STAT 220 Lecture 7
Data Collection — Sampling
Section 1.3 of OpenIntro 4ed and more

Yibi Huang
Department of Statistics
University of Chicago

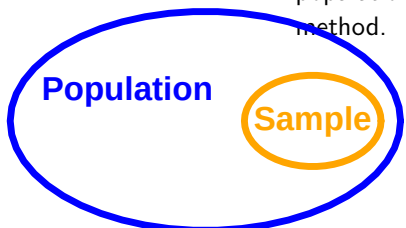
Four Keywords in Sampling

Population: The entire group of individuals in which we are interested but can't usually assess directly.

Example: All humans, all crickets, all working-age people in California

Sample: The part of the population we actually examine and for which we do have data.

How well the sample represents the population depends on the sampling method.



A **parameter** is a number describing a characteristic of the population

A **statistic** is a number describing a characteristic of a sample.

Example: Peanut Allergies

A paper[1] published in 2008 by George Du Toit et al. studied the relationship between peanut allergies and early exposure to peanuts in infancy. The study randomly sampled 3943 Jewish children aged 4-19 in Britain and found 73 (1.85%) of them had peanut allergies. Please identify the **population**, the **sample**, the **parameter** and the **statistic** of this study.

- **Population**: All Jewish children aged 4-19 in Britain
- **Sample**: The 3943 Jewish children aged 4-19 sampled in Britain
- **Parameter**: The proportion of Jewish children aged 4-19 in Britain that had peanut allergies
- **Statistic**: The proportion of the 3943 sampled Jewish children that had peanut allergies

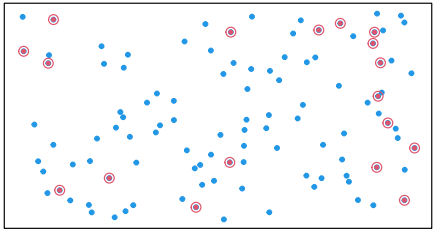
[1] doi:10.1016/j.jaci.2008.08.039

Some Bad (But Commonly Used) Sampling Methods

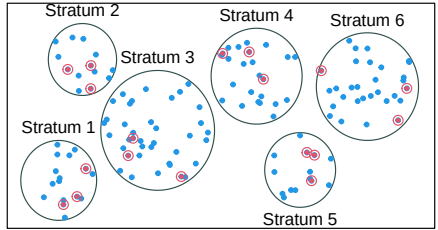
- *Convenience Sampling* — just sampling from those who are easily accessible
  - E.g. "Man on the street" survey (cheap, convenient, popular with TV "journalism")
  - Problem: results may vary greatly with "when and where" the survey is done, lack of representation
- *Voluntary Response Sampling*
  - e.g., internet polls, call-in surveys
  - Only people visit the website/watch the program will be sampled
  - People with strong opinions are more likely to participate

Four Better Sampling Methods

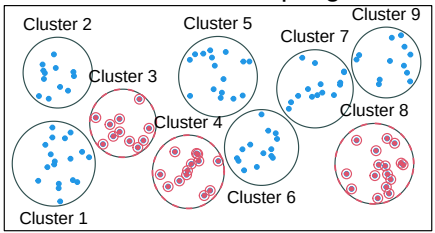
Simple Random Sampling



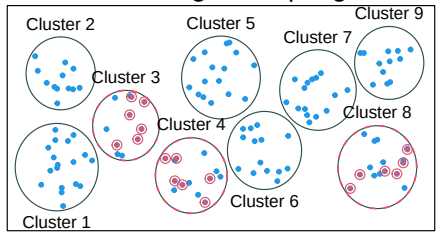
Stratified Sampling



Clustered Sampling

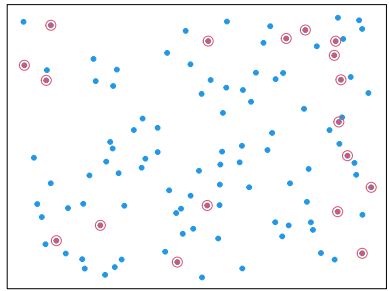


Multistage Sampling



Simple Random Sampling (SRS)

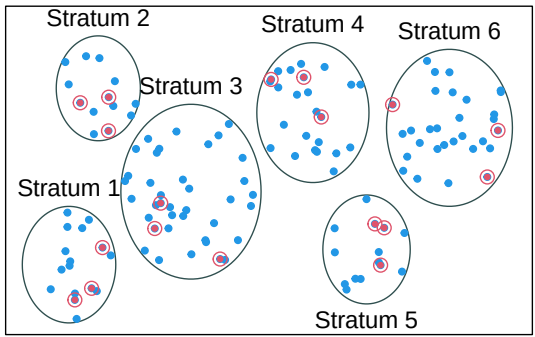
- Basic idea: put the names in a box, shake well, and make draws from the box.
- Need a list of names of all subjects in the population, called the sampling frame
- All subjects have the same chance to be chosen



- Pros: the makeup of the sample will mimic the makeup of the population (age/gender/race/income...), by the Law of Large Number.
- Cons: The need of a sampling frame makes it impractical for large populations

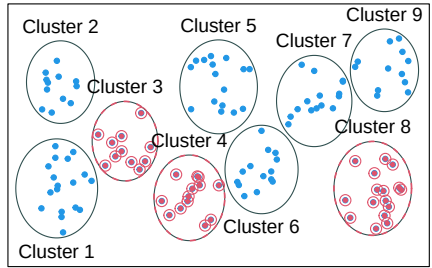
Stratified Sampling

The population is divided into groups called **strata**, and a separate simple random sample is selected from each stratum.



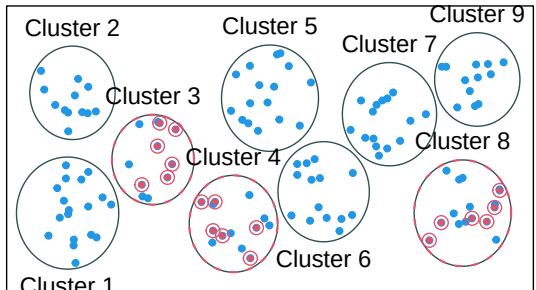
Clustered Sampling

The population is divided into groups, called **clusters**. A sample of clusters is chosen. All subjects in the selected clusters are sampled.



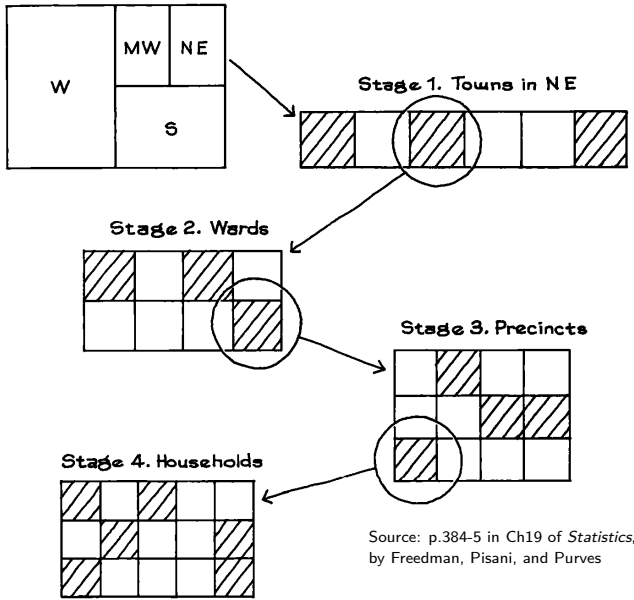
- Ex1: To survey Walmart employees, select several Walmart stores at random and interview every employee in those stores.
  - Here a cluster is a store.
- Ex2: To estimate the percentage of oaks in a forest affected by a disease, divide forests into small regions, randomly pick a few regions, and examine every oak tree in the selected regions.
  - Here a cluster is a region.

Multistage Sampling



- First stage: Divide the population into groups (**clusters**) and randomly select some groups.
- Second stage: Subdivide each selected group into subgroups and randomly pick subgroups within them.
- Third stage: (and so forth...)

Multistage Sampling Used By Gallop Polls (1952-84)



Source: p.384-5 in Ch19 of *Statistics*, 4ed, 2007, by Freedman, Pisani, and Purves

Identify the Sampling Method

1. An airline company randomly selects **two** flights from all international flights that day. All passengers on the two selected flights are surveyed for meal satisfaction.
2. An airline company randomly selects **two** passengers from each of its international flights that day. The selected passengers complete a meal satisfaction survey and receive a reward.
3. An airline company randomly selects 10 flights from all international flights that day, then randomly selects 5 passengers from each chosen flight. Those selected complete a meal satisfaction survey and receive a reward.

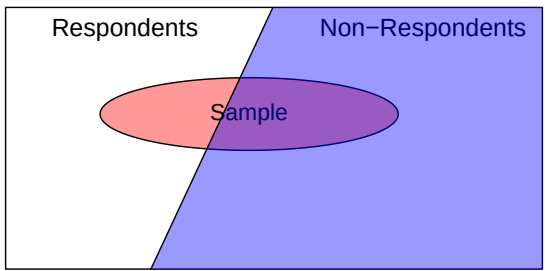
Low Response Rate & Non-Response Bias (1)

Non-respondents can differ significantly from respondents, causing *non-response bias*.

- Non-respondents may have long working hours, live alone, or simply ignore surveys.
- For example, Gallup found that Republicans were more likely to respond than Democrats.

Low Response Rate & Non-Response Bias (2)

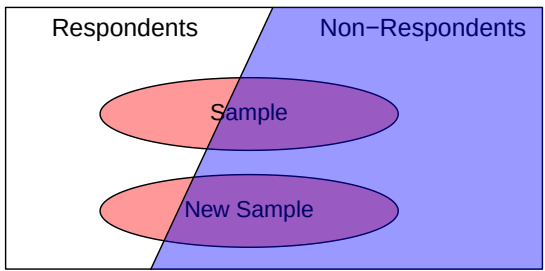
If the response rate is low, taking a new sample to "replace" non-respondents will not solve the problem— since the new sampled subjects are still the respondents, not the non-respondents.



- Efforts must be made to reach non-respondents: more calls, visits, rewards, etc.
- Always check the response rate. If it's low, the results are generally unreliable.

Low Response Rate & Non-Response Bias (2)

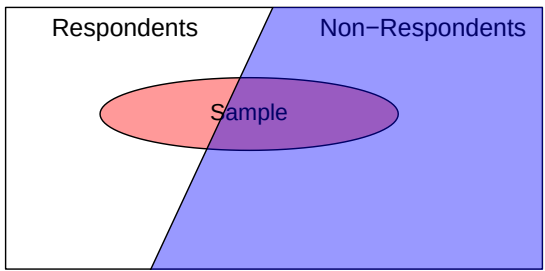
If the response rate is low, taking a new sample to "replace" non-respondents will not solve the problem— since the new sampled subjects are still the respondents, not the non-respondents.



- Efforts must be made to reach non-respondents: more calls, visits, rewards, etc.
- Always check the response rate. If it's low, the results are generally unreliable.

Low Response Rate & Non-Response Bias (2)

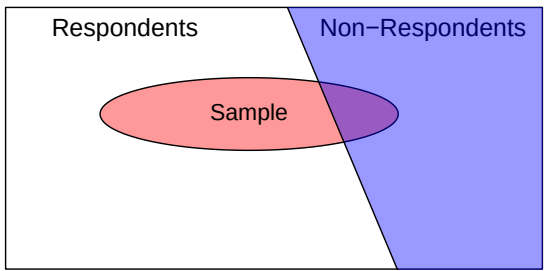
If the response rate is low, taking a new sample to "replace" non-respondents will not solve the problem— since the new sampled subjects are still the respondents, not the non-respondents.



- Efforts must be made to reach non-respondents: more calls, visits, rewards, etc.
- Always check the response rate. If it's low, the results are generally unreliable.

Low Response Rate & Non-Response Bias (2)

If the response rate is low, taking a new sample to "replace" non-respondents will not solve the problem— since the new sampled subjects are still the respondents, not the non-respondents.



- Efforts must be made to reach non-respondents: more calls, visits, rewards, etc.
- Always check the response rate. If it's low, the results are generally unreliable.

Random Assignment vs. Random Sampling

- Whether a result from a sample is generalizable to a larger population depends on whether the sample is randomly sampled from the population
- Whether a cause-and-effect relationship can be inferred depends on whether the subjects are randomly assigned to the treatments (and the control).

Example 1: SSRIs & Bone Fractures

Researchers at McGill University Health Center followed 5008 adults aged 50+ for 5 years to assess if taking SSRIs increases the risk of bone fractures. Results:

Experienced Fracture?	Taking SSRIs? Yes	No	Total
Yes	14	244	258
No	123	4627	4750
Total	137	4871	5008

Bone fracture rates in

- SSRI group: $14/137 \approx 10.2\%$
- no SSRI group: $244/4871 \approx 5.0\%$

Example 1: SSRIs & Bone Fractures (2)

Q1: Can we conclude that taking SSRI increases the risk of bone fractures?

a. Yes, because the study is a comparative experiment
b. Yes, because the sample size 5008 was large
c. No, because the study was not blinded
d. No, because the subjects were not randomized to take SSRI
e. No, because the 5008 subjects were not randomly sampled from the population of interest

Q2: Can we conclude that the 5-year bone fractutre rate for those age 50+ taking SSRIs is around 10.2%?

Example 1: SSRIs & Bone Fractures (2)

Q1: Can we conclude that taking SSRI increases the risk of bone fractures?

a. Yes, because the study is a comparative experiment
b. Yes, because the sample size 5008 was large
c. No, because the study was not blinded
d. No, because the subjects were not randomized to take SSRI
e. No, because the 5008 subjects were not randomly sampled from the population of interest

Q2: Can we conclude that the 5-year bone fractutre rate for those age 50+ taking SSRIs is around 10.2%?

Example: Smart Meter for Saving Energy

A utility company wanted to encourage energy conservation. Executives considered installing electronic meters in homes to display the projected monthly cost based on current usage.

A study was conducted by installing meters in some **randomly selected** households. The executive claimed the meters helped conserve energy because electric usage in these homes dropped by 18% on average compared to the same time period the year before.

Q1: What are the two groups compared in this study?

a. Households with meters vs. households without meters
b. Households with meters vs. the same households in the previous year

Example 2: Smart Meter for Saving Energy (2)

Q2: Can one claim that the meter helps conserve energy?

Q3: If the unselected households also also had meters installed at the same time, would you expect a similar reduction (18%) in electric usage compared to the previous year?