STATISTICS 110

Outline for today:

- Go over syllabus and dates for the quarter
- Overview of basic terminology
- Cover most of Chapter 0
- Overview of coverage in this course and in Stat 111/202

Examples on White Board

1. Ex 0.4: Do students with higher GPA have a better chance of getting into med school? *MedGPA* includes Accept/Deny and GPA
2. Ex 0.6: Do financial incentives help people lose weight? Randomly assigned to get incentive or not (control group) *WeightLossIncentive4* and page 8.

Some Fundamental Definitions

- **Population:** All of the individual ***units*** about which we want information
  - **Examples on white board**
- **Sample:** Units for which we obtain data
  - **Examples on white board**
- **A variable:** Something we measure (for sample) or could measure (for population) on each unit
  - **Examples on white board**

Types of Data (Variables)

- **Categorical:** Data consist of category names
  - Male/Female (two categories = **binary**)
  - Level of education (ordered categories = **ordinal**)
  - Smoker/nonsmoker
  - Opinion on an issue (favor, oppose, no preference)
  - Admit status (for med school example)
- **Quantitative:** Data consist of numbers where ordinary arithmetic makes sense
  - Height, weight, GPA, number of siblings

More Fundamental Definitions

(Population) Parameter:

A number associated with a *population*

- **Example:** Proportion admitted to med school for the *population* of applicants with GPA of at least 3.5.

(Sample) Statistic:

A number associated with a *sample*

- **Example:** Proportion admitted to med school for the observed *sample* of applicants with GPA of at least 3.5.

Description or Decision? How Data Are Used

- **Descriptive Statistics:** using numerical and graphical summaries to characterize a data set (and *only* that data set).
- **Inferential Statistics:** using sample information to make conclusions about a *population*.
- **Models:** Used to approximate the population relationship between two (or more) variables. This course is all about finding good models!

Definitions of Types of Studies

Observational Study:

- Researchers *observe* or *question* participants about opinions, behaviors, or outcomes.
- Participants not asked to do anything different.
- Example: We cannot randomly assign students to have GPA above/below 3.5!

Two special cases:

Sample surveys and *Case-control studies*.

Experiment:

Researchers *manipulate* something and *measure the effect* of the manipulation on some outcome of interest.

Randomized experiments: participants are *randomly assigned* to participate in one condition (called *treatment*) or another.

Sometimes cannot conduct experiment due to practical/ethical issues.

NOT the same thing as random sampling.

Two Important Issues Based on Data Collection Method

- **Extending results to a population:** This can be done if the *data are representative of a larger population* ***for the question of interest.*** Safest to use a ***random sample***.
- **Cause and effect conclusion:** Can *only* be made if data are from a ***randomized experiment,*** **not from an *observational study.***
- **Examples on white board**

Types of Variables (Measured or Not)

- **Explanatory variable** (or independent variable) is one that may explain or may cause differences in a **response variable** (or outcome or dependent variable).
- A **confounding variable** is a variable that:
  - *affects the response variable* and also
  - is *related to the explanatory variable*.
- **Example:** Admit (yes/no) is response variable and GPA is explanatory variable. Possible confounding variable is general ambition.

Example of an Observational Study: *Lead Exposure and Bad Teeth*

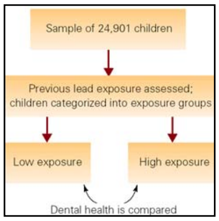
> "Children exposed to lead are more likely to suffer tooth decay …"
> *USA Today*

Observational study involving 24,901 children.

Explanatory variable = level of lead exposure.

Response variable = extent child has missing/decayed teeth.

Possible confounding variables = income level, diet, time since last dental visit.


Sample of 24,901 children
Previous lead exposure assessed; children categorized into exposure groups
Low exposure
High exposure
Dental health is compared

CRUCIAL POINT

This study is an observational study.

We cannot conclude that lead exposure *causes* tooth decay.

It would be unethical to do a randomized experiment, so we need other (non-statistical) ways to establish cause and effect.

Randomized Experiment:

Quitting Smoking with Nicotine Patches

> "After the eight-week period of patch use, almost half (46%) of the nicotine group had quit smoking, while only one-fifth (20%) of the placebo group had." *Newsweek, March 9, 1993, p. 62*

Double-blind, Placebo-controlled Randomized Experiment

240 smokers recruited (volunteers)

Randomized to 22-mg nicotine patch or placebo (**controlled**) patch for 8 weeks.

Double-blind: neither the participants nor the nurses taking the measurements knew who had received the active nicotine patches.

CRUCIAL POINT

This study is a randomized experiment.

We *can* conclude that nicotine patches *cause* people to quit smoking.

Potential confounding variables should be similar in the placebo and nicotine patch groups because of random assignment.

Summary of Types of Studies

Observational study – Data are recorded without "manipulating" any of the variables.

Statistical experiment – One or more of the explanatory variables is/are assigned/controlled for all experimental units.

Should use an experiment if we want to confirm a "cause/effect" relationship.

Cannot conclude cause/effect from an observational study!

Building a Statistical Model: Four-step Process Used by Textbook

1. CHOOSE – Pick a form for the model.
2. FIT – Estimate any parameters.
3. ASSESS – Is the model adequate? Could it be simpler? Are conditions met?
4. USE – Answer the question of interest.

General form of a model (for each individual):

$$Y = \boxed{f(X)} + \varepsilon$$

Individual Random error

"Expected" Y for some combination of predictors

$$\text{Data} = \text{Model} + \text{Error}$$

Simplest Example: Constant Model; predict weight loss for certain diet, based on sample of people

CHOOSE this model: $Y = c + \varepsilon$

where c is an unknown constant.

Terminology:

The constant c is a parameter of this model.

We use data to provide a sample estimate of c.

How should we estimate c from data?

FIT the model: Predicted Value for *Y*

Get an *estimate* for *Y* using the predictors and the model with estimated parameter(s). For the "constant" model, only 1 parameter.

Note: The predicted *Y* is denoted $\hat{Y}$.

Examples: $\hat{Y} = \overline{Y}$ (c = Sample mean)

$\hat{Y} = m$ (c = Sample median)

Assessment Questions

(1) Which estimator (mean or median) is *better*?

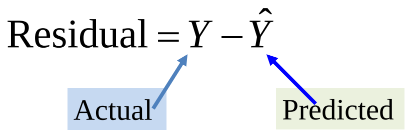
(That is, how can we compare models?)

(2) Is *either* model any good?

(That is, how can we assess fit?)

Assessing Fit: Residuals

Using the predicted value for each sample point the residual is:



$$\text{Residual} = Y - \hat{Y}$$

Assess fit by creating a summary of size of the residuals – want it to be small!

Criteria to Minimize Residuals

Sum of residuals: $\sum (Y - \hat{Y})$

Sum of absolute deviations: $\sum \left| Y - \hat{Y} \right|$

Sum of squared errors: $\sum (Y - \hat{Y})^2$

Use the Model

After choosing a model, fitting it, and assessing that it fits well, you can use it to:

- Predict the *response variable* for an individual in the future, when you only know the value(s) of the explanatory variable(s)
- Estimate the *mean response* for a specific value of the explanatory variable(s)
- Extend results to a population, if appropriate
- Determine causal relationships, if appropriate

Overview of Types of Models

Response	Explanatory	Procedure	Where
Quantitative	One quantitative	Simple linear regression	Chs 1 &2
Quantitative	Multiple	Multiple regr.	Chs 3, 4
Quantitative	One categorical	One-way ANOVA	Ch 5
Quantitative	Binary	Two-sample t	Stat 7
Quantitative	Multiple cat.	ANOVA	Chs 6, 7
Categorical	Categorical	Chi-square	Stat 7
Categorical	Quantitative	Logistic regr.	Stat 111
Categorical	Multiple	Logistic regr.	Stat 111