Post-lecture Notes and then Questions I.3 – Bivariate Data and Correlations

(Note: I've added some post-lecture notes before the questions, because my experience is that some students have some difficulty with the difference and uses of "unreliability" and test/retest reliability. If you're completely at ease with the two concepts, feel free to jump straight to the questions.)

Recall that we now have two ways of assessing the "quality" of a measure. We can use the standard deviation across repeated uses of the measure on the same thing in the same subject – i.e., the "unreliability" of the measure. Or we can use the correlation between pairs of uses of the measure across a group of subjects – i.e., test/retest reliability.

The main advantage of "unreliability" is that it has a direct relationship with most of the inferential statistics that we will use later. There is a relatively simple relationship between the "unreliability" of a measure and whether a difference between two values will be statistical significant. Furthermore, we have a very straight-forward way of reducing "unreliability"; take multiple measures, instead of just one, and keep the mean (a summary score), instead of a single piece of raw data. Using the mean of many redundant measures reduces "unreliability" by the $1/\sqrt{N}$ rule (where N is the number of raw-data measures that were used to create the mean score that you keep).

The largest disadvantage of "unreliability" is that, because it is the standard deviation across multiple measures, it has the same units as the raw data. (E.g., the "unreliability" of response time, which is usually measured in thousandths of a second, also has the units *milliseconds*.) Therefore, there is no way to compare "unreliabilities" across measures. Every measure has its own kind of units, and you can't compare two things with different units – apples to oranges.

Using reliability, instead, avoid this problem. Because reliability is a correlation, it has no units and is "trapped" between –1.00 and +1.00, regardless of what two things are being examined. This allows us to have a single, general rule with regard to what is an acceptable, minimum level of reliability for a psychological measure: +.70. It also allows us to directly compare any two measures, so, when you have more than one option available, you can make a simple decisions as to which to use.

Another advantage of using reliability comes from the fact that correlations are completely unaffected by linear transformations of the data (i.e., changes to the raw data caused by adding, subtracting, multiplying, or dividing the raw data by any particular number). Thus, if the reliability of measuring response time in milliseconds is +.85, then the reliability of measuring response time in seconds is also +.85, even though you just divided all of the raw data by 1000.

A third advantage of using reliability comes from the type of data that are needed to calculate it. You need to sample a large number of random people (measuring each of them twice). Given that our experiments usually involve large samples of random people, there's a nice match between the type of data used to establish reliability and the type of data used when the measure is actually employed.

The down-sides of reliability also come from the same, general properties of correlations. For example, imagine that every time that a particular measure is used an additional time (on a given subject), the value you get is 10% higher than last time it was used. This will have no effect at all on reliability, since an increase of 10% is the same thing as multiplying one set of the values by 1.10 and correlations are unaffected by multiplying all of the scores (of either or both variables) by a fixed number. (Note that having the scores go up by 10% every time will definitely affect “unreliability” since it will make the scores more different from each other, so if this kind of “drift” in the data is an issue you want to keep track of, then maybe “unreliability” is the better measure to use.)

With that said, we still use reliability as the preferred way to quantify the quality of a measure. Additional reasons for focusing on a measure that comes from a correlation will come up next week.

From what kinds of paired data can you calculate a correlation coefficient?

What information does a correlation coefficient provide?

What possible values do correlations take and what do they mean?

How can we measure the strength, instead of the size, of the linear relationship between two variables?

When applying a correlation that was collected using one set of data to a different set of data, what must you keep in mind? [This is a hard question, but try to answer it before looking at the next page.]

Assume that the correlation between depression and anxiety is +.50. From this can you figure out what the correlation between anxiety and depression is and, if so, what is it?

What is “unreliability” and how do we quantify it?

What is the definition of *reliability*? What is the requirement for this in psychology?

What is the biggest reason that we use reliability – instead of something like “unreliability” – as our way of quantifying the quality of our measures (in terms of their statistics)?

Some example multiple-choice questions for this week:

1. Which of the following is **not** part of a complete set of descriptive statistics (as used in psychology)?
   (A) A numerical measure of the center
   (B) A numerical measure of the spread
   (C) A numerical measure of the reliability
   (D) A name for the shape

2. The correlations between two variables (X & Y) _____ .
   (A) has the same units as the first (X) variable
   (B) has the same units as the second (Y) variable
   (C) has the units of both variables (X & Y) combined
   (D) doesn’t have any units

You can calculate something like a correlation coefficient using pairs of any two variables. However, the standard version of the correlation coefficient only works when both variables are numerical.

Correlation coefficients are a measure of the linear – i.e., straight-line – relationship between the variables. They do not measure other forms of association, such as curved relationships.

Correlations range from –1.00 to +1.00. A correlation with an absolute value of 1.00 implies a perfect (straight-line) relationship between the two variables. Negative correlations mean: as one variable increases, the other decreases. Positive correlations mean: as one variable goes up, the other does, too. A zero correlation means that the two variables do not show a linear relationship; they could still be related in some way, just not by a straight line.

The strength of the relationship is given by the correlation squared, which is called the *coefficient of determination*.

You must always keep the ranges of values in mind. You cannot generalize a correlation to a different range of values on either variable. The correlation between two variables for one range of values can be completely different – even opposite! – from the correlation between variables for a different range of values. A great example of this is when one of the variables is age of the subject. Many relationships involving age are exactly reversed for very young ranges of people versus very old ranges of people.

Correlations are bidirectional. If the correlation between depression and anxiety is +.50, then the correlation between anxiety and depression is +.50.

The "unreliability" of a measure is the extent to which you don't get the same answer when you measure the same thing repeatedly under the same conditions. The best way to quantify it is in terms of the standard deviation across the measures.

The term *reliability* (when there's no modifier word in front of it) always means test/retest reliability, which is defined as the correlation, across subjects, between two uses of the measure under the same conditions. In a nutshell: you measure a bunch of people twice (under the same conditions, etc) and find the correlation between the pairs of values. In psychology, we demand that our measures have at least +.70 reliability.

We use reliability to assess the statistical quality of a measure because it is based on correlations and correlations have no units, so we can compare the quality of any two measures, regardless of what they involve, directly. A second reason that we like test/retest reliability in particular is that the data are collected from a variety of people and each person is measure only a few times (twice), which match how experiments are run.

The correct answer to the first question is C. We need measures of center and spread, plus a name for the shape, but a measure of the reliability (or "unreliability") of the measure is not part of the descriptive stats. It's important, but not part of a set of D-stats.

The correct answer to the second question is D. This is absolutely critical. So much of the usefulness of correlation coefficients comes from the fact that they don't have units. This is what allows us to directly compare any two correlations, no matter what kinds of data they came from.