Independent Samples: Comparing Proportions

Lecture 39
Section 11.5

Robb T. Koether

Hampden-Sydney College

Wed, Apr 2, 2008
Outline

1. Confidence Intervals for $\mu_1 - \mu_2$
2. Confidence Intervals for $\mu_1 - \mu_2$ on the TI-83
3. Introduction to Proportions
4. Comparing Proportions
5. Example - Turmoil at City Hall
6. The Hypothesis Testing for $p_1 - p_2$
   - The Hypotheses
   - The Sampling Distribution of $\hat{p}_1 - \hat{p}_2$
   - The Test Statistic
   - The Pooled Estimate of $p$
   - The Rest of the Procedure
7. Summary
Confidence Intervals

- Confidence intervals for $\mu_1 - \mu_2$ use the same theory.
- The point estimate is $\bar{x}_1 - \bar{x}_2$.
- The standard deviation of $\bar{x}_1 - \bar{x}_2$ is approximately

$$sp \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$
The confidence interval is either

\[(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\]

or

\[(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\]

or

\[(\bar{x}_1 - \bar{x}_2) \pm t_{df, \alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\]

depending on the circumstances.
The choice depends on

- Whether \( \sigma \) is known.
- Whether the populations are normal.
- Whether the sample sizes are large.
Example

Find a 95% confidence interval for $\mu_1 - \mu_2$ in Example 11.4, p. 699.

- $\bar{x}_1 - \bar{x}_2 = 3.2$.
- $s_p = 5.052$.
- Use $t_{18,0.025} = 2.101$.
- The confidence interval is

$$3.2 \pm (2.101)(2.259) = 3.2 \pm 4.75.$$
Example

- Find a 95% confidence interval for $\mu_1 - \mu_2$ in Example 11.4, p. 699.
  - $\bar{x}_1 - \bar{x}_2 = 3.2$.  
  - $s_p = 5.052$.  
  - Use $t_{18,0.025} = 2.101$.
  - The confidence interval is $3.2 \pm (2.101)(2.259) = 3.2 \pm 4.75$. 
Example

- Find a 95% confidence interval for $\mu_1 - \mu_2$ in Example 11.4, p. 699.
- $\bar{x}_1 - \bar{x}_2 = 3.2$.
- $s_p = 5.052$.
- Use $t_{18,0.025} = 2.101$.
- The confidence interval is

$$3.2 \pm (2.101)(2.259) = 3.2 \pm 4.75.$$
To find a confidence interval for the difference between means on the TI-83,

- Press STAT > TESTS.
- Choose either 2-SampZInt or 2-SampTInt, depending on circumstances.
- Choose Data or Stats.
- Provide the information that is called for.
- 2-SampTTest will ask whether to use a pooled estimate of $\sigma$. Answer “yes.”
- Press Calculate.

The confidence interval appears.
Example

- Find a 95% confidence interval for $\mu_1 - \mu_2$ in Example 11.4, p. 699 using the TI-83.
We will now do for proportions what we just did for means.

We will test hypotheses concerning the difference between proportions from two different populations.

Again we will encounter the concept of a pooled estimate, this time for $p$.

Some of the formulas get a bit messy, but there will be a formula sheet given out later.
Comparing Proportions

- We wish to compare proportions between two populations.
- We should compare proportions for the same attribute in order for it to make sense.
- For example, we could measure the proportion of NC residents living below the poverty level and the proportion of VA residents living below the poverty level.
  - Populations - people in NC and VA.
  - Variable - whether the resident is living below the poverty level.
Example

The “gender gap” refers to the difference between the proportion of men who vote Republican and the proportion of women who vote Republican.

There are several ways to view this.

- Two populations (males and females), one variable (how one votes).
- Two populations (Dems and Reps), one variable (one’s sex).
- One population, two variables (sex, vote).
Example

- The proportion of patients given treatment $A$ who recovered vs. the proportion of patients given treatment $B$ who recovered.
  - $p_1$ = recovery rate under treatment $A$.
  - $p_2$ = recovery rate under treatment $B$.
- The important questions is, which is greater, $p_1$ or $p_2$?
We will work with the single number $p_1 - p_2$.

To estimate the difference between population proportions $p_1$ and $p_2$, we will use the sample proportions $\hat{p}_1$ and $\hat{p}_2$.

The sample difference $\hat{p}_1 - \hat{p}_2$ is an estimator of the population difference $p_1 - p_2$. 
Test the hypothesis that a higher proportion of men than women believe that Mayor Wilder is doing a good or excellent job as mayor of Richmond.
Let $p_1 =$ proportion of men who believe that Mayor Wilder is doing a good or excellent job.

Let $p_2 =$ proportion of women who believe that Mayor Wilder is doing a good or excellent job.
Turmoil at City Hall

The data:
- 500 people surveyed.
- 48% were male; 52% were female.
- 41% of men rated Wilder’s performance good or excellent.
- 37% of men rated Wilder’s performance good or excellent.
The data:

- 500 people surveyed.
- 48% were male; 52% were female.
- 41% of men rated Wilder’s performance good or excellent.
- 37% of men rated Wilder’s performance good or excellent.
Turmoil at City Hall

The data:

- 500 people surveyed.
- 48% were male; 52% were female.
- 41% of men rated Wilder’s performance good or excellent.
- 37% of men rated Wilder’s performance good or excellent.
Turmoil at City Hall

The data:
- 500 people surveyed.
- 48% were male; 52% were female.
- 41% of men rated Wilder’s performance good or excellent.
- 37% of men rated Wilder’s performance good or excellent.
Turmoil at City Hall

The data:
- 500 people surveyed.
- 48% were male; 52% were female.
- 41% of men rated Wilder’s performance good or excellent.
- 37% of men rated Wilder’s performance good or excellent.
Hypothesis Testing Procedure

The hypotheses.

\[ H_0 : p_1 = p_2 \]
\[ H_1 : p_1 > p_2 \]

The significance level is \( \alpha = 0.05 \).

What is the test statistic?

That depends on the sampling distribution of \( \hat{p}_1 - \hat{p}_2 \).

Here we go again...
The Sampling Distribution of $\hat{p}_1 - \hat{p}_2$

- If the sample sizes are large enough, then $\hat{p}_1$ is $N(p_1, \sigma_{\hat{p}_1})$, where

\[
\sigma_{\hat{p}_1} = \sqrt{\frac{p_1(1 - p_1)}{n_1}}
\]

and $\hat{p}_2$ is $N(p_2, \sigma_{\hat{p}_2})$, where

\[
\sigma_{\hat{p}_2} = \sqrt{\frac{p_2(1 - p_2)}{n_2}}
\]
The Sampling Distribution of $\hat{p}_1 - \hat{p}_2$

- The sample sizes will be large enough if
  - $n_1 p_1 \geq 5$, and $n_1 (1 - p_1) \geq 5$, and
  - $n_2 p_2 \geq 5$, and $n_2 (1 - p_2) \geq 5$. 
Recall the statistical facts.

For any two random variables $X$ and $Y$,

\[ \mu_{X-Y} = \mu_X - \mu_Y \]

\[ \sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y \]

\[ \sigma_{X-Y} = \sqrt{\sigma^2_X + \sigma^2_Y} \]

Furthermore, if $X$ and $Y$ are both normal, then $X - Y$ is normal.
The Sampling Distribution of $\hat{p}_1 - \hat{p}_2$

Therefore, $\hat{p}_1 - \hat{p}_2$ is normal with

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$$

$$\sigma^2_{\hat{p}_1 - \hat{p}_2} = \sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2} = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}$$

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$
Therefore, the test statistic would be

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

except that...
...we do not know the values of $p_1$ and $p_2$. We will approximate $p_1$ and $p_2$ with $\hat{p}_1$ and $\hat{p}_2$. Therefore, the test statistic would be

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

except that...
Pooled Estimate of $p$

...we can do better.

The null hypothesis is

$$H_0 : p_1 = p_2$$

Under that assumption,

$\hat{p}_1$ and $\hat{p}_2$ are both estimators of a common value, which we will call $p$. 
Pooled Estimate of $p$

- Rather than use either $\hat{p}_1$ or $\hat{p}_2$ alone to estimate $p$, we will use the **pooled estimate**.

- The pooled estimate is the proportion that we would get if we pooled the two samples together into one.

- We would have a total count of $x_1 + x_2$ members out of a sample of $n_1 + n_2$, for a pooled proportion of

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$
The Test Statistic

- This leads to a better estimator of the standard deviation of $\hat{p}_1 - \hat{p}_2$.

\[ \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}. \]

- So the test statistic is

\[ Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}, \]

where

\[ \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}. \]
Turmoil at City Hall

- In the survey, we had 240 males (48%) and 260 females (52%).
- 41% of the males, or 98 males, said Wilder is doing good or excellent.
- 37% of the females, or 96 females, said he is doing good or excellent.
- Therefore, altogether, 194 people out of 500, or 38.8%, said he is doing good or excellent.
Now compute

$$
\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{(0.388)(0.612) \left( \frac{1}{240} + \frac{1}{260} \right)} = 0.04362.
$$

For comparison, estimate $\sigma_{\hat{p}_1 - \hat{p}_2}$ without using the pooled estimate. There will not be much difference. Nevertheless, we should always use the pooled estimate.
The Value of the Test Statistic

- Now compute $z$:
  \[ z = \frac{0.04}{0.04362} = 0.9170. \]

- Compute the $p$-value:
  \[ P(Z > 0.9170) = 0.1796. \]

- Accept $H_0$.

- Equal proportions of men and women believe that Mayor Wilder is doing a good or excellent job.
Do equal proportions of whites and blacks believe that Mayor Wilder is doing a good or excellent job?

Do equal proportions of Republicans and Democrats believe that Mayor Wilder is doing a good or excellent job?

Let’s learn how to do this on the TI-83 and then test those hypotheses.
Summary

- The point estimate for $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$.
- The sampling distribution of $\hat{p}_1 - \hat{p}_2$ is normal with

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$$

and

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

- However, we use the pooled estimate

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

in place of both $p_1$ and $p_2$ in the formula.

- Otherwise, the hypothesis-testing procedure is the same as before.