summer 2011, Volume 1, Issue 2

The StatsWhisperer™

The StatsWhisperer Newsletter is published by staff at StatsWhisperer. Visit us at: www.statswhisperer.com

Introduction to this Issue

The presence of missing data is a problem that plagues and befuddles many of us. For the data analyst, the challenges incurred through the presence of missing data can seem like an affliction. Fortunately, as is the case with most afflictions, there are a number of different treatments that may be applied as a remedy.

The goal of the current issue of this newsletter is to provide some small insight into the issues surrounding the treatment of missing data. While this discussion is in no way comprehensive, hopefully these pages can provide some foundational support to those approaching this challenge.

For learning purposes, I have presented a study that I recently completed, where the treatment of missing data was the primary challenge in statistical analysis. This study is a pretest to posttest child mental health intervention where approximately half of the 386 youth that completed measures at pretest, were missing scores at posttest.

Our challenge was identify and employ a credible method toward figuring out if the intervention was effective while missing posttest scores for half of the treatment group (the data presented here are proxy numbers, as the findings of the actual study have not yet been released).

Inside This Issue

The Study Presented as an Example

The study used here to illustrate how the missing data were treated was a pretest to posttest school-based intervention program designed to treat youth for general internalizing and externalizing symptoms.
There were 386 youth in the treatment group at pretest. Data regarding youth mental health were gathered upon these youth via the child, teacher, and parent reports of the Strengths and Difficulties Questionnaire (SDQ), which were completed at pretest and posttest. At posttest, data were available for 205 of these youth.

There were approximately 11,000 youth in the control arm of the study that received no intervention. Data regarding youth mental health were gathered upon these youth via the child report of the SDQ, which was completed at pretest and posttest. The issue of missing data was not a concern for youth in the control group as a very small number were missing posttest data.

Accounting for Missing Data

In order to account for the missing data points involved in this analysis, several steps needed to be taken in order to produce valid inferences regarding the pretest to posttest scores on study outcome measures. Essential, we needed to first:

1) Examine the data for nonresponse bias;
2) Consider why values may be missing; and
3) Use of a specialized statistical method in analysis, which can incorporate uncertainty into statistical parameters in relation to the missing data.

These steps are described here as each was applied to the sample study. Clearly, as methods and results vary from study to study, the directions each step took us will not be the same in all other studies. However, we will try to provide some insight in the case of alternate results.
It is important to note that I performed steps one through three several times with the data structured in different ways. For example, I examined SDQ variables in a continuous form, as well as dichotomized into categories that reflected high vs. low/some needs. I also performed the analysis using the whole sample, as well as only youth that indicated high needs youth at pretest. This was done so I could get a comprehensive view of the data.

In this article, I report an analysis of pretest to posttest changes regarding mean value SDQ scores among youth that indicated a high level of mental health need at pretest (SDQ score $\geq$ 20). This goal of this analysis is to examine how the intervention program impacted the youth that entered the study with the most serious levels of mental health need.

An Examination of Data for Non-Response Bias

When considering missing data, it must be determined if observations with missing values are systematically different from observations with observed values. If such a difference exists, bias can be easily introduced into parameter estimates.

For example, in the current study, it is possible that within the treatment group, youth with high levels of mental health need at pretest may have been more likely to drop out of the study and not be represented at posttest. In that case, there would have been lower levels of youth mental health need at posttest because the high needs cases were not present. This is a big confounder! Therefore, we employed a careful analysis (through the use of chi-squares and t-tests) among youth in the treatment group regarding if youth that completed and did not complete the intervention were significantly different in terms of by demographic characteristics, as well as by levels of mental health need at pretest as reported on the SDQ measure by multiple informants (i.e., child, parent, teacher).

Please see Table 1 (below) for a description of these findings. Chi-square analysis indicated no statistically significant differences between youth that did and did not complete the program by demographic characteristics (e.g., gender) or youth mental health need (high vs. low/some). For example, as per SDQ Score analysis, reports provided by children, parents, and teachers indicated that among children with high need at pretest, about half of children were completers and non-completers.

Table 1
Demographic and Mental Health Need Characteristics of Youth within the Intervention Group Gathered at Pretest Stratified by Program Completers and Non-Completers

	Completers		Non-Completers		Total	
Variable	***n***	**%**	***n***	**%**	***n***	**%**
Demographics (e.g., Gender)						
Gender						
Male	124	53.9%	106	46.1%	230	100.0%
Female	79	55.2%	64	44.8%	143	100.0%
Mental Health Difficulties						
SDQ Score - Child						
High Need	63	56.8%	48	43.2%	111	100.0%
Low/Some Need	140	55.8%	111	44.2%	251	100.0%
SDQ Score - Parent						
High Need	83	52.5%	75	47.5%	158	100.0%
Low/Some Need	112	53.1%	99	46.9%	211	100.0%
SDQ Score – Teacher						
High Need	63	49.2%	65	50.8%	128	100.0%
Low/Some Need	120	58.8%	84	41.2%	204	100.0%

Consider Why Values Are Missing

Second, there is a need to consider the reasons certain observations were missing at posttest. For this study, it is important to note that we had the benefit of reports of the reasons most of the youth in the intervention group had missing data points at posttest.

Specifically, about half of the youth in the treatment group that were missing data did not have the opportunity to provide data at posttest because of the expiration of the school year. Subsequently, project staff members were not able to gather data from them. Thus, it is reasonable to speculate that this percentage of missing youth did not differ from the youth that completed posttest measures. The real difference was that one group participated in the intervention program at a juncture where the end of the school year precluded them from providing posttest data.

Of the remaining youth, most did not provide posttest data because they were suspended or expelled.

In any case, the knowledge of why at least a portion of youth did not provide data is a significant benefit in accounting for missing values.

Consider this example where we know that about half the missing posttest scores are related to the timing of when the surveys were administered and not the traits of the subject. We might say, “The timing was the reason, oh that could happen to anybody.” This realization is exactly what supports the premise that the differences between completers and non-completers may not be significant.

A Specialized Statistical Method

After making careful considerations of nonresponse bias, as well as why values are missing, the fact remains the presence of missing data means that all the values in the dataset are not known. Subsequently, there is uncertainty regarding what these values truly are and this uncertainty needs to be incorporated into the standard errors of parameter estimates so that the respective confidence intervals are not overly precise.

The current analysis employs Mixed-effects regression models (MRMs; Gibbons, Hedeker, & DuToit, 2010) as a means of providing this type of flexible framework for analyzing longitudinal data with missing values. MRMs make use of all observed values through incorporating all available data for a subject within statistical analysis, thereby providing valid inferences in the presence of missing data.

Specifically, random coefficient modeling was performed on key outcomes using the SuperMix program for mixed effects regression models which uses maximum likelihood estimation to model measurements over time within cases. This form of modeling, also known as hierarchical linear modeling or multilevel linear modeling, allows parameters (intercepts and slopes) for measurements over time within cases to vary between cases, while accurately accounting for correlation between measurements within cases.

This procedure also allows for different times and numbers of measurements within people. Therefore, it is an appropriate method to model longitudinal change involving data where there is attrition over time. It is critical that one realizes that this procedure requires making the assumption that the missing data is ignorable (i.e., at least missing at random - MAR). Based upon our considerations of nonresponse bias, as well as why values are missing, this seems to be a reasonable assumption within these data. If data were not missing at random other methods would be considered (see Siddique, Brown, Hedeker, Duan, Gibbons, Miranda, & Lavori, 2008).

Please note that I prepared the data is SPSS and then loaded them into the Supermix software program for multivariate analysis. When data are in SPSS, they must be transposed in order to make the pretest and posttest scores into one new outcome variable (i.e., the scores are now sorted into one common column instead of two).

In order to identify scores over time, a new variable is computed that indicates which score within the new outcome variable is a pretest or posttest score. This new variable can also be used to indicate the variable time (pretest to posttest), which is needed for analysis in the Supermix program. The first analysis I wanted perform was a test for differences between the treatment group vs. control group over time (i.e., from pretest to posttest). Keep in mind, the only control group scores available were child reports on the SDQ.

In order to test this relationship, while in SPSS, I made a variable that indicated if a child was in the treatment or control group. Next, I created a new variable, which was a product term where condition (treatment vs. control) was multiplied by time.

A Specialized Statistical Method (continued)

This is easy to do. In SPSS the process involves using the *transform* function to multiply two variables (here it is group by time) in order to create a third.

Next, the data were transported into the Supermix software program to test for significant differences between pretest to posttest mean value changes in SDQ scores (child reported) between the treatment group vs. control group.

For the parent and teacher reports on the SDQ, there were no control group subject data. Subsequently, changed were just examined over time and not time by condition.

Once in Supermix, randomly varying intercepts were allowed in each model, which also enabled calculation of the intra-class correlation coefficient (ICC), a measure of the percentage of variance between people compared to the total variance.

Then while in Supermix, I selected the youth with scores at pretest that indicated a high level of mental health need. In order to examine the child level scores, I examined changes in continuous SDQ scores from pretest to posttest as a function of the product term time by group. I repeated my pretest to posttest analysis for changes in the parent and teacher scores over time using only time as a predictor. Table 2 (below) presents the results of the model examining changes over time by study group based on child level data.

In Table 2, the *intercept* indicates the mean value score at baseline for one group (in this table it is the Treatment group). If it is significant, it means the pretest scores are significantly different. The line for *Group by Time* indicates the difference between the changes between groups over time. In the table below, a significant effect is indicated. Subsequently, this model indicates that after considering the uncertainty presented by missing data, among youth with high mental health needs at pretest, the treatment group evidenced a significantly higher mean value score. However, this group also evidenced a significantly higher reduction in SDQ scores over time relative to the control group.

The models using parent and teacher data also reflected significant reductions in child SDQ scores from pretest to posttest for the treatment group.

Table 2. Mean Value Changes in SDQ Scores from Pretest to Posttest (child level data)

Variable	Estimate	SE	Z-value	P-value
SDQ Score - Child				
Intercept	23.7027	0.3634	65.2190	0.0000
Time	-7.7718	0.5656	-13.7418	0.0000
Group	-1.1033	0.3833	-2.8781	0.0040
Group by Time	3.4258	0.6048	5.6643	0.0000

A Specialized Statistical Method (continued)

When I describe what I did in this study, most people can follow along just fine as I describe the examination for non-response bias and considering why values are missing. However, my description of the specialized statistical procedure I used is often not greeted with such enthusiasm. Subsequently, I thought it may be useful to mention some other methods one may employ toward accounting for missing values that can be done in SPSS and are not too complicated.

I hesitate to mention entering the mean variable value as the value for the missing cell. Often data are missing because they vary significantly from the mean. For example, when a study participant does not report their income, their income is often high above or below the mean income, which may be the reason they do not wish to report the number. Thus, entering the mean value may throw the analysis way off. Yet, a famous statistician once told me, if you have just a few empty cells in a large dataset, entering the mean response is okay. Hence, we can consider this method, but very sparingly.

A more respected, practical, and reasonable method to generate values when faced with missing data is to use the mean value of the data that are present in a scale.

For example, suppose a scale is comprised of 10 questions that are answered along a 1-5 point Likert-type scale. Then suppose that one study participant responds to 8 questions and then leaves the other two questions blank. A possible solution would be to take the mean value of the 8 questions where data were provided and use the value as the total scale score. Through this method, you would be arriving at a total scale score using only the data provided by the study participant, in light of missing data. Of course there are some measurement sacrifices involved, such as the data that would have been captured by the two missing items are not represented in the total score.

If one employs this method, they should have a cutoff point for missing data. Most analysts would not include a case unless at least 75% of the data are present. Thus, in our example above, with data present for 8 of 10 questions, 80% of the data are provided. If 3 of 10 questions were left unanswered, only 70% of the data would be present. Hence, the level of data would be below the required 75% mark and the method of taking the average of the data that are present would not be appropriate.

Generating the mean value for each case is not complicated. There is a function in SPSS that will take the mean value for the specified group of items at: *Transform*, *Compute Variable*, under function group pick *Statistical*, then under functions pick *Mean*, use the arrow box to put it under *numeric expression*, enter the items to be averaged in the parentheses, then name the *Target Variable*. All that remains is to click ok or better yet, paste and save the syntax before you run it.

Final Comments

It is important to note that this analysis actually presented few challenges after the pattern of missing data were analyzed.

For example, if these data were not missing at random or there was non-response bias, these developments would have needed to have been accounted for. However, for every challenge there are solutions.

However, the best way to handle missing data is not to have any! That is the best advice regarding missing data that I have ever heard!

Feel free to visit statswhisperer on the web at:

www.statswhisperer.com

Dr. William M. O'Bannon, Jr., the founder and CEO can be contacted at:

wb@statswhisperer.com

Information on our upcoming seminars can be viewed at:

http://www.statswhisperer.com/seminars/

Thanks for your interest in our newsletter!

REFERENCES

Gibbons, R., Hedeker, D., & DuToit, S. (2010). Advances in analysis of longitudinal data. Annual Review of Clinical Psychology. Vol.6 2010, pp. 79-107.

Siddique, J., Brown, C. H., Hedeker, D., Duan, N., Gibbons, R., Miranda, J., & Lavori, P. (2008). Missing data in longitudinal clinical trials - Part B, Analytical issues. Psychiatric Annals. Vol.38(12), Dec 2008, pp. 793-801.