Sampling

Benjamin Graham
Schedule

• This Week: Sampling and External Validity
How many kids?

- Fertility rate in the US. could be interesting as an independent or a dependent variable.

- How many children did your mother have?
Landon in a Landslide!

- 1936 Literary Digest Poll: nationwide sample
  - Ralph Landon: 1,296,669 (57%); Franklin Roosevelt 972,897
  - Broken down by state it predicted 370 electoral college votes for Landon

- What went wrong with that poll?
The Power of Random Sampling

• We often can’t measure the entire population we want to know about
  • Instead we measure a sample
  • A census vs. a poll
• We can make valid inferences about the population, based on the sample, if and only if:
  • The sample is a miniature version of your population
• The best way to do this is by random Sampling (AKA probability sampling)
  • Every unit in the population has the same probability of being chosen
Some terminology

- **Sampling frame**: A list of all the units in the population
  - Sampling frame for the analysis is the only population we can make valid inferences about
- **Units**: Sometimes we sample at multiple levels, we have to keep our units straight
  - Households vs. individuals
  - Schools vs. classrooms vs. students
- **Variance**: A statistical way to talk about the diversity of the population
Which sample of heights has a higher variance?

• Group 1:  5’10”  6’  5’11”  6’1”  6’2”
• Group 2:  6’10”  7‘  4’11”  6’1”  5‘

• A: Group 1 has a higher variance
• B: Group 2 has a higher variance

• Which group has the higher mean?
The Weak Law of Large Numbers

• **The Weak Law of Large Numbers:** If you take a random sample of observations from a population....

• The mean of the sample approaches the mean of the population as the size of the sample approaches infinity (or approaches the size of the population).

• That’s great, but we have finite samples...
  • So how big does our sample have to be?
Central Limit Theorem

- **The Central Limit Theorem:** The mean of a sample of independent draws from the same population with expected value $\mu$ and finite variance $\sigma^2$ will be normally distributed with an expected value of $\mu$ and a variance of $\sigma^2 / n$.

- What that means in practice: If we take a random sample from a population and measure something about that sample:
  - The expected value of the mean in that sample is the mean of the population.
  - If we calculate the variance in the sample, so we know how close to the population mean our sample mean is going to be.
Random Sampling Error

• The error caused by observing a random sample, instead of the whole population.
  • The bigger the sample, the smaller the error.
  • The smaller the variance, the smaller the error.
How Do We Measure Unemployment?

• # looking for work/(# of employed people + # looking for work)
  • What is the appropriate population we’re sampling from in this measure?

• How do we collect this data?

• Unemployment rate vs. labor force participation rate
The Power of Random Assignment

• Random Sampling is tied directly to random assignment
  • In an experiment, we randomly assign individuals to the treatment and the control group.
• Here, we have two random samples from the population
  • Across all variables the mean of the sample matches the population
  • Also matches the mean of the other sample
• This means that (with a few caveats) we can attribute any difference between the two groups to the treatment.
Clicker Review

I draw a random sample of 500 students from a list of all the students enrolled in public high schools in California. I have the school nurse weigh each student with the best scale money can buy. The average weight recorded in my sample is:

• A. Exactly equal to the average weight of public high school students in California
• B. Equal to the average weight of public high school students in California, plus or minus random sampling error
• C. Equal to the average weight of public high school students in California, plus or minus random sampling error and measurement error
• D. Equal to the average weight of public high school students in the U.S., plus or minus random sampling error and measurement error
• E. B & D
• F. C & D
Clicker Review

• What is my sampling frame in the study described on the last slide?
  • A. The entire population of public high school students in the U.S.
  • B. The entire population of public high school students in California.
  • C. The 500 students in my sample
Clicker Review

• About whose IQ can I make an inference, based on the measure I took in this sample?
  • A. The entire population of public high school students in the U.S.
  • B. The entire population of public high school students in California.
  • C. The 500 students in my sample
Clicker Review

- Lets assume that Nevada is a much more diverse state than Utah, and that this diversity extends to IQ (and lets assume their populations are the same size). In Nevada, IQ scores are very spread out, with a comparatively large number of very high IQ scores and very low IQ scores, while in Utah most students score very close to the state average. If I measured a random sample of 500 students in each state:
  - A. My estimate of the population average would be more precise in Utah
  - B. My estimate of the population average would be more precise in Nevada
  - C. My estimate of the population average IQ would be equally precise in each state because the sample size and population size in each state is the same.
EXHIBIT 5.5  Stratified Random Sampling

Population: All residents of community X
n = 10,000

- White: n = 7,000 (70%)
- Black: n = 1,500 (15%)
- Hispanic: n = 1,000 (10%)
- Asian: n = 500 (5%)

Random selection:
- 1 in 20 from each stratum
- 1 in 56 from white stratum;
- 1 in 8 from Hispanic stratum;
- 1 in 12 from black stratum;
- 1 in 4 from Asian stratum

Proportionate sample, n = 500
- White: n = 350 (70%)
- Black: n = 75 (15%)
- Hispanic: n = 50 (10%)
- Asian: n = 25 (5%)

Disproportionate sample, n = 500
- White: n = 125 (25%)
- Black: n = 125 (25%)
- Hispanic: n = 125 (25%)
- Asian: n = 125 (25%)
What happens when I can’t take a random sample?

• Lets say I want to measure the fitness of 40-year old Americans. I want to have each individual run a mile on a track while I time them.
  • How would this look if I used a simple random sample?
  • Is there an easier way that is almost as good?

• **Cluster sampling:**
  • First I take a random sample of houses of states, then a random sample of towns, then a random sample of 20 individuals from each town.
  • 50 tracks instead of 1000.
Non-probability Methods (AKA non-random)

- Stratified random sampling and cluster sampling are still probability sampling methods
- Non-probability methods trade generalizability for logistics
- Probability sampling requires a list of all the units in our population
  - Which we may not have
- Probability sampling is only useful with many cases
  - Rule of thumb: at least 30
  - So its not useful for in-depth study of a few cases
Availability Sampling

• AKA convenience sampling
  • Magazine polls of their readership
  • Polling people who walk by
  • Amazon Turk
• What’s the population you can make an inference about?
Quota Sampling

• Quotas to match your sample to the population
  • 50% women
  • x% poor, x% rich, x% college educated, etc.

• Pros:
  • Matches the population better

• Cons
  • May not match the population on other dimensions
    • Especially unobservable dimensions
  • Requires detailed knowledge about the population
Purposive Sampling

• Targets key individuals with important knowledge
  • e.g. Interview the diplomats who negotiated a particular treaty
  • e.g. Interview the leaders of an important social movement
How big of a sample?

- Probability Sampling:
  - **Power Calculations**: If you know the size of the effect you expect and the variance of the sample, you can calculate how large of a sample you need to detect that effect (with 95% confidence).