Corviz Help

USER MANUAL

Advanced Projects R&D Ltd.

Introduction

1 Introduction

Corviz is a handy graphics-oriented routine that allows you to specify any value you like for a correlation coefficient, and have it generated instantly as a graphical image. You can optionally specify minimum and maximum values for each variable, as well as the means and standard deviation for each, and whether or not generated observations are integers or real-valued (decimal) format.

However, it also allows you to place cut-scores on each variable axis, and have the relevant Taylor-Russell statistics automatically computed. All results, including relevant regressions and subset correlations are shown graphically as well as numerically in table format.

The scatterplot graphics provide an option for bubble-plots for integer-valued observations on both variables. You can also compare two tests' validities in terms of their expected efficiencies in selecting 'successful' members of a screened sample.

The program is designed to be used by consultants, trainers, and professionals who might have need for the capability to show correlations to others as 'you can see by eye' effects rather than as abstract indices. To that end, the program allows a user to copy tables and images to the clipboard for immediate cut-and-paste into other documents, as well as being able to save all images and tables into a variety of graphical and file formats.

I have prepared a 23-page application note available for free download, so as to show how this program might be used in competitive sales presentations, training sessions and teaching material generation, and 'what are the likely consequences' cut-score-setting exercises.

1.1 What does Corviz do?

1. Generate correlation data and scatterplots on demand, where sample sizes can vary between 10 and no sensible upper limit *(I've tested it up to 500,000 cases; it takes approximately 25 secs on a Dell M90 Intel core i7-920xm extreme edition processor. The data generation takes about 2 seconds but the table formatting eats up time with this number of cases)*.

2. Makes it trivial for the user to setup cut-scores on one or both variables. These are displayed automatically on a cut-score graph, with associated cut-score consequence statistics displayed in diagnostic tables.

3. Regressions for subset-datasets (above one or both cut-scores) are computed and displayed automatically, as are correlations and descriptive statistics.

4. The selection efficiencies of tests may be compared by entering the relevant data in a table (for test-1 and test-2). The user can also request the required proportion of cases expected to be found above a cut-score, and the program will calculate the cut-score for either criterion or test variable which will achieve this (essentially this minor function computes the Taylor-Russell table value as presented in Taylor, H.C., & Russell, J.T. (1939) The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. Journal of Applied Psychology, 23, 5, 565-578.

1.2 Program Features

1. Generates positive, zero, or negative correlations from data. It will generate data which produce a correlation to within 3-decimal-place accuracy of the specified correlation.

2. Sample sizes can vary between 11 and no serious upper limit (tested to 500,000 cases in a single sample).

3. Graphics and tables can be copied to the clipboard with a single right-hand mouse-click. These can be pasted into any other relevant document such as a word-processor file, an Excel spreadsheet, statistical package, or powerpoint presentation.

4. Graphs can also be saved as bmp, jpg, gif, emf, pcx, pdf, png, or svg format files. Tables can be saved as xls, xlsx, xml, html, or csv format files.

5. Plot types are 'standard' or 'bubble-frequency' plot formats.

6. The program computes the relevant cut scores required to select a proportion of cases from a total sample; ultimately providing the Taylor-Russell table value for a criterion and selection test, with selection ratios and validity index (correlation). However, the program calculates these values for any set of selection ratios and any size validity coefficient.

7. All graphics images are able to edited:

- graphics zoom/shrink/reset.
- inside and outside background colors.
- symbol size, color, and border width, fill/hollow, and color.
- line colors, widths and type (dash or solid line).
- axis and main title text labels, fonts, sizes, normal/italic/bold, and colors.
- axis units, minimum and maximum display values, axis tick-label sizes, fonts, normal/italic/bold, and color.

8. Tables in the clipboard are HTML formatted to preserve shading and colored text effects as per screen image. When pasted into a document, these can be edited as you would any table in a word-processor or Excel spreadsheet.

9. When started, the program automatically checks for updates if you are connected to the internet, and will inform you if an update is available. You can also manually check for updates from the "about" screen.

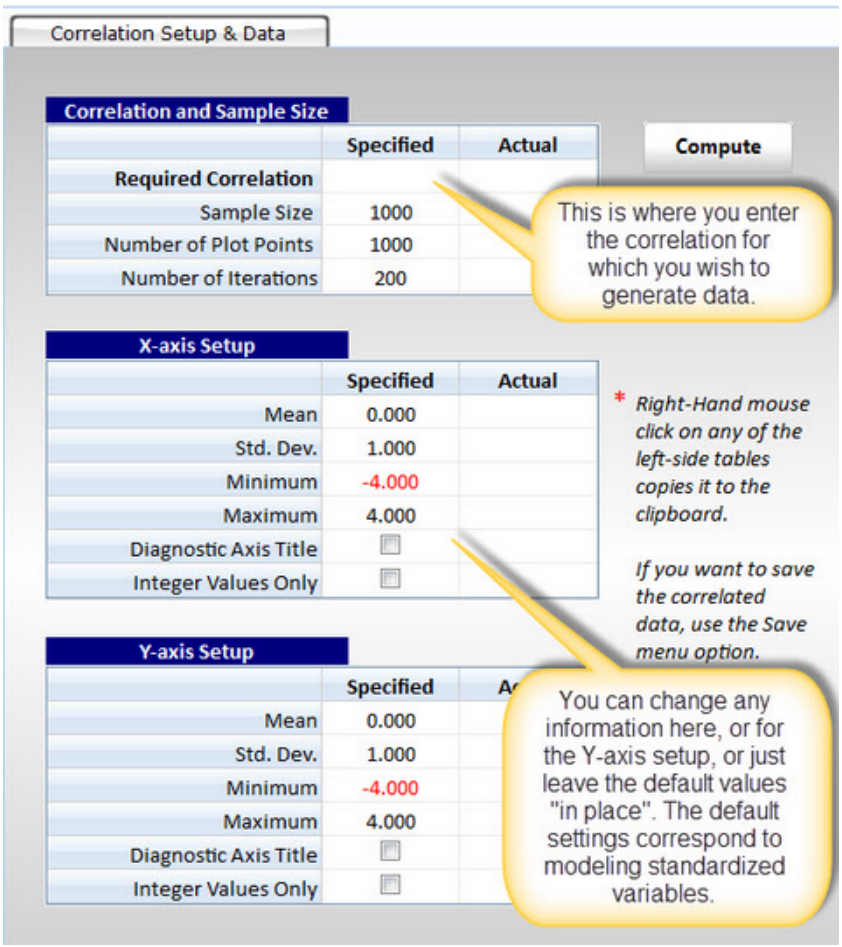
How to Use the Software

2 How to Use the Software

For a quick overview, download the 23-page application note available for free download or view the audio-visual introduction on the program page.

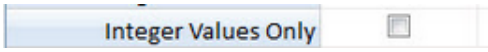
2.1 Generating Data for a Specified Correlation

When you start the program, you see the setup screen:

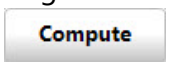


Note, the two columns entitled "**Specified**" and "**Actual**". The "**Specified**" column contains the setup data. The "**Actual**" column will contain the parameters for the data generated after you click on the Compute button.

The Diagnostic Axis Title checkbox, if checked, creates an axis title which includes the mean and standard deviation of the generated data for that particular variable (X or Y).

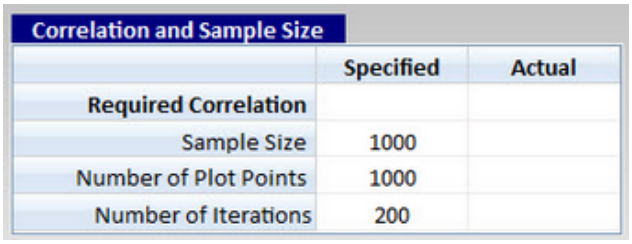
The 

checkbox, if checked, generates data as integers (whole numbers). Otherwise data is generated as continuous, real-valued numbers within the range set by the minimum and maximum values (decimal/fractional numbers).

So, if you just wanted to envisage/generate a correlation scatterplot using the defaults, all you would need to do is type in the correlation value, and click on the 

button.

The initial setup parameters ...



Correlation and Sample Size

	Specified	Actual
Required Correlation		
Sample Size	1000	
Number of Plot Points	1000	
Number of Iterations	200	

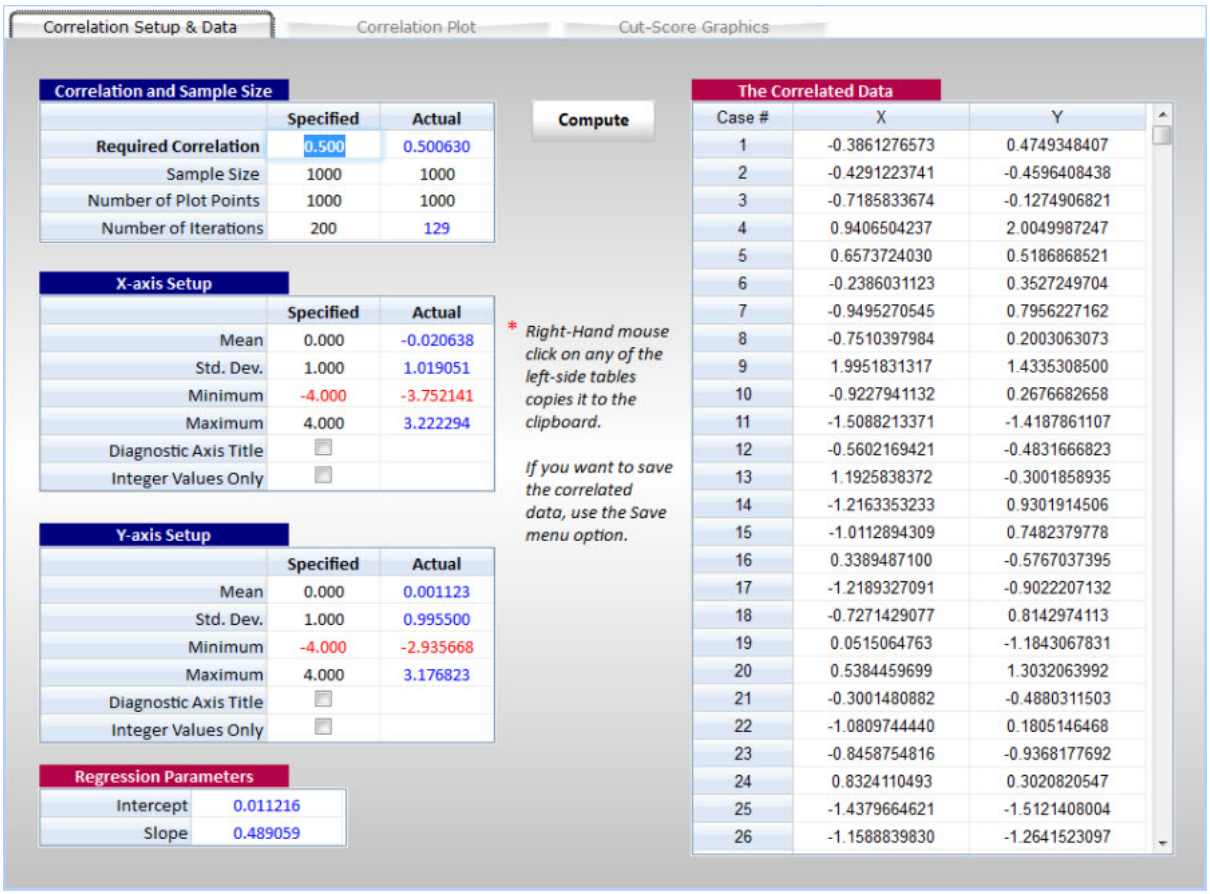
Samplesize: the number of cases you wish to generate with a correlation you specify. This could be the sample size in an study reported in a research document, where you wish to view what the scatterplot might look look for a particular reported correlation. Remember, the sample size must be greater than 10. There is no upper-limit on the sample-size except that which might be considered reasonable (tested to 500,000 cases).

Number of Plot Points: You can specify the number of points plotted in a graph. This is because if you decided to generate a 50,000 case dataset, you do not want to plot all 50,000 data points unless you have a special reason to do so. So, you can specify an upper limit to the number of points you wish to plot. It doesn't matter this value is larger than the number of points you generate.

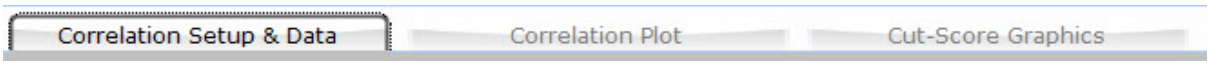
Number of Iterations: CorViz generates a dataset by sampling a pair of random numbers from a normal distribution, and transforming one of them so that a vector of such numbers correlates at the specified level with the other vector of normally distributed numbers. However, it does this under the additional constraints that the numbers for each variable must be generated within a minimum and maximum value bound, possess a specified mean and standard deviation, and must be integers (if specified) or real-valued. Furthermore, the dataset must produce correlated data which generates a correlation accurate to within three decimal places of the specified value. E.g. if a value of 0.5 is specified, then 0.499 or 0.501 would be considered acceptable.To do this, the algorithm needs to generate a dataset, check the correlation is within bounds, and if not, generate another dataset, check, and so on, until a dataset is generated that meets the specification. Depending upon the nature of the constraints imposed on the data, the number of such iterations may vary between 3 or 4, right up to 3000+. This parameter sets the number of

iterations to be carried out before the algorithm informs you that the specified number have taken place, and asks you whether you want to try another ***n*** iterations (***n*** being the value specified)), The default is set to **200**.

Let's assume we set a correlation value of 0.5, stay with the default axis constraints, and click on the **Compute** **button.** What we see is:

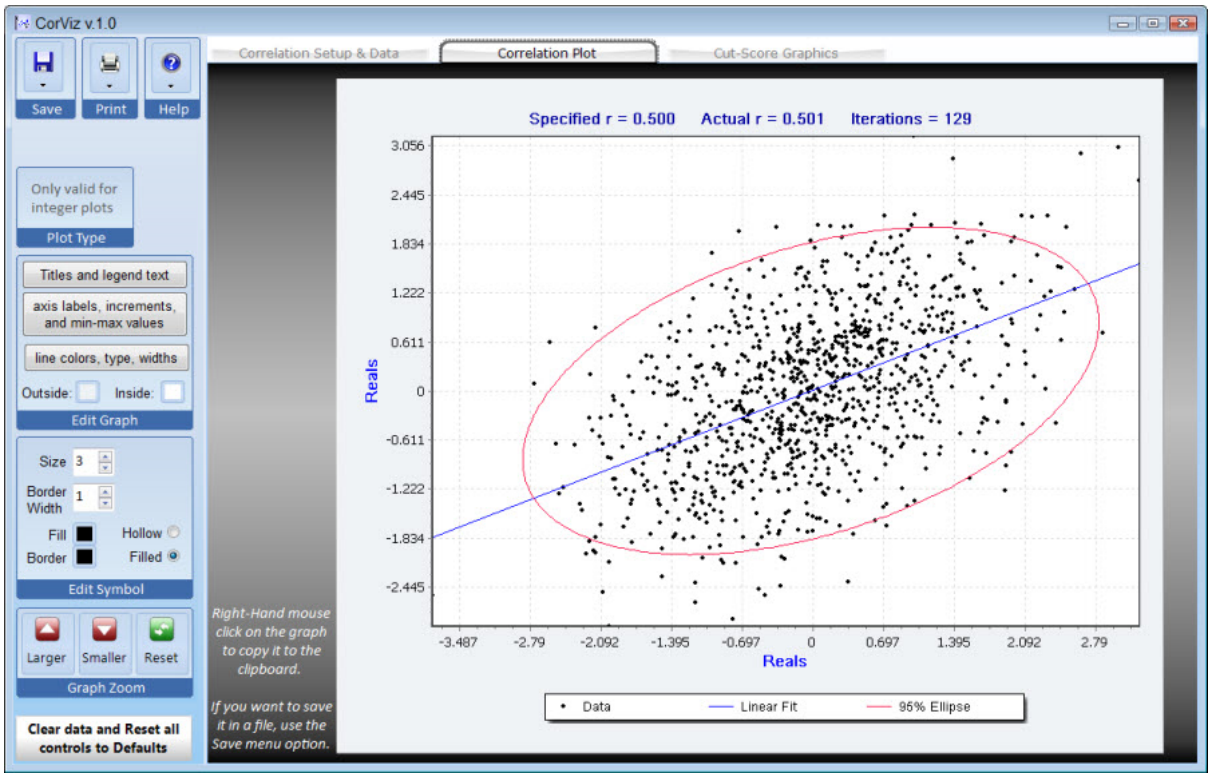


The "**Actual**" column now contains the parameter values of the generated 1000-case dataset. It also shows that data table containing the correlated data, and the linear regression equation for the data. You'll also note two new tabs have appeared:



The "Correlation Plot" tab shows the scatterplot of the data, the "Cut-Score Graphics" tab shows the same graph, but with the added facility of imposing cut-scores along with the automated computation of a variety of regression lines for subsets of the data.

Correlation Plot

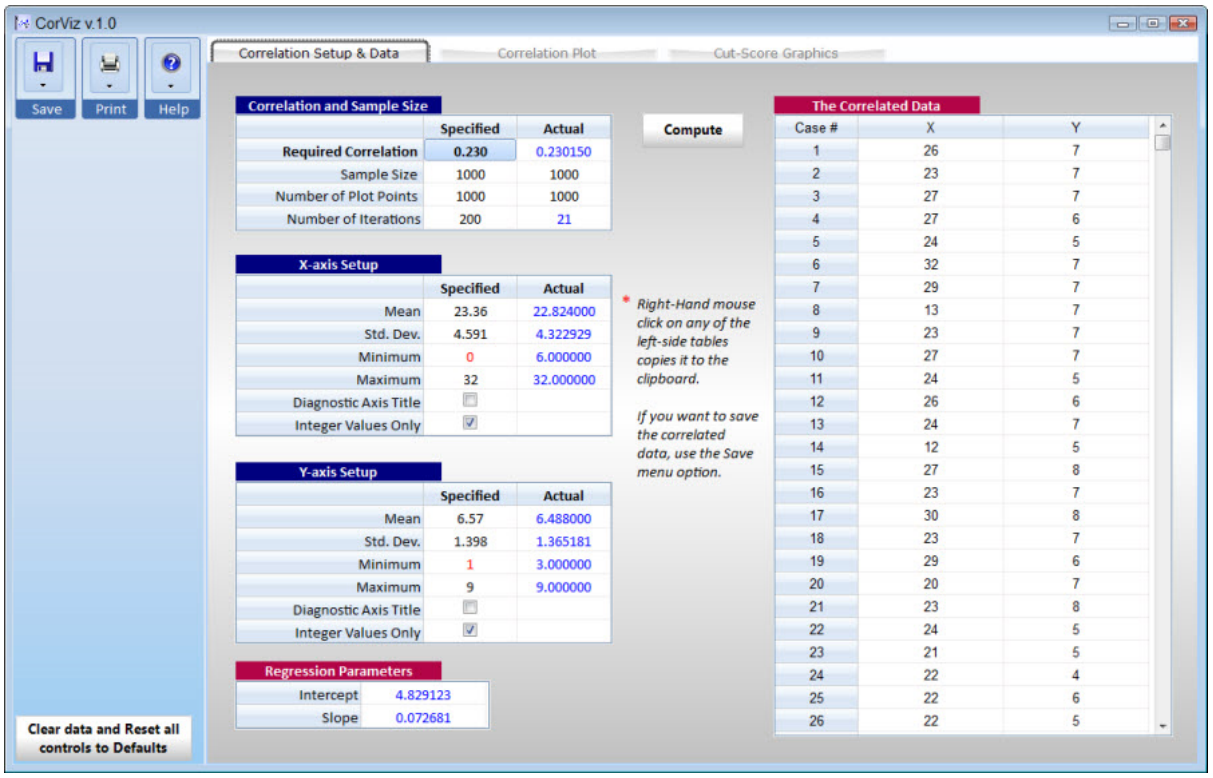


Here we see the scatterplot with the regression trend-line and a 95% **bivariate confidence ellipse**.

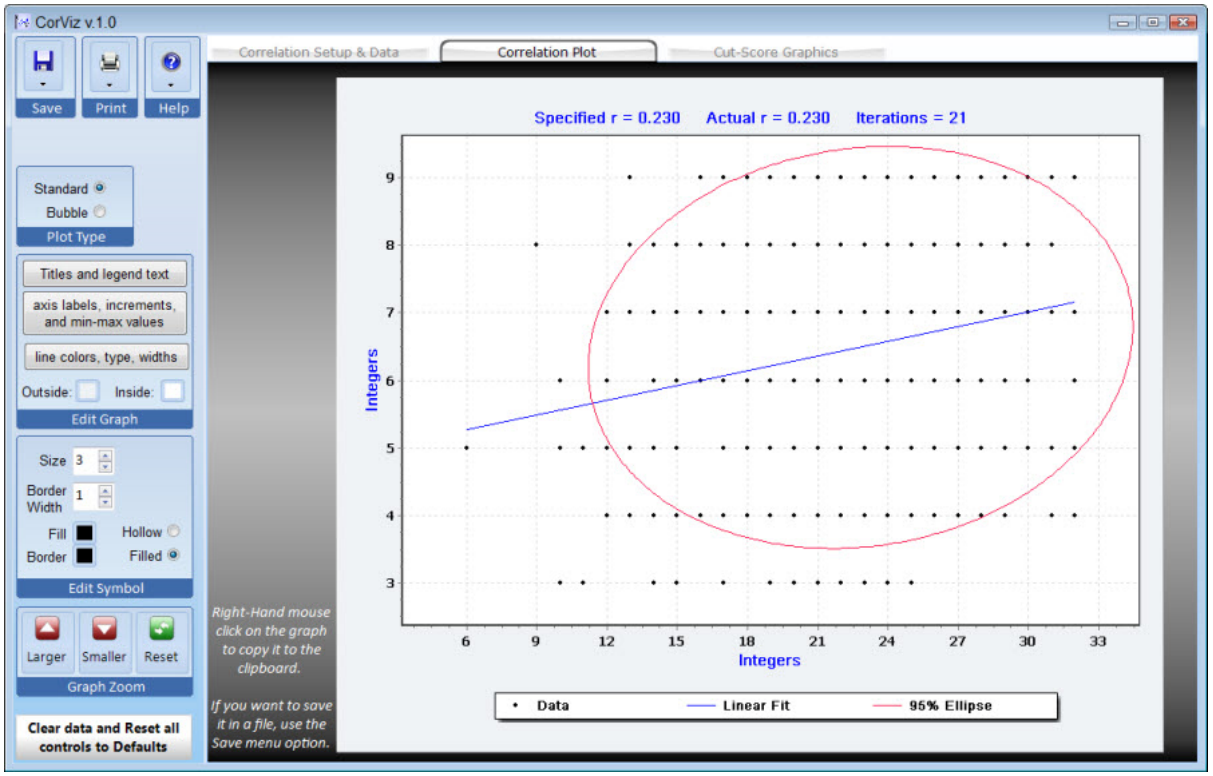
The confidence ellipse is the area within which 95% of all observations would be expected to be found when sampling data from a normal bivariate distribution. It's a useful and simple way of visually encapsulating the variability in the swarm of points in a correlation. As correlations increase in size, so does the ellipse become more narrow. The best way of using this ellipse is as an indication of the inherent variability in the relationship between two variables.

For example, let's say a consultant is recommending a test for 'customer focus', which has some validity data showing a correlation of 0.23 between the test scores and rated customer service performance (by supervisors). The test scores range from 0 to 32, the performance ratings between 1 to 9. Both are **integer** scales. The validity study reported a mean of 23.36 and SD of 4.591 for the test, with a mean of 6.57 and SD of 1.398 for the performance ratings.

The setup screen with the generated data values is:

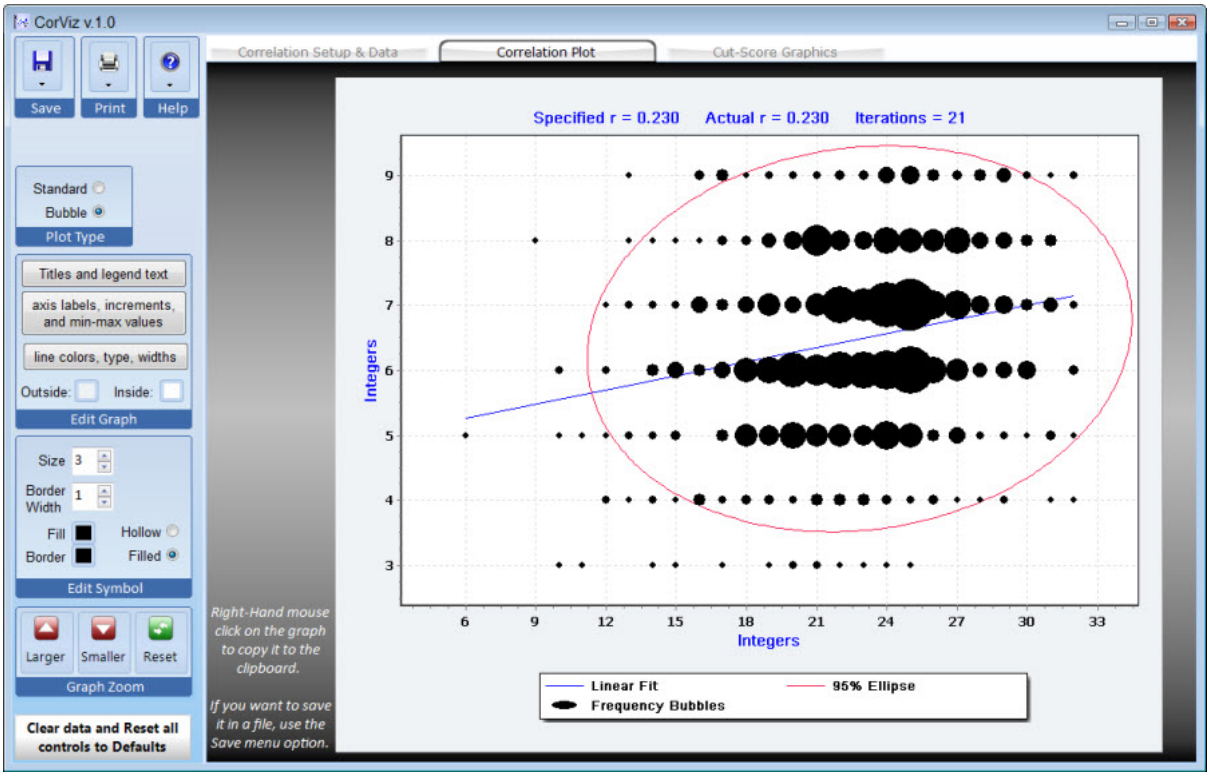


With scatterplot:

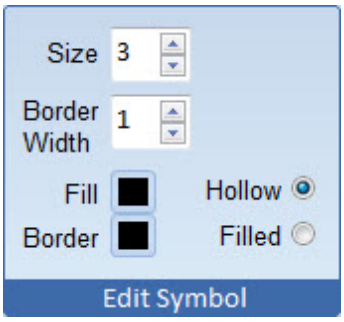


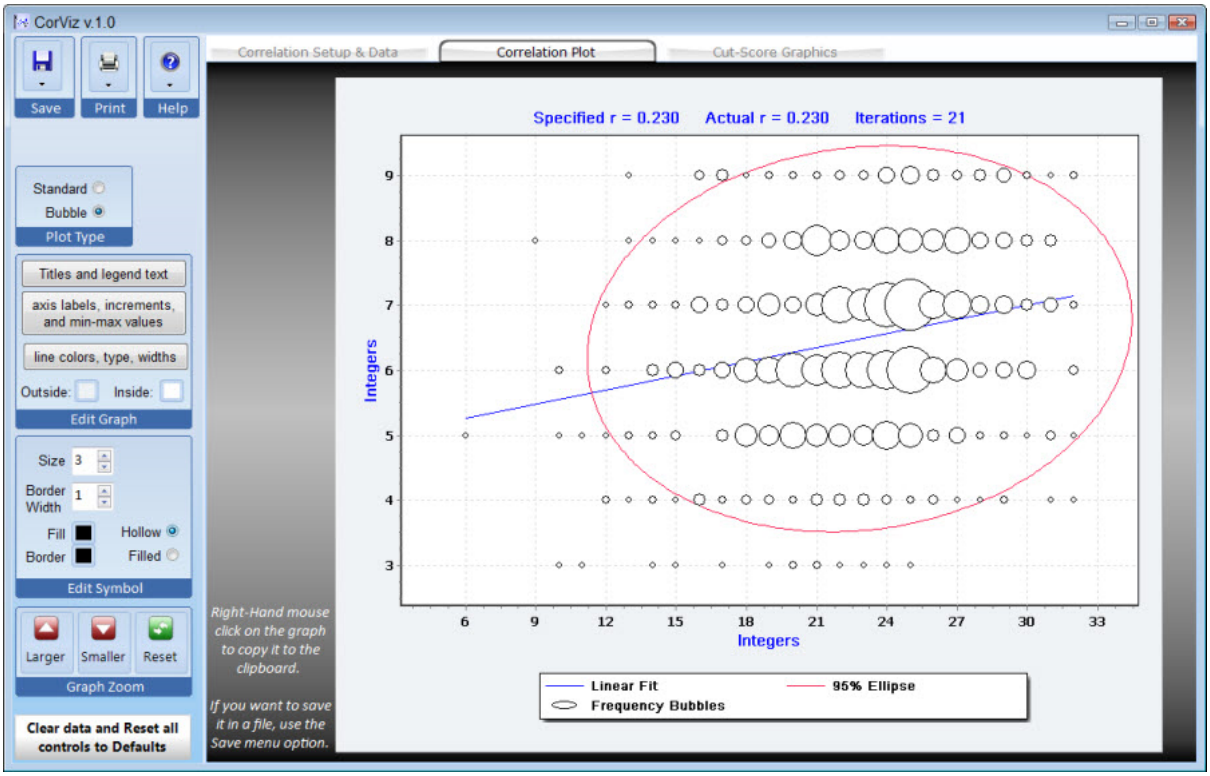
When plotting integers, you notice that standard graphs like the one above do not optimally convey where the bulk of observations are in such a plot. This is because many are exactly the same value in a constrained magnitude range, so they overlay one another. One way of showing where the bulk of observations are in such a plot is to re-plot the points as "**frequency bubbles**". The size of 'bubble' reflects the number of observations at each point; the larger the bubble, the more observations. You'll note that in the "controls" area for graphic images (which appears when a graph is being displayed), an option is presented for choice when both variables being plotted are integers:

If you click on bubble, the graph is redisplayed as:



You may prefer hollow circles .. selected by clicking on the control ..





Anyway, you get the point! You also see how inaccurate a correlation of 0.23 is if expecting, or being told, that you can use it to predict customer service performance from a test score. Yes, there is a trend, and yes, it's better than tossing a coin, but clearly to use such a test will need an appreciation of its propensity for inaccurate classification.

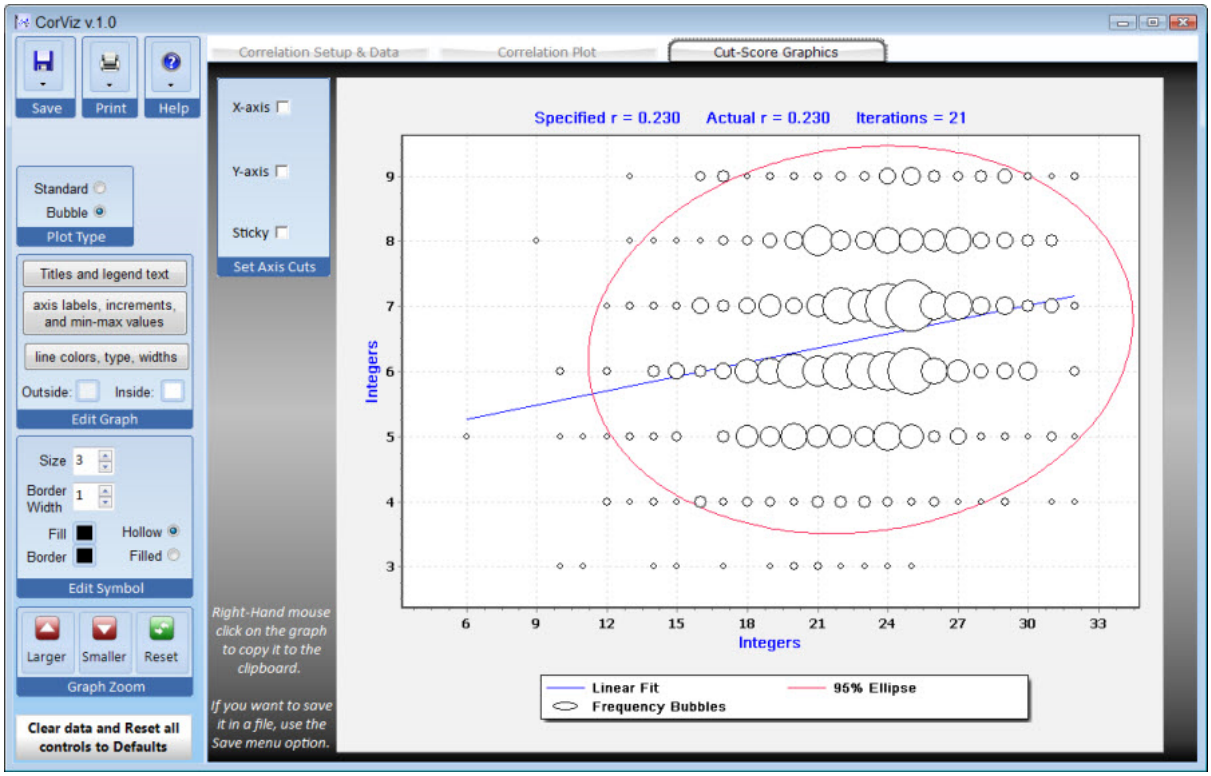
Finally, you can enter new data in the setup-screen (any data, correlations, descriptives etc.) without having to clear any values. The program automatically refreshes variables and other internal arrays etc. Handy when you want to see retain all descriptives but want to see how a new correlation dataset looks compared to a previous value computed under the same constraints.

2.2 Specifying Cut-Scores

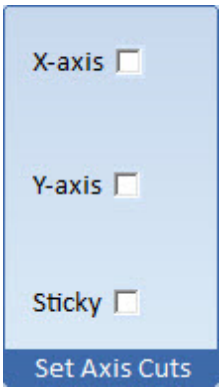
Continuing with our example from the previous section ... a consultant is recommending a test for 'customer focus', which has some validity data showing a correlation of 0.23

between the test scores and rated customer service performance (by supervisors). The test scores range from 0 to 32, the performance ratings between 1 to 9. Both are **integer** scales. The validity study reported a mean of 23.36 and SD of 4.591 for the test, with a mean of 6.57 and SD of 1.398 for the performance ratings.

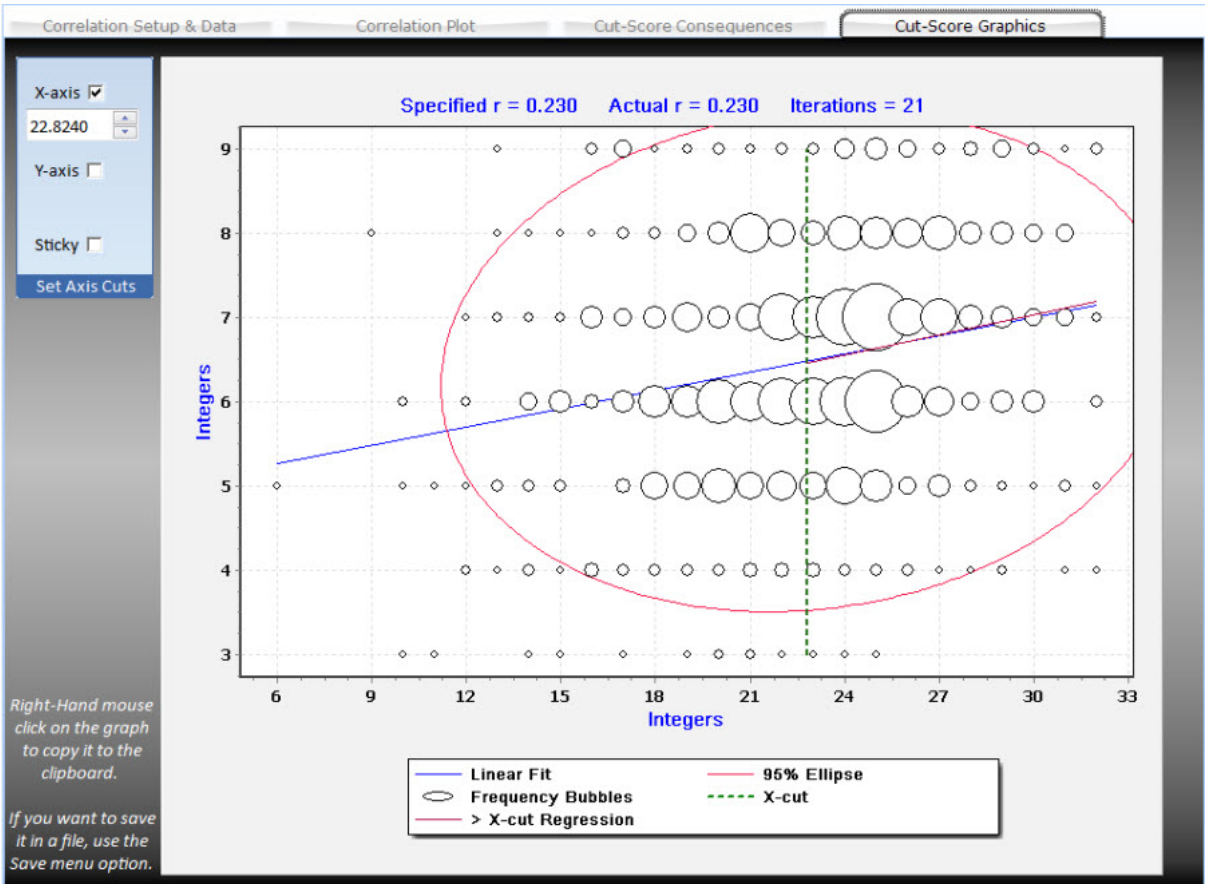
We can view the cut-score plot ..



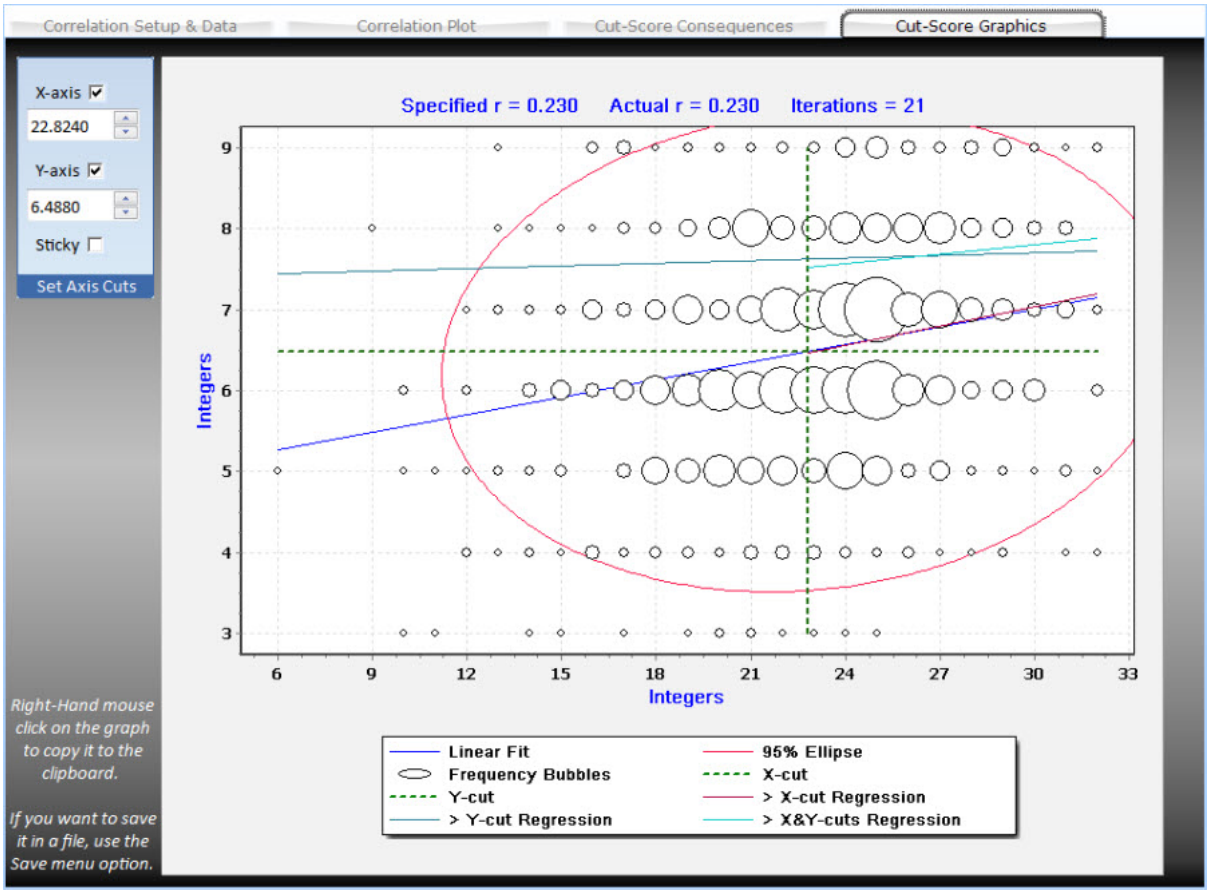
It's the same screen initially as for the scatterplot, except that a new control has appeared ..



This how you initially set a cut-score. By clicking on a checkbox, a green, dashed, cut-score line will immediately be drawn on the graph at the mean value for that variable.



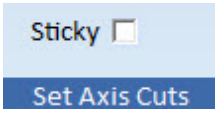
and



You'll also note several regression lines have now appeared - as per the legend. Each of these lines can be edited for width, color, and type (dashed or solid), or even hidden if you prefer less clutter or want to focus on a specific regression.

The new linear regression lines are:
> X-cut Regression: literally a regression computed on the subset of data which exceeds the X-cut value
> Y-cut Regression: a regression computed on the subset of data which exceeds the X-cut value
> X & Y-cuts Regression: a regression computed on the subset of data which exceeds both the X and Y-cut values (the 'selected' subset in the upper right-hand quadrant of the cut-score plot.

You can change the position of the lines by simply typing in new values into the cut-score 'spin-boxes'. The lines and all recalculations will be implemented and redisplayed automatically. You could also click on the arrows at the side of a spin-box which will move a cut-score repetitively in single integer steps, for a 'dynamic' visual consequence effect *(the regression lines reflect attenuation effects as you move through the range of a cut-score)*.



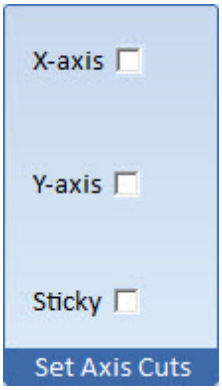
The checkbox, when set, preserves the chosen cut-scores if you decide to generate a new set of data. Purely a convenience.

A new tab has now appeared .. **Cut-Score Consequences**

This is the tab which contains the numerical analysis of the cut-score data - and where you can evaluate cut-score effects and compare test efficiencies.

2.3 Evaluating Cut-Score Effects

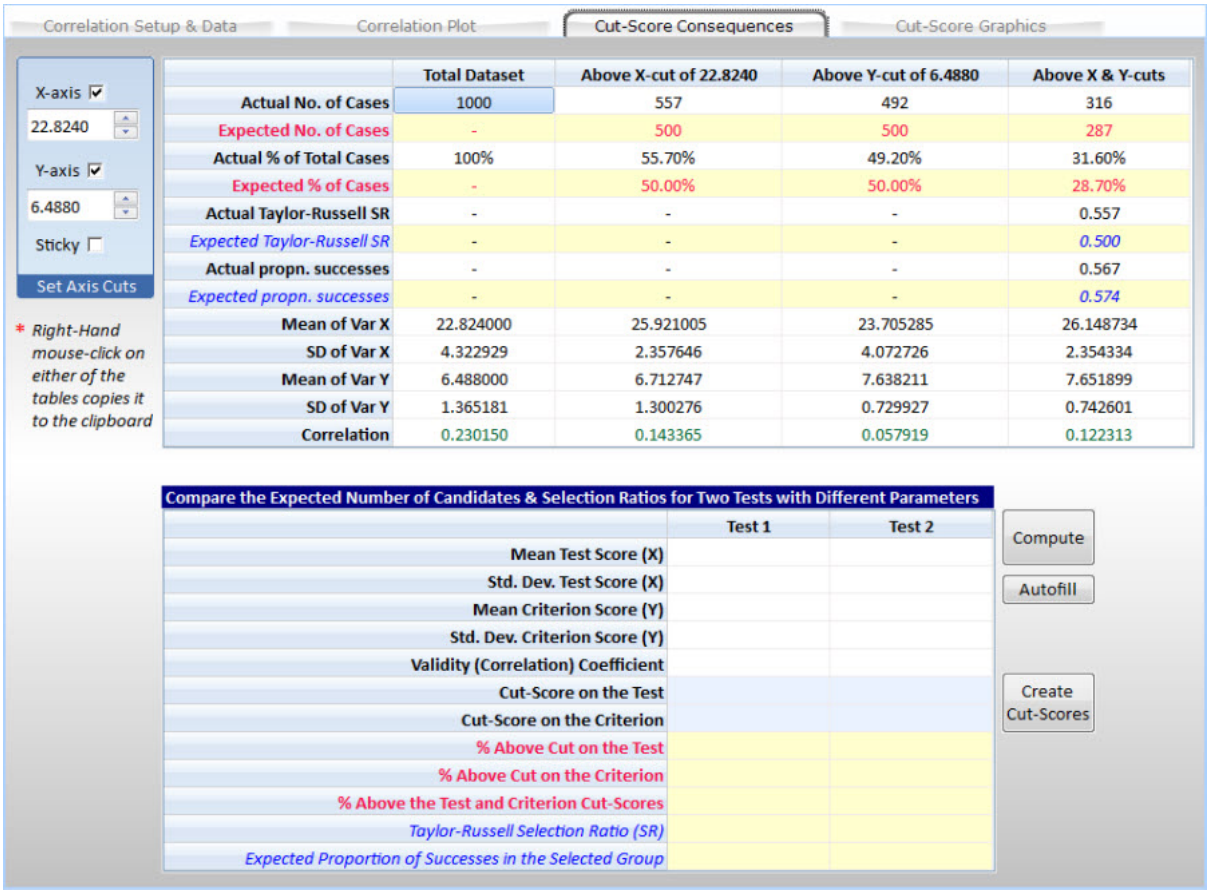
Graphically, one can evaluate cut-score effects by clicking on either of the cut-score checkbox fields:



This in turn enables the tab entitled "Cut-Score Consequences":

And it is in this tab that all cut-score numerical calculation stake place, including the comparison between tow test efficiencies.

Following on from the previous help-section, we click on the "Cut-Score Consequences" tab to reveal:



What we see is that for the displayed cut-scores, the relevant statistics for the chosen score subsets are presented in the upper table. In the final column are two sets of parameters, those computed from the empirical sample. and those computed from a bivariate normal distribution (as expected values). The Taylor-Russell notation Selection Ratio (SR) is also provided.

As you change the values of either cut-score, so will these statistics change.

The lower table is where you can enter any test statistics, and evaluate the change in "**Expected Proportion of Successes in the Selected Group**". That expected proportion refers to the proportion of cases who score above both the test (X) and criterion (Y) cut-scores, considered 'successes' in a selection scenario.

The 23-page application note available for free download provides an example of how this works, and how these values could have been found within the Taylor-Russell tables.

Another example:

The current selection test used by an organization has a validity of 0.30, how does the one you are hoping to sell to the organization compare? Your test has shown validities in several studies of around of 0.37.

Let's assume we continue using the variable values we set up for the example above. If we don't want to enter any variables by hand (or cut-and-paste them into the relevant fields), we can use the Autofill button. This fills the required fields with the setup data we originally specified in the Setup screen:

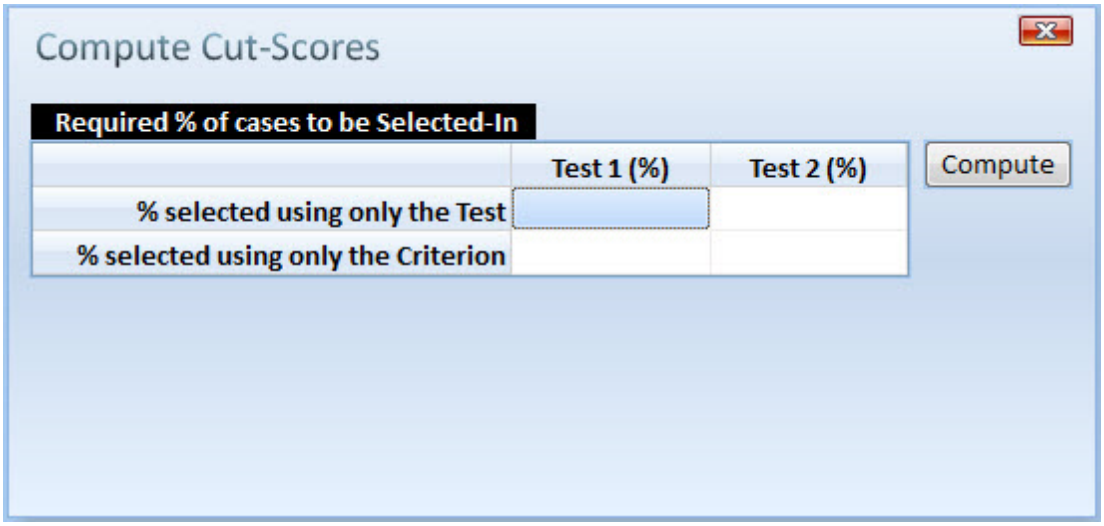
Compare the Expected Number of Candidates & Selection Ratios for Two Tests with Different Parameters		
	Test 1	Test 2
Mean Test Score (X)	23.360000	23.360000
Std. Dev. Test Score (X)	4.591000	4.591000
Mean Criterion Score (Y)	6.570000	6.570000
Std. Dev. Criterion Score (Y)	1.398000	1.398000
Validity (Correlation) Coefficient	0.230000	0.230000
Cut-Score on the Test		
Cut-Score on the Criterion		
% Above Cut on the Test		
% Above Cut on the Criterion		
% Above the Test and Criterion Cut-Scores		
Taylor-Russell Selection Ratio (SR)		
Expected Proportion of Successes in the Selected Group		

While this is OK for Test 1 *(except we need to change the correlation to 0.30)*, we may want/need to change the data for Test 2 (the new test). We keep the Criterion data (Y) the same for both tests.

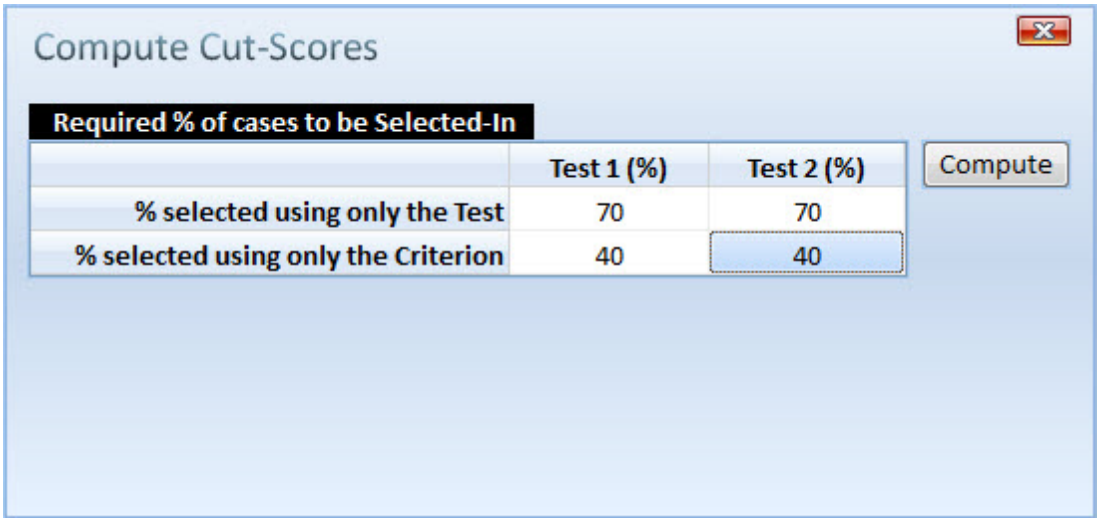
Compare the Expected Number of Candidates & Selection Ratios for Two Tests with Different Parameters		
	Test 1	Test 2
Mean Test Score (X)	23.360000	15.5
Std. Dev. Test Score (X)	4.591000	3.76
Mean Criterion Score (Y)	6.570000	6.570000
Std. Dev. Criterion Score (Y)	1.398000	1.398000
Validity (Correlation) Coefficient	.30	.37
Cut-Score on the Test		
Cut-Score on the Criterion		
% Above Cut on the Test		
% Above Cut on the Criterion		
% Above the Test and Criterion Cut-Scores		
Taylor-Russell Selection Ratio (SR)		
Expected Proportion of Successes in the Selected Group		

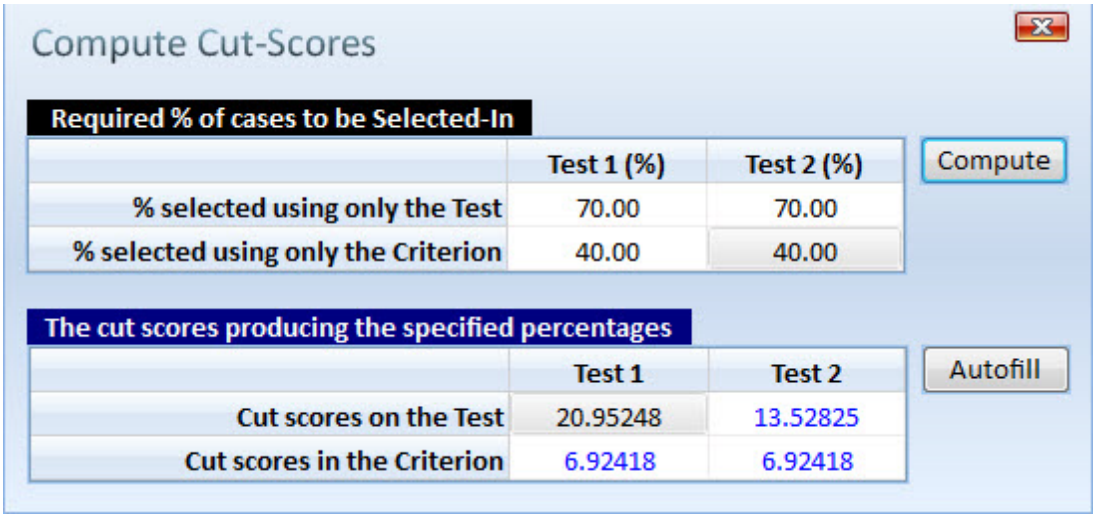
We know that the current criterion selects 40% of candidates as 'acceptable and above' performance. We also know that the current test selects 70% of cases scoring above a cut-score. So, we want to retain these selection proportions (and relevant cut-scores) for our new test, and evaluate how well our test will compare with regard to the "Expected Proportion of Successes in the Selected Group".

To do this, we click on the Create Cut-Scores button This opens up a utility which will calculate the cut-scores on the criterion and/or performance variable which will produce the desired cut-score proportions.



We enter our values .. and click on **Compute** ..





We see that the cut-scores are now provided for us, in the relevant metric of each variable. We can now 'autofill' the table on the cut-score consequence tab by clicking on the Autofill button on this screen.

Compare the Expected Number of Candidates & Selection Ratios for Two Tests with Different Parameters		
	Test 1	Test 2
Mean Test Score (X)	23.360000	15.5
Std. Dev. Test Score (X)	4.591000	3.76
Mean Criterion Score (Y)	6.570000	6.570000
Std. Dev. Criterion Score (Y)	1.398000	1.398000
Validity (Correlation) Coefficient	.30	.37
Cut-Score on the Test	20.952480	13.528250
Cut-Score on the Criterion	6.924180	6.924180
% Above Cut on the Test		
% Above Cut on the Criterion		
% Above the Test and Criterion Cut-Scores		
Taylor-Russell Selection Ratio (SR)		
Expected Proportion of Successes in the Selected Group		

We now click on the **Compute** button and see ...

Compare the Expected Number of Candidates & Selection Ratios for Two Tests with Different Parameters		
	Test 1	Test 2
Mean Test Score (X)	23.360000	15.500000
Std. Dev. Test Score (X)	4.591000	3.760000
Mean Criterion Score (Y)	6.570000	6.570000
Std. Dev. Criterion Score (Y)	1.398000	1.398000
Validity (Correlation) Coefficient	0.300000	0.370000
Cut-Score on the Test	20.952480	13.528250
Cut-Score on the Criterion	6.924180	6.924180
% Above Cut on the Test	70.00%	70.00%
% Above Cut on the Criterion	40.00%	40.00%
% Above the Test and Criterion Cut-Scores	31.99%	32.92%
Taylor-Russell Selection Ratio (SR)	*0.700*	*0.700*
Expected Proportion of Successes in the Selected Group	*0.457*	*0.470*

Compute | Autofill | Create Cut-Scores

Using the current test, we are expecting to select 45.7% of potentially successful cases in the selected group (above both cuts). With the new test, we would expect to see 47%.

Whether or not that increase of 1.3% justifies a change of test supplier is a matter for other considerations *(e.g. the cost-benefit of successes and failures and other less quantifiable advantages/deficits of deploying a new assessment etc.)*.

If the new test possessed a validity of 0.59, then the proportions might be:

Compare the Expected Number of Candidates & Selection Ratios for Two Tests with Different Parameters		
	Test 1	Test 2
Mean Test Score (X)	23.360000	15.500000
Std. Dev. Test Score (X)	4.591000	3.760000
Mean Criterion Score (Y)	6.570000	6.570000
Std. Dev. Criterion Score (Y)	1.398000	1.398000
Validity (Correlation) Coefficient	0.300000	0.600000
Cut-Score on the Test	20.952480	13.528250
Cut-Score on the Criterion	6.924180	6.924180
% Above Cut on the Test	70.00%	70.00%
% Above Cut on the Criterion	40.00%	40.00%
% Above the Test and Criterion Cut-Scores	31.99%	36.00%
Taylor-Russell Selection Ratio (SR)	*0.700*	*0.700*
Expected Proportion of Successes in the Selected Group	*0.457*	*0.514*

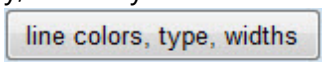
Using the current test, we are expecting to select 45.7% of potentially successful cases in the selected group (above both cuts). With the new test, we would expect to see 51.4%., an increase of 5.7%.

If we shifted the cut-score on the customer-focus selection test so that it selected 50% (instead of 70%) of candidates, while keeping all other values the same, we 'd see a 10.3% lift in successes proportions. That is worth knowing, although the number of candidates we are selecting is dropping, which may be of concern.

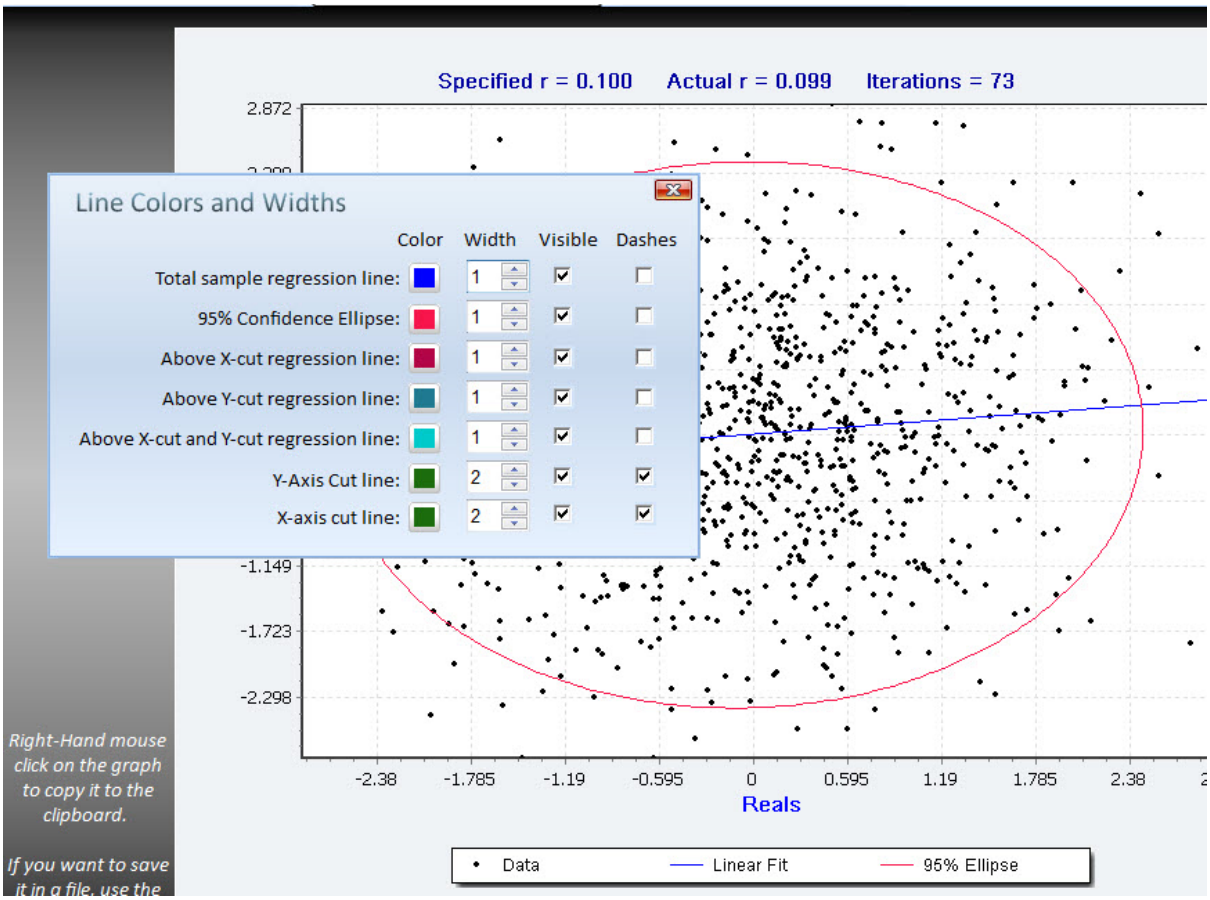
Compare the Expected Number of Candidates & Selection Ratios for Two Tests with Different Parameters		
	Test 1	Test 2
Mean Test Score (X)	23.360000	15.500000
Std. Dev. Test Score (X)	4.591000	3.760000
Mean Criterion Score (Y)	6.570000	6.570000
Std. Dev. Criterion Score (Y)	1.398000	1.398000
Validity (Correlation) Coefficient	0.300000	0.600000
Cut-Score on the Test	23.360000	15.500000
Cut-Score on the Criterion	6.924180	6.924180
% Above Cut on the Test	50.00%	50.00%
% Above Cut on the Criterion	40.00%	40.00%
% Above the Test and Criterion Cut-Scores	24.69%	29.87%
Taylor-Russell Selection Ratio (SR)	*0.500*	*0.500*
Expected Proportion of Successes in the Selected Group	*0.494*	*0.597*

Anyway, you get the idea. This software is built for "what-if" scenario exercises like this.

2.4 Changing Graphic Objects and Text Features

When a graph is on display, a variety of edit functions are available .. these are pretty self-obvious ... e.g. clicking on 

reveals a floating (moveable) edit screen ..



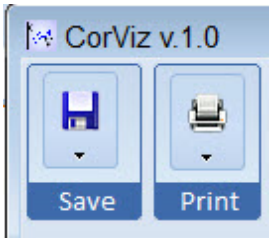
Any changes made on these screens have an immediate impact on both graphs (scatterplot and cut-score plot), regardless of which you happen to be editing. Likewise the other options such as symbol sizes and other features.

If you want to clear all data and reset all graphs back to their default colors, titles, fonts, etc., use the **Clear data and Reset all controls to Defaults** button located at the bottom left of the screen (always on display).

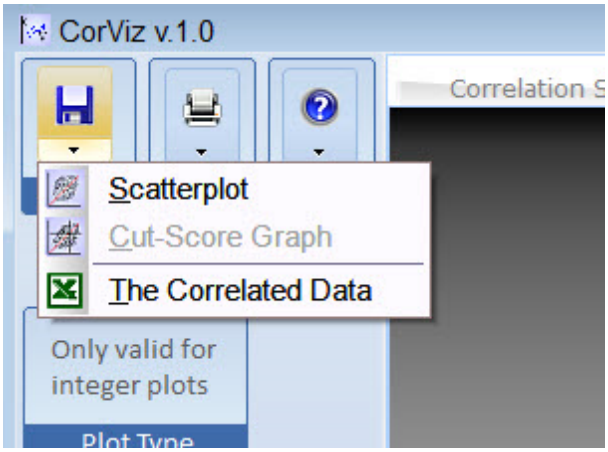
Saving/Printing Graphs and Tables

3 Saving/Printing Graphs and Tables

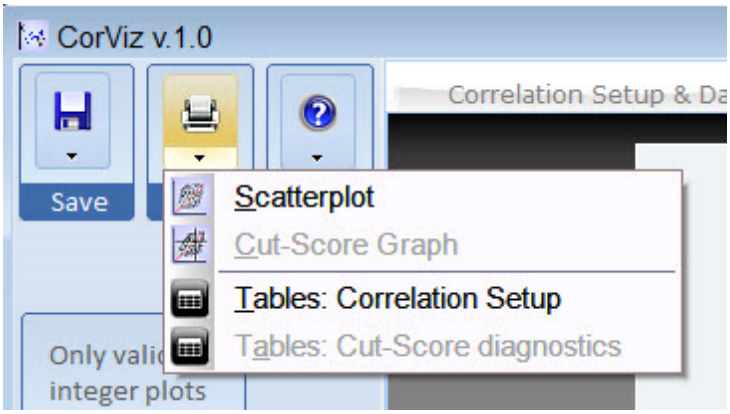
This is accomplished using two menus ..



which expand into...

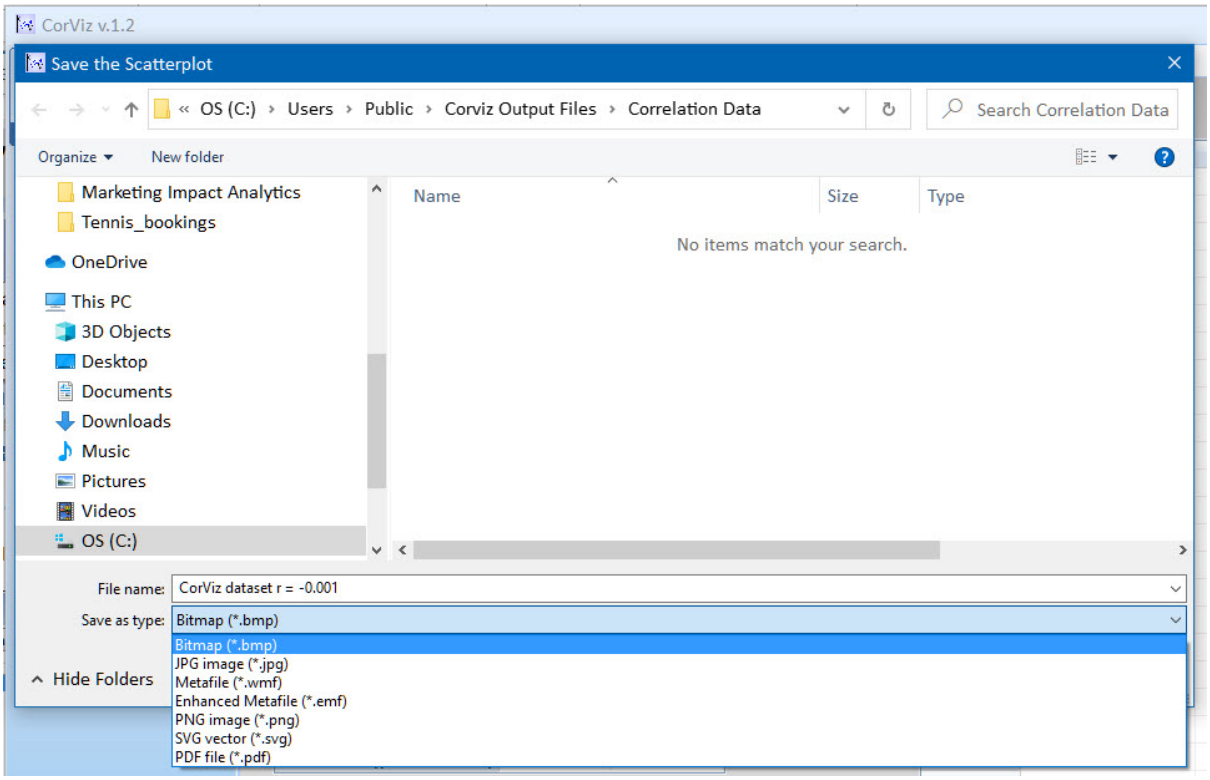


and



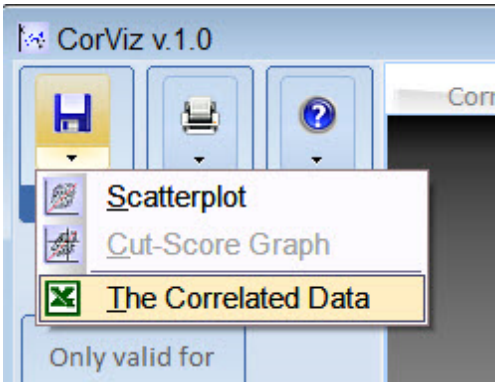
When available items are able to be saved or printed, the options are enabled in the menus.

For example, to save a scatterplot .. a save-file dialog will be presented, from which you can change the directory, filename, or one of 7 image types. If you change the directory, the program will remember that change for next time you wish to save any file.

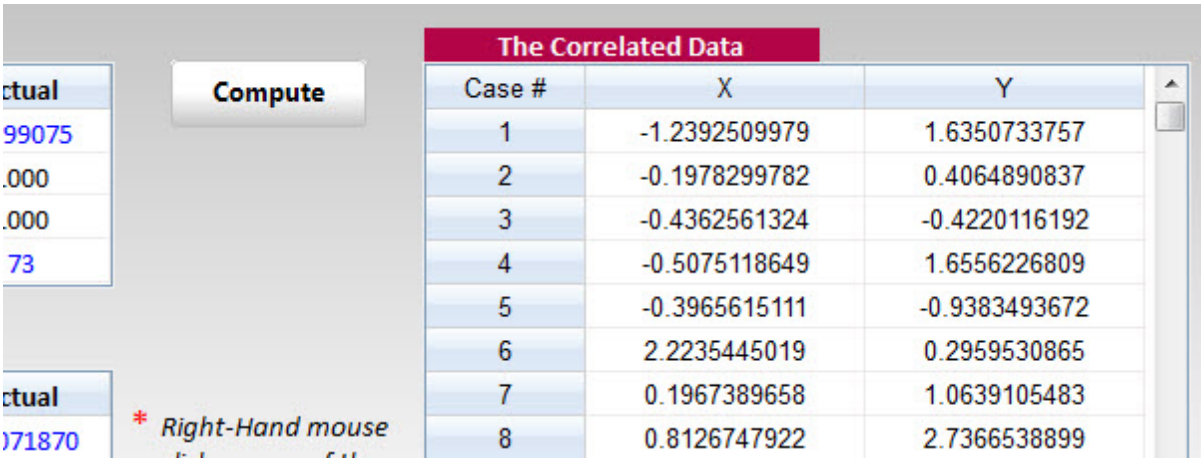


Recommended Correlated Data Save option

To save the Correlated Data table you see on the Setup screen into an Excel file (xls or xlsx), use the dedicated Excel menu option on the Save menu ..



This will save the data as per Correlated Data Table format ..



Note also, you can save all graphs or tables to the clipboard by using a right-hand button mouse click on each ..

As to printing, the Print menu takes care of this .. all essential tables and graphs can be printed. The Tables are nicely formatted, one page containing all the Correlation Setup tables, another for both cut-score tables.

What's new in v1.1

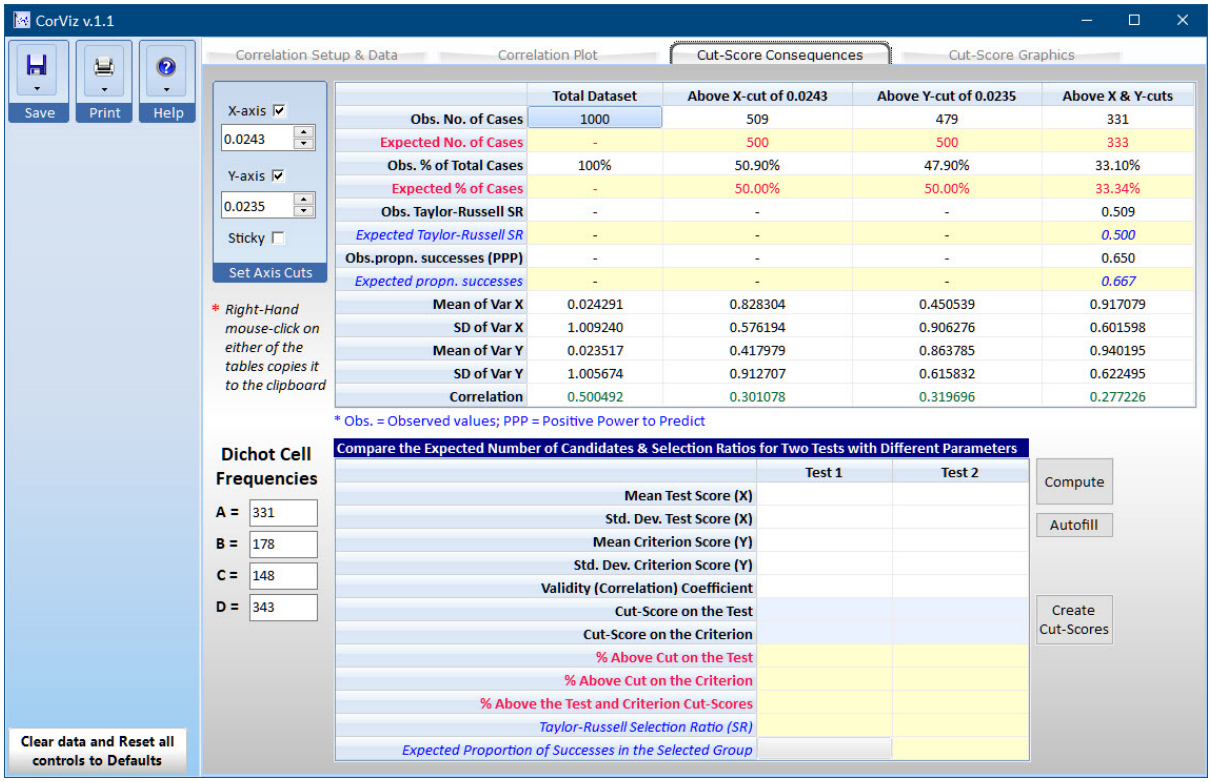
4 What's new in v1.1

[June 11th, 2017, and 11th May, 2021]

1. When computing **both** X and Y cut-score "consequences", I've added the cell frequencies for entry into my 2x2 table analysis program Dichot 3.2a into the on-screen display; so that they may easily be entered into the program for more in-depth actuarial 2x2 analysis. Any change made to either X or Y cut-scores is automatically updated in the cell-frequencies.

2. The cut-score analysis result table now uses 'Obs' instead of "Actual" as the description of results calculated on the sample data (contrasted with the Taylor-Russell expected values).

The new cut-score "consequences screen" now looks like:



Version 1.2

- Just tidied up the saving of the graphics
- Added 'last used directory' memory to the program
- Re-pathed the default output/savefile directory to a public area: C:\Users\Public\Corviz Output Files\Correlation Data

- B -

- C -

- D -

- F -

- I -

- N -

- S -