Chapter 10. Discrete Data Analysis

10.1 Inferences on a Population Proportion

Sample proportion $\hat{p}$

- $X \sim B(n, p)$.
- $\hat{p} = \frac{\mathrm{x}}{\mathrm{n}}$.
- $E(\hat{p}) = p$ and $\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$.

For large n,

- $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) \approx N(p, \frac{\hat{p}(1-\hat{p})}{n})$

10.1.1 Confidence Intervals for Population Proportions

- Two-sided conf. intervals for a population proportion

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

- One-sided conf. intervals for a population proportion with a lower bound

$$\left(\hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \qquad 1\right)$$

- One-sided conf. intervals for a population proportion with a upper bound

$$\left(0, \qquad \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

- These approximate results are safe as long as both x and $\mathrm{n} - x$ are larger than 5.

- Example 57 : Building Tile Cracks

  Random sample n = 1250 of tiles in a certain group of downtown building for cracking. x = 98 are found to be cracked.

$$\hat{p} = \frac{98}{1250} = 0.0784. \; z_{0.005} = 2.576.$$

99% two-sided conf. interval

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = (0.0588,\ 0.0980)$$

10.1.2 Hypothesis Tests on a Population Proportion

- Two-sided hypothesis tests

$$H_0:\ p = p_0 \text{ vs } H_A:\ p \neq p_0$$

p-value$=2 \times \min\{P(X \geq x),\ P(X \leq x)\}$

where $X \sim B(n, p_0)$.

- When np_0 and $n(1-p_0)$ are both larger than 5, a normal approximation may be used to compute the p-value.

p-value$=2 \times \Phi(-|z|)$ where

$$z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$$

- Continuity correction can be used for a better approximation to the p-value.
- A size α hypothesis test rejects H_0 when

$$|z| > z_{\alpha/2} \text{ or p-value} < \alpha.$$

Otherwise, accept H_0.

One-sided hypothesis tests for a population proportion

- For testing

$$H_0:\ p \geq p_0 \text{ vs } H_A:\ p < p_0$$

P-value=$P(X \leq x)$ where $X \sim B(n, p_0)$

The p-value by the normal approximation:

P-value=$\Phi(z)$ where $\mathrm{z} = \frac{\mathrm{x+0.5-np_0}}{\sqrt{np_0(1-p_0)}}$.

- For testing

$$H_0:\ p \leq p_0 \text{ vs } H_A:\ p > p_0$$

P-value=$P(X \geq x)$ where $X \sim B(n, p_0)$

The p-value by the normal approximation:

P-value=$1 - \Phi(z)$ where $\mathrm{z} = \frac{\mathrm{x-0.5-np_0}}{\sqrt{np_0(1-p_0)}}$.

- Example 57 : Building Tile Cracks

10% or more of the building tiles are cracked ?

$$H_0\colon\ p \geq 0.1 \ \text{ vs } \ H_A\colon\ p < 0.1$$

From data: $n = 1250, \mathrm{x} = 98.$

$$z = \frac{x + 0.5 - np_0}{\sqrt{np_0(1 - p_0)}} = -2.50$$

P-value= $\Phi(-2.50) = 0.0062$

Python codes

```
import numpy as np
from statsmodels.stats.proportion import proportions_ztest
from scipy.stats import binom_test
zstat, pvalue = proportions_ztest(45,100,0.5)
print(“Two-sided 1 sample proportions test \n Z = %.4f, p-value = %.4f" %(zstat, pvalue))
```

```
Two-sided 1 sample proportions test
 Z = -1.0050, p-value = 0.3149
```

```
pvalue = binom_test(8,20,0.5)
print("Two-sided exact binomial test \n p-value = %.4f" %pvalue)
```

```
Two-sided exact binomial test
 p-value = 0.5034
```

10.1.3 Sample Size Calculations

- Consider a two-sided $1-\alpha$ level CI for p which is obtained by normal approximation

The interval length L is given by

$$L = 2\, z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

In case $\hat{p}$ is not available,

$$L = 2\, z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq z_{\alpha/2}\sqrt{\frac{1}{n}}$$

- Example 61 : Political Polling

  To determine the proportion p of people who agree with the statement “The city mayor is doing a good job.” within 3% accuracy, how many people do they need to poll?

  A 99% CI for p with length no larger than $\mathrm{L}_0 = 6\%$

$$L = 2\, z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq 0.06$$

  Since $\hat{p}$ is not available,

$$L \leq z_{0.005} \sqrt{\frac{1}{n}} \leq 0.06.$$

  The smallest sample size n satisfying the above inequality is desired.

$$n \geq \frac{z_{0.005}^2}{0.06^2} = \frac{2.576^2}{0.06^2} = 1843.3.$$

10.2 Comparing Two Population Proportions
10.2.1 Confidence Intervals for the Difference Between Two Population Proportions

- Assume $X \sim B(n, p_A)$ and $\mathrm{Y} \sim B(m, p_B)$ and X and Y are independent.
- Approximate two-sided $1 - \alpha$ level CI for $p_A - p_B$ with end-points:

$$\hat{p}_A - \hat{p}_B \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n} + \frac{\hat{p}_B(1-\hat{p}_B)}{m}}$$

- Approximate one-sided $1 - \alpha$ level CI for $p_A - p_B$ with a lower bound:

$$\left(\hat{p}_A - \hat{p}_B - z_{\alpha}\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n} + \frac{\hat{p}_B(1-\hat{p}_B)}{m}},\quad 1\right)$$

- Approximate two-sided $1 - \alpha$ level CI for $p_A - p_B$ with an upper bound:

$$\left(-1,\ \hat{p}_A - \hat{p}_B + z_{\alpha}\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n} + \frac{\hat{p}_B(1-\hat{p}_B)}{m}}\right)$$

- These approximations are reasonable as long as x, $n - x$, y, and $n - y$ are all larger than 5.

- Example 57 : Building Tile Cracks

  Building A : 406 cracked tiles out of n = 6000.

  Building B : 83 cracked tiles out of m = 2000.

$\hat{p}_A = \frac{406}{6000} = 0.0677. \; \hat{p}_B = \frac{83}{2000} = 0.0415.$

A $1-\alpha$ level CI for $p_A - p_B$:

$$\hat{p}_A - \hat{p}_B \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n} + \frac{\hat{p}_B(1-\hat{p}_B)}{m}}$$

$\Longrightarrow (0.0120,\ 0.0404)$ when $\alpha = 0.01$.

10.2.2 Hypothesis Tests on the Difference Between Two Population Proportions

- For testing H_0: $p_A = p_B$ vs H_A: $p_A \neq p_B$

  p-value=$2 \times \Phi(-|z|)$ where

  $$z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n}+\frac{1}{m}\right)}} \text{ and } \hat{p} = \frac{x+y}{n+m}$$

- For testing H_0: $p_A \geq p_B$ vs H_A: $p_A < p_B$

  p-value=$\Phi(z)$

- For testing H_0: $p_A \leq p_B$ vs H_A: $p_A > p_B$

  p-value=$1 - \Phi(z)$

- Conclusion:

  Reject H_0 if p-value is smaller than the sig. level α.

  Otherwise, accept H_0.

- Example 57 : Building Tile Cracks

  Test $H_0 : p_A = p_B$ vs $H_A : p_A \neq p_B$

  $\hat{p}_A = \frac{406}{6000} = 0.0677.\ \hat{p}_B = \frac{83}{2000} = 0.0415.$

  p-value = $2 \times \Phi(-|z|) \approx 0$ where

$$z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n}+\frac{1}{m}\right)}} = 4.3755 \quad \text{and}$$

$$\hat{p} = \frac{x+y}{n+m} = \frac{489}{8000} = 0.0611.$$

Python codes for two-sample test of proportions

```
from statsmodels.stats.proportion import proportions_ztest
stat, pvalue = proportions_ztest([12, 8], [34, 24])
                    #([x1,x2],[n1,n2])
print("2 sample test for equality of proportions \n z = %.4f, p-value = %.4f“  %(stat, pvalue))
    2 sample test for equality of proportions
    z = 0.1547, p-value = 0.8770
```

10.3 Goodness of Fit Tests for One-Way Contingency Tables
10.3.1 One-Way Classifications

Each of n observations is classified into one of k categories or cells.

Cell frequencies: $\mathrm{x}_1, \cdots, x_k$. $\quad \sum_{i=1}^{k} x_i = n$

Cell probabilities: $\mathrm{p}_1, \cdots, p_k$. $\quad \sum_{i=1}^{k} p_i = 1$.

- Test H_0: $\mathrm{p_i} = \mathrm{p_i^*}, \quad \mathrm{i} = 1, \cdots, k$ vs $\mathrm{H_A}$: $\quad not \;\; H_0$.

Under H_0, the expected cell frequency at cell i, e_i, is given by

$$\mathrm{e_i} = \mathrm{np_i^*}$$

Two test statistics:

(1) Pearson's Chi-square statistic:

$$\mathrm{X}^2 = \sum_{i=1}^{k} \frac{(\mathrm{x_i} - \mathrm{e_i})^2}{\mathrm{e_i}}.$$

(2) Likelihood ratio Chi-square statistic:

$$G^2 = 2\sum_{i=1}^{k} x_i \ln(\frac{x_i}{e_i}).$$

Both of the statistics, X^2 and G^2, follow asymptotically Chi-square distribution with $\text{df} = k - 1$.

This asymptotic result is reasonable as long as all the e_i's are larger than 5.

P-value $= P(X^2 \geq obs(X^2))$

Conclusion: Reject H_0 if the p-value is smaller than the sig. level α.
Otherwise, accept H_0.

Mathematics for goodness-of-fit test

- Likelihood function $L(p_1, \cdots, p_k) = f(x^1, x^2, \cdots, x^n; p_1, \cdots, p_k) = \prod_{i=1}^{k} p_i^{x_i}$

  where x^l is the *l*-th observation of data.

- $\log L = \sum_{i=1}^{k} x_i \log p_i$.

- For i $\neq k$, $\frac{\partial \log L}{\partial p_i} = x_i \frac{\partial \log p_i}{\partial p_i} + x_k \frac{\partial \log p_k}{\partial p_i} = \frac{x_{i\cdot}}{p_{i\cdot}} - \frac{x_k}{p_k}$.

  $\frac{\partial \log L}{\partial p_i} = 0. \quad \Longrightarrow \quad \hat{p}_i = \hat{p}_k \frac{x_i}{x_k}$.

  Therefore, $\hat{p}_i = \frac{x_i}{n}$, i = 1,2, $\cdots$,k.

- $\gamma = \log \frac{\prod_{i=1}^{k} \hat{p}_i^{\,x_i}}{\prod_{i=1}^{k} p_i^{x_i}} = \log \prod_{i=1}^{k} \left(\frac{\hat{p}_i}{p_i}\right)^{x_i} = \sum_{i}^{k} x_i \log \frac{\hat{p}_i}{p_i}$

  $G^2 = 2\,\gamma$ when p_i's are the cell probabilities under H_0.

- Example 1 : Machine Breakdowns

  $n = 46$ machine breakdowns.

  $x_1 = 9$: electrical problems

  $x_2 = 24$: mechanical problems

  $x_3 = 13$: operator misuse

  It is suggested that the cell probabilities are

  $$p^*_1 = 0.2,\ p^*_2 = 0.5,\ p^*_3 = 0.3.$$

For testing H_0 : $p_1 = 0.2$, $p_2 = 0.5$, $p_3 = 0.3$ vs H_A: $not \quad H_0$.

	Electrical	Mechanical	Operator misuse	
Observed cell freq.	$x_1 = 9$	$x_2 = 24$	$x_3 = 13$	n = 46
Expected cell freq.	$e_1 = 46*0.2$ $= 9.2$	$e_2 = 46*0.5$ $=23.0$	$e_3 = 46*0.3$ $=13.8$	n = 46

$X^2 = 0.0942.$ $\quad G^2 = 0.0945.$ $\quad$ df=3-1=2.

P-value ≈ $P(X^2 \geq 0.0942) = 0.95.$

- Check for homogeneity

$$H_0: p_1 = p_2 = p_3 = \frac{1}{3} \quad \text{vs } H_A: \ not \ \ H_0$$

P-value=$P(X^2 \geq 7.87) \approx 0.02.$

10.3.2 Testing Distributional Assumptions

Number of errors found in a software product	0	1	2	3	4	5	6	7	8	
Frequency	3	14	20	25	14	6	2	0	1	$n = 85$

H_0 : number of errors, X, has a Poisson distribution with mean $\lambda = 3.0$

Cell	Expected cell frequency		
$X = 0$	$e_1 = 85 \times P(X = 0) = 85 \times \frac{e^{-3} \times 3^0}{0!}$	$= 4.23$	Group
$X = 1$	$e_2 = 85 \times P(X = 1) = 85 \times \frac{e^{-3} \times 3^1}{1!}$	$= 12.70$	
$X = 2$	$e_3 = 85 \times P(X = 2) = 85 \times \frac{e^{-3} \times 3^2}{2!}$	$= 19.04$	
$X = 3$	$e_4 = 85 \times P(X = 3) = 85 \times \frac{e^{-3} \times 3^3}{3!}$	$= 19.04$	
$X = 4$	$e_4 = 85 \times P(X = 4) = 85 \times \frac{e^{-3} \times 3^4}{4!}$	$= 14.28$	
$X = 5$	$e_5 = 85 \times P(X = 5) = 85 \times \frac{e^{-3} \times 3^5}{5!}$	$= 8.57$	
$X = 6$	$e_6 = 85 \times P(X = 6) = 85 \times \frac{e^{-3} \times 3^6}{6!}$	$= 4.28$	Group
$X = 7$	$e_7 = 85 \times P(X = 7) = 85 \times \frac{e^{-3} \times 3^7}{7!}$	$= 1.84$	
$X = 8$	$e_8 = 85 \times P(X = 8) = 85 \times \frac{e^{-3} \times 3^8}{8!}$	$= 0.69$	
$X \geq 9$	$e_9 = 85 \times P(X \geq 9)$	$= 0.33$	
		$n = 85.0$	

- Example 3 : Software Errors

  For some of expected values are smaller than 5, it is appropriate to group the cells.

- Test if (H_0) the data are from a Poisson distribution with mean=3.

				After grouping			
Number of errors	0–1	2	3	4	5	≥ 6	
Observed cell frequency	$x_1 = 17$	$x_2 = 20$	$x_3 = 25$	$x_4 = 14$	$x_5 = 6$	$x_6 = 3$	$n = 85$
Expected cell frequency	$e_1 = 16.93$	$e_2 = 19.04$	$e_3 = 19.04$	$e_4 = 14.28$	$e_5 = 8.57$	$e_6 = 7.14$	$n = 85$

$$X^2 = \frac{(17.00 - 16.93)^2}{16.93} + \frac{(20.0 - 19.04)^2}{19.04} + \frac{(25.00 - 19.04)^2}{19.04}$$
$$+ \frac{(14.00 - 14.28)^2}{14.28} + \frac{(6.00 - 8.57)^2}{8.57} + \frac{(3.00 - 7.14)^2}{7.14}$$
$$= 5.12$$

P-value = $P(X^2 \geq 5.12) = 0.40$ where X^2 follows a Chi-square distribution with df=5.

10.4 Testing for Independence in Two-Way Contingency Tables
10.4.1 Two-Way Classifications

- A two-way (r x c) contingency table.

Second Categorization

First Categorization

	Level 1	Level 2	$\cdots$	Level j	$\cdots$	Level c	
Level 1	x_{11}	x_{12}	$\cdots$		$\cdots$	x_{1c}	$x_{1\cdot}$
Level 2	x_{21}	x_{22}	$\cdots$		$\cdots$	x_{2c}	$x_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$		$\ddots$	$\vdots$	$\vdots$
Level i	$\cdots$	$\cdots$	$\cdots$	x_{ij}	$\cdots$	$\cdots$	$x_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$		$\ddots$	$\vdots$	$\vdots$
Level r	x_{r1}	x_{r2}	$\cdots$		$\cdots$	x_{rc}	$x_{r\cdot}$
	$x_{\cdot 1}$	$x_{\cdot 2}$	$\cdots$	$x_{\cdot j}$	$\cdots$	$x_{\cdot c}$	$n = x_{\cdot\cdot}$

Row marginal frequencies

Column marginal frequencies

Example 57 : Building Tile Cracks

	Location			
Tile Condition		Building A	Building B	
	Undamaged	$x_{11} = 5594$	$x_{12} = 1917$	$x_{1.} = 7511$
	Cracked	$x_{21} = 406$	$x_{22} = 83$	$x_{2.} = 489$
		$x_{.1} = 6000$	$x_{.2} = 2000$	$n = x_{..} = 8000$

Notice that the column marginal frequencies are fixed.
($x_{.1} = 6000$, $x_{.2} = 2000$)

10.4.2 Testing for Independence

- Testing for independence in a Two-way contingency table

$$H_0\text{: Two factors are independent} \quad \text{vs} \quad H_A\text{: not } H_0$$

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(x_{ij}-e_{ij}\right)^2}{e_{ij}}. \qquad G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c}x_{ij}\ln\left(\frac{x_{ij}}{e_{ij}}\right).$$

Here $e_{ij} = \frac{x_{i\cdot}x_{\cdot j}}{n}$.

The two test statistics follow asymptotically Chi-square distribution with df = $\mathrm{rc} - 1 - (\mathrm{r} - 1) - (\mathrm{c} - 1) = (r-1)(c-1)$.

This result is valid as long as all the e_{ij}'s are larger than 5.

P-value= $P(X^2 \geq obs(X^2))$

- Example 57 : Building Tile Cracks

	Building A	Building B	
Undamaged	$x_{11} = 5594$ $e_{11} = 5633.25$	$x_{12} = 1917$ $e_{12} = 1877.75$	$x_{1.} = 7511$
Cracked	$x_{21} = 406$ $e_{21} = 366.75$	$x_{22} = 83$ $e_{22} = 122.25$	$x_{2.} = 489$
	$x_{.1} = 6000$	$x_{.2} = 2000$	$n = x_{..} = 8000$

$$obs(X^2) = 17.896$$

$$P - value = 0.000023 \approx 0$$

Python codes for Independence test

```
import numpy as np
import pandas as pd
from scipy.stats import chi2_contingency
chi, pvalue, dof, expctd = chi2_contingency(np.array[[24,12],[8,10]])
print("Pearson's Chi-squared test \nX-squared = %.4f, p-value = %.4f,
df = %d,"  %(chi, pvalue, dof))
print("Expected cell frequencies:\n", pd.DataFrame(expctd,
index=[1,2], columns=[1,2]))
```

```
Pearson's Chi-squared test
X-squared = 1.6204, p-value = 0.2030, df = 1,
Expected cell frequencies:
           1          2
1  21.333333  14.666667
2  10.666667   7.333333
```

Mathematics for Independence Test (1)

- Likelihood function $L(p_{11}, \cdots, p_{rc}) = f(x^1, x^2, \cdots, x^n; p_{11}, \cdots, p_{rc}) = \prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{x_{ij}}$

  where x^l is the *l*-th observation of data.
- $\log L = \sum_i^r \sum_j^c x_{ij} \log p_{ij}$.
- Under H_0:

$$\log L = \sum_i^r \sum_j^c x_{ij} \log p_{i\cdot} p_{\cdot j} = \sum_i^r \sum_j^c x_{ij} \log p_{i\cdot} + \sum_i^r \sum_j^c x_{ij} \log p_{\cdot j}$$

$$= \sum_{i=1}^r x_{i\cdot} \log p_{i\cdot} + \sum_j^c x_{\cdot j} \log p_{\cdot j}$$

For $\mathrm{i} \neq r$, $\frac{\partial \log L}{\partial p_{i\cdot}} = x_{i\cdot} \frac{\partial \log p_{i\cdot}}{\partial p_{i\cdot}} + x_{r\cdot} \frac{\partial \log p_{r\cdot}}{\partial p_{i\cdot}} = \frac{x_{i\cdot}}{p_{i\cdot}} - \frac{x_{r\cdot}}{p_{r\cdot}}$.

$\frac{\partial \log L}{\partial p_{i\cdot}} = 0. \quad \Longrightarrow \quad \hat{p}_{i\cdot} = \hat{p}_{r\cdot} \frac{x_{i\cdot}}{x_{r\cdot}}$.

Therefore, $\hat{p}_{i\cdot} = \frac{x_{i\cdot}}{n}$, $\mathrm{i} = 1,2, \cdots, r$.

In the same way, we obtain $\hat{p}_{\cdot j} = \frac{x_{\cdot j}}{n}$, $\mathrm{j} = 1,2, \cdots, \mathrm{c}$.

Mathematics for Independence Test (2)

- Under H_0:

$$\mathrm{e_{ij}} = \mathrm{n}\hat{p}_{ij} = n\hat{p}_{i\cdot}\hat{p}_{\cdot j} = \frac{x_{i\cdot}x_{\cdot j}}{n}$$

Example 10.4.a SAT score and occupation

- The following data is a two-dimensional contingency table data of 4353 individuals. They are cross-classified into 4 occupational groups (O) and 5 aptitude levels (A) as measured by a SAT test (Beaton, 1975) The aptitude levels are from low (A1) to high (A5) and the occupational levels are:

  O1 = self-employed, business

  O2 = self-employed, professional

  O3 = teacher

  O4 = salaried, employed

- Test if the SAT score and the occupation are independent with the sig. level 0.05.

✓ Beaton, A.E. (1975). The influence of educational and ability on alary and attitudes. In F.T. Juster (ed.), *Education, Income, and Human Behavior}, pp. 365-396. New York, McGraw-Hill.*

```
import numpy as np
import pandas as pd
from scipy.stats import chi2_contingency
aptocc = np.array([[122,30,20,472],[226,51,66,704],[306,115,96,1072], [130,59,38,501],[50,31,15,249]])
aptocc = pd.DataFrame(data=aptocc, index=['A1','A2','A3','A4','A5'], columns=['O1','O2','O3','O4'])
print("Observed cell frequencies:\n", aptocc)
```

```
Observed cell frequencies:
     O1   O2  O3    O4
A1  122   30  20   472
A2  226   51  66   704
A3  306  115  96  1072
A4  130   59  38   501
A5   50   31  15   249
```

```
chi, pvalue, dof, expctd = chi2_contingency(aptocc)

print("Pearson's Chi-squared test \nX-squared = %.4f, p-value = %.4f, df = %d,"  %(chi, pvalue, dof))

print("Expected cell frequencies:\n", pd.DataFrame(expctd,  index=['A1','A2','A3','A4','A5'],
columns=['O1','O2','O3','O4']))
```

```
Pearson's Chi-squared test
X-squared = 35.7989, p-value = 0.0003, df = 12,
Expected cell frequencies:
            O1          O2         O3           O4
A1  123.385252   42.311969  34.766827   443.535952
A2  200.596830   68.789800  56.523088   721.090283
A3  304.439697  104.400184  85.783368  1094.376752
A4  139.478980   47.830921  39.301631   501.388468
A5   66.099242   22.667126  18.625086   237.608546
```

- Conclusion:

This result strongly suggests that the SAT level and the occupation are not independent at the sig. level 0.05.

Simpson's paradox

Suppose P(A|B'∩ C_i) < P(A´|B' ∩ C_i) for $i = 1,2,\cdots,k$,
with $\sum_{i=1}^{k} P(C_i) = 1$.

It is possible that P(A|B') > P(A´|B')

This phenomenon is called a Simpson's paradox.

Example 41 (Internet commerce).

FIGURE 10.34

Simpson's paradox

		Internet sales	Telephone sales
Product A Sales	New customers	199 (11.10%)	63 (6.71%)
	Repeat customers	1594 (88.90%)	876 (93.29%)
		1793	939
Product B Sales	New customers	243 (11.10%)	138 (9.98%)
	Repeat customers	1946 (88.90%)	1245 (90.02%)
		2189	1383
Product C Sales	New customers	864 (16.15%)	1107 (15.90%)
	Repeat customers	4486 (83.85%)	5855 (84.10%)
		5350	6962
Product D Sales	New customers	128 (38.32%)	180 (36.59%)
	Repeat customers	206 (61.68%)	312 (63.41%)
		334	492
Total Sales	New customers	1434 (14.84%)	1488 (15.22%)
	Repeat customers	8232 (85.16%)	8288 (84.78%)
		9666	9776

Chapter summary

10.1 Inferences on a Population Proportion

10.2 Comparing Two Population Proportions

10.3 Goodness of Fit Tests for One-Way Contingency Tables

Two test statistics. Asymptotic distributions of them.

Conditions for asymptotic distributions.

10.4 Testing for Independence in Two-Way Contingency Tables