Shiken: JALT Testing & Evaluation SIG Newsletter, 1 (1) April 1997 (p. 20-23)

Statistics Corner

Questions and answers about language testing statistics:

Skewness and kurtosis

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

QUESTION: My computer program has a function that provides what it calls "basic statistics." Among those are Skew and Kurtosis. Your book on testing says that abnormally skewed and peaked distributions may be signs of trouble and that problems may then arise in applying testing statistics. What are the acceptable ranges for these two statistics and how will they affect the testing statistics if they are outside those limits? - Paul Jacquith

ANSWER: Probably the most commonly used "number crunching" software program on IBM compatibles and Macs is Excel™ (Microsoft, 1996). In my IBM version of that program, I simply select **Tools**, then select **Data Analysis**, then select **Descriptive Statistics** and hit **enter**. Then I fill in the input range (the numbers I want to analyze), and the output range (where I want to put the resulting statistics) and check **Summary Statistics**. When I hit **return**, the Excel program puts a number of useful descriptive statistics into my spreadsheet, including all of the following: mean, standard error of the mean, median, mode, standard deviation, variance, kurtosis, skewness, range, minimum, maximum, sum, count, largest, smallest, and confidence level. But, your question focuses in on the skew and kurtosis statistics. So I'll narrow the discussion to only those two statistics.

Skewness

Let me begin by talking about skewness. In its help screens, Excel defines SKEW as a function that "returns the skewness of a distribution. Skewness characterizes the degree of asymmetry of a distribution around its mean. Positive skewness indicates a distribution with an asymmetric tail extending towards more positive values. Negative skewness indicates a distribution with an asymmetric tail extending towards more negative values" (Microsoft, 1996). While that definition is accurate, it isn't 100 percent helpful because it doesn't explain what the resulting number actually means.

The skewness statistic is sometimes also called the skewedness statistic. Normal distributions produce a skewness statistic of about zero. (I say "about" because small variations can occur by chance alone). So a skewness statistic of -0.01819 would be an acceptable skewness value for a normally distributed set of test scores because it is very close to zero and is probably just a chance fluctuation from zero. As the skewness statistic departs further from zero, a positive value indicates the possibility of a positively skewed distribution (that is, with scores bunched up on the low end of the score scale) or a negative value indicates the possibility of a negatively skewed distribution (that is, with scores bunched up on the high end of the scale). Values of 2 standard errors of skewness (*ses*) or more (regardless of sign) are probably skewed to a significant degree.

The *ses* can be estimated roughly using the following formula (after Tabachnick and Fidell, 1996): $\sqrt{\frac{6}{N}}$. The square root of 6 over N For example, let's say you are using Excel and calculate a skewness statistic of -.9814 for a particular test administered to 30 students. An approximate estimate of the *ses* for this example would be: $ses = \sqrt{\frac{6}{N}} = \sqrt{\frac{6}{30}} = \sqrt{.20} = .4472$. The square root of 6 over N Since two times the standard error of the skewness is .8944 and the absolute value of the skewness statistic is -.9814, which is greater than .8944, you can assume that the distribution is significantly skewed. Since the sign of the skewness statistic is negative, you know that the distribution is negatively skewed. Alternatively, if the skewness statistic had been positive, you would have known that the distribution was positively skewed. Yet another alternative would be that the skew statistic might fall within the range between - .8944 and + .8944, in which case, you would have to assume that the skewness was within the expected range of chance fluctuations in that statistic, which would further indicate a distribution with no significant skewness problem.

On a norm-referenced test, the existence of positively or negatively skewed distributions as indicated by the skewness statistic is important for you to recognize as a language tester because skewing, one way or the other, will tend to reduce the reliability of the test. Perhaps more importantly, from a decision making point of view, if the scores are scrunched up around any of your cut-points, making a decision will be difficult because many students will be near that cut-point. Skewed distributions will also create problems insofar as they indicate violations of the assumption of normality that underlies many of the other statistics like correlation coefficients, t-tests, etc. used to study test validity.

However, violations of that assumption of normality are only problematic if the test is norm-referenced and being used for norm-referenced. As I have discussed elsewhere (see for instance, Brown, 1996, pp. 138-142), a skewed distribution may actually be a desirable outcome on a criterion-referenced test. For example, a negatively skewed distribution with students all scoring very high on an achievement test at the end of a course may simply indicate that the teaching, materials, and student learning are all functioning very well. This would be especially true if the students had previously scored poorly in a positively skewed distribution (with students generally scoring very low) at the beginning of the course on the same or a similar test. In fact, the difference between the positively skewed distribution at the beginning of the course and the negatively skewed distribution at the end of the course would be an indication of how much the students had learned while the course was going on.

". . . reporting the median along with the mean in skewed distributions is a generally good idea."

You should also note that, when reporting central tendency for skewed distributions, it is a good idea to report the median in addition to the mean. A few very skewed scores (representing only a few students) can dramatically affect the mean, but will have less affect on the median. This is why we rarely read about the average family income (or mean salary) in the United States. Just a few billionaires like Bill Gates would make the average "family income" very high, higher than most people actually make. I guess that means that most of us, even those U.S. citizens working for good salaries in Japan, would be below average in terms of family incomes in the United States. However, in terms of median income, most of us working in Japan would be above the median income in the United States. Hence, median income is reported and makes a lot more sense to most people. The same is true in any skewed distributions of test scores as well. So reporting the median along with the mean in skewed distributions is a generally good idea.

Kurtosis

The Excel help screens tell us that "kurtosis characterizes the relative peakedness or flatness of a distribution compared to the normal distribution. Positive kurtosis indicates a relatively peaked distribution. Negative kurtosis indicates a relatively flat distribution" (Microsoft, 1996). And, once again, that definition doesn't really help us understand the meaning of the numbers resulting from this statistic.

Normal distributions produce a kurtosis statistic of about zero (again, I say "about" because small variations can occur by chance alone). So a kurtosis statistic of 0.09581 would be an acceptable kurtosis value for a mesokurtic (that is, normally high) distribution because it is close to zero. As the kurtosis statistic departs further from zero, a positive value indicates the possibility of a leptokurtic distribution (that is, too tall) or a negative value indicates the possibility of a platykurtic distribution (that is, too flat, or even concave if the value is large enough). Values of 2 standard errors of kurtosis (*sek*) or more (regardless of sign) probably differ from mesokurtic to a significant degree.

The *sek* can be estimated roughly using the following formula (after Tabachnick and Fidell, 1996): $sek = \sqrt{\frac{24}{N}}$. For example, let's say you are using Excel and calculate a kurtosis statistic of + 1.9142 for a particular test administered to 30 students. An approximate estimate of the *sek* for this example would be: $sek = \sqrt{\frac{24}{N}} = \sqrt{\frac{24}{30}} = \sqrt{.80} = .8944$. Since two times the standard error of the kurtosis is 1.7888 and the absolute value of the kurtosis statistic was 1.9142, which is greater than 1.7888, you can assume that the distribution has a significant kurtosis problem. Since the sign of the kurtosis statistic is positive, you know that the distribution is leptokurtic (too tall). Alternatively, if the kurtosis statistic had been negative, you would have known that the distribution was platykurtic (too flat). Yet another alternative would be that the kurtosis statistic might fall within the range between - 1.7888 and + 1.7888, in which case, you would have to assume that the kurtosis was within the expected range of chance fluctuations in that statistic.

The existence of flat or peaked distributions as indicated by the kurtosis statistic is important to you as a language tester insofar as it indicates violations of the assumption of normality that underlies many of the other statistics like correlation coefficients, *t*-tests, etc. used to study the validity of a test.

Another practical implication should also be noted. If a distribution of test scores is very leptokurtic, that is, very tall, it may indicate a problem with the validity of your decision making processes. For instance, at the University of Hawai'i at Manoa, we give a writing placement test for all incoming native-speaker freshmen (or should that be freshpersons?) that produces scores on a scale of 0-20 (each student's score is based on four raters' scores, which each range from 0-5). Yearly, we test about 3400 students. You can imagine how tall the distribution must look when it is plotted out as a histogram: 20 points wide and hundreds of students high. The decision that we are making is a four way decision about the level of instruction that students should take: remedial writing; regular writing with an extra lab tutorial; regular writing; or honors writing. The problem that arises is that very few points separate these four classifications and that hundreds of students are on the borderline. So a wider distribution would help us to spread the students out and make more responsible decisions especially if the revisions resulted in a more reliable measure with fewer students near each cut point.

Conclusion

One last point I would like to make: the skewness and kurtosis statistics, like all the descriptive statistics, are designed to help us think about the distributions of scores that our tests create. Unfortunately, I can give you no hard-and-fast rules about these or any other descriptive statistics because interpreting them depends heavily on the type and purpose of the test being analyzed. Nonetheless, I have tried to provide some basic guidelines here that I hope will serve you well in interpreting the skewness and kurtosis statistics when you encounter them in analyzing your tests. But, please keep in mind that all statistics must be interpreted in terms of the types and purposes of your tests.

"interpreting . . . [statistics] depends heavily on the type and purpose of the test being analyzed."

References

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.

Microsoft [Computer software]. (1996). *Excel*. Redmond, WA: Microsoft Corporation.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.

HTML: http://www.jalt.org/test/bro_1.htm / PDF: http://www.jalt.org/test/PDF/Brown1.pdf