CHAPTER

4

CORE

Displaying and describing relationships between two variables

- What are the statistical tools for displaying and describing relationships between
  - two categorical variables?
  - a numerical and a categorical variable?
  - two numerical variables?
- What is a causal relationship?

So far we have looked at statistical techniques for displaying and describing the distributions of single variables. This is termed **univariate** or **single-variable** data analysis. In this chapter we look at statistical techniques displaying and describing the relationship between two variables. This is termed **bivariate** or **two-variable** data analysis.

4.1 Investigating the relationship between two categorical variables

The two-way frequency table

It has been suggested that males and females have different attitudes to gun control, that is, that attitude to gun control depends on the sex of the person. How might we investigate the relationship between attitude to gun control and sex?

The first thing to note is that these two variables, *Attitude to gun control* ('For' or 'Against') and *Sex* ('Male' or 'Female'), are both categorical variables. Categorical data is usually presented in the form of a frequency table. For example, if we interview a sample of 100 people we might find that there are 58 males and 42 females. We can present this result in a frequency table, see Table 4.1.

Cambridge University Press • Uncorrected Sample pages • 978-0-521-61328-6 • 2008 © Jones, Evans, Lipson
TI-Nspire & Casio ClassPad material in collaboration with Brown and McMenamin

Back to Menu >>>

Table 4.1

Sex	*Frequency*
Male	58
Female	42
Total	100

Table 4.2

Attitude	*Frequency*
For	62
Against	38
Total	100

Similarly, if we recorded their attitude to gun control, we might find 62 'For' and 38 'Against' gun control. Again we could present these results in a table, see Table 4.2.

From Table 4.1, we can see that there were more men than women in our sample. From Table 4.2, we see that more people in the sample were 'For' gun control than 'Against' gun control. However, we cannot tell from the information contained in the tables whether attitude to gun control depends on the sex of the person. To do this we need to form a **two-way frequency table**, as shown in Table 4.3.

The process of forming a two-way frequency table is called **crosstabulation**. In Table 4.3, we have crosstabulated the variables *Attitude to gun control* with *Sex*.

Table 4.3

	Sex		
Attitude	Male	Female	Total
For	32	30	62
Against	26	12	38
Total	58	42	100

Dependent and independent variables in tabulated data

When studying relationships between variables, it is sometimes clear that one of the variables might depend on the other, but not the other way around. For example, a person's attitude to gun control might depend on their sex, but not the other way around. In such situations, we call the variable that depends on the other (*Attitude to gun control*), the **dependent variable** (**DV**) and the variable it depends on (*Sex*) the **independent variable** (**IV**).

In two-way frequency tables, it is conventional to let the categories of the **dependent variable** define the **rows** of the table and the categories of the **independent variable** define the **columns** of the table. The convention was followed when setting up a table to investigate the relationship between *Attitude to gun control* (the DV) and *Sex* (the IV). See Table 4.4.

Table 4.4

IV

		Sex			
	Attitude	Male	Female	Total	
DV	For	32	30	62	Row sum
	Against	26	12	38	Row sum
	Total	58	42	100	
		Column sum	Column sum		

Reading a two-way frequency table

In a two-way frequency table, the regions shaded blue in Table 4.4 are called the **margins** of the table.

- The numbers in the right margin are called **row sums**, for example, $62 = 32 + 30$.

- The numbers in the bottom margin are called **column sums**, for example, $58 = 32 + 26$.
- The number in the right hand corner is called the **grand sum**. If the table has been constructed correctly, both the row sums and column sums should add up to 100, the total number of people.

The regions in the table shaded purple are called the **cells** of the table. It is the numbers in these cells that we look at when investigating the relationship between the two variables.

In Table 4.4, there are **four** cells. These cells represent the four categories of people revealed by the survey, namely, 'males who are for gun control', 'males who are against gun control', 'females who are for gun control' and 'females who are against gun control'.

Thus we see that, for example:

- 32 males are for gun control
- 30 females are for gun control

	Male	Female
For	32	30
Against	26	12

This information tells us that more men are in favour of gun control than women. But is this just due to the fact that there were more men in the sample, or are men really more in favour of gun control than women? To help us answer this question we turn our table entries into percentages.

Percentaging a two-way frequency table

There are several different ways we can percentage a two-way frequency table, each of which will give us different information. To answer our question, we need **column percentages**. These will give us the percentage of males and females for and against gun control.

Column percentages are determined by dividing each of the cell frequencies (the numbers in the purple region) by the column totals.

Thus we find, the percentage of:

- males who are for gun control is: $\frac{32}{58} \times 100 = 55.2\%$
- males who are against gun control is: $\frac{26}{58} \times 100 = 44.8\%$
- females who are for gun control is: $\frac{30}{42} \times 100 = 71.4\%$
- females who are against gun control is: $\frac{12}{42} \times 100 = 28.6\%$

Note: Unless small percentages are involved, it is usual to round percentages to one decimal place in tables.

Entering these percentages in the appropriate places and totalling the columns gives the percentage two-way frequency shown in Table 4.5.

Table 4.5

	Sex	
Attitude	Male	Female
For	55.2%	71.4%
Against	44.8%	28.6%
Total	100.0%	100.0%

Percentaging the table enables us to compare the attitudes of males and females on an equal footing. From the table we see that 55.2% of males in the sample were for gun control compared to 71.4% of the females. This means that the females in the sample were more supportive of gun control than the males. This reverses what the frequencies told us. It is easy to be misled if you just compare frequencies in a two-way frequency table.

Using percentages to identify relationships between variables

The fact that the percentage of 'Males for gun control' differs from the percentage of 'Females for gun control' indicates that a person's attitude to gun control **depends** on their sex. Thus we can say that the variables *Attitude to gun control* and *Sex* are **related** or **associated** (go together). If *Attitude to gun control* and *Sex* were **not related**, we would expect roughly equal percentages of males and females to be 'For' gun control.

We could have also arrived at this conclusion by focusing our attention on the percentages 'against' gun control. We might report our findings as follows.

> **Report**
> From Table 4.5 we see that a higher percentage of females were for gun control than males, 71.4% to 55.2%. This indicates that a person's attitude to gun control is related to their sex.

Note: Finding a single row in the two-way frequency distribution in which percentages are clearly different is sufficient to identify a relationship between the variables.

We will now consider a two-way frequency table which shows no evidence of a relationship between the variables *Attitude to mobile phones in cinemas* and *Sex*.

Table 4.6 shows the distribution of the responses of the same group of people to the question, 'Do you support the banning of mobile phones in cinemas?'

For this data, we might report our findings as follows.

Table 4.6

	Sex	
Mobile banned	Male	Female
Yes	87.9%	85.8%
No	12.1%	14.2%
Total	100.0%	100.0%

> **Report**
> From Table 4.6 we see that the percentage of males and females in support of banning mobile phones in cinemas was similar, 87.9% to 85.8%. This indicates that a person's support for banning mobile phones in cinemas was not related to their sex.

Exercise 4A

1 Complete Tables 1 and 2 by filling in the missing information. Where percentages are required, calculate column percentages.

Table 1

	Age		
Change	Young	Old	*Total*
Yes	23	15	
No	22		
Total	45	85	

Table 2

	Age	
Change	Young(%)	Old(%)
Yes		
No		82.4
Total		100.0

Back to Menu >>>

2 The following pairs of variables are related. Which is likely to be the dependent variable?

a Participates in regular exercise and age
b Level of education and salary level
c Comfort level and temperature
d Time of year and incidence of hay fever
e Age group and musical taste?
f AFL team supported and State of residence

3 A group of 100 people were asked about their attitude to Sunday racing with the following results.

	Sex	
Attitude	Male	Female
For	25	30
Against	20	25
Total	45	55

a How many:
 i people were surveyed?
 ii males were 'Against' Sunday racing?
 iii females were in the survey?
 iv females were 'For' Sunday racing?
 v people in the survey were 'For' Sunday racing?

b Percentage the table by forming column percentages.

c Do the percentages suggest that a person's attitude to Sunday racing is related to their sex? Write a brief report quoting appropriate percentages.

4 A survey was conducted on 242 university students. As part of this survey, data was collected on the students' enrolment status (full-time, part-time) and their drinking behaviour (drinks alcohol; yes, does not drink alcohol; no).

a It is expected that enrolment status and drinking behaviour are related. Which of the two variables would be the dependent variable?

b For analysis purposes, the data was organised into a two-way frequency table as follows:

	Enrolment status		
Drinks alcohol	Full-time	Part-time	*Total*
Yes	124	72	196
No	30	16	46
Total	154	88	242

How many of the students:
i drank alcohol? **ii** were part-time? **iii** were full-time and drank alcohol?

c Percentage the table by calculating column percentages.

d Does the data support the contention that there is a relationship between drinking behaviour and enrolment status? Write a brief report quoting appropriate percentages.

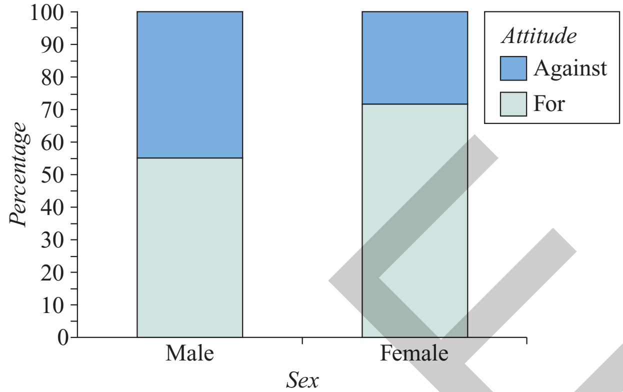
4.2 Using a segmented bar chart to identify relationships in tabulated data

Relationships between categorical variables are identified by comparing percentages. This process can sometimes be made easier by using a percentaged segmented bar chart to display the percentages graphically. For example, the following segmented bar chart is a graphical representation of the information in Table 4.5. Each column in the bar chart corresponds to a column in the purple shaded region of the percentaged table. Each segment corresponds to a cell in the table.

Back to Menu >>>

Table 4.5

Attitude	Sex: Male	Sex: Female
For	55.2%	71.4%
Against	44.8%	28.6%
Total	100.0%	100.0%



From the segmented bar chart, we can see clearly that a greater percentage of females than males favour gun control. This indicates that for this group of people, attitude to gun control is related to sex. If there was no relationship, we would expect the bottom segments in each bar to be roughly equal in length (indicating that similar percentages of males and females were in favour of gun control).

For a two-by-two table (each variable only has two categories), it is relatively easy to see whether the variables are related by comparing percentages. However, when dealing with variables with more than two categories, it is not always so easy to identify trends. In such circumstances, the segmented bar chart is a useful aid. However, we still need to refer to the table for percentages.

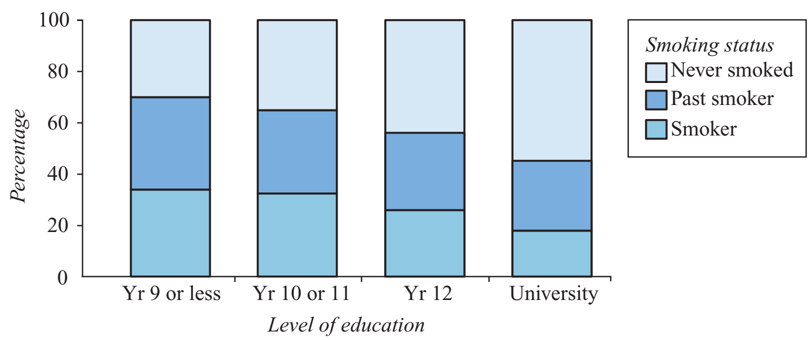
For example Table 4.7 shows the smoking status of adults (smoker, past smoker, never smoked) by level of education (year 9 or less, year 10 or 11, year 12, university).

Table 4.7

	Education level (percentage)			
Smoking status	Year 9 or less	Year 10 or 11	Year 12	University
Smoker	33.9	31.7	26.5	18.4
Past smoker	36.0	33.8	30.9	28.0
Never smoked	30.0	34.5	42.6	53.6
Total	99.9	100.0	100.0	100.0

Source: Hill & White, *Australian Journal of Public Health*, 1995, vol. 9, no. 3, 305–308

The following segmented bar chart is a graph of the information in Table 4.7. Each column represents a column from the purple shaded part of the table.



Back to Menu >>>

From the segmented bar chart, looking at the bottom segment in each column, it is clearly seen that as education level increases there is decrease in the percentage of smokers. Thus we can conclude that there is a relationship between smoking and education level in this sample. We could report this finding as follows.

> **Report**
> From Table 4.7 we see that the percentage of smokers clearly decreases with education level from 33.9% for year 9 or below, to 18.4% for university. This indicates that smoking is related to level of education.

A similar conclusion could be drawn by focusing attention on the top segment of each column, which shows that the percentage of non-smokers increases with education level.

Exercise 4B

1 The table classifies people according to their attitude to Sunday racing and their sex.

	Sex	
Attitude	Male	Female
For	55.6%	54.5%
Against	44.4%	45.5%
Total	100.0%	100.0%

a Display the table graphically in the form of a segmented bar chart.

b Does the segmented bar chart support our previous conclusion (Exercise 4A) that attitude to Sunday racing is not related to sex?

2 As part of the General Social Survey conducted in the US, respondents were asked to say whether they found life exciting, pretty routine or dull. Their marital status was also recorded as married, widowed, divorced, separated or never married. The results are organised below into tabular form:

	Marital status					
Attitude to life	Married	Widowed	Divorced	Separated	Never	*Total*
Exciting	392		77	18	146	
Pretty routine	401	82			124	704
Dull	31		11	4	9	73
Total		151	165	42	279	1461

a How many people were:
   i in the study? **ii** divorced? **iii** separated and found life dull?
   iv married and found life pretty routine?

b Fill in the gaps in the table.

c Turn the frequencies into percentages by calculating column percentages.

d Display the information in the percentaged table using a segmented bar chart.

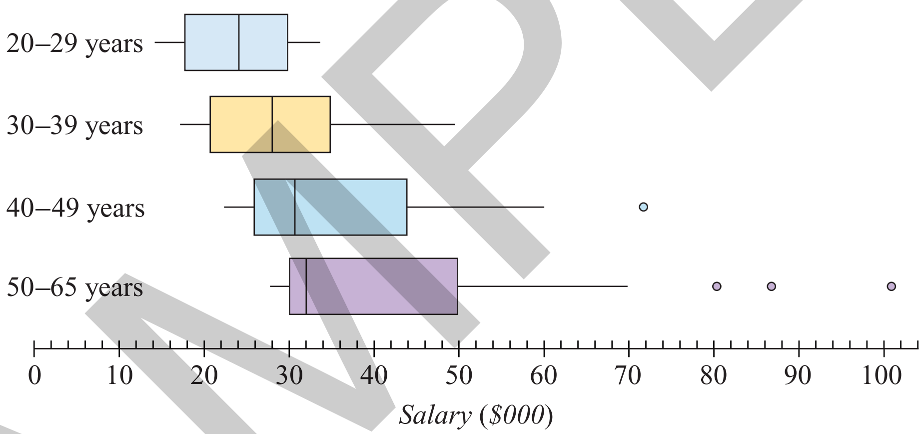
e Does the data support the contention that a person's attitude to life is related to their marital status? Justify your argument by quoting appropriate percentages.

f If attitude to life and marital status are related, which would be the likely independent variable?

4.3 Investigating the relationship between a numerical and a categorical variable

We wish to investigate the relationship between the numerical variable *Salary* (in thousands of dollars), and *Age group* (20–29 years, 30–39 years, 40–49 years, 50–65 years), a categorical variable. The statistical tool that we use to investigate the relationship between a numerical variable and a categorical variable is a series of parallel box plots. In this display, there is one box plot for each category of the categorical variable. Relationships can then identified by comparing the distribution of the numerical variable in terms of shape, centre and spread. You have already learned how to do this in Chapter 2, section 2.5.

The parallel box plots show the salary distribution for four different age groups, 20–29 years, 30–39 years, 40–49 years, 50–65 years. Note that in this situation, the numerical variable *Salary* is the **dependent** variable and the categorical variable *Age group* is the **independent** variable.



There are several ways of deducing the presence of a relationship between salary and age group from this display:

- comparing **medians**

> **Report**
> From the parallel box plots we can see that median salaries increase with age group, from around $24 000 for 20—29-year-olds to around $32 000 for 50—65-year-olds. This is an indication that typical salaries are related to age group.

- comparing **IQRs** and/or **ranges**

> **Report**
> From the parallel box plots we can see that spread of salaries increased with age. For example, the IQR increased from around $12 000 for 20—29-year-olds to around $20 000 for 50—65-year-olds. This is an indication that the spread of salaries is related to age group.

Back to Menu >>>

- comparing **shapes**

> **Report**
> From the parallel box plots we can see that the shape of the distribution of salaries changes with age. It is approximately symmetric for the 20–29-year-olds and becomes progressively more positively skewed with increasing age. We can also see that with increasing age, more outliers begin to appear, indicating salaries well above normal. This is an indication that the shape of the distribution of salaries is related to age group.

Note: Any one of these reports by themselves can be used to claim that there is a relationship between salary and age. However, the use of all three gives a more complete description of this relationship.

Exercise 4C

1 Each of the following variable pairs are related. In each case:

i classify the variable as categorical or numerical

ii name the likely dependent variable

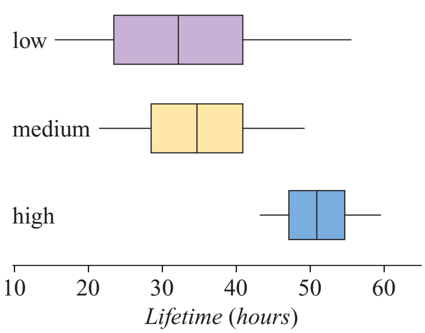
a weight loss (kg) and level of exercise (low, medium, high)

b hours of study (low, medium, high) and test mark

c state of residence and number of sporting teams

d temperature (°C) and season

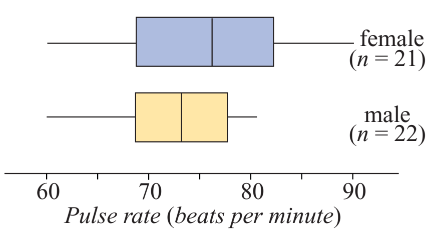
2 The parallel box plots show the distribution of the life time (in hours) of three different priced batteries (low, medium, high).



a The two variables displayed here are battery *Lifetime* and battery *Price* (low, medium, high). Which is the numerical and which is the categorical variable?

b Do the parallel boxplots support the contention that battery lifetime depends on price? Explain.

3 The two parallel box plots show the distribution of pulse rate of 21 adult females and 22 adult males.



a The two variables displayed here are *Pulse rate* and *Sex* (male, female).

i Which is the numerical and which is the categorical variable?

ii Which is the dependent and which is the independent variable?

b Do the parallel box plots support the contention that pulse rate depends on sex? Write a brief report based on centre.

Back to Menu >>>

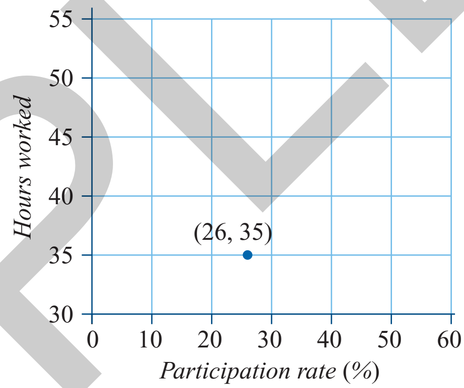
4.4 Investigating the relationship between two numerical variables

The first step in investigating the relationship between two numerical variables is to construct a scatterplot. We will illustrate the process by constructing a scatterplot to display average *Hours worked* (the DV) against university *Participation rate* (the IV) in 9 countries. The data is shown below.

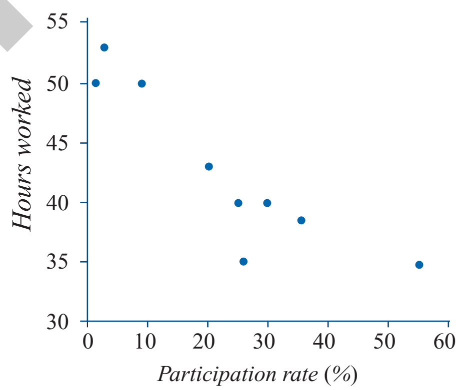
Participation rate (%)	26	20	36	1	25	9	30	3	55
Hours worked	35	43	38	50	40	50	40	53	35

Constructing a scatterplot

In a scatterplot, each point represents a single case, in this instance, a country. The horizontal or x coordinate of the point represents the university participation rate (the IV) and the vertical or y coordinate represents the average working hours (the DV). The scatterplot opposite shows the point for a country for which the university participation rate is 26% and average hours worked is 35.



The scatterplot is completed by plotting the points for each of the remaining countries as shown opposite.



When constructing a scatterplot it is conventional to use the **vertical** or ***y* axis** for the dependent variable (**DV**) and the **horizontal** or ***x* axis** for the independent variable (**IV**).

Following this convention will become very important when we come to fitting lines to scatterplots in the next chapter, so it is a good habit to get into right from the start.

Cambridge University Press • Uncorrected Sample pages • 978-0-521-61328-6 • 2008 © Jones, Evans, Lipson
TI-Nspire & Casio ClassPad material in collaboration with Brown and McMenamin

How to construct a scatterplot using the TI-Nspire CAS

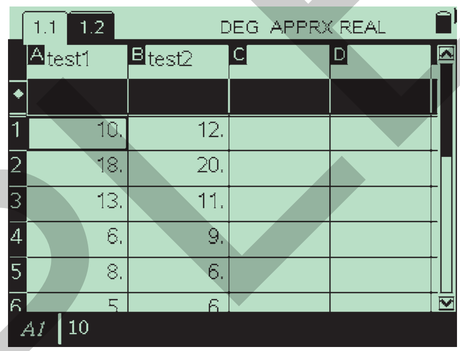
Construct a scatterplot for the set of test scores given below.
Treat Test 1 as the independent (i.e. x) variable.

Test 1 score	10	18	13	6	8	5	12	15	15
Test 2 score	12	20	11	9	6	6	12	13	17

Steps

1 Start a new document by pressing (enter) + (N).

2 Select **3:Add Lists & Spreadsheet**. Enter the data into lists named *test1* and *test2*.



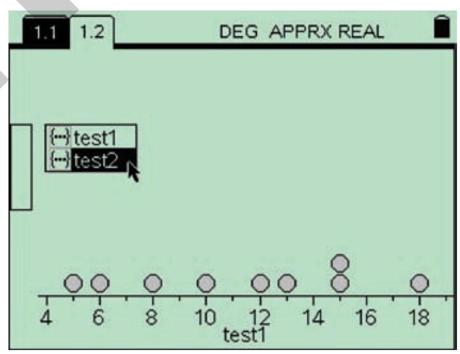
3 Statistical graphing is done through the **Data & Statistics** application.

Press (home) and select **5:Data & Statistics**.

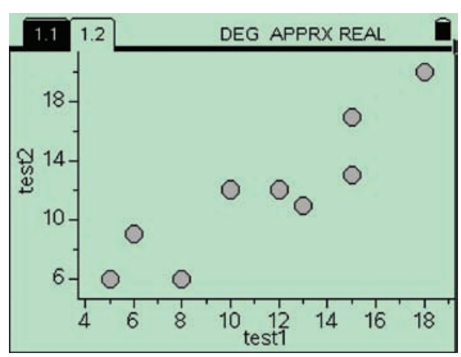
A random display of dots (not shown here) will appear – this is to indicate list data is available for plotting. It is not a statistical plot.

a On this plot, move the cursor to the text box area below the horizontal (or x-) axis. Press (click) when prompted and select the independent variable, *test1*. Press (enter) to paste the variable to that axis.

b Now move the cursor towards the centre of the vertical (or y-) axis until a text box appears (as shown opposite).



c Press (click) when prompted to select the dependent variable, *test2*. Pressing (enter) pastes the variable to that axis and generates a scatterplot as shown opposite. The plot is scaled automatically.

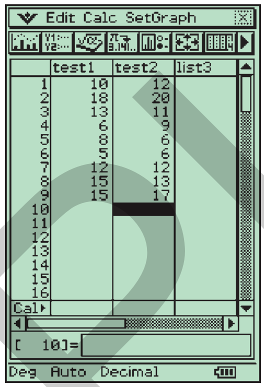


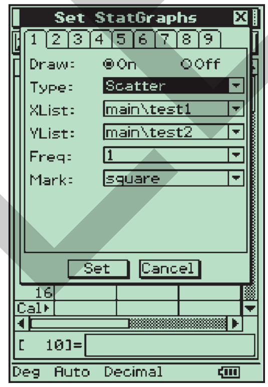
How to construct a scatterplot using the ClassPad

Construct a scatterplot for the set of test scores given below.
Treat Test 1 as the independent (i.e. x) variable.

Test 1 score	10	18	13	6	8	5	12	15	15
Test 2 score	12	20	11	9	6	6	12	13	17

Steps



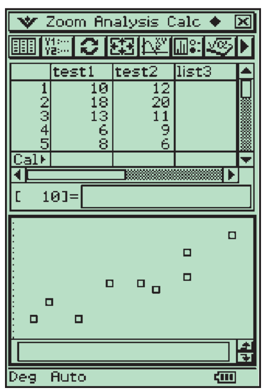


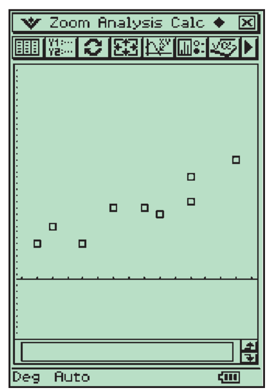
1 Open the **Statistics** application and enter the data into the columns named **test1** and **test2**. Your screen should look like the one shown.

2 Tap to open the **Set StatGraphs** dialog box and complete as given below. For

- **Draw:** select **On**
- **Type:** select **Scatter** ( )
- **XList:** select **main \ test1** ( )
- **YList:** select **main \ test2** ( )
- **Freq:** leave as **1**
- **Mark:** leave as square

Tap SET to confirm your selections.





3 Tap in the toolbar at the top of the screen to plot the scatterplot in the bottom half of the screen.

4 To obtain a full-screen plot, tap Resize from the icon panel.

Note: If you have more than one graph on your screen, tap the data screen, select **StatGraph** and turn off any unwanted graphs.

Exercise 4D

1

Minimum temperature (x)	17.7	19.8	23.3	22.4	22.0	22.0
Maximum temperature (y)	29.4	34.0	34.5	35.0	36.9	36.4

The table above shows the maximum and minimum temperatures (in °C) during a hot week in Melbourne. Using a calculator, construct a scatterplot with *Minimum temperature* as the IV (x-variable). Name variables, *mintemp* and *maxtemp*.

2

Balls faced	29	16	19	62	13	40	16	9	28	26	6
Runs scored	27	8	21	47	3	15	13	2	15	10	2

The table above shows the number of runs scored and the number of balls faced by batsmen in a one-day international cricket match. Use a calculator to construct an appropriate scatterplot. Remember to identify the IV.

3

Temperature (°C)	0	10	50	75	100	150
Diameter (cm)	2.00	2.02	2.11	2.14	2.21	2.28

The table above shows the changing diameter of a metal ball as it is heated. Use a calculator to construct an appropriate scatterplot. Temperature is the IV.

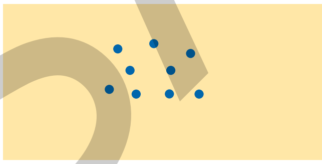
4

Number in theatre	87	102	118	123	135	137
Time (minutes)	0	5	10	15	20	25

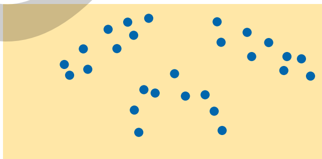
The table above shows the number of people in a theatre at five minute intervals after the advertisements started. Use a calculator to construct an appropriate scatterplot.

4.5 How to interpret a scatterplot

What features do we look for in a scatterplot that will help us identify and describe any relationships present? First we look to see if there is a **clear pattern** in the scatterplot.



In the example opposite, there is **no clear pattern** in the points. The points are just **randomly scattered** across the plot.
Conclude that there is **no relationship**.



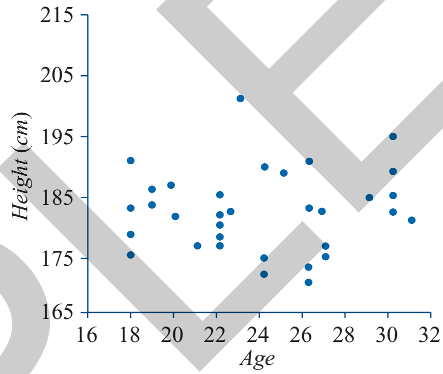
For the three examples opposite, there is a **clear** (but different) **pattern** in each of the sets of points.
Conclude that there is a **relationship**.

Having found a clear pattern, there are several things we look for in the pattern of points. These are:

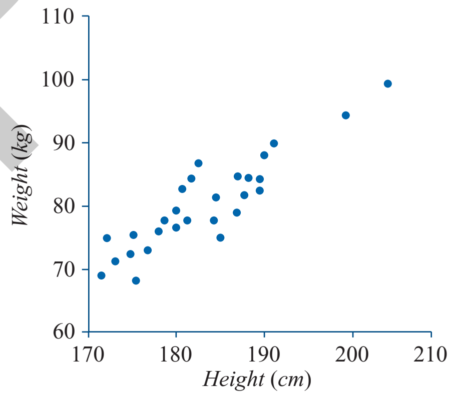
- direction and outliers (if any)
- form
- strength

Direction and outliers

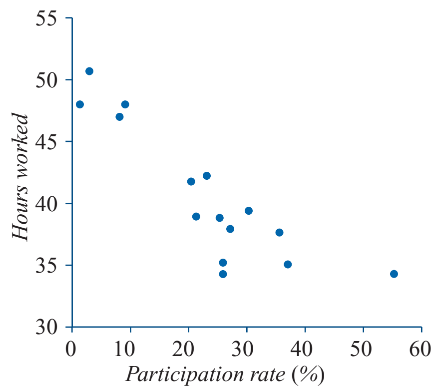
The scatterplot of height against age of a group of footballers (shown opposite) is just a **random** scatter of points. This suggests that there is **no relationship** between the variables *Height* and *Age* for this group of footballers. However, there is an **outlier**, the footballer who is 201 cm tall.



In contrast, there is a **clear pattern** in the scatterplot of weight against height for the same group of footballers (shown opposite). The two **variable**s are **related**. Furthermore, the points seem to **drift upwards** as you move across the plot. When this happens, we say that there is a **positive relationship** between the variables. Tall players tend to be heavy and vice versa. In this scatterplot, there are **no outliers**.



Likewise, the scatterplot of working hours against university participation rates for 15 countries shows a **clear pattern**. The two **variable**s are **related**. However, in this case the points seem to **drift downwards** as you move across the plot. When this happens, we say that there is a **negative** relationship between the variables. Countries with high working hours tend to have low university participation rates and vice versa. In this scatterplot, there are **no outliers**.

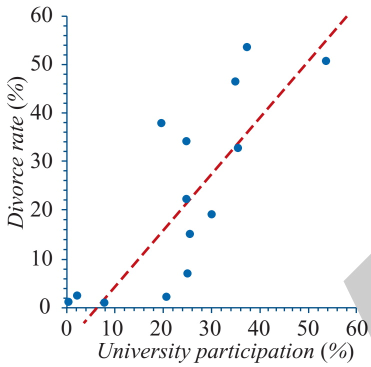


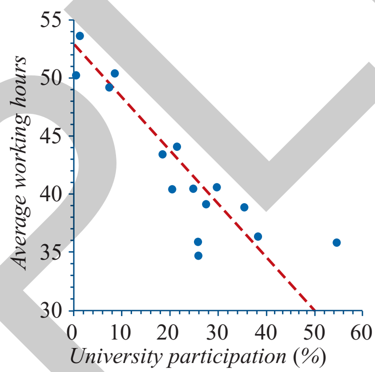
Back to Menu >>>

Form

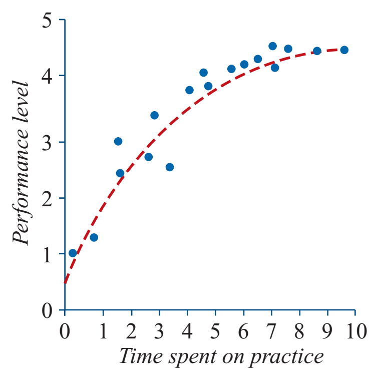
What we are looking for here is whether the pattern in the points has a **linear form**. If the points in a scatterplot can be thought of as random fluctuations around a **straight line**, then we say that the scatterplot has a linear form. If the scatterplot has a **linear form** then we say that the variables involved are **linearly related**.

For example, both of the scatterplots shown below can be described as having a **linear form**; that is, the scatter in the points can be thought of as just random fluctuations around a straight line. We can then say that the relationships between the variables involved are linear. (The dotted straight lines have been added to the graphs to make it easier to see the linear form.)





By contrast, the scatterplot opposite clearly has a non-linear form. This is a plot of performance level against time spent on practising a task. There is a relationship between performance level and time spent in practice, but it is clearly non-linear. The scatterplot shows that while level of performance on a task will increase with practice, there comes a time when the performance level will not improve substantially with extra practice.



While non-linear relationships exist (and we must always check for their presence by examining the scatterplot), many of the relationships we meet in practice are linear or may be made linear by transforming the data (a technique you will meet in Chapter 6). For this reason we will now restrict ourselves to the analysis of scatterplots with linear forms.

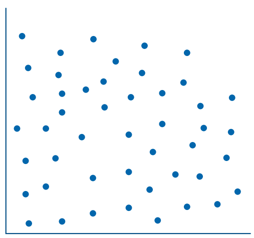
Back to Menu >>>

Strength of a linear relationship: the correlation coefficient

The strength of a linear relationship is an indication of how closely the points in the scatterplot fit a straight line. If the points in the scatterplot lie exactly on a straight line, we say that there is a perfect linear relationship. If there is no fit at all we say there is no relationship. In general, we have an imperfect fit, as seen in all of the scatterplots to date.

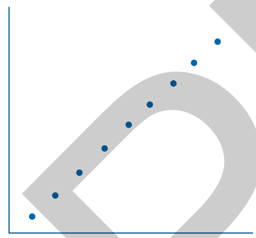
To measure the strength of a linear relationship, a statistician called Carl Pearson developed a **correlation coefficient**, ***r***, which has the following properties:

- If there is **no linear** relationship, $r = 0$.



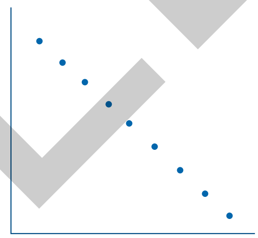
$r = 0$

- If there is a **perfect positive linear** relationship, $r = +1$.



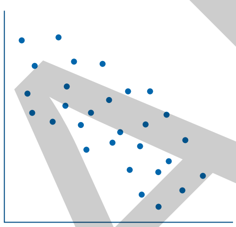
$r = +1$

- If there is a **perfect negative linear** relationship, $r = -1$.

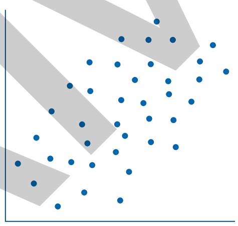


$r = -1$

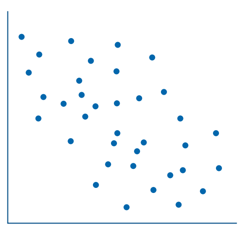
If there is a less than perfect linear relationship, then the correlation coefficient r has a value between -1 and $+1$, or $-1 < r < +1$. The scatterplots below show the approximate values of r for linear relationships of varying strengths.



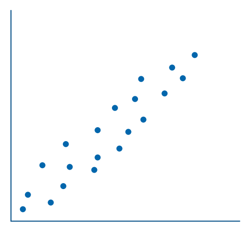
$r = -0.7$



$r = +0.5$



$r = -0.3$



$r = +0.9$

At present, these scatterplots with their associated correlation coefficients should help you get a feel for the relationship between the correlation coefficient and a scatterplot. Later in this chapter, you will learn to calculate its value. At the moment you only have to be able to roughly estimate the value of the correlation coefficient from the scatterplot by comparing it with standard plots such as those given above.

Back to Menu >>>

Guidelines for classifying the strength of a linear relationship

Our reason for estimating the value of the correlation coefficient is to give a measure of the strength of the linear relationship. When doing this, we sometimes find it useful to classify the strength of the linear relationship as **weak**, **moderate** or **strong** as shown opposite.

Strong positive relationship r between 0.75 and 0.99
Moderate positive relationship r between 0.5 and 0.74
Weak positive relationship r between 0.25 and 0.49
No relationship r between –0.24 and +0.24
Weak negative relationship r between –0.25 and –0.49
Moderate negative relationship r between –0.5 and –0.74
Strong negative relationship r between –0.75 and –0.99

For example, the correlation coefficient between scores of a test of verbal skills and a test on mathematical skills is:

$$r_{\text{verbal, mathematical}} = +0.275$$

indicating that there is a **weak** positive linear relationship.

In contrast, the correlation coefficient between carbon monoxide level and traffic volume is

$$r_{\text{CO level, traffic volume}} = +0.985$$

indicating a **strong** positive linear relationship between carbon monoxide level and traffic volume.

Warning!!

If you are using the value of the **correlation coefficient** as a measure of the strength of a relationship, then you are implicitly **assuming**:

1 the variables are **numeric**

2 the **relationship** is **linear**

3 there are **no outliers** in the data. The **correlation coefficient** can give a **misleading** indication of the strength of the linear relationship if there are outliers present.

Exercise 4E

1 For each of the following pairs of variables, indicate whether you expect a relationship to exist between the variables and, if so, whether you would expect the variables to be positively or negatively related:

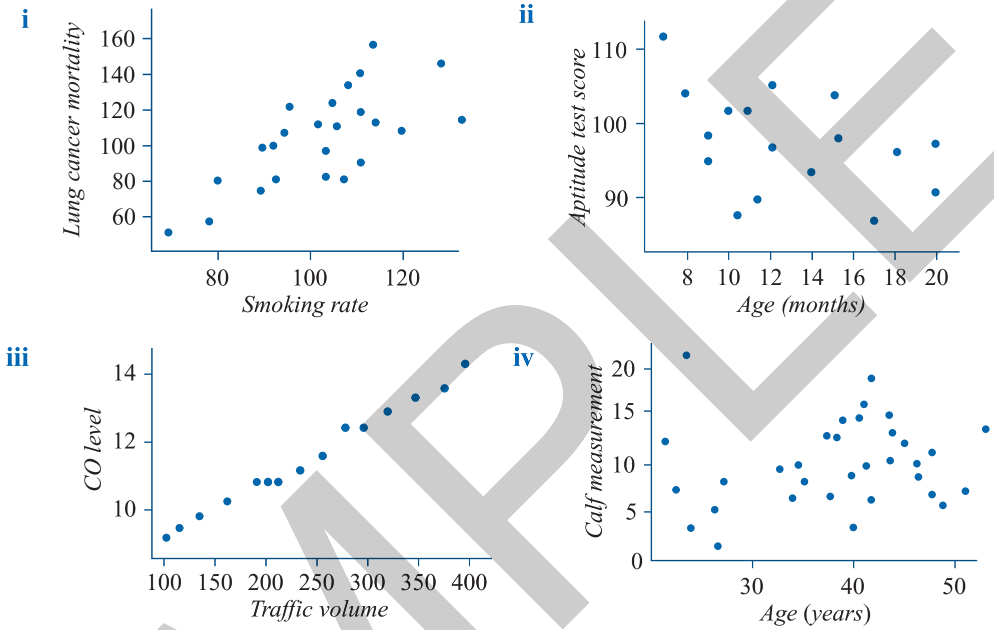
a intelligence and height
b intelligence and salary level
c salary earned and tax paid
d frustration and aggression
e population density and distance from the centre of a city
f time spent watching TV and creativity

Cambridge University Press • Uncorrected Sample pages • 978-0-521-61328-6 • 2008 © Jones, Evans, Lipson
TI-Nspire & Casio ClassPad material in collaboration with Brown and McMenamin

2 For each of the following scatterplots, state whether the variables appear to be related. If the variables appear to be related:

a state whether the relationship is positive or negative

b estimate the strength of the relationship by estimating the value of the correlation coefficient and classifying it as either weak, moderate, strong or no relationship



3 What three assumptions do you make when you use the value of the correlation coefficient as a measure of the strength of a relationship?

4.6 Calculating Pearson's correlation coefficient r

Pearson's correlation coefficient r gives a numerical measure of the degree to which the points in the scatterplot tend to cluster around a straight line.

Formally, if we call the two variables we are working with x and y, and we have n observations, then r is given by:

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}$$

In this formula, $\bar{x}$ and s_x are the mean and standard deviation of the x values and $\bar{y}$ and s_y are the mean and standard deviation of the y values.

Calculating the correlation coefficient using the formula (optional)

In practice, you can always use your calculator to determine the value of the correlation coefficient. However, to understand what is involved when your calculator is doing the calculation for you, it is best that you know how to calculate the correlation coefficient from the formula first.

How to calculate the correlation coefficient using the formula

Use the formula to calculate the correlation coefficient r for the following data.

x	1	3	5	4	7
y	2	5	7	2	9

$\bar{x} = 4,\ s_x = 2.236$
$\bar{y} = 5,\ s_y = 3.082$

Give the answer correct to two decimal places.

Steps

1 Write down the values of the means, standard deviations and n.

$\bar{x} = 4 \quad s_x = 2.236$
$\bar{y} = 5 \quad s_y = 3.082 \quad n = 5$

2 Set up a table like that shown opposite to calculate $\Sigma(x - \bar{x})(y - \bar{y})$.

x	$(x - \bar{x})$	y	$(y - \bar{y})$	$(x - \bar{x}) \times (y - \bar{y})$
1	−3	2	−3	9
3	−1	5	0	0
5	1	7	2	2
4	0	2	-3	0
7	3	9	4	12
Sums	0		0	23

$\therefore \sum(x - \bar{x})(y - \bar{y}) = 23$

3 Write down the formula for r. Substitute the appropriate values and evaluate.

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

$$\therefore r = \frac{23}{(5 - 1) \times 2.236 \times 3.082} = 0.834\ldots$$

4 Write down your answer, giving r correct to two decimal places.

Correct to two decimal places, the correlation coefficient is $r = 0.83$

Back to Menu >>>

Determining the correlation coefficient using a graphics calculator

The graphics calculator automates the process of calculating a correlation coefficient. However, it does it as part of the process of fitting a straight line to the data (the topic of Chapter 5). As a result, more statistical information will be generated than you need at this stage.

How to calculate the correlation coefficient using the TI-Nspire CAS

Determine the value of the correlation coefficient r for the given data. Give the answer correct to 2 decimal places.

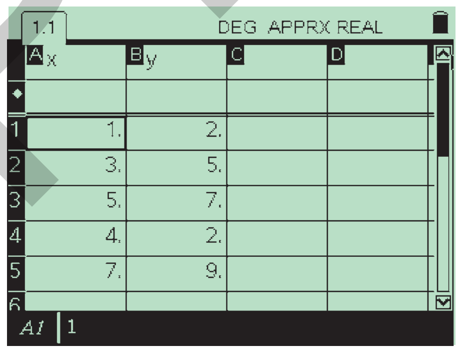
x	1	3	5	4	7
y	2	5	7	2	9

Steps

1 Start a new document by pressing (enter) + (N).

2 Select **3:Add Lists & Spreadsheet**.
Enter the data into lists named x and y.

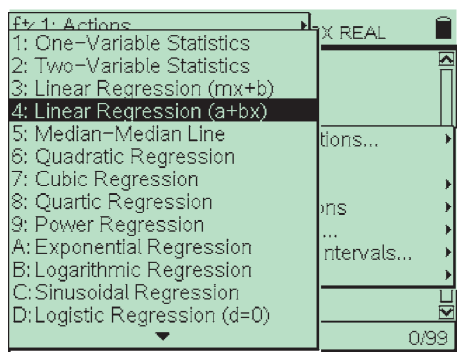
3 Statistical calculations can be done in the **Calculator** application (as used here) or the **Lists & Spreadsheet** application.
Press (home) and select **1:Calculator**.



Method 1

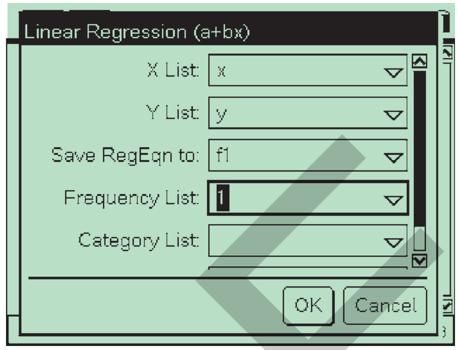
Using the **Linear Regression (a+bx)** command

a Press (menu) **/6:Statistics/1:Stat Calculations/4:Linear Regression (a+bx)** to generate the screen opposite.

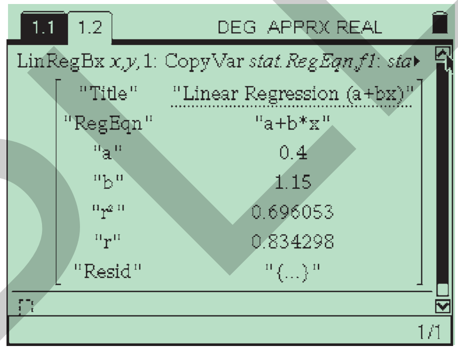


Back to Menu >>>

b Press enter to generate the pop-up screen as shown. To select the variable for the X List entry use the ▼ arrow and enter to select and paste in the list name x. Press tab to move to the Y List entry, use the ▼ arrow twice and enter to select and paste in the list name y.



c Press enter to exit the pop-up screen and generate the results shown in the screen opposite.



The value of the correlation coefficient is $r = 0.8342\ldots$ or 0.83, correct to 2 decimal places.

Method 2

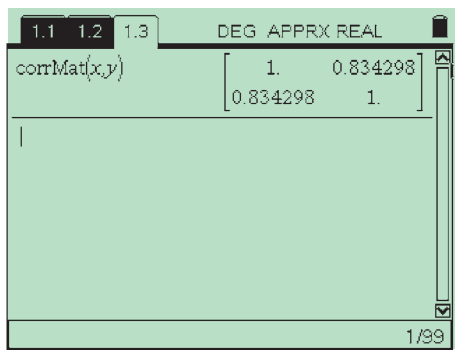
Using the **corrMat(x, y)** command

In the **Calculator** application, type in **corrmat(x, y)** and press enter.

Alternatively

a Press 1 C to access the **Catalog**, scroll down to **corrMat(**and press enter to select and paste the **corrMat(**command onto the **Calculator** screen.

b Complete the command by typing in ***x*,*y*** and press. enter.



The value of the correlation coefficient is $r = 0.8342\ldots$ or 0.83, correct to 2 decimal places.

Back to Menu >>>

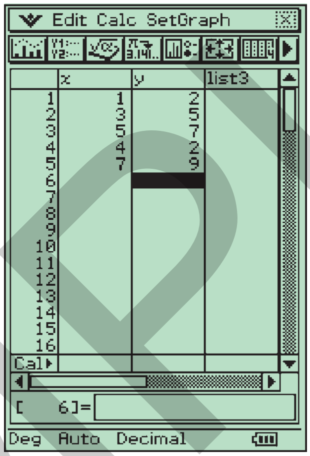
How to calculate the correlation coefficient using the ClassPad

Determine the value of the correlation coefficient r for the given data. Give the answer correct to 2 decimal places..

x	1	3	5	4	7
y	2	5	7	2	9

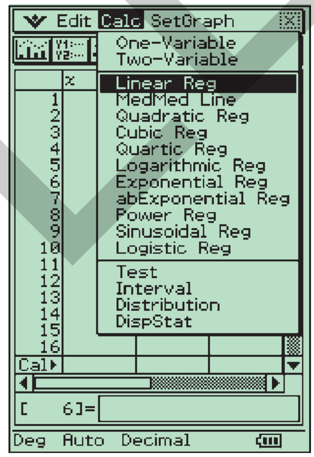
Steps

1 Open the **Statistics** application and enter the data into columns labelled $\boldsymbol{x}$ and $\boldsymbol{y}$. Your screen should look like the one shown.



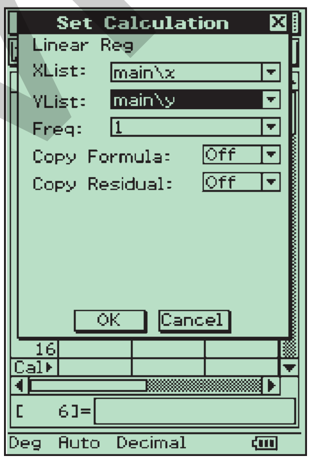
2 Select **Calc** from the menu bar, and then **Linear Reg** and press (EXE).

This opens the **Set Calculation** dialog box shown below (left).



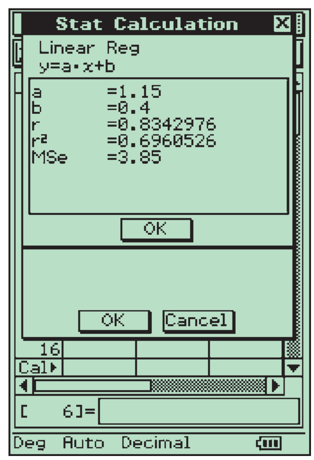
3 Complete the **Set Calculations** dialog box as shown. For

- **XList:** select **main** \ $\boldsymbol{x}$(▾)
- **YList:** select **main** \ $\boldsymbol{y}$(▾)
- **Freq:** leave as **1**
- **Copy Formula:** select **Off**
- **Copy Residual:** select **Off**



4 Tap **OK** to confirm your selections and generate the required results.

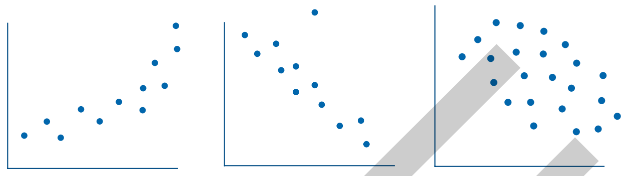
The value of the correlation coefficient is $r = 0.8342\ldots$ or 0.83, correct to 2 decimal places.



Skillsheet TEACHER

Exercise 4F

1 The scatterplots of three sets of related variables are shown opposite.



a For each scatterplot, describe the relationship in terms of direction, form and outliers (if any).

b For which of the scatterplots would it **not** be appropriate to use the correlation coefficient r to give a measure of the strength of the relationship between the variables? Give reasons for your decisions.

2 Use the formula to calculate the correlation coefficient r for this data.

x	2	3	6	3	6
y	1	6	5	4	9

$\bar{x} = 4,\ s_x = 1.871$
$\bar{y} = 5,\ s_y = 2.915$

Give the answer correct to two decimal places.

3 **a** The table below shows the maximum and minimum temperatures during a heat-wave week. *Maximum* and *Minimum temperature* are linearly related variables. There are no outliers. Use your calculator to show that $r = 0.818$ correct to three decimal places.

Day	*Sunday*	*Monday*	*Tuesday*	*Wednesday*	*Thursday*	*Friday*
Max. temperature (°C)	29.4	34.0	34.5	35.0	36.9	36.4
Min. temperature (°C)	17.7	19.8	23.3	22.4	22.0	22.0

b This table shows the number of runs scored and balls faced by batsmen in a cricket match. *Runs scored* and *Balls faced* are linearly related variables. There are no outliers. Use your calculator to show that $r = 0.8782$ correct to four decimal places.

Batsman	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Runs scored	27	8	21	47	3	15	13	2	15	10	2
Balls faced	29	16	19	62	13	40	16	9	28	26	6

c This table shows the hours worked and university participation rate (%) in six countries. *Hours worked* and university *Participation rate* are linearly related variables. There are no outliers. Use your calculator to show that $r = -0.6727$ correct to four decimal places.

Country	*Australia*	*Britain*	*Canada*	*France*	*Sweden*	*US*
Hours worked	35.0	43.0	38.2	39.8	35.6	34.8
Participation rate (%)	26	20	36	25	37	55

d This table shows the number of TVs and cars owned per 1000 people in six countries. *Number of TVs* and *Number of cars* owned are linearly related variables. There are no outliers. Use your calculator to show that $r = 0.82$ correct to two decimal places.

Back to Menu >>>

Country	*Australia*	*Britain*	*Canada*	*France*	*Sweden*	*US*
Number of TV's/1000	378	404	471	354	381	624
Number of cars/1000	417	286	435	370	357	550

4.7 The coefficient of determination

If two variables are related, it is possible to estimate the value of one variable from the value of the other. For example, people's weight and height are related. Thus, given a person's height, we should be able to roughly predict the person's weight. The degree to which we can make such predictions depends on the value of r. If there is a perfect linear relationship ($r = 1$) between two variables then we can exactly predict the value of one variable from the other.

For example, when you buy cheese by the gram there is an exact relationship ($r = 1$) between the weight of cheese you buy and the amount you pay. At the other end of the scale, for adults, there is no relationship between an adult's height and their IQ ($r \approx 0$). Knowing an adult's height will not enable you to predict their IQ any better than guessing.

The coefficient of determination

The **degree** to which one variable can be predicted from another linearly related variable is given by a statistic called the **coefficient of determination**.

The coefficient of determination is **calculated** by squaring the correlation coefficient:

$$\text{coefficient of determination} = \boldsymbol{r^2}$$

Calculating the coefficient of determination

Numerically, the coefficient of determination $= r^2$. Thus, if correlation between weight and height is $r = 0.8$, then the

$$\text{coefficient of determination} = r^2 = 0.8^2 = 0.64 \qquad \text{or} \qquad 0.64 \times 100 = 64\%$$

Note: We have converted the coefficient of determination into a percentage (64%) as this is the most useful form when we come to interpreting the coefficient of determination.

Interpreting the coefficient of determination

We now know how to calculate the coefficient of determination, but what does it tell us?

Interpreting the coefficient of determination

In technical terms, the coefficient of determination tells us that $r^2 \times 100$ percent of the **variation** in the **dependent variable** (DV) is **explained** by the **variation** in the **independent variable** (IV).

But what does this mean in practical terms?

Let us take the relationship between weight and height that we have just been considering as an example. Here the coefficient of determination is 0.64 (or 64%).

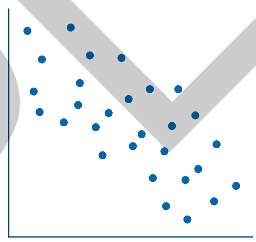
The coefficient of determination tells us that 64% of the variation in people's weight (the DV) is explained by the variation in their height (the IV).

What do we mean by 'explained'?

If we take a group of people, we find that both their weights and heights will vary. One explanation for the variation in people's weights is that their heights vary. Taller people tend to be heavier. Shorter people tend to be lighter. The coefficient of determination tells us that 64% of the variation in people's weights can be explained in this way. The rest of the variation (36%) in their weights will be explained by other factors, for example, sex, lifestyle, build.

Example 1 Calculating the correlation coefficient from the coefficient of determination

For the relationship described by this scatterplot, the coefficient of determination = 0.5210.
Determine the value of the correlation coefficient r.



Solution

1 The coefficient of determination = r^2. Use this information and the value of the coefficient of determination to set up an equation for r. Solve.

$r^2 = 0.5210$
$\therefore r = \pm\sqrt{0.5210} = \pm 0.7218$

2 There are two solutions, one positive, one negative. Use the scatterplot to decide which applies.

Scatterplot indicates a negative relationship.

3 Write down your answer.

$\therefore r = -0.7218$

Example 2 Calculating and interpreting the coefficient of determination

Carbon monoxide (CO) levels in the air and traffic volume are linearly related with:

$$r_{\text{co level, traffic volume}} = +0.985$$

Determine the value of the coefficient of determination, write it in percentage terms and interpret. In this relationship, CO content is the DV.

Solution

The coefficient of determination is:

$$r^2 = (0.985)^2 = 0.970\ldots \quad \text{or} \quad 0.970 \times 100 = 97.0\%$$

Therefore, 97% of the variation in carbon monoxide levels in the atmosphere can be explained by the variation in traffic volume.

Cambridge University Press • Uncorrected Sample pages • 978-0-521-61328-6 • 2008 © Jones, Evans, Lipson
TI-Nspire & Casio ClassPad material in collaboration with Brown and McMenamin

Back to Menu >>>

Clearly, traffic volume is a very good predictor of carbon monoxide levels in the air. Knowing the traffic volume will enable us to predict carbon monoxide levels with a high degree of accuracy. This contrasts with the next example, which concerns the ability to predict mathematical ability from verbal ability.

Example 3 **Calculating and interpreting the coefficient of determination**

Scores on tests of verbal and mathematical ability are linearly related with:

$$r_{\text{mathematical, verbal}} = +0.275$$

Determine the value of the coefficient of determination, write it in percentage terms, and interpret. In this relationship, mathematical ability is the DV.

Solution

The coefficient of determination is:

$$r^2 = (0.275)^2 = 0.0756\ldots \text{ or } 0.076 \times 100 = 7.6\%$$

Therefore, only 7.6% of the variation observed in scores on the test of mathematical ability can be explained by the variation in scores obtained on the test of verbal ability.

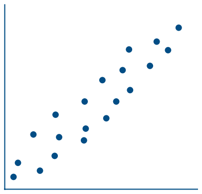
Clearly, scores on the verbal ability test are not good predictors of the scores on the mathematical ability test; 92.4% of the variation in mathematical ability is explained by other factors.

Exercise 4G

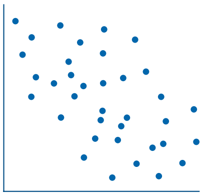
1 For each of the following values of r, calculate the value of the coefficient of determination and convert to a percentage (correct to one decimal place).

a $r = 0.675$ **b** $r = 0.345$ **c** $r = -0.567$ **d** $r = -0.673$ **e** $r = 0.124$

2 **a** For the relationship described by the scatterplot shown opposite, the coefficient of determination = 0.8215. Determine the value of the correlation coefficient r (correct to three decimal places).



b For the relationship described by the scatterplot shown opposite, the coefficient of determination = 0.1243. Determine the value of the correlation coefficient r (correct to three decimal places).



3 For each of the following, determine the value of the coefficient of determination, write it in percentage terms, and interpret.

a Scores on hearing tests (DV) and age are linearly related, with: $r_{\text{hearing, age}} = -0.611$

b Mortality rates (DV) and smoking rates are linearly related, with: $r_{\text{mortality, smoking}} = +0.716$

c Life expectancy (DV) and birth rates are linearly related, with: $r_{\text{life expectancy, birth rate}} = -0.807$

d Daily maximum (DV) and minimum temperatures are linearly related, with: $r_{\text{max, min}} = 0.818$

e Runs scored (DV) and balls faced by a batsman are linearly related, with: $r_{\text{runs, balls}} = 0.8782$

4.8 Correlation and causality

Some statements to consider

A study of primary school children found a high positive correlation between shoe size and reading ability. Can we conclude that having small feet causes a person to have a low level of reading ability? Or, is it just that as children grow older, their reading ability increases as does their shoe size?

The number of days a patient stays in hospital has been shown by a study to be positively correlated to the number of beds in the hospital. Can it be said that these hospitals are encouraging patients to stay in hospital longer than necessary to keep their beds occupied? Or, is it just that bigger hospitals treat more people with serious illnesses and these require longer hospital stays?

While you might establish a relationship between two variables, this in itself is not sufficient to imply that a change in one of the one variables will **cause** a change in the other. For example, if you gathered data about crime rates and unemployment rates in a range of cities you would find that they are highly correlated. But can you then go on and infer that decreasing unemployment will lead to (cause) a decrease in crime rates? It may, but we cannot make such a conclusion on the basis of correlation alone. Many other possible explanations could be found that might equally explain both a high crime rate and a high unemployment rate. Factors such as home background, peer group, education level and economic conditions are possible explanations. Thus, two variables may vary together without one directly being the cause of the other and we must be aware of not reading too much into any relationships we might discover.

Exercise 4H

Consider these reports.

1 A study of primary school children aged 5 to 11 finds a high positive correlation between height and score on a test of mathematics ability. Does this mean that taller people are better at mathematics? What other factors might explain this relationship?

Cambridge University Press • Uncorrected Sample pages • 978-0-521-61328-6 • 2008 © Jones, Evans, Lipson
TI-Nspire & Casio ClassPad material in collaboration with Brown and McMenamin

2 It is known that there is a clear positive correlation between the number of churches in a town and the amount of alcohol consumed by its inhabitants. Does this mean that religion is driving people to drink? What other factors might explain this relationship?

3 There is a strong positive correlation between the amount of ice-cream consumed and the number of drownings each day. Does this mean that the consumption of ice-cream at the beach is dangerous? What other factors might explain this relationship?

4 Students who perform well in music exams are also known to perform well in mathematics exams. Does this mean that in order to do well in mathematics you should take up a musical instrument? What other factors might explain this relationship?

4.9 Which graph?

One of the problems that you will face is choosing a suitable graph to investigate a relationship. The following guidelines might help you in your decision making.

Type of data		Graph
Dependent variable	Independent variable	
Categorical	Categorical	Segmented bar chart
Categorical	Numerical	Parallel box plots
Categorical (two categories only)	Numerical	Back-to-back stemplot (Ch. 2) (box plots preferred)
Numerical	Numerical	Scatterplot

Exercise 4I

1 Which graphical display (parallel box plots, a segmented bar chart, or a scatterplot) would be appropriate to display the relationship between:

a vegetarian (yes, no) and sex (male, female)
b mark obtained on a statistics test out of 100 and time spent studying (in hours)
c number of hours spent at the beach each year and state of residence
d number of CDs purchased per year and income (in dollars)
e runs scored in a cricket game and number of ‘overs’ faced
f attitude to compulsory sport in school (agree, disagree, no opinion) and school type (government, independent)
g income level (high, medium, low) and place of living (urban, rural)
h number of cigarettes smoked per day and sex (male, female)

MC TEST D&D TEST

Cambridge University Press • Uncorrected Sample pages • 978-0-521-61328-6 • 2008 © Jones, Evans, Lipson
TI-Nspire & Casio ClassPad material in collaboration with Brown and McMenamin

Key ideas and chapter summary

Two-way frequency tables

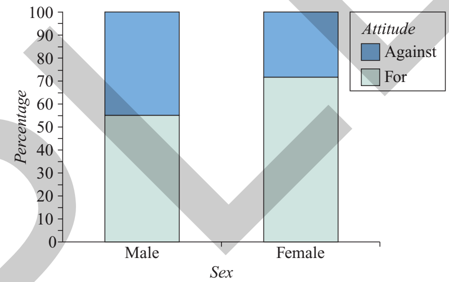
Two-way frequency tables are used as the starting point for investigating the relationship between two **categorical** variables.

Identifying relationships between two categorical variables

Relationships between two categorical variables are **identified** by **comparing** appropriate percentages in a two-way frequency table. When the categories of the **DV** define the rows in the table and the categories of the **IV** define the columns, the appropriate percentages are **column** percentages.

Segmented bar charts

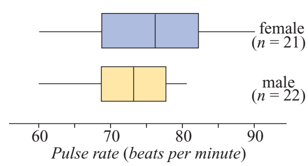
A **segmented bar chart** can be used to graphically display the information contained in a two-way frequency table. It is a useful tool for identifying relationships between two categorical variables.



For example, the clearly higher percentage of females who were 'For' gun control indicates a relationship between attitude to gun control and sex.

Parallel box plots

Parallel box plots can be used to display and describe the relationship between a **numerical** and a **categorical** variable.

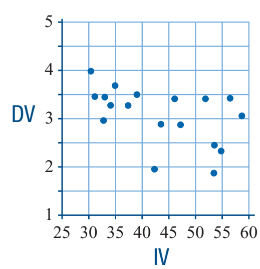


Relationships are **identified** by finding **differences** in the centres, spreads or shapes of the parallel box plots. For example, the difference in the median pulse rate between males and females indicates that the pulse rate depends on sex.

Scatterplots

A **scatterplot** is used to help identify and describe the relationship between two **numerical** variables.

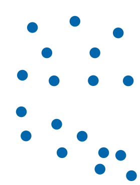
In a scatterplot, the **dependent variable** (**DV**) is plotted on the **vertical** axis and the **independent variable** (**IV**) on the **horizontal** axis.



Identifying relationships between two numerical variables

A **random** cluster of points (no clear pattern) indicates that the variables are **unrelated**.

A **clear pattern** in the scatterplot indicates that the variables are related.



Back to Menu >>>

Describing relationships in scatterplots

Relationships are described in terms of:

- **direction** (positive or negative) and **outliers**
- **form** (linear or non-linear)
- **strength** (weak, moderate or strong)

Correlation coefficient r

The correlation coefficient r gives a measure of the strength of a linear relationship.

Strong positive relationship r between 0.75 and 0.99
Moderate positive relationship r between 0.5 and 0.74
Weak positive relationship r between 0.25 and 0.49
No relationship r between –0.24 and +0.24
Weak negative relationship r between –0.25 and –0.49
Moderate negative relationship r between –0.5 and –0.74
Strong negative relationship r between –0.75 and –0.99

Assumptions made when using r as a measure of strength

- Variables are numeric.
- The underlying relationship between the variables is linear.
- There are no clear outliers.

The coefficient of determination: defined

The **coefficient of determination** $= r^2$

For example, if $r_{\text{pay rate, experience}} = 0.85$, then the coefficient of determination $= r^2 = (0.85)^2 = 0.72$ (or 72%)

The coefficient of determination interpreted

The coefficient of determination above tells us that '72% of the variation in workers salaries (DV) can be explained by the variation in their experience (IV)'.

Which graph?

The graph used to display a relationship between two variables depends on the type of variables:

- **two categorical** variables: segmented bar chart
- a **numerical** and a **categorical** variable: parallel box plots
- **two numerical** variables: scatterplot

Correlation and causation

Correlation does not necessarily imply causation.

Skills check

Having completed this chapter you should be able to:

- interpret the information contained in a two-way frequency table
- identify, where appropriate, the dependent and independent variable in a relationship

Review

- identify a relationship in tabulated data by forming and comparing appropriate percentages
- represent a two-way percentaged frequency table by a segmented bar chart and interpret the chart
- choose among a scatterplot, segmented bar chart and parallel boxplots as a means of graphically displaying the relationship between two variables
- construct a scatterplot
- use a scatterplot to comment on the following aspect of any relationship present:
  - direction (positive or negative association) and possible outliers
  - form (linear or non-linear)
  - strength (weak, moderate, strong)
- calculate and interpret the correlation coefficient r
- know the three key assumptions made when using Pearson's correlation coefficient as a measure of the strength of the relationship between two variables, that is:
  - the variables are numeric
  - the relationship is linear
  - no clear outliers
- calculate and interpret the coefficient of determination
- identify situations where unjustified statements about causality could be (or have been) made

Multiple-choice questions

The information in the following frequency table relates to Questions 1 to 4

	Sex	
Plays sport	Male	Female
Yes	68	79
No	34	
Total	102	175

1 The variables *Plays sport* and *Sex* are:
A both categorical variables
B a categorical and a numerical variable respectively
C a numerical and a categorical variable respectively
D both numerical variables
E neither a numerical nor a categorical variable

2 The number of females who do not play sport is
A 21 **B** 45 **C** 79 **D** 96 **E** 175

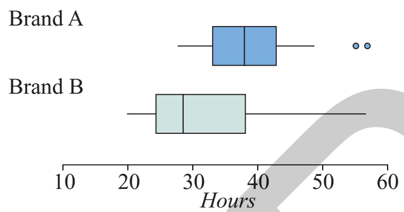
3 The percentage of males who do not play sport is

A 19.4% B 33.3% C 34.0% D 66.7% E 68.0%

4 The variables *Plays sport* and *Sex* appear to be related because

A more females play sport than males
B less males play sport than females
C a higher percentage of females play sport compared to males
D a higher percentage of males play sport compared to females
E both males and females play a lot of sport

The information in the following parallel boxplots relates to Questions 5 and 6



The parallel boxplots above display the distribution of battery life (in hours) for two brands of batteries (Brand A and Brand B).

5 The variables *Battery life* and *Brand* are:

A both categorical variables
B a categorical and a numerical variable respectively
C a numerical and a categorical variable respectively
D both numerical variables
E neither a numerical nor a categorical variable

6 Which of the following statements (there may be more than one) support the contention that *Battery life* and *Brand* are related?

I the median battery life for Brand A is clearly higher than for Brand B
II battery lives for Brand B are more variable than Brand A
III the distribution of battery lives for Brand A is symmetric with outliers but positively skewed for Brand B

A I only B II only C III only D I and II only E I, II and III

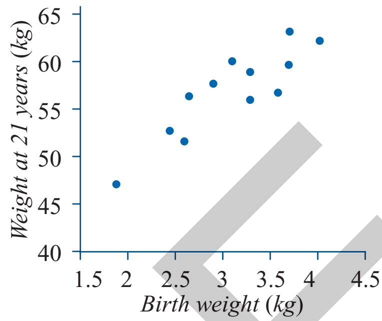
7 The relationship between weight at age 21 (in kg) and weight at birth (in kg) is to be investigated. In this investigation, the variables *Weight at age 21* and *Weight at birth* are:

A both categorical variables
B a categorical and a numerical variable respectively
C a numerical and a categorical variable respectively
D both numerical variables
E neither a numerical nor a categorical variable

Cambridge University Press • Uncorrected Sample pages • 978-0-521-61328-6 • 2008 © Jones, Evans, Lipson
TI-Nspire & Casio ClassPad material in collaboration with Brown and McMenamin

Review

8 The scatterplot opposite shows the weight at age 21 and weight at birth of 12 women. The relationship displayed is best described as a:



A weak positive linear relationship
B weak negative linear relationship
C moderate positive non-linear relationship
D strong positive non-linear relationship
E strong positive linear relationship

9 The variables *Response time* to a drug and *Drug dosage* are linearly related with r = −0.9.
From this information, we can conclude that:
A response times are −0.9 times the drug dosage
B that response times decrease with decreased drug dosage
C that response times decrease with increased drug dosage
D that response times increase with increased drug dosage
E response times are 81% of the drug dosage

10 The birth weight and weight at age 21 of eight women are given in the table below.

Birth weight	1.9	2.4	2.6	2.7	2.9	3.2	3.4	3.6
Weight at 21	47.6	53.1	52.2	56.2	57.6	59.9	55.3	56.7

The value of the correlation coefficient is closest to:
A 0.536 **B** 0.6182 **C** 0.7863 **D** 0.8232 **E** 0.8954

11 The value of a correlation coefficient is $r = -0.7685$. The value of the corresponding coefficient of determination is closest to:
A −0.7685 **B** −0.5906 **C** 0.2315 **D** 0.5906 **E** 0.7685

12 The relationship between heart weight and body weight in a group of mice is linearly related with a correlation coefficient of $r = 0.765$. Heart weight is the DV.
From this information, we can conclude that:
A 58.5% of the variation in heart weights can be explained by the variation in body weights
B 76.5% of the variation in heart weights can be explained by the variation in body weights
C heart weights are 58.5% of body weights
D heart weights are 76.5% of body weights
E 58.5% of the mice had heavy hearts

13 We wish to display the relationship between the variables *Weight* (in kg) of young children and *Level of nutrition* (poor, adequate, good). The most appropriate graphical display would be:
A a histogram **B** parallel box plots **C** a segmented bar chart
D a scatter plot **E** a back-to-back stem plot

14 We wish to display the relationship between the variables *Weight* (under-weight, normal, over-weight) of young children and *Level of nutrition* (poor, adequate, good). The most appropriate graphical display would be:

A a histogram **B** parallel box plots **C** a segmented bar chart
D a scatter plot **E** a back-to-back stem plot

15 There is a strong linear positive relationship ($r = 0.85$) between the amount of *Garbage recycled* and *Salary level*. From this information, we can conclude that:

A the amount of garbage recycled can be increased by increasing people's salaries
B the amount of garbage recycled can be increased by decreasing people's salaries
C increasing the amount of garbage you recycle will increase your salary
D people on high salaries tend to recycle less garbage
E people on high salaries tend to recycle more garbage

Extended-response questions

1 One thousand drivers who had an accident during the past year were classified according to age and the number of accidents.

Number of accidents	*Age < 30*	*Age ≥ 30*
At most one accident	130	170
More than one accident	470	230
Total	600	400

a What are the variables shown in the table? Are they categorical or numerical?
b Determine which is the dependent and which is the independent variable.
c How many drivers under the age of 30 had more than one accident?
d Percentage the cells in the table. Calculate column percentages.
e Use these percentages to comment on the statement: '*Younger drivers (age < 30) are more likely than older drivers (age ≥ 30) to have had more than one accident*.'

2 It was suggested that day and evening students differed in their satisfaction with a course in psychology. The following crosstabulation was obtained:

	Type of student	
Level of satisfaction with course	Day	Evening
Satisfied	90	22
Neutral	18	5
Dissatisfied	12	3
Total	120	30

a Name the dependent variable.
b How many students were involved?

Cambridge University Press • Uncorrected Sample pages • 978-0-521-61328-6 • 2008 © Jones, Evans, Lipson
TI-Nspire & Casio ClassPad material in collaboration with Brown and McMenamin

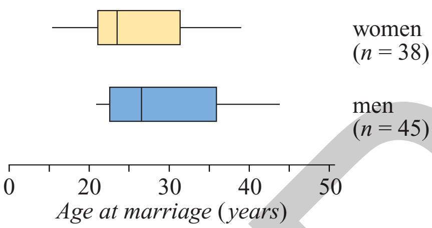
Back to Menu >>>

c Calculate the appropriate column percentages and write them down in an appropriate table.

d Does there appear to be a relationship between satisfaction with the course and the type of student in the sample? Fully explain your answer.

e Comment on the statement:
'*There was greater satisfaction with the psychology course among day students as 90 day students were satisfied with the course while only 22 evening students were satisfied.*'

3 The parallel box plots below compare the distribution of age at marriage of 45 married men and 38 married women.



a The two variables displayed here are *Age at marriage* and *Sex*. Which is the numerical and which is the categorical variable?

b Do the parallel box plots support the contention that the age a person marries depends on their sex? Explain why.

4 The data below gives the hourly pay rate (in dollars per hour) of 10 production-line workers along with their years of experience on initial appointment.

Rate (\$/h)	15.90	15.70	16.10	16.00	16.79	16.45	17.00	17.65	18.10	18.75
Experience (*yrs*)	1.25	1.50	2.00	2.00	2.75	4.00	5.00	6.00	8.00	12.00

a Use a calculator to construct a scatterplot of the data with *Rate* plotted on the vertical axis and *Experience* on the horizontal axis. Why has the vertical axis been used for rate?

b Comment on direction, outliers, form and strength of any relationship revealed.

c Determine the value of the correlation coefficient (r) correct to three decimal places.

d Determine the value of the coefficient of determination (r^2) and interpret.

5 A researcher noted that loss of sleep affected the number of dreams experienced by an individual. He also noted that as soon as people started to dream they exhibited rapid eye movement (REM). To examine this apparent relationship, he kept a group of volunteers awake for various lengths of time by reading them spicy chapters from a statistics book. After they fell asleep, he recorded the number of times REM occurred. The following data was obtained.

Hours of sleep deprivation	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
Number of times REM occurred	10	20	15	30	20	20	25	35

a Name the dependent variable in the study.

b Use a calculator to construct a scatterplot of the data. Name variables, *sleepdep* and *rem*.

c Does there appear to be a relationship between the variables? If so, is it positive or negative?

d Determine the value of r, the Pearson's correlation coefficient, correct to three decimal places. Comment on the nature of the relationship between the variables in this study.

e Calculate the coefficient of determination (r^2) and interpret.