Know where you are going before you get there: Statistical Action Plans work.

Dr. Jon Starkweather, Research and Statistical Support consultant.

Clients often come to RSS wondering what analysis they should do or what analysis they should do next. Clients are often looking for some remedy, fix, or contingency because they have realized their data, for one reason or another, do not meet the assumptions of the analysis they were expecting to conduct. Many of these clients have successfully proposed the research to grant committees, colleagues, or dissertation / thesis committees. However, the proposed analysis or analyses were chosen based on what the researcher or student knew at the time (prior to proposing the study). Often those analyses contain assumptions which are not carefully considered and upon collection of data it is realized the data do not meet those assumptions. The problem becomes apparent only after data has been collected and the researcher realizes they are going to need to learn an analysis unfamiliar to them while under some publication or dissertation / thesis defense deadline.

It is for these reasons and occasions that we (RSS) generally recommend creating a Statistical Action Plan (SAP) prior to proposal of the study. The current article describes the SAP general format and offers suggestions for why it can be helpful. The SAP is nothing more than a document which contains the planned analyses and the order in which they are likely to be conducted. It should also contain alternative analysis in case the data is discovered to violate the assumptions of a planned traditional analysis. It is both a reference to turn to for the analyses to answer the research questions (i.e. hypotheses) and a way to plan contingencies in case the data do not meet standard parametric assumptions. Furthermore, it offers student researchers a guide to analyses they may not be familiar with and need to familiarize themselves with prior to proposal or data collection. Along those same lines, the SAP should include which software will be used to conduct the planned analyses; especially considering the limitations of some software packages (Starkweather, 2013a).

SAP Format

Generally speaking, quantitative data analysis follows a four stage process. The first stage in the process is *Initial Data Analysis* (IDA). During this stage it is critical to know thy data, become intimately familiar with it; so decisions about subsequent analysis can be made appropriately. Generally the focus of this stage is on univariate descriptive statistics and associated plots. Some of the most important operations during IDA include some of the most mundane tasks of a data analyst. Simple operations such as creating a frequency distribution and the appropriate graph for every variable (e.g. bar charts for categorical variables, histograms for continuous or nearly continuous variables). Here the focus should be on the shape of the distributions of each variable (e.g. Are the continuous variables normally distributed? Are the categorical variables evenly balanced, or do you have severely unbalanced categories? Do your survey items display floor or ceiling effects?). Other necessary steps in this process follow from those simple charts and graphs. Inspection of data entry errors, missing values, and outliers will need to be completed – and these should be easy to identify from these simple charts and graphs. Data entry errors need to be corrected by reviewing and comparing the actual data collection materials (e.g. recording devices, surveys and responses, etc.) to the electronic data file values. For example, if your dataset contains a gender or sex variable and you create a bar chart with the bars representing the gender of each participant (male and female), then if one participant has a value other than those two you know you have some sort of data entry error, coding error, or a missing value. Missing values need to be investigated further when identified (Little & Rubin, 1987). What percentage of the data matrix (i.e. number of rows multiplied times number of columns) is missing? A determination must be made as to whether the missing values are missing at random (MAR) or not missing at random. In other words, MAR means; given the observed data, the missingness mechanism does not depend on the unobserved data. Once that determination has been made, an imputation strategy can be decided upon. If the values are missing at random, then there are multitudes of missing value imputation procedures available for single or multiple imputation (e.g. random recursive partitioning, maximum likelihood imputation, sequential nearest neighbors' imputation, etc.; see Starkweather, 2014 and Starkweather, 2010). If the values are not missing at random, then some strategy must be devised to account for the pattern of missing while imputing those values; in other words, there must be a model estimated which controls for the relationships among the variables and imputes values estimated to contain the least bias.

The second stage generally involves *preliminary data analysis*. This stage is primarily concerned with assumption checking and making sure you measured what you think you measured. Bivariate linearity, multivariate normality, and multivariate outliers should be assessed. Again, a good place to start is with relatively simple tables and graphs, such as scatterplots and scatterplot matrices along with associated correlations and correlation matrices. Keep in mind, there is more than one type of correlation and the type used is largely determined by the type of variables being correlated (e.g. Pearson product moment correlation, Spearman's rho, Kendall's tau, point biserial correlation, polychoric correlation, tetrachoric correlation, etc.). One goal at this stage is to understand the nature of the relationships among the variables – not just the variables of most interest to the hypotheses, but also the auxiliary, confounding, demographic and any other type of variables as well. It is important to realize there are three key aspects of any relationship among two (or more) variables: significance (which can be meaningless with large sample sizes), direction (i.e. positive, negative), and magnitude. The level of magnitude which indicates importance varies with each study and / or field, as does the effect size (e.g. R^2 or percentage of variance shared / accounted for). Obviously, this highlights the importance of a thorough literature review and becoming familiar with acceptable effect sizes within the field or subject of study. Similarly, different fields often use (or are only familiar with) different metrics; for example some fields rely upon (and expect) Mahalanobis' distance for assessing multivariate normality and multivariate outliers (Starkweather, 2013b); others might choose Cook's distance or some other measure of multivariate distance or leverage (also called influence).

As mentioned briefly at the beginning of the previous paragraph, this stage (second stage) may also involve more complex analysis such as Item Response Theory (IRT) or factor analysis to make sure you measured the variables of interest appropriately. This step is required if you are using a survey to assess the primary variables of interest. Does the factor structure (or item difficulty, item discrimination, etc.) of your sample conform to what has been established of the items as reported in the literature? Also, in regards to surveys; do not use Cronbach's Alpha unless your data meet the assumptions associated with it, most survey data does not and there are more appropriate statistics to use for reliability (Starkweather, 2012b). Another goal of the second stage in complex research designs might be variable or model selection analysis. For example, if the study is primarily exploratory and involves collection of extremely large data (e.g. genetics); then a variable or model selection technique might be used to reduce the variables down to a set containing the most important variables (e.g. Relative Importance, Bayesian Model Selection; see Starkweather, 2011). This second stage might also contain strategies for developing weights in order to correct for imbalances in the data or to statistically control confounding variables. Weighting strategies (e.g. propensity scores) should not be avoided, they are often very effective (see Kish, 1990); but choosing the *right* weights is essential.

The third stage of the data analysis process generally involves the *primary data analysis*; this is the stage in which the major analyses required to answer the hypothesis or hypotheses of the study are conducted. This is the stage in which the theoretical model is fit to the data – that model may be something as simple as a factorial Analysis of Variance (ANOVA) or it may be very complex, such as a Structural Equation Model (SEM). The main goal of this stage is to determine if the data and model fit well. There are often many measures of model fit (e.g. Root Mean Square Error of Approximation [RMSEA], Normed Fit Index [NFI], Non-Normed Fit Index [NNFI], Akian Information Criterion [AIC], Bayesian Information Criterion [BIC], etc.). Therefore, it is again important to have completed a thorough literature review in order to understand what represents appropriate fit in your discipline. Keep in mind; whenever fitting a model is required or hypothesized, it is generally a good idea to fit some competing models in order to give goodness-of-fit metrics some context. In other words, if you are hypothesizing one model, you had better fit at least one (and probably two) more models in order to have some empirical evidence for the model you hypothesized (being the *best* model). Something else to keep in mind at this stage – with respect to goodness-of-fit measures – a chi-square statistic is virtually meaningless when fitting an even moderately complex model. This is because even moderately complex models require large sample sizes and as everyone knows, chi-square becomes less and less meaningful as sample size increases. Last, but not least, this stage should include extensive evaluation of residual values. All models are capable of producing residuals – the difference between the actual values of the data and the predicted values of the model (e.g. Y minus Y-hat or predicted Y [y – ŷ] in regression, matrix of association minus the reproduced matrix or predicted matrix in many multivariate analyses, etc.). One assumption of common parametric analyses is normally distributed residual values – obviously this cannot be checked until the model has been fit; and should be checked carefully.

The fourth stage of the process involves *secondary* or *subsequent data analysis*. This stage involves analyses for testing secondary hypotheses or individual hypotheses nested within, or of lesser importance than, the larger goal (hypothesis) of the study. Consider something as simple as a one-way ANOVA. The model would consist of two variables: one categorical variable with more than two categories (often called an independent variable), predicting one continuous or nearly continuous outcome variable (often called a dependent variable). Evaluating the effect of the independent variable on the dependent variable would entail interpretation of the omnibus *F* (the main effect) which would inform whether or not the model fit well and a main effect was present. However, the *F* test does not inform where the significant differences lie. In order to identify which group or groups were significantly (and meaningfully – with effect sizes) different from which other group or groups, planned contrasts or post hoc testing would be necessary. These planned contrasts or post hoc tests would be done as the fourth stage of the process or SAP. In a regression setting, the model fit is evaluated with a combination of R^2 type of statistics (and often an ANOVA summary type table with an *F* test) while the simple effects or fourth stage is done with the individual predictor coefficients' (often with *t* tests for each predictor's standardized coefficient). In more complex settings, such as path models or SEM, the fit of the model is evaluated in the third stage and the individual path coefficients or structure coefficients are evaluated in the fourth stage (often with t-tests of the standardized coefficients). The fourth stage might also be the stage in which confounding variables are controlled or mediation and / or moderation are evaluated. This stage may also include post-stratification, as in multilevel regression (also known as Hierarchical Linear Modeling [HLM]) with post-stratification.

Conclusions

Obviously, the main idea of this article was to help researchers, primarily graduate students, better prepare for data collection. It is important to note that although the stages of a Statistical Action Plan are listed and described above as sequential, a researcher may need to return to previous steps throughout the process. Again, this is one of the benefits of forcing one's self to create such a plan – it necessitates thinking about what type of data is needed to answer the research question or specific hypotheses and it motivates consideration of alternative analyses as a contingency if the resulting data does not conform to the assumptions of the planned primary analysis strategy. As many people have recognized over the historical course of science, more effort spent in planning research pays substantial benefits as the study is conducted and analyzed. In essence, it is much better to plan potential contingencies and learn about them (i.e. unfamiliar analysis) prior to data collection than it is after data collection and one is facing a thesis / dissertation defense or publication deadline. Lastly, an Adobe.pdf version of this article can be found here (along with several of the resources listed below). Other, potentially useful resources are also located here.

Until next time; "*a failure to plan at the beginning [of the semester] on* your part *does not represent a crisis at the end [of the semester] on* my part*.*" – Kevin J. Armstrong, PhD.

References / Resources

Kish, L. (1990). Weighting: Why, when, and how? Paper presented at the Proceedings of the Survey Research Methods section of the American Statistical Association. Available at: https://www.amstat.org/sections/SRMS/Proceedings/papers/1990_018.pdf

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: John Wiley & Sons. [Still one of the most important resources for understanding missing values].

Starkweather, J. (2014). Your one-stop multiple missing value imputation shop: R 2.15.0 with the rrp package. Benchmarks Online, February 2014. Available at: http://it.unt.edu/benchmarks/issues/2014/02/rss-matters

Starkweather, J. (2013a). Why R; it's not a question, it's an answer. Benchmarks Online, October 2013. Available at:

http://it.unt.edu/benchmarks/issues/2013/10/rss-matters

Starkweather, J. (2013b). Multivariate outlier detection with Mahalanobis' distance. Benchmarks Online, July 2013. Available at:
http://it.unt.edu/benchmarks/issues/2013/07/rss-matters

Starkweather, J. (2012a). Statistical Resources (updated). Benchmarks Online, July 2012. Available at:
http://it.unt.edu/benchmarks/issues/2012/07/rss-matters

Starkweather, J. (2012b). Step out of the past: Stop using coefficient alpha; there are better ways to calculate reliability. Benchmarks Online, June 2012. Available at:
http://it.unt.edu/benchmarks/issues/2012/06/rss-matters

Starkweather, J. (2011). Sharpening Occam's Razor: Using Bayesian Model Averaging in R to Separate the Wheat from the Chaff. Benchmarks Online, February 2011. Available at:
http://it.unt.edu/benchmarks/issues/2011/02/rss-matters

Starkweather, J. (2010). How to identify and impute multiple missing values using R. Benchmarks Online, November 2010. Available at:
http://web3.unt.edu/benchmarks/issues/2010/11/rss-matters