Econ 103 – Statistics for Economists

Intro and Chapter 1

Mallick Hossain

University of Pennsylvania

Syllabus and Logistics

Where is Everything?

- The Syllabus

Where is Everything?

- The Syllabus
- Course materials are on my website (mallickhossain.com/econ-103)

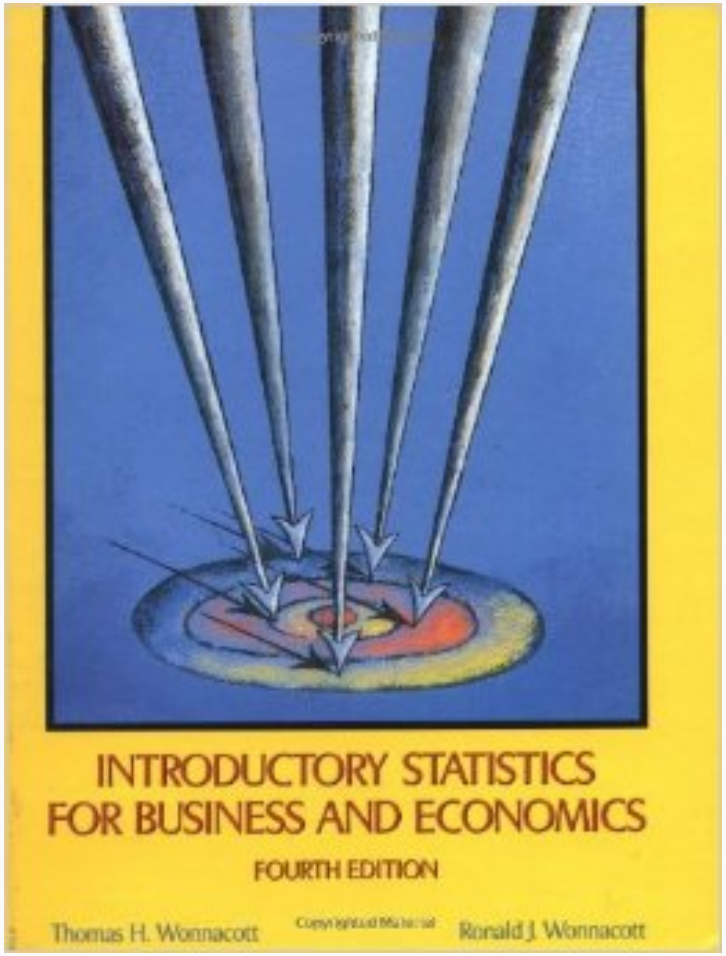
Where is Everything?

- The Syllabus
- Course materials are on my website (mallickhossain.com/econ-103)
- Grades are on Canvas (canvas.upenn.edu)

Where is Everything?

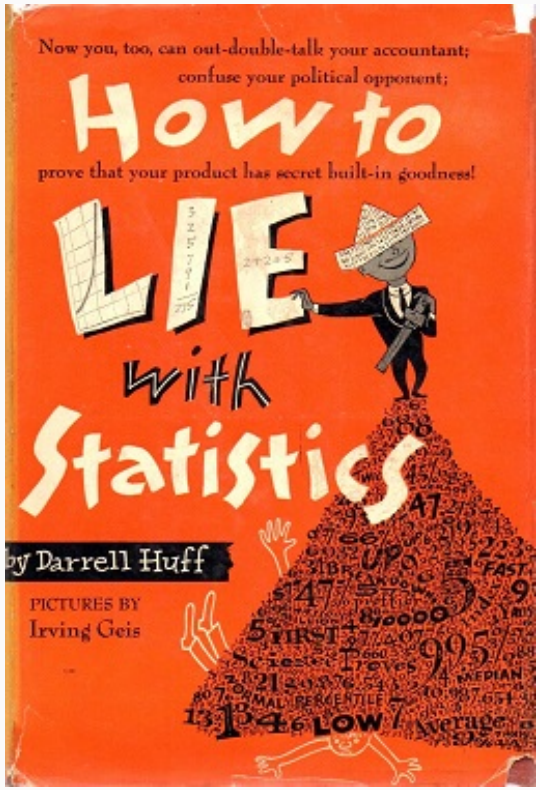
- The Syllabus
- Course materials are on my website (mallickhossain.com/econ-103)
- Grades are on Canvas (canvas.upenn.edu)
- Questions are on Piazza (piazza.com)

Textbook



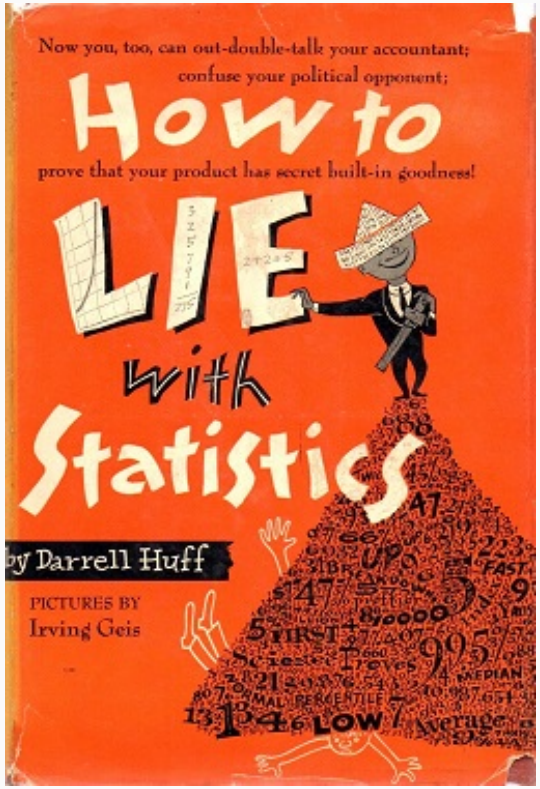
Just get a used copy and save some money

Other Recommendations



Your “Defense Against the Dark (Statistical) Arts” guide. 100% of teachers recommend it*

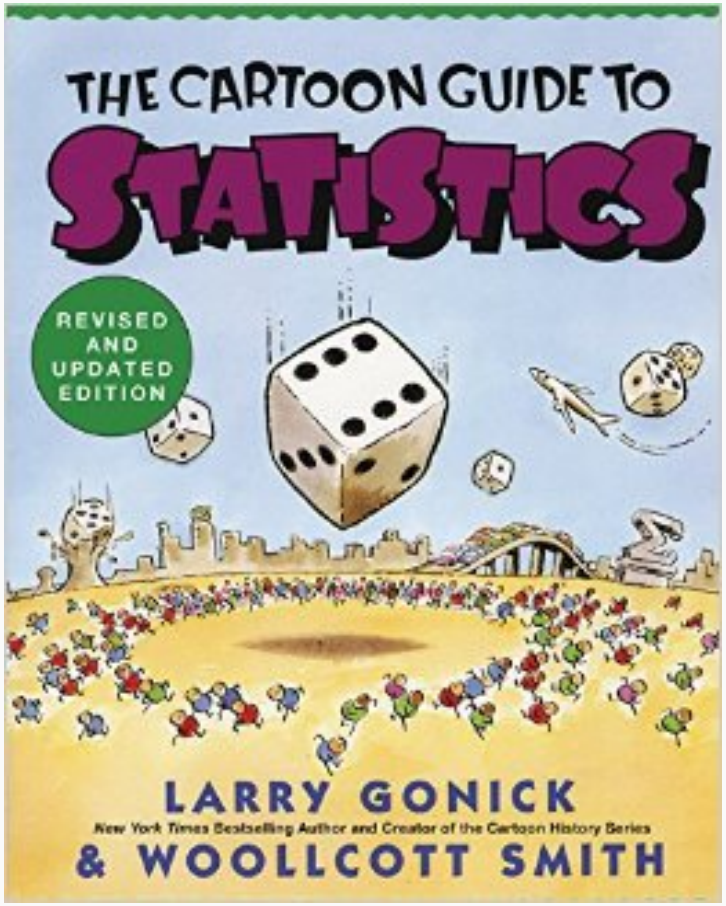
Other Recommendations



Your "Defense Against the Dark (Statistical) Arts" guide. 100% of teachers recommend it*

*based on a sample of Econ 103 teachers named Mallick Hossain ($n = 1$)

Other Recommendations



Everyone loves cartoons![citation needed]

Grading

1. Default Scheme

$$\begin{aligned}\text{Final Grade} = (20\% \times \text{R Project}) + (20\% \times \text{Midterm 1}) \\ + (20\% \times \text{Midterm 2}) + (40\% \times \text{Final})\end{aligned}$$

2. Participation Scheme (must opt-in)

$$\begin{aligned}\text{Final Grade} = (15\% \times \text{Participation}) + (15\% \times \text{R Project}) \\ + (20\% \times \text{Midterm 1}) + (20\% \times \text{Midterm 2}) \\ + (30\% \times \text{Final})\end{aligned}$$

R Project

- Teams are encouraged

R Project

- Teams are encouraged
- Open-ended assignment

R Project

- Teams are encouraged
- Open-ended assignment
- Come up with a question you're interested in exploring

R Project

- Teams are encouraged
- Open-ended assignment
- Come up with a question you're interested in exploring
- Find a dataset that could be used to illustrate your question

R Project

- Teams are encouraged
- Open-ended assignment
- Come up with a question you're interested in exploring
- Find a dataset that could be used to illustrate your question
    - For macro data, the Fed, OECD, or IMF are good resources

R Project

- Teams are encouraged
- Open-ended assignment
- Come up with a question you're interested in exploring
- Find a dataset that could be used to illustrate your question
  - For macro data, the Fed, OECD, or IMF are good resources
  - It does not even have to be economics related! If you want to do something with sports, politics, Twitter, finance, in-class survey, etc., go for it!

R Project

Write a report or do a presentation that contains the following:

- Summary of question

R Project

Write a report or do a presentation that contains the following:

- Summary of question
- Summary of data

R Project

Write a report or do a presentation that contains the following:

- Summary of question
- Summary of data
- Initial analysis (summary stats) of the data

R Project

Write a report or do a presentation that contains the following:

- Summary of question
- Summary of data
- Initial analysis (summary stats) of the data
- Data visualization/hypothesis testing (depending on complexity of data)

R Project

Write a report or do a presentation that contains the following:

- Summary of question
- Summary of data
- Initial analysis (summary stats) of the data
- Data visualization/hypothesis testing (depending on complexity of data)
- Discussion of results

R Project

Write a report or do a presentation that contains the following:

- Summary of question
- Summary of data
- Initial analysis (summary stats) of the data
- Data visualization/hypothesis testing (depending on complexity of data)
- Discussion of results
- Suggestions for further analysis or extension to the project

R Project

Write a report or do a presentation that contains the following:

- Summary of question
- Summary of data
- Initial analysis (summary stats) of the data
- Data visualization/hypothesis testing (depending on complexity of data)
- Discussion of results
- Suggestions for further analysis or extension to the project
- If a team is particularly ambitious (or has previous coding experience), R has the ability to make interactive applications!

Attendance

- I will not take attendance, so show up if it's helpful
- If you opted into the "Participation" grading scheme, part of your score comes from how active you are in class, so ask questions and PARTICIPATE!

How Do I Do Well In the Course?

- Don’t cram

How Do I Do Well In the Course?

- Don’t cram
- Learn concepts, don’t memorize

How Do I Do Well In the Course?

- Don't cram
- Learn concepts, don't memorize
- Review slides shortly after lecture

How Do I Do Well In the Course?

- Don't cram
- Learn concepts, don't memorize
- Review slides shortly after lecture
- Quizzes assess your fundamental understanding

How Do I Do Well In the Course?

- Don't cram
- Learn concepts, don't memorize
- Review slides shortly after lecture
- Quizzes assess your fundamental understanding
- Do the homework

How Do I Do Well In the Course?

- Don't cram
- Learn concepts, don't memorize
- Review slides shortly after lecture
- Quizzes assess your fundamental understanding
- Do the homework
- Learn R

How Do I Do Well In the Course?

- Don't cram
- Learn concepts, don't memorize
- Review slides shortly after lecture
- Quizzes assess your fundamental understanding
- Do the homework
- Learn R, seriously.

Learning Curves

There are two learning curves to be aware of in this course:

1. **Statistics:** This will be a tough, but manageable one (like hiking up a constant moderate-graded mountain)

Learning Curves

There are two learning curves to be aware of in this course:

1. **Statistics:** This will be a tough, but manageable one (like hiking up a constant moderate-graded mountain)
2. **Statistical Programming:** This is probably best illustrated at the link below:

Learning Curves

There are two learning curves to be aware of in this course:

1. **Statistics:** This will be a tough, but manageable one (like hiking up a constant moderate-graded mountain)
2. **Statistical Programming:** This is probably best illustrated at the link below:
    - `http://i.imgur.com/vPkUXWB.gif`

Learning Curves

There are two learning curves to be aware of in this course:

1. **Statistics:** This will be a tough, but manageable one (like hiking up a constant moderate-graded mountain)
2. **Statistical Programming:** This is probably best illustrated at the link below:
    - `http://i.imgur.com/vPkUXWB.gif`
    - The good news is that once you get around the curve, it will be a pleasant ride in a Cadillac

Learning Curves

There are two learning curves to be aware of in this course:

1. **Statistics:** This will be a tough, but manageable one (like hiking up a constant moderate-graded mountain)
2. **Statistical Programming:** This is probably best illustrated at the link below:
   - `http://i.imgur.com/vPkUXWB.gif`
   - The good news is that once you get around the curve, it will be a pleasant ride in a Cadillac
   - Tutorials to make getting around the curve easier

Motivation

Motivation

Real Motivation

▸ Who's Ready to Make Some Science?

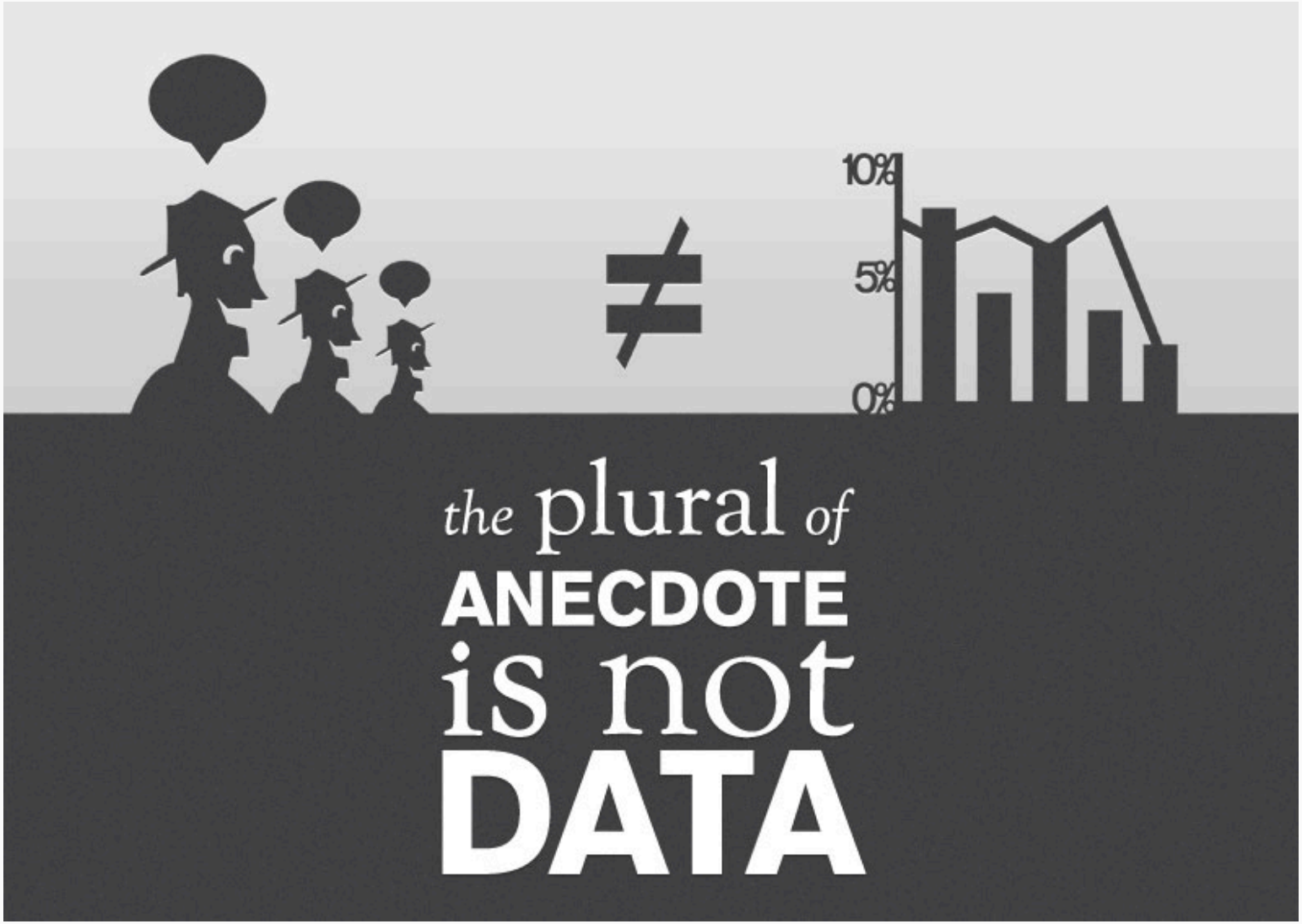
Real Life Examples

“People in this country have had enough of experts”

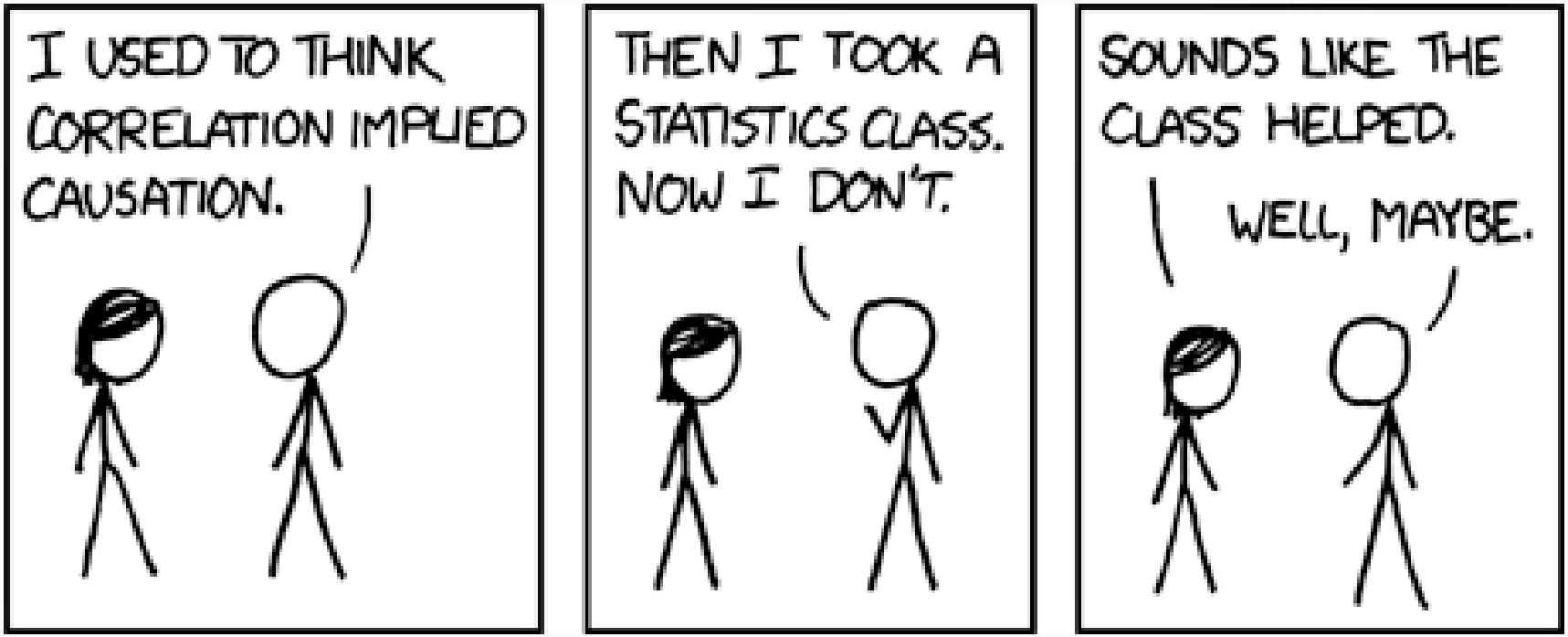
Real Life Examples

"5.3 percent unemployment – that is the biggest joke there is in this country. … The unemployment rate is probably 20 percent, but I will tell you, you have some great economists that will tell you it's a 30, 32. And the highest I've heard so far is 42 percent."

Remember!



Remember!


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Questions

- How many people are employed?

Questions

- How many people are employed?
- How many people have a high school diploma/GED?

Bad Statistics

- Using this class as a representation of the U.S. population
  - U.S. employment-population ratio is 59.7 percent
  - 88 percent of adults (25 and older) have a high-school degree or equivalent

Good (Albeit Useless) Statistics

- Using this class as a representation of this class
  - X percent of Wednesday evening Econ 103 students are employed
  - X percent of Wednesday evening Econ 103 students have a high school diploma or equivalent

Chapter 1: The Nature of Statistics

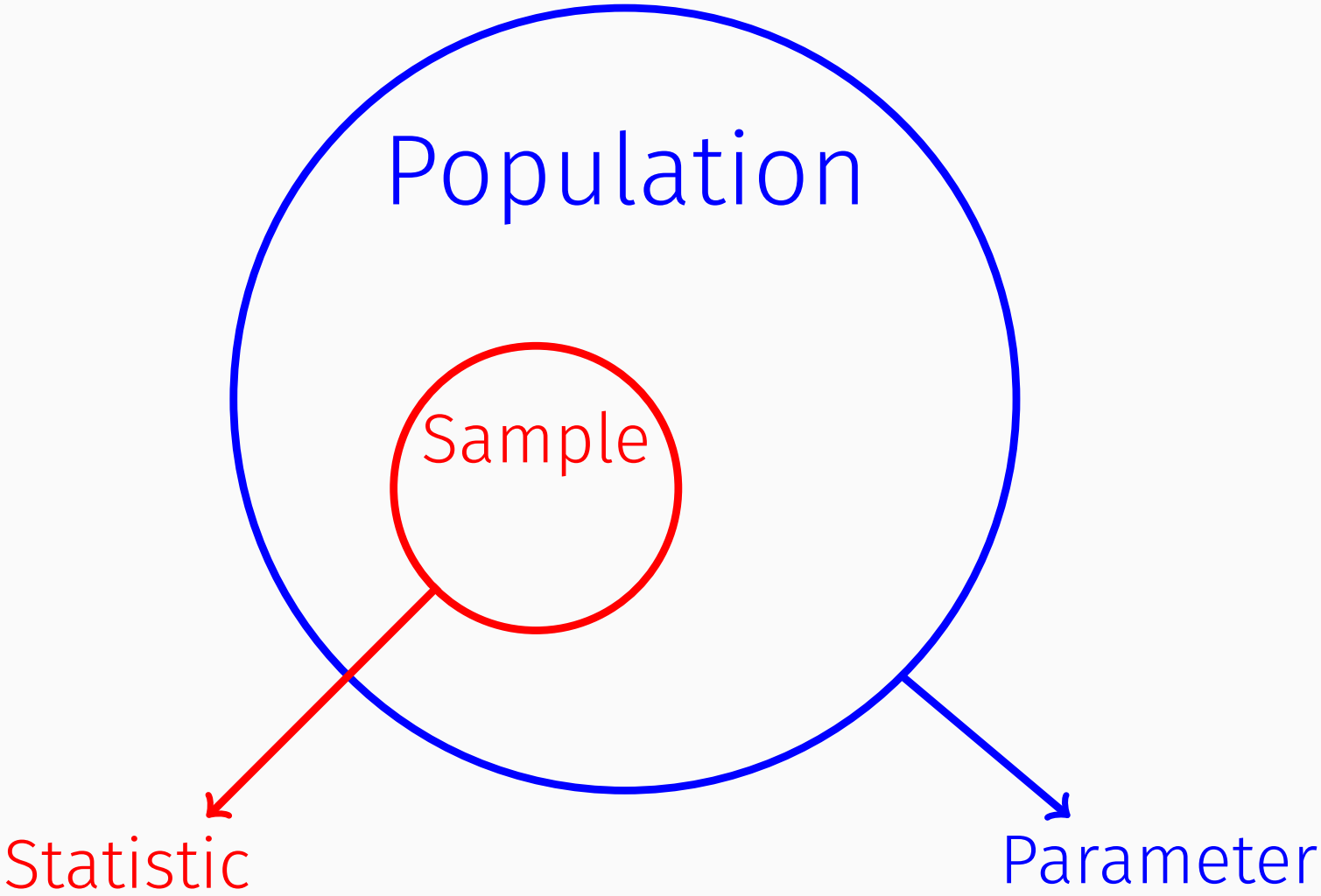
Rule 1: Sample $\neq$ Population

How did this happen?

Definitions

- **Population:** Complete set of all items of interest
- **Parameter:** A specific characteristic of a *population*
- **Sample:** Observed subset of the *population*
- **Statistic:** A specific characteristic of a *sample*
- **Sample Size (n):** Number of items in the *sample*

Essential Distinction You Must Remember!



Kinds of Statistics

- **Descriptive Statistics:** Graphical and numerical summaries of data
- **Inferential Statistics:** Using data to estimate, predict, and quantify uncertainty

Course Outline

1. Descriptive Statistics: summarize data
    - Summary Statistics
    - Graphics
2. Probability: Population $\rightarrow$ Sample
    - Using information about the population to predict properties of a sample
    - Deductive: "safe" argument
        - All ducks waddle, swim, and quack. Donald is a duck. Donald must waddle, swim, and quack.
3. Statistics: Sample $\rightarrow$ Population
    - Using information about the sample to predict properties of the population
    - Inductive: "risky" argument
        - If it walks like a duck, quacks like a duck, and swims like a duck, it's probably a duck
        - When you hear hoofbeats, think horses, not zebras

Why Do Statistics?

- **Ockham's Razor:** If we can predict everything based on the population, just get data on the population and call it a day, right? How hard can this really be?

Why Do Statistics?

- **Ockham's Razor:** If we can predict everything based on the population, just get data on the population and call it a day, right? How hard can this really be?
- What's wrong with this reasoning?

Why Do Statistics?

- **Ockham's Razor:** If we can predict everything based on the population, just get data on the population and call it a day, right? How hard can this really be?
- What's wrong with this reasoning?
  - **Limited resources:** Surveying the whole population is expensive and usually infeasible

Why Do Statistics?

- **Ockham's Razor:** If we can predict everything based on the population, just get data on the population and call it a day, right? How hard can this really be?
- What's wrong with this reasoning?
  - **Limited resources:** Surveying the whole population is expensive and usually infeasible
  - **Scarcity:** Sometimes only a small sample is available

Why Do Statistics?

- **Ockham's Razor:** If we can predict everything based on the population, just get data on the population and call it a day, right? How hard can this really be?
- What's wrong with this reasoning?
  - **Limited resources:** Surveying the whole population is expensive and usually infeasible
  - **Scarcity:** Sometimes only a small sample is available
  - **Destructive testing:** Rating car parts for durability requires testing them until they break. If you tested every part, you'd have no parts to use in cars.

Why Do Statistics?

- **Ockham's Razor:** If we can predict everything based on the population, just get data on the population and call it a day, right? How hard can this really be?
- What's wrong with this reasoning?
  - **Limited resources:** Surveying the whole population is expensive and usually infeasible
  - **Scarcity:** Sometimes only a small sample is available
  - **Destructive testing:** Rating car parts for durability requires testing them until they break. If you tested every part, you'd have no parts to use in cars.
  - **Error reduction:** Getting data on the whole population could aggravate measurement error if done improperly

Sampling and Nonsampling Error

In statistics we use samples to learn about populations, but samples almost never are *exactly* like the population they are drawn from.

1. Sampling Error
    - *Random* differences between sample and population
    - Cancel out on average
    - Decreases as sample size grows
2. Nonsampling Error
    - *Systematic* differences between sample and population
    - Does *not* cancel out on average
    - Does *not* decrease as sample size grows

Example: Historic Polling Mistake

Illustrative Example

Literary Digest – 1936 Presidential Election Poll

FDR versus Kansas Gov. Alf Landon

Data

Sent out over 10 million ballots to those on auto registries and phone books.

2.4 million replied (Compared to less than 45 million votes cast in actual election)

Prediction

Landslide for Landon: *Landonslide*, if you will.

What Could Go Wrong?

FDR versus Kansas Gov. Alf Landon

	Roosevelt	Landon
Literary Digest Prediction:	41%	57%

What Could Go Wrong?

FDR versus Kansas Gov. Alf Landon

	Roosevelt	Landon
Literary Digest Prediction:	41%	57%

The rest is history. President Landon joined the ranks of forgettable presidents like Millard Fillmore and William Henry Harrison

What Could Go Wrong?

FDR versus Kansas Gov. Alf Landon

	Roosevelt	Landon
Literary Digest Prediction:	41%	57%
Actual Result:	61%	37%

Oops...

What Went Wrong? *Non-sampling Error (aka Bias)*

Biased Sample

Sampled car owners and those with telephones

Non-response Bias

Even if sample is unbiased, can't force people to reply.

- Among those who recieved a ballot, Landon supporters were more likely to reply.

In this case, neither effect *alone* was enough to throw off the result but together they did.

Source: Squire (1988)

How Do You Get an Unbiased Sample?

Simple Random Sample

Each member of population is chosen strictly by chance, so that: (1) selection of one individual doesn't influence selection of any other, (2) each individual is just as likely to be chosen, (3) every possible sample of size n has the same chance of selection.

What about non-response bias?

“Americans Divided on Outlook for Next Generation”

PRINCETON, NJ – Americans are evenly divided about whether it is likely (49%) or unlikely (50%) that the next generation of youth in the country will have a better life than their parents. That is a slightly more positive assessment than in early 2011, when the slight majority, 55%, thought it was unlikely the next generation would achieve this goal.

Source: Gallup

Example of Sampling Error

"...evenly divided about whether it is likely (49%) or unlikely (50%) that the next generation of youth in the country will have a better life..."

> *Results for this USA Today/Gallup poll are based on telephone interviews conducted Dec. 14-17, 2012, with a random sample of 1025 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia. For results based on the total sample of national adults, one can say with 95% confidence that the maximum margin of sampling error is ±4 percentage points.*

Quantifying Sampling Error

$$Marginof Error(ME) \approx 2\sqrt{P(1-P)/n}$$

We report: $P \pm ME$ (often called the confidence interval)

Correlation, Causation, RCTs

Swimming Pools and Lead Poisoning

Ask random sample of parents if they have an in-ground swimming pool and whether their child contracted lead poisoning. Compare those who had pools to those who did not. Would this procedure:

(a) Overstate health benefits of swimming (or really, having a swimming pool)

(b) Correctly identify health benefits of swimming

(c) Understate health benefits of swimming

Problem

Parents who own swimming pools may differ systematically from those who don't in *other ways* that impact child's chance of getting lead poisoning!

Wealth influences one's ability to have a swimming pool and to live in a house without lead paint.

Confounder

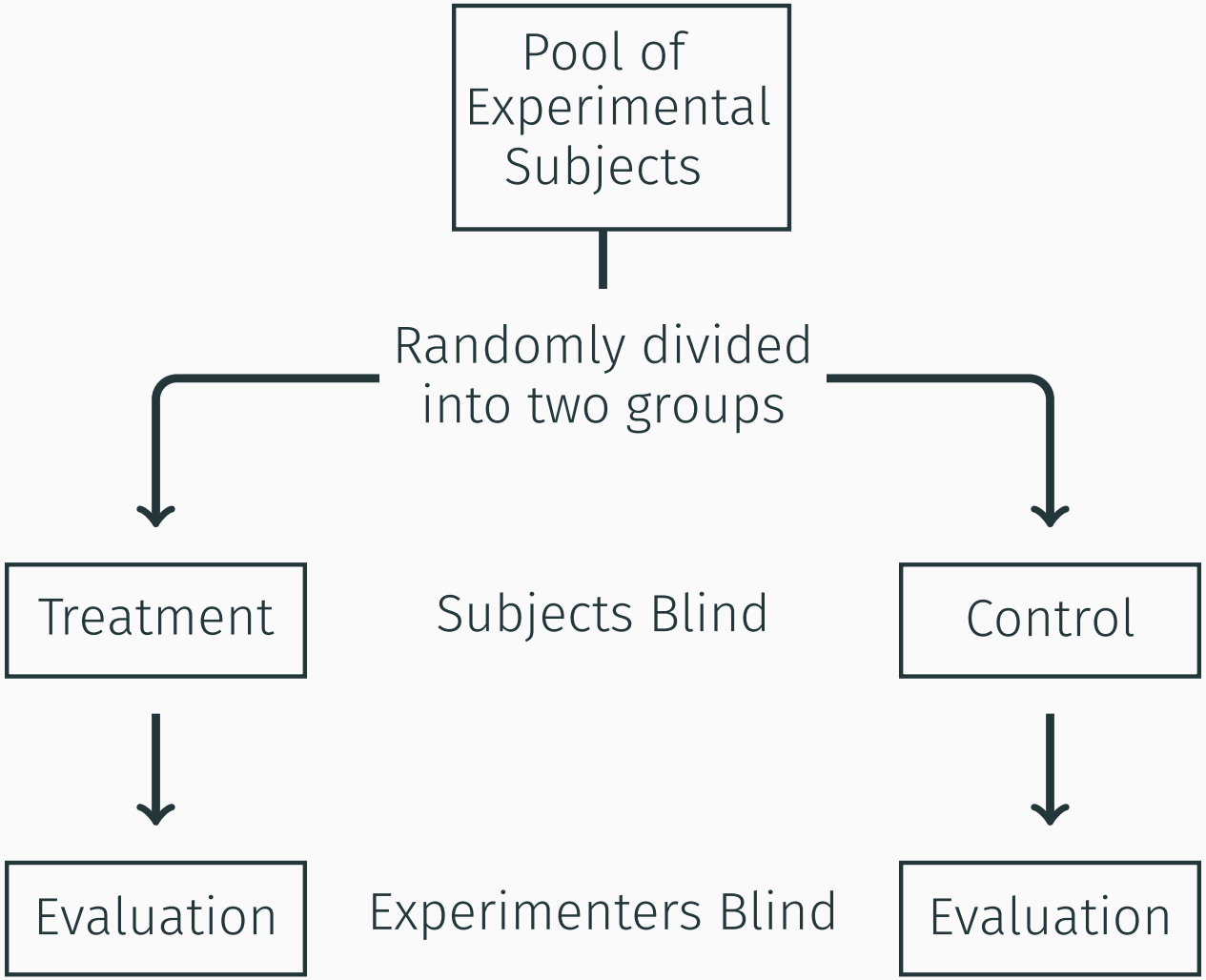
Factor than influences both outcomes and whether subjects are treated or not. Masks true effect of treatment.

Properly Determining Treatment Effectiveness: Randomized Experiments

- Start with group of experimental subjects
- Randomly assign one group to get the “treatment” and the other gets nothing (i.e. the “control” group)
- Random assignment neutralizes the chance of confounding factors since groups are initially equal, on average, and only difference is the treatment.

Double-blind randomized trials are the gold standard

Double-Blind Randomized Trial



Gold Standard: Randomized, Double-blind Experiment

Randomized blind experiments ensure that on average the two groups are initially equal, and continue to be treated equally. Thus a fair comparison is possible.

Randomized, double-blind experiments are generally the best way to untangle causation.

Ockham's Razor II: Randomize everything and fix this whole causation/correlation problem!

What Shall We Solve?

- Does gender affect one's wages?

Ockham's Razor II: Randomize everything and fix this whole causation/correlation problem!

What Shall We Solve?

- Does gender affect one's wages?
- Does the defendant's race affect their sentencing?

Ockham's Razor II: Randomize everything and fix this whole causation/correlation problem!

What Shall We Solve?

- Does gender affect one's wages?
- Does the defendant's race affect their sentencing?
- Does spanking cause criminality?

Randomization is not always possible, practical, or ethical.

Aperture Science and Randomized Trials

▸ Mandatory Testing ▸ Control Groups ▸ Control Groups (ct'd)

How Can We Learn Anything Without Randomized Experiments?

Observational Data

Data that do not come from a randomized experiment.

It is very difficult to untangle cause and effect using observational data because of confounders.

Does Racial Discrimination Affect Criminal Sentencing?

Social scientists have studied the issue for decades, but the seemingly simple question "Does race affect sentencing?" is surprisingly difficult to answer on the basis of empirical evidence. Abrams explains: "The most straightforward way you might look at it is to say, Let's look at what sentences people get and see whether sentence length varies by race. If it looks like people of one race receive longer sentences than another, that might indicate that the criminal justice system is unfair. But the shortcoming to that approach is that it's also possible that sentences can differ for many reasons; for example, it's possible people of different races might have different criminal histories on average, and that could also explain the difference in sentence length."

Source: Penn Law Website

Reducing Bias in Observational Studies

Regression

Technique that allows us to remove influence of confounders.

Works well if we can identify and gather data on all of them.

But...

Does Racial Discrimination Affect Criminal Sentencing?

To address that difficulty [confounders] social scientists have ... applied control variables to standard regression equations, a statistical method for identifying significant correlations between observed events. For instance, controlling for type of crime committed or for the defendant's criminal history, researchers look to see whether the results of their equation still show racial disparity.“The problem with that is you still leave the possibility that any differences you see are due to unobserved variables, differences that might be there but that you can't control for” Abrams says.“That might be demeanor in the courtroom, it might be the quality of the attorney you can afford, it might be some details about the crime that you might not capture in your data. If those things are correlated with race, which they probably are, you're not going to know whether the effect you think you're detecting is really race or is something else.”

Source: Penn Law Website

Related Reading

- Wonnacott: Chapter 1
- How to Lie with Statistics: Chapter 1