Hasselt University
Censtat

Multivariate Data Analysis

Project B: Statistical Validity and Reliability

by

Banerjee Soutrik
Kishtammaragari Manjularani
Lauwers Kris
Rosius Wannes

Diepenbeek, February 27th, 2006

Contents

1 Introduction

The quality of a test is reflected in two statistics called validity and reliability, the first being the most important. In short, validity consists of correlations between the test and the world outside it, while reliability consists of correlations within the test. In other words, with validity and reliability we test ,respectively, whether we are measuring what we intend to measure and whether the same measurement process yields the same results.

In this paper we go deeper into these two concepts and illustrate them briefly with a case study.

2 Validity

Validity refers to the accuracy of measure. A measurement is valid when it measures what it is supposed to measure and performs the functions that it purports to perform. Validity is often defined for an instrument, a scale, or an indicator in a vast range of fields, *viz.*, social sciences, engineering, bio-medical sciences, veterinary and agricultural sciences, and research. As measurement in itself is imperfect, validity is a question of degree.

The classification of validity is generally done as follows :

Face validity: This was developed in the context of questionnaires scales. If the scales apparently 'looked' right for what it is supposed to measure, then it is called face validity.

Content validity: This implies that the questionnaire or the instrument developed to measure something covers all the necessary subject matter. For example, a questionnaire meant to assess hallucination (e.g., Neuro-Psychiatric Inventory) should cover all types of hallucinations encountered in common practice, *viz.*, visual, auditory, tactile, gustatory, olfactory, etc.

Criterion validity: In simple terms it means that we take some known quantity and compare our measurements with it. The known quantity often can be considered as a gold standard for the criterion the instrument is supposed to measure. This is again sub-divided into *predictive validity* meaning to what extent does the test predict what it is supposed to predict, and *concurrent validity* to measure an individual's present standing on the criterion variable. The latter is not meant to predict the future performance of the individual.

Construct validity: It is concerned with the extent to which a particular measure relates to other measures consistent with the theoretically derived hypotheses concerning the concepts (or constructs) that are being measured. In other words, it focuses on the extent

to which a measure performs in accordance with theoretically derived expectations. To illustrate with a previous example of assessing hallucination - a scale may have all the necessary items to evaluate hallucinations, but does it distinguish a hallucination from an illusion or a delusion? Hallucination is the false perception of a stimulus in the absence of a real stimulus, whereas illusion is a wrongly perceived stimulus and delusion is a false fixed firm belief that cannot be removed from the person's mind even after repeated logical explanations. Therefore, the questionnaire must be construct valid to be able to distinguish these concepts.

Convergent validity: It refers to the extent to which different methods of measuring the same trait yield similar results; the fundamental assumptions being that different methods of measuring the same trait should converge on the same result.

Discriminant validity: It refers to the extent to which similar or identical methods measuring different traits lead to different results; that is the traits that are truly distinct from one another should lead to different results even if they are measured by the same method.

2.1 Assessment of validity

The sensitivity, specificity, positive predictive, and negative predictive values are the commonly used indices of *criterion validity* of the information from the questionnaire.

Sensitivity is defined as the probability that the individual with the trait is correctly identified by the questionnaire as having the trait. The comparison is often done with a 'gold standard' for the criterion. *Specificity* is the probability that the individuals without the trait are correctly identified by the questionnaire that the trait is not present.

Positive predictive value is the probability that a positive test will correctly identify people with the specified trait. *Negative predictive value*, on the other hand, is the probability that a negative test will accurately identify people without the trait. The positive and negative predictive values depend on the *prevalence* of the measurement trait in the population. For this reason, one must be cautious in the interpretation of these values.

Positive and negative likelihood ratios are defined as $\frac{\text{sensitivity}}{1-\text{specificity}}$ and $\frac{1-\text{sensitivity}}{\text{specificity}}$, respectively. These can be multiplied with *prior odds* to obtain *posterior odds* regarding a 'new diagnostic test' in question.

Youden's index is sometimes used, which is defined as $\text{sensitivity} + \text{specificity} - 1$.

Receiver Operating Characteristic curve (ROC curve) is often the standard procedure to establish one method as being 'better' than the other methods : choosing the best method, when compared

to a 'gold standard'. In this, 1− specificity is plotted in the X-axis and sensitivity is plotted on the Y-axis, from which the optimum cut-off value can be chosen. Normally, the curve with the maximum area under the curve is the best method.

Control charts are used in the production line for quality control, in which the mean $\pm$ 3 standard deviations are chosen as predefined limits : UCL (Upper Control Limit) and LCL (Lower Control Limit) in order to explore the special causes of variation, whenever the measured value is greater in absolute value than these limits.

Cross-validation refers to obtaining the optimum cut-off value(s) (typically obtained from a ROC curve) from an *experimental group* and applying the cut-off value(s) to separate two (or more) groups on a *validation group* of subjects. This helps in obtaining the cross-validated sensitivity and specificity of the system, and it also has a rôle in the *predictive validity* of the system.

3 Reliability

Reliability is the extent to which an experiment, test, measuring procedure, or an indicator yields the same results on repeated trials. The measurement of any phenomenon always contains a certain amount of error. The goal of error free measurement is never attained in practice. Repeated measurements of the same phenomenon never precisely duplicate each other, they do tend to be consistent from measurement to measurement. This tendency towards consistency found in repeated measurements of the same phenomenon is referred to as reliability. The term reliability is often associated with the *precision* of measurement. In addition, reliability is also sometimes called as *reproducibility* or *consistency* in different or similar contexts.

3.1 Comparison between reliability and validity

An object which is reliable, consistently measures what it is supposed to measure without much loss of precision. However, this does not confirm that the object is valid as it could measure consistently with a certain amount of *bias*. On the contrary, the measurement may be a valid one, indicating that in the long run, the mean measurement will be very close to the theoretical or population mean, but in the process it may not be reliable, hence inconsistent or lack precision. The term validity is often associated with the *accuracy* of measurement. In Figure 1, the concepts of reliability and validity are illustrated.

Reliability is basically an empirical issue, focussing on the performance of empirical measures. Validity, in contrast, is usually more of a theoretically oriented issue because it inevitably raises the question, "valid for what purpose ?"

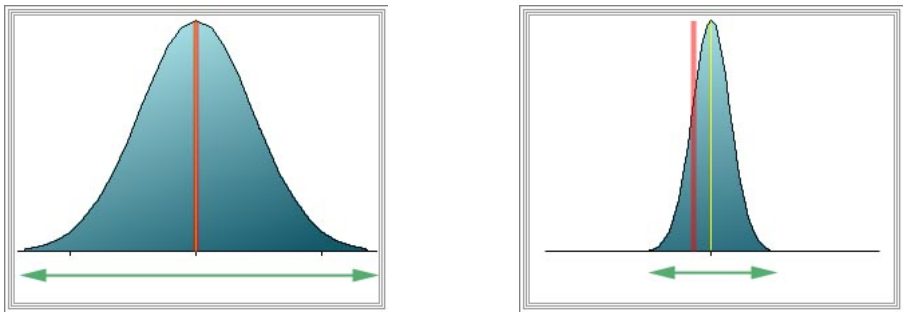

Figure 1: *The figure on the left shows that the measurement is valid, but not reliable (hence, low precision, but accurate) with high variance. The population mean and sample means are the same (overlapping). The figure on the right, on the contrary, shows that the population mean and the sample means do not overlap, but there is a bias (hence, greater precision, but lack of accuracy). There is low variance on the part of the measurement of the sample mean, signifying that it is reliable, but not valid.*

3.2 Random and nonrandom measurement error

There are two basic kinds of errors that affect empirical measurements : random error and nonrandom error. The former is used to designate all of those chance factors that confound the measurement of any phenomenon. *The amount of random error is inversely related to the degree of reliability.* The latter signifies that there is a systematic biasing effect on measuring instruments. In Figure 1 (left), there is a large amount of random error, but no nonrandom error. In Figure 1 (right), there is a smaller amount of random error, but in addition, there is a nonrandom error resulting in a systematic bias. This nonrandom error lies at the heart of the concept of validity. Just as reliability is inversely related to the amount of random error, validity depends on the extent of nonrandom error present in the measurement process.

3.3 Reliability of measurements

Since random error must be considered in the measurement of any phenomenon, we begin with the basic formulation,

$$x = t + e$$

where x is the observed score, t is the true score, and e is the random error. Since the expectation of the random error is zero, i.e., $E(e) = 0$, we have,

$$\begin{aligned} E(x) &= E(t) && (1) \\ VAR(x) &= VAR(t+e) \\ &= VAR(t) + 2COV(t,e) + VAR(e) && (2) \\ &= VAR(t) + VAR(e) \end{aligned}$$

Since the assumption is that the covariance (or correlation) between the trues scores and random errors is zero, it implies that $2COV(t, e) = 0$. Equation (2) signifies that the observed variance is equal to the sum of true score and random error variances. Given this, the ratio of the true to observed variance is called the reliability (ρ_x) of x as a measure of t. In other words, the true score variance equals the observed variance multiplied by the reliability of the measure.

$$\begin{aligned} \rho_x &= VAR(t)/VAR(x) \\ &\Downarrow \\ VAR(t) &= \rho_x VAR(x) \end{aligned} \quad (3)$$

Reliability can also be expressed in terms of the error variance as obtained by combining equations (2) and (3). Equation (4) makes it obvious that the reliability of a measure varies between 0 and 1. In sum, the greater the error variance relative to the observed variance, the closer the reliability is to zero. On the contrary, when the error variance approaches zero, then the reliability approaches unity.

$$\rho_x = 1 - [VAR(e)/VAR(x)] \quad (4)$$

3.4 Parallel measurements

In this section, we elaborate how to estimate the reliability (ρ_x) of a measure. In order to do so, at first we describe what is meant by parallel measurements. *Two measurements are defined parallel, if and only if, they have identical true scores and equal error variances.*

Symbolically, x and x' are parallel, if $x = t + e$ and $x' = t' + e'$, where, $\sigma_e^2 = \sigma_{e'}^2$ and $t = t'$. Or, if the response to the items differ only with respect to random fluctuations, then the items are considered to be parallel. Parallel items are functions of the same true score and the differences between them are the result of purely random error. It can be shown that the correlation ($\rho_{xx'}$) between parallel measures is equal to the true score variance divided by the observed variance.

That is,

$$\begin{aligned} \rho_{xx'} &= \sigma_t^2/\sigma_x^2 \\ &\Downarrow \\ \sigma_t^2 &= \rho_{xx'}\sigma_x^2 \end{aligned} \tag{5}$$

as both the terms in the RHS of equation (5) (below) are observable. Substituting the values of equation (5) into equation (3), we get that the estimate of reliability (ρ_x) is simply the correlation ($\rho_{xx'}$) between parallel measures. In the equation form,

$$\rho_{xx'} = \rho_x. \tag{6}$$

In addition, the greater the number of separate measurements of a given phenomenon, the more accurate (and higher) the estimate of its reliability will be. This estimate will only be accurate if the items are actually parallel - that is identical true scores and equal error variances. It should also be noted that the correlation ($\rho_{(x,t)}$) between the true and observed scores is equal to the square root of the reliability, hence also equal to the square root of the correlation between parallel measures. That is,

$$\rho_{(x,t)} = \sqrt{\rho_{xx'}} = \sqrt{\rho_x} \tag{7}$$

The square root of the reliability of a measure in (7) provides an *upper bound* for its correlation with any other measure, y (which may or may not be parallel). This means that a measure of a reliability of 0.81 can never correlate greater than 0.9 with another variable.

3.5 Assessment of reliability

Test-retest method: This is one of the easiest method to estimate reliability. The same test is given to the same people after a period of time. One then obtains a correlation between the two scores on the two administrations. If the measures were parallel, the correlation between them will be an estimate of the reliability by equation (6). The disadvantages of such procedures are that they are more expensive, reactivity of the subjects to the first test can influence the result of the second test, overestimation of the responses in the retest due to memorisation of test items, which in turn leads to inflated reliability estimates, and lastly the theoretical concept during the first administration might have changed during the retest. It is recommended that the retest should be administered at least after a time period greater than 1 month.

Alternative-form method: This is to overcome the disadvantages encountered in the first variety. Here a new set of test items are employed in the retest, called the alternative form.

The correlation between the alternative forms also provide an estimate of the reliability. The retest is typically spanned two weeks apart. The main difficulty lies with constructing alternative forms that are nearly parallel.

Split-halves method: Both the first two methods have the disadvantage that two tests need to be employed. In this procedure, the items are divided into two halves and correlation is obtained between the scores of the two halves. A correction of the correlation, hence the reliability estimate, is obtained by Spearman-Brown prophecy formula applicable for the *entire* set of items. The main disadvantage with this method is that with the choice of different possible halves, reliability estimates may vary significantly.

Internal consistency method: By far the most popular of these methods was proposed by Cronbach (1951), which is a generalised version of Kuder and Richardson's (1937) formula number 20. The formula of Cronbach to calculate the coefficient α is as follows :

$$\alpha = \frac{N}{N-1}\left(1 - \sum_i \frac{\sigma^2(Y_i)}{\sigma_x^2}\right)$$

where, N is the number of items, $\sum_i \sigma^2(Y_i)$ is equal to the sum of item variances (within variances), and σ_x^2 is equal to the variance of the total composite. If one works with correlation matrix rather than the variance-covariance matrix, then α reduces to the following expression:

$$\alpha = \frac{N\bar{\rho}}{1 + \bar{\rho}(N-1)}$$

where, N is again the number of items, and $\bar{\rho}$ is equal to the mean inter item correlation. Specifically, coefficient α for a test having $2N$ items is equal to the average value of the α coefficients obtained for all possible combinations of items into two split-halves test. Alternatively, α can be considered a unique estimate of the expected correlation of one test with an alternative form containing the same number of items. It has been demonstrated that coefficient α can be derived as the expected correlation between an actual test and a hypothetical alternative form of the same length, one that may never have been constructed. In addition, it has been proven that the coefficient α provides a *lower bound* of the reliability of an unweighted scale of N items. This signifies that the coefficient α is actually conservative in nature. It is equal to the reliability if the test items are parallel. *Thus, the reliability of a scale can never be lower than coefficient α even if the items depart substantially from being parallel measurements.*

While increasing the number of items in a scale can improve α, there are significant limitations to this procedure. First, adding items indefinitely makes progressively less impact on reliability. Second, greater the number of items in a scale, more time and resources are spent constructing the test instrument and causes less acceptability to the subjects who respond the the test items. It should be noted that sometimes adding new

items can in fact reduce the scale's reliability if the additional items substantially lower the average interitem correlation. In other words, this means that multidimensionality will decrease the coefficient α. It is important to realise that even though coefficient α is more complex, it has the same logical status as coefficients arising from the other methods of assessing reliability.

What is a satisfactory level of reliability? The usual opinion is a coefficient above 0.7 is considered satisfactory, and above 0.9 to be very good. The coefficient α also provides the *lower bound* to other lesser used coefficients of reliability estimators like theta and omega, with the latter provides the highest reliability estimates of the three.

We have included a recently published paper by one of the authors in the appendix, which uses the coefficient α to validate a mini-questionnaire.

Cohen's Kappa statistic: With questionnaires which measure conditions or states represented by categorical variables (e.g. 'disease / non-disease', 'mild / moderate / severe'), reproducibility is most commonly and appropriately assessed using the Cohen's kappa statistic. A kappa of 0 to 0.20 indicates slight agreement, 0.21 to 0.40 indicates fair agreement, 0.41 to 0.60 moderate agreement, 0.61 to 0.80 substantial agreement and 0.81 to 1 perfect agreement.

Intra-class correlation coefficient: The Pearson's product-moment correlation coefficient or Spearman's rho are not good for measuring agreement, when the data are continuous. In such cases, similar to Cohen's Kappa statistic for categorical data, the intra-class correlation coefficient is an appropriate choice.

4 Case study: Nocturnal activity of normal vs. demented elderly subjects

The present dataset consists of 27 patients, of which 21 are demented and 6 are controls. Nocturnal activity (total and cumulative) was measured by an unobtrusive, passive, infrared, environmental sensor system, in seconds, from 00:00 until 06:00, for five continuous nights, in all patients during their stay in a geriatric rehabilitation department in Grenoble, France. The pertinent question was to explore if there was intra-group reliability of nocturnal activity. This was assessed by measuring Cronbach's alpha index on each group for the 5-nights' measurement. This data analysis was done in SPSS. The values obtained were 0.82 in the demented group and 0.67 in the control group. The results are illustrated in Figure 2.

Interpretation: This implies that during the 5-night stay in the hospital, the nocturnal activity remained constant (hence reliable) in each of these groups. The higher reliability index in the

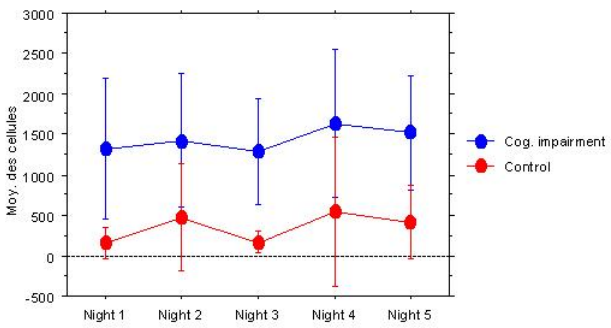

Figure 2: *This Figure shows there is inter-group difference of nocturnal activity, but intra-group reliability. The mean nocturnal activities in seconds are plotted with 95% CI.*

demented group could, partially be, explained by the greater inter-subject variance in that group. The same finding was also found in a previous study by (Lavie et al. 1992).

5 Summary

Apart from obviously depending on the quality of the records obtained, both validity and reliability are affected by sheer numbers. The more records, the higher the validity and reliability.

Understanding how to choose proper research variables and thus control for concerns about validity will make the difference between a solid and a meaningless study.

References

[1] Banerjee S and Couturier P. A mini-questionnaire to assess the acceptability of an environmental, unobtrusive, patient-monitoring device. *J Telemed Telecare* 2006;**12**(1):50-52

[2] Bland JM and Altman DG. Statistical method for assessing agreement between two methods of clinical assessment. *Lancet* 1986;**i**:307-310

[3] Bland JM and Altman DG. Cronbach's alpha. *BMJ* 1997;**314**:572

[4] Bland JM and Altman DG. Validating scales and indexes. *BMJ* 2002;**324**:606-607

[5] Carmines EG and Zeller RA. *Reliability and Validity Assessment*. Quantitative Applications in the Soc. Sc. Series, Sage Publications Inc., 6th reprint, 1983

[6] Lavie P, Aharon-Peretz J, Klein F, Gruner F, Epstein R, Tzischinsky O, and Herer P. Sleep quality in geriatric depressed patients: comparison with elderly demented patients and normal controls and effects of Moclobemide. *Dementia* 1992;**3**:360-366

[7] Saw SM and Ng TP. The design and assessment of questionnaires in clinical research. *Singapore Med J* 2001;**42**(3):131-135

[8] *Statistics*, Schaum's Outlines series, 3rd ed., 1999

[9] SPSS FAQ: What does Cronbach's alpha mean? http://www.ats.ucla.edu/stat/spss/faq/alpha.html (visited on 14-05-2005)

A A mini-questionnaire to assess the acceptability of an environmental, unobtrusive, patient-monitoring device.

See next pages