Chapter 1

Introduction to Data
Learning Objectives

- Distinguish between numerical and categorical variables.
- Find and use rates (including percentages) and understand when and why they are more useful than counts for describing and comparing groups.
- Understand when it is possible to infer a cause-and-effect relationship.
- Explain how confounding variables prevent us from inferring causation, and suggest confounding variables that are likely to occur in some situations.
- Be able to distinguish between observational studies and controlled experiments.
Statistics

...is the science of problem solving using data. A statistical investigative process involves with collecting, organizing, summarizing and analyzing information in order to draw conclusions.
Data

- **Data** are observations that you or someone else records
- **Numeric data**
  - Pain/Satisfaction scale from 1-10
  - Measurements: Weight, Height, Distance, etc.
  - Number of customers during a business day
- **Data can also be non-numeric.**
  - List of song titles stored on I-Pod
  - Ethnicities of each student on campus
  - Party preference in the upcoming election
Sample vs. Population

- The population is the collection of all data values. It includes all outcomes that have ever or could ever occur. Information about the population is usually the goal; however, obtaining all data values from the population is usually impossible.

- A sample is a subset of the population. A sample is used to get a partial understanding of the population.
Variables

A variable is a characteristic of the individuals (sampling units) within a population. Some examples are

- A person’s gender
- The weight of a newborn puppy
- The concentration of CO\textsubscript{2} in the atmosphere

The word variable comes from the fact that the measurements vary from one value to another.
The Process of a Statistical Investigation

Step 1: Identify a Research Objective

Step 2: Collect the information needed to answer the questions.

Step 3: Organize and summarize the information.

Step 4: Draw conclusions from the information.
Example

How can statistical techniques be applied to solve the following problem?

UCLA is interested in revising its recruiting policy. A study will be conducted to find out where the current UCLA students come from and the reasons for choosing UCLA.
Step 1: Identify a Research Objective

• Researcher must determine a question he/she wants answered - question must be detailed.

• Identify the group to be studied. This group is called the population.

• An individual person or object that is a member of the population being studied is the sampling unit.
Step 2: Collect the information needed to answer the questions.

• In conducting research, we typically look at a subset of the population, called a sample.

Step 3: Organize and summarize the information.

• Descriptive statistics consists of organizing and summarizing the information collected. Consists of charts, tables, and numerical summaries.
Step 4: Draw conclusions from the information.

• The information collected from the sample is generalized to the population.

• **Inferential statistics** uses methods that generalize results obtained from a sample to the population and measure their reliability.
Step 1: Identify the research objective.

Consider the Recruiting Example:
To investigate where the current UCLA students come from and the reasons for choosing UCLA. So that the information can be used for revising recruiting policy.

• Determine the target population for the study:
  ___all UCLA students____

• Determine the sampling unit for the study:
  ___1 UCLA student___

• Determine the number of individuals to be chosen for the study (sample size, \( n \)) ___500 (for example)___
Step 2: Collect the information needed to answer the question.

Develop a survey with questions asking their demographic information, and reasons for choosing UCLA.

Select a sample of individuals for the study by using random sampling technique, such as assigning an ID number for each individual in the population, randomly select $n$ individuals as the random sample.

Distribute the survey to the randomly selected individuals to collect the information needed.
Step 3: Organize and summarize the information.

An example of a summary from sample data:
The demographic information from the sample indicated 90% students are from within 200 miles of UCLA. 70% of students are from cities along 405 down south of UCLA.

Three main reasons selecting UCLA are
• It is at the right distance from home.
• Come for the programs in teacher education.
• It is less expensive than others.
Step 4: Draw conclusions from the data.

An example of drawing conclusions from the summary and potential policy revision:
Based on the summary of the random sample of 500 students, we conclude that approximately 90% of all UCLA students live with 200 miles from UCLA, and that most of them are from areas along 405. UCLA may want to provide more information to schools within 200 to 300 miles from UCLA and provide some targeted visits to schools. In addition, UCLA may want to make effort to recruit more diverse students from international community.
Another Example of a Sample and a Population

- **Population**: All babies born in NC in 2004
- **Sample**: These six babies
- **Variables**: Weight, Gender, Mother Smoked?

Birth data from North Carolina 2004

<table>
<thead>
<tr>
<th>Weight</th>
<th>Gender</th>
<th>Smoke</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.69</td>
<td>F</td>
<td>0</td>
</tr>
<tr>
<td>0.88</td>
<td>M</td>
<td>1</td>
</tr>
<tr>
<td>6.00</td>
<td>F</td>
<td>0</td>
</tr>
<tr>
<td>7.19</td>
<td>F</td>
<td>0</td>
</tr>
<tr>
<td>8.06</td>
<td>F</td>
<td>0</td>
</tr>
<tr>
<td>7.94</td>
<td>F</td>
<td>0</td>
</tr>
</tbody>
</table>
Other Questions of Interest

- **How Were the Data Collected:**
  Hospital reports to the government. Medical Care-Giver weighed the infant and surveyed the mother.

- **What Were the Units of Measurement:**
  pounds, F and M for girls and boys, 0→Nonsmoker 1→Smoker

- **Where Were the Data Collected:**
  North Carolina

- **Why Were the Data Collected:**
  To learn about infant health as it relates to mothers smoking.

### Birth data from North Carolina 2004

<table>
<thead>
<tr>
<th>Weight</th>
<th>Gender</th>
<th>Smoke</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.69</td>
<td>F</td>
<td>0</td>
</tr>
<tr>
<td>0.88</td>
<td>M</td>
<td>1</td>
</tr>
<tr>
<td>6.00</td>
<td>F</td>
<td>0</td>
</tr>
<tr>
<td>7.19</td>
<td>F</td>
<td>0</td>
</tr>
<tr>
<td>8.06</td>
<td>F</td>
<td>0</td>
</tr>
<tr>
<td>7.94</td>
<td>F</td>
<td>0</td>
</tr>
</tbody>
</table>
Two Types of Variables

- A **Quantitative or Numerical Variable** describes quantities of the objects of interest. **Data values are numbers.**
  - Weight of an infant
  - Number of sexual partners
  - Time to run a mile

- A **Qualitative or Categorical Variable** describes qualities of the objects of interest. **Data values are usually words.**
  - Skin color
  - Birth city
  - Last Name
### Example: Numerical or Categorical?

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Major</th>
<th>Units</th>
<th>Housing</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>Male</td>
<td>Psychology</td>
<td>16</td>
<td>Dorm</td>
<td>3.6</td>
</tr>
<tr>
<td>21</td>
<td>Male</td>
<td>Nursing</td>
<td>15</td>
<td>Parents</td>
<td>3.1</td>
</tr>
<tr>
<td>20</td>
<td>Female</td>
<td>Business</td>
<td>16</td>
<td>Apartment</td>
<td>2.8</td>
</tr>
</tbody>
</table>

- **Numerical**
  - Age
  - Units
  - GPA

- **Categorical**
  - Gender
  - Major
  - Housing
Numerical or Categorical?

Why are you in college? Answer:
1. Person Growth  2. Career Opportunities
3. Parental Pressure  4. Personal Networking

Results:  1, 4, 3, 2, 2, 1, 2, 3, 3, 1, 4, 2

- **Coding Categorical Data** with Numbers:
  Although the above data values are numbers, the variable is still categorical.

- **Reason for Coding**: Easier to input into a computer.
Coding Yes/No Questions

- We often use 0 for “No” and 1 for “Yes”
- Useful for data with only two possible values
  - True or False
  - Black or White
  - Success or Failure (Bernoulli)
  - Dead or Alive
  - Head or Tail (coin toss)
EXAMPLE: Distinguishing Between Qualitative and Quantitative Variables

Determine whether the following variables are qualitative or quantitative.

(a) Flavors of ice-cream.
(b) Speed of Roger Federer's serve.
(c) Number of times your Internet service went down in the last 30 days.
(d) Zip codes
Types of Quantitative Variables

A **discrete variable** is a quantitative variable that either has a finite number of possible values or a countable number of possible values. The term “countable” means the values result from counting such as 0, 1, 2, 3, and so on.

A **continuous (interval) variable** is a quantitative variable that is defined on an interval of real numbers and can be measured to any desired level of accuracy.
Types of Qualitative Variables

An **ordinal variable** is a qualitative variable that bears an ordering and can be ranked from the lowest level to the highest level. It can be easily coded into discrete numbers. For example: rating of a professor/course, customer satisfaction level.

A **nominal variable** is a qualitative variable that simply classify data into categories without any specific order. For example: ethnicity, First/Last Name, Email etc.
Two-Way Tables

Gender and Seat Belt Practices

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Always</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Always</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

- **Two-Way Tables** show how many times each combination of categories occur.
- The “2” in the above table tells us that there were two men from the sample who do not always wear seat belts.
- A frequency is the number of times the value is observed in a data set.
Two-Way Tables and Frequencies

Gender and Seat Belt Practices

- How many women were surveyed?
  - Women: $3 + 7 = 10$

- How many from the survey always wear a seat belt?
  - Always: $4 + 7 = 11$

- How many women from the survey do not always wear a seat belt?
  - Women AND Not Always: 3

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Always</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Always</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>
Two-Way Tables and Percentages

Gender and Seat Belt Practices

- What percent are men?
  - Men: \[ \frac{2 + 4}{2 + 4 + 3 + 7} \times 100\% = 0.375 \]
- What percent always wear a seat belt?
  - Always: \[ \frac{4 + 7}{2 + 4 + 3 + 7} \times 100\% = 0.6875 \]
Percentages to Counts

Of the 400 students who were surveyed 65% were carrying their calculators. Of those carrying their calculators, 40% were men.

- How many of these 400 students were carrying their calculator?
  \[ 400 \times 0.65 = 260 \]

- How many female students were carrying a calculator?
  \[ 260 \times 0.60 = 156 \]
Organizing and Reporting Categorical Data

- Use a 2-Way Table to display results of a two question survey or two outcome experiment.
- Use percents or rates rather than counts when comparing groups with different sizes.