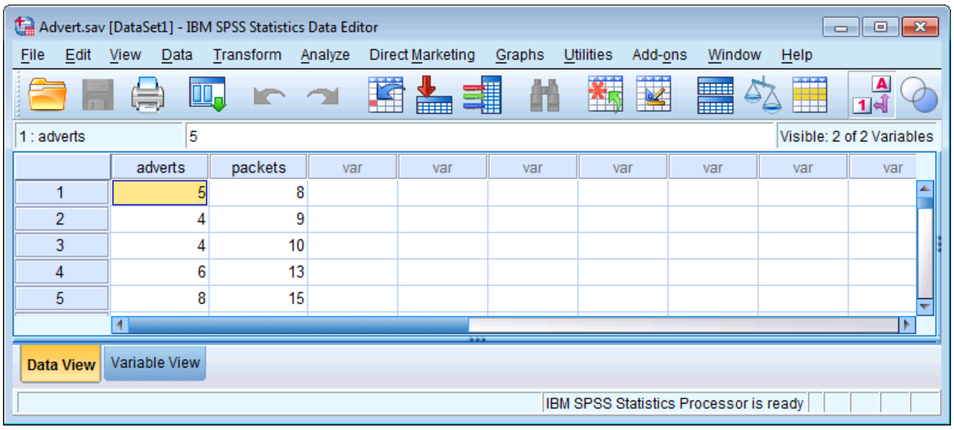
Correlation in IBM SPSS Statistics

Data entry for correlation analysis using SPSS

Imagine we took five people and subjected them to a certain number of advertisements promoting toffee sweets, and then measured how many packets of those sweets each person bought during the next week. The data are in Table 1. We could see how strong the relationship is between these variables.

Data entry when looking at relationships between variables is straightforward because each variable is entered in a separate column. So, for each variable you have measured, create a variable in the data editor with an appropriate name, and enter a participant's scores across one row of the data editor. There may be occasions on which you have one or more categorical variables (such as gender) and these variables can also be entered in a column (but remember to define appropriate value labels). As an example, if we wanted to calculate the correlation between the two variables in Table 1 we would enter these data as in Figure 1. You can see that each variable is entered in a separate column, and each row represents a single individual's data (so the first consumer saw 5 adverts and bought 8 packets).



1 : adverts 5

Visible: 2 of 2 Variables

	adverts	packets	var	var	var	var	var	var	var
1	5	8							
2	4	9							
3	4	10							
4	6	13							
5	8	15							

IBM SPSS Statistics Processor is ready

Figure 1: Data entry for correlation

Table 1: some advertising data

Participant:	1	2	3	4	5	Mean	S
Adverts Watched	5	4	4	6	8	5.4	1.67
Packets Bought	8	9	10	13	15	11.0	2.92

Bivariate correlation

Figure 2 from Field (2013) shows a general procedure when considering computing a bivariate correlation coefficient. In Field (2013), I look at an example relating to exam anxiety: a psychologist was interested in the effects of exam stress and revision on exam performance. She had devised and validated a questionnaire to assess state anxiety relating to exams (called the Exam Anxiety

Questionnaire, or EAQ). This scale produced a measure of anxiety scored out of 100. Anxiety was measured before an exam, and the percentage mark of each student on the exam was used to assess the exam performance. She also measured the number of hours spent revising. These data are in **Exam Anxiety.sav**. In my book I show how to look at scatterplots and other graphs exploring assumptions of the test for these data.

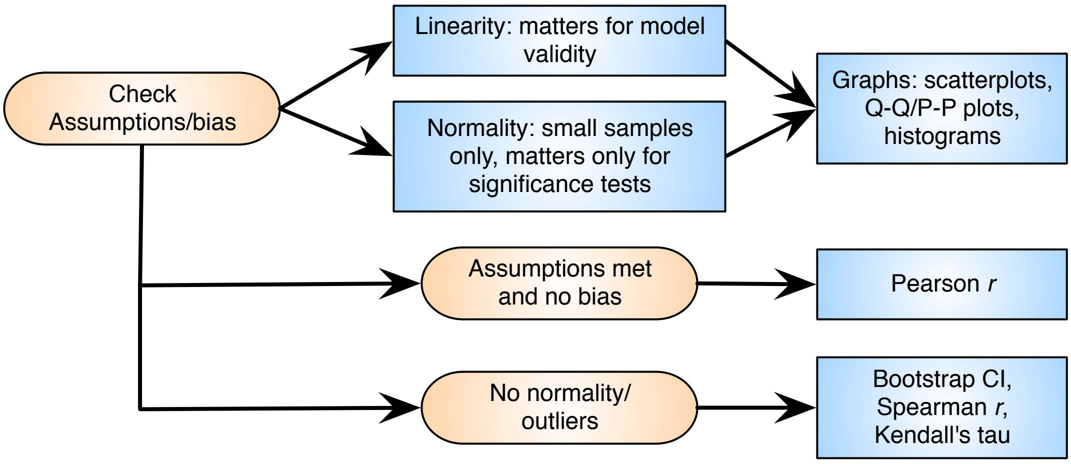
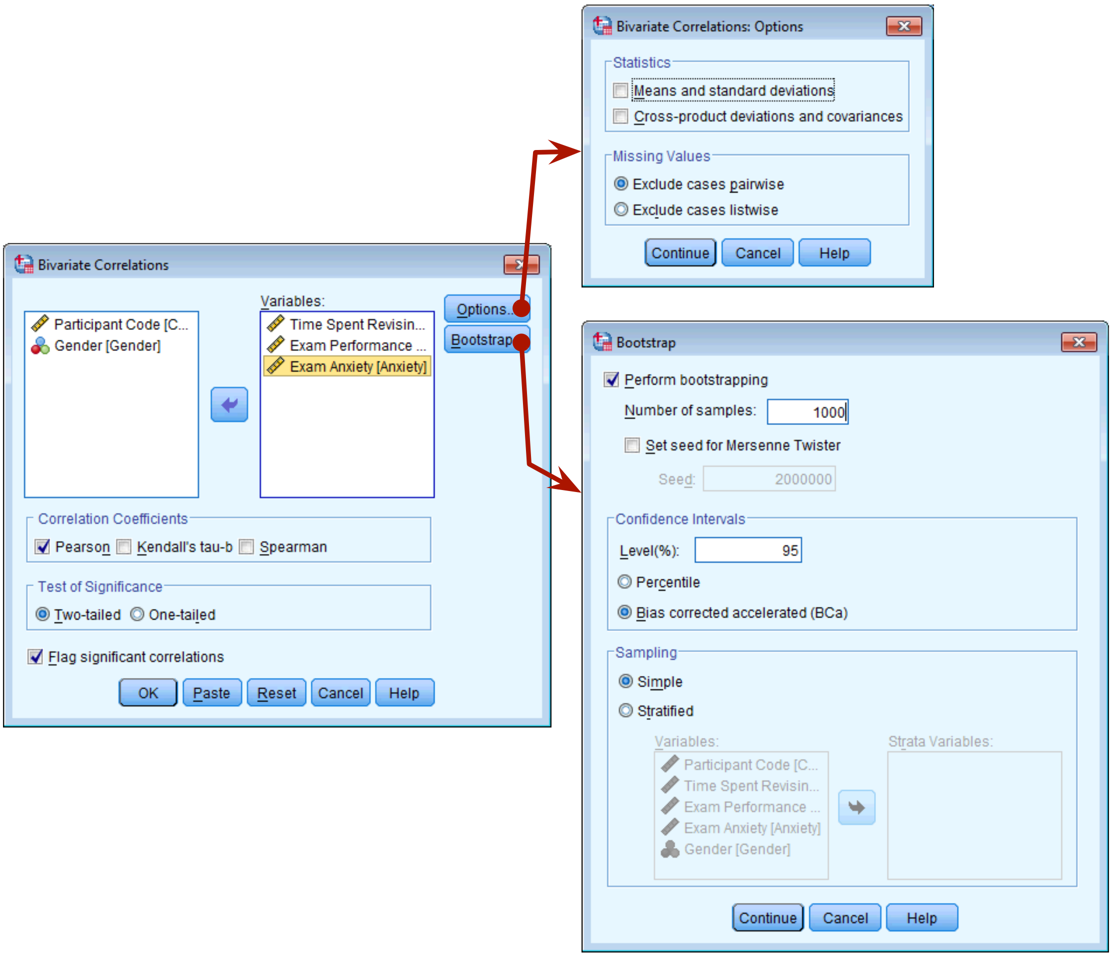

Check Assumptions/bias
Linearity: matters for model validity
Normality: small samples only, matters only for significance tests
Graphs: scatterplots, Q-Q/P-P plots, histograms
Assumptions met and no bias
Pearson *r*
No normality/ outliers
Bootstrap CI, Spearman *r*, Kendall's tau

Figure 2: The general process for conducting correlation analysis

To conduct a bivariate correlation you need to find the *Correlate* option of the *Analyze* menu. The main dialog box is accessed by selecting Analyze Correlate ▸ Bivariate... and is shown in Figure 3. Using the dialog box it is possible to select which of three correlation statistics you wish to perform. The default setting is Pearson's product-moment correlation, but you can also calculate Spearman's correlation and Kendall's correlation—we will see the differences between these correlation coefficients in due course.

Having accessed the main dialog box, you should find that the variables in the data editor are listed on the left-hand side of the dialog box. There is an empty box labelled *Variables* on the right-hand side. You can select any variables from the list using the mouse and transfer them to the *Variables* box by dragging them there or clicking on ➡. SPSS will create a table of correlation coefficients for all of the combinations of variables. This table is called a correlation matrix. For our current example, select the variables **Exam performance**, **Exam anxiety** and **Time spent revising** and transfer them to the *Variables* box by clicking on ➡. Having selected the variables of interest you can choose between three correlation coefficients: Pearson's product-moment correlation coefficient (☑ Pearson), Spearman's rho (☑ Spearman) and Kendall's tau (☑ Kendall's tau-b). Any of these can be selected by clicking on the appropriate tick-box with a mouse.

In addition, it is possible to specify whether or not the test is one- or two-tailed. Therefore, if you have a directional hypothesis (e.g., 'the more anxious someone is about an exam, the worse their mark will be') you could click on ◉ One-tailed, whereas if you have a non-directional hypothesis (i.e., 'I'm not sure whether exam anxiety will improve or reduce exam marks') you could click on ◉ Two-tailed. In my book I advise against one-tailed tests so I would leave the default of ◉ Two-tailed.



Bivariate Correlations

Participant Code [C...
Gender [Gender]

Variables:
Time Spent Revisin...
Exam Performance ...
Exam Anxiety [Anxiety]

Correlation Coefficients
☑ Pearson □ Kendall's tau-b □ Spearman

Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations

Bivariate Correlations: Options

Statistics
□ Means and standard deviations
□ Cross-product deviations and covariances

Missing Values
● Exclude cases pairwise
○ Exclude cases listwise

Bootstrap

☑ Perform bootstrapping
Number of samples: 1000
□ Set seed for Mersenne Twister
Seed: 2000000

Confidence Intervals
Level(%): 95
○ Percentile
● Bias corrected accelerated (BCa)

Sampling
● Simple
○ Stratified

Variables:
Participant Code [C...
Time Spent Revisin...
Exam Performance ...
Exam Anxiety [Anxiety]
Gender [Gender]

Strata Variables:

Figure 3: Dialog box for conducting a bivariate correlation

If you click on Options... then another dialog box appears with two *Statistics* options and two options for missing values. The *Statistics* options are enabled only when Pearson's correlation is selected; if Pearson's correlation is not selected then these options are disabled (they appear in a light grey rather than black and you can't activate them). This deactivation occurs because these two options are meaningful only for interval data and the Pearson correlation is used with those kinds of data. If you select the tick-box labelled *Means and standard deviations* then SPSS will produce the mean and standard deviation of all of the variables selected for analysis. If you activate the tick-box labelled *Cross-product deviations and covariances* then SPSS will give you the values of these statistics for each of the variables in the analysis.

Finally, we can get bootstrapped confidence intervals for the correlation coefficient by clicking Bootstrap.... You select ☑ Perform bootstrapping to activate bootstrapping for the correlation coefficient, and to get a 95% confidence interval click ● Percentile or ● Bias corrected accelerated (BCa). For this analysis, let's ask for a bias corrected (BCa) confidence interval.

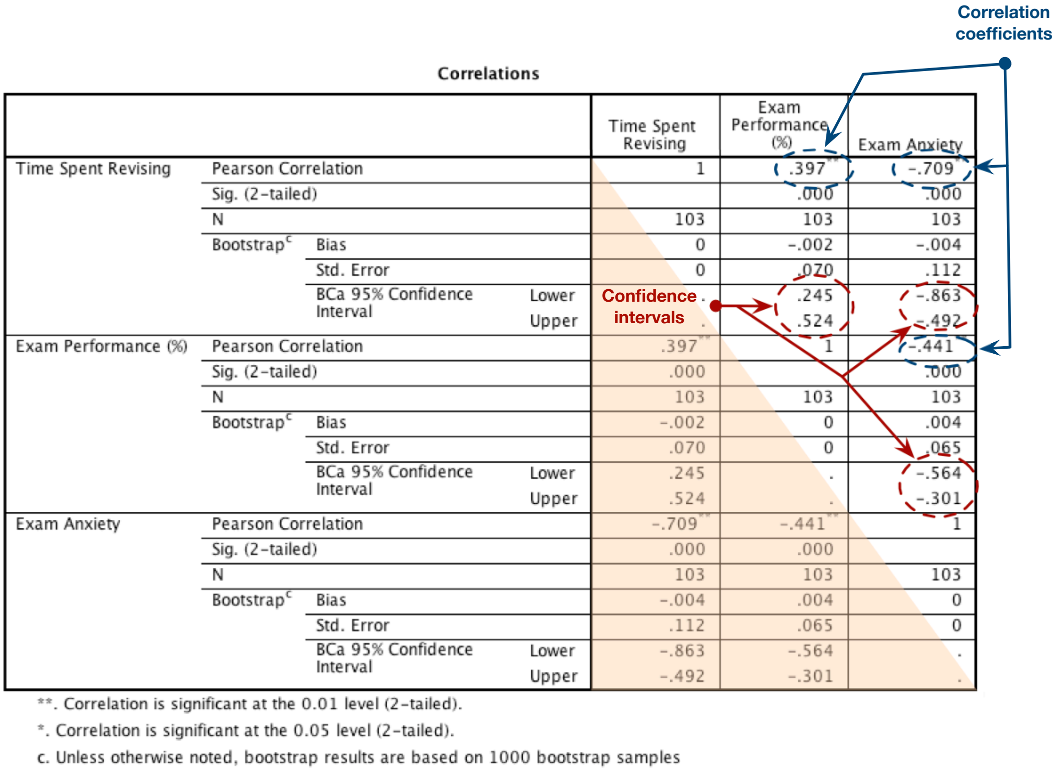
Pearson's correlation coefficient

Running Pearson's r on SPSS

We have already seen how to access the main dialog box and select the variables for analysis earlier in this section (Figure 3). To obtain Pearson's correlation coefficient simply select the appropriate box (☑ Pearson)—SPSS selects this option by default. Click on OK to run the analysis.

Output 1 provides a matrix of results, which looks bewildering, but it's not as bad as it looks. For one thing the information in the top part of the table (not shaded) is the same as in the bottom half (which I have shaded): so we can effectively ignore half of the table. The first row tells us about time spent revising. This row is subdivided so first we are told the correlation coefficients with the other variables: r = .397 with exam performance, and r = –.709 with exam anxiety. The second major row in the table tells us about exam performance, and from this part of the table we can get the correlation coefficient for its relationship with exam anxiety, r = –.441. Directly underneath each correlation coefficient we're told the significance value of the correlation and the sample size (N) on which it is based. The significance values are all less than .001 (as indicated by the double asterisk after the coefficient). This significance value tells us that the probability of getting a correlation coefficient this big in a sample of 103 people if the null hypothesis were true (there was no relationship between these variables) is very low (close to zero in fact). All of the significance values are below the standard criterion of .05 indicating a 'statistically significant' relationship.

Given the lack of normality in some of the variables, we should be more concerned with the bootstrapped confidence intervals than the significance per se: this is because the bootstrap confidence intervals will be unaffected by the distribution of scores, but the significance value might be. These confidence intervals are labelled *BCa 95% Confidence Interval* and you're given two values: the upper boundary and the lower boundary. For the relationship between revision time and exam performance the interval is .245 to .524, for revision time and exam anxiety it is –.863 to –.492, and for exam anxiety and exam performance it is –.564 to –.301. There are two important points here. First, because the confidence intervals are derived empirically using a random sampling procedure (i.e., bootstrapping) the results will be slightly different each time you run the analysis. Therefore, the confidence intervals you get, won't be the same as the ones in utput 1 and that's normal and nothing to worry about. Second, think about what a correlation of zero represents: it is no effect whatsoever. A confidence interval is the boundary between which the population value falls (in 95% of samples), therefore, if this interval crosses zero it means that the population value could be zero (i.e., no effect at all). If it crosses zero it also means that the population value could be a negative number (i.e., a negative relationship) or a positive one (i.e., a positive relationship); in other words, we can't be sure if the true relationship goes in one direction or the complete opposite. For our three correlation coefficients none of them cross zero therefore we can be confident that there is a genuine effect in the population. In psychological terms, this all means that as anxiety about an exam increases, the percentage mark obtained in that exam decreases. Conversely, as the amount of time revising increases, the percentage obtained in the exam increases. Finally, as revision time increases, the student's anxiety about the exam decreases. So there is a complex interrelationship between the three variables



Correlation coefficients
Confidence intervals

Correlations

				Time Spent Revising	Exam Performance (%)	Exam Anxiety
Time Spent Revising	Pearson Correlation			1	.397**	-.709**
	Sig. (2-tailed)				.000	.000
	N			103	103	103
	Bootstrap^{c}	Bias		0	-.002	-.004
		Std. Error		0	.070	.112
		BCa 95% Confidence Interval	Lower	.	.245	-.863
			Upper	.	.524	-.492
Exam Performance (%)	Pearson Correlation			.397**	1	-.441**
	Sig. (2-tailed)			.000		.000
	N			103	103	103
	Bootstrap^{c}	Bias		-.002	0	.004
		Std. Error		.070	0	.065
		BCa 95% Confidence Interval	Lower	.245	.	-.564
			Upper	.524	.	-.301
Exam Anxiety	Pearson Correlation			-.709**	-.441**	1
	Sig. (2-tailed)			.000	.000	
	N			103	103	103
	Bootstrap^{c}	Bias		-.004	.004	0
		Std. Error		.112	.065	0
		BCa 95% Confidence Interval	Lower	-.863	-.564	.
			Upper	-.492	-.301	.

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

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

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

Output 1: Output for a Pearson's correlation

Spearman's Correlation Coefficient

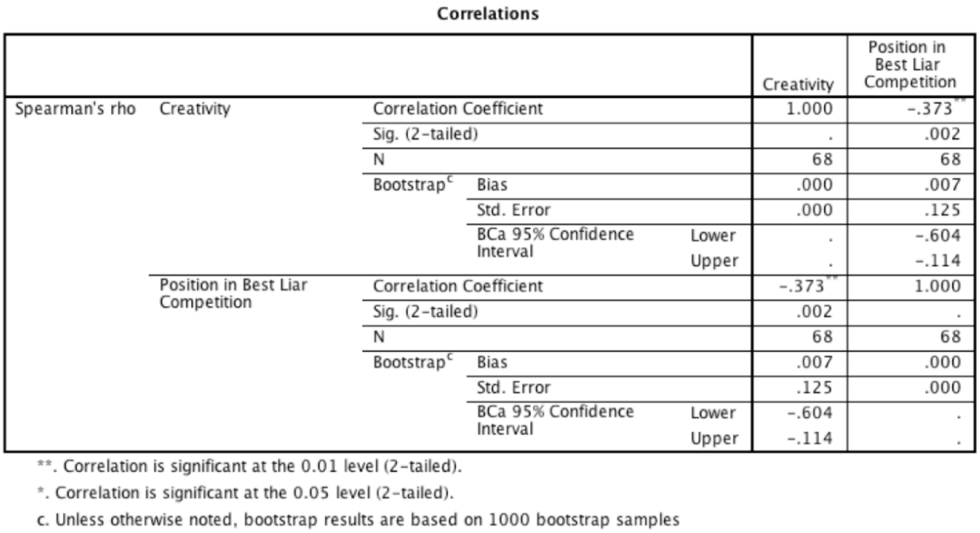
Spearman's correlation coefficient r_s is a non-parametric statistic based on ranked data and so can be useful to minimise the effects of extreme scores or the effects of violations of the assumptions discussed in. Spearman's test works by first ranking the data and then applying Pearson's equation to those ranks.

I was born in England, which has some bizarre traditions. One such oddity is The World's Biggest Liar Competition held annually at the Santon Bridge Inn in Wasdale (in the Lake District). The contest honours a local publican, 'Auld Will Ritson' who in the nineteenth century was famous in the area for his far-fetched stories (one such tale being that Wasdale turnips were big enough to be hollowed out and used as garden sheds). Each year locals are encouraged to attempt to tell the biggest lie in the world (lawyers and politicians are apparently banned from the competition). Over the years there have been tales of mermaid farms, giant moles, and farting sheep blowing holes in the ozone layer. (I am thinking of entering next year and reading out some sections of this book.)

Imagine I wanted to test a theory that more creative people will be able to create taller tales. I gathered together 68 past contestants from this competition and noted where they were placed in the competition (first, second, third, etc.) and also gave them a creativity questionnaire (maximum score 60). The position in the competition is an ordinal variable because the places are categories but have a meaningful order (first place is better than second place and so on). Therefore, Spearman's correlation coefficient should be used (Pearson's r requires interval or ratio data). The data for this study are in the file **The Biggest Liar.sav**. The data are in two columns: one labelled **Creativity** and one labelled **Position** (there's actually a third variable in there but we will ignore it for the time being). For the **Position** variable, each of the categories described above has been coded with a numerical value. First place has been coded with the value 1, with positions being labelled 2, 3 and so on. Note that for each numeric code I have provided a value label (just like we did for coding variables). I have also set the *Measure* property of this variable to Ordinal.

The procedure for doing a Spearman correlation is the same as for a Pearson correlation except that in the *Bivariate Correlations* dialog box (Figure 3), we need to select ☑ Spearman and deselect the option for a Pearson correlation. As with the Pearson correlation we should use the Bootstrap... option to get some robust confidence intervals.

Output 2 shows the output for a Spearman correlation on the variables **Creativity** and **Position**. The output is very similar to that of the Pearson correlation: a matrix is displayed giving the correlation coefficient between the two variables (–.373), underneath is the significance value of this coefficient (.002) and finally the sample size (68). We also have the BCa 95% confidence interval that ranges from –.604 to –.114. The fact that the confidence interval does not cross zero (and the significance is less than .05) tells us that there is a significant negative relationship between creativity scores and how well someone did in the World's Biggest Liar Competition: as creativity increased, position decreased. This might seem contrary to what we predicted until you remember that a low number means that you did well in the competition (a low number such as 1 means you came first, and a high number like 4 means you came fourth). Therefore, our hypothesis is supported: as creativity increased, so did success in the competition.



Correlations

					Creativity	Position in Best Liar Competition
Spearman's rho	Creativity	Correlation Coefficient			1.000	-.373**
		Sig. (2-tailed)			.	.002
		N			68	68
		Bootstrap[c]	Bias		.000	.007
			Std. Error		.000	.125
			BCa 95% Confidence Interval	Lower	.	-.604
				Upper	.	-.114
	Position in Best Liar Competition	Correlation Coefficient			-.373**	1.000
		Sig. (2-tailed)			.002	.
		N			68	68
		Bootstrap[c]	Bias		.007	.000
			Std. Error		.125	.000
			BCa 95% Confidence Interval	Lower	-.604	.
				Upper	-.114	.

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

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

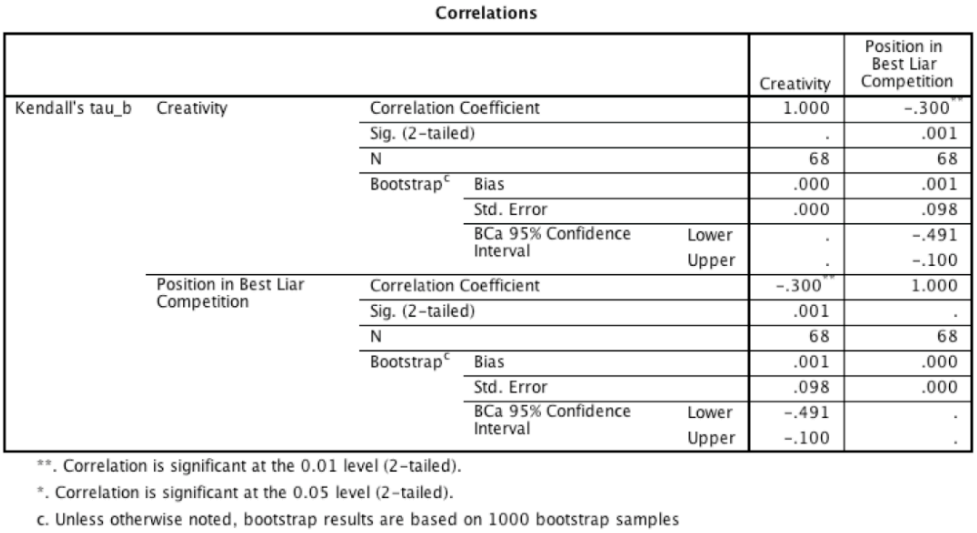
c. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 2

Kendall's tau (non-parametric)

Kendall's tau, τ, is another non-parametric correlation and it should be used rather than Spearman's coefficient when you have a small data set with a large number of tied ranks. This means that if you rank all of the scores and many scores have the same rank, then Kendall's tau should be used. Although Spearman's statistic is the more popular of the two coefficients, there is much to suggest that Kendall's statistic is a better estimate of the correlation in the population. To carry out Kendall's correlation on the world's biggest liar data simply follow the same steps as for Pearson and Spearman correlations but select ☑ Kendall's tau-b and deselect the Pearson and Spearman options. The output is much the same as for Spearman's correlation.

You'll notice from Output 3 that the actual value of the correlation coefficient is closer to zero than the Spearman correlation (it has increased from –.373 to –.300). Despite the difference in the correlation coefficients we can still interpret this result as being a highly significant relationship because the significance value of .001 is less than .05 and the robust confidence interval does not cross zero (–.491 to –.100). However, Kendall's value is a more accurate gauge of what the correlation in the population would be. As with the Spearman correlation, we cannot assume that creativity caused success in the World's Best Liar Competition.



Correlations

					Creativity	Position in Best Liar Competition
Kendall's tau_b	Creativity	Correlation Coefficient			1.000	-.300**
		Sig. (2-tailed)			.	.001
		N			68	68
		Bootstrap[c]	Bias		.000	.001
			Std. Error		.000	.098
			BCa 95% Confidence Interval	Lower	.	-.491
				Upper	.	-.100
	Position in Best Liar Competition	Correlation Coefficient			-.300**	1.000
		Sig. (2-tailed)			.001	.
		N			68	68
		Bootstrap[c]	Bias		.001	.000
			Std. Error		.098	.000
			BCa 95% Confidence Interval	Lower	-.491	.
				Upper	-.100	.

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

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

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

Output 3

References

Field, A. P. (2013). Discovering statistics using IBM SPSS Statistics: And sex and drugs and rock 'n' roll (4th ed.). London: Sage.

Terms of Use