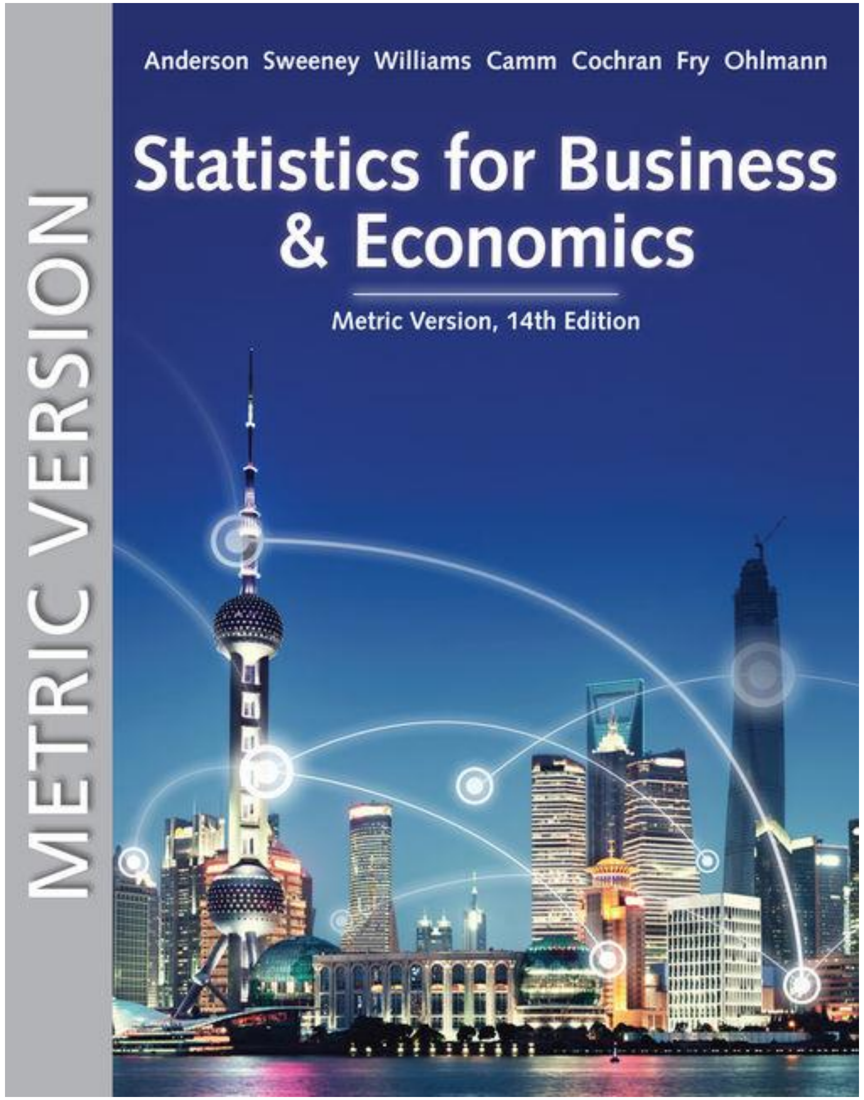


Statistics for Business and Economics (14e) Metric Version

Chapter 12 （卡方檢定）

Chapter 12 - Comparing Multiple Proportions, Test of Independence and Goodness of Fit

12.1 - Testing the Equality of Population Proportions for Three or More Populations

12.2 - Test of Independence

12.3 - Goodness of Fit Test

卡方檢定(Chi-Square Test)

- 卡方檢定主要在處理類別資料的檢定，依序為：

→ Goodness of Fit Test (適合度檢定)

通常用來檢查資料是否為某一特定分配，例如：樣本是否與母體類似。

→ Tests of Independence (獨立性檢定)

→ Tests of Homogeneity(齊一性檢定)

Tests of Goodness of Fit, Independence, and Multiple Proportions

- In this chapter we introduce three additional hypothesis-testing procedures.
- The test statistic and the distribution used are based on the chi-square (χ^2) distribution.
- In all cases, the data are categorical.

Testing the Equality of Population Proportions for Three or More Populations (1 of 12)

Hypotheses:

H_0: $p_1 = p_2 = \cdots = p_k$

H_a: Not all population proportions are equal.

Where p_1 = population proportion for population 1

p_2 = population proportion for population 2

$\vdots$

p_k = population proportion for population k

Testing the Equality of Population Proportions for Three or More Populations (2 of 12)

- If H_0 cannot be rejected, we cannot detect a difference among the k population proportions.
- If H_0 can be rejected, we can conclude that not all k population proportions are equal.
- Further analyses can be done to conclude which population proportions are significantly different from others.

Testing the Equality of Population Proportions for Three or More Populations (3 of 12)

Example: Finger Lakes Homes

Finger Lakes Homes manufactures three models of prefabricated homes: a two-story colonial, a log cabin, and an A-frame. To help in product-line planning, management would like to compare the customer satisfaction with the three home styles.

p_1 = proportion likely to repurchase a Colonial for the population of Colonial owners

p_2 = proportion likely to repurchase a Log Cabin for the population of Log Cabin owners

p_3 = proportion likely to repurchase an A-Frame for the population of A-Frame owners

Testing the Equality of Population Proportions for Three or More Populations (4 of 12)

- We begin by taking a sample of owners from each of the three populations.
- Each sample contains categorical data indicating whether the respondents are likely or not likely to repurchase the home.

Here are the observed frequencies (sample results)

Likely to repurchase	Home owner Colonial	Home Owner Log	Home Owner A-frame	Total
Yes	97	83	80	260
No	38	18	44	100
Total	135	101	124	360

Testing the Equality of Population Proportions for Three or More Populations (5 of 12)

Next, we determine the expected frequencies under the assumption H_0 is correct.

Expected frequencies under the assumption H_0 is true are calculated using this formula:

$$e_{ij} = \frac{(\text{Row } i \text{ Total})(\text{Column } j \text{ Total})}{\text{Total Sample Size}}$$

- If a significant difference exists between the observed and expected frequencies, H_0 can be rejected.

Testing the Equality of Population Proportions for Three or More Populations (6 of 12)

The expected frequencies are:

Likely to repurchase	Home Owner Colonial	Home Owner Log	Home Owner A-frame	Total
Yes	97.50	72.94	89.56	260
No	37.50	28.06	34.44	100
Total	135	101	124	360

Testing the Equality of Population Proportions for Three or More Populations (7 of 12)

Next, compute the value of the chi-square test statistic.

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

where: f_{ij} = observed frequency for the cell in row i and column j

e_{ij} = expected frequency for the cell in row i and column j under the assumption H_0 is true

Note: The test statistic has a chi-square distribution with $k - 1$ degrees of freedom, provided the expected frequency is 5 or more for each cell.

Testing the Equality of Population Proportions for Three or More Populations (8 of 12)

Computation of the Chi-Square Test Statistic

Likely to Repurchase	Home Owner	Obs. Freq. f_{ij}	Exp. Freq. e_{ij}	Diff. $(f_{ij} - e_{ij})$	Sqd. Diff. $(f_{ij} - e_{ij})^2$	Sqd. Diff. / Exp. Freq. $(f_{ij} - e_{ij})^2/e_{ij}$
Yes	Colonial	97	97.50	-0.50	0.2500	0.0026
Yes	Log Cab.	83	72.94	10.06	101.1142	1.3862
Yes	A-Frame	80	89.56	-9.56	91.3086	1.0196
No	Colonial	38	37.50	0.50	0.2500	0.0067
No	Log Cab.	18	28.06	-10.06	101.1142	3.6041
No	A-Frame	44	34.44	9.56	91.3086	2.6509
	Total	360	360		χ^2 =	8.6700

Testing the Equality of Population Proportions for Three or More Populations (9 of 12)

Rejection Rule:

p-value approach: Reject H_0 if the *p*-value $\leq \alpha$

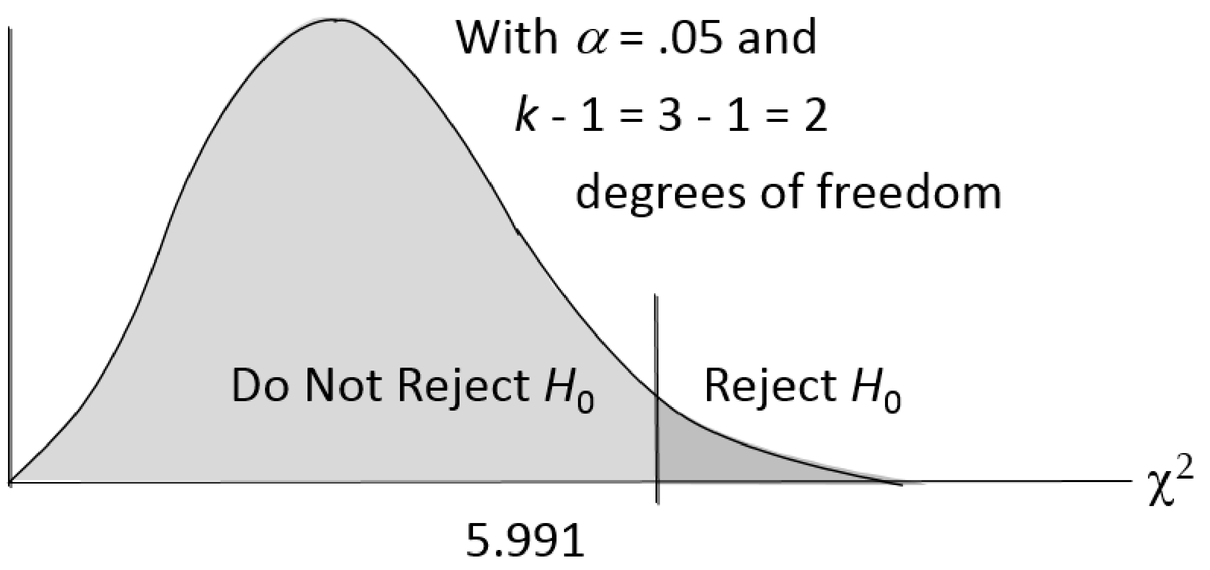
Critical value approach: Reject H_0 if $\chi^2 \geq \chi^2_\alpha$

Where α is the significance level and there are $k - 1$ degrees of freedom.

Testing the Equality of Population Proportions for Three or More Populations (10 of 12)

Rejection Rule (using $\alpha = 0.05$)

Reject H_0 if p-value $\leq .05$ or $\chi^2 > 5.991$



Testing the Equality of Population Proportions for Three or More Populations (11 of 12)

Conclusion using the *p*-value approach:

Area in Upper Tail	.10	.05	.025	.01	.005
χ^2 Value (df = 2)	4.605	5.991	7.378	9.210	10.597

Because $\chi^2 = 8.670$ is between 9.210 and 7.378, the area in the upper tail of the distribution is between 0.01 and 0.025.

Therefore, the *p*-value $\leq \alpha$. We reject the null hypothesis. The actual *p*-value is 0.0131.

Testing the Equality of Population Proportions for Three or More Populations (12 of 12)

- We have concluded that the population proportions for the three populations of home owners are not equal.

- To identify where the differences between population proportions exist, we will rely on a multiple comparisons procedure.

Multiple Comparisons Procedure (1 of 4)

We begin by computing the three sample proportions.

Colonial: $\bar{p}_1 = \frac{97}{135} = 0.7185$

Log Cabin: $\bar{p}_2 = \frac{83}{101} = 0.8218$

A-Frame: $\bar{p}_3 = \frac{80}{124} = 0.6452$

We will use a multiple comparison procedure known as the Marascuilo procedure.

Multiple Comparisons Procedure (2 of 4)

Marascuilo Procedure

We compute the absolute value of the pairwise difference between sample proportions.

Colonial and Log Cabin: $|\bar{p}_1 - \bar{p}_2| = |0.7185 - 0.8218| = 0.1033$

Colonial and A-Frame: $|\bar{p}_1 - \bar{p}_3| = |0.7185 - 0.6452| = 0.0733$

Log Cabin and A-Frame: $|\bar{p}_2 - \bar{p}_3| = |0.8218 - 0.6452| = 0.1766$

Multiple Comparisons Procedure (3 of 4)

Critical Values for the Marascuilo Pairwise Comparison

For each pairwise comparison, compute a critical value as follows:

$$CV_{ij} = \sqrt{\chi^2_{\alpha,k-1}}\sqrt{\frac{\bar{p}_i(1-\bar{p}_i)}{n_i} + \frac{\bar{p}_j(1-\bar{p}_j)}{n_j}}$$

For $\alpha = .05$ and $k = 3$: $\chi^2 = 5.991$

Multiple Comparisons Procedure (4 of 4)

Pairwise Comparison Tests

Pairwise Comparison	$\lvert \bar{p}_i - \bar{p}_j \rvert$	CV_{ij}	Significant if $\lvert \bar{p}_i - \bar{p}_j \rvert > CV_{ij}$
Colonial vs. Log Cabin	.1033	.1329	Not Significant
Colonial vs. A-Frame	.0733	.1415	Not Significant
Log Cabin vs. A-Frame	.1766	.1405	Significant

Test of Independence (1 of 7)

1. Set up the null and alternative hypotheses.

H_0: The column variable is independent of the row variable

H_a: The column variable is not independent of the row variable

2. Select a random sample and record the observed frequency, f_{ij} , for each cell of the contingency table.

3. Compute the expected frequency, e_{ij} , for each cell.

$$e_{ij} = \frac{(\text{Row } i \text{ Total})(\text{Column } j \text{ Total})}{\text{Sample Size}}$$

Test of Independence (2 of 7)

4. Compute the test statistic.

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

5. Determine the rejection rule.

Reject H_0 if $\chi^2 \geq \chi^2_\alpha$

Where α is the significance level with n rows and m columns, there are $(n-1)(m-1)$ degrees of freedom.

Test of Independence (3 of 7)

Example: Finger Lakes Homes (B)

Each home sold by Finger Lakes Homes can be classified according to price and to style. Finger Lakes' manager would like to determine if the price of the home and the style of the home are independent variables.

The number of homes sold for each model and price for the past two years is shown below. For convenience, the price of the home is listed as either *less than $200,000* or *greater than or equal to $200,000*.

Price	Colonial	Log	Split-Level	A-Frame
< $200,000	18	6	19	12
≥ $200,000	12	14	16	3

■ 卡方檢定通常以列聯表(Contingency Table)方式呈現，獨立性檢定可用於行列的類別不只兩種！

Marital Status by Education | n = 300

	Middle school or lower	High school	Bachelor's	Master's	PhD or higher	Total
Never married	18	36	21	9	6	90
Married	12	36	45	36	21	150
Divorced	6	9	9	3	3	30
Widowed	3	9	9	6	3	30
Total	39	90	84	54	33	300

https://spss-tutorials.com/img/chi-square-test-contingency-table.png

Test of Independence (4 of 7)

Hypotheses

H_0: Price of the home is independent of the style of the home that is purchased

H_a: Price of the home is not independent of the style of the home that is purchased

Expected Frequencies

Price	Colonial	Log	Split-Level	A-frame	Total
< \$200 K	18	6	19	12	55
≥ \$200 K	12	14	16	3	45
Total	30	20	35	15	100

Test of Independence (5 of 7)

- Rejection Rule

With $\alpha = 0.05$ and (2 – 1)(4 – 1) = 3 degrees of freedom, $\chi^2_\alpha = 7.815$.

Reject H_0 if the *p*-value≤ 0.05 or if $\chi^2 \geq 7.815$.

- Test Statistic

$$\chi^2 = \frac{(18 - 16.5)^2}{16.5} + \frac{(6 - 11)^2}{11} + \cdots + \frac{(3 - 6.75)^2}{6.75}$$

$$= 0.1364 + 2.2727 + \cdots + 2.0833$$

$$= 9.149$$

Test of Independence (6 of 7)

Conclusion using the *p*-value approach

Area in Upper Tail	.10	.05	.025	.01	.005
χ^2 Value (df = 3)	6.251	7.815	9.348	11.345	12.838

Because $\chi^2 = 9.145$ is between 7.815 and 9.348, the area in the upper tail of the distribution is between 0.025 and 0.05.

Therefore, the *p*-value $\leq \alpha$. We reject the null hypothesis. The actual *p*-value is 0.0274.

Test of Independence (7 of 7)

Conclusion using the critical value approach

$$\chi^2 = 9.145 \geq 7.815$$

We reject at the 0.05 level of significance the assumption that the price of the home is independent of the style of home that is purchased.

Goodness of Fit Test: Multinomial Probability Distribution (1 of 3)

1. State the null and alternative hypotheses.

   H_0: The population follows a multinomial distribution with specified probabilities for each of the k categories

   H_a: The population does not follow a multinomial distribution with specified probabilities for each of the k categories

2. Select a random sample and record the observed frequency, f_i, for each of the k categories.

3. Assuming H_0 is true, compute the expected frequency, e_i, in each category by multiplying the category probability by the sample size.

Goodness of Fit Test: Multinomial Probability Distribution (2 of 3)

4. Compute the value of the test statistic.

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

where:

f_i = observed frequency for category i

e_i = expected frequency for category i

k = number of categories

Note: The test statistic has a chi-square distribution with $k - 1$ degrees of freedom provided that the expected frequencies are 5 or more for all categories.

Goodness of Fit Test: Multinomial Probability Distribution (3 of 3)

Rejection Rule:

p-value approach: Reject H_0 if the *p*-value $\leq \alpha$

Critical value approach: Reject H_0 if $\chi^2 \geq \chi^2_\alpha$

Where α is the significance level and there are k – 1 degrees of freedom.

Multinomial Distribution Goodness of Fit Test (1 of 6)

Example: Finger Lakes Homes (A)

Finger Lakes Homes manufactures four models of prefabricated homes: a two-story colonial, a log cabin, a split-level, and an A-frame. To help in production planning, management would like to determine if previous customer purchases indicate that there is a preference in the style selected.

The number of homes sold of each model for 100 sales over the past two years is shown below.

Model	Colonial	Log	Split-Level	A-Frame
# Sold	30	20	35	15

Multinomial Distribution Goodness of Fit Test (2 of 6)

Hypotheses

H_0: $p_C = p_L = p_S = p_A = .25$

H_a: The population proportions are not $p_C = .25$, $p_L = .25$, $p_S = .25$, and $p_A = .25$

where:

p_C = population proportion that purchase a colonial

p_L = population proportion that purchase a log cabin

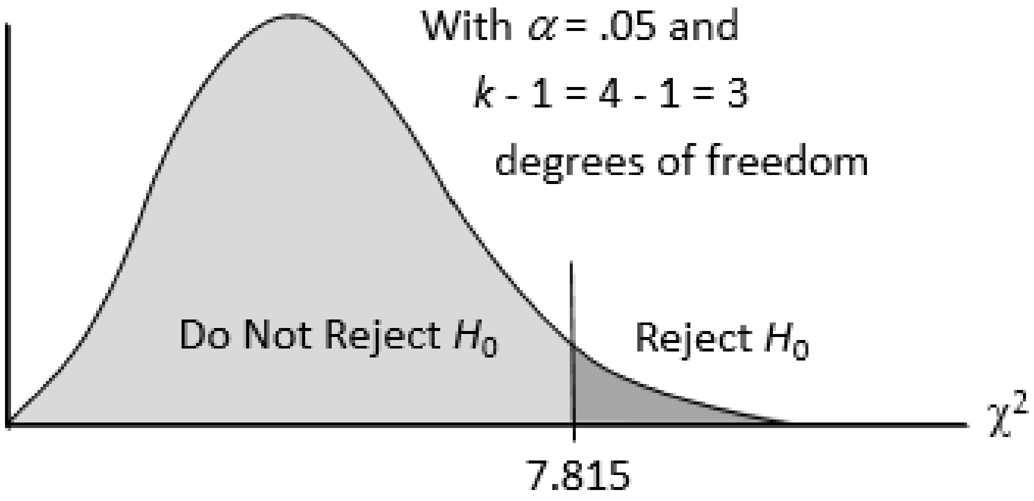
p_S = population proportion that purchase a split-level

p_A = population proportion that purchase an A-frame

Multinomial Distribution Goodness of Fit Test (3 of 6)

Rejection Rule:

Reject H_0 if the p-value ≤ 0.05 or $\chi^2 \geq 7.815$.



Multinomial Distribution Goodness of Fit Test (4 of 6)

Expected Frequencies:

$$e_1 = 0.25(100) = 25$$
$$e_2 = 0.25(100) = 25$$
$$e_3 = 0.25(100) = 25$$
$$e_4 = 0.25(100) = 25$$

Test Statistic:

$$\chi^2 = \frac{(30-25)^2}{25} + \frac{(20-25)^2}{25} + \frac{(35-25)^2}{25} + \frac{(15-25)^2}{25} = 1 + 1 + 4 + 4 = \mathbf{10}$$

Multinomial Distribution Goodness of Fit Test (5 of 6)

Conclusion using the *p*-value approach:

Area in Upper Tail	.10	.05	.025	.01	.005
χ^2 Value (df = 3)	6.251	7.815	9.348	11.345	12.838

Because $\chi^2 = 10$ is between 9.348 and 11.345, the area in the upper tail of the distribution is between 0.01 and 0.025.

Therefore, the *p*-value $\leq \alpha$. We reject the null hypothesis.

Multinomial Distribution Goodness of Fit Test (6 of 6)

Conclusion using the critical value approach

$$\chi^2 = 10 \geq 7.815$$

We reject, at the 0.05 level of significance, the assumption that there is no home style preference.