Econ 103 – Statistics for Economists

Intro and Chapter 1

Mallick Hossain

University of Pennsylvania

Syllabus and Logistics

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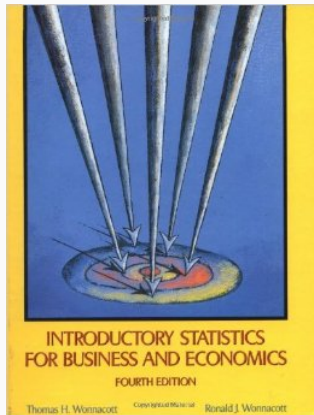
- The Syllabus
- Course materials are on my website
  (`mallickhossain.com/econ-103`)

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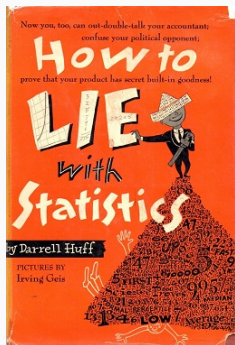
- The Syllabus
- Course materials are on my website (`mallickhossain.com/econ-103`)
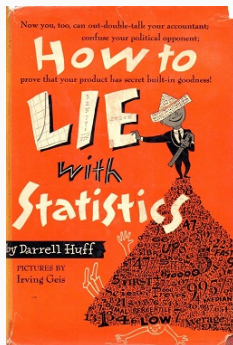- Grades are on Canvas (canvas.upenn.edu)
- Questions are on Piazza (piazza.com)

Just get a used copy and save some money

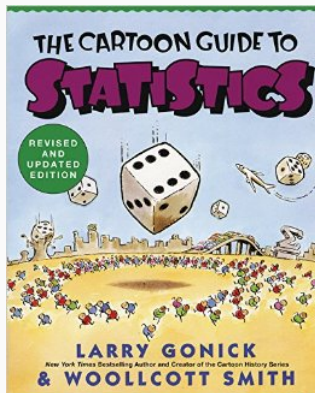Your "Defense Against the Dark (Statistical) Arts" guide. 100% of teachers recommend it*

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

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

Everyone loves cartoons![citation needed]

Grading

1. Default Scheme

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

2. Participation Scheme (must opt-in)

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

Participation

- Your participation grade will be based on the following:
  - In-class participation
  - Piazza participation
  - Homework assignments

Participation

- Your participation grade will be based on the following:
  - In-class participation
  - Piazza participation
  - Homework assignments

In short, get involved in the course!

- 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!

R Project

- Apply the skills and tools you have learned

- Apply the skills and tools you have learned
- Answer questions you are interested in

- Apply the skills and tools you have learned
- Answer questions you are interested in
- Head-start on honors thesis?

- ANYTHING

- ANYTHING, seriously.
    - Political or ideological agendas are welcome

- ANYTHING, seriously.
  - Political or ideological agendas are welcome
  - Guns, gender, inequality, race, climate change, …

- ANYTHING, seriously.
    - Political or ideological agendas are welcome
    - Guns, gender, inequality, race, climate change, …
    - Google trends, Twitter, financial data, macro data, …

- Summary of question

- Summary of question
- Summary of data and why it is relevant for the question

- Summary of question
- Summary of data and why it is relevant for the question
- Tables and charts of summary stats of the data

What Do I Turn In?

- Summary of question
- Summary of data and why it is relevant for the question
- Tables and charts of summary stats of the data
- Data visualization/hypothesis testing

What Do I Turn In?

- Summary of question
- Summary of data and why it is relevant for the question
- Tables and charts of summary stats of the data
- Data visualization/hypothesis testing
- Discussion of results

- Summary of question
- Summary of data and why it is relevant for the question
- Tables and charts of summary stats of the data
- Data visualization/hypothesis testing
- Discussion of results
- Criticism of your findings. What are the biggest flaws in your analysis or in the underlying data?

- Summary of question
- Summary of data and why it is relevant for the question
- Tables and charts of summary stats of the data
- Data visualization/hypothesis testing
- Discussion of results
- Criticism of your findings. What are the biggest flaws in your analysis or in the underlying data?
- Suggestions for further analysis or extension to the project

- **October 12:** Submit project idea and question or spoken with me about the project. Submit names of people in your group.
- **November 16:** (Optional) Submit rough draft for comments.
- **December 7:** Hand in project.

- R Tutorials provide examples of the kinds of exploration I am looking for

- R Tutorials provide examples of the kinds of exploration I am looking for
- CEA blog posts on jobs and GDP and other publications (like the Economic Report of the President)

- R Tutorials provide examples of the kinds of exploration I am looking for
- CEA blog posts on jobs and GDP and other publications (like the Economic Report of the President)
- FiveThirtyEight's report on gun deaths

- Not looking for a Nobel-prize-winning discovery

Do I Need to Discover Something New?

- Not looking for a Nobel-prize-winning discovery
- You should learn something new

- Not looking for a Nobel-prize-winning discovery
- You should learn something new
- Hopefully I'll learn something new

- Not looking for a Nobel-prize-winning discovery
- You should learn something new
- Hopefully I'll learn something new
- Be honest

- Don't cram

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

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

- 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

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

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)

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:

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

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

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

1. Descriptive Statistics: summarize data

1. Descriptive Statistics: summarize data
   - Summary Statistics and graphics

Course Outline

1. Descriptive Statistics: summarize data
   - Summary Statistics and graphics
2. Probability: Population $\rightarrow$ Sample

1. Descriptive Statistics: summarize data
   - Summary Statistics and graphics
2. Probability: Population $\rightarrow$ Sample
   - Deductive: "safe" argument

Course Outline

1. Descriptive Statistics: summarize data
   - Summary Statistics and graphics
2. Probability: Population $\rightarrow$ Sample
   - Deductive: "safe" argument
     - All ducks waddle, swim, and quack. Donald is a duck. Donald must waddle, swim, and quack.

1. Descriptive Statistics: summarize data
   - Summary Statistics and graphics
2. Probability: Population → Sample
   - Deductive: "safe" argument
     - All ducks waddle, swim, and quack. Donald is a duck. Donald must waddle, swim, and quack.
3. Statistics: Sample → Population

Course Outline

1. Descriptive Statistics: summarize data
   - Summary Statistics and graphics
2. Probability: Population → Sample
   - Deductive: "safe" argument
     - All ducks waddle, swim, and quack. Donald is a duck. Donald must waddle, swim, and quack.
3. Statistics: Sample → Population
   - Inductive: "risky" argument

1. Descriptive Statistics: summarize data
   - Summary Statistics and graphics

2. Probability: Population $\rightarrow$ 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
   - Inductive: "risky" argument
     - If it walks like a duck, quacks like a duck, and swims like a duck, it's probably a duck

1. Descriptive Statistics: summarize data
   - Summary Statistics and graphics

2. Probability: Population $\rightarrow$ 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
   - Inductive: "risky" argument
     - If it walks like a duck, quacks like a duck, and swims like a duck, it's probably a duck
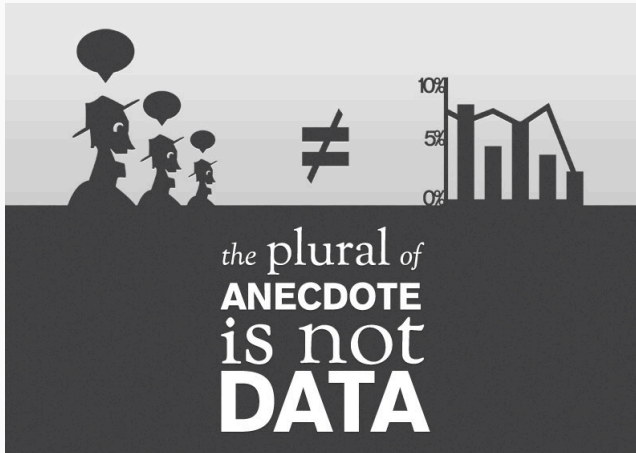   - When you hear hoofbeats, think horses, not zebras

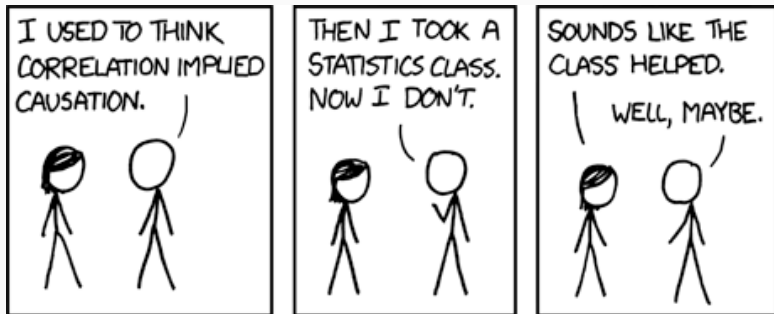# Motivation

"People in this country have had enough of experts"

"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."

- How many people are employed?

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

- 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

- 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
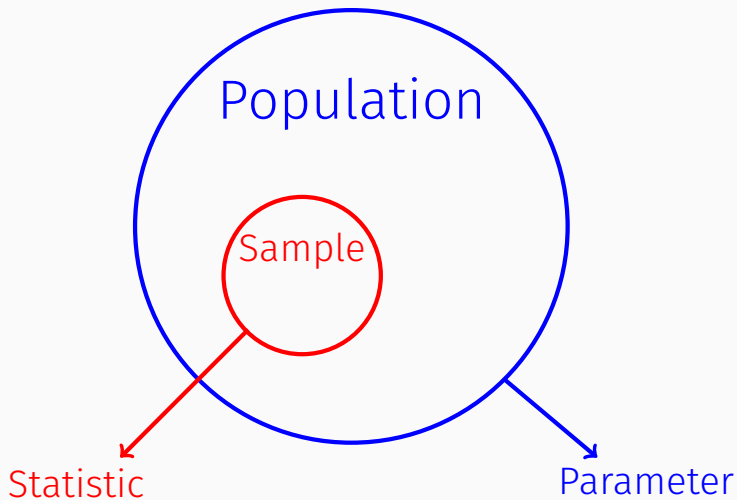
How did this happen?

- **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*

Kinds of Statistics

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

- **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?

- **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?

- **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

- **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

- **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.

- **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

FDR versus Kansas Gov. Alf Landon

Data

Over 10 million ballots sent 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.

FDR versus Kansas Gov. Alf Landon

	Roosevelt	Landon
Literary Digest Prediction:	41%	57%

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

FDR versus Kansas Gov. Alf Landon

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

Oops…

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)

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?

We'll cover how to quantify errors later in the course.

Correlation, Causation, RCTs

Swimming Pools and Lead Poisoning

We asked a random sample of parents if they have an in-ground swimming pool and whether their child suffered from lead poisoning. We compare those who had pools to those who did not and find that those with pools had lower instances of lead poisoning. Does this survey:

(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
(d) None of the above

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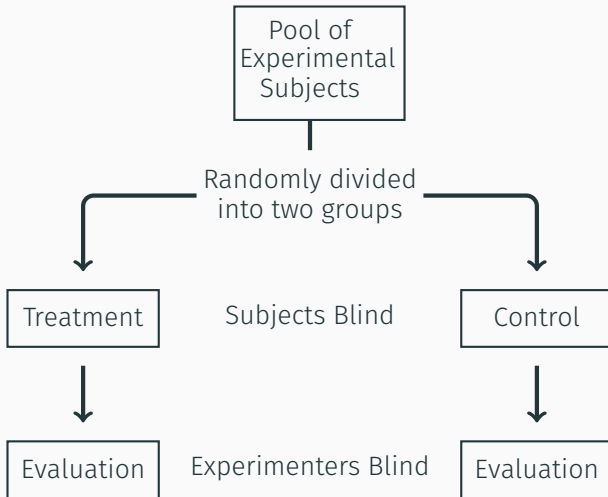
Factor than influences both outcomes and whether subjects are treated or not. Masks true effect of treatment.

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

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!

- Ockham's Razor II: Randomize everything and fix this whole causation/correlation problem!
- What Shall We Solve?

- 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.

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

Regression

Technique that allows us to remove influence of confounders.
Works well if we can identify and gather data on all of them.
But...

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

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

- Math diagnostic
- Chapter 1 Problems
- R Tutorial 1