13.3 Hypothesis Test: Multinomial Population

Motivating example:

Objective:

we want to determine if a new product from company C has changed the market shares.

p_A : **market share for *company A***

p_B : **market share for *company B***

p_C : **market share for *company C***

We want to test

$$H_0: p_A = 0.3,\ p_B = 0.5,\ p_C = 0.2$$

vs.

$$H_a: \text{The population proprotions are not } p_A = 0.3,\ p_B = 0.5,\ p_C = 0.2$$

with $\alpha = 0.05$. We also have the following information:

total sample size: $n = 200$

observed number (company A): $f_1 = 48$

observed number (company B): $f_2 = 98$ $\quad \Rightarrow f_1 + f_2 + f_3 = 48 + 98 + 54 = 200 = n$

observed number (company C): $f_3 = 54$

In addition, as H_0 is true, the expected number of company's products are

expected number (company A): $e_1 = np_A = 200 \cdot 0.3 = 60$

expected number (company B): $e_2 = np_B = 200 \cdot 0.5 = 100$

expected number (company C): $e_3 = np_C = 200 \cdot 0.2 = 40$.

Intuitively, if the differences between f_i and e_i, $i = 1, 2, 3$, are small, that might imply H_0 is true and thus the observed number and the expected number (under H_0) are close. On the other hand, if the differences between f_i and e_i, $i = 1, 2, 3$, are large, that might imply H_0 might not be true, the expected number (under H_0) would be significantly different from the "true" expected number and thus result in the difference between the observed number and the expected number. The following statistic can be used to reflect the difference between the observed number and the expected number,

$$\chi^2 = \sum_{i=1}^{3} \frac{(f_i - e_i)^2}{e_i} = \frac{(f_1 - e_1)^2}{e_1} + \frac{(f_2 - e_2)^2}{e_2} + \frac{(f_3 - e_3)^2}{e_3}$$

$$= \frac{(48 - 60)^2}{60} + \frac{(98 - 100)^2}{100} + \frac{(54 - 40)^2}{40}$$

$$= 7.34$$

General Case:

Suppose there are K populations. We want to test

$$H_0 : p_1 = a_1,\ p_2 = a_2, \ldots,\ p_k = a_k$$

vs.

$$H_a : \text{The population proprotions are not } p_1 = a_1, p_2 = a_2, \ldots, p_k = a_k$$

where

$$\sum_{i=1}^{k} a_i = a_1 + a_2 + \cdots + a_k = 1$$

We also have the following information:

total sample size: n_T

observed numbers: $f_i, i = 1, 2, \ldots, k$.

In addition, as H_0 is true, the expected numbers are

expected numbers: $e_i = n_T a_i, i = 1, 2, \ldots, k$

The test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} = \frac{(f_1 - e_1)^2}{e_1} + \frac{(f_2 - e_2)^2}{e_2} + \cdots + \frac{(f_k - e_k)^2}{e_k}$$

Next question: how large χ^2 must be to reject H_0?

Chi-Square Distribution:

χ_n^2 : the random variable distributed as chi-square distribution with degrees of freedom n .

Example:

$$P(\chi_3^2 > x) = 0.05 \Rightarrow x = 7.814$$

$$P(\chi_{15}^2 > x) = 0.1 \Rightarrow x = 22.3072$$

$$P(\chi_9^2 > x) = 0.9 \Rightarrow x = 4.168$$

Chi-Square Test:

Let

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} = \frac{(f_1 - e_1)^2}{e_1} + \frac{(f_2 - e_2)^2}{e_2} + \cdots + \frac{(f_k - e_k)^2}{e_k}$$

The chi-square test with level of significance α for

$$H_0 : p_1 = a_1, \ p_2 = a_2, \ldots, \ p_k = a_k$$

vs.

$$H_a : \text{The population proprotions are not } p_1 = a_1, \ p_2 = a_2, \ldots, \ p_k = a_k$$

is to

$$\text{reject } H_0 : \quad \chi^2 > \chi_{k-1,\alpha}^2$$

$$\text{not reject } H_0 : \quad \chi^2 \le \chi_{k-1,\alpha}^2 \ ,$$

where $\chi^2_{k-1,\alpha}$ **can be obtained by**

$$P\left(\chi^2_{k-1} > \chi^2_{k-1,\alpha}\right) = \alpha \text{ .}$$

In addition,

$$\text{p-value} = P\left(\chi^2_{k-1} > \chi^2\right).$$

Note:

As H_0 is true, the random variable with sample value χ^2 is χ^2_{k-1}.

χ^2 : the ***sample statistic***

χ^2_{k-1} : the ***random variable*** distributed as chi-square distribution with

degree of freedom ***k-1*** and sample value χ^2.

$\chi^2_{k-1,\alpha}$ **:** the ***critical value*** satisfying $P\left(\chi^2_{k-1} > \chi^2_{k-1,\alpha}\right) = \alpha$.

Motivating Example (continue):

Since $k = 3$,

$$\chi^2 = 7.34 > 5.99 = \chi^2_{2,0.05} = \chi^2_{k-1,\alpha} \text{ ,}$$

thus we reject H_0 .

Example:

The following data are the frequencies of products of throwing a dice 120 times:

Point	1	2	3	4	5	6
Frequency	*13*	*24*	*18*	*22*	*19*	*24*

Please test if the dice is fair (i.e., $H_0 : p_1 = p_2 = \cdots = p_6 = \dfrac{1}{6}$) with $\alpha = 0.05$.

[solution:]

$$e_i = 120 \cdot \frac{1}{6} = 20,\ i = 1,2,\ldots,6.$$

Then,

$$\chi^2 = \sum_{i=1}^{6} \frac{(f_i - e_i)^2}{e_i} = \frac{(13-20)^2}{20} + \frac{(24-20)^2}{20} + \frac{(18-20)^2}{20}$$
$$+ \frac{(22-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20} = 4.5$$

Since

$$\chi^2 = 4.5 < 11.0705 = \chi^2_{5,0.05} = \chi^2_{k-1,\alpha},$$

we do ***not*** reject H_0.

Example:

The following are the number of wrong answers for the number of the students.

Number of wrong answers	**0**	**1**	**2**	**3**
Number of the students	**21**	**31**	**12**	**0**

Suppose X is the random variable representing the number of wrong answers. Please test X is distributed as ***Binomial(3,0.25)*** with $\alpha = 0.05$.

(Note: the distribution function for ***Binomial(3,0.25)*** is

$$f(x) = \binom{3}{x}(0.25)^x(0.75)^{3-x},\ x = 0,1,2,3..$$

[solutions:]

As H_0 is true, the distribution for the number of wrong answers is

$$p_1 = P(X=0) = \binom{3}{0}0.25^0 0.75^3 = \frac{27}{64},$$

$$p_2 = P(X=1) = \binom{3}{1}0.25^1 0.75^2 = \frac{27}{64},$$

$$p_3 = P(X=2) = \binom{3}{2}0.25^2 0.75^1 = \frac{9}{64},$$

$$p_4 = P(X = 3) = \binom{3}{3} 0.25^3 0.75^{01} = \frac{1}{64},$$

Since the sample size $n = 21 + 31 + 12 + 0 = 64$, the expected numbers under H_0 are

$$e_1 = np_1 = 64 \cdot \frac{27}{64} = 27, e_2 = np_2 = 64 \cdot \frac{27}{64} = 27,$$
$$e_3 = np_3 = 64 \cdot \frac{9}{64} = 9, e_4 = np_4 = 64 \cdot \frac{1}{64} = 1,$$

Therefore,

$$\chi^2 = \sum_{k=1}^{4} \frac{(f_k - e_k)^2}{e_k}$$
$$= \frac{(21-27)^2}{27} + \frac{(31-27)^2}{27} + \frac{(12-9)^2}{9} + \frac{(-1)^2}{1}$$
$$= 3.92$$

Since

$$\chi^2 = 3.92 < 7.81 = \chi^2_{3,0.05} = \chi^2_{k-1,\alpha},$$

we do ***not*** reject H_0 .

Online Exercise:

Exercise 13.3.1
Exercise 13.3.2