Sampling Bias
Principle underlying all of statistics

Drawing inferences from a suitable sample of a population is far less labor intensive than, but can be equally as informative as, studying the entire population.

What’s the catch? A suitable sample must look like the population from which it is drawn, which is easier said than done. Also, bigger is better. An unsuitable sample is usually biased in some way.
Who won the 1948 U.S. Presidential Election?

A: Harry Truman

B: Thomas Dewey

C: Someone Else

*iClicker Q: Sampling Bias*
The News Got it Wrong

- Because of a strike at the paper, the *Chicago Tribune* had printed the papers before the election results became official.
  - Almost every pre-election poll had Dewey winning, so the paper assumed a Dewey victory.
  - The *Tribune* was not the only paper to make this mistake.
  - The *Journal of Commerce* had 8 articles about what to expect from the Dewey administration the day after the election.

- Errors in how the pre-election polls were conducted led to the wrong prediction.
  - Pre-election polls were not a random sample of voters; Republicans were more likely to appear in pre-election polls than Democrats.
  - Most of these polls were conducted by telephone; Republicans were more likely to have telephones than Democrats, and were thus overrepresented in the sample.
Types of Bias
Selection Bias

- When each member of the relevant population does not have an equal chance of being selected for the study.
- Iowa straw poll (now defunct!)
  - Republican candidates descend upon Ames, IA in a year before an election
  - Both a poll and a fundraiser: it costs $30 to cast your vote
  - Very poor predictor of future Republican candidate
Self-Selection Bias

- Participants choose whether or not to participate, so the sample biases towards the types of people that opt in
- Examples: Surveys being conducted at airports or highway rest areas
- What would bias someone to participate?
  - Extra time
  - Strong opinions
  - Incentives are not equally appealing to all
- How can studies avoid self-selection bias?
  - Thorough outreach; no barriers to participation
  - Carefully designed incentives
  - Opt out system
Exclusion Bias

- Researchers cannot sample some subgroups of the population: e.g.,
  - *We cannot reach Brown students who don’t read Morning Mail.*
  - Phone surveys cannot reach people without phones.
  - Government records exclude illegal immigrants.
- Exclusion bias is the opposite of selection bias because you cannot self select if you are not presented with the opportunity to do so.
- Not necessarily motivated by an intent to discriminate/gather biased data; it’s just much more difficult to reach people without phones, or email, etc.
Famous Examples of Exclusion Bias

- **Literary Digest Poll of 1936**
  - FDR or Alf Landon (Republican)
  - Ten million subscribers / prospective voters!
  - Prediction: Landon over FDR, with 57% of the popular vote
  - Actual: FDR over Landon with 60% of the popular vote
  - Problem: Magazine subscribers were wealthy Americans
    - Wealthier Americans lean Republican

- **Dewey vs. Truman (1948):** same problem; they didn’t learn!
Survivorship Bias

● Rather than sample everyone, sample only those that survived a treatment.
   Examples:
   ○ Studying the effects of a medicine only in patients that survived the clinical trials.
   ○ Studying test scores in high-school students, and claiming higher scores over the years (each year more and more high school dropouts are excluded).
   ○ Studying mutual funds over the last ten years, but not including mutual funds that do not survive the ten year period of study.

● Need not involve actual death or destruction, only the removal of certain parties from the subject group.
Example of survivorship bias

- During World War 2, the military hired a statistician to tell them where to armour their aircrafts.
- Below, is a diagram of bullet holes on planes after they returned from war.
- Where should you put the armour?
Recall Bias

From the *New York Times Magazine* (2011)

“The diagnosis of breast cancer had not just changed a woman’s present and the future; it had altered her sense of her past. Women with breast cancer had (unconsciously) decided that a higher-fat diet was a likely predisposition for their disease and (unconsciously) recalled a high-fat diet. It was a pattern poignantly familiar to anyone who knows the history of this stigmatized illness: these women, like thousands of women before them, had searched their own memories for a cause and then summoned that cause into memory.”
Publication Bias

From the *New York Times* (2008)

“The makers of antidepressants like Prozac and Paxil never published the results of about a third of the drug trials that they conducted to win government approval, misleading doctors and consumers about the drugs’ true effectiveness.”

- 94% of studies with positive fundings on the effectiveness of the drugs were published
- But only 14% of the studies with non-positive findings were published
Random Sampling

- Everyone in the population is equally likely to be selected, so that the sample is reflective of the entire population.
- And since the sample reflects the population, we can make inferences about the population just by looking at the sample.
Avoiding Bias
Made-up example of selection bias

- We want to find out the political preferences of Brown students.
- We post a morning mail asking students to choose their favorite among candidates in various elections.
- Is there selection bias in this method of sampling?
Made-up example of selection bias

- We want to find out the political preferences of Brown students.
- We post a morning mail asking students to choose their favorite among candidates in various elections.
- Is there selection bias in this method of sampling?
- Who would answer this Morning Mail?
  - Students who have time to read morning mail
  - Students who also have time to act on morning mail content
Made-up example of exclusion bias

- Amy wants to find out which CS department in the US is the best.
- She picks ten of her colleagues to survey, and then asks them to ask ten of their colleagues for their top picks.
- She declares the winner to be the department with the most votes.
- Do you see any potential for exclusion bias in this method of sampling?
Made-up example of exclusion bias

- Amy wants to find out which CS department in the US is the best.
- She picks ten of her colleagues to survey, and then asks them to ask ten of their colleagues for their top picks.
- She declares the winner to be the department with the most votes.
- Do you see any potential for exclusion bias in this method of sampling?
  - Which departments are likely more (less) connected to Amy?
  - What about the voters might exclude certain groups?
Made-up example of survivorship bias

- We want to know if muggles can survive a cruciatus curse.
- We know that Presmout was recently attacked by Death Eaters.
- We ask muggles if they survived a cruciatus curse.
- Who are we missing?
- Is the actual survival rate likely lower or higher?