10.1 Comparing Two Proportions

- ✓ DESCRIBE the shape, center, and spread of the sampling distribution of the difference of two sample proportions.
- ✓ 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.

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.

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

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 $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$ $\quad \sigma_{X-Y} = \sqrt{\sigma^2_X + \sigma^2_Y}$

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 $p_1 - p_2$. $P_1 - P_2$

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

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

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

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

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

Shape?

Normal: $100(.3) \geq 10$, $30 \geq 10$; $200(.5) \geq 10$, $100 \geq 10$. Approx. Normal

Mean: $.7 - .5$; $\mu_{\hat{p}_1 - \hat{p}_2} = .2$

Standard deviation: $\sqrt{\frac{.7(.3)}{100} + \frac{.5(.5)}{200}}$

.0579

In repeated sampling like these we expect to be within .0579 of the true mean difference in the proportions of these two schools.

Confidence Intervals for $p_1 - p_2$

Conditions:

- **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.

Normal

When the conditions are met, an approximate C% confidence interval for $\hat{p}_1 - \hat{p}_2$ is

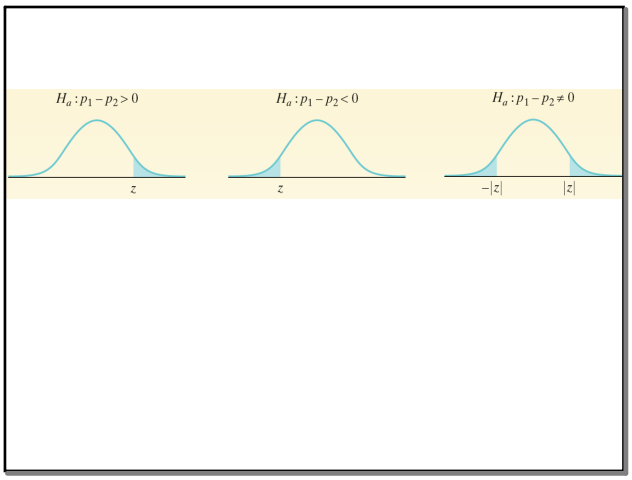
2 Prop Z int $\quad (\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ $\quad$ Margin of Error

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

Significance Tests for $p_1 - p_2$

H_o: $p_1 - p_2 = 0$ or H_o: $p_1 = p_2$

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



(10) $n_1 = 634$ $n_2 = 567$

$x_1 = 368$ $x_2 = 130$

$p_1 = .58$ $p_2 = .23$

a) $\sqrt{\frac{.58(.42)}{634} + \frac{.23(.77)}{567}} = .0264$

In repeated samples like these we expect the mean difference to be .0264 off from the true mean difference in the proportion of young blacks vs young whites who listen to rap music.

b) 95% CI (.299, .403)

We are 95% confident that the interval .299 to .403 contains the true mean difference in the proportions of young blacks vs. young whites listening to rap.

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

$p_1 - p_2 = 0$

Zero not in the interval so we Reject H_0!

We have strong evidence that more young blacks listen to rap music than do young whites.

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

CI 95%

$\alpha = .05$

2-PropZTest

$z = 12.33$ p-value = 0

$0 < .05$

Reject H_0