The Statistics Handbook

Version 0.1

Carlo Occhiena

Feb 2023

Contents

1 Scope of this handbook

"Statistical analysis is the best way to predict events we do not know using information we do know."

We are used to talk 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, 25th of January 2023.

1.1 Versioning & Contributions

- Version 0.1 is the first release ever published and distributed online. It's the version written and verified personally by me but does not include any third-party contributions or revisions.
- I plan to submit the handbook to several SME (Subject Matter Experts).

- Each contribution will be indicated in the Acknowledgments section.
- The feedback from each SME will help raise the version by 1/10, so that with 9 revisions it will progress to version 1.0 of the document.
- Contributions are free and welcome, you can contact me via Linkedin.

1.2 LaTeX & Open Source Repository

In addition to being distributed under a Free Culture CC BY 4.0 license, all materials related to this handbook are available in the GitHub repository at the link: `https://github.com/carloocchiena/the_statistics_handbook`.

This also includes the LaTeX source of this handbook and an Excel with several exercises and applied formulas. This could therefore also be helpful to students and those who want to use this handbook for practical purposes.

1.3 Version History

- 0.1 first version ever distributed; written, checked, implemented by the Author, under his liability.

2 Core Concepts

2.1 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.

2.2 Property and type of data

We should not think that the data is solely a numerical value. There is a multitude of data types, each with specific characteristics.

2.2.1 Continuous vs Discrete

Discrete means data 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 dice: 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.

2.2.2 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. 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.

2.2.3 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.

2.2.4 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

- Parametric assumes 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.

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

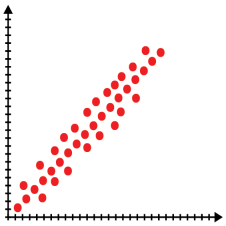
Nonparametric

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

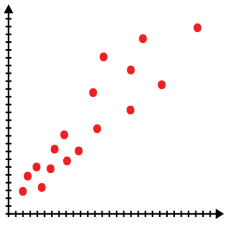
Nonparametric tests should be preferred whenever the dataset is not distributed in a normal (gaussian distribution) way, or, in any case, this specificity is not being demonstrated. A typical example is whenever the dataset is too small to prove a parametric distribution.

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.



Deterministic vs Stochastic

A **deterministic** model produces, for a specific set of inputs, the same exact results. Given the inputs, the result can be predicted accurately.

A **stochastic** model does not produce, for a specific set of inputs, a completely predictable result. The result account for a certain level of unpredictability or randomness.

Stochastic models can be analyzed statistically but may not be predicted precisely (such as Monte Carlo simulations).

Expected Value

The expected value (also called expectation, expectancy, mathematical expectation, mean, average) is a generalization of the weighted average.

Informally, the expected value is the arithmetic mean of a large number of independently selected outcomes of a random variable.

The expected value of a random variable with a finite number of outcomes is a weighted average of all possible outcomes. In the case of a continuum of possible outcomes, the expectation is defined by integration.

The expected value of a random variable X is often denoted by $E(X)$, $E[X]$, or EX, with E also often stylized as E or $\mathbb{E}$.

Linear, Nonlinear, and Monotonic Relationships

Linear:
When variables increase or decrease concurrently and at a constant rate, a positive linear relationship exists. When one variable increases while the other variable decreases, a negative linear relationship exists.

Nonlinear:
If a relationship between two variables is not linear, the rate of increase or decrease can change as one variable changes, causing a "curved pattern" in the data.

Monotonic:
In a monotonic relationship, the variables tend to move in the same relative direction, but not necessarily at a constant rate.

2.3 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, it is necessary to include a statistically relevant sample. A representative sample.

In particular, identifying the sample size, knowing the size of a specific population, is critical to the significance of statistical analysis.

A numerical example of this calculation is provided in the following section: “Calculate the Sample Size from a Population”.

The calculation has also been exemplified on the spreadsheet made available in the GitHub repository of this handbook.

Use “population” when:

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

Use “sample” 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.

2.4 Parameters vs Statistics vs Hyperparameters

Parameters describe the properties of the entire population.

Statistics describe the properties of a sample.

Hyperparameters[1] (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.

Hat symbols over variables ()

The estimated or predicted values in a regression or other predictive model in statistics are referred to as “hat values”.

$\hat{y}$: y is the outcome or dependent variable in the model equation, the ”hat” symbol ($\hat{}$) placed over the variable name is the statistical designation of an estimated value.

Outliers

An outlier is a data point that differs significantly from other observations. In regression analysis, outliers are the farther points from the regression line.

[1] even if slightly out of context this is added for clarity and significance

2.5 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.

2.6 Binomial Distribution

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.

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.

2.6.1 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.

2.7 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, or representative, of the collected data.

2.7.1 Mean

Mean is generically expressed as:

$$\frac{\text{sum of all data points}}{\text{number of data points}}$$

And, more specifically, with the formula:

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Mean hass the same meaning 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:
μ (mu) or even with $\bar{x}$ (x bar).

In the specific context of statistical studies:

- $\bar{x}$ is used for mean of a sample.
- μ is used for mean of the entire population.

Arithmetic Mean
It's the simplest and most common type of average, expressed as the sum of all data points over the count of data points.

Weighted Mean
It's similar to the arithmetic mean, except the fact that each of the data point contributes to the computation with its own weight factor.

$$\mu(x) = \frac{\sum_{i=1}^{k} x_i * n_i}{N}$$

For example, let's calculate 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: $\frac{((8*200)+(3*250)+(8*100))}{(8+3+8)} = 165.75$ grams.

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.

High and low end data points are called "outliers" (a data point that differs significantly from other observations).

2.7.2 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.

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

$mode = \varnothing$

2.7.3 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
$median = \frac{(27+32)}{2} = 29.5$

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

2.8 Measurement of Dispersion

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

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

2.8.1 Variance

Variance is the expectation of the squared deviation of a random variable from its population mean or sample mean. Variance is a measure of dispersion, meaning it is a measure of how far a set of numbers is spread out from their average value.

The always-positive value of variance 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 a better understanding of their impact on the dataset.

Variance is represented by: σ^2 (sigma squared) (when referred to population), s^2 (when referred to sample), $\mathrm{Var}(X), V(X), or \mathbb{V}(X)$

2.8.2 Standard Deviation

Standard deviation is a measure of the amount of variation or dispersion of a set of values. Standard deviation is equal to the square root of variance and it's represented with the Greek letter σ (sigma) or the letter s.

Being square rooted, the standard deviation returns a value that has again the same scale of the initial dataset, hence allowing for better comparisons and understanding of the statistics.

Mean, Variance, and Standard Deviation, are closed linked together.

	POPULATION (N)	SAMPLE (n)
Mean	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Standard Deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

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 to undershoot (and never to overshoot) the parameter being estimated.

It is possible to think of the Bessel's 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's correction to degrees of freedom.

2.9 Quartiles and IQR

A quartile is a type of quantile (quantiles are values that split sorted data or a probability distribution into equal parts) 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.

Quartiles:

- Quartile zero (Q0) corresponds to the first value of the ordered dataset.
- 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.
- Quartile four (Q4) corresponds to the last value of the ordered dataset.

IQR - Interquartile Range

IQR is a measure of statistical dispersion and it is defined as the difference between Q3 and Q1.

As an example, having an ordered dataset as following:
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 = Q3 - Q1 = 9.5 - 2.5 = 7

2.10 Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. The dependent variable y is also called response variable. The independent variable X is also called explanatory or predictor variables.

The resultant is a straight line intersecting the cartesian plane, attempting to minimize the distance (least-squares optimization) between actual output values so that hypothetical predicted values can be estimated.

Linear regression is a mathematical function based on the equation of the line:

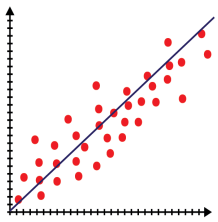
$Y_i = \beta_0 + \beta_1 X_i + u_i$

Where:

- i ranges between observations, $i = 1, \ldots, n$.
- Y_i is the dependent (response) variable.
- X_i is the independent (explanatory) variable.
- $\beta_0 + \beta_1 X$ is the regression function.
- β_0 is the line intercept (the value of y when $x = 0$).
- β_1 is the line angular coefficient.
- u_i is the statistical error.

Linear regression is a fundamental analysis of statistics, both because of its simplicity, interpretive immediacy and breadth of application cases.

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other, but that there is some significant association between the two variables.
A scatterplot can be a helpful tool in determining the strength of the relationship between two variables.
If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model.
A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.



Parameter estimations in the bivariate case

Generalizing the regression line equation, one can, in the case of the two-variable problem, start from:

$\hat{y} = mx + b + \varepsilon_i$,

Where:

- $\hat{y}$ is the dependent (response) variable.
- m is the line angular coefficient.
- b is the line intercept.
- ε_i is the statistical error.

At this point, the regression problem results in the determination of m and b so as to express the functional relationship between y and x as best as possible.

$$m = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$b = \frac{\sum y - m\sum x}{n}$$

3 Data Visualization

Data visualization (data viz) is the graphical representation of data. The main goals of data visualization are to make the phenomena within the dataset more evident, convey the embedded information in the analysis more efficiently, and reinforce cognitive aspects of the provided study (e.g., ease of reporting, memorability).

While data visualization pertains to the field of science and statistics, it has also taken on cross-cutting significance in purely artistic or design-related contexts.

Data visualization is so relevant that it could be considered a discipline within a discipline, with a deep vertical of study and insight that spans mathematical, scientific, statistical, cognitive, and humanistic domains.

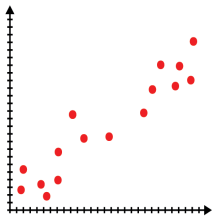
The recent spread of data science has made data viz even more important.

However, this paper will be limited to exploring some of the best-known forms of graphical representation in the field of statistics, and some of their properties.

3.1 Scatter Plot

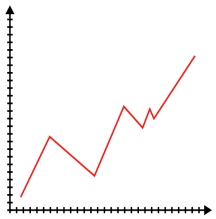
A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. Every data point is displayed as a dot.

Scatter plot has its most significance with continuous distributions.



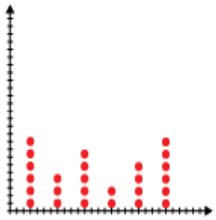
3.2 Line Chart

Line charts show the evolution of a continuous variable (often over a time horizon). A line chart is a way of plotting data points on a line. It is used to show trend data, or the comparison of two data sets.



3.3 Dot Plot

Dot plot[2] is a way to display data frequency piled over data points and along a number line.



3.4 Histograms

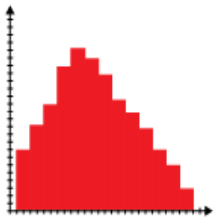
A histogram is a bar chart that groups continuous data into ranges. Ranges are discretional to the creator of the chart. For example, overall user ages (continuous dataset) can be grouped in clusters such as 0-10, 11-20 and such.

Histogram bars are adjacent (no spaces between bars).

Histograms don't have to be confused with bar charts:

- Histograms visualize quantitative data or numerical data. Usually, histograms display continuous variables.
- Bar charts display categorical (discrete) variables.

Correctly labeling horizontal (X) axis of an histogram chart is important in order to make it readable.

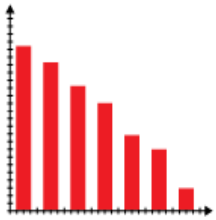


3.5 Bar Plot

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.

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

[2] In some texts, also called Line Plot.

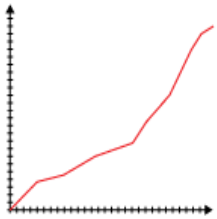


3.6 Ogive

An ogive, 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.

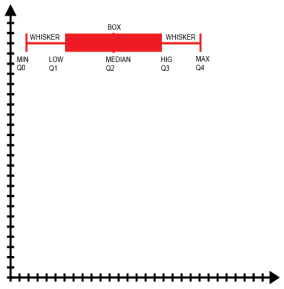
It is usually easier to create this kind of graph from a frequency table.



3.7 Box and Whisker Plot

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.

BWP are also called: box plot, box and whisker diagram, box and whisker plot with outliers[3].



[3] "Diagramma a scatola e baffi in italiano"

Box and whisker vs candlestick chart

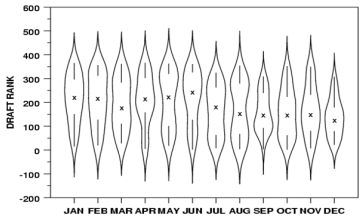
Mathematically speaking there is no difference. Both show an upper and lower boundary and points outside these boundaries.

However,a candlestick chart is mainly used in the finance industry. Its most popular application is to show stock prices. 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.

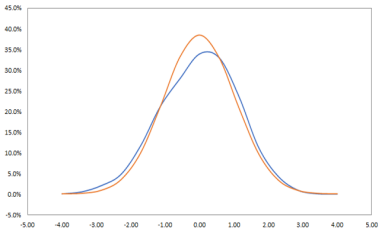
3.8 Violin Plot

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.



3.9 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.



4 Combinatorics

Combinatoric 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.

4.1 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 factorial.

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

An interesting property is also:

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

Example: 5! = 5 * 4! = 120

This leads to: $\frac{n!}{(n-1)!} = \frac{n(n-1)!}{(n-1)!} = n$

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:

$(n - 1)! = \frac{n!}{n}$

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 $n \in N$

4.2 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 permutations refer 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".

4.3 Combinations

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;
- a pear and an orange.

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

More formally, a $k-$*combination* of a set S is a subset of k distinct elements of S.

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:

$$C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

4.3.1 Permutations, Combinations and Dispositions

	REPETITION	NO REPETITION (simple)
Permutations	n^k	$\frac{n!}{(n-k)!}$ where k = cluster size $\frac{n!}{k1! * k2! * kn!}$ where k = items repeated
Combinations	$\frac{(n+k-1)}{k!(n-1)!}$	$\frac{n!}{k!(n-k)!}$
Dispositions	n^k	$\frac{n!}{(n-k)!}$

Examples

Permutations with repetitions
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" being a valid number)?

$n = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9] = 10$
$k = \_, \_, \_, \_, \_, \_, \_ = 7$
Each slot of k allows 10 combinations, so it's $n^k = 10^7$

Permutations with no repetitions
Find all the way the word MAMA can be arranged.

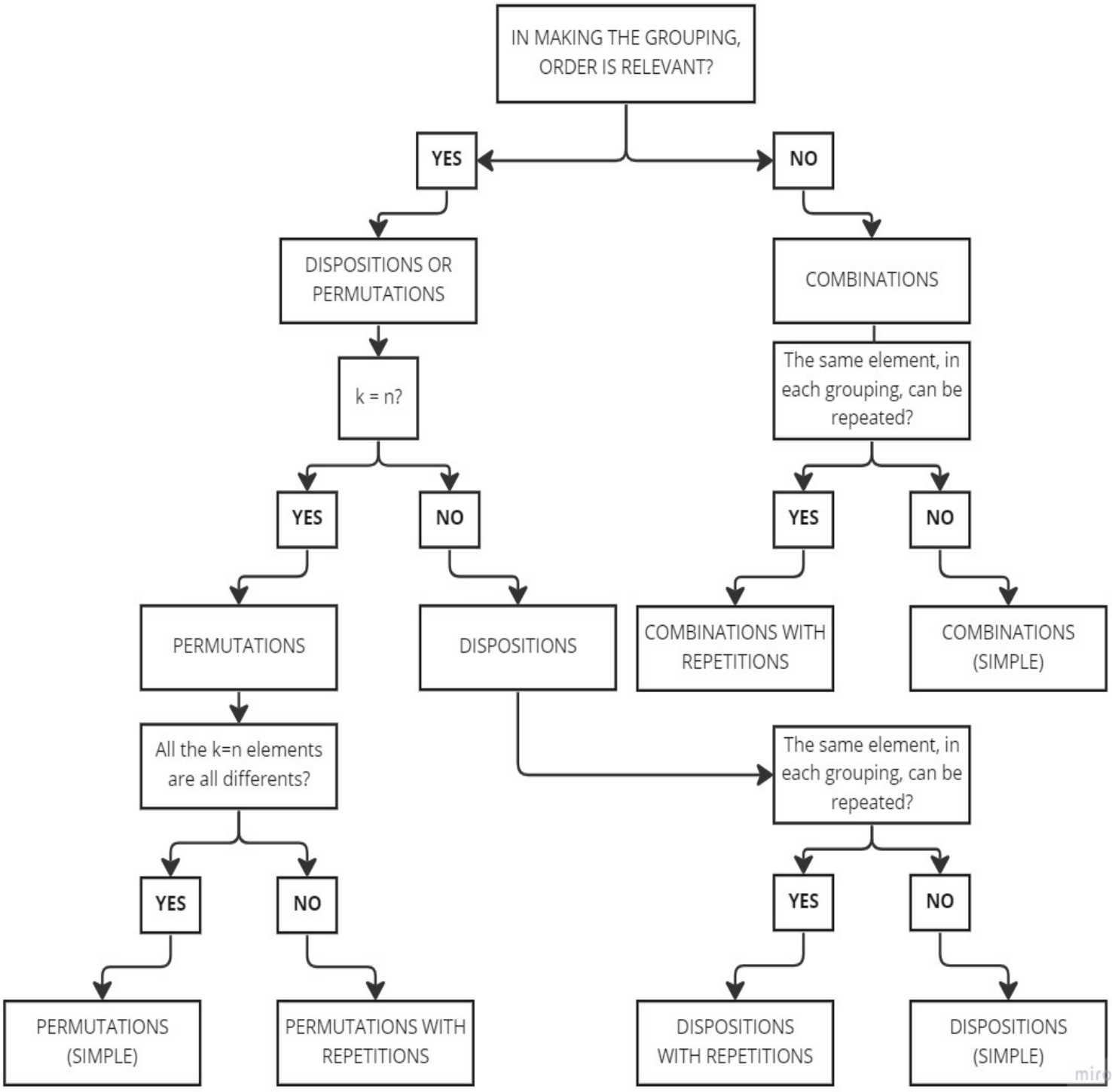
n = 3
k1 = 2 (the letter M is repeated 2 times)
k2 = 2 (the letter A is repeated 2 times)
4! / (2! * 2!) = 24 / 4 = 6

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? A diagram

Key insight:
Is the order relevant?

- YES = PERMUTATIONS (ex. numbers)
- NO = COMBINATIONS (ex. people in teams)


IN MAKING THE GROUPING, ORDER IS RELEVANT?
YES
NO
DISPOSITIONS OR PERMUTATIONS
COMBINATIONS
k = n?
The same element, in each grouping, can be repeated?
YES
NO
YES
NO
PERMUTATIONS
DISPOSITIONS
COMBINATIONS WITH REPETITIONS
COMBINATIONS (SIMPLE)
All the k=n elements are all differents?
The same element, in each grouping, can be repeated?
YES
NO
YES
NO
PERMUTATIONS (SIMPLE)
PERMUTATIONS WITH REPETITIONS
DISPOSITIONS WITH REPETITIONS
DISPOSITIONS (SIMPLE)

5 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.

Probability Notation

$P(A)$	Individual probability	The probability of event A happening
$P(A')$	Complement	The probability of event A not happening
$P(A')$	Complement	The probability of event A not happening
$P(A \cup B)$	Union	The probability of both A and B happening for both datasets (all elements of A plus all elements of B).
$P(A \cap B)$	Union	The probability of both A and B happening for both datasets (all elements of A plus all elements of B).
$P(A\|B)$	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 $P \cup Q$, and $P \cap Q$?

$P \cup Q = \{1, 2, 3, 5, 7, 9\}$
$P \cap Q = \{3, 5, 7\}$

5.1 Simple Probability

Simple probability define how likely a specific event A is going to happen in the given scenario.

$P(A)$ = target events / total events

And, consequentially:

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

5.1.1 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.
- Expected probability is the theoretical probability coming from applying the probability formula to the scenario.

The expected probability of having head over a coin toss is 50%. However over a 100 tosses, the experimental probability may vary (ex. resulting in 30% heads).

5.1.2 Law of Large Numbers

The law of large numbers, or Bernoulli's theorem (since its first formulation is due to Jakob Bernoulli), describes the behavior of the mean of a sequence of n trials of a random variable, independent and characterized by the same probability distribution (n measurements of the same magnitude, n tosses of the same coin, etc.), as the numerosity of the sequence itself n tends to infinity.

Regression toward the mean

Regression toward the mean (also called reversion to the mean, and reversion to mediocrity) is the phenomenon where if one sample of a random variable is extreme, the next sampling of the same random variable is more probable to be closer to its mean.

This is linked to the law of large numbers. Increasing the size of the sample and the length of the observations, the event outcomes will tend toward the population mean. Law of large number is explaning the whole phenomena, while regression toward the mean is useful to understand the expected behaviour of a single observation.

However, in no sense does the future event "compensate for" or "even out" the previous event.

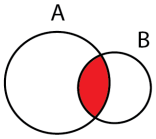
5.1.3 Probability Addition Rule

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)

Or, with sets notation as:

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

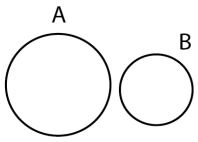


If A and B are two mutually exclusive events,
$P(A \cap B) = 0$. Then the probability that either one of the events will occur is:

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

Or, with sets notation as:

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



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.

5.1.4 Conditional Probability for Independent and Dependent Events

Independent event probability
The probability of A and B happening.

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

Tossing two coins A and B, what is the probability of having two head values?
Coins are independent each other, so:

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

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

$P(A \cap B \cap C) = 2/100 * 2/100 * 2/100 = \frac{8}{100^3} = 1/125'000$

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

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

What's the probability of drafting two Kings in a row from a standard deck of cards?
$P(A \cap B) = 4/52 * 3/51 = 1/13 * 1/17 = 1/221$

Reminder

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

5.2 Bayes Theorem

Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

$P(A \mid B) = \dfrac{P(B \mid A)P(A)}{P(B)}$

where:

- A and B are events and $P(B) \neq 0$.
- $P(A)$ 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 1

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 call:

- $P(B)$ the probability of a product being defective.
- $P(A)$ the probability of a product coming from line 1.

Hence:

$P(A) = 1/2$
(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 = $[(1/2) * (3/100)] + [(1/2) * (1/100)] = 3/200 + 1/200 = 4/200 = 1/50$

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 2

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: $(100\% - 99\%) = 1\%$

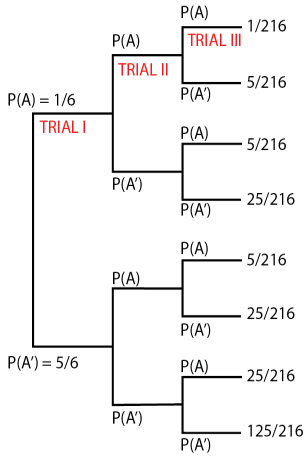
	SICK	NOT SICK	TOTAL
TESTED POS	0.99[1]	9.99[3]	10.98
TESTED NEG	0.01[2]	989.01[4]	989.02
TOTAL	1	999	1'000

[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%

5.2.1 Tree Diagrams

A tree diagram is a type of diagram that can be useful as an aid in computing probabilities.
For example, consider an experiment of tossing a six-sided dice. 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, 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.



5.3 Discrete Probability

Discrete probability deals with events with a finite or countable number of occurrences.
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: $2^3 = 8$

Possible outcomes:

EVENT	N. OF HEADS
HHH	3
THH	2
HTH	2
TTH	1
HHT	1
THT	1
HTT	1
TTT	0

HEADS IN 3 COIN FLIPS (X)	P(X)
0	1/8
1	3/8
2	3/8
3	1/8

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

Mean $= 3/2$

Variance $= \sigma^2 = (0 - 3/2)^2 * 1/8 + (1 - 3/2)^2 * 3/8 + (2 - 3/2)^2 * 3/8 + (3 - 3/2)^2 * 1/8 = 0,75$

Standard Deviation $= \sigma = \sqrt{\sigma^2} = 0.866$

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 heads as a result. This is a case of theoretical probability vs experimental probability).

5.3.1 Transforming Random Variables

How the distribution of a random variable changes when the variable is transformed in a deterministic way?

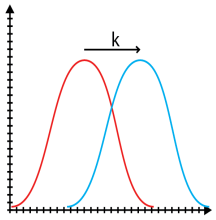
Shifting Data

Shifthing data means adding a constant $k \in \mathbb{R}$ to each member of a dataset.

X	Y
3	3 + K
3	3 + K
7	7 + K
10	10 + K
12	12 + K

Dataset	Shifted, K
Mean: 6	Mean: 6 + K
Median:	Median: 3 + K
Mode: 3	Mode: 3 + K
Range: 10	Range: 10
IQR: 8	IQR: 8
St. Dev: σ	St. Dev: σ

Impact of K-shifting: 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.

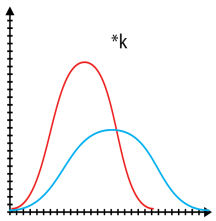


Scaling Data

Scaling data means multipying each member of a dataset by a constant k.

Dataset	Scaled, K
Mean: 6	Mean: 6K
Median:	Median: 3K
Mode: 3	Mode: 3K
Range: 10	Range: 10K
IQR: 8	IQR: 8K
St. Dev: σ	St. Dev: σ K

Impact of K-scaling: 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 standard deviation, IQR and range).



5.3.2 Linear Combinations of Random Variables

Let X_1 and X_2 be two independent random variables. Let a and b be scalars. Then a linear combination of the variables X_1 and X_2 and is defined to be any other random variable of the form $Y = aX_1 + bX_2$.

Base assumptions:

- Variables must be independent.
- Variables must have matching units of measurements.

	$\sum$	Δ
Combination	$\sum = X + Y$	$\Delta = X - Y$
Mean	$\mu = \mu x + \mu y$	$\mu = \mu x - \mu y$
Variance	$\sigma^2 = \sigma^2 x + \sigma^2 y$	$\sigma^2 = \sigma^2 x + \sigma^2 y$

Example: Time in hours 4 managers manage timesheets:
$X = [1, 2, 2, 3]$

Time in hours 4 HRs manages payrolls on such timesheets:
$Y = [2, 3, 5, 6]$

$\mu X = (1 + 2 + 2 + 3) \ / \ 4 = 2$

$\mu Y = (2 + 3 + 5 + 6) \ / \ 4 = 4$

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

$\sigma X = \sqrt{0.5} = 0.70$

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

$\sigma Y = \sqrt{2.5} = 1.58$

$X + Y$
$\mu = 2 + 4 = 6$

$\sigma^2 = 0.5 + 2.5 = 3$

$\sigma = \sqrt{3} = 1.73$

$X - Y$
$\mu = |2 - 4| = 2$

$\sigma^2 = 0.5 + 2.5 = 3$

$\sigma = \sqrt{3} = 1.73$

5.3.3 Fair Game

A fair game[4] 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

Is state lottery a fair game?
The Italian national lottery ("Lotto") is a bingo where 5 numbers are drawn out of a poll of 90. Repetition is not allowed (extracted numbers are discarded)[5].

Players win if they have all the 5 numbers on the 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[6] of 5 numbers are $C_n^k = \binom{n}{k}$ hence:

$\frac{90!}{5!(90-5)!} = 43'949'268$

This is obviously matching the official lottery statement.

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

[4] Italian translation: gioco equo
[5] `https://www.lotto-italia.it/lotto/come-dove-giocare/il-gioco/premi-del-lotto`
[6] see Combinatorics chapter for details

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

6 Joint Distributions

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).

6.1 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.

Expressed in mathematical notation, covariance formulas are:

Population Covariance:

$$cov(X, Y) = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$$

Sample Covariance:

$$cov(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

6.2 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 be close to 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 measures just 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).

6.2.1 Pearson Correlation

Pearson correlation is the most widely used correlation statistic to measure the degree of the relationship between linearly related variables.

For the Pearson 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.
- Data have no outliers.
- Variables linearity.
- Variables homoscedasticity (homogeneity of variance).

For a population
Pearson correlation is expressed with the Greek letter ρ (rho) when referred to population.

Given a pair of random variables (X, Y), the formula for ρ is:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance.
- σ_X is the standard deviation of X.
- σ_Y is the standard deviation of Y.

For a sample

Pearson's correlation coefficient, when applied to a sample, is commonly represented by r_{xy} and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient. We can obtain a formula for r_{xy} by substituting estimates of the covariances and variances based on a sample into the formula above.

Given paired data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ consisting of n pairs, r_{xy} is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n is sample size.
- x_i, y_i are the individual sample points indexed with i.
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$.

Rearranging gives us this formula for r_{xy}:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\ \sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

where n, x_i, y_i are defined as above.

This formula suggests a convenient single-pass algorithm for calculating sample correlations.

6.2.2 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 τ (tau)

$$\tau = \frac{n_c - n_d}{n_c + n_d} = \frac{n_c - n_d}{n(n-1)/2}$$

where:

- n_c is the number of concordant pairs.
- n_d is the number of discordant pairs.
- n is the number of pairs.

Kendall rank assumptions:

- Pairs of observations are independent.
- Two variables should be measured on an ordinal, interval or ratio scale.
- Monotonic relationship between the two variables.

6.2.3 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, ρ and r.

For a sample of size n, the n raw scores X_i, Y_i are converted to ranks $R(X_i), R(Y_i)$, and r_s is computed as:

$$r_s = \rho_{R(X),R(Y)} = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}$$

where:

- ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables.
- $cov(R(X), R(Y))$ is the covariance of the rank variables.
- $\sigma_{R(X)}, \sigma_{R(Y)}$ are the standard deviations of the rank variables.

Spearman rank assumptions:

- Pairs of observations are independent.
- Two variables should be measured on an ordinal, interval or ratio scale.
- Monotonic relationship between the two variables.

6.2.4 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 r_{pb}.

7 Data Distributions

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.

7.1 Probability Mass Function (PMF)

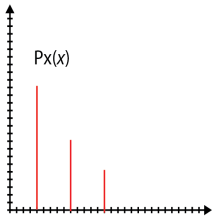
Probability mass function is a function that gives the probability that a discrete random variable is exactly equal to some value.

$f(x) = P[X = x]$

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

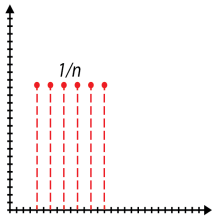
Example:
What is the probability of picking a specific ball out of a jar with 100 balls, all the balls being equal? It's 1/100, or 1%.

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).



PMF Overview

- Notation: $\mathcal{U}\{a, b\}$ or unif$\{a, b\}$
- Parameters: a, b integers with $b \geq a$, $n = b - a + 1$
- Support: $k \in \{a, a + 1, \ldots, b - 1, b\}$
- PMF: $\dfrac{1}{n}$

- CDF: $\frac{\lfloor k \rfloor - a + 1}{n}$
- Mean: $\frac{a+b}{2}$
- Median: $\frac{a+b}{2}$
- Mode: N/A
- Variance: $\frac{n^2 - 1}{12}$

7.2 Probability Density Function (PDF)

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).

$$Pr[a \leq X \leq b] = \int_a^b f_X(x)dx$$

with:

- $P(x = c) = 0$ (the probability P of x to be equal to a constant c is zero).

More technically, 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.

For $f(x)$ to be a legitimate PDF, it must satisfy the following two conditions:

- $f(x) \geq 0, \forall x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

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:

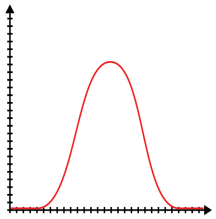
$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

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).

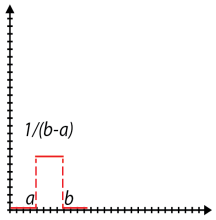
PDF Visualization



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 (ex. $[a, b]$) or open (ex. (a, b)). Therefore, the distribution is often abbreviated $U(a, b)$, where U stands for uniform distribution.



CUD Overview

- Notation: $\mathcal{U}_{[a,b]}$
- Parameters: $-\infty < a < b < \infty$
- Support: $x \in [a, b]$
- PDF: $\begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$
- CDF: $\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x > b \end{cases}$
- Mean: $\frac{1}{2}(a + b)$
- Median: $\frac{1}{2}(a + b)$

- Mode: any value in (a, b)
- Variance: $\frac{1}{12}(b-a)^2$

7.3 Cumulative Distribution Functions (CDF)

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

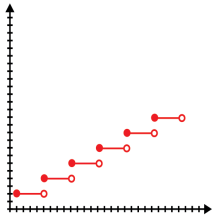
$F(x) = P(X \leq x), \forall x \in \mathbb{R}$

CDF expresses the cumulative probability of a given event.

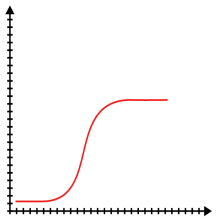
Discrete Cumulative Distribution Function
A cumulative distribution function in a discrete set.

As an example, rolling a standard dice:

- the CDF of having no numeric input is 0.
- the CDF of a number from 1 to 6 is 1 (6/6).
- the CDF of a number less or equal to 3 is 3/6.



Continuous Cumulative Distribution Function
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, ex. computing the PDF of a function of a random variable.



7.4 Hypergeometric Distribution

The hypergeometric distribution is a discrete probability distribution that describes the probability of k successes (random draws for which the object drawn has a specified feature) in n draws,

without replacement, from a finite population of size N that contains exactly K objects with that feature, wherein each draw is either a success or a failure.

In contrast, the binomial distribution describes the probability of k successes in n draws with replacement.

The following conditions characterize the hypergeometric distribution:

- The result of each draw (the elements of the population being sampled) can be classified into one of two mutually exclusive categories.
- The probability of a success changes on each draw, as each draw decreases the population (sampling without replacement from a finite population).

(e.g. Pass/Fail or Employed/Unemployed).

A random variable X follows the hypergeometric distribution if its probability mass function (pmf) is given by:

$$p_X(k) = \Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where

- N is the population size.
- K is the number of success states in the population.
- n is the number of draws (i.e. quantity drawn in each trial).
- k is the number of observed successes.
- $\binom{a}{b}$ is a binomial coefficient.

7.5 Binomial Distribution

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

$$f(k, n, p) = \Pr(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, \ldots, n$ where, as already explained:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

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.

Binomial Distribution Overview

- Notation: $B(n, p)$
- Parameters:
  - $n \in \{0, 1, 2, \ldots\}$ – number of trials
  - $p \in [0, 1]$ – success probability for each trial
  - $q = 1 - p$
- Support: $k \in \{0, 1, \ldots, n\}$ – number of successes
- PMF: $\binom{n}{k} p^k q^{n-k}$
- CDF: $I_q(n - k, 1 + k)$ (the regularized incomplete beta function)
- Mean: np
- Median: $\lfloor np \rfloor$ or $\lceil np \rceil$
- Mode: $\lfloor (n + 1)p \rfloor$ or $\lceil (n + 1)p \rceil - 1$
- Variance: npq

Example of a binomial distribution:
According to a report, 80% of prospects at company α will result in a signed contract. Each prospect is independent of each other.

It's asked to calculate the probability to close a deal from a round of 3 prospects they are working on.

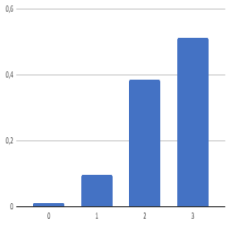
Hence:

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

$$\binom{n}{k} = \binom{3}{1} = \frac{3!}{1!(3-1)!} = 3$$

and then $B = 3 * (0.8^1)(1 - 0.8)^(3 - 1) = 0.096$

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



7.6 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 (boolean) question.

Such questions lead to outcomes that are boolean-valued: a single bit whose value is success, yes, true, 1 with probability p and failure, no, false, 0 with probability $q = 1 - p$.

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 $p \neq 1/2$

Bernouilli Distribution Overview

- Notation: $X \sim B(1, p)$
- Parameters:
  - $0 \leq p \leq 1$
  - $q = 1 - p$
- Support: $k \in \{0, 1\}$
- PMF: $\begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$
- CDF: $\begin{cases} 0 & \text{if } k < 0 \\ 1 - p & \text{if } 0 \leq k < 1 \\ 1 & \text{if } k \geq 1 \end{cases}$
- Mean: p

- Median: $\begin{cases} 0 & \text{if } p < 1/2 \\ [0,1] & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
- Mode: $\begin{cases} 0 & \text{if } p < 1/2 \\ 0,1 & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
- Variance: $p(1-p) = pq$

7.7 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 $\lambda > 0$, if it has a probability mass function given by:

$$f(k;\lambda) = \Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where:

- k is the number of occurrences ($k = 0, 1, 2, \ldots$).
- e is Euler's number ($e = 2.71828\ldots$).
- ! 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 (λ) is known and assumed to be constant.
- Occurrences occur in an interval and are discrete and countable.
- 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 (it means that intervals are not overlapping).
- Probability is proportional to interval size.

Some examples of Poisson distribution are:

- The number of chewing gum on a single tile over a sidewalk.
- The number of planes that fly over a specific 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 time an event is investigated is k and is assumed to be $k = 2$ (we want to calculate the probability that in a year 2 soldiers will die from a horse kick).

Hence:

$$\mathrm{P}(X = k \text{ where } k = 2) = \frac{0.61^2 e^{-0.61}}{2!} = 0.101$$

Poisson Distribution Overview

- Notation: $Pois(\lambda)$
- Parameters: $\lambda \in (0, \infty)$ (rate)
- Support: $k \in \mathbb{N}_0$ (Natural numbers starting from 0)
- PMF: $\frac{\lambda^k e^{-\lambda}}{k!}$
- CDF: $\frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !}$, or $e^{-\lambda} \sum_{j=0}^{\lfloor k \rfloor} \frac{\lambda^j}{j!}$,
  or $Q(\lfloor k+1 \rfloor, \lambda)$
  (for $k \geq 0$, where $\Gamma(x, y)$ is the upper incomplete gamma function, $\lfloor k \rfloor$ is the floor function, and Q is the regularized gamma function).
- Mean: λ
- Median: $\approx \left\lfloor \lambda + \frac{1}{3} - \frac{1}{50\lambda} \right\rfloor$
- Mode: $\lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$
- Variance: λ

8 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:

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

where:

- σ is the standard deviation of the distribution.
- π is the mathematical constant (3.14159...).
- e is Euler's number ($e = 2.71828\ldots$).
- μ is the mean of the distribution.

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 distribution used in business, statistics, biology, etc., since so many real-life datasets end up resembling a normal distribution (see also: Central Limit Theorem).

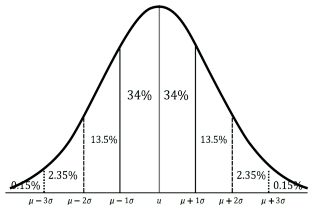
Normal distributions are referred to with capital letter N, such as $N(5, 9)$ means a normal distribution with mean 5 and variance 9.

Normal distributions have unique properties of mean and standard deviation.

The Empirical Rule

The below chart helps understanding the empirical rule.

The empirical rule (also known as three-sigma rule or 68-95-99.7 rule) states that for a normal distribution, almost all observed data will fall within the range of 3σ from the mean μ.



In particular, it can be stated that:

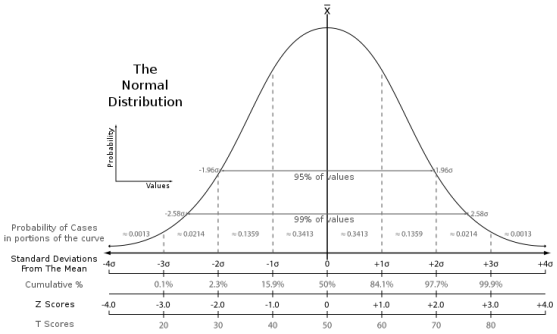
- 68% of observations will fall within the first standard deviations ($\mu \pm \sigma$).
- 95% of observations will fall within the first two standard deviations ($\mu \pm 2\sigma$).
- 99.7% of observations will fall within the first three standard deviations ($\mu \pm 3\sigma$)

8.1 Z-Score

Z-score (also known as 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.

Less technically, z-scores represents how a given data point is far from the mean.

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-score standardization formula is:

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

8.1.1 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 hundredth values on the table header. To extrapolate data from z-table:

- Turn data into a normal distribution.
- Find the matching z-score to the left of the table and align it with the z-score at the top of the table.
- The result gives you the probability.

The image below shows the probability for a z-score of 1.2 + 0.05 = 1.25, that is 0.89435.

z-table	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08
0	0,5	0,503989	0,507978	0,511966	0,515953	0,519939	0,523922	0,527903	0,531881
0,1	0,539828	0,543795	0,547758	0,551717	0,55567	0,559618	0,563559	0,567495	0,571424
0,2	0,57926	0,583166	0,587064	0,590954	0,594835	0,598706	0,602568	0,60642	0,610261
0,3	0,617911	0,62172	0,625516	0,6293	0,633072	0,636831	0,640576	0,644309	0,648027
0,4	0,655422	0,659097	0,662757	0,666402	0,670031	0,673645	0,677242	0,680822	0,684386
0,5	0,691462	0,694974	0,698468	0,701944	0,705401	0,70884	0,71226	0,715661	0,719043
0,6	0,725747	0,729069	0,732371	0,735653	0,738914	0,742154	0,745373	0,748571	0,751748
0,7	0,758036	0,761148	0,764238	0,767305	0,77035	0,773373	0,776373	0,77935	0,782305
0,8	0,788145	0,79103	0,793892	0,796731	0,799546	0,802337	0,805105	0,80785	0,81057
0,9	0,81594	0,818589	0,821214	0,823814	0,826391	0,828944	0,831472	0,833977	0,836457
1	0,841345	0,843752	0,846136	0,848495	0,85083	0,853141	0,855428	0,85769	0,859929
1,1	0,864334	0,8665	0,868643	0,870762	0,872857	0,874928	0,876976	0,879	0,881
1,2	0,88493	0,886861	0,888768	0,890651	0,892512	0,89435	0,896165	0,897958	0,899727
1,3	0,9032	0,904902	0,906582	0,908241	0,909877	0,911492	0,913085	0,914657	0,916207
1,4	0,919243	0,92073	0,922196	0,923641	0,925066	0,926471	0,927855	0,929219	0,930563
1,5	0,933193	0,934478	0,935745	0,936992	0,93822	0,939429	0,94062	0,941792	0,942947
1,6	0,945201	0,946301	0,947384	0,948449	0,949497	0,950529	0,951543	0,95254	0,953521

A z-table can also be created from the ground-up, in example:

1. with Excel using the formula:
   (a) =+NORM.S.DIST($A2 + B1$;TRUE), having "A" column = [0, 0.1, 0.2, ..., 3.4] and row 1 = [0, 0.01, 0.02, ..., 0.09]
2. with a programming language, such as R or Python (here is script for generating a z-table in Python: gist.github.com/carloocchiena/)

8.2 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.

8.3 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.

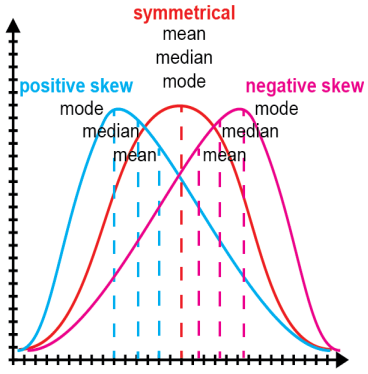
Would be useful to remember that for a symmetric distribution, and, specifically, for a normal distribution, median = mean = mode.

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

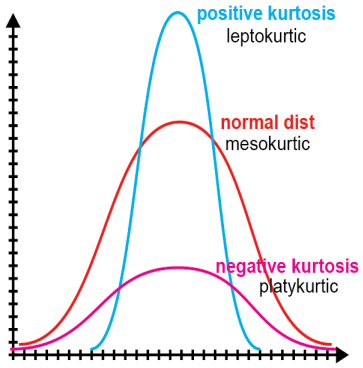


8.4 Kurtosis

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.

It can also be stated more simply that skewness represents the vertical distance of the curve from a normal distribution, while kurtosis represents the horizontal distance of the curve from a normal distribution.

- **Skewness**: lack of symmetry in the distribution.
- **Kurtosis**: height and sharpness of the central peak.



8.5 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.

9 Sampling

Sampling is the selection of a subset (a **sample**) of individuals from within a statistical **population** to estimate characteristics of the whole population. Statisticians attempt to collect samples that are representative of the population in question. Sampling has lower costs and faster data collection than measuring the entire population and can provide insights in cases where it is infeasible to measure an entire population.

Sampling allows to test a hypothesis about general characteristics of a population. Samples are used to make inferences about a population.

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

9.1 Sampling Methodologies

The representativeness of the selected sample is a crucial element in creating a good sampling. Therefore, there are different types of sampling that can be chosen appropriately according to the characteristics of the population.

Some of the discriminating criteria in choosing a sampling method may be:

- Nature and quality of the frame.
- Availability of auxiliary information about units on the frame.
- Accuracy requirements, and the need to measure accuracy.
- Whether detailed analysis of the sample is expected.
- Cost/operational concerns.

In general terms, it can be stated that a good sampling is:

- Representative of the population.
- Unbiased.

9.1.1 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.

9.1.2 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 (i.e. one out of every 10 members).

9.1.3 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.

9.1.4 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.

9.2 Central Limit Theorem

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.

It implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions.

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

- The mean of the distribution of sample means is the **expected value of** M and is always equal to the population mean μ.
- The standard deviation of the distribution of sample means is **the standard error of** M (SE) and is computed by: $\sigma M = \frac{\sigma}{\sqrt{n}}$. As the sample size increases, the error decreases. As the sample size decreases, the error increases. At the extreme, when n = 1, the error is equal to the standard deviation.
- The shape of the distribution of sample means tends to be normal. It is guaranteed to be normal if either:
  - The population from which the samples are obtained is normal.
  - The sample size (n) is $n \geq 30$.

9.2.1 Sampling Distribution of the Sample Mean

The distribution of sample means is defined as the set of means from all the possible random samples of a specific size (n) selected from a specific population.

If repeated random samples of a given size n are taken from a population of values for a quantitative variable, the mean of all sample means is population mean.

$\bar{x} = \mu$

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

This statement can leads 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 (combinations)**.

The idea is that when we think of taking a sample, there are a large number of possible samples, from which we will choose just one. However, we 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 ("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."

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. There are other candidates for an estimator of the population mean, such as the median, the mode, or the geometric average. While they might be unbiased in certain situations, none of them are "in general" an unbiased estimator of the mean, like the sample mean is.[7]

Finite Population Correction Factor (FPC)

Finite Population Correction Factor (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).

$FPC = \sqrt{\frac{N-n}{N-1}}$

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

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

$\sigma M = \frac{\sigma}{\sqrt{n}}$

By applying the finite population correction, the formula becomes:

$\sigma M = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

9.3 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.

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".

[7] This explanation has been provided in an online debate from Dwight Galster, Statistics PhD from NDSU.

The formula to calculate the T-distribution is:

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

where:

- $\bar{x}$ is the sample mean.
- μ is the population mean.
- s is the standard deviation.
- n is the size of the given sample.

9.3.1 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 $n - 1$.

9.3.2 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 on a T-table[8] with a process similar to the Z-scores.

When to use Z-distribution and when T-distribution?

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

9.4 Mean Estimation and Confidence Intervals

Estimation, in statistics, is concerned with inference about the numerical value of unknown population values from incomplete data such as a sample from a population.

- 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".

[8] https://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf

- If the estimate is made from a range of values where the parameter is expected to lie, it's called an "interval estimate".

9.4.1 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, at $t0$, the accuracy of a particular approximation is not known precisely.

9.4.2 Interval Estimation

Interval estimation is the use of sample data to estimate an interval of plausible values of a parameter of interest. This is in contrast to point estimation, which gives a single value.

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.

9.4.3 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 α (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:

$CI = \bar{x} \pm MOE$

where:

- $\bar{x}$ is the sample mean.
- $MOE = z_y SE = z_y \frac{\sigma}{\sqrt{n}}$ is the Standard Error, z_y being the quantile or z-score.

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

9.4.4 Confidence Level (CL)

The confidence level measures the level of trust[9] 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.

9.4.5 Confidence Intervals 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

These values are particularly useful in estimating the sample size from a population, assuming it is following a normal distribution.

9.4.6 Calculate the Sample Size from a Population

In identifying the sample size, one must make sure that it is significant for the population size and the estimated margin of error and confidence level.

This dimension can be obtained by applying the properties of the normal distribution, with the following formula:

$$n = N \frac{\frac{z^2 p(1-p)}{MOE^2}}{N - 1 + \frac{z^2 p(1-p)}{MOE^2}}$$

where:

- N is the population size.
- n is the sample size.
- p is the p-value.
- z is the z-score for the given confidence level.
- MOE is the margin of error.

Calculating Sample Size: an Example

- $N = 500$
- $CI = 95\%$
- $MOE = 2\%$
- $p = 0.5$ (conservative value for unknown or uncertain scenarios).

[9] a note for Italian readers: "Confidence" translates to "Fiducia" and not to "Confidenza" as we're used to say

- $z = 1.96$ (z-value associated with such CI, see previous section).

$n = 414$

There are myriads of tools and online calculators for measuring sample size (or verifying the calculation).

Among the many, a good one is: `https://www.checkmarket.com/sample-size-calculator/`.

10 Hypothesis Testing

Hypothesis testing is a method used in statistical inference to validate specific assumptions made over a population parameters 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.

10.1 Null & Alternative Hypotheses

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.

10.1.1 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 α 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 happens 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 = $1 - \beta$
Accepted	Correct decision True negative Probability = 1 - α	Type II error False negative Probability = β

Tradeoff between Type I and Type II errors

Type I and Type II errors influence each other.

- Setting a lower LOS decreases Type I but increases incidence of Type II errors.
- Increasing power of a test decreases 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[10].

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.
- Type II errors 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.

10.2 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, with t-scores being used where the sample size is small and the population's standard deviation is unknown.

One-sample z-test

$$z = \frac{\overline{x} - \mu_0}{(\sigma/\sqrt{n})}$$

With:

[10] 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.

- Normal population or n large) and σ known.
- z being the distance from the mean in relation to the standard deviation of the mean.

One-sample t-test

$t = \frac{\overline{x} - \mu_0}{(s/\sqrt{n})}$

With:

- Normal population or n large and σ unknown.
- $df = n - 1$

10.2.1 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 $\alpha/2$ (both sides of the distribution, right and left), more extreme values are needed in order to incur in a rejection of the null hypothesis. Hence, two-tailed tests are more conservative, and one should assumes 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 is equal to α (one side of the distribution, right or left).

10.3 P-Value and Critical Value

p-value and critical value are meant to do the same thing: support or reject the null hypothesis. Both are measure of evidence to differentiate between randomness and causality. The same result is then obtained following different approaches.

The test process will follow this roadmap:

- State Null ($H0$) and Alternate (Ha) hypotheses.
- Choose LOS (level of significance, α).
- Calculate test statistics from the sample.

Then, for **p-value** approach:

- Compute p-value.
- Compare p-value to α level.
- Reject null if $p \leq \alpha$ level.

And for **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).

Try it with Python:
I made an interactive Jupyter Notebook to support the understanding of both approaches; it can be found in the GitHub repository of the handbook.

P-value and Critical Value: a comparison

P-value: 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. This means measure the statistical significance of the observations.

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.[11]

Common errors regarding the use of p-value:

- The p-value is not the probability that the null hypothesis is true or the probability that the null hypothesis is false. It is not related to either.
- The p-value is not the probability that an observation is a random event. The calculation of p-value is based on the assumption that every observation is a random event, a random result. The phrase "the result is due to random chance" usually means that the null hypothesis is probably correct, but remember that p-value cannot be used to represent the probability that a hypothesis is true.
- The p-value is not the probability of rejecting the null hypothesis when it is true.
- The p-value is not the probability that replicating the experiment would yield the same conclusion. To quantify the replicability of an experiment, the concept of p-rep was introduced.
- The significance level α is not determined by the p-value. The significance level is decided by the person conducting the experiment before seeing the data.

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.

[11] Hartmann, K., Krois, J., Waske, B. (2018): E-Learning Project SOGA: Statistics and Geospatial Data Analysis. Department of Earth Sciences, Freie Universitaet Berlin.

- If the test statistic is not as extreme as the critical value, the null hypothesis is not rejected.

10.4 A/B Testing

A/B testing has become particularly well known due to its extensive use in marketing and its declinations (product validation, UX, CTAs). From a statistical point of view, it can be defined as a two-sample hypothesis test (two variables that are independent of each other).

From a practical point of view, implementing an A/B testing strategy requires grounding most of skill set of a statistician:

- Identifying the target population.
- Identify a sample.
- Formulate $H0$ null and Ha alternative hypotheses.
- Define a confidence interval.
- Perform the experiment.
- Collect the results.
- Assess the statistical significance of observations.
- Validate or reject the hypothesis made.

Common use-cases requiring A/B testing are[12]:

- Email marketing (open rates, click-thru rates).
- User Interfaces (buttons, image size, background colors).
- Programming routines (APIs performances, HTTP routing).
- Advertising campaign.

[12]You may want to check this Jupyter Notebook (link) created by F.Casalegno, with a neat real world example.

11 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 jargon) and one or more **independent variables** (often called "predictors", "covariates", "explanatory variables" or "features").

Regression analysis is primarily used for two conceptually distinct purposes:

- For prediction and forecasting, where its use has substantial overlap with the field of machine learning.
- 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 expressed as:

$y = mx + b$

where:

- y is the independent variable.
- m is the slope.
- x is the independent variable.
- b is the intercept (the expected mean value of y when all x=0, where the function crosses the x-axis).

To solve the linear regression function in a context where there are many possible features of x, such as in a prediction under uncertainty, a specific methodology is needed.

Ordinarry least squares (OLS) is one of the possible approaches.

11.1 Ordinary Least Squares (OLS)

Ordinary least squares (OLS)[13], also known as Multiple Regression, is a type of linear least squares 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.

Assuming that in the case scenario the dependent and independent variables are linearly related, and impact of different variables are additive, the equation of a typical linear regression can be written as below:

$\hat{y} = \beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_n x_n$

or, expressed as a summation:

[13] Minimi Quadrati Ordinari in Italian translation

$$\hat{y} = \sum_{i=0}^{n} \beta_i x_i$$

OLS allows us to directly solve the $y = mx + b$ equation for the slope m and the intercept b in a context where there are many possible features of x, such as in a prediction under uncertainty, hence:

$y = x_1, x_2, x_3, \ldots, x_n$

Thus the linear regression function will translate to:

$b = \hat{y} - m\hat{x}$

$$m = \frac{\sum(x_i - \hat{x})(y_i - \hat{y})}{\sum(x_i - \hat{x})^2}$$

where:

- x are the independent variables.
- $\hat{x}$ is the average of the independent variables.
- y are the dependent variables.
- $\hat{y}$ is the average of dependent variables.

OLS is the default regression method for continuous dependent variables, but not the only one. Other methodologies are, in example:

- Quantile Regression.
- LAD (Least Absolute Deviation).
- GLS (Generalized Least Squares).
- SGD (Stochastic Grandient Descent) (with some caveat for linear regression).

11.2 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.

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).

r can be calculated with correlation formula such as Pearson's.[14]

[14] see: 6.2 Correlation

Correlation Coefficient r	Association Strength
+1.0	Perfect positive
+0.8 to 1.0	Very strong positive
+0.6 to 0.8	Strong positive
+0.4 to 0.6	Moderate positive
+0.2 to 0.4	Weak positive
0.0 to 0.2	Very weak or neutral
0.0 to -0.2	Very weak or neutral
-0.2 to -0.4	Weak negative
-0.4 to -0.6	Moderate negative
-0.6 to -0.8	Strong negative
-0.8 to -1.0	Very strong negative
-1.0	Perfect negative

11.3 Line Fitting, Residuals and Errors

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.

Residual have not to be confused with errors.

An **error** is the (often not observable) difference between the observed value and the true value (generated by the data generating process).

A **residual** is the difference between the observed value and the predicted value (by the model).

11.4 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.

11.5 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 MAE, MSE, RMSE, and R-Squared metric.

MAE - Mean Absolute Error
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).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

MSE - Mean Squared Error
MSE represent the average of the squared distances between the initial dataset and the prediction.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

RMSE - Root Mean Squared Error
RMSE is the square root of the MSE.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

R^2 - 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.

It's 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.

The capitalized R^2 notation suggests that it is referred to a multiple linear regression model (a model with more than one independent variable).

If, on the other hand, a simple linear regression model has been constructed, i.e., with only one independent variable, r^2 is usually preferred.

Coefficient of determination is a scale-free score, this mean that the value will always be within a range of 1 and 0[15].

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Usually, the higher the R^2 value (closer to 1), the higher the model has a predictive value. However, the significance of the R^2 ratio cannot be inferred by excluding the reference context. In some fields such as the behavioral sciences, it is usual to observe R^2 values of less than 50%. This does not mean that the regression model is not performing accurately, but that, by its very nature, the dependent variable being analyzed depends on so many different factors, many of which have not been measured. On the other hand, a high R^2 is a necessary but not sufficient condition for accurate predictions.

11.6 Chi-Square Test

Chi-square is among the most common nonparametric 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.

[15]There are cases where R^2 can yield negative values. This can arise when the predictions that are being compared to the corresponding outcomes have not been derived from a model-fitting procedure using those data

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.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where:

- χ^2 is the chi-square test statistic.
- O is the observed frequency.
- E is the expected frequency.

Results of the chi-square equation have to be checked against a chi-square table[16], in order to get the test results.

$\chi^2 >$ right tail probability $\implies H0 = FALSE$ (null hypothesis can be rejected).

11.7 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, F-test value can be calculated as a ratio of two variances.

$$F = \frac{\frac{SSB}{m-1}}{\frac{SSW}{m(n-1)}}$$

where:

- F is the F-test value.
- SSB is the variance of the group mean from the global mean (Grand Mean).
- SSW is the variance of each datapoint from the mean of each group.
- m is the number of clusters (groups).
- n is the number of items of each cluster.

[16] https://cdn.scribbr.com/wp-content/uploads/2022/05/Chi-square-table.pdf

The F test value has then to be compared on an F-table[17]. Note that there is an F-table for each value of α (for each level of confidence), and that both degrees of freedom (DF) (for clusters and for the sample) are needed in order to calculate the value.

[17] https://statisticsbyjim.com/wp-content/uploads/2022/02/F-table_Alpha10.png

12 License

Attribution 4.0 International (CC BY 4.0)[18]

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 diffusion 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 (and one of the goals of this work was to refer back to the canonical definitions, as further detailed in the following section) so where in good faith I have traced the work of licensed material I am happy to take action to amend it. Again starting from the premise that the purpose of this paper is not commercial but rather informative and educational. See the **Contact** section for any needs.

This work was originally released by the Author as a free downloadable pdf.

Machine-readable license metadata:
<arel="license"href="http://creativecommons.org/licenses/by/4.0/">
<imgalt="LicenzaCreativeCommons"style="border-width:0"src="https://i.
creativecommons.org/l/by/4.0/88x31.png"/>
<spanxmlns:dct="http://
purl.org/dc/terms/"href="http://purl.org/dc/dcmitype/Text"property="dct:
title"rel="dct:type">TheStatisticsHandbookdi<spanxmlns:cc="http://
creativecommons.org/ns#"property="cc:attributionName">CarloOcchiena
èdistribuitoconLicenza<arel="license"href="http://creativecommons.org/licenses/
by/4.0/">CreativeCommonsAttribuzione4.0Internazionale.

13 Source Code & Additional Materials

The source of this handbook, and some additional materials can be found on an open repository on my GitHub, at the following link:

- `https://github.com/carloocchiena/the_statistics_handbook`

The repository is including:

- LICENSE: the license definition for this handbook.
- README: the readme file.
- The Statistics Handbook - main.pdf: this handbook.
- Statistical_Workbook.xlsx: the .xlsx file with exercises and demonstrations.
- datachart_template.ai: the .ai file with the charts and image source file.

[18] `https://creativecommons.org/licenses/by/4.0/`

- main.tex: is the latex source code of the handbook.
- z-table.xlsx: the z-table generated with an Excel formula.
- z_score_table_t_score_analysis.ipynb: interactive notebook to explore z-score and t-score testing.

14 Bibliography & Sources

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

Some of them were institutional, such as University papers and websites. 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.

Books:

- Gambini A., Argomenti di statistica descrittiva.
- Perisco L, Di Bella E., Mosto L., Applicazioni di probabilità e statistica.
- C.Gosio, Matematica Finanziaria.
- A. Pascucci, W.J. Runggaldier Finanza Matematica.
- S. Fiorenzani, E. Edoli, S. Ravelli, The Handbook of Energy Trading.

Online sources:

- `http://www.stat.yale.edu/`
- `https://corporatefinanceinstitute.com/`
- `https://datascience.eu/it/matematica-e-statistica`
- `https://imstat.org/`
- `https://it.scienza.matematica.narkive.com/XNiiE9wo/probabilita-di-uscita-di-un-numero-su-una-ruota-del-lotto`
- `https://it.wikipedia.org/wiki/Statistica`
- `https://matematica.unibocconi.it/search/node/statistica`

- https://math.stackexchange.com/
- https://meetheskilled.com/test-di-ipotesi-one-sample-t-test/
- https://ocw.mit.edu/courses/1-151-probability-and-statistics-in-engineering-spring-2005/pages/lecture-notes/
- https://paolapozzolo.it
- https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html
- https://statisticsbyjim.com/hypothesis-testing/hypothesis-tests-significance-levels-alpha-p-values/
- https://statology.org
- https://thestatsgeek.com/
- https://www.britannica.com/browse/Mathematics
- https://www.datasciencecentral.com/
- https://www.datatechnotes.com/
- https://www.geo.fu-berlin.de/
- https://www.hwupgrade.it/forum/archive/index.php/t-1294440.html
- https://www.investopedia.com/math-and-statistics-4689831
- https://www.khanacademy.org/
- https://www.kristakingmath.com/
- https://www.matematicamente.it/forum/viewtopic.php?t=16622
- https://www.scribbr.com/statistics/type-i-and-type-ii-errors/
- https://www.statisticshowto.com/
- https://www.wallstreetmojo.com/t-distribution-formula/
- https://www.westga.edu/academics/research/vrc/assets/docs/confidence_intervals_notes.pdf
- https://www.youmath.it/

14.1 Images

All images used in this book were created by the author, either through a spreadsheet or a vector graphics software. In any case, the source files are included in the GitHub repository.

The only images drawn from the web are:

- The violin graph.

- the Gaussian with z-scores

These images appear to be part of the public domain; however, I have no problem removing them if that definition is incorrect.

14.2 Canonical Formulas and Definitions

As much as I wrote every single line of that manual in my own hand, I found it counterproductive if not misleading to paraphrase mathematical formulas and definitions.

I was therefore very careful to resort to canonical definitions and formulas.

This work was done by comparing different sources, especially comparing university papers with what could be found on the Web.

I think the final result is appreciable, nevertheless I consider it a possible point of discussion and improvement on which I gladly accept comparisons.

14.3 Datasets

A very good source of datasets ready to be used, and with a fair amount of metadata, can be found on Kaggle `https://www.kaggle.com/datasets`.

This has been my number one source of dataset for this handbook.

Other relevant sources are:

- `https://data.gov/`
- `https://archive.ics.uci.edu/ml/datasets.php`
- `https://datahub.io/collections`
- `https://datasetsearch.research.google.com/`

15 Acknowledgement

This work was inspired by all the people who taught me love for studying and sciences. I was welcomed into their circles without prejudice, although I often came across as an ant in their presence.

To all of them goes my deepest gratitude and appreciation!

16 Contacts

- Linkedin: `https://www.linkedin.com/in/carloocchiena/`
- Twitter: `https://twitter.com/carloocchiena`