Questions and answers about language testing statistics:

What do distributions, assumptions, significance vs. meaningfulness, multiple statistical tests, causality, and null results have in common?

James Dean Brown
brownj@hawaii.edu
University of Hawai'i at Mānoa

Question:

The field of statistics and research design seems so complicated with different assumptions, and problems associated with each form of analysis. Is there anything simple? I mean are there any principles that are worth knowing that apply across the board to many types of statistical analyses?

Answer:

Fortunately, a number of issues are common to the most frequently reported forms of statistical analysis. I will discuss a number of those issues in the following six categories: distributions underlie everything else, assumptions must be examined, statistical significance does not assure meaningfulness, multiple statistical tests cloud interpretations, causal interpretations are risky, and null results do not mean sameness.

Distributions underlie everything else

Statistical studies investigate variables, and those variables are operationalized (i.e., observed and quantified) into scales that are nominal, ordinal, interval, or ratio (for definitions and examples of these different types of scales, see Brown, 2011a). The variables of focus in the majority of language studies are observed or measured as interval or ratio scales (known collectively as continuous scales). For many statistical analyses, such continuous scales need to be normally distributed, or if they are not, the researcher needs to consider what the effect might be of that lack of normality.

As I will explain in the next section, most statistical analysis make certain assumptions, the first of which in many cases is the assumption of normality (i.e., for the statistics to work well, the distributions in the continuous scales must be normal, or approximately normal. This is particularly important for correlational statistics, or statistics that involve correlation in any way (e.g., reliability estimates, regression analysis, factor analysis, structural equation modeling, analysis of covariance, etc.). To insure that their statistics can function appropriately, researchers always need to check the assumptions that underlie those statistics. Sadly, that is not often the case in second language research. At very least, researchers should provide descriptive statistics (including means, standard deviations, minimum and maximum values, numbers of people and items, reliability estimates, etc.) so that readers can examine for themselves the degree to which important assumptions like normality, equality of variances, reliability, and so forth have been met. That is why I report descriptive statistics and reliability estimates in my own studies before I do anything else. If all quantitative researchers would do the same, that habit would go a long way

toward increasing the quality and interpretability of the quantitative research in our field because the distributions of data (normal or otherwise) underlie everything else in statistical analyses.

Assumptions must be examined

Why do statistical tests have assumptions? The various statistical tests that researchers use were all created and tested for application under certain conditions, and they were found to work under those conditions. If those conditions do not obtain, that is, if the assumptions are not met, researchers cannot be sure if their statistics are being properly applied and accurately doing what they were designed to do. For example, the common Pearson product-moment correlation coefficient assumes that (a) the data for both variables are on a continuous scale, (b) the observations within those scales are independent of each other, (c) the distributions for the scales are normal, and (d) the relationship between the two scales is linear (for explanations of how these assumptions are defined, how they can be checked, and how the results should be interpreted when violations of the assumptions occur, see Brown, 2001, pp. 140-143). If the assumptions are met, all is well, and the researcher can interpret the results within the limits of probability that the statistics indicate. However, if the assumptions are not met, the researcher cannot be sure of the interpretations. For example, in the case of the correlation coefficient, if the distribution for one of the scales (or both) is skewed (i.e., non-normal with values scrunched up at one or the other end of the scale), it may not be appropriate to use a correlation coefficient at all, or it may be wise to adjust for the violation of the assumption by normalizing the variables. Alternatively, it may be necessary to interpret the resulting correlation coefficient very cautiously, while recognizing the likely effects of the skewing. In my experience, the likely effect when one (or both) variables is skewed is that the magnitude of any resulting correlation coefficient will tend to be depressed (i.e., will tend to provide an underestimate of the actual state of affairs). In any case, ignoring the assumptions of the seemingly simple correlation coefficient is ill-advised.

I don't want to get down in the weeds here by discussing the assumptions of every statistical procedure. The point is that for virtually every form of statistical analysis, two things are true: there is a standard error for that statistic (see Brown, 2011b), and there are assumptions that should be considered in setting up, conducting, and interpreting the analysis of that statistic (for an overview of the assumptions underlying a wide variety of statistical analyses, see Brown, 1992).

Statistical significance does not assure meaningfulness

One of the biggest problems in second language quantitative research occurs when researchers treat statistical significance as though it indicates meaningfulness. I have spent 35 years chanting that statistical significance and meaningfulness are different things, yet nothing seems to change. It is a fact that a study with a sufficiently large sample size can produce statistics (e.g., correlation coefficients, t-tests, etc.) that are statistically significant for even small degrees of relationship or small mean differences. Those p-values that lead to interpretations of statistical significance (e.g., $p < .05$ for a particular correlation coefficient) only reveal the probability that the statistic occurred by chance alone (e.g., $p < .05$ for a correlation coefficient means that there is only a 5% chance that correlation coefficient of this magnitude would occur by chance alone). That p-value does not mean that the correlation or mean difference or whatever is being tested is large, interesting, noteworthy, or meaningful. These characteristics can only be determined by looking at things like the magnitude of the correlation within the particular research context or the size of the mean difference in the context. For instance, it is perfectly valid to ask if a significant (with p

< .01) correlation of .40 found in a particular study is also meaningful and interesting. But the researcher cannot answer that question without considering the magnitude of the statistical results within the context of the specific research situation. Sometimes, a small correlation is very interesting because the researcher is looking for any sign of relationship. In such a situation, .40 would be meaningful. Other times (e.g., when costs or other stakes are very high), only a strong correlation of say .90 or higher will be meaningful. Similarly, a mean difference of 10 points on a 20 point scale might seem very interesting, but on a 1000 point scale 10 points might be far from interesting, especially if it took 300 hours of instruction to produce that one percent difference. So clearly, interpreting the meaningfulness of any statistic is different from, and additional to, first deciding whether that result has a high probability of being a non-chance statistical finding. In other words, while significance is a precondition for interpreting a statistic result at all (after all nobody wants to interpret a result that is due to chance alone), the degree to which the same statistic is interesting or meaningful will depend on the magnitude of the results and the context in which they were found. That is why statistical significance, though a precondition for meaningfulness, does not assure meaningfulness.

Multiple statistical tests cloud interpretations

Multiple statistical tests are another big problem in our research that my chanting does not seem to have affected. This phenomenon occurs when researchers perform multiple statistical tests without adjusting their p-values for that fact. During the last 35 years, I have observed multiple statistical tests in so many second language research studies that I can't even guess how many there are out there. Yet, I continue to staunchly believe (because of my training and experience with statistics) that multiple statistical tests create important problems in interpreting statistical results. I have explained this issue elsewhere in more detail (e.g., Brown, 1990, 2001, pp. 169-171, 2008), and I am not alone in holding this view (e.g., Dayton, 1970, pp. 37-49; Kirk, 1968, pp. 69-98; Shavelson, 1981, pp. 447-448; and so forth).

In brief, the problem is that conducting multiple statistical tests seriously clouds the interpretation of resulting statistical tests, usually by increasing the probability of finding spuriously significant results (i.e., results that are not really significant, popularly known as "false positives"). This problem is amplified by the fact that researchers who produce spuriously significant results do not know which of their results are spuriously significant, so even results that might actually be significant cannot be trusted. The kindest way to put this problem is that multiple statistical tests cloud interpretations. Sadly, with proper use of the analysis of variance (ANOVA) family of statistics, the effects of such multiple comparisons can be controlled (by including all of the comparisons in one omnibus ANOVA design) or minimized (by using the Bonferroni adjustments when multiple comparisons cannot be avoided) [For more on the latter topic, see Brown, 2001, pp. 169-171, 2008].

Causal interpretations are risky

Another axiom that I live by is that it is irresponsible to interpret significant statistics, even ones that appear to be meaningful, especially correlation coefficients, as indicating causality. Just because two sets of numbers seem to be related does not mean that either variable is causing the other. There are many reasons for two sets of numbers to be correlated without either causing the other. Most notably a third factor may be causing both of the variables of interest to be related. For example, when I was young and stupid, I smoked and drank coffee like my life depended on it. In fact, the numbers of cigarettes per hour and the number of cups of coffee per hour were

probably significantly correlated (at say $p < .01$). Does that mean that the coffee was causing the cigarettes or vice versa? No, of course not. There was simply a relationship. A third variable was probably causing both (e.g., fatigue, or need for stimulation, or social pressures, or advertising, or some combination of these factors). The message should be clear: be very careful if you are tempted to interpret causation based on any statistic. There may always be an alternative explanation that you overlooked for your result. That is why causal interpretations are so risky.

Null results do not mean sameness

Researchers are often tempted to interpret a lack of statistical significance (e.g., the probability is greater than 5%, or $p > .05$) as showing statistical sameness. For example, a researcher may use two ESL classes as experimental groups with one group getting some specific instructional treatment and the other group serving as a control group that gets some unrelated "placebo" treatment. Since the two groups were samples of convenience (i.e., not randomly assigned), the teacher/researcher will be tempted to compare the two groups on some form of pretest to see if they are the same at the beginning of the experiment. Naturally, they are never exactly the same, so the researcher performs a t-test to see if the difference is significant and infers (or counts on the reader to infer) from a non-significant result (i.e., $p > .05$) that the two groups were therefore statistically the same at the beginning of the study. This is not a correct inference, that is, the $p > .05$ does not indicate the probability that the two groups were the same on average. It does indicate that the researcher was unable to establish that the mean difference was statistically significant. Such a result can easily occur simply because the research design lacked sufficient power to detect a statistically significant result. Many factors can contribute to a lack of power: a sample size that is too small, measurement that lacks reliability, limited variation in ability levels for the construct being measured, etc. To determine if this is the case, procedures known as power analysis need to be included to defend any conclusion about the probability of sameness for the means of two groups. The bottom line is that a finding of no statistically significant mean difference indicates that the study was unable to establish significance, not that the two means are the same. [For further explanation of this issue, see Brown (2007a; 2007b).]

Conclusion

In the title of this column, I asked the following question: What do distributions, assumptions, significance vs. meaningfulness, multiple statistical tests, causality, and null results have in common? The simple answer is that these are six of my pet statistical peeves. To recap briefly, my pet statistical peeves are that researchers in our field often:

1. Forget to consider the potential effects of their data distributions on their statistical results (and foolishly forget to report descriptive statistics)
2. Fail to check the assumptions for the statistics they use, much less consider what violations of those assumptions mean for the interpretation of their results
3. Act as if statistical significance means that the results of their study are interesting and meaningful, which is flat out not true
4. Let multiple statistical tests cloud their interpretations
5. Make unjustified causal interpretations of their results
6. And, treat non-significant results as though they indicate the sameness of two groups

Why should anyone care about my pet statistical peeves? These peeves have developed over 35 years of experience in the ESL/EFL/Applied Linguistics field, and they are based on reading thousands of statistical studies in which I have witnessed researchers overinterpreting, underinterpreting, and/or misinterpreting their statistical results because the researchers were either ignorant of these six sets of issues or willfully ignored them. More importantly such overinterpretation, underinterpretation, and/or misinterpretation of statistical results means that the interpretations were wrong in important ways. And yet, they serve as the knowledge base of our field.

In direct answer to your question, the six sets of issues covered in this column serve as principals that are worth knowing because they are important to the quality of the statistical research in our field and because they "apply across the board to many types of statistical analyses." As a consumer of statistical studies, you can help improve the quality of the research in our field by paying attention to these issues whenever you pick up a professional journal and read quantitative research studies. My guess is that you already read such studies critically in terms of their content, but you might now want to also read them critically in terms of their statistical research methods. You can help increase the quality of the quantitative research in our field by being a critical reader, by spreading the word about these problems to your colleagues, and by complaining in letters to the editors of professional journals where you see researchers ignore these six sets of issues. Together we can help improve the statistical research methods used in the research of our field by refusing to tolerate shoddy work. How can that help but be good for the field, and good for our knowledge about second language learning and teaching?

References

Brown, J. D. (1990). The use of multiple t-tests in language research. *TESOL Quarterly, 24*(4), 770-773.

Brown, J. D. (1992). Statistics as a foreign language—Part 2: More things to look for in reading statistical language studies. *TESOL Quarterly, 26*(4), 629-664.

Brown, J. D. (2001). *Using surveys in language programs.* Cambridge: Cambridge University.

Brown, J. D. (2007a). Statistics Corner. Questions and answers about language testing statistics: Sample size and power. *Shiken: JALT Testing & Evaluation SIG Newsletter, 11*(1), 31-35. Also retrieved from the World Wide Web at http://www.jalt.org/test/bro_25.htm

Brown, J. D. (2007b). Statistics Corner. Questions and answers about language testing statistics: Sample size and statistical precision. *Shiken: JALT Testing & Evaluation SIG Newsletter, 11*(2), 21-24. Also retrieved from the World Wide Web at http://www.jalt.org/test/bro_26.htm

Brown, J. D. (2008). Statistics Corner. Questions and answers about language testing statistics: The Bonferroni adjustment. *Shiken: JALT Testing & Evaluation SIG Newsletter, 12*(1), 23-28. Also retrieved from the World Wide Web at http://www.jalt.org/test/bro_27.htm

Brown, J. D. (2011a). Statistics Corner. Questions and answers about language testing statistics: Likert items and scales of measurement? *Shiken: JALT Testing & Evaluation SIG Newsletter, 15*(1), 10-14. Also retrieved from the World Wide Web at http://www.jalt.org/test/bro_34.htm

Brown, J. D. (2011b). Statistics Corner. Questions and answers about language testing statistics: Confidence intervals, limits, and levels? *Shiken: JALT Testing & Evaluation SIG Newsletter, 15*(2), 23-27. Also retrieved from the World Wide Web at http://www.jalt.org/test/bro_35.htm

Dayton, C. M. (1970). *The design of educational experiments*. New York: McGraw-Hill.

Kirk, R. E. (1968). Experimental design: Procedures for the behavioral sciences. Belmont, CA: Brooks/Cole.

Shavelson, R. J. (1981). Statistical reasoning for the behavioral sciences. Boston: Allyn & Bacon.

Where to Submit Questions:

Please submit questions for this column to the following e-mail or snail-mail addresses:

brownj@hawaii.edu.

JD Brown
Department of Second Language Studies
University of Hawai‘i at Mānoa
1890 East-West Road
Honolulu, HI 96822
USA

Your question can remain anonymous if you so desire.