Guidelines for the Final Project
Statistics 100 Spring 2012

What is the project?

The final project will be a brief presentation and poster display on a data analysis of any topic of interest to you. Your presentations will be given to your fellow students (in small groups); posters will be displayed in a class poster session in the style of a professional conference. The presentation and poster sessions will be scheduled at the end of reading period. The content of your project should be roughly equivalent to a 10-15 page paper, though you will not be writing a paper. The project is worth 15% of your final grade. The grade for the project will be based entirely on your presentation on project day. The preliminary proposals described in these guidelines will not be graded; they are for your benefit and are designed to help you get started.

More details about the presentations and the organization of the poster day will be provided closer to the day itself. These guidelines describe the substantive aspects of the project.

What types of projects are possible?

The most important part of a project is to start with an interesting question. Here are some examples:

- Do marriages end in divorce more often for women in high salaried jobs than in lower salaried jobs? Three years ago, students used data from from the Murray Center can be used to examine this question.
- Has the rate of property crimes increased in the last two years as the economic recession has continued? You may find data for this at the Department of Justice web site.
- Is the food at Annenberg worse than in the upper class houses (see the project mentioned below).
- Will the gap between men and women elite swimmers disappear in the coming years? Two women from the Harvard swimming team used world record data to make predictions on record times in the next 10 years.
- Can Harvard students taste the difference cookies made from a long cherished recipe vs those made from frozen dough? Two years ago, two students tested cookies made from an old family recipe vs. cookies made from frozen dough to see if students could tell the difference. Differences were measured in a blind, taste-testing experiment.

A project may be one of the following:

(1) An analysis of an existing dataset. You might apply some technique we have learned in the course to a dataset that has already been collected. The Web is a vast source of datasets on almost any subject, such as demographics, disease, economics, geography, entertainment, science, etc. Just search for your favorite topic in a web search engine and see what comes up. Here are some specific suggestions about places where you can find some interesting data.

- *Sports:* You may find Baseball Reference or ESPN useful. Major league baseball has a web site that project teams have used to devise a new way of rating players

- *Social Sciences:* There are many, many sources of information on social and political issues.

  - The course web page has a link to the Gapminder web site, a site with a great deal of information on global health and development.
  - The Murray Research Center at Radcliffe is an excellent source of social science datasets, especially for issues of sociology and psychology relevant to women. If you want to use data from the Murray Center, you have to submit a request and wait for the dataset to be prepared and sent to you. The process can take several weeks.
  - The Census Bureau has a large amount of data available on its web site that can be combined in interesting ways. The 2010 Statistical Abstract of the United States provides data on state level characteristics that can be used to build interesting data sets.

(2) An original experiment, observational study or survey. Experiments may be the most difficult to conceive, but they yield the most interpretable results. Three years ago, a student team explored the question of whether the food in upper class houses was better than in Annenberg. Rather than conduct a survey, the team convinced Annenberg and 3 upper class houses to donate 40 pieces of cooked chicken from each location. Forty participants ate the chicken in random order and rated the quality. The analysis showed that the upper class houses were the clear winner.

Observational studies are best done on well defined populations. A particularly interesting project a few years ago consisted of a combined survey and strength test of the strength and endurance of men on the Harvard Rowing Team (using a rowing machine). A regression model was used to determine predictors of strength (body weight was the strongest predictor, though having a girlfriend seemed to result in a slight strength decrease).

Surveys are the easiest to conduct, but by far the most difficult to interpret because of item response bias and non-respondent bias. Internet-based surveys are notoriously unreliable and often misleading. For this reason, we typically discourage surveys. Surveys conducted in Annenberg and upper class dining halls can also be disruptive.

If you choose to do a survey, we will expect you to think clearly about how you will design the survey and maximize response rates. We will also want to know what steps you took to field test your survey. The best surveys also have a defined population (e.g., a random sample of students in upper class houses, or in 2 or 3 entry-ways in a freshman dorm). It will also be essential that you obtain prior permission from the House master or the dorm proctor.

How can you begin designing a project before you know what methods of analysis Statistics 100 will cover?

The last 2/3 of Statistics 100 covers formal methods of inference for the data types that are studied in Units 1 and 2. After we finish Unit 2, you will have a very good idea of the kind of data you should gather in your project. A large component of the poster for the project consists of graphical and numerical summaries of the sort we are studying in Units 1 and 2.

Unit 3 in Statistics 100 examines a range of designs for both experimental and observational studies, so this unit will help you conceptualize the project. The due date for the project proposal is timed to coincide with the end of Unit 3.

How much data should you collect?

It is not possible to give a single answer to this question, since the work to collect data varies by data source. Projects that use data from the web typically have a sample size between 100 and 200. Surveys are more difficult to do; these projects are typically based in 50 - 100 respondents. Because experiments are done in controlled settings, small well-designed experiments often work well; samples between 20 and 50 are often sufficient.

Project requirements and due dates:

- The project must be done in groups of two or threestudents. You are responsible for forming your own group, but the TFs will help with match-making if you wish. On project day, you will be asked to submit a signed sheet indicating each team member's contribution to the project. All students in a team sign the same sheet, and each team member must commit to doing their share of the work.
- A preliminary proposal is due on Friday March 9. Based on the March 9 proposal, you will receive advice and comments from one of the TFs on how to improve the project, or how to make it more manageable. The TFs and I will provide feedback by March 26. Send the project proposal by email to the TF of the student whose last 4 digits of the Harvard ID is the largest number.
- A Project proposal revision is due on Friday April 13. The revision should include changes you have made to your proposal in response to the initial review and a report on the status of data collection.
- Late proposals, updates or projects will not be accepted.

What should go in the proposal?

Your initial proposal should not be longer than two pages and should include the following elements, in a numbered list:

1. The names and Harvard ID numbers of the members of your team.
2. The issue you wish to study, phrased as a question. Examples might be "Do women in high-salaried, high-stress jobs experience divorce at a higher rate than the general population?" or

"Can students type text messages faster on an iPhone than on a Blackberry?" Phrasing the issue as a question is a good way to make the issue specific, and all good research is driven by a question.

3. Your data source. Be specific this section. For the divorce question above, the Murray Center at Radcliffe is a good data source (but this project has already been done). For the texting question, a randomized experiment would be a good data source.

4. If your team is going to do an experiment, briefly describe the experimental design. If you are going to do a survey, list two or three sample questions. If you are going to use data from the web, please provide the names of the sites and urls that you expect to be helpful.

5. A very brief description of how you will conduct a small pilot of your data collection. For surveys, it is always best to ask a small number of volunteers to answer your survey to be sure the questions are not ambiguous.

6. A brief list of 3 or 4 of main variables you will collect. Specify which variable will be the response variable

7. An approximate timeline for the work you will do. The timeline for a pilot study, data collection and analysis should begin on approximately April 13 and end on the first day of reading period (April 26).

You should plan to meet frequently with your project group between March 26 and April 15 to refine your plans and begin data collection. Make appointments with the TF working with you on the project or with Professor Harrington to discuss any aspects of the project you need help with.