Chapter 10: Comparing Two Populations or Groups

Section 10.1
Comparing Two Proportions

The Practice of Statistics, 4th edition – For AP*

STARNES, YATES, MOORE

Chapter 10
Comparing Two Populations or Groups

- **10.1 Comparing Two Proportions**
- 10.2 Comparing Two Means

Section 10.1
Comparing Two Proportions

Learning Objectives

After this section, you should be able to…

- ✓ DETERMINE whether the conditions for performing inference are met.
- ✓ CONSTRUCT and INTERPRET a confidence interval to compare two proportions.
- ✓ PERFORM a significance test to compare two proportions.
- ✓ INTERPRET the results of inference procedures in a randomized experiment.

Introduction

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.

What if we want to compare the effectiveness of Treatment 1 and Treatment 2 in a completely randomized experiment? This time, the parameters p_1 *and* p_2 that we want to compare are the true proportions of successful outcomes for each treatment. We use the proportions of successes in the two treatment groups to make the comparison. Here's a table that summarizes these two situations.

Population or treatment	Parameter	Statistic	Sample size
1	p_1	$\hat{p}_1$	n_1
2	p_2	$\hat{p}_2$	n_2

The Sampling Distribution of a Difference Between Two Proportions

In Chapter 7, we saw that the sampling distribution of a sample proportion has the following properties:

Shape Approximately Normal if $np \geq 10$ and $n(1 - p) \geq 10$

Center $\mu_{\hat{p}} = p$

Spread $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$ if the sample is no more than 10% of the population

To explore the sampling distribution of the difference between two proportions, let's start with two populations having a known proportion of successes.

- ✓ 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.

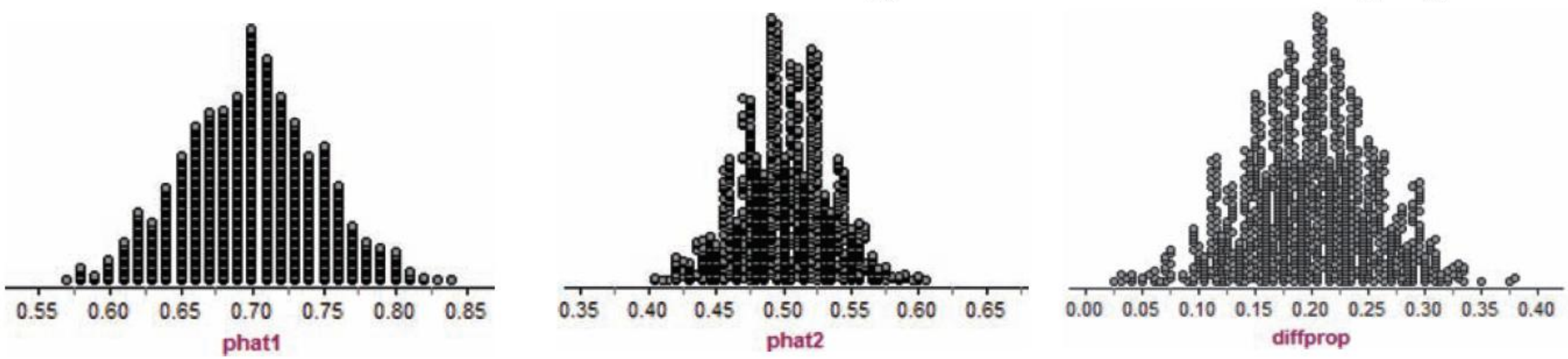
School 2's counselor takes an SRS of 200 students and records the sample proportion that did their homework.

What can we say about the difference $\hat{p}_1 - \hat{p}_2$ **in the sample proportions**?

The Sampling Distribution of a Difference Between Two Proportions

Using Fathom software, we generated an SRS of 100 students from School 1 and a separate SRS of 200 students from School 2. The difference in sample proportions was then calculated and plotted. We repeated this process 1000 times. The results are below:

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



Shape: Approximately Normal
Center: $\mu_{\hat{p}_1} = p_1 = 0.70$
Spread: $\sigma_{\hat{p}_1} = \sqrt{\frac{p_1(1-p_1)}{n_1}} = \sqrt{\frac{0.7(0.3)}{100}} = 0.0458$

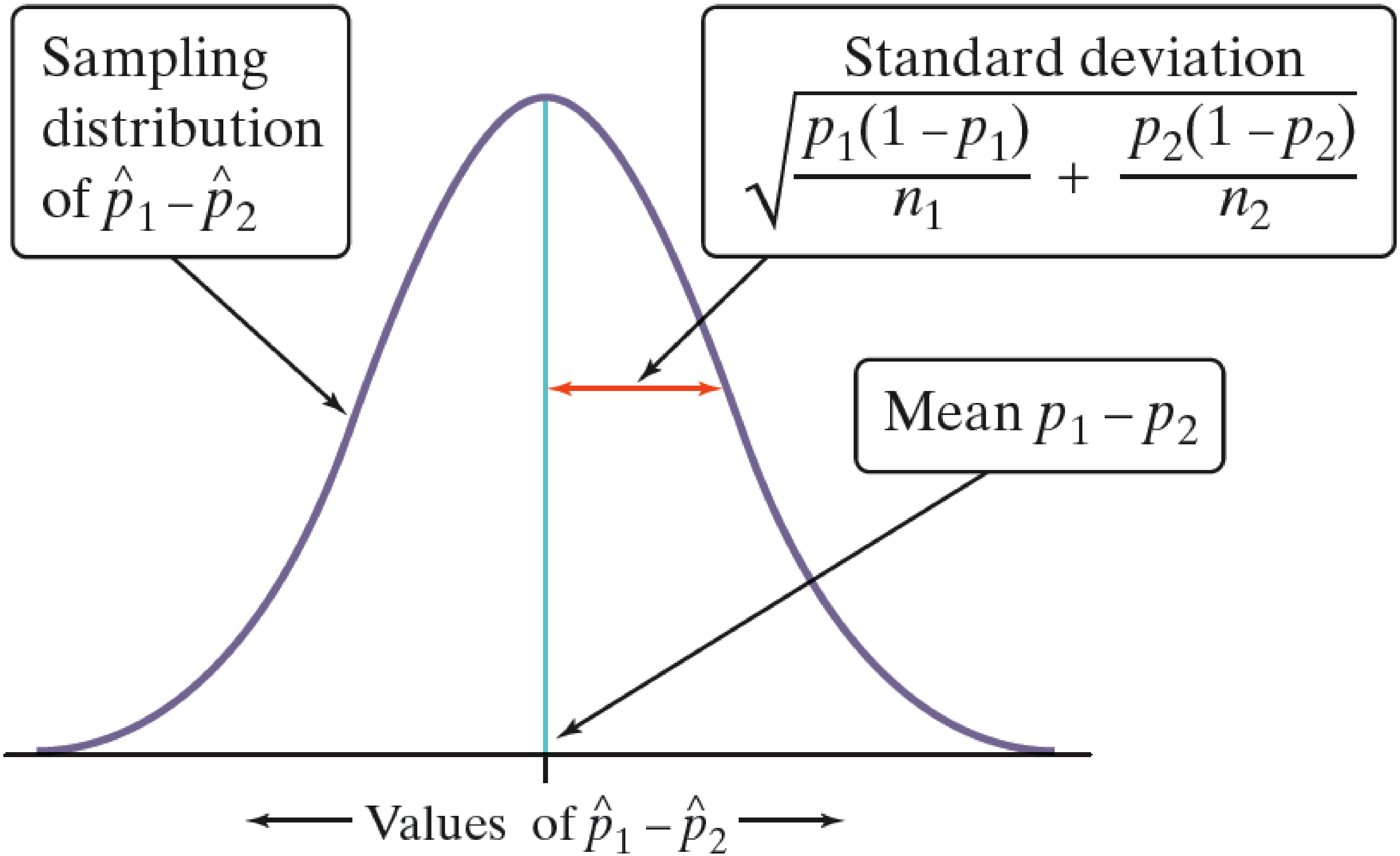
Shape: Approximately Normal
Center: $\mu_{\hat{p}_2} = p_2 = 0.50$
Spread: $\sigma_{\hat{p}_2} = \sqrt{\frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{0.5(0.5)}{200}} = 0.0354$

Shape: Approximately Normal
Center: $\mu_{\hat{p}_1-\hat{p}_2} = 0.20$
Spread: $\sigma_{\hat{p}_1-\hat{p}_2} = 0.058$

What do you notice about the shape, center, and spread of the sampling distribution of $\hat{p}_1 - \hat{p}_2$?

The Sampling Distribution of a Difference Between Two Proportions

Both $\hat{p}_1$ and $\hat{p}_2$ are random variables. The statistic $\hat{p}_1 - \hat{p}_2$ is the difference



Example: Who Does More Homework?

Suppose that there are two large high schools, each with more than 2000 students, in a certain town. At School 1, 70% of students did their homework last night. Only 50% of the students at School 2 did their homework last night. The counselor at School 1 takes an SRS of 100 students and records the proportion that did homework. School 2's counselor takes an SRS of 200 students and records the proportion that did homework. School 1's counselor and School 2's counselor meet to discuss the results of their homework surveys. After the meeting, they both report to their principals that $\hat{p}_1 - \hat{p}_2 = 0.10$.

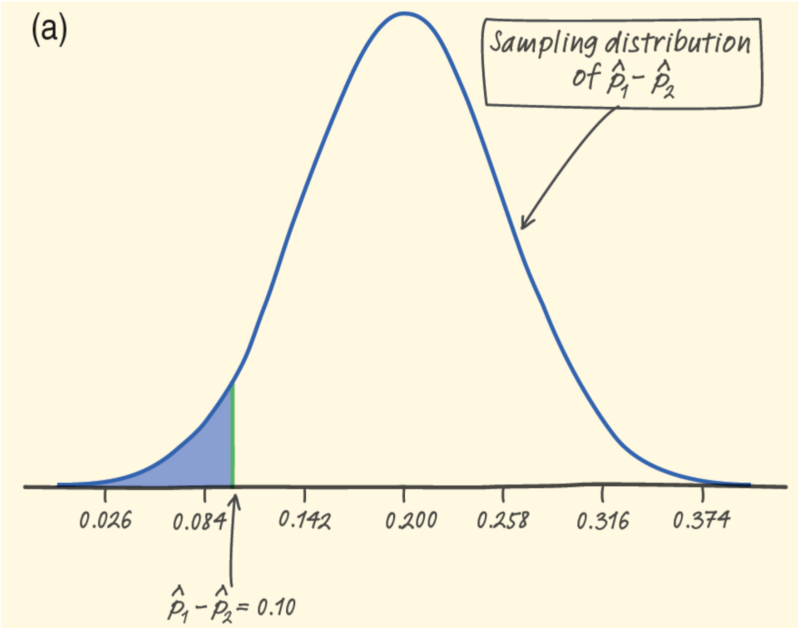
a) Describe the shape, center, and spread of the sampling distribution of $\hat{p}_1 - \hat{p}_2$.

Because $n_1 p_1 = 100(0.7) = 70$, $n_1(1-p_1) = 100(0.30) = 30$, $n_2 p_2 = 200(0.5) = 100$ and $n_2(1-p_2) = 200(0.5) = 100$ are all at least 10, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately Normal.

Its mean is $p_1 - p_2 = 0.70 - 0.50 = 0.20$.

Its standard deviation is

$$\sqrt{\frac{0.7(0.3)}{100} + \frac{0.5(0.5)}{200}} = 0.058.$$

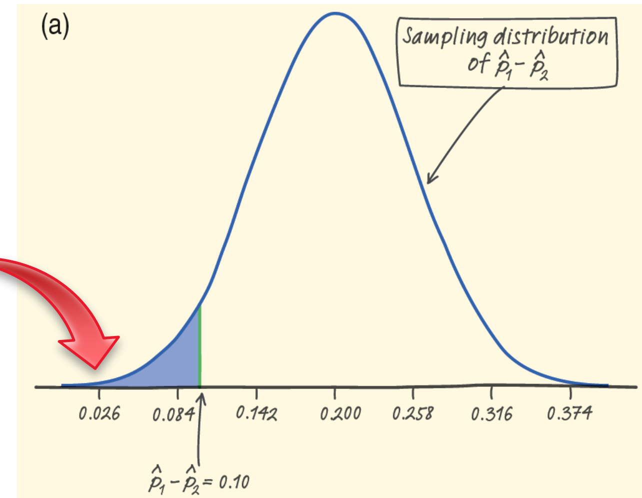


Example: Who Does More Homework?

b) Find the probability of getting a difference in sample proportions $\hat{p}_1 - \hat{p}_2$ of 0.10 or less from the two surveys.

Standardize: When $\hat{p}_1 - \hat{p}_2 = 0.10,$

$$z = \frac{0.10 - 0.20}{0.058} = -1.72$$



Use Table A: The area to the left of $z = -1.72$ under the standard Normal curve is 0.0427.

c) Does the result in part (b) give us reason to doubt the counselors' reported value?

There is only about a 4% chance of getting a difference in sample proportior as small as or smaller than the value of 0.10 reported by the counselors. This does seem suspicious!

Alternate Example: Who Does More Homework? Part 2

- Suppose that two counselors at School 1, Michelle and Julie, independently take a random sample of 100 students from their school and record the proportion of students that did their homework last night. When they are finished, they find that the difference in their proportions, $\hat{p}_M - \hat{p}_J = 0.08.$ They were surprised to get a difference this big, considering they were sampling from the same population.

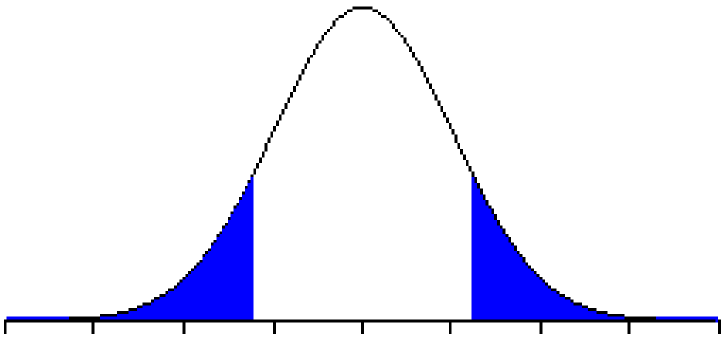
a) Describe the shape, center, and spread of the sampling distribution of $\hat{p}_M - \hat{p}_J$.

Since $n_M p_M = 70, n_M(1 - p_M) = 30, n_J p_J = 70, n_J(1 - p_J) = 300$ are all at least 10, the sampling distribution of $\hat{p}_M - \hat{p}_J$ is approximately Normal.

Its mean is $p_{\mathrm{M}} - p_J = 0.70 - 0.70 = 0.$

Its standard deviation is

$$\sqrt{\frac{0.7(0.3)}{100} + \frac{0.7(0.3)}{100}} = 0.065.$$

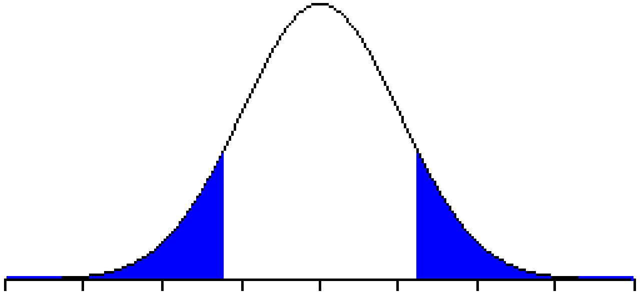


Alternate Example: Who Does More Homework? Part 2

b) Find the probability of getting two proportions that are at least 0.08 apart.

We want to calculate the probability that $\hat{p}_M - \hat{p}_J \le -0.08$ or $\hat{p}_M - \hat{p}_J \ge 0.08$. The figure shows the sampling distribution of $\hat{p}_M - \hat{p}_J$ with the desired area shaded. Using technology, the probability is

$$\text{normalcdf}(-100, -0.08, 0, 0.065) + \text{normalcdf}(0.08, 100, 0, 0.065) = 0.2184.$$



(c) Should the counselors have been surprised to get a difference this big? Explain.

Since the probability isn't very small, we shouldn't be surprised to get a difference of sample proportions of 0.08 or bigger, just by chance, even when sampling from the same population.

Comparing Two Proportions

Confidence Intervals for $p_1 - p_2$

When data come from two random samples or two groups in a randomized experiment, 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 $p_1 - p_2$:

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

When the Independent condition is met, the standard deviation of the statistic $\hat{p}_1 - \hat{p}_2$ is:

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Because we dont 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** of the statistic $\hat{p}_1 - \hat{p}_2$:

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

If the Normal condition is met, we find the critical value *z** for the given confidence level from the standard Normal curve. Our confidence interval for $p_1 - p_2$ is:

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

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

Two-Sample z Interval for $p_1 - p_2$

Two-Sample z Interval for a Difference Between Proportions

When the Random, Normal, and Independent conditions are met, an approximate level 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 area C between $-z^*$ and z^*.

Random The data are produced by a random sample of size n_1 from Population 1 and a random sample of size n_2 from Population 2 or by two groups of size n_1 and n_2 in a randomized experiment.

Normal The counts of "successes" and "failures" in each sample or group -- $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$ and $n_2(1-\hat{p}_2)$ -- are all at least 10.

Independent Both the samples or groups themselves and the individual observations in each sample or group are independent. When sampling without replacement, check that the two populations are at least 10 times as large as the corresponding samples (the 10% condition).

Example: Teens and Adults on Social Networks

As part of the Pew Internet and American Life Project, researchers conducted two surveys in late 2009. The first survey asked a random sample of 800 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, 73% of teens and 47% of adults said that they use social-networking sites. Use these results to 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: Our parameters of interest are p_1 = the proportion of all U.S. teens who use social networking sites and p_2 = the proportion of all U.S. adults who use social-networking sites. We want to estimate the difference $p_1 - p_2$ at a 95% confidence level.

Plan: We should use a two-sample z interval for $p_1 - p_2$ if the conditions are satisfied.

✓ **Random** The data come from a random sample of 800 U.S. teens and a separate random sample of 2253 U.S. adults.

✓ **Normal** We check the counts of "successes" and "failures" and note the Normal condition is met since they are all at least 10:

$$n_1\hat{p}_1 = 800(0.73) = 584 \qquad n_1(1-\hat{p}_1) = 800(1-0.73) = 216$$
$$n_2\hat{p}_2 = 2253(0.47) = 1058.91 \Rightarrow 1059 \qquad n_2(1-\hat{p}_2) = 2253(1-0.47) = 1194.09 \Rightarrow 1194$$

✓ **Independent** We clearly have two independent samples—one of teens and one of adults. Individual responses in the two samples also have to be independent. The researchers are sampling without replacement, so we check the 10% condition: there are at least 10(800) = 8000 U.S. teens and at least 10(2253) = 22,530 U.S. adults.

Example: Teens and Adults on Social Networks

Do: Since the conditions are satisfied, we can construct a two-sample *z* interval for the difference $p_1 - p_2$.

STEP 4

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = (0.73 - 0.47) \pm 1.96\sqrt{\frac{0.73(0.27)}{800} + \frac{0.47(0.53)}{2253}}$$

$$= 0.26 \pm 0.037$$

$$= (0.223,\ 0.297)$$

Conclude: We are 95% confident that the interval from 0.223 to 0.297 captures the true difference in the proportion of all U.S. teens and adults who use social-networking sites. This interval suggests that more teens than adults in the United States engage in social networking by between 22.3 and 29.7 percentage points.

Alternate Example: Presidential approval

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 August 2009 to September 2010? According to a CNN poll of 1024 randomly selected U.S. adults on September 1-2, 2010, 50% approved of Obama's job performance. A CNN poll of 1010 randomly selected U.S. adults on August 28-30, 2009 showed that 53% approved of Obama's job performance. Use the results of these polls to construct and interpret a 90% confidence interval for the change in Obama's approval rating among all US adults.

State: We want to estimate $p_{2010} - p_{2009}$ at the 90% confidence level where p_{2010} = the true proportion of all U.S. adults who approved of President Obama's job performance in September 2010 and p_{2009} = the true proportion of all U.S. adults who approved of President Obama's job performance in August 2009.

Plan: We should use a two-sample z interval for $p_{2010} - p_{2009}$ if the conditions are satisfied.

✓ **Random** The data come from a random sample of 800 U.S. teens and a separate random sample of 2253 U.S. adults.

✓ **Normal**

$$n_{2010}\hat{p}_{2010} = 512 \qquad n_{2010}(1-\hat{p}_{2010}) = 512$$

$$n_{2009}\hat{p}_{2009} = 535 \qquad n_{2009}(1-\hat{p}_{2009}) = 475, \text{ all are at least } 10$$

✓ **Independent** The samples were taken independently and there are more than 10(1024) = 10,240 U.S. adults in 2010 and 10(1010) = 10,100 U.S. adults in 2009.

Alternate Example: Presidential approval

Do: Since the conditions are satisfied, we can construct a two-sample *z* interval for the difference $p_{2010} - p_{2009}$.

STEP 4

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

$$= (0.50 - 0.53) \pm 1.645\sqrt{\frac{0.5(1-0.5)}{1024} + \frac{0.53(1-0.53)}{1010}}$$

$$= 0.03 \pm 0.036$$

$$= (-0.066, 0.006)$$

Conclude: We are 95% confident that the interval from –0.066 to 0.006 captures the true change in the proportion of U.S. adults who approve of President Obama's job performance from August 2009 to September 2010. That is, it is plausible that his job approval has fallen by up to 6.6 percentage points or increased by up to 0.6 percentage points.

(b) Based on your interval, is there convincing evidence that Obama's job approval rating has changed?

Since 0 is included in the interval, it is plausible that there has been no change in President Obama's approval rating. Thus, we do not have convincing evidence that his approval rating has changed.

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. Significance tests help us decide which explanation makes more sense. The null hypothesis has the general form

$$H_0: p_1 - p_2 = \text{hypothesized value}$$

We'll restrict ourselves to situations in which the hypothesized difference is 0. Then the null hypothesis says that there is no difference between the two parameters:

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

The alternative hypothesis says what kind of difference we expect.

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

If the Random, Normal, and Independent conditions are met, we can proceed with calculations.

Alternate Example – Hearing Loss

Are teenagers going deaf? In a study of 3000 randomly selected teenagers in 1988-1994, 15% showed some hearing loss. In a similar study of 1800 teenagers in 2005-2006, 19.5% showed some hearing loss. Does these data give convincing evidence that the proportion of all teens with hearing loss has increased? (Source: *Arizona Daily Star*, 8-18-2010).

Problem: State the hypotheses we are interested in testing. Define any parameters you use.

Solution: If p_1 = the proportion of all teenagers with hearing loss in 2005 - 2006 and p_2 = the proportion of all teenagers with hearing loss in 1988-1994,

$$H_0: p_1 - p_2 = 0 \text{ and } \mathrm{H_a}: p_1 - p_2 > 0$$

Significance Tests for $p_1 - p_2$

To do a test, standardize $\hat{p}_1 - \hat{p}_2$ to get a z statistic:

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\text{standard deviation of statistic}}$$

If H_0: $p_1 = p_2$ is true, the two parameters are the same. We call their common value p. But now 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:

$$\hat{p}_C = \frac{\text{count of successes in both samples combined}}{\text{count of individuals in both samples combined}} = \frac{X_1 + X_2}{n_1 + n_2}$$

Use $\hat{p}_C$ in place of both p_1 and p_2 in the expression for the denominator of the test statistic:

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

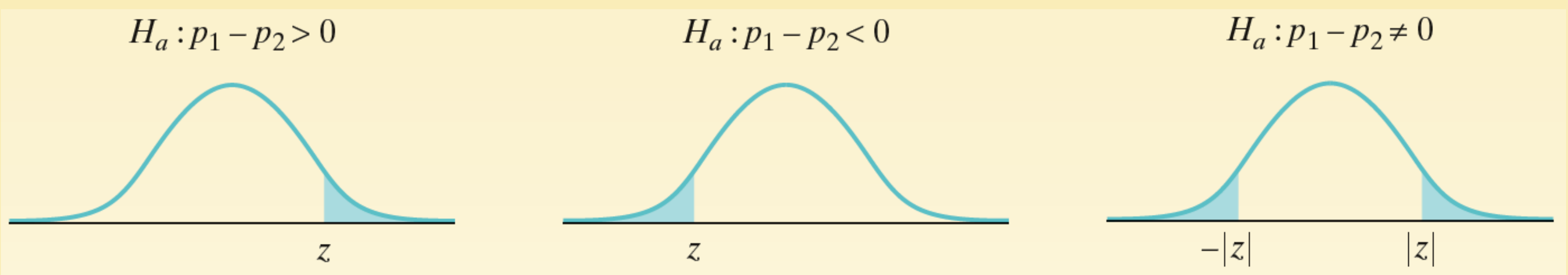
Two-Sample *z* Test for The Difference Between Two Proportions

Two-Sample *z* Test for the Difference Between Proportions

Suppose the Random, Normal, and Independent 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

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

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



Example: Hungry Children

- 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 of a difference in the population proportions? Carry out a significance test at the $\alpha = 0.05$ level to support your answer.

State: Our hypotheses are

$$H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 \neq 0$$

where p_1 = the true proportion of students at School 1 who did not eat breakfast, and p_2 = the true proportion of students at School 2 who did not eat breakfast.

Plan: We should perform a two-sample z test for $p_1 - p_2$ if the conditions are satisfied.

✓ **Random** The data were produced using two simple random samples—of 80 students from School 1 and 150 students from School 2.

✓ **Normal** We check the counts of "successes" and "failures" and note the Normal condition is met since they are all at least 10:

$$n_1\hat{p}_1 = 19,\ n_1(1-\hat{p}_1) = 61,\ n_2\hat{p}_2 = 26,\ n_2(1-\hat{p}_2) = 124$$

✓ **Independent** We clearly have two independent samples—one from each school. Individual responses in the two samples also have to be independent. The researchers are sampling without replacement, so we check the 10% condition: there are at least 10(80) = 800 students at School 1 and at least 10(150) = 1500 students at School 2.

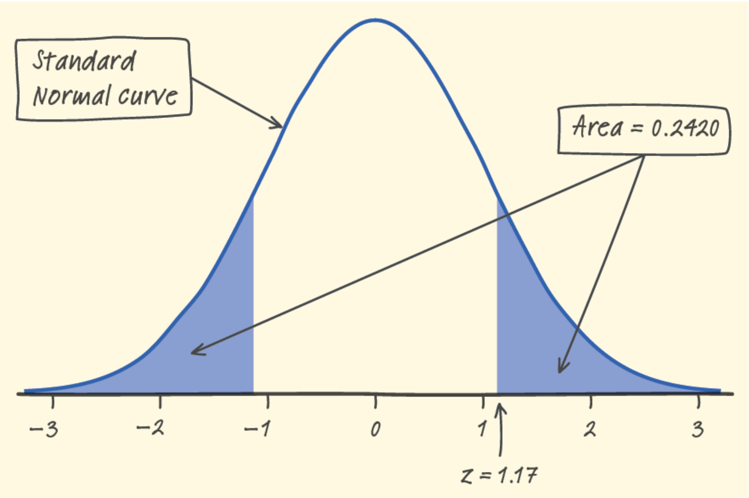
Example: Hungry Children

Do: Since the conditions are satisfied, we can perform a two-sample *z* test for the difference $p_1 - p_2$.

$$\hat{p}_C = \frac{X_1 + X_2}{n_1 + n_2} = \frac{19 + 26}{80 + 150} = \frac{45}{230} = 0.1957$$

Test statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_C(1-\hat{p}_C)}{n_1} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_2}}} = \frac{(0.2375 - 0.1733) - 0}{\sqrt{\frac{0.1957(1-0.1957)}{80} + \frac{0.1957(1-0.1957)}{150}}} = 1.17$$



P-value Using Table A or normalcdf, the desired *P*-value is
$2P(z \geq 1.17) = 2(1 - 0.8790) = 0.2420$.

Conclude: Since our *P*-value, 0.2420, is greater than the chosen significance level of $\alpha = 0.05$, we fail to reject H_0. There is not sufficient evidence to conclude that the proportions of students at the two schools who didn't eat breakfast are different.

Alternate Example: Hearing loss

- Are teenagers going deaf? In a study of 3000 randomly selected teenagers in 1988-1994, 15% showed some hearing loss. In a similar study of 1800 teenagers in 2005-2006, 19.5% showed some hearing loss. (Source: *Arizona Daily Star*, 8-18-2010). Does these data give convincing evidence that the proportion of all teens with hearing loss has increased?

State: We will test

$$H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 > 0$$

at the 0.05 significance level, where p_1 = the true proportion of all teenagers with hearing loss in 2005-2006 , and p_2 = the true proportion of all teenagers with hearing loss in 1988-1994.

Plan: We should perform a two-sample z test for $p_1 - p_2$ if the conditions are satisfied.

✓ **Random** The data came from separate random samples.

✓**Normal** $n_1\hat{p}_1 = 351, n_1(1-\hat{p}_1) = 1449, n_2\hat{p}_2 = 450, n_2(1-\hat{p}_2) = 2550$ are all at least 10.

✓ **Independent** The samples were taken independently and there were more than 10(1800) = 18,000 teenagers in 2005-2006 and 10(3000) = 30,000 teenagers in 1988-1994.

Example: Hungry Children

Do:

$$\hat{p}_C = \frac{X_1 + X_2}{n_1 + n_2} = \frac{450 + 351}{3000 + 1800} = 0.167$$

Test statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_C(1-\hat{p}_C)}{n_1} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_2}}} = \frac{(0.195 - 0.15) - 0}{\sqrt{\frac{0.167(1-0.167)}{1800} + \frac{0.167(1-0.167)}{3000}}} = 4.05$$

P-value Using Table A or normalcdf, the desired *P*-value is about 0.

Conclude: Since the *P*-value is less than 0.05, we reject H_0. We have convincing evidence that the proportion of all teens with hearing loss has increased from 1988-1994 to 2005-2006.

(b) Between the two studies, Apple introduced the iPod. If the results of the test are statistically significant, can we blame iPods for the increased hearing loss in teenagers?

No. Since we didn't do an experiment where we randomly assigned some teens to listen to iPods and other teens to avoid listening to iPods, we cannot conclude that iPods are the cause. It is possible that teens who listen to iPods also like to listen to music in their cars and perhaps the car stereos are causing the hearing loss.

Example: Significance Test in an Experiment

- 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 the apparent benefit of gemfibrozil statistically significant? Perform an appropriate test to find out.

State: Our hypotheses are

$$H_0: p_1 - p_2 = 0 \qquad OR \qquad H_0: p_1 = p_2$$
$$H_a: p_1 - p_2 < 0 \qquad\qquad\quad H_a: p_1 < p_2$$

where p_1 is the actual heart attack rate for middle-aged men like the ones in this study who take gemfibrozil, and p_2 is the actual heart attack rate for middle-aged men like the ones in this study who take only a placebo. No significance level was specified, so we'll use $\alpha = 0.01$ to reduce the risk of making a Type I error (concluding that gemfibrozil reduces heart attack risk when it actually doesn't).

Example: Cholesterol and Heart Attacks

Plan: We should perform a two-sample z test for $p_1 - p_2$ if the conditions are satisfied.

✓ **Random** The data come from two groups in a randomized experiment

✓ **Normal** The number of successes (heart attacks!) and failures in the two groups are 56, 1995, 84, and 1946. These are all at least 10, so the Normal condition is met.

✓ **Independent** Due to the random assignment, these two groups of men can be viewed as independent. Individual observations in each group should also be independent: knowing whether one subject has a heart attack gives no information about whether another subject does.

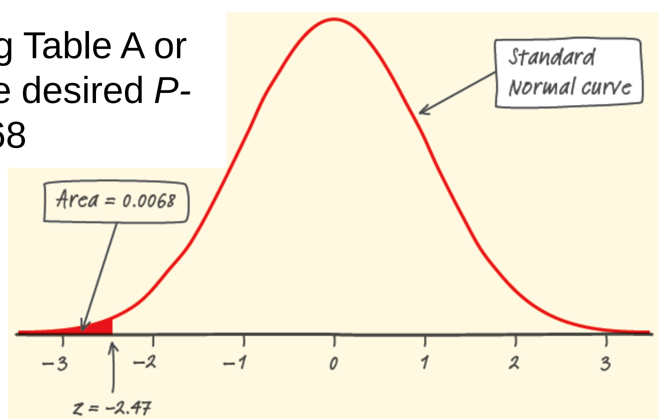
Do: Since the conditions are satisfied, we can perform a two-sample *z* test for the difference $p_1 - p_2$.

$$\hat{p}_C = \frac{X_1 + X_2}{n_1 + n_2} = \frac{56 + 84}{2051 + 2030} = \frac{140}{4081} = 0.0343$$

Test statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_C(1-\hat{p}_C)}{n_1} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_2}}} = \frac{(0.0273 - 0.0414) - 0}{\sqrt{\frac{0.0343(1-0.0343)}{2051} + \frac{0.0343(1-0.0343)}{2030}}} = -2.47$$

***P*-value** Using Table A or normalcdf, the desired *P*-value is 0.0068



Conclude: Since the *P*-value, 0.0068, is less than 0.01, the results are statistically significant at the $\alpha = 0.01$ level. We can reject H_0 and conclude that there is convincing evidence of a lower heart attack rate for middle-aged men like these who take gemfibrozil than for those who take only a placebo.

Alternate Example: Cash for quitters

- In an effort to reduce health care costs, General Motors sponsored a study to help employees stop smoking. In the study, half of the subjects were randomly assigned to receive up to $750 for quitting smoking for a year while the other half were simply encouraged to use traditional methods to stop smoking. None of the 878 volunteers knew that there was a financial incentive when they signed up. At the end of one year, 15% of those in the financial rewards group had quit smoking while only 5% in the traditional group had quit smoking. Do the results of this study give convincing evidence that a financial incentive helps people quit smoking? (Source: *Arizona Daily Star*, 2-11-09).

State: We will test

$$H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 > 0$$

at the 0.05 significance level, where p_1 = the true quitting rate for employees like these who get a financial incentive to quit smoking and p_2 = the true quitting rate for employees like these who don't get a financial incentive to quit smoking.

Example: Cholesterol and Heart Attacks

Plan: We should perform a two-sample z test for $p_1 - p_2$ if the conditions are satisfied.

✓ **Random** The treatments were randomly assigned.

✓ **Normal** $n_1\hat{p}_1 = 66, n_1(1-\hat{p}_1) = 373, n_2\hat{p}_2 = 22, n_2(1-\hat{p}_2) = 417$ are all at least 10.

✓ **Independent** The random assignment allows us to view these two groups as independent. We must assume that each employee's decision to quit is independent of other employee's decisions.

Do:

$$\hat{p}_C = \frac{X_1 + X_2}{n_1 + n_2} = \frac{66+22}{439+439} = 0.100$$

Test statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_C(1-\hat{p}_C)}{n_1} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_2}}} = \frac{(0.15-0.05)-0}{\sqrt{\frac{0.1(1-0.1)}{439} + \frac{0.1(1-0.1)}{439}}} = 4.94$$

***P*-value** is about 0.

Conclude: Since the *P*-value is less than 0.05, we reject H_0. We have convincing evidence that financial incentives help employees like these quit smoking.

+

Section 10.1
Comparing Two Proportions

Summary

In this section, we learned that…

- Choose an SRS of size n_1 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 $n_1p_1,\ n_1(1-p_1),\ n_2p_2$ 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 $p_1 - p_2$. That is, the difference in sample proportions is an unbiased estimator c the difference in population proportions.

  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% conditic

- Confidence intervals and tests to compare the proportions p_1 and p_2 of successes for two populations or treatments are based on the difference between the sample proportions.

- When the Random, Normal, and Independent conditions are met, we can use two-sample z procedures to estimate and test claims about p_1 - p_2.

Section 10.1
Comparing Two Proportions

Summary

In this section, we learned that…

- The conditions for two-sample *z* procedures are:

  Random The data are produced by a random sample of size n_1 from Population 1 and a random sample of size n_2 from Population 2 or by two groups of size n_1 and n_2 in a randomized experiment.

  Normal The counts of "successes" and "failures" in each sample or group -- $n_1\hat{p}_1,\ n_1(1-\hat{p}_1),\ n_2\hat{p}_2$ and $n_2(1-\hat{p}_2)$ -- are all at least 10.

  Independent Both the samples or groups themselves and the individual observations in each sample or group are independent. When sampling without replacement, check that the two populations are at least 10 times as large as the corresponding samples (the 10% condition).

- An approximate level *C* confidence interval for $p_1 - p_2$ 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 standard Normal critical value. This is called a **two-sample z interval** for $p_1 - p_2$.

+

Section 10.1
Comparing Two Proportions

Summary

In this section, we learned that…

- ✓ **Significance tests** of H_0: p_1 - p_2 = 0 use the **pooled (combined) sample proportion**

$$\hat{p}_C = \frac{\text{count of successes in both samples combined}}{\text{count of individuals in both samples combined}} = \frac{X_1 + X_2}{n_1 + n_2}$$

- ✓ The **two-sample *z* test for p_1 - p_2** uses the test statistic

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

  with *P*-values calculated from the standard Normal distribution.

- ✓ Inference about the difference p_1 - p_2 in the effectiveness of two treatments in a completely randomized experiment is based on the **randomization distribution** of the difference of sample proportions. When the Random, Normal, and Independent conditions are met, our usual inference procedures based on the sampling distribution will be approximately correct.

Looking Ahead…

In the next Section…

We'll learn how to compare two population means.

We'll learn about

- ✓ **The sampling distribution for the difference of means**
- ✓ **The two-sample *t* procedures**
- ✓ **Comparing two means from raw data and randomized experiments**
- ✓ **Interpreting computer output for two-sample *t* procedures**