16.5: Instrument reliability and validity

Introduction

What's in a **measure**? We've talked about measurement extensively, e.g., Chapter 3.3. As review, a measure is simply the result of some process used to quantify an object or item of interest. Instead of a number, a measure may return a classification: for a given sample unit, the unit may be categorized as meeting the definition and therefore given a "yes" or it does not ("no"), a dichotomous response. The method used to obtain the measurement is called the **instrument**. An instrument may indeed be an instrument like a sphygmomanometer or a thermocycler equipped with fluorescent optics. Instrument in this context, however, also includes questionnaires or surveys intended to determine people's responses on a particular topic.

In biology and biomedical research, there are thousands of kinds of measurements one has to choose among depending on the question at hand. In many cases choices are straightforward: in morphometrics, the instruments of choice will be lengths and areas and shapes quantified by rulers and application of well-defined geometry equations. Where multiple measurement approaches apply, **reliability analysis** can help decide which method to use, or, importantly, whether the different approaches agree. For example, Kruse et al (2017) compared ultra sound and magnetic resonance imaging measurements of Achilles tendon cross-sectional area; they found that although both methods were **internally consistent**, the methods consistently yielded different results.

In other arenas, the choice of instrument will be less clear. For example, doctors use a questionnaire to rank cardiac patients for attention in perioperative care, the care a surgical patient receives from admittance to release from the hospital, to improve patient outcomes. The questionnaire will include a number of questions intended to provide a summary picture of each patient so that if resources are limited, the most at risk patients may get priority. To the extent that the questionnaire in fact is a useful discriminant, then the instrument may benefit both hospital and patients.

In conducting measures one selects instruments that provide valid results. That is, provided the instruments are maintained and well-calibrated, use of a sphygmomanometer by a trained technician will return accurate and valid measures of a subject's blood pressure. Survey questions also can be evaluated for validity, although the extent to which survey questions measure what is intended may be more complex. For example, if the intent is to ascertain a subject's chance (i.e., risk) of graduating from college in the next year, how useful would the following question be if administered to a room filled with first-year students?

Survey question: How old are you now?

Simple enough question, but immediately, several questions come to mind. Do we want our responses in years, months, days, hours, minutes, or seconds? What about for those individuals that know only approximately when they were born (i.e., in many parts of the world, registration of birth is irregular)? So, we may even wonder about how necessary it is we ask this question of college students in the first place. E.g., do we really want to trigger our subjects for a case in which most of our subjects are about the same age?

Perhaps we decide this is important information to ask. When do you start counting? Most Western cultures start the clock at zero when the baby is born. In China and many other Eastern Asian countries, people are born at one. In India, once a person reaches a year plus six months, the person would be considered two years old, whereas in the USA, the person would still be considered one until the second birthday. Thus, depending on the person's culture identity, responses to this simple question may differ by as much as a year.

Types of reliability

Regardless of instrument, all measures contain error. Hence, even a valid instrument may not return an accurate measure for each subject. The concept of instrument **reliability** is concerned with error of measurement. Reliability may be defined in at least four contexts:

- internal consistency
- inter-rater (also called inter-observer reliability)
- parallel-forms
- test-retest

For an instrument to show internal consistency, this implies that the survey has multiple questions that pertain to the same concept or topic, but written in different ways to reveal effects of word choice, for example. Inter-rater reliability refers to an instrument that when used by different observers (e.g., science fair judges), the observers give the same or at least consistent scores for the

same test. Parallel forms reliability implies that two surveys on, perhaps, scientific literacy in high school students yield the same conclusions even though each survey has different questions. Test-retest reliability is a straight-forward concept — if the instrument is repeated, are the same scores achieved?

Reliability estimators

In general, correlation-type measures can be used to quantify the extent of reliability (also termed reproducibility or repeatability). The product moment correlation is used to quantify the relationship between two measures where there is clear distinction between the two variables. For example, to quantify the association between body weight and height, the proper correlation to calculate would be the product moment correlation because it is clear that a measure of weight goes with the variable weight whereas height measures goes with the variable height.

But it is less clear which variable should go first in the calculation when you have repeat measures of essentially the same thing. Which goes first, the first observation of sprint running speed over a 100 meters of the second measure of the same person's performance? Logically, we may say take the first as the X variable and the second as the Y variable, but there is no mathematical justification.

For a more challenging example, consider measures of body mass on male and female birds that are mated, and the researcher wants to assess whether there is a correlation between male and female weight — which variable goes first in the analysis, male weight or female weight? In such cases the intraclass correlation coefficient may be used (Chapter 12.3, Chapter 16.4). The intraclass correlation can be estimated as the ratio of the variance of interest over the sum of the variance of interest plus the variance error. Interest in the case of sprint running would be the two (or more) trials; for the bird weights, the variable of interest is weights of male and female birds within a mated pair. The formula for ICC in this context is given by

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_e^2} = \frac{MS_B - MS_e}{MS_B + (k-1)MS_e}$$

where k is the number of repeat measures, MS refers to the mean squares from the one-way ANOVA, B refers to variability between (among) subjects and e is the error or within-subjects variability.

Cronbach's alpha is a reliability measure that quantifies the internal consistency of items in a survey or instrument by calculating the average among these items. Cronbach's alpha will tend to increase as the intercorrelations among test items increase, and in this sense can be taken as an internal consistency estimate of the reliability of test scores.

$$Cronbach's\ \alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{n} s_i^2}{s_T^2}\right)$$

where k is the number of items, s_i^2 is the variance of the i-th item, and s_T^2 is the variance of the total score after summing all items.

Cronbach's alpha is one of the oldest measures, and at least in part because of how long ago it was introduced, is very common as a measure of reliability. However, there are other estimators and in some aspects these perform better than Cronbach's alpha.

Reliability statistics like Cronbach's alpha are available in the R package `psych` . See also R package `agRee` .

Example. Judging of posters

```
library(psych)
alpha(myJudge)
```

Output from R

```
Reliability analysis
Call: alpha(x = myJudge)

raw_alpha std.alpha G6(smc) average_r S/N   ase  mean   sd median_r
     0.74      0.74    0.59       0.59 2.9  0.21   2.8  1.1     0.59
```

```
lower alpha upper 95% confidence boundaries
0.33 0.74 1.15
```

Questions

[pending]
