Unit 5: Inference for categorical data

2. Comparing two proportions

Sta 101 - Spring 2015

Duke University, Department of Statistical Science

March 18, 2015

Dr. Çetinkaya-Rundel

Slides posted at http://bitly.com/sta101sp15

1. Housekeeping

2. Main ideas
   1. CLT also describes the distribution of $\hat{p}_1 - \hat{p}_2$
   2. For HT where $H_0 : p_1 = p_2$, pool!
   3. When S-F fails, simulate!

3. Applications
   1. Two population proportions, small sample
   2. Single population proportion, small sample

4. Summary

- MT2 Review - Monday, March 23, 7-8pm
- OH next week - Monday and Tuesday 3-5pm
- MT review materials to be posted on Sakai over the weekend
- RA5 opens on Sunday (but will also cover material from Monday's class so you might want to wait to take it) and will close at midnight on Monday
- Mao's office hours moved to tomorrow 7-9pm, Tori is there tonight 7-9pm.

1. Housekeeping

2. Main ideas
    1. CLT also describes the distribution of $\hat{p}_1 - \hat{p}_2$
    2. For HT where $H_0 : p_1 = p_2$, pool!
    3. When S-F fails, simulate!

3. Applications
    1. Two population proportions, small sample
    2. Single population proportion, small sample

4. Summary

1. Housekeeping

2. Main ideas
    1. CLT also describes the distribution of $\hat{p}_1 - \hat{p}_2$
    2. For HT where $H_0 : p_1 = p_2$, pool!
    3. When S-F fails, simulate!

3. Applications
    1. Two population proportions, small sample
    2. Single population proportion, small sample

4. Summary

$$(\hat{p}_1 - \hat{p}_2) \sim N\left(mean = (p_1 - p_2), SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

Conditions:

- Independence: Random sample/assignment + 10% rule
- Sample size / skew: At least 10 successes and failures

1. Housekeeping

2. Main ideas
    1. CLT also describes the distribution of $\hat{p}_1 - \hat{p}_2$
    2. For HT where $H_0 : p_1 = p_2$, pool!
    3. When S-F fails, simulate!

3. Applications
    1. Two population proportions, small sample
    2. Single population proportion, small sample

4. Summary

As with working with a single proportion,

- When doing a HT where $H_0 : p_1 = p_2$ (almost always for HT), use expected counts / proportions for S-F condition and calculation of the standard error.
- Otherwise use observed counts / proportions for S-F condition and calculation of the standard error.

As with working with a single proportion,

- When doing a HT where $H_0 : p_1 = p_2$ (almost always for HT), use expected counts / proportions for S-F condition and calculation of the standard error.
- Otherwise use observed counts / proportions for S-F condition and calculation of the standard error.

Expected proportion of success for both groups when $H_0 : p_1 = p_2$ is defined as the *pooled proportion*:

$$\hat{p}_{pool} = \frac{total\ successes}{total\ sample\ size} = \frac{suc_1 + suc_2}{n_1 + n_2}$$

Clicker question

Suppose in group 1 30 out of 50 observations are successes, and in group 2 20 out of 60 observations are successes. What is the pooled proportion?

(a) $\frac{30}{50}$

(b) $\frac{20}{60}$

(c) $\frac{30}{50} + \frac{20}{60}$

(d) $\frac{30+20}{50+60}$

(e) $\frac{\frac{30}{50}+\frac{20}{60}}{2}$

Clicker question

Suppose in group 1 30 out of 50 observations are successes, and in group 2 20 out of 60 observations are successes. What is the pooled proportion?

(a) $\frac{30}{50}$

(b) $\frac{20}{60}$

(c) $\frac{30}{50} + \frac{20}{60}$

(d) $\frac{30+20}{50+60}$

(e) $\frac{\frac{30}{50}+\frac{20}{60}}{2}$

1. Housekeeping

2. Main ideas
    1. CLT also describes the distribution of $\hat{p}_1 - \hat{p}_2$
    2. For HT where $H_0 : p_1 = p_2$, pool!
    3. When S-F fails, simulate!

3. Applications
    1. Two population proportions, small sample
    2. Single population proportion, small sample

4. Summary

- If the S-F condition is met, can do theoretical inference: Z test, Z interval
- If the S-F condition is not met, must use simulation based methods: randomization test, bootstrap interval

1. Housekeeping

2. Main ideas
    1. CLT also describes the distribution of $\hat{p}_1 - \hat{p}_2$
    2. For HT where $H_0 : p_1 = p_2$, pool!
    3. When S-F fails, simulate!

3. Applications
    1. Two population proportions, small sample
    2. Single population proportion, small sample

4. Summary

1. Housekeeping

2. Main ideas
    1. CLT also describes the distribution of $\hat{p}_1 - \hat{p}_2$
    2. For HT where $H_0 : p_1 = p_2$, pool!
    3. When S-F fails, simulate!

3. Applications
    1. Two population proportions, small sample
    2. Single population proportion, small sample

4. Summary

``Healthy adults immunized with an experimental malaria vaccine, called PfSPZ may be completely protected from infection, according to government researchers." reported Time magazine in Aug 2013. The vaccine contains weakened forms of the live parasite -- *Plasmodium falciparum* -- responsible for causing malaria. In a randomized trial, none of the six patients who received the vaccine developed malaria, while five of the six who were not vaccinated became infected. Do these data provide convincing evidence of a difference in rate of malaria infection?

``Healthy adults immunized with an experimental malaria vaccine, called PfSPZ may be completely protected from infection, according to government researchers." reported Time magazine in Aug 2013. The vaccine contains weakened forms of the live parasite -- *Plasmodium falciparum* -- responsible for causing malaria. In a randomized trial, none of the six patients who received the vaccine developed malaria, while five of the six who were not vaccinated became infected. Do these data provide convincing evidence of a difference in rate of malaria infection?

		Outcome		
		Malaria	No malaria	
Group	Vaccine	0	6	6
	No vaccine	5	1	6
	Total	5	7	12

http://healthland.time.com/2013/08/09/malaria-vaccine-shows-strongest-protection-yet-against-parasite/

		Outcome		
		Malaria	No malaria	
Group	Vaccine	0	6	6
	No vaccine	5	1	6
	Total	5	7	12

		Outcome		
		Malaria	No malaria	
Group	Vaccine	0	6	6
	No vaccine	5	1	6
	Total	5	7	12

$$H_0 : p_T = p_C \qquad H_A : p_T \neq p_C$$

		Outcome		
		Malaria	No malaria	
Group	Vaccine	0	6	6
	No vaccine	5	1	6
	Total	5	7	12

$$H_0 : p_T = p_C \qquad H_A : p_T \neq p_C$$

Conditions:

1. Independence: Patients are randomly assigned to treatment groups
2. Success-failure: ?

Clicker question

Assuming that the null hypothesis ($H_0 : p_T = p_C$) is true, which of the following is the pooled proportion of patients with malaria in the two groups?

(a) $\frac{6}{12} = 0.5$

(b) $\frac{5}{12} = 0.417$

(c) $\frac{0}{5} = 0$

(d) $\frac{6}{7} = 0.857$

(e) $\frac{7}{12} = 0.583$

		Outcome		
		Malaria	No malaria	
Group	Vaccine	0	6	6
	No vaccine	5	1	6
	Total	5	7	12

Clicker question

Assuming that the null hypothesis ($H_0 : p_T = p_C$) is true, which of the following is the pooled proportion of patients with malaria in the two groups?

(a) $\frac{6}{12} = 0.5$

(b) $\frac{5}{12} = 0.417$

(c) $\frac{0}{5} = 0$

(d) $\frac{6}{7} = 0.857$

(e) $\frac{7}{12} = 0.583$

		Outcome		
		Malaria	No malaria	
Group	Vaccine	0	6	6
	No vaccine	5	1	6
	Total	5	7	12

Clicker question

Assuming that the null hypothesis ($H_0 : p_T = p_C$) is true, how many patients would we expect to get infected with malaria in the vaccine group?

(a) $0.417 \times 12 = 5$

(b) $0.417 \times 6 = 2.5$

(c) $0.417 \times 5 = 2.085$

(d) $0.5 \times 6 = 3$

(e) $0.583 \times 12 = 7$

		Outcome		
		Malaria	No malaria	
Group	Vaccine	0	6	6
	No vaccine	5	1	6
	Total	5	7	12

Clicker question

Assuming that the null hypothesis ($H_0 : p_T = p_C$) is true, how many patients would we expect to get infected with malaria in the vaccine group?

(a) $0.417 \times 12 = 5$

(b) $0.417 \times 6 = 2.5$

(c) $0.417 \times 5 = 2.085$

(d) $0.5 \times 6 = 3$

(e) $0.583 \times 12 = 7$

		Outcome		
		Malaria	No malaria	
Group	Vaccine	0	6	6
	No vaccine	5	1	6
	Total	5	7	12

Clicker question

Assuming that the null hypothesis ($H_0 : p_T = p_C$) is true, how many patients would we expect to get infected with malaria in the vaccine group?

(a) $0.417 \times 12 = 5$

(b) $0.417 \times 6 = 2.5$

(c) $0.417 \times 5 = 2.085$

(d) $0.5 \times 6 = 3$

(e) $0.583 \times 12 = 7$

		Outcome		
		Malaria	No malaria	
Group	Vaccine	0	6	6
	No vaccine	5	1	6
	Total	5	7	12

$$\hat{p}_{pool} = 5/12 = 0.417$$
$$1 - 0.417 = 0.583$$

$$Exp\ S_T = 0.417 \times 6 = 2.5 \qquad Exp\ S_C = 0.417 \times 6 = 2.5$$
$$Exp\ F_T = 0.583 \times 6 = 3.5 \qquad Exp\ F_C = 0.583 \times 6 = 3.5$$

1. Use 12 index cards, where each card represents an experimental unit.

1. Use 12 index cards, where each card represents an experimental unit.
2. Mark 5 of the cards as ``malaria" and the remaining 7 as ``no malaria".

1. Use 12 index cards, where each card represents an experimental unit.
2. Mark 5 of the cards as ``malaria" and the remaining 7 as ``no malaria".
3. Shuffle the cards and split into two groups of size 6, for vaccine and no vaccine.

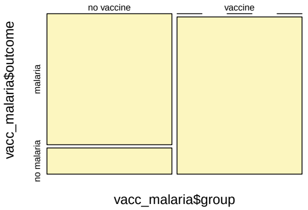
1. Use 12 index cards, where each card represents an experimental unit.
2. Mark 5 of the cards as ``malaria" and the remaining 7 as ``no malaria".
3. Shuffle the cards and split into two groups of size 6, for vaccine and no vaccine.
4. Calculate the difference between the proportions of ``malaria" in the vaccine and no vaccine decks, and record this number.

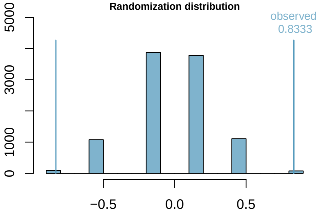
1. Use 12 index cards, where each card represents an experimental unit.
2. Mark 5 of the cards as ``malaria" and the remaining 7 as ``no malaria".
3. Shuffle the cards and split into two groups of size 6, for vaccine and no vaccine.
4. Calculate the difference between the proportions of ``malaria" in the vaccine and no vaccine decks, and record this number.
5. Repeat steps (3) and (4) many times to build a randomization distribution of differences in simulated proportions.

```
download("https://stat.duke.edu/~mc301/data/vacc_malaria.csv", destfile = "vacc_malaria.csv")
vacc_malaria = read.csv("vacc_malaria.csv")

inference(vacc_malaria$outcome, vacc_malaria$group, success = "malaria", est = "proportion",
    type = "ht", null = 0, alternative = "twosided", method = "simulation", seed = 1028)
```

```
download("https://stat.duke.edu/~mc301/data/vacc_malaria.csv", destfile = "vacc_malaria.csv")
vacc_malaria = read.csv("vacc_malaria.csv")

inference(vacc_malaria$outcome, vacc_malaria$group, success = "malaria", est = "proportion",
    type = "ht", null = 0, alternative = "twosided", method = "simulation", seed = 1028)
```

```
Response variable: categorical, Explanatory variable: categorical
Difference between two proportions -- success: malaria
Summary statistics:
            x
y             no vaccine vaccine Sum
  malaria              5       0   5
  no malaria           1       6   7
  Sum                  6       6  12
Observed difference between proportions (no vaccine-vaccine) = 0.8333
H0: p_no vaccine - p_vaccine = 0
HA: p_no vaccine - p_vaccine != 0
p-value =  0.0152
```





1. Housekeeping

2. Main ideas
    1. CLT also describes the distribution of $\hat{p}_1 - \hat{p}_2$
    2. For HT where $H_0 : p_1 = p_2$, pool!
    3. When S-F fails, simulate!

3. Applications
    1. Two population proportions, small sample
    2. Single population proportion, small sample

4. Summary

Application exercise: App Ex 5.2

See course website for details.

1. Housekeeping

2. Main ideas
    1. CLT also describes the distribution of $\hat{p}_1 - \hat{p}_2$
    2. For HT where $H_0 : p_1 = p_2$, pool!
    3. When S-F fails, simulate!

3. Applications
    1. Two population proportions, small sample
    2. Single population proportion, small sample

4. Summary

1. CLT also describes the distribution of $\hat{p}_1 - \hat{p}_2$
2. For HT where $H_0 : p_1 = p_2$, pool!
3. When S-F fails, simulate!