"Final-Project Options"

As stated in the course syllabus, 15% of the course grade will consist of a data analysis project of your choosing. Pick one of the types of projects described below.

- Work in a group of 3 or 4 people max.
- Hand in a written report on the last day of class.
- Be prepared to give a short (≈ 10 min) presentation either to the class or to me.

Statistical investigations are generally designed to answer an essential question. So give some thought to structuring a question for which you don't know the answer but that can be attacked through data analysis. The project must go beyond merely descriptive statistics in order to draw inferences from data. Plan to consult with me for approval sometime before or on Monday, July 24th to discuss your idea. Don't hesitate to be creative, even if you're not sure your idea is feasible, and we'll work out an approach together.

Scope: As a broad target, final projects should involve approximately as much work as two homework assignments. For groups of more than one person, the total work should scale roughly linearly with the group size, and be distributed roughly equally.

Grading: This assignment will be worth 15% of your final grade. It may be worth more if it is particularly excellent. The criteria for grading the presentations are as follows (Presentation and write-up is worth 70 points of the total assignment):

- Appropriate Length (10%)
- Clarity of Presentation (25%)
- Clear connection between the data and your conclusion, and clear explanation of your data analysis (50%)
- Written Summary (10%)
- Individual Contribution (5%)

Other Grading and Deadlines:

Monday, July 24th ,2017 - Hand in a paragraph or a sentence or two about what you/your group are/is going to do for your final project. (30 points)
In-class presentations and Final Write-ups due: August 9, 2017 (70 points)

Regression Analysis Option

Later in the course you will be introduced to linear regression analysis, which consists of finding a predictive relationship between two variables.

Identify some measurable quantity that you think could be related to a few other quantities (1 quantity to at least 2) in some possibly predictable ways. You should either collect relevant data or find a source of available data that you can analyze. Choose an example for which a linear relationship is at least plausible with one of the variables, but don't pick an example having an obvious and strong linear relationship (like, for example, stress vs strain for a uniform material). Find a situation for which the answer is not already well known.

- Use techniques discussed in Chapter 12 to perform regression analyses and find regression equations that allow you to (at least partially) predict the value of one variable from the values of the other variables.
- Use a variety of techniques to judge the appropriateness of your models and the strengths of the relationships. (Remember that the coefficient of determination r^2 can be compared for different types of models, but the correlation coefficient r is only used for linear models.)
- Include appropriate inferences (tests of hypothesis) regarding the model parameters. We won't discuss these in class for a few weeks, but Sections 12.3 gives some background on these topics.
- Use your models to make predictions, and discuss the uncertainty involved ("limits of prediction" in Section 12.3).
- Summarize your findings, and include a discussion of any possible lurking or confounding variables that may affect the results.

Simulation Option

As we will see in a future classes, simulation is a technique by which you can create, using probabilities, a "virtual experiment" that allows you to see what might happen in a particular situation. (We have considered simple situations including simulating free throw contests, traffic patterns, and completion times for a four-activity construction project.)

Formulate a question you might want to answer through simulation, for which the answer is not already known. Remember that simulation is often used to answer the question "What if...?" in order to test out a new way of doing things, or to get a handle on what might happen in a situation before it is actually tried. (For example, the manager of a new product line for a manufacturing company might be interested in finding out how long the development of the product might take. A complex product design schedule might consist of hundreds of interrelated activities that could be simulated in a manner similar to the four-activity problem you submitted for homework. Or, in another example, the manager of a trucking company that has one loading dock might want to simulate the loading times for trucks if an extra loading dock were added, in order to see if the idle time of trucks waiting in line to be loaded or unloaded could be reduced significantly.)

- The situation should have several interconnected components that must be separately simulated.
- Information on the probability of occurrence of all the possible values of the quantities to be simulated must be available. This may consist of theoretically determined probabilities (as in the case of a random variable that follows a known probability distribution, like a binomial or Poisson distribution for example); or it may come from historical empirical data, in other words data that have been observed in the past, or from data you could collect yourself in a pilot study. In any case, you should justify your assignments of probabilities.
- Carry out enough simulations of the situation so that a statistical summary of the results allows you to draw meaningful conclusions. Include a consideration of the uncertainty (accuracy) of any numerical results.

Hypothesis Test Option

A test of hypothesis, or hypothesis test, is the proper terminology for what is often simply called a "statistical test." Such a test is usually conducted to answer a yes-or-no question. Examples might include a study to determine whether taking regular supplements of fish oil helps prevent heart disease, or whether adding silicon to concrete increases its strength, or whether baseball bats made of maple wood are more likely to break than ones made of ash.

We will study the basic protocol for performing hypothesis tests at the end of the semester; they are presented in the textbook beginning in Chapter 8. A statistical test always results in

- a conclusion that answers the question that was initially posed, proving or disproving the stated hypothesis; and
- a measure of the probability that the conclusion is correct.

This option will require a considerable data collection component or finding good data to compare.

Statistical Machine Learning

Machine learning techniques are algorithms that are constructed or studied that can learn from data and make various predictions on data. Machine learning algorithms are obtained by building a model from example inputs and outputs (supervised learning involves outputs, unsupervised learning involves no outputs) in order to make a predictive model, which can be used to predict an unknown output (could be a class, hence classification problem) for a given set of inputs. Examples applications include face-recognition, spam filters, toxicity and search engines.

The machine learning analysis should minimally include

- Basic understanding of the algorithm used (e.g. Bayesian classification, Support Vectors, Principle component Analysis, k-nearest neighbors, Decision Trees, Logistical Regression ect.)
- Use your models to make predictions, and discuss the accuracy of the model.

"Piggy-back" Option

You may also choose to attach a significant statistical analysis to a project that you are working on in another course. See me to discuss this option.

Data Sources

Here are just a few sites that are possible sources of data. You can find many others through Internet searches.

http://www.data.gov – lots of U.S. government data, census info, etc.
http://www.statsci.org/datasets.html
http://lib.stat.cmu.edu/DASL/
http://library.med.cornell.edu/Guides/findingstatistics.html
http://tstation.info – data on neighborhoods around MBTA stations
http://www.reddit.com/r/dataisbeautiful/wiki/index
http://zoomprospector.com – economic and demographic data on U.S.cities and towns
Data sets in R: data(), then help(*data set name*)
http://lib.stat.cmu.edu/DASL/
http://stat.ethz.ch/R-manual/R-patched/library/datasets/html/00Index.html This is the R Datasets Package
http://archive.ics.uci.edu/ml/ These data sets are mainly for machine learning, but if anyone wants to analyze these sets using machine learning you are more than welcome.
http://www.statsci.org/datasets.html
http://library.med.cornell.edu/Guides/findingstatistics.html
http://www.itl.nist.gov/div898/strd/