Comparing Two Proportions, Part 1
Difference in Proportions

JC Wang

March 1, 2016
Outline

1. Difference Between Independent Proportions
   - Example and Notation
   - Standard Error of Difference in Sample Proportions

2. Confidence Interval for Difference in Proportions
   - Confidence Interval for Difference in Proportions
   - iClicker Questions

3. Statistical Significance
   - Statistical Significance
Change in Student Retention Rate

Has retention rate at WMU been changing?

- A random sample of 200 entering students in 1989 → 74% were still enrolled 3 years later.
- Another random sample of 200 entering students in 1999 → 66% were still enrolled 3 years later.
- An 8% change in 3-year retention rate was observed.
- The 8% difference is based on random sampling, and is only an estimate of the true difference.
- What is the likely size of the error of estimation?
Change in Student Retention Rate

Has retention rate at WMU been changing?

- A random sample of 200 entering students in 1989 $\Rightarrow$ 74% were still enrolled 3 years later.
- Another random sample of 200 entering students in 1999 $\Rightarrow$ 66% were still enrolled 3 years later.
- An 8% change in 3-year retention rate was observed.
- The 8% difference is based on random sampling, and is only an estimate of the true difference.
- What is the likely size of the error of estimation?
Change in Student Retention Rate

Has retention rate at WMU been changing?

- A random sample of 200 entering students in 1989 → 74% were still enrolled 3 years later.
- Another random sample of 200 entering students in 1999 → 66% were still enrolled 3 years later.

- An 8% change in 3-year retention rate was observed.
- The 8% difference is based on random sampling, and is only an estimate of the true difference.
- What is the likely size of the error of estimation?
Change in Student Retention Rate

Has retention rate at WMU been changing?

- A random sample of 200 entering students in 1989 → 74% were still enrolled 3 years later.
- Another random sample of 200 entering students in 1999 → 66% were still enrolled 3 years later.
- An 8% change in 3-year retention rate was observed.

- The 8% difference is based on random sampling, and is only an estimate of the true difference.
- What is the likely size of the error of estimation?
Has retention rate at WMU been changing?

- A random sample of 200 entering students in 1989 $\rightarrow$ 74% were still enrolled 3 years later.
- Another random sample of 200 entering students in 1999 $\rightarrow$ 66% were still enrolled 3 years later.
- An 8% change in 3-year retention rate was observed.
- The 8% difference is based on random sampling, and is only an estimate of the true difference.

What is the likely size of the error of estimation?
Change in Student Retention Rate

Has retention rate at WMU been changing?

- A random sample of 200 entering students in 1989 \(\Rightarrow\) 74% were still enrolled 3 years later.

- Another random sample of 200 entering students in 1999 \(\Rightarrow\) 66% were still enrolled 3 years later.

- An 8% change in 3-year retention rate was observed.

- The 8% difference is based on random sampling, and is only an estimate of the true difference.

- What is the likely size of the error of estimation?
A categorical variable with binary responses (‘success’ and ‘failure’) is of interest for two independent populations.

- Population 1 has proportion $p_1$ of successes.
- Population 2 has proportion $p_2$ of successes.
- Sample of size $n_1$ is taken from population 1: $X$ successes observed in the sample with sample proportion of $\hat{p}_1 = \frac{X}{n_1}$.
- Sample of size $n_2$ is taken from population 2: $Y$ successes observed in the sample with sample proportion of $\hat{p}_2 = \frac{Y}{n_2}$.
- The two samples are independent.
Notation

A categorical variable with binary responses ('success' and 'failure') is of interest for two independent populations.

- Population 1 has proportion $p_1$ of successes.
- Population 2 has proportion $p_2$ of successes.
- Sample of size $n_1$ is taken from population 1: $X$ successes observed in the sample with sample proportion of $\hat{p}_1 = \frac{X}{n_1}$.
- Sample of size $n_2$ is taken from population 2: $Y$ successes observed in the sample with sample proportion of $\hat{p}_2 = \frac{Y}{n_2}$.
- The two samples are independent.
Notation

A categorical variable with binary responses (‘success’ and ‘failure’) is of interest for two independent populations.

- Population 1 has proportion $p_1$ of successes.
- Population 2 has proportion $p_2$ of successes.
- Sample of size $n_1$ is taken from population 1: $X$ successes observed in the sample with sample proportion of $\hat{p}_1 = \frac{X}{n_1}$.
- Sample of size $n_2$ is taken from population 2: $Y$ successes observed in the sample with sample proportion of $\hat{p}_2 = \frac{Y}{n_2}$.
- The two samples are independent.
Notation

A categorical variable with binary responses (‘success’ and ‘failure’) is of interest for two independent populations.

- Population 1 has proportion $p_1$ of successes.
- Population 2 has proportion $p_2$ of successes.
- Sample of size $n_1$ is taken from population 1: $X$ successes observed in the sample with sample proportion of $\hat{p}_1 = \frac{X}{n_1}$.
- Sample of size $n_2$ is taken from population 2: $Y$ successes observed in the sample with sample proportion of $\hat{p}_2 = \frac{Y}{n_2}$.
- The two samples are independent.
Notation

A categorical variable with binary responses (‘success’ and ‘failure’) is of interest for two independent populations.

- Population 1 has proportion $p_1$ of successes.
- Population 2 has proportion $p_2$ of successes.
- Sample of size $n_1$ is taken from population 1: $X$ successes observed in the sample with sample proportion of $\hat{p}_1 = \frac{X}{n_1}$.
- Sample of size $n_2$ is taken from population 2: $Y$ successes observed in the sample with sample proportion of $\hat{p}_2 = \frac{Y}{n_2}$.
- The two samples are independent.
Notation

A categorical variable with binary responses (‘success’ and ‘failure’) is of interest for two independent populations.

- Population 1 has proportion $p_1$ of successes.
- Population 2 has proportion $p_2$ of successes.
- Sample of size $n_1$ is taken from population 1: $X$ successes observed in the sample with sample proportion of $\hat{p}_1 = \frac{X}{n_1}$.
- Sample of size $n_2$ is taken from population 2: $Y$ successes observed in the sample with sample proportion of $\hat{p}_2 = \frac{Y}{n_2}$.
- The two samples are independent.
Standard Error of Difference of independent sample proportions

The SE (Standard Error) of the difference in the sample proportions of two independent samples is

$$SE\ of\ (\hat{p}_1 - \hat{p}_2) = \sqrt{(SE\ of\ \hat{p}_1)^2 + (SE\ of\ \hat{p}_2)^2}$$

where

$$SE\ of\ \hat{p}_1 = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1}}$$

$$SE\ of\ \hat{p}_2 = \sqrt{\frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$
Change in Student Retention Rate

revisited

- **for 1989 sample**: $\hat{p}_1 = 0.74$ give or take (i.e., with a standard error of)

  $$
  \text{SE of } \hat{p}_1 = \sqrt{\frac{0.74(1 - 0.74)}{200}} = \sqrt{0.000962} = 0.031
  $$

- **for 1999 sample**: $\hat{p}_2 = 0.66$ give or take (i.e., with a standard error of)

  $$
  \text{SE of } \hat{p}_2 = \sqrt{\frac{0.66(1 - 0.66)}{200}} = \sqrt{0.001122} = 0.033
  $$

- and hence the SE for the difference in sample proportions:

  $$
  \text{SE of } (\hat{p}_1 - \hat{p}_2) = \sqrt{(0.031)^2 + (0.033)^2}
  $$

  $$
  = \sqrt{0.000962 + 0.001122} = \sqrt{0.002084} = 0.0456
  $$
Change in Student Retention Rate

**revisited**

- **for 1989 sample:** \( \hat{p}_1 = 0.74 \) give or take (i.e., with a standard error of)

\[
\text{SE of } \hat{p}_1 = \sqrt{\frac{0.74(1 - 0.74)}{200}} = \sqrt{.000962} = .031
\]

- **for 1999 sample:** \( \hat{p}_2 = 0.66 \) give or take (i.e., with a standard error of)

\[
\text{SE of } \hat{p}_2 = \sqrt{\frac{0.66(1 - 0.66)}{200}} = \sqrt{.001122} = .033
\]

and hence the SE for the difference in sample proportions:

\[
\text{SE of } (\hat{p}_1 - \hat{p}_2) = \sqrt{(.031)^2 + (.033)^2} = \sqrt{.000962 + .001122} = \sqrt{.002084} = .0456
\]
Change in Student Retention Rate revisited

- **for 1989 sample:** \( \hat{p}_1 = 0.74 \) give or take (i.e., with a standard error of)
  \[
  \text{SE of } \hat{p}_1 = \sqrt{\frac{0.74(1 - 0.74)}{200}} = \sqrt{0.000962} = 0.031
  \]

- **for 1999 sample:** \( \hat{p}_2 = 0.66 \) give or take (i.e., with a standard error of)
  \[
  \text{SE of } \hat{p}_2 = \sqrt{\frac{0.66(1 - 0.66)}{200}} = \sqrt{0.001122} = 0.033
  \]

- and hence the SE for the difference in sample proportions:
  \[
  \text{SE of } (\hat{p}_1 - \hat{p}_2) = \sqrt{(0.031)^2 + (0.033)^2}
  \]
  \[
  = \sqrt{0.000962 + 0.001122} = \sqrt{0.002084} = 0.0456
  \]
Calculation of the SE (of $\hat{p}_1 - \hat{p}_2$)

1. Calculate $(SE_1)^2$, the squared SE of $\hat{p}_1$:

$$(SE_1)^2 = \frac{\hat{p}_1(1 - \hat{p}_1)}{n_1}$$

keeping 6 decimal places to the right of the decimal point.

2. Calculate $(SE_2)^2$, the squared SE of $\hat{p}_2$:

$$(SE_2)^2 = \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}$$

keeping 6 decimal places to the right of the decimal point.

3. Calculate $(SE_1)^2 + (SE_2)^2$.

4. $SE = \sqrt{(SE_1)^2 + (SE_2)^2}$. 
Calculation of the SE (of $\hat{p}_1 - \hat{p}_2$)

1. Calculate $(SE_1)^2$, the squared SE of $\hat{p}_1$:

$$(SE_1)^2 = \frac{\hat{p}_1(1 - \hat{p}_1)}{n_1}$$

keeping 6 decimal places to the right of the decimal point.

2. Calculate $(SE_2)^2$, the squared SE of $\hat{p}_2$:

$$(SE_2)^2 = \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}$$

keeping 6 decimal places to the right of the decimal point.

3. Calculate $(SE_1)^2 + (SE_2)^2$.

4. $SE = \sqrt{(SE_1)^2 + (SE_2)^2}$. 
Calculation of the SE (of $\hat{p}_1 - \hat{p}_2$)

1. Calculate $(SE_1)^2$, the squared SE of $\hat{p}_1$:

$$(SE_1)^2 = \frac{\hat{p}_1(1 - \hat{p}_1)}{n_1}$$

keeping 6 decimal places to the right of the decimal point.

2. Calculate $(SE_2)^2$, the squared SE of $\hat{p}_2$:

$$(SE_2)^2 = \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}$$

keeping 6 decimal places to the right of the decimal point.

3. Calculate $(SE_1)^2 + (SE_2)^2$.

4. $SE = \sqrt{(SE_1)^2 + (SE_2)^2}$. 
Calculation of the SE (of $\hat{p}_1 - \hat{p}_2$)

1. Calculate $(SE_1)^2$, the squared SE of $\hat{p}_1$:

\[(SE_1)^2 = \frac{\hat{p}_1(1 - \hat{p}_1)}{n_1}\]

keeping 6 decimal places to the right of the decimal point.

2. Calculate $(SE_2)^2$, the squared SE of $\hat{p}_2$:

\[(SE_2)^2 = \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}\]

keeping 6 decimal places to the right of the decimal point.

3. Calculate $(SE_1)^2 + (SE_2)^2$.

4. $SE = \sqrt{(SE_1)^2 + (SE_2)^2}$. 
Calculation of the SE (of $\hat{p}_1 - \hat{p}_2$)

Change in Retention Rate Example

1. 

$$(SE_1)^2 = \frac{0.74(1 - 0.74)}{200} = 0.000962.$$ 

2. 

$$(SE_2)^2 = \frac{0.66(1 - 0.66)}{200} = 0.001122.$$ 

3. 

$$(SE_1)^2 + (SE_2)^2 = 0.000962 + 0.001122 = 0.002084.$$ 

4. 

$$SE = \sqrt{0.002084} = 0.0456.$$
Calculation of the SE (of $\hat{p}_1 - \hat{p}_2$)
Change in Retention Rate Example

1. $$(SE_1)^2 = \frac{0.74(1 - 0.74)}{200} = 0.000962.$$  
2. $$(SE_2)^2 = \frac{0.66(1 - 0.66)}{200} = 0.001122.$$  
3. $$(SE_1)^2 + (SE_2)^2 = 0.000962 + 0.001122 = 0.002084.$$  
4. $$SE = \sqrt{0.002084} = 0.0456.$$
Calculation of the SE (of $\hat{p}_1 - \hat{p}_2$)

Change in Retention Rate Example

1. $(SE_1)^2 = \frac{0.74(1 - 0.74)}{200} = 0.000962$.

2. $(SE_2)^2 = \frac{0.66(1 - 0.66)}{200} = 0.001122$.

3. $(SE_1)^2 + (SE_2)^2 = 0.000962 + 0.001122 = 0.002084$.

4. $SE = \sqrt{0.002084} = 0.0456$. 
Calculation of the SE (of $\hat{p}_1 - \hat{p}_2$)

Change in Retention Rate Example

1. $(SE_1)^2 = \frac{0.74(1 - 0.74)}{200} = 0.000962.$

2. $(SE_2)^2 = \frac{0.66(1 - 0.66)}{200} = 0.001122.$

3. $(SE_1)^2 + (SE_2)^2 = 0.000962 + 0.001122 = 0.002084.$

4. $SE = \sqrt{0.002084} = 0.0456.$
Outline

1. Difference Between Independent Proportions
   - Example and Notation
   - Standard Error of Difference in Sample Proportions

2. Confidence Interval for Difference in Proportions
   - Confidence Interval for Difference in Proportions
   - iClicker Questions

3. Statistical Significance
   - Statistical Significance
Confidence Interval
for the Difference in Proportion

A 95% confidence interval for the true difference $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm [1.96 \times \text{SE of } (\hat{p}_1 - \hat{p}_2)]$$

That is

$$(\hat{p}_1 - \hat{p}_2) \pm \left[1.96 \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}\right]$$

If the interval excludes 0, then we say that the difference in sample proportions is statistically significant.

If the interval includes 0 then the difference is statistically insignificant.
Change in Student Detention Rate

95% confidence interval

Recall that the standard error of the difference in the sample proportions is

\[ SE \ of \ (\hat{p}_1 - \hat{p}_2) = 0.0456 \]

So, a 95% c.i. (confidence interval) for \( p_1 - p_2 \) is

\[ (.74 - .66) \pm [1.96 \times .0456] = .08 \pm .089 = (-.009, .169) \]

If we round it off to \((-0.01, 0.17)\), or, in percentages, \((-1\%, 17\%)\), we say that the drop in retention rate from 1989 to 1999 is between \(-1\%\) and \(17\%\) with 95% confidence. Note: 0% is contained in this interval and hence there is still a probability that there might not be a real change in retention rate, just chance variability.
A 95% confidence interval was constructed for difference in the proportions $p_1 - p_2$ in two independent populations: $(-0.08, 0.26)$. Which of the following is true?

A. The difference in the proportions is significant.
B. $p_1$ differs from $p_2$ significantly.
C. The difference in the proportions is insignificant.
D. None of the previous.
iClicker Question 13.2

A study of the television viewing preferences of children, each child is asked if the Sesame Street is the program he or she likes the best among others. Of 200 girls surveyed, 85 like the Sesame Street the best; of 100 boys surveyed, 30 like the Sesame Street the best. A 95% confidence interval for the difference in the percentages of children like the Sesame Street the best between girls and boys is (1.2%, 23.8%). Which of the following is true?

A. The two percentages differ significantly.
B. The two percentages do not differ significantly.
C. The two proportions do not differ significantly.
D. None of the previous.
Outline

1. Difference Between Independent Proportions
   - Example and Notation
   - Standard Error of Difference in Sample Proportions

2. Confidence Interval for Difference in Proportions
   - Confidence Interval for Difference in Proportions
   - iClicker Questions

3. Statistical Significance
   - Statistical Significance
Cooks or Chefs

According to a 2009 occupation survey by the Census Bureau, regular cooks were a separate classification from chefs or head cooks:

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
<th>Total</th>
<th>%Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooks</td>
<td>441</td>
<td>762</td>
<td>1203</td>
<td>37</td>
</tr>
<tr>
<td>Chefs or Head Cooks</td>
<td>45</td>
<td>245</td>
<td>290</td>
<td>16</td>
</tr>
</tbody>
</table>

The difference in percentage is approximately 21%.

Is the difference in percentages just luck of the draw, or due to something else besides chance?
Cooks or Chefs, continued

If chance was at work, how likely we get a difference in proportions of 0.21?

The chance of this occurs is small $\Rightarrow < 0.0001$. That is, less than 1 in 10,000. This chance of getting 0.21 by chance is called a $P$-value.

But how do we know that this P-value is less than 0.0001?
Cooks or Chefs, continued

The SE for the difference in proportion is

$$\text{SE of } (\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{.37(1-.37)}{1203} + \frac{.16(1-.16)}{290}} = .026$$

And hence the chance to get a difference beyond ±0.078 (= 3SE) is 0.003 (= 1 − .997 by the empirical rule), or 3 in 1,000.

Similarly, the chance to get a difference beyond ±.104 (= 4SE) is 0.00006 < 0.0001, or less than 1 in 10,000.

Now, in our example, a difference of 0.21 is beyond 8SE. This cannot be just chance variability. Something else is at work.

Note: the probability of 0.00006 above was obtained by computer.
Statistical Significance, The P-Value

Rule of thumb for P-value for the difference:

- If P-value ≤ .05, the difference is *statistically significant*. (difference is at least 1.96SE in absolute value)
- If P-value ≤ .01, the difference is called *highly significant*. (difference is at least 2.58SE in absolute value)
- If P-value >.05, the difference is *insignificant*. (difference is less than 1.96SE in absolute value)