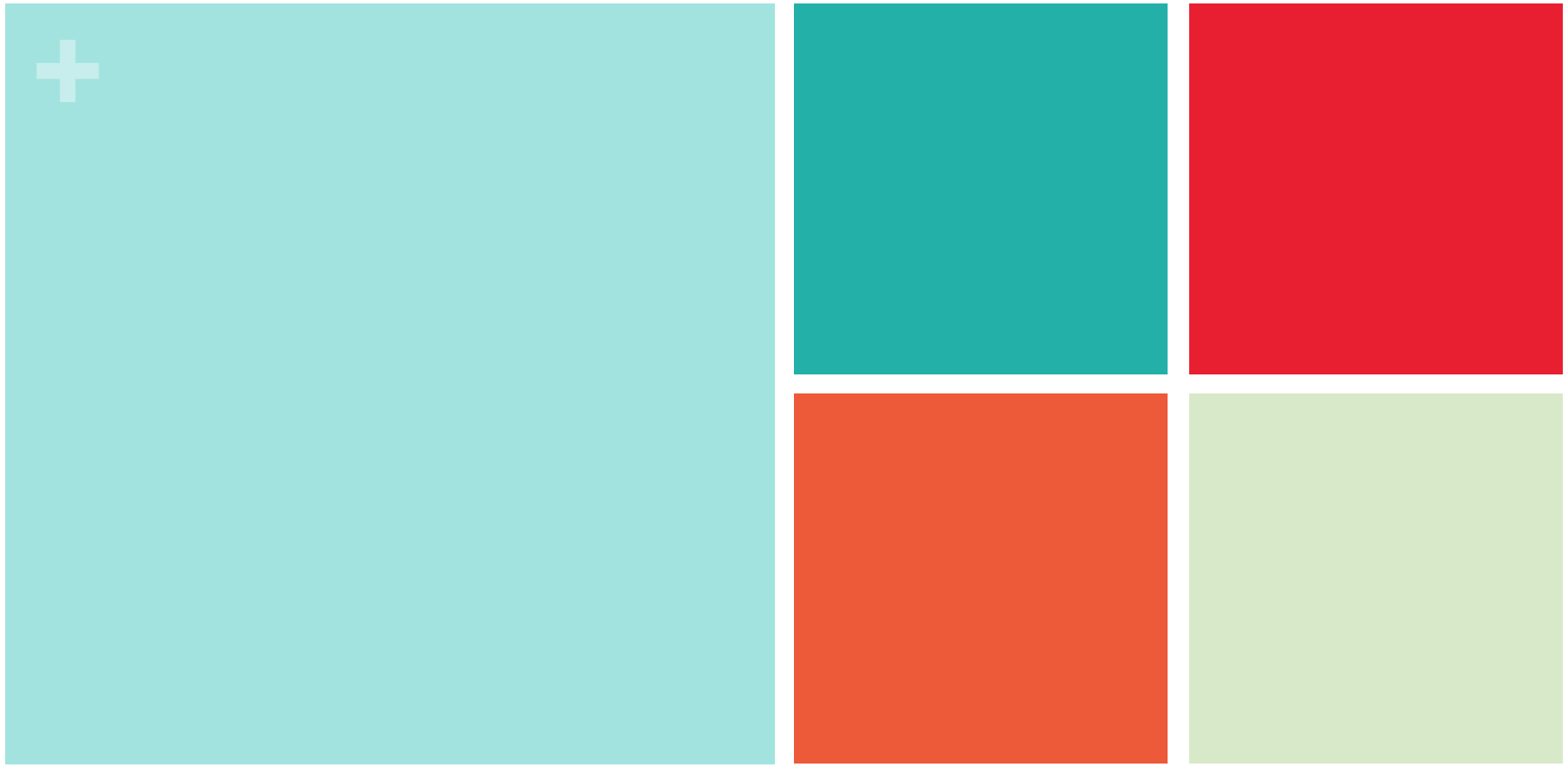


Unit 6 - Comparing Two Populations or Groups

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 does the CLT say about the Sampling Distribution of a Difference Between Two Proportions

The Sampling Distribution of the Difference Between Sample Proportions

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.

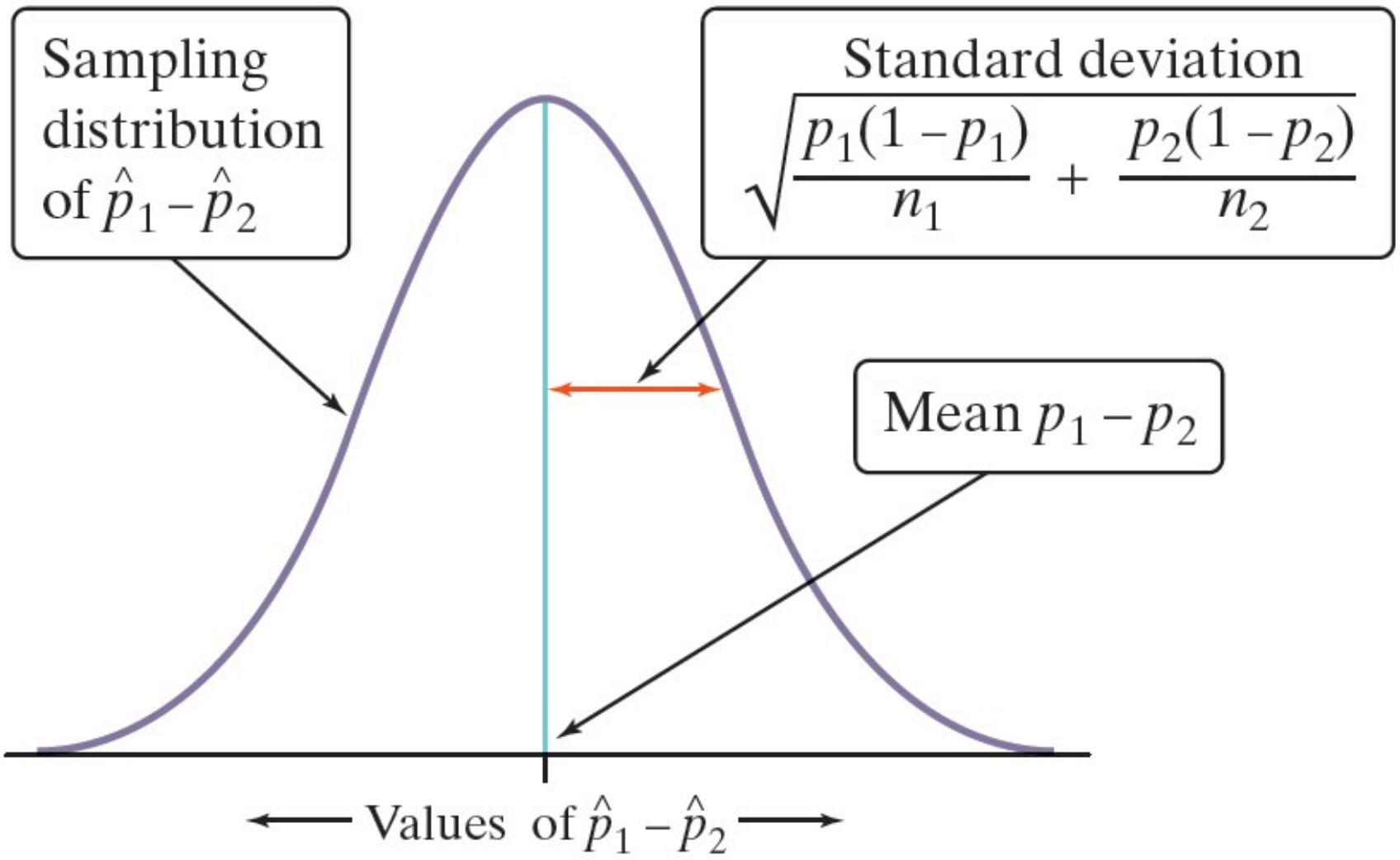
Center The mean of the sampling distribution is $p_1 - p_2$. That is, the difference in sample proportions is an unbiased estimator of the difference in population propotions.

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).

The Sampling Distribution of a Difference Between Two Proportions



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.

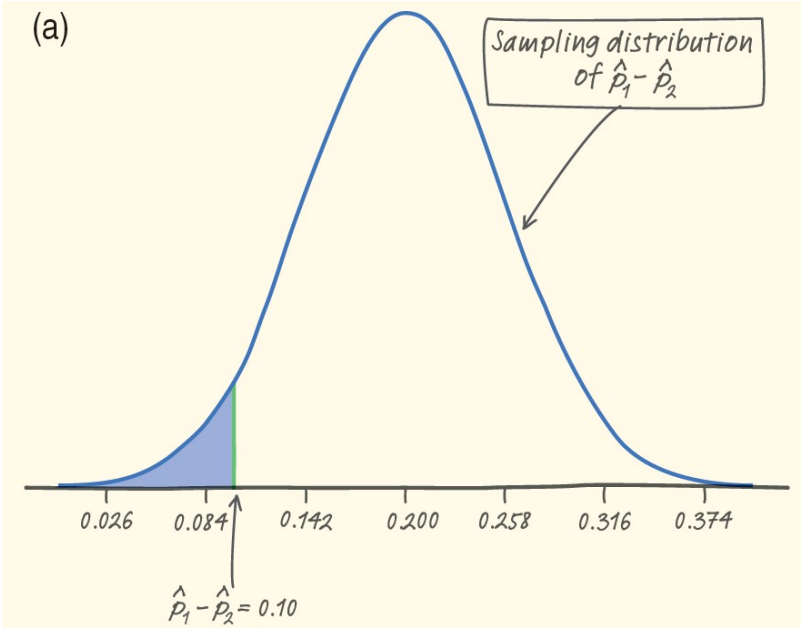
a) Describe the shape, center, and spread of the sampling distribution of $\hat{p}_1 - \hat{p}_2$.

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

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



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 $p_1 - p_2$:

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

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** 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 $p_1 - p_2$

Conditions for Two-Sample *z* Interval for a Difference Between Proportions

The 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.

Conditions 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).

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-proportion 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 $p_1 - p_2$.

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

$$= 0.26 \pm 0.037$$

$$= (0.223,\ 0.297)$$

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.

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:

$H_0: p_1 - p_2 = 0$ or, alternatively, $H_0: p_1 = p_2$

The alternative hypothesis says what kind of difference we expect.

$H_a: p_1 - p_2 > 0$, $H_a: p_1 - p_2 < 0$, or $H_a: p_1 - p_2 \neq 0$

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

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:

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

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

If the following conditions are met, we can proceed with a two-sample *z* test for the difference between two proportions:

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: 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.

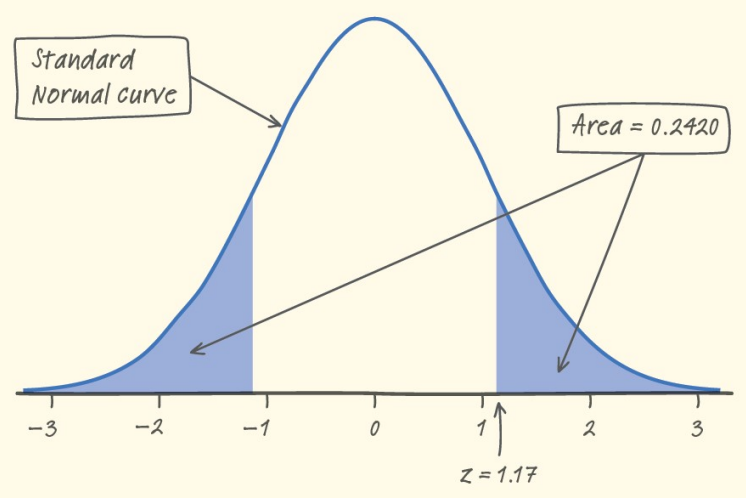
Example: Hungry Children

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.

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

$H_0: p_1 - p_2 = 0$ OR $H_0: p_1 = p_2$

$H_a: p_1 - p_2 < 0$ $H_a: p_1 < p_2$

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.

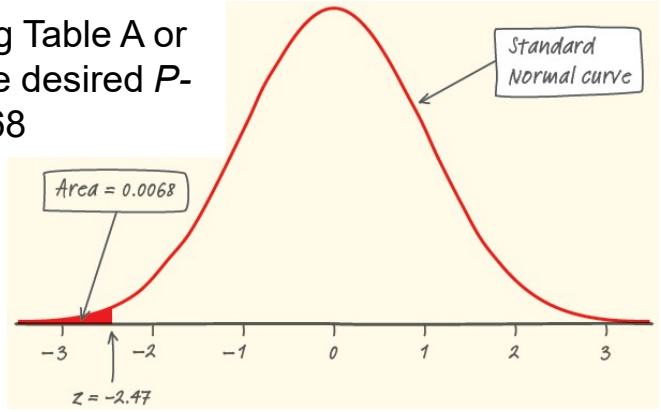
Do: Since the conditions are satisfied, we can perform a two-sample *z* test for the difference $p_1 - p_2$.

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

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.

Homework

Chapter 10. #’s, 9b, 11, 15, 16
Topic 24 additional notes

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 of 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% condition).

- 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.