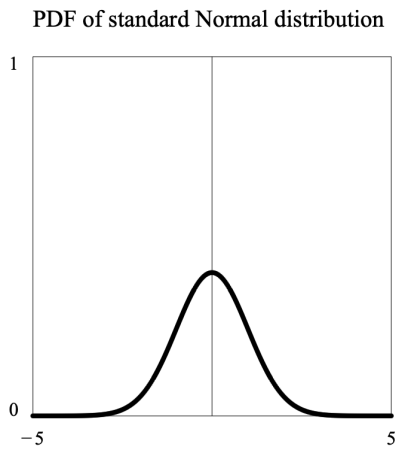
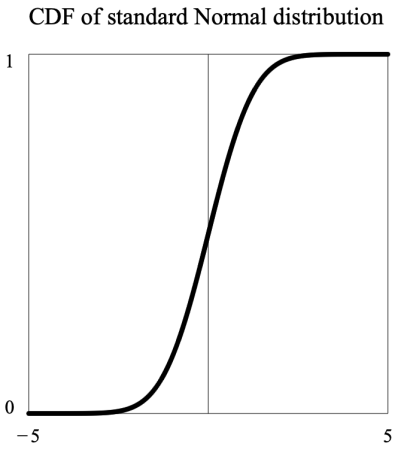
Exploring Math  with EIGENMATH

Statistics with EIGENMATH





Dr. Wolfgang Lindner
Dr.W.G.Lindner@gmail.com
Leichlingen, Germany
2026

Contents

0.1 Preface

This collection of small scripts show the use of the free CAS EIGENMATH to implement some well known concepts and routines of elementary statistics. We focus on comprehension of statistical concepts by direct translation of mathematical formulas, qualitative simple pictures and worked examples. This makes this 'handbook' and the accompanying EIGENMATH worksheets a perfect candidate for an learning environment at an 'action and process level' in the sense of the APOS Theory, cf. the short summary at [9].

The statistical concepts are presented in an CAS-language, that lies between the semiprofessional slang, in which mathematical concepts are presented and roughly explained and the high precision of the formal mathematical language, which is not so easy to grasp at a first attempt. The EIGENMATH language allows to make the concepts of statisticss executable, to coin math formulas and processes into 'run'able functions/procedures and therefore to allow own experiments. The results of one's thinking in the CAS language is immediately returned to the screen and helps to check the right understanding.

For using a script in the EIGENMATH language, no installation of any software is necessary, *everything runs directly online*: only run the script by loading the corresponding `html`-file into your browser and run the script by pressing the [Run] button below on the right: the calculation is made, allowing further free inputs from the user. Also, a click on a link in this text is enough to invoke the corresponding script - and by a click on the `RUN` button on the right bottom of the EIGENMATH frame □. the calculation is made.

If you own an APPLE iMac, there is the option to install the app EIGENMATH free of charge from the APPLE AppStore and run the scripts by opening the corresponding `txt`-file into the app's workspace and press [Run], cf. [18].

0.1.1 Why EIGENMATH?

EIGENMATH is a small but well designed and powerful computer algebra system (CAS), that can be used to solve problems in mathematics and the natural and engineering sciences. It is a personal resource for students, teachers and scientists. EIGENMATH is compact, capable and free.

EIGENMATH ...

- focuses straight to the point, no frills, no gimmicks,
- makes learning easy: approximately 50 pages manual [18], only about 100 commands
  ▷ INDEX - but you will use only a good dozen of them for your work - that's all!
- has an intuitive user playground (IDE) on the iMac with a two window frame,
- has an online version EIGENMATH^{online} ▷Demo, which runs in your browser,

- has mathematical oriented output in professional looking LaTeX printing,
- allows distraction-free experiences in Mathematics,
- allows to embed qualitativ rough plots to help visual understanding,
- is ideally suited for rapid prototyping.

The new versions ≥3.42 of EIGENMATH has new commands like enhanced `for` and `loop` commands for better control structures and new functions `erf()` resp. `tgamma()`, which are very useful in statistics. Using versions ≥ 3.50 of EIGENMATH we have also the new functions `incbeta()`, `fdist()`, `tdist()` and the undocumented function `tdistinv()`, which are of immense value for a compact programming of statistical distribution functions.

0.1.2 Why statistics with EIGENMATH?

Professional and university statisticians use free software like R, PYTHON, OCTAVE or free spreadsheets like LIBREOFFICE CALC etc. Besides possible problems of installation of the software by the novice user, it is often not easy to look into the source code of the relevant procedures, because they are cluttered into diverse packages or dependency structures and are not useable stand-alone. So, here are some reasons for the use of EIGENMATH ...

- allows to program the statistics formulas very close to their mathematical formulation,
- code is more compact compared to code in `C`, `R`, OCTAVE, MATHEMATICA ..
- each worksheet is totally independent of other imported code, it's self-contained,
- it's easy to include comments, tests, tables and qualitative simple plots in the worksheet,
- .. and it is just fun and illuminating to use EIGENMATH.

0.1.3 Some remarks on the content of this Manual

The order of the following suite of worksheets with definitions, checks, exercises (problems) etc. follows a standard presentation. Sometimes you have to enlarge the EIGENMATH window or scroll down inside it in order to view the whole calculation or to make further EIGENMATH calculations.

1 First we define the standard functions of descriptive statistics, i.e. `mean`, `var`, `std`, `sem`, `mad`, `rms`, `median`, `mode`, `madM`, `quantile`, `moment`, `skew`, `kurtosis`. Ergo, the user should be able to easily follow the calculations in standard texts like [13].

2 We implement some important statistical distributions as helper functions to enable the statistical test in chapter 4. and 5. For each distribution presented in chapter 2. and chapter 3. we give

- the definition resp. coding in EIGENMATH notation for
  - the probability density functions f named "...**PD**F",
  - the cumulative distribution functions F named "....**CDF**" and
  - the quantile functions F^{-1} named "....**INV**", i.e. the INVerse of the CDF.
- a check of the calculations against the statistics software R,
- a short table of the distributions values,
- a qualitative plot to have a visual impression or to verify one's result on a graph,
- a solution of a prototypical problem resp. application,
- a link to an applett on BOGNAR's homepage for further information ('help').

To invoke an EIGENMATH sheet click on a link starting with "▷ ...".

3 Last we implement some standard functions of test statistics, i.e. *one-parameter Gauss test*, *two-parameter Gauss test*, *one-sample t-test*, *two-sample t-test*, *F-test*, *ChiSquare test* and *Z test*. The user should easily follow these calculations in standard texts like [13] or in free spreadsheets. All these calculations are demonstrated and recalculated as examples using EIGENMATH functions.

4 There is also a collection of all implemented functions in the file ▷`statsBox.txt` to facilitate the use of the statistics functions in EIGENMATH. You load this library in your EIGENMATH sheet with the command `run()`, cf. run

0.1.4 Some remarks on the didactical contcept of this Manual

This small booklet is dedicated to the novice – w.r.t. the subject matter as well w.r.t. the use of CAS. Therefore, I roughly follow the pedagogical APOS theory, cf. [9]: this means each section of this book is divided into the following 4 steps:

M: we start with a short *motivation* of the mathematical concept using colloquial *words*.[1]

V: we present an adequate concrete *visualization* of the mathematical concept using a prototypical example. The user should pause and think and reflect a while about it.

[1]Some of these sentences are produced using Google's AI answer. Therefore, I would appreciate any information regarding the original source of the citation and would add the appropriate credit to the quote.

D: we give a precise mathematical *definition* using mathematical *symbols*. This definition should be memorized along with the visualization and the prototypical example.

E: we solve a concrete *example* often w.r.t. part $\mathbf{V}^{\cdot}$ and prepare for the use of the CAS EIGENMATH. The reader will fully understand the mathematical concept in question when he can translate the mathematical definition $\mathbf{D}^{\cdot}$ into a working code snippet. This simultaneously enables to conduct own experiments with own data using the mathematical concept, allowing to *observe the effects of our own actions* in the CAS EIGENMATH.

Remark. The presented code in this book and in the appendices is 'pedagogical' code, aimed at the novice learner who is willing to learn statistics and a CAS like EIGENMATH simultaneously. Therefore there is *no guaranty for correct results using the presented statistical functions*. Nevertheless, our code is checked with the professional industry strength code from MATLAB (online) or the academical strength code from R.

Any feedback from the user is very welcome.

I want to thank George WEIGT for the development of the free CAS EIGENMATH and EIGENMATH^{online} over nearly 20 years and for his friendly support with tips and hints while writing these notes.

Wolfgang Lindner
Leichlingen, Germany
January 2026

1 Desriptive Statistics

We collect the well known measures of tendency, dispersion and deviation. We translate the mathematical definitions into EIGENMATH code and demonstrate exemplary calls and examples. We define and code the standard functions of descriptive statistics, i.e. `mean`, `var`, `std`, `sem`, `mad`, `rms`, `median`, `mode`, `madM`, `quantile`, `moment`, `skew`, `kurtosis`. Ergo the user should easily follow the calculations in standard texts like [13]. All these calculations could now simply be done with EIGENMATH.

For independent work, there is a text file containing all relevant EIGENMATH definitions of all chapters, cf. ▷`statsBox.txt` or an html-file, which can be looked at ▷ EIGENMATH : `StatsBox.html`.

1.1 `mean` - the arithmetic Mean

The arithmetic mean $\bar{X}$ of a list X of numbers is the sum of all of the numbers in the list divided by their count.

Definition For a vector (list) $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$ of numbers the arithmetic mean $\bar{X}$ or `mean(X)` (or simply *mean* or *average*) is defined by

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n} =: mean(X)$$

Examples [wiki] The arithmetic mean of the five values: 4, 36, 45, 50, 75 is:

$$\bar{X} = \frac{4 + 36 + 45 + 50 + 75}{5} = \frac{210}{5} = 42$$

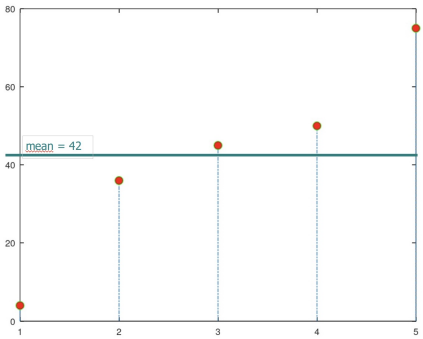
If the five values are interpreted as hourly earnings in of five employees, then the arithmetic mean corresponds to the hourly earning, that everyone would receive if the total earnings were distributed equally among all employees.

Check with EIGENMATH :

| `mean((4,36,45,50,75))` | 42

∘ EIGENMATH definition and examples are in the notebook ▷ EIGENMATH : `mean`

Mental image



To reproduce the figure above with octave/Matlab do

```
# OCTAVE
X =[1,2,3,4,5];  Y = [4,36,45,50,75];
stem(X,Y,'LineStyle','-.', 'MarkerFaceColor','red', 'MarkerEdgeColor','green')
```

General information

General mathematical information about the concept is here ▷ WIKIPEDIA : Mean
Syntax and semantic of the function is here ▷ MATLAB : `mean`

1.2 `var` - the variance

The *variance* is the mean squared deviation from the mean. Variance is a measure of how far the observed values x_i in a dataset X fall from the arithmetic mean $\mu := \bar{X}$ and is therefore a measure of spread.

Definition

• For a vector $X := (x_1, x_2, ..., x_n)$, the (population) *variance* is defined as

$$var(X) := \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \mu) = \frac{(x_1 - \mu) + (x_2 - \mu) + \cdots + (x_n - \mu)}{n}$$

where μ is the mean of X.

• The (sample) *variance* is defined as $\text{var0}(X) := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)$.

Examples The (sample!) variance of the five values: 4, 36, 45, 50, 75 is:

$$\text{var}(X) = \frac{(4-42)^2 + (36-42)^2 + (45-42)^2 + (50-42)^2 + (75-42)^2}{5} = 528.4$$

Check with EIGENMATH :

```
| X=(4,36,45,50,75)
| var0(X)                | 660.5
```

◦ EIGENMATH definition and examples are in the notebook ▷ `var`

Mental image

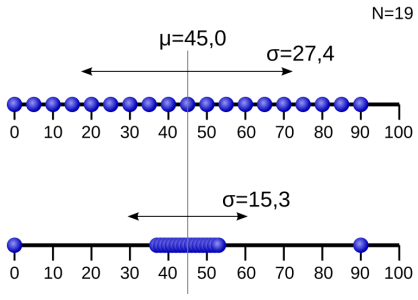


Figure 1: Figure illustrating the concept of variance for two different sets of 19 numbers (0, 5, ..., 90) and (0, 37, 38, ... 53, 90). μ denotes the mean and σ denotes the square root of the variance as a measure of the spread.

General information

General mathematical information about the concept is here ▷ WIKIPEDIA : Variance
Syntax and semantic of the function is here ▷ MATLAB : `var`

1.3 `sd` - the Standard deviation

The *standard deviation* is the square root of the mean squared deviation from the mean. A large standard deviation indicates that the data points x_i in a dataset X can spread far from the mean $\mu := \bar{X}$ and is therefore a measure of spread.

Definition

• For a vector $X := (x_1, x_2, ..., x_n)$, the (population) *standard deviation* is defined as

$$sd(X) := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

where μ is the mean of X.

• The (sample) *standard deviation* is defined as $\mathrm{sd}(X) := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$.

Mental image

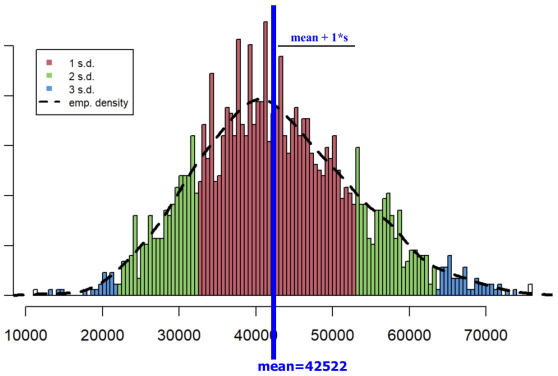


Figure 2: Table of annual salary of employees, $\mu = mean, s = sd(X)$: ▷ var
⊓: 68 % of the data fall inside the red region $\mu \pm s$.
⊓: 95 % of the data fall inside the red+green region $\mu \pm 2 * s$.
⊓: 99.7 % of the data fall inside the red+green+blue region $\mu \pm 3 * s$.
⌒: the empirical density function

Example The (population) standard deviation of the five values: 4, 36, 45, 50, 75 is:

$$\mathrm{sd}(X) = \sqrt{\mathrm{var}(X)} = \sqrt{528.4} = 22.98$$

∘ EIGENMATH definition and examples are in the notebook ▷EIGENMATH : `sd`

∘ General mathematical **Information** about the concept is ▷ WIKIPEDIA : std.deviation
∘ Syntax and semantic of the function is in ▷ MATLAB : `std`

1.4 `sem` - the Standard error of Mean

For a given sample X, the standard error of the mean $sem(X)$ equals the standard deviation divided by the square root of the sample size $dim(X)$.
In other words, the *standard error of the mean* is a measure of the dispersion of sample means around the population mean. The SEM is a measure of how much a sample mean is likely to differ from the true population mean, see example 2. This concept is heavily used in hypothesis testing and calculating confidence intervals..

Definition
For a vector $X = (x_1, x_2, ..., x_n)$, the *standard error of the mean* is defined as

$$sem(X) := \frac{sd(X)}{\sqrt{n}}$$

where $sd(X)$ is the standard deviation of the population

Examples 1. The standard error of the mean of the five values $X = (4, 36, 45, 50, 75)$ is $sem(X) = 11.49$
2. In the table below, a measurement of $N = 10$ resistance values of a production of resistors with a $\mu_0 = 100\ \Omega$ 'guaranteed' value is shown. A reasonable assumption is that the production is normally distributed. Let us assume that the spread of the production is well known with sigma $\sigma = 0.5\,\Omega$. To check whether the target value $100\ \Omega$ is still being met by the production, one has to determine first of all the test statistic $T = \frac{(\overline{X}-\mu 0)}{sem(X)}$.
Calculate T for the following sample of resistors.[2] *Solution* at ▷ EIGENMATH
○ EIGENMATH definition and examples are in the notebook ▷ `sem`

Sample of a production of 100 resistors.

n :	1	2	3	4	5	6	7	8	9	10
Ω :	100.1	101.2	99.5	99.0	100.7	100.0	101.2	99.2	99.0	98.7

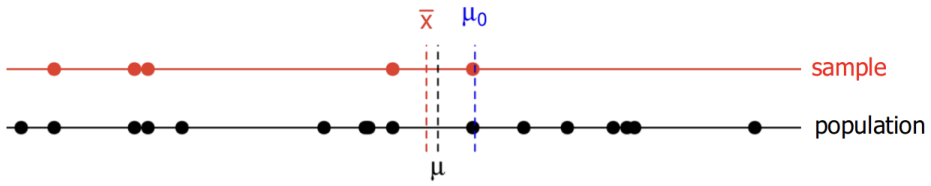


Figure 3: Visualization of sample vs. population and there resp. means $\overline{X}$ vs. μ. μ_0 is the theoretical mean, i.e. the target value.

○ Mathematical `Information` about the concept is ▷ WIKIPEDIA : std.error of mean
○ Syntax and semantic of the function is ▷ MATLAB :– no function provided –

[2]This example concerning resistors quality is in BEUCHER [1, p. 216]. See chapter §4.1 of this script.

1.5 mad - the average absolute deviation

The average absolute deviation *mad* of a data set is the average of the absolute deviations from a central point. It is a means of statistical dispersion or variability. The central point can be e.g. mean, *median*, mode.

Definition

For a list $X = (x_1, x_2, \ldots, x_n)$, the *mean/median/.. absolute deviation* `mad` is defined as

$$mad(X) := \frac{1}{n} \sum_{i=1}^{n} |x_i - m(X)|$$

where $n = length(X)$ and $m \in \{mean, median, mode, ...\}$.

Examples

The *mad* w.r.t. the *mean* of the five values 4, 36, 45, 50, 75 is 17.6.
The *mad* w.r.t. the *median* of the five values 4, 36, 45, 50, 75 is 9.

◦ Check with EIGENMATH :

| `mad( (4,36,45,50,75) )` | 17.6

◦ EIGENMATH definition and examples are in the notebook ▷ `mad`

Mental image

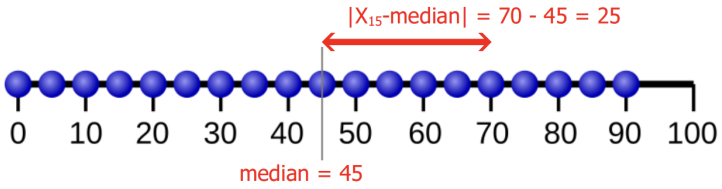


Figure 4: Figure illustrating the concept of *mad* for the data set of the 19 numbers $X = (0, 5, 10, ..., 90)$.

General information

General mathematical information about the concept is ▷ WIKIPEDIA : average abs. dev.
Syntax and semantic of the function is here ▷MATLAB : `mad`

1.6 rms - the Root mean square

The *root mean square* (abbrev. *rms*) of a set of values is the square root of the set's mean square.

Definition
For a vector $X = (x_1, x_2, \ldots, x_n)$, the *rms* is defined as

$$rms(X) := \sqrt{\frac{1}{n}\left({x_1}^2 + {x_2}^2 + \cdots + {x_n}^2\right)}$$

where $n = length(X)$.

Examples

1. The *rms* of the five values 4, 36, 45, 50, 75 is 47.88.

◦ Check with EIGENMATH :

```
X = (4,36,45,50,75)
rms(X)                    | 47.879
```

2. The *rms* of $X = (0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90)$ is 52.68.
◦ EIGENMATH definition and examples are in the notebook ▷ `rms`

Mental image

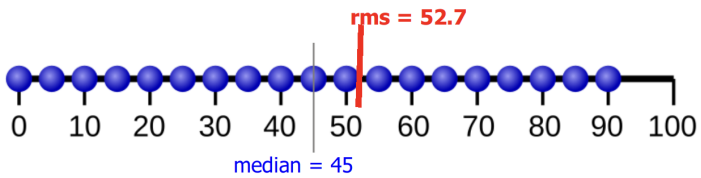


Figure 5: Figure illustrating the concept of *rms* for the data set of the 19 numbers $X = (0, 5, 10, ..., 90)$.

General information

Mathematical information about the concept is ▷ WIKIPEDIA : Root mean square
Syntax and semantic of the function is here ▷ MATLAB : `rms`

1.7 `median` - the Median

The *median* is the central value of an *ordered* data set and divides it into a part with small and with large data points. Therefore, the median is a suitable measure for determining the distribution of a data set. The median minimizes the total distance to all other data elements. It is the solution to the optimization problem $min_{a\in\mathbb{R}} \sum_{\nu=1}^{n} |x_\nu - a|$.
If a data element change, this only causes a change in the median, if the older value moves from one half of the ordered data set to the other half.

Definition
For a *sorted* vector $X = (x_1, x_2, \ldots, x_n)$, the *median* $\tilde{X}$ =`median(X)` is defined as

$$\tilde{X} := \begin{cases} x_{(n+1)/2} & \text{if } n = dim(X) \text{ is odd} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & \text{if } n = dim(X) \text{ is even} \end{cases}$$

where n =length(X).

Examples
The median of the five values 4, 36, **45**, 50, 75 is 45. $n = dim(X) = 5$.
The median of $X = (36, 4, 75, 45, 50) \xrightarrow{sort} (4, 36, \mathbf{45}, 50, 75) \xrightarrow{\sim} 45 = \text{median}(\text{sort}(X))$.

Check with EIGENMATH :

| `median((36,4,75,45,50))` | 45

∘ EIGENMATH definition and examples are in the notebook ▷ EIGENMATH : `median`

Mental image

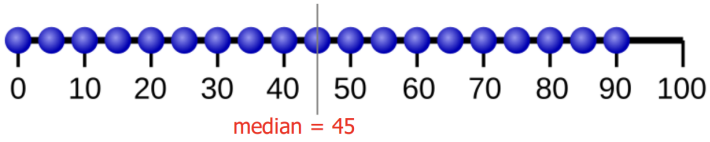


Figure 6: Figure illustrating the concept of *median* for the data set of the 19 numbers $X = (0, 5, 10, ..., 90)$.

General information
General mathematical information about the concept is ▷ WIKIPEDIA : Median
Syntax and semantic of the function is here ▷ MATLAB : `median`

1.8 `mode` - the Mode

The *mode* is the value $\overset{\approx}{X}$ in a data set X that appears most often, i.e. the maximal value in the frequency distribution of X.

Definition
For a vector $X := (x_1, x_2, ..., x_n)$, the *mode* is defined as

$$mode(X) := max \text{"freq"}(X)$$

where *freq* is the frequency table of X (not defined here).

Examples

1. The *mode* of the five values 4, 36, 45, 50, 75 is 4.

2. $freq(1,3,6,6,6,6,7,7,12,12,17) = \begin{pmatrix} 1 & 3 & \mathbf{6} & 7 & 12 & 17 \\ 1 & 1 & \mathbf{4} & 2 & 2 & 1 \end{pmatrix} \Rightarrow mode((1,3,6,6,6,6,7,7,...,17)) = \mathbf{6}$

3. $mode((1,1,2,4,4)) = 1$

Check with EIGENMATH :

| `mode((1,1,2,4,4))` | 1

○ EIGENMATH definition and examples are in the notebook ▷ `mode`

Mental image

Figure 7: Figure illustrating the concept of *mode* for a sets of 9 numbers (37, 38, 38, 40, 42, 42, 42, 45, 46). The most sold length of the 9 shoes was 42.

General information

General mathematical information about the concept is ▷ WIKIPEDIA : Mode
Syntax and semantic of the function is here ▷ MATLAB : `mode`

1.9 `quantile` - the Quantile

A *quantile* is a score at or below which a given percentage of the all scores exists, i.e., a score in the k-th percentile would be above approximately k % of all scores in its set.
[We quote ▷QUANTILE] One definition of percentile or *quantile* Q_p is that the p-th percentile of a list X of n *ordered* values (sorted from least to greatest) is the smallest value in the list such that no more than p percent of the data is strictly less than that value and at least p percent of the data is less than or equal to that value. This is obtained by first calculating the ordinal rank and then taking the value from the ordered list that corresponds to that rank. The ordinal rank N is calculated using the formula $N = \lceil p * 0.01 * n \rceil$. [3]

Definition (The *nearest-rank method* for a quantile)
For a vector $X := (x_1, x_2, ..., x_n)$, the *quantile* Q_p is defined as

$$Q_p(X, p) := \begin{cases} max(X) & \text{if } p = 100 \\ X_{\lceil p*0.01*n \rceil} & \text{if } p \in [0, 100[\end{cases}$$

where $n = dim(X)$ is the length of X.

Mental image

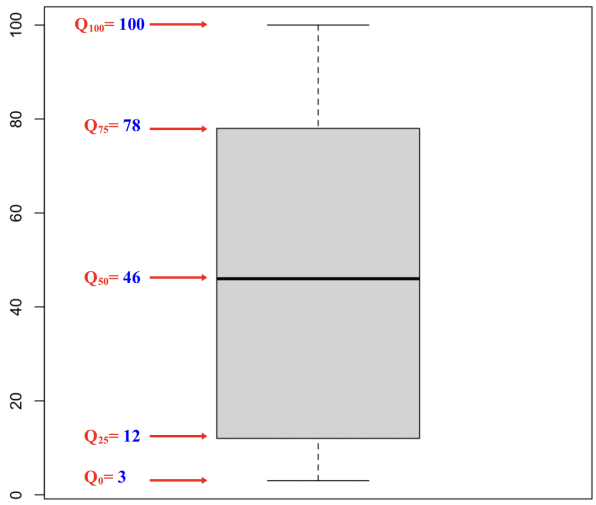


Figure 8: Figure illustrating the concept of *quantile* for the sets of numbers $(12, 3, 4, 56, 78, 18, 46, 78, 100)$: $\begin{pmatrix} 0\% & 25\% & 50\% & 75\% & 1000\% \\ 3 & 12 & 46 & 78 & 100 \end{pmatrix}$, visualizied in a so-called *boxplot*. $Q_{50} = 46$ is the $median(X)$.

[3] `https://en.wikipedia.org/wiki/Floor_and_ceiling_functions`, the ceiling function maps x to the least integer greater than or equal to x, denoted $\lceil x \rceil$ or $ceil(x)$.

Examples

The quantile values for the ordered list $X = (12, 3, 4, 56, 78, 18, 46, 78, 100)$ are $Q_0 = 3, Q_{25} = 12, Q_{50} = 46, Q_{75} = 78, Q_{100} = 100$.

Check with EIGENMATH :

```
X = (12,3,4,56,78,18,46,78,100)
boxplot(X)
```

0%	25%	50%	75%	100%
3	12	46	78	100

∘ EIGENMATH definition and examples are in the notebook ▷`quantile`

Remark.
∘ Using the nearest-rank method on lists with fewer than 100 distinct values can result in the same value being used for more than one percentile.
∘ A percentile calculated using the nearest-rank method will always be a member of the original ordered list. This is not always the case using R or MATLAB/OCTAVE, which uses advanced sophisticated algorithms.
∘ The 100th percentile is defined to be the largest value in the ordered list, the 0th percentile to be the smallest value.

General information

General mathematical information about the concept is ▷ WIKIPEDIA : quantile
Syntax and semantic of the function is here ▷ MATLAB : `prctile`

1.10 `moment` - the rst Moment

In statistics, moments are parameters that describe the shape of a probability distribution by measuring different aspects, such as its central tendency, spread, and asymmetry. The most common moments are the first moment (mean), which indicates the center; the second moment (variance), which measures spread; skewness, the third moment, which shows asymmetry; and kurtosis, the fourth moment, which describes the peakedness or flatness of the distribution. Cf. ▷ *Google search: statistics moments. AI overview.*
The *moment* concept generalizes the concepts of variance, skewness and kurtosis. The r-th raw moment of a population can be estimated using the r-th raw sample moment, applied to a sample $x1, ..., x_n$ drawn from the population. The *(central) moment of order r about A*, `moment(X,A,r)` computes a sample version of a population value. The first-order central moment is zero, the second-order central moment is the variance computed using a divisor of n rather than $n-1$, where n is the length of the data vector X.

Definition

For a data vector $X := (x_1, x_2, ..., x_n)$, the *r-st moment about A* is defined as

$$moment(X, A, r) := \frac{1}{n} \sum_{i=1}^{n} (x_i - A)^r$$

where n is the length of X.

Mental image

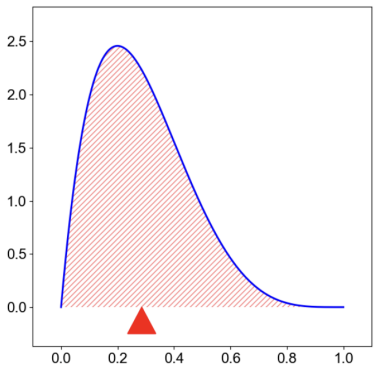


Figure 9: Imagine the data distributed under the blue graph. Then the moment may be interpreted as 'center of mass', balanced upon the 'expected' value △, the moment e.g. the mean; Cf. GUNDERSEN

Examples

The 1st *moment* of the five values 4, 36, 45, 50, 75 is 45, i.e. its mean.
The 2nd moment of the five values 4, 36, 45, 50, 75 is 2292.4.

The 3rd moment about $A = 4$ of the five values 4, 36, 45, 50, 75 is 111387.

Check with EIGENMATH :

```
X = (4, 36, 45, 50, 75)
moment0(X,1)                    | 42
moment0(X,2)                    | 2292.4
moment0(X-4,3)                  | 111387
```

∘ EIGENMATH definition and examples are in the notebook ▷ `moment`

Remark.
When `center` and `absolute` are both FALSE in function `moment`() of R, the moment is simply

$$sum(X^{order})/length(x)$$

General information

General mathematical information about the concept is ▷ WIKIPEDIA : Moment
Syntax and semantic of the function is here ▷ MATLAB : `moment`

1.11 `skew` - the Skew.ness

Skewness is a *measure of the asymmetry* of the data around the sample mean. If skewness is negative, the data spreads out more to the left of the mean than to the right. If skewness is positive, the data spreads out more to the right.
The skewness of the normal distribution (or any perfectly symmetric distribution) is zero. The *relative position of arithmetic mean and the median* to each other is also characterized by the skewness of a data set. If the data set is *skewed right* (steep to the left), the arithmetic mean lies to the *right* of the median. If the data set is skewed left (steep to the right), the arithmetic mean lies to the left of the median. If the data set is symmetrical, then the arithmetic mean and median are approximately the same.

Definition

For a data vector $X := (x_1, x_2, ..., x_n)$, the *skewness* `skew(X)` is defined as

$$\begin{aligned} skew(X) &:= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{X}}{s}\right)^3 = \frac{m_3}{s^3} \\ skewness(X) &:= \frac{\sum_{i=1}^{n}(x_i - \overline{X})^3}{(n-1)s^3} \end{aligned}$$

where n is the number of observations, m_3 ist the 3rd moment and $s = sd(X)$ the standard deviation and $\overline{X}$ is the mean of X. – The $1st$ formula is used by MatLab, the $2nd$ by R.

Mental image

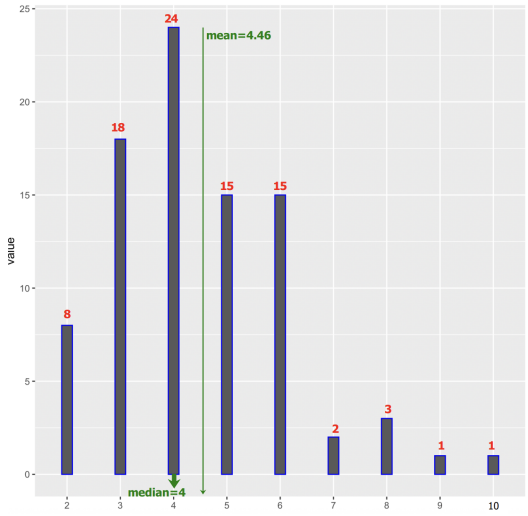


Figure illustrating the concept of *skew*ness for a distribution of numbers.
Figure 10: The frequency distribution has positive skewness of 0.808 (skew right). The *mean*= 4.46 is placed *right* of the *median*= 4.

Examples

The *skewness* of the five values $4, 36, 45, 50, 75$ is -0.306 resp. -0.2741.

Check with EIGENMATH :

`X = (4, 36, 45, 50, 75)`		
`skew(X)`	-0.306	[MatLAB]
`skewR(X)`	-0.274	[R]
`skew1(X)`	-0.219	[R, type=3]
`skew2(X)`	-0.456	[R, type=2]

◦ EIGENMATH definition and examples are in the notebook ▷ `skew`

General information

General mathematical information about the concept is ▷ WIKIPEDIA : Skewness
Syntax and semantic of the function is ▷MATLAB : `Skewness`

1.12 `kurtosis` - the Kurtosis

The kurtosis, the fourth moment, describes the *peakedness or flatness* of the distribution. Cf. ▷ *Google search: statistics moments. AI overview.*
[MatLab:] Kurtosis is a measure of how outlier-prone a distribution is. *The kurtosis of the normal distribution is 3.* Distributions that are more outlier-prone than the normal distribution have kurtosis greater than 3; distributions that are less outlier-prone have kurtosis less than 3. Some definitions of kurtosis subtract 3 from the computed value (measure only the 'excess'), so that the normal distribution has kurtosis of 0.
∘ Our kurtosis function `kurtosis(X)` does not use this convention.

Definition

For a data vector $X := (x_1, x_2, ..., x_n)$, the *kurtosis* `kurtosis(X)` is defined as

$$\begin{aligned} kurtosis(X) &:= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2\right]^2} = \frac{m_4}{m_2^2} \\ kurtosisM(X) &:= \frac{\sum_{i=1}^{n}(x_i - \bar{X})^4}{(n-1)s^4} \end{aligned}$$

where n is the length of X and $\bar{X}$ is the mean of X , $s = sd(X)$ the standard deviation and m_r is the rst moment of X. – The $1st$ formula is used by R, the $2nd$ by MatLAB.

Mental image

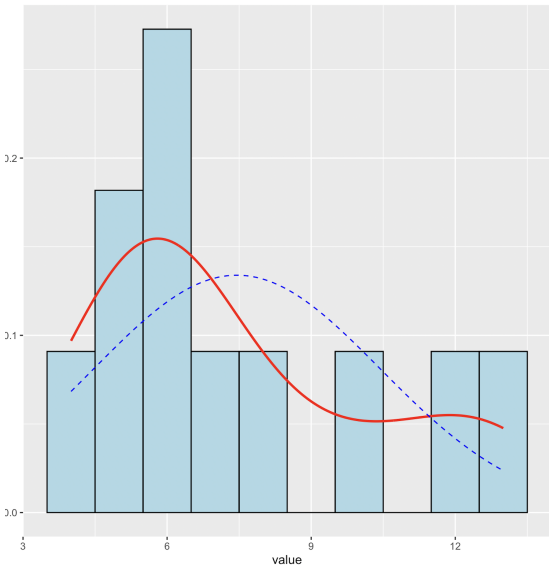


Figure illustrating the concept of *kurtosis* for a set of numbers.
Figure 11: The normal distribution is noted in '---', the distribution of the number set in '---'. The kurtosis is 2.295 resp.2.226. See example.2.

Examples

1. The *kurtosis* of the five values $4, 36, 45, 50, 75$ is 1.9129 resp. 1.7390.
2. The *kurtosis* of the sample

```
X=(4,4,4,5,5,5,5,5,5,6,6,6,6,6,6,6,6,6,7,7,7,8,8,8,10,10,10,12,12,12,13,13,13)
```

with the frequency distribution $\binom{3\,6\,9\,3\,3\,3\,3\,3}{4\,5\,6\,7\,8\,10\,12\,13}$, e.g. value 5 is 6 times, shown in figure.11 is 2.295 resp. 2.226.

Check with EIGENMATH :

```
X =(4,4,4,5,5,5,5,5,5,6,6,6,6,6,6,6,6,6,
    7,7,7,8,8,8,10,10,10,12,12,12,13,13,13)
kurtosis(X)                          | 2.2958    [R; library(moments)]
```

∘ EIGENMATH definition and examples are in the notebook ▷`kurtosis`

General information

General mathematical information about the concept is ▷ WIKIPEDIA : Kurtosis
Syntax and semantic of the function is ▷MATLAB : `kurtosis`

1.13 `cov` - the Covariance

Covariance is a statistical measure of the joint variability of two random variables, indicating how much they change together. A positive covariance signifies that variables tend to move in the same direction, a negative covariance indicates they move in opposite directions, and a zero covariance suggests no linear relationship.[4]
There are different methods used to compute covariance: the *population* `cov` dividing the sum of the distance products $(x_i - \bar{X}) \cdot (y_i - \bar{Y})$ by n vs. the *sample* `cov` dividing by $n-1$.

Definition

• For two data vectors $X := (x_1, x_2, ..., x_n)$ and $Y := (y_1, y_2, ..., y_n)$, the (population) *covariance* `cov(X,Y)` *between* X *and* Y is defined as

$$\text{cov}(X,Y) := \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{X}) \cdot (y_i - \bar{Y})$$

where n is the length of X and Y and $\bar{X}$ resp. $\bar{Y}$ is the mean of X resp. Y.
• The *sample* covariance is defined as $covR(X,Y) := \frac{1}{n-1} \cdot \ldots$, e.g. in R: typ "pearson".

Mental image

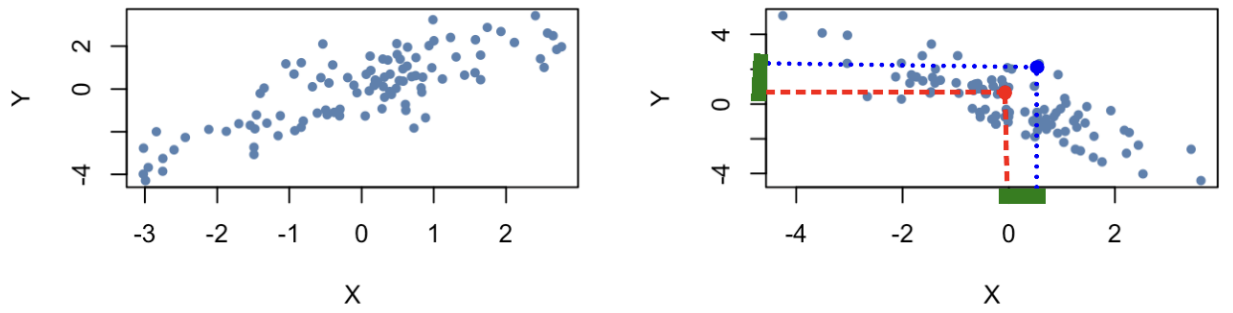


Figure 12: Figure illustrating the concept of *covariance* for a set of numbers. Left: positive `cov`. Right: negative `cov`. • $= (\bar{X}, \bar{Y})$. • $= (x_i, y_i)$ a data point. ⋈ × ⋈ $= (x_i - \bar{X}) \cdot (y_i - \bar{Y})$ a distance product w.r.t. the mean.

Examples

The covariance of the two data sets $1, 3, 5, 10$ and $2, 4, 6, 20$ is 23 resp. 30.66 (R: method 'pearson').

[4] cf. ▷ *Google search: covariance.*

Check with EIGENMATH :

`X = (1,3,5,10)`	
`Y = (2,4,6,20)`	
`cov(X,Y)`	23
`covR(X,Y)`	30.66

◦ EIGENMATH definition and examples are in the notebook ▷EIGENMATH : `cov`

General information

General mathematical information about the concept is ▷ WIKIPEDIA : Covariance
Syntax and semantic of the function is ▷MATLAB : `cov`

2 Discrete distributions

We implement some discrete resp. continuous statistical distributions as helper functions to do the statistical tests in chapter 4 and 5. For each distribution presented in chapter 2 and chapter 3 we give the definition resp. coding in EIGENMATH notation for the probability density function f (named "...**PDF**"), the cumulative distribution function F (named "....**CDF**") and the quantile function F^{-1} (named "....**INV**") i.e. the INVerse of the CDF.

2.1 Binomial distribution

In probability theory and statistics, the binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes-no question, and each with its own Boolean-valued outcome: success or failure. The binomial distribution is used when there are exactly two mutually exclusive outcomes of a trial. Cf. WIKIPEDIA: *Binomial_distribution*

Definition Notation: $X \sim Binomial(n, p)$

- The *probability density* function for the Binomial distribution is defined as:

  $\texttt{binoPDF}(\texttt{k}, \texttt{n}, \texttt{p}) := f(k, n, p) := \Pr(X = k)^5 = \binom{n}{k} p^k (1-p)^{n-k}$

- The *cumulative* binomial distribution function can be expressed as:

  $\texttt{binoCDF}(\texttt{k}, \texttt{n}, \texttt{p}) := F(k; n, p) = \Pr(X \leq k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$

- The *quantile* function (inverse cumulative distribution function) for $Bin(n, p)$ is[6]

  $\texttt{binoINV}(\alpha, n, p) := \inf\{k \in \mathbb{R} : \alpha \leq F(k; n, p)\}$

  i.e. we must find the *smallest* k such that $\alpha \leq \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$ for given α.

Examples

1. *binoPDF*: Suppose a biased coin comes up heads with probability 0.3 when tossed. What is the probability of seeing exactly 4 heads in 6 tosses?

   | `binoPDF(4, 6, 0.3)` | 0.059535

[5] Pr(X=k): represents the probability of getting exactly k successes
[6] there is no closed term for the inverse binomial distribution

2. *binoCDF*: The cumulative probability of a binomial distribution with 10 trials and a probability of success 0.5 for 4 successes is 0.3769.

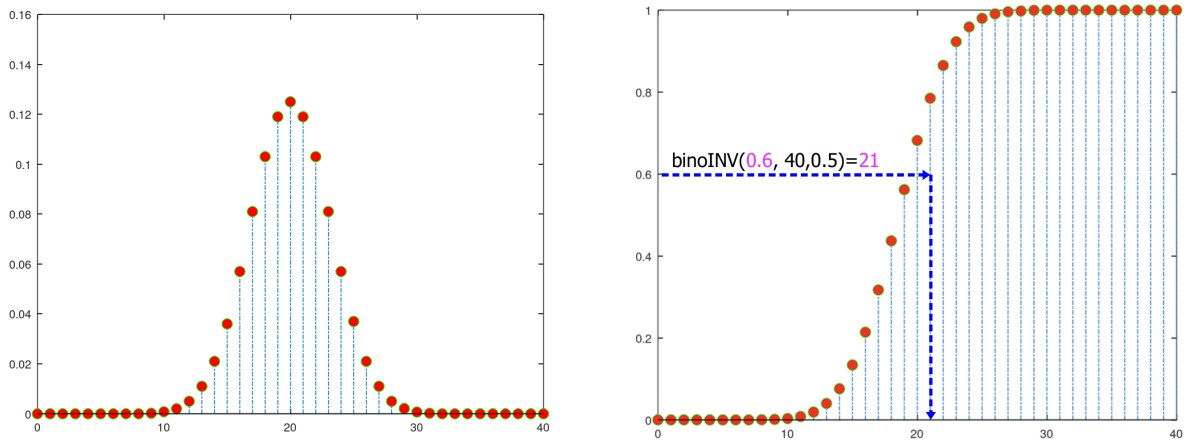
   | `binoCDF(4, 10, 0.5)` | 0.3769

3. *binoINV*: Given a number of trials n=100, the probability of success p=0.3, the cumulative area of the binomial distribution α=0.7, find the first value x such that $\alpha = 0.7 \leq$ binoCDF(x,100,0.3). Result: 32

   | `binoINV(0.7, 100, 0.3)` | 32

∘ EIGENMATH definition and examples are in the notebook ▷EIGENMATH `binomial distr.`

Graphical representation



Left figure: • = plot for `binoPDF(k, 40.0.5)` for $k = 0, \ldots, 40$.
Check `binoPDF`$(20, 40, 0.5) = 0.125$ on the graph.
Figure 13: Right figure: • = plot for `binoCDF(k, 40.0.5)` for $k = 0, \ldots, 40$.
Check `binoCDF`$(20, 40, 0.5) = 0.563$ on graph.
Check `binoINV`$(0.6, 40, 0.5) = 21$ on graph via $y = 0.6 \longrightarrow\downarrow 21 = k$

Remark. A representation of `binoCDF(k,n,p)` in terms of the *incomplete beta function* $I_x(a, b)$ is

$$\begin{aligned} \texttt{binoCDF}(\texttt{k}, \texttt{n}, \texttt{p}) &= I_{1-p}(n-k, k+1) \\ &= (n-k)\binom{n}{k}\int_0^{1-p} t^{n-k-1}(1-t)^k\, dt, \end{aligned}$$

We use it in our accompanying EIGENMATH notebook to plot the *continuous* (!) graph of `binoCDF(.)`, look at figure.13, right plot.

∘ General mathematical information about the concept is ▷ WIKIPEDIA : Binomial distr.
∘ Syntax and semantic of the function is here ▷ MATLAB : `binopdf`
[†] *Bognar's app*

2.2 Geometric distribution

A geometric distribution is a discrete probability distribution that describes the chances of achieving success in a series of independent trials, each having two possible outcomes. The geometric distribution thus helps measure the probability of success after a given number of trials. In the binomial distribution, the number of trials is fixed, and we count the number of "successes". Whereas, in the geometric distribution, the number of "successes" is fixed, and we count the number of trials needed to obtain the desired number of "successes".
The geometric distribution is a discrete analog of the exponential distribution.
It is discrete, i.e. existing only on the nonnegative integers.

Definition Notation: $X \sim Geometric(n, p)$

- The probability density function of a geometric distribution is defined as:

$$\texttt{geoPDF}(\texttt{k},\texttt{p}) := \Pr(X = k) = p(1-p)^k, \quad k = 0, 1, 2, 3, \ldots \text{ and } 0 < p < 1$$

  where k is number of failures before the first success and p is the probability of success on a given trial.

- The cumulative geometric distribution can be expressed as:

$$\texttt{geoCDF}(\texttt{n},\texttt{p}) := \Pr(X \leq k) = 1 - (1-p)^n, \quad n = 1, 2, 3, ...$$

- The *quantile* function (inverse cumulative geometric distribution) is[7]

$$\texttt{geoINV}(u, p) := \lceil \frac{log(1-u)}{log(1-p) - 1} \rceil$$

  i.e. we must find the *smallest* n such that $u \leq 1 - (1-p)^n$ for given u.

Examples from MatLAB solved with Eigenmath.

▷ *geoPDF*: A man asking for help with a probability of getting help is 0.5 .Calculate the probability hat the person experiences 5 'failures' before the first success.

| `geoPDF(5, 0.5)` | 0.015625

▷ *geoCDF*: The probability of getting the help is 0.6. Calculate the probability that the person will have to talk to 8 or less people to find someone who helps.

| `geoCDF(8,0.6)` | .9997

[7] $\lceil \ldots \rceil$ is the ceiling function.

▷ *geoINV*: Suppose the probability of a five-year-old car battery not starting in cold weather is 0.03. If we want no more than a ten percent chance that the car does not start, what is the maximum number of days in a row that we should try to start the car?

| `geoINV(0.1, 0.03)` | 3

○ EIGENMATH definition and accompanying examples are in the ▷ `geometric distr.`

Graphical representation

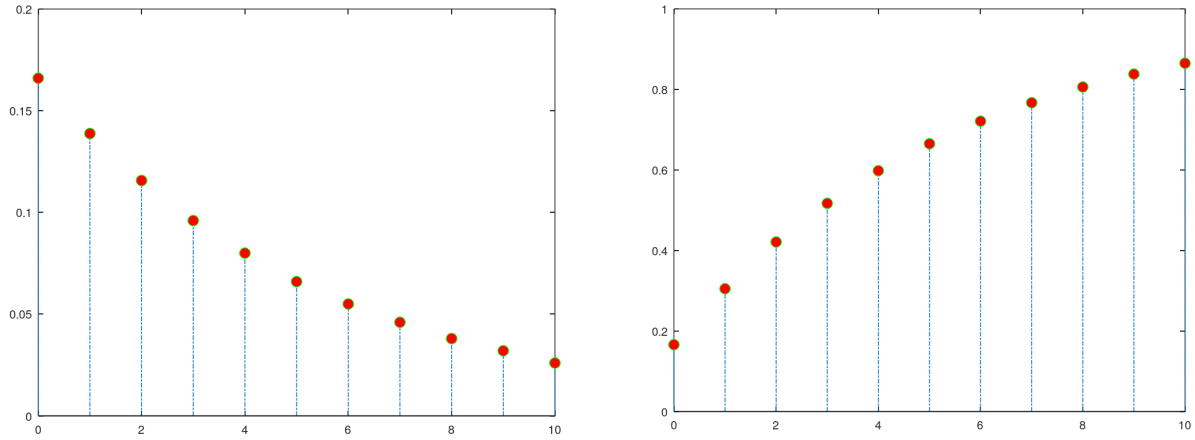


Figure 14: Experiment: rolling a six-sided die.
Left figure: • = plot for `geoPDF(k, 1/6)` for $k = 0, \ldots, 10$.
Check `geoPDF`$(4, 1/6) = 0.08$ on the graph.
Right figure: • = plot for `geoCDF(k, 1/6)` for $k = 0, \ldots, 10$.
Check `geoCDF`$(5, 1/6) = 0.665$ on graph.
Check `geoINV`$(0.7, 1/6) = 6$ on the CDF graph via $y = 0.6 \longrightarrow\downarrow 6 = k$

General information

General mathematical information about the concept is here ▷ WIKI : Geometric_distr.
Syntax and semantic of the function is here ▷ MATLAB : `geopdf`
[+] *Bognar's app*

2.3 Negative binomial distribution

The simplest motivation for the negative binomial is the case of successive random trials, each having a constant probability p of success. The number of extra trials you must perform in order to observe a given number r of successes has a negative binomial distribution. The negative binomial distribution is a discrete, i.e. existing only on the nonnegative integers.

Definition Notation: $X \sim NB(r, p)$

- The probability density function of a geometric distribution is defined as:

$$\texttt{nbinPDF}(\texttt{k}, \texttt{r}, \texttt{p}) := \binom{k+r-1}{k} p^r (1-p)^k =: f(k, r, p) = \Pr(X = k)$$

  where k is number of failures before the first success and p is the probability of success on a given trial.

- The cumulative geometric distribution can be expressed as:

$$\texttt{nbinCDF}(\texttt{k}, \texttt{r}, \texttt{p}) := \sum_{i=0}^{k} \binom{i+r-1}{r-1} p^r (1-p)^i =: F(k, r, p) = \Pr(X \leq k)$$

- The *quantile* function (inverse cumulative geometric distribution) is[8]

$$\texttt{nbinINV}(\alpha, r, p) := \inf\{k \in \mathbb{R} : \alpha \leq F(k; r, p)\}$$

  i.e. we must find the *smallest* k such that $\alpha \leq \sum_{i=0}^{k} \binom{i+r-1}{r-1} p^r (1-p)^i$ for given α.

  `nbinINV(alpha, size, prob)` returns the number of trials (or failures before the *size*-th success) such that the probability of observing that many or fewer failures is at least p.

Examples form Excel, Datacamp and MatLAB solved with EIGENMATH.

▷ *nbinPDF*: In quality control, if we need 3 defective units and each unit has a 10% chance of being defective, what is the probability of getting exactly 5 non-defective units before finding the third defective one?

| `nbinPDF(5, 3, 0.1)` | 0.0124

▷ *nbinCDF*: You have to identify five people who have excellent reflexes, you know the probability that any one candidate meets this requirement is 0.25. What is the probability that you will interview a certain number of unsuitable candidates before identifying five suitable candidates.

| `nbinCDF(10, 5, 0.25)` | 0.3135

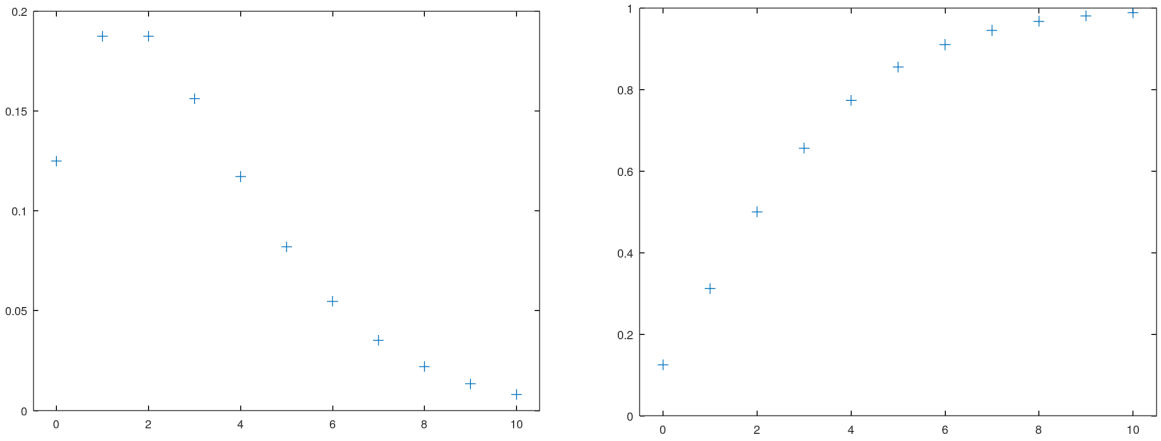
[8] $\lceil \ldots \rceil$ is the ceiling function.

▷ *nbinINV*: How many times would you need to flip a fair coin to have a 99% probability of having observed 10 heads?

| `nbinINV(0.99, 10, 0.5)` | 23

○ EIGENMATH definition and accompanying examples are in ▷ `nbin` distribution

Graphical representation



Left figure: + = plot for `nbinPDF((x,3,0.5))` for $x = 0, 1, 2, \ldots, 10$.
Check `nbinPDF`$((3, 3, 0.5) = 0.15$ on the graph.
Figure 15: Right figure: + = plot for `nbinCDF((x,3,0.5)` for $k = 0, 1, 2, \ldots, 10$.
Check `nbinCDF`$(4, 3, 0.5)) = 0.77$ on graph.
Check `nbinINV`$(0.75, 3, 0.5) = 4$ on graph via $y = 0.75 \longrightarrow\downarrow 4 = x$

General information

General mathematical information about the concept is here ▷ nbin.distribution
Syntax and semantic of the function is here ▷ MATLAB : `nbinpdf`
[†] cf. *Bognar's app*

2.4 Hypergeometric distribution

'Think of an urn with two colors of marbles, red and green. Define drawing a green marble as a success and drawing a red marble as a failure. Let N describe the number of all marbles in the urn and K describe the number of green marbles, then $N - K$ corresponds to the number of red marbles. Now, standing next to the urn, you close your eyes and draw n marbles without replacement. Define X as a random variable whose outcome is k, the number of green marbles drawn in the experiment.', cf. [†3]
The *hypergeometric* distribution is discrete, i.e. existing only on the nonnegative integers.

Definition Notation: $X \sim hg(N, R, n)$

- The probability density function of a hypergeometric distribution is defined as:

$$\texttt{hgPDF}(\texttt{N}, \texttt{R}, \texttt{n}, \texttt{r}) := \Pr(X = r) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}$$

  where r is number of failures before the first success and p is the probability of success on a given trial.

- The cumulative hypergeometric distribution can be expressed as:

$$\texttt{hgCDF}(\texttt{N}, \texttt{R}, \texttt{n}, \texttt{x}) := \Pr(X \leq x) = \sum_{r=0}^{x} \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}$$

- The *quantile* function (inverse cumulative hypergeometric distribution) is

$$\texttt{hgINV}(\alpha, N, R, n) := \inf\{k \in \mathbb{R} : \alpha \leq \texttt{hgCDF}(N, R, n, k)\}$$

  i.e. we must find the *smallest* k such that $\alpha \leq \Pr(X \leq k)$ for given α.

  `hgINV(alpha, N,R,n):` You can think of α as the probability of observing r defective items in n drawings without replacement from a group of N items where R are defective.

Examples

▷ *hgPDF*: What is the probability of selecting 14 red marbles from a sample of 20 taken from an urn containing 70 red marbles and 30 green marbles?

| `hgPDF(100, 70, 20, 14)` | 0.21409

▷ *hgCDF*: (MatLAB) Suppose you have a lot of 100 floppy disks and you know that 20 of them are defective. What is the probability of drawing zero to two defective floppies if you select 10 at random?

| `hgCDF(100, 20, 10, 2)` | 0.6812

▷ *hgINV*: (MatLAB) Suppose you are the Quality Assurance manager for a floppy disk manufacturer. The production line turns out floppy disks in batches of 1,000. You want to sample 50 disks from each batch to see if they have defects. You accept 99% of the batches if there are no more than 10 defective disks in the batch. What is the maximum number of defective disks should you allow in your sample of 50?

| `hgINV(0.99, 1000, 10, 50)` | 3

○ EIGENMATH definition and accompanying examples are in the notebook ▷hypergeometric

Graphical representation

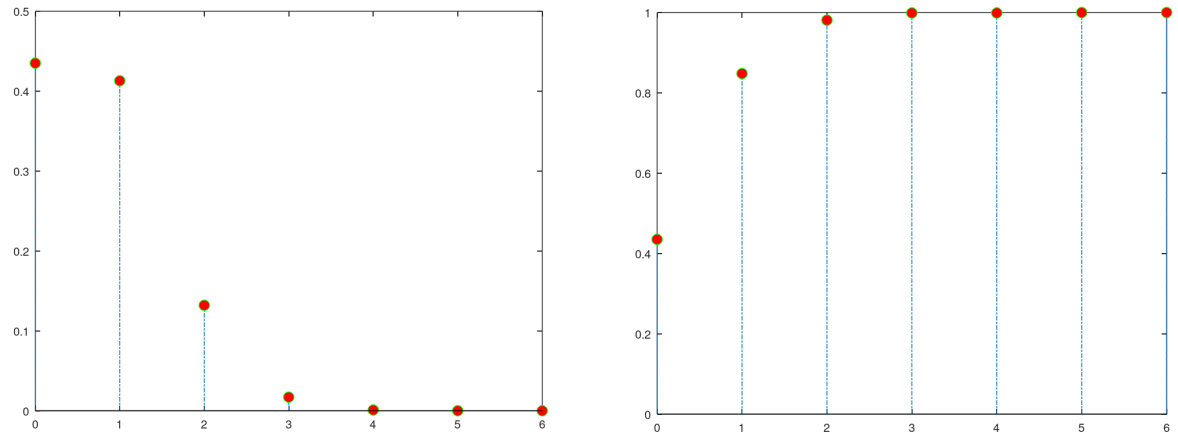


Figure 16: National lottery: from 49 including 6 red take 6 and get x red.
Left figure: • = plot for `hgPDF(49,6,6,x)` for $x = 0, 1, 2, \ldots, 6$.
Check `hgPDF`$hgPDF(49, 6, 6, 2) = 0.139$ on the graph.
Right figure: • = plot for `hgCDF(49,6,6,x)` for $x = 0, 1, 2, \ldots, 6$.
Check `hgCDF`$(49, 6, 6, 1) = 0.84$ on graph.
Check `hgINV`$(0.75, 49, 6, 6) = 1$ on graph via $y = 0.75 \longrightarrow\downarrow 1 = x$

General information

General mathematical information about the concept is here ▷ hypergeometric. distr.
Syntax and semantic of the function is here ▷MATLAB : `hygepdf`
[†] cf. *Bognar's app*

2.5 Poisson distribution

The Poisson distribution can be applied to systems with a large number of possible events, each of which is rare. The Poisson probability *density* function lets you obtain the probability of an event occurring within a given time or space interval exactly k times if on average the event occurs λ times within that interval. The Poisson distribution is discrete, i.e. existing only on the nonnegative integers.

Definition Notation: $X \sim Poisson(k, \lambda)$

- The probability density function of a Poisson distribution is defined as:

$$\texttt{poissonPDF(k,}\lambda) := \frac{\lambda^k e^{-\lambda}}{k!} = f(k, \lambda) = \Pr(X = k)$$

  where k is number of failures before the first success and p is the probability of success on a given trial.

- The cumulative Poisson distribution can be expressed as:

$$\texttt{poissonCDF(x,}\lambda) := \sum_{k=0}^{x} \frac{\lambda^k}{k!} e^{-\lambda} = F(x; \lambda) = \Pr(X \leq x)$$

- The *quantile* function (inverse cumulative Poisson distribution) is

$$\texttt{poissonINV}(\alpha, \lambda) := \inf\{x \in \mathbb{R} : \alpha \leq \sum_{k=0}^{x} \frac{\lambda^k}{k!} e^{-\lambda}\}$$

  i.e. we must find the *smallest* k such that $\alpha \leq \Pr(X \leq k)$ for given α.

  `poissonINV(`α, λ`)` returns the smallest value k such that the Poisson CDF evaluated at k equals or exceeds p, using mean parameters in lambda.

Examples

▷ *poissonPDF*: What is the probability of making 2 sales in a week if the average sales rate is 3 per week?

| `poissonPDF(3,2)` | 0.2240

▷ *poissonCDF*: (MatLAB) A computer hard disk manufacturing facility performs random tests of individual hard disks. The policy is to shut down the manufacturing process if an inspector finds more than four bad sectors on a disk. Assuming that on average a disk has two bad sectors, find the probability of a manufacturing process shutdown after the first inspection.

| `1 - poissonCDF(4,2)` | 0.0526

▷ *poissonINV*: If the average number of defects is two, what is the 95th percentile of the number of defects?

| `poissonINV(0.95, 2)` | 5

○ EIGENMATH definition and accompanying examples are in the notebook ▷POISSON

Graphical representation

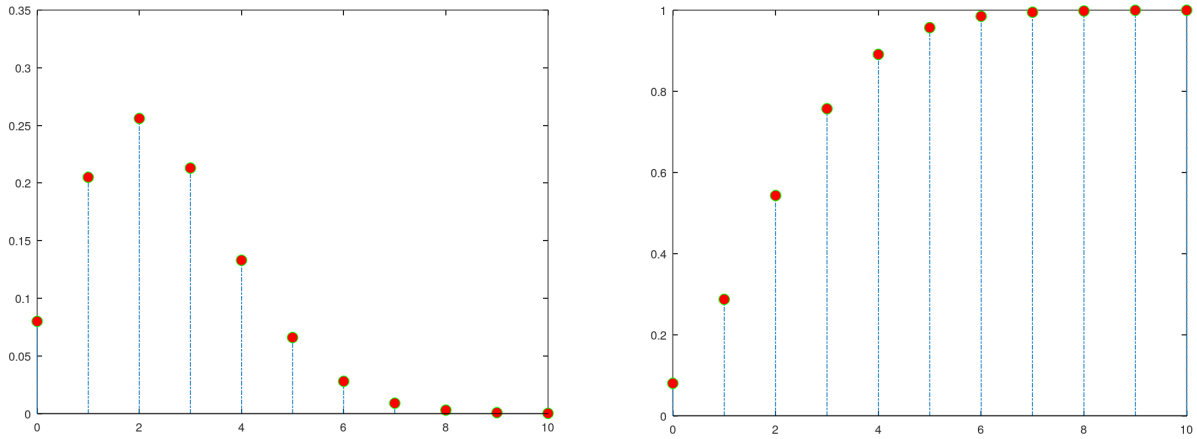


Figure 17: Plot of POISSON distribution for parameter $\lambda = 2.5$
Left figure: • = plot of `poissonPDF(2.5, k)` for $k = 0, \ldots, 10$.
Check `poissonPDF`$(2.5, 3) = 0.21$ on the graph.
Right figure: • = plot of `poissonCDF(2.5, k)` for $k = 0, \ldots, 10$.
Check `poissonCDF`$(2.5, 3) = 0.75$ on graph.
Check `poissonINV`$(0.7, 2.5) = 3$ on the CDF via $y = 0.7 \longrightarrow\downarrow 3 = k$

General information

General mathematical information about the concept is here ▷ WIKI : Poisson.distr.
Syntax and semantic of the function is here ▷ MATLAB : `poissonpdf`
[†] cf. *Bognar's app*

⋈

You should now be able to implement other discrete distributions along the above examples using CAS EIGENMATH.

⋈

3 Continuous distributions

3.1 Normal distribution

In probability theory and statistics, the *Normal Distribution*, also called the *Gaussian* Distribution by physicists, is the most significant continuous probability distribution. Social scientists refer to it as the *bell* curve, because of its curved flaring shape. A large number of random variables are either nearly or exactly represented by the normal distribution, in every physical science and economics. In a normal distribution, the mean, median and mode are equal. The total area under the curve is equal to 1. The normal distribution curve is symmetric at the centre.
The *standard normal* distribution is one of the forms of the normal distribution. It occurs when a normal random variable has mean $\mu = 0$ and a standard deviation $\sigma = 1$.

Definition Notation: $X \sim \mathcal{N}(\mu, \sigma)$

- The probability *density* function of the *normal* distribution is defined by

$$\phi(\mathtt{x}, \mu, \sigma) := \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} := \Pr(X = x) =: \mathtt{normPDF}(\mathtt{x}, \mu, \sigma)$$

- The *cumulative* normal distribution is defined by[9]

$$\Phi(x, \mu, \sigma) := \frac{1}{2} + \frac{1}{2} \cdot \operatorname{erf}\left(\frac{x - \mu}{\sigma\sqrt{2}}\right) := \Pr(X \leq x) =: \mathtt{normCDF}(\mathtt{x}, \mu, \sigma)$$

- The *inverse* normal distribution ('quantile function') is defined by

$$\Phi^{-1}(p, \mu, \sigma) := \mu + \sigma \cdot \Psi^{-1}(p) = \mu + \sigma\sqrt{2}\operatorname{erf}^{-1}(2p - 1) = \mathtt{normINV}(\mathtt{p}, \mu, \sigma)$$

  with $p \in (0, 1)$, using the *inverse standard* normal function $\Psi^{-1}(p) = \sqrt{2}\operatorname{erf}^{-1}(2p - 1)$. Because there is no build-in inverse error-function `erfinv`, we implement Φ^{-1} using Φ and the defining relation $x = \Phi^{-1}(p)$ with $\Phi(x) = p$ and a simple search method..

Examples

▷ *normPDF*: A sample of 30 students has an average test score of 78 with a standard deviation of 12. Assuming the distribution of test scores is normal, what is the probability that the sample mean score is greater than 82?

| `normPDF(82, 78, 12)` | 0.0314

[9] using the build-in error-function `erf`

▷ *normCDF*: A factory produces 100 Ω resistors, distributed via $N(100, 6^2)$. What proportion of the resistors have a maximum deviation of 10% from the mean value? Cf. [6, p. 50]

| `normCDF(110,100,6) - normCDF(90,100,6)` | 0.9044

▷ *normINV*: Find the height at which 80% of the population falls below in a normal distribution with mean 170 and standard deviation 5.

| `normINV(0.80, 170, 5)` | 174.2

○ EIGENMATH code and accompanying examples are in the notebook ▷ normal-distribution

Graphical representation

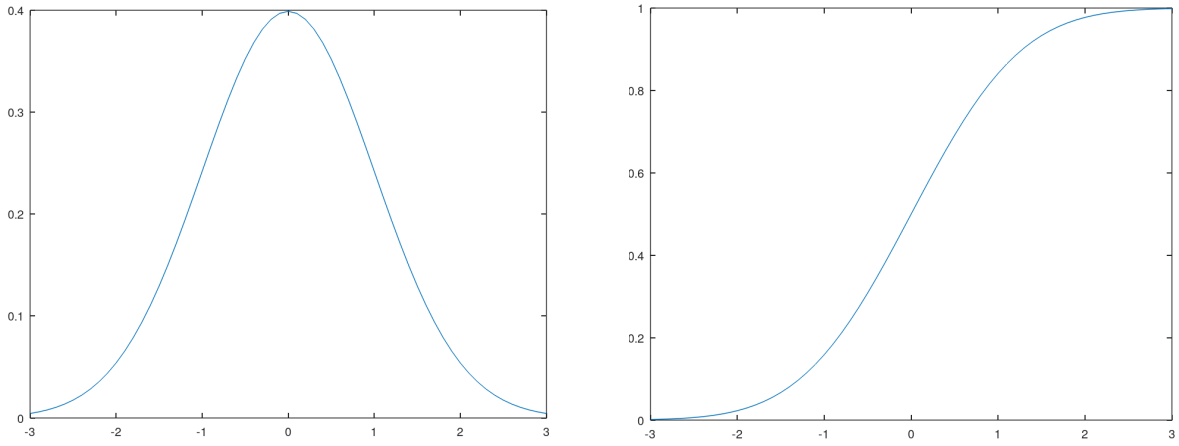


Figure 18: Standard normal distribution with parameters $\mu = 0$ and σ equal to 1.
Left figure: ≀ = plot of `normPDF(x)`.
Check `normPDF`$(1) = 0.24$ on the graph.
Right figure: ≀ = plot of `normCDF(x)`.
Check `normCDF`$(1) = 0.84$ on graph.
Check `normINV`$(0.6) \approx 0.25$ on the CDF via $y = 0.6 \longrightarrow\downarrow 0.25 = x$

General information

General mathematical information about the concept is here ▷ Normal.distribution
Syntax and semantic of the function is here ▷ MATLAB : `normalpdf`
[⊕] cf. *Bognar's app*

3.2 Exponential distribution

The exponential distribution is a probability distribution that is used to model the time we must wait until a certain event occurs. The exponential distribution is the continous analog of the geometric distribution.

Definition Notation: $X \sim Exp(\lambda)$

- The probability *density* function of the *Exponential* distribution is

$$\texttt{expPDF}(\texttt{x}, \lambda) := f(x, \lambda) = \Pr(X = x) = \begin{cases} \lambda e^{-\lambda x} & : x \geq 0, \\ 0 & : x < 0 \end{cases}$$

  where $\lambda > 0$ is the *'rate'* parameter and x is the time until the next event occurs.

- The *cumulative* Exponential distribution is

$$\texttt{expCDF}(\texttt{x}, \lambda) := F(x, \lambda) = \Pr(X \leq x) = \begin{cases} 1 - e^{-\lambda x} & : x \geq 0, \\ 0 & : x < 0 \end{cases}$$

- The *inverse*[10] (quantile) Exponential distribution is

$$\texttt{expINV}(\texttt{p}, \lambda) := F^{-1}(p, \lambda) = \frac{-\ln(1-p)}{\lambda}, \qquad 0 \leq p < 1$$

Examples

▷ *expPDF*: Compute the density of the observed value 5 in the exponential distribution specified by mean 3.

| `float( expPDF(5,1/3) )` | 0.0629

▷ *expCDF*: The lifespan of a light bulb is exponentially distributed and averages 1,000 hours. What is the probability, that it will last 800 hours?

| `float( 1 - expCDF(800, 1/1000))` | 0.4493

▷ *expINV*: (MatLAB) Assume that the lifetime of light bulbs are exponentially distributed with a mean of 700 hours. Find the median lifetime.

| `expINV(0.50, 1/700)` | 485.2

○ EIGENMATH code and accompanying examples are in the notebook
  ▷ EIGENMATH : exponential-distribution

[10] The 'quantile' function of a distribution is the inverse of the cumulative distribution function.

Graphical representation

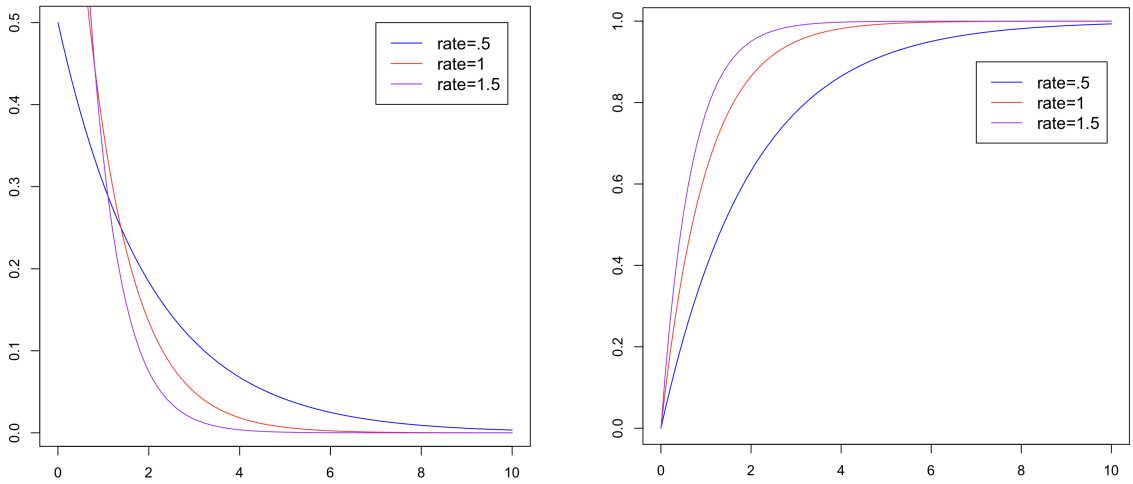


Figure 19: Exponential distributions with different rates λ.
Left figure: ≀ = plot of `expPDF(x,λ)`, $\lambda = 1; 1.5; 5$.
Check `normPDF`$(1) = 0.24$ on the graph.
Right figure: ≀ = plot of `expPDF(x,λ)`, $\lambda = 1; 1.5; 5$.
Check `expPDF` $(2, 1) \approx 0.14$ on the graph.
Check `expINV`$(0.8, 1) \approx 1.6$ on the CDF via $y = 0.8 \longrightarrow\downarrow 1.6 = x$

General information

General mathematical information about the concept is here ▷ Exponential.distribution
Syntax and semantic of the function is here ▷ MATLAB : `explpdf`
[⊕] cf. *Bognar's app*

3.3 Student's t-distribution

As we know the normal distribution assumes two important characteristics about the dataset: a large sample size and knowledge of the population standard deviation. However, if we do not meet these two criteria, and we have a small sample size (i.e. the sample size is 30 or less than 30) or an unknown population standard deviation, then we use the t-distribution. It is similar to the standard normal distribution ('Z'-distribution), but it has heavier tails. The t-score represents the number of standard deviations the sample mean is away from the population mean. ▷ [g4g]

Definition Notation: $X \sim T(t, \nu)$

- The probability density function of Student's t-distribution is defined as:

$$\texttt{tPDF}(\texttt{t}, \nu) := \Pr(X = k) = \frac{1}{\sqrt{\nu}\,\mathrm{B}\left(\frac{1}{2}, \frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

  where ν is the number of degrees of freedom, $t \in (0, 1)$ is a probability value and B is the beta function.

- The *cumulative* probability of Student's t-distributiont can be expressed as

$$\texttt{tCDF}(\texttt{t}, \nu) := \Pr(X \leq k) \;=\; = I\left(\frac{t + \sqrt{t^2 + \nu}}{2\sqrt{t^2 + \nu}}, \frac{\nu}{2}, \frac{\nu}{2}\right)$$

  using the regularized incomplete beta function I.

- The *quantile* function (inverse cumulative Student's t-distribution) is approximately

$$\texttt{tINV}(\alpha, \nu) \approx \sqrt{\nu \cdot \exp\left(c \cdot u_\alpha^2\right)} \quad \text{, where } c := \frac{\nu - \frac{5}{6}}{(\nu - \frac{2}{3} + \frac{1}{10\nu})^2}$$

  u_α is the α-quantile of the standard normal distribution.

  We give this approximation formula by Preizer & Pratt, [6, p. 70] only as reference. [11] The approximation error for $0.5 < \alpha < 0.99$ and $\nu \geq 3$ is maximal 0.08.

  Instead we implement Student's t inverse function using Student's t CDF via the defining relation $F := \Pr(X \leq k)$ and $x = F^{-1}(p, \nu)$ with $F(x, \nu) = p$ and a simple search method.

[11] See also G. W.Hill: ACM Algorithm 396. `https://dl.acm.org/doi/10.1145/355598.355599`. A formula for the quantile function of the t-distribution does not exist in a closed form.

Graphical representation

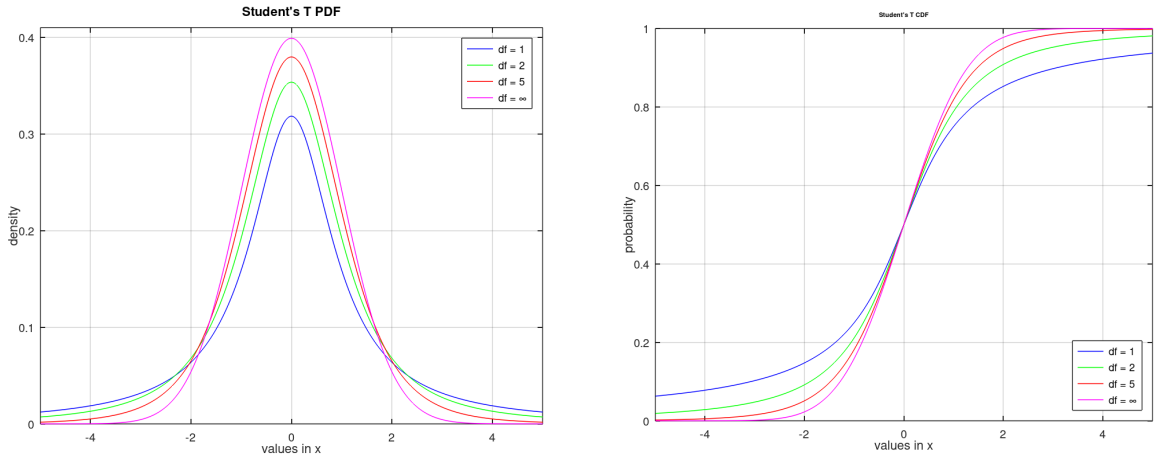


Figure 20: Student's t distributions with different degrees of freedom $df := \nu$.
Left figure: ≀ = plot of `tPDF(x,`ν`)`, $\nu = 1; 2; 5$.
Check `tPDF`$(0, 1) = 0.3$ on the graph.
Right figure: ≀ = plot of `tCDF(x,`ν`)`, $\nu = 1; 2; 5$.
Check `tCDF` $(0, 1) \approx 0.5$ on the graph.
Check `tINV`$(0.2, 1) \approx -1$ on the CDF via $y = 0.2 \longrightarrow\downarrow -1 = x$

Examples

▷ *tPDF*: (MatLAB) The mode of the Student's t distribution is at x = 0. Compute the pdf at the mode for degree of freedom 3.

| `tPDF(0,3)` | 0.3675

▷ *tCDF*: (MatLAB) Determine the probability that an observation from the Student's t distribution with degrees of freedom 99 falls on the interval [10, .., `Inf`].

| `1 - tCDF(10,99)` | 0

▷ *tINV*: (MatLAB) Compute the 99th percentile of the Student's t distribution for 3 degrees of freedom.

| `tINV(.99, 3)` | 4.5407

○ EIGENMATH code and accompanying examples are in the notebook ▷ Student-t

General information

General mathematical information about the concept is here ▷ Student-t.distribution
Syntax and semantic of the function is here ▷ MATLAB : `tpdf`
[⊕] cf. *Bognar's app*

3.4 Snedecor's F distribution

The confidence interval for the ratio of two variances requires the use of the probability distribution known as the `F`-distribution.
In particular, this distribution arises from ratios of sums of squares when sampling from a normal distribution, and is important in estimation and in hypothesis testing in the two-sample normal model. The distribution function and its quantile function do not have simple, closed-form representations. ▷ [g4g]

Definition Notation: $X \sim F(d_1, d_2)$

- The probability density function of the F-distribution is defined as

$$\texttt{fPDF}(\texttt{x}, \texttt{d}_1, \texttt{d}_2) := \Pr(X = k) = \frac{\sqrt{\frac{(d_1 x)^{d_1} \cdot d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x \cdot \mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

  where d_1, d_2 are the 'degrees of freedom', x is a probability value and B is the `beta` function.

- The *cumulative* probability of the F-distribution can be expressed as

$$\texttt{fCDF}(\texttt{x}, \texttt{d}_1, \texttt{d}_2) := \Pr(X \leq k) = I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$$

  where I_x is the regularized *incomplete* beta function.

- The *quantile* function (inverse cumulative the F-distribution) is approximately

$$\texttt{fINV}(\alpha, f_1, f_2) \approx \text{invers}(\Phi\left(\frac{x^{1/3} \cdot (1 - \frac{2}{9 f_2}) - (1 - \frac{2}{9 f_1})}{\sqrt{\frac{2}{9 f_1} + x^{2/3} \cdot \frac{2}{9 f_2}}}\right))$$

  i.e. we have to invert the $\Phi(..)$ expression to $x = \Phi^{-1}(..)$, so that we find the *smallest* k such that $\alpha \leq \Pr(X \leq k)$ for given α. We do not use this approximation formula by E. Paulson, [6, p. 72], for the implementation, but cite it here for your information.

  Instead we implement Snedecor's F inverse function using the F CDF via the defining relation $F := \Pr(X \leq k)$ and $x = F^{-1}(p, d_1, d_2)$ with $F(x, d_1, d_2) = p$ and a simple search method.

Examples

▷ *fPDF*: Calculate the density of a F-curve with $d1 = 10$, $d2 = 20$ at the value of 1.2.

| `fPDF(1.2, 10, 20)` | 0.5626

Graphical representation

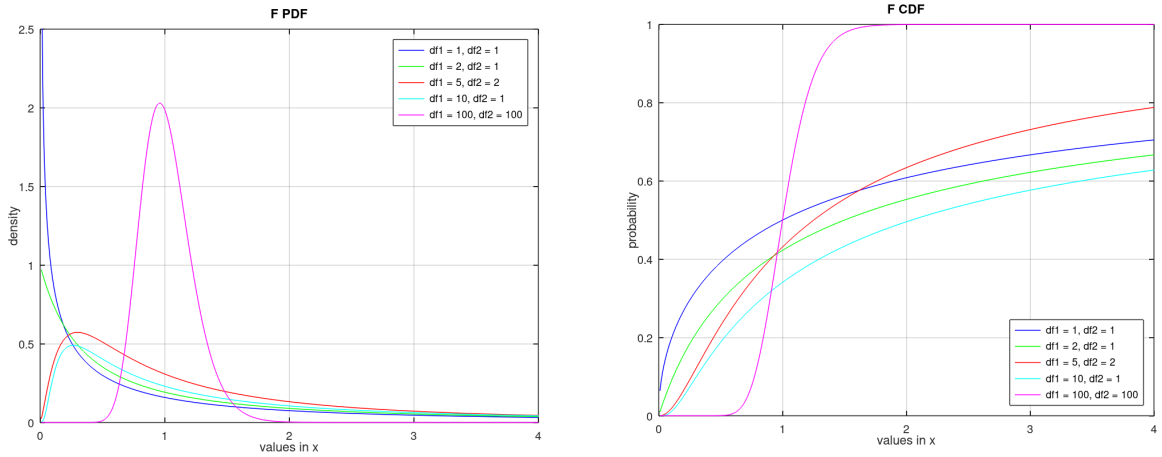


Figure 21: SNEDECOR's F distribution with different degrees of freedom d_i.
Left figure: ≀ = plot of `FPDF(x,`d_1, d_2`)`.
Check `FPDF`$(0, 1) = 0.3$ on the graph.
Right figure: ≀ = plot of `FCDF(x,`d_1, d_2`)`.
Check `FCDF` $(1, 5, 2) \approx 0.4$ on the graph.
Check `FINV`$(0.6, 100, 100) \approx 1$ on the CDF via $y = 0.6 \longrightarrow\downarrow 1 = x$

▷ *fCDF*: Calculate the area under the F-curve for the interval $[0, 1.5]$ with $d1 = 10$, $d2 = 20$.

| `fCDF(1.5, 10, 20)` | 0.7890

▷ *fINV*: Let X be an F random variable with 4 numerator degrees of freedom and 5 denominator degrees of freedom. What is the upper 5th percentile?

| `fINV(0.95, 4, 5)` | 5.1921

○ EIGENMATH code and accompanying examples are in the notebook ▷ F-distribution

General information

General mathematical information about the concept is here ▷ wiki : *F distribution*
Syntax and semantic of the function is here ▷ MATLAB : `fpdf`
[†] cf. *Bognar's app*

3.5 Chi-Square distribution

The χ^2-distribution is the distribution of a sum of the squares of independent standard normal random variables. The chi-squared distribution is used in the common chi-squared tests for goodness of fit of an observed distribution to a theoretical one, the independence of two criteria of classification of qualitative data, and in finding the confidence interval for estimating the population standard deviation of a normal distribution from a sample standard deviation. ▷ [wiki]

Definition Notation: $X \sim \chi^2(k)$

- The probability density function of the *Chi-squared distribution* χ^2 is defined as

$$\texttt{chiPDF}(\texttt{x},\texttt{k}) := \Pr(X = k) = \begin{cases} \dfrac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} & : x > 0 \\ 0 & : \text{otherwise.} \end{cases}$$

  where k is the number of degrees of freedom, and Γ is the Gamma function.

- The *cumulative* probability of χ^2-distribution can be expressed as

$$\texttt{chiCDF}(\texttt{x},\texttt{k}) := \Pr(X \leq k) = P\left(\frac{k}{2}, \frac{x}{2}\right)$$

  where $P(s,t)$ is the *regularized* gamma function.

- The *quantile* function (inverse cumulative of χ^2-distribution) is approximately

$$\texttt{chiINV}(\alpha, k) \approx k \cdot \left(1 - \frac{2}{9k} + u_\alpha \cdot \sqrt{\frac{2}{9k}}\right)^3$$

  i.e. we will find the *smallest* k such that $\alpha \leq \Pr(X \leq k)$ for given α.
  u_α is the α-quantile of the standard normal distribution. – We do not use this approximation formula by Wilson & Hilferty, [6, p. 73], for our implementation, but cite it here for your information.

  Instead we implement the χ inverse function using the χ CDF via the defining relation $F := \Pr(X \leq k)$ and $x = F^{-1}(p, k)$ with $F(x, k) = p$ and a simple search method..

Graphical representation

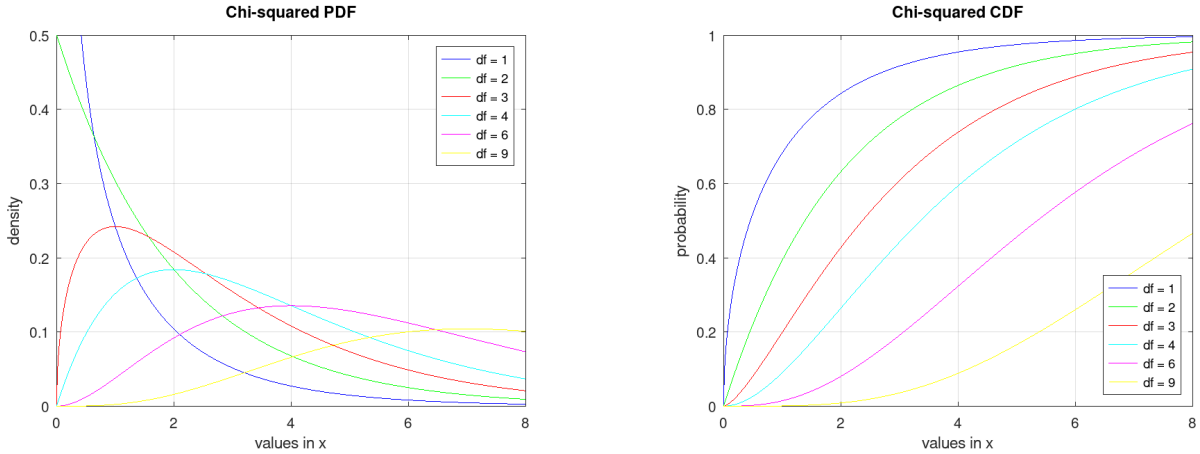


Figure 22: χ^2-distribution with different degrees of freedom k. ▷wiki.
Left figure: ≀ = plot of `chiPDF(x,`k`)`, $k = 1; 2; 3, 4; 6; 9$.
Check `chiPDF`$(2, 4) = 0.18$ on the graph.
Right figure: ≀ = plot of `chiCDF(x,`ν`)`, $\nu = 1; 2; ....$
Check `chiCDF` $(4, 4) \approx 0.6$ on the graph.
Check `chiINV`$(0.6, 4) \approx 4$ on the CDF via $y = 0.6 \longrightarrow\downarrow 4 = x$

Examples

▷ *chiPDF*: Let X follow a χ^2-distribution with 3 *df*. What is the density if the value of X is 2?

| `chi2PDF(2, 3)` | 0.2075

▷ *chiCDF*: Let X be a chi-square random variable with 3 degrees of freedom.. Compute the probability $P(0.35 \leq X \leq 7.81)$.

| `chi2CDF(7.81, 3) - chi2CDF(0.35, 3)` | 0.9002

▷ *chiINV*: Find the 95th percentile for the chi-square distribution with 10 degrees of freedom.

| `chi2INV(0.95, 10)` | 18.307

○ EIGENMATH code and accompanying examples are in the notebook ▷ chi2

General information

General mathematical information about the concept is here ▷ WIKI : Chi squared
Syntax and semantic of the function is here ▷ MATLAB : `chi2`
[†] cf. *Bognar's app*

3.6 PARETO distribution

▷ [g4g:]The PARETO distribution is a power-law probability distribution that is used in description of social, quality control, scientific, geophysical, actuarial, and many other types of observable phenomena; the principle originally applied to describing the distribution of wealth in a society, fitting the trend that a large portion of wealth is held by a small fraction of the population. The PARETO principle or "80:20 rule" stating that 80% of outcomes are due to 20% of causes was named in honour of PARETO.
▷ [MatLab:] Fitting a parametric distribution to data sometimes results in a model that agrees well with the data in high density regions, but poorly in areas of low density. For unimodal distributions, such as the normal or Student's t, these low density regions are known as the "tails" of the distribution. One reason why a model might fit poorly in the tails is that by definition, there are fewer data in the tails on which to base a choice of model, and so models are often chosen based on their ability to fit data near the mode.

If X is a random variable with a Pareto distribution, then the probability that X is greater than some number x, i.e., the 'survival' function (also called 'tail' function) is given by $\frac{\alpha \cdot x_m^\alpha}{x^{\alpha+1}}$. Here x_m is the (positive) minimum possible value of X, and α is a positive parameter. The Pareto distribution is characterized by this *scale* parameter x_m and this *shape* parameter α, which is known as the 'tail index'.

Definition Notation: $X \sim Pareto(\alpha, x_m)$

- The probability density function of the PARETO-distribution is defined as

$$\texttt{paretoPDF}(x, \mu, \alpha) := \Pr(X = k) = \begin{cases} \frac{\alpha \cdot \mu^\alpha}{x^{\alpha+1}} & : x \geq \mu \\ 0 & : x < x_m \end{cases}$$

  where μ is the *scale* parameter and the *shape* parameter is α.

- The *cumulative* probability of a PARETO distribution[12] with parameters α and x_m is

$$\texttt{paretoCDF}(\texttt{x}, \mu, \alpha) := \Pr(X \leq k) = \begin{cases} 1 - \left(\frac{\mu}{x}\right)^\alpha & : x \geq x_\text{m} \\ 0 & : x < x_\text{m} \end{cases}$$

- The *quantile* function (inverse cumulative PARETO -distribution) is[13]

$$\texttt{paretoINV}(p, \mu, \alpha) := \mu \cdot (1 - p)^{-\frac{1}{\alpha}} =: x_p$$

  where p is a a probability value and x_p is the p's quantile.

[12]The CDF function for the Pareto distribution returns the probability that an observation from a Pareto distribution with the shape parameter α and the scale parameter x_m, is less than or equal to x.

[13]This formula takes a probability p with $(0 < p < 1)$ and returns the value of x at which the CDF is equal to p.

Graphical representation

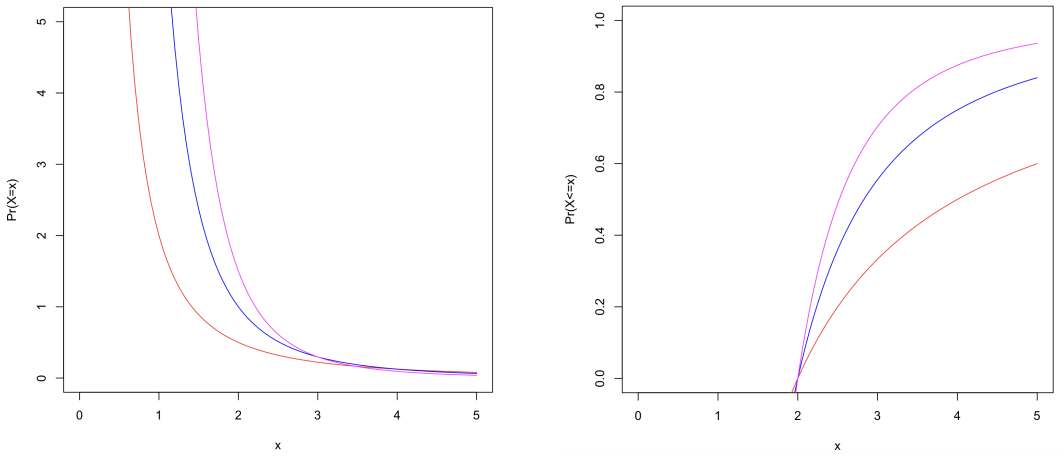


Figure 23: PARETO distributions with different shape parameters α. ▷wiki.
Left figure: ≀ = plot of `paretoPDF(x,`α`)`, $\alpha = 1; 2; 3$.
Check `paretoPDF`$(1, 1) = 1$ on the graph.
Right figure: ≀ = plot of `paretoCDF(x,`α`)`, $\alpha = 1; 2; 3$.
Check `paretoCDF` $(2, 1) \approx 0.5$ on the graph.
Check `paretoINV`$(0.6, 1) \approx 3$ on the CDF via $y = 0.6 \longrightarrow\downarrow 3 = x$

Examples In the case where the shape parameter is $\alpha = log45 = 1.160964$, we get the famous Pareto principle, aka the 80-20 rule, which states that 80% of the outcomes are due to 20% of the causes.
E.g. 20% of the workers do 80% of the work. 80% of the wealth is owned by 20% of the people.

▷ *paretoPDF*: If X follows a Pareto distribution with shape parameter $\alpha = 5$ and scale parameter $\mu = 2$, find the probability density of the distribution at $x = 3$.

| `paretoPDF(3, 2, 5)` | 0.2194

▷ *paretoCDF*: Suppose X follows a Pareto distribution with shape parameter $\alpha = 2.5$ and scale parameter $\mu = 1$. What is the probability that $X \geq 5$?

| `1 - paretoCDF(5, 1, 2.5)` | 0.0178

▷ *paretoINV*: Calculate the 25'th percentile of a Pareto distribution with parameters location=1 and shape=2.

| `paretoINV(0.25, 1,2)` | 1.1547

○ EIGENMATH code and accompanying examples are in the notebook ▷ PARETO

General information

General mathematical information about the concept is here ▷ WIKI : Pareto
Syntax and semantic of the function is here ▷ MATLAB : `gppdf`
[†] cf. *Bognar's app*

3.7 Weibull distribution

The Weibull distribution is a continuous probability distribution. It models a broad range of random variables, largely in the nature of a time to failure or time between events. Examples are maximum one-day rainfalls and the time a user spends on a web page. ▷ [wiki]

Definition Notation: $X \sim Weibull(\lambda, k)$

- The probability density function of a Weibull random variable is

$$\texttt{weibullPDF}(\texttt{x}, \lambda, \texttt{k}) := \Pr(X = k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & : x \geq 0, \\ 0 & : x < 0, \end{cases}$$

  where $k > 0$ is the *shape* parameter and $\lambda > 0$ is the *scale* parameter.

- The cumulative Weibull distribution is

$$\texttt{weibullCDF}(\texttt{x}, \lambda, \texttt{k}) := F(p, \lambda) = \Pr(X \leq k) = \begin{cases} 1 - e^{-(x/\lambda)^k}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

- The quantile function (inverse cumulative Weibull distribution) is

$$\texttt{weibullINV}(\texttt{p}, \lambda, \texttt{k}) := F^{-1}(p, \lambda) = \lambda(-\ln(1-p))^{\frac{1}{k}}$$

Examples

▷ (MatLAB) Compute the density of the observed value 3 in the Weibull distribution with unit scale and shape.

| `weibullPDF(3,1,1)` | 0.0497

▷ (MatLAB) What is the probability that a value from a Weibull distribution with parameters $a = 0.15$ and $b = 0.8$ is less than 0.5?

| `weibullCDF(0.5, 0.15, 0.8)` | 0.9272

▷ (MatLAB) The lifetimes (in hours) of a batch of light bulbs has a Weibull distribution with parameters a = 200 and b = 6. Find the median lifetime of the bulbs.

| `weibullINV(0.5, 200, 6)` | 188.15

○ Eigenmath code and accompanying examples are in the notebook ▷ Weibull

Graphical representation

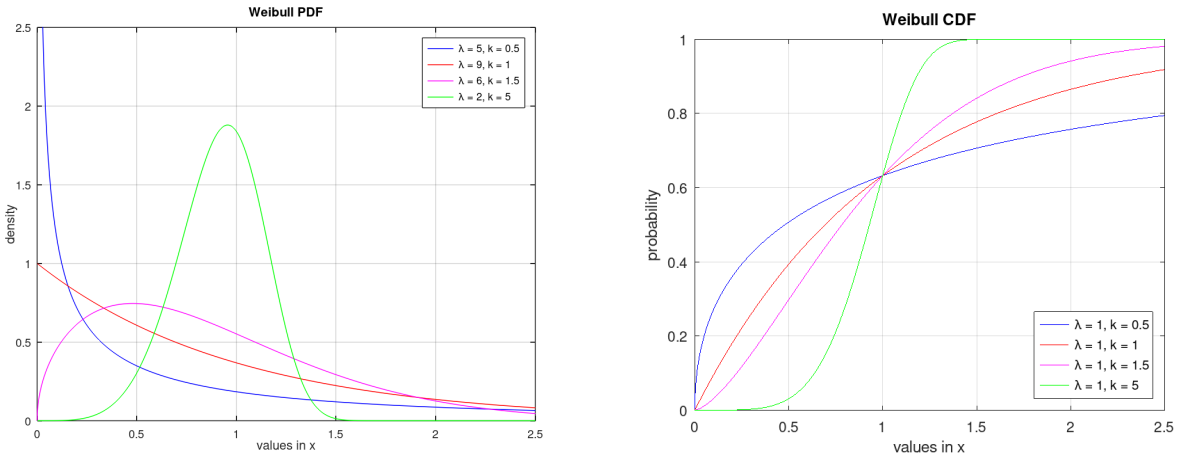


Figure 24: WEIBULL distributions with different parameters λ, k.
Left figure: ≀ = plot of `weibullPDF(x,`λ, k`)`, $\lambda = 1, k = 0.5; ..; 5$.
Check `weibullPDF`$(1, 1, 5) = 1.8$ on the graph.
Right figure: ≀ = plot of `weibullCDF(x,`λ`,k)`, $\lambda = 1, k = 0.5; ..; 5$.
Check `weibullCDF` $(1, 1, 5) \approx 0.6$ on the graph.
Check `weibullINV`$(0.8, 1, 5) \approx 1.2$ on the CDF via $y = 0.8 \longrightarrow\downarrow 1.2 = x$

General information

General mathematical information about the concept is here ▷ WIKI : Weibull
Syntax and semantic of the function is here ▷ MATLAB : `wblpdf`
⊕ cf. *Bognar's app*

4 Test Statistics

We now show, how the distributions are used to implement diverse parameter tests. We translate the mathematical concepts and definitions in EIGENMATH code and demonstrate exemplary calls and examples.

A: Parameter Tests We do all tests first *semi-automatic* in EIGENMATH, so that the user can follow the essential steps and can do the steps also by hand. This *procedural* way is followed by a *fully automatic solution* using EIGENMATH.

4.1 One Sample Z-Test alias GAUSS test

The z-test is a parametric hypothesis test used to determine whether a sample data set comes from a normal distributed population with a particular mean and a known standard deviation. The one sample GAUSS–test tests the sample for a certain mean value μ_0. For this, the parameters of the normal distribution must *not* estimated from the sample.

Mental image

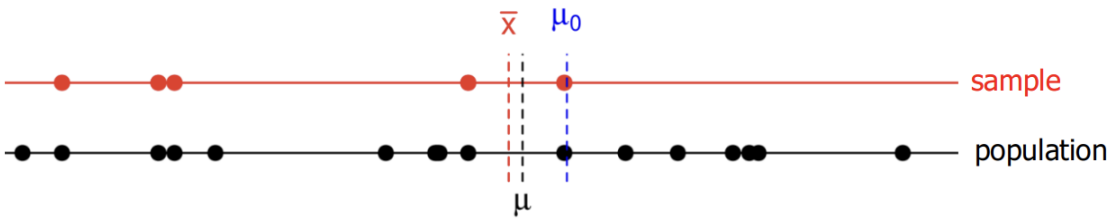


Figure 25: Visualization of Z-test for a Hypothesized Mean μ_0:
• : sample X with sample mean $\overline{X}$.
• : population with population mean μ.

Procedure *One sample Z-test alias* GAUSS *-test*

1. **Assumptions** The sample $X = (x_1, \cdots, x_n)$ must be $\mathcal{N}(\mu, \sigma^2)$ distributed.
2. **Null hypothesis** $H_0 : \mu = \mu_0$ (two-sided)
3. **Test statistics** Calculate the normal distributed test value
$$T := \frac{\sqrt{n}}{\sigma} \cdot (\overline{X} - \mu_0).$$
4. **Decision** Reject H_0, if $|T| > u_{1-\alpha/2}$.
u_α is the α-quantile of the standard normal distribution, i.e. `stdnormalINV`(α).

Remark. one-sided test $\mu > \mu_0$: $T > u_{1-\alpha}$ resp. one-sided test $\mu \leq \mu_0$: $T < u_\alpha$

Example (*female blood pressure*)

The female blood pressure of a certain population is known to follow Gaussian (alias: normal) distribution with mean 124.6 and standard deviation 14.5 measured in units of mmHg. In order to test the effect of a food product on the female blood pressure, a clinical trial was performed in which 12 female volunteers of this population consumed the product for 3 months and their blood pressure were measured in the end. The readings are as follows:

Sample of female blood pressure

n :	1	2	3	4	5	6	7	8	9	10	11	12
mmHg	141.5	152.3	121.2	123.0	151.6	124.8	138.9	137.4	145.6	135.6	135.4	121.5

Let $\alpha = 0.05$ be the probability of rejecting the null hypothesis.
Can we conclude from this data, that the population mean of the data set from which these random observations are drawn is not equal to (ie., different from) 124.6?

Solution:

```
X = (141.5, 152.3, 121.2,123.0,151.6,124.8,138.9,137.4,145.6,135.6,135.4,121.5)
-- given data:  mu0 = 124.6, sigma = 14.5, alpha = 0.05
-------- mu0   sigma alpha type
ztest(X, 124.6, 14.5, 0.05, 1)
```

Z-test

Z	2.65979
quantil	1.64485
CI.left	Inf
CI.right	128.848
p	0.00390943

∘ Follow the solution of this example using ▷ EIGENMATH: One Sample z-test.

∘ Study this example at ▷ CountBio.
∘ *Exercise*: Study using EIGENMATH the ▷ FU Berlin script.
∘ *Exercise*: Follow using EIGENMATH the ' *Example of One Sample z-test in R*' ▷ Tutorial.

General information

General mathematical information about the concept is here ▷ WIKIPEDIA : Z-Test
Syntax and semantic of the implementation is here ▷MATLAB : `Z-Test`

4.2 Two Sample Z-Test

Two sample Z-Test for two samples with two means and two known variances is to test the null hypothesis that there is no difference between the means of the two independent samples. The assumptions are: 1. Normal but independent populations. 2. Variances for the populations are known.

Mental image

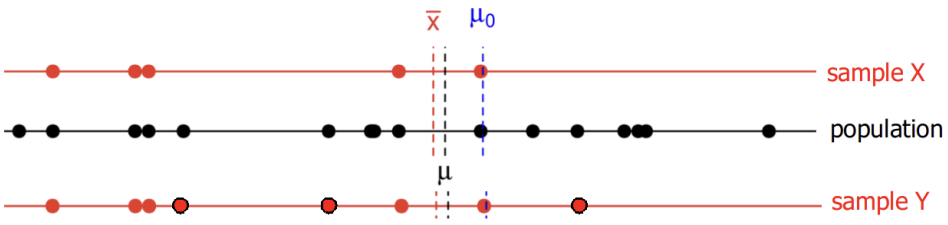


Figure 26: Visualization of Z-test for a Hypothesized Mean μ_0:
• : samples X and Y with sample means $\overline{X}$ resp. $\overline{Y}$
• : population with population mean μ.

Procedure *Two sample Z-test*

1. **Assumptions** $X = (x_1, \cdots, x_m) \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y = (y_1, \cdots, y_n) \sim \mathcal{N}(\mu_2, \sigma_2^2)$ normal distributed. X and Y independent.

2. **Null hypothesis** $H_0 : \mu_1 = \mu_2$ (two-sided)

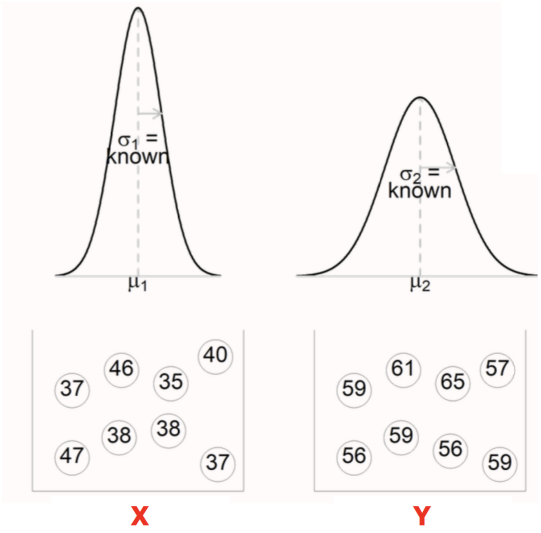
3. **Test statistics** Calculate the normal distributed test value
$$T := \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{m} - \frac{\sigma_2^2}{n}}}.$$

4. **Decision** Reject H_0, if $|T| > u_{1-\alpha/2}$.
u_α is the α-quantile of the standard normal distribution, i.e. `stdnormalINV`(α).

Remark. one-sided test of $H_0 : \mu_1 > \mu_2 \Rightarrow T > u_{1-\alpha}$
one-sided test of $H_0 : \mu_1 \leq \mu_2 \Rightarrow T < u_\alpha$
1. The Z-score T (aka the 'test statistics') represents the number of standard deviations that the difference between the two sample means is from zero.
2. The 'critical values' are based on the standard normal distribution and are used to determine whether the calculated z-score is statistically significant. If the calculated Z-score is greater than the critical value, the null hypothesis is rejected, and the alternative hypothesis is accepted.
3. For a two-tail test with a significance level of $\alpha = 0.05$, the critical value is 1.96. You can find the critical values using the EIGENMATH function `stdnormINV`$(\alpha/2)$ for the two-tail test.

Example *two samples drawn from a population.*

The figure show two samples X and Y, blindly drawn from a box with balls numbered (30) until (70) (the 'population'). Check using e.g. a QQ-plot that X and Y are approximately normal distributed.
Calculate (μ_1, σ_1) and (μ_2, σ_2).
Let $\alpha = 0.05$ be the probability of rejecting the null hypothesis $\mu_1 = \mu_2$.
Calculate the value 'test statistic' T and the critical value.
Decide, wether the null hypothesis can be accepted or must be rejected.



Solution:

```
X=(47,37,46,38,35,38,40,37.)                    | '.' to get float results
Y=(59,56,61,59,65,56,57,59.)
--    X,Y, muX, muY, sigmaX, sigmaY, alpha, altHyp
ztest2(X,Y, 0, 0, 4.11552, 2.78388, 0.05, 0)
```

__ two sample Z-test __

Reject the null hypothesis.

Z	−10.9581
Z score	1.95996
CI.left	−22.693
CI.right	−15.807
p	6.07282×10^{-28}

∘ Follow the solution of this example using ▷ EIGENMATH: two-sample-z-test.

∘ Study this example at ▷ statkat.
∘ *Exercise*: Study using EIGENMATH the example in ▷ openeducator.
∘ *Exercise*: Follow using EIGENMATH the '*Examples of Two Sample Z-test*' ▷ g4g.
∘ *Exercise*: Solve using EIGENMATH *Example.1 of*' ▷ statstutorial.

General information

General mathematical information about the concept is here ▷ WIKIPEDIA : Z-Test
Syntax and semantic of the implementation is here ▷MATLAB : `Z-Test`

4.3 One Sample t-Test

The one sample t-test is a parametric hypothesis test used to determine whether a sample data set X comes from a normal distributed population with a particular mean μ_0. In contrast to the one-sample Z-test, the standard deviation σ of the population is estimated using the sample's standard deviation s.

Mental image

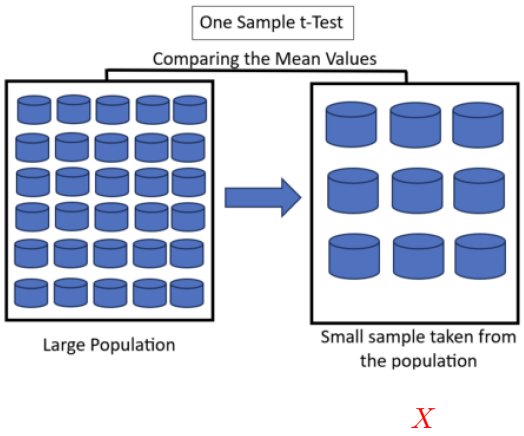


Figure 27: Visualization of one-sample-t-test for a hypothesized mean μ_0: ▷ wiki

Procedure *One sample t-test*

1. **Assumptions** independent sample $X = (x_1, \cdots, x_n)$ is $\mathcal{N}(\mu, \sigma^2)$ distributed.
2. **Null hypothesis** $H_0 : \mu = \mu_0$ (two-sided)
3. **Test statistics** Calculate the t-distributed test value with $n-1$ degrees of freedom.

$$T := \frac{\sqrt{n}}{s} \cdot (\overline{X} - \mu_0) \qquad \text{where } s := sd(X).$$

4. **Decision** Reject H_0, if $|T| > t_{1-\alpha/2\,;\,n-1} = \texttt{tINV}(1-\alpha/2\,,\,n-1)$.
   $t_{\alpha,\nu}$ is the α-quantile of the t- distribution, i.e. $\texttt{tINV}(\alpha, \nu)$.

Remark. one-sided test $\mu > \mu_0$: $T > t_{1-\alpha,\,n-1}$ resp. one-sided test $\mu \leq \mu_0 : T < t_{\alpha,\,n-1}$

Example *Students weight in Europe.*

A students data set consists of 8239 rows, each of them representing a particular student, and 16 columns, each of them corresponding to a variable/feature related to that particular student. We examine the average weight of a random sample of students from the students data set and compare it to the average weight of all European adults. WALPOLE et al. (2012) published data on the average body mass (kg) per region, including Europe. They report the average body mass for the European adult population to be 70.8 kg. We therefore set μ_0, the population mean, accordingly to $\mu_0 = 70.8$. Further, we take a random sample (X) with a sample size of $n = 9$. The sample consists of the weights in kg of 9 randomly picked students from the students data set.

Sample of students weight in Europe

n:	1	2	3	4	5	6	7	8	9
kg:	64.4	68.5	64.8	58.9	64.5	68.6	68.7	62.9	73.5

The null hypothesis H_0 states that the average weight of students equals the average weight of European adults as reported by WALPOLE. In other words, there is no difference between the mean weight of students and the mean weight of European adults. Let $\alpha = 0.05$ be the significance level of rejecting the null hypothesis.

∘ Compute the value of the test statistic T and determine the critical value.
Conclude, whether the null hypothesis H_0 is rejected or accepted.

Solution:

```
X = (64.4 , 68.5 , 64.8 , 58.9 , 64.5 , 68.6 , 68.7 , 62.9 , 73.5)
mu0 = 70.8
alpha = 0.05
ttest(X, mu0, alpha, 0)
```

One Sample t-test

both: reject H0

Significance level:	0.05
Degrees of freedom:	8
Test statistic:	−3.34583
Critical value:	−2.306

∘ Study the solution of this example in ▷ EIGENMATH: t-Test
∘ Study this example at its source ▷ FU Berlin SOGA-R script

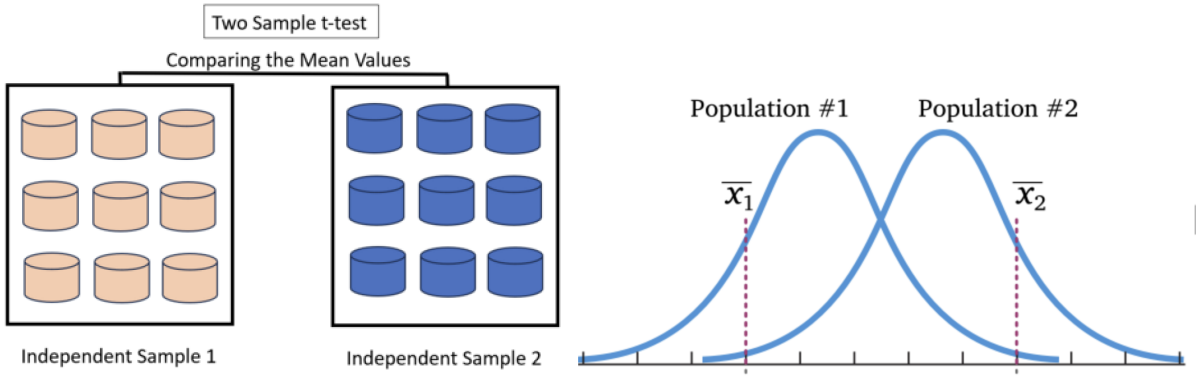
General information

General mathematical information about the concept is here ▷ WIKIPEDIA : t-Test
Syntax and semantic of the implementation is here ▷MATLAB : `t-test`

4.4 Two Sample t-Test

The *two-sample t-test* tests two normally distributed samples for the same mean value. In contrast to the two-sample Z-test, only the sample variances are used.

Mental image



Visualization of two-sample-t-test for 2 populations: ▷ wiki

Figure 28: Left: the two samples X_1 and X_2 with sample means $\overline{x_1}$ and $\overline{x_2}$.
Right: the two populations with their normal probability distribution.

Procedure *Two Sample t-test*

1. **Assumptions** independent samples $X = (x_1, \cdots, x_n)$ is $\mathcal{N}(\mu_1, \sigma_1^2)$ and $Y = (y_1, \cdots, y_n)$ is $\mathcal{N}(\mu_2, \sigma_2^2)$ distributed.

2. **Null hypothesis** $H_0 : \mu_1 = \mu_2$ (two-sided)

3. **Test statistics** T is approximately t-distributed

$$T := \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad \text{where } s_1 := sd(X), n_1 = dim(X) \ldots$$

   if the number of degrees of freedom f following Welch is chosen as

$$f := \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}}$$

4. **Decision** Reject H_0, if $|T| > t_{1-\alpha/2\,;\,f} = \texttt{tINV}(1 - \alpha/2\,,\, f)$.
   $t_{\alpha,\nu}$ is the α-quantile of the t-distribution, i.e. $\texttt{tINV}(\alpha, \nu)$.

Remark. one-sided test $\mu_1 > \mu_2$: $T > t_{1-\alpha,f}$ resp. one-sided test $\mu_1 \leq \mu_2 : T < t_{\alpha,f}$

Example *Students test scores.*

Suppose we have data from two groups (Group A and Group B), each representing the test scores of different students. We want to know if there is a significant difference between the mean test scores of the two groups.

Samples of students scores

A:	88	92	94	78	88	95
B:	75	80	79	88	85	92

The null hypothesis H_0 states that the mean test score of the two student groups A and B are equal, i.e. there is no difference between the mean scores of both student groups. Let $\alpha = 0.05$ be the significance level of rejecting the null hypothesis.

∘ *Solution* with EIGENMATH of this example in ▷ ttest2.

```
A = (88, 92, 94, 78, 88, 95.)          | '.' to get float results
B = (75, 80, 79, 88, 85, 92.)
M
```

..... two sample t test .....

$$M = \begin{bmatrix} \text{Significance level:} & 0.05 \\ \text{Degrees of freedom:} & 9.99765 \\ \text{Test statistic:} & 1.66056 \\ \text{p-value:} & 0.127795 \\ \text{CI left:} & -2.39869 \\ \text{CI right:} & 14.3987 \\ \text{mean A:} & 89.1667 \\ \text{mean B:} & 83.1667 \end{bmatrix}$$

95 percent confidence interval left limit
95 percent confidence interval rigth limit
sample estimates: mean of A
sample estimates: mean of B

ok

$c_{ritVal} = -2.22821$

reject H0

true difference in means is not equal to 0!

∘ Determine the value of the test statistic T and determine the critical value.
Argue, whether the null hypothesis H_0 is rejected or accepted.
∘ Study this example at its source ▷ Statistik Nachhilfe

General information

General mathematical information about the concept is here ▷ WIKIPEDIA : t-Test
Syntax and semantic of the implementation is here ▷MATLAB : `t-test`

4.5 Paired t-Test alias Differences t-Test

The *Differences t-Test* checks two dependent and normally distributed samples for the same mean value. As common with all tests of the t-tests group, it uses the t-distribution to calculate the test statistics, i.e. the *critical* resp. p-value.

Mental image

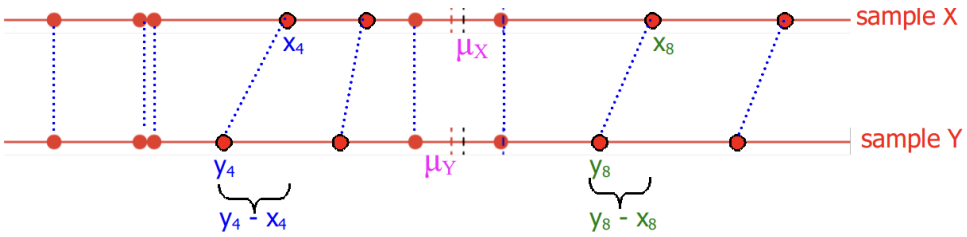


Figure 29: Visualization of paired-t-test for 2 dependent samples
above: the sample X with sample mean μ_X.
bottom: the sample Y with sample mean μ_Y.

Procedure *Paired alias Differences t-test*

1. **Assumptions** dependent samples $X = (x_1, \cdots, x_n)$ and $Y = (y_1, \cdots, y_n)$, where $d := (x_1 - y_1, \cdots, x_n - y_n)$ is $\mathcal{N}(\mu_d, \sigma_d^2)$ distributed.

2. **Null hypothesis** $H_0 : \mu_d = 0$ (two-sided)

3. **Test statistics** T is for H_0 approximately t-distributed

$$T := \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} d_i}{\frac{1}{n(n-1)} \cdot \sqrt{\sum_{i=1}^{n} d_i^2 - \frac{1}{n} \cdot (\sum_{i=1}^{n} d_i)^2}}$$

   with $df = n - 1$ degrees of freedom.

4. **Decision** Reject H_0, if $|T| > t_{\alpha/2\,;n-1} = \texttt{tINV}(\alpha/2\,,\, n-1)$.
   $t_{\alpha,\nu}$ is the α-quantile of the t-distribution, i.e. $\texttt{tINV}(\alpha, \nu)$.

Remark. one-sided test $\mu_1 > 0$: $T > t_{1-\alpha;\, n-1}$ resp. one-sided test $\mu_1 \leq 0$: $T < t_{\alpha,\, n-1}$

Example *Monthly Rainfall.*

We use the example in [6, p. 99] The monthly rainfall in millimeters over the course of two years is considered. We assume that the differences in rainfall are normally distributed. The null hypothesis is the assumption of the same rainfall with a 5% probability of error.

Rainfall in mm

month :	1	2	3	4	5	6	7	8	9	10	11	12
A:	52.4	37.4	41.2	71.5	51.3	21.6	17.7	21.2	41.3	32.0	53.0	61.4
B:	47.8	42.0	33.2	41.0	29.5	28.1	17.4	21.5	41.4	51.3	49.5	53.7

We have $\alpha = 0.05$ as the significance level of rejecting the null hypothesis.

∘ *Solution* with EIGENMATH of this example in ▷ paired t-test.

```
X = (52.4,37.4,41.2,71.5,51.3,21.6,17.7,21.2,41.3,32.0,53.0,61.4)
Y = (47.8,42.0,33.2,41.0,29.5,28.1,17.4,21.5,41.4,51.3,49.5,53.7)
T                | T = 1.02197
p                | p_Value = 0.3287
                 | alternative hypothesis: true mean difference is not equal to 0
```

∘ Read the value of the test statistic T and determine the critical value.
Argue, whether the null hypothesis H_0 is rejected or accepted.

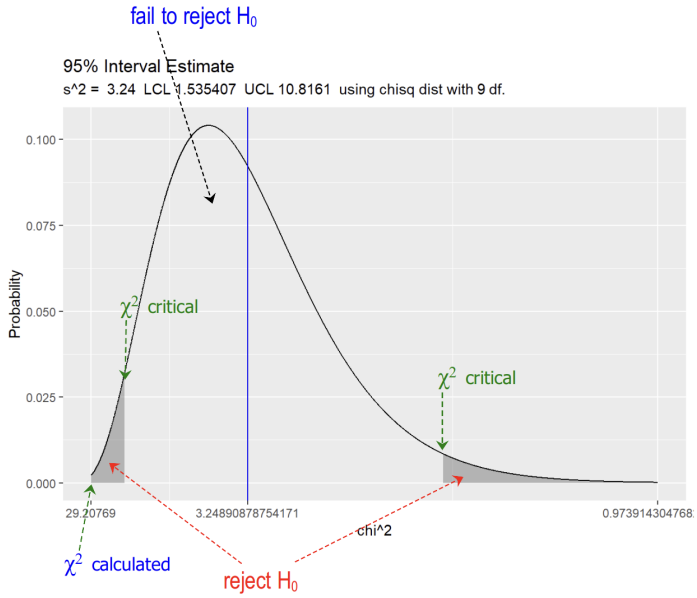
General information

General mathematical information about the concept is here ▷ WIKIPEDIA : t-Test
Syntax and semantic of the implementation is here ▷MATLAB : `t-test`

4.6 Chi-Squared Test on Variance

The Chi-Squared Test on Variance is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying. ▷ AI
The χ^2-variance test checks a normally distributed sample for a given variance σ_0^2.

Mental image



Visualization of Chi-Squared Test on Variance, cf. ▷ Foley, enhanced.
Figure 30: LCL = Lower Confidence Level; UCL = Upper CL; chisq = `chi2INV`. The values shown are from the solution of the example, see below.

Procedure *Chi-Squared Test on Variance*

1. **Assumptions** a normal distributed sample $X = (x_1, \cdots, x_n)$ with $\mathcal{N}(\mu, \sigma^2)$.
2. **Null hypothesis** $H_0 : \sigma^2 = \sigma_0^2$ (two-sided)
3. **Test statistics** T is for H_0 χ^2-distributed with $df = n - 1$ degrees of freedom.

$$T := \frac{(n-1) \cdot s^2}{\sigma_0^2}$$

where s is the sample standard deviation, σ_0 is the hypothesized standard deviation.

4. **Decision** Reject H_0, if $T > \chi^2_{1-\alpha/2\,;\,n-1} = \texttt{tINV}(1 - \alpha/2\,,\, n-1)$.
$\chi^2_{\alpha,\nu}$ is the α-quantile of the χ^2-distribution, i.e. $\texttt{tINV}(\alpha, \nu)$.

Remark. one-sided test $\sigma^2 > \sigma_0^2$: $T > \chi^2_{1-\alpha;\, n-1}$ resp. one-sided test $\sigma^2 \leq \sigma_0^2 : T < \chi^2_{\alpha,\, n-1}$.

Remark. The further the ratio $\frac{s^2}{\sigma_0^2}$ deviates from 1, the more likely you are to reject the null hypothesis.

Example *The size of prey.*

We use the example in ▷ M. Foley's *RPubs*: The size of prey (millimeters) of two species of net-casting spiders, deinopis (X) and menneus (Y) are sampled for 10 spiders each species. What is the difference in the variance of the prey of the two species? The null hypothesis is the assumption of the same mean with a 5% probability of error.

The size of prey (millimeters) of two species

X:	12.43	11.71	14.41	11.05	9.53	11.66	9.33	11.71	14.35	13.81

We have $\alpha = 0.05$ as the significance level of rejecting the null hypothesis.

◦ SOLUTION of this example in ▷ Eigenmath: χ^2-test

```
X = (12.43, 11.71, 14.41, 11.05, 9.53, 11.66, 9.33, 11.71, 14.35, 13.81)
vartest(X, 0.05, 1)
```

vartest

var, T, p, alpha, LCL, UCL:	
3.2453	estimated variance
29.2077	test statistic Chi-Squared
0.00119556	p-value
0.05	alpha
1.5354	95% Confidence Interval left
10.816	95% CI right

◦ Result: the p-value $= 0.001196 < \alpha = 0.05$, so reject H0 that $s^2 = \sigma^2$.

◦ Read the value of the test statistic T and determine the critical value.
Argue, whether the null hypothesis H_0 is rejected or accepted.
◦ Look at the solution of this example in ▷ Eigenmath: χ^2-test. Study `vartest()`.

General information
General mathematical information about the concept is here ▷ WikipediA : χ^2-test
Syntax and semantic of the implementation is here ▷Matlab : `vartest`

4.7 F test

The F-test tests two normally distributed independent samples for equal variance σ^2.

Mental image

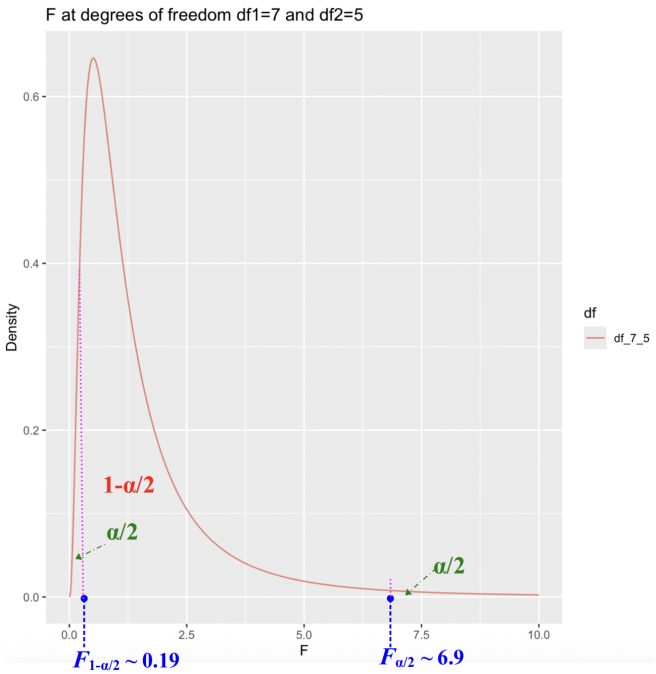


Figure 31: Visualization of F-test on equal variances, F_α is the α-quantile of F. The values shown are from the solution of the example, see below. $1-\alpha/2$ gives the area spanned inside between the limits $|\cdots|$.

Procedure *F–test for equal variances*

1. **Assumptions** normal distributed samples $X = (x_1, \cdots, x_m)$ with $\mathcal{N}(\mu_1, \sigma_1^2)$ and $Y = (y_1, \cdots, y_n)$ with $\mathcal{N}(\mu_2, \sigma_2^2)$

2. **Null hypothesis** $H_0 : \sigma_1^2 = \sigma_2^2$ (two-sided)

3. **Test statistics**
$$T := \frac{s_1^2}{s_2^2}$$
T is for H_0 F-distributed with $df_1 = m-1$ and $df_2 = n-1$ degrees of freedom, where s_i are the sample's standard deviations.

4. **Decision** Reject H_0, if $T < F_{\alpha/2,\, m-1,\, n-1} =: \texttt{fINV}(\alpha/2,\ m-1,\ n-1)$.
$F_{\alpha,\nu}$ is the α-quantile of the F-distribution, i.e. $\texttt{fINV}(\alpha, \nu)$ in EIGENMATH.

Remark. one-sided test $\sigma_1^2 > \sigma_2^2$: $T > F_{1-\alpha;\,m-1,\,n-1}$ resp. one-sided test $\sigma_1^2 \leq \sigma_2^2$: $T < F_{\alpha,\,m-1,\,n-1}$.

Remark. Note, that $F_{\alpha,\,m-1,\,n-1}$ is the *critical value* of the F distribution with $m-1$ and $n-1$ degrees of freedom and a significance level of α.

Remark. It should be noted that the F-test is *not robust*, i.e. it is very sensitive to small deviations from the normal distribution.

Example *Groundwater sulfate concentrations.*
We use the example from ▷ M. GIMOND : Groundwater sulfate concentrations are monitored at a contaminated site over the course of a year. Those concentrations are compared to ones measured at background sites for the same time period. We seek to compare the concentration of sulfates between background sites and a contaminated well (data taken from MILLARD et al., p. 418). Did the two samples have equal variances?
The concentrations of sulfate (in ppm) for both sites are as follows:

Groundwater sulfate concentrations in ppm

Contaminated:	600	590	590	630	610	630		
Background:	560	530	570	490	510	550	550	530

∘ SOLUTION of this example in ▷ EIGENMATH: F-test.

```
X = (560, 530, 570, 490, 510, 550, 550, 530) -- Background
Y = (600, 590, 590, 630, 610, 630)           -- Contaminated
varTest(X, Y)           -- F test to compare two variances
```

F-test

T, p, lCI, uCI:	Interpretation:
2.11634	F test statistic value
0.426341	p-value
0.308818	95% Confidence Interval left
11.1852	95% CI right

Result: the p-value $= 0.4263 > \alpha = 0.05$, and with such a high p, we cannot reject the null hypothesis and therefore state that the variances between both populations are the same.

General information
General mathematical information about the concept is here ▷ WIKIPEDIA : F-test
Syntax and semantic of the implementation is here ▷MATLAB : `vartest2`

B: Parameter tests

We quote HERMANN[6, p.134]: "In contrast to the parameter tests of the last Chapter, non-parametric tests do not require the presence of a normal distribution of the data or a specific parameter. The parameterfree tests are therefore based on a nominal or ordinal scale of the data as they are usually given in sociology, pedagogy and psychology.
Since a nominal or original scale, in contrast to interval scales, do not allow a mean concept, only combinatorial arrangements such as rank or sign distributions, iterations ('runs') or information statistics can be evaluated. This make parameterfree test methods universally applicable, but leads, for example, to information loss with normally distributed data."

Hereinafter, we do all tests first in a *semi-automatic* way, so that the user can follow the essential steps and can do these steps also by hand. This *procedural* way is accompanied by a *fully automatic solution* using functions in CAS EIGENMATH.

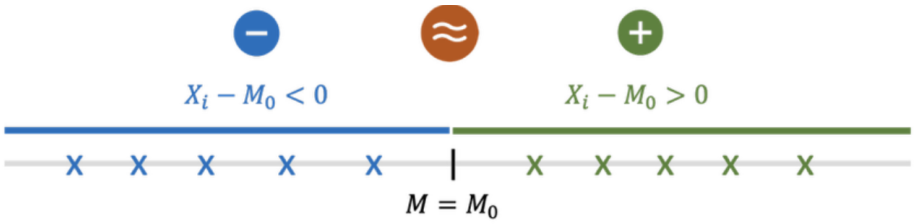
We discuss using EIGENMATH

1. Sign tests
2. WILCOXON tests
3. MANN–WHITNEY U test
4. PEARSON χ^2 test
5. FISHER test
6. MCNEMAR test

4.8 One Sample Sign Test

The *sign test* checks the symmetry of a sample with respect to a central value, e.g., the median, by counting the signs that result from the differences to the median.

Mental image



Visualization of F-test on equal variances, F_α is the α-quantile of F.
Figure 32: The values shown are from the solution of the example, see below.
$1 - \alpha/2$ gives the area spanned inside between the limits $|\cdots|$.

Procedure *One-Sample sign–test*

1. **Assumptions** ordinal distributed independent sample $X = (x_1, \cdots, x_m)$ with continous CDF symmetric to the median $x_{0.5}$

2. **Null hypothesis** $H_0 : x_{0.5} = x_0$ (two-sided)

3. **Test statistics**

$$T := \sum_{i=1}^{n} y_i \qquad \text{where} \quad y_i = \begin{cases} 1, & x_i - x_0 \geq 0, \\ 0, & x_i - x_0 < 0. \end{cases}$$

   T for H_0 is $\sum_{k=0}^{n} B(k, n, 0.5)$-distributed, where $B(k, n, p)$ is the density of the binomial distribution.

   The T-Statistic the number of positive differences between the data and the hypothesized median $x_{0.5}$.

4. **Decision** Reject H_0, if $T < B_{\alpha/2} =: \texttt{binINV}(\alpha/2)$ or $T > B_{1-\alpha/2}$.
   B_α is the α-quantile of the binomial B-distribution, i.e. $\texttt{binINV}(\alpha)$ in EIGENMATH.

Example *Average age of men at first marriage.*

HERRMANN [6, p.135] gives the following example of a one-sample sign test: The following table shows the average age of men at first marriage (Federal Republic of Germany, 1971-1984):

average age of men at first marriage

year :	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
age:	26.0	25.6	25.5	25.6	25.3	25.6	25.7	25.9	26.0	26.1	26.3	26.6	26.9	27.0

The table is to be examined for a trend.
Check, if the age of 1st marriage has median 26.
Use the one-sample-sign-test to analyze the experiment.

◦ SOLUTION of this example with EIGENMATH in ▷ signtest1.

```
X = (26.0, 25.6, 25.5, 25.6, 25.3, 25.6, 25.7,
     25.9, 26.0, 26.1, 26.3, 26.6, 26.9, 27.0)
signtest(X, 26.0, "both")
```

signtest

s	p
5	0.774414

Interpretation:
5 differences with a positive sign,
i.e. 5 data elements have a value > median.
probability p for $s = 5$ is $0.7744 > 0.05 = \alpha$

Result: The null hypothesis, $Ho : median = 26$, can be rejected.

◦ Study function `signtest()`.

General information

General mathematical information about the concept is here ▷ WIKIPEDIA : Sign-test
Syntax and semantic of the implementation is here ▷MATLAB : `signtest`

4.9 Two Sample Sign Test

The *Two Sample Sign test* tests for two paired samples for the same central tendency by counting the signs that result from the differences between corresponding data items.

Mental image

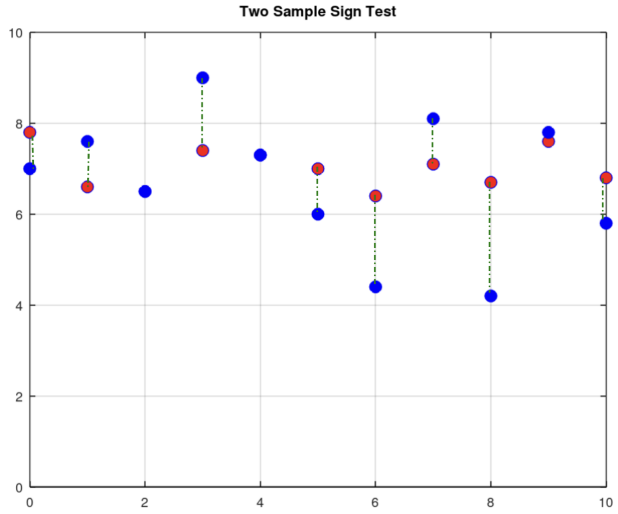


Figure 33: Visualization of *sign*2-test on two samples X and Y with ⋮ $= x_i - y_i$.
●: sample $X = (7.8, 6.6, 6.5, 7.4, 7.3, 7.0, 6.4, 7.1, 6.7, 7.6, 6.8)$.
●: sample $Y = (7.0, 7.6, 6.5, 9.0, 7.3, 6.0, 4.4, 8.1, 4.2, 7.8, 5.8)$.

Procedure *Two-Sample sign–test*

1. **Assumptions** two samples $X = (x_1, ..., x_n)$ and $Y = (y_1, ...y_n)$ with independent differences $D = (x_1 - y_1, \cdots, x_n - y_n)$

2. **Null hypothesis** $H_0 : Pr(X < Y) = Pr(X > Y)$ (two-sided)

3. **Test statistics**

$$T := \sum_{i=1}^{n} d_i \qquad \text{where} \quad d_i = \begin{cases} 1, & x_i - y_i > 0, \\ 0, & x_i - y_i < 0. \end{cases}$$

   T for H_0 is $\sum_{k=0}^{T} B(k, n, 0.5)$-distributed, where $B(k, n, p)$ is the density of the binomial distribution.

   The T-Statistic the number of positive differences $d_i > 0$ between corresponding data items in X and Y.

4. **Decision** Reject H_0, if $T < B_{\alpha/2} =: \texttt{binINV}(\alpha/2)$ or $T > B_{1-\alpha/2}$.
   B_α is the α-quantile of the binomial B-distribution, i.e. $\texttt{binINV}(\alpha)$ in EIGENMATH.

Example *IQ of twin pairs.*

Herrmann [6, p.142] gives the following example of a two-sample sign test: Given the following observation table of weights and intelligence quotients (IQ) of single twins after Churchill & Willerman. Test the null hypothesis that the IQ scores of the twins are independent of their birth weight.
IQ_ℓ designate the lighter twin pairs and $IQ_\hbar$ the heavier twin pairs.

IQ of twin pairs

pair:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$IQ_\hbar$:	97	79	100	100	100	124	95	80	91	108	91	90	104	119
IQ_ℓ:	97	70	101	106	85	123	84	70	84	106	97	90	92	104

Use the two-sample-sign-test to analyze the experiment.

∘ SOLUTION and explanation of this example with Eigenmath in ▷ signtest2.

```
IGh = (97,79,100,100,100,124,95,80,91,108,91,90,104,119)
IQl = (97,70,101,106,85,123,84,70,84,106,97,90,92,104)
signtest2(IGh, IQl, "both")
```

signtest2

T	p
9	0.145996

Interpretation:
$T = 9$ differences with a positive sign,
i.e. 9 data pairs have a difference greater 0.
probability for $T = 8$ is $p = 0.1459 > 0.05 = \alpha$

Result: The null hypothesis, Ho: "IQ independent of weight", can not be rejected with a 5 % probability of error.

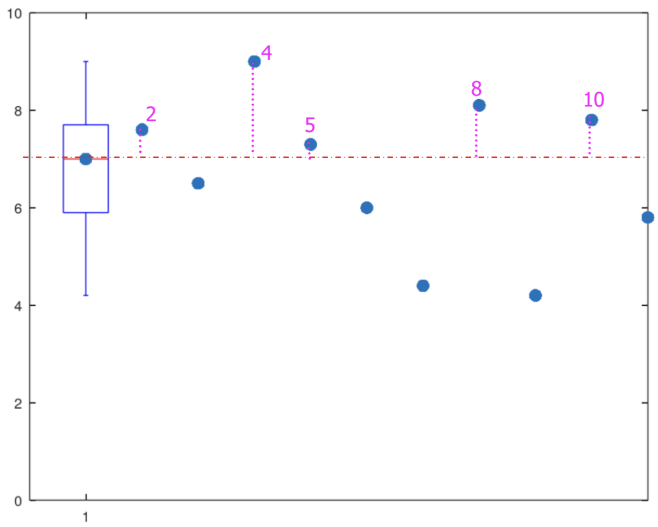
General information

General mathematical information about the concept is here ▷ WikipediA : Sign-test
Syntax and semantic of the implementation is here ▷MATLAB : `signtest`

4.10 One Sample WILCOXON Test

The *One Sample Sign* test only registers the signs of the differences to the median. The *One Sample Sign* WILCOXON test also considers the absolute values of the differences. These values are sorted and ranked, and the *rank sum* of the positive differences is calculated. This sum represents the WILCOXON test statistic W.

Mental image



Visualization of WILCOXON-One-Sample test on the sample X.
Figure 34: •: sample $X = (7.0, 7.6, 6.5, 9.0, 7.3, 6.0, 4.4, 8.1, 4.2, 7.8, 5.8)$.
i: items with index $i = 2, 4, 5, 8, 10$ have positive distance to the median.

Procedure *One-Sample*WILCOXON*–test*

1. **Assumptions** one sample $X = (x_1, ..., x_n)$ symmetrical w.r.t. the median.
2. **Null hypothesis** H_0 : median$(X) = x_0$ (two-sided)
3. **Test statistics** with $d_i := x_i - x_0$

$$W := \sum_{i=1}^{n} c_i \cdot Rank(|d_i|) \qquad \text{where} \quad c_i := \begin{cases} 1, & x_i - x_0 > 0, \\ 0, & x_i - x_0 < 0. \end{cases}$$

4. **Decision** Reject H_0, if $W \leq w_{\alpha/2}$ or $W \leq w_{1-\alpha/2}$.
w_α is the critical value of the tabulated WILCOXON-distribution, ▷ U-test

Example *median length of pygmy sunfish*

We quote the example from PennState Eberly College at ▷ Example 2.2.
Let X_i denote the length, in centimeters, of a randomly selected pygmy sunfish, $i = 1, .., 10$.
If we obtain the data set `5.0 3.9 5.2 5.5 2.8 6.1 6.4 2.6 1.7 4.3` can we conclude that the median length of pygmy sunfish differs significantly from 3.7 centimeters?

Solution step-by-step and maybe by hand using 3.7 for the hypothesized median.

1. We construct a table with the item numbers No_i, the items X_i itself, the items centered around 3.7 i.e. $X_i - 3.7$, their absolute values $\mid X_i - 3.7 \mid$, the ranked absolute values R_i and the signed ranked absolute values $signed\, R_i$:

No	1	2	3	4	5	6	7	8	9	10
Xi	5	3.9	5.2	5.5	2.8	6.1	6.4	2.6	1.7	4.3
Xi-3.7	1.3	0.2	1.5	1.8	−0.9	2.4	2.7	−1.1	−2	0.6
|Xi-3.7|	1.3	0.2	1.5	1.8	0.9	2.4	2.7	1.1	2	0.6
Rank Ri	5	1	6	7	3	9	10	4	8	2
signed Ri	5	1	6	7	0	9	10	0	0	2

2. $W = sum(signed\, R_i) = 40$
3. Check the ▷ `w-table`; with $n = 10$, a small sample size, the upper and lower percentiles of the Wilcoxon signed rank statistic is: $n = 10 : Pr(T >= W = 40) = 0.116$
4. Therefore, our P-value is $2 \times 0.116 = 0.232$. Because our P-value is large, we cannot reject the null hypothesis.

∘ SOLUTION and explanation of this example with EIGENMATH in ▷ signrank.

```
X = (5.0, 3.9, 5.2, 5.5, 2.8, 6.1, 6.4, 2.6, 1.7, 4.3)
signrank(X, 3.7)
```

_ Wilcoxon test _

W	T
40	1.27412

General information

General mathematical information about the concept is here ▷ WIKIPEDIA : signrank
Syntax and semantic of the implementation is here ▷MATLAB : `signrank`.

4.11 Two Sample Wilcoxon Test

The *Two Sample* Wilcoxon *Test* tests for two paired samples X and Y for the same central tendency by counting the rank sums of their differences. It is a non-parametric alternative to the paired t-test, used when you have paired or dependent data, such as two measurements from the same individual.

Mental image

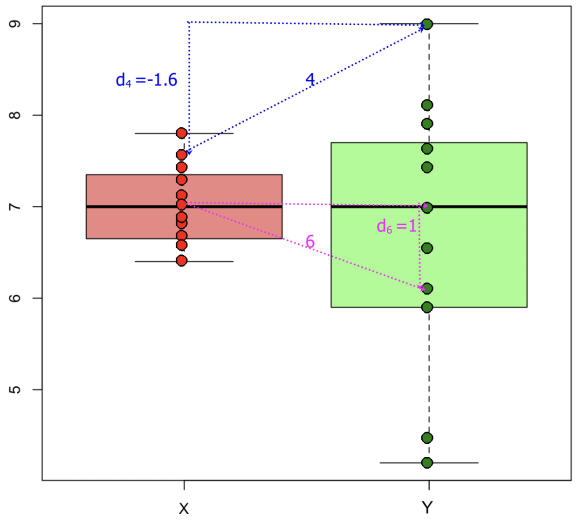


Figure 35: Visualization of Two Sample Wilcoxon–test on two samples X and Y
●: sample $X = (7.8, 6.6, 6.5, 7.4, 7.3, 7.0, 6.4, 7.1, 6.7, 7.6, 6.8)$.
●: sample $Y = (7.0, 7.6, 6.5, 9.0, 7.3, 6.0, 4.4, 8.1, 4.2, 7.8, 5.8)$.
●: For pair 6 we have $d_6 = x_6 - y_6 = 7 - 6 = +1$.

Procedure *Two-Sample*Wilcoxon*–test*

1. **Assumptions** two samples $X = (x_1, ..., x_n)$ and $Y = (y_1, ...y_n)$ with differences $D = (x_1 - y_1, \cdots, x_n - y_n)$
2. **Null hypothesis** H_0 : median$(D) := \tilde{d} = 0$ (two-sided)
3. **Test statistics** W with $d_i := x_i - \tilde{x}$ is calculated by

$$W := \sum_{i=1}^{n} z_i \cdot Rank(|d_i|) \qquad \text{where} \quad z_i := \begin{cases} 1, & x_i - \tilde{d} > 0, \\ 0, & x_i - \tilde{d} < 0. \end{cases}$$

4. **Decision** Reject H_0, if $W \leq w_{\alpha/2}$ or $W \leq w_{1-\alpha/2}$.
w_α is the critical value of the tabulated Wilcoxon-distribution, ▷ U-test

Example *Figure.35*

We do the two-sample WILCOXON–test for the data of figure.35.
We have to test the null hypothesis that the median of the differences $X - Y$ is 0, i.e. to test if the median is the same for the first and second sample.

Solution along the rows of the 'rank' table:

1	2	3	4	5	6	7	8	9	10	11	item number
7.8	6.6	6.5	7.4	7.3	7	6.4	7.1	6.7	7.6	6.8	sample X
7	7.6	6.5	9	7.3	6	4.4	8.1	4.2	7.8	5.8	sample Y
0.8	−1	0	−1.6	0	1	2	−1	2.5	−0.2	1	differences $d_i := x_i - y_i$
4	6.5	1.5	9	1.5	6.5	10	6.5	11	3	6.5	rank'ing of $\|d_i\|$
↓					↓	↓		↓		↓	

1. rank'sums: $T^+ = 4+6.5+10+11+6.5 = 38$[14] and analog $T^- = 6.5+9+6.5+3 = 25$.
2. test statistic $T := min(T^+, T^-) = min(38, 25) = 25$.[15]
3. tabulated critical value $w_{crit}(0.05; 9) = 6$. $[n = 11-2 = 9$, because of 'ties' $(7.3, 7.3)$ and $(6.5, 6.5)]$
4. Because $T = 25 > 6 = w_{crit}$, accept H_0.
5. Check the plausibility of the decision by Figure.35.

∘ SOLUTION and explanation of this example with EIGENMATH in ▷ signrank2.

```
X = [7.8, 6.6, 6.5, 7.4, 7.3, 7.0, 6.4, 7.1, 6.7, 7.6, 6.8];
Y = [7.0, 7.6, 6.5, 9.0, 7.3, 6.0, 4.4, 8.1, 4.2, 7.8, 5.8];
signrank2(X,Y)
```

signrank2

Wp	Wm	T
38	25	25

General Information
General mathematical information about the concept is here ▷ WIKIPEDIA : rank test
Syntax and semantic of the implementation is here ▷MATLAB : `signrank(x,y)`

[14] If one or more differences $x_i - y_i = 0$, which is the case here, the corresponding observations are not included in the test, i.e. these pairs are deleted from the sample. Then we get the same value as OCTAVE: $W = T^+ =$ `signedrank` $= 28$.

[15] In contrast to the often used $T := min(T^+, T^-)$, in our definition in the text only T^+ is used to define $W := T^+$. This seems to be the same choice as in OCTAVE/MATLAB.

4.12 Mann–Whitney U test

The Mann–Whitney U test is designed for independent samples. Since the sample sizes no longer necessarily are equal, it is not possible to calculate pairwise differences. Therefore, the two samples are combined into a single total sample, and the rank values of all sample items are determined. The test statistic W is then the smaller sum of the ranks from the two samples.

Mental image

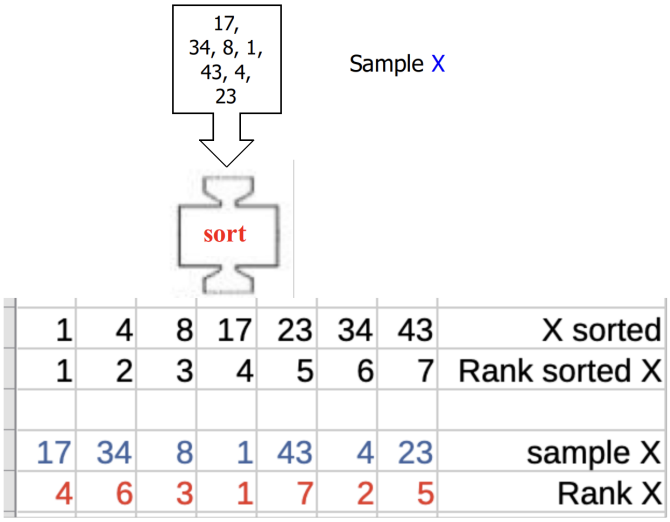


Figure 36: Visualization of U-test process on sample $X : 17, 34, 8, 1, 43, 4, 23$.
1: $X = (17, 34, 8, 1, 43, 4, 23) \xrightarrow{sort} (1, 4, 8, 17, 23, 34, 43) = X_{sorted}$.
2: $X_{sorted} = (1, 4, 8, 17, 23, 34, 43) \xrightarrow{Rank} (1, 2, 3, 4, 5, 6, 7) = X_{ranked}$.

Procedure Mann–Whitney U *test*

1. **Assumptions** two independent samples $X = (x_1, ..., x_m)$ and $Y = (y_1, ...y_n)$ with $m < n$ and concepts (e.g. `median`) F resp. G.

2. **Null hypothesis** $H_0 : F = G$ (two-sided)

3. **Test statistics**

$$U := W - \frac{m * (m + n + 1)}{2} \qquad \text{where} \quad W := \sum_{i=1}^{m} Rank(X_i)$$

4. **Decision** Reject H_0, if $U < w_{\alpha/2}$ or $U > w_{1-\alpha/2}$.
w_α is the critical value of the tabulated Wilcoxon-distribution, ▷ Wilcoxon : U-test

Example *Butterflies on sunny vs cloudy days.*

Zoefel [19, p.104] gives the following example of a Mann–Whitney U test: As part of a biological study, butterflies were observed at various times within a fixed timeframe. The weather conditions were recorded, with a general classification into sunny and cloudy periods. The results are shown in the following table:

butterflies on sunny vs cloudy days

sunny:	6	15	35	35	62	73	98	112
cloudy:	1	4	8	17	23	34	43	

An U test should explain whether the difference between the two median values of the samples is significant.

Solution along the columns of the rank table:

sun:	rank:	cloud:	rank:
6	3	1	1
15	5	4	2
35	9.5	8	4
35	9.5	17	6
62	12	23	7
73	13	34	8
98	14	43	11
112	15		
	___		___
sum:	81		39

1. ranksum's $R_1 = 81$ and $R_2 = 39$.
2. U values
   $U_1 = R_1 - n_1(n_1 + 1) \cdot 0.5 = 11$ and
   $U_2 = R_2 - n_2(n_2 + 1) \cdot 0.5 = 45$.
3. The tabulated critical value for
   $U = min(U1, U2) = 11$ is 10.
4. Because $U = 11 > 10 = U_{crit}$,
   reject H_0.

∘ *Solution* and explanation of this example with Eigenmath in ▷ U test.

```
X = (6, 15, 35, 35, 62, 73, 98, 112)
Y = (1, 4, 8, 17, 23, 34, 43)
ranksum(X,Y)
```

ranksum

U	p	Ho
11	0.97543	reject H0

General information

General mathematical information about the concept is here ▷ WikipediA : U-test
Syntax and semantic of the implementation is here ▷MATLAB : `ranksum`

4.13 PEARSON's Chi-squared test & Contingency Tables

Many experimental results can be represented in the form of so called '*four-field tables*' or multi-field tables, also called *contingency tables*. The tests developed for this purpose do not require any specific parameters of the populations, they are therefore *non-parametric test* procedures. A very popular test is the *chi-square test* for four-field tables.

Mental image

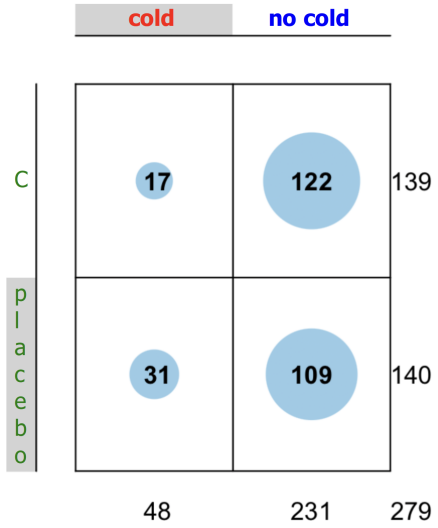


Figure 37: Visualization of Four-field Table for am experimental outcome.
•: experiment outcome as frequency table $T = \binom{a\,b}{c\,d} = \binom{17\,122}{31\,109}$.

Procedure PEARSON's *independence test*

a	b
c	d

1. **Assumptions** The frequencies of the dichotomous characteristics A and B are counted on $n = a+b+c+d$ subjects forming a random sample; the observations are made independently.

2. **Null hypothesis** H_0 : the characteristics A and B are independent (two-sided)

3. **Test statistics**

$$\chi^2 := \frac{n \cdot (ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad \text{is approximately } \chi^2_1 \text{ distributed.}$$

4. **Decision** Reject H_0, if $\chi^2 > \chi^2_{1;1-\alpha/2}$.
▷ U-test

Example *Does vitamin C help against colds?*

We do the example in [6, p.153, 160]. Studies by the physician G. Ritzel from 1961, who tested the influence of vitamin C on cold prophylaxis on skier in a double-blind, randomized experiment, was given in the following four-field table.

	cold	no cold
C	17	122
placebo	31	109

Test the null hypothesis $Pr[cold] = Pr[no\ cold]$ for a 5% probability of significance.

Solution

1. n = 279.
2. test statistic $\chi = 4.1407$.
3. the p-value $1 - \chi(0.05; 1) = 0.0418$.
4. Because $p = 0.042 < 0.05 = \alpha$, accept H_0.

∘ *Solution* and explanation of this example with Eigenmath in ▷ χ^2-test.

```
chisqTest(17,122,  31,109)
```

ChiSquared test

Chi	p	phi
4.14068	0.0418644	0.121824

General information

General mathematical information about the concept is here ▷ Contingency table
Syntax and semantic of the implementation is here ▷Matlab : `crosstab`

4.14 FISHER test

The FISHER test is used for 2×2 contingency tables where the total frequency is less than or equal to 40. This test is also called the *exact* FISHER test because it calculates the exact probability according to the hypergeometric distribution.

Mental image

height	age of 31 US presidents
small (X)	67,79,80,85,90
tall (Y)	53,56,60,60,63,63,64,64,65,66,67,67,68, 70,71,71,72,73,73,74,77,78,78,83,88,90

Sample	$\geq \tilde{x}$	$< \tilde{x}$
X	1	4
Y	14	12

X = sample of small US presidents
Y = sample of tall US presidents
$\tilde{x}$: median of the total sample $X \cup Y$ is 70.

Figure 38: Motivation for exact FISHER test w.r.t. the natural death year before/after median. Is the death independent of age and height?

Procedure FISHER*'s independence test*

1. **Assumptions** The frequencies of the dichotomous characteristics A and B are counted on $n := a + b + c + d$ subjects forming a random sample; the observations are made independently.

2. **Null hypothesis** H_0 : the characteristics A and B are independent (two-sided)

3. **Test statistics** given the corner distribution

		sum
a	b	a+b
c	d	c+d
a+c	b+d	n

, the test statistics P is

$$P := \sum_{x=0}^{min(a+b,a+c)} \frac{\binom{a+c}{x} \cdot \binom{b+d}{a+b-x}}{\binom{n}{a+b}} = \frac{(a+b)! \cdot (c+d)! \cdot (a+c)! \cdot (b+d)!}{n! \cdot a! \cdot b! \cdot c! \cdot d!}$$

4. **Decision** Reject H_0, if $P < \texttt{hypINV}(\alpha/2)$ or $P > \texttt{hypINV}(1-\alpha/2)$.

Example *Is natural death year independent of age and height?*

We do the example in [6, p.163, 160]. There was a random sample of 31 US presidents drawn, who died of natural causes. The whole sample was divided in tall (X) and small (Y) persons:

	age at year of death												
tall:	67	79	80	85	90								
small:	53	56	60	60	63	63	64	64	65	66	67	67	68
	70	71	71	72	73	73	74	77	78	78	83	88	90

Is the death independent of age and height?

Solution

1. To solve this, both sets of data are combined into a single sample and it is counted how many elements in the sample are above or below the median $\tilde{x} = 70$. This yields the following contingency table:

Sample	$\geq \tilde{x}$	$< \tilde{x}$
X	1	4
Y	14	12

2. Ho: the death independent of age and height
3. n = 31.
4. median$(X \cup Y) =: \tilde{x} = 70$.
5. test statistic $p = 0.1864$ (p-value)
6. Because $p = 0.1864 > 0.05 = \alpha$, reject H_0. So the FISHER test tells us that there is no statistically significant difference in the association of height and age w.r.t. the natural death.

∘ *Solution* and explanation of this example with EIGENMATH in ▷ Fisher-test.

```
fishertest(1,4, 14,12)
```

__ FISHER test __

Ho	p
1	0.18638

General information

General mathematical information about the concept ▷ WIKIPEDIA : Fisher exact test
Syntax and semantic of the implementation is here ▷MATLAB : `fishertest`

4.15 McNemar test

McNemar's test is a non-parametric test used to analyze paired nominal data. It is a test on a 2×2 contingency table and checks the marginal homogeneity of two dichotomous variables. The test requires one nominal variable with two categories (dichotomous) and one independent variable with two dependent groups, cf. *google* ▷ *McNemar test*

Mental image

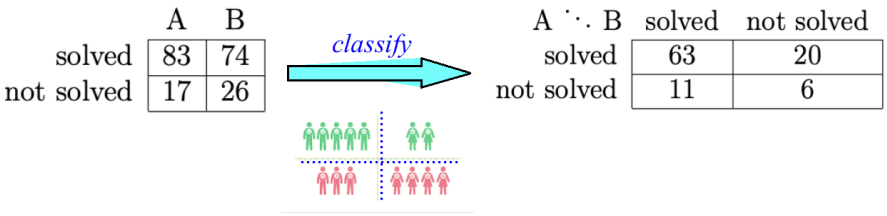


Figure 39: Visualization of McNemar test: first split the outcome 'solved vs. not solved' w.r.t. the groups A an B.
Left: experiment outcome as frequency table $\left(\begin{smallmatrix} a' & b' \\ c' & d' \end{smallmatrix}\right) = \left(\begin{smallmatrix} 83 & 74 \\ 17 & 26 \end{smallmatrix}\right)$.
Right: experiment rearranged as new frequency table $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \left(\begin{smallmatrix} 63 & 20 \\ 11 & 6 \end{smallmatrix}\right)$.
⇒: McNemar needs 'splitted' table `[A\B]`, not 'summary' table `[A|B]`.

Procedure McNemar*'s test*

1. **Assumptions** The frequencies of the dichotomous characteristics A and B are counted on $n := a + b + c + d$ subjects forming a random sample; the observations are made independently.

2. **Null hypothesis** H_0 : the characteristics A and B are equal distributed. (two-sided)

3. **Test statistics** given the '*splitted*' frequency table

A ˙·. B	+	-
+	a	b
-	c	d

   with $n > 40$, the test statistics

$$\chi := \frac{(b-c)^2}{b+c} \quad \text{is approximately } \chi^2_1 \quad \text{distributed.}$$

   Remark: for $b + c < 40$ the Yates correction $\chi := \frac{(|b-c|-1)^2}{b+c}$ is used.

4. **Decision** Reject H_0, if $\chi > \texttt{chiINV}(1; 1-\alpha)$.

Example *Survey of tea and coffee drinking*

We do the example of S. MANGIAFICO ▷ McNemar test. Consider a survey of tea and coffee drinking, in which each respondent is asked both if they drink coffee, and if they drink tea. Is coffee more popular than tea? That is, is it more common for someone to drink coffee and not tea than to drink tea and not coffee?

	Tea	
Coffee	Yes	No
Yes	37	17
No	9	25

Solution

1. $a = 37, b = 17, c = 9, d = 25$.
2. n = 88.
3. $\chi = 1.8846$.
4. p-value $p = 0.1698$
5. Because $p = 0.1698 > 0.05 = \alpha$, we reject H_0.

∘ *Solution* and explanation of this example with EIGENMATH in ▷ McNemar-test.

```
mcnemarTest(37,17,   9,25, "corrected")
```

____ McNEMAR ____

chi	p	Ho
1.88462	0.169811	0

Remark. Neither coffee nor tea is more popular, specifically because neither the 9 nor the 17 in the table are large relative to the other.
$Ho = 0$ means 'reject Ho'.

General information

General mathematical information is here ▷ WIKIPEDIA/GE : McNemar's test
Syntax and semantic of the implementation is here ▷R : `mcnemar.test`

5 Correlation and Bootstrap

We collect the well known measures of correlation using EIGENMATH. We translate the mathematical definitions in EIGENMATH code and demonstrate exemplary calls and show examples.

5.1 PEARSON's ρ Correlation coefficient

The correlation coefficient ρ of two random variables X and Y is a measure of their linear dependence. If each variable has n scalar observations, then the PEARSON correlation coefficient ρ is defined as $\frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y}$.

Mental image

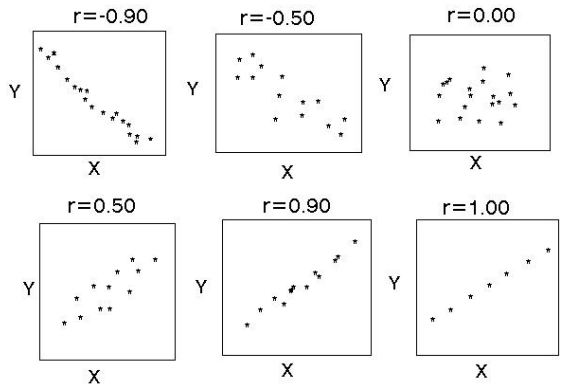


Figure 40: Visualization of PEARSON's correlation coefficient ρ: 6 typical cases. *row*1: from strong negative correlation ('-0.9') until no correlation ('0'). *row*2: from mean positive correlation ('+0.5') until strong positive correlation ('+1').

Procedure PEARSON*'s correlation coefficient for linear correlated samples*

1. **Assumptions** two independent random samples $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ with means $\bar{X}$ resp. $\bar{Y}$

2. **correlation** ρ coefficient is given by

$$\rho(X,Y) := \frac{\sum_{i=1}^{n}(x_i - \bar{X}) \cdot (y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{Y})}} = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

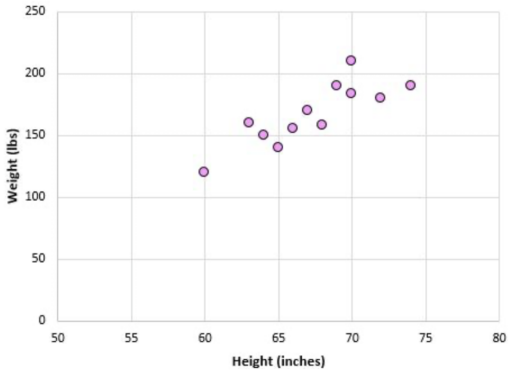
Remark: for $n < 30$ the OLKIN–PRATT correction $\rho := \rho \cdot (1 + \frac{(1-\rho)^2}{2n-6})$ is used.

Example *Height and weight of 12 persons*

We do the example of Z. Bobbitt from statology/pearson. The dataset below shows the height and weight of 12 individuals. The scatterplot of this dataset depicts the values of these two variables. Verify: The Pearson correlation coefficient for these two variables is $\rho = 0.836$.

Solution by hand using ρ-formula.

Height (inches)	Weight (lbs)
60	120
65	140
72	180
70	184
74	190
63	160
66	155
68	158
67	170
69	190
70	210
64	150



1. $\bar{X} = 67.33$ and $\bar{Y} = 167.25$.
2. $cov(X, Y) = 916$.
3. $\sigma_X = 174.67$ and $\sigma_Y = 6874.25$.
4. $\rho = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y} = 0.836$. Positive correlation: As X variable increases, the Y tends to increase as well.

∘ *Solution* and explanation of this example with Eigenmath in ▷ Pearson

```
X = (60,65,72,70,74,63,66,68,67,69,70,64)
Y = (120,140,180,184,190,160,155,158,170,190,210,150)
corrcoef(X,Y)                    | 0.8359
```

General Information

General mathematical information is here ▷ WikipediA : Pearson corrcoef
Syntax and semantic of the implementation is here ▷MatLAB : `corrcoef`

5.2 SPEARMAN's ρ_S rank correlation coefficient

SPEARMAN 's rank correlation coefficient ρ_S is a number ranging from -1 to 1 that indicates how strongly two *sets of ranks*(!) are correlated. The difference between the PEARSON correlation ρ and the SPEARMAN ρ_S correlation is that the PEARSON is most appropriate for measurements taken from an interval scale, while the SPEARMAN is more appropriate for measurements taken from ordinal scales.

Mental image

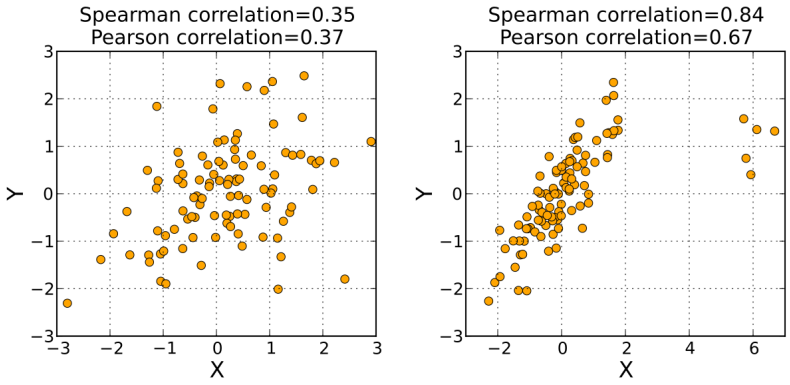


Figure 41: Visualization of SPEARMAN 's rank coeff. ρ_S of two samples X and Y Left: For roughly elliptically distributed data with no prominent outliers, Spearman's ρ_S and Pearson's ρ correlation give similar values. Right: The Spearman correlation is less sensitive than Pearson's to strong outliers, because ρ_S limits the outlier to the value of its rank.
Figures and comments are cited from ▷ WIKIPEDIA : Spearman's ..

Procedure SPEARMAN*'s rank correlation coefficient*

1. **Assumptions** two independent random samples $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$
2. **correlation** ρ_S rank coefficient by SPEARMAN is given by

$$\rho_S(X, Y) := \frac{cov(cX, cY)}{\sigma_{cX} \cdot \sigma_{cY}}$$

where cX resp. cY are the ranks of X and Y *corresponding* to their sample values.

Remark. If X and Y have no 'ties', then on uses the simpler formula

$$\rho_S := 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)}$$

where $d_i := Rank(x_i) - Rank(y_i)$.

Example *Example from exceldemy*

We do the example by R.R. SUPROV in ▷ exceldemy with their data sets.
Calculate SPEARMAN's ρ_S in two ways using both formulas from above.

Solution along the columns of the 'rank' table:

Student Name	Math	Economics	R_{math}	$R_{economics}$
Ali	70	90	10	3
Beatriz	78	94	8	1
Charles	90	79	2	8
Diya	87	86	3	5
Eric	84	84	6	6
Fatima	86	83	4	7
Gabriel	91	88	1	4
Hanna	74	92	9	2
Rodriguez	83	76	7	9
Robert	85	75	5	10

1. $R_{math} \equiv cX$ and $R_{econ} \equiv cY$
2. $cov(cX, cY) = -3.45$.
3. $\sigma_{cX} = 2.8722$ and $\sigma_{cY} = 2.8722$.
4. $\rho_S = \frac{cov(cX,cY)}{\sigma_{cX} \cdot \sigma_{cY}} = -0.4181$.
   Negative correlation: As X variable increases, the Y tends to decrease.

∘ *Solution* and explanation of this example with EIGENMATH in ▷ Spearman.

```
X = (70,78,90,87,84,86,91,74,83,85) -- math
Y = (120,140,180,184,190,160,155,158,170,190,210,150) -- econ
spearman(X,Y)                          | −0.4182
```

General Information

General mathematical information about the concept is here ▷ WIKIPEDIA : spearman
Syntax and semantic of the implementation is here ▷MATLAB : . . . `spearman`

5.3 KENDALL's τ rank correlation coefficient

KENDALL's τ is a non-parametric rank correlation coefficient that measures the similarity between two rankings by assessing the number of concordant and discordant pairs. It ranges from -1 to $+1$, where $+1$ indicates identical rankings, -1 indicates the reverse of the other, and 0 indicates no relationship.Â quote ▷ KENDALL.

Mental image

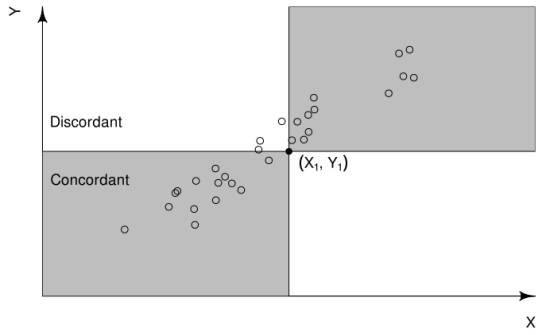


Figure 42: Visualization of KENDALL 's rank corrcoeff τ of two samples X and Y. All points in the gray area are *concordant* and all points in the white area are *discordant* with respect to point (X_1, Y_1). With $n = 30$ points, there are a total of $\binom{30}{2} = 435$ possible point pairs. In this example there are 395 concordant point pairs and 40 discordant point pairs, leading to a Kendall rank correlation coefficient of $\tau = 0.816$.
Figures and comments are cited from ▷ WIKIPEDIA : Kendall's ..

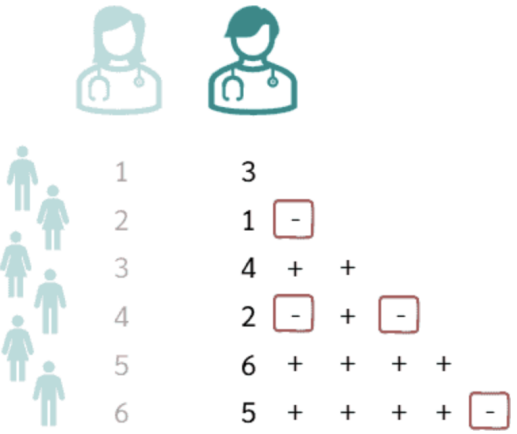
Procedure KENDALL*'s rank correlation coefficient* τ

1. **Assumptions** two random samples $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$
2. **correlation** τ rank coefficient by KENDALL is calculated by these steps:
   a. The data for each sample X and Y is first converted into ranks..
   b. Then examine every possible pair of observations to determine if the ranks of the pair are in the same order ('*concordant*') or a different order ('*discordant*').
   c. The final value τ is based on the difference between the number C of concordant pairs and the number D of discordant pairs:
   $$\tau := \frac{C - D}{C + D}$$
   Positive correlation: As X variable increases, the Y tends also to increase.

Example *Example : two doctors rank 6 patients.*

We do the example from ▷ numiqo. – Suppose two doctors rank 6 patients by descending physical health. One of the two doctors, in this case the female, is now defined as the reference (doc1) and the patients are sorted from 1 to 6.
Calculate KENDALL's τ iusing the formula from above.

Solution along the columns of the 'rank' table in EIGENMATH:



doc1:	doc2:	p:	q:
1	3	3	2
2	1	3	1
3	4	2	1
4	2	2	0
5	6	1	0
6	5	—	—
		11	4

$\tau = \frac{7}{15}$

1st column "-+-++" ↑ gives ... ... $(p, q) = (3, 2)$ in row 2 of matrix.

1. Why $p = 3$? – Because below $doc2 = 3$ are 3 items greater than $doc2 = 3 : 4, 6, 5$.
2. Why $q = 2$? – Because below $doc2 = 3$ there are 2 items smaller than $doc2 = 3 : 1, 2$.
3. $C := \Sigma_p = 3 + 3 + 2 + 2 + 1 = 11$. $D := \Sigma_q = 2 + 1 + 1 + 0 + 0 = 4$.
4. $\tau = \frac{C-D}{C+D} = 0.4666$.
   Positive correlation: As $doc1$ variable increases, the $doc2$ tends also to increase.

∘ *Solution* and explanation of this example with EIGENMATH in ▷ Kendall-τ.

```
X=(1,2,3,4,5,6)
Y=(3,1,4,2,6,5)
kendall(X,Y)
```

___ KENDALL tau ___

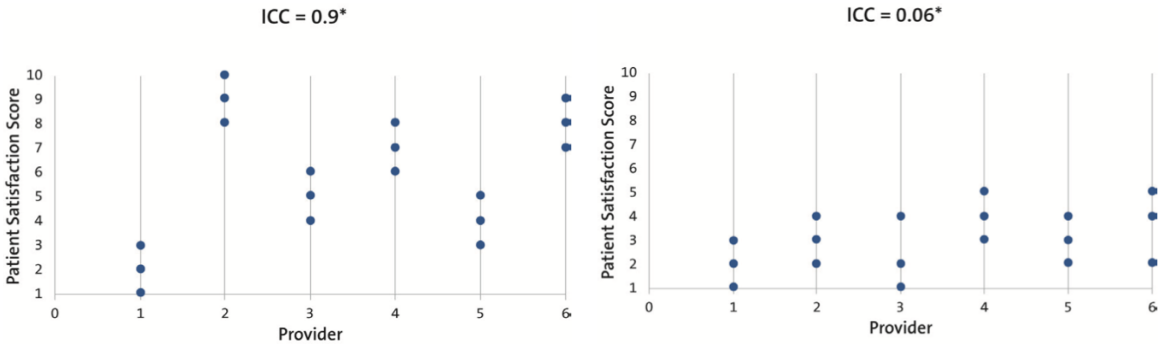
C	D	tau
11	4	0.466667

General Information
∘ General mathematical information about the concept is here ▷ WIKIPEDIA : Kendall
∘ Syntax and semantic of the implementation is here ▷MATLAB : ...`corr()`

5.4 ICC - Intraclass Correlation Coefficient

The intraclass correlation coefficient ICC is a descriptive statistic that measures how strongly units within the same group are similar to each other. It is used for quantitative measurements organized into groups and is often used to assess agreement between multiple raters or measurements, with a value between 0 (no agreement) and 1 (perfect agreement). An ICC helps determine reliability, like in cases where two raters evaluate the same set of patients or a single rater measures the same subjects multiple times. ▷ *Google*

Mental image



Visualization of intraclass correlation coefficient ICC.
Left: Suppose we have 6 providers, each with 3 eligible participants for a pragmatic cluster-randomized trial. The outcome is patient satisfaction rated on a scale from 1 to 10 with an outcome distribution as shown.
Figure 43: Right: Here no patient provides a satisfaction score above 5, the overall variability of the data is lower than in the left figure, and there is much lower between-provider variability in these data. Here, the ICC is lower because the outcomes across different clusters are not likely to be different from each other. – Figures and comments are cited from ▷ NIH : ICC sheet

Procedure `icc` *correlation coefficient*

1. **Assumptions** k correlating variables in n cases with observation matrix $X = (x_{ij})_{i=1..k}^{j=1..n}$
2. **correlation** `icc` coefficient is given by

$$icc(X) := \frac{MQ_{row} - MQ_{rest}}{MQ_{row} + (k-1) \cdot MQ_{rest}}$$

where we use the following 9 terms

$$\begin{array}{lll} T_j := \sum_{i=1}^{k} x_{ij} & S := \sum_{j=1}^{k} T_j & SAQ_{total} := \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}^2 - \frac{S^2}{k \cdot n} \\ SAQ_{row} := \frac{1}{k} \cdot \sum_{j=1}^{n} T_j^2 - \frac{S^2}{k \cdot n} & df_{row} := n - 1 & MQ_{row} := \frac{SAQ_{row}}{df_{row}} \\ SAQ_{rest} := SAQ_{total} - SAQ_{row} & df_{rest} := n \cdot (k-1) & MQ_{rest} := \frac{SAQ_{rest}}{df_{rest}} \end{array}$$

Example *Example and text from* ZOEFEL[19, p.133 ff]
Fourteen women were asked about their weight; the given answers were then compared to the actual measured weight. The measured values yield a mean of 57.6 kg, while the mean of the estimated values is 54.6 kg. The actual weight values are therefore significantly underestimated. The `icc` attempts to combine both aspects, the correlation and the differences in the means, into a single measure. `icc` only achieves high values, when both the direction and the levels of the variables involved match. We have the data

$$X = \begin{pmatrix} 48 & 48 & 50 & 50 & 52 & 58 & 63 & 56 & 48 & 63 & 58 & 58 & 52 & 60 \\ 51 & 50 & 50 & 52 & 53 & 63 & 70 & 70 & 49 & 63 & 59 & 58 & 56 & 62 \end{pmatrix} \begin{matrix} \text{weight estimated} \\ \text{weight measured} \end{matrix}$$

Calculate the `icc` of X using the formulas from above.

Solution along the 9 constituting terms of the `icc`:

1. $k = 2$ and $n = 14$ and $T_1 = 99, T_2 = 98, \ldots$ and $S = 1570$.
2. $SAQ_{total} = 1131.9$.
3. $SAQ_{row} = 976.9$ and $df_{row} = 13$ and $MQ_{row} = 75.1$.
4. $SAQ_{rest} = 155$ and $df_{rest} = 14$ and $MQ_{rest} = 11.07$.
5. $\texttt{icc}(X) = 0.743$.

∘ *Solution* and explanation of this example with EIGENMATH in ▷ ICC.

```
| X=(48,48,50,50,52,58,63,56,48,63,58,58,52,60)
| Y=(51,50,50,52,53,63,70,70,49,63,59,58,56,62)
| icc(X,Y)                    | 0.743
```

∘ *Exercise*: Write the 'early ICC formula' ▷ iccOld in EIGENMATH and solve the example above with it. Compare both solutions.

General Information
General mathematical information about the concept ▷ WIKIPEDIA : ICC
Syntax and semantic of the implementation ▷R : `icc`

5.5 `regression` - the linear regression line

Simple Linear regression models the relation between a dependent, or response, variable y and an independent, or predictor, variable x and considers the independent variable using the relation $y = ax + b$, where b is the y-intercept, and a is the slope (or regression coefficient), i.e. there is a linear relation between x and y ($x \sim y$).

Mental image

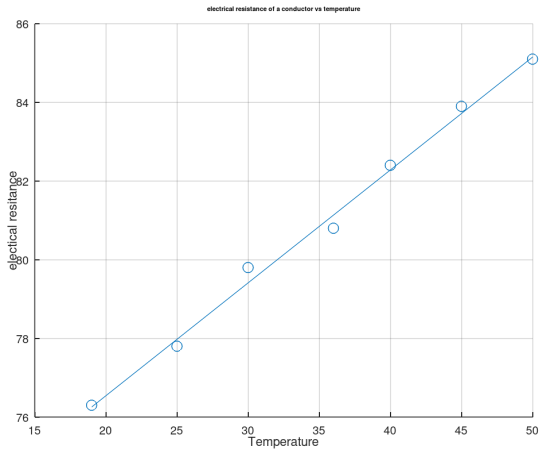


Figure 44: Visualization of linear regression (line):
electrical resistance of a conductor as a function of temperature.
–: regression line $y = 0.28 \cdot x + 70.8$ of the measurement $X \mapsto Y$ with
○: $(19, 25, 30, 36, 40, 45, 50) \mapsto (76.3, 77.8, 79.8, 80.8, 82.4, 83.9, 85.1)$.

Procedure *linear regression line* $y = ax + b$

1. **Assumptions** a 2-dimensional sample (X, Y) with a linear connection $y_i \approx a \cdot x_i + b$.
2. **regression line** coefficients a and b are calculated by

$$\begin{aligned} a &:= \frac{\sum_{i=1}^{n}(x_i - \bar{X}) \cdot (y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2} \\ b &:= \frac{1}{n} \cdot (\bar{Y} - a \cdot \bar{X}) \end{aligned}$$

where $\bar{X}$ resp. $\bar{Y}$ are the mean values of $X = (x_1, ..., x_n)$ resp. $Y = (y_1, ..., y_n)$.

Example *electrical resistance of a conductor as a function of temperature*

We do the example of figure.44, cf. [6, p.193]. We have the electrical resistance (Y) of a conductor as a function of temperature (X):

Temperature:	19	25	30	36	40	45	50
Resistance:	76.3	77.8	79.8	80.8	82.4	83.9	85.1

Calculate the equation $y = ax + b$ form the measurement.

Solution

1. n = 7.
2. meanX = 35 and meanY = 80.87.
3. numerator of $a = cov(X, Y) = 210.5$
4. denumerator of $a = \Sigma(X - \text{meanX})^2 = 732$
5. $a = \frac{numerator}{denumerator} = 0.287$
6. $b = 70.80$

∘ *Solution* and explanation of this example with EIGENMATH in ▷ Pearson-regression.

```
X = (19,25,30,36,40,45,50)
Y = (76.3, 77.8, 79.8, 80.8, 82.4, 83.9, 85.1)
regression(X,Y)
```

y = a*x + b

a	b
0.287568	70.8065

General Information

General mathematical information is here ▷ WIKIPEDIA : Linear regression
Syntax and semantic of the implementation is here ▷R : `lm`

5.6 `anova1` - One-way Analysis Of Variance

One-way ANOVA is a statistical test used to compare the means of three or more independent groups to determine if there is a statistically significant difference between them. It works by partitioning the total variability in the data into two sources: the variability between the groups and the variability within the groups. If the variability between groups is large compared to the variability within groups, it suggests that the group means are significantly different.

Mental image

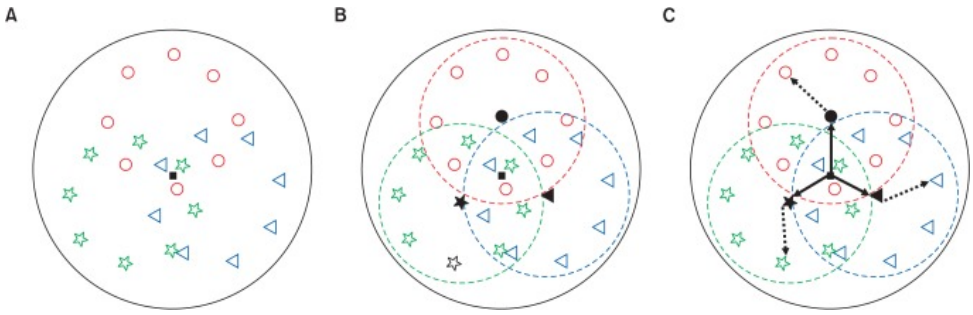


Figure 45: Visualization of *One-way Analysis Of Variance* test of three groups.
A: A solid black square "□" is suggested as a general representative value such as mean of overall data.
B: It looks reasonable to divide the data into three groups and explain the data with three different means of groups: •, ⋆, △.
C: To evaluate the efficiency or validity of dividing three groups, the distances from group means to overall mean and the distances from group means to each data are compared. Distance between group means and overall mean (solid arrows) stands for the inter-group variance and distance between group means and each group data (dotted arrows) stands for the intra-group variances. – Figures and comments are cited from T.K. Kim ▷ NIH : ANOVA

Procedure *One-way Analysis Of Variance*

1. **Assumptions** a data matrix ('table') $X = [x_{ij}]_{j=1..b}^{i=1..a}$, where x_{ij} denotes the observation no. j in group i. The rows of the data matrix are the 'treatments'.

2. **Null hypothesis** H_0 : the variances of all groups are equal (two-sided)

3. **Test statistics** `anova1(X)` is F-distributed with $df_1 = a - 1$ and $df_2 = a \cdot (b - 1)$ degrees of freedom and is given by

$$F := \frac{MS_B}{MS_W} = \texttt{anova1(X)}$$

where we use the following 'ANOVA1 table' to calculate F

calculate from left to right:	Variation	df	Mean square	F
Between treatments	$V_B := \frac{1}{b} \cdot \sum_j T_j^2 - \frac{T^2}{ab}$	$a-1$	$MS_B := \frac{V_B}{a-1}$	$F := \frac{SS_B}{SS_W}$
Within treatments	$V_W := V - V_B$	$a \cdot (b-1)$	$MS_W := \frac{V_W}{a(b-1)}$	
Total treatments	$V := \sum_{j,k} x_{jk}^2 - \frac{T^2}{ab}$	$ab-1$		ANOVA1

and the helper terms $T := \sum_{i,j} x_{ij}$ = the total sum of all values in X
and $T_j := \sum_{i=1}^{a} x_{ij}$ = the total sum of all values in the jth treatment (row) of X.

4. **Decision** Reject H_0, if $\chi >$ `chiINV`$(1; 1-\alpha)$.

Example *Yields in bushels per acre*

We do the example 16.4 of M. SPIEGEL [13, p.415]. The table shows the yields in bushels per acre of a certain variety of wheat grown in a particular type of soil treated with chemicals A,B and C. Do an ANOVA1 analysis of the treatments.

```
A:  3  4  5  4
B:  2  4  3  3
C:  4  6  5  5
```

∘ *Solution* and explanation of this example with EIGENMATH in ▷ ANOVA1.

```
X = ((3,4,5,4),(2,4,3,3),(4,6,5,5))
Y = (76.3, 77.8, 79.8, 80.8, 82.4, 83.9, 85.1)
anova1(X)
```

Source	df	SS	MS	F
Between	2	8	4	6
Within	9	6	$\frac{2}{3}$	-
Total	11	14	-	ANOVA

General Information
∘ General mathematical information is here ▷ WIKIPEDIA : One-way analysis of variance
∘ Syntax and semantic of the implementation is here ▷MATLAB : `anova1`

5.7 `boot1` - the bootstrap method for dependent samples

The purpose of bootstrapping in statistics is to estimate the sampling distribution of a statistic by repeatedly resampling from the original data. This allows statisticians to perform hypothesis testing, calculate standard errors, and construct confidence intervals without needing to make strong assumptions about the underlying data distribution, a task that can be complex or impossible with traditional methods.

Mental image

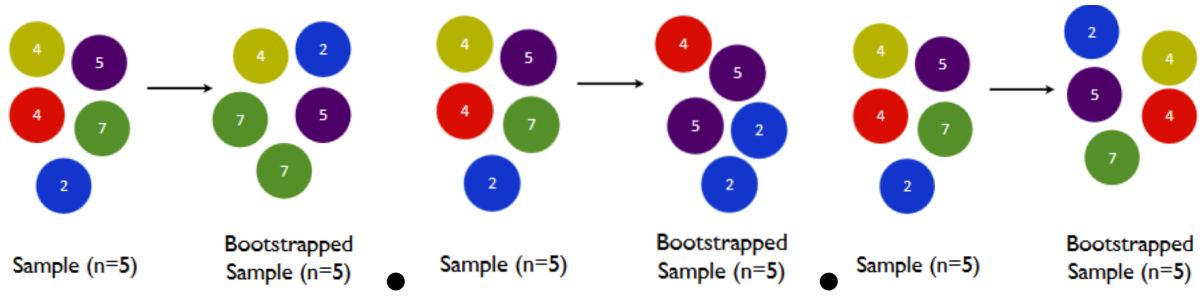


Figure 46: Visualization of `bootstrap`.
Left: From sample $X = \{4, 5, 4, 7, 2\}$ with mean $\bar{X} = 4.4$ we draw a new ('bootstrapped') sample giving $X' = \{4, 2, 7, 5, 7\}$ with mean $\overline{X'} = 5$.
Middle: From the original sample X we draw ('resample') a new sample $X'' = \{4, 5, 5, 2, 2\}$ mean $\overline{X''} = 3.6$
Right: From the original sample X we resample again a new sample $X''' = \{2, 4, 5, 4, 7\}$ with mean $\overline{X'''} = 4.4$.
From the resamples, the statistic T (here: *mean*) is calculated with $T = mean(\overline{X'}, \overline{X''}, \overline{X'''}) = 4.33$ to estimate the distribution of T.
▷ Figures are cited from ▷ G.J. Scott: The bootstrap

Procedure *the bootstrap*

1. **Assumptions** a data set (sample) $X = (x_1, \ldots, x_n)$
2. **bootstrap** The bootstrap method is summarized by the following steps:
   Choose a statistics T to be studied, e.g. $T := mean$.
   a. Choose the number n of bootstrap samples to take.
   b. For each bootstrap sample, draw a replacement sample X_i of the size n you selected.
   c. Calculate the statistics T_i for the samples X_i.
   d. Find a summary statistic T (called a bootstrap statistic) for each of the n samples T_i.

Example *bootstrap the test statistic sd*

We follow the toy example of Figure.46. Consider the sample $X = \{4, 5, 4, 7, 2\}$. Using the bootstrap samples X_1, X_2, X_3 from above, estimate the standard deviation of the bootstrap distribution.

Solution We follow the recipe bootstrap.

1. Choose $T = sd = \frac{\sqrt{\sum (X-\bar{X})^2}}{\sqrt{length(X)-1}}$ with $length(X) = 5$.
2. $n = 3$
3. $T_1 := sd(X_1) = 2.1213$, $T_2 := sd(X_2) = 1.5165$, $T_3 := sd(X_3) = 1.8165$.
4. $T = mean(T_1, T_2, T_3) = mean(2.1213, 1.5165, 1.8165) = 1.8181$

∘ *Solution* and explanation of this example with EIGENMATH in ▷ bootstrap1.

```
X = (4,5,4,7,2)
boot1(X,1000)
```

boot1

est.mean	est.sd
4.4158	0.526604

General Information

General mathematical information is here ▷ WIKIPEDIA : Bootstrapping
Syntax and semantic of the implementation is here ▷MATLAB : `bootstrp`

5.8 `boot2` - the bootstrap method for independent samples

The bootstrap method for independent samples X and Y is a resampling technique that estimates the sampling distribution of a statistic θ ($\overset{e.g.}{=}$ *mean*) without relying on strong assumptions like normality. It involves repeatedly drawing samples with replacement from each of the two original, independent samples X and Y, calculating the statistic of interest like the *difference in means for each resampled pair*, and using the resulting distribution of these statistics to e.g. to calculate the estimated difference in means for X and Y..

Mental image

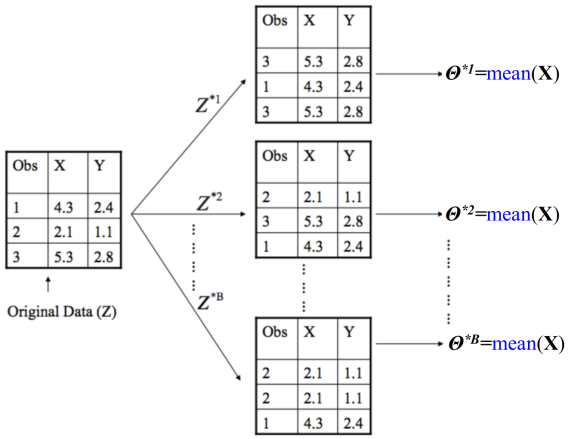


Figure 47: Visualization of `bootstrap2`. We generate new samples Z^{*1}, Z^{*2}, Z^{*B} directly from the population `obs`ervations (leftmost histogram; *obs*).
If we choose $\theta \overset{p.d.}{=} mean$; then $\bar{X} = 3.83$ and $\bar{Y} = 2.1$.
We generate the distribution of sample means B-times, e.g. $B = 3$.
Z^{*1}: $\theta^{*X1} = \bar{X}^1 = mean(5.3, 4.3, 5.3) = 4.96$ and $\theta^{*Y1} = \bar{Y}^1 = 2.33$.
Z^{*2}: $\theta^{*X2} = \bar{X}^2 = 3.9$ and $\theta^{*Y2} = \bar{Y}^2 = 2.1$.
Z^{*B}: $\theta^{*XB} = \bar{X}^B = 2.83$ and $\theta^{*YB} = \bar{Y}^B = 1.53$.

Procedure *bootstrap2 method for independent samples*

1. **Assumptions** two data sets (samples) $X = (x_1, \ldots, x_m)$ and $Y = (y_1, \ldots, y_n)$.
2. **bootstrap2** The bootstrap2 method is summarized by the following steps:
   Choose a statistics θ to be studied resp. estimated, e.g. $\theta := mean$. [16]
   a. *Choose the number* B of bootstrap samples to take.
   b. *Resample with replacement*: Create new 'bootstrap' samples by drawing observations with replacement from each of your original, independent samples.[17]

[16] For independent samples, this is often the difference between the two sample means.
[17] For example, if you have two samples, one of size m and one of size n, you will create b pairs of new samples, each with size m and n respectively.

c. *Calculate the statistic*: For each of the B pairs of bootstrap samples, calculate the statistic θ_i of interest. This gives you a collection of B values for your statistic.

d. *Use the bootstrap distribution*: The B calculated statistics form a *bootstrap distribution* approximating the true sampling distribution of the statistic θ.

e. *Summarize the distribution*: Use the bootstrap distribution to estimate the difference in the means..

Example *the bootstrap mean of Figure.46*

We follow the toy example of Figure.46.
Calculate the bootstrap statistics θ^{*X} and θ^{*Y} and their difference.

Solution

1. The Bootstrap stats: $\theta^{*X} = mean(\theta^{*X1}, \theta^{*X2}, \theta^{*X3}) = 3.90$
2. $\theta^{*Y} = 1.98$
3. The difference is $d = \theta^{*X} - \theta^{*Y} = 1.92$, ergo X and Y are independent.

∘ *Solution* and explanation of this example with EIGENMATH in ▷ bootstrap1.

```
X = (4.3, 2.1, 5.3)
Y = (2.4, 1.1, 2.8)
boot2(X,Y)
```

3.9

2.1

D = 1.8

boot2

e.meanX	e.meanY	significance:
4.63333	3.06667	0.0599

General Information

General mathematical information is here ▷ WIKIPEDIA : Bootstrapping
Syntax and semantic of the implementation is here ▷MATLAB : `bootstrp` and ▷R : `bootstrap`

5.9 `bootCI` - Confidence Interval using bootstrap

Bootstrapping is a statistical resampling technique that uses an existing dataset to estimate the properties of an estimator, such as constructing an confidence interval. These '*bootstrap samples*' are then used to build an approximate sampling distribution for the statistic of interest, which helps in situations where analytical methods are complex or difficult to apply. For example, *to find a confidence interval*, you take an initial sample, then repeatedly create new bootstrap samples of the same size by randomly drawing from your original sample with replacement. You calculate the mean for each of these new samples and use the distribution of these means to find the confidence interval, such as the $2.5th$ and $97.5th$ percentiles for a 95% confidence interval.

Mental image

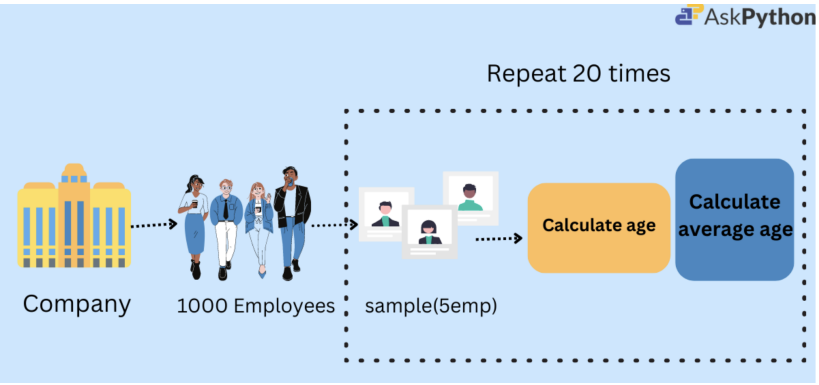


Figure 48: Visualization of `bootCI`: Instead of taking multiple samples directly from the whole employees population, we only draw a single, representative *sample* (5emp). We then generate e.g. 20 'new' samples by *repeatedly* create new bootstrap samples of the same size. – Picture found at ▷ ASKPYTHON : bootstrap-sampling

Procedure `bootstrap` *confidence interval*

1. Choose a statistics T to be studied, e.g. $T := mean$.
2. Choose the number n of bootstrap samples to take.
3. For $i = 1, ..., n$, draw a replacement sample X_i ('bootstrap sample') of size n.
4. Calculate the statistics T_i for the samples X_i, e.g. $T_i := mean(X_i)$.
5. For the k th percentile with $k \in \{1, ..., 99\}$, the k *percent confidence interval* for the summary statistic $T := T_1, \ldots, T_n$ [18] of the n resamples T_i is defined by:

$$CI_k := \left(percentile(T, \frac{100-k}{2}), \quad percentile(T, \frac{100+k}{2}) \right)$$

[18] the 'bootstrap statistic', e.g. the vector of the bootstrap means T_i

Example *The 'lemployees' example from askpython*

We follow the example of Figure.48 by hand on the toy example by A. YADAV for the fictive ages sample $5emp = (25, 30, 35, 40, 45, 50, 55, 60, 65, 70)$.
Calculate the 95% confidence interval for the mean of $5emp$.

Solution We do only $n = 3$ resamples to show the principle of the procedure.

1. $X := (25, 30, 35, 40, 45, 50, 55, 60, 65, 70) = 5emp$, $n = 3$. So we iterate 3-times:
2. $X_1 = (40, 60, 40, 35, 55, 65, 70, 45, 35, 25), T_1 = mean(X_1) = 47.$
3. $X_2 = (70, 60, 40, 55, 55, 40, 50, 25, 50, 55), T_1 = 50.$
4. $X_3 = (70, 70, 50, 70, 30, 50, 30, 30, 55, 60), T_1 = 51.5.$
5. $T(X) := mean(T_1, T_2, T_3) = mean(47, 50, 51.5) = 51.16.$
6. $CI_{95} = (47; 51.5) =:$(CI.l,CI.h) – this interval catch 51.16 with no surprise.

∘ *Solution* and explanation of this example with EIGENMATH in ▷ bootstrapCI.
We check the solution using our function `bootci` and 100 resamples:

```
X = (25, 30, 35, 40, 45, 50, 55, 60, 65, 70)
alpha = 0.05
bootci(X,100)
```

bootCI

est.Mean	CI.l	CI.h
48.02	38.5	50

∘ The solution using `Python` is here ▷ *AskPython*.

General Information

Syntax and semantic of the implementation is here ▷R : `boot.ci`
or here ▷MATLAB : `bootci`

5.10 `jackknife` - The Jackknife method

Jackknife is a statistical resampling technique that estimates the bias and variance of a statistic θ *by repeatedly leaving out one observation at a time* from the original sample $X = (x_1, x_2, ..., x_n)$. This creates n smaller subsamples X_i, each with one observation removed. The statistic of interest is then calculated for each subsample, and the resulting set of estimates θ_i is used to derive bias-corrected estimates and confidence intervals for the original statistic. ▷ Google AI : Jackknife resampling

Mental image

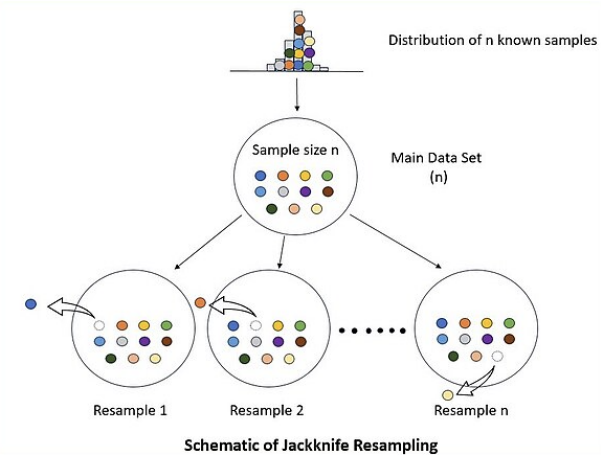


Figure 49: Visualization of `jackknife`: The 'leave-one-out' jackknife algorithm. Instead of taking multiple samples directly from the population, we only have a single, representative *sample*. We then generate 'new' samples by *repeatedly leaving out one observation at a time* from the original *sample*. – Picture found at ▷ WikipediA : Jackknife resampling

Procedure `jackknife`*'s resampling*

- The jackknife method in words:

A *jackknife sample* is a 'leave-one-out' resample of the data. If there are n observations, then there are n jackknife samples, each of size n. If the original data are $X = (x_1, x_2, ..., x_n)$, then the ith jackknife sample is $X_i := (x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$. You then compute n jackknife replicates. A *jackknife replicate* is the statistic of interest (e.g. estimate of the standard error) computed on a jackknife resample.

- The jackknife method is summarized by the following steps:

  a. Compute a statistic, θ, on the original sample of size n, e.g. $\theta = \bar{X}$ =mean(X).

  b. For $i = 1$ to n, repeat the following:

  ▷ Leave out the ith observation x_i from X to form the new ith jackknife sample X_i.

▷ Compute the ith jackknife replicate statistic, θ_i, by computing the statistic on the ith jackknife sample X_i, e.g. $\theta_i := \bar{X}_i$ as estimate of mean $\bar{X}$ of X

▷ Compute the mean of the jackknife replicates: $\bar{\theta_i} := \frac{1}{n} \cdot \sum_{i=1}^{n} \theta_i$.
Estimate the bias $J_i := n\theta - (n-1)\bar{\theta_i}$, the 'pseudovariable's.

▷ Estimate the standard error $\mathrm{SE}_\theta := \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} (\theta_i - \bar{\theta})^2}$

- **Test statistics** Let σ_J be the standard deviation of the *pseudovariable*'s J_i.
Then we have

$$T := \frac{\sqrt{n} \cdot (J_\theta - \theta)}{\sqrt{\sigma_J^2}} \quad \text{is approximately } t\text{-distributed.}$$

Example *The 'leave-one-out' jackknife*

We follow the jackknife procedure by hand on the toy example $X = (1, 2, 3, 4)$.
Calculate the standard error of the pseudovariables J_i.

Solution

1. $X = (1, 2, 3, 4), n = 4$. So we iterate 4-times.
2. $stats(y) := mean(y) = \bar{y}$, defines the statistics of interest.
3. $\theta := stats(X) = mean(X) = 2.5$.
4. calculate pseudovariables J_i:
   $J_1 : \quad X_1 = (2, 3, 4) \to \theta \cdot n - (n-1) \cdot stats(X_1) = 2.5 \times 4 - 3 \times 3 = 1$
   $J_2 : \quad X_2 = (1, 3, 4) \to \theta \cdot n - (n-1) \cdot stats(X_2) = 2.5 \times 4 - 3 \times 8/3 = 2$
   $J_3 : \quad 1\ 2\ 4 \to \cdots = 3$
   $J_4 : \quad 1\ 2\ 3 \to \cdots = 4$
5. $\bar{J} = mean(J_1, J_2, J_3, J_4) = mean(1, 2, 3, 4) = 2.5$
6. bias $= \bar{J} - \theta = 4 - 4 = 0$
7. SE $= \sqrt{\frac{variance(J)}{4}} = 0.6455$

◦ *Solution* and explanation of this example with EIGENMATH in ▷ jackknife.

```
X = (1,2,3,4)
alpha = 0.05
jack(X)
```

jackknife

SE	CI.l	CI.h
0.645497	4.55426	0.44574

◦ The corresponding *R* code is partially found at ▷ jackknifing

General Information

General mathematical information is here ▷ WIKIPEDIA : Jackknife resampling
Syntax and semantic of the implementation is here ▷MATLAB : `jackknife`

6 Appendix - the statsbox

- If you want to run the EIGENMATH functions from this book for own work, you can load them all with

  ▷ statsbox.html.

- If you want to run the EIGENMATH functions from this text using the iMac `eigenmath.app`, you can extract all functions from the above file `statsbox.html` or downoad them into your work folder with

  ▷ statsbox.txt

  and then use the EIGENMATH function `run("/Users/yourFolder/statsbox.txt")` to load the functions into the memory.

7 Bibliography

References

[1] Beucher, O. (2005): *Wahrscheinlichkeitsrechnung und Statistik mit MATLAB.* Berlin: Springer.

[2] Bognar, M.: *Probability Distribution Applets.* url: `https://homepage.divms.uiowa.edu/~mbognar/`

[3] Dalgaard, P. (2002): *Introductory Statistics with* `R`. New York: Springer.

[4] DIALEKT-Projekt (2000): *Statistik interaktive! - Deskriptive Statistik.* Berlin: Springer.

[5] Heiberger, R.M., Holland, B. (2004): *Statistical Analysis and Data Display - An Intermediate Course with Examples in S-Plus, R and SAS.* New York: Springer.

[6] Herrmann, D. (1991): *Statistik in C.* Braunschweig: Vieweg.

[7] Herrmann, D. (1991): *C++ für Naturwissenschafteler.* Bonn: Addison-Wesley.

[8] LibreOffice (2025): *LibreOffice Calc.* url: `https://www.libreoffice.org/download/download-libreoffice/`

[9] Lindner, W. (2025): *Diverse Memos, e.g. APOS Theory and A.C.E. Teaching Cycle.* url: `https://lindnerdrwg.github.io/`

[10] Press, W.H., Flanery, B.P., Teucholsky, S.A., Vetterling, W.T. (1988): *Numerical Recipes in C.* Cambridge: Cambridge University Press.

[11] SoftMaker (2024): *PlanMaker Free 2024.* url: `https://www.freeoffice.com/en/?option=com_content&view=article&id=13&Itemid=129&tmpl=component`

[12] Rose, C. & Smith, M.D. (2002): *Mathematical Statistics with* Mathematica. New York: Springer.

[13] Spiegel, M.R. & Stephens, L.J. (^{4}2011): *Statistics.* New York: McGraw-Hill.

[14] Sprott, J.C.(1989): *Numerical Recipes - Routines and Examples in BASIC.* Cambridge: Cambridge University Press.

[15] Venables, W.N., Ripley, B.D. (^{3}1999): *Modern Applied Statistics with S-Plus.* New York: Springer.

[16] Vetterling, W.T., Teucholsky, S.A., Press, W.H., Flanery, B.P.,(1989): *Numerical Recipes Example Book (Pascal).* Cambridge: Cambridge University Press.

[17] Weigt, G. (2025): Eigenmath *online Demo.*
url: `https://georgeweigt.github.io/eigenmath-demo.html`

[18] Weigt, G. (2025): Eigenmath *Manual.*
url: `https://georgeweigt.github.io/eigenmath.pdf`

[19] Zöfel, P. (1991): *Statistik verstehen.* Bonn: Addison-Wesley.

⋈

Dr. Wolfgang Lindner
Leichlingen, Germany
dr.w.g.Lindner@gmail.com
2026