Topic 1 - Introduction

STAT 511
Professor Bruce Craig

Background Reading
Devore : Section 1.1

Outline

- Class Policies / Schedule
- Class Website
- Overview of Material
- Statistical Software

Class Website
www.stat.purdue.edu/~bacraig/stat511.html

- Course syllabus / Announcements
- Exam and homework schedule
- Lecture notes
- Homework assignments
- Data sets for class and homework
- Sample SAS programs
- Information about group projects

Class Policies

- Attendance
  - Not required but you are responsible for announcements and lecture material
  - If you have to leave early or arrive late, notify me in advance and sit near door
- Class participation encouraged
- Questions welcome at all times
Exams

- There will be two evening exams and a final
  - Must notify me early if scheduling conflict (set up alternate time)
  - Will be allowed one-page of notes and calculator
  - I will supply necessary tables
  - Each worth 20% of your grade

Homework

- Expect “weekly” homework assignments
  - Will be due Tues before class
  - Individual vs group effort
  - Worst grade will be dropped
  - Represents 25% of your grade
  - Answer key posted after due date

Project

- Group / Team projects
  - Teams determined after week 3
  - Will analyze more complicated data sets / perform simulations
  - Represents 15% of your grade
  - Check web site for upcoming details

Communication

- Office Hours
  - Mon 3:00-4:30
  - Thu 11:45-12:45
  - By appt.
  - By email - bacraig@stat.purdue.edu
Statistical Software

- Computer Software
  - Will use PC-SAS in class
  - Available on computer lab machines
  - Can get copy and put on home PC
  - Syntax Help / Examples available
  - Will provide template programs
  - Software Consulting Service (MATH G175)
  - Computer lab will be reserved M-R
  - Can use other programs at own risk

Overview

To learn a variety of ways to describe data; To understand the mathematical foundation of statistical inference for random samples; To conceptually understand the application of statistical methods such as hypothesis tests, regression, and ANOVA; To properly apply these methods to real world problems using statistical software and draw valid conclusions; To obtain the background necessary to study additional statistical methods in the future.

Schedule

- Descriptive Statistics (1.5 wks)
- Mathematical/Probabilistic Foundation (1.5 wks)
- Probability Distributions (2 wks)
- Sampling Distributions (2 wks)
- Statistical Inference (1 sample) (2 wks)
- Statistical Inference (2 samples) (1 wk)
- Analysis of Variance (1.5 wks)
- Linear Regression (1.5 wks)
- Categorical Data (2 wks)

What is Statistics?

- Science of learning from information (data)
  - Collection
  - Description
  - Interpretation

- Bombarded with statistics daily
  - Polls (e.g., Nielsen rating, popularity/opinion)
  - Sports (e.g., FT percentage, batting average)
  - Economic indicators (e.g., GNP, CPI)
  - Health (e.g., cohort studies, risk factors)
**Why Statistics?**

- Cannot escape interpretations of data
- Numerical literacy important to everyone
  - Express your conclusions numerically
  - Critically read and comprehend other results
  - Develop sound methods for trustworthy results
- Uncertainty exhibited in most processes
  - Different properties of "similar" material
  - Uncontrollable/controllable factors
  - Measurement/human error

**Background Information**

In addition to the data themselves, should also consider the science of problem

- **What purpose do the data have?**
  - Want to answer specific questions/hypotheses
  - Want to generalize results to population (polls)
- **What is the population of interest?**
  - Is the data set "representative" of population
  - How is the "representative" sample obtained
  - How many individuals in data set
- **What is the information?**
  - What variables are in the data set
  - What are the exact definitions of variables
  - What are the units of measure

**Definitions**

- **A data set** is a collection of observations
- Observations concern a set of **individuals**
- **Individuals** can be
  - People, objects, or animals
  - Experimental trials (under identical conditions)
- **Information** is organized as set of variables
- **Variable** is a characteristic of an individual
  - Person - Hair color, height, blood type
  - Object - Density, strength, concentration
  - Trial - H/T of coin flip
Definitions

- An **observation** is a realization of a variable.
  - Hair color → blond
  - Height → 62 inches

- **Univariate** data set consists of values from one variable

- **Multivariate** data set consists of values from more than one variable

Examples

- Will denote variables by upper case letters ($X$)
- Will denote observations by lower case letters ($x$)

The number of accidents on Interstate 65 in the month of December was 10

Individual = Interstate 65  
Variable $Y$ = the number of accidents in Dec  
Observation $y = 10$

The following table (WSJ - 1997) summarizes a week’s beer advertising on TV

<table>
<thead>
<tr>
<th>Advertiser</th>
<th>Show (Network)</th>
<th>Date (Time)</th>
<th>% Viewers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coors Light</td>
<td>Hit List (BET)</td>
<td>Sept 2 (8:00 pm)</td>
<td>51</td>
</tr>
<tr>
<td>Molson</td>
<td>Singled Out (MTV)</td>
<td>Sept 2 (7:00 pm)</td>
<td>52</td>
</tr>
<tr>
<td>Molson Ice</td>
<td>Beavis and Butthead (MTV)</td>
<td>Sept 2 (11:30 pm)</td>
<td>48</td>
</tr>
<tr>
<td>Foster’s</td>
<td>Singled Out (MTV)</td>
<td>Sept 3 (11:00 pm)</td>
<td>46</td>
</tr>
<tr>
<td>Molson</td>
<td>Real World (MTV)</td>
<td>Sept 3 (8:30 pm)</td>
<td>45</td>
</tr>
<tr>
<td>Foster’s</td>
<td>Melrose Place (E!)</td>
<td>Sept 2 (7:00 pm)</td>
<td>41</td>
</tr>
<tr>
<td>Miller</td>
<td>Unreal (BET)</td>
<td>Sept 5 (8:00 pm)</td>
<td>65</td>
</tr>
<tr>
<td>Schlitz</td>
<td>Yo MTV (MTV)</td>
<td>Sept 5 (10:00 pm)</td>
<td>50</td>
</tr>
<tr>
<td>Molson</td>
<td>Beavis and Butthead (MTV)</td>
<td>Sept 6 (10:30 pm)</td>
<td>69</td>
</tr>
<tr>
<td>Budweiser</td>
<td>Video Music Awards (MTV)</td>
<td>Sept 7 (8:30 pm)</td>
<td>46</td>
</tr>
</tbody>
</table>

What are the variables and observations?

Inferential Statistics

- Study often focused on well-defined collection of indivs known as the **population**

- Cannot study entire population (census)

- Want to generalize **sample** to population

- Process of drawing conclusions about population is called **statistical inference**

- Conclusions valid only if sample considered representative of population
  - Representative : closely resembles population
  - When sample non-representative, results biased
  - Strive for an unbiased sample of population

Defining the Population

- Should be done prior to collecting sample

- Should always consider
  - How will individuals be selected?
  - What variables are of interest?

**Example:** Want to determine average level of radon in homes built on reclaimed land. Plan to select 50 homes and measure radon level. How to select homes?

  - Type of house?
  - Location of house?
  - Age of house?