Section 7.2
Sample Proportions

Learning Objectives

After this section, you should be able to…

✓ FIND the mean and standard deviation of the sampling distribution of a sample proportion

✓ DETERMINE whether or not it is appropriate to use the Normal approximation to calculate probabilities involving the sample proportion

✓ CALCULATE probabilities involving the sample proportion

✓ EVALUATE a claim about a population proportion using the sampling distribution of the sample proportion
The Sampling Distribution for the Statistic $\hat{p}$

Consider the approximate sampling distributions generated by a simulation in which SRSs of Reese’s Pieces are drawn from a population whose proportion of orange candies is 0.15.

What happens to $\hat{p}$ as the sample size increases from 25 to 50?

What do you notice about the shape, center, and spread?

How good is the statistic $\hat{p}$ as an estimate of the parameter $p$?

The sampling distribution of $\hat{p}$ answers this question.
The Sampling Distribution for the Statistic \( \hat{p} \)

You should have noticed the sampling distribution has the following characteristics for shape, center, and spread:

**Shape**: In some cases, the sampling distribution of \( \hat{p} \) can be approximated by a Normal curve. This seems to depend on both the sample size \( n \) and the population proportion \( p \).

**Center**: The mean of the distribution is \( \mu_{\hat{p}} = p \). This makes sense because the sample proportion \( \hat{p} \) is an unbiased estimator of \( p \).

**Spread**: For a specific value of \( p \), the standard deviation \( \sigma_{\hat{p}} \) gets smaller as \( n \) gets larger. The value of \( \sigma_{\hat{p}} \) depends on both \( n \) and \( p \).
The Connection between THE STATISTIC \( \hat{p} \) and a random variable \( X \)

There is an important connection between the sample proportion \( \hat{p} \) and the number of "successes" for the random variable \( X \) in the sample.

\[
\hat{p} = \frac{\text{count of successes in sample}}{\text{size of sample}} = \frac{X}{n}
\]

**REMEMBER:** for a binomial random variable \( X \), the mean and standard deviation are:

\[
\mu_X = np \quad \sigma_X = \sqrt{np(1-p)}
\]

Since \( \hat{p} = X / n \)  

**THEN**  \( \hat{p} = (1/n) \cdot X \)

we are just multiplying the random variable \( X \) by a constant \((1/n)\) to get the random variable \( \hat{p} \).

Now we can use algebra to calculate \( \mu_{\hat{p}} \) and \( \sigma_{\hat{p}} \).
The Connection between THE STATISTIC $\hat{p}$ and a random variable $X$

Binomial random variable $X$ are:

$$
\mu_X = np \\
\sigma_X = \sqrt{np(1 - p)}
$$

Since $\hat{p} = X/n$ then $\hat{p} = (1/n) \cdot X$

Therefore...

$$\mu_{\hat{p}} = \frac{1}{n}(np) = p$$

$\hat{p}$ is an unbiased estimator for $p$

$$\sigma_{\hat{p}} = \frac{1}{n} \sqrt{np(1 - p)} = \sqrt{\frac{np(1 - p)}{n^2}} = \sqrt{\frac{p(1 - p)}{n}}$$

As sample size increases, the spread decreases.
Using the Normal Approximation for $\hat{p}$

Inference about a population proportion $p$ is based on the sampling distribution of $\hat{p}$. When the sample size is large enough.

You must check the following 2 conditions have been met

\[ np \geq 10 \quad \text{and} \quad n(1 - p) \geq 10 \]

Then the sampling distribution of $\hat{p}$ is approximately Normal.
We can summarize the facts about the sampling distribution of \( \hat{p} \) as follows:

### Sampling Distribution of a Sample Proportion

Choose an SRS of size \( n \) from a population of size \( N \) with proportion \( p \) of successes. Let \( \hat{p} \) be the sample proportion of successes. Then:

The **mean** of the sampling distribution of \( \hat{p} \) is \( \mu_{\hat{p}} = p \)

The **standard deviation** of the sampling distribution of \( \hat{p} \) is

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

as long as the 10\% condition is satisfied: \( n \leq (1/10)N \).

As \( n \) increases, the sampling distribution becomes **approximately Normal**. Before you perform Normal calculations, check that the Normal condition is satisfied: \( np \geq 10 \) and \( n(1-p) \geq 10 \).
Example 1:

CHECK YOUR UNDERSTANDING
About 75% of young adult Internet users (ages 18 to 29) watch online video. Suppose that a sample survey contacts an SRS of 1000 young adult Internet users and calculates the proportion \( \hat{p} \) in this sample who watch online video.

1. What is the mean of the sampling distribution of \( \hat{p} \)? Explain.
2. Find the standard deviation of the sampling distribution of \( \hat{p} \). Check that the 10% condition is met.
3. Is the sampling distribution of \( \hat{p} \) approximately Normal? Check that the Normal condition is met.
4. If the sample size were 9000 rather than 1000, how would this change the sampling distribution of \( \hat{p} \)?

See next slide for worked out solution
**Example 1:**

**CHECK YOUR UNDERSTANDING**

About 75% of young adult Internet users (ages 18 to 29) watch online video. Suppose that a sample survey contacts an SRS of 1000 young adult Internet users and calculates the proportion \( \hat{p} \) in this sample who watch online video.

1. What is the mean of the sampling distribution of \( \hat{p} \)? Explain.
2. Find the standard deviation of the sampling distribution of \( \hat{p} \). Check that the 10% condition is met.
3. Is the sampling distribution of \( \hat{p} \) approximately Normal? Check that the Normal condition is met.
4. If the sample size was increased, would the standard deviation of \( \hat{p} \) decrease? Explain.

**a) Given information:** \( p = 0.75 \) (the population parameter for a proportion)

1. The mean of the sampling distribution \( \mu_{\hat{p}} \) is the same as the population proportion \( \mu_{\hat{p}} = p \).

2. 10% Condition: SRS = 1000 AND IT IS FAIR TO ASSUME THE POPULATION IS OVER 10,000 young adults.

\[
\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.75(0.25)}{1000}} = 0.0137
\]

3. The sampling distribution is approximately normal because Normal conditions met: \( np = 1000(0.75) = 750 > 10 \checkmark \)

\( nq = 1000(0.25) = 250 > 10 \checkmark \)

4. SRS \( n = 9000 \)

\[
\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.75(0.25)}{9000}} = 0.0046 \text{ (NOTICE IT DECREASES)}
\]
Example 2:

A polling organization asks an SRS of 1500 first-year college students how far away their home is. Suppose that 35% of all first-year students actually attend college within 50 miles of home. What is the probability that the random sample of 1500 students will give a result within 2 percentage points of this true value?

So what are they asking?
Draw a picture!
Example 2:

A polling organization asks an SRS of 1500 first-year college students how far away their home is. Suppose that 35% of all first-year students actually attend college within 50 miles of home. What is the probability that the random sample of 1500 students will give a result within 2 percentage points of this true value?

**STATE:** We want to find the probability that the sample proportion falls between 0.33 and 0.37 (within 2 percentage points, or 0.02, of 0.35).

**PLAN:** We have an SRS of size \( n = 1500 \) drawn from a population in which the proportion \( p = 0.35 \) attend college within 50 miles of home.

Keep Going!
**Example 2 (Cont):**

Since we know \( p = 0.35 \) and \( n = 1500 \) then we can find the mean and standard deviation:

\[
\begin{align*}
\mu_{\hat{p}} &= 0.35 \\
\sigma_{\hat{p}} &= \sqrt{\frac{(0.35)(0.65)}{1500}} = 0.0123
\end{align*}
\]

Can we use the normal model?

• Since \( np = 1500(0.35) = 525 \) and \( n(1 - p) = 1500(0.65) = 975 \)
  • And both are both greater than 10, we can use the normal model.

• Next standardize to find the desired probability.

\[
\begin{align*}
z &= \frac{0.33 - 0.35}{0.0123} = -1.63 \\
z &= \frac{0.37 - 0.35}{0.0123} = 1.63
\end{align*}
\]

\[
P(0.33 \leq \hat{p} \leq 0.37) = P(-1.63 \leq Z \leq 1.63) = 0.9484 - 0.0516 = 0.8968
\]

**CONCLUDE:** About 90% of all SRSs of size 1500 will give a result within 2 percentage points of the truth about the population.
Example 3: The Superintendent of a large school wants to know the proportion of high school students in her district are planning to attend a four-year college or university. Suppose that 80% of all high school students in her district are planning to attend a four-year college or university. What is the probability that an SRS of size 125 will give a result within 7 percentage points of the true value?
Example 3: The Superintendent of a large school wants to know the proportion of high school students in her district are planning to attend a four-year college or university. Suppose that 80% of all high school students in her district are planning to attend a four-year college or university. What is the probability that an SRS of size 125 will give a result within 7 percentage points of the true value?

\( \hat{p} = 0.8 \) = proportion of HS students planning to attend 4-yr college

SR55, n = 125

Find probability \( \hat{p} = 0.8 \pm 7\% \rightarrow P(0.73 \leq \hat{p} \leq 0.87) \)

Check Conditions

1. 10% Condition - Is the school district large enough?
   \( n = 125 \times 0.1 = 12.5 \) (we assume the school has 1,250 HS students which seems reasonable for a large school)

2. Normal Condition - np = 125(0.8) = 100 > 10, ng = 125(0.2) = 25 > 10

   We can use the normal approximation

3. Find mean and std dev:

   \[ \mu_{\hat{p}} = 0.8, \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.8(0.2)}{125}} = 0.036 \]

4. State model \( \hat{p} \sim N(0.8, 0.036) \)

5. Find probability by using Z-scores

\[ P(0.73 \leq \hat{p} \leq 0.87) = P(-1.94 \leq Z \leq 1.94) \]

\[ Z = \frac{0.73 - 0.8}{0.036} = -1.94 \]

\[ Z = \frac{0.87 - 0.8}{0.036} = 1.94 \]

What would you guess the probability?
Since the Z scores are ±1.94 (about 2 std deviations), remember the 68-95-99.7 rule!

The probability should be around 95%.

Find $P(-1.94 \leq \hat{p} \leq 1.94) = 0.9476$

$N(0,1) \rightarrow \text{normal cdf}(\hat{p} = (-1.94, 1.94, 0, 1))$

Conclude: About 95% of all SESSs of size 125 will give a sample proportion within 7 points of the true population proportion of high school students who are planning to attend a 4 year college or university.
Sample Proportions

Summary

In this section, we learned that...

- When we want information about the population proportion \( p \) of successes, we often take an SRS and use the sample proportion \( \hat{p} \) to estimate the unknown parameter \( p \). The sampling distribution of \( \hat{p} \) describes how the statistic varies in all possible samples from the population.

- The mean of the sampling distribution of \( \hat{p} \) is equal to the population proportion \( p \). That is, \( \hat{p} \) is an unbiased estimator of \( p \).

- The standard deviation of the sampling distribution of \( \hat{p} \) is \( \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \) for an SRS of size \( n \). This formula can be used if the population is at least 10 times as large as the sample (the 10% condition). The standard deviation of \( \hat{p} \) gets smaller as the sample size \( n \) gets larger.

- When the sample size \( n \) is larger, the sampling distribution of \( \hat{p} \) is close to a Normal distribution with mean \( p \) and standard deviation \( \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \).

- In practice, use this Normal approximation when both \( np \geq 10 \) and \( n(1-p) \geq 10 \) (the Normal condition).