SHIKEN: JALT Testing & Evaluation SIG Newsletter. 9 (2) October 2005 (p. 2 – 7)

Some preliminary thoughts on statistics and background information on SPSS (Part 3)

by H.P.L. Molloy and T Newfields

This article introduces some additional descriptive statistics concepts for novice researchers. The first article in this series, online at http://jalt.org/test/mn_1.htm, highlighted alpha values and beta values and distinguished between Type 1 and Type 2 errors. The second article, online at http://jalt.org/test/mn_2.htm, mentioned how to calculate the using the SPSS Statistical Software Package. This final segment describes a few basic descriptive statistics which can be calculated with SPSS, how to interpret various types of curve distributions, and concludes by mentioning outliers.

7. Descriptive statistics

Besides histograms, there is another way to look at the shape of any data: through numbers. The numbers that are most important are those that show us where the middle of our data are and how widely or tightly they are clustered around those centers.

7.1 Mean, Median, and Mode

Mean refers to an average; median is the middle; mode is the most frequent number. Each is a way of seeing where the center of your data is. Ideally, the mean, median, and mode are more or less the same. If they are not, you may have problems.

If the median is much higher or lower than the mean, you either have too many high or low scores. In such cases, your distribution is said to be positively or negatively skewed and your histogram will look off-center. If the skew is not too big, you will probably be OK, but if it is very large, you may not be able to use parametric statistics.

If the mode is much higher or lower than the mean, you may have some outliers – that is, a few freakishly large or small numbers. Your histogram will look as if one or both of the ends of the curve are too long. You may have to do something about those numbers before proceeding. If you have two modes (or, more often, a histogram that has two peaks), you are probably dealing with two groups, not one.

"If you have two modes (or, more often, a histogram that has two peaks), you are probably dealing with two groups, not one."

7.2 Standard deviation

Standard deviation can be thought of as "average differentness." It shows how far the typical distribution you have is from the mean. If the standard deviation is small compared with the mean, your numbers are grouped tightly around the mean (and your histogram will look tall); if it is large, your numbers are not so tightly grouped (and your histogram will look flat). This “tallness” or “flatness” is known as kurtosis and explained in depth by J. D. Brown in an earlier *SHIKEN* article.

Standard deviation can be thought of as dividing your numbers into groups. One standard deviation above and below the mean should usually cover about 68% of your numbers. For example, if your mean is 25 and the standard deviation is 3, that means that 68% of your numbers are between 22 and 28. Two standard deviations usually cover usually cover 95% of the scores, and so on.

What does this mean in practical terms? By checking how many standard deviations away from the mean a given score is, we can see potential problems. Again, if you have a mean of 25 and a standard deviation of 3, a score of 13 may indicate some sort of problem. A score of 13 is 4 standard deviations below the mean, where we expect only 0.01 of scores to be this far from the mean. If you have 1000 or more scores, one or two scores so far away from the mean are to be expected, but if you have only 100 scores, there may be a problem with that score.

7.3 Descriptive statistics in SPSS

Menu If you didn't get the descriptive statistics with the histogram in SPSS, you can also do the following once your numbers are in the program:

```
1.      Choose the "Analyze" menu.
2.      Choose the "Descriptive statistics" item.
3.      Choose the "Descriptives" item from the submenu. Click it.
4.      On the left side of the dialogue box, highlight the variables you are interested in.
5.      Click the arrow between the white boxes.
6.      Click on the "Options" button.
7.      Click on the statistics you want to see. I recommend clicking "Mean,"
        "Std. deviation," "Variance," "S.E. mean," "Kurtosis," and "Skewedness."
8.      Click on "Continue."
9.      Click on "OK."
```

Syntax To do the same thing using syntax, type this:

```
DESCRIPTIVES
VARIABLES=[variable 1] [variable 2] [variable 3] [variable n]
/STATISTICS=MEAN STDDEV VARIANCE MIN MAX SEMEAN KURTOSIS SKEWNESS.
```

Here, replace "[variable 1]" (and so on) with the name(s) of the variables you're interested in. The names you should use are those are the tops of the columns in SPSS.

7.4 Recommendations for using SPSS

An easy way to learn about syntax is to ask SPSS to print the syntax for every procedure you use with the menus. Here is how to do it:

In SPSS, go to the "Edit" menu. Choose "Options." A dialog box will appear. Choose the "Draft Viewer" page. On that page, in the top left corner (under "Display Output Items," check "Display commands in log" and "Log." Click "OK." Now when you run a test from the menu, the syntax for that test will appear just above your results.

7.5 The normal curve

The normal curve (also known as the *z* distribution and the bell curve) is the familiar inverted-u histogram. Most parametric statistical procedures work only if your data are distributed like this normal curve. If your data are not distributed normally, you should not use parametric statistics.

Although, as mentioned earlier, there are other distributions used in educational statistics, the normal distribution is by far the most important distribution. This is partly because it has been proved (using the central limit theorem) that with histograms of anything it is possible to obtain a normal curve eventually, "eventually" here meaning when there is a large enough *N* size. It helps to keep in mind that everything resembles a normal curve.

The normal curve also comes into play directly in significance tests: a significance test can be thought of as placing your result somewhere in the normal distribution of all possible outcomes of your test. For example, if you run a t-test and SPSS tells you the significance ("sig.") of the result is 0.04, that means that your result is somewhere toward one end of the curve of all possible results of that test, far away from the "average" result. In this case, the "average" result can be thought of as the result made by random numbers or the result consistent with the null hypothesis ("there is no difference"). The farther away from the middle your result is, the more likely it is that your result is not random, that is, that it is the result of a real difference.

7.6. Skew

Skew refers to the amount your distribution is shifted to the left or the right on your histogram. Negative skew means that more than 50% the scores are toward the right of your histogram; positive skew means the opposite. Hence, with negative skew the median will be higher than the mean. This means that some of your numbers are unusually low.

How much skew is too much? How much makes your data unacceptably different from the normal curve? In your SPSS "descriptives" or "frequencies" output, check two numbers: those for "Skewness" and for "Std. Error of Skewedness." If the "Skewedness" number is more than two times as large as the "Std. Error of Skewedness" number, your data may be unacceptably skewed. (The problem is not so great with large N sizes and with certain statistical procedures.)

7.7. Kurtosis

Kurtosis refers to the degree to which your histogram is flat or peaked. It also can be thought of as referring to the degree to which your numbers are squashed together or spread out. If your standard deviation is very small, your data may be "leptokurtic," and the data in your histogram will look all bunched together toward the middle. This makes your histogram look pointed, not bell-shaped. If your standard deviation is very large, your data may be "platykurtic," and the data in your histogram will look flattened and spread out widely.

How much kurtosis is too much? How much makes your data unacceptably different from the normal curve? In your SPSS "descriptives" or "frequencies" output, check two numbers: those for "Kurtosis" and for "Std. Error of Kurtosis." If the "Kurtosis" number is more than two times as large as the "Std. Error of Kurtosis" number, your data may be unacceptably kurtic. (The problem is not so great with large N sizes and with certain statistical procedures.)

7.8 Outliers (weirdoes)

Outliers are data that do not fit with the rest of your data. They come in two flavors: univariate outliers and multivariate outliers. If you have outliers in your data, you have problems. The problems take two forms: first, you may not have been measuring correctly or measuring a few of the wrong people. Second, running statistical tests with outliers, especially multivariate outliers, can give you inaccurate results.

Here is an example. Imagine that you have measured two things: annual income (as measured in US thousands of dollars), and value of automobile (measured the same way). Here are your data:

Case no.	Income	Autovalue
1	23	15
2	15	14
3	22	14
4	19	30
5	30	20
6	24	29
7	34	16
8	25	13
9	16	22
10	17	17
11	18	30
12	132	129

You are interested in how well annual income and automobile value correlate (that is, if the two variables go up and down together). If you run these numbers using the SPSS "Analyze," "Correlate," "Bivariate" procedure, with the obvious outlier (the last case), you will find the Pearson correlation is very high: 0.957, which may lead you to believe that income indeed does vary with car price and vice versa. The association seems to be very tight.

Now run it again, this time with the outlier (Case number 12) deleted. The Pearson correlation is now 0.201, a completely different picture.

The outlier case, with an outlier on both variables, has made the variables seem much more strongly correlated than they "really" are. I shan't explain why this happens here, but it is important to remember what effect outliers can have. Very many of the common statistical procedures are based on a (mathematical) form of correlation known as regression, and outliers can have similarly strong effects on your results in those statistical procedures.

7.9 Univariate outliers

Univariate outliers are numbers that are very far from the mean in your data. In the example mentioned above, the mean was 25 and the standard deviation was 3. If you have a score of 13, you may have an outlier. A score of 13 is 4 standard deviations below the mean, where we can expect only 0.01% of scores to be. Is this score an outlier? Maybe. It depends on how many scores you have, that is, on your *N* size. If you have an *N* size of 10,000, some scores 4 standard deviations away from the mean are to be expected and are not outliers. If you have an *N* size of 100, a score 4 standard deviations from the mean is probably an outlier.

To check for outliers, compare the maximum score(s) you have with the mean and the standard deviation. If the maximum score is more than 3 standard deviations above the mean, it may be an outlier. Check the skewedness of your distribution as well: if your distribution is both unacceptably skewed and you have one or more scores far above the mean, those scores are probably outliers. Do the same with the minimum scores.

In the example above, Case number 12 is a univariate outlier for both variables.

Where do outliers come from? There are two main sources. One is mistaken data entry: you may just have put the wrong numbers in your computer, scored a test incorrectly, put a decimal point in the wrong place, or something like that. Go back and check the original data. The other is mistakenly including data from another group. For example, if you give an English vocabulary test to a group of first-year junior-high school English students and one of us happens to take the test as well, our scores will probably be much, much higher than any of the others. We don't belong to the group that you think you are testing.

What should you do about univariate outliers? There are two usual choices. One is to simply delete them. If you believe a mistake has been made in data entry or the score comes from someone who is not a member of the group you are interested in, this choice is acceptable. You should, however, be able to justify the decision.

The second choice is to change the score. If (again) you have a mean of 25, a standard deviation of 3 and one score of 7 (that is, six standard deviations below the mean), you might want to change that score to 16 (that is, three standard deviations below the norm). Why would you want to do this? Isn't it cheating? You may have reason to believe that the score is not truly representative. (Perhaps the student who made that score was sick that day.) You may also wish to preserve other scores that person has contributed it you are working with multiple scores.

7.10 Multivariate outliers

Multivariate outliers can occur when you are measuring several things for each participant. Imagine that you are doing a study in which you have given three different tests to a group of students. You are doing a multivariate study.

Multivariate outliers are different from univariate outliers. Univariate outliers mean scores too far away from the mean; multivariate outliers mean a strange pattern of scores.

It is possible for a person to be a multivariate outlier without being a univariate outlier for any variable. For example, imagine that you measure three things in a group of people: years of education, income, and value of that person's automobile. Most people you check will follow a certain pattern: high scores on each variable will go together, and low scores will go together. If you have one person who has many years of education and a really expensive car but a very low income, that person does not fit the expected pattern. The person may be a multivariate outlier.

It is also possible for a person to be a univariate outlier on every variable but not a multivariate outlier. In the example above, case number 12 is a univariate outlier for both variables, but not a multivariate outlier.

"Detecting multivariate outliers is difficult, especially if you have a large N size and many variables. SPSS offers a way to check called the Mahalonobis distance, which can be thought of as a kind of multidimensional standard deviation."

Detecting multivariate outliers is difficult, especially if you have a large N size and many variables. SPSS offers a way to check called the Mahalonobis distance, which can be thought of as a kind of multidimensional standard deviation.

To check the Mahalonobis distance, you have to run the SPSS "Regression" procedure. Alas, it doesn't seem to work with the menus, so you have to use syntax.

Enter the following in your syntax window, and click "Run," then "All."

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT [variable 1]
  /METHOD=ENTER [variable 2] [variable 3] [variable n]
  /RESIDUALS HIST(ZRESID) NORM(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(2)
  /RESIDUALS=OUTLIERS(Mahal) .
```

Here, replace "[variable 1]" with the name of one of your variables. It doesn't matter which one. Replace "[variable 2]," "[variable 3]," and so on with the names of the rest of your variables. The names you should use are those are the tops of the columns in SPSS.

Toward the bottom of the output, you will find a table titled "Outlier statistics." This contains the ten most outlying cases (or combinations of variables). The column marked "Case Number" shows which line in the SPSS data window has been checked. "Statistic" is the number you are interested in.

Is the statistic bad or good? You have to check it against a chi-square distribution. Find a chi-square table, and look at line for degrees of freedom equal to the number of variables you have. (In the case above, the degrees of freedom would be 4). It is recommended that you look at the $p < 0.001$ column. That is, you're looking for outliers that appear only one in a thousand times. If the "Statistic" number is higher than the chi-square value, that case is a multivariate outlier.

This method of checking for multivariate outliers only shows the top ten cases. If you delete one case (because it is the only one that is a multivariate outlier), you have to run the "Regression" procedure again, because there may be other multivariate outliers in the new set of data.

8. Conclusion

This article has covered a few basic statistical concepts for novice users of the SPSS software. After reviewing each of the articles in this series, you should be able to answer these questions:

1. Why should research begin with a null hypothesis rather than an alternative hypothesis?
2. How do Type 1 errors differ from Type 2 errors?
3. What are five ways to enhance the statistical power of a research project?
4. How do counting, ordering, and measuring differ?
5. When are parametric and nonparametric statistics used?
6. How do univariate and multivariate outliers differ?

For more detailed information about SPSS and research statistics, it might be worth referring to either *Psychological Statistics Using SPSS for Windows* (2001) by Robert C. Gardner or *Data Analysis for the Behavioral Sciences Using SPSS* (2002) by Weinberg and Abramowitz. Japanese readers may prefer Oshio's *SPSS to Amos Ni Yoru Shinri Chousha Deeta Kaiseki* (2004).

HTML: http://jalt.org/test/mn_3.htm / **PDF**: http://jalt.org/test/PDF/Molloy-Newfields3.pdf