Reliability

Andrew Johnson

Introduction

In this demonstration, we will explore some of the basic principles of reliability, including:

1) calculating the standard error of measurement
2) calculating confidence intervals around true scores, taking reliability into account
3) calculating Cronbach's alpha, and establishing a confidence interval around this estimate of reliability
4) using the Spearman-Brown prophecy to estimate the effects of scale length on reliability

There are a number of different ways in which these tasks can be completed, but we will use the `psychometric` library in this demonstration.

```
library(psychometric)
```

```
## Loading required package: multilevel
## Loading required package: nlme
## Loading required package: MASS
```

Calculating the Standard Error of Measurement

The standard error of measurement (SEM) is the expected standard deviation of a test when an individual completes a large number of tests, and is typically assessed using the formula:

$$\sigma\sqrt{1 - r_{xx}}$$

For example, consider a test with a standard deviation of 15, and a reliability of 0.8:

```
SE.Meas(s=15, rxx=.8)
```

```
## [1] 6.708204
```

The SEM is 6.7082039. Or, consider a test with a standard deviation of 5, and a reliability of 0.7:

```
SE.Meas(s=5, rxx=.7)
```

```
## [1] 2.738613
```

The SEM for this test would be calculated to be 2.7386128. Note that the standard deviation in this second example is smaller than the standard deviation in the first example, *even though the reliability estimate is smaller*. This underscores the importance of the standard deviation in estimating the error of the measure.

Using the Standard Error of Measurement to Calculate Confidence Intervals

The SEM is not a particularly exciting calculation on its own, but it can be used in some interesting ways, particularly with regards to the establishment of confidence interval around observed scores. This practice is a nod to an important property of any observed score - namely that the observed score is merely a measurement of the true score that you were interested in assessing, and the extent to which this observation is close to the true score is going to be a function of the variability of individuals on the measure, *as well as the ability of measure to capture the variability of that true score.* In other words, the extent to which we are successful in measuring a true score is going to depend, at least in part, upon the reliability of the measurement instrument. If we wanted to use the SEM in the calculation of a confidence interval around an observed score, we would apply the following formula:

$$x \pm z_{\frac{\alpha}{2}} * SEM$$

Thus, for a 95% confidence interval, we would use the following formula:

$$x \pm 1.96 * SEM$$

Example of a Confidence Interval Calculated Using SEM

To illustrate the utility of these confidence intervals, let's consider a measure with a mean of 100, and a standard deviation of 15. You have tested two individuals on this measure, and obtained the following two scores: 105 and 115. Can we be 95% confident that these two measurements are suggestive of a difference between the individuals, if the measure has a reliability of 0.8?

One method that we might employ, in identifying whether or not there is an expected overlap between these two measurements, is the calculation of a confidence interval. If the 95% confidence intervals of these two measurements overlap, we cannot be 95% confident that the scores are entirely distinct. Let's compute the confidence intervals for each observation.

```
CI.obs(obs=105, s=15, rxx=.8)
```

```
##    SE.Meas      LCL OBS      UCL
## 1 6.708204 91.85216 105 118.1478
```

```
CI.obs(obs=115, s=15, rxx=.8)
```

```
##    SE.Meas      LCL OBS      UCL
## 1 6.708204 101.8522 115 128.1478
```

As we can see, the confidence intervals overlap - the upper limit on the confidence interval for the observed score of 105 is 118, and the lower limit on the confidence interval for the observed score of 115 is 102. Furthermore (as you can see below), given the standard deviation of the measure (15), the observations would need to be 27 points apart on this measure (with this reliability), in order for us to be 95% certain that there is no overlap between the two 95% confidence intervals!

```
CI.obs(obs=105, s=15, rxx=.8)
```

```
##    SE.Meas      LCL OBS      UCL
## 1 6.708204 91.85216 105 118.1478
```

```
CI.obs(obs=132, s=15, rxx=.8)
```

```
##    SE.Meas      LCL OBS      UCL
## 1 6.708204 118.8522 132 145.1478
```

Obviously, this is an absurdly high difference between observations on this measure. Consider that this is the distribution that would be expected on a conventionally normed IQ test - and thus you are (effectively) saying that we can only tell the difference between two individuals that are almost two standard deviations apart on the measure. Cohen (1990) suggested that a confidence interval of 80% is suitable for most situations, and this would suggest a markedly narrower confidence interval for these observations:

```
CI.obs(obs=105, s=15, rxx=.8, level=.80)
```

```
##    SE.Meas      LCL OBS      UCL
## 1 6.708204 96.40309 105 113.5969
```

```
CI.obs(obs=115, s=15, rxx=.8, level=.80)
```

```
##    SE.Meas      LCL OBS      UCL
## 1 6.708204 106.4031 115 123.5969
```

but even this suggests that we would need to see a difference of at least 19 points in order to see no overlap. Consider, however, the effect of markedly improving the measurement of the scales. If we were able to measure the scores with a measure that has a reliability of 0.95, our 80% confidence intervals for these measurements would not overlap:

```
CI.obs(obs=105, s=15, rxx=.95, level=.80)
```

```
##    SE.Meas      LCL OBS      UCL
## 1 3.354102 100.7015 105 109.2985
```

```
CI.obs(obs=115, s=15, rxx=.95, level=.80)
```

```
##    SE.Meas      LCL OBS      UCL
## 1 3.354102 110.7015 115 119.2985
```

This illustrates the importance of reliability in establishing the difference between scores.

Calculating Alpha

Cronbach's alpha may not be the best method for estimating the internal consistency of a measure, but it is certainly the most ubiquitous. We will learn other methods for calculating coefficient alpha when we discuss item analysis, but a quick method can be found in the `psychometric` library.

To illustrate this "quick method" of calculating Cronbach's alpha, we will use the `bfi` dataset found within the `psych` package. It is made up of 25 self-report personality items from the International Personality Item Pool, as well as gender, education level, and age, for 2800 subjects. The personality items are split into 5 categories: Agreeableness (A), Conscientiousness (C), Extraversion (E), Neuroticism (N), and Openness (O). Each item was answered on a six-point scale. We will be re-visiting this dataset for our demonstrations of item analysis, factor analysis, and item response theory.

We only need the `psych` package for the dataset, and so we will detach it immediately after loading the data.

```
library(psych)
data(bfi)
detach("package:psych")
```

The `alpha` function in the `psychometric` package allows us to calculate Cronbach's alpha for each individual scale. To do this, we need to specify (in parentheses) all of the items that we would like to use in our calculation of alpha.

```
alpha(data.frame(bfi[,1:5]))
```

```
## [1] 0.4306169
```

Hmmm... this alpha (0.43) is quite low. Although this may be indicative of a poorly constructed scale, it may also suggest that we have failed to take into account the "keying" of the items within the measure (i.e., whether all of the items are assessing the construct in the same "direction"). For this scale, we need to ensure that all of the items are assessing "agreeableness", rather than "disagreeableness".

Because the `alpha` function in the `psychometric` package doesn't have a facility for automatically reverse-keying items on the measure, we have to ensure that this is done when specifying the items to be used in the calculation. The first item on this scale is negatively-keyed, as we can see by consulting the data description found in the help file (`?bfi`).

To proceed, we need to reverse-key this item so that it is measured on the same scale as the other four items. There are 6 response options for each item, and so we can reverse-key this item by subtracting each value from 7.

```
alpha(data.frame(A1=7-bfi[,1],bfi[,2:5]))
```

```
## [1] 0.7037559
```

The resulting alpha (0.70) is considerably better.

We can also calculate the 95% confidence interval for this alpha (of 0.7), given that we know that there are 5 items on the measure, and approximately 2800 participants in this dataset.

```
alpha.CI(alpha=.7, k=5, N=2800, level=.95)
```

```
##         LCL ALPHA       UCL
## 1 0.6820654   0.7 0.7172182
```

This suggests that the reliability of the agreeableness measure ranges from 0.68 to 0.72, given a sample size of 2800 participants.

Applying the Spearman-Brown Prophecy

But what if we wanted to estimate the reliability of this measure with a larger number of items? This is a highly plausible scenario - a reliability of 0.70 is barely acceptable for most applications, and five items is a comparatively small scale length. You will recall that the Spearman-Brown prophecy can be used to estimate the reliability of a measure after adding (or removing) a number of parallel items to the scale. To do this, we need to calculate a value for "k", the ratio of the number of items on the new scale, to the number of items on the original scale. Thus, if we wanted to estimate the reliability of our measure (with a reliability of 0.70), if we had 10 items instead of 5, we would calculate a value of 2 for "k". To estimate this new reliability, we can use the `SBrel` function in the `psychometric` package.

```
k <- 10 / 5
SBrel(Nlength=k,rxx=0.7)
```

```
## [1] 0.8235294
```

This suggests that our new reliability would be estimated to be 0.82.

We can also estimate the length of the scale, were it to be predicted to have a reliability of 0.90. For this, we can use the `SBlength` function in the `psychometric` package.

```
SBlength(rxxp = 0.90, rxx = 0.70)
```

```
## [1] 3.857143
```

This is the factor by which the length of the measure will need to increase, meaning that we need to multiply this value by the original number of items (5) in order to find out how long the test must be. Note that we must always round up for this calculation:

```
5 * SBlength(rxxp = 0.90, rxx = 0.70)
```

```
## [1] 19.28571
```

This means that we would predict a need for a measure with 20 items, if we are to see a Cronbach's alpha of 0.90. We would need to add 15 parallel items to our existing measure.