Chapter 4

Describing the Relation Between Two Variables

We look at scatter diagrams, linear correlation and regression for paired (bivariate) quantitative data sets and contingency tables for paired qualitative data, related to qualitative-quantitative analysis of experimental and observed study data.

4.1 Scatter Diagrams and Correlation

Scatter diagram is graph of paired *sampled* data and linear correlation is a measure of linearity of scatter plot.

Exercise 4.1 (Scatter Diagrams and Correlation)

1. *Scatter Diagram: Reading Ability Versus Brightness.*

brightness, x	1	2	3	4	5	6	7	8	9	10
ability to read, y	70	70	75	88	91	94	100	92	90	85

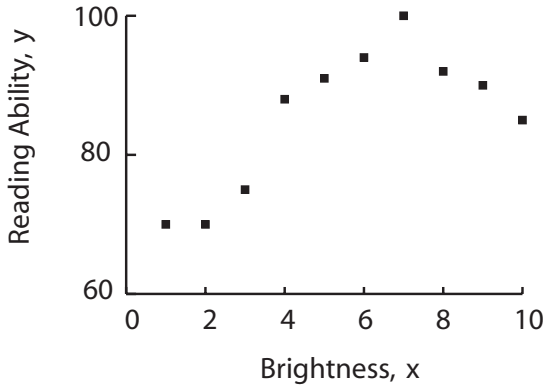

Figure 4.1 (Scatter Diagram, Reading Ability Versus Brightness)

(StatCrunch: Relabel var1 as brightness and var2 as reading ability. Type data into two columns. Graphics, Scatter Plot, X variable: brightness, Y variable: reading ability, Create Graph!) Notice scatter plot may be misleading because y-axis ranges 60 to 80, rather than 0 to 80.

(a) There are (circle one) **10** / **20** / **30** data points.
One particular data point is (circle one) **(70, 75)** / **(75, 2)** / **(2, 70)**.
Data point (9,90) means (circle one)
   i. for brightness 9, reading ability is 90.
   ii. for reading ability 9, brightness is 90.

(b) Reading ability **positively** / **not** / **negatively** *associated* to brightness.
As brightness increases, reading ability (circle one) **increases** / **decreases**.

(c) Association **linear** / **nonlinear (curved)** because straight line cannot be drawn on graph where all points of scatter fall on or near line.

(d) "Reading ability" is **response** / **explanatory** variable and "brightness" is **response** / **explanatory** variable because reading ability depends on brightness, not the reverse

Sometimes it is not so obvious which is response variable and which is explanatory variable. For example, it is not immediately clear which is explanatory variable and response variable for a scatter plot of husband's IQ scores and wife's IQ scores. If you were interested in knowing husband's IQ score, *given* the wife's IQ score, say, then wives's IQ score would be explanatory variable and husband's IQ score would be response variable..

(e) Scatter diagrams drawn for quantitative data, not qualitative data because (circle one or more)
   i. qualitative data has no order,
   ii. distance between qualitative data points is not meaningful.

(f) Another ten individuals *sampled* gives **same** / **different** scatter plot. Data here is a **sample** / **population**. Data here is **observed** / **known**.

2. *Scatter Diagram: Grain Yield (tons) versus Distance From Water (feet).*

dist, x	0	10	20	30	45	50	70	80	100	120	140	160	170	190
yield, y	500	590	410	470	450	480	510	450	360	400	300	410	280	350

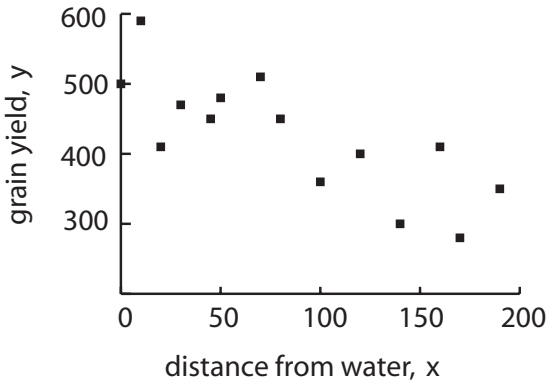

Figure 4.2 (Scatter Diagram, Grain Yield Versus Distance from Water)

(StatCrunch: Relabel var3 as distance and var4 as grain yield. Type data into two columns. Graphics, Scatter Plot, X variable: distance, Y variable: grain yield, Create Graph!)

(a) Scatter diagram has **pattern** / **no pattern (randomly scattered)** with (choose one) **positive** / **negative** association,
which is (choose one) **linear** / **nonlinear**, that is a
(choose one) **weak** / **moderate** / **strong** (non)linear relationship,
where grain yield is (choose one) **response** / **explanatory** variable.

(b) *Review.* Second random sample would be **same** / **different** scatter plot of (distance, yield) points. Any statistics calculated from second plot would be **same** / **different** from statistics calculated from first plot.

3. *Scatter Diagram: Pizza Sales ($1000s) versus Student Number (1000s).*

student number, x	2	6	8	8	12	16	20	20	22	26
pizza sales, y	58	105	88	118	117	137	157	169	149	202

(StatCrunch: Relabel var5 as number and var6 as pizza sales. Type data into two columns. Graphics, Scatter Plot, X variable: number, Y variable: pizza sales, Create Graph! Data, Save data, 4.1 three scatter plots.)
Scatter diagram has **pattern** / **no pattern (randomly scattered)** with
(choose one) **positive** / **negative** association,
which is (choose one) **linear** / **nonlinear**, that is a
(choose one) **weak** / **moderate** / **strong** (non)linear relationship,
where student number is (choose one) **response** / **explanatory** variable.

4. *More Scatter Diagrams*

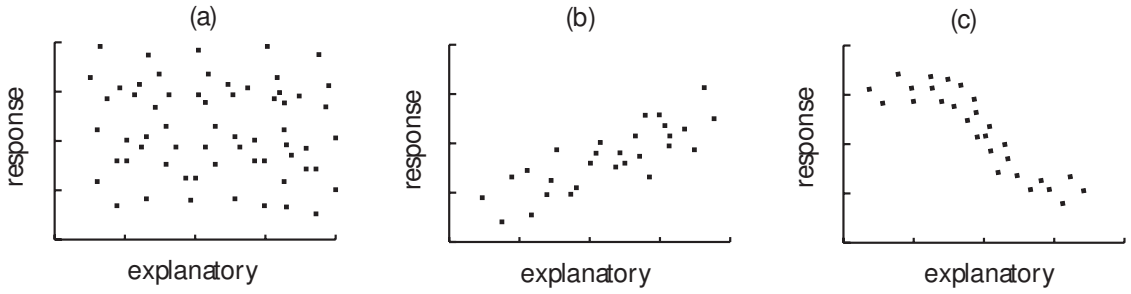

Figure 4.3 (More Scatter Diagrams)

Describe each scatter plot.

(a) Scatter diagram (a) has **pattern** / **no pattern (randomly scattered)**.

(b) Scatter diagram (b) has **pattern / no pattern (randomly scattered)**
with (choose one) **positive / negative** association,
which is (choose one) **linear / nonlinear**, that is a
(choose one) **weak / moderate / strong** (non)linear relationship.

(c) Scatter diagram (c) has **pattern / no pattern (randomly scattered)**
with (choose one) **positive / negative** association,
which is (choose one) **linear / nonlinear**, that is a
(choose one) **weak / moderate / strong** (non)linear relationship.

5. *Linear Correlation Coefficient: Using StatCrunch.*
Linear correlation coefficient statistic, r, measures *linearity* of scatter diagram.

$r = +1$	x and y perfectly positively linear
$r \geq 0.8$ or $r \leq -0.8$	x and y strongly linear
$0.5 \leq r \leq 0.8$ or $-0.8 \leq r \leq -0.5$	x and y moderately linear
$-0.5 \leq r \leq 0.5$, $r \neq 0$	x and y weakly linear
$r = 0$	x and y uncorrelated
$r = -1$	x and y perfectly negatively linear

(a) *Reading ability versus brightness*

brightness, x	1	2	3	4	5	6	7	8	9	10
reading ability, y	70	70	75	88	91	94	100	92	90	85

In this case, $r \approx$ (circle one) **0.704 / 0.723 / 0.734**.
(Stat, Summary Stats, Correlation, Select Columns: brightness, reading ability, then Calculate.)
So, association between reading ability and brightness is (circle one)
positive strong linear
negative moderate linear
positive moderate linear

(b) *Grain yield versus distance from water*

dist, x	0	10	20	30	45	50	70	80	100	120	140	160	170	190
yield, y	500	590	410	470	450	480	510	450	360	400	300	410	280	350

In this case, $r \approx$ (circle one) **-0.724 / -0.785 / -0.950**.
(StatCrunch: Stat, Summary Stats, Correlation, Select Columns: distance, grain yield, Calculate.)
So, association between grain yield and distance from water is (circle one)
positive strong linear
negative moderate linear
positive moderate linear

(c) *Annual pizza sales versus student number*

student number, x	2	6	8	8	12	16	20	20	22	26
pizza sales, y	58	105	88	118	117	137	157	169	149	202

In this case, $r \approx$ (circle one) **0.724** / **0.843** / **0.950**.

(Stat, Summary Stats, Correlation, Select Columns: number, pizza sales, Calculate. Data, Save data, 4.1 three scatter plots.)

So, association between pizza sales and student number is (circle one)
positive strong linear
negative moderate linear
positive moderate linear

6. *Linear correlation coefficient: understanding.*

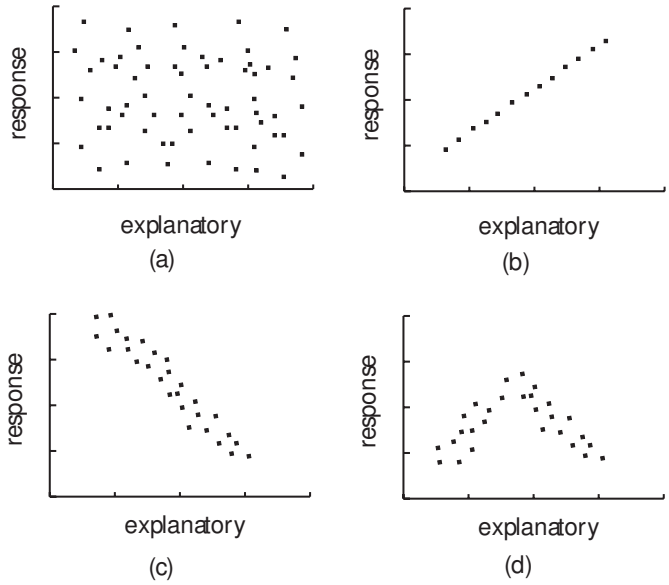

Figure 4.4 (Scatter Diagrams and Possible Correlation Coefficients)

Match correlation coefficients with scatter plots.

(a) scatter diagram (a): $\boldsymbol{r = -0.7}$ / $\boldsymbol{r = 0}$ / $\boldsymbol{r = 0.3}$
(b) scatter diagram (b): $\boldsymbol{r = -0.7}$ / $\boldsymbol{r = 0.1}$ / $\boldsymbol{r = 1}$
(c) scatter diagram (c): $\boldsymbol{r = -0.7}$ / $\boldsymbol{r = 0}$ / $\boldsymbol{r = 0.7}$
(d) scatter diagram (d): $\boldsymbol{r = -0.7}$ / $\boldsymbol{r = 0}$ / $\boldsymbol{r = 0.7}$

When $r \neq 0$, x and y are *linearly* related to one another. If $r = 0$, x and y are *nonlinearly* related to one another, which *often* means diagram (a) or sometimes means diagram (d) where positive and negative associated data points cancel one another out. Always show scatter diagram with correlation r.

7. *Linear Correlation Coefficient: Properties (Reading Ability Versus Brightness).*

brightness, x	1	2	3	4	5	6	7	8	9	10
reading ability, y	70	70	75	88	91	94	100	92	90	85

(a) As brightness increases, reading ability **increases** / **decreases** because $r \approx 0.704$ is positive.

(b) The more positive r is (the closer r is to 1), the (circle one)

i. more linear the scatter plot.

ii. steeper the slope of the scatter plot.

iii. larger the reading ability value.

iv. brighter the brightness.

(c) If 0.5 is added to *all* x values, 1 becomes 1.5, 2 becomes 2.5 and so on, r **changes from 0.704 to 0.892**.
remains the same, at 0.704.

(d) The r–value calculated after accidently reversing point (1,70) with point (70,1) **equals** / **does not equal** r value before reversing this point.

(e) **True** / **False** The r–value remains same whether or not brightness is measured in foot candles or lumens.

(f) Ability to read and brightness are mistakenly reversed:

ability to read, y	70	70	75	88	91	94	100	92	90	85
brightness, x	1	2	3	4	5	6	7	8	9	10

The r value (circle one) **remains unchanged** / **changes**.

(g) Compare original scatter diagram with one without outlier $(7, 130)$.

brightness, x	1	2	3	4	5	6	7	8	9	10
ability to read, y	70	70	75	88	91	94	100	92	90	85

brightness, x	1	2	3	4	5	6	**7**	8	9	10
ability to read, y	70	70	75	88	91	94	**130**	92	90	85

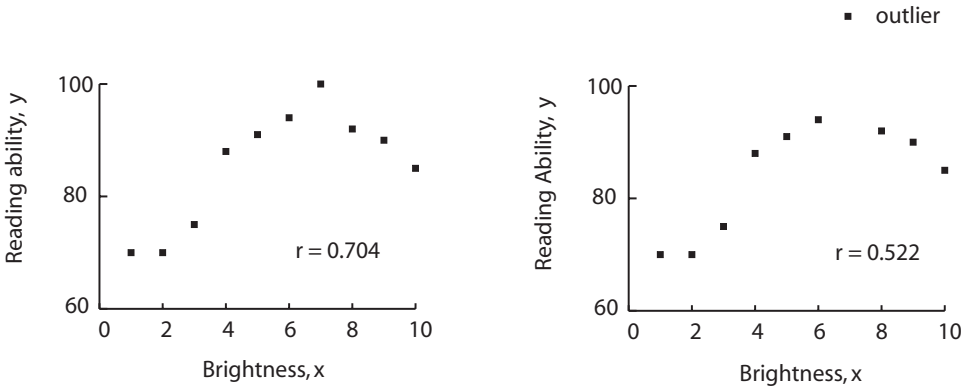

Figure 4.5 (Correlation and Outliers)

The correlation coefficient is (circle one) **resistant** / **sensitive** to outliers.

(h) Identify statistical items in example.

terms	reading/lighting example
(a) population	**(a)** all reading/brightness levels
(b) sample	**(b)** correlation of 10 reading/brightness levels, r
(c) statistic	**(c)** correlation of all reading/brightness levels, ρ
(d) parameter	**(d)** 10 reading/brightness levels

terms	(a)	(b)	(c)	(d)
reading/brightness example				

Notice *population* parameter for linear correlation coefficient is ρ.

(i) Brightness increase **causes** / **is associated with** reading ability increase.

8. *Linear correlation coefficient: correlation, not causation (chimpanzees).*
In a study of chimpanzees it was found there was a positive correlation between tallness and intelligence. Circle true or false:

True / **False** Taller chimpanzees were also more intelligent, on average.

True / **False** Intelligent chimpanzees were also taller, on average.

True / **False** The data show that tallness causes intelligence.

True / **False** The data show that intelligence causes tallness.

In general, although two variables may be highly correlated, this does not necessarily mean that an increase (or decrease) in one variable *causes* an increase or decrease in other variable. It may be that the chimpanzees were bred for both intelligence and tallness: breeding is a *lurking variable* which may explain the correlation between intelligence and tallness.

9. *Linear Correlation Coefficient: Formulas.*
Definitional formula:

$$r = \frac{\sum \left(\frac{\sum x_i - \bar{x}}{s_x} \right) \left(\frac{\sum y_i - \bar{y}}{s_y} \right)}{n-1}$$

Computational formula:

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{\left(\sum x_i \right)^2}{n} \right) \left(\sum y_i^2 - \frac{\left(\sum y_i \right)^2}{n} \right)}}$$

Computational formula generally easier to calculate, but less accurate than definitional formula.

(a) *Computational formula.*
If $\sum_{i=1}^n x_i = -13$, $\sum_{i=1}^n y_i = 12$, $\sum_{i=1}^n x_i^2 = 160$, $\sum_{i=1}^n y_i^2 = 930$, $\sum_{i=1}^n x_i y_i = -345$, and $n = 5$, then
$SS_{xy} = \sum_{i=1}^n x_i y_i - [\sum_{i=1}^n x_i \sum_{i=1}^n y_i / n] = -345 - [(-13)(12)/5] =$
(circle one) **−189 / −234 / −313.8**
$SS_x = \sum_{i=1}^n x_i^2 - [(\sum_{i=1}^n x_i)^2/n] = 160 - [(-13)^2/5] =$
(circle one) **110.2 / 126.2 / 231.3**
and $SS_y = \sum_{i=1}^n y_i^2 - [(\sum_{i=1}^n y_i)^2/n] = 930 - [(12)^2/5] =$
(circle one) **640.2 / 901.2 / 960.8**
and so $r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{-313.8}{\sqrt{(126.2)(901.2)}} \approx$
(circle one) **−0.560 / −0.621 / −0.93**

(b) *Definitional formula: circumference versus heights of trees.*

circumference, x	2.1	1.7	1.1	1.5	2.7
height, y	40	37	35	36	42

(StatCrunch: Relabel var7 as circumference and var8 as height. Type data into two columns. Data, Save data, 4.1 four scatter plots. Stat, Summary Stats, columns, Select Columns: circumference, height, then Next, Statistics: Sum, Other statistic: sum(x∧2), Calculate. Data, Compute Expression, sum(circumference*height), Compute.)
Since $n =$ **5 / 10 / 15**
$\sum_{i=1}^n x_i =$ (sum(x) for circumference) **9.1 / 10.3 / 11.4**
$\sum_{i=1}^n y_i =$ (sum(x) for height) **89 / 134 / 190**
$\sum_{i=1}^n x_i^2 =$ (sum(x∧2) for circumference) **18.05 / 20.34 / 21.34**
$\sum_{i=1}^n y_i^2 =$ (sum(x∧2) for height) **5189 / 6434 / 7254**
$\sum_{i=1}^n x_i y_i =$ (sum(circumference*height)) **352.8 / 634.1 / 745.4**
then $SS_{xy} = \sum_{i=1}^n x_i y_i - [\sum_{i=1}^n x_i \sum_{i=1}^n y_i / n] = 352.8 - [(9.1)(190)/5] =$
5 / 6 / 7
$SS_x = \sum_{i=1}^n x_i^2 - [(\sum_{i=1}^n x_i)^2/n] = 18.05 - [(9.1)^2/5] =$
1.321 / 1.488 / 2.233
and $SS_y = \sum_{i=1}^n y_i^2 - [(\sum_{i=1}^n y_i)^2/n] = 7254 - [(190)^2/5] =$
23 / 34 / 60
and so $r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{7}{\sqrt{(1.488)(34)}} \approx$ **0.560 / 0.621 / 0.984**
and $r^2 =$ (circle one) **0.314 / 0.723 / 0.968**.

10. *Linear correlation coefficient: more properties, r can be fooled or confused.*

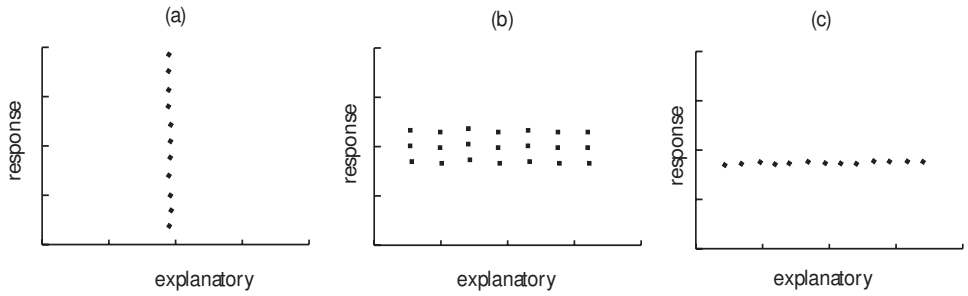

Figure 4.6 (Correlation can be fooled or confused)

It appears as though scatter diagrams (a) and (c) are perfectly linear, that $r = 1$. In fact, $r \neq 1$ in all three cases. Match r with scatter diagrams.

(a) $\boldsymbol{r = 0}$ / $\boldsymbol{r = \infty}$ / $\boldsymbol{r}$ **undefined** since $SS_x = 0$, $SS_y = 22.75$, $SS_{xy} = 0$.
Try vertical line of points, (3,1), (3,2), (3,5), (3,7); r is undefined.

(b) $\boldsymbol{r = 0}$ / $\boldsymbol{r = \infty}$ / $\boldsymbol{r}$ **undefined** since $SS_x = 45.5$, $SS_y = 2$, $SS_{xy} = 0$.
For two horizontal line of points, (1,3), (2,3), (5,3), (7,3) and (1,4), (2,4), (5,4), (7,4), $r = 0$.

(c) $\boldsymbol{r = 0}$ / $\boldsymbol{r = \infty}$ / $\boldsymbol{r}$ **undefined** since $SS_x = 22.75$, $SS_y = 0$, $SS_{xy} = 0$.
Try horizontal line of points, (1,3), (2,3), (5,3) and (7,3); r is undefined.

(d) **True / False** Roughly, SS_x measures how "wide" scatter is in x–direction, SS_y measures how "wide" scatter plot is in y–direction and SS_{xy} measures if slope of scatter is positive (increasing), negative (decreasing) or zero (perfectly vertical or horizontal scatter plot).

4.2 Least-Squares Regression

We fit a *least-squares regression* line,

$$\hat{y} = b_1 x + b_0$$

where b_1 is slope and b_0 is y–intercept, to paired quantitative data.

Exercise 4.2 (Least-Squares Regression)

1. *Least–Squares Line: Calculation, Prediction and Understanding.*

   (a) *Reading ability versus brightness.*
   Create scatter diagram, calculate least-squares regression line and superimpose line on scatter diagram.

brightness, x	1	2	3	4	5	6	7	8	9	10
reading ability, y	70	70	75	88	91	94	100	92	90	85

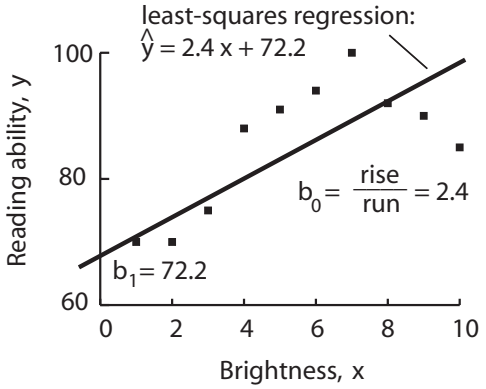

Figure 4.7 (Least–squares Line, reading ability versus brightness)

i. *Calculating least-squares regression line.* Choose *two.*
$\hat{y} = 72.2 + 2.418x$
$\hat{y} = 2.418x + 72.2$
$\hat{y} = 72.2x + 2.418$
$\hat{y} = 47.04x + 2.944$
(StatCrunch: Stat, Regression, Simple Linear, X-Variable: brightness, Y-Variable: reading ability, choose plot the fitted line, Calculate. Next (in Simple Linear Regression box) gives plotted regression.)

ii. *Slope and y-intercept of least-squares regression line,* $\hat{y} = 2.418x + 72.2$.
Slope is b_1 = (circle one) **72.2** / **2.418**.
Slope, $b_1 = 2.418$, means, on average, reading ability increases 2.418 units for an increase of *one* unit of brightness.

The *y–intercept* is b_0 = (circle one) **72.2** / **2.418**.
The y–intercept, $b_0 = 72.2$, means average reading ability is 72.2, if brightness is zero.

iii. *Prediction.*
At brightness $x = 6.5$, predicted reading ability is
$\hat{y} \approx 2.418x + 72.2 = 2.418(6.5) + 72.2 \approx$ **83.9** / **85.5** / **87.9**.
(StatCrunch: Click Options (in Simple Linear Regression box!), check Predict Y for X = 6.5, Compute!. Predicted reading ability given under Predicted values: 87.91818, in Simple Regression box.)

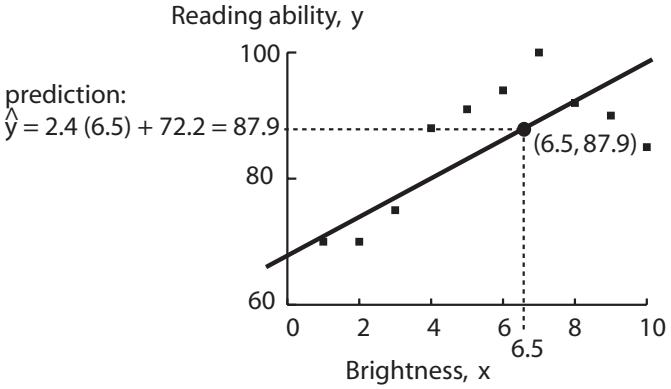

Figure 4.8 (Least–Squares Line: Prediction)

iv. *More Prediction.*
At brightness $x = 5.5$, $\hat{y} \approx 2.418(5.5) + 72.2 \approx$ **83.9** / **85.5** / **87.6**.
At brightness $x = 7.5$, $\hat{y} \approx 2.418(7.5) + 72.2 \approx$ **83.9** / **89.5** / **90.4**.
(StatCrunch: Click Options (in Simple Linear Regression box!), Edit, check Predict Y for X = 5.5, Calculate. Then repeat for 7.5.)

v. *Residual.*
At brightness $x = 7$, $\hat{y} \approx 2.418(7) + 72.2 \approx$ **87.9** / **89.1** / **120.6**.
Observed value, $y = 100$ compared to predicted $\hat{y} = 89.1$;
difference between two is *residual*:
$y - \hat{y} = 100 - 89.1 =$ (circle one) **9.2** / **10.9** / **12.6**.
(Click Options, Edit, choose Save residuals, Compute! Look in Residuals column in StatCrunch spreadsheet opposite brightness = 7.)

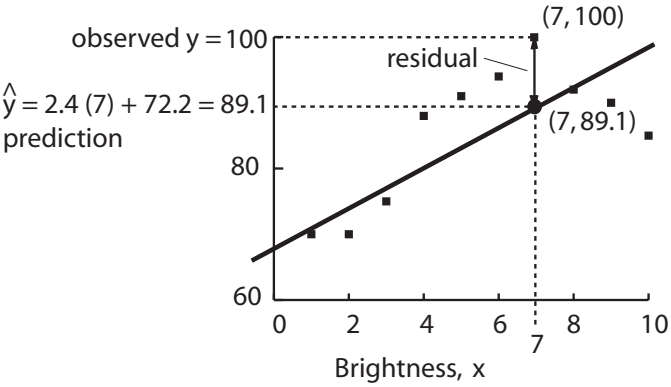

Figure 4.9 (Least–Squares Line: Residual)

Residual for $x = 7$ is vertical distance between observed (7,100) and predicted (7, 89.1) on least-squares regression line.

vi. *More Residuals.*
At brightness $x = 8$, $y - \hat{y} \approx 92 - 91.5 =$ **−0.5** / **0.5** / **1.5**.
At brightness $x = 3$, $y - \hat{y} \approx 75 - 79.5 =$ **−4.5** / **−3.5** / **−1.5**.
There are (circle one) **1** / **5** / **10** residuals on scatter diagram.
(StatCrunch: Look in Residuals column in StatCrunch spreadsheet, beside brightness $x = 8$ and

$x = 3$.)

(b) *Grain yield (tons) versus distance from water (feet)*

dist, x	0	10	20	30	45	50	70	80	100	120	140	160	170	190
yield, y	500	590	410	470	450	480	510	450	360	400	300	410	280	350

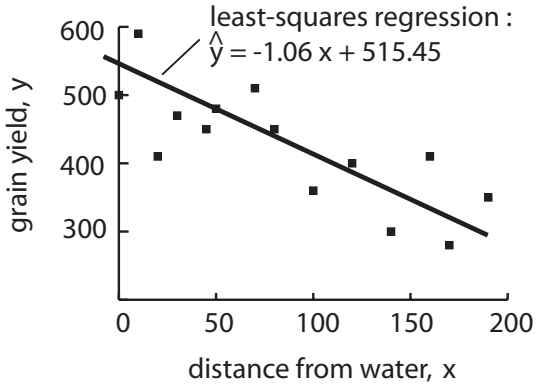

Figure 4.10 (Least–squares regression, grain yield versus distance)

i. The least–squares line is (circle one)
$\hat{y} = 515.45 - 1.56x$
$\hat{y} = 535.45x - 2.56x$
$\hat{y} = 515.45 - 1.06x$.
(StatCrunch: Edit, Columns, Delete, Residuals, Delete. Stat, Regression, Simple Linear, X-Variable: brightness, Y-Variable: reading ability, choose Save Residuals, choose plot the fitted line, Calculate.)

ii. *Slope and y-intercept.*
Slope is b_1 = (circle one) **515.45 / −1.06**.
Slope, $b_1 = -1.06$, means, on average, grain yield decreases 1.06 tons for an increase of one foot away from water.

The *y–intercept* is b_0 = (circle one) **515.45 / −1.06**.
The y–intercept, $b_0 = 515.45$, means average grain yield is 515.45 at water's edge.

iii. *Prediction.*
At distance $x = 100$,
$\hat{y} = -1.06x + 515.45 = -1.06(100) + 515.45 =$ **400 / 407.3 / 409.5**.
At distance $x = 165$,
$\hat{y} = -1.06x + 515.45 = -1.06(165) + 515.45 =$ **340.5 / 367.0 / 404.8**
(StatCrunch: Click Options (in Simple Linear Regression box!), choose Predict Y for X = 100, Calculate. Then, for X = 165.)

iv. *Residual.*
At distance $x = 100$,

$y - \hat{y} \approx 360 - 409.5 =$ **-49.5 / -36.5 / -25.5**.
At distance $x = 140$,
$y - \hat{y} \approx 300 - 367 =$ **-67 / -55 / -25**.
(StatCrunch: Look in Residuals column in data, beside distance $x = 100$ and $x = 140$.)

v. *Review.* Second random sample gives **same** / **different** scatter diagram. Statistics calculated from second plot **same** / **different** from statistics calculated from first plot. So, slope, b_1, and y–intercept, b_0, and predicted values, $\hat{y} = b_1 x + b_0$, all **statistics** / **parameters**.

vi. Identify statistical items in example.

terms	grain yield/water example
(a) population	**(a)** all (yield, distance) amounts
(b) sample	**(b)** $b_0, b_1, \hat{y}$
(c) statistics	**(c)** α, β, μ_x
(d) parameters	**(d)** 14 (yield, distance) amounts

terms	(a)	(b)	(c)	(d)
example				

(c) *Height versus circumference of trees.*

circumference, x	2.1	1.7	1.1	1.5	2.7
height, y	40	37	35	36	42

i. Least–squares line is (circle *two*!)
$\hat{y} = 29.438 + 4.704x$
$\hat{y} = 4.704x + 29.438$
$\hat{y} = 2.944 + 47.04x$.
(StatCrunch: Edit, Columns, Delete, Residuals, Delete. Stat, Regression, Simple Linear, X-Variable: circumference, Y-Variable: height, choose Save Residuals and Predicted values, Calculate.)

ii. *Residuals.* Fill in blanks.

circumference, x	2.1	1.7	1.1	1.5	2.7	total
observed height, y	40	37	35	36	42	190
predicted height, $\hat{y}$	39.3	37.4	34.6	_____	_____	190
residual, $y - \hat{y}$	0.7	-0.4	0.4	_____	_____	0
residual^2, $(y - \hat{y})^2$	0.5	0.2	0.2	_____	_____	1.1

Total residuals^2 measure how close points are to least-squares line.
(StatCrunch: Look in Fitted Values in StatCrunch spreadsheet for missing values in predicted height; look in Residuals in StatCrunch spreadsheet for missing values in residual $y - \hat{y}$. For missing values in residual^2, Data, Compute expression, Expression: Residuals^2, Compute. OR use SS Error (1.06) in ANOVA table given at bottom of simple linear regression results.)

2. *Least–squares regression line: formulas.*

Formula for least–squares regression line is

$$\hat{y} = b_1 x + b_0$$

where (definitional formulas)

$$b_1 = r \cdot \frac{s_y}{s_x}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

or where (computational formulas)

$$b_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

(a) *Computation formula.*
If $\sum_{i=1}^n x_i = -13$, $\sum_{i=1}^n y_i = 12$, $\sum_{i=1}^n x_i^2 = 160$,
$\sum_{i=1}^n x_i y_i = -345$, and $n = 5$, then
$SS_{xy} = \sum_{i=1}^n x_i y_i - [\sum_{i=1}^n x_i \sum_{i=1}^n y_i / n] = -345 - [(-13)(12)/5] =$
(circle one) **-189 / -234 / -313.8**
$SS_x = \sum_{i=1}^n x_i^2 - [(\sum_{i=1}^n x_i)^2 / n] = 160 - [(-13)^2/5] =$
(circle one) **110.2 / 126.2 / 231.3**
and so $b_1 = \frac{SS_{xy}}{SS_x} = \frac{-313.8}{126.2} =$ (circle one) **-0.19 / -1.34 / -2.49**
and $b_0 = \bar{y} - b\bar{x} = \frac{12}{5} - (-2.49)\frac{-13}{5} =$ (circle one) **-1.5 / -2.3 / -4.1**
and so $\hat{y} = b_1 x + b_0 =$
(circle one) $\boldsymbol{\hat{y} = 1.5x - 1.5}$ / $\boldsymbol{\hat{y} = -1.5x - 1.5}$ / $\boldsymbol{\hat{y} = -2.49x - 4.1}$.

(b) *Height versus circumference of trees: definitional formula.*

circumference, x	2.1	1.7	1.1	1.5	2.7
height, y	40	37	35	36	42

(StatCrunch: Edit, Columns, Delete, Residuals, Fitted Value, Residuals^2, Delete. Stat, Summary Stats, Columns, circumference, height, Calculate gives $\bar{x}, \bar{y}$ (Mean) and s_x, s_y (Std. Dev.). Stat, Summary Stats, Correlation, circumference, height, Calculate gives r.)
So $b_1 = r \cdot \frac{s_y}{s_x} \approx 0.9841 \cdot \frac{2.9155}{0.6099} \approx$ (circle one) **3.704 / 4.704 / 5.704**
and $b_0 = \bar{y} - b_1 \bar{x} = 38 - 4.704(1.82) =$ (circle one) **28.44 / 29.44 / 30.44**
and so least–squares line $\hat{y} = b_1 x + b_0$ is (circle one)
$\boldsymbol{\hat{y} = 4.704x + 29.438}$
$\boldsymbol{\hat{y} = 29.438x + 4.704}$
$\boldsymbol{\hat{y} = 47.04x + 2.944}$.

4.3 Diagnostics on the Least-Squares Regression

Various diagnostic analyzes are described which assess "fit" of least-squares line to data, for example, to detect

- any patterns, other than linearity, in data,
- whether variance of residuals against exploratory variable is constant or not,
- there are outliers,
- influential points, which, when removed, change slope or y-intercept a lot.

Diagnostic analyzes include scatter diagrams, residual plots and boxplots of residuals. Coefficient of determination, R^2, a measure of proportion of scatter explained by least-squares regression, is also discussed. This analysis is *exploratory*; we will look at significance of diagnostic analysis later.

Exercise 4.3 (Diagnostics on the Least-Squares Regression)

1. *Diagnostic analyzes: reading ability versus brightness.* Consider table of residuals, scatter diagram and residual plot.

brightness, x	1	2	3	4	5	6	7	8	9	10
ability to read, y	70	70	75	88	91	94	100	92	90	85
predicted, $\hat{y}$	74.6	77.0	79.5	81.9	84.3	86.7	89.1	91.5	94.0	96.4
residual, $y - \hat{y}$	-4.6	-7.0	-4.5	6.1	6.7	7.3	10.9	0.5	-4.0	-8.6

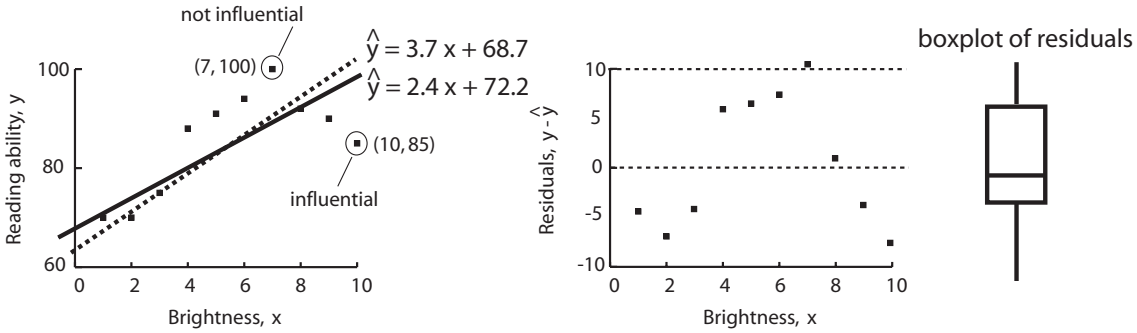

Figure 4.11 (Scatter diagram and residual plot, reading ability vs brightness)

(StatCrunch: Stat, Regression, Simple Linear, X-Variable: brightness, Y-Variable: reading ability, Next, Next, check Save Residuals, Next, check Plot the fitted line, check Residuals vs. X-values, Calculate gives Simple linear regression results, Next gives plotted regression, Next gives Residual plot. Graphics, Boxplot, Residuals, Next, check Use fences to identify outliers, Create Graph! gives boxplot.)

(a) *Pattern?*
According to either scatter diagram or residual plot,
there (choose one) **is a** / **is no** pattern (around line): points are curved.

(b) *Constant variance?*
According to residual plot, residuals vary -10 and 10 over entire range of brightness; that is, data variance is (choose one) **constant** / **variable**.

(c) *Outliers?*
According to boxplot of residuals, there **are** / **are no** outliers
(StatCrunch: No outliers "•"s in boxplot.)

(d) *Influential points?*
least-squares line is $\hat{y} = 2.418x + 72.2$, $r = 0.704$
(StatCrunch: Stat, Regression, Simple Linear, X-Variable: brightness, Y-Variable: reading ability, Calculate.)

point $(x, y) = (7, 100)$ influential?
removing this point, $\hat{y} = 2.192x + 72.2$, $r = 0.721$
so $(7, 100)$ **is** / **is not** influential since b_0, b_1, r do not change much
(StatCrunch: Click Row 7, Edit, Rows, Delete, Delete Rows! Stat, Regression, Simple Linear, X-Variable: brightness, Y-Variable: reading ability, Calculate.)

point $(x, y) = (10, 85)$ influential?
removing this point, $\hat{y} = 3.367x + 68.7$, $r = 0.836$
so $(10, 85)$ **is** / **is not** influential since all three b_0, b_1, r change a lot
(StatCrunch: Edit, Undo Delete Rows, Unclick Row 7 (7,100), Click Row 10 (10, 85), Edit, Rows, Delete, Delete Rows! Stat, Regression, Simple Linear, X-Variable: brightness, Y-Variable: reading ability, Calculate.)

(e) *Coefficient of determination, R^2.*
Since $R^2 =$ (choose one) **0.496** / **0.523** / **0.539**, least-squares line explains 49.6% of variability in reading ability.
(StatCrunch: Edit, Undo Delete Rows, Unclick Row 10 (10, 85), Stat, Regression, Simple Linear, X-Variable: brightness, Y-Variable: reading ability, Calculate. R-sq gives R^2.)

2. *Diagnostic analyzes: grain yield versus distance from water.* Consider scatter diagram and residual plot. Do not use StatCrunch, use plots and regressions given below instead.

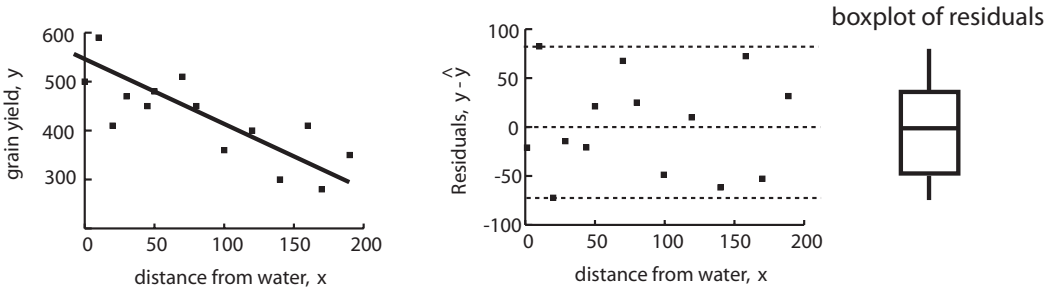

Figure 4.12 (Scatter diagram and residual plot, reading ability versus brightness)

(a) *Pattern?*
According to either scatter diagram or residual plot,
there (choose one) **is a** / **is no** pattern (around line).

(b) *Constant variance?*
According to residual plot, residuals vary -85 and 85 over entire range of distances; that is, data variance is (choose one) **constant** / **variable**.

(c) *Outliers?*
According to boxplot of residuals, there **are** / **are no** outliers.

(d) *Influential points?*
least-squares line is $\hat{y} = -1.06x + 515.45$, $r = -0.785$

point $(x, y) = (20, 140)$ influential?
removing this point, $\hat{y} = -1.18x + 533.10$, $r = -0.839$
so $(20, 140)$ **is** / **is not** influential since b_0, b_1, r do not change much

point $(x, y) = (190, 350)$ influential?
removing this point, $\hat{y} = -1.16x + 520.58$, $r = -0.781$
so $(190, 350)$ **is** / **is not** influential since b_0, b_1, r do not change much

(e) *Coefficient of determination R^2.*
Since $R^2 =$ (choose one) **0.496** / **−0.616** / **0.616**, least-squares line explains 61.0% of variability in grain yield.
(StatCrunch: Stat, Regression, Simple Linear, X-Variable: distance, Y-Variable: grain yield, Calculate. R-sq gives R^2.)

3. *Understanding coefficient of determination, R^2.*
True / **False**

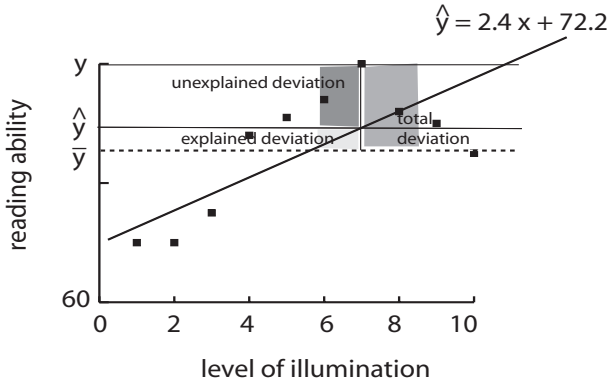

Figure 4.13 (Understanding Coefficient of Determination)

If the average of the y variable, $\bar{y}$, is a kind of baseline and since

$$\underbrace{(y-\bar{y})}_{\text{total deviation}} = \underbrace{(\hat{y}-\bar{y})}_{\text{explained deviation}} + \underbrace{(y-\hat{y})}_{\text{unexplained deviation}}$$

then taking sum of squares over all data points

$$\underbrace{\sum(y-\bar{y})^2}_{\text{total variation}} = \underbrace{\sum(\hat{y}-\bar{y})^2}_{\text{explained variation}} + \underbrace{\sum(y-\hat{y})^2}_{\text{unexplained variation}}$$

and so coefficient of determination is a measure of proportion of scatter diagram explained by least-squares line.

$$R^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} = \frac{\text{explained variation}}{\text{total variation}}$$

4.4 Contingency Tables and Association

We look at contingency tables to determine the 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.

Exercise 4.4 (Contingency Tables and Association)

1. *Contingency 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) *Marginal row (father) frequency, marginal column (son) frequency.* Fill in blanks.

	son attended college	son did not attend college	row total
father attended college	18	7	_____
father did not attend college	22	33	_____

	son attended college	son did not attend college
father attended college	18	7
father did not attend college	22	33
column totals	_____	_____

(b) *Relative marginal row (father) frequency, marginal column (son) frequency.* Complete following relative marginal frequency contingency table.

	son attended college	son did not attend college	
father attended college	18	7	$\frac{25}{80} = 0.3125$
father did not attend college	22	33	$\frac{55}{80} =$ _____
	$\frac{40}{80} = 0.5$	$\frac{40}{80} =$ _____	$\frac{25}{80} = 1$

(c) *Son's attendance conditional on father attendance distribution.* 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	$\frac{22}{55} =$ _____	$\frac{33}{55} = 0.6$	$\frac{55}{55} =$ _____
	$\frac{40}{80} = 0.5$	$\frac{40}{80} =$ _____	$\frac{80}{80} = 1$

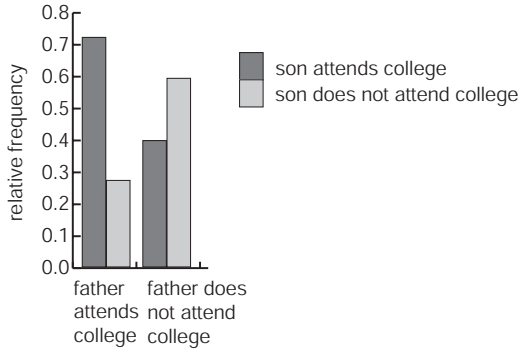

Figure 4.14 (Bar graph: son conditional on father.)

(Blank data table. Relabel var1 father, var2 son attends, var3 son does not attend, type father attends and father does not attend under father column, type 0.72 and 0.4 under son attends and 0.28 and 0.6 under son does not attend. Chart, Columns, choose son attends, son does not attend, Row labels in: father, Plot: vertical bars (split).)

Response variable is **father's attendance / son's attendance** because calculation of son's attendance *conditional on* father's attendance.

There is **an** / **no** association: son attends college more likely if father attends college, less likely if father does not attend college.

2. *Contingency table: 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 conditional on drug distribution.*
Complete conditional table.

flu symptoms →	reduced	not reduced	
drug	$\frac{100}{150} =$ _____	$\frac{50}{150} = 0.33$	$\frac{150}{150} =$ _____
no drug	$\frac{200}{300} =$ _____	$\frac{100}{300} = 0.33$	$\frac{300}{300} = 1$
	$\frac{300}{450} =$ _____	$\frac{150}{450} = 0.33$	$\frac{450}{450} = 1$

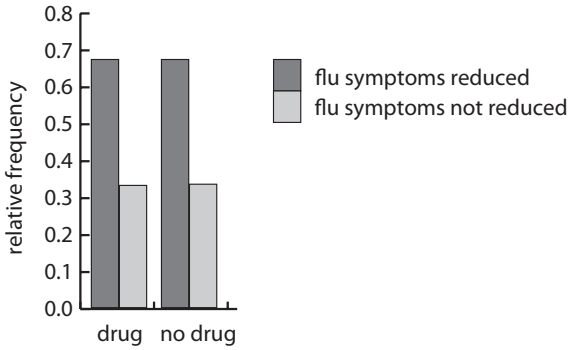

Figure 4.15 (Bar graph: flu symptoms conditional on drug.)

Response variable is (choose one) **flu symptoms** / **drug** because
flu symptom counts is *divided* by drug count row totals.
There is (choose one) **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 tabulate effect of drug on males and, separate from this, tabulate 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	$\frac{120}{150} = 0.8$	$\frac{30}{150} = 0.2$	$\frac{150}{150} = 1$

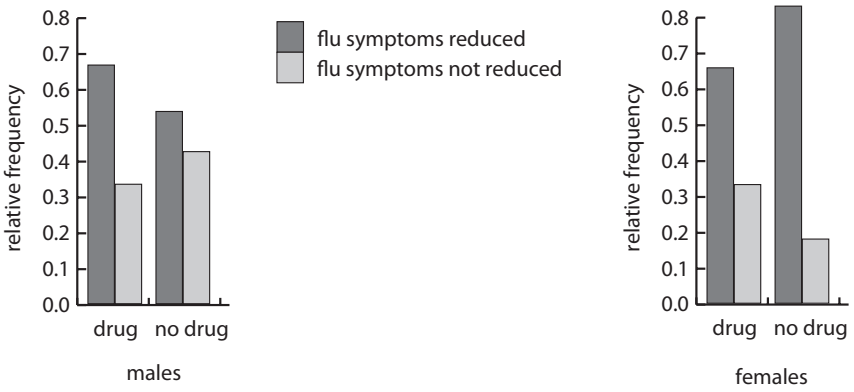

Figure 4.16 (Bar graph: flu conditional on drug, males/females.)

There is (choose one) **an** / **no** association for *males*:
more likely flu symptoms reduced when taking drug than not taking drug.
There is (choose one) **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.