Synthesis and Review for Exam 1

The Big Picture

Population

Sampling

Sample

Interval estimation

Descriptive statistics

Statistical Inference

Observational Studies

• A third variable that is associated with both the explanatory variable and the response variable is called a **confounding variable**
• There are almost always confounding variables in observational studies

Observational studies can almost never be used to establish causation

Sampling

Population

Sample

GOAL: Select a sample that is similar to the population
HOW? RANDOM SAMPLE!

Randomized Experiments

• In a **randomized experiment** the explanatory variable for each unit is determined randomly, before the response variable is measured
• Because the explanatory variable is randomly assigned, it is not associated with any other variables.
• Confounding variables are eliminated!!!
• Randomized experiments make it possible to infer causation!
Chapter 1: Data Collection

Was the sample randomly selected?
Yes
No

Was the explanatory variable randomly assigned?
Yes
No

Possible to generalize to the population
Should not generalize to the population
Possible to make conclusions about causality
Can not make conclusions about causality

Chapter 2: Descriptive Statistics

- In order to make sense of data, we need ways to summarize and visualize it.
- Summarizing and visualizing variables and relationships between two variables is often known as descriptive statistics (also known as exploratory data analysis).
- Type of summary statistics and visualization methods depend on the type of variable(s) being analyzed (categorical or quantitative).

Chapter 3: Confidence Intervals

- A confidence interval for a parameter is an interval computed from sample data by a method that will capture the parameter for a specified proportion of all samples.
- A 95% confidence interval will contain the true parameter for 95% of all samples.

Confidence Intervals

- The parameter is fixed
- The statistic is random (depends on the sample)
- The interval is random (depends on the statistic)
- 95% of 95% confidence intervals will capture the truth

Confidence Intervals

Sample

Bootstrap Sample

Calculate statistic for each bootstrap sample

Bootstrap Distribution

Standard Error (SE): standard deviation of bootstrap distribution

Margin of Error (ME)
(95% CI: ME = 2×SE)
**Percentile Method**

- Lower Bound
- Observed Statistic
- Upper Bound

**Statistics: Unlocking the Power of Data**

---

**Question of the Day**

**What happens when you switch to organic food?**

---

**The Organic Effect**

- A study took a Swedish family who ate conventionally (non-organic) and had them eat only organic food for 2 weeks.
- The resulting video publicizing the study has had over 30 million views between Youtube, Facebook, and other sites.
- Here we look at data from the original study.


---

**Data**

- Measurements on pesticide levels taken the week before switching to organic (while on regular non-organic diet).
- Ate only organic for two weeks.
- Measurements on pesticide levels taken again the second of those two weeks.
- Explanatory variable: before or after switching to organic food.
- Response variables:
  - Pesticide detected or not
  - Pesticide concentration (μg / g crt)

---

**Generalize?**

Can we generalize to a population of all people?

- a) Yes
- b) No
**Study Design**

Is this an experiment or an observational study?

a) Experiment  

b) Observational study

---

**Causality?**

Can we make conclusions about causality?

a) Yes  

b) No

---

**Variables**

We first want to analyze detection of pesticides (pesticide detected or not) before and after eating organic.  

This is a question involving  

a) One categorical variable  

b) Two categorical variables  

c) One quantitative variable  

d) Two quantitative variables  

e) One quantitative and one categorical variable

---

**Visualization**

In analyzing detection of pesticides (pesticide detected or not) before and after eating organic, how might we visualize this data?

a) Bar chart  

b) Histogram  

c) Side-by-side boxplots  

d) Segmented bar charts  

e) Scatterplot

---

**Detection of Pesticides**

- There were 12 pesticides measured 20 times each, before and after eating organic, yielding 240 measurements in each group.  
- Before eating organic, 111 measurements detected pesticide.  
- After eating organic, only 24 measurements detected pesticide.  
- Create a two-way table summarizing this data.
**Parameter and Statistic**

In analyzing detection of pesticides (pesticide detected or not) before and after eating organic, what would be an appropriate parameter and statistic for inference?

- a) Single proportion
- b) Difference in proportions
- c) Single mean
- d) Difference in means
- e) Correlation

**Detection of Pesticides**

- $p_1 - p_2$: Proportion detecting pesticides before eating organic – proportion detecting pesticides after eating organic. Calculate the relevant statistic.

<table>
<thead>
<tr>
<th></th>
<th>Detected</th>
<th>Not Detected</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before eating organic</td>
<td>111</td>
<td>129</td>
<td>240</td>
</tr>
<tr>
<td>After eating organic</td>
<td>24</td>
<td>216</td>
<td>240</td>
</tr>
<tr>
<td>TOTAL</td>
<td>135</td>
<td>345</td>
<td>480</td>
</tr>
</tbody>
</table>

**Bootstrap Distribution**

This is a bootstrap distribution based on 1000 simulations. Approximate a 99% confidence interval.

- a) (0.33, 0.4)
- b) (0.3, 0.42)
- c) (0.27, 0.45)
- d) (0.25, 0.5)

**Interpretation**

- Interpret the confidence interval in context.

**Variables**

We’ll next analyze the response variable concentration of pesticide (measured in μg/g creatinine) to see if it differs before and after eating organic.

This is a question involving

- a) One categorical variable
- b) Two categorical variables
- c) One quantitative variable
- d) Two quantitative variables
- e) One quantitative and one categorical variable

**Visualization**

In analyzing concentration of pesticides before and after eating organic, how might we visualize this data?

- a) Bar chart
- b) Histogram
- c) Side-by-side boxplots
- d) Segmented bar charts
- e) Scatterplot
Parameter and Statistic

In analyzing pesticide concentration (for any one pesticide) before and after eating organic, what would be an appropriate parameter and statistic for inference?

- a) Single proportion
- b) Difference in proportions
- c) Single mean
- d) Difference in means
- e) Correlation

95% Confidence Interval

- Use bootstrap distribution from StatKey to create a 95% confidence interval using the standard error method.
- Interpret the confidence interval in context.

Variables

Actually, this data is paired (before and after measurements on the same people), so it makes sense to look at the differences in pesticide concentration, before – after.

This is a question involving

- a) One categorical variable
- b) Two categorical variables
- c) One quantitative variable
- d) Two quantitative variables
- e) One quantitative and one categorical variable
Visualization

In analyzing the differences in concentration of pesticides before and after eating organic, how might we visualize this data?

a) Bar chart  
b) Histogram  
c) Side-by-side boxplots  
d) Segmented bar charts  
e) Scatterplot

Parameter and Statistic

In analyzing the differences in concentration of each pesticide before and after eating organic, what would be an appropriate parameter and statistic for inference?

a) Single proportion  
b) Difference in proportions  
c) Single mean  
d) Difference in means  
e) Correlation

3-PBA

- Insecticide found in grains, fruits, vegetables
- Difference in 3-PBA concentration before and after:

Confidence Interval

- Using the statistic and standard error, calculate a 95% confidence interval and interpret in context.
## EXAM 1
- In-class on Friday, 10/9
- Covers everything we have done in class or lab so far (Chapters 1 – 3 in book, except not 2.6)
- Bring a non-cell phone calculator and a one-sided page of notes (8 ½ x 11 paper)
- Optional review problems posted on WileyPlus (doing lots of problems is the best way to study!)

## Exam Topics: Ch 1-3
- **Chapter 1: Data Collection**
  - Data (cases, variables, etc.)
  - Sampling
  - Observational studies and confounding
  - Randomized experiments
- **Chapter 2: Descriptive Statistics (except not 2.6)**
  - Summary statistics for one or two variable(s)
  - Graphical displays for one or two variable(s)
  - Conditional probability
- **Chapter 3: Confidence Intervals**
  - Sampling distributions
  - Confidence intervals
  - Margin of error
  - Standard error and 95% confidence intervals
  - Bootstrapping
  - Percentile method for any level of confidence