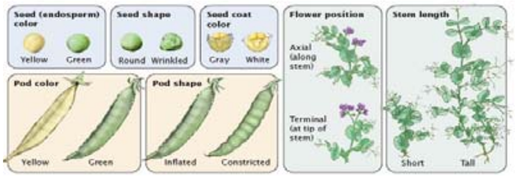
Chapter 11

Chi-Square Tests and

Other Nonparametric Statistics

1

Statistics and Heredity



2

Introduction

- A specific *frequency distributions*
  - If a sample of buyers is given a choice of automobile colors, will each color be selected with the same frequency?
- *Independence* of two variables
  - Whether gender is related to voting preference?
- *Homogeneity of proportions*
  - Is the proportion of high school seniors who attend college immediately after graduating the same for the northern, southern, eastern, and western parts of the United States?

3

Chapter 11 Overview

Introduction

- 11-1 Test for Goodness of Fit
- 11-2 Tests Using Contingency Tables

4

Chapter 11 Objectives

1. Test a distribution for goodness of fit, using chi-square.
2. Test two variables for independence, using chi-square.
3. Test proportions for homogeneity, using chi-square.

5

Section 11-1

Test for Goodness of Fit

6

11.1 Test for Goodness of Fit

- The chi-square statistic can be used to see whether a frequency distribution fits a specific pattern. This is referred to as the chi-square goodness-of-fit test.

Example 1

適合度檢定 —多項式母體比例的檢定

擲一個骰子 100 次，其出現點數次數之資料如下：

點 數	1	2	3	4	5	6
次 數	20	24	10	15	14	17

試檢定上述資料是否服從於分立均等分配？($\alpha = 0.05$)

Example 2

適合度檢定 —多項式母體比例的檢定

依孟德爾研究豌豆的外型，得出圓而黃、圓而綠、皺而黃、皺而綠之遺傳比例為 9：3：3：1，茲觀察 200 棵豌豆外型，其資料如下：

豌 豆	圓而黃	圓而綠	皺而黃	皺而綠
數 目	110	35	40	15

試檢定此豌豆實驗是否符合遺傳理論？($\alpha = 0.05$)

Example 3: Fruit Soda Flavors

A market analyst wished to see whether consumers have any preference among five flavors of a new fruit soda. A sample of 100 people provided the following data. Is there enough evidence to reject the claim that there is no preference in the selection of fruit soda flavors, using the data shown previously? Let $\alpha = 0.05$.

Cherry	Strawberry	Orange	Lime	Grape
32	28	16	14	10

基本概念

- 類別資料概述：

  類別資料 (categorical data)，是指根據觀察值 (即個體) 的屬性，按照不同類別加以分類 (分組)，而各分類 (分組) 的個數乃依點計方式計數而得，故又稱為點計資料 (count data)。

- 類別資料的種類：
  - 單分類列聯表：

    單分類列聯表 (one-way contingency table)，係指統計資料只依單一標準加以分類 。
  - 雙分類列聯表：

    雙分類列聯表 (two-way contingency table)，係指統計資料依兩標準加以分類 。

基本概念

- 常見的卡方檢定類型：

  (1) 多項式母體比例的檢定。
  (2) 母體分配型態的檢定。
  適合度檢定。(goodness-of-fit test)

  (3) 獨立性檢定。
  (4) 齊一性檢定。
  列聯表檢定。(contingency table)

卡方檢定

- 適合度檢定
  - 檢定母體分配是否為某一特定分配或理論分配的統計方法。
  - 如學生成績分配是否為常態分配，或孟德爾的碗豆實驗發現其外型是否符合特定比例。
- 獨立性檢定
  - 檢定兩個屬性間是否獨立的統計方法。
  - 研究顧客之教育程度與其對某門市之服務滿意度是否有關。
- 齊一性檢定
  - 檢定兩個或兩個以上母體的某一特性的分配(各類別的比例)是否相同或相近。
  - 檢定某項新產品在台北、新竹、台中等地之之名度是否相同。

Test for Goodness of Fit

Formula for the test for goodness of fit:

$$\chi^2_{STAT} = \sum \frac{(O-E)^2}{E}$$

where
d.f. = number of categories minus 1
O = observed frequency
E = expected frequency

Assumptions for Goodness of Fit

1. The data are obtained from a random sample.
2. The expected frequency for each category must be 5 or more.

Basic Concept

- **Chi-square test**

Category	1	2		k	Total
Observation, o_i	o_1	o_2	…	o_k	n
Expectation, e_i (assume H_0 is true)	e_1	e_2	…	e_k	n

$$\chi^2 = \frac{(o_1-e_1)^2}{e_1} + \frac{(o_2-e_2)^2}{e_2} + \cdots + \frac{(o_k-e_k)^2}{e_k} = \sum_{i=1}^{k} \frac{(o_i-e_i)^2}{e_i}$$

If $\chi^2 > \chi^2_{(\alpha,f)}$,

The difference between observation and expectation is significant.

→Reject H_0

→Chi-square is always right-tailed test.

Example 1

擲一個骰子 100 次，其出現點數次數之資料如下：

點 數	1	2	3	4	5	6
次 數	20	24	10	15	14	17

試檢定上述資料是否服從於分立均等分配？($\alpha = 0.05$)

Example 2

依孟德爾研究豌豆的外型，得出圓而黃、圓而綠、皺而黃、皺而綠之遺傳比例為 9：3：3：1，茲觀察 200 棵豌豆外型，其資料如下：

豌　豆	圓而黃	圓而綠	皺而黃	皺而綠
數　目	110	35	40	15

試檢定此豌豆實驗是否符合遺傳理論？($\alpha = 0.05$)

Example 3: Fruit Soda Flavors

A market analyst wished to see whether consumers have any preference among five flavors of a new fruit soda. A sample of 100 people provided the following data. Is there enough evidence to reject the claim that there is no preference in the selection of fruit soda flavors, using the data shown previously? Let $\alpha = 0.05$.

Cherry	Strawberry	Orange	Lime	Grape
32	28	16	14	10

Step 1: State the hypotheses and identify the claim.

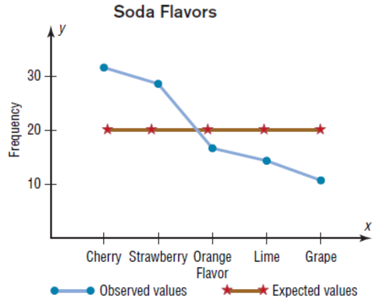
Example 3: Fruit Soda Flavors

	Cherry	Strawberry	Orange	Lime	Grape
Observed	32	28	16	14	10
Expected					

Step 2: Find the critical value.

Step 3: Compute the test value.

Graphs of the Observed and Expected Values for Soda Flavors



Results of the Goodness-of-Fit Test

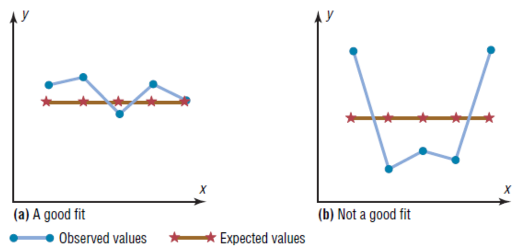


Figure 11–2 Graphs of the Observed and Expected Values for Soda Flavors

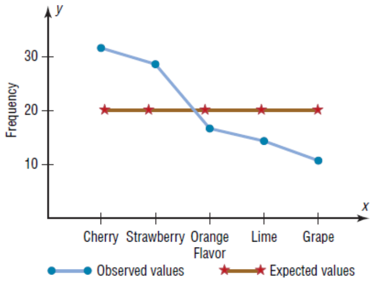
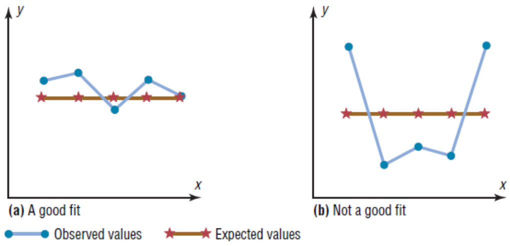


Figure 11–3 Results of the Goodness-of-Fit Test



Procedure

Chi-Square Goodness-of-Fit Test

Purpose To perform a hypothesis test for the distribution of a variable

Assumptions†

1. All expected frequencies are 1 or greater
2. At most 20% of the expected frequencies are less than 5
3. Simple random sample

STEP 1 The null and alternative hypotheses are

H_0: The variable has the specified distribution.
H_a: The variable does not have the specified distribution.

STEP 2 Calculate the expected frequency for each possible value of the variable by using the formula $E = np$, where n is the sample size and p is the relative frequency (or probability) given for the value in the null hypothesis.

STEP 3 Determine whether the expected frequencies satisfy Assumptions 1 and 2. If they do not, this procedure should not be used.

STEP 4 Decide on the significance level, α.

STEP 5 Compute the value of the test statistic

$$\chi^2 = \Sigma(O - E)^2/E,$$

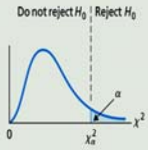
where O and E represent observed and expected frequencies, respectively. Denote the value of the test statistic χ_0^2.

Procedure (cont.)

CRITICAL-VALUE APPROACH or **P-VALUE APPROACH**

CRITICAL-VALUE APPROACH

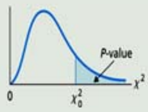
STEP 6 The critical value is χ_α^2 with df $= k - 1$, where k is the number of possible values for the variable. Use Table VII to find the critical value.



STEP 7 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

***P*-VALUE APPROACH**

STEP 6 The χ^2-statistic has df $= k - 1$, where k is the number of possible values for the variable. Use Table VII to estimate the *P*-value, or obtain it exactly by using technology.



STEP 7 If $P \le \alpha$, reject H_0; otherwise, do not reject H_0.

STEP 8 Interpret the results of the hypothesis test.

Example 11-2: Education Level of Adults

- The Census Bureau of the U.S. government found that 13% of adults did not finish high school, 30% graduated from high school only, 29% had some college education but did not obtain a bachelor’s degree, and 28% were college graduates. To see if these proportions were consistent with those people who lived in the Lincoln County area, a local researcher selected a random sample of 300 adults and found that 43 did not finish high school, 76 were high school graduates only, 96 had some college education, and 85 were college graduates. At α = 0.10, test the claim that the proportions are the same for the adults in Lincoln County as those stated by the Census Bureau.

Example 11-2: Education Level of Adults

Example: Retirees

The Russel Reynold Association surveyed retired senior executives who had returned to work. They found that after returning to work, **38%** were employed by another organization, **32%** were self-employed, **23%** were either freelancing or consulting, and **7%** had formed their own companies. To see if these percentages are consistent with those of Allegheny County residents, a local researcher surveyed **300** retired executives who had returned to work and found that **122** were working for another company, **85** were self-employed, **76** were either freelancing or consulting, and **17** had formed their own companies. At ***α* = 0.10**, test the claim that the percentages are the same for those people in Allegheny County.

Example: Retirees

	New Company	Self-Employed	Free-lancing	Owns Company
Observed	122	85	76	17

Step 1: State the hypotheses and identify the claim.

H_0: The retired executives who returned to work are distributed as follows: 38% are employed by another organization, 32% are self-employed, 23% are either freelancing or consulting, and 7% have formed their own companies (claim).

H_1: The distribution is not the same as stated in the null hypothesis.

Example: Retirees

	New Company	Self-Employed	Free-lancing	Owns Company
Observed	122	85	76	17
Expected				

Example 11-3: Firearm Deaths

A researcher read that firearm-related deaths for people aged 1 to 18 were distributed as follows: 74% were accidental, 16% were homicides, and 10% were suicides. In her district, there were 68 accidental deaths, 27 homicides, and 5 suicides during the past year. At $\alpha = 0.10$, test the claim that the percentages are equal.

Accidental	Homicides	Suicides
68	27	5

Example 11-3: Firearm Deaths

	Accidental	Homicides	Suicides
Observed	68	27	5

Step 1: State the hypotheses and identify the claim.

H_0: Deaths due to firearms for people aged 1 through 18 are distributed as follows: 74% accidental, 16% homicides, and 10% suicides (claim).

H_1: The distribution is not the same as stated in the null hypothesis.

Example 11-3: Firearm Deaths

Exercise: Blood Types

- Human blood is grouped into four types: A, B, AB, and O. The percents of Americans with each type are as follows: O, 43%; A, 40%; B, 12%; and AB, 5%. At a recent blood drive at a large university, the donors were classified as shown below. At the 0.05 level of significance, is there sufficient evidence to conclude that the proportions differ from those stated above?

Exercise: Blood Types (*Solution*)

Section 11-2

Tests Using Contingency Tables

Contingency Tables

列聯表——性別與偏好車型表

性別 \ 車型	迷你車	小型車	中型車
男	15	36	49
女	34	50	16

類別	很滿意	滿意	普通	不滿意	很不滿意	列合計
台北市	50	95	125	140	90	500
台中市	48	62	131	127	132	500
高雄市	60	75	130	110	125	500
行合計	158	232	386	377	347	1500

11.2 Tests Using Contingency Tables

- When data can be tabulated in table form in terms of frequencies, several types of hypotheses can be tested by using the chi-square test.
- The **test of independence of variables** is used to determine whether two variables are independent of or related to each other when a single sample is selected.
- The **test of homogeneity of proportions** is used to determine whether the proportions for a variable are equal when several samples are selected from different populations.

Test for Independence

- The chi-square goodness-of-fit test can be used to test the independence of two variables.
- The hypotheses are:
  - H_0: There is no relationship between two variables.
  - H_1: There is a relationship between two variables.
- If the null hypothesis is rejected, there is some relationship between the variables.

Test for Independence

- In order to test the null hypothesis, one must compute the expected frequencies, assuming the null hypothesis is true.
- When data are arranged in table form for the independence test, the table is called a **contingency table**.

Illustration –1/5

p.625

- Suppose a new postoperative procedure is administered to a number of patients in a large hospital. One can ask the question, Do the doctors feel differently about this procedure from the nurses, or do they feel basically the same way?
- As the survey indicates, 100 nurses prefer the new procedure, 80 prefer the old procedure, and 20 have no preference; 50 doctors prefer the new procedure, 120 like the old procedure, and 30 have no preference.

Group	Prefer new procedure	Prefer old procedure	No preference
Nurses	100	80	20
Doctors	50	120	30

Illustration –2/5

- Since the main question is whether there is a difference in opinion, the null hypothesis is stated as follows:

H_0: The opinion about the procedure is *independent* of the profession.

The alternative hypothesis is stated as follows:

H_1: The opinion about the procedure is *dependent* on the profession.

The degrees of freedom for any contingency table are $(R-1)\times(C-1)$; that is, d.f. $(R-1)\times(C-1)$. *In this case,* $(2-1)\times(3-1)=2$.

Group	Prefer new procedure	Prefer old procedure	No preference
Nurses	100	80	20
Doctors	50	120	30

Illustration –3/5

1. Find the sum of each row and each column, and find the grand total, as shown.

Group	Prefer new procedure	Prefer old procedure	No preference	Total
Nurses	100	80	20	Row 1 sum 200
Doctors	+50	+120	+30	Row 2 sum 200
Total	150	200	50	400
	Column 1 sum	Column 2 sum	Column 3 sum	Grand total

2. For each cell, multiply the corresponding row sum by the column sum and divide by the grand total, to get the expected value:

$$\text{Expected value} = \frac{\text{row sum} \times \text{column sum}}{\text{grand total}}$$

For example, for $C_{1,2}$, the expected value, denoted by $E_{1,2}$, is (refer to the previous tables)

$$E_{1,2} = \frac{(200)(200)}{400} = 100$$

Illustration –4/5

The expected values can now be placed in the corresponding cells along with the observed values, as shown.

Group	Prefer new procedure	Prefer old procedure	No preference	Total
Nurses	100 (75)	80 (100)	20 (25)	200
Doctors	50 (75)	120 (100)	30 (25)	200
Total	150	200	50	400

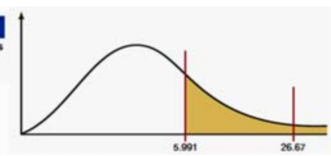
The formula for the test value for the independence test is the same as the one used for the goodness-of-fit test. It is

$$\chi^2 = \sum\frac{(O-E)^2}{E} = \frac{(100-75)^2}{75} + \frac{(80-100)^2}{100} + \frac{(20-25)^2}{25} + \frac{(50-75)^2}{75} + \frac{(120-100)^2}{100} + \frac{(30-25)^2}{25} = 26.67$$

Illustration –5/5

The final steps are to make the decision and summarize the results. This test is always a right-tailed test, and the degrees of freedom are $(R-1)(C-1) = (2-1)(3-1) = 2$. If $\alpha = 0.05$, the critical value from Table G is 5.991. Hence, the decision is to reject the null hypothesis, since $26.67 > 5.991$. See Figure 11–6.

Figure 11–6
Critical and Test Values for the Postoperative Procedures Example



The conclusion is that there is enough evidence to support the claim that opinion is related to (dependent on) profession—that is, that the doctors and nurses differ in their opinions about the procedure.

Example: College Education and Place of Residence

A sociologist wishes to see whether the number of years of college a person has completed is related to her or his place of residence. A sample of 88 people is selected and classified as shown. At α = 0.05, can the sociologist conclude that a person's location is dependent on the number of years of college?

Location	No College	Four-Year Degree	Advanced Degree	Total
Urban	15	12	8	35
Suburban	8	15	9	32
Rural	6	8	7	21
Total	29	35	24	88

Example: College Education and Place of Residence

Step 1: State the hypotheses and identify the claim.

Step 2: Find the critical value.

Example: College Education and Place of Residence

Compute the expected values.

$$E = \frac{(\text{row sum})(\text{column sum})}{\text{grand total}} \qquad E_{1,1} = \frac{(35)(29)}{88} = 11.53$$

Location	No College	Four-Year Degree	Advanced Degree	Total
Urban	15	12	8	35
Suburban	8	15	9	32
Rural	6	8	7	21
Total	29	35	24	88

Example: College Education and Place of Residence

Step 3: Compute the test value.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Example 11-6: Sports Preference

- A researcher wished to see if there is a difference in the favorite sport of males and the favorite sport of females. She selected a sample of 32 males and 48 females and asked them which of three sports was their favorite. The results are shown.

Gender	Football	Baseball	Hockey	Total
Male	18	10	4	32
Female	20	16	12	48
Total	38	26	16	80

Example 11-6: Sports Preference

Example 11-6: Sports Preference

Compute the expected values.

$$E = \frac{(\text{row sum})(\text{column sum})}{\text{grand total}} \qquad E_{1,1} = \frac{(27)(23)}{68} = 9.13$$

Gender	Alcohol Consumption			Total
	Low	Moderate	High	
Male	10 (9.13)	9	8	27
Female	13	16	12	41
Total	23	25	20	68

Example 11-6: Sports Preference

Step 3: Compute the test value.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Contingency Tables

	Column 1	Column 2	Column 3
Row 1	$C_{1,1}$	$C_{1,2}$	$C_{1,3}$
Row 2	$C_{2,1}$	$C_{2,2}$	$C_{2,3}$

- The degrees of freedom for any contingency table are d.f. = (rows – 1) (columns – 1) = $(R-1)(C-1)$.

Test for Independence

The formula for the test for independence:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where

d.f. = $(R-1)(C-1)$

O = observed frequency

E = expected frequency = $\dfrac{(\text{row sum})(\text{column sum})}{\text{grand total}}$

Chi-Square Independence Test

Purpose To perform a hypothesis test to decide whether two variables are associated

Assumptions

1. All expected frequencies are 1 or greater
2. At most 20% of the expected frequencies are less than 5
3. Simple random sample

STEP 1 The null and alternative hypotheses are

H_0: The two variables are not associated.

H_a: The two variables are associated.

STEP 2 Calculate the expected frequencies by using the formula $E = RC/n$, where R = row total, C = column total, and n = sample size. Place each expected frequency below its corresponding observed frequency in the contingency table.

STEP 3 Determine whether the expected frequencies satisfy Assumptions 1 and 2. If they do not, this procedure should not be used.

STEP 4 Decide on the significance level, α.

STEP 5 Compute the value of the test statistic

$$\chi^2 = \Sigma(O-E)^2/E,$$

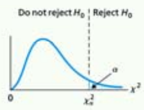
where O and E represent observed and expected frequencies, respectively. Denote the value of the test statistic χ^2_0.

Procedure (cont.)

CRITICAL-VALUE APPROACH or ***P*-VALUE APPROACH**

Critical-value approach

STEP 6 The critical value is χ^2_α with df $= (r-1) \times (c-1)$, where r and c are the number of possible values for the two variables. Use Table VII to find the critical value.



STEP 7 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

***P*-value approach**

STEP 6 The χ^2-statistic has df $= (r-1)(c-1)$, where r and c are the number of possible values for the two variables. Use Table VII to estimate the P-value, or obtain it exactly by using technology.

STEP 7 If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.

STEP 8 Interpret the results of the hypothesis test.

Example: Alcohol and Gender

A researcher wishes to determine whether there is a relationship between the gender of an individual and the amount of alcohol consumed. A sample of 68 people is selected, and the following data are obtained. At $\alpha = 0.10$, can the researcher conclude that alcohol consumption is related to gender?

	Alcohol Consumption			
Gender	**Low**	**Moderate**	**High**	**Total**
Male	10	9	8	27
Female	13	16	12	41
Total	23	25	20	68

Example: Alcohol and Gender

Step 1: State the hypotheses and identify the claim.

H_0: The amount of alcohol that a person consumes is independent of the individual's gender.

H_1: The amount of alcohol that a person consumes is dependent on the individual's gender (claim).

Step 2: Find the critical value.

Example: Alcohol and Gender

Compute the expected values.

$$E = \frac{(\text{row sum})(\text{column sum})}{\text{grand total}} \qquad E_{1,1} = \frac{(27)(23)}{68} = 9.13$$

	Alcohol Consumption			
Gender	**Low**	**Moderate**	**High**	**Total**
Male	10 (9.13)	9	8	27
Female	13	16	12	41
Total	23	25	20	68

Example: Alcohol and Gender

Step 3: Compute the test value.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Test for Homogeneity of Proportions

- **Homogeneity of proportions test** is used when samples are selected from several different populations and the researcher is interested in determining whether the proportions of elements that have a common characteristic are the same for each population.

Test for Homogeneity of Proportions

- The hypotheses are:
  - H_0: $p_1 = p_2 = p_3 = \dots = p_n$
  - H_1: At least one proportion is different from the others.
- When the null hypothesis is rejected, it can be assumed that the proportions are not all equal.

Assumptions for Homogeneity of Proportions

1. The data are obtained from a random sample.
2. The expected frequency for each category must be 5 or more.

Example 11-7: Money and Happiness

- A psychologist randomly selected 100 people from each of four income groups and asked them if they were "very happy." For people who made less than $30,000, 24% responded yes. For people who made $30,000 to $74,999, 33% responded yes. For people who made $75,000 to $90,999, 38% responded yes, and for people who made $100,000 or more, 49% responded yes. At *α=0.05, test the claim that there is no* difference in the proportion of people in each economic group who were very happy.

Household income	Less than $30,000 (24%)	$30,000–$74,999 (33%)	$75,000–$99,999 (38%)	$100,000 or more (49%)	Total
Yes	24	33	38	49	144
No	76	67	62	51	256
	100	100	100	100	400

Example 11-7: Money and Happiness

Step 3: Compute the test value.

Household income	Less than $30,000 (24%)	$30,000–$74,999 (33%)	$75,000–$99,999 (38%)	$100,000 or more (49%)	Total
Yes	24 (36)	33 (36)	38 (36)	49 (36)	144
No	76 (64)	67 (64)	62 (64)	51 (64)	256
	100	100	100	100	400

Example: Lost Luggage

A researcher selected 100 passengers from each of 3 airlines and asked them if the airline had lost their luggage on their last flight. The data are shown in the table. At $\alpha = 0.05$, test the claim that the proportion of passengers from each airline who lost luggage on the flight is the same for each airline.

	Airline 1	Airline 2	Airline 3	Total
Yes	10	7	4	21
No	90	93	96	279
Total	100	100	100	300

Example: Lost Luggage

Step 1: State the hypotheses.

Step 2: Find the critical value.

Example: Lost Luggage

Compute the expected values.

$$E = \frac{(\text{row sum})(\text{column sum})}{\text{grand total}} \qquad E_{1,1} = \frac{(21)(100)}{300} = 7$$

	Airline 1	Airline 2	Airline 3	Total
Yes	10 (7)	7 (7)	4 (7)	21
No	90 (93)	93 (93)	96 (93)	279
Total	100	100	100	300

Example: Luggage

Step 3: Compute the test value.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

獨立性檢定與齊一性檢定之比較

1. 一般常將齊一性檢定視為獨立性檢定的延伸，其所使用的檢定統計量、公式、判斷準則皆相同。
2. 獨立性檢定只從一母體分配中抽取一組隨機樣本，而使用不同變數對該資料加以分類；齊一性檢定則可視為由不同母體分別各抽取一組樣本而相互比較。
3. 獨立性檢定是檢定不同變數的分類是否相互獨立；而齊一性檢定則是檢定不同的隨機樣本在各個母體的比例是否一致。

因為齊一性檢定與獨立性檢定所使用的檢定統計量相同，在此不另外說明檢定的步驟 。

卡方檢定

- 適合度檢定
  - 檢定母體分配是否為某一特定分配或理論分配的統計方法。
  - 如學生成績分配是否為常態分配，或孟德爾的碗豆實驗發現其外型是否符合特定比例。
- 獨立性檢定
  - 檢定兩個屬性間是否獨立的統計方法。
  - 研究民中之教育程度與其對某門市之服務滿意度是否有關。
- 齊一性檢定
  - 檢定兩個或兩個以上母體的某一特性的分配(各類別的比例)是否相同或相近。
  - 檢定某項新產品在台北、新竹、台中等地之之名度是否相同。