STAB22 TUT21

Chong Chen
University of Toronto, Scarborough
Department of Computer and Mathematical Sciences

March 8, 2019

1 Observation and Experiment

In an **observational study** we observe individuals and measure variables of interest but do not attempt to influence the responses.

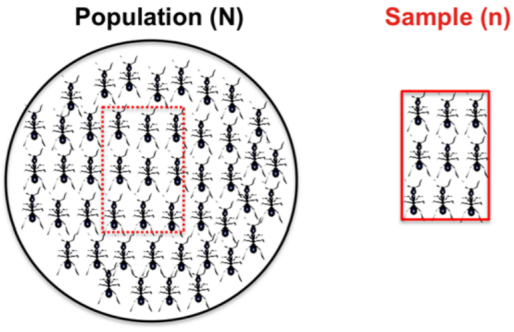
In an **experiment** we deliberately impose some treatment on individuals and we observe their responses.

NOTE: The best way to see the effects of a change is to do an intervention – where we actually impose the change.

2 Population and Sample

The entire group of individuals that we want information about is called the **population**.

A **sample** is a part of the population that we actually examine in order to gather information.



3 Parameter and Statistic

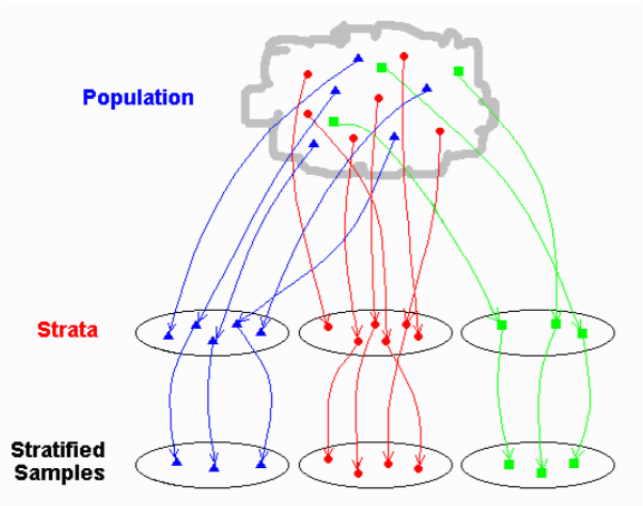
A **parameter** is a number that describes the population. A parameter is a fixed number, but in practice we do not know its value.

A **statistic** is a number that describes a sample. The value of a statistic is known when we have taken a sample, but it can change from sample to sample.

4 Sampling methods

A **simple random sample (SRS)** of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.

4.1 Stratified Random Sample



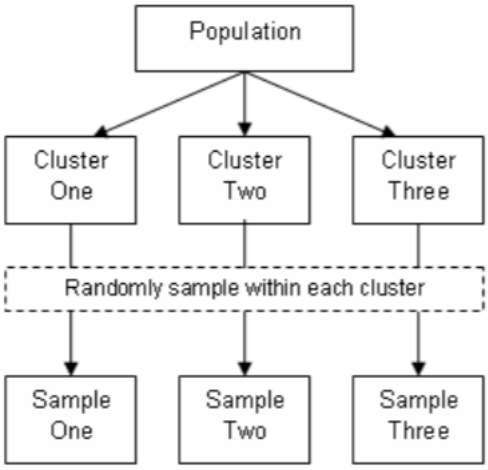
- Divide the population into groups of similar individuals, called strata.
- Then choose a separate SRS in each stratum and combine these SRSs to form the full sample.

4.2 Cluster Sampling

Sometimes stratifying isn't practical and simple random sampling is difficult.

Splitting the population into similar parts or clusters can make sampling more practical.

Then we could select one or a few clusters at random and perform a census (or take a sample if the clusters are large) within each of them.



4.3 Systematic Sampling

Systematic sampling is a statistical method involving the selection of elements from an ordered sampling frame.

4.4 Voluntary Response Sampling

5 Examples

5.1

A computer network manager wants to test the reliability of some new and expensive fiber-optic Ethernet cables that the computer department just received. The computer department received 40 boxes containing 10 cables each. The manager does not have the time to test every cable in each box. He chooses 5 boxes at random and tests all cables within these 5 boxes.
Use this information for this question and the next question.
Which sampling method is applied by the manager to select cables?

(a). Simple random sampling
(b). Stratified sampling
(c). Cluster sampling
(d). Systematic sampling

5.2

Use the information in the previous question.
What is the population in this question?

(a). the 5 boxes of cables chosen at random
(b). the 50 cables within the chosen boxes
(c). the entire shipment of 40 boxes
(d). the entire shipment of 400 cables
(e). the proportion of reliable cables

5.3

Suppose a research team wants to determine the GPA of college students across the U.S. The research team has difficulty collecting data from all 21 million college students; it decides to take a random sample of the population by using $4,000$ college students. The team researches the demographics of college students in the U.S and finds the percentage of students in different majors: 12% major in English, 28% major in science, 24% major in computer science, 21% major in engineering, and 15% major in mathematics. The team uses simple random sampling to select 560 students among English majors. Then, they use simple random sampling independently to select $1,135$ students among science majors. They repeat the same operation to select 800 students among computer science majors, then 1090 students among engineering majors, and 415 students among math majors.
Use this information for this question and the next question.
Which sampling method is applied by the research team to select the sample of $4,000$ college students?

(a). Simple random sampling
(b). Stratified sampling
(c). Cluster sampling
(d). Multistage sampling
(e). Systematic sampling

5.4

Use the information in the previous question.
What is the parameter in this question?

(a) the number of college students in the U.S.
(b) the number of college students selected
(c) the GPA of the 4,000 college students selected in the sample
(d) the GPA of college students across the U.S.