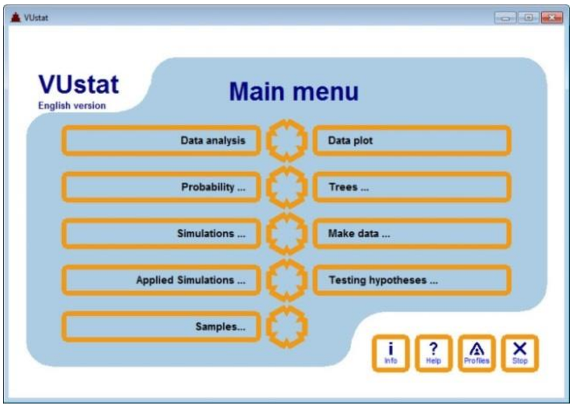
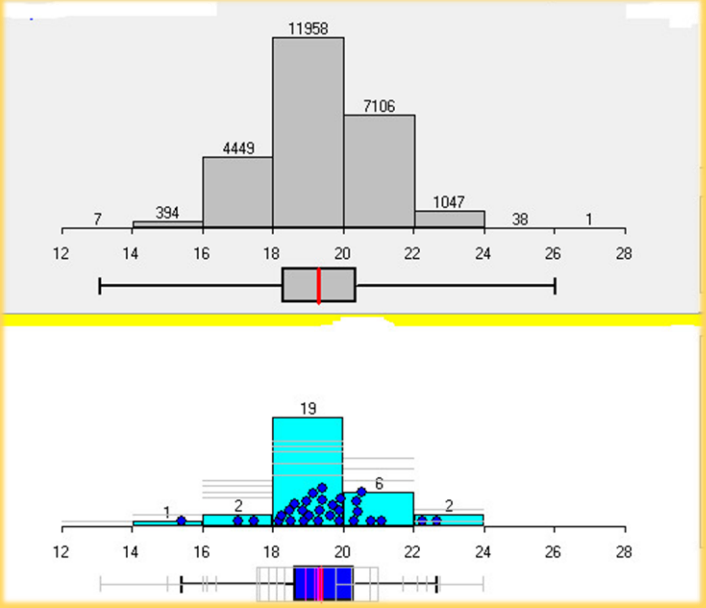
Visual Understanding

educational software for maths



VUstat is a program with a wealth of possibilities to illustrate ideas and concepts in statistics education. VUstat is a statistical package for different levels of education for students and teachers. It contains data, tools and simulations to enhance the ability to teach statistics and probability to students of various levels. VUstat consists of many different modules. Many parts of the program can be turned off by a special tool called *profiles*, so that the student is not confronted with too much complexity at lower levels.

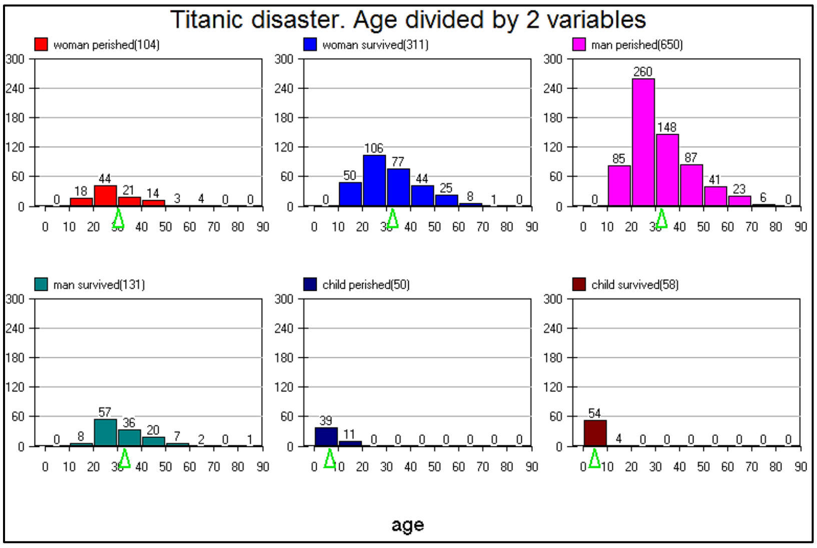
Visual Understanding software
www.vusoft.eu



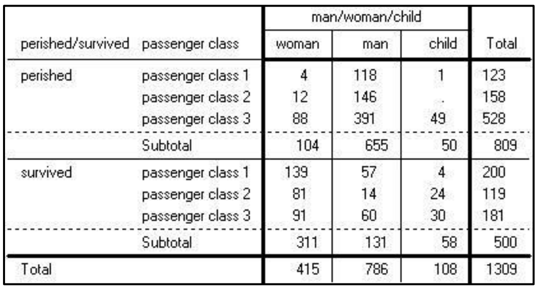
"You don't have to eat the whole ox to know that the meat is tough"
(David Moore about sampling)

Data analysis

Data analysis enables you to make the more general tables and graphs of raw data input. It has advanced options, but its main purpose is to make the analysis of data easier to understand for the student who is neither a statistics expert nor a computer whiz. The selection and dividing of data into subgroups is made straightforward.

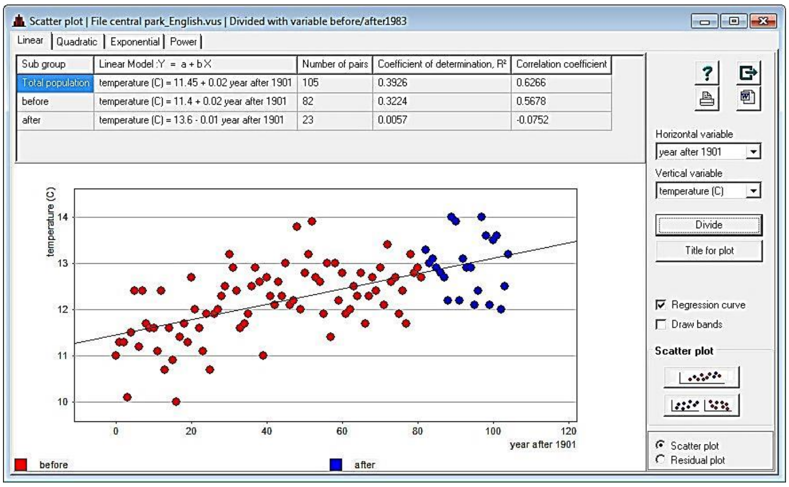


Histogram from a dataset of the passengers of the Titanic. The diagrams show passengers age divided by the variables sex and survived/perished. In the diagrams we also show mean values (optional).



		man/woman/child			
perished/survived	passenger class	woman	man	child	Total
perished	passenger class 1	4	118	1	123
	passenger class 2	12	146	.	158
	passenger class 3	88	391	49	528
	Subtotal	104	655	50	809
survived	passenger class 1	139	57	4	200
	passenger class 2	81	14	24	119
	passenger class 3	91	60	30	181
	Subtotal	311	131	58	500
Total		415	786	108	1309

Contingency table (cross table) divided by variable passenger class. The table can be copied and pasted in text format into other applications, e.g. Microsoft Word®.



Scatter plot | File central park_English.vus | Divided with variable before/after1983

Sub group	Linear Model :Y = a + bX	Number of pairs	Coefficient of determination, R²	Correlation coefficient
Total population	temperature (C) = 11.45 + 0.02 year after 1901	105	0.3926	0.6266
before	temperature (C) = 11.4 + 0.02 year after 1901	82	0.3224	0.5678
after	temperature (C) = 13.6 - 0.01 year after 1901	23	0.0057	-0.0752

Horizontal variable: year after 1901

Vertical variable: temperature (C)

☑ Regression curve

☐ Draw bands

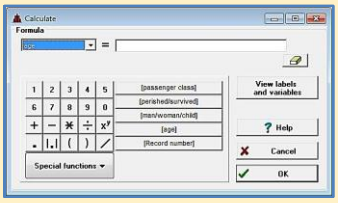
● Scatter plot

○ Residual plot

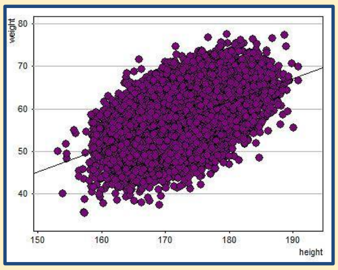
Climate data for more than 100 years. Scatter plot with data divided into two variables. Regression line and correlation coefficient are calculated.

Data analysis

- ✓ Import of data from the Web
- ✓ Import/export from and to Excel®
- ✓ Built-in calculator for recalculating data



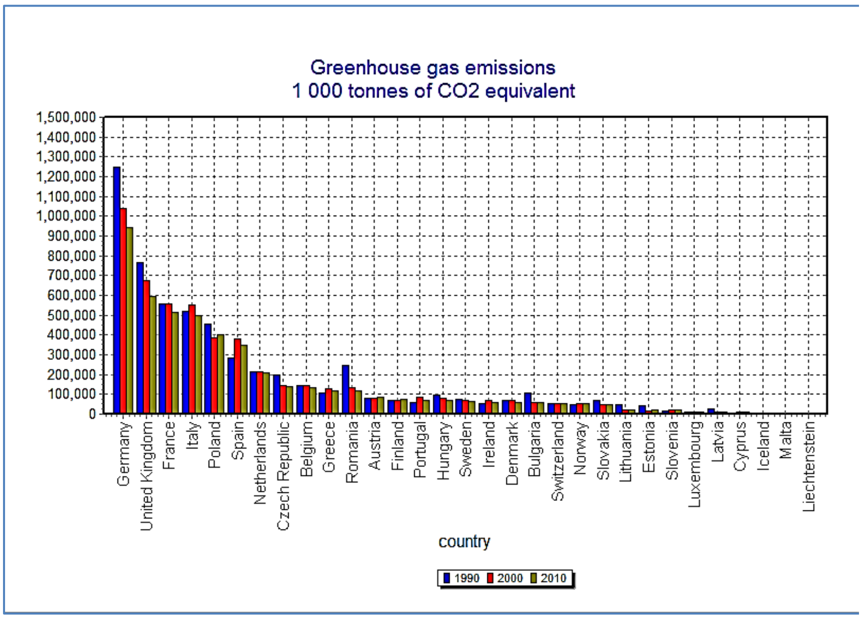
- ✓ Sorting data
- ✓ 5 different types of variables: text, integers, real numbers, labels and multiple response
- ✓ Options for filtering data
- ✓ Drawing samples from a dataset
- ✓ 8 different types of diagrams: Bar chart/Histogram, Dotplot, Pie chart, Frequency polygon, Box plot, Scatter plot, Time series, Lorenz curve



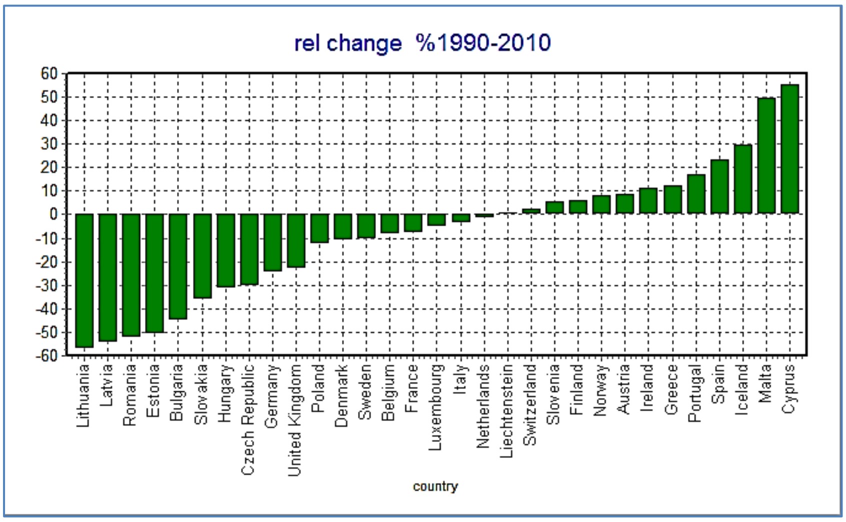
- ✓ 3 types of tables: Frequency table, Cross table, Stem and leaf plot
- ✓ Data can be divided for up to two variables.
- ✓ Aggregation of data

Data plot

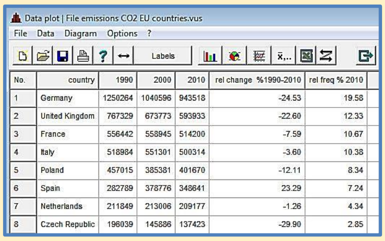
The purpose of this feature is to make graphs of ordered data. There are built-in features for design of diagrams. Some of the tools in the module Data Analysis are here too. Dataplot is especially useful for analyzing *frequency tables*. Traditionally, much teaching in the lower grades is concentrated on frequency tables. The Dataplot module allows students to do more than drawing graphs; it allows them to concentrate on the more useful task of comparing and interpreting data.



Emissions from all European countries (no figures for Croatia), from highest to lowest.

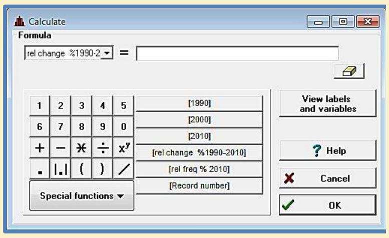


Once the files have been opened in modules Data plot and Data analysis you can toggle between them with a shortcut button.

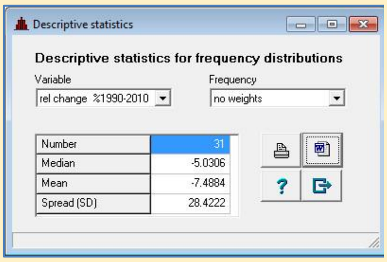


No.	country	1990	2000	2010	rel change %1990-2010	rel freq % 2010
1	Germany	1250264	1040596	943518	-24.53	19.58
2	United Kingdom	767329	673773	593933	-22.60	12.33
3	France	556442	558945	514200	-7.59	10.67
4	Italy	518984	551301	500314	-3.60	10.38
5	Poland	457015	385381	401670	-12.11	8.34
6	Spain	282789	378776	348641	23.29	7.24
7	Netherlands	211849	213006	209177	-1.26	4.34
8	Czech Republic	196039	145886	137423	-29.90	2.85

The Data table with the first four variables imported as an Excel file (from Eurostat). The last two variables are calculated with the built-in calculator.



Calculator



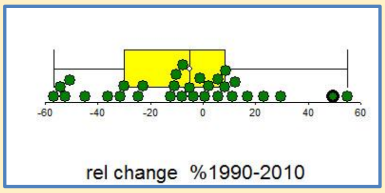
Descriptive statistics for frequency distributions

Variable: rel change %1990-2010

Frequency: no weights

Number	31
Median	-5.0306
Mean	-7.4884
Spread (SD)	28.4222

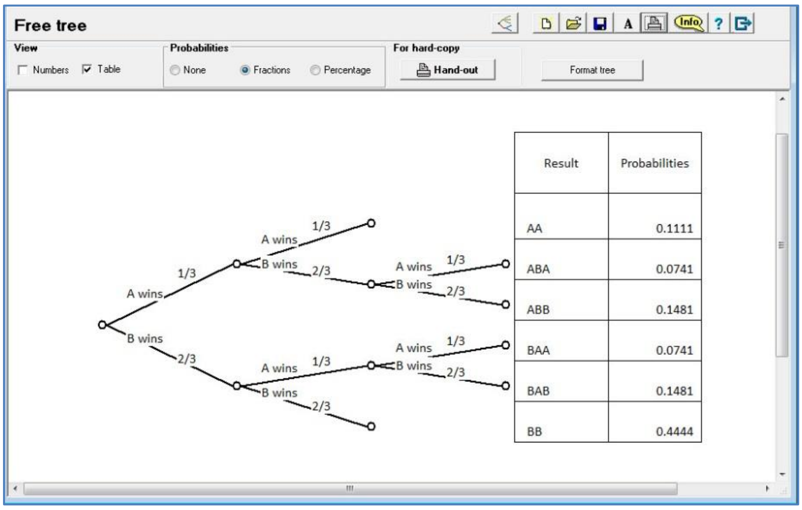
Summary statistics for the variable *rel change.*



The file can of course also be opened in the module Data analysis. Here is a Dotplot and a Box plot of the distribution of the variable *rel change*.

Trees

A **tree diagram** visualizes a probability tree from different points of view. Both a regular tree and a free tree can be constructed.



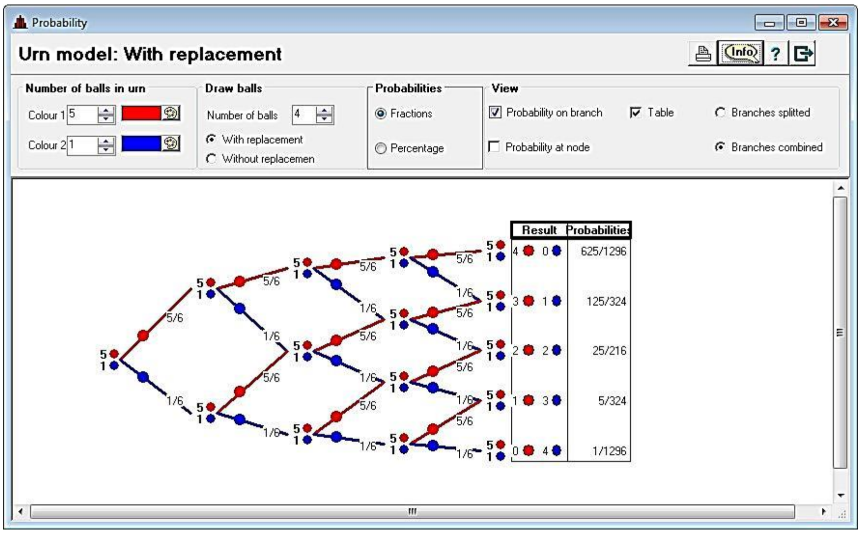
View: □ Numbers ☑ Table

Probabilities: ○ None ● Fractions ○ Percentage

Result	Probabilities
AA	0.1111
ABA	0.0741
ABB	0.1481
BAA	0.0741
BAB	0.1481
BB	0.4444

Free tree: Probability tree for a tennis match in 3 sets.

The module **Urn** shows the probabilities when balls are drawn from an urn with or without replacement. A lot of settings, as you can see on the screen.



Number of balls in urn: Colour 1: 5; Colour 2: 1

Draw balls: Number of balls: 4; ● With replacement ○ Without replacemen

Probabilities: ● Fractions ○ Percentage

View: ☑ Probability on branch ☑ Table ○ Branches splitted □ Probability at node ● Branches combined

Result		Probabilities
4 ●	0 ●	625/1296
3 ●	1 ●	125/324
2 ●	2 ●	25/216
1 ●	3 ●	5/324
0 ●	4 ●	1/1296

A demonstration of the famous Chevalier De Mere´s problem. The screen shows the possible outcomes and the probability of getting at least one "6" in four throws of a single 6-sided die.

Table and tree

Table and tree. Data in cross tables (absolute or relative values in rows or columns) can be represented in trees and diagrams in different ways.

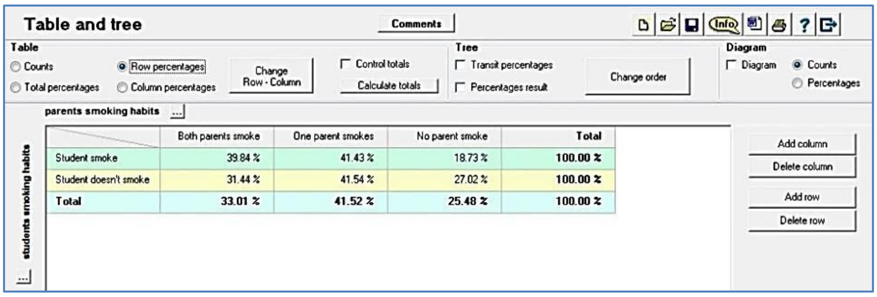


Table: ○ Counts ● Row percentages ○ Total percentages ○ Column percentages □ Control totals

Tree: □ Transit percentages □ Percentages result

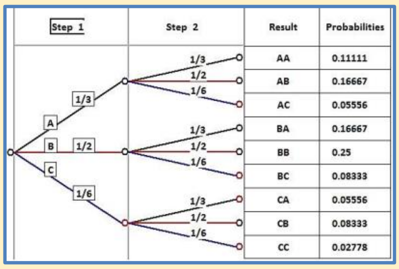
Diagram: □ Diagram ● Counts ○ Percentages

parents smoking habits / students smoking habits

	Both parents smoke	One parent smokes	No parent smoke	Total
Student smoke	39.84 %	41.43 %	18.73 %	100.00 %
Student doesn't smoke	31.44 %	41.54 %	27.02 %	100.00 %
Total	33.01 %	41.52 %	25.48 %	100.00 %

Trees

- ✓ Two different types of probability trees: a regular tree and a tree which you can design.

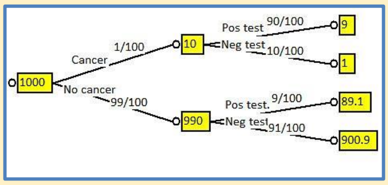


Step 1	Step 2	Result	Probabilities
		AA	0.11111
		AB	0.16667
		AC	0.05556
		BA	0.16667
		BB	0.25
		BC	0.08333
		CA	0.05556
		CB	0.08333
		CC	0.02778

Regular tree with three branches and two steps.

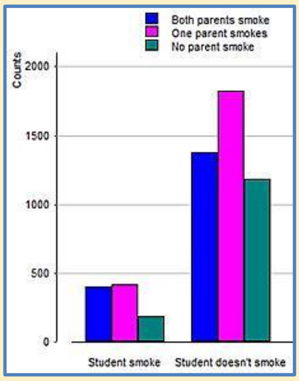
Assume you conduct breast cancer screening using mammography in a certain region. You know the following information about the women in this region:

- The probability that a woman has breast cancer is 1 % (prevalence)
- If a woman has breast cancer, the probability that she tests positive is 90% (sensitivity)
- If a woman does not have breast cancer, the probability that she nevertheless tests positive is 9% (false-positive rate)



From the tree diagram it is much easier to calculate the probability that a woman who tests positive has breast cancer if we work with *natural frequencies*.

$$P = \frac{9}{9+89} \approx 9\ \%$$

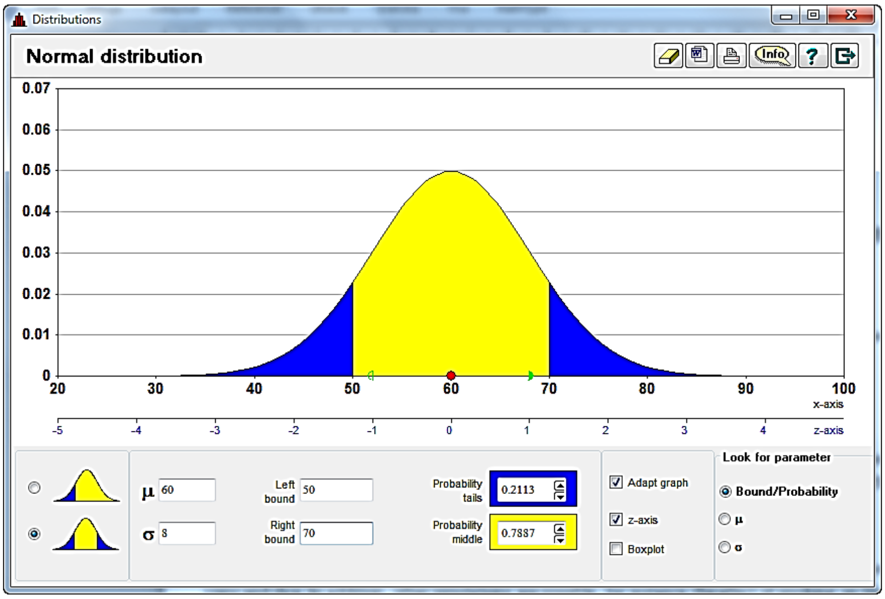


Probability

This section contains several options that are useful in counting problems and understanding of probability. The modules in this section are Distributions, Galtons board, Grid, Combinatorics and Frequency grid.

Distributions

The binomial, Poisson, hypergeometric, and normal distributions can easily be manipulated by students using VUstat. The central limit theorem is demonstrated with skewed and normal dice and you can compare plots of different distributions.



○ ● μ 60 σ 8

Left bound 50

Right bound 70

Probability tails 0.2113

Probability middle 0.7887

☑ Adapt graph

☑ z-axis

□ Boxplot

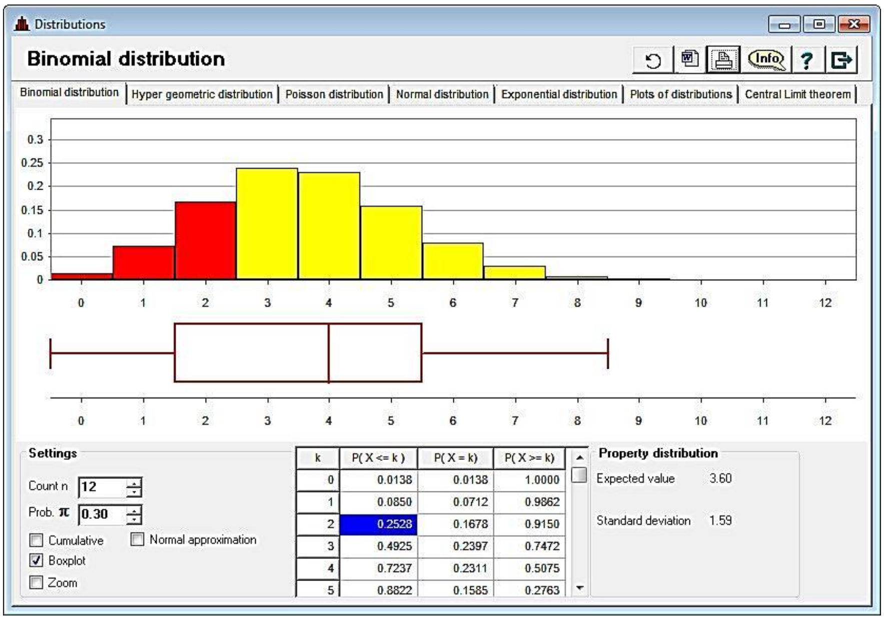
Look for parameter

● Bound/Probability

○ μ

○ σ

Normal distribution. You can change parameters by changing values in the frames or by dragging the bounds between the blue and yellow areas, the red points representing μ or the green arrows representing σ.



Settings

Count n 12

Prob. π 0.30

□ Cumulative □ Normal approximation

☑ Boxplot

□ Zoom

k	P(X <= k)	P(X = k)	P(X >= k)
0	0.0138	0.0138	1.0000
1	0.0850	0.0712	0.9862
2	0.2528	0.1678	0.9150
3	0.4925	0.2397	0.7472
4	0.7237	0.2311	0.5075
5	0.8822	0.1585	0.2763

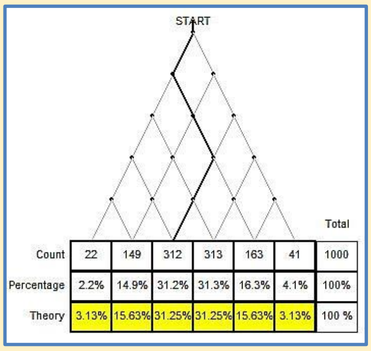
Property distribution

Expected value 3.60

Standard deviation 1.59

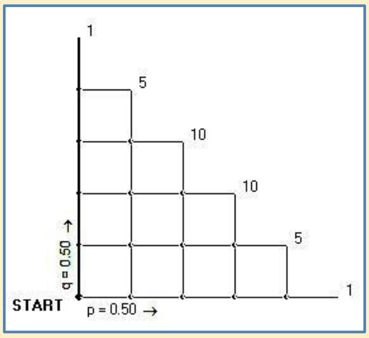
Binomial distribution. The number of trials is 12. The probability of success is 0.3 and we can see the probabilities for success in three different ways in the table. If we highlight a cell in the table we can see the corresponding bins in red color in the diagram. Here we have highlighted P(X≤2).

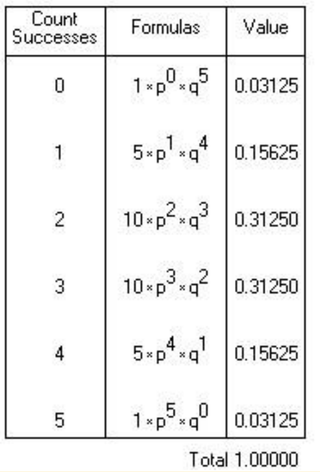
In this module there are two small applications for the binomial distribution: *Galton's board* and *Grid*.



							Total
Count	22	149	312	313	163	41	1000
Percentage	2.2%	14.9%	31.2%	31.3%	16.3%	4.1%	100%
Theory	3.13%	15.63%	31.25%	31.25%	15.63%	3.13%	100 %

Galton's board





Count Successes	Formulas	Value
0	$1 \times p^0 \times q^5$	0.03125
1	$5 \times p^1 \times q^4$	0.15625
2	$10 \times p^2 \times q^3$	0.31250
3	$10 \times p^3 \times q^2$	0.31250
4	$5 \times p^4 \times q^1$	0.15625
5	$1 \times p^5 \times q^0$	0.03125
	Total	1.00000

In the grid model a "success" is taking a step to the right. The number of trials is the total number of steps taken in the grid to get from start to an end point. You can see the result in different views. Here we show a probability table with formulas.

Combinatorics

The module Combinatorics deals with things that can be enumerated. For example, how many different ice creams with three flavors there are if you have six different flavors to choose from or how many handshakes take place if ten people meet and everyone shakes hands with everyone else.

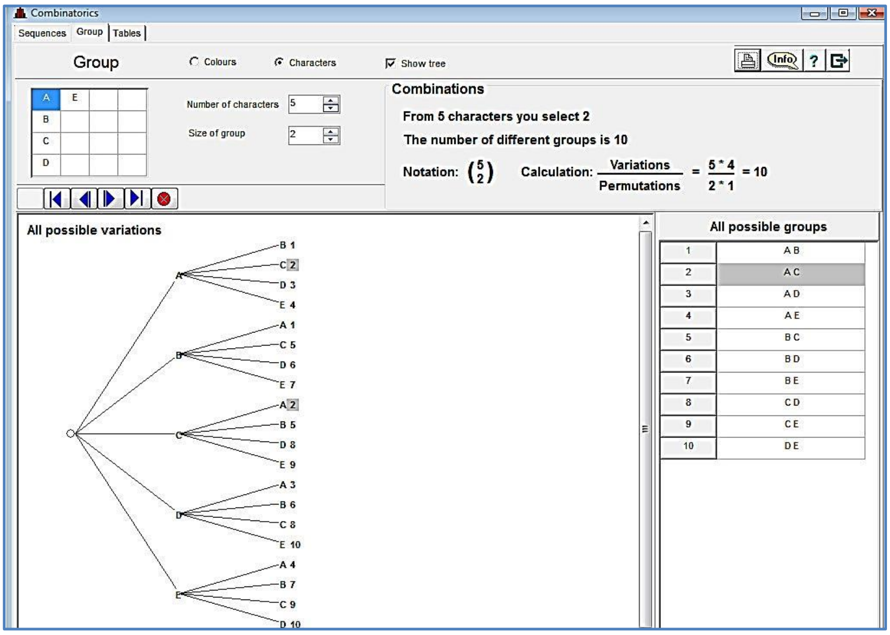
We illustrate concepts as variation, permutation and combination in three views and we use colors or characters as objects.

An example: Suppose we have two points in a plane. How many paths are there that pass only through the lines of the grid? Only horizontal (r) or vertical movements (u) are permitted.

All paths between the points have the same length, 3 right + 2 up. We can calculate the number of all the paths between the points as combinations: from five characters, you choose two.

$$\binom{5}{2} = 10$$

The possible ten paths are:
uurrr, ururr, urrur, urrru, rrruu
rruru, rruur, ruurr, rurur, rurru



Group ○ Colours ● Characters ☑ Show tree

A	E		
B			
C			
D			

Number of characters 5

Size of group 2

Combinations

From 5 characters you select 2

The number of different groups is 10

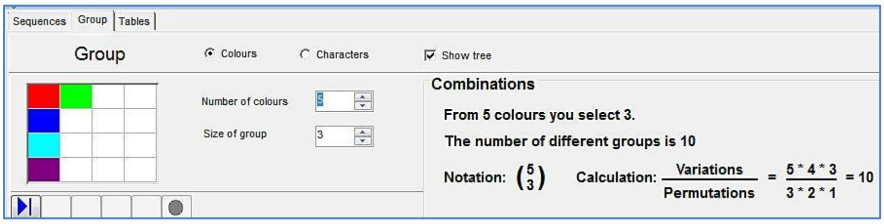
Notation: $\binom{5}{2}$ Calculation: $\frac{\text{Variations}}{\text{Permutations}} = \frac{5 * 4}{2 * 1} = 10$

All possible variations

All possible groups

1	A B
2	A C
3	A D
4	A E
5	B C
6	B D
7	B E
8	C D
9	C E
10	D E

*Option **Groups** gives you number of combinations. You can see all variations in a tree and all possible groups in a table.*



Group ● Colours ○ Characters ☑ Show tree

🟥	🟩		
🟦			
🟦			
🟪			

Number of colours 5

Size of group 3

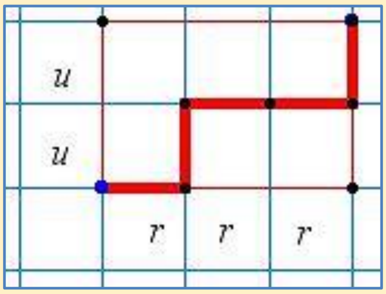
Combinations

From 5 colours you select 3.

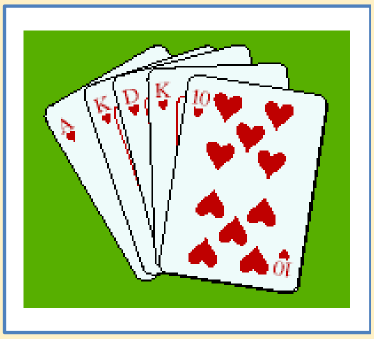
The number of different groups is 10

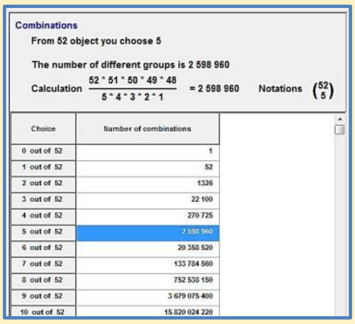
Notation: $\binom{5}{3}$ Calculation: $\frac{\text{Variations}}{\text{Permutations}} = \frac{5 * 4 * 3}{3 * 2 * 1} = 10$

If we choose 3 out of 5 we get the same result. Here we have colors as objects.



The path here is rurru.





Combinations

From 52 object you choose 5

The number of different groups is 2 598 960

Calculation $\frac{52 * 51 * 50 * 49 * 48}{5 * 4 * 3 * 2 * 1}$ = 2 598 960 Notations $\binom{52}{5}$

Choice	Number of combinations
0 out of 52	1
1 out of 52	52
2 out of 52	1326
3 out of 52	22 100
4 out of 52	270 725
5 out of 52	2 598 960
6 out of 52	20 358 520
7 out of 52	133 784 560
8 out of 52	752 538 150
9 out of 52	3 679 075 400
10 out of 52	15 820 024 220

From option *Tables*. Here you can show tables for both variations and combinations. The probability that five random cards drawn (out of 52) will give some unique hand where order does not matter is

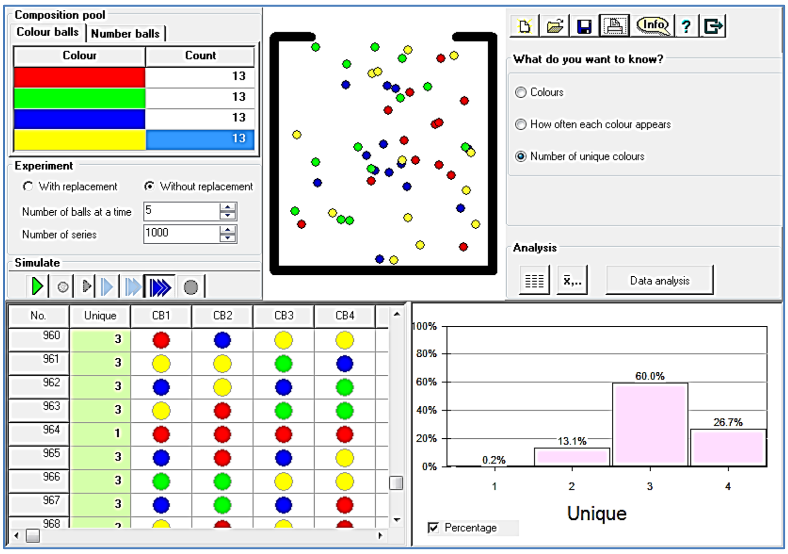
$$\frac{1}{2\,598\,960}$$

Simulations

In this module, a lot of simulations can be run. These include classical simulations, such as throwing coins and dice. There are also a lot of simulations with random numbers. In addition, we have a package of *Applied simulations*, for instance the effect of smoking on life expectancy, the effect of overselling of airline tickets on airline profit, the classical game Roulette and waiting times in a fast-food shop. There is also a module where students can create their own experimental data. Theses data can also be further analyzed.

Ball pool

The ball pool is a random generator based on the functionality of an urn model. You can choose from two options. One model shows a vase with colored balls. In the other model balls are numbered as in a lottery or bingo machine. The simulation process contains several steps.



Composition pool

Colour balls | Number balls

Colour	Count
🟥	13
🟩	13
🟦	13
🟨	13

Experiment

○ With replacement ● Without replacement

Number of balls at a time: 5

Number of series: 1000

What do you want to know?

○ Colours

○ How often each colour appears

● Number of unique colours

No.	Unique	CB1	CB2	CB3	CB4
960	3	🔴	🔵	🟡	🟡
961	3	🟡	🟡	🟢	🔵
962	3	🔵	🟡	🔵	🟢
963	3	🟡	🔴	🟢	🟢
964	1	🔴	🔴	🔴	🔴
965	3	🔵	🔴	🔵	🟡
966	3	🟢	🟢	🟡	🟡
967	3	🔵	🟢	🔵	🔴
968	2	🟡	🔴	🟡	🔴

In the simulation above we use a deck of cards where we draw five cards. How many different colors do we get? The result shows that we have "1 unique",which means five cards of the same suit, in 0.2 % of the simulations.

1. Balls in the pool

You choose a model and fill in which balls should fill the pool.

In *Color balls,* you give color and number of each color. In *Number balls,* you fill in the numbers. More times the same number means more of the same balls. You can see the balls displayed in the pool.

2. Picking balls

You fill in:

- How the balls are to be picked.
- How many balls at a time: this is actually the size of the sample. You can choose *with* or *without* replacement.
- How many times you want to run the simulation.

3. Simulation

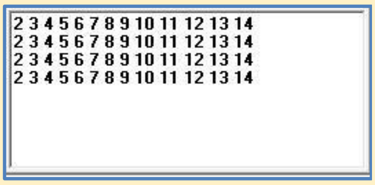
First have a small number of simulations (e.g. ten) and establish whether you really obtain what you expect.

4. What do you want to know?

Often you want to view the outcome variable regarding a certain problem. Before that, you can choose from several options. With buttons on the right below the chart, you can scroll through the variables.

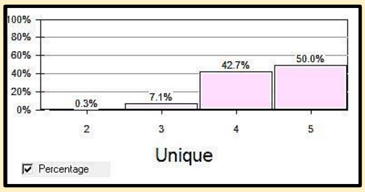
5. Analyze the data using the tools

The tools available are histogram, frequency table, and statistics. For more advanced problems you can send the dataset to the Data analysis module.



```
2 3 4 5 6 7 8 9 10 11 12 13 14
2 3 4 5 6 7 8 9 10 11 12 13 14
2 3 4 5 6 7 8 9 10 11 12 13 14
2 3 4 5 6 7 8 9 10 11 12 13 14
```

A simulation with five balls at a time from this set of numbers (ranks in a deck of cards) gives the result you can see in the bar chart.

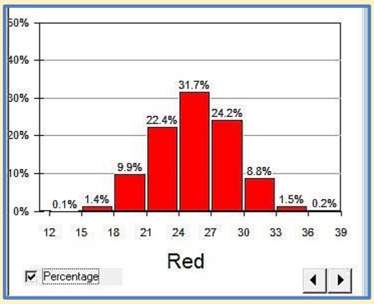


In about 43 % of the simulations we have four unique colors, which means that we have a "pair" or two cards of the same rank.

Another example:

What about the surplus of boys?

In a large hospital 200 children are born each month. In a smaller hospital it´s about 50. Assume that the probability of giving birth to a boy or girl is equally large. *In which hospital is it more likely that more than 60 % of newborns in a given month are boys*?



A simulation for the small hospital gives the result

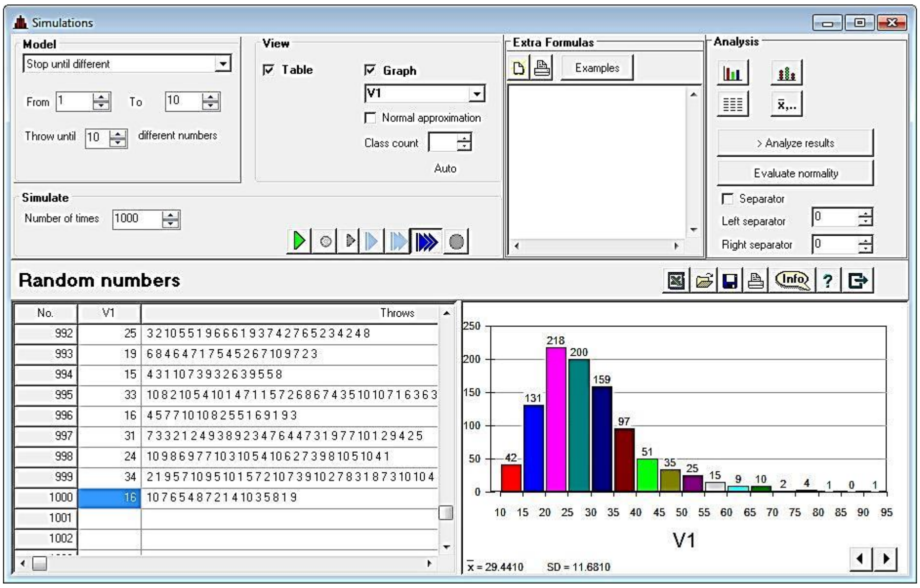
8.3 %+1.5 % + 0.2 % = 10.5 %

A simulation for the larger hospital gave the result 0.3 %.

Random numbers

In this module there are a lot of settings for simulations of random numbers. You can see the result in a table and a graph. You can also analyze the result further by pressing the graphs icons or the button “Analyze results” in the upper right corner.

Another option is to use extra formulas to create new variables.



Model: Stop until different

From 1 To 10

Throw until 10 different numbers

View: ☑ Table ☑ Graph V1

☐ Normal approximation

Class count

Auto

Simulate: Number of times 1000

Extra Formulas

Analysis

> Analyze results

Evaluate normality

☐ Separator

Left separator 0

Right separator 0

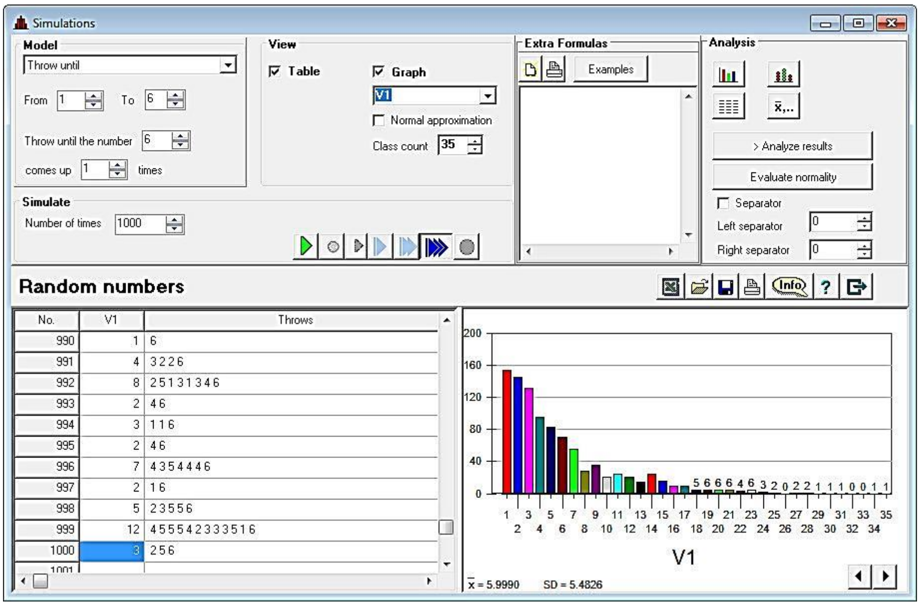
Random numbers

No.	V1	Throws
992	25	3 2 10 5 5 1 9 6 6 6 1 9 3 7 4 2 7 6 5 2 3 4 2 4 8
993	19	6 8 4 6 4 7 1 7 5 4 5 2 6 7 10 9 7 2 3
994	15	4 3 1 10 7 3 9 3 2 6 3 9 5 5 8
995	33	10 8 2 10 5 4 10 1 4 7 1 1 5 7 2 6 8 6 7 4 3 5 10 10 7 1 6 3 6 3
996	16	4 5 7 7 10 10 8 2 5 5 1 6 9 1 9 3
997	31	7 3 3 2 1 2 4 9 3 8 9 2 3 4 7 6 4 4 7 3 1 9 7 7 10 1 2 9 4 2 5
998	24	10 9 8 6 9 7 7 10 3 10 5 4 10 6 2 7 3 9 8 10 5 10 4 1
999	34	2 1 9 5 7 10 9 5 10 1 5 7 2 10 7 3 9 10 2 7 8 3 1 8 7 3 10 10 4
1000	16	10 7 6 5 4 8 7 2 1 4 10 3 5 8 1 9
1001		
1002		

$\bar{x}$ = 29.4410 SD = 11.6810

A simulation with Random numbers module and the choice ”Stop until .. different”. It illustrates the famous “Coupon collectors problem”. We simulate a sequence of numbers from 1 to 10, and we want to know how many times we have to draw until we got all 10. The simulation was done 1000 times and we can observe the result in the diagram.

Another example: You throw a dice. How many times must you on average throw until six comes up? Now, we use the model "Throw until ". We see that the mean becomes 6 times. If the question was "In which throw do you guess that a six comes up?” we would respond differently. Look at the diagram.



Model: Throw until

From 1 To 6

Throw until the number 6

comes up 1 times

View: ☑ Table ☑ Graph V1

☐ Normal approximation

Class count 35

Simulate: Number of times 1000

Extra Formulas

Analysis

> Analyze results

Evaluate normality

☐ Separator

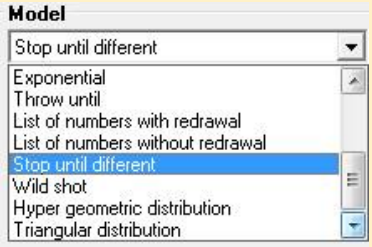
Left separator 0

Right separator 0

Random numbers

No.	V1	Throws
990	1	6
991	4	3 2 2 6
992	8	2 5 1 3 1 3 4 6
993	2	4 6
994	3	1 1 6
995	2	4 6
996	7	4 3 5 4 4 4 6
997	2	1 6
998	5	2 3 5 5 6
999	12	4 5 5 5 4 2 3 3 3 5 1 6
1000	3	2 5 6
1001		

$\bar{x}$ = 5.9990 SD = 5.4826

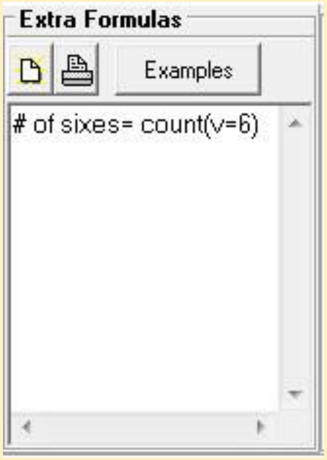


Model

Stop until different

- Exponential
- Throw until
- List of numbers with redrawal
- List of numbers without redrawal
- Stop until different
- Wild shot
- Hyper geometric distribution
- Triangular distribution

There are a lot of options to generate random numbers.

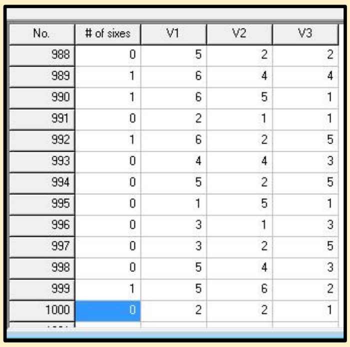


Extra Formulas

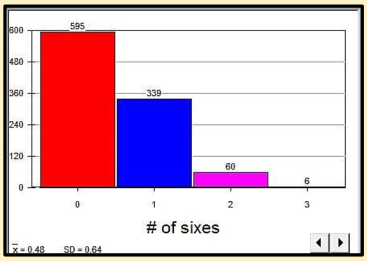
Examples

```
# of sixes= count(v=6)
```

You can also create new variables by creating extra formulas. The diagram below shows the distributions for # of sixes when you throw a dice three times and run the simulation 1000 times.



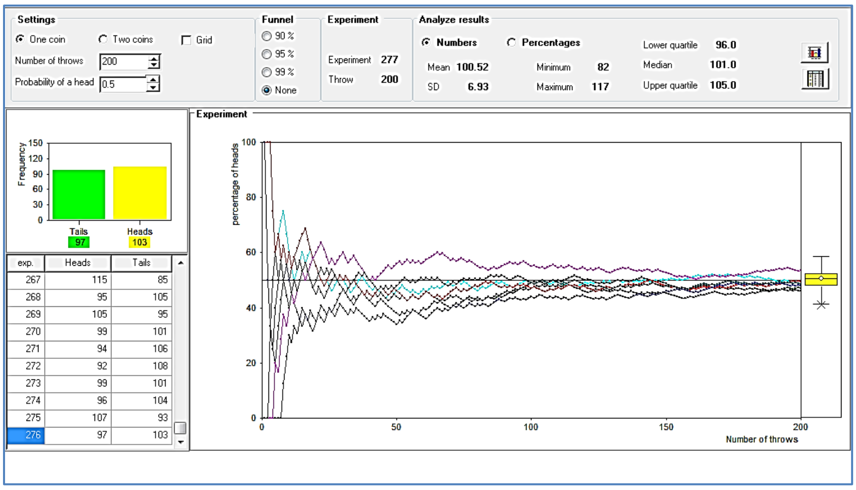
No.	# of sixes	V1	V2	V3
988	0	5	2	2
989	1	6	4	4
990	1	6	5	1
991	0	2	1	1
992	1	6	2	5
993	0	4	4	3
994	0	5	2	5
995	0	1	5	1
996	0	3	1	3
997	0	3	2	5
998	0	5	4	3
999	1	5	6	2
1000	0	2	2	1



$\bar{x}$ = 0.48 SD = 0.64

Coins

In the module Coins you can throw one or two coins.With one coin you can see how the relative frequency stabilizes after many throws. When you have two coins you will see the frequencies for zero, one or two heads.



Settings
● One coin ○ Two coins □ Grid
Number of throws 200
Probability of a head 0.5

Funnel
○ 90 %
○ 95 %
○ 99 %
● None

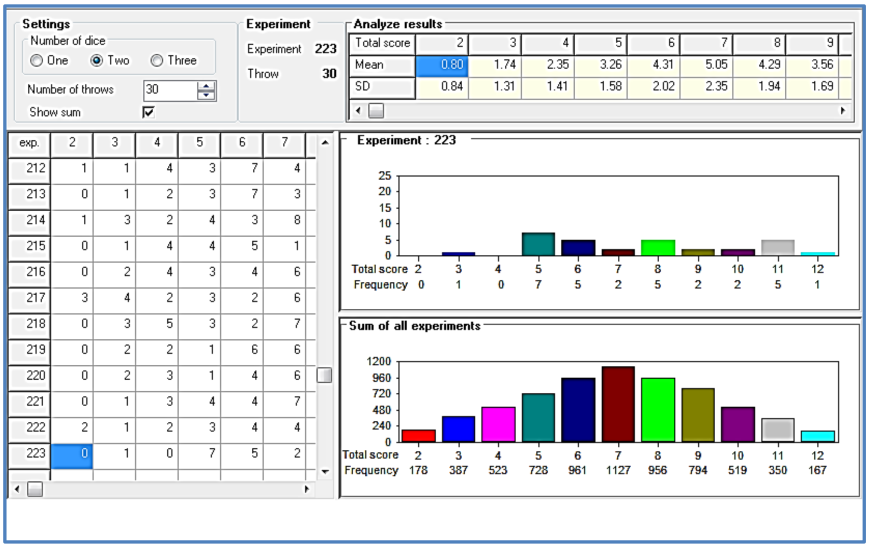
Experiment
Experiment 277
Throw 200

Analyze results
● Numbers ○ Percentages
Mean 100.52
SD 6.93
Minimum 82
Maximum 117
Lower quartile 96.0
Median 101.0
Upper quartile 105.0

exp.	Heads	Tails
267	115	85
268	95	105
269	105	95
270	99	101
271	94	106
272	92	108
273	99	101
274	96	104
275	107	93
276	97	103

Dice

This simulation consists of a number of throws with one, two or three dice. If you tick the box "Show sum of all experiments" you also get a graph which shows a summary for all performed experiments. We can see that "7" has the highest frequency.



Settings
Number of dice
○ One ● Two ○ Three
Number of throws 30
Show sum ☑

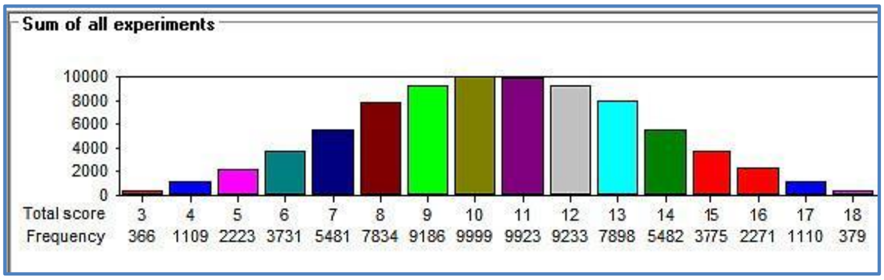
Experiment
Experiment 223
Throw 30

Analyze results

Total score	2	3	4	5	6	7	8	9
Mean	0.80	1.74	2.35	3.26	4.31	5.05	4.29	3.56
SD	0.84	1.31	1.41	1.58	2.02	2.35	1.94	1.69

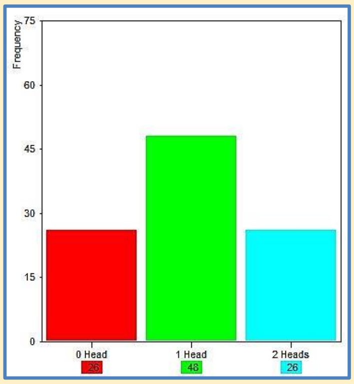
exp.	2	3	4	5	6	7
212	1	1	4	3	7	4
213	0	1	2	3	7	3
214	1	3	2	4	3	8
215	0	1	4	4	5	1
216	0	2	4	3	4	6
217	3	4	2	3	2	6
218	0	3	5	3	2	7
219	0	2	2	1	6	6
220	0	2	3	1	4	6
221	0	1	3	4	4	7
222	2	1	2	3	4	4
223	0	1	0	7	5	2

Experiment : 223
25
20
15
10
5
0
Total score 2 3 4 5 6 7 8 9 10 11 12
Frequency 0 1 0 7 5 2 5 2 2 5 1
Sum of all experiments
1200
960
720
480
240
0
Total score 2 3 4 5 6 7 8 9 10 11 12
Frequency 178 387 523 728 961 1127 956 794 519 350 167

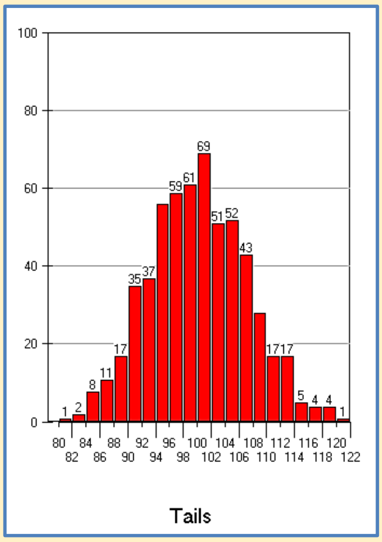


Summary for a simulation with three dice. 1000 dice were thrown in each experiment and the simulation was performed 80 times. In the table to the right you can see the number of outcomes for different sums. Observe the symmetry in both the diagram and the table.

Coins



Simulation of 100 throws with two coins . The bars are frequencies for zero, one and two heads.



Here is a histogram for the distribution of number of tails for 200 throws with one coin.

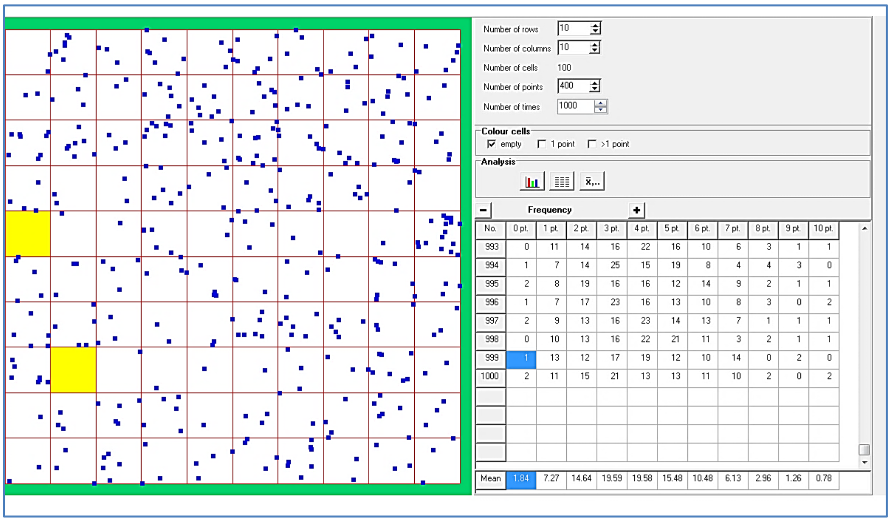
# of dots	# of outcomes
3	1
4	3
5	6
6	10
7	15
8	21
9	25
10	27
11	27
12	25
13	21
14	15
15	10
16	6
17	3
18	1

Raindrops

When it´s raining, the drops hit the ground evenly or unevenly distributed. Rain can be regarded as a random process. You can simulate the process with a surface where rain falls completely randomly. Because the drops fall randomly, they sometimes fall "on each other".

A good application of this process is as folllows: "A baker bakes a large dough which contains 400 raisins. From the dough she then bakes 100 buns. How great is the probability that you pick a bun with no raisins at all? "

This simulation is shown in the screenshot below. The grid has 10 columns and 10 rows, e.g. 100 cells. We repeat the simulation 1000 times.
The mean after 1000 simulations is 1.83. The probability that you will not get any raisins in your bun then becomes 1.84/100 = 0.0184 or about 2 %. The probability that you will get exactly 4 raisins is 0.196. The expected value is of course 4.



Number of rows: 10
Number of columns: 10
Number of cells: 100
Number of points: 400
Number of times: 1000

Colour cells: ☑ empty ☐ 1 point ☐ >1 point

Analysis

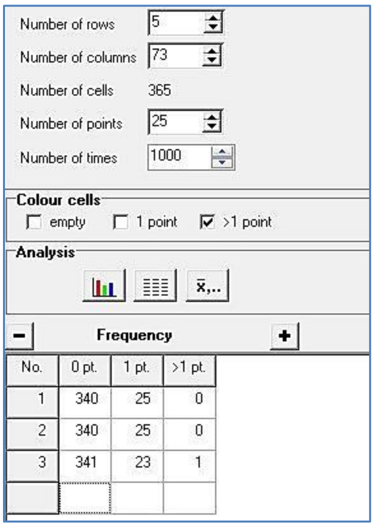
Frequency

No.	0 pt.	1 pt.	2 pt.	3 pt.	4 pt.	5 pt.	6 pt.	7 pt.	8 pt.	9 pt.	10 pt.
993	0	11	14	16	22	16	10	6	3	1	1
994	1	7	14	25	15	19	8	4	4	3	0
995	2	8	19	16	16	12	14	9	2	1	1
996	1	7	17	23	16	13	10	8	3	0	2
997	2	9	13	16	23	14	13	7	1	1	1
998	0	10	13	16	22	21	11	3	2	1	1
999	1	13	12	17	19	12	10	14	0	2	0
1000	2	11	15	21	13	13	11	10	2	0	2
Mean	1.84	7.27	14.64	19.59	19.58	15.48	10.48	6.13	2.96	1.26	0.78

We also show a simulation of the famous *birthday paradox*. The grid is divided into **365** cells (73 × 5). Then let **25** drops fall every time we simulate. We want to do a simulation which calculates the probability that *two or more students* have the same birthday in a group of about 25 students. We repeat this **1000** times.

In the screen to the right we have all the settings. We show the columns 0 pt, 1 pt and > 1pt. It is the *last column* we are interested in here. For the analysis we press the button for the bar chart. We are interested in the variable *> 1pt*. From the diagram, we can see that in 58.6 % of the simulations we get two hits or more. The simulation calculates that the probability that two or more have the same birthday in a class of 25 students is approximately 58.6 %. The theoretical probability is

$$1-\frac{364}{365}\cdot\frac{363}{365}\cdot\frac{362}{365}\cdot\ldots\cdot\frac{342}{365}\cdot\frac{341}{365}\approx 0.57$$



Number of rows: 5
Number of columns: 73
Number of cells: 365
Number of points: 25
Number of times: 1000

Colour cells: ☐ empty ☐ 1 point ☑ >1 point

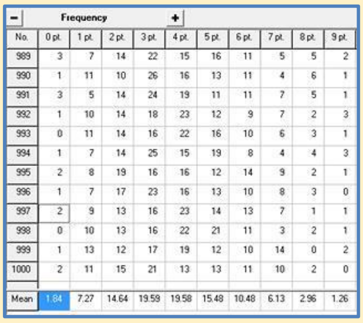
Analysis

Frequency

No.	0 pt.	1 pt.	>1 pt.
1	340	25	0
2	340	25	0
3	341	23	1

The settings for the birthday paradox

Explanation of the table
The numbers in each column show, for each experiment, in how many cells we have 0 drops, 1 drop, 2 drops etc. You can reduce or increase the view for the number of columns with the plus and minus button.
The bottom of the table contains average values for each outcome.

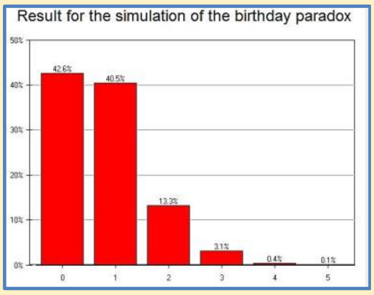


Frequency

No.	0 pt.	1 pt.	2 pt.	3 pt.	4 pt.	5 pt.	6 pt.	7 pt.	8 pt.	9 pt.
989	3	7	14	22	15	16	11	5	5	2
990	1	11	10	26	16	13	11	4	6	1
991	3	5	14	24	19	11	11	7	5	1
992	1	10	14	18	23	12	9	7	2	3
993	0	11	14	16	22	16	10	6	3	1
994	1	7	14	25	15	19	8	4	4	3
995	2	8	19	16	16	12	14	9	2	1
996	1	7	17	23	16	13	10	8	3	0
997	2	9	13	16	23	14	13	7	1	1
998	0	10	13	16	22	21	11	3	2	1
999	1	13	12	17	19	12	10	14	0	2
1000	2	11	15	21	13	13	11	10	2	0
Mean	1.84	7.27	14.64	19.59	19.58	15.48	10.48	6.13	2.96	1.26

0 pt.	Freq.
0	145
1	303
2	270
3	170
4	81
5	24
6	6
7	1
Total	1000

You can analyze all your data in a bar chart, frequency table or as summary statistics. Above is a table for the variable “*0 pt*”.



In the bar chart can we see that in 57.64 % of the simulations we have ***two or more*** persons who share the same birthday.

Applied Simulations

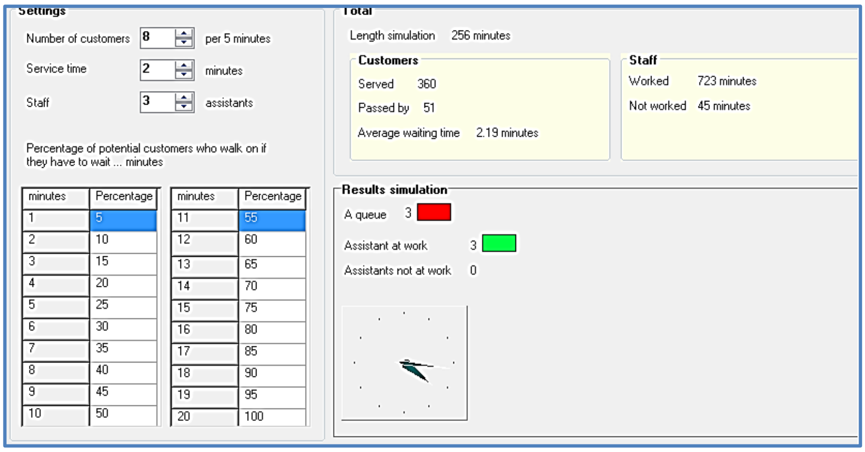
This module contains a number of simulations of specific situations which occur in practice and in games. Examples are for instance the effect of smoking on life expectancy, the effect of overselling of airline tickets on airline profit, the game Roulette and waiting times in a fast-food shop. Here are two examples of these simulations.

Example Fast-food shop

Everyone has experienced having to wait in a queue at a checkout or at a booking office. Queuing is a problem that has to do with capacity (e.g. the number of assistants) and costs. If the capacity is large enough, nobody has to wait. But this is an expensive solution. If the capacity is too low, potential customers walk away, and this costs money too.

This simulation gives you the opportunity to study this problem. In this simulation customers arrive and are served inside by the staff, or they have to wait. If customers have to wait too long, some will walk away.

There are several variables that can be chosen. In the simulation you can see how many customers are served and for how much of the time the staff is at work.



Settings

Number of customers 8 per 5 minutes

Service time 2 minutes

Staff 3 assistants

Percentage of potential customers who walk on if they have to wait ... minutes

minutes	Percentage	minutes	Percentage
1	5	11	55
2	10	12	60
3	15	13	65
4	20	14	70
5	25	15	75
6	30	16	80
7	35	17	85
8	40	18	90
9	45	19	95
10	50	20	100

Total

Length simulation 256 minutes

Customers

Served 360

Passed by 51

Average waiting time 2.19 minutes

Staff

Worked 723 minutes

Not worked 45 minutes

Results simulation

A queue 3 🟥

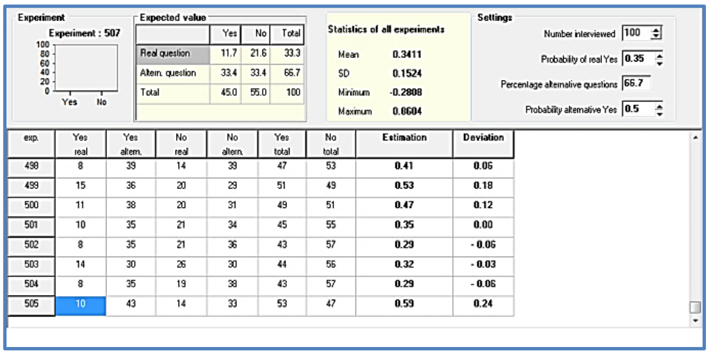
Assistant at work 3 🟩

Assistants not at work 0

Example Randomized response

Randomized response is a technique which tries to obtain answers to sensitive questions like "*Have you ever taken illegal drugs*". The simulation shows how this technique works for different values of the parameters. The parameters are:

- The probability of "yes" with the real question; in practice this is the value you are looking for.
- The number of persons interviewed
- The percentage of people answering the alternative question
- The probability of "yes" for the alternative question



Experiment

Experiment : 507

Expected value

	Yes	No	Total
Real question	11.7	21.6	33.3
Altern. question	33.4	33.4	66.7
Total	45.0	55.0	100

Statistics of all experiments

Mean 0.3411

SD 0.1524

Minimum -0.2808

Maximum 0.8604

Settings

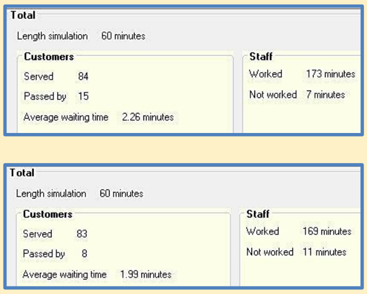
Number interviewed 100

Probability of real Yes 0.35

Percentage alternalive questions 66.7

Probability alternative Yes 0.5

exp.	Yes real	Yes altern.	No real	No altern.	Yes total	No total	Estimation	Deviation
498	8	39	14	39	47	53	0.41	0.06
499	15	36	20	29	51	49	0.53	0.18
500	11	38	20	31	49	51	0.47	0.12
501	10	35	21	34	45	55	0.35	0.00
502	8	35	21	36	43	57	0.29	- 0.06
503	14	30	26	30	44	56	0.32	- 0.03
504	8	35	19	38	43	57	0.29	- 0.06
505	10	43	14	33	53	47	0.59	0.24



Total

Length simulation 60 minutes

Customers

Served 84

Passed by 15

Average waiting time 2.26 minutes

Staff

Worked 173 minutes

Not worked 7 minutes

Total

Length simulation 60 minutes

Customers

Served 83

Passed by 8

Average waiting time 1.99 minutes

Staff

Worked 169 minutes

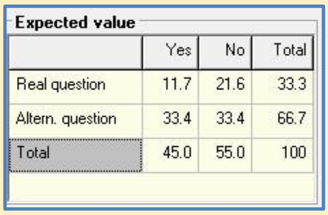
Not worked 11 minutes

Simulations with the same settings give different results.

How Randomized response works:
You get a bag with three cards. Then you randomly draw a card, which only you can see, and respond in accordance with what you can read on the card:

First card: "say YES"
Second card: "say NO"
Third card: "Do you use illegal drugs?"

On average,If you are honest, one-third will answer YES, one-third will answer NO and one-third will respond if they use illegal drugs.



Expected value

	Yes	No	Total
Real question	11.7	21.6	33.3
Altern. question	33.4	33.4	66.7
Total	45.0	55.0	100

In the table for Expected values, you can see that with 100 persons in the simulation 45 percent said YES, and 55 percent said NO. The figures tell you what percentage of persons have taken illegal drugs!

Samples

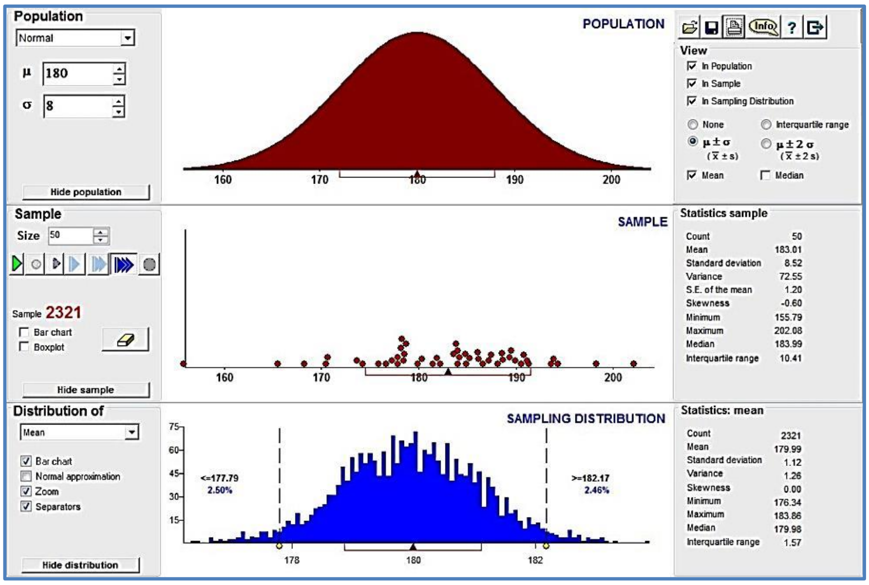
Sampling is to gain information about the whole by examining a part. That is why sampling is a basic issue for statistics. We need to understand the process of drawing samples to know what you can do with the results. Important concepts include variability, uncertainty, sampling distribution, randomization, confidence, square root *n* law, central limit theorem .. Statistics is not mathematics.
Simulations are excellent to visualize the concepts and so lead to understanding. This will enhance statistical reasoning with insight.

Sampling distribution

The idea with this module is that students should get some insight in the the sample process and then be able to interpret the result you get from the sample distribution. You have a lot of settings for the population, the sample and for the sample distribution and can view the result numerically and graphically in different ways.

If you want to get information about a large unknown population, you need to draw one or more samples. The data you collect can be used to make estimates of mean and median and the distribution of the population.

Of course, these estimates are uncertain due to random fluctuations in the data collected. This uncertainty will be less if you draw a large sample, but in practice such sampling is often too costly. *The idea behind this module is to gain insight into the sampling process and be able to interpret the results you get from the sampling distribution.*



Population: Normal
μ 180
σ 8

View
☑ In Population
☑ In Sample
☑ In Sampling Distribution
○ None ○ Interquartile range
● μ ± σ ($\bar{x} \pm s$) ○ μ ± 2 σ ($\bar{x} \pm 2 s$)
☑ Mean □ Median

Sample
Size 50
Sample 2321
□ Bar chart
□ Boxplot

Statistics sample

Count	50
Mean	183.01
Standard deviation	8.52
Variance	72.55
S.E. of the mean	1.20
Skewness	-0.60
Minimum	155.79
Maximum	202.08
Median	183.99
Interquartile range	10.41

Distribution of: Mean
☑ Bar chart
□ Normal approximation
☑ Zoom
☑ Separators

Statistics: mean

Count	2321
Mean	179.99
Standard deviation	1.12
Variance	1.26
Skewness	0.00
Minimum	176.34
Maximum	183.86
Median	179.98
Interquartile range	1.57

Sampling from a normal distribution with zooming. Sample size is 50.

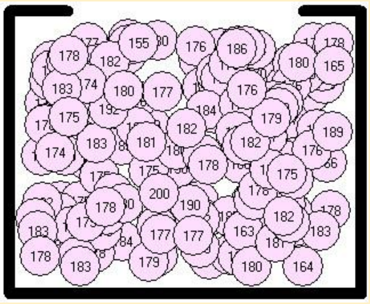
One can show that in about 19 cases out of 20, e.g. 95 % of all samples, the sample mean will be in the range

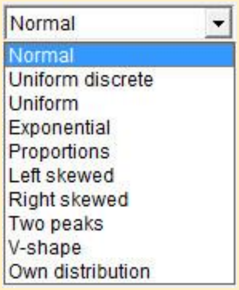
$$\bar{x} \pm 2 \cdot \frac{s}{\sqrt{n}}$$

where $\bar{x}$ is the mean of the sampling distribution, *s* is the standard deviation and *n* is the sample size. In this example we have the interval

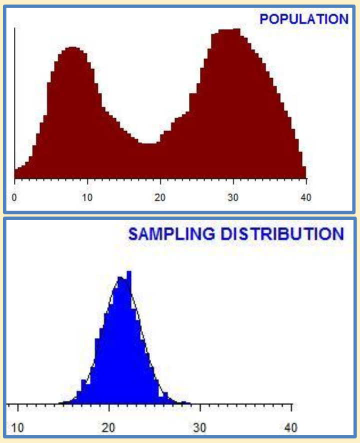
$$180.0 \pm 2 \cdot \frac{8}{\sqrt{50}} \approx 180.0 \pm 2.3$$

We can check this in the histogram above. About 95 % of the total area of the blue bars is in the range 177.7 - 182.3. We call this a confidence interval at 95 % confidence level.

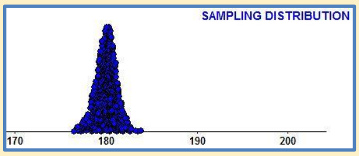




For the population you can choose different models. You can also make your own distribution with the help of the mouse.



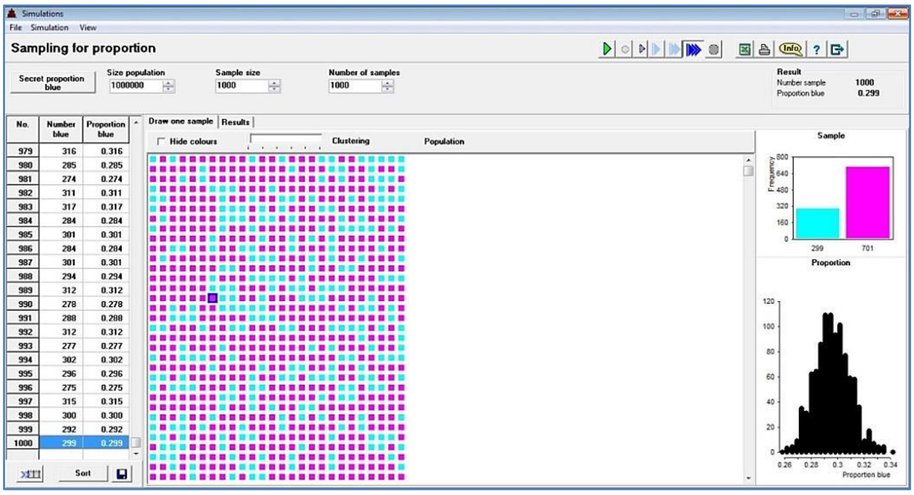
It is surprising to see is that the sampling distribution of the mean is more or less a normal distribution, whatever the population distribution may be.



We can see without zooming that the spread is significantly less.

Sampling for proportion

It surprises many that rather small samples can give a good picture of a population. Think of polls. Facing parliamentary elections polling institutes usually ask about 1000 people which party they would vote for if the election were today. People say that they don't believe in polls because they have never been asked about their opinion! In this simulation, you can study the relationship between *sample size*, *population size* and *population characteristics*.



Simulations

File Simulation View

Sampling for proportion

Secret proportion blue

Size population: 1000000

Sample size: 1000

Number of samples: 1000

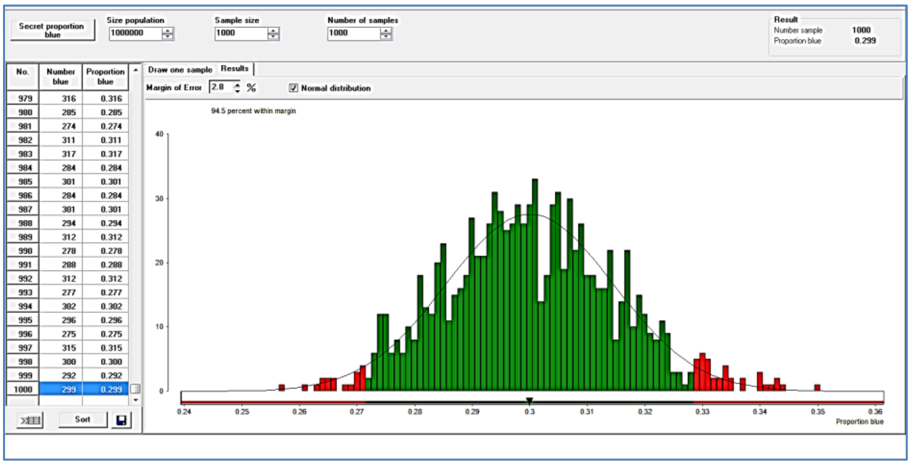
Result
Number sample 1000
Proportion blue 0.299

No.	Number blue	Proportion blue
979	316	0.316
980	285	0.285
981	274	0.274
982	311	0.311
983	317	0.317
984	284	0.284
985	301	0.301
986	284	0.284
987	301	0.301
988	294	0.294
989	312	0.312
990	278	0.278
991	288	0.288
992	312	0.312
993	277	0.277
994	302	0.302
995	296	0.296
996	275	0.275
997	315	0.315
998	300	0.300
999	292	0.292
1000	299	0.299

The population consists of red and blue squares. They represent the proponents and the opponents. Population characteristics of the colours are displayed. Before the sample is drawn, the proportion in the "Secret proportion blue" is entered. In reality, this is an *unknown parameter* of the population and it is the one to be estimated.

You can now enter the size of the population and the sample. Because you can draw many samples, you can then make assessments of the reliability of the estimate of the proportion.

The result of all trials are displayed on the left. The result is also displayed with a bar chart for each sample. You can also see a preview of the distribution of the proportion in each sample.



Secret proportion blue

Size population: 1000000

Sample size: 1000

Number of samples: 1000

Result
Number sample 1000
Proportion blue 0.299

Margin of Error 2.8 %

☑ Normal distribution

94.5 percent within margin

No.	Number blue	Proportion blue
979	316	0.316
980	285	0.285
981	274	0.274
982	311	0.311
983	317	0.317
984	284	0.284
985	301	0.301
986	284	0.284
987	301	0.301
988	294	0.294
989	312	0.312
990	278	0.278
991	288	0.288
992	312	0.312
993	277	0.277
994	302	0.302
995	296	0.296
996	275	0.275
997	315	0.315
998	300	0.300
999	292	0.292
1000	299	0.299

If you press the button **Results** you see the distribution of the proportions. This is called the *Sampling distribution*. If you change *Margin of error* the diagram is updated and you can see how many are within the margin by the green bars in the diagram. This value is also shown as text in the upper left corner.

Survey of a sample of 1000 people. The question is what you can expect as to how close the proportion of the sample will come to the proportion in the entire population. One can show that if you take many samples the distribution will become roughly normally distributed with a mean of *p* and a standard deviation

$$\sqrt{\frac{p(1-p)}{1000}}.$$

This means approximately 95 % probability of getting the sample proportion in the range

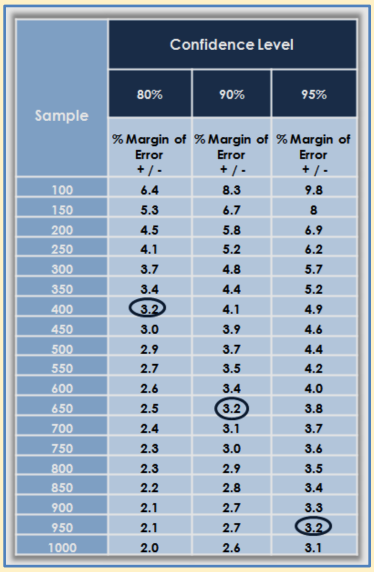
$$p \pm 2 \cdot \sqrt{\frac{p(1-p)}{1000}}$$

and in our example

$$0.30 \pm 2 \cdot \sqrt{\frac{0.30(1-0.30)}{1000}} \approx$$

$$\approx 0.30 \pm 0.03$$

Therefore, if the real proportion in the population is 30 %, it is only one time in twenty that a sample would give a percentage that deviates more than 3 percentage points from the true value. This is known as the *margin of error*.



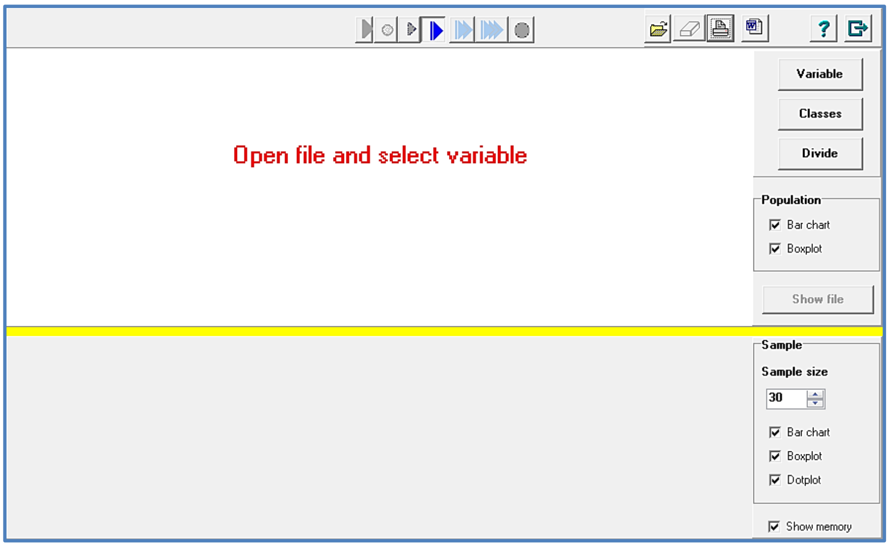
Sample	Confidence Level		
	80%	90%	95%
	% Margin of Error + / -	% Margin of Error + / -	% Margin of Error + / -
100	6.4	8.3	9.8
150	5.3	6.7	8
200	4.5	5.8	6.9
250	4.1	5.2	6.2
300	3.7	4.8	5.7
350	3.4	4.4	5.2
400	3.2	4.1	4.9
450	3.0	3.9	4.6
500	2.9	3.7	4.4
550	2.7	3.5	4.2
600	2.6	3.4	4.0
650	2.5	3.2	3.8
700	2.4	3.1	3.7
750	2.3	3.0	3.6
800	2.3	2.9	3.5
850	2.2	2.8	3.4
900	2.1	2.7	3.3
950	2.1	2.7	3.2
1000	2.0	2.6	3.1

Dancing samples

The purpose of this module is to demonstrate the influence of the size of the sample on the sampling results. To see the effect of this module, it makes sense to use a large file. The file in the example was created for a nationwide survey of students, with 50 000 participants. In this module it is possible to draw many samples.

In the top half of the screen is the population. The bottom half shows the samples drawn. The results are presented using a box plot, histogram or dotplot. If memory is turned on, boxplot and histogram show the results of the last 30 sample drawings. It is clear that with a small sample of 30 individuals the variation is very large. With a sample size of 300 there is much less variation.

The idea of this module comes from Journal of the Royal Statistical Society:Towards more accessible conceptions of statistical inference. Authors: C. J. Wild, M. Pfannkuch, M. Regan, N. J. Horton.

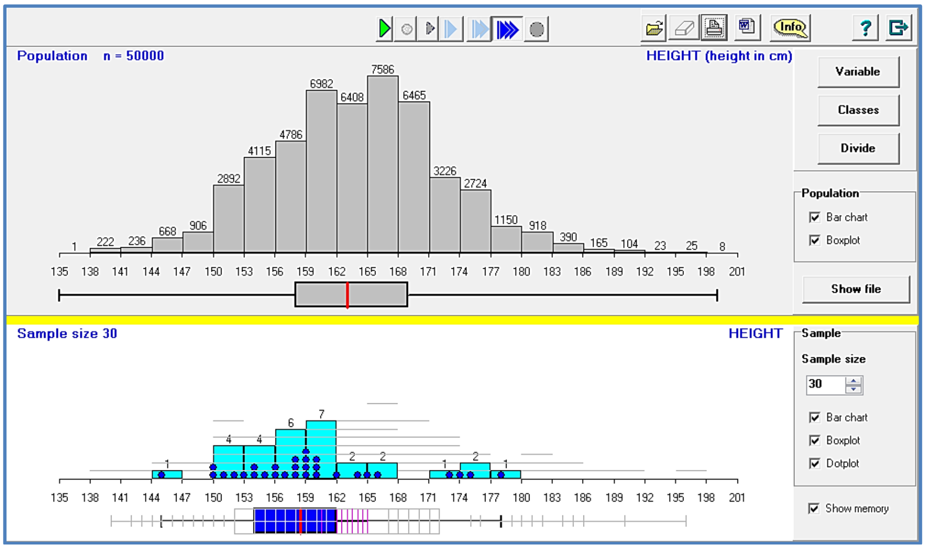


Open file and select variable

Population
☑ Bar chart
☑ Boxplot

Sample size 30
☑ Bar chart
☑ Boxplot
☑ Dotplot
☑ Show memory

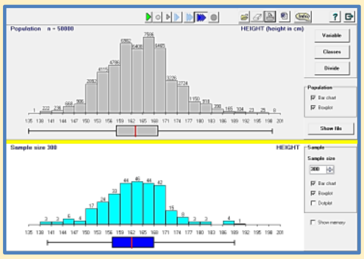
You start by opening a file. Then you select a variable which can be divided. You can also set the number of classes in the histogram.



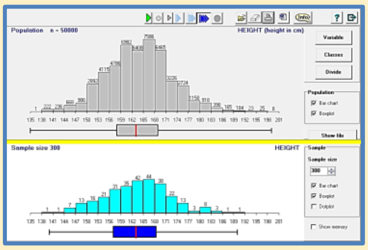
Population
☑ Bar chart
☑ Boxplot

Sample
Sample size 30
☑ Bar chart
☑ Boxplot
☑ Dotplot
☑ Show memory

Sample size here is 30. Large variation between samples.

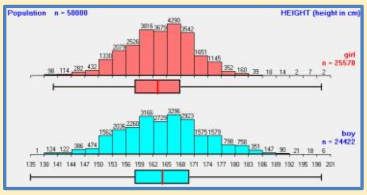


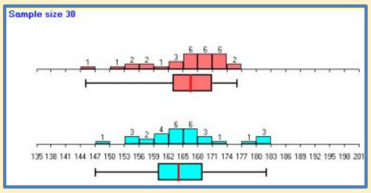
sample 1



sample 2

With larger samples the variation is much less.





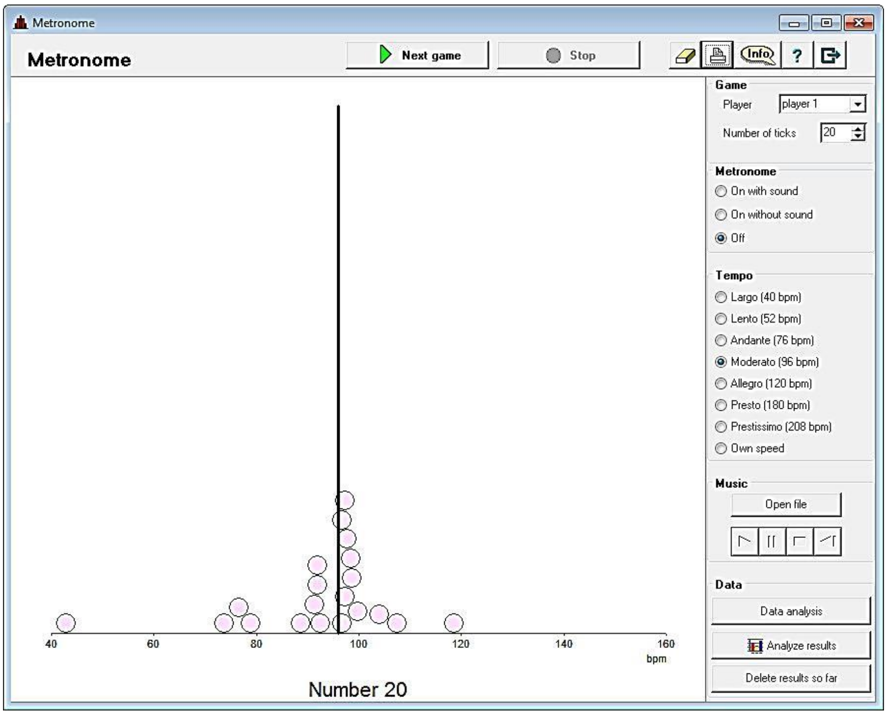
Data for height could also be divided to a variable. The division variable here is gender.

Make data

VUstat has three modules with which one can create one´s own data from *experiments*.

- In **Reaction time** we measure your response speed at the keyboard. Data are number of milliseconds which can be analysed directly or in Data analysis. Several players can compete against each other.
- Every human being shows unconsciously regularities in his/her behavior. If these regularities can be found, then you can predict his or her behavior based on this knowledge. This is the case in module **Artificial intelligence** when typing zero and one on a computer. In this simulation the computer tries to predict your behavior from the zeroes and ones you type.
- **Metronome:** a player needs to maintain a certain metronome tempo as far as possible and does this by holding the tempo by tapping the space bar. After each tap, BPM (beats per minute) is calculated and displayed as a dot. The screen displays a dotplot of the scores. Several players can compete against each other and data can be analyzed in different ways.

Metronome



Game
Player: player 1
Number of ticks: 20

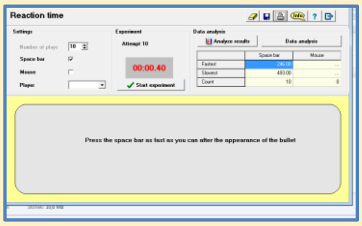
Metronome
○ On with sound
○ On without sound
● Off

Tempo
○ Largo (40 bpm)
○ Lento (52 bpm)
○ Andante (76 bpm)
● Moderato (96 bpm)
○ Allegro (120 bpm)
○ Presto (180 bpm)
○ Prestissimo (208 bpm)
○ Own speed

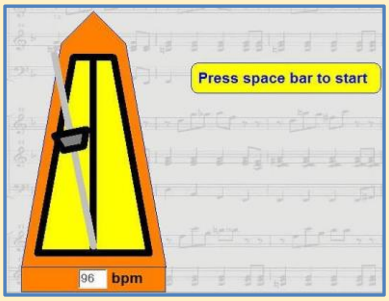
Music

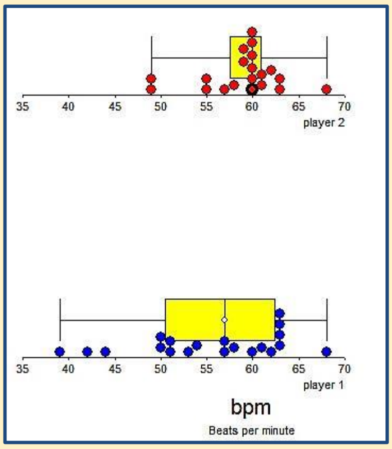
Data

For each game round, you should turn on the metronome with sound, so that new players can listen in and learn the tempo.There are 8 different tempos.



Reaction time. Here we test the response speed at the keyboard. There are three settings for the experiment. You use just the spacebar (easy), just the mouse (a little more difficult) or a mixture (very difficult).





Player	player 2	player 1	
bpm	Freq.	Freq.	Total
39 - 41	0	1	1
42 - 44	0	2	2
45 - 47	0	0	0
48 - 50	2	2	4
51 - 53	0	3	3
54 - 56	2	1	3
57 - 59	4	3	7
60 - 62	9	3	12
63 - 65	2	4	6
66 - 68	1	1	2
Total	**20**	**20**	**40**

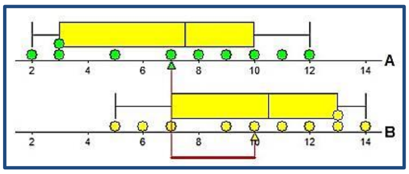
You can compare the results for two or more players in different ways.

Testing hypotheses

The classic way of introducing statistical inference runs mostly through techniques of hypothesis testing. The level of mathematics and abstract language involved makes this approach difficult for students. With simulations, an intuitive inference approach is possible, which focuses on basic concepts such as sampling and chance. The *randomization test* is very suitable for such an intuitive introduction into statistical inference. The classic ways of hypotheses testing can be also be demonstrated through visualization in modules such as binomial test, *z*-test and *t*-test.

Randomization test - Which battery brand is best?

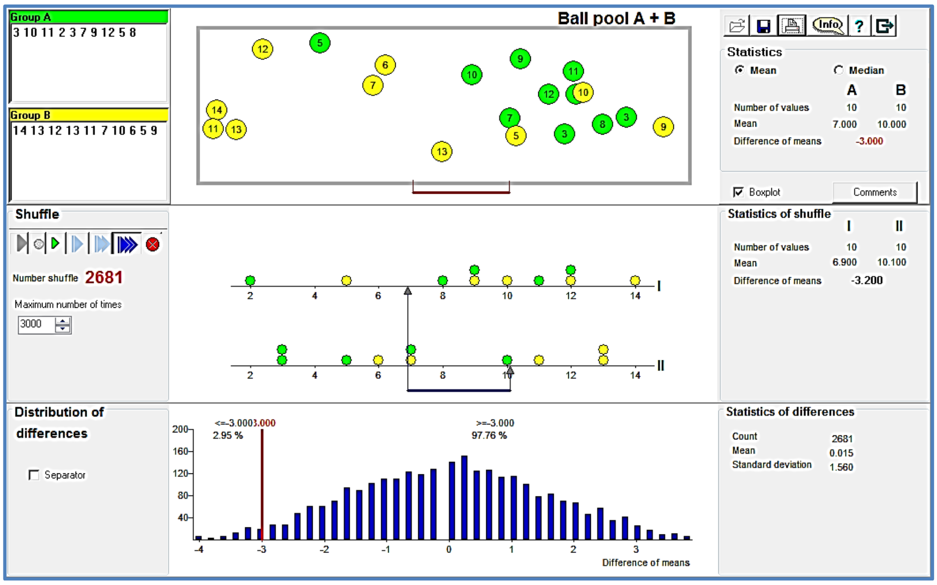
For two brands of batteries, A and B, 10 lifetimes are listed. Though the mean of sample B is 3 hours ahead of the mean of brand A, the dotplots of the two samples show too much overlap and contain too few data to provide a basis for any conclusion that brand B is better.



Here is the distribution of the samples for Brand A and Brand B

We put both samples in a ball pool of A+B.
A sample of 10 from this pool results in a remainder of 10. These two samples I and II have also a difference between the means. Since the data values were randomly divided into two groups this difference between the means can be attributed to chance.This procedure can be repeated a lot of times with test statistic: the difference between the means. The observed difference is 3 hours in favor of brand B. *The question is whether this difference is significant.*

The distribution of differences indicates that about 3 % of the simulations show a greater difference than the observed difference. Then the conclusion is that brand B is significantly better than brand A.
Although the mean is most commonly used it can be unstable because of the small dataset. The median is more stable. However, when the median is the key, the difference is not significant. This can be seen easily by switching from mean to median.



Group A
3 10 11 2 3 7 9 12 5 8

Group B
14 13 12 13 11 7 10 6 5 9

Statistics
● Mean ○ Median

	A	B
Number of values	10	10
Mean	7.000	10.000
Difference of means	-3.000	

☑ Boxplot
Comments

Number shuffle 2681
Maximum number of times 3000

Statistics of shuffle

	I	II
Number of values	10	10
Mean	6.900	10.100
Difference of means	-3.200	

Distribution of differences
☐ Separator

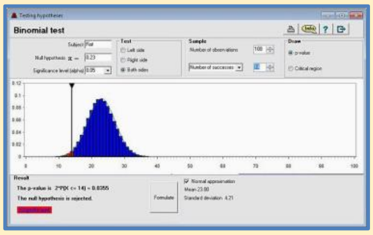
<=-3.0003.000
2.95 %
>=-3.000
97.76 %

Statistics of differences

Count	2681
Mean	0.015
Standard deviation	1.560

Available modules:
Alternative test
Binomial test
z-test
t-test
Randomization test
Confidence intervals

Example of Binomial test
A Fiat car dealer claims that Fiat´s share of the car market is about 23 %. A consumer organization is to consider whether this claim is correct. A random sample of 100 car owners are asked to mark the car they drive. How many Fiat owners will be needed to prove that the dealer is wrong at a significance level of 5 %?



Binomial test

Subject Fiat
Null hypothesis π = 0.23
Significance level (alpha) 0.05

Test
○ Left side
○ Right side
● Both sides

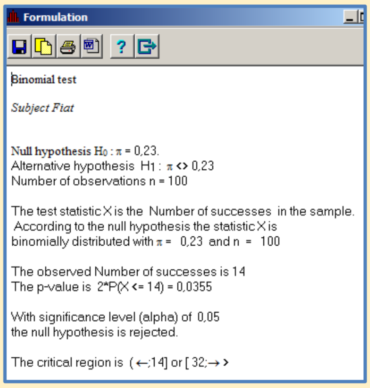
Sample
Number of observations 100
Number of successes 14

Draw
● p-value
○ Critical region

Result
The p-value is 2*P(X <= 14) = 0.0355
The null hypothesis is rejected.

Formulate

☑ Normal approximation
Mean 23.00
Standard deviation 4.21



Formulation

Binomial test

Subject Fiat

Null hypothesis H_0 : π = 0,23.
Alternative hypothesis H_1 : π <> 0,23
Number of observations n = 100

The test statistic X is the Number of successes in the sample.
According to the null hypothesis the statistic X is
binomially distributed with π = 0,23 and n = 100

The observed Number of successes is 14
The p-value is 2*P(X <= 14) = 0,0355

With significance level (alpha) of 0,05
the null hypothesis is rejected.

The critical region is (←;14] or [32;→ >

You can display a full list of test data with comments, copy it and paste into other applications.