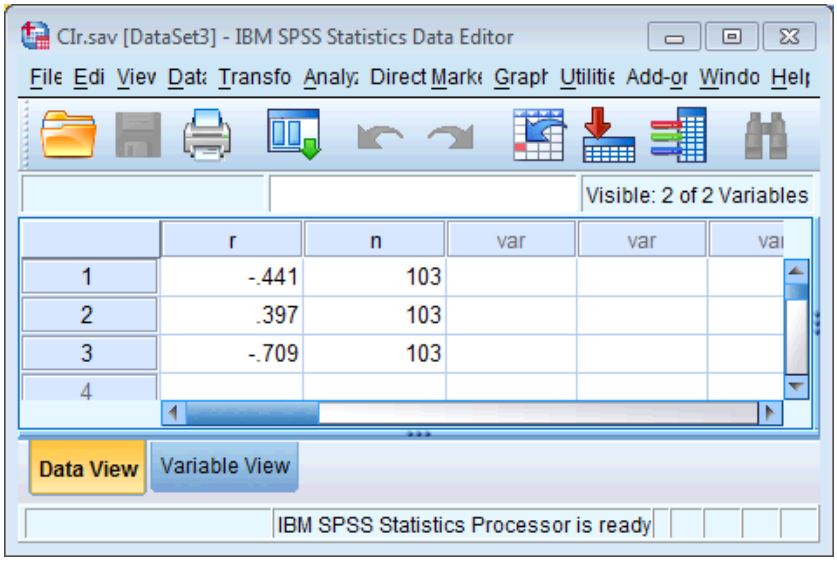
Chapter 7: Correlation

Oliver Twisted

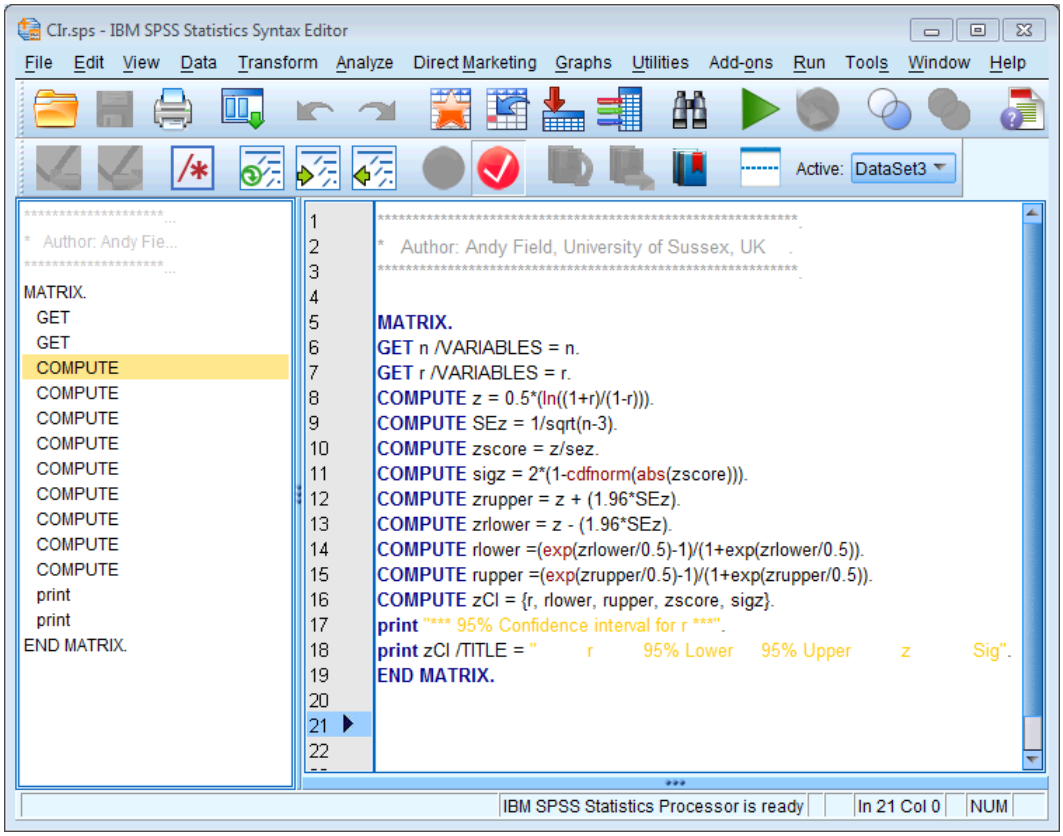
Please, Sir, can I have some more … confidence intervals?

To use this syntax open the data file **CIr.sav**. The data editor looks like this:



The values in the table are the correlations (in column r) and sample sizes (in column n) for the exam anxiety data (row 1 is the correlation of exam anxiety and exam performance, row 2 is revision and exam performance and row 3 is for the correlation between exam anxiety and revision). Use this data file to enter new correlations and sample sizes. You can enter as many different *r*s and *n*s as you like.

Now, open the syntax file **CIr.sps**. It will look like this:



Click on Run All to run these commands.

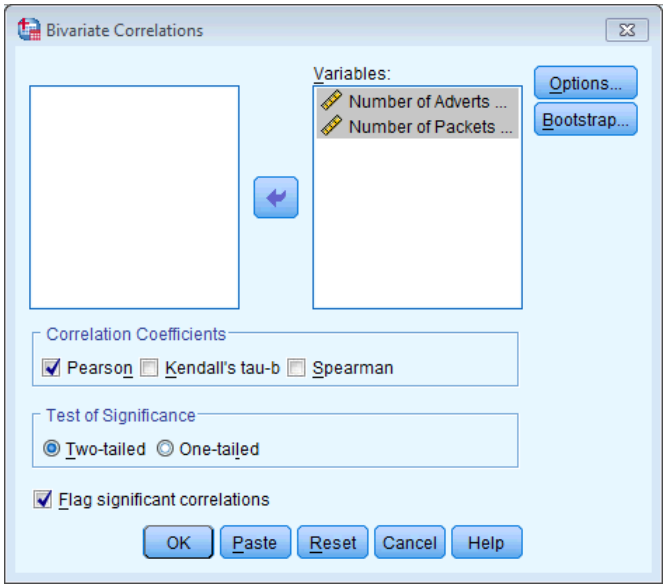
The output will look like this:

r	95% Lower	95% Upper	z	Sig
-.441000000	-.584632169	-.270563342	-4.734715614	.000002194
.397000000	.220405793	.548394607	4.200825813	.000026594
-.709000000	-.793632649	-.597448663	-8.851702080	.000000000

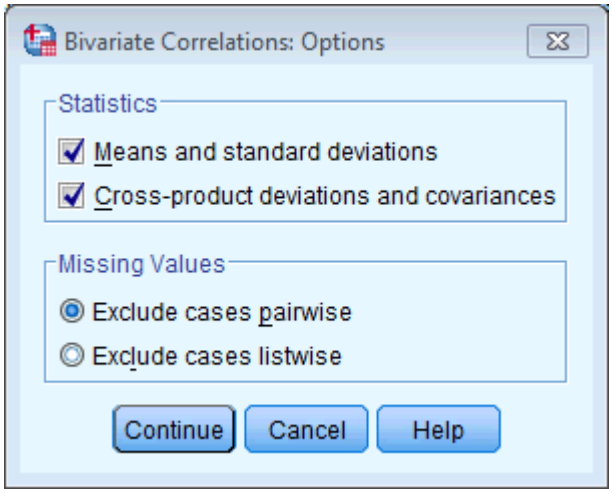
The first column is the original *r*, the next two columns show the lower and upper bounds of the 95% confidence interval for each *r* (these bounds are reported in the *r* metric so can be copied straight from the table). The value of *z* and the significance of *r* are included also, but you can also use the significance value from the original SPSS analysis.

Please, Sir, can I have some more … options?

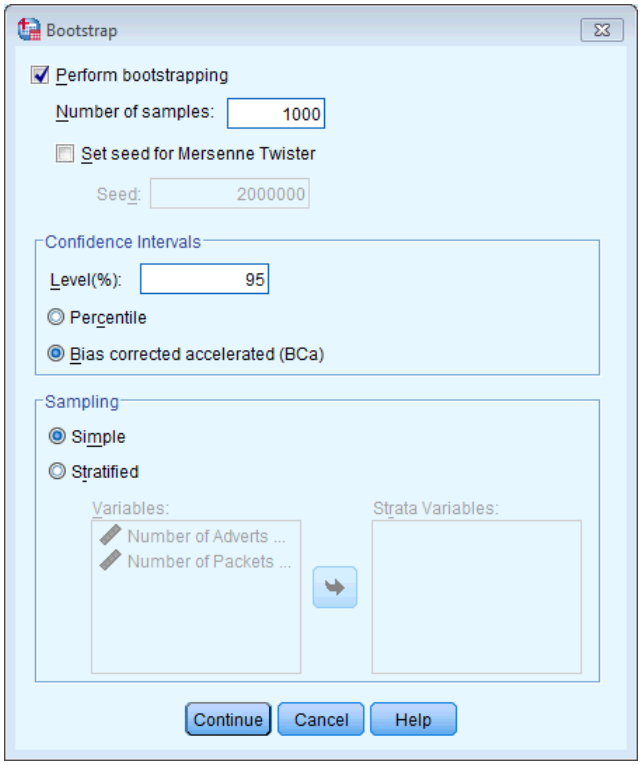
To illustrate what these options do, let's run the data from the advert example in the chapter. First, enter the data as in Figure 7.4 in the chapter. Then select Analyze Correlate Bivariate... to get this dialog box:



Select the variables **Number of Adverts Watched [adverts]** and **Number of Packets Bought [packets]** and transfer them to the variables list by clicking on  . Now click on  and another dialog box appears with two *Statistics* options and two *Missing Values* options. Select them for the advert data:



Next click on  to get some robust confidence intervals. Your completed dialog box should look like mine below:



Leave the default options in the main dialog box as they are. The resulting output from SPSS is shown below. The section labelled *Sum of Squares and Cross-products* shows us the cross-product (17 in this example) that we calculated from equation (7.2) in the chapter, and the sums of squares for each variable. The sums of squares are calculated from the top half of equation (7.1). The value of the covariance between the two variables is 4.25, which is the same value as was calculated from equation (7.2). The covariance within each variable is the same as the variance for each variable (so the variance for the number of adverts seen is 2.8, and the variance for the number of packets bought is 8.5). These variances can be calculated manually from equation (7.1). Also note that the Pearson correlation coefficient between the two variables is .871, which is the same as the value we calculated in the chapter. Underneath is the significance value of this coefficient (.054). We also have the BCa 95% confidence interval that ranges from –1.00 to 1.00. The fact that the confidence interval crosses zero (and the significance is greater than .05) tells us that there was not a significant relationship between the number of adverts watched and packets bought.

Correlations

				Number of Adverts Watched	Number of Packets Bought
Number of Adverts Watched	Pearson Correlation			1	.871
	Sig. (2-tailed)				.054
	Sum of Squares and Cross-products			11.200	17.000
	Covariance			2.800	4.250
	N			5	5
	Bootstrap[d]	Bias		0[e]	-.123[e]
		Std. Error		0[e]	.475[e]
		BCa 95% Confidence Interval	Lower	.[e]	-1.000[e]
			Upper	.[e]	1.000[e]
Number of Packets Bought	Pearson Correlation			.871	1
	Sig. (2-tailed)			.054	
	Sum of Squares and Cross-products			17.000	34.000
	Covariance			4.250	8.500
	N			5	5
	Bootstrap[d]	Bias		-.123[e]	0
		Std. Error		.475[e]	0
		BCa 95% Confidence Interval	Lower	-1.000[e]	.
			Upper	1.000[e]	.

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

c. Cannot be computed because at least one of the variables is constant.

d. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

e. Based on 985 samples

Please, Sir, can I have some more … biserial correlation?

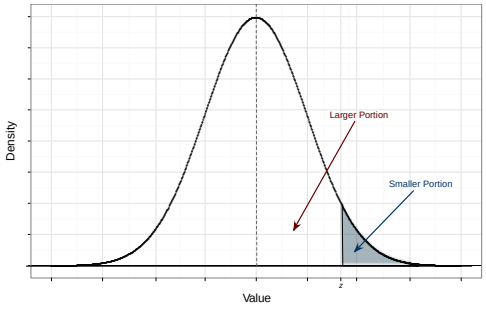
Imagine now that we wanted to convert the point-biserial correlation into the biserial correlation coefficient (r_b) (because some of the male cats were neutered and so there might be a continuum of maleness that underlies the gender variable). We must use equation (1) below in which *p* is the proportion of cases that fell into the largest category and *q* is the proportion of cases that fell into the smallest category. Therefore, *p* and *q* are simply the number of male and female cats. In this equation, *y* is the ordinate of the normal distribution at the point where there is a proportion *p* (or 100*p*%) of the area on one side and a proportion *q* on the other (this will become clearer as we do an example):

$$r_b = \frac{r_{pb}\sqrt{pq}}{y} \quad (1)$$

To calculate *p* and *q*, access the *Frequencies* dialog box using Analyze Descriptive Statistics Frequencies... and select the variable **Gender**. There is no need to click on any further options as the defaults will give you what you need to know (namely the percentage of male and female cats). It turns out that 53.3% (.533 as a proportion) of the sample was female (this is *p*, because it is the largest portion) while the remaining 46.7% (.467 as a proportion) were male (this is *q* because it is the smallest portion). To calculate *y,* we use these values and the values of the normal distribution displayed in the Appendix. The extract from the table

shown below shows how to find *y* when the normal curve is split with .467 as the smaller portion and .533 as the larger portion. It shows which columns represent *p* and *q*, and we look for our values in these columns (the exact values of .533 and .467 are not in the table, so we use the nearest values that we can find, which are .5319 and .4681 respectively). The ordinate value is in the column *y* and is .3977.

A.1. Table of the standard normal distribution



z	Larger Portion	Smaller Portion	y	z	Larger Portion	Smaller Portion	y
.00	.50000	.50000	.3989	.12	.54776	.45224	.3961
.01	.50399	.49601	.3989	.13	.55172	.44828	.3956
.02	.507	Value of *q*	.3989	.14	.55567	.44433	.3951
.03	.51197	.4 803	.3988	.15	.55962	.44038	.3945
.04	Value of *p*	.48405		16	.56356	.43644	.3939
.05	.51994	.48006	.398	.17	.56749	.43251	.3932
.06	.52392	.47608	.3982	.18	.57142	.42858	.3925
.07	.52790	.47210	.3980	.19	.57535	.42465	.3918
.08	.53188	.46812	.3977	.20	.57926	.42074	.3910
.09	.53586	.46414	.3973	.21	.58317	.41683	.3902
.10	.53983	.46017	.3970	.22	.58706	.41294	.3894
.11	.54380	.45620	.3965	.23	.59095	.40905	.3885

Value of *y*

Getting the 'ordinate' of the normal distribution

If we replace these values in equation (1) we get .475 (see below), which is quite a lot higher than the value of the point-biserial correlation (.378). This finding just shows you that whether you assume an underlying continuum or not can make a big difference to the size of effect that you get:

$$r_b = \frac{r_{pb}\sqrt{pq}}{y}$$
$$= \frac{0.378\sqrt{0.533 \times 0.467}}{0.3977}$$
$$= .475$$

If this process freaks you out, you can also convert the point-biserial *r* to the biserial *r* using a table published by Terrell (1982b) in which you can use the value of the point-biserial correlation (i.e., Pearson's *r*) and *p*, which is just the proportion of people in the largest group (in the above example, .533). This spares you the trouble of having to work out *y* in

the above equation (which you're also spared from using). Using Terrell's table we get a value in this example of .48, which is the same as we calculated to 2 decimal places.

To get the significance of the biserial correlation we need to first work out its standard error. If we assume the null hypothesis (that the biserial correlation in the population is zero) then the standard error is given by (Terrell, 1982a):

$$SE_{r_b} = \frac{\sqrt{pq}}{y\sqrt{N}} \quad (2)$$

This equation is fairly straightforward because it uses the values of *p*, *q* and *y* that we already used to calculate the biserial *r*. The only additional value is the sample size (*N*), which in this example was 60. So, our standard error is:

$$SE_{r_b} = \frac{\sqrt{0.533 \times 0.467}}{0.3977\sqrt{60}} = 0.162$$

The standard error helps us because we can create a *z*-score. To get a *z*-score we take the biserial correlation, subtract the mean in the population and divide by the standard error. We have assumed that the mean in the population is 0 (the null hypothesis), so we can simply divide the biserial correlation by its standard error:

$$z_{r_b} = \frac{r_b - \bar{r}_b}{SE_{r_b}} = \frac{r_b - 0}{SE_{r_b}} = \frac{r_b}{SE_{r_b}} = \frac{0.475}{0.162} = 2.93$$

We can look up this value of *z* (2.93) in the table for the normal distribution in the Appendix and get the one-tailed probability from the column labelled 'Smaller Portion'. In this case the value is .00169. To get the two-tailed probability we multiply this value by 2, which gives us .00338.

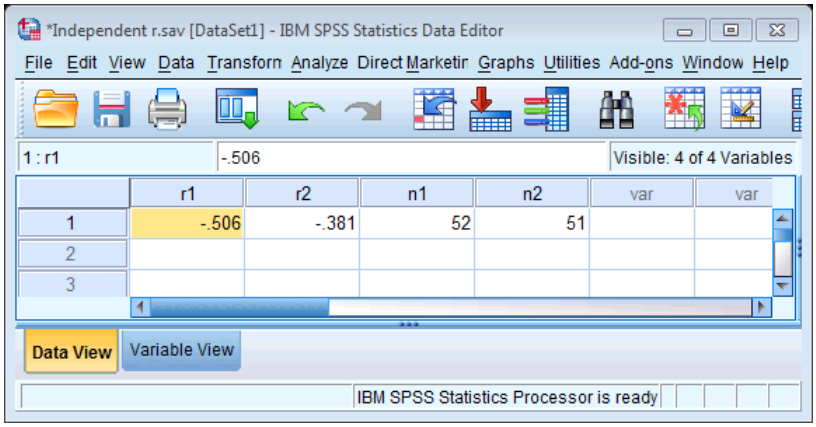
References

Terrell, C. D. (1982a). Significance tables for the biserial and the point biserial. *Educational and Psychological Measurement*, *42*, 975–981.

Terrell, C. D. (1982b). Table for converting the point biserial to the biserial. *Educational and Psychological Measurement*, *42*, 982–986.

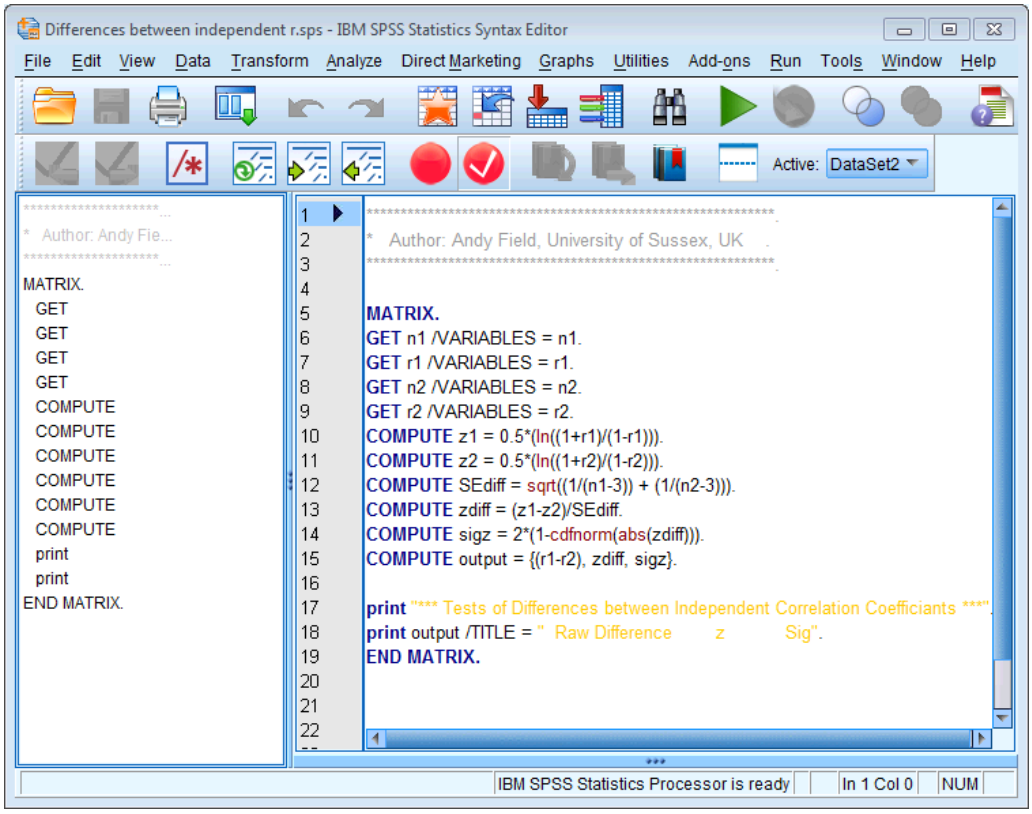
Please, Sir, can I have some more … comparing of correlations?

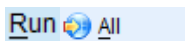
If you want to compare two independent *r*s (i.e. *r*s measuring the same thing in two different samples) then open the data file **Independent r.sav**. The data editor looks like this:



The values in the data editor are the correlations for the example in the book. Column r1 has the correlation for the males, r2 the correlation for the females, and n1 and n2 are the respective sample sizes (see the book chapter). Use this data file to enter new correlations and sample sizes. You can enter as many different *r*s and *n*s as you like down the rows.

Now, open the syntax file **Differences between independent r.sps**. It will look like this:



Click on  to run these commands.

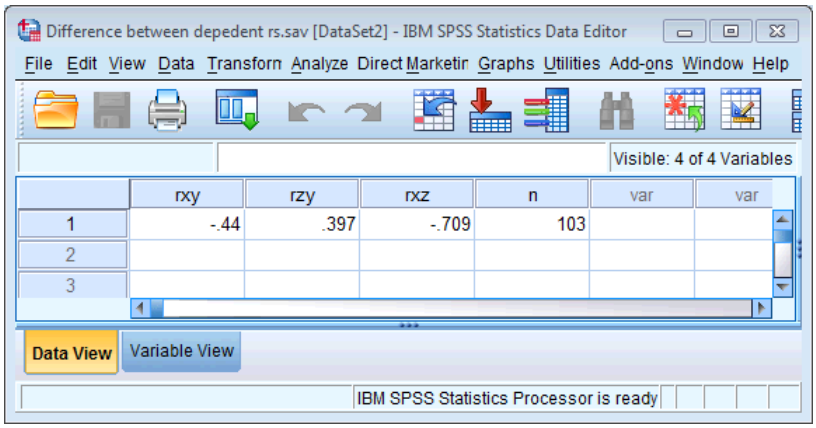
The output will look like this:

```
  Raw Difference          z              Sig

-.1250000000   -.7687093063    .4420658990
```

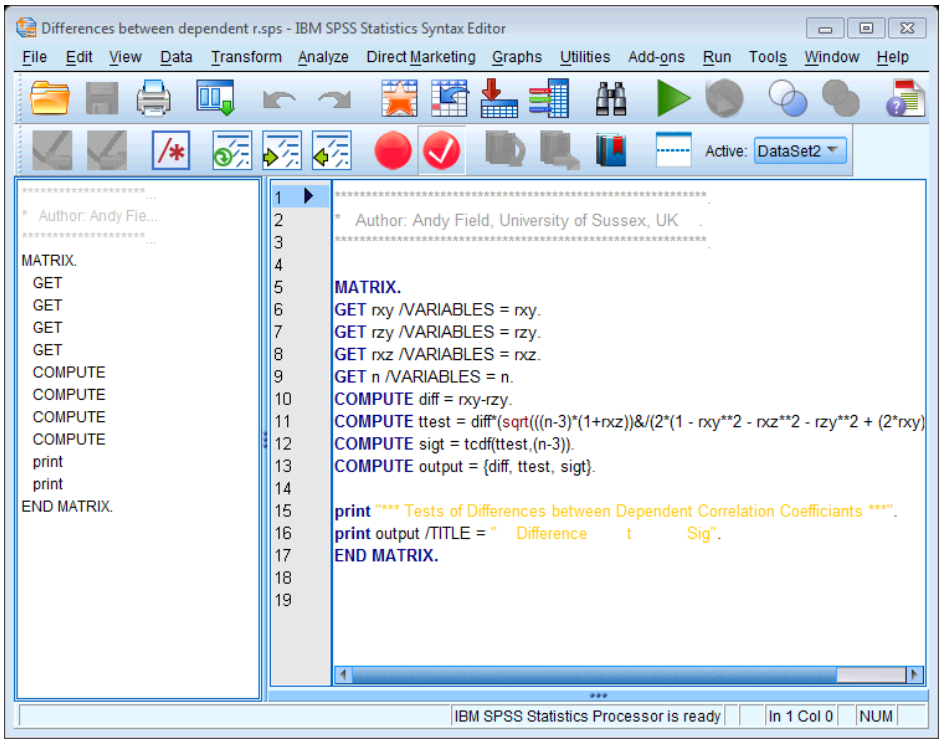
The first column is the differences between the male and female *rs*, the next two columns show the value of *z* and its two-tailed significance.

If you want to compare dependent *r*s then open the data file **Differences between dependent rs.sav**. The data editor looks like this:



The values in the data editor are the correlations for the example in the book for the exam anxiety data. The columns are labelled rxy, rzy, rxz which correspond to the explanations in the book chapter. Just type in the correlation coefficients for the variables that you want to compare, and enter the sample size in the column labelled *n*. You can enter as many different *r*s and *n*s as you like down the rows.

Now, open the syntax file **Differences between dependent r.sps**. It will look like this:



Click on Run All to run these commands.

The output will look like this:

```
    Difference         t            Sig

  -.838000000  -5.095768225     .000000822
```

The first column is the differences between the two correlations being compared (rxy and rzy), the next two columns show the value of *t* and its two-tailed significance. The value of *t* should correspond to the value calculated in the chapter.