Biost / Stat 579: Data Analysis and Report Writing
Emerson, Fall 2011

Project Assignment
September 30, 2011

General Comments:

For the project, students will be assigned a data analysis that is to be completed within a 1-2 week period at the end of the quarter. The data sets and their descriptions will be posted on the class web pages at the time that the project is actually assigned.

After turning in your paper, there will be an oral defense of your analysis and write-up.

Ground Rules:

1. Prior to your oral defense, you are not to discuss your data analysis or paper with anyone other than me.
2. The report you submit is to be your own work. I take plagiarism very seriously. Thus you should not copy information you obtain from other works into your report. This prohibition extends to the documentation of the dataset which I provided. Use your own words. I have many anecdotes of recognizing my wording that appeared in papers that I had refereed several years earlier. I also have much experience with seeing the same wording appearing in different papers received from the same class. These instances are usually easily traced these days to web pages. In any case, you are forewarned: This is something I notice when grading papers.

Requirements for the Manuscript:

Your paper should be no more 20 pages in length (so 1 to 20 single sided sheets of paper or the equivalent printed double sided), counting figures and tables. It may not use fonts less than 10 points for the main text.

In this report, you should describe the results of your analysis and the conclusions you would reach from those results. This report should look like a formal report to a statistically naïve client (i.e., the researcher who brought you the data and/or involved you in the analysis) or an interested reader of the scientific/medical literature. Because a statistical analysis aims to answer a scientific question, you should organize your report in the manner which is customarily used in science. To wit:

1. *Summary:* Provide a concise description of the question, the data used to try to answer it, and the conclusions of your analysis. Give a brief description of the study design/sampling scheme. Give the most pertinent estimates, confidence intervals, and P values. **Note that estimates and confidence intervals regarding the main question of interest are also important even when there is no statistically significant effect.** Don't give too much detail here, but do note any significant problems that were encountered. The basic goal is to have all the key information in your summary, and the rest of your report is the supporting detail. (Many journals now require abstracts that are organized just like the rest of the paper, with a sentence or two for each of Background, Methods, Results, Discusson.)

2. *Background*: Provide a description of the scientific motivation for the analysis. Use your own words rather than copying the description provided by the client or the description from some other source. By providing your understanding of the problem, the client may be able to correct any misconceptions that you had about the science. You don't have to go into great detail here, but do give all the facts that entered into your decision process during the analysis. Generally this will include a statement about the overall goal you are trying to address (e.g., the disease and the public health impact of the disease), the current state of knowledge (e.g., conclusions reached in previous studies), and the specific aims of the current study.

3. *Questions of Interest*: List the specific questions that your client posed as well as the questions that you answered. Highlight discrepancies between the two categories of questions.

4. *Source of the Data*: Describe the source and sampling methods for the data, if known. Describe the variables that are available and their meaning for the analysis. Highlight patterns of missing data as well as possible confounding by measured or unmeasured variables. This should not be a detailed presentation of descriptive statistics, however. That will come under Results.

5. *Statistical Methods*: Describe the methods used for the analysis at two levels. 1) Give a low-level technical description of the analysis for the client to use in the manuscript. Include references for non-standard techniques. You may want to describe the software used, and you certainly want to describe any methods used for assessing the appropriateness of your models. (Note that your goal should be to use methods that require as few assumptions as possible.) Explain how you handled common problems like missing data, multiple comparisons, etc. 2) Explain the basic philosophy behind the analysis techniques in layman's terms. Provide interpretations for all parameter estimates. Motivate transformations. Describe the use of P values and confidence intervals if they play an important role in your analysis. Explain why you didn't use more common techniques if necessary.

6. *Results*: Provide the pertinent results of your analyses. Do not include all the dead-end analyses you might have done unless they provide insight into the question. Do lead the client up to the analyses gradually.
    a. Start off with descriptive statistics. This is an area often given short shrift. The goal is to describe the basic characteristics of the sample used to address the question (materials and methods), as well as to present simple descriptive statistics (non-model based) that address the questions. Tables and plots are the key tools. If there are any characteristics of the data that present technical problems that needed to be addressed in the modeling (validity of any assumptions), try to present descriptive statistics illustrating those issues. Avoid statistical jargon and the methods that can only be described with statistical jargon. (Residual plots are almost always a bad idea in the paper, and most often not that useful otherwise—you should have chosen a model with fewer assumptions.) The basic idea is to presage all the issues you will talk about when presenting the models used in statistical inference, insofar as possible with simple descriptive statistics.
    b. Then go to the major analyses used to answer the primary questions. Present summaries of the statistical inference obtained from these models (point estimates, CI, P values). Highlight any particular issues that materially affected

the models used to answer the question (confounding, interactions, nonlinearities, etc.) Tables can often be used to good effect here.

c. Leave exploratory analyses (if any) for last and highlight the exploratory nature of those analyses.

d. Present the results of your analyses in tables and publishing quality figures. DO NOT INCLUDE OUTPUT FROM STATISTICAL PROGRAMS. (Such means little to me and nothing to a client). When possible, use words instead of cryptic variable names. Use forms of estimates that have some meaning to a statistically naive researcher. Thus, if you log transform your response, present geometric mean ratios rather than linear regression parameters. Present confidence intervals rather than the values of Z, t, F, or chi squared statistics.

7. *Discussion*: Discuss the conclusions which you feel can be drawn from the analyses. Suggest directions for future studies and analyses. Highlight the limitations of the data and your analyses. Sometimes particularly speculative analyses are reported here—this is especially true of informal meta-analyses that might compare the newly reported results with what had been previously observed in the literature.

The major theme of the above is to write to the client and the scientific community rather than to a statistician. If you cannot explain your findings in a straightforward manner, then the analysis is of little value to anyone.

Also, lead your reader to all the proper results. You spent a long time analyzing the data. Now provide a brief tour through the high points of your work. Statistical diagnostics, which take a lot of our time, can most often be avoided through carefully wording your question and choosing your methods. If you have to use diagnostic methods, the technical results can usually be summarized in a single sentence (``Similar trends were observed at other time points." or ``We found no evidence to suggest that the final model did not fit the data adequately.") If the assumptions of your initial, pre-specified model become a scientific issue (e.g., linearity of effect or influence of particular cases), there are almost straightforward scientific summaries that can be presented (e.g., stratified smooths, tests of nonlinearity, evaluation of influence). You are reporting your major results and impressions of the data. If the client wanted to see every detail, he/she would have to do the analysis himself/herself.

It is probably most useful to first consider the tables and figures you will present. For instance, in a clinical trial, I would tend to include

1. Table 1: Descriptive statistics for the patient characteristics by treatment group. The purpose of such a table is to allow the reader to assess the comparability of treatment groups with respect to other predictors of response such as age, sex, etc., while at the same time giving them an idea of the types of patients used in the clinical trial.

2. Table 2: If missing data or sampling scheme issues are a major problem, a table depicting the availability of data by variable and time will be important. Many journals require a “CONSORT graph” that is a specific way of showing how the patient population might have fared as they progressed through screening, treatment, drop out, and assessment of the final outcome. Other times it will be important to show adherence to the timing of measurements, etc., as well as the ways in which the patients with complete data differ from patients with incomplete data.

3. Table 3: Descriptive statistics for outcomes by treatment group. While we are ultimately interested in making inference about some summary measure (along with its precision as measured by a CI or a SE), we need to recognize that excessively high or low outcomes may indicate possible toxic treatments for individual patients (so ranges of the data and/or SD are also of interest). Hence, this table might focus more on the data itself, rather than the inference. (The inference is further described below.)

4. Figure 1: A graphical display of outcomes. This could either be primarily descriptive (e.g., by showing the (possibly jittered) data) by treatment group with superimposed smooths, or it could be primarily inferential (by showing point estimates with standard error bars or confidence intervals). With time to event data, it is not uncommon to display the survival curves, which also serves to depict the range of the data. In this case, consideration might also be given to the censoring distribution. In studies with repeated measurements over time, estimates of the outcomes (with SD or SE bars depending on you goals) might be displayed.

5. Table 4: Inferential statistics presenting results by treatment group. This table would typically include point estimates, confidence intervals, and P values. It might also include exploratory analyses displaying analogous results after adjustment for specific baseline covariates of within important subgroups. Sometimes, other types of sensitivity analyses (e.g., to explore impact of missing data) should be included.

Reports of observational data analyses do not really differ that much, but

1. In Table 1, we are no longer protected by randomization, so inference about differences between groups defined by the predictor of interest might be more important.
2. There is heightened importance of Table 2, as the measurement scheme may not have been protocolized at all.
3. In Table 4, there is more of a tendency to explore multiple models in order to try to tease out confounding relationships. (It is best if all those models have been pre-specified, as multiple comparison issues in stepwise model building are, to my mind, the biggest source of nonreproducible results.)