Summer 2012, Volume 2, Issue 2

THE STATSWHISPERER

The StatsWhisperer Newsletter is published by Dr. William Bannon and the staff at StatsWhisperer.™ For other resources in learning statistics visit us on the web at: www.statswhisperer.com

Introduction to this Issue

INSIDE THIS ISSUE

The current newsletter issue is a rather concentrated piece that discusses removing outliers in response to a variable distribution failing to approximate a normal curve. When I first began learning statistics, one of the most frustrating situations I would encounter was being told to examine data for issues that needed to be resolved, but not being told how to resolve those issues. I encountered this event most often when I read materials concerning the testing if all assumptions for a statistical test had been made, especially when it came to checking variables for normality.

Most of the time, a text would remind the reader that when doing certain procedures, such as multiple regression, one must check the assumption that independent variables are normally distributed. However, the text would often fail to address what to do if a variable distribution was non-normal. Essentially, the text might help you identify that you have a problem, but then neglect to give any guidance concerning how to solve it.

Today, we will make one small advance in addressing the issue of treating a non-normally distributed variable. Specifically, we will discuss how to identify outliers and the utility of removing those values toward approximating a normal distribution. Also, we will discuss when is it desirable to include or exclude an outlier from analysis. It is important to realize it is not always desirable to remove an outlier from analysis.

Putting Out Fires by Removing Outliers

What exactly does that title mean? The phrase "putting out fires" has several meanings, including meaning that one is alarmingly busy with a project and has several ancillary challenges (or fires) popping up that they must deal with, which is our interpretation here.

When doing a project involving statistical analysis, a non-normally distributed variable is often like a small fire that must be dealt with so the larger project can progress. One method of dealing with the issue, or putting out the fire, is to remove the outlier score(s) from the variable distribution. The removal of outliers often changes the non-normal variable distribution into a distribution that better approximates a normal curve. Hence, this method solves the problem or puts out the fire.

However, it is important state up front that this is only one method of treating a non-normal distribution. Although not discussed in the current newsletter, there are other methods (e.g., Log Odds transformation) to be considered.

The Potentially Mighty Influence of an Outlier

Many times when we read even the most well done empirical research studies in top peer reviewed journals, the methods and analysis sections do not describe if an examination of outlier scores was conducted. Subsequently, one might be tempted to think, is the examination of outliers really that important or necessary? The answer of course is yes!

But just what are outliers? In brief, outliers are atypical, infrequent observations. In a distribution of values within a variable, such as the range of ages in the study variable reflecting participant age, an outlier is often a score or value that is either really high or low relative to all the other scores/values. Using the variable age again as an example, an outlier would be someone either much younger of older relative to the rest of the study participants.

But what about the potential influence of an outlier? Well one outlier can be responsible for the statistically significant findings of an empirical test. Again, let's take a very common procedure that assumes a normal distribution, multiple regression, as an example. Because of the way that the regression line is determined (minimizing the sum of *squares of distances* of data points from the line), outliers have a profound influence on the slope of the regression line, the value of the correlation coefficient, and resulting values reflecting statistical significance. In fact, an outlier is capable of considerably changing the slope of the regression line and the resulting statistical parameters.

This is often a real downer. Many times researchers will conduct a fast preliminary multiple regression model that yields statistically significant results that support their hypothesis. After a brief peaking of their interest, the researchers will go and make a careful analysis of outliers only to find that with the removal of one outlier, their statistically significant findings disappear and the hypothesis is not supported.

Thus, it would not be entirely inappropriate to say that if we are presented with a multiple regression model within an empirical study, but are not presented with a summary of an examination for outliers, we may perceive an unanswered question that could dramatically impact study findings.

It is also important to consider that outliers often have different levels of influence relative to the number of study participants in a sample. If a sample size is relatively small, then very big or small scores/values within a variable will have a greater influence on findings.

Let's not forget, an outlier is not an invalid score/value. An outlier is a naturally occurring value within the population. If we surveyed the whole population, an outlier would likely have a stable place and effect in the distribution. However, within most empirical studies it is not tenable to survey the whole population, so we have our smaller sample to work with. Subsequently, the outlier may represent an atypical value/score within the study that can seem somewhat out of place. Then when this value/score is combined with other more typical scores/values, the outlier almost has an overshadowing effect that can have a large and disproportionate impact of study findings.

How to Spot an Outlier: The Boxplot

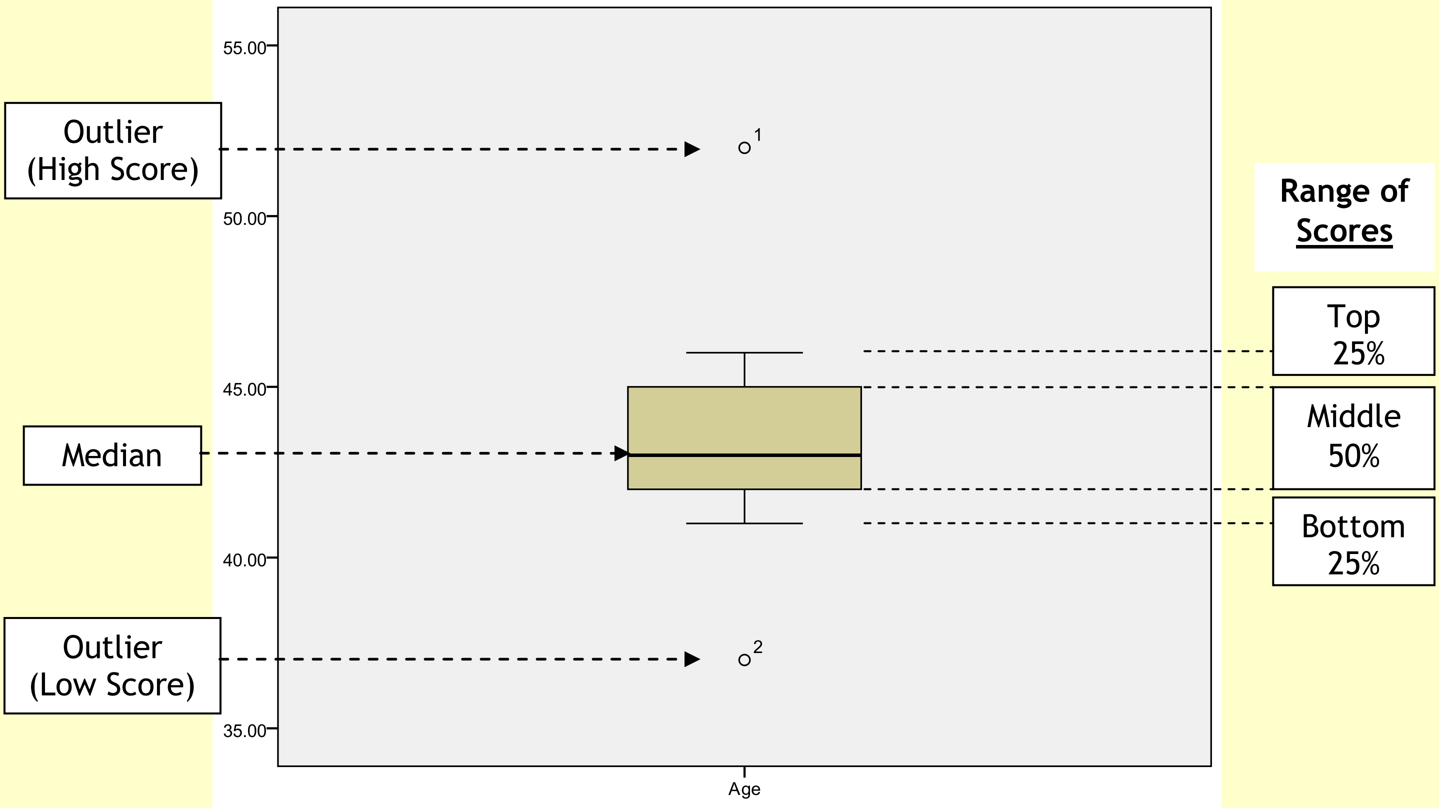
A computer generated boxplot is a common method for identifying outliers. The computation of boxplots is a widely available function in standard software packages such as SPSS. For example, such a boxplot, known as the "Stem-and-Leaf" plot is produced through when examining variables through the **Explore** function in SPSS. Specifically, to perform this test in SPSS:

Click **Analyze** > **Descriptive Statistics** > **Explore...** on the top menu.

A dialogue box will then open that reads "**Explore**" in the upper left hand corner. Then click to highlight the variable in the left hand column that needs to be tested for normality and then transfer that variable into the left hand column into the "**Dependent List**" box clicking the arrow button to the left of the "**Dependent List**" box.

Here we will use a variable common to most studies as an example, which is the variable *Age*. Next, click the button that reads "**Plots**" in the upper right hand corner. In the new dialogue box, click the box next to the word "Stem-and-Leaf" then click the "**Continue**" button. Then finally click the "**OK**" button in the bottom left hand corner of the dialogue box. A Stem-and-Leaf box will be produced similar to Figure 1 below.

Figure 1. A Stem-and-Leaf Boxplot Presenting the Study Variable *Age*



How to Spot an Outlier: The Boxplot (continued)

Again, Figure 1, which presents the distribution for the variable *Age* in our study, indicates the presence of two outlier values. Specifically, within the figure, the black line within the box in the center of the figure indicates the median score for the variable *Age*. The space within the box itself indicates the middle (50%) of scores. The line (shaped like a T) above the box indicates the top 25% of scores and the line below the box indicates the bottom 25% of scores.

The circles with the numbers next to them above and below the figure and T-shaped lines are what SPSS has identified as outliers. The first outlier is well above the other scores (the circle with the number 1 next to it) and the second is well below the other scores (the circle with the number 2 next to it).

The first thing to note is that the numbers next to the circles correspond to the variable number in the light blue/gray column to the far left in the Data View portion of the SPSS database.

Specifically, when you are in the Data View portion of the SPSS database, you will see that all cases are numbered in a light blue/gray column on the left that starts at the top with the number one. The outlying cases are identified via this column. For example, since the software indicated that numbers 1 and 2 are outlying cases, to find these cases we need only go the columns in the Data View of the database that lists the variable age, and observe lines 1 and 2 as per the numbered column of the left to see what these values are. Subsequently, now you have identified the outliers.

How to Spot an Outlier: Using Standard Deviations

You can also identify outliers through noting the number of standard deviations a score is from the mean score. To do this, first calculate the mean and standard deviation of the variable being examined for outliers. To perform this test in SPSS:

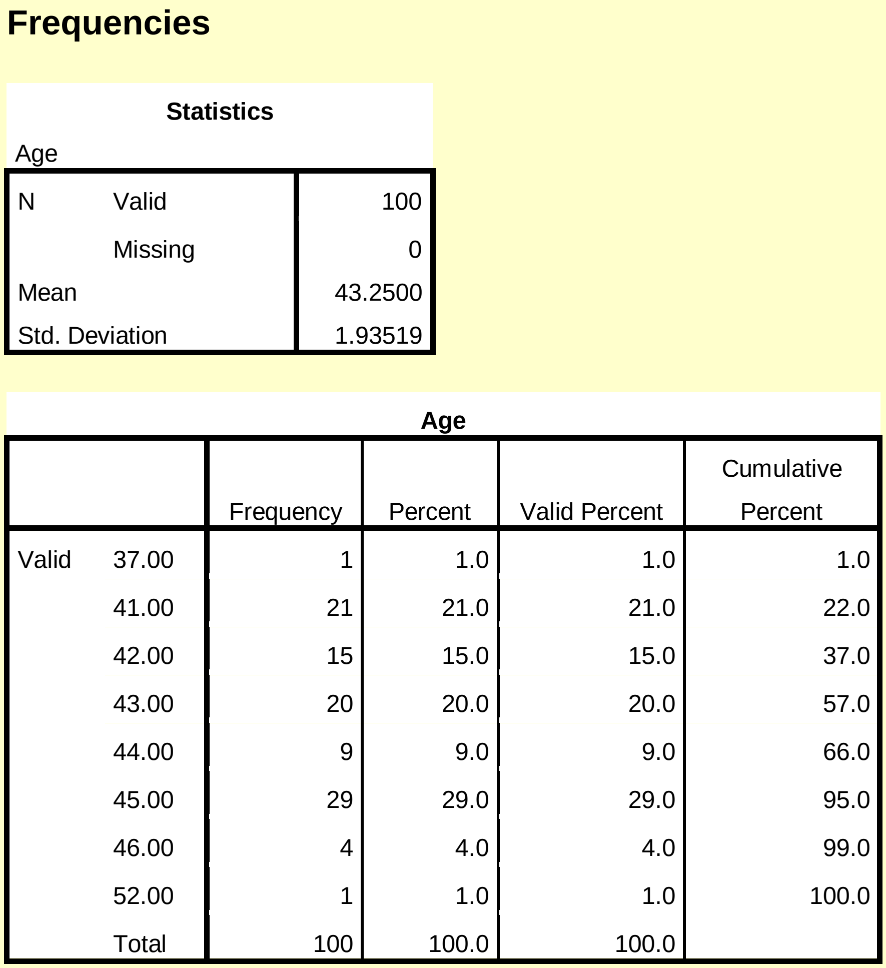
Click **Analyze** > **Descriptive Statistics** > **Frequencies...** on the top menu.

A dialogue box will then open that reads "**Frequencies**" in the upper left hand corner. Then click to highlight the variable in the left hand column that needs to be tested for normality and then transfer that variable into the left hand column into the "**Variable(s)**" box clicking the arrow button to the left of the "**Variable(s)**" box.

Next, click the button that reads "**Statistics**" in the upper right hand corner. In the new dialogue box, click the box next to the word "**Mean**" in the upper right hand corner and "**Standard Deviation**" in the lower left hand corner. Then click the "**Continue**" button. Then finally click the "**OK**" button in the bottom left hand corner of the dialogue box.

The output will contain two boxes. Please see Figure 2 below, which lists these two boxes using the study variable *Age* as an example.

Figure 2. SPSS Output Presenting the Mean, Standard Deviation, and Frequency Distribution for the Study Variable *Age*



Frequencies

Statistics

Age

N	Valid	100
	Missing	0
Mean		43.2500
Std. Deviation		1.93519

Age

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	37.00	1	1.0	1.0	1.0
	41.00	21	21.0	21.0	22.0
	42.00	15	15.0	15.0	37.0
	43.00	20	20.0	20.0	57.0
	44.00	9	9.0	9.0	66.0
	45.00	29	29.0	29.0	95.0
	46.00	4	4.0	4.0	99.0
	52.00	1	1.0	1.0	100.0
	Total	100	100.0	100.0	

How to Spot an Outlier: Using Standard Deviations (Cont.)

You may note the second box lists the distribution of scores, while the first box lists the number of cases along with the mean value and value for standard deviation. The mean value for the variable *Age* is reported as 43.25, while the value for the standard deviation is 1.93519. To identify which cases are outliers we must first decide how many standard deviations we will use to define a case as an outlier. Typically, a researcher will pick 3 standard deviations. Please recall in a normal distribution 99.7% of values will be between −3 and +3 standard deviations of the mean value. Thus, we are defining an outlier as a case outside of this interval. At times, a researcher may use an interval of −2 to +2 or −4 to +4 standard deviations depending on the distribution of scores. However, defining outliers as values outside −3 and +3 standard deviations is common.

How to Spot an Outlier: Using Standard Deviations (Cont.)

To compute this −3 to +3 interval, it is first necessary to multiply the value of the standard deviation by 3. In our example, the value of the standard deviation is 1.93519, which when multiplied by three produces a value of 5.80557. Recall our mean value for the variable *Age* is 43.25 years. Thus if we add three standard deviations to the mean we produce an upper bound limit of 49.05557 years (43.25 + 5.80577 = 49.05557). Furthermore, if we subtract three standard deviations to the mean we produce a lower bound limit of 37.44423 years (43.25 – 5.80577 = 37.44423). Subsequently, we will classify any values below 37.44423 years (−3 Standards Deviations) and above 49.05557 years (+3 Standards Deviations) as outliers.

If we again note the second table in Figure 2, we can see that there is one case in the distribution that is below this lower bound mark of 37.44423 years (i.e., the study participant that is 37 years of age) and one case in the distribution that is above this upper bound mark of 49.05557 years (i.e., the study participant that is 52 years of age).

It is interesting to note that the two cases identified as outliers via noting the number of standard deviations from the mean are the same two cases identified as outliers by the SPSS software using the boxplot method. This is also a common finding.

Creating a Version of the Variable that does not Contain the Outlier Values

Lastly, once the outliers have been identified within a variable, those outlier scores must be removed so that the new more normally distributed version (hopefully) of the variable can be used in statistical analysis. The simplest (and thus not the most sophisticated) way to do this is to cut and paste a copy of the variable next to original and then manually delete the outlier scores. To do this go to the:

Data View section of the SPSS database

At the top of the column, click on the name of the variable of interest (e.g., *Age*)

Press the right click button, then click “Insert Variable,” a new variable will appear

Right click on the variable to be copied (e.g., *Age*) and click copy

Then right click on the top of the newly created variable and click paste

Now rename the new variable, for example, *AgeNoOutliers*

Now you can manually remove the outlier scores from the copy of the original (but renamed) version of the variable being revised. Do this carefully. In the Data View portion of the SPSS database, make sure you note not only the number of the light blue/gray left hand column, but also the ID number of the cases with the outlier scores in the “ID” variable created within the database. It is important to note that if the cases are moved to different lines, through sorting cases, adding/removing cases, etc. the physical location of the cases will change in the database and the numbers in the light blue/gray left hand column will no longer correspond to the number used to identify the outlier scores. Thus, it is important to have another

Creating a Version of the Variable that does not Contain the Outlier Values (Continued)

indicator of the outliers score besides the numbers in the gray column to the left or the actual outlier score. Thus, note the outlier score and the ID number of that score. Then you can use the sort function (in the Data View portion of the SPSS database, right click on the name of the variable, then click sort ascending) to organize the scores, then scroll to and delete the outlier score, while noting the variable ID number to be sure you are deleting the appropriate case score.

Then, rerun the original and revised versions of the variables in a frequency function (Analyze, Descriptive statistics, Frequencies…). Next, compare the original and revised versions of the variable by each respective frequency distribution to verify that the revised version (e.g., *AgeNoOutlier*) of the variable does not contain the outlier scores that should still be within the original version (e.g., *Age*) of the variable. Subsequently, you should now have a version of the original non-normally distributed variable with the outliers removed to better approximate a normal variable distribution.

Of course, now you must go and test if the distribution of the revised variable does now approximate a normal distribution with the outlying cases excluded. The descriptions of how to test for the normality of a distribution are available in most texts and include the use of examining the values of skewness and kurtosis, as well as applying statistical tests such as the Kolmogorov-Smirnov and Shapiro-Wilk (both available in SPSS). This is obviously a topic for another newsletter.

Summary

The modification of any variable is a big compromise in the world or research and statistics. The point of research is to generalize from our sample to the larger population. When we begin to remove study participants from out sample that have outlying scores, we are removing members of the study population. Thus, in a way we are changing our sample and how the sample might relate to the study population. Subsequently, I prefer not to remove anyone from my sample unless it is necessary.

Although, situations where it is appropriate to remove outliers do exist. As we pointed out initially, one outlier can severely distort statistical findings. One outlier can make it seem like the whole sample is evidencing a relationship between the independent and dependent variable, when it is actually just that one outlying score. This is certainly not permissible.

However, often you are faced with a dilemma. Would you change the sample by removing study participants who are outliers or leave in the outliers and violate the assumption of normalicy for multiple regression?

Of course, the methods recommended on how to deal with this issue vary to extremes. For example, many researchers would say the correct way to deal with a non-normal distribution is to remove the outliers from the analysis and just use the

Summary (Continued)

transformed variables regardless of subsequent effects.

Still other researchers would say that eliminating study participants and changing variables distorts the sample. Also, many researchers point out that many statistical tests (such as multiple regression) are fairly robust, which means the procedure is not heavily impacted by violating test assumptions. Therefore, leaving the outliers in the analysis is just fine, even if the outliers make a distribution somewhat non-normal. However, I prefer a more middle of the road approach where both schools of thought described above are considered.

Specifically, I will perform the analysis two ways. Let's take multiple regression as an example. I will first perform the analysis with the inclusion of the original variable with all scores (including the outliers), even if the variable is not normally distributed. Next, I will repeat the same analysis which is identical in every way, except that I will remove the original version of the variable and enter version of the variable with the outliers removed.

Typically, the two versions of the analysis will be similar, in that either both are statistically significant or neither are statistically significant. Also, I pay close attention to how this variable, which is an independent variable, impacts the other independent variables in the model. In other words does the original or modified version of the variable effect whether or not other independent variables are associated with the dependent variable at a statistically significant level.

If the original variable has a different effect on the dependent variable, relative to the modified version of the variable, I will use the modified version of the variable (with the outliers removed). This is because if the original variable has a different effect on the dependent variable (or even other independent variables) relative to the modified version of the variable, it is likely that the non-normality of the distribution of scores might be responsible for the differential effect.

However, as I said, most of the time the two versions of the variable, i.e., original and modified, have very similar effects on the dependent variable, as well as the other independent variables. Subsequently, in that case it would be prudent to assume that the multiple regression procedure is responding in a robustness manner and that it is permissible to use the original variable. Certainly in my opinion it is preferable to use the original variable version as it represents more fidelity to the responses of the study participants.

In end, if you need to repeat the analysis with the original and modified versions of a variable, it would be smart to report this process. Specifically, there should be a separate section reporting the normality of variables in the Results section. Here a sentence could be added that describes that both versions of the variables were tested in analysis and yielded similar results (if they did of course).

Research Pathways Classified

Statistical Consultation and Analysis/Speeches and Seminars/Dissertation Services

William Bannon Associates, Inc., provides a wide array of services to professionals, faculty, and students engaged in research and both quantitative and qualitative analysis.

This newsletter, *Research Pathways*, is our flagship publication. It is but one way that we strive to contribute to the learning, growth, and excellence of individuals engaged in research and data analysis.

We also provide statistical consultation and analysis on research studies, as well as dissertation projects. A partial description of these services, as well as rates and testimonials can be seen on line at: www.williambannonassociates.org.

Dr. William M. Bannon, Jr., the president of the company, is also available for speaking engagements. Dr. Bannon has created a new seminar that presents techniques that make statistical methods more teachable, learnable, enjoyable, and accessible to the student, professional, and/or faculty member.

If you have any questions or requests, please feel free to email Dr. Bannon directly at: wb@williambannonassociates.org or contact him by phone at (718) 791-5329.

We look forward to hearing from you!

Personal Editor/Tutor/Coach

Published writer and experienced editor offers professional and creative editing, tutoring, and coaching assistance.

Credentials include:

- MFA in Creative Writing from The New School University
- Over 10 years experience in private editing, tutoring, and coaching creative and academic writers
- Personal editor to Stacey Patton, Simon & Schuster author of memoir, “That Mean Old Yesterday” (2007)
- Other editing clients include The New Yorker, The Feminist Press, Four Way Books, Salonika Magazine, etc.
- Taught writing instruction at The New School University, 14th Street Y in New York City, Berkeley College
- Certified proofreader
- Published poet

Rates are flexible/References available upon request

email Kathleen at kekky@mindspring.com

If you have questions regarding the material presented in this newsletter feel free to contact the staff at WBA, Inc. at the following email address: wb@williambannonassociates.org.

WBA, Inc. is headed by Dr. William Bannon, who is an Assistant Professor at the Mount Sinai School of Medicine, as well as the president of WBA, Inc.

Those of us at WBA, Inc. enjoy providing written materials, as well as private statistical and research consultation to clients. For further information, as well as archived copies of newsletters, visit us on the web at:
www.williambannonassociates.org