Section 11.1: Inference about Two Proportions

The Difference Between Two Population Proportions

In Section 8.2, we discussed the distribution of one sample proportion, $\hat{p}$. What we'll need to do now is develop some similar theory regarding the distribution of the difference in two sample proportions, $\hat{p}_1 - \hat{p}_2$.

> **The Sampling Distribution of the Difference between Two Proportions**
>
> Suppose simple random samples size n_1 and n_2 are taken from two populations. The distribution of $\hat{p}_1 - \hat{p}_2$ where $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$, is aproximately normal with mean $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$ and standard deviation
> $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$, provided:
>
> 1. $n_1\hat{p}_1(1-\hat{p}_1) \geq 10$
> 2. $n_2\hat{p}_2(1-\hat{p}_2) \geq 10$
> 3. both sample sizes are less than 5% of their respective populations.
>
> The standardized version is then
>
> $$z_0 = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$
>
> which has an approximate standard normal distribution.

The thing is, in most of our hypothesis testing, the null hypothesis assumes that the proportions are the same ($p_1 = p_2$), so we can call $p = p_1 = p_2$.

Since $p_1 = p_2$, we can substitute 0 for $p_1 - p_2$, and substitute p for both p_1 and p_2. In that case, we can rewrite the above standardized z the following way:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{(\hat{p}_1 - \hat{p}_2) - (0)}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}} = \frac{\hat{p}_1 - \hat{p}_2}{p(1-p)\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Which leads us to our hypothesis test for the difference between two proportions.

Performing a Hypothesis Test Regarding $p_1 - p_2$

> ***Step 1*:** State the null and alternative hypotheses.
>
> | **Two-Tailed** | **Left-Tailed** | **Right-Tailed** |
> |---|---|---|
> | H_0: $p_1 - p_2 = 0$ | H_0: $p_1 - p_2 = 0$ | H_0: $p_1 - p_2 = 0$ |
> | H_1: $p_1 - p_2 \neq 0$ | H_1: $p_1 - p_2 < 0$ | H_1: $p_1 - p_2 > 0$ |
>
> ***Step 2*:** Decide on a level of significance, α.
>
> ***Step 3*:** Compute the test statistic, $z = \frac{\hat{p}_1 - \hat{p}_2}{\hat{p}(1-\hat{p})\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$.
>
> ***Step 4*:** Determine the *P*-value.
>
> ***Step 5*:** Reject the null hypothesis if the *P*-value is less than the level of significance, α.
>
> ***Step 6*:** State the conclusion.

Hypothesis Testing Regarding $p_1 - p_2$ Using StatCrunch

> 1. Go to **Stat** > **Proportions** > **Two Sample** > **with summary**
> 2. Enter the number of successes and observations for each sample and press **Next**.
> 3. Set the null proportion difference and the alternative hypothesis.
> 4. Click on **Calculate**.
>
> The results should appear.

Example 1

Problem: Suppose a researcher believes that college faculty vote at a higher rate than college students. She collects data from 200 college faculty and 200 college students using simple random sampling. If 167 of the faculty and 138 of the students voted in the 2008 Presidential election, is there enough evidence at the 5% level of significance to support the researcher's claim?

Solution:

First, we need to check the conditions. Both sample sizes are clearly less than 5% of their respective populations. In addition,

$$n_1\hat{p}_1(1-\hat{p}_1) = 200\left(\frac{167}{200}\right)\left(1-\frac{167}{200}\right) \approx 27.6 \geq 10$$

$$n_2\hat{p}_2(1-\hat{p}_2) = 200\left(\frac{138}{200}\right)\left(1-\frac{138}{200}\right) \approx 42.8 \geq 10$$

So our conditions are satisfied.

***Step 1*:**

Let's take the two portions in the order we receive them, so
$p_1 = p_f$ (faculty) and $p_1 = p_s$ (students)

Our hypotheses are then:
H_0: $p_f - p_s = 0$
H_1: $p_f - p_s > 0$ (since the researcher claims that faculty vote at a *higher* rate)

***Step 2*:** $\alpha = 0.05$ (given)

***Step 3*:** (we'll use StatCrunch)

***Step 4*:** Using StatCrunch:

Hypothesis test results:
p_1 : proportion of successes for population 1
p_2 : proportion of successes for population 2
$p_1 - p_2$: difference in proportions
$H_0 : p_1 - p_2 = 0$
$H_A : p_1 - p_2 > 0$

Difference	Count1	Total1	Count2	Total2	Samp	Z-Stat	P-value
$p_1 - p_2$	167	200	138	200		3.4073462	0.0003

(Trimmed to fit on this page.)

***Step 5*:** Since the *P*-value $< \alpha$, we reject the null hypothesis.

***Step 6*:** Based on these results, there is very strong evidence (certainly enough at the 5% level of significance) to support the researcher's claim

Confidence Intervals about the Difference Between Two Proportions

We can also find a confidence interval for the difference in two population proportions.

In general, a **(1-α)100% confidence interval for p_1-p_2** is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Note: The following conditions must be true:

1. $n_1\hat{p}_1(1-\hat{p}_1) \geq 10$
2. $n_2\hat{p}_2(1-\hat{p}_2) \geq 10$
3. both sample sizes are less than 5% of their respective populations.

Confidence Intervals About p_1-p_2 Using StatCrunch

1. Go to **Stat** > **Proportions** > **Two Sample** > **with summary**
2. Enter the number of successes and observations for each sample and press **Next**.
3. Click the Confidence Interval button, and set the confidence level.
4. Click on **Calculate**.

Example 2

Problem: Considering the data from Example 1, find a 99% confidence interval for the difference between the proportion of faculty and the proportion of students who voted in the 2008 Presidential election.

Solution: From Example 1, we know that the conditions for performing inference are met, so we'll use StatCrunch to find the confidence interval.

99% confidence interval results:
p_1 : proportion of successes for population 1
p_2 : proportion of successes for population 2
$p_1 - p_2$: difference in proportions

Difference	Count1	Total1	Count2	Total2	.	L. Limit	U. Limit
$p_1 - p_2$	168	200	137	20	346	0.047218394	0.26278162

(Timmed to fit on this page.)

So we can say that we're 99% confident that the difference between the proportion of faculty who vote and the proportion of students who vote is between 3.7% and 25.3%.

Determining the Sample Size Needed

In Section 9.3, we learned how to find the necessary sample size if a specific margin of error is desired. We can do a similar analysis for the difference in two proportions. From the confidence interval formula, we know that the margin of error is:

$$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

The **sample size required** to obtain a (1-α)100% confidence interval for p_1-p_2 with a margin of error *E* is:

$$n = n_1 = n_2 = [\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)]\left(\frac{z_{\alpha/2}}{E}\right)^2$$

rounded up to the next integer, if $\hat{p}_1$ and $\hat{p}_2$ are esimates for p_1 and p_2, respectively.

If no prior estimate is available, use $\hat{p}_1 = \hat{p}_2 = 0.5$, which yields the following formula:

$$n = n_1 = n_2 = 0.5\left(\frac{z_{\alpha/2}}{E}\right)^2$$

again rounded up to the next integer.

Note: As in Section 9.3, the desired margin of error should be expressed as a decimal.

Example 3

Suppose we want to study the success rates for students in Mth098 Intermediate Algebra at ECC. We want to compare the success rates of students who place directly into Mth098 with those who first took Mth096 Beginning Algebra. From past experience, we know that a typical success rate for students in this class is about 65%. How large of a sample size is necessary to create a 95% confidence interval for the difference of the two passing rates with a maximum error of 2%?

[reveal answer]

$$n = n_1 = n_2 = [0.65(1-0.65) + 0.65(1-0.65)]\left(\frac{1.96}{0.02}\right)^2$$

$$\approx 4369.82 \rightarrow 4370$$

So we would need a sample size of 4,370 students - from each population!