Shiken: JALT Testing & Evaluation SIG Newsletter, 11 (2) August 2007 (p. 21 - 24)

Statistics Corner

Questions and answers about language testing statistics:

Sample size and statistical precision

James Dean Brown (University of Hawai'i at Manoa)

QUESTION: One topic I think many people would be interested in is something about sampling sizes and calculating sampling errors. It seems many teachers just randomly select a number between 20 - 50 when determining the size of their samples without knowing *why* (or how) a sample size for a survey should estimate – or how to calculate the error of measurement that is probably due to sampling error. Since most of the research studies language teachers conduct involve small samples, some information about calculating sampling errors is probably needed.

ANSWER: As I pointed out in my last two columns (Brown, 2006, 2007), you seem to be asking several questions simultaneously: one about *sampling* and *generalizability*, a second about sample *size* and *power*, and a third about *sample size* and *statistical precision*. I addressed sample size and generalizability in my Brown (2006) column. I addressed sample size and power in my Brown (2007) column. I will attempt to address the issues involved in sample size and statistical precision in the present column.

What are Samples and Populations, Statistics and Parameters

As I pointed out in Brown, (2006, p. 24), a *population* is "the entire group of people that a particular study is interested in." For instance, we might be interested in the population of all university EFL students in Japan. Given that few language researchers have the resources to study such a large population in its entirety, they typically use samples (i.e., subgroups drawn from the population to represent the population). In samples, we calculate *statistics* like the sample mean and standard deviation (commonly symbolized as *M* and *SD,* or *S*). But these sample statistics represent the population *parameters* for the population mean and standard deviation (commonly symbolized by the Greek letters μ and σ). One way of thinking about the relationship between such statistics and the parameters they represent is *statistical precision*, that is, how precisely the statistics from a sample represent the parameters in a population.

What is Statistical Precision?

Cohen (1988, p. 6) defines the *statistical precision* of a sample statistic as "the closeness with which

it can be expected to approximate the relevant population value. It is necessarily an estimated value in practice, since the population value is generally unknown" (Cohen, 1988, p. 6). This precision is usually estimated using a standard error, that is, the amount of chance fluctuation (or lack of precision) we can expect in sample estimates. We can use the *standard error* as an estimate of the precision of a statistic in two ways: descriptively or inferentially (for more on these two ways of looking at the standard error, see Thompson, 2006, pp. 154-155).

How is Precision Used Descriptively?

Descriptively, when precision is estimated using a *standard error*, it is thought of as the amount of fluctuation from the population parameter that we can expect by chance alone in sample estimates. For example, for a sample mean (M), we can calculate the standard error of the mean (SE_M), which provides an estimate of how much fluctuation from the population parameter that we can expect in sample estimates of M. Since standard errors are distributed normally, we can expect sample means to vary by chance ±1 SE_M 68% of the time, ±2 SE_M 95% of the time, and ±3 SE_M 98% of the time (for a review of how these percentages work, see Brown, 1988, pp. 80-85; or 2005, pp. 116-123). For example, if the mean for a sample turned out to be 78 with a conveniently round SE_M of 2, we would expect such sample means to vary by chance between 76 and 80 (68% of the time), between 74 and 82 (95% of the time), and between 72 and 84 (98% of the time). The following equation can be used to calculate the standard error of the mean (SE_M):

$$SE_M = \sqrt{\frac{S^2}{n}}$$

Where: SE_M = standard error of the mean S = standard deviation n = group size

So, the SE_M for a group of 30 students with a mean of 50 and a standard deviation of 10 would be 1.83 as follows:

$$SE_M = \sqrt{\frac{S^2}{n}} = \sqrt{\frac{10^2}{30}} = \sqrt{\frac{100}{30}} = \sqrt{3.3333} = 1.8257418 \approx 1.83$$

The standard error for a Pearson product-moment correlation coefficient (r), or the SE_r, is calculated as follows:

$$SE_r = \frac{1 - r^2}{\sqrt{n-1}}$$

So, the SE_r for a group of 30 students whose scores on two different tests correlate at .80 would be .067 as follows:

$$SE_r = \frac{1 - r^2}{\sqrt{n-1}} = \frac{1 - .80^2}{\sqrt{30-1}} = \frac{1 - .64}{\sqrt{29}} = \frac{.36}{5.3851648} = .0668503 \approx .067$$

Such standard errors can be calculated for all statistical estimates of parameters and can all be

interpreted the same way the degree to which the statistical estimates are likely to fluctuate, or put another way, as the degree to which the statistical estimates are precise.

How is Precision Used Inferentially?

Inferentially, the standard error is also commonly used in estimating the statistical significance of differences between or among parameter estimates. For example, a *t*-test can be used to estimate the probability that an observed difference between two means (say between treatment-group and control-group means) is statistically significant (i.e., that the difference is due to other than chance factors. One formula for the *t*-test (where the two samples are independent and are the same size) is as follows:

$$t = \frac{M_T - M_C}{\sqrt{\frac{S_T^2}{n_T} + \frac{S_C^2}{n_C}}}$$

Notice that the numerator represents the difference between the treatment-group mean and the control-group mean (M_T - M_C), and that the denominator contains the two standard errors for the treatment and control groups. Thus the *t*-test is simply a ratio of the mean difference to the square root of the sum of their standard errors, or put another way, the t-test is the mean difference in relation to the precision with which the two means were estimate than chance factors.

What is the Relationship Between Sample Size and Precision?

Though several factors *can* affect the precision of a parameter estimate, sample size is always a factor. As Cohen (1988, p. 6), put it, "depending upon the statistics in question, and the specific statistical model on which the test is based, reliability [i.e., precision] may or may not be directly dependent upon the unit of measurement, the population value, and the shape of the population distribution. However, it is *always* dependent upon the size of the sample."

Look at any of the equations above for various permutations of standard error, and notice that all of them have the n-size in the denominator. Hence, with all other factors held steady, as sample size increases, the standard error decreases, or gets more precise. Put another way, as the sample size increases so does the statistical precision of the parameter estimate. This has ramifications for both the descriptive and inferential uses of the standard error. Descriptively, as sample size goes up, parameter estimates become more precise. Inferentially, as sample sizes go up, parameter estimates are more precise, so differences between or among parameter estimates can be smaller and still turn out to be statistically significant.

Conclusion

As you put it in your question, "since most of the research studies language teachers in Japan conduct involve small samples, some information about calculating sampling errors is probably needed." I would extend this notion beyond your intended meaning to suggest that studies would benefit greatly from having more precision and this can be achieved most directly by increasing sample sizes. I would also argue that, when small sample sizes are absolutely unavoidable, standard errors ought to be calculated, reported, and included in the researcher's thinking about any statistical results.

References

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design*. Cambridge: Cambridge University Press.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (New edition). New York: McGraw-Hill.

Brown, J. D. (2006). Statistics Corner. Questions and answers about language testing statistics: Generalizability from second language research samples. *Shiken: JALT Testing & Evaluation SIG Newsletter, 10* (2), 24-27. Retrieved from the World Wide Web at http://www.jalt.org/test/bro_24.htm

Brown, J. D. (2007). Statistics Corner. Questions and answers about language testing statistics: Sample size and power. *Shiken: JALT Testing & Evaluation SIG Newsletter, 11* (1), 24-27. Retrieved from the World Wide Web at http://www.jalt.org/test/bro_25.htm

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford.

Last Next

http://www.jalt.org/test/bro_26.htm (HTML)

http://www.jalt.org/test/PDF/Brown26.pdf (PDF)