Objectives:
• Describe the shape, center, and spread of the sampling distribution of $\hat{p}_1 - \hat{p}_2$.
• Determine whether the conditions are met for doing inference about $p_1 - p_2$.
• Construct and interpret a confidence interval to compare two proportions.
• Perform a significance test to compare two proportions.

Sampling Distribution of a Difference between Two Proportions
Suppose we want to compare the proportions of individuals with a certain characteristic in Population 1 and Population 2. Let’s call these parameters of interest $p_1$ and $p_2$. The ideal strategy is to take a separate random sample from each population and to compare the sample proportions with that characteristic.

<table>
<thead>
<tr>
<th>Population or treatment</th>
<th>Parameter</th>
<th>Statistic</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$p_1$</td>
<td>$\hat{p}_1$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>2</td>
<td>$p_2$</td>
<td>$\hat{p}_2$</td>
<td>$n_2$</td>
</tr>
</tbody>
</table>

When the Population Parameter is known (Rare)
Suppose there are two large high schools, each with over 2000 students. At School 1, 70% of students did their homework last night. At School 2, 50% of students did their homework last night. Suppose the counselor at School 1 takes an SRS of 100 students and records the sample proportion that did their homework. School 2’s counselor takes an SRS of 200 students and records the sample proportion that did their homework.

Simulation of Sampling Distributions
(a) Approximate sampling distribution of $\hat{p}_1$
(b) Approximate sampling distribution of $\hat{p}_2$
(c) Approximate sampling distribution of $\hat{p}_1 - \hat{p}_2$

The Sampling Distribution of the Difference between Sample Proportions
Choose an SRS of size $n_2$ from Population 1 with proportion of successes $p_1$ and an independent SRS of size $n_2$ from Population 2 with proportion of successes $p_2$.

**Shape** When $n_1 p_1$, $n_1 (1 - p_1)$, $n_2 p_2$, and $n_2 (1 - p_2)$ are all at least 10, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately Normal.

**Center** The mean of the sampling distribution is $p_1 - p_2$.

**Spread** The standard deviation of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is

$$\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

as long as each sample is no more than 10% of its population (10% condition).
Comparing Two Proportions

Example Problem When Population Proportion is known (rare):

Your teacher brings two bags of colored goldfish crackers to class. Bag 1 has 25% red crackers and Bag 2 has 35% red crackers. Each bag contains more than 1000 crackers. Using a paper cup, your teacher takes an SRS of 50 crackers from Bag 1 and a separate SRS of 40 crackers from Bag 2. Let \( \hat{p}_1 - \hat{p}_2 \) be the difference in the sample proportions of red crackers.

a) What is the shape of the sampling distribution of \( \hat{p}_1 - \hat{p}_2 \)? Why?

\[
\hat{p}_1 = \frac{50(0.25)}{50} = 0.25 \\
\hat{p}_2 = \frac{40(0.35)}{40} = 0.35
\]

b) Find the mean of the sampling distribution. Show your work.

\[
\mu_{\hat{p}_1 - \hat{p}_2} = \hat{p}_1 - \hat{p}_2 = 0.25 - 0.35 = -0.10
\]

c) Find the standard deviation of the sampling distribution. Show your work.

\[
\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{(0.25)(0.75)}{50} + \frac{(0.35)(0.65)}{40}} = 0.0971
\]

Independent Practice When Population Proportion is known (rare):

A researcher reports that 80% of high school graduates, but only 40% of high school dropouts, would pass a basic literacy test. Assume that the researcher’s claim is true. Suppose we give a basic literacy test to a random sample of 60 high school graduates and a separate random sample of 75 high school dropouts. Let \( \hat{p}_G = \text{sample proportion of graduates and } \hat{p}_D = \text{sample proportion of dropouts who pass the test.} \)

a) What is the shape of the sampling distribution of \( \hat{p}_G - \hat{p}_D \)? Why?

Approximately Normal \( \frac{60}{60} \text{ all are } \geq 10 \)

\[
\hat{p}_G = \frac{60(0.8)}{60} = 0.8 \\
\hat{p}_D = \frac{75(0.4)}{75} = 0.4
\]

b) Find the mean of the sampling distribution. Show your work.

\[
\mu_{\hat{p}_G - \hat{p}_D} = 0.8 - 0.4 = 0.4
\]

c) Find the standard deviation of the sampling distribution. Show your work.

\[
\sigma_{\hat{p}_G - \hat{p}_D} = \sqrt{\frac{(0.8)(0.2)}{60} + \frac{(0.4)(0.6)}{75}} = 0.0766
\]
Confidence Intervals for $p_1 - p_2$:
When data come from two independent random samples or two groups in a randomized experiment (the Random condition), the statistic $\hat{p}_1 - \hat{p}_2$ is our best guess for the value of $p_1 - p_2$. We can use our familiar formula to calculate a confidence interval for $p_1 - p_2$:

$$\text{statistic} \pm (\text{critical value}) \cdot (\text{standard deviation of statistic})$$

Because we don’t know the values of the parameters $p_1$ and $p_2$, we replace them in the standard deviation formula with the sample proportions. The result is the standard Error

$$\text{standard error of the statistic } \hat{p}_1 - \hat{p}_2 : \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Two-Sample z Interval for a Difference between Two Proportions
When the conditions are met, an approximate C% confidence interval for $\hat{p}_1 - \hat{p}_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $z^*$ is the critical value for the standard Normal curve with C% of its area between $-z^*$ and $z^*$.

Example: Confidence Interval for a Difference between two proportions
As part of the Pew Internet and American Life Project, researchers conducted two surveys in 2012. The first survey asked a random sample of 799 U.S. teens about their use of social media and the Internet. A second survey posed similar questions to a random sample of 2253 U.S. adults. In these two studies, 80% of teens and 69% of adults used social-networking sites.

a) Calculate the standard error of the sampling distribution of the difference in the sample proportions (teens – adults). What information does this value provide?

$$\text{SE}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = 0.0172$$

If we took many SRS's of teens (n=799) and adults (n=2253), the difference in sample proportions who use social media will typically be 0.0172 from the true difference in proportions of all teens and adults who use social media.

b) Construct and interpret a 95% confidence interval for the difference between the proportion of all U.S. teens and adults who use social-networking sites.

State: $\hat{p}_1$: Proportion of teens on social media

$\hat{p}_2$: Proportion of adults on social media

Estimate $\hat{p}_1 - \hat{p}_2$ with 95% Confidence

Plan: z-Sample z Interval

Randomization Both are independent SRS's

90% Cond. $n_1 = 799$: at least 7999 teens in pop.

$n_2 = 2253$: at least 22533 adults in pop.

Large Counts $\hat{p}_1(1-\hat{p}_1) = 159 > 10$ $\hat{p}_2(1-\hat{p}_2) = 160 > 10$

$2253(0.69) = 1555 > 10$ $2253(0.31) = 698 > 10$

All are at least 10, it is safe to assume the sampling dist. of $\hat{p}_1 - \hat{p}_2$ is approx Normal

Do: $$(\hat{p}_1 - \hat{p}_2) \pm z^* \text{SE}_{\hat{p}_1 - \hat{p}_2}$$

$\text{SE}_{\hat{p}_1 - \hat{p}_2} = 0.0172$ $z^* = \pm 1.96 \text{ Norm}(0.05, 0.1) = 1.96$

$$(0.05 - 0.1) = 1.96 $$

$0.05 \pm 0.034$

$= (0.016, 0.084)$

Conclude: We are 95% confident the interval from 0.016 to 0.084 captures the true difference in proportions of all teens and adults who use social networking sites.
Comparing Two Proportions

Independent Practice: Confidence Interval for a Difference between two proportions

Many news organizations conduct polls asking adults in the United States if they approve of the job the president is doing. How did President Obama’s approval rating change from October 2012 to October 2013? According to a Gallup poll of 1500 randomly selected U.S. adults on October 2–4, 2012, 52% approved of Obama’s job performance. A Gallup poll of 1500 randomly selected U.S. adults on October 5–7, 2013, showed that 46% approved of Obama’s job performance.

a) Calculate the standard error of the sampling distribution of the difference in the sample proportions (2013 – 2012). Interpret this value.

\[
SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{(\hat{p}_1)(1-\hat{p}_1)}{n_1} + \frac{(\hat{p}_2)(1-\hat{p}_2)}{n_2}}
\]

\[
SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{(0.52)(0.48)}{1500} + \frac{(0.46)(0.54)}{1500}} = 0.0182
\]

If we were to take many SRSs of 1500 adults in October 2012 and 2013, the difference in sample proportions that approve of Obama will typically be 0.0182 from the true difference in proportions.

b) Use the results of these polls to construct and interpret a 90% confidence interval for the change in Obama’s approval rating among all U.S. adults from October 2012 to October 2013.

\[
\text{State: } \hat{p}_1 = \text{Proportion of adults who approve Obama in 2012} \quad \text{DO: } (\hat{p}_1 - \hat{p}_2) \pm Z_{SE_{\hat{p}_1 - \hat{p}_2}}
\]

\[
SE_{\hat{p}_1 - \hat{p}_2} = 0.0182 \quad Z_{0.05} = 1.645
\]

\[
(0.52 - 0.46) \pm 1.645(0.0182) = 0.06 \pm 0.03 = (0.03, 0.09)
\]

Conclude: We are 95% confident the interval from 0.03 to 0.09 captures the true change in proportion of U.S. adults who approve Obama’s job performance from Oct 2012 to Oct 2013.

c) Based on your interval, is there convincing evidence that Obama’s job approval rating has changed?

Because 0 is not in the interval, we have convincing evidence that Obama’s approval rating has changed from 2012 to 2013.
Comparing Two Proportions

Significance Tests for $p_1 - p_2$
An observed difference between two sample proportions can reflect an actual difference in the parameters, or it may just be due to chance variation in random sampling or random assignment. The hypotheses may be be:

**Null Hypothesis:**

$$H_0: p_1 - p_2 = 0 \text{ or } H_0: p_1 = p_2$$

**Alternative Hypothesis:**

- $H_a: p_1 - p_2 \neq 0 \text{ or } H_a: p_1 \neq p_2$
- $H_a: p_1 - p_2 > 0 \text{ or } H_a: p_1 > p_2$
- $H_a: p_1 - p_2 < 0 \text{ or } H_a: p_1 < p_2$

The conditions for a significance test are the same for a two-sample z Interval for a difference of proportions.

If $H_0: p_1 = p_2$ is true, the two parameters are the same. We call their common value $p$. We need a way to estimate $p$, so it makes sense to combine the data from the two samples. This pooled (or combined) sample proportion is:

$$\hat{p}_c = \text{pooled (combined)} = \frac{\text{count of successes in both samples combined}}{\text{count of individuals in both samples combined}} = \frac{x_1 + x_2}{n_1 + n_2}$$

Suppose the conditions are met. To test the hypothesis $H_0: p_1 - p_2 = 0$, first find the pooled proportion $\hat{p}_c$ of successes in both samples combined. Then compute the z statistic

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_1} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_2}}}$$

Find the P-value by calculating the probability of getting a z statistic this large or larger in the direction specified by the alternative hypothesis $H_a$:

Inference for Experiments:
Many important statistical results come from randomized comparative experiments. Defining the parameters in experimental settings is more challenging.

- Most experiments on people use recruited volunteers as subjects.
- When subjects are not randomly selected, researchers cannot generalize the results of an experiment to some larger populations of interest.
- Researchers can draw cause-and-effect conclusions that apply to people like those who took part in the experiment.
- Unless the experimental units are randomly selected, we don’t need to check the 10% condition when performing inference about an experiment.
Comparing Two Proportions

Example: Significance Tests for $p_1 - p_2$ (Inference for Populations)

Researchers designed a survey to compare the proportions of children who come to school without eating breakfast in two low-income elementary schools. An SRS of 80 students from School 1 found that 19 had not eaten breakfast. At School 2, an SRS of 150 students included 26 who had not had breakfast. More than 1500 students attend each school. Do these data give convincing evidence at the $\alpha = 0.05$ level of a difference in the population proportions? (Use the 4 Step Process)

**State:** $H_0: p_1 - p_2 = 0$

$H_a: p_1 - p_2 \neq 0$

$p_1$: Proportion of students at School 1 who did not eat breakfast.

$p_2$: Proportion of students at School 2 who did not eat breakfast.

**Plan:** 2-Sample Z Test

Random → Both are independent random samples

10% Cond → 80 < 10% of students at School 1

150 < 10% of students at School 2

Large Counts → $n_1 \hat{p}_1 = 19 \checkmark \quad n_2 \hat{p}_2 = 26 \checkmark$

$n_1(1-\hat{p}_1) = 61 \checkmark \quad n_2(1-\hat{p}_2) = 124 \checkmark$

All are at least 10.

**Do:**

$\hat{p}_1 = \frac{19}{80} = 0.2375 \quad \hat{p}_2 = \frac{26}{150} = 0.1733$

Pooled $\hat{p} = \frac{19 + 26}{80 + 150} = \frac{45}{230} = 0.196$

$z = \frac{(0.2375 - 0.1733) - 0}{\frac{0.196}{\sqrt{80}} + \frac{0.196}{\sqrt{150}}} = \frac{0.0642}{0.055} = 1.17$

$\text{normal cdf}(1.17, \infty, 0, 1) = 0.1210$

$P-value = 2(0.1210) = 0.2420$

**Conclude:**

Because the $P-value = 0.2420 > \alpha = 0.05$, we fail to reject $H_0$. There is not convincing evidence of a difference in the true proportions of students at School 1 and School 2 who skipped breakfast today.

Independent Practice: Significance Tests for $p_1 - p_2$ (Inference for Experiments)

High levels of cholesterol in the blood are associated with higher risk of heart attacks. Will using a drug to lower blood cholesterol reduce heart attacks? The Helsinki Heart Study recruited middle-aged men with high cholesterol but no history of other serious medical problems to investigate this question. The volunteer subjects were assigned at random to one of two treatments: 2051 men took the drug gemfibrozil to reduce their cholesterol levels, and a control group of 2030 men took a placebo. During the next five years, 56 men in the gemfibrozil group and 84 men in the placebo group had heart attacks. Is this difference statistically significant at the $\alpha = 0.01$ level?

**State:** We hope to show that gemfibrozil reduces heart attacks.

$H_0: p_2 = p_1$

$H_a: p_2 < p_1$

$\alpha = 0.05$

$p_1$: % of heart attacks of all males who use gemfibrozil

$p_2$: % of heart attacks of all males who placebo

**Plan:** 2-Sample Z Test

Random → Randomized Experiment

10% Cond → no sampling is done

Large Counts → $n_1 \hat{p}_1 = 56 \quad n_1(1-\hat{p}_1) = 1995$

$n_2 \hat{p}_2 = 84 \quad n_2(1-\hat{p}_2) = 1996$

**Do:**

$\hat{p}_1 = \frac{56}{2051} = 0.0273 \quad \hat{p}_2 = \frac{84}{2030} = 0.0414$

Pooled $\hat{p} = \frac{56 + 84}{2051 + 2030} = \frac{140}{4081} = 0.0343$

$z = \frac{(0.0273 - 0.0414) - 0}{\sqrt{\frac{0.0273(0.9727)}{2051} + \frac{0.0414(0.9586)}{2030}}} = -2.475$

$\text{normal cdf}(-2.475, \infty, 0, 1) = 0.0067$

**Conclude:**

We $P-value = 0.0067 < \alpha = 0.01$, we reject $H_0$. We have convincing evidence the rate of heart attacks is lower for middle-aged men who take gemfibrozil. Our results are statistically significant.