Lecture 13

Inference for Two Proportions

agenda

Reminders	Mid-Semester Survey 1 due Thursday.
Review	Hypothesis testing for one proportion
New stuff	Inference for two proportions
R activity	Hypothesis testing for two proportions in R

Hypothesis testing for two proportions

$p_1 = p_2$

1) Hypotheses $H_0 : p_1 - p_2 = 0$

$H_A : p_1 - p_2 < / > / \neq 0$

$p_1 \overset{<}{\underset{>}{\neq}} p_2$

2) Test conditions:
    - Independence
    - Large counts

$$n_1\hat{p}_1 \geq 5 \quad n_1(1-\hat{p}_1) \geq 5$$

$$n_2\hat{p}_2 \geq 5 \quad n_2(1-\hat{p}_2) \geq 5$$

3) Calculate test statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_2} - \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{\text{statistic} - \text{null val}}{\text{se}}$$

4) Compute p-value

```
<: pnorm(z)
>: pnorm(z, lower.tail = FALSE)
≠: 2*pnorm(-abs(z))
```

5) Decision and conclusion in context

Hypothesis testing for a difference in proportions

Two prop z-test

Goal: to compare proportions measured across two independent groups

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

Goal: learn about $p_1 - p_2$

Let $X_1, X_2, \cdots, X_{n_1} \overset{iid}{\sim} Bern(p_1)$ and $Y_1, Y_2, \cdots, Y_{n_2} \overset{iid}{\sim} Bern(p_2)$.

$$E[\hat{p}_1] = E\left[\frac{1}{n_1}\sum X_i\right] = \frac{1}{n_1}\sum_{i=1}^{n_1} E[X_i] = \frac{1}{n_1}\sum_{n_1} p_1 = p_1$$

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

$$var[\hat{p}_1] = var\left[\frac{1}{n_1}\sum_{i=1}^{n_1} X_i\right] = \frac{1}{n_1^2}\sum_{i=1}^{n_1} var[X_i] = \frac{1}{n_1^2}\sum_{i=1}^{n_1} p_1(1-p_1)$$

$$= \frac{\cancel{n_1}\, p_1(1-p_1)}{n_1^{\cancel{2}}}$$

$$= \frac{p_1(1-p_1)}{n_1}$$

$$\boxed{\mathrm{var}[X_1 + X_2] = \mathrm{var}[X_1] + \mathrm{var}[X_2] - 2\,\overset{0}{\cancel{\mathrm{cov}[X_1, X_2]}}}$$

$$\mathrm{var}[\hat{p}_1 - \hat{p}_2] = \mathrm{var}[\hat{p}_1] + \mathrm{var}[\hat{p}_2]$$

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

$$\boxed{\begin{array}{l}\mathrm{var}[X - Y] = \mathrm{var}[X] + \mathrm{var}[Y] \\ \mathrm{var}[X + (-1)Y]\end{array}}$$

Suppose we are doing a study to see if Vitamin C is helpful in preventing colds. We collect a sample of 279 participants and randomly assign 139 to receive a vitamin C supplement (treatment) while 140 receive a placebo (control). 17 of the participants in the treatment group and 31 of the participants in the control group contract a cold.

What is the parameter of interest? Define in symbols and interpret the parameter in context.

Let 1 = vitamin C group and 2 = control group.

$p_1 - p_2$ = true difference in proportions of those caught a cold in vit C vs./minus control grap

Hypothesis Testing

null

alternative

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

$$H_A: p_1 - p_2 \neq 0 \quad \begin{matrix} < \\ \\ > \end{matrix}$$

$H_0: p_1 - p_2 = 0$

$H_A: \underline{p_1 < p_2} \rightarrow p_1 - p_2 < 0$

Check Conditions

Independence

Observational units are randomly selected AND each group is independent of one another

1. Within : assume random ✓
2. Between : randomly assigned

Large counts

$$n_1\hat{p}_1 \geq 5 \quad n_1(1-\hat{p}_1) \geq 5$$

$$n_2\hat{p}_2 \geq 5 \quad n_2(1-\hat{p}_2) \geq 5$$

$$\left\{\begin{array}{l} n_1\hat{p}_1 = 139\left(\frac{17}{139}\right) = 17 \geq 5 \checkmark \\ n_1(1-\hat{p}_1) = 139\left(1-\frac{17}{139}\right) = 122 \geq 5 \checkmark \\ n_2\hat{p}_2 = 140\left(\frac{31}{140}\right) = 31 \geq 5 \checkmark \\ n_2(1-\hat{p}_2) = 140\left(\frac{31}{140}\right) = 109 \geq 5 \checkmark \end{array}\right.$$

$$z = \frac{\text{statistic - null value}}{\text{standard error of statistic}}$$

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

$\hat{p}_1 = 0.12$

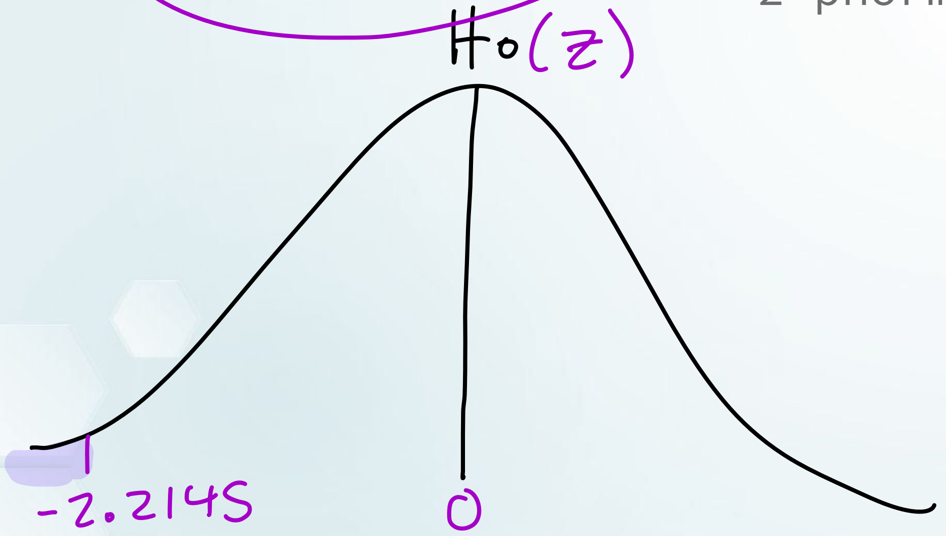
$\hat{p}_2 = 0.22$

$$z = \frac{0.12 - 0.22 - 0}{\sqrt{\frac{0.12(0.88)}{139} + \frac{0.22(0.78)}{140}}} = \boxed{-2.2145}$$

Calculate test statistic

Compute p-value

<: pnorm(z) | ≠: pnorm(z, lower.tail = FALSE) | >: 2*pnorm(-abs(z))



pnorm(-2.2145)

= 0.014

Conclusion in context

Reject the null: We have enough evidence that the difference in proportion of ________ is less than/different from/ more than 0.

Fail to reject the null: We do not have enough evidence that the difference in proportion of ________ is less than/different from/ more than 0.

Confidence intervals

statistic ± null value * se

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

$= (l, u)$

To R!

please download and open ht_two_prop.Rmd in the Week 9 module