

How to Conduct Empirical Academic Research: A (very) General Guide

As published in Benchmarks RSS Matters, December 2011
<http://web3.unt.edu/benchmarks/issues/2011/12/rss-matters>

Jon Starkweather, PhD

Jon Starkweather, PhD
jonathan.starkweather@unt.edu
Consultant
Research and Statistical Support



<http://www.unt.edu>



<http://www.unt.edu/rss>

RSS hosts a number of “Short Courses”.
A list of them is available at:
<http://www.unt.edu/rss/Instructional.htm>

How to Conduct Empirical Academic Research: A (very) General Guide

This month's article was motivated by an interaction with a student who reported being "stuck" on their dissertation and not knowing what to do next. A dissertation, or thesis for that matter, is not an immovable object; nor is graduation an unattainable goal. The author of this article was reminded that some students spend a year or more between the completion of research methods and/or statistics courses and the beginning of work on their dissertation. Recognition of the phenomena known as being stuck and the often lengthy time between methods/statistics courses and dissertation work motivated the writing of this article. The article is meant to provide a very general framework, or guide, to the process of conducting empirical research (specifically a dissertation or thesis). Keep in mind, if this process were extremely easy (and free), everyone would have an advanced degree. A meaningful and successful study takes a great deal of effort, time, and resources (e.g. caffeine, money, etc.).

Before Data Collection

There are several key ingredients which must be present in order for a meaningful study to be completed. The first, of course, is *thought*. A necessary step in conducting a study is careful, critical, and repeated thought about what will be accomplished, why it is meaningful to accomplish it, and how it will be accomplished. Choosing a research topic is also a critical decision in the process and should not be taken lightly. If one chooses a topic in which one has no personal interest (i.e. intrinsic motivation), then one is unlikely to be able to muster the self-discipline to work on the research when distractions are present. Equally important is choosing the scale or scope of the research. Passion is a great asset because it motivates work, but passion can also lead to an overly ambitious study (i.e. one which cannot possibly be completed in the allotted time frame). When choosing a research topic, make sure it meets the approval of any and all collaborators (e.g. a dissertation advisor). Peers and advisors are invaluable resources during the entirety of the research process; they can often point out advantages and disadvantages for you. Do not be afraid to ask others the question: "Am I making sense?" The answer can only improve your project or increase your confidence.

Once a general area of interest, or topic, is decided upon; a thorough review of the literature should be conducted. Generally, the word 'literature' in this context refers to peer reviewed academic journal articles; with specific emphasis on empirical studies. Where should you look to find this literature, who could you consult? If you are working on a dissertation, your advisor should be familiar with the resources you will need to turn to (e.g. what journals, electronic databases, societies/associations are likely to be oriented toward your topic). Also, remember library professionals (e.g. reference librarians) are

Side Note 1: Choosing a Dissertation Advisor.

When choosing a dissertation advisor, make sure that person's research interests are well matched with your own. It is preferred that your dissertation advisor be at least familiar with, if not an expert on, the domain in which you wish to conduct your study. If a prospective advisor has been doing research on the mating habits of the Great Blue Heron and you are interested in conducting research into the thermodynamics of the Gulf Stream current then you might not get the support or advice you will likely need. Occasionally, you may also want to consider the values and beliefs of a prospective dissertation advisor. If a prospective advisor has been doing externally funded research on the efficient extraction of petroleum and natural gas reserves for the last 20 years and you are interested in conducting a study of the impacts of hydraulic fracturing on drinking water then you may not get the support or advice you will need and perhaps you should choose a different advisor.

experts you can contact to learning how and where to search for information. Becoming familiar with the literature will acquaint you with the concepts, terms, measures/instruments, methods, and results related to your chosen topic. Becoming familiar with the research which has been completed on, or around, the topic will also allow you to transition from an area of interest to a *research question*. The research question should be just that; a question, stated in lay terms (i.e. even people not associated with your topic, or even your field, should be able to understand the question). The research question should be constructed in such a way that the research you conduct should answer that question. For example, do animals raised in zoos suffer negative health effects due to lack of exercise or predation?

The research question should then flow naturally into formal statement of hypotheses. Again, concerted effort (i.e. thought) should be expended on developing the hypotheses. Often collaboration is involved in the development of hypotheses. Hypotheses should focus on the strength and direction of expected effects. Keep in mind; formal hypotheses should not to be confused with null and alternative hypotheses. Formal hypotheses should be concise sentences which convey expected findings; for example, one might hypothesize that animals raised in zoos have on average significantly greater body weight than similarly aged animals of the same species which were raised in the wild. Generally, a meaningful research project will have multiple formal hypotheses. Often they are structured hierarchically; meaning a central thesis is conveyed in a main effects hypothesis and subordinate hypotheses are used for more narrow or lower level effects of interest.

Once formal hypotheses have been constructed, the research design can be attacked. Research design includes determining how variables will be measured, what instruments (if any are necessary) will be used, will you develop your own instruments or use existing ones, will random sampling (and/or random assignment) be employed, what procedures will be followed (pretest – post test; experimental manipulation, etc.), how will internal and external validity be achieved, etc. in order to gather the data *necessary* to test the hypotheses. In this context, the word ‘necessary’ refers to both the *amount* of data and the *appropriateness* of the data. The ‘amount’ of data determines the power of the study and is commonly constrained by practical concerns such as time and funding. However, many applications are available (e.g. G^*Power^3 ¹) for determining a priori sample size for a given design, desired power, and desired effect size. The ‘appropriateness’ of the data has two meanings. First, obviously you need to collect data which will be meaningful for answering your hypotheses; for example you are not going to measure animals’ weight with a thermometer. Second, ‘appropriateness’ refers to whether or not the data will adhere to the assumptions of a given analysis. For instance, a simple independent *t*-test (which is typically used to evaluate mean differences) requires a categorical (i.e. factor) variable (e.g. animal sex; male or female) and a continuous or nearly continuous (i.e. numeric) variable (e.g. adult weight; kilograms). As another example, consider studying the effects of chemotherapy on hair loss. Here, you would find it beneficial to collect hair loss data by measuring the number of hairs per square inch of scalp, rather than simply rating hair loss as extreme, moderate, or slight (i.e. scale of

Side Note 2: New Data vs. Archival Data.

New data in this context is defined as data you collect. Archival data is defined as existing data which someone else collected. There are benefits and costs associated with each. Generally, the main benefit of using archival data is that of time. The time associated with collecting archival data is drastically lower than the time associated with collecting new data. The primary benefit of collecting new data is control; meaning, you will have control over what is collected (i.e. how variables are measured and what the measurements represent). It is the opinion of this author that students conducting a dissertation should collect their own data and not rely upon archival data. Often, archival data is like the carrion of the research world, it has been picked over for years and likely has no meaningful effects left in it *undiscovered*.

¹<http://www.psych.uni-duesseldorf.de/abteilungen/aap/gpower3/>

measurement is important). Clearly, there is a relationship between formal hypotheses, research design, and types of analysis. However, keep in mind; the data may not conform to expectations, which means the initial analysis chosen may not be the analysis most appropriate once the data has been collected. Therefore, again, careful thought and collaboration should be exercised during the consideration of design and choice of primary analysis, secondary analysis, and possibly alternative analytic techniques in case the data does not conform to assumptions (e.g. linearity). It is often the case that a particular hypothesis and data combination can be addressed with more than one, and often several, statistical analyses. Therefore, it is important to consider the strengths and weaknesses of alternative or competing research designs and statistical analyses.

Many questions will have to be addressed as you (and your advisor or collaborators) develop the design of the study. The following represent some likely questions to consider during this phase of the process. Will you be attempting to identify mean/median differences and/or the strength and direction of relationships? Will you be modeling latent variables, manifest variables, or both? Will you be using a covariance decomposition technique, a variance or components based technique, a qualitative technique or a ...? Will you be taking a Frequentist or Bayesian approach to data analysis? Will you be conducting a pilot study? Will you be doing simulations prior to data collection? Will you need Institutional Review Board (IRB²) approval? Will you need approval from other institutions (e.g. hospitals, schools, zoos, other universities)? Will your study be funded (e.g. grants)? Will you be handling sensitive information (e.g. health records)? Will you be collecting data from a vulnerable population (e.g. children)? How will you safeguard the data and insure it is kept confidential? Will you need to develop an Informed Consent form? Will your study involve any level of deception? If gathering data from human participants, will they be compensated (e.g. paid money, given extra credit, etc.)? Will your participants (humans) or subjects (non-humans) be treated safely, ethically, and respectfully? Of course they will, but you will still need to think about how they will be treated (e.g. will they benefit emotionally, physically, intellectually, and/or financially from participation in your study?).

Once the topic has been chosen, the literature review completed, formal hypotheses formulated, research design and proposed analyses decided (by you and your collaborators/advisor); you should prepare to propose the study in written and oral form. The proposal stage involves writing a formal proposal manuscript and presenting the proposed research, including all of the above information (often the bulk of the manuscript is the literature review). For students, oral presentation of the written proposal will be conducted as a method of gaining approval from a dissertation committee to proceed with the study. Students can find assistance with the process of writing by contacting the Writing Lab³. Once the committee has approved the study, very few deviations should be made from what was approved. If collecting new data, generally the next step would be IRB approval. Then, of course, data collection can proceed.

After Data Collection

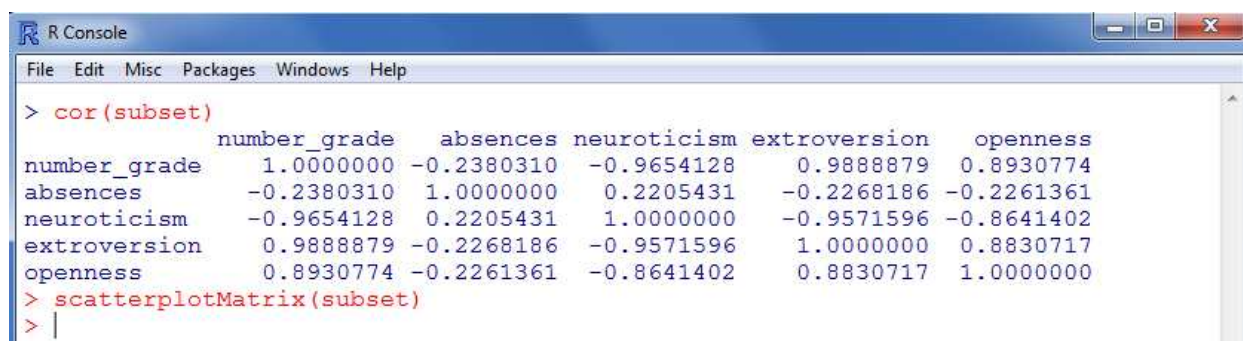
Once the data has been collected, the first step will commonly be to convert the data into a stable electronic format. It is generally recommended that the data be preserved in the most basic format possible; because, versions of software and operating systems change over time and it may be the case that future versions are not capable of opening a particular file format. Next to binary code form, the basic text format (filename.txt) is the obvious choice; using one of the common delimiters (e.g. comma delimited, space delimited, tab delimited, etc.). If one is using a traditional paper and pencil based survey, one can utilize

²<http://research.unt.edu/faculty-resources/research-integrity-and-compliance/use-of-huma>

³<http://www.unt.edu/writinglab/>

the services of Data Entry⁴ to have the paper surveys (or ScanTrons) digitized. If one is using a software program to enter the data (e.g. Microsoft Office Excel), then it is strongly recommended that the data be converted into text (.txt) files to be preserved. The second benefit of preserving data in text file format is that all popular statistical computing software is capable of opening text data files (for a comparison of statistical software, see Wikipedia⁵). This can be extremely important when multiple collaborators use different software (e.g. one collaborator using Open Office Calc and SAS on a Mac, and one collaborator using Microsoft Office Excel and IBM SPSS on a Windows PC).

Next, the data will likely be imported into one of the common statistical software packages for analysis (of course, RSS staff strongly recommends using R⁶). However, prior to conducting the primary and secondary analysis; one should do thorough initial data analysis. Initial data analysis refers to a wide variety of procedures which allow the researcher to become intimately familiar with the data (i.e. variable distributions, relationships, etc.). Initial data analysis ranges from rather mundane tasks such as recoding/reverse coding variables, reviewing histograms and bar charts for every variable; to more complex tasks like evaluating multivariate outliers and missing data. Whole books have been written on the subject of missing values (e.g. Little & Rubin, 2002), because, missing values are an important issue for virtually every dataset collected. Initial data analysis should also include an evaluation of the relationships between each pair of variables, with correlation matrices and scatterplot matrices commonly used. Testing the assumptions of planned parametric analyses should also be rigorously investigated (i.e. linearity, homoscedasticity, etc.). It should be noted that in this discussion of initial data analysis, the use of graphs is repeatedly mentioned. Graphs are important because they can convey information more clearly than simple numeric output; for example consider a five variable correlation matrix augmented with the same five variable relationships displayed in a scatterplot matrix:



```
R Console
File Edit Misc Packages Windows Help
> cor(subset)
      number_grade  absences  neuroticism  extroversion  openness
number_grade  1.0000000 -0.2380310  -0.9654128   0.9888879  0.8930774
absences      -0.2380310  1.0000000   0.2205431  -0.2268186 -0.2261361
neuroticism   -0.9654128  0.2205431   1.0000000  -0.9571596 -0.8641402
extroversion   0.9888879 -0.2268186  -0.9571596   1.0000000  0.8830717
openness       0.8930774 -0.2261361  -0.8641402   0.8830717  1.0000000
> scatterplotMatrix(subset)
> |
>
```

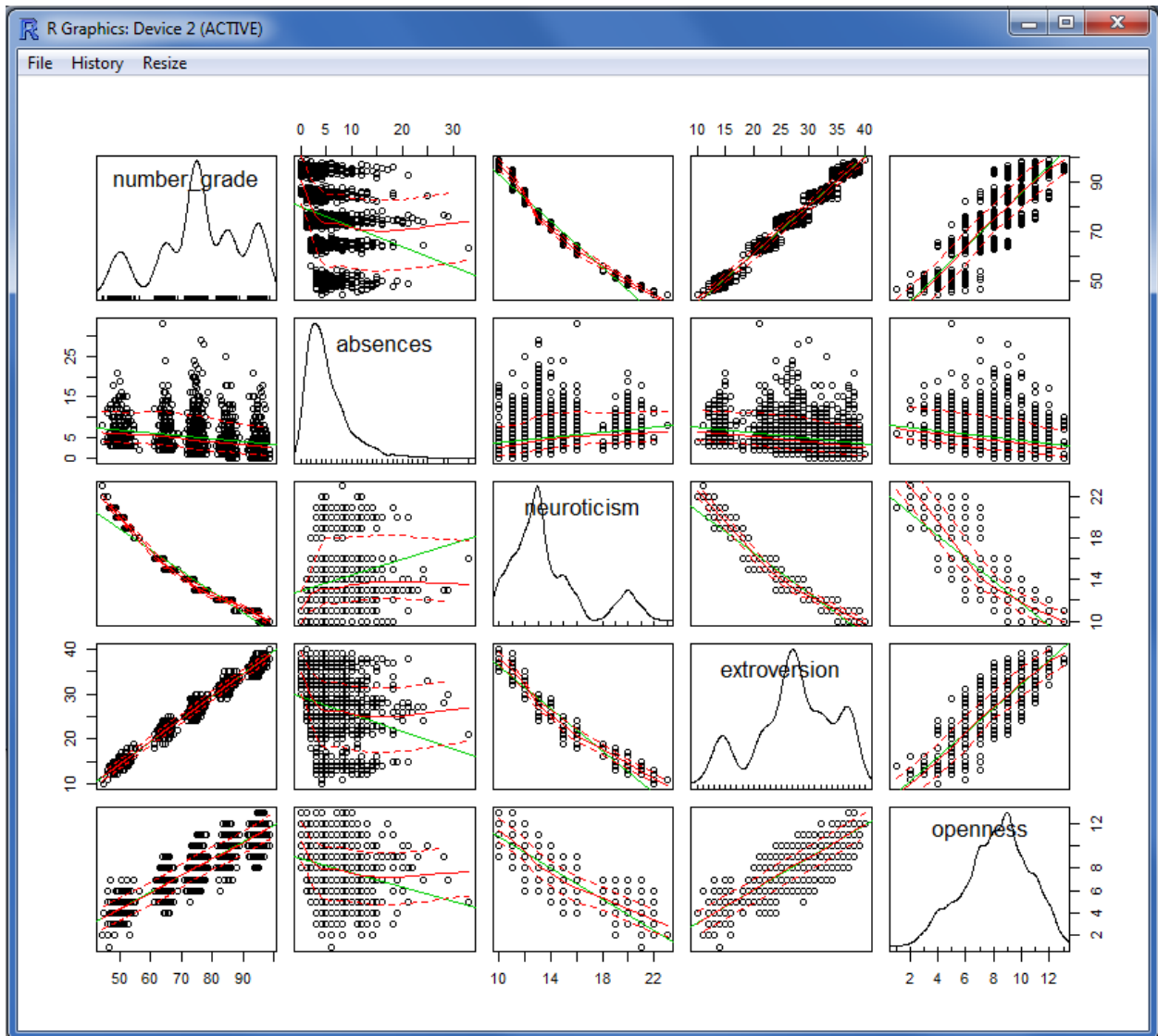
For those interested, these two screen captures can be replicated using this⁷ script.

⁴<http://www.unt.edu/ACS/datamanage.htm>

⁵http://en.wikipedia.org/wiki/Comparison_of_statistical_packages

⁶<http://cran.r-project.org/>

⁷http://www.unt.edu/rss/class/Jon/R_SC/Module6/M6_Funkiness.R



Initial data analysis may also employ parametric statistics, nonparametric statistics, transformations, optimal scaling techniques, variable selection techniques, matching, propensity score analysis, model comparison, etc. The point being made here is that initial data analysis is a necessary step, and one which requires critical thought, as well as time and effort – like all data analysis, it requires the tenacity and curiosity of a very good detective.

Primary, and secondary, analyses can commence once the initial data analysis is completed – although one may need to return to initial data analysis periodically during the course of alternative analyses (i.e. if proposed primary analyses are replaced). Due to the extremely wide array of analyses one might employ, specific techniques will not be covered here. However, there are three key concepts which should be kept in mind while conducting primary analyses. First, virtually all inferential statistics are model based and with models comes the possibility of model specification error. One could say there are two types of model specification error; errors of form and variable selection errors. Errors of form include specifying the wrong type of model, such as imposing a linear model when an exponential model or quadratic model might be more appropriate. Variable selection errors are errors of inclusion and errors of omission (e.g. meaningless variables in the model and meaningful variables left out of the model). Second, virtually all inferential statistics are based on some form of measurement and with measurement comes the possibility of measurement error. Measurement error is more prevalent

among the so-called soft sciences, as opposed to the hard sciences such as physics, biology, chemistry, etc.; however, measurement error should be investigated and modeled or acknowledged when discovered. Third, inferential statistics are, by their very name and nature, used to make inferences from a sample to a population. In other words, unless you are working with the entire population of interest, you are going to be computing or calculating *sample statistics* rather than *population parameters*. Therefore, sampling bias and/or non-response should be investigated and reported.

Given the rapid expansion of sophisticated modern methods, the data analyst should be open to using such robust techniques as booting (i.e. bootstrap resampling), bagging (i.e. bootstrapped aggregation), and boosting (i.e. using multiple models) to increase the precision and decrease the bias of statistical estimates. There are also modern sophisticated techniques to allow for statistical control of so-called nuisance variables or confounding variables; techniques such as nearest neighbor matching, balancing, random stratification and propensity score analysis. It should also be noted that there has been an expansion of optimization techniques in recent years, such that maximum likelihood, which is rather commonly known, has been joined by ant-colony optimization and genetic optimization algorithms. Both of which can be applied to certain situations with amazing speed and produce *optimal* results (i.e. optimize on the most probable estimate of a parameter). Also, for particularly large datasets and associated complex computation, UNT's High Performance Computing (HPC⁸) center is available for jobs which require serious computing power.

Of course, once the data has been analyzed and interpreted, it is time to write up the results and prepare the final presentation. Again, students can get assistance from the Writing Lab if they are having difficulty with the writing process. Students should turn to their dissertation advisor for advice on formatting the manuscript. For example, some departments use the Modern Language Association (MLS) style, some use the Chicago style, some use the American Psychological Association (APA) style, and still others use a style of their own creation or an amalgamation of several styles. Students may also, at some point, want to contact the Graduate Reader in order to prepare their completed dissertation (or thesis) for submission to the Toulouse Graduate School. Another thing to consider, when writing up an empirical research manuscript, is the journal in which one wishes to publish the results. It is often the case that journals have their own formatting idiosyncrasies and therefore, it is often a good idea to consult their web site to review their submission guidelines long in advance of actually submitting a manuscript for review.

Conclusions

It is important to note that this article represents a very general guide to the conduct of empirical research and it is aimed more toward students conducting a dissertation than that of the professional researcher. For students, it is important to note that your dissertation (or thesis) advisor should be able to offer you suggestions and guide your progress. However, not all questions have easy or readily available answers; students should be proactive in seeking out information through any or all available sources.

Side Note 3: RSS Can Help.

Of course, RSS can help with choices of research design and statistical analysis. However, it is important to remember that RSS staff will recommend and suggest; but it is ultimately the responsibility of the researcher to make decisions concerning what will be done. RSS has available literally walls full of books and articles related to research design and statistical analysis as well as the experience to be able to communicate the strengths and weaknesses of various choices. Please review our entire website (particularly the FAQ page), as well as last month's article which dealt directly with statistical resources, prior to contacting us for a consultation.

⁸<http://citc.unt.edu/hpc/>

Do not expect your advisor (or anyone else) to do your work for you. Completing a dissertation is hard work and should be a learning process. Remember, a meaningful study is one that contributes to a better understanding of the phenomena under investigation. Lastly, a couple of *sound-bytes* of wisdom: Do not be afraid of your own ignorance; Albert Einstein once quipped something to the effect of: “if we already knew the answers, it would not be called *re*-search.” Do not be afraid of non-significant results; as Thomas Edison once said, “I have not failed; I’ve just found 10,000 ways that won’t work!”

References, Resources, and perhaps useful Links.

Clark, M. (2007). What is Statistics? *Benchmarks: RSS Matters*, September 2007. Available at: <http://www.unt.edu/benchmarks/archives/2007/september07/rss.htm>

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate Data Analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall. Chapter 2 available at: http://www.unt.edu/rss/class/Jon/ResourcesWkshp/1998_HairEtAl_Ch2.pdf

Herrington, R. (2007). How long should my analysis take? *Benchmarks: RSS Matters*, July 2007. Available at: <http://www.unt.edu/benchmarks/archives/2007/july07/rss.htm>

Kirk, R. E. (1995). *Experimental Design* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken, NJ: John Wiley and Sons, Inc.

Mahoney, M. J. (2004). *Scientist as Subject: The Psychological Imperative*. Percheron Press. <http://www.amazon.com/Scientist-Subject-Psychological-Imperative-Foundation>

Mertler, C. A., & Vannatta, R. A. (2002). *Advanced and Multivariate Statistical Methods: Practical Application and Interpretation* (2nd ed.). Los Angeles, CA: Pyrczak Publishing. Chapter 3 available at: http://www.unt.edu/rss/class/Jon/ResourcesWkshp/2002_MertlerVannatta_Ch3.pdf

Pedhazur, E. J. (1997). *Multiple Regression in Behavioral Research* (3rd ed.). Crawfordsville, IN: R.R. Donnelley (for Wadsworth – Thomson Learning, Inc.). Chapter 3 available at: http://www.unt.edu/rss/class/Jon/ResourcesWkshp/1997_Pedhazur_Ch3.pdf

Raykov, T., & Marcoulides, G. A. (2008). *An Introduction to Applied Multivariate Analysis*. New York: Routledge (Taylor & Francis Group). Chapter 3 available at: http://www.unt.edu/rss/class/Jon/ResourcesWkshp/2008_RaykovMarcoulides_Ch3.pdf

Starkweather, J. (2011). Go forth and propagate: Book recommendations for learning and teaching Bayesian statistics. *Benchmarks: RSS Matters*, September 2011. Available at: <http://web3.unt.edu/benchmarks/issues/2011/09/rss-matters>

Starkweather, J. (2011). Statistical resources. *Benchmarks: RSS Matters*, November 2011. Available at: <http://web3.unt.edu/benchmarks/issues/2011/10/rss-matters>

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics* (4th ed.). Needham, MA: Allyn & Bacon. Chapter 4 available at:
http://www.unt.edu/rss/class/Jon/ResourcesWkshp/2001_TabachnickFidell_Ch4.pdf

This article was last updated on June 15, 2012.
This document was created using L^AT_EX