Chapter 8

Inference for Proportions

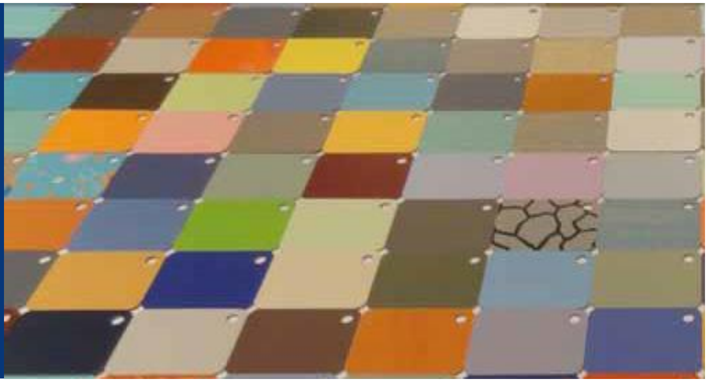
Introduction to the Practice of

STATISTICS EIGHTH EDITION

Moore / McCabe / Craig

Lecture Presentation Slides

Chapter 8
Inference for Proportions



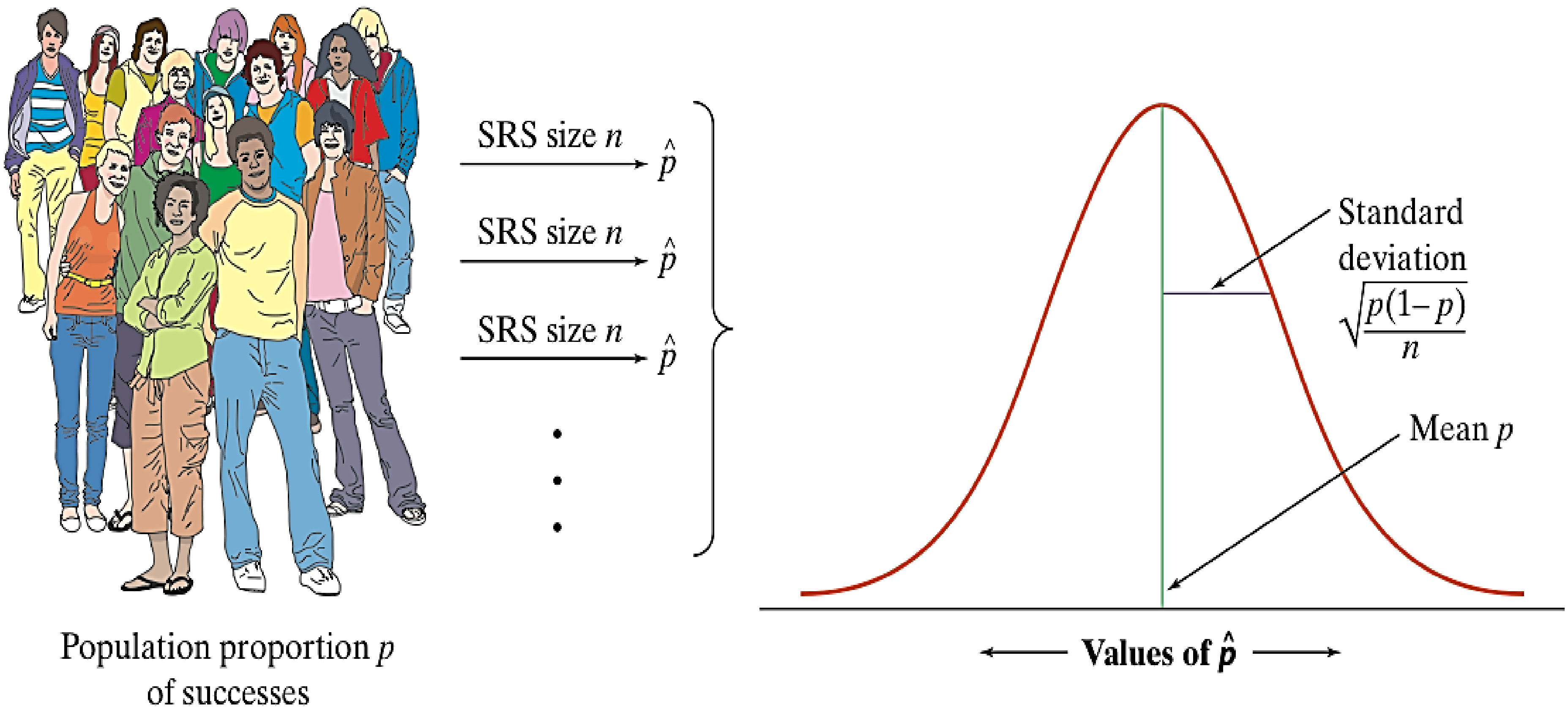
8.1 Inference for a Single Proportion

8.2 Comparing Two Proportions

8.1 Inference for a Single Proportion

- Large-sample confidence interval for a single proportion
- Significance test for a single proportion
- Choosing a sample size

Sampling Distribution of a Sample Proportion



Large-Sample Confidence Interval for a Proportion

We can use the same path from sampling distribution to confidence interval as we did with means to construct a confidence interval for an unknown population proportion *p*:

$$\textbf{statistic} \pm \textbf{(critical value)} \times \textbf{(standard deviation of statistic)}$$

The sample proportion $\hat{p}$ is the statistic we use to estimate p.
When the Independent condition is met, the standard deviation of the sampling distibution of $\hat{p}$ is:

$$s_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Since we don’t know p, we replace it with the sample proportion $\hat{p}$.
This gives us the **standard error (SE)** of the sample proportion:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

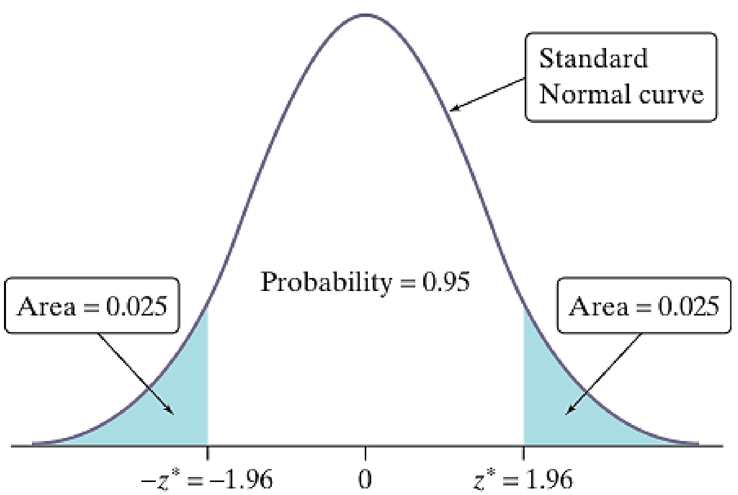
Large-Sample Confidence Interval for a Proportion

How do we find the critical value for our confidence interval?

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

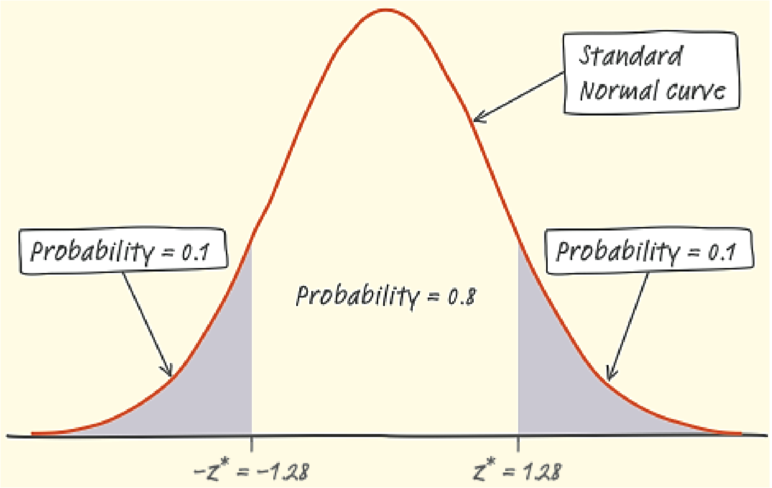
If the Normal condition is met, we can use a Normal curve. To find a level *C* confidence interval, we need to catch the central area *C* under the standard Normal curve.

For example, to find a 95% confidence interval, we use a critical value of 2 based on the 68-95-99.7 rule. Using a Standard Normal Table or a calculator, we can get a more accurate critical value.
Note, the **critical value *z**** is actually 1.96 for a 95% confidence level.



Large-Sample Confidence Interval for a Proportion

Find the critical value z^* for an 80% confidence interval. Assume that the Normal condition is met.



Since we want to capture the central 80% of the standard Normal distribution, we leave out 20%, or 10% in each tail.
Search Table A to find the point z^* with area 0.1 to its left.

z	**.07**	**.08**	**.09**
– 1.3	.0853	.0838	.0823
– 1.2	.1020	.1003	.0985
– 1.1	.1210	.1190	.1170

The closest entry is $z = -1.28$.

So, the **critical value z^*** for an 80% confidence interval is $z^* = 1.28$.

Large-Sample Confidence Interval for a Proportion

Once we find the critical value *z**, our confidence interval for the population proportion *p* is:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

One-Sample *z* Interval for a Population Proportion

Choose an SRS of size *n* from a large population that contains an unknown proportion *p* of successes. An approximate level *C* **confidence interval for *p*** is:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where *z** is the critical value for the standard Normal density curve with area *C* between – *z** and *z**.

Use this interval only when the numbers of successes and failures in the sample are both at least 15.

Example

Your instructor claims 50% of the beads in a container are red. A random sample of 251 beads is selected, of which 107 are red. Calculate and interpret a 90% confidence interval for the proportion of red beads in the container. Use your interval to comment on this claim.

z	**.03**	**.04**	**.05**
– 1.7	.0418	.0409	.0401
– 1.6	.0516	.0505	.0495
– 1.5	.0630	.0618	.0606

✓ sample proportion = 107/251 = 0.426

✓ This is an SRS and there are 107 successes and 144 failures. Both are greater than 15.

✓ For a 90% confidence level, z^* = 1.645

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= 0.426 \pm 1.645\sqrt{\frac{(0.426)(1-0.426)}{251}}$$

$$= 0.426 \pm 0.051$$

$$= (0.375,\ 0.477)$$

We are 90% confident that the interval from 0.375 to 0.477 captures the actual proportion of red beads in the container.

Since this interval gives a range of plausible values for p and since 0.5 is not contained in the interval, we have reason to doubt the claim.

Significance Test for a Proportion

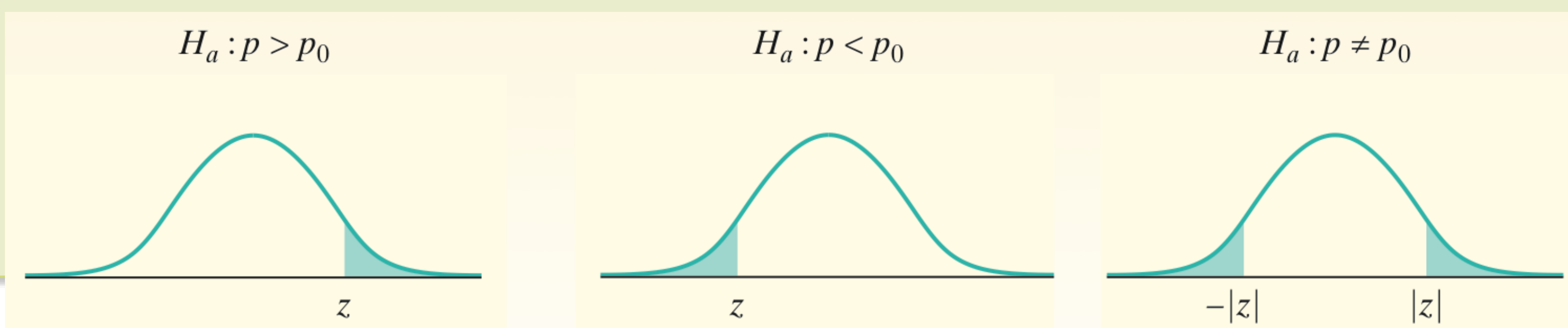
The z statistic has approximately the standard Normal distribution when H_0 is true. P-values therefore come from the standard Normal distribution. Here is a summary of the details for a ***z* test for a proportion.**

z Test for a Proportion

Choose an SRS of size n from a large population that contains an unknown proportion p of successes. To test the hypothesis H_0: $p = p_0$, compute the z statistic:

$$z = \frac{\hat{p} - p}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$$

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

A potato-chip producer has just received a truckload of potatoes from its main supplier. If the producer determines that more than 8% of the potatoes in the shipment have blemishes, the truck will be sent away to get another load from the supplier. A supervisor selects a random sample of 500 potatoes from the truck. An inspection reveals that 47 of the potatoes have blemishes. Carry out a significance test at the $\alpha = 0.10$ significance level. What should the producer conclude?

We want to perform a test at the $\alpha = 0.10$ significance level of

$$H_0: p = 0.08$$
$$H_a: p > 0.08$$

where p is the actual proportion of potatoes in this shipment with blemishes.

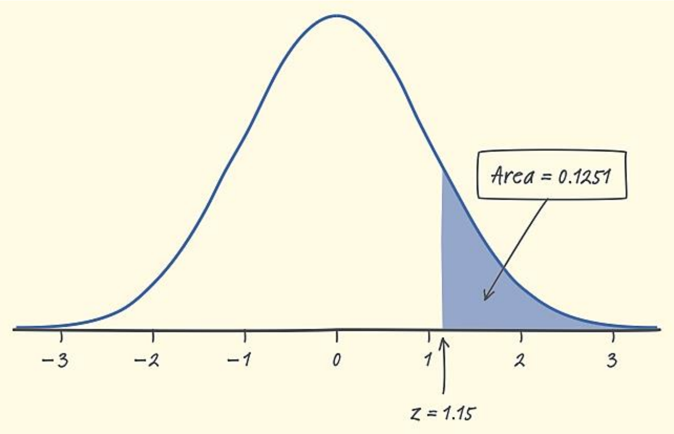
If conditions are met, we should do a one-sample z test for the population proportion p.

- ✓ **Random:** The supervisor took a random sample of 500 potatoes from the shipment.
- ✓ **Normal:** Assuming H_0: $p = 0.08$ is true, the expected numbers of blemished and unblemished potatoes are $np_0 = 500(0.08) = 40$ and $n(1 - p_0) = 500(0.92) = 460$, respectively. Because both of these values are at least 10, we should be safe doing Normal calculations.

Example

The sample proportion of blemished potatoes is $\hat{p} = 47/500 = 0.094$.

Test statistic $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} = \dfrac{0.094 - 0.08}{\sqrt{\dfrac{0.08(0.92)}{500}}} = 1.15$



***P*-value** The desired *P*-value is:

$P(z \geq 1.15) = 1 - 0.8749 = 0.1251$

Since our *P*-value, 0.1251, is greater than the chosen significance level of $\alpha = 0.10$, we fail to reject H_0. There is not sufficient evidence to conclude that the shipment contains more than 8% blemished potatoes. The producer will use this truckload of potatoes to make potato chips.

Choosing the Sample Size

In planning a study, we may want to choose a sample size that allows us to estimate a population proportion within a given margin of error.

$$m = z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

✓ z^* is the standard Normal critical value for the level of confidence we want.

Because the margin of error involves the sample proportion $\hat{p}$, we have to guess the latter value when choosing n. There are two ways to do this:

- Use a guess for $\hat{p}$ based on past experience or a pilot study.
- Use $\hat{p} = 0.5$ as the guess. The margin of error is largest when $\hat{p} = 0.5$.

Sample Size for Desired Margin of Error

To determine the sample size *n* that will yield a level *C* confidence interval for a population proportion *p* with a maximum margin of error, solve the following:

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1 - p^*)$$

where p^* is a guessed value for the sample proportion. The margin of error will always be less than or equal to m if you take the guess p^* to be 0.5.

Example

Suppose you wish to determine what percent of voters favor a particular candidate. Determine the sample size needed to estimate *p* within 0.03 with 95% confidence.

- The critical value for 95% confidence is $z^* = 1.96$.
- Since the company president wants a margin of error of no more than 0.03, we need to solve the equation:

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1 - p^*)$$

$$n = \left(\frac{1.96}{0.03}\right)^2 0.5(1 - 0.5)$$

$$n = 1067.1$$

We round up to 1068 respondents to ensure the margin of error is no more than 0.03 at 95% confidence.

8.2 Comparing Two Proportions

- Large-sample confidence interval for a difference in proportions
- Plus-four confidence interval for a difference in proportions
- Significance test for a difference in proportions
- Relative risk

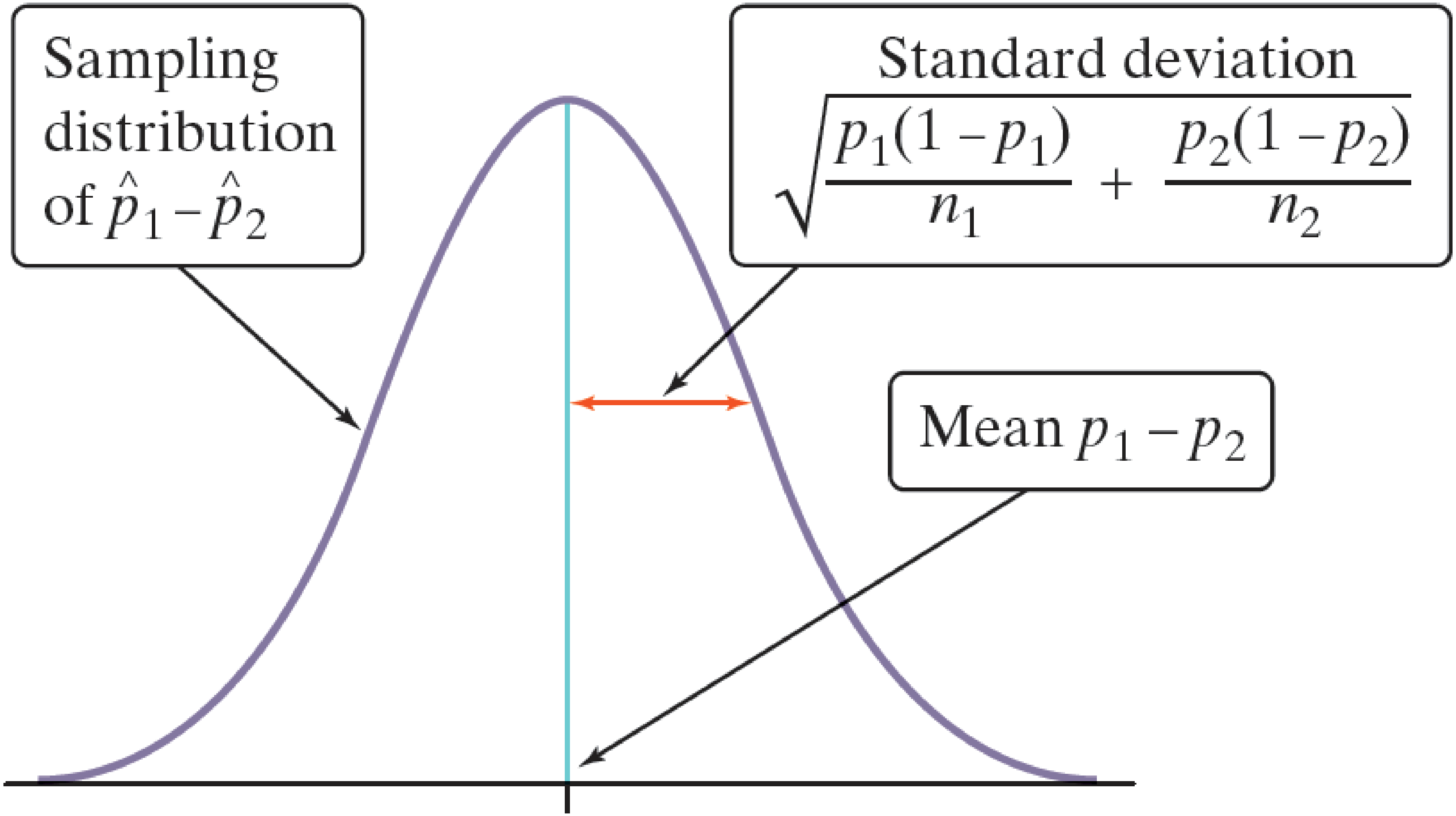
Two-Sample Problems: Proportions

Suppose we want to compare the proportions of individuals having 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

Sampling Distribution of a Difference Between Proportions



Large-Sample Confidence Interval for Comparing Proportions

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 $p_1 - p_2$:

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

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 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 $\hat{p}_1 - \hat{p}_2$:

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

Large-Sample Confidence Interval for Comparing Proportions

Large-Sample Confidence Interval for Comparing Proportions

When the Random and Normal 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 sizes 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.

Example

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.

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.

We should use a large-sample confidence 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 because they are all greater than 10.

Example

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

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

$$= 0.26 \pm 0.037$$

$$= (0.223,\ 0.297)$$

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 Test for Comparing Proportions

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.

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:

$$\hat{p} = \frac{\text{count of successes in both samples combined}}{\text{count of individuals in both samples combined}}$$

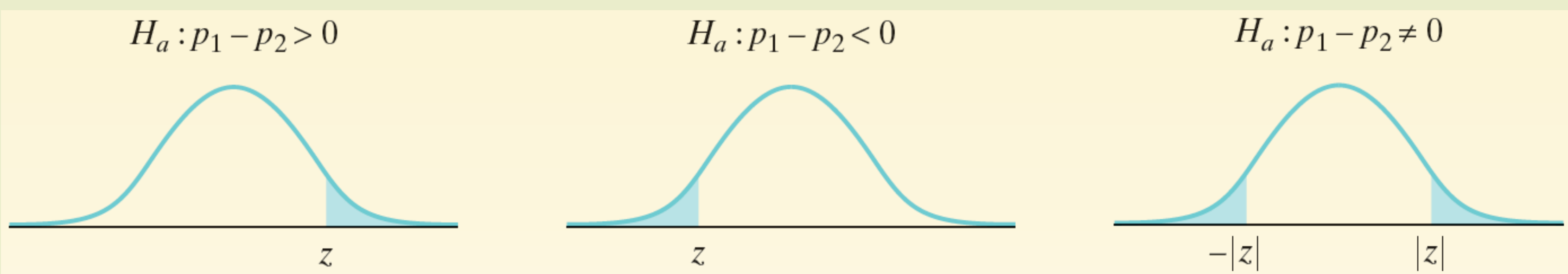
Significance Test for Comparing Proportions

Significance Test for Comparing Two Proportions

Draw an SRS of size n_1 from a large population having proportion p_1 of successes, and draw an independent SRS of size n_2 from a large population having proportion p_2 of successes. To test the hypothesis $H_0\colon p_1 - p_2 = 0,$ first find the pooled proportion $\hat{p}$ of successes in both samples combined. Then compute the z statistic

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

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

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.

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.

We should perform a significance test for $p_1 - p_2$ if the conditions are satisfied.

✓**Random:** The data were produced using two simple random samples—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 because they are all greater than 5.

Example

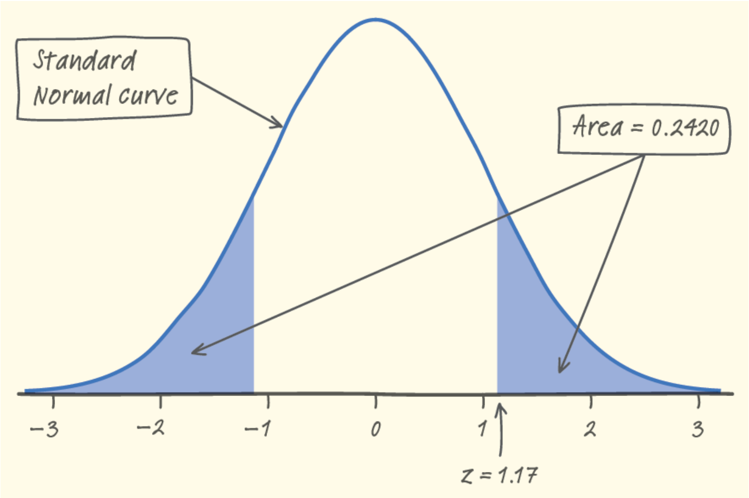
✓**Independent:** The samples were taken independently of each other, since they were at two different schools. Also, the population sizes are at least 10 times the sample sizes.

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

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

Test statistic :

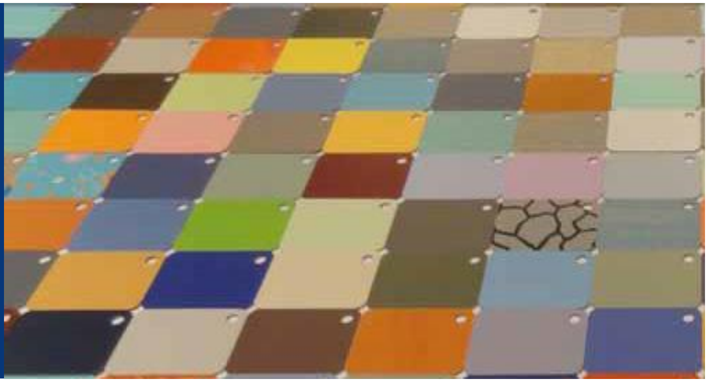
$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(0.2375 - 0.1733)}{\sqrt{0.1957(1-0.1957)\left(\frac{1}{80} + \frac{1}{150}\right)}} = 1.17$$



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

Since our P-value, 0.2420, is greater than the chosen significance level of α = 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.

Chapter 8
Inference for Proportions



8.1 Inference for a Single Proportion

8.2 Comparing Two Proportions