Chapter 14

Describing Relationships: Scatterplots and Correlation

We look at scatterplots and linear correlation for paired (bivariate) quantitative data sets. Scatterplot is graph of paired *sampled* data and linear correlation is a measure of linearity of scatterplot. Formula for linear correlation coefficient is

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

Exercise 14.1 (Describing Relationships: Scatterplots and Correlation)

1. *Scatterplot: 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

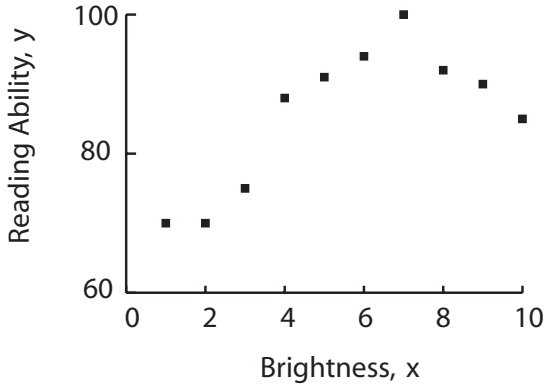

Figure 14.1 (Scatterplot, Reading Ability Versus Brightness)

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 on y–axis and "brightness" is **response** / **explanatory** variable on x–axis because reading ability depends on brightness, not the reverse.

Sometimes the response variable and explanatory variable are not distinguished from one another. 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. *Scatterplot: 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

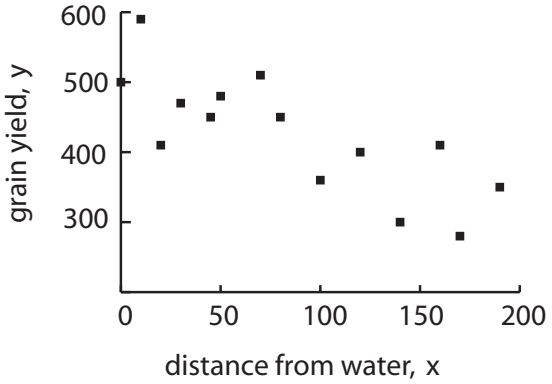

Figure 14.2 (Scatterplot, Grain Yield Versus Distance from Water)

(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. *Scatterplot: 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

Use data in table to complete following scatterplot.

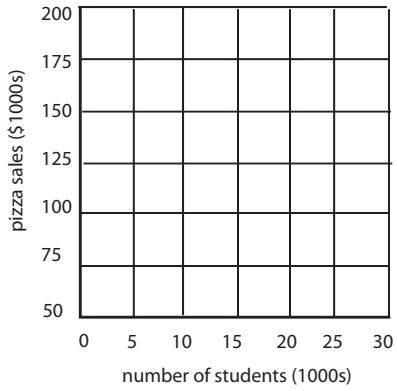

Figure 14.3 (Scatterplot of pizza sales versus student number)

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 Scatterplots*

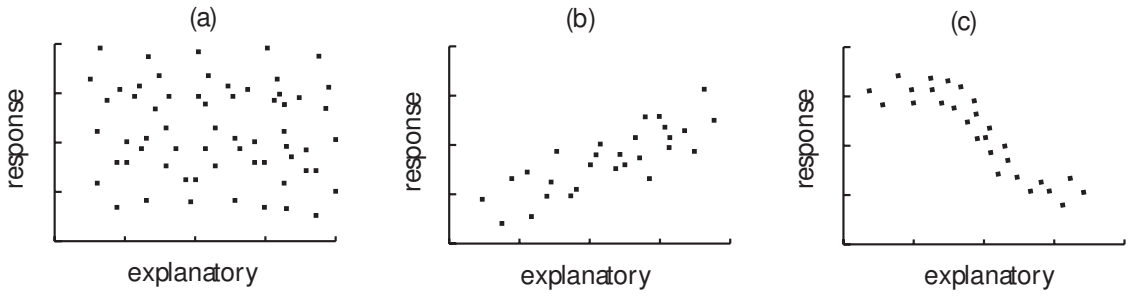

Figure 14.4 (More Scatterplots)

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.*
Linear correlation coefficient statistic, r, measures *linearity* of scatterplot.

$r = +1$	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

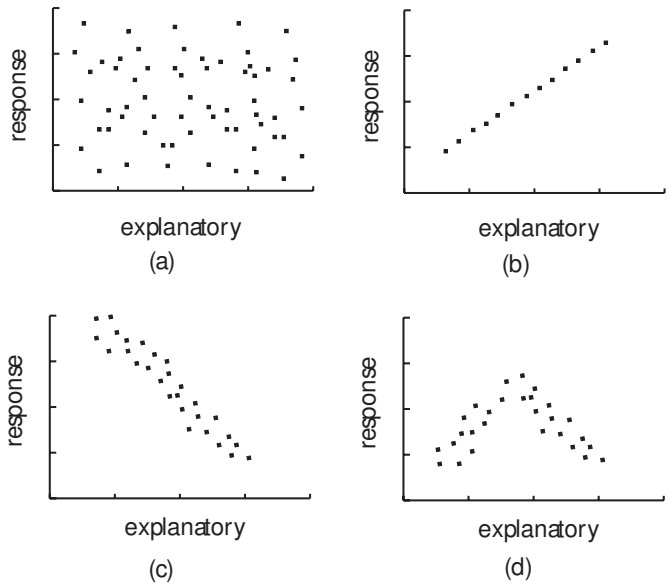

Figure 14.5 (Scatterplots and Possible Correlation Coefficients)

Match correlation coefficients with scatter plots.

(a) scatterplot (a): $\boldsymbol{r = -0.7}$ / $\boldsymbol{r = 0}$ / $\boldsymbol{r = 0.3}$

(b) scatterplot (b): $\boldsymbol{r = -0.7}$ / $\boldsymbol{r = 0.1}$ / $\boldsymbol{r = 1}$

(c) scatterplot (c): $\boldsymbol{r = -0.7}$ / $\boldsymbol{r = 0}$ / $\boldsymbol{r = 0.7}$

(d) scatterplot (d): $\boldsymbol{r = -0.7}$ / $\boldsymbol{r = 0}$ / $\boldsymbol{r = 0.7}$

(e) Correlation coefficient r (choose one) **has units** / **is unit-less**.

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 scatterplot with correlation r.

6. *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 scatterplot with or 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

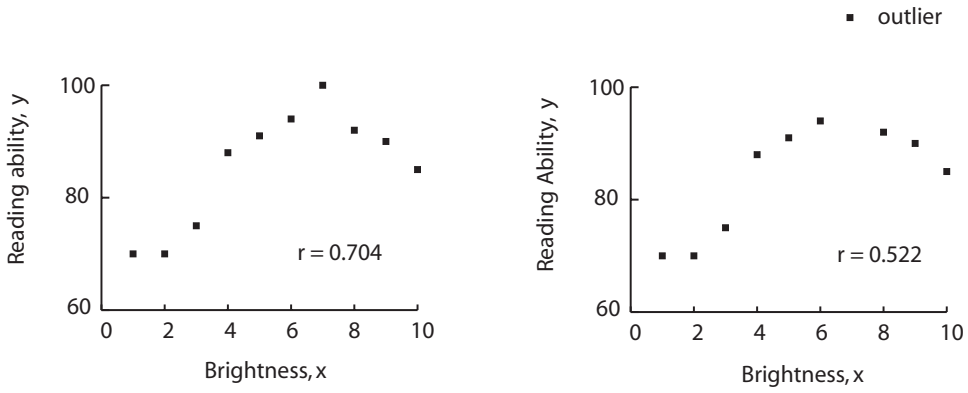

Figure 14.6 (Correlation and Outliers)

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.

7. *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.

8. *Linear correlation coefficient: formula.*
Formula for linear correlation coefficient is

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

(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

where

mean brightness $\bar{x} = 5.5$ SD brightness $s_x \approx 3.03$
mean reading ability $\bar{y} = 85.5$ SD reading ability $s_y \approx 10.39$

and where

x	standard score $\frac{x-\bar{x}}{s_x}$	y	standard score $\frac{y-\bar{y}}{s_y}$	product $\frac{x-\bar{x}}{s_x} \times \frac{y-\bar{y}}{s_y}$
1	$\frac{1-5.5}{3.03} \approx -1.49$	70	$\frac{70-85.5}{10.39} \approx -1.49$	2.22
2	$\frac{2-5.5}{3.03} \approx -1.16$	70	$\frac{70-85.5}{10.39} \approx -1.49$	1.72
3	$\frac{3-5.5}{3.03} \approx -0.83$	75	$\frac{75-85.5}{10.39} \approx -1.01$	0.83
4	$\frac{4-5.5}{3.03} \approx -0.50$	88	$\frac{88-85.5}{10.39} \approx 0.24$	-0.12
5	$\frac{5-5.5}{3.03} \approx -0.17$	91	$\frac{91-85.5}{10.39} \approx 0.53$	-0.09
6	$\frac{6-5.5}{3.03} \approx 0.17$	94	$\frac{94-85.5}{10.39} \approx 0.82$	0.14
7	$\frac{7-5.5}{3.03} \approx 0.50$	100	$\frac{100-85.5}{10.39} \approx 1.39$	0.69
8	$\frac{8-5.5}{3.03} \approx 0.83$	92	$\frac{92-85.5}{10.39} \approx 0.63$	0.52
9	$\frac{9-5.5}{3.03} \approx 1.16$	90	$\frac{90-85.5}{10.39} \approx 0.43$	0.50
10	$\frac{10-5.5}{3.03} \approx 1.49$	95	$\frac{95-85.5}{10.39} \approx -0.05$	-0.07

So

$$\begin{aligned} r &= \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right) \\ &= \frac{1}{10-1}(2.22+1.72+0.83-0.12-0.09+0.14+0.69+0.52+0.50-0.07) \end{aligned}$$

and so $r \approx$ (circle one) **-0.7 / 0 / 0.7**,
and also association between reading ability and brightness is (circle one)
positive strong linear
negative moderate linear
positive moderate linear

(b) *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

where

mean number of students $\bar{x} = 14$ SD number of students $s_x \approx 3.94$
mean pizza sales $\bar{y} = 130$ SD pizza sales $s_y \approx 41.81$

and where

x	standard score $\frac{x-\bar{x}}{s_x}$	y	standard score $\frac{y-y}{s_y}$	product $\frac{x-\bar{x}}{s_x} \times \frac{y-y}{s_y}$
2	$\frac{2-14}{3.94} \approx -1.51$	58	$\frac{58-130}{41.81} \approx -1.72$	2.60
6	$\frac{6-14}{3.94} \approx -1.01$	105	$\frac{105-130}{41.81} \approx -0.60$	0.60
8	$\frac{8-14}{3.94} \approx -0.76$	88	$\frac{88-130}{41.81} \approx -1.01$	0.76
8	$\frac{8-14}{3.94} \approx -0.76$	118	$\frac{118-130}{41.81} \approx -0.29$	0.22
12	$\frac{12-14}{3.94} \approx -0.25$	117	$\frac{117-130}{41.81} \approx -0.31$	0.08
16	$\frac{16-14}{3.94} \approx 0.25$	137	$\frac{137-130}{41.81} \approx 0.17$	0.04
20	$\frac{20-14}{3.94} \approx 0.76$	157	$\frac{157-130}{41.81} \approx 0.65$	0.49
20	$\frac{20-14}{3.94} \approx 0.76$	169	$\frac{169-130}{41.81} \approx 0.93$	0.70
22	$\frac{22-14}{3.94} \approx 1.01$	149	$\frac{149-130}{41.81} \approx 0.45$	0.46
26	$\frac{26-14}{3.94} \approx 1.51$	202	$\frac{202-130}{41.81} \approx 1.72$	2.60

So

$$\begin{aligned} r &= \frac{1}{n-1} \sum \left(\frac{x-\bar{x}}{s_x} \right) \left(\frac{y-\bar{y}}{s_y} \right) \\ &= \frac{1}{10-1} (2.60 + 0.60 + 0.76 + 0.22 + 0.08 + 0.04 + 0.49 + 0.70 + 0.46 + 2.60) \end{aligned}$$

and so $r \approx$ (circle one) **0.724** / **0.843** / **0.950**.
So, association between pizza sales and student number is (circle one)
positive strong linear
negative moderate linear
positive moderate linear

(c) *A last example*

x	2	6	8	8
y	58	105	88	118

where

$$\text{mean } \bar{x} = 6 \qquad \text{SD } s_x \approx 2.83$$
$$\text{mean } \bar{y} = 92.25 \quad \text{SD } s_y \approx 25.93$$

and where

x	standard score $\frac{x-\bar{x}}{s_x}$	y	standard score $\frac{y-\bar{y}}{s_y}$	product $\frac{x-\bar{x}}{s_x} \times \frac{y-\bar{y}}{s_y}$
2	$\frac{2-6}{2.83} \approx -1.41$	58	$\frac{58-92.25}{25.93} \approx -1.32$	$-1.41 \times -1.32 \approx 1.86$
6	$\frac{6-6}{2.83} \approx$	105	$\frac{105-92.25}{25.93} \approx$	
8	$\frac{8-6}{2.83} \approx$	88	$\frac{88-92.25}{25.93} \approx$	
8	$\frac{8-6}{2.83} \approx$	118	$\frac{118-92.25}{25.93} \approx$	

So

$$\begin{aligned} r &= \frac{1}{n-1} \sum \left(\frac{x-\bar{x}}{s_x} \right) \left(\frac{y-\bar{y}}{s_y} \right) \\ &= \frac{1}{4-1} (1.86 + 0 - 0.11 + 0.70) \end{aligned}$$

and so $r \approx$ (circle one) **0.82** / **0.87** / **0.95**.
So, association between x and y is (circle one)
positive strong linear
negative moderate linear
positive moderate linear