

**Biost 518: Applied Biostatistics II**  
**Biost 515: Biostatistics II**  
Emerson, Winter 2015

**Project Assignment**  
February 2, 2015

**General Comments:**

For the project, students are assigned to a writing group of approximately 4 students. The project deals with the analysis related to sex discrimination in university salaries. You will receive email notification of your group assignment.

The data set and its descriptions are posted on the class web pages (salary.doc and salary.txt).

Each writing group will submit a short paper describing the results of a statistical analysis to a scientist collaborator. Papers should be submitted electronically as MS-Word documents, labeled only with the group number (and otherwise anonymized).

You should also note that I will post anonymous versions of the papers on the course web pages at some future date.

**Due Dates:**

- 9:30 am, Wednesday, February 18, 2015: Each student should submit an electronic version of an abbreviated Statistical Analysis Plan (SAP) via the Catalyst drop box.
  - The abbreviated statistical analysis plan is ideally written after looking at the documentation for the file, but prior to looking at the data. It should include:
    - A description of the statistical methods you will use for the descriptive statistics you will present in your paper regarding the subjects used in your analysis and the sampling frame (available measurements). This should include a “mock-up” of any tables or figures that will be used.
    - A description of the statistical methods you will use for the descriptive statistic you will present in your paper regarding the association between the POI and the response variable. Again, include “mock-up” of any tables or figures that will be used.
    - A description of the statistical methods you will use for statistical inference you will use to answer the primary question and any secondary questions.
    - A description of the figures and tables you will use to present those inferential statistics in your paper.
  - Under the title of your SAP, it should be labeled with your name.
  - The file you submit should be a MS Word document.
  - The file name must follow the following very strict format. If your last name is Emerson, your file should be named: sapemerson.doc (or sapemerson.docx if you are using a more recent version of MS Word).
  - *After all SAPs have been submitted, I will assign groups and distribute the SAPs to the members. Later I will hold an extra session to discuss some general issues regarding the data analyses. Hence, the abbreviated Statistical Analysis Plans must be submitted on time.*
- 5:00 pm, Wednesday, March 4, 2015: Each group should submit an electronic version of a first draft of their paper via the Catalyst dropbox.

- Your paper should be labeled only with your group number, NOT your names.
- The file you submit should be a MS Word document.
- The file name must follow the following very strict format. If you are Group *kk*, your file should be named: *draftkk.doc* (or *draftkk.docx* if you are using more recent versions of MS Word). You need to use lower case. If you are group 1 – 9, please use *draft09.doc*, etc. If you fail to name your file correctly, I will return it to you.
- 7:00 pm, Wednesday, March 4, 2015: Copies of another group's paper to be refereed will be distributed to each group via email.
- 11:59 pm, Friday, March 6, 2015: Each group should submit an electronic version of its referee report via the Catalyst dropbox. Clearly indicate the group number of the authors, as well as the group number (NOT your names) of the referees.
  - The heading of your referee report should say "Comments on the paper authored by Group *kk* as Refereed by Group *nn*" DO NOT sign your names.
  - The file you submit should be a MS Word document. It should be a separate document with general comments, followed by enumerated specific comments. **It should NOT be comments added to the original manuscript.**
  - The file name must follow the following very strict format. If you are Group *nn* refereeing the paper by Group *kk*, your file should be named: *refereennauthorkk.doc* (or *refereennauthorkk.docx* if you are using a more recent version of MS Word). You need to use lower case. If you are group 1 – 9, please use *referee01author09.doc*, etc. If you fail to name your file correctly, I will return it to you.
  - *Again, the deadline is strict. Failure to have the referees' report available at the prescribed time will be synonymous with failure on the project.*
- 7:00 am, Saturday: March 7, 2015: Copies of the referees' report will be emailed to the author group. You will probably receive some comments from one of the TAs, as well as possibly my comments on your Abstract.
- 5:00 pm, Friday, March 13, 2015: Each group should submit an electronic version of their final report via the Catalyst dropbox.
  - Your paper should be labeled only with your group number, NOT your names.
  - The file you submit should be a MS Word document.
  - The file name must follow the following very strict format. If you are Group *kk*, your file should be named: *finalkk.doc* (or *finalkk.docx* if you are using more recent versions of MS Word). You need to use lower case. If you are group 1 – 9, please use *final09.doc*, etc. If you fail to name your file correctly, I will return it to you.
  - The group should also provide a file describing their response to the referees. The file name must follow the following very strict format. If you are Group *kk*, your file should be named: *responsekk.doc* (or *responsekk.docx* if you are using more recent versions of MS Word). You need to use lower case. If you are group 1 – 9, please use *final09.doc*, etc. If you fail to name your file correctly, I will return it to you.
  - The group should also provide a file describing the contribution of each member to the final paper. The file name must follow the following very strict format. If you are Group *kk*, your file should be named: *contributionskk.doc* (or *contributionskk.docx* if you are using more recent versions of MS Word). You need to use lower case. If you are group 1 – 9, please use *final09.doc*, etc. If you fail to name your file correctly, I will return it to you.

- Various times, Monday - Friday, March 16-20, 2015: Each group can at their option meet for approximately 30-45 minutes with the instructor and TAs to defend their report and to receive feedback.

### Ground Rules:

1. You are not to discuss your data analysis or paper with anyone other than the course instructor or course TAs.
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:

*In contrast to recent years, I am asking this year for a paper that might represent the first report of a data analysis to a collaborator, as opposed to a paper that might be submitted to a journal. Hence, the supposition is that you are the statistical analyst, and your collaborator is the expert on the area of application.*

*I have posted an example paper on the web. I wrote this as an example of a write-up for a parametric survival analysis almost 20 years ago. While I modified it a little for these purposes, I remind you of the dictum: "If you are not embarrassed of what you did 6 months ago, you are not learning anything."*

Your paper should be more than 0 and fewer than 13 pages in length (so 1 to 12 single sided sheets of paper or the equivalent printed double sided), not counting figures and tables. It may contain at most six tables and at most four figures (though each figure may have multiple panels to display different endpoints). It may not use fonts less than 10 points for the main text. (Do not feel compelled to hit the maximum on any of these, I am just trying to give you some flexibility while avoiding a proliferation of information that exceeds the size of the original data files: If you present more statistics than original data points, you have clearly failed at summarizing the results.)

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). 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.

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. (You do not need to do a literature search, though you may if you really want. However, the goal of this project is the statistical analysis and its correct interpretation. I usually hold my collaborators responsible for having done the literature search.)
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 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. Highlight the limitations of the data and your analyses. Sometimes particularly speculative analyses are reported here. But you do not need to give all the discussion that would eventually appear in a scientific journal. Suggest directions for future analyses that might be possible prior to publication of these results, but you do not at this stage need to suggest what next experiment the scientific field needs to consider.

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 studies such as these, I would tend to include

1. Table 1: Descriptive statistics for the subject characteristics at time of study inclusion, perhaps broken down by any primary predictor of interest (if there is one) or by outcome group. The purpose of such a table is to allow the reader to assess the comparability of important 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 subjects used in the study.
2. Table 2: Descriptive statistics for the “subject disposition” detailing the intensity of follow-up and availability of data. It should be anticipated that subjects will vary in their available data due to their employment history. Any missing data that results from such varied participation can have major impact on the generalizability (at least) and credibility of the trial results. At the very minimum, we would want to know what data is missing. (In an observational longitudinal study, this becomes extremely important.)
3. Table 3: Descriptive statistics for outcomes by primary 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 important variability for individual subjects (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: Any relevant 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.
5. Table 4: Inferential statistics presenting results by primary group. This table would typically include point estimates, confidence intervals, and P values. When the primary question involves some amount of exploration, this table might present separately the univariate analyses and adjusted analyses.

I note that you need not follow this scheme. But you do need the information displayed somehow. I also note that there are four somewhat related questions that you are asked to answer, and each of these questions may require its own "Table 4".