Mathematics 231

Lecture 15

Announcements

- Reading
  - Today — Rest of Chapter 3
  - Next class — M&M 4.0- 237-254
    M&M 4.2

Sampling Design

- In sample surveys we want to obtain information from a part of a group to draw conclusions about the whole group.
- Population → Sample
  - **Population**: Entire group of individuals we desire information on.
  - **Sample**: Part of population we actually collect data from.
  - **Sampling Design**: Method used to choose sample from population.

Parameter and Statistic

- **Parameter:** Number that describes the population.
- **Statistic:** Number that describes a sample.

- We use a statistic to estimate an unknown parameter.

Simple Random Survey (SRS)

- In an SRS of size *n*:

1. Each individual has an *equal* chance of being chosen.
2. Every set of *n* individuals has an equal chance of being the sample chosen.

Stratified Samples

- **Basic Idea:** Sample important groups separately, then combine those samples.

1. Divide population into groups of similar individuals, called **strata**.
2. Choose a separate SRS within each strata.
3. Combine these SRS's to form the full sample.

Stratified Samples

- Strata for sampling are similar to blocks in experiments.
- Stratified sampling designs can provide more precise information than an SRS of the same size.
- For example, if all individuals within each stratum are identical, only need one individual from each stratum to perfectly describe the population.

Multistage Samples

- Basic Idea: Choose sample in stages.

- Often used for national surveys (U.S. households).
- Not practical to do SRS from list of all U.S. households (cost, inconvenience, time).

Multistage Samples

- To take a nationwide multistage sample:

1. Take sample from the 3000 counties in the U.S.
2. Take a sample of townships within each county chosen.
3. Take a sample of city blocks within each township chosen.
4. Take a sample of households within each city block.

- At each stage, take random sample (e.g., an SRS)

Pitfalls of Sample Surveys

- **Selection Bias:** Some groups in population are over- or under-represented in sample.
- **Nonresponse Bias:** Nonrespondents may differ in important ways from respondents.
- **Response Bias:** e.g., wording of question, ordering of questions, telescoping in the recall of events.

1936 Literary Digest Poll

- Literary Digest had predicted the winner of every U.S. presidential election since 1916.
- In 1936, Literary Digest mailed questionnaires to 10 million people (25% of voters).
- 2.4 million people responded, the largest number of people ever replying to a poll.
- **Prediction**: Roosevelt 43%, Landon 57%
- **Actual Result**: Roosevelt 62%, Landon 38%

Selection and Nonresponse Bias

- **Selection Bias:** People surveyed came from telephone books, club memberships, mail order lists, automobile ownership lists.
- **Nonresponse Bias:** 76% did not respond.

- The Gallup Poll predicted Roosevelt's victory with a sample of 50,000 people.

Response Bias

- Wording of question can deliberately bias:
  - Do you favor, or do you not favor, increased restrictions on public smoking?
  - Do you favor Gestapo-like police tactics to prevent smoking in public?
  - Do you think smokers have the right to impose their filthy habits on the rest of us, polluting our precious air?

Response Bias

- Social Desirability:
  - Surveys of smoking underestimate the prevalence of smoking and do not match cigarette sales.
- Uninformed:
  - Survey by the American Jewish Committee on attitudes toward various ethnic groups.
  - "30% of respondents expressed an opinion about the Wisians…"

Statistical Inference

- Basic Idea: Use sample (statistic) estimate to infer conclusion about the population (parameter).
- Need to distinguish between sample and population values (statistics vs. parameters).
- Parameters are numbers that describe (unknown) characteristics of the population.
- Statistics are numbers that describe a sample.

Statistical Inference

- Statistics: Numbers that describe a sample.
- We use statistics to estimate unknown parameters.
- Although a statistic is known once we have selected our sample, it can change from sample to sample.
- The is referred to as **sampling variability**.

Sampling Distribution

- Question: What would happen if the sample or experiment were repeated many times?
- Consider the following "thought experiment":
- Take repeated samples of the same size from the same population.
  - 1st sample, calculate the statistic of interest
  - 2nd sample, calculate the statistic of interest
  - 3rd sample, calculate the statistic of interest

  and so on…

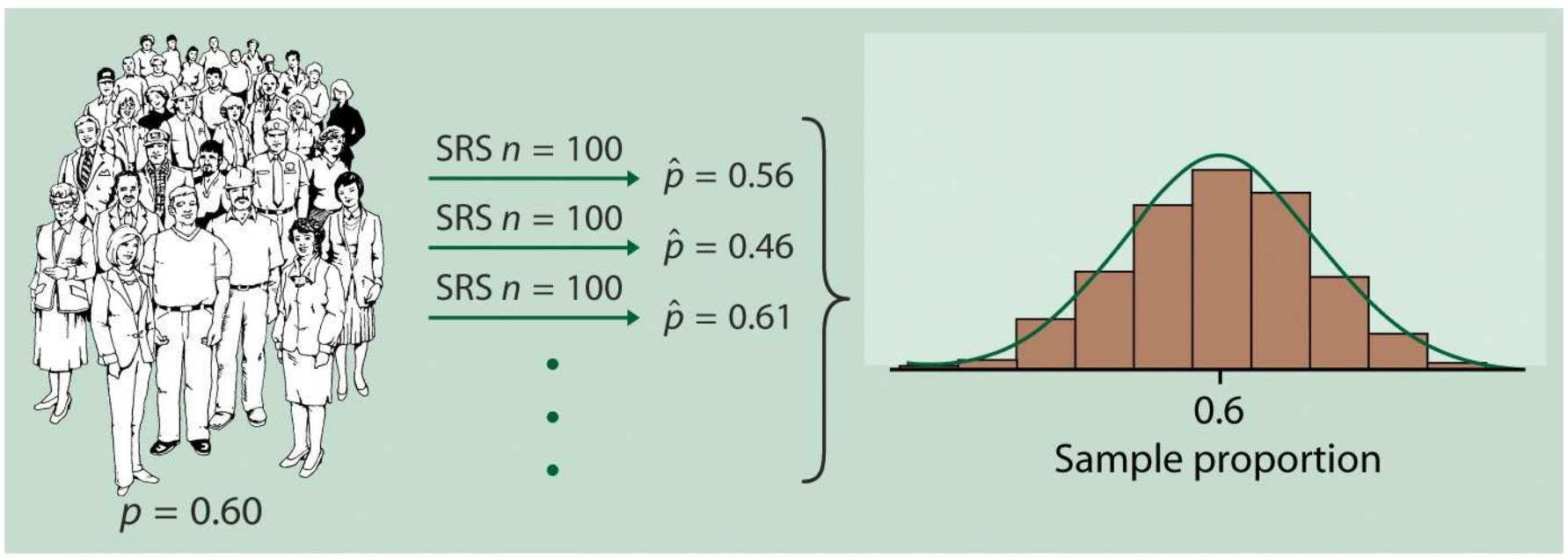
Sampling Distribution

- The statistic will vary from sample to sample due to **sampling variability**.
- **Sampling distribution** of a **statistic** is the distribution of values taken by the statistic in all possible samples of the same size from the same population.
- The sampling variation has a predictable pattern.

Sampling Distribution

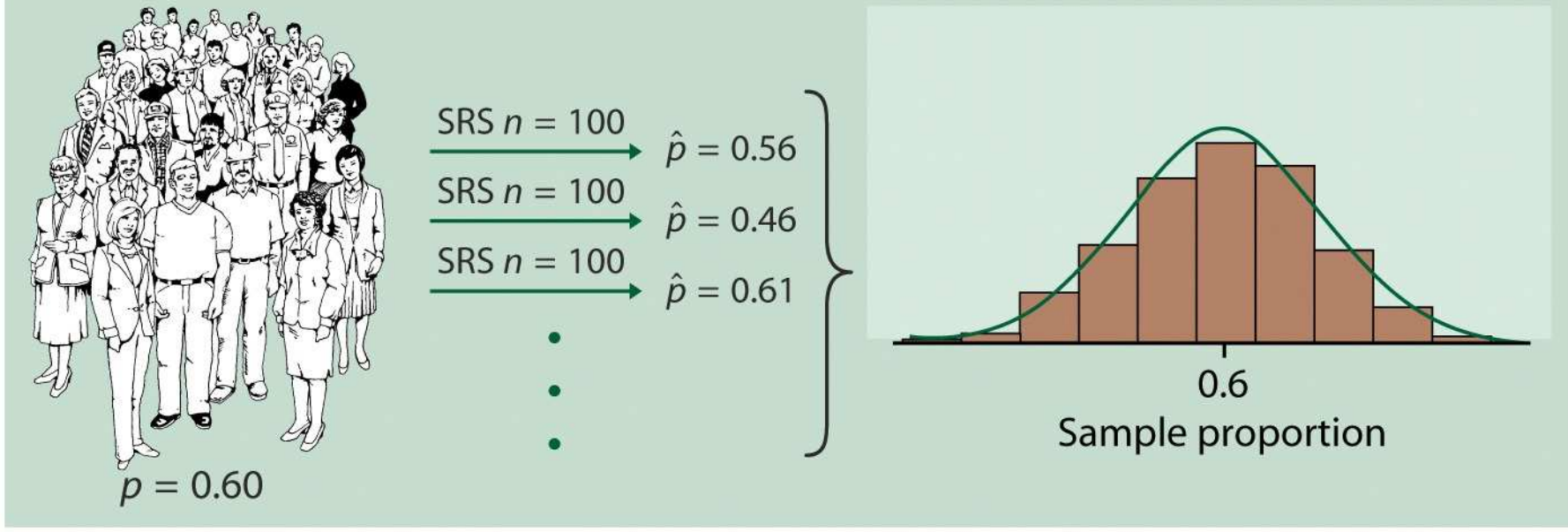
- Example: Opinion Poll

1. Take a large number of samples of size n (e.g., n = 100) from the population.
2. Calculate the sample statistics for each sample (e.g., the proportion of people supporting Bush).
3. Make a histogram of the sample proportions.
4. Examine the distribution and determine center, spread, and shape.


SRS $n = 100$
$\hat{p} = 0.56$
SRS $n = 100$
$\hat{p} = 0.46$
SRS $n = 100$
$\hat{p} = 0.61$
$p = 0.60$
0.6
Sample proportion

Sampling Distribution

- **Center**: Values are centered at the true population parameter.
- **Spread**: Samples of size 1000 are much less variable than samples of size 100.
- **Shape**: Sampling distribution is approximately normal under certain conditions. Moreover, approximation gets better as the sample size, n, gets larger.


SRS $n = 100$
$\hat{p} = 0.56$
SRS $n = 100$
$\hat{p} = 0.46$
SRS $n = 100$
$\hat{p} = 0.61$
$p = 0.60$
0.6
Sample proportion

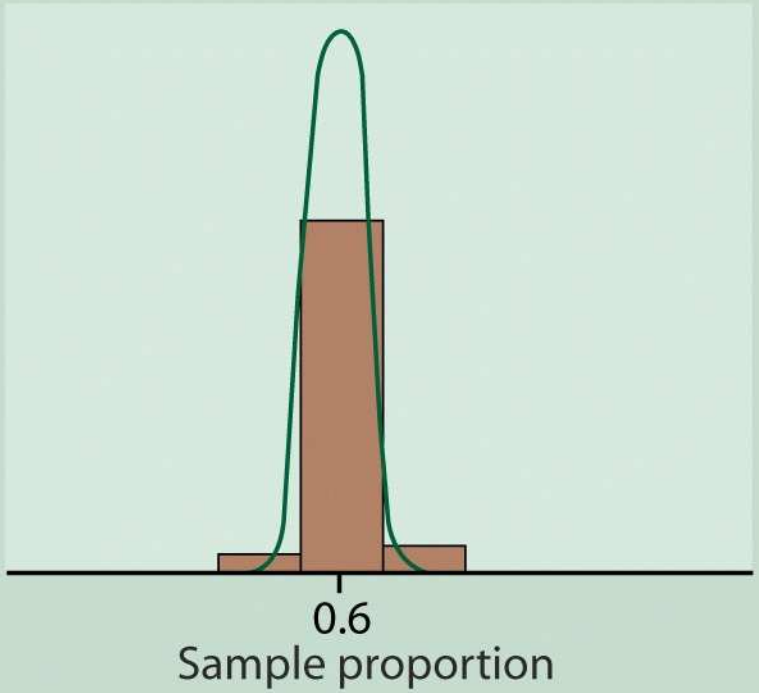
$p = 0.60$

SRS $n = 2500$ → $\hat{p} = 0.609$

SRS $n = 2500$ → $\hat{p} = 0.625$

SRS $n = 2500$ → $\hat{p} = 0.579$

•
•
•



Simplified Example

- Partying habits of statistics students at one of those inferior schools (e.g., Bates).
- Suppose there are only 5 statistics students who went to Westminster High School, then the entire population consists of these 5 students:

Name:	Jerry	Mary	Gary	Larry	Sue
Parties:	2	4	4	6	8

Simplified Example

- Take an SRS of two students from this population.

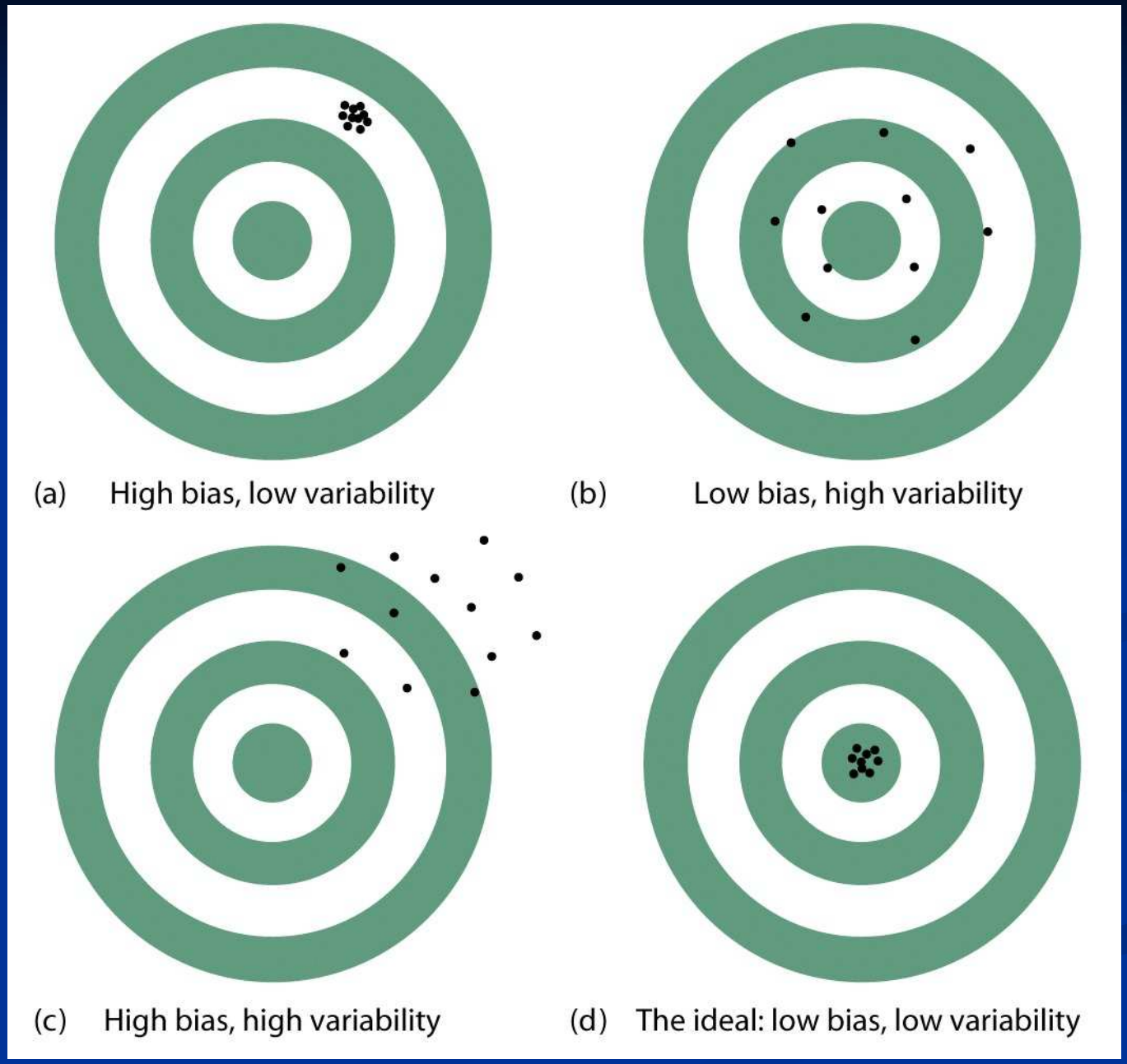
Names	# Parties	Mean
Jerry, Mary	2,4	3
Jerry, Gary	2,4	3
Jerry, Larry	2,6	4
Jerry, Sue	2,8	5
Mary, Gary	4,4	4
Mary, Larry	4,6	5
Mary, Sue	4,8	6
Gary, Larry	4,6	5
Gary, Sue	4,8	6
Larry, Sue	6,8	7

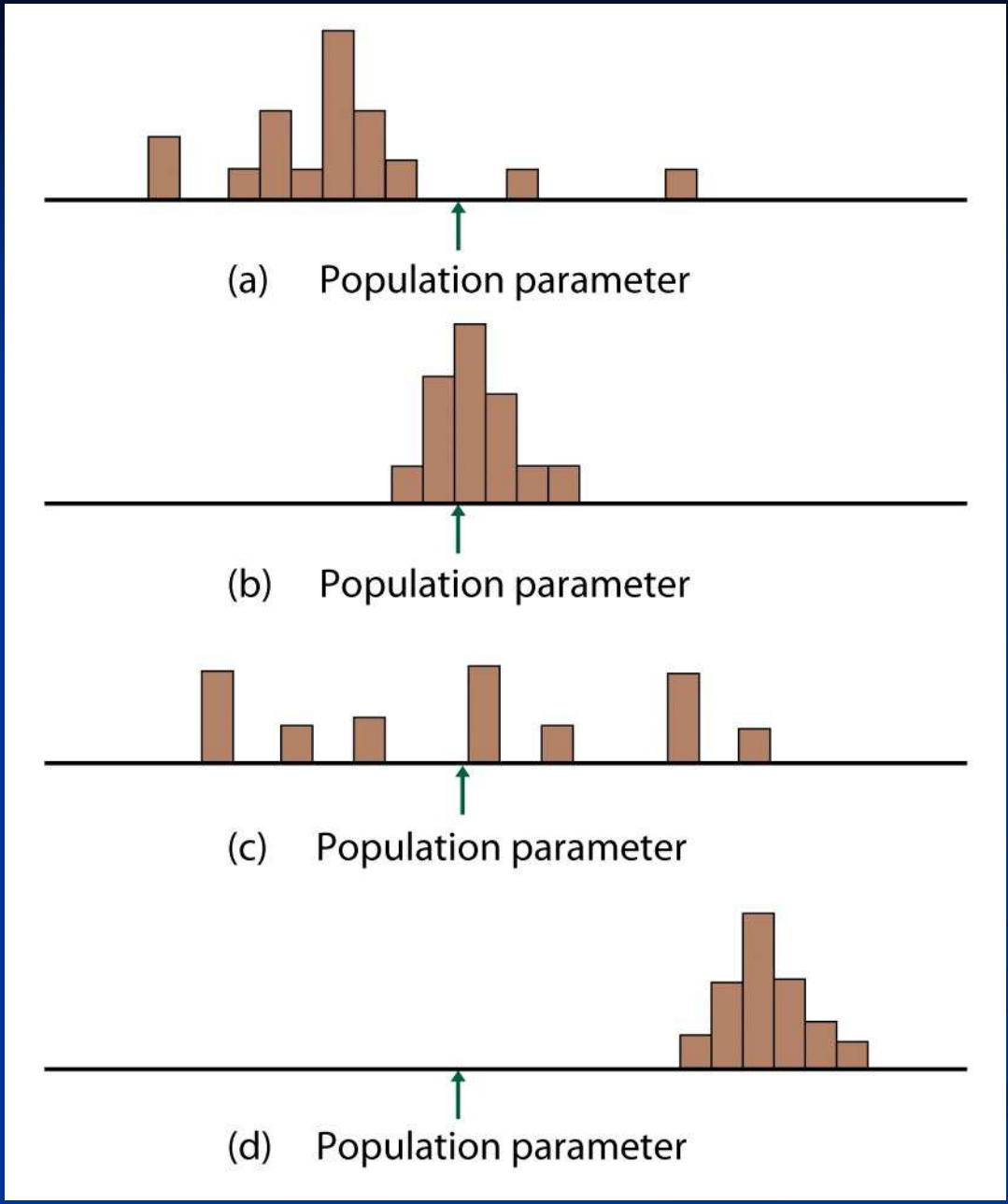
Bias and Variability

- **Bias**: Concerns the center of the sampling distribution.
- A statistic is said to be **unbiased** if the mean of its sampling distribution is equal to the true value of the parameter.
- Bias is reduced by using random sampling.
- If randomization of not done properly, then bias can be introduced. This is BAD.

Bias and Variability

- The variability of a statistic is described by the spread of its sampling distribution.
- Variability is reduced by using a larger sample size, n.
- Results of a sample survey usually come with a **margin of error**.
- This sets bounds on the size of the likely error.


(a) High bias, low variability
(b) Low bias, high variability
(c) High bias, high variability
(d) The ideal: low bias, low variability


(a) Population parameter
(b) Population parameter
(c) Population parameter
(d) Population parameter

Population Size Doesn't Matter

- **Population size doesn't matter.**
- The variability of a statistic from a random sample does not depend on the size of the population (provided the population is substantially larger than the sample).
- Important consequences for surveys:
  - An SRS of 2500 from the more than 210 million adults in U.S. gives results as precise as an SRS of 2500 from the 665,000 inhabitants of San Francisco.

Population Size Doesn't Matter

- Intuition:
  - Imagine you're a chef tasting soup. As long as the soup is well mixed (ensuring a random sample), the variability of the results depends only on the size of the spoon (sample) and not on the size of the pot (population).