Data Analysis Project

Math 130 Section 01
Fall 2011

The goal of this project is to try to give you experience of using statistics in a practical setting. The main idea is to find a data set you find interesting, and to summarize it and make some inferences. Statistics is almost always a collaborative effort, and so I'd like you to work in teams of two, three, or four for this project.

Data Analysis Project:

- Brainstorm a question of interest (it can be silly – the idea is to have fun practicing statistics). If you have a project/data from some previous/current research you've done, that could be used.
- Design your own study based on that question of interest.
- Collect data
- Summarize your data with appropriate numerical and graphical methods
- Use appropriate inference procedures to make statements about the population of interest (choices below)
- Communicate your work effectively to others in a short class presentation
- Prepare a report of your project that conveys your data analysis process, results, and conclusions

Choices for inference procedures:

1. Hypothesis testing for a difference in two population means or a population mean difference
   (covered next week in class) i.e. $\mu_1 - \mu_2$ or μ_d scenarios
2. ANOVA – testing the equality of 3 or more population means, with multiple comparisons if appropriate (week after next in class)
3. Regression – examining a relationship between two quantitative variables and making an inference about the slope of a regression line (after ANOVA in class)

Human Subjects:

Your data may or may not involve human subjects. There are some additional guidelines for questions pertaining to studies involving human subjects that you are expected to follow if your data set does involve human subjects. The most important thing is that the data are anonymous, and can't identify anyone personally. If you decide to collect your own data and it involves humans, then the school's Internal Review Board (IRB) may need to approve it.

For the purposes of this project:

- Studies involving human subjects must focus on adults (college students are fine).
- Questions CANNOT be sensitive questions (e.g., drug use, sexual behavior) or questions about illegal behavior (e.g., drug use, underage alcohol – i.e. NO alcohol related topics) and cannot involve deception.
- Data MUST be collected anonymously so that it cannot be traced back to the individuals who provided the responses.
- If you think your question may not meet these guidelines, pick another question that does.

Basically these guidelines will allow you to administer surveys to your fellow students or observe some action undertaken by them so long as the questions are appropriate and you collect the data anonymously. It also allows for observations of adults who are not college students (example: you observe how long drivers stop at a stop sign and record their gender to test for differences in duration of average stop time).

Example Questions of Interest:

You may not use these examples. Similar ones are fine but try to find something interesting you want to investigate. Can you identify the inference procedures corresponding to these?

1. Compare prices of common grocery items at two different supermarkets to see whether one store is cheaper to shop at on average.
2. Compare the duration of stops at a stop sign for male and female drivers in Massachusetts to see if there is a gender difference in average stop time.
3. Investigate differences in average number of caffeinated beverages ingested in a week across the four class years at Amherst.
4. Explore the relationship between students estimated and actual caloric intake of their evening meal in Valentine (would take substantial effort).

Time Schedule and Deadlines:

1. One page (maximum) description of project as **a proposal is due to me on or before November 9th.** Proposals should include your proposed topic, choice of inference procedure, and group member names (one copy per group). An example proposal is given at the end of these instructions.
2. You will have my feedback in class on or by November 14th.
3. Preliminary and formal analysis can be done any time. You can work on writing up different parts (see below) as you do them as well.
4. **Class presentations are in class December 13th and 14th.**
5. **Final reports are due to me by Wednesday, December 14th at 5 p.m.** One copy per group. You can email your final report or turn in a hard copy in my mailbox in the mathematics/computer science office.

Class Presentations:

Communication is an important part of statistical work. Each group will make a short presentation to the class.

- 8 – 10 minutes long. A good rule of thumb is to spend about 1 minute per slide.
- Describe the data set you used. Where did it come from? What is it about?
- Brief summary of results
- Conclusion from your inference procedure
- Anything else you want to share that you found interesting, etc.

The format is up to you. You can have one presenter or multiple presenters. If you want to use Powerpoint, you should have it available on someone's U:drive or a USB flash drive or email it to me before class. Group order for project presentations will be set in advance after proposals by random assignment.

Reports – should contain the following sections (at a minimum):

- Introduction – Describe the data and what question of interest you are considering
- Methods
    - Give the hypotheses in terms of statistical notation (can ask me for help with Equation Editor if you want to do this in Word)
    - Describe the inference procedure that is appropriate for your question
    - Discuss the assumptions that inference procedure has
    - Discuss **how** you will check those assumptions (do not discuss results of the checks until results)
- Results
    - Preliminary Analysis
        - Graphical description of data
        - Numerical description of data
    - Formal Analysis
        - Assumption checking – report on whether or not the conditions check out
        - Inference procedure test statistic , p-value, decision and/or CI if appropriate

- Conclusions
    - Real world conclusion from inference procedure
    - Suggestions/Cautions for future analysis
- Appendix
    - Relevant graphs if not included above
    - Inference output (the piece of Rcmdr output with the test output)
    - Data Set (if longer than 2 pages, send me an electronic copy instead)

Do NOT simply fill in replies to the bullet points. This is a basic framework to help you with a starting structure as a guide, but you need to provide a coherent report. You may of course include points not listed here.

You should include relevant graphs; you may put them with the related content or in the appendix and refer to them. Note that you can resize graphs.

Length – there is no set limit, but a 5-6 page report and appendix combined is reasonable.

Assessment – This assignment is worth 50 points (~8% of your grade). Your project grade will be primarily based on your communication. Can we understand where the data came from and what they are about? Are your analysis methods reasonable? All group members will receive the same grade.

Miscellaneous – You may come see me for help at any time; this is especially important if you need help organizing the data into the appropriate format for analysis in R. However, I will not proofread drafts for completeness. (Since I cannot possibly proofread for all groups, I cannot do it for anyone.)

Example Proposal:
Group members:
Amy
Herle
Lacey

As students living on a budget, and perhaps not always eating at Valentine, we'd like to study the prices of some common grocery items at Big Y vs. Stop N Shop to see whether one store is cheaper to shop at on average. A quick Yahoo search for "common grocery list" reveals many results. We may adapt a list of those items, but plan to obtain results for more than 40 different items. To address "brand" issues, we will always compare store brands if available, and if not, at the first store we go to, we'll use a random number table to pick a brand (assuming fewer than 10 brands of most common items), and check for that brand at the other store.
Our setup for this project is therefore a paired t-test for average price. The observations are paired because we get one price from each observation at each store.
I.E. our data set will look like:

	BigY	StopNShop
Flour		
Water		
Sugar		
Yogurt		
Cheese		
etc.		

Likely comments (what I might comment if you turned this in):

Might be more interesting to compare “snack” or “party” items, rather than an entire grocery list

Are you doing Big Y- Stop or Stop-Big Y? Be sure to specify. Do you have a direction picked?

What happens if the brand you chose because there was no store brand is not at the other store?

What happens if something is on sale when you check prices?

Are you counting “card” discounts (both stores have those in some fashion)?

What happens if the store brands are not the same size!?! Are you going to record units as well?