Stat 3660: Introduction to Statistics
Summer 1 2014 Syllabus

Chapters 1 – 4
Introduction to Statistics
Agenda – Part 1

• General introductions and review of syllabus (what are your expectations of this class?)
• Introduction to statistics (what is statistics?)
• Controlled Experimentation
• Observational Studies
• Simpson’s Paradox
• Issues Related to Statistical Studies
Chapter 1 – Introduction to Statistics

• Statistics is about obtaining and analyzing data

• Statistics the ART and SCIENCE of
  – Making conclusions (decisions and judgments) from data (numbers, information)
Chapter 1 – Introduction to Statistics

• The SCIENCE part of statistics deals with
  – Methods and procedures for getting information (data) and analyzing it

• The ART part of statistics comes in when we realize that elements of this process cannot be made objective and we have to rely on intuition and judgment to some degree
  – Like Charlie the tuna said, “it’s a matter of ‘good taste’”
Statistics in Health and Medicine

• Clinical Trials
  – Evaluate safety and effectiveness of new drugs or therapies

• Epidemiology
  – Investigate distribution and determinants of health and disease

• Biostatistics
  – Development and application of data-analytic techniques to health related research
Types of Studies

1. Randomized studies (Controlled Experiments)

2. Observational studies

3. Anecdotal studies (no comparison)
Chapter 2 – Controlled Experiments

• Study whether a treatment causes a change in response
  – Testing new procedures
  – Testing new drugs

• Principals of good experimental design:
  1. Make a comparison
     ✓ setup a treatment group and a control group)
  2. Randomize the group membership of the subjects
  3. Make the experiment blind
     ✓ Subjects don’t know which group they are in
  4. Make the experiment double blind
     ✓ Those assessing the results don’t know group identify
Example of Controlled Experiments

• We want to test a new drug that helps controls a skin rash

• Select 200 people who have this particular skin rash and randomly divide these people into two groups
  – 100 people will receive the new drug (treatment)
  – 100 people will receive an old drug (control)

• We also perform a double-blind experiment
  – Patients and doctors checking on the skin rash do not know which group the patient is in
Randomization

- Among 200 people, randomize into Treatment and Control groups

Toss Coin

0.5
Treatment

0.5
Control

Age, Diet, Exercise Habits, Attitudes, etc. should all be relatively balanced in each group
Example of Controlled Experiments

• After a period of time we record either
  – A ‘success’, no skin rash present
  – Or ‘failure’, skin rash still present, for each person

• The data may look like

<table>
<thead>
<tr>
<th></th>
<th>Treatment Group</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>75% success</td>
<td>60% success</td>
<td></td>
</tr>
</tbody>
</table>

• We can’t make any statistical conclusions yet (we don’t know these methods yet), but we will learn them soon enough
Chapter 2 – Controlled Experiments

• In comparing two groups the goal is to have
  – ALL THINGS BEING EQUAL BETWEEN THE TWO GROUPS EXCEPT THE
    PRESENCE OF THE TREATMENT.
    ✓ Here the word equal is intended to mean homogeneous (homogeneity),
      that is that there is no significant difference between the to groups (e.g.,
      one group has more males, etc)

    ✓ The test can then be optimized to reduce uncontrolled (unexpected)
      variation that cannot be eliminated.

    ✓ This tends to reduce experimental error and can make decisions and
      predictions more accurate and repeatable.
Chapter 3 – Observational Studies

- Often we have data that was collected in a more haphazard manner
- We observe results that are happening in the world without any explicit control on our part
- There are potential troubles in analyzing observational data as
  - other factors that we have not controlled may be affecting the main response
  - Biases may be built into the data
  - Seemingly obvious conclusions can be wrong
- Sometimes we can partially correct for such potential biases with various adjustments or by stratification
Example of Observational Studies

• In an Australian study, researchers found that for men, an additional 2 inches of height brings an additional $1,000 in annual income

• This ‘height premium’ is less for women in Australia, while for men in the UK and USA is higher

• There are many proposed theories to explain this phenomena, but that data for this study was observational and obtained in some vague fashion
Example: Grade School Study

‘In a large study of children in grades 1-3, it was found that children who weighed less than 40 pounds knew 50% few words than children who weighed over 40 pounds.’

Possible Scenarios for this study

A. Weight → Vocabulary
B. Vocabulary → Weight
C. Age → Vocabulary
D. More...
Example: Grade School Study

• ‘Age’ is a **lurking variable** here

• The effects of ‘Age’ and ‘Weight’ are **confounded**
Grade School Study

- **Expert Conclusion:**
  - Age is the variable which ‘derives’ the relationship between Vocabulary and Weight (or Height, Shoe Size)

- **Statistical techniques cannot tell cause-and-effect. On paper, the following plots all look the same**
  - Vocabulary vs. Weight
  - Vocabulary vs. Height
  - Vocabulary vs. Age

- **Moral:**
  - Data Analysis should include numerical methods as well as information about the subject matter
Example: Does Antibiotic Exposure During Infancy Lead to Development of Asthma?

Marra F. et al., J. of the ACCP, March 2006, pp. 610-618

• Objective:
  – To determine the association between antibiotic exposure in the first year of life and the development of childhood asthma.

• Conclusion:
  – Exposure to at least one course of antibiotics in the first year of life appears to be a risk factor for the development of childhood asthma.
Example: Does Antibiotic Exposure During Infancy Lead to Development of Asthma?

• Possible Scenarios for this study
  A. Antibiotic $\rightarrow$ Asthma
  B. Asthma $\rightarrow$ Antibiotic
  C. Infections, Allergies $\leftrightarrow$ Antibiotics $\leftrightarrow$ Asthma

• In cause-and-effect studies, beware of *direction* and *lurking variables*
Simpson’s Paradox (comparing aggregate scores)

- The local newspaper examined two hospitals in our town and found that over the last six months
  - 80% of the patients survived at Hospital A
  - while 90% survived at Hospital B

<table>
<thead>
<tr>
<th></th>
<th>Lived</th>
<th>Died</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital 1</td>
<td>800</td>
<td>200</td>
<td>1000</td>
</tr>
<tr>
<td>Hospital 2</td>
<td>900</td>
<td>100</td>
<td>1000</td>
</tr>
</tbody>
</table>
Simpson’s Paradox (comparing aggregate scores)

- On closer investigation, the patients were categorized by their condition.

<table>
<thead>
<tr>
<th>Overall</th>
<th>Lived</th>
<th>Died</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital 1</td>
<td>800</td>
<td>200</td>
<td>1000</td>
</tr>
<tr>
<td>Hospital 2</td>
<td>900</td>
<td>100</td>
<td>1000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fair condition</th>
<th>Lived</th>
<th>Died</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital 1</td>
<td>590</td>
<td>10</td>
<td>600</td>
</tr>
<tr>
<td>Hospital 2</td>
<td>870</td>
<td>30</td>
<td>900</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Poor condition</th>
<th>Lived</th>
<th>Died</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital 1</td>
<td>210</td>
<td>190</td>
<td>400</td>
</tr>
<tr>
<td>Hospital 2</td>
<td>30</td>
<td>70</td>
<td>100</td>
</tr>
</tbody>
</table>
Simpson’s Paradox

<table>
<thead>
<tr>
<th></th>
<th>Lived</th>
<th>Died</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital 1</td>
<td>590</td>
<td>10</td>
<td>600</td>
</tr>
<tr>
<td>Hospital 2</td>
<td>870</td>
<td>30</td>
<td>900</td>
</tr>
</tbody>
</table>

- Among patients in fair condition,
  - 98.3% of the patients survived at Hospital A
  - while 96.6% survived at Hospital B

<table>
<thead>
<tr>
<th></th>
<th>Lived</th>
<th>Died</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital 1</td>
<td>210</td>
<td>190</td>
<td>400</td>
</tr>
<tr>
<td>Hospital 2</td>
<td>30</td>
<td>70</td>
<td>100</td>
</tr>
</tbody>
</table>

- Among patients in poor condition,
  - 52.5% of the patients survived at Hospital A
  - while 30% survived at Hospital B
Chapter 4 – Issues Related to Statistical Studies

• Key stages of a Study:
  1. Statement of Objective
  2. Identify the target Population(s)
  3. Select the Frame(s)
  4. Select the Sampling and Measurement Methods
  5. Organization of Data Management
  6. Perform Appropriate Statistical Analysis(es) to support the original objective of the study
  7. Draw conclusions and decide on follow-up as appropriate
Example: Key stages of a Study

• We want to test a new drug that helps controls a skin rash

• Select 200 people who have this particular skin rash and randomly divide these people into two groups
  – 100 people will receive the new drug (treatment)
  – 100 people will receive an old drug (control)

• We also perform a double-blind experiment
  – Patients and doctors checking on the skin rash do not know which group the patient is in
Objectives – Part 2

• By the end of this material, you will be able to:
  – Distinguish between three applications of statistics and identify situations in which each is appropriate
  – Distinguish between discrete and continuous variables and cite examples of each
  – Identify and describe three levels of measurement and cite examples of variable from each type.
The Basic Types of Statistics

• Descriptive Statistics (Chapters 2 – 5)
  – Summarizing or describing the distribution of a single variable (univariate) through calculation of relevant statistics (percents, ratios, means, standard deviations, etc.), or graphical methods.

• Inferential Statistics (Chapters 6 – 11)
  – Inferential statistical techniques use information (estimation and tests of hypotheses) from a sample to generalize to a population.

• Bivariate and Multivariate Descriptive Statistics (Chapters 12 – 16)
  – Understanding the relationship (measures of association) between two or more variables allows us to quantify the strength and direction of the relationship.
Types of Variables (data)

Mind Map of Data Types

- **Random Variable**
  - **Categorical**
    - **Ordinal**
      - with order
      - eg. military rank
    - **Nominal**
      - without order
      - eg. color of eyes
  - **Interval**
    - temperature
    - eg. degrees C
  - **Numerical**
    - **Continuous**
      - measuring process
      - eg. distance to GR
    - **Ratio**
      - counting process
    - **Discrete**
      - eg. number of widgets in the inventory
Types of Data

• Categorical (qualitative data)
  – characteristics that cannot be “measured”
    • Eye color
    • Gender
    • Race
    • Political affiliation

• Numerical (quantitative data)
  – Characteristics that can be measured”
    • Height, weight, etc.
    • Time it takes to perform a task
    • Counts
Categorical Data

- **Random Variable**
  - **Categorical**
    - **Ordinal** with order
      - eg. military rank
    - **Nominal** without order
      - eg. color of eyes
  - **Numerical**
    - **Interval**
      - eg. degrees C
    - **Continuous** measuring process
      - eg. distance to GR
    - **Ratio** counting process
      - Discrete eg. number of widgets in the inventory
Categorical Data

• Types of **CATEGORICAL** data

  – Since categorical data are not measurement (numerical) they cannot have numerical operations (addition, subtraction, multiplication or division) performed on them.

  – So the only available method of describing categorical data is by the frequency within a specific category (counts, proportions or percentages).
Categorical Data

- **Types of **CATEGORICAL** data**
  - There are two (2) distinct of categorical data:
  - Nominal data:
    - Nominal data are non-measurement data that **CANNOT** be ordered, e.g., eye color, race, gender, etc.
  - Ordinal data:
    - Ordinal data are non-measurement data that **CAN** be ordered, e.g., military rank, place in a race, working class, student class standing, etc.
Example of Categorical Data

<table>
<thead>
<tr>
<th>Subject</th>
<th>Gender</th>
<th>Political Affiliation</th>
<th>Level of Education</th>
<th>Age Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Female</td>
<td>Republican</td>
<td>HS Grad</td>
<td>Young Adult</td>
</tr>
<tr>
<td>2</td>
<td>Male</td>
<td>Democrat</td>
<td>Masters</td>
<td>Older Adult</td>
</tr>
<tr>
<td>3</td>
<td>Male</td>
<td>Democrat</td>
<td>HS Grad</td>
<td>Middle Age</td>
</tr>
<tr>
<td>4</td>
<td>Female</td>
<td>Republican</td>
<td>Bachelor Degree</td>
<td>Middle Age</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>Republican</td>
<td>Bachelor Degree</td>
<td>Older Adult</td>
</tr>
</tbody>
</table>

Above is a sample of 5 individuals taken from a population and the following categorical data was observed.
- Gender is nominal since there is no obvious order to the data.
- Political Affiliation is also nominal data
- Level of Education is ordinal data as Bachelor Degree comes after HS Graduate but before Masters
- Age group is also ordinal data as Middle age comes after Young Adult but before Older Adult.
Numerical Data

Random Variable

Categorical
  Ordinal
    with order
    eg. military rank
  Nominal
    without order
    eg. color of eyes

Numerical
  Interval
    temperature
    eg. degrees C
  Continuous
    measuring process
    eg. distance to GR
  Ratio
    counting process
  Discrete
    eg. number of widgets in the inventory
Numerical Data

• Types of **NUMERICAL** data
  – Numerical (or measurements) data can have numerical operations (addition, subtraction, multiplication or division)
  – Numerical summaries may be calculated (e.g., mean, median, standard deviation, variance, etc).
  – Interval and ratio data are slightly different, but statistically they are analyzed in the same manner and so we refer to numerical data.
  – There are two (2) distinct types of numerical data (discrete and continuous) and these are described on the next slides.
Numerical Data

• Discrete numerical data
  – A variable is said to be **DISCRETE** if it has a basic unit of measurement that cannot be subdivided.
  – For example, the number of people per household is a discrete variable. The basic unit is people, a variable that will always be measured in whole numbers (you should never find 2.5 people in a specific household). Below is a sample of three (3) households with discrete data obtained on them.

<table>
<thead>
<tr>
<th>Sample Household</th>
<th>Number of Pets in Household</th>
<th>Number of People in Household</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>
Numerical Data

• Notice there appears to be similarities between discrete data and summaries associated with categorical data. But, the difference becomes clear when you recall the original data where the categorical data appears like (green, blue, etc.) and the discrete data appears like (4, 3, 0, etc.).

• For example in the prior example of data on individuals the data is sometimes reports as a summary as follows:

<table>
<thead>
<tr>
<th>Gender</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>3</td>
</tr>
<tr>
<td>Female</td>
<td>2</td>
</tr>
</tbody>
</table>

• But the data is still of the form ‘male’ and ‘female’ and not numeric versus the counts associated with households which have counts associated with each sample household. This is an important distinction and understanding this distinction will improve understanding later.
Numerical Data

- Continuous numerical data
  - A variable is **CONTINUOUS** if it has scores that can be subdivided infinitely (at least theoretically). Generally continuous variables are measurements, e.g., length, weight, time, etc.
  - In a sense, when we measure a continuous variable, we are always approximating and rounding off the scores. This sometimes results in discrete and continuous variables appearing the same though we will process them differently. This distinction is very important as different statistical techniques (calculated statistics, graphic methods, etc.) will be applied based on the types of data involved.
## Types of Data (summary)

<table>
<thead>
<tr>
<th>Level of Measurement</th>
<th>Examples</th>
<th>Measurement Procedures</th>
<th>Mathematical Operations Permitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Sex, race, religion, martial status</td>
<td>Classification into categories (counts by category)</td>
<td>Counting number in each category, comparing size of categories</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Social class, attitude, opinion scales</td>
<td>Classification into categories plus ranking of categories with respect to each other.</td>
<td>All of the foregoing plus judgments of “greater than” and “less than”</td>
</tr>
<tr>
<td>Numerical</td>
<td>Age, number of children, income</td>
<td>All of the foregoing plus description of distances between scores in terms of equal units</td>
<td>All of the foregoing plus all other mathematical operations (addition, subtraction, multiplication, division, square roots, etc.)</td>
</tr>
</tbody>
</table>
QUESTIONS?