Introduction to Statistics (MAT/SST 115.03 2008S)

Project Overview

Summary: In the project, you will explore a large data set using statistical techniques and then report on aspects of that data set in four ways: In a short note, in a "lightning presentation", in a two or three page paper, and in a poster.

Purposes: To give you the opportunity to use statistical techniques in the context of a more open problem. To encourage you to communicate with statistics in a variety of ways.

Collaboration: Students should do this assignment in teams of size two or three. (If you really prefer not to work with other people, that will also be possible, but please talk to me asap.) You may discuss the assignment with anyone you wish. You may discuss the assignment with anyone you wish. You may also obtain help from anyone you wish, but you should clearly document that help. You may choose a different group for the first and second parts of the project.

Due Dates:

- Lightning talk: Wednesday, 16 April 2008
- Memo using descriptive statistics: Monday, 21 April 2008
- Poster (as Powerpoint or PDF): Friday, 25 April 2008
- Paper: Wednesday, 30 April 2008.
- Presentation of Poster: Thursday, 1 May 2008

Citation: The idea for the project comes from previous faculty teaching statistics. I have also borrowed text describing the project from handouts by Kathy Kamp, Kent McClelland, and Katherine McClelland.

Warning: So that this assignment is a learning experience for everyone, I may spend class time publicly critiquing your work.

Introduction

While you have done a variety of kinds of statistical analysis this semester on a wide variety of data sets, all of your work was somewhat constrained: You were told what patterns to look for and what kind of statistical tests to use. Certainly, this approach is appropriate to help you learn the tests and techniques and to make sure that you can interpret the statistical results that others create.

However, you should also have some experience in a more open-ended setting, one in which you identify the research question, determine how to analyze the data, analyze the data, and express the results in a way that others can understand them.

The two parts and four components of the project are intended to give you this experience, using a fairly large data set that is ripe for exploration: Information from a first-year survey of Grinnell students.

The first part of the project (aka “The Mini-Project”) asks you to use simple descriptive statistics to look at the effects of sex on one variable of your choice. You will present the results of the mini-project in two forms: In a short memo to a campus person of your choice and in a lightning presentation (two-to-three minutes) to the class.

The second part of the project asks you to pick both the explanatory (or possibly explanatory) and response variables for a hypothesis and to use more sophisticated statistical techniques (particularly those we will see in Topics 21, 24, and 25) to explore the relationships between the variables. You will represent the results of the second part of the project in two forms: In a two-to-three page paper and in a poster presentation.

You’ll note that the two parts of the project ask you to communicate the information in four ways: in informal writing, in spoken form (with visual aids), in formal writing, and in a formal presentation that requires both writing and speaking.

About the Data Set

Most of you will remember taking a survey during orientation. The 2004 survey (for the class of 2008) was particularly interesting in that it included not just the standard questions, but a variety of questions on beliefs and values. We will be working with those data. To protect confidentiality, identifying questions have been removed.

This is a census of the students for the year in question, but, according to the person doing Institutional Research at the time, a single year can be viewed as a sample because the results are quite consistent from year to year, in the short term, at least. For the purposes of practice with inferential data analysis, we will treat this as a SRS from all recent Grinnell Students.

You should remember that these data were provided to us as a privilege for use in a scholarly way, which we must not abuse by sensationalizing findings or otherwise misusing the data. If you are not sure how certain questions were put into columns or how things were coded, just ask.

Unfortunatley, the data set provided for this project does not meet all the technical conditions necessary for proper use of inferential procedures. Your report should clearly demonstrate your awareness of the way or ways in which the data set may fall short.

Reading Your Data

The data set is stored in the file `/home/rebelsky/Stats115/Data/first-year-survey-200x.csv`. You read it in with

```
FYS = read.csv("/home/rebelsky/Stats115/Data/first-year-survey-200x.csv")
```

While you should know enough R to explore the data, I will provide handouts that suggest some typical ways to work with the data. We will also go through a simple analysis in class.

The Mini-project

The mini-project is intended to give you an opportunity to explore the data set, experience massaging the data, and a bit of experience using simple descriptive statistics. You should find some surprising or interesting information for one variable in the survey, look at how reported sex affects that variable, and report your results. We will practice this approach in class.

In a *memo* to someone in authority on campus (President Osgood, a member of Academic Advising, a faculty member, etc.), you will explain what you have found. Your explanation should provide a simple analysis of the data and at least one helpful graph. (You will not send the memo; we are just using the memo form to give you a way to frame the information.)

In a *lightning presentation* of two or three minutes in class, you will explain what you have found to your colleagues. In that presentation, you may use any graphs and data you find useful. (I will make sure to have a document projector for you to show them; you simply need to bring them on paper.)

The Poster Project

In the second project (aka "The Poster Project"), you will find a few variables and explore their relationships. In this project, once you have found and explored the variables of interest, you should make a hypothesis about relationships between the variables (e.g., a more sophisticated version of something like "a student's number of hours per week of community service in high school predicts whether or not Grinnell was their first choice").

You should formulate a suitable research question that can be answered by an analysis of a limited number of these variables, typically a total of no more than five or six. Most questions either (1) consider the effects of a group of explanatory variables on a single response variable or (2) consider the effect of a single explanatory variable on multiple response variables. For your own sake, you should keep things simple and avoid the temptation to explore everything that might be interesting in the data. The emphasis should be on quality, not quantity. Your exploration of the data should begin with a careful univariate analysis of each of the selected variables, so that you can tell immediately if there are any problems with the variables you have chosen. After that preliminary analysis, you will apply inferential procedures to better understand how the data support and fail to support your hypotheses.

You will first present the results of your analysis in a *poster*, like the ones you see scattered throughout the halls of the science building. The posters will be part of a poster session. During this session, one member of your group will be responsible for standing by the poster and answering questions that visitors (other students, faculty members, passerby) may have. Other members of your group will be responsible for looking at and assessing the posters of peers. You will rotate the responsibility of who stands at the poster. [More details to follow.]

You will also present these results in a *short paper*, of two or three pages. In writing the paper, you should assume that your reader is an alum of 115 (that is, a student at Grinnell College who knows basic statistical techniques).

Both poster and paper should include the following sections.

1. An *Informative Title* in which you orient readers immediately to the information you are presenting.

2. An *Introduction* in which you introduce your central research question and the specific questions to be answered with the data. If you have approached the data with certain expectations or theories, you should tell the reader about them here.

3. A *Methods* section in which you very briefly explain the source of the data, describe the sample and its limitations, describe the variables used in your analyses, and explain any re-codings or other changes you have made to the variables.

4. A *Results* section in which you provide concise answers to your research questions, using a combination of your fine prose and elegant statistical displays and summary statistics.

5. A *Discussion* section in which you briefly summarize your answers to the research questions and indicate, in the bigger picture, what your findings might mean. You should also use your *Discussion* section to suggest possible followup studies.
