

## FILE OPERATIONS

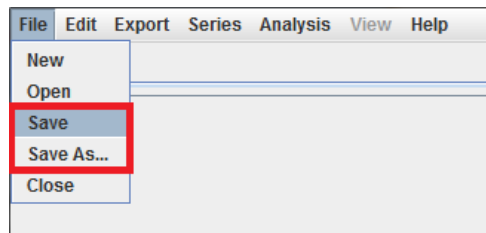
2R Data is capable of reading and writing **.2rd** files, which contain the complete description of a specific data model:

1. Series names.
2. Series data values.

These files are the means for 2R Data users to save their work, as well as the medium of distribution of data models that could be of interest to other 2R Data users.

## SAVING DATA MODELS

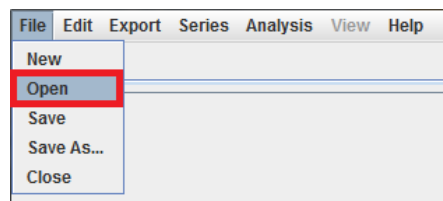
At any given time, a user can decide to save a data model's information for later use. In order to do this, the user must navigate through the **File** menu and select the **Save** or **Save As...** option:



The difference between **Save** and **Save As...** is that, while **Save** will only ask for the file's name and destination once and will then overwrite that same file on any subsequent uses, **Save As...** will ask for the file's name and destination every time it is invoked. Thus, **Save As...** is to be used whenever a user wants to save modifications made to a file without modifying the base file.

## OPENING DATA MODELS

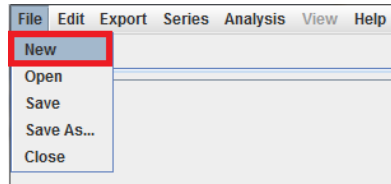
In order to load the information contained inside a **.2rd** file, the user must navigate through the **File** menu and select the **Open** option:



If no errors occur, the loaded model is shown in the **Main** tab (series names and data values).

## STARTING NEW DATA MODELS

If a user is done working with a data model and wants to start a new one, the **New** option from the **File** menu provides this functionality:



Before the new model is created, the user is given the chance to save his current work.

## SERIES MANAGEMENT

### ADDING A SERIES

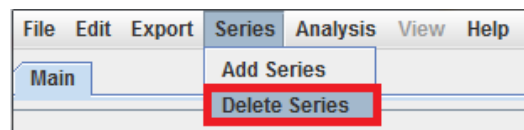
In order to add a series to a data model, a user must navigate through the **Series** menu and select the **Add Series** option:



An input dialog will appear asking for the new series' name. If the user enters a name that hasn't yet been used in the current data model, the new series shows up in the **Main** tab ready for editing.

### REMOVING A SERIES

If a data model contains a data series that is no longer needed, the user can request its elimination by navigation through the **Series** menu and selecting the **Delete Series** option:



The user is then prompted to enter the name of the series to be deleted. If the input matches one of the current data model's series, that series and its corresponding data values are eliminated from the **Main** tab.

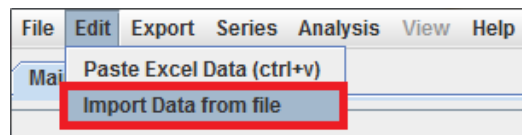
## INPUTTING AND MODIFYING DATA

2R Data offers a straight-forward editing interface in the **Main** tab. The following table summarizes the different actions that a user can carry out in 2R Data's spreadsheet editor:

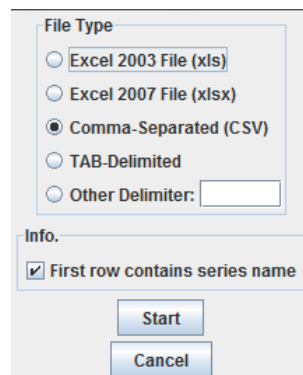
Action	How To
<b>Add values to a series</b>	<ol style="list-style-type: none"><li>1. Select the series' last cell (the one that is always blank).</li><li>2. Write a value to be added.</li><li>3. Hit the <b>ENTER</b> key.</li><li>4. Repeat steps 2 and 3 until all the values have been added.</li></ol>
<b>Modify a specific value of a series</b>	There are two ways of going about this task: <ul style="list-style-type: none"><li>• Double-click over the cell that contains the value to be edited, make the desired changes, and then hit the <b>ENTER</b> key.</li><li>• To overwrite a value, select the appropriate cell (single click) and write the new value. Then, hit the <b>ENTER</b> key.</li></ul>
<b>Delete a value from a series</b>	<ol style="list-style-type: none"><li>1. Select the cell that contains the value to be deleted (single click).</li><li>2. Hit the <b>DEL</b> or <b>DELETE</b> key.</li></ol>

## IMPORTING DATA FROM A FILE

Given that, in real life, most data analysis is done based on files generated by specialized equipment and other types of external sources, 2R Data gives users the option to import data from a wide assortment of file formats. To initiate the data import module, navigate through the **Edit** menu and select the **Import Data from file** option:



A dialog box appears asking for information regarding the external file's format:



An important thing to have in mind before importing data is that **2R Data assumes that the series are organized by columns in the external file**. Thus, each column must represent a series of data and each row must contain at most one data value of a given series.

The **First row contains series name** checkbox is to be checked if and only if you want the text from the first row to be taken as the name of the different series contained in the external file.

**Warning:** if a series from the external file has the same name as one of the existing series in the current 2R Data model, all of the data values of the existing series are deleted and the ones from the external file take their place.

When the **First row contains series name** checkbox is left unchecked, 2R Data will ask for the name to be assigned to the data of each of the columns in the external file. **Warning:** if a series from the external file is given the same name as one of the existing series in the current 2R Data model, all of the data values of the existing series are deleted and the ones from the external file take their place.

Once the dialog options are correctly set, clicking the **Start** button will open a file selection dialog which the user should use to select the file to be imported.

## FILE TYPES

The following table summarizes and exemplifies the different types of external files that 2R Data supports for data importing tasks. **Note that 2R Data assumes “.” (dot) to be the decimal separator and DOES NOT support numbers with 1000 separator [usually “,” (comma)].**

File Type	Description and Examples																																																								
<b>Excel 2003 File (.xls)</b>	Files created with Microsoft Excel and saved with the XLS extension. It is a proprietary format.  Examples: <table border="1"> <thead> <tr> <th>First row contains series names</th> <th>First row doesn't contain series names</th> </tr> </thead> <tbody> <tr> <td> <table border="1"> <thead> <tr> <th></th> <th>A</th> <th>B</th> <th>C</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>x</td> <td>y</td> <td>z</td> </tr> <tr> <td>2</td> <td>1.45</td> <td>54.9292</td> <td>1.3</td> </tr> <tr> <td>3</td> <td>2.34</td> <td>5.2</td> <td>1.7</td> </tr> <tr> <td>4</td> <td>3.3</td> <td>2</td> <td>1.2</td> </tr> <tr> <td>5</td> <td>1.3</td> <td></td> <td>1.7</td> </tr> <tr> <td>6</td> <td></td> <td></td> <td>1.9</td> </tr> </tbody> </table> </td> <td> <table border="1"> <thead> <tr> <th></th> <th>A</th> <th>B</th> <th>C</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>1.45</td> <td>54.9292</td> <td>1.3</td> </tr> <tr> <td>2</td> <td>2.34</td> <td>5.2</td> <td>1.7</td> </tr> <tr> <td>3</td> <td>3.3</td> <td>2</td> <td>1.2</td> </tr> <tr> <td>4</td> <td>1.3</td> <td></td> <td>1.7</td> </tr> <tr> <td>5</td> <td></td> <td></td> <td>1.9</td> </tr> </tbody> </table> </td> </tr> </tbody> </table>	First row contains series names	First row doesn't contain series names	<table border="1"> <thead> <tr> <th></th> <th>A</th> <th>B</th> <th>C</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>x</td> <td>y</td> <td>z</td> </tr> <tr> <td>2</td> <td>1.45</td> <td>54.9292</td> <td>1.3</td> </tr> <tr> <td>3</td> <td>2.34</td> <td>5.2</td> <td>1.7</td> </tr> <tr> <td>4</td> <td>3.3</td> <td>2</td> <td>1.2</td> </tr> <tr> <td>5</td> <td>1.3</td> <td></td> <td>1.7</td> </tr> <tr> <td>6</td> <td></td> <td></td> <td>1.9</td> </tr> </tbody> </table>		A	B	C	1	x	y	z	2	1.45	54.9292	1.3	3	2.34	5.2	1.7	4	3.3	2	1.2	5	1.3		1.7	6			1.9	<table border="1"> <thead> <tr> <th></th> <th>A</th> <th>B</th> <th>C</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>1.45</td> <td>54.9292</td> <td>1.3</td> </tr> <tr> <td>2</td> <td>2.34</td> <td>5.2</td> <td>1.7</td> </tr> <tr> <td>3</td> <td>3.3</td> <td>2</td> <td>1.2</td> </tr> <tr> <td>4</td> <td>1.3</td> <td></td> <td>1.7</td> </tr> <tr> <td>5</td> <td></td> <td></td> <td>1.9</td> </tr> </tbody> </table>		A	B	C	1	1.45	54.9292	1.3	2	2.34	5.2	1.7	3	3.3	2	1.2	4	1.3		1.7	5			1.9
First row contains series names	First row doesn't contain series names																																																								
<table border="1"> <thead> <tr> <th></th> <th>A</th> <th>B</th> <th>C</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>x</td> <td>y</td> <td>z</td> </tr> <tr> <td>2</td> <td>1.45</td> <td>54.9292</td> <td>1.3</td> </tr> <tr> <td>3</td> <td>2.34</td> <td>5.2</td> <td>1.7</td> </tr> <tr> <td>4</td> <td>3.3</td> <td>2</td> <td>1.2</td> </tr> <tr> <td>5</td> <td>1.3</td> <td></td> <td>1.7</td> </tr> <tr> <td>6</td> <td></td> <td></td> <td>1.9</td> </tr> </tbody> </table>		A	B	C	1	x	y	z	2	1.45	54.9292	1.3	3	2.34	5.2	1.7	4	3.3	2	1.2	5	1.3		1.7	6			1.9	<table border="1"> <thead> <tr> <th></th> <th>A</th> <th>B</th> <th>C</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>1.45</td> <td>54.9292</td> <td>1.3</td> </tr> <tr> <td>2</td> <td>2.34</td> <td>5.2</td> <td>1.7</td> </tr> <tr> <td>3</td> <td>3.3</td> <td>2</td> <td>1.2</td> </tr> <tr> <td>4</td> <td>1.3</td> <td></td> <td>1.7</td> </tr> <tr> <td>5</td> <td></td> <td></td> <td>1.9</td> </tr> </tbody> </table>		A	B	C	1	1.45	54.9292	1.3	2	2.34	5.2	1.7	3	3.3	2	1.2	4	1.3		1.7	5			1.9				
	A	B	C																																																						
1	x	y	z																																																						
2	1.45	54.9292	1.3																																																						
3	2.34	5.2	1.7																																																						
4	3.3	2	1.2																																																						
5	1.3		1.7																																																						
6			1.9																																																						
	A	B	C																																																						
1	1.45	54.9292	1.3																																																						
2	2.34	5.2	1.7																																																						
3	3.3	2	1.2																																																						
4	1.3		1.7																																																						
5			1.9																																																						
<b>Excel 2007 File (.xlsx)</b>	Same as the Excel 2003 file, but saved with the XLSX extension. It is an open XML-based format.																																																								
<b>Comma-Separated (CSV)</b>	Simple text-based format that uses the “,” (comma) character to separate values in one column from the values in the next column.  Examples (same data as the examples in the Excel 2003 File description): <table border="1"> <thead> <tr> <th>First row contains series names</th> <th>First row doesn't contain series names</th> </tr> </thead> <tbody> <tr> <td>x,y,z</td> <td>1.45,54.9292,1.3</td> </tr> <tr> <td>1.45,54.9292,1.3</td> <td>2.34,5.2,1.7</td> </tr> <tr> <td>2.34,5.2,1.7</td> <td>3.3,2,1.2</td> </tr> <tr> <td>3.3,2,1.2</td> <td>1.3,,1.7</td> </tr> <tr> <td>1.3,,1.7</td> <td>,,1.9</td> </tr> <tr> <td>,,1.9</td> <td></td> </tr> </tbody> </table>	First row contains series names	First row doesn't contain series names	x,y,z	1.45,54.9292,1.3	1.45,54.9292,1.3	2.34,5.2,1.7	2.34,5.2,1.7	3.3,2,1.2	3.3,2,1.2	1.3,,1.7	1.3,,1.7	,,1.9	,,1.9																																											
First row contains series names	First row doesn't contain series names																																																								
x,y,z	1.45,54.9292,1.3																																																								
1.45,54.9292,1.3	2.34,5.2,1.7																																																								
2.34,5.2,1.7	3.3,2,1.2																																																								
3.3,2,1.2	1.3,,1.7																																																								
1.3,,1.7	,,1.9																																																								
,,1.9																																																									

**TAB-Delimited**

Simple text-based format that uses the TAB character to separate values in one column from the values in the next column.

**Examples (same data as the examples in the Excel 2003 File description):**

First row contains series names			First row doesn't contain series names		
x	y	z	1.45	54.9292	1.3
1.45	54.9292	1.3	2.34	5.2	1.7
2.34	5.2	1.7	3.3	2	1.2
3.3	2	1.2	1.3		1.7
1.3		1.7			1.9
		1.9			

**Other Delimiter**

Simple text-based format that employs a user-defined character sequence to separate values in one column from the values in the next column.

**Examples using "&&" as the delimiter:**

**(same data as the examples in the Excel 2003 File description):**

First row contains series names			First row doesn't contain series names		
x&&y&&z			1.45&&54.9292&&1.3		
1.45&&54.9292&&1.3			2.34&&5.2&&1.7		
2.34&&5.2&&1.7			3.3&&2&&1.2		
3.3&&2&&1.2			1.3&&&&1.7		
1.3&&&&1.7			&&&&1.9		
&&&&1.9					

## PASTING DATA FROM EXCEL

Given that Microsoft Excel is the de facto standard for spreadsheet management, 2R Data has been enhanced to allow an intuitive flow of information with that program. Therefore, users can copy and paste data from Excel into the 2R Data user interface with ease, both by using hotkeys and by using explicit menu options.

### METHOD 1 - HOTKEYS

1. In Microsoft Excel, highlight the data values that you wish to paste into a 2R Data model:

	A	B	C
1	1	3	1
2	2	5	3
3	4	6	5
4	65	7	6
5	7	8	87
6	4		8
7	3		
8	2		
9	1		
10	8		
11	0		

2. Then, press the **CTRL** and the **C** keys **simultaneously** (ctrl+c).

3. Now, open 2R Data and start a new data model or open an existing one. In order to be able to paste the data into the program, you must already have some series in your current file.

4. In 2R Data's **Main** tab, select the cell that will act as the top-left cell of the imported data. **Warning:** the Excel data will overwrite data values of the current model if the position of the imported values coincides with the position of existing values.

x	y	z
1.0	5.0	
2.0	6.0	
3.0	7.0	
4.0		

5. Press the **CTRL** and the **V** keys **simultaneously** (ctrl+v).

6. The data is copied to 2R Data:

x	y	z
1.0	5.0	1.0
2.0	6.0	3.0
1.0	3.0	5.0
2.0	5.0	6.0
4.0	6.0	87.0
65.0	7.0	8.0
7.0	8.0	
4.0		
3.0		
2.0		
1.0		
8.0		
0.0		

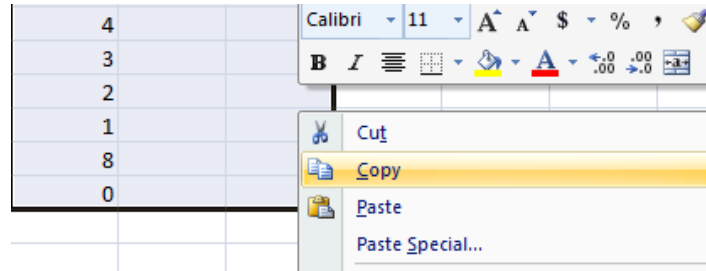
Note that, as the warning from step 4 suggests, some values from the **x** and **y** series got overwritten by the Excel data.

## METHOD 2 – EXPLICIT MENU OPTIONS

1. In Microsoft Excel, highlight the data values that you wish to paste into a 2R Data model:

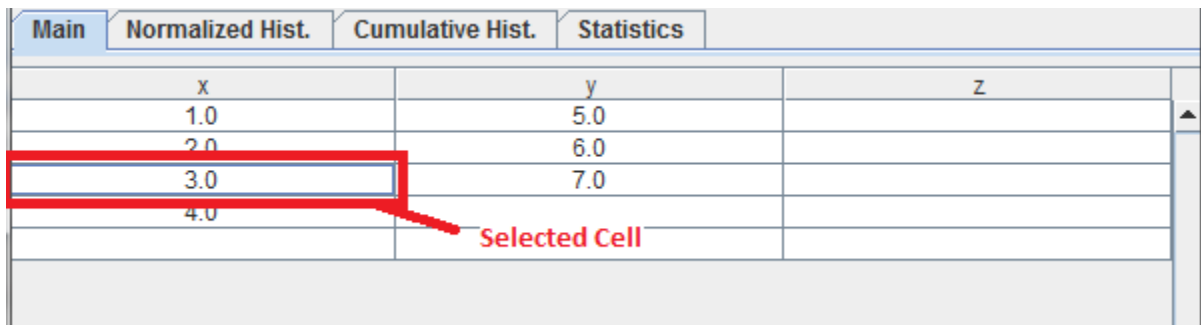
	A	B	C
1	1	3	1
2	2	5	3
3	4	6	5
4	65	7	6
5	7	8	87
6	4		8
7	3		
8	2		
9	1		
10	8		
11	0		

2. Right click over the selection and select the **Copy** option.

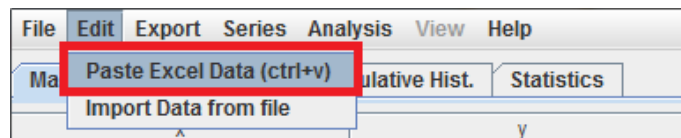


3. Now, open 2R Data and start a new data model or open an existing one. In order to be able to paste the data into the program, you must already have some series in your current file.

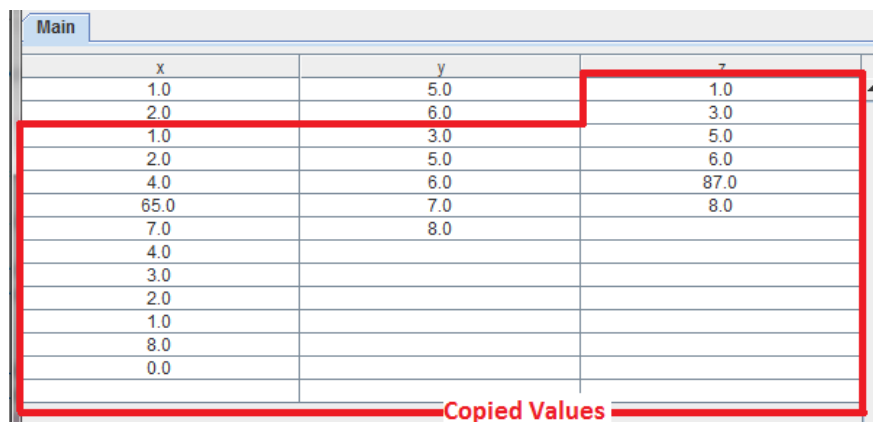
4. In 2R Data's **Main** tab, select the cell that will act as the top-left cell of the imported data. **Warning:** the Excel data will overwrite data values of the current model if the position of the imported values coincides with the position of existing values.



5. Navigate through the **Edit** menu and click over the **Paste Excel Data** option.



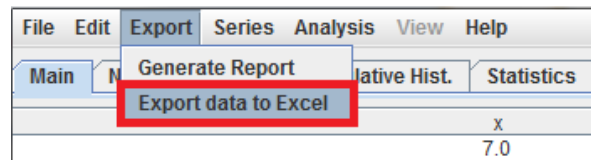
6. The data is copied to 2R Data:



Note that, as the warning from step 4 suggests, some values from the **x** and **y** series got overwritten by the Excel data.

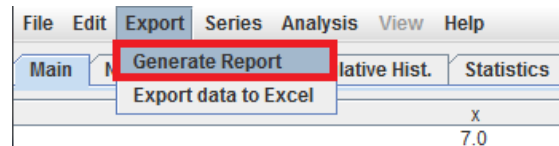
## EXPORTING DATA

Even though 2R Data provides a variety of ways to analyze one or more data series, some users might want to carry out operations on a data model's information with other software packages. For that reason, 2R Data users are given the option to export a model's series and their data values to XLSX (Excel 2007) format. The option to do this is located under the **Export** menu and is labeled as **Export data to Excel**.



## REPORT GENERATION

2R Data is capable of exporting all the useful information that results from a data analysis to PDF format. To do this, a user must navigate through the **Export** menu and select the **Generate Report** option. This option will only be enabled if a successful analysis has been completed.

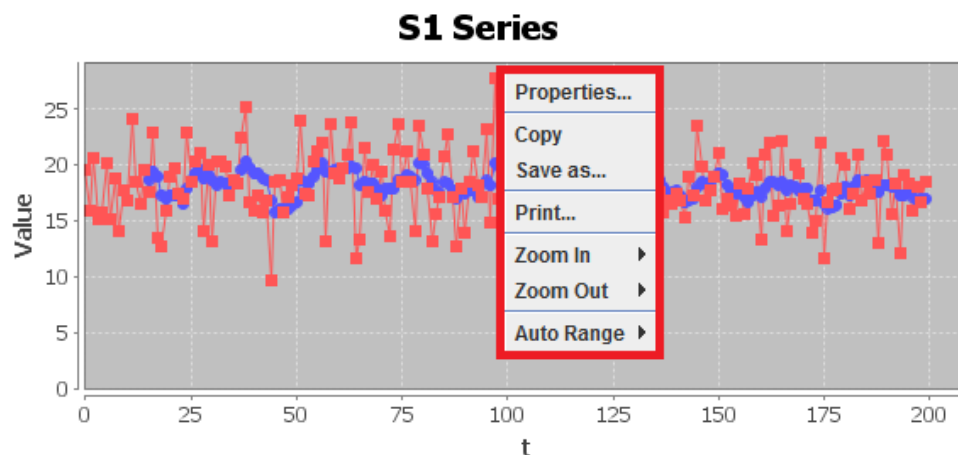


The user is then prompted for the new report's name and location before 2R Data creates the PDF file.

## BASICS

### GRAPHS

All of the graphs generated in 2R Soft provide a wide array of options in the form of a context menu. **The context menu appears when you right-click over a graph:**

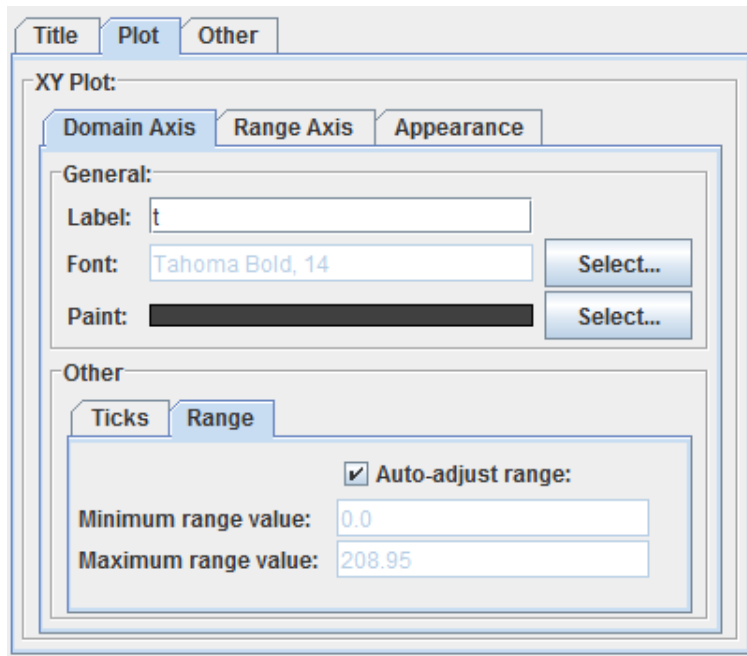




---

## PROPERTIES PANE

If you select the **Properties...** option, a properties pane appears. The properties pane lets you change the graph title, axis names, axis ranges, and font size.

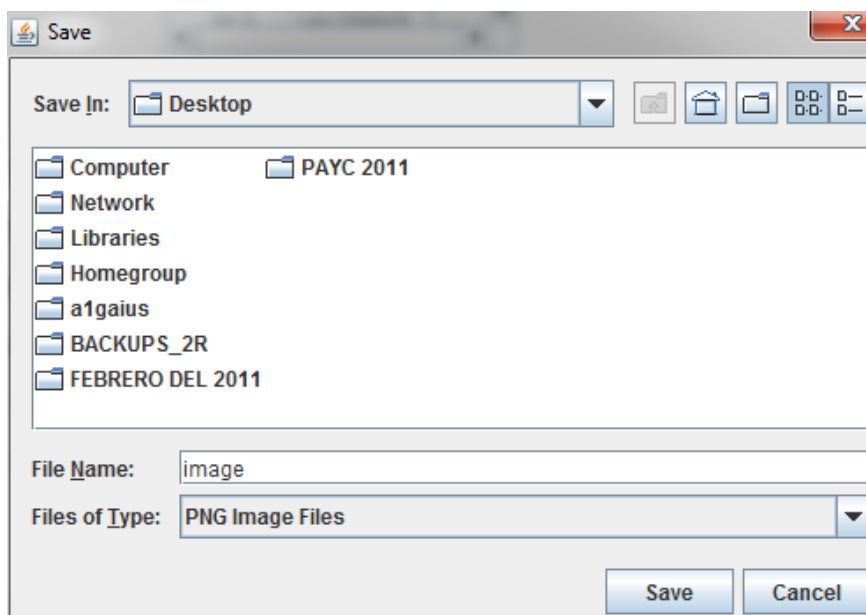


---

## COPY AND SAVE AS...

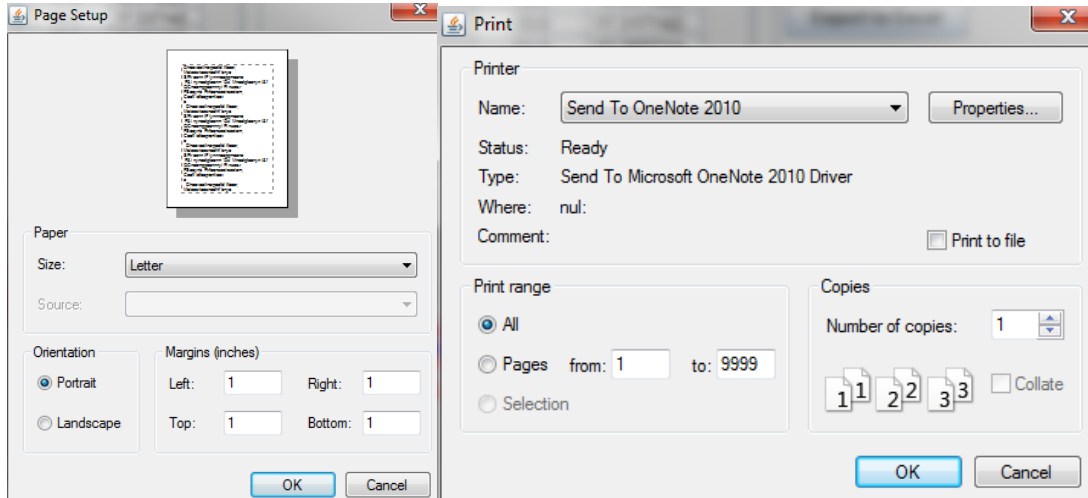
If you select **COPY**, the graph is copied to the system clipboard, so you can **PASTE** it anywhere else (Microsoft Word, Microsoft PowerPoint, etc).

Meanwhile, if **Save As...** is selected, 2R Soft will save the graph as a **PNG** image file in your hard disk after selecting the desired output folder and file name:



## PRINT

The **Print** option does just that: it sends the graph to the printer of your choice (local or networked):

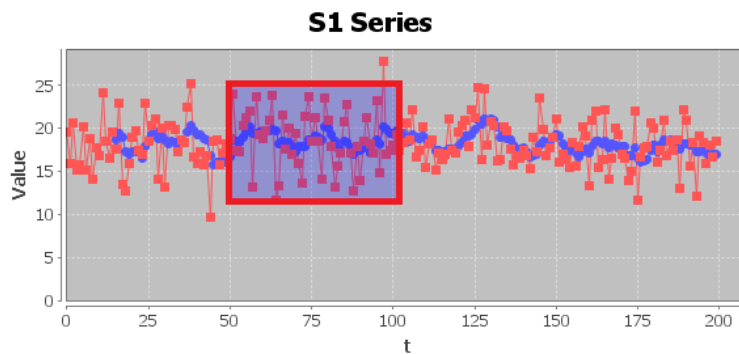


## SCALE OPTIONS

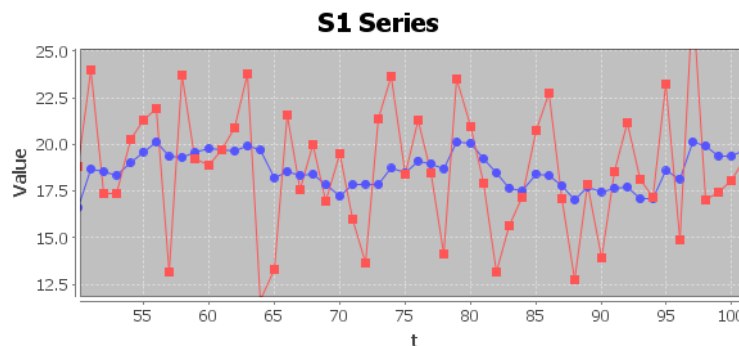
The **Auto Range**, **Zoom In**, and **Zoom Out** options are a quick way to inspect the graph. If a very specific range is needed for an axis, we highly recommend the [Properties Pane](#).

## MANUAL ZOOM IN

For user convenience, all **2R Soft** graphs support **manual zoom in by regions**. If you're interested in a specific region, **hold your left-click and drag the mouse** to generate a highlighted box around that region:

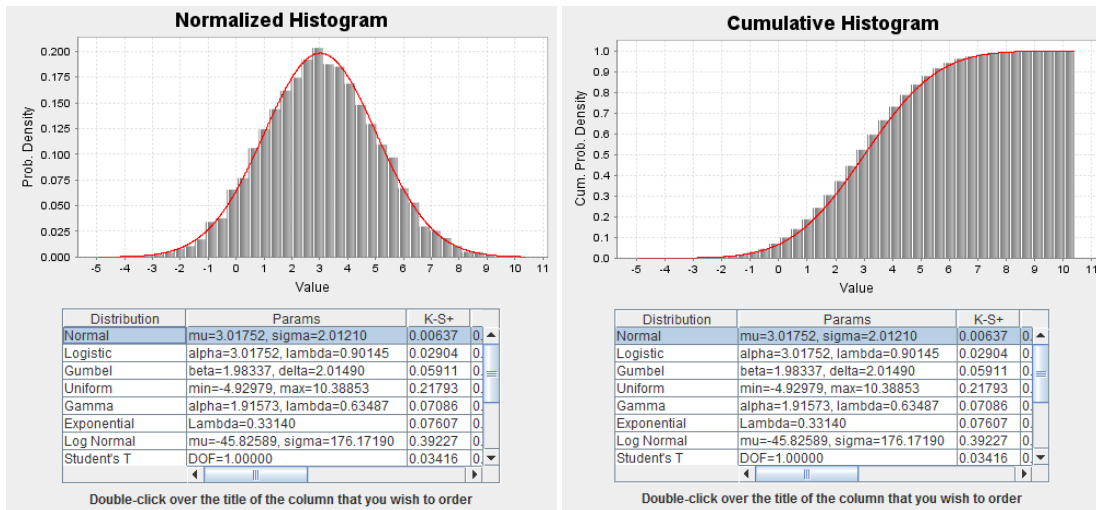


The end result:



## NORMALIZED AND CUMULATIVE HISTOGRAMS

Normalized and Cumulative histograms are found all throughout 2R Soft. These two types of graphs are of great importance, considering that they show the stochastic tendencies of a data set. With them, a goodness-of-fit table is displayed with various probability distributions and the best estimates for their parameters, along with different goodness-of-fit tests:



The distribution selected in the goodness-of-fit table is juxtaposed with the histograms (red curve).

## PROBABILITY DISTRIBUTION TYPES

When declaring a variable with a known distribution, the user can select one of many types of probability distributions.

Distribution	Description	Parameters
<b>Beta</b>	<p>The <i>beta</i> distribution has shape parameters <math>\alpha &gt; 0</math> and <math>\beta &gt; 0</math> over the interval <math>(a, b)</math>, where <math>a &lt; b</math>.</p> <p>It has density:  <math display="block">f(x) = (x - a)^{\alpha-1} (b - x)^{\beta-1} / [B(\alpha, \beta)(b - a)^{\alpha+\beta-1}]</math>                     for <math>a &lt; x &lt; b</math>, and 0 elsewhere.</p> <p>It has the following distribution function:  <math display="block">F(x) = I_{a, \theta}(x) = \int_a^x (\xi - a)^{\alpha-1} (b - \xi)^{\beta-1} / [B(\alpha, \beta)(b - a)^{\alpha+\beta-1}] d\xi, \text{ for } a &lt; x &lt; b</math></p> <p>(Simard)</p>	<p>Alpha – shape parameter, <math>\alpha &gt; 0</math></p> <p>Beta – shape parameter, <math>\beta &gt; 0</math></p> <p>a – lower bound of the interval</p> <p>b – upper bound of the interval, <math>b &gt; a</math></p>
<b>Binomial</b>	<p>The binomial distribution with parameters <math>n</math> and <math>p</math>, where <math>n</math> is a positive integer and <math>0 \leq p \leq 1</math>. Its mass function is given by:  <math display="block">p(x) = nCr(n, x)p^x(1 - p)^{n-x} = n! / [x!(n - x)!] p^x(1 - p)^{n-x} \text{ for } x = 0, 1, 2, \dots, n,</math></p> <p>and its distribution function is:  <math display="block">F(x) = \sum_{j=0}^x nCr(n, j) p^j(1 - p)^{n-j} \text{ for } x = 0, 1, 2, \dots, n,</math>                     where <math>nCr(n, x)</math> is the number of possible combinations of <math>x</math> elements chosen among a set of <math>n</math> elements.</p> <p>(Simard)</p>	<p><math>p</math> – probability of success on each trial (<math>0 \leq p \leq 1</math>)</p> <p><math>n</math> – number of trials (integer), <math>n &gt; 0</math></p>
<b>Chi Square</b>	<p>The <i>chi-square</i> distribution with <math>n</math> degrees of freedom, where <math>n</math> is a positive integer. Its density is:  <math display="block">f(x) = x^{(n/2)-1} e^{-x/2} / (2^{n/2} \Gamma(n/2)), \text{ for } x &gt; 0</math>                     where <math>\Gamma(x)</math> is the gamma function. The <i>chi-square</i> distribution is a special case of the <i>gamma</i> distribution with shape parameter <math>n/2</math> and scale parameter <math>1/2</math>.</p> <p>(Six)</p>	<p><math>n</math> – degrees of freedom (integer), <math>n &gt; 0</math></p>

<b>Deterministic</b>	Distribution that represents a constant value, <i>val</i> . Consequently: $f(x) = \begin{cases} 1 & \text{if } x = val \\ 0 & \text{if } x \neq val \end{cases}, \quad F(x) = \begin{cases} 1 & \text{if } x \geq val \\ 0 & \text{if } x < val \end{cases}$	Value – any real number
<b>Discrete Uniform</b>	The <i>discrete uniform</i> distribution over the integers in the range $[i, j]$ . Its mass function is given by: $p(x) = 1/(j - i + 1) \quad \text{for } x = i, i + 1, \dots, j$ and 0 elsewhere.  The distribution function is: $F(x) = (\text{floor}(x) - i + 1)/(j - i + 1) \quad \text{for } i \leq x \leq j$ and its inverse is: $F^{-1}(u) = i + (j - i + 1)u \quad \text{for } 0 \leq u \leq 1.$  (Simard)	Min. – lower bound (integer) Max. – upper bound (integer) (Max. > Min.)
<b>Exponential</b>	The <i>exponential</i> distribution with mean $1/\lambda$ where $\lambda > 0$ . Its density is: $f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0,$ its distribution function is: $F(x) = 1 - e^{-\lambda x}, \quad \text{for } x \geq 0,$ and its inverse distribution function is: $F^{-1}(u) = -\ln(1 - u)/\lambda, \quad \text{for } 0 < u < 1$  (Simard)	Lambda – rate parameter, lambda > 0
<b>F-Distribution</b>	The Fisher F distribution with $n$ and $m$ degrees of freedom, where $n$ and $m$ are positive integers. Its density is: $f(x) = \frac{\Gamma((n+m)/2) n^{n/2} m^{m/2} / [\Gamma(n/2) \Gamma(m/2)] x^{(n-2)/2} / (m + nx)^{(n+m)/2}, \text{ for } x > 0.$ where $\Gamma(x)$ is the gamma function  (Simard)	D.O.F. 1 – the $n$ degrees of freedom (integer), D.O.F. 1 > 0 D.O.F. 2 – the $m$ degrees of freedom (integer), D.O.F. 2 > 0
<b>Gamma</b>	The <i>gamma</i> distribution with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$ . The density is: $f(x) = \lambda^\alpha x^{\alpha-1} e^{-\lambda x} / \Gamma(\alpha), \quad \text{for } x > 0,$ where $\Gamma$ is the gamma function, defined by: $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$  In particular, $\Gamma(n) = (n - 1)!$ when $n$ is a positive integer.  (Simard)	Alpha – shape parameter, alpha > 0 Lambda – scale parameter, lambda > 0
<b>Geometric</b>	The <i>geometric</i> distribution with parameter $p$ , where $0 < p < 1$ . Its mass function is: $p(x) = p(1 - p)^x, \quad \text{for } x = 0, 1, 2, \dots$ The distribution function is given by: $F(x) = 1 - (1 - p)^{x+1}, \quad \text{for } x = 0, 1, 2, \dots$ and its inverse is: $F^{-1}(u) = \text{floor}(\ln(1 - u)/\ln(1 - p)), \quad \text{for } 0 \leq u < 1$  (Simard)	p – probability of success on each trial (0 < p < 1)
<b>Gumbel</b>	The Gumbel distribution, with location parameter $\delta$ and scale parameter $\beta \neq 0$ . Using the notation $z = (x - \delta)/\beta$ , it has density: $f(x) = e^{-z} e^{-e^{-z}} /  \beta , \quad \text{for } -\infty < x < \infty.$ and distribution function: $F(x) = e^{-e^{-z}}, \quad \text{for } \beta > 0$ $F(x) = 1 - e^{-e^{-z}}, \quad \text{for } \beta < 0$  (Simard)	Beta – scale parameter, beta $\neq$ 0 Delta – location parameter, any real number
<b>Hypergeometric</b>	The <i>hypergeometric</i> distribution with $k$ elements chosen among $l$ , $m$ being of one type, and $l - m$ of the other. The parameters $m$ , $k$ and $l$ are positive integers where $1 \leq m \leq l$ and $1 \leq k \leq l$ . Its mass function is given by: $p(x) = \frac{nCr(m, x)nCr(l - m, k - x)}{nCr(l, k)}$ for $\max(0, k - l + m) \leq x \leq \min(k, m)$ where $nCr(n, x)$ is the number of possible combinations of $x$ elements chosen among a set of $n$ elements.  (Simard)	m – number of elements of one type (integer), $m > 0$ l – total elements (integer), $l > 0$ k – number of elements chosen among l (integer), $k > 0$

<b>Logistic</b>	<p>The <i>logistic</i> distribution. It has location parameter <math>\alpha</math> and scale parameter <math>\lambda &gt; 0</math>. The density is:  <math>f(x) = (\lambda e^{-\lambda(x-\alpha)}) / ((1 + e^{-\lambda(x-\alpha)})^2)</math> for <math>-\infty &lt; x &lt; \infty</math>.  and the distribution function is:  <math>F(x) = 1 / [1 + e^{-\lambda(x-\alpha)}]</math> for <math>-\infty &lt; x &lt; \infty</math>.</p> <p>For <math>\lambda = 1</math> and <math>\alpha = 0</math>, one can write:  <math>F(x) = (1 + \tanh(x/2)) / 2</math>.</p> <p>The inverse distribution function is given by:  <math>F^{-1}(u) = \ln(u / (1 - u)) / \lambda + \alpha</math> for <math>0 &lt; u &lt; 1</math></p> <p>(Simard)</p>	<p>Alpha – location parameter, any real number  Lambda – scale parameter, lambda &gt; 0</p>
<b>Lognormal</b>	<p>The <i>lognormal</i> distribution. It has scale parameter <math>\mu</math> and shape parameter <math>\sigma &gt; 0</math>. The density is:  <math>f(x) = ((2\pi)^{-1/2} \sigma^{-1}) e^{-(\ln(x)-\mu)^2 / (2\sigma^2)}</math> for <math>x &gt; 0</math>, and 0 elsewhere.</p> <p>The distribution function is:  <math>F(x) = \Phi((\ln(x)-\mu)/\sigma)</math> for <math>x &gt; 0</math>,  where <math>\Phi</math> is the standard normal distribution function.</p> <p>Its inverse is given by:  <math>F^{-1}(u) = e^{\mu + \sigma \Phi^{-1}(u)}</math> for <math>0 &lt; u &lt; 1</math></p> <p>If <math>\ln(Y)</math> has a <i>normal</i> distribution, then <math>Y</math> has a <i>lognormal</i> distribution with the same parameters.</p> <p>(Simard)</p>	<p>log mu – scale parameter, any real number  log sigma – shape parameter, log sigma &gt; 0</p>
<b>Negative Binomial</b>	<p>The negative binomial distribution with real parameters <math>\gamma</math> and <math>p</math>, where <math>\gamma &gt; 0</math> and <math>0 &lt; p &lt; 1</math>. Its mass function is:  <math>p(x) = \Gamma(\gamma + x) / (x! \Gamma(\gamma)) p^\gamma (1 - p)^x</math>, for <math>x = 0, 1, 2, \dots</math>  where <math>\Gamma</math> is the gamma function.</p> <p>If <math>\gamma</math> is an integer, <math>p(x)</math> can be interpreted as the probability of having <math>x</math> failures before the <math>\gamma</math>-th success in a sequence of independent Bernoulli trials with probability of success <math>p</math>.</p> <p>(Simard)</p>	<p>Gamma – number of failures until the experiment is stopped, Gamma &gt; 0  p – success probability in each experiment, <math>0 \leq p \leq 1</math></p>
<b>Normal</b>	<p>The <i>normal</i> distribution. It has mean <math>\mu</math> and variance <math>\sigma^2</math>. Its density function is:  <math>f(x) = e^{-x^2 / (2\sigma^2)} / ((2\pi)^{1/2} \sigma)</math> for <math>-\infty &lt; x &lt; \infty</math>, where <math>\sigma &gt; 0</math>.</p> <p>When <math>\mu = 0</math> and <math>\sigma = 1</math>, we have the <i>standard normal</i> distribution, with corresponding distribution function:  <math>F(x) = \Phi(x) = \int_{-\infty}^x e^{-t^2/2} dt / (2\pi)^{1/2}</math> for <math>-\infty &lt; x &lt; \infty</math>.</p> <p>(Simard)</p>	<p>Mean – self-explanatory, any real number  Standard Deviation – self-explanatory, Std. Dev. &gt; 0</p>
<b>Pareto</b>	<p>The <i>Pareto</i> family, with shape parameter <math>\alpha &gt; 0</math> and location parameter <math>\beta &gt; 0</math>. The density for this type of Pareto distribution is:  <math>f(x) = \alpha \beta^\alpha / x^{\alpha+1}</math> for <math>x \geq \beta</math>, and 0 otherwise.</p> <p>The distribution function is:  <math>F(x) = 1 - (\beta/x)^\alpha</math> for <math>x \geq \beta</math>,</p> <p>and the inverse distribution function is:  <math>F^{-1}(u) = \beta(1 - u)^{-1/\alpha}</math> for <math>0 &lt; u &lt; 1</math></p> <p>(Simard)</p>	<p>Alpha – shape parameter, alpha &gt; 0  Beta – location parameter, beta &gt; 0</p>
<b>Poisson</b>	<p>The <i>Poisson</i> distribution with mean <math>\lambda \geq 0</math>. The mass function is:  <math>p(x) = e^{-\lambda} \lambda^x / (x!)</math>, for <math>x = 0, 1, \dots</math>  and the distribution function is:  <math>F(x) = e^{-\lambda} \sum_{i=0}^x \lambda^i / (i!)</math>, for <math>x = 0, 1, \dots</math></p> <p>(Simard)</p>	<p>Lambda – mean, lambda <math>\geq 0</math></p>
<b>Student's T</b>	<p>The <i>Student-t</i> distribution with <math>n</math> degrees of freedom, where <math>n</math> is a positive integer. Its density is:  <math>f(x) = [\Gamma((n+1)/2) / (\Gamma(n/2)(\pi n)^{1/2})] [1 + x^2/n]^{-(n+1)/2}</math> for <math>-\infty &lt; x &lt; \infty</math>,  where <math>\Gamma(x)</math> is the gamma function</p> <p>(Simard)</p>	<p>D.O.F – degrees of freedom (integer), D.O.F &gt; 0</p>

## Triangular

The *triangular* distribution with domain  $[a, b]$  and mode (or shape parameter)  $m$ , where  $a \leq m \leq b$ . The density function is:

$f(x) = 2(x - a)/[(b - a)(m - a)]$	for $a \leq x \leq m$ ,
$f(x) = 2(b - x)/[(b - a)(b - m)]$	for $m \leq x \leq b$ ,
$f(x) = 0$	elsewhere,

the distribution function is:

$F(x) = 0$	for $x < a$ ,
$F(x) = (x - a)^2/[(b - a)(m - a)]$	if $a \leq x \leq m$ ,
$F(x) = 1 - (b - x)^2/[(b - a)(b - m)]$	if $m \leq x \leq b$ ,
$F(x) = 1$	for $x > b$ ,

and the inverse distribution function is given by:

$F^{-1}(u) = a + ((b - a)(m - a)u)^{1/2}$	if $0 \leq u \leq (m - a)/(b - a)$ ,
$F^{-1}(u) = b - ((b - a)(b - m)(1 - u))^{1/2}$	if $(m - a)/(b - a) \leq u \leq 1$

$a$  – lower bound of the domain, any real number  
 $b$  – upper bound of the domain, any real number  
 mode – shape parameter, any real number

“ $a \leq mode \leq b$ ” must be satisfied

(Simard)

## Uniform

The *uniform* distribution over the interval  $[a, b]$ . Its density is:  
 $f(x) = 1/(b - a)$  for  $a \leq x \leq b$ , and 0 elsewhere.

Min. – lower bound  
 Max. – upper bound  
 (Max. > Min.)

The distribution function is:

$$F(x) = (x - a)/(b - a) \quad \text{for } a \leq x \leq b$$

and its inverse is:

$$F^{-1}(u) = a + (b - a)u \quad \text{for } 0 \leq u \leq 1$$

(Simard)

## Weibull

The *Weibull* distribution with shape parameter  $\alpha > 0$ , location parameter  $\delta$ , and scale parameter  $\lambda > 0$ . The density function is:  
 $f(x) = \alpha\lambda^\alpha(x - \delta)^{\alpha-1}e^{-(\lambda(x-\delta))^\alpha}$  for  $x > \delta$ .

Alpha – shape parameter, alpha > 0  
 Lambda – scale parameter, lambda > 0  
 Delta – location parameter, any real number

the distribution function is:

$$F(x) = 1 - e^{-(\lambda(x-\delta))^\alpha} \quad \text{for } x > \delta,$$

and the inverse distribution function is:

$$F^{-1}(u) = (-\ln(1 - u))^{1/\alpha}/\lambda + \delta \quad \text{for } 0 \leq u < 1$$

(Simard)

## GOODNESS-OF-FIT (GOF) STATISTICS

To decide whether to accept or reject a proposed probability distribution for the generated data, it is necessary to at least use one goodness-of-fit statistic as a criterion.

Col.	Test Type	Explanation	Critical Values																																																
A-D	Anderson-Darling	The Anderson-Darling test is defined as:	The critical values for the Anderson-Darling test are dependent on the specific distribution that is being tested. (SEMATECH2)																																																
		<p>H<sub>0</sub>: The data follow a specified distribution.</p> <p>H<sub>a</sub>: The data do not follow the specified distribution</p> <p>Test Statistic: The Anderson-Darling test statistic is defined as</p> $A^2 = -N - S$ <p>where</p> $S = \sum_{i=1}^N \frac{(2i-1)}{N} [\ln F(Y_i) + \ln(1 - F(Y_{N+1-i}))]$ <p>F is the cumulative distribution function of the specified distribution. Note that the Y<sub>i</sub> are the ordered data.</p> <p>The test is a one-sided test and the hypothesis that the distribution is of a specific form is rejected if the test statistic, A, is greater than the critical value.</p> <p>(SEMATECH2)</p>																																																	
K-S	Kolmogorov-Smirnov	The Kolmogorov-Smirnov test is defined by:	An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested. Therefore, the critical values are universal. (SEMATECH1)																																																
		<p>H<sub>0</sub>: The data follow a specified distribution</p> <p>H<sub>a</sub>: The data do not follow the specified distribution</p> <p>Test Statistic: The Kolmogorov-Smirnov test statistic is defined as</p> $D = \max_{1 \leq i \leq N} \left( F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right)$ <p>where F is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution (i.e., no discrete distributions such as the binomial or Poisson), and it must be fully specified.</p> <p>(SEMATECH1)</p>																																																	
K-S+	Kolmogorov-Smirnov+	Given a sample of n independent uniforms U <sub>i</sub> over [0, 1], the Kolmogorov-Smirnov+ statistic D <sub>n</sub> <sup>+</sup> and the Kolmogorov-Smirnov- statistic D <sub>n</sub> <sup>-</sup> , are defined by	The same from the K-S test.																																																
K-S-	Kolmogorov-Smirnov-	$D_n^+ = \max_{1 \leq j \leq n} (j/n - U_{(j)}),$ $D_n^- = \max_{1 \leq j \leq n} (U_{(j)} - (j-1)/n),$ <p>where the U<sub>(j)</sub> are the U<sub>i</sub> sorted in increasing order. Both statistics follows the same distribution function, i.e. F<sub>n</sub>(x) = P[D<sub>n</sub><sup>+</sup> &lt;= x] = P[D<sub>n</sub><sup>-</sup> &lt;= x]</p> <p>(Simard)</p>																																																	
CVM	Cramér-von Mises	Given a sample of n independent uniforms U <sub>i</sub> over [0, 1], the Cramér-von Mises statistic W <sub>n</sub> <sup>2</sup> is defined by W <sub>n</sub> <sup>2</sup> = 1/12n + ∑ <sub>j=1</sub> <sup>n</sup> (U <sub>(j)</sub> - (j-0.5)/n) <sup>2</sup> , where the U <sub>(j)</sub> are the U <sub>i</sub> sorted in increasing order. The distribution function (the cumulative probabilities) is defined as F <sub>n</sub> (x) = P[W <sub>n</sub> <sup>2</sup> <= x]	As with the K-S test, the critical values are universal:																																																
			<table border="1"> <thead> <tr> <th colspan="6">Significance</th> </tr> <tr> <th>N</th> <th>0.20</th> <th>0.15</th> <th>0.10</th> <th>0.05</th> <th>0.01</th> </tr> </thead> <tbody> <tr> <td>2</td> <td>0.138</td> <td>0.149</td> <td>0.162</td> <td>0.175</td> <td>0.186</td> </tr> <tr> <td>10</td> <td>0.125</td> <td>0.142</td> <td>0.167</td> <td>0.212</td> <td>0.32</td> </tr> <tr> <td>20</td> <td>0.128</td> <td>0.146</td> <td>0.172</td> <td>0.217</td> <td>0.33</td> </tr> <tr> <td>30</td> <td>0.128</td> <td>0.146</td> <td>0.172</td> <td>0.218</td> <td>0.33</td> </tr> <tr> <td>60</td> <td>0.128</td> <td>0.147</td> <td>0.173</td> <td>0.220</td> <td>0.33</td> </tr> <tr> <td>100</td> <td>0.129</td> <td>0.147</td> <td>0.173</td> <td>0.220</td> <td>0.34</td> </tr> </tbody> </table>	Significance						N	0.20	0.15	0.10	0.05	0.01	2	0.138	0.149	0.162	0.175	0.186	10	0.125	0.142	0.167	0.212	0.32	20	0.128	0.146	0.172	0.217	0.33	30	0.128	0.146	0.172	0.218	0.33	60	0.128	0.147	0.173	0.220	0.33	100	0.129	0.147	0.173	0.220	0.34
Significance																																																			
N	0.20	0.15	0.10	0.05	0.01																																														
2	0.138	0.149	0.162	0.175	0.186																																														
10	0.125	0.142	0.167	0.212	0.32																																														
20	0.128	0.146	0.172	0.217	0.33																																														
30	0.128	0.146	0.172	0.218	0.33																																														
60	0.128	0.147	0.173	0.220	0.33																																														
100	0.129	0.147	0.173	0.220	0.34																																														
			(ReliaSoftCorp)																																																

**WG Watson G**

Given a sample of  $n$  independent uniforms  $U_i$  over  $[0, 1]$ , the  $G$  statistic is defined by

$$G_n = (n)^{1/2} \max_{1 \leq j \leq n} [j/n - U_{(j)} + \text{bar}(U)_n - 1/2]$$

$$= (n)^{1/2} (D_n^+ + \text{bar}(U)_n - 1/2),$$

where the  $U_{(j)}$  are the  $U_i$  sorted in increasing order,  $\text{bar}(U)_n$  is the average of the observations  $U_i$ , and  $D_n^+$  is the Kolmogorov-Smirnov+ statistic. The distribution function (the cumulative probabilities) is defined as  $F_n(x) = P[G_n \leq x]$

(Simard)

From 2R Soft WG CDF code:

N	Significance				
	0.20	0.15	0.10	0.05	0.01
2	0.609	0.636	0.670	0.718	0.787
10	0.692	0.728	0.773	0.843	0.979
20	0.711	0.747	0.794	0.866	1.010
30	0.719	0.755	0.802	0.875	1.021
60	0.728	0.765	0.813	0.887	1.034
100	0.734	0.770	0.818	0.893	1.041
1000	0.745	0.782	0.831	0.906	1.055
10000	0.749	0.786	0.834	0.909	1.059

**WU Watson U**

Given a sample of  $n$  independent uniforms  $u_i$  over  $[0, 1]$ , the Watson statistic  $U_n^2$  is defined by

$$W_n^2 = 1/12n + \sum_{j=1}^n [u_{(j)} - (j-0.5)/n]^2,$$

$$U_n^2 = W_n^2 - n(\text{bar}(u)_n - 1/2)^2.$$

where the  $u_{(j)}$  are the  $u_i$  sorted in increasing order, and  $\text{bar}(u)_n$  is the average of the observations  $u_i$ . The distribution function (the cumulative probabilities) is defined as  $F_n(x) = P[U_n^2 \leq x]$

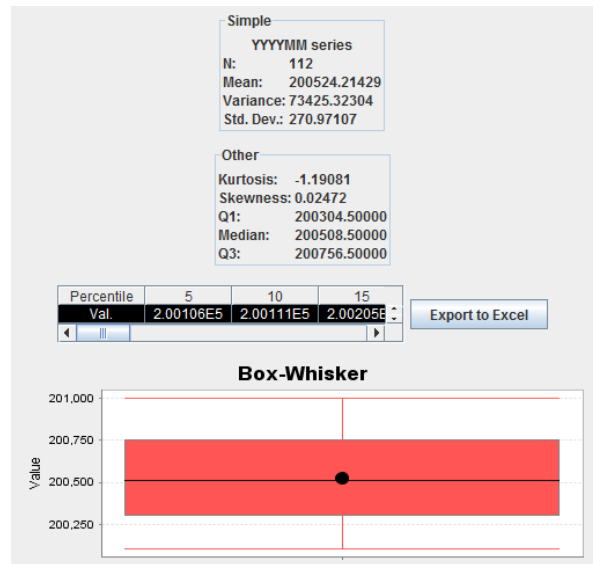
(Simard)

From 2R Soft WU CDF code:

N	Significance				
	0.20	0.15	0.10	0.05	0.01
2	0.122	0.132	0.143	0.154	0.164
10	0.116	0.130	0.150	0.183	0.255
20	0.116	0.131	0.151	0.185	0.262
30	0.117	0.131	0.151	0.185	0.264
60	0.117	0.131	0.151	0.186	0.266
100	0.117	0.131	0.152	0.186	0.267
1000	0.117	0.131	0.152	0.187	0.268
10000	0.117	0.131	0.152	0.187	0.268

## STATISTICS

When appropriate, 2R Soft calculates an array of statistics that can be both relevant and pertinent for users interested in analyzing the behavior of a data set.



The screenshot above is an example of what the user should expect to find in a **Statistics** tab. The following subsections of this document explain each of the statistics that are calculated by 2R Soft.

### MEAN

The arithmetic mean of a set of values is the quantity commonly called "the" mean or the average. Given a set of samples  $\{x_i\}$ , the arithmetic mean is: (Weisstein2)

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$



---

## VARIANCE

For a series of data, the sample variance may be computed as: (Weisstein3)

$$s_N^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

where  $\bar{x}$  is the sample mean.

Note that the sample variance  $s_N^2$  defined above is *not* an unbiased estimator for the population variance,  $\sigma^2$ . In order to obtain an unbiased estimator for  $\sigma^2$ , it is necessary to instead define a "bias-corrected sample variance": (Weisstein3)

$$s_{N-1}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$$

2R Soft uses the bias-corrected sample variance.

From the mathematical definition of variance, it is clear that this statistic expresses the dispersion of the data with respect to the mean. The larger the variance, the more spread out is the data.

---

## STANDARD DEVIATION

The standard deviation formula is very simple: it is the square root of the variance. It is the most commonly used measure of spread (Lane2). Hence, an unbiased estimator of the population standard deviation,  $\sigma$ , is given by:

$$s_{N-1} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

---

## KURTOSIS

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case. For univariate data  $Y_1, Y_2, \dots, Y_N$ , the formula for kurtosis is: (SEMATECH3)

$$kurtosis = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^4}{(N-1)s^4}$$

where  $\bar{Y}$  is the mean,  $s$  is the standard deviation, and  $N$  is the number of data points.

---

## SKEWNESS

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. For univariate data  $Y_1, Y_2, \dots, Y_N$ , the formula for skewness is: (SEMATECH3)

$$skewness = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^3}{(N - 1)s^3}$$

where  $\bar{Y}$  is the mean,  $s$  is the standard deviation, and  $N$  is the number of data points. The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail. Some measurements have a lower bound and are skewed right. For example, in reliability studies, failure times cannot be negative. (SEMATECH3)

---

## QUARTILES (Q1, MEDIAN, Q3)

The quartiles of a data set are formed by the two boundaries on either side of the median, which divide the set into four equal sections. The lowest 25% of the data being found below the first quartile value, also called the lower quartile (Q1). The median, or second quartile divides the set into two equal sections. The lowest 75% of the data set should be found below the third quartile, also called the upper quartile (Q3). These three numbers are measures of the dispersion of the data, while the mean, median and mode are measures of central tendency. (EncyclopediaOfStatistics)

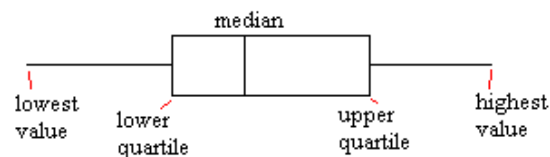
---

## BOX-WHISKER DIAGRAM

Given some data, a **box and whisker diagram** (or box plot) can be drawn to show the spread of the data. The diagram shows the quartiles of the data, using these as an indication of the spread. (MathsRevision.net)

The diagram is made up of a "box", which lies between the upper and lower quartiles. The median can also be indicated by dividing the box into two. (MathsRevision.net)

The "whiskers" are straight line extending from the ends of the box to the maximum and minimum values: (MathsRevision.net)



In 2R Soft, the Box-Whisker diagram also contains a black dot, which marks the arithmetic mean of the generated values.

## ANOVA

Variance Analysis, also known as **ANOVA**, generally makes reference to a set of experimental situations and statistical procedures for the analysis of quantitative answers of experimental units. The simplest problem tackled by way of ANOVA is known as **One-Way ANOVA**, where the input data comes from sampling more than two different populations and a single factor characterizes those populations (usually the mean). Meanwhile, **Two-Way ANOVA** is employed when two factors of interest are in play and, as expected, it translates into a more complex analysis. (Devore 1999)

### ONE-WAY ANOVA

#### INTRODUCTION

**Note:** This section is based on (Devore 1999). Refer to that text for a more in-depth treatment of the topic.

#### NULL AND ALTERNATIVE HYPOTHESES

One-way ANOVA focuses on the comparison of two or more population means. Let:

$$I = \text{number of populations being compared}$$
$$\mu_i = \text{mean of population } i$$

Then:

$$\text{Null Hypothesis, } H_0: \mu_1 = \mu_2 = \dots = \mu_I$$
$$\text{Alternative Hypothesis, } H_a: \text{at least two of the } \mu_i \text{ are different}$$

If  $I=4$ , the null hypothesis is true if the four  $\mu_i$  are identical.  $H_a$  would be true if, for example,  $\mu_1 = \mu_2 \neq \mu_3 = \mu_4$ , if  $\mu_1 = \mu_3 = \mu_4 \neq \mu_2$ , if the four  $\mu_i$  differ from each other, etc.

#### NOTATION

From this point on, we will be using the following notation:

$X_{ij}$  = the random variable representing the  $j$ -th measurement taken from the  $i$ -th population.

$x_{ij}$  = the observed value of  $X_{ij}$  when the experiment is carried out.

$J_i$  = the sample size of the  $i$ -th population

$n = \sum_i J_i$  = total number of observations in the ANOVA model

Individual sample means will be represented by  $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_I$ . Thus:

$$\bar{X}_i = \frac{\sum_{j=1}^{J_i} X_{ij}}{J_i}$$

In a similar way, the mean of all of the observations, called **grand mean**, is:

$$\bar{X}_{..} = \frac{\sum_{i=1}^I \sum_{j=1}^{J_i} X_{ij}}{n}$$

## ASSUMPTIONS

Let us assume that the  $I$  populations are normally distributed with the same variance,  $\sigma^2$ . That is, each  $X_{ij}$  is normally distributed with:

$$E(X_{ij}) = \mu_i \quad V(X_{ij}) = \sigma^2$$

The standard deviations of the  $I$  samples ( $s_1, s_2, \dots, s_I$ ) will generally differ a bit even when the corresponding  $\sigma$  are identical. In terms of sample standard deviations,  $s$ , a practical rule is: if the largest  $s$  isn't much larger than twice the value of the smallest, it's reasonable to suppose equal  $\sigma^2$ .

## TEST STATISTIC

If the null hypothesis is true, the  $J_i$  observations of every sample come from a normally distributed population with the *same* mean  $\mu$ , in which case the sample means  $\bar{x}_1, \dots, \bar{x}_I$  must be reasonably close to each other. The testing procedure is based on the comparison of a measure of differences between the  $\bar{x}_i$  with a measure of variation calculated within each one of the samples.

Let:

$$\begin{aligned} SST &= \sum_{i=1}^I \sum_{j=1}^{J_i} (X_{ij} - \bar{X}_{..})^2 && \text{with } n - 1 \text{ degrees of freedom} \\ SSTr &= \sum_{i=1}^I \sum_{j=1}^{J_i} (\bar{X}_{i.} - \bar{X}_{..})^2 && \text{with } I - 1 \text{ degrees of freedom} \\ SSE &= SST - SSTr && \text{with } n - I \text{ degrees of freedom} \end{aligned}$$

Then, the **mean square of treatments** is given by:

$$MSTr = \frac{SSTr}{I - 1}$$

And the **mean square of error** is:

$$MSE = \frac{SSE}{n - I}$$

**The test statistic for one-way ANOVA is:**

$$F = \frac{MSTr}{MSE}$$

## THE F TEST

When  $H_0$  is true:

$$E(MSTr) = E(MSE) = \sigma^2$$

Meanwhile, when  $H_0$  is false:

$$E(MSTr) > E(MSE) = \sigma^2$$

That is, both statistics are unbiased for estimating the common population variance  $\sigma^2$  when  $H_0$  is true, but  $MSTr$  tends to overestimate  $\sigma^2$  when  $H_0$  is false.

When  $H_0$  is true and the basic ANOVA assumptions are satisfied, the F statistic has an F distribution with  $v_1 = I - 1$  and  $v_2 = n - I$ . If the calculated value of F is represented by  $f$ , the rejection region  $f \geq F_{\alpha, I-1, n-I}$  specifies an F-test with significance level of  $\alpha$ .

### TUKEY PROCEDURE

When the calculated value of F,  $f$ , in a one-way ANOVA is outside the rejection region, the analysis ends because no significant differences have been detected between the  $\mu_i$ . On the other hand, when  $H_0$  is rejected, further analysis can be made. Under such condition, it is useful to know which of the  $\mu_i$  are dissimilar with one another. The Tukey procedure provides the means to draw that sort of conclusions.

Let:

$$w_{ij} = Q_{\alpha, I, n-I} \times \sqrt{\frac{MSE}{2} \left( \frac{1}{J_i} + \frac{1}{J_j} \right)}$$

Where  $Q_{\alpha, I, n-I}$  is the critical upper tail value of the studentized distribution with I degrees of freedom in the numerator and n-I degrees of freedom in the denominator, with significance level of  $\alpha$ .

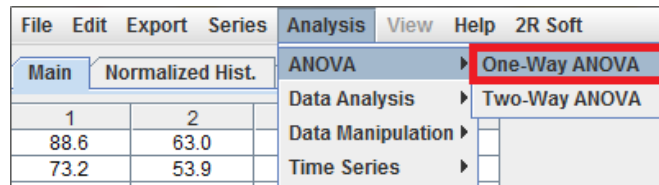
Then,  $1 - \alpha$  is *approximately* the probability that:

$$\bar{X}_i - \bar{X}_j - w_{ij} \leq \mu_i - \mu_j \leq \bar{X}_i - \bar{X}_j + w_{ij} \quad \text{for all } i \neq j$$

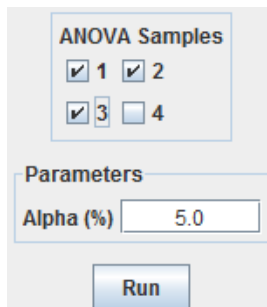
**Any pair of sample means ( $\bar{x}_i, \bar{x}_j$ ) with a distance greater than  $w_{ij}$  between them is  $\alpha$ -significantly different.**

### INPUT

In order to run a One-Way ANOVA analysis, the user must navigate through the **Analysis** menu and select the **One-Way ANOVA** option from the **ANOVA** submenu:



The user then needs to select the data series that will be treated as samples in the ANOVA analysis:



- The series (or samples) are expected to be associated with separate populations or different treatments of the same population.
- The **Alpha (%)** parameter is the significance level,  $\alpha$ , to be used for both the F test and the Tukey procedure (when appropriate).

## OUTPUT

### ANOVA TABLES

An ANOVA table is always the main result of an ANOVA analysis. A standard ANOVA table is summarized below:

Variation Source	Degrees of Freedom	Sum of Squares	Mean Square	$f$	F_max (rejection region limit)
Model (Samples)	$l-1$	SSTr	MSTr	MSTr / MSE	$F_{\alpha, l-1, n-l}$
Error	$n-l$	SSE	MSE		
Total	$n-1$	SST			

Refer to the [One-Way ANOVA Introduction](#) section for information regarding the contents of the table.

### WHEN NULL HYPOTHESIS CANNOT BE REJECTED

When the calculated value of  $F$ ,  $f$ , in a one-way ANOVA is outside the rejection region, the analysis ends because no significant differences have been detected between the  $\mu_i$ . In this case, an ANOVA table is shown and nothing else:

$f < F_{\max}(\alpha)$ , so  $H_0$  cannot be rejected with significance  $\alpha=5.0\%$ , where  $H_0: \mu_1=\mu_2=\dots=\mu_l$

Var. Source	Degrees of ...	Sum of Squ...	Mean Square	F Value	F_max (alp...
Model	1	4.63761E2	4.63761E2	5.24542E0	5.31766E0
Error	8	7.073E2	8.84125E1		
Total	9	1.17106E3			

Export to Excel

### WHEN NULL HYPOTHESIS IS REJECTED

When the calculated value of  $F$ ,  $f$ , in a one-way ANOVA falls inside the rejection region, the Tukey procedure is also performed:

**Tukey Procedure Results**  
Means with the same letter are not significantly different with  $\alpha=5.0\%$   
Critical value of studentized range=4.04609E0

Grouping	Mean	N	Series
C	79.28	5	1
A	61.54	5	2
AB	47.92	5	3
B	32.76	5	4

Pair	Dif.	Min. Sig. Dif.	Sig. Dif.?
1,2	1.774E1	1.74463E1	YES
1,3	3.136E1	1.74463E1	YES
1,4	4.652E1	1.74463E1	YES
2,3	1.362E1	1.74463E1	NO
2,4	2.878E1	1.74463E1	YES
3,4	1.516E1	1.74463E1	NO

The **Grouping** and **Series** columns contain the most useful information from the first table. If two series have a **Grouping** letter in common, they are **NOT** significantly different with significance level  $\alpha$ . In the example above, series 2 and 3 aren't significantly different, since they both contain the letter **A** in their **Grouping**. Meanwhile, series 1 is significantly different from every other series, given that the grouping letter **C** doesn't appear anywhere else.

In the second table, conclusions regarding each pair of samples are explicitly shown. If the **Sig. Dif.?** column indicates **YES** for a pair, the two samples are significantly different with significance level  $\alpha$ ; if it indicates **NO**, the pair of samples isn't significantly different with significance level  $\alpha$ . While the **Dif.** column shows the distance between the sample means,  $|\bar{x}_i - \bar{x}_j|$ , the **Min. Sig. Dif.** shows the Tukey minimum distance,  $w_{ij}$ , for the two series to be considered significantly different with significance level  $\alpha$ . Refer to the [Tukey Procedure](#) section for more information on the meaning of these results.

## TWO-WAY ANOVA

### INTRODUCTION

**Note:** This section is based on (Devore 1999). Refer to that text for a more in-depth treatment of the topic.

One-Way ANOVA is used to test for similarity of population means. Nonetheless, many experiments deal with two or more factors of interest. In Two-Way ANOVA, conclusions can be drawn regarding two factors, A and B. We use  $I$  to represent the number of levels of factor A and  $J$  to represent the number of levels of factor B, for a total of  $IJ$  possible combinations.

**Due to the fact that 2R Soft aims for simplicity, the only type of Two-Way ANOVA analysis that can be performed with this software suite is with datasets where  $K_{ij}=1$ , being  $K_{ij}$  the number of measurements made for every combination (i,j) of factors A and B.**

### NOTATION

From this point on, we will be using the following notation:

$X_{ij}$  = the random variable representing the measurement when factor A is at level "i" and factor B is at level "j".

$x_{ij}$  = the observed value of  $X_{ij}$  when the experiment is carried out.

$I$  = number of levels of factor A.

$J$  = number of levels of factor B.

$$\bar{X}_i = \text{mean of the measurements when factor A is kept at level } i = \frac{\sum_{j=1}^J X_{ij}}{J}$$

$$\bar{X}_j = \text{mean of the measurements when factor B is kept at level } j = \frac{\sum_{i=1}^I X_{ij}}{I}$$

The mean of all of the observations, called **grand mean**, is:

$$\bar{X}_{..} = \frac{\sum_{i=1}^I \sum_{j=1}^J X_{ij}}{IJ}$$

## THE MODEL

Suppose the existence of  $I$  parameters  $\alpha_1, \alpha_2, \dots, \alpha_I$  and  $J$  parameters  $\beta_1, \beta_2, \dots, \beta_J$  such that:

$$X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad (i = 1, \dots, I \text{ and } j = 1, \dots, J)$$

$$\text{where } \sum_{i=1}^I \alpha_i = 0, \sum_{j=1}^J \beta_j = 0,$$

and the  $\epsilon_{ij}$  are assumed to be independent and normally distributed with mean 0 and common variance  $\sigma^2$

In the model,  $\epsilon_{ij}$  is the random quantity by which the measurement,  $x_{ij}$ , differs from its predicted value (reality vs model).

There are two hypotheses of interest in a Two-Way ANOVA:

$$H_{0A}: \alpha_1 = \alpha_2 = \dots = \alpha_I = 0 \\ \text{versus } H_{0A}: \text{at least one } \alpha_i \neq 0$$

$$H_{0B}: \beta_1 = \beta_2 = \dots = \beta_J = 0 \\ \text{versus } H_{0B}: \text{at least one } \beta_j \neq 0$$

**The first hypothesis,  $H_{0A}$ , states that the various levels of factor A have no effect over the real average measure; the second hypothesis,  $H_{0B}$ , rules out any effect from factor B.**

## TEST PROCEDURE

The test procedure here is very similar to the [F Test](#) carried out in a One-Way ANOVA. In Two-Way ANOVA, the total variation is divided into the unexplained variability (or error), SSE, and into other two parts, SSA and SSB, that can be explained by the possible falseness of the two null hypothesis,  $H_{0A}$  and  $H_{0B}$  respectively.

The calculations involved are:

$$SST = \sum_{i=1}^I \sum_{j=1}^J (X_{ij} - \bar{X}_{..})^2 \text{ with } IJ - 1 \text{ degrees of freedom}$$

$$SSA = \sum_{i=1}^I \sum_{j=1}^J (\bar{X}_{i.} - \bar{X}_{..})^2 \text{ with } I - 1 \text{ degrees of freedom}$$

$$SSB = \sum_{i=1}^I \sum_{j=1}^J (\bar{X}_{.j} - \bar{X}_{..})^2 \text{ with } J - 1 \text{ degrees of freedom}$$

$$SSE = SST - SSA - SSB \text{ with } (I - 1)(J - 1) \text{ degrees of freedom}$$

$$MSA = \frac{SSA}{I - 1}, MSB = \frac{SSB}{J - 1}, MSE = \frac{SSE}{(I - 1)(J - 1)}$$



## F TEST

The resulting F test is:

Hypothesis	Test Statistic	Rejection Region
$H_{0A}$ versus $H_{aA}$	$f_A = \frac{MSA}{MSE}$	$f_A \geq F_{\alpha, I-1, (I-1)(J-1)}$
$H_{0B}$ versus $H_{aB}$	$f_B = \frac{MSB}{MSE}$	$f_B \geq F_{\alpha, J-1, (I-1)(J-1)}$

When  $H_{0A}$  is true and the basic ANOVA assumptions are satisfied, the  $F_A$  statistic has an F distribution with  $\nu_1 = I - 1$  and  $\nu_2 = (I - 1)(J - 1)$ . If the calculated value of  $F_A$  is represented by  $f_A$ , the rejection region  $f_A \geq F_{\alpha, I-1, (I-1)(J-1)}$  specifies an F-test with significance level of  $\alpha$ .

When  $H_{0B}$  is true and the basic ANOVA assumptions are satisfied, the  $F_B$  statistic has an F distribution with  $\nu_1 = J - 1$  and  $\nu_2 = (I - 1)(J - 1)$ . If the calculated value of  $F_B$  is represented by  $f_B$ , the rejection region  $f_B \geq F_{\alpha, J-1, (I-1)(J-1)}$  specifies an F-test with significance level of  $\alpha$ .

## TUKEY PROCEDURE

When  $H_{0A}$  or  $H_{0B}$  are rejected, further analysis can be made. Under such condition, it is useful to detect important differences between the levels of factor A or B in the samples. The Tukey procedure provides the means to draw that sort of conclusions.

Let:

$$w = \begin{cases} Q_{\alpha, I, (I-1)(J-1)} \times \sqrt{\frac{MSE}{J}} & \text{for comparisons of factor A levels} \\ Q_{\alpha, J, (I-1)(J-1)} \times \sqrt{\frac{MSE}{I}} & \text{for comparisons of factor B levels} \end{cases}$$

Where  $Q_{\alpha, x, y}$  is the critical upper tail value of the studentized distribution with x degrees of freedom in the numerator and y degrees of freedom in the denominator, with significance level of  $\alpha$ .

Then,  $1 - \alpha$  is approximately the probability that:

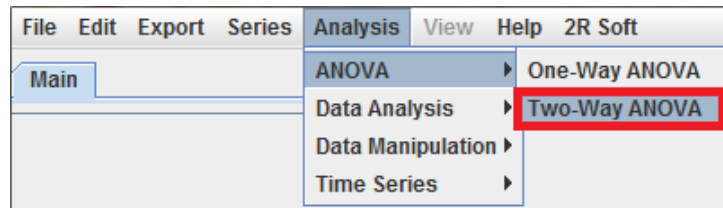
$$\text{With Factor A: } \bar{X}_i - \bar{X}_j - w \leq \mu_i - \mu_j \leq \bar{X}_i - \bar{X}_j + w \quad \text{for all } i \neq j$$

$$\text{With Factor B: } \bar{X}_i - \bar{X}_j - w \leq \mu_i - \mu_j \leq \bar{X}_i - \bar{X}_j + w \quad \text{for all } i \neq j$$

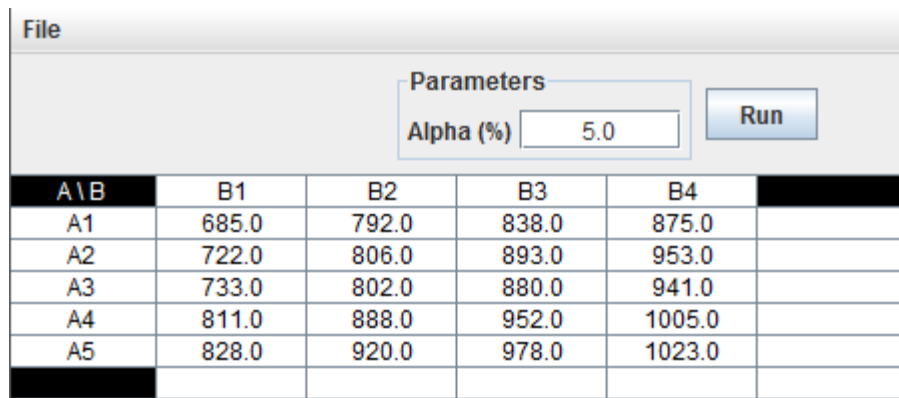
Any pair of sample means with a distance greater than  $w$  between them is  $\alpha$ -significantly different in terms of the factor being studied (A or B).

## INPUT

In order to run a One-Way ANOVA analysis, the user must navigate through the **Analysis** menu and select the **Two-Way ANOVA** option from the **ANOVA** submenu:



This action opens a new graphical interface, which receives the sample data in a matrix that makes a 1-to-1 mapping between levels of factor A and levels of factor B:



A\B	B1	B2	B3	B4	
A1	685.0	792.0	838.0	875.0	
A2	722.0	806.0	893.0	953.0	
A3	733.0	802.0	880.0	941.0	
A4	811.0	888.0	952.0	1005.0	
A5	828.0	920.0	978.0	1023.0	

- The **Alpha (%)** parameter is the significance level,  $\alpha$ , to be used for both the F tests and the Tukey procedures (when appropriate).

**Two-Way ANOVA analysis uses a unique file format (.2ra) to save the data disjointedly from the one found in 2R Data's Main tab. For this reason, the Two-Way ANOVA window contains a separate File menu. Jump to the [Saving and Loading](#) section for further information.**

## DATA EDITING FUNCTIONS

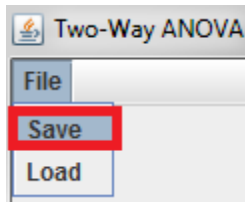
The following table summarizes the data input and modification procedures that are supported by the input matrix:

Action	How To
<b>Add factor levels to the matrix</b>	<ol style="list-style-type: none"><li>5. To add a new factor A level to the matrix, select the last cell in the B1 column (the one that is always blank). To add a new Factor B level to the matrix, select the last cell in the A1 row (the one that is always blank).</li><li>6. Write a value to be added.</li><li>7. Hit the <b>ENTER</b> key.</li></ol>
<b>Modify a specific value of the matrix</b>	<p>There are two ways of going about this task:</p> <ul style="list-style-type: none"><li>• Double-click over the cell that contains the value to be edited, make the desired changes, and then hit the <b>ENTER</b> key.</li><li>• To overwrite a value, select the appropriate cell (single click) and write the new value. Then, hit the <b>ENTER</b> key.</li></ul>

## SAVING AND LOADING TWO-WAY ANOVA DATA

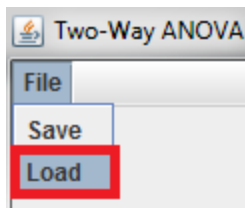
As mentioned before, the Two-Way ANOVA interface features its own, independent file management options to save and load the data matrix along with the Alpha parameter.

To save the current data, select the **Save** option from the **File** menu in the **Two-Way ANOVA Window**:



The user is then prompted to select the file name and the folder in which to save the **.2ra** file.

On the other hand, to load a data file, select the **Load** option from the **File** menu in the **Two-Way ANOVA Window**:



The user is then prompted to select the **.2ra** file to be loaded.

## OUTPUT

### ANOVA TABLES

An ANOVA table is always the main result of an ANOVA analysis. A standard ANOVA table is summarized below:

Variation Source	Degrees of Freedom	Sum of Squares	Mean Square	$f$	F_max (rejection region limit)
Factor A	I-1	SSA	MSA	MSA / MSE	$F_{\alpha, I-1, (I-1)(J-1)}$
Factor B	J-1	SSB	MSB	MSB / MSE	$F_{\alpha, J-1, (I-1)(J-1)}$
Error	(I-1)(J-1)	SSE	MSE		
Total	IJ-1	SST			

Refer to the [Two-Way ANOVA Introduction](#) section for information regarding the contents of the table.

## MAIN RESULTS

The main results after running the Two-Way ANOVA analysis are two statements regarding the rejection or acceptance of the null hypotheses, along with an ANOVA table:

$f_a \geq F_{\max}(\alpha)_a$ , so  $H_{0A}$  is rejected with significance  $\alpha=5.0\%$ , where  $H_{0A}: \alpha_1=\alpha_2=\dots=\alpha_I=0$

$f_b \geq F_{\max}(\alpha)_b$ , so  $H_{0B}$  is rejected with significance  $\alpha=5.0\%$ , where  $H_{0B}: \beta_1=\beta_2=\dots=\beta_J=0$

Var. Source	Degrees of ...	Sum of Squ...	Mean Square	F Value	F_max (alp...
Factor A	4	5.3231E4	1.33078E4	9.55673E1	3.25917E0
Factor B	3	1.16218E5	3.87392E4	2.78199E2	3.49029E0
Error	12	1.671E3	1.3925E2		
Total	19	1.7112E5			

In the image shown above, the rejection of both hypotheses is explicitly stated.

## TUKEY RESULTS

For every rejected null hypothesis, a Tukey procedure will be run with respect to the corresponding factor (A or B). Then, a table will show the Tukey groupings with respect to that specific factor. The image below is an example of a Tukey analysis with respect to Factor A:

**Tukey Procedure Results - Factor A**  
Means with the same letter are not significantly different with  $\alpha=5.0\%$   
Critical value of studentized range=4.50771E0 Minimum significant difference=2.65964E1

Grouping	Mean	N	Series
A	937.25	4	A5
A	914.0	4	A4
B	843.5	4	A2
B	839.0	4	A3
C	797.5	4	A1

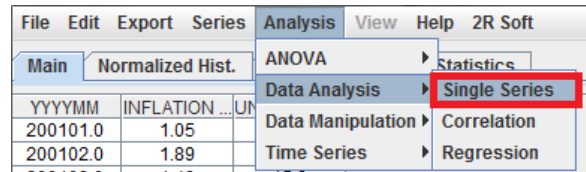
The **Grouping** and **Series** columns contain the most useful information from the Tukey table. If two series have a **Grouping** letter in common, they are **NOT** significantly different with significance level  $\alpha$ . In the example above, series A5 and A4 aren't significantly different, since they both contain the letter **A** in their **Grouping**. Meanwhile, series A1 is significantly different from every other series, given that the grouping letter **C** doesn't appear anywhere else.

Refer to the [Tukey Procedure](#) section for more information on the meaning of these results.

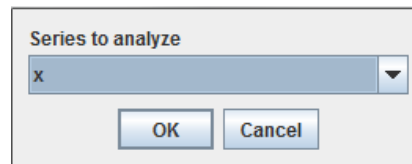
## DATA ANALYSIS

### SINGLE SERIES ANALYSIS

As the name suggests, one particular series is the input for this type of analysis. No values are generated, so 2R Data only focuses on the calculation of statistics and goodness-of-fit indicators associated with the selected series. In order to run a Single Series Analysis, the user must navigate through the **Analysis** menu and select the **Single Series** option from the **Data Analysis** submenu:



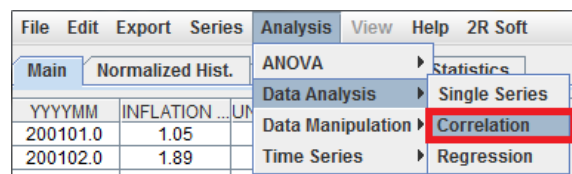
The user then has to select the series to be analyzed from a combo box and left-click over the **OK** button to start the process:



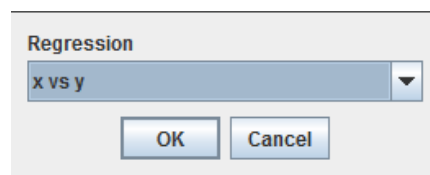
Results are shown in three separate tabs: **Normalized Hist.**, **Cumulative Hist.** and **Statistics**. Refer to the [Normalized and Cumulative Histograms](#) and [Statistics](#) sections for more information on how to interpret those results.

### CORRELATION ANALYSIS

The correlation analysis is a 2-variable analysis that tests for linearity and dependence of one variable with respect to another. To run this algorithm, a user must navigate through the **Analysis** menu and select the **Correlation** option from the **Data Analysis** submenu:



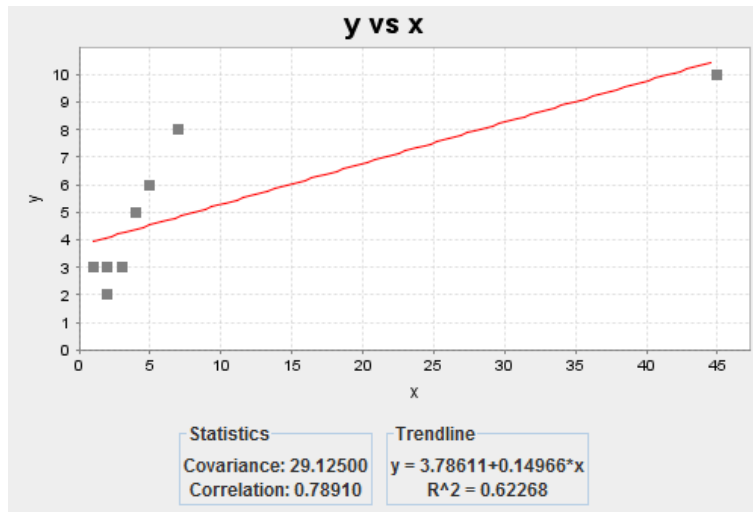
A new dialog will open with all the possible 2-variable permutations that can be analyzed:



**Warning:** in this case, **the order of the variables does matter**. The results of the “y vs x” correlation analysis aren’t necessarily equal to the ones obtained in an “x vs y” correlation analysis. The variable to the left is the **dependent**

variable, while the variable to the right is the **independent** variable. Consequently, 2R Data will try to express the **dependent** variable in terms of the **independent** variable in purely linear terms, which will be explained in the following sub-subsections.

A typical result screen for this analysis type is shown below:



In addition to a scatter plot with the dependent variable in the vertical axis and the independent variable in the horizontal axis, 2R Data shows the **Covariance** and the **Correlation Coefficient** between the two variables as well as the result from a **linear regression** and its corresponding **Coefficient of Determination (R<sup>2</sup>)**.

---

## CORRELATION COEFFICIENT

The correlation coefficient is a single number between -1 and 1 that describes the dependence between two variables. The formula used to compute the correlation coefficient of two variables from experimental data is (Lane):

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}}$$

When two variables are independent from each other, their correlation coefficient is equal to 0. On the other hand, if two variables are perfectly dependent, their correlation coefficient is equal to 1 (as one increases the other one also increases) or -1 (as one increases the other one decreases, and vice-versa).

## COVARIANCE

The difference between the covariance and the correlation coefficient is that the latter is a normalized value, while the former is not. Hence, the correlation coefficient is more valuable when drawing conclusions, since the only meaningful characteristic of a covariance is its sign (positive or negative). A positive covariance indicates that as one variable increases the other one also increases, and a negative covariance indicates that as one variable increases the other one decreases (and vice-versa). Mathematically:

$$\text{cov}(X, Y) = (\sqrt{\text{var}(x) \times \text{var}(y)}) \times \text{corr}(X, Y) \text{ (PlanetMath)}$$

Where *cov* is covariance, *corr* is correlation, and *var* is variance.

## REGRESSION AND COEFFICIENT OF DETERMINATION

Essentially, regression analysis attempts to measure the correspondence between the dependent and independent variables, thereby establishing the latter's predictive value. The proportion of unexplained variations is termed the *coefficient of determination*. (Answers.com)

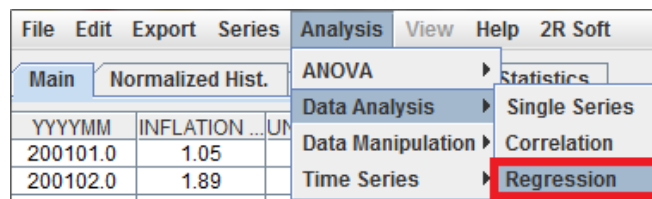
The coefficient of determination measures how good the estimated regression equation is, designated as  $R^2$  (read as R-squared). The higher the R-squared, the more confidence one can have in the equation. Statistically, the coefficient of determination represents the proportion of the total variation in the y variable that is explained by the regression equation. It has the range of values between 0 and 1, where 1 is the ideal value (AllBusiness). It is computed as:

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

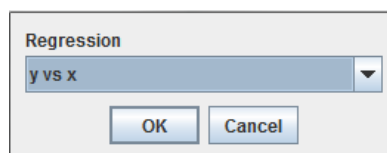
Being  $y$  the real dataset and  $y'$  the set of values predicted by the regression equation.

## REGRESSION ANALYSIS

The regression analysis performed by 2R Data is exclusively a 2-variable analysis, although some software packages are capable of running regressions with 3 or more variables. Its objective is to find an equation that explains one variable in terms of another variable in a reliable way. To run this algorithm, the user must navigate through the **Analysis** menu and select the **Regression** option from the **Data Analysis** submenu:

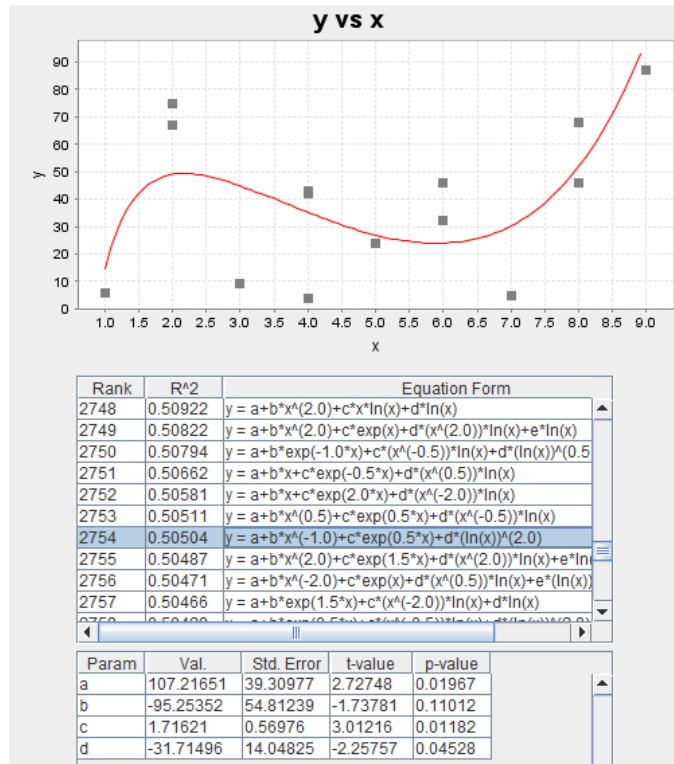


A new dialog will open with all the possible 2-variable permutations that can be analyzed:



**Warning:** in this case, **the order of the variables does matter**. The results of the “y vs x” regression analysis aren’t necessarily equal to the ones obtained in an “x vs y” regression analysis. The variable to the left is the **dependent** variable, while the variable to the right is the **independent** variable. Consequently, 2R Data will try to express the **dependent** variable in terms of the **independent** variable.

A typical result screen for this analysis type is shown below:



The result contains a scatter plot with the dependent variable in the vertical axis and the independent variable in the horizontal axis and two different tables: one with all the calculated regression equations ranked by coefficient of determination and another one that displays the values of the parameters associated with the regression selected in the first table. Additionally, the selected regression is drawn on top of the scatter plot (red curve).

## REGRESSION AND COEFFICIENT OF DETERMINATION

Essentially, regression analysis attempts to measure the correspondence between the dependent and independent variables, thereby establishing the latter's predictive value. The proportion of unexplained variations is termed the *coefficient of determination*. (Answers.com)

The coefficient of determination measures how good the estimated regression equation is, designated as R<sup>2</sup> (read as R-squared). The higher the R-squared, the more confidence one can have in the equation. Statistically, the coefficient of determination represents the proportion of the total variation in the y variable that is explained by the regression equation. It has the range of values between 0 and 1, where 1 is the ideal value (AllBusiness). It is computed as:

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

Being y' is the real dataset and y the set of values predicted by the regression equation.



## PARAMETERS AND THEIR STATISTICS

Although the coefficient of determination is a good statistic for ranking purposes, the acceptance of a regression should also take the parameters and their related errors into account. The following table explains the different statistics that 2R Data calculates for each estimated parameter in a regression equation:

Statistic	Description
<b>Std. Error</b>	Standard deviation estimate of the best estimate of the parameter.
<b>t-value</b>	A one sample t-test is a hypothesis test for answering the questions about the mean where the data is a random sample of independent observations from an underlying normal distribution $N(\mu, \sigma^2)$ , where $\sigma^2$ is unknown.

The null hypothesis for the one sample t-test is:

$H_0: \mu = \mu_0$ , where  $\mu_0$  is known.

That is, the sample has been drawn from a population of given mean and unknown variance (which therefore has to be estimated from the sample).

This null hypothesis,  $H_0$  is tested against one of the following alternative hypotheses, depending on the question posed:

$H_1: \mu$  is not equal to  $\mu$

$H_1: \mu > \mu$

$H_1: \mu < \mu$

(McColl)

The t-value for acceptance depends on the confidence level of the test (see table below). If the calculated t is greater than value shown, reject the null hypothesis: (ANU)

df	Upper tail probability $p$											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
$z^*$	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level $C$											

**p-value** The probability value (p-value) of a statistical hypothesis test is the probability of getting a value of the test statistic as extreme as or more extreme than that observed by chance alone, if the null hypothesis  $H_0$ , is true.

It is the probability of wrongly rejecting the null hypothesis if it is in fact true.

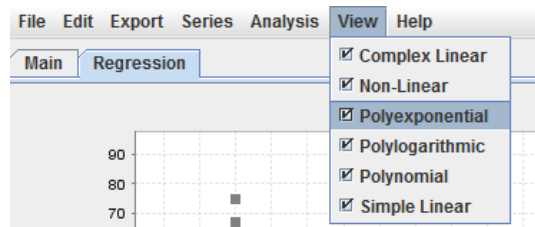
It is equal to the significance level of the test for which we would only just reject the null hypothesis. The p-value is compared with the actual significance level of our test and, if it is smaller, the result is significant. That is, if the null hypothesis were to be rejected at the 5% significance level, this would be reported as " $p < 0.05$ ".

Small p-values suggest that the null hypothesis is unlikely to be true. The smaller it is, the more convincing is the rejection of the null hypothesis. It indicates the strength of evidence for say, rejecting the null hypothesis  $H_0$ , rather than simply concluding "Reject  $H_0$ " or "Do not reject  $H_0$ ".

(McColl)

## FILTERS

2R Data tries out over 6,000 different equation forms during a regression analysis. If a user is interested in a particular subset of such equation forms, he can filter out unwanted types of regressions by using the **View** menu and deselecting the appropriate checkboxes.



The table containing regressions ranked by coefficient of determination ( $R^2$ ) will only show equations of the types that are ticked in the **View** menu.

## TYPES OF EQUATION FORMS

The table shown below summarizes the characteristics of each of the equation forms that can be filtered out through 2R Data's **View** menu.

Equation Type	Corresponding Equation Forms
<b>Complex Linear</b>	Linear combinations with 4 or more unknown coefficients (including the constant term).  Example: $a+b*x^{-1.5}+c*e^{-1.5*x}+d*\ln(x)$ where $x$ is the independent variable and $a$ , $b$ , $c$ , and $d$ are unknown parameters.
<b>Non-Linear</b>	Any equation in which an unknown parameter plays a role different from being a coefficient in a linear combination.  Example: $a*e^{b*x}$ where $x$ is the independent variable and $a$ and $b$ are unknown parameters. Clearly, $b$ is playing a role different from a linear coefficient, such as $a$ .

**Polyexponential**

$$\sum_i a_i * e^{i*x}$$

where **x** is the independent variable and  $a_i$  are unknown coefficients.

Example:

$$a+b*e^x+c*e^{2*x}+d*e^{3*x}$$

where **x** is the independent variable and **a**, **b**, **c**, and **d** are unknown parameters.

**Polylogarithmic**

$$\sum_i a_i * [\ln(x)]^i$$

where **x** is the independent variable and  $a_i$  are unknown coefficients.

Example:

$$a+b*\ln(x)+c*[\ln(x)]^2+d*[\ln(x)]^3$$

where **x** is the independent variable and **a**, **b**, **c**, and **d** are unknown parameters.

**Polynomial**

$$\sum_i a_i * x^i$$

where **x** is the independent variable and  $a_i$  are unknown coefficients.

Example:

$$a+b*x+c*x^2+d*x^3$$

where **x** is the independent variable and **a**, **b**, **c**, and **d** are unknown parameters.

**Simple Linear**

Linear combinations with less than 4 unknown coefficients (including the constant term).

Example:

$$a+b*e^{-1.5*x}$$

where **x** is the independent variable and **a** and **b** are unknown parameters.

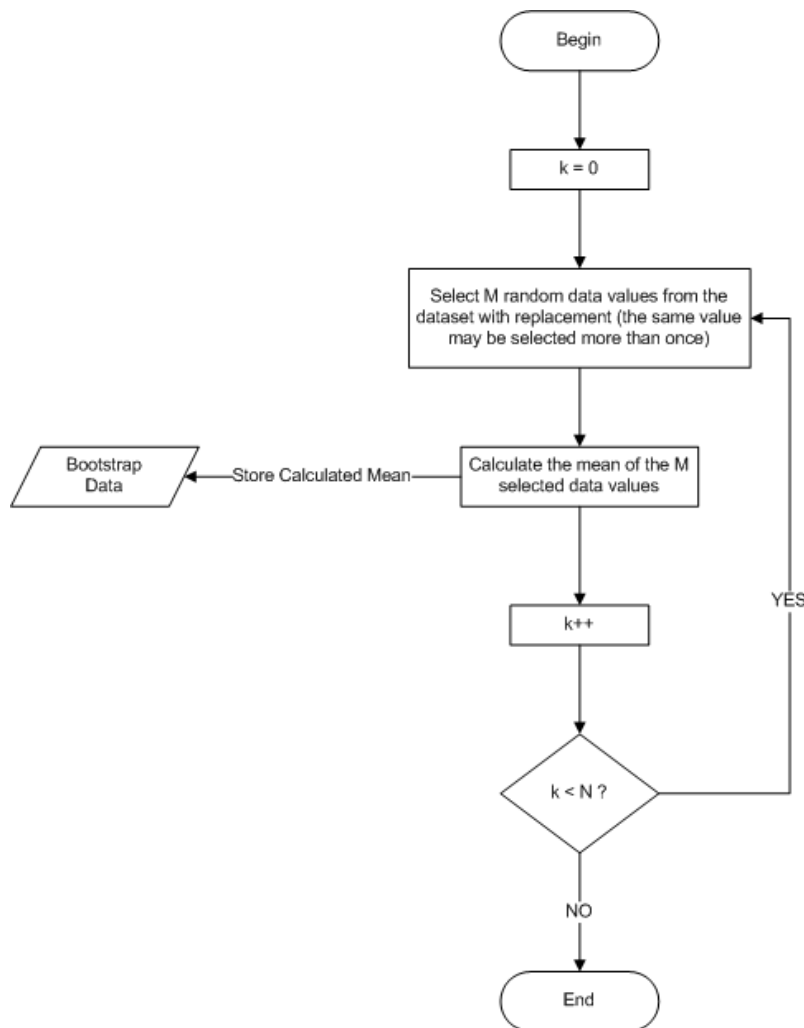
---

## DATA MANIPULATION

### BOOTSTRAP ANALYSIS

Unlike the Single Series analysis, the Bootstrap algorithm does involve generating values before calculating statistics and goodness-of-fit indicators. Bootstrapping is a resampling method. The idea: We have just one dataset. When we compute a statistic on the data, we only know that one statistic -- we don't see how variable that statistic is. The bootstrap creates a large number of datasets that we might have seen and computes the statistic on each of these datasets. Thus we get a distribution of the statistic. Key is the strategy to create data that "we might have seen" (BurnsStatistics).

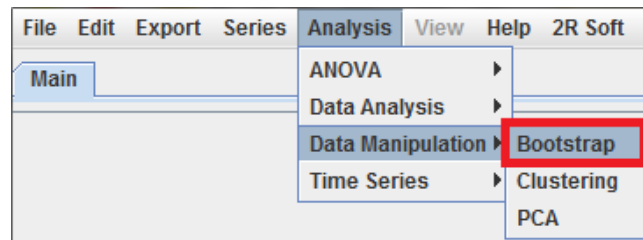
The algorithm that 2R Data runs for bootstrapping is summarized below: (let  $N$  be the number of bootstrap iterations to be executed on the dataset and let  $M$  be the number of data values in the dataset)



Given that every bootstrap iteration is completely independent from the rest, 2R Data can speed up this processor-intensive algorithm with the use of a thread pool, which can execute various iterations in parallel if more than one CPU core is present.

**Important:** in 2R Data's Bootstrap analysis, the **mean** of the series is the statistic being studied, as can be seen in the flow chart above. **Therefore, all the resulting statistics and goodness-of-fit indicators make reference to the mean of the dataset and not the dataset itself.**

To begin a Bootstrap analysis **of the mean of a data series**, the user must navigate through the **Analysis** menu and select the **Bootstrap** option from the **Data Manipulation** submenu:



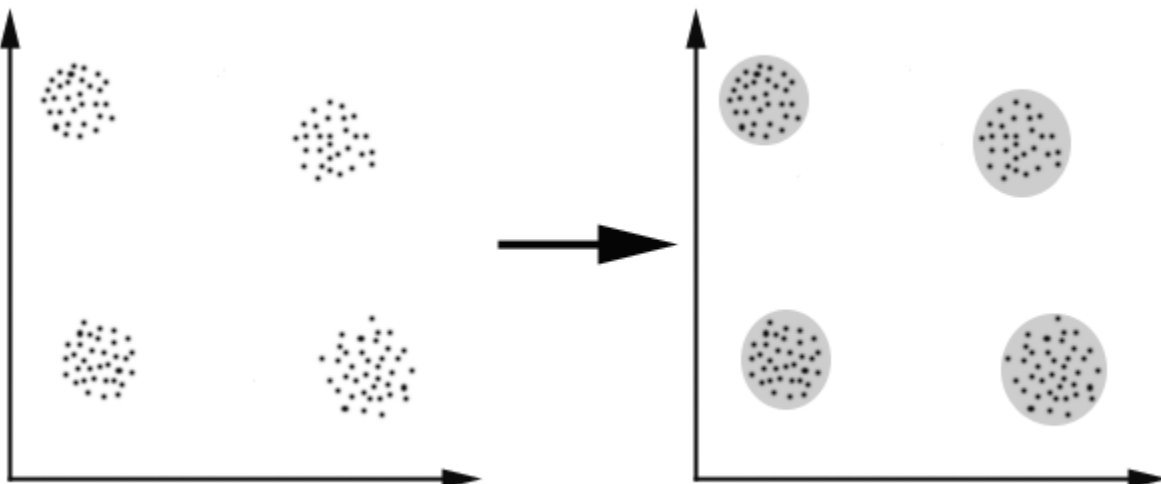
Afterwards, the user is prompted to select the appropriate series and enter the number of bootstrap iterations to be executed, N, before the algorithm begins.

Results are shown in three separate tabs: **Normalized Hist.**, **Cumulative Hist.** and **Statistics**. Refer to the [Normalized and Cumulative Histograms](#) and [Statistics](#) sections for more information on how to interpret those results.

## CLUSTERING ANALYSIS

In the words of (Dipartimento Di Elettronica E Informazione):

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A *cluster* is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. We can show this with a simple graphical example:



---

## INTRODUCTION TO CLUSTERING ALGORITHMS

Clustering algorithms can be classified into 4 different groups (Dipartimento Di Elettronica E Informazione):

1. **Exclusive clustering:** if a certain datum belongs to a definite cluster then it could not be included in another cluster.
2. **Overlapping clustering:** uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership.
3. **Hierarchical clustering:** based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted.
4. **Probabilistic clustering:** a completely probabilistic approach.

An important step in any clustering is to select a **distance measure**, which will determine how the *similarity of two elements is calculated*. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another (Temporis). Common distance measures include the Euclidean distance, Manhattan distance, Mahalanobis distance, and Hamming distance. **In the case of 2R Soft, Euclidean distance is always used as the distance measure during clustering analysis.**

---

### EUCLIDEAN DISTANCE

If  $a = (x_1, x_2, \dots, x_n)$  and  $b = (y_1, y_2, \dots, y_n)$ , the Euclidean distance from  $a$  to  $b$  is defined as:

$$d(a, b) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$$

**Geometrically, the result is equal to the length of the segment joining  $a$  and  $b$ .**

---

### DENDROGRAM / HIERARCHICAL CLUSTERING

Given a set of  $N$  items to be clustered and an  $N \times N$  distance (or similarity) matrix, the basic process of hierarchical clustering defined by (Johnson 1967) is:

1. Start by assigning each item to a cluster, so that if you have  $N$  items, you now have  $N$  clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size  $N$ .

Step 3 can be done in different ways, which is what distinguishes the various clustering algorithms (**single linkage, complete linkage, etc**) from each other.

Jump to the [Running a Clustering Analysis in 2R Soft](#) section to learn how to run Hierarchical Clustering algorithms on 2R Data datasets.

---

## K-MEANS CLUSTERING

K-Means is an exclusive clustering algorithm for partitioning  $\mathbf{N}$  data points into  $\mathbf{K}$  disjoint subsets  $\mathbf{S}_j$  containing  $\mathbf{N}_j$  data points so as to minimize the sum-of-squares criterion (Weisstein):

$$J = \sum_{j=1}^K \sum_{n \in S_j} |x_n - \mu_j|^2$$

Where  $x_n$  is a vector representing the  $n$ th data point and  $\mu_j$  is the geometric centroid (average of all coordinates) of the data points in  $\mathbf{S}_j$ .

The algorithm consists on a simple re-estimation procedure as follows (Weisstein). Initially, the data points are assigned at random to the  $\mathbf{K}$  sets.

- For step 1, the centroid is computed for each set.
- In step 2, every point is re-assigned to the cluster whose centroid is closest to that point.
- These two steps are alternated until a stopping criterion is met, i.e., when there is no further change in the assignment of the data points.

Jump to the [Running a Clustering Analysis in 2R Soft](#) section to learn how to run the K-means algorithm on 2R Data datasets.

---

## K-MEDOIDS CLUSTERING

K-medoids is a clustering algorithm that is very much like [k-means](#). The main difference between the two algorithms is the cluster center they use. K-means uses the average of all instances in a cluster, while k-medoids uses the instance that is the closest to the mean, i.e. the most 'central' point of the cluster. Using an actual point of the data set to cluster makes the k-medoids algorithm more robust to outliers than the k-means algorithm. (Abeel)

Jump to the [Running a Clustering Analysis in 2R Soft](#) section to learn how to run the K-medoids algorithm on 2R Data datasets.

---

## SPECTRAL CLUSTERING

**Note:** This sub-section is based on (Nugent and Stanberry). Refer to that source for a more in-depth treatment of the topic.

Spectral clustering algorithms partition points using eigenvectors of matrices derived from the data. They obtain a data representation in the low-dimensional space that can be easily clustered and a variety of methods exist that use the eigenvectors differently.

**2R Soft uses the NJW (Ng, Jordan, and Weiss) algorithm.**

## NJW ALGORITHM

**Motivation:** Given a set of points  $S = \{s_1, \dots, s_n\} \in R^l$ , we would like to cluster them into  $k$  subsets:

1. Form the affinity matrix  $A \in R^{n \times n}$  by applying a kernel (Gaussian, Inverse Quadratic, etc) to a measure of similarity:

With Gaussian Kernel:

$$A_{ij} = \begin{cases} e^{-\frac{\|s_i - s_j\|^2}{2\sigma^2}} & \text{if } i \neq j, \\ 0 & \text{if } i = j \end{cases} \quad \text{scaling parameter } \sigma \text{ chosen by the user}$$

With Inverse Quadratic Kernel:

$$A_{ij} = \begin{cases} \frac{1}{\sqrt{\|s_i - s_j\|^2 + c}} & \text{if } i \neq j, \\ 0 & \text{if } i = j \end{cases} \quad \text{scaling parameter } c \text{ chosen by the user}$$

In both cases,  $\|s_i - s_j\|$  stands for a basic measure of similarity. If  $\|s_i - s_j\|$  is large, points  $s_i$  and  $s_j$  are said to be "closer" to each other in accordance with the distance measure being used. **In 2R Soft, [Euclidean Distance](#) is the unit measure employed, so the result of  $\|s_i - s_j\|$  is taken to be the INVERSE of that distance. More distance between two points leads to a lower  $\|s_i - s_j\|$  (similarity).**

2. Define  $D$  a diagonal matrix whose  $(i,i)$  element is the sum of  $A$ 's row  $i$ :

$$D_{ii} = \sum_k a_{ik}$$

3. Form the matrix  $L$ , where:

$$L = D^{-1/2} A D^{-1/2}$$

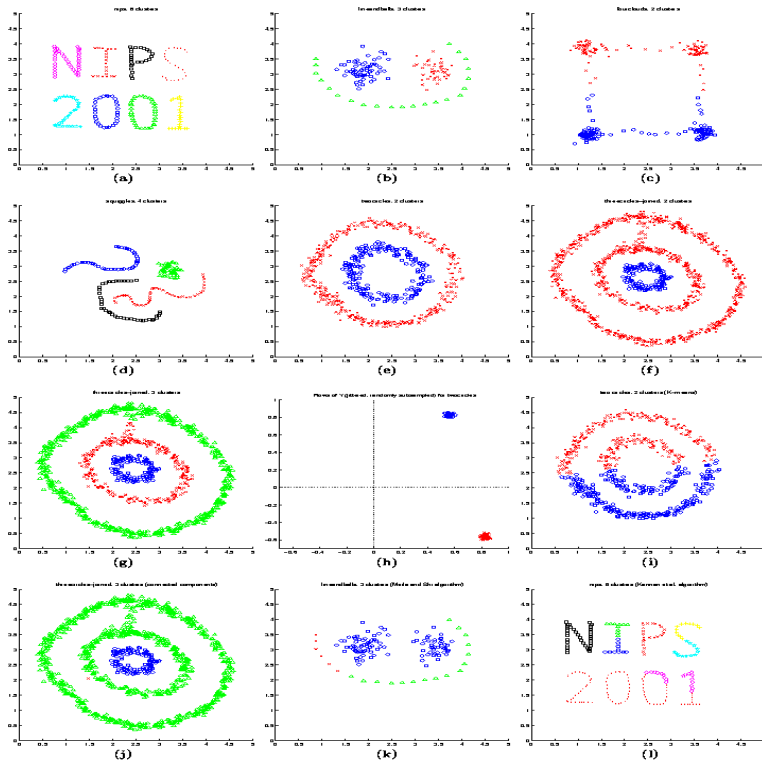
4. Stack the  $k$  largest eigenvectors of  $L$  to form the columns of the new matrix  $X$ .
5. Renormalize each of  $X$ 's rows to have unit length. The result of this operation is matrix  $Y$ :

$$Y_{ij} = X_{ij} / \left( \sum_j X_{ij}^2 \right)^{1/2}$$
$$Y \in R^{n \times k}$$

6. Cluster the rows of matrix  $Y$  as points in  $R^k$  via K-means.
7. For every  $i$ , assign point  $s_i$  to cluster  $j$  iff row  $i$  of  $Y$  was assigned to cluster  $j$ .



The answer to “If we eventually use K-means, why not just apply K-means to the original data?” is that **this method allows us to cluster non-convex regions:**

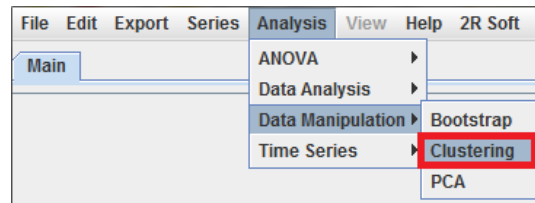


Jump to the [Running a Clustering Analysis in 2R Soft](#) section to learn how to run the NJW Spectral Clustering algorithm on 2R Data datasets.

---

## RUNNING A CLUSTERING ANALYSIS IN 2R SOFT

To begin a clustering analysis in 2R Data, the user must navigate through the **Analysis** menu and select the **Clustering** option from the **Data Manipulation** submenu:



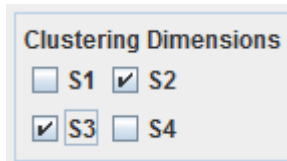
---

### INPUT

The input window's options vary depending on the algorithm to be used and on whether or not the user plans to perform a multi-level clustering. This section covers each region of the window separately.

---

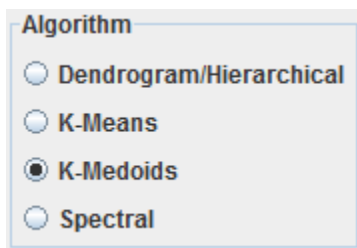
### CLUSTERING DIMENSIONS



In this region of the input window, the user must select the data series that will be part of the clustering analysis. **Each selected series will act as a unique space dimension.** In the example above, series S2 and S3 are selected, so point  $x_i$  will be taken to be a 2-dimensional coordinate  $(S2_i, S3_i)$  for every row "i" of data.

---

### ALGORITHM



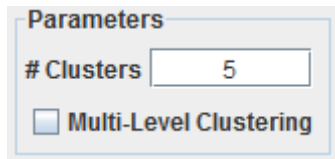
The algorithm dictates how clustering is to be carried out and the type of results that will be shown. Click on the name of a specific algorithm to be redirected to the section of this document that explains the logic behind it:

- [Dendrogram/Hierarchical](#): a dendrogram is generated as the final result. Jump to the [Dendrogram](#) section for more information.
- [K-Means](#), [K-Medoids](#) and [Spectral](#): a tree structure will show the resulting arrangement of clusters and their members. The options to analyze a specific cluster and to graph a 2-dimensional or 1-dimensional cluster are also provided. Jump to the [Tree View](#) section for more information.

## PARAMETERS (FOR K-MEANS, K-MEDOIDS AND SPECTRAL)

---

When running a one-level clustering analysis, meaning that clusters are only calculated from the source data and no sub-clusters within those clusters are to be calculated, a single parameter has to be entered:



Parameters

# Clusters

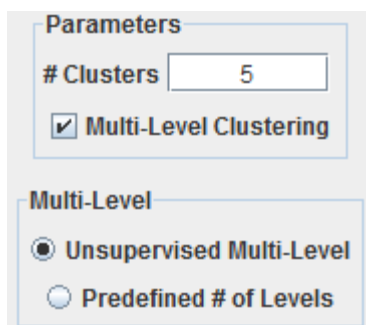
Multi-Level Clustering

- **# Clusters:** the number of clusters that the user wants to obtain from the clustering process.

## MULTI-LEVEL CLUSTERING

---

If sub-clusters within clusters are to be calculated, the user will have to activate the **Multi-Level Clustering** checkbox. This will require additional information to be entered in the **Multi-Level** region of the clustering window:



Parameters

# Clusters

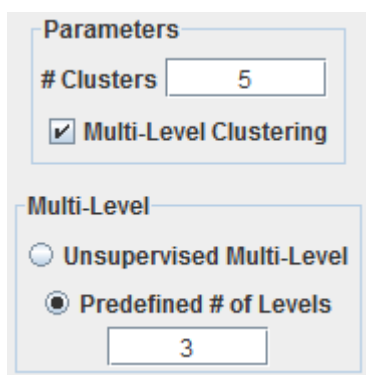
Multi-Level Clustering

Multi-Level

Unsupervised Multi-Level

Predefined # of Levels

- **Unsupervised Multi-Level:** for each calculated cluster, sub-clusters will be found until a point is reached where the sub-sub-...-sub-cluster cannot be divided into **#Clusters** clusters (5 in the screen above). Therefore, for the example above, a tree of clusters is obtained in which each “node” (cluster) has either 0 or 5 “nodes” (clusters) as childs.



Parameters

# Clusters

Multi-Level Clustering

Multi-Level

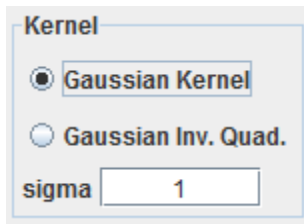
Unsupervised Multi-Level

Predefined # of Levels

- **Predefined # of Levels:** this type of multi-level clustering stops after **N** levels of clusters are found (limited depth of search). For each calculated cluster, sub-clusters will be found until a point is reached where the sub-sub-...-sub-cluster cannot be divided into **#Clusters** clusters (5 in the screen above) **or until N levels of clusters have already been calculated**. Therefore, for the example above, a tree of clusters is obtained in which each “node” (cluster) has either 0 or 5 “nodes” (clusters) as childs, **and the maximum depth of search is of sub-sub-clusters with respect to the original clusters**.

## KERNEL (SPECTRAL ONLY)

The Kernel region is displayed when the Spectral algorithm is selected.



Kernel

Gaussian Kernel

Gaussian Inv. Quad.

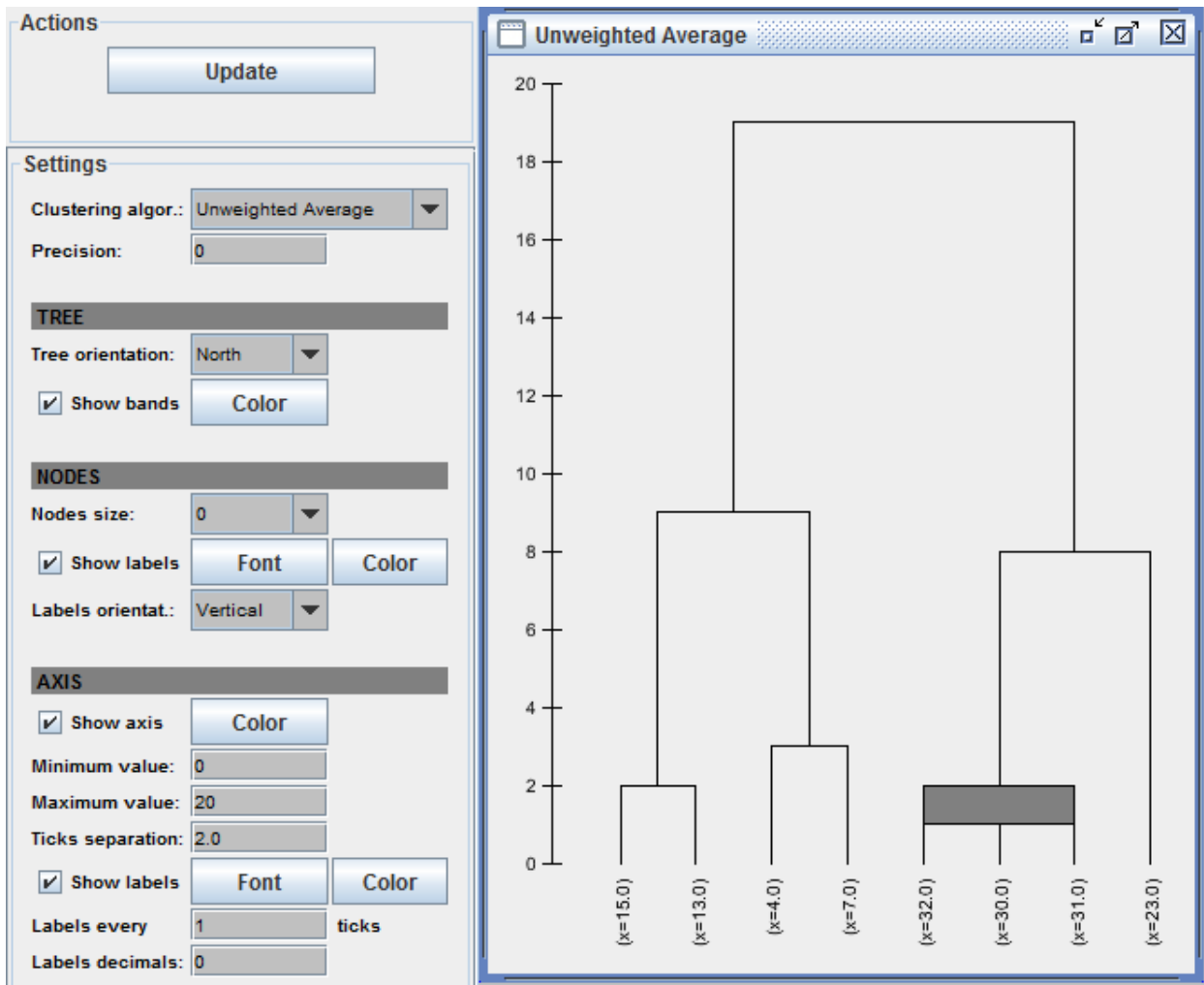
sigma

For information on what the Kernel does and on the effects of “sigma” and “c”, refer to the [Spectral Clustering](#) section.

## OUTPUT

### DENDROGRAM (HIERARCHICAL ONLY)

When a hierarchical clustering analysis is performed, a dendrogram is the end result. A new window will become visible with a variety of options. **For option changes to take effect, the “Update” button must be used.**



- **Clustering algorithm:** this setting dictates how distances are computed on step 3 of the [clustering algorithm](#).
- **Precision:** the amount of decimal places to use in the axis.
- **Tree**
  - **Tree Orientation:** defines the direction in which the dendrogram is generated.
  - **Show bands:** when selected, bands are displayed to indicate the level of heterogeneity inside the clusters.
- **Nodes**
  - **Nodes size:** the size of the dendrogram terminations.
  - **Show labels:** the numerical value of each termination is shown when this is selected.
  - **Labels Orientation:** the direction in which the labels are to be displayed.
- **Axis**
  - **Show axis:** the value axis is shown when this is selected.
  - **Show labels:** the axis is complemented with labels when this is enabled.

### TREE VIEW (FOR K-MEANS, K-MEDOIDS AND SPECTRAL)

After a non-hierarchical clustering analysis is run, the results are displayed in the form of a tree view:

The screenshot displays a tree view of clustering results on the left and a control panel on the right. The tree view shows a hierarchy of clusters. The root is 'Clusters', which contains 'Cluster 1(<dist.>=5.17031E0)', 'Cluster 2(<dist.>=9.56087E0)', and 'Cluster 3(<dist.>=4.19292E0)'. 'Cluster 1' is expanded to show a 'Members' folder containing ten data points with S2 and S3 values, and a sub-cluster 'Cluster 1\_1(<dist.>=3.03192E0)'. 'Cluster 1\_1' is further expanded to show its own 'Members' folder and two sub-clusters: 'Cluster 1\_1\_1(<dist.>=0E0)' and 'Cluster 1\_1\_2(<dist.>=1.47922E0)'. The control panel on the right has four buttons: 'Graph Real', 'Graph Fictitious', 'Analyze Real', and 'Analyze Fictitious'.

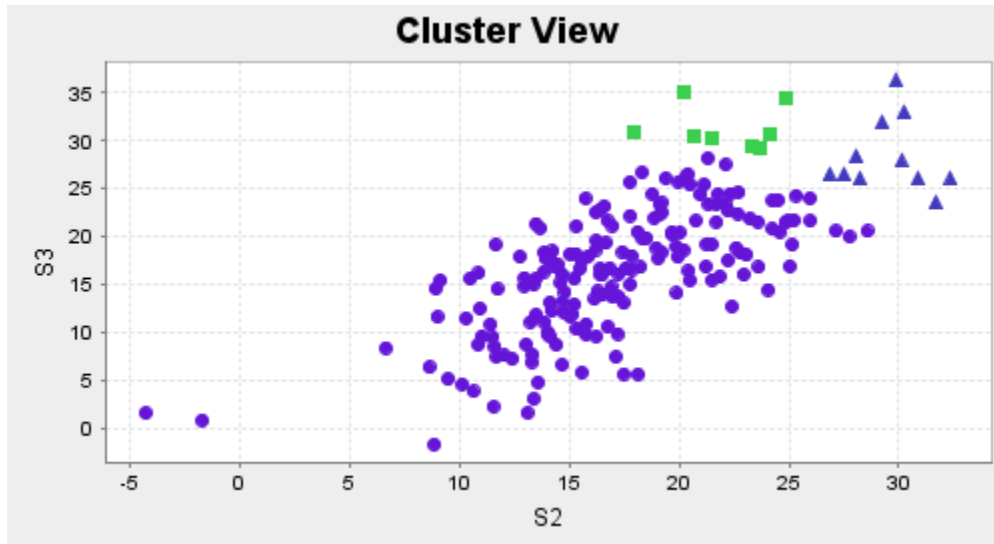
**Nodes with a folder icon (clusters and members) can be expanded and contracted by double-clicking on them.**

Every cluster contains:

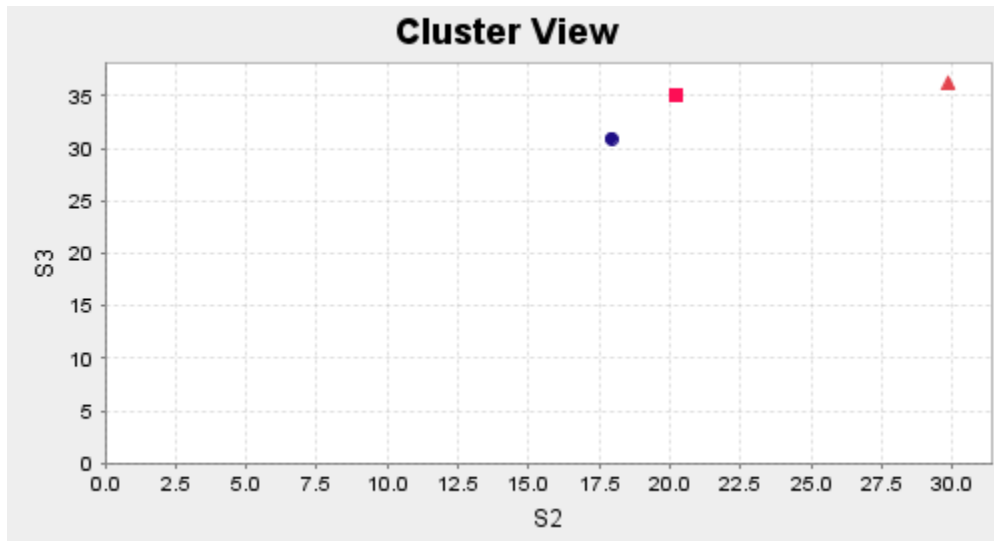
- The average distance between pairs of data points inside the cluster (**<dist.>**).
- A **Members** folder with a list of all the data points that comprise the cluster.
- A list of sub-clusters when applicable (if running a [multi-level clustering](#)).

The tree view provides 4 important options. **They are applied on the selected cluster only. The selected cluster is the one highlighted by the user by clicking on it.**

- **Graph Real:** a graph of the selected cluster's data points is shown (only for 2-D and 1-D analyses). The members of a cluster are represented with a unique shape and color.



- **Graph Fictitious:** a graph of the selected cluster's sub-clusters is generated, where each sub-cluster is represented as a single data point corresponding to its geometric centroid. Every sub-cluster's geometric centroid is graphed with a unique shape and color.



- **Analyze Real:** the selected cluster's data points are exported to a new 2R Data window for further analysis.
- **Analyze Fictitious:** the selected cluster's sub-clusters are exported to a new 2R Data window for further analysis, where each sub-cluster is represented as a single data point corresponding to its geometric centroid.

## PCA ANALYSIS (PRINCIPAL COMPONENT ANALYSIS)

PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analyzing data. The other main advantage of PCA is that once you have found these patterns in the data, and you compress the data by reducing the number of dimensions, without much loss of information. (Smith)

The PCA algorithm is comprised by the following steps:

1. **Arrange the data in a matrix**, where each column represents a data dimension (a separate dataset, such as a sensor's readings) and each row represents a data value for each dimension. Clearly, all the dimensions must contain the same amount of observations.
2. **Subtract the mean across each dimension**. For example, the mean of the values in the  $n$ th column is calculated and then subtracted from each of the values in that same column.
3. **Calculate the covariance matrix**, such that the cell  $(i,j)$  contains the covariance between dimension  $i$  and dimension  $j$ . The resulting covariance matrix must be symmetric
4. **Calculate the eigenvalues and eigenvectors of the covariance matrix**.

Let  $\mathbf{A}$  be a linear transformation represented by a matrix  $\mathbf{A}$ . If there is a vector  $\mathbf{X} \in \mathbb{R}^n \neq 0$  such that  $\mathbf{AX} = \lambda\mathbf{X}$  for some scalar  $\lambda$ , then  $\lambda$  is called the eigenvalue of  $\mathbf{A}$  with corresponding (right) eigenvector  $\mathbf{X}$ . (Weisstein4)

5. **Choose components and form a feature vector**. Once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives you the components in order of significance. Now, if you like, you can decide to *ignore* the components of lesser significance. You do lose some information, but if the eigenvalues are small, you don't lose much. If you leave out some components, the final data set will have fewer dimensions than the original. To be precise, if you originally have dimensions in your data, and so you calculate eigenvectors and eigenvalues, and then you choose only the first eigenvectors, then the final data set has only dimensions. What needs to be done now is you need to form a *feature vector*, which is just a fancy name for a matrix of vectors. This is constructed by taking the eigenvectors that you want to keep from the list of eigenvectors, and forming a matrix with these eigenvectors in the columns. (Smith)

$$FeatureVector = (eig_1 \ eig_2 \ eig_3 \ .. \ eig_n)$$

6. **Calculate the transformed values**.

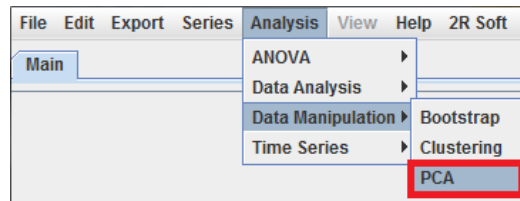
$$FinalData = RowFeatureVector \times RowDataAdjust$$

*RowFeatureVector* is the *FeatureVector* transposed, while *RowDataAdjust* is the mean-adjusted data transposed. (Smith)

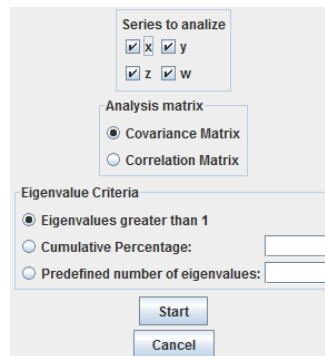
The final data obtained gives the original data solely in terms of the eigenvectors that were chosen in the previous step. If the amount of dimensions of the transformed data is less than the original amount of dimensions, some data is lost, but the most significant features are retained in the new dimensions.

## PCA IN 2R DATA

To start a PCA analysis in 2R Data, the user must have declared two or more data series with the same amount of data points. Then, the **PCA** option under the **Data Manipulation** submenu of the **Analysis** menu can be invoked:



A new screen will appear for the user to indicate the series to be included in the analysis, the matrix to be used for the analysis, and the criteria that should be used for selecting the eigenvalues:

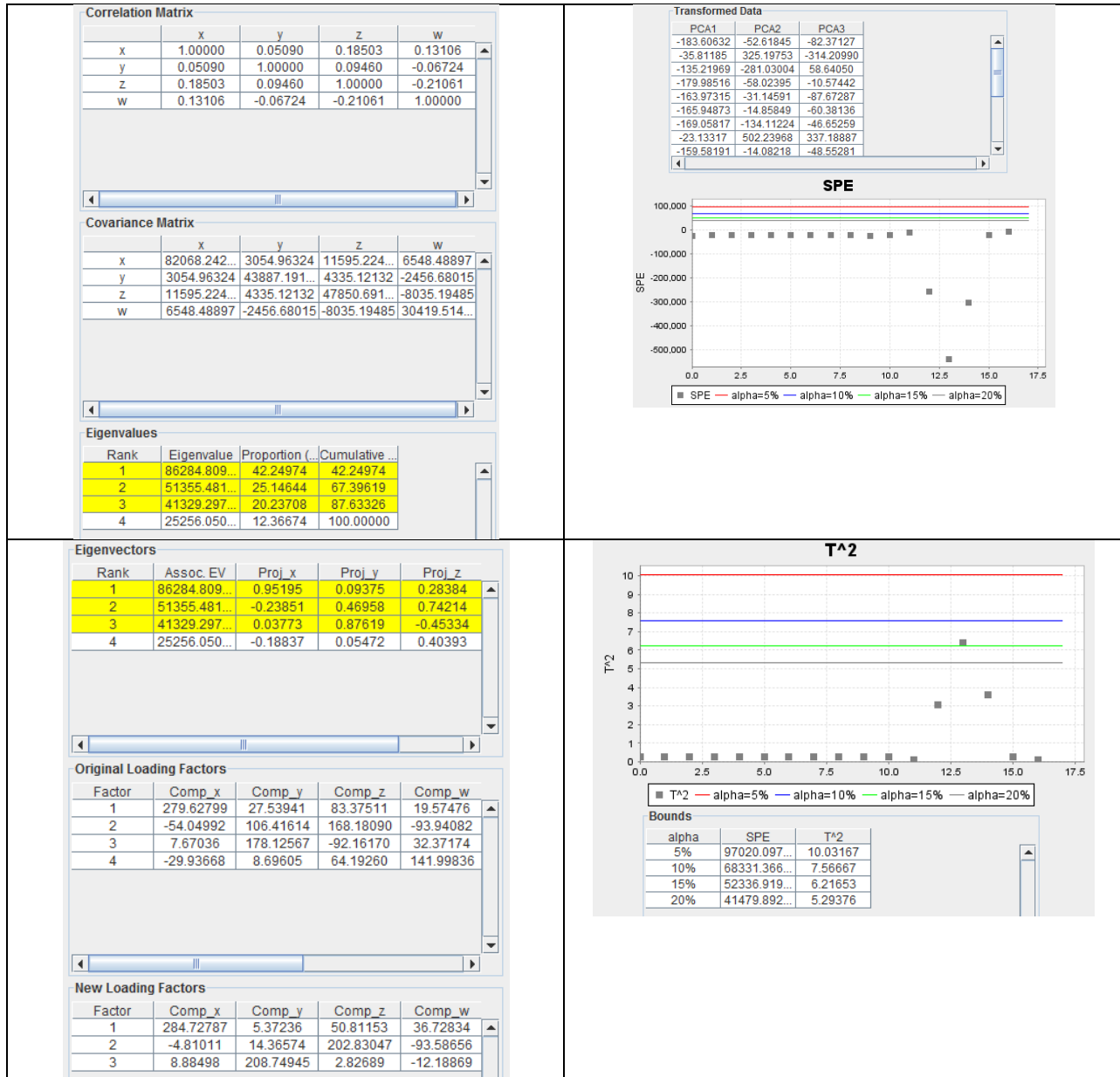


The first two inputs (series to analyze and analysis matrix) are self-explanatory. It should be noted that the covariance matrix is the default analysis matrix. The eigenvalue criteria are explained below:

Criterion	Explanation
<b>Eigenvalues greater than 1</b>	In essence this is like saying that, unless a factor extracts at least as much as the equivalent of one original variable, we drop it. (StatSoft)
<b>Cumulative percentage</b>	The user enters the minimum percentage of the data that should be explained by the new dimensions. Thus, 2R Data retains eigenvalues until the cumulative loading percentage is greater than or equal to such percentage.
<b>Predefined number of eigenvalues</b>	If the user expects a specific amount of dimensions for the transformed data, 2R Data will retain that number of eigenvalues during the analysis.



After setting up the PCA analysis and pressing the **Start** button, the user is presented with a variety of results:



**Note:** The retained eigenvalues are highlighted with **yellow** in the visual interface.

Result	Explanation
<b>Correlation Matrix</b>	The correlation matrix, such that the cell (i,j) contains the correlation between dimension i and dimension j
<b>Covariance Matrix</b>	The covariance matrix, such that the cell (i,j) contains the covariance between dimension i and dimension j.
<b>Eigenvalues</b>	Eigenvalues of the analysis matrix (covariance or correlation matrix, depending on what the user chose in the PCA analysis settings screen).
<b>Eigenvectors</b>	Eigenvectors of the analysis matrix (covariance or correlation matrix, depending on what the user chose in the PCA analysis settings screen).
<b>Original Loading Factors</b>	Correlations between observed variables and factors. The unrotated factor loading matrix is the matrix product of the eigenvector with the square root of the eigenvalue matrix. (Pickering)
<b>New Loading Factors</b>	Loading factors after a varimax rotation. Varimax maximises the variance of loadings within factors across variables (simplifies factors: some variables have high loadings, others have low). (Pickering)
<b>Transformed Data</b>	Data points in the new dimensions, where $PCA_N$ is associated with the nth most significant eigenvalue.

**SPE** Let  $\mathbf{x} \in \mathbb{R}^m$  denote a sample vector of m sensors. Assuming that there are N samples for each sensor, a data matrix  $X \in \mathbb{R}^{N \times m}$  is composed with each row representing a sample. The matrix X is scaled to zero mean for covariance-based PCA and, in addition, to unit variance for correlation-based PCA. The matrix X can be decomposed into a score matrix T and a loading matrix P by either the NIPALS or the singular value decomposition (SVD) algorithm: (Qin 2003)

$$\begin{aligned} \mathbf{X} &= \mathbf{TP}^T + \tilde{\mathbf{X}} = \mathbf{TP}^T + \tilde{\mathbf{T}}\tilde{\mathbf{P}}^T \\ &= [\mathbf{T} \ \tilde{\mathbf{T}}] [\mathbf{P} \ \tilde{\mathbf{P}}]^T \equiv \tilde{\mathbf{T}}\tilde{\mathbf{P}}^T \end{aligned}$$

The SPE index measures the projection of the sample vector on the residual subspace: (Qin 2003)

$$SPE \equiv \|\tilde{\mathbf{x}}\|^2 = \|(\mathbf{I} - \mathbf{PP}^T)\mathbf{x}\|^2$$

The process is considered normal if  $SPE \leq \delta_\alpha$  where  $\delta_\alpha$  denotes the upper control limit for SPE with a significance level  $\alpha$ . (Qin 2003)

$$\delta_\alpha^2 = g\lambda_{h;\alpha}^2$$

$$\begin{aligned} g &= \theta_2/\theta_1, & h &= \theta_1^2/\theta_2 \\ \theta_i &= \sum_{j=i+1}^m \lambda_j^i, & i &= 1, 2, 3 \end{aligned}$$

**T<sup>2</sup>** Hotelling's T<sup>2</sup> statistic measures variations in the PCS: (Qin 2003)

$$T^2 = \mathbf{x}^T \mathbf{P} \mathbf{\Lambda}^{-1} \mathbf{P}^T \mathbf{x}, \quad \bar{\mathbf{\Lambda}} = \frac{1}{N-1} \tilde{\mathbf{T}}^T \tilde{\mathbf{T}} = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_m\}$$

Under the condition that the process is normal and the data follow a multivariate normal distribution, the T<sub>2</sub> statistic is related to an F distribution considering that the population mean and covariance are estimated from data: (Qin 2003)

$$\frac{N(N-1)}{I(N^2-1)} T^2 \sim F_{I, N-I}$$

where  $F_{I, N-I}$  is an F distribution with I and N-I degrees of freedom. For a given significance level  $\alpha$  the process is considered normal if: (Qin 2003)

$$T^2 \leq T_\alpha^2 \equiv \frac{I(N^2-1)}{N(N-1)} F_{I, N-I; \alpha}$$

**Bounds** Upper control limits for SPE and T<sup>2</sup> with different significance levels.

## FACTORS VS COMPONENTS

2R Data performs Factor Analysis in parallel with the PCA Analysis. The differences between factors and components are: (Pickering)

- Factor Analysis (FA) gives factors; PCA yields components.
- Same overall stages.
- Differ in the variance that is analyzed.
- PCA: all variance of observed variables (shared; unique; and error) is analyzed.
- FA: only shared variance is analyzed.
- Theoretically, factors are the underlying (*latent*) variables that CAUSE the covariation between observed variables. The labels for factors are attempts to name these causal latent variables.
- Components are just empirically determined aggregates of the variables without presumed theory. Labels are used but they are just a short-hand for the component.

## TIME SERIES

Time series are used to keep record of the evolution of a specific measure as time passes by. Their analysis has been the topic of research for a long time and has resulted in entire books dedicated to that topic. **Given that 2R Soft aims for simplicity, 2R Data only supports fixed-interval time series, meaning that the time elapsed between one measurement and its adjacent values is always the same.**

## AUTO-CORRELATION

### THEORY

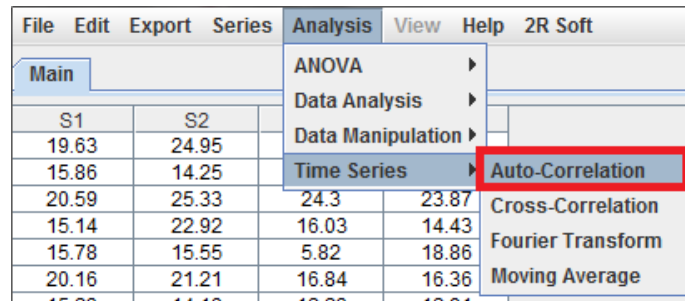
When the correlation is calculated between a series and a lagged version of itself it is called **autocorrelation**. A high correlation is likely to indicate a periodicity in the signal of the corresponding time duration. The correlation coefficient at lag  $k$  of a series  $x_0, x_1, x_2, \dots, x_{N-1}$  is normally given as (Bourke 1996):

$$\text{autocorr}(k) = \frac{\sum_{i=0}^{N-1} (x_i - mx) (x_{i+k} - mx)}{\sum_{i=0}^{N-1} (x_i - mx)^2}$$

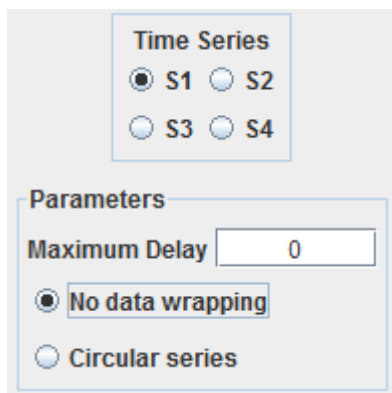
Where  $mx$  is the mean of the series. When the term  $i+k$  extends past the length of the series  $N$  two options are available. The series can either be considered to be 0 or in the usual Fourier approach the series is assumed to wrap, in this case the index into the series is  $(i+k) \bmod N$ . The denominator in the expression above serves to normalize the correlation coefficients such that  $-1 \leq \text{autocorr}(k) \leq 1$ , the bounds indicating maximum correlation and 0 indicating no correlation. A high negative correlation indicates a high correlation but of the inverse of the lagged series (Bourke 1996).

## INPUT

To begin an auto-correlation analysis, the user must navigate through the **Analysis** menu and select the **Auto-Correlation** option from the **Time Series** submenu:



A new window will pop up:



- **Time Series:** the fixed-interval time series that will be analyzed.
- **Maximum Delay:** the maximum lag to be used for the analysis in terms of measurement intervals. For example, if the value entered is 3, the auto-correlation function will test lag values of -3,-2,-1,0,1,2, and 3 to calculate the desired correlogram (or auto-correlation series).
- **No Data Wrapping:** as mentioned in the [Theory](#) section, the lagged series will go past the maximum term in the original series. If no data wrapping is used, the value for all points beyond the maximum term will be taken to be zero (0).
- **Circular Series:** as mentioned in the [Theory](#) section, the lagged series will go past the maximum term in the original series. If a circular series is used, the series becomes a periodic series with period N, where N is the length of the original series. Thus, it will start over right after the maximum term.

## OUTPUT

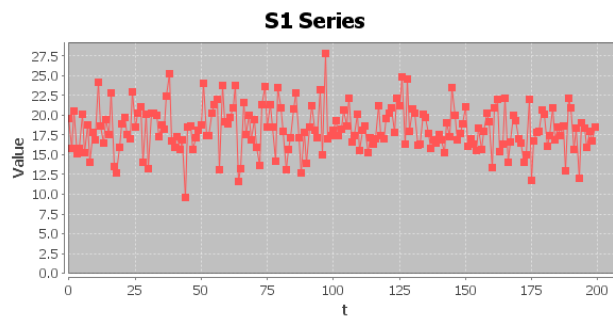
2R Data will show 3 results when the calculations are finished:

- A table with the auto-correlation value,  $r$ , for every amount of delay. This table can be exported to Excel.

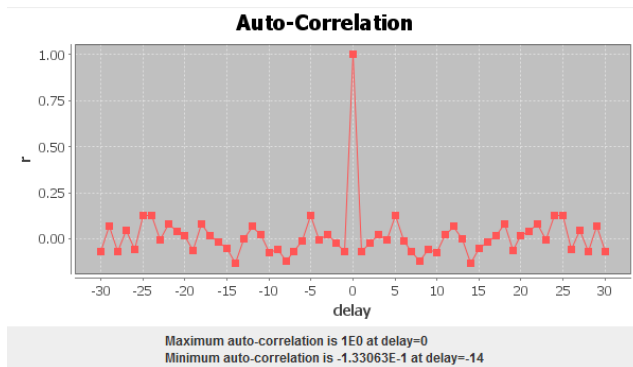
Delay	r
-30.0	-0.0680553...
-29.0	0.0688726...
-28.0	-0.0676005...
-27.0	0.0473514...
-26.0	-0.0588832...
-25.0	0.1289058...
-24.0	0.1302273...
-23.0	-0.0061946...
-22.0	0.0814607...
-21.0	0.0433586...
-20.0	0.0404476...

Export to Excel

- A graph showing the original time series.



- A correlogram (or auto-correlation series) graphically showing the auto-correlation values as a function of lag. The maximum and minimum auto-correlations are explicitly pointed out below the correlogram.



## CROSS-CORRELATION

### THEORY

Cross correlation is a standard method of estimating the degree to which two series are correlated. Consider two series  $x(i)$  and  $y(i)$  where  $i=0,1,2,\dots,N-1$ . The cross correlation,  $r$ , at delay  $d$  is defined as (Bourke 1996):

$$r(d) = \frac{\sum_i [(x(i) - m_x) * (y(i-d) - m_y)]}{\sqrt{\sum_i (x(i) - m_x)^2} \sqrt{\sum_i (y(i-d) - m_y)^2}}$$

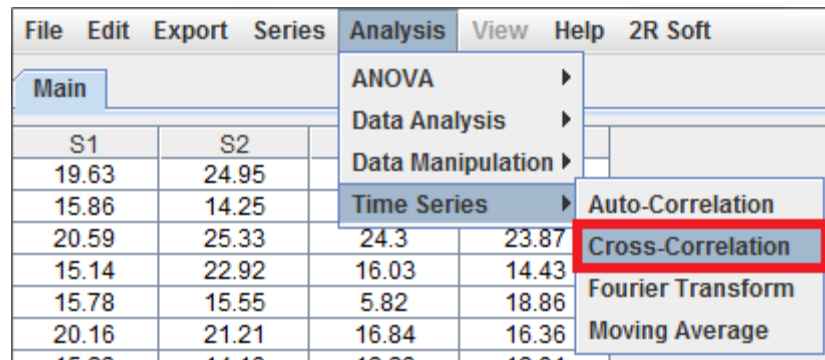
Where  $m_x$  and  $m_y$  are the means of the corresponding series. If the above is computed for all delays  $d=0,1,2,\dots,N-1$ , then it results in a cross correlation series of twice the length as the original series (Bourke 1996).

There is the issue of what to do when the index into the series is less than 0 or greater than or equal to the number of points ( $i-d < 0$  or  $i-d \geq N$ ). The most common approaches are to either ignore these points or assuming the series  $x$  and  $y$  are zero for  $i < 0$  and  $i \geq N$ . In many signal processing applications the series is assumed to be circular in which case the out of range indexes are "wrapped" back within range, ie:  $x(-1) = x(N-1)$ ,  $x(N+5) = x(5)$  (Bourke 1996).

The range of delays  $d$  and thus the length of the cross correlation series can be less than  $N$ , for example the aim may be to test correlation at short delays only. The denominator in the expression above serves to normalize the correlation coefficients such that  $-1 \leq r(d) \leq 1$ , the bounds indicating maximum correlation and 0 indicating no correlation. A high negative correlation indicates a high correlation but of the inverse of one of the series (Bourke 1996).

### INPUT

To begin a cross-correlation analysis, the user must navigate through the **Analysis** menu and select the **Cross-Correlation** option from the **Time Series** submenu:



A new window will pop up:

Time Series (Pick 2)

S1  S2

S3  S4

Parameters

Maximum Delay

No data wrapping

Circular series

- **Time Series:** the two fixed-interval time series that will be juxtaposed.
- **Maximum Delay:** the maximum lag to be used for the analysis in terms of measurement intervals. For example, if the value entered is 3, the cross-correlation function will test lag values of -3,-2,-1,0,1,2, and 3 to calculate the desired cross-correlation series.
- **No Data Wrapping:** as mentioned in the [Theory](#) section, the lagged series will go past the maximum term in the original series. If no data wrapping is used, the value for all points beyond the maximum term will be taken to be zero (0).
- **Circular Series:** as mentioned in the [Theory](#) section, the lagged series will go past the maximum term in the original series. If a circular series is used, the series becomes a periodic series with period N, where N is the length of the original series. Thus, it will start over right after the maximum term.

---

## OUTPUT

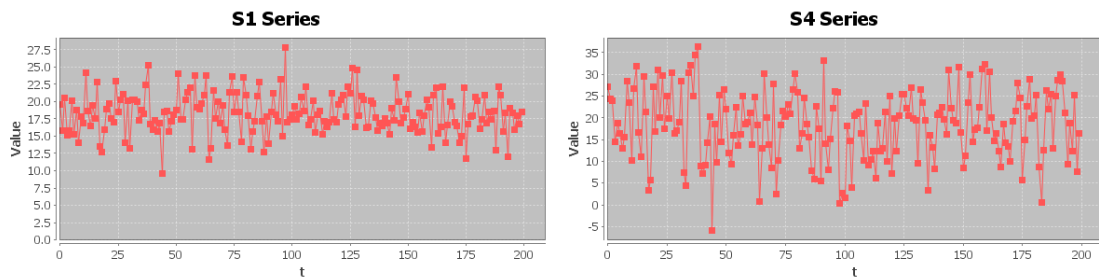
2R Data will show 3 results when the calculations are finished:

- A table with the cross-correlation value,  $r$ , for every amount of delay. This table can be exported to Excel.

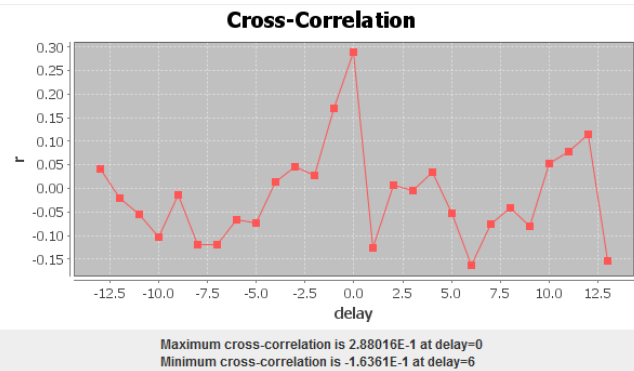
Delay	r
-13.0	0.0411533...
-12.0	-0.0215481...
-11.0	-0.0540871...
-10.0	-0.1033212...
-9.0	-0.0141858...
-8.0	-0.1183583...
-7.0	-0.1204626...
-6.0	-0.0658630...
-5.0	-0.0726633...
-4.0	0.0126413...
-3.0	0.0466488...

Export to Excel

- Graphs showing the original time series.



- A cross-correlation plot graphically showing the cross-correlation values as a function of lag. The maximum and minimum cross-correlations are explicitly pointed out below the cross-correlation plot.



## FOURIER TRANSFORM

### THEORY

**Note:** this section is based on (Handley 2007). Refer to that text for a more in-depth treatment of the topic.

### FOURIER SERIES

Any periodic function can be expressed as the sum of a series of sines and cosines (of varying amplitudes). A function  $f(x)$  can be expressed as a series of sines and cosines:

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos(nx) + \sum_{n=1}^{\infty} b_n \sin(nx),$$

Where:

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) dx, \quad a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx) dx, \quad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(nx) dx, \quad n = 1, 2, 3, \dots$$



## FOURIER TRANSFORM

Fourier Series can be generalized to complex numbers, and further generalized to derive the *Fourier Transform*.

Forward Fourier Transform:

$$F(k) = \int_{-\infty}^{\infty} f(x)e^{-2\pi i k x} dk$$
, Note:  $e^{xi} = \cos(x) + i \sin(x)$

Fourier Transform maps a time series (e.g. audio samples) into the series of frequencies (their amplitudes and phases) that composed the time series.

If we wish to find the frequency spectrum of a function that we have *sampled*, the continuous Fourier Transform is not so useful. We need a discrete version:

Discrete Forward Fourier Transform:

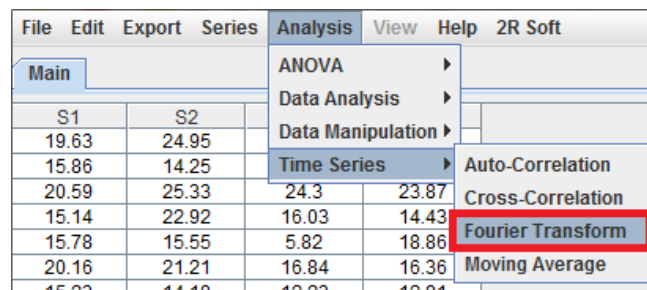
$$F_n = \sum_{k=0}^{N-1} f_k e^{-2\pi i n k / N}$$
, so the complex numbers  $f_0 \dots f_N$  are transformed into complex numbers  $F_0 \dots F_N$ .

## INTERPRETATION

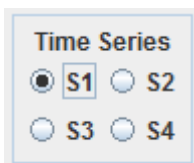
- The magnitude of the complex number for a DFT component is the power at that frequency.

## INPUT

To begin a Fourier Transform analysis, the user must select the **Fourier Transform** option found under the **Time Series** submenu of the **Analysis** menu:



The program then asks the user to select the series that is to undergo a Fourier analysis:



- **Time Series:** the fixed-interval time series that will be analyzed.

## OUTPUT

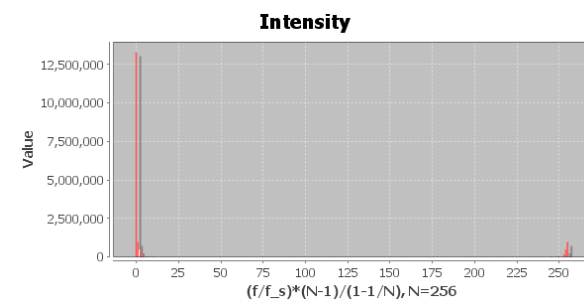
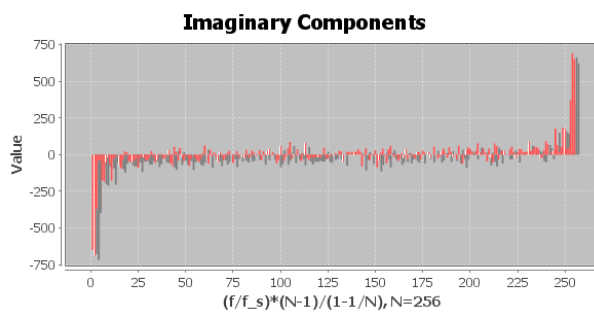
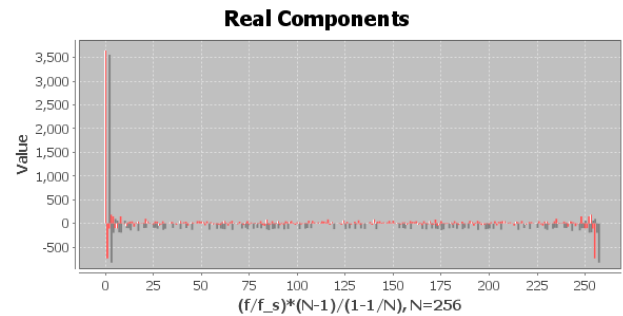
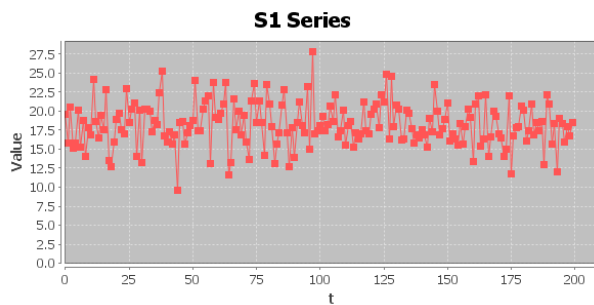
In the results, the variable **f\_s** is used to represent the **sampling frequency** (samples per unit of time), which should be known by the user. **The calculated series covers frequencies between 0 and f\_s.**

- The first result is a table showing the real and imaginary components for the different frequencies, along with the associated intensities ( $\text{img}^2 + \text{real}^2$ ). This data can be easily exported to Excel for further analysis.

f/f_s	real	img	I
0.0	3647.5	0.0	1.3304256...
0.00390625	-738.53294...	-650.49599...	968575.95...
0.0078125	-109.17594...	-691.60520...	490237.15...
0.01171875	183.59145...	-370.83999...	171228.12...
0.015625	148.50952...	-43.046854...	23908.111...
0.01953125	-99.891384...	-39.911950...	11571.252...
0.0234375	-105.35849...	-172.70219...	40926.458...
0.02734375	14.007429...	-183.54421...	33884.687...
0.03125	146.93692...	-51.907293...	24284.827...
0.03515625	-9.2303827...	-16.479043...	356.75882...
0.0390625	10.707694...	65.000400...	4495.0004...

[Export to Excel](#)

- 4 graphs are also generated: one showing the original series, and the other three showing the real components, imaginary components, and associated intensities of the various frequencies:



## MOVING AVERAGE

### THEORY

Inherent in the collection of data taken over time is some form of random variation. There exist methods for reducing or canceling the effect due to random variation. An often-used technique in industry is "smoothing". This technique, when properly applied, reveals more clearly the underlying trend, seasonal and cyclic components (SEMATECH). Moving Average is a popular smoothing method. It computes the mean of successive smaller sets of past data. The general expression for the moving average is:

$$M_t = [X_t + X_{t-1} + \dots + X_{t-N+1}] / N, \text{ where } N \text{ is the size of the smaller sets of past data.}$$

Here's an example of a Moving Average when N=3:

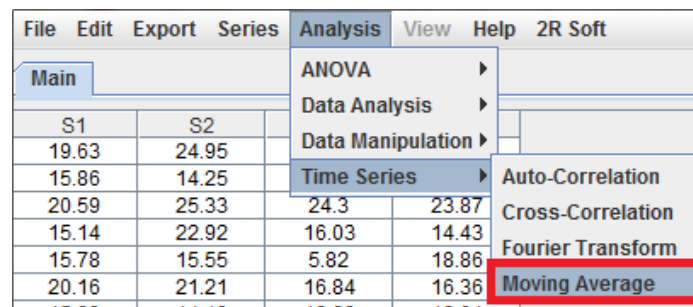
Point	Value	MA
1	9	9
2	8	8.5
3	9	8.67
4	12	9.67
5	9	10.00
6	12	11.00
7	11	10.67
8	7	10.00
9	13	10.33
10	9	9.67
11	11	11.00
12	10	10.00

$$\text{E.g. } M_6 = (X_6 + X_5 + X_4) / 3 = (12 + 9 + 12) / 3 = 33 / 3 = 11$$

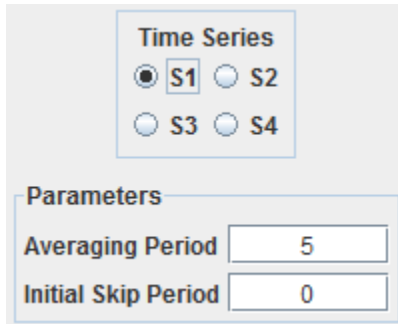
Note that the first two data points don't have 2 values before them to take into account, so they only consider the average of the available values.

### INPUT

To begin a Moving Average analysis, the user must select the **Moving Average** option found under the **Time Series** submenu of the **Analysis** menu:



The user is prompted for some information:

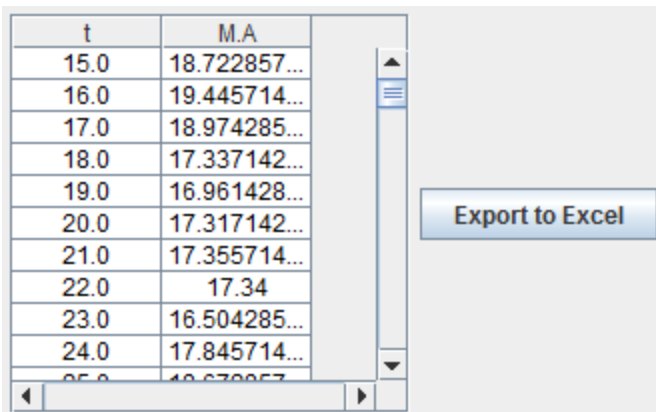


The screenshot shows a software interface with two sections. The top section is titled "Time Series" and contains four radio buttons labeled S1, S2, S3, and S4. The S1 radio button is selected. The bottom section is titled "Parameters" and contains two input fields: "Averaging Period" with the value 5, and "Initial Skip Period" with the value 0.

- **Time Series:** the fixed-interval time series that will be analyzed.
- **Averaging Period:** the size of the smaller sets of past data. Refer to the [Theory](#) section for details.
- **Initial Skip Period:** the point from which the moving average series will begin. E.g. If the initial skip period is X, the moving average series starts at  $t=X$ .

## OUTPUT

- A table shows the moving average value for every point in the analysis range. This data can be easily exported to Excel for further analysis.

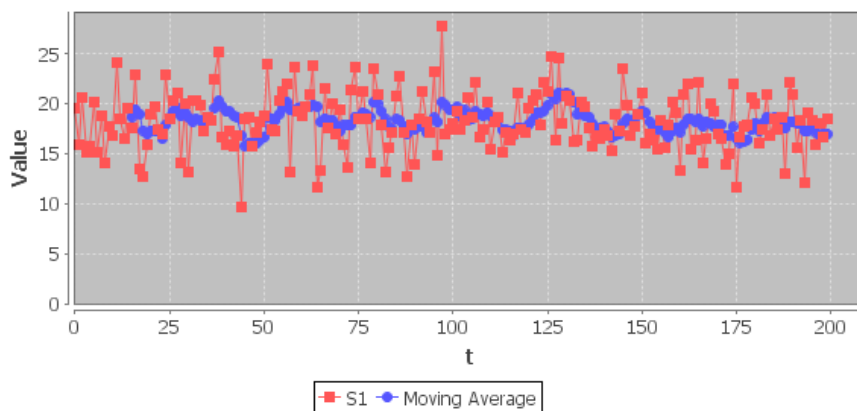


The screenshot shows a data table with two columns: 't' and 'M.A'. The table contains 11 rows of data. To the right of the table is a button labeled "Export to Excel".

t	M.A
15.0	18.722857...
16.0	19.445714...
17.0	18.974285...
18.0	17.337142...
19.0	16.961428...
20.0	17.317142...
21.0	17.355714...
22.0	17.34
23.0	16.504285...
24.0	17.845714...
25.0	18.678571...

- The results are also plotted. The moving average is shown in BLUE, while the original series is displayed in RED. Note that the random variations are reduced in the moving average series.

### S1 Series



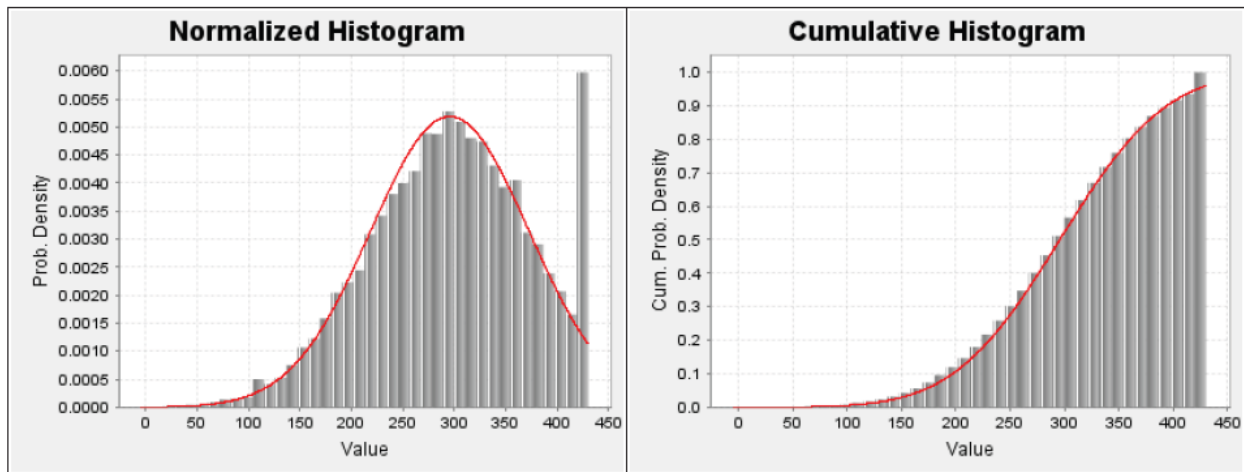
## EXAMPLES

### EXAMPLE 1 – INTERNATIONAL EXAM – SINGLE SERIES ANALYSIS

An organization in charge of an international exam, such as the TOEFL or the GRE, has received all of the candidate scores for one season and wants to define the different grade bounds to be able to send the score reports to the test-takers. The final result for each candidate is expressed as a numeric value between 1 and 9 and the organization has decided that, given the size of the population, the fairest method to define the grade bounds is to fit a normal distribution to the sample and take predefined percentiles as the grade bounds.

The exam scores range from 0 to 430 and the sample data can be found in the Example 1 data model.

The results of a single-series analysis over the Example 1 data model are shown below:



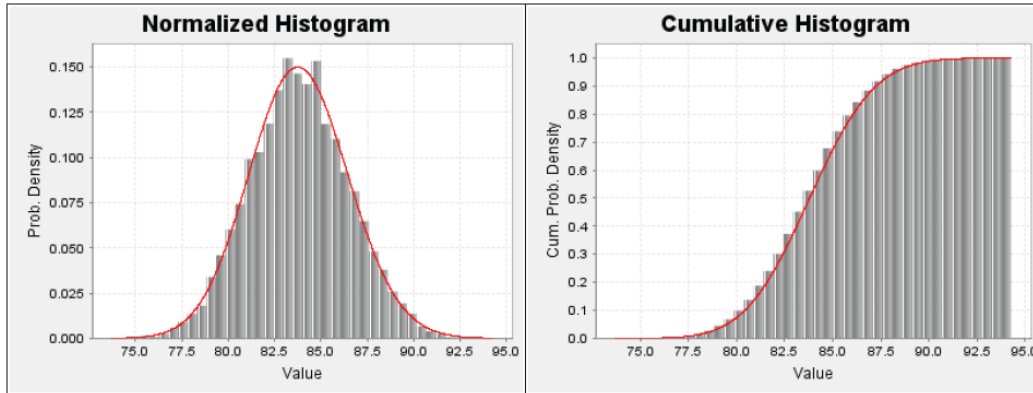
The best-fit normal distribution (red curve in graphs above) has a mean score of 295.69 and standard deviation of 76.93. Consequently, the grade bounds can now be calculated:

Final Grade	Grade Bound Description	Lower Bound	Upper Bound
1	Lower than 20 <sup>th</sup> percentile	0	231
2	20 <sup>th</sup> percentile to 30 <sup>th</sup> percentile	231	255
3	30 <sup>th</sup> percentile to 40 <sup>th</sup> percentile	255	276
4	40 <sup>th</sup> percentile to 50 <sup>th</sup> percentile	276	296
5	50 <sup>th</sup> percentile to 60 <sup>th</sup> percentile	296	315
6	60 <sup>th</sup> percentile to 70 <sup>th</sup> percentile	315	336
7	70 <sup>th</sup> percentile to 80 <sup>th</sup> percentile	336	360
8	80 <sup>th</sup> percentile to 90 <sup>th</sup> percentile	360	394
9	90 <sup>th</sup> percentile or higher	394	430

### EXAMPLE 2 – ADMISSIONS OFFICE - BOOTSTRAPPING

A university's admissions office wants to calculate a confidence interval with  $\alpha = 5\%$  for the amount of students that send applications to enter the Civil Engineering major each semester. The Civil Engineering Faculty opened 10 years ago, so they only have 20 data points to carry out the analysis, which are included in the Example 2 data

model. After conducting a Bootstrap analysis with 10,000 iterations, the admissions office obtained the following results:



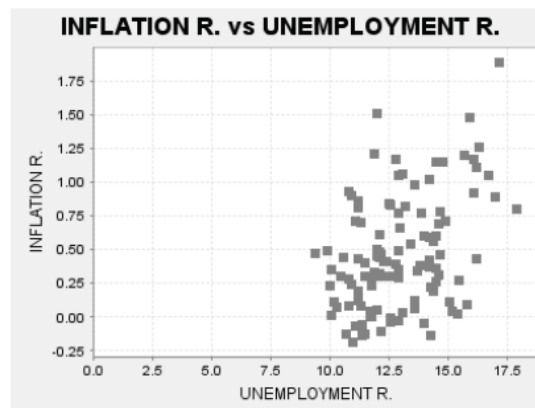
The probability distribution with the best (lowest) Anderson-Darling statistic is a Gamma distribution with an alpha parameter of 991.58 and a lambda parameter of 11.83. Thus, a confidence interval with  $\alpha = 5\%$  is that between the 5<sup>th</sup> percentile and the 95<sup>th</sup> percentile of such distribution:

Confidence Interval = [79, 88] applications per semester

### EXAMPLE 3 – INFLATION AND UNEMPLOYMENT RATES – CORRELATION

According to the Phillips Curve, the rate of inflation and the unemployment rate follow an inverse relationship (when one is high the other one is low, and vice-versa). To test this theory in practice, a student has decided to analyze the correlation between the two macroeconomic measurements (unemployment and inflation) in Colombia with historic data from the past 10 years. The Example 3 data model contains such information.

After running a correlation analysis over the Example 3 data model for inflation rate vs unemployment rate, the student obtained the results shown below:



## Statistics

Covariance: 0.32213
Correlation: 0.41117

## Trendline

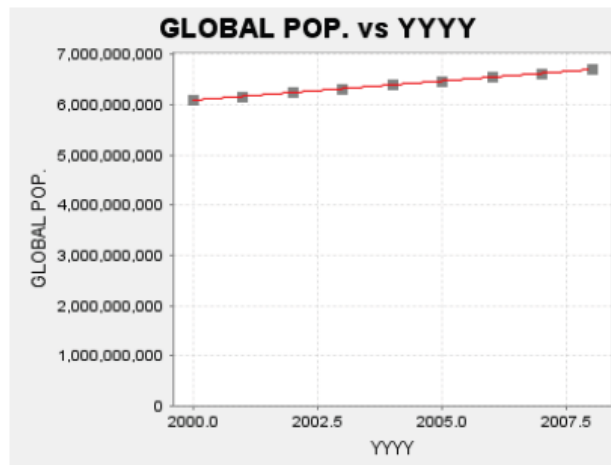
INFLATION R. = -0.72962+0.09248*UNEMPLOYMENT R.
R <sup>2</sup> = 0.16906

The correlation coefficient (0.411) indicates that there is a positive correlation between the two data series, which is exactly the opposite of what was expected. Such relation between inflation and unemployment rates isn't linear, since the determination coefficient ( $R^2$ ) for the linear regression, 0.169, is nowhere near 1.0. Therefore, one can conclude that Colombia's economy doesn't hold true to what the Phillips Curve states, and that the relation between these two variables must be non-linear.

### EXAMPLE 4 – WORLD POPULATION - REGRESSION

A student wants to estimate the global population by the year 2020 based on world population data from the past 10 years. The Example 4 data model contains the input data for this estimate.

The result of a regression analysis over the Example 4 model is shown below:



## Selected Regression

Equation Form: GLOBAL POP. = a+b*YYYY+c*YYYY <sup>2</sup>
R <sup>2</sup> : 1.00000

## Calculated Parameters

Param	Val.	Std. Error	t-value	p-value
a	-75638459904.47890	123227987737.79543	-0.61381	0.56188
b	5469855.95589	122982168.90417	0.04448	0.96597
c	17696.05195	30684.16897	0.57672	0.58510

The regression is carried out in the form “GLOBAL POP. vs YYYY”, taking into account that time is the independent variable and population is the dependent variable. According to the screenshot above, a second-degree polynomial regression fits the data with a determination coefficient ( $R^2$ ) of 1.0, so it can be used as the base regression for the subsequent analysis. Not only does the equation “ $GP=-75638459904.48+5469855.96*Y+17696.05*Y^2$ ” explain the dependence of global population with time; it also expresses a coherent behavior (always increasing, faster than linear, etc).

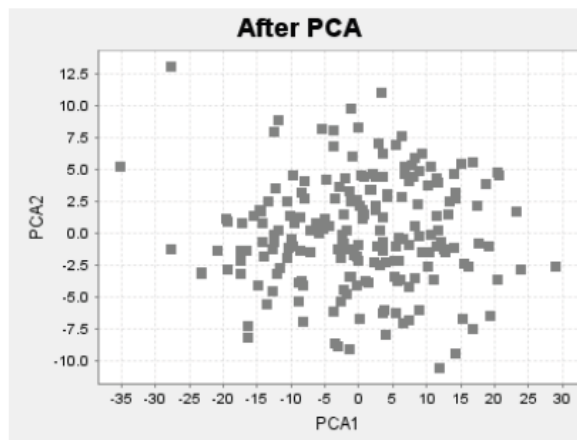
To estimate the world population by the year 2020, one must replace “Y” for “2020” in the regression equation:

$$GP(2020)=-75638459904.48+5469855.96*2020+17696.05*2020^2 = 7,617,611,555 \text{ people}$$

### EXAMPLE 5 – TEMPERATURE SENSORS - PCA

A meteorologist has recollected historic temperature measurements with the help of 4 sensors spread along a small village. Given that temperature differs slightly within a specific geographic region, the meteorologist expects the 4 sensors to share around 80% of data in common and wants to be able to visualize the information in a 2-dimensional plane with the minimum amount of data loss. The Example 5 data model contains the input data for this analysis expressed in the Celsius temperature scale.

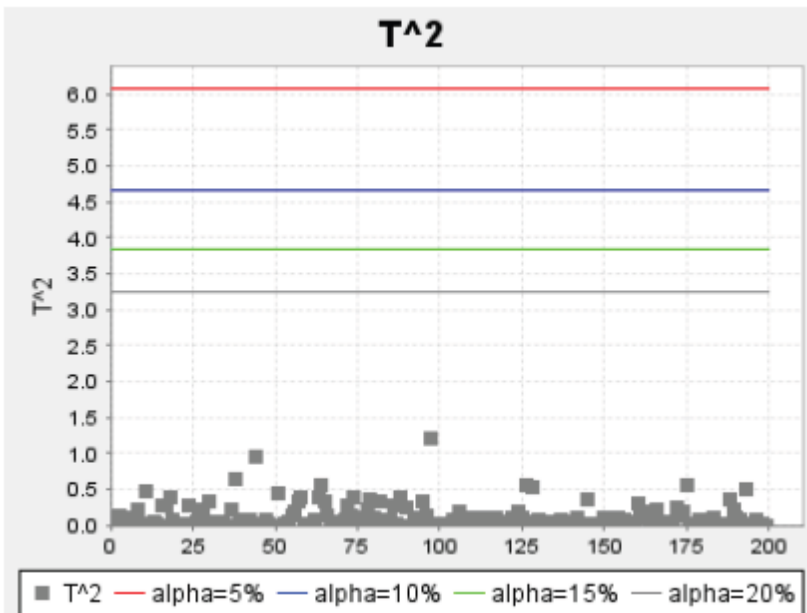
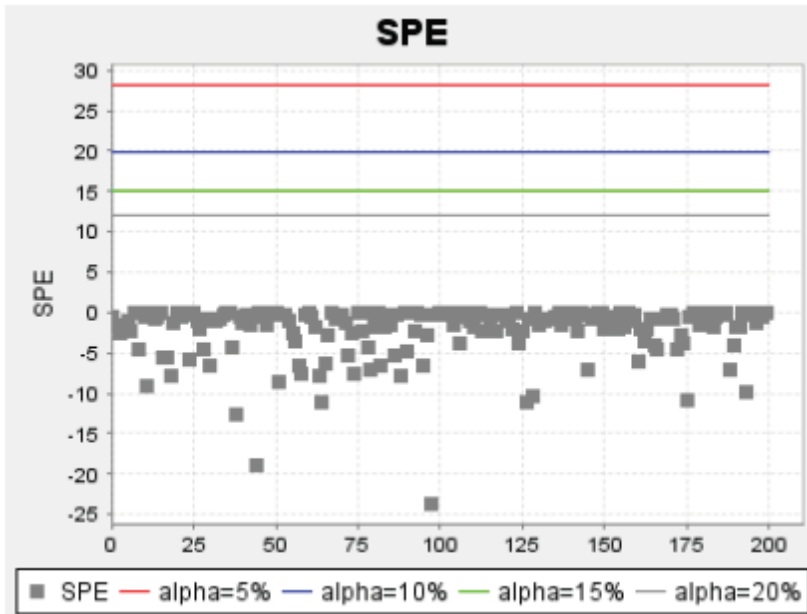
After restricting the PCA result to exactly 2 eigenvalues, the meteorologist obtained a new set of coordinates, PCA1 and PCA2:



### Eigenvalues

Rank	Eigenvalue	Proportion (%)	Cumulative %
1	123.16258	79.44042	79.44042
2	17.89882	11.54482	90.98524
3	8.48158	5.47066	96.45590
4	5.49469	3.54410	100.00000





As the eigenvalues table shows, the 2 new dimensions have retained 91% of the original data, which is impressive taking into account that the original data had twice the amount of dimensions. Furthermore, the SPE and T<sup>2</sup> graphs indicate that all the resulting data points are reliable, even with an alpha as high as 20%.

## BIBLIOGRAPHY

Abeel, T. "K-Medoids." from <http://java-sourceforge.net/api/0.1.6/net/sf/javaml/clustering/KMedoids.html>.

AllBusiness. "R-Squared." Retrieved 05/17/2010, from <http://www.allbusiness.com/glossaries/r-squared-r-squared/4954639-1.html>.

Annis, C. "Goodness-of-Fit tests for Statistical Distributions." Retrieved 05/17/2010, from <http://www.statisticalengineering.com/goodness.htm>.

Answers.com. "Regression Analysis." Retrieved 05/17/2010, from <http://www.answers.com/topic/regression-analysis>.

ANU, A. N. U. "t-table." Retrieved 05/17/2010, from [http://engnet.anu.edu.au/DEcourses/engn2226/web2226\\_Exam/t\\_table\\_2.jpg](http://engnet.anu.edu.au/DEcourses/engn2226/web2226_Exam/t_table_2.jpg).

Bourke, P. (1996). "Cross Correlation." from <http://paulbourke.net/miscellaneous/correlate/>.

BurnsStatistics. "The Statistical Bootstrap and Other Resampling Methods." Retrieved 05/17/2010, from [http://www.burns-stat.com/pages/Tutor/bootstrap\\_resampling.html#bootstrap](http://www.burns-stat.com/pages/Tutor/bootstrap_resampling.html#bootstrap).

Devore, J. L. (1999). Probability and Statistics for Engineering and the Sciences, Duxbury Pr.

Dipartimento Di Elettronica E Informazione, P. D. M. "A Tutorial on Clustering Algorithms." 2011, from [http://home.dei.polimi.it/matteucc/Clustering/tutorial\\_html/](http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/).

EncyclopediaOfStatistics. "Quartiles." Retrieved 05/17/2010, from <http://books.google.com/books?id=56LkkdZPpyoC&pg=PT19&lpg=PT19>.

ERI. "Kolmogorov-Smirnov Test." Retrieved 05/17/2010, from <http://www.eridlc.com/onlinetextbook/index.cfm?fuseaction=textbook.appendix&FileName=Table7>.

Handley, M. (2007) Fourier Transforms.

Johnson, S. C. (1967). "Hierarchical Clustering Schemes." Psychometrika.

Lane2, D. "Standard Deviation and Variance." Retrieved 05/17/2010, from <http://davidmlane.com/hyperstat/A16252.html>.

Lane, D. "Computing Pearson's Correlation Coefficient." Retrieved 05/17/2010, from <http://davidmlane.com/hyperstat/A51911.html>.

MathsRevision.net. "Box and Whisker Diagrams." Retrieved 05/17/2010, from <http://www.mathsrevision.net/alevel/pages.php?page=50>.

McColl, V. J. E. J. H. "Statistics Glossary." Retrieved 05/17/2010, from [http://www.stats.gla.ac.uk/steps/glossary/hypothesis\\_testing.html#ts](http://www.stats.gla.ac.uk/steps/glossary/hypothesis_testing.html#ts).

Nugent, R. and L. Stanberry Cluster Analysis and Other Unsupervised Learning Methods (Stat 593 E).

Pickering, A. "PCA and FA." Retrieved 05/17/2010, from <http://homepages.gold.ac.uk/aphome/lec7ohp.doc>.

PlanetMath. "Covariance." Retrieved 05/17/2010, from <http://planetmath.org/encyclopedia/Covariance.html>.

Qin, S. J. (2003). "Statistical process monitoring: basics and beyond." Journal of Chemometrics **17**(8-9): 480-502.

ReliaSoftCorp. "Critical Values for Cramér-von Mises Test." Retrieved 05/17/2010, from [http://www.weibull.com/RelGrowthWeb/Appendix\\_B\\_Critical\\_Values\\_for\\_Cramer-von\\_Mises\\_Test.htm](http://www.weibull.com/RelGrowthWeb/Appendix_B_Critical_Values_for_Cramer-von_Mises_Test.htm).

SEMATECH1, N. "Kolmogorov-Smirnov Goodness-of-Fit Test." Retrieved 05/17/2010, from <http://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm>.

SEMATECH2, N. "Anderson-Darling Test." Retrieved 05/17/2010, from <http://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm>.

SEMATECH3, N. "Measures of Skewness and Kurtosis." Retrieved 05/17/2010, from <http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm>.

SEMATECH, N. "Engineering Statistics Handbook." from <http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc42.htm>.

Simard, R. "Package umontreal.iro.lecuyer.probdist." Retrieved 05/17/2010, from <http://www.iro.umontreal.ca/~simardr/ssj/doc/html/umontreal/iro/lecuyer/probdist/package-summary.html>.

Smith, L. I. A tutorial on Principal Components Analysis.

StatSoft. "Principal Components and Factor Analysis." Retrieved 05/17/2010, from <http://www.statsoft.com/textbook/principal-components-factor-analysis/>.

Temporis, S. "Cluster Analysis." 2011, from <http://www.spiritus-temporis.com/cluster-analysis/distance-measure.html>.

Weisstein2, E. W. "Arithmetic Mean." Retrieved 05/17/2010, from <http://mathworld.wolfram.com/ArithmeticMean.html>.

Weisstein3, E. W. "Variance." Retrieved 05/17/2010, from <http://mathworld.wolfram.com/Variance.html>.

Weisstein4, E. W. "Eigenvalue." Retrieved 05/17/2010, from <http://mathworld.wolfram.com/Eigenvalue.html>.

Weisstein, E. W. "K-Means Clustering Algorithm." from <http://mathworld.wolfram.com/K-MeansClusteringAlgorithm.html>.