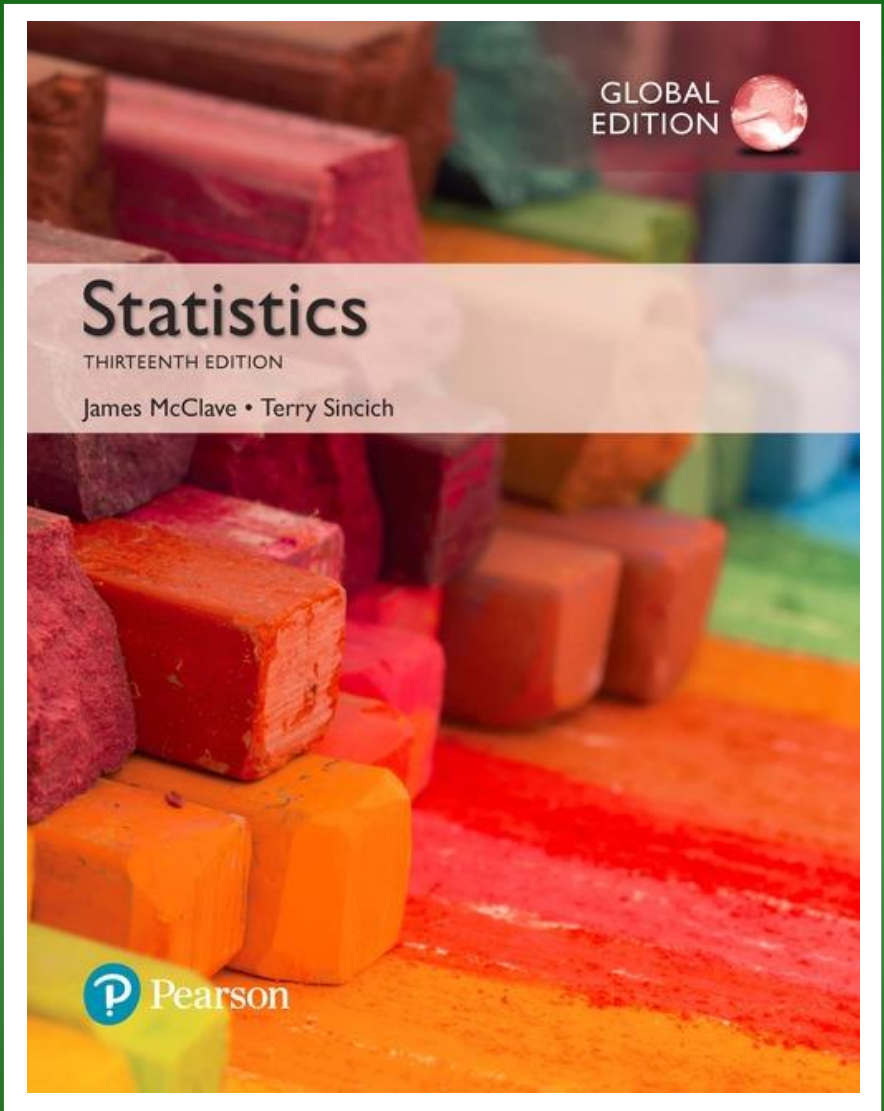


Chapter 6

Sampling Distributions

Known vs. unknown parameters

- Previously, we assumed that we knew the probability distribution of a random variable, and using this knowledge, we were able to compute the mean, variance, and probabilities associated with the random variable.
- However, in most practical applications, this information is not available.

Known vs. unknown parameters

- In most situations, the true mean and standard deviation are unknown quantities that have to be estimated.
- Numerical quantities that describe probability distributions are called **parameters**.

Definition

A **parameter** is a numerical descriptive measure of a population. Because it is based on the observations in the population, its value is almost always unknown.

- Thus, p, the probability of a success in a binomial experiment, and μ and σ, the mean and standard deviation, respectively, of a normal distribution, are examples of parameters.

Definition

- We have also discussed the sample mean $\overline{x}$, sample variance s^2, sample standard deviation s, and the like, which are numerical descriptive measures calculated from the sample.
- We will often use the information contained in these sample statistics to make inferences about the parameters of a population.

A **sample statistic** is a numerical descriptive measure of a sample. It is calculated from the observations in the sample.

Table 6.1

Table 6.1 List of Population Parameters and Corresponding Sample Statistics		
	Population Parameter	Sample Statistic
Mean:	μ	$\bar{x}$
Median:	η	M
Variance:	σ^2	s^2
Standard deviation:	σ	s
Binomial proportion:	p	$\hat{p}$

- Note that the term **statistic** refers to a **sample quantity** and the term **parameter** refers to a **population quantity**.

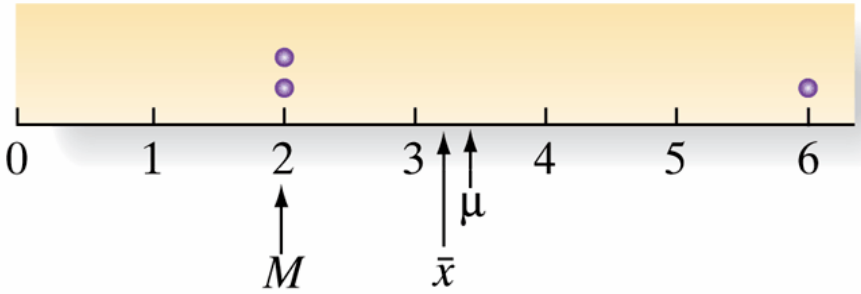
6.1

The Concept of a Sampling Distribution

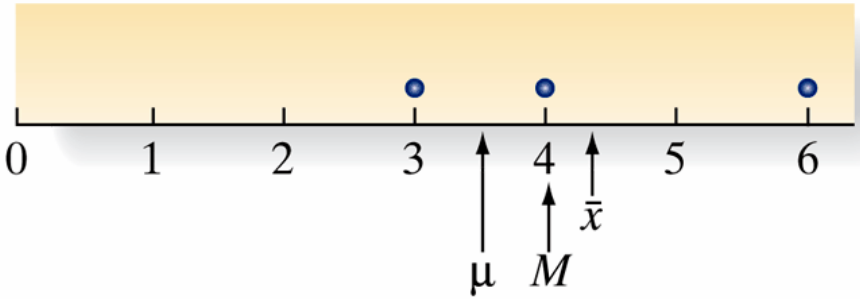
The Concept of a Sampling Distribution

- If we want to estimate a parameter of a population—say, the population mean μ—we can use a number of sample statistics for our estimate.
- Two possibilities are the sample mean $\bar{x}$ and the sample median M.
- Which of these do you think will provide a better estimate of μ?

Figure 6.1 Comparing the sample mean $(\bar{x})$ and sample median (M) as estimators of the population mean (μ)



a. Sample 1: $\bar{x}$ is closer than M to μ



b. Sample 2: M is closer than $\bar{x}$ to μ

- This simple example illustrates an important point: **Neither the sample mean nor the sample median will always fall closer to the population mean.**
- Consequently, we cannot compare these two sample statistics or, in general, any two sample statistics on the basis of their performance with a single sample

The Concept of a Sampling Distribution

- **Sample statistics** are themselves **random variables** because different samples can lead to different values for the sample statistics.
- As random variables, **sample statistics must be judged and compared** on the basis of their **probability distributions** (i.e., the collection of values and associated probabilities of each statistic that would be obtained if the sampling experiment were repeated a *very large number of times*).

Conceptual example

- Suppose it is known that the daily high temperature recorded for all past months of January has a mean $\mu = 10\,^{\circ}C$ and a standard deviation $\sigma = 5\,^{\circ}C$.
- The first sample of 25 temperature measurements have a mean $\overline{x} = 9.8$, the second sample a mean $\overline{x} = 11.4$, the third sample a mean $\overline{x} = 10.5$, etc.
- If the sampling experiment were repeated a very large number of times, the resulting histogram of sample means would be approximately the probability distribution of $\overline{x}$.

Definition

- If $\overline{x}$ is a good estimator of μ, we would expect the values of $\overline{x}$ to cluster around μ as shown in Figure 6.2.
- This probability distribution is called a *sampling distribution* because it is generated by repeating a sampling experiment a very large number of times.

The **sampling distribution** of a sample statistic calculated from a sample of n temperature measurements is the probability distribution of the statistic.

Figure 6.2 Sampling distribution for $\bar{x}$ based on a sample of $n = 25$ temperature measurements

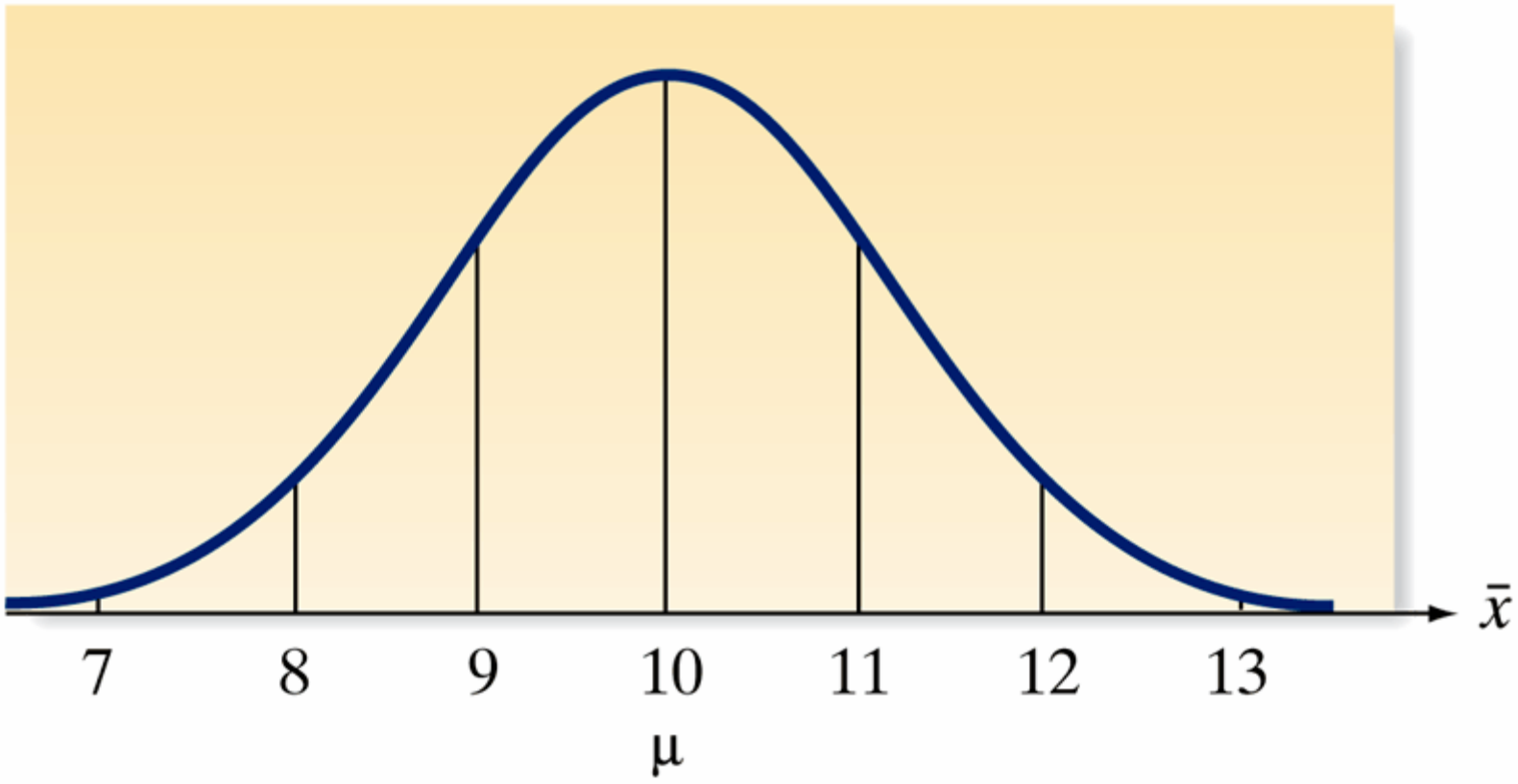
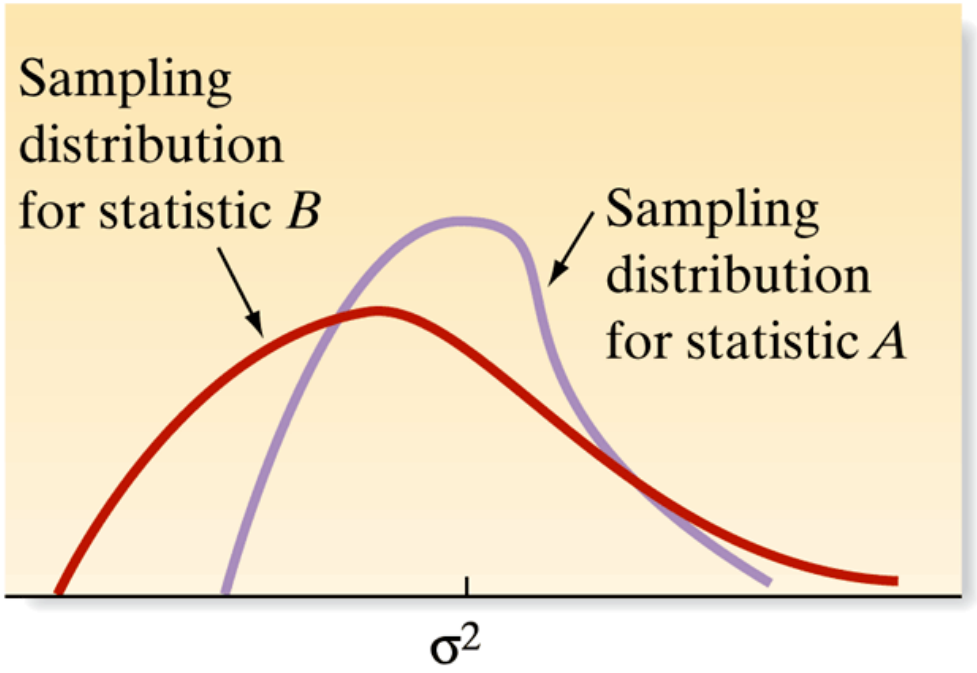


Figure 6.3 Two sampling distributions for estimating the population variance σ^2

- You would prefer statistic A over statistic B. You would do so because the sampling distribution for statistic A centers over s^2 and has less spread (variation) than the sampling distribution for statistic B.



Problem

- In a single toss of a fair coin, let x equal the number of heads observed. Now consider a sample of $n = 2$ tosses. Find the sampling distribution of $\overline{x}$, the sample mean.

Solution

- For one coin toss, the result is either a head (H) or tail (T).
- The value of x is either $x = 1$ or $x = 0$.
- The four possible outcomes (sample points) for two coin tosses and corresponding probabilities are listed Table 6.2.
- Since a fair coin is tossed, these four possible outcomes are equally likely.

Table 6.2

Table 6.2 Outcomes for $n = 2$ Coin Tosses		
Outcome (Toss 1, Toss 2)	Probability	$\bar{x}$
H H $(x = 1, x = 1)$	$1/4$	1
H T $(x = 1, x = 0)$	$1/4$	.5
T H $(x = 0, x = 1)$	$1/4$	.5
T T $(x = 0, x = 0)$	$1/4$	0

- The value of x for each outcome is also listed in Table 6.2.

Solution

- You can see that the values are either 0, .5, or 1.

$\bar{x}$	0	.5	1
$p(\bar{x})$	$1/4$	$1/2$	$1/4$

Problem

- The rolling machine of a steel manufacturer produces sheets of steel of varying thickness. The thickness of a steel sheet follows a uniform distribution with values between 150 and 200 millimeters. Suppose we perform the following experiment over and over again: Randomly sample 11 steel sheets from the production line and record the thickness x of each. Calculate the two sample statistics
  - $\overline{x}$ = Sample mean = $\frac{\sum x}{11}$
  - M = Median = Sixth sample measurement when the 11 thicknesses are arranged in ascending order
- Find approximations to the sampling distributions of $\overline{x}$ and M.

Figure 6.4 Uniform distribution for thickness of steel sheets

- The population of thicknesses follows the uniform distribution shown in Figure 6.4.

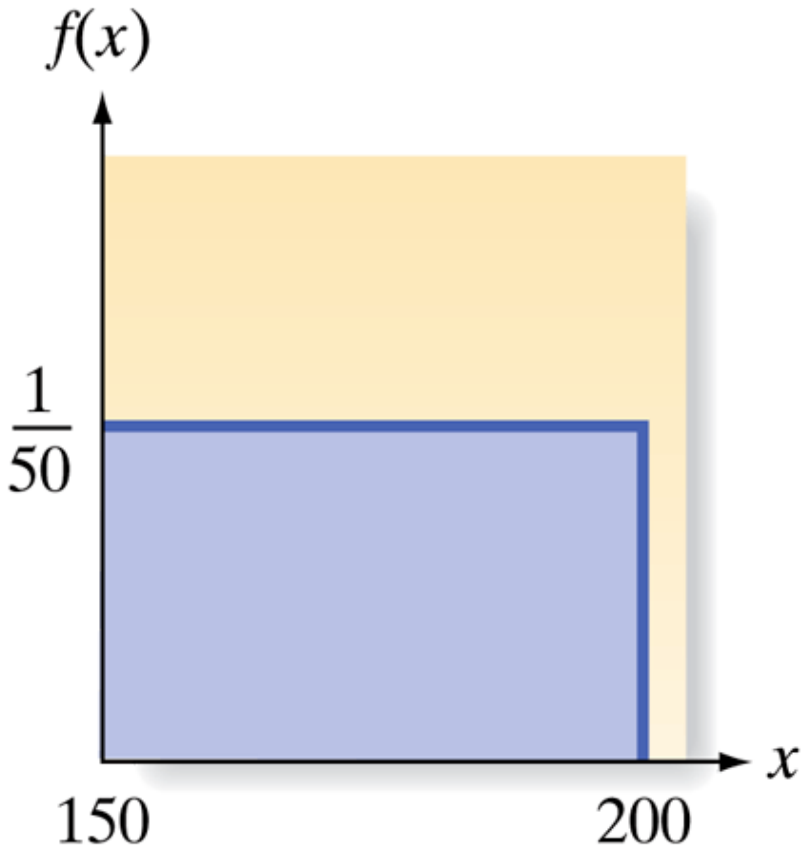
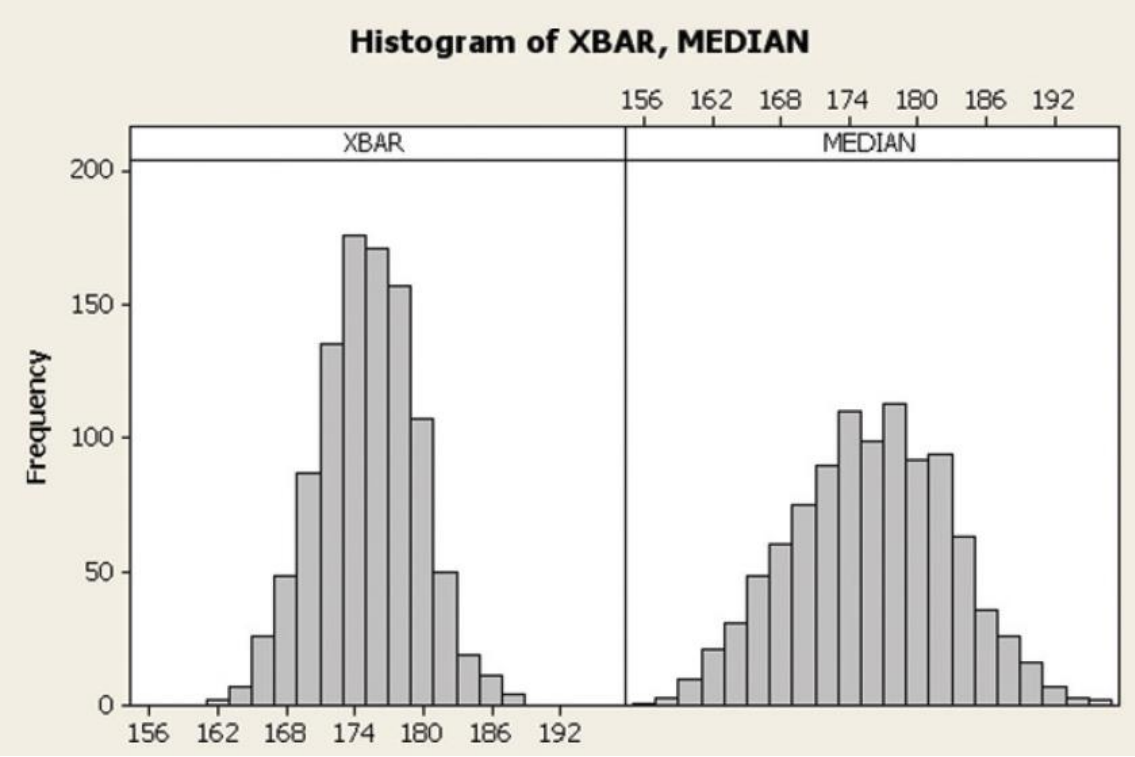


Table 6.4

Table 6.4 First 10 Samples of $n = 11$ Thickness Measurements from Uniform Distribution													
Sample	Thickness Measurements											Mean	Median
1	173	171	187	151	188	181	182	157	162	169	193	174.00	173
2	181	190	182	171	187	177	162	172	188	200	193	182.09	182
3	192	195	187	187	172	164	164	189	179	182	173	180.36	182
4	173	157	150	154	168	174	171	182	200	181	187	172.45	173
5	169	160	167	170	197	159	174	174	161	173	160	169.46	169
6	179	170	167	174	173	178	173	170	173	198	187	176.55	173
7	166	177	162	171	154	177	154	179	175	185	193	172.09	175
8	164	199	152	153	163	156	184	151	198	167	180	169.73	164
9	181	193	151	166	180	199	180	184	182	181	175	179.27	181
10	155	199	199	171	172	157	173	187	190	185	150	176.18	173

Figure 6.5 Histograms for sample mean and sample median

- You can see that the values of x tend to cluster around m to a greater extent than do the values of M.
- Thus, on the basis of the observed sampling distributions, we conclude that x contains more information about μ than M does —at least for samples of $n = 11$ measurements from the uniform distribution.



6.2

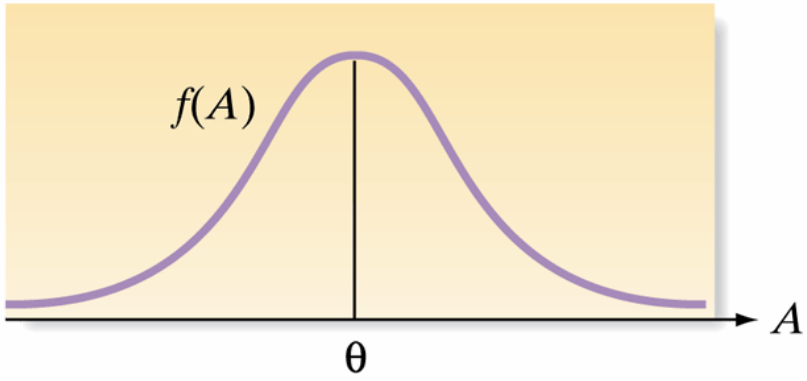
Properties of Sampling Distributions: Unbiasedness and Minimum Variance

Definition

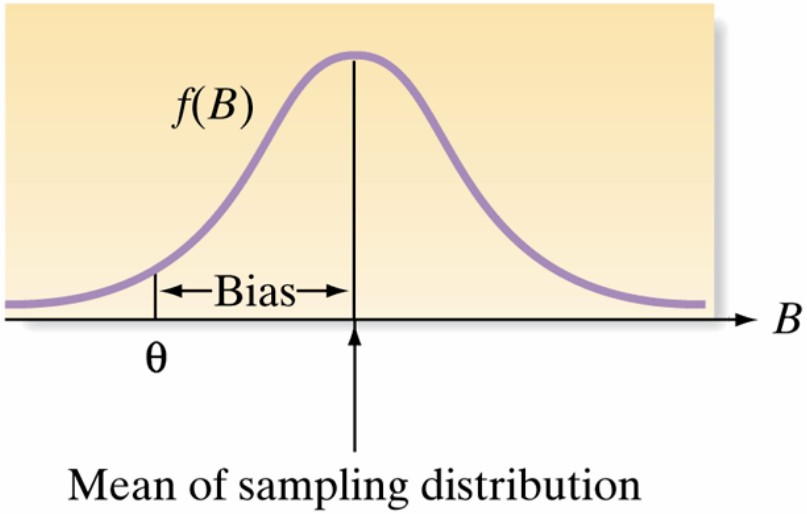
A **point estimator** of a population parameter is a rule or formula that tells us how to use the sample data to calculate a single number that can be used as an *estimate* of the population parameter.

- For example, the sample mean $\overline{x}$ is a point estimator of the population mean μ.
- Similarly, the sample variance s^2 is a point estimator of the population variance σ^2.

Figure 6.6 Sampling distributions of unbiased and biased estimators



a. Unbiased sample statistic for the parameter θ



b. Biased sample statistic for the parameter θ

Definition

Unbiased and Biased Estimators

If the sampling distribution of a sample statistic $(\hat{\theta})$ has a mean equal to the population parameter (θ) the statistic is intended to estimate, the statistic is said to be an **unbiased estimate** of the parameter. If the mean of the sampling distribution is not equal to the parameter, the statistic is said to be a **biased estimate** of the parameter.

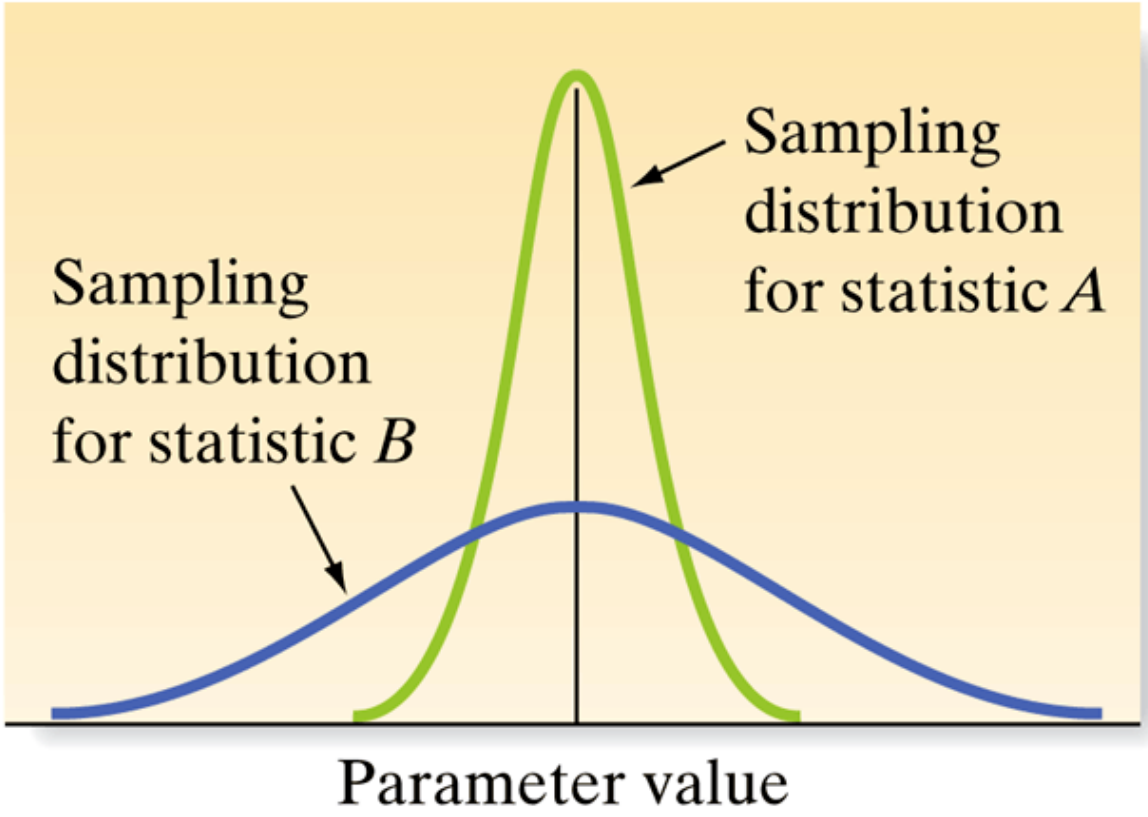
$$\text{Unbiased:} \quad E(\hat{\theta}) = \theta$$

$$\text{Biased:} \quad E(\hat{\theta}) \neq \theta$$

Figure 6.7 Sampling distributions for two unbiased estimators

Naturally, we will choose the sample statistic that has the smaller standard deviation.

Note: The standard deviation of the sampling distribution of a statistic is also called the **standard error of the statistic.**



6.3

The Sampling Distribution of $\bar{x}$ and the Central Limit Theorem

Figure 6.8 Sampled uniform population

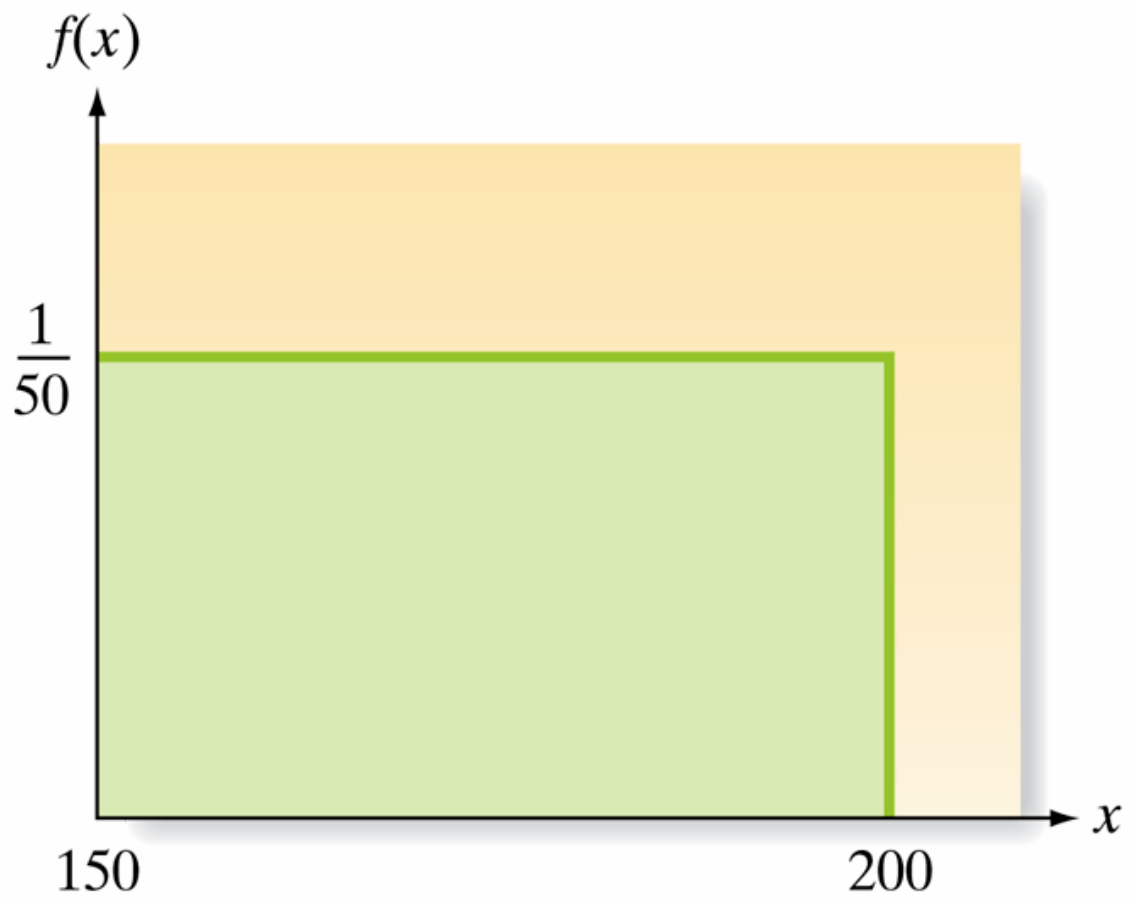
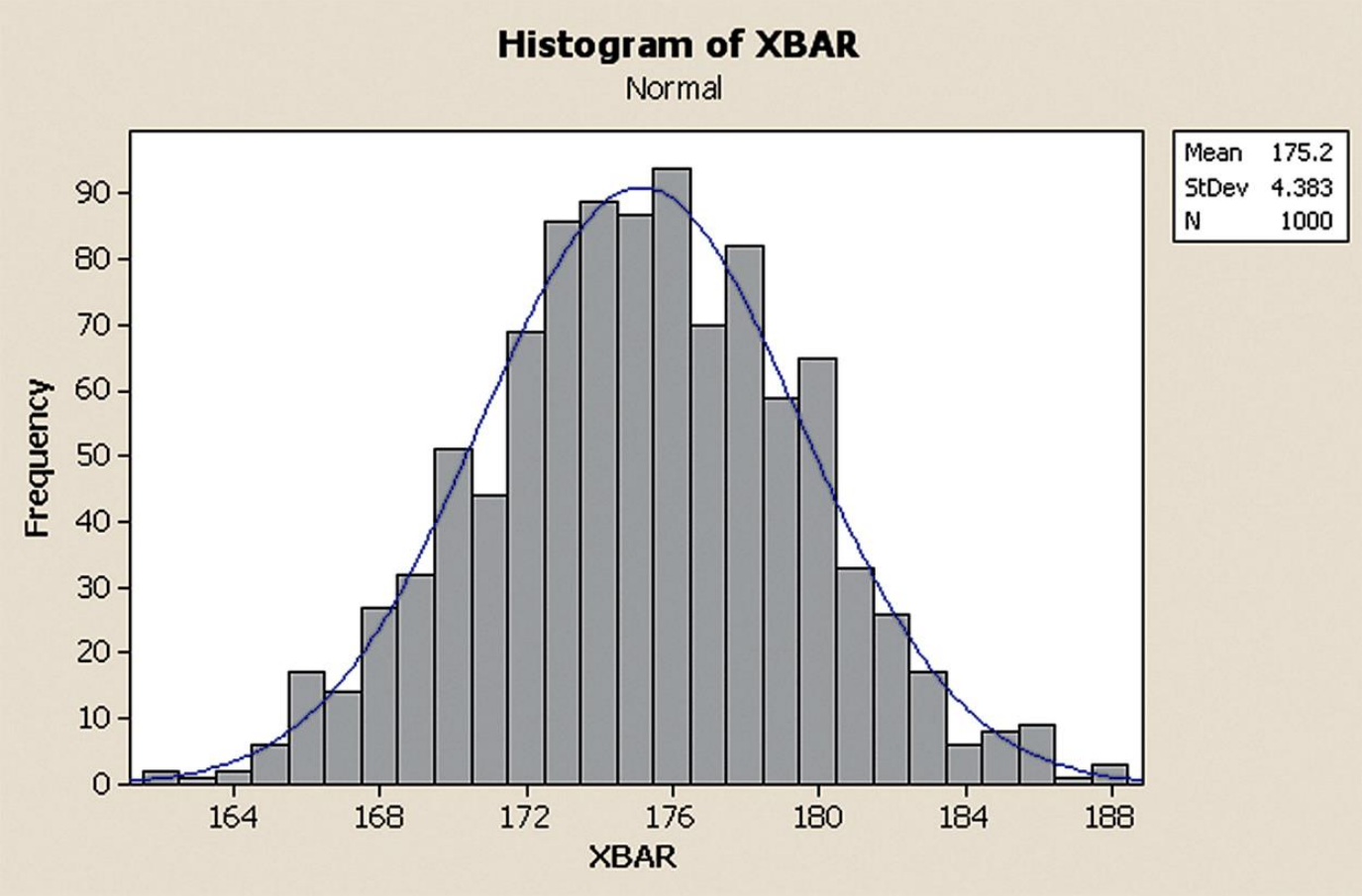


Figure 6.9 MINITAB histogram for sample mean in 1,000 samples



Definition

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

1. The mean of the sampling distribution of $\bar{x}$ equals the mean of the sampled population. That is, $\mu_{\bar{x}} = E(\bar{x}) = \mu$.
2. The standard deviation of the sampling distribution of $\bar{x}$ equals

$$\frac{\text{Standard deviation of sampled population}}{\text{Square root of sample size}}$$

That is, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$*

The standard deviation $\sigma_{\bar{x}}$ is often referred to as the **standard error of the mean**.

Theorem

Theorem 6.1

If a random sample of n observations is selected from a population with a normal distribution, the sampling distribution of $\bar{x}$ will be a normal distribution.

Theorem

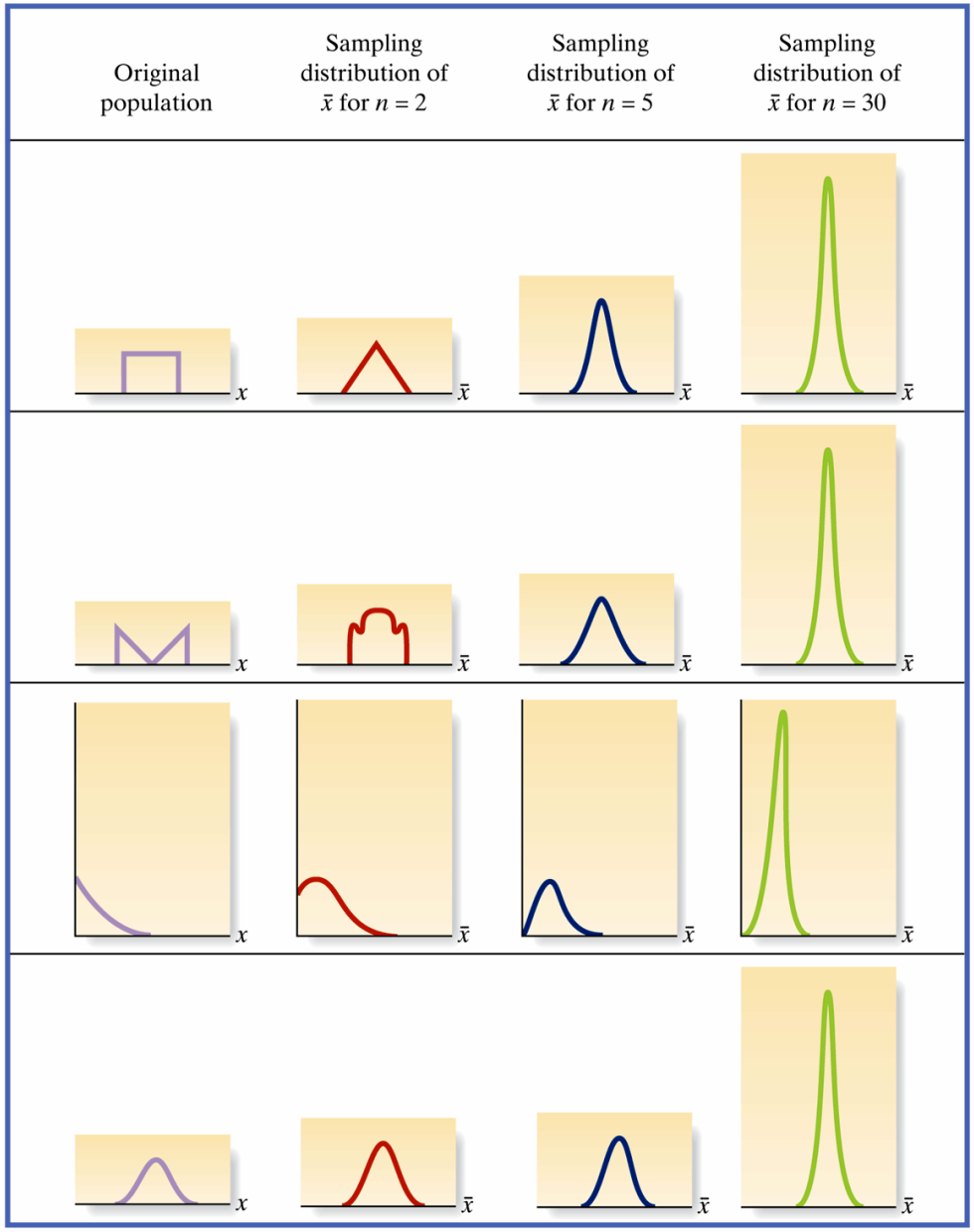
Theorem 6.2: Central Limit Theorem

Consider a random sample of n observations selected from a population (*any* population) with mean μ and standard deviation σ. Then, when n is sufficiently large, the sampling distribution of $\bar{x}$ will be approximately a normal distribution with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. The larger the sample size, the better will be the normal approximation to the sampling distribution of $\bar{x}$.*

*Moreover, because of the Central Limit Theorem, the sum of a random sample of n observations, Σx, will possess a sampling distribution that is approximately normal for large samples. This distribution will have a mean equal to $n\mu$ and a variance equal to $n\sigma^2$. Proof of the Central Limit Theorem is beyond the scope of this book, but it can be found in many mathematical statistics texts.

Figure 6.10

Sampling distributions of $\bar{x}$ for different populations and different sample sizes



Sufficiently large samples have a sampling distribution of $\bar{x}$ that is approximately normal

- Generally speaking, the greater the skewness of the sampled population distribution, the larger the sample size must be before the normal distribution is an adequate approximation to the sampling distribution of x.
- For most sampled populations, sample sizes of $n \geq 30$ will suffice for the normal approximation to be reasonable.

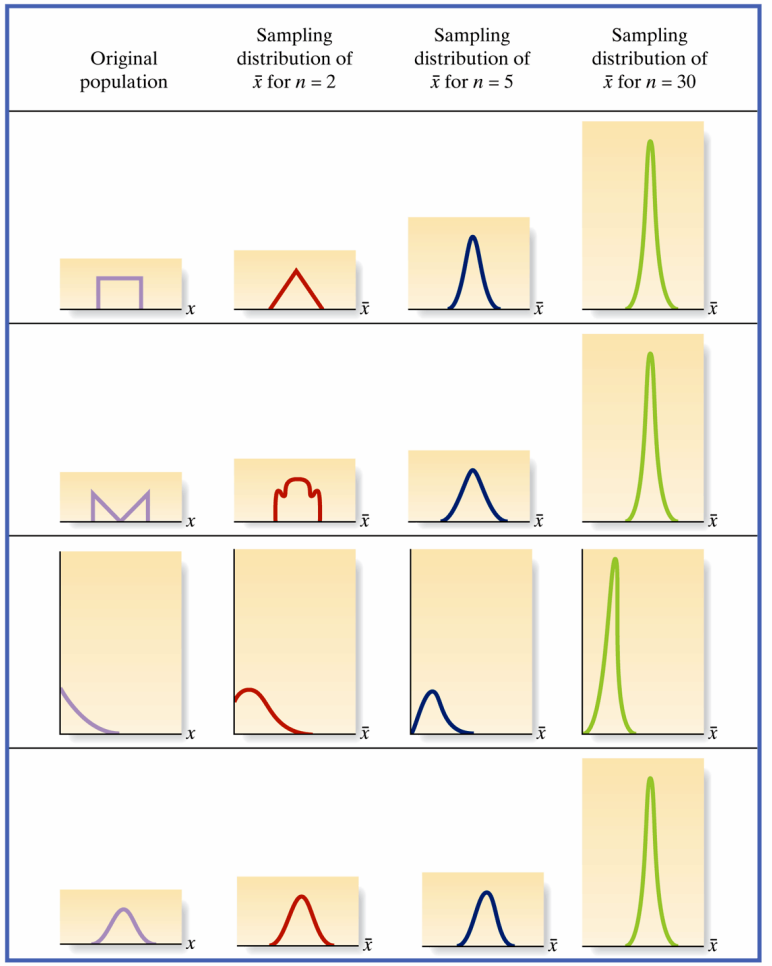
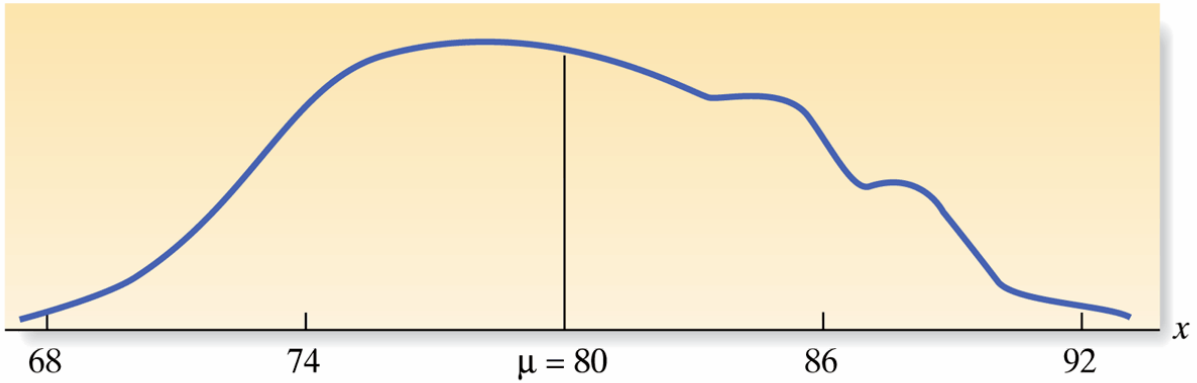
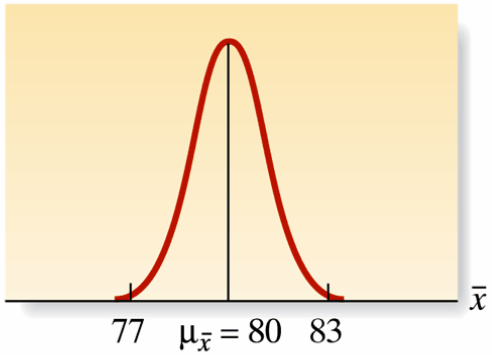


Figure 6.11 A population relative frequency distribution and the sampling distribution for $\overline{x}$

- Suppose we have selected a random sample of $n = 36$ observations from a population with mean equal to 80 and standard deviation equal to 6. It is known that the population is not extremely skewed.
  - Sketch the relative frequency distributions for the population and for the sampling distribution of the sample mean $\overline{x}$.
  - Find the probability that $\overline{x}$ will be larger than 82.

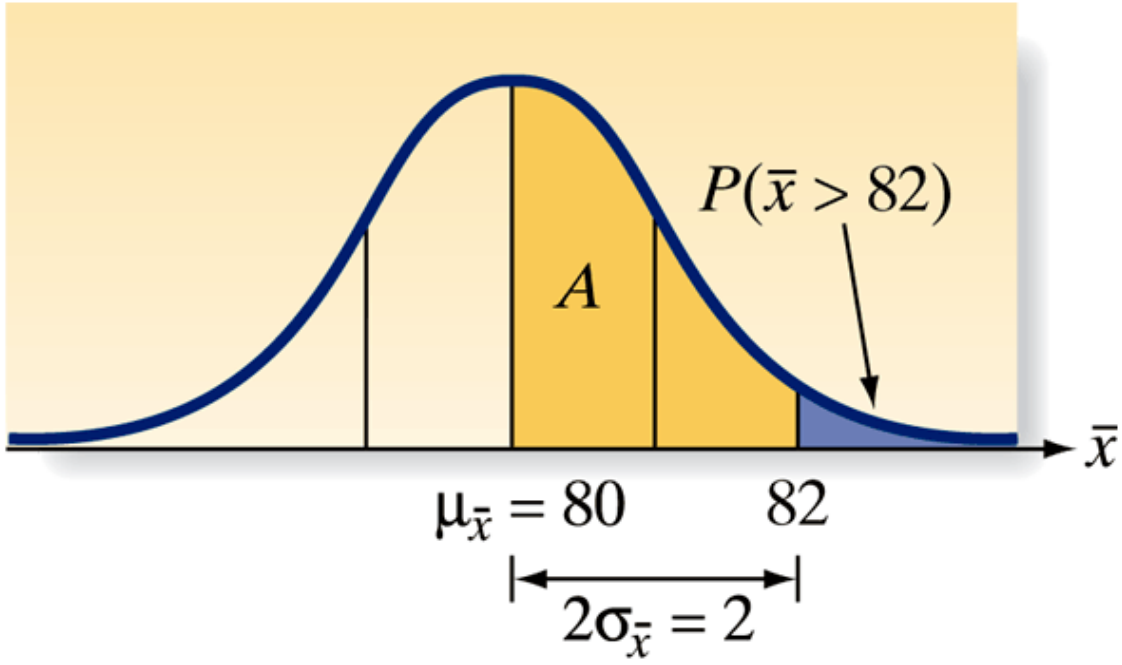


a. Population relative frequency distribution



b. Sampling distribution of $\overline{x}$

Figure 6.12 The sampling distribution of $\bar{x}$

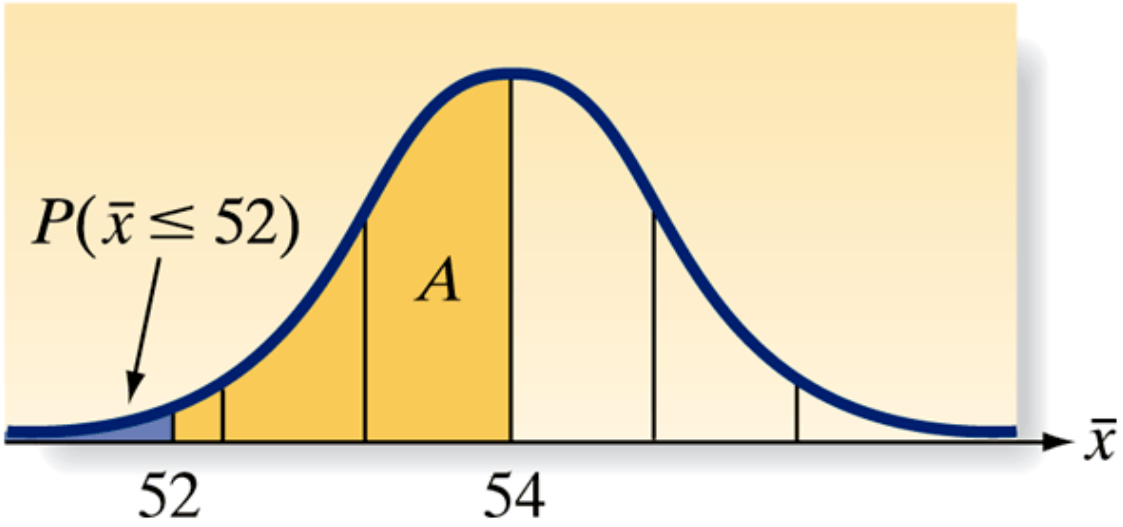


$$P(\bar{x}>82) = P(z>2) = .5 - .4772 = .0228$$

Problem

- A manufacturer of automobile batteries claims that the distribution of the lengths of life of its best battery has a mean of 54 months and a standard deviation of 6 months. Suppose a consumer group decides to check the claim by purchasing a sample of 50 of the batteries and subjecting them to tests that estimate the battery's life.
  - Assuming that the manufacturer's claim is true, describe the sampling distribution of the mean lifetime of a sample of 50 batteries.
  - Assuming that the manufacturer's claim is true, what is the probability that the consumer group's sample has a mean life of 52 or fewer months?

Figure 6.13 The sampling distribution of $\bar{x}$ in Example 6.9 for $n = 50$



$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{50}} = .85 \text{ month}$$

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{52 - 54}{.85} = -2.35$$

$$P(\bar{x} \leq 52) = .5 - A = .5 - .4906 = .0094$$