Chapter 3: Sampling

Halima Bensmail

CS502

Monday 9-12pm

Room B013

Without data one can't do statistics. Actually, one definition of statistics is the "science of data." However, not all data are created equal. If the collection of data is not carefully planned, they can give a misleading or biased description of the phenomenon of interest.

This chapter discusses concepts and techniques for data collection that reduce the possibility of obtaining biased results. This branch of statistics is called sampling.

Statistical data can be obtained in two ways: observing or experimenting

OBSERVATION VERSUS EXPERIMENT

An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses.

An **experiment** deliberately imposes some treatment on individuals to observe their responses.

Give at least two examples of each.

<u>Experiment</u>

<u>Observational study</u>

When collecting data one needs to be aware of possible confounders. Confounders can render the data useless.

CONFOUNDING

Two variables (explanatory variables or lurking variables) are **confounded** when their effects on a response variable cannot be distinguished from each other.

- “Does job training work? A state institutes a job-training program for manufacturing workers who lose their jobs. After five years, the state reviews how well the program works. Critics claim that because the states unemployment rate for manufacturing workers was 6% when the program began and is 10% five years later, the program is ineffective…”
- Offer a rebuttal. What can be the confouding variable?

Situation:

Population of N individuals (or items)
e.g.

- Students at this university

- Light bulbs produced by a company on one day

Such information about population

- Amount of money students spent on books this quarter

- Percentage of students who bought more than 10 books in this quarter
- Lifetime of light bulbs

Full data collection is often not possible because it is e.g.

- too expensive
- too time consuming
- not sensible (e.g. testing every produced light bulbs for Its lifetime)

Statistical approach:

- collect information from part of the population (sample)
- use information on sample to draw conclusions on whole population

Fundamental concepts in statistics

POPULATION, SAMPLE

The **population** in a statistical study is the entire group of individuals about which we want information.

A **sample** is a part of the population from which we actually collect information, used to draw conclusions about the whole.

VOLUNTARY RESPONSE SAMPLE

A **voluntary response sample** consists of people who choose themselves by responding to a general appeal. Voluntary response samples are biased because people with strong opinions, especially negative opinions, are most likely to respond.

Voluntary response samples almost always cause a bias called "selection bias"

- Should the U.N. headquarters stay in NY?

Nightline poll with 186,000 respondents: 67% say No

Another poll regarding the same question: 72% say Yes

BIAS

The design of a study is **biased** if it systematically favors certain outcomes.

A telephone poll conducted on weekdays
from 7 pm to 9 pm
is almost certain to have bias. Why?

One method to select a sample randomly is simple random sampling.

SIMPLE RANDOM SAMPLE

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.

Assuring "randomness"

RANDOM DIGITS

A **table of random digits** is a long string of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 with these two properties:

1. Each entry in the table is equally likely to be any of the 10 digits 0 through 9.
2. The entries are independent of each other. That is, knowledge of one part of the table gives no information about any other part.

Nowadays, random samples are selected using the computer

CHOOSING AN SRS

Choose an SRS in two steps:

Step 1: Label. Assign a numerical label to every individual in the population.

Step 2: Table. Use Table B to select labels at random.

Suppose we begin at line 130 in Table B:

69051 64817 87174 09517 84534 06489 87201 97245

The first 10 two-digit groups in this line are :

69 05 16 48 17 87 17 40 95 17

We have 30 small business
We want to interview 5 clients each from a company
Selected randomly

01 A-1Plumbing	16 JLRecords
02 AccentPrinting	17 JohnsonCommodities
03Action Sport Shop	18 KeiserConstruction
04 AndersonConstruction	19 Liu'sChineseRestaurant
05 BaileyTrucking	20 MagicTan
06 BalloonsInc.	21 PeerlessMachine
07 BennettHardware	22 PhotoArts
08Best's Camera Shop	23River City Books
09Blue Print Specialties	24 RiversideTavern
10Central Tree Service	25 RusticBoutique
11 ClassicFlowers	26 SatelliteServices
12 ComputerAnswers	27 ScotchWash
13 Darlene'sDolls	28 SewingCenter
14 FleischRealty	29 TireSpecialties
15 HernandezElectronics	30Von's Video Store

Suppose we begin at line 130 in Table B:

69051 64817 87174 09517 84534 06489

The first 5 two-digit groups in this line are :

69 05 16 48 17 87 17 40 95 17

Label 00 and 31 to 99 are not used in this example so we ignore them

Using software for randomization. Let's do a randomization similar to the one we did before, but this time using Excel. Suppose we have to select 10 experimental units

Or subjects that can be assigned to two groups of 5/5.

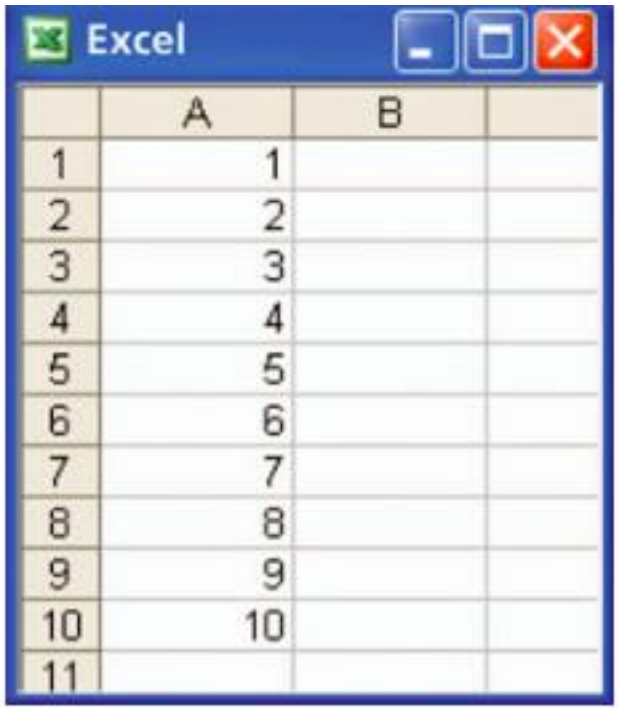
We will assign 5 to the treatment group and 5 to the control group.

We first create a data set with the numbers 1 to 10 in the first column.

Then we use RAND() to generate 10 random numbers in the second column.

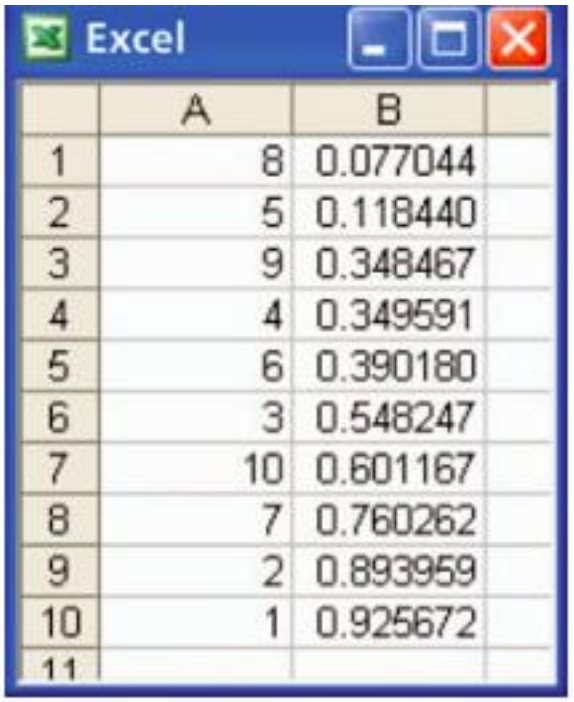
Finally, we sort the data set based on the numbers in the second column.

The first 5 labels (8, 5, 9, 4, and 6) are assigned to the experimental group. The remaining 5 labels (3, 10, 7, 2, and 1) correspond to the control group.



	A	B
1	1	
2	2	
3	3	
4	4	
5	5	
6	6	
7	7	
8	8	
9	9	
10	10	
11		

	A	B
1	1	0.925672
2	2	0.893959
3	3	0.548247
4	4	0.349591
5	5	0.118440
6	6	0.390180
7	7	0.760262
8	8	0.077044
9	9	0.348467
10	10	0.601167



	A	B
1	8	0.077044
2	5	0.118440
3	9	0.348467
4	4	0.349591
5	6	0.390180
6	3	0.548247
7	10	0.601167
8	7	0.760262
9	2	0.893959
10	1	0.925672
11		

PROBABILITY SAMPLE

A **probability sample** is a sample chosen by chance. We must know what samples are possible and what chance, or probability, each possible sample has.

STRATIFIED RANDOM SAMPLE

To select a **stratified random sample,** first 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.

Stratified sampling can yield more precise Information than simple random sampling. They can also be easier to carry out.

Examples:

Population: Students at this university

Objective: Amount of money spent on books this quarter

Knowledge: Students e.g. humanities spent more money on books

Use knowledge to build samples

Divide sample into groups of similar individuals called strata

Choose simply random sample within each group

Size of samples in each group e.g. proportional to size of population

can reduce variability of estimate significantly

UNDERCOVERAGE AND NONRESPONSE

Undercoverage occurs when some groups in the population are left out of the process of choosing the sample.

Nonresponse occurs when an individual chosen for the sample can't be contacted or refuses to cooperate.

These two problems – **<u>two ways to be systematically excluded</u>** - exist in many samples. They can produce serious biases.

Fundamental concepts of statistics, cont.

PARAMETERS AND STATISTICS

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. We often use a statistic to estimate an unknown parameter.

Parameter - number that describes the population, e.g.

$$\mu_{\text{pop}} = \frac{1}{N}\sum_{j=1}^{N} \tilde{x}_j \qquad \text{population mean}$$

$$\sigma^2_{\text{pop}} = \frac{1}{N}\sum_{j=1}^{N} (\tilde{x}_j - \mu_{\text{pop}})^2 \qquad \text{population variance}$$

Estimate population parameter from sampled values:

$$\hat{\mu}_{\text{pop}} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{sample mean}$$

$$\hat{\sigma}^2_{\text{pop}} = s^2 = \frac{1}{n-1}\sum_{j=1}^{N} (x_j - \bar{x})^2 \qquad \text{sample variance}$$

In practice, the objective is to generalize or infer results to a population from a sample. In other words, we use **statistics (**from samples) to estimate **parameters (**from populations**)**.

Statistical inference is the branch of statistics that develops methods to use statistics to infer from a sample to a population. An important feature of these methods is that they give ways to assess the reliability of these inferences. E.g.,

Suppose we are interested in the amount of money students at this university have spent on books this quarter.

Idea: Ask 20 students about the amount they have spent and take the average.

The value we obtain will vary from sample to sample, that is, if we asked another 20 students we would get a different answer.

In our example, the sampling distribution of the average amount obtained from the sample depends on the way we choose the sample from the population:

- Ask 20 students in this class.
- Ask 20 students in your department.
- Ask 20 students in the University bookshop.
- Select randomly 20 students from the register of the university.

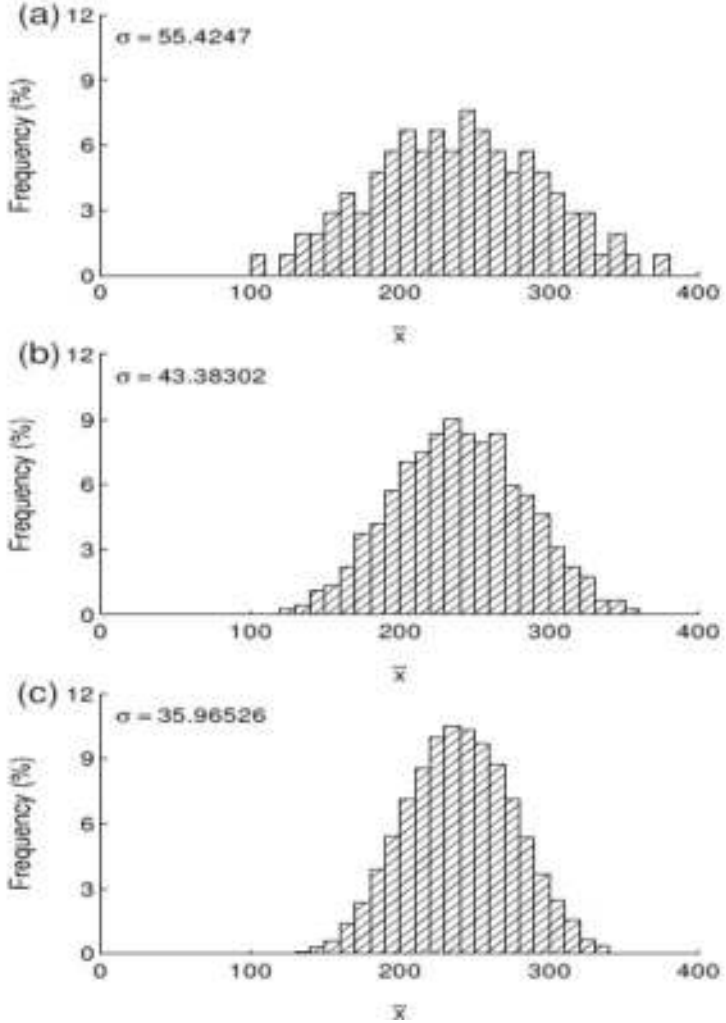
SAMPLING DISTRIBUTION

The **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.

Example:

Consider a population of 20 students who spent the following amounts on books:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
100	120	150	180	200	220	220	240	260	280	290	300	310	350	400



Sampling distribution of

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

for sample sizes

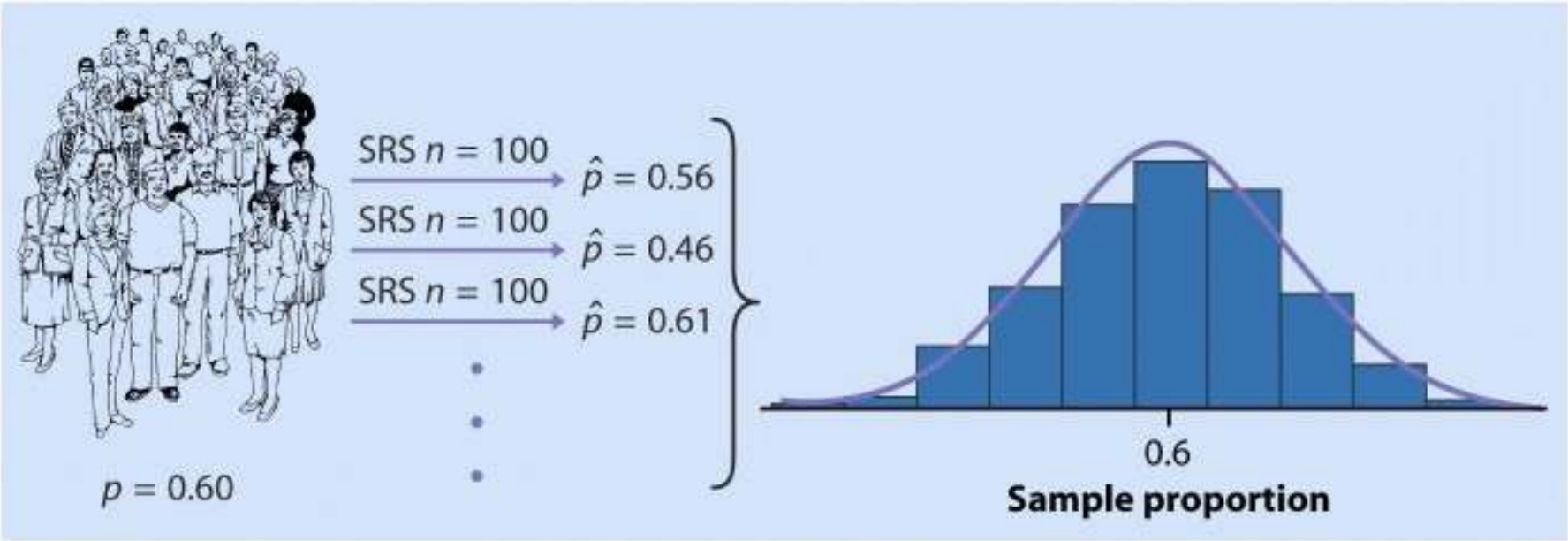
(a) $n = 2$

(b) $n = 3$

(c) $n = 4$

Sample size = 100

Conducting a survey on shopping
Question: “I like buying new clothes, but shopping is often frustrating and time-consuming.”
A sample of 250 was asked, 150 answered yes, estimated p= 150/250=0.6

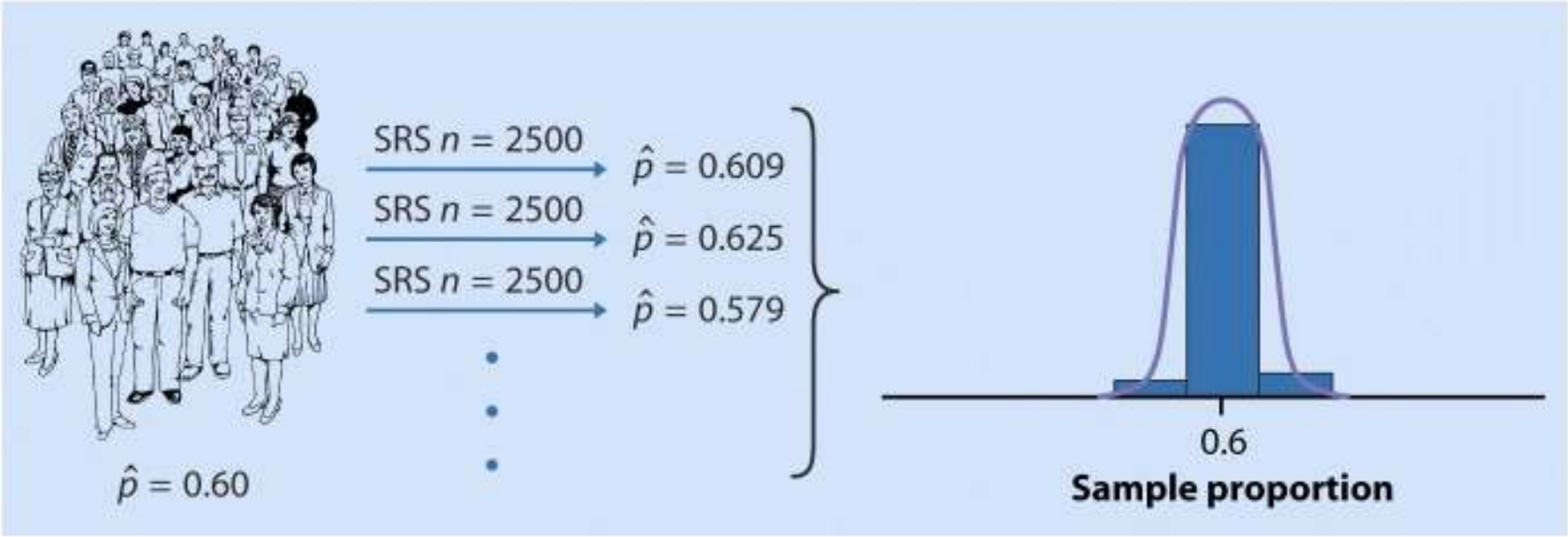


Repeated sampling

Sampling Distribution

Sampling 1000 times

Sample size = 2,500



Repeated sampling

Sampling Distribution

What has changed about the distribution of sample proportions?

Good estimators have low bias and variability

BIAS AND VARIABILITY

Bias concerns the center of the sampling distribution. A statistic used to estimate a parameter is **unbiased** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

The **variability of a statistic** is described by the spread of its sampling distribution. This spread is determined by the sampling design and the sample size *n*. Statistics from larger samples have smaller spreads.

We can think of the following:

True value as the Bull’s eyes on a target

Sample statistics as an arrow fired at the bull’s eyes

Bias means: The arrow is off

(a) Large bias, small variability

(b) Small bias, large variability

(c) Large bias, large variability

(d) Small bias, small variability

MANAGING BIAS AND VARIABILITY

To reduce bias, use random sampling. When we start with a list of the entire population, simple random sampling produces **unbiased estimates**—the values of a statistic computed from an SRS neither consistently overestimate nor consistently underestimate the value of the population parameter.

To reduce the variability of a statistic from an SRS, use a larger sample. You can make the variability as small as you want by taking a large enough sample.

Sample of 25 movies released in the 1990s

Movie name (year)	Domestic Gross ($ millions)
Aces: Iron Eagle III (1992)	2.5
BASEketball (1998)	7.0
Body of Evidence (1993)	13.3
Car 54, Where Are You? (1994)	1.2
City of Angels (1998)	78.9
Coneheads (1993)	21.3
Days of Thunder (1990)	82.7
Death Warrant (1990)	16.9
Desperate Hours (1990)	2.7
Ernest Scared Stupid (1991)	14.1
Executive Decision (1996)	68.8
For Love of the Game (1999)	35.2
I Come in Peace (1990)	4.3
Jumanji (1995)	100.2
Kika (1993)	2.1
Miami Rhapsody (1995)	5.2
Mighty, The (1998)	2.6
Perez Family, The (1995)	2.8
Revenge (1990)	15.7
Shine (1996)	35.8
So I Married an Axe Murderer ((1993)	11.6
Thinner (1996)	15.2
Wedding singer (1998)	80.2
Wing commander (1999)	11.6
Xizao (1999)	1.2

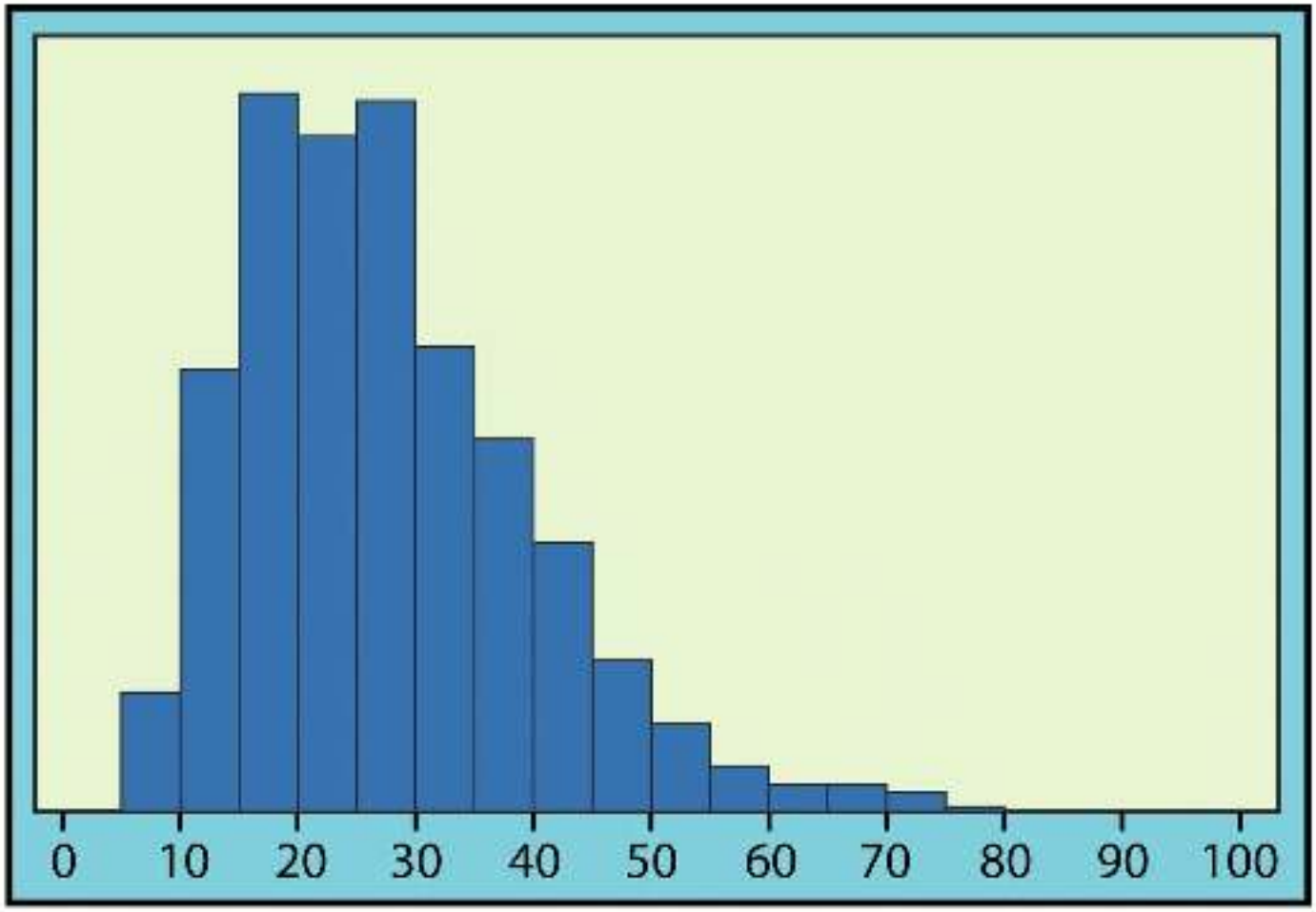
- **Total of 1986 during 1990 were released**
- **We used a SRS to choose 25 movies**
- **Gross sales mean = 33.916 million$**
- **Gross sales std = 50.2697 million dollars**
- **Gross sales median = 16.1 million dollars**
- **Question**
- **Re-calculate the same using the table**
- **How does it compare with the first sample**

The sample(x, n, replace = FALSE, prob = NULL) function takes a sample from a vector x of size n. This sample can be with or without replacement, and the probabilities of selecting each element to the sample can be either the same for each element, or a vector informed by the user.

```
mean(sample(x, 25))
[1] 21.73
> mean(sample(x, 25))
[1] 25.38
> mean(sample(x, 25))
```

Sampling distribution for $\bar{X}$ based on samples of n = 25 movies

We can generate up to 1000 samples



Sampling distribution for $\bar{X}$ based on samples of n = 100 movies
(How is this similar or different from the previous histogram?)

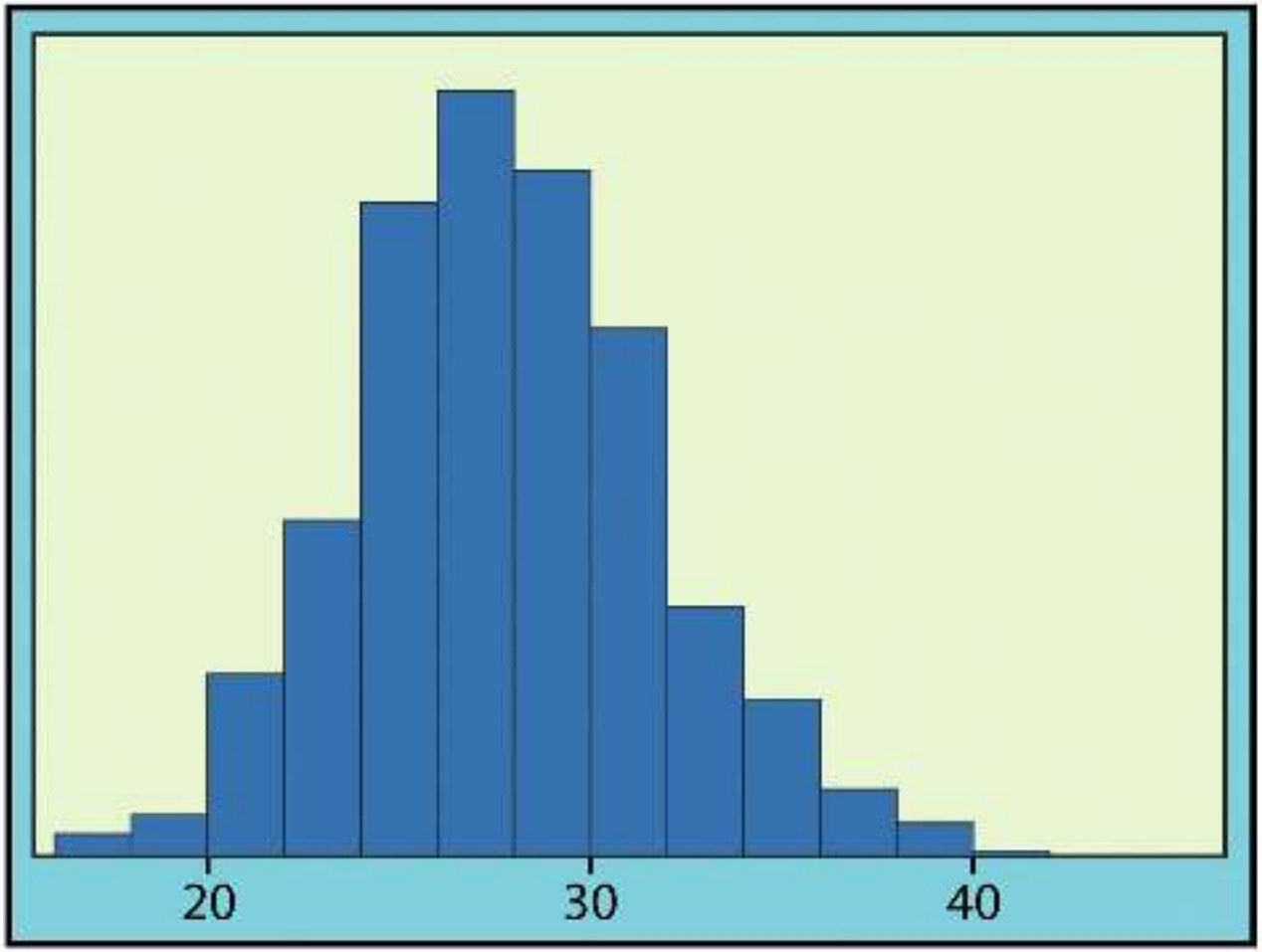
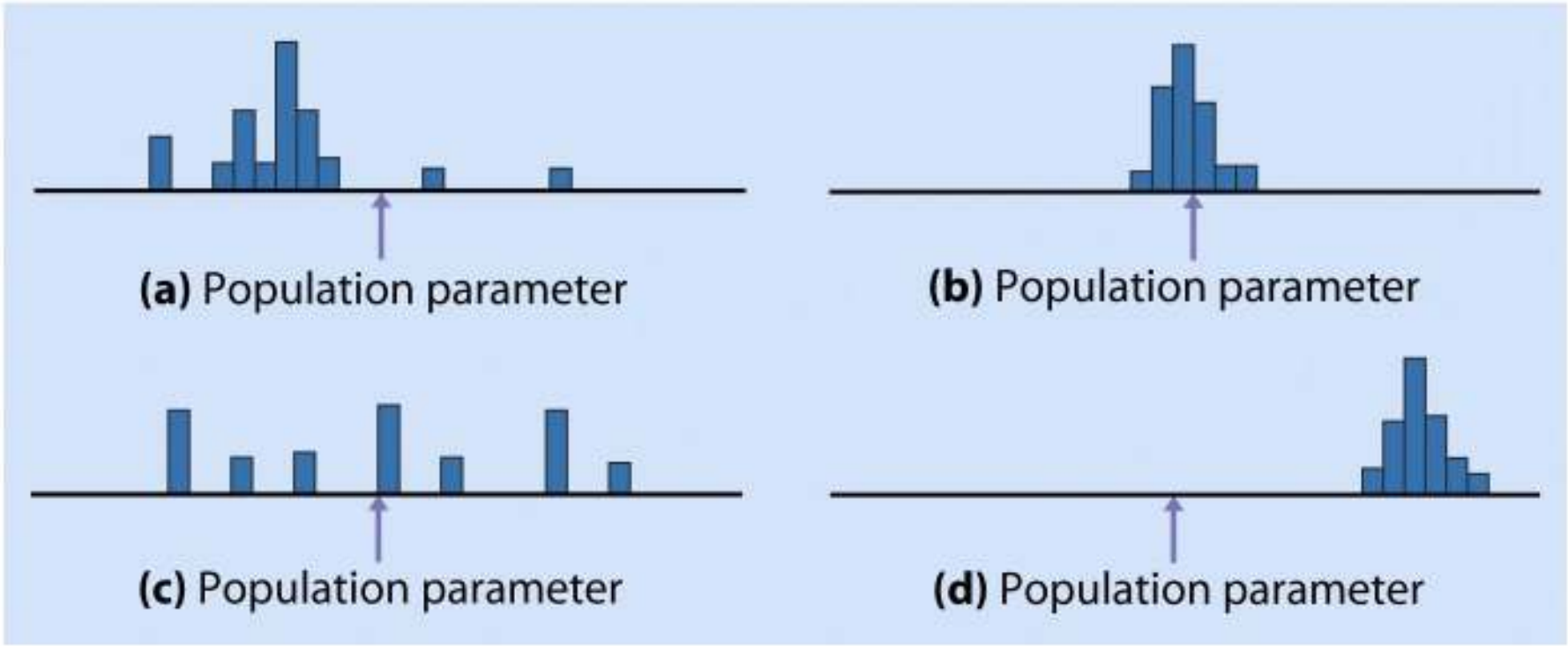


Figure 3.13 (p. 215)

POPULATION SIZE DOESN'T MATTER

The variability of a statistic from a random sample does not depend on the size of the population, as long as the population is at least 100 times larger than the sample.

Which of the following sampling distributions is more desirable? Why?



STATISTICS IN SUMMARY
Simple Random Sample