Empirical Research and the Problem of Error

Raymond J. Ballard, Ph.D.
Professor of Economics
Texas A&M University – Commerce
2600 South Neal Street, Commerce, TX 75428 USA

Dale R. Funderburk
Professor of Economics
Texas A&M University – Commerce
2600 South Neal Street, Commerce, TX 75428 USA

Abstract

In this paper we discuss the error when doing empirical research. The concepts of validity and reliability are discussed along with various ways of measuring reliability. We discuss the measurement of repeatability and reproducibility of empirical measurements. We then discuss reliability where one or several characteristics of the empirical counterpart to a theoretical variable are needed.

Keywords: Estimation, validity, reliability, error, Pareto Principle

Introduction

Assume that someone tells you the probability of a certain event occurring is zero. You ask that person where he got his information and he replies, "My ouiji board told me." From that one might conclude that the person may be coherent, even if not very realistic. But if he tells you that he estimated the probability of the event occurrence and arrived at the claimed probability of zero, he is both unrealistic and inconsistent. The fact is, anything that is estimated has an estimation error. Thus the probability cannot be zero if it is estimated. As the French mathematician and astronomer Pierre-Simon Laplace showed some two centuries ago, even an estimation error approaching infinity leads to a probability that approaches 0.50.

Once a researcher has a theory about anything, he/she must be concerned about measurement of the real world counterparts of the theorized variables. (It is assumed that the variables of interest have real world counterparts that are capable of measurement.) Given that one has correctly identified the real world variables that match the theoretical variables, only then can the researcher begin their empirical work.

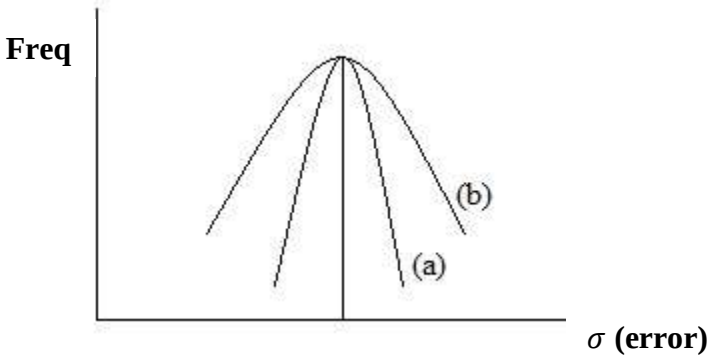
Validity and reliability

To ensure that one has the correct empirical counterpart, he must determine whether the variable to be measured is both valid and reliable. In empirical work the researcher should be concerned with both types of error. Validity has to do with the degree to which a study accurately reflects or assesses the specific concept that the researcher is attempting to measure. Reliability, on the other hand, has to do with the extent to which an experiment, test, or any measuring procedure yields the same results on repeated tests. Validity answers the question of whether the observed (empirical) variable is a true counterpart of the variable in the theory. For example, suppose a theory posits that the length of a person's left thumb is directional proportional to their intelligence. If we repeatedly measure the length of an individual's left thumb a thousand times we may argue that the measurement is reliable, given that we get the same (or very nearly the same) length each time. So, the measurement is reliable. But is the length of a person's left thumb a valid measure of their intelligence? Probably not! The point is that the measurement of the length of the left thumb can be reliably measured, but that mean it is valid for measuring intelligence? Reliability of measurement without validity of the measuring stick leaves the empirical researcher lacking in the relevance realm.

Perfect reliability (measurement) is not possible in the real world, since all measurements are subject to error. If a measure is to be considered "reliable enough" then that measure should result in similar rank ordering as say, one hundred measures being taken.

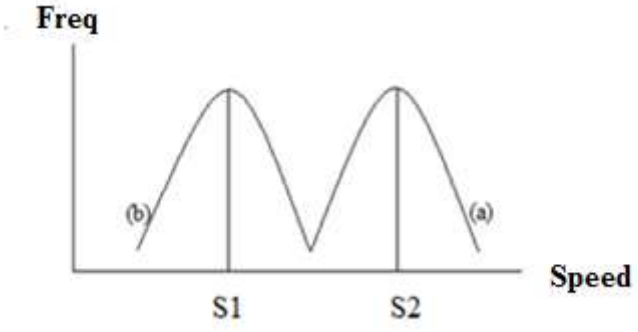
Note that rank correlations should be close to 1.0, but will always be something less than 1.0. If a variable (empirical counterpart to the theoretical variable) is valid, then the rank correlation between two measures represents the size of random measurement errors. Since empirical research always involves error, one should be concerned with both the size and source of error. Reliability analysis is concerned with both the size of errors and with random error. Systematic error affects validity. As previously mentioned, measurement of a variable can be reliable but not valid. Errors are either random or systematic. Random errors create variation in measurements that are repeated many times under the same conditions. These errors are expected to follow a symmetrical distribution around the true value of the variable. The standard error of the measure (e.g., the length of a person's left thumb) is the standard deviation of the distribution of errors. See Figure 1, where distribution (a) provides a more reliable measure than distribution (b).

Figure 1.



The question of validity requires one to check for systematic error. Systematic error occurs when a measure consistently overestimates or underestimates what one is trying to measure (e.g. intelligence). Of course this does not affect the reliability of the measurement. For example to determine the mean speed of vehicles on a highway a policeman uses a radar gun to measure speed over a period of two weeks. The results show that average (mean) speed is 60 mph in a 40 mph zone. Are people really on average going that fast? To answer that question one would want to check the radar gun to see if it was correctly calibrated. If it is found to consistently overestimate vehicle speed by 15 mph that would present no particular problem since systematic errors can be factored out. In the above example we can subtract 15 mph from all the data on speed and arrive at the approximate mean speed. See Figure 2.

Figure 2



It should be noted that validity is affected by misspecification (measuring the wrong thing). Again one can reliably measure the length of a person's left thumb, but is that a meaningful way to measure the person's intelligence? Probably not. This type of invalidity is not like calibration error. Validity is a matter of degree. If sources of nonrandom errors are known (e.g., the radar gun illustrated systematic error), the researcher can model their effects or adjust the data by the known systematic error (e.g. miscalibration). However, many measures are not one dimensional. The act of measurement itself may affect the measurement. (This is well known in quantum physics.) If researchers have a well-defined theory about something, they are then able to list sets of outcomes that can be observed.

A valid indicator for a concept will reflect the relationship in its correlations with these variables. But if the association fails to be confirmed, then either the measure is invalid, the theory about the concept is invalid, or the predictions are not correctly measured.

It follows from the definition of parallel measures that the correlation of observed data and their population values can be written as Equation (1).

Equation 1 $R_{\bar{t},t} = \sqrt{r_{1\,2}} = \sqrt{Reliability}$

Note that Equation (1) limits the size of the correlation one variable can have with another variable. Its size is limited to the square root of its reliability. If your measure has a reliability of 0.81 it cannot correlate more than 0.90 with any other variable (e.g. your measures of the length of a person's left thumb with a reliability of 0.81 then no other variable will have a correlation greater than 0.90. It is reasonable that a measure can be parallel to itself when two samples are taken from the same population. We would expect differences between the samples only in terms of random error.

In summary thus far, reliability is a measure of the amount of nonsystematic error (i.e. random error) associated with measuring. The smaller the nonsystematic (random) error is, the greater the reliability of the measuring. Validity is an indicator of systematic variation. (e.g., perhaps the measuring device is miscalibrated, as in the previously mentioned example of the miscalibrated speed detection gun). Are we actually measuring what we claim to be measuring (mean speed), or is it inaccurate due to a calibration error (systematic error) which can be adjusted out of the data?

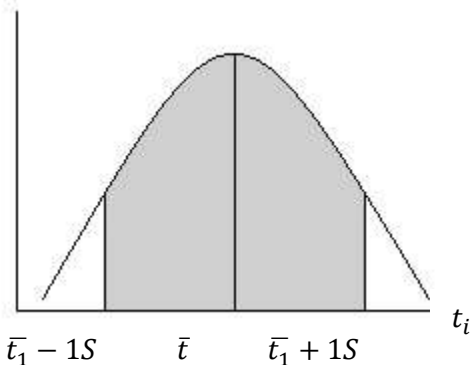
The following discusses additional ways to determine the reliability of a measure (data). If one takes many observations (measurements), the central limit theorem would lead us to expect the mean to be the true value of the variable being measured. Now, relating reliability to a probability, let $\bar{t}$ represent the true value of the variable of interest. The reliability of the measured variable (t) can be represented as a probability. For simplicity, let us assume measurements are normally distributed. In order to be useful in empirical studies, it would be important to sample the population of interest so as to get an idea of what the population distribution looks like. Again for simplicity, assume the population of interest is approximately normally distributed.

Let $\bar{t}$ represent the true value or measure of the variable you are looking for. Let R represent reliability of a measure. We can then write Equation (2) to relate reliability to probability.

Equation 2 $R(t_1) = P(-1s \leq \hat{t} \leq 1s)$

Let (s) represent standard deviation. Equation (2) is the probability that a particular measure or observation (t_i) is within one standard deviation of true value of variable of interest. See Figure 3:

Figure 3



It may be asked, for example, what is the reliability of an observation or a measure that equals 50 and the standard deviation (based on sampling) is 10, so what is the reliability of the observation or measure of 50? To find the percentage area between $\bar{t}$ + S and $\bar{t}$ - S, we determine how many standard deviations on either side of $\bar{t}$ our measure (50) is from the true value of the measure. Given the above that t_i = 50 and S = 10, the reliability or probability that the observed or measured value is within one standard deviation of the true value ($\bar{t}$) is about 0.68.

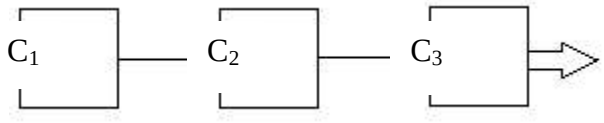
This relatively low reliability is due to a relatively large standard deviation. Note we cannot change the standard deviation. This is an extreme assertion that the standard deviation is one fifth the expected value ($\bar{t}$). One would of course attempt to find measures without this much variation.

Observations with More than One Dimension

The reliability of observations that have several dimensions to them can easily be determined. For illustrative purposes assume that an observable variable has three dimensions (characteristics), and then consider the reliability of the observed variable.

Let C_1 represent one characteristic, C_2 another characteristic, and C_3 another characteristic. We can represent this as Figure (4).

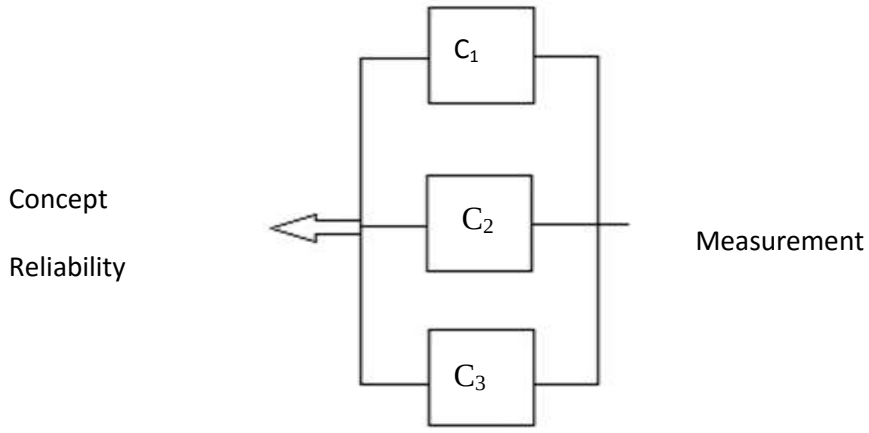
Figure 4



Now assume that we wish to measure one observation that has three characteristics that will affect the overall reliability of the measure. Again let ($\bar{t}$) represent the true value in the population variable and assume that values in the population are approximately normally distributed around the true value ($\bar{t}$). Empirically, one could make several measurements of each characteristic of the variable of interest. The overall reliability of the observed (measured) variable of interest will depend on the reliabilities of each of the three characteristics. Again, assume the true measure of the variable of interest ($\bar{t}$) follows a normal distribution. Also, assume the calculated reliabilities for the three characteristics of the variable of interest yielded the following: (C_1=0.90), (C_2=0.95), and (C_3=0.98). Given the assumptions we have made that the reliability of the measured variable of interest depends on each of its three characteristics, reliability of the variable of interest is then calculated as C_1x C_2x C_3 or 0.90x0.95x0.98=0.85 rounded up.

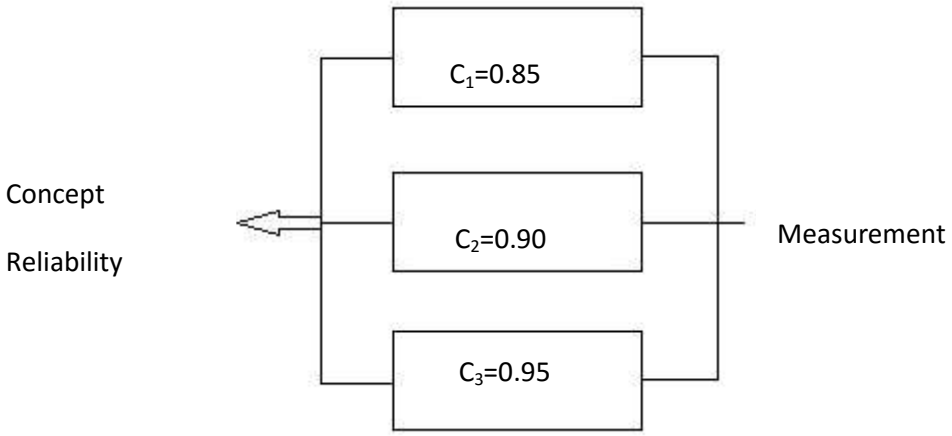
If only one of the characteristics is reliable enough (meaning that the reliabilities of the other two are lower than required) the reliability of the measured concept can be calculated. Again, assume that the variable of interest has three characteristics, C_1, C_2, C_3 and you can use the measure of the concept if one or more of the characteristics meets your reliability requirement. This situation is illustrated in Figure (5) below:

Figure 5



For illustration purpose we assume that the reliability of each characteristic has been calculated. Let the reliability of C_1=0.85, C_2=0.90, and C_3=0.95. We wish to find the overall reliability of the concept itself that has three characteristics, see Figure (6)

Figure 6



Given the above reliabilities of the characteristics, we wish to determine the reliability of the concept itself. Let say you must have the reliability of the concept to be 0.90; we can write Equation (3).

Equation 3 R(Concept) = 1-[(1-0.85) ×(1-0.90) ×(1-0.95)]
R(Concept) = 1-[(0.15) ×(1.0) ×(0.05)]=0.999

It can be seen that the reliability of the measured concept is almost 100%. Given that the assumed required concept reliability was 0.90; one would conclude that the concept can be reliably measured.

In summary thus far, in empirical studies one must first be confident that the theoretical counterparts are correctly related to the concept being measured, else we encounter the problem of systematic errors (measuring the wrong empirical counterpart to one's theoretical variables). Now, a measure can be reliable but not valid (e.g., one can reliably measure the length of a person's left thumb—in that we get approximately the same length when we measure say one thousand times). However if one claims that the length of the left thumb is directly proportional to the person's intelligence, the result no doubt will involve systematic error (i.e. non-random error).

Sources of variation tend to follow the Pareto Principle since one to three sources of variation are usually responsible for most of the variation. In such a situation one would wish to find the root cause of variation and adjust for it. The Pareto Principle suggests that usually no more than one, two, or three factors cause the problem. There may be a number of small contributing factors, but only one, two or three factors that really affect variation.

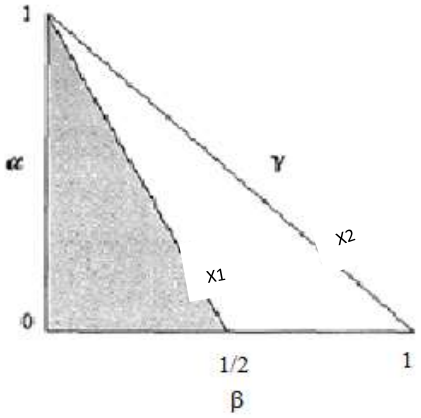
Measurement Repeatability and Reproducibility

Measurement repeatability: a measure's repeatability refers to variation in your measuring when you make repeated measurements on a sample using the same measurement condition. Repeatability is measured by the standard deviation of the repeated measures on the same sample of data. It can be said that measures are repeatable if the values of the measures are similar under the same measurement conditions. On the other hand, measurement reproducibility is measured as the variation of the means of the measurements when they are measured by different people. Reproducibility can be determined by the standard deviation of samples taken by other people. So measurements are said to be reproducible if values found by other people are close. Note, in order to minimize systematic error (shown in Figure 2 above) samples taken by others should be randomized and one can use charting to measure changes (differences) in these mean values of other's measurements. One would first need to calculate the centering of the measurements, i.e., find the mean of all the sample means that others derived $(\bar{X})$, and. then find the mean spread (standard deviation) for the measurements. One then needs to decide on acceptable limits that are adequate for the measuring, e.g. ±30.

Another View of Measurement Error

Another way to view measurement error is shown graphically in Figure (7) below. This is based on Pythagoras's theorem ($\alpha=\sqrt{\alpha^2 - \beta^2}$. Let Figure (7) represent a right triangle with the area being proportional to measurement error. Assume also α = the true population value and ½ β *x* α = measurement error, with γ representing the hypotenuse of the right triangle.

Figure 7



Note that X2 represents larger measurement error than X1 above. The maximum measurement error is given by the area ½ β.α (i.e. ½ ×1×1= ½ or 50% of the triangle). We will assume a maximum measurement error of ±50%. The shaded are of the triangle represents a measurement error of ±25% (i.e. ½×½×1=$^1/_4$). Ideally we would like to have zero measurement error (i.e. ½×0 ×1 = 0). But, this is not possible in reality. How much error is acceptable will depend on both the cost, and benefits of reducing measurement error. As mentioned previously, one can study the repeatability and reproducibility of one's measuring to determine whether or not your measurement will be adequate for your purposes.

Concluding Considerations

In summary, when one engages in empirical research they must answer the following questions:

a) Have I chosen the correct theoretical counterpart to measure?
b) If the theoretical counterpart has different characteristics, do they all need to be measured or will measuring one of the characteristics be sufficient?
c) Is the characteristic I am measuring stable over time?

References

Carmines, E. and R Zeller (1979). Reliability and Validity Assessment. Beverly Hills: Sage Publications

Ghiselli, E. and J Campbell (1981) Measurement Theory for the Behavioral Sciences San Francisco: Freeman

Halpern, J., and J Pearl (2005). "Causes and Explanations: A structural-Model Approach" British Journal of Philosophy of Science. 2005

Lentner, M. and T Bishop (1993) Experimental Design and Analysis. Blacksburg Va.: Valley Book Company

Laplace, P. S. (2009). *Essai Philosophique Sur les Probabilites*. New York, NY: Cambridge University Press. Translation of 5th edition, originally published in 1829

Montgomery, D. (1991) Design and Analysis of Experiments. 3rd edition. New York: John Wiley & Sons

Schmidt, S., and R Launsby (1992) Understanding Industrial Designed Experiments. 3rd edition Air Academy Press