Chapter 7
Sampling and Sampling Distributions

- Selecting a Sample
- Point Estimation
- Introduction to Sampling Distributions
- Sampling Distribution of $\bar{x}$
- Sampling Distribution of $\bar{p}$
- Other Sampling Methods
- Big data and Errors in Sampling

CENGAGE

Introduction (1 of 2)

- An element is the entity on which data are collected.
- A population is a collection of all the elements of interest.
- A sample is a subset of the population.
- The sampled population is the population from which the sample is drawn.
- A frame is a list of the elements that the sample will be selected from.

Introduction (2 of 2)

- The reason we select a sample is to collect data to answer a research question about a population.
- The sample results provide only estimates of the values of the population characteristics.
- The reason is simply that the sample contains only a portion of the population.
- With proper sampling methods, the sample results can provide “good” estimates of the population characteristics.

Selecting a Sample

- Sampling from a Finite Population
- Sampling from an Infinite Population

Sampling from a Finite Population (1 of 4)

- Finite populations are often defined by lists such as:
  - Organization membership roster
  - Credit card account numbers
  - Inventory product numbers
  - A simple random sample of size *n* from a finite population of size *N* is a sample selected such that each possible sample of size *n* has the same probability of being selected.

Sampling from a Finite Population (2 of 4)

- In large sampling projects, computer-generated random numbers are often used to automate the sample selection process.

Sampling from a Finite Population (3 of 4)

Example: National Baseball League teams

There are 15 teams that played in the 2019 National Baseball League. Suppose we want to select a simple random sample of 5 teams to conduct in-depth interviews about how they manage their minor legal franchises.

Sampling from a Finite Population (4 of 4)

Example: National Baseball League teams

Step 1: Assign a random number to each of the 15 teams in the population.

The random numbers generated by Excel's *RAND* function follow a uniform probability distribution between 0 and 1.

Step 2: Select the five teams corresponding to the 5 smallest random numbers as the sample.

Sampling from a Finite Population Using Excel (1 of 4)

- Excel Formula Worksheet

	A	B
1	Team	Random Numbers
2	Arizona	=RAND()
3	Atlanta	=RAND()
4	Chicago	=RAND()
5	Cincinnati	=RAND()
6	Colorado	=RAND()
7	Los Angeles	=RAND()
8	Miami	=RAND()
9	Milwaukee	=RAND()
10	New York	=RAND()
11	Philadelphia	=RAND()
12	Pittsburgh	=RAND()
13	San Diego	=RAND()
14	San Francisco	=RAND()
15	St. Louis	=RAND()
16	Washington	=RAND()
17		
18		

Sampling from a Finite Population Using Excel (2 of 4)

- Excel Value Worksheet

	A	B
1	Team	Random Numbers
2	Arizona	0.850862
3	Atlanta	0.706245
4	Chicago	0.724789
5	Cincinnati	0.614784
6	Colorado	0.553815
7	Los Angeles	0.857324
8	Miami	0.179123
9	Milwaukee	0.525636
10	New York	0.471490
11	Philadelphia	0.523103
12	Pittsburgh	0.851552
13	San Diego	0.806185
14	San Francisco	0.327713
15	St. Louis	0.374168
16	Washington	0.066942
17		
18		

Sampling from Finite Population Using Excel (3 of 4)

Use Excel's sort procedure to select the five teams corresponding to the five smallest random numbers.

Steps:

- Select any cell in the range B2:B16
- Click on **Home** tab on the Ribbon
- In the **Editing** group click **Sort & Filter**
- Choose **Sort smallest to largest**

Sampling from Finite Population Using Excel (4 of 4)

- Excel Value Sheet (Sorted)

	A	B	C
1	Team	Random Numbers	
2	Washington	0.066942	
3	Miami	0.179123	
4	San Francisco	0.327713	
5	St. Louis	0.374168	
6	New York	0.471490	
7	Philadelphia	0.523103	
8	Milwaukee	0.525636	
9	Colorado	0.553815	
10	Cincinnati	0.614784	
11	Atlanta	0.706245	
12	Chicago	0.724789	
13	San Diego	0.806185	
14	Arizona	0.850862	
15	Pittsburgh	0.851552	
16	Los Angeles	0.857324	
17			
18			

Sampling from an Infinite Population (1 of 3)

- Sometimes we want to select a sample but find it is not possible to obtain a list of all elements in the population.
- As a result, we cannot construct a frame for the population.
- Hence, we cannot use the random number selection procedure.
- Most often this situation occurs in infinite population cases.

Sampling from an Infinite Population (2 of 3)

- Populations are often generated by an ongoing process where there is no upper limit on the number of units that can be generated.
- Some **examples** of on-going processes, with infinite populations, are:
  - Parts being manufactured on a production line
  - Transactions occurring at a bank
  - Telephone calls arriving at a technical help desk
  - Customers entering a store

Sampling from an Infinite Population (3 of 3)

- In the case of an infinite population, we must select a random sample in order to make valid statistical inferences about the population from which the sample is taken.

- A random sample from an infinite population is a sample selected such that the following conditions are satisfied.

  1. Each element selected comes from the population of interest.
  2. Each element is selected independently.

Point Estimation (1 of 5)

- Point estimation is a form of statistical inference.
- In point estimation we use the data from the sample to compute a value of a sample statistic that serves as an estimate of a population parameter.
- We refer to $\bar{x}$ as the point estimator of the population mean μ.
- s is the point estimator of the population standard deviation σ.
- $\bar{p}$ is the point estimator of the population proportion p.

Point Estimation (2 of 5)

Example: EAI Employee Data

Out of a total number of 2,500 employees, a simple random sample of 30 employees and corresponding data on annual salary and management training program participation are shown in the given table. $x_1, x_2 \ldots x_n$ is used to denote annual salary of the employees, and the participation in the management training program is indicated by a yes / no.

Point Estimation (3 of 5)

Annual Salary ($)	Management Training Program	Annual Salary ($)	Management Training Program	Annual Salary ($)	Management Training Program
X_1 = 69094.30	YES	65,922.60	YES	65,120.90	YES
X_2 = 73263.90	YES	77,268.40	NO	71,753.00	YES
69,643.50	YES	75,688.40	YES	74,391.80	NO
69,894.90	YES	71,564.70	NO	70,164.20	NO
67,621.60	NO	76,188.20	NO	72,973.60	NO
75,924.00	YES	71,766.00	YES	70,241.30	NO
69,092.30	YES	72,541.30	NO	72,793.90	NO
71,404.40	YES	64980.00	YES	70,979.40	YES
70,957.70	YES	71,932.60	YES	75,860.90	YES
75,109.70	YES	72,973.00	YES	77,309.10	NO

Point Estimation (4 of 5)

To estimate the value of population parameter, we can compute the corresponding characteristic of the sample, referred to as a **sample statistic.**

Note: Different random numbers will identify different sample which would result different point estimates.

Point Estimation (5 of 5)

Example: EAI Employee Data

- $\bar{x}$ as Point Estimator of μ

$$\bar{x} = \frac{\sum x_i}{n} = \frac{2{,}154{,}420}{30} = \$71{,}814$$

- s as Point Estimator of σ

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{325{,}009{,}260}{29}} = \$3348$$

- $\bar{p}$ as Point Estimator of p

$$\bar{p} = \frac{x}{n} = {}^{19}\!/_{30} = .63$$

Summary of Point Estimates Obtained from a Simple Random Sample

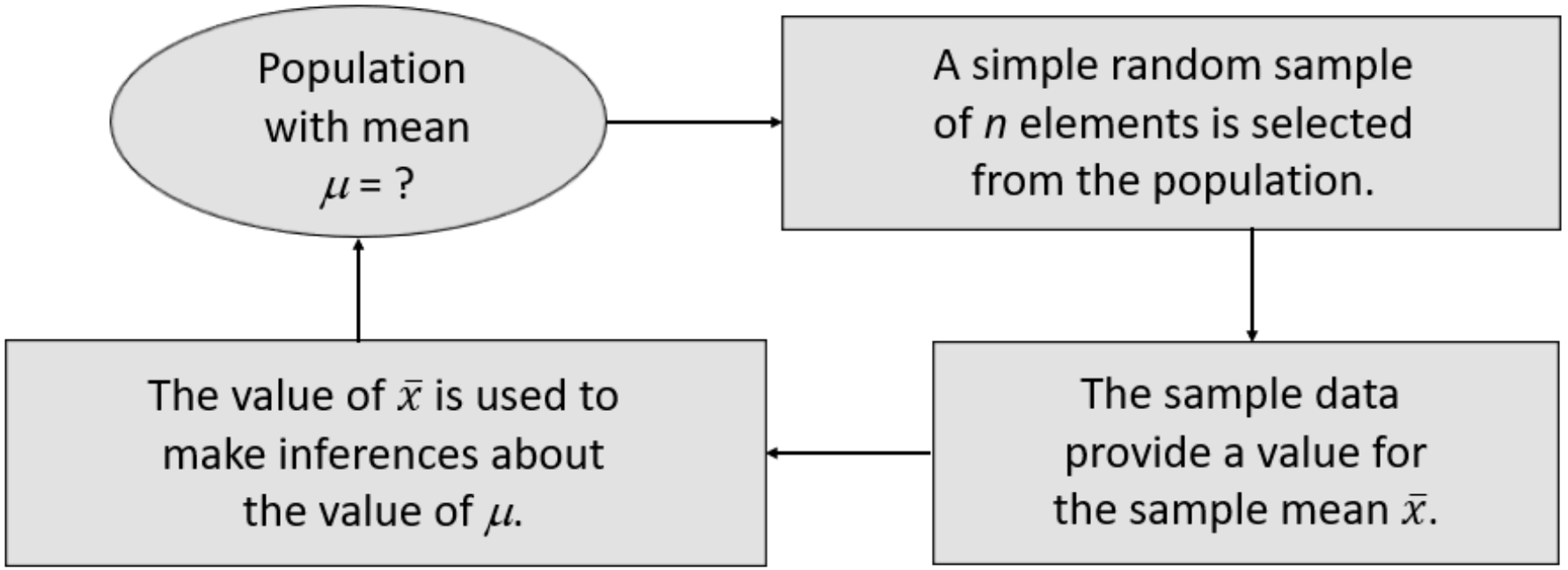
Population Parameter	Parameter value	Point estimator	Point estimate
μ = Population mean annual salary	$ 71,800	$\bar{x}$ = Sample mean annual salary	$ 71,814
σ = Population standard deviation for annual salary	$ 4,000	s = Sample standard deviation for annual salary	$ 3,348
p = Population proportion having completed MTP	.60	$\bar{p}$ = Sample proportion having completed the MTP	.63

Practical Advice

- The target population is the population we want to make inferences about.
- The sampled population is the population from which the sample is actually taken.
- Whenever a sample is used to make inferences about a population, we should make sure that the targeted population and the sampled population are in close agreement.

Sampling Distribution of $\bar{x}$ (1 of 13)

- Process of Statistical Inference



Sampling Distribution of $\bar{x}$ (2 of 13)

- The sampling distribution of $\bar{x}$ is the probability distribution of all possible values of the sample mean $\bar{x}$.
- Expected Value of $\bar{x}$

$$E(\bar{x}) = \mu$$

where: μ = the population mean

When the expected value of the point estimator equals the population parameter, we say the point estimator is unbiased.

Sampling Distribution of $\bar{x}$ (3 of 13)

- We will use the following notations to define the standard deviation of the sampling distribution of $\bar{x}$.

$\sigma_{\bar{x}}$ = the standard deviation of $\bar{x}$

σ = the standard deviation of the population

n = the sample size

N = the population size

Sampling Distribution of $\bar{x}$ (4 of 13)

Standard Deviation of $\bar{x}$

Finite Population

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Infinite Population

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- A finite population is treated as being infinite if $n/N \leq .05$.
- $\sqrt{(N-n)/(N-1)}$ is the finite population correction factor.
  - $\sigma_{\bar{x}}$ is referred to as the standard error of the mean.

Sampling Distribution of $\bar{x}$ (5 of 13)

- When the population has a normal distribution, the sampling distribution of $\bar{x}$ is normally distributed for any sample size.
- In most applications, the sampling distribution of $\bar{x}$ can be approximated by a normal distribution whenever the sample is size 30 or more.
- In cases where the population is highly skewed or outliers are present, samples of size 50 may be needed.

Sampling Distribution of $\bar{x}$ (6 of 13)

- The sampling distribution of $\bar{x}$ can be used to provide probability information about how close the sample mean $\bar{x}$ is to the population mean μ.

Central Limit Theorem

When the population from which we are selecting a random sample does not have a normal distribution, the central limit theorem is helpful in identifying the shape of the sampling distribution of $\bar{x}$.

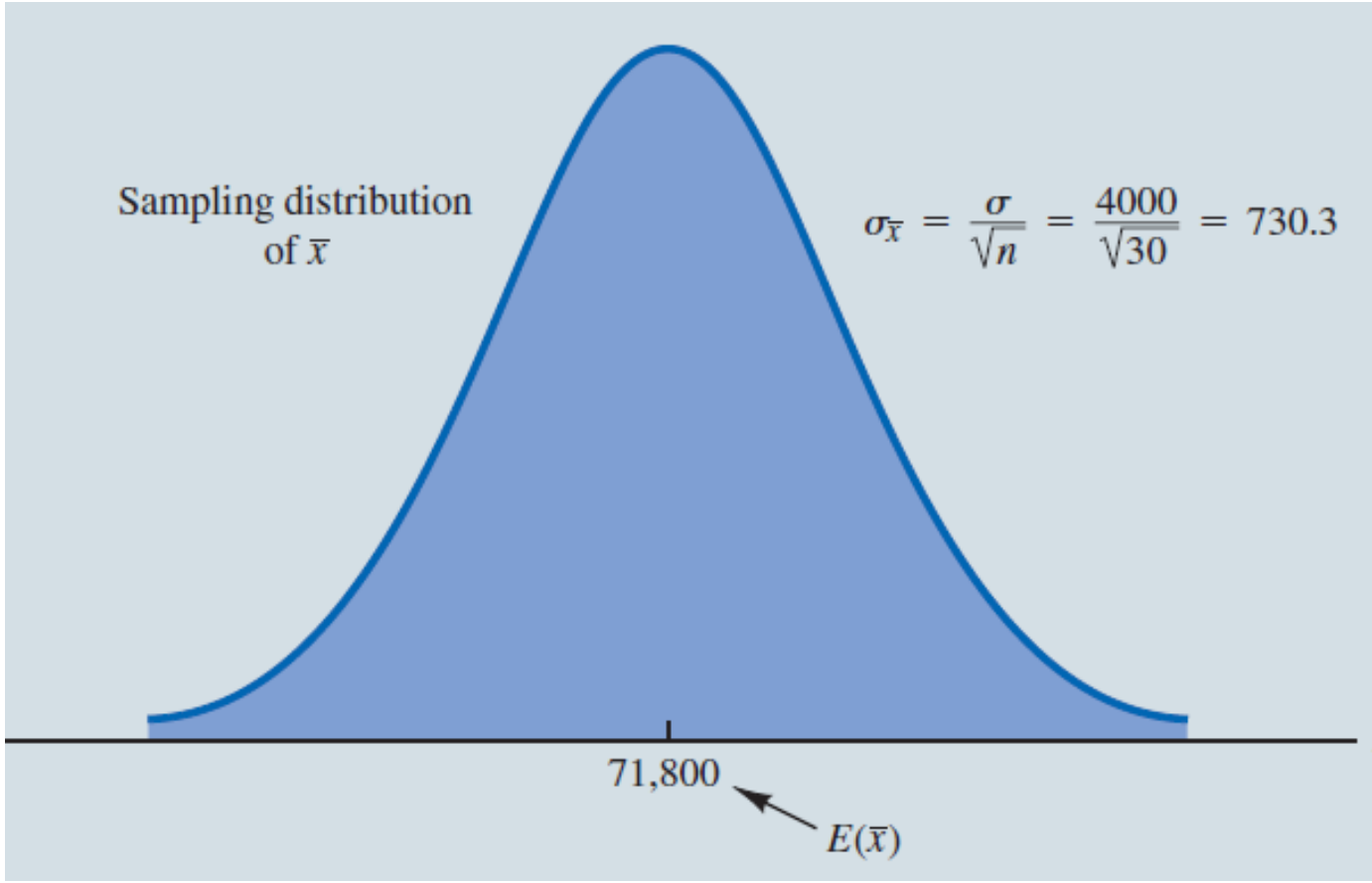
CENTRAL LIMIT THEOREM

In selecting random samples of size n from a population, the sampling distribution of the sample mean $\bar{x}$ can be approximated by a normal distribution as the sample size becomes large.

Sampling Distribution of $\bar{x}$ (7 of 13)

Example: EAI Employee Data

Note: $n/N = 30/2500 = .012$
Because sample is less than 5% of the population size, we ignore the finite population correction factor.



Sampling Distribution of $\bar{x}$ (8 of 13)

Example: EAI Employee Data

Suppose the personnel director believes the sample mean will be an acceptable estimate of population mean if the sample mean is within $500 of the population mean.

What is the probability that the sample mean computed using a simple random sample of 30 employees will be within $500 of population mean?

Sampling Distribution of $\bar{x}$ (9 of 13)

Example: EAI Employee Data

Step 1: Calculate the *z*-value at the upper endpoint of the interval.

$$z = (72{,}300 - 71{,}800)/730.30 = .68$$

Step 2: Find the area under the curve to the left of the upper endpoint.

$$P(z \leq .68) = .7517$$

Sampling Distribution of $\bar{x}$ (10 of 13)

Example: EAI Employee Data

Cumulative Probabilities for
the Standard Normal Distribution

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.	.	.	.	.	.	.	.	.	.	.
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	7.486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
.	.	.	.	.	.	.	.	.	.	.

Sampling Distribution of $\bar{x}$ (11 of 13)

Example: EAI Employee Data

Step 3: Calculate the *z*-value at the lower endpoint of the interval.

$$z = (71{,}300 - 71{,}800)/730.30 = -.68$$

Step 4: Find the area under the curve to the left of the lower endpoint.

$$P(z \leq -.68) = .2483$$

Sampling Distribution of $\bar{x}$ (12 of 13)

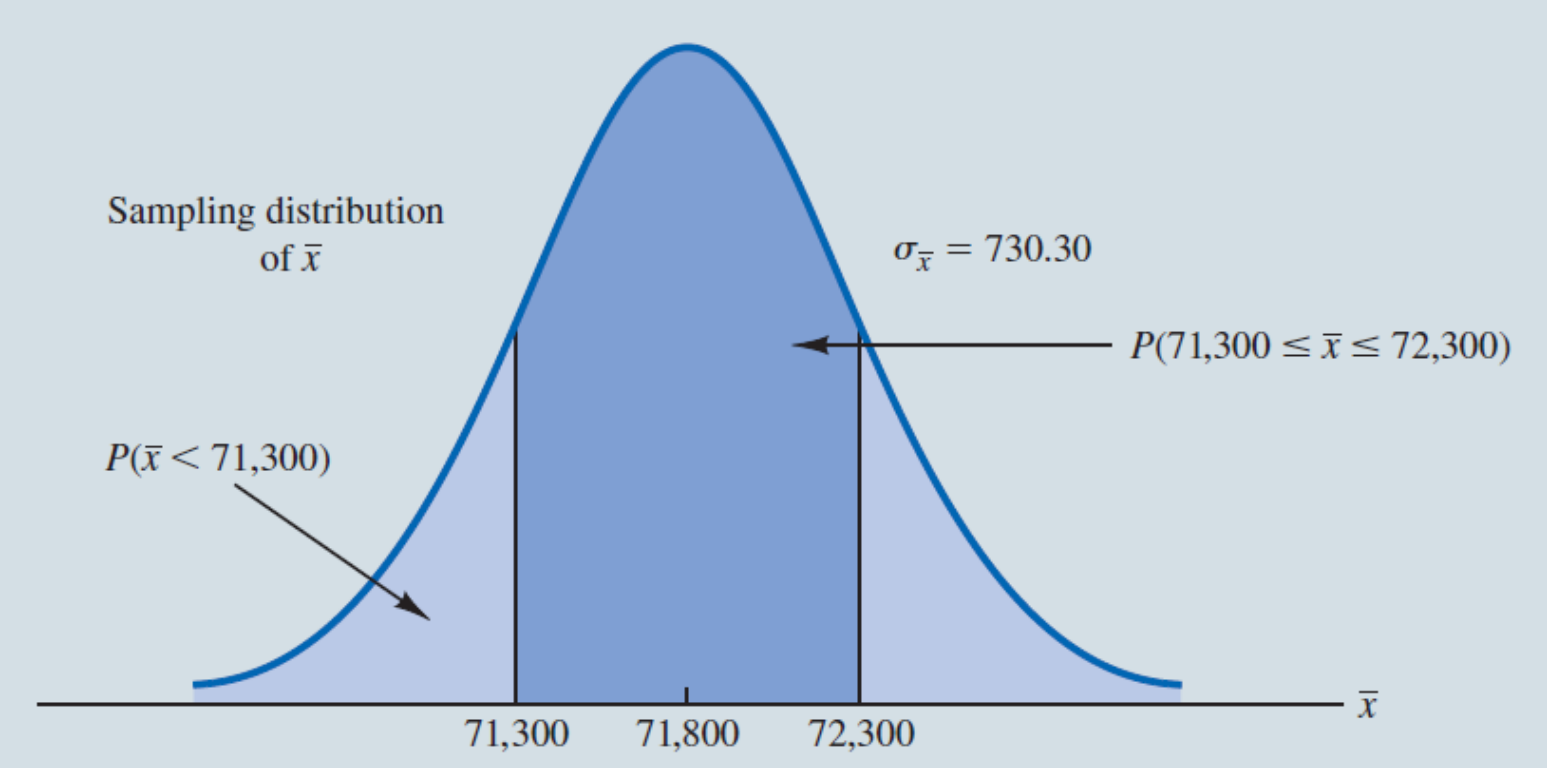
Example: EAI Employee Data

Step 5: Calculate the area under the curve between the lower and upper endpoints of the interval.

$$P(71{,}300 \le \bar{x} \le 72{,}300) = P(z \le .68) - P(z \le -.68)$$
$$= .7517 - .2483$$
$$= .5034$$

Sampling Distribution of $\bar{x}$ (13 of 13)

Example: EAI Employee Data



Use of Excel to Compute Sampling Distribution of $\bar{x}$ (1 of 2)

- Function used: NORM.DIST
- We do not have to make separate computation of *z* value.
- Evaluating NORM.DIST function at each end point of the interval provides cumulative probability at the specified end point of the interval.
- The result obtained using NORM.DIST is more accurate.

Use of Excel to Compute Sampling Distribution of $\bar{x}$ (2 of 2)

Example: EAI Employee Data

- Upper endpoint: Entering the formula =*NORM.DIST(72300,71800,730.30,TRUE)* into a cell provides a cumulative probability of .7532.
- Lower endpoint: Entering the formula =*NORM.DIST(71300,71800,730.30,TRUE)* into a cell provides a cumulative probability of .2468
- So, $P\left(71{,}300 \le \bar{x} \le 72{,}300\right) = .7532 - .2468 = .5064$
- Results above are more accurate than those in the normal table, since the table values are rounded to two decimal places.

Relationship Between the Sample Size and the Sampling Distribution of $\bar{x}$ (1 of 4)

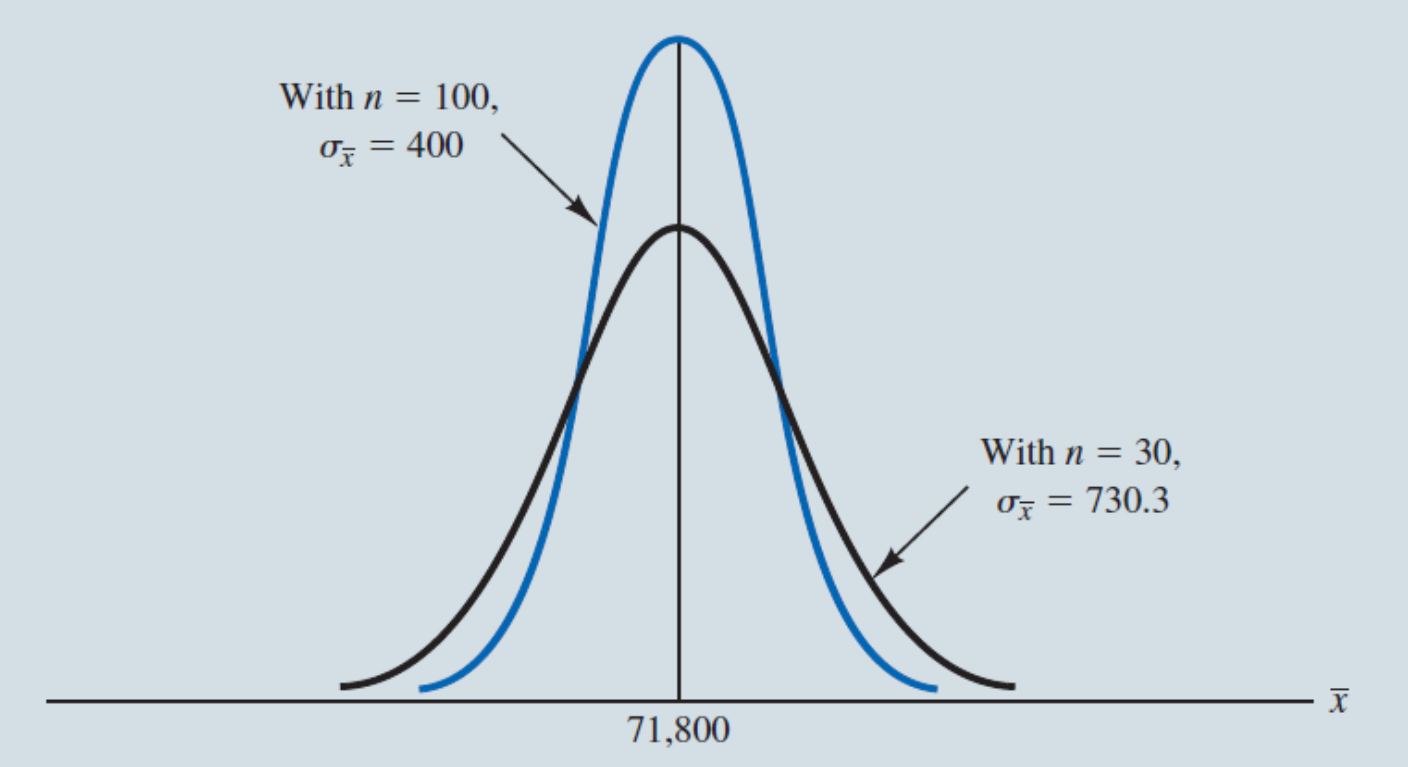
Example: EAI Employee Data

- Suppose we select a simple random sample of 100 employees instead of the 30 originally considered.
- $E(\bar{x}) = m$ regardless of the sample size. In our example, $E(\bar{x})$ remains at 71,800.
- Whenever the sample size is increased, the standard error of the mean $\sigma_{\bar{x}}$ is decreased. With the increase in the sample size to n = 100, the standard error of the mean is decreased from 730.3 to 400.

$$\sigma_{\bar{x}} = \left(\frac{\sigma}{\sqrt{n}}\right) = \left(\frac{4000}{\sqrt{100}}\right) = 400$$

Relationship Between the Sample Size and the Sampling Distribution of $\bar{x}$ (2 of 4)

Example: EAI Employee Data



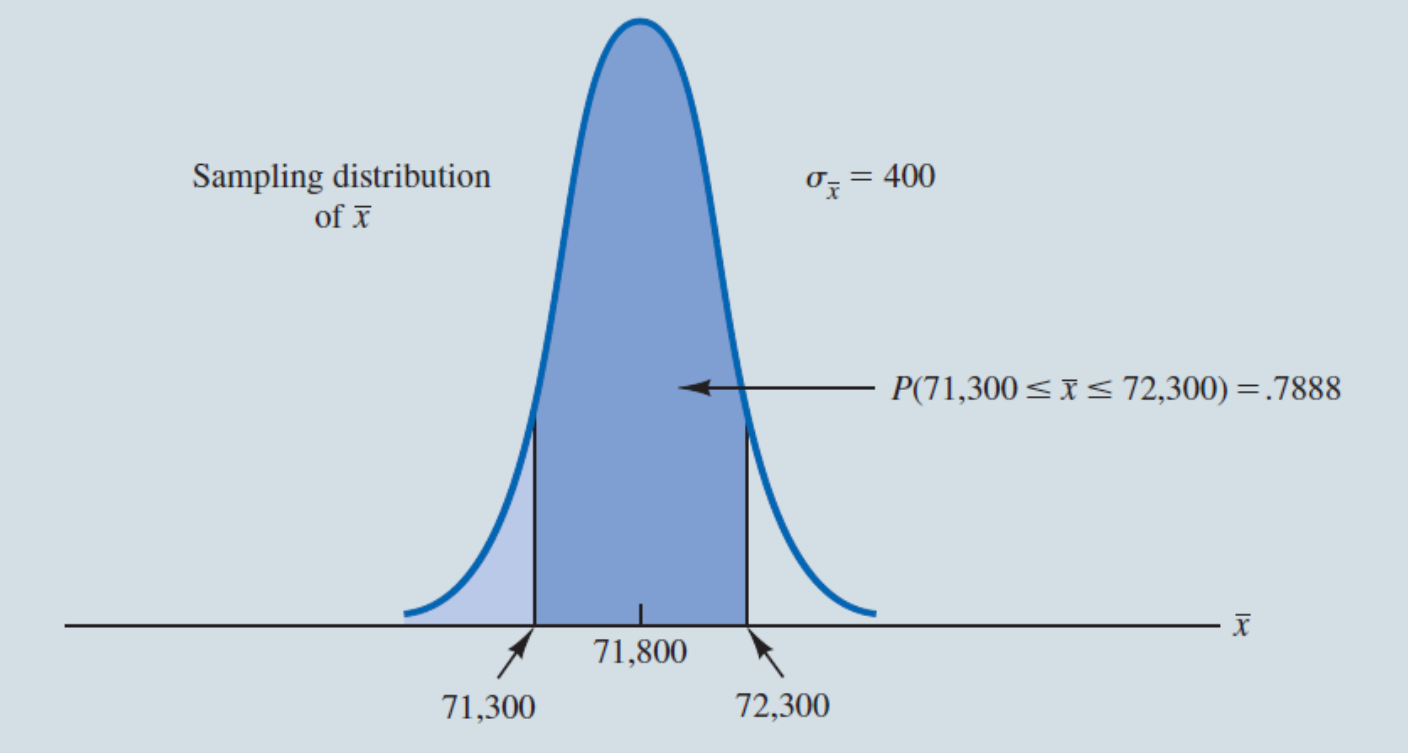
Relationship Between the Sample Size and the Sampling Distribution of $\bar{x}$ (3 of 4)

Example: EAI Employee data

- Recall that when $n = 30$, $P(71300 \le \bar{x} \le 72300) = .5034$ we follow the same steps to solve for $P(71300 \le \bar{x} \le 72300)$ when $n = 100$ as we showed earlier when $n = 30$.
- Now, using Excel with $n = 100$, $P(71300 \le \bar{x} \le 72300) = .7888$
- Because the sampling distribution with $n = 100$ has a smaller standard error, the values of $\bar{x}$ have less variability and tend to be closer to the population mean than the values of $\bar{x}$ with $n = 30$.

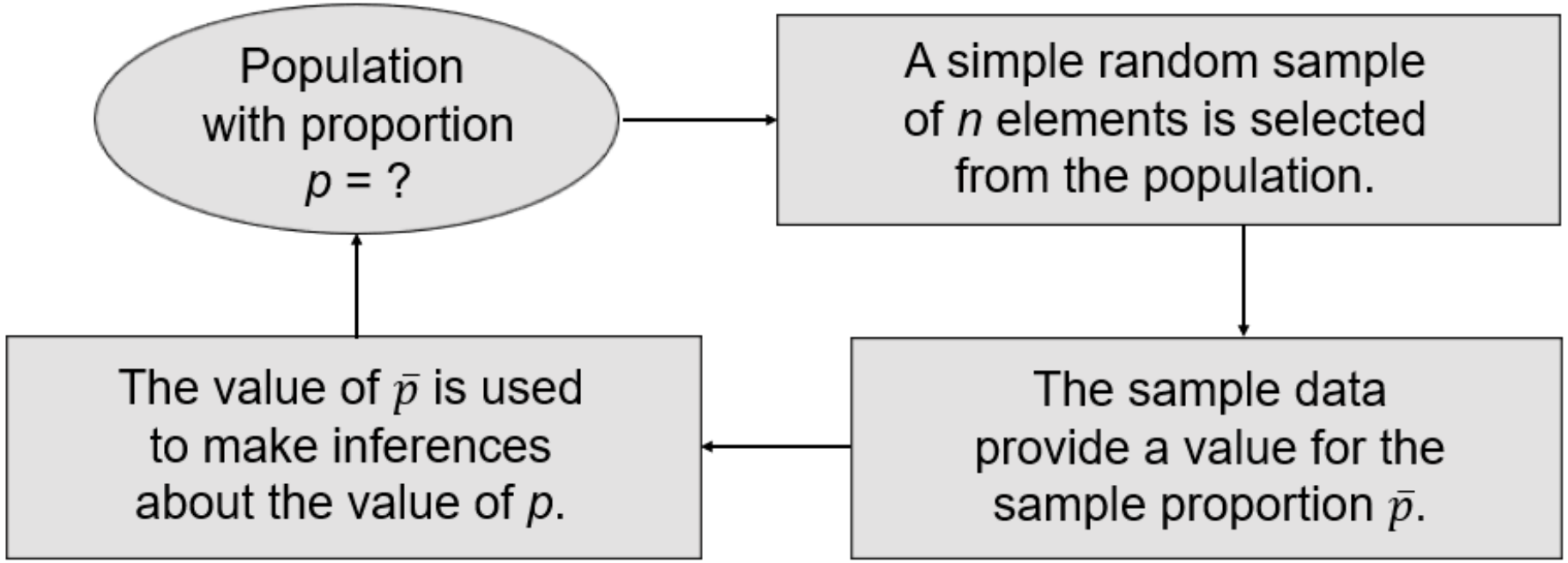
Relationship Between the Sample Size and the Sampling Distribution of $\bar{x}$ (4 of 4)

Example: EAI Employee Data



Sampling Distribution of $\bar{p}$ (1 of 9)

- Making inferences about a population proportion



Sampling Distribution of $\bar{p}$ (2 of 9)

- The sampling distribution of $\bar{p}$ is the probability distribution of all possible values of the sample proportion $\bar{p}$.
- Expected Value of $\bar{p}$

$$E(\bar{p}) = p$$

where: p = the population proportion

Sampling Distribution of $\bar{p}$ (3 of 9)

Standard Deviation of $\bar{p}$

Finite Population

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}$$

Infinite Population

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

- $\sigma_{\bar{p}}$ is referred to as the standard error of the proportion.
- $\sqrt{(N-n)/(N-1)}$ is the finite population correction factor.

Sampling Distribution of $\bar{p}$ (4 of 9)

- The sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever the sample size is large enough to satisfy the two conditions:

$$np \geq 5 \text{ and } n(1-p) \geq 5$$

- When these conditions are satisfied, the probability distribution of x in the sample proportion, $\bar{p} = x/n$, can be approximated by a normal distribution (because n is a constant, the sampling distribution of $\bar{p}$ can also be approximated by a normal distribution).

Sampling Distribution of $\bar{p}$ (5 of 9)

Example: EAI Employee Data

For the EAI study we know that the population proportion of employees who participated in the management training program is $p = .6$.

What is the probability that a simple random sample of 30 employees will provide an estimate of the population proportion of employees attending management program that is within plus or minus .05 of the actual population proportion?

Sampling Distribution of $\bar{p}$ (6 of 9)

Example: EAI Employee Data

For our example, with $n = 30$ and $p = .6,$ the normal distribution is an acceptable approximation because:

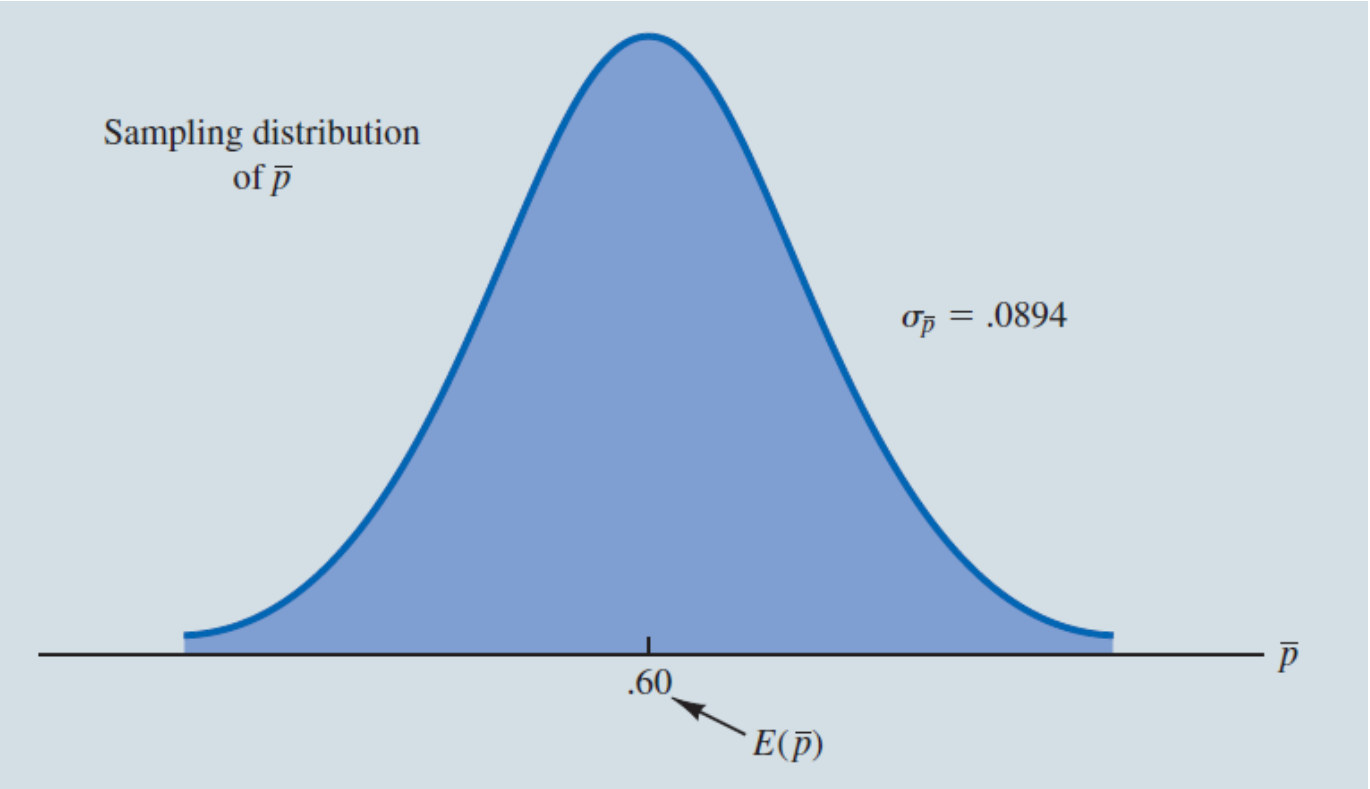
$$np = 30(.6) = 18 \geq 5$$

and

$$n(1-p) = 30(.4) = 12 \geq 5$$

Sampling Distribution of $\bar{p}$ (7 of 9)

- **Example:** EAI Employee Data



$$\sigma_{\bar{p}} = \sqrt{\frac{.6(1-.6)}{30}}$$

$$= .0894$$

Using Excel for the Sampling Distribution of $\bar{p}$

Example: EAI Employee Data

Using the mean of .60 and a standard error of $\sigma_{\bar{p}} = .0894$, we can use Excel's NORM.DIST function to make the calculation.

- Entering the formula = NORM.DIST(.65,.60,.0894,TRUE) in a cell provides the cumulative probability corresponding to $\bar{p}$ =.65. The value is **.7120.**

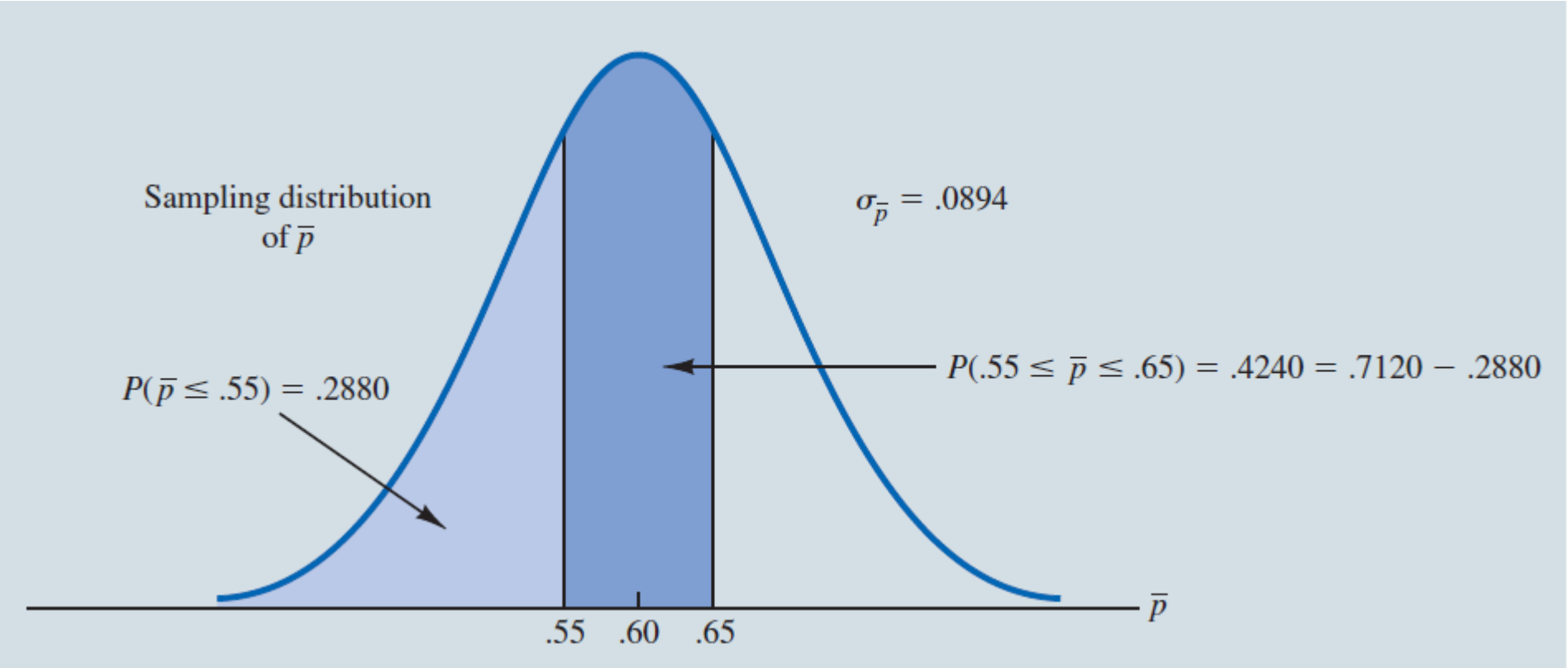
Sampling Distribution of $\bar{p}$ (8 of 9)

Example: EAI Employee Data

- Entering the formula = NORM.DIST(.55,.60,.0894,TRUE) in a cell provides the cumulative probability corresponding to $\bar{p} = .55$. The value is **.2880.**
- So, the probability of $\bar{p}$ being in the interval from .55 to .65 is given by **.7120** – **.2880** = .4240.

Sampling Distribution of $\bar{p}$ (9 of 9)

- **Example:** EAI Employee Data



Other Sampling Methods

- Stratified Random Sampling
- Cluster Sampling
- Systematic Sampling
- Convenience Sampling
- Judgment Sampling

Stratified Random Sampling (1 of 2)

- The population is first divided into groups of elements called strata.
- Each element in the population belongs to one and only one stratum.
- Best results are obtained when the elements within each stratum are as much alike as possible (i.e. a homogeneous group).

Stratified Random Sampling (2 of 2)

- A simple random sample is taken from each stratum.
- Formulas are available for combining the stratum sample results into one population parameter estimate.
- Advantage: If strata are homogeneous, this method is as “precise” as simple random sampling but with a smaller total sample size.

Example: The basis for forming the strata might be department, location, age, industry type, and so on.

Cluster Sampling (1 of 2)

- The population is first divided into separate groups of elements called clusters.
- Ideally, each cluster is a representative small-scale version of the population (i.e. heterogeneous group).
- A simple random sample of the clusters is then taken.
- All elements within each sampled (chosen) cluster form the sample.

Cluster Sampling (2 of 2)

- Advantage: The close proximity of elements can be cost effective (i.e., many sample observations can be obtained in a short time).
- Disadvantage: This method generally requires a larger total sample size than simple or stratified random sampling.

Example: A primary application is area sampling, where clusters are city blocks or other well-defined areas.

Systematic Sampling (1 of 2)

- If a sample size of *n* is desired from a population containing *N* elements, we might sample one element for every *N*/*n* elements in the population.
- We randomly select one of the first *N*/*n* elements from the population list.
- We then select every *N*/*n*th element that follows in the population list.

Systematic Sampling (2 of 2)

- This method has the properties of a simple random sample, especially if the list of the population elements is a random ordering.
- Advantage: The sample usually will be easier to identify than it would be if simple random sampling were used.

Example: Selecting every 100^{th} listing in a telephone book after the first randomly selected listing

Convenience Sampling (1 of 2)

- It is a nonprobability sampling technique. Items are included in the sample without known probabilities of being selected.
- The sample is identified primarily by convenience.

Example: A professor conducting research might use student volunteers to constitute a sample.

Convenience Sampling (2 of 2)

- Advantage: Sample selection and data collection are relatively easy.
- Disadvantage: It is impossible to determine how representative of the population the sample is.

Judgment Sampling (1 of 2)

- The person most knowledgeable on the subject of the study selects elements of the population that he or she feels are most representative of the population.
- It is a nonprobability sampling technique.

Example: A reporter might sample three or four senators, judging them as reflecting the general opinion of the senate.

Judgment Sampling (2 of 2)

- Advantage: It is a relatively easy way of selecting a sample.
- Disadvantage: The quality of the sample results depends on the judgment of the person selecting the sample.

Recommendation

- It is recommended that probability sampling methods (simple random, stratified, cluster, or systematic) be used.
- For these methods, formulas are available for evaluating the “goodness” of the sample results in terms of the closeness of the results to the population parameters being estimated.
- An evaluation of the goodness cannot be made with nonprobability (convenience or judgment) sampling methods.

Errors in Sampling (1 of 2)

- The difference between the value of sample statistic and the corresponding value of the population parameters is called the sampling error.
- Deviations of the sample from the population that occur for reasons other than random sampling are referred to as nonsampling errors.
- Nonsampling errors can occur in a sample or a census.

Errors in Sampling (2 of 2)

- Reasons for Nonsampling Errors
  - Coverage error
  - Non-response error
    - Interviewer error
    - Processing error
  - Measurement error

Steps to Minimize Nonsampling Errors

- Carefully define the target population and design the data collection procedure.
- Carefully design the data collection process and train the data collectors.
- Pretest the data collection procedure.
- Use stratified random sampling when population-level information about an important qualitative characteristic is available.
- Use systematic sampling when population-level information about an important quantitative characteristic is available.