Chapter 10: Comparing Two Populations or Groups

Section 10.1
Comparing 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.

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.

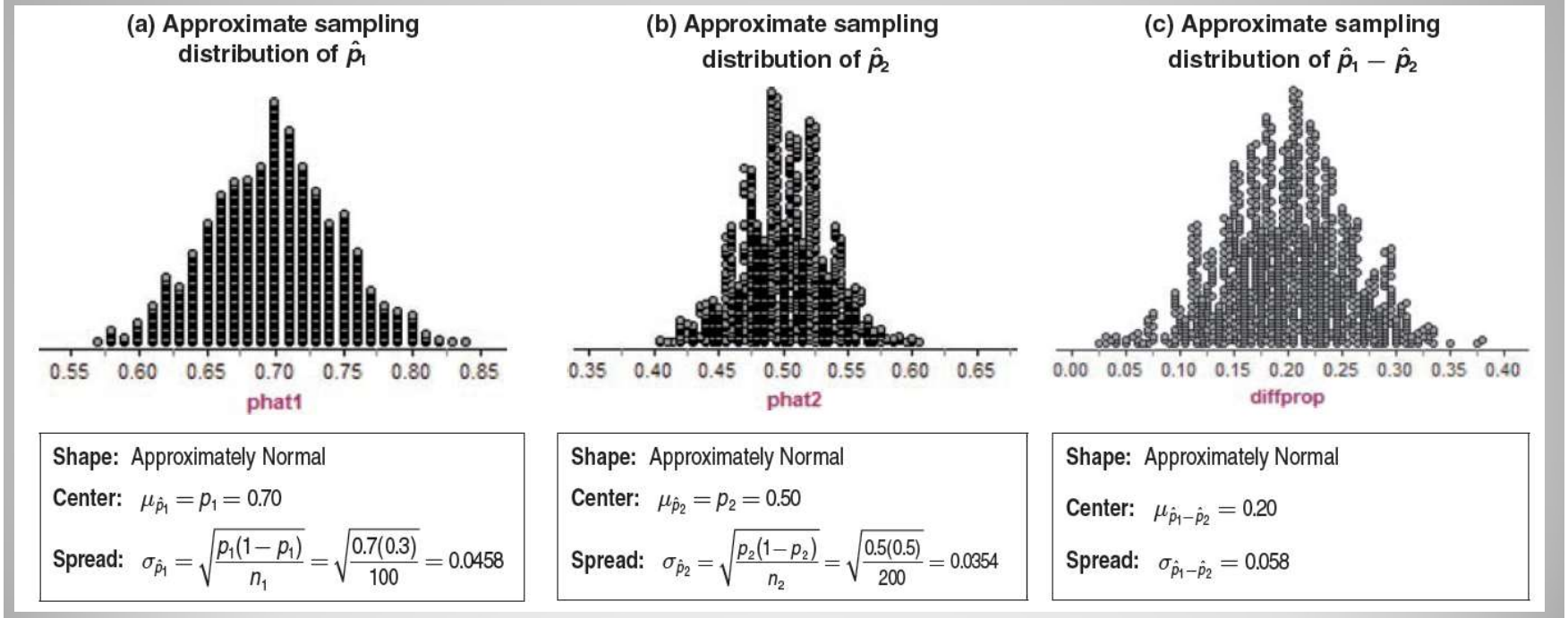
Suppose the counselor at School 1 takes an SRS of 100 students and records the sample proportion that did their homework.

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:



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

Both $\hat{p}_1$ and $\hat{p}_2$ are random variables. The statistic $\hat{p}_1 - \hat{p}_2$ is the difference of these two random variables. In Chapter 6, we learned that for any two independent random variables X and Y,

$\mu_{X-Y} = \mu_X - \mu_Y$ and $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$

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 $n_1 p_1$, $n_1(1-p_1)$, $n_2 p_2$ 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 $p_1 - p_2$. 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).

Example 1: 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?

Because $n_1p_1 = 50(0.25) = 12.5$, $n_1(1 - p_1) = 50(0.75) = 37.5$, $n_2p_2 = 40(0.35) = 14$, and $n_2(1 - p_2) = 40(0.65) = 26$ are all at least 10, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately Normal.

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

The mean is $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2 = 0.25 - 0.35 = -0.10$.

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

Because there are at least 10(50) = 500 crackers in Bag 1 and 10(40) = 400 crackers in Bag 2, the standard deviation is

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{(0.25)(0.75)}{50} + \frac{(0.35)(0.65)}{40}} = 0.0971$$

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 $p_1 - p_2$:

statistic ± (critical value) × (standard deviation of statistic)

When the 10% condition is met, the standard deviation of the statistic $\hat{p}_1 - \hat{p}_2$ is:

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

If the Large Counts condition is met, we find the critical value z^* for the given confidence level from the standard Normal curve.

Conditions For Constructing A Confidence Interval About A Difference In Proportions

Random: The data come from two independent random samples or from two groups in a randomized experiment.

10%: When sampling without replacement, check that $n_1 \leq (1/10)N_1$ and $n_2 \leq (1/10)N_2$.

Large Counts: 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.

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** (also called the *estimated standard deviation*) of the statistic $\hat{p}_1 - \hat{p}_2$:

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

When the conditions are met, our confidence interval for $p_1 - p_2$ is:

statistic ± (critical value) · (standard deviation of statistic)

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

This is often called a **two-sample *z* interval for a difference between two proportions**.

Two-Sample z Interval for a Difference Between Two Proportions

When the Random, Normal, and Independent conditions are met, an approximate level C confidence interval for $\hat{p}_1 - \hat{p}_2$ is

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

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

Example 2: 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?

The sample proportions of teens and adults who use social-networking sites are $\hat{p}_1 = 0.80$ and $\hat{p}_2 = 0.69$, respectively. The standard error of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is

$$SE_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{(0.80)(0.20)}{799} + \frac{(0.69)(0.31)}{2253}} = 0.0172$$

If we were to take many random samples of 799 teens and 2253 adults, the difference in the sample proportions of teens and adults who use social-networking sites will typically be 0.0172 from the true difference in proportions of all teens and adults who use social-networking sites.

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: 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 met.

✓ **Random** The data come from independent random samples of 799 U.S. teens and 2253 U.S. adults.

✓**10%:** The researchers are sampling without replacement, so we must check the 10% condition: there are at least 10(799) = 7990 U.S. teens and at least 10(2253) = 22,530 U.S. adults.

✓ **Large Counts:** We check the counts of "successes" and "failures":

$n_1\hat{p}_1 = 799(0.80) = 639.2 \rightarrow 639$ $\quad n_1(1-\hat{p}_1) = 799(1-0.80) = 159.8 \rightarrow 160$

$n_2\hat{p}_2 = 2253(0.69) = 1554.57 \rightarrow 1555$ $\quad n_2(1-\hat{p}_2) = 2253(1-0.69) = 698.43 \rightarrow 698$

Note that the observed counts have to be whole numbers! Because all four values are at least 10, this condition is met.

Do: We know that $n_1 = 799$, $\hat{p}_1 = 0.80$, $n_2 = 2253$, and $\hat{p}_2 = 0.69$. For a 95% confidence level, the critical value is $z^* = 1.96$. So our 95% confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = (0.80 - 0.69) \pm 1.96\sqrt{\frac{0.80(0.20)}{799} + \frac{0.69(0.31)}{2253}}$$

$$= 0.11 \pm 0.034$$

$$= (0.076,\ 0.144)$$

CONCLUDE: We are 95% confident that the interval from 0.07588 to 0.14324 captures the true difference in the proportion of all U.S. teens and adults who use social-networking sites.

AP EXAM TIP The formula for the two-sample z interval for p1 − p2 often leads to calculation errors by students. As a result, we recommend using the calculator's 2-PropZIntfeature to compute the confidence interval on the AP® exam. Be sure to name the procedure (two-proportion z interval) and to give the interval (0.076, 0.143) as part of the "Do" step.

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

Example 3: 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? State appropriate hypotheses for a significance test to answer this question. Define any parameters you use.

Our hypotheses are

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

OR

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

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.

The conditions for performing a significance test about $p_1 - p_2$ are the same as for constructing a confidence interval.

Note: It would also be correct to check the Normal condition using the pooled (combined) sample proportion $\hat{p}_C$. Just confirm that $n_1\hat{p}_C, n_1(1-\hat{p}_C), n_2\hat{p}_C$, and $n_2(1-\hat{p}_C)$ are all at least 10.

If the conditions are met, we can proceed with calculations. 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:

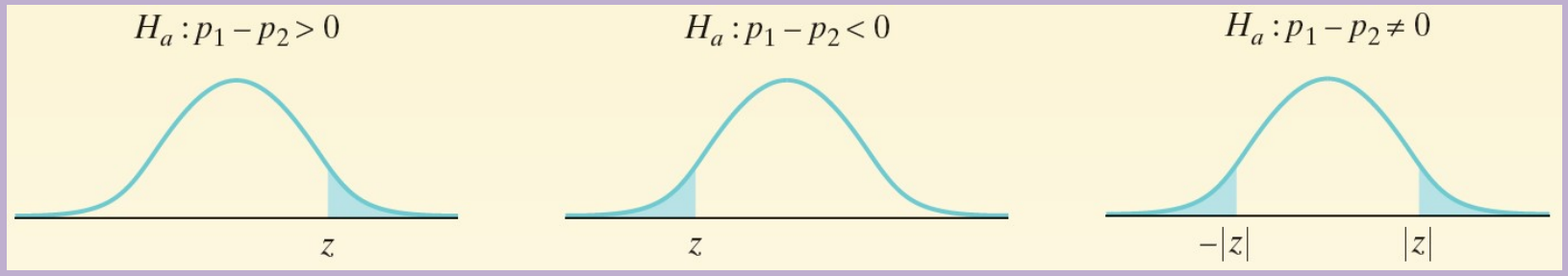
$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_C(1-\hat{p}_C)}{n_1} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_2}}}$$

Two-Sample z Test for the Difference Between Proportions

Suppose the conditions are met. To test the hypothesis H_0: $p_1 - p_2 = 0$, first find the pooled proportion 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:



Example 4: 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?

State: Our hypotheses are

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

OR

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

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: If conditions are met, we should perform a two-sample z test for $p_1 - p_2$.

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

✓**10%:** 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.

✓ **Large Counts:** We check the counts of "successes" and "failures":

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

All four values are at least 10, so the sampling distribution is approximately Normal.

Do: We know that $n_1 = 80$, $\hat{p}_1 = \frac{19}{80} = 0.2375$, $n_2 = 150$, and $\hat{p}_2 = \frac{26}{150} = 0.1733$. Our point estimate for the difference in population proportions is $p_1 - p_2 = 0.2375 - 0.1733 = 0.0642$. The pooled proportion of students who didn't eat breakfast in the two samples is

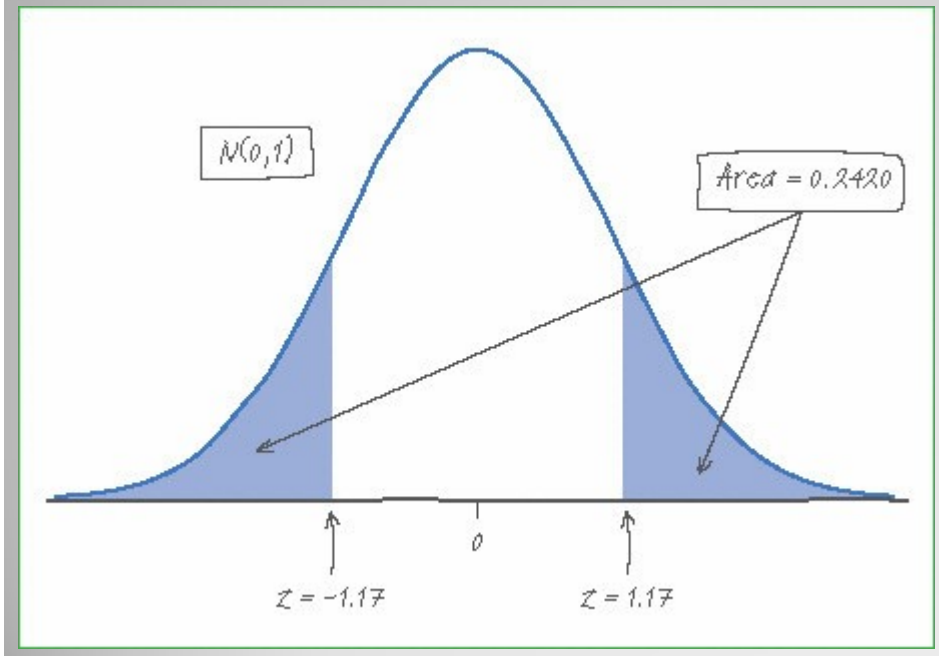
$$\hat{p}_C = \frac{X_1 + X_2}{n_1 + n_2} = \frac{19 + 26}{80 + 150} = \frac{45}{230} = 0.1957$$

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** The figure below displays the *P*-value as an area under the standard Normal curve for this two-tailed test. Using **Table A**, the desired *P*-value is $2P(Z \geq 1.17) = 2(1 - 0.8790) = 0.2420$.

Using Technology: $2P(Z \geq 1.17)$ = 2normalcdf(lower: 1.17, upper: 99999, μ: 0, σ: 1) = 0.2427



Conclude: We fail to reject H_0. Since our *P*-value, 0.2427, is greater than $\alpha = 0.05$. There is not enough evidence to suggest that the proportions of students at the two schools who didn't eat breakfast are different.

AP EXAM TIP: The formula for the two-sample z statistic for a test about $p_1 - p_2$ often leads to calculation errors by students. As a result, we recommend using the calculator's 2-PropZTest feature to perform calculations on the AP® exam. Be sure to name the procedure (two-proportion z test) and to report the test statistic($z = 1.17$) and P-value (0.2427) as part of the "Do" step.

Example 5: 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, so we have a one-sided alternative:

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

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. We'll use $\alpha = 0.01$.

Plan: If conditions are met, we will do a two-sample z test for $p_1 - p_2$.

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

10%: Don't need to check because there was no sampling.

***Large Counts*:** 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 this condition is met.

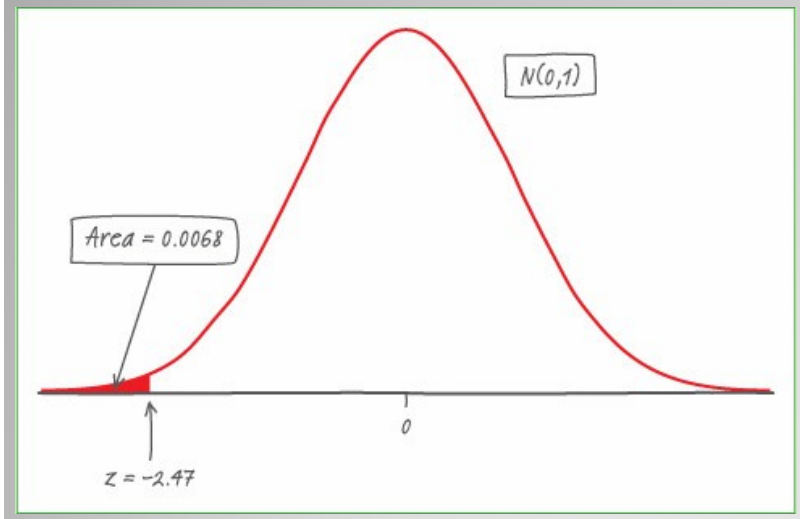
Do: Since the conditions are satisfied, we can perform a two-sample z test for the difference $p_1 - p_2$.

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

$$= \frac{140}{4081} = 0.0343$$

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** $P(z < -2.47)$ = normalcdf(lower: –9999, upper –2.47, μ: 0, σ: 1) = 0.0068

Conclude: Reject H_0. Since the *P*-value, 0.0068, is less than $\alpha = 0.01$. There is enough evidence to suggest that there is a lower heart attack rate for middle-aged men like these who take gemfibrozil than for those who take only a placebo.