Chapter 6 Inference for categorical data

Huamei Dong
03/17/2016

Last class, we talked about the inference for a single proportion, that is, confidence interval and hypothesis test about one proportion. Today we will learn the inference for the difference of two proportions.

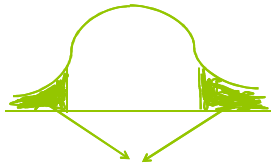
1. **Quick Review**
2. **Sampling distribution of differences of two proportions**
3. **Confidence interval for p_1-p_2**
4. **Hypothesis test when H_0: $p_1=p_2$**

1.Quick Review

(1) Confidence interval for one proportion

$$(\hat{p}-z^*SE_{\hat{p}}, \hat{p}+z^*SE_{\hat{p}})=(\hat{p}-z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p}+z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$$

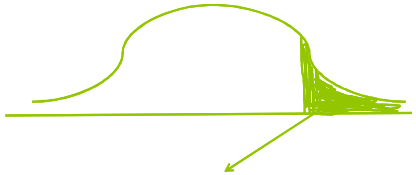
(2) For H_0: p=0.5, H_A: p≠0.5 and significant level is 0.05,



Calculate Z value using your sample data: $Z=\frac{\hat{p}-0.5}{\sqrt{\frac{0.5(1-0.5)}{n}}}$

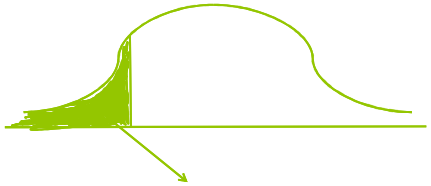
Find the p-value which should be the area of two tails. If the two tails' area is less than 0.05 (or one tail's area is less than 0.025) , we reject H_0.

(3) For H_0: p=0.5, H_A: p>0.5 and significant level is 0.05,



Calculate Z value using your sample data. Find the p-value which is the right tail Area. If this right tail area is less than 0.05, we reject H_0.

(4) For H_0: p=0.5, H_A: p<0.5 and significant level is 0.05, we



Calculate Z value using your sample data. Find the p-value which is the left tail Area. If this left tail area is less than 0.05, we reject H_0.

2. Sampling distribution of the difference of two proportions

Sampling distribution of $\hat{p}_1-\hat{p}_2$:

Assume the true proportion for population 1 is p_1 , sample size from population 1 is n_1 , the true proportion for population 2 is p_2 sample size from population 2 is n_2 .

(1) If samples from population 1 are independent of each other and

$$n_1p_1\geq 10,\ n_1(1-p_1)\geq 10$$

(2) If samples from population 2 are independent of each other and

$$n_2p_2\geq 10,\ n_2(1-p_2)\geq 10$$

(3) Samples from population 1 and samples from population 2 are independent.

Then the sampling distribution of $\hat{p}_1-\hat{p}_2$ is nearly normal with mean p_1-p_2

And standard error

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

3. Confidence interval for p_1-p_2

Constructing a confidence interval for a proportion.

(1) Verify the observations are independent and verify the success-failure condition using n_1 and $\hat{p}_1$, n_2 and $\hat{p}_2$

(2) If the condition are met, the sampling distribution of $\hat{p}_1-\hat{p}_2$ is nearly normal.

(3) Standard error can be approximated by

$$SE\approx\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

(4) Confidence interval is

$$(\hat{p}_1-\hat{p}_2-Z^*\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\ ,\ \hat{p}_1-\hat{p}_2+Z^*\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\ )$$

4. Hypothesis test when H_0: $p_1=p_2$

If we do the hypothesis test when H_0: $p_1=p_2$, we use

$$\hat{p}=\frac{total\ number\ of\ successes\ from\ both\ populations}{total\ number\ of\ cases\ from\ both\ populations}=\frac{\hat{p}_1n_1+\hat{p}_2n_2}{n_1+n_2}$$

to verify the success-failure condition, that is to verify

$$n_1\hat{p}\geq 10,\ n_1(1-\hat{p})\geq 10\ \ and\ \ n_2\hat{p}\geq 10,\ n_2(1-\hat{p})\geq 10$$

We also calculate standard error using $\hat{p}$, that is

$$SE=\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1}+\frac{\hat{p}(1-\hat{p})}{n_2}}$$

Example1 There were 50 patients in the experiment who did not receive the blood thinner and 40 patients who did.(1) What is the observed survival rate in the control group? (2) And in the treatment group?(3) Provide a point estimate of the difference in survival proportion of the two groups $\hat{p}_t - \hat{p}_c$.(4) Find 95% confidence interval for $p_t - p_c$ (5) Complete hypothesis test whether the blood thinner are helpful or harmful at significant level of 0.05.

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

Answer: (1) $\hat{p}_c = \frac{11}{50} = 0.22$

(2) $\hat{p}_t = \frac{14}{40} = 0.35$

(3) $\hat{p}_t - \hat{p}_c = 0.35 - 0.22 = 0.13$

(4) To find 95% confidence interval, we need check the success-failure condition:

We can just check the table to see if the successes are bigger or equal to 10, the failures are bigger or equal to 10. Or you can use

$$n_c=50,\ \hat{p}_c=\frac{11}{50}=0.22,\ n_t=40,\ \hat{p}_t=\frac{14}{40}$$

$$to\ check\ if\ n_c\hat{p}_c \geq 10,\ n_c(1-\hat{p}_c)\geq 10,\ n_t\hat{p}_t \geq 10,\ n_t(1-\hat{p}_t)\geq 10$$

Standard error is

$$SE \approx \sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_c}+\frac{\hat{p}_t(1-\hat{p}_t)}{n_t}} = \sqrt{\frac{(0.22)(1-0.22)}{50}+\frac{0.35(1-0.35)}{40}} = 0.0955$$

So the confidence interval is

$$(\hat{p}_t-\hat{p}_c-1.96SE,\ \hat{p}_t-\hat{p}_c+1.96SE)$$
$$that\ is\ (0.13-1.96\times 0.0955,\ 0.13+1.96\times 0.0955)$$
$$which\ is\ (-0.057,\ 0.317)$$

(5) $H_0: p_t = p_c$ $H_A: p_t \neq p_c$

Check the success-failure condition using

$$\hat{p} = \frac{total\ successes\ in\ both\ samples}{total\ cases\ in\ both\ samples} = \frac{11+14}{50+40} = 0.278$$

$$and\ n_t=40\ ,\ n_c=50$$

$$Here\ n_t\hat{p}=40(0.278)=11.12\geq 10,\ n_t(1-\hat{p})=40(1-0.278)\geq 10$$
$$n_c\hat{p}=50(0.278)=13.9\geq 10,\ n_c(1-\hat{p})=50(1-0.278)\geq 10$$

So the conditions is satisfied.

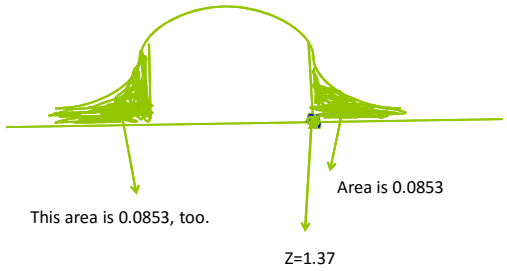
Now we need estimate the standard error using $\hat{p}$

$$SE \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n_t}+\frac{\hat{p}(1-\hat{p})}{n_c}} = \sqrt{\frac{0.278(1-0.278)}{40}+\frac{0.278(1-0.278)}{50}} = 0.095$$

$$Z=\frac{(\hat{p}_t-\hat{p}_c)-null\ \ value}{SE}=\frac{0.13-0}{0.095}=1.37$$

If we use the Z table, we find the one side tail is 0.0853. So the p-value is 0.176.

So we don't reject H_0.



Homework on 03/17/16: (due 03/24/16)

1. Using the data (breast_cancel.txt)
(a) To calculate a 95% confidence interval for the difference between the proportion of women under 55 for those who underwent a radical mastectomy and the proportion of women under 55 for those who underwent a partial mastectomy accompanied by radiation therapy.
(b) Test whether two proportions are equal at significance level of 0.05.

2. Using the data you sampled on 03/15/16 and the data your classmate sampled:
(a) To calculate a 95% confidence interval for the difference between two proportions
(b) Test whether two proportions are equal at significance level of 0.05.