Stat 250 Gunderson Lecture Notes
6: Learning about the Difference in Population Proportions
Part 1: Distribution for a Difference in Sample Proportions

The Independent Samples Scenario

Two samples are said to be **independent samples** when the measurements in one sample are not related to the measurements in the other sample. Independent samples are generated in a variety of ways. Some common ways:

- **Random samples are taken separately from two populations** and the same response variable is recorded for each individual.
- **One random sample** is taken and a variable is recorded for each individual, but then **units are categorized as belonging to one population or another**, e.g. male/female.
- **Participants are randomly assigned to one of two treatment conditions**, and the same response variable, such as weight loss, is recorded for each individual unit.

If the **response variable is categorical**, a researcher might compare two independent groups by looking at the **difference between the two proportions**.

There are usually two questions of interest about a difference in two population proportions. First, we want to estimate the value of the difference. Second, often we want to test the hypothesis that the difference is 0, which would indicate that the two proportions are equal. In either case, we will need to know about the sampling distribution for the difference in two sample proportions (from independent samples).

Sampling Distribution for the Difference in Two Sample Proportions

Example: Driving Safely

Question of interest: How much of a difference is there between men and women with regard to the proportion who have driven a car when they had too much alcohol to drive safely?

Study: Time magazine reported the results of a poll of adult Americans. One question asked was: **"Have you ever driven a car when you probably had too much alcohol to drive safely?"**

Let p_1 be the **population proportion** of **men** who would respond yes.

Let p_2 be the **population proportion** of **women** who would respond yes.

We want to learn about p_1 and p_2 and how they compare to each other. We could estimate the difference $p_1 - p_2$ with the corresponding difference in the sample proportions $\hat{p}_1 - \hat{p}_2$.

Will it be a good estimate? How close can we expect the difference in sample proportions to be to the true difference in population proportions (on average)?

Imagine repeating the study many times, each time taking two independent random samples of sizes n_1 and n_2, and computing the value of $\hat{p}_1 - \hat{p}_2$. What kind of values could you get for $\hat{p}_1 - \hat{p}_2$? What would the distribution of the possible $\hat{p}_1 - \hat{p}_2$ values look like? What can we say about the **distribution of the** difference in two sample proportions?

Using results about how to work with differences of independent random variables and recalling the form of the sampling distribution for a sample proportion, the sampling distribution of the difference in two sample proportions $\hat{p}_1 - \hat{p}_2$ can be determined.

First recall that when working with the difference in two independent random variables:

- the mean of the difference is just the difference in the two means
- the variance of the difference is the sum of the variances

Next, remember that the standard deviation of a sample proportion is $\sqrt{\frac{p(1-p)}{n}}$.

So what would the *variance* of a single sample proportion be?

So let's apply these ideas to our newest parameter of interest, the difference in two sample proportions $\hat{p}_1 - \hat{p}_2$.

Sampling Distribution of the Difference in Two (Independent) Sample Proportions

If the two sample proportions are based on independent random samples from two populations and if all of the quantities $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1-\hat{p}_2)$ are at least 10,

Then the distribution for the possible $\hat{p}_1 - \hat{p}_2$ will be (approximately) …

Since the population proportions of p_1 and p_2 are not known, we will use the data to compute the standard error of the difference in sample proportions.

Standard Error of the Difference in Sample Proportions

$$\text{s.e.}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

The standard error of $\hat{p}_1 - \hat{p}_2$ estimates, roughly, the average distance of the possible $\hat{p}_1 - \hat{p}_2$ values from $p_1 - p_2$. The possible $\hat{p}_1 - \hat{p}_2$ values result from considering all possible independent random samples of the same sizes from the same two populations.

Moreover, we can use this standard error to produce a range of values that we can be quite confident will contain the difference in the population proportions $p_1 - p_2$:

$$\hat{p}_1 - \hat{p}_2 \pm \text{(a few)s.e.}(\hat{p}_1 - \hat{p}_2).$$

This is the basis for confidence interval for the difference in population proportions discussed next in Part 2.

If we are interested in testing hypotheses about the difference in the population rates, we will need to construct a null standard error of the difference in the sample proportions and use it to compute a standardized test statistic. That test statistic will have the following basic form:

$$\frac{\text{Sample statistic} - \text{Null value}}{\text{(Null) standard error}}$$

This is the basis for the hypothesis testing about the difference in population proportions covered in Part 3 of this section of notes.

Additional Notes

A place to … jot down questions you may have and ask during office hours, take a few extra notes, write out an extra problem or summary completed in lecture, create your own summary about these concepts.

Stat 250 Gunderson Lecture Notes
6: Learning about the Difference in Population Proportions

Part 2: Confidence Interval for a Difference in Population Proportions

We have **two populations** from which independent samples are available, (or one population for which two groups formed using a categorical variable). The response variable is also **categorical** and we are interested in comparing the proportions for the two populations.

- Let p_1 be the population proportion for the first population.
- Let p_2 be the population proportion for the second population.

Parameter: the difference in the population proportions $p_1 - p_2$.

Sample estimate: the difference in the sample proportions $\hat{p}_1 - \hat{p}_2$.

Standard error: $\text{s.e.}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

So we have our estimate of the difference in the two population proportions, namely $\hat{p}_1 - \hat{p}_2$, and we have its standard error. To make our confidence interval, we need to know the multiplier.

Sample Estimate ± Multiplier x Standard error

As in the case for estimating one population proportion, we assume the sample sizes are sufficiently large so the multiplier will be a z^* value found from using the standard normal distribution.

Two Independent-Samples *z* Confidence Interval for $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \text{s.e.}(\hat{p}_1 - \hat{p}_2)$$

where $\text{s.e.}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

and z^* is the appropriate multiplier from the N(0,1) distribution.

This interval requires that the sample proportions are based on independent random samples from the two populations.

Also, all of the quantities $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1-\hat{p}_2)$ be preferably at least 10.

Try It! Do Older People Snore More than Younger?

Researchers at the National Sleep Foundation were interested in comparing the proportion of people who snore for two age populations (1 = older adults defined as over 50 years old and 2 = younger adults defined as between 18 and 30 years old). The following data was obtained from adults who participated in a sleep lab study.

	"Snore?"		
Group	**Yes**	**No**	**Total**
1 = older adults (over 50 years old)	**168**	**312**	**480**
2 = younger adults (between 18 and 30 years old)	**45**	**135**	**180**

Let p_2 represent the population proportion of all younger adults who snore. Provide an estimate for this population proportion p_2. Include the appropriate symbol.

We wish to provide a 90% confidence interval to estimate the difference in snoring rates for the two population proportions of adults. One of the conditions for that confidence interval to be valid involves having two independent random samples, which is reasonable from the design of the study. Validate the remaining assumption.

Provide the 90% confidence interval and give an interpretation of this interval in context.

Interpretation this interval.
With 95% confidence we estimate the difference in snoring rates for the two population

of adults to be somewhere between _____ and _____.

What value do you notice is *not* in this interval? _____
Does there appear to be a significant difference between
the population rates of snoring for older versus younger adults? **Yes** **No**

Stat 250 Gunderson Lecture Notes
6: Learning about the Difference in Population Proportions

Part 3: Testing about a Difference in Population Proportions

Testing Hypotheses about the Difference in Two Population Proportions

We have two populations from which independent samples are available, (or one population for which two groups can be formed using a categorical variable). The response variable is also **categorical** and we are interested in comparing the proportions for the two populations.

- Let p_1 be the population proportion for the first population.
- Let p_2 be the population proportion for the second population.

Parameter: the difference in the population proportions $p_1 - p_2$.

Sample estimate: the difference in the sample proportions $\hat{p}_1 - \hat{p}_2$.

Standard error of $\hat{p}_1 - \hat{p}_2$: $\text{s.e.}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Recall that the multiplier in the confidence interval was a *z** value. So we will be computing a *Z* test statistic for performing a significance test.

The standard error used in constructing the confidence interval for the difference between two population proportions is not the same as that used for the standardized z test statistic.

We will need to construct the null standard error, the standard error for the statistic when the null hypothesis is true. Let's start with what the hypotheses will look like.

Possible null and alternative hypotheses.

1. **H_0:** **versus H_a:**

2. **H_0:** **versus H_a:**

3. **H_0:** **versus H_a:**

Next we need to determine the test statistic and understand the conditions required for the test to be valid. The general form of the test statistic is:

$$\textbf{Test statistic} = \frac{\textbf{Sample statistic} - \textbf{Null value}}{\textbf{Standard error}}$$

In the case of two population proportions, if the null hypothesis is true, we have $p_1 - p_2 = 0$ or that the two population proportions are the same, $p_1 = p_2 = p$. What is a reasonable way to **estimate the common population proportion *p***?

The general standard error for $\hat{p}_1 - \hat{p}_2$ is given by:

$$\text{s.e.}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

but if the null hypothesis is true, then $\hat{p}$ is the best estimate for each population proportion and should be used in the standard error.

So, the **null standard error** for $\hat{p}_1 - \hat{p}_2$ is given by:

And the corresponding test statistic is:

If the null hypothesis is true, this *z*-statistic will have a _____ distribution. This distribution is used to find the *p*-value for the test.

Conditions: This test requires that the sample proportions are based on independent random samples from the two populations. Also, all of the quantities $n_1\hat{p}$, $n_1(1-\hat{p})$, $n_2\hat{p}$, and $n_2(1-\hat{p})$ be preferably at least 10. Note these are checked with the estimate of the common population proportion $\hat{p}$.

Try It! Taking More Pictures with Cell

Cell phones can now be used for many purposes besides making calls. An initial study found that more than 75% of *young adults* (defined as 18-25 years old) use their cell phones for taking pictures at least 2 times per week. This study also suggested that the proportion of young women in this age group who use their cell phone to take pictures is higher than that for young men in this age group. A follow-up study was conducted to investigate this conjecture. The researchers which to use a 5% significance level.

Stated the hypotheses: H_0: _____ versus H_a: _____where

p_1 represents the population proportion of all young women 18-25 years old who report using their cell phone to take pictures at least 2 times per week, and

p_2 represents the population proportion of all young men 18-25 years old who report using their cell phone to take pictures at least 2 times per week.

Here are the results:

Age group = 18 – 25 year olds	**Young Women**	**Young Men**
Number who report using phone to take pictures at least 2 times/week	417	369
Sample Size	521	492
Percent	80%	75%

We can assume these samples are independent random samples. Verify the remaining condition necessary to conduct the Z test.

Conduct the test.

Using a 5% significance level which is the appropriate conclusion?

- There is sufficient evidence to demonstrate the population proportion of all young women 18-25 years old who take pictures with their phone at least twice per week is greater than that of the population of all young men 18-25 years old.
- There is not sufficient evidence to demonstrate the population proportion of all young women 18-25 years old who take pictures with their phone at least twice per week is greater than that of the population of all young men 18-25 years old.

Additional Notes

A place to … jot down questions you may have and ask during office hours, take a few extra notes, write out an extra problem or summary completed in lecture, create your own summary about these concepts.

Two Population Proportions

Parameter $p_1 - p_2$

Statistic $\hat{p}_1 - \hat{p}_2$

Standard Error

$$\text{s.e.}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Confidence Interval

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \text{s.e.}(\hat{p}_1 - \hat{p}_2)$$

Large-Sample z-Test

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$