Eight things you need to know about interpreting correlations:

A correlation coefficient is a single number that represents the degree of association between two sets of measurements. It ranges from +1 (perfect positive correlation) through 0 (no correlation at all) to -1 (perfect negative correlation). Correlations are easy to calculate, but their interpretation is fraught with difficulties because the apparent size of the correlation can be affected by so many different things. The following are some of the issues that you need to take into account when interpreting the results of a correlation test.

1. Correlation does not imply causality:

This is the single most important thing to remember about correlations. If there is a strong correlation between two variables, it's easy to jump to the conclusion that one of the variables causes the change in the other. However this is not a valid conclusion. If you have two variables, X and Y, it might be that X causes Y; that Y causes X; or that a third factor, Z (or even a whole set of other factors) gives rise to the changes in both X and Y.

For example, suppose there is a correlation between how many slices of pizza I eat (variable X), and how happy I am (variable Y). It might be that X causes Y - so that the more pizza slices I eat, the happier I become. But it might equally well be that Y causes X - the happier I am, the more pizza I eat. Or it might be that a third factor (price of pizzas) affects both X and Y, so that changes in the price of pizza affect both how much pizza I eat, *and* how happy I am . (When pizza is expensive, I eat little of it and I am miserable).

Sometimes the direction of causality is fairly obvious. For example, the more miles I drive, the greater my CO^2 emissions: it makes no sense to think that my CO^2 emissions cause my mileage. But in many cases, the causal relationship is ambiguous. That's why we do experiments, because a well-designed experiment is able to establish cause and effect relationships.

2. The size of a correlation can be influenced by the size of your sample:

Correlations are usually worked out on the basis of samples from populations. If I am interested in whether there is a correlation between how many times a week someone reads the "Daily Mail" and their score on a Fascism scale, I can't manage to get to the whole population of "Daily Mail" readers to find out. There are (sadly) too many of them. Therefore I would have to be content with taking a random sample of "Daily Mail" readers; asking each one how many times a week they read it and obtaining their Fascism score; and then calculating the correlation between these two samples of measurements. We generally want to infer what the true population correlation is, on the basis of this correlation calculated from our sample. How well does a sample correlation reflect the true state of affairs in the parent population? It depends on the size of your sample. All other things being equal, the larger the sample, the more stable (reliable) the obtained correlation.

Correlations obtained with small samples are quite unreliable. This can be shown as follows. Suppose we take two variables that we *know* are not correlated at all with each other in the parent population (Pearson's $r = 0$). Now suppose we take two samples of measurements from these variables and perform a correlation test on them. What happens? Most of the time, the *r* derived from the samples will be similar to the true value of *r* in the population: our correlation test will produce a value of *r* that is 0, or close to 0. Now suppose we repeat this procedure many times over. Because samples vary randomly, from time to time we will get a sample correlation coefficient that is much larger or smaller than the true population figure. In other words, on occasion we will get freakily large values of r that have really occurred by chance but which might fool us into thinking that there was a strong correlation in the parent population. The smaller the sample size, the greater the likelihood of obtaining a spuriously-large correlation coefficient in this way.

Look at the following table. It shows the limits within which 80% of Pearson's *r* values are likely to fall, if you performed many separate correlation tests between samples from a population in which there was really no correlation at all between the two variables concerned. For example, suppose you performed lots of correlations with a sample size of 5 measurements for each variable. Remember that the true relationship, in the population, is that *r* = 0. However, approximately 80% of the time, the sample correlation will be somewhere between -.69 to .69. This in turn means that 22% of correlations will be even more extreme values: 11% of the time, you are likely to get a correlation bigger than .69, and 11% of the time you are likely to get a correlation bigger than -.69 - purely by chance!

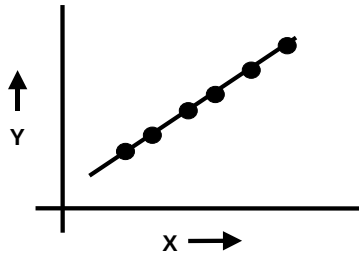
As sample size goes up, so correlation coefficients fluctuate less around the "true" figure for the population *r.* Notice that it's only once you get up to a sample size of 100 or so that the sample correlation coefficients start to consistently resemble the true population correlation coefficient. In short, the message is - be very wary of correlations based on small sample sizes. You need a large sample before you can be really sure that your sample *r* is an accurate reflection of the population *r*.

Limits within which 80% of sample r's will fall, when the true (population) correlation is 0:

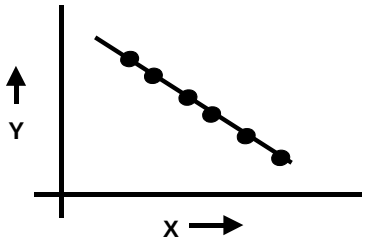
Sample size:	**80% limits for *r*:**
5	**-.69 to +.69**
15	**-.35 to +.35**
25	**-.26 to +.26**
50	**-.18 to +.18**
100	**-.13 to +.13**
200	**-.09 to +.09**

3. Linearity of the relationship:

It's important to keep in mind what a correlation test actually does. Pearson's *r* measures the strength of the *linear* relationship between two variables: for a given increase (or decrease) in variable X, is there a constant corresponding increase (or decrease) in variable Y? If there is, then the values on a scatterplot of the relationship between X and Y will fall on (or close to) a straight line. You will get something like this:

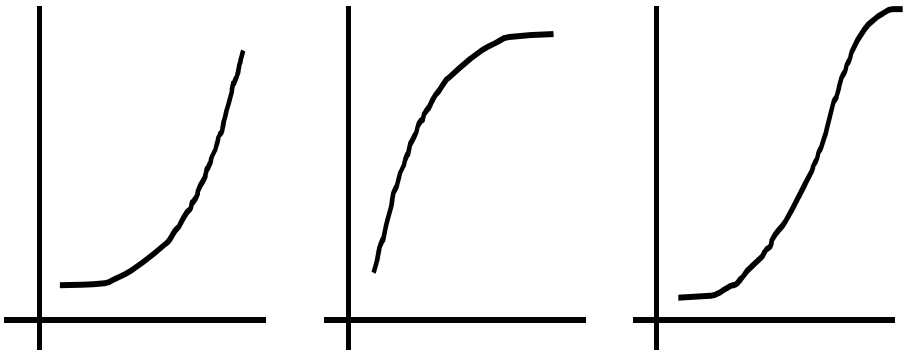


This graph shows a perfect positive correlation between X and Y (*r* = +1): as X increases, so too does Y, by a constant amount.



This graph shows a perfect negative correlation between X and Y (*r* = -1): as X increases, Y decreases by a constant amount.

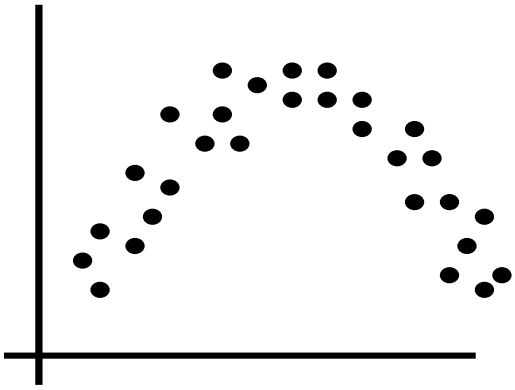
Spearman's *rho* measures something different: it is an indicator of **monotonicity**. It reflects the extent to which an increase (or decrease) in X is associated with *some kind* of increase (or decrease) in Y, but the amount of increase does not have to constant over the whole range of values.



Examples of monotonic relationships between two variables: in all three cases, X increases as Y increases, but the rate of increase in Y differs depending on the value of X (e.g. in the first graph, Y doesn't increase by much when values of X are small; however Y increases a great deal when values of X are large).

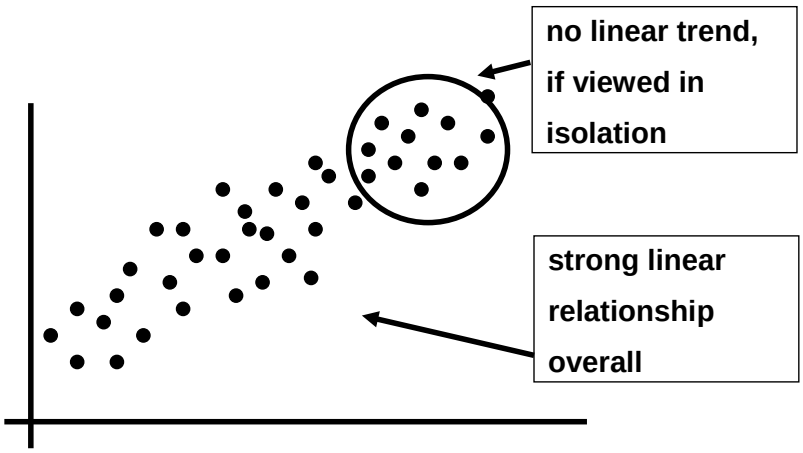
If the data show a monotonic but non-linear relationship, as in the figures above, Pearson's *r* will underestimate the true strength of the relationship. In fact, in this situation, Spearman's *rho* will be larger than Pearson's *r,* because these data show a monotonic relationship and that's what *rho* tests for. This is why you should always look at the scatterplot of the data before calculating a correlation coefficient. The scatterplot will show you what the relationship between X and Y really looks like. In the following example there is a really clear relationship between X and Y. However it is neither linear

nor monotonic: it's curvilinear and so both *rho* and *r* will be very small. If you only looked at the correlation test result, you might be misled into thinking that there was no relationship between X and Y in this situation.



3. Range of talent (variability):

The smaller the amount of variability in X and/or Y, the lower the apparent correlation.

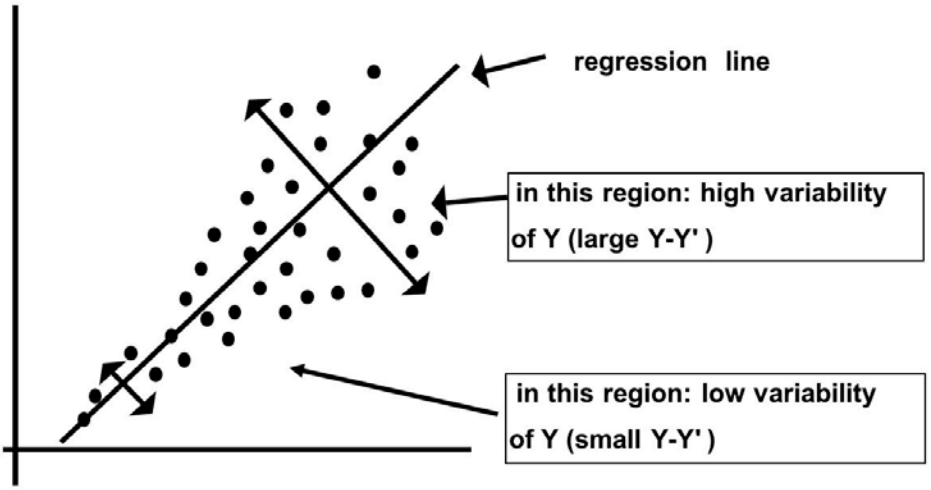


This phenomenon often crops up in correlational research. For example, if you look at the relationship between IQ and educational attainment in the general population, you'll find that the correlation appears to be stronger than it is within a subset of that population (e.g. university students). That's partly because the amount of variability in both IQ and educational attainment is greater in the general population than in amongst students. Therefore, in assessing a correlation, you often need to think about the range of scores on the two variables being correlated. If you have two tests, and one or other of them produces consistently high or consistently low scores, then this will reduce the apparent size of the correlation between the two sets of scores.

4. Homoscedasticity (equal variability):

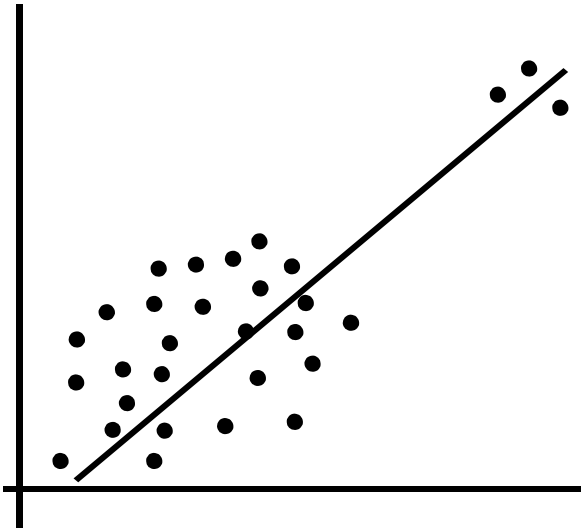
Pearson's *r* describes the *average* strength of the relationship between X and Y. Hence scores should have a constant amount of variability at all points in their distribution. In the following graph, the data lack homoscedasticity. For small values of X and Y, there is a strong relationship: the points are all very close to the regression line. However for large values of X and Y, the relationship is much weaker. This makes *r* pretty meaningless as a summary of the relationship between the two

variables: for small values of X and Y, *r underestimates* the strength of the relationship, whereas for large values of X and Y, *r overestimates* it.



5. Effects of discontinuous distributions:

A few outliers can distort things considerably. if you consider the bulk of the scores here, at the bottom left of the graph, there is clearly no correlation between X and Y. If the outliers at the top right of the figure are included, you increase the size of the correlation coefficient, making it seem as if there is a relationship between X and Y. Even if these were not outliers, it would still be misleading to cite the value of *r* in this case. That's because it's implicit in the use of *r* that the relationship applies across the whole range of values for X and Y. Here, we have no information at all about what happens when X and Y have intermediate values: we only know what happens when the values of X and Y are either low or high. Since r is an estimate of the average strength of the relationship, it would be misleading to use it in this case.



6. Deciding what is a "good" correlation:

We saw earlier that an apparently strong correlation could occur purely by chance, especially if our sample is quite small. Any observed correlation can occur for one of two reasons:

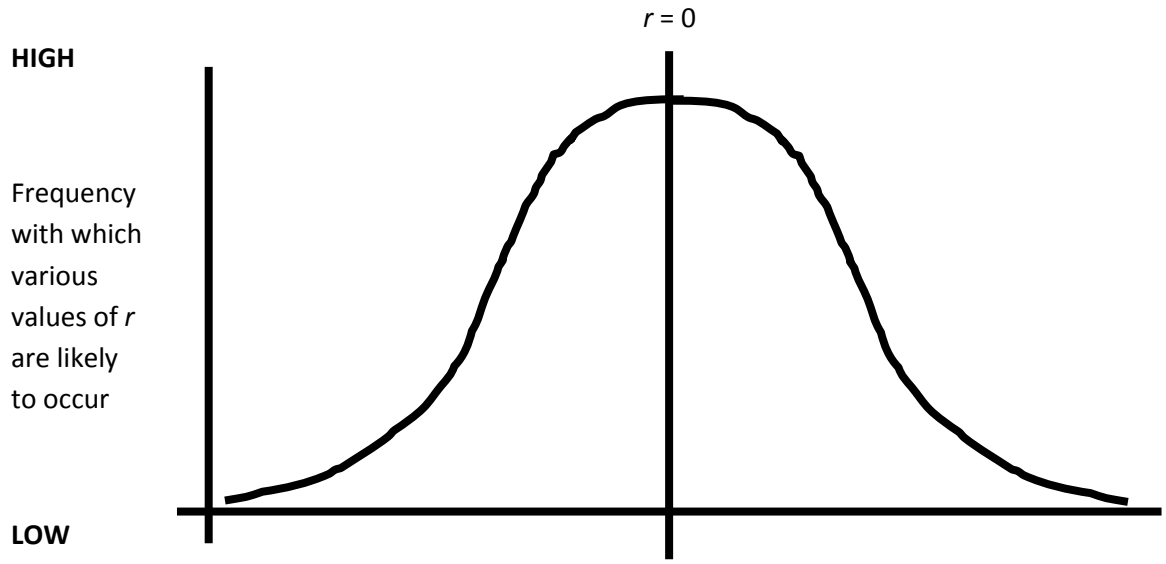
(a) because of sampling variation (so that it's just a fluke); or

(b) because there is a genuine association between the two variables concerned.

How can we tell which of these possibilities is the correct interpretation for our observed value of *r* or *rho*?

We rely on the fact that although sampling variation *can* throw up large correlation coefficients by chance, in practice it is unlikely to do so. The bigger the value of a correlation coefficient, the less likely it is to have occurred merely by chance, and the more likely it is that it has occurred because it represents a genuine relationship between the two variables in question. In fact, the probabilities of obtaining various sizes of correlation between two actually-uncorrelated variables are normally distributed:

Distribution of *r*'s obtained using samples drawn from two *uncorrelated* populations of scores:



large negative values of *r*, highly unlikely to occur by chance when the population *r* is really 0.	small values of *r*, (accurately reflecting the true state of affairs in the population, where *r = 0*).	large positive values of *r*, highly unlikely to occur by chance when the population *r* is really 0.

So if there is really no relationship between X and Y, but we keep taking samples of X and Y and work out the correlation between them each time, this is what we are likely to get. If the sample size is reasonably large, then most of the time we will obtain a small correlation between the two variables (representing the true state of affairs in the population). Spuriously large positive or negative correlations will occur, but not very often - and the larger the correlation, the less likely it is to occur by chance.

In practice, we make a somewhat arbitrary decision:

(a) If our sample correlation is so large that it is likely to occur by chance only 5 times in a hundred tries (it has a $p < .05$), we will *assume* that it reflects a genuine correlation in the population from which the sample came.

(b) If a correlation like ours is likely to occur by chance more often than this, we *assume* it has arisen merely by chance, and that it is not evidence for a correlation in the parent population.

To make this judgement, the size of our sample needs to be taken into account. As mentioned earlier, large correlation coefficients are more likely to occur purely by chance if the sample size is small. Therefore, we need to be more cautious in deciding that our observed value of the correlation coefficient represents a "real" relationship when the sample size is small than when it is large. For example, a correlation of .20 could quite easily occur by chance with a sample size of 5, but is much less likely to occur by chance with a sample size of 200.

SPSS will take sample size into account when telling you whether or not a given correlation is statistically significant (i.e. whether or not it is likely to have occurred by chance). For Spearman's *rho*, the convention is to report N, the number of pairs of scores, as well as the value of *rho* itself. For Pearson's *r*, you report the "degrees of freedom", which is the number of pairs of scores minus 2 (so if you had 20 pairs of scores, you would have 18 d.f.; if you had 11 pairs of scores, you would have 9 d.f.; and so on). This enables the reader to get an idea of the sample size on which your correlation test was based.

The more old-fashioned way of assessing the strength of a correlation is to calculate it by hand, and then compare your obtained correlation coefficient to those in a table of "critical values". There are tables of this sort in the back of many statistics books and on my website.

There are different tables for *r* and for *rho.* For *r*, you need to know the d.f. in order to use the table, while for *rho* you need to know N. Essentially, in both cases, the table shows you various possible values of the correlation coefficient and how likely these are to have occurred by chance. You compare your obtained correlation to the ones in the table. If your obtained value is bigger than the value in the table, then by implication your obtained correlation is even *le*ss likely to have occurred by chance.

An illustration of how to use these tables with Pearson's r:

Suppose we take a sample of 20 people and measure their eye-separation and back hairiness. Our sample *r* is .75. Does this reflect a true correlation between eye-separation and hairiness in the parent population, or has our *r* arisen merely by chance (i.e. because we have a freaky sample)?

Step 1:

Calculate the "degrees of freedom" (d.f. = the number of pairs of scores, minus 2). Here, we have 20 pairs of scores, so d.f. = 18.

Step 2:

Find a table of "critical values for Pearson's *r*". Here's part of such a table.

	Level of significance (two-tailed)		
d.f.	**.05**	**.01**	**.001**
17	**.4555**	**.5751**	**.6932**
18	**.4438**	**.5614**	**.6787**
19	**.4329**	**.5487**	**.6652**
20	**.4227**	**.5368**	**.6524**

Step 3:

Move along the row of values for 18 d.f., from left to right, comparing our obtained *r* with the ones in the table. With 18 d.f., a correlation of .4438 or larger will occur *by chance* with a probability of .05: i.e., if we took 100 samples of 20 people, about 5 of those samples are likely to produce an *r* of .4438 or larger (even though there is actually no correlation in the population). Our sample *r* of .75 is larger than .4438, and so even *less* likely to occur by chance than this.

With 18 d.f., a correlation of .5614 or larger will occur *by chance* with a probability of .01: i.e., if we took 100 samples of 20 people, about 1 of those 100 samples is likely to give an *r* of .5614 or larger (again, even though there is actually no correlation in the population). Our *r* of .75 is larger than .5614, and so even *less* likely to occur by chance than this.

With 18 d.f., a correlation of .6787 or larger will occur *by chance* with a probability of .001: i.e., if we took *1000* samples of 20 people, about 1 of those 1000 samples is likely to give an *r* of .6787 or larger (again, even though there is actually no correlation in the population). Our *r* of .75 is larger than .6787, and so even *less* likely to occur by chance than this.

We can therefore consider the two possible interpretations of our obtained value of r. The first (the "null hypothesis") is that there is really no relationship between eye-separation and back-hairiness in the population: our obtained value of r has arisen merely by chance, as a sampling "fluke". This interpretation might be true, but the probability of obtaining a value of *r* as large as ours, purely by chance, is less than one in a thousand ($p < .001$). It seems more plausible to accept the "alternative hypothesis", that the value of *r* that we have obtained from our sample represents a real relationship between eye-separation and back-hairiness in the population. You always have to bear in mind that any sample correlation could *in principle* occur due to chance or because it reflects a true relationship in the population from which the sample was taken. We can never be certain which of these two interpretations is the correct one, because freaky results *can* occur merely by chance. However, in this case, because our *r* of .75 is *so* unlikely to occur by chance, we can safely assume that there almost certainly is a genuine relationship between eye-separation and back-hairiness.

7. How to report the results of a correlation test:

For the above example, you might write something as follows.

"There was a strong positive correlation between eye-separation and back-hairiness, $r(18) = .75$, $p < .001$".

You report Spearman's *rho* similarly, except that you report N instead of d.f. :

"There was a negative correlation between Daily Mail reading and intelligence, $r_s(98) = .21$, $p < .05$".

8. Statistical significance versus importance:

Do not confuse statistical significance with practical importance. They are quite different issues. We have just assessed "statistical significance" - the likelihood that our obtained correlation has arisen merely by chance. Our *r* of .75 is "highly significant" (i.e., highly unlikely to have arisen by chance). However, a weak correlation can be statistically significant, if the sample size is large enough. Suppose that actually we tested 102 people, and that the relationship between eye-separation and back-hairiness gave rise to an *r* of .19. With 100 d.f., this *r* would be statistically "significant" in the sense that it is unlikely to have arisen by chance (*r*'s bigger than this will occur by chance only 5 in a 100 times).

In order to assess practical importance, square the correlation coefficient to get R^2, the **coefficient of determination**. This shows how much of the variation in one of the variables is associated with variation in the other. An *r* of .19 produces an R^2 value of only 3.61% (.19 * .19 = .0361, which as a percentage = 3.61%). R^2 shows that this is not a *strong* relationship in a practical sense; knowledge of one of the variables would account for only 3.61% of the variance in the other. This would be completely useless for predictive purposes, as 96.39% of the variation in back-hairiness occurs for reasons that are unconnected with that variable's relationship to eye-separation!

R^2 is an example of a measure of **effect size**. Other statistical tests, such as t-tests and ANOVA, also have their own measures of effect size. It is becoming increasingly common in psychology for researchers to report effect size alongside more conventional measures of statistical significance. This is because researchers want to convey some idea of the size of an effect, as well as the likelihood of it occurring merely by chance.