Chapter 10 Comparing Two Proportions

A more common situation than making an inference about a single proportion is comparing two population proportions. Comparison is at the basis of many observational studies and experiments.

Example:

- A recent medical study focused on the relationship between estrogen use and premature death rates among post-menopausal women. Researchers searched the medical records of a large health care maintenance organization for women born between 1900 and 1915 who had taken estrogen supplements for at least one year starting in 1969. There were 232 such women and 53 of them had died prematurely (not defined) from all causes. The researchers also selected a sample of records of women born between 1900 and 1915 who had not taken estrogen supplements at all. There were 222 women in this sample and 87 of them had died prematurely from all causes.

- Assuming that these samples of women are representative samples from the populations of women born between 1900 and 1915 who did and did not take estrogen supplements, do these data provide evidence that the premature death rates for these two populations are different? How big might the difference in the two population death rates be?

The first of these questions is asking for a hypothesis test and the second is asking for a confidence interval for the difference between the population proportions. Since there are two populations, there are two parameters:

1. $p_1 = \text{proportion of estrogen takers born between 1900 and 1915 who died prematurely}$
2. $p_2 = \text{proportion of estrogen non-takers born between 1900 and 1915 who died prematurely}$

The sample proportions are

- $\hat{p}_1 = 53/232 = 0.2284$ and $\hat{p}_2 = 87/222 = 0.3919$

- We estimate the difference in the population proportions, $p_1 - p_2$, by the difference in the sample proportions,

$$\hat{p}_1 - \hat{p}_2 = 0.2284 - 0.3919 \approx -0.163$$

Confidence interval for the difference between two proportions

In order to form a confidence interval for $p_1 - p_2$ we need to know the sampling distribution of $\hat{p}_1 - \hat{p}_2$. What do we know about the sampling distributions of $\hat{p}_1$ and $\hat{p}_2$ individually? From Chapter 19:

$$E(\hat{p}_1) = p_1 \text{ and } SD(\hat{p}_1) = \sqrt{\frac{p_1q_1}{n_1}}$$

$$E(\hat{p}_2) = p_2 \text{ and } SD(\hat{p}_2) = \sqrt{\frac{p_2q_2}{n_2}}$$

In addition, if the sample sizes are large enough, we know that the sampling distributions of $\hat{p}_1$
and \( \hat{p}_2 \) are approximately normal.

What do these results tell us about the sampling distribution of \( \hat{p}_1 - \hat{p}_2 \)? Using the results from Chapter 16 on the mean and variance of the difference between two random variables:

\[
E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 \quad \text{and} \quad \text{Var}(\hat{p}_1 - \hat{p}_2) = \text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}.
\]

The result about the variance is valid only if \( \hat{p}_1 \) and \( \hat{p}_2 \) are independent. That means we assume that the two samples are drawn independently from their respective populations. We also know that the difference between two independent normal random variables is normal. So, if the sample sizes are large enough, the sampling distribution of \( \hat{p}_1 - \hat{p}_2 \) is approximately

\[
N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right).
\]

Making a confidence interval for the difference

- To make a confidence interval for \( p_1 - p_2 \), we will need to estimate \( \text{SD}(\hat{p}_1 - \hat{p}_2) \) because it depends on the unknown population proportions. We estimate it by

\[
\text{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}.
\]

- A confidence interval for \( p_1 - p_2 \) is then

\[
\hat{p}_1 - \hat{p}_2 \pm z^* \text{SE}(\hat{p}_1 - \hat{p}_2),
\]

where \( z^* \) is the critical value from the standard normal distribution for the level of confidence you have chosen.

- The sample size assumption is that the number of successes and number of failures is greater than 10 in both groups (individually).

- We also assume that these are random samples of less than 10% of the respective populations and that the samples were selected independently of each other.

**Example:** We want to know the true difference between the death rates of estrogen users and nonusers born between 1900 and 1915. We first check the conditions necessary for using the confidence interval procedure.

- **Randomization condition:** we assume that the two samples are like random samples from the populations of estrogen users and nonusers born between 1900 and 1915. We should be very careful of generalizing our results when all our subjects came from the records of one health care maintenance organization because we don’t know how representative these women are of women in general, even in this age group. In practice, we might go ahead and report our confidence interval, but then hope that other researchers do similar studies on other groups of subjects.

- **10% condition:** these samples are much less than 10% of the populations
• Independent samples condition: we assume that the samples were drawn independently of each other. These are not independent samples from the general population of women born between 1900 and 1915 because the two samples are from the same health care maintenance organization. We have to assume that the difference between the two groups at this organization is representative of the difference in the larger populations.

• Success/failure condition: There are more than 10 “successes” and 10 “failures” in each sample so this condition is satisfied.

A 95% confidence interval for the difference between the premature death rates of the populations of estrogen users and nonusers is

\[
\hat{p}_1 - \hat{p}_2 \pm z^\ast \text{SE}(\hat{p}_1 - \hat{p}_2) = 0.2284 - 0.3919 \pm 1.96 \sqrt{\frac{0.2284(0.7716) + 0.3919(0.6081)}{232} + \frac{0.2284(0.7716) + 0.3919(0.6081)}{222}} \\
= -0.163 \pm 1.96(0.0428) = -0.163 \pm 0.084 = -0.247 \text{ to } -0.080
\]

Conclusion: We are 95% confident that the premature death rate among all estrogen users born between 1900 and 1915 is from about .25 to .08 lower than for non-users.

Hypothesis test for two proportions

• The natural hypothesis to test when you’re comparing two groups is whether or not the difference in the population proportions is zero. So the null hypothesis is

\[H_0: p_1 - p_2 = 0 \quad \text{(or, equivalently, } H_0: p_1 = p_2)\]

• The alternative hypothesis can be one-sided or two sided:

\[H_A: p_1 - p_2 > 0 \quad \text{(or } H_A: p_1 > p_2)\]

\[H_A: p_1 - p_2 < 0 \quad \text{(or } H_A: p_1 < p_2)\]

\[H_A: p_1 - p_2 \neq 0 \quad \text{(or } H_A: p_1 \neq p_2)\]

To carry out the test, we need to know the sampling distribution of \(\hat{p}_1 - \hat{p}_2\) when \(H_0\) is true; that is, when \(p_1 = p_2\). If we let \(p\) denote the common value of \(p_1\) and \(p_2\) when \(H_0\) is true then

\[E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 = 0 \quad \text{and } \text{SD}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}\]

Unfortunately, \(\text{SD}(\hat{p}_1 - \hat{p}_2)\) depends on \(p\) and the null hypothesis doesn’t specify what the common value of \(p_1\) and \(p_2\) is. Hence, we must estimate it from the data. Since we’re assuming the two populations have the same \(p\), it makes sense to estimate it by pooling the two samples together:

\[\hat{p}_{\text{pooled}} = \frac{\text{Success}_1 + \text{Success}_2}{n_1 + n_2}\]

to give us
The test statistic is then
\[ z = \frac{\hat{p}_1 - \hat{p}_2}{\text{SE}_{\text{pooled}}(\hat{p}_1 - \hat{p}_2)} \]

The P-value is computed just as for the one-proportion z-test.

**Example:** In the estrogen study, do the data provide evidence that there is a difference in the premature death rates of the two populations?

- We are testing \( H_0 : p_1 - p_2 = 0 \) versus \( H_A : p_1 - p_2 \neq 0 \).
- We already checked the conditions when we calculated the confidence interval. We have to be careful about how far we generalize our conclusion from this test.
- The pooled sample proportion is
  \[ \hat{p}_{\text{pooled}} = \frac{53 + 87}{232 + 222} = \frac{140}{454} = 0.3084. \]

So,
\[ \text{SE}_{\text{pooled}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{0.3084(0.6916)}{232} + \frac{0.3084(0.6916)}{222}} = 0.04282 \]

- The test statistic is
  \[ z = \frac{0.2284 - 0.3919}{0.04282} = -3.82. \]
- The P-value is
  \[ \text{P-value } 2P(Z < -3.82) = 2(0.0001) = 0.0002. \]

There’s a .0002 probability of observing a sample difference this big or bigger if the premature death rates of estrogen users and nonusers were equal.

**Conclusion:** The data provide very strong evidence that the premature death rate is lower among estrogen users than among nonusers.

**Question:** Can we conclude from this study that estrogen helps prevent premature death?