Sampling and Sampling Distributions

Introduction to Sampling

Sampling is the process of collecting data to answer a research question about a population.

Two important terms used in sampling are:

- 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.

Because the sample contains only a portion of the population, results provide only estimates of the values of the population characteristics.

With proper sampling methods, the sample results can provide “good” estimates of the population characteristics.

The Electronics Associates Sampling Problem

The numerical characteristics of a population are called **parameters**.

Using descriptive statistics, we compute the **population mean**, $\mu = \$51{,}800$, and the **population standard deviation**, $\sigma = \$4{,}000$, for the 2,500 manager salaries included in EAI the data set (DATAfile: *EAI.*)

The data for the training program status also show that 1,500 of the 2,500 managers completed the training program. Thus, we have the **population proportion**, $p = 1500/2500 = 0.6$

If the information on all the EAI managers is not readily available in the company's database, estimates of the population parameters can be obtained using a sample of managers drawn from the population.

The time and cost of developing a profile on an adequate sample of 30 managers would be substantially less than collecting salary data on the entire population of 2,500 managers.

Let us explore the possibility of identifying a sample of 30 managers for the EAI study.

Sampling from a Finite Population

A **simple random sample** of size *n* is selected from a finite population of size *N* so that each possible sample of size *n* has the same probability of being selected.

Replacing each sampled element before selecting subsequent elements is called **sampling with replacement**. An element can appear in the sample more than once.

Sampling without replacement is the procedure used most often.

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

To select a simple random sample from the finite population of EAI managers, follow this procedure:

Step 1: Assign a random number to each of the 2,500 managers.

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

Step 2: Select the 30 applicants corresponding to the 30 smallest random numbers.

Sampling from an Infinite Population

Sometimes we want to select a sample but find that 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.

When a frame is not available, such as in the case of an infinite population, we cannot use the random number selection procedure previously outlined.

To select a sample from an infinite population, we must still follow a random procedure in order to make valid statistical inferences about the population.

A **random sample** of size *n* is selected from an infinite population so that

1. each element selected comes from the same population
2. each element is selected independently

An ongoing process often generates an infinite population because there is no upper limit on the number of units it can create.

Examples of ongoing processes include parts manufactured on a production line and transactions occurring at a bank

A Simple Random Sample for EAI

A **sample statistic** that estimates the value of a population parameter is called a **point estimator**.

In **point estimation**, the value of a sample statistic is the **point estimate** of a population.

The annual salaries and training program status for a simple random sample of 30 EAI Managers are reported in the table to the right.

Note: different random numbers would have identified a different sample.

Manager Number	Annual Salary ($)	Management Training Program	Manager Number	Annual Salary ($)	Management Training Program
1	69,094.30	Yes	16	71,766.00	Yes
2	73,263.90	Yes	17	72,541.30	No
3	69,643.50	Yes	18	64,980.00	Yes
4	69,894.90	Yes	19	71,932.60	Yes
5	67,621.60	No	20	72,973.00	Yes
6	75,924.00	Yes	21	65,120.90	Yes
7	69,092.30	Yes	22	71,753.00	Yes
8	71,404.40	Yes	23	74,391.80	No
9	70,957.70	Yes	24	70,164.20	No
10	75,109.70	Yes	25	72,973.60	No
11	65,922.60	Yes	26	70,241.30	No
12	77,268.40	No	27	72,793.90	No
13	75,688.80	Yes	28	70,979.40	Yes
14	71,564.70	No	29	75,860.90	Yes
15	76,188.20	No	30	77,309.10	No

Point Estimation

Let us compute sample statistics for a simple random sample of EAI managers.

$\bar{x}$ is the point estimator of the population mean, $\mu = \$71{,}800$.

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

s is the point estimator of the population standard deviation, $\sigma = \$4{,}000$

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{325{,}009{,}260}{29}} = \$3{,}348$$

$\bar{p}$ is the point estimator of the population proportion, $p = 0.6$.

$$\bar{p} = \frac{x}{n} = \frac{19}{30} = 0.633$$

Practical Advice

The subject matter of most of the rest of the book is concerned with statistical inference.

In statistical inference, the subject of most of the rest of the book, we use a sample statistic as a point estimator of a population parameter.

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.

In the EAI example, we drew a simple random sample from the sampled population of 2,500 managers to make point estimates of parameters of an identical target population.

However, it is not always easy to obtain a close correspondence between the sampled and target populations as we did in the EAI example.

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, or we cannot guarantee that the sample is representative of the target population.

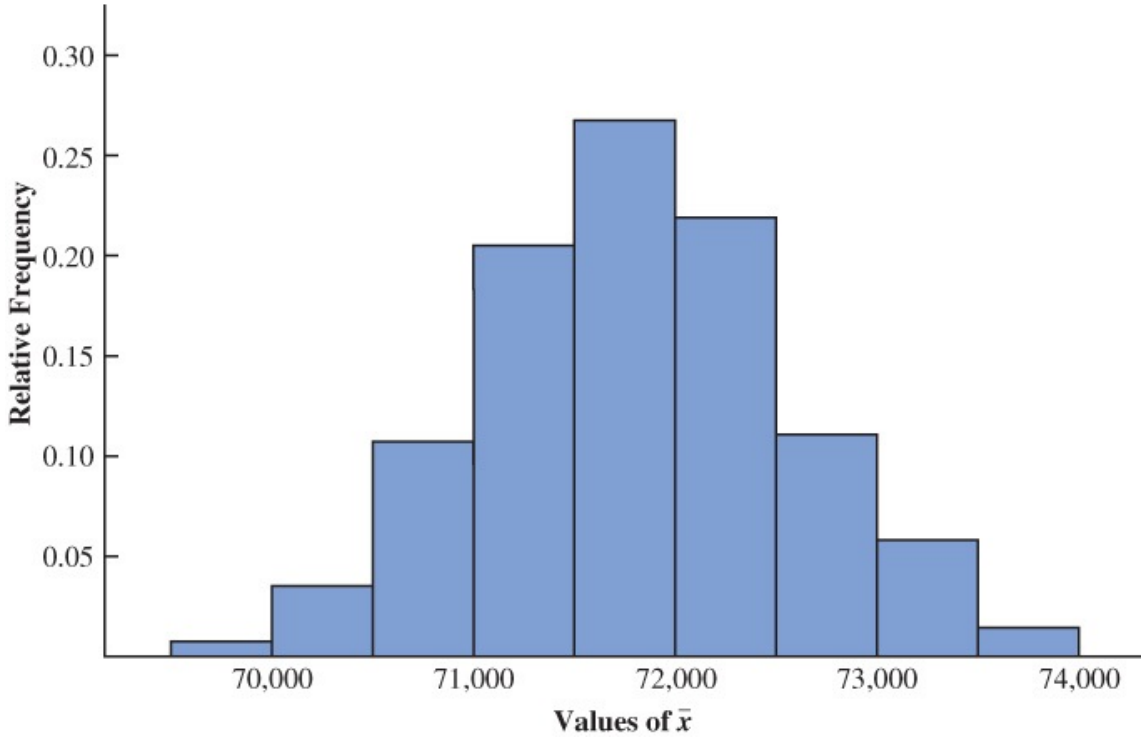
Introduction to Sampling Distributions

Suppose we repeat the process of selecting a simple random sample of 30 EAI managers over and over again, each time computing the values of $\bar{x}$ and $\bar{p}$.

Because each random sample likely contains a unique random selection of 30 managers, each point estimate of $\bar{x}$ and $\bar{p}$ will be different.

As a random variable, $\bar{x}$ will have an expected value, a standard deviation, and a probability distribution, called the **sampling distribution** of $\bar{x}$.

The relative frequency histogram of 500 values of $\bar{x}$, as shown in the figure on the bottom right, approximates the sampling distribution of $\bar{x}$.



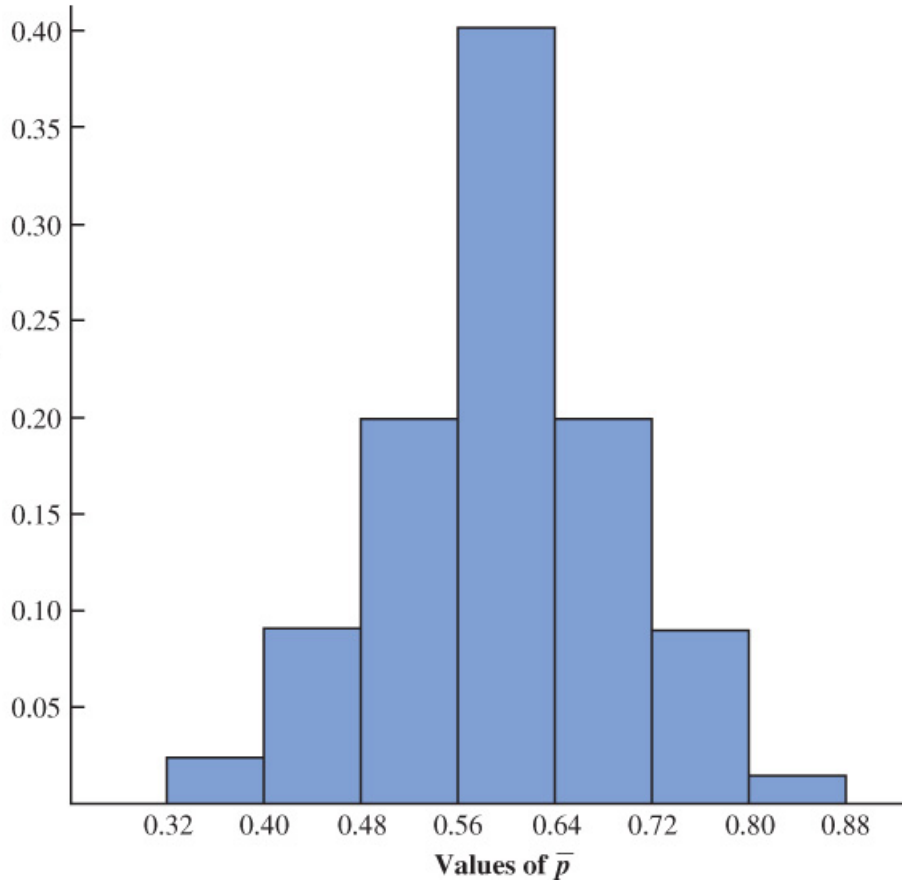
Approximation of a Sampling Distribution

$\bar{p}$ is also a random variable. If a value of $\bar{p}$ was computed for every possible sample of size 30, the resulting probability distribution would be the sampling distribution of $\bar{p}$.

The relative frequency histogram of the 500 $\bar{p}$ values in the figure to the right provides a general idea of the appearance of the sampling distribution of $\bar{p}$.

From the approximations of the sampling distributions of $\bar{x}$ and $\bar{p}$, we observe the bell-shaped appearance and the concentration of most values of the sample statistic around their respective population parameter.

In practice, to build a sampling distribution we select only one simple random sample from the population.



Sampling Distribution of $\bar{x}$

The **sampling distribution of $\bar{x}$**, the probability distribution of all possible values of the sample mean $\bar{x}$, is defined by

$$E(\bar{x}) = \mu$$ the epected value of $\bar{x}$

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

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$$ the **standard error** of $\bar{x}$

Where n is the sample size, N the population size, and $\sqrt{(N-n)/(N-1)}$ is the finite population correction factor.

When $n/N < 0.05$, we can neglect the population correction factor, and $\sigma_{\bar{x}}$ simplifies as

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

Form of the Sampling Distribution of $\bar{x}$

When the population has a normal distribution, the sampling distribution of $\bar{x}$ is normally distributed for any sample size.

When the population does not have a normal distribution, the **central limit theorem** helps identify the sampling distribution of $\bar{x}$.

Central Limit Theorem

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

In most applications, the sampling distribution of $\bar{x}$ can be approximated by a normal distribution whenever the sample is size 30 or more.

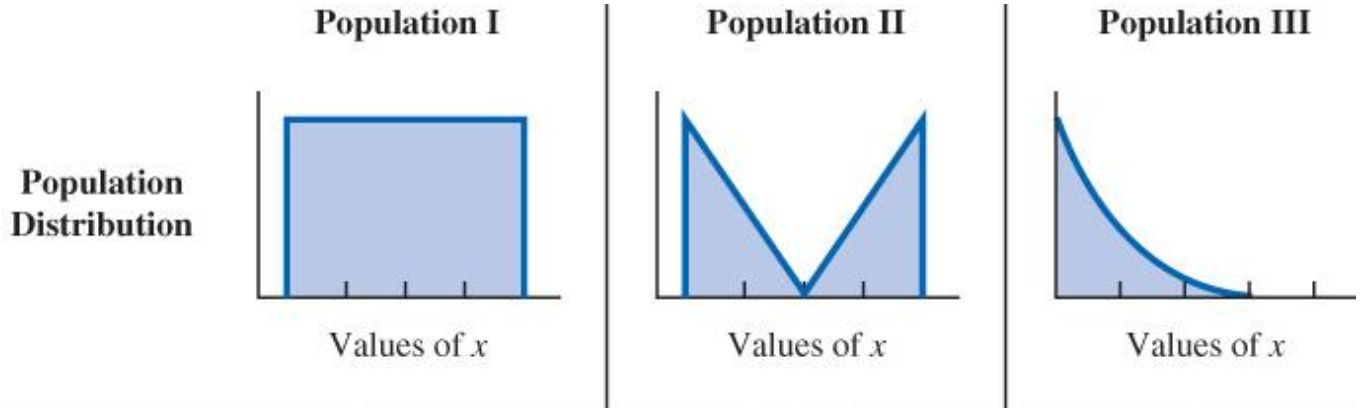
In cases where the population is highly skewed or outliers are present, samples of size 50 may be needed.

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.

Sampling from Different Population Distributions

Let us consider three different population distributions.

- Population I: uniform distribution
- Population II: rabbit-eared distribution
- Population III: exponential distribution (right-skewed)



In the next slide, we will examine how the central limit theorem works by increasing the sample size n of the sampling distribution of $\bar{x}$ for each of the three population distributions.

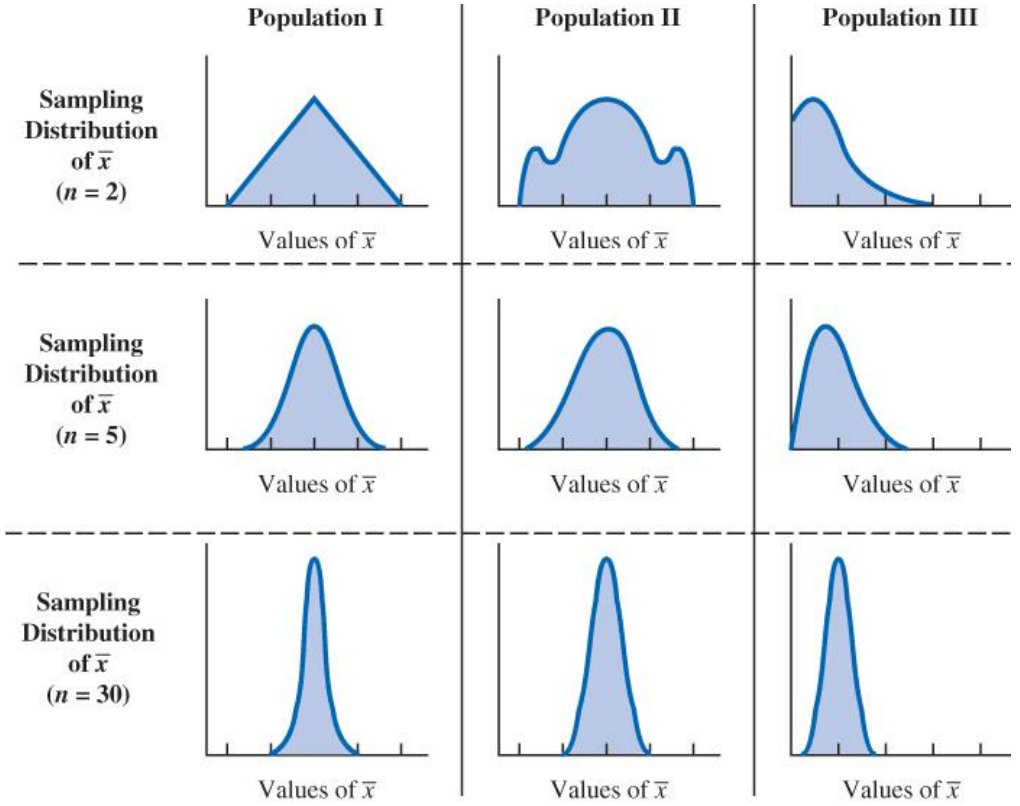
Illustration of the Central Limit Theorem

The three panels shows what happens to the sampling distribution of $\bar{x}$ for the three population distributions introduced in the previous slide when the sample size n is increased.

$n = 2$ the shape of each sampling distribution is different from the shape of the corresponding population distribution.

$n = 5$ the shapes of sampling distributions I and II begin to look similar, but III is still skewed to the right.

$n = 30$ the shape of each sampling distribution has already become approximately normal.



Sampling Distribution of $\bar{x}$ for the EAI Problem

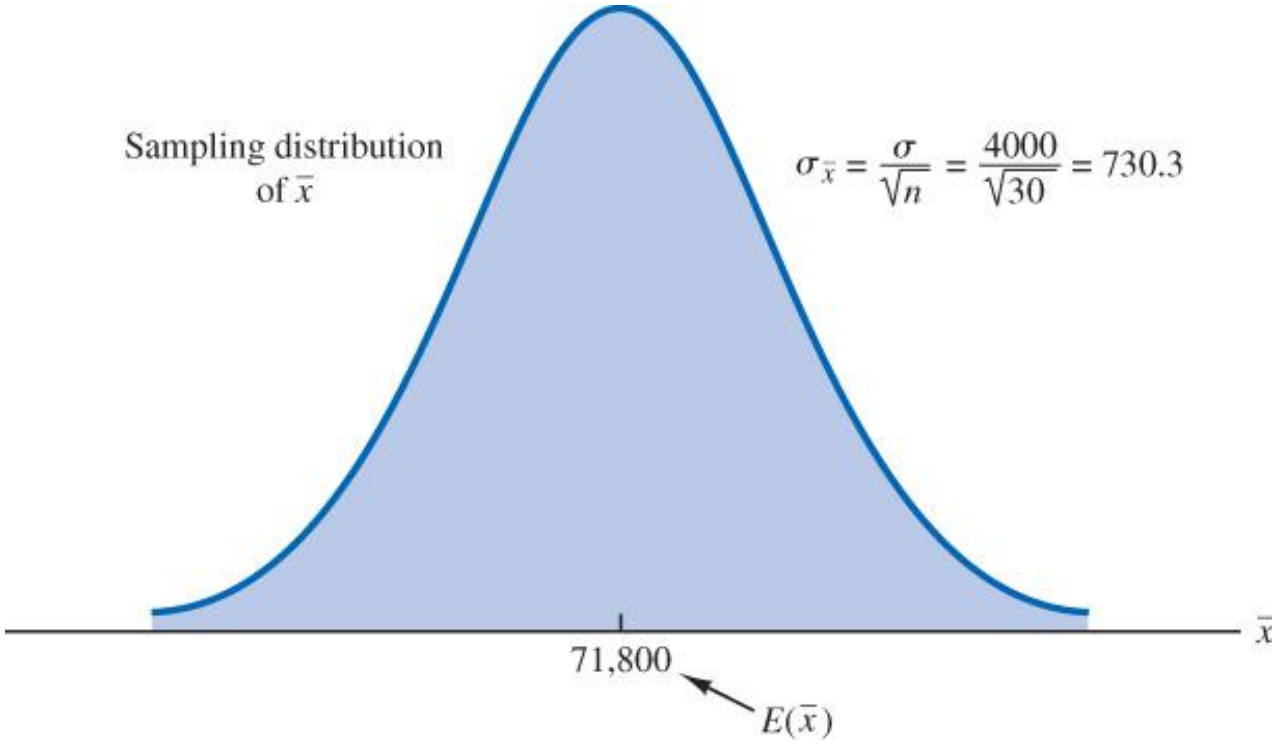
The sampling distribution of $\bar{x}$ for the EAI problem is characterized by expected value and standard error of the mean equal to

$$E(\bar{x}) = \mu = \$71{,}800$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{30}} = \$730.3$$

Because $n/N = 30/2500 = 0.012 < 0.05$, we do not need the finite population correction factor.

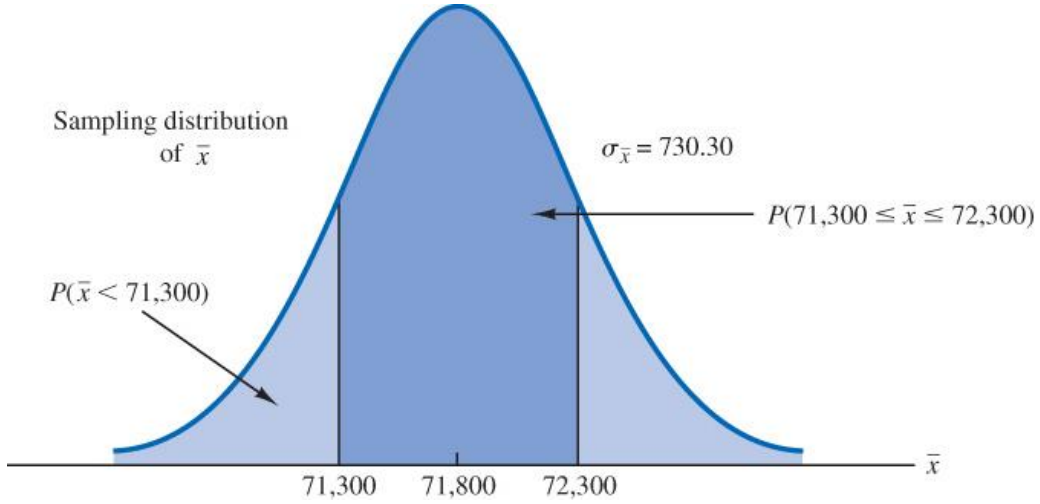
With a sample of 30 managers, the central limit theorem enables us to conclude that the sampling distribution follows the normal distribution.



Application of the Sampling Distribution of $\bar{x}$

EAI wants to know the probability that the sample mean computed using a simple random sample of 30 EAI managers will be within \$500 of the population mean.

Because μ = \$71,800, we want to know the probability that $\bar{x}$ is between \$71,300 and \$72,300, represented in the figure to the right.



At $\bar{x} = \$72{,}300$, we have:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{72300 - 71800}{730.3} = 0.68$$

And at $\bar{x} = \$71{,}300$, we have

$$z = -0.68$$

Therefore, the desired probability is

$$P(71300 \le \bar{x} \le 72300) = P(-0.68 \le z \le 0.68) = P(z \le 0.68) - P(z \le -0.68)$$
$$= 0.7517 - 0.2483 = 0.5034 \approx 50\%$$

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

Suppose the personnel director at EAI selects a larger simple random sample of 100 managers.

$E(\bar{x}) = \mu$, so it is still \$71,800, but the standard error of the mean is now narrower, with

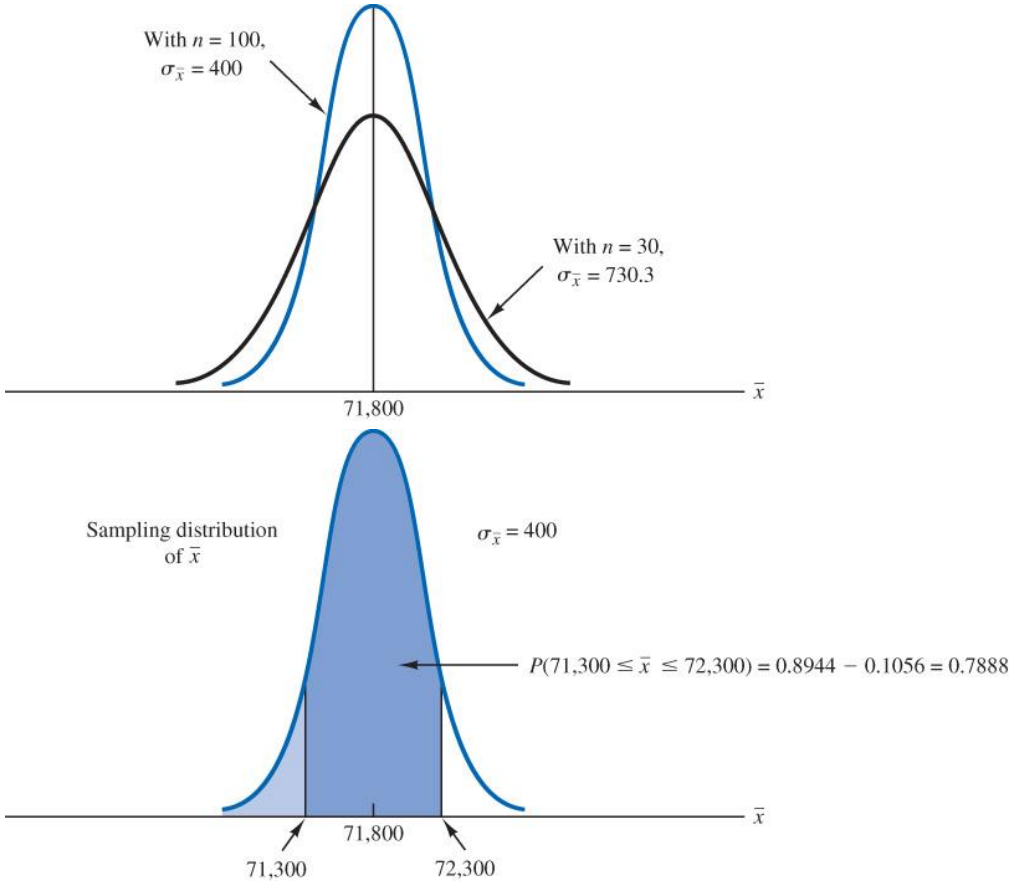
$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 4000/\sqrt{100} = \$400$$

At $\bar{x} = \$72{,}300$, we have:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{72300 - 71800}{400} = 1.25$$

The probability that $\bar{x}$ is within \$500 of μ becomes

$$P(71300 \le \bar{x} \le 72300) = P(-1.25 \le z \le 1.25)$$
$$= P(z \le 1.25) - P(z \le -1.25) = 0.9472 - 0.0528$$
$$= 0.8944 \approx 89\%$$



Sampling Distribution of $\bar{p}$

The **sampling distribution of $\bar{p}$**, the probability distribution of all possible values of the sample proportion $\bar{p}$, is defined as

$$\boldsymbol{E(\bar{p}) = p}$$ the expected value of $\bar{p}$

$$\boldsymbol{\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}}$$ the standard error of $\bar{p}$

Where p is the population proportion and $\sqrt{(N-n)/(N-1)}$ the finite population correction factor.

When $n/N < 0.05$, we can neglect the population correction factor, and $\sigma_{\bar{p}}$ simplifies as

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

Form of the Sampling Distribution of $\bar{p}$

In Chapter 6, we showed that a binomial distribution can be approximated by a normal distribution whenever the sample size is large enough to satisfy the following two conditions:

$$np \geq 5$$

$$n(1-p) \geq 5$$

Because n is a constant, if the probability distribution of x in the sample proportion $\bar{p} = x/n$ can be approximated by a normal distribution, it follows that the sampling distribution of $\bar{p}$ can also be approximated by a normal distribution.

Thus, we formally state that

The sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever $np \geq 5$ and $n(1-p) \geq 5$

Practical Value of the Sampling Distribution of $\bar{p}$

What is the probability of obtaining a value of $\bar{p}$ that is within 0.05 of the population proportion, p, of EAI managers who participated in the training program?

The expected value and standard error of the sampling distribution of $\bar{p}$ can be written as

$$E(\bar{p}) = p = 0.6$$

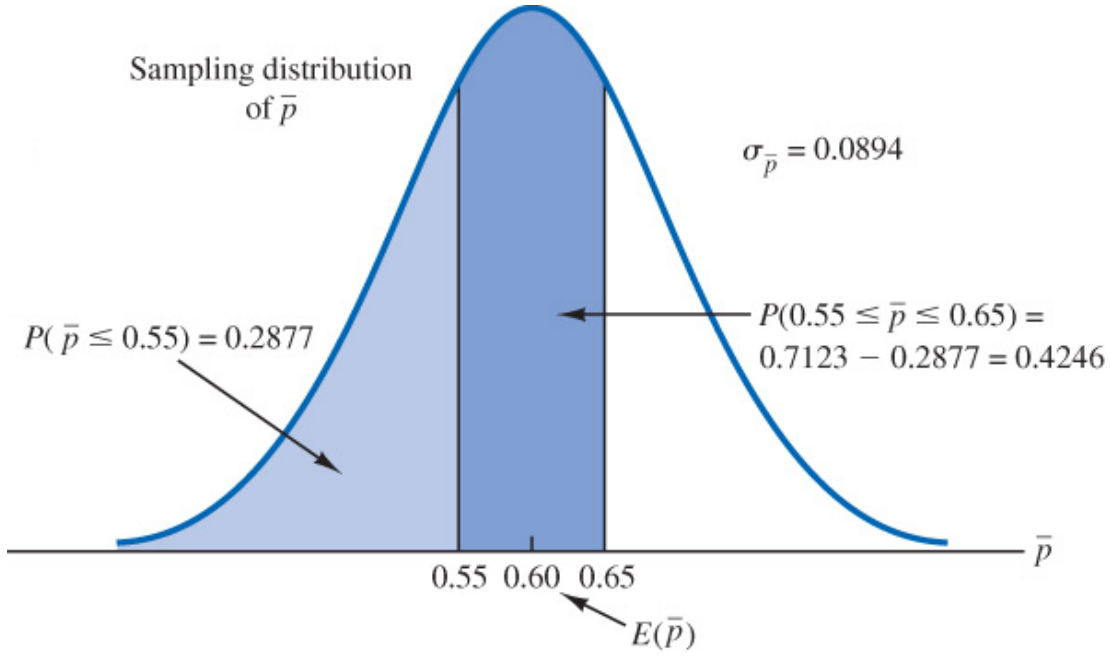
$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.6(1-0.6)}{30}} = 0.0894$$

At $\bar{p} = 0.65$, we have:

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0.65 - 0.6}{0.0894} = 0.56$$

Thus, the probability that $\bar{p}$ is within 0.05 of p is

$$P(0.55 \le \bar{p} \le 0.65) = P(-0.56 \le z \le 0.56)$$
$$= 0.7123 - 0.2877 = 0.4246 \approx 42\%$$



Relationship Between the Sample Size and the Sampling Distribution of $\bar{p}$

Increasing the sample size will also narrow the sampling distribution of $\bar{p}$.

If we increased the sample size to *n* = 100, the standard error of the proportion becomes

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.6(1-0.6)}{100}} = 0.049$$

We do not have to consider the finite population correction factor for the standard error of the proportion because $n/N = 100/2{,}500 = 0.04 < 0.05$. So, at $\bar{p} = 0.65$, we have

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0.65 - 0.6}{0.049} = 1.02$$

The probability that $\bar{p}$ is within 0.05 of *p* becomes

$$P(0.55 \le \bar{p} \le 0.65) = P(-1.02 \le z \le 1.02) = 0.8461 - 0.1539 = 0.6922 \approx 69\%$$

Properties of Point Estimators

Before using a sample statistic as a point estimator, statisticians check to see whether the sample statistic demonstrates certain properties associated with good point estimators.

In this section, we discuss three properties of good point estimators:

- Unbiased
- Efficiency
- Consistency

To represent a generic point estimator and population parameter such as the population mean, population standard deviation, or population proportion, we use the following notation:

- θ = the population parameter of interest
- $\hat{\theta}$ = the sample statistic or point estimator of θ

Unbiased Point Estimator

The sample statistic $\hat{\theta}$ is an **unbiased** estimator of the population parameter θ if $E(\hat{\theta}) = \theta$.

Where $E(\hat{\theta})$ is the expected value of the sample statistic $\hat{\theta}$.

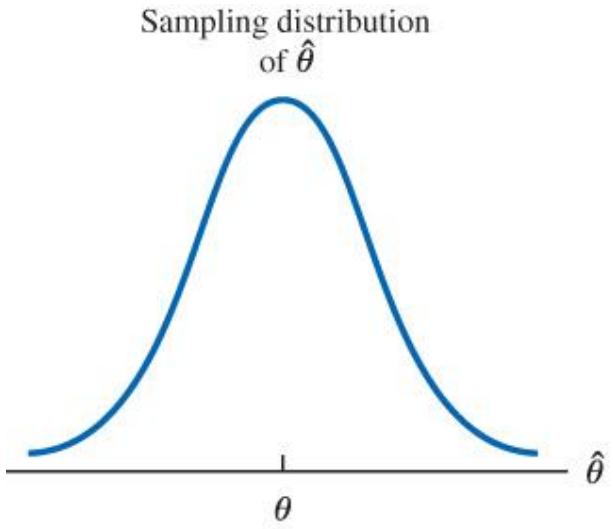
We have previously stated that

$$E(\bar{x}) = \mu \quad \text{and} \quad E(\bar{p}) = p$$

Thus, both $\bar{x}$ and $\bar{p}$ are unbiased estimators of the population parameters μ and p.

It can also be shown that the sample variance is an unbiased estimator of the population variance

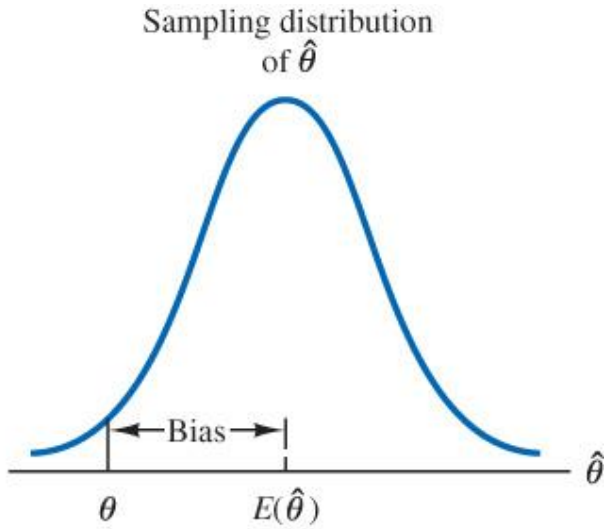
$$E(s^2) = \sigma^2$$



Parameter θ is located at the mean of the sampling distribution;

$E(\hat{\theta}) = \theta$

Panel A: Unbiased Estimator



Parameter θ is not located at the mean of the sampling distribution;

$E(\hat{\theta}) \neq \theta$

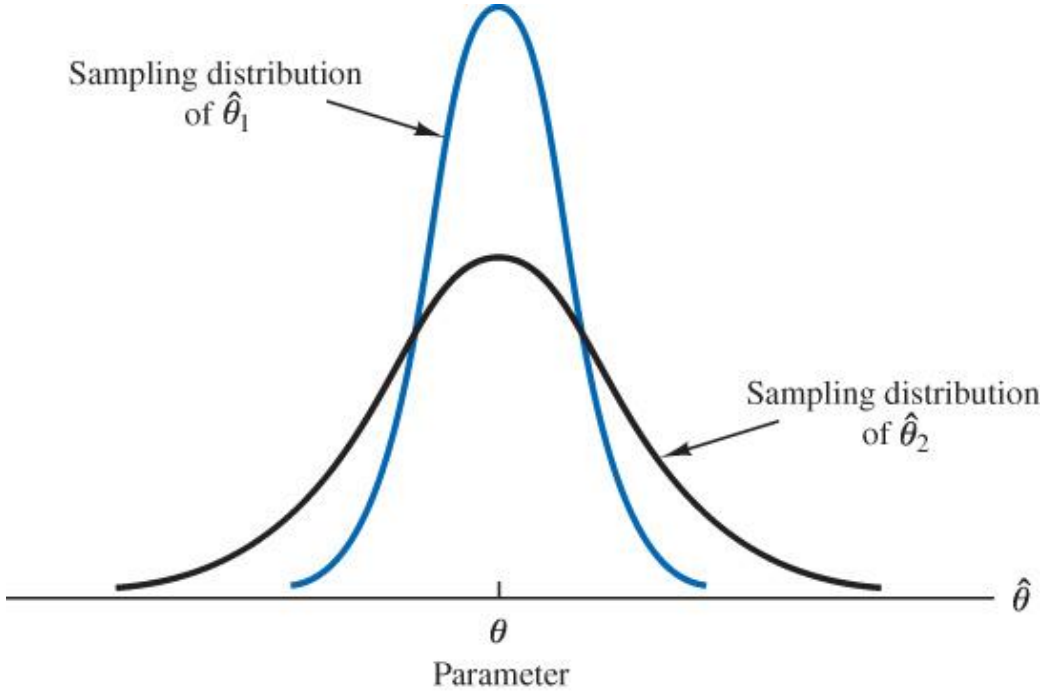
Panel B: Biased Estimator

Efficient and Consistent Point Estimator

Efficiency

A point estimator with a smaller standard error is said to have greater **relative efficiency**.

In the figure to the right, the sampling distribution of $\hat{\theta}_1$ is the preferred point estimator of θ because it has a smaller standard error than the sampling distribution of $\hat{\theta}_2$ does.



Consistency

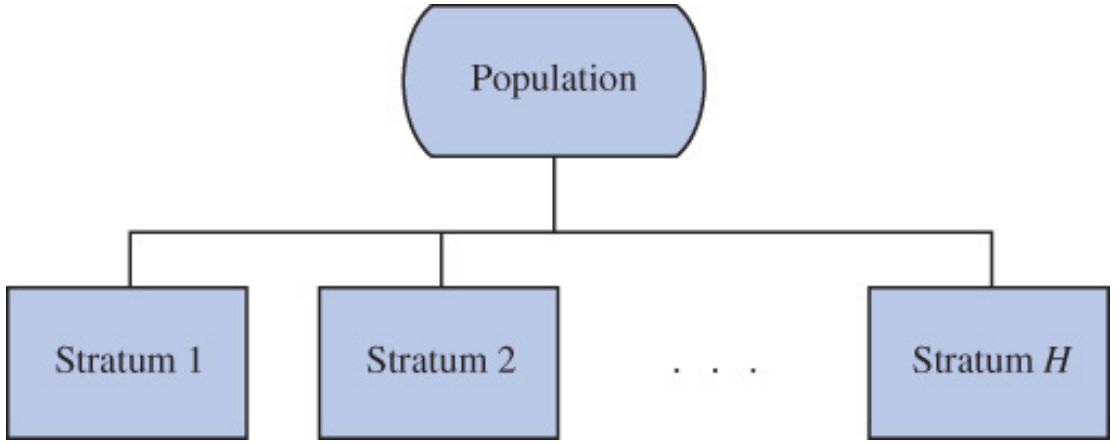
A point estimator is **consistent** if the values of the point estimator tend to become closer to the population parameter as the sample size becomes larger. In other words, a large sample size tends to provide a better point estimate than a small sample size.

Stratified Random Sampling

The population is first divided into groups of elements called *strata*.

Each element in the population must belong 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.)



A simple random sample is then taken from each stratum. Formulas are used (i.e., weighted average) for combining the stratum sample results into one population parameter estimate.

Advantage: if the strata are homogeneous, this method may provides results that are more "precise" than simple random sampling.

Disadvantage: may require a larger total sample size than simple random sampling.

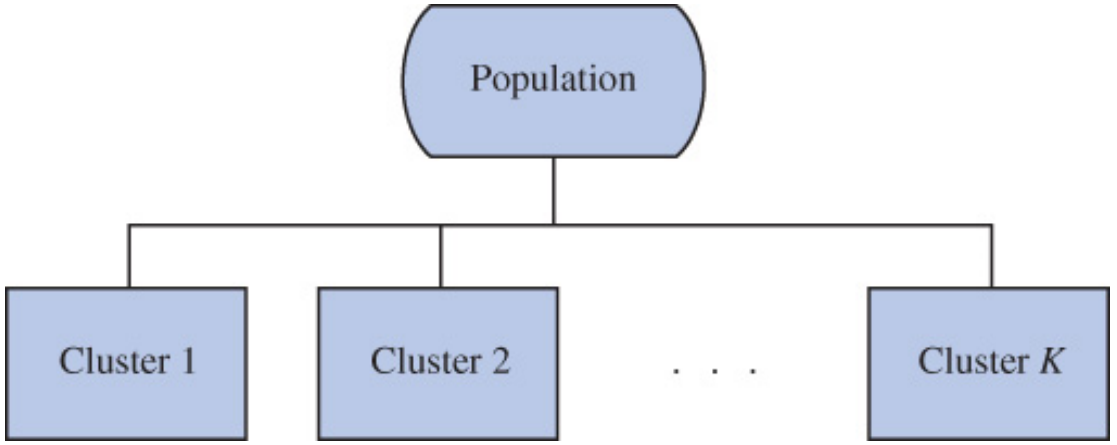
Example: the basis for forming the strata might be department, location, age, industry type, etc.

Cluster Sampling

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.

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: typical in area sampling, with clusters being well-defined geographical areas.

Systematic Sampling

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.

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.

Nonprobability Sampling Methods

Convenience Sampling

Items are included in the sample without known probabilities of being selected.

The sample is identified primarily by convenience.

It is a *nonprobability sampling* technique.

Advantage: sample selection and data collection are relatively easy.

Disadvantage: it is impossible to determine how representative of the population the sample is.

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

Judgment Sampling

The person most knowledgeable on the subject of the study selects elements of the population that they feel as most representative of the population.

It is a nonprobability sampling technique.

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 (*see notes.)

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

Sampling Error

Sampling error is the natural deviation of the sample from the population that results from random sampling.

On average, random samples will be representative of the population from which they are collected.

However, the random collection of sample data does not ensure that every single sample will be perfectly representative of the population of interest.

Because the standard error of the sampling distribution of a sample statistic becomes smaller as the size of the sample increases, we can limit the impact of sampling error by collecting larger sample sizes.

Non-sampling Error

Deviations of the sample from the population that occur for reasons other than random sampling are referred to as **non-sampling error**.

Types of non-sampling error may include:

- **Coverage error** may occur when the research objective and the population from which the sample is to be drawn are not aligned.
- **Nonresponse error** may occur in surveys when some segments of the population are either more or less likely to respond to the survey mechanism.
- **Measurement error** may result from the incorrect measurement of the population characteristic of interest.
- **Interviewer error** may be caused by introducing a bias toward positive or negative responses during a telephone or in-person survey.
- **Processing error** may be introduced during data recording and preparation.

Big Data and Sampling Error

Big data is defined as any set of data that is too large or too complex to be handled by standard data-processing techniques and typical desktop software.

The **Four V's** are the four attributes that characterize the processes that generate big data:

- **Volume** – the amount of data generated
- **Variety** – the diversity in types and structures of data generated
- **Veracity** – the reliability of the data generated
- **Velocity** – the speed at which the data are generated

The large structural dimension of a data set may present the following challenges:

- **Tall data** – a data set has so many observations that make traditional statistical inference obsolete.
- **Wide data** – a data set has so many variables that render infeasible their simultaneous consideration.