Data Collection and Random Sampling

Eco 2470: Economic Statistics

Fall, 2019. Chaoyi Chen

(Chapter 5)

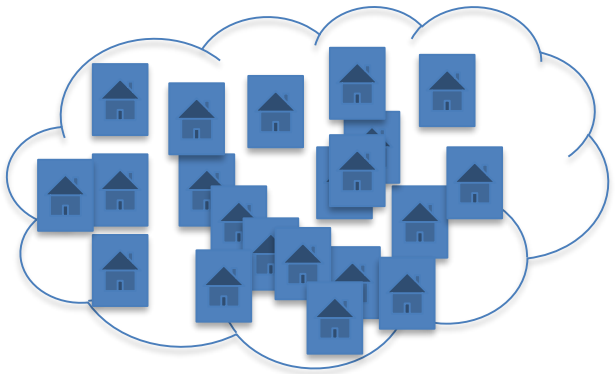
Ways data are collected

- Controlled experiments (e.g. pharmaceutical drug testing)
- Observational data
- Survey data

Data collection: obtain a sample

- A smaller number of individuals drawn from our population.
- For practical and cost saving reasons.

population



sample

What makes a good sample?

- A good sample satisfies just one property:

Apart from its smaller size, it should be similar to the population. Put another way, it should be representative of the population.

- For example, it should have similar looking histogram, quartiles, means, variances, etc.

Why is it so important that the sample be similar to the population?

- Because, we want to able to draw conclusions about the overall population, based on what we learn from the sample.

- E.g. We might use the mean of the sample to draw conclusions about the mean of the population.

- E.g. I pick 10 students in the course to ask how the course could be improved. Are the views of these 10 students representative (similar to) the views of the overall class?

How can we check if the sample and population are similar?

- Answer: Usually we cannot!

- Why? Remember the reason for working with the sample is that we are unable to get or afford the population.

- Otherwise, we would just work with the population directly.

Then, how can we tell if our sample is a good sample?

- We need to ensure that it was collected in a good way. We pay particular attention to two issues:

1. The sample size (n): larger is better (but also more expensive)
2. Is it a random sample? (definition coming soon)

Why is the sample size important for obtaining a representative Sample?

- Because unrepresentative individual(s) may happen to enter your sample by chance.
- With a large sample the impact of the unrepresentative individual will be averaged out,
- but this won’t happen if the sample size is too small.

Examples

- I pick just one or two students to ask how the class can be improved. They say: harder exams, more homework, and lower marks please.

- In an income survey, just 50 Canadians picked at random. One happens to be a CEO.

Sampling errors

- These are examples of Sampling Errors.

- Sampling Errors: Differences between population and sample that occur because of the observations that happened to be picked for our sample.

What is random sample?

Random sample: A sample in which

1. all population observations have an equal chance of being chosen. AND
2. the fact that one observation is chosen does not effect the chance of another individual being chosen (except to the extent that there is one less observation to choose from).

Examples of Random and Non-random sampling

Which examples involve random sampling? Which do not?

- E.g: I write the name of every student in the course on an equally sized piece of paper, put them in a hat, mix them thoroughly, and draw 10 with a blind fold on.
- E.g. Same as above, but I only draw students attending lecture today.

Examples (continued)

- E.g. I select the students with the ten highest midterm marks.

- E.g. I draw one student at random. Then I ask him/her to recruit nine friends or classmates.

Why is Random Sampling Important for Obtaining a Representative Sample?

Let's answer this with an example:

- Suppose we are interested in alcohol consumption by University of Guelph Students.

- Consider three sampling methods to select the students in our sample:

1. Put all students names into a hat and draw students
2. Go to the library on Saturday evening to find students.
3. Go to the student bar on Monday night to find students.

Nonsampling Errors

- nonsampling errors: Differences between population and sample that occur due to a flaw in the sampling method.
- Usually by causing it to deviate from a random sample.
- We just saw two nonsampling error examples on the last slide:

2. Go to the library on Saturday evening to find students.

3. Go to the student bar on Monday night to find students.

Fancier Sampling Methods

<u>Stratified Random Sampling</u>

1. Step 1. Divide the population into two (or more) mutually exclusive groups (stratas).
2. Randomly sample from each strata

E.g. Break the population into male and female. Then separately draw men and women using random sampling.

Why use Stratified Random Sampling?

- Usually, because we want to make sure that we get enough observations from each group we are interested in.
- E.g. Suppose you are collecting data to study differences between the “one percent” and “everyone else”.
  - Only 1% percent of your data falls into the one percent.
  - Without large budget may not get enough one percenters.
  - Solution: use stratified random sampling.

Cluster Sampling

- Cluster Sampling: Random sample of groups or clusters of observations.

- E.g. Draw townships, postal codes, or city blocks at random then survey the residents.

Why Cluster Sampling?

- Because in some cases random sampling may be difficult/costly because:
  - Not easy to draw up a full list of population members to draw from.
  - Population widely dispersed and costly to survey.

Cluster Sampling Example

- E.g. You want to survey family housing arrangements in the townships of northern Canada and need to go in person.
- Random sample hard to conduct because:
  - May be hard to obtain list of individuals living in northern Canada.
  - Expensive/time consuming to get to every individual.
- Cluster Sampling Sampling Solution: Random sample townships instead of households.

Do Fancy Sampling Methods Lead to Nonsampling Error?

- Stratified and Cluster Sampling are not random samples
- and so not directly representative of the population.
- But, they differ from a random sample in known ways.
- Fancy statistical methods can be used to adjust for the fact that they are not a random sample.
- We won't cover these in this course, but here's an example.

Example

- E.g. Suppose true population is 50% male.
- Collect Stratified Sample 30 men and 70 women.
- Take mean sample weight.
- Is it likely to be a good estimate of average population weight?
- But, what if instead we used:

½ (sample mean men)+1/2(sample mean women)?

What if there is no way to get a random sample?

- This is a big challenge in Economics
- Econometrics is a field which develops and applies statistical methods to economic problems.
- Many of the methods in econometrics have developed in response to the difficulty of obtaining random samples in economic data.
- To understand why, let's look at the main types of data available.

Types of Data

- **Experimental Data:**
  - Example: Pharmaceutical Drug Tests.
  - Approximate random sampling possible
- **Survey Data:**
  - Opinion polls
  - Market research
  - Approximate random sample possible with a high response rate (proportion completing the survey)
- **Observational Data:**
  - Common to Economics
  - Example, Daily Closing value of the TSX.
  - Just record what you observe.
  - Often too little control over data collection to ensure a random sample.

Examples of Economic Data that do NOT come from random samples

- Yearly Canadian GDP.
  - We do not choose at random.
  - We get data for every year.
  - And last year's value is correlated to this year's value.
- Bank of Canada Interest Rate
  - The Bank of Canada is not selecting its interest rate at random so that we can have a random sample.
  - It is conducting monetary policy in response to macroeconomic developments.

Observational Data: Direction of causation harder to determine

- Suppose instead of running a clinical trial (experiment), where it is determined **at random** which patients get the medicine
- The drug companies simply observe patients who **choose** or **don't choose** to take the medicine and compared outcomes.
- The second method may result in reserve causality. Why?