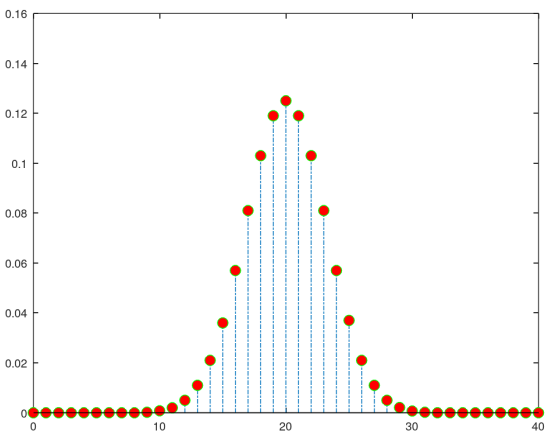
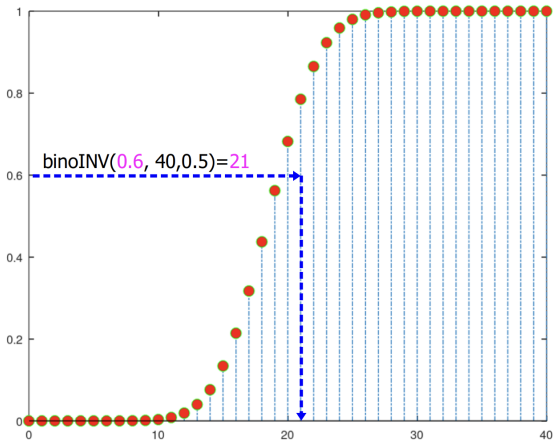
Exploring Math with Yamwi

# Statistics - A Maxima Companion




**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 Maxima 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 Maxima 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 [11].

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 Maxima language allows to make the concepts of statistics 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.

**Why statistics with Maxima?**
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 Maxima ...

- allows to program the statistics formulas very close to their mathematical formulation,
- each worksheet is totally independent of other imported code, it's self-contained,
- it's easy to include comments, tests, tables and plots in the worksheet,
- .. and it is just fun and illuminating to use Maxima in it's new Yamwi version..

**Some remarks on the content of this booklet**
The order of the following suite of worksheets with definitions, checks, exercises (problems) etc. follows a standard presentation.

**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 [17].

**2** We implement some important statistical distributions as helper functions to enable 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 MAXIMA notation for
  - the probability density functions $f$ named "...**PDF**",
  - the cumulative distribution functions $F$ named "....**CDF**" and
  - the quantile functions $F^{-1}$ named "....**INV**", i.e. the INVerse of the CDF.
- often a check of the calculations against the statistics software R or MATLAB,
- a short table of the distributions values,
- a qualitative plot to have a visual impression or to verify one's result on a graph,
- solution of a prototypical problem resp. application followed by a set of exercises.

To run an MAXIMA sheet click on a link starting with  .

**3** Then we implement some standard functions of test statistics, i.e. *Gauss tests*, STUDENT *t-tests*, *F-test*, *ChiSquare test*, WILCOXON *tests*, etc. The user should easily follow these calculations in standard texts like [17] or in free spreadsheets. All these calculations are demonstrated and recalculated as examples using MAXIMA functions.

**4** Last we implement some rank'ing tests like SPEARMAN *rank correlation*, do *regression* and *one-way ANOVA* and close with *bootstrap* and *jackknife* methods. Again, these calculations are demonstrated and calculated using small self-made MAXIMA functions.

**Some remarks on the didactical contcept of this booklet**
This small booklet is dedicated to the novice – w.r.t. the subject matter as well w.r.t. the use of CAS MAXIMA. Therefore, I roughly follow the educational APOS theory, cf. [11]: 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.

**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 MAXIMA. 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, allowing to *observe the effects of own actions* in the CAS MAXIMA.

[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.

**Remark.** The presented code in this booklet is 'educational' code, aimed at the novice learner who is willing to learn statistics and a CAS like MAXIMA simultaneously. I have tried to program all concepts *as small as possible*, 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) aka OCTAVE or the academical strength code from R.

The code in the MAXIMA statistics packages *descriptive*, *distrib* and *stats* is professional code contributed by Mario M. RIOTORTO. Use these functions, if my educational code runs into problems (e.g. I have dropped checks against input errors by the user etc.)

For the inspection or running an MAXIMA script no installation is necessary, *everything runs directly online*: a click on a link like ▷ Click here. in this text is enough to invoke the corresponding script and execute it automatically.
To profit from these features, we benefit from the excellent online version of MAXIMA, called YAMWI, which is great work by M. RIOTORTO, Leo BUTLER and M. GOSSE.
The code lines were tested on an `iMac 24"`, Apple M3, Tahoe 26.3.1 using Safari.

I am very grateful to Dr. Leo R. BUTLER[2] and Michel GOSSE[3] for their friendly support with tips and hints while writing these notes.[4]

Any feedback from the user is very welcome.

Wolfgang Lindner
Leichlingen, Germany
July 2026

[2]University of Manitoba, for sharing his expert knowledge of MAXIMA constructions
[3]Honorary Mathematics Inspector, for finding hidden bugs and mistakes
[4]This booklet is based on my earlier book [10], where I use the small compact CAS EIGENMATH [19] designed for the use by physicists.

# 1 Desriptive Statistics

We collect the well known measures of tendency, dispersion and deviation. We translate the mathematical definitions into MAXIMA 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 [17].
All these calculations could now simply be done with MAXIMA.

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

## 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

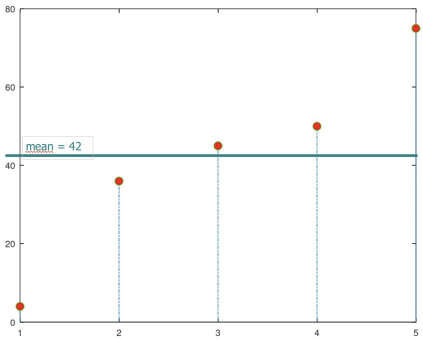
$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n} =: mean(X)$$

**Example** [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.

**Mental image**

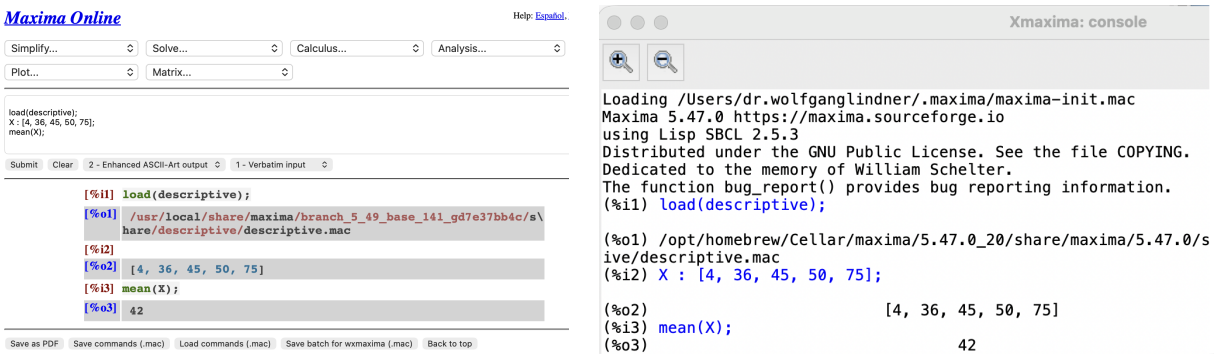



**Use Maxima** Copy the following 3 commands to YAMWI's console (figure.1 left) or e.g. in XMAXIMA's console (figure.1 right) line-by-line or use the direct link below.

```
load(descriptive);
L : [4, 36, 45, 50, 75];
mean(L);
```

```
| 42
```

▷ `mean`
You get:




Left figure: Screenshot of YAMWI showing code and output lines.
Figure 1: Right figure: Screenshot of XMAXIMA with input lines in blue.
The inputs are shown in the `(%i..)` lines, the outputs in the `(%o..)`.

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

### 1.1.1 Exercises

**Exercise 1.** The check with MAXIMA of the example invokes and then uses the package `descriptive` by M.M. RIOTORTO. Write a user-defined function `mean1` using the mathematical definition of *mean* and
**a.** the build'in function `apply`,
**b.** the build'in function `lsum` ('list sum').
*Solution*: `Ex.1`

**Exercise 2.** The definition, which comes next to the mathematical definition of *mean*, use the build'in function `sum`. Write `mean2()` using `sum`.
*Solution*: `Ex.2`

```
[%i1] M : [1,2,3];
[%o1] [1, 2, 3]
[%i2] mean1(L) :=sum( L[i],i,1,length(L))  /length(L);
                    sum(L , i, 1, length(L))
                         i
[%o2] mean1(L) := ------------------------
                          length(L)
[%i3] mean1(M);
[%o3] 2
```

**Exercise 3.** (Spiegel, p.74, P 3.19 - grouped data with guessed mean A)
Calculate the arithmetic mean of the numbers $5, 8, 11, 9, 12, 6, 14, 10$ choosing as the 'guessed mean' the number $A = 9$.

*Solution*: `Ex.3`

```
[%i1] X : [5,8,11,9,12,6,14,10];
[%o1] [5, 8, 11, 9, 12, 6, 14, 10]
[%i2] X-9;
[%o2] [- 4, - 1, 2, 0, 3, - 3, 5, 1]
[%i3] apply("+", X-9);
[%o3] 3
[%i4] 9+apply("+", X-9);
[%o4] 12
[%i5] meanA(X,A)    :=float(A+apply("+", X-A))/length(X);
                                 float(A + apply("+", X - A))
[%o5] meanA(X, A) := ----------------------------------------
                                      length(X)
[%i6] meanA(X,9);
[%o6] 1.5
```

**Exercise 4.** (mean w.r.t a frequency)
100 students took part in a test. The marks X = (61,64,67,70,73) were distributed with the frequency f = (5,18,42,27, 8). In other words, we had the frequency tableau $\begin{pmatrix} 61 & 64 & 67 & 70 & 73 \\ 5 & 18 & 42 & 27 & 8 \end{pmatrix}$, i.e. 5 students get 61 marks, 18 students get 64 marks etc. Calculate the mean mark of the 100 students.
*Hint*: use formula `meanF(X,f) = dot(f,X)/sum(f)` for a mean of $X$ w.r.t. a frequency $f$.

*Solution*: `Ex.4`

```
[%i1] Sum(M) := apply("+",M)$
[%i2] f :[ 5,18,42,27, 8] $
[%i3] Sum(f);
[%o3] 100
[%i4] X : [61,64,67,70,73]$
[%i5] (X.f)/Sum(f), float;
[%o5] 67.45
[%i6] dot(a,b) := sum (a[i]*b[i], i, 1, length(a));
                          i  i
[%o6] dot(a, b) := sum(a  b , i, 1, length(a))
[%i7] dot(X,f);
[%o7] 6745
[%i8] meanF(X,f) := float(dot(X,f)/Sum(f))$
[%i9] meanF(X,f);
[%o9] 67.45
```

**Exercise 5.** Let $X = (1, 3, 6, 6, 6, 6, 7, 7, 12, 12, 17) = \begin{pmatrix} 1 & 3 & 6 & 7 & 12 & 17 \\ 1 & 1 & 4 & 2 & 2 & 1 \end{pmatrix}$.
Calculate the mean of $X$.
*Solution*: `Ex.5`

## 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 $\mu$ is the mean of $X$.

• The (sample) *variance* is defined as $var1(X) := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)$.

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

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

**Use Maxima**

```
load(descriptive) $
X : [4, 36, 45, 50, 75] $
var(X);                 /* = 528.4 */
var1(X);               /* = 660.5 */
```

▷ `Var`

**Mental image**

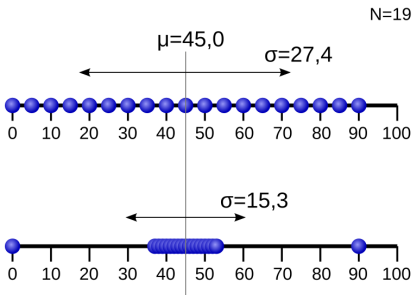



Figure 2: Figure illustrating the concept of variance for two different sets of 19 numbers (0, 5, ..., 90) and (0, 37, 38, ... 53, 90). $\mu$ denotes the mean and $\sigma$ denotes the square root of the variance as a measure of the spread.

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

### 1.2.1 Exercises

**Exercise 6.** The MAXIMA check of the example uses the build'in function *var* of MAXIMA package `descriptive`.
Write a user-defined function `Var` using the mathematical definition of *var*.

*Solution*: `Ex.6`

**Exercise 7.** (example from wikipedia, ibid.)
For a set of numbers 10, 15, 30, 45, 57, 52, 63, 72, 81, 93, 102, 105, if this set is the *whole data population* for some measurement, then variance is the population variance 932.743 as the sum of the squared deviations about the mean of this set, divided by 12 as the number of the set members. If the set is a *sample* from the whole population, then the unbiased sample variance can be calculated as 1017.538 that is the sum of the squared deviations about the mean of the sample, divided by 11 instead of 12.
Check this using your user-defined function `Var` and the predefined function `var`.

*Solution*: `Ex.7`

## 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 $\mu$ is the mean of $X$.

• The (sample) *standard deviation* is defined as $sd1(X) := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$.

**Mental image**

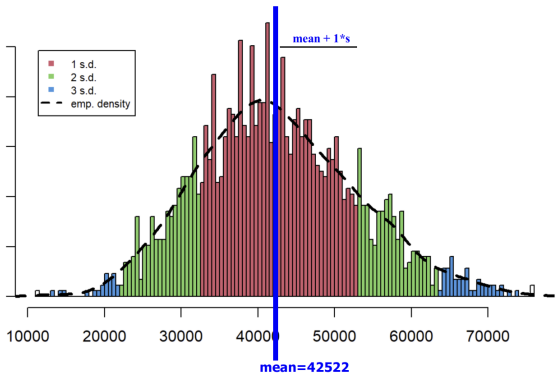



Table of annual salary of employees, $\mu = mean, s = sd(X)$: ▷ var
⊓: 68 % of the data fall inside the red region $\mu \pm s$.
Figure 3: ⊓: 95 % of the data fall inside the red+green region $\mu \pm 2 * s$.
⊓: 99.7 % of the data fall inside the red+green+blue region $\mu \pm 3 * s$.
⌒: the empirical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

**Example**

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

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

**Use Maxima**

```
load(descriptive) $
X : [4, 36, 45, 50, 75] $
std(X), numer;                /* =22.98 */
std1(X), numer;              /* = 25.70 */
```

▷ `std`

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

### 1.3.1 Exercises

**Exercise 8.** The MAXIMA check of the example uses the build'in function *std* of MAXIMA package `descriptive`.
Write a user-defined function `sd` using the mathematical definition of *sd*.
*Solution*: `Ex.8`

**Exercise 9.** (example from wikipedia, ibid.)
Suppose that the entire population of interest is eight students in a particular class.
Their marks are the following eight values: 2,4,4,4,5,5,7,9.
∘ The *std* formula of MAXIMA is valid only if the eight values with which we began form the complete population. If the values instead were a random sample drawn from some large parent population (for example, there were 8 students randomly and independently chosen from a student population of 2 million), then one divides by 7 (which is n - 1) instead of 8 (which is n) in the denominator of the *sd* formula, and we have to use function *std*1 of MAXIMA.

Check this using your user-defined function `sd` and the predefined function `std` resp. `std1`.
*Solution*: `Ex.9`

**Exercise 10.** (example from cf. Spiegel p. 106, P 4.10)
Find the std. deviation of 12,6,7,3,15,10,18,5.
*Solution*: `Ex.10`

**Exercise 11.** (students marks)
Find the std. deviation of the height of the students marks $X = (61, 64, 67, 70, 73)$ with frequency $f = (5, 18, 42, 27, 8)$, e.g. mark 61 was reached 5 times.
*Hint*: We have to adapt the definition of *std* w.r.t. frequency $f$.
Define e.g. 'mean w. frequency' by $meanF(X, f) = \frac{dot(f,X)}{\sum f}$.
*Solution*: `Ex.11`

**Exercise 12.** (Z-score alias standard.score)
The Z-score function is used to transform a list ('vector') of numbers into a new list with $mean = 0$ and $std = 1$, cf. [1] or [2]
Task: Define function 'zscore' and transform the list 6,2,8,7,5 into Z-scored numbers. Compare with function *standardize* of package *descriptive*.
Check mean and std.dev of the transformed numbers.
*Solution*: `Ex.12`

```
load ("descriptive")$
zscore(X) := float( (X-mean(X))/std(X) );

X : [6,2,8,7,5]$
zscore(X);
```

**Exercise 13.** (example from geeksforgeeks)
Using MAXIMA do Example 1 and 2 in [3].
*Solution*: `Ex.13`

## 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

**Mental image**

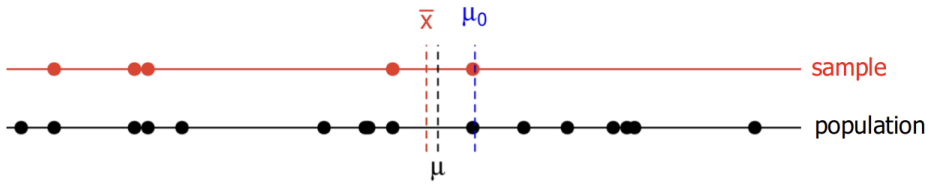



Figure 4: Visualization of sample vs. population and there resp. means $\overline{X}$ vs. $\mu$. $\mu_0$ is the theoretical mean, i.e. the target value.

**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.[5]

*Sample of a production of 100 resistors.*

| $n$ : | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\Omega$ : | 100.1 | 101.2 | 99.5 | 99.0 | 100.7 | 100.0 | 101.2 | 99.2 | 99.0 | 98.7 |

[5]This example concerning resistors quality is in Beucher [1, p. 219]. See chapter §4.1 of this script.

**Use Maxima** for example.1:

```
load(descriptive) $
X : [4, 36, 45, 50, 75] $
std1(X)/ sqrt(length(X)), numer;                    /* =11.49 */
```

▷ `std`

**General information**
◦ Mathematical `Information` about the concept is ▷ WIKIPEDIA : std.error of mean
◦ Syntax and semantic of the function in MAXIMA : – no build-in function provided –

### 1.4.1 Exercises

**Exercise 14.** There is no build'in function *sem* for MAXIMA.
Write a user-defined function `sem` using the mathematical definition of *sem* and use it to calculate the std.err.mean and the test statistic $T$ for the values of example2.
MATLAB: By default, the standard deviation is normalized by $n-1$, where $n$ is the number of observations.
*Solution*: `Ex.14`

**Exercise 15.** (check)
Verify, that $sem([1,2,3,4]) = 0.6454$.

**Exercise 16.** (example from soton.ac.uk)
Read: [1]. Program the right most explicit term for SE in [1] in MAXIMA.
Calculate the sem (=SE) of examples 1 and 2 in [1] using MAXIMA.

**Exercise 17.** (example from .investopedia)
Do the example in [2].
Read: handout.

## 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 $m \in \{mean, median, mode, ...\}$.

**Mental image**

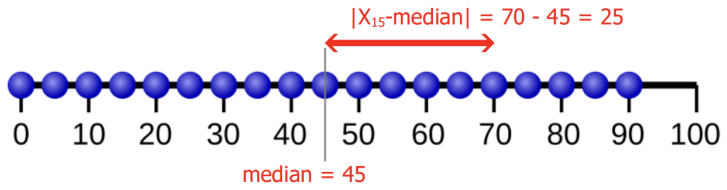



Figure 5: Figure illustrating the concept of *mad* for the data set of the 19 numbers $X = (0, 5, 10, ..., 90)$.

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

**Use Maxima** where `mad()` is called `mean_deviation()`:

```
load(descriptive) $
L: [4,36,45,50,75];
mad(L) := mean_deviation(L);
mad(L), numer;
```

▷ `mad`

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

### 1.5.1 Exercises

**Exercise 18.** There is no build'in function *mad* for MAXIMA.
Write a user-defined function `mad` using the mathematical definition of *mad* and use it to calculate the *mad* for example 1. and 2. For 2. use *median*() instead of *mean*.
*Solution*: `Ex.18`

**Exercise 19.** (check, wiki ibid.)
Verify, that for example, for the data set $2, 2, 3, 4, 14$ the *mad* is 3.6.
*Solution*: `Ex.19`

**Exercise 20.** (mean squared deviation)
Calculate for the contrast to the mean absolute deviation (*mad*) the 'mean squared deviation' `dsquare` defined by $dsquare(X) = \frac{\sum \bar{X}^2}{dim(X)}$ using MAXIMA; you may use this 'pseudo code' (in fact runnable EIGENMATH$^{online}$ code) as pattern:

```
X=(4,36,45,50,75)
mX    = mean(X)
Xbar  = X-mX
Xbar2 = Xbar^2
dsquare = sum(Xbar^2)/dim(X)
```

*Solution*: `Ex.20`

**Exercise 21.** (median abs.dev. around the median) For the sample $2, 2, 3, 4, 14$ the number 3 is the median, so the absolute deviations from the median are $1, 1, 0, 1, 11$ (reordered as $0, 1, 1, 1, 11$) with a median of 1, in this case unaffected by the value of the outlier 14, so the median absolute deviation is 1.
Define a function *madMed*() to calculate the 'median abs.dev. around the median' and calculate the *madMed* and verify, that `madMed( [1,2,3,4,5,6,7,8,9] ) = 2`.
Hint: use the definition `madMed(X) = median(abs(X - median(X)))`.
*Solution*: `Ex.21`

**Exercise 22.** (test vehicles of one car type, cf. Beucher p.176, P 64)
For 40 test vehicles of one car type, the fuel consumption $X$ was determined in liters per 100 km of driving in the city center. This sample yielded the following results (table TX with 5 rows):

```
TX = zero(5, 8)
TX[1] = (10.1,10.6,10.9,10.0,10.4,10.5, 9.7,10.5)
TX[2] = (10.4,10.1,10.8, 9.2,10.2,10.3,10.5, 9.2)
TX[3] = (10.2,10.5, 9.4,10.2, 9.6,10.2, 9.7,10.2)
TX[4] = (10.8, 9.9,10.5,10.6, 9.8,10.7,11.2,10.8)
TX[5] = ( 9.9,10.0,10.5,10.4,11.4,10.4,10.1,10.4)
```

Divide the sample into the following five approx. equidistance characteristic classes:
" $< 9.5$", "$9.5 - 10.0$", "$10.0 - 10.5$", "$10.5 - 11.0$", " $> 11.0$".

For the following task use functions of MAXIMA package descriptive.

a. Determine the list of absolute and relative frequencies.

b. Determine the empirical distributions corresponding to the class classification and represent this in a histogram.

c. Collect the known 'measures of dispersion' i.e. mean, std.dev., variance, and m.a.deviation in a 'score' table.

*Solution*: `Ex.22`

## 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 $n = length(X)$.

### Mental image

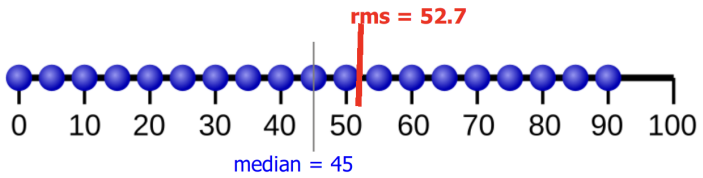



Figure 6: Figure illustrating the concept of *rms* for the data set of the 19 numbers $X = (0, 5, 10, ..., 90)$.

### Examples

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

**Use R** to check 1. (where there is also no build-in function for *rms*()):

```
> X <- c(4, 36, 45, 50, 75)
> sqrt(mean(X^2))
  [1] 47.87901
```

### General information

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

### 1.6.1 Exercises

**Exercise 23.** There is no build'in function *rms* for MAXIMA.
Write a user-defined function `rms` using the mathematical definition of *rms* or the term in the `R` check and calculate the *rms* for examples 1. and 2.
*Solution*: `Ex.23`

**Exercise 24.** (check, wiki ibid.)
Verify, that for example, for the data set $X = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$ the $rms(X)$ is 6.204837.
*Solution*: `Ex.24`

**Exercise 25.** (rms in MatLab)
Check some of the examples in ▷ MATLAB : `rms`
*Solution*: `Ex.25`

## 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).

**Mental image**

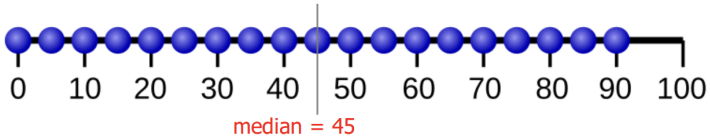



Figure 7: Figure illustrating the concept of *median* for the data set of the 19 numbers $X = (0, 5, 10, ..., 90)$.

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

**Use** MAXIMA to check 1. :

```
load(descriptive)$
median([36,4,75,45,50]);  /* = 45 */
```

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

### 1.7.1 Exercises

**Exercise 26.** (a user-defined function `Median`)
Write a user-defined function `Median` using the mathematical definition of *median* and check your function on example 1.
Hint: you must take into account that the list $X$ for the use of $Median$ has to be sorted! So you must use MAXIMA's $sort$(X) function first.
*Solution*: `Ex.26`

**Exercise 27.** (check median for a data set)
Verify, that for example, for the data set

```
X = (1,2,3,4,  5,  6,7,8,9)
    -- 4 elem. ^   4 elements right
```

your Median(X) is 5.
*Solution*: `Ex.27`

**Exercise 28.** (*median* in `MatLab` and $\mathcal{R}$)
Check some of the examples in ▷ MATLAB : `median` or in ▷ R : `median`

## 1.8 `mode` - the Mode

The *mode* is the value $\tilde{\tilde{X}}$ 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).

**Mental image**

Figure 8: 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.

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

**Use** Octave e.g. with SageMathCell or MatLab$^{online}$ to check 1. :

```
octave:1> X=[1,2,1,4,4]
octave:2> mode(X)
          ans = 1
```

**General information**

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

### 1.8.1 Exercises

**Exercise 29.** (a user-defined function `mode`, contributed by M. R. Riotorto)
There is no build-in function *mode* for Maxima.
Write a user-defined function *mode*() using the above incomplete mathematical 'definition' of *mode* and calculate the *mode*'s for examples 1., 2. and 3. .
Hint: you may get an idea looking at example 2.
*Solution*: `Ex.29`

**Exercise 30.** (onother user-defined function `mode`)
Write a user-defined function *mode*1() using the pseudocode in the following link and check your code for examples 1., 2. and 3. .
*Solution*: `Ex.30`

**Exercise 31.** (*mode* in `MatLab`)
Check some of the examples in ▷ 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 $N = \lceil p * 0.01 * n \rceil$. [6]

**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**

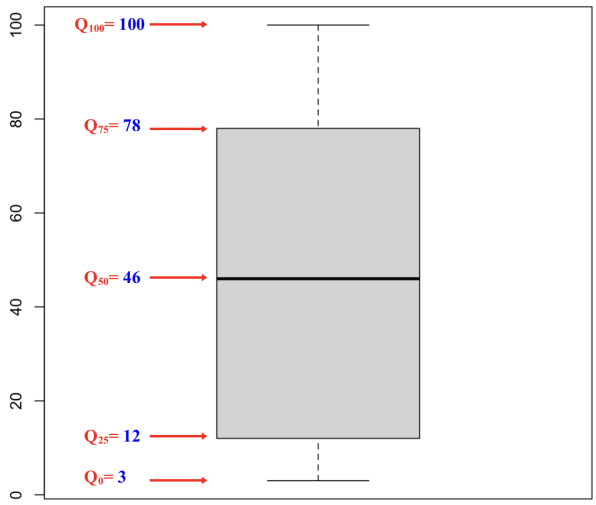



Figure 9: Figure illustrating the concept of *quantile* for the sets of numbers $(12, 3, 4, 56, 78, 18, 46, 78, 100)$: $\begin{pmatrix} 0\% & 25\% & 50\% & 75\% & 1000\% \\ 3 & 12 & 46 & 78 & 100 \end{pmatrix}$, visualizied in a so-called *boxplot*. $Q_{50} = 46$ is the $median(X)$.

[6] `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 $ceil(x)$.

### Example

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

**Use** MAXIMA to check:

```
load ("descriptive")$
X : [12, 3, 4, 56, 78, 18, 46, 78, 100]$
[quantile(X, 1/4), quantile(X, 1/4), quantile(X, 1/2),
 quantile(X, 3/4), quantile(X, 1/1)], numer;
```

*Check* Example

- OCTAVE uses `prctile(X,[0,25,50,75,100])` and gives `ans = 3 10 46 78 100`, where $Q_0 = 3$ does not corresponds with $Q_0\,|_{Maxima}$ =12.

**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 ▷ MAXIMA : `quantile`
∘ Syntax and semantic of the function is here ▷ MATLAB : `prctile`

### 1.9.1 Exercises

In CAS MAXIMA, quantiles are calculated using the built-in functions from the *descriptive* package. To find the $p$-quantile, use the naming convention `quantile(X,p)`, with $p$ a number in $]0,1[$ and $X$ is the sample list.

**Exercise 32.** (boxplot of Fig.9)
Reproduce the boxplot fig.9 for the data set $(12, 3, 4, 56, 78, 18, 46, 78, 100)$, cf. ▷ `boxplot`.
*Solution*: `Ex.32`

**Exercise 33.** (a user-defined function `Quantile`)
Write a user-defined function `Quantile` using the mathematical definition for $Q_p$ using the *nearest-rank method* for a quantile and check your function on example 1.
Compare the result w.r.t. the CAS MAXIMA, MATLAB and $\mathcal{R}$.
*Solution*: `Ex.33`

**Exercise 34.** (five-number summary)
The "five-number summary" (consisting of the minimum, 1st Quartile, Median, 3rd Quartile, and maximum) can be computed using the *quantile* function. These five statistics break down the position and spread of a dataset.
Program the "five-number summary" as a MAXIMA function.
*Solution*: `Ex.34`

**Exercise 35.** (example from MATLAB)
Calculate the quantile of data set $X$ for percentage $p = 42\%$,
where $X : 0.5377, 1.8339, -2.2588, 0.8622, 0.3188, -1.3077, -0.4336$.
*Result*: quantile(X,42) should be -0.1026 or -0.0424
*Solution*: `Ex.35`

**Exercise 36.** (example from $\mathcal{R}$)
Reproduce the following calculation in MAXIMA:

```
R>  df <- c(6, 3, 2, 10, 1)
R>  quantile(df)
    [1]   0%  25%  50%  75% 100%
           1    2    3    6   10
R> quantile(df, .4)
   [2]   2.6     --  which is NOT in the data !
```

*Solution*: `Ex.36`

**Exercise 37.** (example with 'pseudo'code)
Do the following calculation in MAXIMA:

```
A = (6, 3, 2, 10, 1)
percentile(A, 40)
Qp(A,40)
boxplot(A)
```

*Solution*: `Ex.37`

## 1.10 `moment` - the r$^{st}$ 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

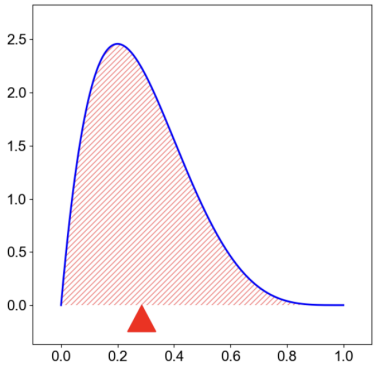



Figure 10: Imagine the data distributed under the blue graph. Then the moment may be interpreted as 'center of mass', balanced upon the 'expected' value △, the moment e.g. the mean; cf. GUNDERSEN

### Examples

1. The 1st *moment* of the five values 4, 36, 45, 50, 75 is 45, i.e. its mean.
2. The 2nd moment of the five values 4, 36, 45, 50, 75 is 2292.4.
3. The 3rd moment about $A = 4$ of the five values 4, 36, 45, 50, 75 is 111387.

**Use** MAXIMA to check:

```
load ("descriptive")$
moment(M,r) := float( noncentral_moment (M,r));

X : [4, 36, 45, 50, 75]$
moment(X,0);
moment(X,1);       /* = 42   */
moment(X,2);      /* = 2292.4 */
moment(X-4,3);   /* = 111387.2 */
```

*Check* Example

### General information

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

### 1.10.1 Exercises

**Exercise 38.** (user-defined function `moment`)
Write a user-defined function `moment(X,A,r)` using the mathematical definition for *moment* and check your function on the three examples.
Do not inspect the source code of 'noncentral_moment' in package *descriptive*.
*Solution*: `Ex.38`

**Exercise 39.** (moment about the mean)
Let $X = (0.5377, 1.8339, -2.2588, 0.8622, 0.3188, -1.3077)$ be a small sample.
Let $A = mean(X)$ be the mean of $X$.
Calculate the $moment(X, A, 3)$ and $moment(X, 0, 3)$.
Result:$-1.1143$ using $\mathcal{R}$with center=TRUE; $-1.1274$ using $\mathcal{R}$with center=FALSE.
*Solution*: `Ex.39`

**Exercise 40.** (example from Spiegel, p.129, P 5.1)
Calculate the first, second, third and fourth moments of the data set $2, 3, 7, 8, 10$.
Hint: Let $A = 0$ be the 'noncentral' value of $X$.
*Solution*: `Ex.40`

**Exercise 41.** (moments for grouped data, cf. Spiegel, p.131 P 5.6)
Students marks in a class have the following frequency table

```
class mark X:  61 64 67 70 73
frequency  f:   5 18 42 27  8
```

i.e. class mark 67 comes with frequency 42.

Task: find the first four moments of the class marks.

Hint: the solution needs *meanF* for grouped data and a variant of moment w.r.t. mean of grouped data; here is the pseudocode:

```
meanF(X,f)      = float( dot(f,X) / sum(f) )
momentMF(X,f,r) = float( dot(f,(X-meanF(X,f))^r) / sum(f) )
```

Write `meanF(X,f)` and `momentMF(X,f,r)` in MAXIMA and then do `momentMF(X,f, 1)` ...
.
*Solution*: `Ex.41`

**Exercise 42.** (moments for grouped data, cf. Spiegel, p.137 P5.27)
Find the first four moments for the distribution $X : f$, where the
date is $X = (12, 14, 16, 18, 20, 22)$ with frequency $f = (1, 4, 6, 10, 7, 2)$.

Result:momentMF(X,f,1) = 0; momentMF(X,f,2) = 5.97; momentMF(X,f,3) = -0.397; momentMF(X,f,4) = 89.22
*Solution*: `Ex.42`

## 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

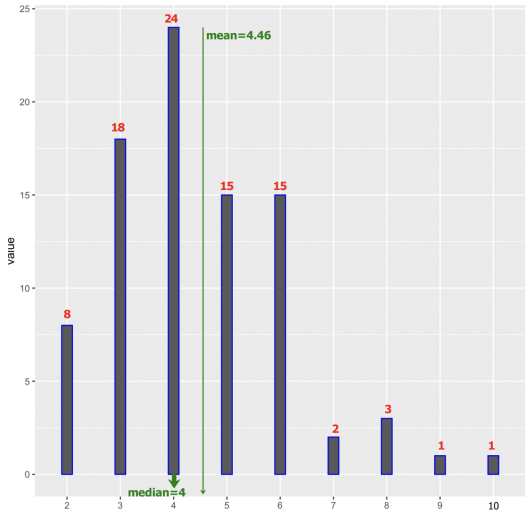



Figure illustrating the concept of *skew*ness for a distribution of numbers.
Figure 11: The frequency distribution has positive skewness of 0.808 (skew right). The *mean*= 4.46 is placed *right* of the *median*= 4.

### Example

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

**Use** MAXIMA to check:

```
load ("descriptive")$
X : [4,36,45,50,75]$
skewness (X), numer;     /* = -0.306 = by octave:1> skewness(X) */
```

*Check* Example

*Warning*: if you use $\mathcal{R}$ to check you get different results w.r.t. options:

| | | |
|---|---|---|
| `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] |

### General information

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

### 1.11.1 Exercises

**Exercise 43.** (user-defined function `skewness`)
Write a user-defined function `skew(X)` using the mathematical definition for *skew* in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
Also, do not inspect the source code of 'skewness' in package *descriptive*.

**Remark.** 1. The software $\mathcal{R}$ use different realizations for *skewness*, which we also define in this exercise. So you get very different results. Study the code lines with care.
2. Because we have our own self-defined function *skew*, we are able to adapt their code for other features, i.e. to implement versions of *skew* which have identical values w.r.t. other stats software. Look at the warning after the example.
*Solution*: `Ex.43`

**Exercise 44.** (skewness example from MATLAB)
Let $X = (0.5377, 1.8339, -2.2588, 0.8622, 0.3188, -1.3077, -0.4336, 0.3426, 3.5784, 2.7694)$.

a. Verify using exercise.43.
(you should at minimum do $skew(X) := skewness(X)$, which is build-in in MAXIMA):

```
skew(X)       -- 0.1060  R: typ=1, Matlab: default
skewR(X)      -- 0.1006  R: ok
skew1(X)      -- 0.0905  R: typ=3, Matlab: ./.
skew2(X)      -- 0.1258  R: typ=2, Matlab: flag = 0
```

b. Verify the results of $\mathcal{R}$ using our user-defined functions of exercise 43:

```
R> library(e1071)
   skewness(X)                  -- default, we use skew1
   [1] 0.09058831
   skewness(X, type=3)          -- we use skew1
   [1] 0.09058831
   skewness(X, type=1)          -- we use skew
   [1] 0.1060983
   skewness(X, type=2)          -- we use skew2
   [1] 0.1258171
```

*Solution*: `Ex.44`

**Exercise 45.** (data set with sharp peak, cf. Spiegel P. 5.44)
The following data set $X$ has a sharp peak at 50. Verify: skew(X) = 0.

```
X=(10,
   20,20,
   30,30,30,
   40,40,40,40,
   50,50,50,50,50,50,50,50,50,50,50,50,50,50,50,50,50,50,50,
                                  50,50,50,50,50,50,50,50,50,50,
   60,60,60,60,
   70,70,70,
   80,80,
   90)
```

*Solution*: `Ex.45`

**Exercise 46.** (a skew data set)
A skew data set is $sk = (5, 10, 15, 20, 30, 50, 80, 150, 300, 500)$.
a. Calculate $skew(sk)$. (Bonus: calculate $skewR(sk), skew1(sk), skew2(sk)$)
b. Argue, which value of the different versions (e.g. `skew()` or `skewR()` or `skew1/2()`) of $skew$ is best suited to describe the shape of $Y$.
*Solution*: `Ex.46`

**Exercise 47.** (Skewness of the experimental data of figure.9)
Calculate the skewness of the experimental data of figure.9.

```
X =(2,2,2,2,2,2,2,2,
    3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,
    4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,
    5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,6,
    6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,
    7,7,
    8,8,8,
    9,
    10)
```

*Solution*: `Ex.47`

## 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 $r$st moment of $X$. – The $1st$ formula is used by R, the $2nd$ by MatLAB.

### Mental image

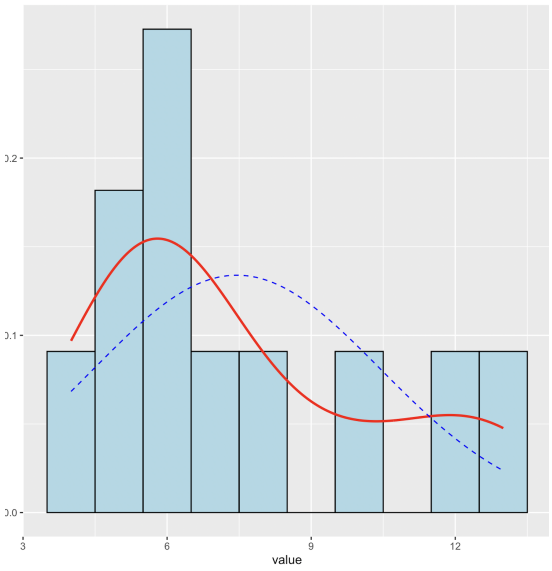



Figure illustrating the concept of *kurtosis* for a set of numbers.
Figure 12: The normal distribution is noted in '---', the distribution of the number set in '---'. The kurtosis is 2.295 resp.2.226. See example.2.

### 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.

**Use** MAXIMA to check example.1:

```
load ("descriptive")$
X : [4,36,45,50,75]$
kurtosis (X), numer;     /* = - 0.65297 */
```

*Check* Example

Check example.2 with $\mathcal{R}$:

```
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)]
```

### General information

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

### 1.12.1 Exercises

**Exercise 48.** (user-defined function `kurtosis` in plain MAXIMA)
Write a user-defined function `kurtosis(X)` using the mathematical definition for *kurtosis* in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
Also, do not inspect the source code of 'kurtosis' in package *descriptive*.

a. Check your function on the examples 1. and 2.
*Solution*: `Ex.48`

b. This implementation of kurtosis gives compatibility with MATLAB's *kurtosis* with 'no flag'. Verify this with your `kurtosis(X)` w.r.t. the following commands:

```
MATLAB>> Y = [4, 5, 5, 6, 6, 6, 7, 8, 10, 12, 13]
      >> kurtosis(Y)
         ans =    2.2959
```

c. Check, that the build-in `kurtosis(X)` of MAXIMA package *descriptive* is _*not*_ compatible with MATLAB.

d. Check, that our `kurtosis(X)` *is* compatible with $\mathcal{R}$'s *kurtosis* from library 'moments':

```
## check compatibility with R and library(moments)
R> X = c(2,3,4,4,4,5,6,7,8,9,10)
R> library(moments)
R> kurtosis(X)
  [1] 1.912949
```

• The software $\mathcal{R}$ use different realizations for *kurtosis*, which we discuss in the following exercises. Be warned: you get very different results w.r.t. *kurtosis* using different software!

**Exercise 49.** (MATLAB's *kurtosis*0 with $flag = 0$, i.e. bias-corrected equation)
Reconstruct MATLAB's `kurtosis(X, flag)`. That is: When $X$ represents a sample from a population, the *kurtosis* of $X$ is 'biased', meaning it tends to differ from the population kurtosis by a systematic amount based on the sample size. If this is the case:
◦ .. set $flag = 0$ to correct for this systematic bias.
◦ .. this bias-corrected equation requires that $X$ contain at least four elements.

Here is a 'pseudo'code for your help:

```
(* MatLab's kurtosis with flag=0, i.e. bias-corrected equation *)
kurtosis0(X,n) = do( n=dim(X),
             test( n<4 , "n must be >=4",
             (n-1)/((n-2)*(n-3))*((n+1)*kurtosis(X) -3*(n-1))+3 ))
```

a. Write MAXIMA function *kurtosisO*, which implements MATLAB's `kurtosis(X, flag)`.
b. Check your *kurtosisO* function w.r.t. this MATLAB call:

```
MATLAB>> Y = [4, 5, 5, 6, 6, 6, 7, 8, 10, 12, 13]
      >> kurtosis(Y)
         ans =    2.2959
      >> kurtosis(Y,0)
         ans =    2.6598
```

*Solution*: Ex.49

**Exercise 50** (*Bonus*). (*kurtosis* compatibility functions w.r.t. MatLab and $\mathcal{R}$)
The following file implements variants of kurtosis. Study the code and use appropriate versions for the following exercises.

*Solution*: Ex.50

**Exercise 51.** (kurtosis of data with different implementations)
Calculate the kurtosis of $data = (2, 3, 4, 4, 4, 5, 6, 7, 8, 9, 10)$ and compare the result with $\mathcal{R}$:'e1071' and $\mathcal{R}$:'moment'.

```
# kurtosis with R and library(moments)
R> X = c(2,3,4,4,4,5,6,7,8,9,10)
R> library(moments)
R> kurtosis(X)
    [1] 1.912949

#kurtosis with R and library(e1071)
R> X = c(2,3,4,4,4,5,6,7,8,9,10)
R> library(e1071)
R> kurtosis(X)
   [1] -1.41905          --> .. + 3 = 1.581
```

*Solution*: Ex.51

**Exercise 52.** (kurtosis of a skew data set)
Let $Y = (4, 5, 5, 6, 6, 6, 7, 8, 10, 12, 13)$ be a data set.
Calculate the kurtosis of $Y$ with different functions.

*Solution*: Ex.52

**Exercise 53.** (kurtosis for experimental data of figure 12)
Let $Y$ be the data set of fig.12:

```
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
```

a. Calculate the *kurtosis* of $Y$ and argue, which value of the different versions (e.g. `kurtosis()` or `kurtosisM()` or `kurtosisR3()` etc.) of *kurtosis* is best suited to describe the shape of $Y$.
b. Which connection is seen between our self-build *kurtosis* function and the kurtosis of package *descriptive*?
*Solution*: `Ex.53`

**Exercise 54.** (kurtosis of a frequency table, cf. Spiegel, p.135, P 5.13)
Let $X = (61, 64, 67, 70, 73)$ be the marks with a frequency distribution $f = (5, 18, 42, 27, 8)$. Adapt the kurtosis to median and frequency of the group and calculate kurtosis of $X : f$.
Hint: you may use this pseudocode:

```
.............................................................
momentMF(X,f,r) = float( dot( f, (X-meanF(X,f))^r)/sum(f) )
kurtosisMF(X,f) = momentMF(X,f,4)/momentMF(X,f,2)^2
.............................................................
```

*Solution*: `Ex.54`

**Exercise 55.** (a left skew data set, cf. Spiegel, data Xleft: p.126, col.5-6)
The 'obitage data' set is

```
X = (102,55,70,95,73,79,60,73,89,85,
          72,92,76,93,76,97,10,70,85,25,83,
          58,10,92,82,87,104,75,80,66,93,
          90,84,73,98,79,35,71,90,71,63,
          58,82,72,93,44,65,77,81,77)
```

a. Calculate the kurtosis of $X$ using *kurtosis*(X) and *kurtosisR*3(X) (i.e. giving the same value as if calling $\mathcal{R}$:'e1071' Typ 3).
b. Give a tabulated summary of the results.
*Solution*: `Ex.55`

Summary of the different implementations of *kurtosis*:

| To calculate KURTOSIS | *use package* | use MAXIMA |
|---|---|---|
| with self-build kurtosis : | | `kurtosis` |
| with build-in kurtosis : | descriptive | `kurtosis` |
| analog MATLAB: | no flag | `kurtosis` |
| MATLAB: | flag = 0 | `kurtosis0` |
| analog $\mathcal{R}$: | moments | `kurtosis` |
| $\mathcal{R}$: | e1071, typ 1 | `kurtosisR1` |
| $\mathcal{R}$: | e1071, typ 2 | `kurtosisR2` |
| $\mathcal{R}$: | e1071, typ 3 | `kurtosisR3` |

## 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.[7]
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

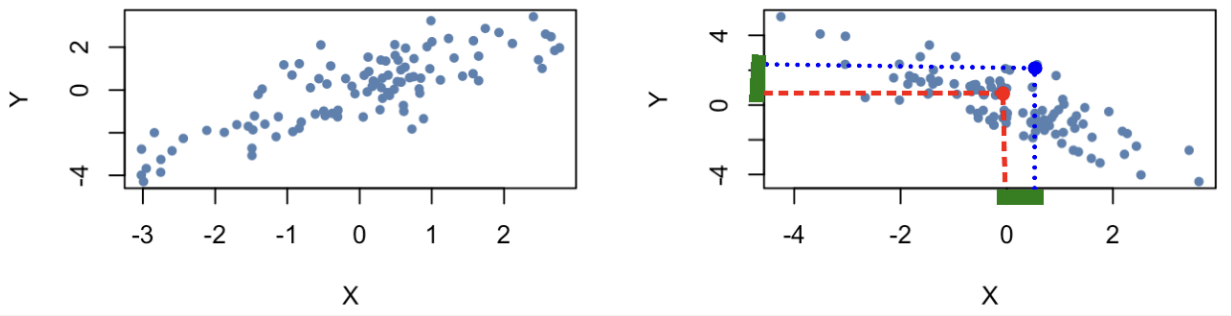



Figure 13: Figure illustrating the concept of *covariance* for a set of numbers. Left: positive `cov`. Right: negative `cov`. • $= (\bar{X}, \bar{Y})$. • $= (x_i, y_i)$ a data point. $\bowtie \times \bowtie = (x_i - \bar{X}) \cdot (y_i - \bar{Y})$ a distance product w.r.t. the mean.

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

### Examples

The covariance of the two data sets $1, 3, 5, 10$ and $2, 4, 6, 20$ is 23 resp. 30.66.

**Use** MAXIMA to check the example:

```
(* .. user-defined functions are here .. *)
X : [1,3,5,10]$
Y : [2,4,6,20]$
cov(X,Y), numer;    /* = 23    */
covR(X,Y), numer ; /* = 30.66 */
```

- Check with $\mathcal{R}$:

```
> X<-c(1,3,5,10)
> Y<-c(2,4,6,20)
> cov(X, Y)         # default:  method = "pearson"
[1] 30.66667
```

### General information

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

### 1.13.1 Exercises

**Exercise 56.** (user-defined function `cov` in plain MAXIMA)
Write a user-defined function `cov(X)` using the mathematical definition for *cov* in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
Check your function on the example.
*Solution*: `Ex.56`

**Remark.** 1. WARNING: There are two different methods to compute the covariance, which leads to different results w.r.t population vs. sample.
2. This implementation of *covR* gives compatibility with $\mathcal{R}$'s *cov* with the option 'method = pearson'.

**Exercise 57.** (Calculate the covariance of two data vectors)
Let $X = (16.50, 19.16, 7.64, 8.14, 8.14, 2.48, 9.23, 4.78, 6.93, 11.91, 11.60, 13.97)$
and $Y = (16.50, 19.16, 7.64, 8.14, 8.14, 2.48, 9.23, 4.78, 6.93, 11.91, 11.60, 13.97)$.
Determine `cov(X,Y)` and `covR(X,Y)`.
*Solution*: `Ex.57`

**Exercise 58.** (Calculate the covariance of two data vectors)
Let $a = (2, 4, 6, 8, 10)$ and $b = (1, 11, 3, 33, 5)$.
Calculate $covR(a, b)$ and $cov(a, b)$.
Result: covR = 15, with method = "pearson"

**Exercise 59.** (Calculate the covariance and interpret the result)
Task: calculate
a. cov( (2,3,4,3), (2,3,4,3) )
b. covR( (2,3,4,3), (2,3,4,3) )
c. cov( (1,2,3,4), (4,3,2,1) )
d. Plot a diagram a la Fig.13 for each task and interpret the results.

**Exercise 60.** (example from r-bloggers cf. cov)
Suppose we have two vectors, $x$ and $y$, representing the number of hours studied and the corresponding test scores, respectively, for a group of students.
We want to measure the covariance between these two variables.
*Solution*:

1. Create example vectors $x = (5, 7, 3, 6, 8)$ and $y = (65, 80, 50, 70, 90)$
2. Calculate the covariance $cov(x, y)$ resp. $covR(x, y) = 29$.

   ok: `R> covariance(x,y) [1] 29`
3. Interpretation: – In this example, the resulting covariance value will help us understand the relationship between the hours studied and the corresponding test scores. What this particular example is saying is that for every unit increase in $x$ there is a 29 unit increase in $y$.

Task: do this exercise in MAXIMA.

# 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 MAXIMA 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*[8] function for the Binomial distribution is defined as:

  $\texttt{binoPDF}(\texttt{k}, \texttt{n}, \texttt{p}) := f(k, n, p) := \Pr(X = k)$[9] $= \binom{n}{k} p^k (1-p)^{n-k}$

- 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[10]

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

  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 $\alpha$.

**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? *Check* example.

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

[8] We adopt the naming convention of MATLAB. So there is no confusion with the use of functions of MAXIMA package *distib*, where e.g. *binoPDF* is noted as *pdf_binomial*.

[9] Pr(X=k): represents the probability of getting exactly $k$ successes

[10] there is no closed term for the inverse binomial distribution

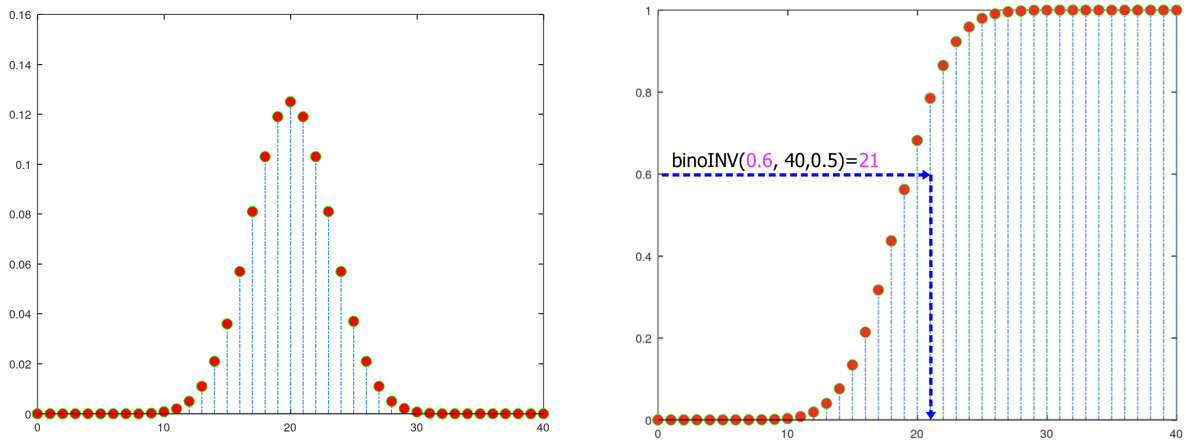
2. *binoCDF*: The cumulative probability of a binomial distribution with 10 trials and a probability of success 0.5 for 4 successes is 0.3769. *Check* example.

| `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 $\alpha$=0.7, find the first value $x$ such that $\alpha = 0.7 \leq$ binoCDF(x,100,0.3). Result: 32 *Check* example.

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

### 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 14: 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

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

MAXIMA's *distrib* function *cdf_binomial* is defined using $I_x(a, b)$.
We use it to plot the *continuous* graph of `binoCDF(.)`, look at figure.14, right plot.

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

### 2.1.1 Exercises

**Exercise 61.** (user-defined function `binoPDF` in plain MAXIMA)
Write a user-defined function `binoPDF()` using the mathematical definition for *binoPDF* in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
Check your function on example.1.
*Solution*: `Ex.61`

**Exercise 62.** (Table and plot for binoPDF)
a. Print 10 values for the binomial PDF for $n = 10$ and $p = 1/6$ and e.g. $i = 1$ *thru* 10.
b. Plot the Binomial PDF values from a. - if necessary by paper and pencil.
c. Do a plot for the binomial PDF for $n = 40$ and $p = 0.5$.
*Solution*: `Ex.62`

**Exercise 63.** (user-defined function `binoCDF` in plain MAXIMA)
Write a user-defined function `binoCDF()` using the mathematical definition for *binoCDF* in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
Check your function on the example.2.
*Solution*: `Ex.63`

**Exercise 64.** (Table and plot for binoCDF)
a. Print 10 values for the binomial CDF for $n = 10$ and $p = 1/6$ and e.g. $i = 1$ *thru* 10.
b. Plot the Binomial CDF values from a. - if necessary by paper and pencil.
c. Do a plot for the binomial CDF for $n = 40$ and $p = 0.5$.
*Solution*: `Ex.64`

**Exercise 65.** (a continous version of *choose* and *binoCDF* via Gamma function)
Pocket calculators often have a button nCr (=' n Choose r') to calculate the binomial coefficient $\binom{n}{r}$, in MAXIMA `binomial(n,r)`. We now extend this function and define a 'continous' version of *choose*() via Gamma function $\Gamma$ and also extend the cumulative binomial probability to `binoCDF2`, e.g. to plot graphs of the `binoCDF`:

$$\begin{aligned} nCk(n,k) &= \Gamma(n+1)/(\Gamma(k+1)\cdot\Gamma(n-k+1)) \\ binoCDF2(k,n,p) &= \sum_{j=0}^{k} nCk(n,j)\cdot p^j\cdot(1-p)^{n-j} \end{aligned}$$

where $\Gamma(x)$ is `gamma(x)`$|_{Maxima}$.

**Remark** (from docu)**.** In Maxima, the *gamma* function is written as $gamma(z)$. It evaluates symbolically, numerically, and can be expanded or simplified for various inputs.
For integer arguments, Maxima automatically evaluates to the exact $factorial\,value$. For example, $gamma(5)$; evaluates to 24. Use $float(gamma(x))$ to evaluate the gamma function for real numbers to receive a floating-point result.

a. Define `nCk(n,k)` and `binoCDF2(k,n,p)` in MAXIMA and
check it e.g. by `binoCDF2(4, 6, 0.3); /* ok: 0.9890 */;`

b. Draw `binoCDF2(x, 10, 1/6)` using $xrange = (0, 10)$ and $yrange = (0, 1)$.

*Solution*: `Ex.65`

**Exercise 66.** (user-defined function `binoINV` in plain MAXIMA)
Write a user-defined function `binoINV()` using the mathematical definition for $binoINV$ in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.

• The inverse binomial cumulative distribution function `binoINV()` results in the minimum number of successes, such that the cumulative binomial probability for that minimum number of successes is $\geq$ the given cumulative probability. Because there is no closed formula for `binoINV()` we do a simple search to solve the following typical problem to understand the construction of `binoINV()`, cf. finding values of $binoCDF$

a. Given a number of trials $n = 100$, the probability of success $p = 0.3$, the cumulative area of the binomial distribution $A = 0.7$,
find the value $x$ such that `A = 0.7` $\leq$ `binoCDF(100,0.3,x)`.

To solve 'Task: search for k with binoCDF $\geq 0.7$ for 1st time' do a simple search via a loop e.g.

```
for(k,30,35,  print((k, binoCDF(k,100,0.3))) )
```

and read off 'ok : k=32'.

*Solution*: `Ex.66a`

b. Now program `binoINV()` using e.g. the following pseudocode[11]

```
k=0
loop(k=k+1, x=binoCDF(k,100,0.3), x,k, test(x>0.95, break))
return(k)
```

*Solution*: `Ex.66b`

[11] which is in fact runnable code in CAS EIGENMATH$^{online}$

**Exercise 67.** (problem from Texas Instruments, cf. cf.: TI)
Assume the toss of a fair coin 30 times. What is the minimum number of heads you must observe such that the cumulative binomial probability for that number of observed heads is at least 0.95?
*Hint*:
• (idea:) set $x = binCDF(k, 30, 0.5)$ and use a table of values (starting at 0 and increment by 1) to find when the cumulative binomial probability is at or just above the given cumulative binomial probability. This gives you a view of all values to make decisions. For this example, search in the table to find the cumulative binomial probability just larger than 0.95. So, the number of successes is 19.
• (motivation for $binoINV$, quoting TI:) The results on the screen first show that the minimum number of successes to obtain at least the given cumulative binomial probability of 0.95 is 19. Next, the cumulative probability for up to 19 is computed using $binoCDF$ and is approximately 0.9506314271 which meets the criteria of $0.9506314271 \geq 0.95$.
Abstracting this process we get our small function $binoINV$.

```
choose(n,x)    := binomial(n,x)$
binoCDF(k,n,p) := sum( choose(n,j)*p^j*(1-p)^(n-j) , j,0,k) $

binoINV(alpha,n,p) := block( [x,k:0],
                        do( k : k+1,
                            x : binoCDF(k,n,p),
                            if x>alpha then return(k)))$

fpprintprec : 7$
binoINV(0.95, 30, 0.5);
```

*Solution*: Ex.67

## 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[12]

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

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

**Examples** from MatLAB solved with Maxima.

1. *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. *Check*.

   | `geoPDF(5, 0.5)` | 0.015625

2. *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. *Check*.

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

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

3. *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? *Check*.

| `geoINV(0.1, 0.03)` | 3 |
|---|---|

### Graphical representation

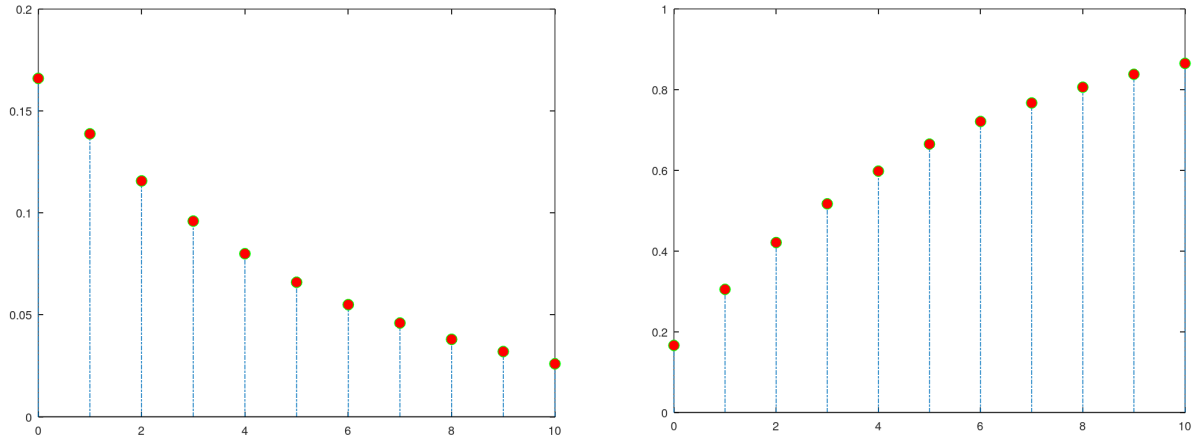



Figure 15: Experiment: rolling a six-sided die.
Left figure: • = plot for `geoPDF(k, 1/6)` for $k = 0, \ldots, 10$.
Check `geoPDF`$(4, 1/6) = 0.08$ on the graph.
Right figure: • = plot for `geoCDF(k, 1/6)` for $k = 0, \ldots, 10$.
Check `geoCDF`$(5, 1/6) = 0.665$ on graph.
Check `geoINV`$(0.7, 1/6) = 6$ 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`
○ Online calculator *Bognar's app*

### 2.2.1 Exercises

**Exercise 68.** (user-defined function `geoPDF` in plain MAXIMA)
Write a user-defined function `geoPDF()` using the mathematical definition for $geoPDF$ in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
Check your function on example.1.
*Solution*: `Ex.68`

**Exercise 69.** (Table and plot for geoPDF)
a. Print 10 values for the geometric PDF for $k = 1..10$ and $p = 1/6$.
b. Plot the geometric PDF values from a. - if necessary by paper and pencil.
*Solution*: `Ex.69`

**Exercise 70.** (user-defined function `geoCDF` in plain MAXIMA)
Write a user-defined function `geoCDF()` using the mathematical definition for $geoCDF$ in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
Check your function on the example.2.
*Solution*: `Ex.70`

**Exercise 71.** (Table and plot for geoCDF)
a. Print 10 values for the geometric CDF for $n = 1..10$ and $p = 1/6$.
b. Plot the geometric CDF values from a. - if necessary by paper and pencil.
c. Do a plot for the geometric CDF for $n = 40$ and $p = 0.5$.
*Solution*: `Ex.71`

**Exercise 72.** (user-defined function `geoINV` using explicit formula)
Write a user-defined function `geoINV()` using the mathematical definition for $geoINV$ in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
*Solution*: `Ex.72`

**Exercise 73.** (alternative user-defined function `geoINV` using loop construct)
Write a user-defined function `geoINV()` using a loop construct for $geoINV$ in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.

> The inverse geometric cumulative distribution function `geoINV()` results in the minimum number of successes, such that the cumulative geometric probability for that minimum number of successes is $\geq$ the given cumulative probability. Although there is a closed formula for `binoINV()` in a., we do a simple search to solve the typical problem of example.3 to understand the construction of `binoINV()`, see exercise 66.

*Solution*: `Ex.73`

**Exercise 74.** (inverse probability)
Verify, that the inverse probability at $pin = 0.25, p = 0.6835$ is 3.

## 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[13]

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

  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 $\alpha$.

  `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 MAXIMA.

1. *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? *Check.*

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

2. *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. *Check.*

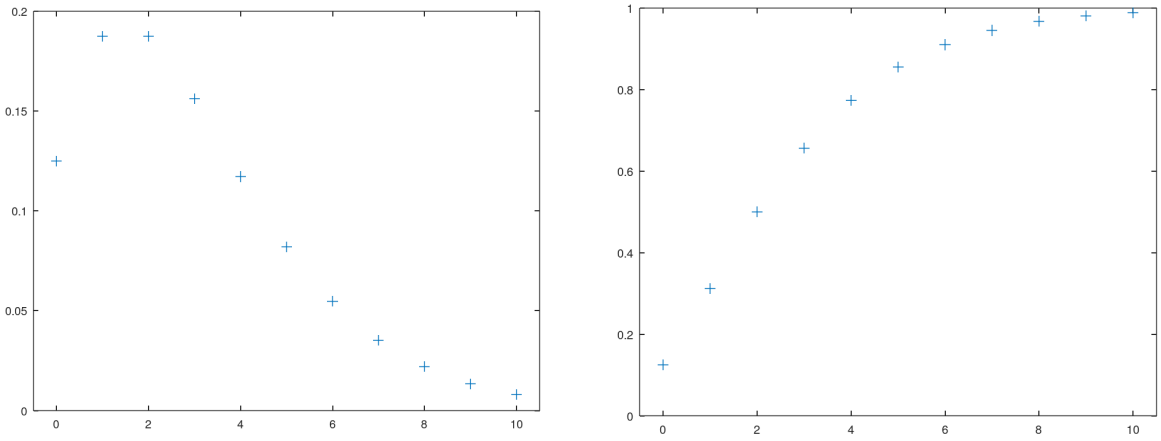
   | `nbinCDF(10, 5, 0.25)` | 0.3135

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

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

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

### 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 16: 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`
∘ Online calculator *Bognar's app*

### 2.3.1 Exercises

**Exercise 75.** (user-defined function `nbinPDF` in plain MAXIMA)
Write a user-defined function `nbinPDF()` using the mathematical definition for $nbinPDF$ in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
Check your function on example.1.
*Solution*: `Ex.75`

**Exercise 76.** (Table and plot for nbinPDF)
a. Print 10 values for the negative-binomial PDF for $k = 1..10$ and $p = 1/6$.
b. Plot the negative-binomial PDF values from a. - if necessary by paper and pencil.
*Solution*: `Ex.76`

**Exercise 77.** (candy problem, cf.wiki)
Pat Collis is required to sell candy bars to raise money for the 6th grade field trip. Pat is (somewhat harshly) not supposed to return home until five candy bars have been sold. So the child goes door to door, selling candy bars. At each house, there is a 0.6 probability of selling one candy bar and a 0.4 probability of selling nothing.
a. What's the probability of selling the last candy bar at the $n$-th house?
b. What's the probability that Pat finishes on or before reaching the eighth house?
c. What's the probability that Pat exhausts all 30 houses that happen to stand in the neighborhood?
*Solution*: `Ex.77`

**Exercise 78.** (user-defined function `nbinCDF` in plain MAXIMA)
Write a user-defined function `nbinCDF()` using the mathematical definition for $nbinCDF$ in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
Check your function on the example.2.
*Solution*: `Ex.78`

**Exercise 79.** (Table and plot for nbinPDF)
a. Print 10 values for the negative-binomial CDF for $n = 1..10$ and $p = 1/6$.
b. Plot the negative-binomial CDF values from a. - if necessary by paper and pencil.
c. Do a plot for the negative-binomial CDF for $n = 40$ and $p = 0.5$.
*Solution*: `Ex.79`

**Exercise 80.** (user-defined function `nbinINV`)
Write a user-defined function `nbinINV()` using the mathematical definition for $nbinINV$ in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
a. Check your function on example.3.
b. Verify, that your function solves `nbinINV(0.99,5,0.25) = 38`
and `nbinINV(0.95, 10, 0.7) = 9` correctly.
*Solution*: `Ex.80`

**Exercise 81.** (inverse probability)
Verify, that the inverse probability at $pin = 0.25, p = 0.6835$ is 3.

## 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 $\alpha$.

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

**Examples**

1. *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? *Check.*

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

2. *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 two defective floppies if you select 10 at random? *Check.*

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

3. *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?

   *Check.*

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

### Graphical representation

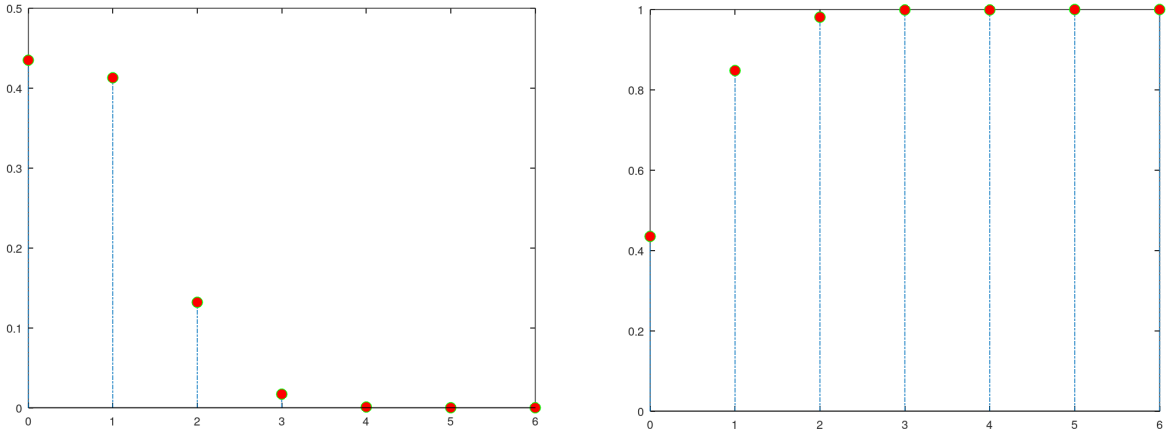



Figure 17: National lottery: from 49 including 6 red take 6 and get $x$ red.
**Left figure:** • = plot for `hgPDF(49,6,6,x)` for $x = 0, 1, 2, \ldots, 6$.
Check `hgPDF`$hgPDF(49, 6, 6, 2) = 0.139$ on the graph.
**Right figure:** • = plot for `hgCDF(49,6,6,x)` for $x = 0, 1, 2, \ldots, 6$.
Check `hgCDF`$(49, 6, 6, 1) = 0.84$ 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`
○ Online calculator *Bognar's app*

### 2.4.1 Exercises

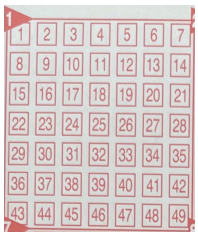
Unfortunately, the parameter list for the set of hypergeometric functions varies from one software package to another; therefore, we will address this peculiarity in the exercises – specifically regarding $\mathcal{R}$, MatLab, and `Excel`. We have the

| LEXICON | *this book* | distrib |
|---|---|---|
| follows convention of | MatLab | $\mathcal{R}$ |
| parameter list | $N, R, n, r$ | $x, n1, n2, n$ |
| mental model | urn | |

**Exercise 82.** (a user-defined function `hgPDF` in plain Maxima)
Write a user-defined function `hgPDF()` using the mathematical definition for $hgPDF$ in 'plain' Maxima, i.e. without using build-in functions from Maxima packages.
Check your function on example.1.
*Solution*: `Ex.82`

**Exercise 83.** (Table and plot for hgPDF)
a. Print 10 values for the hypergeometric PDF for $k = 1..10$ and $p = 1/6$.
b. Plot the hypergeometric PDF values from a. - if necessary by paper and pencil.
*Solution*: `Ex.83`

**Exercise 84.** (national lottery)
In a national lottery box there are 49 numbered balls.



6 balls are drawn.
a. What is the probability to predict 4 numbers correct of the 6?
b. Calculate the probabilities for exact $k = 0, .., 6$ 'red balls' (successes).
c. Plot the hypergeometrical PDF for the lottery outcomes.
*Solution*: `Ex.84`

**Exercise 85.** (example from MatLab, hygepdf) Reproduce

```
MATLAB> p = hygepdf(0:5,100,20,10)
        p = 0.0951  0.2679  0.3182  0.2092  0.0841  0.0215
```

*Solution*: `Ex.85`

**Exercise 86.** (Wikipedia ibid.: working example)
The WikiPedia definition of the hypergeometric PDF definition use another parameter sequence:

```
hgPDFwiki(r,N,R,n) = hgPDF(N,R,n,r)
```

We are interested in calculating the probability of drawing $r$ red marbles in $n$ draws, given that there are $R$ red marbles out of a total of $N$ marbles.

For this example, assume that there are 5 red and 45 green marbles in the urn. Standing next to the urn, you close your eyes and draw 10 marbles without replacement.
What is the probability that exactly 4 of the 10 are red?
*Solution*: `Ex.86`

**Exercise 87.** (example with MS EXCEL, cf. Excel)
MS EXCEL use yet another parameter sequence for the hypergeometric PDF:

```
hgPDFexcel(x, n, k, m) = hgPDF(m,k,n,x)
```

in words: `hgPDFexcel(x,n,k,m)` is the probability of getting $x$ successes from a sample of size $n$, where the size of the population is $m$ of which $k$ are successes.
EXAMPLES:
a. A bag contains 12 balls, 8 red and 4 blue. You reach into the bag and pick 3 balls at random without replacement.
What is the probability that at least 2 of the balls will be blue?
b. A warehouse contains 500 used computers. A random sample of 100 of these is tested and three of them are found to be defective.
What is the most likely percentage of defective computers in the warehouse?
*Solution*: `Ex.87`

**Exercise 88.** (user-defined function `hgCDF` in plain MAXIMA)
Write a user-defined function `hgCDF()` using the mathematical definition for $hgCDF$ in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
Check your function on the example.2.
• *Description*: $hgCDF(N, R, n, x)$ computes the hypergeometric CDF at the value in $x$ using the corresponding size of the population $N$, the number of the desired items in the population $R$, and the number of samples drawn $n$.
*Solution*: `Ex.88`

**Exercise 89.** (table and plot of the hypergeometric CDF)
a. Print 9 values for the cumulative hypergeometric CDF for $n = 1..9$ and the parameter set $N = 100, R = 20, n = 10$.
b. Plot the hypergeometric CDF values from a. - if necessary by paper and pencil.
c. Do a plot for the hypergeometric CDF for $n = 40$ and $p = 0.5$.
*Solution*: `Ex.89`

**Exercise 90.** (defective floppy disks, cf. hygecdf)
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?
*Solution*: `Ex.90`

**Exercise 91.** (draw at least 3 numbers right)
What is the probability to draw at least 3 numbers right by drawing of 6 out of 49?
*Solution*: `Ex.91`

**Exercise 92.** (user-defined function `hgINV`)
Write a user-defined function `hgINV()` using the mathematical definition for $hgINV$ in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.

Check your function on example.3.
*Solution*: `Ex.92`

• This function is used to find the n.th quantile, that is if P(x ¡= k) is given, it finds k.
• $hgINV(alpha, N, R, n)$ returns the smallest integer $x$ such that the hypergeometric CDF evaluated at $x$ equals or exceeds $P$. You may think of $alpha$ as the probability of observing $x$ defective items in $n$ drawings without replacement from a group of $N$ items where $R$ are defective.

**Exercise 93.** (Table and plot for hgINV)
a. Print 10 values for the hypergeometric INV for $\alpha = 0.1 \ldots 1$ *step* 0.1
and $N = 100, R = 10, n = 50$.
b. Plot the hypergeometric INV values from a. - if necessary by paper and pencil.
c. Do a plot for the hypergeometric INV for $n = 40$ and $p = 0.5$.
*Solution*: `Ex.93`

**Exercise 94.** (A manufacturer produces shirts, cf. real-statistics)
A manufacturer produces shirts in batches of 3,000. They decide to sample 50 shirts from each batch and count the number of defective shirts in each sample. If they want to make sure that 99.5% of the batches have no more than 20 defects,
What is the maximum number of defects that they should allow in each sample?
*Solution*: `Ex.94`

**Exercise 95.** (example from $\mathcal{R}$)
Warning: watch the other parameter list (same as in *distrib* ;)
Reproduce in MAXIMA:

```
R> qhyper(p, m, n, k, lower.tail = TRUE, log.p = FALSE)
R> qhyper(p = 0.90, m = 5, n = 11, k = 4)
   [1] 2
```

*Solution*: `Ex.95`

## 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 $\lambda$ 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:

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

  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:

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

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

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

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

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

**Examples**

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

   | `poissonPDF(2,3)` | 0.2240

2. *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. *Check.*

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

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

| `poissonINV(0.95, 2)` | 5 |
|---|---|

### Graphical representation

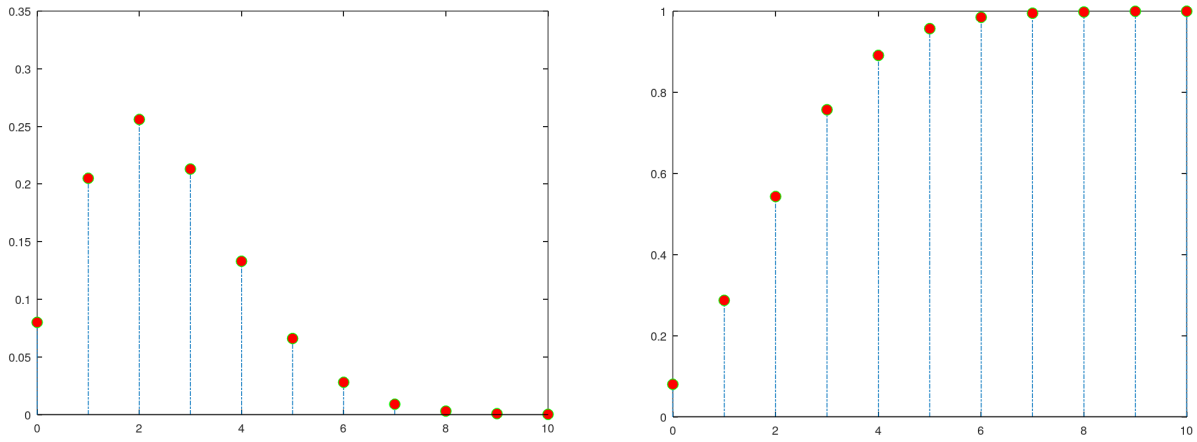



Figure 18: Plot of Poisson distribution for parameter $\lambda = 2.5$
**Left figure:** • = plot of `poissonPDF(k,2.5)` for $k = 0, \ldots, 10$.
Check `poissonPDF`$(2.5, 3) = 0.21$ on the graph.
**Right figure:** • = plot of `poissonCDF(k,2.5)` for $k = 0, \ldots, 10$.
Check `poissonCDF`$(2.5, 3) = 0.75$ 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 : `poisspdf`
○ Online calculator *Bognar's app*

### 2.5.1 Exercises

**Exercise 96.** (a user-defined function `poissonPDF` in plain MAXIMA)
Write a user-defined function `poissonPDF()` using the mathematical definition for *poissonPDF* in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
Check your function on example.1.
*Solution*: `Ex.96`

**Exercise 97.** (Table and plot for poissonPDF)
a. Calculate the 11 values of the poissonPDF shown in fig.18.left.
b. Plot the poisson PDF values from a. - if necessary by paper and pencil. It's fig.18.
*Solution*: `Ex.97`

**Exercise 98.** (examples from wiki, cf. poisson)
Solve the examples ("Beispiele") in [ibid.]
*Solution*: `Ex.98`

**Exercise 99.** (examples from Hermann [7, p.42])
a. On average, a book contains one typo error on every page. What is the probability that there are two typographical errors on a given page? (solution: 18.4 %)
b. What is the probability that at a party with 25 people, everyone will have different birthdays?
*Hint*: 25 people gives $\binom{25}{2} = 300$ birthday-pairs. So $\lambda = 300/365$. (Solution: 56%)
*Solution*: `Ex.99`

**Exercise 100.** (example from Hermann [7, p.42]
In a city, there are an average of two accidents per day.
What is the probability of more than four accidents per day?
Solution: *lambda* = 2. Pr(X¡4)=1-Pr(X¡=4) = .. =5.2%
*Solution*: `Ex.100`

**Exercise 101.** (user-defined function `poissonCDF` in plain MAXIMA)
Write a user-defined function `poissonCDF()` using the mathematical definition for *poissonCDF* in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
Check your function on the example.2.
• An alternative implementation uses the incomplete Gamma function, cf. package *distrib*.
*Solution*: `Ex.101`

**Exercise 102.** (table and plot of the poisson CDF)
a. Calculate the 11 values of the poissonCDF shown in fig.18.right.
b. Plot the poisson CDF values from a. - if necessary by paper and pencil.
*Solution*: `Ex.102`

**Exercise 103.** (user-defined function `poissonINV`)
Write a user-defined function `poissonINV()` using the mathematical definition for $poissonINV$ in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
Check your function on example.3.
*Solution*: `Ex.103`

**Remark.**
• This function is used to find the n.th quantile, that is if P(x ¡= k) is given, it finds k.
• `X=poissonINV(alpha, lambda)` returns the smallest value $X$ such that the Poisson CDF evaluated at $X$ equals or exceeds $P$, using mean parameter in lambda. [ibid.]

**Exercise 104.** (Table and plot for poissonINV)
a. Print 10 values for the poisson INV for $\alpha = 0.1 \ldots 1$ *step* 0.1 and $\lambda = 2.5$.
b. Plot the poisson INV values from a. - if necessary by paper and pencil.
*Solution*: `Ex.104`

**Exercise 105.** (defects in a machine part, cf. poissinv)
Suppose the average number of defects in a machine part is two.
a.Use the inverse of the Poisson distribution to determine the 95th percentile of the number of defects.
b. Calculate the median number of defects per machine part.
*Solution*: `Ex.105`

⋈

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

⋈

# 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[14]

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

- The *cumulative* normal distribution is defined by[15]

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

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

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

  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 $\Phi^{-1}$ using $\Phi$ and the defining relation $x = \Phi^{-1}(p)$ with $\Phi(x) = p$ and a simple search method..

**Examples**

1. *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? *Check.*

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

[14] we adopt the notation convention of MatLab , i.e. `normXXX` with `XXX` ∈ {`PDF`,`CDF`,`INV`}
[15] using the build-in error-function `erf`

2. *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. [7, p. 50] *Check*.

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

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

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

### Graphical representation

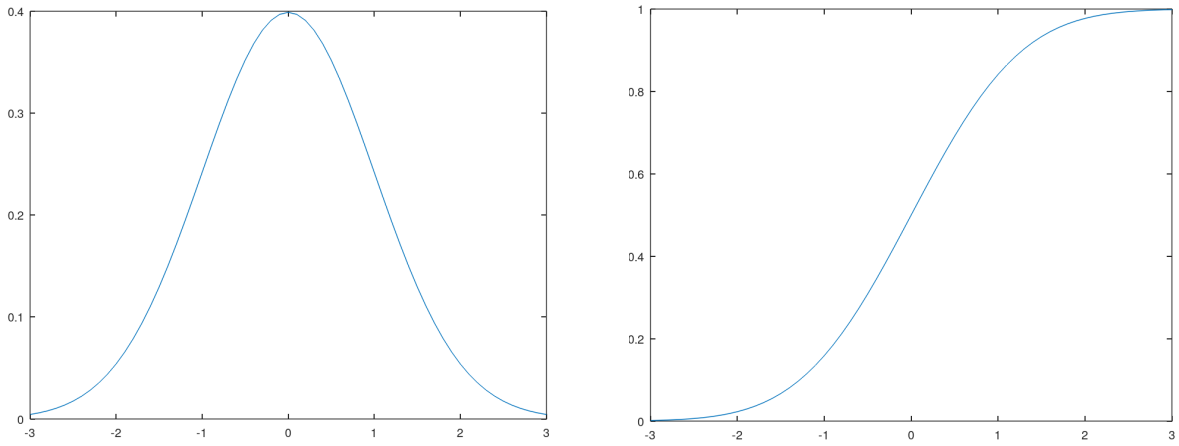



Figure 19: Standard normal distribution with parameters $\mu = 0$ and $\sigma$ equal to 1.
**Left** figure: ≀ = plot of `stdnormPDF(x)`.
Check `stdnormPDF`$(1) = 0.24$ on the graph.
**Right** figure: ≀ = plot of `stdnormCDF(x)`.
Check `stdnormCDF`$(1) = 0.84$ on graph.
Check `stdnormINV`$(0.6) \approx 0.25$ 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 : `normpdf`
○ Online calculator *Bognar's app*

### 3.1.1 Exercises

**Exercise 106.** (a user-defined function normPDF in plain MAXIMA)
A user-defined function `normPDF()` make use of the mathematical definition for $\phi(x, \mu, \sigma)$, no build-in function from a MAXIMA package is needed.
Check your function on example.1.
*Solution*: `Ex.106`

**Exercise 107.** (Table and plot for [standard] normal PDF)
a. Reproduce the following $\mathcal{R}$ session in MAXIMA, where `dnorm`|$\mathcal{R}$ is the density of the normal distribution, i.e. our normal PDF.

```
R> # Create 10 numbers between 0 and 2 incrementing by 0.1:
R> x <- seq(0, 1, by = .1)
R> # Choose the mean as 0 and standard deviation as 1:
R> dnorm(x, mean = 0, sd = 1)
   [1] 0.39894228 0.39695255 0.39104269 0.38138782 0.36827014 ..
```

b. Plot the PDF of the Standard Normal distribution $f(x) = normPDF(x, 0, 1)$ for $xrange = (-5, 5)$ and $yrange = (0, 1)$.
*Solution*: `Ex.107`

**Exercise 108.** (user-defined function normCDF in plain MAXIMA)
A user-defined function `normCDF()` make use of the mathematical definition for $\Phi(x, \mu, \sigma)$, using the build-in function `erf` in plain MAXIMA.
Check your function on example.2.
*Solution*: `Ex.108`

**Exercise 109.** (table and plot of the normal CDF)
a. Calculate the 7 values of the normCDF at $-3, -2, .., 2, 3$ shown in fig.19.right.
b. Plot the normal CDF values from a. - if necessary by paper and pencil.
*Solution*: `Ex.109`

**Exercise 110.** (examples from MATLAB)
Construct a table for normCDF for x = -2,-1,0,1,2 and $\mu = 2; \sigma = 1$ and verify the MATLAB session:

```
MatLAB> x = [-2,-1,0,1,2]; mu = 2; sigma = 1;
      > p = normcdf(x,mu,sigma)
        p = 1x5
            0.0000    0.0013    0.0228    0.1587    0.5000
```

*Solution*: `Ex.110`

**Exercise 111.** (sugar in packs from [7, p.50]
A manufacturer fills sugar in packs, which have a distribution of $\mathcal{N}(1000, 12^2)$.
What percentage of the packs meet the printed minimum weight of 980 $g$?
*Solution*: `Ex.111`

**Exercise 112.** (examples from $\mathcal{R}$)
a. Calculate the probability of a value less than or equal to 1.96 in a standard normal distribution. `[ R> pnorm(1.96) [1] 0.9750021 ]`
b. Determine the probability of a value less than or equal to 2 in a normal distribution with mean 1 and standard deviation 0.5? `[ R> pnorm(2, mean = 1, sd = 0.5) [1] 0.9772499`
*Solution*: `Ex.112`

**Exercise 113.** (resistors from Hermann [7, p.51]
A factory produces 100 Ω resistors, distributed via $\mathcal{N}(100, 6^2)$.
What proportion of the resistors have a maximum deviation of 10% from the mean value?
*Solution*: `Ex.113`

**Exercise 114.** (assortment of screws from [7, p.42]
From an assortment of screws distributed via N(10,0.3$\hat{2}$) in a all screws larger than 10.5 mm or smaller than 9.5 mm are removed.
What is the probability that one of the remaining screws is smaller than 10.3 mm?
*Solution*: `Ex.114`

**Exercise 115.** (men's height from [7, p.50]
When measuring men, the mean height was 1.73 cm with a spread of 7 cm. Men between the heights of 166 cm and 180 cm are therefore considered to be of 'normal' height.
What is the probability that a man is shorter than 1.85 m according to this distribution?
*Solution*: `Ex.115`

**Exercise 116.** (alternativ definition of standard normal distribution using `erfc(z)`)
`erfc(z)` is the so-called Complementary Error Function. This function is used by MatLab to build `normcdf()`, cf. normcdf.
a. Write function `stdnormCDF1` via `erfc(z)`:

```
stdnormCDF1(x)        = 1/2*erfc(-x/sqrt(2))
normCDF1(x,mu,sigma) = stdnormCDF1((x-mu)/sigma)
```

b. Solve exercise.115 using `normCDF1`.
*Solution*: `Ex.116`

**Exercise 117.** (user-defined function `normINV`)
Look up the build-in function `normINV()`, called 'quantile_normal' in package 'distrib' and so realizing the mathematical definition for $normINV$.
a. Check your function on example.3.
*Solution*: `Ex.117`

**Remark.**
• This function is used to find the $n^{th}$ quantile, that is if $P(x \leq k)$ is given, it finds $k$.
• The inverse normal distribution or normINV allows you to find the value of $x$ of a normal distribution given a probability $p$, the mean $\mu$ and the standard deviation $\sigma$ of the

distribution. Simply put, it works backward from the area under the normal distribution curve to find the x-value corresponding to that area.
• There is no simple algebraic formula; instead, the one may use an iterative search technique to find the $x$, that satisfies `p=normCDF(x,mu,sigma)`. The standard normal version is a special case where the mean is 0 and the standard deviation is 1.

*b. `X=normalINV(alpha, lambda)` returns the smallest value $X$ such that the normal CDF evaluated at $X$ equals or exceeds $P$, using mean parameter in lambda.
Use this algorithm to implement the inverse of the standard cumulative normal distribution using an iterative bisection method. Pseudocode:

```
stdnormINV(x, a,b,k,t,y) = test(x=0, "-Inf", x=1, "+Inf", and(0<x,x<1),
               do( a = -10, b = 10,    -- x in [a,b]
                   for(k,1,100,  t = (a+b)/2,
                                 y = stdnormCDF(t),
                       test(abs(x-y)<0.0000001, break,
                       y<x,a=t, b=t)),
                   float(t) ))
```

or use the implementation of 'quantile_binomial' in package 'distrib' as pattern:

```
quantile_binomial(q,n,p):=  /* partition method */
      block([a:0, b:n, m],
        while (b-a>1) do (
               m: 0.5*(a+b),
               if cdf_binomial(m,n,p) < q then a: m else b: m ),
        floor(b))$
```

**Exercise 118.** (Table and plot for normINV)
a. Print 9 values for the normal INV for $\alpha = 0.1 \ldots 0.9$ *step* 0.1.
b. Plot the normal INV values from a. - if necessary by paper and pencil.
*Solution*: `Ex.118`

**Exercise 119.** (MatLab: norminv)
Compute the inverse of cdf values evaluated at the probability values in $p$ for the normal distribution with mean mu=2 and standard deviation sigma=1.

```
p = 0:0.25:1; mu = 2; sigma = 1;
norminv(p,mu,sigma)
ans = -Inf  1.3255   2.0000   2.6745    Inf
```

*Solution*: `Ex.119`

**Exercise 120.** (Compare with R qnorm list)
Reproduce wit MAXIMA:

```
R> # Create 10 numbers between 0 and 2 incrementing by 0.1.
R> x <- seq(0, 1, by = .1)
R> # Choose the mean as 0 and standard deviation as 1.
R> qnorm(x, mean = 0, sd = 1)
  [1] -Inf -1.2815516 -0.8416212 -0.5244005 -0.2533471  0.0 ..
```

*Solution*: `Ex.120`

**Exercise 121.** (confidence intervall, cf. MATLAB)
Find an interval that contains 95% of the values from a standard normal distribution.

```
Matlab> x = norminv([0.025     0.975])
        x = 1x2     -1.9600    1.9600
                    Interval =[-1.96, 1.96]
```

Use MAXIMA.
*Solution*: `Ex.121`

**Exercise 122.** (some Quantile of normal/GAUSS distribution)
Verify, that for $X = (0.9, 0.95, 0.975, 0.99)$ we get the following quantile of the normal aka GAUSS distribution: $R = (1.281552, 1.644854, 1.959964, 2.326348)$ in CAS $\mathcal{R}$.
*Solution*: `Ex.122`

**Exercise 123.** (The 68-95-99.7 rule)
Let us determine the area under the standard normal curve for $\pm 1$ standard deviation, for $\pm 2$ standard deviations, and for $\pm 3$ standard deviations.

**Remark.** Awesome, we just confirmed the Empirical Rule, also known as the "68-95-99.7" rule, which relates to the CHEBYSHEV's theorem.
For a bell-shaped distribution the 3 rules are, that approximately
∘ 68% of the observations lie within 1 std.deviation of the mean,
∘ 95% of the observations lie within 2 std.deviations of the mean,
∘ 99.7% of the observations lie within 3 std.deviations of the mean.
*Solution*: `Ex.123`

## 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*[16] (quantile) Exponential distribution is

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

**Examples**

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

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

2. *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? *Check*.

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

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

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

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

### Graphical representation

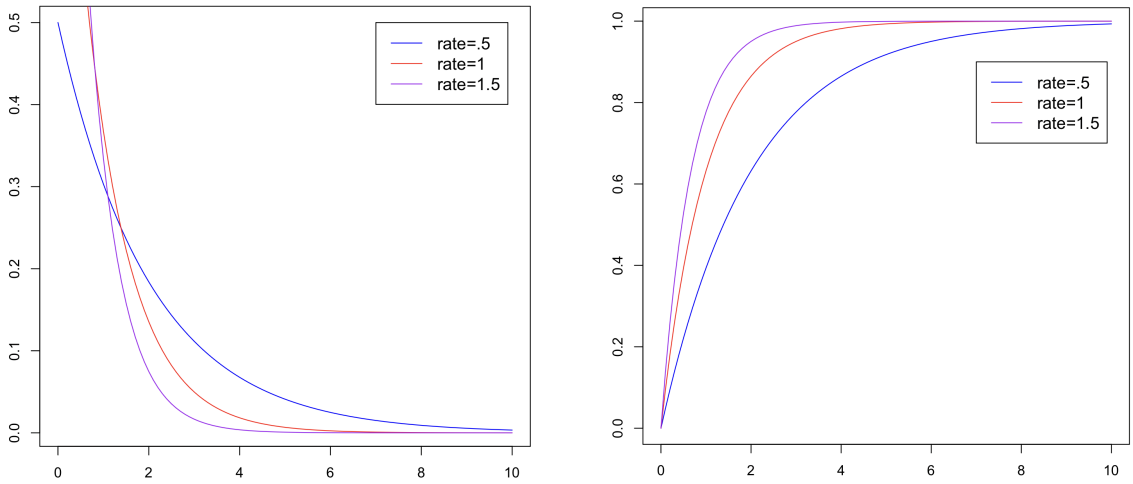



Figure 20: Exponential distributions with different rates $\lambda$.
**Left** figure: ≀ = plot of `expPDF(x,λ)`, $\lambda = 1; 1.5; 5$.
Check `expPDF`$(1) = 0.24$ on the graph.
**Right** figure: ≀ = plot of `expPDF(x,λ)`, $\lambda = 1; 1.5; 5$ .
Check `expPDF` $(2, 1) \approx 0.14$ on the graph.
Check `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`
∘ Online calculator *Bognar's app*

### 3.2.1 Exercises

**Exercise 124.** (a user-defined function expPDF in plain MAXIMA)
Write a user-defined function `expPDF()` using the mathematical definition for $expPDF$ in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
`expPDF` returns the pdf of the exponential distribution with rate $L$, evaluated at the values in $x$.
a. Check your function on example.1.
b. Verify: `R> dexp(5,1/2)  [1] 0.0410425`
*Solution*: `Ex.124`

**Exercise 125.** (Table and plot for expPDF)
a. Calculate the 11 values of the expPDF shown in fig.20.left for $\lambda = 5$.
b. Plot the expPDF values from a. - if necessary by paper and pencil. It's fig.18.
*Solution*: `Ex.125`

**Exercise 126.** (lifetime of a product; from [?, p.68])
The lifetime of a product with the parameter $L = 1/2$ is exponentially distributed.
What is the probability that the product will fail in a period between three and five years?
*Solution*: `Ex.126`

**Exercise 127.** (user-defined function expCDF in plain MAXIMA)
Write a user-defined function `expCDF()` using the mathematical definition for $expCDF$ in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.
Check your function on the example $expCDF(4, 2)$.
*Solution*: `Ex.127`

**Exercise 128.** (table and plot of the expCDF)
a. Calculate the 11 values of the expCDF shown in fig.18.right.
b. Plot the expCDF values from a. - if necessary by paper and pencil.
*Solution*: `Ex.128`

**Exercise 129.** (lifespan of a light bulb, cf. [7, p.63])
The lifespan of a light bulb is exponentially distributed and averages 1,000 hours.
What is the probability, that it will last 800 hours?
*Solution*: `Ex.129`

**Exercise 130.** (mean time between failures, cf. [7, p.63])
A printer has a mean time between failures ('MTBF') of 2,000 hours.
What is the probability that the first 500 hours of the trip will be trouble-free?
*Solution*: `Ex.130`

**Exercise 131.** (refrigeration unit, cf. [7, p.63])
A refrigeration unit has a mean exponentially distributed lifespan of ten years.
What is the probability that it will last another five years, if it survives the first five years?
*Solution*: `Ex.131`

**Exercise 132.** (user-defined function `expINV`)
Write a user-defined function `expINV()` using the mathematical definition for $expINV$ in 'plain' MAXIMA, i.e. without using build-in functions from MAXIMA packages.

**Remark.**
• from MATLAB: The result $x$ is the value such that an observation from an exponential distribution with parameter $m$ will falls in the range $[0, x]$ with probability $p$.
• the quantile function (or inverse CDF) is given by $F^{-1}(u) = -(1/L) * ln(1-u)$, where $0 < u < 1$. This inverse function takes a uniform random variable (between 0 and 1) as input and produces a value that would be expected from the exponential distribution.
• This function is used to find the $n^{th}$ quantile, that is if $P(x \leq k)$ is given, it finds $k$.
• `x=expINV(p, lambda)` returns the smallest value $x$ such that `expCDF` evaluated at $x$ equals or exceeds $p$, using mean parameter in lambda.

Check your function on example.3.
*Solution*: `Ex.132`

**Exercise 133.** (Table and plot for expINV)
a. Print 10 values for the expINV for $p = 0.1 \ldots 1$ *step* 0.1 and $\lambda = 2.5$.
b. Plot the expINV values from a. - if necessary by paper and pencil.
*Solution*: `Ex.133`

**Exercise 134.** (median lifetime, from expinv)
Assume that the lifetime of light bulbs are exponentially distributed with a mean of 700 hours.
Find the median lifetime using expinv.
Verify:

```
MATLAB> expinv(0.50,700)
        ans = 485.2030
.. Half of the light bulbs will burn out within the first 485 hours of use.

R> qexp(0.5, 1/700)
   [1] 485.203
```

*Solution*: `Ex.134`

## 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 $\nu$ 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

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

  $u_\alpha$ is the $\alpha$-quantile of the standard normal distribution.

  We give this approximation formula by Peizer & Pratt, [7, p. 70] only as reference. [17] 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.

[17] 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

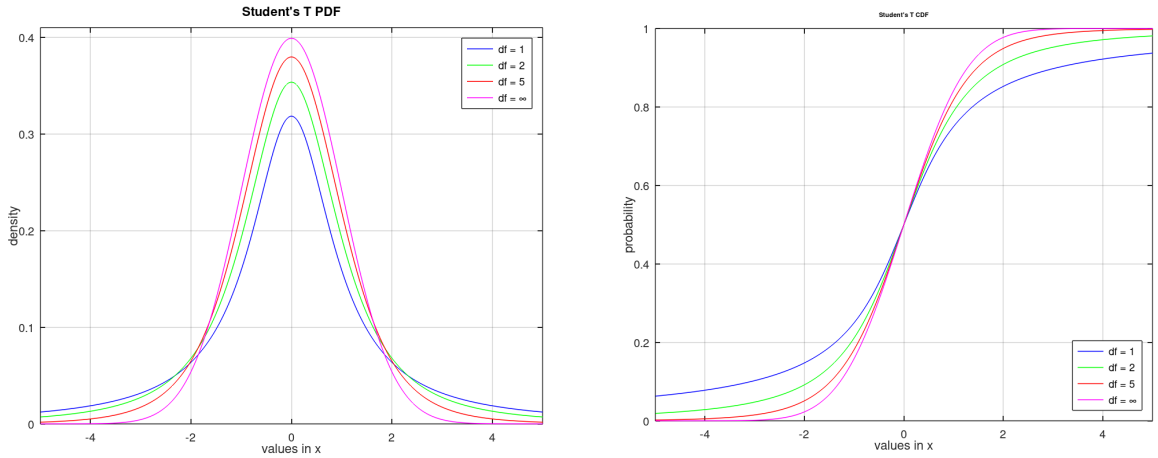



Figure 21: Student's t distributions with different degrees of freedom $df := \nu$.
**Left** figure: ≀ = plot of `tPDF(x,`$\nu$`)`, $\nu = 1; 2; 5$.
Check `tPDF`$(0, 1) = 0.3$ on the graph.
**Right** figure: ≀ = plot of `tCDF(x,`$\nu$`)`, $\nu = 1; 2; 5$.
Check `tCDF` $(0, 1) \approx 0.5$ on the graph.
Check `tINV`$(0.2, 1) \approx -1$ on the CDF via $y = 0.2 \longrightarrow\downarrow -1 = x$

### Examples

1. *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. *Check*.

   | `tPDF(0,3)` | 0.3675

2. *tCDF*: (MatLAB) Determine the probability that an observation from the Student's t distribution with degrees of freedom 99 falls on the interval $[10, .., \texttt{Inf}]$. *Check*.

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

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

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

### General information

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

### 3.3.1 Exercises

**Exercise 135.** (a user-defined function tPDF in plain MAXIMA)
A user-defined function `tPDF()` for `student_t` distribution using the mathematical definition for $tPDF$ in 'plain' MAXIMA needs the special function $Beta(a, b)$.
There is no simple term. Look e.g. at the definition in *distrib* by M. RIOTORTO.
Check our function on example.1.
*Solution*: `Ex.135`

**Exercise 136.** (Table and plot for tPDF)
a. Calculate the 11 values of the tPDF shown in fig.21.left.
b. Plot the tPDF values from a.
*Solution*: `Ex.136`

**Exercise 137.** (degree of freedom table for Student's t distribution from MATLAB)
The value of the pdf at the mode is an increasing function of the degrees of freedom. Let the mode of the Student's t distribution at x = 0.
Compute the pdf at the mode for degrees of freedom 1 to 6 with MAXIMA.

```
MATLAB>  tpdf(0,1:6)
         ans =  0.3183  0.3536  0.3676  0.3750  0.3796  0.3827
```

*Solution*: `Ex.137`

**Exercise 138.** (t distribution converges to the standard normal distribution, cf. poisson)
The t distribution converges to the standard normal distribution as the degrees of freedom approach infinity.
Compute the difference between the pdfs of the standard normal distribution and the Student's t distribution pdf with 30 degrees of freedom.
*Solution*: `Ex.138`

**Exercise 139.** (user-defined function tCDF in plain MAXIMA)
A user-defined function `tCDF()` for cumulate `student_t` distribution using the mathematical definition for $tCDF$ in 'plain' MAXIMA needs the special function $betainc(a, b, c)$.
There is no simple term. Look e.g. at the definition in *distrib* by M. RIOTORTO.
Check your function on the example.2.
*Solution*: `Ex.139`

**Exercise 140.** (table and plot of the student-t-CDF)
a. Calculate the 11 values of the tCDF shown in fig.21.right.
b. Plot the poisson CDF values from a. - if necessary by paper and pencil.
*Solution*: `Ex.140`

**Exercise 141.** (example from MATLAB, cf. tcdf)
Determine the probability that an observation from the Student's t distribution with degrees of freedom 99 falls on the interval [10 Inf].
*Solution*: `Ex.141`

**Exercise 142.** (user-defined function tINV)
A user-defined function `tINV()` for inverse `student_t` distribution using the mathematical definition for $tINV$ in 'plain' MAXIMA needs the special function $iibeta(a, b, c)$.
There is no simple term. Look e.g. at the definition in *distrib* by M. RIOTORTO.
heck your function on example.3.
*Solution*: `Ex.142`

**Remark.**
• This function is used to find the n.th quantile, that is if P(x ¡= k) is given, it finds k.
• `X=poissonINV(alpha, lambda)` returns the smallest value $X$ such that the Poisson CDF evaluated at $X$ equals or exceeds $P$, using mean parameter in lambda. [ibid.]

**Exercise 143.** (user-defined search method for `tINV`)
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.
Check your function on example.3.

**Exercise 144.** (an approximation for tINV)
Because there is no simple term for the inverse `student_t` distribution `tINV()` we try the following approximation after PREIZER & PRATT, cf. [7, p.71]:

```
tINV1(x,f,  q,c,t) =                -- 0<x<1
  do( PI = 3.1415926535898,
    test( f==1, tan(PI*(x-0.5)),
          f==2, do(t=2.*x-1., sqrt(2.)*t/sqrt(1.-t*t) ),
          do( q = normINV(x),
              c = f -2/3 + 1/(10*f),
              c = (f-5/6)/(c*c),
              t = sqrt(f*exp(c*q*q)-f),
              test(x>0.5, t, -t) )))
```

Write `tINV1()` in MAXIMA and test it w.r.t.

```
MATLAB> tinv(.95,50)       ans = 1.6759
     R> qt(.95,50)         [1] 1.675905
```

*Solution*: `Ex.144`

**Exercise 145.** (Table and plot for tINV)
a. Print 10 values for the tINV for $\alpha = 0.1 \ldots 1$ *step* 0.1 and $f = 5$.
b. Plot the tINV values from a.
*Solution*: `Ex.145`

**Exercise 146.** (the 99th percentile of the Student's t distribution, cf. tinv)
Compute the 99th percentile of the Student's t distribution for 1 to 6 degrees of freedom (df), i.e. verify

```
MATLAB> percentile = tinv(0.99,1:6)
       31.8205  6.9646  4.5407  3.7469  3.3649  3.1427
```

*Solution*: `Ex.146`

**Exercise 147.** (Quantiles of the t distribution, cf. [7, p.74ff])
Produce a table for the quantiles of t-distribution for $f = 1..20$ and $X = (0.9, 0.95, 0.975, 0.99)$.
The result should be

quantils of t-distribution for f=1..20

| f | 0.9 | 0.95 | 0.975 | 0.99 |
|---|---|---|---|---|
| 1 | 3.07768 | 6.31375 | 12.7062 | 31.8205 |
| 2 | 1.88562 | 2.91999 | 4.30265 | 6.96456 |
| 3 | 1.63774 | 2.35336 | 3.18245 | 4.5407 |
| 4 | 1.53321 | 2.13185 | 2.77645 | 3.74695 |
| 5 | 1.47588 | 2.01505 | 2.57058 | 3.36493 |
| 6 | 1.43976 | 1.94318 | 2.44691 | 3.14267 |
| 7 | 1.41492 | 1.89458 | 2.36462 | 2.99795 |
| 8 | 1.39682 | 1.85955 | 2.306 | 2.89646 |

...

*Solution*: `Ex.147`

## 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 $\alpha$. We do not use this approximation formula by E. PAULSON, [7, 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.

### Graphical representation

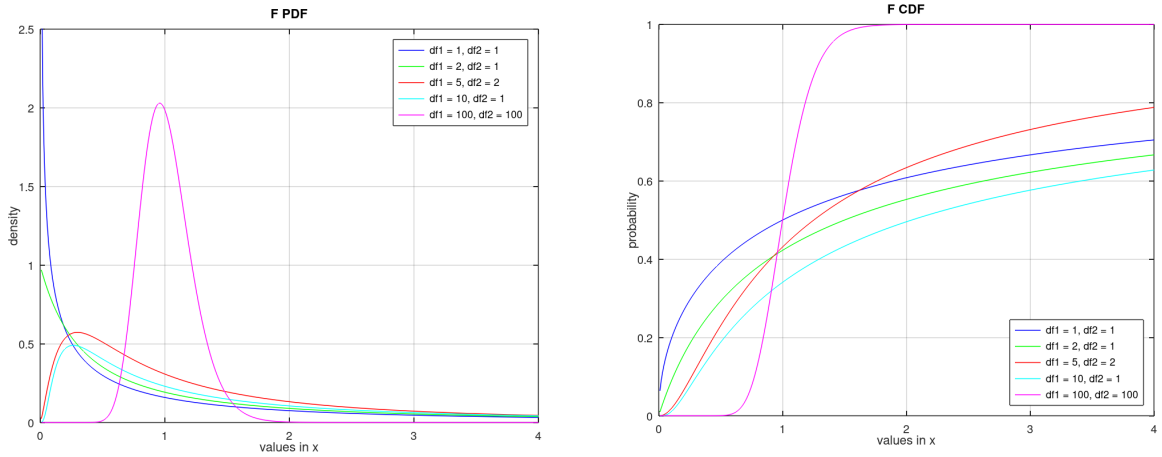



Figure 22: SNEDECOR's F distribution with different degrees of freedom $d_i$.
**Left** figure: ≀ = plot of `FPDF(x,`$d_1, d_2$`)`.
Check `FPDF`$(0, 1) = 0.3$ on the graph.
**Right** figure: ≀ = plot of `FCDF(x,`$d_1, d_2$`)`.
Check `FCDF` $(1, 5, 2) \approx 0.4$ on the graph.
Check `FINV`$(0.6, 100, 100) \approx 1$ on the CDF via $y = 0.6 \longrightarrow\downarrow 1 = x$

### Examples

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

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

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

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

3. *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? *Check*.

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

### General information

- General mathematical information about the concept is here ▷ wiki : *F distribution*
- Syntax and semantic of the function is here ▷ MATLAB : `fpdf`
- Online calculator *Bognar's app*

### 3.4.1 Exercises

**Exercise 148.** (a user-defined function fPDF in plain MAXIMA)
A user-defined function `fPDF()` for the `f` distribution using the mathematical definition for `fPDF` in 'plain' MAXIMA needs the special function $Beta(a, b)$ or $\Gamma(x)$.
There is no simple term. Look e.g. at the definition in *distrib* by M. RIOTORTO.
Check the function on example.1.
*Solution*: `Ex.148`

**Exercise 149.** (Table and plot for `fPDF`)
a. Calculate the 9 values of the `fPDF` plot (red') shown in fig.22.left.
b. Plot the `fPDF` values from a.
*Solution*: `Ex.149`

**Exercise 150.** (examples from MATLAB, cf. fpdf)
Verify the example from MATLAB:

```
MatLAB>  y = fpdf(1:6,2,2)
         y = 0.2500   0.1111   0.0625   0.0400   0.0278   0.0204
```

*Solution*: `Ex.150`

**Exercise 151.** (user-defined function fCDF in plain MAXIMA) A user-defined function `fCDF()` for the `f` distribution using the mathematical definition for `fCDF in 'plain'` MAXIMA `needs the special function` *betainc*$(a, b)$ `or` $\Gamma(x)$.
`There is no simple term. Look e.g. at the definition in` *distrib* `by M.` RIOTORTO.

`Check your function on example.2.` *Solution*: `Ex.151`

**Exercise 152.** (table and plot of the `fCDF`)
a. Calculate the 9 values of the `fCDF` shown in fig.22.right.
b. Plot the `fCDF` values from a.
*Solution*: `Ex.152`

**Exercise 153.** (user-defined function fINV)
A user-defined function `fINV()` for the `f` distribution using the mathematical definition for `fINV` in 'plain' MAXIMA needs the special function *betainc*$(a, b)$ or $\Gamma(x)$.
There is no simple term. Look e.g. at the definition in *distrib* by M. RIOTORTO.
Check your function on example.3.
*Solution*: `Ex.153`

**Remark.**
• This function is used to find the n.th quantile, that is if $P(x \leq k)$ is given, it finds $k$.
• `X=fINV(alpha, lambda)` returns the smallest value $X$ such that the `fCDF` evaluated at $X$ equals or exceeds $P$, using mean parameter in lambda. [ibid.]

**Exercise 154.** (Table and plot for `fINV` )
a. Print 10 values for the `fINV` for $\alpha = 0.1 \ldots 1$ *step* 0.1 and $\lambda = 2.5$.
b. Plot the `fINV` values from a.
*Solution*: `Ex.154`

**Exercise 155.** (Table from MatLAB)
Reproduce the following table with MAXIMA:

```
octave:1> pkg load statistics
octave:2> X     = [0.25, 0.5, 0.75, 0.999]
octave:3> finv(X,10,20)
          ans = 0.6564   0.9663   1.3995   5.0752
```

*Solution*: `Ex.155`

**Exercise 156.** (user-defined search method for `fINV`)
Implement the $f$ inverse function using `fCDF` 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, e.g.

```
--  works only if df1>6 & df2>4
fINV1(x,df1,df2, a,b,k,t,y) = do(
       a = -10.0, b = 10.0,
       for(k,1,100,  t = (a + b) / 2,
                     y = fCDF(t,df1,df2),
                     test(abs(x - y) < 0.000001, break,
                          y < x, a = t, b = t)),  t)
```

a. Write `fINV1()` in MAXIMA and test it w.r.t. the table of exercise.155.
b. WARNING: $df1 > 6$ and $df2 > 4$ must be fulfilled!
Check: `fINV(0.95,8,1)` w.r.t. `R> qf(0.95,8,1) [1] 238.8827`
ChecK `fINV(0.6,2,9)` w.r.t. `R> qf(0.6,2,9)  [1] 1.016246`
or `octave:1> finv(0.6, 2, 9)  ans = 1.0162`
*Solution*: `Ex.156`

**Exercise 157.** (an approximation for `fINV` after PAULSON)
Because there is no simple term for the inverse `f` distribution `fINV()` implement the approximation after PAULSON, cf. [7, p.71], cited in the definition above.
Write `fINV2()` in MAXIMA and test it w.r.t. `R> qt(.95,50)  [1] 1.6759`.
*Solution*: `Ex.157`

## 3.5 Chi-Square distribution

The $\chi^2$-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* $\chi^2$ is defined by

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

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

- The *cumulative* probability of $\chi^2$-distribution can be expressed by

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

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

- The *quantile* function (inverse cumulative of $\chi^2$-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 $\alpha$.
  $u_\alpha$ is the $\alpha$-quantile of the standard normal distribution. – We do not use this approximation formula by Wilson & Hilferty, [7, p. 73], for our implementation, but cite it here for your information.

  Instead we implement the $\chi$ inverse function using the $\chi$ 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

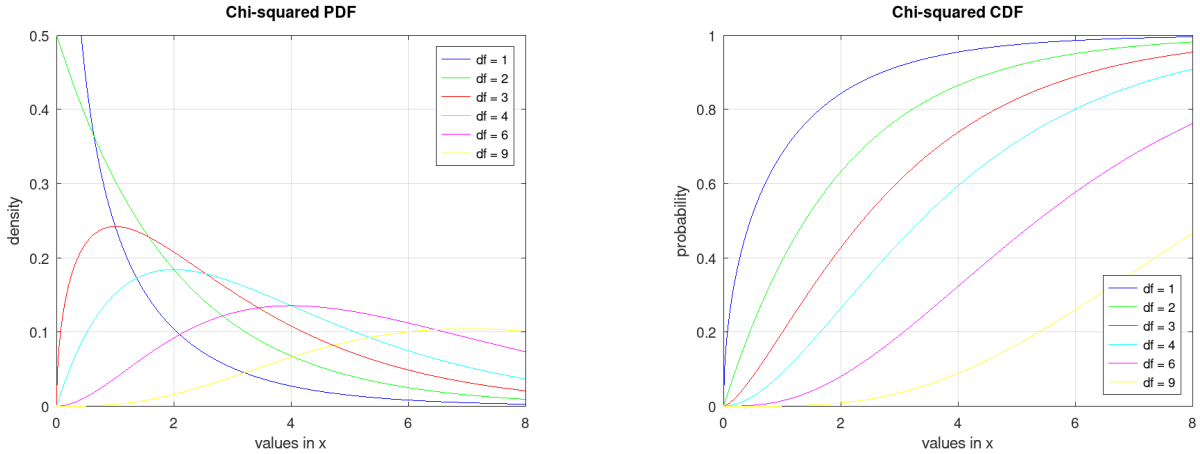



Figure 23: $\chi^2$-distribution with different degrees of freedom $k$. ▷wiki.
**Left** figure: ≀ = plot of `chiPDF(x,`$k$`)`, $k = 1; 2; 3, 4; 6; 9$.
Check `chiPDF`$(2, 4) = 0.18$ on the graph.
**Right** figure: ≀ = plot of `chiCDF(x,`$\nu$`)`, $\nu = 1; 2; ....$
Check `chiCDF` $(4, 4) \approx 0.6$ on the graph.
Check `chiINV`$(0.6, 4) \approx 4$ on the CDF via $y = 0.6 \longrightarrow\downarrow 4 = x$

### Examples

1. *chiPDF*: Let $X$ follow a $\chi^2$-distribution with 3 *df*. What is the density if the value of $X$ is 2? *Check*.

   | `chi2PDF(2, 3)` | 0.2075

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

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

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

   | `chi2INV(0.95, 10)` | 18.307

### General information

∘ General mathematical information about the concept is here ▷ Wiki : Chi squared
∘ Syntax and semantic of the function is here ▷ Matlab : `chi2`
∘ Online calculator *Bognar's app*

### 3.5.1 Exercises

**Exercise 158.** (a user-defined function chi2PDF in plain MAXIMA)
A user-defined function `chi2PDF()` for the `f` distribution using the mathematical definition for `chi2PDF` in 'plain' MAXIMA uses the special function $\Gamma(x)$.
Check the function on example.1.
*Solution*: `Ex.158`

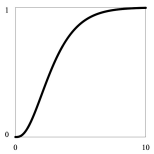
**Exercise 159.** (Table and plot for `chi2PDF`)
a. Do the following calculation in MAXIMA:

```
"    x       chi2PDF(x, 6)          Chi-squared table for df=6"
T = zero(2,9)
for(i,1,9, T[1,i] = 0.1*i)
for(i,1,9, T[2,i] = chi2PDF(0.1*i ,6))
transpose(T)
```

b. Calculate the values of the `chi2PDF` plot (red') shown in fig.23.left.
c. Plot the `chi2PDF` values from a.
*Solution*: `Ex.159`

**Exercise 160.** (user-defined function chi2CDF in plain MAXIMA)
A user-defined function `chi2CDF()` for the `chi2` distribution using the mathematical definition for `chi2CDF` in 'plain' MAXIMA needs the special function `GammaP(x)`, the incomplete Gamma function.
∘ There is no simple term. Look e.g. at the definition in *distrib* by M. RIOTORTO.
Check our function on example.2. *Solution*: `Ex.161`

**Exercise 161.** (table and plot of the `chi2CDF`)
a. Calculate the 9 values of the `chi2CDF` shown in fig.22.right.
b. Plot the `chi2CDF` values from a.
c. Check with MAXIMA: `octave> gammainc(3,2)  ans = 0.8009`
*Solution*: `Ex.162.c`
d. Plot the incomplete gamma function 'GammaP' for $k = 3$, cf. gammainc



**Exercise 162.** ($\chi^2$ table for $df = 3$)
Do the following calculation in MAXIMA:

```
Values for chi^2 with 3 degrees of freedom, i.e. df=3
 x   Chi^2(x)   P value (df=3)
 x   Chi^2cdf   P-value
 1   0.198748   0.801252
 2   0.427593   0.572407
```

```
 4   0.738536   0.261464
 8   0.953988   0.046012
16   0.998866   0.001134
32   0.999999   0.000001
```

*Solution*: Ex.162

**Exercise 163.** (user-defined function chi2INV)
A user-defined function `chi2INV()` for the `f` distribution using the mathematical definition for `chi2INV` in 'plain' MAXIMA needs the special function `iigamma`, the inverse incomplete gamma function, with source in `numdist.lisp`.
There is no simple term. Therfore we use the definition in *distrib* by M. RIOTORTO.
Check this function on example.3 and w.r.t.

```
MATLAB> chi2inv(0.95,10)              ans = 18.3070
     R> qchisq(x = 0.95, df = 10)     [1]   18.30704
```

*Solution*: Ex.163

**Remark.**
• This function is used to find the n.th quantile, that is if $P(x \leq k)$ is given, it finds $k$.
• `X=chi2INV(alpha, lambda)` returns the smallest value $X$ such that the `chi2INV` evaluated at $X$ equals or exceeds $P$, using mean parameter in lambda. [ibid.]

**Exercise 164.** ($\chi^2$ quantiles for $k = 3$ )
Let $X = (0.9, 0.95, 0.975, 0.99)$.
Calculate chi2INV$(x, 10)$ for $x$ in $(1, 2, 3, 4)$.
Result: $(15.987, 18.307, 20.483, 23.209)$.
*Solution*: Ex.164

**Exercise 165.** (Table from MatLAB)
Reproduce the following table with MAXIMA:

```
octave> chi2inv(0.1:0.1:1.0, 10)
ans = 4.8652 6.1791 7.2672 8.2955 9.3418  10.4732 11.7807 13.4420
      15.9872  Inf (is our right limit b=20 ;)
```

*Solution*: Ex.165

**Exercise 166.** (user-defined search method for `chi2INV`)
Implement the $f$ inverse function using `chi2INV` 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,.
Write `chiINV1()` in MAXIMA and test it w.r.t. the table of exercise.164.
*Solution*: Ex.166

**Exercise 167.** (an approximation for `chi2INV` after WILSON & HILFERTY)
Because there is no simple term for the inverse `chi2INV` distribution `fchi2NV()` implement the approximation after WILSON & HILFERTY, cf. [7, p.73], cited in the definition above.
Write `chi2INV2()` in MAXIMA and test it w.r.t. `R> qt(.95,50)  [1] 1.6759`.
*Solution*: Ex.167

## 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 $\alpha$ is a positive parameter. The Pareto distribution is characterized by this *scale* parameter $x_m$ and this *shape* parameter $\alpha$, 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 $\mu$ is the *scale* parameter and the *shape* parameter is $\alpha$.

- The *cumulative* probability of a PARETO distribution[18] with parameters $\alpha$ and $x_m$ is

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

- The *quantile* function (inverse cumulative PARETO -distribution) is[19]

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

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

[18] The CDF function for the Pareto distribution returns the probability that an observation from a Pareto distribution with the shape parameter $\alpha$ and the scale parameter $x_m$, is less than or equal to x.
[19] 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

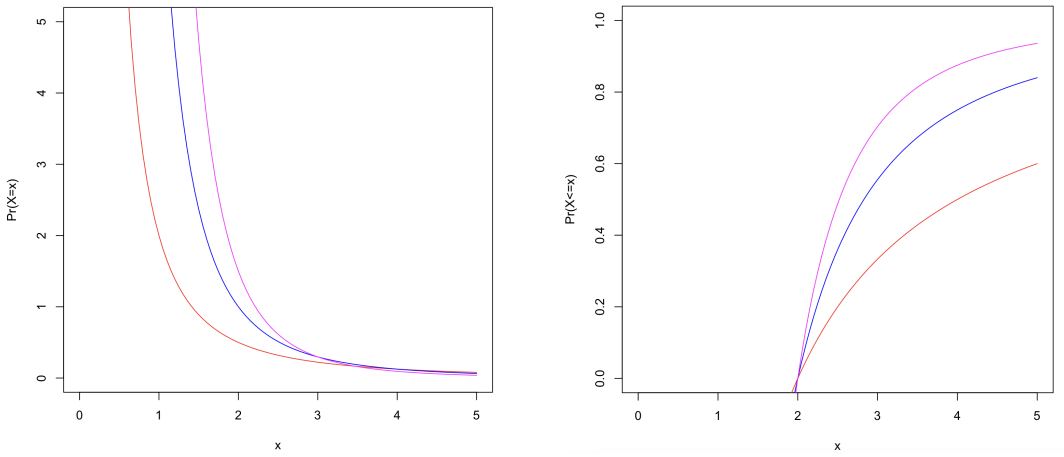



Figure 24: PARETO distributions with different shape parameters $\alpha$. ▷wiki.
**Left** figure: ≀ = plot of `paretoPDF(x,`$\alpha$`)`, $\alpha = 1; 2; 3$.
Check `paretoPDF`$(1, 1) = 1$ on the graph.
**Right** figure: ≀ = plot of `paretoCDF(x,`$\alpha$`)`, $\alpha = 1; 2; 3$.
Check `paretoCDF` $(2, 1) \approx 0.5$ 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.

▷ We adopt the $\mathcal{R}$ convention $(x, \mu, \alpha)$, whereas `pareto`$|_{distrib}$ use $(x, \alpha, \mu)$!

1. *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 $x = 3$. *Check*.

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

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

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

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

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

### General information

○ General mathematical information about the concept is here ▷ WIKI : Pareto
○ Syntax and semantic of the function is here ▷ MATLAB : `gppdf`
○ Online calculator *Bognar's app*

### 3.6.1 Exercises

**Exercise 168.** (a user-defined function paretoPDF in plain MAXIMA)
Write a user-defined function `paretoPDF()` for the `Pareto` distribution using the mathematical definition for `paretoPDF` in 'plain' MAXIMA.
Check the function on example.1.
*Solution*: `Ex.168`

**Exercise 169.** (Table and plot for `paretoPDF`)
a. Calculate 9 values of the `paretoPDF` plot (red') shown in fig.24.left.
b. Plot the `paretoPDF` values from a.
c. MATLAB: Compute the density of the observed value 3 in the Pareto distributions with scale parameter 2 and shape parameters 1 through 5.
*Solution*: `Ex.169.c`

**Exercise 170.** (user-defined function paretoCDF in plain MAXIMA) Write a user-defined function `paretoCDF()` for the cumulative `pareto` distribution using the mathematical definition for `paretoPDF` in 'plain' MAXIMA .
Check your function on example.2.
*Solution*: `Ex.170`

**Exercise 171.** (table and plot of the `paretoCDF`)
a. Calculate 9 values of the `paretoCDF` shown in fig.24.right.
b. Plot the `paretoCDF` values from a.
c. $\mathcal{R}$: calculate the cdf's of a Pareto distribution with parameters location=2 and shape=1, evaluated at 3, 4, and 5.
*Solution*: `Ex.171.c`

**Exercise 172.** (user-defined function paretoINV)
Write a user-defined function `paretoINV()` for the inverse `pareto` distribution using the mathematical definition for `fINV` in 'plain' MAXIMA.
Check your function on example.3.
*Solution*: `Ex.172`

**Remark.**
`paretoINV()` solves the task: find $x$, given $y = p$, shape $m$, scale $a$,
s.t. `paretoINV()` $(x, m, a) = p = y$.
The result $u$ is the value such that an observation from an Pareto distribution with parameters $a, m$ will fall in the range $[0, x]$ with probability $p$.

**Exercise 173.** (Table and special value for `paretoINV` )
a. Print 10 values for the `paretoINV` for $\alpha = 0.1 \ldots 1$ *step* 0.1 and $\mu = 2.5$.
b. MATLAB: Determine the 25'th percentile of a Pareto distribution with parameters location=1 and shape=1.
*Solution*: `Ex.173.b`

## 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**

▷ We adopt the $\mathcal{R}$ convention $(x, \lambda, k)$, whereas $\texttt{weibull}|_{distrib}$ use $(x, k, \lambda)$!

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

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

2. (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? *Check.*

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

3. (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. *Check.*

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

**General information**

- ○ General mathematical information about the concept is here ▷ WIKI : Weibull
- ○ Syntax and semantic of the function is here ▷ MATLAB : `wblpdf`
- ○ Online calculator *Bognar's app*

### Graphical representation

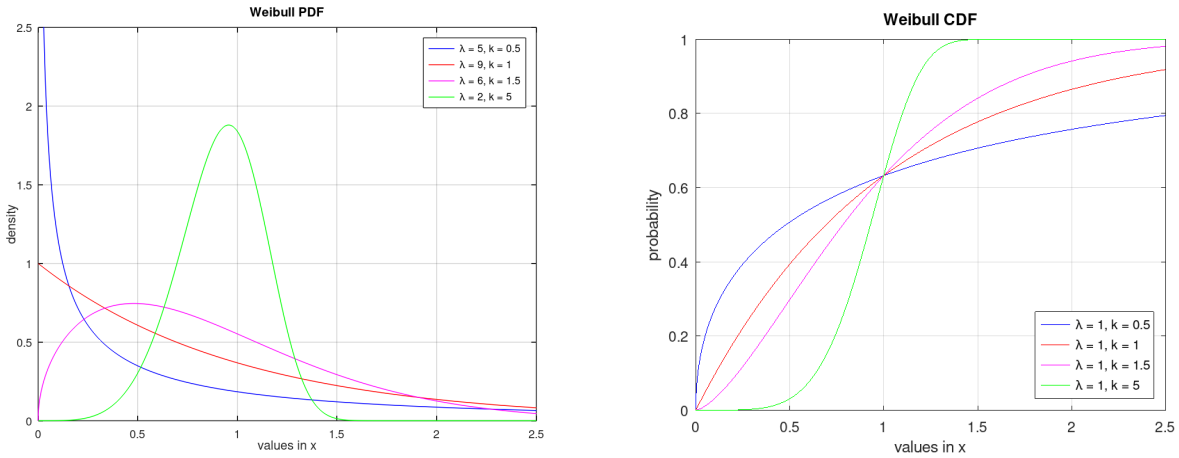



Figure 25: WEIBULL distributions with different parameters $\lambda, k$.
**Left** figure: ≀ = plot of `weibullPDF(x,`$\lambda, k$`)`, $\lambda = 1, k = 0.5; ..; 5$.
Check `weibullPDF`$(1, 1, 5) = 1.8$ on the graph.
**Right** figure: ≀ = plot of `weibullCDF(x,`$\lambda$`,k)`, $\lambda = 1, k = 0.5; ..; 5$.
Check `weibullCDF` $(1, 1, 5) \approx 0.6$ on the graph.
Check `weibullINV`$(0.8, 1, 5) \approx 1.2$ on the CDF via $y = 0.8 \longrightarrow\downarrow 1.2 = x$

### 3.7.1 Exercises

**Exercise 174.** (a user-defined function `weibullPDF` in plain MAXIMA)
Write a user-defined function `weibullPDF()` for the density of the `weibull` distribution using the mathematical definition for `weibullPDF` in 'plain' MAXIMA.
Check the function on example.1.
*Solution*: `Ex.174`

**Exercise 175.** (Table and plot for `weibullPDF`)
a. Calculate the 9 values of the `weibullPDF` plot ('green') shown in fig.25.left.
b. Plot the `weibullPDF` values from a.
*Solution*: `Ex.175`

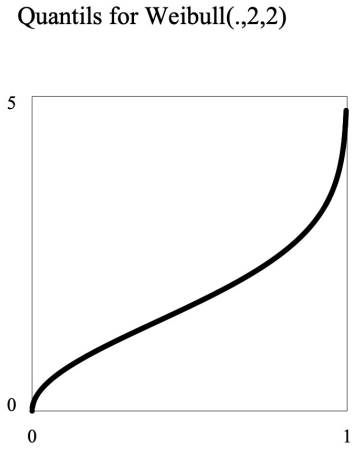
**Exercise 176.** (example from MATLAB, cf. wblpdf)
Compute the density of the observed value 3 in the Weibull distributions with scale parameter 2 and shape parameters 1 through 5.
*Solution*: `Ex.176`

**Exercise 177.** (user-defined function `weibullCDF` in plain MAXIMA) Write a user-defined function `weibullCDF()` for the cumulative `weibullCDF` distribution using the mathematical definition for `weibullCDF` in 'plain' MAXIMA.
Check your function on example.2.
*Solution*: `Ex.177`

**Exercise 178.** (table and plot of the `weibullCDF`)
a. Calculate the 9 values of the `weibullCDF` 'green' shown in fig.25.right.
b. Plot the `weibullCDF` values from a.
*Solution*: `Ex.178`

**Exercise 179.** (user-defined function `weibullINV`)
A user-defined function `weibullINV()` for the inverse `weibullINV` distribution using the mathematical definition for `weibullINV` in 'plain' MAXIMA .
Check your function on example.3.
*Solution*: `Ex.179`

**Exercise 180.** (Table and plot for `weibullINV` )
a. Print 10 values for the `weibullINV` for $\lambda = 0.1 \ldots 1$ *step* 0.1 and $k = 2.5$.
b. Plot the `weibullINV` values from a.
*Solution*: `Ex.180`




**Exercise 181.** (1st optional parameterization used by wiki: weibull1)
Wikipedia use the transformation $b := \lambda^( - k)$, e.g. [6], p.12,64 wiki:weibull.
a. Write the three weibull-distribution functions `weibull1XXX` w.r.t. Wikipedia.
b. Solve the problem:
A particular switch has a Weibull distributed lifetime with $k := \alpha = 0.01$ (1/year) and $\beta = 2$.
What guarantee period can be given for the switch if the survival probability during this time is to be 99%?
*Solution*: `Ex.181`

**Exercise 182.** (2nd optional parameterization: weibull2)
In wiki2 the shape parameter $k$ is the same as in the standard case, while the scale parameter ($\lambda$ is replaced with a rate parameter $\beta = 1/\lambda$.
a. Write the three weibull-distribution functions `weibull2XXX` w.r.t. Wikipedia.
b. Calculate $1 - weibullCDF2(1, 0.01, 2)$
*Solution*: `Ex.182`

**Exercise 183.** (example from Beucher )
Translate the following MATLAB/Octave snippet to MAXIMA, cf. [**?**], warning: $(a, b) = (b, T)$:

```
x=(0:1:10);                % values for x:  0 until 10 step 1
b = 0.444;  T=0.767;       % parameter of Weibull-distribution
a = 1/(T^b);               % adaption of parameters to MATLAB
                           % calculate the Weibull-density
pdfwb = wblpdf(x, a, b)
pdfwb =
 Columns 1 through 6:
   Inf  1.6312e-01  7.8813e-02  4.8763e-02  3.3664e-02  2.4765e-02
 Columns 7 through 11:
   1.9002e-02  1.5028e-02  1.2160e-02  1.0018e-02  8.3746e-03
```

*Solution*: `Ex.183`

# 4 Test Statistics

We now show, how the distributions of chapter 3 are used to implement diverse parameter tests. We translate the mathematical concepts and definitions in MAXIMA code and demonstrate exemplary calls and examples.

**A: Parameter Tests** We do all tests first *semi-automatic* in MAXIMA, 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 plain MAXIMA.

## 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 $\mu_0$. For this, the parameters of the normal distribution must *not* estimated from the sample.

**Mental image**

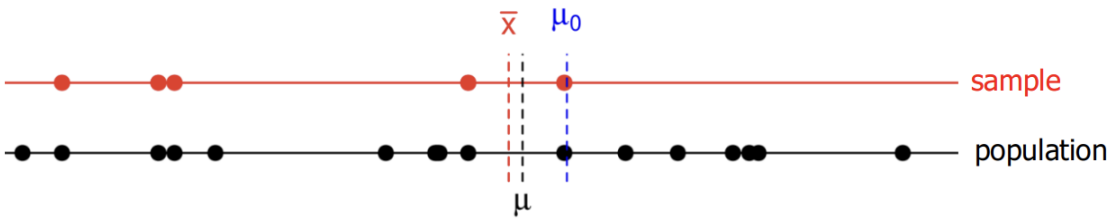



Figure 26: Visualization of $Z$-test for a Hypothesized Mean $\mu_0$:
• : sample $X$ with sample mean $\overline{X}$.
• : population with population mean $\mu$.

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

**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]$
....... mu0    sigma alpha type
ztest(X, 124.6, 14.5,  0.05, 0);
```

| Z | quantil | CI.left | CI.right | p |
|---|---|---|---|---|
| 2.6598 | 1.96 | 127.53 | 143.94 | 0.0078189 |

*Solution*: step-by-step, `Example.4.1.a`
*Solution*: automatic, `Example.4.1.b`

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

### 4.1.1 Exercises

**Exercise 184.** (example at CountBio)
Study this example at ▷ CountBio.

**Exercise 185.** (example at FU Berlin)
Study using MAXIMA the example ▷ FU Berlin script.

**Exercise 186.** (One Sample z-test in $\mathcal{R}$)
Do using MAXIMA the ' *Example of One Sample z-test in R*' ▷ Tutorial.

**Exercise 187.** (One Sample z-test of a sample with given parameters)
Suppose, a sample of $n = 50$ items has $\bar{x} = 105$, $\mu = 100$ and $\sigma = 15$.
Determine the z-test on the significance level $\alpha = 0.05$.
Use only build-in functions from 'distrib'. *Solution*: `Ex.187`

## 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**

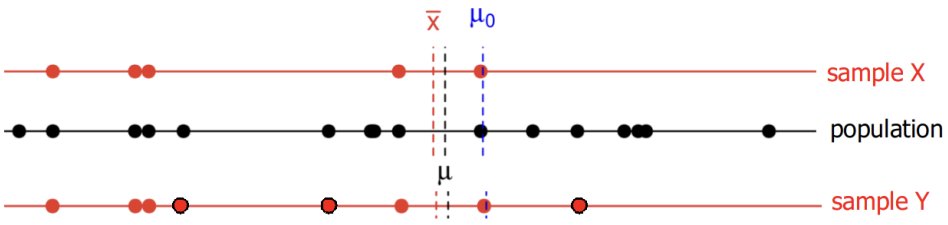



Figure 27: Visualization of $Z$-test for a Hypothesized Mean $\mu_0$:
• : samples $X$ and $Y$ with sample means $\overline{X}$ resp. $\overline{Y}$
• : population with population mean $\mu$.

**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)

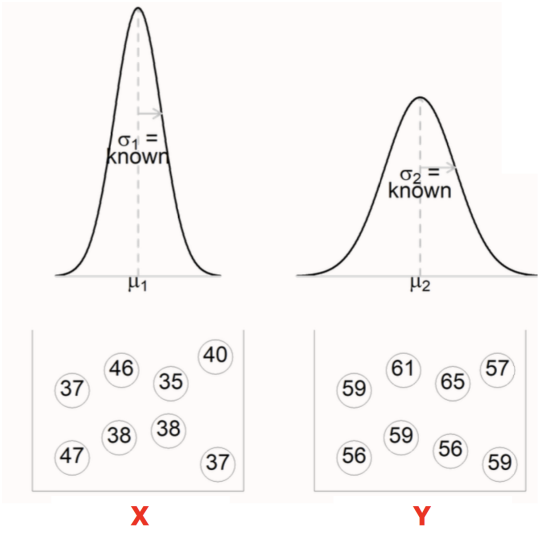
3. **Test statistics** Calculate the normal distributed test value
$$T := \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{m} - \frac{\sigma_2^2}{n}}}.$$

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

**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 MAXIMA 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 $(\mu_1, \sigma_1)$ and $(\mu_2, \sigma_2)$.
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 : [59,56,61,59,65,56,57,50];
/*--   X,Y, muX, muY, sigmaX,  sigmaY,   alpha, altHyp */
ztest2(X,Y, 0,   0,   4.11552, 2.78388, 0.05, 0);
```

| Z | Z.score | CI.left | CI.right | p |
|---|---|---|---|---|
| − 10.318 | 1.96 | −21.568 | −14.682 | 0.0 |

*Solution*: step-by-step, `example.4.2.a`
*Solution*: automatic, `example.4.2.b`

### General information

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

### 4.2.1 Exercises

**Exercise 188.** Study the example at ▷ statkat.
Use MAXIMA to follow and control the argumentation.

**Exercise 189.** (Sperling data" samples from ▷ openeducator.)
We have the "Sperling data" samples from US vs. Sweden:

```
US =(69.12,66.88,74.82,67.00,69.12,65.00,71.00,66.76,72.12,72.94,
    69.18,66.18,64.94,71.76,70.12,71.00,71.88,65.24,70.06,71.94,
    72.12,66.88,73.82,74.00,71.18,67.88,65.94,68.88,68.00,75.12)

Sw  =(74.56,71.89,73.00,67.78,72.22,68.00,73.56,75.00,68.22,69.00,
    68.00,72.00,73.56,72.56,75.00,68.33,71.67,72.44,75.00,71.89,
    72.00,70.00,69.22,74.44,68.00,73.89,70.00,70.44,70.22,73.33)
```

with population means 69.98 vs. 70.43 and stdDevs 3.12 vs. 2.44.
a. Do a step-by-step two-sample-z-test analog example.4.2.1
b. Do the two-sample-z-test for US vs. Sw using the function `ztest2()` from example.4.2.2.

**Exercise 190.** (Two-sampled z-test for two medications ▷ g4g.)
A researcher wants to compare the effectiveness of two different medications for reducing blood pressure. Medication A is tested on 50 patients, resulting in a mean reduction of 15 mmHg with a standard deviation of 3 mmHg. Medication B is tested on 60 patients, resulting in a mean reduction of 13 mmHg with a standard deviation of 4 mmHg.
At a 1% significance level, is there a significant difference between the two medications?
a. Do a step-by-step two-sample-z-test analog example.4.2.1.
b. Do the two-sample-z-test for US vs. Sw using the function ztest2() from example.4.2.2
.

**Exercise 191.** (Example from REAL STATISTICS ▷ real-statistics.)
Here are the data

```
X=(82.67,90.11,89.20,119.15,83.01,93.61,88.42,97.02,126.11,127.96,
89.03,94.51,93.32,89.26,110.36,92.52,112.87,64.05,80.06,74.13,
109.13,81.59,94.99,101.34,104.82,106.92,80.50,106.31,85.46,103.69)

Y=(106.18,100.86,129.85,100.30,87.56,96.87,112.57,148.36,131.62,
114.60,95.47,108.66,83.32,117.64,96.90,66.46,87.80,115.52,102.34,
97.01,117.51,115.64,97.22,131.04,101.58,103.80,111.99,119.34,95.10,
107.62)
```

Do the two-sample-z-test following the text.

**Exercise 192.** (Examples from www.countbio.com, countbio)
Do example 1 and example 2 in www.countbio.com.

## 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 $\mu_0$. In contrast to the one-sample $Z$-test, the standard deviation $\sigma$ of the population is estimated using the sample's standard deviation $s$.

**Mental image**

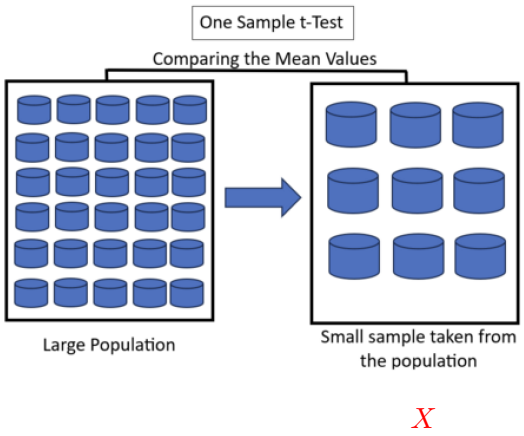



Figure 28: Visualization of one-sample-$t$-test for a hypothesized mean $\mu_0$: ▷ wiki

**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 $\alpha$-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, cf. geo.fu-berlin.*

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 $\mu_0$, 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);
```

["both: reject H0"]

| | |
|---|---|
| "Significance level:" | 0.05 |
| "Degrees of freedom:" | 8 |
| "Test statistic:" | −3.3458 |
| "Critical value:" | −2.306 |

*Solution*: step-by-step, `example.4.3.1`
*Solution*: automatic, `example.4.3.2`

### General information

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

### 4.3.1 Exercises

**Exercise 193.** (p-value and confidence interval)
is to calculate for the example "Students weight in Europe".

**Exercise 194.** (Example from statstutorial, cf. screws)
A random sample of screws have weights

```
29.99, 30.01, 29.97, 30.11, 29.99, 30.02.
```

a. Calculate a 95% confidence interval for the population's mean weight.
Assume the population is distributed as $\mathcal{N}(\mu, \sigma^2)$.
b. Do a step-by-step one-sample-t-test analog example.4.3.1
c. Do the one-sample-t-test using the function `ttest()` from example.4.3.2.
∘ Result: If we sampled many times, our interval would capture the true mean weight 95% of the time; thus, we are 95% confident that the true mean weight of all screws will fall between 29.96 and 30.07.

**Exercise 195.** (example from $\mathcal{R}$) Study this example at its source $\mathcal{R}$vignette

**Exercise 196.** (example from [7, p.81, p.96])
This is a normal $\mathcal{N}(100, 3^2)$-distributed sample of length 100.

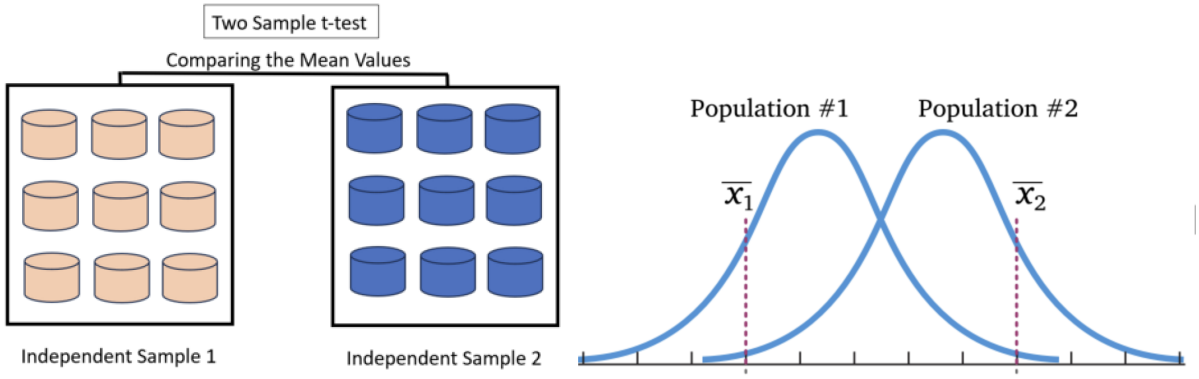
```
X = (96.3, 95.8, 91.7, 91.6, 98.7,      105.0, 99.4, 101.1, 92.9, 97.2,
     110.6, 94.9, 101.2, 92.3, 106.1,   96.5, 107.9, 101.5, 99.0, 97.1,
     90.8, 101.1, 101.0, 100., 106.9,   101.2, 95.7, 92.6, 101.1, 104.9,
     101.1, 104.5, 105.5, 114.1, 91.8,  95.8, 93.5, 98.2, 98.8, 98.5,
     108.1, 99.9, 105.1, 95.6, 98.3,    96.5, 94.9, 101.3, 94.7, 103.7,
     99.0, 98.5, 94.4, 104.7, 93.2,     97.3, 104.3, 101.3, 96.9, 107.0,
     95.0, 101.0, 103.4, 100.5, 100.2,  102.3, 95.1, 95.8, 102.9, 95.0,
     102.7, 101.0, 105.5, 97.9, 104.0,  103.3, 97.7, 88.5, 95.5, 100.5,
     102.3, 101.0, 100.0, 106.0, 102.8, 106.7, 106.6, 99.2, 112.2, 100.1,
     99.3, 100.9, 99.8, 96.0, 97.9,     93.0, 93.7, 97.0, 95.8, 99.9)
```

Use the std.Dev $\sigma = 4.1$ and check the hypothesis $H_o : \mu = 100$.

## 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 29: 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 $\alpha$-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*, cf 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 MAXIMA of this example in

```
A : [88, 92, 94, 78, 88, 95.]$
B : [75, 80, 79, 88, 85, 92.]$
ttest2(A, B, 0.05, 0) ;
```

```
["both: accept H0"]

("Significance level:"      0.05
 "Degrees of freedom:"      9.9977
 "Test statistic:"          1.6606
 "Critical value:"         −2.2282
 "CI left:"                −2.3987
 "CI right:"               14.399
 "mean A:"                 89.167
 "mean B:"                 83.167
 "p-value:"                0.12779)
```

*Solution*: step-by-step, `example.4.4.1`
*Solution*: automatic, `example.4.4.2`

### General information

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

### 4.4.1 Exercises

**Exercise 197.** (OCTAVE $t$-test tutorial)
Reproduce the following tutorial in MAXIMA.
First generate some toy data

```
N = 20;
d = [randn(N,1) randn(N,1)+0.7];      \% Each column is one dataset
```

```
d =
  -0.269943   0.727796
   0.234784   0.656962
   0.286618   0.107063
  -0.246098  -0.796628
  -0.365948   0.844111
  -0.830328  -0.988360
   0.555346   1.202737
  -0.431598   0.282257
   0.025352   1.049475
   1.304216  -0.617381
  -0.511775   0.190412
  -0.184626   0.299169
  -0.335331   0.386240
  -0.149371   0.809719
  -0.038682   1.536842
  -0.281365  -0.632293
  -1.210282   0.250702
  -0.557427   1.395904
  -0.203508   0.083293
   0.050512   0.206742
```

Then verify the following calculation

```
% Paired-samples t-test
octave:6>  [h,p,ci,stats] = ttest(d(:,1), d(:,2));
octave:7> h
  h = 1
octave:9> p
  p = 0.017092
octave:10> ci
  ci = -0.9143
       -0.1011
octave:11> stats
  stats = scalar structure containing the fields:
    tstat = -2.6133
    df = 19
    sd = 0.8688
```

**Exercise 198.** Follow the example at 'MATLAB.help' using MAXIMA : ▷ ttest2/MATLAB

**Exercise 199.** Follow this example using MAXIMA : ▷ theopeneducator

## 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**

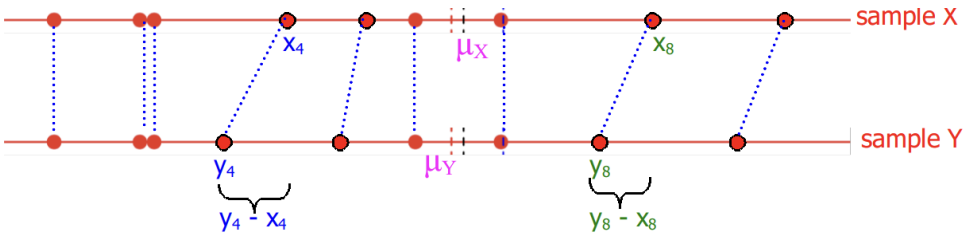



Figure 30: Visualization of paired-$t$-test for 2 dependent samples
above: the sample $X$ with sample mean $\mu_X$.
bottom: the sample $Y$ with sample mean $\mu_Y$.

**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 $\alpha$-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 [7, 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 MAXIMA :

```
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]$
alternative hypothesis: true mean difference is not equal to 0
```

```
..... accept H0

[ T:        1.022   ]
[ p-value:  0.32874 ]
```

*Solution*: step-by-step, `example.4.5.1`

*Solution*: automatic, `example.4.5.2`

### General information

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

### 4.5.1 Exercises

**Exercise 200.** (Lowering cholesterol levels, cf. ttest)
To test a new therapy for lowering cholesterol levels, cholesterol levels are measured in ten subjects before and after treatment. The following measurement results were obtained:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Vor der Behandlung: | 223 | 259 | 248 | 220 | 287 | 191 | 229 | 270 | 245 | 201 |
| Nach der Behandlung: | 220 | 244 | 243 | 211 | 299 | 170 | 210 | 276 | 252 | 189 |
| Differenz: | 3 | 15 | 5 | 9 | –12 | 21 | 19 | –6 | –7 | 12 |

| LEXICON | *German* | English |
|---|---|---|
| | Vor der Behandlung | before treatment |
| | Nach der Behandlung | after treatment |

Do the example 2 ('Beispiel 2') w.r.t. lowering cholesterol levels.

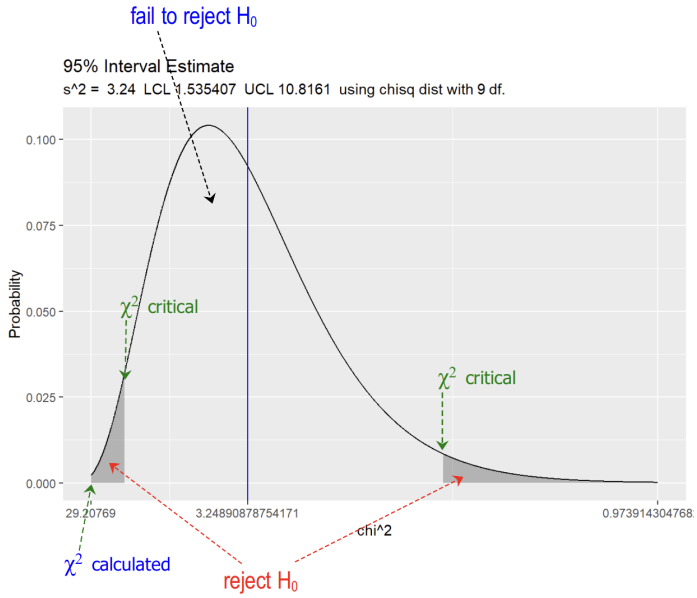
**Exercise 201.** (example from $\mathcal{R}$)
Ceck our example alternatively with $\mathcal{R}$.

```
R> X = c(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 = c(47.8,42.0,33.2,41.0,29.5,28.1,17.4,21.5,41.4,51.3,49.5,53.7)
 > result <- t.test(X, Y, paired = TRUE)
 > print(result)

  Paired t-test

    data:  X and Y
    t = 1.022, df = 11, p-value = 0.3287
    alternative hypothesis: true mean difference is not equal to 0
    95 percent confidence interval:
                     -4.383924 11.983924
    sample estimates:
    mean difference
                3.8
```

## 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 $\chi^2$-variance test checks a normally distributed sample for a given variance $\sigma_0^2$.

### Mental image




Visualization of Chi-Squared Test on Variance, cf. ▷ Foley, enhanced.
Figure 31: 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 $H_0$ $\chi^2$-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, $\sigma_0$ 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 $\alpha$-quantile of the $\chi^2$-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 MAXIMA:

```
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);
```

..... reject H0

| T=chi2 | p | alpha | lCL | uCL |
|---|---|---|---|---|
| 29.208 | 0.0011956 | 0.05 | 1.5354 | 10.816 |

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

*Solution*: step-by-step, `example.4.6.1`
*Solution*: automatic, `example.4.6.2`

**General information**
General mathematical information about the concept is here ▷ WIKIPEDIA : $\chi^2$-test
Syntax and semantic of the implementation is here ▷MATLAB : `vartest`

### 4.6.1 Exercises

**Exercise 202.** (the NIST example, cf. GEAR)
A chi-square test was performed for the GEAR.DAT data set. The observed variance for the 100 measurements of gear diameter is 0.00003969 (the standard deviation is 0.0063). The following are the data used for the chi-square test for the variance example.
We will test the null hypothesis that the true variance is equal to 0.01.

a. Do the example from NIST with the following first 20 items from a. of the data GEAR.DAT. The following data is the gear diameter:

```
1.006  0.996  0.998  1.000  0.992
0.993  1.002  0.999  0.994  1.000
0.998  1.006  1.000  1.002  0.997
0.998  0.996  1.000  1.006  0.988
```

b. Calculate the test statistic value for the sample of 20 items and compare it with the value of 0.3903 for the whole sample.
c. Check, wether the test statistic value is much smaller than the lower critical value, so we also may reject the null hypothesis and conclude that the variance is not equal to 0.01.

**Exercise 203.** (example from $\mathcal{R}$)
Check our solution using the `varTest(..)` function of `R`.

```
R> library(EnvStats)
 > X <- c(12.43, 11.71, 14.41, 11.05, 9.53,  11.66, 9.33, 11.71, 14.35, 13.81)
 > alpha <- 0.05
 > sigma = 1.00
 > result <- varTest(X, "two.sided", 0.95, sigma.squared = 1)
 > print(result)

    Null Hypothesis:                 variance = 1       [=sigma.squared]
    Alternative Hypothesis:          True variance is not equal to 1
    Test Name:                       Chi-Squared Test on Variance
    Estimated Parameter(s):          variance = 3.245299
    Data:                            X
    Test Statistic:                  Chi-Squared = 29.20769
    Test Statistic Parameter:        df = 9
    P-value:                         0.001195555
    95\% Confidence Interval:        LCL =  1.535407
                                     UCL = 10.816103
```

## 4.7 F test

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

**Mental image**

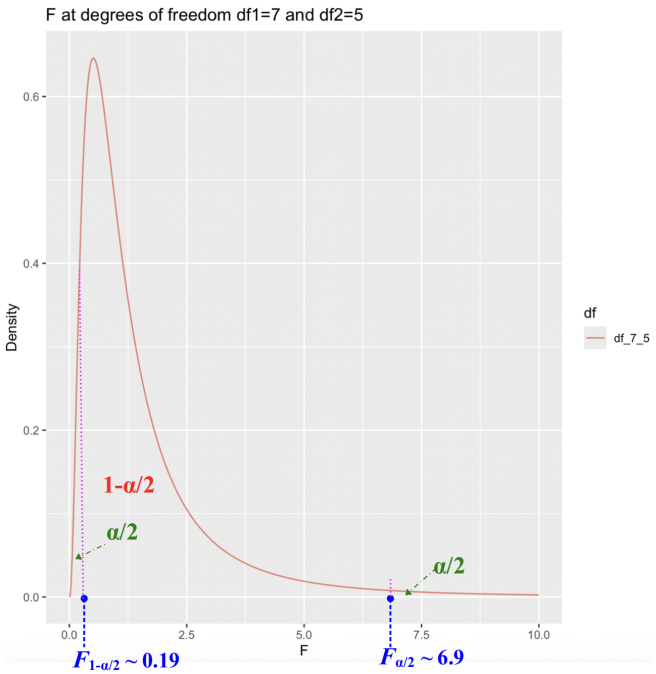



Figure 32: Visualization of $F$-test on equal variances, $F_\alpha$ is the $\alpha$-quantile of $F$. The values shown are from the solution of the example, see below. $1-\alpha/2$ gives the area spanned inside between the limits $|\cdots|$.

**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 $\alpha$-quantile of the $F$-distribution, i.e. $\texttt{fINV}(\alpha, \nu)$ in MAXIMA.

**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 $\alpha$.

**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 MAXIMA:

```
X : [560, 530, 570, 490, 510, 550, 550, 530]; /*-- Background */
Y : [600, 590, 590, 630, 610, 630];           /*- Contaminated */
varTest(X, Y, 0.05);     /*-- F test to compare two variances */
```

```
... accept H0

[   T      p-value    CI.l      CI.h   ]
[ 2.1163   0.42634    0.30882   11.185 ]
```

*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.

*Solution*: step-by-step, `example.4.7.1`
*Solution*: automatic, `example.4.7.2`

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

### 4.7.1 Exercises

**Exercise 204.** (Prices of shares, cf. f-test)
The prices of two shares A and B over eleven months at the beginning of each month are

```
A: 44.32 49.28  65.12 69.44 59.20 70.56
   60.60 87.16 102.60 97.96 93.00

B: 59.35 54.24 59.61 58.68 67.28 60.28
   65.93 73.63 82.84 76.01 78.55
```

Verify: Your risk when investing in shares B of Bernhard is significantly lower than when investing in shares A of Albert.

**Exercise 205.** (groundwater example in $\mathcal{R}$)
Check the following solution using `var.test(..)`[20] function of base $\mathcal{R}$.

```
R> Backg <- c(560, 530, 570, 490, 510, 550, 550, 530)
 > Conta <- c(600, 590, 590, 630, 610, 630)
 > var.test(Backg, Conta, alternative="two.sided")

            F test to compare two variances

  data:  Backg and Conta
  F = 2.1163,  num df = 7,  denom df = 5,  p-value = 0.4263
  alternative hypothesis: true ratio of variances is not equal to 1
  95 percent confidence interval:
            0.3088156   11.1853404
  sample estimates:
  ratio of variances
           2.116337
```

[20] GIMOND, ibid.: "Note that the `var.test()` computes the $F$ ratio using the first variable name in the list as the numerator. For example, had we reversed the order of variables (i.e. var.test(Conta, Backg, alternative="two-sided")), the returned $F$ value would be the inverse of the original $F$ value, or $1/2.12 = 0.47$. The $p$ value would have stayed the same however."

## B: Parameter tests

We quote Hermann[7, 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 Maxima.

We discuss using Maxima the

1. Sign tests
2. Wilcoxon tests
3. Mann–Whitney U test
4. Pearson $\chi^2$ 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**

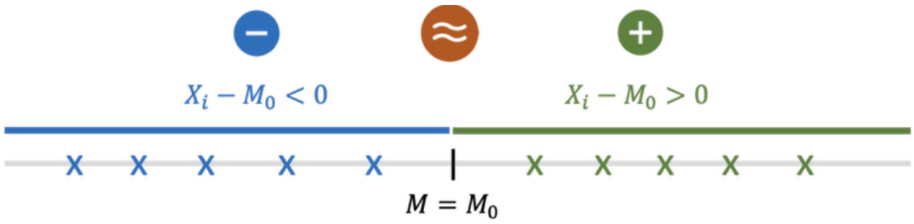



Figure 33: Visualization of $F$-test on equal variances, $F_\alpha$ is the $\alpha$-quantile of $F$. The values shown are from the solution of the example, see below. $1 - \alpha/2$ gives the area spanned inside between the limits $|\cdots|$.

**Procedure** *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 $T > B_{1-\alpha/2}$.
   $B_\alpha$ is the $\alpha$-quantile of the binomial $B$-distribution, i.e. $\texttt{binINV}(\alpha)$ in MAXIMA.

**Example** *Average age of men at first marriage.*

HERRMANN [7, 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 MAXIMA :

```
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") ;
```

| s | p |
|---|---|
| 5 | 0.77441 |

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.

*Solution*: step-by-step, `example.4.8.1`
*Solution*: automatic, `example.4.8.2`

### General information

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

### 4.8.1 Exercises

**Exercise 206.** (Remark and surplus)
We can do the decision about the $H0$ in the 'age' example using alpha-quantiles of the binomial distribution.
a. Look up our function `binoINV()` or the 'distrib' version `quantile_binomial(q,n,p)`.
b. Then do:

```
...
   alpha : 0.05 $
   nt  : 12      /*-- n total */$
   T   : 5       /*-- T =s test statistic */$
   qt1 :binoINV(alpha/2, nt, 0.5)         /*-- alpha quantile */ ;
   qt2 : binoINV(1-alpha/2, nt, 0.5)     /* -- alpha quantile */;
   /*   5   3       5   9     */
   if  (T < qt1) or (T > qt2)  then print("reject H0") else print("accept H0");
```

**Exercise 207.** (Check example 'Average age of men..' with OCTAVE)
We use OCTAVE function `signtest()`.

```
octave:1> pkg load statistics
octave:2> 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];
octave:3> [p,h,stats] = signtest(X,26)
          ans =
                p     = 0.7744
                h     = 0
                stats =
                        zval = NaN
                        sign = 5
                        (wL: i.e. 5 data elements have a value > median)
```

**Exercise 208.** (example by SPRENT from WIKIPEDIA )
Do the cited example of a sign test for the median of a single sample in MAXIMA.

**Exercise 209.** (example by $\mathcal{R}$)
Reproduce the following example from the $\mathcal{R}$documentation in MAXIMA:

```
R> x <- c(7.8, 6.6, 6.5, 7.4, 7.3, 7., 6.4, 7.1, 6.7, 7.6, 6.8)
R> SIGN.test(x, md = 6.5)
%  Verify:
   s = 9, p-value = 0.02148
   alternative hypothesis: true median is not equal to 6.5
   95 percent confidence interval:
      6.571273 7.457455
   sample estimates: median of x
                    7
```

## 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**

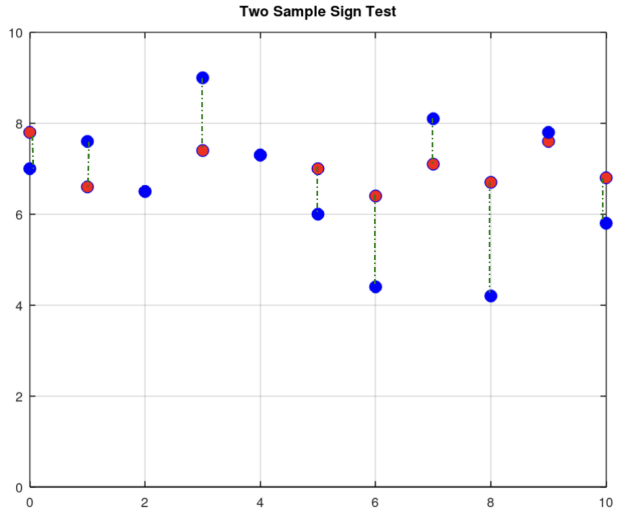



Figure 34: Visualization of *sign*2-test on two samples $X$ and $Y$ with ⋮ $= x_i - y_i$.
●: sample $X = (7.8, 6.6, 6.5, 7.4, 7.3, 7.0, 6.4, 7.1, 6.7, 7.6, 6.8)$.
●: sample $Y = (7.0, 7.6, 6.5, 9.0, 7.3, 6.0, 4.4, 8.1, 4.2, 7.8, 5.8)$.

**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 $T > B_{1-\alpha/2}$.
   $B_\alpha$ is the $\alpha$-quantile of the binomial $B$-distribution, i.e. $\texttt{binINV}(\alpha)$ in MAXIMA.

**Example** *IQ of twin pairs.*

HERRMANN [7, 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.
$IQ_\ell$ 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 |
| $IQ_\ell$: | 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 MAXIMA :

```
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");
```

$$\begin{bmatrix} s & p \\ 9 & 0.146 \end{bmatrix}$$

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.

*Solution*: step-by-step, `example.4.9.1`
*Solution*: automatic, `example.4.9.2`

### General information

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

### 4.9.1 Exercises

**Exercise 210.** (Fig. 34) Do the sign test for the data of Fig.34, i.e.

```
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)
```

a. Check the differences on the plot.
b. How to determine the test statistic T on the graph?

**Exercise 211.** (Example from MatLab, cf. signtest)
Do the example from the MATLAB documentation in Maximausing the data from section 'Medians of Paired Samples'.
a. Test the hypothesis of zero median for the difference between the two paired samples using the exact methods.
b. z-Statistic: For a large sample, signtest uses the z-statistic to approximate the p-value. Program the z-statistic in Maxima.
c. Check the z-statistic for the data of a.

**Exercise 212.** (example from Octave)
Check the following solution of the example 4.9 using `signtest` function of `statistics` package of Octave.

```
octave:1> pkg load statistics
octave:2> IQh = [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];
octave:3> [p,h,stats] = signtest(IQh, IQl)
         ans =
                p     = 0.1460
                h     = 0
                stats = scalar structure containing the fields:
                          zval = NaN
                          sign = 9
                          (wL: i.e. 9 data pairs have a difference greater 0)
```

## 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**

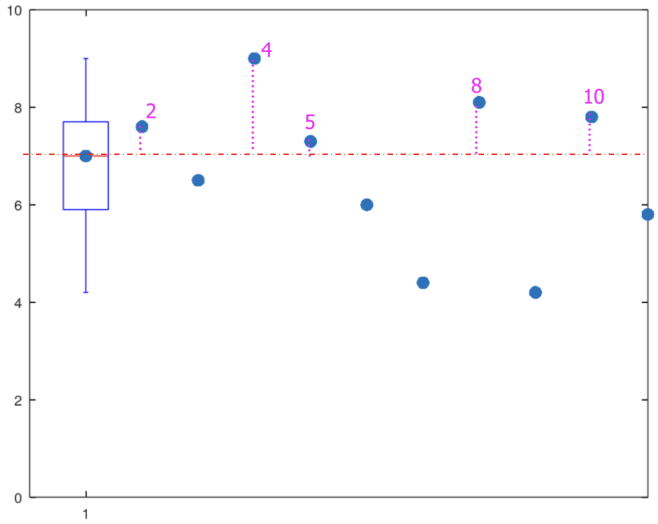



Figure 35: Visualization of WILCOXON-One-Sample test on the sample $X$.
•: sample $X = (7.0, 7.6, 6.5, 9.0, 7.3, 6.0, 4.4, 8.1, 4.2, 7.8, 5.8)$.
$i$: items with index $i = 2, 4, 5, 8, 10$ have positive distance to the median.

**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 $W \leq w_{1-\alpha/2}$.
$w_\alpha$ 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 $\text{No}_i$, 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 MAXIMA

```
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);
```

| W | T |
|---|---|
| 40.0 | 1.2741 |

*Solution*: step-by-step, `example.4.10.1`
*Solution*: automatic, `example.4.10.2`

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

### 4.10.1 Exercises

**Exercise 213.** (pygmy sunfish, example from psu.edu)
Check the solution of example.4.10 using `wilcox.test()` function of f `R`.

```
R>  X=c(5.0, 3.9, 5.2, 5.5, 2.8, 6.1, 6.4, 2.6, 1.7, 4.3)
R> >    wilcox.test(X, mu = 3.7)

        Wilcoxon signed rank exact test

  data:  X
  V = 40, p-value = 0.2324
  alternative hypothesis: true location is not equal to 3.7
```

Interpretation: Based on the results of the test, (at the significance level of 0.05) we reject the null hypothesis.

**Exercise 214.** (Example at Fig.35)
Calculate the WILCOXON test statistic $W$ for the sample $X$ with hypothesized median $x_0 = 7$, where

$$X = (7.0, 7.6, 6.5, 9.0, 7.3, 6.0, 4.4, 8.1, 4.2, 7.8, 5.8)$$

from figure.35.

**Exercise 215.** (placebo-controlled clinical trial, cf. nbisweden)
Let's imagine we are a part of team analyzing results of a placebo-controlled clinical trial to test the effectiveness of a sleeping drug. We have collected data on 10 patients when they took – a sleeping drug and when they took a placebo.
The hours of sleep recorded for each study participant:

```
X=(6.1, 6.0, 8.2, 7.6, 6.5, 5.4, 6.9, 6.7, 7.4, 5.8)     -- drug
Y=(5.2, 7.9, 3.9, 4.7, 5.3, 7.4, 4.2, 6.1, 3.8, 7.3)     -- placebo
```

Before we investigate the effect of drug, a senior statistician ask us: 'Is the median sleeping time without taking the drug significantly less than the recommended 7 h of sleep?'
For the *solution* we therefore consider only $Y$.

## 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**

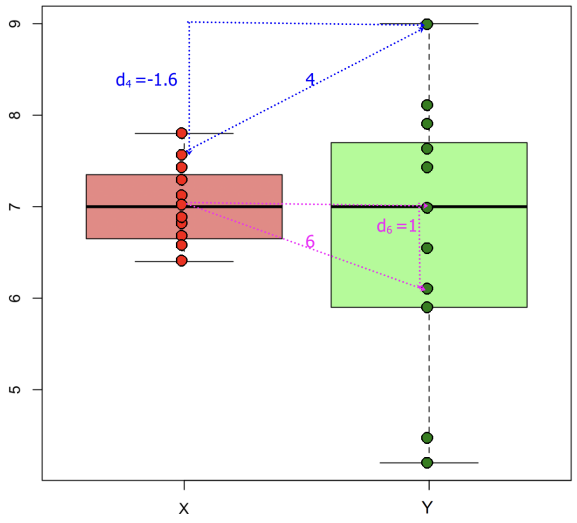



Figure 36: Visualization of Two Sample Wilcoxon–test on two samples $X$ and $Y$
●: sample $X = (7.8, 6.6, 6.5, 7.4, 7.3, 7.0, 6.4, 7.1, 6.7, 7.6, 6.8)$.
●: sample $Y = (7.0, 7.6, 6.5, 9.0, 7.3, 6.0, 4.4, 8.1, 4.2, 7.8, 5.8)$.
●: For pair 6 we have $d_6 = x_6 - y_6 = 7 - 6 = +1$.

**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 $W \leq w_{1-\alpha/2}$.
$w_\alpha$ 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: $T^+ = 4+6.5+10+11+6.5 = 38$[21] and analog $T^- = 6.5+9+6.5+3 = 25$.
2. test statistic $T := min(T^+, T^-) = min(38, 25) = 25$.[22]
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 MAXIMA

```
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) ;
```

| Wp | Wm | T |
|---|---|---|
| 38 | 25 | 25 |

*Solution*: step-by-step, `example.4.11.1`
*Solution*: automatic, `example.4.11.2`

**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)`

[21] 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` $= 28$.

[22] 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.

### 4.11.1 Exercises

**Exercise 216.** (sleep study, cf.psu.edu)
Going back to our sleep study, now we are ready to examine whether there is enough evidence to reject a null hypothesis of median of the differences between the paired observations is equal to 0.
a. Check the solution using `wilcox.test()` function of f $\mathcal{R}$.

```
############## check using R - our W is R's V ;)
R> X=c(6.1, 6.0, 8.2, 7.6, 6.5, 5.4, 6.9, 6.7, 7.4, 5.8)
R> Y=c(5.2, 7.9, 3.9, 4.7, 5.3, 7.4, 4.2, 6.1, 3.8, 7.3)
R>
R> wilcox.test(x = Y, y = X, alternative = "two.sided", mu = 0,
      paired = TRUE, exact = F)

  Wilcoxon signed rank test with continuity correction
  data:  Y and X
  V = 15, p-value = 0.2213
  alternative hypothesis: true location shift is not equal to 0
```

b. Do the test using MAXIMA.

**Exercise 217.** (Example at Fig.36 using OCTAVE)
Check our solution using `signrank` function of MATLAB/OCTAVE.

```
octave:1> pkg load statistics
octave:2> X = [7.8, 6.6, 6.5, 7.4, 7.3, 7.0, 6.4, 7.1, 6.7, 7.6, 6.8];
octave:3> Y = [7.0, 7.6, 6.5, 9.0, 7.3, 6.0, 4.4, 8.1, 4.2, 7.8, 5.8];
octave:4> [p,h,stats] = signrank(X,Y)
           p = 0.5625
           h = 0
           stats = [...]
                   signedrank = 28      zval = NaN
```

## 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**

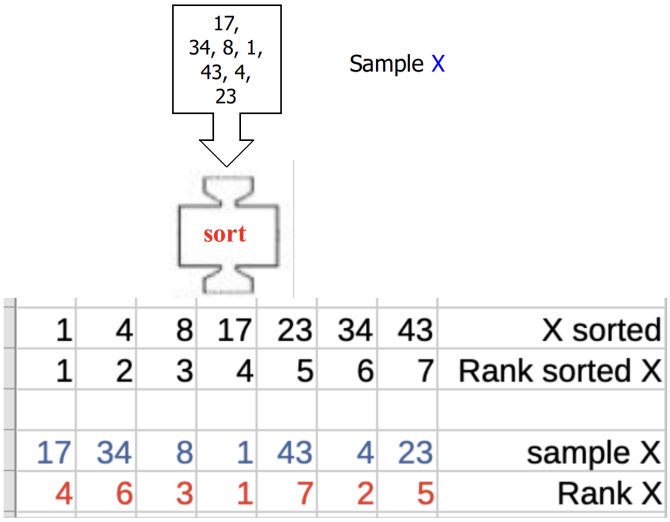



Figure 37: Visualization of $U$-test process on sample $X : 17, 34, 8, 1, 43, 4, 23$.
1: $X = (17, 34, 8, 1, 43, 4, 23) \xrightarrow{sort} (1, 4, 8, 17, 23, 34, 43) = X_{sorted}$.
2: $X_{sorted} = (1, 4, 8, 17, 23, 34, 43) \xrightarrow{Rank} (1, 2, 3, 4, 5, 6, 7) = X_{ranked}$.

**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 $U > w_{1-\alpha/2}$.
$w_\alpha$ is the critical value of the tabulated Wilcoxon-distribution, ▷ Wilcoxon : U-test

**Example** *Butterflies on sunny vs cloudy days.*

Zoefel [22, 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 Maxima

```
X : [6, 15, 35, 35, 62, 73, 98, 112]$
Y : [1, 4, 8, 17, 23, 34, 43]§
ranksum(X,Y)$
```

| U | p | Ho |
|---|---|---|
| 11.0 | 0.97543 | reject H0 |

*Solution*: step-by-step, `example.4.12.1`
*Solution*: automatic, `example.4.12.2`

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

### 4.12.1 Exercises

**Exercise 218.** (Redo the Example..4.12 using OCTAVE)
∘ Check the solution using `ranksum` function of `statistics` package of `Octave`.

```
octave:1> pkg load statistics
octave:2> x=[6, 15, 35, 35, 62, 73, 98, 112];
octave:3>             y=[1, 4, 8, 17, 23, 34, 43];
octave:4>            [p, h, stats] = ranksum(x, y)
            ans : p = 0.050039      h = 0         stats = ranksum = 81
```

**Exercise 219.** (Redo the Example..4.12 using $\mathcal{R}$)
∘ Check the solution using `wilcox.test` function of $\mathcal{R}$.

```
R> X=c(6, 15, 35, 35, 62, 73, 98, 112)
R> Y=c(1, 4, 8, 17, 23, 34, 43)
R> wilcox.test(x = Y,y = X, alternative="two.sided",
                            paired=FALSE, exact = F)

Wilcoxon rank sum test with continuity correction
  data:  Y and X
  W = 11, p-value = 0.05598
  alternative hypothesis: true location shift is not equal to 0
```

**Exercise 220.** (two sorts of alloy, cf. SPIEGEL [17, p. 447, P454])
See the example in Spiegel w.r.t. two sorts of alloy

```
A1 = (18.3, 16.4, 22.7, 17.8, 18.9, 25.3, 16.1, 24.2)
A2 = (12.6, 14.1, 20.5, 10.7, 15.9, 19.6, 12.9, 15.2, 11.8, 14.7)
```

Do a $U$-test to check if $A1$ and $A2$ come from the same population.
Result: reject Ho at the 0.05 level.

## 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**

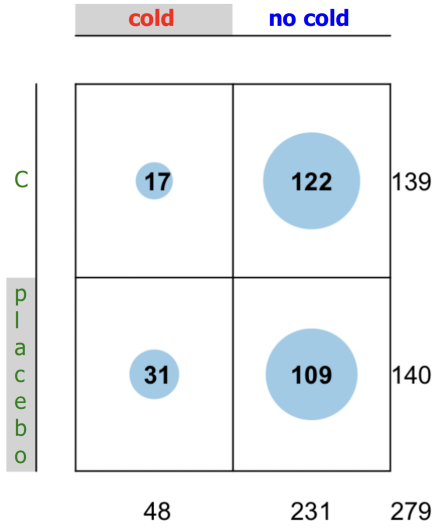



Figure 38: Visualization of Four-field Table for am experimental outcome.
●: experiment outcome as frequency table $T = \binom{a\,b}{c\,d} = \binom{17\,122}{31\,109}$.

**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 [7, 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 MAXIMA :

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

| Chi | p | phi: |
|---|---|---|
| 4.1407 | 0.041864 | 0.12182 |

*Solution*: step-by-step, `example.4.13.1`
*Solution*: automatic, `example.4.13.2`

### General information

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

### 4.13.1 Exercises

**Exercise 221.** (red-green color blindness, cf. [7, p.161])
In a red-green color blindness study, the following four-field table was obtained.

```
                    male   female
normal             :[8324    9031]
red-green-blind    :[ 725      40]
```

Check via a chi-squared test statistic for the dependence of blindness on gender.
DECISION: The $\chi^2$ test statistic is 640.3.
The associated p-value p of the $\chi^2$ distribution is 1.
With a 5color blindness of the gender can just be rejected.

**Exercise 222.** (Example from MatLab, cf. signtest)
Do the example from the MATLAB documentation in Maximausing the data from section 'Medians of Paired Samples'.
a. Test the hypothesis of zero median for the difference between the two paired samples using the exact methods.
b. z-Statistic: For a large sample, signtest uses the z-statistic to approximate the p-value. Program the z-statistic in Maxima.
c. Check the z-statistic for the data of a.

**Exercise 223.** (check with $\mathcal{R}$)
Check the solution of the example 4.13 using `chisq.test` function of $\mathcal{R}$.

```
   # Step 1: Creating a contingency table
R> data <- matrix(c(17, 122, 31, 109), nrow = 2)
   # Step 2: Applying the chi-square test function
   > chisq.test(data)

        Pearson's Chi-squared test with Yates' continuity correction
  data:  data
  X-squared = 4.1407, df = 1, p-value = 0.04186
```

or if one only wants the p- value

```
######## Check with R
R> x <- matrix(c(17, 122, 31, 109), ncol = 2)
 > chisq.test(x)$p.value
   [1] 0.04186438
```

## 4.14 FISHER test

The FISHER test is used for $2 \times 2$ 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 39: 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 [7, 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 is 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 MAXIMA :

```
fishertest(1,4, 14,12, 0.05);
```

$$\begin{bmatrix} \text{Ho} & \text{p} \\ 1 & 0.18638 \end{bmatrix}$$

*Solution*: step-by-step, `example.4.14.1`
*Solution*: automatic, `example.4.14.2`

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

### 4.14.1 Exercises

**Exercise 224.** (vitamin C, cf. [7, p.154])
Redo the example 4.13 with the Fisher test.

```
             Cold notCold
Vitamin C:[17    122] a+b 139
Placebo  :[31    109] c+d 140
           48    231
          a+c    b+d
```

fishertest(17,122, 31,109, 0.05);

**Exercise 225.** (woman exampe, cf. the exact test in woman)
Do the 'woman' example in wikipedia.

**Exercise 226.** (students example, cf. students)
Do the 'students' example in WikiPediA.

**Exercise 227.** (cancer example in cancer)
Do the 'cancer' example in WikiPediA.

**Exercise 228.** (two formulas)
Calculate $P$ for the following four-fields-table.

```
                 Men   Women  Row Total
Studying          0    10      : 10       a=0  b=10
Non-studying     12     2      : 14       c=12 d=2
                 ................
Column Total     12    12      : 24
```

**Exercise 229.** (check with $\mathcal{R}$)
Check the solution of the example 4.14 using `fisher.test` function of $\mathcal{R}$.

```
R> fisher.test(rbind(c(1,4),c(14,12)), alternative="less")$p.value
  [1] 0.1863799

R> fisher.test(rbind(c(1,4),c(14,12)))$p.value
          # default alternative="both.sided"
  [1] 0.3325918
```

## 4.15 McNemar test

McNemar's test is a non-parametric test used to analyze paired nominal data. It is a test on a $2 \times 2$ 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**

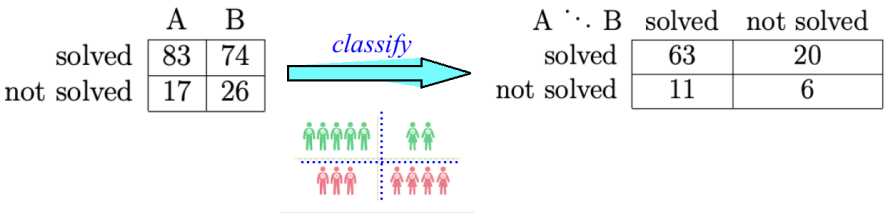



Figure 40: Visualization of McNemar test: first split the outcome 'solved vs. not solved' w.r.t. the groups $A$ an $B$.
Left: experiment outcome as frequency table $\left(\begin{smallmatrix} a' & b' \\ c' & d' \end{smallmatrix}\right) = \left(\begin{smallmatrix} 83 & 74 \\ 17 & 26 \end{smallmatrix}\right)$.
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

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

   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 MAXIMA :

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

| chi | p | Ho |
|---|---|---|
| 1.8846 | 0.16981 | 1 |

**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'.

*Solution*: step-by-step, `example.4.15.1`
*Solution*: automatic, `example.4.15.2`

### General information

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

### 4.15.1 Exercises

**Exercise 230.** ('disease' example in wikipedia)
Do the 'disease' example in wikipedia disease

**Exercise 231.** (students example in wiki)
Do the example in wikipedia/GE studs

**Exercise 232.** (Presidents performance)
Redo the AGRESTI example in R-manual with MAXIMA : agresti

**Exercise 233.** (EXCEL example in real-statistics)
Do the example in real-statistics/EXCEL with MAXIMA: real.stats

**Exercise 234.** (MCNEMAR test for Fig.40)
Do a MCNEMAR test for the values of fig.40.

**Remark.** 1. See the implementation of mcnemarTest in McNemar.
2. The smaller the value of $p$, the greater the evidence for rejecting the null hypothesis.
3. ibid.: Python code excerpt

```
chi2_statistic = (abs(n_min - n_max) - corr) ** 2 /  (n_min + n_max)
pvalue = chi2.sf(chi2_statistic, 1)
```

(!) This pvalue is not in coincidence with R, see below.

**Exercise 235.** (check example.4.15 with $\mathcal{R}$)
Check the solution of the example 4.15 using `mcnemar.test` function of $\mathcal{R}$.

```
R> Matrix = as.matrix(read.table(header=TRUE, row.names=1,
                      text=" Coffee Yes No  Yes 37 17 No 9 25"))
R> mcnemar.test(Matrix)
   McNemar's Chi-squared test with continuity correction
   McNemar's chi-squared = 1.8846, df = 1, p-value = 0.1698
```

# 5 Correlation and Bootstrap

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

## 5.1 PEARSON's $\rho$ Correlation coefficient

The correlation coefficient $\rho$ 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 $\rho$ is defined as $\frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y}$.

**Mental image**

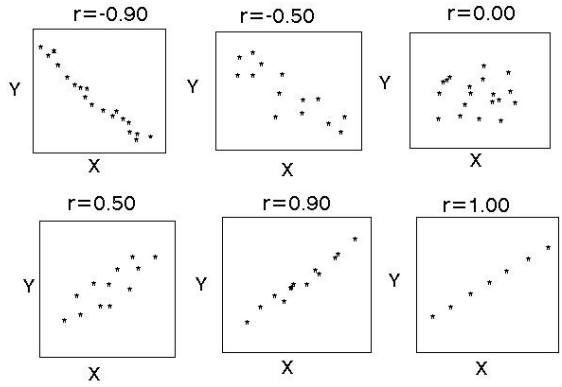



Figure 41: Visualization of PEARSON's correlation coefficient $\rho$: 6 typical cases. *row1:* from strong negative correlation ('-0.9') until no correlation ('0'). *row2:* from mean positive correlation ('+0.5') until strong positive correlation ('+1').

**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** $\rho$ coefficient is given by

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

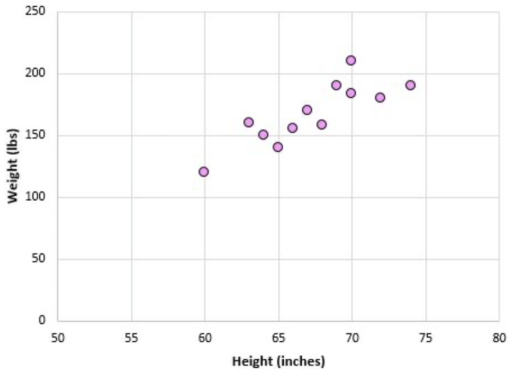
Remark: for $n < 30$ the OLKIN–PRATT correction $\rho := \rho \cdot (1 + \frac{(1-\rho)^2}{2n-6})$ is used.

**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 $\rho$-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 Maxima

```
| 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
```

*Solution*: step-by-step, `example.5.1.1`
*Solution*: automatic, `example.5.1.2`

### General Information

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

### 5.1.1 Exercises

**Exercise 236.** (daily calorie intake, cf. [22, p.125])
Use data from Zoefel:

```
X=(3113, 3216, 3495, 2021, 1916, 2375, 2751, 2288, 1932, 2166, 2217,
   3482, 2036, 2581, 2639)
Y=(68, 74,73,53,59,60,57,47, 46, 50,55, 74,41,71, 67)
```

In a study of 15 randomly selected countries, using the variables 'daily calorie intake' $X$ and 'life expectancy' $Y$ of men, is their a positive correlation between calorie intake and life expectancy?
*Hint*: the test $t$ statistics is

```
n = length(X)
r = corrcoef(X,Y)
t = abs(r) * sqrt(n-2)/sqrt(1 - r^2)
```

**Exercise 237.** (another check)
Verify that the corrcoef of $X = (1, 3, 4, 6, 8, 9, 11, 14)$ and $Y = (1, 2, 4, 4, 5, 7, 8, 9)$ is 0.977.

**Exercise 238.** (Check SPIEGEL example 14.12, cf. [17, p.361])
We have $X = (20, 5, 8, 10, 13, 7, 13, 5, 25, 14)$
and $Y = (2.35, 3.8, 3.5, 2.75, 3.25, 3.4, 2.9, 3.5, 2.25, 2.75)$
with $r = 0.9097$.

**Exercise 239.** (example 4.15 with $\mathcal{R}$ and `Octave`)
◦ Check the solution of the example 4.15 using `cor` function of $\mathcal{R}$.

```
R> X = c(60,65,72,70,74,63,66,68,67,69,70,64)
 > Y = c(120,140,180,184,190,160,155,158,170,190,210,150)
 > cor(X, Y, method = 'pearson')
   [1] 0.8359452
```

◦ Check the solution using `corrcoef()` function of `Octave`:

```
octave:1> X = [60,65,72,70,74,63,66,68,67,69,70,64];
octave:2> Y = [120,140,180,184,190,160,155,158,170,190,210,150];
octave:3> corrcoef(X,Y)
          ans = 1.0000   0.8359
                0.8359   1.0000
```

*Remark*: (MatLAB) For two input arguments, result is a 2-by-2 matrix with ones along the diagonal and the correlation coefficients along the off-diagonal.

## 5.2 SPEARMAN's $\rho_S$ rank correlation coefficient

SPEARMAN 's rank correlation coefficient $\rho_S$ is a number ranging from $-1$ to 1 that indicates how strongly two *sets of ranks*(!) are correlated. The difference between the PEARSON correlation $\rho$ and the SPEARMAN $\rho_S$ 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**

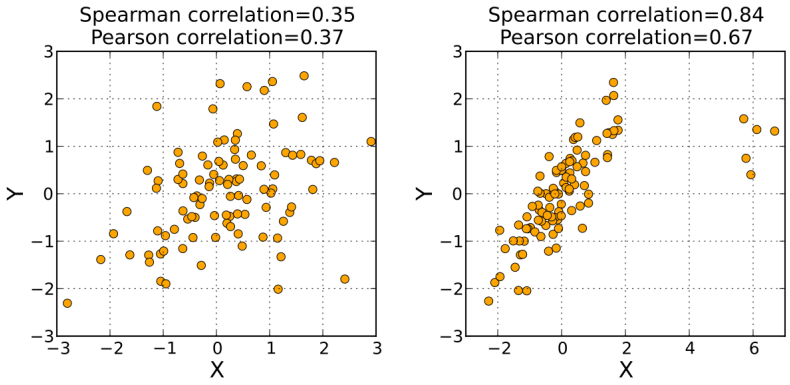



Figure 42: Visualization of SPEARMAN 's rank coeff. $\rho_S$ of two samples $X$ and $Y$
Left: For roughly elliptically distributed data with no prominent outliers, Spearman's $\rho_S$ and Pearson's $\rho$ correlation give similar values.
Right: The Spearman correlation is less sensitive than Pearson's to strong outliers, because $\rho_S$ 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** $\rho_S$ 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 $\rho_S$ 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 MAXIMA 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
```

*Solution*: step-by-step, `example: 5.2.1 - father vs son`
*Solution*: automatic I, `example.5.2.2`
*Solution*: automatic II, `example.5.2.3`
*Solution*: automatic III using cov, `example.5.2.4`

### General Information

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

### 5.2.1 Exercises

**Exercise 240.** (redo example.5.2 using OCTAVE)
Check the following solution using `spearman` function of `statistics` package of `Octave`.

```
octave:1> pkg load statistics
octave:2> X = [70,78,90,87,84,86,91,74,83,85];
octave:3> Y = [90,94,79,86,84,83,88,92,76,75];
octave:4> spearman(X, Y)
               ans = -0.4182
```

**Exercise 241.** (check the Father–Son–example using $\mathcal{R}$)
Check the solution of the example using $\mathcal{R}$.

```
R> Far = c(65,63,67,64,68,62,70,66,68,67,69,71)
R> Son = c(68,66,68,65,69,66,68,65,71,67,68,70)
R> cor.test(Far, Son, method = "spearman")

   Spearman's rank correlation rho
   data:  Far and Son
   S = 74.285, p-value = 0.005903
    alternative hypothesis: true rho is not equal to 0
    sample estimates:
    rho = 0.7402623
```

**Exercise 242.** (significance)
Use the FACT

> The test for significance of SPEARMAN's $\rho_S$ use the statistics
>
> $$T = \mid \rho_S \mid \cdot \frac{\sqrt{n-2}}{\sqrt{1-\rho_S^2}}$$
>
> which is $t$-distributed with $df = n - 2$.

and do an $t$ test for the significance of the $\rho_S$.

Result: T = 2.905

## 5.3 KENDALL's $\tau$ rank correlation coefficient

KENDALL's $\tau$ 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**

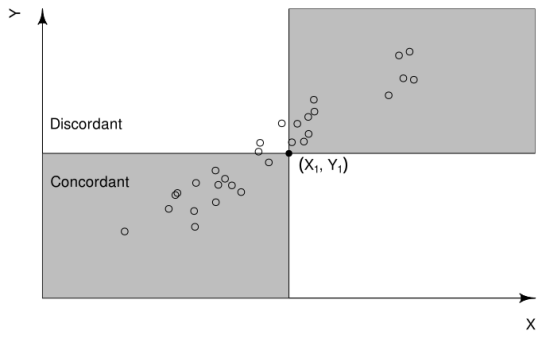



Figure 43: Visualization of KENDALL 's rank corrcoeff $\tau$ of two samples $X$ and $Y$. All points in the gray area are *concordant* and all points in the white area are *discordant* with respect to point $(X_1, Y_1)$. With $n = 30$ points, there are a total of $\binom{30}{2} = 435$ possible point pairs. 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 ..

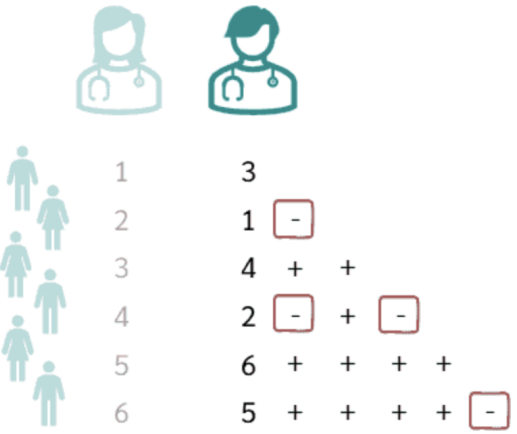
**Procedure** KENDALL*'s rank correlation coefficient* $\tau$

1. **Assumptions** two random samples $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$
2. **correlation** $\tau$ 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 $\tau$ 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 $\tau$ iusing the formula from above.

*Solution* along the columns of the 'rank' table




in MAXIMA:

| 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. $\tau = \frac{C-D}{C+D} = \frac{7}{15} = 0.4666$.
   Positive correlation: As $doc1$ variable increases, the $doc2$ tends also to increase.

∘ *Solution* and explanation of this example with MAXIMA

```
X : [1,2,3,4,5,6]$
Y : [3,1,4,2,6,5]$
kendall(X,Y)$
```

| C | D | tau |
|---|---|---|
| 4 | 11 | 0.46667 |

*Solution*: step-by-step, `example.5.3.1`
*Solution*: automatic, `example.5.3.2`

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

### 5.3.1 Exercises

**Exercise 243.** (example.5.3.1 using $\mathcal{R}$)
◦ Check the following solution using `cor(..,"kendall")` function of `R`.

```
R> X=c(1,2,3,4,5,6)
R> Y=c(3,1,4,2,6,5)
R> cor(X,Y, method="kendall")
   [1] 0.4666667
```

**Exercise 244.** (league standings by [22, p.129ff])
The following table compares the league standings $X$ of the clubs with their stadium sizes $Y$. The aim is to check whether stadium size correlates with league position.

```
X = (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18)
Y = (30,72,65,80,72,25,41,45,59,15,70,23,58,24,64,22,26,28) -- in 10^3
```

[We cite ZOEFEL, a.a.o.] According to KENDALL, the first step is to write down the ranking of the first variable $X$ in ascending order and assign the other accordingly. In this context, one speaks of an anchor row $X$ and a comparison row $Y$. In the example, the league position already noted in ascending order is the anchor row. The comparison row $Y$ consists of the outermost positions in terms of stadium size. It is that the largest stadium is assigned to rank no 1.

With a positive correlation, the comparison series is expected to rise in a similarly monotonic manner as the anchor series. The number of disturbances ('inversion' w.r.t to this monotonicity) represents a progression for the strength of the correlation, since such a disturbance occurs when a lower ranking follows a higher ranking in the comparison series. E.g. this happens ten times for the club in first place and three times for the club in third place.

The number of inversions is entered in the column labeled $I$, the number of proversions is entered in the column labeled $P$, meaning that in the comparison series, a single-place entry follows – the higher one.

Now the sums of I and P are calculated and this gives the KENDALL $\tau$ as $\tau := \sum P - \sum I$

a. Follow the algorithm by hand – but MAXIMA is your friend.
b. Make a 'video' a la example.5.3.1.
c. Verify

$$\tau = \frac{99 - 53}{99 + 53}$$

d. Verify using $\mathcal{R}$: `cor(X,Y, method="kendall")` gives `[1] -0.301641`.

**Exercise 245.** (significance of Kendall-$\tau$)
Use the FACT

The test for significance of KENDALL $\tau$ use the statistics

$$Z = \mid \tau \mid \sqrt{\frac{2 \cdot (2 \cdot n + 5)}{9 \cdot n \cdot (n - 1)}}$$

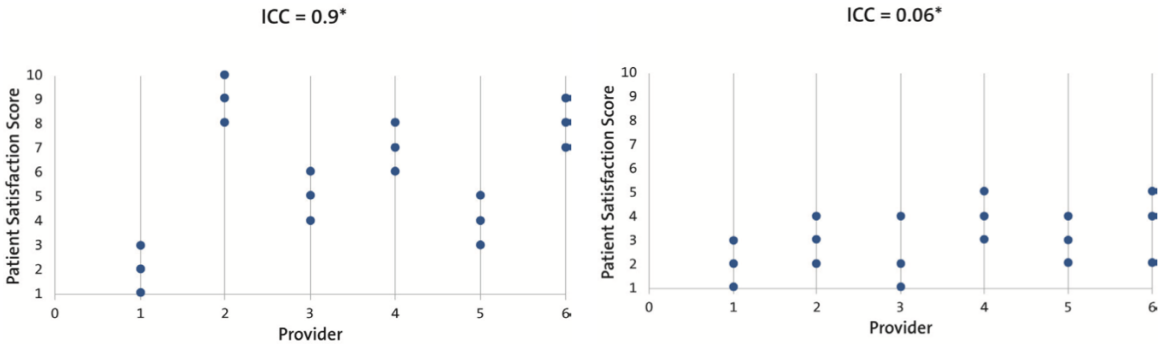
which is std.normal-distributed.

and do an $F$ test for the significance of the $\tau$.
Solution should be 1.74 for the ZOEFEL example.

## 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 44: 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

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

**Example** *Example and text from* ZOEFEL[22, 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 MAXIMA :

```
|  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
```

*Solution*: step-by-step, `example.5.4.1`
*Solution*: automatic, `example.5.4.2`

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

### 5.4.1 Exercises

**Exercise 246.** (early ICC after PEARSON/FISHER)
Program the 'early ICC formula' ▷ iccOld in MAXIMA and solve our example.5.4 with it. Compare both solutions.
You may use the following pseudocode:

```
### ----------------------   ICCp   ------------------------
###     iccP(X,Y)
###     method = Intraclass Correlation after PEARSON/FISHER
### ......................................................
iccP(X,Y,  Z,T,S,n) = do(
          n = dim(X),
          Xbar = 1/(2*n)*sum(i,1,n, X[i]+Y[i]),
          sigma2 = 1/(2*n)*( sum(i,1,n, (X[i]-Xbar)^2) +
                             sum(i,1,n, (Y[i]-Xbar)^2) ),
          r = 1/(n*sigma2)*sum(i,1,n, (X[i]-Xbar)*(Y[i]-Xbar)),
          float(r))
```

**Exercise 247.** (significance of ICC)
Use the fact, that the test for significance of ICC use the statistics

```
F = MQrow/MQrest
```

which is $F$-distributed with $(n-1; n\cdot(k-1))$ degrees of freedom.
Do an $F$ test for the significance of the ICC of example.5.4.

**Exercise 248.** (Check example.5.4 with $\mathcal{R}$)
Check the solution of example.5.4 using package `icc` of $\mathcal{R}$.

```
R> install.packages('irr')
R> library('irr')
R> data <- data.frame( X = c(48,48,50,50,52,58,63,56,48,63,58,58,52,60),
                       Y = c(51,50,50,52,53,63,70,70,49,63,59,58,56,62))
R> icc(data, model = "oneway", type = "agreement", unit = "single")

           Single Score Intraclass Correlation
   Model: oneway   Type : agreement
   Subjects = 14  Raters = 2
     ICC(1) = 0.743

   F-Test, H0: r0 = 0 ; H1: r0 > 0;  F(13,14) = 6.79 , p = 0.00053
   95%-Confidence Interval for ICC Population Values:  0.385 < ICC < 0.909
```

## 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

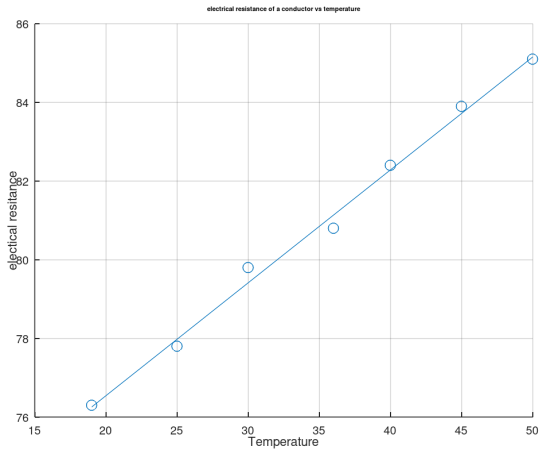



Figure 45: Visualization of linear regression (line):
electrical resistance of a conductor as a function of temperature.
–: regression line $y = 0.28 \cdot x + 70.8$ of the measurement $X \mapsto Y$ with
○: $(19, 25, 30, 36, 40, 45, 50) \mapsto (76.3, 77.8, 79.8, 80.8, 82.4, 83.9, 85.1)$.

### 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. [7, 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 MAXIMA in

```
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);
```

| a | b |
|---|---|
| 0.28757 | 70.807 |

*Solution*: step-by-step, `example.5.5.1`
*Solution*: automatic, `example.5.5.2`

### General Information

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

### 5.5.1 Exercises

**Exercise 249.** (example from [17, p.361, P 14.12])
Is there a connection between grade point average (GPA) and hours of TV watched per week. There is a sample of 10 hight school studs:

```
TVhours = (20,5,8,10,13,7,13,5,25,14)
GPA     = (2.35, 3.8, 3.5, 2.75, 3.25, 3.4, 2.9, 3.5, 2.25, 2.75)
```

Result: The variable TVhours and GPA are negatively correlated. i.e. as more hours of TV are watched, the lower is the GPA :(

**Exercise 250.** (check example.5.5 with $\mathcal{R}$)
Check the solution using `lm` function of `R`:

```
R> temperature <- c(19,25,30,36,40,45,50)
R> resistance  <- c(76.3, 77.8, 79.8, 80.8, 82.4, 83.9, 85.1)
R> lm(resistance ~ temperature)
R> plot(temperature, resistance)+abline(lm(resistance ~ temperature))

   Call:
       lm(formula = resistance ~ temperature)
       Coefficients:
             (Intercept)  temperature
              70.8065     0.2876
```

## 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**

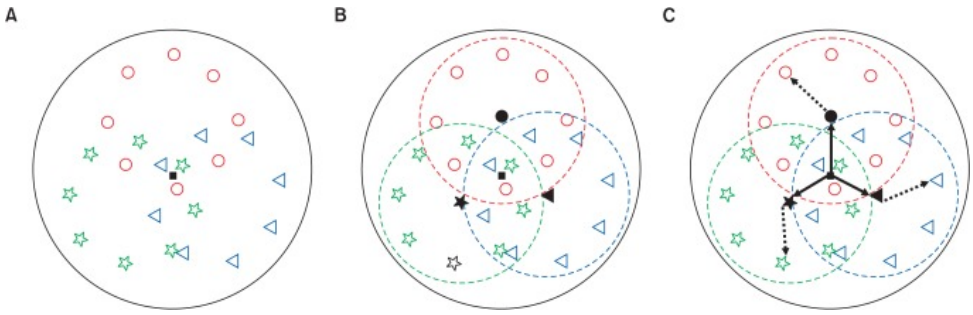



Figure 46: Visualization of *One-way Analysis Of Variance* test of three groups.
A: A solid black square "□" is suggested as a general representative value such as mean of overall data.
B: It looks reasonable to divide the data into three groups and explain the data with three different means of groups: •, ⋆, △.
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 $j$th treatment (row) of $X$.

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

**Example** *Yields in bushels per acre*

We do the example 16.4 of M. SPIEGEL [17, 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 MAXIMA :

```
X : [[3,4,5,4],[2,4,3,3],[4,6,5,5]]$
anova1(X);
```

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Between | 2 | 8 | 4.0 | 6.0 |
| Within | 9 | 6 | 0.66667 | - |
| Total | 11 | 14 | - | ANOVA |

*Solution*: step-by-step, `example.5.6.1`
*Solution*: automatic, `example.5.6.2`

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

### 5.6.1 Exercises

**Exercise 251.** (mineral water company, cf. [22, p.179])
A mineral water company claims that its mineral water lowers cholesterol levels. Eight test subjects drank this water for three weeks. The values at the beginning of the trial, as well as after one week, two weeks, and three weeks, are included in the following table.

```
A  =(267,248,321,272,355,264,270,252)  -- values at begin of test
W1 =(238,232,307,295,348,260,266,249)  -- after 1 week
W2 =(191,246,295,270,330,262,295,220)  -- after 2 weeks
W3 =(206,207,282,269,275,281,263,219)  -- after 3 weeks
```

The aim is to determine whether these mean values differ significantly from one another, i.e., whether the mineral water actually has a significant effect on cholesterol levels. These are dependent samples.
Result: According to the F-table, this is a significant value at $(3, 21)$ degrees of freedom ($p < 0.05$). Therefore, a significant decrease in cholesterol levels is observed over the time period.

**Exercise 252.** (blood pressure from numiqo, cf. anova)
Do the example in numiqo.

**Exercise 253.** (salaries of people from excel-easy, cf. salary)
Do the example in excel-easy.

**Exercise 254.** (Example from *statsdirect*, cf. statsdirect)
Do the example in www.statsdirect.com: *analysis_of_variance/one_way*.

**Exercise 255.** (Example from geeksforgeeks, cf. anova1)
Do the example in geeksforgeeks.org.

**Exercise 256.** (Example from kent.edu, cf. signtest)
Do the example in kent.edu : SPSS/OneWayANOVA.

**Exercise 257.** (Example from statsandr.com, cf. anova-by-hand)
Do the example in statsandr.com: *how-to-one-way-anova-by-hand*.

**Exercise 258.** (check with $\mathcal{R}$)
Check the solution of the example.5.6 using using `anova1` function of `octave`.

```
octave:1> pkg load statistics
octave:2> X=[3,4,5,4; 2,4,3,3; 4,6,5,5]
octave:3> anova1(X)

    ANOVA Table
    Source        SS      df        MS       F      Prob>F
    -----------------------------------------------------
    Groups      8.0000     2     4.0000     6.00    0.0221 =  p value
    Error       6.0000     9     0.6667
    Total      14.0000    11
```

## 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

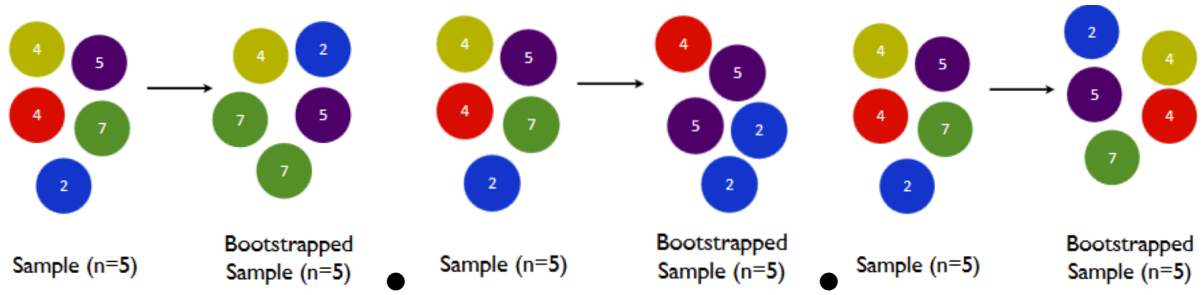



Figure 47: Visualization of `bootstrap`.
Left: From sample $X = \{4, 5, 4, 7, 2\}$ with mean $\bar{X} = 4.4$ we draw a new ('bootstrapped') sample giving $X' = \{4, 2, 7, 5, 7\}$ with mean $\overline{X'} = 5$.
Middle: From the original sample $X$ we draw ('resample') a new sample $X'' = \{4, 5, 5, 2, 2\}$ 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.47. 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 MAXIMA

```
X : [4,5,4,7,2]$
boot1(X,1000);
```

| est.mean | est.sd |
|---|---|
| 4.476 | 0.51367 |

*Solution*: step-by-step, `example.5.7.1`
*Solution*: automatic, `example.5.7.2`

**General Information**

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

### 5.7.1 Exercises

**Exercise 259.** (Marihuana, cf. [7, p.205])

> "From a given sample, any number of further samples can be drawn using random numbers. Compare the number of random samples that meet a given condition with the total number of samples drawn. This allows a statement to be made about the significance level of the test."

The following sample of memory test scores after smoking a regular cigarette and a Marijuana cigarette is considered.

```
person: 1   2  3  4  5  6  7  8  9
cigar: -3  10 -3  3  4 -3  2 -1 -1
marih.: 5 -17 -7 -3 -7 -9 -6  1 -3
```

The related samples are combinatorially interpreted as follows: First, the memory performances $X$ after smoking a regular cigarette and $Y$ after smoking a marijuana cigarette are combined into a new sample of 18 elements.
In our example, after inputting a sample of size n, 1000 pairs of samples of size n/2 are drawn, a counter counts the number of sample pairs whose sums S resp. T satisfy the condition $|T - S| >= |D|$ where $D := \sum(Xi - Yi)$.
Use the pseudocode to program a adapted bootstrap method

```
...
z = 0,                  -- counter
n = dim(X),             -- number of samples
k = n/2,
...
    D = abs(sum(T-S)),
    test(abs(T[j]-S[j]>D), z=z+1) ),
estT = mean(T),  -- bootstrap sample means
estS = mean(S),  -- i.e. estimated values
p=0.5*z/IT,      -- calculate probability
print(("est.mean", "est.sd","p"),(estT, estS, p ))
...
```

*Result*: For the marijuana example dataset, the program randomly returns values around 0.015 as the significance level. The null hypothesis, that memory performance after smoking ordinary and marijuana cigarettes differs only by chance, can be rejected with 1.6% probability of error.

**Exercise 260.** (Bootstrap Samples of Observations from MatLab, cf. bootstrp)
Do two bootstrap examples from MatLab.
a. Take bootstrap samples of patient data, compute the mean measurements for each data sample, and visualize the results.
b. Create 50 bootstrap samples from the numbers 1 through 6. To create each sample, bootstrp randomly chooses with replacement from the numbers 1 through 6, six times. This process is similar to rolling a die six times. For each sample, the custom function *countfun* (to be programed by you) counts the number of 1s in the sample.

## 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 $\theta$ ($\overset{e.g.}{=}$ *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**

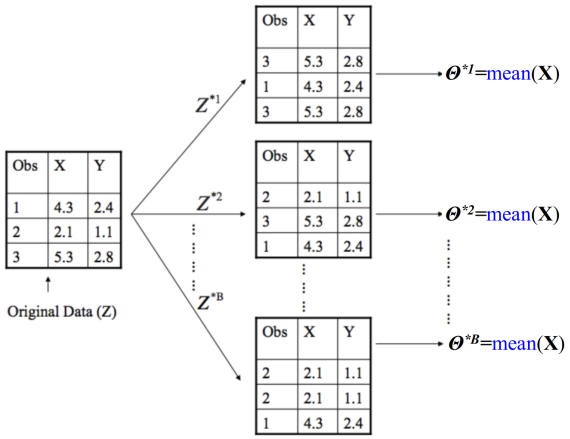



Figure 48: Visualization of `bootstrap2`. We generate new samples $Z^{*1}, Z^{*2}, Z^{*B}$ directly from the population `obs`ervations (leftmost histogram; *obs*).
If we choose $\theta \overset{p.d.}{=} mean$; then $\bar{X} = 3.83$ and $\bar{Y} = 2.1$.
We generate the distribution of sample means $B$-times, 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* $\theta$ to be studied resp. estimated, e.g. $\theta := mean$. [23]
   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.[24]

[23] For independent samples, this is often the difference between the two sample means.
[24] 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 $\theta_i$ 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 $\theta$.

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 $\theta^{*X}$ and $\theta^{*Y}$ 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 MAXIMA

```
X : [4.3, 2.1, 5.3]$
Y : [2.4, 1.1, 2.8]$
boot2(X,Y, 1000)$
```

| e.meanX | e.meanY | significance |
|---|---|---|
| 1.8667 | 4.1333 | 0.055 |

*Solution*: automatic, `example.5.8.2`

**General Information**

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

### 5.8.1 Exercises

**Exercise 261.** (31 US presidents, cf. [7, p.207])
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.

```
X = (85,79,90,67,80)

Y = (67,56,53,77,88,57,63,78,70,
     66,83,67,60,74,65,64,73,71,
     60,90,63,71,64,78,68,72)
```

The null hypothesis is: age and height are independent.
Use `boot2` to test the hypothesis.

Result: We get values around 0.013. With a 5% probability of error, the null hypothesis, that age and size differ only by chance, can be rejected.

**Exercise 262.** (Example from MatLab, cf. bootstrp)
Do an example from the MatLab documentation in Maximausing the data from section 'Medians of Paired Samples'.

## 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**

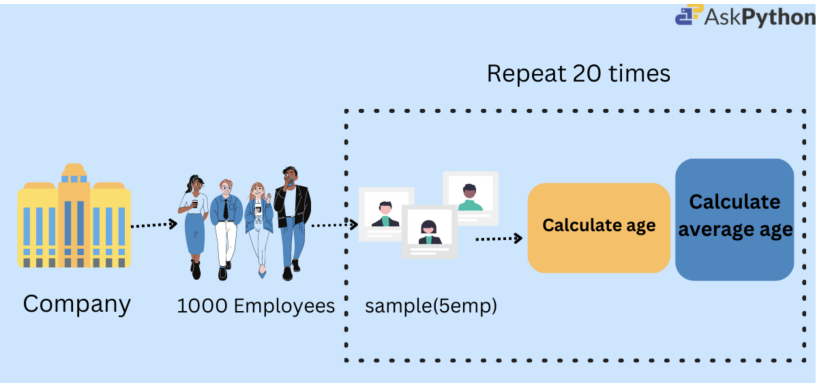



Figure 49: Visualization of `bootCI`: Instead of taking multiple samples directly from the whole employees population, we only draw a single, representative *sample* (5emp). We then generate e.g. 20 'new' samples by *repeatedly* create new bootstrap samples of the same size. – Picture found at ▷ ASKPYTHON : 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 $T := T_1, \ldots, T_n$ [25] of the $n$ resamples $T_i$ is defined by:

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

[25] 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 MAXIMA.
We check the solution using our function `bootci` and 100 resamples:

```
X : [25, 30, 35, 40, 45, 50, 55, 60, 65, 70]$
bootci(X,100);
```

| est.Mean | CI.l | CI.h |
|---|---|---|
| 48.135 | 39.5 | 50.5 |

*Solution*: step-by-step, `example.5.9.1`
*Solution*: automatic, `example.5.9.2`

### General Information

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

### 5.9.1 Exercises

**Exercise 263.** (Check the solution using $\mathcal{R}$)
Check the solution of example.5.9 using function `boot.ci` of `R` and 999 resamples:

```
R> X <- c(25, 30, 35, 40, 45, 50, 55, 60, 65, 70)
R> library(boot)
R> mean.fun <- function(d, i) { m <- mean(X[i]) }
R> X.boot   <- boot(X, mean.fun, R = 999)
R> boot.ci(X.boot, type = c("perc"))            # method = "percentile"

        BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
        Based on 999 bootstrap replicates
    CALL :       boot.ci(boot.out = X.boot, type = c("perc"))
    Intervals :  Level     Percentile
                 95%       (38, 56 )
    Calculations and Intervals on Original Scale
```

**Exercise 264.** (Example from askPython)
Ten university students were recruited to play the seminal video game pac-man. After equal amounts of practice the students were then asked to play the first stage and the time to completion was recorded.
The following data in seconds were recorded.

```
X = (63.1,64.6,63.4,65.6,63.4,65.7,64.6,63.1,65.9,63.3)
```

Calculate the 95% confidence interval CI, which catch the true population mean.

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

**Exercise 265.** (cytokine responses)
In these data we have reason to believe that the confidence intervals will not always be symmetric about the mean. Let the data be the cytokine responses (in %) for a sample measured by 16 laboratories. For the purposes of this example we will assume independence among the samples.

```
X = (0.4043, 0.5958, 0.5876, 0.5939, 0.6053, 0.4649, 0.5722, 0.6234,
     0.5208, 0.6117, 0.6137, 0.7397, 0.6981, 0.7757, 0.7596, 0.532)
```

Do a bootCI for X with 100 and 1000 iterations.

**Exercise 266.** (A 95% Confidence Interval for Movie Run Time, cf. usu.edu)
Do the example 'Movie Run Time' of the *usu.edu*.

**Exercise 267.** (cholesterol)
Serum cholesterol (mmol/L) measured on a sample of 86 stroke patients (data of Markus et al., 1995):

```
X=(3.7, 4.8, 5.4, 5.6, 6.1, 6.4, 7.0, 7.6, 8.7,
   3.8, 4.9, 5.4, 5.6, 6.1, 6.5, 7.0, 7.6, 8.9,
   3.8, 4.9, 5.5, 5.7, 6.1, 6.5, 7.1, 7.6, 9.3,
   4.4, 4.9, 5.5, 5.7, 6.2, 6.6, 7.1, 7.7, 9.5,
   4.5, 5.0, 5.5, 5.7, 6.3, 6.7, 7.2, 7.8, 10.2,
   4.5, 5.1, 5.6, 5.8, 6.3, 6.7, 7.3, 7.8, 10.4,
   4.5, 5.1, 5.6, 5.8, 6.4, 6.8, 7.4, 7.8,
   4.7, 5.2, 5.6, 5.9, 6.4, 6.8, 7.4, 8.2,
   4.7, 5.3, 5.6, 6.0, 6.4, 7.0, 7.5, 8.3,
   4.8, 5.3, 5.6, 6.1, 6.4, 7.0, 7.5, 8.6)
```

Calculate the CI using `bootci(X,100))`.

**Exercise 268.** (CI for correlation coefficient after EFRON, cf. [7, p.209 ff])

> The bootstrap method was first applied in determining the non-parametric determination of a correlation coefficient. In 1973, Efron determined the average grades of all law school graduates (great point average, GPA) and at the law school admission test (LSAT) at 82 American law schools. By means of a random sample with 15 students, the following values were obtained.

EFRON used the data LSAT and GPA

```
LSAT = (576,635,558,578,666, 580,555,661,651,605, 653,575,545,572,594)

GPA = (3.39,3.30,2.81,3.03,3.44, 3.07,3.00,3.43,3.36,3.13,
       3.12,2.74,2.76,2.88,2.96)
```

◦ EFRON developed the idea of repeatedly drawing a random sample from the given sample to determine the variance. The correlation coefficient of 1000 samples he drew had a variance of 0.254.

a. First calculate the parametric Pearson correlation coefficient.
b. Second calculate the Confidence Interval CI for *rho*; remember: we need the inverse of the standard normal distribution to calculate quantiles.
c. Now define the confidence interval function for Pearson rho: `CIrho(rho,n,alpha)` and calculate the CI for rho at level 68.3% for rho = 0.7763.
d. Third calculate the CI using bootstrap method. Therefore define the following specialized bootstrap function: `bootCIrho(X,Y, rho)` and use `bootCIrho` on the Efron data.
The parametric range of deviation is 0.2347 and the non-parametric bootstrap range is 2*sigma 0.27, which is a approx. 0.05 greater.

## 5.10 `jackknife` - The Jackknife method

Jackknife is a statistical resampling technique that estimates the bias and variance of a statistic $\theta$ *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 $\theta_i$ is used to derive bias-corrected estimates and confidence intervals for the original statistic. ▷ Google AI : Jackknife resampling

**Mental image**

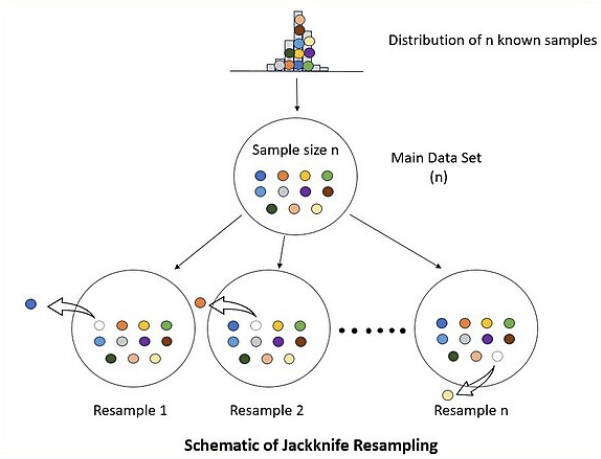



Figure 50: Visualization of `jackknife`: The 'leave-one-out' jackknife algorithm. Instead of taking multiple samples directly from the population, we only have a single, representative *sample*. We then generate 'new' samples by *repeatedly leaving out one observation at a time* from the original *sample*. – 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 $i$th 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, $\theta$, 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 $i$th observation $x_i$ from $X$ to form the new $i$th jackknife sample $X_i$.

▷ Compute the $i$th jackknife replicate statistic, $\theta_i$, by computing the statistic on the $i$th 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 $\sigma_J$ 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 $J_i$:
   $J_1$ : $X_1 = (2, 3, 4) \to \theta \cdot n - (n-1) \cdot stats(X_1) = 2.5 \times 4 - 3 \times 3 = 1$
   $J_2$ : $X_2 = (1, 3, 4) \to \theta \cdot n - (n-1) \cdot stats(X_2) = 2.5 \times 4 - 3 \times 8/3 = 2$
   $J_3$ : 1 2 4 $\to \cdots = 3$
   $J_4$ : 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 MAXIMA

```
X : [1,2,3,4]$
jack(X, 0.05)$
```

| SE | CI.l | CI.h |
|---|---|---|
| 0.6455 | 4.5543 | 0.44574 |

- *Solution*: step-by-step, `example.5.10.1`
- *Solution*: automatic, `example.5.10.2`

### General Information

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

### 5.10.1 Exercises

**Exercise 269.** (nonparametric 'leave-one-out' jackknife algorithm in $\mathcal{R}$, cf. jackknifing) The corresponding $R$ code is partially found at *influentialpoints.com*:

```
R> ## The nonparametric 'leave-one-out' jackknife algorithm in R
R> X     <- c(1,2,3,4)               # assign data
R> stats <- function(y){mean(y)}     # define statistics of interest as mean
R> n     <- length(X)
R> theta <- stats(X)                 # the total mean of all of X
R>          # now calculate 'pseudo-values' (pv)
R> pv = i <- 0; while(i < n){i<-i+1 ; pv[i] <- theta*n - (n-1)*stats(X[-i]);
R>              print(X[-i]); print(pv[i])}    # show process

R> J    <- mean(pv)           # estimate statistic as mean of pseudo-values
R> bias <- mean(pv) - th      # estimate bias
R> se   <- sqrt(var(pv)/n)    # estimate standard error SE
R> se

R> # calculate the confidence interval
R> alpha <- 1 - 0.95
R> ci    <- qt(alpha/2,n-1,lower.tail=FALSE)*se
R> CI    <- c(J - ci, J + ci)         # Confidence Interval
R> CI
           The nonparametric 'leave-one-out' jackknife
   [1] 2 3 4   # leave 1 out in X=(1,2,3,4)
   [1] 1
   [2] 1 3 4   # leave 2 out
   [2] 2
   [3] 1 2 4   # leave 3 out
   [3] 3
   [4] 1 2 3   # leave 4 out
   [4] 4
   [5] 0.6454972              SE
   [6] 0.4457398 4.5542602    CI
```

**Exercise 270.** (birth weights from cf. [7, p.212])
The birth weights of 72 boys and 48 girls born in the first half of 1981 at the Freiburg (Germany) clinic were examined.
The 72 boys data were

```
X = (3700,3800,3040,3080,3170, .., 2880 ,3150, 3160)
```

The 48 girls data were

```
Y = (3700,3800,3040,3080,3170, .., 2880 ,3150, 3160)
```

The 120 combined data is

```
Z = (2990,3700,3800, 3040,3080,3350,2930,3355,3360,3170.,
     3220,3850,3080,3440,3840,3320,2640,3020,2850,3700,
     4630,3670,3400,3580,3270,3360,3600,3220,4060,2670,
     3550,3690,2510,1980,3700,3730,3170,3580,3720,2910,
     3500,3430,3510,1520,3420,4630,3470,3400,3290,3250,
     2420,3580,3100,3480,3830,3050,3680,3330,3650,3200,
     3810,3350,3390,3550,3350,3200,4190,3050,4020,2880,
     3020,3010,3560,3030,3200,2830,3120,3000,4010,3050,
     3980,3530,3245,3415,3660,4350,3740,3220,3350,3350,
     3780,3620,3310,3630,3290,3800,4130,3380,3100,3430,
     3620,3060,2610,3560,3720,3030,3630,3400,3580,3700,
     3470,2810,2330,1960,3300,2880,3660,3150,3160,3300)
```

From this sample of size 120, there are 120 jackknife samples generated and their mean is determined. This is intended to provide a non-parametric 95% confidence interval for the sample mean.
The available point estimate for the mean is 3352.8.
The non-parametric 95% confidence interval for the mean is $[3267.9, 3437.7]$.

Task: verify the results above using
- .. 3352.8 jack estimation of mean, theta = jac
- .. $[3267.88, 3437.71]$ CI, our code
- .. $[3267.89, 3437.69]$ CI jknife, $\mathcal{R}$ code
- .. $[3297.3, 3458.1]$ parametric CI

# 6 Appendix - the statsBox

If you want to run the MAXIMA functions from chapter 1 – 3 of this book for own work, you can load them all with

- ▷ statsBox.txt in your editor

or with

- ▷ statsBox.mac in MAXIMA or YAMWI.

# 7 Bibliography

# References

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

[2] Bognar, M.: *Probability Distribution Applets.* url: `https://mabognar.github.io/apps/`

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

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

[5] Gosse, M. (2026): Maxima *par l'Example.* url: `https://maxima-french-doc.fr/wp-content/uploads/2026/04/guide-eng.pdf`

[6] 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.

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

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

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

[10] Lindner, W. (2025): *Statistics with* Eigenmath. url: `https://lindnerdrwg.github.io/Statistics-with-Eigenmath.pdf`

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

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

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

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

[15] SoftMaker (2024): *PlanMaker Free 2024.* url: `https://www.freeoffice.com/en/?option=com_content`

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

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

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

[19] Weigt, G. (2025): Eigenmath *Command List.* url: `https://georgeweigt.github.io/`

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

[21] Woollett, E. (2024): *Statsitics with* Maxima. url: `https://home.csulb.edu/~woollett/stat.html`

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

⋈

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