New View of Statistics: Measures of Reliability

Summarizing Data:
PRECISION OF MEASUREMENT

How precise are your measurements? An important question, because the lower the precision, the more subjects you'll need in your study to make up for the "noise" in your measurements. Even with a larger sample, noisy data can be hard to interpret. And if you are an applied scientist in the business of testing and assessing clients, you need special care when interpreting results of noisy tests.

The two most important aspects of precision are **reliability** and **validity**. Reliability refers to the reproducibility of a measurement. You quantify reliability simply by taking several measurements on the same subjects. Poor reliability degrades the precision of a single measurement and reduces your ability to track changes in measurements in the clinic or in experimental studies. Validity refers to the agreement between the value of a measurement and its true value. You quantify validity by comparing your measurements with values that are as close to the true values as possible. Poor validity also degrades the precision of a single measurement, and it reduces your ability to characterize relationships between variables in descriptive studies.

The concepts of reliability and validity are related. For example, a little thought will satisfy you that measurements can be reliable but not valid, and that a valid measurement must be reliable. But we usually deal with these two concepts separately, either because most researchers study them separately, or because bringing the two concepts together is mathematically difficult. I've had a shot at combining them, but there's much more work to do.

Here's the route map for this excursion. We begin with measures of reliability, then there are separate pages for applications of reliability and calculations for reliability. We'll deal with measures of validity and calculations for validity on one page, followed by applications of validity. Along the way there are three spreadsheets for various calculations: the precision of a subject's true value using reliability or validity, calculating reliability between pairs of trials, and calculating validity. Then there's a quick and easy page on precision in reporting measurements, and finally a page devoted to the all-important question of mean ± SD vs mean ± SEM. Some of the material on these pages is in Hopkins (2000).

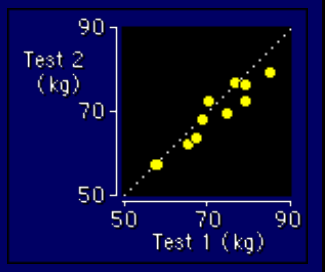
Update Oct 2011: view this slideshow on validity and reliability for an overview of the important principles.

For a **Powerpoint presentation** (slide show) on the essentials of reliability and some of its uses (assessing individuals, estimating sample sizes, estimating individual responses), click here. This presentation was part of a mini-symposium entitled "Reliability, a Crucial Issue for Clinicians and Researchers" at the 2001 annual meeting of the American College of Sports Medicine in Baltimore.

MEASURES OF RELIABILITY

The most common form of reliability is **retest reliability**, which refers to the reproducibility of values of a variable when you measure the same subjects twice or more. Let's get down to the detail of how we quantify it. The data below, and the figure, show an example of high reliability for measurement of weight, for 10 people weighed twice with a gap of two weeks between tests. I'll use this example to explain the three important components of retest reliability: change in the mean, typical error, and retest correlation. I'll finish this page with two other measures of reliability: kappa coefficient and alpha reliability.

Test 1	Test 2
57.5	57.4
65.6	63.2
67.0	66.5
68.5	69.9
70.8	72.8
72.2	70.1
74.9	75.6
76.0	75.2
76.1	72.8
83.1	79.0



Change in the Mean

The dotted line in the figure is the line representing identical weights on retest. Notice that most of the subjects are below the line: they were a bit lighter in the second test. To put a number on the change in weight, you subtract the mean of all the subjects for Test 1 (71.2 kg) from that for Test 2 (70.3 kg). The result (-0.9 kg) is the **change in the mean**: the difference between the means for two tests. The change consists of two components: a **random change** and a **systematic change**.

Random change in the mean is due to so-called **sampling error**. This kind of change arises purely from the typical error, which is like a randomly selected number added to or subtracted from the true value every time you take a measurement. The random change is smaller with larger sample sizes, because the random errors from all the measurements contributing to the mean tend to cancel out more.

Systematic change in the mean is a non-random change in the value between two trials. If the drop in weight in our example is a systematic change, it could be due to changes in the the subjects' behavior between trials. In tests of human performance that depend on effort or motivation, subjects might also perform the second trial better because they want to improve. Performance can be worse in a second trial if fatigue from the first trial is present at the time of the second trial. Performance can also decline in a series of trials, owing to loss of motivation.

Systematic change in the mean is an important issue when subjects perform a series of trials as part of a monitoring program. The subjects are usually monitored to determine the effects of an intervention (e.g., a change in diet or training), so it is important to perform enough trials to make learning effects or other systematic changes negligible before applying the intervention.

Systematic change is less of a worry for researchers performing a controlled study, because only the relative change in means for both groups provides evidence of an effect. Even so, the magnitude of the systematic change is likely to differ between individuals, and these individual differences make the test less reliable by increasing the typical error. You should therefore choose or design tests or equipment with small learning effects, or you should get subjects to perform practice (familiarization) trials to reduce learning effects.

How do you tell whether an observed change in the mean is a reproducible systematic effect? You work out and interpret the confidence limits for the mean, which represent the likely range of the true (systematic) change.

Typical Error of Measurement

Notice that our subjects didn't have exactly the same weight in the first and second tests. Sure, part of the problem is that everyone got a bit lighter, but even when you take the shift in the mean out of the picture, the weights on retest aren't exactly the same. To see what I mean, imagine that you reweighed one subject many times, with two weeks between each weighing. You might get something like:

> *72.2, 70.1, 68.5, 69.9, 67.9, 69.6...*

The first few weights show a slight trend downwards--our subjects decided to lose a bit of weight, remember--then the weights level off, apart from a random variation of about a kilogram. That random variation is the **typical error**. We quantify it as the standard deviation in each subject's measurements between tests, after any shifts in the mean have been taken into account. The official name is the **within-subject standard deviation**, or the **standard error of measurement**. From now on I will refer to it as the **typical error of measurement**, or simply typical error, because its value is indeed the typical error or variation in a subject's value from measurement to measurement.

We talk about variation in measurements as *error,* but it's important to realize that only part of the variation is due to error in the sense of **technological error** arising from the apparatus. In fact, in the above example the variation is due almost entirely to **biological variation** in the weight of the subject. If we were to reweigh the subject with two minutes between weighings rather than two weeks, we'd get pure technological error: the noise in the scales. (We might have to take into account the fact that the subject would be getting slightly lighter all the time, through evaporation or trips to the bathroom.) *Measurement error* is a statistical term that covers variation from whatever source. It would be better to talk about measurement *variation* or typical *variation,* rather than *error,* but I might have trouble convincing my colleagues...

I've explained the notion of typical error as variation for one subject, but in practice you calculate the average typical error for all the subjects. You can calculate it even when there are only two tests, and even when there is a shift in the mean between those tests. See the page on calculations for reliability and the reliability spreadsheet for details. For the weight data shown in the figure, the typical error is 1.4 kg.

You can derive a closely related measure of error simply by calculating each subject's standard deviation, then averaging them. The result is the **total error of measurement**, which is a form of typical error contaminated by change in the mean. On its own the total error is not a good measure of reliability, because you don't know how much of the total error is due to change in the mean and how much is due to typical error. Some researchers and anthropometrists have used this measure, nevertheless.

An important form of the typical error is the **coefficient of variation**: the typical error expressed as a percent of the subject's mean score. For the above data, the coefficient of variation is 2.0%. The coefficient of variation is particularly useful for representing the reliability of athletic events or performance tests. For most events and tests, the coefficient of variation is between 1% and 5%, depending on things like the nature of the event or test, the time between tests, and the experience of the athlete. For example, if the coefficient of variation for a runner performing a 10,000-m time trial is 2.0%, a runner who does the test in 30 minutes has a typical variation from test to test of 0.6 minutes.

If you use the coefficient of variation rather than the raw typical error, it makes sense to represent any changes in the mean between tests as **percent changes**. In our example of body weights, the shift in the mean of -0.9 kg is -1.2%. The percent shifts, and the coefficient of variation, can be derived by analysis of the log-transformed variable. See the page on calculations for reliability for details.

All standard methods for calculating the typical error are based on the assumption that the typical error has the same

average magnitude for every subject. If the typical error varies between subjects, statisticians say the data display heteroscedasticity, or **non-uniform error**. In this situation the analysis provides you with some kind of average typical error that will be too high for some subjects and too low for others. To get rid of heteroscedasticity, you have to either do separate analyses for subgroups of subjects with similar typical errors (e.g., males and females), or find a way to transform the variable to make the typical error for the transformed variable uniform. Log transformation often makes the error uniform when larger values of the original variable have more error. You should check for non-uniform error whenever you calculate reliability statistics. I explain how on the calculations page.

Another form of within-subject variation promoted by some statisticians is **reliability limits of agreement**, which represent the 95% likely range for the difference between a subject's scores in two tests. For example, if the limits of agreement for a measurement of weight are ±2.5 kg, there's a 95% chance that the difference between a subject's scores for two weighings will be within -2.5 kg and +2.5 kg (after any learning effect or other systematic change in the mean on retest has been taken out of the picture). Equivalently, if you reweighed a large number of subjects, 95% of them would have difference scores within -2.5 kg and +2.5 kg. The range defined by the limits of agreement is regarded as a kind of **reference range** for changes between pairs of measurements: in our example, any change between -2.5 and +2.5 kg is deemed to be normal variation; anything else is unusual enough to be indicative that a real change has occurred.

For a normally distributed variable, the limits of agreement are ±2.77 times the typical error. The 2.77 comes from the standard deviation of the difference score (which is root2 times the typical error) multiplied by 1.96 (which includes 95% of observations of the difference score). So even though they are very different in definition, the fact that the typical error and limits of agreement are proportional makes their properties similar. Which is the better measure of reliability? I prefer typical error, because limits of agreement are harder to understand, they are harder to apply to the error of a single measurement, they are too large as a reference range for making a decision about a change in a subject's measurements (more about this issue on the next page), and they have to be converted into a typical error for most statistical calculations.

Retest Correlation

When you plot test and retest values, it's obvious that the closer the values are to a straight line, the higher the reliability. A **retest correlation** is therefore one way to quantify reliability: a correlation of 1.00 represents perfect agreement between tests, whereas 0.00 represents no agreement whatever. In our example the correlation is 0.95, which represents very high reliability.

OK, do we need the correlation coefficient? Why can't we just use the typical error? Hmmm... Well, the two are certainly related, because a small typical error usually means a high correlation. But they also measure different things. The typical error is a pure measure of variation within each subject, whereas the correlation coefficient tells us something about the reproducibility of the rank order of subjects on retest. A high correlation means the subjects will mostly keep their same places between tests, whereas a low correlation means they will be all mixed up. Even a correlation as high as 0.95 implies some loss of order, as you can see in our example in the columns of weights. I've rank-ordered the weights in the first column (Test 1) to show you that the ordering is degraded somewhat in the second column (Test 2). It might help you understand if you think about the possibility of *negative* correlations for reliability. Such things exist and are even worse than zero, because they imply that the rank order of subjects in the first test tends to be *reversed* in the second test.

There is another important difference between typical error and retest correlation. Typical error can be estimated from a sample of subjects that is not particularly representative of the population you want to study. For example, the sample can be homogeneous relative to the population, or you can do multiple retests on just a few subjects. Either way, you can usually assume the resulting typical error applies to any subject in the population. But the retest correlation is sensitive to the nature of the sample used to estimate it. For example, if the sample is homogeneous, the correlation will be low. Or if multiple tests are performed on only a few subjects, the resulting estimate of correlation will be "noisy" (take my word for it). So whenever you interpret a correlation, remember to take into

consideration the sample that was used to calculate it.

How do you calculate the retest correlation? The usual Pearson correlation coefficient is acceptable for two tests, but it overestimates the true correlation for small sample sizes (less than ~15). A better measure of the retest correlation is the **intraclass correlation coefficient** or ICC. It does not have this bias with small samples, and it also has the advantage that it can be calculated as a single correlation when you have more than two tests. In fact, the intraclass correlation is equivalent to the appropriate average of the Pearson correlations between all pairs of tests. You use analysis of variance or repeated measures to do the calculation, as detailed in reliability calculations.

Pearson and intraclass correlations are unaffected by any shift in mean on retest. So, in our example, the fact that the weights are down a bit in the second test has no effect on the correlation coefficient. And that's the way it should be. The question of any change in the mean value on retest should be kept separate.

By the way, I don't know what *intraclass* means. I presume the *intra* refers to the way typical error enters into the calculation of the correlation.

Kappa Coefficient: Reliability of Nominal Variables

Reliability can also be defined for nominal variables, to represent the consistency with which something is classified on several occasions. For example, how consistent are subjects in their choice of favorite sport, or in agreeing or disagreeing with a statement? The best measure is something called the **kappa coefficient**. It is analogous to a correlation coefficient and has the same range of values (-1 to +1). As far as I know, there is nothing analogous to typical error or change in the mean for nominal variables.

Alpha Reliability

Sport psychologists often produce a variable by effectively averaging the scores of two or more items from a multi-item questionnaire or inventory. The **alpha reliability** of the variable is derived by assuming each item represents a retest of a single item. For example, if there are five items, it's as if the five scores are the retest scores for one item. But the reliability is calculated in such a way that it represents the reliability of the *mean* of the items, not the reliability of any single item. So, for example, the alpha reliability of 10 items would be higher than that of 5 similar items.

Alpha reliability should be regarded as a measure of internal consistency of the mean of the items at the time of administration of the questionnaire. It is not test-retest reliability. For that, the questionnaire has to be administered on two or more occasions.

Go to: Next · Previous · Contents · Search · Home

Hopkins WG (2000). Measures of reliability in sports medicine and science. Sports Medicine 30, 1-15 (PDF reprint)

Last updated 4 Oct 2011