Chapter 6 Inference for categorical data

Huamei Dong

03/22/2016

1. **Review of hypothesis test when H_0: $p_1=p_2$ or $p_1-p_2=0$**
2. **Hypothesis test when H_0: p_1-p_2=some non-zero number**
3. **Summary of inferences for proportions**
4. **Testing for goodness of fit using chi-square**
5. **Chi-square distribution and p-value**
6. **Test for independence in two-way table using chi-square**

1. Review of hypothesis test for H_0: p_1-p_2=0 or p_1=p_2

We have learned the hypothesis test for H_0: p_1-p_2=0 or p_1=p_2. In the test, we use

$$\hat{p} = \frac{total\ number\ of\ successes\ from\ both\ populations}{total\ number\ of\ cases\ from\ both\ populations} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

to calculate the standard error

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

In this test, we assume H_0 is true and try to find p-value. If H_0 is true, the two population proportions are equal and we should use one sample proportion, the pooled proportion estimate, to calculate standard error.

2. Hypothesis test for H_0: p_1-p_2=c (some constant not equal to 0)

When we test for H_0: p_1-p_2=some non-zero number, we still use $\hat{p}_1$ and $\hat{p}_2$ to estimate the standard error

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

Example 1 There were 50 patients in the experiment who did not receive the blood thinner and 40 patients who did.

	Survived	Died	Total
control	11	39	50
treatment	14	26	40
total	25	65	90

Does this provide convincing evidence for the claim that blood thinners improve survival rate more than 8% using significant level 0f 0.05?

Answer: (1) H_0: $p_t - p_c = 0.08$, H_A: $p_t - p_c > 0.08$

(2) Check the success-failure condition: Using

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

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

(3) Point estimate for $p_t - p_c$ is

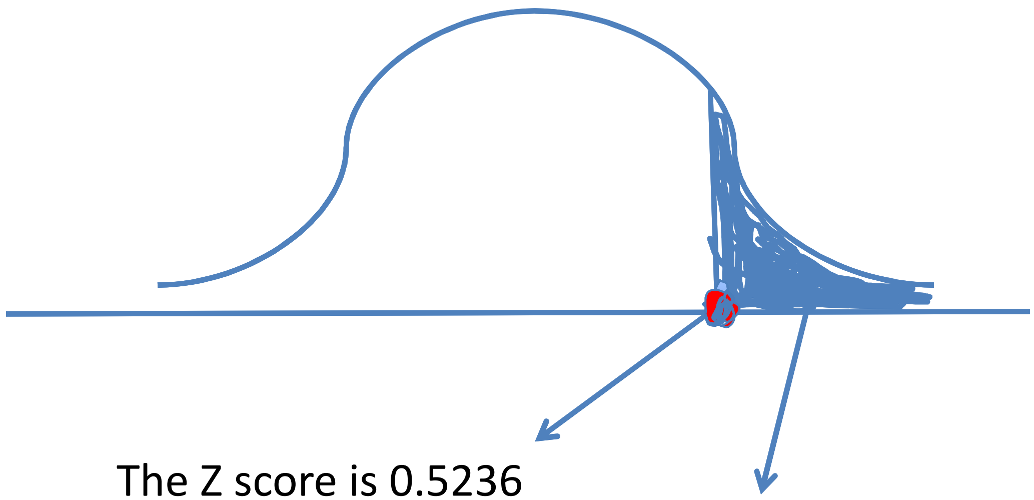
$$\hat{p}_t - \hat{p}_c = 0.35 - 0.22 = 0.13$$

(4) Standard error is

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

(5) Now we calculate Z score and find the p-value.

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.13 - 0.08}{0.0955} = 0.5236$$



(6) Since p-value is 0.3 which is big than 0.05, we don't reject $H_{0.}$ That is we don't have convincing evidence for improvement of 8% survival rate.

3. Summary of inferences for proportions

$\hat{p}_1$: Point estimate for p_1, $\hat{p}_2$: Point estimate for p_2	
95% confidence interval for p_1 $\left(\hat{p}_1 - 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}},\ \hat{p}_1 + 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}\right)$	
Hypothesis test for H_0: $p_1=0.5$ Using $SE=\sqrt{\frac{0.5(1-0.5)}{n_1}}$ to calculate z score $Z=\frac{\hat{p}_1 - null\ \ value}{SE}=\frac{\hat{p}_1-0.5}{SE}$ and p-value	
$\hat{p}_1-\hat{p}_2$ Point estimate for p_1-p_2	
95% confidence interval for p_1-p_2: Here Z*=1.96 $\left(\hat{p}_1-\hat{p}_2-Z^*\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\ \hat{p}_1-\hat{p}_2+Z^*\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$	
Hypothesis test for H_0: p_1-p_2=0.2 using $SE\approx\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ and $Z=\frac{point\ \ estimate - null\ \ value}{SE}=\frac{\hat{p}_1-\hat{p}_2-0.2}{SE}$ to get p-value	
Hypothesis test for H_0: p_1-p_2=0 or p_1=p_2 using $SE=\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1}+\frac{\hat{p}(1-\hat{p})}{n_2}}$ Here $\hat{p}$ is pooled proportion estimate. $Z=\frac{point\ \ estimate - null\ \ vaule}{SE}=\frac{\hat{p}_1-\hat{p}_2-0}{SE}$	

4. Testing for goodness of fit using chi-square

Given a sample of cases that can be classified into several groups, how can we test if the sample is representative of the general population?

Example 2 We consider data from a random sample of 275 jurors in a small county as in the following table. We would like to determine if these jurors are racially representative of the population.

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00

How should we do the test? The idea is that if the jury is representative of the population, then the proportion in the sample should roughly reflect the population of registered voters. Let's check the following table.

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected counts	198	19.25	33	24.75	275

If the more the differences between the observed data and expected data are, the stronger evidence we have for not fit.

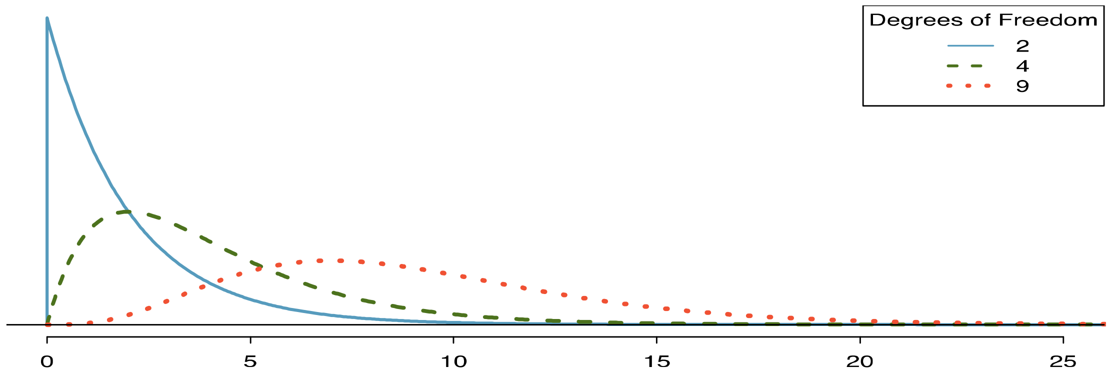
Chi-square test for one-way table

Suppose we are to evaluate whether there is convincing evidence that a set of observed counts O_1, O_2, ..., O_k in k categories are unusually different from what might be expected under a null hypothesis. Call the *expected counts* that are based on the null hypothesis E_1, E_2, ..., E_k. If each expected count is at least 5 and the null hypothesis is true, then the test statistic below follows a chi-square distribution with $k-1$ degrees of freedom:

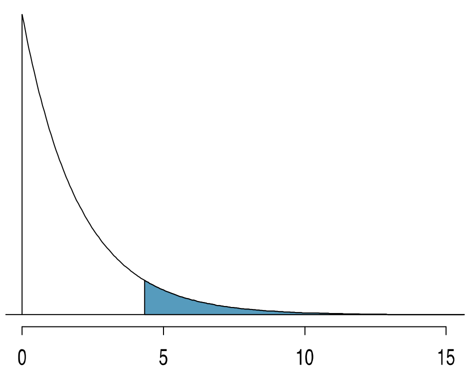
$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k}$$

The p-value for this test statistic is found by looking at the upper tail of this chi-square distribution. We consider the upper tail because larger values of X^2 would provide greater evidence against the null hypothesis.

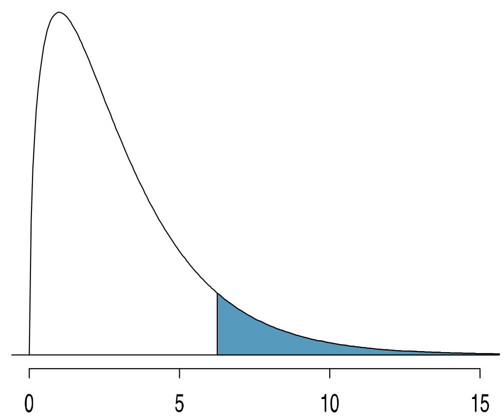
5. Chi-square distribution and p-value



Three chi-square distributions with different degrees of freedom



Chi-square distribution with 2 degree of freedom, area above 4.3 shaded



Chi-square distribution with 3 degree of freedom, area above 6.25 shaded

<u>Example 2</u> We consider data from a random sample of 275 jurors in a small county as in the following table. We would like to test at 5% significant level if these jurors are racially representative of the population.

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00

Answer: (1) H_0: The jury is representative of the population.

H_A: The jury is not representative of the population.

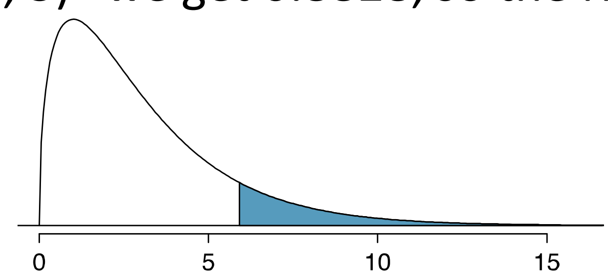
(2) Calculate X^2 :

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected counts	198	19.25	33	24.75	275

$$X^2 = \frac{(205-198)^2}{198} + \frac{(26-19.25)^2}{19.25} + \frac{(25-33)^2}{33} + \frac{(19-24.75)^2}{24.75} = 5.89$$

(3)Using R or table to find the p-value, which is the right tail area for Chi-square.

Using R: “ pchisq(5.89, 3)” we get 0.8828, so the right tail is 0.1172>0.05. We don’t reject H_0.



6.Test for independence in two-way table using chi-square

Test of two-way table is very similar to the test of one-way table. We still use chi-square test. There are two modifications here.

(1) Calculation of the expected count:

Computing expected counts in a two-way table

To identify the expected count for the i^{th} row and j^{th} column, compute

$$\text{Expected Count}_{\text{row } i, \text{ col } j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}$$

(2)

Computing degrees of freedom for a two-way table

When applying the chi-square test to a two-way table, we use

$$df = (R - 1) \times (C - 1)$$

where R is the number of rows in the table and C is the number of columns.

Example 3 The following table are the results of a Pew Research Poll. We would like to test if there are actually differences in the approval rating of Barack Obama, Democrats in Congress, and Republicans in Congress.

		Congress		
	Obama	Democrats	Republicans	Total
Approve	842	736	541	2119
Disapprove	616	646	842	2104
Total	1458	1382	1383	4223

Answer: (1) H_0: There is no difference in approval rating between three groups.

H_A: There is some difference in approval rating between three groups.

(2)

	Obama	Democrats	Republican	Total
Approval	842 (E=2119x1458/4223 =731.6)	736 (E=2119x1382/4223 =693.45)	541 (E=2119x1383/4223 =693.96)	2119
Disapprove	616 (E=2104x1458/4223 =726.4)	646 (E=2104x1382/4223 =688.55)	842 (E=2104x1383/4223 =689.04)	2104
total	1458	1382	1383	4223

For first cell, we calculate $(842-731.6)^2/731.6=16.7$. Similarly we calculate all the cells, and add all the results together. Then we have

$X^2=16.7+.....+34.0=106.4$

Degree of freedom=(2-1)(3-1)=2.

Using R: pchisq(106.4, 2)=1. So the right tail area is $0<0.05$. We reject H_0.

Homework on 03/22/16 (due 03/29/16)

(1) Try to finish the following table and do one-way chi-square test. I have 33.3% (or 1/3) black dice, 40% (or 2/5) white dice, and 26.7 % (or 4/15) color dice. Try to sample 60 dice in total and finish one way test.

black	**white**	color	**total**
33.3%	40%	26.7%	1.00

(2) Using the data you and all your classmates collected on 03/15/16 to do the two-way table chi-square test.

	Yours	Classmate 1	Classmate 2	Classmates 3	total
Black					
White					
total					

B.3 Chi-Square Probability Table

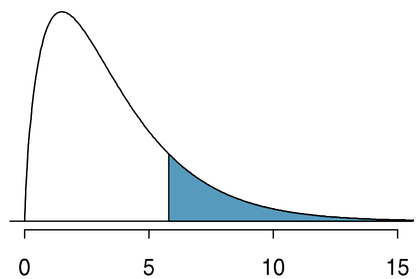


Figure B.2: Areas in the chi-square table always refer to the right tail.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26
	12	14.01	15.81	18.55	21.03	24.05	26.22	28.30	32.91
	13	15.12	16.98	19.81	22.36	25.47	27.69	29.82	34.53
	14	16.22	18.15	21.06	23.68	26.87	29.14	31.32	36.12
	15	17.32	19.31	22.31	25.00	28.26	30.58	32.80	37.70
	16	18.42	20.47	23.54	26.30	29.63	32.00	34.27	39.25
	17	19.51	21.61	24.77	27.59	31.00	33.41	35.72	40.79
	18	20.60	22.76	25.99	28.87	32.35	34.81	37.16	42.31
	19	21.69	23.90	27.20	30.14	33.69	36.19	38.58	43.82
	20	22.77	25.04	28.41	31.41	35.02	37.57	40.00	45.31
	25	28.17	30.68	34.38	37.65	41.57	44.31	46.93	52.62
	30	33.53	36.25	40.26	43.77	47.96	50.89	53.67	59.70
	40	44.16	47.27	51.81	55.76	60.44	63.69	66.77	73.40
	50	54.72	58.16	63.17	67.50	72.61	76.15	79.49	86.66