Chapter 24

Two-Way Tables and the Chi-Square Test

We look at two-way tables to determine association of paired qualitative data. We look at marginal distributions, conditional distributions and bar graphs. We also discuss Simpson's Paradox, analogous to lurking variables in paired quantitative data. We perform a chi-square test using statistic

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i},$$

where E_i = (row total) × (column total) ÷ (table total), which is approximately chi-square, $(r-1)(c-1)$ degrees of freedom, provided expected counts $E_i \geq 1$ and at least 80% of expected counts are greater than 5.

Exercise 24.1 (Two-Way Tables and the Chi-Square Test)

1. *Two-way table: association between fathers, sons and attending college.*
   Data from a sample of 80 families in a midwestern city gives record of college attendance by fathers and their oldest sons.

	son attended college	son did not attend college	
father attended college	18	7	25
father did not attend college	22	33	55
	40	40	80

   (a) *Some (marginal) percentage questions.*
   Proportion of fathers who attended college $\frac{25}{80}$ = **0.3125 / 0.5 / 0.6875**
   Proportion of sons who attended college $\frac{40}{80}$ = **0.3125 / 0.5 / 0.6875**

   (b) *Some (conditional) percentage questions.*
   Proportion of sons who attended college,

if fathers attended college $\frac{18}{25}$ = **0.28 / 0.5 / 0.72**

Proportion of sons who attended college,
if fathers did not attend college $\frac{22}{55}$ = **0.28 / 0.4 / 0.72**

Percentage of sons who attended college,
if fathers did not attend college **28% / 40% / 72%**

(c) *Son's attendance associated with father's attendance?*
Is son's attendance (response) influenced by father's attendance (explanatory)? Complete conditional table.

divide by *row* totals	son attended college	son did not attend college	
father attended college	$\frac{18}{25}$ = ______	$\frac{7}{25} = 0.28$	$\frac{25}{25} = 1$
father did not attend college	$\frac{22}{55}$ = ______	$\frac{33}{55} = 0.6$	$\frac{55}{55}$ ______

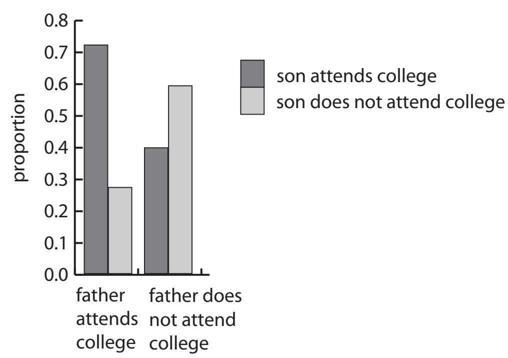

Figure 24.1 (Bar graph: son conditional on father.)

Response variable is **father's attendance / son's attendance** because son's attendance *divided by* father's attendance.
There *appears* to be **an / no** association: son attends college more likely if father attends college, less likely if father does not attend college.

2. *Two-way and three-way tables: association between drug, flu symptoms and gender lurking variable.* Are flu symptoms (response) influenced by drug (explanatory)?

flu symptoms →	reduced	not reduced	totals
drug	100	50	150
no drug	200	100	300
totals	300	150	450

(a) *Flu symptoms associated with drug?*
Complete conditional table.

flu symptoms →	reduced	not reduced	
drug	$\frac{100}{150} =$ ______	$\frac{50}{150} = 0.33$	$\frac{150}{150} =$ ______
no drug	$\frac{200}{300} =$ ______	$\frac{100}{300} = 0.33$	$\frac{300}{300} = 1$

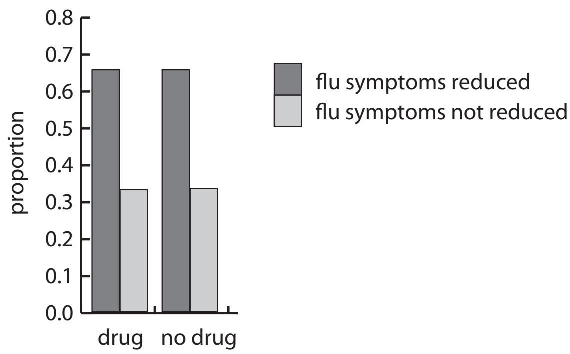

Figure 24.2 (Bar graph: flu symptoms not associated with drug.)

Response variable is **flu symptoms** / **drug** because
flu symptom counts is *divided* by drug count row totals.
There appears to be **an** / **no** association:
flu symptoms same whether drug given or not.

(b) *Lurking variable: gender.* Doctors suspect gender is confounding results. Consequently, *to control for gender*, they create a *three-way* table by tabulating effect of drug on males and, separate from this, tabulating effect of drug on females.

male	reduced	not reduced	subtotals
drug	80	40	120
no drug	100	80	180
subtotals	180	120	300

female	reduced	not reduced	subtotals
drug	20	10	30
no drug	100	20	120
subtotals	120	30	150

Complete conditional table for both males and females.

males	reduced	not reduced	subtotals
drug	$\frac{80}{120} =$ ______	$\frac{40}{120} =$ ______	$\frac{120}{120} =$ ______
no drug	$\frac{100}{180} = 0.55$	$\frac{80}{180} = 0.44$	$\frac{180}{180} =$ ______
subtotals	$\frac{180}{300} = 0.6$	$\frac{120}{300} = 0.4$	300 $\frac{300}{300} = 1$

females	reduced	not reduced	subtotals
drug	$\frac{20}{30} =$ ______	$\frac{10}{30} =$ ______	$\frac{30}{30} =$ ______
no drug	$\frac{100}{120} = 0.83$	$\frac{20}{120} = 0.17$	$\frac{120}{120} =$ ______
subtotals	$\frac{120}{150} = 0.8$	$\frac{30}{150} = 0.2$	$\frac{150}{150} = 1$

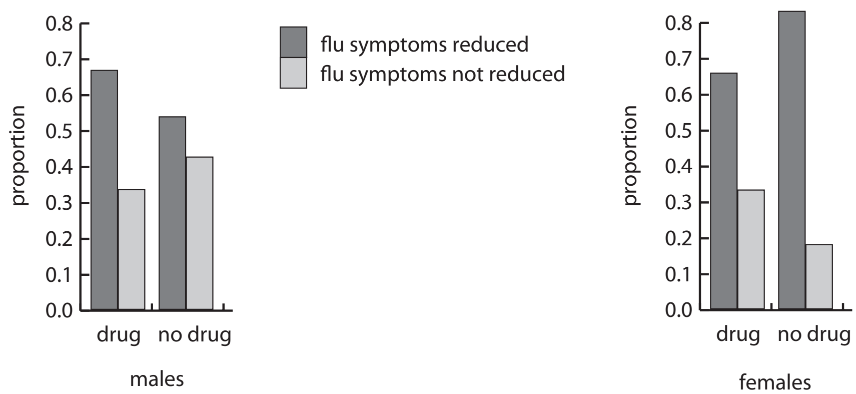

Figure 24.3 (Bar graph: flu associated with drug, males/females.)

There appears to be **an** / **no** association for *males*:
more likely flu symptoms reduced when taking drug than not taking drug.
There appears to be **an** / **no** association for *females*:
less likely flu symptoms reduced when taking drug than not taking drug.

(c) **True** / **False** Although combined study demonstrates *no* association between drug and reduced flu symptoms, a positive association between drug and reduced flu symptoms occurs for males, whereas a negative association between drug and reduced flu symptoms occurs for females. This is an example of *Simpson's paradox* where association changes with introduction of third (lurking) variable.

3. *Chi-square test: fathers, sons and college.*
Random sample of college attendance by fathers and their oldest sons in a midwestern city recorded in table below. Test whether or not a son attends college is associated with whether or not father attends college at $\alpha = 0.01$.
No matter how this question is worded, null hypothesis for test is *always* "not associated" (or "not related") and alternative hypothesis is *always* "associated" (or "related").

observed, O_i	son attended college	son did not attend college	
father attended college	18	7	25
father did not attend college	22	33	55
	40	40	80

(a) *Statement.* Choose one.

i. H_0 : son attends equals father attending
versus H_a : son attends does not equal to father attending

ii. H_0 : son attends not associated with father attending
versus H_a : son attends associated with father attending

iii. H_0 : son attends associated with father attending
versus H_a : son attends not associated with father attending

(b) *Test.*

attendance	observed, O_i	expected, E_i, if not associated	$\frac{(O_i-E_i)^2}{E_i}$
both father and son	18	$\frac{25\cdot 40}{80} \approx 12.5$	$\frac{(18-12.5)^2}{12.5} \approx 2.42$
not father, son does	22	$\frac{55\cdot 40}{80} \approx 27.5$	$\frac{(22-27.5)^2}{27.5} \approx 1.1$
father does, not son	7	$\frac{25\cdot 40}{80} \approx 12.5$	$\frac{(7-12.5)^2}{12.5} \approx 2.42$
neither father nor son	33	$\frac{55\cdot 40}{80} \approx 27.5$	$\frac{(33-27.5)^2}{27.5} \approx 1.1$

Observed test statistic is

$$\sum \frac{(O_i - E_i)^2}{E_i} = 2.42 + 1.1 + 2.42 + 1.1 = 7.04$$

with degrees of freedom

$$\text{(number of rows} - 1) \times \text{(number of columns} - 1) \\ = (2-1) \times (2-1) =$$

(circle one) **1** / **2** / **3** df,
and so, using table 24.1,

(i) P-value < 0.001
(ii) $0.01 >$ P-value > 0.001
(iii) $0.05 >$ P-value > 0.01
(iv) $0.10 >$ P-value > 0.05
(v) $0.15 >$ P-value > 0.10

(c) *Conclusion.*
Since $0.01 >$ P-value > 0.0001 which is less than 0.01,
do not reject / **reject** null H_0 : not associated.
Observed data indicates whether or not a son attends college
not associated with / **associated with**
whether or not father attends college

4. *Chi-square test: flu symptoms and drug.*
Consider *observed* data from a random sample of 354 patients in an investigation of effect of a new drug on reducing flu symptoms. Test whether or not reduction of flu symptoms is associated with whether or not drug is administered at $\alpha = 0.01$.

observed, O_i	drug	no drug	subtotals
flu symptoms reduced	100	50	150
flu symptoms not reduced	200	100	300
subtotals	300	150	450

(a) *Statement.* Choose one.

i. H_0 : flu symptoms equals of drug
versus H_a : flu symptoms does not equal drug

ii. H_0 : flu symptoms associated with drug
versus H_a : flu symptoms not associated with drug

iii. H_0 : flu symptoms not associated with drug
versus H_a : flu symptoms associated with drug

(b) *Test.*

flu study	observed, O_i	expected, E_i	$\frac{(O_i-E_i)^2}{E_i}$
drug given, flu reduced	100	$\frac{150\cdot 300}{450} \approx 100$	$\frac{(100-100)^2}{100} \approx 0$
drug not given, flu reduced	200	$\frac{300\cdot 300}{450} \approx 200$	$\frac{(200-200)^2}{200} \approx 0$
drug given, flu not reduced	50	$\frac{150\cdot 150}{450} \approx 50$	$\frac{(50-50)^2}{50} \approx 0$
drug not given, flu not reduced	100	$\frac{300\cdot 150}{450} \approx 100$	$\frac{(100-100)^2}{100} \approx 0$

Observed test statistic is

$$\sum \frac{(O_i - E_i)^2}{E_i} = 0 + 0 + 0 + 0 =$$

0 / **21.33** / **25.46**,
with degrees of freedom

$$(\text{number of rows} - 1) \times (\text{number of columns} - 1)$$
$$= (2 - 1) \times (2 - 1) =$$

(circle one) **1** / **2** / **3** df,
and so, using table 24.1,

(i) P-value < 0.001

(ii) $0.01 >$ P-value > 0.001

(iii) $0.05 >$ P-value > 0.01

(iv) $0.10 >$ P-value > 0.05

(v) $0.15 >$ P-value > 0.10

(vi) P-value > 0.25

(c) *Conclusion.*
Since P-value $> 0.25 > \alpha = 0.01$,

do not reject / **reject** null H_0 : not associated.
Data indicates flu symptoms are
not associated with / **associated with** drug.

(d) *How are flu symptoms and drug associated?*

flu symptoms →	reduced	not reduced	
drug	$\frac{100}{150} = 0.33$	$\frac{50}{150} = 0.33$	$\frac{150}{150} = 1$
no drug	$\frac{200}{300} = 0.67$	$\frac{100}{300} = 0.67$	$\frac{300}{300} = 1$

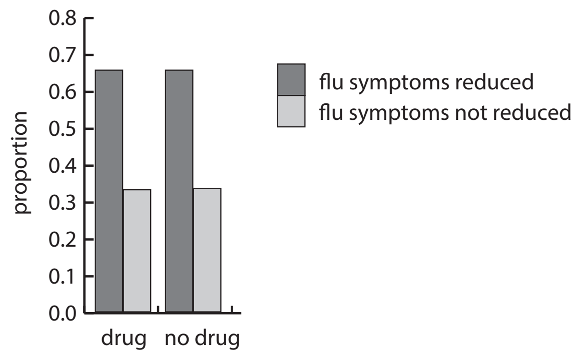

Figure 24.4 (Bar graph: flu symptoms not associated with drug.)

This confirms there is **an** / **no** association:
flu symptoms same whether drug given or not.

(e) *Check assumptions.*

i. All E_i **are** / **are not** greater than 1.

ii. At least 80% of E_i should be more than 5.
In fact, **0%** / **50%** / **100%** of $E_i > 5$.

5. *Chi-square test: plant growth and nutrition.*
Consider *observed* data from a random sample of 390 plants in an investigation of effect of nutritional level on plant growth. Test if proportion plant growth is associated with nutrition levels at $\alpha = 0.05$.

O_i	nutritional level →	poor	adequate	excellent	row totals
plant	below average	70	95	35	200
growth	above average	90	30	70	190
	column totals	160	125	105	390

(a) *Statement.* Choose one.

i. H_0 : plant growth not associated with nutrition
versus H_a : plant growth associated with nutrition

ii. H_0 : plant growth associated with nutrition
versus H_a : plant growth dependent on nutrition

iii. H_0 : plant growth not equal to nutrition
versus H_a : plant growth equal to nutrition

(b) *Test.*

plant study	O_i	E_i	$\frac{(O_i-E_i)^2}{E_i}$
below plant, poor nutrition	70	$\frac{(200)(160)}{390}$ = ______	$\frac{(70-82.1)^2}{82.1} \approx$ ______
above plant, poor nutrition	90	$\frac{(190)(160)}{390}$ = ______	$\frac{(90-77.9)^2}{77.9} \approx$ ______
below plant, adequate nutrition	95	$\frac{(200)(125)}{390}$ = ______	$\frac{(95-64.1)^2}{64.1} \approx$ ______
above plant, adequate nutrition	30	$\frac{(190)(125)}{390}$ = ______	$\frac{(30-60.9)^2}{60.9} \approx$ ______
below plant, excellent nutrition	35	$\frac{(200)(105)}{390}$ = ______	$\frac{(35-53.8)^2}{53.8} \approx$ ______
above plant, excellent nutrition	70	$\frac{(190)(105)}{390}$ = ______	$\frac{(70-51.2)^2}{51.2} \approx$ ______

Observed test statistic is

$$\sum \frac{(O_i - E_i)^2}{E_i} \approx 1.77 + 1.86 + 14.9 + 15.7 + 6.6 + 6.9 =$$

(circle one) **32.2** / **41.3** / **47.7**,
with degrees of freedom

$$(\text{number of rows} - 1) \times (\text{number of columns} - 1)$$
$$= (2-1) \times (3-1) =$$

(circle one) **1** / **2** / **3** df,
and so, using table 24.1,

(i) P-value < 0.001
(ii) 0.01 > P-value > 0.001
(iii) 0.05 > P-value > 0.01
(iv) 0.10 > P-value > 0.05
(v) 0.15 > P-value > 0.10

(c) *Conclusion.*
Since P-value = 0.00 < α = 0.05,
do not reject / **reject** null H_0 : no association.
Data indicates plant growth
associated with / **not associated with**
for different nutrition levels.

(d) *How is plant growth and nutrition associated?*

O_i	nutritional level →	poor	adequate	excellent	row totals
plant	below average	$\frac{70}{160} \approx 0.44$	$\frac{95}{125} = 0.76$	$\frac{35}{105} \approx 0.33$	200
growth	above average	$\frac{90}{160} \approx 0.56$	$\frac{30}{125} = 0.24$	$\frac{70}{105} \approx 0.67$	190
	column totals	160	125	105	390

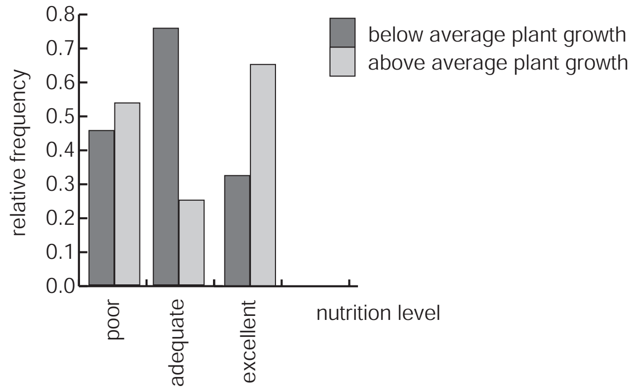

Figure 24.5 (Bar graph: plant growth associated with nutrition.)

Bar graph indicates plant growth
associated with / not associated with
nutrition levels.