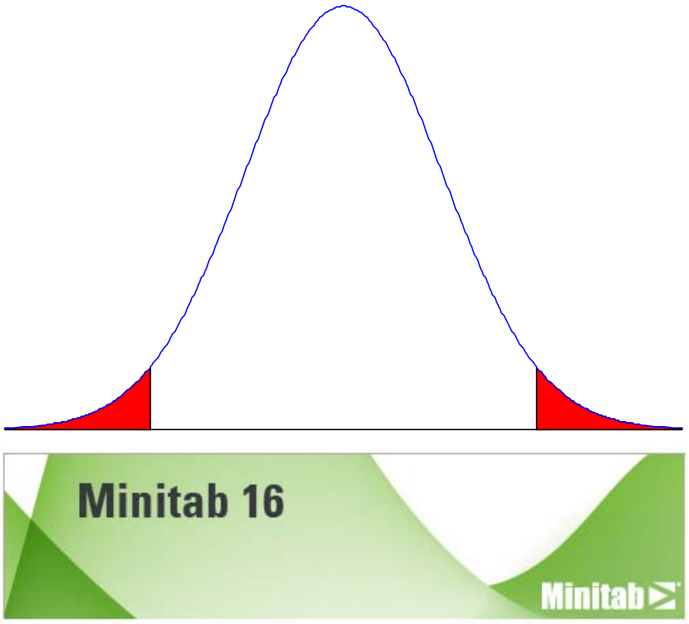
MINITAB, a primer

release 16

Jan-Eric Englund



SLU Alnarp
Område Agrosystem

Swedish University of Agricultural Sciences
Department of Agrosystems

Kompendium 2011
Course notes

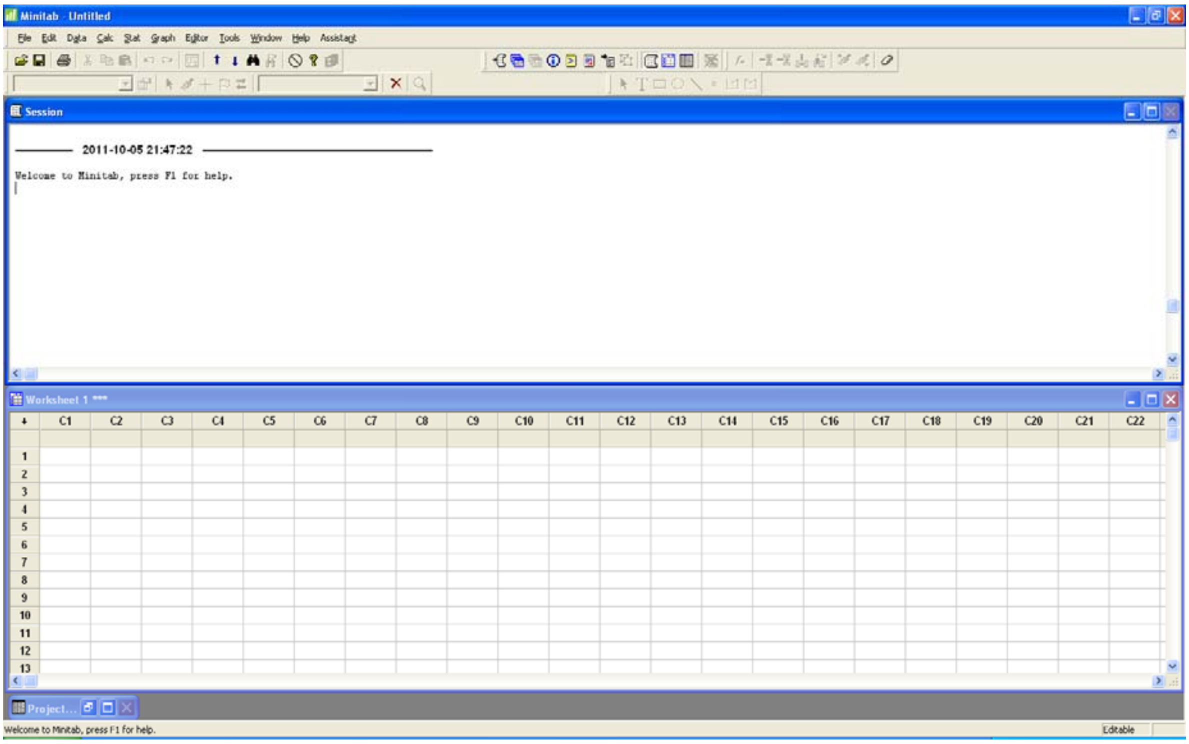
Jan-Eric Englund is lecturer in statistics at the Division of Statistics in Alnarp.

Minitab, a primer, describes the statistical methods used in a basic course in statistics, and is not at all an ambition to be a complete manual for Minitab

SLU
P.O. Box 104
SE-230 53 ALNARP
SWEDEN
Phone: +46 40 415000 (operator)

1. Introduction

When you have opened Minitab, there are two windows, the one above is the Session Window and the one below is the Worksheet. The Worksheet is in some sense similar to the sheets in Excel, but there are important differences.



Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Session
Worksheet 1 ***
C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11 C12 C13 C14 C15 C16 C17 C18 C19 C20 C21 C22
1 2 3 4 5 6 7 8 9 10 11 12 13
Project...
Editable

```
———— 2011-10-05 21:47:22 ————————————

Welcome to Minitab, press F1 for help.
```

Welcome to Minitab, press F1 for help.

When you analyse your data in Minitab, you start by reading your data, choose a menu to decide what to do, and then you get the output in the Session Window. If you have asked the program to do a graph, this is produced in a new window.

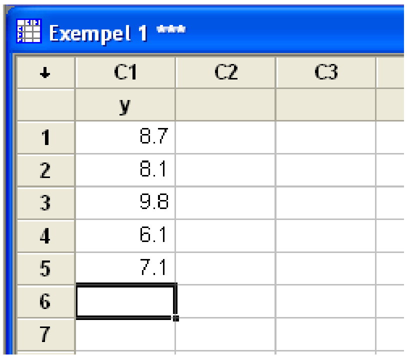
First check whether your version of Minitab uses decimal point or decimal comma, if you have wrong here the title of the column changes into `C1-T`, indicating that it is a column with text, and this type of columns cannot be used for numerical calculations[1]. In the following we assume that the program uses a decimal point.

Example 1. Introductory example

Calculate the mean and the standard deviation for the following dataset: 8.7, 8.1, 9.8, 6.1, 7.1. (In the examples it is assumed that Minitab uses decimal point and not decimal comma.)

Read the values into column `C1` as (the name of the variable is here *y*)

[1] If you by mistake have got a numerical column to be a text column you can change it with the command Data → Code → Text to Numeric…

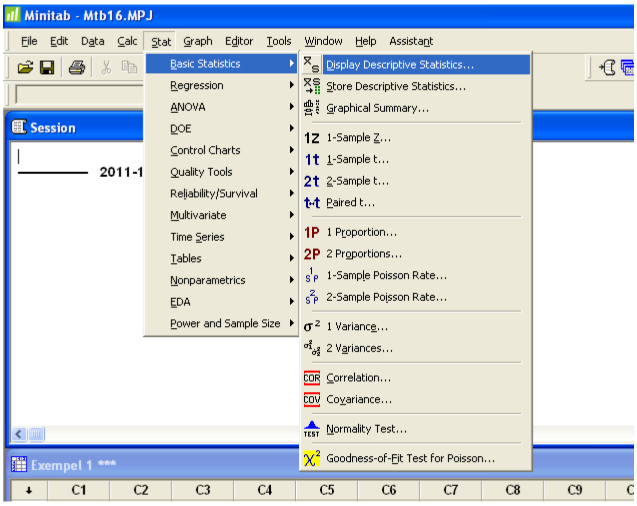


Exempel 1 ***

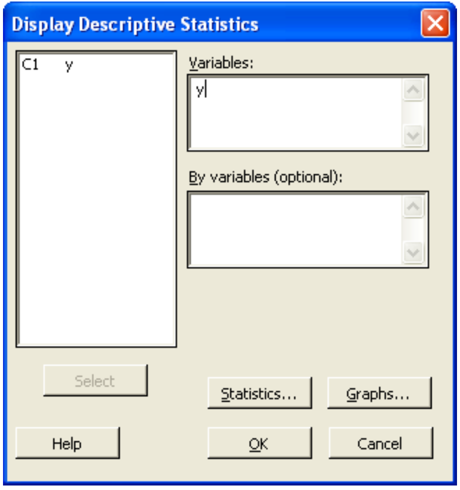
↓	C1	C2	C3
	y		
1	8.7		
2	8.1		
3	9.8		
4	6.1		
5	7.1		
6			
7			

When the values are in the Worksheet you can do the analysis by choosing a menu. In this case we choose

Stat ▶ Basic Statistics ▶ Display Descriptive Statistics…



In the window you now can use different ways to choose the variable, but the easiest way to do it is to first of all put the marker[2] in the white area marked Variables: click ones on the line C1 y and then click Select (you can also double-click C1 y).


Display Descriptive Statistics

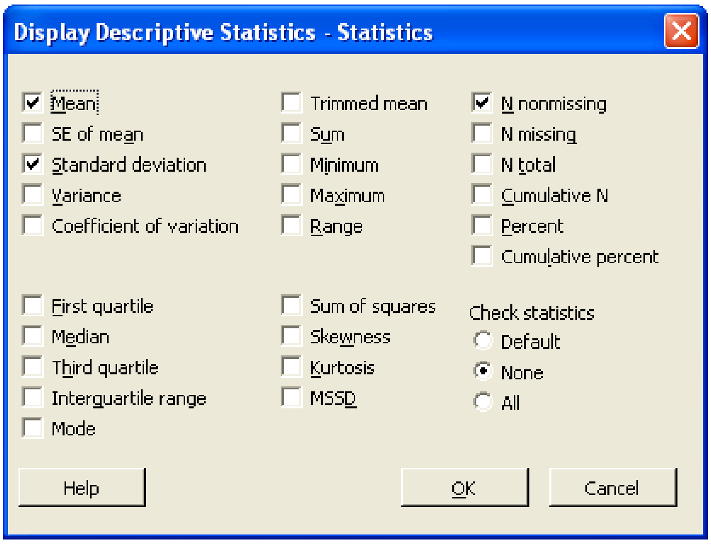
C1 y

Variables: y

By variables (optional):

[2] The marker has to be in the box marked Variables, otherwise you have no variables to choose.

To decide what types of statistical measurements you want to calculate, you click on Statistics… to get the menu

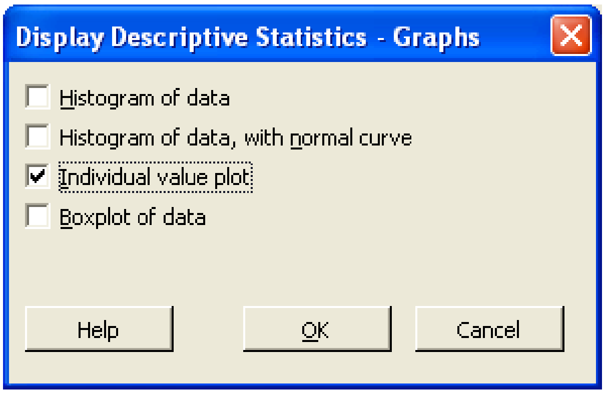


Display Descriptive Statistics - Statistics

☑ Mean
□ SE of mean
☑ Standard deviation
□ Variance
□ Coefficient of variation
□ Trimmed mean
□ Sum
□ Minimum
□ Maximum
□ Range
☑ N nonmissing
□ N missing
□ N total
□ Cumulative N
□ Percent
□ Cumulative percent
□ First quartile
□ Median
□ Third quartile
□ Interquartile range
□ Mode
□ Sum of squares
□ Skewness
□ Kurtosis
□ MSSD

Check statistics
○ Default
● None
○ All

In this case we choose Mean (= average), Standard deviation and N nonmissing (the number of observations which are not missing). Press OK when you have done you choice. To get different types of graphs to describe the dataset you can choose Graphs… . In this case we choose an “Individual value plot” and therefore have the menu



Display Descriptive Statistics - Graphs

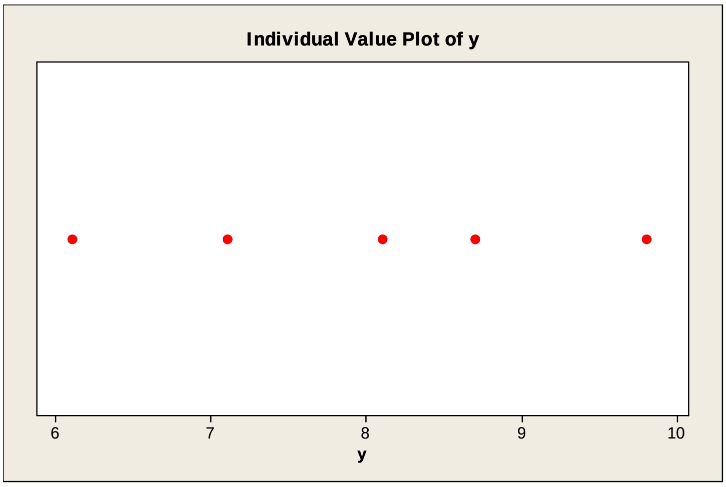
□ Histogram of data
□ Histogram of data, with normal curve
☑ Individual value plot
□ Boxplot of data

and press OK . Choose OK and look at the result. In the Session Window now is

```
Descriptive Statistics: y

Variable  N   Mean  StDev
y         5  7.960  1.428
```

You can see the result under the Headings Mean and StDev (= Standard Deviation), 7.960 resp. 1.428.

You also have a graphical illustration as



The picture is its own window and you can edit this picture, for example if you want another title or change the scales on the axes. ■

There are three types of windows, Session with the extension `.TXT`, Worksheet with the extension `.MTW` and Graphs with the extension `.MGF`. Furthermore, you can save the total project with the extension `.MPJ`. It is often very convenient to save the complete project, when you open it again the program remembers all the menus, graphs and outputs.

2. Reading the observations

When you have your observations and will use a computer to analyse the dataset, it is good to have the data in a way to make the further analysis as easy as possible. There are in fact rather general rules for how this should be done to suit most computer packages.

The dataset is read into a matrix, where the rows ("horizontally") is observations and the columns ("vertically") is the values for one of the variables. This way to enter the observations makes it easy to add new observations.

$$\begin{array}{r|c|c|c|c}
 & \text{Variabel 1} & \text{Variabel 2} & \text{Variabel 3} & \dots \\
 & \downarrow & \downarrow & \downarrow & \\ \hline
\text{Observation 1} \rightarrow & & & & \dots \\ \hline
\text{Observation 2} \rightarrow & & & & \dots \\ \hline
\text{Observation 3} \rightarrow & & & & \dots \\ \hline
\vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

Example 2. Children

From a class in the school you have a sample of five children where name, sex, age and weight are observed. A matrix with five rows and four columns can describe the dataset, and one more row has been added to denote the names of the variables.

NAMN	KÖN	ÅLDER	VIKT
Lisa	F	6	20
Stina	F	7	25
Lars	M	6	17
Peter	M	8	30
Anna	F	7	27

(Here we have used Swedish names for the variables, but the English names are often to be preferred because some packages don't accept the Swedish letters å, ä och ö.) ■

You can also use one or more variable to classify the observations into groups. In the following example the treatment and the week are variables used to split the dataset into different groups.

Exempel 3. Leek

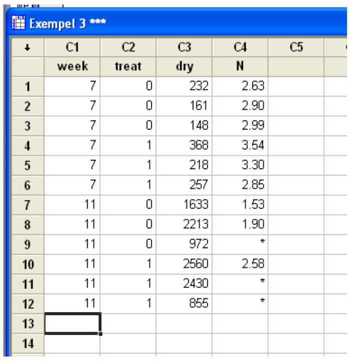
In an experiment to investigate how the content of nitrogen in leek is changed, they have compared one field without fertilization and one field with clover/oats. The dry weight of the leek in kg/ha and the content of nitrogen in percent of the dry weight is measured on the places per treatment and the measurement has been repeated after 7 and 11 weeks.

To simplify the entering of the dataset the treatment is quantified as

$$\begin{cases} 0 = \text{no fertilization} \\ 1 = \text{clover/oats} \end{cases}.$$

To get an overview of the result from a statistical package the dataset is entered according to the table.

If you will enter new variables into the Worksheet, you can continue in the same sheet, and put them in C2 if you already have something in C1, but you also can make a new Worksheet to have a better overview and structure for the different experiments. In this case we make a new Worksheet with the name Exempel 3 by choosing **File▶ New...**.[3].



Exempel 3 ***

↓	C1	C2	C3	C4	C5
	week	treat	dry	N	
1	7	0	232	2.63	
2	7	0	161	2.90	
3	7	0	148	2.99	
4	7	1	368	3.54	
5	7	1	218	3.30	
6	7	1	257	2.85	
7	11	0	1633	1.53	
8	11	0	2213	1.90	
9	11	0	972	*	
10	11	1	2560	2.58	
11	11	1	2430	*	
12	11	1	855	*	
13					
14					

[3] Three stars (***) after Worksheet 2 tell you that this is the worksheet in use.

An asterisk has denoted missing values, and this is the “missing value code” in Minitab. Other programs use other symbols, SAS uses a point and in Excel you write “N/A”. It often is the missing values that complicate the transportation of datasets between different packages. ■

3. Minitab with Word and Excel

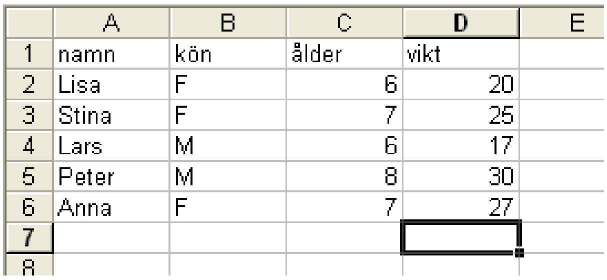
3.1. From Excel to Minitab

To move a dataset from Excel to Minitab it is important to remember the advices given above for the entering of datasets. If you do like this, you have a rectangular scheme, where the observations are in the rows and the variables are in the columns. It is important to remember that the Worksheet in Excel cannot contain averages and standard deviations, but only the dataset.

It is often preferable to read large datasets into Excel because this program has more facilities for data entry and editing.

Example 2 (cont). Children

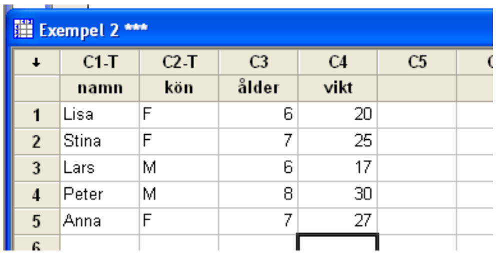
To read the dataset with children in Excel you make a worksheet according to



	A	B	C	D	E
1	namn	kön	ålder	vikt	
2	Lisa	F	6	20	
3	Stina	F	7	25	
4	Lars	M	6	17	
5	Peter	M	8	30	
6	Anna	F	7	27	
7					
8					

You now can choose to open a saved Worksheet from Excel directly into Minitab, but if it is small datasets it is often easier to copy and paste directly into Minitab.

To paste you highlight all observations, the variable names included, and choose to copy. Enter Minitab and put the marker in the line for the position of the variable names (here we paste the data in a new worksheet). The non-numerical variables get the names C1-T and C2-T.



Exempel 2 ***

↓	C1-T	C2-T	C3	C4	C5
	namn	kön	ålder	vikt	
1	Lisa	F	6	20	
2	Stina	F	7	25	
3	Lars	M	6	17	
4	Peter	M	8	30	
5	Anna	F	7	27	
6					

Now you can let Minitab work with this dataset. ■

Note that one reason for having a column denoted “T” might be that you have a decimal comma instead of a decimal point (or in the opposite direction depending on the settings).

3.2. From Minitab to Word

The results in the Session Window or in a Graph Window can immediately be printed by choosing printing the window. Then you have the output as A4, but you often wish to put the results in a document as in the example above.

To paste a part of the Session Window into Word you highlight the text and copy as usual. Then you go to Word without closing Minitab and paste the text.

It is a good exercise to copy the previous result and move it into Word and see the formatting is preserved[4].

Windows with graphs (“Graph Window”) is easier to copy, you just have to enter the graphical window, choose copy, and then go to Word and paste. If you don't want the link back to Minitab you can choose Paste Special.

3.3. Read data into Minitab

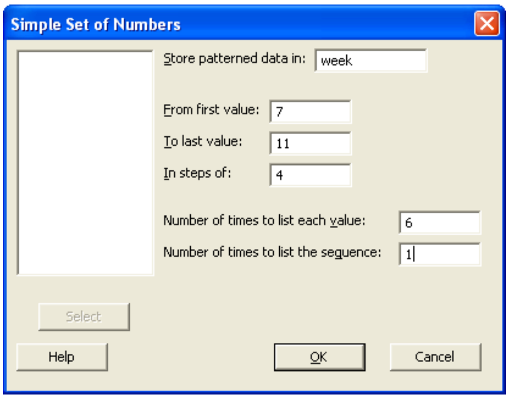
There is one situation when Minitab is to be preferred when you read your data, and this is when you have numbers who are repeated in a regular pattern. As an example we choose the previous example with nitrogen in leek.

Example 3 (cont). Leek

We will have one column with six 7 followed by six 11 and another column with three 0 followed by three 1 repeated twice. This is accomplished by choosing

Calc ▶ Make Patterned Data ▶ Simple Set of Numbers…

To make the first column you choose



Store patterned data in: week

From first value: 7

To last value: 11

In steps of: 4

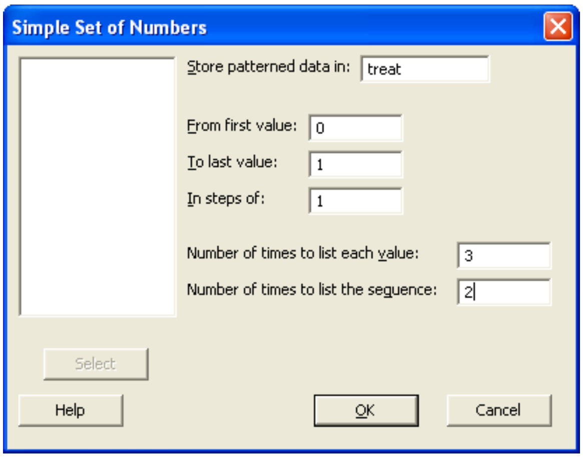
Number of times to list each value: 6

Number of times to list the sequence: 1

To make the column with 0 and 1 repeated twice you again choose

Calc ▶ Make Patterned Data ▶ Simple Set of Numbers…

[4] The Session Window uses Courier to make positions of the table columns correct.


Simple Set of Numbers

Store patterned data in: treat

From first value: 0

To last value: 1

In steps of: 1

Number of times to list each value: 3

Number of times to list the sequence: 2

(It is a good exercise to realise that this give you the required column.) ■

3.4. Edit Session Window in Minitab

You soon realise that the result is added to the Session Window, and if you want a single result you soon realise that you have a lot of results. However, it is easy to erase the output in the Session Window. If you will guarantee not to erase by mistake you can make the Session Window “read-only” by choosing

Editor ▶ Output Editable

when the marker is in the Session Window [5].

3.5. Edit Worksheet in Minitab

Copy parts of columns

Sometimes you want only a part of a column. In the example with leek you perhaps will separate the measurements done after seven weeks from the measurements done after eleven weeks.

Example 3 (cont). Leek

To copy the results for seven weeks to a new Worksheet you choose

Data ▶ Subset Worksheet…

and tell which lines you wish to copy. Here is one alternative, Brushed rows, and this means that you in a picture can use **Editor ▶ Brush** and mark the observations you want to exclude from the analysis. You can also use

Data ▶ Copy ▶ Columns to Columns…

to copy parts of the dataset to new columns or a new worksheet. ■

[5] Note that the content of the menues are changed according to the type of Window you are in.

Mathematical functions

Sometimes you wish to transform the observations or use a function not available in the menus. If you, for example, will use the logarithm of the values you can do this by using

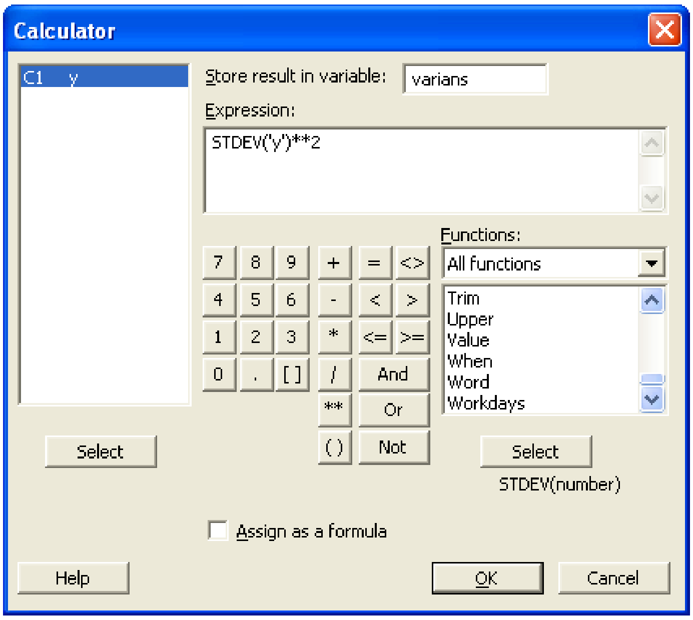
Calc ▶ Calculator …

Example 1 (cont). Introductory example

In this example we calculated the mean and the standard deviation, and you can also have the variance. As an example of the calculator we do this calculation, even though it is easier to do it in the Stat menu. Choose

Calc ▶ Calculator …

and to have the answer in a column named varians you write (** is the notation to "raised to the power of" in Minitab)



Calculator

C1 y

Store result in variable: varians

Expression: STDEV('y')**2

Functions: All functions

Trim
Upper
Value
When
Word
Workdays

STDEV(number)

□ Assign as a formula

You can write the variable directly in the box, but note that Minitab does not warn you if you replace an already used variable name. If you will transform the observations in a column or add two columns, this is also easily done. With the description above this could be done without too much problems. ■

4. Descriptive statistics

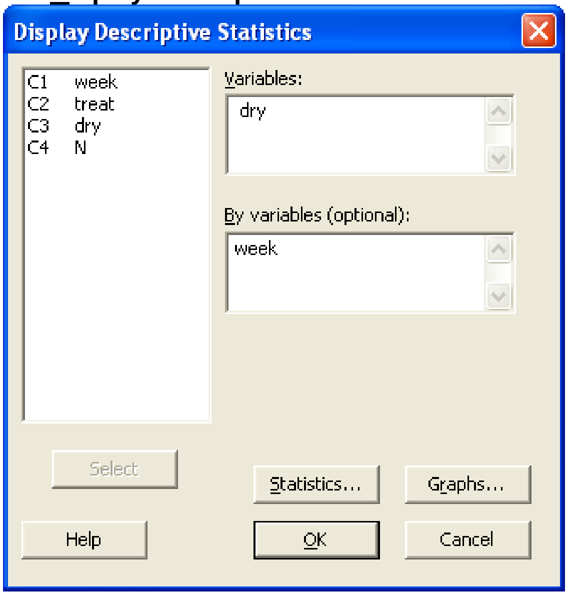
4.1. Numerical methods

To investigate how the yield is changed from the 7th to the 11th week you can use some numerical measures to get some information about this. Here we don’t consider the fact that the leeks have different treatments; the aim is to give an example.

Example 3 (cont). Leek

You get the descriptive measures by using

Stat ▶ Basic Statistics ▶ Display Descriptive Statistics…



Display Descriptive Statistics

C1 week
C2 treat
C3 dry
C4 N

Variables: dry

By variables (optional): week

The result is (Minitab remember what you did last, so if “Individual value plot” still is crossed, you get this one also this time, and you also get the same descriptive statistics as you earlier choose with Statistics…)

```
Descriptive Statistics: dry

Variable  week  N   Mean  StDev
dry          7  6  230.7   79.3
            11  6   1777    741
```

You could have chosen other descriptive measures. Here is an explanation of some of these different measures:

N is the number of observations used in the calculations.

N* is the number of observations with a “missing value” (coded as *).

Mean is the arithmetic mean.

Median is as usual the value “in the centre”, or the mean of the two values in the center if it is an even number of values.

Tr Mean, is “trimmed mean” where you have removed a number of small and large values and then calculated the mean. This is to prevent that “outliers” have too much influence when you calculate the mean. Usually you remove 5% of the largest and 5% of the smallest observations when you calculate the trimmed mean.

StDev is the standard deviation. It is a considerably larger variation among the older leeks.

SE Mean is the Standard Error. The formula is $StDev/\sqrt{N}$ and this is in some cases more useful than the standard deviation.

Min and Max is of course minimum and maximum of the values.

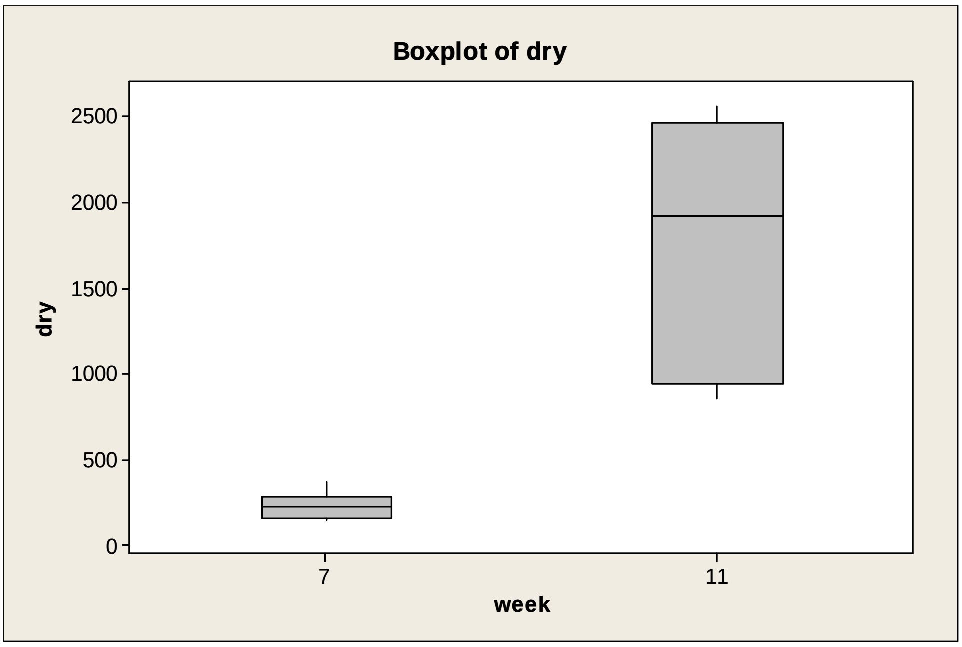
Q1 and Q3 are the quartiles. You get the quartile by ordering the observations and split them into four parts. The three points that separates the material is the lower quartile, the median and the upper quartile. The quartiles are also illustrated in a boxplot.

4.2. Boxplot

The boxplot gives an illustrative graphical presentation of the median and the quartiles (or box-and-whisker-plot). To get it you choose

Stat ▶ Basic Statistics ▶ Display Descriptive Statistics…

Then choose Graphs… . and "Boxplot of data". Press OK and you have the picture



In the picture the line inside the box is the median, the lower part of the box is the lower quartile and the upper part is the upper quartile.

4.3. Scatterplot

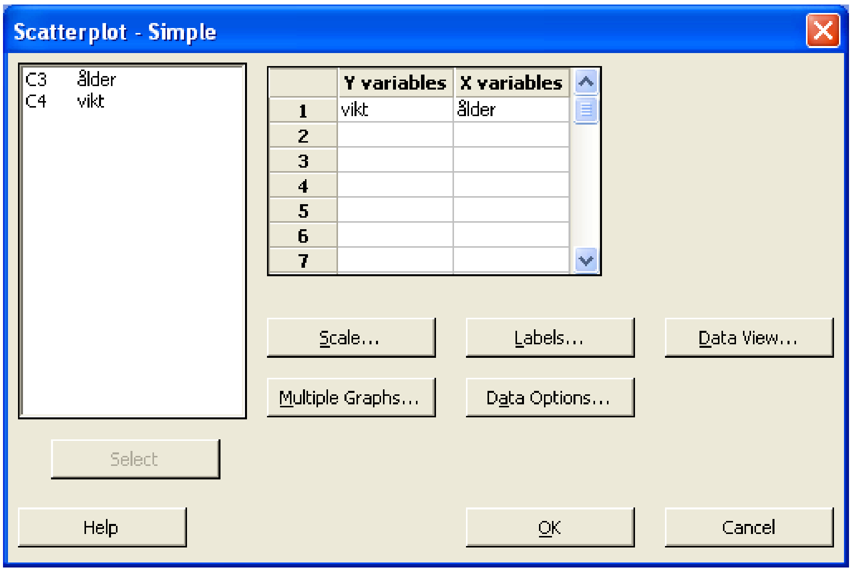
If you want to illustrate two-dimensional variables you can use a scatterplot to see if there is a relation between the variables.

Example 2 (forts). Children

To draw the variable vikt against ålder in a scatterplot you use

Graph ▶ **ScatterPlot**…

and then choose the marked box ”Simple”. Fill in as follows:

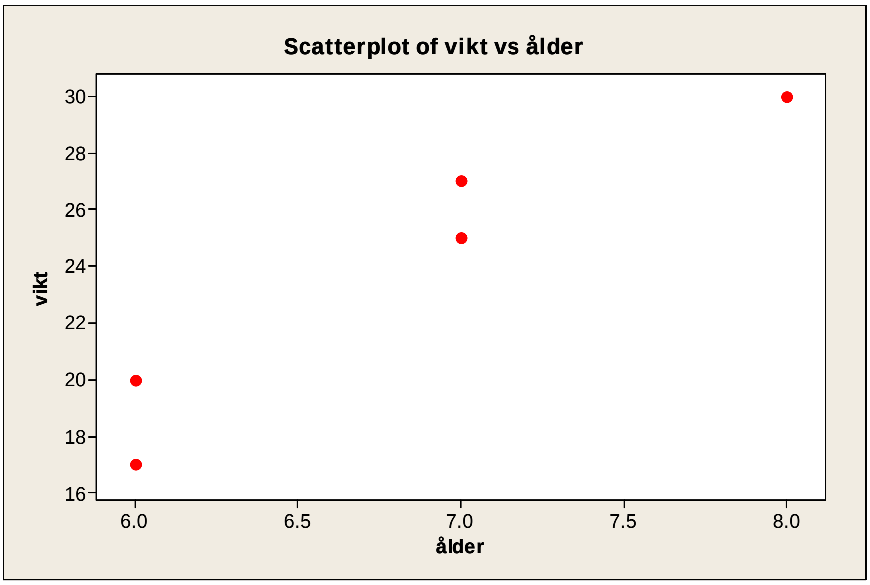


Scatterplot - Simple

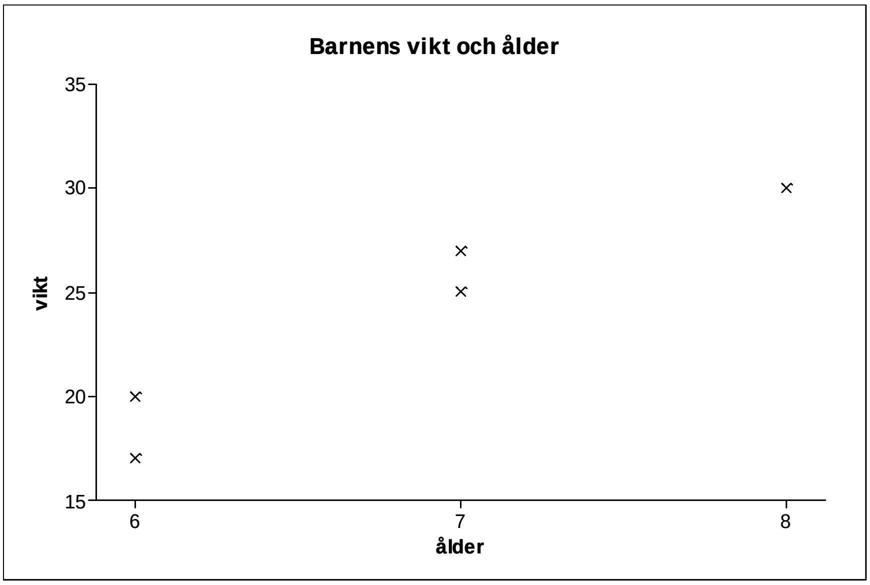
C3 ålder
C4 vikt

	Y variables	X variables
1	vikt	ålder
2		
3		
4		
5		
6		
7		

Press OK , and the result is



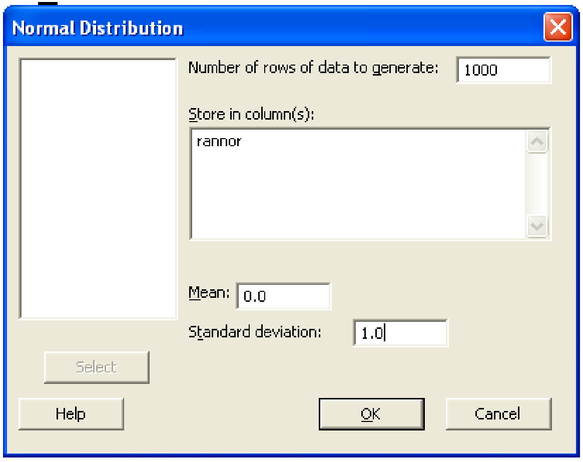
The picture given by Minitab can be improved a lot. By choosing the different alternatives you can edit the figure. An example of what you can have after some work is the following figure:



4.4. Histogram

A histogram is very easy to do, as soon as you know where to find it! As an example we check if the random number generator in Minitab give you normally distributed observations. Random numbers from the normal distribution with mean 0 and standard deviation 1 you have by

Calc ▶ Random Data ▶ Normal…


Normal Distribution

Number of rows of data to generate: 1000

Store in column(s): rannor

Mean: 0.0

Standard deviation: 1.0

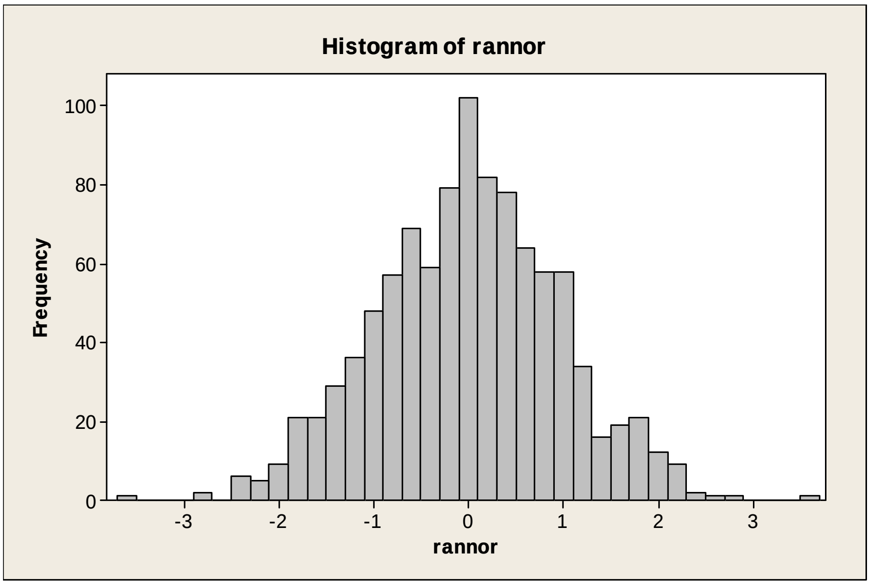
Now the computer has randomised 1000 observations of normally distributed random variables with mean 0 and standard deviation 1, and they are in the column named rannor. To draw a histogram of these observations you can use the menu for descriptive statistics, but here we use an alternative with more possibilities to change the appearance of the picture (in fact this is also true for the Boxplot we made earlier):

Graph ▶ Histogram…

Choose ”Simple”.

Press OK and choose rannor as your variable:

One example of this type of histogram is below, but it is random numbers, so if you do this you probably don’t have exactly the same histogram. However, with 1000 observations the bell-shaped curve should be there.



5. Statistical methods for one sample

The idea with experiments is usually to compare one or more treatments, but sometimes you have only one sample, and we start there, and go on with more samples later on.

When you do the analysis it is often based on the assumption that the observations are from a simple random sample, that is that the observations are independent.

5.1. Normal distribution and one sample

Usually you assume that the sample is from a normal distribution, or at least that it is approximately normally distributed. This means that if you collect a lot of observations and make a histogram, this histogram will look like the bell-shaped curve. Based on one sample from a normal distribution you can do test of hypotheses or you can do a confidence interval for the (population) mean. This is very easy to do in Minitab, you only have to decide whether the standard deviation is known or if it should be estimated from the sample. In practice it is not very often that the standard deviation is known, and therefore we only treat the situation with unknown standard deviation.

Example 4. One sample

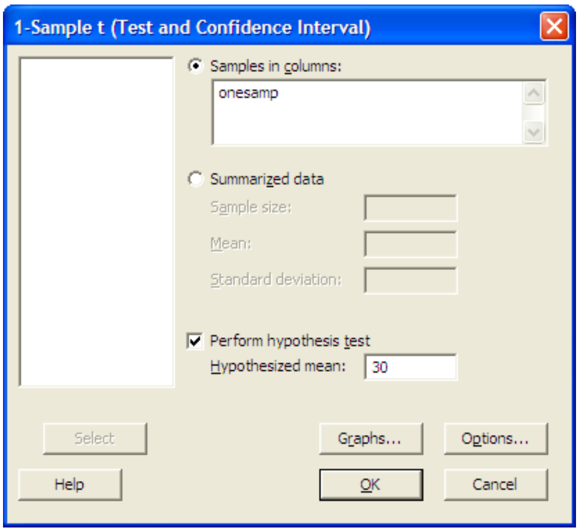
Test if it might be that the sample 33.5, 32.0, 32.5, 36.5 is sampled from a population with mean 30.

These values are read in a column called onesamp:

The null hypothesis in this case is that the mean is 30 while the alternative hypothesis is that the mean not is 30. Choose

Stat ▶ **Basic Statistics** ▶ **1-Sample t**…

(**1-Sample Z** when the standard deviation is known)



1-Sample t (Test and Confidence Interval)

● Samples in columns: onesamp

○ Summarized data

Sample size:

Mean:

Standard deviation:

☑ Perform hypothesis test

Hypothesized mean: 30

You have to write the null hypothesis to get a result of your test. To change the alternative hypothesis or the confidence level you use Options… but in this case it is the way we want it, not equal is exactly the alternative hypothesis we have. The result is in the Session Window

One-Sample T: onesamp

```
Test of mu = 30 vs not = 30


Variable  N   Mean  StDev  SE Mean       95% CI          T      P
onesamp   4  33.63   2.02     1.01  (30.42; 36.83)  3.60  0.037
```

The value P is 0.037 and the most interesting. This so-called P-value tell you if the hypothesis can be rejected or not, and the rule usually is to reject the null hypothesis if the P-value is less than 0.05. Here we can reject the null hypothesis, that is, we have reason to suspect that the mean *not* is 30. However, in statistics we can never be sure, we have chosen the level so that in 1 case out of 20 rejects the null hypothesis even if it is true. ■

An alternative to hypotesis testing is to do a confidence interval. A confidence interval with confidence level 95% is also given in the printout above as (30.42; 36.83), and if you will change the confidence level you can choose Options…. The confidence level 95% tells you that the given confidence interval has the probability 95% to cover the true value of the mean (that is, it covers the true mean in 19 cases out of 20).

It is of course good if the width of the interval is small. The width of the interval depends on the number of observations in the sample (and this might be possible to increase), and the variation between different observations (and this is hard to influence for the experimenter). In the example above, the interval probably is too wide to be useful in practice.

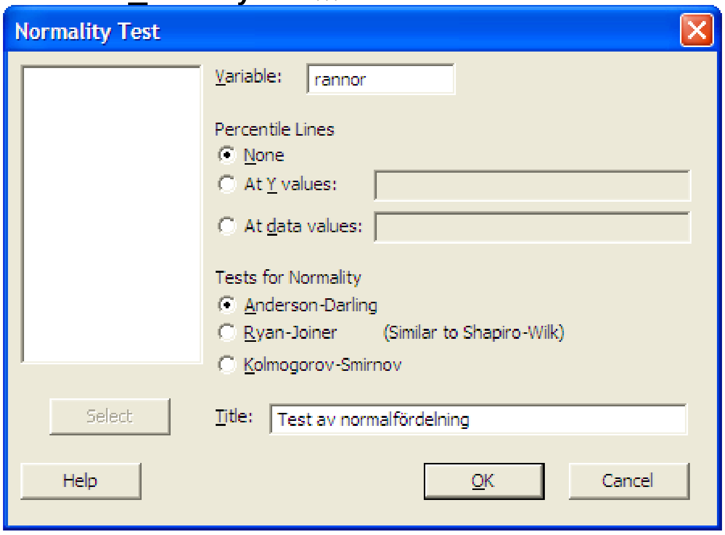
An assumption for the test and the interval to be a satisfactory solution is that the population is normally distributed. One method to check this is to use a so-called normal probability plot. Before the era or computers this was a very time-consuming activity, but now it is very easy to do.

Normal probability plot

A normal probability plot has a nonlinear scale on the *y*-axis, and this scale is done to have the points to approximately follow a straight line if the observations are from a normally distributed population. When you test whether the points are close to the line, you use asymptotic results, that is, the results are only valid if the number of observations is large. We earlier used a data set with 1000 normally distributed observations to exemplify the histogram.

The data set is already in the variable rannor. Choose

Stat ▶ Basic Statistics ▶ Normality test …



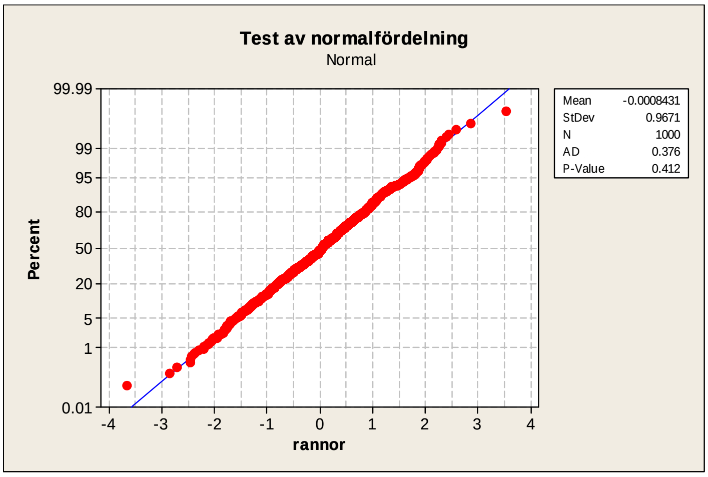
Variable: rannor

Percentile Lines
● None
○ At Y values:
○ At data values:

Tests for Normality
● Anderson-Darling
○ Ryan-Joiner (Similar to Shapiro-Wilk)
○ Kolmogorov-Smirnov

Title: Test av normalfördelning

(Here we also have given a title for the graph.) The result is



The P-value gives us a hint if the hypothesis concerning a normal distribution is true or not. A value below 0.05 rejects the null hypotheses that the population is normally distributed, and a P-value above 0.05 does not reject the hypothesis. Usually, this means that you continue to assume that the population is normally distributed. However, note that you don't have proved this, to say that you don't reject is not the to say that the alternative hypotheses is true. In this case we don't reject the hypotheses, and this is in accordance with the fact that the data were simulated from a normal distribution. ■

If you don't assume that the population is normally distributed, you have two alternatives:

1. If you have a lot of observations in the sample it might work even if it is not a normal distribution of the population, if there are not too many "outliers".
2. Use a non-parametric method, which does not assume that the population is normally distributed.

5.2. Non-normal population and one sample

Non-parametric tests are of course in the menu "Nonparametrics". In the case with one sample you have two different tests. Note that nonparametric tests don't test the mean, but the median. To test if the population has the median 30 is equivalent to test if half of the population is below 30 and half of the population is above 30.

Sign test

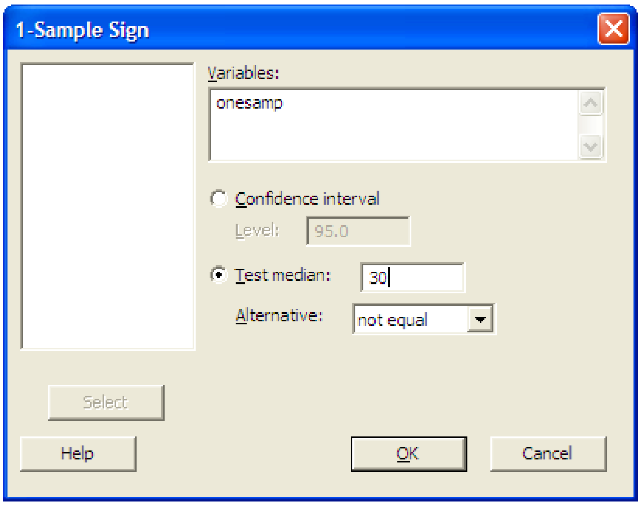
The sign test is easy to understand, but unfortunately not very powerful. The idea is that if the median is 30, half of the observations in the sample should be smaller than 30 and the rest of them should be above 30. If you for example have 20 observations in the sample and all are below 30, it is hard to believe that the median is 30.

Example 4 (cont). One sample

To decide whether the median is 30 with a test, choose

Stat ► Nonparametrics ► 1-Sample Sign…

and tell the program that you will test if the median is 30.



Variables: onesamp

○ Confidence interval

Level: 95.0

● Test median: 30

Alternative: not equal

The output in the Session Window is

```
Sign Test for Median: onesamp

Sign test of median =  30.00 versus not = 30.00

         N  Below  Equal  Above       P  Median
onesamp  4      0      0      4  0.1250   33.00
```

Even if all the observations are above 30, you cannot reject the null hypotheses (the P-value is 0.1250, and this is larger than 0.05). The sign test is not powerful enough in this case. ■

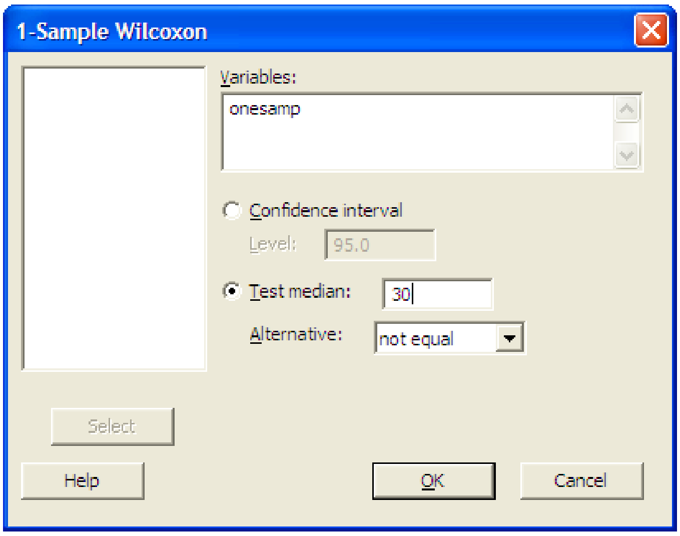
Wilcoxon's signed rank test

This test is more powerful because it also uses the size of the values, but in this test you also have the assumption that the distribution of the population will be symmetric. This also has as the consequence that the mean and the median are identical.

Example 4 (cont). One sample

To decide whether the median is 30, choose

Stat ▶ **Nonparametrics** ▶ **1-Sample Wilcoxon**…



Variables: onesamp

○ Confidence interval

Level: 95.0

● Test median: 30

Alternative: not equal

Wilcoxon Signed Rank Test: onesamp

```
Test of median = 30.00 versus median not = 30.00

              N for   Wilcoxon           Estimated
           N   Test  Statistic       P      Median
onesamp    4      4       10.0   0.100       33.25
```

You cannot reject the null hypotheses even if the P-value is a little bit lower in this test. ■

The example is illustrative. If you can assume that the distribution of the population is normal, you should use a test based on the normal assumption because these tests are more powerful.

There is also another reason to use a non-parametric test, and this is that a non-parametric test is more robust against “outliers” or bad data. If you by mistake have printed 365 and not 36.5 as your last observation, the P-value in the test based on the normal assumption will change to 0.38. It might be surprising that the test not rejects the null hypotheses if you move one of the values further away from 30, but the reason is that you increase the estimated standard deviation considerably. If you do the sign test or the Wilcoxon's signed rank test, the P-value is the same even if you introduce this misprint.

6. Statistical methods for two samples

The most common problem in practice is to compare one or more populations. In the previous section, one of the main results were that if you assume that the population is normally distributed, you could use “better” tests than you could do without this assumption. With two samples, it is similar; if you have a normally distributed population you use the famous *t*-test.

6.1. Normal distribution and two samples

If you assume that the two populations are normally distributed, there are two parameters for each population, the mean and the standard deviation. If you can assume that the standard deviations in the two populations are the same, the only difference is the means, and to test if the means are the same is equivalent to test if the two samples are from the same population. There is also another reason why you often assume that the standard deviations are the same; the theory is simplified if you have the same the same standard deviations (this is called homoscedasticity). In a computer package it is not more complicated with different standard deviations, but if you have more than two samples it is more complicated also by the computer.

You also can do a confidence interval for the difference between the two population means. Here you usually check if the interval covers 0; if so the means are not significantly different.

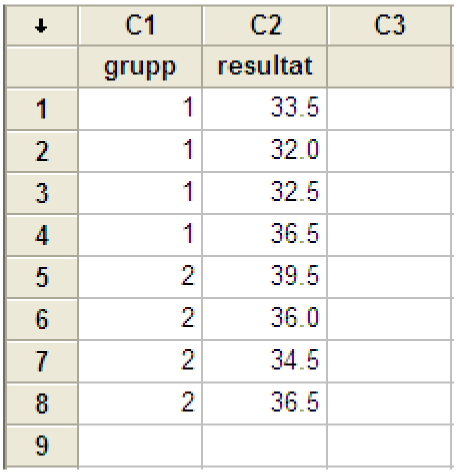
Example 5. Two samples

Test whether the samples 33.5, 32.0, 32.5, 36.5 and 39.5, 36.0, 34.5, 36.5, respectively, are from the populations with the same mean.

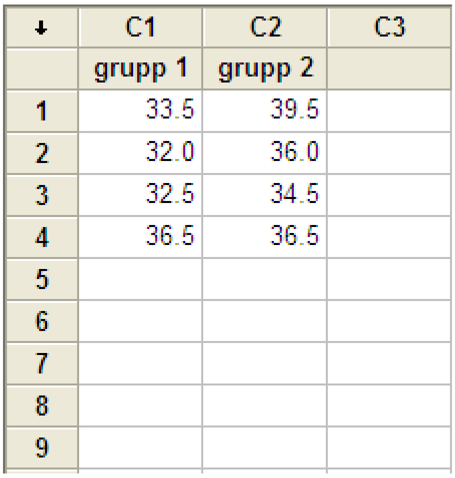
The null hypothesis is in this case that the means are the same, the alternative hypotheses is that the means are different. Choose

Stat ▶ Basic Statistics ▶ 2-Sample t…

Now you have to choose whether the data is in one column (Samples in one column), in two columns (Samples in different columns) or if you already have calculated the means and standard deviations (Summarized data). If you read the data according to the recommendations in earlier chapters, you have one column, but you also have one more column telling what sample that observation is from. The dataset is then as described to the left (Samples in one column), but two different columns (Samples in different columns) is to the right.

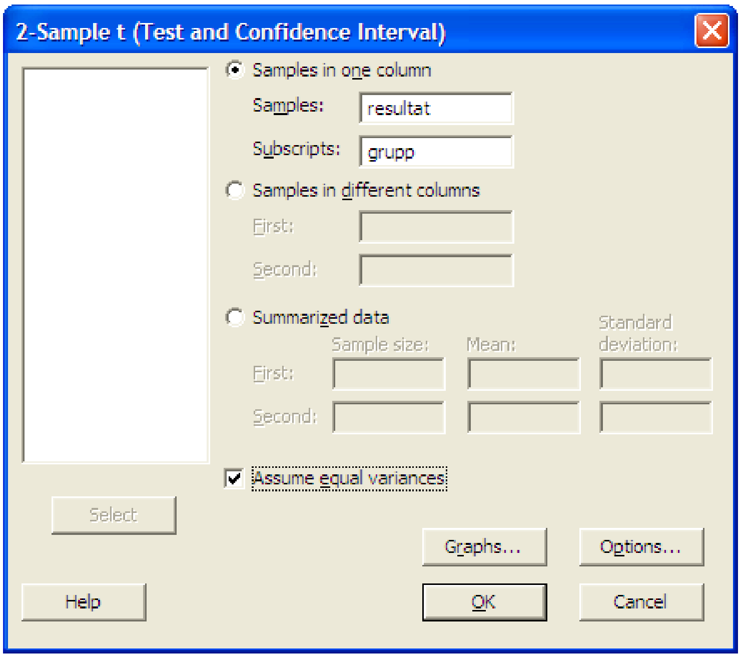


↓	C1	C2	C3
	grupp	resultat	
1	1	33.5	
2	1	32.0	
3	1	32.5	
4	1	36.5	
5	2	39.5	
6	2	36.0	
7	2	34.5	
8	2	36.5	
9			



↓	C1	C2	C3
	grupp 1	grupp 2	
1	33.5	39.5	
2	32.0	36.0	
3	32.5	34.5	
4	36.5	36.5	
5			
6			
7			
8			
9			

Here we choose one column, and the window looks like



● Samples in one column

Samples: resultat

Subscripts: grupp

☑ Assume equal variances

The alternative hypotheses and the significance level can be changed if you choose Options… . In the last line in the menu you can choose whether the variances (and the standard deviations) are the same or not (Assume equal variances). Note that the default is that the variances *not* are equal but it is most common in practice to assume that the variances are equal. If you have chosen equal variances, the printout is

Two-Sample T-Test and CI: resultat; grupp

```
Two-sample T for resultat

grupp  N   Mean  StDev  SE Mean
1      4  33.63   2.02      1.0
2      4  36.63   2.10      1.0


Difference = mu (1) - mu (2)
Estimate for difference:  -3.00
95% CI for difference:  (-6.56; 0.56)
T-Test of difference = 0 (vs not =): T-Value = -2.06  P-Value = 0.085  DF = 6
Both use Pooled StDev = 2.0565
```

The “difficult” part of the printout is the line with *t*-test. The interesting P-value is 0.085, which shows that you *cannot* reject the hypotheses that the samples are from populations with the same mean. The value of T is a test quantity that is of no importance, and the same is true for the DF. Since we have assumed that the variances (and the standard deviations) are the same, you also have an estimate of the common standard deviation, and in this case this is 2.06. Here it is exactly the average of the standard deviation of the two sample standard deviations, but this is just a coincidence. The “pooled” standard deviation is calculated by a more complicated formula. Also note that the confidence interval covers 0, and this also shows that the null hypotheses about the same means cannot be rejected at level 0.05 (since the confidence interval has the confidence level 95%). ■

You can also do a test to check whether it makes sense to assume that the variances are the same, and this test is discussed in the following section.

6.2. Test of equal variances

Example 5 (cont). Two samples

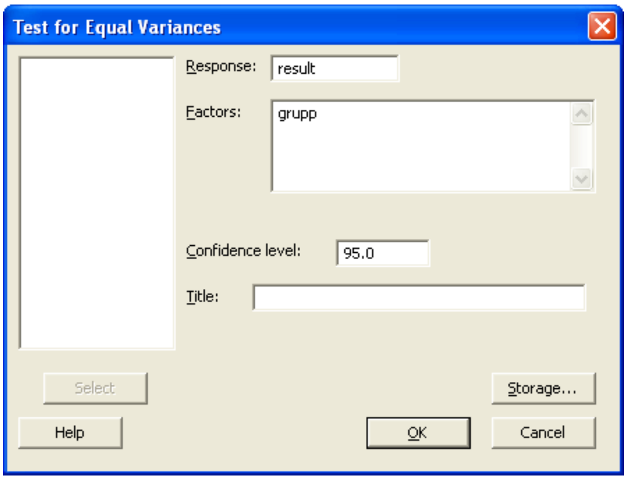
Test if the samples 33.5, 32.0, 32.5, 36.5 and 39.5, 36.0, 34.5, 36.5, respectively, can be from to populations with the same variance.

The null hypothesis is in this case that the variances are the same, and the alternative hypothesis is that the variances are different. Choose

Stat ▶ ANOVA ▶ Test for Equal Variances….

(You can use **Stat ▶ Basic Statistics ▶ 2 Variances…** if you only have two groups.)

With more than two groups you have to read the values in two columns and fill in as



Response: result

Factors: grupp

Confidence level: 95.0

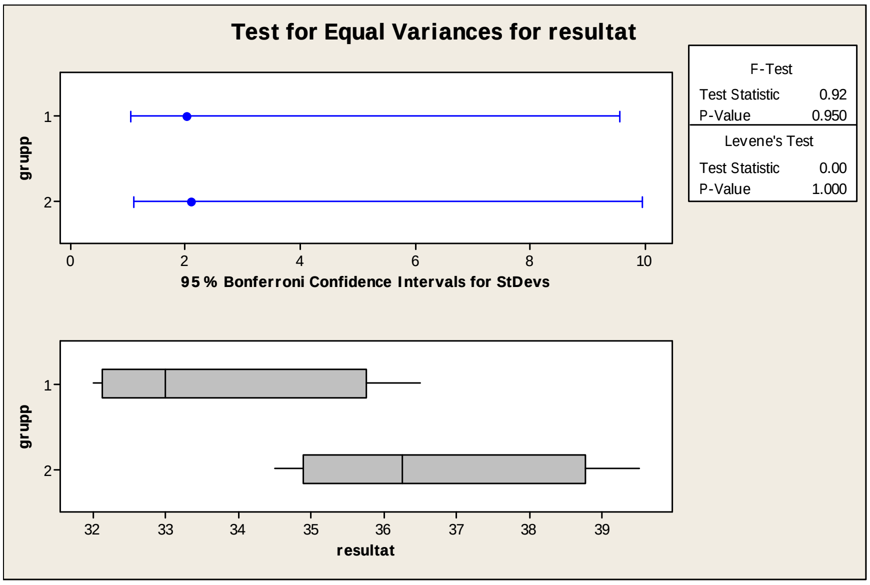
Title:

When you let Minitab do the calculations you have one output in the Session Window and also one figure. The result in the Session Window is

Test for Equal Variances: resultat versus grupp

```
95% Bonferroni confidence intervals for standard deviations

grupp  N    Lower    StDev    Upper
    1  4  1.05929  2.01556  9.54568
    2  4  1.10190  2.09662  9.92958


F-Test (Normal Distribution)
Test statistic = 0.92; p-value = 0.950


Levene's Test (Any Continuous Distribution)
Test statistic = 0.00; p-value = 1.000
```

Here you have two P-values, one corresponding to an F-test (with more than two groups this is called Bartlett’s test) used when you have normal populations, and one named Levene's test which can accept some deviances from the normal distribution. Both the P-values are above 0.05, and therefore you *cannot* reject the null hypothesis that the variances (or the standard deviations) are equal. The figure also gives a graphical illustration of the variances in the different groups.



Note that this test not is restricted to two samples. However, then the F-test is substituted with Bartlett's test, but this is *not* treated in the section with more than two samples. ■

6.3. Non-normal distribution and two samples

The test Minitab uses in this case is called Mann-Whitney's test. When you will do this test, you realise that Minitab not is consequent. In this case Minitab assumes that you have the two samples in different columns, to read it in one column does not work with Mann-Whitney’s method! However, if you have your data in one column it is easy to split it into two columns by using

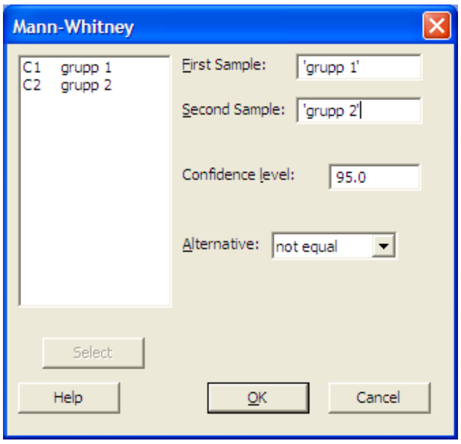
Data ▶ Unstack Columns…

Example 5 (cont). Two samples

Test if the samples 33.5, 32.0, 32.5, 36.5 and 39.5, 36.0, 34.5, 36.5, respectively, are from populations with the same median.

The null hypothesis is in this case that the medians are the same, the alternative hypothesis that the medians are different. Choose

Stat ▶ Nonparametrics ▶ Mann-Whitney…



Mann-Whitney

C1 grupp 1
C2 grupp 2

First Sample: 'grupp 1'

Second Sample: 'grupp 2'

Confidence level: 95.0

Alternative: not equal

If you wish, change the alternative hypothesis and the level of significance, but in practice the default values “95%” and “not equal” are the most common ones.

Mann-Whitney Test and CI: grupp 1; grupp 2

```
         N  Median
grupp 1  4  33.000
grupp 2  4  36.250


Point estimate for ETA1-ETA2 is -3.000
97.0 Percent CI for ETA1-ETA2 is (-7.501;2.001)
W = 12.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.1489
The test is significant at 0.1465 (adjusted for ties)
```

Here you have two P-values, 0.1489 and 0.1465 (adjusted for ties). The P-value adjusted for ties is when you have adjusted the test due to the fact that you have two or more identical values (=”ties”). (There are two values of 36.5.) Here you have the strange conclusion that the test is significant at level 0.1465, but this means that you cannot reject the hypothesis that there is a difference between the population medians. ■

As for one sample you lose some power if you use Mann-Whitney's test if you have a the possibility to make a *t*-test.

7. Statistical methods for more than two samples

The most common method in statistics is ANalysis Of Variance (often called ANOVA). The standard assumptions in ANOVA are a normally distributed population and equal variances in all populations. It is also important to remember that even if the name may indicate that it is a test of variances, it is actually a test of the means; accomplished by comparing the variances.

7.1. Normal distribution and more than one sample

ANOVA is an extension of the *t*-test, and the difference is that you have more than two samples. The ANOVA can be extended in different directions, but in this section we just treat the most basic problem, to see if a number of samples is from the same population. This usually is called one-way ANOVA.

Example 6. More than two samples

You have four samples, and you will see whether these samples are from populations with identical means. There are reasons to believe that the samples are from normally distributed populations with equal standard deviations. The null hypothesis of the same means therefore is equivalent to test if all samples are from the same population.

Sample 1: 33.5, 32.0, 32.5, 36.5
Sample 2: 39.5, 36.0, 34.5, 36.5
Sample 3: 32.0, 32.5, 33.5, 34.5
Sample 4: 36.0, 35.5, 33.5, 37.0

You can choose if you will read the data in one column with one column expressing the sample, or to read the samples in four columns.

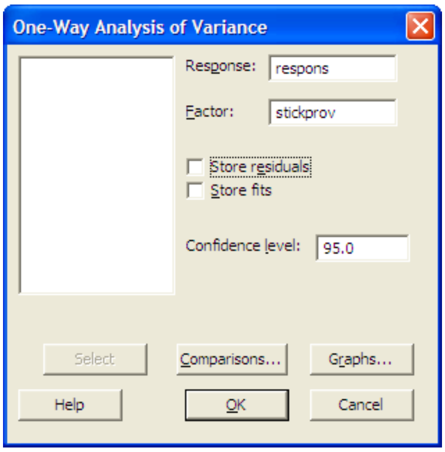
If you have the data in four columns, choose

Stat ▶ **ANOVA** ▶ **Oneway (Unstacked)** …

while if you have one column, choose

Stat ▶ **ANOVA** ▶ **Oneway**…

If you have the alternative with one column you have more options later on, and this method is therefore to be preferred. In this example the factors are in the variable stickprov and the responses are in the variable respons.



Response: respons
Factor: stickprov
□ Store residuals
□ Store fits
Confidence level: 95.0

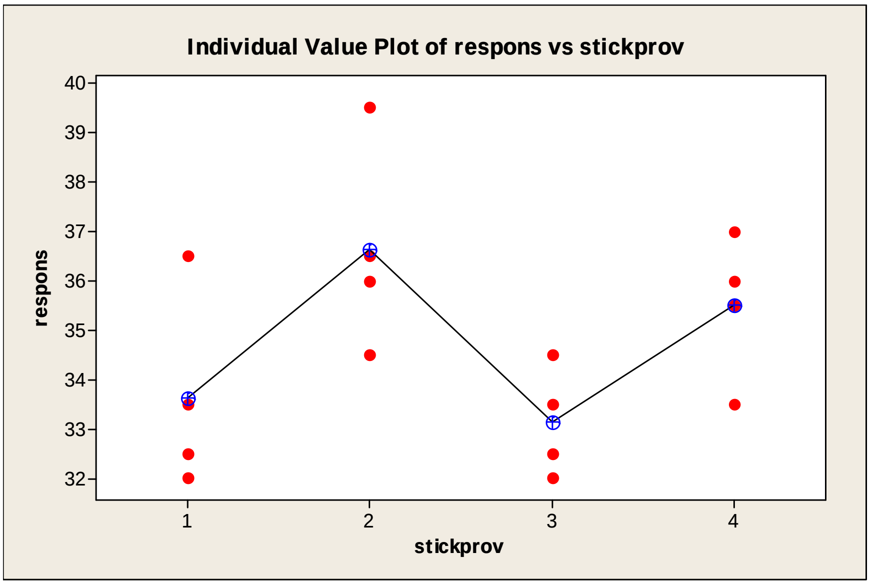
The result in the Session Window is

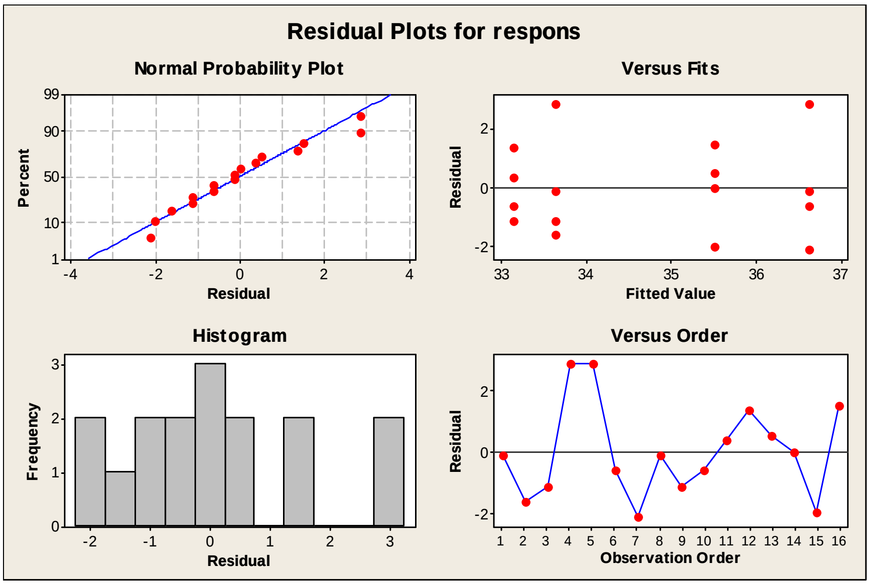
One-way ANOVA: respons versus stickprov

```
Source     DF     SS     MS     F      P
stickprov   3  31.92  10.64  3.59  0.046
Error      12  35.56   2.96
Total      15  67.48
```

```
S = 1.721   R-Sq = 47.30%   R-Sq(adj) = 34.13%


                         Individual 95% CIs For Mean Based on
                         Pooled StDev
Level  N    Mean  StDev  ----+---------+---------+---------+-----
1      4  33.625  2.016     (--------*---------)
2      4  36.625  2.097                    (--------*---------)
3      4  33.125  1.109  (---------*--------)
4      4  35.500  1.472              (--------*---------)
                         ----+---------+---------+---------+-----
                          32.0      34.0      36.0      38.0

Pooled StDev = 1.721
```

The most important number is the P-value 0.046 that shows that the null hypothesis that the samples are from different populations can be rejected. You also have illustrations of the confidence intervals for each sample. Note that this in *not* the same interval you have if you only look at one sample, because the pooled standard deviation (`Pooled StDev`) from all four samples is used in the intervals.





You should make some kind of graph of your data, and in this case you can do it by choosing the alternatives in Graphs... and use Individual Value Plot. To see if the observations are normally distributed you can choose Four in one. Note that it is the residuals and not the observations that are illustrated in the normal probability plot. (The residual is the observation minus the value the model says it should be. In one-way ANOVA the residual for an observation is the response minus the mean of all observations with the same treatment as this response.) ■

In an earlier section with two samples there was a test of equal variances, and this method is applicable also here with

Stat ▶ ANOVA ▶ Test for Equal Variances...

It is often not sufficient to test if there is a significant difference between the treatments, you will also know where the differences are. This is produced by Comparisons... in the ANOVA-menu and then you choose the required test. If you would like to do Fisher's test this is the only place you can find it, but if you would like to do Tukey's test the output is easier if you use the menu General Linear Model... and not One-way... In a later section block designs are discussed and there is an illustration of Tukey's test if you use General Linear Model...

7.2. Non-normal distribution and more than two samples

If you can't assume that the population is normally distributed and will make a non-parametric test, the test corresponding to the one-way ANOVA is Kruskal-Wallis' test. If you know how to read the data for a one-way ANOVA in one column, it is similar with Kruskal-Wallis' test. Choose

Stat ▶ Nonparametrics ▶ Kruskal-Wallis...

and choose your variables. The result is

Kruskal-Wallis Test: respons versus stickprov

```
Kruskal-Wallis Test on respons

stickprov   N  Median  Ave Rank      Z
1           4   33.00       6.1  -1.15
2           4   36.25      12.4   1.88
3           4   33.00       4.9  -1.76
4           4   35.75      10.6   1.03
Overall    16               8.5

H = 6.76  DF = 3  P = 0.080
H = 6.85  DF = 3  P = 0.077  (adjusted for ties)

* NOTE * One or more small samples
```

Kruskal-Wallis’ test is based on asymptotic results, and to trust the result your sample shouldn’t be too small, and this is not the case here. You therefore have the warning in the last line. Here you also have two P-values, and the difference is that the second one has taken into consideration that some values are identical. But as mentioned, there is a warning that the P-values are not reliable. ■

8. Block designs

If you have a block design with normally distributed observations you cannot use the menu for one-way ANOVA and have to use the more advanced procedure General Linear Model. You also have to enter the values the way described earlier for the procedure to work.

Example 10. Block design
The concentration of nitrogene in the soil was measured for four different fertilizers (A, B, C eller D). The experiment was a block design with three blocks (I, II, III). Result:

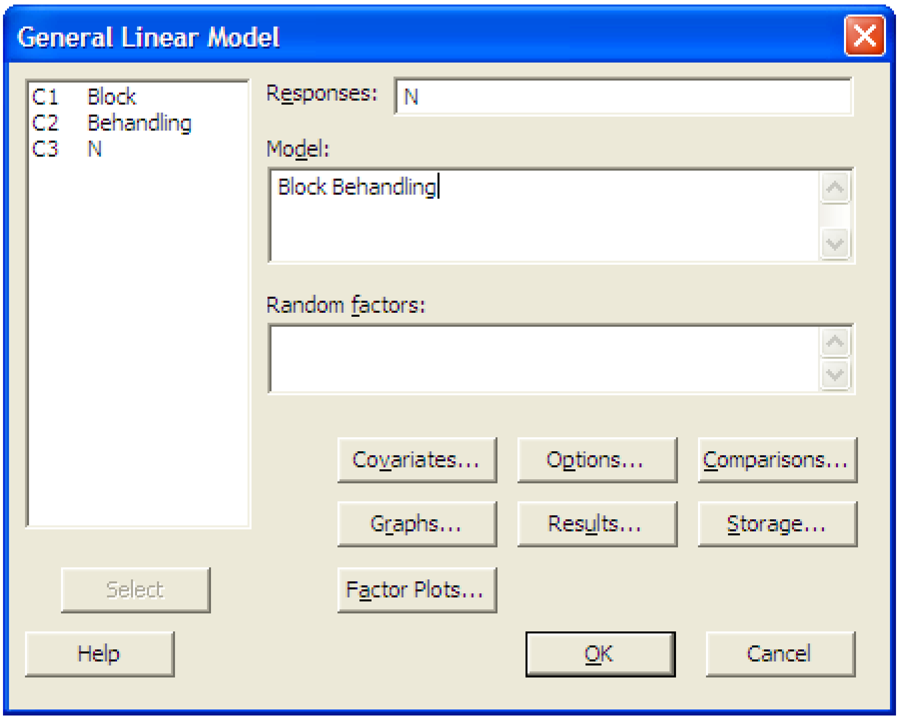
Block	Behandling	*N*
I	A	3.85
I	B	10.70
I	C	8.60
I	D	10.95
II	A	4.25
II	B	11.85
II	C	9.35
II	D	11.80
III	A	5.60
III	B	9.15
III	C	8.40
III	D	7.55

The data was entered into Minitab and then the columns Block and Behandling turns out to be text columns, but it works in a block design (and it also works in a one-way ANOVA but it doesn't work in regression)).

To see if there are any differences between the fertilizers we make an ANOVA in General Linear Models … (and if there are differences, we want to find out where the differences are).

Stat ▶ ANOVA ▶ General Linear Model…

The menu should be filled in as


General Linear Model

C1 Block
C2 Behandling
C3 N

Responses: N

Model: Block Behandling

Random factors:

and if you only want to find out if there are differences, and not make a picture of the data, this is enough.

The result is

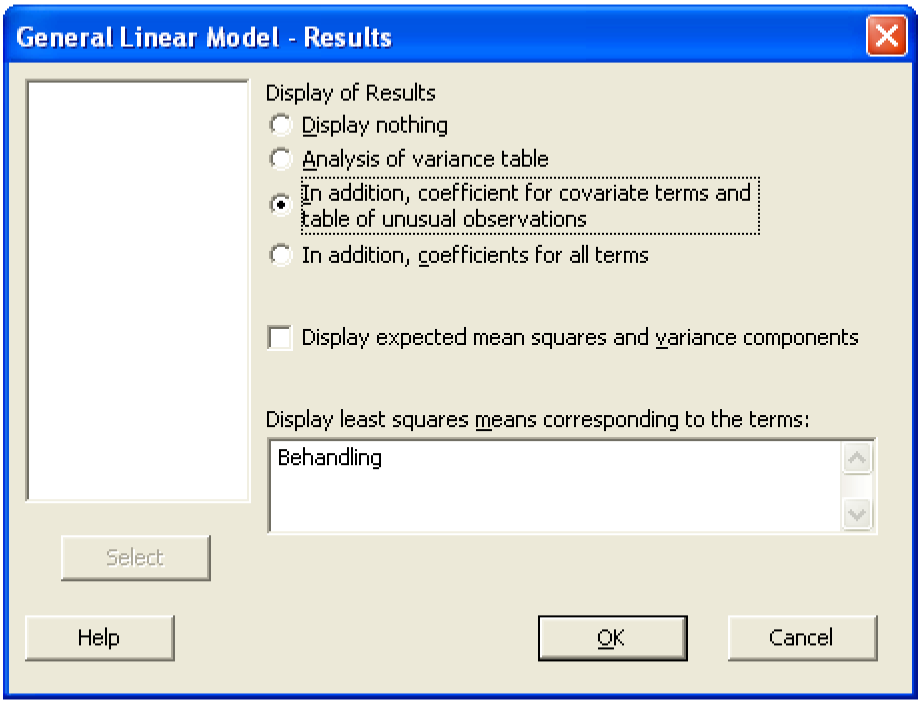
General Linear Model: N versus Block; Behandling

```
Factor       Type   Levels  Values
Block        fixed       3  I; II; III
Behandling   fixed       4  A; B; C; D


Analysis of Variance for N, using Adjusted SS for Tests

Source      DF  Seq SS  Adj SS  Adj MS      F      P
Block        2   5.365   5.365   2.683   1.52  0.293
Behandling   3  67.147  67.147  22.382  12.66  0.005
Error        6  10.605  10.605   1.767
Total       11  83.117


S = 1.32945   R-Sq = 87.24%   R-Sq(adj) = 76.61%
```

From this output you can see that there is a significant difference between the treatments (the P-value for Behandling is 0.005, which means that it is significiant at level 0.01). From the output you also can see that it was no big difference between the blocks.

The output contains no information about the means for different treatments, and to have this you can use what is called least squares means. If the design is balanced the result is identical to the means, but if the design is unbalanced, least squares means tries to compensate for the unbalance. These least square means are produced by choosing Results...

and then in Display least squares means corresponding to the terms: you fill in what way you want to calculate your mean, in this example the variable describing your treatment is Behandling.



General Linear Model - Results

Display of Results

○ Display nothing

○ Analysis of variance table

● In addition, coefficient for covariate terms and table of unusual observations

○ In addition, coefficients for all terms

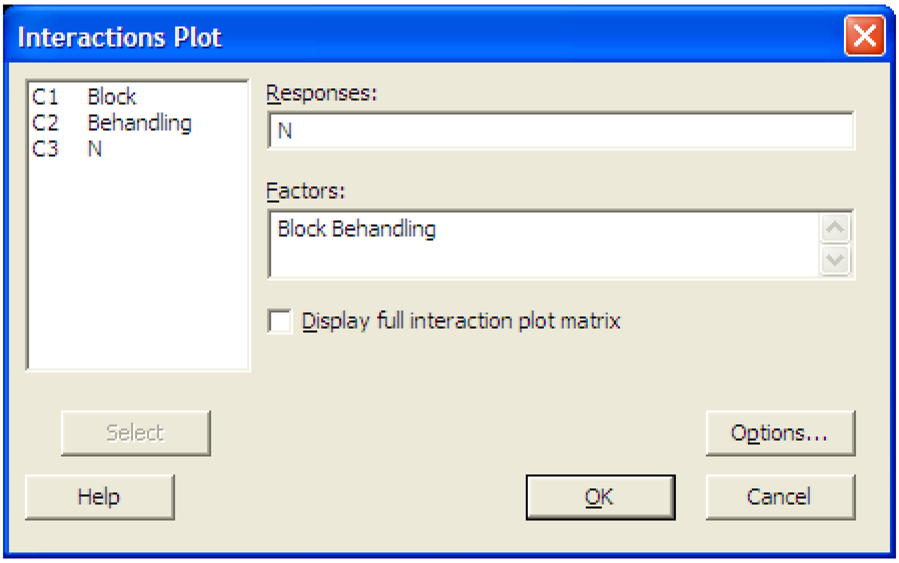
□ Display expected mean squares and variance components

Display least squares means corresponding to the terms: Behandling

If you want a picture of the dataset you can do an interactions plot in the General Linear Model… but you can also find it separately in the ANOVA-menu:

Stat ▶ ANOVA ▶ Interactions Plot…

You enter



Interactions Plot

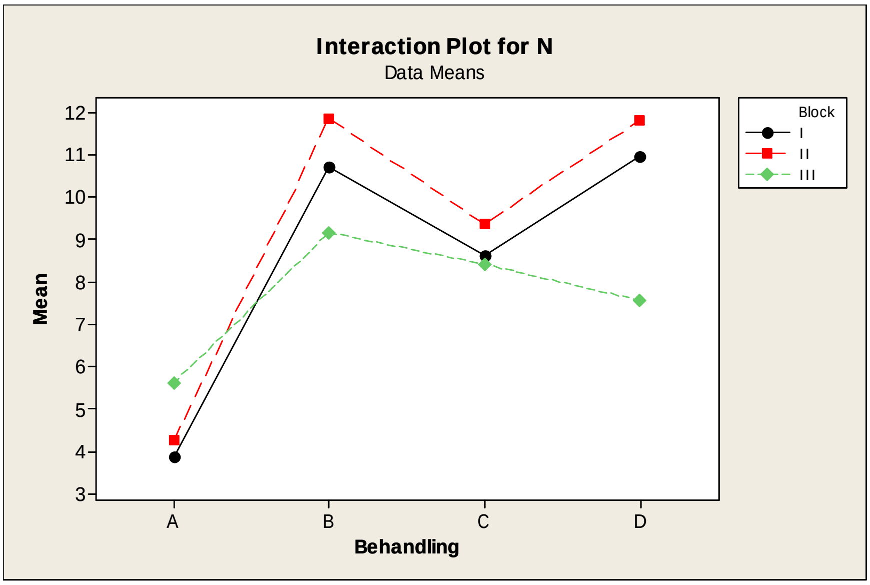
C1 Block
C2 Behandling
C3 N

Responses: N

Factors: Block Behandling

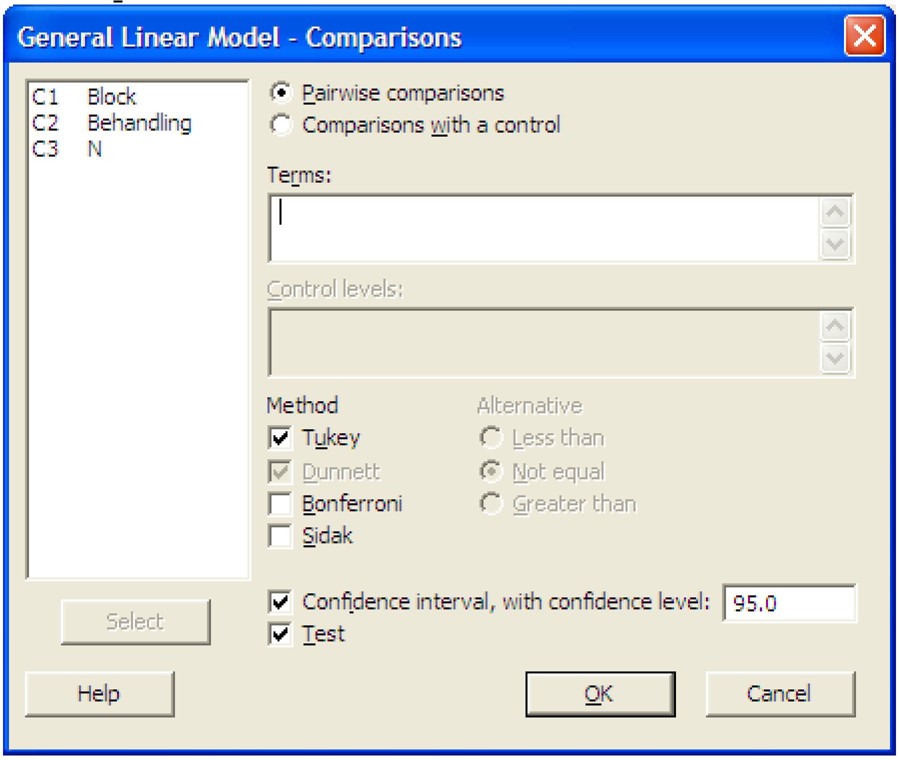
□ Display full interaction plot matrix

and the result is



Even if the axis for the mean starts at 3 and therefore is a bit misleading, you can see that there is no big difference between the blocks, but in all blocks you have the same pattern between the treatments. If you look at the figure you can guess that treatment A will have a significantly smaller mean than the other treatments, but it is not obvious whether there is a difference between the other treatments.

To compare the different levels you once again use General Linear Model… but chooses this time the button Comparisons… , and because you want to compare the treatments you write that in the box for Terms:



General Linear Model - Comparisons

C1 Block
C2 Behandling
C3 N

● Pairwise comparisons
○ Comparisons with a control

Terms:

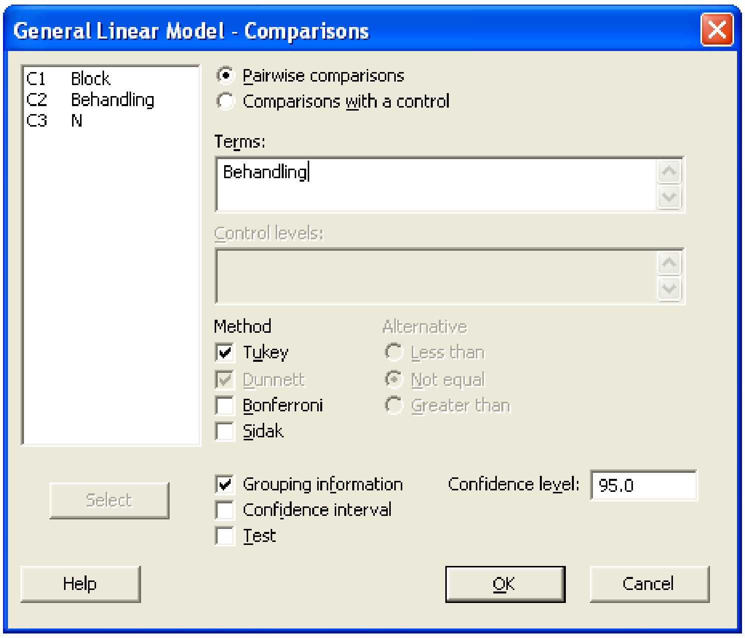
Control levels:

Method
☑ Tukey
☑ Dunnett
☐ Bonferroni
☐ Sidak

Alternative
○ Less than
● Not equal
○ Greater than

☑ Confidence interval, with confidence level: 95.0
☑ Test

You have to fill in the box at Terms: to get any result at all, and then you can decide what type of test you will do (Dunnett’s test is when you have a control and is therefore grey if you haven’t used Comparison with a control). Also note that Fishers test not is an alternative, Minitab doesn’t like to use it when you have a more complicated model. If you fill in as



C1 Block
C2 Behandling
C3 N

● Pairwise comparisons
○ Comparisons with a control

Terms: Behandling

Control levels:

Method
☑ Tukey
☑ Dunnett
□ Bonferroni
□ Sidak

Alternative
○ Less than
● Not equal
○ Greater than

☑ Grouping information
□ Confidence interval
□ Test

Confidence level: 95.0

the result is (below the result of the ANOVA we already has produced)

```
Grouping Information Using Tukey Method and 95.0% Confidence

Behandling  N  Mean  Grouping
B           3  10.6  A
D           3  10.1  A
C           3   8.8  A
A           3   4.6    B
```

In conclusion: Treatment A is significantly different from the other treatments, but there is no significant difference between B, C och D.

It is often easier to look at the P-values in the table than to see if the confidence intervals cover 0 or not. If you use one-way ANOVA you only have the confidence intervals and not the P-values and therefore it is often easier to use General Linear Model… (if you don’t want to use Fisher’s test). ■

9. Correlation

If you have two-dimensional variables you can ask whether they are related. The most common value to measure this relation is the correlation coefficient, usually denoted by ρ (greek letter "rho"). There is "robust alternatives" to this so called Pearson's correlation coefficient, and one of them is Spearman's correlation coefficient. If you will use Spearman's correlation coefficient in Minitab you are asked to do the ranking by yourself. The P-value for the correlation coefficient is the P-value for the test that the correlation coefficient is 0 against the alternative that it is not 0.

Example 7. Cotton

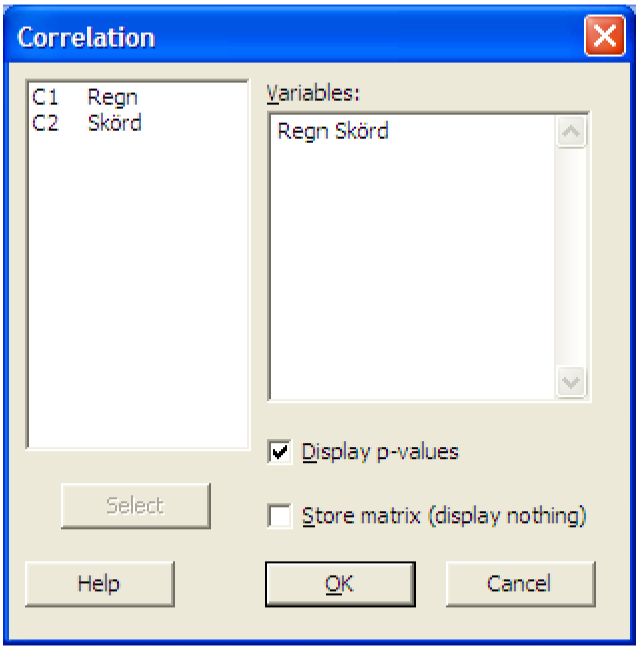
You have observed rain (”regn”) and yield of cotton (“skörd”). To see if they are correlated you calculate the correlation coefficient.

Regn (inches)	Skörd (lb/acres)
7.12	1037
63.54	380
47.38	416
45.92	427
8.68	619
50.86	388
44.46	321

The correlation coefficient is calculated in Minitab by

Stat ▶ **Basic Statistics** ▶ **Correlation**…

and if you want the P-value you mark Display p-values.



Variables: Regn Skörd

☑ Display p-values

□ Store matrix (display nothing)

You can have more than two variables in Variables: and have the correlation coefficient for all pairs of variables. In this case with two variables the result is

Correlations: Regn; Skörd

```
Pearson correlation of Regn and Skörd = -0.836
P-Value = 0.019
```

The correlation coefficient is always between -1 and 1, and here we have a negative correlation. It seems that small amount of rain give you a better yield and vice versa, and the P-value shows that the correlation coefficient is significantly different from 0. ■

10. Regression

Regression is similar to correlation, but here you have one variable that is independent (the *x*-variable) while the the response *y* is a dependendent variable (depending on *x*). There is a not exact linear relationship between the variables, but they are close to a straight line (if it is "simple linear regression"). By calculating the regression line you can have a picture of the estimated line. Here is an example.

Example 8. Carbon dioxide

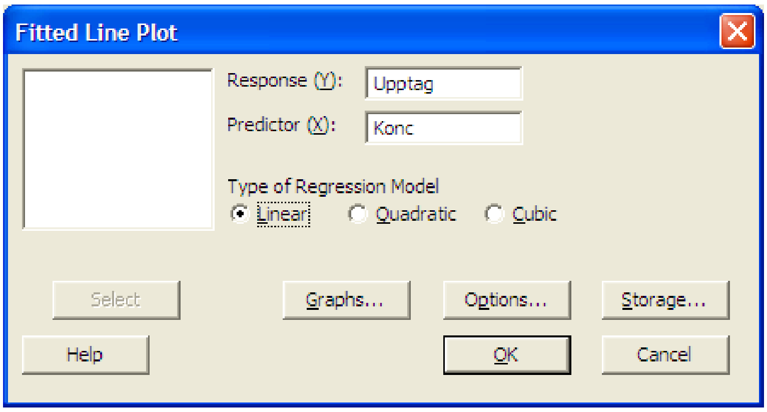
These data is for the concentration of carbon dioxide (*x*) and uptake in leafs of wheat (*y*).

Konc (*x*)	75	100	100	120	130	160	160	190	200
Upptag (*y*)	0.00	0.65	0.50	1.00	0.95	1.80	2.10	2.80	2.50

The first thing to do is to plot the data to see if a regression model makes sense in this case, but we leve it for the reader. In fact, Minitab gives you a picture of the data when you do a simple linear regression.

It seems to make sense to find a line that fits to the data set, and this is accomplished by the regression. To get the equation for the line you choose

Stat ▶ Regression ▶ Fitted Line Plot…



Response (Y): Upptag

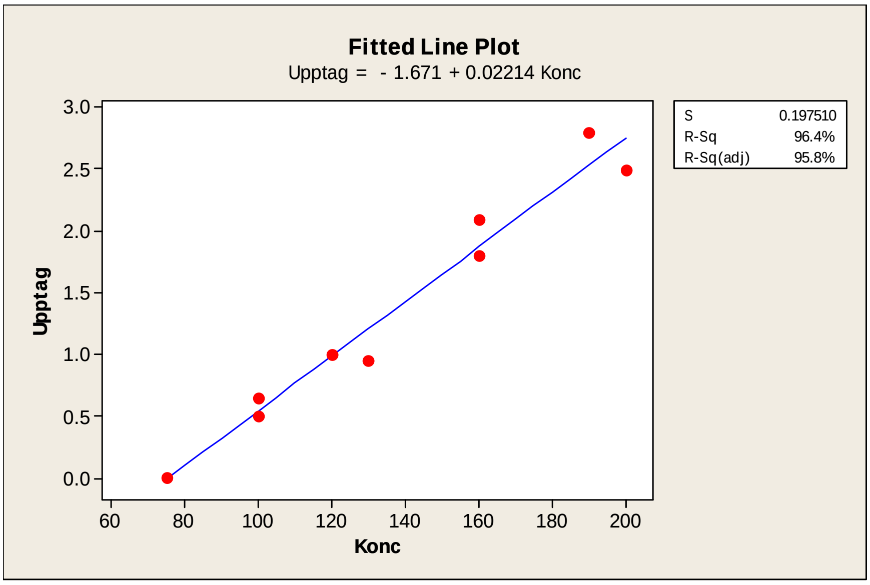
Predictor (X): Konc

Type of Regression Model: ● Linear ○ Quadratic ○ Cubic

The result is in the Session Window but also in a Graph.

```
Regression Analysis: Upptag versus Konc

The regression equation is
Upptag = - 1.671 + 0.02214 Konc


S = 0.197510   R-Sq = 96.4%   R-Sq(adj) = 95.8%


Analysis of Variance

Source      DF       SS       MS       F      P
Regression   1  7.23193  7.23193  185.38  0.000
Error        7  0.27307  0.03901
Total        8  7.50500
```

The regression equation is $y = -1.67 + 0.0221x$, and the important number is the P-value 0.000. If the regression coefficient is 0 it is no idea to use the value of x to predict the value of y. In this case the P-value tells you that the regression coefficient is significantly different from 0. You also have a picture of the data set and the line as



When you do regression you could also use the menu

Stat ▶ Regression ▶ Regression …

and this is more general. You can do more advanced regressions and the output in the Session Window is contains more information. On the other hand, you have no picture of the estimated line when you use this menu. ■

11. Test of homogeneity

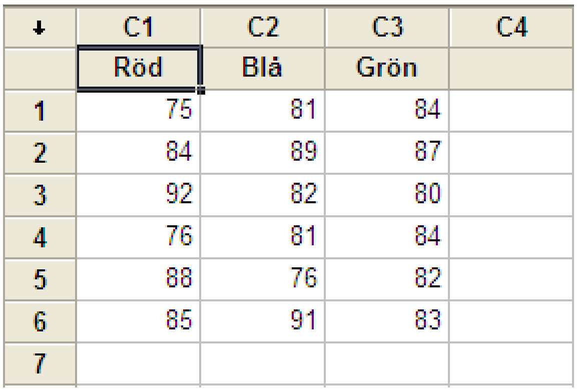
If you have categorical data (data in classes) you will find you methods under Tables in Minitab. Here you have some possibilities, but the most common test is the Chi-square-test (χ^2-test). You usually split these into goodness-of-fit-tests and tests of homogeneity. Goodness-of-fit-test might occur when you have a die and will know whether it is fair or not. Unfortunately, you cannot do this type of tests in Minitab, but you have tests of homogeneity, and therefore this section is about that type of tests. The test of homogeneity tests if two or more populations are equal. However, now the populations are not normally distributed, but contain elements that are categorized.

Example 9. Dice

There are three dice, one red, one blue and one green. To see if they are similar, you throw them 500 times and get the result

	1	2	3	4	5	6
Röd	75	84	92	76	88	85
Blå	81	89	82	81	76	91
Grön	84	87	80	84	82	83

You read the data into Minitab in three columns as

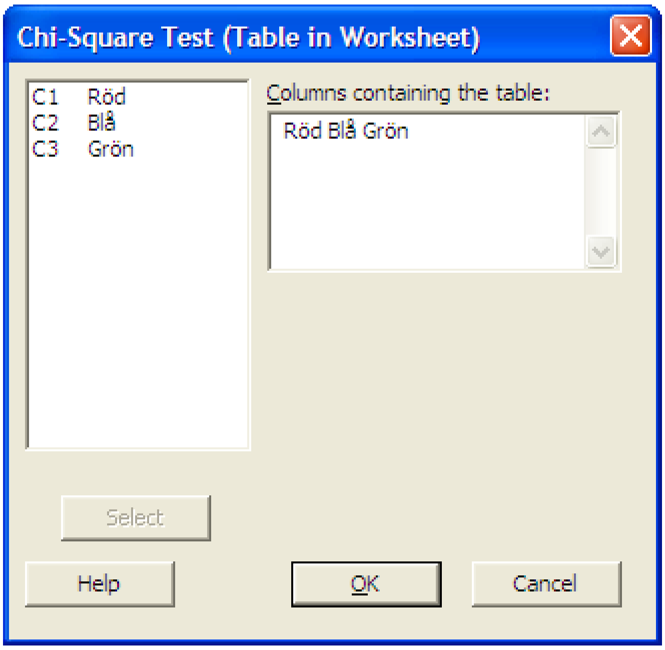


↓	C1	C2	C3	C4
	Röd	Blå	Grön	
1	75	81	84	
2	84	89	87	
3	92	82	80	
4	76	81	84	
5	88	76	82	
6	85	91	83	
7				

You choose

Stat ▶ Tables ▶ Chi-Square Test (Two-Way Table in Worksheet)…

and give the name of the columns as your variables.



C1 Röd
C2 Blå
C3 Grön

Columns containing the table: Röd Blå Grön

The result is

Chi-Square Test: Röd; Blå; Grön

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

        Röd    Blå   Grön  Total
    1    75     81     84    240
      80.00  80.00  80.00
      0.313  0.013  0.200

    2    84     89     87    260
      86.67  86.67  86.67
      0.082  0.063  0.001

    3    92     82     80    254
      84.67  84.67  84.67
      0.635  0.084  0.257

    4    76     81     84    241
      80.33  80.33  80.33
      0.234  0.006  0.167

    5    88     76     82    246
      82.00  82.00  82.00
      0.439  0.439  0.000

    6    85     91     83    259
      86.33  86.33  86.33
      0.021  0.252  0.129

Total   500    500    500   1500

Chi-Sq = 3.334; DF = 10; P-Value = 0.972
```

The interesting value is the last one, 0.972, that shows that there is no significant difference between the three dice. Note that we don’t say anything about whether they are fair or not, just that they are similar. ■