Lecture Set 1 – Goals of Course and Intro to Statistics

Administrative - General

- Website: http://statweb.calpoly.edu/rottesen
- Office hours
- Homework, Quizzes and Activities
- Exams and Final
- Working together
- This class moves really fast!

Administrative – General (cont')

- The book for this course – Statistics the Exploration and Analysis of Data
  - Homework will be primarily assigned from the text
  - Quizzes will be based on the homework
  - You are responsible for keeping up with reading
  - Some sections may be covered by reading only

Administrative – General (cont')

- Two types of notes for this course
  - Type I: the power point slides that I will use during lecture (go to my website under stat217 – ppt)
    - requires current version of adobe acrobat (downloadable)
    - formatted to print handouts 6 per page
  - Type II: questions that should be answered within each lecture (go to my website under stat217 – handouts)
    - requires adobe acrobat
    - you are responsible for answering these questions (ie. you need to ask if you don’t understand or didn’t catch something!)
    - NOTE: naming convention has nothing to do with the chapter we are in

Administrative – General (cont')

- Statistical software used this course – Minitab
  - Can be found in: the statistics studio (02-208); Liberal Arts Lab (10-128); Library (35-114c); Business (03-3rd floor)
  - For help go to my website under stat217 – MTB (try this first) or use the help menu in Minitab

My Vision

- Optimal strategy for this course
  - A single and organized binder dedicated to only this course
  - Read assigned chapters before each class
  - Come to class with notes printed and reviewed so you know what questions to ask
  - Work on homework as much as possible every day
    - After you’ve completed each problem ask yourself “what does this mean?”
  - Don’t spend hours struggling, come to office hours or make an appointment with me if you need help
  - Stay on top of the material!
**Statistics**

• What do you think of when you hear “statistics”?
• Three important reasons to learn statistics:
  – To be informed
  – To understand issues and be able to make decisions
  – To be able to evaluate decisions that affect your life
• **Statistics** is the scientific discipline that provides methods to help us make sense of data
• To utilize statistics we need to understand:
  – how the data was collected
  – why it was collected
  – how to analyze and interpret the data
  APPROPRIATELY!

---

**Statistics Example**

Example: A researcher collected data on nutritional information and grocery shelf location for 76 breakfast cereals. Her research goal was to try to identify how grocery stores might be marketing cereals to children based on which are the best and worst cereals for kids to consume.

What points seem important or may be missing from this information?

---

**Statistics Example (cont’)**

Suppose that the data on sugar content (g) were:

<table>
<thead>
<tr>
<th>Shelf 1</th>
<th>Shelf 2</th>
<th>Shelf 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Sample Size: 19, 21, 36

Mean: 5.11, 9.62, 6.53

– What are the important features of this data?
– Can we say that one shelf contains definitively better cereals for kids?

---

**Statistics Example (cont’)**

Another look at the data from a visual standpoint:

– Are there any aspects of the data that make you question whether a true difference in the shelves exists?

---

**Statistics Example**

Example: A poll was conducted during September 6-8, 2005 by Ipsos Public Affairs where 1,002 adults were asked information regarding voting affiliation and opinion about the government. Results from the survey showed that 32% thought that the country is on the right track, 65% wrong track, and 3% were unsure.

What points seem important or may be missing from this information?

---

**Statistics Example (cont’)**

The margin of error was reported to be ± 3.1%.

Therefore, researchers concluded that about 62 to 68% of Americans feel that the country is on the wrong track.

Statistically speaking what does this mean?

Would the conclusion be as interesting if the margin of error were ± 25%?

Would the conclusion be as interesting if the percentage who felt that the country was on the wrong track was 51% ± 3.1%?
Variability

- When collecting data there is always going to be variability in the measurements that are collected.
  - Think of an example where there is no variability in the data collected.
- We need to be able to understand this variability in order to properly collect, analyze and interpret data.

Statistical Jargon

- **Definition**: *A population* is an entire collection of individuals or objects about which information is desired.
- **Definition**: *A sample* is a subset of the population, selected for study in some prescribed manner.

![Population and Sample diagram]

Population

Sample

Sample size is denoted by $n$.

Statistical Jargon (con't)

- **Definition**: *Descriptive statistics* consists of numerical, graphical, and tabular methods for organizing and summarizing data.
- **Definition**: *Inferential statistics* involves generalizations from a sample to the population from which it was selected.
- Were the last two examples (cereal and political) examples of descriptive or inferential statistics?
- How can you tell?

Variables

- There are two types of variables:
  1. categorical (aka qualitative)
  2. numerical (aka quantitative)

Variables (con't)

- **Categorical (qualitative) variables** are variables that are classified into groups.
- There are two types of categorical variables:
  - Ordinal (arranged in a meaningful order)
  - Not ordinal (no meaningful order)
- **What type of categorical variable** are following:
  - gender (M/F)?
  - size of soda (small, medium, large)?
  - political affiliation (democrat, republican, independent, green party, other)?
Variables (cont')

- Quantitative variables are variables that have a meaningful numerical value.
- There are two types of quantitative variables:
  - Continuous (lies on an interval scale with infinite possible values)
  - Discrete (space between each value, countable)
- What type of quantitative variable are following:
  - weight (lbs.)?
  - height (in.)?
  - number of cars in the library parking lot?

Using Statistical Jargon

Example: Most breast cancer patients (>80%) are over the age of 50 at diagnosis. A researcher at a particular New York cancer center believes that his patients are even older than the norm, typically older than 65 years at diagnosis. To investigate he reviews the ages of a random sample of 100 of his female patients diagnosed with breast cancer.

Using Statistical Jargon (cont')

- Identify the following:
  - Population
  - Sample
  - Sample size
  - Variable of interest
    - numerical or categorical?
  - Other variables
    - numerical or categorical?
  - Individuals measured

Describing Data

- There are two ways to describe a data set:
  - Graphs and tables
  - Numerically
- Both are important for analyzing data

Graphs and Tables Intro

- **Definition:** A frequency distribution is a table that displays the possible categories along with their associated frequencies (counts)
- **Definition:** A relative frequency distribution is a table that displays the possible categories along with their associated frequencies (percents = counts/n)

<table>
<thead>
<tr>
<th>Stage</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>197</td>
<td>0.09</td>
</tr>
<tr>
<td>I</td>
<td>691</td>
<td>0.33</td>
</tr>
<tr>
<td>II</td>
<td>753</td>
<td>0.34</td>
</tr>
<tr>
<td>III</td>
<td>314</td>
<td>0.15</td>
</tr>
<tr>
<td>IV</td>
<td>157</td>
<td>0.08</td>
</tr>
<tr>
<td>Total</td>
<td>2092</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Graphs and Tables Intro (cont')

- Categorical variables
  - Easier to deal with than quantitative variables
Example: Stage of disease at diagnosis of breast cancer in a random sample of US women.
• Example: Stage of disease (cont'):

<table>
<thead>
<tr>
<th>Stage</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>200</td>
</tr>
<tr>
<td>6</td>
<td>300</td>
</tr>
<tr>
<td>7</td>
<td>400</td>
</tr>
<tr>
<td>8</td>
<td>500</td>
</tr>
<tr>
<td>9</td>
<td>600</td>
</tr>
<tr>
<td>10</td>
<td>700</td>
</tr>
</tbody>
</table>

Chart of Frequency vs Stage

• A widely used graphical display of numerical data is called a dot plot.
  – Looks just like its name

<table>
<thead>
<tr>
<th>Calories</th>
</tr>
</thead>
<tbody>
<tr>
<td>192</td>
</tr>
<tr>
<td>176</td>
</tr>
<tr>
<td>160</td>
</tr>
<tr>
<td>144</td>
</tr>
<tr>
<td>128</td>
</tr>
<tr>
<td>112</td>
</tr>
</tbody>
</table>

Dotplot of Calories

• Does the dot plot seem to accentuate anything interesting about the hotdog data?

<table>
<thead>
<tr>
<th>Type</th>
<th>Calories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beef</td>
<td>192</td>
</tr>
<tr>
<td>Meat</td>
<td>176</td>
</tr>
<tr>
<td>Poultry</td>
<td>160</td>
</tr>
</tbody>
</table>

Dotplot of Calories vs Type

• Numerical variables
Example: People who are concerned about their health may prefer hot dogs that are low in salt and calories. The "Hot dogs" datafile contains data on the sodium and calories contained in each of 54 major hot dog brands. The hot dogs are also classified by type: beef, poultry, and meat (mostly pork and beef, but up to 15% poultry meat). For now we will focus on the calories of these sampled hotdogs.