10.1 Comparing 2 proportions

We compare 2 proportions by finding their difference, $p_1 - p_2$ between the parameters. The statistic used to estimate the parameter is $\hat{p}_1 - \hat{p}_2$, called the <u>2 sample proportion</u>.

Sampling Distribution of $\hat{p}_1 - \hat{p}_2$

choose an SRS of size n_1 from population 1 with proportion of success p_1 and and an SRS of size n_2 from population 2 with proportion of success p_2.

<u>shape</u>: when $n_1 p_1$ and $n_1(1-p_1) \geq 10$ and $n_2 p_2$ and $n_2(1-p_2) \geq 10$ the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal.

<u>center</u>: mean of the sampling distribution, $\mu_{\hat{p}_1 - \hat{p}_2}$ is:

$$\hat{p}_1 - \hat{p}_2 = p_1 - p_2$$

<u>spread</u>: the standard deviation of the sampling distribution $\sigma_{\hat{p}_1 - \hat{p}_2}$ is:

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

assuming <u>each</u> sample is less than 10% of the respective populations.

Confidence Interval for $\hat{p}_1 - \hat{p}_2$

Conditions

1. Data come from 2 independent random samples.
2. 10% condition: $n_1 \leq \frac{1}{10} N_1$ AND $n_2 \leq \frac{1}{10} N_2$
3. successes + failures in each sample are greater than 10

   $n_1 \hat{p}_1 \geq 10$ $\quad n_1(1-\hat{p}_1) \geq 10$ AND $n_2 \hat{p}_2 \geq 10$ $\quad n_2(1-\hat{p}_2) \geq 10$

Two Sample z interval for the difference of 2 proportions

When conditions are met, a C% CI for $p_1 - p_2$ is:

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

Check your understanding on page 619

p_1 = true proportion of adults that go online everyday
p_2 = true proportion of teens that go online everyday
We are going to construct CI for $p_1 - p_2$ with 90% confidence

We are going to use the 2 sample z interval for the difference of 2 proportions.

Conditions: Random: they are both independent random samples
10%: 799 teens is less than 10% of all teens
2253 adults is less than 10% of all adults.

Large counts: 503 successes (internet daily) for teens + 296 failures (no internet daily) for teens which are both greater than 10.

1532 successes for adults + 721 failures are both greater than 10.

Using calc., use 2 proportion z confidence interval
(.01803, .08287)

x_1 = 1532 success for adults
n_1 = 2253 sample for adults
x_2 = 503 successes for teens
n_2 = 799 sample size for teens
CL = .90

We are 90% confident the interval from .01803 to .08287 captures the true difference in proportions of adults who use internet daily to teens who use the internet daily.

Significance Tests for $p_1 - p_2$

The null hypothesis will typically be no difference, $H_0: p_1 - p_2 = 0$

- The conditions for significance tests are same as for CI's.

Test Statistic

we will use z since $\hat{p}_1 - \hat{p}_2$ has approximately the normal sampling distribution

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}(H_0)}{\text{SD of statistic}}$$

$$= \frac{\hat{p}_1 - \hat{p}_2 - 0}{\text{SD of statistic}}$$

Standard Deviation of a statistic

If $H_0: p_1 - p_2 = 0$, then $p_1 = p_2$. Since the proportions are the same, we call them p. This is called the pooled sample proportion and is calculated by:

$$\hat{p}_c = \frac{\text{count of success of both samples combined}}{\text{sample size of both samples combined}}$$

$$= \frac{X_1 + X_2}{n_1 + n_2}$$

therefore the test statistic is:

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

Two sample z test for the difference of 2 proportions

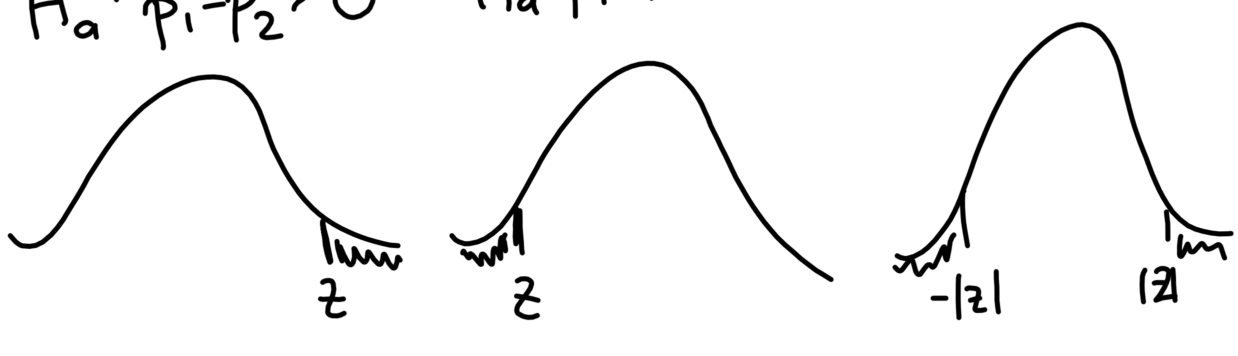
Suppose conditions are met. To test $H_0: p_1 - p_2 = 0$

find $\hat{p}_c$. Then compute

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

Find the p value by calculating the probability of getting a z statistic this large or larger in the direction of H_a.

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



In a study of 3000 randomly selected teenagers in 1988-1994, 15% showed some hearing loss. In a similar study of 1800 teens in 2005-2006, 19.5% showed some hearing loss. Do these data give convincing evidence the proportion of all teens with hearing loss has increased?

We are going to test the claim

$H_0: p_1 - p_2 = 0$

$H_a: p_1 - p_2 > 0$

with a 5% significance level where

p_1 = true proportion of teens in 2005-2006 with hearing loss

p_2 = true proportion of teens in 1988-1994 with hearing loss.

Random: Independent random samples from both populations.

10%: 3000 is less than 10% of all teen in '88-'94
1800 is less than 10% of all teens in '05-'06.

large counts: $.15(3000) = 450 \geq 10$ $(.85)(3000) = 2550 \geq 10$

$.195(1800) = 351 \geq 10$ $(.805)(1800) = 1449 \geq 10$

Use 2 sample z test for the difference in 2 proportions

on calc, 2 proportion z test:

$x_1 = 351$ success

$n_1 = 1800$ sample size

$x_2 = 450$ successes

$n_2 = 3000$ sample size

$p_1 > p_2$

$z = 4.05$

P value = .000025

Since our P value of .000025 < .05, we have convincing evidence to reject the H_0 meaning there is evidence that true proportion of teens with hearing loss has increased.