Fall 2011, Volume 1, Issue 3

The StatsWhisperer

The StatsWhisperer Newsletter is published by staff at StatsWhisperer.™ For many more free resources in learning statistics, including webinars and subscribing to this newsletter, visit us on the web at: www.statswhisperer.com

Introduction to this Issue

The process of conducting a statistical analysis often seems like a vast nebulous undertaking that often intimidates even the most daring individuals charged with this task. However, there are a few steps that are often common to this process.

Perhaps the greatest friend to the learner is the nuemonic phrase. Thus, I have devised such a companion for the person approaching the process of statistical analysis. The phrase is PUB Maker, which represents the steps, **P**reparation of data, **U**nivariate analysis, **B**ivariate analysis, and **M**ultivariate analysis. This seems an appropriate term as most researchers eventually want their analysis to be made into the basis of a publication.

These are the essential phases of most projects involving statistical analysis. One should bear in mind that there are a great many variations and diverse tasks incorporated within these steps. However, before one can approach the more finite nuances of statistical analysis, he or she might do well to study these basic steps.

Many sources of statistical instruction describe basic statistical tests and processes. However, few sources describe how to combine techniques to create a competent statistical analysis project. This issue of *Research Pathways* is designed to serve as a small source of instruction regarding how all statistical tests and processes may be combined to generate a publishable research study.

Inside This Issue on Addressing Non-Normal Data

P is for Preparation

Data preparation is vital to performing a brilliant statistical analysis project. Why? Because this is the phase where you assess, note, and describe the *quality of the materials* you will use in your statistical analysis.

How important is affirming that you are working with the right materials you need for a project? Well, I like to reference an old sports coach I knew in high school. He used to say, to paraphrase, that you could not make chicken salad out of chicken excrement. Clearly, he was right. You really can't begin a project with the wrong materials and produce the outcome you want. Likewise, you really can't begin a statistical analysis with flawed data and produce quality results.

Although the preparation of data is one of the most important phase of statistical analysis, it is also seemingly the most neglected phase of statistical analysis. This is often done to the point where this phase is barely mentioned or described in prestigious peer-reviewed journals. Thus,

P is for Preparation (continued)

perhaps one of the most obvious hallmarks of a well-grounded and professional statistical analysis is the presence of a thorough data preparation phase as a component of statistical analysis.

That being said, one might wonder, just what is the process of data preparation? Well, like most things in life, the process is specific to a particular project. However, there are a series of general steps that one should consider while approaching data preparation. A few are presented in a very abbreviated form here:

Step 1: Data Cleaning

Data cleaning refers to performing checks to best assure that data have been transferred accurately from the survey completed by the study participant to the software (e.g., SPSS, SAS, Excel) database that will be used to perform statistical analysis. Certainly, if there is a disconnect at this point, all steps in statistical analysis will examine useless data that does not reflect the study population.

The most obvious way to assure that all survey data has been entered into the software program accurately is to have someone sit down and compare the survey hard copies with the data entered into the software. However, this stepo is often too time consuming. As an alternative, many projects randomly select 10% of the survey hard copies and compare those raw data with what has been entered into the software for those surveys. If there are little or no errors found among this random selection of surveys, a researcher often assumes the data have been entered accurately. If there are many errors, perhaps all the surveys should be checked in the same manner.

Next, scores should be checked to be sure that all values reported fall within the legitimate maximum and minimum scores for that each variable. For example, let's assume the response to an item such as "Are you happy?" could range from 1=Very to 4=Not at all. Well if there is a numeric value of 5 entered in the software database for this item, we may have a problem.

Also, evidence of invalid responding should be examined. For example, if a study participant reports that they have never consumed alcohol in item 1, then reports that they drink a drink a day in item 2, then obviously the conflict indicates that one of the data points is innaccurate and needs to be addressed.

Tests for Mediating Effects

How does one test for the presence of a mediator variable through statistics? Over the past two decades one of the most common methods for testing for a mediator has been to follow the criteria set forth by Baron and Kenny (1986).

One of the benefits of this method is that it can be done using linear regression, which is a function of most statistical software, including SPSS and SAS. Thereby, specific advanced statistical software programs are not necessary. Please see the reference list for the fine article by Preacher & Hayes (2004) for instruction on SPSS and SAS procedures for estimating indirect effects in simple mediation models.

Essentially, the criteria for a mediated relationship set forth by Baron and Kenny states the following:

Step 1: Show that the independent variable is correlated with the outcome. Enter the independent variable in a linear regression equation with the dependent variable (estimate and test path c in Figure 1). This step establishes that there is a significant direct effect that may be mediated.

Step 2: Show that the independent variable is correlated with the mediator variable. Enter the independent variable in a linear regression equation with the mediator variable entered as the dependent variable (estimate and test path a in Figure 1). This step essentially involves treating the mediator as if it were an outcome variable.

Step 3: Show that the mediator affects the outcome variable. Enter the mediator varable as the independent variable in a linear regression equation with the dependent variable (estimate and test path b in Figure 1). It is not sufficient just to correlate the mediator with the outcome; the mediator and the outcome may be correlated because they are both caused by the independent variable. Thus, the independent variable must be controlled in establishing the effect of the mediator on the dependent variable.

Step 4: Establish that the mediator variable mediates the relationship between the independent and dependent variables. Enter the independent and mediator variables as 2 independent variables in a linear regression equation and regress them upon the dependent variable. If the effect between the independent and dependent variable becomes zero when the mediator is added we say that perfect mediation has occurred. If the effect between the independent and dependent variable becomes is significantly attenuated (e.g., reduced betas and p values) when the mediator is added, but not set to zero, we say that partial mediation has occurred.

Next, the statistical significance of the mediational relationship should be assessed. A clear and accessible method of accomplishing this procedure is to use the test proposed by Sobel (1982). This test of the indirect effect is given by dividing standard errors by the square root of variance and treating the ratio as a Z test. Please go to the excellent website (by Preacher and Leonardelli): http://www.people.ku.edu/~preacher/sobel/sobel.htm for a web page that can help you calculate this test simply by entering the (unstandardized) regression coefficients and standard errors from your output derivative of steps 1 to 4 described above.

Testing for Moderating Effects

A moderator variable is (also referred to as a test of interaction), in general terms, a variable that affects the direction and/or strength of the relation between dependent and independent variables (Kraemer, Wilson, Fairburn, & Agras, 2002). I have found that moderated effects are often best demonstrated through an example. In turn, let's approach how we may arrive at and conduct a test for moderating effects.

Suppose I have found a significant direct effect of interest to me, such as greater exposure to community stressors is related to increased anxiety among a sample of youth of African-American descent residing in an inner-city environment. I then conceive that an important cultural construct, such as *cultural pride* may impact this relationship. I may posit that higher levels of cultural pride should act as a buffer between youth exposure to community stressors and their resulting levels of anxiety. Thus, what we would say here is that cultural pride moderates the causal effect of exposure to community stressors on youth anxiety.

How would we test for such a relationship? Let us use SPSS, which is a standard statistical software program. We may use a Univariate General Linear Model to test this relationship. To do this, after opening your SPSS program, go to *Analyze*, then *General Linear Model*, and then *Univariate*.

To build our model, select youth anxiety and move it (via the arrows) to the dependent variable box. Next, select cultural pride and exposure to community stressors and move them into the fixed factor box.

Then select the box marked Model in the upper right hand corner. Within the new box that appears, in the upper right hard portion, select Custom. One at a time, highlight and move the variables cultural pride and exposure to community stressors over into to the right hand side of the box labeled Model.

Then highlight cultural pride and exposure to community stressors at the same time. Set the Build Term function between the boxes to interaction, and move the variables to the right. Click Continue, then once returned to the preceding box, click OK (or click paste to save the syntax, then run the syntax).

Your output should report a Tests of Between-Subjects Effects, where as a final factor the product term cultural pride*exposure to community stressors, is listed. The p value to the right will indicate if this term is significant. Specifically, the p value will indicate if cultural pride does or does not significantly moderate the relationship between exposure to community stressors and youth anxiety.

For the sake of our learning here, let's assume the moderating relationship is significant. Now we may wonder, what does that tell us? How does cultural pride impact the relationship between the independent and dependent variables? At this point, it seems that all we have is a relationship that statistical software has told us is significant, but no story to tell! It is time to take the final step, and not only identify, but express the moderated relationship.

What does the Moderator Say?

One of the simplest methods of depicting the relation between a moderator and the independent-dependent variable relationship it influences is to examine the extreme cases.

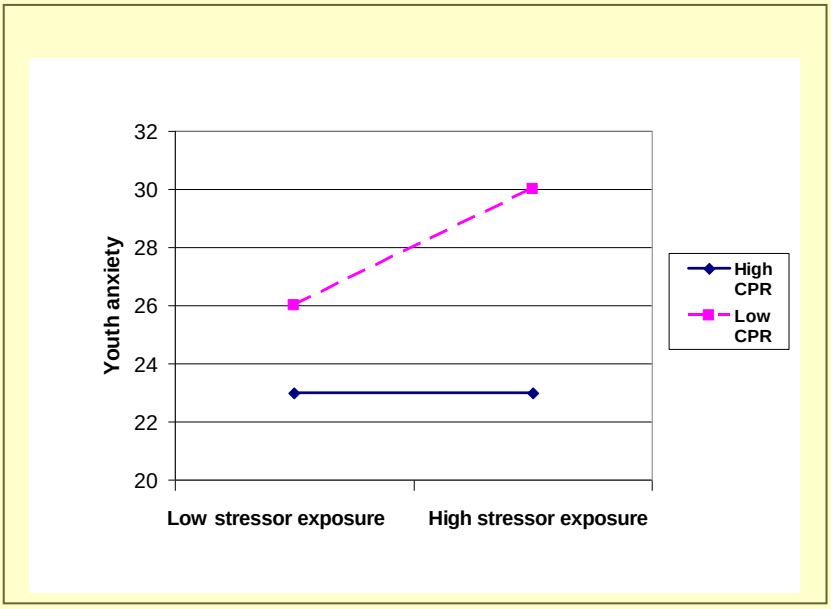
This is accomplished through creating two new variables out of the independent and moderator variables, and then plotting the new variables on a graph. For example, one could identify the cases one standard deviation outside of the mean for the independent variable and moderator variables. In each variable, the cases above the mean are labeled in a low category, and the cases above the mean are labeled in a high category. Then take the mean of each of the four categories and plot them on a graph to display the variable relationship.

For example, in our analysis we would plot the means for Low/High exposure to community stressors and Low/High Cultural Pride. The independent variable is put on the Y axis and the moderator displayed in categories. The dependent variable is put on the X axis. Please see Figure 2 for the final product expressing our relationship.

Now we have our story. We see for youth with high cultural pride, initially, in the presence of low stressors their mean level of anxiety is below that of youth with low cultural pride. Then as each group moves to the high stressor category, we see that youth with high cultural pride are unaffected, while youth with low cultural pride experience a significant increase in their mean level of anxiety scores. This supports out posit that cultural pride does seem to act as a buffer between exposure to community stressors and youth anxiety.

This is one example of a moderated relationship. As you progress in your learning you will enjoy learning several more!

Figure 2. The moderating effect of cultural pride (CPR) on the association of exposure to community stressors and child anxiety



Reference List

Baron, R. M. and Kenny, D. A. (1986) The Moderator-Mediator Variable Distinction in Social Psychological Research - Conceptual, Strategic, and Statistical Considerations, *Journal of Personality and Social Psychology, 51*(6), 1173-1182.

Kraemer H. C., Wilson G. T., Fairburn C. G., & Agras W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry, 59,* 877-883.

Preacher, K. J. and Hayes, A F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers, 36*, 717—731.

Sobel, M. E. (1982). Asymptotic intervals for indirect effects in structural equations models. In S. Leinhart (Ed.), *Sociological methodology 1982* (pp.290-312). San Francisco: Jossey-Bass.

This newsletter is published by William M. Bannon, Jr., Ph.D., as a resource for academics and students studying social science research.

Dr. Bannon is available to respond to any questions regarding the materials and/or articles presented. Feel free to call, (212) 933-1999, or Email him, wb@williambannonassociates.org, with any requests.

Dr. Bannon is an Assistant Professor at the Mount Sinai School of Medicine and the president of William Bannon Associates, Inc., which is a firm that enjoys providing statistical and research consultation to students and academics. Visit us on the web at: www.williambannonassociates.org

William M. Bannon, Jr., Ph.D.
125 West 72nd St., 3F
NY, NY 10023

Phone:
(212) 933-1999

Fax:
(646) 596-9610

E-Mail:
wb@williambannonassociates.org

William M. Bannon, Jr., Ph.D.
125 West 72nd Street, 3F
NY, NY 10023

University of Albany, School of Social Welfare
135 Western Avenue, Richardson Hall,
Albany, NY 12222