# Chapter 3: Inference for Two Proportions

## The Goal:

To evaluate the association between two categorical (binary) variables.

Is there a significant difference between two populaiton proportions $(p_1 - p_2)$?

This often leads to hypothesis formulations we have already seen with the randomization tests:

$H_0 : p_1 - p_2 = 0$

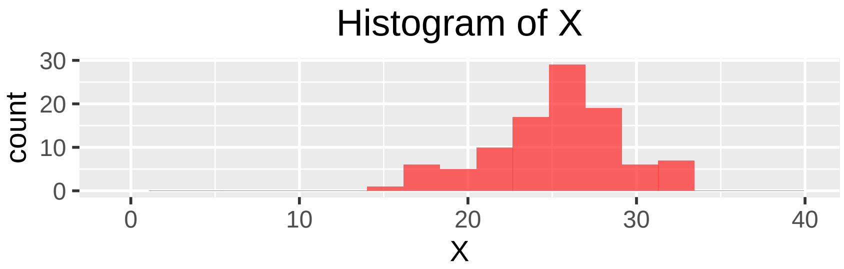
$H_a : p_1 - p_2 \neq 0$ (or $>$ or $<$)

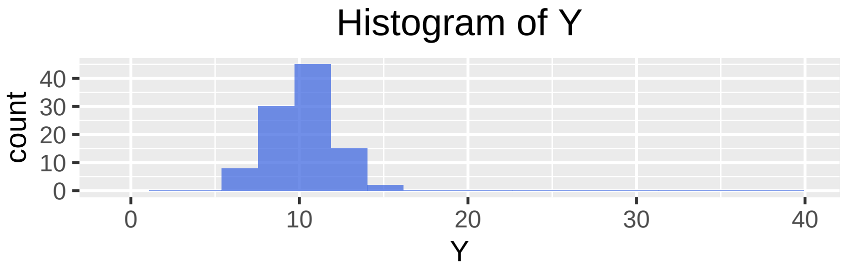
### The Approach

Just like last time, we will use the Central Limit Theorem to understand the sampling distribution for the differences in sample proportions $(\hat{p}_1 - \hat{p}_2)$.

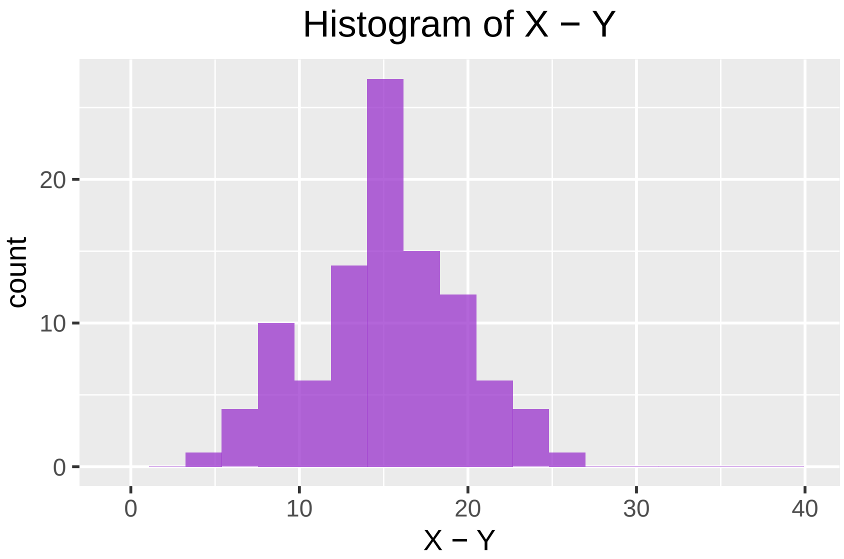
How can we know what the possible values will be for the difference in two statistics? Well, it turns out that if both of those statistics are Normally distributed, then their difference will be Normally distributed.

Consider `X` which is 100 values simulated from a Normal population $N(25, 4)$. And also `Y` which is 100 values simulated from a Normal population $N(10, 2)$. Below are their histograms.





If we take the difference of `X - Y`, we know that the differences will form a Normal distribution as well.



Not only do we know the shape, but we also know what to expect for the center. The center of `X - Y` will be the difference in the individual centers.

```
mean(X)
```

```
## [1] 25.34306
```

```
mean(Y)
```

```
## [1] 10.11779
```

Even more amazing is the fact that we can understand the spread of `X - Y` just by knowing the individual standard deviations. But in the case of spread, since we are incorporating a second population, we *increase* the amount of variation we observe (even though we are taking a difference). Notice in the histograms above that the purple plot seems to be more spread out than both the blue and red ones.

### Always Check the Conditions

This only works if the Central Limit Theorem applies for each sample. So we need to check:

- **Sample size**
  - are there at least 10 successes and 10 failures in *each* group?
- **Independence**
  - is it reasonable to assume each participant/case is independent of the others?
  - is the sample less than 10% of the total population (when the sample is too large it's harder to expect everyone to be independent from each other.)

## For Samples (Extra topic)

### Extra topic on the theory behind the sampling distribution

The **point estimate** for the difference in two population proportions ($p_1 - p_2$) is the difference in sample proportions.

Point estimate: $\hat{p_1} - \hat{p_2}$

The **standard error** for the difference in two population proportions is a combination of the individual spreads:

For **confidence intervals**: $SE_{p_1-p_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

When we are thinking about hypothesis tests, we need to assume that the two groups we are comparing have the same proportion (like assuming they are from the same population). So we combine our two samples into one *pooled proportion.*

$$\hat{p}_{pooled} = \frac{X_1 + X_2}{n_1 + n_2}$$

where $X_1$ and $X_2$ are the number of successes in each sample respectively.

For **hypothesis tests**: $SE_{\hat{p}_{pooled}} = \sqrt{\frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_1} + \frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_2}}$

## Example - Melting Ice Cap

The General Social Survey asked the following question to U.S. adults in 2010:

*Scientists predict that global warming may have big effects on the polar regions within the next 100 years. One of the possible effects is that the northern ice cap may completely melt. Would this bother you a great deal, some, a little, or not at all if it actually happened?*

a) A great deal

b) Some

c) A little

d) Not at all

The same question was also posed to 105 Duke University students in an introductory stats class.

### Results (Get the data)

Below is the data from the two groups that answered this question:

| | **U.S. adults** | **Duke** |
|---|---|---|
| A great deal | 454 | 69 |
| Some | 124 | 30 |
| A little | 52 | 4 |
| Not at all | 50 | 2 |
| **Total** | 680 | 105 |

### Give data context

If we are concerned with the proportion of people who are concerned a great deal by the melting ice caps, think about these questions:

- What is the **Research Question**?
- What is the **response variable**?
- What is the **explanatory variable** (if there is one)?

- How can I represent this question as a relationship between these two variables?

  How could you represent this question as a relationship between these two variables?

- **Parameter of interest**: Difference between the proportions of *all* Duke students and *all* U.S. adults who would be bothered a great deal by the northern ice cap completely melting.

$$p_{Duke} - p_{US}$$

Our hypotheses:

$H_0$ : There is no difference in the proportion of all Duke students and all U.S. adults who are greatly concerned in the melting ice cap (they are *independent*).

$$p_{Duke} - p_{US} = 0$$

$H_a$ : There is a significant difference in the proportion of all Duke students and all U.S. adults who are greatly concerned in the melting ice cap (they are *dependent*).

$$p_{Duke} - p_{US} \neq 0$$

### The sample evidence

First let's tidy up the table from above. We want the number of people who are greatly concerned and who aren't. (Sometimes it's useful to force variables into a binary form).

| | **U.S. adults** | **Duke** |
|---|---|---|
| A great deal | 454 | 69 |
| Not a great deal | 226 | 36 |
| **Total** | 680 | 105 |

Remember that the sample will be our starting point to evaluate the null hypothesis claim that the proportions are equal.

So we can calculate the difference in the sample proportions as:

$$\hat{p}_{Duke} - \hat{p}_{US} = 0.657 - 0.668 = -0.011$$

(Remember it's OK that this difference is negative, it just implies that the second group has a larger proportion.)

### Evaluate the evidence

We will do this in R. If you'd like to know how to perform the calculations by hand, you can look back at the extra topics section above to find the pooled proportion ($\hat{p}_{pooled}$), the standard error ($SE_{\hat{p}_{pooled}}$), and calculate the test statistic by finding the z-score of our sample difference.

Good news! we get to use the same function as we did for a single proportion test, `prop.test`. Only this time, we have two samples. So two "number of successes", `x` and two "sample sizes", `n`.

Fill in the values for `x` and `n` below. Remember to keep the same order (Duke, U.S.). Also fill in the `alt =` base on our alternative hypothesis.

```
prop.test(x = c(___, ___), n = c(___, ___), alt = "_____", correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c out of c69 out of 105454 out of 680
## X-squared = 0.045133, df = 1, p-value = 0.8318
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1079538  0.0869454
## sample estimates:
##    prop 1    prop 2
## 0.6571429 0.6676471
```

### Read the output (Interpret)

Once again, from the output, we are interested in the **p-value** as a way to measure the strength of our evidence. In this case, `p-value = 0.8318`. This can be interpreted as meaning, when we assume that there is no difference between Duke students and U.S. adults in the proportion of those that care a great deal about the melting ice cap, we would expect to observe the difference in survey results we have, or a larger difference, about 83% of the time. Since this means our survey is *not unusual* under the null hypothesis, we **fail to reject**. Our sample doesn't show that one group is more concerned than the other.

### Bonus interpretation

Notice the confidence interval we get with the output from `prop.test`.

```
95 percent confidence interval: -0.1079538  0.0869454
```

From this we can also say that we are 95% confident the true difference in the proportions of all Duke students and all U.S. adults is somewhere between -11% and 9%. In other words, we can't really be sure which group has a larger proportion, just like we concluded in the hypothesis test.

> When dealing with the difference in two parameters, always be on the lookout for whether **zero** is in your interval.

| Interval "contains" zero | Zero is "outside" of interval |
|---|---|
| We can't say there is a significant difference<br>Either group could have the larger proportion | There is evidence of a significant difference<br>Can determine which group is larger by the sign (+/-) of upper and lower bound |

## Another Example - Texting and Flu Vaccines

A 2012 JAMA article (Stockwell et al.) reported the results of a randomized controlled study "to evaluate targeted text reminders for low-income, urban parents to promote receipt of influenza vaccination among children and adolescents." They found that "a higher proportion of children and adolescents in the intervention group (43.6%; n=1653) compared with the usual care group (39.9%; n=1509) had received influenza vaccine (difference, 3.7% [95% CI, 1.5% - 5.9%]; relative rate ratio [RRR], 1.09 [95% CI, 1.04 - 1.15]; P=.001)." Data relevant to this study is stored as *textflu.csv*, and R code for data analysis is stored as *TwoProportions.Rmd*

### Does texting increase vaccination rates?

Try to answer these on your own first.

- What is the **Research Question**?
- What is the **response variable**?
- What is the **explanatory variable** (if there is one)?
- How can I represent this question as a relationship between these two variables?

- What is the **Research Question**?
  - Does sending a text message reminder change the proportion of people who will get their children vaccinated?
- What is the **response variable**?
  - Vaccine (Categorical: Yes/No)
- What is the **explanatory variable** (if there is one)?
  - Text Intervention (Categorical: Intervention/Usual care)
- How can I represent this question as a relationship between these two variables?

$H_0 : p_{text} - p_{usual} = 0$

$H_0 : p_{text} - p_{usual} \neq 0$

### Sample information

| | **Text Intervention** | **Usual care** |
|---|---|---|
| $\hat{p}$ | 0.436 | 0.399 |
| $n$ | 1653 | 1509 |

### Check the conditions

- Sample size: everything looks good
  - $n(\hat{p}_{text}) = 1653(0.436) > 10$
  - $n(1 - \hat{p}_{text}) = 1653(1 - 0.436) > 10$
  - $n(\hat{p}_{usual}) = 1653(0.399) > 10$
  - $n(\hat{p}_{usual}) = 1653(1 - 0.399) > 10$
- Independence:
  - reasonable to assume subjects are independent from one another
  - the samples are large, but we are certainly under 10% of the population

### Run the test in R

In this case, the data are available in a .csv file. How can we use raw data with `prop.test`? First read in the data as before.

```
flu <- read.csv("~/Stats 212b S20/Class/Data/textflu.csv")
```

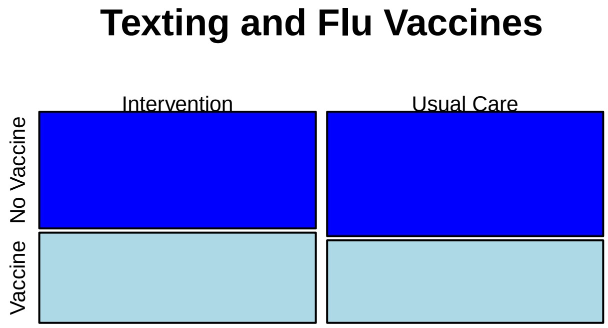
Check out the variable names and the first couple observations.

```
## [1] "Group"  "Result"
##          Group  Result
## 1 Intervention Vaccine
## 2 Intervention Vaccine
## 3 Intervention Vaccine
## 4 Intervention Vaccine
## 5 Intervention Vaccine
## 6 Intervention Vaccine
```

Do some EDA for two categorical variables…

```
##
##                No Vaccine Vaccine
##   Intervention       2137    1653
##   Usual Care         2275    1509
```

```
##
##                No Vaccine Vaccine  Sum
##   Intervention       2137    1653 3790
##   Usual Care         2275    1509 3784
##   Sum                4412    3162 7574
```



```
##
##                No Vaccine   Vaccine
##   Intervention  0.5638522 0.4361478
##   Usual Care    0.6012156 0.3987844
```

```
## [1] 0.0373634
```

From the tables above we can get the number of successes and the sample size. Remember to put the numbers in the order you want (`Text intervention`, `Usual care`) for the outcome you're interested in (`Vaccine`).

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c out of c1653 out of 37901509 out of 3784
## X-squared = 10.87, df = 1, p-value = 0.0009776
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.0151675 0.0595593
## sample estimates:
##    prop 1    prop 2
## 0.4361478 0.3987844
```

**Compare to the report**

*They found that "a higher proportion of children and adolescents in the intervention group compared with the usual care group had received influenza vaccine (difference, 3.7% [95% CI, 1.5% - 5.9%]"*

Notice that the difference in our `sample estimates` is $0.436 - 0.399 = 0.037$ or 3.7%. And the 95% confidence interval from our output is `[0.015, 0.059]`. We replicated the exact analysis these researchers used for their study. Notice that **`zero`** is not in the interval, so we can conclude that the text intervention group has a higher overall proportion of getting vaccinated.

What statements can we make about causation and generalizability?

This was a randomized controlled experiment so we can conclude that sending text reminders to get your kids vaccinated really does cause the rate of vaccinations to increase.