Introduction to Biostatistics
Syllabus

- Review
- Questions?
Questions

- Website or Email list – Subir
- STATA – Jeff
- Homeworks and Graduate Projects
  - Specific TA for each assignment
    - Check the website
- Distance issues – Lynne
- Course issues - Scott
Homework

- Register (if not already)
- Get textbook
- Visit the course website and get notes and other vital information
- Get on the course email list
- Get access to statistical software
  - STATA for most students
Statistician’s Image

- Dull, dry, humorless
- Speaks in technical jargon that no one understands
- Wears thick glasses and carries a calculator in the pocket
- Inflexible (… always says “You can’t do that!”)
Statistician’s Image

- Spends Saturday nights in the library
- Favorite Movie: Revenge of the Nerds
- Doesn’t play golf
- A necessary evil
- SMART!
“A statistician is a person that is good with numbers but that lacks the personality to become an accountant.”
The opposite sex ignores us because we are boring.
In God we trust.
All others must bring data.
Challenges

- Statistical ideas can be difficult and intimidating
- Thus:
  - statistical results are often “skipped-over” when reading scientific literature
  - Data is often mis-interpreted
Mis-Interpretation of Data

“On average, my class is doing well. Half of my students think that 2+2=3, the other half thinks that 2+2=5.”
You may think that:

- A Bar Chart is a map of the locations of the nearest taverns
- A p-value is the result of a urinalysis
- Martingale residuals are the droppings of a rare bird
- A t-test is a blinded taste test between black and green tea
Data

- Pieces of information
- Scales of Measurement
  - Nominal – unordered categories
  - Ordinal – ordered categories
  - Discrete – only whole numbers are possible, order and magnitude matters
  - Continuous – any value is conceivable
Data

- The vast majority of errors in research arise from a poor planning (e.g., data collection)
- Fancy statistical methods cannot rescue garbage data.
- Collect exact values whenever possible.
Statistics

- The science of collecting, monitoring, analyzing, summarizing, and interpreting data.
  - This includes design issues as well.
Biostatistics

- Statistics applied to biological (life) problems, including:
  - Public health
  - Medicine
  - Ecological and environmental
- Much more statistics than biology, however biostatisticians must learn the biology also.
What Do Biostatisticians Do?

- Identify and develop treatments for disease and estimate their effects.
- Identify risk factors for diseases.
- Design, monitor, analyze, interpret, and report results of clinical studies.
- Develop statistical methodologies to address questions arising from medical/public health data.
Why Can it be Interesting?

- Combines rigors of mathematics with uncertainties of the real world.
- Can make contribution to advancement of science, statistics, medicine, and public health.
- Can study diseases/health problems in which you may have an interest (cancers, HIV, reproductive health, ...).
A Challenge

- Much of life is composed of a systematic component (i.e., signal) and a random component (i.e., error or noise).

- Example:
  - Smoking is associated with lung cancer.
  - Yet not everyone that smokes, gets lung cancer, and not everyone that gets lung cancer, smokes.
  - Yet we know that there is an association (a systematic component)
A Challenge

- Our challenge is to identify the systematic component (separate it from the random component), estimate it, and perhaps make inferences with it.
The Big Picture

Populations and Samples

Population Parameters:
\( \mu, \sigma, \sigma^2 \)

Sample / Statistics:
\( \bar{x}, s, s^2 \)
Populations and Parameters

- Population – a group of individuals that we would like to know something about
- Parameter - a characteristic of the population in which we have a particular interest
  - Often denoted with Greek letters (, \( \theta \), \( F \), \( D \))
  - Examples:
    - The proportion of the population that would respond to a certain drug
    - The association between a risk factor and a disease in a population
Samples and Statistics

- Sample – a subset of a population (hopefully representative)
- Statistic – a characteristic of the sample
  - Examples:
    - The observed proportion of the sample that responds to treatment
    - The observed association between a risk factor and a disease in this sample
Populations and Samples

- Studying populations is too expensive and time-consuming, and thus impractical
- If a sample is representative of the population, then by observing the sample we can learn something about the population
  - And thus by looking at the characteristics of the sample (statistics), we may learn something about the characteristics of the population (parameters).
Statistical Analyses

- Descriptive Statistics
  - Describe the sample

- Inference
  - Make inferences about the population using what is observed in the sample
  - Primarily performed in two ways:
    - Hypothesis testing
    - Estimation
Issues

- Samples are random
  - If we had chosen a different sample, then we would obtain different statistics (sampling variation or random variation)
  - However, note that we are trying to estimate the same (unchanged) population parameters.
Step I – Descriptive Statistics

Describe the Sample

Begin one variable at a time.
Nominal Data

- Mutually exclusive unordered categories
- Examples
  - Sex (male, female)
  - Race/ethnicity (white, black, latino, asian, native american, etc.)
- Can summarize in:
  - Tables – using counts and percentages
  - Bar Chart
Nominal Data

- Example Table
  - Insert accrual_site_treatment
- Bar Graphs
Ordinal Data

- Ordered Categories
- Examples
  - Injury – mild, moderate, severe
  - Income – low, medium, high
Discrete Data

- If there are many different discrete values, then discrete data is often treated as continuous.
- If there are very few discrete values, then discrete data is often treated as ordinal.
Continuous Data

- Any value on the continuum is possible (even fractions or decimals)
- Examples:
  - Height
  - Weight
  - Many “discrete” variables are often treated as continuous
    - Examples: CD4 count, viral load
Survival Data

- **Time to an event (continuous variable)**
  - The event does not have to be survival
- **Concept of “Censoring”**
  - If we follow a person until the event, then the survival time is clear.
  - If we follow someone for a length of time but the event does not occur, the time is censored (but we still have partial information; namely that the event did not occur during the follow up period).
- **Examples:** time to response, time to relapse, time to death
Dataset Structure

- Think of data as a rectangular matrix of rows and columns.
- Rows represent the “experimental unit” (e.g., person)
- Columns represent “variables” measured on the experimental unit
- Insert dataset_example
Data Summaries

- It is ALWAYS a good idea to summarize your data
  - You become familiar with the data and the characteristics of the people that you are studying
  - You can also identify problems with data collection or errors in the data (data management issues).
Visual Data Summaries

- Some visual ways to summarize data (one variable at a time):
  - Tables
  - Graphs
    - Bar charts
    - Histograms
    - Box plots
Frequency Tables

- Summarizes a variable with counts and percentages
- The variable is categorical
  - Note that you can take a continuous variable and create categories with it
    - How do you create categories for a continuous variable?
      - Choose cutoffs that are biologically meaningful
      - Natural breaks in the data
      - Precedent from prior research
Example: Serum Cholesterol Levels

- Insert Cholesterol_Example
- How to choose the categories?
  - Talk to physician about risk categories
  - May look at National Cholesterol Education Program (NCEP) guidelines and categories
Graphical Summaries

- **Histograms**
  - Continuous or ordinal data on horizontal axis

- **Bar Graphs**
  - Nominal data
    - No order to horizontal axis

- **Box Plots**
  - Continuous data
More Examples

- Insert accrual_by_month
- Insert More_Examples