Math 11
Calculus-Based Introductory Probability and Statistics

Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today:

- Two-Sample Proportion Inference

Statistics in the Large

Where we stand: We know how to build C.I.'s and run hypothesis tests for one proportion.
So we can build a range of plausible values for a single proportion, or compare a single proportion to a known value.

Statistics in the Large

Where we stand: We know how to build C.I.'s and run hypothesis tests for one proportion.
So we can build a range of plausible values for a single proportion, or compare a single proportion to a known value.

Today: Extend these ideas to two parameters (two populations)

Statistics in the Large

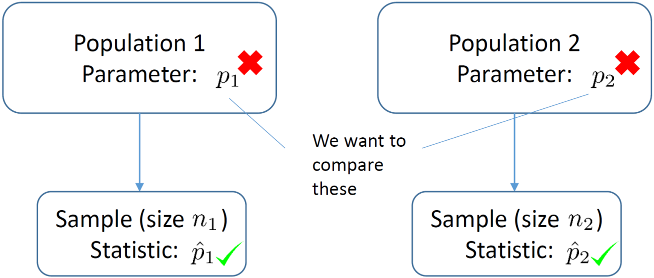
Where we stand: We know how to build C.I.'s and run hypothesis tests for one proportion.
So we can build a range of plausible values for a single proportion, or compare a single proportion to a known value.

Today: Extend these ideas to two parameters (two populations)

Nice thing: The approach we use closely follows the same ideas as for inference of a single proportion.

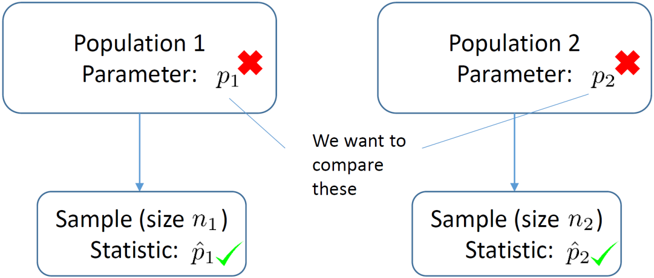
Two-Sample Proportion Setting

Before, we either had one population, or two, but we knew the parameter from the second population. Now, we don't know anything about either population.



Two-Sample Proportion Setting

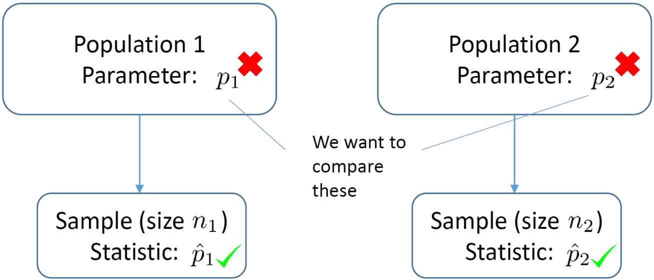
Before, we either had one population, or two, but we knew the parameter from the second population. Now, we don't know anything about either population.



Typically, when we compare p_1 and p_2 (or μ_1 and μ_2), we think about $p_1 - p_2$.
For example, if you care about $p_1 > p_2$, then explore $p_1 - p_2 > 0$.

Two-Sample Proportion Setting

Before, we either had one population, or two, but we knew the parameter from the second population. Now, we don't know anything about either population.

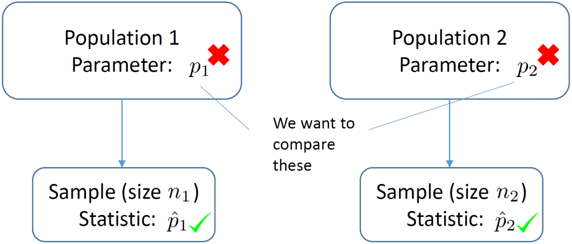


Typically, when we compare p_1 and p_2 (or μ_1 and μ_2), we think about $p_1 - p_2$.
For example, if you care about $p_1 > p_2$, then explore $p_1 - p_2 > 0$.

We might try to explore this using a confidence interval about $\hat{p}_1 - \hat{p}_2$, or we might run a hypothesis test with $H_0 : p_1 - p_2 = 0$.

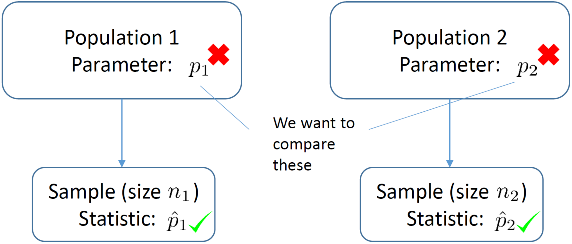
Sampling Distribution Galore!

Sampling Distribution Galore!

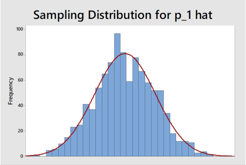


(**Note**: the samples may have different sizes)

Sampling Distribution Galore!



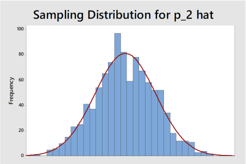
(**Note**: the samples may have different sizes)



Normal distribution

$E(\hat{p}_1) = p_1$

$SE(\hat{p}_1) = \sqrt{\frac{p_1 q_1}{n_1}}.$

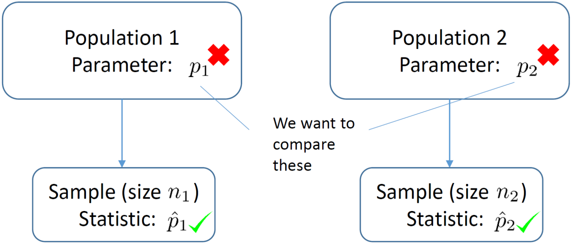


Normal distribution

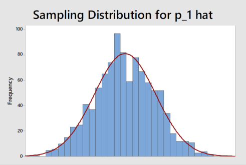
$E(\hat{p}_2) = p_2$

$SE(\hat{p}_2) = \sqrt{\frac{p_2 q_2}{n_2}}.$

Sampling Distribution Galore!



(**Note**: the samples may have different sizes)



Normal distribution

$E(\hat{p}_1) = p_1$

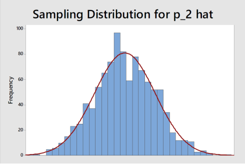
$SE(\hat{p}_1) = \sqrt{\frac{p_1 q_1}{n_1}}$.

What does the sampling distribution of $\hat{p}_1 - \hat{p}_2$ look like?

Shape?

Center?

Spread?



Normal distribution

$E(\hat{p}_2) = p_2$

$SE(\hat{p}_2) = \sqrt{\frac{p_2 q_2}{n_2}}$.

Sampling Distribution of the Difference

If X and Y are independent random variables with Normal distributions, then $X - Y$ is also Normal. In addition,

$$E(X - Y) = E(X) - E(Y),$$

and

$$SD(X - Y) = \sqrt{Var(X - Y)} = \sqrt{SD(X)^2 + SD(Y)^2}.$$

Sampling Distribution of the Difference

If X and Y are independent random variables with Normal distributions, then $X - Y$ is also Normal. In addition,

$$E(X - Y) = E(X) - E(Y),$$

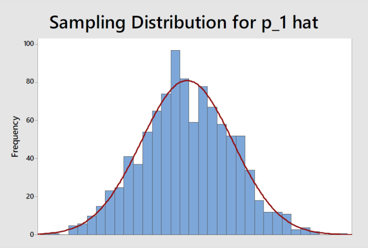
and

$$SD(X - Y) = \sqrt{Var(X - Y)} = \sqrt{SD(X)^2 + SD(Y)^2}.$$

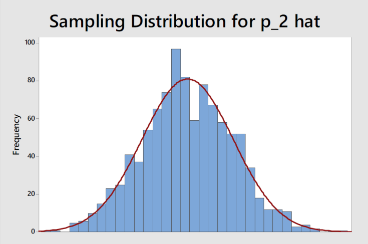
So if $\hat{p}_1 \simeq N\left(p_1, \sqrt{\frac{p_1 q_1}{n_1}}\right)$ and $\hat{p}_2 \simeq N\left(p_2, \sqrt{\frac{p_2 q_2}{n_2}}\right)$ are independent, we get

$$\hat{p}_1 - \hat{p}_2 \simeq N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$$

Sampling Distribution of the Difference



Normal distribution
$E(\hat{p}_1) = p_1$
$SE(\hat{p}_1) = \sqrt{\frac{p_1 q_1}{n_1}}$.



Normal distribution
$E(\hat{p}_2) = p_2$
$SE(\hat{p}_2) = \sqrt{\frac{p_2 q_2}{n_2}}$.

For unpaired data, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is:

- Normal
- $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$
- $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$

Drill, Baby, Drill

A 2010 survey asked 827 random voters in California how they feel about drilling for oil off the coast of CA. Of the 438 college graduates in the sample, 154 approved. Of the 389 who didn't graduate from college, 132 we in favor.

Find a 95% C.I. for the **difference in the proportions** of college and non-college California grads who support drilling.

Drill, Baby, Drill

A 2010 survey asked 827 random voters in California how they feel about drilling for oil off the coast of CA. Of the 438 college graduates in the sample, 154 approved. Of the 389 who didn't graduate from college, 132 we in favor.
Find a 95% C.I. for the **difference in the proportions** of college and non-college California grads who support drilling.

Let p_1 be the proportion of CA college grads that support drilling.
Let p_2 be the proportion of CA non-college grads that support drilling.

Drill, Baby, Drill

A 2010 survey asked 827 random voters in California how they feel about drilling for oil off the coast of CA. Of the 438 college graduates in the sample, 154 approved. Of the 389 who didn't graduate from college, 132 we in favor.
Find a 95% C.I. for the **difference in the proportions** of college and non-college California grads who support drilling.

Let p_1 be the proportion of CA college grads that support drilling.
Let p_2 be the proportion of CA non-college grads that support drilling.

We found $\hat{p}_1 = \dfrac{154}{438} \simeq 35.16\%$ and $\hat{p}_2 = \dfrac{132}{389} \simeq 33.93\%$. So

$$\hat{p}_1 - \hat{p}_2 \simeq 1.23\%.$$

Recall that the sampling distribution of the difference $\hat{p}_1 - \hat{p}_2$ is

$$N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right).$$

Recall that the sampling distribution of the difference $\hat{p}_1 - \hat{p}_2$ is

$$N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right).$$

To build a confidence interval, we will need to estimate the SE since we don't know p_1 or p_2.

Recall that the sampling distribution of the difference $\hat{p}_1 - \hat{p}_2$ is

$$N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right).$$

To build a confidence interval, we will need to estimate the SE since we don't know p_1 or p_2.

As usual, we use the point estimate $SE_{\hat{p}_1 - \hat{p}_2} \simeq \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$.

As before, we start at our estimate and reach out a certain number of SE's:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times SE_{\hat{p}_1 - \hat{p}_2}.$$

We found $\hat{p}_1 \simeq 35.16\%$ and $\hat{p}_2 \simeq 33.93\%$, so $\hat{p}_1 - \hat{p}_2 \simeq 1.23\%$.

We find $SE = \sqrt{\dfrac{35.16 \times 64.84}{438} + \dfrac{33.93 \times 66.07}{389}} \simeq 3.312\%$.

We found $\hat{p}_1 \simeq 35.16\%$ and $\hat{p}_2 \simeq 33.93\%$, so $\hat{p}_1 - \hat{p}_2 \simeq 1.23\%$.

We find $SE = \sqrt{\dfrac{35.16 \times 64.84}{438} + \dfrac{33.93 \times 66.07}{389}} \simeq 3.312\%$.

For a 95% C.I., we must reach $z^* = 1.96$ SE's:

$$\begin{aligned}(\hat{p}_1 - \hat{p}_2) \pm z^* \times SE_{\hat{p}_1 - \hat{p}_2} &= 1.23 \pm 1.96 \times 3.312 \\ &= (-5.26\%, 7.72\%).\end{aligned}$$

We found $\hat{p}_1 \simeq 35.16\%$ and $\hat{p}_2 \simeq 33.93\%$, so $\hat{p}_1 - \hat{p}_2 \simeq 1.23\%$.

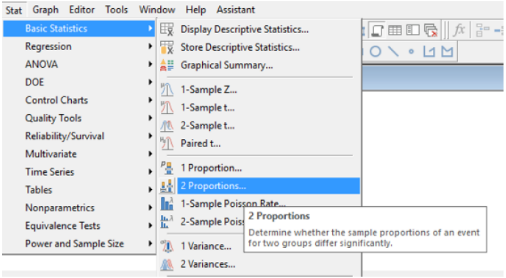
We find $SE = \sqrt{\dfrac{35.16 \times 64.84}{438} + \dfrac{33.93 \times 66.07}{389}} \simeq 3.312\%$.

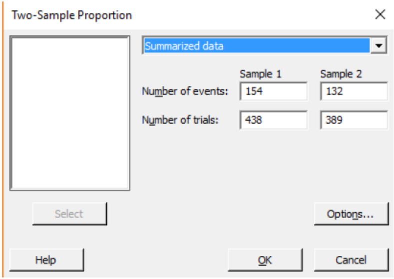
For a 95% C.I., we must reach $z^* = 1.96$ SE's:

$$\begin{aligned}(\hat{p}_1 - \hat{p}_2) \pm z^* \times SE_{\hat{p}_1 - \hat{p}_2} &= 1.23 \pm 1.96 \times 3.312 \\ &= (-5.26\%, 7.72\%).\end{aligned}$$

Note: Minitab can create confidence intervals for the difference in two proportions (see next slide). You should only use this to check your answers on homework, not to completely do problems.

Using Minitab





Test and CI for Two Proportions

```
Sample     X    N  Sample p
1        154  438  0.351598
2        132  389  0.339332


Difference = p (1) - p (2)
Estimate for difference:  0.0122666
95% CI for difference:  (-0.0526437, 0.0771768)
```

But Wait! When is the Sampling Distribution What We Claim?

To get each of the individual sampling distributions to be Normal, **in each sample** we need:

- Independence (usually shown through Randomization and <10% Conditions)
- At least 10 successes and failures

To use the $Var(X-Y) = Var(X)+Var(Y)$ formula to find $SE_{\hat{p}_1-\hat{p}_2}$, we need

- Independence between the two samples

Below is given two samples (A and B) and a proportion of interest that you want to compare across the two groups. Which of the following setups <u>will violate</u> the independence required between the two samples?

1. A: Random Californians,
   B: Random Texans;
   percent with college degree in CA vs TX residents
2. A: Random married men,
   B: The wives of those married men;
   percent with college degrees in married men and married women
3. A: Random adults that have kids,
   B: Kids of those adults;
   percent that believe in God in adults vs kids.
4. A: Random people in Canada,
   B: Random people in the U.S.;
   percent that enjoy ice hockey in Canada vs U.S.
5. A: Random people not on antidepressants,
   B: Those same people after taking antidepressants;
   percent of people that are happy off and on antidepressants.

Below is given two samples (A and B) and a proportion of interest that you want to compare across the two groups. Which of the following setups will violate the independence required between the two samples?

1. A: Random Californians,
   B: Random Texans;
   percent with college degree in CA vs TX residents
2. A: Random married men,
   B: The wives of those married men;
   percent with college degrees in married men and married women
3. A: Random adults that have kids,
   B: Kids of those adults;
   percent that believe in God in adults vs kids.
4. A: Random people in Canada,
   B: Random people in the U.S.;
   percent that enjoy ice hockey in Canada vs U.S.
5. A: Random people not on antidepressants,
   B: Those same people after taking antidepressants;
   percent of people that are happy off and on antidepressants.

Answer: 2,3 and 5.

Your Turn!

Suppose X and Y are independent random variables where $X = N(4, 3)$ and $Y = N(2, 1)$.
What will the distribution of $X - Y$ look like?

1. $N(2, 2)$
2. $N(2, 4)$
3. $N(2, \sqrt{10})$
4. $N(-2, 4)$
5. $N(-2, -2)$

Your Turn!

Suppose X and Y are independent random variables where $X = N(4,3)$ and $Y = N(2,1)$.
What will the distribution of $X - Y$ look like?

1. $N(2,2)$
2. $N(2,4)$
3. $N(2,\sqrt{10})$
4. $N(-2,4)$
5. $N(-2,-2)$

Answer: 3. $N(4-2,\sqrt{3^2+1^2}) = N(2,\sqrt{10})$

Your Turn!

You create a 90% C.I. for a difference in the proportion of Democrats and Republicans that enjoy the TV personality Stephen Colbert. You find the C.I. for $p_{dem} - p_{rep}$ is $(1\%, 5\%)$. What is the be way to report this?

1. 90% of the time, Democrats are about 1 to 5% more likely to enjoy S. Colbert.
2. 90% of the time, the percentage difference in those who enjoy S. Colbert (Democrats vs Republicans) will be between 1 and 5%.
3. The difference in the percent of Democrats and Republicans who enjoy S. Colbert is between 1 and 5%.
4. I am 90% confident that the percentage of Democrats who enjoy S. Colbert is 1 to 5% higher than the percentage of Republicans who enjoy Colbert.

Your Turn!

You create a 90% C.I. for a difference in the proportion of Democrats and Republicans that enjoy the TV personality Stephen Colbert. You find the C.I. for $p_{dem} - p_{rep}$ is $(1\%, 5\%)$. What is the be way to report this?

1. 90% of the time, Democrats are about 1 to 5% more likely to enjoy S. Colbert.
2. 90% of the time, the percentage difference in those who enjoy S. Colbert (Democrats vs Republicans) will be between 1 and 5%.
3. The difference in the percent of Democrats and Republicans who enjoy S. Colbert is between 1 and 5%.
4. I am 90% confident that the percentage of Democrats who enjoy S. Colbert is 1 to 5% higher than the percentage of Republicans who enjoy Colbert.

Answer: 4.

Does sexual orientation affect how much people prefer a certain color? In 2001, researchers explored this question with thousands of college students (`source`). Suppose the 95% C.I. for

$$p_{\text{LGBT male that likes pink}} - p_{\text{Straight male that likes pink}}$$

was calculated as $(-0.03, 0.04)$. Which of the following statements are true?

1. There is not a statistically significant difference in the percent of college-aged straight males and college-aged LGBT males who like pink.
2. The probability the true parameter difference lies in this interval is 0.95.
3. The 95% C.I. for difference in the other order

$$p_{\text{Straight male that likes pink}} - p_{\text{LGBT male that likes pink}}$$

   is $(-0.04, 0.03)$.
4. We are 95% confident that the difference in the observed proportions is in the stated interval.

Does sexual orientation affect how much people prefer a certain color? In 2001, researchers explored this question with thousands of college students (`source`). Suppose the 95% C.I. for

$$p_{\text{LGBT male that likes pink}} - p_{\text{Straight male that likes pink}}$$

was calculated as $(-0.03, 0.04)$. Which of the following statements are true?

1. There is not a statistically significant difference in the percent of college-aged straight males and college-aged LGBT males who like pink.
2. The probability the true parameter difference lies in this interval is 0.95.
3. The 95% C.I. for difference in the other order

   $$p_{\text{Straight male that likes pink}} - p_{\text{LGBT male that likes pink}}$$

   is $(-0.04, 0.03)$.
4. We are 95% confident that the difference in the observed proportions is in the stated interval.

Answer: 1.,3.

Difference in Proportions: Hypothesis Testing

We are usually interested in whether the proportions are different in our two populations.
Thus, we set H_0: $p_1 - p_2 = 0$ (or equivalently $p_1 = p_2$).

Difference in Proportions: Hypothesis Testing

We are usually interested in whether the proportions are different in our two populations.
Thus, we set H_0: $p_1 - p_2 = 0$ (or equivalently $p_1 = p_2$).

Common alternative hypotheses are:
H_A: $p_1 - p_2 > 0$
H_A: $p_1 - p_2 \neq 0$
H_A: $p_1 - p_2 < 0$

Difference in Proportions: Hypothesis Testing

We are usually interested in whether the proportions are different in our two populations.
Thus, we set H_0: $p_1 - p_2 = 0$ (or equivalently $p_1 = p_2$).

Common alternative hypotheses are:
H_A: $p_1 - p_2 > 0$
H_A: $p_1 - p_2 \neq 0$
H_A: $p_1 - p_2 < 0$

In one-sample hypothesis testing, we calculate $Z = \frac{\hat{p} - \text{null value}}{SE_{\hat{p}}}$, so you might expect we would do something similar when we have two samples:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{SE_{\hat{p}_1 - \hat{p}_2}} = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE_{\hat{p}_1 - \hat{p}_2}}.$$

Pooling Our Data

This is almost correct. But notice that $SE = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$.
This formula acts like we have two different populations going on. But if we assume H_0, then our populations are really the same (in relation to the idea we are measuring) since $p_1 = p_2$.

Pooling Our Data

This is almost correct. But notice that $SE = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$.
This formula acts like we have two different populations going on. But if we assume H_0, then our populations are really the same (in relation to the idea we are measuring) since $p_1 = p_2$.

Instead of using $\hat{p}_1$ and $\hat{p}_2$ in this formula, we create a single statistic

$$\hat{p}_{pooled} = \frac{\#\text{ Sucesses}_1 + \#\text{ Sucesses}_2}{n_1 + n_2}.$$

Pooling Our Data

This is almost correct. But notice that $SE = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$.
This formula acts like we have two different populations going on. But if we assume H_0, then our populations are really the same (in relation to the idea we are measuring) since $p_1 = p_2$.

Instead of using $\hat{p}_1$ and $\hat{p}_2$ in this formula, we create a single statistic

$$\hat{p}_{pooled} = \frac{\#\ \text{Sucesses}_1 + \#\ \text{Sucesses}_2}{n_1 + n_2}.$$

Example: If we had done hypothesis testing for the California drilling example, we would have written

$$\hat{p}_{pooled} = \frac{154 + 132}{438 + 389} \simeq 34.58\%.$$

Pooling Our Data

So, we actually use $Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE_{pooled}}$, where

$$SE_{pooled} = \sqrt{\frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_1} + \frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_2}}$$

Pooling Our Data

So, we actually use $Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - 0}{SE_{pooled}}$, where

$$SE_{pooled} = \sqrt{\frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_1} + \frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_2}}$$

Why do we pool?

Pooling Our Data

So, we actually use $Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE_{pooled}}$, where

$$SE_{pooled} = \sqrt{\frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_1} + \frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_2}}$$

Why do we pool?

The simple answer is that when you find the SE, you want to do this with the best info you have available.
Usually, this involves just using $\hat{p}_1$ and $\hat{p}_2$ in place of p_1 and p_2. If you are hypothesis testing, you assume momentarily $p_1 = p_2$ and get better approximations by using $\hat{p}_{pooled}$ in place of both $\hat{p}_1$ and $\hat{p}_2$.

Back to Drilling

Test the claim that CA college grads (Population 1, sample: 153 of 438 supported) are more interested in drilling than CA non-college grads (Population 2, sample: 132 of 389 supported).

Back to Drilling

Test the claim that CA college grads (Population 1, sample: 153 of 438 supported) are more interested in drilling than CA non-college grads (Population 2, sample: 132 of 389 supported).

We set H_0: $p_1 - p_2 = 0$ and H_A: $p_1 - p_2 > 0$.

From before, $\hat{p}_1 - \hat{p}_2 = 1.23\%$ and $\hat{p}_{pooled} = 34.58\%$, so that

$$SE_{pooled} = \sqrt{\frac{34.58 \times 65.42}{438} + \frac{34.58 \times 65.42}{389}} \simeq 3.31\%.$$

Back to Drilling

Test the claim that CA college grads (Population 1, sample: 153 of 438 supported) are more interested in drilling than CA non-college grads (Population 2, sample: 132 of 389 supported).

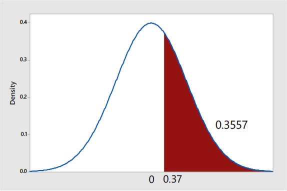
We set H_0: $p_1 - p_2 = 0$ and H_A: $p_1 - p_2 > 0$.

From before, $\hat{p}_1 - \hat{p}_2 = 1.23\%$ and $\hat{p}_{pooled} = 34.58\%$, so that

$$SE_{pooled} = \sqrt{\frac{34.58 \times 65.42}{438} + \frac{34.58 \times 65.42}{389}} \simeq 3.31\%.$$

Our z-score is $\dfrac{1.23 - 0}{3.31} \simeq 0.37$

Our p-value is $p = 0.3557$.



Back to Drilling

Test the claim that CA college grads (Population 1, sample: 153 of 438 supported) are more interested in drilling than CA non-college grads (Population 2, sample: 132 of 389 supported).

We set H_0: $p_1 - p_2 = 0$ and H_A: $p_1 - p_2 > 0$.

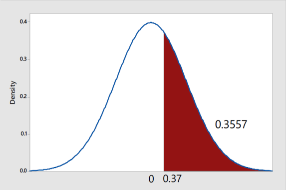
From before, $\hat{p}_1 - \hat{p}_2 = 1.23\%$ and $\hat{p}_{pooled} = 34.58\%$, so that

$$SE_{pooled} = \sqrt{\frac{34.58 \times 65.42}{438} + \frac{34.58 \times 65.42}{389}} \simeq 3.31\%.$$

Our z-score is $\dfrac{1.23 - 0}{3.31} \simeq 0.37$

Our p-value is $p = 0.3557$.

Since $0.356 > 0.05$, we do not reject the null. It is possible that both populations support drilling equally.



Back to Drilling

Test the claim that CA college grads (Population 1, sample: 153 of 438 supported) are more interested in drilling than CA non-college grads (Population 2, sample: 132 of 389 supported).

We set H_0: $p_1 - p_2 = 0$ and H_A: $p_1 - p_2 > 0$.

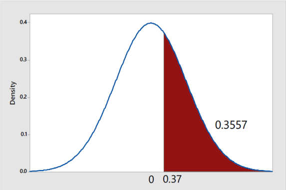
From before, $\hat{p}_1 - \hat{p}_2 = 1.23\%$ and $\hat{p}_{pooled} = 34.58\%$, so that

$$SE_{pooled} = \sqrt{\frac{34.58 \times 65.42}{438} + \frac{34.58 \times 65.42}{389}} \simeq 3.31\%.$$

Our z-score is $\dfrac{1.23 - 0}{3.31} \simeq 0.37$

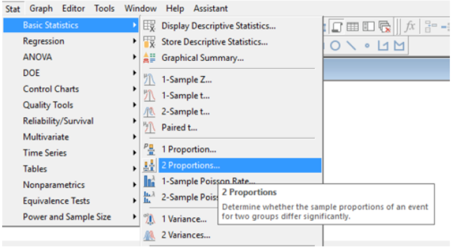
Our p-value is $p = 0.3557$.

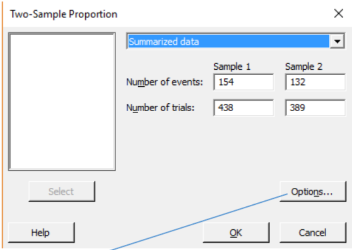
Since $0.356 > 0.05$, we do not reject the null. It is possible that both populations support drilling equally.

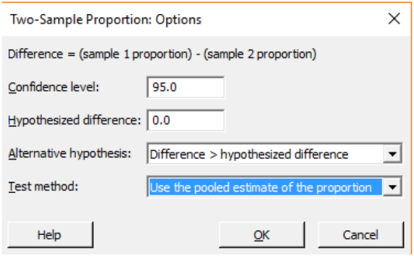


Note: Minitab also does hypothesis testing (next slide). Again, only use this to check answers.

Using Minitab








```
Difference = p (1) - p (2)
Estimate for difference:  0.0122666
95% lower bound for difference:  -0.0422079
Test for difference = 0 (vs > 0):  Z = 0.37  P-Value = 0.356
```

Sleep Time

		Transportation Professionals			
	Control	Pilots	Truck Drivers	Train Operators	Bus/Taxi/Limo Drivers
Less than 6 hours of sleep	35	19	35	29	21
6 to 8 hours of sleep	193	132	117	119	131
More than 8 hours	64	51	51	32	58
Total	292	202	203	180	210

A 2012 study from the National Sleep Foundation explored how much sleep various professions get. The above data explore sleep times for the transportation sector.

Do these data suggest that average Americans (control) are less sleep deprived (< 6 hours/night) than train operators? Do a 95% C.I. and hypothesis test.

Sleep Time

		Transportation Professionals			
	Control	Pilots	Truck Drivers	Train Operators	Bus/Taxi/Limo Drivers
Less than 6 hours of sleep	35	19	35	29	21
6 to 8 hours of sleep	193	132	117	119	131
More than 8 hours	64	51	51	32	58
Total	292	202	203	180	210

A 2012 study from the National Sleep Foundation explored how much sleep various professions get. The above data explore sleep times for the transportation sector.

Do these data suggest that average Americans (control) are less sleep deprived ($<$ 6 hours/night) than train operators? Do a 95% C.I. and hypothesis test.

Let p_T be the proportion of train operators that get <6 hours of sleep/night, and p_C the same idea in the control group.

$$\hat{p}_T = \frac{29}{180} \simeq 0.161, \qquad \hat{p}_C = \frac{35}{292} \simeq 0.120, \quad \text{so } \hat{p}_T - \hat{p}_C = 0.041.$$

Sleep Time

$$\hat{p}_T = \frac{29}{180} \simeq 0.161, \qquad \hat{p}_C = \frac{35}{292} \simeq 0.120, \quad \text{so } \hat{p}_T - \hat{p}_C = 0.041.$$

Confidence Interval: **Do not** use a pooled estimate for the C.I's:

$$\begin{aligned} SE &\simeq \sqrt{\frac{\hat{p}_T\hat{q}_T}{n_T} + \frac{\hat{p}_C\hat{q}_C}{n_C}} \\ &= \sqrt{\frac{0.161 \times 0.839}{180} + \frac{0.12 \times 0.88}{292}} \simeq 0.033. \end{aligned}$$

So,

$$\begin{aligned} CI &= \hat{p}_T - \hat{p}_C \pm z^* \times SE \\ &= 0.041 \pm 1.96 \times 0.033 \\ &= (-0.024, 0.106). \end{aligned}$$

Sleep Time

$$\hat{p}_T = \frac{29}{180} \simeq 0.161, \qquad \hat{p}_C = \frac{35}{292} \simeq 0.120, \quad \text{so } \hat{p}_T - \hat{p}_C = 0.041.$$

Hypothesis Test: Set H_0: $p_T - p_C = 0$ and $p_T - p_C > 0$.

Under the null, you can (and should!) pool the data and get

$$\hat{p}_{pooled} = \frac{29 + 35}{180 + 292} \simeq 0.135.$$

We get a slightly better estimate for the SE:

$$\begin{aligned} SE_{pooled} &\simeq \sqrt{\frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_T} + \frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_C}} \\ &= \sqrt{\frac{0.135 \times 0.865}{180} + \frac{0.135 \times 0.865}{292}} \simeq 0.032. \end{aligned}$$

(Compare with 0.033 from CI slide. Pooled estimates only differ slightly.)

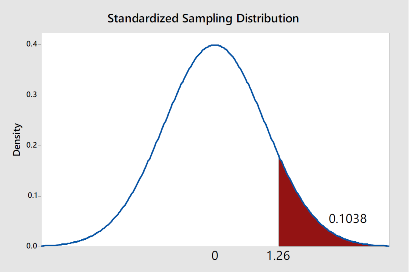
Sleep Time

The z-score of data is $Z = \dfrac{(\hat{p}_T - \hat{p}_C) - 0}{SE_{pooled}} \simeq \dfrac{0.041}{0.032} \simeq 1.26.$

Sleep Time

The z-score of data is $Z = \dfrac{(\hat{p}_T - \hat{p}_C) - 0}{SE_{pooled}} \simeq \dfrac{0.041}{0.032} \simeq 1.26.$

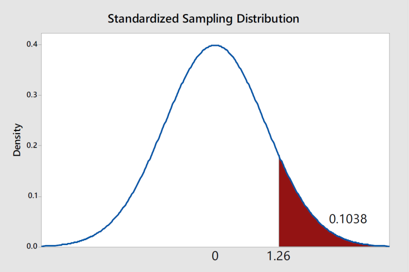
Using Minitab, we find the p-value is $p = 0.1038$.



Sleep Time

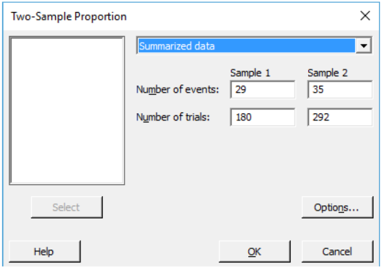
The z-score of data is $Z = \dfrac{(\hat{p}_T - \hat{p}_C) - 0}{SE_{pooled}} \simeq \dfrac{0.041}{0.032} \simeq 1.26.$

Using Minitab, we find the p-value is $p = 0.1038$.



Since $p = 0.10 > 0.05$, we do not reject the null.
It appears that average Americans are not less sleep deprived than train operators.

Checking Your Answer in Minitab




```
Sample    X    N  Sample p
1        29  180  0.161111
2        35  292  0.119863


Difference = p (1) - p (2)
Estimate for difference:  0.0412481
95% CI for difference:  (-0.0241144, 0.106611)      Test for difference = 0 (vs > 0):  Z = 1.27  P-Value = 0.102
```