Comparing Proportions for Two Independent Samples: An Example

Brady T. West

Example: Comparing Proportions in Two Groups

Research Question:

Considering elderly Hispanic adults (age 80+) living in the U.S. in 2015-2016, did the **proportions of males and females who smoked vary significantly**?

Inference Approaches:

- Form a confidence interval for the difference in the two proportions
- Perform a chi-square test to test the significance of the difference in the two population proportions
- Be sure to check assumptions!

Approach 1: Form a Confidence Interval

- **Males**: Proportion = 0.565, n = 16 (small sample!)
- **Females**: Proportion = 0.250, n = 32

Approach 1: Form a Confidence Interval

Males: Proportion = 0.565, n = 16 (small sample!)
Females: Proportion = 0.250, n = 32

- **Best Point Estimate:** Difference in sample proportions is

$$0.565 - 0.250 = 0.315$$

- **Interpretation:** In 2015-2016, we estimate that the percentage of all **male** elderly Hispanics who smoked was 31.5 percentage points ***higher*** than for all **female** elderly Hispanics

Approach 1: Form a Confidence Interval

Males: Proportion = 0.565, n = 16 (small sample!)
Females: Proportion = 0.250, n = 32

Assumptions:

1. Is sampling distribution of difference in sample proportions normal? **Unlikely…**
2. Do we have at least 10 outcomes (smoker/non-smoker) in each group? **No…**
3. Are the two samples independent? **Yes**
4. Are the observations independent? **Yes, assumed for now…**

Approach 1: Form a Confidence Interval

- Calculate ***estimated* standard error of the difference:**

$$\text{Estimated SE} = \sqrt{\frac{0.565(1-0.565)}{16} + \frac{0.250(1-0.250)}{32}} = \mathbf{0.146}$$

- Determine the "few": a critical value for a 95% CI: **z = 1.96**
- Add and subtract margin of error from best estimate of the difference:

$$\mathbf{0.315 \pm 1.96\ x\ 0.146}$$

- 95% CI for difference in population proportion is **(0.027, 0.598)**
- Interval doesn't include 0 → **Significant difference!**

Approach 1: Check Robustness

Given concerns about some assumptions due to small sample sizes,
can compute an **exact 95% confidence interval**
for the difference in population proportions
(more computationally intensive)

Resulting exact 95% confidence interval: **(0.015, 0.574)**

Similar result, but evidence is not as overwhelming!

Approach 2: Chi-square Test

- **Null:** Equal population proportions of elderly male and female Hispanics who smoke
- **Alternative:** Males and Females have different population proportions who smoke

Alternative allows male proportion to be *either* greater or less than female proportion → **two-tailed test!** Need more evidence against null hypothesis to reject it!

Significance Level = 5%

Approach 2: Chi-square Test

Assumptions:

- Are all **expected counts** for each cell of the 2 x 2 table *under the null hypothesis* **greater than 5**?
- **Yes**: if overall sample rate of smokers, $\frac{17}{48} = 0.354$, is applied to each sample size of 16 (males) and 32 (females)

  → expect about 6 males and about 11 females to be smokers!
- ***Note: we assume independent observations for now!***

Approach 2: Chi-square Test

Resulting test statistic and p-value:

χ^2 = 4.554, df = 1, p-value = 0.033

Reject null hypothesis →

support the population proportions of smokers are **different**!

Note: if had initially selected 1% significance level
(Type I error rate = 0.01),
we would have **failed to reject** the null hypothesis!
Evidence is not overwhelming.

Chi-square Test: Check Robustness

Consider **Fisher's Exact Z Test** given small sample sizes…

p-value is 0.054 →

do not have overwhelming evidence against the null hypothesis

Overall Conclusion

Conclusion: We have weak evidence of a significant difference in the population proportions of smokers for elderly male and female Hispanics living in the U.S. in 2015-2016 (*small sample sizes and limited statistical power*)

Notes:

1. If same difference in proportions (0.315) were to emerge with larger sample sizes in each group, would likely find it significant, regardless of significance level → big difference in reality!
2. Need to allow for larger degree of uncertainty in our analysis due to small samples sizes.