Logistics

- Office Hours.
- Mailing list, pubh6450-list@lists.umn.edu. You are encouraged to use it for all course-related communications.
- Class website: http://www.biostat.umn.edu/~brad/ph6450.html/

Grading

- There will be about 11 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, potentially through the class mailing list.
- Exams will be open book and open notes.
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 use the Unix version of SAS by telnet to the biostat workstation “saturn” (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.
- If you are inclined you can use other statistical software, but no support for other packages is promised.

SAS: PC versus Unix

- 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.

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!

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.

Working with Unix SAS

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!

EDA and Descriptive Statistics

Part I
Exploratory Data Analysis and Descriptive Statistics

Tips:
▶ Use sentence structure for displaying 2 to 5 numbers, tables for displaying more numerical information, and graphs for complex relationships.
▶ The weak criterion for a good table: the structure should be obvious after 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.

**Example**

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Count</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**

**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>

**Summarizing Frequency Data**

**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>
EDA and Descriptive Statistics

Tables
Graphs for Categorical Variables
Histogram

Pies and Bars

Tips:
- Never use a pie chart.
- Always think of alternatives to a bar chart.
- Never use a “3-D” bar chart.

Highest level of education, Americans aged 25-34

Example

<table>
<thead>
<tr>
<th>Education</th>
<th>Count (Millions)</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than high school</td>
<td>4.7</td>
<td>12.3</td>
</tr>
<tr>
<td>High school graduate</td>
<td>11.8</td>
<td>30.7</td>
</tr>
<tr>
<td>Some college</td>
<td>10.9</td>
<td>28.3</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>8.5</td>
<td>22.1</td>
</tr>
<tr>
<td>Advanced degree</td>
<td>2.5</td>
<td>6.6</td>
</tr>
</tbody>
</table>
Histogram

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.
EDA and Descriptive Statistics

Tables
Graphs for Categorical Variables
Histogram

Percent of Hispanics by States

Tip: Sometimes it is necessary to adjust bin-width to get the best result.

What to look for in a histogram

▶ Shape
▶ Location
▶ Spread
▶ Outliers
Normal (Gaussian) Distribution

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

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

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

\(\mu\): location.

\(\sigma\): spread.

Symmetry and Modality

PH 6450 Fall 2005