

Lecture 2:  
Sampling Methods  
STAT 630, Fall 2021

## Topics:

- ▶ Sampling design terminology
- ▶ Sampling methods:
  - ▶ Simple random sampling
  - ▶ Stratified sampling
  - ▶ Cluster sampling
  - ▶ Systematic sampling
- ▶ Problems with survey sampling

# Sampling Design Terminology

- ▶ **Population:** The complete collection of individuals, or cases, that we want to study.
- ▶ **Sample:** A subset of the population.
- ▶ **Sampling frame:** The list of all cases from which the sample was taken (e.g., list of street addresses or telephone numbers)
- ▶ A sample is called **representative** if it accurately reflects characteristics of the population. Random sampling strategies are used to collect representative samples.

# Sampling Design Terminology

**Example:** Public opinion polls (such as Gallop or the Washington Post) are used to predict which candidate will win the next election.

- ▶ Population: all registered voters
- ▶ Sampling frame: list of telephone numbers for voters that can be interviewed
- ▶ Sample: subset of voters interviewed by telephone

# Sampling Design Terminology

**Example:** The Environmental Protection Agency (EPA) samples lakes across the U.S. and assesses their condition (good, fair, or poor according to an aquatic health index).

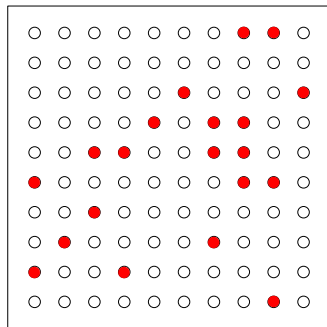
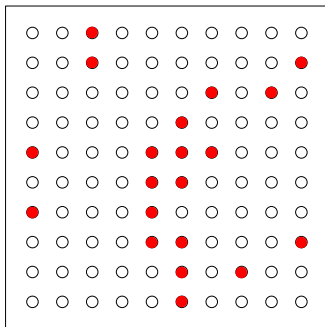
- ▶ Population: all lakes in the U.S.
- ▶ Sampling frame: list of lakes and their locations from a Geographic Information System (GIS) database
- ▶ Sample: subset of lakes selected from the database

# Simple Random Sampling

- ▶ A **simple random sample** (SRS) of size  $n$  is taken when every possible subset of  $n$  distinct cases from the population has the same probability of being selected.
- ▶ For SRS each individual, or case, in the population has the same probability of being included in the final sample.
- ▶ One way to select a SRS of 10 students from this class: write the names of all the students on separate pieces of paper, and place the pieces in a hat and stir. Then draw out 10 pieces from the hat.

# Simple Random Sampling

Two simple random samples of size  $n = 20$  from a population with  $N = 100$  cases.



# Simple Random Sampling

Using R to take a SRS of size  $n = 10$  from a population with  $N = 100$  cases:

```
> sample(1:100, size = 10)
[1] 48 42 49 77 45 96 33 64 98 65
```

```
> sample(1:100, size = 10)
[1] 78 62 58 33 36 15  6 64 41  2
```

```
> sample(1:100, size = 10)
[1] 98 40 53 27  8 29  7 84 59 11
```



# Simple Random Sampling

**Example:** Let  $\{a,b,c,d,e\}$  be a population of size  $N = 5$ . List all possible samples of size  $n = 2$  from this population. For SRS what is the probability of selecting each sample of size  $n = 2$ ?

*Solution:*

There are 10 possible samples of size  $n = 2$ :

$\{a, b\}$ ;  $\{a, c\}$ ;  $\{a, d\}$ ;  $\{a, e\}$ ;  $\{b, c\}$ ;  $\{b, d\}$ ;  $\{b, e\}$ ;  
 $\{c, d\}$ ;  $\{c, e\}$ ;  $\{d, e\}$

For SRS,  $1/10$  is the probability of selecting each sample of size  $n = 2$ .

# Simple Random Sampling

Can use combinations to count # of samples:

$$\binom{5}{2} = \frac{5!}{2!3!} = \frac{5 \cdot 4}{2} = 10$$

In general, there are  $\binom{N}{n} = \frac{N!}{n!(N-n)!}$  samples of size  $n$  from a population with  $N$  cases.

# Simple Random Sampling

**Example:** Suppose a population consists of  $N = 20$  individuals. How many possible samples of size  $n = 4$  can we select from this population? Assume sampling is done **without replacement**.

*Solution:*

$$\binom{20}{4} = \frac{20!}{4!16!} = \frac{20 \cdot 19 \cdot 18 \cdot 17}{4 \cdot 3 \cdot 2 \cdot 1} = \boxed{4845}$$

R command:

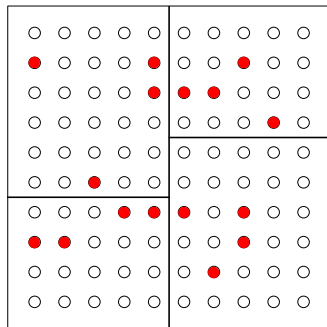
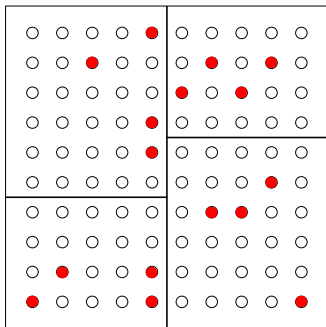
```
> choose(20, 4)
```

# Stratified Sampling

- ▶ For **stratified sampling** the population is divided into distinct groups called **strata**. Then a SRS is selected from from each strata.
- ▶ The strata are selected so that the cases within each strata are similar in some way. For example, the strata might be different ethnic or age groups when surveying people.
- ▶ Commonly used in geographic sampling where the strata can be states or counties.

# Stratified Sampling

Two stratified random samples. Cases are grouped into 4 strata, and a SRS of size 4 is selected within each strata.

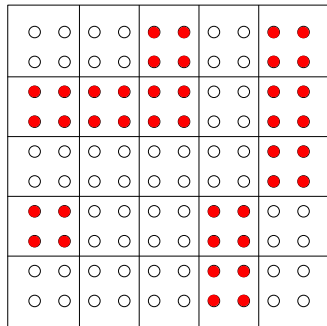
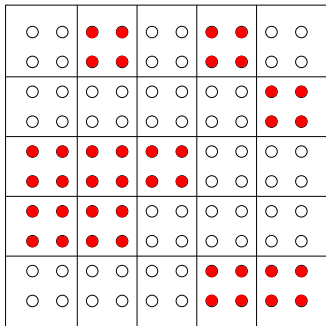


# Cluster Sampling

- ▶ For **cluster sampling** the population is divided up into groups called clusters. Then a fixed number of clusters are randomly sampled, and all cases within each of the selected clusters are included in the sample.
- ▶ For example, suppose we want to survey church members. Instead of taking a SRS of individual church members, we take a random sample of churches (the clusters) and sample all individuals in the selected churches.
- ▶ Unlike stratified sampling, cluster sampling works best when there is a lot variability within a cluster, and the cases within each cluster are representative of the population.

# Cluster Sampling

Two cluster samples. There are 25 clusters and 10 clusters are randomly selected. All cases within each of the selected clusters are included in the sample.



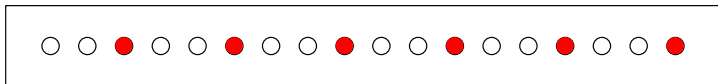
# Systematic Sampling

- ▶ A **systematic sample** is drawn by selecting cases systematically from a sample frame.
- ▶ For example, suppose we have a alphabetical list of names of all students attending CSUEB. We then select a student at the beginning of the list and proceed to select every 10th name thereafter.



# Systematic Sampling

A systematic sample. Every third case is included in the sample.



# Census

- ▶ A **census** is taken if every individual in the population is included in the sample. That is, the sample and the population are the same.
- ▶ Taking a census is more costly and time consuming than random sampling.
- ▶ For large populations, data collection and processing for a census is complex and may be prone to errors.

## Example

Identify the type of sampling design:

- ▶ The selection of 200 people to serve as potential jurors in a trial is conducted by assigning a number to each of 140,000 registered voters in the county. The R command `sample(1:140000, 200)` is used to take a sample of 200 numbers between 1 and 140,000. People having these 200 numbers are sent postcards notifying them of jury duty.  
*Ans: Simple random sampling (SRS)*
- ▶ Suppose you are selecting microchips from a production line for inspection. As the chips process past the inspection point, every 100th chip is selected for inspection.  
*Ans: Systematic sampling*

# Example

Identify the type of sampling design:

- ▶ In a survey on household income, 1000 households are randomly selected in each of the 50 states in the U.S.

*Ans: Stratified sampling*

- ▶ A survey is conducted to find the average weight of cows in a region. A list of all farms is available for the region, and 50 farms are selected at random. Then the weight of each cow at the 50 selected farms is recorded.

*Ans: Cluster sampling*

# Problems with Survey Sampling

A sample is **biased** if it is not representative of the population. Statistics from biased samples tend to overestimate or underestimate the population parameter. Some sources of bias for survey sampling include:

- ▶ **Nonresponse**: failing to obtain responses from some individuals selected for the sample. There may be differences between those that respond and do not respond to a survey.
- ▶ Taking a **sample of convenience** by only including individuals that are easily accessible in the sample.
- ▶ Allowing the sample to consist entirely of volunteers.
- ▶ Wording a survey question in such a way that it influences the response.
- ▶ **Undercoverage**: Using a sample frame that does not include a portion of the population.

## Historical Example: Landon vs. FDR, 1936

- ▶ Literary Digest polled 10 million Americans, and 2.4 million responded
- ▶ Prediction: 43% for FDR
- ▶ Result: 62% for FDR



- ▶ The magazine was so discredited by the poll that it was discontinued.

# Historical Example: Landon vs. FDR, 1936

What went wrong?

- ▶ The magazine had surveyed
  - ▶ its own readers,
  - ▶ registered automobile owners, and
  - ▶ registered telephone users.
- ▶ The sample frame consisted of individuals that were wealthier than the majority of voters, and therefore more likely to support the Republicans (example of undercoverage).
- ▶ Nonresponse: 10 million sampled, but 2.4 million responded. Persons supporting Landon were more likely to have responded to the survey.