POLI 30: Political Inquiry

Heidi McNamara

University of California, San Diego

h1hall@ucsd.edu

May 13, 2016

Reminders

Problem sets handed back at the end of class today.

Email: h1hall@ucsd.edu

Office Hours: Wed. 11-1, SSB 343

Hannah's SPSS OH: Tuesdays 12:30-1:45, Sequoyah 142

Website: heidimcnamara.com

Quiz

Describe the difference between sampling bias and sampling variability.

Hint: You might want to think about whether they are related to your observed sample or other (potentially unobserved) samples from the same population. You might also want to think about whether or not they are accounted for in a sample statistic's margin of error.

Reminder: Difference of Proportions Test

Last week (and on your homework) we ran a *difference of proportions* test with the following formula:

$$(\hat{p}_2 - \hat{p}_1) \pm 2 * \sqrt{(\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{N_2}})^2 + (\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{N_1}})^2}$$

Difference of Means Test

You can also run a *difference of means* test. The formula is the same, but the calculation of the standard error is slightly different (as you know).

$$(\bar{X}_2 - \bar{X}_1) \pm 2 * \sqrt{(SE_1)^2 + (SE_2)^2}$$

Where:

$$SE_i = \frac{\hat{\sigma}_i}{\sqrt{N_i}}$$

Difference of Means Test, Step 1: Sample

Take a sample (hypothetical data) of students' GPA by hair color:

Student	Hair Color	GPA
1	Brown	3.2
2	Red	3.9
3	Brown	4.0
4	Brown	2.5
5	Brown	3.0
6	Red	3.4
7	Brown	3.7
8	Red	3.6
9	Brown	3.0
10	Brown	2.7

Difference of Means Test, Step 2: Separate Sample

Take a sample (hypothetical data) of students' GPA by hair color:

Student	Hair Color	GPA
1	Brown	3.2
3	Brown	4.0
4	Brown	2.5
5	Brown	3.0
7	Brown	3.7
9	Brown	3.0
10	Brown	2.7
Mean		3.16
2	Red	3.9
6	Red	3.4
8	Red	3.6
Mean		3.63

Difference of Means Test, Step 3: Difference from Mean

Take a sample (hypothetical data) of students' GPA by hair color:

Student	Hair Color	GPA	$(X_i - \bar{X})$
1	Brown	3.2	0.04
3	Brown	4.0	0.84
4	Brown	2.5	-0.64
5	Brown	3.0	-0.16
7	Brown	3.7	0.54
9	Brown	3.0	-0.16
10	Brown	2.7	-0.46
Mean		3.16	
2	Red	3.9	0.27
6	Red	3.4	-0.23
8	Red	3.6	-0.03
Mean		3.63	

Difference of Means Test, Step 4: Square those

Take a sample (hypothetical data) of students' GPA by hair color:

Student	Hair Color	GPA	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	Brown	3.2	0.04	0.0016
3	Brown	4.0	0.84	0.7056
4	Brown	2.5	-0.64	0.4096
5	Brown	3.0	-0.16	0.0256
7	Brown	3.7	0.54	0.2916
9	Brown	3.0	-0.16	0.0256
10	Brown	2.7	-0.46	0.2116
Mean		3.16		
2	Red	3.9	0.27	0.0729
6	Red	3.4	-0.23	0.0529
8	Red	3.6	-0.03	0.0.0009
Mean		3.63		

Difference of Means Test, Step 5: Calculate SD

Student	Hair Color	GPA	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	Brown	3.2	0.04	0.0016
3	Brown	4.0	0.84	0.7056
4	Brown	2.5	-0.64	0.4096
5	Brown	3.0	-0.16	0.0256
7	Brown	3.7	0.54	0.2916
9	Brown	3.0	-0.16	0.0256
10	Brown	2.7	-0.46	0.2116
Mean		3.16	**Sum**	1.6712
2	Red	3.9	0.27	0.0729
6	Red	3.4	-0.23	0.0529
8	Red	3.6	-0.03	0.0.0009
Mean		3.63	**Sum**	0.1267

Brown Hair $\hat{\sigma}$: $\sqrt{\frac{1.6712}{7-1}} = \sqrt{0.2785} = 0.52$

Red Hair $\hat{\sigma}$: $\sqrt{\frac{0.1267}{3-1}} = \sqrt{0.06335} = 0.25$

Now take those standard deviations to calculate the standard errors

Where:

$$SE_i = \frac{\hat{\sigma}_i}{\sqrt{N_i}}$$

Now take those standard deviations to calculate the standard errors

Where:

$$SE_i = \frac{\hat{\sigma}_i}{\sqrt{N_i}}$$

$$SE_{BrownHair} = \frac{0.52}{\sqrt{7}} = 0.19$$

$$SE_{RedHair} = \frac{0.25}{\sqrt{3}} = 0.14$$

Plug in what we know...

Formula:

$$(\bar{X}_2 - \bar{X}_1) \pm 2 * \sqrt{(SE_1)^2 + (SE_2)^2}$$

Plug in what we know...

Formula:

$$(\bar{X}_2 - \bar{X}_1) \pm 2 * \sqrt{(SE_1)^2 + (SE_2)^2}$$

Filled in:

$$(3.63 - 3.16) \pm 2 * \sqrt{(0.19)^2 + (0.14)^2} =$$

$$0.47 \pm 2 * \sqrt{0.0557} = 0.47 \pm 0.472$$

So our 95% Confidence Interval...

So our 95% Confidence Interval...

We can say with 95 percent confidence that the true difference in mean GPAs between students with red hair and students with brown hair is somewhere between -0.002 and 0.942 points.

So can we reject our null hypothesis that there is not true population difference in GPA's based on students' hair color?

So our 95% Confidence Interval...

We can say with 95 percent confidence that the true difference in mean GPAs between students with red hair and students with brown hair is somewhere between -0.002 and 0.942 points.

So can we reject our null hypothesis that there is not true population difference in GPA's based on students' hair color?

No. We fail to reject the null hypothesis.

What do you do when you have more than two categories?

What do you do when you have more than two categories?

Chi-square test!

More Hypothetical Hair Color Data...

Hair Color	# with GPA <3.5	# with GPA >=3.5
Brown	5	7
Blonde	7	8
Black	4	6
Red	8	6

First Step: Calculate Totals

Hair Color	# with GPA <3.5	# with GPA >=3.5	Total
Brown	5	7	12
Blonde	7	8	15
Black	4	6	10
Red	8	6	14
Total	24	27	51

Second Step: Calculate Expected Numbers

	Observed		Expected	
Hair Color	GPA $<$3.5	GPA $>=$3.5	GPA $<$3.5	GPA $>=$3.5
Brown	5	7		
Blonde	7	8		
Black	4	6		
Red	8	6		

Second Step: Calculate Expected Numbers

$$ExpectedCount = \frac{RowTotal * ColumnTotal}{TableTotal}$$

Second Step: Calculate Expected Numbers

	Observed		Expected	
Hair Color	GPA <3.5	GPA >=3.5	GPA <3.5	GPA >=3.5
Brown	5	7	6	6
Blonde	7	8	7	8
Black	4	6	5	5
Red	8	6	7	7

Third Step: Calculate Chi-Square Statistic

Formula:

$$\chi^2 = \sum \frac{(ObservedCount_i - ExpectedCount_i)^2}{ExpectedCount_i}$$

Third Step: Calculate Chi-Square Statistic

	Observed		Expected		$\frac{(Observed - Expected)^2}{Expected}$	
Hair	<3.5	>=3.5	<3.5	>=3.5	< 3.5	$>= 3.5$
Brown	5	7	6	6	0.1667	0.1667
Blonde	7	8	7	8	0	0
Black	4	6	5	5	0.2	0.2
Red	8	6	7	7	0.1428	0.1428

Third Step: Calculate Chi-Square Statistic

	Observed		Expected		$\frac{(Observed-Expected)^2}{Expected}$	
Hair	<3.5	>=3.5	<3.5	>=3.5	< 3.5	$>= 3.5$
Brown	5	7	6	6	0.1667	0.1667
Blonde	7	8	7	8	0	0
Black	4	6	5	5	0.2	0.2
Red	8	6	7	7	0.1428	0.1428
Sum					0.5095	0.5095

$\chi^2 = 1.019$

Fourth Step: Calculate Degrees of Freedom

Formula:

$$df = (\# \text{ of rows - } 1) * (\# \text{ of columns - } 1)$$

Fourth Step: Calculate Degrees of Freedom

Formula:

$$\text{df} = (\#\text{ of rows - 1}) * (\#\text{ of columns - 1})$$

In our case:

$$\text{df} = (4 - 1) * (2 - 1) = 3$$

Reject or Fail to Reject the Null?

$\chi^2 = 1.019$
Degrees of freedom = 3

Degrees of Freedom	95% Threshold
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07
6	12.59
7	14.07
8	15.51
9	16.92

Reject or Fail to Reject the Null?

$\chi^2 = 1.019$
Degrees of freedom = 3

Degrees of Freedom	95% Threshold
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07
6	12.59
7	14.07
8	15.51
9	16.92

Because 1.019 is smaller than 7.81, we fail to reject the null.

Now try one on your own... Bowling Alone?

Robert Putnam's hypothesis is that those who watch more TV are more likely to bowl alone. Using the observed data in this table, conduct a chi-square test of Putnam's hypothesis. Tell me what the null hypothesis is (substantively), and construct a table showing me what data you would expect to see if the null hypothesis were true (expected counts). Calculate the chi-square statistic (if you need to round some of the numbers, go ahead), and compare it to the appropriate threshold value. First, tell me whether or not we can reject the null hypothesis. Second, tell me if the data in this table support Putnam's hypothesis about the relationship between television watching and bowling alone.

Time spend watching TV	# Who Bowl Alone	# in League
Zero hours per week	10	2
1-10 hours per week	18	2
$>$ 10 hours per week	2	6

Homework 3 Hints

Part A: Review section slides (available at heidimcnamara.com) from this week and last week for step-by-step processes and formulas for each of these tests of significance. Remember to calculate 99.7% confidence intervals, you multiple the standard errors by 3 (instead of the 2 used for 95% confidence intervals).

Part B: If you didn't save your data after the last problem set, remember to re-recode any data you recoded before (or do it now, if you didn't last time but should have). Have some fun making graphs. Explore different types and turn in the one that portrays the most information in the best way. Include titles and labels that actually say what the variables are, rather than just "V15". Be sure to pick an intervening or confounding variable that is **actually in your dataset**.