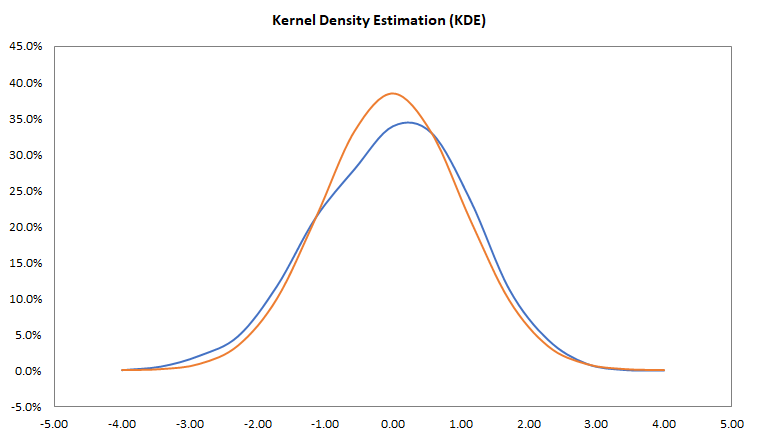
A violin plot is a method of plotting numeric data. It is similar to a box plot, with the addition of a rotated kernel density plot on each side.



## KDE Plot

KDE Plot described as Kernel Density Estimate is used for visualizing the Probability Density of a continuous variable. It depicts the probability density at different values in a continuous variable. We can also plot a single graph for multiple samples which helps in more efficient data visualization.

Kernel density estimates are closely related to histograms, but can be endowed with properties such as smoothness or continuity by using a suitable kernel.



# Combinatorics

Combinatorics is an area of mathematics primarily concerned with counting, both as a means and an end in obtaining results, and certain properties of finite structures. It is closely related to many other areas of mathematics and has many applications ranging from logic to statistical physics and from evolutionary biology to computer science.

## Factorials

In mathematics, the factorial of a non-negative integer **n**, denoted by **n!**, is the product of all positive integers less than or equal to **n**. The factorial of n also equals the product of n with the next smaller factoria.

5! = 5 \* 4 \* 3 \* 2 \* 1 = 120

### Properties

**n! = n \* (n-1)!**

5! = 5 \* (4!)

Practical application:

n! / (n-1)! = n \* (n-1)! / (n-1)! = n

### Factorials and 0

Factorials deal only with natural numbers, hence 0 is omitted in the series (otherwise n! = 0).

But why **0! = 1** ?  
  
It’s proven that:

(n-1)! = n! / n

this means that:

4! = 24

3! = 24 / 4 = 6

2! = 6 / 3 = 2

1! = 2 / 2 = 1

0! = 1 / 1 = 1

And, following the same logic:

-1! = 1 / 0 = ND

that’s why n! if

## Permutations

A permutation of a set of objects is an arrangement of the objects in a **certain order**.

Permutations differ from combinations, which are selections of some members of a set regardless of order.

Usually we refer to permutations referring to all the possible arrangements (all the possible permutations of a set of objects).

Permutations are calculated as factorials (n!)

Permutations are relevant when working with numbers, since 575 is not equal to 577 nor 557.

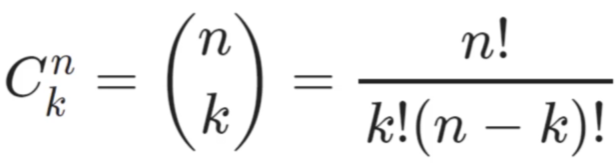
## Combinations

In mathematics, a combination is a selection of items from a set that has distinct members, such that the order of selection does not matter (unlike permutations). For example, given three fruits, say an apple, an orange and a pear, there are three combinations of two that can be drawn from this set: an apple and a pear; an apple and an orange; or a pear and an orange. More formally, a k-combination of a set S is a subset of k distinct elements of S.

Combinations is an **unordered** selection of objects from a set of objects.

Combinations are relevant when working with products, or people: apple and orange is equal to orange and apple. A team with Mark and Tom is equal to a team with Tom and Mark.

The number of combinations from a set of **n** objects taken **k** a time is:



## Dispositions

Dispositions are not specifically mentioned in the course, but since I feel they are relevant, I’d like to explore them.

Dispositions is in fact a particular sub-case of permutations, when we need to select objects from a dataset of *n* elements and arrange them in *k* places, being *n > k* and, as in permutations, the order is relevant.   
  
This case in the course is being managed under permutations with no repetitions.

## Permutations vs Combinations with or without repeats

|  | **Repeats** | **No Repeats (also: simple)** |
| --- | --- | --- |
| **Permutations** | n^k | n! / (n - k)!  where k = cluster size  n! / (k1! \* k2! \* kn!)  where k = items repeated |
| **Combinations** | (n + k - 1)! / k!(n -1)! | n! / k!(n - k)! |
| **Dispositions** | n^k | n! / (n - k)! |

Very good chart with explanation: <https://www.matematika.it/public/allegati/39/15_04_Calcolo_combinatorio_1_0.pdf>

**Permutations with repetitions example**

How many phone numbers of 7 digits can we generate using all the numbers from 0 to 9, allowing every specific case (such as all zeros)?

n = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9] = 10

k = \_ \_ \_ \_ \_ \_ \_ = 7

Each slot of k allows 10 combinations, so it’s n to the power of k: 10^7

**Permutations with no repetitions example**

Find all the way I can arrange the word MAMA

n = 3

k1 = 2 (the letter M is repeated 2 times)

k2 = 2 (the letter A is repeated 2 times)

4! / (2! \* 2!) = 24 / 4 = 6

The example shown in the lecture is speaking specifically to the case of a cluster that is a subset of the original dataset of items we don’t want to repeat within the permutations, p.e.

How can we arrange 5 students in 3 chairs?

n = 5 (all the students we have to pick from)

k = 3 (seats available)

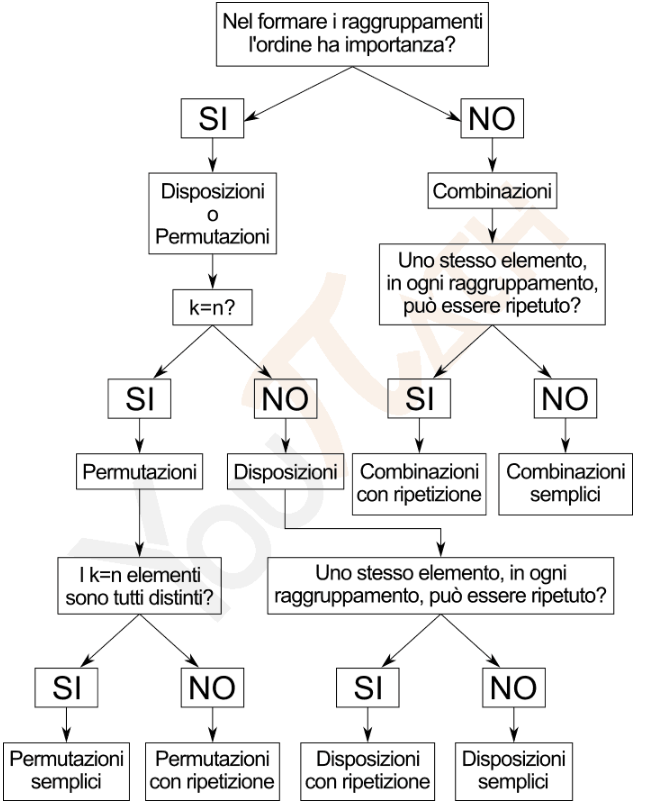
5! / (5 - 3)! = 120 / 2 = 60

## When to use permutations, combinations or dispositions?

Useful link with explanation (with the diagram quoted below) <https://www.youmath.it/lezioni/probabilita/calcolo-combinatorio/1218-risolvere-i-problemi-di-calcolo-combinatorio.html>

Key insight::

* Is the order relevant?
  + YES = PERMUTATIONS (es. lottery numbers)
  + NO = COMBINATIONS (es. people in teams)



# Probability

Probability is the branch of mathematics that deals with how likely an event is to occur, or how likely is that a given proposition is true.

## Notation

| P(A) | Individual probability | The probability of event A happening |
| --- | --- | --- |
| P(A’) | Complement | The probability of event A not happening |
| P(A B) | Union | The probability of both A and B happening for both datasets (all elements of A plus all elements of B). |
| P(A B) | Intersection | The probability of event A and B happening for the elements at the intersection of the datasets (only the elements A and B have in common). |
| P(A | B) | Dependent | The probability of A given that B has occurred |

**Example:**

If P = { 1,3,5,7,9} and Q = { 2,3,5,7}

What are P ∪ Q, and P ∩ Q

**Solution:**

P ∪ Q = { 1,2,3,5,7,9}

P ∩ Q = { 3,5,7}

## Simple Probability

P(A) = matching outcomes / total outcomes

P(A’) = 1 - P(A)

### Experimental and Expected probability

Experimental probability is the probability resulting from empirical experimentations, such as flipping a coin 100 times and recording the results in a datasheet.  
No matter the result of such an experiment, we still know that the expected probability of a coin toss being H or T is still 50%.

### Law of Large Numbers

The more a specific case an experiment is run, the more the experimental probability will match the expected probability.

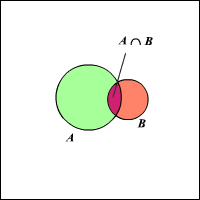
### Addition Rule

If A and B are two events in a probability experiment, then the probability that either one of the events will occur is:

P(A or B)=P(A)+P(B)−P(A and B)

This can be represented in a Venn diagram as:

P(A∪B)=P(A)+P(B)−P(A∩B)



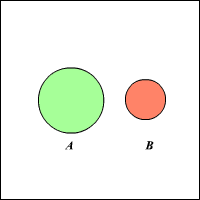
If A and B are two mutually exclusive events,

P(A∩B)=0 . Then the probability that either one of the events will occur is:

P(A or B)=P(A)+P(B)

This can be represented in a Venn diagram as:

P(A∪B)=P(A)+P(B)



Hence, as a recap:

P(A U B) = P(A) + P(B) - P(A B) (for intersections)

P(A U B) = P(A) + P(B) (for mutually exclusive events)

### Fundamental rule for addition or product in probability calculation:

Given two **independent** events, the probability of them **occurring both** is given by the **product** of the individual probabilities.

Example: having a head out of two coin flips.

The probability of two or more **alternative** events occurring is equal to the sum of the individual probabilities.

Example: having 1 or 2 out of a dice roll.

### Conditional probability for Independent and Dependent events

**Independent event probability**

The probability of A and B happening.

P(A B) = P(A) \* P(B)

Tossing two coins A and B, what is the probability of having two head values?

Coins are independent each other, so:

P(A B) = 1 / 2 \* 1 / 2 = 1 / 4

Defective rate in a production line is 2%. What is the probability of having 3 defective products in a row?

P(A B C) = 2 / 100 \* 2 / 100 \* 2 / 100 = 8 / 100^3 = 1 / 125’000

**Dependent event probability**

The probability of A and B, given that A has already occurred.

P(A B) = P(A) \* P(B | A)

What’s the probability of drafting two Kings in a row from a standard deck of cards?

P(A B) = 4 / 52 \* 3 / 51 = 1 / 13 \* 1 / 17 = 1 / 221

**Reminder**

If P(B) = P(B | A) then the events must be independent.

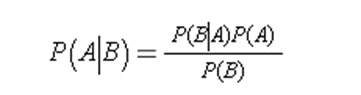
## Bayes Theorem

P(A) is the probability of event A

P(B) is the probability of event B

P(A | B) is the probability of observing event A if B is true

P(B | A) is the probability of observing event B if A is true.



**Example A)**

We have two assembly lines, 1 and 2.

Line 1 has a defective rate of 3%, Line 2 of 1%.

Given a defective part, what is the probability that it came from line 1?

Let’s understand our variables first.  
Let’s call:

P(B) the probability of a product being defective.

P(A) the probability of a product coming from line 1.

**P(A)** = 1 / 2

(with the available data, we must assume we have a 50% likelihood from two lines).

**P(B | A)** = probability of B(defect) if A(product is coming from line 1) has occurred = 3 / 100 (this is the info provided by the context already).

**P(B)** = overall probability of having a defective product = (brackets are only for added clarity) = [(1 / 2) \* (3 / 100)] + [(1 / 2) \* (1 / 100)] = 3 / 200 + 1/ 200 = 4 / 200 = 1 / 50

Applying Bayes Theorem, then:

**P(A | B)** = [(3 / 100) \* (1 / 2)] / (1 / 50) = (3 / 200) / (1 / 50) = (3 / 200) \* (50 / 1) = 150 / 200 = **3 / 4 = 75%**

**Example B)**

You’re tested for a disease that occurs 1 out of 1’000 people.

Test accuracy is 99%.

You are tested positive.  
What is the change you actually have the disease?  
  
Population: 1’000

Incidence: (1/1’000) = 0.001

Accuracy: 99%

False positive\negative: (100% - 99%) = 1%

|  | SICK | NOT SICK | TOTAL |
| --- | --- | --- | --- |
| TESTED POS | 0.99 [1] | 9.99 [3] | 10.98 |
| TESTED NEG | 0.01 [2] | 989.01 [4] | 989.02 |
| TOTAL | 1 | 999 | 1’000 |

[1] = 1’000 \* 0.001 \* 99%

[2] = 1’000 \* 0.001 \* 1%

[3] = (1’000 \* (1 - 0.001)) \* 1%

[4] = (1’000 \* (1 - 0.001)) \* 99%

P(A) = probability of being sick = 0.001  
P(B) = probability of having a positive test = 10.98 / 1’000

P(B|A) = probability of having a positive test being sick = 0.99

P(A|B) = probability of being sick having a positive test = 10.98 / 0.99 (or applying Bayesian formula) = 9%

### Tree Diagrams and Bayes’ Theorem

A type of diagram that can be used as an aid in computing probabilities is a tree

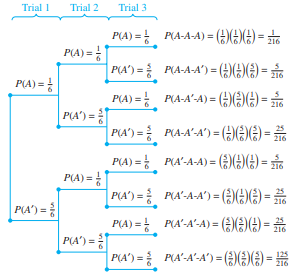
diagram. For example, consider an experiment of tossing a six-sided die. Each

time the experiment is repeated, the probability of obtaining a 1 (event ) is P(A) = 1 / 6

If you are only concerned with whether the number is 1 or not 1

(event or event ), and the experiment is repeated three times, then eight different sequences of events are possible. The tree diagram below shows

the probabilities of these eight sequences of events.



## Discrete Probability

A discrete probability distribution counts occurrences that have countable or finite outcomes.

This is in contrast to a continuous distribution, where outcomes can fall anywhere on a continuum.

Common examples of discrete distribution include the binomial, Poisson, and Bernoulli distributions.

Example of of discrete probability:

What is the probability of having head out of 3 coin flips?

Number of variable: 2 (head or tail).

Number of events: 3 flips.

Total number of combinations: 2^3 = 8

Possibile outcomes:

| Events | Number of head outcomes |
| --- | --- |
| HHH | 3 |
| THH | 2 |
| HTH | 2 |
| TTH | 1 |
| HHT | 2 |
| THT | 1 |
| HTT | 1 |
| TTT | 0 |

| Number of head outcomes in 3 coin flips (X) | Probability P(X) |
| --- | --- |
| 0 | 1/8 |
| 1 | 3/8 |
| 2 | 3/8 |
| 3 | 1/8 |

P(X) = 0 \* ⅛ + 1 \* ⅜ + 2 \* ⅜ + 3 \* ⅛ = 12 / 8 = 3 / 2 = 150 %

Mean = 3 / 2

Variance (2 ) = (0 - 3/2)^2 \* ⅛ + (1 - 3/2)^2 \* ⅜ + (2 - 3/2)^2 \* ⅜ + (3 - 3/2)^2 \* ⅛ = 0,75

Standard Deviation () = 0,866

Note that in practical terms, a rational number is not making sense in a discrete probability calculation (we can’t have half of a coin flip or 1.5 head as a result. This is a case of theoretical probability vs experimental probability).

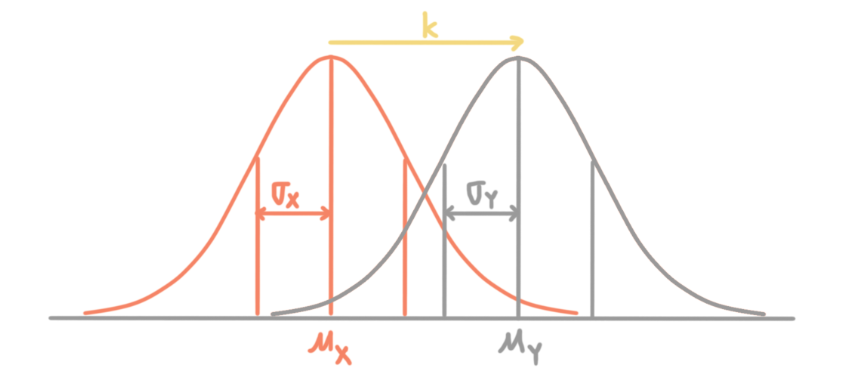
## Transforming Random Variables

This section studies how the distribution of a random variable changes when the variable is transformed in a deterministic way.

| **X** | **Y** |
| --- | --- |
| 3 | 3 + K |
| 3 | 3 + K |
| 7 | 7 + K |
| 10 | 10 + K |
| 12 | 12 + K |

| **Data Set** | **Shifted, K** |
| --- | --- |
| Mean: 6 | Mean: 6+K |
| Median: 7 | Median: 7+K |
| Mode: 3 | Mode: 3+K |
| Range: 10 | Range: 10 |
| IQR: 8 | IQR: 8 |
| St.Dev: | St.Dev: |

The meaning of this chart is to highlight the impact of a constant K **added (shifting)** to the whole dataset. While mean, median and mode will change of a K-addictive factor, range, IQR and standard deviation will stay the same since in fact the shape (and, more precisely, the distance between each datapoint) will not change.

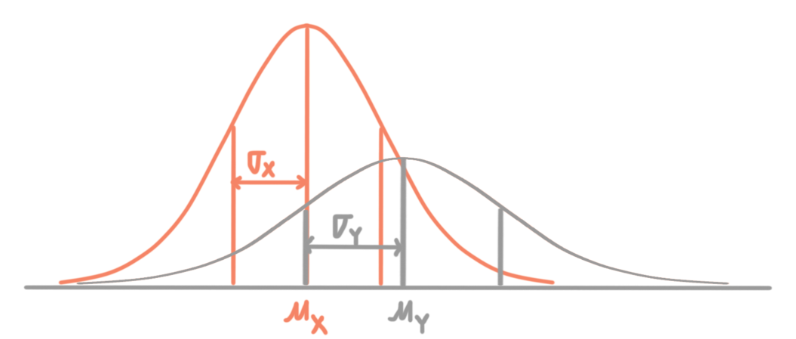


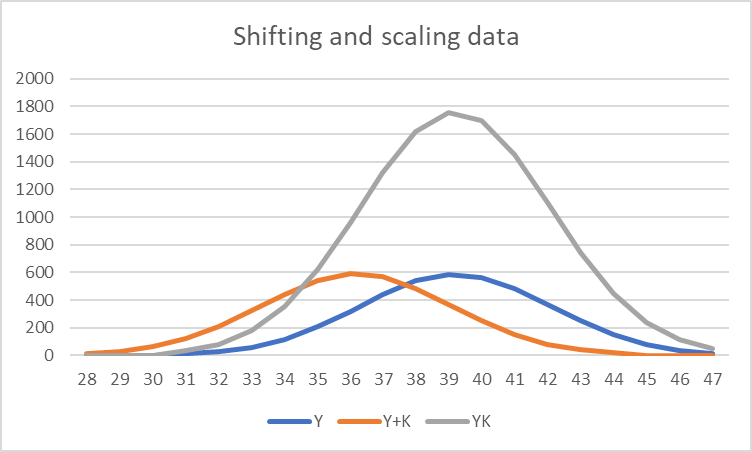
| **Data Set** | **Scaled, K** |
| --- | --- |
| Mean: 6 | Mean: 6K |
| Median: 7 | Median: 7K |
| Mode: 3 | Mode: 3K |
| Range: 10 | Range: 10K |
| IQR: 8 | IQR: 8K |
| St.Dev: | St.Dev: K |

The meaning of this chart is to highlight the impact of a constant K **multiplied (scaled)** to the whole dataset. In that case all the values are impacted by a K-magnitude factor.

Consider that both assumptions will be valid in a mixed case such as:

N(X) = 10x - 2 (\*x - 2 applies for mean, median and mode, while only \*x applies to st dev, IQR and range)





## Linear Combinations of Random Variables

Let 1 2 X X and be two independent random variables. Let a and b be scalars. Then a linear

combination of the variablesX1 X2 and is defined to be any other random variable of the form Y= aX1+ bX2.

Base assumptions:

* Variables must be independent.
* Variables have to have matching units.

|  | Sum S | Difference D |
| --- | --- | --- |
| Combination | S = X + Y | D = X - Y |
| Mean | µs = µx + µy | µd = µx - µy |
| Variance | ^2s = ^2x + ^2y | ^2d = ^2x + ^2y |

Example:

Time in hours 4 managers manage timesheet

X = [1, 2, 2, 3]

Time in hours 4 HRs manages payrolls on such timesheets

Y = [2, 3, 5, 6]

µX = (1+2+2+3) / 4 = 2

µY = (2+3+5+6) / 4 = 4

^2X = [(1 - 2)^2 + (2 - 2)^2 + (2 - 2)^2 + (3 - 2)^2] / 4 = 0.5

X = SQR(0.5) = 0.70

^2Y = [(2 - 4)^2 + (3 - 4)^2 + (5 - 4)^2 + (6 - 4)^2] / 4 = 2.5

Y = SQR(2.5) = 1.58

X + Y

µS = 2 + 4 = 6

^2S = 0.5 + 2.5 = 3

S = SQR(3) = 1.73

X - Y

µD = |2 - 4| = 2

^2D = 0.5 + 2.5 = 3

D = SQR(3) = 1.73

### Fair Game

A fair game (“gioco equo”) is a game (a bet, as an example, or a lottery) in which the cost of playing the game equals the expected winnings of the game, so that net value of the game equals zero.

W = B/p = 1

Where:  
W = win

B = Bet

p = probability

**Example: is the state lottery fair?**

The Italian national lottery (“Lotto[[2]](#footnote-1)”) is a bingo where 5 numbers are drawn out of a poll of 90. Repetition is not allowed (extracted numbers are discarded).

You win if you have all the 5 numbers on your card, in no specific order (there are also some minor prizes such as “ambo” for two numbers, “terno” for three numbers, and such. But the logic is the same).

A five is paying 6’000’000 times the bet.  
  
It this fair? Intuitively, it’s not.

But we are here not to assume but to investigate.

All the combinations of 5 numbers are C(n,k) hence:

90! / 5!(90-5)! = 43’949’268

This is obviously matching the official lottery statement.

(see Combinatorics chapter for details)

The probability of winning is then 1/43’949’268 = 0,00000228%

Fair game will assume:

W = B/p = 1

But we have:

W = 6’000’000

B = 1

p = 0,00000228%

W = B/p = 1/ 0,00000228% = 43’949’268

How unfair is that game?

43’949’268 / 6’000’000 = 7,32

The Italian lotto is more than 7 times unfair for the player.

# Joint Distribution

Joint distributions allow us to mathematically quantify the relationship between two distributions of data.

Given two random variables that are defined on the same probability space, the joint probability distribution is the corresponding probability distribution on all possible pairs of outputs.

The joint distribution can just as well be considered for any given number of random variables.

The joint distribution encodes the marginal distributions, i.e. the distributions of each of the individual random variables. It also encodes the conditional probability distributions, which deal with how the outputs of one random variable are distributed when given information on the outputs of the other random variable(s).

## Covariance

Covariance is a numerical value that provides a measure of how much two variables vary together.

It evaluates how the variables change together, providing the **direction** of the variation.

It’s a measure of the variance between two variables. However, the metric does not assess the dependency between variables.

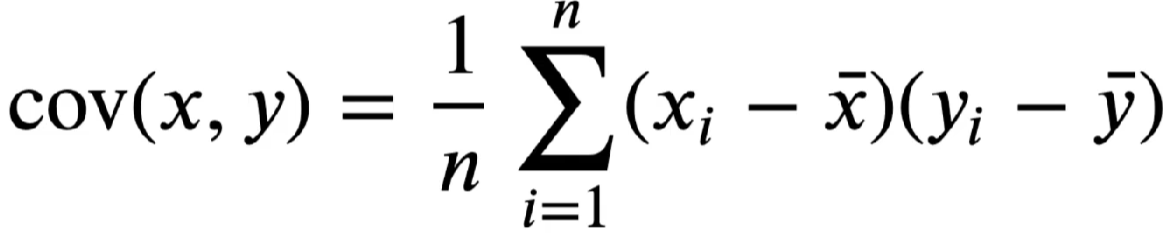
A **positive covariance** means that it’s expected both variables have a concordant behavior (X grows, Y grows, X decreases, Y decreases).

A **negative covariance** means that it’s expected both variables have a discordant behavior

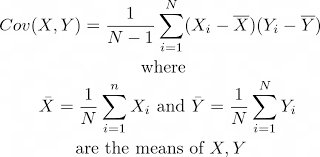
(X grows, Y decreases, X decreases, Y grows).

A **neutral covariance** means that the variables have no relations with each other.

**Population Covariance**:



**Sample Covariance**



## Correlation

Correlation is a metric used to measure the **strength** of a statistical relationship between two random variables. It’s also called a **measure of relationship**.

The correlation coefficient is a dimensionless metric and its value ranges from -1 to +1.

The closer it is to +1 or -1, the more closely the two variables are related.

If there is no relationship at all between two variables, then the correlation coefficient will certainly be 0. However, if it is 0 then we can only say that there is no linear relationship. There could exist other functional relationships between the variables.

+1: positive correlation (X grows, Y grows, X decreases, Y decreases).

-1: negative correlation (X grows, Y decreases, X decreases, Y grows).

While Covariance just measures the direction of variation between two variables, Correlation explores the strength and relation of the variation, in a standardized and comparable format.

In statistics application, there are three kind of correlation being applied:

* Pearson (Parametric method).
* Spearman (Nonparametric Method).
* Kendall (Nonparametric Method).

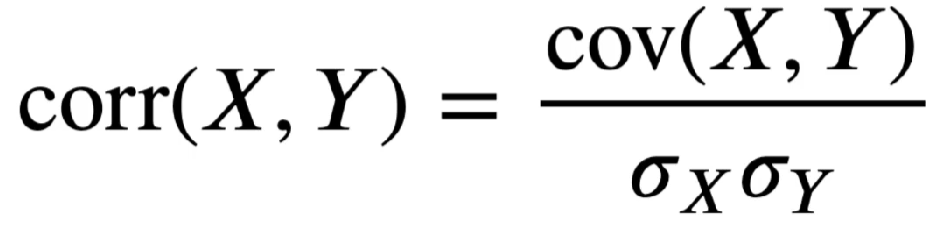
### Pearson correlation

Pearson r correlation is the most widely used correlation statistic to measure the degree of the relationship between linearly related variables. For the Pearson r correlation, both variables should be normally distributed (normally distributed variables have a bell-shaped curve).

Pearson correlation assumptions:

* Each observation should have a pair of values.
* Each variable should be continuous.
* It should be the absence of outliers.
* It assumes linearity and homoscedasticity (same finite and homogeneity of variance)..

Pearson correlation is expressed with the greek letter when referred to population or r when referred to a sample.



### For a sample

Pearson's correlation coefficient, when applied to a sample, is commonly represented by {\displaystyle r\_{xy}}r\_{xy} and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient. We can obtain a formula for {\displaystyle r\_{xy}}r\_{xy} by substituting estimates of the covariances and variances based on a sample into the formula above. Given paired data {\displaystyle \left\{(x\_{1},y\_{1}),\ldots ,(x\_{n},y\_{n})\right\}}{\displaystyle \left\{(x\_{1},y\_{1}),\ldots ,(x\_{n},y\_{n})\right\}} consisting of {\displaystyle n}n pairs, {\displaystyle r\_{xy}}r\_{xy} is defined as:

{\displaystyle r\_{xy}={\frac {\sum \_{i=1}^{n}(x\_{i}-{\bar {x}})(y\_{i}-{\bar {y}})}{{\sqrt {\sum \_{i=1}^{n}(x\_{i}-{\bar {x}})^{2}}}{\sqrt {\sum \_{i=1}^{n}(y\_{i}-{\bar {y}})^{2}}}}}}{\displaystyle r\_{xy}={\frac {\sum \_{i=1}^{n}(x\_{i}-{\bar {x}})(y\_{i}-{\bar {y}})}{{\sqrt {\sum \_{i=1}^{n}(x\_{i}-{\bar {x}})^{2}}}{\sqrt {\sum \_{i=1}^{n}(y\_{i}-{\bar {y}})^{2}}}}}}

where:

{\displaystyle n}n is sample size

{\displaystyle x\_{i},y\_{i}}x\_{i},y\_{i} are the individual sample points indexed with i

{\textstyle {\bar {x}}={\frac {1}{n}}\sum \_{i=1}^{n}x\_{i}}{\textstyle {\bar {x}}={\frac {1}{n}}\sum \_{i=1}^{n}x\_{i}} (the sample mean); and analogously for {\displaystyle {\bar {y}}}{\bar {y}}

Rearranging gives us this formula for {\displaystyle r\_{xy}}r\_{xy}:

{\displaystyle r\_{xy}={\frac {n\sum x\_{i}y\_{i}-\sum x\_{i}\sum y\_{i}}{{\sqrt {n\sum x\_{i}^{2}-\left(\sum x\_{i}\right)^{2}}}~{\sqrt {n\sum y\_{i}^{2}-\left(\sum y\_{i}\right)^{2}}}}}.}{\displaystyle r\_{xy}={\frac {n\sum x\_{i}y\_{i}-\sum x\_{i}\sum y\_{i}}{{\sqrt {n\sum x\_{i}^{2}-\left(\sum x\_{i}\right)^{2}}}~{\sqrt {n\sum y\_{i}^{2}-\left(\sum y\_{i}\right)^{2}}}}}.}

where {\displaystyle n,x\_{i},y\_{i}}n,x\_{i},y\_{i} are defined as above.

This formula suggests a convenient single-pass algorithm for calculating sample correlations, though depending on the numbers involved, it can sometimes be numerically unstable.

Rearranging again gives us this formula for {\displaystyle r\_{xy}}r\_{xy}:

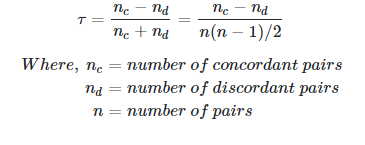
{\displaystyle r\_{xy}={\frac {\sum \_{i}x\_{i}y\_{i}-n{\bar {x}}{\bar {y}}}{{\sqrt {\sum \_{i}x\_{i}^{2}-n{\bar {x}}^{2}}}~{\sqrt {\sum \_{i}y\_{i}^{2}-n{\bar {y}}^{2}}}}}.}{\displaystyle r\_{xy}={\frac {\sum \_{i}x\_{i}y\_{i}-n{\bar {x}}{\bar {y}}}{{\sqrt {\sum \_{i}x\_{i}^{2}-n{\bar {x}}^{2}}}~{\sqrt {\sum \_{i}y\_{i}^{2}-n{\bar {y}}^{2}}}}}.}

where {\displaystyle n,x\_{i},y\_{i},{\bar {x}},{\bar {y}}}n,x\_{i},y\_{i},{\bar {x}},{\bar {y}} are defined as above.

### 

### Kendall rank correlation

Kendall rank correlation is a non-parametric test that measures the strength of dependence between two quantitative or qualitative ordinal statistical variables.   
  
Kendall rank correlation is expressed with greek letter tau ()



Kendall rank assumptions

* Pairs of observations are independent.
* Two variables should be measured on an ordinal, interval or ratio scale.
* It assumes that there is a monotonic relationship between the two variables.

### Spearman rank correlation

Spearman rank is a non parametric measure of rank correlation (measure of statistical dependence between the rankings of two variables). Spearman rank is also defined as the Pearson correlation between the rank variables.

Spearman rank is denoted with same greeks as Pearson, or r



Spearman rank assumption

* Pairs of observations are independent.
* Two variables should be measured on an ordinal, interval or ratio scale.
* It assumes that there is a monotonic relationship between the two variables.

### Point-biserial correlation coefficient

The Point-Biserial correlation coefficient is used when one of the given variable is dichotomous (such as “head or tail” on a coin flip).

Point-biserial correlation coefficient is expressed with rpb.

# Data Distribution

In statistics, and specifically to the field of descriptive statistics, a distribution is a representation of how different modes of a character are distributed across the statistical units that make up the collective under study.

## Probability Mass Function (PMF)

Probability mass function is a function that gives the probability that a discrete random variable is exactly equal to some value.

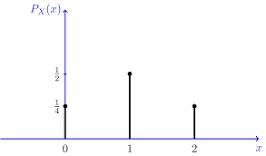
f(x) = P[X = x]

where X is the **discrete random variable** and x is the target value.

Example:

What is the probability of picking a specific ball out of a jar with 100 balls, all the balls being equal?   
It’s 1/100, or 1%.

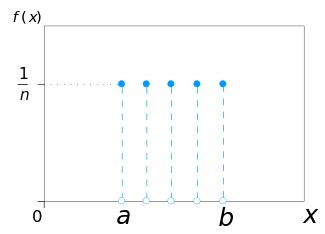
**PMF visualization**



### 

### Discrete uniform distribution

Discrete uniform distribution refers to discrete events where all the events have an equal chance of occurring (such as dice rolls).



## Probability Density Function (PDF)

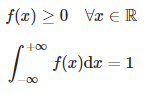
In probability theory, a probability density function (PDF), or density of a **continuous random variable**, is a function whose value at any given sample (or point) in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a relative likelihood that the value of the random variable would be close to that sample.

Density functions require switching from exact outcomes (such as for discrete variables) to approximations or interval ranges for an infinite set of values within the probability interval (in a continuous interval, the values of the variable are so small to be practically uncountable).



More technical: In a continuous distribution (e.g. continuous uniform, normal, and others), the probability is calculated by integration, as an area under the probability density function and with:

P(x = c) = 0



If a random variable X is given and its distribution admits a probability density function f, then the expected value of X (if the expected value exists) can be calculated as

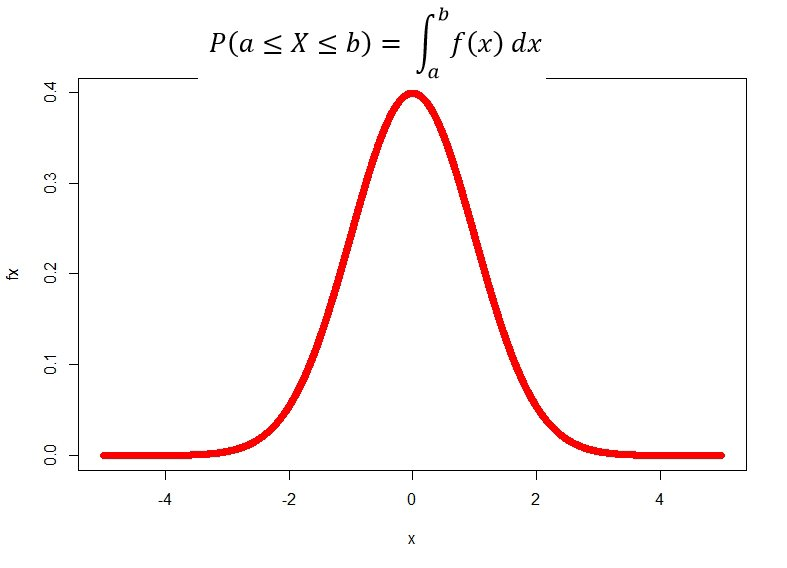


Not every probability distribution has a density function: the distributions of discrete random variables do not, for example.

Note also that within a PDF, the probability of having a specific, exact value (such as in PMF) is always equal to zero. The expectation is for an approximation (a < X < b) and not for an equality (X = a).

There are many kinds of probability density functions (actually more than 100, going from very common normal distribution, Pareto distribution, uniform distribution, to other less common distributions such as Polya-Gamma).

**PDF visualization**

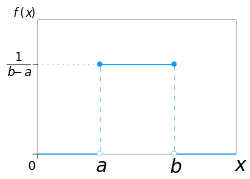


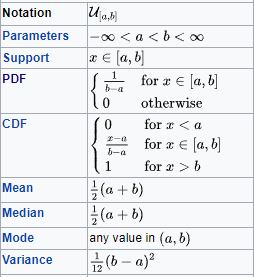
### Continuous uniform distribution

In probability theory and statistics, the continuous uniform distribution or rectangular distribution is a family of symmetric probability distributions. The distribution describes an experiment where there is an arbitrary outcome that lies between certain bounds.

The bounds are defined by the parameters, a and b, which are the minimum and maximum values. The interval can either be closed (e.g. [a, b]) or open (e.g. (a, b)).

Therefore, the distribution is often abbreviated U (a, b), where U stands for uniform distribution.





## Cumulative Distribution Functions (CDF)

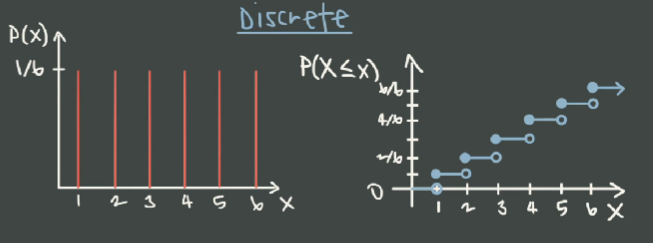
The cumulative distribution function (CDF) is the probability that the variable takes a value less than or equal to x.

F(x) = P(X <= x)

CDF expresses the cumulative probability of a given event.

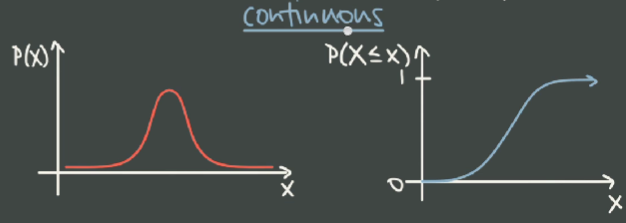
## Discrete CDF

Example: rolling a dice, the CDF of having no numeric output is 0, while the CDF of having a number between 1 and 6 is 1 (6/6), the CDF of having a number less or equal to 3 is 3/6 (1/ 6+ 2/6 + 3/6).



## Continuous CDF

The cumulative distribution function, CDF, or cumulant is a function derived from the probability density function for a continuous random variable. It gives the probability of finding the random variable at a value less than or equal to a given cutoff. Many questions and computations about probability distribution functions are convenient to rephrase or perform in terms of CDFs, e.g. computing the PDF of a function of a random variable.

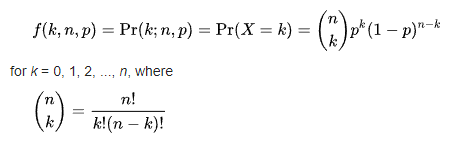


## Binomial Distribution

Binomial distribution expresses the discrete probability distribution of an experiment that is repeated multiple times, having only two possible outcome: positive or negative.

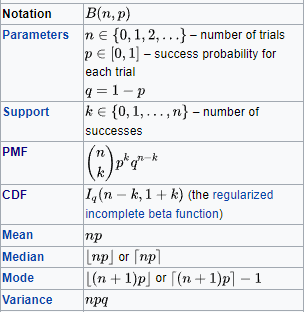
Binomial refers to the distribution having only two possible outcomes, positive or negative.

Binomial distribution is expressed as B(n,p), with n being trials and p being the probability of success.



Binomial distribution should respect the following criteria:

* Outcome is binomial (positive or negative, 1 or 0, + or - etc.).
* Each event is independent.
* The number of trials n is fixed.
* Success\failure rate p is constant.



Example of a binomial distribution in the real world:

According to a report, 80% of prospects at your company will result in a signed contract. Each prospect is independent of each other.  
Your boss asks you to calculate the probability to close a deal from a round of 3 prospects you are working on.

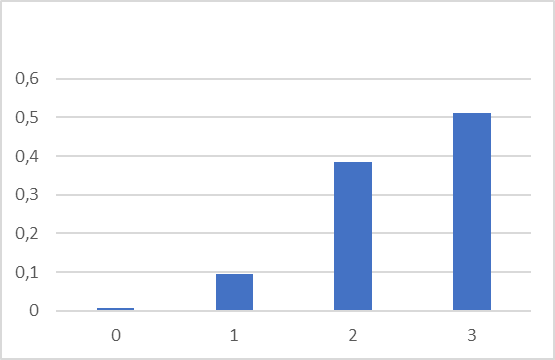
Hence:

* n = Number of trials = 3
* Number of outcomes = binary positive\negative (contract or no contract)
* p = Probability of success = 0.8
* Trials are independent = yes
* k = Target result = 1 contract closed = 1

(n, k) = (3,1) = 3! / (1! \* (3 -1)!) = 6 / 2 = 3

and then B = 3 \* 0.8^1(1-0.8)^(3-1) = 0,096

Calculating then the values for each value of k , hence P(n,k) = P(3,0), P(3,1), P(3,2), P(3,3) we should be able to plot a chart of the distribution.



## Bernoulli Distribution

The Bernoulli distribution is the discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability q = 1 - p.

Less formally, it can be thought of as a model for the set of possible outcomes of any single experiment that asks a yes–no question.

Such questions lead to outcomes that are boolean-valued: a single bit whose value is success/yes/true/one with probability p and failure/no/false/zero with probability q.

Bernoulli distribution can be see as a specific case of binomial distribution, where:

* Binomial distribution: n trials
* Bernoulli distribution: one trial.

It can be used to represent a (possibly biased) coin toss where 1 and 0 would represent "heads" and "tails", respectively, and p would be the probability of the coin landing on heads (or vice versa where 1 would represent tails and p would be the probability of tails). In particular, unfair coins would have



## Poisson Distribution

Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.

A discrete random variable X is said to have a Poisson distribution, with parameter >0, if it has a probability mass function given by:



where

* k is the number of occurrences
* e is Euler's number
* ! is the factorial function.

Poisson distribution should respect the following criteria:

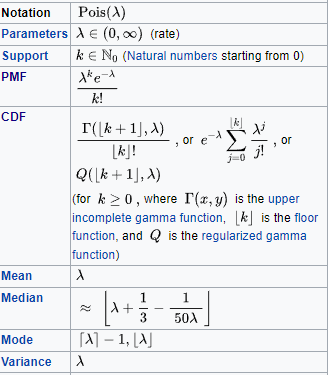
* The mean number of events occurring within a given interval of time or space lambda ( ) is known and assumed to be constant.
* Occurrences occur in an interval and are discrete and countables.
* Events occur independently one to another.
* The average rate at which events occur is independent of any occurrences (assumed to be constant).
* Two events cannot occur exactly at the same instant. At each atomic interval, either or one event occurs or no event occurs (also: intervals are no overlapping).
* Probability is proportional to interval size.

Example: the number of chewing gum on a single tile over a sidewalk, or the number of planes that fly over your house in an hour.

One of the first applications of Poisson distribution was the investigation of deaths by horse kick in soldiers in 1800. Researchers were interested in inferring the deaths by horse kick in a year.

* Event = death by kick
* Time interval = 1 year
* Lambda = 0.61
* The number of times an event is investigated is K and is assumed to be = 2 (we want to calculate the probability that in a year 2 soldiers will die from a horse kick).

P(X = k where k = 2) = e^(- ) ^K / k! = 0.101



# Normal Distribution

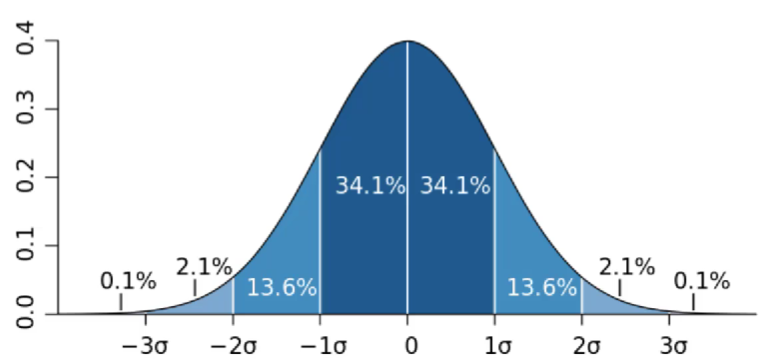
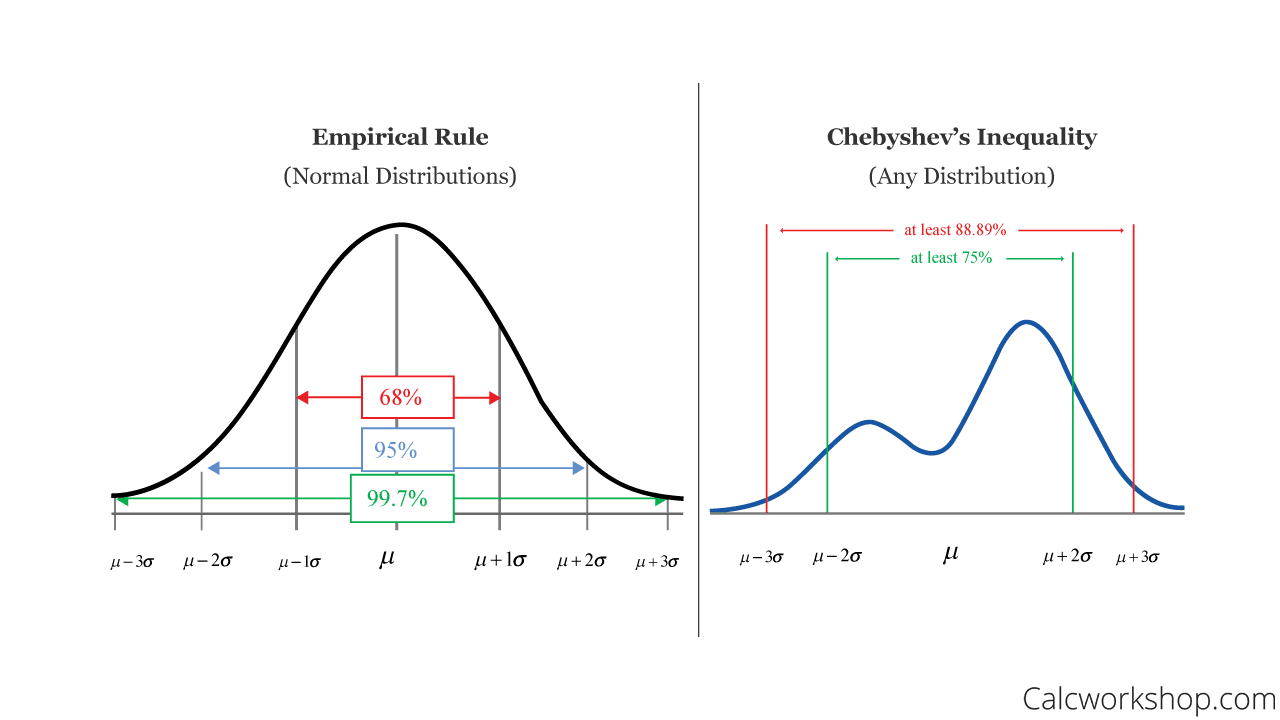
In statistics, a normal distribution or Gaussian distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is:



A random variable with a Gaussian distribution is said to be normally distributed and is called a normal deviate.

Normal distribution is one of the most common distributions used in business, statistics, biology, etc., since so many real-life datasets end up resembling a normal distribution (Central Limit Theorem).

Normal distributions are referred to with capital letter N, such as N(5,9) means a normal distribution with mean 5 and variance 9.

Normal distributions have unique properties of mean and standard deviation.  
  


## Z-Score

Z-score (even called standard score) is the number of standard deviations by which the value of an observed value or data point is above or below the mean of the distribution.

Negative z-scores fall on the left of the distribution (-4 being the further) while positive z-scores fall on the right of the distribution (+4 being the further).

z = (x – μ) / σ

## 

### Z-Tables

A z-table, or standard normal table, reveals what percentage of values fall below a certain z-score in a normal distribution. It allows to translate specifics z-scores to their statistic relevance in a normal distribution.  
  
The table has z decimal values on the row header, and decimal places on the table header.  
  
So, for example, -3.4 has to be found at “Col:1, Row:1” of the table; -3.9 has to be found at “Col:10, Row:1” of the table.  
  
To have a visual explanation of this process, check: <https://www.z-table.com/>

## Normality test

In theory, a normal distribution follows precisely a Gaussian curve. But real world data is rarely so precise and could be the case we may be willing to test if the distribution we are working with is effectively normal.

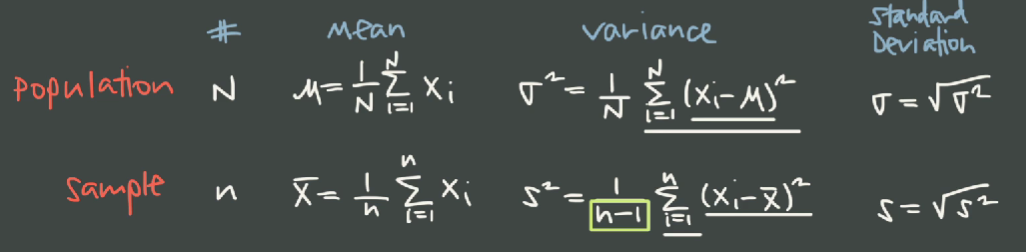
Normality tests are used to determine if a data set is well-modeled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed.

Normality tests return a probability of normality. More specifically, these tests operate using a hypothesis paradigm, where you posit an hypothesis that your particular data sample is normally distributed.

Normality tests are such as:

* D'Agostino's K-squared test,
* Jarque–Bera test,
* Anderson–Darling test,
* Kolmogorov–Smirnov test
* Shapiro–Wilk test

## Mean, Variance and Standard Deviation for normal distribution



Remember that higher values of variance and standard dev mean a flattened, wide bell curve. Smaller values of variance and standard dev mean a peaked, more compressed bell curve.

## Skewed Distribution (Skewness)

Skewness is a measure of the asymmetry of the probability distribution of a random variable in respect to the mean of the distribution.

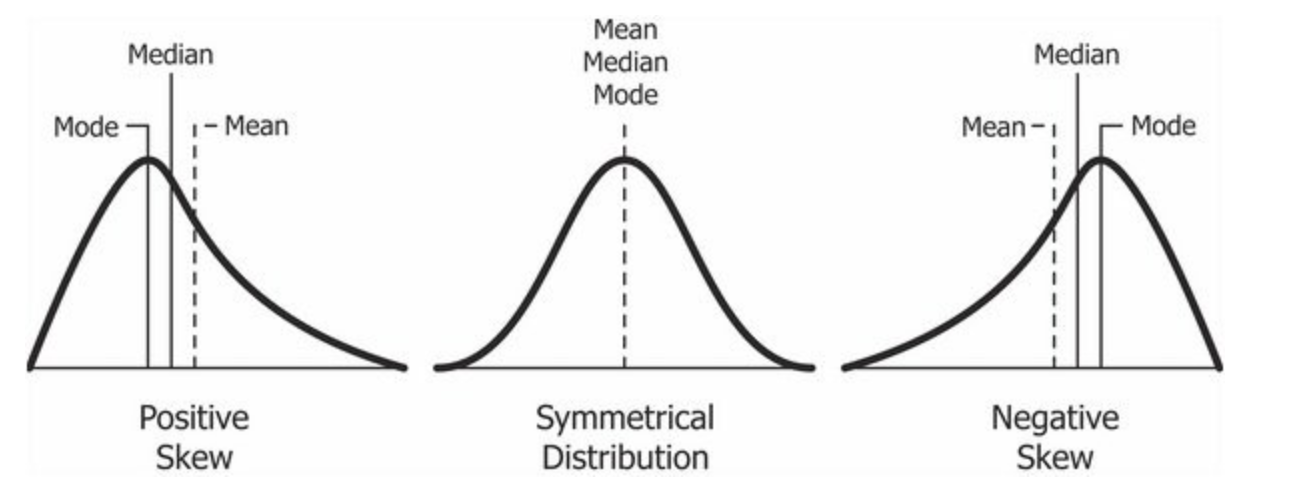
Would be useful to remember that for a symmetric distribution, and, specifically, for a normal distribution, median = mean = mode.

**Negative Skew**

The left tail is longer; the mass of the distribution is concentrated on the right of the figure. The distribution is said to be left-skewed, left-tailed, or skewed to the left, despite the fact that the curve itself appears to be skewed or leaning to the right; left instead refers to the left tail being drawn out and, often, the mean being skewed to the left of a typical center of the data. A left-skewed distribution usually appears as a right-leaning curve.

**Positive Skew**

The right tail is longer; the mass of the distribution is concentrated on the left of the figure. The distribution is said to be right-skewed, right-tailed, or skewed to the right, despite the fact that the curve itself appears to be skewed or leaning to the left; right instead refers to the right tail being drawn out and, often, the mean being skewed to the right of a typical center of the data. A right-skewed distribution usually appears as a left-leaning curve

[[3]](#footnote-2)

## Kurtosis

While skewness refers to the tendency of a distribution to propagate in respect to its mean, kurtosis measures the sharpness of the curve in respect of its distribution.

**Skewness:** lack of symmetry in the distribution

**Kurtosis:** height and sharpness of the central peak

## Standard Normal Distribution

The standard normal distribution, also known as the z-distribution (Z), is a particular case of normal distribution. Standard Normal Distribution has a mean = 0 and the standard deviation = 1.

Z = N(0,1)

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.

# Sampling

Sampling allows us to test a hypothesis about general characteristics of a population.

Samples are used to make inferences about a population.

We can think of sampling as grabbing data instances from a larger data distribution.

## Sampling Methodologies

A good sample is:

* Representative of the population.
* Unbiased.

### Simple Random Sample (SRS)

A simple random sample (or SRS) is a subset of individuals (a sample) chosen from a larger set (a population) in which a subset of individuals are chosen randomly, all with the same probability. It is a process of selecting a sample in a random way.

### Systematic Random Sample

Systematic random sample is a subclass of SRS. Every member of the population is labeled with a number, and the sample is picked using a fixed interval (p.e. one out of every 10 members).

### Stratified Random Sample

Stratified random sample is a subclass of SRS, where the population can be further classified into subpopulations. Subpopulations can’t overlap and they should be proportional in order for the sample to be relevant.

### Clustered Random Sample

Clustered random sample is a sampling methodology that can be used whenever the population can be aggregated into mutually homogeneous yet internally heterogeneous groupings.

After the population is divided into clusters, then a SRS is performed on each cluster.

It’s a sampling method often used in marketing research.

## Central Limit Theorem

In probability theory, the central limit theorem (CLT) establishes that, in many situations, when independent random variables are summed up, their properly normalized sum tends toward a normal distribution around the mean of the original dataset even if the original variables themselves are not normally distributed.

The CLT establishes a relationship between the data from the whole population (parameters) and the data from the sample (statistics).

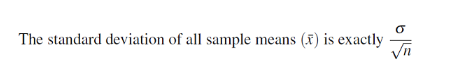
### Sampling Distribution of the Sample Mean (SDSM)

The mean of all repeated samples (where all = all the combinations from the sample) from a population of value for a quantitative variable is equal to the population mean.

xbar = mu

And the standard deviation of the sample is equal to:

sigma / squareroot(n)



The standard deviation of the sample is also referred to as “Standard Error (SE)”.

### More on mean of the sampling being equal to the mean of the population

This statement can lead to wrong assumptions - since the mean of the sample typically is not equal to the mean of the population *unless* samples are taken listing all the possible samples over the population (that being, combinations).

The idea is that when you think of taking a sample, there are a large number of possible samples, from which you will choose just one. However, you need to imagine having all the samples available, and knowing the mean of each one. This list of all the possible sample means makes up the distribution of the sampling means (aka “the sampling distribution of the means”). Now, this distribution itself has a mean, which is how we get the somewhat confusing phrase “mean of … means.”[[4]](#footnote-3)

So, while the mean of one sample is an estimate of the population mean and rarely equal to it, the mean of (all the means of all the possible samples), is exactly equal to the population mean. This is the basis of an important result in statistical theory, that the sample mean is an unbiased estimator of the population mean.

### Finite Population Correction Factor (FPC)

FPC is used to adjust sampling bias whenever sampling is done without replacement and over more than the 5% of a finite population (both frequent real case scenarios).  
  
Example: you have to apply FPC if picking 600 people (>5%) from a city telephone address book of 10’000 members (population is finite and whenever a person is picked from the list, can’t be picked again).

FPC = √(N-n) / (N-1)

To apply a finite population correction, simply multiply it by the standard error that you would have originally used.

For example, the standard error of a mean is calculated as:

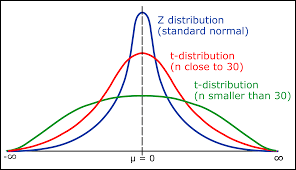
Standard error of mean: s / √n

By applying the finite population correction, the formula becomes:

Standard error of mean: s / √n \* √(N-n) / (N-1)

## The Student's T-Distribution

Student's t-distribution a family of continuous probability distributions that arise when estimating the mean of a normally distributed population in situations where the sample size is small and the population's standard deviation is unknown.



t = (x̄ – μ) / (s/√n)

Where,

x̄ is the sample mean

μ is the population mean

s is the standard deviation

n is the size of the given sample

### Degrees of freedom (DF)

The number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary (independent, known, available data).

Degrees of freedom (DF) is equal to the size of the sample n minus 1.

DF = n - 1

The explanation for this lies in the fact that the punctual features of a dataset can be computed backward knowing the global features of the dataset.

For example, having a dataset of 15 elements, and knowing the mean, we are free to delete one of these values, since knowing the mean we are able to calculate it backwards.

But what happens if, for example, we delete two values? We are able to trace the global value of the two variables but not to assign their value in a timely and independent way.

That is why DF is calculated as n - 1.

### T-Score

The T-distribution (and those associated T-score values) is used in hypothesis testing

when determining if one should reject or accept the null hypothesis.

Values of T-score have to be compared to the T-table with a process similar to the Z-scores

You can find a practical T-table here: <https://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf>

### Why “Student Distribution”?

The t-distribution was developed by English statistician William Sealy Gosset.

At the time (1912) he published more than 20 academic papers, mostly using the pseudonym “Student”.

The t-distribution was originally called “Test of statistical significance” or “Student’s z” (for its similarity to Z distribution).

In the end he could have called it “Gosset distribution”, but he passed to history as “Student”.

### When to use Z or T distribution?

If sample < 30, t-distribution will provide a more accurate value.  
  
If sample > 30, z-distribution will provide a more accurate one.

## Confidence interval for the mean

Estimates about the population can be made sampling from the entire population dataset.  
  
If the estimate is made from a single value, it is called a “point estimate”.  
  
If the estimate is made from a range of values where the parameter is expected to lie, it’s called an “interval estimate”.

### Point Estimation

Point estimation involves the use of sample data to calculate a single value (called point estimate since it identifies a point in some parameter space) which is to serve as a "best guess" or "best estimate" of an unknown population parameter (for example, the population mean).

More informally, point estimation is the process of finding an appropriate value of a population parameter, such as the mean of the population, from random samples of the population.

Hence, the accuracy of a particular approximation is not known precisely.

### Interval estimation

Interval estimation involves the evaluation of a parameter of a population (for example, the population mean) by computing an interval, or range of values, within which the parameter is most likely to be located.

Intervals are commonly chosen within **confidence intervals (CI)**, such that the parameter falls within a **confidence level (CL)** of 95% or 99% (confidence coefficient). For example, out of all intervals computed at the 95% level, 95% of them should contain the parameter's true value.

The end points of such an interval are called upper and lower confidence limits.

### Confidence interval (CI)

A confidence interval refers to the probability that a population parameter will fall between a set of values for a certain probability. That probability is known as confidence level.

Thus, if a point estimate is generated from a statistical model of 10.00 with a 95% confidence interval of 9.50 - 10.50, it can be inferred that there is a 95% probability that the true value falls within that range.

The percentage falling outside the confidence level is called alpha value () or level of significance (LOS). So the alpha value associated with a 90% confidence level is 10%.

Mathematically speaking, the confidence interval can be defined as  
  
CI = x bar (sample mean) +/- margin of error.

where

Margin of error = z \* SE

where  
SE = standard error = population standard deviation / squareroot(n)

If standard deviation is not known, then we need to use a t-score instead of a t-score.

### Confidence Level (CL)

The confidence level measures the level of trust[[5]](#footnote-4) of the accuracy of the provided interval.

Higher confidence levels means a wider range of values.  
Lower confidence levels means a tighter range of values.

### Confidence interval and Z-scores

Since we already know that the most common confidence intervals are 90%, 95% and 99%, we can already get confidence with the Z-values associated with such levels.

90% CL = Z +/- 1.65

95% CL = Z +/- 1.96

99% CL = Z +/- 2.58

# Hypothesis Testing

Hypothesis testing is a method used in statistical inference used to validate specific assumptions made over a population starting from a data sample.

The process of testing an hypothesis is the following:

* State null and alternative hypotheses.
* Determine the levels of significance.
* Calculate the statistics.
* Find critical values.
* Determine regions of acceptance and rejections.
* State the conclusion.

## Null & Alternative Hypothesis

**Null hypothesis (H0)** states that no difference or relationship exists between two sets of data or variables being analyzed.

**Alternative hypothesis (H1, Ha)** states that there exists a relationship between two sets of data or variables being analyzed.

The two hypotheses are tested together.

## Type I and Type II Errors

**Type I error** is a false positive conclusion (rejecting an actually true null hypothesis).

Type I error is equal to the alpha probability (the percentage falling outside the level of significance). These errors mean that the results are assumed to be statistically significant while they are actually not.

**Type II error** is a false negative conclusion (accepting an actually false null hypothesis). This has not to be mixed with accepting the null hypothesis. It may happen whenever the analysis has not enough statistical power to detect an effect of such a size.

Statistical power is determined by:

* Size of the effect.
* Measurement errors.
* Sample size.
* Significance level.

| **Null hypothesis is:** | **True** | **False** |
| --- | --- | --- |
| **Rejected** | Type I Error False positive  Probability = LOS = α | Correct decision  True positive Probability = 1-β |
| **Accepted** | Correct decision  True negative  Probability = 1 - α | Type II error  False negative  Probability = β |

### Trade-off between Type I and Type II errors

Type I and Type II errors influence each other.

* Setting a LOS decreases Type I but increases incidence of Type II errors.
* Increasing power of a test decrease Type II but increases incidence of a Type I error.

**What’s worse?**  
  
The example of the innocent convicted to jail is often used to explain why - usually - Type I errors (rejecting H0 when it’s in fact true) may lead to worse consequences than a Type II error.

In more general terms, we may say that Type I errors may lead to a change in a given process, hence, to an active, wrong, action. While Type II may lead to overlooking an action that would have been positive. A passive point of view.  
  
But, obviously, it depends on the context and on the scope of the analysis.

Sources: Neyman, J.; Pearson, E.S. (1967) [1933]. “The testing of statistical hypotheses in relation to probabilities a priori”. Joint Statistical Papers. Cambridge University Press. pp. 186–202.

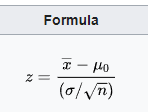
## Test Statistics

A test statistic is a statistic measurement (hence, a quantity derived from a sample taken from a population), specifically used in hypothesis testing.

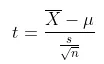
Test statistics can be done via different methods such as t-value, z-value, F-value, x^2-value.

Here we focus on z-test and t-test. The choice criteria is the same as for z-scores and t-scores - t-scores being used where the sample size is small and the population's standard deviation is unknown.

| **Formula** | **Assumptions or notes** |
| --- | --- |
| One-sample [z-test](https://en.wikipedia.org/wiki/Z-test) | {\displaystyle z={\frac {{\overline {x}}-\mu \_{0}}{({\sigma }/{\sqrt {n}})}}} |



One-sample t-test



### Two-Tailed and One-Tailed Tests

**Two-tailed test** is used whenever the estimated value can be greater or smaller than a certain range of values (for example, a target score from an exercise).

If the estimated value exists in the critical areas, the alternative hypothesis is accepted over the null hypothesis.

Noting that in a two-tailed test the range of rejection is equal to α/2 (both sides of the distribution, right and left), we need more extreme values in order to incur in a rejection of the null hypothesis. Hence, two-tailed tests are more conservative, and we should assume to have strong evidence to state a specific direction in the hypothesis testing.

**One-tailed test** is used whenever the estimated value can differ from the reference value only in one direction, greater or smaller. but not both. For example, the defective rate of a machine.

If the estimated value exists in one of the one-sided critical areas, depending on the direction of interest, the alternative hypothesis is accepted over the null hypothesis.

Range of rejection = α (one side of the distribution, right or left).

## P-Value and Critical Value

p-value and critical value are meant to do the same thing: support or reject the null hypothesis.

The test process will follow this roadmap:

1. State Null and Alternate Hypothesis.
2. Choose LOS (level of significance, α).
3. Calculate test statistics from the sample

then, for **P-Value** approach:

* Compute p-value
* Compare p-value to alpha level
* Reject null if p-value <= alpha level

if **Critical Value** approach:

* Find critical value.
* Compare test statistics to critical value.
* Reject null if the test value falls in the rejection region (|t| < critical value)

**P-Value[[6]](#footnote-5) (compare areas):**

For the p-value approach, the likelihood (p-value) of the numerical value of the test statistic is compared to the specified significance level (α) of the hypothesis test.

The p-value corresponds to the probability of observing sample data at least as extreme as the actually obtained test statistic. Small p-values provide evidence against the null hypothesis. The smaller (closer to 0) the p-value, the stronger is the evidence against the null hypothesis.

**Critical Value (compare scores):**

It is determined whether or not the observed test statistic is more extreme than a defined critical value. Therefore the observed test statistic (calculated on the basis of sample data) is compared to the critical value, some kind of cutoff value.

* If the test statistic is more extreme than the critical value, the null hypothesis is rejected.
* If the test statistic is not as extreme as the critical value, the null hypothesis is not rejected.

## A/B Testing

A/B Testing is a way to test two and two only independent variables (two-sample hypothesis testing, split testing or bucket testing).

# Regression Analysis

Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable, or a 'label' in machine learning parlance) and one or more independent variables (often called 'predictors', 'covariates', 'explanatory variables' or 'features').

Regression analysis is primarily used for two conceptually distinct purposes.

First, regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning.

Second, in some situations regression analysis can be used to infer causal relationships between the independent and dependent variables. Importantly, regressions by themselves only reveal relationships between a dependent variable and a collection of independent variables in a fixed dataset.

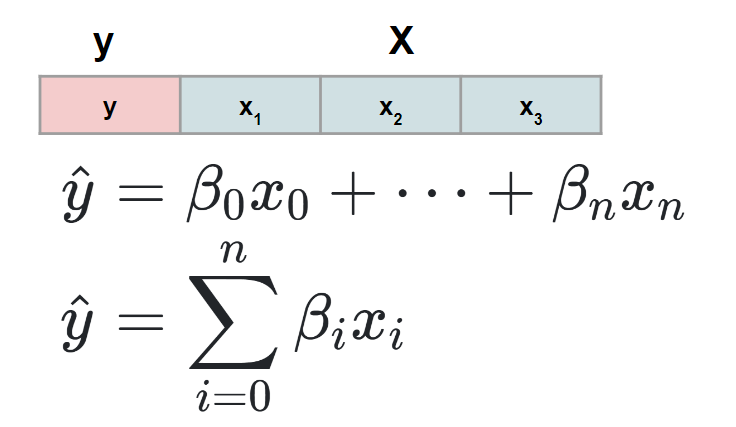
Linear regression line is on the form of  
  
y = mx + b

But how can we solve that in a context where there are many possible features of x, such as in a prediction under uncertainty? OLS is one of the possible approaches.

## Ordinary Least Squares (OLS)[[7]](#footnote-6)

Ordinary least squares (OLS) is a t method for choosing the unknown parameters in a linear regression model by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear regression function.

OLS allows us to directly solve the y = mx + b equation for the slope m and the intercept b.  
  
Thus our linear regression function will translate to:



We need then to solve for the beta coefficients, usually done with computational techniques using stochastic gradient descent.

## Correlation Coefficient

The sample correlation coefficient (r) is a measure of the closeness of association of the points in a scatter plot to a linear regression line based on those points, as in the example above for accumulated saving over time. Possible values of the correlation coefficient range from -1 to +1, with -1 indicating a perfectly linear negative, i.e., inverse, correlation (sloping downward) and +1 indicating a perfectly linear positive correlation (sloping upward).

Correlation strength

| **Correlation Coefficient (r)** | **Description**  **(Rough Guideline )** |
| --- | --- |
| +1.0 | Perfect positive + association |
| +0.8 to 1.0 | Very strong + association |
| +0.6 to 0.8 | Strong + association |
| +0.4 to 0.6 | Moderate + association |
| +0.2 to 0.4 | Weak + association |
| 0.0 to +0.2 | Very weak + or no association |
| 0.0 to -0.2 | Very weak - or no association |
| -0.2 to – 0.4 | Weak - association |
| -0.4 to -0.6 | Moderate - association |
| -0.6 to -0.8 | Strong - association |
| -0.8 to -1.0 | Very strong - association |
| -1.0 | Perfect negative association |

## Line Fitting and Residuals

**Line fitting** is the process of constructing a straight line that has the best fit to a series of data points.  
  
The definition of best fit in itself varies depending on the methodology used, for examples:

* Linear regression minimizes the vertical distance.
* Orthogonal regression minimizes the perpendicular distance.

**Residuals** are the differences between each data point and the estimated value, given by the regression line equation.  
  
Residuals have an overall sum and mean of zero.

## Linear Regression Trendlines

There are four different ways in which you can describe a statistics trend, analyzing the trendlines (linear regression lines).

**Form:**

The shape that the trend is following.

It could be:

* Linear: a straight line.
* Exponential: a parabolic line.
* Sinusoidal: a upward-downward horizontal S curve shaped line.
* Logarithmic: following a logarithmic distribution.
* No correlation: for scattered data with no evident trend and correlation.

**Direction:**

The orientation of the trend.  
It could be:

* Positive: the trendline is pointing upward.
* Negative: the trendline is pointing downward.

**Strength:**

How tightly clustered (distance of data points from the predicted value of the regression equations, or residual) the data points are from the trendline.

It could be:

* Strong: tightly clustered, minimal distance.
* Weak: sparse, mid-high distance from the trendline.

**Outliers:**

How many and how far far are the farther points from the regression line.

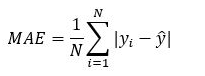
## Regression model evaluation metrics

Regression analysis is just the first part of the evaluation process during a statistical analysis. Evaluating the model accuracy is a paramount step in order to appreciate the accuracy of our analysis, and how representative it is of the dataset examined.

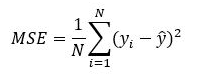
The main metrics used to evaluate accuracy of a regression analysis are the MSE, MAE, RMSE, and R-Squared metric.

**MEAN ABSOLUTE ERROR:** the MAE represents the **average of the absolute distances** between the initial dataset and the prediction.

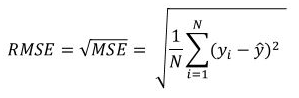
MAE is a rather simple metrics that returns a value in the same unit as the output variable. It is more robust to outliers than MSE (less impacted by value that are very far from the mean).



**MEAN SQUARED ERROR:** the MSE represent the **average of the squared distances** between the initial dataset and the prediction.

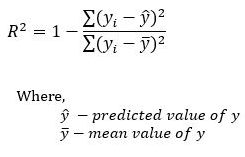


**ROOT MEAN SQUARED ERROR:** the RMSE is the square root of the MSE.



**R-SQUARED:** R-Squared, also known as **coefficient of determination**, is a ratio, the value of which ranges between 0 and 1, and which represents the accuracy of the prediction model with respect to the original values. is obtained as a ratio of the distance of the original points from the prediction values minus the distance of the original points from the dataset mean.

Coefficient of determination is a scale-free score, this mean that the value will always be within a range of 1 and 0.



## Chi-Square Test

Chi-square is among the most common [nonparametric](#_u5uyz0wilnt3) tests used in statistics.

Chi-square is one of the hypothesis testing tests used in statistics to decide whether or not to reject the null hypothesis.

Depending on the starting assumptions used such tests are considered parametric or nonparametric.

The chi-square test is widely used to test that the frequencies of observed values fit the theoretical frequencies of a fixed probability distribution. For example, it is well known that the result of 100 tosses of a coin follows the uniform distribution, and it is difficult to obtain a result that differs significantly from obtaining 50 heads and 50 tails. The chi-square test makes it possible to determine, after fixing the maximum permissible error, whether the discrepancies between the observed and theoretical frequencies can be attributed entirely to chance or whether it is safe to assume that the coin is being cheated.

\begin{equation*} X^2=\sum{\frac{(O-E)^2}{E}} \end{equation*}

* Where:
* Χ2 is the chi-square test statistic
* Σ is the summation operator (it means “take the sum of”)
* O is the observed frequency
* E is the expected frequency

Results of the chi-square equation have to be looked up on a chi-square table, such as [here](https://cdn.scribbr.com/wp-content/uploads/2022/05/Chi-square-table.pdf) in order to get the test results. Chi-square test > upper tail probability as shown in the table = H0 (null hypothesis) can be rejected.

## Analysis of Variance (ANOVA)

ANOVA is a set of statistical techniques that allow to compare mean variance between groups.

ANOVA can be calculated as follow:

* calculate the SST (SUM SQUARES TOTAL) (variance of each datapoint from the Gran Mean, being Grand Mean the mean of the whole dataset).
* calculate the SSW (SUM SQUARES WITHIN) (variance of each datapoint from the mean of each group).
* calculate the SSB (SUM SQUARES BETWEEN) (variance of the group mean from the Grand Mean)

At this point we can compute the F-test. F test is a ratio of two variances.

F test formula is:

F = (SSB / m-1) / (SSW / m(n-1)

The F test value has then to be compared on the F table, keeping in mind that there’s an F table for each value of alpha (for each level of confidence), and that we need both DF to cross the exact value (m-1 and m(n-1)).

# License

[**Attribution 4.0 International (CC BY 4.0)**](https://creativecommons.org/licenses/by/4.0/)

This document is distributed under a Creative Common, Free Culture, License

This work was created as a means of learning and diffusion for study purposes, so I hope for its free dissemination and dissemination.

While producing it, I mainly observed a two-pronged approach:

- Maintaining the accuracy of definitions, formulas, and descriptions.

- Describe everything with proprietary wording that does not infringe upon the intellectual property of the sources I have drawn on.

For obvious reasons however, mathematical definitions are free up to a point, so where in good faith I have traced the work of licensed material I am happy to take action to remove it.

You can contact me at the email \_\_\_\_\_\_\_\_\_\_\_\_\_

Still moving from the initial desire to disseminate knowledge in a streamlined and unconstrained manner.

This work was originally released by the author as a freely downloadable pdf.

**Machine-readable license metadata:**  
*<a rel="license" href="http://creativecommons.org/licenses/by/4.0/"><img alt="Licenza Creative Commons" style="border-width:0" src="https://i.creativecommons.org/l/by/4.0/88x31.png" /></a><br /><span xmlns:dct="http://purl.org/dc/terms/" href="http://purl.org/dc/dcmitype/Text" property="dct:title" rel="dct:type">The Statistics Handbook</span> di<span xmlns:cc="http://creativecommons.org/ns#" property="cc:attributionName"> Carlo Occhiena</span> è distribuito con Licenza <a rel="license" href="http://creativecommons.org/licenses/by/4.0/">Creative Commons Attribuzione 4.0 Internazionale</a>.*

# Source Code

The source of this paper can be found on an open repository on my GitHub, at the following link:

# 

# Bibliography & Sources

This workbook was only made possible through research, consultations and studies of numerous sources.

Some of them were institutional, such as University sources.

Others were related to the dissemination of collective knowledge, such as Wikipedia, Britannica, Khan, Youmath, Statology.

Some others were related to the work of specific science spreaders such as P. Pozzolo, K.King.

I have not neglected to cite even minor sources that have been helpful in finding meaningful exercises or comparing the goodness of the solutions I have devised. For example, 20-year-old forum threads.

The cue is to keep exploring and DYOR (do your own research).

Happy reading!

Links checked in December 2022.

* Gambini A., Argomenti di statistica descrittiva
* Perisco L, DI Bella E., Mosto L., Applicazioni di probabilità e statistica
* <https://it.wikipedia.org/wiki/Statistica>
* <https://math.stackexchange.com/>
* <https://corporatefinanceinstitute.com/>
* <https://statology.org>
* <https://www.britannica.com/browse/Mathematics>
* <https://www.statisticshowto.com/>
* <https://www.kristakingmath.com/>
* <https://www.khanacademy.org/>
* <https://www.youmath.it/>
* <https://ocw.mit.edu/courses/1-151-probability-and-statistics-in-engineering-spring-2005/pages/lecture-notes/>
* <https://corporatefinanceinstitute.com/>
* <https://paolapozzolo.it>
* <https://www.investopedia.com/math-and-statistics-4689831>
* <https://thestatsgeek.com/>
* <https://www.hwupgrade.it/forum/archive/index.php/t-1294440.html>
* [https://www.matematicamente.it/forum/viewtopic.php?t=16622](https://www.matematicamente.it/forum/viewtopic.php?t=166220)
* <https://it.scienza.matematica.narkive.com/XNiiE9wo/probabilita-di-uscita-di-un-numero-su-una-ruota-del-lotto>
* <https://matematica.unibocconi.it/search/node/statistica>
* <https://www.wallstreetmojo.com/t-distribution-formula/>
* <https://statisticsbyjim.com/hypothesis-testing/hypothesis-tests-significance-levels-alpha-p-values/>
* <https://www.westga.edu/academics/research/vrc/assets/docs/confidence_intervals_notes.pdf>
* <https://www.scribbr.com/statistics/type-i-and-type-ii-errors/>
* <https://www.datasciencecentral.com/>
* <https://www.geo.fu-berlin.de/>
* <https://meetheskilled.com/test-di-ipotesi-one-sample-t-test/>
* <https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html>
* <https://www.datatechnotes.com/>
* <https://imstat.org/>
* <http://www.stat.yale.edu/>

# Acknowledgement

1. even if slightly out of context this is added for clarity and significance [↑](#footnote-ref-0)
2. <https://www.lotto-italia.it/lotto/come-dove-giocare/il-gioco/premi-del-lotto> [↑](#footnote-ref-1)
3. By Diva Jain - https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=84219892 [↑](#footnote-ref-2)
4. <https://www.quora.com/Is-the-mean-of-the-sampling-distribution-of-means-exactly-equal-to-the-population-mean> [↑](#footnote-ref-3)
5. (a note for Italian readers: note that “Confidence” translates to “Fiducia” and not to “Confidenza” as we’re used to say) [↑](#footnote-ref-4)
6. Hartmann, K., Krois, J., Waske, B. (2018): E-Learning Project SOGA: Statistics and Geospatial Data Analysis. Department of Earth Sciences, Freie Universitaet Berlin. [↑](#footnote-ref-5)
7. Minimi Quadrati Ordinari in Italian [↑](#footnote-ref-6)