

Biost 517: Applied Biostatistics I

Emerson, Fall 2005

Project Assignment

November 24, 2005

General Comments:

For the project, students have been assigned to a writing group of 3 – 4 students. You should have received email notification of your group assignment and which question your group is to address. Do not provide analyses of the other question. In both cases, you need only consider analysis methods based on comparing two groups at a time.

The data set and its description are posted on the class web pages.

Each writing group will submit a short paper describing the results of a statistical analysis of the appropriate aspect of a clinical trial. The paper will then be anonymously refereed by another student group in the class, and a revised paper will be resubmitted. Grading of the project will be primarily based on the revised paper, though cases of egregious nonperformance on the first submission and/or the referee's report will lower the grade on the project.

In order to have the referee process remain as anonymous as possible, each paper should be identified only by the group number assigned in the email. Your group number should not be divulged to anyone.

Due Dates:

- 10:30 am, Monday, December 5, 2005: Each group should submit the name of their corresponding author by email (semerson@u.washington.edu). The corresponding author will be responsible for picking up the paper to be refereed at the end of class on Wednesday, December 7.
- 9:30 am, Wednesday, December 7, 2005: Submission of five copies of first draft. Your paper should be labeled only with your group number, NOT your names. The copies should be submitted in a manilla envelope labeled with the name of the corresponding author, but not your group number. (These envelopes will be used at the end of the class to distribute the papers to be refereed by each group.)
 - *Please note that this deadline is strict. Failure to have the papers available at the start of class on Wednesday will be synonymous with failure on the project. If there is a chance of being struck by a meteor on the way to class, you are strongly urged to make other arrangements to get the papers to class on time.*
- 10:20 am, Wednesday, December 7, 2005: Copies of the paper to be refereed will be distributed to the corresponding author of each group.
- 9:30 am, Friday, December 9, 2005: Submission of five copies of your group's referees' report. The copies should be submitted in a manilla envelope labeled with the name of the corresponding author, but not your group number. (These envelopes will be used at the end of the class to distribute the referees report to each group.)
 - *Again, this deadline is strict. Failure to have the referees' report available at the start of class on Friday will be synonymous with failure on the project.*
- 10:20 am, Friday, December 9, 2005: Copies of the referees' report will be distributed to the corresponding author of each group.

- 8:30 am, Wednesday, December 14, 2005: Submission of two copies of your group's final paper. The paper should be labeled only with your group number, not your names. Along with the paper, you should supply completed and signed forms for each group member declaring the Authorship Responsibility, Criteria, and Contributions for that group member. (We will use the forms used by the *Journal of the American Medical Association*, copies of which will be distributed to you.)

Ground Rules:

1. You are not to discuss your data analysis or paper with anyone other than the course instructor or course TAs.
2. Until you have handed in your final paper, you are not to access, read, or otherwise reference any paper dealing with either methotrexate or primary biliary cirrhosis that was published in the medical or scientific literature during 2005. This prohibition extends to any web pages referencing such publications. (If for any reason this prohibition presents problems in the performance of duties related to your employment or requirements for other classes, please discuss this with me.)
-- Rationale: The results of this clinical trial were published very recently. For the purposes of this project, I want to see how you would analyze this data based on information available prior to that publication. And I certainly do not want to see you just duplicate the analyses and/or description that I already did. (I do note that the data you have been given is a slightly different cut of the data that was actually analyzed in the publication.)
3. 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 more than 0 and fewer than 6 pages in length (so 1 to 5 single sided sheets of paper or the equivalent printed double sided), not counting figures and tables. It may contain at most three tables and at most one figure (though the figure may have multiple panels to display different endpoints). 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 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.

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 certainly want to describe the methods used for assessing the appropriateness of your models. 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 in previous years. 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. 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 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.") 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. In a clinical trial such as this, 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: 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 (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.)
3. 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.

4. Table 3: Inferential statistics presenting results by treatment group. This table would typically include point estimates, confidence intervals, and P values.

I note that you need not follow this scheme. It is quite possible to combine the information of Table 2 with that of Table 1, Figure 1, and/or Table 3. But you do need that information displayed somehow. (I might well choose to give baseline measurements of plasma/serum levels in Table 1, display the data in my figures, and then just give tabular inferential statistics.) And you editors always like you better if you do not use tables for things that can be presented in the text. (My personal opinion, however, is that tables often lay out the pertinent information in a more logical manner.)