Introductory Statistics – Day 20

Comparing Two Proportions

Open the NCBabySmoke data set from Moodle (from OpenIntro).

- Previously, we've worked with single proportions.
- Today, we want to compare proportions from two subgroups of a population and determine if there is a statistically significant difference between those proportions.

Is a difference in smoking rates among mature and younger moms.

- What should our null and alternative hypotheses be?

$$H_0 : p_{mature} = p_{younger}$$

or

$$H_0 : p_{mature} - p_{younger} = 0$$

$$H_A : p_{mature} - p_{younger} \neq 0$$

Use a pivot table in Excel to organize your data.

- Find the proportion of mature moms and younger moms who smoke, as well as the point estimate (the difference between them).

  $\hat{p}_{mature} = \frac{11}{133} = 0.0827$ and $\hat{p}_{younger} = \frac{115}{867} = 0.1326$

  The difference in rates of smoking based on age is

  $\hat{p}_{mature} - \hat{p}_{younger} = -0.0499.$

- What do we still need in order to answer our research question?

  $$SE_{pooled} = \sqrt{p_{pooled}(1 - p_{pooled})(\frac{1}{n_1} + \frac{1}{n_2})}$$

  where the pooled proportion is $p_{pooled} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$

  For this problem, $p_{pooled} = \dfrac{11 + 115}{133 + 867} = \dfrac{126}{999} = 0.126.$

  Therefore $SE_{pooled} = \sqrt{0.126(1 - 0.126)(\dfrac{1}{133} + \dfrac{1}{867})} = 0.0309$

- Let's find the p-value.

- Let's find the p-value.

$$2 \times norm.dist(-0.0499, 0, 0.0309, 1) = 2 \times 0.05317 = 0.1063$$

- Conclusion?

  With a p-value of 0.1063, we do not have sufficient evidence to say the difference we found was significant. We cannot claim that there is a significant difference in the rates of smoking between younger and more mature new moms.

Confidence Intervals with 2 proportions:

In order to follow-up with a confidence interval, we can use the same confidence interval strategy as before, but we will need a different formula for the standard error SE.
Confidence Interval:

$$\text{point estimate} \pm z_{critical} \times SE$$
$$(\hat{p_1} - \hat{p_2}) \pm \ z_{critical} \times SE$$

Why is the standard error that we used in our hypothesis test not appropriate for a confidence interval?

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

Comparing proportions

Use the NCBabySmoke data set to answer each of the questions on the slides for today.

column name	description and units
fage	father's age
mage	mother's age
mature	under 35 vs. 35 or older
weeks	length of pregnancy
premie	premie or full term
visits	number of doctor visits
marital	married or not married
gained	weight gained by mom (lbs)
weight	weight of baby (lbs)
lowbirthweight	low is ≤ 5.5 lbs
gender	baby's gender
habit	smoking habit of mom
whitemom	white or not white

Activity A. Is there a difference in smoking prevalence between new moms who are married and not married?

1. State the null and alternative hypotheses
2. Find and label the proportions ($p_{married}$ and $p_{notmarried}$) and the sample size $n_{married}$ and $n_{notmarried}$ using a pivot table in Excel.
3. Find the difference in proportions, the standard error, and the p-value. Label each in your Excel sheet. Note: You will need to use the pooled proportion in your calculations for SE because your null hypothesis is that there is no difference between the two proportions.
4. State your conclusions in a complete sentence related to the context of the problem.
5. Are you surprised by the conclusions? Did you expect something different?
6. If you rejected the null hypothesis, what should you do to follow-up?

Activity B. Is there a difference in rates of low weight babies between smoking moms and non-smoking moms?

1. State the null and alternative hypotheses
2. Find and label the proportions ($p_{smoking}$ and $p_{nonsmoking}$) and the sample size $n_{smoking}$ and $n_{nonsmoking}$.
3. Find the difference in proportions, the standard error, and the p-value. Label each in your Excel sheet.
4. State your conclusions in a complete sentence related to the context of the problem.
5. Are you surprised by the conclusions? Did you expect something different?
6. If you rejected the null hypothesis, what should you do to follow-up?

Activity C. Generate two more research questions that you could ask of this data. Choose one question which could be answered with a one proportion hypothesis test and choose a second question that requires a two proportion hypothesis test.