Chapter 7 Sampling Distributions

In Chapter 1 we introduced random sampling. In this chapter we continue our discussion of random sampling by explaining what a random sample is and how to select a random sample. In addition, we discuss two probability distributions that are related to random sampling. To understand these distributions, note that if we select a random sample, then we use the sample mean as the point estimate of the population mean and the sample proportion as the point estimate of the population proportion. Two probability distributions that help us assess how accurate the sample mean and sample proportion are likely to be as point estimates are **the sampling distribution of the sample mean** and **the sampling distribution of the sample proportion.** After discussing random sampling in the first section of this chapter, we consider these sampling distributions in the next two sections. Moreover, using the car mileage case, the e-billing case, and the cheese spread case, we demonstrate how sampling distributions can be used to make statistical inferences.

The discussions of random sampling and of sampling distributions given in the first three sections of this chapter are necessary for understanding the rest of this book. The last three sections of this chapter consider advanced aspects of sampling and are optional. In the first optional section, we discuss three alternatives to random sampling—**stratified random sampling, cluster sampling,** and **systematic sampling.** In the second optional section, we discuss issues related to designing surveys and the errors that can occur in survey sampling. In the last optional section, we derive the mean and variance of the sample mean.

Outline

- The Sampling Distribution of the Sample Mean
- The Sampling Distribution of the Sample Proportion

7.1 Sampling Distribution of the Sample Mean

The sampling distribution of the sample mean $\bar{X}$ is the probability distribution of the population of the sample means obtainable from all possible samples of size n from a population of size *N*

Sample Mean

Example: The law firm of Hoya and Associates has five partners. At their weekly partners meeting each reported the number of hours they billed clients for their services last week.

Partner	Hours
Dunn	22
Hardy	26
Kiers	30
Malory	26
Tillman	22

A total of 10 different samples

If two partners are selected randomly, how many different samples are possible?

Partners	Total	Mean
1,2	48	24
1,3	52	26
1,4	48	24
1,5	44	22
2,3	56	28
2,4	52	26
2,5	48	24
3,4	56	28
3,5	52	26
4,5	48	24

As a sampling distribution

Sample Mean	*Frequency*	*Relative Frequency probability*
22	1	1/10
24	4	4/10
26	3	3/10
28	2	2/10

Different samples of the same size from the same population will yield different sample means

Sample Mean

- Let there be a population of units of size N
- Consider all its samples of a fixed size n ($n<N$)
- For all possible samples of size n, we obtain a population of sample means.

  That is, $\bar{x}$ ***is a random variable*** which may have all these means as its values.

- Before we draw the sample, the sample mean $\bar{x}$ is a random variable.
- We consider the probability distribution of the random variable $\bar{x}$, i.e., the probability distribution for the population of sample means

The **sampling distribution of the sample mean** is the probability distribution of the population of the sample means obtainable from all possible samples of size n from a population of size N.

Example 7.1

The Stock Return Case

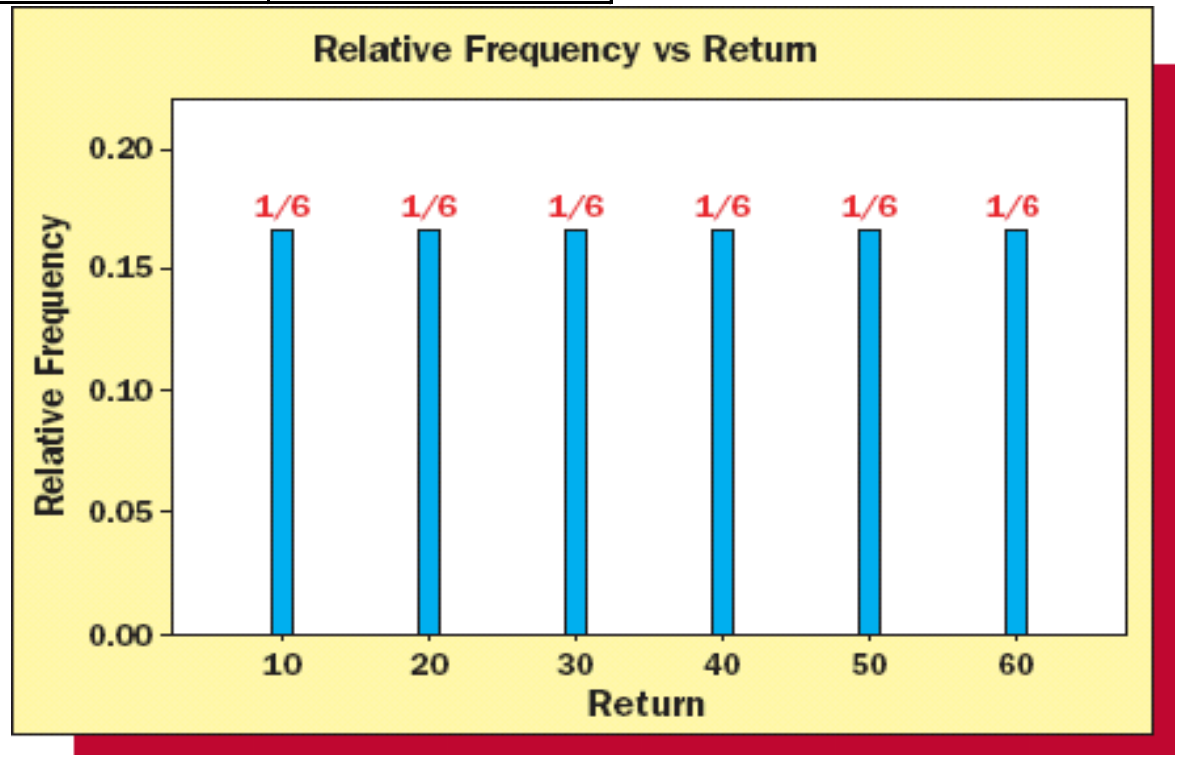
- We have a population of the percent returns from six stocks
  - In order, the values of % return are:

    10%, 20%, 30%, 40%, 50%, and 60%

    - Label each stock A, B, C, …, F in order of increasing % return
    - The mean rate of return is 35% with a standard deviation of 17.078%
  - Any one stock of these stocks is as likely to be picked as any other of the six
    - Discrete uniform distribution with $N = 6$
    - Each stock has a probability of being picked of 1/6

The Stock Return Case # 2

Stock	*% Return*	*Frequency*	*Relative Frequency*
Stock A	10	1	1/6
Stock B	20	1	1/6
Stock C	30	1	1/6
Stock D	40	1	1/6
Stock E	50	1	1/6
Stock F	60	1	1/6
Total		**6**	**1**



The Stock Return Case # 3

- Now, select all possible samples of size $n = 2$ from this population of stocks of size $N = 6$
  - That is, select all possible pairs of stocks
- How to select?
  - Sample randomly
  - Sample without replacement
  - Sample without regard to order

Recall: Combination

- The combination is a way of selecting items from a collection, such that the order of selection does not matter.
- For example, given three elements, say A, B and C, there are three combinations of two that can be drawn from this set: AB, AC, BC.
- Combination Formula: $$C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- For example, n=3, k=2,

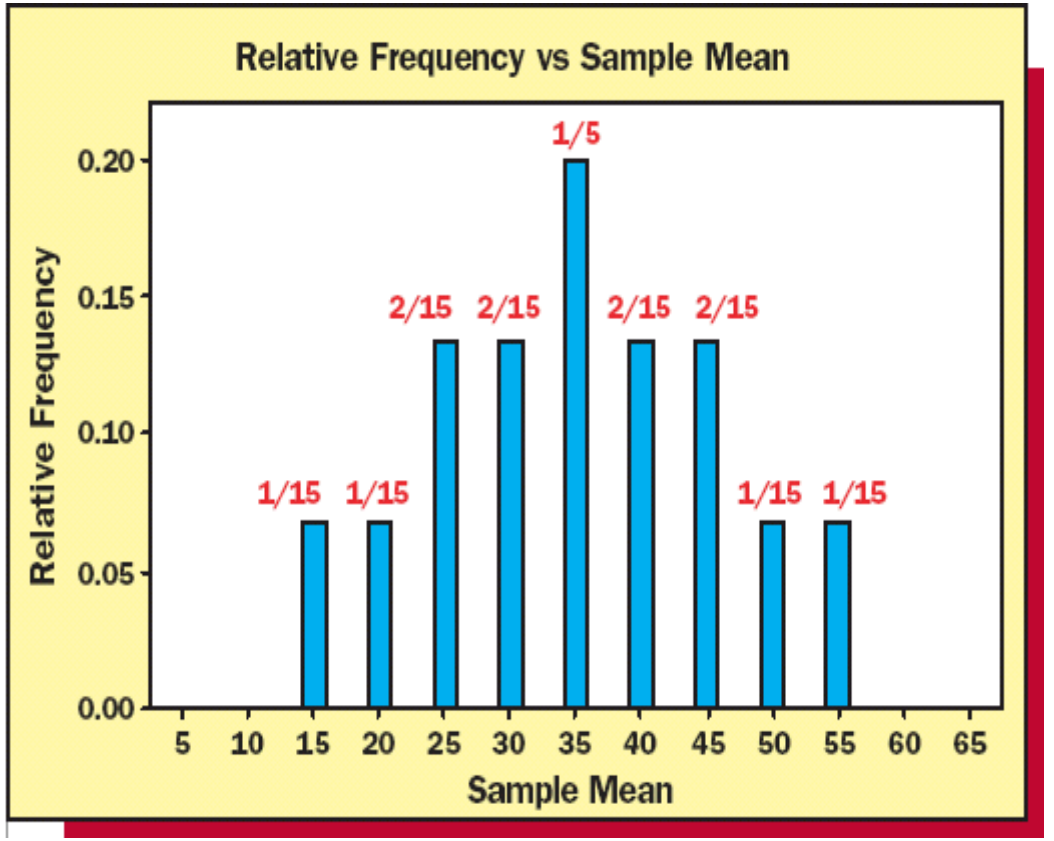
$$C_3^2 = \binom{3}{2} = \frac{3!}{2!\,(3-2)!} = 3$$

The Stock Return Case # 4

- Result: There are $C_6^2 = 15$ possible samples of size $n = 2$
- Calculate the sample mean of each and every sample
- For example, if we choose the two stocks with returns 10% and 20%, then the sample mean is 15%

The Stock Return Case # 5

Sample Mean	Frequency	Relative Frequency
15	1	1/15
20	1	1/15
25	2	2/15
30	2	2/15
35	3	3/15
40	2	2/15
45	2	2/15
50	1	1/15
55	1	1/15

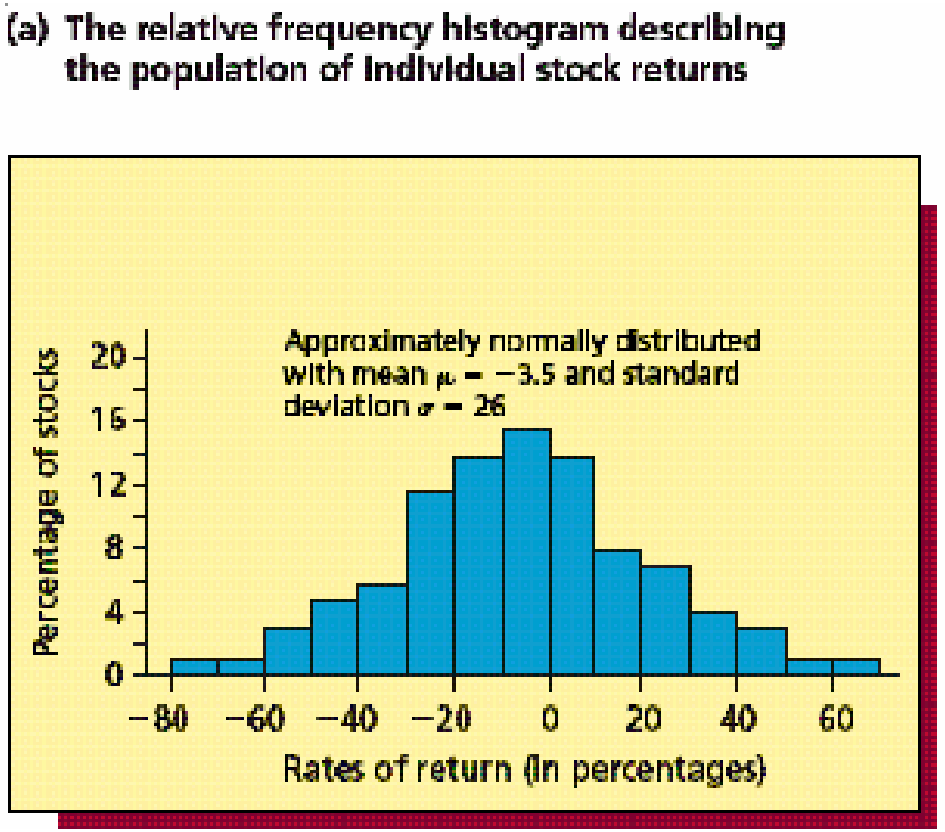


Observations

- The population of $N = 6$ stock returns has a uniform distribution.
- But the histogram of $n = 15$ sample mean returns:
  1. Seems to be centered over the same mean return of 35%, and
  2. Appears to be bell-shaped and less spread out than the histogram of individual returns

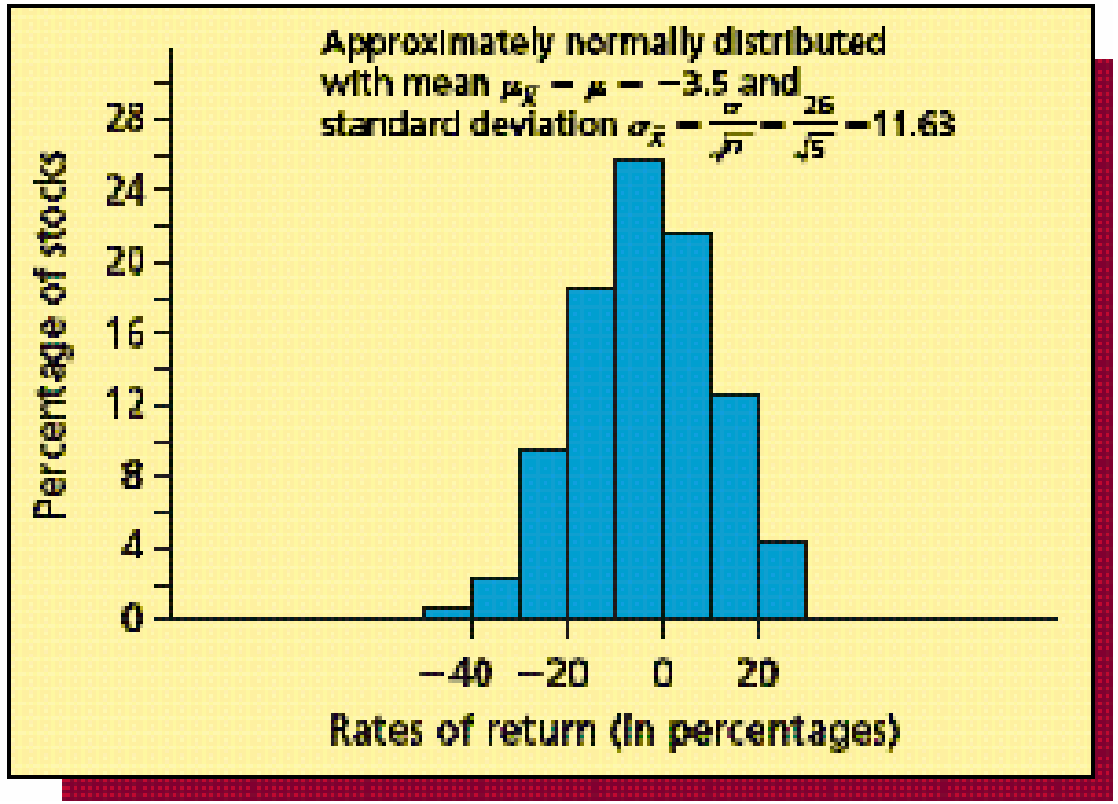
Sampling all the stocks

- Consider the population of returns of all *1,815* stocks listed on *NYSE* for *1987*
  - The mean rate of return μ was *–3.5%* with a standard deviation σ of *26%*



Draw all possible random samples of size $n = 5$ and calculate the sample mean return of each Sample with a computer

(b) The relative frequency histogram describing the population of all possible sample mean returns when $n = 5$



- Observations
  - Both histograms appear to be bell-shaped and centered over the same mean of *–3.5%*
  - The histogram of the sample mean returns looks less spread out than that of the individual returns
- Statistics
  - Mean of all sample means: $\mu_{\bar{x}} = \mu = -3.5\%$
  - Standard deviation of all possible means:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{26}{\sqrt{5}} = 11.63\%$$

General Conclusions

- If the population of individual items is normal, then the population of all sample means is also normal
- Even if the population of individual items is not normal, there are circumstances that the population of all sample means is normal (see *Central Limit Theorem* later)

General Conclusions

- The mean of all possible sample means equals the population mean
  - That is, $\mu_{\bar{x}} = \mu$

- The standard deviation $\sigma_{\bar{x}}$ of all sample means is less than the standard deviation of the population
  - That is, $\sigma_{\bar{x}} < \sigma$

- The empirical rule holds for the sampling distribution of the sample mean
  - *68.26%* of all possible sample means are within (plus or minus) one standard deviation $\sigma_{\bar{x}}$ of $\mu_{\bar{x}}$
  - *95.44%* of all possible observed values of *x* are within (plus or minus) two $\sigma_{\bar{x}}$ of $\mu_{\bar{x}}$
    - In the example, *95.44%* of all possible sample mean returns are in the interval *[-3.5 ± (2×11.63)] = [-3.5 ± 23.26]*
    - That is, *95.44%* of all possible sample means are between *-26.76%* and *19.76%*
  - 99.73% of all possible observed values of *x* are within (plus or minus) three $\sigma_{\bar{x}}$ of $\mu_{\bar{x}}$

Properties of the Sampling Distribution of the Sample Mean #1

- If the population being sampled is normal, then so is the sampling distribution of the sample mean $\bar{x}$
- The mean $\mu_{\bar{x}}$ of the sampling distribution of $\bar{x}$ is $\mu_{\bar{x}} = \mu$

  That is, the mean of all possible sample means is the same as the population mean

Properties of the Sampling Distribution of the Sample Mean #2

- The variance $\sigma_{\bar{x}}^2$ of the sampling distribution of $\bar{x}$ is

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

  - That is, the variance $\sigma_{\bar{x}}^2$ of the sampling distribution of $\bar{x}$ is
    - directly proportional to the variance of the population, and
    - inversely proportional to the sample size

Properties of the Sampling Distribution of the Sample Mean #3

- The standard deviation $\sigma_{\bar{x}}$ of the sampling distribution of $\bar{x}$ is

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

  - That is, the standard deviation $\sigma_{\bar{x}}$ of the sampling distribution of $\bar{x}$ is
    - directly proportional to the standard deviation of the population, and
    - inversely proportional to the square root of the sample size

The Sampling Distribution of $\bar{x}$

Assume that the population from which we will randomly select a sample of n measurements has mean μ and standard deviation σ. Then, the population of all possible sample means

1. Has a normal distribution, if the sampled population has a normal distribution.
2. Has mean $\mu_{\bar{x}} = \mu$.
3. Has standard deviation $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$.

The formula for $\sigma_{\bar{x}}$ in (3) holds exactly if the sampled population is infinite. If the sampled population is finite, this formula holds approximately under conditions to be discussed at the end of this section.

Stated equivalently, the sampling distribution of $\bar{x}$ has mean $\mu_{\bar{x}} = \mu$, has standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ (if the sampled population is infinite), and is a normal distribution (if the sampled population has a normal distribution).[2]

Notes

- $\bar{x}$ is the point estimate of μ, and the larger the sample size n, the more accurate the estimate, because when n increases, $\sigma_{\bar{x}}$ decreases, $\bar{x}$ is more clustered to the population
  - In order to reduce $\sigma_{\bar{x}}$, take bigger samples!

Example 7.2

Car Mileage Case

EXAMPLE 7.3 The Car Mileage Case: Estimating Mean Mileage

Part 1: Basic concepts Consider the infinite population of the mileages of all of the new mid-size cars that could potentially be produced by this year's manufacturing process. If we assume that this population is normally distributed with mean μ and standard deviation $\sigma = .8$ (see Figure 7.3(a)), and if the automaker will randomly select a sample of n cars and test them as prescribed by the EPA, it follows that the population of all possible sample means is normally distributed with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n} = .8/\sqrt{n}$. In order to show

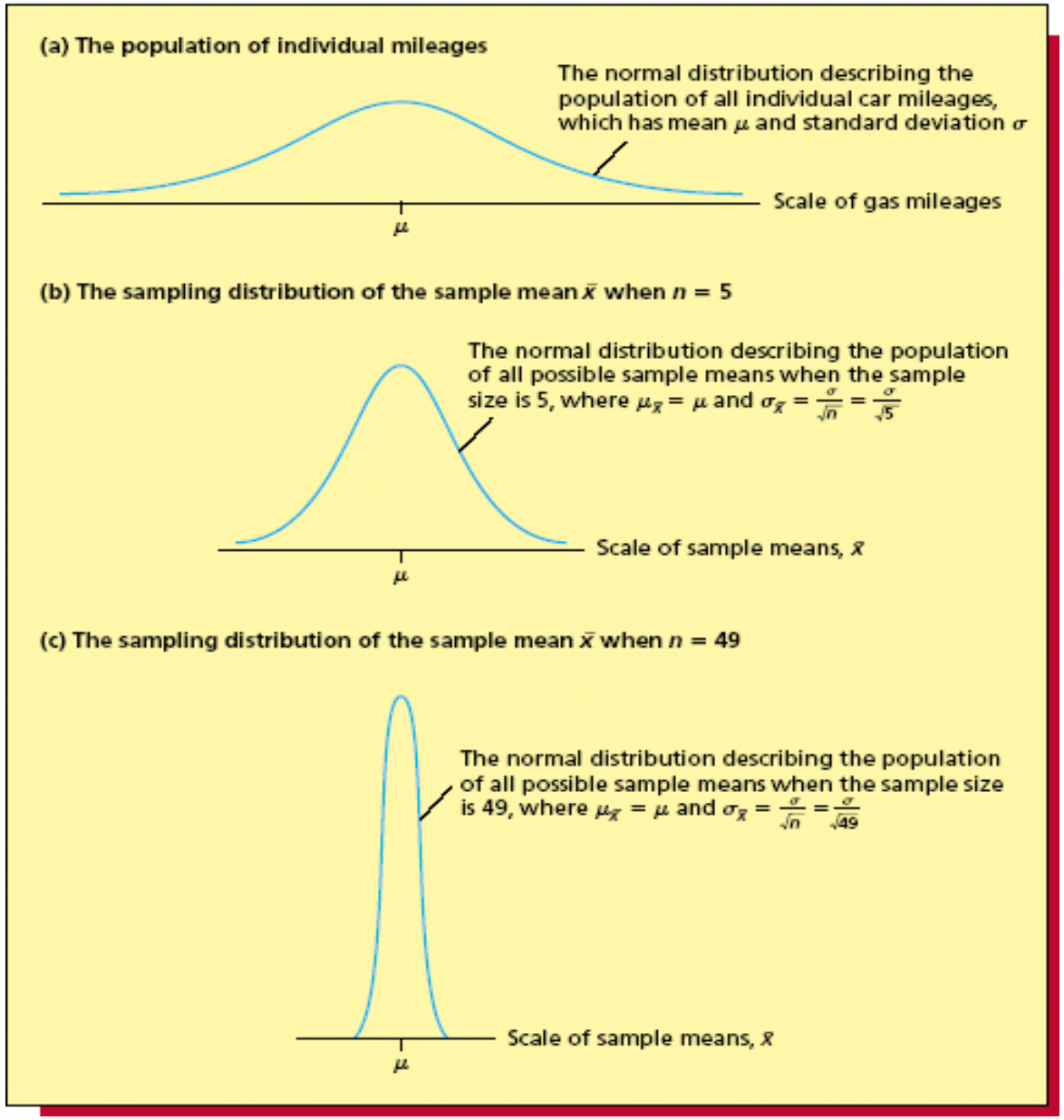
Example 7.2

Car Mileage Case

- Population of all midsize cars of a particular make and model
  - Population is normal with mean μ and standard deviation σ
  - Draw all possible samples of size n
  - Then the sampling distribution of the sample mean is normal with mean $\mu_{\bar{x}} = \mu$ and standard deviation

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

  - In particular, draw samples of size:
    - $n = 5$
    - $n = 49$



$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{5}} = \frac{\sigma}{2.2361}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{49}} = \frac{\sigma}{7}$$

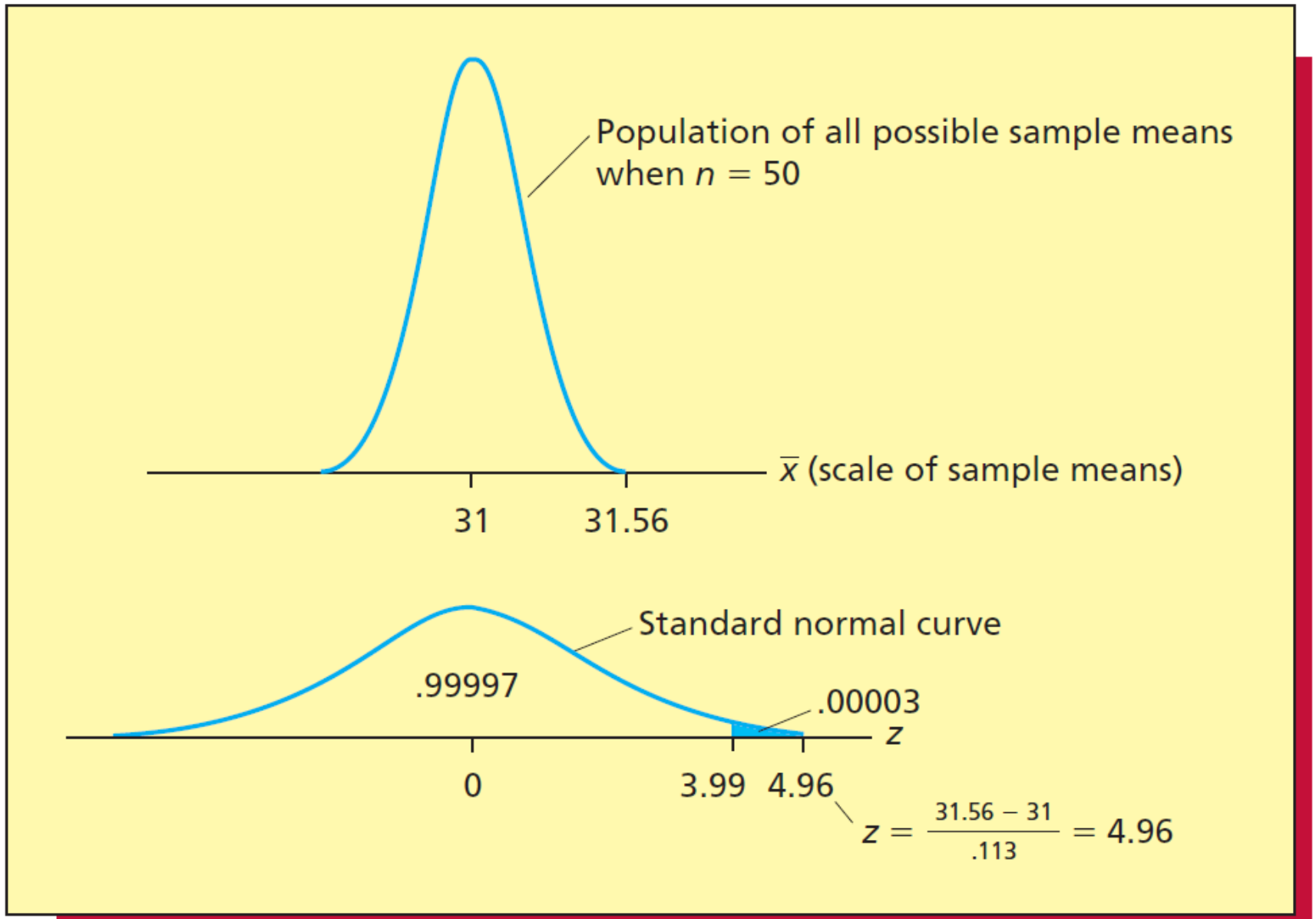
So, all possible sample means for $n=49$ will be more closely clustered around μ than the case of $n=5$

Part 2: Statistical inference Recall from Chapter 3 that the automaker has randomly selected a sample of $n = 50$ mileages, which has mean $\bar{x} = 31.56$. We now ask the following question: If the population mean mileage μ exactly equals 31 mpg (the minimum standard for the tax credit), what is the probability of observing a sample mean mileage that is greater than or equal to 31.56 mpg? To find this probability, recall from Chapter 2 that a histogram of the 50 mileages indicates that the population of all individual mileages is normally distributed. Assuming that the population standard deviation σ is known to equal .8 mpg, it follows that the sampling distribution of the sample mean $\bar{x}$ is a normal distribution, with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n} = .8/\sqrt{50} = .113$. Therefore,

$$P(\bar{x} \geq 31.56 \quad \text{if} \quad \mu = 31) = P\left(z \geq \frac{31.56 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right) = P\left(z \geq \frac{31.56 - 31}{.113}\right)$$

$$= P(z \geq 4.96)$$

FIGURE 7.4 The Probability That $\bar{x} \geq 31.56$ When $\mu = 31$ in the Car Mileage Case



To find $P(z \geq 4.96)$, notice that the largest z value given in Table A.3 (page 791) is 3.99, which gives a right-hand tail area of .00003. Therefore, because $P(z \geq 3.99) = .00003$, it follows that $P(z \geq 4.96)$ is less than .00003 (see Figure 7.4). The fact that this probability is less than .00003 says that, if μ equals 31, then fewer than 3 in 100,000 of all possible sample means are at least as large as the sample mean $\bar{x} = 31.56$ that we have actually observed. Therefore, if we are to believe that μ equals 31, then we must believe that we have observed a sample mean that can be described as a smaller than 3 in 100,000 chance. Because it is extremely difficult to believe that such a small chance would occur, we have extremely strong evidence that μ does not equal 31 and that μ is, in fact, larger than 31. This evidence would probably convince the federal government that the midsize model's mean mileage μ exceeds 31 mpg and thus that the midsize model deserves the tax credit.

To conclude this subsection, it is important to make two comments. First, the formula $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ follows, in theory, from the formula for $\sigma^2_{\bar{x}}$, the variance of the population of all possible sample means. The formula for $\sigma^2_{\bar{x}}$ is $\sigma^2_{\bar{x}} = \sigma^2/n$. Second, in addition to holding exactly if the sampled population is infinite, **the formula $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ holds approximately if the sampled population is finite and much larger than (say, at least 20 times) the size of the sample.** For example, if we define the population of the mileages of all new midsize cars to be the population of the mileages of all cars that will actually be produced this year, then the population is

Central Limit Theorem#1

- If the population is non-normal, what is the shape of the sampling distribution of the sample means?

- In fact the sampling distribution is approximately normal if the sample size is large enough, even if the population is non-normal

  by the "Central Limit Theorem"

- No matter what is the probability distribution that describes the population, if the sample size *n* is large enough, then the population of all possible sample means is ***approximately*** normal with mean

$$\mu_{\bar{x}} = \mu$$

and standard deviation

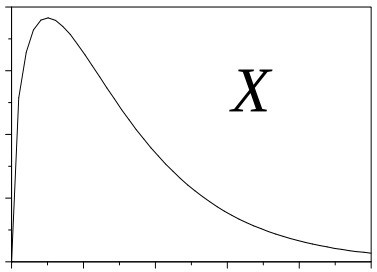
$$\sigma_{\bar{x}} = \sigma/\sqrt{n}$$

- Further, the larger the sample size *n*, the closer the sampling distribution of the sample means is to being normal
  - In other words, the larger *n*, the better the approximation

The Central Limit Theorem

If the sample size n is sufficiently large, then the population of all possible sample means is approximately normally distributed (with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$), no matter what probability distribution describes the sampled population. Furthermore, the larger the sample size n is, the more nearly normally distributed is the population of all possible sample means.

Random Sample (x_1, x_2, …, x_n)

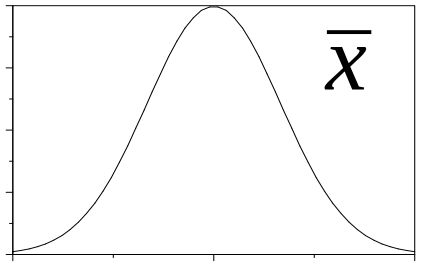


Population Distribution

(μ, σ)

(right-skewed)

as $n \rightarrow$ large

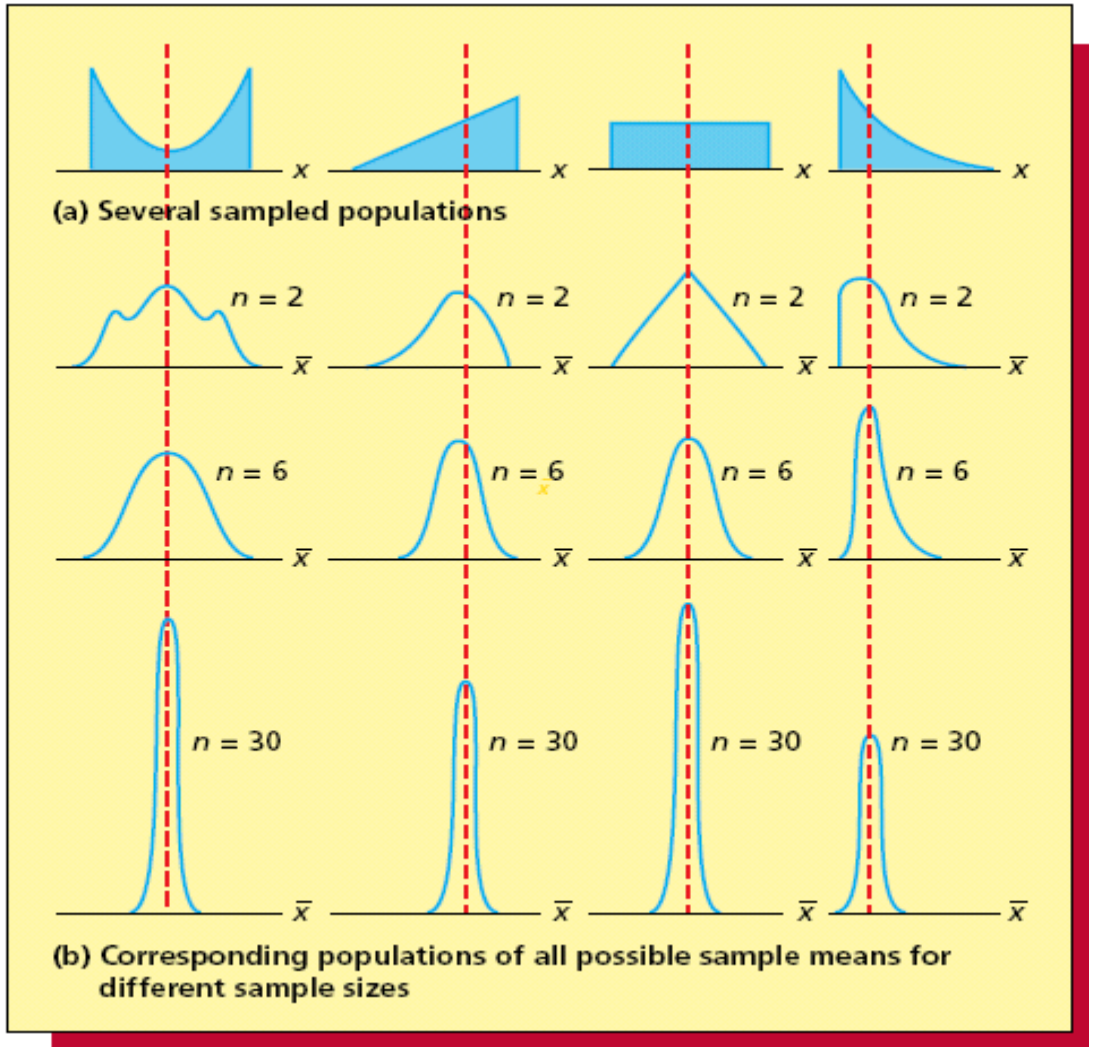


Sampling Distribution of Sample Means

$\left(\mu_{\bar{x}} = \mu, \sigma_{\bar{x}} = \sigma/\sqrt{n}\right)$

(nearly normal)

Effect of the Sample Size



The larger the sample size, the more nearly normally distributed is the population of all possible sample means

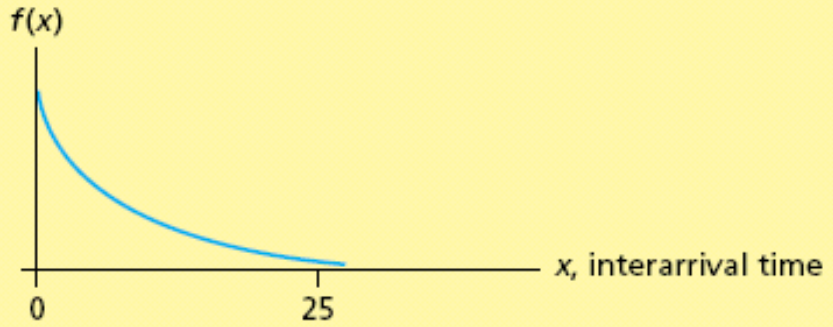
Also, as the sample size increases, the spread of the sampling distribution decreases

How Large?

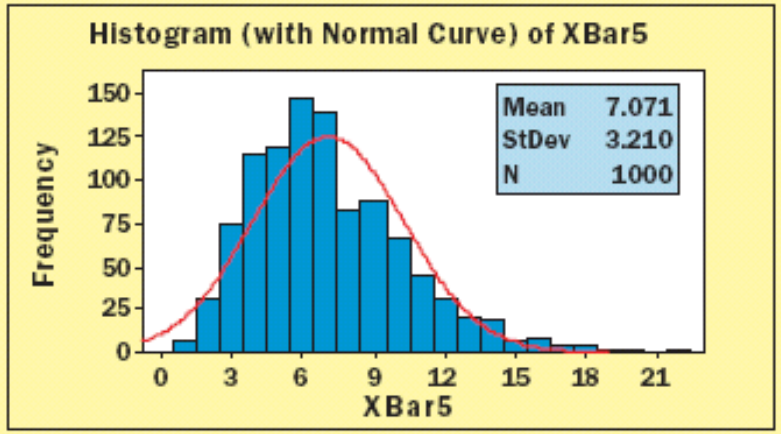
- How large is “large enough?”
- If the sample size n is at least 30, then for most sampled populations, the sampling distribution of sample means is approximately normal

  Refer to Figure 7.6 on next slide
  - Shown in Fig 7.6(a) is an exponential (right skewed) distribution
  - In Figure 7.6(b), 1,000 samples of size $n = 5$
    - Slightly skewed to right
  - In Figure 7.6(c), 1,000 samples with $n = 30$
    - Approximately bell-shaped and normal
- If the population is normal, the sampling distribution of $\bar{x}$ is normal regardless of the sample size

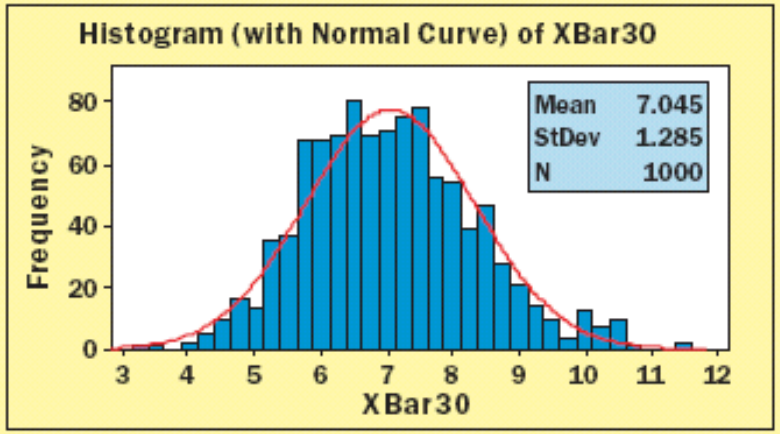
Example: Central Limit Theorem Simulation



(a) The exponential distribution describing the emergency room interarrival times



(b) A histogram of 1,000 sample means based on samples of size 5



(c) A histogram of 1,000 sample means based on samples of size 30

Example:

The foreman of a bottling plant has observed that the amount of soda in each “32-ounce” bottle is actually a normally distributed random variable, with a mean of 32.2 ounces and a standard deviation of .3 ounce.

If a customer buys a carton of four bottles, what is the probability that the *mean amount of the four bottles* will be greater than 32 ounces?

Solution:

Solution:

We want to find P($\bar{X}$ > 32), where X is normally distributed with μ =32.2 and σ =.3

X is normally distributed, therefore so will $\bar{X}$

$\mu_{\bar{X}} = \mu = 32.2$ oz.

$\sigma_{\bar{X}} = \sigma/\sqrt{n} = .3/\sqrt{4} = .15$

$$P(\bar{X} > 32) = P\left(\frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} > \frac{32 - 32.2}{.15}\right) = P(Z > -1.33) = .9082$$

"There is about a 91% chance the mean of the four bottles will exceed 32oz."

Example 2

A light bulb manufacturer claims that the lifespan of its light bulbs has a mean of 54 months and a standard deviation of 6 months. Your consumer advocacy group tests 50 of them. Assuming the manufacturer's claims are true, what is the probability that it finds a mean lifetime of 52 months or less?

Solution: $\bar{x}$ is approximately normally distributed.

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

We are seeking P($\bar{x}$ ≤52). Convert to Z-scores (Standard normal distribution)

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{52 - 54}{0.85} = -2.35$$

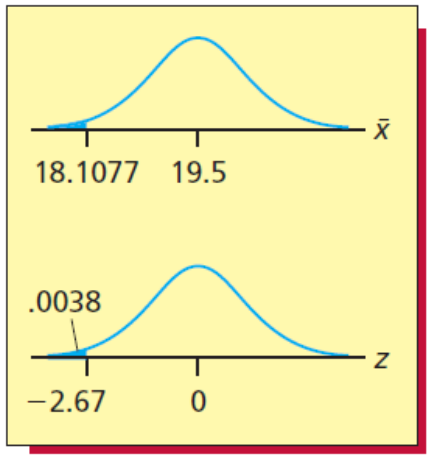
P(Z≤-2.35)=0.0094.

Thus, the probability of 52 months or less is 0.0094, or 0.94%.

Example 7.3

EXAMPLE 7.4 The e-billing Case: Reducing Mean Bill Payment Time

Recall that a management consulting firm has installed a new computer-based electronic billing system in a Hamilton, Ohio, trucking company. Because of the previously discussed advantages of the new billing system, and because the trucking company's clients are receptive to using this system, the management consulting firm believes that the new system will reduce the mean bill payment time by more than 50 percent. The mean payment time using the old billing system was approximately equal to, but no less than, 39 days. Therefore, if μ denotes the new mean payment time, the consulting firm believes that μ will be less than 19.5 days. To assess whether μ is less than 19.5 days, the consulting firm has randomly selected a sample of $n = 65$ invoices processed using the new billing system and has determined the payment times for these invoices. The mean of the 65 payment times is $\bar{x} = 18.1077$ days, which is less than 19.5 days. Therefore, we ask the following question: If the population mean payment time is 19.5 days, what is the probability of observing a sample mean payment time that is less than or equal to 18.1077 days? To find this



viation σ of payment times is the same for different applications and equals 4.2 days. Assuming that σ also equals 4.2 days for the trucking company, it follows that $\sigma_{\bar{x}}$ equals $4.2/\sqrt{65} = .5209$ and that

$$P(\bar{x} \leq 18.1077 \text{ if } \mu = 19.5) = P\left(z \leq \frac{18.1077 - 19.5}{.5209}\right) = P(z \leq -2.67)$$

which is the area under the standard normal curve to the left of −2.67. The normal table tells us that this area equals .0038. This probability says that, if μ equals 19.5, then only .0038 of all possible sample means are at least as small as the sample mean $\bar{x} = 18.1077$ that we have actually observed. Therefore, if we are to believe that μ equals 19.5, we must believe that we have observed a sample mean that can be described as a 38 in 10,000 chance. It is very difficult to believe that such a small chance would occur, so we have very strong evidence that μ does not equal 19.5 and is, in fact, less than 19.5. We conclude that the new billing system has reduced the mean bill payment time by more than 50 percent.

BI

Unbiasedness

- $\bar{x}$
- s^2

> The **sampling distribution of a sample statistic** is the probability distribution of the population of all possible values of the sample statistic.

> A sample statistic is an **unbiased point estimate** of a population parameter if the mean of the population of all possible values of the sample statistic equals the population parameter.

Unbiased Estimates

- A sample statistic is an ***unbiased*** point estimate of a population parameter if the mean of all possible values of the sample statistic equals the population parameter
- $\bar{x}$ is an unbiased estimate of μ because $\mu_{\bar{x}} = \mu$
  - In general, the sample mean is always an unbiased estimate of μ
  - The sample median is often an unbiased estimate of μ
    - But not always

- The sample variance s^2 is an unbiased estimate of σ^2 if the sampled population is infinite
  - That is why s^2 has a divisor of $n-1$

(if we used n as the divisor when estimating σ^2 , we would not obtain an unbiased estimate)

However, s is not an unbiased estimate of σ

- Even so, since there is no easy way to calculate an unbiased point estimate of σ, the usual practice is to use s as an estimate of σ

Minimum-variance estimates

median are unbiased point estimates of μ. In fact, there are many unbiased point estimates of μ. However, it can be shown that the variance of the population of all possible sample means is smaller than the variance of the population of all possible values of any other unbiased point estimate of μ. For this reason, **we call the sample mean a minimum-variance unbiased point estimate of μ.** When we use the sample mean as the point estimate of μ, we are more likely to obtain a point estimate close to μ than if we used any other unbiased sample statistic as the point estimate of μ. This is one reason why we use the sample mean as the point estimate of the population mean.

Technical Note: If we randomly select a sample of size n without replacement from a finite population of size N, then it can be shown that $\sigma_{\bar{x}} = (\sigma/\sqrt{n})\sqrt{(N-n)/(N-1)}$, where the quantity $\sqrt{(N-n)/(N-1)}$ is called the **finite population multiplier.** If the size of the sampled population is at least 20 times the size of the sample (that is, **if $N \geq 20n$**), **then the finite population multiplier is approximately equal to one, and $\sigma_{\bar{x}}$ approximately equals $\sigma/\sqrt{n}$.** However, if the population size N is smaller than 20 times the size of the sample, then the finite population multiplier is substantially less than one, and we must include this multiplier in the calculation of $\sigma_{\bar{x}}$. For instance, in Example 7.2, where the standard deviation σ of the population of $N = 6$ car mileages can be calculated to be 1.7078, and where $N = 6$ is only three times the sample size $n = 2$, it follows that

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \left(\frac{1.7078}{\sqrt{2}}\right)\sqrt{\frac{6-2}{6-1}} = 1.2076(.8944) = 1.08$$

We will see how this formula can be used to make statistical inferences in Section 8.5.

7.2 The Sampling Distribution of the Sample Proportion

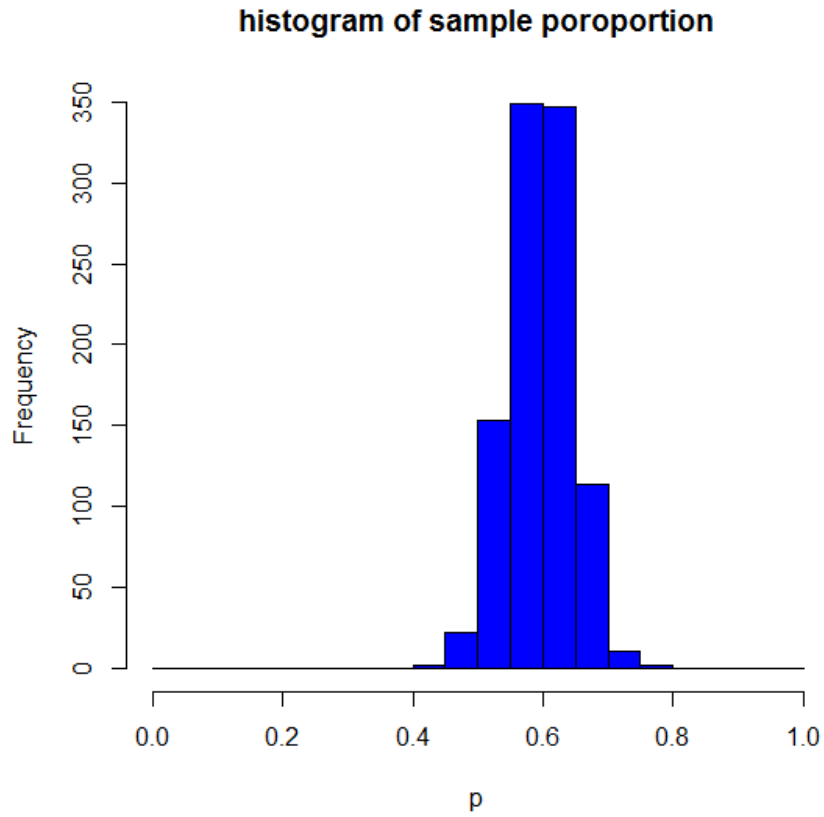
- For a population of units, we select samples of size n, and calculate its proportion $\hat{p}$ for the units of the sample to be fall into a particular category.

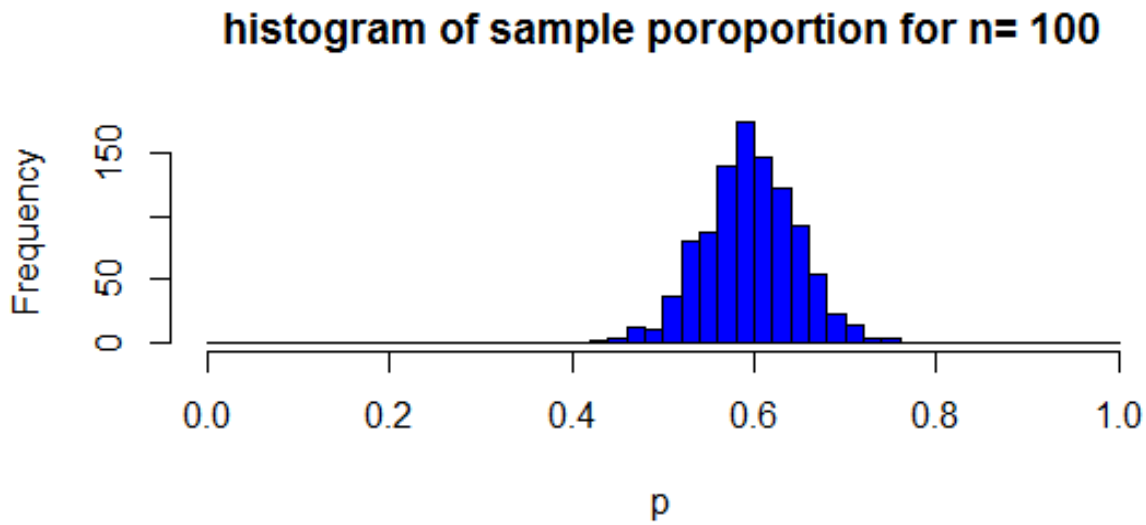
- $\hat{p}$ is a random variable and has its probability distribution.
- The probability distribution of all possible sample proportions is called the sampling distribution of the sample proportion

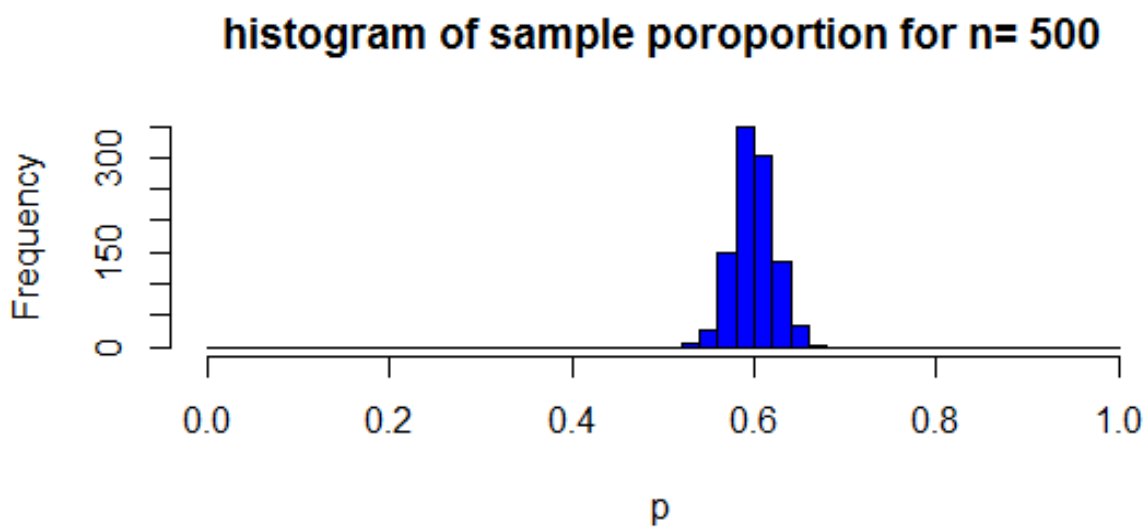
There are 200000 eligible voters in York County, South Carolina.
Among all voters, 120000 voters all voters preferred Candidate ***A***. So the proportion of voters preferred Candidate ***A*** is $\boldsymbol{p}$=0.6

Survey: Select 100 York County voters and return the proportion $\hat{p}$ of voters preferred Candidate ***A***

Conduct the *survey 1000 times, the distribution of sampling proportion looks like:*




histogram of sample poroportion for n= 100
Frequency
0
50
150
0.0
0.2
0.4
0.6
0.8
1.0
p


histogram of sample poroportion for n= 500
Frequency
0
150
300
0.0
0.2
0.4
0.6
0.8
1.0
p

The sampling distribution of $\hat{p}$ is

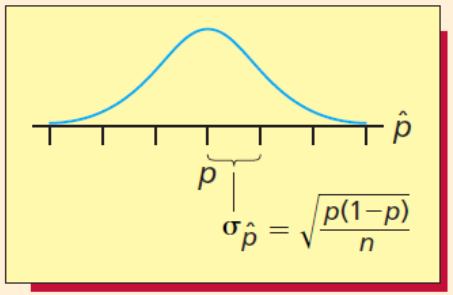
- approximately normal, if *n* is large (meet the conditions that $np \geq 5$ and $n(1-p) \geq 5$)
- has mean $\mu_{\hat{p}} = p$
- has standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

where *p* is the population proportion for the category

The Sampling Distribution of the Sample Proportion $\hat{p}$

The population of all possible sample proportions

1. Approximately has a normal distribution, if the sample size n is large.
2. Has mean $\mu_{\hat{p}} = p$.
3. Has standard deviation $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$.



Stated equivalently, the sampling distribution of $\hat{p}$ has mean $\mu_{\hat{p}} = p$, has standard deviation $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$, and is approximately a normal distribution (if the sample size n is large).

Property 1 in the box says that, if n is large, then the population of all possible sample proportions approximately has a normal distribution. Here, it can be shown that ***n* should be considered large if both *np* and *n*(1 − *p*) are at least 5.**[3] Property 2, which says that $\mu_{\hat{p}} = p$, is valid for any sample size and tells us that $\hat{p}$ is an unbiased estimate of p. That is, although the sample proportion $\hat{p}$ that we calculate probably does not equal p, the average of all the different sample proportions that we could have calculated (from all the different possible samples) is equal to p. Property 3, which says that

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

[3]Some statisticians suggest using the more conservative rule that both np and $n(1-p)$ must be at least 10.

Example 7.4

The Cheese Spread Case

- A food processing company developed a new cheese spread spout which may save production cost. If only less than *10%* of current purchasers do not accept the design, the company would adopt and use the new spout.
- *1000* current purchasers are randomly selected and inquired, and *63* of them say they would stop buying the cheese spread if the new spout were used. So, the sample proportion $\hat{p} = 0.063$.
- if $p=0.1$, what is the probability of observing a sample of size *1000* with sample proportion $\hat{p} \leq 0.063$?

- If p=0.10, since n=1000, $np \geq 5$ and $n(1-p) \geq 5$, $\hat{p}$ is approximately normal with

$$\mu_{\hat{p}} = p = 0.1,$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = 0.094868,$$

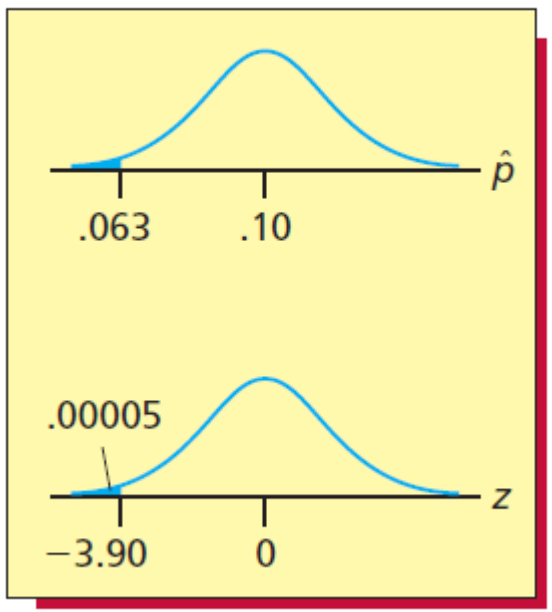
$$P\left(\hat{p} \leq 0.063 \text{ if } p = 0.1\right) = P\left(z \leq \frac{0.063 - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right)$$

$$= P\left(z \leq \frac{0.063 - 0.10}{0.094868}\right)$$

$$= P(z \leq -3.90) \leq 0.001.$$

- So, if $p=0.1$, the chance of observing at most 63 out of 1000 randomly selected customers do not accept the new design is less than 0.001
- But such observation does occur. This means that we have extremely strong evidence that $p \neq 0.1$, and p is in fact less than 0.1
- Therefore, the company can adopt the new design

have actually observed. That is, if we are to believe that p equals .10, we must believe that we have observed a sample proportion that can be described as a 5 in 100,000 chance. It follows that we have extremely strong evidence that p does not equal .10 and is, in fact, less than .10. In other words, we have extremely strong evidence that fewer than 10 percent of current purchasers would stop buying the cheese spread if the new spout were used. It seems that introducing the new spout will be profitable.



Sample Mean $\bar{x}$ (Central Limit Theorem)

- Sample mean is approximately ***normally*** distributed if:
  - Sample size is ***larger than 30*** (without assuming the population also has a normal distribution);
  - Sample size is ***less than 30***, and the *population* also has a *normal distribution*.

When sample size is large:

$$\bar{x} \sim N(\mu, \frac{\sigma^2}{n}) \text{ approximately} \qquad \mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$\bar{x}$ – Sample mean μ– Population mean

$\mu_{\bar{x}}$ – Mean of sample mean σ– Population standard deviation

$\sigma_{\bar{x}}$ – Standard deviation of sample mean n- sample size

Sample Proportion

the sampling distribution of $\hat{p}$ is

- approximately normal, if n is large (meet the conditions that $np \geq 5$ and $n(1-p) \geq 5$)
- has mean $\mu_{\hat{p}} = p$
- has standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

where p is the population proportion for the category

Chapter Summary

We began this chapter by defining a random sample and by explaining how to use a **random number table** or **computer-generated random numbers** to select a **random sample.** We then discussed **sampling distributions.** A **sampling distribution** is the probability distribution that describes the population of all possible values of a sample statistic. In this chapter we studied the properties of two important sampling distributions—the sampling distribution of the sample mean, $\bar{x}$, and the sampling distribution of the sample proportion, $\hat{p}$.

Because different samples that can be randomly selected from a population give different sample means, there is a population of sample means corresponding to a particular sample size. The probability distribution describing the population of all possible sample means is called the **sampling distribution of the sample mean, $\bar{x}$.** We studied the properties of this sampling distribution when the sampled population is and is not normally distributed. We found that, when the sampled population has a normal distribution, then the sampling distribution of the sample mean is a normal distribution. Furthermore, the **Central Limit Theorem** tells us that, if the sampled population is not normally distributed, then the sampling distribution of the sample mean is approximately a normal distribution when the sample size is large (at least 30). We also saw that the mean of the sampling distribution of $\bar{x}$ always equals the mean of the sampled population, and we presented formulas for the variance and the standard deviation of this sampling distribution. Finally, we explained that the sample mean is a **minimum-variance unbiased point estimate** of the mean of a normally distributed population.

We also studied the properties of the **sampling distribution of the sample proportion $\hat{p}$.** We found that, if the sample size is large, then this sampling distribution is approximately a normal distribution, and we gave a rule for determining whether the sample size is large. We found that the mean of the sampling distribution of $\hat{p}$ is the population proportion p, and we gave formulas for the variance and the standard deviation of this sampling distribution.

Thank you!