Lecture 1
Outline

Course Overview
- Logistics
- Computing Info
- Objectives

EDA and Descriptive Statistics
- Tables
- Graphs for Categorical Variables
- Histogram
Part I

Course Overview
Logistics

▶ Mailing list: pubh6450-list@lists.umn.edu. You are encouraged to use it for all course-related communications that you wish to share with the whole class.

▶ For private communications to me, use my personal email address: brad@biostat.umn.edu. I check my email every day – some would say obsessively.

▶ Class website:
  http://www.biostat.umn.edu/~brad/ph6450.html/

▶ Office Hours:
  ▶ Brad: Tu-Th 3:30-5:00, in Mayo A427 (Brad’s office)
  ▶ TAs: TBA (see course website), in Mayo A452 (TA room)
Grading

- There will be 12 weekly homeworks (from which you can drop two worst scores) and 3 exams.
- Homeworks will be graded on the scale of 6 to 10 if you turn in anything.
- You are encouraged to discuss homeworks with other students, the TAs, and the instructor.
- All 3 exams (2 plus the final) will be *open book* and *open notes*.
- If you have a problems with the grading or with the instruction, please see your TA; if the problem remains, please see the instructor. I am very eager to resolve any problems that arise in a timely fashion.
Computing Information

- The primary computing environment will be PC SAS as available at the Mayo Computing Lab (Mayo C381).
- If you need access to SAS elsewhere, you can find it at several campus PC labs (Coffman basement, SPH student lounge, etc.) or you can purchase it from ADCS for $100; see http://www.biostat.umn.edu/~susant/PH6450SAS.html
- As a last resort, you can use the Unix version of SAS by telnet to the biostat workstation “sатурна” (instructions on the class website).
- In the lab sessions, we will go through the computing part of the homework using PC SAS. Detailed instructions for using Unix SAS will also be available on the web.
Statistical Software

- SAS (http://www.sas.com)
- SPSS (http://www.spss.com/spssbi/spss/)
- Stata (http://www.stata.com)
- Splus (http://www.insightful.com/products/splus/)
- R (http://www.r-project.org) – **Free!**
PC SAS (interactive mode):

Pros: more intuitive; can do most analysis via “point-and-click”; online help is readily available.

Cons: you need sit in front of the PC that has SAS installed.

Unix SAS (batch mode):

Pros: you can access SAS on saturn remotely; fewer “bells and whistles”.

Cons: Some working knowledge of UNIX required; no menu system; can’t see graphics directly.
A screenshot of PC SAS session. Three windows are shown: Editor, Log, and Browser; Output window is hidden.
Working with PC SAS

The “script-driven” method involves iterative steps:

1. Edit the SAS program in the Editor window.
2. “Submit” the program for execution.
3. Examine the log information in the Log window.
4. If something is wrong, “recall” the program in the editor window and go to step 1.
5. Otherwise, check the results in the Output window.
1. Edit the SAS program using any text editor (say, emacs or pico), and save it in a file example.sas.
2. Execute the program by typing: sas example.sas.
3. Check the log file example.log.
4. If something is wrong, debug the program (back to step 1).
5. Otherwise, check the result in the output file example.lst.
What is Statistics

Definition

▷ Statistics is concerned with the *collection* of data and with their *analysis* and *interpretation* (Lehmann and Casella, 1998).
▷ Statistics is “the technology of the scientific method” (Mood, 1950).
▷ Biostatistics is “the fundamental language of public health” (Louis, 1991!)
Science and Statistics

- Biostatistics/Biometry/Biometrics
- Econometrics
- Psychometrics
- Environmetrics
- Pharmacometrics
- Chemometrics
- Technometrics (engineering)
Objectives

▶ Techniques (point estimates, confidence intervals, and hypothesis tests)
▶ Computing (SAS: graphical displays, descriptive and inferential summaries)
▶ Reasoning (the “art and science of data analysis”)
▶ Asking for help: when and how.
Aspects of Statistical Reasoning

- formulate and ask the right question(s)
- construct testable subject matter hypotheses
- choose the right study design
- collect data effectively
- choose appropriate statistical methods
- summarize and interpret information
- present study results
- understand the limitations of your inferences!
Part II

Exploratory Data Analysis and Descriptive Statistics
Tables

Tips:

▶ Use sentence structure for displaying 2 to 5 numbers, tables for displaying more *numerical* information, and graphs for complex *relationships*.

▶ A working definition of a “good table”: its structure is obvious once you have described it.
Making Good Tables

Tips:

▶ Arrange the rows and columns in a meaningful way.
▶ Limit the number of significant digits (the rule of 2 informative digits).
▶ Make the table as self-contained as possible.
▶ Use white space and lines to organize rows and columns (don’t use too many lines).
▶ Use the table caption to convey crucial information.
Frequency Table

Example
Number of major hurricanes from 1944 to 2000 (MM Table 1.7)

<table>
<thead>
<tr>
<th>Year</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1944</td>
<td>3</td>
</tr>
<tr>
<td>1945</td>
<td>2</td>
</tr>
<tr>
<td>1946</td>
<td>1</td>
</tr>
<tr>
<td>1947</td>
<td>2</td>
</tr>
<tr>
<td>1948</td>
<td>4</td>
</tr>
<tr>
<td>1949</td>
<td>3</td>
</tr>
<tr>
<td>1950</td>
<td>7</td>
</tr>
<tr>
<td>1951</td>
<td>2</td>
</tr>
<tr>
<td>1952</td>
<td>3</td>
</tr>
<tr>
<td>1953</td>
<td>3</td>
</tr>
</tbody>
</table>

...
Frequency Distribution

An alternate summary of the previous data, giving the numbers of data points (here, years) with the given hurricane count:

Example

<table>
<thead>
<tr>
<th>Count</th>
<th>No. of Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>
We will sometimes combine categories, to convey more meaningful information and ensure table is not too “sparse” (i.e., too many very small counts):

**Example**

<table>
<thead>
<tr>
<th>Count</th>
<th>No. of Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 3</td>
<td>9</td>
</tr>
<tr>
<td>1–3</td>
<td>43</td>
</tr>
<tr>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>
Tips:

▶ Try not to use pie charts.
▶ Always think of alternatives to a bar chart.
▶ *Never* use a “3-D” bar chart – misleading! (see below)
EDA and Descriptive Statistics

Graphs for Categorical Variables

Histogram

Pie chart showing the distribution of educational levels:
- Bachelor's degree: 22%
- Some college: 28%
- High school graduate: 31%
- Less than high school: 12%
- Advanced degree: 7%
Tip: A *histogram* is a good way to display the empirical frequency distribution for a continuous variable.

A “true” histogram should have total area 1.
Percent of Hispanics by States (Figure 1.4)
Percent of Hispanics by States
Tip: Sometimes it is necessary to adjust bin-width to get the best result.
Percent of Hispanics by States

Percent of Hispanic adults

Percent of Hispanic adults
What to look for in a histogram

▶ Shape
▶ Location
▶ Spread
▶ Outliers
Stem and Leaf Plot

Basically a histogram with more information!

Example
Suppose the data are ages of women diagnosed with breast cancer: 36, 39, 40, 45, 51, 52, 52, 57, 59, 60, 61, 74. Stem and leaf plot:

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>6 9</td>
</tr>
<tr>
<td>4</td>
<td>0 5</td>
</tr>
<tr>
<td>5</td>
<td>1 2 2 7 9</td>
</tr>
<tr>
<td>6</td>
<td>0 1</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

The **stem** is the leading digit (here, tens) while the **leaf** is the next most significant digit (here, ones). As with a histogram, the **shape** of the data’s distribution is still visible (tip your head sideways)!
Normal (Gaussian) Distribution

\[ \mathcal{N}(\mu, \sigma^2) \] density:

\[ f(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \]

\[-\infty < x < \infty.\]

- \( \mu \): location.
- \( \sigma \): spread.
Symmetry and Modality

<table>
<thead>
<tr>
<th>0</th>
<th>2</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.5</td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.05</td>
<td>0.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>-4</th>
<th>-2</th>
<th>0</th>
<th>2</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
</tr>
</tbody>
</table>